Beginning Data Science in R

Data Analysis, Visualization, and Modelling for the Data Scientist

Thomas Mailund

Beginning Data Science in R

Data Analysis, Visualization, and Modelling for the Data Scientist

Thomas Mailund

Apress®

Beginning Data Science in R: Data Analysis, Visualization, and Modelling for the Data Scientist

Thomas Mailund
Aarhus, Denmark

Contents at a Glance

Contents

About the Author

Thomas Mailund is an associate professor in bioinformatics at Aarhus University, Denmark. His background is in math and computer science, but for the last decade, his main focus has been on genetics and evolutionary studies, particularly comparative genomics, speciation, and gene flow between emerging species.

About the Technical Reviewer

Andrew Moskowitz is a doctoral candidate in Quantitative Psychology at UCLA and self-employed statistical consultant. His quantitative research focuses mainly on hypothesis testing and effect sizes in mixed effects models. While at UCLA, Andrew has collaborated with a number of faculty, students, and enterprises to help them derive meaning from data across an array of fields ranging from psychological services and health care delivery to marketing.

Acknowledgments

I would like to thank Asger Hobolth for many useful comments on earlier versions of this manuscript. He helped me improve the writing and the presentation of the material.

Introduction

Welcome to *Introduction to Data Science with R*. This book was written as a set of lecture notes for two classes I teach, *Data Science: Visualization and Analysis* and *Data Science: Software Development and Testing*. The book is written to fit the structure of these classes, where each class consists of seven weeks of lectures and project work. This means that there are 14 chapters with the core material, where the first seven focus on data analysis and the last seven on developing reusable software for data science.

What Is Data Science?

Oh boy! That is a difficult question. I don't know if it is easy to find someone who is entirely sure what data science is, but I am pretty sure that it would be difficult to find two people with fewer than three opinions about it. It is certainly a popular buzzword, and everyone wants to have data scientists these days, so data science skills are useful to have on the CV. But what *is* it?

Since I can't really give you an agreed upon definition, I will just give you my own: *Data science is the science of learning from data*.

This is a very broad definition—almost too broad to be useful. I realize this. But then, I think data science is an incredibly general field. I don't have a problem with that. Of course, you could argue that any *science* is all about getting information out of data, and you might be right. Although I would say that there is more to science than just transforming raw data into useful information. The sciences are focusing on answering specific questions about the world while data science is focusing on how to manipulate data efficiently and effectively. The primary focus is not which questions to ask of the data but how we can answer them, whatever they may be. It is more like computer science and mathematics than it is like natural sciences, in this way. It isn't so much about studying the natural world as it is about how to compute data efficiently.

Included in data science is the design of experiments. With the right data, we can address the questions we are interested in. With a poor design of experiments or a poor choice of which data we gather, this can be difficult. Study design might be the most important aspect of data science, but is not the topic of this book. In this book I focus on the analysis of data, once gathered.

Computer science is also mainly the study of computations—as is hinted at in the name—but is a bit broader in this focus. Although *datalogy*, an earlier name for data science, was also suggested for computer science, and for example in Denmark it *is* the name for computer science, using the name "computer science" puts the focus on computation while using the name "data science" puts the focus on data. But of course, the fields overlap. If you are writing a sorting algorithm, are you then focusing on the computation or the data? Is that even a meaningful question to ask?

There is a huge overlap between computer science and data science and naturally the skillsets you need overlap as well. To efficiently manipulate data you need the tools for doing that, so computer programming skills are a must and some knowledge about algorithms and data structures usually is as well. For data science, though, the focus is always on the data. In a data analysis project, the focus is on how the data flows from its raw form through various manipulations until it is summarized in some useful form. Although the difference can be subtle, the focus is not about what operations a program does during the analysis, but about how the data flows and is transformed. It is also focused on *why* we do certain transformations of the

data, what purpose those changes serve, and how they help us gain knowledge about the data. It is as much about deciding what to do with the data as it is about how to do it efficiently.

Statistics is of course also closely related to data science. So closely linked, in fact, that many consider data science just a fancy word for statistics that looks slightly more modern and sexy. I can't say that I strongly disagree with this—data science *does* sound sexier than statistics—but just as data science is slightly different from computer science, data science is also slightly different from statistics. Just, perhaps, somewhat less different than computer science is.

A large part of doing statistics is building mathematical models for your data and fitting the models to the data to learn about the data in this way. That is also what we do in data science. As long as the focus is on the data, I am happy to call statistics data science. If the focus changes to the models and the mathematics, then we are drifting away from data science into something else—just as if the focus changes from the data to computations we are drifting from data science to computer science.

Data science is also related to machine learning and artificial intelligence, and again there are huge overlaps. Perhaps not surprising since something like machine learning has its home both in computer science and in statistics; if it is focusing on data analysis, it is also at home in data science. To be honest, it has never been clear to me when a mathematical model changes from being a plain old statistical model to becoming machine learning anyway.

For this book, we are just going to go with my definition and, as long as we are focusing on analyzing data, we are going to call it data science.

Prerequisites for Reading this Book

In the first seven chapters in this book, the focus is on data analysis and not programming. For those seven chapters, I do not assume a detailed familiarity with topics such as software design, algorithms, data structures, and such. I do not expect you to have any experience with the R programming language either. I do, however, expect that you have had *some* experience with programming, mathematical modeling, and statistics.

Programming R can be quite tricky at times if you are familiar with a scripting language or object-oriented languages. R is a functional language that does not allow you to modify data, and while it does have systems for object-oriented programming, it handles this programming paradigm very differently from languages you are likely to have seen such as Java or Python.

For the data analysis part of this book, the first seven chapters, we will only use R for very straightforward programming tasks, so none of this should pose a problem. We will have to write simple scripts for manipulating and summarizing data so you should be familiar with how to write basic expressions like function calls, `if` statements, loops, and so on. These things you will have to be comfortable with. I will introduce every such construction in the book when we need them so you will see how they are expressed in R, but I will not spend much time explaining them. I mostly will just expect you to be able to pick it up from examples.

Similarly, I do not expect you to know already how to fit data and compare models in R. I do expect that you have had enough introduction to statistics to be comfortable with basic terms like parameter estimation, model fitting, explanatory and response variables, and model comparison. If not, I expect you to be at least able to pick up what we are talking about when you need to.

I won't expect you to know a lot about statistics and programming, but this isn't *Data Science for Dummies,* so I do expect you to be able to figure out examples without me explaining everything in detail.

After the first seven chapters is a short description of a data analysis project, one of my students did in an earlier class. It shows how such a project could look, but I suggest that you do not wait until you have finished the first seven chapters to start doing such analysis yourself. To get the most benefit out of reading this book, you should be applying what you learn continuously. Already when you begin reading, I suggest that you find a dataset that you would be interested in finding out more about and then apply what you learn in each chapter to that data.

For the final seven chapters of the book, the focus *is* on programming. To read this part you should be familiar with object-oriented programming. I will explain how it is handled in R and how it differs from languages such as Python, Java or C++ but I expect you to be familiar with terms such as class hierarchies, inheritance, and polymorphic methods. I will not expect you to be already familiar with functional programming (but if you are, there should still be plenty to learn in those chapters if you are not already familiar with R programming as well).

Plan for the Book

In the book, we cover basic data manipulation—filtering and selecting relevant data; transforming data into shapes readily analyzable; summarizing data; visualizing data in informative ways both for exploring data and presenting results; and model building. These are the key aspects of doing analysis in data science. After this we will cover how to develop R code that is reusable and works well with existing packages, and that is easy to extend, and we will see how to build new R packages that other people will be able to use in their projects. These are the essential skills you will need to develop your own methods and share them with the world.

We will do all this using the programming language R (`https://www.r-project.org/about.html`). R is one of the most popular (and open source) data analysis programming languages around at the moment. Of course, popularity doesn't imply quality, but because R is so popular it has a rich ecosystem of extensions (called "packages" in R) for just about any kind of analysis you could be interested in. People who develop statistical methods often implement them as R packages, so you can quite often get the state of the art techniques very easily in R. The popularity also means that there is a large community of people who can help if you have problems. Most problems you run into can be solved with a few minutes on Google because you are unlikely to be the first to run into any particular issue. There are also plenty of online tutorials for learning more about R and specialized packages, there are plenty of videos with talks about R and popular R packages, and there are plenty of books you can buy if you want to learn more.

Data Analysis and Visualization

The topics focusing on data analysis and visualization are covered in the first seven chapters:

- Chapter 1, Introduction to R programming. In which you learn how to work with data and write data pipelines.

- Chapter 2, **Reproducible** analysis. In which you find out how to integrate documentation and analysis in a single document and how to use such documents to produce reproducible research.

- Chapter 3, Data manipulation. In which you learn how to import, tidy up, and transform data, and compute summaries from data.

- Chapter 4, **Visualizing** and exploring data. In which you learn how to make plots for exploring data features and for presenting data features and analysis results.

- Chapter 5, **Working** with large datasets. In which you learn how to deal with data where the number of observations make the usual approaches too slow.

- Chapter 6, **Supervised** learning. In which you learn how to train models when you have datasets with known classes or regression values.

- Chapter 7, **Unsupervised** learning. In which you learn how to search for patterns you are not aware of in data.

These chapters are followed by the first project, where you see the various techniques in use.

Software Development

Software and package development is then covered in the following seven chapters:

- Chapter 8, **More** R programming. In which you'll return to the basics of R programming and get a few more details than the tutorial in Chapter 1.

- Chapter 9, **Advanced** R **programming**. In which you explore more advanced features of the R programming language, in particular, functional programming.

- Chapter 10, **Object** oriented programming. In which you learn how R models object orientation and how you can use it to write more generic code.

- Chapter 11, B**uilding** an R package. In which you learn the necessary components of an R package and how to program your own.

- Chapter 12, T**esting** and checking. In which you learn techniques for testing your R code and checking the consistency of your R packages.

- Chapter 13, **Version** control. In which you learn how to manage code under version control and how to collaborate using GitHub.

- Chapter 14, **Profiling** and optimizing. In which you learn how to identify hotspots of code where inefficient solutions are slowing you down and techniques for alleviating this.

These chapters are then followed by the second project, where you'll build a package for Bayesian linear regression.

Getting R and RStudio

You will need to install R on your computer to do the exercises in this book. I suggest that you get an integrated environment since it can be slightly easier to keep track of a project when you have your plots, documentation, code, etc., all in the same program.

I personally use RStudio (https://www.rstudio.com/products/RStudio), which I warmly recommend. You can get it for free—just follow the link—and I will assume that you have it when I need to refer to the software environment you are using in the following chapters. There won't be much RStudio specifics, though, and most tools for working with R have the same features, so if you want to use something else you can probably follow the notes without any difficulties.

Projects

You cannot learn how to analyze data without analyzing data, and you cannot learn how to develop software without developing software either. Typing in examples from the book is nothing like writing code on your own. Even doing exercises from the book—which you really ought to do—is not the same as working on your own projects. Exercises, after all, cover small isolated aspects of problems you have just been introduced to. In the real world, there is not a chapter of material presented before every task you have to deal with. You need to work out by yourself what needs to be done and how. If you only do the exercises in this book, you will miss the most important lessons in analyzing data. How to explore the data and get a feeling for it; how to do the detective work necessary to pull out some understanding from the data; and how to deal with all the noise and weirdness found in any dataset. And for developing a package, you need to think through how to design and implement its functionality so that the various functions and data structures fit well together.

In this book, I go through a data analysis project to show you what that can look like. To actually learn how to analyze data, you need to do it yourself as well, and you need to do it with a dataset that I haven't analyzed for you. You might have a dataset lying around you have worked on before, a dataset from something you are just interested in, or you can probably find something interesting at a public data repository, e.g., one of these:

- RDataMining.com

- UCI machine learning repository (`http://archive.ics.uci.edu/ml/`)

- KDNuggets (`http://www.kdnuggets.com/datasets/index.html`)

- Reddit r/datasets (`https://www.reddit.com/r/datasets`)

- GitHub awesome public datasets (`https://github.com/caesar0301/awesome-public-datasets`)

I suggest that you find yourself a dataset and that after each lesson, you use the skills you have learned to explore this dataset. Pick data that is structured as a table with observations as rows and variables as columns, since that is the form of the data we consider in this book. At the end of the first seven chapters, you will have analyzed this data, you can write a report about your analysis that others can evaluate to follow and maybe modify it. You will be doing reproducible science.

For the programming topics, I describe another project illustrating the design and implementation issues involved in making an R package. There, you should be able to learn from just implementing your own version of the project I use, but you will, of course, be more challenged by working on a project without any of my help at all. Whichever you do, to get the full benefit of this book you should make your own package while reading the programming chapters.

CHAPTER 1

■ ■ ■ ■

Introduction to R Programming

We will use R for our data analysis so we need to know the basics of programming in the R language. R is a full programming language with both functional programming and object oriented programming features. Learning the language is far beyond the scope of this chapter and is something we return to later. The good news, though, is that to use R for data analysis, you rarely need to do much programming. At least, if you do the *right* kind of programming, you won't need much.

For manipulating data—and how to do this is the topic of the next chapter—you mainly just have to string together a couple of operations. Operations such as "group the data by this feature" followed by "calculate the mean value of these features within each group" and then "plot these means". This used to be much more complicated to do in R, but a couple of new ideas on how to structure such data flow—and some clever implementations of these in a couple of packages such as `magrittr` and `dplyr`—has significantly simplified it. We will see some of this at the end of this chapter and more in the next chapter. First, though, you need to get a taste for R.

Basic Interaction with R

Start by downloading RStudio if you haven't done so already (`https://www.rstudio.com/products/RStudio`). If you open it, you should see a window similar to Figure 1-1. Well, except that you will be in an empty project while the figure shows (on the top right) that this RStudio is opened in a project called "Data Science". You always want to be working on a project. Projects keep track of the state of your analysis by remembering variables and functions you have written and keep track of which files you have opened and such. Choose File ➤ New Project to create a project. You can create a project from an existing directory, but if this is the first time you are working with R you probably just want to create an empty project in a new directory, so do that.
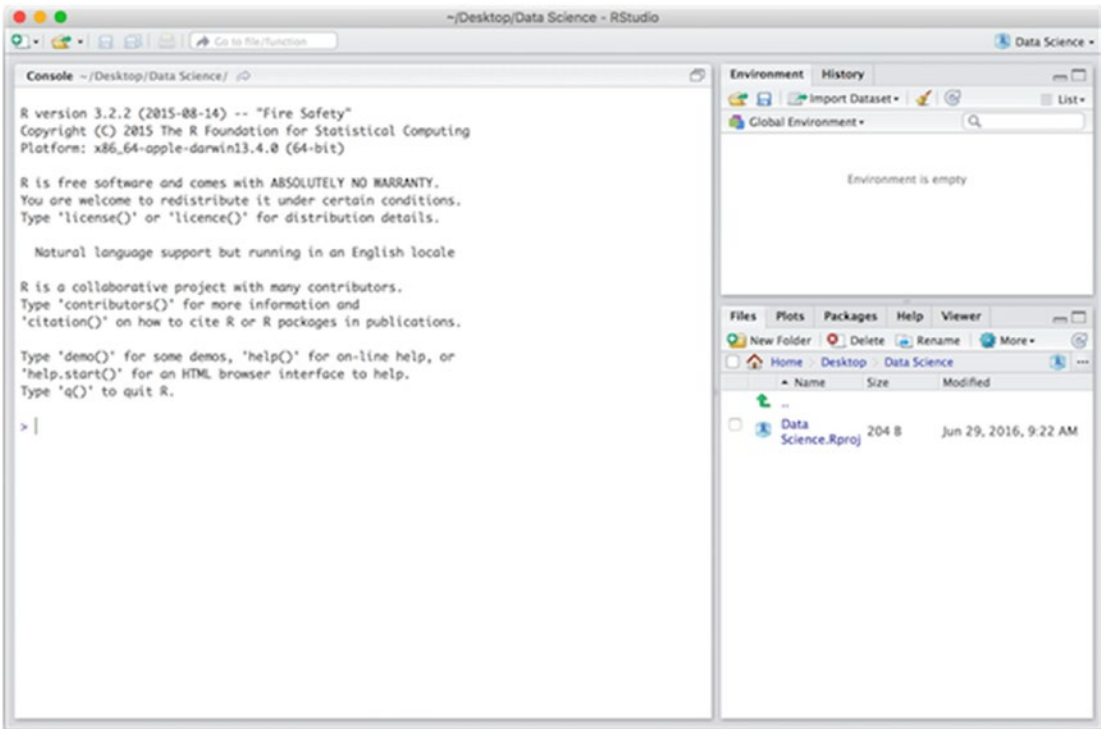
Figure 1-1. *RStudio*

Once you have opened RStudio, you can type R expressions into the console, which is the frame on the left of the RStudio window. When you write an expression there, R will read it, evaluate it, and print the result. When you assign values to variables, and you will see how to do this shortly, they will appear in the Environment frame on the top right. At the bottom right, you have the directory where the project lives, and files you create will go there.

To create a new file, choose File ➤ New File. You can select several different file types. We are interested in the *R* Script and R Markdown types. The former is the file type for pure R code, while the latter is used for creating reports where documentation text is mixed with R code. For data analysis projects, I recommend using Markdown files. Writing documentation for what you are doing is really helpful when you need to go back to a project several months down the line.

For most of this chapter, you can just write R code in the console, or you can create an R Script file. If you create an R Script file, it will show up on the top left, as shown in Figure 1-2. You can evaluate single expressions using the Run button on the top-right of this frame, or evaluate the entire file using the Source button. For longer expressions, you might want to write them in an R Script file for now. In the next chapter, we talk about R Markdown, which is the better solution for data science projects.
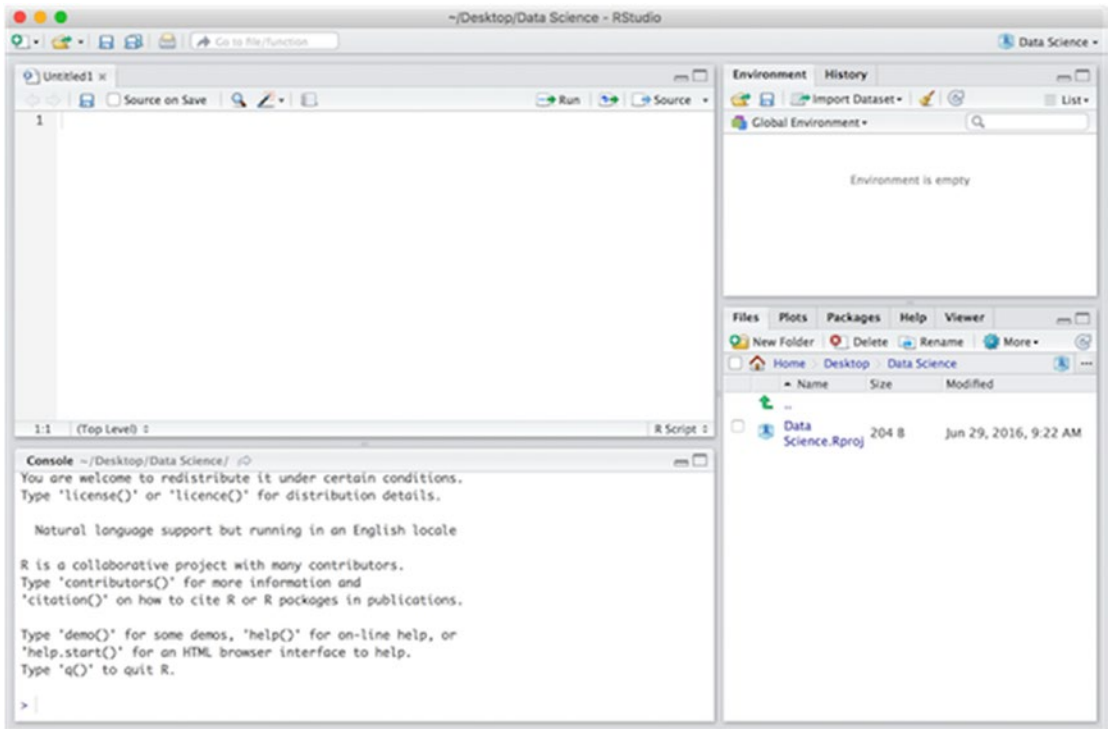
Figure 1-2. *RStudio with a new R Script file*

Using R as a Calculator

You can use the R console as a calculator where you just type in an expression you want calculated, press Enter, and R gives you the result. You can play around with that a little bit to get familiar with how to write expressions in R—there is some explanation for how to write them below—moving from using R as a calculator in this sense to writing more sophisticated analysis programs is only a question of degree. A data analysis program is really little more than a sequence of calculations, after all.

Simple Expressions

Simple arithmetic expressions are written, as in most other programming languages, in the typical mathematical notation that you are used to.

```
1 + 2
## [1] 3
4 / 2
## [1] 2
(2 + 2) * 3
## [1] 12
```

It also works pretty much as you are used to. Except, perhaps, that you might be used to integers behaving as integers in a division. At least in some programming languages, division between integers is integer division, but in R, you can divide integers and if there is a remainder you will get a floating-point number back as the result.

```
4 / 3
## [1] 1.333333
```

When you write numbers like 4 and 3, they are interpreted as floating-point numbers. To explicitly get an integer, you must write 4L and 3L.

```
class(4)
## [1] "numeric"
class(4L)
## [1] "integer"
```

You will still get a floating-point if you divide two integers, although there is no need to tell R explicitly that you want floating-point division. If you want integer division, on the other hand, you need a different operator, %/%:

```
4 %/% 3
## [1] 1
```

In many languages % is used to get the remainder of a division, but this doesn't quite work with R, where % is used to construct infix operators. So in R, the operator for this is %%:

```
4 %% 3
## [1] 1
```

In addition to the basic arithmetic operators—addition, subtraction, multiplication, division, and the modulus operator you just saw—you also have an exponentiation operator for taking powers. For this, you can use ^ or ** as infix operators:

```
2^2
## [1] 4
2^3
## [1] 8
2**2
## [1] 4
2**3
## [1] 8
```

There are some other data types besides numbers, but we won't go into an exhaustive list here. There are two types you do need to know about early, though, since they are frequently used and since *not* knowing about how they work can lead to all kinds of grief. Those are strings and "factors".

Strings work as you would expect. You write them in quotes, either double quotes or single quotes, and that is about it.

```
"hello,"
## [1] "hello,"
'world!'
## [1] "world!"
```

Strings are not particularly tricky, but I mention them because they look a lot like factors, but factors are *not* like strings, they just look sufficiently like them to cause some confusion. I explain factors a little later in this chapter when you have seen how functions and vectors work.

Assignments

To assign a value to a variable, you use the arrow operators. So you assign the value 2 to the variable x, you would write the following:

```
x <- 2
```

You can test that x now holds the value 2 by evaluating x.

```
x
## [1] 2
```

And of course, you can now use x in expressions:

```
2 * x
## [1] 4
```

You can assign with arrows in both directions, so you could also write the following:

```
2 -> x
```

An assignment won't print anything if you write it into the R terminal, but you can get R to print it just by putting the assignment in parentheses.

```
x <- "invisible"
(y <- "visible")
## [1] "visible"
```

Actually, All of the Above Are Vectors of Values…

If you were wondering why all the values printed above had a [1] in front of them, I am going to explain that right now. It is because we are usually not working with single values anywhere in R. We are working with vectors of values (and you will hear more about vectors in the next section). The vectors we have seen have length one—they consist of a single value—so there is nothing wrong about thinking about them as individual values. But they really are vectors.

The [1] does not indicate that we are looking at a vector of length one, though. The [1] tells you that the first value after [1] is the first value in the vector. With longer vectors, you get the index each time R moves to the next line of output. This is just done to make it easier to count your way into a particular index.

You will see this if you make a longer vector, for example, you can make one of length 50 using the : operator:

```
1:50
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
## [16] 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
## [31] 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45
## [46] 46 47 48 49 50
```

Because we are essentially always working on vectors, there is one caveat I want to warn you about. If you want to know the length of a string, you might—reasonably enough—think you can get that using the length function. You would be wrong. That function gives you the length of a vector, so if you give it a single string, it will always return 1.

```
length("qax")
## [1] 1
length("quux")
## [1] 1
length(c("foo", "barz"))
## [1] 2
```

In the last expression, we used the function c() to concatenate two strings. This creates a vector of two strings, and thus the result of calling length on that is 2. To get the length of the actual string, you want nchar instead:

```
nchar("qax")
## [1] 3
nchar("quux")
## [1] 4
nchar(c("foo", "barz"))
## [1] 3 4
```

Indexing Vectors

If you have a vector and want the i'th element of that vector, you can index the vector to get it like this:

```
(v <- 1:5)
## [1] 1 2 3 4 5
v[1]
## [1] 1
v[3]
## [1] 3
```

We have parentheses around the first expression to see the output of the operation. An assignment is usually silent in R, but by putting the expression in parentheses, we make sure that R prints the result, which is the vector of integers from 1 to 5. Notice here that the first element is at index 1. Many programming languages start indexing at 0, but R starts indexing at 1. A vector of length n is thus indexed from 1 to n, unlike in zero-indexed languages, where the indices go from 0 to $n-1$.

If you want to extract a subvector, you can also do this with indexing. You just use a vector of the indices you want inside the square brackets. You can use the : operator for this or the concatenate function, c():

```
v[1:3]
## [1] 1 2 3
v[c(1,3,5)]
## [1] 1 3 5
```

You can even use a vector of Boolean values to pick out those values that are "true":

```
v[c(TRUE, FALSE, TRUE, FALSE, TRUE)]
## [1] 1 3 5
```

This is particularly useful when you combine it with expressions. You can, for example, get a vector of Boolean values telling you which values of a vector are even numbers and then use that vector to pick them out.

```
v %% 2 == 0
## [1] FALSE  TRUE FALSE  TRUE FALSE
v[v %% 2 == 0]
## [1] 2 4
```

You can get the complement of a vector of indices if you just change the sign of them:

```
v[-(1:3)]
## [1] 4 5
```

It is also possible to give vector indices names and, if you do, you can use those to index the vector. You can set the names of a vector when constructing it or use the names() function.

```
v <- c("A" = 1, "B" = 2, "C" = 3)
v
## A B C
## 1 2 3
v["A"]
## A
## 1
names(v) <- c("x", "y", "z")
v
## x y z
## 1 2 3
v["x"]
## x
## 1
```

This can be very useful for making tables where you can look up a value by a key.

Vectorized Expressions

Now, the reason that the expressions you saw above worked with vector values instead of single values (although vectors containing only a single value) is that in R, arithmetic expressions actually all work component-wise on vectors. When you write an expression such as:

```
x ** 2 - y
```

You are actually telling R to take each element in the vector x, square it, and subtract it by the elements in y.

```
(x <- 1:3)
## [1] 1 2 3
x ** 2
## [1] 1 4 9
y <- 6:8
x ** 2 - y
## [1] -5 -3  1
```

7

This also works if the vectors have different lengths, as they actually do in the previous example. The vector 2 is a vector of length 1 containing the number 2. The way expressions work, when vectors do not have the same length, is that you just repeat the shorter vector as many times as you need.

```
(x <- 1:4)
## [1] 1 2 3 4
(y <- 1:2)
## [1] 1 2
x - y
## [1] 0 0 2 2
```

If the length of the longer vector is not a multiple of the length of the shorter, you get a warning. The expression still repeats the shorter vector a number of times, just not an integer number of times.

```
(x <- 1:4)
## [1] 1 2 3 4
(y <- 1:3)
## [1] 1 2 3
x - y
## Warning in x - y: longer object length is not a
## multiple of shorter object length
## [1] 0 0 0 3
```

Here, y is used once against the 1:3 part of x, and the first element of y is then used for the 4 in x.

Comments

You probably don't want to write comments when you are just interacting with the R terminal, but in your code you do. Comments are just everything that follows #. From a # to the end of the line, the R parser just skips the text.

```
# This is a comment.
```

If you write your analysis code in R Markdown documents, which we cover in the next chapter, you won't have much need for comments. In those kinds of files, you mix text and R code in a different way. But if you develop R code, you will have a need for it. So now you know how to write comments.

Functions

You have already seen the use of functions, although you probably didn't think much about it, when you saw expressions such as:

```
length("qax")
```

You didn't think about it because there wasn't anything surprising about it. You just use the usual mathematical notation for functions: $f(x)$. If you want to call a function, you simply use this notation and give the function its parameters in parentheses.

In R, you can also use the names of the parameters when calling a function, in addition to the positions. If you have a function $f(x,y)$ of two parameters, x and y, calling $f(5,10)$ means calling f with parameter x set to 5 and parameter y set to 10. In R, you can specify this explicitly, and these two function calls are equivalent:

```
f(5, 10)
f(x = 5, y = 10)
```

If you specify the names of the parameters, the order doesn't matter anymore, so another equivalent function call would be:

```
f(y = 10, x = 5)
```

You can combine the two ways of passing parameters to functions as long as you put all the positional parameters before the named ones.

```
f(5, y = 10)
```

Except for maybe making the code slightly more readable—it is usually easier to remember what parameters do than which order they come in—there is not much need for this by itself. Where it really becomes useful is when combined with default parameters.

A lot of functions in R take many parameters. More than you really can remember the use for and definitely the order of. They are a lot like programs that take a lot of options but where you usually just use the defaults unless you really need to tweak an option. These functions take a lot of parameters, but most of them have useful default values, and you usually do not have to specify the values to set them to. When you *do* need it, though, you can specify it with a named parameter.

Getting Documentation for Functions

Since it can easily be hard to remember the details of what a function does, and especially what all the parameters to a function do, you often have to look up the documentation for functions. Luckily, this is very easy to do in R and in RStudio. Whenever you want to know what a function does, you can just ask R, and it will tell you (assuming that the author of the function has written the documentation).

Take the function length from the example you saw earlier. If you want to know what the function does, just write ?length in the R terminal. If you do this in RStudio, it will show you the documentation in the frame on the right, as shown in Figure 1-3.
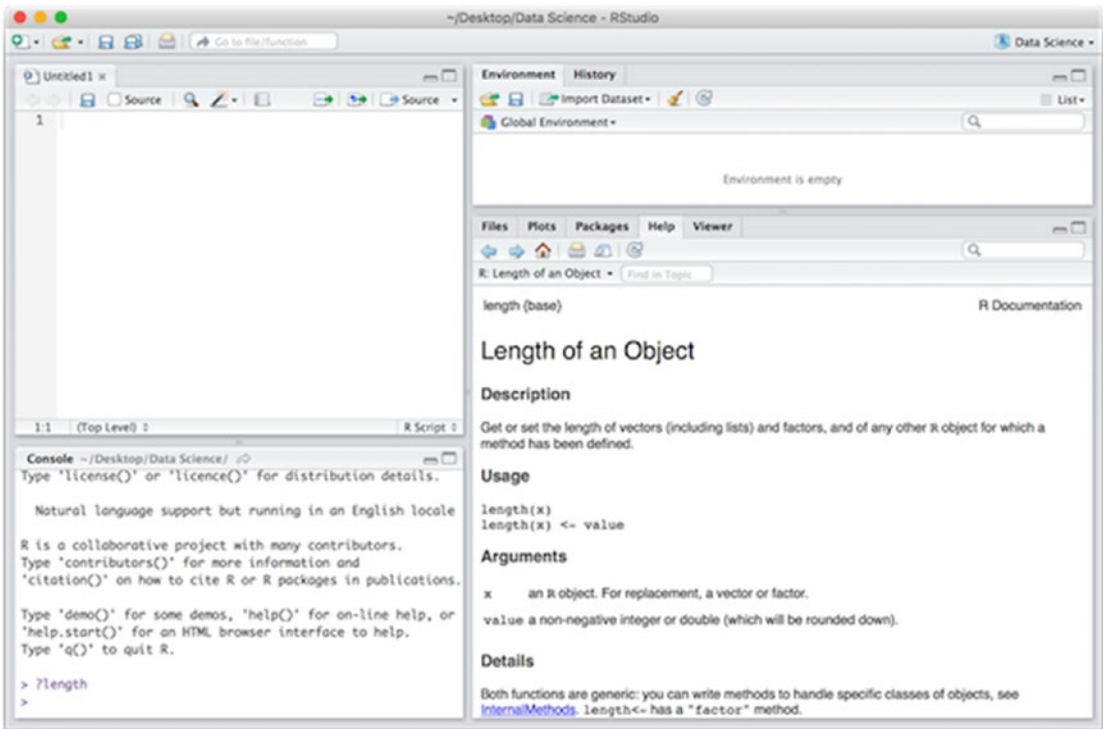
Figure 1-3. *RStudio's help frame*

Try looking up the documentation for a few functions. For example, the `nchar` function you also saw previously.

All infix operators, like + or %%, are also functions in R, and you can read the documentation for them as well. But you cannot write ?+ in the R terminal and get the information. The R parser doesn't know how to deal with that. If you want help on an infix operator, you need to quote it, and you do that using backquotes. So to read the documentation for +, you need to write:

```
?`+`
```

You probably do not need help to figure out what addition does, but people can write their own infix operators, so this is useful to know when you need help on those.

Writing Your Own Functions

You can easily write your own functions. You use `function` expressions to define a function and an assignment to give a function a name. For example, to write a function that computes the square of a number, you can write:

```
square <- function(x) x**2
square(1:4)
## [1]  1  4  9 16
```

The function(x) x**2 expression defines the function, and anywhere you need a function, you can write the function explicitly like this. Assigning the function to a name lets you use the name to refer to the function, just like assigning any other value, like a number or a string to a name, will let you use the name for the value.

Functions you write yourself works just like any function already part of R or part of an R package. With one exception, though: you will not have documentation for your own functions unless you write it, and that is beyond the scope of this chapter (but covered in Chapter 11).

The square function just does a simple arithmetic operation on its input. Sometimes you want the function to do more than a single thing. If you want the function to do several operations on its input, you need several statements for the function, and in that case you need to give it a "body" of several statements, and such a body has to go in curly brackets.

```
square_and_subtract <- function(x, y) {
    squared <- x ** 2
    squared - y
}
square_and_subtract(1:5, rev(1:5))
## [1] -4  0  6 14 24
```

(Check the documentation for rev to see what is going on here. Make sure you understand what this example is doing.)

In this simple example, we didn't really need several statements. We could just have written the function as:

```
square_and_subtract <- function(x, y) x ** 2 - y
```

As long as there is only a single expression in the function, we don't need the curly brackets. For more complex functions you will need it, though.

The result of a function—what it returns as its value when you call it—is the last statement or expression (there really isn't any difference between statements and expressions in R; they are the same thing). You can make the return value explicit, though, using the return() expression.

```
square_and_subtract <- function(x, y) return(x ** 2 - y)
```

This is usually only used when you want to return a value before the end of the function—and to see examples of this, you really need control structures, so you will have to wait a little bit to see an example—so it isn't used as much as in many other languages.

One important point here, though, if you are used to programming in other languages: the return() expression needs to include the parentheses. In most programming languages, you could just write:

```
square_and_subtract <- function(x, y) return x ** 2 - y
```

This doesn't work for R. Try it, and you will get an error.

Vectorized Expressions and Functions

Many functions work with vectorized expressions just as arithmetic expressions. In fact, any function you write that is defined just using such expressions will work on vectors, just like the square function.

This doesn't always work. Not all functions take a single value and return a single value, and in those cases, you cannot use them in vectorized expressions. Take for example the function sum, which adds all the values in a vector you give it as an argument (check ?sum now to see the documentation).

```
sum(1:4)
## [1] 10
```

This function summarizes its input into a single value. There are many similar functions, and naturally, these cannot be used element-wise on vectors.

Whether a function works on vector expressions or not depends on how it is defined. Most functions in R either work on vectors or summarizes vectors like sum. When you write your own functions, whether the function works element-wise on vectors or not depends on what you put in the body of the function. If you write a function that just does arithmetic on the input, like square, it will work in vectorized expressions. If you write a function that does some summary of the data, it will not. For example, if we write a function to compute the average of its input like this:

```
average <- function(x) {
    n <- length(x)
    sum(x) / n
}
average(1:5)
## [1] 3
```

This function will not give you values element-wise. Pretty obvious, really. It gets a little more complicated when the function you write contains control structures, which we will get to in the next section. In any case, this would be a nicer implementation since it only involves one expression:

```
average <- function(x) sum(x) / length(x)
```

Oh, one more thing: don't use this average function to compute the mean value of a vector. R already has a function for that, mean, that deals much better with special cases like missing data and vectors of length zero. Check out ?mean.

A Quick Look at Control Structures

While you get very far just using expressions, for many computations you need more complex programming. Not that it is particularly complex, but you do need to be able to select a choice of what to do based on data—selection or if statements—and ways of iterating through data—looping or for statements.

If statements work like this:

```
if (<boolean expression>) <expression>
```

If the Boolean expression evaluates to true, the expression is evaluated; if not, it will not.

```
# this won't do anything
if (2 > 3) "false"
# this will
if (3 > 2) "true"
## [1] "true"
```

For expressions like these, where we do not alter the program state by evaluating the expression, there isn't much of an effect in evaluating the if expression. If we, for example, assign it to a variable, there will be an effect.

```
x <- "foo"
if (2 > 3) x <- "bar"
x
## [1] "foo"
if (3 > 2) x <- "baz"
x
## [1] "baz"
```

If you want to have effects for both true and false expressions, you have this:

```
if (<boolean expression>) <true expression> else <false expression>
if (2 > 3) "bar" else "baz"
## [1] "baz"
```

If you want newlines in if statements, whether you have an else part or not, you need curly brackets. This won't work:

```
if (2 > 3)
    x <- "bar"
```

But this will:

```
if (2 > 3) {
    x <- "bar"
}
```

An if statement works like an expression.

```
if (2 > 3) "bar" else "baz"
```

This evaluates to the result of the expression in the if or the else part.

```
x <- if (2 > 3) "bar" else "baz"
x
## [1] "baz"
```

You cannot use it for vectorized expressions, though, since the Boolean expression, if you give it a vector, will evaluate the first element in the vector:

```
x <- 1:5
if (x > 3) "bar" else "baz"
## Warning in if (x > 3) "bar" else "baz": the
## condition has length > 1 and only the first
## element will be used
## [1] "baz"
```

If you want a vectorized version of if statements, you can instead use the ifelse function:

```
x <- 1:5
ifelse(x > 3, "bar", "baz")
## [1] "baz" "baz" "baz" "bar" "bar"
```

(Read the ?ifelse documentation to get the details of this function.)

This, of course, also has consequences for writing functions that use if statements. If your function contains a body that isn't vectorized, your function won't be either.

```
maybe_square <- function(x) {
    if (x %% 2 == 0) {
        x ** 2
    } else {
        x
    }
}
maybe_square(1:5)
## Warning in if (x%%2 == 0) {: the condition has
## length > 1 and only the first element will be used
## [1] 1 2 3 4 5
```

If you want a vectorized function, you need to use ifelse(), or you can use the Vectorize() function to translate a function that isn't vectorized into one that is.

```
maybe_square <- function(x) {
    ifelse (x %% 2 == 0, x ** 2, x)
}
maybe_square(1:5)
## [1]  1  4  3 16  5
maybe_square <- function(x) {
    if (x %% 2 == 0) {
        x ** 2
    } else {
        x
    }
}
maybe_square <- Vectorize(maybe_square)
maybe_square(1:5)
## [1]  1  4  3 16  5
```

The `Vectorize` function is what is known as a "functor"—a function that takes a function as input and returns a function. It is beyond the scope of this chapter to cover how functions can be manipulated like other data, but it is a very powerful feature of R that we return to in Chapter 9 about advanced R programming.

To loop over elements in a vector, you use `for` statements.

```
x <- 1:5
total <- 0
for (element in x) total <- total + element
total
## [1] 15
```

As with `if` statements, if you want the body to contain more than one expression, you need to put it in curly brackets.

The `for` statement runs through the elements of a vector. If you want the indices instead, you can use the `seq_along()` function, which, when given a vector as input, returns a vector of indices.

```
x <- 1:5
total <- 0
for (index in seq_along(x)) {
    element <- x[index]
    total <- total + element
}
total
## [1] 15
```

There are also `while` statements for looping. These repeat as long as an expression is true.

```
x <- 1:5
total <- 0
index <- 1
while (index <= length(x)) {
    element <- x[index]
    index <- index + 1
    total <- total + element
}
total
## [1] 15
```

If you are used to zero-indexed vectors, pay attention to the `index <= length(x)` here. You would normally write `index < length(x)` in zero-indexed languages. Here that would miss the last element.

There is also a `repeat` statement that loops until you explicitly exit using the `break` statement.

```
x <- 1:5
total <- 0
index <- 1
repeat {
  element <- x[index]
  total <- total + element
  index <- index + 1
```

```
  if (index > length(x)) break
}
total
## [1] 15
```

There is also a next statement that makes the loop jump to the next iteration.

Now that I have told you about loops, I feel I should also say that they generally are not used as much in R as in many other programming languages. Many actively discourage using loops, and they have a reputation for leading to slow code. The latter is not justified in itself, but it is easier to write slow code using loops than the alternative. Instead, you use functions to take over the looping functionality. There is usually a function for doing whatever you want to accomplish using a loop and when there is not, you can usually get what you want by combining the three functions—Map, Filter, and Reduce.

But that is beyond the scope of this chapter; we return to it later in the book.

Factors

Now let us return to data types and the factors I hinted at earlier. Factors are essentially just vectors, but of categorical values. That just means that the elements of a factor should be considered as categories or classes and not as numbers. For example categories such as "small", "medium", and "large" *could* be encoded as numbers but there isn't really any natural numbers to assign to them. We could encode soft drink sizes like 1, 2, and 3 for "small", "medium", and "large". By doing this, we are implicitly saying that the difference between "small" and "medium" is half of the difference between "small" and "large", which may not be the case. Data with sizes "small", "medium", and "large" should be encoded as categorical data, not numbers, and in R that means encoding them as factors.

A factor is usually constructed by giving it a list of strings. These are translated into the different categories, and the factor becomes a vector of these categories.

```
f <- factor(c("small", "small", "medium",
              "large", "small", "large"))
f
## [1] small  small  medium large  small  large
## Levels: large medium small
```

The categories are called "levels".

```
levels(f)
## [1] "large"  "medium" "small"
```

By default, these are ordered alphabetically, which in this example gives us the order "large", "medium", "small". You can change this order by specifying the levels when you create the factor.

```
ff <- factor(c("small", "small", "medium",
               "large", "small", "large"),
             levels = c("small", "medium", "large"))
ff
## [1] small  small  medium large  small  large
## Levels: small medium large
```

Changing the order of the levels like this changes how many functions handle the factor. The order of factor levels mostly affects how summary information is printed and how factors are plotted.

```
summary(f)
## large medium  small
##     2      1      3
summary(ff)
##  small medium  large
##      3      1      2
```

The order the levels are given shouldn't be an "ordering" of the categories, though. It is just used for displaying results; there is not an order semantics given to the levels unless you explicitly specify one.

Some categorical data has a natural order. Like "small", "medium" and "large". Other categories are not naturally ordered. There is no natural way of ordering "red", "green", and "blue". When we print data, it will always come out ordered since text always comes out ordered. When we plot data, it is usually also ordered. But in many mathematical models, we would treat ordered categorical data different from unordered categorical data, so the distinction is sometimes important.

By default, factors do not treat the levels as ordered, so they assume that categorical data is like "red", "green", and "blue", rather than ordered like "small", "medium", and "large". If you want to specify that the levels are actually ordered, you can do that using the ordered argument with the factor() function.

```
of <- factor(c("small", "small", "medium",
               "large", "small", "large"),
             levels = c("small", "medium", "large"),
             ordered = TRUE)
of
## [1] small  small  medium large  small  large
## Levels: small < medium < large
```

You can also use the ordered() function:

```
ordered(ff)
## [1] small  small  medium large  small  large
## Levels: small < medium < large
ordered(f, levels = c("small", "medium", "large"))
## [1] small  small  medium large  small  large
## Levels: small < medium < large
```

A factor is actually not stored as strings, even though we create it from a vector of strings. It is stored as a vector of integers where the integers are indices into the levels. This can bite you if you try to use a factor to index with.

Read the following code carefully. We have the vector v that can be indexed with the letters A, B, C, and D. We create a factor, ff, that consists of these four letters in that order. When we index with it, we get what we would expect. Since ff is the letters A to D, we pick out the values from v with those labels and in that order.

```
v <- 1:4
names(v) <- LETTERS[1:4]
v
## A B C D
## 1 2 3 4
(ff <- factor(LETTERS[1:4]))
```

```
## [1] A B C D
## Levels: A B C D
v[ff]
## A B C D
## 1 2 3 4
```

We are lucky to get the expected result, though. Because this expression is *not* indexing using the names we might expect it to use. Read the following even more carefully!

```
(ff <- factor(LETTERS[1:4], levels = rev(LETTERS[1:4])))
## [1] A B C D
## Levels: D C B A
v[ff]
## D C B A
## 4 3 2 1
```

This time ff is still a vector with the categories A to D in that order, but we have specified that the levels are D, C, B, and A, in *that* order. So the numerical values that the categories are stored as are actually these:

```
as.numeric(ff)
## [1] 4 3 2 1
```

What we get when we use it to index into v are *those* numerical indices, so we get the values pulled out of v in the reversed order from what we would expect if we didn't know this (which you now know).

The easiest way to deal with a factor as the actual labels it has is to translate it into a vector of strings. You can use that vector to index:

```
as.vector(ff)
## [1] "A" "B" "C" "D"
v[as.vector(ff)]
## A B C D
## 1 2 3 4
```

If you ever find yourself using a factor to index something—or in any other way treat a factor as if it was a vector of strings—you really should stop and make sure that you explicitly convert it into a vector of strings. Treating a factor as if it were a vector of strings, when in fact, it is a vector of integers, only leads to tears and suffering in the long run.

Data Frames

The vectors we have seen, whatever their type, are just sequences of data. There is no structure to them except for the sequence order, which may or may not be relevant for how to interpret the data. That is not how data we want to analyze looks. What we usually have is several variables that are related as part of the same observations. For each observed data point, you have a value for each of these variables (or missing data indications if some variables were not observed). Essentially, what you have is a table with a row per observation and a column per variable. The data type for such tables in R is the data.frame.

A data frame is a collection of vectors, where all must be of the same length, and you treat it as a two-dimensional table. We usually think of data frames as having each row correspond to some observation and each column to some property of the observations. Treating data frames that way makes them extremely useful for statistical modeling and fitting.

You can create a data frame explicitly using the data.frame function, but usually you will read in the data frame from files.

```
df <- data.frame(a = 1:4, b = letters[1:4])
df
##   a b
## 1 1 a
## 2 2 b
## 3 3 c
## 4 4 d
```

To get to the individual elements in a data frame, you must index it. Since it is a two-dimensional data structure, you should give it two indices.

```
df[1,1]
## [1] 1
```

You can, however, leave one of these empty, in which case you get an entire column or an entire row.

```
df[1,]
##   a b
## 1 1 a
df[,1]
## [1] 1 2 3 4
```

If the rows or columns are named, you can also use the names to index. This is mostly used for column names since it is the columns that correspond to the observed variables in a dataset. There are two ways to get to a column, but explicitly indexing:

```
df[,"a"]
## [1] 1 2 3 4
```

Or using the $column_name notation that does the same thing but lets you get at a column without having to use the [] operation and quote the name of a column.

```
df$b
## [1] a b c d
## Levels: a b c d
```

By default, a data frame will consider a character vector as a factor, and you need to tell it explicitly not to if you want a character vector.

```
df <- data.frame(a = 1:4, b = letters[1:4], stringsAsFactors = FALSE)
```

Functions for reading in data from various text formats will typically also convert string vectors to factors, and you need to prevent this explicitly. The readr package (see https://github.com/hadley/readr) is a notable exception where the default is to treat character vectors as character vectors.

You can combine two data frames row-wise or column-wise by using the `rbind` and `cbind` functions:

```
df2 <- data.frame(a = 5:7, b = letters[5:7])
rbind(df, df2)
##   a b
## 1 1 a
## 2 2 b
## 3 3 c
## 4 4 d
## 5 5 e
## 6 6 f
## 7 7 g
df3 <- data.frame(c = 5:8, d = letters[5:8])
cbind(df, df3)
##   a b c d
## 1 1 a 5 e
## 2 2 b 6 f
## 3 3 c 7 g
## 4 4 d 8 h
```

For more sophisticated manipulation of data frames, you really should use the dplyr package (see https://github.com/hadley/dplyr) or similar. We return to this in Chapter 3.

Dealing with Missing Values

Most datasets have missing values—parameters that weren't observed or that were incorrectly recorded and had to be masked out. How you deal with missing data in an analysis depends on the data and the analysis, but it must be addressed, even if all you do is remove all observations with missing data.

Missing data is represented in R by the special value NA (not available). Values of any type can be missing and represented as NA, and importantly R knows that NA means missing values and treats NAs accordingly. You should always represent missing data as NA instead of some special number (like -1 or 999 or whatever). R knows how to work with NA but has no way of knowing that -1 means anything besides minus one.

Operations that involve NA are themselves NA. You cannot operate on missing data and get anything but more missing values in return. This also means that if you compare two NAs, you get NA. Because NA is missing information, it is not even equal to itself.

```
NA + 5
## [1] NA
NA == NA
## [1] NA
NA != NA
## [1] NA
```

If you want to check if a value is missing, you must use the function `is.na`:

```
is.na(NA)
## [1] TRUE
is.na(4)
## [1] FALSE
```

Functions such as sum() will by default return NA if its input contains NA:

```
v <- c(1,NA,2)
sum(v)
## [1] NA
```
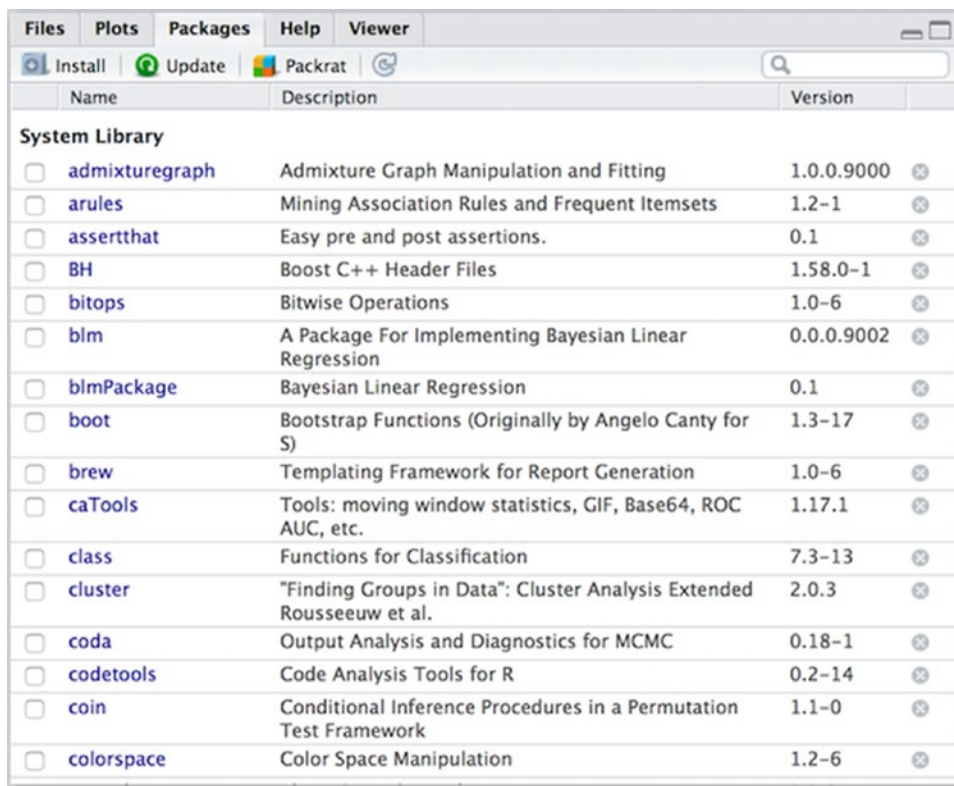
If you want just to ignore the NA values, there is often a parameter for specifying this:

```
sum(v, na.rm = TRUE)
## [1] 3
```

Using R Packages

Out of the box, R has a lot of functionality but where the real power comes in is through its package mechanism and the large collection of packages available for download and use.

When you install RStudio, it also installs a set of default packages. You can see which packages are installed by clicking on the Packages tab in the lower-right frame, as shown in Figure 1-4.

Figure 1-4. *RStudio packages*

From here you can update packages—new versions of important packages are regularly released—and you can install new packages. Try installing the package magrittr (see https://github.com/smbache/magrittr). We are going to use it shortly.

You can also install packages from the R console. Just write:

```
install.packages("magrittr")
```

Once you have installed a package, you have access to the functionality in it. You can get function f in package by writing package::f() or you can load all functions from a package into your global namespace to access them without using the package:: prefix.

Loading the functionality from the magrittr package is done like this:

```
library(magrittr)
```

Data Pipelines (or Pointless Programming)

Most data analysis consists of reading in some data, performing some operations on that data and, in the process, transforming it from its raw form into something we can start to make meaning out of. Then you do some summarizing or visualization toward the end.

These steps in an analysis are typically expressed as a sequence of function calls that each change the data from one form to another. It could look like this pseudo-code:

```
my_data <- read_data("/some/path/some_file.data")
clean_data <- remove_dodgy_data(my_data)
data_summaries <- summarize(clean_data)
plot_important_things(data_summaries)
```

There isn't really anything wrong with writing a data analysis in this way. But there are typically many more steps involved than just these. When there are, you either have to get very inventive in naming the variables you are saving data in or you have to overwrite variable names by reassigning to a variable after modifying the data. Both having many variable names and reassigning to variables can be problematic.

If you have may variables, it is easier accidentally to call a function on the wrong variable. For example, you might summarize the my_data variable instead of the clean_data. While you would get an error if you called a function with a variable name that doesn't exist, you won't necessarily get a simple error. If you just call a function with incorrect data, you might not notice it, but it would probably give you the wrong result. It would not be an error easy to debug later.

There is slightly less of a problem with reassigning to a variable. It is mostly an issue when you work with R interactively. There, if you want to go back and change part of the program you are writing, you have to go all the way back to the start, where the data is imported. You cannot just start somewhere in the middle of the function calls with a variable that doesn't refer to the same data as it did when you ran the program from scratch. It is less of a problem if you always run your R scripts from the beginning, but the typical use of R is to work with it in an interactive console or Markdown document, and this can be a problem.

A solution, then, is not to call the functions one at a time and assign each temporary result to a variable. Instead of having four statements in the previous example, one per function call, you would just feed the result of the first function call into the next.

```
plot_important_things(
    summarize(
        remove_dodgy_data(
            read_data("/some/path/some_file.data"))))
```

You get rid of all the variables, but the readability suffers, to put it mildly. You have to read the code from right to left and inside out.

Writing Pipelines of Function Calls

The magrittr package implements a trick to alleviate this problem. It does this by introducing a "pipe operator," %>%, that lets you write the functions you want to combine from left to right. You get the same effect as if you were calling one after the other and sending the result from one function to the input of the next function.

The operator works such that writing:

```
x %>% f
```

Is equivalent to writing:

```
f(x)
```

And writing:

```
x %>% f %>% g %>% h
```

Is equivalent to writing:

```
h(g(f(x)))
```

The previous example would become:

```
read_data("/some/path/some_file.data") %>%
    remove_dodgy_data %>%
    summarize %>%
    plot_important_things
```

Reading code like this might still take some getting used to, but it is much easier to read than combining functions from the inside and out.

If you have ever used pipelines in UNIX shells, you should immediately see the similarities. It is the same approach to combining functions/programs. By combining several functions, each of which does something relatively simple, you can create very powerful pipelines.

Writing pipelines using the %>% operator is a relatively new idiom introduced to R programming, but one that is very powerful and is being used more and more in different R packages.

Incidentally, if you are wondering why the package that implements pipes in R is called magrittr, it refers to Belgian artist René Magritte who famously painted a pipe and wrote: "Ceci n'est pas une pipe" ("This is not a pipe") below it. But enough about Belgian surrealists.

Writing Functions that Work with Pipelines

The %>% operator actually does something very simple, which in turn makes it simple to write new functions that work well with it. It simply takes whatever is computed on the left side of it and inserts it as the first argument to the function given on the right side, and it does this left to right. So x %>% f becomes f(x), x %>% f %>% g becomes f(x) %>% g and then g(f(x)), and x %>% f(y) becomes f(x,y). If you are already providing parameters to a function in the pipeline, the left side of %>% is just inserted before those parameters in the pipeline.

If you want to write functions that work well with pipelines, you should, therefore, make sure that the most likely parameter to come through a pipeline is the first parameter of your function. Write your functions so the first parameter is the data it operates on, and you have done most of the work.

For example, if you wanted a function that would sample n random rows of a data frame, you could write it such that it takes the data frame as the first argument and the parameter n as its second argument and then you could simply pop it right into a pipeline:

```
subsample_rows <- function(d, n) {
  rows <- sample(nrow(d), n)
  d[rows,]
}

d <- data.frame(x = rnorm(100), y = rnorm(100))
d %>% subsample_rows(n = 3)
##              x          y
## 46  0.5622234 -0.4184033
## 17 -0.5973131 -1.5549958
## 38 -2.0004727 -1.0736909
```

The magical "." argument

Now, you cannot always be so lucky that all the functions you want to call in a pipeline take the left side of the %>% as its first parameter. If this is the case, you can still use the function, though, because magrittr interprets . in a special way. If you use . in a function call in a pipeline, then that is where the left side of the %>% operation goes instead of as default first parameter of the right side. So if you need the data to go as the second parameter, you put a . there, since x %>% f(y, .) is equivalent to f(y, x). The same goes when you need to provide the left side as a named parameter since x %>% f(y, z = .) is equivalent to f(y, z = x), something that is particularly useful when the left side should be given to a model-fitting function. Functions fitting a model to data are usually taking a model specification as their first parameter and the data they are fitting as a named parameter called data.

```
d <- data.frame(x = rnorm(10), y = rnorm(10))
d %>% lm(y ~ x, data = .)
##
## Call:
## lm(formula = y ~ x, data = .)
##
## Coefficients:
## (Intercept)            x
##      0.0899       0.1469
```

We return to model fitting, and what an expression such as y ~ x means, in a later chapter, so don't worry if it looks a little strange for now. If you are interested, you can always check the documentation of the lm() function.

The magrittr package does more with `.` than just changing the order of parameters. You can use `.` more than once when calling a function and you can use it in expressions or in function calls:

```
rnorm(4) %>% data.frame(x = ., is_negative = . < 0)
##           x is_negative
## 1 -0.6187822        TRUE
## 2 -1.5446573        TRUE
## 3 -2.4387665        TRUE
## 4 -1.7097824        TRUE
rnorm(4) %>% data.frame(x = ., y = abs(.))
##            x         y
## 1  1.5754641 1.5754641
## 2 -0.2162109 0.2162109
## 3 -0.1151102 0.1151102
## 4 -0.4561123 0.4561123
```

There is one caveat: If `.` *only* appears in function calls, it will still be given as the first expression to the function on the right side of %>%.

```
rnorm(4) %>% data.frame(x = sin(.), y = cos(.))
##          .          x          y
## 1 -1.471748 -0.9950987  0.09888622
## 2 -1.732544 -0.9869474 -0.16104285
## 3  0.642917  0.5995326  0.80035036
## 4  2.081730  0.8722884 -0.48899182
```

The reason is that it is more common to see expressions with function calls like this when the full data is also needed than when it is not. So by default, `f(g(.),h(.))` gets translated into `f(.,g(.),h(.))`. If you want to avoid this behavior, you can put curly brackets around the function call, since `{f(g(.),h(.))}` is equivalent to `f(g(.),h(.))`. (The meaning of the curly brackets is explained soon.) You can get both the behavior `f(.,g(.),h(.))` and the behavior `{f(g(.),h(.))}` in function calls in a pipeline; the default is just the most common case.

Defining Functions Using .

While `.` is mainly used for providing parameters to functions in a pipeline, it can also be used as a short-hand for defining new functions. Writing:

```
. %>% f
```

Is equivalent to writing:

```
function(.) f(.)
```

And is a quick way of defining a function as a combination of other functions. For example:

```
f <- . %>% cos %>% sin
```

Is equivalent to

```
f <- function(.) sin(cos(.))
```

25

Defining functions from combining other functions is called "tacit" or "point-free" programming (or sometimes even pointless programming, although that is a little harsh), referring to the way you are not storing the intermediate steps (points) of a computation. You write:

```
f <- . %>% cos %>% sin
```

Instead of:

```
f <- function(x) {
    y <- cos(x)
    z <- sin(y)
    z
}
```

Naturally, this is mostly used when you have a sub-pipeline that you intend to call on more than one dataset. You can just write a function specifying the pipeline like you would write an actual pipeline. You just give it . as the very first left side, instead of a dataset, and you are defining a function instead of running data through a pipeline.

Anonymous Functions

Pipelines are great when you can call existing functions one after another, but what happens if you need a step in the pipeline where there is no function doing what you want? You can, of course, always write such a missing function but if you need to write functions time and time again for doing small tasks in pipelines, you have a similar problem to when you needed to save all the intermediate steps in an analysis in variables. You do not want to have a lot of functions defined because there is a risk that you use the wrong one in a pipeline—especially if you have many similar functions, as you are likely to have if you need a function for every time you need a little bit of data-tweaking in your pipelines.

Again, magrittr has the solution: lambda expressions. This is a computer science term for anonymous functions, that is, functions that you do not give a name.

When you define a function in R, you actually always create an anonymous function. Any expression of the form function(x) expression is a function, but it doesn't have a name unless you assign it to a variable.

As an example, consider a function that plots the variable y against the variable x and fits and plots a linear model of y against x. You can define and name such a function to get the following code:

```
plot_and_fit <- function(d) {
  plot(y ~ x, data = d)
  abline(lm(y ~ x, data = d))
}

x <- rnorm(20)
y <- x + rnorm(20)
data.frame(x, y) %>% plot_and_fit
```

Since giving the function a name doesn't affect how the function works, it isn't necessary to do so. You can just put the code that defined the function where the name of the function goes to get this:

```
data.frame(x, y) %>% (function(d) {
  plot(y ~ x, data = d)
  abline(lm(y ~ x, data = d))
})
```

It does the exact same thing, but without defining a function. It is just not that readable either. Using
`.` and curly brackets, you can improve the readability (slightly) by just writing the body of the function and
referring to the input of it—what was called d above—as `.`:

```
data.frame(x, y) %>% {
  plot(y ~ x, data = .)
  abline(lm(y ~ x, data = .))
}
```

Other Pipeline Operations

The `%>%` operator is a very powerful mechanism for specifying data analysis pipelines, but there are some
special cases where slightly different behavior is needed.

One case is when you need to refer to the parameters in a data frame you get from the left side of the
pipe expression directly. In many functions, you can get to the parameters of a data frame just by naming
them, as you have seen with `lm` and `plot`, but there are cases where that is not so simple.

You can do that by indexing `.` like this:

```
d <- data.frame(x = rnorm(10), y = 4 + rnorm(10))
d %>% {data.frame(mean_x = mean(.$x), mean_y = mean(.$y))}
##      mean_x   mean_y
## 1 0.4167151 3.911174
```

But if you use the operator `%$%` instead of `%>%`, you can get to the variables just by naming them instead.

```
d %$% data.frame(mean_x = mean(x), mean_y = mean(y))
##      mean_x   mean_y
## 1 0.4167151 3.911174
```

Another common case is when you want to output or plot some intermediate result of a pipeline. You
can of course write the first part of a pipeline, run data through it, and store the result in a parameter, output
or plot what you want, and then continue from the stored data. But you can also use the `%T>%` (tee) operator.
It works like the `%>%` operator but where `%>%` passes the result of the right side of the expression on, `%T>%`
passes on the result of the left side. The right side is computed but not passed on, which is perfect if you only
want a step for its side-effect, like printing some summary.

```
d <- data.frame(x = rnorm(10), y = rnorm(10))
d %T>% plot(y ~ x, data = .) %>% lm(y ~ x, data = .)
```

The final operator is `%<>%`, which does something I warned against earlier—it assigns the result of a
pipeline back to a variable on the left. Sometimes you *do* want this behavior—for instance if you do some
data cleaning right after loading the data and you never want to use anything between the raw and the
cleaned data, you can use `%<>%`.

```
d <- read_my_data("/path/to/data")
d %<>% clean_data
```

I use it sparingly and prefer to just pass this case through a pipeline, as follows:

```
d <- read_my_data("/path/to/data") %>% clean_data
```

Coding and Naming Conventions

People have been developing R code for a long time, and they haven't been all that consistent in how they do it. So as you use R packages, you will see many different conventions on how code is written and especially how variables and functions are named.

How you choose to write your code is entirely up to you as long as you are consistent with it. It helps somewhat if your code matches the packages you use, just to make everything easier to read, but it really is up to you.

A few words on naming is worth going through, though. There are three ways people typically name their variables, data, or functions, and these are:

```
underscore_notation(x, y)
camelBackNotation(x, y)
dot.notation(x, y)
```

You are probably familiar with the first two notations, but if you have used Python or Java or C/C++ before, the dot notation looks like method calls in object oriented programming. It is not. The dot in the name doesn't mean method call. R just allows you to use dots in variable and function names.

I will mostly use the underscore notation in this book, but you can do whatever you want. I recommend that you stay away from the dot notation, though. There are good reasons for this. R actually put some interpretation into what dots mean in function names, so you can get into some trouble. The built-in functions in R often use dots in function names, but it is a dangerous approach, so you should probably stay away from it unless you are absolutely sure that you are avoiding its pitfalls.

Exercises

Try the following exercises to become more comfortable with the concepts discussed in this chapter.

Mean of Positive Values

You can simulate values from the normal distribution using the rnorm() function. Its first argument is the number of samples you want, and if you do not specify other values, it will sample from the $N(0,1)$ distribution.

Write a pipeline that takes samples from this function as input, removes the negative values, and computes the mean of the rest. Hint: One way to remove values is to replace them with missing values (NA); if a vector has missing values, the mean() function can ignore them if you give it the option na.rm = TRUE.

Root Mean Square Error

If you have "true" values, $\mathbf{t} = (t_1, ..., tn)$ and "predicted" values $\mathbf{y} = (y_1, ..., yn)$, then the root mean square error is defined as $\mathrm{RMSE}(\mathbf{t},\mathbf{y}) = \sqrt{\dfrac{1}{n}\sum_{i=1}^{n}(t_i - y_i)^2}$.

Write a pipeline that computes this from a data frame containing the **t** and **y** values. Remember that you can do this by first computing the square difference in one expression, then computing the mean of that in the next step, and finally computing the square root of this. The R function for computing the square root is sqrt().

CHAPTER 2

■ ■ ■

Reproducible Analysis

The typical data analysis workflow looks like this: you collect your data and you put it in a file or spreadsheet or database. Then you run some analyses, written in various scripts, perhaps saving some intermediate results along the way or maybe always working on the raw data. You create some plots or tables of relevant summaries of the data, and then you go and write a report about the results in a text editor or word processor. It is the typical workflow. Most people doing data analysis do this or variations thereof. But it is also a workflow that has many potential problems.

There is a separation between the analysis scripts and the data, and there is a separation between the analysis and the documentation of the analysis.

If all analyses are done on the raw data then issue number one is not a major problem. But it is common to have scripts for different parts of the analysis, with one script storing intermediate results that are then read by the next script. The scripts describe a workflow of data analysis and, to reproduce an analysis, you have to run all the scripts in the right order. Often enough, this correct order is only described in a text file or, even worse, only in the head of the data scientist who wrote the workflow. What is even worse, it won't stay there for long and is likely to be lost before it is needed again.

Ideally, you always want to have your analysis scripts written in a way in which you can rerun any part of your workflow, completely automatically, at any time.

For issue number two, the problem is that even if the workflow is automated and easy to run again, the documentation quickly drifts away from the actual analysis scripts. If you change the scripts, you won't necessarily remember to update the documentation. You probably don't forget to update figures and tables and such, but not necessarily the documentation of the exact analysis run. Options to functions and filtering choices and such. If the documentation drifts far enough from the actual analysis, it becomes completely useless. You can trust automated scripts to represent the real data analysis at any time—that is the benefit of having automated analysis workflows in the first place—but the documentation can easily end up being pure fiction.

What you want is a way to have dynamic documentation. Reports that describe the analysis workflow in a form that can be understood both by machines and humans. Machines use the report as an automated workflow that can redo the analysis at any time. We humans use it as documentation that always accurately describes the analysis workflow that we run.

Literate Programming and Integration of Workflow and Documentation

One way to achieve the goal of having automated workflows and documentation that is always up to date is something called "literate programming". Literate programming is an approach to software development, proposed by Stanford computer scientist Donald Knuth, which never became popular for programming, possibly because most programmers do not like to write documentation.

The idea in literate programming is that the documentation of a program—in the sense of the documentation of how the program works and how algorithms and data structures in the program works—is written together with the code implementing the program. Tools such as Javadoc and Roxygen (`http://roxygen.org`) do something similar. They have documentation of classes and methods written together with the code in the form of comments. Literate programming differs slightly from this. With Javadoc and Roxygen, the code is the primary document, and the documentation is comments added to it. With literate programming, the documentation is the primary text for humans to read and the code is part of this documentation, included where it falls naturally to have it. The computer code is extracted automatically from this document when the program runs.

Literate programming never became a huge success for writing programs, but for doing data science, it is having a comeback. The results of a data analysis project is typically a report describing models and analysis results, and it is natural to think of this document as the primary product. So the documentation is already the main focus. The only thing needed to use literate programming is a way of putting the analysis code inside the documentation report.

Many programming languages have support for this. Mathematica (`https://www.wolfram.com/mathematica/`) has always had notebooks where you could write code together with documentation. Jupyter (`http://jupyter.org`), the descendant of iPython Notebook, lets you write notebooks with documentation and graphics interspersed with executable code. And in R there are several ways of writing documents that are used both as automated analysis scripts as well as for generating reports. The most popular of these approaches is R Markdown (for writing these documents) and `knitr` (for running the analysis and generating the reports).

Creating an R Markdown/knitr Document in RStudio

To create a new R Markdown document, go to the File menu, choose New File and then R Markdown. Now RStudio will bring up a window where you can decide which kind of document you want to make and add some information, such as title and author name. It doesn't matter so much what you do here; you can change it later. But try making an HTML document.

The result is a new file with some boilerplate text in it, as shown in Figure 2-1. At the top of the file, between two lines containing just `---` is some meta-information for the document, and after the second `---` is the actual text. It consists of a mix of text, formatted in the Markdown language, and R code.
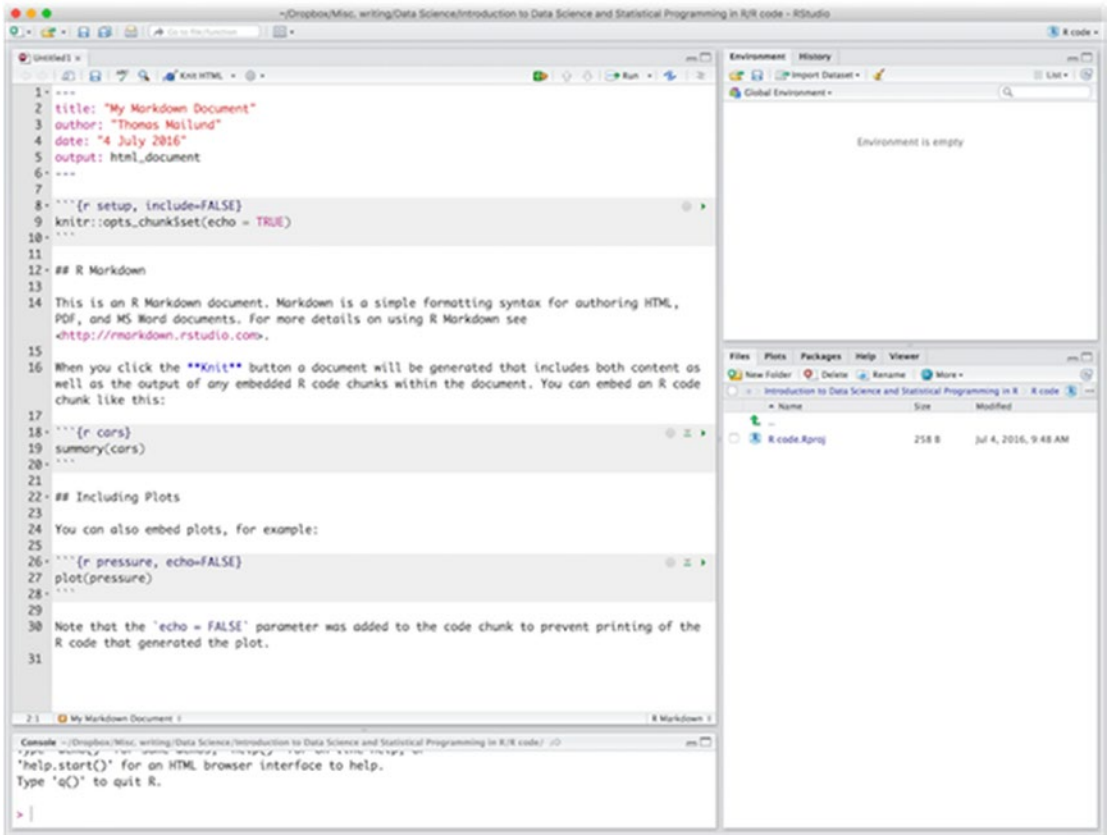
Figure 2-1. *A new R Markdown file*

In the toolbar above the open file, there is a menu option called Knit HTML. If you click it, it will translate the R Markdown into an HTML document and open it, as shown in Figure 2-2. You have to save the file first, though. If you click the Knit HTML button before saving, you will be asked to save the file.
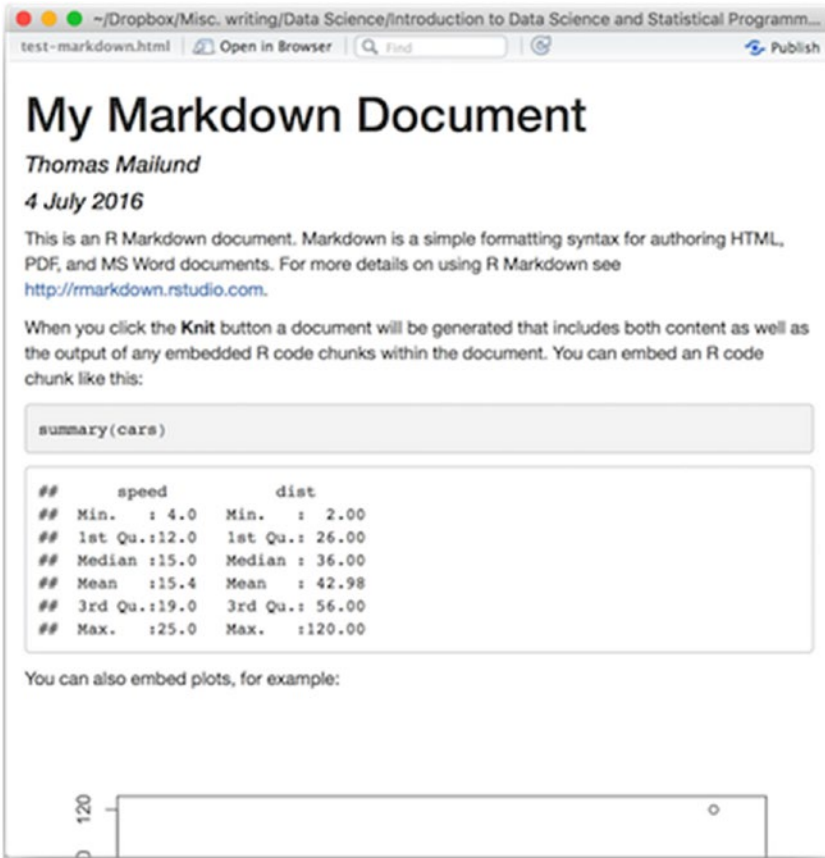
Figure 2-2. *Compiled Markdown file*

The newly created HTML file is also written to disk with the name you gave the R Markdown file. The R Markdown file will have suffix `.rmd` and the HTML file will have the same prefix, but with the suffix `.html`.

If you click the gear logo next to Knit HTML (in earlier versions of R Studio this is a down-pointing arrow), you get some additional options. You can ask to see the HTML document in the pane to the right in RStudio instead of in a new window. Having the document in a panel instead of a separate window can be convenient if you are on a laptop and do not have a lot of screen space. You can also generate a file or a Word file instead of an HTML file.

If you decide to produce a file in a different output format, RStudio will remember this. It will update the Knit HTML to Knit or Knit Word and it will update the metadata in the header of your file. If you manually update the header, this is reflected in the Knit X button. If you click the gear icon one step farther right, you get some more options for how the document should be formatted.

The actual steps involved in creating a document involves two tools and three languages, but it is all integrated so you typically will not notice. There is the R code embedded in the document. The R code is first processed by the `knitr` package that evaluates it and handles the results such as data and plots according to options you give it. The result is a Markdown document (notice no R). This Markdown document is then processed by the tool `pandoc`, which is responsible for generating the output file. For this, it uses the metadata in the header, which is written in a language called YAML, whereas the actual formatting is written in the the Markdown language.

You usually don't have to worry about `pandoc` working as the back-end of document processing. If you just write R Markdown documents, then RStudio will let you compile them into different types of output documents. But because the pipeline goes from R Markdown via `knitr` to Markdown and then via `pandoc` to the various output formats, you do have access to a very powerful tool for creating documents. I have written this book in R Markdown where each chapter is a separate document that I can run through `knitr` independently. I then have `pandoc` with some options take the resulting Markdown documents, combine them, and produce both output and Epub output. With `pandoc`, it is possible to use different templates for setting up the formatting, and having it depend on the output document you create by using different templates for different formats. It is a very powerful, but also a very complicated, tool and it is far beyond what we can cover in this book. Just know that it is there if you want to take document writing in R Markdown further than what you can readily do in RStudio.

As I mentioned, there are actually three languages involved in an R Markdown document. We will handle them in order—first the header language, which is YAML, then the text formatting language, which is Markdown, and then finally how R is embedded in a document.

The YAML Language

YAML is a language for specifying key-value data. YAML stands for the (recursive) acronym *YAML Ain't Markup Language*. So yes, when I called this section "The YAML Language," I shouldn't have included *language* since the *L* stands for language, but I did. I stand by that choice. The acronym used to stand for Yet Another Markup Language but since "markup language" typically refers to commands used to mark up text for either specifying formatting or for putting structured information in a text, which YAML doesn't do, the acronym was changed. YAML is used for giving options in various forms to a computer program processing a document, not so much for marking up text, so it isn't really a markup language.

In your R Markdown document, the YAML is used in the header, which is everything that goes between the first and second line with three dashes. In the document you create when you make a new R Markdown file, it can look like this:

```
---
title: "My Markdown Document"
author: "Thomas Mailund"
date: "17 July 2016"
output: html_document
---
```

You usually do not have to modify this header manually. If you use the GUI, it will adjust the header for you when you change options. You do need to alter it to include bibliographies, though, which we get to later. And you can always add anything you want to the header if you need to and it can be easier than using the GUI. But you don't have to modify the header that often.

YAML gives you a way to specify key-value mappings. You write `key:` and then the `value` afterward. So above, you have the key `title` referring to `My Markdown Document`, the key `author` to refer to `"Thomas Mailund"` and so on. You don't necessarily need to quote the values unless they have a colon in them, but you always can.

The YAML header is used both by RStudio and `pandoc`. RStudio uses the `output` key for determining which output format to translate your document into and this choice is reflected in the **Knit** toolbar button. `pandoc`, on the other hand, uses the `title`, `author`, and `date` to put that information into the generated document.

You can have slightly more structure to the key-value specifications. If a key should refer to a list of values, you use -. So if you have more than one author, you can use something like this:

```
---
...
author:
 - "Thomas Mailund"
 - "Christian Storm"
...
---
```

Or you can have more nested key-value structure, so if you change the output theme (using Output Options after clicking on the tooth-wheel in the toolbar), you might see something like this:

```
output:
  html_document:
    theme: united
```

How the options are used depends on the tool-chain used to format your document. The YAML header just provides specifications. Which options you have available and what they do is not part of the language.

For pandoc, it depends on the templates used to generate the final document (see later), so there isn't even a complete list that I can give you for pandoc. Anyone who writes a new template can decide on new options to use. The YAML header gives you a way to provide options to such templates, but there isn't a fixed set of keywords to use. It all depends on how tools later in the process interpret them.

The Markdown Language

The Markdown language *is* a markup language—the name is a pun. It was originally developed to make it easy to write web pages. HTML, the language used to format web pages, is also a markup language but is not always easily human readable. Markdown intended to solve this by formatting text with very simple markup commands—familiar from e-mails back in the day before e-mails were also HTML documents—and then have tools for translating Markdown into HTML.

Markdown has gone far beyond just writing web pages, but it is still a very simple and intuitive language for writing human-readable text with markup commands that can then be translated into other document formats.

In Markdown, you write plain text as plain text. So the body of text is just written without any markup. You will need to write it in a text editor so the text is actually text, not a word processor where the file format already contains a lot of markup information that isn't readily seen onscreen. If you are writing code, you should already know about text editors. If not, just use RStudio to write R Markdown files, and you will be okay.

Markup commands are used when you want something else than just plain text. There aren't many commands to learn—the philosophy is that when writing you should focus on the text and not the formatting—so they are very quickly learned.

Formatting Text

First, there are section headers. You can have different levels of headers—think chapters, sections, subsections, etc.— and you specify them using # starting at the beginning of a new line.

```
# Header 1
## Header 2
### Header 3
```

For the first two, you can also use this format:

```
Header 1
========

Header 2
--------
```

To have lists in your document, you write them as you have probably often seen them in raw text documents. A list with bullets (and not numbers) is written like this:

```
 * this is a
 * bulleted
 * list
```

The result looks like this:

- this is a

- bulleted

- list

You can have sublists just by indenting. You need to move the indented line in so there is a space between where the text starts at the outer lever and where the bullet is at the next level. Otherwise, the line goes at the outer level. The output of this:

```
 * This is the first line
   * This is a sub-line
   * This is another sub-line
  * This actually goes to the outer level
 * This is definitely at the outer level
```

Is this list:

- This is the first line

  - This is a sub-line
  - This is another sub-line

- This actually goes to the outer level

- Back to the outer level

If you prefer, you can use - instead of * for these lists and you can mix the two.

```
- First line
* Second line
    - nested line
```

- First line

- Second line

  – nested line

To have numbered lists, just use numbers instead of * and -.

```
1. This is a
2. numbered
3. list
```

The result looks like this:

1. This is a

2. numbered

3. list

You don't actually need to get the numbers right, you just need to use numbers. So

```
1. This is a
3. numbered
2. list
```

Would produce the same (correctly numbered) output. You will start counting at the first number, though, so

```
4. This is a
4. numbered
4. list
```

Produces:

4. This is a
5. numbered
6. list

To construct tables, you also use a typical text representation with vertical and horizontal lines. Vertical lines separate columns and horizontal lines separate headers from the table body. This code:

```
| First Header  | Second Header | Third Header   |
| :------------ | :-----------: | -------------: |
| First row     | Centered text | Right justified |
| Second row    | *Some data*   | *Some data*    |
| Third row     | *Some data*   | *Some data*    |
```

Will result in this table:

First Header	Second Header	Third Header
First row	Centered text	Right justified
Second row	*Some data*	*Some data*
Third row	*Some data*	*Some data*

The : in the line separating the header from the body determines the justification of the column. Put it on the left to get left justification, on both sides to get the text centered, and on the right to get the text right justified.

Inside text, you use markup codes to make text italic or boldface. You use either *this* or _this_ to make *this* italic, while you use **this** or __this__ to make **this** boldface.

Since Markdown was developed to make HTML documents it, of course, has an easy way to insert links. You use the notation [link text](link URL) to put link text into the document as a link to link URL. This notation is also used to make cross-references inside a document—similar to how HTML documents have anchors and internal links—but more on that later.

To insert images into a document, you use a notation similar to the link notation, but you just put a ! before the link. So ![Image description](URL to image) will insert the image pointed to by URL to image with a caption saying Image description. The URL here will typically be a local file, but it can be a remote file referred to via HTTP.

With long URLs, the marked-up text can be hard to read even with this simple notation and it is possible to remove the URLs from the actual text and place them later in the document, for example, after the paragraph referring to the URL or at the end of the document. For this, you use the notation [link text][link tag] and define the link tag as the URL you want later.

```
This is some text [with a link][1].
The link tag is defined below the paragraph.

[1]: interesting-url-of-some-sort-we-dont-want-inline
```

You can use a string here for the tag. Using numbers is easy, but for long documents, you won't be able to remember what each number refers to.

```
This is some text [with a link][interesting].
The link tag is defined below the paragraph.

[interesting]: interesting-url-of-some-sort-we-dont-want-inline
```

You can make block quotes in text using notation you will be familiar with from e-mails.

```
> This is a
> block quote
```

Gives you this:
This is a block quote

To put verbatim input as part of your text, you can either do it inline or as a block. In both cases you use backticks `. Inline in the text, you use single backticks `foo`. To create a block of text you write:

```
```

block of
text
```
```

You can also just indent text with four spaces, which is how I managed to make a block of verbatim text that includes three backticks.

Markdown is used a lot by people who document programs, so there is a notation for getting code highlighted in verbatim blocks. The convention is to write the name of the programming language after the three backticks, then the program used for formatting the document will highlight the code when it can. For R code you write r, so this block:

```r
f <- function(x) ifelse(x %% 2 == 0, x**2, x**3)
f(2)
```

Is formatted like this:

```r
f <- function(x) ifelse(x %% 2 == 0, x**2, x**3)
f(2)
```

The only thing this markup of blocks does is highlight the code. It doesn't try to evaluate the code. Evaluating code happens before the Markdown document is formatted and we return to that shortly.

Cross-Referencing

Out of the box, there is not a lot of support for making cross references in Markdown documents. You can make cross-references to sections but not figures or tables. There are ways of doing it with extensions to pandoc—I use it in this book—but out of the box from RStudio, you cannot. Although, with the work being done for making book-writing and lengthy reports in Bookdown (https://bookdown.org/yihui/bookdown/), that might change soon.[1]

The easiest way to reference a section is to put the name of the section in square brackets. If I write [Cross referencing] here, I get a link to this cross-referencing section. Of course, you don't always want the name of the section to be the text of the link, so you can also write [this section][Cross referencing] to get a link to this section.

This approach naturally works only if all section titles are unique. If they are not, you cannot refer to them simply by their names. Instead, you can tag them to give them a unique identifier. You do this by writing the identifier after the title of the section. To put a name after a section header, you write:

```
### Cross referencing {#section-cross-ref}
```

Then you can refer to the section using [this](#section-cross-ref). Here you do need the # sign in the identifier. That markup is leftover from HTML, where anchors use #.

---

[1]In any case, having cross-references to sections but not figures is still better than Word, where the feature is there but buggy to the point of uselessness, in my experience.

Bibliographies

Often you want to cite books or papers in a report. You can, of course, handle citations manually, but a better approach is to have a file with the citation information and then refer to it using markup tags. To add a bibliography, you use a tag in the YAML header called bibliography.

```
---
...
bibliography: bibliography.bib
...
---
```

You can use several different formats here; see the R Markdown documentation (http://rmarkdown. rstudio.com/authoring_bibliographies_and_citations.html) for a list. The suffix .bib is used for BibLaTeX. The format for the citation file is the same as BibTeX, and you get citation information in that format from nearly every site that will give you bibliography information.

To cite something from the bibliography, you use [@smith04] where smith04 is the identifier used in the bibliography file. You can cite more than one paper inside square brackets separated by a semicolon, [@smith04; doe99], and you can add text such as chapters or page numbers [@smith04, chapter 4]. To suppress the author name(s) in the citation, say when you mention the name already in the text, you put - before the @, so you write As Smith showed [-@smith04].... For in-text citations, similar to \citet{} in natbib, you just leave out the brackets: @smith04 showed that... and you can combine that with additional citation information as @smith04 [chapter 4] showed that....

To specify the citation style to use, you use the csl tag in the YAML header.

```
---
...
bibliography: bibliography.bib
csl: biomed-central.csl
...
---
```

Check out the citation styles list at https://github.com/citation-style-language/styles for a large number of different formats. There should be most, if not all, of your heart desires there.

Controlling the Output (Templates/Stylesheets)

The pandoc tool has a powerful mechanism for formatting the documents it generates. This is achieved using stylesheets in CSS for HTML and from using templates for how to format the output for all output formats. The template mechanism lets you write an HTML or LaTeX document, say, that determines where various part of the text goes and where variables from the YAML header is used. This mechanism is far beyond what we can cover in this chapter, but I just want to mention it if you want to start writing papers using R Markdown. You can do this, you just need to have a template for formatting the document in the style a journal wants. Often they provide LaTeX templates, and you can modify these to work with Markdown.

There isn't much support for this in RStudio, but for HTML documents, you can use the Output Options command (click on the tooth-wheel) to choose different output formatting.

Running R Code in Markdown Documents

The formatting so far is all Markdown (and YAML). Where it combines with R and makes it R Markdown is through `knitr`. When you format a document, the first step evaluates R code to create a Markdown document. This translates an `.rmd` document into an `.md` document, but this intermediate document is deleted afterward unless you explicitly tell RStudio not to do so. It does that by running all the R code you want to be executed and putting it into the Markdown document.

The simplest R code you can evaluate is part of a text. If you want an R expression evaluated, you use backticks but add r right after the first. So to evaluate 2 + 2 and put the result in your Markdown document, you write `` `r `` and then the expression 2 + 2 and get the result 4 inserted into the text. You can write any R expression there to get it evaluated. This is useful for inserting short summary statistics like means and standard deviations directly into the text and ensuring that the summaries are always up to date with the actual data you are analyzing.

For longer chunks of code, you use the block-quotes, the three backticks. Instead of just writing:

```
```r
2 + 2
```
```

which will only display the code (highlighted as R code), you put the r in curly brackets.

This will insert the code in your document but will also show the result of evaluating it right after the code block. The boilerplate code you get when creating an R Markdown document in RStudio shows you examples of this (see Figure 2-3).

Figure 2-3. *Code chunk in RStudio*

You can name code chunks by putting a name right after r. You don't have to name all chunks, and if you have a lot of chunks, you probably won't bother naming all of them. But if you give them a name, they are easily located by clicking on the structure button in the bar below the document (see Figure 2-4). You can also use the name to refer to chunks when caching results, which we will cover later.

You can also embed plots,

```
Untitled
Chunk 1: setup
R Markdown
  Chunk 2: cars
Including Plots
  Chunk 3: pressure
```

Untitled ⬍

Figure 2-4. *Document structure with chunk names*

If you have the most recent version of RStudio, you should see a toolbar to the right on every code chunk (see Figure 2-5). The rightmost option, the Play button, will let you evaluate the chunk. The results will be shown below the chunk unless you have disabled that option. The middle button evaluates all previous chunks down to and including the current one. This is useful when the current chunk depends on previous results. The tooth-wheel lets you set options for the chunk.

```
17
18 ▾ ```{r cars}
19    summary(cars)
20    ```
21
```

Figure 2-5. *Code chunk toolbar*

The chunk options, shown in Figure 2-6, control the output you get when evaluating a code chunk. The Output drop-down selects what output the chunk should generate in the resulting document, while the Show Warnings and Show Messages buttons determine whether warnings and messages, respectively, should be included in the output. The Use Custom Figure Size button is used to determine the size of figures you generate. We return to these later.

```
Name:   cars
Output: (Use document default)  ◆
◉ Show warnings
◉ Show messages
☐ Use custom figure size
(?) Chunk options        Revert  Apply
```

Figure 2-6. *Code chunk options*

If you modify these options, you will see that the options are included in the top line of the chunk. You can of course also manually control the options here, and there are more options than what you can control with the window in the GUI. You can read the knitr documentation for all the details (http://yihui.name/knitr/).

This dialog box will handle most of your needs, though, except for displaying tables or when you want to cache results of chunks, both of which we return to later.

Using Chunks when Analyzing Data (Without Compiling Documents)

Before continuing, though, I want to stress that working with data analysis in an R Markdown document is useful for more than just creating documents. I personally do all my analysis in these documents because I can combine documentation and code, regardless of whether I want to generate a report at the end. The combination of explanatory text and analysis code is just convenient to have.

The way code chunks are evaluated as separate pieces of analysis is also part of this. You can evaluate chunks individually, or all chunks down to a point, and I find that very convenient when doing an analysis. There are keyboard shortcuts for evaluating all chunks, all previous chunks, or just the current chunk (see Figure 2-7), which makes it very easy to write a bit of code for an exploratory analysis and evaluate just that piece of code. If you are familiar with Jupyter or similar notebooks, you will recognize the workflow.

Figure 2-7. *Options for evaluating chunks*

Even without the option for generating final documents from a Markdown document, I would still be using them just for this feature.

Caching Results

Sometimes part of an analysis is very time-consuming. Here I mean in CPU time, not thinking time—it is also true for thinking time, but you don't need to think the same things over and over. If you are not careful, however, you will need to run the same analysis on the computer again and again.

If you have such time-consuming steps in your analysis, then compiling documents will be very slow. Each time you compile the document, all the analysis is done from scratch. This is the functionality you want since this makes sure that the analysis does not have results left over from code that isn't part of the document, but it limits the usability of the workflows if they take hours to compile.

To alleviate this, you can cache the results of a chunk. To cache the result of a chunk, you should add the option cache=TRUE to it. This means adding that in the header of the chunk similar to how output options are added. You will need to give the chunk a name to use this. Chunks without names are actually given a default name, but this name changes according to how many nameless chunks you have earlier in the document and you can't have that if you use the name to remember results. So you need to name it. A named chunk that is set to be cached will not only be when you compile a document if it has changed since the last time it was evaluated. If it hasn't changed, the cached results of the last evaluation will just be reused.

R cannot cache everything, so if you load libraries in a cached chunk they won't be loaded unless the chunk is being evaluated. That means there are some limits to what you can do, but generally it is a very useful feature.

Since other chunks can depend on a cached chunk, there can also be problems if a cached chunk depends on another chunk, cached or not. The chunk will only be re-evaluated if you have changed the code inside it, so if it depends on something you have changed, it will remember results based on outdated data. You have to be careful about that.

You can set up dependencies between chunks, though, to fix this problem. If a chunk is dependent on the results of another chunk, you can specify this using the chunk option dependson=other. Then, if the chunk other (and you need to name such chunks) is modified, the cache is considered invalid, and the depending chunk will be evaluated again.

Displaying Data

Since you are writing a report on data analysis, you naturally want to include some results. That means displaying data in some form or other.

You can simply include the results of evaluating R expressions in a code chunk, but often you want to display the data using tables or graphics, especially if the report is something you want to show to people not familiar with R. Luckily, both tables and graphics are easy to display.

To make a table, you can use the function kable() from the knitr package. Try adding a chunk like this to the boilerplate document you have:

```
library(knitr)
kable(head(cars))
```

The library(knitr) imports functions from the knitr package so you get access to the kable() function. You don't need to include it in every chunk you use kable() in, just in any chunk before you use the function—the setup chunk is a good place—but adding it in the chunk, you write now will work.

The function kable() will create a table from a data frame in the Markdown format so it will be formatted in the later step of the document compilation. Don't worry too much about the details in the code here; the head() function just picks out the first lines of the cars data so the table doesn't get too long.

Using kable() should generate a table in your output document. Depending on your setup, you might have to give the chunk the output option result="asis" to make it work, but it usually should give you a table even without this.

We will cover how to summarize data in later chapters. Usually, you don't want to make tables of full datasets, but for now, you can try just getting the first few lines of the cars data.

Adding graphics to the output is just as simple. You simply make a plot in a code chunk, and the result will be included in the document you generate. The boilerplate R Markdown document already gives you an example of this. We will cover plotting in much more detail later.

Exercises

Try the following exercises to become more comfortable with the concepts discussed in this chapter.

Create an R Markdown Document

Go to the File menu and create an R Markdown document. Read through the boilerplate text to see how it is structured. Evaluate the chunks. Compile the document.

Produce Different Output

Create from the same R Markdown document an HTML document, a document, and a Word document.

Add Caching

Add a cached code chunk to your document. Make the code there sample random numbers, e.g., using rnorm(). When you recompile the document, you should see that the random numbers do not change.

Make another cached chunk that uses the results of the first cached chunk. Say, compute the mean of the random numbers. Set up dependencies and verify that if you modify the first chunk, the second chunk gets evaluated.

CHAPTER 3

■ ■ ■

Data Manipulation

Data science is as much about manipulating data as it is about fitting models to data. Data rarely arrives in a form that we can directly feed into the statistical models or machine learning algorithms we want to analyze them with. The first stages of data analysis are almost always figuring out how to load the data into R and then figuring out how to transform it into a shape you can readily analyze. The code in this chapter, and all the following, assumes that the packages `magrittr` and `ggplot2` have been loaded (just to avoid explicitly doing so in each example).

Data Already in R

There are some datasets already built into R or available in R packages. Those are useful for learning how to use new methods. If you already know a dataset and what it can tell you, it is easier to evaluate how a new method performs. It's also useful for benchmarking methods you implement. They are of course less helpful when it comes to analyzing new data.

Distributed together with R is the package `dataset`. You can load the package into R using the `library()` function and get a list of the datasets in it, together with a short description of each, like this:

```
library(datasets)
library(help = "datasets")
```

To load an actual dataset into R's memory, use the `data()` function. The datasets are all relatively small, so they are ideal for quickly testing the code you are working with. For example, to experiment with plotting x-y plots (see Figure 3-1), you could use the `cars` dataset that consists of only two columns—a speed and a breaking distance:

```
data(cars)
head(cars)
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10
cars %>% qplot(speed, dist, data = .)
```

Figure 3-1. *Plot of the cars dataset*

Don't worry about the plotting function for now; we return to plotting in the next chapter.

If you are developing new analysis or plotting code, usually one of these datasets is useful for testing it.

Another package with several useful datasets is mlbench. It contains datasets for machine learning benchmarks so these datasets are aimed at testing how new methods perform on known datasets. This package is not distributed together with R, but you can install it, load it, and get a list of the datasets in it like this:

```
install.packages("mlbench")
library(mlbench)
library(help = "mlbench")
```

In this book, I use data from one of those two packages when giving examples of data analyses.

The packages are convenient for me for giving examples, and if you are developing new functionality for R they are suitable for testing, but if you are interested in data analysis, presumably you are interested in your *own* data, and there they are of course useless. You need to know how to get your own data into R. We get to that shortly, but first I want to say a few words about how you can examine a dataset and get a quick overview.

Quickly Reviewing Data

I have already used the function head(), which shows the first *n* lines of a data frame where *n* is an option with default 6. You can use another *n* to get more or less this:

```
cars %>% head(3)
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
```

The similar function tail() gives you the last *n* lines:

```
cars %>% tail(3)
##    speed dist
## 48    24   93
## 49    24  120
## 50    25   85
```

To get summary statistics for all the columns in a data frame, you can use the summary() function:

```
cars %>% summary
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

It isn't that exciting for the cars dataset, so let's see it on another built-in dataset:

```
data(iris)
iris %>% summary
##   Sepal.Length    Sepal.Width     Petal.Length
##  Min.   :4.300   Min.   :2.000   Min.   :1.000
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600
##  Median :5.800   Median :3.000   Median :4.350
##  Mean   :5.843   Mean   :3.057   Mean   :3.758
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100
##  Max.   :7.900   Max.   :4.400   Max.   :6.900
##   Petal.Width          Species
##  Min.   :0.100   setosa    :50
##  1st Qu.:0.300   versicolor:50
##  Median :1.300   virginica :50
##  Mean   :1.199
##  3rd Qu.:1.800
##  Max.   :2.500
```

The summary you get depends on the types the columns have. Numerical data is summarized by their quartiles and meanwhile categorical, and Boolean data is summarized by counts of each category or TRUE/ FALSE values. In the `iris` dataset there is one column, `Species`, that is categorical, and its summary is the count of each level.

To see the type of each column, you can use the `str()` function. This gives you the structure of a data type and is much more general than you need here, but it does give you an overview of the types of columns in a data frame and is very useful for that.

Reading Data

There are several packages for reading data in different file formats, from Excel to JSON to XML and so on. If you have data in a particular format, try to Google for how to read it into R. If it is a standard data format, the chances are that there is a package that can help you.

Quite often, though, data is available in a text table of some kind. Most tools can import and export those. R has plenty of built-in functions for reading such data. Use this to get a list of them:

```
?read.table
```

These functions are all variations of the `read.table()` function, just using different default options. For instance, while `read.table()` assumes that the data is given in whitespace-separated columns, the `read.csv()` function assumes that the data is represented as comma-separated values, so the difference between the two functions is in what they consider being separating data columns.

The `read.table()` function takes a lot of arguments. These are used to adjust it to the specific details of the text file you are reading. (The other functions take the same arguments, they just have different defaults.) The options I find I use the most are these:

- `header`—This is a Boolean value telling the function whether it should consider the first line in the input file a header line. If it's set to true, it uses the first line to set the column names of the data frame it constructs; if it is set to false the first line is interpreted as the first row in the data frame.

- `col.names`—If the first line is not used to specify the header, you can use this option to name the columns. You need to give it a vector of strings with a string for each column in the input.

- `dec`—This is the decimal point used in numbers. I get spreadsheets that use both `.` and `,` for decimal points, so this is an important parameter to me. How important it will be to you probably depends on how many nationalities you collaborate with.

- `comment.char`—By default, the function assumes that # is the start of a comment and ignores the rest of a line when it sees it. If # is actually used in your data, you need to change this. The same goes if comments are indicated with a different symbol.

- `stringsAsFactors`—By default, the function will assume that columns containing strings should really be interpreted as factors. Needless to say, this isn't always correct. Sometimes a string is a string. You can set this parameter to FALSE to make the function interpret strings as strings. This is an all or nothing option, though. If it is TRUE, *all* columns with strings will be interpreted as factors and if it is FALSE, *none* of them will.

- colClasses—Lets you specify which type each column should have, so here you can specify that some columns should be factors, and others should be strings. You have to specify all columns, though, which is cumbersome and somewhat annoying since R, in general, is pretty good at determining the right types for a column. The option will only take you so far in any case. You can tell it that a column should be an ordered factor but not what the levels should be and such. I mainly use it for specifying which columns should be factors and which should be strings, but using it will also speed up the function for large datasets since R then doesn't have to figure out the column types itself.

For reading in tables of data, read.table() and friends will usually get you there with the right options. If you are having problems reading data, check the documentation carefully to see if you cannot tweak the functions to get the data loaded. It isn't always possible, but it usually is. When it really isn't, I usually give up and write a script in another language to format the data into a form, I can load into R. For raw text processing, R isn't really the right tool and rather than forcing all steps in an analysis into R, I will be pragmatic and choose the best tools for the task, and R isn't always it. But before taking drastic measures and go programming in another language, you should carefully check if you cannot tweak one of the read.table() functions first.

Examples of Reading and Formatting Datasets

Rather than discussing the import of data in the abstract, let's now see a couple of examples of how data can be read in and formatted.

Breast Cancer Dataset

As a first example of reading data from a text file, we consider the BreastCancer dataset from mlbench. Then we have something to compare our results with. The first couple of lines from this dataset are:

```
library(mlbench)
data(BreastCancer)
BreastCancer %>% head(3)
##        Id Cl.thickness Cell.size Cell.shape
## 1 1000025            5         1          1
## 2 1002945            5         4          4
## 3 1015425            3         1          1
##   Marg.adhesion Epith.c.size Bare.nuclei
## 1             1            2           1
## 2             5            7          10
## 3             1            2           2
##   Bl.cromatin Normal.nucleoli Mitoses  Class
## 1           3               1       1 benign
## 2           3               2       1 benign
## 3           3               1       1 benign
```

The data can be found at https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin +(Original), where there is also a description of the data. I have made this tinyURL to the actual data file http://tinyurl.com/kw4xtts. While R can read data from URLs, it cannot deal with the magic behind tinyURL and the real URL is too long to fit on the page of this book, so I have saved it in a variable, data_url, that I will use. To run the code yourself, you simply need to use the tinyURL. It will send you to the real URL, and then you can copy that into your code.

To download the data, you could go to the URL and save the file. Explicitly downloading data outside of the R code has pros and cons. It is pretty simple, and you can look at the data before you start parsing it, but on the other hand, it gives you a step in the analysis workflow that is not automatically reproducible. Even if the URL is described in the documentation and uses a link that doesn't change over time, it *is* a manual step in the workflow. And a step that people could make mistakes in.

Instead, I am going to read the data directly from the URL. Of course, this is also a risky step in a workflow because I am not in control of the server the data is on, and I cannot guarantee that the data will always be there and that it won't change over time. It is a bit of a risk either way. I will usually add the code to my workflow for downloading the data, but I will also store the data in a file. If I leave the code for downloading the data and saving it to my local disk in a cached Markdown chunk, it will only be run the one time I need it.

I can read the data and get it as a vector of lines using the readLines() function. I can always use that to scan the first one or two lines to see what the file looks like.

```
lines <- readLines(data_url)
lines[1:5]
## [1] "1000025,5,1,1,1,2,1,3,1,1,2"
## [2] "1002945,5,4,4,5,7,10,3,2,1,2"
## [3] "1015425,3,1,1,1,2,2,3,1,1,2"
## [4] "1016277,6,8,8,1,3,4,3,7,1,2"
## [5] "1017023,4,1,1,3,2,1,3,1,1,2"
```

For this data, it seems to be a comma-separated values file without a header line. So I save the data with the .csv suffix. None of the functions for writing or reading data in R cares about the suffixes, but it is easier for me to remember what the file contains that way.

```
writeLines(lines, con = "data/raw-breast-cancer.csv")
```

For that function to succeed, I first need to make a data/ directory. I suggest you have a data/ directory for all your projects, always, since you want your directories and files structured when you are working on a project.

The file I just wrote to disk can then read in using the read.csv() function.

```
raw_breast_cancer <- read.csv("data/raw-breast-cancer.csv")
raw_breast_cancer %>% head(3)
##   X1000025 X5 X1 X1.1 X1.2 X2 X1.3 X3 X1.4 X1.5
## 1  1002945  5  4    4    5  7   10  3    2    1
## 2  1015425  3  1    1    1  2    2  3    1    1
## 3  1016277  6  8    8    1  3    4  3    7    1
##   X2.1
## 1    2
## 2    2
## 3    2
```

Of course, I wouldn't write exactly these steps into a workflow. Once I have discovered that the data at the end of the URL is a .csv file, I would just read it directly from the URL.

```
raw_breast_cancer <- read.csv(data_url)
raw_breast_cancer %>% head(3)
##   X1000025 X5 X1 X1.1 X1.2 X2 X1.3 X3 X1.4 X1.5
## 1  1002945  5  4    4    5  7   10  3    2    1
## 2  1015425  3  1    1    1  2    2  3    1    1
## 3  1016277  6  8    8    1  3    4  3    7    1
##   X2.1
## 1    2
## 2    2
## 3    2
```

The good news is that this data looks similar to the BreastCancer data. The bad news is that it appears that the first line in BreastCancer seems to have been turned into column names in raw_breast_cancer. The read.csv() function interpreted the first line as a header. This we can fix using the header parameter.

```
raw_breast_cancer <- read.csv(data_url, header = FALSE)
raw_breast_cancer %>% head(3)
##        V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11
## 1 1000025  5  1  1  1  2  1  3  1   1   2
## 2 1002945  5  4  4  5  7 10  3  2   1   2
## 3 1015425  3  1  1  1  2  2  3  1   1   2
```

Now the first line is no longer interpreted as header names. That is good, but the names you actually get are not that informative about what the columns contain.

If you read the description of the data from the web site, you can see what each column is and choose names that are appropriate. I am going to cheat here and just take the names from the BreastCancer dataset.

I can set the names explicitly like this:

```
names(raw_breast_cancer) <- names(BreastCancer)
raw_breast_cancer %>% head(3)
##        Id Cl.thickness Cell.size Cell.shape
## 1 1000025            5         1          1
## 2 1002945            5         4          4
## 3 1015425            3         1          1
##   Marg.adhesion Epith.c.size Bare.nuclei
## 1             1            2           1
## 2             5            7          10
## 3             1            2           2
##   Bl.cromatin Normal.nucleoli Mitoses Class
## 1           3               1       1     2
## 2           3               2       1     2
## 3           3               1       1     2
```

Or I could set them where I load the data:

```
raw_breast_cancer <- read.csv(data_url, header = FALSE,
                              col.names = names(BreastCancer))
raw_breast_cancer %>% head(3)
##        Id Cl.thickness Cell.size Cell.shape
## 1 1000025            5         1          1
## 2 1002945            5         4          4
## 3 1015425            3         1          1
##   Marg.adhesion Epith.c.size Bare.nuclei
## 1             1            2           1
## 2             5            7          10
## 3             1            2           2
##   Bl.cromatin Normal.nucleoli Mitoses Class
## 1           3               1       1     2
## 2           3               2       1     2
## 3           3               1       1     2
```

Okay, we are getting somewhere. The `Class` column is not right. It encodes the classes as numbers (the web page documentation specifies 2 for benign and 4 for malignant), but in R it would be more appropriate with a factor.

You can translate the numbers into a factor by first translating the numbers into strings and then the strings into factors. I don't like modifying the original data—even if I have it in a file—so I am going to copy it first and then do the modifications.

```
formatted_breast_cancer <- raw_breast_cancer
```

It is easy enough to map the numbers to strings using `ifelse()`:

```
map_class <- function(x) {
    ifelse(x == 2, "bening",
    ifelse(x == 4, "malignant",
           NA))
}
mapped <- formatted_breast_cancer$Class %>% map_class
mapped %>% table
## .
##    bening malignant
##       458       241
```

I *could* have made it simpler with this:

```
map_class <- function(x) {
    ifelse(x == 2, "bening", "malignant")
}
mapped <- formatted_breast_cancer$Class %>% map_class
mapped %>% table
## .
##    bening malignant
##       458       241
```

Since 2 and 4 are the only numbers in the data:

```
formatted_breast_cancer$Class %>% unique
## [1] 2 4
```

But it is always a little risky to assume that there are no unexpected values, so I always prefer to have "weird values" as something I handle explicitly by setting them to NA.

Nested `ifelse()` are easy enough to program, but if there are many different possible values, it also becomes somewhat cumbersome. Another option is to use a table to map between values. To avoid confusion between a table as the one you are going to implement and the function `table()`, which counts how many times a given value appears in a vector, I am going to call the table you create a dictionary. A *dictionary* is a table where you can look up words, and that is what you are implementing.

For this, you can use named values in a vector. Remember that you can index in a vector both using numbers *and* using names.

You can create a vector where you use names as the indices. Use the keys you want to map from as the indices and the names you want as results as the values. We want to map from numbers to strings, which poses a small problem. If we index into a vector with numbers, R will think we want to get positions in the vector. If we make the vector `v <- c(2 = "benign", 4 = "malignant")`, which we can't, it is a syntax error and for good reasons, then how should `v[2]` be interpreted? Do we want the value at index 2, `"malignant"`, or the value that has *key* 2, `"benign"`? When we use a vector as a table, we need to have strings as keys. That also means that the numbers in the vector we want to map from should be converted to strings before we look up in the dictionary. The code looks like this:

```
dict <- c("2" = "benign", "4" = "malignant")
map_class <- function(x) dict[as.character(x)]

mapped <- formatted_breast_cancer$Class %>% map_class
mapped %>% table
## .
##    benign malignant
##       458       241
```

That worked fine, but if we look at the actual vector instead of summarizing it, we will see that it looks a little strange.

```
mapped[1:5]
##         2        2        2        2        2
## "benign" "benign" "benign" "benign" "benign"
```

This is because when we create a vector by mapping in this way we preserve the names of the values. Remember that the dictionary we made to map our keys to values has the keys as names; these names are passed on to the resulting vector. We can get rid of them using the `unname()` function.

```
mapped %<>% unname
mapped[1:5]
## [1] "benign" "benign" "benign" "benign" "benign"
```

Now we just need to translate this vector of strings into a factor, and we will have our `Class` column.

The `BreastCancer` dataset actually represent the `Id` column as strings and all the other columns as categorical (some ordered, some not), but I am not going to bother with that. If you want to transform the data this way, you know how to do it.

The entire reading of data and formatting can be done like this:

```
read.csv(data_url, header = FALSE,
         col.names = names(BreastCancer)) ->
  raw_breast_cancer ->
  formatted_breast_cancer

dict <- c("2" = "benign", "4" = "malignant")
map_class <- function(x) dict[as.character(x)]
formatted_breast_cancer$Class <-
  formatted_breast_cancer$Class %>%
  map_class %>%
  unname %>%
  factor(levels = c("benign", "malignant"))
```

It is not strictly necessary to specify the levels in the factor() call, but I prefer always to do so explicitly. If there is an unexpected string in the input to factor(), it would end up being one of the levels, and I wouldn't know about it until much later. Specifying the levels explicitly alleviates that problem.

If you don't like writing and naming a function just to map the class representation—and why would you want to pollute your namespace with a map_class() function you won't remember what does a few weeks later?—you can use a lambda expression:

```
raw_breast_cancer$Class %>%
  { dict <- c("2" = "benign", "4" = "malignant")
    dict[as.character(.)]
  } %>%
  unname %>%
  factor(levels = c("benign", "malignant")) %>%
  table
## .
##    benign malignant
##       458       241
```

Now, you don't want to spend time parsing input data files all the time so I recommend putting all the code you write to read in data and transforming it into the form you want in a cached code chunk in an R Markup document. This way you will only evaluate the code when you change it.

You can also explicitly save data using the save() function.

```
formatted_breast_cancer %>%
  save(file = "data/formatted-breast-cancer.rda")
```

Here I use the suffix .rda for the data. It stands for R data, and your computer will probably recognize it. If you click on a file with that suffix, it will be opened in RStudio (or whatever tool you use to work on R). The actual R functions for saving and loading data do not care what suffix you use, but it is easier to recognize the files for what they are if you stick to a fixed suffix.

The data is saved together with the name of the data frame, so when you load it again using the load() function, you don't have to assign the loaded data to a variable. It will be loaded into the name you used when you saved the data.

```
load("data/formatted-breast-cancer.rda")
```

This is both good and bad. I would probably have preferred to control which name the data is assigned to so I have explicit control over the variables in my code, but save() and load() are designed to save more than one variable, so this is how they work.

I personally do not use these functions that much. I prefer to write my analysis pipelines in Markdown documents, and there it is easier just to cache the import code.

Boston Housing Dataset

For the second example of loading data, we take another dataset from the mlbench package. The BostonHousing data contains information about crime rates and some explanatory variables we can use to predict crime rates.

```
library(mlbench)
data(BostonHousing)
str(BostonHousing)
## 'data.frame':    506 obs. of  14 variables:
## $ crim   : num  0.00632 0.02731 0.02729 0.03237..
## $ zn     : num  18 0 0 0 0 12.5 12.5 ...
## $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7..
## $ chas   : Factor w/ 2 levels "0","1": 1 1 1 1 ..
## $ nox    : num  0.538 0.469 0.469 0.458 0.458 0..
## $ rm     : num  6.58 6.42 7.18 7 ...
## $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 6..
## $ dis    : num  4.09 4.97 4.97 6.06 ...
## $ rad    : num  1 2 2 3 3 3 5 5 ...
## $ tax    : num  296 242 242 222 222 222 311 311..
## $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 1..
## $ b      : num  397 397 393 395 ...
## $ lstat  : num  4.98 9.14 4.03 2.94 ...
## $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22...
```

As before, the link to the actual data is pretty long so I will give you a tinyURL to it: http://tinyurl.com/zq2u8vx. I have also saved the original URL in the variable data_url.

I have already looked at the file at the end of the URL and seen that it consists of whitespace-separated columns of data, so the function you need to load it is read.table().

```
boston_housing <- read.table(data_url)
str(boston_housing)
## 'data.frame':    506 obs. of  14 variables:
## $ V1 : num  0.00632 0.02731 0.02729 0.03237 ...
## $ V2 : num  18 0 0 0 0 12.5 12.5 ...
## $ V3 : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 ..
## $ V4 : int  0 0 0 0 0 0 0 0 ...
## $ V5 : num  0.538 0.469 0.469 0.458 0.458 0.458..
## $ V6 : num  6.58 6.42 7.18 7 ...
## $ V7 : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 ..
## $ V8 : num  4.09 4.97 4.97 6.06 ...
## $ V9 : int  1 2 2 3 3 3 5 5 ...
## $ V10: num  296 242 242 222 222 222 311 311 ...
## $ V11: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 ..
## $ V12: num  397 397 393 395 ...
## $ V13: num  4.98 9.14 4.03 2.94 ...
## $ V14: num  24 21.6 34.7 33.4 36.2 28.7 22.9 27..
```

If we compare the data that we have loaded with the data from `mlbench`, we see that we have integers and numeric data in our imported data but that it should be a factor for the `chas` variable and numeric for all the rest. We can use the `colClasses` parameter for `read.table()` to fix this. We just need to make a vector of strings for the classes; a vector that is `"numeric"` for all columns except for the `"chas"` column, which should be `"factor"`.

```
col_classes <- rep("numeric", length(BostonHousing))
col_classes[which("chas" == names(BostonHousing))] <- "factor"
```

We should also name the columns, but again we can cheat and get the names from `BostonHousing`:

```
boston_housing <- read.table(data_url,
                             col.names = names(BostonHousing),
                             colClasses = col_classes)
str(boston_housing)
## 'data.frame':    506 obs. of  14 variables:
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237..
##  $ zn     : num  18 0 0 0 0 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7..
##  $ chas   : Factor w/ 2 levels "0","1": 1 1 1 1 ..
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0..
##  $ rm     : num  6.58 6.42 7.18 7 ...
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 6..
##  $ dis    : num  4.09 4.97 4.97 6.06 ...
##  $ rad    : num  1 2 2 3 3 3 5 5 ...
##  $ tax    : num  296 242 242 222 222 222 311 311..
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 1..
##  $ b      : num  397 397 393 395 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22...
```

The levels in the `"chas"` factor are 0 and 1. It is not really good levels as they are very easily confused with numbers—they will print like numbers—but they are not. The numerical values in the factor are actually 1 for "0" and 2 for "1", so that can be confusing. But it is the same levels as the `mlbench` data frame, so I will just leave it the way it is as well.

The `readr` Package

The `read.table()` class of functions will usually get you to where you want to go with importing data. I use these in almost all my work. But there is a package aimed at importing data that tries to speed up the importing and be more consistent in how data is imported, so I think I should mention it.

That package is `readr`.

```
library(readr)
```

It implements the same class of import functions as the built-in functions. It just uses underscores instead of dots in the function names. So where you would use `read.table()`, the `readr` package gives you `read_table()`. Similarly, it gives you `read_csv()` as a substitute for `read.csv()`.

The readr package has different defaults for how to read data. For instance, it doesn't by default consider string columns as factors. Other than that, its main claim to fame is that it's faster than the built-in R functions. This shouldn't concern you much if you put your data import code in a cached code chunk. In any case, if loading data is an issue, you need to read Chapter 5.

Anyway, because the package exists, and because it is popular, I thought I should mention it.

Let's look at how to import data using the functions in the package. We return to the breast cancer data we imported earlier. We downloaded the breast cancer data and put it in a file called data/raw-breast-cancer.csv, so we can try to read it from that file. Obviously, since it is a CSV file, we will use the read_csv() function.

```
raw_breast_cancer <- read_csv("data/raw-breast-cancer.csv")
## Warning: Duplicated column names deduplicated: '1'
## => '1_1' [4], '1' => '1_2' [5], '1' => '1_3' [7],
## '1' => '1_4' [9], '1' => '1_5' [10], '2' =>
## '2_1' [11]
## Parsed with column specification:
## cols(
##   `1000025` = col_integer(),
##   `5` = col_integer(),
##   `1` = col_integer(),
##   `1_1` = col_integer(),
##   `1_2` = col_integer(),
##   `2` = col_integer(),
##   `1_3` = col_character(),
##   `3` = col_integer(),
##   `1_4` = col_integer(),
##   `1_5` = col_integer(),
##   `2_1` = col_integer()
## )
raw_breast_cancer %>% head(3)
## # A tibble: 3 × 11
##   `1000025`   `5`   `1` `1_1` `1_2`   `2` `1_3`
##       <int> <int> <int> <int> <int> <int> <chr>
## 1   1002945     5     4     4     5     7    10
## 2   1015425     3     1     1     1     2     2
## 3   1016277     6     8     8     1     3     4
## # ... with 4 more variables: `3` <int>,
## #   `1_4` <int>, `1_5` <int>, `2_1` <int>
```

The function works similar to the read.csv() function and interprets the first as the column names. We don't want that, but this function doesn't have the option to tell it that the first line is not the names of the columns. Instead, we can inform it what the names of the columns are and then it will read the first line as actual data.

```
raw_breast_cancer <- read_csv("data/raw-breast-cancer.csv",
                             col_names = names(BreastCancer))
## Parsed with column specification:
## cols(
##   Id = col_integer(),
##   Cl.thickness = col_integer(),
##   Cell.size = col_integer(),
##   Cell.shape = col_integer(),
##   Marg.adhesion = col_integer(),
##   Epith.c.size = col_integer(),
##   Bare.nuclei = col_character(),
##   Bl.cromatin = col_integer(),
##   Normal.nucleoli = col_integer(),
##   Mitoses = col_integer(),
##   Class = col_integer()
## )
raw_breast_cancer %>% head(3)
## # A tibble: 3 × 11
##        Id Cl.thickness Cell.size Cell.shape
##     <int>        <int>     <int>      <int>
## 1 1000025            5         1          1
## 2 1002945            5         4          4
## 3 1015425            3         1          1
## # ... with 7 more variables: Marg.adhesion <int>,
## #   Epith.c.size <int>, Bare.nuclei <chr>,
## #   Bl.cromatin <int>, Normal.nucleoli <int>,
## #   Mitoses <int>, Class <int>
```

It imports data similarly to the functions we have already seen, but the printed results are slightly different. This is just because it represents data frames slightly differently. There are different ways of representing data frames, we will also read about this more soon, and the readr packages loads data in a different format. That is why it prints differently. The way you interact with this representation is the same as with any other data frame, so it doesn't matter for this use.

Which functions you use to import data doesn't much matter. You can use the built-in functions or the readr functions. It is up to you.

Manipulating Data with `dplyr`

Data frames are ideal for representing data where each row is an observation of different parameters you want to fit in models. Nearly all packages that implement statistical models or machine learning algorithms in R work on data frames. But to actually manipulate a data frame, you often have to write a lot of code to filter data, rearrange data, and summarize it in various ways. A few years ago, manipulating data frames required a lot more programming than actually analyzing data. That has improved dramatically with the dplyr package (pronounced "d plier" where "plier" is pronounced as the tool, "pliers").

This package provides a number of convenient functions that let you modify data frames in various ways and string them together in pipes using the %>% operator. As far as I know, this operator was first used in this package, and if you only import dplyr you get the operator as well. However, magrittr implements several extensions so I suggest you always import magrittr as well.

If you import dplyr, you get functions that let you build pipelines for data frame manipulation.

Some Useful `dplyr` Functions

I will not be able to go through all of the `dplyr` functionality in this chapter. In any case, it is updated frequently enough that, by the time you read this, there will probably be more functionality than at the time I wrote the chapter. So be sure to check the package's documentation for updates.

This section describes the functions I use on a regular basis. They all take a data frame or equivalent as the first argument so they work perfectly with pipelines. When I say "data frame equivalent," I mean that they take as an argument anything that works like a data frame. Quite often there are better representations of data frames than the built-in data structure. For large datasets, it is often better to use a different representation than the built-in data frame; something we will return to in Chapter 5. Some alternative data structures are better because they can work with data on disk. R's data frames have to be loaded into memory, and others are just faster to do some operations on. Or maybe they just print better. If you write the name of a data frame into the R terminal, it will print the entire data. Other representations will automatically give you the head of the data.

The `dplyr` package has several representations. One I use a lot is the `tbl_df` representation. I use it just because I prefer the output when I print such tables.

```
iris %>% tbl_df
## # A tibble: 150 × 5
##    Sepal.Length Sepal.Width Petal.Length
##           <dbl>       <dbl>        <dbl>
## 1           5.1         3.5          1.4
## 2           4.9         3.0          1.4
## 3           4.7         3.2          1.3
## 4           4.6         3.1          1.5
## 5           5.0         3.6          1.4
## 6           5.4         3.9          1.7
## 7           4.6         3.4          1.4
## 8           5.0         3.4          1.5
## 9           4.4         2.9          1.4
## 10          4.9         3.1          1.5
## # ... with 140 more rows, and 2 more variables:
## #   Petal.Width <dbl>, Species <fctr>
```

It only prints the first ten rows, and it doesn't print all columns. The output is a little easier to read than if you get the entire data frame.

Now on to the `dplyr` functions.

select(): Pick Selected Columns and Get Rid of the Rest

The `select()` function selects columns of the data frame. It is equivalent to indexing columns in the data.

You can use it to pick out a single column:

```
iris %>% tbl_df %>% select(Petal.Width) %>% head(3)
## # A tibble: 3 × 1
##    Petal.Width
##          <dbl>
## 1          0.2
## 2          0.2
## 3          0.2
```

Or pick several columns:

```
iris %>% tbl_df %>%
  select(Sepal.Width, Petal.Length) %>% head(3)
## # A tibble: 3 × 2
##   Sepal.Width Petal.Length
##         <dbl>        <dbl>
## 1         3.5          1.4
## 2         3.0          1.4
## 3         3.2          1.3
```

You can even give it ranges of columns:

```
iris %>% tbl_df %>%
  select(Sepal.Length:Petal.Length) %>% head(3)
## # A tibble: 3 × 3
##   Sepal.Length Sepal.Width Petal.Length
##          <dbl>       <dbl>        <dbl>
## 1          5.1         3.5          1.4
## 2          4.9         3.0          1.4
## 3          4.7         3.2          1.3
```

How that works depends on the order of the columns in the data frame and it is not something I find all that useful.

The real usefulness comes with pattern matching on column names. There are different ways to pick columns based on the column names:

```
iris %>% tbl_df %>%
  select(starts_with("Petal")) %>% head(3)
## # A tibble: 3 × 2
##   Petal.Length Petal.Width
##          <dbl>       <dbl>
## 1          1.4         0.2
## 2          1.4         0.2
## 3          1.3         0.2
iris %>% tbl_df %>%
  select(ends_with("Width")) %>% head(3)
## # A tibble: 3 × 2
##   Sepal.Width Petal.Width
##         <dbl>       <dbl>
## 1         3.5         0.2
## 2         3.0         0.2
## 3         3.2         0.2
iris %>% tbl_df %>%
  select(contains("etal")) %>% head(3)
## # A tibble: 3 × 2
##   Petal.Length Petal.Width
##          <dbl>       <dbl>
## 1          1.4         0.2
## 2          1.4         0.2
## 3          1.3         0.2
```

```
iris %>% tbl_df %>%
  select(matches(".t.")) %>% head(3)
## # A tibble: 3 × 4
##   Sepal.Length Sepal.Width Petal.Length
##          <dbl>       <dbl>        <dbl>
## 1          5.1         3.5          1.4
## 2          4.9         3.0          1.4
## 3          4.7         3.2          1.3
## # ... with 1 more variables: Petal.Width <dbl>
```

The matches function searches for a regular expression, and in this example it will select any name that contains a t except if it is the first or last letter.

Check out the documentation for dplyr to see which options you have for selecting columns.

You can also use select() to remove columns. The previous examples select the columns you want to include, but if you use - before the selection criteria, you will exclude, instead of include, the columns you specify.

```
iris %>% tbl_df %>%
  select(-starts_with("Petal")) %>% head(3)
## # A tibble: 3 × 3
##   Sepal.Length Sepal.Width Species
##          <dbl>       <dbl>  <fctr>
## 1          5.1         3.5  setosa
## 2          4.9         3.0  setosa
## 3          4.7         3.2  setosa
```

mutate():Add Computed Values to Your Data Frame

The mutate() function lets you add a column to your data frame by specifying an expression for how to compute it:

```
iris %>% tbl_df %>%
  mutate(Petal.Width.plus.Length = Petal.Width + Petal.Length) %>%
  select(Species, Petal.Width.plus.Length) %>%
  head(3)
## # A tibble: 3 × 2
##   Species Petal.Width.plus.Length
##    <fctr>                   <dbl>
## 1  setosa                     1.6
## 2  setosa                     1.6
## 3  setosa                     1.5
```

You can add more columns than one by specifying them in the mutate() function:

```
iris %>% tbl_df %>%
  mutate(Petal.Width.plus.Length = Petal.Width + Petal.Length,
         Sepal.Width.plus.Length = Sepal.Width + Sepal.Length) %>%
  select(Petal.Width.plus.Length, Sepal.Width.plus.Length) %>%
  head(3)
```

```
## # A tibble: 3 × 2
##    Petal.Width.plus.Length Sepal.Width.plus.Length
##                      <dbl>                   <dbl>
## 1                      1.6                     8.6
## 2                      1.6                     7.9
## 3                      1.5                     7.9
```

You could of course also just call `mutate()` several times in your pipeline.

Transmute(): Add Computed Values to Your Data Frame and Get Rid of All Other Columns

The `transmute()` function works just like the `mutate()` function, except it is combined with a `select()` so the result is a data frame that only contains the new columns you make.

```
iris %>% tbl_df %>%
  transmute(Petal.Width.plus.Length = Petal.Width + Petal.Length) %>%
  head(3)
## # A tibble: 3 × 1
##    Petal.Width.plus.Length
##                      <dbl>
## 1                      1.6
## 2                      1.6
## 3                      1.5
```

arrange(): Reorder Your Data Frame by Sorting Columns

The `arrange()` function just reorders the data frame by sorting columns according to what you specify:

```
iris %>% tbl_df %>%
  arrange(Sepal.Length) %>%
  head(3)
## # A tibble: 3 × 5
##    Sepal.Length Sepal.Width Petal.Length
##           <dbl>       <dbl>        <dbl>
## 1           4.3         3.0          1.1
## 2           4.4         2.9          1.4
## 3           4.4         3.0          1.3
## # ... with 2 more variables: Petal.Width <dbl>,
## #   Species <fctr>
```

By default, it orders numerical values in increasing order, but you can ask for decreasing order using the desc() function:

```
iris %>% tbl_df %>%
  arrange(desc(Sepal.Length)) %>%
  head(3)
## # A tibble: 3 × 5
##   Sepal.Length Sepal.Width Petal.Length
##          <dbl>       <dbl>        <dbl>
## 1          7.9         3.8          6.4
## 2          7.7         3.8          6.7
## 3          7.7         2.6          6.9
## # ... with 2 more variables: Petal.Width <dbl>,
## #   Species <fctr>
```

filter(): Pick Selected Rows and Get Rid of the Rest

The filter() function lets you pick out rows based on logical expressions. You give the function a predicate, specifying what a row should satisfy to be included.

```
iris %>% tbl_df %>%
  filter(Sepal.Length > 5) %>%
  head(3)
## # A tibble: 3 × 5
##   Sepal.Length Sepal.Width Petal.Length
##          <dbl>       <dbl>        <dbl>
## 1          5.1         3.5          1.4
## 2          5.4         3.9          1.7
## 3          5.4         3.7          1.5
## # ... with 2 more variables: Petal.Width <dbl>,
## #   Species <fctr>
```

You can get as inventive as you want here with the logical expressions:

```
iris %>% tbl_df %>%
  filter(Sepal.Length > 5 & Species == "virginica") %>%
  select(Species, Sepal.Length) %>%
  head(3)
## # A tibble: 3 × 2
##     Species Sepal.Length
##      <fctr>        <dbl>
## 1 virginica          6.3
## 2 virginica          5.8
## 3 virginica          7.1
```

group_by(): Split Your Data Into Subtables Based on Column Values

The group_by() function tells dplyr that you want to work on data separated into different subsets.

By itself, it isn't that useful. It just tells dplyr that, in future computations, it should consider different subsets of the data as separate datasets. It is used with the summarise() function, where you want to compute summary statistics.

You can group by one or more variables; you just specify the columns you want to group by as separate arguments to the function. It works best when grouping by factors or discrete numbers; there isn't much fun in grouping by real numbers.

```
iris %>% tbl_df %>% group_by(Species) %>% head(3)
## Source: local data frame [3 x 5]
## Groups: Species [1]
##
##   Sepal.Length Sepal.Width Petal.Length
##          <dbl>       <dbl>        <dbl>
## 1          5.1         3.5          1.4
## 2          4.9         3.0          1.4
## 3          4.7         3.2          1.3
## # ... with 2 more variables: Petal.Width <dbl>,
## #   Species <fctr>
```

Not much is happening here. You have restructured the data frame such that there are groupings, but until you do something with the new data, there isn't much to see. The power of group_by() is when it's combined with the summarise() function.

summarise/summarize(): Calculate Summary Statistics

The spelling of this function depends on which side of the pond you are on. It is the same function regardless of how you spell it.

The summarise() function is used to compute summary statistics from your data frame. It lets you compute different statistics by expressing what you want to summarize. For example, you can ask for the mean of values:

```
iris %>%
  summarise(Mean.Petal.Length = mean(Petal.Length),
            Mean.Sepal.Length = mean(Sepal.Length))
##   Mean.Petal.Length Mean.Sepal.Length
## 1             3.758          5.843333
```

Where it is really powerful is in the combination with group_by(). There you can split the data into different groups and compute the summaries for each group:

```
iris %>%
  group_by(Species) %>%
  summarise(Mean.Petal.Length = mean(Petal.Length))
## # A tibble: 3 × 2
##      Species Mean.Petal.Length
##       <fctr>             <dbl>
## 1     setosa             1.462
## 2 versicolor             4.260
## 3  virginica             5.552
```

A summary function worth mentioning here is n(), which just counts how many observations you have in a subset of your data:

```
iris %>%
  summarise(Observations = n())
##   Observations
## 1          150
```

Again, this is more interesting when combined with group_by():

```
iris %>%
  group_by(Species) %>%
  summarise(Number.Of.Species = n())
## # A tibble: 3 × 2
##      Species Number.Of.Species
##       <fctr>             <int>
## 1     setosa                50
## 2 versicolor                50
## 3  virginica                50
```

You can combine summary statistics simply by specifying more than one in the summary() function:

```
iris %>%
  group_by(Species) %>%
  summarise(Number.Of.Samples = n(),
            Mean.Petal.Length = mean(Petal.Length))
## # A tibble: 3 × 3
##      Species Number.Of.Samples Mean.Petal.Length
##       <fctr>             <int>             <dbl>
## 1     setosa                50             1.462
## 2 versicolor                50             4.260
## 3  virginica                50             5.552
```

Breast Cancer Data Manipulation

To get a little more feeling for how the dplyr package can help you explore data, let's see it in action.

We'll return to the breast cancer data. We start with the modifications we used to transform the raw data we imported from the CVS file (stored in the variable raw_breast_cancer).

When we formatted this dataset, we had to structure the factor for the Class variable. We did this by explicitly assigning to the Class variable using formatted_breast_cancer$Class, but we can do it directly as a data frame transformation using the mutate() function from dplyr.

```
formatted_breast_cancer <-
  raw_breast_cancer %>%
  mutate(Class = Class %>% {
      c("2" = "benign", "4" = "malignant")[as.character(.)]
    } %>%
    unname %>%
    factor(levels = c("benign", "malignant")) )
```

Here, we cannot assign to a `dict` variable inside the `mutate()` function, so I had to put the dictionary construction before the subscript. It isn't pretty, and it may not be that readable. This is one of the cases where I would probably use a function to do the mapping.

```
format_class <- . %>% {
  dict <- c("2" = "benign", "4" = "malignant")
  dict[as.character(.)]
} %>% unname %>% factor(levels = c("benign", "malignant"))

formatted_breast_cancer <-
  raw_breast_cancer %>% mutate(Class = format_class(Class))
```

Now whether this is more readable, I don't know. It might not be if you are not used to writing code as pipelines like this, but once you get used to reading pipeline code, it isn't too bad. In any case, it makes the transformation very clear, and there can be no doubt that we are creating the `formatted_breast_cancer` data frame by doing transformations on the `raw_breast_cancer` data frame.

Now let's look a little at the actual data. This is a very crude analysis of the data we can do for exploratory purposes. It is not a proper analysis, but we will return to that in Chapter 6.

We could be interested in how the different parameters affect the response variable, the `Class` variable. For instance, are cell thickness different for benign and malignant tumors? To check that, we can group the data by the `Cell` parameter and look at the mean cell thickness.

```
formatted_breast_cancer %>%
  group_by(Class) %>%
  summarise(mean.thickness = mean(Cl.thickness))
## # A tibble: 2 × 2
##       Class mean.thickness
##      <fctr>          <dbl>
## 1    benign       2.956332
## 2 malignant       7.195021
```

It looks like there is a difference. Now whether this difference is significant requires a proper test—after all, we are just comparing means here, and the variance could be huge. But just by exploring the data, we get a hint that there might be something to work with.

We could ask the same question for other variables, like cell size:

```
formatted_breast_cancer %>%
  group_by(Class) %>%
  summarise(mean.size = mean(Cell.size))
## # A tibble: 2 × 2
##       Class mean.size
##      <fctr>     <dbl>
## 1    benign  1.325328
## 2 malignant  6.572614
```

Another way of looking at this could be to count, for each cell size, how many benign tumors and how many malignant tumors we see. Here, we need to group by both cell size and class and then count, and we would probably want to arrange the data so we get the information in order of increasing or decreasing cell size:

```
formatted_breast_cancer %>%
  arrange(Cell.size) %>%
  group_by(Cell.size, Class) %>%
  summarise(ClassCount = n())
## Source: local data frame [18 x 3]
## Groups: Cell.size [?]
##
##    Cell.size      Class ClassCount
##        <int>     <fctr>      <int>
## 1          1     benign        380
## 2          1  malignant          4
## 3          2     benign         37
## 4          2  malignant          8
## 5          3     benign         27
## 6          3  malignant         25
## 7          4     benign          9
## 8          4  malignant         31
## 9          5  malignant         30
## 10         6     benign          2
## 11         6  malignant         25
## 12         7     benign          1
## 13         7  malignant         18
## 14         8     benign          1
## 15         8  malignant         28
## 16         9     benign          1
## 17         9  malignant          5
## 18        10  malignant         67
```

Here again, we get some useful information. It looks like there are more benign tumors compared to malignant tumors when the cell size is small and more malignant tumors when the cell size is large. Again, this is something we can start to work from when we later want to build statistical models.

This kind of grouping only works because the cell size is measured as discrete numbers. It wouldn't be helpful to group by a floating-point number. There plotting is more useful. But for this data, we have the cell size as integers, so we can explore the data just by building tables in this way.

We can also try to look at combined parameters. We have already seen that both cell size and cell thickness seem to be associated with how benign or malignant a tumor is, so let's try to see how the cell thickness behaves as a function of both class and cell size.

```
formatted_breast_cancer %>%
  group_by(Class, as.factor(Cell.size)) %>%
  summarise(mean.thickness = mean(Cl.thickness))
## Source: local data frame [18 x 3]
## Groups: Class [?]
##
##          Class `as.factor(Cell.size)`
##          <fctr>                 <fctr>
## 1     benign                       1
## 2     benign                       2
## 3     benign                       3
## 4     benign                       4
## 5     benign                       6
## 6     benign                       7
## 7     benign                       8
## 8     benign                       9
## 9  malignant                       1
## 10 malignant                       2
## 11 malignant                       3
## 12 malignant                       4
## 13 malignant                       5
## 14 malignant                       6
## 15 malignant                       7
## 16 malignant                       8
## 17 malignant                       9
## 18 malignant                      10
## # ... with 1 more variables: mean.thickness <dbl>
```

I am not sure how much I learn from this. It seems that for the benign tumors, the thickness increases with the cell size but for the malignant ones, there isn't that pattern.

Maybe we can learn more by ordering the data in a different way. What if we look at the numbers of benign and malignant tumors for each cell size and see what the thickness is?

```
formatted_breast_cancer %>%
  group_by(as.factor(Cell.size), Class) %>%
  summarise(mean.thickness = mean(Cl.thickness))
## Source: local data frame [18 x 3]
## Groups: as.factor(Cell.size) [?]
##
##    `as.factor(Cell.size)`     Class
##                    <fctr>    <fctr>
## 1                      1     benign
## 2                      1 malignant
## 3                      2     benign
## 4                      2 malignant
## 5                      3     benign
## 6                      3 malignant
## 7                      4     benign
## 8                      4 malignant
## 9                      5 malignant
## 10                     6     benign
```

```
## 11                     6 malignant
## 12                     7    benign
## 13                     7 malignant
## 14                     8    benign
## 15                     8 malignant
## 16                     9    benign
## 17                     9 malignant
## 18                    10 malignant
## # ... with 1 more variables: mean.thickness <dbl>
```

I am not sure how much we learned from that either, but at least it looks like for each cell size where we have both benign and malignant tumors, the thickness is greater with the malignant tumors than with the benign. That is something at least. A place to start the analysis. You can learn more when you start plotting data and when do a proper statistical analysis. We return to that process in later chapters. For now, we leave it at that.

Tidying Data with `tidyr`

I am not really sure where the concept of "tidy data" comes from. Hadley Wickham, the author of many of the essential packages you will use in your R data analysis, describes tidy data as such (from http://vita.had.co.nz/papers/tidy-data.pdf):

> *Tidy data is a standard way of mapping the meaning of a dataset to its structure. A dataset is messy or tidy depending on how rows, columns, and tables are matched up with observations, variables, and types.*

In my experience, tidy data means that I can plot or summarize the data efficiently. It mostly comes down to which data is represented as columns in a data frame and which is not.

In practice, this means that I have columns in my data frame that I can work with for the analysis I want to do. For example, if I want to look at the `iris` dataset and see how the `Petal.Length` varies among species, I can look at the `Species` column against the `Petal.Length` column.

```
iris %>% select(Species, Petal.Length) %>% head(3)
##   Species Petal.Length
## 1  setosa          1.4
## 2  setosa          1.4
## 3  setosa          1.3
```

I have a column specifying the `Species` and another specifying the `Petal.Length` and it is easy enough to look at their correlation. I can plot one against the other (we cover visualization in the next chapter). I can let the x-axis be species and the y-axis be `Petal.Length` (see Figure 3-2).

```
iris %>% select(Species, Petal.Length) %>%
  qplot(Species, Petal.Length, geom = "boxplot", data = .)
```

Figure 3-2. *Plotting species versus petal length*

This works because I have a column for the x-axis and another for the y-axis. But what happens if I want to plot the different measurements of the irises to see how those are? Each measurement is a separate column. They are `Petal.Length`, `Petal.Width`, and so on.

Now I have a bit of a problem because the different measurements are in different columns in my data frame. I cannot easily map them to an x-axis and a y-axis.

The `tidyr` package addresses that.

```
library(tidyr)
```

It has a function, `gather()`, that modifies the data frame so columns become names in a factor and other columns become values.

What it does is essentially transform the data frame so that you get one column containing the name of your original columns and another column containing the values in those columns.

In the `iris` dataset, we have observations for sepal length and sepal width. If we want to examine `Species` versus `Sepal.Length` or `Sepal.Width`, we can readily do this. We have more of a problem if we want to examine for each species both measurements at the same time. The data frame just doesn't have the structure we need for that.

If we want to see `Sepal.Length` and `Sepal.Width` as two measurements, we can plot against their values. We would need to make a column in our data frame that tells us if a measurement is a length or a width and another column that shows us what the measurement actually is. The `gather()` function from `tidyr` lets you do that.

```
iris %>%
  gather(key = Attribute, value = Measurement,
         Sepal.Length, Sepal.Width) %>%
  select(Species, Attribute, Measurement) %>%
  head(3)
##   Species    Attribute Measurement
## 1  setosa Sepal.Length         5.1
## 2  setosa Sepal.Length         4.9
## 3  setosa Sepal.Length         4.7
```

This code tells `gather()` to make a column called `Attributes` that contains the names of columns from the input data frame and another called `Measurement` that will contain the values of the key columns. From the resulting data frame, you can see that the `Attribute` column contains the `Sepal.Length` and `Sepal.Width` names (well, you can see it if you don't run it through `head()`; in the output here you only see `Sepal.Length`), and another column that shows the `Measurements`.

This transforms the data into a form where we can plot the attributes against measurements (see Figure 3-3 for the result).

```
iris %>%
  gather(key = Attribute, value = Measurement,
         Sepal.Length, Sepal.Width) %>%
  select(Species, Attribute, Measurement) %>%
  qplot(Attribute, Measurement,
        geom = "boxplot",
        facets = . ~ Species, data = .)
```

Figure 3-3. *Plot measurements versus values*

The tidyr package contains functions for both mapping columns to values and for mapping back from values to columns. The gather() function is the one I regularly use so that is the only one I will mention here.

Exercises

It is time to put what you have learned into practice. There are only a few exercises, but I hope you will do them. You can't learn without doing exercises, after all.

Importing Data

To get a feeling of the steps in importing and transforming data, you need to try it yourself. So try finding a dataset you want to import. You can do that from one of the repositories I listed in the Introduction:

- RDataMining.com (http://www.rdatamining.com/resources/data)

- UCI machine learning repository (http://archive.ics.uci.edu/ml/)

- KDNuggets (http://www.kdnuggets.com/datasets/index.html)

- Reddit r/datasets (https://www.reddit.com/r/datasets)

- GitHub awesome public datasets (https://github.com/caesar0301/awesome-public-datasets)

Or maybe you already have a dataset you would like to analyze.

Have a look at your dataset and figure out which `import` function you need. You might have to set a few parameters in the function to get the data loaded correctly, but with a bit of effort, you should be able to. For column names, you should either choose some appropriate ones from reading the data description or if you are loading something in that is already in `mlbench`, you can cheat as I did in the examples.

Using `dplyr`

Now take the data you just imported and examine various summaries. It is not so important what you look at in the data, as it is that you try to summarize different aspects of it. We will look at proper analyses later. For now, just use `dplyr` to explore your data.

Using `tidyr`

Look at the `dplyr` example in this chapter. There I plotted `Sepal.Length` and `Sepal.Width` for each species. Do the same thing for `Petal.Length` and `Petal.Width`.

If there is something similar you can do with the dataset you imported in the first exercise, try doing it with that dataset as well.

CHAPTER 4

■ ■ ■

Visualizing Data

Nothing really tells a story about your data as powerfully as good plots. Graphics capture your data much better than summary statistics and often show you features that you would not be able to glean from summaries alone.

R has very powerful tools for visualizing data. Unfortunately, it also has more tools than you'll really know what to do with. There are several different frameworks for visualizing data, and they are usually not particularly compatible, so you cannot easily combine the various approaches.

In this chapter, we look at graphics in R. We cannot possibly cover all the plotting functionality, so I will focus on a few frameworks. First, the basic graphics framework. It is not something I frequently use or recommend that you use, but it is the default for many packages so you need to know about it. Secondly, we discuss the ggplot2 framework, which is my preferred approach to visualizing data. It defines a small domain-specific language for constructing data and is perfect for exploring data as long as you have it in a data frame (and with a little bit more work for creating publication-ready plots).

Basic Graphics

The basic plotting system is implemented in the graphics package. You usually do not have to include the package:

```r
library(graphics)
```

It is already loaded when you start up R. But you can use this to get a list of the functions implemented in the package:

```r
library(help = "graphics")
```

This list isn't exhaustive, though, since the main plotting function, plot(), is generic and many packages write extensions to it to specialize plots.

In any case, you create basic plots using the plot() function. This function is a so-called generic function, which means that what it does depends on the input it gets. So you can give it different first arguments to get plots of various objects.

The simplest plot you can make is a scatterplot, which plot points for *x* and *y* values, as shown in Figure 4-1.

```r
x <- rnorm(50)
y <- rnorm(50)
plot(x, y)
```
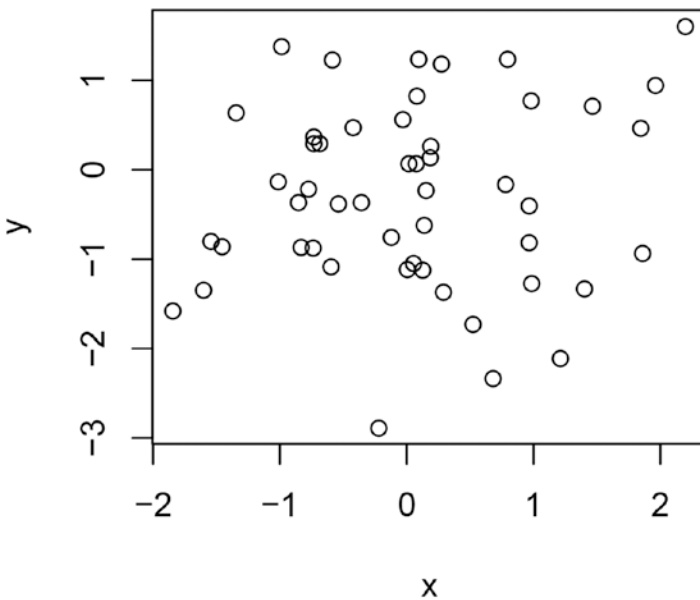
Figure 4-1. *A typical scatterplot*

The `plot()` function takes a data argument you can use to plot data from a data frame, but you cannot write code like this to plot the `cars` data from the `datasets` package:

```
data(cars)
cars %>% plot(speed, dist, data = .)
```

Despite giving `plot()` the data frame, it will not recognize the variables for the x and y parameters, and so adding plots to pipelines requires that you use the `%$%` operator to give `plot()` access to the variables in a data frame. So, for instance, we can plot the `cars` data like this:

```
cars %$% plot(speed, dist, main="Cars data",
              xlab="Speed", ylab="Stopping distance")
```

Figure 4-2 uses `main` as a title and `xlab` and `ylab` specify the axes labels.

Cars data
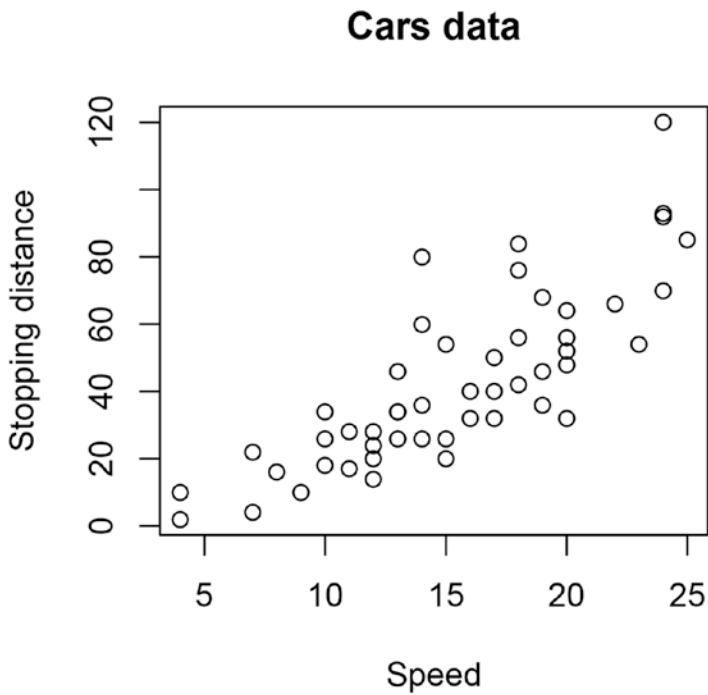
Figure 4-2. *A scatterplot of speed and distance for cars*

The data argument of `plot()` is used when the variables of the plot are specified as a formula. It is combined with a formula that the data parameter of the `plot()` function is used. If the x and y values are specified in a formula, you can give the function a data frame that holds the variables and plot from that, as follows:

```
cars %>% plot(dist ~ speed, data = .)
```

By default, the plot shows the data as points, but you can specify a type parameter to display the data in other ways, such as lines or histograms (see Figure 4-3).

```
cars %$% plot(speed, dist, main="Cars data", type="h",
              xlab="Speed", ylab="Stopping distance")
```
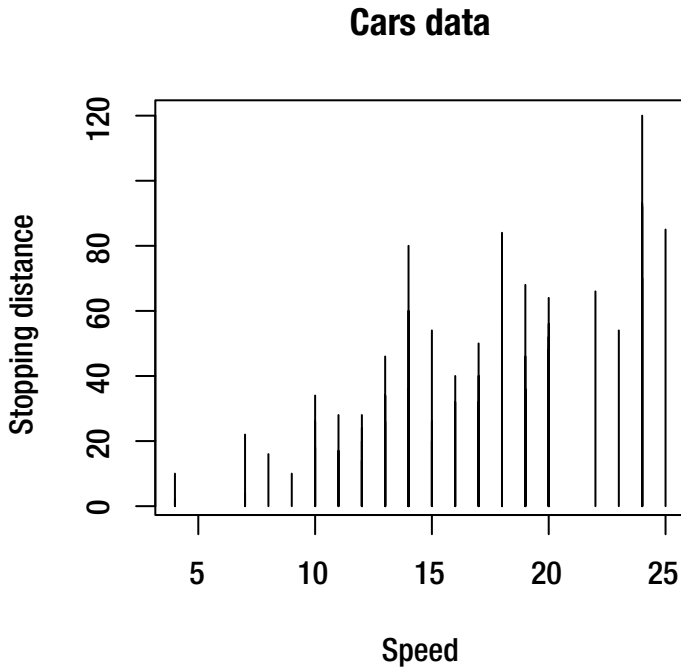

Figure 4-3. *Histogram plot of speed and distance for cars*

To get a histogram of a single variable, use the function `hist()` instead of `plot()`, as shown in Figure 4-4.

```
cars %$% hist(speed)
```
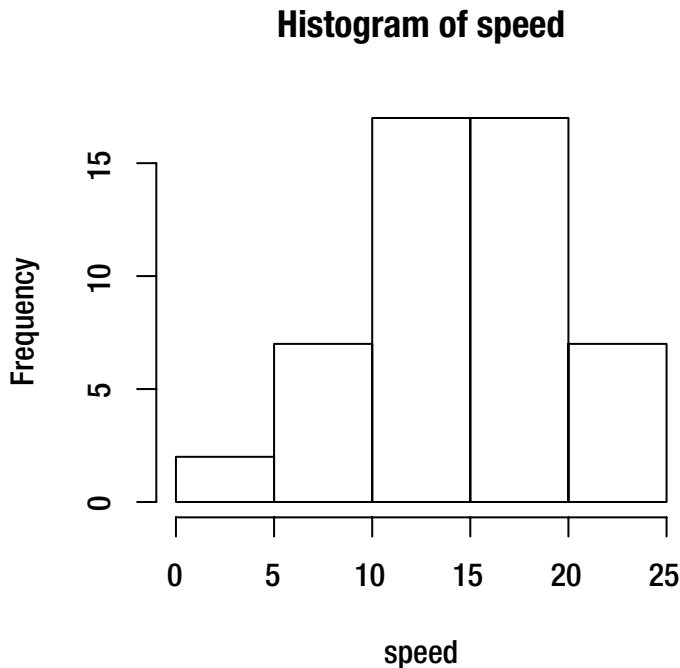
Histogram of speed

Figure 4-4. *Histogram for cars speed*

What is meant by `plot()` being a *generic* function (something we cover in much greater detail in Chapter 10) is that it will have different functionality depending on the parameters you give it.

Different kinds of objects can have their own plotting functionality, though, and they often do. This is why you probably will use basic graphics from time to time even if you follow my advice and use ggplot2 for your own plotting.

Linear regression, for example, created with the `lm()` function, has its own plotting routine. Try evaluating the following expression:

```
cars %>% lm(dist ~ speed, data = .) %>% plot
```

It will give you several summary plots for visualizing the quality of the linear fit.

Many model-fitting algorithms return a fitted object that has specialized plotting functionality like this, so when you have fitted a model, you can always try to call `plot()` on it and see if you get something useful out of that.

Functions like `plot()` and `hist()` and a few more creates new plots, but there is also a large number of functions for annotating a plot. Functions such as `lines()` or `points()` add lines and points, respectively, to the current plot rather than making a new plot.

You can see them in action if you plot the `longley` dataset and want to see both the unemployment rate and people in the armed forces over the years.

```
data(longley)
```

Check the documentation for `longley` (`?longley`) for a description of the data. The data has various statistics for each year from 1947 to 1962, including the number of people unemployed (variable `Unemployed`) and the number of people in the armed forces (variable `Armed.Forces`). To plot both of these on the same plot, you first plot `Unemployed` against years (variable `Year`) and then add lines for `Armed.Forces`. See Figure 4-5.

```
longley %>% plot(Unemployed ~ Year, data = ., type = 'l')
longley %>% lines(Armed.Forces ~ Year, data = ., col = "blue")
```
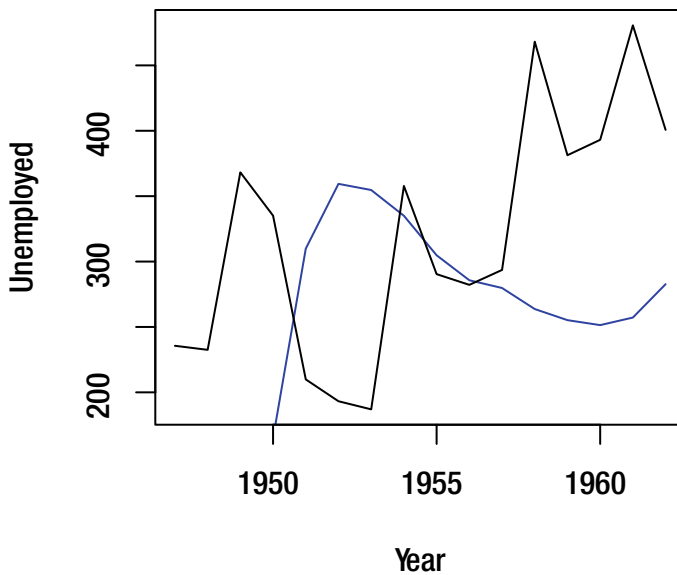

Figure 4-5. *Longley data showing Unemployed and Armed.Forces (the rounded line). The y-axis doesn't cover all of the Armed.Forces variable.*

This almost gets you what you want, but the y-axis is chosen by the `plot()` function to match the range of y-values in the call to `plot()` and `Armed.Forces` doesn't quite fit into this range. To fit both, you have to set the limits of the y-axis, which you do with parameter `ylim` (see Figure 4-6).

```
longley %$% plot(Unemployed ~ Year, type = 'l',
                 ylim = range(c(Unemployed, Armed.Forces)))
longley %>% lines(Armed.Forces ~ Year, data = ., col = "blue")
```
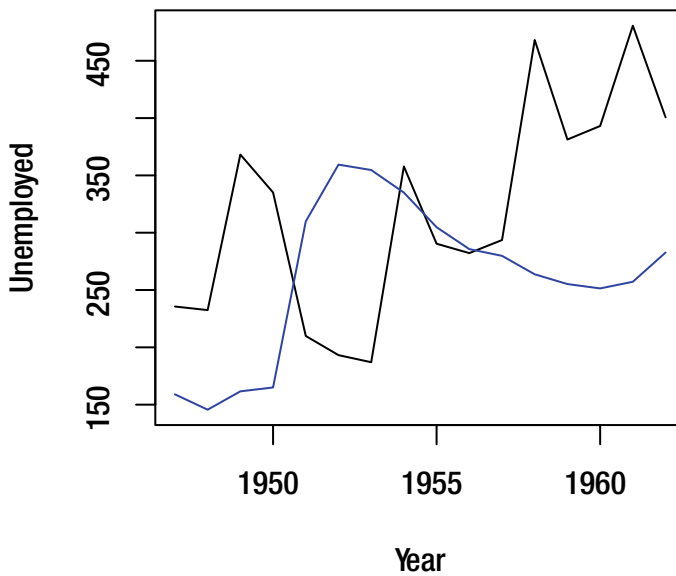
Figure 4-6. *Longley data showing Unemployed and Armed.Forces. The y-axis is wide enough to hold all the data.*

Like plot(), the other plotting functions are usually generic. This means you can sometimes give them objects such as fitted models. The abline() function is one such case. It plots lines of the form $y = a + bx$, but there is a variant of it that takes a linear model as input and plot the best fitting line defined by the model. So you can plot the cars data together with the best-fitted line using the combination of the lm() and abline() functions (see Figure 4-7).

```
cars %>% plot(dist ~ speed, data = .)
cars %>% lm(dist ~ speed, data = .) %>% abline(col = "red")
```
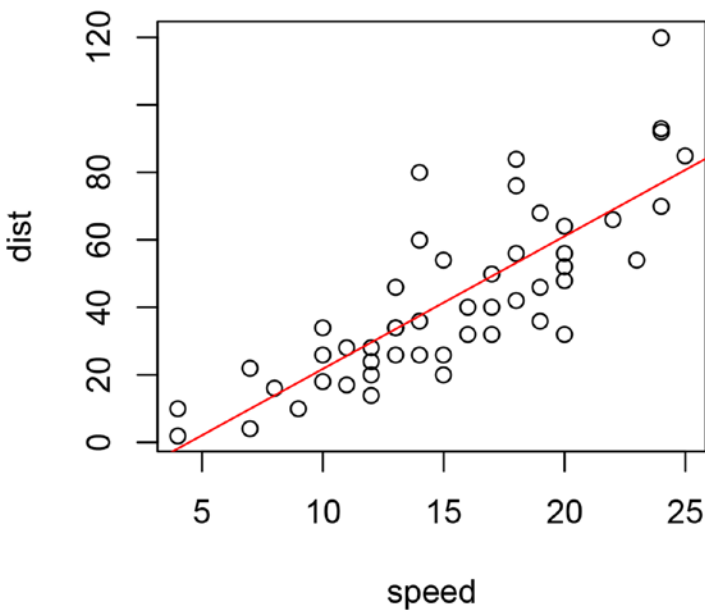
Figure 4-7. *The cars data points annotated with the best fitting line*

   Plotting using the basic graphics usually follows this pattern. First, there is a call to plot() that sets up the canvas to plot on—possibly adjusting the axes to make sure that later points will fit in on it. Then any additional data points are plotted—like the second time series you saw in the longley data. Finally, there might be some annotation like adding text labels or margin notes (see the text() and mtext() functions for this).
   If you want to select the shape of points or their color according to other data features, e.g., plotting the iris data with data points in different shapes according to the Species variable, you need to map features to columns (see Figure 4-8).

```
shape_map <- c("setosa" = 1,
               "versicolor" = 2,
               "virginica" = 3)
iris %$% plot(Petal.Length ~ Petal.Width,
              pch = shape_map[Species])
```
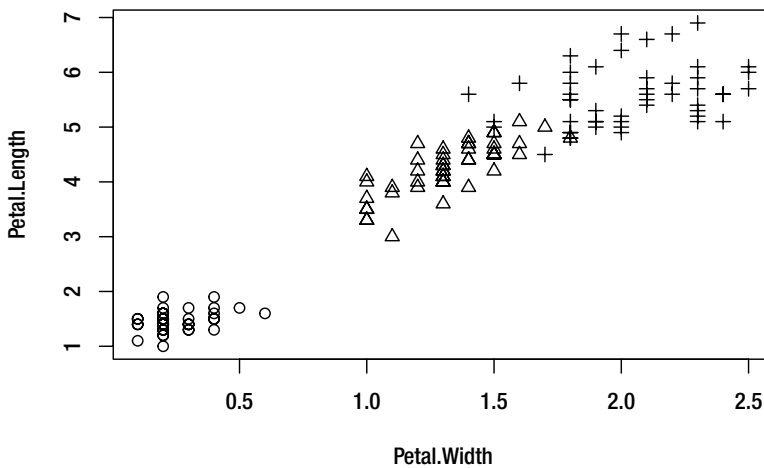
Figure 4-8. *Iris data plotted with different colors for different species*

The basic graphics system has many functions for making publication quality plots, but most of them work at a relatively low level. You have to map variables to colors or shapes explicitly if you want a variable to determine how points should be displayed. You have to set the `xlim` and `ylim` parameters to have the right *x* and *y* axis if the first points you plot do not cover the entire range of the data you want to plot. If you change an axis—say log-transform or if you flip the *x* and *y* axis—then you will usually need to update several function calls. If you want to have different subplots—so-called *facets*—for different subsets of your data, you have to subset and plot this explicitly.

So while the basic graphics system is powerful for making good-looking final plots, it is not necessarily optimal for exploring data, when you often want to try different ways of visualizing it.

The Grammar of Graphics and the ggplot2 Package

The `ggplot2` package provides an alternative to the basic graphics that is based on what is called the "grammar of graphics". The idea here is that the system gives you a small domain-specific language for creating plots (similar to how `dplyr` provides a domain-specific language for manipulating data frames). You construct plots through a list of function calls—similar to how you would do so with basic graphic—but these function calls do not directly write on a canvas independently of each other. Rather they all manipulate a plot by either modifying it—scaling axes or splitting data into subsets that are plotted on different facets—or adding layers of visualization to the plot.

To use it you, of course, need to import the library:

```
library(ggplot2)
```

You can get a list of functions it defines using this:

```
library(help = "ggplot2")
```

I can only give a very brief tutorial-like introduction to the package here. There are full books written about ggplot2 if you want to learn more. After reading this chapter, you should be able to construct basic plots, and you should be able to find information about how to make more intricate plots by searching online.

We ease into ggplot2 by first introducing the qplot() function (it stands for quick plot). This function works similarly to plot()—although it handles things a little differently—but creates the same kind of objects that the other ggplot2 functions operate on and so it can be combined with those.

Using qplot()

The qplot() function can be used to plot simple scatterplots the same way as the plot() function. To plot the cars data (see Figure 4-9), you can use:

```
cars %>% qplot(speed, dist, data = .)
```
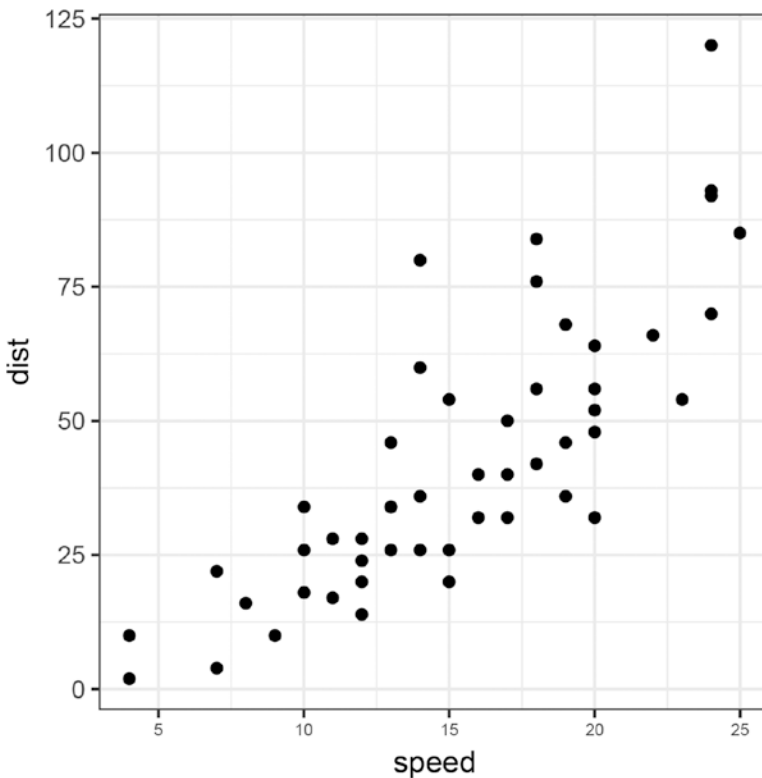

Figure 4-9. *Plot of the cars data using qplot (ggplot2)*

What happens is slightly different, though. The qplot() function actually creates a ggplot object rather than directly plotting. When such objects are printed, the effect of printing is that they are plotted. That sounds a bit confusing, but it is what happens. The function used for printing R objects is a generic function, so the effect of printing an object depends on what the object implements for the print() function. For ggplot objects, this function plots the object. It works well with the kind of code we write, though, because in the previous code, the result of the entire expression is the return value of qplot() and when this is evaluated at the outermost level in the R prompt, the result is printed. So the ggplot object is plotted.

The previous code is equivalent to this:

```
p <- cars %>% qplot(speed, dist, data = .)
p
```

Which is also equivalent to this:

```
p <- cars %>% qplot(speed, dist, data = .)
print(p)
```

The reason that it is the print() function rather than the plot() function—which would otherwise be more natural—is that the print() function is automatically called when we evaluate an expression at the R prompt. By using print(), we don't need to print objects explicitly, we just need the plotting code to be at the outermost level of the program. If you create a plot inside a function, however, it isn't automatically printed, and you do need to do this explicitly.

I mention all these details about objects being created and printed because the typical pattern for using ggplot2 is to build such a ggplot object, do various operations on it to modify it, and then finally plot it by printing it.

When you're using qplot(), some transformations of the plotting object are done before qplot() returns the object. The *quick* in *quick plot* consists of qplot() guessing at what kind of plot you are likely to want and then doing transformations on a plot to get there. To get the full control of the final plot, we skip qplot() and do all the transformations explicitly. I personally never use qplot() anymore myself, but to get started and get familiar with ggplot2 it is not a bad function to use.

With qplot(), you can make the visualization of data points depend on data variables in a simpler way than you can with plot(). To color the iris data according to Species in plot(), we needed to code up a mapping and then transform the Species column to get the colors. With qplot(), you just specify that you want the colors to depend on the Species variable, as shown in Figure 4-10.

```
iris %>% qplot(Petal.Width, Petal.Length ,
               color = Species, data = .)
```
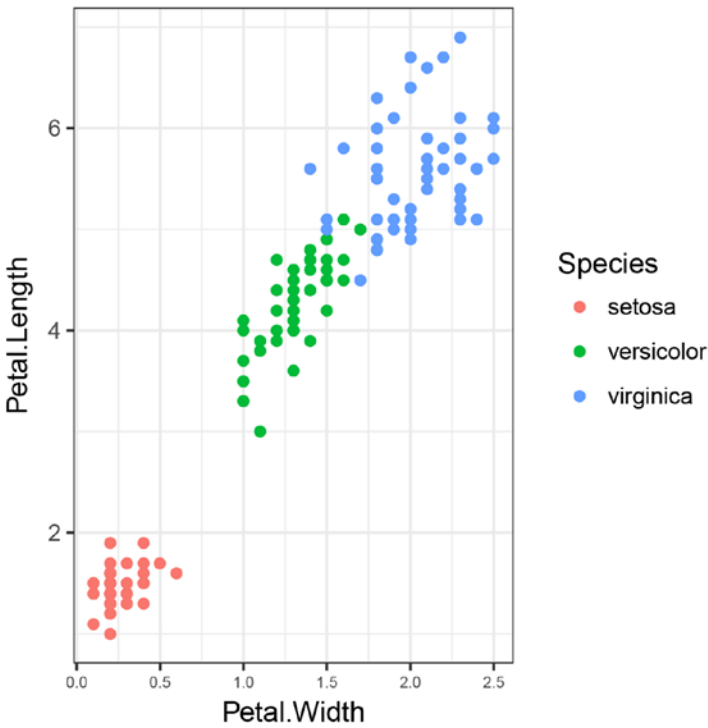
Figure 4-10. *Plot of iris data with colors determined by the species. Plotted with qplot (ggplot2).*

We get the legend for free when we are mapping the color like this, but we can modify it by doing operations on the ggplot object that qplot() returns, should we want to.

You can also use qplot() for other types of plots other than scatterplots. If you give it a single variable to plot, it will assume that you want a histogram instead of a scatterplot (see Figure 4-11).

```
cars %>% qplot(speed, data = ., bins = 10)
```
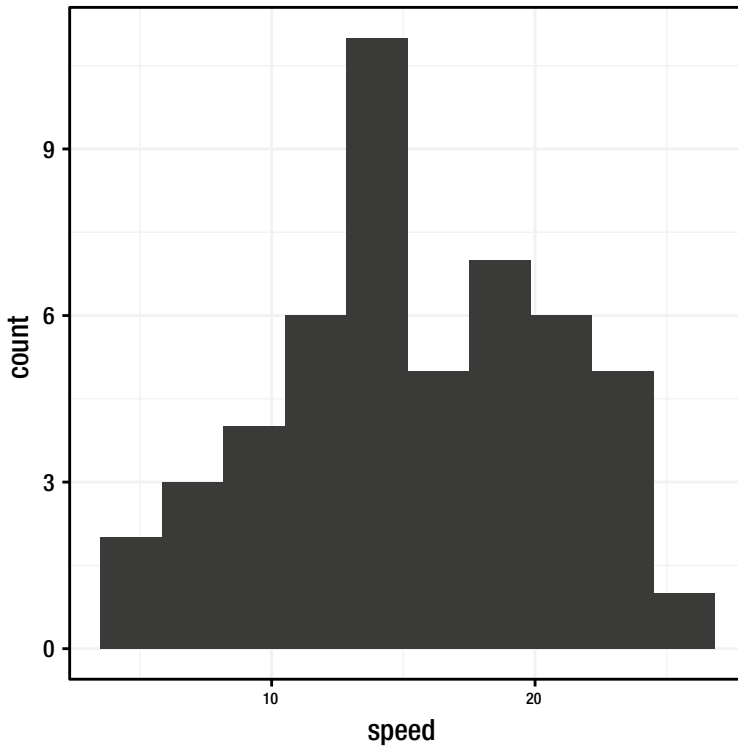
Figure 4-11. *Histogram of car speed created using qplot (ggplot2)*

If you want a density plot instead, you simply ask for it (see Figure 4-12):

```
cars %>% qplot(speed, data = ., geom = "density")
```
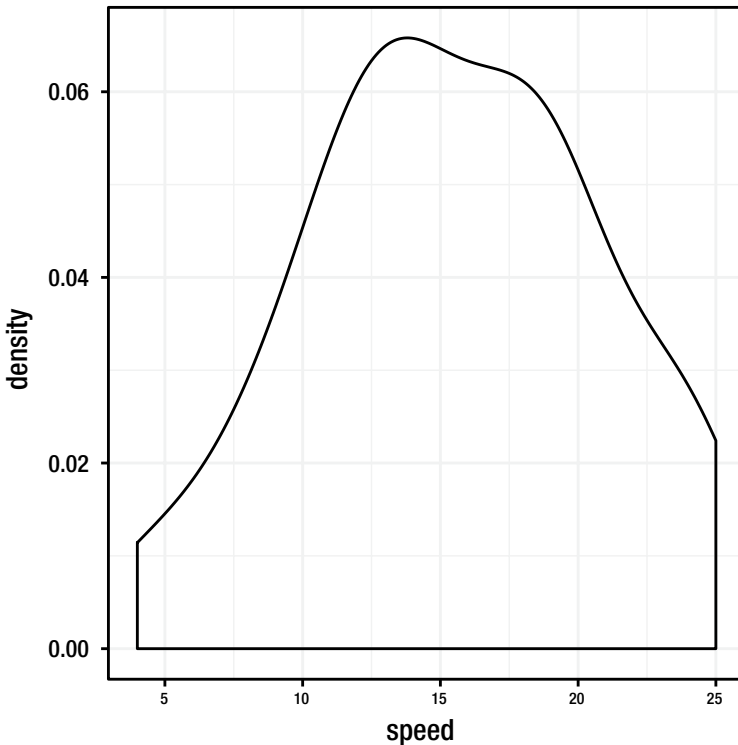
Figure 4-12. *Density of car speed created using qplot (ggplot2)*

Similarly, you can get lines, boxplots, violin plots, etc. by specifying a geometry. Geometries determine how the underlying data should be visualized. They might involve calculating some summary statistics, which they do when we create a histogram or a density plot, or they might just visualize the raw data, as we do with scatterplots, but they all describe how data should be visualized. Building a plot with ggplot2 involves adding geometries to your data, typically more than one geometry. To see how this is done, though, we leave qplot() and look at how we can create the plots we made previously with qplot() using geometries instead.

Using Geometries

By stringing together several geometry commands, you can display the same data in different ways—e.g., scatterplots combined with smoothed lines—or put several data sources on the same plot. Before you look at more complex constructions, though, let's look at how the qplot() plots could be made by explicitly calling geometry functions.

You start with the scatterplot for cars where we used the following:

```
cars %>% qplot(speed, dist, data = .)
```

To create this plot using explicit geometries we want a ggplot object, we need to map the speed parameter from the data frame to the x-axis and the dist parameter to the y-axis, and we need to plot the data as points.

```
ggplot(cars) + geom_point(aes(x = speed, y = dist))
```

We create an object using the ggplot() function. We give it the cars data as input. When we give this object the data frame, following operations can access the data. It is possible to override which data frame the data we plot comes from, but unless otherwise specified, we have access to the data we gave ggplot() when we created the initial object. Next, we do two things in the same function call. We specify that we want the x- and y-values to be plotted as points by calling geom_point() and we map speed to the x-values and dist to the y-values using the "aesthetics" function aes(). Aesthetics are responsible for mapping from data to graphics. With the geom_point() geometry, the plot needs to have x- and y-values. The aesthetics tell the function which variables in the data should be used for these.

The aes() function defines the mapping from data to graphics just for the geom_point() function. Sometimes we want to have different mappings for different geometries and sometimes we do not. If we want to share aesthetics between functions, we can set it in the ggplot() function call instead. Then, like the data, the following functions can access it, and we don't have to specify it for each subsequent function call.

```
ggplot(cars, aes(x = speed, y = dist)) + geom_point()
```

The ggplot() and geom_point() functions are combined using +. You use + to string together a series of commands to modify a ggplot object in a way very similar to how we use %>% to string together a sequence of data manipulations. The only reason that these are two different operators here are historical; if the %>% operator had been in common use when ggplot2 was developed, it would most likely have used that. As it is, you use +. Because + works slightly different in ggplot2 than %>% does in magrittr, you cannot just use a function name when the function doesn't take any arguments, so you need to include the parentheses in geom_point().

Since ggplot() takes a data frame as its first argument, it is a typical pattern to first modify data in a string of %>% operations and then give it to ggplot() and follow that with a series of + operations. Doing that with cars would provide us with this simple pipeline—in larger applications, more steps are included in both the %>% pipeline and the + plot composition.

```
cars %>% ggplot(aes(x = speed, y = dist)) + geom_point()
```

For the iris data, we used the following qplot() call to create a scatterplot with colors determined by the Species variable.

```
iris %>% qplot(Petal.Width, Petal.Length ,
               color = Species, data = .)
```

The corresponding code using ggplot() and geom_point() looks like this:

```
iris %>% ggplot +
  geom_point(aes(x = Petal.Width, y = Petal.Length,
                 color = Species))
```

Here we could also have put the aesthetics in the `ggplot()` call instead of in the `geom_point()` call.

When you specify the color as an aesthetic, you let it depend on another variable in the data. If you instead want to hard-wire a color—or any graphics parameter in general—you simply have to move the parameter assignment outside the `aes()` call. If `geom_point()` gets assigned a color parameter, it will use that color for the points; if it doesn't, it will get the color from the aesthetics. See Figure 4-13.

```
iris %>% ggplot +
  geom_point(aes(x = Petal.Width, y = Petal.Length),
                 color = "red")
```
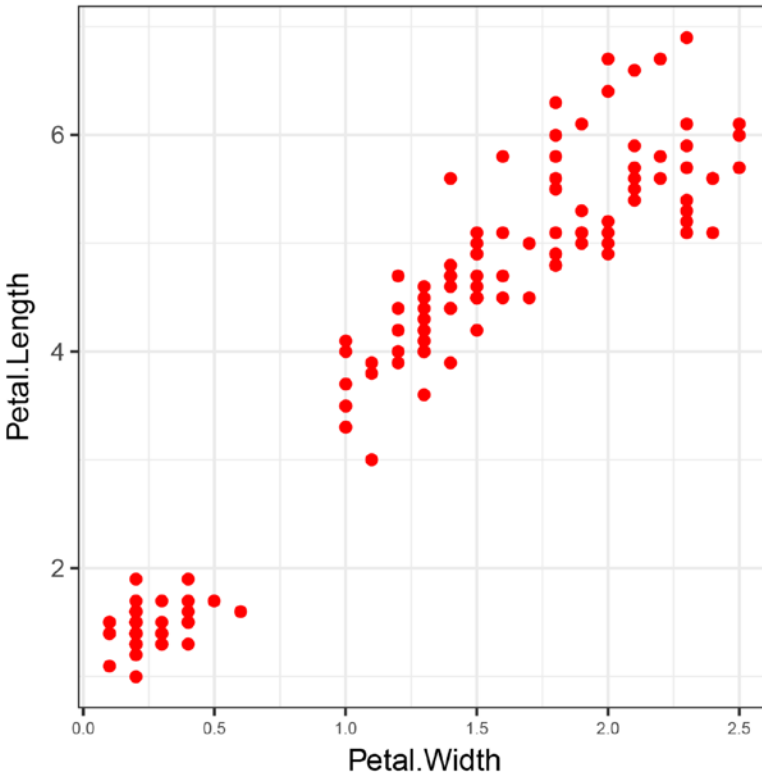

Figure 4-13. *Iris data where the color of the points is hardwired*

The `qplot()` code for plotting a histogram and a density plot:

```
cars %>% qplot(speed, data = ., bins = 10)
cars %>% qplot(speed, data = ., geom = "density")
```

Can be constructed using `geom_histogram()` and `geom_density()`, respectively:

```
cars %>% ggplot + geom_histogram(aes(x = speed), bins = 10)
cars %>% ggplot + geom_density(aes(x = speed))
```

You can combine more geometries to display the data in more than one way. This isn't always meaningful and depends on how data is summarized—combining scatterplots and histograms might not be so useful. However, you can, for example, make a plot showing the car speed both as a histogram and a density (see Figure 4-14).

```
cars %>% ggplot(aes(x = speed, y = ..count..)) +
  geom_histogram(bins = 10) +
  geom_density()
```
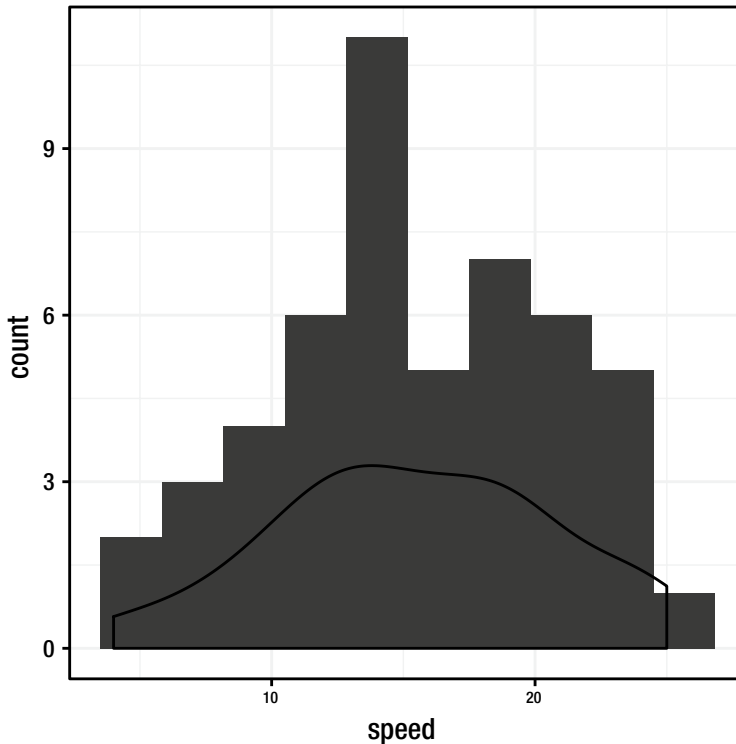

Figure 4-14. *Combined histogram and density plot for speed from the cars data*

It just requires you to call both geom_historam() and geom_density(). You do also need to add an extra aesthetics option for the y-value. This is because histograms by default will show the counts of how many observations fall within a bin on the y-axis, while densities integrate to one. By setting y = ..count.., you tell both geometries to use counts as the y-axis. To get densities instead, you can use y = ..density...

You can also use combinations of geometries to show summary statistics of data together with a scatterplot. We added the result of a linear fit of the data to the scatterplot we created for the cars data with plot(). To do the same with ggplot2, you add a geom_smooth() call, as shown in Figure 4-15.

```
cars %>% ggplot(aes(x = speed, y = dist)) +
  geom_point() + geom_smooth(method = "lm")
```
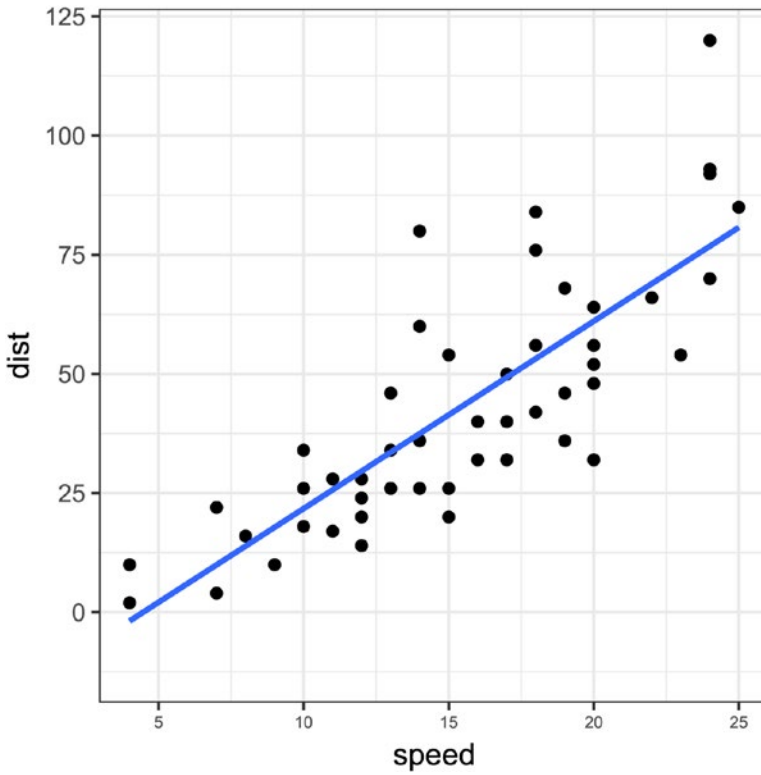
Figure 4-15. *Cars data plotted with a linear model smoothing*

You tell the geom_smooth() call to use the linear model method. If you didn't do this, it would instead plot a loess smoothing, as illustrated in Figure 4-16.

```
cars %>% ggplot(aes(x = speed, y = dist)) +
  geom_point() + geom_smooth()
## `geom_smooth()` using method = 'loess'
```
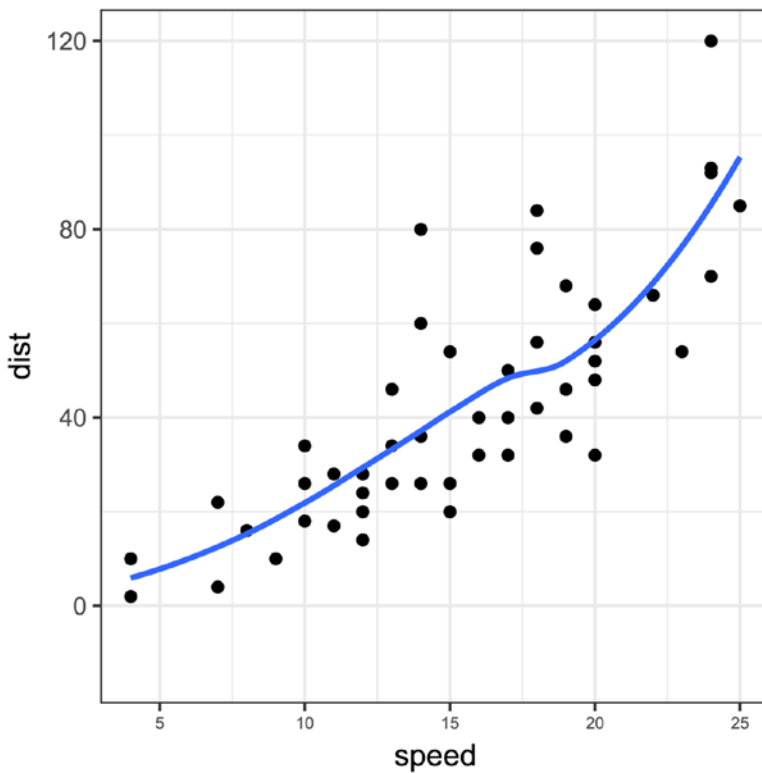
Figure 4-16. *Cars data plotted with a loess smoothing*

You can also use more than one geometry to plot more than one variable. For the `longley` data, you could use two different `geom_line()` to plot the `Unemployed` and the `Armed.Forces` data. See Figure 4-17.

```
longley %>% ggplot(aes(x = Year)) +
  geom_line(aes(y = Unemployed)) +
  geom_line(aes(y = Armed.Forces), color = "blue")
```
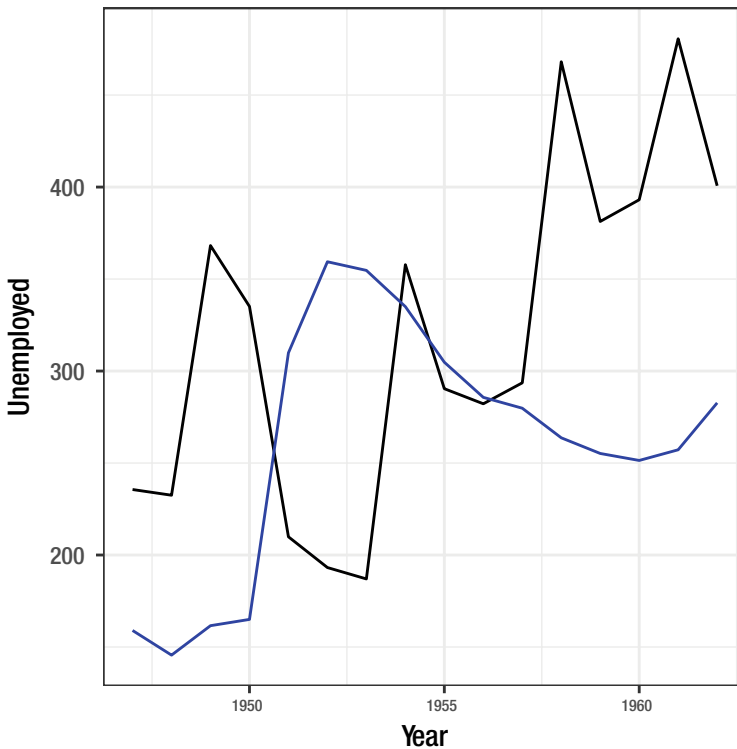
Figure 4-17. *Longley data plotted with ggplot2*

Here, we set the x-value aesthetics in the ggplot() function since it is shared by the two geom_line() geometries, but we set the y-value in the two calls, and we set the color for the Armed.Forces data, hard-wiring it instead of setting it as an aesthetic. Because we are modifying a plot rather than just drawing on a canvas with the second geom_line() call, the y-axis is adjusted to fit both lines. You, therefore, do not need to set the y-axis limit anywhere.

You can also combine geom_line() and geom_point() to get both lines and points for your data, as shown in Figure 4-18.

```
longley %>% ggplot(aes(x = Year)) +
  geom_point(aes(y = Unemployed)) +
  geom_point(aes(y = Armed.Forces), color = "blue") +
  geom_line(aes(y = Unemployed)) +
  geom_line(aes(y = Armed.Forces), color = "blue")
```
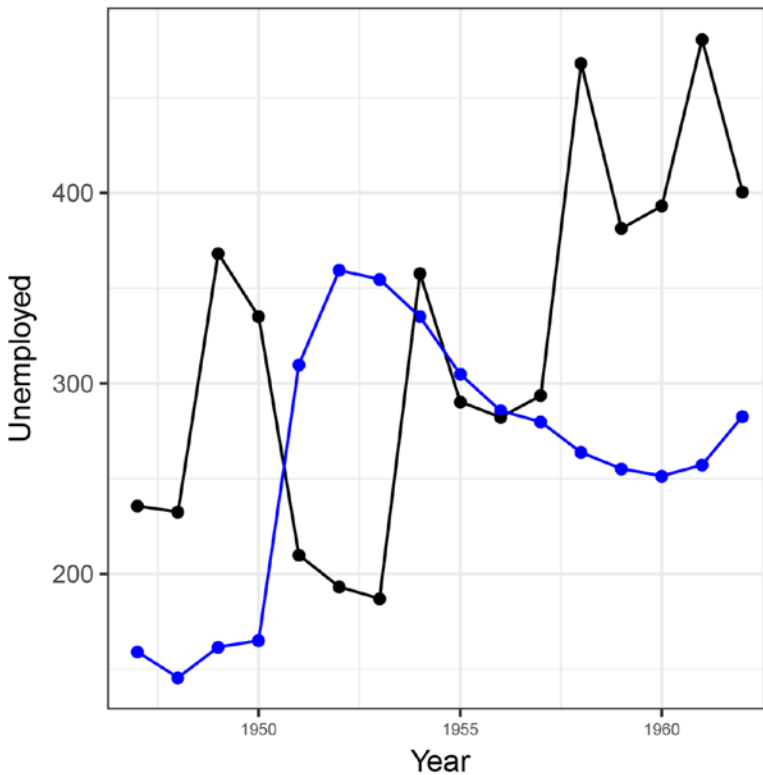
Figure 4-18. *Longley data plotted with ggplot2 using both points and lines*

Plotting two variables using different aesthetics like this is fine for most applications, but it is not always the optimal way to do it. The problem is that we are representing that the two measures, Unemployment and Armed.Forces, are two different measures we have per year and that we can plot together in the plotting code. The data is not reflecting this as something we can compute on. Should we want to split the two measures into subplots instead of plotting them in the same frame, we would need to write new plotting code. A better way is to reformat the data frame so we have one column telling us whether an observation is Unemployment or Armed.Forces and another using the values and then setting the color according to the first column and the y-axis according to the other. You can do this with the gather() function from the tidyr package, as shown in Figure 4-19.

```
longley %>% gather(key, value, Unemployed, Armed.Forces) %>%
  ggplot(aes(x = Year, y = value, color = key)) + geom_line()
```
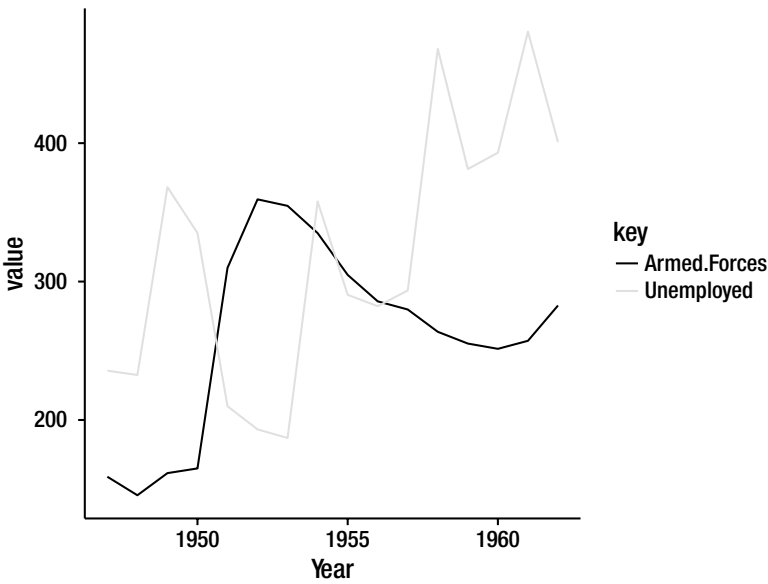
Figure 4-19. *Longley data plotted using tidy data*

Once you have transformed the data, you can change the plot with little extra code. If, for instance, you want the two values on different facets, we can simply specify this (instead of setting the colors). See Figure 4-20.

```
longley %>% gather(key, value, Unemployed, Armed.Forces) %>%
  ggplot(aes(x = Year, y = value)) + geom_line() +
  facet_grid(key ~ .)
```
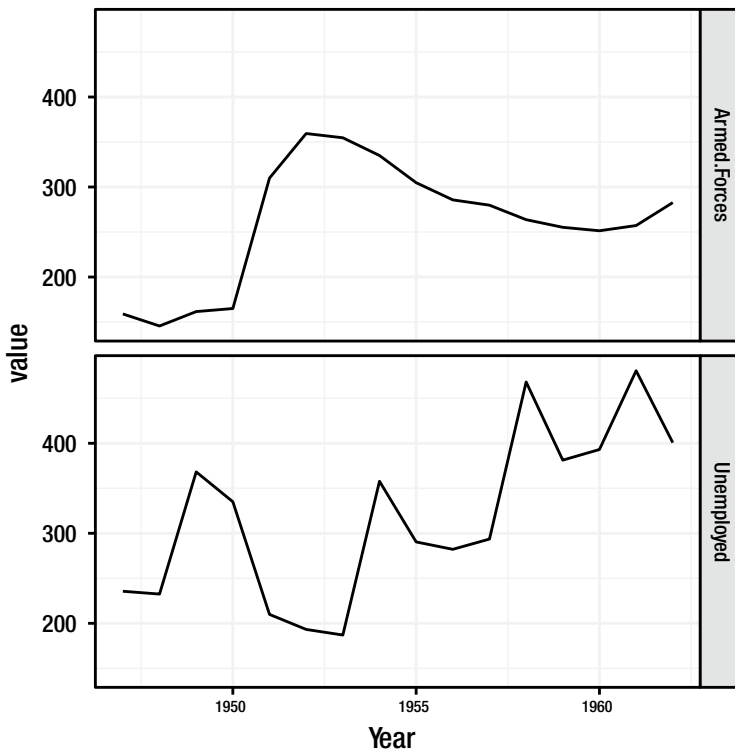
Figure 4-20. *Longley data plotted using facets*

Facets

Facets are subplots showing different subsets of the data. In the previous example, we show the `Armed.Forces` variable in one subplot and the `Unemployed` variable in another. You can specify facets using one of two functions—`facet_grid()` creates facets from a formula `rows ~ columns` and `facet_wrap()` creates facets from a formula `~ variables`. The former creates a row for the variables on the left side of the formula and a column for the variables on the right side and builds facets based on this. In the previous example, we used "`key ~ .`", so we get a row per key. Had we used "`. ~ key`" instead, we would get a column per key. `facet_wrap()` doesn't explicitly set up rows and columns, it simply makes a facet per combination of variables on the right side of the formula and wraps the facets in a grid to display them.

By default, ggplot2 will try to put values on the same axes when you create facets using `facet_grid()`. So in the previous example, the `Armed.Forces` values are shown on the same x- and y-axis as `Unemployment` even though the y-values, as we have seen, are not covering the same range. The parameter `scales` can be used to change this. Facets within a column will always have the same x-axis, however, and facets within a row will have the same y-axis.

We can see this in action with the `iris` data. We can plot the four measurements for each species in different facets, but they are on slightly different scales so we will only get a good look at the range of values for the largest range. We can fix this by setting the y-axis free. Contrast Figure 4-21 and Figure 4-22.

```
iris %>% gather(Measurement, Value, -Species) %>%
  ggplot(aes(x = Species, y = Value)) +
  geom_boxplot() +
  facet_grid(Measurement ~ .)
```
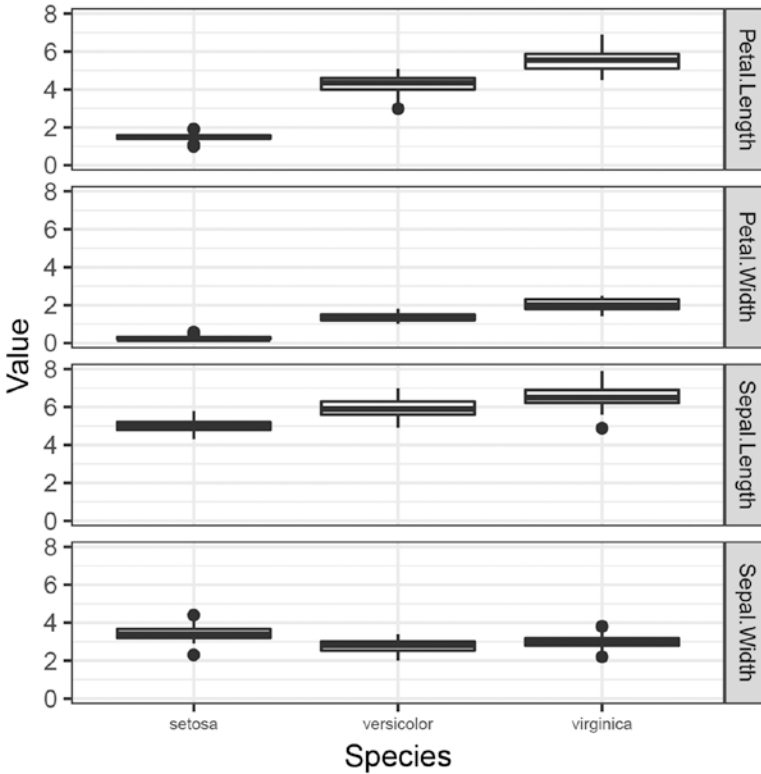

Figure 4-21. *Iris measures plotted on the same y-axis*

```
iris %>% gather(Measurement, Value, -Species) %>%
  ggplot(aes(x = Species, y = Value)) +
  geom_boxplot() +
  facet_grid(Measurement ~ ., scale = "free_y")
```
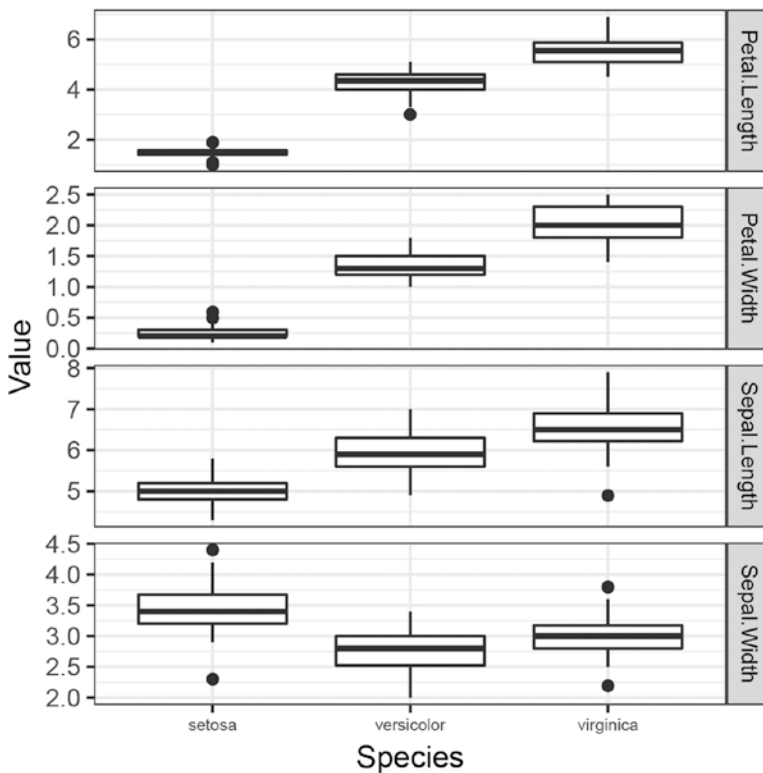
Figure 4-22. *Iris measures plotted on different y-axes*

By default, all the facets will have the same size. You can modify this using the space variable. This is mainly useful for categorical values if one facet has many more levels than another.

The labels used for facets are taken from the factors in the variables used to construct the facet. This is a good default but for print quality plots, you often want to modify the labels a little. You can do this using the labeller parameter to facet_grid(). This parameter takes a function as an argument that is responsible for constructing labels. The easiest way to construct this function is using another function, labeller(). You can give labeller() a named argument and specify a factor to make labels, with lookup tables that map the levels to the labels. For the iris data, we can use this to remove the dots in the measurement names, as shown in Figure 4-23.

```
label_map <- c(Petal.Width = "Petal Width",
               Petal.Length = "Petal Length",
               Sepal.Width = "Sepal Width",
               Sepal.Length = "Sepal Length")

iris %>% gather(Measurement, Value, -Species) %>%
  ggplot(aes(x = Species, y = Value)) +
  geom_boxplot() +
  facet_grid(Measurement ~ ., scale = "free_y",
             labeller = labeller(Measurement = label_map))
```

99
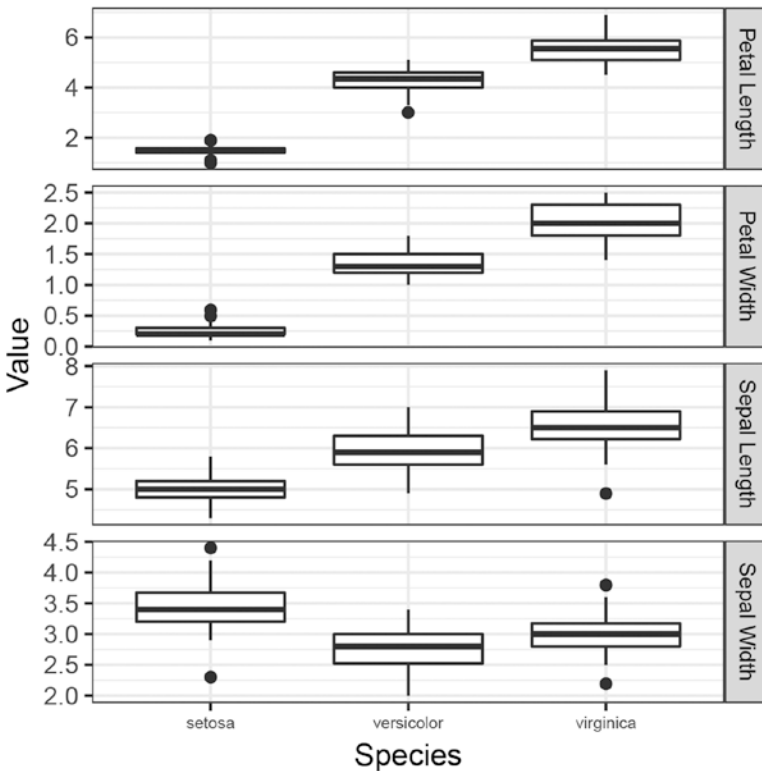
Figure 4-23. *Iris measures with measure labels adjusted*

Scaling

Geometries specify part of how data should be visualized and *scales* another. The geometries tell ggplot2 how you want your data mapped to visual components, like points or densities, and scales tell ggplot2 how dimensions should be visualized. The simplest scales to think about are the x- and y-axes, where values are mapped to positions on the plot as you are familiar with, but scales also apply to visual properties such as colors.

The simplest way to use scales is to put labels on the axes. You can also do this using the xlab() and ylab() functions, and if setting labels were all you were interested in, you could do this. However, in this example, you see a different use of scales. To set the labels in the cars scatterplot, you write:

```
cars %>% ggplot(aes(x = speed, y = dist)) +
  geom_point() + geom_smooth(method = "lm") +
  scale_x_continuous("Speed") +
  scale_y_continuous("Stopping Distance")
```

Both the x- and y-axis are showing a continuous value, so we scale like that and give the scale a name as the parameter. This will then be the name placed on the axes labels. In general, you can use the scale_x/y_continuous() functions to control the axis graphics. For instance, to set the breakpoints shown, if you wanted to plot the longley data with a tickmark for every year instead of every five years, you could set the breakpoints to every year:

```
longley %>% gather(key, value, Unemployed, Armed.Forces) %>%
  ggplot(aes(x = Year, y = value)) + geom_line() +
  scale_x_continuous(breaks = 1947:1962) +
  facet_grid(key ~ .)
```

You can also use the scale to modify the labels shown at tickmarks or set limits on the values displayed.

Scales are also the way to transform data shown on an axis. If you want to log-transform the x- or y-axis, you can use the scale_x/y_log10() functions, for instance. This usually leads to a nicer plot since the plotting code then knows that you want to show data on a log scale rather than showing transformed data on a linear scale.

To reverse an axis, use scale_x/y_reverse(). This is better than reversing the data mapped in the aesthetic since all the plotting code will just be updated to the reversed axis; you don't need to update x- or y-values in all the function geometry calls. For instance, to show the speed in the cars data in decreasing instead of increasing order, you could write:

```
cars %>% ggplot(aes(x = speed, y = dist)) +
  geom_point() + geom_smooth(method = "lm") +
  scale_x_reverse("Speed") +
  scale_y_continuous("Stopping Distance")
```

Neither axis has to be continuous, though. If you map a factor to x or y in the aesthetics, you get a discrete axis. As an example, Figure 4-24 shows the iris data plotted with the factor Species on the x-axis.

```
iris %>% ggplot(aes(x = Species, y = Petal.Length)) +
  geom_boxplot() + geom_jitter(width = 0.1, height = 0.1)
```
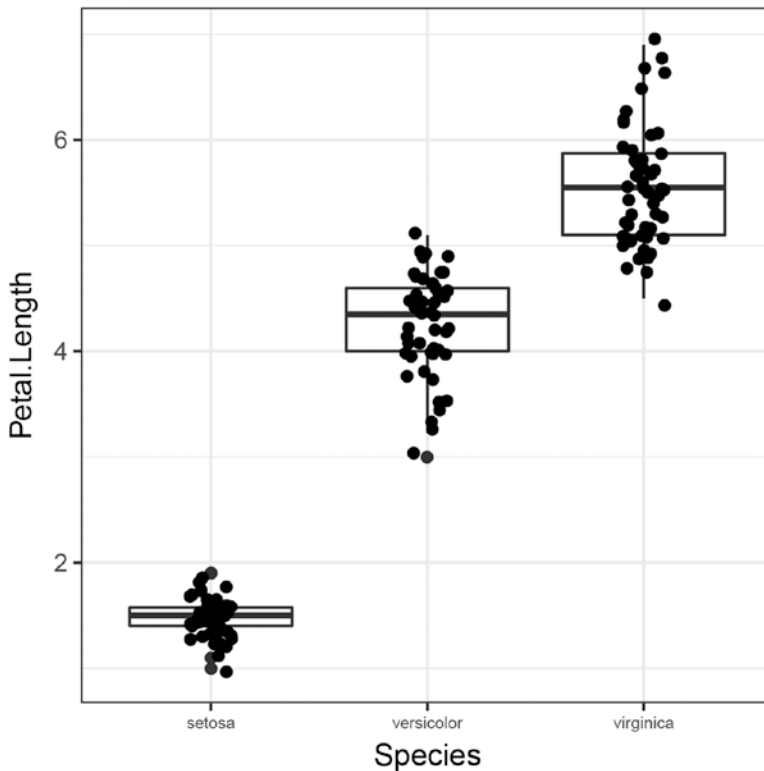
Figure 4-24. *Iris data plotted with a factor on the x-axis*

Since Species is a factor, the x-axis will be discrete, and you can show the data as a boxplot and the individual data points using the jitter geometry. If you want to modify the x-axis, you need to use scale_x_discrete() instead of scale_x_continuous().

You could, for instance, use this to modify the labels on the axis to put the species in capital letters:

```
iris %>% ggplot(aes(x = Species, y = Petal.Length)) +
  geom_boxplot() + geom_jitter(width = 0.1, height = 0.1) +
  scale_x_discrete(labels = c("setosa" = "Setosa",
                              "versicolor" = "Versicolor",
                              "virginica" = "Virginica"))
```

You just provide a map from the data levels to labels. There is more than one way to set the labels, but this is by far the easiest.

Scales are also used to control colors. You use the various scale_color_ functions to control the color of lines and points, and you use the scale_fill_ functions to control the color of filled areas.

You can plot the `iris` measurements per species and give them a different color for each species. Since it is the boxes you want to color, you need to use the `fill` aesthetics. Otherwise, you would just color the lines around the boxes. See Figure 4-25.

```
iris %>% gather(Measurement, Value, -Species) %>%
  ggplot(aes(x = Species, y = Value, fill = Species)) +
  geom_boxplot() +
  facet_grid(Measurement ~ ., scale = "free_y",
             labeller = labeller(Measurement = label_map))
```
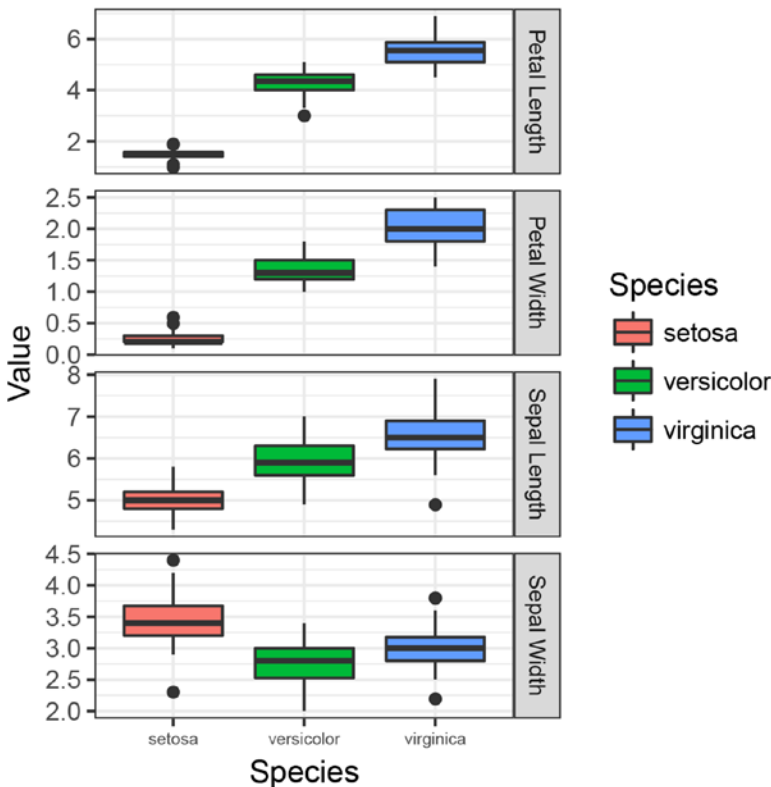

Figure 4-25. *Iris data plotted with default fill colors*

There are different ways to modify color scales. There are two classes, as there are for axes—`discrete` and `continuous`. The `Species` variable in `iris` is discrete, so to modify the fill color, you need one of the functions for that. The simplest is just to give a color per species explicitly. You can do that with the `scale_fill_manual()` function, as shown in Figure 4-26.

```
iris %>% gather(Measurement, Value, -Species) %>%
  ggplot(aes(x = Species, y = Value, fill = Species)) +
  geom_boxplot() +
  scale_fill_manual(values = c("red", "green", "blue")) +
  facet_grid(Measurement ~ ., scale = "free_y",
             labeller = labeller(Measurement = label_map))
```

103
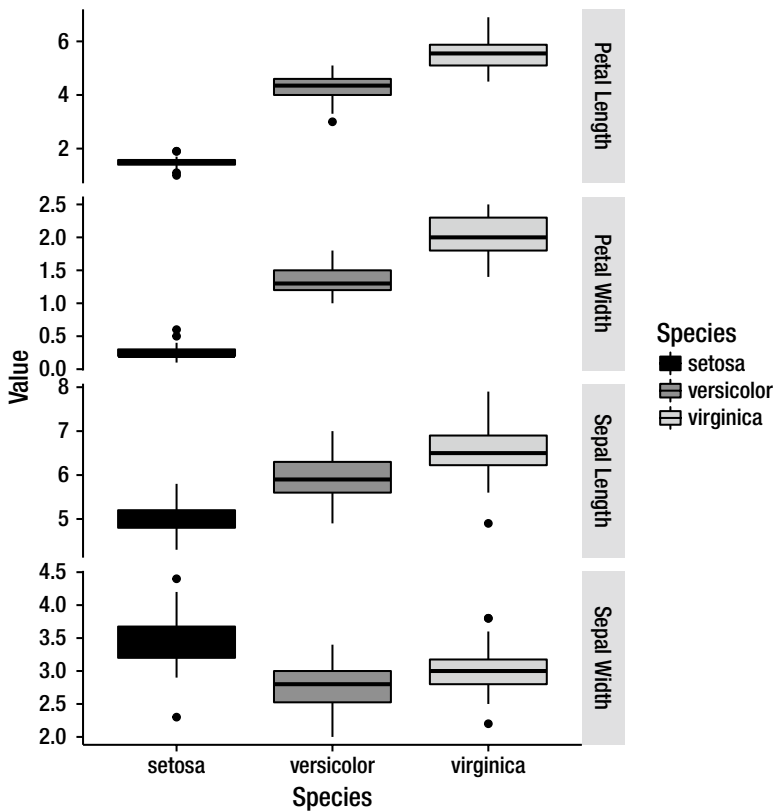
Figure 4-26. *Iris data plotted with custom fill colors*

Explicitly setting colors is risky business, though, unless you have a good feeling for how colors work together and which combinations can be problematic for color-blind people. It is better to use one of the "brewer" choices. These are methods for constructing good combinations of colors (see http://colorbrewer2.org) and you can use them with the scale_fill_brewer() function, as shown in Figure 4-27.

```
iris %>% gather(Measurement, Value, -Species) %>%
  ggplot(aes(x = Species, y = Value, fill = Species)) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Greens") +
  facet_grid(Measurement ~ ., scale = "free_y",
             labeller = labeller(Measurement = label_map))
```
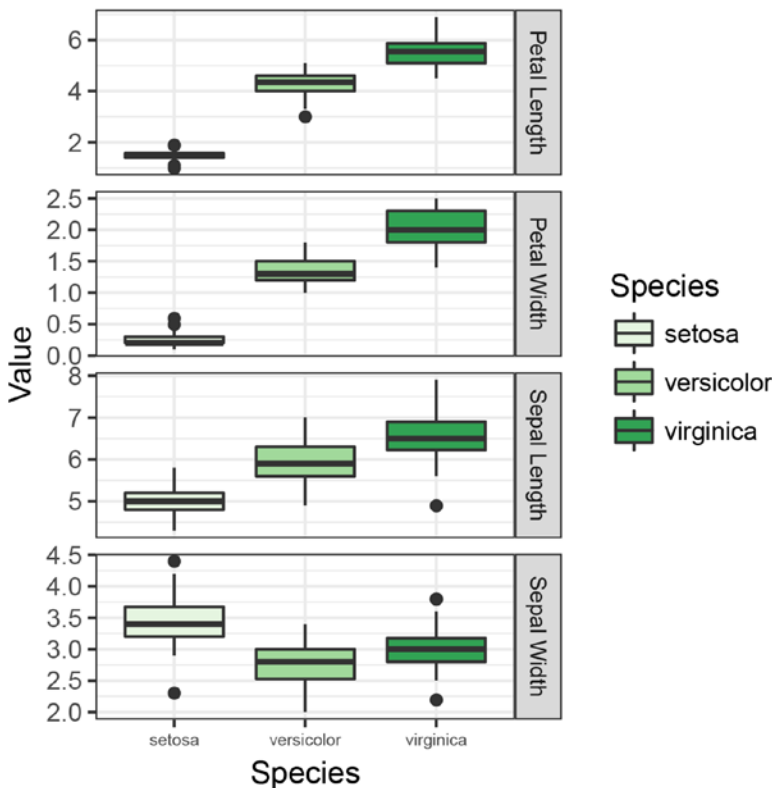
Figure 4-27. *Iris data plotted with a brewer fill color*

Themes and Other Graphics Transformations

Most of using `ggplot2` consist of specifying geometries and scales to control how data is mapped to visual components, but you also have control over how the final plot will look through functions that only concern the final visual result.

Most of this is done by modifying the so-called theme. If you have tried the examples I gave in this chapter yourself, the results might look different from the figures in this book. This is because I have set up a default theme for the book using this command:

```
theme_set(theme_bw())
```

The `theme_bw()` sets up the look of the figures you see here. You can add a theme to a plot using + as you would any other `ggplot2` modification or set it as default as I have done here. There are several themes you can use; look for functions that start with `theme_`. They all can be modified for more control over a plot.

Besides themes, various other functions also affect the way a plot looks. There is far too much to cover here on all the things you can do with themes and graphics transformations, but I can show you an example that will give you an idea of what you can achieve.

You can for instance change coordinate systems using various `coord_` functions—the simplest is just flipping x and y with `coord_flip()`. This can of course also be achieved just by changing the aesthetics, but flipping the coordinates of a complex plot can be easier than updating aesthetics several places. For the `iris` plot we looked at before, I might want to change the axes.

I also want to put the measurement labels on the left instead of on the right. You can control the placement of facet labels using the switch option to facet_grid(). Giving the switch parameter the value y will switch the location of that label.

```
iris %>% gather(Measurement, Value, -Species) %>%
  ggplot(aes(x = Species, y = Value, fill = Species)) +
  geom_boxplot() +
  scale_x_discrete(labels = c("setosa" = "Setosa",
                              "versicolor" = "Versicolor",
                              "virginica" = "Virginica")) +
  scale_fill_brewer(palette = "Greens") +
  facet_grid(Measurement ~ ., switch = "y",
             labeller = labeller(Measurement = label_map)) +
  coord_flip()
```
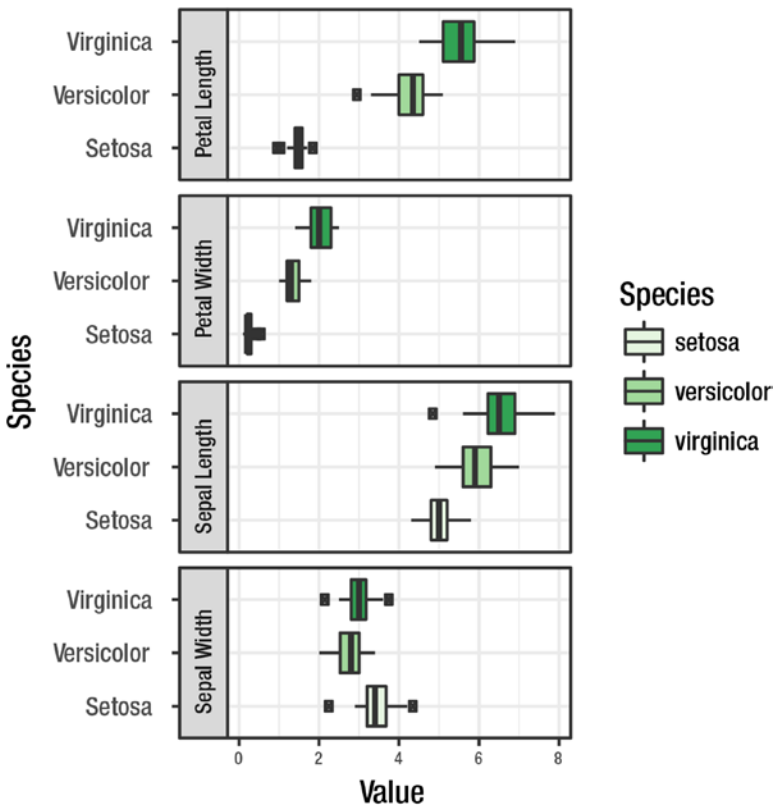

Figure 4-28. *Iris with flipped coordinates and switched facet labels*

If I just flip the coordinates the axis labels on the new x-axis will be wrong if I tell the `facet_grid()` function to have a free y-axis. With a free y-axis, it would have different ranges for the y-values, which is what we want, but after flipping the coordinates, we will see the values for only one of the y-axes. The other values will be plotted as if they were on the same axis, but they won't be. So I have removed the `scale` parameter of `facet_grid()`. Try to put it back and see what happens.

The result so far is shown in Figure 4-28. We flipped coordinates and moved labels, but the labels look ugly with the color background. You can remove it by modifying the theme using `theme(strip.background = element_blank())`. It just sets the `strip.background`, which is the graphical property of facet labels, to a blank element, so in effect it removes the background color. We can also move the legend label using a theme modification such as `theme(legend.position="top")`.

```
iris %>% gather(Measurement, Value, -Species) %>%
  ggplot(aes(x = Species, y = Value, fill = Species)) +
  geom_boxplot() +
  scale_x_discrete(labels = c("setosa" = "Setosa",
                              "versicolor" = "Versicolor",
                              "virginica" = "Virginica")) +
  scale_fill_brewer(palette = "Greens") +
  facet_grid(Measurement ~ ., switch = "y",
             labeller = labeller(Measurement = label_map)) +
  coord_flip() +
  theme(strip.background = element_blank()) +
  theme(legend.position="top")
```
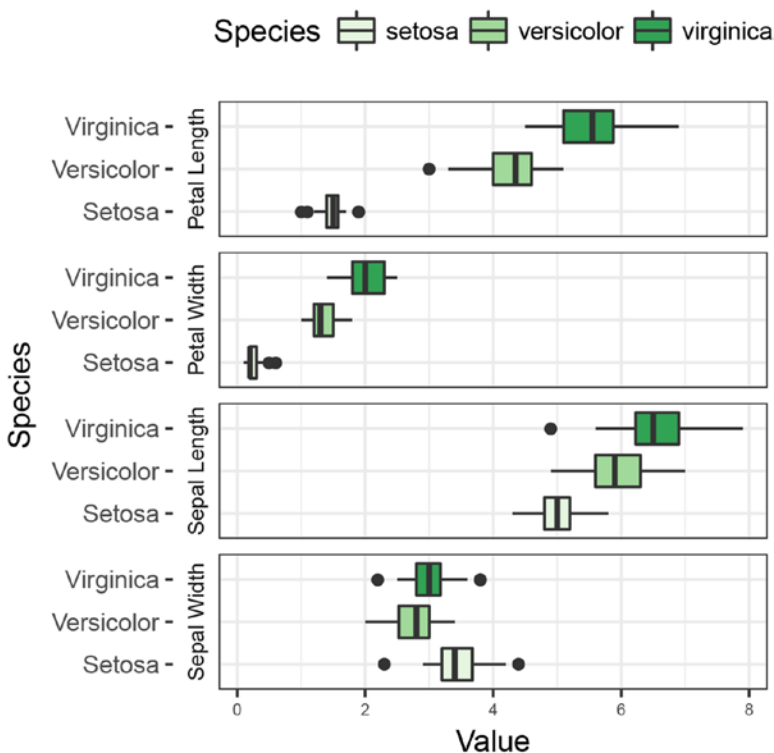

Figure 4-29. *Iris data with theme modifications*

The result is now shown in Figure 4-29. It is pretty close to something we could print. We want the labeled species to be in capital letters just like the axis labels.

Well, you know how to do that using the labels parameter to a scale, so the final plotting code could look like this:

```
label_map <- c(Petal.Width = "Petal Width",
               Petal.Length = "Petal Length",
               Sepal.Width = "Sepal Width",
               Sepal.Length = "Sepal Length")
species_map <- c(setosa = "Setosa",
                 versicolor = "Versicolor",
                 virginica = "Virginica")

iris %>% gather(Measurement, Value, -Species) %>%
  ggplot(aes(x = Species, y = Value, fill = Species)) +
  geom_boxplot() +
  scale_x_discrete(labels = species_map) +
  scale_fill_brewer(palette = "Greens", labels = species_map) +
  facet_grid(Measurement ~ ., switch = "y",
             labeller = labeller(Measurement = label_map)) +
  coord_flip() +
  theme(strip.background = element_blank()) +
  theme(legend.position="top")
```

The final result is shown in Figure 4-30.
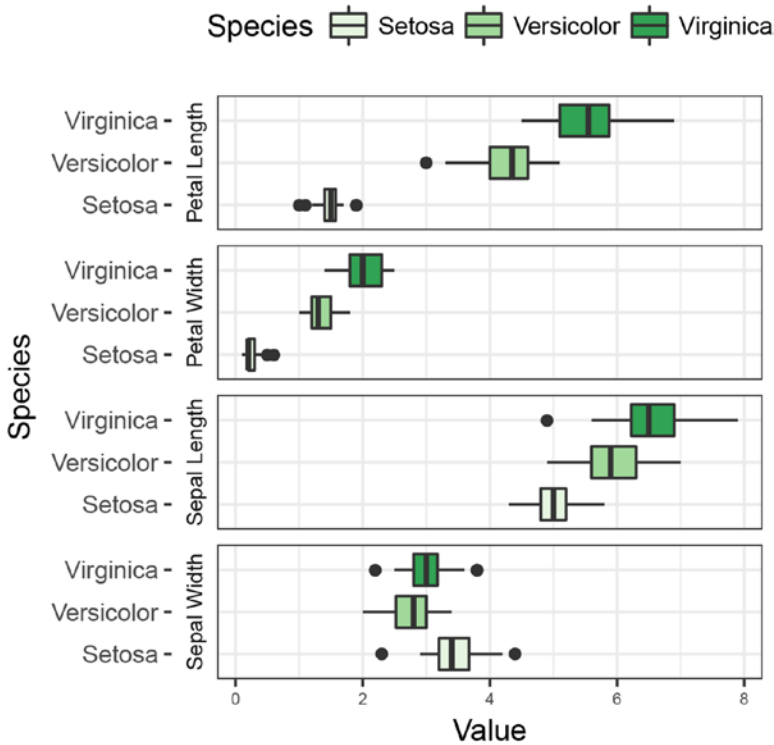
Figure 4-30. *Final version of iris plot*

Figures with Multiple Plots

Facets cover many situations where you want to have multiple panels in the same plot, but not all. You use facets when you want to display different subsets of the data in separate panels, but essentially have the same plot for the subsets. Sometimes you want to combine different types of plots, or plots of different datasets, as subplots in different panels. For that, you need to combine otherwise independent plots.

The ggplot2 package doesn't directly support combining multiple plots, but it can be achieved using the underlying graphics system, grid. Working with basic grid you have many low-level tools for modifying graphics, but for just combining plots you want more high-level functions, and you can get them from the gridExtra package.

To combine plots, you first create them as you normally would. So, for example, you could make two plots of the iris data like this:

```
petal <- iris %>% ggplot() +
  geom_point(aes(x = Petal.Width, y = Petal.Length,
                 color = Species)) +
  theme(legend.position="none")

sepal <- iris %>% ggplot() +
  geom_point(aes(x = Sepal.Width, y = Sepal.Length,
                 color = Species)) +
  theme(legend.position="none")
```

109

You then import the gridExtra package:

```
library(gridExtra)
```

You then use the grid.arrange() function to create a grid of plots, putting in the two plots you just created, as shown in Figure 4-31.

```
grid.arrange(petal, sepal, ncol = 2)
```
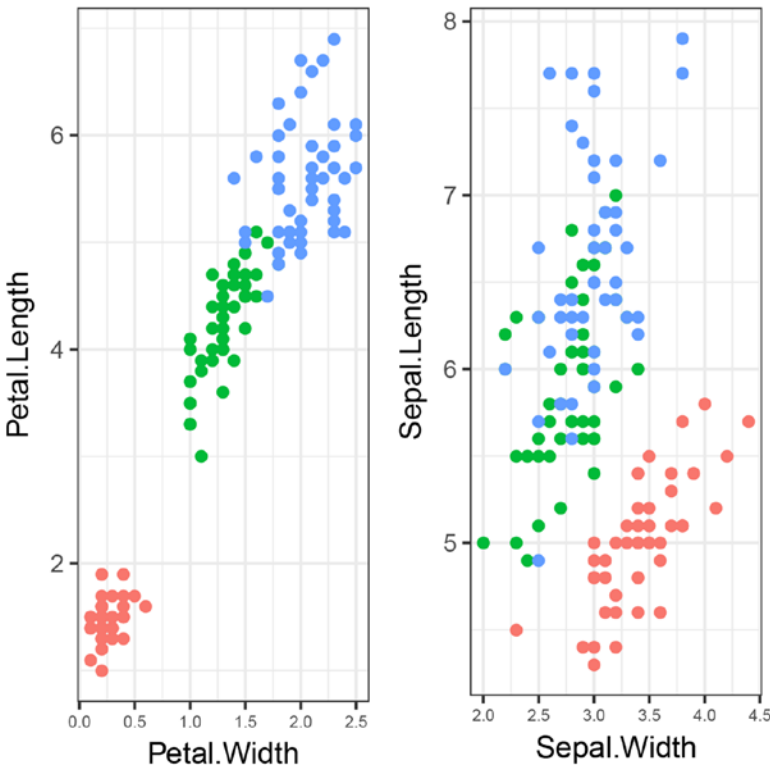

Figure 4-31. *Combining two plots of the iris data using grid.arrange*

Another approach I like to use is the plot_grid() function from the cowplot package. This package contains several functions developed by Claus O. Wilke (where the cow comes from) for his plotting needs. Loading it will redefine the default ggplot2 theme. You can use the theme_set() function to change it back if you don't like the theme that cowplot provides.

Anyway, to create a plot with subplots using cowplot, you have to import the package as follows:

```
library(cowplot)
```

If you don't want the theme it chooses, you need to change it using theme_set(). Otherwise, you can combine the plots defined before using plot_grid(). See Figure 4-32.

```
plot_grid(petal, sepal, labels = c("A", "B"))
```
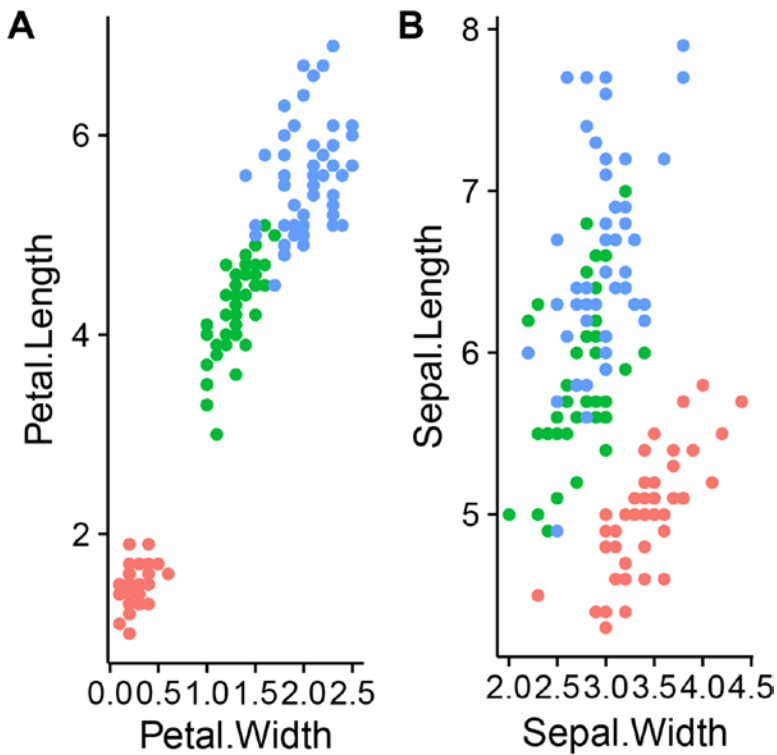
Figure 4-32. *Combining two plots of the iris data using cowplot*

Exercises

In the previous chapter, you should have imported a dataset and used `dplyr` and `tidyr` to explore it using summary statistics. Now do the same thing using plotting. If you looked at summary statistics, try representing these as boxplots or smoothed scatterplots. If you have different variables that you gathered using `tidyr`, try to plot the data similar to what you saw with `iris`.

CHAPTER 5

■ ■ ■

Working with Large Datasets

The concept of *Big Data* refers to very large datasets, sets of sizes where you need data warehouses to store the data, where you typically need sophisticated algorithms to handle the data, and distributed computations to get anywhere with it. At the very least, we talk many gigabytes of data but also are often dealing with terabytes or exabytes.

Dealing with Big Data is also part of data science, but it is beyond the scope of this book. This chapter is on large datasets and how to deal with data that slows down your analysis, but it is not about datasets so large that you cannot analyze them on your own desktop computer.

If we ignore the Big Data issue, what a large dataset is depends very much on what you want to do with the data. That comes down to the complexity of what you are trying to achieve. Some algorithms are fast and can scan through data in linear time—meaning that the time it takes to analyze the data is linear in the number of data points—while others take exponential time and cannot actually be applied to datasets with more than a few tens or hundreds of data points. The science of what you can do with data in a given amount of time, or a given amount of space (be it RAM or disk space or whatever you need), is called *complexity theory* and is one of the fundamental topics in computer science. In practical terms, though, it usually boils down to how long you are willing to wait for an analysis to be done and it is a very subjective decision.

In this chapter, we consider several cases I have found in my own work where data gets a bit too large to do what I want, and I have to deal with it in various ways. Your cases are likely to be different, but maybe you can get some inspiration, at least, from these cases.

Subsample Your Data Before You Analyze the Full Dataset

The first point I want to make, though, is this: You very rarely need to analyze a complete dataset to get at least an idea of how the data behaves. Unless you are looking for very rare events, you will get as much a feeling for the data looking at a few thousands of data points as you would from looking at a few million.

Sometimes you do need very large data to find what you are looking for. This is, for example, the case when looking for associations between genetic variation and common diseases where the association can be very weak, and you need lots of data to distinguish between chance associations and true associations. But for most signals in data that are of practical importance, you will see the signals in smaller datasets. So before you throw the full power of all your data at an analysis, especially if that analysis turns out to be very slow, you should explore a smaller sample of your data.

Here it is important that you pick a random sample. There is often structure in data beyond the columns in a data frame. This could be structure caused by when the data was collected. If the data is ordered by when the data was collected, then the first data points you have can be different from later data points. This isn't explicitly represented in the data, but the structure is there nevertheless. Randomizing your data alleviates problems that can arise from this. Randomizing might remove a subtle signal, but with the power of statistics, we can deal with random noise. It is much harder to deal with consistent biases we just don't know about.

If you have a large dataset, and your analysis is being slowed down because of it, don't be afraid to pick a random subset and analyze that. It is possible that you will see signals in the subsample that is not present in the full dataset, but it is much less likely than you might fear. When you are looking for signals in your data, you always have to worry about false signals. But it is not more liable to pop up in a smaller dataset than in a larger. And with a larger dataset to check your results against later, you are less likely to stick with wrong results at the end of your analysis.

Getting spurious results is mostly a concern with traditional hypothesis testing. If you set a threshold for when a signal is significant at 5% for p-values, you will see spurious results one time out of twenty. If you don't correct for multiple testing, you will be almost guaranteed to see false results. These are unlikely to survive when you later throw the complete data at your models.

In any case, you are more likely to have statistically significant deviations from a null model, which is completely irrelevant to your analysis, with large datasets. We typically use very simple null models when analyzing data and any complex dataset is not generated from a simple null model. With enough data, the chances are that anything you look at will have significant deviations from your simple null model. The real world does not draw samples from a simple linear model. There is always some extra complexity. You won't see it with a few data points but with enough data, you can reject any null model. It doesn't mean that what you see has any practical importance.

If you have signals you discover in a smaller subset of your data, and these signals persist when you look at the full dataset, you can trust them that much more.

So if the data size slows you down, downsample and analyze a subset of it.

You can use the dplyr functions sample_n() and sample_frac() to sample from a data frame. Use sample_n() to get a fixed number of rows and sample_frac() to get a fraction of the data:

```
iris %>% sample_n(size = 5)
##     Sepal.Length Sepal.Width Petal.Length
## 15           5.8         4.0          1.2
## 59           6.6         2.9          4.6
## 52           6.4         3.2          4.5
## 128          6.1         3.0          4.9
## 141          6.7         3.1          5.6
##     Petal.Width    Species
## 15          0.2     setosa
## 59          1.3 versicolor
## 52          1.5 versicolor
## 128         1.8  virginica
## 141         2.4  virginica
iris %>% sample_frac(size = 0.02)
##     Sepal.Length Sepal.Width Petal.Length
## 61           5.0         2.0          3.5
## 127          6.2         2.8          4.8
## 48           4.6         3.2          1.4
##     Petal.Width    Species
## 61          1.0 versicolor
## 127         1.8  virginica
## 48          0.2     setosa
```

Of course, to sample using dplyr, you need your data in a form that dplyr can manipulate, and if the data is too large even to load into R, then you cannot have it in a data frame to sample from, to begin with. Luckily, dplyr has support for using data that is stored on disk rather than in RAM, in various backend formats, as you will see soon. It is, for example, possible to connect a database to dplyr and sample from a large dataset this way.

Running Out of Memory During Analysis

R can be very wasteful of RAM. Even if your dataset is small enough to fit in memory and small enough that the analysis time is not a substantial problem, it is easy to run out of memory because R remembers more than is immediately obvious.

In R, all objects are immutable,[1] so whenever you modify an object, you are actually creating a new object. The implementation of this is smart enough that you only have independent copies of data when it actually is different. Having two different variables to refer to the same data frame doesn't mean that the data frame is represented twice, but if you modify the data frame in one of the variables, then R will create a copy with the modifications and you now have the data twice, accessible through the two variables. If you refer to the data frame through only one variable, then R is smart enough not to make a copy, though.

You can examine memory usage and memory changes using the `pryr` package:

```
library(pryr)
```

For example, you can see what the cost is of creating a new vector:

```
mem_change(x <- rnorm(10000))
## 80.5 kB
```

Modifying this vector—which R doesn't actually allow, so what happens is that a new copy is made with the modification—doesn't significantly increase the memory usage because R is smart about only copying when more than one variable refers to an object:

```
mem_change(x[1] <- 0)
## 1.3 kB
```

If we assign the vector to another variable, we do not use twice the memory, because both variables will just refer to the same object:

```
mem_change(y <- x)
## 1.36 kB
```

But if we modify one of the vectors, we will have to make a copy so the other vector remains the same:

```
mem_change(x[1] <- 0)
## 81.4 kB
```

This is another reason, besides polluting the namespace, for using pipelines rather than assigning to many variables during an analysis. You are fine if you assign back to a variable, though, so the `%<>%` operator does not lead to a lot of copying.

Even when using pipelines, you still have to be careful, though. Many functions in R will still copy data.

If a function does any modification to data, it is copied to a local variable. There might be some sharing so for example just referring to a data frame in a local variable does not create a copy, but if you, for example, split a data frame into training and test data in a function then you will be copying and now represent all the data twice. This memory is freed after the function finishes its computations so it is really only a problem if you are very close to the limit of RAM.

---

[1]This is not entirely true; it is *possible* to make mutable objects, but it requires some work. Unless you go out of your way to create mutable objects, this statement is true.

If such copied data is saved in the output of the function, however, it is *not* freed when the function returns. It is, for example, not unusual that model fitting functions will save the entire fitting data in the returned object. The linear regression function, lm(), will store not only the input data frame but also the response variable and all explanatory variables, essentially making copies so all the data is stored twice (and it does so in a way that does not reuse memory). You have to tell it explicitly not to, using parameters model, x, y, and qr if you want to avoid this.

When you have problems with running out of memory in a data analysis in R, it is usually not that you cannot represent your data initially but that you end up having many copies. You can avoid this to some extent by not storing temporary data frames in variables and by not implicitly storing data frames in the output of functions, or you can explicitly remove stored data using the rm() function to free up memory.

Too Large to Plot

The first point where I typically run into problems with large datasets is not that I run out of RAM but when I am plotting. Especially when making scatterplots; box plots and histograms summarize the data and are usually not a problem.

There are two problems when making scatterplots with a lot of data. The first is that if you create files from scatterplots, you will create a plot that contains every single individual point. That can be a very large file. Worse, it will take forever to plot, since a viewer will have to consider every single point. You can avoid this problem by creating raster graphics instead of PDFs, but that takes us to the second issue. With too many points, a scatterplot is just not informative any longer. Points will overlap, and you cannot see how many individual data points fall on the plot. This usually becomes a problem long before the computational time becomes an issue.

If, for example, we have a data frame with 10000 points:

```
d <- data.frame(x = rnorm(10000), y = rnorm(10000))
```

We can still make a scatterplot, and if the plot is saved as raster graphic instead of PDF, the file will not be too large to watch or print:

```
d %>% ggplot(aes(x = x, y = y)) +
  geom_point()
```

The result will just not be all that informative; see Figure 5-1. The points are shown on top of each other making it hard to see if the big black cloud of points has different densities.
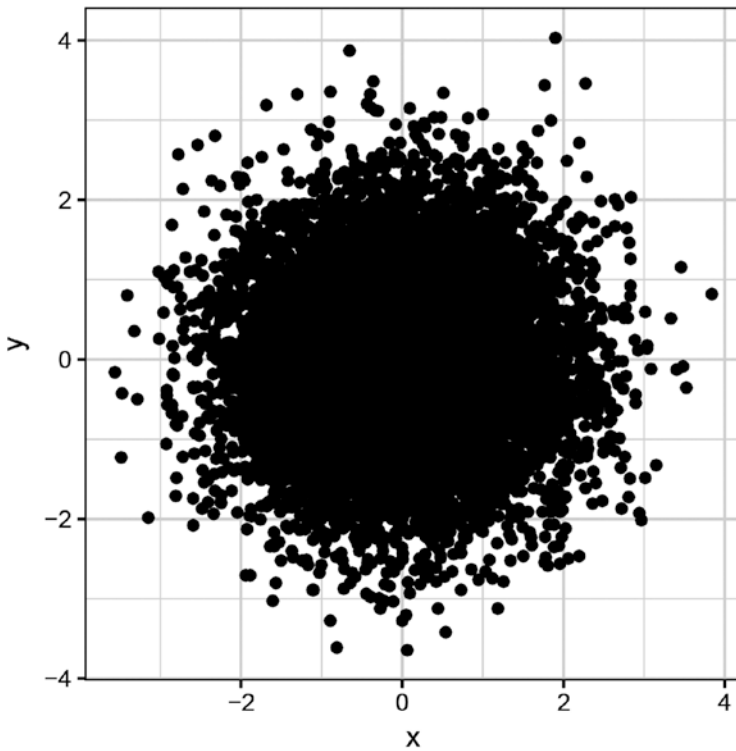
Figure 5-1. *A scatterplot with too many points*

The solution is to represent points in a way such that they still visible even when there are many overlapping points. If the points are overlapping because the actually have the same x- or y-coordinates, you can jitter them, as you saw in the previous chapter. Another solution to the same problem is plotting the points with alpha levels so each point is partly transparent. You can see the density of points because they are partly transparent, but you still end up with a plot with very many points, as shown in Figure 5-2.

```
d %>% ggplot(aes(x = x, y = y)) +
  geom_point(alpha = 0.2)
```
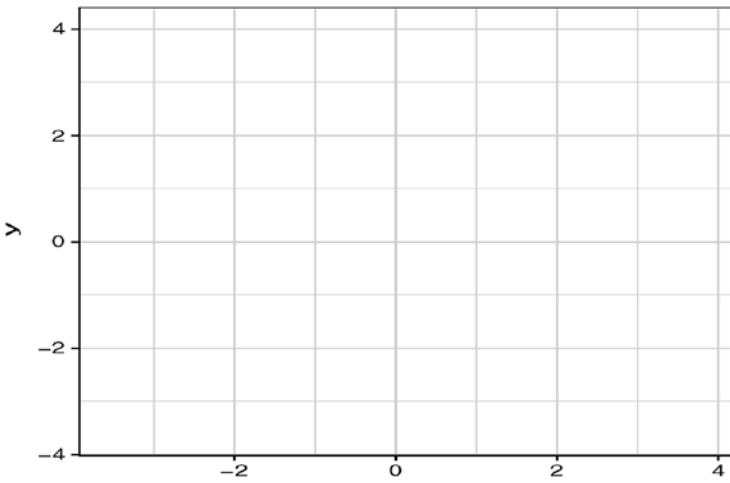
Figure 5-2. *A scatterplot with alpha values*

This doesn't solve the problem that files will draw every single point and cause printing and file-size problems. A scatterplot with transparency is just a way of showing the 2D density, though, and you can do that directly using the geom_density_2d() function, as shown in Figure 5-3.
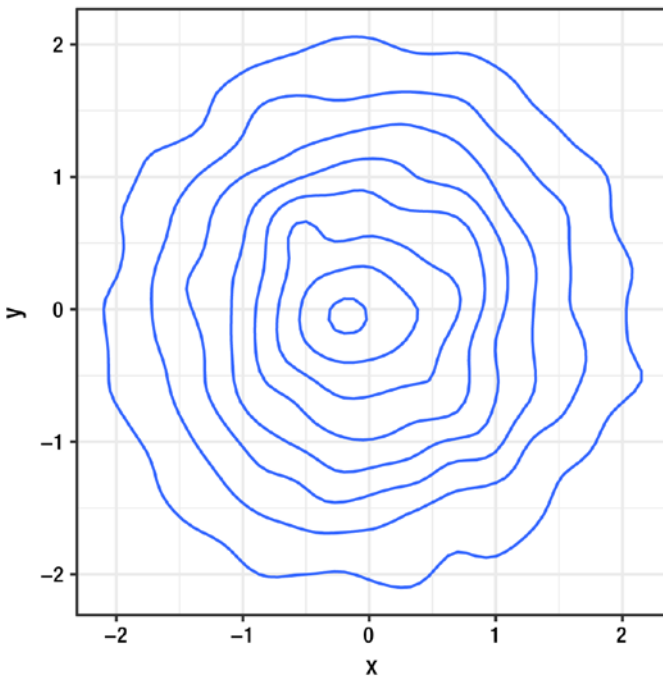
Figure 5-3. *A 2D density plot*

```
d %>% ggplot(aes(x = x, y = y)) +
  geom_density_2d()
```

Figure 5-3 shows the contour of the density.

An alternative way of showing a 2D density is using a so-called hex-plot. This is the 2D equivalent of a histogram. The 2D plane is split into hexagonal bins, and the plot shows the count of points falling into each bin.

To use it, you need to install the package hexbin and use the ggplot2 function geom_hex(), as shown in Figure 5-4.
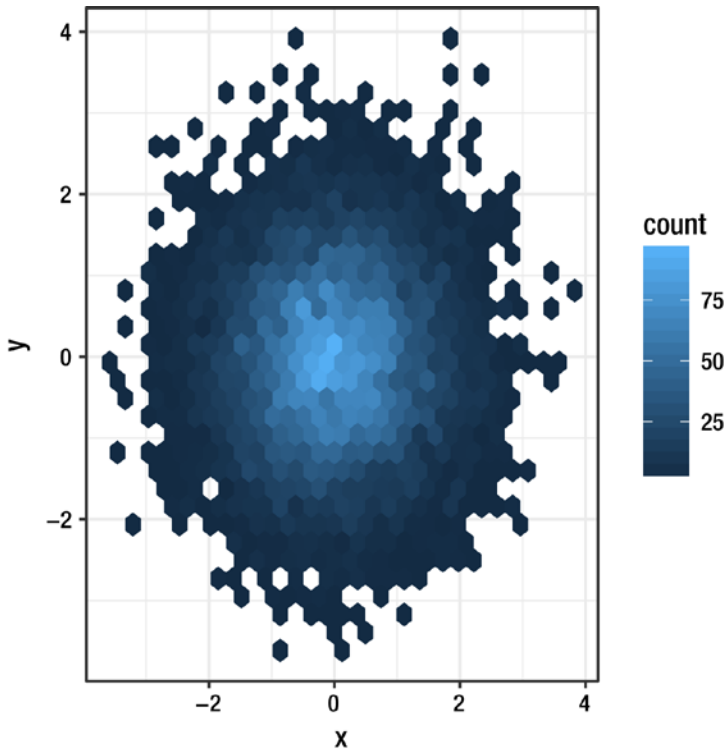
Figure 5-4. *A hex plot*

```
d %>% ggplot(aes(x = x, y = y)) +
  geom_hex()
```

The colors used by geom_hex() are the fill colors, so you can change them using the scale_fill functions. You can also combine hex and 2D density plots to get both the bins and contours displayed, as shown in Figure 5-5.

```
d %>% ggplot(aes(x = x, y = y)) +
  geom_hex() +
```
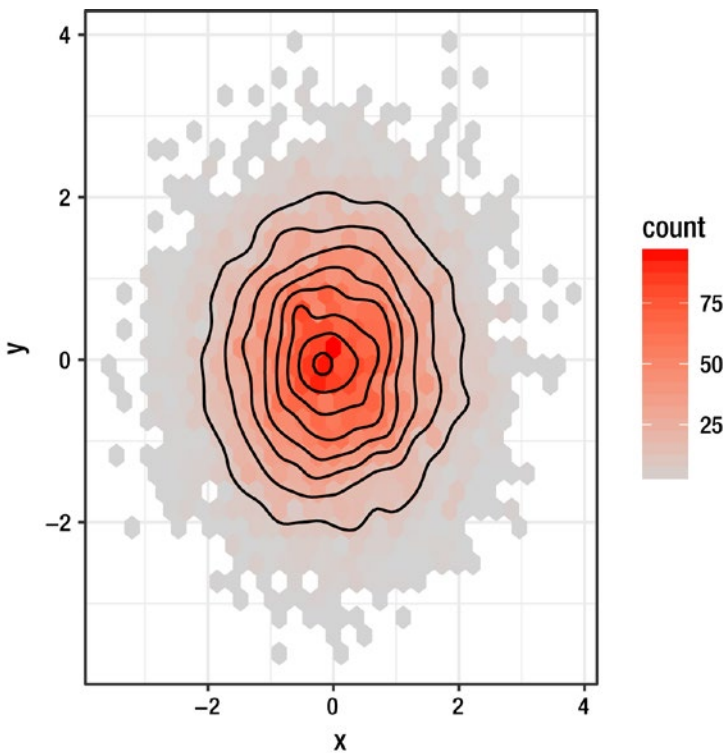
Figure 5-5. *A plot combining hex and 2D density*

```
  scale_fill_gradient(low = "lightgray", high = "red") +
  geom_density2d(color = "black")
```

Too Slow to Analyze

When plotting data, the problem is usually only in scatterplots. Otherwise, you don't have to worry about having too many points or too large plot files. Even when plotting lots of points, the real problem doesn't show up until you create a PDF plot and load it into your viewer or send it to the printer.

   With enough data points, though, most analyzes will slow down, and that can be a problem.

   The easy solution is again to subsample your data and work with that. It will show you the relevant signals in your data without slowing down your analysis.

   If that is not a solution for you, you need to pick analysis algorithms that work more efficiently. That typically means linear time algorithms. Unfortunately, many standard algorithms are not linear time, and even if they are, the implementation does not necessarily make it easy to fit data in batches where the model parameters can be updated one batch at a time. You often need to find packages specifically written for that, or make your own.

   One package that both provides a memory efficient linear model fitting (it avoids creating a model matrix that would have rows for each data point and solving equations for that) and functionality for updating the model in batches is the biglm package:

library(biglm)

You can use it for linear regression using the `biglm()` function instead of the `lm()` function, and you can use the `bigglm()` function for generalized linear regression instead of the `glm()` function, (see Chapter 6 for details).

If you are using a data frame format that stores the data on disk and has support for `biglm` (see the next section), the package will split the data into chunks that it can load into memory and analyze. If you do not have a package that handles this automatically, you can split the data into chunks yourself. As a toy example, we can consider the `cars` dataset and try to fit a linear model of stopping distance as a function of speed, but do this in batches of 10 data points. Of course, we can easily fit such a small dataset without splitting it into batches—we don't even need to use the `biglm()` function for it—but as an example, it will do.

Defining the slice indices requires some arithmetic and after that we can extract subsets of the data using the `slice()` function from `dplyr`. We can create a linear model from the first slice and then update using the following code:

```
slice_size <- 10
n <- nrow(cars)
slice <- cars %>% slice(1:slice_size)
model <- biglm(dist ~ speed, data = slice)
for (i in 1:(n/slice_size-1)) {
  slice <- cars %>% slice((i*slice_size+1):((i+1)*slice_size))
  model <- update(model, moredata = slice)
}
model
## Large data regression model: biglm(dist ~ speed, data = slice)
## Sample size =  50
```

Bayesian model fitting methods have a (somewhat justified) reputation for being slow, but Bayesian models based on conjugate priors are actually ideal for this. Having a conjugate prior means that the posterior distribution you get out of analyzing one dataset can be used as the prior distribution for the next dataset. This way you can split the data into slices, fit the first slice with a real prior, and the subsequent slices with the result of the previous model fits.

The Bayesian linear regression model in project 2 is one such model. There we implement an `update()` function that fits a new model based on a dataset and a previously fitted model. Using it on the `cars` data, splitting the data into chunks of size 10, would look like very similar to the `biglm` example.

Even better are models where you can analyze slices independently and then combine the results to get a model for the full dataset. These can not only be analyzed in batches, but the slices can be handled in parallel, exploiting multiple cores or multiple computer nodes. For gradient descent optimization approaches, you can compute gradients for slices independently and then combine them to make a step in the optimization.

There are no general solutions for dealing with data that is too large to be efficiently analyzed, though. It requires thinking about the algorithms used and usually also some custom implementation of these unless you are lucky and can find a package that can handle data in batches.

Too Large to Load

R wants to keep the data it works on in memory. So if your computer doesn't have the RAM to hold it, you are out of luck. At least if you work with the default data representations like `data.frames`. R usually also wants to use 32-bit integers for indices, and since it uses both positive and negative numbers for indices, you are limited to indexing around 2 billion data points. Even if you can hold more in memory.

There are different packages for dealing with this. One such is the `ff` package. It works with the kind of tables we have used so far but uses memory mapped files to represent the data and load data chunks into memory as needed.

```
library(ff)
```

It essentially creates flat files and has functionality for mapping chunks of these into memory when analyzing them.

It represents data frames as objects of the class `ffdf`. These behave just like data frames if you use them as such and you can translate a data frame into an `ffdf` object using the `as.ffdf()` function.

You can, for example, convert the `cars` data into an `ffdf` object using the following:

```
ffcars <- as.ffdf(cars)
summary(ffcars)
##       Length Class     Mode
## speed 50     ff_vector list
## dist  50     ff_vector list
```

Of course, if you can already represent a data frame in memory, there is no need for this translation, but `ff` also has functions for creating `ffdf` objects from files. If, for example, you have a large file as comma-separated values, you can use `read.csv.ffdf()`.

With `ff`, you get various functions for computing summary statistics efficiently from the memory mapped flat files. These are implemented as generic functions (generic functions are covered in Chapter 10) and this means that for most common summaries, you can work efficiently with `ffdf` objects. Not every function supports this, however, so sometimes functions will (implicitly) work on an `ffdf` object as if it were a plain `data.frame` object, which can result in the flat file being loaded into memory. This usually doesn't work if the data is too large to fit.

To deal with data that you cannot load into memory, you will have to analyze it in batches. This means that you need special functions for analyzing data, and quite often this means that you have to implement analysis algorithms yourself.

For linear models and generalized linear models, the `biglm` package implements them as generic functions. This means that the code that is actually run depends on the format in which the input data is provided. If you just give them an `ffdf` object, they will treat it as a `data.frame` object and not exploit that the data can be fitted in chunks. The `ffbase` package deals with this by implementing a special `bigglm()` function that works on `ffdf` objects. Although this is for generalized linear models, you can still use it for linear regression, since linear models are special cases of generalized linear models.

To fit a linear model (or generalized linear), just load the package:

```
library(ffbase)
```

If the data is represented as an `ffdf` object ,you use the special function for fitting the data:

```
model <- bigglm(dist ~ speed, data = ffcars)
summary(model)
## Large data regression model: bigglm(dist ~ speed, data = ffcars)
## Sample size =  50
##               Coef     (95%    CI)    SE
## (Intercept) -17.5791 -31.0960 -4.0622 6.7584
## speed         3.9324   3.1014  4.7634 0.4155
##                  p
## (Intercept) 0.0093
## speed       0.0000
```

The function takes a parameter, `chunksize`, to control how many data points are loaded into memory at a time. There is a more sensible default than 10 that we used previously, but generally you can just use the `bigglm()` function on `ffdf` objects like you would use `lm()` or `glm()` on `data.frame` objects.

You cannot use `ffdf` objects together with `dplyr`, which is the main drawback of using `ff` to represent data, but there is a development version of the package `ffbase2` that supports this. See the GitHub repository at https://github.com/edwindj/ffbase2 to learn more.

The `dplyr` package does provide support for different backends, such as relational databases. If you can work with data as flat files, there is no benefit to putting it in databases, but large datasets usually are stored in databases that are accessed through the *Structured Query Language* (SQL). This is a language that is worth learning but beyond the scope of this book, and in any case, `dplyr` can be used to access such databases. This means that you can write `dplyr` pipelines of data manipulation function calls. These calls will be translated into SQL expressions that are then sent to the database system, and you can get the results back.

With `dplyr` you can access commonly used database systems such as MySQL or PostgreSQL. These systems require that you set up a server for the data, though, so a simpler solution, if your data is not already stored in a database, is to use LiteSQL.

LiteSQL works just on your filesystem but provides a file format and ways of accessing it using SQL. You can open or create a LiteSQL file using the `src_sqlite()` function:

```
iris_db <- src_sqlite("iris_db.sqlite3", create = TRUE)
```

You load a dataset into it using `copy_to()`:

```
iris_sqlite <- copy_to(iris_db, iris, temporary = FALSE)
```

Of course, if you can already represent a data frame in RAM, you wouldn't usually copy it to a database. It only slows down analysis to go through a database system compared to keeping the data in memory. But the point is, of course, that you can populate the database outside of R and then access it using `dplyr`.

The `temporary` option to the function here ensures that the table you fill into the database survives between sessions. If you do not set `temporary` to `FALSE`, it will exist only as long as you have the database open; after you close it, it will be deleted. This is useful for many operations, but not what we want here.

Once you have a connection to a database, you can pull out a table using `tbl()`:

```
iris_sqlite <- tbl(iris_db, "iris")
```

Then you can use `dplyr` functions to make a query to it:

```
iris_sqlite %>% group_by(Species) %>%
  summarise(mean.Petal.Length = mean(Petal.Length))
## Source:   query [?? x 2]
## Database: sqlite 3.8.6 [iris_db.sqlite3]
##
##      Species mean.Petal.Length
##        <chr>             <dbl>
## 1    setosa             1.462
## 2 versicolor             4.260
## 3  virginica             5.552
```

Using `dplyr` with SQL databases is beyond the scope of this book, so I will just refer you to the documentation for it at https://cran.r-project.org/web/packages/dplyr/vignettes/databases.html.

Manipulating data using `dplyr` with a database backend is only useful for doing analysis exclusively using `dplyr`, of course. To fit models and such, you still have to batch data, so some custom code is usually still required.

Exercises

Try the following exercises to become more comfortable with the concepts discussed in this chapter.

Subsampling

Take the dataset you worked on the last two chapters and pick a subset of the data. Summarize it and compare to the results you get with the full data. Plot the subsamples and compare that to the plots you created with the full data.

Hex and 2D Density Plots

If you have used any scatterplots to look at your data, translate them into hex or 2D density plots.

CHAPTER 6

■ ■ ■

Supervised Learning

This chapter and the next concern the mathematical modeling of data that is the essential core of data science. We can call this statistics, or we can call it machine learning. At its core, it is the same thing. It is all about extracting information out of data.

Machine Learning

Machine learning is the discipline of developing and applying models and algorithms for learning from data. Traditional algorithms implement fixed rules for solving particular problems. Like sorting numbers or finding the shortest route between two cities. To develop algorithms like that, you need a deep understand of the problem you are trying to solve. A thorough understanding that you can rarely obtain unless the problem is particularly simple or you have abstracted away all the interesting cases. Far more often, you can collect examples of good or bad solutions to the problem you want to solve without being able to explain exactly why a given solution is good or bad. Or you can obtain data that provides examples of relationships between data you are interested in without necessarily understanding the underlying reasons for these relationships.

This is where machine learning can help. Machine learning concerns learning from data; you do not explicitly develop an algorithm for solving a particular problem. Instead, you use a generic learning algorithm that you feed examples of solutions to, and let it learn how to solve the problem from those examples.

This might sound very abstract, but most statistical modeling is indeed examples of this. Take for example a linear model $y = \alpha x + \beta + \varepsilon$ where ε is the stochastic noise (usually assumed to be normal distributed). When you want to model a linear relationship between x and y, you don't figure out α and β from first principle. You can write an algorithm for sorting numbers without having studied the numbers beforehand, but you cannot usually figure out what the linear relationship is between y and x without looking at data. When you fit the linear model, you are doing machine learning. (Well, I suppose if you do it by hand it isn't *machine* learning, but you are not likely to fit linear models by hand that often.) People typically do not call simple models like linear regression machine learning, but that is mostly because the term "machine learning" is much younger than these models. Linear regression is as much machine learning as neural networks are.

Supervised Learning

Supervised learning is used when you have variables you want to predict using other variables. Situations like linear regression where you have some input variables, for example, x, and you want a model that predicts output (or response) variables, $y = f(x)$.

Unsupervised learning, the topic of Chapter 7, is instead concerned with discovering patterns in data when you don't necessarily know what kind of questions you are interested in learning. When you don't have x and y values and want to know how they are related, but instead have a collection of data, and you want to discover what patterns there are in the data.

For the simplest case of supervised learning, we have one response variable, y, and one input variable, x, and we want to figure out a function, f, mapping input to output, i.e., so that $y = f(x)$. What we have to work with is example data of matching x and y. We can write that as vectors $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ where we want to figure out a function f such that $y_i = f(x_i)$.

We will typically accept that there might be some noise in our observations, so f doesn't map perfectly from x to y. So we can chance the setup slightly and assume that the data we have is $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{t} = (t_1, \ldots, t_n)$, where $\mathbf{t}$ is *target* values and where $t_i = y_i + \varepsilon_i$, $y_i = f(x_i)$, and ε_i is the error in the observation t_i.

How we model the error ε_i and the function f are choices that are up to us. It is only modelling, after all, and we can do whatever we want. Not all models are equally good, of course, so we need to be a little careful with what we choose and how we evaluate if the choice is good or bad, but in principle, we can do anything.

The way most machine learning works is that an algorithm, implicitly or explicitly, defines a class of parameterized functions $f(-; \theta)$, each mapping input to output $f(-; \theta): x \mapsto f(x; \theta) = y(\theta)$ (now the value we get for the output depends on the parameters of the function, θ), and the learning consists of choosing parameters θ such that we minimize the errors, i.e., so that $f(x_i; \theta)$ is as close to t_i as we can get. We want to get close for all our data points, or at least get close on average, so if we let $\mathbf{y}(\theta)$ denote the vector $(y(\theta)_1, \ldots, y(\theta)_n) = (f(x_1; \theta), \ldots, f(x_n; \theta))$ we want to minimize the distance from $\mathbf{y}(\theta)$ to $\mathbf{t}$, $\| \mathbf{y}(\theta) - \mathbf{t} \|$, for some distance measure $\|\cdot\|$.

Regression versus Classification

There are two types of supervised learning: regression and classification. Regression is used when the output variable we try to target is a number. Classification is used when we try to target categorical variables.

Take linear regression, $y = \alpha x + \beta$ (or $t = \alpha x + \beta + \varepsilon$). It is regression because the variable we are trying to target is a number. The parametrized class of functions, f_θ, are all lines. If we let $\theta = \theta_1, \theta_0$ and $\alpha = \theta_1$, $\beta = \theta_0$ then $y(\theta) = f(x; \theta) = \theta_1 x + \theta_0$ Fitting a linear model consists of finding the best θ, where *best* is defined as the θ that gets $\mathbf{y}(\theta)$ closest to $\mathbf{t}$. The distance measure used in linear regression is the squared Euclidean distance

$$\left| \mathbf{y}^{(\theta)} - \mathbf{t} \right|^2 = \sum_{i=1}^n \left(y_i(\theta) - t_i \right)^2 .$$

The reason it is the squared distance instead of just the distance is mostly mathematical convenience—it is easier to maximize θ that way—but also related to us interpreting the error term ε as normal distributed. Whenever you are fitting data in linear regression, you are minimizing this distance; you are finding the parameters θ that bests fit the data in the sense of:

$$\hat{\theta} = \underset{\theta_1, \theta_0}{\arg \min} \sum_{i=1}^n \left(\theta_1 x_i + \theta_0 - t_i \right)^2$$

For an example of classification, assume that the targets t_i are binary, encoded as 0 and 1, but that the input variables x_i are still real numbers. A common way of defining the mapping function $f(-; \theta)$ is to let it map x to the unit interval $[0, 1]$ and interpret the resulting $y(\theta)$ as the probability that t is 1. In a classification setting, you would then predict 0 if $f(x; \theta) < 0.5$ and predict 1 if $f(x; \theta) > 0.5$ (and have some strategy for dealing with $f(x; \theta) < 0.5$). In linear classification, the function f_θ could look like this:

$$f(x; \theta) = \sigma \left(\theta_1 x + \theta_0 \right)$$

where σ is a sigmoid function (a function mapping R → [0, 1] that is "S-shaped"). A common choice of σ is the logistic function $\sigma : z \mapsto \dfrac{1}{1+e^{-z}}$ in which case we call the fitting of $f(-;\theta)$ *logistic regression*.

Whether we are doing regression or classification, and whether we have linear models or not, we are simply trying to find parameters θ so that our predictions $y(\theta)$ are as close to our targets $\mathbf{t}$ as possible. The details that differ between different machine learning methods is how the class of prediction functions $f(-;\theta)$ is defined, what kind of parameters θ we have, and how we measure the distance between $y(\theta)$ and $\mathbf{t}$. There are *a lot* of different choices here and a lot of different machine learning algorithms. Many of them are already implemented in R, however, so we rarely will have to implement our own. We just need to find the right package that implements the learning algorithms we need.

Inference versus Prediction

A question always worth considering when we fit parameters of a model is this: do we care about the model parameters or do we just want to make a function that is good at predicting?

If you were taught statistics the same way I was, your introduction to linear regression was mostly focused on the model parameters. You inferred the parameters θ_1 and θ_0 mostly to figure out if $\theta_1 \neq 0$, i.e, to find out if there was a (linear) relationship between x and y or not. When we fit our function to data to learn about the parameters, we say we are doing inference, and we are inferring the parameters.

This focus on model parameters makes sense in many situations. In a linear model, the coefficient θ_1 tells us if there is a significant correlation between x and y, meaning we are statistically relatively certain that the correlation exists, and whether it is substantial, meaning that θ_1 is large enough to care about in practical situations.

When we care about model parameters, we usually want to know more than just the best fitting parameters, $\hat{\theta}$. We want to know how certain we are that the "true parameters" are close to our estimated parameters. This usually means estimating not just the best parameters but also confidence intervals or posterior distributions of parameters. How easy it is to estimate these depends very much on the models and algorithms used.

I put "true parameters" in quotes, where I talked about how close estimates were to the true parameters, for a good reason. True parameters only exist if the data you are analzsing were simulated from a function f_θ where some true θ exist. When you are estimating parameters, $\hat{\theta}$, you are looking for the best choice of parameters *assuming* that the data were generated by a function f_θ. Outside of statistics textbooks, there is no reason to think that your data was generated from a function in the class of functions you consider. Unless we are trying to model causal relationships—modeling how we think the world actually works as forces of nature—that is usually not an underlying assumption of model fitting. A lot of the theory we have for doing statistics on inferred parameters *does* assume that we have the right class of functions, which is where you get confidence intervals. In practice, data does not come from these sorts of functions so treat the results you get from theory with some skepticism.

We can get more empirical distributions of parameters directly from data if we have a lot of data–which we usually do have when doing data science–using sampling methods. I will briefly return to that later in this chapter.

We don't always care about the model parameters, though. For linear regression, it is easy to interpret what the parameters mean but in many machine learning models the parameters aren't that interpretable, and we don't really care about them. All we care about is if the model we have fitted is good at predicting the target values. To evaluate how well we expect a function to be able to predict is also something that we sometimes have theoretical results regarding, but as for parameter estimation, we shouldn't trust these too much. It is much better to use the actual data to estimate this and as for getting empirical distributions of model parameters it is something we return to later.

Whether you care about model parameters or not depends on your application and quite often on how you think your model relates to reality.

Specifying Models

The general pattern for specifying models in R is using what is called "formulas". The simplest form is y ~ x, which we should interpret as saying $y = f(x)$. Implicitly there is assumed some class of functions indexed with model parameters, $f(-; \theta)$, and which class of functions we are working with depends on which R functions we use.

Linear Regression

If we take a simple linear regression, $f_\theta(x) = \theta_1 x + \theta_0$, we need the function `lm()`.

For an example, we can use the built-in dataset `cars`, which just contains two variables, speed and breaking distance, where we can consider speed the x value and breaking distance the y value.

```
cars %>% head
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10
```

If we plot the dataset (see Figure 6-1), we see that there is a very clear linear relationship between speed and distance.

```
cars %>% ggplot(aes(x = speed, y = dist)) +
  geom_point() +
  geom_smooth(method = "lm")
```
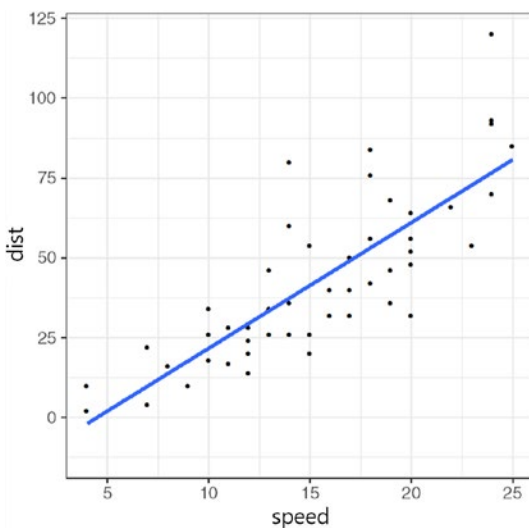

Figure 6-1. *Plot of breaking distance versus speed for cars*

In this plot, I used the method `"lm"` for the smoothed statistics to see the fit. By default the `geom_ smooth()` function would have given us a loess curve, but since we are interested in linear fits, we tell it to use the `lm` method. By default `geom_smooth()` will also plot the uncertainty of the fit. This is the gray area in the plot. This is the area where the line is likely to be (assuming that the data is generated by a linear model). Do not confuse this with where data points are likely to be, though. If target values are given by $t = \theta_1 x + \theta_0 + \varepsilon$ where ε has a very large variance, then even if we knew θ_1 and θ_0 with high certainty we still wouldn't be able to predict with high accuracy where any individual point would fall. There is a difference between prediction accuracy and inference accuracy. We might know model parameters with very high accuracy without being able to predict very well. We might also be able to predict very well without knowing all model parameters well. If a given model parameter has little influence on where target variables fall, then the training data gives us little information about that parameter. This usually doesn't happen unless the model is more complicated than it needs to be, though, since we often want to remove parameters that do not affect the data.

To actually fit the data and get information about the fit, we use the `lm()` function with the model specification, `dist ~ speed`, and we can use the `summary()` function to see information about the fit:

```
cars %>% lm(dist ~ speed, data = .) %>% summary
##
## Call:
## lm(formula = dist ~ speed, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123
## speed         3.9324     0.4155   9.464 1.49e-12
##
## (Intercept) *
## speed       ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Or we can use the `coefficients()` function to get the point estimates and the `confint()` function to confidence intervals for the parameters:

```
cars %>% lm(dist ~ speed, data = .) %>% coefficients
## (Intercept)       speed
##  -17.579095    3.932409
cars %>% lm(dist ~ speed, data = .) %>% confint
##                 2.5 %     97.5 %
## (Intercept) -31.167850  -3.990340
## speed         3.096964   4.767853
```

Here, (Intercept) is θ_0 and speed is θ_1.

To illustrate the fitting procedure and drive the point home, we can explicitly draw models with different parameters, i.e., draw lines with different choices of θ. To simplify matters, I am going to set $\theta_0 = 0$. Then I can plot the lines $y = \theta_1 x$ for different choices of θ_1 and visually see the fit; see Figure 6-2.

```
predict_dist <- function(speed, theta_1)
  data.frame(speed = speed,
             dist = theta_1 * speed,
             theta = as.factor(theta_1))

cars %>% ggplot(aes(x = speed, y = dist, colour = theta)) +
  geom_point(colour = "black") +
  geom_line(data = predict_dist(cars$speed, 2)) +
  geom_line(data = predict_dist(cars$speed, 3)) +
  geom_line(data = predict_dist(cars$speed, 4)) +
  scale_color_discrete(name=expression(theta[1]))
```
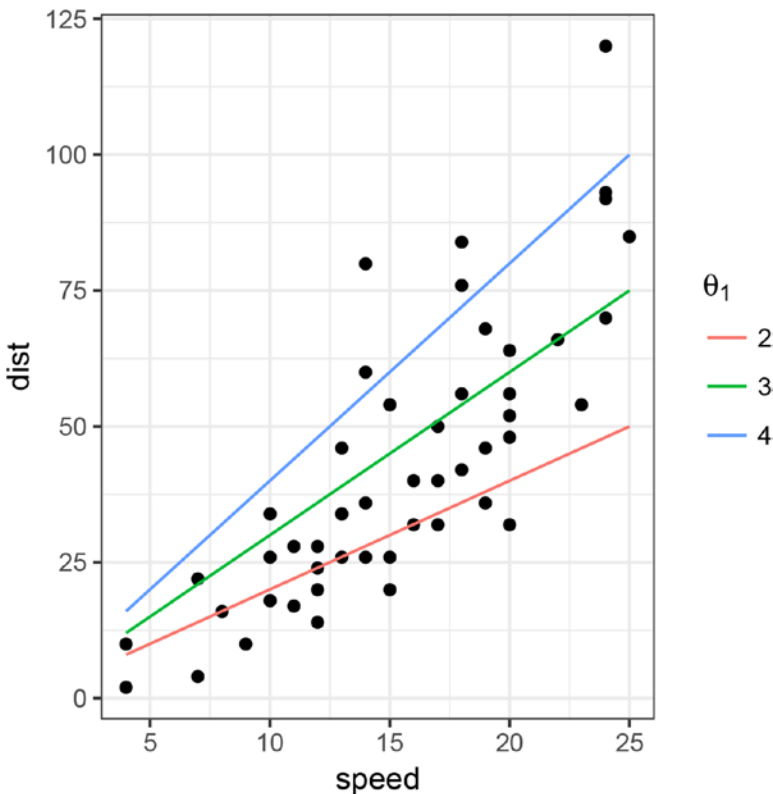

Figure 6-2. *Prediction lines for different choices of parameters*

In this plot, I want to color the lines according to their θ_1 parameter but since the cars data frame doesn't have a theta column I hardwire that the dots should be plotted in black. The lines are plotted according to their theta value, which I set in the predict_dist() function.

Each of the lines shows a choice of model. Given an input value x they all produce an output value. $y(\theta) = f(x; \theta)$. So we can fix θ and consider the mapping $x \mapsto \theta_1 x$. This is the function we use when predicting the output for a given value of x. If we fix x instead we can also see it as a function of θ: $\theta_1 \mapsto \theta_1 x$. This is what we use when we fit parameters to the data, because if we keep our dataset fixed this mapping defines an error function, that is, a function that given parameters gives us a measure of how far our predicted values are from our target values. If, as before, our input values and target values are vectors $\mathbf{x}$ and $\mathbf{t}$, then the error function is as follows:

$$E_{x,t}\left(\theta_i\right) = \sum_{i=1}^{n}\left(\theta_1 x_i - t_i\right)^2$$

We can plot the errors against different choices of θ_1 (see Figure 6-3). Where this function is minimized, we find our best estimate for θ_1.

```r
thetas <- seq(0, 5, length.out = 50)
fitting_error <- Vectorize(function(theta)
  sum((theta * cars$speed - cars$dist)**2)
)

data.frame(thetas = thetas, errors = fitting_error(thetas)) %>%
  ggplot(aes(x = thetas, y = errors)) +
  geom_line() +
  xlab(expression(theta[1])) + ylab("")
```
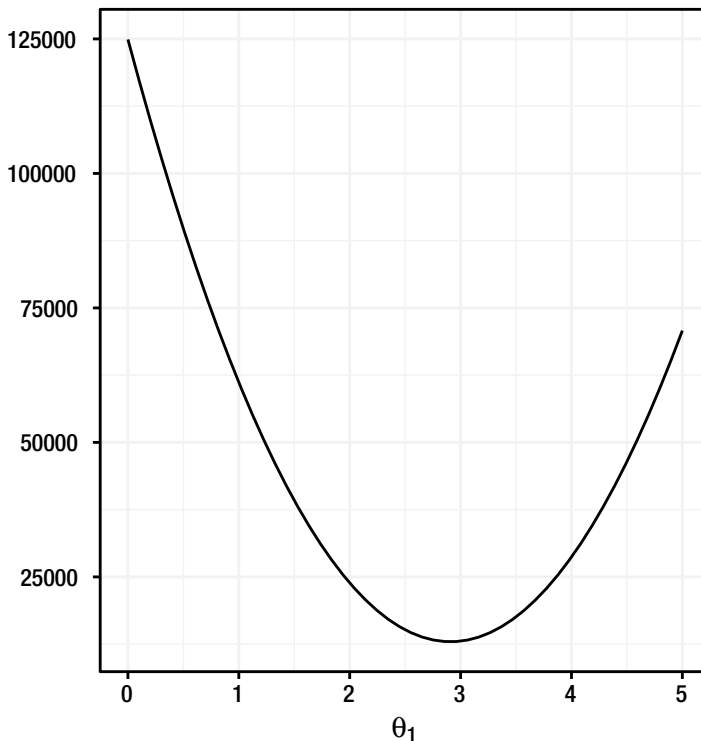

Figure 6-3. *Error values for different choices of parameters*

To wrap up this example, we can also plot and fit the best model where $\theta_0 = 0$. The formula needed to remove the intercept is of the form "y ~ x - 1". It is the "- 1" that removes the intercept.

```
cars %>% lm(dist ~ speed - 1, data = .) %>% coefficients
##    speed
## 2.909132
```

We can also plot this regression line, together with the confidence interval for where it lies, using geom_smooth(). See Figure 6-4. Here, though, we need to use the formula y ~ x - 1 rather than dist ~ speed - 1. This is because the geom_smooth() function works on the ggplot2 layers that have x- and y-coordinates and not the data in the data frame as such. We map the speed variable to the x-axis and the dist variable to the y-variable in the aesthetics, but it is x and y that geom_smooth() works on.

```
cars %>% ggplot(aes(x = speed, y = dist)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x - 1)
```
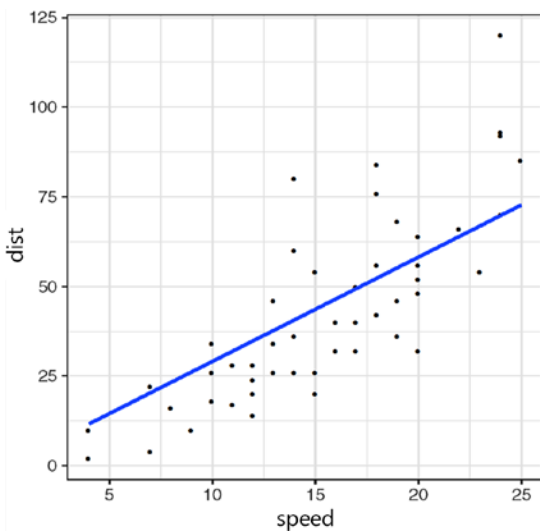

Figure 6-4. *Best regression line going through* (*0,0*)

Logistic Regression (Classification, Really)

Using other statistical models works the same way. We specify the class of functions, f_θ, using a formula and use a function to fit its parameters. Consider binary classification and logistic regression.

Here we can use the breast cancer data from the `mlbench` library discussed in Chapter 3 and ask if the clump thickness has an effect on the risk of a tumor being malignant. That is, we want to see if we can predict the `Class` variable from the `Cl.thickness` variable.

```
library(mlbench)
data("BreastCancer")
BreastCancer %>% head
##        Id Cl.thickness Cell.size Cell.shape
## 1 1000025            5         1          1
## 2 1002945            5         4          4
## 3 1015425            3         1          1
## 4 1016277            6         8          8
## 5 1017023            4         1          1
## 6 1017122            8        10         10
##   Marg.adhesion Epith.c.size Bare.nuclei
## 1             1            2           1
## 2             5            7          10
## 3             1            2           2
## 4             1            3           4
## 5             3            2           1
## 6             8            7          10
##   Bl.cromatin Normal.nucleoli Mitoses     Class
## 1           3               1       1    benign
## 2           3               2       1    benign
## 3           3               1       1    benign
## 4           3               7       1    benign
## 5           3               1       1    benign
## 6           9               7       1 malignant
```

We can plot the data against the fit, as shown in Figure 6-5. Since the malignant status is either 0 or 1, the points would overlap but if we add a little jitter to the plot we can still see them, and if we make them slightly transparent, we can see the density of the points.

```
BreastCancer %>%
  ggplot(aes(x = Cl.thickness, y = Class)) +
  geom_jitter(height = 0.05, width = 0.3, alpha=0.4)
```
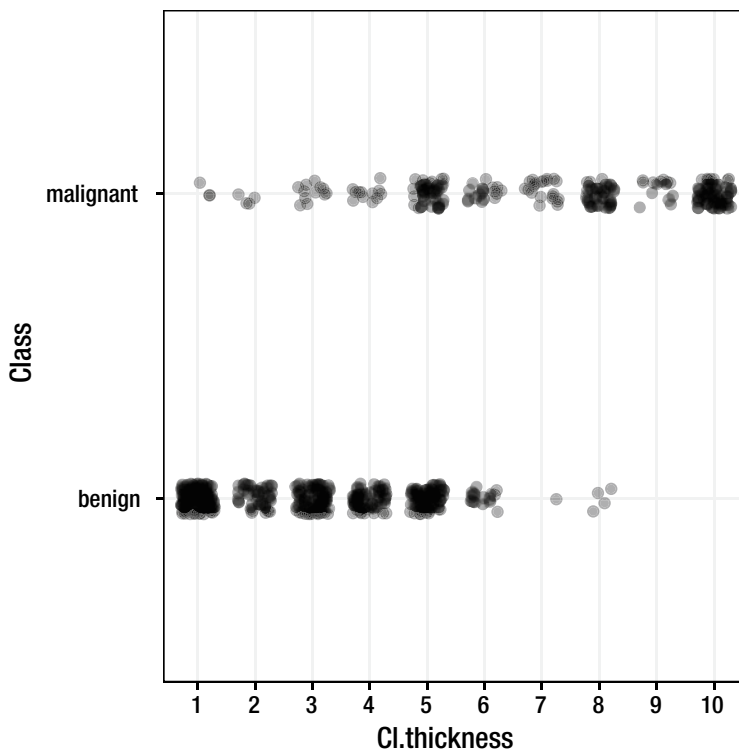
Figure 6-5. *Breast cancer class versus clump thickness*

For classification we still specify the prediction function $y = f(x)$ using the formula y ~ x. The outcome parameter for y ~ x is just binary now. To fit a logistic regression we need to use the glm() function (*generalized* linear model) with the family set to "binomial". This specifies that we use the logistic function to map from the linear space of x and θ to the unit interval. Aside from that, fitting and getting results is very similar.

We cannot directly fit the breast cancer data with logistic regression, though. There are two problems. The first is that the breast cancer dataset considers the clump thickness ordered factors, but for logistic regression we need the input variable to be numeric. While generally, it is not advisable to directly translate categorical data into numeric data, judging from the plot it seems okay in this case. Using the function as. numeric() will do this, but remember that this is a risky approach when working with factors! It actually would work for this dataset, but we will use the safer approach of first translating the factor into strings and then into numbers. The second problem is that the glm() function expects the response variable to be numerical, coding the classes like 0 or 1, while the BreastCancer data encodes the classes as a factor. Generally, it varies a little from algorithm to algorithm whether a factor or a numerical encoding is expected for classification, so you always need to check the documentation for that, but in any case, it is simple enough to translate between the two representations.

We can translate the input variable to numerical values and the response variable to 0 and 1 and plot the data together with a fitted model, as shown in Figure 6-6. For the geom_smooth() function, we specify that the method is glm and that the family is binomial. To specify the family, we need to pass this argument

on to the smoothing method, and that is done by giving the method.args parameter a list of named
parameters; here we just give it list(family = "binomial).

```
BreastCancer %>%
  mutate(Cl.thickness.numeric =
                as.numeric(as.character(Cl.thickness))) %>%
  mutate(IsMalignant = ifelse(Class == "benign", 0, 1)) %>%
  ggplot(aes(x = Cl.thickness.numeric, y = IsMalignant)) +
```
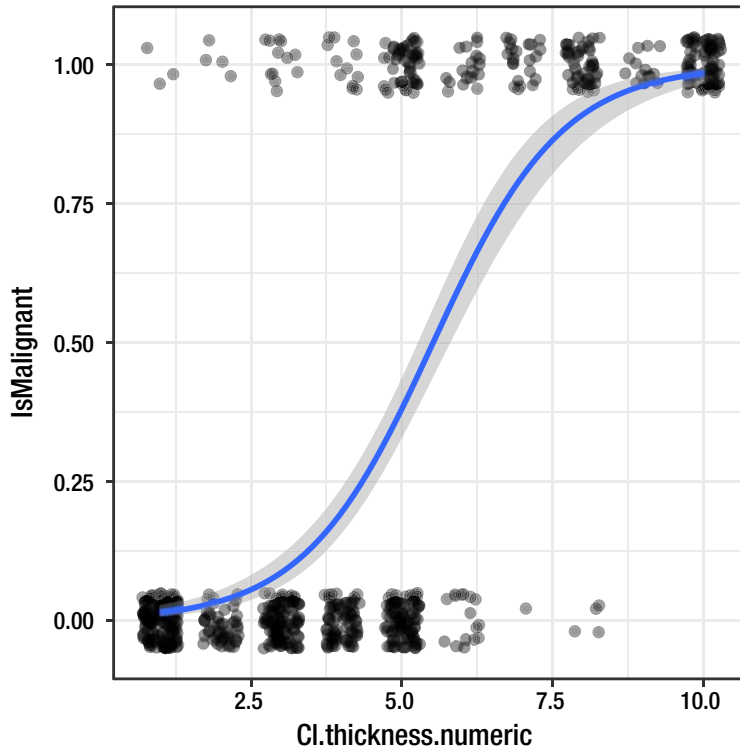

Figure 6-6. *Logistic regression fit to breast cancer data*

```
  geom_jitter(height = 0.05, width = 0.3, alpha=0.4) +
  geom_smooth(method = "glm",
              method.args = list(family = "binomial"))
```

To actually get the fitted object, we use glm() like we used lm() for the linear regression.

```
BreastCancer %>%
  mutate(Cl.thickness.numeric =
                as.numeric(as.character(Cl.thickness))) %>%
  mutate(IsMalignant = ifelse(Class == "benign", 0, 1)) %>%
  glm(IsMalignant ~ Cl.thickness.numeric,
      family = "binomial",
      data = .)
```

```
##
## Call:  glm(formula = IsMalignant ~ Cl.thickness.numeric, family = "binomial",
##     data = .)
##
## Coefficients:
##         (Intercept)  Cl.thickness.numeric
##             -5.1602                 0.9355
##
## Degrees of Freedom: 698 Total (i.e. Null);  697 Residual
## Null Deviance:        900.5
## Residual Deviance: 464.1      AIC: 468.1
```

Model Matrices and Formula

Most statistical models and machine learning algorithms actually creates a map not from a single value, $f(-;\theta): x \mapsto y$, but from a vector, $f(-;\theta): \mathbf{x} \mapsto y$. When we fit a line for single x and y values we are actually also working with fitting a vector because we have both the x values and the intercept to fit. That is why the model has two parameters, θ_0 and θ_0. For each x value, we are actually using the vector $(1, x)$ where the 1 is used to fit the intercept.

We shouldn't confuse this with the vector we have as input to the model fitting, though. If we have data $(\mathbf{x}, \mathbf{t})$ to fit, then we already have a vector for our input data. But what the linear model actually sees is a matrix for $\mathbf{x}$, so we'll call that X. This matrix, know as the *model matrix*, has a row per value in $\mathbf{x}$ and it has two columns, one for the intercept and one for the x values.

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

We can see what model matrix R generates for a given dataset and formula using the `model.matrix()` function. For the cars data, if we want to fit `dist` versus `speed` we get this:

```
cars %>%
  model.matrix(dist ~ speed, data = .) %>%
  head(5)
##   (Intercept) speed
## 1           1     4
## 2           1     4
## 3           1     7
## 4           1     7
## 5           1     8
```

If we remove the intercept, we simply get this:

```
cars %>%
  model.matrix(dist ~ speed - 1, data = .) %>%
  head(5)
##   speed
## 1     4
```

```
## 2      4
## 3      7
## 4      7
## 5      8
```

Pretty much all learning algorithms work on a model matrix so, in R, they are implemented to take a formula to specify the model and then build the model matrix from that and the input data.

For linear regression, the map is a pretty simple one. If we let the parameters $\theta = (\theta_0, \theta_1)$ then it is just multiplying that with the model matrix, X.

$$\theta^T X = (\theta_0, \theta_1) \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} \theta_0 + \theta_1 x_1 \\ \theta_0 + \theta_1 x_2 \\ \theta_0 + \theta_1 x_3 \\ \vdots \\ \theta_0 + \theta_1 x_n \end{bmatrix}$$

This combination of formulas and model matrices is a powerful tool for specifying models. Since all the algorithms we use for fitting data works on model matrices anyway, there is no reason to hold back on how complex formulas to give them. The formulas will just be translated into model matrices anyhow, and they can all deal with them.

If you want to fit more than one parameter, no problem. You just give write y ~ x + z and the model matrix will have three columns.

$$X = \begin{bmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ 1 & x_3 & z_3 \\ \vdots & \vdots & \vdots \\ 1 & x_n & z_n \end{bmatrix}$$

Our model fitting functions are just as happy to fit this model matrix like the one we get from just a single variable.

So if we wanted to fit the breast cancer data to both cell thickness and cell size, we can do that just by adding both explanatory variables in the formula.

```
BreastCancer %>%
  mutate(Cl.thickness.numeric =
           as.numeric(as.character(Cl.thickness)),
         Cell.size.numeric =
           as.numeric(as.character(Cell.size))) %>%
  mutate(IsMalignant = ifelse(Class == "benign", 0, 1)) %>%
  model.matrix(IsMalignant ~ Cl.thickness.numeric + Cell.size.numeric,
               data = .) %>%
  head(5)
##   (Intercept) Cl.thickness.numeric
## 1           1                    5
## 2           1                    5
## 3           1                    3
## 4           1                    6
## 5           1                    4
##   Cell.size.numeric
```

```
## 1                    1
## 2                    4
## 3                    1
## 4                    8
## 5                    1
```

Then the generalized linear model fitting function will happily work with that:

```
BreastCancer %>%
  mutate(Cl.thickness.numeric =
            as.numeric(as.character(Cl.thickness)),
         Cell.size.numeric =
            as.numeric(as.character(Cell.size))) %>%
  mutate(IsMalignant = ifelse(Class == "benign", 0, 1)) %>%
  glm(IsMalignant ~ Cl.thickness.numeric + Cell.size.numeric,
      family = "binomial",
      data = .)
##
## Call:  glm(formula = IsMalignant ~ Cl.thickness.numeric + Cell.size.numeric,
##      family = "binomial", data = .)
##
## Coefficients:
##         (Intercept)  Cl.thickness.numeric
##             -7.1517                0.6174
##    Cell.size.numeric
##              1.1751
##
## Degrees of Freedom: 698 Total (i.e. Null);  696 Residual
## Null Deviance:        900.5
## Residual Deviance: 212.3      AIC: 218.3
```

Translating data into model matrices also works for factors, they are just represented as a binary vector for each level:

```
BreastCancer %>%
  mutate(IsMalignant = ifelse(Class == "benign", 0, 1)) %>%
  model.matrix(IsMalignant ~ Bare.nuclei, data = .) %>%
  head(5)
##   (Intercept) Bare.nuclei2 Bare.nuclei3
## 1           1            0            0
## 2           1            0            0
## 3           1            1            0
## 4           1            0            0
## 5           1            0            0
##   Bare.nuclei4 Bare.nuclei5 Bare.nuclei6
## 1            0            0            0
## 2            0            0            0
## 3            0            0            0
## 4            1            0            0
## 5            0            0            0
```

```
##    Bare.nuclei7 Bare.nuclei8 Bare.nuclei9
## 1             0            0            0
## 2             0            0            0
## 3             0            0            0
## 4             0            0            0
## 5             0            0            0
##    Bare.nuclei10
## 1              0
## 2              1
## 3              0
## 4              0
## 5              0
```

The translation for ordered factors gets a little more complicated, but R will happily do it for you:

```
BreastCancer %>%
  mutate(IsMalignant = ifelse(Class == "benign", 0, 1)) %>%
  model.matrix(IsMalignant ~ Cl.thickness, data = .) %>%
  head(5)
##   (Intercept) Cl.thickness.L Cl.thickness.Q
## 1           1    -0.05504819    -0.34815531
## 2           1    -0.05504819    -0.34815531
## 3           1    -0.27524094    -0.08703883
## 4           1     0.05504819    -0.34815531
## 5           1    -0.16514456    -0.26111648
##   Cl.thickness.C Cl.thickness^4 Cl.thickness^5
## 1      0.1295501     0.33658092    -0.21483446
## 2      0.1295501     0.33658092    -0.21483446
## 3      0.3778543    -0.31788198    -0.03580574
## 4     -0.1295501     0.33658092     0.21483446
## 5      0.3346710     0.05609682    -0.39386318
##   Cl.thickness^6 Cl.thickness^7 Cl.thickness^8
## 1     -0.3113996      0.3278724      0.2617852
## 2     -0.3113996      0.3278724      0.2617852
## 3      0.3892495     -0.5035184      0.3739788
## 4     -0.3113996     -0.3278724      0.2617852
## 5      0.2335497      0.2459043     -0.5235703
##   Cl.thickness^9
## 1     -0.5714300
## 2     -0.5714300
## 3     -0.1632657
## 4      0.5714300
## 5      0.3809534
```

If you want to include interactions between your parameters, you specify that using * instead of +:

```
BreastCancer %>%
  mutate(Cl.thickness.numeric =
           as.numeric(as.character(Cl.thickness)),
```

139

```
        Cell.size.numeric =
          as.numeric(as.character(Cell.size))) %>%
  mutate(IsMalignant = ifelse(Class == "benign", 0, 1)) %>%
  model.matrix(IsMalignant ~ Cl.thickness.numeric * Cell.size.numeric,
            data = .) %>%
  head(5)
##   (Intercept) Cl.thickness.numeric
## 1           1                    5
## 2           1                    5
## 3           1                    3
## 4           1                    6
## 5           1                    4
##   Cell.size.numeric
## 1                 1
## 2                 4
## 3                 1
## 4                 8
## 5                 1
##   Cl.thickness.numeric:Cell.size.numeric
## 1                                       5
## 2                                      20
## 3                                       3
## 4                                      48
## 5                                       4
```

How interactions are modeled depends a little bit on whether your parameters are factors or numeric but for numeric values, the model matrix will just contain a new column with the two values multiplied. For factors, you will get a new column for each level of the factor:

```
BreastCancer %>%
  mutate(Cl.thickness.numeric =
          as.numeric(as.character(Cl.thickness))) %>%
  mutate(IsMalignant = ifelse(Class == "benign", 0, 1)) %>%
  model.matrix(IsMalignant ~ Cl.thickness.numeric * Bare.nuclei, data = .) %>%
  head(3)
##   (Intercept) Cl.thickness.numeric Bare.nuclei2
## 1           1                    5            0
## 2           1                    5            0
## 3           1                    3            1
##   Bare.nuclei3 Bare.nuclei4 Bare.nuclei5
## 1            0            0            0
## 2            0            0            0
## 3            0            0            0
##   Bare.nuclei6 Bare.nuclei7 Bare.nuclei8
## 1            0            0            0
## 2            0            0            0
## 3            0            0            0
##   Bare.nuclei9 Bare.nuclei10
## 1            0             0
```

```
## 2                  0                    1
## 3                  0                    0
##    Cl.thickness.numeric:Bare.nuclei2
## 1                                     0
## 2                                     0
## 3                                     3
##    Cl.thickness.numeric:Bare.nuclei3
## 1                                     0
## 2                                     0
## 3                                     0
##    Cl.thickness.numeric:Bare.nuclei4
## 1                                     0
## 2                                     0
## 3                                     0
##    Cl.thickness.numeric:Bare.nuclei5
## 1                                     0
## 2                                     0
## 3                                     0
##    Cl.thickness.numeric:Bare.nuclei6
## 1                                     0
## 2                                     0
## 3                                     0
##    Cl.thickness.numeric:Bare.nuclei7
## 1                                     0
## 2                                     0
## 3                                     0
##    Cl.thickness.numeric:Bare.nuclei8
## 1                                     0
## 2                                     0
## 3                                     0
##    Cl.thickness.numeric:Bare.nuclei9
## 1                                     0
## 2                                     0
## 3                                     0
##    Cl.thickness.numeric:Bare.nuclei10
## 1                                      0
## 2                                      5
## 3                                      0
```

The interaction columns all have : in their name, and you can specify an interaction term directly by writing that in the model formula as well:

```
BreastCancer %>%
  mutate(Cl.thickness.numeric =
            as.numeric(as.character(Cl.thickness))) %>%
  mutate(IsMalignant = ifelse(Class == "benign", 0, 1)) %>%
  model.matrix(IsMalignant ~ Cl.thickness.numeric : Bare.nuclei, data = .) %>%
  head(3)
##   (Intercept) Cl.thickness.numeric:Bare.nuclei1
## 1           1                                 5
```

141

```
## 2              1                                    0
## 3              1                                    0
##    Cl.thickness.numeric:Bare.nuclei2
## 1                                0
## 2                                0
## 3                                3
##    Cl.thickness.numeric:Bare.nuclei3
## 1                                0
## 2                                0
## 3                                0
##    Cl.thickness.numeric:Bare.nuclei4
## 1                                0
## 2                                0
## 3                                0
##    Cl.thickness.numeric:Bare.nuclei5
## 1                                0
## 2                                0
## 3                                0
##    Cl.thickness.numeric:Bare.nuclei6
## 1                                0
## 2                                0
## 3                                0
##    Cl.thickness.numeric:Bare.nuclei7
## 1                                0
## 2                                0
## 3                                0
##    Cl.thickness.numeric:Bare.nuclei8
## 1                                0
## 2                                0
## 3                                0
##    Cl.thickness.numeric:Bare.nuclei9
## 1                                0
## 2                                0
## 3                                0
##    Cl.thickness.numeric:Bare.nuclei10
## 1                                0
## 2                                5
## 3                                0
```

If you want to use all the variables in your data except the response variable, you can even use the formula y ~ . where the . will give you all parameters in your data except y.

Using formulas and model matrices also means that we do not have to use are data raw. We can transform it before we give it to our learning algorithms. In general, we can transform our data using a function ϕ. It is called phi because we call what it produces *features* of our data and the point of it is to pull out the relevant features of the data to give to the learning algorithm. It usually maps from vectors to vectors, so you can use it to transform each row in your raw data into the rows of the model matrix, which we will then call Φ instead of X.

$$\Phi = \begin{bmatrix} -\phi(\boldsymbol{x}_1)- \\ -\phi(\boldsymbol{x}_2)- \\ -\phi(\boldsymbol{x}_3)- \\ \dots \\ -\phi(\boldsymbol{x}_n)- \end{bmatrix}$$

If this sounds very abstract, perhaps it will help to see some examples. We go back to the cars data but this time, we want to fit a polynomial to the data instead of a line. If d denotes breaking distance and s the speed, then we want to fit $d = \theta_0 + \theta_1 s + \theta_2 s^2 + \dots + \theta_n s^n$. We'll just do $n = 2$ so we want to fit a second-degree polynomial. Don't be confused about the higher degrees of the polynomial, it is still a linear model. The *linear* in linear model refers to the θ parameters, not the data. We just need to map the single s parameter into a vector with the different polynomial degrees, so 1 for the intercept, s for the linear component, and s^2 for the squared component. So $\phi(s) = (1, s, s^2)$.

We can write that as a formula. There we don't need to specify the intercept term explicitly—it will be included by default and if we don't want it we have to remove it explicitly—but we need speed, and we need speed^2.

```
cars %>%
  model.matrix(dist ~ speed + speed^2, data = .) %>%
  head
##   (Intercept) speed
## 1           1     4
## 2           1     4
## 3           1     7
## 4           1     7
## 5           1     8
## 6           1     9
```

Now this doesn't quite work, and the reason is that multiplication is interpreted as interaction terms even if it is interaction with the parameter itself. And interaction with itself doesn't go into the model matrix because that would just be silly.

To avoid that problem we need to tell R that the speed^2 term should be interpreted just the way it is. We do that using the identity function, I():

```
cars %>%
  model.matrix(dist ~ speed + I(speed^2), data = .) %>%
  head
##   (Intercept) speed I(speed^2)
## 1           1     4         16
## 2           1     4         16
## 3           1     7         49
## 4           1     7         49
## 5           1     8         64
## 6           1     9         81
```

Now our model matrix has three columns, which is precisely what we want.

We can fit the polynomial using the linear model function like this:

```
cars %>% lm(dist ~ speed + I(speed^2), data = .) %>%
  summary
##
```

```
## Call:
## lm(formula = dist ~ speed + I(speed^2), data = .)
##
## Residuals:
##      Min      1Q  Median      3Q      Max
## -28.720  -9.184  -3.188   4.628   45.152
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.47014   14.81716   0.167    0.868
## speed        0.91329    2.03422   0.449    0.656
## I(speed^2)   0.09996    0.06597   1.515    0.136
##
## Residual standard error: 15.18 on 47 degrees of freedom
## Multiple R-squared:  0.6673, Adjusted R-squared:  0.6532
```
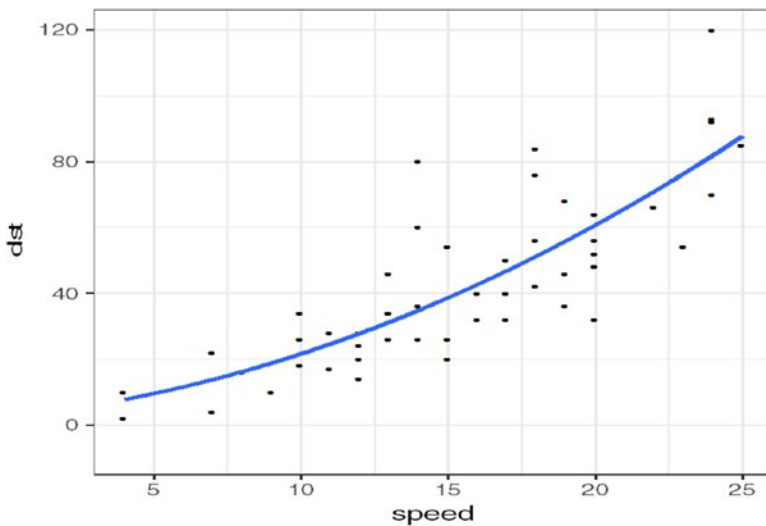

Figure 6-7. *The cars data fitted to a second degree polynomial*

```
## F-statistic: 47.14 on 2 and 47 DF,  p-value: 5.852e-12
```

Or we can plot it like this (see Figure 6-7):

```
cars %>% ggplot(aes(x = speed, y = dist)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x + I(x^2))
```

This is a slightly better fitting model, but that wasn't the point. You can see how you can transform data in a formula to have different features to give to your fitting algorithms.

Validating Models

How did I know the polynomial fit was better than the linear fit? Well, theoretically a second-degree polynomial should always be a better fit than a line since a line is a special case of a polynomial. We just set θ_2 to zero. If the best-fitted polynomial doesn't have $\theta_2 = 0$ then that is because we can fit the data better if it is not.

The result of fitting the polynomial tells me, in the output from the `summary()` function, that the variables are not significant. It tells me that both from the linear and the squared component, though, so it isn't that useful. Clearly, the points are on a line so it cannot be correct that there isn't a linear component. I cannot use the summary that much because it is only telling me that when I have both components, then neither of them are statistically significant. That doesn't mean much.

But should I even care, though? If I know that the more complex model always fits better then shouldn't I just always use it? The problem with that idea is that while the most complex model will always fit the training data better, it will not necessarily generalize better. If I use a high enough degree polynomial—if I have a degree that is the same as the number of data points—I can fit the data perfectly. But it will be fitting both the systematic relationship between x and y and *also* the statistical errors in our targets t. It might be utterly useless for predicting point number $n + 1$.

What I really need to know is whether one or the other model is better at predicting the distance from the speed.

We can fit the two models and get their predictions using the `predict()` function. It takes the fitted model as the first argument and data to predict on as the second:

```
line <- cars %>% lm(dist ~ speed, data = .)
poly <- cars %>% lm(dist ~ speed + I(speed^2), data = .)

predict(line, cars) %>% head
##         1         2         3         4         5
## -1.849460 -1.849460  9.947766  9.947766 13.880175
##         6
## 17.812584
predict(poly, cars) %>% head
##         1         2         3         4         5
##  7.722637  7.722637 13.761157 13.761157 16.173834
##         6
## 18.786430
```

Evaluating Regression Models

To compare the two models, we need a measure of how well they fit. Since both models are fitting the squared distances from predictions to targets, a fair measure would be looking at the mean squared error. The unit of that would be distance squared, though, so we usually use the square root of this mean distance to measure the quality of the predictions, which would give us the errors in the distance unit.

```
rmse <- function(x,t) sqrt(mean(sum((t - x)^2)))

rmse(predict(line, cars), cars$dist)
## [1] 106.5529
rmse(predict(poly, cars), cars$dist)
## [1] 104.0419
```

Now clearly the polynomial fits slightly better, which it should based on theory, but there is a bit of a cheat here. We are looking at how the models work on the data we used to fit them. The more complex model will always be better at this. That is the problem we are dealing with. The more complex model might be overfitting the data and capturing the statistical noise we don't want it to capture. What we really want to know is how well the models generalizes; how well do they work on data they haven't already seen and used to fit their parameters?

We have used all the data we have to fit the models. That is generally a good idea. You want to use all the data available to get the best-fitted model. But to compare models we need to have data that isn't used in the fitting.

We can split the data into two sets, one we use for training and the other we use to test the models. There are 50 data points so I can take the first 25 to train my models on and the next 25 to test them on.

```
training_data <- cars[1:25,]
test_data <- cars[26:50,]

line <- training_data %>% lm(dist ~ speed, data = .)
poly <- training_data %>% lm(dist ~ speed + I(speed^2), data = .)

rmse(predict(line, test_data), test_data$dist)
## [1] 88.89189
rmse(predict(poly, test_data), test_data$dist)
## [1] 83.84263
```

The second-degree polynomial is still better, but I am also still cheating. There is more structure in my dataset than just the speed and distances. The data frame is sorted according to the distance so the training set has all the short distances and the test data all the long distances. The are not similar. That is not good.

In general, you cannot know if there is such structure in your data. In this particular case, it is easy to see because the structure is that obvious, but sometimes it is subtler. So when you split your data into training and test data, you will want to sample data points randomly. That gets rid of the structure that is in the order of the data points.

We can use the sample() function to sample randomly zeros and ones:

```
sampled_cars <- cars %>%
  mutate(training = sample(0:1, nrow(cars), replace = TRUE))

sampled_cars %>% head
##   speed dist training
## 1     4    2        1
## 2     4   10        0
## 3     7    4        0
## 4     7   22        0
## 5     8   16        1
## 6     9   10        1
```

This doesn't give us 50/50 training and test data since which data point gets into each category will depend on the random samples, but it will be roughly half the data we get for training:

```
training_data <- sampled_cars %>% filter(training == 1)
test_data <- sampled_cars %>% filter(training == 0)
```

```
training_data %>% head
##    speed dist training
## 1     4     2        1
## 2     8    16        1
## 3     9    10        1
## 4    11    28        1
## 5    12    20        1
## 6    13    26        1
test_data %>% head
##    speed dist training
## 1     4    10        0
## 2     7     4        0
## 3     7    22        0
## 4    10    18        0
## 5    10    26        0
## 6    10    34        0
```

Now we can get a better estimate of how the functions are working:

```
line <- training_data %>% lm(dist ~ speed, data = .)
poly <- training_data %>% lm(dist ~ speed + I(speed^2), data = .)
```

```
rmse(predict(line, test_data), test_data$dist)
## [1] 82.45426
rmse(predict(poly, test_data), test_data$dist)
## [1] 81.2045
```

Now, of course, the accuracy scores depend on the random sampling when we create the training and test data so you might want to use more samples. We return to that in the next section.

Now, once you have figured out what the best model is you will still want to train it on all the data you have. Splitting the data is just a tool for evaluating how well different models work. For the final model, you choose to work with you will always want to fit it with all the data you have.

Evaluating Classification Models

If you want to do classification rather than regression, then the root mean square error is not the function to use to evaluate your model. With classification, you want to know how many data points are classified correctly and how many are not.

As an example, we can take the breast cancer data and fit a model:

```
formatted_data <- BreastCancer %>%
  mutate(Cl.thickness.numeric =
           as.numeric(as.character(Cl.thickness)),
         Cell.size.numeric =
           as.numeric(as.character(Cell.size))) %>%
  mutate(IsMalignant = ifelse(Class == "benign", 0, 1))

fitted_model <- formatted_data %>%
  glm(IsMalignant ~ Cl.thickness.numeric + Cell.size.numeric, data = .)
```

To get its prediction, we can again use `predict()`, but we will see that for this particular model the predictions are probabilities of a tumor being malignant. By default, the model we created with `glm()` will be in "logit" units, but we can use the `type` parameter to get it in the input unit used with probabilities.

```
predict(fitted_model, formatted_data, type = "response") %>% head
##         1          2          3          4
## 0.05266571 0.65374326 0.01591478 0.99740926
##         5          6
## 0.02911157 0.99992795
```

We would need to translate that into actual predictions. The natural choice here is to split the probabilities at 50%. If we are more certain that a tumor is malignant than benign, we will classify it as malignant.

```
classify <- function(probability) ifelse(probability < 0.5, 0, 1)
classified_malignant <- classify(predict(fitted_model, formatted_data))
```

Where you want to put the threshold of how to classify depends on your data and the consequences of the classification. In a clinical situation, maybe you want to examine further a tumor with less than 50% probability that it is malignant, or maybe you don't want to tell patients that a tumor might be malignant if it is only 50% probable. The classification should take into account how sure you are about the classification, and that depends a lot on the situation you are in. Of course, you don't want to bet against the best knowledge you have, so I am not suggesting that you should classify everything below probability 75% as the "false" class, for instance. The only thing you gain from this is making worse predictions than you could. But sometimes you want to leave some data unpredicted. So here you can use the probabilities the model predicts to leave some data points as `NA`. How you want to use that your prediction gives you probabilities instead of just classes—assuming it does, it depends on the algorithm used for classifying—is up to you and the situation you are analyzing.

Confusion Matrix

In any case, if we just put the classification threshold at 50/50 then we can compare the predicted classification against the actual classification using the `table()` function, as follows:

```
table(formatted_data$IsMalignant, classified_malignant)
##    classified_malignant
##        0   1
##   0 450   8
##   1  42 199
```

This table, contrasting predictions against true classes, is known as the *confusion matrix*. The rows count how many zeros and ones we see in the `formatted_data$IsMalignant` argument and the columns how many zeros and ones we see in the `classified_malignant` argument. So the first row is where the data says the tumors *are not* malignant and the second row is where the data says that the tumors *are* malignant. The first column is where the predictions say the tumors are not malignant while the second column is where the predictions say that they are.

This, of course, depends on the order of the arguments to `table()`, it doesn't know which argument contains the data classes and which contains the model predictions. It can be a little hard to remember which dimension, rows or columns, are the predictions but you can provide a parameter, `dnn` (*dimnames* names), to make the table remember it for you.

```
table(formatted_data$IsMalignant, classified_malignant,
      dnn=c("Data", "Predictions"))
##      Predictions
## Data   0   1
##    0 450   8
##    1  42 199
```

The correct predictions are on the diagonal, and the off-diagonal values are where our model predicts incorrectly.

The first row is where the data says that tumors are not malignant. The first element, where the model predicts that the tumor is benign, and the data agrees, is called the *true negatives*. The element to the right of it, where the model says a tumor is malignant but the data says it is not, is called the *false positives*.

The second row is where the data says that tumors are malignant. The first column is where the prediction says that it isn't a malignant tumor, and these are called the *false negatives*. The second column is the cases where both the model and the data says that the tumor is malignant. That is the *true positives*.

The terms *positives* and *negatives* are a bit tricky here. I managed to sneak them past you by having the classes called zeros and ones that you already associate with true and false and positive and negative, and by having a dataset where it was more natural to think of malignant tumors as being the ones we want to predict.

The classes do not have to be zeros and ones. That was just easier in this particular model where I had to translate the classes into zeros and ones for the logistic classification anyway. But really, the classes are "benign" and "malignant".

```
classify <- function(probability)
  ifelse(probability < 0.5, "benign", "malignant")
classified <- classify(predict(fitted_model, formatted_data))

table(formatted_data$Class, classified,
      dnn=c("Data", "Predictions"))
##             Predictions
## Data         benign malignant
##    benign       450         8
##    malignant     42       199
```

What is *positive* and what is *negative* now depends on whether we want to predict malignant or benign tumors. Of course, we really want to predict both well, but the terminology considers one class true and the other false.

The terms carry over into several of the terms used in classification described next, where the classes and predictions are not so explicitly stated. In the confusion matrix we can always see exactly what the true classes are and what the predicted classes are, but once we start summarizing it in various ways, this information is no longer explicitly available. The summaries still will often depend on which class we consider "positive" and which we consider "negative", though.

Since which class is which really is arbitrary, so it is always worth a thought deciding which you want to call which and definitely something you want to make explicit in any documentation of your analysis.

Accuracy

The simplest measure of how well a classification is doing is the *accuracy*. It measures how many classes it gets right out of the total, so it is the diagonal values of the confusion matrix divided by the total.

```
confusion_matrix <- table(formatted_data$Class, classified,
```

```
                                dnn=c("Data", "Predictions"))
```

```
(accuracy <- sum(diag(confusion_matrix))/sum(confusion_matrix))
## [1] 0.9284692
```

This measure of the classification accuracy is pretty simple to understand, but you have to be careful in what you consider a good accuracy. Of course "good" is a subjective term, so I'll get technical and think in terms of "better than chance". That means that your baseline for what you consider "good" is randomly guessing. This, at least, is not subjective.

It is still something you have to consider a bit carefully, though. Because what does randomly guessing mean? We naturally think of a random guess as one that chooses either class with the same 50% probability. If the data has the same number of observations for each of the two classes, then that would be a good strategy and would get the average accuracy of 0.5. So better than chance would, in that case, be better than 0.5. The data doesn't have to have the same number of instances for each class. The breast cancer data does not. The breast cancer data has more benign tumors than malignant tumors.

```
table(BreastCancer$Class)
##
##    benign malignant
##       458       241
```

Here, you would be better off guessing more benign than malignant. If you had to guess and already knew that you were more than twice as likely to have a benign than a malignant tumor, you would always guess benign.

```
tbl <- table(BreastCancer$Class)
tbl["benign"] / sum(tbl)
##    benign
## 0.6552217
```

Always guessing "benign" is a lot better than 50/50. Of course, it is arguable whether this is guessing but it *is* a strategy for guessing and you want your model to do better than this simple strategy.

Always guessing the most frequent class–assuming that the frequency of the classes in the dataset is representative for the frequency in new data as well (which *is* a strong assumption)—is the best strategy for guessing.

If you actually want to see "random" guessing, you can get an estimate of this by simply permuting the classes in the data. The function sample() can do this:

```
table(BreastCancer$Class, sample(BreastCancer$Class))
##
##               benign malignant
##    benign        311       147
##    malignant     147        94
```

This gives you an estimate for random guessing, but since it is random, you would want to get more than one to get a feeling for how much it varies with the guess.

```
accuracy <- function(confusion_matrix)
  sum(diag(confusion_matrix))/sum(confusion_matrix)
```

```r
replicate(8, accuracy(table(BreastCancer$Class,
                            sample(BreastCancer$Class))))
## [1] 0.5193133 0.5507868 0.5650930 0.5422031
## [5] 0.5565093 0.5479256 0.5708155 0.5450644
```

As you can see, even random permutations do better than 50/50–but the better guess is still just the most frequent class and at the very least you would want to beat that.

Sensitivity and Specificity

We want a classifier to have a high accuracy, but accuracy isn't everything. The costs in real life of misclassifying often have different consequences when you classify something like a benign tumor as malignant from when you classify a malignant tumor as benign. In a clinical setting, you have to weight the false positives against the false negatives and the consequences they have. You are interested in more than pure accuracy.

We usually use two measures of the predictions of a classifier that takes that into account. The *specificity* and the *sensitivity* of the model. The first measure captures how often the model predicts a negative case correctly. In the breast cancer data, this is how often, when the model predicts a tumor as benign, it actually is.

```r
(specificity <- confusion_matrix[1,1]/
  (confusion_matrix[1,1] + confusion_matrix[1,2]))
## [1] 0.9825328
```

The sensitivity does the same thing but for the positives. It captures how well, when the data has the positive class, your model predicts this correctly.

```r
(sensitivity <- confusion_matrix[2,2]/
  (confusion_matrix[2,1] + confusion_matrix[2,2]))
## [1] 0.8257261
```

If your accuracy is 100%, then both of these will also be 100%. But there is usually a trade-off between the two. Using the "best guessing" strategy of always picking the most frequent class will set one of the two to 100% but at the cost of the other. In the breast cancer data the best guess is always benign, the negative case, and always guessing benign will give us a specificity of 100%

This strategy can always achieve 100% for one of the two measures but at the cost of setting the other to 0%. If you only ever guess at one class, you are perfect when the data is actually from that class, but you are always wrong when the data is from the other class.

Because of this, we are never interested in optimizing either measure alone. That is trivial. We want to optimize both. We might consider specificity more important than sensitivity or vice versa, but even if we want one to be 100% we also want the other to be as good as we can get it.

To evaluate how much better than chance we are doing, we can again compare to random permutations. This tells us how well we are doing compared to random guesses for both.

```r
specificity <- function(confusion_matrix)
  confusion_matrix[1,1]/(confusion_matrix[1,1]+confusion_matrix[1,2])

sensitivity <- function(confusion_matrix)
  confusion_matrix[2,2]/(confusion_matrix[2,1]+confusion_matrix[2,2])
```

```
prediction_summary <- function(confusion_matrix)
  c("accuracy" = accuracy(confusion_matrix),
    "specificity" = specificity(confusion_matrix),
    "sensitivity" = sensitivity(confusion_matrix))

random_prediction_summary <- function()
  prediction_summary(table(BreastCancer$Class,
                           sample(BreastCancer$Class)))

replicate(3, random_prediction_summary())
##                  [,1]      [,2]      [,3]
## accuracy    0.5565093 0.5422031 0.5336195
## specificity 0.6615721 0.6506550 0.6441048
## sensitivity 0.3568465 0.3360996 0.3236515
```

Other Measures

The specificity is also known as the *true negative rate* since it measures how many of the negative classifications are true. Similarly, the sensitivity is known as the *true positive rate*. There are analogue measures for getting thins wrong. The *false negative rate* is the analogue of the true negative rate, but instead of dividing the true negatives by all the negatives, it divides the false negatives by all the negatives. The *false positive rate* similarly divides the false positives by all the positives. Having these two measures together with sensitivity and specificity are not really adding much. The true negative rate is just one minus the false negative rate and similar for the true positive rate and false positive rate. They just focus on when the model gets things wrong instead of when it gets things right.

All four measures split the confusing matrix into the two rows. They look at when the data says the class is true and when the data says the class is false. We can also look at the columns instead, and consider when the predictions are true and when the predictions are false.

When we look at the column where the predictions are false–for the breast cancer when the tumors are predicted as benign–we have the *false omission rate*, which is the false negatives divided by all the predicted negatives:

```
confusion_matrix[2,1] / sum(confusion_matrix[,1])
## [1] 0.08536585
```

The *negative predictive value* is instead the true negatives divided by the predicted negatives.

```
confusion_matrix[1,1] / sum(confusion_matrix[,1])
## [1] 0.9146341
```

These two will always sum to one so we are really only interested in one of them, but which we choose is determined by which we find more important.

For the predicted positives we have the *positive predictive values* and *false discovery rate*.

```
confusion_matrix[2,2] / sum(confusion_matrix[,2])
## [1] 0.9613527
confusion_matrix[1,2] / sum(confusion_matrix[,2])
## [1] 0.03864734
```

The false discovery rate, usually abbreviated FDR, is the one most frequently used. It is closely related to the threshold used on p-values (the significance thresholds) in classical hypothesis testing. Remember that

if you have a 5% significance threshold in classical hypothesis testing, it means that when the null hypothesis is true, you will predict it is false 5% of the time. This means that your false discovery rate is 5%.

The classical approach is to pick an acceptable false discovery rate, by convention this is 5% but there is nothing magical about that number–it is simply convention–and then that threshold determines how extreme a test statistic has to be before we switch from predicting a negative to predicting a positive. This approach entirely ignores the cases where the data is from the positive class. It has its uses, but not for classification where you have data from both the positive class and the negative class so we will not consider it more here. You will have seen it in statistics classes, and you can learn more about it in any statistics textbook.

More Than Two Classes

All of the above considers a situation where we have two classes, one we call positive and one we call negative. This is a common case, which is the reason we have so many measures for dealing with it, but it is not the only case. Quite often we need to classify data into more than two classes.

The only measure you can reuse there is the accuracy. The accuracy is always the sum along the diagonal divided by the total number of observations. Accuracy still isn't everything in those cases. Some classes are perhaps more important to get right than others–or just harder to get right than others–so you have to use a lot of sound judgment when evaluating a classification. There are just fewer rules of thumbs to use here, so you are more left to your own judgment. Sampling approaches

To validate classifiers I suggested splitting the data into a training dataset and a test dataset. I also mentioned that there might be hidden structures in your dataset so you always want to make this split a random split of the data.

Generally, there are a lot of benefits you can get out of randomly splitting your data, or randomly subsampling from your data. We have mostly considered prediction in this chapter, where splitting the data into training and a test data lets us evaluate how well a model does at predicting on unseen data. But randomly splitting or subsampling from data is also very useful for inference. When we do inference, we can typically get confidence intervals for model parameters, but these are based on theoretical results that assume that the data is from some (usually) simple distribution. Data is generally not. If you want to know how a parameter is distributed from the empirical distribution of the data, you will want to subsample and see what distribution you get.

Random Permutations of Your Data

With the `cars` data we split the observations into two equally sized datasets. Since this data is ordered by the stopping distance, splitting it into the first half and the second half makes the datasets different in distributions.

The simplest approach to avoiding this problem is to reorder your data randomly before you split it. Using the `sample()` function we can get a random permutation of any input vector—we saw that earlier—and we can exploit this to get a random order of your dataset.

Using `sample(1:n)` we get a random permutation of the numbers from 1 to *n*. We can select rows in a data frame by giving it a vector of indices for the rows. Combining these two observations we can get a random order of `cars` observations this way:

```
permuted_cars <- cars[sample(1:nrow(cars)),]
permuted_cars %>% head(3)
##    speed dist
## 34    18   76
## 25    15   26
## 29    17   32
```

The numbers to the left of the data frame are the original row numbers (it really is the row names, but it is the same in this case).

We can write a simple function for doing this for general data frames:

```
permute_rows <- function(df) df[sample(1:nrow(df)),]
```

Using this, we can add it to a data analysis pipeline, where we would write:

```
permuted_cars <- cars %>% permute_rows
```

Splitting the data into two sets, training and testing, is one approach to subsampling, but a general version of this is used in something called *cross-validation*. Here the idea is to get more than one result out of the random permutation we use. If we use a single training/test split, we only get one estimate of how a model performs on a dataset. Using more gives us an idea about the variance of this.

We can split a dataset into *n* groups like this:

```
group_data <- function(df, n) {
  groups <- rep(1:n, each = nrow(df)/n)
  split(df, groups)
}
```

You don't need to understand the details of this function for now, but it is a good exercise to try to figure it out, so you are welcome to hit the documentation and see if you can work it out.

The result is a list, a data structure we haven't explored yet (but feel free to skip ahead to Chapter 8 to read about it). It is necessary to use a list here since vectors or data frames cannot hold complex data, so if we combined the result in one of those data structures, they would just be merged back into a single data frame here.

As it is, we get something that contains *n* data structures that each have a data frame of the same form as the cars data:

```
cars %>% permute_rows %>% group_data(5) %>% head(1)
## $`1`
##    speed dist
## 10    11   17
## 9     10   34
## 7     10   18
## 43    20   64
## 46    24   70
## 26    15   54
## 39    20   32
## 29    17   32
## 17    13   34
## 21    14   36
```

All you really need to know for now is that to get an entry in a list, you need to use [[]] indexing instead of [] indexing:

```
grouped_cars <- cars %>% permute_rows %>% group_data(5)
grouped_cars[[1]]
##    speed dist
## 30    17   40
## 12    12   14
```

```
## 10     11    17
## 24     15    20
## 27     16    32
## 4       7    22
## 18     13    34
## 28     16    40
## 5       8    16
## 42     20    56
```

If you use [ ] you will also get the data, but the result will be a list with one element, which is not what you want (but is what head( ) gave you previously):

```
grouped_cars[1]
## $`1`
##     speed dist
## 30     17    40
## 12     12    14
## 10     11    17
## 24     15    20
## 27     16    32
## 4       7    22
## 18     13    34
## 28     16    40
## 5       8    16
## 42     20    56
```

We can use the different groups to get estimates of the model parameters in the linear model for cars:

```
grouped_cars[[1]] %>%
  lm(dist ~ speed, data = .) %>%
  .$coefficients
## (Intercept)       speed
##   -7.004651    2.674419
```

With a bit of programming, we can get the estimates for each group:

```
estimates <- grouped_cars[[1]] %>%
  lm(dist ~ speed, data = .) %>%
  .$coefficients

for (i in 2:length(grouped_cars)) {
  group_estimates <- grouped_cars[[i]] %>%
    lm(dist ~ speed, data = .) %>%
    .$coefficients
  estimates <- rbind(estimates, group_estimates)
}

estimates
##                 (Intercept)     speed
## estimates         -7.004651 2.674419
## group_estimates  -25.709091 4.366234
## group_estimates  -20.037464 4.741457
## group_estimates  -18.849797 4.336942
```

155

```
## group_estimates  -13.846071 3.207831
```

Right away, I will stress that this is not the best way to do this, but it shows you how it could be done. We will get to better approaches shortly. Still, you can see how splitting the data this way lets us get distributions for model parameters.

There are several reasons why this isn't the optimal way of coding this. The row names are ugly, but that is easy to fix. The way we combine the estimates in the data frame is inefficient—although it doesn't matter much with such a small dataset—and later in the book, we will see why. The main reason, though, is that explicit loops like this make it hard to follow the data transformations since it isn't a pipeline of processing.

The package purrr lets us work on lists using pipelines. You import the package as follows:

```
library(purrr)
```

Then you have access to the function map() that lets you apply a function to each element of the list:

```
estimates <- grouped_cars %>%
  map(. %>% lm(dist ~ speed, data = .) %>% .$coefficients)

estimates
## $`1`
## (Intercept)        speed
##    -7.004651     2.674419
##
## $`2`
## (Intercept)        speed
##   -25.709091     4.366234
##
## $`3`
## (Intercept)        speed
##   -20.037464     4.741457
##
## $`4`
## (Intercept)        speed
##   -18.849797     4.336942
##
## $`5`
## (Intercept)        speed
##   -13.846071     3.207831
```

The result is another list, but we really want a data frame, and we can get that using the piece of magical invocation called do.call("rbind", .):

```
estimates <- grouped_cars %>%
  map(. %>% lm(dist ~ speed, data = .) %>% .$coefficients) %>%
  do.call("rbind", .)

estimates
```

```
##   (Intercept)    speed
## 1   -7.004651 2.674419
## 2  -25.709091 4.366234
## 3  -20.037464 4.741457
## 4  -18.849797 4.336942
## 5  -13.846071 3.207831
```

There isn't much to say about this. It just combines the elements in a list using the rbind() function, and the result is a data frame. It is not particularly pretty, but it is just the invocation you need here.

Cross-Validation

A problem with splitting the data into many small groups is that we get a large variance in estimates. Instead of working with each little dataset independently we can remove one of the datasets and work on all the others. This will mean that our estimates are no longer independent, but the variance goes down. The idea of removing a subset of the data and then cycle through the groups evaluating a function for each group that is left out, is called *cross-validation*. Well, it is called cross-validation when we use it to validate prediction, but it works equally well for inferring parameters.

If we already have the grouped data frames in a list, we can remove one element from the list using [-i] indexing—just as we can for vectors—and the result is a list containing all the other elements. We can then combine the elements in the list into a single data frame using the do.call("rbind", .) magical invocation.

So we can write a function that takes the grouped data frames and gives us another list of data frames that contains data where a single group is left out.

```
cross_validation_groups <- function(grouped_df) {
  result <- vector(mode = "list", length = length(grouped_df))
  for (i in seq_along(grouped_df)) {
    result[[i]] <- grouped_df[-i] %>% do.call("rbind", .)
  }
  result
}
```

The vector(mode = "list", length = length(grouped_df)) is a little misleading here. It doesn't actually create a vector but a list. It is not my fault, but just how R creates lists.

The function does have a for loop, something I suggested you avoid in general, but by constructing a list up front and then assigning to elements of the list, I prevent the performance penalties often seen when working with loops. By isolating the loop in a function, we can still write data processing pipelines without using loops.

We could have combined this with the group_data() function, but I prefer to write functions that do one simple thing and combine them instead using pipelines. We can use this function and all the stuff we did previously to get estimates using cross-validation.

```
cars %>%
  permute_rows %>%
  group_data(5) %>%
  cross_validation_groups %>%
  map(. %>% lm(dist ~ speed, data = .) %>% .$coefficients) %>%
```

```
  do.call("rbind", .)
##      (Intercept)    speed
## [1,]  -16.42502 3.911860
## [2,]  -17.92765 3.835678
## [3,]  -17.97865 3.976987
## [4,]  -17.51737 4.010904
## [5,]  -17.58658 3.898504
```

Where cross-validation is typically used is when leaving out a subset of the data for testing and using the rest for training.

We can write a simple function for splitting the data this way, similar to the `cross_validation_groups()` function. It cannot return a list of data frames but needs to return a list of lists, each list containing a training data frame and a test data frame. It looks like this:

```
cross_validation_split <- function(grouped_df) {
  result <- vector(mode = "list", length = length(grouped_df))
  for (i in seq_along(grouped_df)) {
    training <- grouped_df[-i] %>% do.call("rbind", .)
    test <- grouped_df[[i]]
    result[[i]] <- list(training = training, test = test)
  }
  result
}
```

Don't worry if you don't understand all the details of it. After reading later programming chapters, you will. Right now, I hope you just get the gist of it.

I will not show you the result. It is just long and not that pretty, but if you want to see it, you can type:

```
cars %>%
  permute_rows %>%
  group_data(5) %>%
  cross_validation_split
```

As we have seen, we can index into a list using `[[]]`. We can also use the `$name` indexing like we can for data frames, so if we have a list `lst` with a `training` dataset and a `test` dataset, we can get them as `lst$training` and `lst$test`:

```
prediction_accuracy_cars <- function(test_and_training) {
  result <- vector(mode = "numeric",
                   length = length(test_and_training))
  for (i in seq_along(test_and_training)) {
    training <- test_and_training[[i]]$training
    test <- test_and_training[[i]]$test
    model <- training %>% lm(dist ~ speed, data = .)
    predictions <- test %>% predict(model, data = .)
    targets <- test$dist
    result[i] <- rmse(targets, predictions)
  }
  result
}
```

You should be able to understand most of this function even though we haven't covered much R programming yet, but if you do not, then don't worry.

You can then add this function to your data analysis pipeline to get the cross-validation accuracy for your different groups:

```
cars %>%
  permute_rows %>%
  group_data(5) %>%
  cross_validation_split %>%
  prediction_accuracy_cars
## [1] 197.0962 244.4090 192.4210 228.2972 179.1443
```

The prediction accuracy function isn't general. It is hardwired to use a linear model and to the formula `dist ~ speed`. It is possible to make a more general function, but that requires a lot more R programming skills ,so we will leave the example here.

Selecting Random Training and Testing Data

In the example earlier where I split the data `cars` into training and test data using `sample(0:1, n, replacement = TRUE)`. I didn't permute the data and then deterministically split it afterwards. Instead, I sampled training and test based on probabilities of picking any given row as training and test.

What I did was adding a column to the data frame where I randomly picked whether an observation should be used for the training or for the test data. Since it required first adding a new column and then selecting rows based on it, it doesn't work well as part of a data analysis pipeline. We can do better, and slightly generalize the approach at the same time.

To do this, I shamelessly steal two functions from the documentation of the `purrr` package. They do the same thing as the grouping function I wrote previously. If you do not quite follow the example, do not worry. But I suggest you try to read the documentation for any function you do not understand and at least try to work out what is going on. Follow it as far as you can, but don't sweat it if there are things you do not fully understand. After finishing the entire book, you can always return to the example.

The previous grouping function defined groups by splitting the data into *n* equally sized groups. The first function here instead samples from groups specified by probabilities. It creates a vector naming the groups, just as I did previously. It just names the groups based on named values in a probability vector and creates a group vector based on probabilities given by this vector:

```
random_group <- function(n, probs) {
  probs <- probs / sum(probs)
  g <- findInterval(seq(0, 1, length = n), c(0, cumsum(probs)),
                    rightmost.closed = TRUE)
  names(probs)[sample(g)]
}
```

If we pull the function apart, we see that it first normalizes a probability vector. This just means that if we give it a vector that doesn't sum to one, it will still work. To use it, it makes the code easier to read if it already sums to one, but the function can deal with it, even if it doesn't.

The second line, which is where it is hardest to read, just splits the unit interval into *n* subintervals and assign a group to each subinterval based on the probability vector. This means that the first chunk of the *n*

intervals is assigned to the first group, the second chunk to the second group, and so on. It is not doing any sampling yet, it just partitions the unit interval into *n* subinterval and assigns each subinterval to a group.

The third line is where it is sampling. It now takes the *n* subintervals, permutes them, and returns the names of the probability vector each one falls into.

We can see it in action by calling it a few times. We give it a probability vector where we call the first probability "training" and the second "test":

```
random_group(8, c(training = 0.5, test = 0.5))
## [1] "test"     "training" "training" "test"
## [5] "training" "test"     "test"     "training"
random_group(8, c(training = 0.5, test = 0.5))
## [1] "training" "training" "test"     "training"
## [5] "training" "test"     "test"     "test"
```

We get different classes out when we sample, but each class is picked with 0.5 probability. We don't have to pick them 50/50, though; we can choose more training than test data, for example.

```
random_group(8, c(training = 0.8, test = 0.2))
## [1] "training" "training" "test"     "training"
## [5] "training" "test"     "training" "training"
```

The second function just uses this random grouping to split the dataset. It works exactly like the cross-validation splitting we saw earlier.

```
partition <- function(df, n, probs) {
  replicate(n, split(df, random_group(nrow(df), probs)), FALSE)
}
```

The function replicates the subsampling *n* times. Here *n* is not the number of observations you have in the data frame, but a parameter to the function. It lets you pick how many subsamples of the data you want.

We can use it to pick four random partitions. Here with training and test select with 50/50 probability.

```
random_cars <- cars %>% partition(4, c(training = 0.5, test = 0.5))
```

If you evaluate it on your computer and look at random_cars, you will see that resulting values are a lot longer now. This is because we are not looking at smaller datasets this time; we have as many observations as we did before (which is 50), but we have randomly partitioned them.

We can combine this partition() function with the accuracy prediction from before.

```
random_cars %>% prediction_accuracy_cars
## [1] 93.75803 81.01278 70.82501 80.13141
```

Examples of Supervised Learning Packages

So far in this chapter, we looked at classical statistical methods for regression (linear models) and classification (logistic regression) but there are many machine learning algorithms for both, and many are available as R packages.

They all work similarly to the classical algorithms. You give the algorithms a dataset and a formula specifying the model matrix. From this, they do their magic. All the ideas presented in this chapter can be used together with them.

Next I go through a few packages, but there are many more. A Google search should help you find a package if there is a particular algorithm you are interested in applying.

I present their use with the same two datasets we have used previously, the `cars` data where we aim at predicting the stopping distance from the speed and the `BreastCancer` where we try to predict the class from the cell thickness. For both these cases, the classical models–a linear model and a logistic regression–are more ideal solutions and these models will not out-compete them, but for more complex datasets they can usually be quite powerful.

Decision Trees

Decision trees work by building a tree from the input data, splitting on a parameter in each inner node according to a variable value. This can be splitting on whether a numerical value is above or below a certain threshold or which level a factor has.

Decision trees are implemented in the `rpart` package and models are fitted, just as linear models are:

```r
library(rpart)

model <- cars %>% rpart(dist ~ speed, data = .)
rmse(predict(model, cars), cars$dist)
## [1] 117.1626
```

Building a classifying model works very similar. We do not need to translate the cell thickness into a numerical value, though; we can use the data frame as it is (but you can experiment with translating factors into numbers if you are interested in exploring this).

```r
model <- BreastCancer %>%
  rpart(Class ~ Cl.thickness, data = .)
```

The predictions when we used the `glm()` function were probabilities for the tumor being malignant. The predictions made using the decision tree gives you the probabilities *both* for being benign and being malignant:

```r
predict(model, BreastCancer) %>% head
##       benign malignant
## 1 0.82815356 0.1718464
## 2 0.82815356 0.1718464
## 3 0.82815356 0.1718464
## 4 0.82815356 0.1718464
## 5 0.82815356 0.1718464
```

```
## 6 0.03289474 0.9671053
```

To get a confusion matrix, we need to translate these probabilities into the corresponding classes. The output of predict() is not a data frame but a matrix so we first convert it into a data frame using the function as.data.frame() and then we use the %$% operator in the pipeline to get access to the columns by name in the next step.

```
predicted_class <- predict(model, BreastCancer) %>%
  as.data.frame %$%
  ifelse(benign > 0.5, "benign", "malignant")

table(BreastCancer$Class, predicted_class)
##             predicted_class
##              benign malignant
##   benign        453         5
##   malignant      94       147
```

Another implementation of decision trees is the ctree() function from the party package:

```
library(party)
model <- cars %>% ctree(dist ~ speed, data = .)
rmse(predict(model, cars), cars$dist)
## [1] 117.1626
model <- BreastCancer %>%
  ctree(Class ~ Cl.thickness, data = .)

predict(model, BreastCancer) %>% head
## [1] benign    benign    benign    benign
## [5] benign    malignant
## Levels: benign malignant
table(BreastCancer$Class, predict(model, BreastCancer))
##
##             benign malignant
##   benign        453         5
##   malignant      94       147
```

I like this package slightly more since it can make plots of the fitted models; see Figure 6-8:

```
cars %>% ctree(dist ~ speed, data = .) %>% plot
```
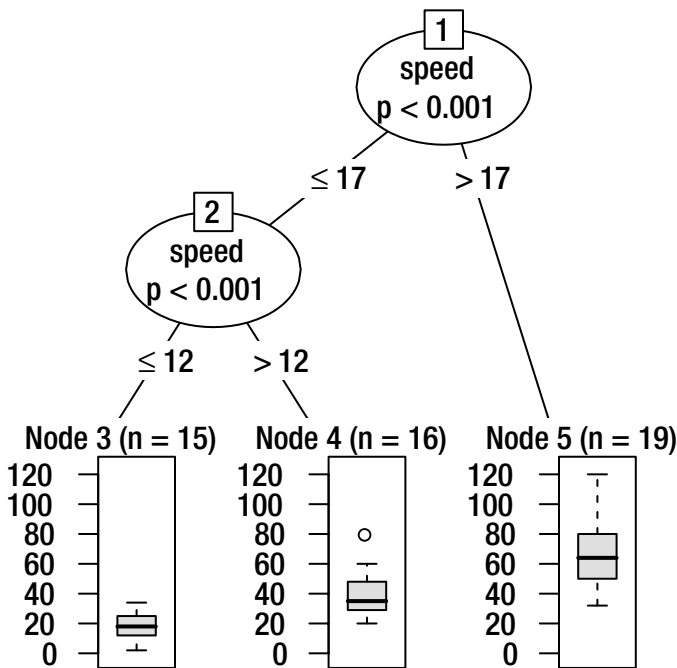
Figure 6-8. *Plot of the cars decision tree*

Random Forests

Random forests generalize decision trees by building several of them and combining them. They are implemented in the randomForest package, as follows:

```
library(randomForest)
model <- cars %>% randomForest(dist ~ speed, data = .)
rmse(predict(model, cars), cars$dist)
## [1] 83.5496
```

For classification, the predictions are the actual classes as a factor, so no translation is needed to get a confusion matrix:

```
model <- BreastCancer %>%
  randomForest(Class ~ Cl.thickness, data = .)

predict(model, BreastCancer) %>% head
##        1        2        3        4        5
##   benign   benign   benign malignant   benign
##        6
## malignant
## Levels: benign malignant
table(BreastCancer$Class, predict(model, BreastCancer))
##
##            benign malignant
```

```
##    benign     437        21
##    malignant   76       165
```

Neural Networks

You can use a package called `nnet` to construct neural networks.

```
library(nnet)
```

You can use it for both classification and regression. We can see it in action on the `cars` dataset:

```
model <- cars %>% nnet(dist ~ speed, data = ., size = 5)
## # weights: 16
## initial  value 122462.309952
## final  value 120655.000000
## converged
rmse(predict(model, cars), cars$dist)
## [1] 347.3543
```

The neural networks require a `size` parameter specifying how many nodes you want in the inner layer of the network. Here I have used five.

For classification, you use a similar call:

```
model <- BreastCancer %>%
  nnet(Class ~ Cl.thickness, data = ., size = 5)
## # weights: 56
## initial  value 718.232444
## iter  10 value 227.662158
## iter  20 value 225.222217
## iter  30 value 225.099474
## iter  40 value 225.098372
## final  value 225.098275
## converged
```

The output of the `predict()` function is probabilities for the tumor being malignant:

```
predict(model, BreastCancer) %>% head
##         [,1]
## 1 0.3461458
## 2 0.3461458
## 3 0.1111090
## 4 0.5294166
## 5 0.1499927
## 6 0.9130590
```

We need to translate it into classes and, for this, we can use a lambda expression:

```
predicted_class <- predict(model, BreastCancer) %>%
  { ifelse(. < 0.5, "benign", "malignant") }

table(BreastCancer$Class, predicted_class)
##           predicted_class
##            benign malignant
##   benign      437        21
##   malignant    76       165
```

Support Vector Machines

Another popular method is support vector machines. These are implemented in the ksvn() function in the kernlab package:

```
library(kernlab)
model <- cars %>% ksvm(dist ~ speed, data = .)
rmse(predict(model, cars), cars$dist)
## [1] 102.3646
```

For classification, the output is again a factor we can use directly to get a confusion matrix:

```
model <- BreastCancer %>%
  ksvm(Class ~ Cl.thickness, data = .)

predict(model, BreastCancer) %>% head
## [1] benign    benign    benign    malignant
## [5] benign    malignant
## Levels: benign malignant
table(BreastCancer$Class, predict(model, BreastCancer))
##
##             benign malignant
##   benign      437        21
##   malignant    76       165
```

Naive Bayes

Naive Bayes essentially assumes that each explanatory variable is independent of the others and uses the distribution of these for each category of data to construct the distribution of the response variable given the explanatory variables.

Naive Bayes is implemented in the e1071 package:

```
library(e1071)
```

The package doesn't support regression analysis—after all, it needs to look at conditional distributions for each output variable value—but we can use it for classification. The function we need is `naiveBayes()` and we can use the `predict()` output directly to get a confusion matrix:

```
model <- BreastCancer %>%
  naiveBayes(Class ~ Cl.thickness, data = .)

predict(model, BreastCancer) %>% head
## [1] benign    benign    benign    malignant
## [5] benign    malignant
## Levels: benign malignant
table(BreastCancer$Class, predict(model, BreastCancer))
##
##             benign malignant
##   benign       437        21
##   malignant     76       165
```

Exercises

Try the following exercises to become more comfortable with the concepts discussed in this chapter.

Fitting Polynomials

Use the `cars` data to fit higher degree polynomials and use training and test data to explore how they generalize. At which degree do you get the better generalization?

Evaluating Different Classification Measures

Earlier I wrote functions for computing the accuracy, specificity (true negative rate), and sensitivity (true positive rate) of a classification. Write similar functions for the other measures described previously. Combine them in a `prediction_summary()` function like I did earlier.

Breast Cancer Classification

You saw how to use the `glm()` function to predict the classes for the breast cancer data. Use it to make predictions for training and test data, randomly splitting the data in these two classes, and evaluate all the measures with your `predict_summary()` function.

If you can, try to make functions similar to the ones I used to split data and evaluate models for the `cars` data.

Leave-One-Out Cross-Validation (Slightly More Difficult)

The code I wrote previously splits the data into *n* groups and constructs training and test data based on that. This is called *n*-fold cross-validation. There is another common approach to cross-validation called *leave one out* cross-validation. The idea here is to remove a single data observation and use that for testing and all the rest of the data for training.

This isn't used that much if you have a lot of data—leaving out a single data point will not change the trained model much if you have lots of data points anyway—but for smaller datasets it can be useful.

Try to program a function for constructing subsampled training and test data for this strategy.

Decision Trees

Use the BreastCancer data to predict the tumour class, but try including more of the explanatory variables. Use cross-validation or sampling of training/test data to explore how it affects the prediction accuracy.

Random Forests

Use the BreastCancer data to predict the tumour class, but try including more of the explanatory variables. Use cross-validation or sampling of training/test data to explore how it affects the prediction accuracy.

Neural Networks

The size parameter for the nnet function specifies the complexity of the model. Test how the accuracy depends on this variable for classification on the BreastCancer data.

We only used the cell thickness variable to predict the tumour class. Include the other explanatory variables and explore if having more information improves the prediction power.

Support Vector Machines

Use the BreastCancer data to predict the tumour class, but try including more of the explanatory variables. Use cross-validation or sampling of training/test data to explore how it affects the prediction accuracy.

Compare Classification Algorithms

Compare the logistic regression, the neural networks, the decision trees, the random forests, and the support vector machines in how well they classify tumors in the BreastCancer data. For each, take the best model you obtained in your experiments.

■ ■ ■

Unsupervised Learning

For supervised learning, we have one or more targets we want to predict using a set of explanatory variables. But not all data analysis consists of making prediction models. Sometimes we are just trying to find out what structure is actually in the data we analyze. There can be several reasons for this. Sometimes unknown structures can tell us more about the data. Sometimes we want to explicitly *avoid* an unknown structure (if we have datasets that are supposed to be similar, we don't want to discover later that there are systematic differences). Whatever the reason, unsupervised learning concerns finding unknown structures in data.

Dimensionality Reduction

Dimensionality reduction, as the name hints at, are methods used when you have high-dimensional data and want to map it down into fewer dimensions. The purpose here is usually to visualize data to try and spot patterns from plots. The analysis usually just transform the data and doesn't add anything to it. It possibly removes some information, but by reducing the number of dimensions it can be easier to analyze.

The type of data where this is necessary is when the data has lots of columns. Not necessarily many observations, but each observation have very many variables, and there is often little information in any single column. One example is genetic data where there is often hundreds of thousands, if not millions, of genetic positions observed in each individual, and at each of these positions we have a count of how many of a given genetic variant is present at these markers, a number from 0 to 2. There is little information in any single marker, but combined they can be used to tell a lot about an individual. The first example we shall see in this chapter, principal component analysis, is frequently used to map thousands of genetic markers into a few more informative dimensions to reveal relationships between different individuals.

I will not use data with very high dimensionality but illustrate them with smaller datasets where the methods can still be useful.

Principal Component Analysis

Principal component analysis (PCA) maps your data from one vector space to another of the same dimensionality as the first. So it doesn't reduce the number of dimensions as such. However, it chooses the coordinate system of the new space such that the most information is in the first coordinate, the second most information in the second coordinate, and so on.

In it simplest form it is just a linear transformation. It changes the basis of your vector space such that the most variance in the data is along the first basis vector, and each basis vector then has increasingly less of the variance. The basis of the new vector space is called the components and the name "principal component" refers to looking at the first few, the most important, the *principal* components.

There might be some transformations of the data first to normalize it, but the final step of the transformation is always such a linear map. Hence after the transformation there is exactly the same amount of information in your data, it is just represented along different dimensions.

Because the PCA just transforms your data, your data has to be numerical vectors, to begin with. For categorical data, you will need to modify the data first. One approach is to represent factors as a binary vector for each level, as is done with model matrices in supervised learning. If you have a lot of factors in your data, though, PCA might not be the right tool.

It is beyond the scope of this book to cover the theory of PCA in any detail—but many other textbooks do—so we will only dig into how it is used in R.

To illustrate this, I will use the `iris` dataset. It is not high-dimensional, but it will do as a first example.

Remember that this data contains four measurements—sepal length and width and petal length and width—for flowers from three different species:

```
iris %>% head
##   Sepal.Length Sepal.Width Petal.Length
## 1          5.1         3.5          1.4
## 2          4.9         3.0          1.4
## 3          4.7         3.2          1.3
## 4          4.6         3.1          1.5
## 5          5.0         3.6          1.4
## 6          5.4         3.9          1.7
##   Petal.Width Species
## 1         0.2  setosa
## 2         0.2  setosa
## 3         0.2  setosa
## 4         0.2  setosa
## 5         0.2  setosa
## 6         0.4  setosa
```

To see if there is information in the data that would enable us to distinguish between the three species based on the measurements, we could try to plot some of the measurements against each other. See Figures 7-1 and 7-2.

```
iris %>% ggplot() +
  geom_point(aes(x = Sepal.Length, y = Sepal.Width, colour = Species))
```
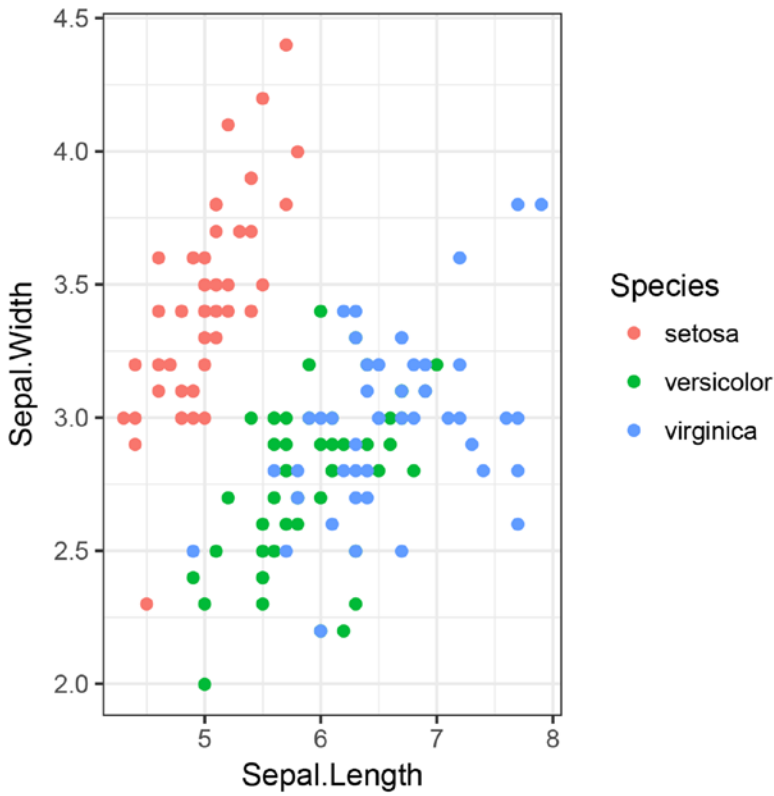
Figure 7-1. *Plot of iris sepal length versus sepal width*

```
iris %>% ggplot() +
  geom_point(aes(x = Petal.Length, y = Petal.Width, colour = Species))
```
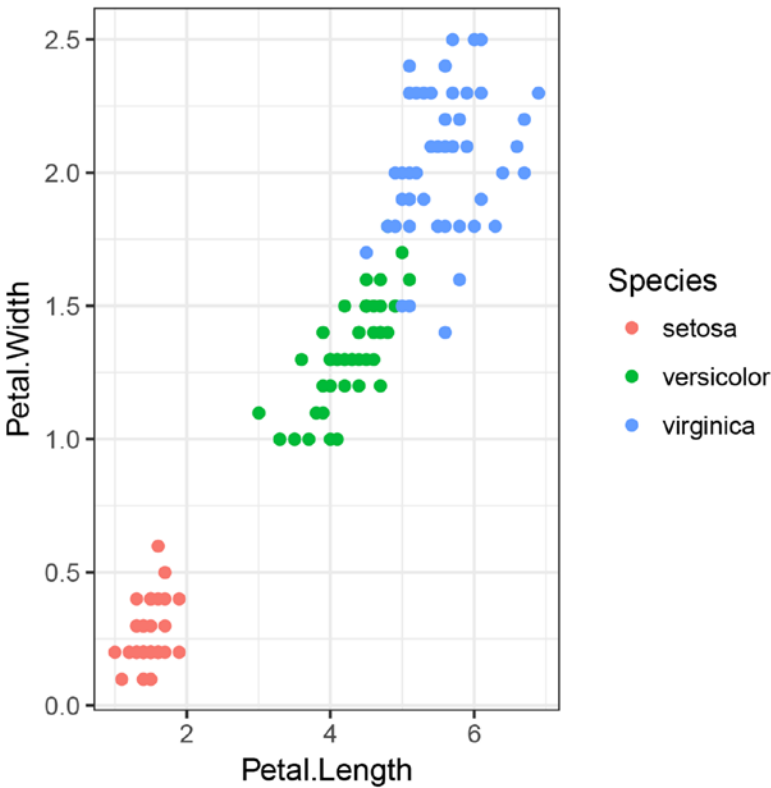
Figure 7-2. *Plot of iris petal length versus petal width*

It does look as if we should be able to distinguish the species. Setosa stands out on both plots, but Versicolor and Virginia overlap on the first.

Since this is such a simple dataset, and since there is obviously structure if we just plot a few dimensions against each other, this is not a case where we would usually pull out the cannon that is PCA, but this is a section on PCA so we will.

Since PCA only works on numerical data, we need to remove the Species parameter, but after that, we can do the transformation using the prcomp function:

```
pca <- iris %>% select(-Species) %>% prcomp
pca
## Standard deviations:
## [1] 2.0562689 0.4926162 0.2796596 0.1543862
##
## Rotation:
##                     PC1         PC2         PC3
## Sepal.Length   0.36138659 -0.65658877  0.58202985
## Sepal.Width   -0.08452251 -0.73016143 -0.59791083
## Petal.Length   0.85667061  0.17337266 -0.07623608
## Petal.Width    0.35828920  0.07548102 -0.54583143
```

```
##                      PC4
## Sepal.Length  0.3154872
## Sepal.Width  -0.3197231
## Petal.Length -0.4798390
## Petal.Width   0.7536574
```

The object that this produces contains different information about the result. The standard deviations tell us how much variance is in each component and the rotation what the linear transformation is. If we plot the `pca` object, we will see how much of the variance in the data is on each component, as shown in Figure 7-3.
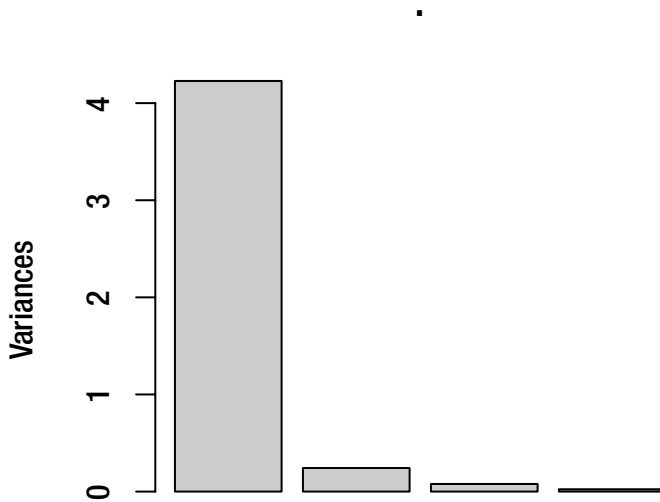
***Figure 7-3.** Plot of the variance on each principal component for the iris dataset*

```
pca %>% plot
```

The first thing you want to look at after making the transformation is how the variance is distributed along the components. If the first few components do not contain most of the variance, the transformation has done little for you. When it does, there is some hope that plotting the first few components will tell you about the data.

To map the data to the new space spanned by the principal components, we use the `predict()` function:

```
mapped_iris <- pca %>% predict(iris)
mapped_iris %>% head
##            PC1        PC2         PC3
## [1,] -2.684126 -0.3193972  0.02791483
## [2,] -2.714142  0.1770012  0.21046427
## [3,] -2.888991  0.1449494 -0.01790026
## [4,] -2.745343  0.3182990 -0.03155937
## [5,] -2.728717 -0.3267545 -0.09007924
## [6,] -2.280860 -0.7413304 -0.16867766
```

```
##                 PC4
## [1,]  0.002262437
## [2,]  0.099026550
## [3,]  0.019968390
## [4,] -0.075575817
## [5,] -0.061258593
## [6,] -0.024200858
```

This can also be used with new data that wasn't used to create the pca object. Here, we just give it the same data we used before. We don't actually have to remove the Species variable; it will figure out which of the columns to use based on their names. We can now plot the first two components against each other, as shown in Figure 7-4.

```
mapped_iris %>%
  as.data.frame %>%
  cbind(Species = iris$Species) %>%
  ggplot() +
  geom_point(aes(x = PC1, y = PC2, colour = Species))
```
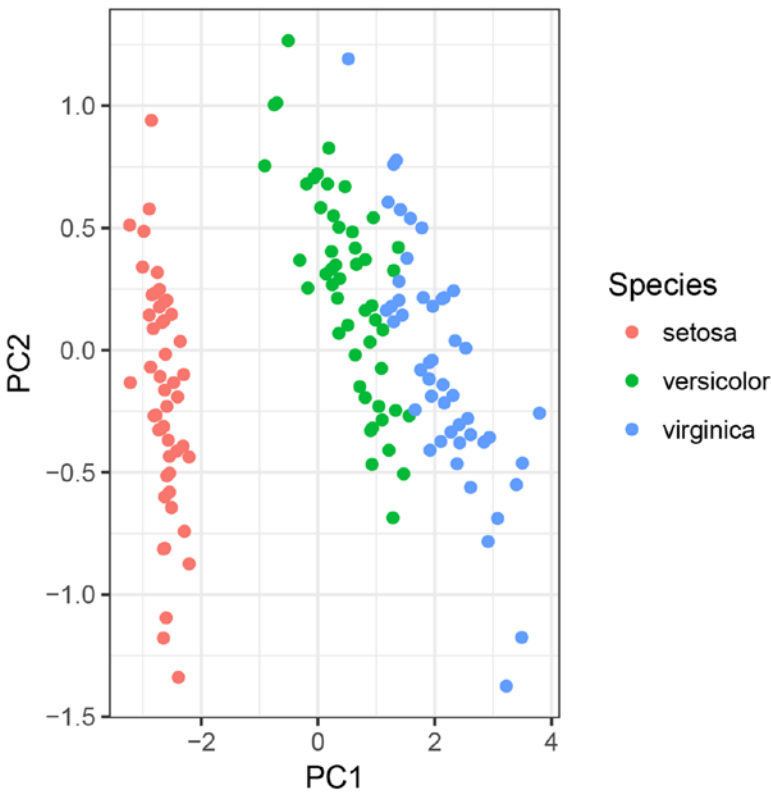

Figure 7-4. *Plot of first two principal components for the iris dataset*

The `mapped_iris` object returned from the `predict()` function is not a data frame but a matrix. That won't work with `ggplot()` so we need to transform it back into a data frame, and we do that with `as.data.frame`. Since we want to color the plot according to species, we need to add that information again—remember the `pca` object does not know about this factor data—so we do that with `cbind()`. After that, we plot.

We didn't gain much from this. There was about as much information in the original columns as there is in the transformed data. But now that we have seen PCA in action we can try it out on a little more interesting example.

We will look at the `HouseVotes84` data from the `mlbench` package:

```
library(mlbench)
data(HouseVotes84)
HouseVotes84 %>% head
##         Class   V1 V2 V3   V4    V5 V6 V7 V8 V9 V10
## 1 republican    n  y  n    y     y  y  n  n  n   y
## 2 republican    n  y  n    y     y  y  n  n  n   n
## 3   democrat <NA>  y  y <NA>     y  y  n  n  n   n
## 4   democrat    n  y  y    n  <NA>  y  n  n  n   n
## 5   democrat    y  y  y    n     y  y  n  n  n   n
## 6   democrat    n  y  y    n     y  y  n  n  n   n
##     V11  V12 V13 V14 V15  V16
## 1 <NA>    y   y   y   n    y
## 2    n    y   y   y   n <NA>
## 3    y    n   y   y   n    n
## 4    y    n   y   n   n    y
## 5    y <NA>   y   y   y    y
## 6    n    n   y   y   y    y
```

The data contains the votes cast for both republicans and democrats on 16 different proposals. The types of votes are yea, nay, and missing/unknown. Now, since votes are unlikely to be accidentally lost, missing data here means someone actively decided not to vote, so it isn't really missing. There is probably some information in that as well.

Now an interesting question we could ask is whether there are differences in voting patterns between republicans and democrats. We would expect that, but can we see it from the data?

The individual columns are binary (well, trinary if we consider the missing data as actually informative) and do not look very different between the two groups, so there is little information in each individual column. We can try doing a PCA on the data.

```
HouseVotes84 %>% select(-Class) %>% prcomp
## Error in colMeans(x, na.rm = TRUE): 'x' must be numeric
```

Okay, R is complaining that the data isn't numeric. We know that PCA needs numeric data, but we are giving it factors. We need to change that so we can try to map the votes into zeros and ones.

We can use the function `apply()`. This function is used to apply a function to a matrix and what it does depends on the dimensions we tell it to work on. It can summarize data along rows or along columns, but if we tell it to work on both dimensions, that is the `c(1,2)` argument to the function, it will apply the function to each element in the matrix. The transformation to do is just a function we give `apply()`. Here you can use a function defined elsewhere or an anonymous function. I used an anonymous function and instead of writing it as `function(x) {...}` I used a lambda expression:

```
HouseVotes84 %>%
  select(-Class) %>%
  apply(c(1,2), . %>% { ifelse(as.character(.) == "n", 0, 1) }) %>%
  prcomp
## Error in svd(x, nu = 0): infinite or missing values in 'x'
```

That doesn't work either, but now the problem is the missing data. We have mapped nay to 0 and yea to 1, but missing data remains missing.

We should always think carefully about how we deal with missing data, especially in a case like this where it might actually be informative. One approach we could take is to translate each column into three binary columns indicating if a vote was cast as yea, nay, or not cast.

I have left that as an exercise. Here I will just say that if someone abstained from voting, then they are equally likely to have voted yea or nay and translate missing data into 0.5.

Since I want to map the data onto the principal components afterward, and since I don't want to write the data transformations twice, I save it in a variable and then perform the PCA:

```
vote_patterns <- HouseVotes84 %>%
  select(-Class) %>%
  apply(c(1,2), . %>% { ifelse(as.character(.) == "n", 0, 1) }) %>%
  apply(c(1,2), . %>% { ifelse(is.na(.), 0.5, .) })

pca <- vote_patterns %>% prcomp
```

Now we can map the vote patterns onto the principal components and plot the first against the second, as shown in Figure 7-5.

```
mapped_votes <- pca %>% predict(vote_patterns)
mapped_votes %>%
  as.data.frame %>%
  cbind(Class = HouseVotes84$Class) %>%
  ggplot() +
  geom_point(aes(x = PC1, y = PC2, colour = Class))
```
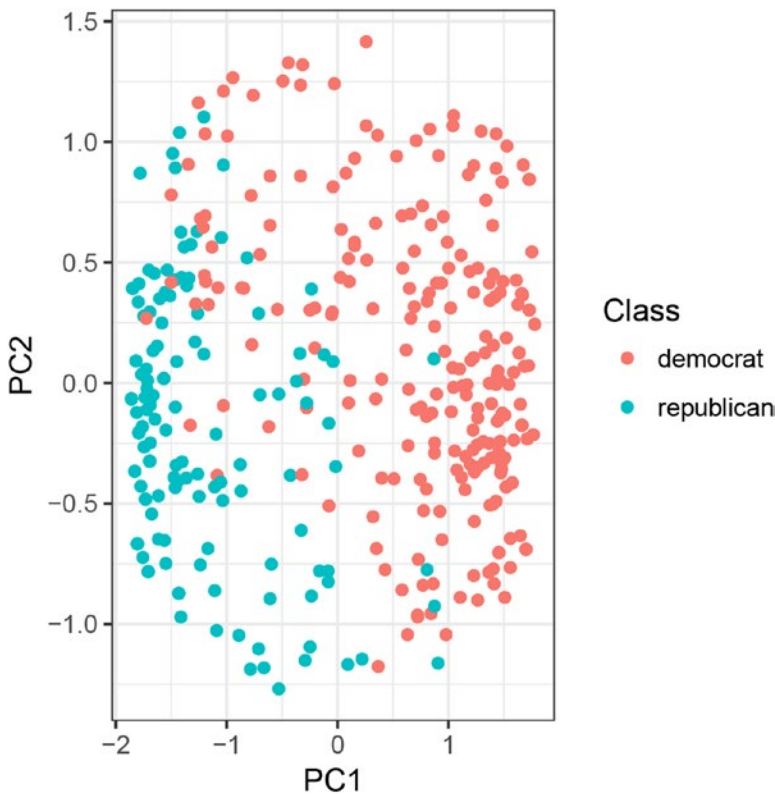
Figure 7-5. *Plot of first two principal components for the house votes dataset*

It looks like there is a clear separation in the voting patterns, at least on the first principal component. This is not something we could immediately see from the original data.

Multidimensional Scaling

Sometimes it is easier to have a measure of distance between objects than representing them as numerical vectors. Consider for example strings. You *could* translate them into numbers based on their encoding, but the space of possible strings is vast—infinite if you do not restrict their length—so it is not a practical approach. However, there are many measures of how different two strings are. For strings, at least, it is easier to define a distance measure than a mapping into numeric values.

When what we have is a distance measure we can represent our data as a distance matrix, one that contains all pair-wise distances. Obviously, this is not a feasible solution if you have very many data points—the number of pairs grows proportionally to the number of data points squared—but up to a few thousand data points, it is not a significant problem. Multidimensional scaling takes such a matrix of all pair-wise distances and maps each data point into a linear space while preserving the pair-wise distances as well as possible.

Consider the iris dataset again. For this dataset, of course, we do have the data points represented as numerical vectors, but it is a dataset we are familiar with so it is good to see the new method in use on it.

We can create a distance matrix using the dist() function:

```
iris_dist <- iris %>% select(-Species) %>% dist
```

To create a representation of these distances in a two-dimensional space we use the function cmdscale(). It takes a parameter, k, that specifies the dimensionality we want to place the points in. Give it a high enough k and it can perfectly preserve all pair-wise distances, but we wouldn't be able to visualize it. We are best served with low dimensionality and to plot the data we chose two. The result is a matrix with one row per original data point and one column per dimension we asked for; here, of course, two.

```
mds_iris <- iris_dist %>% cmdscale(k=2)
mds_iris %>% head
##            [,1]       [,2]
## [1,] -2.684126  0.3193972
## [2,] -2.714142 -0.1770012
## [3,] -2.888991 -0.1449494
## [4,] -2.745343 -0.3182990
## [5,] -2.728717  0.3267545
## [6,] -2.280860  0.7413304
```

We can translate this matrix into a data frame and plot it (see Figure 7-6).

```
mds_iris %>%
  as.data.frame %>%
  cbind(Species = iris$Species) %>%
  ggplot() +
  geom_point(aes(x = V1, y = V2, colour = Species))
```
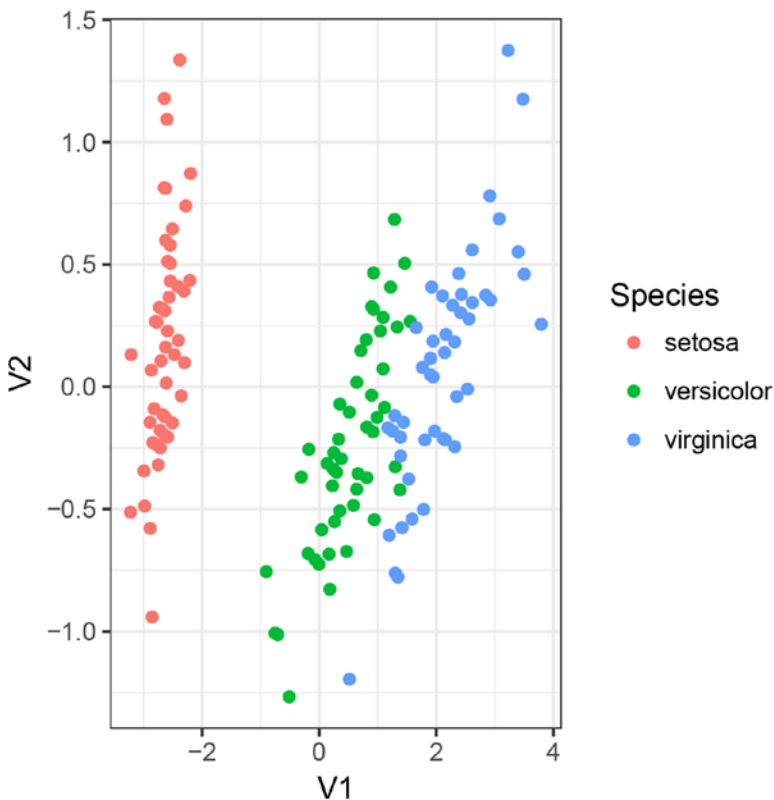
Figure 7-6. *Multidimensional scaling plot for iris data*

This expression uses names V1 and V2 for the x and y axes. This exploits that a data frame we have not provided column names for will name them V*n* where *n* is an increasing integer.

The plot looks essentially the same as the PCA plot earlier, which is not a coincidence, except that it is upside down.

We can do exactly the same thing with the voting data—here we can reuse the cleaned data that has translated the factors into numbers—and the result is shown in Figure 7-7.

```
mds_votes <- vote_patterns %>% dist %>% cmdscale(k = 2)

mds_votes %>%
  as.data.frame %>%
  cbind(Class = HouseVotes84$Class) %>%
  ggplot() +
  geom_point(aes(x = V1, y = V2, colour = Class))
```
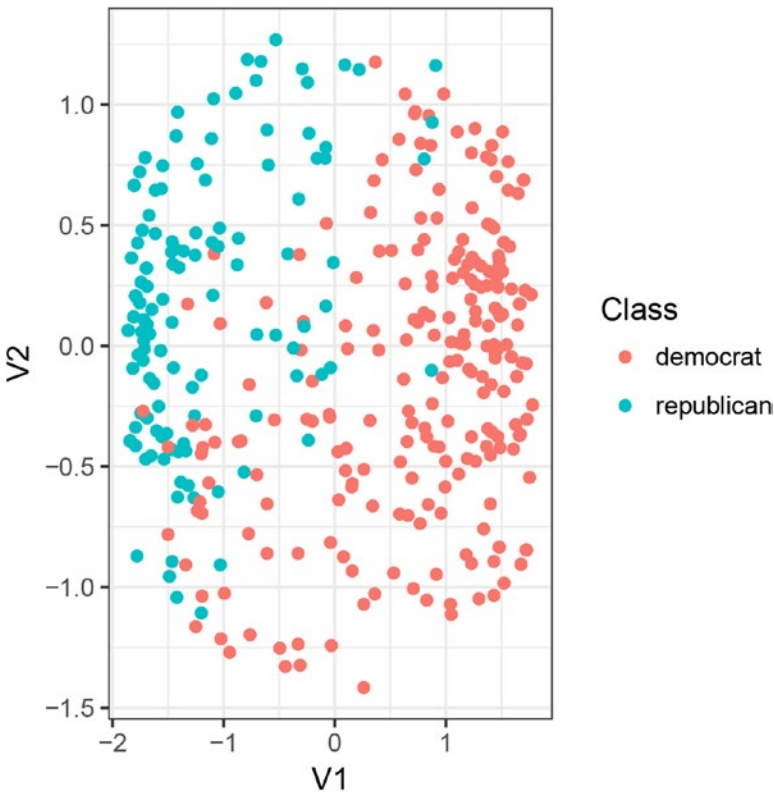
Figure 7-7. *Multidimensional scaling plot for house voting data*

Should you ever have the need for computing a distance matrix between strings, by the way, you might want to look at the stringdist package. As an example illustrating this, we can simulate some strings. The following code first has a function for simulating random strings over the letters "A", "C", "G", and "T" and the second function then adds a random length to that. We then create 10 strings using these functions:

```
random_ngram <- function(n)
  sample(c('A','C','G','T'), size = n, replace = TRUE) %>%
  paste0(collapse = "")

random_string <- function(m) {
  n <- max(1, m + sample(c(-1,1), size = 1) * rgeom(1, 1/2))
  random_ngram(n)
}

strings <- replicate(10, random_string(5))
```

Using the stringdist package, we can compute the all-pairs distance matrix:

```
library(stringdist)
string_dist <- stringdistmatrix(strings)
```

We can now plot the strings in two-dimensional space, roughly preserving their distances, as shown in Figure 7-8.

```
string_dist %>%
  cmdscale(k = 2) %>%
  as.data.frame %>%
  cbind(String = strings) %>%
  ggplot(aes(x = V1, y = V2)) +
  geom_point() +
  geom_label(aes(label = String),
             hjust = 0, nudge_y = -0.1)
```
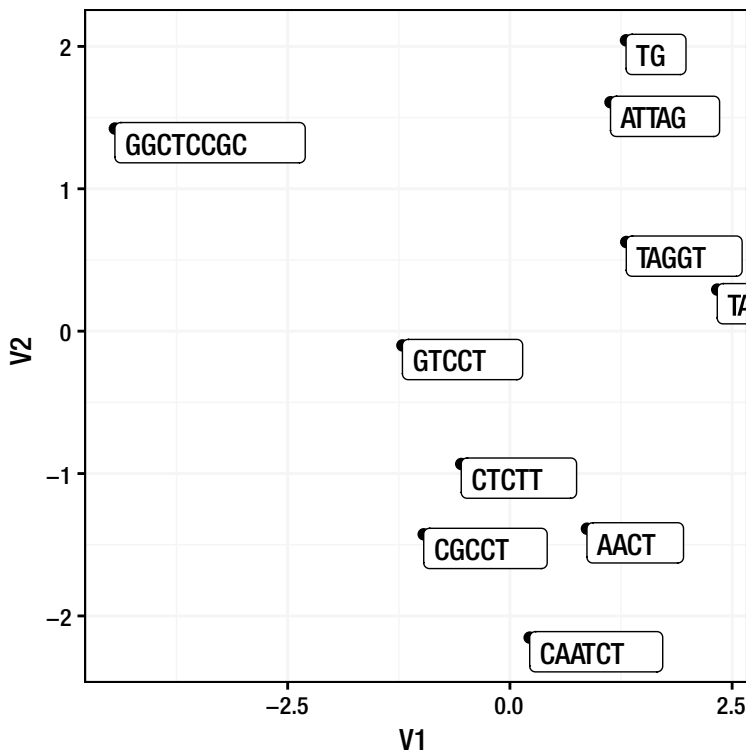

Figure 7-8. *Multidimensionality reduction for random strings*

Clustering

Clustering methods seek to find similarities between data points and group data according to these similarities. Such clusters can either have a hierarchical structure or not; when the structure is hierarchical, each data point will be associated with several clusters, ordered from the more specific to the more general, and when the structure is not hierarchical any data point is typically only assigned a single cluster. The next sections describe two of the most popular clustering algorithms, one of each kind of clustering.

k-Means Clustering

In *k*-means clustering you attempt to separate the data into *k* clusters, where you determine the number *k*. The data usually has to be in the form of numeric vectors. Strictly speaking, the method will work as long as you have a way of computing the mean of a set of data points and the distance between pairs of data points. The R function for *k*-means clustering, kmeans, wants numerical data.

The algorithm essentially works by first guessing at *k* "centers" of proposed clusters. Then each data point is assigned to the center it is closest to, creating a grouping of the data, and then all centers are moved to the mean position of their clusters. This is repeated until an equilibrium is reached. Because the initial centers are randomly chosen, different calls to the function will not necessarily lead to the same result. At the very least, expect the labeling of clusters to be different between the various calls.

Let's see it in action. We use the iris dataset, and we remove the Species column to get a numerical matrix to give to the function:

```
clusters <- iris %>%
  select(-Species) %>%
  kmeans(centers = 3)
```

We need to specify *k*, the number of centers in the parameters to kmeans(), and we choose three. We know that there are three species, so this is a natural choice. Life isn't always that simple, but here it is the obvious choice.

The function returns an object with information about the clustering. The two most interesting pieces of information are the centers, the variable centers (excuse the difference in spelling here, it is a UK versus US thing), and the cluster assignment, the variable cluster.

Let's look at the center:

```
clusters$centers
##   Sepal.Length Sepal.Width Petal.Length
## 1     5.006000    3.428000     1.462000
## 2     6.850000    3.073684     5.742105
## 3     5.901613    2.748387     4.393548
##   Petal.Width
## 1    0.246000
## 2    2.071053
## 3    1.433871
```

These are simply vectors of the same form as the input data points. They are the center of mass for each of the three clusters we have computed.

The cluster assignment is simply an integer vector with a number for each data point specifying which cluster that data point is assigned to:

```
clusters$cluster %>% head
## [1] 1 1 1 1 1 1
clusters$cluster %>% table
## .
##  1  2  3
## 50 38 62
```

There are 50 data points for each species so if the clustering perfectly matched the species we should see 50 points for each cluster as well. The clustering is not perfect, but we can try plotting the data and see how well the clustering matches the species class.

We can first plot how many data points from each species are assigned to each cluster (see Figure 7-9).

```
iris %>%
  cbind(Cluster = clusters$cluster) %>%
  ggplot() +
  geom_bar(aes(x = Species, fill = as.factor(Cluster)),
           position = "dodge") +
  scale_fill_discrete("Cluster")
```
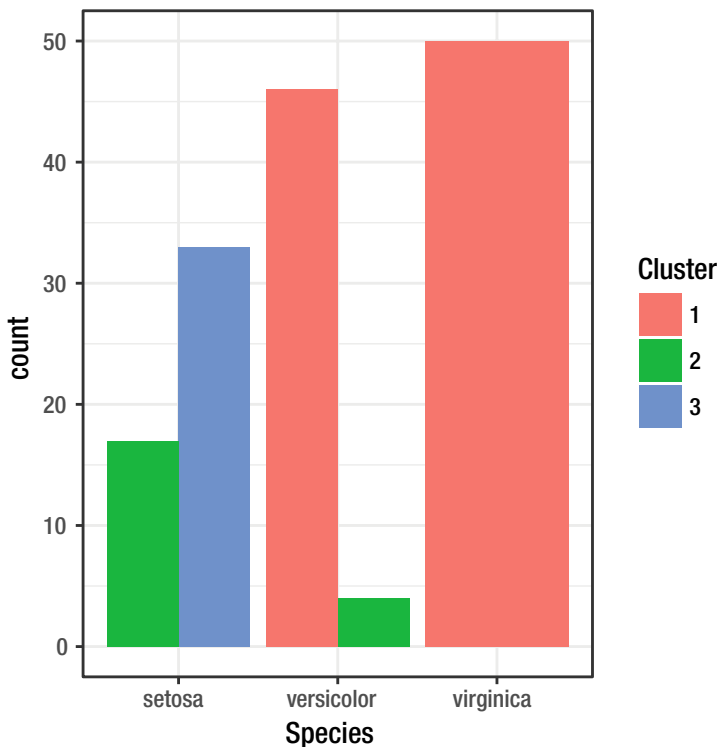

Figure 7-9. *Cluster assignments for the three iris species*

We first combine the `iris` dataset with the cluster association from `clusters` and then make a bar plot. The `position` argument is `"dodge"` so the cluster assignments are plotted next to each other instead of stacked on top of each other.

Not unexpectedly, from what we have learned of the data by plotting it earlier, Setosa seems clearly distinct from the other two species, which, according to the four measurements we have available at least, overlap in features.

There is a bit of luck involved here as well. A different starting point for where `kmeans()` placed the first centers will affect the final result, and had it put two clusters inside the cloud of the Setosa data points it would have split those points into two clusters and merged the Versicolor and Virginia points into a single cluster, for instance.

It is always a good idea to visually examine how the clustering result matches where the actual data points fall. We can do this by plotting the individual data points and see how the classification and clustering looks. We could plot the points for any pair of features, but we have seen how to map the data onto principal components, so we could try to plot the data on the first two of these. As you remember, we can map data points from the four features to the principal components using the predict() function. This works both for the original data used to make the PCA, as well as the centers we get from the *k*-means clustering:

```
pca <- iris %>%
  select(-Species) %>%
  prcomp

mapped_iris <- pca %>%
  predict(iris)

mapped_centers <- pca %>%
  predict(clusters$centers)
```

We can plot the mapped data points, PC1, against PC2 (see Figure 7-10). To display the principal components together with the species information, we need to add a Species column. We also need to add the cluster information since that isn't included in the mapped vectors. This is a numeric vector, but we want to treat it as categorical, so we need to translate it using as.factor():

```
mapped_iris %>%
  as.data.frame %>%
  cbind(Species = iris$Species,
        Clusters = as.factor(clusters$cluster)) %>%
  ggplot() +
  geom_point(aes(x = PC1, y = PC2,
                 colour = Species, shape = Clusters)) +
  geom_point(aes(x = PC1, y = PC2),
             size = 5, shape = "X",
             data = as.data.frame(mapped_centers))
```
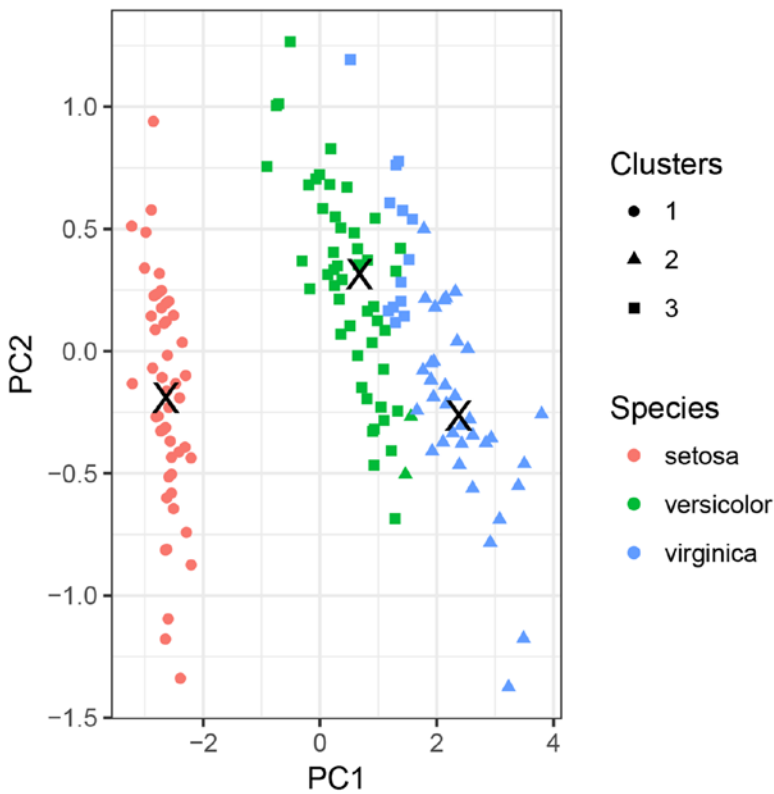
Figure 7-10. *Clusters and species for iris*

In the plot, I also show the centers. I use the `data` argument to `geom_point()` to give it this data, and I set the size to 5 and set the shape to `"X"`.

As mentioned, there is some luck involved in getting a good clustering like this. The result of a second run of the `kmeans()` function is shown in Figures 7-11 and 7-12.
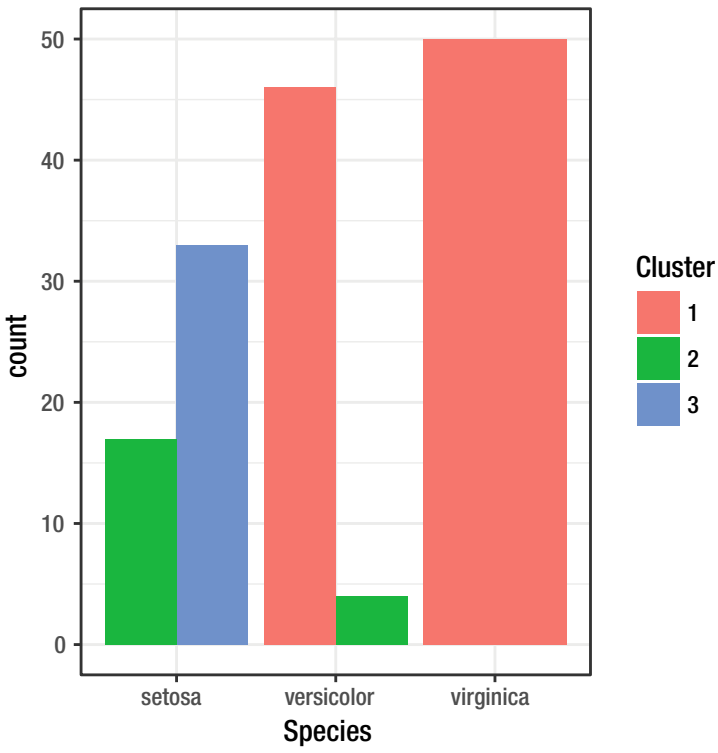
Figure 7-11. *A bad cluster assignment for the three iris species*
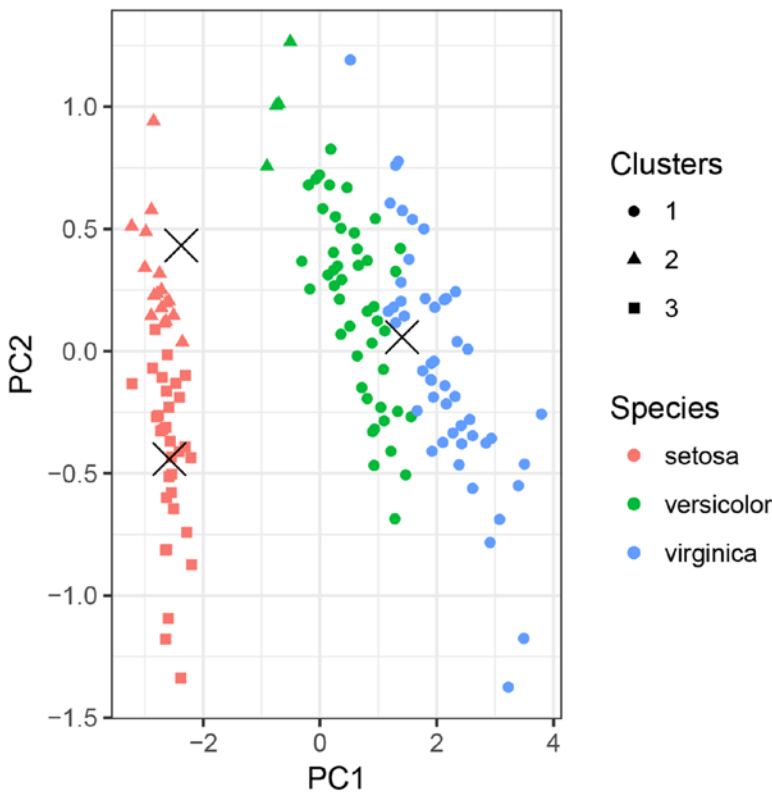
Figure 7-12. *Clusters and species for iris for a bad clustering*

If you go back and look at Figure 7-10 and think that some of the square points are closer to the center of the "triangular cluster" than the center of the "square cluster", or vice versa, you are right. Don't be too disturbed by this; two things are deceptive here. One is that the axes are not on the same scale, so distances along the x-axis are farther than distances along the y-axis. A second is that the distances used to group data points are in the four-dimensional space of the original features, while the plot is a projection onto the two-dimensional plane of the first two principal components.

There *is* something to worry about, though, concerning distances. The algorithm is based on the distance from cluster centers to data points, but if you have one axis in centimeters and another in meters, a distance along one axis is numerically a hundred times farther than along the other. This is not merely solved by representing all features in the same unit. First of all, that isn't always possible. There is no meaningful way of translating time or weight into a distance. Even if it was, what is being measured is also relevant for the unit we consider. The height of a person is meaningfully measured in meters, but you do not want something like cell size to be measured in meters.

This is also an issue for principal component analysis. Obviously, a method that tries to create a vector space basis based on the variance in the data is going to be affected by the units used in the input data. The usual solution is to rescale all input features so they are centered at zero and have variance one. You subtract from each data point the mean of the feature and divide by the standard deviation. This means that measured in standard deviations, all dimensions have the same variation.

The `prcomp()` function takes parameters to do the scaling. Parameter `center`, which defaults to `TRUE`, translates the data points to mean zero, and parameter `scale.` (notice the `.`), which defaults to `FALSE`, scales the data points to have variance one at all dimensions.

The kmeans() functions do not take these parameters, but you can explicitly rescale a numerical data frame using the scale() function. I have left this as an exercise.

Now let's consider how the clustering does at predicting the species more formally. This returns us to familiar territory: We can build a confusion matrix between species and clusters.

```
table(iris$Species, clusters$cluster)
##
##                 1  2  3
##   setosa       50  0  0
##   versicolor    0  2 48
##   virginica     0 36 14
```

One problem here is that the clustering doesn't know about the species, so even if there were a one-to-one corresponding between clusters and species, the confusion matrix would only be diagonal if the clusters and species were in the same order.

We can associate each species to the cluster most of its members are assigned to. This isn't a perfect solution—two species could be assigned to the same cluster this way, and we still wouldn't be able to construct a confusion matrix—but it will work for us in the case we consider here. We can count how many observations from each cluster is seen in each species like this:

```
tbl <- table(iris$Species, clusters$cluster)
(counts <- apply(tbl, 1, which.max))
##     setosa versicolor  virginica
##          1          3          2
```

Build a table mapping species of clusters to get the confusion matrix like this:

```
map <- rep(NA, each = 3)
map[counts] <- names(counts)
table(iris$Species, map[clusters$cluster])
##
##               setosa versicolor virginica
##   setosa          50          0         0
##   versicolor       0         48         2
##   virginica        0         14        36
```

A final word on k-means is this: Since k is a parameter that needs to be specified, how do you pick it? Here we knew that there were three species so we picked three for k as well. But when we don't know if there is any clustering in the data, to begin with, or if there is a lot, how do we choose k? Unfortunately, there isn't a general answer to this. There are several rules of thumbs, but there is no perfect solution you can always apply.

Hierarchical Clustering

Hierarchical clustering is a technique you can use when you have a distance matrix of your data. Here the idea is that you build up a tree structure of nested clusters by iteratively merging clusters. You start with putting each data point in their own singleton clusters. Then iteratively you find two clusters that are close together and merge them into a new cluster. You continue this until all data points are in the same large cluster. Different algorithms exist, and they mainly vary in how they choose which cluster to merge next

and how they compute the distance between clusters. In R the function `hclust()` implements several algorithms—the parameter `method` determines which is used—and we can see it in use with the `iris` dataset. We first need a distance matrix. This time, I first scale the data.

```
iris_dist <- iris %>% select(-Species) %>% scale %>% dist
```

Now the clustering is constructed by calling `hclust()` on the distance matrix.

```
clustering <- hclust(iris_dist)
```

We can plot the result using the generic `plot()` function, as shown in Figure 7-13. There is not much control over how the clustering is displayed using this function, but if you are interested in plotting trees, you should look at the ape package.

```
plot(clustering)
```
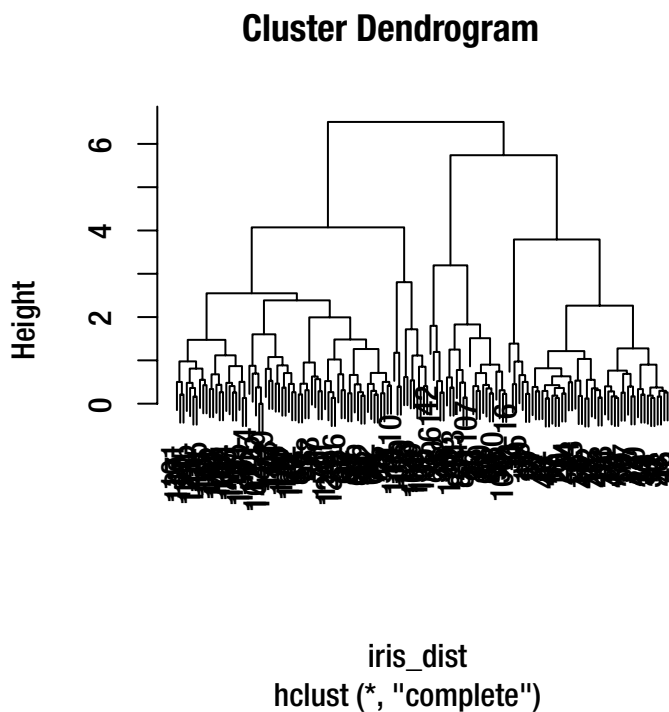

Figure 7-13. *Hierarchical clustering of iris data*

To create plots that work well with `ggplot2` graphics, you want the `ggdendro` package (see Figure 7-14).

```
library(ggdendro)
ggdendrogram(clustering) + theme_dendro()
```
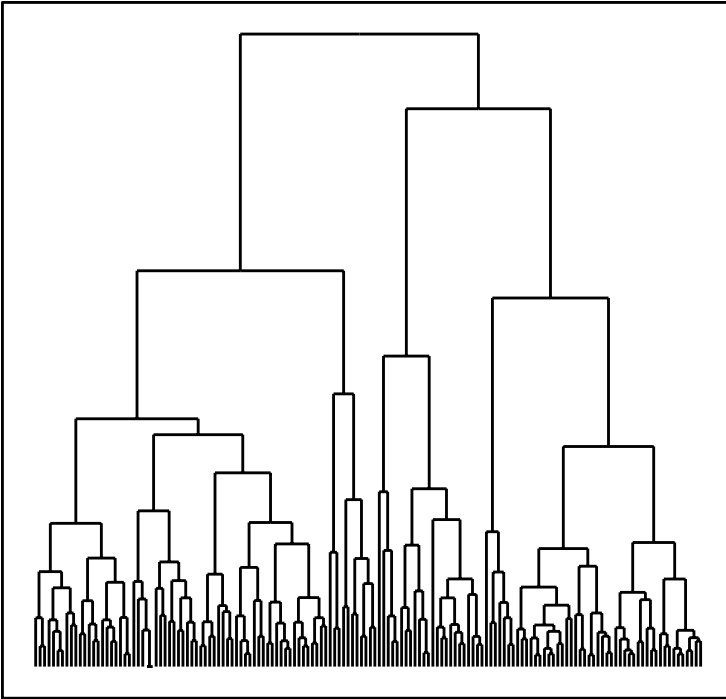

Figure 7-14. *Hierarchical clustering of iris data plotted with ggdendro*

Using `ggdendro`, you can get access to the raw plotting segments, which gives you control over much of the visualization of the tree.

Only visualizing the clustering is rarely enough, so to work with the result we need to be able to extract the actual groupings. The `cutree()` function—it stands for *cut tree* but there is only one *t*—lets you do this. You can give it a parameter `h` to cut the tree into clusters by splitting the tree at height `h`, or you can give it parameter `k` to cut the tree at the level where there is exactly `k` clusters.

Since we are working with the `iris` data, it is natural to want to split the data into three clusters:

```
clusters <- clustering %>% cutree(k = 3)
```

The result is in the same format as we had for *k*-means clustering, i.e., a vector with integers specifying which cluster each data point belongs to. Since we have the information in the familiar format, we can try plotting the clustering information as a barplot (see Figure 7-15):

```
iris %>%
  cbind(Cluster = clusters) %>%
  ggplot() +
  geom_bar(aes(x = Species, fill = as.factor(Cluster)),
           position = "dodge") +
  scale_fill_discrete("Cluster")
```
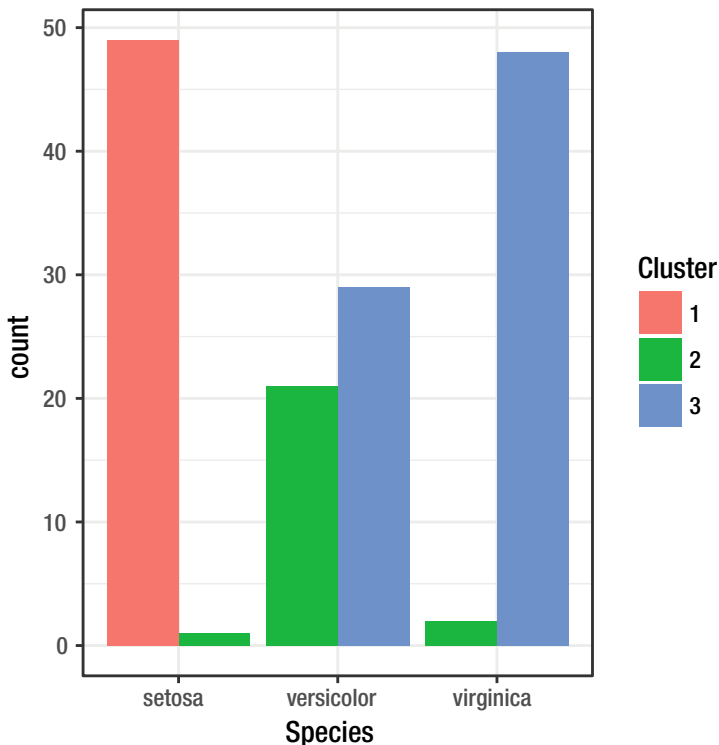

Figure 7-15. *Iris clustering as a barplot*

Or we can plot the individual plots together with species and cluster information (see Figure 7-16).

```
mapped_iris %>%
  as.data.frame %>%
  cbind(Species = iris$Species,
        Clusters = as.factor(clusters)) %>%
  ggplot() +
  geom_point(aes(x = PC1, y = PC2,
                 shape = Species, colour = Clusters))
```
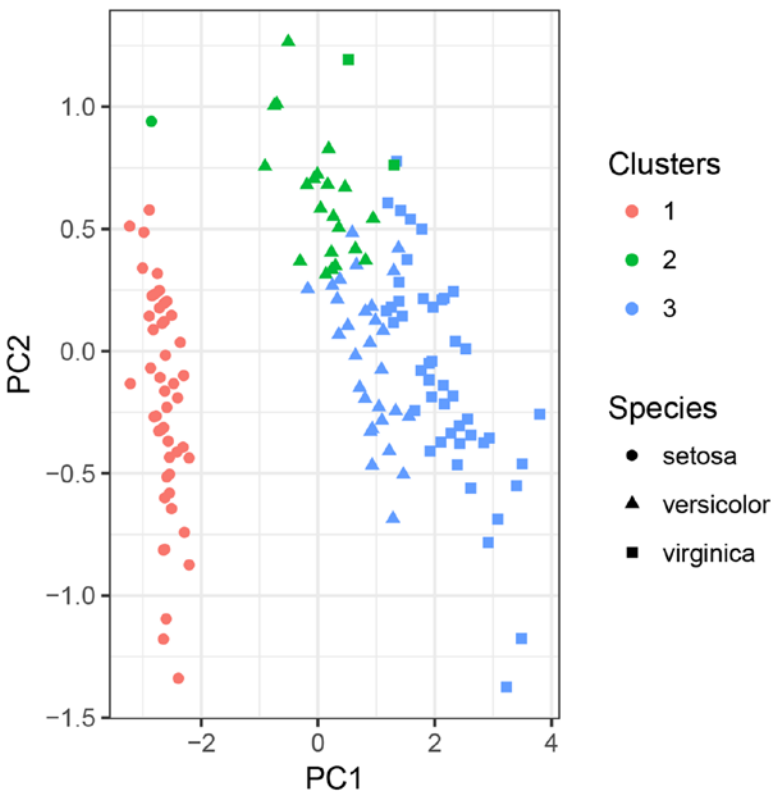
Figure 7-16. *Iris points plotted with species and hierarchical clustering information*

Constructing a confusion matrix if we want to use the clustering for a form of classification is of course done similarly, but hierarchical clustering lends itself much less to classification than *k*-means clustering does. With *k*-means clustering it is simple to take a new data point and see which cluster center it is nearest. With hierarchical clustering, you would need to rebuild the entire tree to see where it falls.

Association Rules

The last unsupervised learning method we will see is aimed at categorical data, ordered or unordered. Just like you have to translate factors into numerical data to use methods such as PCA, you will need to translate numerical data into factors to use association rules. This typically isn't a problem, and you can use the function cut() to split a numerical vector into a factor and combine it with ordered() if you want it ordered.

Association rules searches for patterns in your data by picking out subsets of the data, *X* and *Y*, based on predicates on the input variables and evaluate rules $X \Rightarrow Y$. Picking *X* and *Y* is a brute force choice (which is why you need to break the numerical vectors into discrete classes).[1]

---

[1]The algorithm *could* do it for you by considering each point between two input values, but it doesn't, so you have to break the data.

Any statement $X \Rightarrow Y$ is called a *rule* and the algorithm evaluates all rules (at least up to a certain size) to figure out how good each rule is.

The association rules algorithm is implemented in the `arules` package:

```
library(arules)
```

To see it in action, we use the `income` dataset from the `kernlab` package:

```
library(kernlab)
data(income)
income %>% head
##            INCOME SEX MARITAL.STATUS   AGE
## 1        [75.000-  F        Married 45-54
## 2        [75.000-  M        Married 45-54
## 3        [75.000-  F        Married 25-34
## 4        -10.000)  F         Single 14-17
## 5        -10.000)  F         Single 14-17
## 6 [50.000-75.000)  M        Married 55-64
##               EDUCATION              OCCUPATION
## 1 1 to 3 years of college             Homemaker
## 2        College graduate             Homemaker
## 3        College graduate Professional/Managerial
## 4           Grades 9 to 11  Student, HS or College
## 5           Grades 9 to 11  Student, HS or College
## 6 1 to 3 years of college               Retired
##        AREA DUAL.INCOMES HOUSEHOLD.SIZE UNDER18
## 1 10+ years           No          Three    None
## 2 10+ years           No           Five     Two
## 3 10+ years          Yes          Three     One
## 4 10+ years  Not Married           Four     Two
## 5 4-6 years  Not Married           Four     Two
## 6 10+ years           No            Two    None
##   HOUSEHOLDER HOME.TYPE ETHNIC.CLASS LANGUAGE
## 1        Own     House        White     <NA>
## 2        Own     House        White  English
## 3       Rent Apartment        White  English
## 4     Family     House        White  English
## 5     Family     House        White  English
## 6        Own     House        White  English
```

This data contains income information together with several explanatory variables and is already in a form the `arules` can deal with: all columns are factorial.

The same data is actually also available in the `arules` package as the `Income` dataset, but here it is representing in a different format than a data frame so we will use this version of the data.

```
data(Income)
Income %>% head
## transactions in sparse format with
##  6 transactions (rows) and
##  50 items (columns)
```

To construct the rules, we use the `apriori()` function. It takes various arguments for controlling which rules the function will return, but we can use it with all default parameters:

```
rules <- income %>% apriori
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval
##         0.8    0.1     1 none FALSE
##  originalSupport maxtime support minlen maxlen
##             TRUE       5     0.1      1     10
##  target   ext
##   rules FALSE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 899
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[84 item(s), 8993 transaction(s)] done [0.00s].
## sorting and recoding items ... [42 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 5 6 done [0.02s].
## writing ... [785 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

The `rules` object we create this way is not a simple object like a data frame, but it will let us take the head() of it and we can use the function inspect() to see the individual rules:

```
rules %>% head %>% inspect(linebreak = FALSE)
##     lhs
## [1] {}                                =>
## [2] {EDUCATION=Grad Study}            =>
## [3] {OCCUPATION=Clerical/Service Worker} =>
## [4] {INCOME=[30.000-40.000)}          =>
## [5] {UNDER18=Two}                     =>
## [6] {INCOME=[50.000-75.000)}          =>
##     rhs               support   confidence
## [1] {LANGUAGE=English} 0.8666741 0.8666741
## [2] {LANGUAGE=English} 0.1000778 0.9316770
## [3] {LANGUAGE=English} 0.1046369 0.8860640
## [4] {LANGUAGE=English} 0.1111976 0.9009009
## [5] {LANGUAGE=English} 0.1073057 0.8405923
## [6] {LANGUAGE=English} 0.1329923 0.9143731
##     lift
## [1] 1.0000000
## [2] 1.0750027
## [3] 1.0223728
## [4] 1.0394921
## [5] 0.9699059
## [6] 1.0550368
```

194

The `linebreak = FALSE` here splits the rules over several lines. I find it confusing that to break the lines you have to set `linebreak` to FALSE, but that is how it is.

Each rule has a right side, `rhs`, and a left side, `lhs`. For a rule $X \Rightarrow Y$, X is the `rhs` and Y the `lhs`. The quality of a rule is measured by the following three columns:

- `support`: The fraction of the data where both X and Y holds true. Think of it as $\Pr(X,Y)$.

- `confidence`: The fraction of times where X is true that Y is also true. Think of it as $\Pr(Y|X)$.

- `lift`: How much better than random is the rule, in the sense that how much better is it compared to X and Y being independent. Think $\Pr(X,Y)/\Pr(X)\Pr(Y)$.

Good rules should have high enough support to be interesting—if a rule only affects a tiny number of data points out of the whole data it isn't that important—so you want both support and confidence to be high. It should also tell you more than what you would expect by random chance, which is captured by lift.

You can use the `sort()` function to rearrange the data according to the quality measures:

```
rules %>% sort(by = "lift") %>%
  head %>% inspect(linebreak = FALSE)
##     lhs
## [1] {MARITAL.STATUS=Married,OCCUPATION=Professional/Managerial,LANGUAGE=English}
## [2] {MARITAL.STATUS=Married,OCCUPATION=Professional/Managerial}
## [3] {DUAL.INCOMES=No,HOUSEHOLDER=Own}
## [4] {AREA=10+ years,DUAL.INCOMES=Yes,HOME.TYPE=House}
## [5] {DUAL.INCOMES=Yes,HOUSEHOLDER=Own,HOME.TYPE=House,LANGUAGE=English}
## [6] {DUAL.INCOMES=Yes,HOUSEHOLDER=Own,HOME.TYPE=House}
##        rhs                     support
## [1] => {DUAL.INCOMES=Yes}       0.1091960
## [2] => {DUAL.INCOMES=Yes}       0.1176471
## [3] => {MARITAL.STATUS=Married} 0.1016346
## [4] => {MARITAL.STATUS=Married} 0.1003002
## [5] => {MARITAL.STATUS=Married} 0.1098632
## [6] => {MARITAL.STATUS=Married} 0.1209830
##     confidence lift
## [1] 0.8069022  3.281986
## [2] 0.8033409  3.267501
## [3] 0.9713071  2.619965
## [4] 0.9605964  2.591075
## [5] 0.9601555  2.589886
## [6] 0.9594356  2.587944
```

You can combine this with the `subset()` function to filter the rules:

```
rules %>% subset(support > 0.5) %>% sort(by = "lift") %>%
  head %>% inspect(linebreak = FALSE)
##     lhs
## [1] {ETHNIC.CLASS=White}       =>
## [2] {AREA=10+ years}           =>
## [3] {UNDER18=None}             =>
## [4] {}                         =>
## [5] {DUAL.INCOMES=Not Married} =>
```

```
##      rhs                 support    confidence
## [1] {LANGUAGE=English} 0.6110308 0.9456204
## [2] {LANGUAGE=English} 0.5098410 0.8847935
## [3] {LANGUAGE=English} 0.5609919 0.8813767
## [4] {LANGUAGE=English} 0.8666741 0.8666741
## [5] {LANGUAGE=English} 0.5207384 0.8611622
##      lift
## [1] 1.0910911
## [2] 1.0209069
## [3] 1.0169644
## [4] 1.0000000
## [5] 0.9936402
```

Exercises

Try the following exercises to become more comfortable with the concepts discussed in this chapter.

Dealing with Missing Data in the HouseVotes84 Data

In the PCA analysis, we translated missing data into 0.5. This was to move things along but probably not an appropriate decision. People who do not cast a vote are not necessarily undecided and therefore equally likely to vote yea or nay; there can be conflicts of interests or other reasons. So we should instead translate each column into three binary columns.

You can use the `transmute()` function from `dplyr` to add new columns and remove old ones—it is a bit of typing since you have to do it 16 times, but it will get the job done.

If you feel more like trying to code your way out of this transformation, you should look at the `mutate_at()` function from `dplyr`. You can combine it with column name matches and multiple functions to build the three binary vectors (for the `ifelse()` calls you have to remember that comparing with NA always gives you NA so you need always to check for that first). After you have created the new columns, you can remove the old ones using `select()` combined with `match()`.

Try to do the transformation and then the PCA again. Does anything change?

Rescaling for k-Means Clustering

Use the `scale()` function to rescale the `iris` dataset, then redo the *k*-means clustering analysis.

Varying k

Analyze the `iris` data with `kmeans()` with *k* ranging from 1 to 10. Plot the clusters for each *k*, coloring the data points according to the clustering.

Project 1

To see a data analysis in action, I use an analysis that my student, Dan Søndergaard, did the first year I held the data science class. I am redoing his analysis here with his permission.

The data contains physicochemical features measured from Portuguese Vinho Verde wines, and the goal was to try to predict wine quality from these measurements. The data is available from https://archive.ics.uci.edu/ml/datasets/Wine+Quality.

Importing Data

If we go to the data folder, we can see that the data is split into three files. The measurements from red wine, white wine, and a description of the data (the file `winequality.names`). To avoid showing large URLs, I will not list the code for reading the files, but it is in this form:

```
read.table(URL, header=TRUE, sep=';')
```

That there is a header that describes the columns, and that fields are separated by semicolons we get from looking at the files.

We load the red and white wine data into separate data frames called `red` and `white`.

We can combine the two data frames using this:

```
wines <- rbind(data.frame(type = "red", red),
               data.frame(type = "white", white))
```

Then we'll see the summary:

```
summary(wines)
##     type      fixed.acidity    volatile.acidity
## red  :1599   Min.   : 3.800   Min.   :0.0800
## white:4898   1st Qu.: 6.400   1st Qu.:0.2300
##              Median : 7.000   Median :0.2900
##              Mean   : 7.215   Mean   :0.3397
##              3rd Qu.: 7.700   3rd Qu.:0.4000
##              Max.   :15.900   Max.   :1.5800
##   citric.acid     residual.sugar
## Min.   :0.0000   Min.   : 0.600
## 1st Qu.:0.2500   1st Qu.: 1.800
## Median :0.3100   Median : 3.000
## Mean   :0.3186   Mean   : 5.443
## 3rd Qu.:0.3900   3rd Qu.: 8.100
## Max.   :1.6600   Max.   :65.800
##   chlorides       free.sulfur.dioxide
## Min.   :0.00900   Min.   :  1.00
## 1st Qu.:0.03800   1st Qu.: 17.00
## Median :0.04700   Median : 29.00
## Mean   :0.05603   Mean   : 30.53
## 3rd Qu.:0.06500   3rd Qu.: 41.00
## Max.   :0.61100   Max.   :289.00
## total.sulfur.dioxide    density
## Min.   :  6.0         Min.   :0.9871
## 1st Qu.: 77.0         1st Qu.:0.9923
## Median :118.0         Median :0.9949
## Mean   :115.7         Mean   :0.9947
## 3rd Qu.:156.0         3rd Qu.:0.9970
## Max.   :440.0         Max.   :1.0390
```

```
##        pH             sulfates         alcohol
## Min.   :2.720   Min.   :0.2200   Min.   : 8.00
## 1st Qu.:3.110   1st Qu.:0.4300   1st Qu.: 9.50
## Median :3.210   Median :0.5100   Median :10.30
## Mean   :3.219   Mean   :0.5313   Mean   :10.49
## 3rd Qu.:3.320   3rd Qu.:0.6000   3rd Qu.:11.30
## Max.   :4.010   Max.   :2.0000   Max.   :14.90
##    quality
## Min.   :3.000
## 1st Qu.:5.000
## Median :6.000
## Mean   :5.818
## 3rd Qu.:6.000
## Max.   :9.000
```

There are 11 measurements for each wine, and each wine has an associated quality score based on sensory data. At least three wine experts judged and scored the wine on a scale between 0 and 10. No wine achieved a score below 3 or above 9. There are no missing values. There is not really any measurement that we want to translate into categorical data. The quality scores are given as discrete values, but they are ordered categories, and we might as well consider them as numerical values for now.

Exploring the Data

With the data loaded, we first want to do some exploratory analysis to get a feeling for it.

Distribution of Quality Scores

The first thing Dan did was look at the distribution of quality scores for both types of wine, as shown in Figure 7-17.

```
ggplot(wines) +
  geom_bar(aes(x = factor(quality), fill = type),
           position = 'dodge') +
  xlab('Quality') + ylab('Frequency')
```
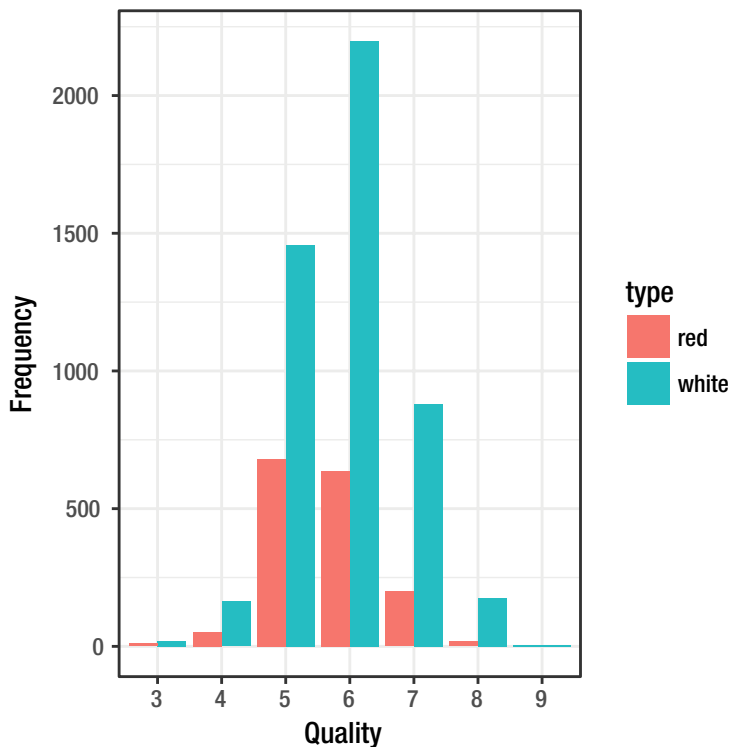
Figure 7-17. *Distribution of wine qualities*

There are very few wines with extremely low or high scores. The quality scores also seem normal-distributed, if we ignore that they are discrete. This might make the analysis easier.

Is This Wine Red or White?

The dataset has two types of wine: red and white. As Dan noticed, these types are typically described by very different words by wine experts, but several experiments have shown that even the best wine experts cannot distinguish red from white if the color is obscured or the experts blindfolded (see http://io9.com/wine-tasting-is-bullshit-heres-why-496098276). It is, therefore, interesting to see if the physicochemical features available in the data can help decide whether a wine is red or white.

Dan used the Naive Bayes method to explore this, so we need the e1071 package.

```
library(e1071)
```

He used a five-fold cross-validation to study this, but I will just use the partition() function from Chapter 6.

```
random_group <- function(n, probs) {
  probs <- probs / sum(probs)
  g <- findInterval(seq(0, 1, length = n), c(0, cumsum(probs)),
                    rightmost.closed = TRUE)
  names(probs)[sample(g)]
}
```

```
partition <- function(df, n, probs) {
  replicate(n, split(df, random_group(nrow(df), probs)), FALSE)
}
```

I will use a variation of the prediction accuracy function we wrote there for cars but using wines and the accuracy() function instead of rmse():

```
accuracy <- function(confusion_matrix)
  sum(diag(confusion_matrix))/sum(confusion_matrix)

prediction_accuracy_wines <- function(test_and_training) {
  result <- vector(mode = "numeric",
                   length = length(test_and_training))
  for (i in seq_along(test_and_training)) {
    training <- test_and_training[[i]]$training
    test <- test_and_training[[i]]$test
    model <- training %>% naiveBayes(type ~ ., data = .)
    predictions <- test %>% predict(model, newdata = .)
    targets <- test$type
    confusion_matrix <- table(targets, predictions)
    result[i] <- accuracy(confusion_matrix)
  }
  result
}
```

We get the following accuracy when we split the data randomly into training and test data 50/50:

```
random_wines <- wines %>%
    partition(4, c(training = 0.5, test = 0.5))
random_wines %>% prediction_accuracy_wines
## [1] 0.9747615 0.9726070 0.9756848 0.9729147
```

This is a pretty good accuracy, so this raises the question of why experts cannot tell red and white wine apart.

Dan looked into this by determining the most significant features that divide red and white wines by building a decision tree:

```
library('party')
tree <- ctree(type ~ ., data = wines,
              control = ctree_control(minsplit = 4420))
```

The plot of the tree is too large for me to show here in the book with the size limit for figures, but try to plot it yourself.

He limited the number of splits made to get only the most important features. From the tree, we see that the total amount of sulfur dioxide, a chemical compound often added to wines to prevent oxidation and bacterial activity, which may ruin the wine, is chosen as the root split.

Sulfur dioxide is also naturally present in wine in moderate amounts. In the EU the quantity of sulfur dioxide is restricted to 160 ppm for red wine and 210 ppm for white wines, so by law, we actually expect a significant difference of sulfur dioxide in the two types of wine. So he looked into that:

```
wines %>%
  group_by(type) %>%
  summarise(total.mean = mean(total.sulfur.dioxide),
            total.sd = sd(total.sulfur.dioxide),
            free.mean = mean(free.sulfur.dioxide),
            free.sd = sd(free.sulfur.dioxide))
## # A tibble: 2 × 5
##     type total.mean total.sd free.mean  free.sd
##   <fctr>      <dbl>    <dbl>     <dbl>    <dbl>
## 1    red   46.46779 32.89532  15.87492 10.46016
## 2  white  138.36066 42.49806  35.30808 17.00714
```

The average amount of total sulfur dioxide is indeed lower in red wines, and thus it makes sense that this feature is picked as a significant feature in the tree. If the amount of total sulfur dioxide in a wine is less than or equal to 67 ppm, we can say that it is a red wine with high certainty, which also fits with the summary statistics.

Another significant feature suggested by the tree is the volatile acidity, also known as the *vinegar taint*. In finished (bottled) wine a high volatile acidity is often caused by malicious bacterial activity, which can be limited by the use of sulfur dioxide, as described earlier. Therefore we expect a strong relationship between these features (see Figure 7-18).

```
qplot(total.sulfur.dioxide, volatile.acidity, data=wines,
      color = type,
      xlab = 'Total sulfur dioxide',
      ylab = 'Volatile acidity (VA)')
```
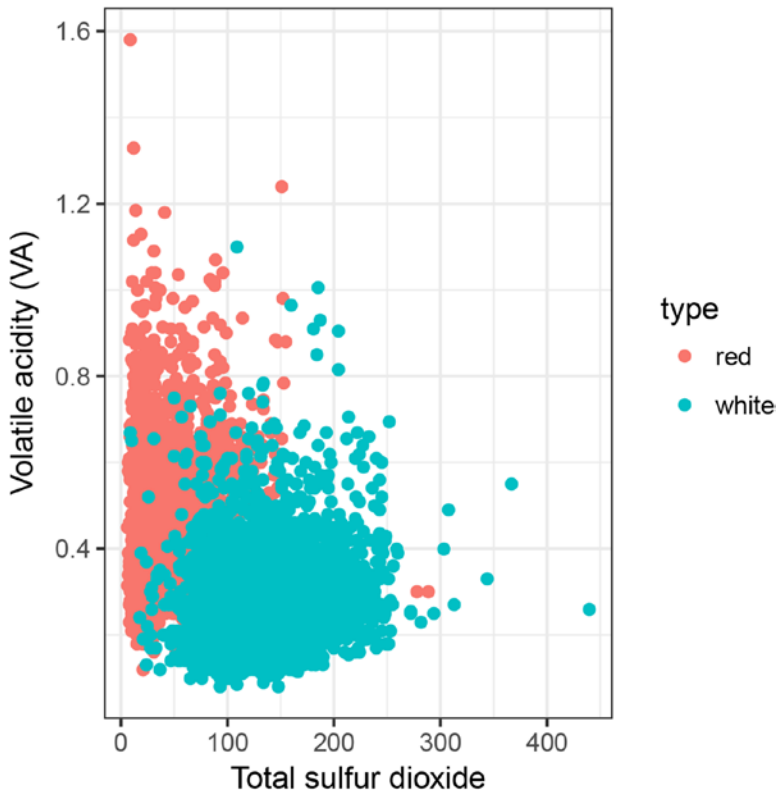
Figure 7-18. *Sulfur dioxide versus volatile acidity*

The plot shows that the amount of volatile acidity as a function of the amount of sulfur dioxide. It also shows that, especially for red wines, the volatile acidity is low for wines with a high amount of sulfur dioxide. The pattern for white wine is not as clear. However, Dan observed, as you can clearly see in the plot, a clear difference between red and white wines when considering the `total.sulfur.dioxide` and `volatile. acidity` features together.

So why can humans not taste the difference between red and white wines? It turns out that sulfur dioxide cannot be detected by humans in free concentrations of less than 50 ppm. Although the difference in total sulfur dioxide is very significant between the two types of wine, the free amount is on average below the detection threshold, and thus humans cannot use it to distinguish between red and white.

```
wines %>%
  group_by(type) %>%
  summarise(mean = mean(volatile.acidity),
            sd = sd(volatile.acidity))
## # A tibble: 2 × 3
##     type      mean         sd
##   <fctr>     <dbl>      <dbl>
## 1    red 0.5278205 0.1790597
## 2  white 0.2782411 0.1007945
```

Similarly, acetic acid (which causes volatile acidity) has a detection threshold of 0.7 g/L, and again we see that the average amount is below this threshold and thus is undetectable by the human taste buds.

So Dan concluded that some of the most significant features which we have found to tell the types apart only appear in small concentrations in wine that cannot be tasted by humans.

Fitting Models

Regardless of whether we can tell red wine and white wine apart, the real question we want to explore is whether the measurements will let us predict quality. Some of the measures might be below human tasting ability, but the quality is based on human tasters, so can we predict the quality based on the measurements?

Before we build a model, though, we need something to compare the accuracy against that can be our null-model. If we are not doing better than a simplistic model, then the model construction is not worth it.

Of course, first, we need to decide whether we want to predict the precise quality as categories or whether we consider it a regression problem. Dan looked at both options, but since we should mostly look at the quality as a number, I will only look at the latter.

For regression, the quality measure should be the root mean square error and the simplest model we can think of is just to predict the mean quality for all wines.

```
rmse <- function(x,t) sqrt(mean(sum((t - x)^2)))

null_prediction <- function(df) {
    rep(mean(wines$quality), each = nrow(df))
}

rmse(null_prediction(wines), wines$quality)
## [1] 70.38242
```

This is what we have to beat to have any model worth considering.

We do want to compare models with training and test datasets, though, so not use the mean for the entire data. So we need a function for comparing the results with split data.

To compare different models using `rmse()` as the quality measure we need to modify our prediction accuracy function. We can give it as parameter the function used to create a model that works with predictions. It could look like this:

```
prediction_accuracy_wines <- function(test_and_training,
                                      model_function) {
  result <- vector(mode = "numeric",
                   length = length(test_and_training))
  for (i in seq_along(test_and_training)) {
    training <- test_and_training[[i]]$training
    test <- test_and_training[[i]]$test
    model <- training %>% model_function(quality ~ ., data = .)
    predictions <- test %>% predict(model, newdata = .)
    targets <- test$quality
    result[i] <- rmse(predictions, targets)
  }
  result
}
```

Here we are hardwiring the formula to include all variables except for `quality` which is potentially leading to overfitting, but we are not worried about that right now.

To get this to work we need a `model_function()` that returns an object that works with `predict()`. To get this to work, we need to use generic functions, something we will not cover until Chapter 10, but it mostly involves creating a "class" and defining what `predict()` will do on objects of that class.

```
null_model <- function(formula, data) {
    structure(list(mean = mean(data$quality)),
            class = "null_model")
}

predict.null_model <- function(model, newdata) {
    rep(model$mean, each = nrow(newdata))
}
```

This `null_model()` function creates an object of class `null_model` and defines what the `predict()` function should do on objects of this class. We can use it to test how well the null model will perform on data:

```
test_and_training <- wines %>%
    partition(4, c(training = 0.5, test = 0.5))
test_and_training %>% prediction_accuracy_wines(null_model)
## [1] 49.77236 50.16679 50.11079 49.59682
```

Don't be too confused about these numbers being much better than the one we get if we use the entire dataset. That is simply because the `rmse()` function will always give a larger value if there is more data and we are giving it only half the data that we did when we looked at the entire dataset.

We can instead compare it with a simple linear model:

```
test_and_training %>% prediction_accuracy_wines(lm)
## [1] 42.30591 41.96099 41.72510 41.61227
```

Dan also tried different models for testing the prediction accuracy, but I have left that as an exercise.

Exercises

Try the following exercises to become more comfortable with the concepts discussed in this chapter.

Exploring Other Formulas

The `prediction_accuracy_wines()` function is hardwired to use the formula `quality ~ .` that uses all explanatory variables. Using all variables can lead to over-fitting so it is possible that using fewer variables can give better results on the test data. Add a parameter to the function for the formula and explore using different formulas.

Exploring Different Models

Try using different models than the null model and the linear model. Any model that can do regression and defines a `predict()` function should be applicable. Try it out.

Analyzing Your Own Dataset

Find a dataset you are interested in investigating and go for it. To learn how to interpret data, you must use your intuition on what is worth exploring and the only way to build that intuition is to analyze data.

■ ■ ■

More R Programming

In this chapter, we leave data analysis and return to programming and software development, topics that are the focus of the remaining chapters of the book. Chapter 1 gave a tutorial introduction to R programming but left out a lot of details. This chapter covers many of those details while the next two chapters will cover more advanced aspects of R programming: functional programming and object oriented programming.

Expressions

We begin the chapter by going back to expressions. Everything we do in R involves evaluating expressions. Most expressions we evaluate to do a computation and get the result, but some expressions have side-effects—like assignments—and those we usually evaluate because of the side-effects.

Arithmetic Expressions

We saw the arithmetic expressions in Chapter 1, so we will just give a very short reminder here. The arithmetic expressions are operators that involve numbers and consist of the unary operators + and -:

```
+ x
- x
```

where + doesn't really do anything, while - changes the sign of its operand. Then there are the infix operators for addition, subtraction, multiplication, and division:

```
x + y
x - y
x * y
x / y
```

Division will return a floating-point number even if both its operands are integers, so if you want to do integer division, you need the special operator for that:

```
x %/% y
```

If you want the remainder of integer division, you need this infix operator instead:

```
x %% y
```

Finally, there are operators for exponentiation. To compute x^y, you can use either of these two operators:

```
x ^ y
x ** y
```

In all these examples, x and y can be numbers or variables referring to numbers (actually, vectors of numbers since R always works on vectors), or they can be other expressions evaluating to numbers. If you compose expressions from infix operators, you have the same precedence rules you know from arithmetic. Exponentiation goes before multiplication that goes before addition, for example. This means that you will need to use parentheses if you need to evaluate the expressions in another order.

Since the rules are the ones you are used to, this is not likely to cause you troubles, except if you combine these expressions with the operator :. This isn't really an arithmetic operator but it *is* an infix operator for generating sequences, and it has a higher precedence than multiplication but lower than exponentiation. This means that 1:2**2 will evaluate the 2**2 expression first to get 1:4 and then construct the sequence:

```
1:2**2
## [1] 1 2 3 4
```

The expression 1:2*2 will evaluate the : expression first to create a vector containing 1 and 2 and then multiply this vector with 2:

```
1:2*2
## [1] 2 4
```

Since the unary - operator has higher precedence than : it also means that -1:2 will give you the sequence from -1 to 2 and not the sequence containing -1 and -2. For that, you need parentheses:

```
-1:2
## [1] -1  0  1  2
-(1:2)
## [1] -1 -2
```

Functions are evaluated before the operators:

```
1:sqrt(4)
## [1] 1 2
```

Boolean Expressions

For Boolean values—those that are either TRUE or FALSE—you also have logical operators. The operator ! negates a value:

```
!TRUE
## [1] FALSE
!FALSE
## [1] TRUE
```

The | and || are logical "or" operators while & and && are logical "and" operators. The difference between | and || or & and && are how they deal with vectors. The one-character version will apply the operator element-wise and create a vector while the two-character version will only look at the first value in vectors.

```
TRUE | FALSE
## [1] TRUE
FALSE | FALSE
## [1] FALSE
TRUE || FALSE
## [1] TRUE
FALSE || FALSE
## [1] FALSE
x <- c(TRUE, FALSE, TRUE, FALSE)
y <- c(TRUE, TRUE, FALSE, FALSE)

x | y
## [1]  TRUE  TRUE  TRUE FALSE
x || y
## [1] TRUE
x & y
## [1]  TRUE FALSE FALSE FALSE
x && y
## [1] TRUE
```

We typically use the two-character version in control structures like if—since these do not operate on vectors in any case—while we use the one-character version when we need to compute with Boolean arithmetic, when we want our expressions to work as vectorized expressions.

Incidentally, all the arithmetic operators work like the | and & operators when operating on more than one value, i.e., they operate element-wise on vectors. We saw that in Chapter 1 when we talked about vector expressions.

Basic Data Types

There are a few basic types in R: numeric, integer, complex, logical, and character.

The Numeric Type

The numeric type is what you get any time you write a number into R. You can test if an object is numeric using the is.numeric function or by getting the class object.

```
is.numeric(2)
## [1] TRUE
class(2)
## [1] "numeric"
```

The Integer Type

The integer type is used for, well, integers. Surprisingly, the 2 is *not* an integer in R. It is a numeric type which is the larger type that contains all floating-point numbers as well as integers. To get an integer you have to make the value explicitly an integer, and you can do that using the function as.integer or writing L after the literal.

```
is.integer(2)
## [1] FALSE
is.integer(2L)
## [1] TRUE
x <- as.integer(2)
is.integer(x)
## [1] TRUE
class(x)
## [1] "integer"
```

If you translate a non-integer into an integer, you just get the integer part.

```
as.integer(3.2)
## [1] 3
as.integer(9.9)
## [1] 9
```

The Complex Type

If you ever find that you need to work with complex numbers, R has those as well. You construct them by adding an imaginary number—a number followed by i—to any number or explicitly using the function as.complex. The imaginary number can be zero, 0i, which creates a complex number that only has a non-zero real part.

```
1 + 0i
## [1] 1+0i
is.complex(1 + 0i)
## [1] TRUE
class(1 + 0i)
## [1] "complex"
sqrt(as.complex(-1))
## [1] 0+1i
```

The Logical Type

Logical values are what you get if you explicitly type in TRUE or FALSE, but it is also what you get if you make, for example, a comparison.

```
x <- 5 > 4
x
## [1] TRUE
class(x)
## [1] "logical"
```

```
is.logical(x)
## [1] TRUE
```

The Character Type

Finally, characters are what you get when you type in a string such as "hello, world".

```
x <- "hello, world"
class(x)
## [1] "character"
is.character(x)
## [1] TRUE
```

Unlike in some languages, character doesn't mean a single character but any text. So it is not like in C or Java where you have single character types, 'c', and multi-character strings, "string", they are both just characters.

You can, similar to the other types, explicitly convert a value into a character (string) using as. character:

```
as.character(3.14)
## [1] "3.14"
```

I will not go further into string handling in R here. There are of course lots of functions for manipulating strings—and even though there are all those functions I still find it a lot harder to manipulate strings in R than in scripting languages such as Python—but those are beyond the scope of this book.

Data Structures

From the basic types, you can construct other data structures, essentially by concatenating simpler types into more complex ones. The basic building blocks here are vectors, which are sequences of values all of the same type, and lists, which are sequences where the values can have different types.

Vectors

We have already seen vectors many times in this book, so you should be familiar with them. Whenever we have seen expressions involving single numbers, we have actually seen vectors containing a single value, so we have never seen anything that *wasn't* a vector. But we now consider more technical aspects of vectors.

What I have called vectors up till now is technically known as "atomic sequences". Those are any sequences of the basic types described previously. You create these by concatenating basic values using the c function.

```
v <- c(1, 2, 3)
```

or through some other operator or function, e.g., the : operator or the rep function

```
1:3
## [1] 1 2 3
rep("foo", 3)
## [1] "foo" "foo" "foo"
```

We can test if something is this kind of vector using the is.atomic function:

```
v <- 1:3
is.atomic(v)
## [1] TRUE
```

The reason I mention that "atomic sequences" is the technically correct term for what we have called vectors until now is that there is also something in R that is *explicitly* called a vector. In practice, there is no confusion because all the atomic sequences I have called vectors are also vectors.

```
v <- 1:3
is.vector(v)
## [1] TRUE
```

It is just that R only consider such a sequence a vector—in the sense that is.vector returns TRUE—if the object doesn't have any attributes (except for one, names, which it is allowed to have).

Attributes are meta-information associated with an object, and not something we will deal with much here, but you just have to know that is.vector will be FALSE if something that is a perfectly good vector gets an attribute.

```
v <- 1:3
is.vector(v)
## [1] TRUE
attr(v, "foo") <- "bar"
v
## [1] 1 2 3
## attr(,"foo")
## [1] "bar"
is.vector(v)
## [1] FALSE
```

So if you want to test if something is the kind of vector I am talking about here, use is.atomic instead.

When you concatenate (atomic) vectors, you always get another vector back. So when you combine several c() calls you don't get any kind of tree structure if you do something like this:

```
c(1, 2, c(3, 4), c(5, 6, 7))
## [1] 1 2 3 4 5 6 7
```

The type might change, if you try to concatenate vectors of different types, R will try to translate the type into the most general type of the vectors.

```
c(1, 2, 3, "foo")
## [1] "1"   "2"   "3"   "foo"
```

Matrix

If you want a matrix instead of a vector, what you really want is just a two-dimensional vector. You can set the dimensions of a vector using the dim function—it sets one of those attributes we talked about previously—where you specify the number of rows and the number of columns you want the matrix to have.

```
v <- 1:6
attributes(v)
## NULL
dim(v) <- c(2, 3)
attributes(v)
## $dim
## [1] 2 3
dim(v)
## [1] 2 3
v
##      [,1] [,2] [,3]
## [1,]    1    3    5
## [2,]    2    4    6
```

When you do this, the values in the vector will go in the matrix column-wise, i.e., the values in the vector will go down the first column first and then on to the next column and so forth.

You can use the convenience function `matrix` to create matrices and there you can specify if you want the values to go by column or by row using the `byrow` parameter.

```
v <- 1:6
matrix(data = v, nrow = 2, ncol = 3, byrow = FALSE)
##      [,1] [,2] [,3]
## [1,]    1    3    5
## [2,]    2    4    6
matrix(data = v, nrow = 2, ncol = 3, byrow = TRUE)
##      [,1] [,2] [,3]
## [1,]    1    2    3
## [2,]    4    5    6
```

Once you have a matrix, there is a lot of support for doing linear algebra in R, but there are a few things you need to know. First, the * operator will not do matrix multiplication. You use * if you want to make element-wise multiplication; for matrix multiplication you need the operator %*% instead.

```
(A <- matrix(1:4, nrow = 2))
##      [,1] [,2]
## [1,]    1    3
## [2,]    2    4
(B <- matrix(5:8, nrow = 2))
##      [,1] [,2]
## [1,]    5    7
## [2,]    6    8
A * B
##      [,1] [,2]
## [1,]    5   21
## [2,]   12   32
A %*% B
##      [,1] [,2]
## [1,]   23   31
## [2,]   34   46
```

211

If you want to transpose a matrix, you use the t function and, if you want to invert it, you use the solve function.

```
t(A)
##      [,1] [,2]
## [1,]    1    2
## [2,]    3    4
solve(A)
##      [,1] [,2]
## [1,]   -2  1.5
## [2,]    1 -0.5
solve(A) %*% A
##      [,1] [,2]
## [1,]    1    0
## [2,]    0    1
```

The solve function is really aimed at solving a set of linear equations, and it does that if it gets a vector argument as well, but you can check the documentation for the function to see how this is done.

You can also get higher dimensional vectors, called *arrays*, by setting the dimension attribute with more than two dimensions as arguments or you can use the array function.

Lists

Lists, like vectors, are sequences, but unlike vectors, the elements of a list can be any kind of objects, and they do not have to be the same type of objects. This means that you can construct more complex data structures out of lists.

For example, we can make a list of two vectors:

```
list(1:3, 5:8)
## [[1]]
## [1] 1 2 3
##
## [[2]]
## [1] 5 6 7 8
```

Notice how the vectors do not get concatenated like they would if we combined them with c(). The result of this command is a list of two elements that happens to be both vectors.

They didn't have to have the same type either, we could make a list like this, which also consist of two vectors but vectors of different types:

```
list(1:3, c(TRUE, FALSE))
## [[1]]
## [1] 1 2 3
##
## [[2]]
## [1]  TRUE FALSE
```

Since lists can contain other lists, you can build tree-like data structures quite naturally.

```
list(list(), list(list(), list()))
## [[1]]
## list()
##
## [[2]]
## [[2]][[1]]
## list()
##
## [[2]][[2]]
## list()
```

You can flatten a list into a vector using the function unlist(). This will force the elements in the list to be converted into the same type, of course, since that is required of vectors.

```
unlist(list(1:4, 5:7))
## [1] 1 2 3 4 5 6 7
```

Indexing

We saw basic indexing in Chapter 1, but there is much more to indexing in R that that. Type ?`[[` into the R prompt and prepare to be amazed.

We have already seen the basic indexing. If you want the *n*th element of a vector v, you use v[n]:

```
v <- 1:4
v[2]
## [1] 2
```

But this you already knew. You also know that you can get a subsequence out of the vector using a range of indices:

```
v[2:3]
## [1] 2 3
```

This is really just a special case of using a vector of indices:

```
v[c(1,1,4,3,2)]
## [1] 1 1 4 3 2
```

Here we are indexing with positive numbers, which makes sense since the elements in the vector have positive indices, but it is also possible to use negative numbers to index in R. If you do that it is interpreted as specifying the complement of the values you want. So if you want all elements *except* the first element, you can use:

```
v[-1]
## [1] 2 3 4
```

You can also use multiple negative indices to remove some values:

```
v[-(1:2)]
## [1] 3 4
```

213

You cannot combine positive and negative indices. I don't even know how that would even make sense, but in any case, you just can't.

Another way to index is to use a Boolean vector. This vector should be the same length as the vector you index into, and it will pick out the elements where the Boolean vector is true.

```
v[v %% 2 == 0]
## [1] 2 4
```

If you want to assign to a vector you just assign to elements you index; as long as the vector to the right of the assignment operator has the same length as the elements the indexing pulls out you will be assigning to the vector.

```
v[v %% 2 == 0] <- 13
v
## [1]  1 13  3 13
```

If the vector has more than one dimension—remember that matrices and arrays are really just vectors with more dimensions—then you subset them by subsetting each dimension. If you leave out a dimension, you will get whole range of values in that dimension, which is a simple way to of getting rows and columns of a matrix:

```
 m <- matrix(1:6, nrow = 2, byrow = TRUE)
 m
##      [,1] [,2] [,3]
## [1,]    1    2    3
## [2,]    4    5    6
 m[1,]
## [1] 1 2 3
 m[,1]
## [1] 1 4
```

You can also index out a submatrix this way by providing ranges in one or more dimensions:

```
m[1:2,1:2]
##      [,1] [,2]
## [1,]    1    2
## [2,]    4    5
```

When you pull out a one-dimensional submatrix—as we did with m[1,]—the result is a vector, not a matrix. Sometimes that is what you want; sometimes you don't really care if you get a matrix or a vector, but sometimes you want to do linear algebra, and then you definitely want that the submatrix you pull out is a matrix. You can tell R that it shouldn't reduce a one-dimensional matrix to a row by giving the indexing an option drop=FALSE:

```
m[1,,drop=FALSE]
##      [,1] [,2] [,3]
## [1,]    1    2    3
m[,1,drop=FALSE]
##      [,1]
## [1,]    1
## [2,]    4
```

If this looks weird to you (giving indexing an option) then what you need to know is that everything in R involves function calls. Indexing into a matrix is just another function call, and functions can take named arguments. That is all that is happening here.

When you subset a list using [ ] the result is always another list. If this surprises you, just remember that when you subset a vector you also always get a vector back. You just don't think so much about it because the way we see single values are always as vectors of length one, so we are more used to that.

Anyway, you will always get a list out of subsetting a list with [ ]. Even if you are subsetting a single element you are not getting that element; you are getting a list containing that one element.

```
L <- list(1,2,3)
L[1]
## [[1]]
## [1] 1
L[2:3]
## [[1]]
## [1] 2
##
## [[2]]
## [1] 3
```

If you want to get to the actual element in there, you need to use the [[ ]] operator instead.

```
L <- list(1,2,3)
L[[1]]
## [1] 1
```

Named Values

The elements in a vector or a list can have names. These are attributes that do not affect the values of the elements but can be used to refer to them.

You can set these names when you create the vector or list:

```
v <- c(a = 1, b = 2, c = 3, d = 4)
v
## a b c d
## 1 2 3 4
L <- list(a = 1:5, b = c(TRUE, FALSE))
L
## $a
## [1] 1 2 3 4 5
##
## $b
## [1]  TRUE FALSE
```

Or you can set the names using the names<- function. That weird name, by the way, means that you are dealing with the names() function combined with assignment:

```
names(v) <- LETTERS[1:4]
v
## A B C D
## 1 2 3 4
```

215

You can use names to index vectors and lists (where the [ ] and [[ ]] returns either a list or the element of the list, as before):

```
v["A"]
## A
## 1
L["a"]
## $a
## [1] 1 2 3 4 5
L[["a"]]
## [1] 1 2 3 4 5
```

When you have named values, you can also use a third indexing operator, the $ operator. It essentially works like [[ ]] except that you don't have to put the name in quotes:

```
L$a
## [1] 1 2 3 4 5
```

There is never really any good time to introduce the [[ ]] operator for vectors but here goes: if you use the [[ ]] operator on a vector it will only let you extract a single element, and if the vector has names, it will remove the name.

Factors

The factor type we saw in Chapter 1 is technically also a vector type, but it isn't a primitive type in the same sense as the previous types. It is stored as a vector of integers—the levels in the factor—and has associated attributes such as the levels. It is implemented using the class system we return to in Chapter 10, and we will not discuss if further here.

Formulas

Another data type is the formula. We saw these in Chapter 6 and we can create them using the ~ operator. Like factors, the result is an object defined using a class. You will see how you can use formulas to implement your own statistical models via model matrices in Project 2.

Control Structures

Control structures determine the flow of execution of a program. You can get far by just having one statement or expression after another, but eventually you will have to do one thing instead of another depending on the results of a calculation, and this is where control structures come in.

Like many other programming languages you have two kinds of control structures in R: select (if statements) and loops (for, while, or repeat statements).

Selection Statements

If statements look like this:

```
if (boolean) {
    # do something
}
```

Or like this:

```
if (boolean) {
    # do one thing
} else {
    # do another thing
}
```

You can string them together like this:

```
if (x < 0) {
    # handle negative x
} else if (x > 0) {
    # handle positive x
} else {
    # handle if x is zero
}
```

In all the examples here, I put the statements you do if the condition is true or if it is false in curly brackets. Strictly speaking, this isn't necessary if we are talking about a single statement. This would work just fine:

```
if (x > 0) "positive" else if (x < 0) "negative" else "zero"
```

But it would fail if you put newlines in between the statements; the R parser would be confused about that and there you *do* need curly brackets. This would be a syntax error:

```
if (x > 0)
    print("positive")
else if (x < 0)
    print("negative")
else
    print("zero")
```

While this would be okay:

```
if (x > 0) {
    print("positive")
} else if (x < 0) {
    print("negative")
} else {
    print("zero")
}
```

I recommend always using curly brackets since they work fine when you only have a single statement so you are not doing anything wrong in that case, and they are the only thing that works when you have more than one statement or when you have newlines in the if statement.

Loops

The most common looping construction you will use is probably the for loop. You use the for loop to iterate through the elements of a sequence; the construction works like this:

```
for (i in 1:4) {
    print(i)
}
## [1] 1
## [1] 2
## [1] 3
## [1] 4
```

Keep in mind, though, that it is the elements in the sequence you are iterating through, so the variable you assign to the iteration variable are the elements in the sequence and *not* the index into the sequence. If you want to loop through the indices into the sequence, you can use the seq_along function:

```
x <- c("foo", "bar", "baz")
for (i in seq_along(x)) {
    print(i)
    print(x[i])
}
## [1] 1
## [1] "foo"
## [1] 2
## [1] "bar"
## [1] 3
## [1] "baz"
```

You will sometimes see code that uses this construction:

```
for (i in 1:length(x)) {
    # do stuff
}
```

Don't do that. It won't work if the sequence x is empty.

```
x <- c()
1:length(x)
## [1] 1 0
```

If you want to jump to the next iteration of a loop, you can use the next keyword. For example, the following will only print every second element of x:

```
for (i in seq_along(x)) {
    if (i %% 2 == 0) {
        next
    }
    print(x[i])
}
```

If you want to terminate the loop completely, you can use break.

```
for (i in 1:100) {
    if (i %% 2 == 0) {
        next
    }
    if (i > 5) {
        break
    }
    print(i)
}
## [1] 1
## [1] 3
## [1] 5
```

The two other loop constructs you won't use as often. They are the while and repeat loops.
The while loop iterates as long as a Boolean expression is true and looks like this:

```
i <- 1
while (i < 5) {
    print(i)
    i <- i + 1
}
## [1] 1
## [1] 2
## [1] 3
## [1] 4
```

The repeat look simply goes on forever, at least until you break out of the loop.

```
i <- 1
repeat {
    print(i)
    i <- i + 1
    if (i > 5) break
}
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
```

A Word of Warning About Looping

If you read more about R you will soon run into the statement that loops are slow in R. It isn't really as bad as some make it out to be, but it is somewhat justified. Because R is an extremely dynamic language—functions and variables can change at any time during program execution—it is hard for the interpreter to optimize code before it runs it, unlike in some other languages (but not that different from other dynamic languages such as Python). There haven't been many attempts at optimizing loops either, though, because there are typically better solutions in R than to use an explicit loop statement.

R is a so-called functional language (among other things) and in functional languages, you typically don't use loops. The way looping constructions work, you need to change the value of a looping variable or a Boolean expression while you execute the code and changing variables is considered "impure" in function languages (so, obviously, R is not a pure functional language). Instead, recursive functions are used for looping. Most functional languages don't even have looping constructions—and pure functional languages certainly do not. R is a bit more pragmatic, but you *are* typically better off with using alternatives to loops.

We get more into that in the next chapter.

Functions

You define functions this way:

```
name <- function(arguments) expression
```

Where name can be any variable name, arguments is a list of formal arguments to the function, and expression is what the function will do when you call it. It says expression because you might as well think about the body of a function as an expression, but typically it is a sequence of statements enclosed by curly brackets:

```
name <- function(arguments) { statements }
```

It is just that such a sequence of statements is also an expression; the result of executing a series of statements is the value of the last statement.

The following function will print a statement and return 5 because the statements in the function body are first a print statement and then just the value 5 that will be the return value of the function:

```
f <- function() {
    print("hello, world")
    5
}
f()
## [1] "hello, world"
## [1] 5
```

We usually don't write functions without arguments—like I just did previously—but have one or more formal arguments. The arguments, in their simplest form, are just variable names. They are assigned values when you call the function, and these can then be used inside the function's body[1]:

```
plus <- function(x, y) {
    print(paste(x, "+", y, "is", x + y))
    x + y
}

div <- function(x, y) {
    print(paste(x, "/", y, "is", x / y))
    x / y
}
```

---

[1]I am actually lying here because the arguments to a function are not assigned values but expressions that haven't been evaluated yet. See the *Lazy Evaluation* section later.

```
plus(2, 2)
## [1] "2 + 2 is 4"
## [1] 4
div(6, 2)
## [1] "6 / 2 is 3"
## [1] 3
```

Named Arguments

The order of arguments matters when you call a function because it determines which argument gets set to which value:

```
div(6,2)
## [1] "6 / 2 is 3"
## [1] 3
div(2,6)
## [1] "2 / 6 is 0.333333333333333"
## [1] 0.3333333
```

If a function has many arguments, though, it can be hard always to remember the order, so there is an alternative way to specify which variable is given which values: named arguments. It means that when you call a function, you can make explicit which parameter each argument should be set to.

```
div(x = 6, y = 2)
## [1] "6 / 2 is 3"
## [1] 3
div(y = 2, x = 6)
## [1] "6 / 2 is 3"
## [1] 3
```

This makes explicit which parameter gets assigned which value, and you can think of it as an assignment operator. You shouldn't, though, because although you *can* use = as an assignment operator you *cannot* use <- for specifying named variables. It looks like you can, but it doesn't do what you want it to do (unless you want something really weird):

```
div(x <- 6, y <- 2)
## [1] "6 / 2 is 3"
## [1] 3
div(y <- 2, x <- 6)
## [1] "2 / 6 is 0.333333333333333"
## [1] 0.3333333
```

The assignment operator <- returns a value and that is passed along to the function as positional arguments. So in the second function call above you are assigning 2 to y and 6 to x in the scope *outside* the function, but the values you pass to the function are positional so inside the function you have given 2 to x and 6 to y.

Don't confuse the two assignment operators: the code most likely will run, but it is unlikely to do what you want it to do!

Default Parameters

When you define a function, you can provide default values to parameters like this:

```
pow <- function(x, y = 2) x^y
pow(2)
## [1] 4
pow(3)
## [1] 9
pow(2, 3)
## [1] 8
pow(3, 3)
## [1] 27
```

Default parameter values will be used whenever you do not provide the parameter at the function call.

Return Values

The return value of a function is the last expression in the statements executed in the function body. If the function is a sequence of statements, this is just the last statement in the sequence, but by using control structures, you can have different statements as the last statement:

```
safer_div <- function(x, y) {
    if (y == 0) {
        NA
    } else {
        x / y
    }
}
safer_div(4, 2)
## [1] 2
safer_div(4, 0)
## [1] NA
```

It is also possible to return explicitly from a function—similarly to breaking from a loop—using the return() statement.

```
safer_div <- function(x, y) {
    if (y == 0) {
        return(NA)
    }
    x / y
}
safer_div(4, 2)
## [1] 2
safer_div(4, 0)
## [1] NA
```

Notice that the `return()` statement has the return value in parentheses. Many programming languages would allow you to write this:

```
safer_div <- function(x, y) {
    if (y == 0) {
        return NA
    }
    x / y
}
```

But this is an error in R.

Lazy Evaluation

Several places I have written about providing values to the function parameters when we call a function. In many programming languages this is exactly how function calls work—the expressions provided for each parameter are evaluated, and the results are assigned to the function parameters so the function can use them in the function body—but in R it is actually the expressions that are assigned to the function parameters. And the expressions are not evaluated until they are needed; something called *lazy evaluation*.

There are some benefits to this way of handling function parameters and some weird consequences as well.

The first benefit is that it makes default parameters more flexible. We can write a function like this:

```
f <- function(x, y = x^2) y + x
```

Where y has a default value that depends on the other parameter x. At the time where the function is declared the value of x is not know but y is not evaluated there so it doesn't matter. Whenever we call the function, x is known inside the body of the function, and that is where we need the value of y so that is where the expression will be evaluated.

```
f(2)
## [1] 6
```

Since y isn't evaluated before it is used, it *does* also mean that if you assign a different value to x before you use y then y evaluates to a value that depends on the new value of x. Not the value of x at the time the function was called!

```
g <- function(x, y = x^2) { x <- 0; y + x }
g(2)
## [1] 0
```

If, on the other hand, y is evaluated before we assign to x then it will evaluate to the value that depends on x at the time we evaluate it and remain that value. It is evaluated the first time it is needed, and the result is then remembered for any later time we refer to y.

```
h <- function(x, y = x^2) { y; x <- 0; y + x }
h(2)
## [1] 4
```

So lazy evaluation lets you specify default parameters that depend on other parameters in a context where those parameters are unknown, but it comes at the prize of the value of the parameter depending on the context at the first time it gets evaluated.

If it was just to be able to specify variables this way we could, of course, have a solution that doesn't involve the weirdness that we pay for it. This is what most programming languages have done, after all, but there are other benefits of lazy evaluation: you only evaluate an expression if you actually need it.

Scoping

Scope in R is lexical. This means that if a variable is used in a function but not defined in the function or part of the function's parameters, then R will start searching outward in the code from where the function was created. This essentially means searching outward and upward from the point in the code where the function is specified, since a function is created when the code is executed where the function is defined.

Consider this code:

```
x <- "x"
f <- function(y) {
    g <- function() c(x, y)
    g()
}
f("y")
## [1] "x" "y"
```

Here we have a global variable x and a function f that takes a parameter argument y. Inside f, we define the function g that neither defines nor take as formal arguments variables x and y but does return them. We evaluate g as the last statement in f so that becomes the result of calling f at the last line.

Inside g, we have not defined x or y, so to find their values R will search outward from where g is created. It will find y as the argument of the function f so get it from there and continue outward to find x at the global level.

The variables that g refers to are the variables and not the values at the time that g is created, so if we update the variables after we create g, we also change the value that g will return:

```
x <- "x"
f <- function(y) {
  g <- function() c(x, y)
  y <- "z"
  g()
}
f("y")
## [1] "x" "z"
```

This isn't just the lazy evaluation madness—it is not that g hasn't evaluated y yet and it, therefore, can be changed. It does look up the value of y when it needs it:

```
x <- "x"
f <- function(y) {
  g <- function() c(x, y)
  g()
  y <- "z"
  g()
}
f("y")
## [1] "x" "z"
```

If we return the function g from f rather than the result of evaluating it, we see another feature of R's scoping—something called *closures*. R remembers the values of variables inside a function that we have returned from and that is no longer an active part of our computation. In the example below, we returned the function g at which point there is no longer an active f function. So there is not really an active instance of the parameter y any longer. Yet g refers to a y, so the parameter we gave to f is actually remembered.

```
x <- "x"
f <- function(y) {
  g <- function() c(x, y)
  g
}
g <- f("y")
g()
## [1] "x" "y"
```

We can see how this works if we invoke f twice, with different values for parameter y:

```
x <- "x"
f <- function(y) {
  g <- function() c(x, y)
  g
}
g <- f("y")
h <- f("z")
g()
## [1] "x" "y"
h()
## [1] "x" "z"
```

This creates two different functions. Inside f they are both called g, but they are two different functions because they are created in two different calls to f, and they remember two different y parameters because the two instances of f were invoked with different values for y.

When looking outward from the point where a function is defined, it is looking for the values of variables at the time a function is invoked, not the values at the time where the function is created. Variables do not necessarily have to be defined at the time the function is created; they just need to be defined when the function is eventually invoked.

Consider this code:

```
f <- function() {
  g <- function() c(y, z)
  y <- "y"
  g
}
h <- f()
h()
## Error in h(): object 'z' not found
z <- "z"
h()
## [1] "y" "z"
```

Where the function g is defined—inside function f—it refers to variables y and z that are not defined yet. This doesn't matter because we only create the function g; we do not invoke it. We then set the variable y inside the context of the invocation of f and return g. Outside of the function call, we name the return value of f() h. If we call h at this point it will remember that y was defined inside f—and it will remember its value at the point in time where we returned from f. There still isn't a value set for z so calling h results in an error. Since z isn't defined in the enclosing scopes of where the inner function refers to it, it must be defined at the outermost global scope, but it isn't. If we do set it there, the error goes away because now R can find the variable by searching outward from where the function was created.

I shouldn't really be telling you this because the feature I am about to tell you about is dangerous. I will show you a way of making functions have even more side-effects than they otherwise have, and functions really shouldn't have side-effects at all. Anyway, this *is* a feature of the language—and if you are very careful with how you use it—it can be very useful when you just feel the need to make functions have side-effects.

This is the problem: what if you want to assign to a variable in a scope outside the function where you want the assignment to be made? You cannot just assign to the variable because if you assign to a variable that isn't found in the current scope, then you *create* that variable in the current scope.

```
f <- function() {
    x <- NULL
    set <- function(val) { x <- val }
    get <- function() x
    list(set = set, get = get)
}

x <- f()
x$get()
## NULL
x$set(5)
x$get()
## NULL
```

In this code—that I urge you to read carefully because there are a few neat ideas in it—we have created a getter and setter function; the getter tells us what the variable x is, and the setter is supposed to update it. But setting x in the body of set creates a local variable inside that function; it doesn't assign to the x one level up.

There is a separate assignment operator, <<-, you can use for that. It will not create a new local variable but instead search outward to find an existing variable and assign to that. If it gets all the way to the outermost global scope, though, it will create the variable there if it doesn't already exist.

If we use that assignment operator in the previous example, we get the behavior we were aiming for.

```
f <- function() {
    x <- NULL
    set <- function(val) { x <<- val }
    get <- function() x
    list(set = set, get = get)
}

x <- f()
x$get()
## NULL
x$set(5)
x$get()
## [1] 5
```

If we hadn't set the variable x inside the body of f in this example, both the getter and setter would be referring to a global variable, in case you are wondering, and the first call to get would cause an error if there was no global variable. While this example shows you have to create an object where functions have side-effects, it *is* quite a bit better to let functions modify variables that are hidden away in a closure like this than it is to work with global variables.

Function Names Are Different from Variable Names

One final note on scopes—which I am not sure should be considered a feature or a bug—is that if R sees something that looks like a function call, it is going to go searching for a function, even if searching outward from a function creation would get to a non-function first.

```
n <- function(x) x
f <- function(n) n(n)
f(5)
## [1] 5
```

Under the scoping rule that says that you should search outward, the n inside the f function should refer to the parameter to f. But it is clear that the first n is a function call and the second is its argument, so when we call f it sees that the parameter isn't a function so it searches further outward and finds the *function* n. It calls that function with its argument. So the two n's inside f actually refers to different things.

Of course, if we call f with something that is actually a function, then it recognizes that n is a function and it calls that function with itself as the argument.

```
f(function(x) 15)
## [1] 15
```

Interesting, right?

Recursive Functions

The final topic we will cover in this chapter before we get to the exercises is recursive functions. Some people find this a difficult topic, but in a functional programming language, it is one of the most basic building blocks so it is really worth spending some time wrapping your head around, even though you are much less likely to need recursions in R than you are in most pure functional languages.

At the most basic level, though, it is just that we can define a function's computations in terms of calls to the same function—we allow a function to call itself, just with new parameters.

Consider the factorial operator $n! = n \times (n - 1) \times \ldots \times 3 \times 2 \times 1$. We can rewrite the factorial of n in terms of n and a smaller factorial, the factorial of $n - 1$ and get $n! = n \times (n - 1)!$. This is a classical case of where recursion is useful: we define the value for some n in terms of the calculations on some smaller value. As a function, we would write factorial(n) equals n * factorial(n-1).

There are two aspects to a recursive function, though. Solving a problem for size n involves breaking down the problem into something you can do right away and combine that with calls of the function with a smaller size, here $n - 1$. This part we call the "step" of the recursion. We cannot keep reducing the problem into smaller and smaller bits forever—that would be an infinite recursion which is as bad as an infinite loop in that we never get anywhere—at some point we need to have reduced the problem to a size small enough that we can handle it directly. That is called the basis of the recursion.

For factorial, we have a natural basis in 1 since 1! = 1. So we can write a recursive implementation of the factorial function like this:

```
factorial <- function(n) {
    if (n == 1) {
        1
    } else {
        n * factorial(n - 1)
    }
}
```

It is actually a general algorithmic strategy, called *divide and conquer*, to break down a problem into sub-problems that you can handle recursively and then combining the results some way.

Consider sorting a sequence of numbers. We could sort a sequence using this strategy by first noticing that we have a simple basis—it is easy to sort an empty sequence or a sequence with a single element since we don't have to do anything there. For the step, we can break the sequence into two equally sized pieces and sort them recursively. Now we have two sorted sequences and if we merge these two we have combined them into a single sorted sequence.

Let's get started.

We need to be able to merge two sequences so we can solve that problem first. This is something we should be able to do with a recursive function because if either sequence is empty, we have a basis case where we can just return the other sequence. If there are elements in both sequences, we can pick the sequence whose first element is smallest, pick that out as the first element we need in our final result and just concatenate the merging of the remaining numbers.

```
merge <- function(x, y) {
  if (length(x) == 0) return(y)
  if (length(y) == 0) return(x)

  if (x[1] < y[1]) {
    c(x[1], merge(x[-1], y))
  } else {
    c(y[1], merge(x, y[-1]))
  }
}
```

A quick disclaimer here: normally this algorithm would run in linear time but because of the way we call recursively we are actually copying vectors whenever we are removing the first element, making it a quadratic time algorithm. Implementing a linear time merge function is left as an exercise.

Using this function, we can implement a sorting function. This algorithm is called merge sort so that is what we call the function:

```
merge_sort <- function(x) {
  if (length(x) < 2) return(x)

  n <- length(x)
  m <- n %/% 2

  merge(merge_sort(x[1:m]), merge_sort(x[(m+1):n]))
}
```

So here, using two simple recursive functions, we solved a real algorithmic problem in a few lines of code. This is typically the way to go in a functional programming language like R. Of course, when things are easier done using loops you shouldn't stick to the pure functional recursions. Use what is easiest in any situation you are in, unless you find that it is too slow. Only then do you start getting clever.

Exercises

Try the following exercises to become more comfortable with the concepts discussed in this chapter.

Fibonacci Numbers

The *Fibonacci* numbers are defined as follows: The first two Fibonacci numbers are 1, $F_1 = F_2 = 1$. For larger Fibonacci numbers, they are defined as $F_i = F_{i-1} + F_{i-2}$.

Implement a recursive function that computes the n'th Fibonacci number.

The recursive function for Fibonacci numbers is usually quite inefficient because you are recomputing the same numbers several times in the recursive calls. So implement another version that computes the n'th Fibonacci number iteratively (that is, start from the bottom and compute the numbers up to n, without calling recursively).

Outer Product

The outer product of two vectors, **v** and **w**, is a matrix defined as follows:

$$\boldsymbol{v} \otimes \boldsymbol{w} = \boldsymbol{v}\boldsymbol{w}^T = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} \begin{bmatrix} w_1 & w_2 & w_3 & w_4 \end{bmatrix} = \begin{bmatrix} v_1w_1 & v_1w_2 & v_1w_3 & v_1w_4 \\ v_2w_1 & v_2w_2 & v_2w_3 & v_2w_4 \\ v_3w_1 & v_3w_2 & v_3w_3 & v_3w_4 \end{bmatrix}$$

Write a function that computes the outer product of two vectors.

There actually is a built-in function, `outer`, that you are overwriting here. You can get to it using the name `base::outer` even after you have overwritten it. You can use it to check that your own function is doing the right thing.

Linear Time Merge

The merge function we used copies vectors in its recursive calls. This makes it slower than it has to be. Implement a linear time merge function.

Before you start, though, you should be aware of something. If you plan to append to a vector by writing something like this:

```
v <- c(v, element)
```

You will end up with a quadratic time algorithm again. This is because when you do this, you are actually creating a new vector where you first copy all the elements in the old `v` vector into the first elements and then add the `element` at the end. If you do this n times, you have spent on average order n^2 per operation. It is because people do something like this in loops, more than the R interpreter, that has given R its reputation for slow loops. You should never append to vectors unless there is no way to avoid it.

In the case of the merge function, we already know how long the result should be, so you can pre-allocate a result vector and copy single elements into it. You can create a vector of length *n* like this:

```
n <- 5
v <- vector(length = n)
```

Should you ever need it, you can make a list of length *n* like this:

```
vector("list", length = n)
```

Binary Search

Binary search is a classical algorithm for finding out if an element is contained in a sorted sequence. It is a simple recursive function. The basic case handles a sequence of one element. There you can directly compare the element you are searching for with the element in the sequence to determine if they are the same. If you have more than one element in the sequence, pick the middle one. If it is the element you are searching for, you are done and can return that the element is contained in the sequence. If it is smaller than the element you are searching for then, you know that *if* the element is in the list then it has to be in the last half of the sequence, and you can search there. If it is larger than the element you are searching for, then you know that if it is in the sequence, it must be in the first half of the sequence, and you search recursively there.

If you implement this exactly as described, you have to call recursively with a subsequence. This involves copying that subsequence for the function call which makes the implementation *much* less efficient than it needs to be. Try to implement binary search without this.

More Sorting

In the merge sort we implemented, we solve the sorting problem by splitting a sequence in two, sorting each subsequence, and then merging them. If implemented correctly this algorithm will run in time $O(n\log n)$, which is optimal for sorting algorithms if we assume that the only operations we can do on the elements we sort are comparing them.

If the elements we have are all integers between 1 and *n* and we have *m* of them, we can sort them in time $O(n + m)$ using bucket sort instead. This algorithm first creates a vector of counts for each number between 1 and *n*. This takes time $O(n)$. It then runs through the *m* elements in our sequence, updating the counter for number *i* each time it sees *i*. This runs in time $O(m)$. Finally, it runs through these numbers from 1 up to *n* and outputting each number, the number of times indicated by the counters, in time $O(n + m)$.

Implement bucket sort.

Another algorithm that works by recursion, and that runs in expected time $O(n\log n)$, is quick sort. Its worst case complexity is actual $O(n^2)$ but on average it runs in time $O(n\log n)$ and with a smaller overhead than merge sort (if you implement it correctly).

It works as follows: the basis case—a single element—is the same as merge sort. When you have more than one element you pick one of the elements in the sequence at random; call it the pivot. Now split the sequence into those elements that are smaller than the pivot, those that are equal to the pivot, and those that are larger. Sort the sequences of smaller and larger elements recursively. Then output all the sorted smaller elements, then the elements equal to the pivot, and then the sorted larger elements.

Implement quick sort.

Selecting the k Smallest Element

If you have *n* elements, and you want the *k* smallest, an easy solution is to sort the elements and then pick number *k*. This works well and in most cases is easily fast enough, but it is actually possible to do it faster. See, we don't actually need to sort the elements completely, we just need to have the *k* smallest element moved to position *k* in the sequence.

The quick sort algorithm from the previous exercise can be modified to solve this problem. Whenever we split a sequence into those smaller than, equal to, and larger than the pivot, we sort the smaller and larger elements recursively. If we are only interested in finding the element that would eventually end up at position *k* in the sorted lists we don't need to sort the sequence that doesn't overlap this index. If we have *m* < *k* elements smaller than the pivot, we can just put them at the front of the sequence without sorting them. We need them there to make sure that the *k*'th smallest element ends up at the right index, but we don't need them sorted. Similar, if *k* < *m* we don't need to sort the larger elements. If we sorted them, they would all end up at indices larger than *k* and we don't really care about those. Of course, if there are *m* < *k* elements smaller than the pivot and *l* equal to the pivot, with *m* + *l* ≥ *k*, then the *k* smallest element is equal to the pivot, and we can return that.

Implement this algorithm.

■ ■ ■

Advanced R Programming

This chapter gets into more details of some aspects of R. This chapter is called "Advanced R Programming" only because it is additional elements on top of the quick introduction you got in the last chapter. Except, perhaps, for the functional programming toward the end, we will not cover anything that is conceptually more complex that we did in the previous chapter. It is just a few more technical details we will dig into.

I stole the title from Hadley Wickham's excellent book of the same name (see `http://adv-r.had.co.nz`) and most of what I cover here, he does in his book as well. He does cover a lot more, though, so this is a book you should get if you want really to drill into the advanced aspects of R programming.

Working with Vectors and Vectorizing Functions

We start out by returning to expressions. In the previous chapter, you saw expressions on single (scalar) values, but you also saw that R doesn't really have scalar values; all the primitive data you have is actually vectors of data. What this means is that the expressions you use in R are actually operating on vectors, not on single values.

When you write this:

```
(x <- 2 / 3)
## [1] 0.6666667
(y <- x ** 2)
## [1] 0.4444444
```

The expressions you write *are*, of course, working on single values—the vectors x and y have length 1, but it is really just a special case of working on vectors.

```
(x <- 1:4 / 3)
## [1] 0.3333333 0.6666667 1.0000000 1.3333333
(y <- x ** 2)
## [1] 0.1111111 0.4444444 1.0000000 1.7777778
```

R works on vectors using two rules: operations are done element-wise, and vectors are repeated as needed.

When you write an expression such as x + y, you are really saying that you want to create a new vector that consists of the element-wise sum of the elements in vectors x and y. So for x and y like this:

```
x <- 1:5
y <- 6:10
```

Writing this:

```
(z <- x + y)
## [1]  7  9 11 13 15
```

Amounts to writing this:

```
z <- vector(length = length(x))
for (i in seq_along(x)) {
    z[i] <- x[i] + y[i]
}
z
## [1]  7  9 11 13 15
```

This is the case for all arithmetic expressions or for logical expressions involving | or & (but not || or &&; these do not operate on vectors element-wise). It is also the case for most functions you can call, such as sqrt or sin:

```
sqrt((1:5)**2)
## [1] 1 2 3 4 5
sin(sqrt((1:5)**2))
## [1]  0.8414710  0.9092974  0.1411200 -0.7568025
## [5] -0.9589243
```

When you have an expression that involves vectors of different lengths, you cannot directly evaluate expressions element-wise. When this is the case, R will try to repeat the shorter vector(s) to create vectors of the same length. For this to work, the shorter vector(s) should have a length divisible in the length of the longest vector, i.e., you should be able to repeat the shorter vector(s) an integer number of times to get the length of the longest vector. If this is possible, R repeats vectors as necessary to make all vectors the same length as the longest and then does operations element-wise:

```
x <- 1:10
y <- 1:2
x + y
##  [1]  2  4  4  6  6  8  8 10 10 12
z <- 1:3
x + z
## Warning in x + z: longer object length is not a
## multiple of shorter object length
##  [1]  2  4  6  5  7  9  8 10 12 11
```

If the shorter vector(s) cannot be repeated an integer number of times to match up, R will still repeat as many times as needed to match the longest vector, but you will get a warning. Most of the time something like this happens, it is caused by buggy code.

```
z <- 1:3
x + z
## Warning in x + z: longer object length is not a
## multiple of shorter object length
##  [1]  2  4  6  5  7  9  8 10 12 11
```

In the expression you saw a while back, different vectors are repeated:

```
(x <- 1:4 / 3)
## [1] 0.3333333 0.6666667 1.0000000 1.3333333
(y <- x ** 2)
## [1] 0.1111111 0.4444444 1.0000000 1.7777778
```

When we divide 1:4 by 3 we need to repeat the (length one) vector 3 four times to be able to divide the 1:4 vector with the 3 vector. When we compute x ** 2, we must repeat 2 four times as well.

Whenever you consider writing a loop over vectors to do some calculations for each element, you should always consider using such vectorized expressions instead. It is typically much less error prone and since it involves implicit looping handled by the R runtime system, it is almost guaranteed to be faster than an explicit loop.

ifelse

Control structures are not vectorized. For example, if statements are not. If you want to compute a vector y from vector x such that y[i] == 5 if x[i] is even and y[i] == 15 if x[i] is odd—for example—you cannot write this as a vector expression:

```
x <- 1:10
if (x %% 2 == 0) 5 else 15
## Warning in if (x%%2 == 0) 5 else 15: the condition
## has length > 1 and only the first element will be
## used
## [1] 15
```

Instead, you can use the ifelse function that works like a vectorized selection; if the condition in its first element is true, it returns the value in its second argument; otherwise, it returns the value in its third argument. It does this as vector operations:

```
x <- 1:10
ifelse(x %% 2 == 0, 5, 15)
## [1] 15  5 15  5 15  5 15  5 15  5
```

Vectorizing Functions

When you write your own functions, you can write them so that they can also be used to work on vectors, that is, you can write them so that they can take vectors as input and return vectors as output. If you write them this way, then they can be used in vectorized expressions the same way as built-in functions such as sqrt and sin.

The easiest way to make your function work on vectors is to write the body of the function using expressions that work on vectors.

```
f <- function(x, y) sqrt(x ** y)
f(1:6, 2)
## [1] 1 2 3 4 5 6
f(1:6, 1:2)
## [1] 1.000000 2.000000 1.732051 4.000000 2.236068
## [6] 6.000000
```

If you write a function where you cannot write the body this way, but where you still want to be able to use it in vector expressions, you can typically get there using the Vectorize function.

As an example, say we have a table mapping keys to some values. We can imagine that we want to map names in a class to the roles the participants in the class have. In R, we would use a list to implement that kind of tables, and we can easily write a function that uses such a table to map names to roles.

```
role_table <- list("Thomas" = "Instructor",
                   "Henrik" = "Student",
                   "Kristian" = "Student",
                   "Randi" = "Student",
                   "Heidi" = "Student",
                   "Manfred" = "Student")

map_to_role <- function(name) role_table[[name]]
```

This works the way it is supposed to when we call it with a single name:

```
map_to_role("Thomas")
## [1] "Instructor"
map_to_role("Henrik")
## [1] "Student"
```

But it fails when we call the function with a vector because we cannot index the list with a vector in this way.

```
x <- c("Thomas", "Henrik", "Randi")
map_to_role(x)
## Error in role_table[[name]]: recursive indexing failed at level 2
```

So we have a function that maps a single value to a single value but doesn't work for a vector. The easy way to make such a function work on vectors is to use the Vectorize function. This function will wrap your function so it can work on vectors and what it will do on those vectors is what you would expect: it will calculate its value for each of the elements in the vector, and the result will be the vector of all the results.

```
map_to_role <- Vectorize(map_to_role)
map_to_role(x)
##      Thomas      Henrik       Randi
## "Instructor"   "Student"   "Student"
```

In this particular example with a table, the reason it fails is that we are using the [[index operator. Had we used the [operator, we would be fine (except that the result would be a list rather than a vector).

```
role_table[c("Thomas", "Henrik", "Randi")]
## $Thomas
## [1] "Instructor"
##
## $Henrik
## [1] "Student"
##
## $Randi
## [1] "Student"
```

So we could also have handled vector input directly by indexing differently and then flattening the list

```
map_to_role_2 <- function(names) unlist(role_table[names])
```

```
x <- c("Thomas", "Henrik", "Randi")
map_to_role_2(x)
##      Thomas      Henrik       Randi
## "Instructor"   "Student"   "Student"
```

It's not always that easy to rewrite a function to work on vector input, though, and when we cannot readily do that then, the Vectorize function can be very helpful.

As a side note, the issue with using [[with a vector of values isn't just that it this doesn't work. It actually does work, but it does something else than what we are trying to do here. If you give [[a vector of indices it is used to do what is called recursive indexing. It is a shortcut for looking up in the list using the first variable and pulling out the vector or list found there. It then takes that sequence and looks up using the next index and so on. Take as an example the following code:

```
x <- list("first" = list("second" = "foo"), "third" = "bar")
x[[c("first", "second")]]
## [1] "foo"
```

Here we have a list of two elements, the first of which is a list with a single element. You can look up the index "first" in the first list and get the list stored at that index. This list we can then index with the "second" index to get "foo" out.

The result is analogous to this:

```
x[["first"]][["second"]]
## [1] "foo"
```

This can be a useful feature—although, to be honest, I haven't found much use for it in my own programming—but it is not the effect we wanted in our mapping to roles example.

The apply Family

Vectorizing a function makes it possible to use it implicitly on vectors. We simply give it a vector as input, and we get a vector back as output. Notice though that it isn't really a vectorized function just because it takes a vector as input— many functions take vectors as input and return a single value as output, e.g., sum and mean—but we use those differently than vectorized functions. If you want that kind of function, you *do* have to handle explicitly how it deals with a sequence as input.

Vectorized functions can be used on vectors of data exactly the same way as on single values with exactly the same syntax. It is an implicit way of operating on vectors. But you can also make it more explicit when calling a function on all the elements in a vector, which gives you a bit more control of exactly *how* it is called. This, in turn, lets you work with those functions that do not just map from vectors to vectors but also from vectors to single values.

There are many ways of doing this—because it is a common thing to do in R—and you will see some general functions for working on sequences and calling functions on them in various ways later. In most of the code you will read, though, the functions that do this are named something with apply in their name and those functions are what we will look at here.

Let's start with apply. This is a function for operating on vectors, matrices (two-dimensional vectors), or arrays (higher-order dimensional vectors).

apply

This function is easiest explained with a matrix, I think, so let's make one.

```
m <- matrix(1:6, nrow=2, byrow=TRUE)
m
##      [,1] [,2] [,3]
## [1,]    1    2    3
## [2,]    4    5    6
```

The apply function takes (at least) three parameters. The first is the vector/matrix/array, the second which dimension(s) we should marginalize along, and the third the function we should apply.

What is meant by marginalization here is that you fix an index in some subset of the dimensions and pull out all values with that index. If we are marginalizing over rows, we will extract all the rows, so for each row, we will have a vector with an element per column, which is what we will pass the function.

We can illustrate this using the paste function that creates a string of its input by concatenating it.[1]

If you marginalize on rows, it will be called on each of the two rows and will produce two strings:

```
apply(m, 1, function(x) paste(x, collapse = ":"))
## [1] "1:2:3" "4:5:6"
```

If you marginalize on columns, it will be called on each of the three columns and produce three strings:

```
apply(m, 2, function(x) paste(x, collapse = ":"))
## [1] "1:4" "2:5" "3:6"
```

If you marginalize on both rows and columns, it will be called on each single element instead:

```
apply(m, c(1, 2), function(x) paste(x, collapse = ":"))
##      [,1] [,2] [,3]
## [1,] "1"  "2"  "3"
## [2,] "4"  "5"  "6"
```

The output here is two-dimensional. That is of course because we are marginalizing over two dimensions, so we get an output that corresponds to the margins.

We can get higher-dimensional output in other ways. If the function produces vectors (or higher dimensional vectors) as output then the output of apply will also be higher dimensional. Consider a function that takes a vector as input and duplicates it by concatenating it with itself. If we apply it to rows or columns, we get a vector for each row/column so the output has to be two-dimensional.

```
apply(m, 1, function(x) c(x,x))
##      [,1] [,2]
## [1,]    1    4
## [2,]    2    5
## [3,]    3    6
## [4,]    1    4
## [5,]    2    5
## [6,]    3    6
```

---

[1]So this is a case of a function that takes a vector as input but outputs a single value; it is not a vectorized function as those we talked about previously.

```
apply(m, 2, function(x) c(x,x))
##      [,1] [,2] [,3]
## [1,]    1    2    3
## [2,]    4    5    6
## [3,]    1    2    3
## [4,]    4    5    6
```

What `apply` does here is that it creates a matrix as its result, where the results of applying the function are collected as columns from left to right. The result of calling the function on the two rows is a matrix with two columns, the first column containing the result of applying the function to the first row and the second column the result of applying it to the second row. Likewise, for columns, the result is a vector with three columns, one for each column in the input matrix.

If you marginalize over more than one dimension and get multidimensional output through that—and at the same time produce more than one value—the two effects combine and you get even higher dimensional output:

```
apply(m, c(1,2), function(x) c(x,x))
## , , 1
##
##      [,1] [,2]
## [1,]    1    4
## [2,]    1    4
##
## , , 2
##
##      [,1] [,2]
## [1,]    2    5
## [2,]    2    5
##
## , , 3
##
##      [,1] [,2]
## [1,]    3    6
## [2,]    3    6
```

I admit that this output looks rather confusing. What happens, though, is the same thing as you saw when you marginalized on rows or columns. You get output for each margin you call the function on—in this case, each of the six cells in the input—and it gets collected "column-wise," except that this is at higher dimensions so it gets collected at the highest dimension (which is the columns for two-dimensional matrices). So to get to the result of the six values the function was called with, you need to index these the same way they were indexed in the input matrix—that is what the margins were—but you need to do it in the highest dimensions. So you can get the six concatenations of input values this way:

```
x <- apply(m, c(1,2), function(x) c(x,x))
k <- dim(x)[3]
n <- dim(x)[2]
for (i in 1:n) {
  for (j in 1:k) {
    print(x[,i,j])
  }
}
```

```
## [1] 1 1
## [1] 2 2
## [1] 3 3
## [1] 4 4
## [1] 5 5
## [1] 6 6
```

So what happens if the function to apply takes arguments besides those you get from the matrix?

```
sumpow <- function(x, n) sum(x) ** n
apply(m, 1, sumpow)
## Error in FUN(newX[, i], ...): argument "n" is missing, with no default
```

If it does, you can give these arguments as additional arguments to apply; they will be passed on to the function in the order you give them to apply.

```
apply(m, 1, sumpow, 2)
## [1]  36 225
```

It helps readability a lot, though, to explicitly name such parameters.

```
apply(m, 1, sumpow, n = 2)
## [1]  36 225
```

lapply

The lapply function is used for mapping over a list. Given a list as input, it will apply the function to each element in the list and output a list of the same length as the input containing the results of applying the function.

```
(l <- list(1, 2, 3))
## [[1]]
## [1] 1
##
## [[2]]
## [1] 2
##
## [[3]]
## [1] 3
lapply(l, function(x) x**2)
## [[1]]
## [1] 1
##
## [[2]]
## [1] 4
##
## [[3]]
## [1] 9
```

If the elements in the input list have names, these are preserved in the output vector.

```
l <- list(a=1, b=2, c=3)
lapply(l, function(x) x**2)
## $a
## [1] 1
##
## $b
## [1] 4
##
## $c
## [1] 9
```

If the input you provide is a vector instead of a list, it will just convert it into a list, and you will always get a list as output.

```
lapply(1:3, function(x) x**2)
## [[1]]
## [1] 1
##
## [[2]]
## [1] 4
##
## [[3]]
## [1] 9
```

Of course, if the elements of the list are more complex than a single number, you will still just apply the function to the elements.

```
lapply(list(a=1:3, b=4:6), function(x) x**2)
## $a
## [1] 1 4 9
##
## $b
## [1] 16 25 36
```

sapply and vapply

The sapply function does the same as lapply, but tries to simplify the output. Essentially, it attempts to convert the list returned from lapply into a vector of some sort. It uses some heuristics for this and guesses as to what you want as output, simplifies when it can, but gives you a list when it cannot figure it out.

```
sapply(1:3, function(x) x**2)
## [1] 1 4 9
```

The guessing is great for interactive work, but can be unsafe when writing programs. It isn't a problem that it guesses and can produce different types of output when you can see what it creates, but that is not safe deep in the guts of a program.

The function vapply essentially does the same as sapply but without the guessing. You have to tell it what you want as output, and if it cannot produce that, it will give you an error rather than produce output that your program may or may not know what to do with.

The difference in interface between the two functions is just that vapply expects a third parameter that should be a value of the type the output should be.

```
vapply(1:3, function(x) x**2, 1)
## [1] 1 4 9
```

Advanced Functions

We now get to some special cases for functions. I call the section "Advanced Functions," but not because they really are that advanced, only that they require a little bit more than the basic functions you have already seen.

Special Names

But first a word on names. Functions can have the same kind of names that variables have—after all, when we name a function, we are really just naming a variable that happens to hold a function—but we cannot have all kinds of names to the right of the assignment operator. For example, if is a function in R, but you cannot write if to the left of an assignment.

Functions with special names, that is names that you couldn't normally put before an assignment, can be referred to by putting them in backticks, so the function if we can refer to as if.

Any function can be referred to by its name in backticks and furthermore you can use backticks to refer to a function in a context where you usually couldn't use its name. This works for calling functions where you can use for example infix operators as normal function calls:

```
2 + 2
## [1] 4
`+`(2, 2)
## [1] 4
```

Or when assigning to a variable name for a function:

```
`%or die%` <- function(test, msg) if (!test) stop(msg)
```

```
x <- 5
(x != 0) %or die% "x should not be zero"
```

```
x <- 0
(x != 0) %or die% "x should not be zero"
## Error in (x != 0) %or die% "x should not be zero": x should not be zero
```

Infix Operators

If the last example looks weird to you, it may just be because you don't know about R's infix operators. In R, any variable that starts and ends with % is considered an infix operator, so calling x %foo% y amounts to calling `%foo%`(x,y). Several built-in infix operators do not have this type of name, + and * are two, but this naming convention makes it possible to create your own infix operators. You saw this come to good use in the dplyr package with the %>% pipe operator.

Replacement Functions

Replacement functions are functions that pretend to be modifying variables. You saw one early when you assigned names to a vector.

```
v <- 1:4
names(v) <- c("a", "b", "c", "d")
v
## a b c d
## 1 2 3 4
```

What happens here is that R recognizes that you are assigning to a function call and goes looking for a function named names<-. It calls this function with the vector v and the vector of names and the result of the function call gets assigned back to the variable v.

So what I just wrote means that this:

```
names(v) <- c("a", "b", "c", "d")
```

Is short for this:

```
v <- `names<-`(v, c("a", "b", "c", "d"))
```

Replacement functions are generally used to modify various attributes of an object, and you can write your own just by using the convention that their names must end with <-:

```
`foo<-` <- function(x, value) {
  x$foo <- value
  x
}

`bar<-` <- function(x, value) {
  x$bar <- value
  x
}

x <- list(foo=1, bar=2)

x$foo
## [1] 1
foo(x) <- 3
x$foo
## [1] 3
x$bar
## [1] 2
bar(x) <- 3
x$bar
## [1] 3
```

Keep in mind that it is just shorthand for calling a function and then reassigning the result to a variable. It is not actually modifying any data. This means that if you have two variables referring to the same object, only the one you call the replacement function on will be affected. The replacement function returns a copy that is assigned the first variable and the other variable still refers to the old object.

```
y <- x
foo(x) <- 5
x
## $foo
## [1] 5
##
## $bar
## [1] 3
y
## $foo
## [1] 3
##
## $bar
## [1] 3
```

Because replacement functions are just syntactic sugar on a function call and then a reassignment, you cannot give a replacement function as its first argument, some expression that cannot be assigned to.

There are a few more rules regarding replacement functions. First, the parameter for the value you are assigning has to be called value. You cannot give it another name.

```
`foo<-` <- function(x, val) {
  x$foo <- val
  x
}

x <- list(foo=1, bar=2)
foo(x) <- 3
## Error in `foo<-`(`*tmp*`, value = 3): unused argument (value = 3)
```

The way R rewrites the expression assumes that you called the value parameter value, so do that.

You don't have to call the first parameter x, though:

```
`foo<-` <- function(y, value) {
  y$foo <- value
  y
}

x <- list(foo=1, bar=2)
foo(x) <- 3
x$foo
## [1] 3
```

You should also have the value parameter as the last parameter if you have more than two parameters. And you are allowed to do so, as long as the object you are modifying is the first and the value parameter is the last:

```
`modify<-` <- function(x, variable, value) {
  x[variable] <- value
  x
}

x <- list(foo=1, bar=2)
modify(x, "foo") <- 3
modify(x, "bar") <- 4
x$foo
## [1] 3
x$bar
## [1] 4
```

How Mutable Is Data Anyway?

We just saw that a replacement function creates a new copy so if we use it to modify an object, we are not actually changing it at all. Other variables that refer to the same object will see the old value and not the updated one. So we can reasonably ask: what does it take actually to modify an object?

The short, and almost always correct, answer is that you cannot modify objects *ever*.[2] Whenever you "modify" an object, you are creating a new copy and assigning that new copy back to the variable you used to refer to the old value.

This is also the case for assigning to an index in a vector or list. You will be creating a copy, and while it looks like you are modifying it, if you look at the old object through another reference, you will find that it hasn't changed.

```
x <- 1:4
f <- function(x) {
  x[2] <- 5
  x
}
x
## [1] 1 2 3 4
f(x)
## [1] 1 5 3 4
x
## [1] 1 2 3 4
```

Unless you changed the [function (which I urge you not to do), it is a so-called primitive function. This means that it is written in C and from C you actually *can* modify an object. This is important for efficiency reasons. If there is only one reference to a vector then assigning to it will not make a new copy and you *will* modify the vector in place as a constant time operation. If you have two references to the vector, then when you assign to it the first time, a copy is created that you can then modify in place. This approach to have immutable objects and still have some efficiency is called *copy on write*.

---

[2]It *is* possible to do depending on what you consider an object. You can modify a closure by assigning to local variables inside a function scope, as you saw last week. This is because namespaces are objects that can be changed. One of the object orientation systems in R—RC—also allows for mutable objects, but we won't look at RC in this book. In general, you are better off thinking that every object is immutable, and any modification you are doing is actually creating a new object because generally, that is what is happening.

To write correct programs, always keep in mind that you are not modifying objects but creating copies—other references to the value you "modify" will still see the old value. To write efficient programs, also keep in mind that for primitive functions you *can* do efficient updates (updates in constant time instead of time proportional to the size of the object you are modifying) as long as you only have one reference to that object.

Functional Programming

There are many definitions of what it means for a language to be a functional programming language, and there have been many language wars over whether any given feature is "pure" or not. I won't go into such discussions, but some features, I think everyone would agree, are needed. You should be able to pass functions along as parameters to other functions, you should be able to create anonymous functions, and you should have closures.

Anonymous Functions

In R it is pretty easy to create anonymous functions: just don't assign the function definition to a variable name.
Instead of doing this:

```
square <- function(x) x^2
```

You simply do this:

```
function(x) x^2
```

In other languages where function definitions have a different syntax than variable assignment, you will have a different syntax for anonymous functions, but in R it is really as simple as this.

Why would you want an anonymous function?

There are two common cases:

- You want to use a one-off function and don't need to give it a name

- You want to create a closure

Both cases are typically used when a function is passed as an argument to anther function or when returned from a function. The first case is something you would use together with functions like apply. If you want to compute the sum of squares over the rows of a matrix, you can create a named function and apply it, as follows:

```
m <- matrix(1:6, nrow=3)
sum_of_squares <- function(x) sum(x^2)
apply(m, 1, sum_of_squares)
## [1] 17 29 45
```

If this is the only time you need this sum of squares function, there isn't really any need to assign it a variable; you can just use the function definition direction:

```
apply(m, 1, function(x) sum(x^2))
## [1] 17 29 45
```

Of course, in this example, you could do even better by just exploiting that ^ is vectorized and write this:

```
apply(m^2, 1, sum)
## [1] 17 29 45
```

Using anonymous functions to create closures is what you do when you write a function that returns a function (more about that next). You *could* name the function as follows:

```
f <- function(x) {
  g <- function(y) x + y
  g
}
```

But there really isn't much point if you just want to return it:

```
f <- function(x) function(y) x + y
```

Functions Taking Functions as Arguments

You already saw this in all the apply examples. You give as an argument to apply a function to be called across dimensions. In general, if some sub-computation of a function should be parameterized then you do this by taking a function as one of its parameters.

Say you want to write a function that works like (s/v)apply but only apply an input function on elements that satisfy a predicate. You can implement such a function by taking the vector and two functions as input:

```
apply_if <- function(x, p, f) {
  result <- vector(length=length(x))
  n <- 0
  for (i in seq_along(x)) {
    if (p(x[i])) {
      n <- n + 1
      result[n] <- f(x[i])
    }
  }
  head(result, n)
}
apply_if(1:8, function(x) x %% 2 == 0, function(x) x^2)
## [1]  4 16 36 64
```

This isn't the most elegant way to solve this particular problem—we get back to the example in the exercises—but it illustrates the use of functions as parameters.

Functions Returning Functions (and Closures)

You create closures when you create a function inside another function and return it. Because this inner function can refer to the parameters and local variables inside the surrounding function even after you have returned from it, you can use such inner functions to specialize generic functions. It can work as a template mechanism for describing a family of functions.

You can, for instance, write a generic power function and specialize it for squaring or cubing numbers:

```
power <- function(n) function(x) x^n
square <- power(2)
cube <- power(3)
x <- 1:4
square(x)
## [1]  1  4  9 16
cube(x)
## [1]  1  8 27 64
```

This works because the functions returned by power(2) and power(3) live in a context—the closure—where n is known to be 2 and 3, respectively. We have fixed that part of the function we return.

Filter, Map, and Reduce

Three patterns are used again and again in functional programming: filtering, mapping, and reducing. In R, all three are implemented in different functions, but you can write all your programs using the Filter, Map, and Reduce functions.

The Filter function takes a predicate and a vector or list and returns all the elements that satisfy the predicate.

```
is_even <- function(x) x %% 2 == 0
Filter(is_even, 1:8)
## [1] 2 4 6 8
Filter(is_even, as.list(1:8))
## [[1]]
## [1] 2
##
## [[2]]
## [1] 4
##
## [[3]]
## [1] 6
##
## [[4]]
## [1] 8
```

The Map function works like lapply: it applies a function to every element of a vector or list and returns a list of the result. Use unlist to convert it into a vector if that is what you want.

```
square <- function(x) x^2
Map(square, 1:4)
## [[1]]
## [1] 1
##
## [[2]]
## [1] 4
##
## [[3]]
## [1] 9
```

```
##
## [[4]]
## [1] 16
```
```
unlist(Map(square, 1:4))
## [1]  1  4  9 16
```

You can do slightly more with Map, though, since Map can be applied to more than one sequence. If you give Map more arguments then these are applied to the function calls as well.

```
plus <- function(x, y) x + y
unlist(Map(plus, 0:3, 3:0))
## [1] 3 3 3 3
```

These constructions should be very familiar to you by now so we will leave it at that.

The Reduce function might look less familiar. You can describe what it does in terms of adding or multiplying numbers, and it is in a way a generalization of this. When you write an expression like this:

```
a + b + c
```

or

```
a * b * c
```

You can think of this as a series of function calls:

```
`+`(`+`(a, b), c)
```

or

```
`*`(`*`(a, b), c)
```

The Reduce function generalizes this.

```
Reduce(f, c(a, b, c))
```

It's evaluated as follows:

```
f(f(a, b), c)
```

You can see this by constructing a function that captures how it is called:

```
add_parenthesis <- function(a, b) paste("(", a, ", ", b, ")", sep = "")
Reduce(add_parenthesis, 1:4)
## [1] "(((1, 2), 3), 4)"
```

Using Reduce, you could thus easily write your own sum function:

```
mysum <- function(x) Reduce(`+`, x)
sum(1:4)
## [1] 10
mysum(1:4)
## [1] 10
```

There are a few additional parameters to the Reduce function—to give it an additional initial value instead of just the leftmost elements in the first function call, or to make it apply the function from right to left instead of left to right—but you can check its documentation for details.

Function Operations: Functions as Input and Output

Functions can, of course, take functions as input and return functions as output.

This lets you modify functions and create new functions from existing functions.

First, consider two old friends, the factorial and the Fibonacci numbers. You have computed those recursively and using tables. What if you could build a generic function for caching results?

Here is an attempt:

```
cached <- function(f) {
  force(f)
  table <- list()

  function(n) {
    key <- as.character(n)
    if (key %in% names(table)) {
      print(paste("I have already computed the value for", n))
      table[[key]]

    } else {
      print(paste("Going to compute the value for", n))
      res <- f(n)
      print(paste("That turned out to be", res))
      table[key] <<- res
      print(table)
      res
    }
  }
}
```

I added some output so it is easier to see what it does below.

It takes a function f and will give you another function back that works like f but remembers functions it has already computed. First, it remembers what the input function was by forcing it. This is necessary for the way we intend to use this cached function. The plan is to replace the function in the global scope with a cached version so the function out there will refer to the cached version. If you don't force f here, the lazy evaluation means that when you eventually evaluate f, you are referring to the cached version and will end up in an infinite recursion. You can try removing the force(f) call and see what happens.

Next, we create a table—we are using a list, which is the best choice for tables in R in general. A list lets us use strings for indices, and doing that you don't need to have all values between 1 and *n* stored to have an element with key *n* in the table.

The rest of the code builds a function that first looks in the table to see if the key is there. If it is, you have already computed the value you want and can get it from the table. If the key is not there, you compute it, put it in the table, and return.

You can try it out on the `factorial` function:

```
factorial <- function(n) {
  if (n == 1) {
    1
  } else {
    n * factorial(n - 1)
  }
}

factorial <- cached(factorial)
factorial(4)
## [1] "Going to compute the value for 4"
## [1] "Going to compute the value for 3"
## [1] "Going to compute the value for 2"
## [1] "Going to compute the value for 1"
## [1] "That turned out to be 1"
## $`1`
## [1] 1
##
## [1] "That turned out to be 2"
## $`1`
## [1] 1
##
## $`2`
## [1] 2
##
## [1] "That turned out to be 6"
## $`1`
## [1] 1
##
## $`2`
## [1] 2
##
## $`3`
## [1] 6
##
## [1] "That turned out to be 24"
## $`1`
## [1] 1
##
## $`2`
## [1] 2
##
## $`3`
## [1] 6
##
## $`4`
## [1] 24
## [1] 24
```

```
factorial(1)
## [1] "I have already computed the value for 1"
## [1] 1
factorial(2)
## [1] "I have already computed the value for 2"
## [1] 2
factorial(3)
## [1] "I have already computed the value for 3"
## [1] 6
factorial(4)
## [1] "I have already computed the value for 4"
## [1] 24
```

And on fibonacci:

```
fibonacci <- function(n) {
  if (n == 1 || n == 2) {
    1
  } else {
    fibonacci(n-1) + fibonacci(n-2)
  }
}
```

```
fibonacci <- cached(fibonacci)
fibonacci(4)
## [1] "Going to compute the value for 4"
## [1] "Going to compute the value for 3"
## [1] "Going to compute the value for 2"
## [1] "That turned out to be 1"
## $`2`
## [1] 1
##
## [1] "Going to compute the value for 1"
## [1] "That turned out to be 1"
## $`2`
## [1] 1
##
## $`1`
## [1] 1
##
## [1] "That turned out to be 2"
## $`2`
## [1] 1
##
## $`1`
## [1] 1
##
## $`3`
## [1] 2
##
## [1] "I have already computed the value for 2"
```

```
## [1] "That turned out to be 3"
## $`2`
## [1] 1
##
## $`1`
## [1] 1
##
## $`3`
## [1] 2
##
## $`4`
## [1] 3
## [1] 3
fibonacci(1)
## [1] "I have already computed the value for 1"
## [1] 1
fibonacci(2)
## [1] "I have already computed the value for 2"
## [1] 1
fibonacci(3)
## [1] "I have already computed the value for 3"
## [1] 2
fibonacci(4)
## [1] "I have already computed the value for 4"
## [1] 3
```

Ellipsis Parameters

Before you see any more examples of function operations, you need to know about a special function parameter, the ellipsis or "three-dots" parameter.

This is a magical parameter that lets you write a function that can take any number of named arguments and pass them on to other functions.

Without it, you would get an error if you provided a parameter to a function that it doesn't know about.

```
f <- function(a, b) NULL
f(a = 1, b = 2, c = 3)
## Error in f(a = 1, b = 2, c = 3): unused argument (c = 3)
```

With it, you can provide any named parameter you want.

```
g <- function(a, b, ...) NULL
g(a = 1, b = 2, c = 3)
## NULL
```

Of course, it isn't much of a feature to allow a function to take arguments that it doesn't know what to do with. But you can pass those arguments to other functions that maybe *do* know what to do with them, and that is the purpose of the … parameter.

You can see this in effect with a very simple function that just passes the ... parameter on to list. This works exactly like calling list directly with the same parameters, so nothing magical is going on here, but it shows how the named parameters are being passed along.

```r
tolist <- function(...) list(...)

tolist()
## list()
tolist(a = 1)
## $a
## [1] 1
tolist(a = 1, b = 2)
## $a
## [1] 1
##
## $b
## [1] 2
```

This parameter has some uses in itself because it lets you write a function that calls other functions, and you can provide those functions parameters without explicitly passing them along. It is particularly important for generic functions (a topic we cover in the next chapter) and for modifying functions in function operators.

Most of what you can do with function operators is beyond the scope of this book, so if you are interested in learning more, you should check out the chapter about them in Hadley Wickham's *Advanced R Programming* book (see http://adv-r.had.co.nz/Function-operators.html).

Here we will just have a quick second example, taken from *Advanced R Programming*, that modifies a function. It wraps a function to time how long it takes to run.

The following function wraps the function f into a function that times it and returns the time usage rather than the result of the function. It will work for any function since it just passes all parameters from the closure we create to the function we wrap (although the error profile will be different since the wrapping function will accept *any* named parameter while the original function f might not allow that).

```r
time_it <- function(f) {
  force(f)
  function(...) {
    system.time(f(...))
  }
}
```

You can try it out like this:

```r
ti_mean <- time_it(mean)
ti_mean(runif(1e6))
##    user  system elapsed
##   0.025   0.002   0.026
```

Exercises

Try the following exercises to become more comfortable with the concepts discussed in this chapter.

between

Write a vectorized function that takes a vector x and two numbers, lower and upper, and replaces all elements in x smaller than lower or greater than upper with NA.

apply_if

Consider the function apply_if you implemented in this chapter. There we use a loop. Implement it using Filter and Map instead.

For the specific instance we used in the example:

```
apply_if(v, function(x) x %% 2 == 0, function(x) x^2)
```

We only have vectorized functions. Rewrite this function call using a vectorized expression.

power

We previously defined the generic power function and the instances square and cube this way:

```
power <- function(n) function(x) x^n
square <- power(2)
cube <- power(3)
```

If you instead defined this:

```
power <- function(x, n) x^n
```

How would you then define square and cube?

Row and Column Sums

Using apply, write the rowsum and colsum functions to compute the row sums and column sums, respectively, of a matrix.

Factorial Again

Write a vectorized factorial function. It should take a vector as input and compute the factorial of each element in the vector.

Try to make a version that remembers factorials it has already computed so you don't need to recompute them (without using the cached function, of course).

Function Composition

For two functions *f* and *g*, the function composition creates a new function $f \circ g$ such that $(f \circ g)(x) = f(g(x))$.

There isn't an operator for this in R, but you can make your own. To avoid clashing with the outer product operator, %o%, call it %.%.

Implement this operator.

Using this operator, you should, for example, be able to combine `Map` and `unlist` once and for all to get a function for the `unlist(Map(...))` pattern

```
uMap <- unlist %.% Map
```

So this function works exactly like first calling `Map` and then `unlist`:

```
plus <- function(x, y) x + y
unlist(Map(plus, 0:3, 3:0))
## [1] 3 3 3 3
uMap(plus, 0:3, 3:0)
## [1] 3 3 3 3
```

With it, you can build functions by stringing together other functions (not unlike how you can create pipelines in magrittr—see https://cran.r-project.org/web/packages/magrittr/vignettes/magrittr.html).

For example, you can compute the root mean square error function like this:

```
error <- function(truth) function(x) x - truth
square <- function(x) x^2

rmse <- function(truth)
  sqrt %.% mean %.% square %.% error(truth)

mu <- 0.4
x <- rnorm(10, mean = 0.4)
rmse(mu)(x)
## [1] 0.8976526
```

Combining a sequence of functions like this requires that you read the operations from right to left, so I personally prefer the approach in magrittr, but you can see the similarity.

■ ■ ■

Object Oriented Programming

This chapter looks at R's flavor of object oriented programming. Actually, R has three different systems for object oriented programming: S3, S4, and RC. We will only look at S3, which is the simplest and (I believe) the most widely used.

Immutable Objects and Polymorphic Functions

Object orientation in S3 is quite different from what you might have seen in Java or Python. Naturally so, since data in R is immutable and the underlying model in OO in languages such as Java and Python is that you have objects with states that you can call methods to change the state. You don't have a state as such in S3; you have immutable objects. Just like all other data in R.

What's the point then, of having object orientation if we don't have object states? What we get from the S3 system is polymorphic functions, called "generic" functions in R. These are functions whose functionality depends on the class of an object—similar to methods in Java or Python where methods defined in a class can be changed in a subclass to refine behavior.

You can define a function `foo` to be polymorphic and then define specialized functions, say `foo.A` and `foo.B`. Then calling `foo(x)` on an object `x` from class `A` will actually call `foo.A(x)` and for an object from class `B` will actually call `foo.B(x)`. The names `foo.A` and `foo.B` were not chosen at random here, as you will see, since it is precisely how you name functions that determine which function is called.

We do not have objects with states; we simply have a mechanism for enabling a function to depend on the class an object has. This is often called "dynamic dispatch" or "polymorphic methods". Here of course, since we don't have states, we can call it polymorphic functions.

Data Structures

Before we get to making actual classes and objects, though, we should look at data structures. We discussed the various built-in data structures in R in Chapters 1 and 8. Those built-in data types are the basic building blocks of data in R, but we never discussed how we can build something more complex from them.

More important than any object oriented system is the idea of keeping related data together so we can treat it as a whole. If we are working on several pieces of data that somehow belongs together, we don't want it scattered out in several different variables, perhaps in different scopes, where we have little chance of keeping it consistent. Even with immutable data, keeping the data that different variables refer to would be a nightmare.

For data we analyze, we therefore typically keep it in a data frame. This is a simple idea for keeping data together. All the data we are working on is in the same data frame, and we can call functions with the data frame and know that they are getting all the data in a consistent state. At least as consistent as we can guarantee with data frames; we cannot promise that the data itself is not messed up somehow, but we can write functions under the assumption that data frames behave a certain way.

What about something like a fitted model? If we fit a model to some data, that fit is stored variables capturing the fit. We certainly would like to keep those together when we do work with the model because we would not like accidentally to use a mix of variables fitted to two different models. We might also want to keep other data together with the fitted model—e.g., some information about what was actually fitted—if we want to check that in the R shell later. Or the data it was fitted to.

The only option we have for collecting heterogeneous data together as a single object is a list. And that is how you do it in R.

Example: Bayesian Linear Model Fitting

Project two, described in the last chapter of the book, concerns Bayesian linear models. To represent such, we would wrap data for a model in a list. For fitting data, assume that you have a function like the one described here (refer to Chapter 9 for details of the mathematics).

It takes the model specification in the form of a formula as its parameter model and the prior precision alpha and the "precision" of the data beta. It then computes the mean and the covariance matrix for the model fitted to the data. The mathematics behind the code is explained in Chapter 9. It then wraps up the fitted model together with some related data—the formula used to fit the model and the data used in the model fit (here assumed to be in the variable frame)—and puts them in a list, which the function returns.

```r
blm <- function(model, alpha = 1, beta = 1, ...) {

    # Here goes the mathematics for computing the fit.
    frame <- model.frame(model, ...)
    phi <- model.matrix(frame)
    no_params <- ncol(phi)
    target <- model.response(frame)

    covar <- solve(diag(alpha, no_params) +
                   beta * t(phi) %*% phi)
    mean <- beta * covar %*% t(phi) %*% target

    list(formula = model,
         frame = frame,
         mean = mean,
         covar = covar)
}
```

You can see it in action by simulating some data and calling the function:

```r
# fake some data for our linear model
x <- rnorm(10)
a <- 1 ; b <- 1.3
w0 <- 0.2 ; w1 <- 3
y <- rnorm(10, mean = w0 + w1 * x, sd = sqrt(1/b))

# fit a model
model <- blm(y ~ x, alpha = a, beta = b)
model
## $formula
## y ~ x
##
```

```
## $frame
##             y              x
## 1    5.9784195   1.73343698
## 2    0.5044947  -0.45442222
## 3   -3.6050449  -1.47534377
## 4    1.7420036   0.81883381
## 5   -0.9105827   0.03838943
## 6   -3.1266983  -1.14989951
## 7    5.9018405   1.78225548
## 8    2.2878459   1.29476972
## 9    1.0121812   0.39513461
## 10  -1.7562905  -0.72161442
##
## $mean
##                   [,1]
## (Intercept) 0.2063805
## x           2.5671043
##
## $covar
##             (Intercept)           x
## (Intercept)  0.07399730 -0.01223202
## x           -0.01223202  0.05824769
```

It collects the relevant data of a model fit like this together in a list, so we always know we are working on the values that belong together. This makes further analysis of the fitted model *much* easier to program.

Classes

The output we got when we wrote:

```
model
```

is what we get if we call the print function on a list. It just shows us everything that is contained in the list. The print function is an example of a polymorphic function, however, so when you call print(x) on an object x, the behavior depends on the *class* of the object x.

If you want to know what class an object has, you can use the class function:

```
class(model)
## [1] "list"
```

If you want to change it, you can use the class<- replacement function:

```
class(model) <- "blm"
```

You can use any name for a class; here I've used blm for Bayesian linear model.

By convention, we usually call the class and the function that creates elements of that class the same name, so since we are creating this type of object with the blm function, convention demands that we call the class of the object blm as well. It is just a convention, though, and you can call the class anything.

We can always assign a class to an object in this way, but changing the class of an existing object is considered bad style. We keep the data that belongs together in a list to make sure that the data is consistent, but the functionality we want to provide for a class is as much a part of the class as the data, so we also need to make sure that the functions that operate on objects of a given class always get data that is consistent with that class. We cannot do that if we go around changing the class of objects willy-nilly.

The function that creates the object should assign the class and then we should leave the class of the object alone. We can set the class with the `class<-` function and then return it using the `blm` function.

```
blm <- function(model, alpha = 1, beta = 1, ...) {

  # stuff happens here...

  object <- list(formula = model,
                 frame = frame,
                 mean = mean,
                 covar = covar)
  class(object) <- "blm"
  object
}
```

The class is represented by an attribute of the object; however, and there is a function that sets these for us, called `structure`, and using that we can create the object and set the class at the same time, which is a little better.

```
blm <- function(model, alpha = 1, beta = 1, ...) {

  # stuff happens here...

  structure(list(formula = model,
                 frame = frame,
                 mean = mean,
                 covar = covar),
            class = "blm")
}
```

Now that we gave the `model` object a class, let's try printing it again.

```
model
## $formula
## y ~ x
##
## $frame
##            y           x
## 1   5.9784195   1.73343698
## 2   0.5044947  -0.45442222
## 3  -3.6050449  -1.47534377
## 4   1.7420036   0.81883381
## 5  -0.9105827   0.03838943
## 6  -3.1266983  -1.14989951
## 7   5.9018405   1.78225548
## 8   2.2878459   1.29476972
```

```
## 9   1.0121812  0.39513461
## 10 -1.7562905 -0.72161442
##
## $mean
##                  [,1]
## (Intercept) 0.2063805
## x           2.5671043
##
## $covar
##             (Intercept)          x
## (Intercept)  0.07399730 -0.01223202
## x           -0.01223202  0.05824769
##
## attr(,"class")
## [1] "blm"
```

The only difference from before is that it has added information about the `"class"` attribute toward the end. It still just prints everything that is contained in the object. This is because we haven't told it to treat any object of class `blm` any differently yet.

Polymorphic Functions

The `print` function is a *polymorphic function*. This means that what happens when it is called depends on the class of its first parameter. When we call `print`, R will get the class of the object, let's say it is `blm` as in our case, and see if it can find a function named `print.blm`. If it can, then it will call this function with the parameters you called `print` with. If it cannot, it will instead try to find the function `print.default` and call that.

We haven't defined a `print` function for the class `blm`, so we saw the output of the default `print` function instead.

Let's try to define a `blm`-specific `print` function.

```
print.blm <- function(x, ...) {
  print(x$formula)
}
```

Here, we just tell it to print the formula we used for specifying the model rather than the full collection of data we put in the list.

If we print the model *now*, this is what happens:

```
model
## y ~ x
```

That is how easy it is to provide your own class-specific `print` function. And that is how easy it is to define your own class-specific polymorphic function in general. You just take the function name and append `.classname` to it, and if you define a function with that name, then that function will be called when you call a polymorphic function on an object with that class.

One thing you do have to be careful about, though, is the interface to the function. By that I mean the parameters the function takes (and their order). Each polymorphic function takes some arguments. You can see which by checking the function documentation.

```
?print
```

When you define your specialized function, you can add more parameters to your function, but you should define it such that you at least take the same parameters as the generic function does. R will not complain if you do not define it that way, but it is bound to lead to problems later when someone calls the function with assumptions about which parameters it takes based on the generic interface and then runs into your implementation of a specialized function that behaves a different way. Don't do that. Implement your function so it takes the same parameters as the generic function. This includes using the same names for parameters. Someone might provide named parameters to the generic function, and that will work only if you call the parameters the same names as the generic function. That is why we used x as the input parameter for the `print.blm` function.

Defining Your Own Polymorphic Functions

To define a class-specific version of a polymorphic function, you just need to write a function with the right name. There is a little bit more to do if you want to define your very own polymorphic function. Then you also need to write the generic function—the function you will actually call with objects, and that is responsible for dispatching the function call to class-specific functions.

You do this using the `UseMethod` function. The generic function typically just does this and looks like this:

```
foo <- function(x, ...) UseMethod("foo")
```

You specify a function with the parameters the generic function should accept and then just call `UseMethod` with the name of the function to dispatch to. Then it does it magic and finds out which class-specific function to call and forwards the parameters to there.

When you write the generic function, it is also good style to define the default function as well.

```
foo.default <- function(x, ...) print("default foo")
```

With that, we can call the function with all types of objects. If you don't want that to be possible, a safe default function would be one that throws an error.

```
foo("a string")
## [1] "default foo"
foo(12)
## [1] "default foo"
```

And of course, with the generic function in place, we can define class-specific functions just like before.

```
foo.blm <- function(x, ...) print("blm foo")
foo(model)
## [1] "blm foo"
```

You can add more parameters to more specialized functions when the generic function takes … as an argument. The generic function will just ignore the extra parameters, but the concrete function that is called might be able to do something about it.

```
foo.blm <- function(x, upper = FALSE, ...) {
  if (upper) {
    print("BLM FOO")
  } else {
    print("blm foo")
  }
}
```

262

```
foo("a string")
## [1] "default foo"
foo(model)
## [1] "blm foo"
foo("a string", upper = TRUE)
## [1] "default foo"
foo(model, upper = TRUE)
## [1] "BLM FOO"
```

Class Hierarchies

Polymorphic functions are one aspect of object oriented programming, another is inheritance. This is the mechanism used to build more specialized classes out of more general classes.

The best way to think about this is as levels of specialization. You have some general class of objects, say furniture, and within that class are more specific categories, say chairs, and within *that* class even more specific types of objects, say kitchen chairs. A kitchen chair is also a chair, and a chair is also furniture. If there is something you can do to *all* furniture, then you can definitely also do it to chairs. For example, you can set furniture on fire; you can set a chair on fire. It is not the case, however, that everything you can do to chairs you can do to all furniture. You can throw a chair at unwelcome guests, but you cannot throw a piano at them.

The way specialization like this works is that there are some operations you can do for the general classes. Those operations can be done on all instances of those classes, including those that are really objects of more specialized classes.

The operations might not do exactly the same thing—like we can specialize print, an operation we can call on *all* objects, to do something special for blm objects—but there is some meaningful way of doing the operation. Quite often the way a class is specialized is exactly by doing an operation that can be done by all objects from the general class, but just in a more specialized way.

The specialized classes, however, can potentially do more so they might have more operations that are meaningful to do to them. That is fine. As long as we can treat all objects of a specialized class the same as we can treat objects of the more general class.

This kind of specialization is partly interface and partly implementation.

Specialization as Interface

The interface is which functions we can call on objects of a given class. It is a kind of protocol for how we interact with objects of the class. If we imagine some general class of "fitted models," we might specify that for all models we should be able to get the fitted parameters and for all models we should be able to make predictions for new values. In R, such functions exist, called coef and predict, and any model is expected to implement them.

This means that I can write code that interacts with a fitted model through these general model functions, and as long as I stick to the interface they provide, I could be working on *any* kind of model. If, at some point, I find out that I want to replace a linear regression model with a decision tree regression model, I can just plug in a different fitted model and communicate with it through the same polymorphic functions. The actual functions that will be called when I call the generic functions coef and predict will, of course, be different, but the interface is the same.

R will not enforce such interfaces for you. Classes in R are not typed in the same way as they are in, for example, Java, where it would be a type error to declare something as an object satisfying a certain interface if it does in fact not. R doesn't care. Not until you call a function that isn't there; then you might be in trouble, of course. But it is up to you to implement an interface to fit the kind of class or protocol you think your class should match.

If you implement the functions that a certain interface expects (and these functions actually do something resembling what the interface expects the functions to do and are not just named the same things),[1] you have a specialization of that interface. You can do the same operations as every other class that implements the interface, but of course, your operations are uniquely fitted to your actual class.

You might implement more functions, making your class capable of more than the more general class of objects, but that is just fine. Other classes might implement those operations as well, so now you have more than one class with the more specialized operations—a new category that is more general and can be specialized further.

You have a hierarchy of classes defined by which functions they provide implementations of.

Specialization in Implementations

Specialization by providing general or more specialized interfaces—in the case of R by providing implementations of polymorphic functions—is the essential feature of the concept of class hierarchies in object oriented programming. It is what lets you treat objects of different kinds as a more general class.

There is another aspect of class hierarchies, though, that has to do with code reuse. You already get a lot of this just by providing interfaces to work with objects, of course, since you can write code that works on a general interface and then reuse it on all objects that implement this interface. But there is another type of reuse you get when you build a hierarchy of classes where you go from abstract, general classes to more specialized and concrete classes. When you are specializing a class, you are taking functionality that exists for the more abstract class and defining a new class that implements the same interface *except* for a few differences here and there.

When you refine a class in this way, you don't want to implement new versions of all the polymorphic functions in its interface. Many of them will do exactly the same as the implementation for their more general class.

Let's say you want to have a class of objects where you can call functions foo and bar. You can call that class A and define it as follows:

```
foo <- function(object, ...) UseMethod("foo")
foo.default <- function(object, ...) stop("foo not implemented")

bar <- function(object, ...) UseMethod("bar")
bar.default <- function(object, ...) stop("bar not implemented")

A <- function(f, b) structure(list(foo = f, bar = b), class = "A")
foo.A <- function(object, ...) paste("A::foo ->", object$foo)
bar.A <- function(object, ...) paste("A::bar ->", object$bar)

a <- A("qux", "qax")
foo(a)
## [1] "A::foo -> qux"
bar(a)
## [1] "A::bar -> qax"
```

For a refinement of that, you might want to change how bar works and add another function called baz:

```
baz <- function(object, ...) UseMethod("baz")
baz.default <- function(object, ...) stop("baz not implemented")
```

---

[1] To *draw* means something very different when you are a gunslinger compared to when you are an artist, after all.

```
B <- function(f, b, bb) {
  a <- A(f, b)
  a$baz <- bb
  class(a) <- "B"
  a
}

bar.B <- function(object, ...) paste("B::bar ->", object$bar)
baz.B <- function(object, ...) paste("B::baz ->", object$baz)
```

We want to leave the foo function just the way it is, but if we define the class B as shown, calling foo on a B object gives us an error because it will be calling the foo.default function.

```
b <- B("qux", "qax", "quux")
foo(b)
## Error in foo.default(b): foo not implemented
```

This is because we haven't told R that we consider the class B a specialization of class A. We wrote the constructor function—the function where we make the object, the function B—such that all B objects contain the data that is also found in an A object. We never told R that we intended B objects also to be A objects.

We could, of course, make sure that foo called on a B object behaves the same way as when called on an A object by defining foo.B such that it calls foo.A. This wouldn't be too much work for a single function, but if there are many polymorphic functions that work on A objects, we would have to implement B versions for all of them. Tedious and error-prone work.

If only there were a way of telling R that the class B is really an extension of the class A. And there is. The class attribute of an object doesn't have to be a string. It can be a vector of strings. If, for B objects, we say that the class is B first and A second, like this,

```
B <- function(f, b, bb) {
  a <- A(f, b)
  a$baz <- bb
  class(a) <- c("B", "A")
  a
}
```

Then calling foo on a B object—where foo.B is not defined—will call foo.A as its second choice and before defaulting to foo.default:

```
b <- B("qux", "qax", "quux")
foo(b)
## [1] "A::foo -> qux"
bar(b)
## [1] "B::bar -> qax"
baz(b)
## [1] "B::baz -> quux"
```

The way the class attribute is used with polymorphic functions is that R will look for functions with the class names appended in the order of the class attributes. The first it finds will be the one that is called, and if it finds no specialized version, it will go for the .default version. When we set the class of B objects to be the vector c("B", "A"), we are saying that R should call .B functions first, if it can find one, but otherwise call the .A function.

This is a very flexible system that lets you implement multiple inheritances from classes that are otherwise not related, but you do so at your own peril. The semantics of how this works—functions are searched for in the order of the class names in the vector—the actual code that will be run can be hard to work out if these vectors get too complicated.

Another quick word of caution is this: if you give an object a list of classes, you should include the classes all the way up the class hierarchy. If you define a new class, C, intended as a specialization of B, we cannot just say that it is an object of class c("C", "B") if we also want it to behave like an A object.

```
C <- function(f, b, bb) {
  b <- B(f, b, bb)
  class(b) <- c("C", "B")
  b
}

c <- C("foo", "bar", "baz")
foo(c)
## Error in foo.default(c): foo not implemented
```

When we call foo(c) here, R will try the functions, in turn: foo.C, foo.B, and foo.default. The only one that is defined is the last, and that throws an error if called.

So what we have defined here is an object that can behave like B but only in cases where B differs from A's behavior! Our intention is that B is a special type of A, so every object that is a B object we should also be able to treat as an A object. Well, with C objects that doesn't work.

We don't have a real class hierarchy here like we would find in languages like Python, C++, or Java. We just have a mechanism for calling polymorphic functions, and the semantic here is just to look for them by appending the names of the classes found in the class attribute vector. Your intentions might very well be that you have a class hierarchy with A being the most general class, B a specialization of that, and C the most specialized class, but that is not what you are telling R. Because you cannot. You are telling R how it should look for dynamic functions, and with the code, you told it to look for .C functions first, then .B functions, and you didn't tell it any more, so the next step it will take is to look for .default functions. Not .A functions. It doesn't know that this is where you want it to look.

If you add this to the class attribute it will work, though:

```
C <- function(f, b, bb) {
  b <- B(f, b, bb)
  class(b) <- c("C", "B", "A")
  b
}

c <- C("foo", "bar", "baz")
foo(c)
## [1] "A::foo -> foo"
bar(c)
## [1] "B::bar -> bar"
baz(c)
## [1] "B::baz -> baz"
```

You are slightly better off getting the class attribute from the object you create in the constructor, though. If, at some point, you changed the class attribute of the object returned from the B() constructor, you don't want to have to change the class vector in all classes that are extending the class.

```
C <- function(f, b, bb) {
  b <- B(f, b, bb)
  class(b) <- c("C", class(b))
  b
}
```

Exercises

Try the following exercises to become more comfortable with the concepts discussed in this chapter.

Shapes

Imagine that we need to handle some geometric shapes for a program. These could be circles, squares, triangles, etc. We need to know the shapes' circumferences and areas. These properties can be calculated from properties of the shapes, but the calculations are different for each shape.

So for our shapes, we want (at least) an interface that gives us two functions: circumference and area. The default functions, where we have no additional information about an object aside from the fact that it is a shape, are meaningless so should raise an error (check the stop function for this), but each specialized shape should implement these two functions.

Implement this protocol/interface and the two functions for at least circles and rectangles; by all means, more shapes if you want to.

Polynomials

Write a class that lets you represent polynomial objects. An n-degree polynomial is on the form $c_0 + c_1 x + c_2 x^2 + \ldots + c_n x^n$ and can be represented by the $n+1$ coefficients $(c_0, c_1, \ldots, c_n)$. Write the interface such that you can evaluate polynomials in any point x, i.e. with a function evaluate_polynomial(poly, x) that gives you the value of the polynomial at the point x.

The function uniroot (built into R) lets you find the roots of a general function. Use it to write a function that finds the roots of your polynomials. This function works by numerically finding the points where the polynomial is zero. For lines and quadratic polynomials, though, there are analytical solutions. Write special cases for polynomials such that calling the root finding function on the special cases exploits that solutions are known there.

■ ■ ■

Building an R Package

You now know how to write functions and create classes in R, but neither functions nor classes is the unit you use for collecting and distributing R code. That unit is the *package*. It is packages that you load and import into your namespace when you write this:

`library``(something)`

And it is packages you download when you write this:

`install.packages``("something")`

The topic of this chapter is how to make your own packages. In the space available, I can only give a very broad overview of the structure of R packages, but it should be enough to get you started. If you want to read more, I warmly recommend Hadley Wickham's book *R Packages*.

Creating an R Package

I am going to assume that you use RStudio for this. If you don't, you can look at the devtools package. It provides functions for doing everything you can do through the GUI in RStudio.

To create a new package, choose File ➤ New Project and you should get a dialog box that asks you whether your new project should be in a new director, in an existing directory, or checked out of a version control repository. Pick the New Directory.

After that, you get the choice between an empty project, a package, or a shiny application. Not surprisingly, you want to pick R Package.

Now you get to a dialog box where you can set the details of the package. You can choose the Type of the package (where you can choose between a plain package or one that uses Rcpp to make C++ extensions); you can specify the Name of the package; and you can provide existing source files to include in the package. Further, you need to choose a location to put the new package and whether you want to use a git repository for the package.

Choose a plain package and click Yes to create a git repository (we return to git later). You now just need to pick a name and a place to put your package. Where you put it is up to you, but there are some guidelines for package names, discussed next.

Package Names

A package name can consist of letters, numbers, and `.`, but *must* start with a letter and must not have `.` as the last character. You cannot use other characters, such as underscores or dashes.

Whenever you build software that you intend for other people to be able to use, be careful with the name you give it. Give it a name that is easy to remember and easy to Google.

For experimenting with packages, you can just create one called `test`. Create it and have a look at the result.

The Structure of an R Package

In the directory that RStudio built for you, you should have two directories, `R` and `man`, three text files, `.Rbuildignore`, `DESCRIPTION`, and `NAMESPACE`, and one project file (its name will be the name of your package followed by `.Rproj`).

The last of these files is used by RStudio, and all you need to know about it is that if you open this file in RStudio, you get an open version of the state of the project you had last time you worked on it.

Inside the R directory you have an example file, called `R/hello.R`, and inside the `man` directory, you have an example documentation[1] file, called `man/hello.Rd`.

The text files and the two directories are part of what an R package looks like, and they must always be there with exactly those names. There are a few more directories that also have standard names,[2] but they are not required, and we don't have them here for now.

.Rbuildignore

The directory you created contains the source code for the package, but it isn't the *actual* package. The package is something you need to build and install from this source code. We will get to how to do that shortly.

The `.Buildignore` file tells R what not to include when it builds a package. Files that are not mentioned here will automatically be included. This isn't a disaster as such, but it does lead to messy packages for others to use, and if you upload a package to CRAN,[3] the filters there enforce a strict directory and file structure and you will not be allowed to include files or directories that do not follow that structure.

The automatically generated `.Buildignore` file looks like this:

```
^.*\.Rproj$
^\.Rproj\.user$
```

These are two regular expressions that prevent R from including the RStudio files in compiled packages.

The `^` character here matches the beginning of a filename, while `$` matches the end. A non-escaped `.` matches any character, while an escaped `\.` matches an actual dot. The `*` specifies that the previous symbol can be repeated any number of times. So the first regular expression specifies any filename that ends in `.Rproj` and the second expression specifies any filename that ends in `.Rproj.user`.

---

[1]`man` stands for *manual* and the abbreviation `man` is a legacy from UNIX.
[2]e.g., `vignettes/` for documentation vignettes, `data/` for data you want to include with your package, and `src/` for C/C++ extensions.
[3]CRAN is the official depository for R package and the place where the `install.packages` function finds them.

Description

This file contains meta-information about your package. If you called your package `test` and created it the same day I did (November 11 2015), it should now look like this:

```
Package: test
Type: Package
Title: What the Package Does (Title Case)
Version: 0.1
Date: 2015-11-22
Author: Who wrote it
Maintainer: Who to complain to <yourfault@somewhere.net>
Description: More about what it does (maybe more than one line)
License: What license is it under?
LazyData: TRUE
```

You need to update it to describe your new package.

I give a short description of the metadata next, but you can also read more about it in Hadley Wickham's *R Packages* book.

Title

The title field is pretty self-explanatory. You need to give your package a title. Here, (`Title Case`) means that you need to use capital first letters in the words there like you would for the title of a book.

If you read the documentation for a package on CRAN, it will look like this: `packagename: This is the Title`. Don't include the package name in your title here, because that is automatically added to the documentation page. You just want the title.

Version

This is just a number to track which version of your package people have installed. Whenever you make changes to your package and release them, this number should go up.

The version numbers are not only used to indicate that you have updated a version, but they are also necessary for specifying dependencies between packages sometimes. If a feature was introduced in version 1.2 but didn't exist in version 1.1, then other packages that use this feature need to know whether they have access to version 1.2 or higher. We return to dependencies later in this chapter.

There are some conventions for version numbers but nothing that is strictly enforced. The convention here is that a *released* version has the numbering scheme `major.minor.patch`, so the version `1.2.3` means that the major version number is 1, the minor 2, and that this is patched version 3. Patches are smaller changes, typically bug fixes and such, while minor revisions usually include some new functionality. The difference between what is considered minor and major is subjective, but any time the interface changes— you change the way a function is called such that the old types of calls are now incorrect—you definitely should increase the major version number.

If you have a development version of your package that you are distributing for those adventurous enough to work with a beta release, the convention is to add a development release number as well. Then the version number looks like `major.minor.patch.develop-number`, where by convention the last number starts at `9000` and is increased with every new release.

You are just beginning developing your new package, so change the version number to `0.0.0.9000`.

Description

This field should describe the package. It is typically a one-paragraph short description. To make R parse the `DESCRIPTION` file correctly, you must indent the lines following `Description:` if the description spans over multiple lines.

Author and Maintainer

Delete these two fields. There is a better way to specify the same information that makes sure that it is provided in a more structured form. You should use a new field called `Authors@R:` instead.

This field takes an R expression specifying one or more authors, where the author information is provided by a call to the function `person`—which is how we make sure that it is structured appropriately. Check the documentation for the function (`?person`) for more details.

You are single author, so you should use something like this:

```
Authors@R: person("First Name", "Last Name",
                  email = "your.email@your.domain.com",
                  role = c("aut", "cre"))
```

The roles here means *author* and *creator*. The documentation for the `person` function lists other options.

If there is more than one person involved as author or maintainer or another sort of contributor, you can list a sequence of `persons` by concatenating them with the `c` function.

License

This specifies the software license the package is released under. It can really be anything, but if you want to put your package on CRAN, you have to pick one of the licenses that CRAN accepts from https://cran.r-project.org/web/licenses/.

You specify which of the recognized licenses by their abbreviation, so to specify that your package is released under the GPL version 2 license, you write:

```
License: GPL-2
```

Type, Date, LazyData

The `Type` and `LazyData` fields are not essential. You can delete them if you want. `Type` is just saying that you have a package but we sort of know that already. `LazyData` tells R that if you include data in your package, it should load it lazily. Again, this is not something that is of particular importance (unless you plan to include extremely large datasets with your package; if you do that, Google for the documentation of `LazyData`).

The `Date` of course includes the date. This should be the last time you modified the package, i.e., the last time you updated the version.

URL and BugReports

If you have a web page for the package and a URL for reporting bugs, these are the fields you want to use. They are not required for a package, but are of course very helpful for the users to have.

Dependencies

If your package has dependencies, you have three fields you can specify them in: Depends, Imports, and Suggests.[4]

With Depends, you can specify both packages that need to be installed for your package to work and which version of R is required for your package to work. For packages, though, it is better to use Imports and Suggests than Depends, so use Depends only to specify which version of R you need.

You specify it this like:

```
Depends: R (>= 2.10)
```

This is saying that you need R to work (not surprisingly, but the syntax is the same for packages) and it has to be at least version 2.10.

The syntax for dependencies is a comma-separated list of package names (or R as shown previously) with optional version number requirements in parentheses after the package name.

Imports and Suggests fields could look like this:

```
Imports:
  ggplot2,
  dplyr (>= 0.4.3),
  pracma
Suggests:
  testthat,
  knitr
```

This specifies that you import three packages—ggplot2, dplyr, and pracma—and use testthat and knitr in some functions if these packages are available. We require that dplyr has at least version 0.4.3 but do not place any demands on the versions of the other packages.

The difference between Imports and Suggests is that requirements in Imports *must* be installed for your package to be installed (or they will be installed if you tell R to install with dependencies), while requirements in Suggests do not.

Using an Imported Package

Packages in the Imports or Suggests lists are not imported into your namespace the way they would be if you call library(package). This is to avoid contaminating your package namespace, and you shouldn't break that by calling library yourself. If you want to use functions from other packages, you must do so by explicitly accessing them through their package namespace or by explicitly importing them at a single-function level.

The way to access a function from another package without importing the package namespace is using the :: notation. If you want to get to the filter function in dplyr without importing dplyr, you can get the function using the name dplyr::filter.

If you access names from a package that you have listed in your Imports field, you know that it exists even if it isn't imported into your namespace, so you just need to use the long name.

An alternative way of importing functions is using Roxygen—which we will discuss later in the chapter—where you can import the namespace of another package or just the name of a single function in another package for a single function at a time.

---

[4]There are a few more fields, such as for linking to external C/C++ code, but these three fields are the most important ones.

Using a Suggested Package

You access functions in a suggested package—the packages named in the Suggests field—using the `::` notation, just as you do for imported packages. There is just one more complication: the package might not be installed on the computer where your package is installed. That is the difference between suggesting a dependency and requiring it by putting it in the Imports field.

The purpose of suggesting packages instead of importing them is that the functionality your package provides doesn't strictly depend on the other package, but you can do more, or do things more efficiently, if a suggested package is there.

So you need a way of checking if a package is installed before you use it and that way is the function `requireNamespace`. It returns TRUE if the namespace (package) you ask for is installed and FALSE otherwise. You can use it like this:

```r
if (requireNamespace("package", quietly = TRUE)) {
    # use package functionality
} else {
    # do something that doesn't involve the package
    # or give up and throw an exception with stop()
}
```

The `quietly` option is to prevent it from printing warnings. You are handling the cases where the package is not installed, so there is no need for it to print warnings.

NAMESPACE

The NAMESPACE file provides information about which of the functions you implement in your package should be exported to the namespace of the users when they write library(test).

Each package has its own namespace. It is similar to how each function has a namespace in its body where we can define and access local variables. Functions you write in a package will look for other functions first in the package namespace and then in the global namespace.

Users who want to use your package can access your functions by loading them into their namespace using this:

library(test)

Or they can explicitly ask for a function in your namespace, as follows:

test::**function_name**()

But they can only get access to functions (and other objects) explicitly exported.[5] If a function is not explicitly exported, it is considered an implementation detail of the package that code outside the package should not be able to be accessed.

The NAMESPACE file is where you specify what should be exported from the package.[6]

The auto-generated file looks like this:

exportPattern("^[[:alpha:]]+")

---

[5]Strictly speaking, this is not true. You can actually get to internal functions if you use the `:::` operator instead of the `::` operator. So if function_name is not exported but still implemented in the test package, you can access it with test:::function_name. But you shouldn't. You should keep your damned dirty paws away from internal functions!
[6]It is also used to import selected functions or packages, but using Roxygen's @import and @importFrom functions are better solutions for that.

It is just exporting anything that has an alphanumeric name. This is definitely too much but we ignore it for now. We are not going to edit this file manually since we can export functions (and all other objects) much easier using Roxygen, as described shortly.

R/ and man/

The R/ directory is where you should put all your R code, and the man/ directory is where the package documentation goes. There is one example file in both directories just after RStudio generates your new package. You can look at them and then delete them afterward.

All the R code you write for a package should go in files in the R/ directory to be loaded into the package. All documentation will go in man/, but we are not going to write the documentation there manually. Instead, we will use Roxygen to document functions, and then Roxygen will automatically make the files that go in man/.

Roxygen

Roxygen is a system for writing documentation for your packages, and if you are familiar with Javadoc, you will recognize its syntax. It does a few things more, however, including handling your namespace import and export, as we will see.

To use it, you first have to install it, so run the following:

```
install.packages("roxygen2")
```

Now go into the Build menu and select Configure Build Tools. Pick Build Tools and check Generate Documentation with Roxygen. In the dialog box that pops up, check Build & Reload. This ensures that Roxygen is used to generate documentation and that the documentation is generated when you build the package. This will also ensure that Roxygen handles the import and export of namespaces.

Documenting Functions

You can see how Roxygen works through an example:

```
#' Add two numbers
#'
#' This function adds two numbers together.
#'
#' @param x  A number
#' @param y  Another number
#' @return The sum of x and y
#'
#' @export
add <- function(x, y) x + y
```

The documentation for this function, add, is provided in comments above the function, but comments starting with the characters #' instead of just #. This is what tells Roxygen that these comments are part of the documentation that it should process.

The first line becomes the title of the documentation for the function. It should be followed by an empty line (still in #' comments).

The text that follows is a description of the function. It is a bit silly with the documentation for this simple function, but normally you will have a few paragraphs describing what the function does and how it is supposed to be used. You can write as much documentation here as you think is necessary.

The lines that start with an @ tag—e.g., @param and @return—contain information for Roxygen. They provide information that is used to make individual sections in the documentation.

The @param tags are used for describing parameters. That tag is followed by the name of a parameter and then a short description of the parameter.

The @return tag provides a description of what the function returns.

After you have written some comments in Roxygen, you can build it by choosing Build ➤ Document. After you build the documentation, take a look at the NAMESPACE file and the man/ directory. In the NAMESPACE file, you should see that the function has been exported; in the man/ directory, there should be a file documenting the function.

Import and Export

In the NAMESPACE file, you should see that your documented function is explicitly exported. That is because we provided the @export tag with the documentation. It tells Roxygen to export it from the package namespace.

This is the easiest way to handle the namespace export so, if, for nothing else, you should use Roxygen for this rather than manually editing the NAMESPACE file.

Roxygen will also ensure that polymorphic functions and other kinds of objects are correctly exported if you use the @export tag. This requires different kinds of commands in the NAMESPACE file. You don't have to worry about it as long as you use Roxygen.

Roxygen can also handle import of namespaces. Remember that the packages you list in your Imports field in the DESCRIPTION file are guaranteed to be installed on the computer where your package is installed but that the namespaces of these packages are *not* imported. You have to use the :: notation to access them.

Well, with Roxygen you can use the @importFrom package object tag to import object (typically a function) into your namespace in a function that you give that tag to. For normal functions, I don't really see the point of using this feature since it isn't shorter than just using the :: notation. For infix functions, though, it makes them easier to use since then you can actually use the infix function as an infix operator.

So in the following function, you can use the %>% operator from dplyr because you import it explicitly. You cannot really get to infix operators otherwise.

```r
#' Example of using dplyr
#'
#' @param data A data frame containing a column named A
#' @param p    A predicate function
#' @return The data frame filtered to those rows where p is true on A
#'
#' @importFrom dplyr filter
#' @importFrom dplyr %>%
#' @export
filter_on_A <- function(data, p) {
  data %>% filter(p(A))
}
```

If you write a function that uses a lot of functionality from a package, you can also import the entire namespace of that package. That is similar to using `library(package)` and is done with the `@import` tag.

```
#' @import dplyr
#' @export
filter_on_A <- function(data, p) {
  data %>% filter(p(A))
}
```

Package Scope Versus Global Scope

A quick comment is in order about the namespace of a package when you load it with `library(package)`. I mentioned it already, but I just want to make it entirely clear. A package has its own namespace where its functions live. Functions that are called from other functions written inside a package are first looked for in the package namespace before they are looked for in the global namespace.

If you write a function that uses another function from your package and someone redefines the function in the global namespace after loading your package, it doesn't change what function is found inside your package.

It doesn't matter if a function is exported or local to a package for this to work. R will always look in a package namespace before looking in the global namespace.

Internal Functions

You might not want to export all functions you write in a package. If there are some functions, you consider implementation details of your package design, you *shouldn't* export them. If you do, people might start to use them, and you don't want that if it is functionality you might change later on when you refine your package.

Making functions local, though, is pretty easy. You just don't use the `@export` tag. Then they are not exported from the package namespace when the package is loaded and then they cannot be accessed from outside the package.[7]

File Load Order

Usually, it shouldn't matter in how many files you write your package functionality. It is usually easiest to find the right file to edit if you have one file for each (major) function or class, but it is mostly a matter of taste.

It also shouldn't matter in which files you place various functions—whether internal or exported—since they will all be present in the package namespace. And if you stick to using functions (and S3 polymorphic functions), the order in which files are processed when building packages shouldn't matter.

It does matter for S4 classes and such and in case you ever run into it being an issue, I will just quickly mention that package files are processed in alphabetical order. Alphabetical for the environment you are in, though, since alphabetical order actually depends on which language you are in, so you shouldn't rely on this.

---

[7]Except through the `:::` operator, of course, but people who use this to access the internals of your package knows – or should know – that they are accessing implementation details that could change in the future so it is their own fault if their code is broken sometime down the line.

Instead, you can use Roxygen. It can also make sure that one file is processed before another. You can use the `@include` field to create a dependency between a function and another file.

```
#' @import otherfile.R
```

I have never had the need for this myself and you probably won't either, but now you know.

Adding Data to Your Package

It is not uncommon for packages to include some data, either data used by the package implementation or more commonly data used for example purposes.

This data goes in the `data/` directory. You don't have this directory in your freshly made package, but it is where data should go if you want to include data in your package.

You cannot use any old format for your data. It has to be in a file that R can read, typically `.RData` files. The easiest way to add data files, though, is using functionality from the `devtools` package. If you don't have it installed then type this:

```
install.packages("devtools")
```

Then you can then use the `use_data` function to create a data file.

For example, I have a small test dataset in my admixturegraph package (see https://github.com/mailund/admixture_graph) that I made using this command:

```
bears <- read.table("bears.txt")
devtools::use_data(bears)
```

This data won't be directly available once a package is loaded, but you can get it using the `data` function:

```
library(admixturegraph)
data(bears)
bears
##      W  X      Y      Z      D Z.value
## 1  BLK PB Sweden   Adm1 0.1258    12.8
## 2  BLK PB  Kenai   Adm1 0.0685     5.9
## 3  BLK PB Denali   Adm1 0.0160     1.3
## 4  BLK PB Sweden   Adm2 0.1231    12.2
## 5  BLK PB  Kenai   Adm2 0.0669     6.1
## 6  BLK PB Denali   Adm2 0.0139     1.1
## 7  BLK PB Sweden    Bar 0.1613    14.7
## 8  BLK PB  Kenai    Bar 0.1091     8.9
## 9  BLK PB Denali    Bar 0.0573     4.3
## 10 BLK PB Sweden   Chi1 0.1786    17.7
## 11 BLK PB  Kenai   Chi1 0.1278    11.3
## 12 BLK PB Denali   Chi1 0.0777     6.4
## 13 BLK PB Sweden   Chi2 0.1819    18.3
## 14 BLK PB  Kenai   Chi2 0.1323    12.1
## 15 BLK PB Denali   Chi2 0.0819     6.7
## 16 BLK PB Sweden Denali 0.1267    14.3
## 17 BLK PB  Kenai Denali 0.0571     5.6
## 18 BLK PB Sweden  Kenai 0.0719     9.6
```

You cannot add documentation for data directly in the data file, so you need to put it in an R file in the R/ directory. I usually have a file called data.R that I use for documenting my package data.

For the bears data, my documentation looks like this:

```
#' Statistics for populations of bears
#'
#' Computed $f_4(W,X;Y,Z)$ statistics for different
#' populations of bears.
#'
#' @format A data frame with 19 rows and 6 variables:
#' \describe{
#'   \item{W}{The W population}
#'   \item{X}{The X population}
#'   \item{Y}{The Y population}
#'   \item{Z}{The Z population}
#'   \item{D}{The $D$ ($f_4(W,X;Y,Z)$) statistics}
#'   \item{Z.value}{The blocked jacknife Z values}
#' }
#'
#' @source \url{http://onlinelibrary.wiley.com/doi/10.1111/mec.13038/abstract}
#' @name bears
#' @docType data
#' @keywords data
NULL
```

The NULL after the documentation is needed because Roxygen wants an object after documentation comments, but it is the @name tag that tells it that this documentation is actually for the bears object. The @docType tells it that this is data that we are documenting.

The @source tag tells us where the data is from; if you have generated it yourself for your package, you don't need this tag.

The @format tag is the only complicated tag here. It describes the data, which is a data frame, and it uses markup that looks very different from Roxygen markup text. The documentation used by R is actually closer to LaTeX than the formatting we have been using, and the data description reflects this.

You have to put your description inside curly brackets marked up with \description{} and inside it, you have an item per data frame column. This has the format \item{column name}{column description}.

Building an R Package

In the frame to the upper right in RStudio, you should have a tab that says Build. Select it.

Inside the tab there are three choices in the toolbar—Build & Reload, Check, and More. They all do just what they say: the first builds and (re)loads your package, the second checks it, meaning it runs unit tests if you have written any and then checks for consistency with CRAN rules, and the third gives you various other options in a drop-down menu.

You use Build & Reload to recompile your package when you make changes to it. It loads all your R code (and various other things) to build the package and then it installs it and reloads it into your terminal so you can test the new functionality.

A package you have built and installed this way can also be used in other projects afterward.

When you have to send a package to someone, you can make a source package in the More drop-down menu. It creates an archive file (.tar.gz).

Exercises

In the last chapter, you wrote functions for working with *shapes* and *polynomials*. Now try to make a package for each with documentation and correct exporting of the functions. If you haven't implemented all the functionality for those exercises, this is your chance to do so.

CHAPTER 12

■ ■ ■

Testing and Package Checking

Without testing, there is little guarantee that your code will work at all. You probably test your code when you write it by calling your functions with a couple of chosen parameters, but to build robust software you need to approach testing more rigorously. And to prevent bugs from creeping into your code over time, you should test often. Ideally, you should check all your code anytime you make *any* changes to it.

There are different ways of testing software—software testing is almost a science in itself—but the kind of testing we do when we want to make sure that the code we just wrote is working as intended is called *unit testing*. The testing we do when we want to ensure that changes to the code do not break anything is called *regression testing*.

Unit Testing

Unit testing is called that because it tests functional units—in R, that essentially means single functions or a few related functions. Whenever you write a new functional unit, you should write test code for that unit as well. The test code is used to check that the new code is actually working as intended and if you write the tests such that they can be run automatically later on you have also made regression tests for the unit at the same time. Whenever you make any changes to your code, you can run all your automated tests, and that will check each unit and make sure that everything works as it did before.

Most programmers do not like to write tests. It is exciting to write new functionality but to probe new features for errors is a lot less interesting. However, you really *do* need the tests, and you will be happy that you have them in the long run. Don't delay writing tests until after you have written all your functions. That is leaving the worst for last, and that is not the way to motivate you to write the tests. Instead, you can write your unit tests while you write your functions; some even suggest writing them *before* you write your functions, something called *test-driven programming*. The idea here is that you write the tests that specify how your function should work, and you know that your function works as intended when it passes the tests you wrote for it.

I have never found test-driven programming that useful for myself. It doesn't match the way I work, because I like to explore different interfaces and uses of a function while I am implementing it, but some prefer to work that way. I *do*, however, combine my testing with my programming in the sense that I write small scripts calling my functions and fitting them together while I experiment with the functions. I write that code in a way that makes it easy for me to take the experiments, and then use them as automated tests later.

Take for example the shapes exercise we had earlier, where you had to write functions for computing the area and circumference of different shapes. I have written a version where I specify rectangles by `width` and `height`.[1] A test of the two functions could then look like this:

```
area <- function(x) UseMethod("area")
circumference <- function(x) UseMethod("circumference")

rectangle <- function(width, height) {
    structure(list(width = width, height = height),
              class = c("rectangle", "shape"))
}
area.rectangle <- function(x) x$height * x$width
circumference.rectangle <- function(x) 2 * x$height + 2 * x$width
r <- rectangle(width = 2, height = 4)
area(r)
## [1] 8
circumference(r)
## [1] 12
```

The area is 2×4 and the circumference is 2×2+2×4, so this looks fine. But I am testing the code by calling the functions and looking at the printed output. I don't want to test the functions that way forever—I cannot automate my testing this way because I then have to sit and look at the output of my tests. But they are okay tests. I just need to automate them.

Automating Testing

All it takes to automate the test is to check the result of the functions in code rather than looking at it, so code that resembles the following would be an automated test:

```
r <- rectangle(width = 2, height = 4)
if (area(r) != 2*4) {
    stop("Area not computed correctly!")
}
if (circumference(r) != 2*2 + 2*4) {
    stop("Circumference not computed correctly!")
}
```

It is a little more code, yes, but it is essentially the same test, and this is something I can run automatically later on. If it doesn't complain about an error, then the tests are passed, and all is good.

You can write your own test this way. Put them in a directory called `tests/` (which is where R expect tests to live) and then run these tests whenever you want to check the status of your code, i.e., whenever you make modifications to it.

Scripts in the `tests/` directory will also be automatically run whenever you do a consistency check of the package (something we return to shortly). That is what happens when you click Check in the Build tab on the right in RStudio or select Check Package in the Build menu, but it does a lot more than just run tests so it is not the most efficient way of running the tests.

There are some frameworks for formalizing this type of testing in R. I use a framework called `testthat`. Using this framework it is easy to run tests (without the full package check) and easy to write tests in a more structured manner—of course at the cost of having a bit more code to write for each test.

---

[1] I know that a rectangle doesn't have to have sides parallel with those two dimensions, but there is no need to make the example more complicated than it has to be.

Using `testthat`

The testthat framework (see https://github.com/hadley/testthat) provides functions for writing unit tests and makes sure that each test is run in a clean environment (so you don't have functions defined in one test leak into another because of typos and such). It needs a few modifications to your DESCRIPTION file and your directory structure, but you can automatically make these adjustments by running the following:

```
devtools::use_testthat()
```

This adds testthat to the Suggests packages, makes the directory tests/testthat and the file tests/testthat.R. You can have a look at the file, but it isn't that interesting. Its purpose is to make sure that the package testing—that runs all scripts in the tests/ directory—will also run all the testthat tests.

The testthat tests should all go in the tests/thestthat directory and in files whose names start with test. Otherwise, testthat cannot find them. The tests are organized in contexts and tests to make the output of running the tests more readable—if a test fails, you don't just want to know that some test failed somewhere, but you want some information about which test failed and where, and that is provided by the contexts.

At the top of your test files, you set a context using the context function. It just gives a name to the following batch of tests. This context is printed during testing so you can see how the tests are progressing and if you keep to one context per file, you can see in which files tests are failing.

The next level of tests is wrapped in calls to the test_that function. This function takes a string as its first argument, which should describe what is being tested. Its second argument is a statement that will be the test. The statement is typically more than one single statement, and in that case, it is wrapped in {} brackets.

At the beginning of the test statements you can create some objects or whatever you need for the tests and after that you can do the actual tests. Here, testthat also provides a whole suite of functions for testing if values are equal, almost equal, if an expression raises a working, triggers an error, and much more. All these functions start with expect_ and you can check the documentation for them in the testthat documentation.

The test for computing the area and circumference of rectangles would look like this in a testthat test:

```
context("Testing area and circumference")

test_that("we compute the correct area and circumference", {
  r <- rectangle(width = 2, height = 4)

  expect_equal(area(r), 2*4)
  expect_equal(circumference(r), 2*2 + 2*4)
})
```

Try to add this test to your shapes packet from last chapter's exercises and see how it works. Try modifying it to trigger an error and see how that works.

You should always worry a little bit when testing equality of numbers, especially if they are floating-point numbers. Computers do not treat floating-point numbers the way mathematics treat real numbers. Because floating-point numbers have to be represented in finite memory, the exact number you get depends on how you compute it, even if mathematically two expressions should be identical.

For the tests we do with the rectangle, this is unlikely to be a problem. There aren't really that many ways to compute the two quantities we test for and we *would* expect to get exactly these numbers. But how about the quantities for circles?

```
circle <- function(radius) {
    structure(list(r = radius),
              class = c("circle", "shape"))
}
area.circle <- function(x) pi * x$r**2
circumference.circle <- function(x) 2 * pi * x$r
test_that("we compute the correct area and circumference", {
  radius <- 2
  circ <- circle(radius = radius)

  expect_equal(area(circ), pi * radius^2)
  expect_equal(circumference(circ), 2 * radius * pi)
})
```

Here I use the built-in `pi` but what if the implementation used something else? Here we are definitely working with floating-point numbers, and we shouldn't ever test for exact equality. Well, the good news is that `expect_equal` doesn't. It tests for equality within some tolerance of floating-point uncertainty—that can be modified using an additional parameter to the function—so all is good. To check *exact* equality, you should instead use the function `expect_identical`, but it is usually `expect_equal` that you want.

Writing Good Tests

The easiest way to get some tests written for your code is to take the experiments you make when developing the code and translate them into unit tests like this right away—or even put your checks in a unit test file, to begin with. By writing the tests at the same time as you write the functions—or at least immediately after—you don't build a backlog of untested functionality (and it can be very hard to force yourself to go and spend hours just writing tests later on). Also, it doesn't really take *that* much longer to take the informal testing you write to check your functions while you write them and put them into a `testthat` file and get a formal unit test.

If this is all you do, at least you know that the functionality that was tested when you developed your code is still there in the future, or you will be warned if it breaks at some point because the tests will start to fail.

If you are writing tests anyway, you might as well be a *little* more systematic about it. We always tend to check for the common cases—the cases we have in mind when we write the function—and forget about special cases. Special cases are frequently where bugs hide, however, so it is always a good idea to put them in your unit tests as well.

Special cases are situations such as empty vectors and lists or `NULL` as a list. If you implement a function that takes a vector as input, make sure that it also works if that vector is empty. If it is not a meaningful value for the function to take, and you cannot think of a reasonable value to return if the input is empty, make sure the function throws an error rather than just do something that it wasn't designed to do.

For numbers, exceptional cases are often zero or negative numbers. If your functions can handle these cases, excellent (but make sure you test it!); if they cannot handle these special situations, throw an error.

For the shapes, it isn't meaningful to have non-positive dimensions, so in my implementation I raise an error if I get that and a test for it, for rectangles, could look like this:

```
test_that("Dimensions are positive", {
  expect_error(rectangle(width = -1, height =  4))
  expect_error(rectangle(width =  2, height = -1))
  expect_error(rectangle(width = -1, height = -1))
```

```
    expect_error(rectangle(width =  0, height =  4))
    expect_error(rectangle(width =  2, height =  0))
    expect_error(rectangle(width =  0, height =  0))
})
```

When you are developing your code and corresponding unit tests, it is *always* a good idea to think a little bit about what the special cases could be and make sure that you have tests for how you choose to handle them.

Using Random Numbers in Tests

Another good approach to testing is to use random data. With tests we manually set up, we have a tendency to avoid pathological cases because we simply cannot think them up. Random data doesn't have this problem. Using random data in tests can, therefore, be more efficient, but of course, it makes the tests non-reproducible which makes debugging extremely hard.

You can, of course, set the random number generator seed. That makes the test deterministic and reproducible, but defeats the purpose of having random tests to begin with.

I don't really have a good solution to this, but I sometimes use this trick: I pick a random seed and remember it and set the seed. Since I now know what the random seed was, I can set it again if the test fails and debug from there.

You can save the seed by putting it in the name of the test. Then if the test fails, you can get the seed from the error message.

```
seed <- as.integer(1000 * rnorm(1))
test_that(paste("The test works with seed", seed), {
  set.seed(seed)
  # test code that uses random numbers
})
```

Testing Random Results

Another issue that pops up when we are working with random numbers is what the expected value that a function returns should be. If the function is not deterministic but depends on random numbers, we don't necessarily have an expected output.

If all we can do to test the result in such cases is statistical, then that is what we must do. If a function is doing something useful, it probably isn't completely random and that means that we can do *some* testing on it, even if that test can sometimes fail.

As a toy example, we can consider estimating the mean of a set of data by sampling from it. It is a silly example since it is probably much faster to just compute the mean in the first place in this example, but let's consider it for fun anyway.

If we sample n elements, the standard error of the mean should be $s/\sqrt{n}$, where s is the sample standard error. This means that the difference between the true mean and the sample mean should be distributed as $N\left(0, s/\sqrt{n}\right)$.

In the following code, I normalize the distance between the two means by dividing it with $s/\sqrt{n}$, which should make it distributed as $Z \sim N(0,1)$. I then pick a threshold for significance that should only be reached one time in a thousand. I actually pick one that is only reached one in two thousand, but I am only testing the positive value for Z, so there is another implicit one in two thousand at the negative end of the distribution.

```
seed <- as.integer(1000 * rnorm(1))
test_that(paste("Sample mean is close to true, seed", seed), {
  set.seed(seed)

  data <- rnorm(10000)
  sample_size <- 100
  samples <- sample(data, size = sample_size, replace = TRUE)

  true_mean <- mean(data)
  sample_mean <- mean(samples)

  standard_error <- sd(samples) / sqrt(sample_size)
  Z <- (true_mean - sample_mean) / standard_error
  threshold <- qnorm(1 - 1/2000)

  expect_less_than(abs(Z), threshold)
})
```

This test is expected to fail one time in a thousand, but we cannot get absolute certainty when the results are actually random. If this test failed a single time, I wouldn't worry about it, but if I see it fail a couple of times, it becomes less likely that it is just a fluke, so then I would explore what is going on.

Checking a Package for Consistency

The package check you can do by clicking Check in the Build tab on the right in RStudio, or the Check Package in the Build menu, runs your unit tests and does a whole lot more.

It calls a script that runs a large number of consistency checks to make sure that your package is in tiptop shape. It verifies that all your functions are documented, that your code follows certain standards, that your files are in the right directories (and that there aren't files where there shouldn't be[2]), that all the necessary meta-information is provided, and many many more things. You can check for a longer list of the tests done when a package is being checked at http://r-pkgs.had.co.nz/check.html.

You should try to run a check for your packages. It will write a lot of output and, in the end, it will inform you about how many errors, warnings, and notes it found.

In the output, every test that isn't declared to be OK is something you should look into. It might not be an error, but if the check raises any flags, you will not be allowed to put it on CRAN. At least not without a very good excuse.

Exercise

You have written two packages—for shapes and for polynomials—and your exercise now is to write unit tests for these and get them to a point where they can pass a package check.

---

[2]If there are, you should have a look at `.Rbuildignore`. If you have a file just the place you want it but the check is complaining, you can just add the filename to `.Rbuildignore` and it will stop complaining. If you have a `README.Rmd` file, for example, it will probably complain, but then you can add a line to `.Rbuildignore` that says `^README.Rmd$`.

CHAPTER 13

■ ■ ■

Version Control

Version control, in its simplest form, is used to track changes to your software. It is also an efficient way of collaborating on software development since it allows several developers to make changes to the software and merge them with changes from other developers. RStudio supports two version control systems, Subversion and git. Of these, git is the most widely used, and although these things are very subjective of course, I think that it is also the better system. It is certainly the system we use here.

Version Control and Repositories

There are two main purposes of using a version control system when you develop software. One is simply to keep track of changes, so that you can later check when which modifications were made to your source code, and if you discover that they were in error, revert to earlier versions to try a different approach. It provides a log for your software development that allows you to go back in time and try again when you find that what you have done so far leads to some place you don't want to go.

The other job a version control system typically does is that it makes it easier for you to collaborate with others. Here the idea is that you share some global repository of all code and code changes—the log that the version control system keeps of all changes—and each developer works on a copy when modifying the code and submits that code to the repository when he or she is done changing the code. In early version control systems, it was necessary to lock files when you wanted to modify them to prevent conflicts with other developers who might also be editing the same files. These days, version control systems are more lenient when it comes to the concurrent editing of the same files, and they will typically just merge changes as long as there are not changes in overlapping lines (in which case you will have to resolve conflicts manually).

With this type of version control, different developers can work concurrently on different parts of the code without worrying about conflicts. Should there be conflicts, these will be recognized when you attempt to push changes to the global repository, and you will be told to resolve the conflicts.

The version control system git allows even more concurrent and independent development than this, by not even having a single global repository as such. At least in theory. In practice, having a global repository for the *official* version of your software is a good idea, and people do have that. The system just doesn't enforce a single global repository, but instead, is built around having many repositories that can communicate changes to each other.

Whenever you are working with git, you will have a local repository together with your source code. You can use this repository as the log system or create branches for different features or releases, as you will see later. You make changes to your source code like normal and can then commit them to your local repository without any conflict with other people's changes. However, you can't see their changes and they can't see yours because you are working on different local repositories. To make changes to another repository, you have to push your changes there and to get changes from another repository, you have to pull them from there.

This is where you typically use a global repository. You make changes to your local repository while developing a feature but when you are done you push those changes to the global repository. Or if you do not have permission to make changes to the global repository—perhaps because you cloned someone else's code and made changes to that—ask someone who *does* have permission to pull your changes into the repository. This is known as a "pull request".

Using git in RStudio

This is all very theoretical and, if it is hard for me to write, it is probably also hard for you to understand. Instead, let's see git in practice.

RStudio has some rudimentary tools for interacting with git: it lets you create repositories, commit to them, and push changes to other repositories. It does not support the full range of what you can do with git— for that, you need other tools or to use the command-line version of git—but for day-to-day version control, it suffices for most tasks.

Installing git

If you haven't installed git already on your computer, you can download it from http://git-scm.com. There should be versions for Windows, OS X, and Linux, although your platform might have better ways of installing it. For example, on a Debian/Ubuntu system, you should be able to use the following:

```
sudo apt-get install git-core
```

On a Red Hat/Fedora system, you should be able to use the following:

```
sudo yum install git-core
```

You have to Google around to check how best to install git on other systems.

Once git is installed, you want to tell it who you are. It needs this to be able to tag changes to your code with your name. It isn't frightfully important if you are the only one working on the code, but if more people are collaborating on the software development, it is necessary to identify who made which changes. You tell git who you are by running the following commands in a terminal[1]:

```
git config --global user.name "YOUR FULL NAME" git config --global user.email "YOUR EMAIL
ADDRESS"
```

You also might have to tell RStudio where the git command you installed can be found. You do that by choosing Tools ➤ Global Options. In the window that pops up you should find, on the icons on the left, a panel with Git/SVN and in there you can tell RStudio where the git command can be found.

The git you have installed is a command-line tool. RStudio has some GUI to work with git, but you can't do everything from the GUI. There are a few GUI tools that allow you to do a lot more with git than RStudio, and I recommend getting one of those. I find it easier using them than the command lines myself since I am getting old and forget the exact commands.

---

[1]Not the R terminal. You need to run this in an actual shell terminal for it to work. How you open a terminal depends on your platform. I can't help you there. If you don't know how to, it is time to fire up Google once again.

Some good choices are:

- SourceTree (https://www.sourcetreeapp.com), for Windows and OS X

- GitHub Desktop (https://desktop.github.com), for Linux, Windows, and OS X (for working with GitHub repositories)

- GitG (https://wiki.gnome.org/Apps/Gitg/), for Linux

Sometimes, though, you do need to use the command-line version. There is a very nice interactive web tutorial for the command-line git program here: try.github.io.

Making Changes to Files, Staging Files, and Committing Changes

If you checked that your project should use git when you created your package, you should have a Git tab on the top-right of RStudio, next to the Build tab (see Figure 13-1). Click on it.

In the main part of this panel, there is a list of files. There are three columns—Staged, Status, and Path—the latter contains the names of modified files (or directories).
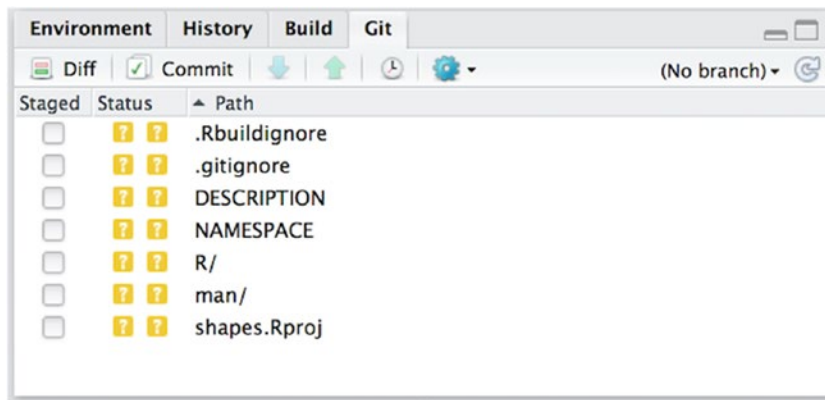
Figure 13-1. *Git panel showing modified files*

If this is the first time you have accessed this panel, the status will contain a yellow question mark for all files you have modified since you created the object (including files that RStudio made during the package creation). This status means that git doesn't know about these files yet. It can see that the files are there, but you have never told it what to do about them. We will do something about that now.

The Staged column has check boxes for all the files. If you click one, the status for that file changes to a green "A". This means that you have staged the file to be added to the git repository. Do this for all of the files. When you do it for a directory, all the files in that directory will also be staged for adding. This is also what we want for now.

The process of committing changes to git involves staging changes to be committed before we actually commit them. What we just did was tell git that next time we commit changes, we want these files added. Generally, committing will only affect changes we have staged. This lets you commit only some of the changes you have made to your source code, which can be helpful at times. You might have made several changes to many files but at some point, you only want to commit a particular bug fix and not changes for a new feature that you are not quite done with yet. Staging only the changes you want to commit allows for this.

Anyway, we have staged everything and to commit the changes you now have to click the Commit button in the toolbar. This opens a new window that shows you the changes you are about to commit and lets you write a commit message (on the upper right). This message is what goes into the change log. Give a short and meaningful description of your changes here. You will want it if you need to find the changes in your log at some later time. Then click Commit and close the window. The Git panel should now be empty. This is because there are no more changes since the last commit, and the panel only shows the files that have changed between your current version of your software and the version that is committed to git.

To do what you just did in the terminal instead, you would stage files using the `git add` command as follows:

```
git add filename
```

You would then commit staged changes using the `git commit` command:

```
git commit -m "message"
```

Now try modifying a file. After you have done that, you should see the file displayed in the Git panel again, but this time with a status that is a blue M, as shown in Figure 13-2. This, not surprisingly, stands for *modified*.
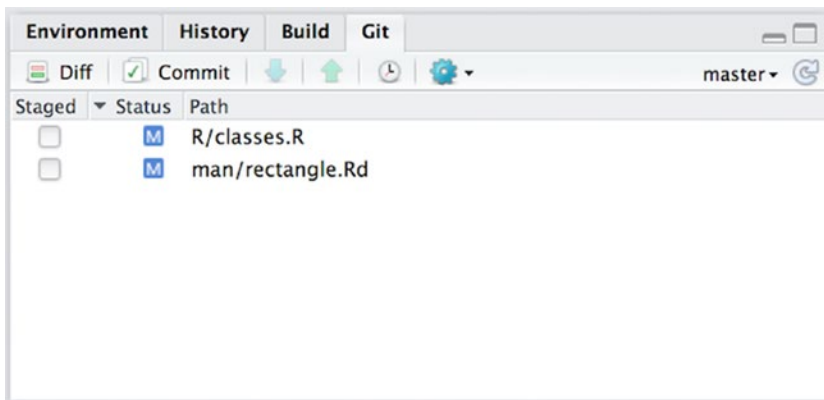
Figure 13-2. *Modified files in the Git panel*

If you stage a file for commit here, the status is still M, but RStudio indicates that it is now staged by moving the M to the left a little (see Figure 13-3). Not that you really need that feedback, because you can also see that it is staged from the checked staged button of course.
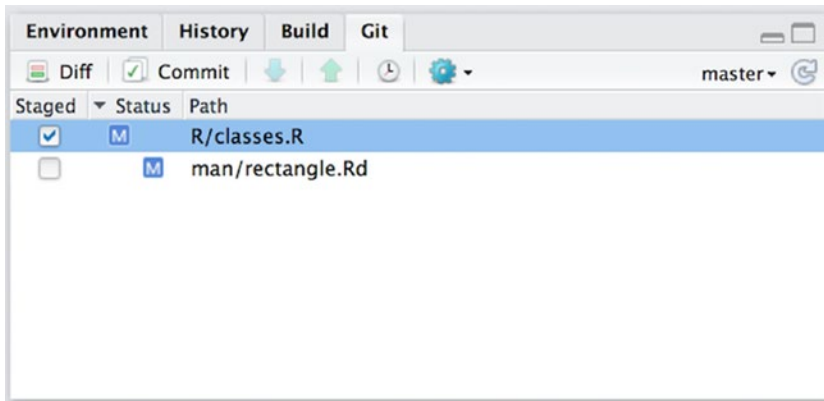
Figure 13-3. *Modified files with one file staged for commit*

Committing modified files works exactly like committing added files.

In the terminal, you use `git add` for staging modified files as well. You don't have a separate command for staging adding new files and staging modified files. It is `git add` for both.

Adding git to an Existing Project

If you didn't create your project with a git repository associated with it—and you have just learned about git now, so unless you have always just clicked the git button when creating projects you probably have many projects without git associated—you can still set up git for an existing directory. Simply choose Build ➤ Configure Build Tools and then click the Git/SVN option in the dialog box. If you choose Git as the version control, RStudio will configure git for you.

You can also do this from the command line. Go to the directory where the project is and run this command:

```
git init
```

This sets up an empty repository. You can then add the files already in the directory, just as you saw previously.

Bare Repositories and Cloning Repositories

Most of the material in this section is not something you will ever have to worry about if you use a repository server such as GitHub. There, creating a repository and interacting with it is handled through a web interface and you won't have to worry about the details, except for "cloning" a repository. We will create a so-called "bare" repository manually here and see how we can communicate changes in different local repositories through this.

The repositories we made when we created R projects or used `git init` in a directory are used for version control of the source code in the project directory. They are not really set up for collaboration between developers. While it *is* technically possible to merge changes in one repository to another, it is a bit cumbersome and not something you want to deal with on a daily basis.

To synchronize changes between different repositories, we want a *bare repository*. This is a repository that doesn't include the local source code; it isn't really special, but it prevents you from making local changes to it. You can only update it with changes from other repositories.

To create a bare repository, you need to use the command-line version of git. Create a directory where you want the repository, and then go in there and type:

```
git --bare init
```

The repository now contains the various files that git needs to work with. Your local repositories also include these files; they are just hidden in a subdirectory called `.git/` when you have the local source code as well.

We are not going to do anything with this repository directly. We just need it to see how we work with other repositories connected to it.

Go to a directory where you want the working source code version of the repository and make a copy of the bare repository by writing the following:

```
git clone /path/to/bare/repository
```

You will get a warning that you have cloned an empty repository. We already know that so don't worry about it. We are going to add to it soon.

To see how to communicate between repositories, though, you need to make another working copy. You can either go to another directory and repeat the `clone` command or clone the repository with another name, as follows:

```
git clone /path/to/bare/repository name
```

You now have two clones of the bare repository. You'll see how to push changes from a clone to the cloned repository and how to pull updates in the cloned repository into the clone.

As mentioned, going through a bare repository is not the only way to move changes from one repository to another, but it *is* the easiest way to work with git and the one you will be using if you use a server such as GitHub. If you do, and you'll see later how to do this, then GitHub will make the bare repository for you. You'll just need to clone it somewhere on your own computer to work with it.

Pushing Local Changes and Fetching and Pulling Remote Changes

Go into one of the clones you just made. It will look like an empty directory because you haven't made any changes to it yet. In fact, it does contain a hidden directory, `.git/`, where git keeps its magic, but you do not need to know about that.

Try to make some files, add them to git, and commit the changes.

```
touch foo bar
git add foo bar
git commit -m "added foo and bar"
```

If you now check the log as follows:

```
git log
```

You will see that you have made changes. If you look in the other clone of the bare repository, though, you don't see those changes.

There are two reasons for this: 1) we have only made changes to the cloned repository but never pushed them to the bare repository the two clones are connected to, and 2) even if we *had* done that, we haven't pulled the changes down into the other clone.

The first of these operations is done using `git push`. This will push the changes you made in your local repository up to the repository you cloned it from.[2]

`git` push

You don't need to push changes up to the global (bare) repository after each commit; you probably don't want to do that, in fact. The idea with this workflow is that you make frequent commits to your local code to make the version control fine grained, but you push these changes up when you have finished a feature—or at least gotten it to a stage where it is meaningful for others to work on your code. It isn't a major issue if you commit code that doesn't quite work to your local repository—although generally, you would want to avoid that—but it will not be popular if you push code that doesn't work onto others.

After pushing the changes in the first cloned repository, they are still not visible in the second repository. You need to pull them down.

The following command makes changes to the global repository and makes it possible for you to check them out before merging them with your own code:

`git` fetch

This can be useful because you can then check out the changes and make sure they aren't breaking anything before you merge them with your code. After running the `fetch` command, you can check out branches from the global repository, make changes there, and merge them into your own code using the branching mechanism described next. In most cases, however, you just want to merge the changes made to the global repository into your current code, and you don't really want to modify it before you do so. In that case, this command is sufficient:

`git` pull

It will both fetch the latest changes and merge them into your local repository in a single operation. This is by far the most common operation for merging changes others have made and pushed to the global repository with your own changes.

Go to the repository clone without the changes and run the command. Check that you now have the changes there.

The general workflow for collaborating with others on a project is to make changes and commit them to your own repository. You use this repository to make changes you are not ready to share yet, and you are the only one who can see them. Then, when you are ready to share with your collaborators, you can push the changes to the shared repository, and when you need changes others have made, you can pull them.

If you try to push to the global repository, and someone else has pushed changes that you haven't pulled yet, you will get an error. Don't worry about that. Just pull the changes; after that, you can push your changes.

If you pull changes into your repository, and you have committed changes there that haven't been pushed yet, it becomes a merge operation, which requires a commit message. There is a default message for this that you can use.

You have your two repositories to experiment with, so try to make variations of pushing and pulling and pulling changes into a repository where you have committed changes. This explanation will hopefully make a lot more sense after you have experimented a bit on your own.

RStudio has some basic support for pushing and pulling. If you make a new RStudio project and choose to put it in an existing directory, you can try to make one that sits in your cloned repositories. If you do this, you will find that the Git panel now has two new buttons: Push and Pull.

---

[2]If you didn't have a bare repository, you could still have connected the clones to see changes made to them, but pushing changes would be much more complicated. With a bare repository that both are cloned from, pushing changes upward is as easy as `git push`.

Handling Conflicts

If it happens that someone has pushed changes to the global repository that overlaps lines that you have been editing in your local repository, you will get a so-called *conflict* when you pull changes.

git will inform you about this, whether you pull from RStudio or use the command line. It will tell you which files are involved, and if you open a file with a conflict, you will see that git has marked the conflict with text that looks like this:

```
<<<<<<< HEAD
your version of the code
=======
the remote version of the code
>>>>>>> 9a0e21ccd38f7598c05fe1e21e2b32091bb0839b
```

It shows you the version of the changes you have made and the version of the changes that are in the global repository. Because there are changes both places, git doesn't know how to merge the remote repository into your repository in the pull command.

You have to go into the file and edit it so it contains the version you want, which could be a merge of the two revisions. Get rid of the <<<</====/>>>> markup lines when you are done making the changes.

Once you have edited the file with conflicts, you need to stage it—by running the `git add filename` on the command line or clicking the file in the Staged column in the Git panel in RStudio—and commit it. This tells git that you handled the conflict and will let you push your own changes if you want to do this.

Working with Branches

Branches are a feature of most version control systems, which allow you to work on different versions of your code at the same time. A typical example is having a branch for developing new features and another branch for the stable version of your software. When you are working on implementing new features, the code is in a state of flux, the implementation of the new feature might be buggy, and the interface could be changing between different designs. You don't your users to use such a version of your software—at least not without being aware that the package they are using is unstable and that the interface could change at a moment's notice. So you want the development code to be separate from the released code.

If you just made releases at certain times and then implemented new features between making releases that wouldn't be much of an issue. People should be using the version you have released and not the commits that fall between released versions. But the world is not that simple if you make a release with a bug in it—and let's face it, that is not impossible—and you want to fix that bug when it is discovered. You probably don't want to wait with fixing the bug until you are done with all the new features you are working. So you want to make changes to the code in the release. If there are more bugs, you will commit more bug fixes onto the release code. And all this while you are still making changes to your development code. Of course, those bug fixes you make to the released code will also need to be merged into the development code. After all, you don't want the next release to reintroduce bugs you already fixed.

This is where branches come in. RStudio has very limited support for branches, and it doesn't help you create them.[3] For that, you need to use the command line.

To create a branch, you use the `git branch name` command. To create a development branch called `develop`, for lack of imagination, you would use this command:

```
git branch develop
```

---

[3]Some of the other GUIs that work with git have excellent support for working with branches. You should check them out.

This simply creates the branch. You are not magically moved to the branch or anything. It just tells git that you have a new branch (and it branches off your current position in the list of commits done to the repository).

In RStudio, you can see which branch you are on in the Git panel. In the project you have experimented on so far—and any project you made where you created a git repository with `git init` or clicked the git selection in the dialog box when you created the project—you will be on the master branch (see Figure 13-4). This is the default branch and is typically used for released versions.
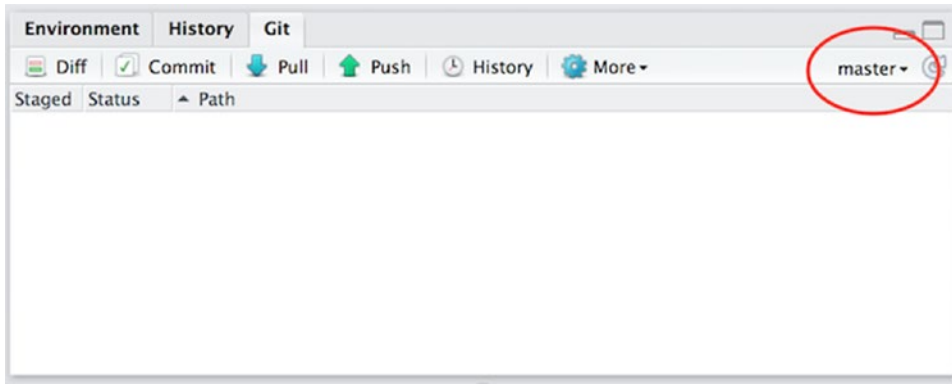
Figure 13-4. *Git panel when the code is on the master branch*

If you click on the branch drop-down in the Git panel, you get a list of the branches in your repository, as shown in Figure 13-5. You will have a branch called `origin/master`. This is the master branch on the central repository and the one you merge with when pulling data. Ignore it, it is not important here. If you ran the `git branch develop` command, you should also have a develop branch (see Figure 13-6). If you select it, you move to that branch.
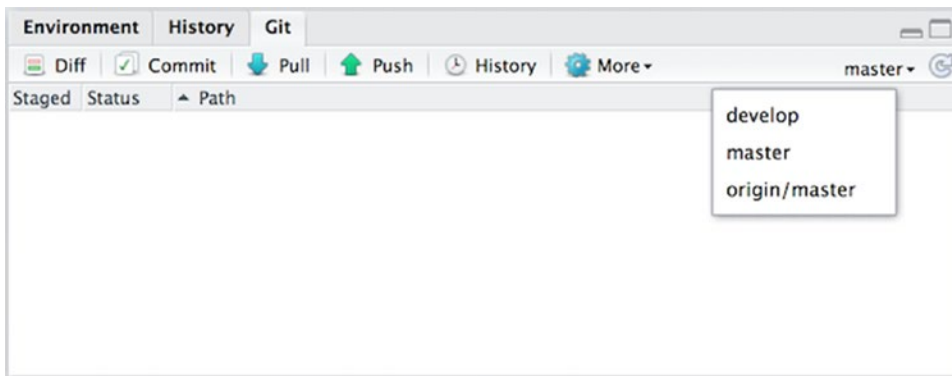
Figure 13-5. *Selecting a branch to switch to*
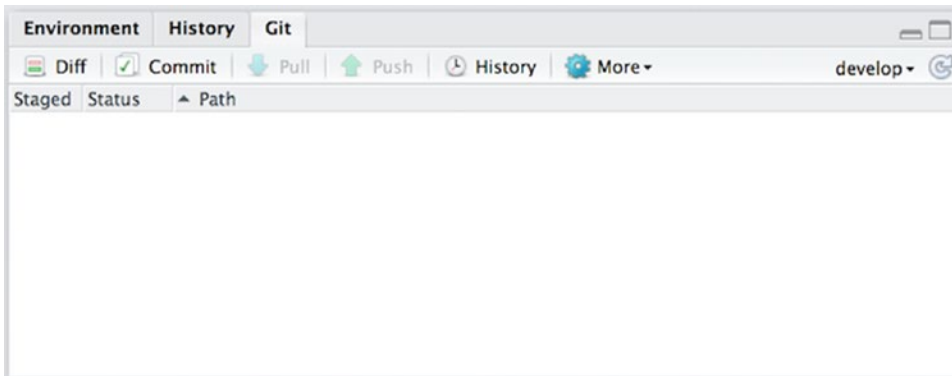
Figure 13-6. *After switching to the develop branch*

You can also get a list of branches on the command line with the following:

git branch

You can also switch to a branch using this command[4]:

git checkout branchname

If you switch to the develop branch, you will see that the Pull and Push buttons are greyed out. You can make changes to your code and commit them when you're on a given branch, but you cannot (yet) push and pull. We will get to that shortly.

If you make some changes to your code and commit it while on the develop branch and then switch to the master branch, you will see that those changes are not there. You can see both by looking at the files and by looking at the git history (using git log or clicking the History button in the Git panel). Similarly, changes you make in master will not show up in develop. This is exactly what you want. The two branches are independent, and you can switch between working on the development branch and the release version of your software by switching branches.

When you make changes to one branch and you want those changes to be added to another, you need to merge branches. Actually, you need to merge one branch into another, it is not a symmetric operation. To do this, check out the branch you want to modify and run this command to merge the changes in otherbranch with the current branch:

git merge otherbranch

So for example, if you fix a bug in the master branch and want to merge it into the develop branch, you would use this command:

git checkout develop
git merge master

If a merge causes conflicts, you resolve them the same was as if a pull causes conflicts. Not surprisingly since a pull command is actually just a shortcut for fetching and merging.

---

[4]You can also combine creating and checking out a branch using git checkout -b branchname if you want. That command creates the branch first and then checks it out. To change between branches later, though, use the checkout command without the -b option.

Typical Workflows Involve Lots of Branches

Git is optimized for working with lots of branches (unlike some version control systems, where creating and merging branches can be rather slow operations). This is reflected in how many people use branches when working with git: you create many branches and work on a graph of different versions of your code and merge them together whenever you need to.

Having a development branch and a master branch is a typical core of the repository structure, but it is also very common to make a branch for each new feature you implement. Typically, you branch these off the `develop` branch when you start working on the feature and merge them back into `develop` when you are done. It is also common to have a separate branch for each bug fix—typically branched off `master` when you start implementing the fix and then branched back into the `master` as well as into `develop` when you are done. See Atlassian's Git Tutorial at `https://www.atlassian.com/git/tutorials/comparing-workflows` for different workflows that exploit various branches.

If you create a lot of branches for each feature or bug fix, you don't want to keep them around after you are done with them—unlike with the `develop` and `master` branches, which you probably will want to keep around forever. To delete a branch, use this command:

```
git branch -d branchname
```

Pushing Branches to the Global Repository

You can work on as many branches as you like in your local repository but they are not automatically found in the global repository. The `develop` branch you made earlier exists only in the local repository and you cannot push changes made to it to the global repository—you can see this in RStudio since the push (and pull) buttons are greyed out.

If you want a branch to exist on the global repository as well—so you can push to it, and so collaborators can check it out—you need to create a branch in that repository and set up a link between your local repository and the global repository.

You can do that for the `develop` branch by checking it out and running this command:

```
git push --set-upstream origin develop
```

This pushes the changes and also remembers that the branch is linked to the `develop` branch in `origin`. The name `origin` refers to the repository you cloned when you created this repository.[5]

Whether you want a branch you are working on, also to be found in the global repository is a matter of taste. If you are working on a feature that you want to share when it is completed but not before, you probably don't want to push that branch to the global repository. For the `develop` and `master` branches, though, you definitely want those to be in the global repository.

GitHub

GitHub (`https://github.com`) is a server for hosting git repositories. Open projects are hosted for free, closed projects for a fee. You can think of it as a place to have your bare/global repository with some extra benefits. There are ways for automatically installing packages that are hosted on GitHub; there is web support for tracking bugs and feature requests, and there is support for sharing fixes and features in hosted projects through a web interface.

---

[5]It is slightly more complex than this; you can have links to other repositories and pull from them or push to them (if they are bare repositories), and `origin` is just a default link to the one you cloned for. It is beyond the scope of these notes, however, to go into more details. If you always work with a single global repository that you push to and pull from, then you don't need to know any more about links to remote repositories.

To use it, you first need to go to the home page and sign up. This is free, and you just need to pick a username and a password.

Once you have created an account on GitHub, you can create new repositories by clicking the big + in the upper-right corner of the home page, as shown in Figure 13-7.

Figure 13-7. *The button to create a new repository at the GitHub home page is found on the upper right of the home page*

Clicking it, you get to a page where you can choose the name of the repository, create a short description, pick a license, and decide whether you want to add a README.md file to the repository. I recommend that you always have a README.md file—it works as the documentation for your package since it is displayed on the home page for the repository at GitHub. You probably want to set up a README.rmd file to generate it, though, as you saw in Chapter 11. For now, though, you might as well just say yes to have one generated.

Once you have generated the repository, you go to a page with an overview of the code in the repository, as shown in Figure 13-8.
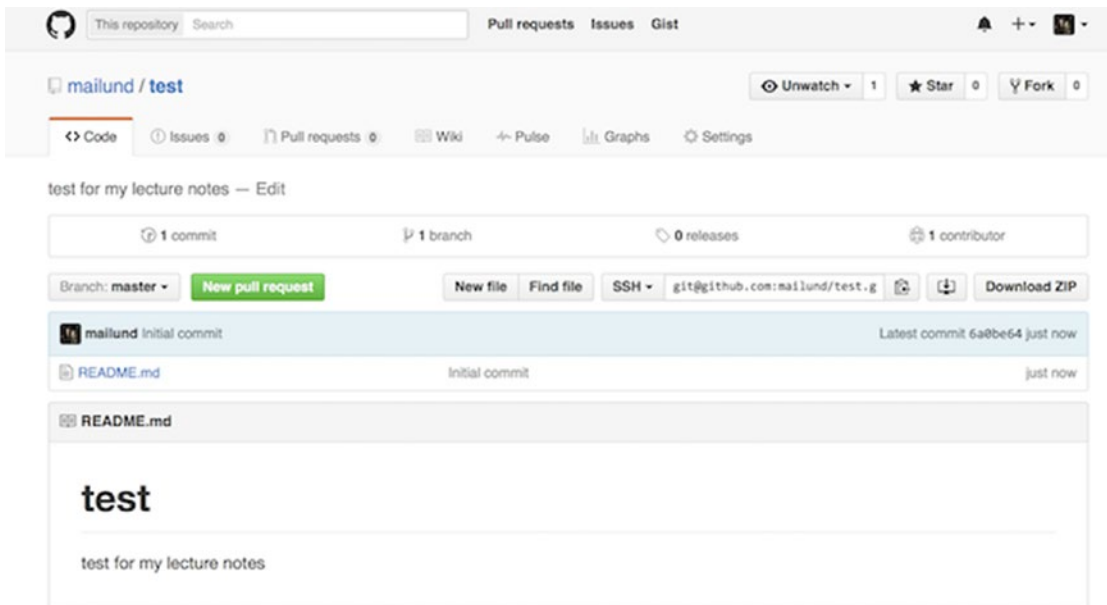
Figure 13-8. *New GitHub repository containing only a README.md file*

You can explore the web page and the features implemented there later—it is a good idea to know what it supports—but for now you can just use the repository here as a remote global repository. To clone it, you need the address in the field next to the button that says **SSH**. In my test repository, it is `git@github.com:mailund/test.git`. This is an address you can use to clone the repository using the SSH protocol.

```
git clone git@github.com:mailund/test.git
```

This is a protocol that you will have access to on many machines, but it involves you having to deal with a public/private key protocol. Check the documentation at https://help.github.com/articles/generating-ssh-keys/ for setting up the SSH key at GitHub for learning more about this. It is mostly automated by now, and you should be able to set it up just by making a push and answering yes to the question you get there.

It is not the easiest protocol to work with, though, if you are on a machine that has HTTPS, which is the protocol used by your web browser for secure communication. You will almost certainly have that on your own machine, but depending on how firewalls are set up, you might not have access to it on computer clusters and such and then you need to use the SSH protocol. To use HTTPS instead of SSH, just click the SSH drop-down and choose HTTPS instead. This gives you a slightly different address, which you can use to clone. In my repository, I get https://github.com/mailund/test.git.

```
git clone https://github.com/mailund/test.git
```

If nothing goes wrong, you should be able to use the cloned repository just as the repositories you looked at previously, when you made your own bare/global repository.

You can also check out the repository and make an RStudio project at the same time by choosing File ➤ New Project in RStudio and selecting Version Control (the third option) in the dialog box that pops up. In the next window, choose Git and then use the HTTPS address as the Repository URL.

Moving an Existing Repository to GitHub

If you have already used git locally in a project and want to move it to GitHub, there is a little more you must do—t least if you want to move your repository including all the history stored in it and not just the current version of the source code in it.

First, you need to make a bare version of your repository. This is, as you saw a while ago, just a version of the repository without the source code associated.

If your repository is called `repo`, you can make a bare version of it, called `repo.git`, by cloning it:

```
git clone --bare repo repo.git
```

To move this to GitHub, create an empty repository there and get the HTTPS address of it. Then go into the bare repository we just made and run the following command:

```
cd repo.git
git push --mirror ‹https address at github›
```

Now just delete the bare repository used to move the code to GitHub and clone the version *from* GitHub. Now you have a version from there that you can work on.

```
rm -rf repo.git
git clone ‹https address at github›
```

Installing Packages from GitHub

A very nice extra benefit you get from having your R packages on GitHub—in addition to having version control—is that other people can install your package directly from there. The requirements for putting packages on CRAN are much stricter than for putting R packages on GitHub, and you are not allowed to upload new versions to CRAN very often, so for development versions of your R package, GitHub is an excellent alternative.

To install a package from GitHub, you need to have the devtools package installed, as follows:

```
install.packages("devtools")
```

After that, you can install a package named packagename written by the GitHub user username with this command:

```
devtools::install_github("username/packagename")
```

Collaborating on GitHub

The repositories you make on GitHub are by default only editable by yourself. Anyone can clone them to get the source code, but only you can push changes to the repository. This is, of course, useful to prevent random people from messing with your code but prevents collaborations.

One way to collaborate with others is to give them write permissions to the repository. On the repository home page, you must select the Settings entry in the toolbar and then pick Collaborators in the menu on the left. After that, you get to a page where you can add collaborators identified by their user account on GitHub. Collaborators can push changes to the repository just as you can yourself. To avoid too much confusion, when different collaborators are updating the code, it is useful to have some discipline in how changes are merged into the master (and/or the develop) branch. One approach that is recommended and supported by GitHub is to make changes in separate branches and then use so-called *pull requests* to discuss changes before they are merged into the main branches.

Pull Requests

The workflow for making pull requests is to implement your new features or bug fixes or whatever you are implementing on separate branches from develop or master. Then, instead of merging them directly, you create what is called a *pull request*. You can start a pull request by switching to the branch on the repository home page and clicking the big green New Pull request button, or if you just made changes, you should also see a green Compare & Pull Request button that lets you start a pull request.

Clicking the button takes you to a page where you can name the pull request and write a description of what the changes in the code are doing. You also decide which branch you want to merge the pull into. Above the title you give the pull request, you can select two branches—the one you want to merge into (Base) and the branch you have your new changes on (Compare). You should pick the one you branched out of when you made the new branch. After that, you can create the pull request.

This simply creates a web interface for having a discussion about the changes you made. It is possible to see the changes on the web page and comment on them and make comments to the branch in general. At the same time, anyone can check out the branch and make their own modifications. As long as the pull request is open, the discussion is going, and people can improve on the branch.

When you are done, you can merge the pull request (using the big green Merge Pull Request button on the web page that contains the discussion about the pull request).

Forking Repositories Instead of Cloning

Making changes to separate branches and then making pull requests to merge in the changes still requires writing access to the repository. This is excellent for collaborating with a few friends, but not ideal for getting fixes from random strangers—or for making fixes to packages other people write; people who won't necessarily want to give you full write access to *their* software.

Not to worry, it is still possible to collaborate with people on GitHub without having write access to each other's repositories. The way that pull-requests work, there is actually no need for branches to be merged to be part of the same base repository. You can merge branches from anywhere if you want to.

If you want to make changes to a repository that you do not have write access to, you can clone it and make changes to the repository you get as the clone, but you cannot push those changes back to the repository you cloned it from. And other users on GitHub can't see the local changes you made (they are on your personal computer, not on the GitHub server). What you want is a repository *on GitHub* that is a clone of the repository you want to modify *and* that is a bare repository so you can push changes into it. You then want to clone *that* repository to your own computer. Changes you make to your own computer can be pushed to the bare repository you have on GitHub—because it is a bare repository and because you have writing access to it—and other users on GitHub can see the repository you have there.

Making such a repository on GitHub is called *forking* the repository. Technically, forking isn't different from cloning—except that you're making a bare repository—and the terminology is taken from open source software where forking a project means making your own version and developing it independently of previous versions.

Anyway, whenever you go to a repository home page on GitHub, you should see the Fork button at the top right—to the right of the name and branch of the repository you are looking at. Clicking the Fork button will make a copy of the repository that you have writing access to. You cannot fork your own repositories, although I'm not sure why you are not allowed to and, in most cases, you don't want to do that anyway. You can also fork any repository at other user's accounts.

Once you have made the copy, you can clone it to your computer and make changes to it, as you can with any other repositories. The only way this repository is different from a repository you made yourself, is that when you make pull requests, GitHub knows that you forked it off another repository. So when you make a pull request, you can choose not only the Base and *Compare* branches, but also the *base fork* and the *head fork*—the former being the repository you want to merge changes into, and the latter the repository where you made your changes. If someone forks your project and you make a pull request in the original repository, you won't see the *base fork* and *head fork* choices by default, but clicking on the Compare Across Forks link when you make pull requests will enable them there as well.

If you make a pull request with your changes to someone else's repository, the procedure is exactly the same as when you make a pull request to your own projects, except that you cannot merge the pull request after the discussion about the changes. Only someone with permission to write to the repository can do that.

The same goes if someone else wants to make changes to your code. They can start a pull request with their changes to your code but only you can decide to merge the changes into the repository (or not) following the pull discussion.

This is a very flexible way of collaborating—even with strangers—on source code development and one of the great strengths of git and GitHub.

Exercises

Take any of the packages you wrote earlier and create a repository on GitHub to host it. Push your code there.

■ ■ ■

Profiling and Optimizing

In this last chapter, we briefly consider what to do when you find that your code is running too slow, and, in particular, how to figure out *why* it is running too slow.

Before you start worrying about your code's performance, though, it is important to consider if it is worth speeding it up. It takes you time to improve performance, and it is only worth it if the improved performance saves you time when this extra programming is included. For an analysis you can run in a day, there is no point in spending one day making it faster, even *much* faster, because you still end up spending the same time, or more, to finally get the analysis done.

Any code you just need to run a few times during an analysis is usually not worth optimizing. We rarely need to run an analysis just *once*—optimistically we might hope to, but in reality, we usually have to run it again and again when data or ideas change—but we don't expect to run it hundreds or thousands of times. So even if it will take a few hours to rerun an analysis, your time is probably better spent working on something else while it runs. It is rarely worth it to spend a lot of time making it faster. The CPU time is cheap compared to your own.

If you are developing a package, though, you often do have to consider performance to some extent. A package, if it is worth developing, will have more users and the total time spend on running your code makes it worthwhile, up to a point, to make that code fast.

Profiling

Before you can make your code faster, you need to figure out why it is slow, to begin with. You might have a few ideas about where the code is slow, but it is actually surprisingly hard to guess at this. Quite often, I have found, it is nowhere near where I thought it would be, that most of the time is actually spend. On two separate occasions, I have tried working really hard on speeding up an algorithm only to find out later that the reason my program was slow was the code used for reading the program's input. The parser was slow. The algorithm was lightning fast in comparison. That was in C, where the abstractions are pretty low-level and where it is usually pretty easy to glance from the code how much time it will take to run. In R, where the abstractions are very high-level, it can be *very* hard to guess how much time a single line of code will take to run.

The point is, if you find that your code is slow, you shouldn't be guessing at where it is slow. You should measure the running time and get to know for sure. You need to profile your code to know which parts of it is taking up most of the running time. Otherwise, you might end up optimizing code that uses only a few percentages of the total running time and leaving the real time-wasters alone.

In common code, there are only a few real bottlenecks. If you can identify these and improve *their* performance, your work will be done. The rest will run fast enough. Figuring out where those bottlenecks are requires profiling.

We are going to use the `profvis` package for profiling. In the most recent versions of RStudio, there is support for this, if your version has it you should have a Profile item in the main menu. We will just use the package in our R code here.

A Graph-Flow Algorithm

For an example of some code, imagine you want to profile a small graph algorithm. It is an algorithm for smoothing out weights put on nodes in a graph. It is part of a method used for propagating weights of evidence for nodes in a graph and has been used to boost searching for disease-gene associations using gene-gene interaction networks. The idea is, that if a gene is a neighbor to another gene in this interaction network, then it is more likely to have a similar association with a disease as the other gene. So genes with known association are given an initial weight, and other genes get a higher weight if they are connected to such genes than if they are not.

The details of what the algorithm is used for is not so important, though. All it does is to smooth out weights between nodes. Initially all nodes, n, are assigned a weight $w(n)$. Then in one iteration of smoothing, this weight is updated as $w'(n) = \alpha w(n) + (1 - \alpha)\dfrac{1}{|N(n)|}\displaystyle\sum_{v \in N(n)} w(v)$, where α is a number between zero and one and $N(n)$ denotes the neighbors of node n. If this is iterated enough times, the weights in a graph become equal for all connected nodes in the graph, but if stopped earlier, it is just a slight smoothing, depending on the value of α.

To implement this, we need both a representation of graphs and the smoothing algorithm. We start with representing the graph. There are many ways to do this, but a simple format is a so-called *incidence matrix*. This is a matrix that has entry $M_{i,j} = 0$ if nodes i and j are not directly connected and $M_{i,j} = 1$ if they are. Since we want to work on a non-directed graph in this algorithm, we will have $M_{i,j} = M_{j,i}$.

We can implement this representation using a constructor function that looks like this:

```
graph <- function(n, edges) {
  m <- matrix(0, nrow = n, ncol = n)

  no_edges <- length(edges)
  if (no_edges >= 1) {
    for (i in seq(1, no_edges, by = 2)) {
      m[edges[i], edges[i+1]] <- m[edges[i+1], edges[i]] <- 1
    }
  }

  structure(m, class = "graph")
}
```

Here I require that the number of nodes is given as an argument n and that edges are specified as a vector where each pair corresponds to an edge. This is not an optimal way of representing edges if graphs should be coded by hand, but since this algorithm is supposed to be used for very large graphs, I assume we can write code elsewhere for reading in a graph representation and creating such an edge vector.

There is not much to the function. It just creates the incidence matrix and then iterates through the edges to set it up. There is a special case to handle if the edges vector is empty. Then the seq() call will return a list going from one to zero. So we avoid this. We might also want to check that the length of the edge vector is a multiple of two, but I haven't bothered. I am going to assume that the code that generates the vector will take care of that.

Even though the graph representation is just a matrix, I give it a class in case I want to write generic functions for it later.

With this graph representation, the smoothing function can look like this:

```
smooth_weights <- function(graph, node_weights, alpha) {
  if (length(node_weights) != nrow(graph))
    stop("Incorrect number of nodes")

  no_nodes <- length(node_weights)
  new_weights <- vector("numeric", no_nodes)

  for (i in 1:no_nodes) {
    neighbour_weights <- 0
    n <- 0
    for (j in 1:no_nodes) {
      if (i != j && graph[i, j] == 1) {
        neighbour_weights <- neighbour_weights + node_weights[j]
        n <- n + 1
      }
    }

    if (n > 0) {
      new_weights[i] <-
        alpha * node_weights[i] +
        (1 - alpha) * neighbour_weights / n
    } else {
      new_weights[i] <- node_weights[i]
    }

  }
  new_weights
}
```

It creates the new weights vector we should return and then iterate through the matrix in nested loops. If the incidence matrix says that there is a connection between *i* and *j*, and *i ≠ j* –, we don't want to add a node's own weight if there is a self-loop—we use it to calculate the mean. If there is something to update—which there will be if there are any neighbors to node *i*, we do the update.

The code is not particularly elegant, but it is a straightforward implementation of the idea.

To profile this code, we use the profvis() function from profvis. It takes an expression as its single argument so to profile more than a single function call we give it a code block, translating the sequence of statements into an expression.

I just generate a random graph with 1000 nodes and 300 edges and random weights. We are not testing the code here, only profiling it. While if this was real code and not just an example we should, of course, have unit tests—this is especially important if you start rewriting code to optimize it. Otherwise, you might end up getting faster but incorrect code for all your efforts.

```
profvis::profvis({
  n <- 1000
  nodes <- 1:n
  edges <- sample(nodes, 600, replace = TRUE)
  weights <- rnorm(n)
  g <- graph(n, edges)
  smooth_weights(g, weights, 0.8)
})
```

Running this code will open a new tab showing the results, as shown in Figure 14-1. The top half of the tab shows your code with annotations showing memory usage first and time usage second as horizontal bars. The bottom half of the window shows the time usage plus callstack.
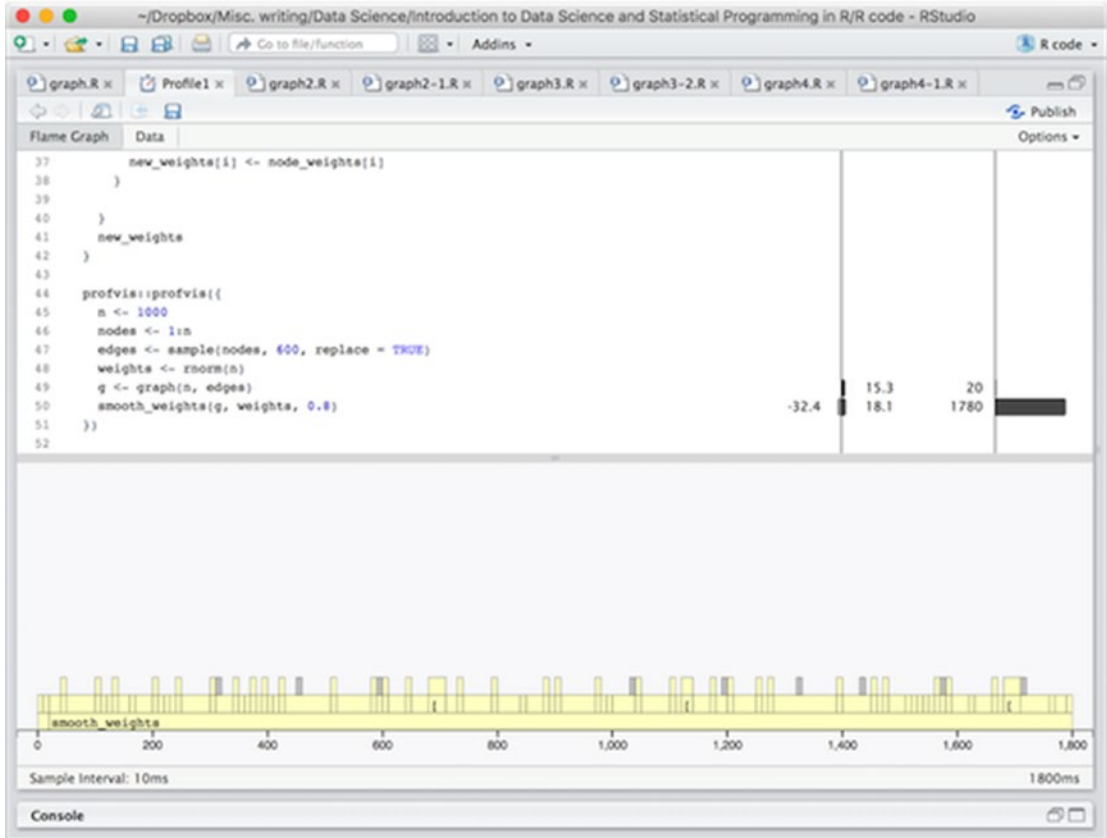
Figure 14-1. *Window showing profile results*

We can see that the total execution took about 1800 ms. The way to read the graph is that, from left to right, you can see what was executed at any point in the run with functions called directly in the code block we gave profvis() at the bottom and code they called directly above that and further function calls stacked even higher.

We can also see that by far the most time was spent in the smooth_weights() function since that stretches almost all the way from the leftmost part of the graph and all the way to the rightmost.

If you move your mouse pointer into the window, either in the code or in the bottom graph, it will highlight what you are pointing at, as shown in Figure 14-2. You can use this to figure out where the time is being spent.
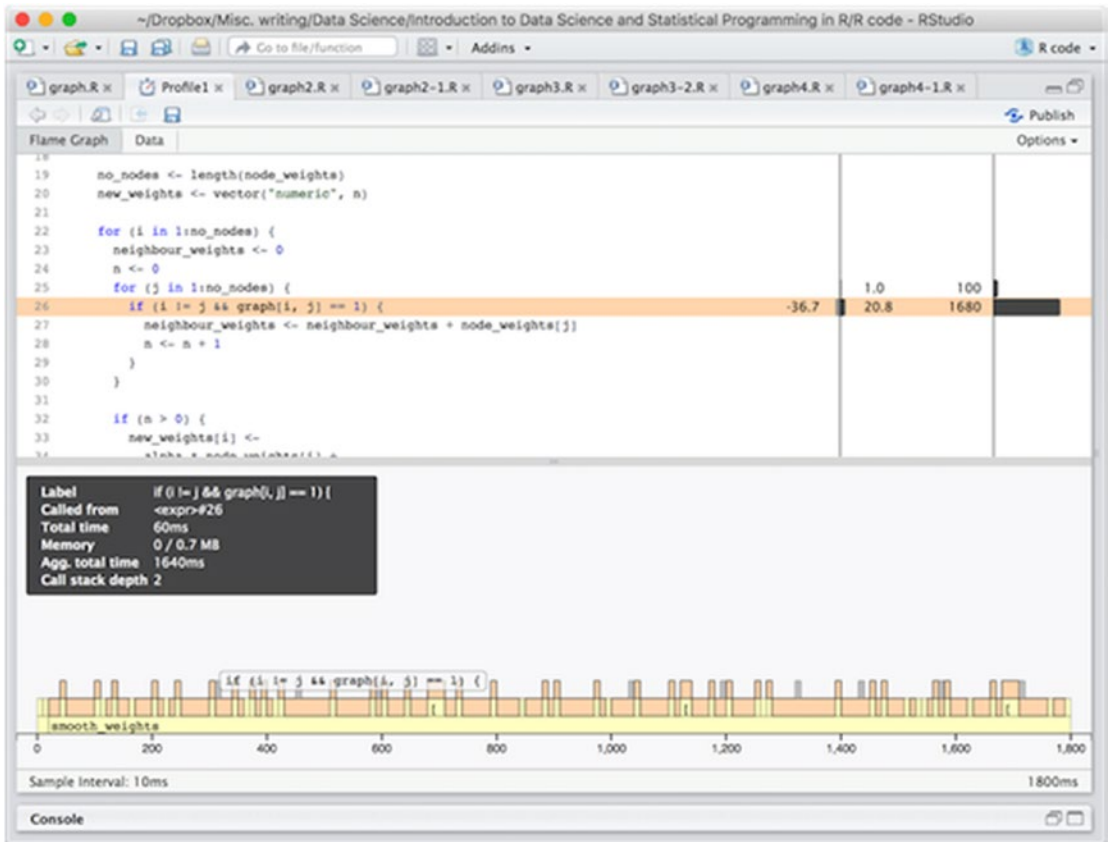
Figure 14-2. *Highlighting executing code from the profiling window*

In this particular case, it looks like most of the time is spent in the inner loop, checking if an edge exists. Since this is the inner part of a double loop, this might not be so surprising. The reason that it is not all the body of the inner loop, but the `if` statement is probably that we check the `if` expression in each iteration but we do not execute its body unless it is true. And with 1000 nodes and 300 edges it is only true with probability around $300/(1000*1000) = 3 \times 10\text{-}4$ (it can be less since some edges could be identical or self-loops).

Now if we had a performance problem with this code, this is where we should concentrate our optimization efforts. With 1000 nodes we don't really have a problem. 1800 ms is not a long time, after all. But the application I have in mind has around 30,000 nodes so it might be worth optimizing a little bit.

If you need to optimize something, the first you should be thinking is—is there a better algorithm or a better data structure? Algorithmic improvements are much more likely to give substantial performance improvements compared to just changing details of an implementation.

In this case, if the graphs we are working on are sparse, meaning they have few actual edges compared to all possible edges, then an incidence matrix is not a good representation. We *could* speed the code up by using vector expressions to replace the inner loop and hacks like that, but we are much better off considering another representation of the graph.

Here, of course, we should first figure out if the simulated data we have used is representative of the actual data we need to analyze. If the actual data is a dense graph and we do performance profiling on a sparse graph, we are not getting the right impression of where the time is being spent and where we can reasonably optimize. But the application I have in mind, I claim, is one that uses sparse graphs.

With sparse graphs, we should represent edges in a different format. Instead of a matrix, we will represent the edges as a list where, for each node, we have a vector of that node's neighbors.

We can implement that representation like this:

```
graph <- function(n, edges) {
  neighbours <- vector("list", length = n)

  for (i in seq_along(neighbours)) {
    neighbours[[i]] <- vector("integer", length = 0)
  }

  no_edges <- length(edges)
  if (no_edges >= 1) {
    for (i in seq(1, no_edges, by = 2)) {
      n1 <- edges[i]
      n2 <- edges[i+1]
      neighbours[[n1]] <- c(n2, neighbours[[n1]])
      neighbours[[n2]] <- c(n1, neighbours[[n2]])
    }
  }

  for (i in seq_along(neighbours)) {
    neighbours[[i]] <- unique(neighbours[[i]])
  }

  structure(neighbours, class = "graph")
}
```

We first generate the list of edge vectors, then we initialize them all as empty integer vectors. We then iterate through the input edges and updating the edge vectors. The way we update the vectors is potentially computationally slow since we force a copy of the previous vector in each update, but we don't know the length of these vectors a priori, so this is the easy solution, and we can worry about it later if the profiling says it is a problem.

Now, if the edges we get as input contains the same pair of nodes twice, we will get the same edge represented twice. This means that the same neighbor to a node will be used twice when calculating the mean of the neighbor weights. If we want to allow such multi-edges in the application that is fine, but we don't, so we explicitly make sure that the same neighbor is only represented once by calling the unique() function on all the vectors at the end.

With this graph representation, we can update the smoothing function to this:

```
smooth_weights <- function(graph, node_weights, alpha) {
  if (length(node_weights) != length(graph))
    stop("Incorrect number of nodes")

  no_nodes <- length(node_weights)
  new_weights <- vector("numeric", no_nodes)

  for (i in 1:no_nodes) {
    neighbour_weights <- 0
    n <- 0
    for (j in graph[[i]]) {
      if (i != j) {
```

```
      neighbour_weights <- neighbour_weights + node_weights[j]
      n <- n + 1
    }
  }

  if (n > 0) {
    new_weights[i] <-
      alpha * node_weights[i] +
      (1 - alpha) * neighbour_weights / n
  } else {
    new_weights[i] <- node_weights[i]
  }

  }
  new_weights
}
```

Very little changes. We just make sure that *j* only iterates through the nodes we know to be neighbors of node *i*.

The profiling code is the same as before, and if we run it, we get the results shown in Figure 14-3.
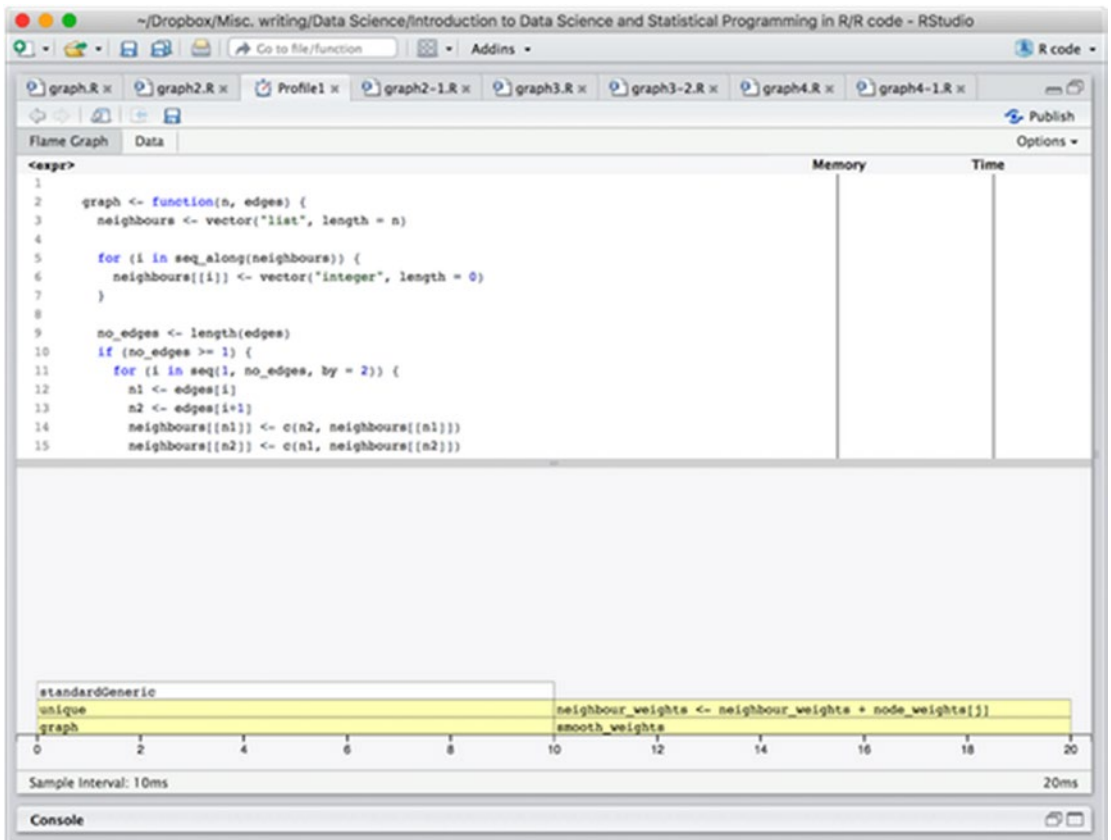
Figure 14-3. Profiling results after the first change

We see that we got a substantial performance improvement. The execution time is now 20 ms instead of 1800 ms. We can also see that half the time is spent on constructing the graph and only half on smoothing it. In the construction, nearly all the time is spent in unique() while in the smoothing function, the time is spent in actually computing the mean of the neighbors.

It should be said here, though, that the profiler works by sampling what code is executing at certain time points. It doesn't have an infinite resolution, it samples every 10 ms as it says at the bottom left, so in fact, it has only sampled twice in this run. The result we see is just because the samples happened to hit those two places in the graph construction and the smoothing, respectively. We are not actually seeing fine details here.

To get more details, and get closer to the size the actual input is expected to be, we can try increasing the size of the graph to 10,000 nodes and 600 edges.

```
profvis::profvis({
  n <- 10000
  nodes <- 1:n
  edges <- sample(nodes, 1200, replace = TRUE)
  weights <- rnorm(n)
  g <- graph(n, edges)
  smooth_weights(g, weights, 0.8)
})
```

The result of this profiling is shown in Figure 14-4.
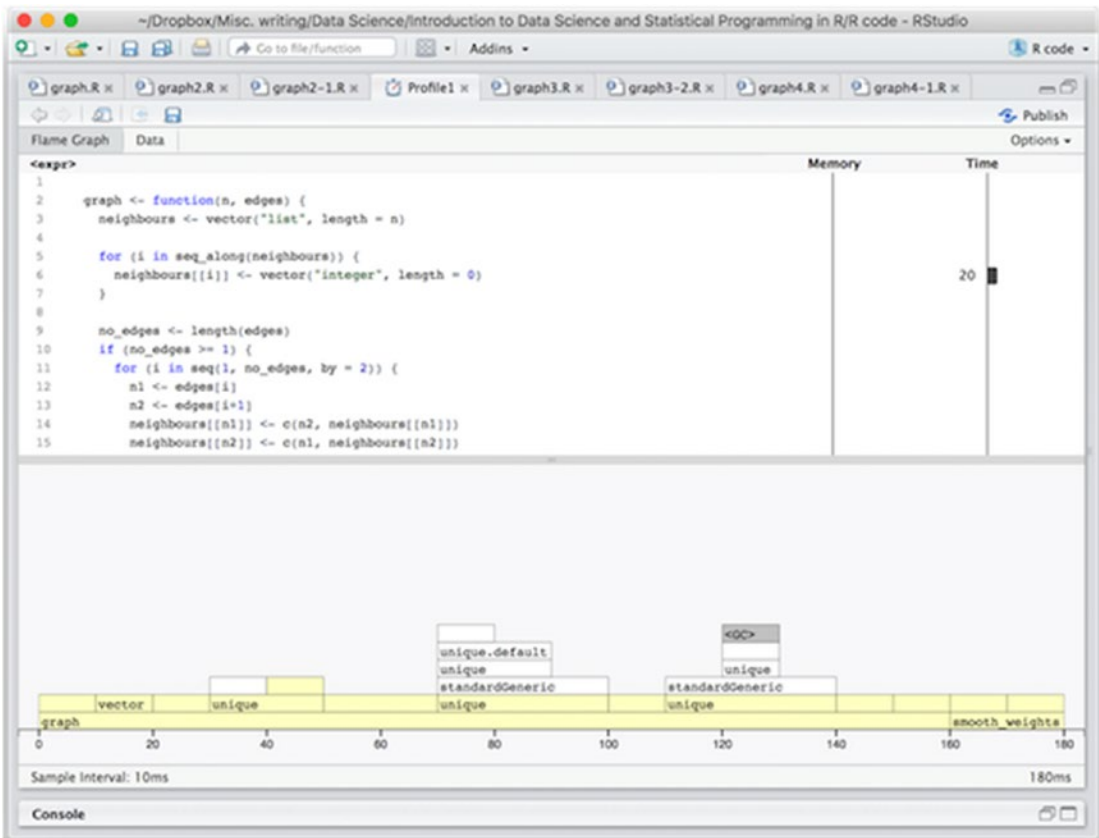
Figure 14-4. *Profiling results with a larger graph*

To our surprise, we see that for the larger graph we are actually spending more time constructing the graph than smoothing it. We also see that this time is spent calling the unique() function.

Now, these calls are necessary to avoid duplicated edges, but they are not necessarily going to be something we often see—in the random graph they will be very unlikely, at least—so most of these calls are not really doing anything.

If we could remove all the duplicated edges in a single call to unique() we should save some time. We can do this, but it requires a little more work in the construction function.

We want to make the edges unique, and there are two issues here. One is that we don't actually represent them as pairs we can call unique() on, and calling unique() on the edges vector is certainly not a solution. The other issue is that the same edge can be represented in two different ways: (i, j) and (j, i).

We can solve the first problem by translating the vector into a matrix. If we call unique() on a matrix we get the unique rows, so we just represent the pairs in that way. The second issue we can solve by making sure that edges are represented in a canonical form, say requiring that $i < j$ for edges (i, j).

```r
graph <- function(n, edges) {
  neighbours <- vector("list", length = n)

  for (i in seq_along(neighbours)) {
    neighbours[[i]] <- vector("integer", length = 0)
  }

  no_edges <- length(edges)
  if (no_edges >= 1) {
    sources <- seq(1, no_edges, by = 2)
    destinations <- seq(2, no_edges, by = 2)

    edge_matrix <- matrix(NA, nrow = length(sources), ncol = 2)
    edge_matrix[,1] <- edges[sources]
    edge_matrix[,2] <- edges[destinations]

    for (i in 1:nrow(edge_matrix)) {
      if (edge_matrix[i,1] > edge_matrix[i,2]) {
        edge_matrix[i,] <- c(edge_matrix[i,2], edge_matrix[i,1])
      }
    }

    edge_matrix <- unique(edge_matrix)

    for (i in seq(1, nrow(edge_matrix))) {
      n1 <- edge_matrix[i, 1]
      n2 <- edge_matrix[i, 2]
      neighbours[[n1]] <- c(n2, neighbours[[n1]])
      neighbours[[n2]] <- c(n1, neighbours[[n2]])
    }
  }

  structure(neighbours, class = "graph")
}
```

The running time is cut in half and relatively less time is spent constructing the graph compared to before. The time spent in executing the code is also so short again that we cannot be too certain about the profiling samples to say much more.

The graph size is not quite at the expected size for the application I had in mind when I wrote this code. We can boost it up to the full size of around 20,000 nodes and 50,000 edges and profile for that size. Results are shown in Figure 14-5.
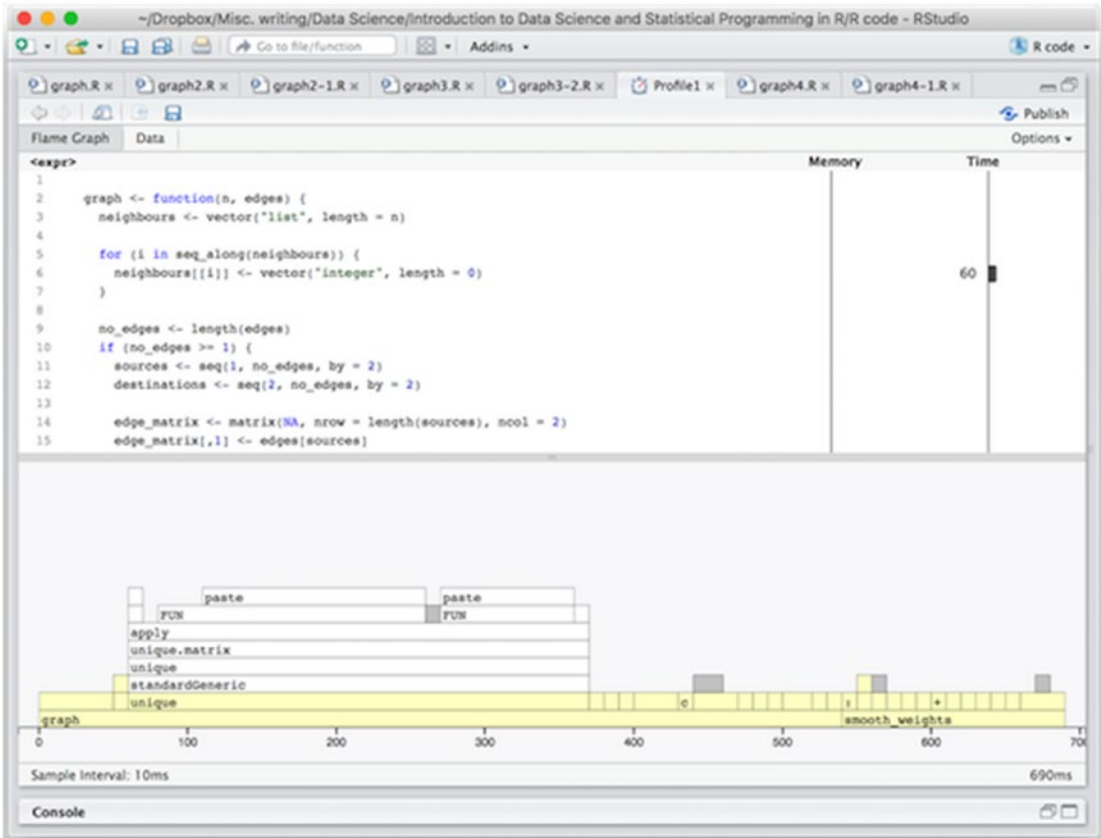
Figure 14-5. *Profiling results on a full-time graph*

On a full-size graph, we still spend most of the time in constructing the graph and not in smoothing it—and about half of the constructing time in the `unique()` function—but this is a little misleading. We don't expect to call the smoothing function just once on a graph. Each call to the smoothing function will smooth the weights out a little more, and we expect to run it around ten times, say, in the real application.

We can rename the function to `flow_weights_iteration()` and then write a `smooth_weights()` function that runs it for a number of iterations:

```
flow_weights_iteration <- function(graph, node_weights, alpha) {
  if (length(node_weights) != length(graph))
    stop("Incorrect number of nodes")
```

```
  no_nodes <- length(node_weights)
  new_weights <- vector("numeric", n)

  for (i in 1:no_nodes) {
    neighbour_weights <- 0
    n <- 0
    for (j in graph[[i]]) {
      if (i != j) {
        neighbour_weights <- neighbour_weights + node_weights[j]
        n <- n + 1
      }
    }

    if (n > 0) {
      new_weights[i] <- (alpha * node_weights[i] + (1 - alpha)
                          * neighbour_weights / n)
    } else {
      new_weights[i] <- node_weights[i]
    }

  }
  new_weights
}
smooth_weights <- function(graph, node_weights, alpha, no_iterations) {
  new_weights <- node_weights
  replicate(no_iterations, {
      new_weights <- flow_weights_iteration(graph, new_weights, alpha)
  })
  new_weights
}
```

We can then profile with 10 iterations:

```
profvis::profvis({
  n <- 20000
  nodes <- 1:n
  edges <- sample(nodes, 100000, replace = TRUE)
  weights <- rnorm(n)
  g <- graph(n, edges)
  smooth_weights(g, weights, 0.8, 10)
})
```

The results are shown in Figure 14-6. Obviously, if we run the smoothing function more times the smoothing is going to take up more of the total time, so there are no real surprises here. There aren't really any obvious hotspots any longer to dig into. I used the replicate() function for the iterations, and it does have a little overhead because it does more than just loop—it creates a vector of the results—and I can gain a few more milliseconds by replacing it with an explicit loop:

```
smooth_weights <- function(graph, node_weights,
                            alpha, no_iterations) {
  new_weights <- node_weights
```

```
  for (i in 1:no_iterations) {
    new_weights <-
      smooth_weights_iteration(graph, new_weights, alpha)
  }
  new_weights
}
```

I haven't shown the results, so you have to trust me on that. There is nothing major to attack any longer, now, however.
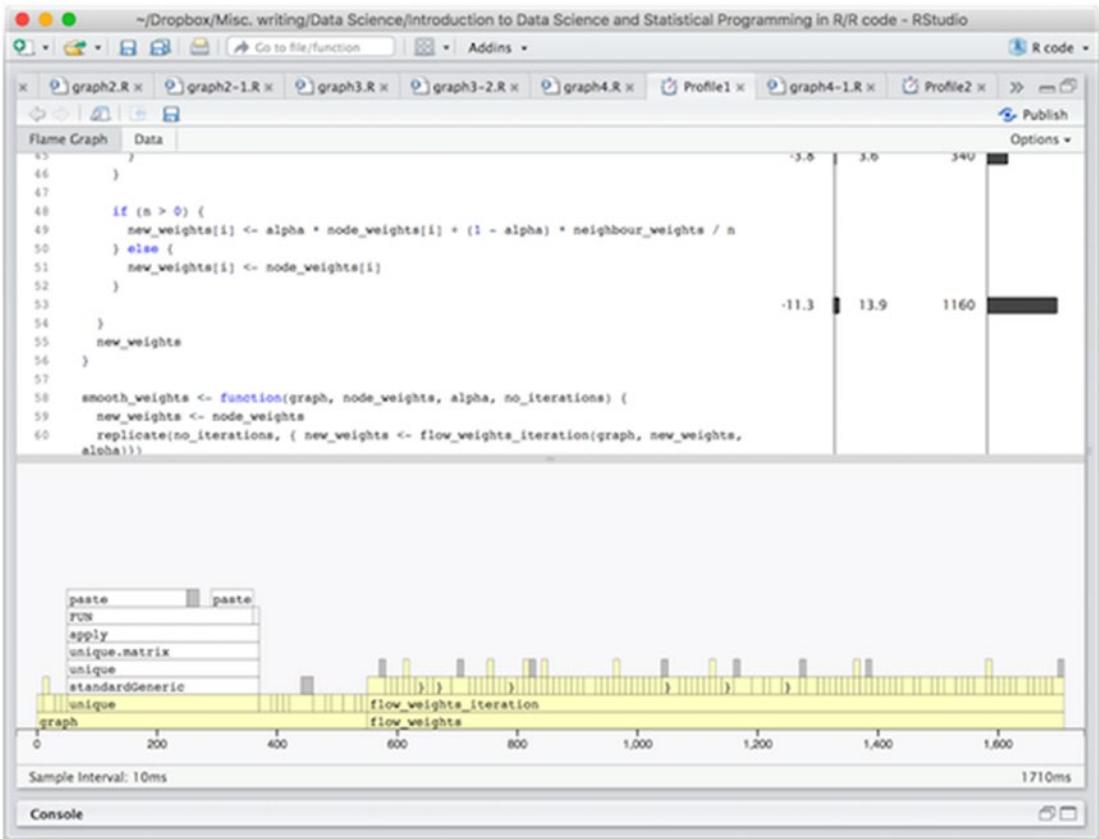
Figure 14-6. *Profiling results with multiple smoothing iterations*

If you are in that situation where there is nothing more obvious to try to speed up, you have to consider if any more optimization is really necessary. From this point an onwards, unless you can come up with a better algorithm, which is hard, further optimizations are going to be very hard and unlikely to be worth the effort. You are probably better off spending your time on something else while the computations run than wasting days on trying to squeeze a little more performance out of it.

Of course, in some cases, you really *have* to improve performance more to do your analysis in reasonable time, and there are some last resorts you can go to such as parallelizing your code or moving time-critical parts of it to C++. But for now, we can analyze full graphs in fewer than two seconds so we definitely should not spend more time on optimizing this particular code.

Speeding Up Your Code

If you really do have a performance problem, what do you do? I will assume that you are not working on a problem that other people have already solved—if there is already a package available you could have used then you should have used it instead of writing your own code, of course. But there might be similar problems you can adapt to your needs, so before you do anything else, do a little bit of research to find out if anyone else has solved a similar problem, and if so, how they did it. There are very few really unique problems in life, and it is silly not to learn from others' experiences.

It can take a little time to figure out what to search for, though, since similar problems pop up in very different fields. There might be a solution out there that you just don't know how to search for because it is described in terms entirely different from your own field. It might help to ask on mailing lists or stack overflow (see http://stackoverflow.com), but don't burn your Karma by asking help with every little thing you should be able to figure out yourself with a little bit of work.

If you really cannot find an existing solution you can adapt, the next step is to start thinking about algorithms and data structures. Improving these usually have *much* more of an impact on performance than micro-optimizations ever can. Your first attempts at any optimization should be to figure out if you could use better data structures or better algorithms.

It is, of course, a more daunting task to reimplement complex data structures or algorithms—and you shouldn't if you can find solutions already implemented—but it is usually where you gain the most performance. Of course, there is always a trade-off between how much time you spend on reimplementing an algorithm versus how much you gain, but with experience, you will get better at judging this. Well, slightly better. If in doubt, it is probably better to live with slow code than spend a lot of time trying to improve it.

And before you do *anything* make sure you have unit tests that ensure that new implementations do not break old functionality! Your new code can be as fast as lightning, and it is worthless if it isn't correct.

If you have explored existing packages and new algorithms and data structures and there still is a performance problem you reach the level of micro-optimizations. This is where you use slightly different functions and expressions to try to improve the performance, and you are not likely to get massive improvements at this level of changes. But if you have code that is executed thousands or millions of times, those small gains can still stack up. So if your profiling highlights a few hotspots for performance you can try to rewrite code there.

The sampling profiler is not terribly useful at this level of optimization. It samples at the level of milliseconds, and that is typically a much coarser grained measurement than what you need here. Instead, you can use the `microbenchmark` package that lets you evaluate and compare expressions. The `microbenchmark()` function runs a sequence of expressions several times and computes statistics on the execution time in units down to nanoseconds. If you want to gain some performance through micro-optimization, you can use it to evaluate different alternatives to your computations.

For example, we can use it to compare an R implementation of `sum()` against the built-in `sum()` function:

```
library(microbenchmark)
mysum <- function(sequence) {
  s <- 0
  for (x in sequence) s <- s + x
  s
}

microbenchmark(
  sum(1:10),
  mysum(1:10)
)
## Unit: nanoseconds
```

```
##            expr  min   lq    mean median     uq
##     sum(1:10)  194  202  300.10  233.5  349.5
##  mysum(1:10) 1396 1592 2280.47 1750.0 1966.5
##    max neval cld
##   2107   100  a
##  11511   100   b
```

The first column in the output is the expressions evaluated, then you have the minimum, lower quarter, mean, median, upper quarter, and maximum time observed when evaluating it, and then the number of evaluations used. The last column ranks the performance, here showing that sum() is a and mysum() is b so the first is faster. This ranking takes the variation in evaluation time into account and does not just rank by the mean.

There are a few rules of thumbs for speeding up the code in micro-optimization, but you should always measure. Intuition is often a quite bad substitute for measurement.

One rule of thumb is to use built-in functions when you can. Functions such as sum() are actually implemented in C and highly optimized, so your own implementation will have a hard time competing with it, as you saw previously.

Another rule of thumb is to use the simplest functions that get the work done. More general functions introduce various overheads that simpler functions avoid.

You can add together all numbers in a sequence using Reduce(), but using such a general function is going to be relatively slow compared to specialized functions.

```
microbenchmark(
  sum(1:10),
  mysum(1:10),
  Reduce(`+`, 1:10, 0)
)
## Unit: nanoseconds
##                     expr  min   lq    mean median
##                sum(1:10)  207  258  356.03  324.5
##              mysum(1:10) 1611 1892 2667.25 2111.0
##  Reduce(`+`, 1:10, 0) 4485 5285 6593.07 6092.0
##      uq    max neval cld
##   409.0   1643   100 a
##  2369.0  11455   100  b
##  6662.5  15497   100   c
```

We use such general functions for programming convenience. They give us abstract building blocks. We rarely get performance boosts out of them and sometimes they can slow things down.

Thirdly, do as little as you can get away with. Many functions in R have more functionality than we necessarily think about. A function such as read.table() not only reads in data, it also figures out what type each column should have. If you tell it what the types of each column are using the colClasses argument, it gets much faster because it doesn't have to figure it out itself. For factor() you can give it the allowed categories using the levels argument so it doesn't have to work it out itself.

```
x <- sample(LETTERS, 1000, replace = TRUE)
microbenchmark(
  factor(x, levels = LETTERS),
  factor(x)
)
## Unit: microseconds
```

```
##                                    expr    min      lq
##   factor(x, levels = LETTERS) 19.211 20.8975
##                     factor(x) 59.458 61.9575
##     mean  median     uq     max neval cld
##  22.03447 21.6175 22.610  32.981   100  a
##  66.70901 62.9135 67.946 132.306   100   b
```

It is not just when providing input, to help functions avoid figuring something out, this is in effect. Functions often also return more than you are necessarily interested in. Functions like unlist(), for instance, will preserve the names of a list into the resulting vector. Unless you really need those names, you should get rid of them since it is expensive dragging those names along with you. If you are just interested in a numerical vector, you should use use.names = FALSE:

```r
x <- rnorm(1000)
names(x) <- paste("n", 1:1000)
microbenchmark(
  unlist(Map(function(x) x**2, x), use.names = FALSE),
  unlist(Map(function(x) x**2, x))
)
## Unit: microseconds
##                                                 expr
##   unlist(Map(function(x) x^2, x), use.names = FALSE)
##                     unlist(Map(function(x) x^2, x))
##     min      lq     mean   median      uq
##  484.866 574.248 704.2379 660.3140 716.2325
##  659.355 722.974 825.7712 813.3855 891.4630
##       max neval cld
##  3141.598   100  a
##  1477.028   100   b
```

Fourthly, when you can, use vector expressions instead of loops. Not just because this makes the code easier to read but because the implicit loop in vector expressions is handled much faster by the runtime system of R than your explicit loops will.

Most importantly, though, is to *always* measure when you try to improve performance and only replace simple code with more complex code if there is a substantial improvement that makes this worthwhile.

Parallel Execution

Sometimes you can speed things up, not by doing them faster, but by doing many things in parallel. Most computers today have more than one core, which means that you should be able to run more computations in parallel.

These are usually based on some variation of lapply() or Map() or similar, see the parallel package as an example, but also check the foreach package, which provides a higher level looping construct that can also be used to run code in parallel.

If we consider our graph smoothing, we could think that since each node is an independent computation we should be able to speed the function up by running these calculations in parallel. If we move the inner loop into a local function, we can replace the outer look with a call to Map():

```r
smooth_weights_iteration_map <- function(graph, node_weights, alpha) {
  if (length(node_weights) != length(graph))
    stop("Incorrect number of nodes")
```

317

```
  handle_i <- function(i) {
    neighbour_weights <- 0
    n <- 0
    for (j in graph[[i]]) {
      if (i != j) {
        neighbour_weights <- neighbour_weights + node_weights[j]
        n <- n + 1
      }
    }

    if (n > 0) {
      alpha * node_weights[i] + (1 - alpha) * neighbour_weights / n
    } else {
      node_weights[i]
    }
  }

  unlist(Map(handle_i, 1:length(node_weights)))
}
```

This is not likely to speed anything up—the extra overhead in the high-level Map() function will do the opposite if anything—but it lets us replace Map() with one of the functions from parallel, for example clusterMap():

```
  unlist(clusterMap(cl, inner_loop, 1:length(node_weights)))
```

Here cl is the "cluster" that just consists of two cores I have on my laptop:

```
cl <- makeCluster(2, type = "FORK")
microbenchmark(
  original_smooth(),
  using_map(),
  using_cluster_map(),
  times = 5
)
```

Where the three functions refer to the three different versions of the algorithm, gave me these result. On my two-core laptop, we could expect the parallel version to run up to two times faster. In fact, it runs several orders of magnitude slower:

```
Unit: milliseconds
                expr         min           lq
   original_smooth()    33.58665     33.73139
        using_map()    33.12904     34.84107
 using_cluster_map() 14261.97728 14442.85032
       mean        median           uq          max
    35.88307     34.25118     36.62977     41.21634
    38.31528     40.50315     41.28726     41.81587
 15198.55138 14556.09176 14913.24566 17818.59187
 neval cld
     5   a
     5   a
     5    b
```

Using type FORK for setting up the cluster only works on UNIX machines, so on Windows you will have to use another type. Using the default parameter, PSOCK, you can set up parallelization but the different cores that run your computations will not be aware of libraries you have imported or functions you have defined in the main script. If you cannot use the type FORK you will need to explicitly inform the cores of which values and functions they should be aware of. Check the documentation for the functions clusterExport and clusterCall.

I am not entirely sure what the problem we are seeing here is, but most likely the individual tasks are very short, and the communication overhead between threads (which are actually processes here) ends up taking much more time than the actual computation. At least my profiling seems to indicate that. With really lightweight threads some of the communication could be avoided, but that is not what we have here.

Parallelization works better when each task runs longer so the threads don't have to communicate so often.

For an example where parallelization works better, we can consider fitting a model on training data and testing its accuracy on test data. We can use the cars data we have looked at before and the partition() function from Chapter 6.

We write a function that evaluates a single train/test partition and then calls it n times, either sequentially or in parallel.

```
test_rmse <- function(data) {
  model <- data$training %>% lm(dist ~ speed, data = .)
  predictions <- data$test %>% predict(model, data = .)
  rmse(data$test$dist, predictions)
}

sample_rmse <- function (n) {
  random_cars <- cars %>%
    partition(n, c(training = 0.5, test = 0.5))
  unlist(Map(test_rmse, random_cars))
}

sample_rmse_parallel <- function (n) {
  random_cars <- cars %>%
    partition(n, c(training = 0.5, test = 0.5))
  unlist(clusterMap(cl, test_rmse, random_cars))
}
```

When I do this for 10 training/test partitions, the two functions take roughly the same time. Maybe the parallel version is a *little* slower, but it is not much overhead this time.

```
microbenchmark(
  sample_rmse(10),
  sample_rmse_parallel(10),
  times = 5
)
Unit: milliseconds
                     expr      min       lq
          sample_rmse(10) 28.72092 29.62857
 sample_rmse_parallel(10) 26.08682 27.15047
    mean   median       uq      max neval cld
 31.57316 33.05759 33.21979 33.23894     5    a
 34.75991 28.17528 29.37144 63.01556     5    a
```

If I create 1000 train/test partitions instead, however, the parallel version starts running faster than the sequential version.

```
microbenchmark(
  sample_rmse(1000),
  sample_rmse_parallel(1000),
  times = 5
)
Unit: seconds
                      expr      min       lq
          sample_rmse(1000) 3.229113 3.277292
 sample_rmse_parallel(1000) 2.570328 2.595402
     mean   median       uq      max neval cld
 3.459333 3.508533 3.536792 3.744934     5   b
 2.921574 2.721095 3.185070 3.535977     5   a
```

Since my laptop only has two cores it will never be able to run more than twice as fast and reaching the possible optimal speed-up from parallelization is rarely possible. The communication overhead between threads adds to the time used for the parallel version, and there are parts of the code that just has to be sequential such as preparing data that all threads should work on.

If you have a machine with many cores, and you can split your analysis into reasonably large independent chunks, though, there is often something to be gained.

Switching to C++

This is a drastic step, but by switching to a language such as C++, you have more fine-grained control over the computer, just because you can program at a much lower level, and you do not have the overhead from the runtime system that R does. Of course, this also means that you don't have many of the features that R does either, so you don't want to program an entire analysis in C++, but you might want to translate the time-critical code to C++.

Luckily, the Rcpp package makes integrating R and C++ pretty straightforward. Assuming that you can program in both languages, of course. The only thing to really be careful about is that C++ index from 0 and R from 1. Rcpp takes care of converting this so a 1 indexed vector from R can be accessed as a 0 indexed vector in C++, but when translating code you have to keep it in mind.

A full description of this framework for communicating between C++ and R is far beyond the scope of this book. For that, I will refer you to the excellent book, *Seamless R and C++ Integration with Rcpp* by Dirk Eddelbuettel. Here I will just give you a taste of how Rcpp can be used to speed up a function.

We will focus on the smoothing function again. It is a relatively simple function that is not using any of R's advanced features so it is ideal to translate into C++. We can do so almost verbatim, just remembering that we should index from zero instead of one.

```
NumericVector
smooth_weights_iteration_cpp(List g,
                             NumericVector node_weights,
                             double alpha)
{
  NumericVector new_weights(g.length());

  for (int i = 0; i < g.length(); ++i) {
```

```
    IntegerVector neighbours = g[i];
    double neighbour_sum = 0.0;
    int n = 0;

    for (int j = 0; j < neighbours.length(); ++j) {
      neighbour_sum += node_weights[j];
      ++n;
    }

    if (n > 0) {
      new_weights[i] = alpha * node_weights[i] +
        (1-alpha) * (neighbour_sum / n);
    } else {
      new_weights[i] = node_weights[i];
    }
  }

  return new_weights;
}
```

The types List, NumericVector, and IntegerVector correspond to the R types and except for how we create the new_weights vector the code very closely follows the R code.

There are several ways you can compile this function and wrap it into an R function, but the easiest is just to put it in a string and give it to the function cppFunction():

```
cppFunction("
NumericVector
smooth_weights_iteration_cpp(List g,
                             NumericVector node_weights,
                             double alpha)
{
  NumericVector new_weights(g.length());

  for (int i = 0; i < g.length(); ++i) {

    // The body here is just the C++ code
    // shown above...

  }

  return new_weights;
}
")
```

That creates a function with the same name as the C++ function that can be called directly from R and Rcpp will take care of converting types as needed.

```
smooth_weights_cpp <- function(graph, node_weights,
                               alpha, no_iterations) {
  new_weights <- node_weights
  for (i in 1:no_iterations) {
```

```
    new_weights <-
      smooth_weights_iteration_cpp(graph, new_weights, alpha)
  }
  new_weights
}
```

If we compare the R and C++ function, we see that we get a substantial performance boost from this.

```
microbenchmark(
  smooth_weights(g, weights, 0.8, 10),
  smooth_weights_cpp(g, weights, 0.8, 10),
  times = 5
)
Unit: milliseconds
                                    expr
     smooth_weights(g, weights, 0.8, 10)
 smooth_weights_cpp(g, weights, 0.8, 10)
       min          lq       mean      median
 1561.78635 1570.23346 1629.12784 1572.3979
   32.77344   33.38822   35.57017   36.5103
        uq         max neval cld
 1694.31571 1746.90573     5   b
   37.29083   37.88803     5   a
```

To translate a function into C++, you are not necessarily prevented from using R's more advanced features. You can call R functions from C++ just as easily as you can call C++ functions from R. Using R types translated into C++ can in many cases be used with vector expressions just like in R. Be aware, though, that the runtime overhead of using advanced features are the same when you use them in C++ as in R. You will likely not gain much performance from translating such functions. Translating low-level code like nested loops often gives you substantial performance boosts, though. If you have a few performance hotspots in your code that are relatively simple, just very time-consuming because they do a lot of work, it is worth considering translating these to C++ and Rcpp makes it easy.

Don't go overboard, though. It is harder to profile and debug code in C++, and it is harder to refactor your code if it is a mix of C++ and R. Use it, but use it only when you really need it.

Exercises

Find some code you have written and try to profile it. If there are performance hotspots you can identify, then try to optimize them. First, think algorithmic changes, then changes to the R expressions—checked using microbenchmark()—and if everything else fails try parallelizing or implementing them in C++.

Project 2

The project for this chapter is building an R package for Bayesian linear regression. The model we will work with is somewhat a toy example of what we could imagine we could build an R package for, and the goal is not to develop all the bells and whistles of Bayesian linear regression. We will just build enough to see the various aspects that go into building a real R package.

Bayesian Linear Regression

In linear regression we assume that we have predictor variables x and target variables y where $y = w_0 + w_1 x$ + ε where $\varepsilon \sim N(0, \sigma^2)$. That is, we have a line with intercept w_0 and incline w_1 such that the target variables are normally distributed around the point given by the line. We sometimes write σ^2 as $1/\beta$ and call β the precision. I will do this here and assume that β is a know quantity; we are going to consider a Bayesian approach to estimating the weights $\boldsymbol{w}^T = (w_0, w_1)$.

Since we assume that we know the precision parameter β, *if* we knew the true weights of the model then whenever we had an x value we would know the distribution of y values: $y \sim N(w_0 + w_1 x, 1/\beta)$.

For notational purposes I am going to define a function that maps x to a vector: $\phi : x \mapsto (1, x)^T$. Then we have $\boldsymbol{w}^T \phi(x) = w_0 + w_1 x$ and $y \sim N(\boldsymbol{w}^T \phi(x), 1/\beta)$.

Of course, we do not know the values of the weights but have to estimate them. In a Bayesian approach, we do not consider the weights as fixed but unknown values; we consider them as random variables from some distribution we have partial knowledge about. Learning about the weights means estimating the posterior distribution for the vector $\mathbf{w}$ conditional on observed x and y values.

We will assume some prior distribution for $\mathbf{w}$, $p(\mathbf{w})$. If we observe a sequence of matching x and y values, $\mathbf{x}^T = (x_1, x_2, ..., x_n)$ and $\mathbf{y}^T = (y_1, y_2, ..., y_n)$, we want to update this prior distribution for the weights $\mathbf{w}$ to the posterior distribution $p(\mathbf{w} \mid \mathbf{x}, \mathbf{y})$

We will assume that the prior distribution of $\mathbf{w}$ is a normal distribution with mean zero and diagonal covariance matrix with some (known hyperparameter) precision α, such as:

$$p(\boldsymbol{w}|\alpha) = N\left(\mathbf{0}, \alpha^{-1} \boldsymbol{I}\right)$$

For reasons that I don't have time or space to go into here, this is a good choice of prior for a linear model since it means that the posterior will also be a normal distribution. It also means that, given $\mathbf{x}$ and $\mathbf{y}$, we can compute the mean and covariance matrix for the posterior with some simple matrix arithmetic.

But first, we need to define our model matrix. This is a matrix that captures that the linear model we are trying to find is a line, i.e., that $y = w_0 + w_1 x$. We define the model matrix for the observed vector $\mathbf{x}$ as such:

$$\phi_x = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

In general, we would have a row per observation with the various features of the observation we want to include in the model, but for a simple line it is the incline and intercept, so for observation i it is 1 and x_i.

The posterior distribution for the weights, $p(\mathbf{w} \mid \mathbf{x}, \mathbf{y}, \alpha, \beta)$, is then given by

$$p(\boldsymbol{w} | \boldsymbol{x}, \boldsymbol{y}, \alpha, \beta) = N\left(\boldsymbol{m}_{x,y}, \boldsymbol{S}_{x,y}\right)$$

where

$$\boldsymbol{m}_{x,y} = \beta \boldsymbol{S}_{x,y} \Phi_x^T \boldsymbol{y}$$

and

$$\boldsymbol{S}_{x,y}^{-1} = \alpha \boldsymbol{I} + \beta \Phi_x^T \Phi_x$$

Exercises: Priors and Posteriors

Try the following exercises to become more comfortable with the concepts discussed in this chapter.

Sample from a Multivariate Normal Distribution

If you want to sample from a multivariate normal distribution, the `mvrnorm` function from the `MASS` package is what you want.

```
library(MASS)
mvrnorm(n = 5, mu = c(0,0), Sigma = diag(1, nrow = 2))
##            [,1]       [,2]
## [1,]  0.6420163 -0.9853573
## [2,]  0.2112605  1.0362092
## [3,]  2.2689703 -0.1181916
## [4,] -0.9177489  0.6836801
## [5,] -0.8123927  0.7117685
```

You need to provide it with a mean vector, `mu`, and a covariance matrix, `Sigma`.

The prior distribution for our weight vectors is $N(\mathbf{0}, \mathbf{S}_0)$ with

$$\mathbf{0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

and

$$\mathbf{S}_0 = \frac{1}{\alpha} I = \begin{pmatrix} 1/\alpha & 0 \\ 0 & 1/\alpha \end{pmatrix}$$

You can use the `diag` function to create the diagonal covariance matrix.

Write a function called `make_prior(alpha)` that constructs this prior distribution and another function `sample_from_prior(n, alpha)` that samples n weight vectors $\mathbf{w}_i$ from it. My version returns the samples as a data frame with a column for the w_1 samples and another for the corresponding w_0 samples. You can return the samples in any form that is convenient for you.

If you can sample from the prior you can plot the results, both as points in $\mathbf{w}$ space and as lines in (x, y) space, as shown in Figures 14-7 and 14-8.
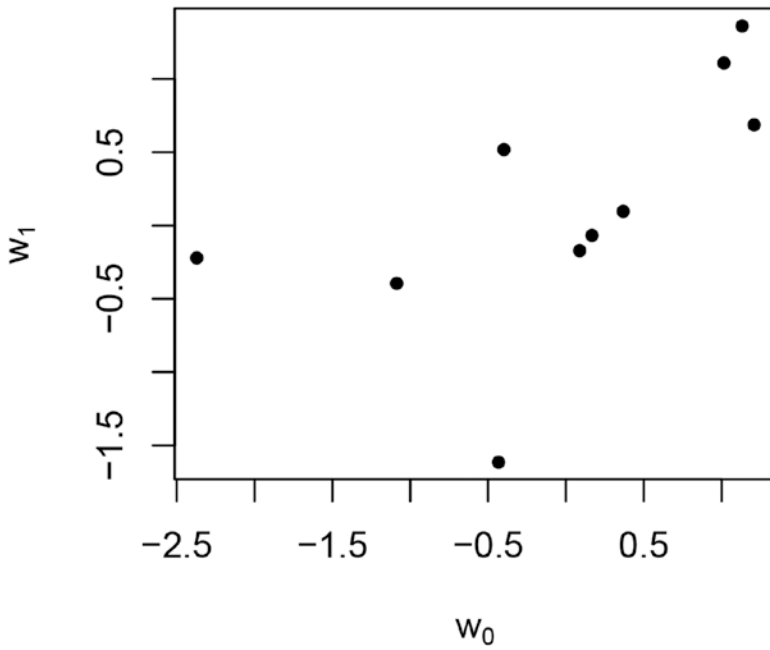
Figure 14-7. *Weight vectors sampled from the prior distribution*
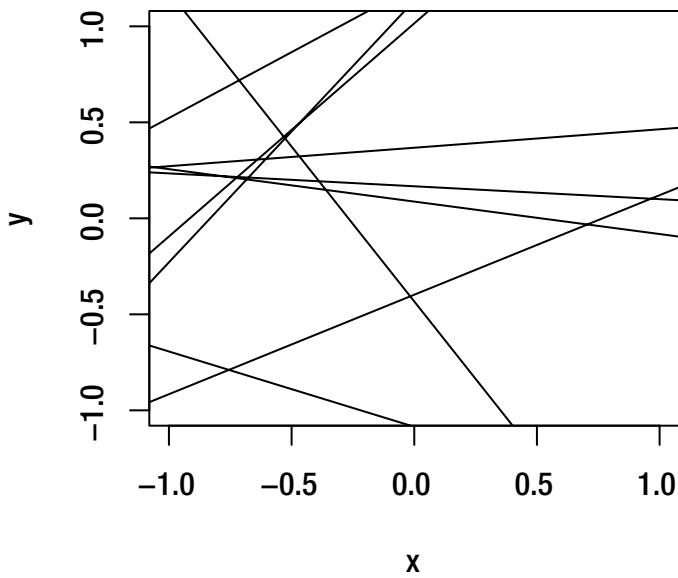
Figure 14-8. *Weight vectors sampled from the prior distribution represented as lines*

Computing the Posterior Distribution

If we fix the parameters of the model, β and $\mathbf{w} = (w_0, w_1)^T$, we can simulate (x, y) values. We can pick some random x values first and then simulate corresponding y values (see Figure 14-9).

```
w0 <- 0.3 ; w1 <- 1.1 ; beta <- 1.3
x <- rnorm(50)
y <- rnorm(50, w1 * x + w0, 1/beta)
```
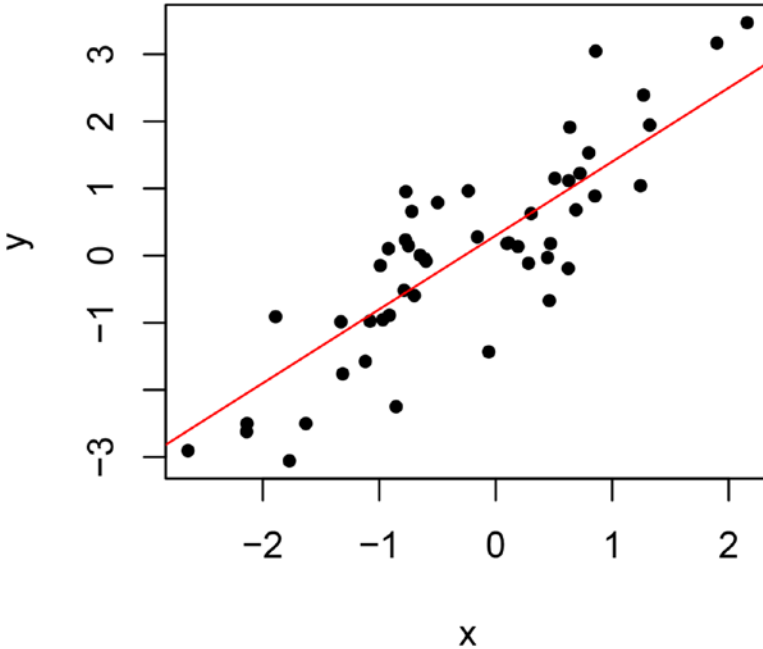

Figure 14-9. *Randomly sampled (x, y) values*

Write a function, `make_posterior(x, y, alpha, beta)`, that computes the posterior distribution for the weights and a function `sample_from_posterior` that lets you sample from the posterior.

Using this sampling function we can see how the posterior distribution gets centered around the real value as the number of (x, y) points increases. In the plots here, I have sampled 10 weights from the posterior in each case but increased the number of (x, y) points the posterior is based on. See Figures 14-10 and 14-11.
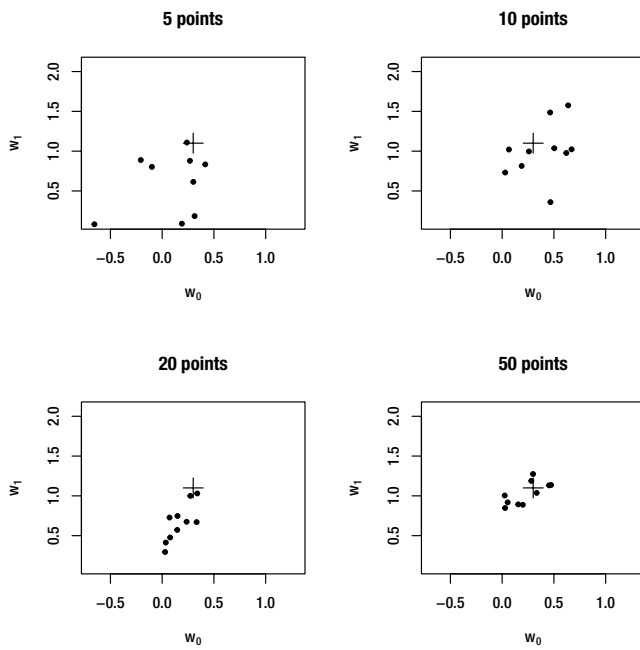
Figure 14-10. *Samples from the posterior. The true value is shown as a cross.*
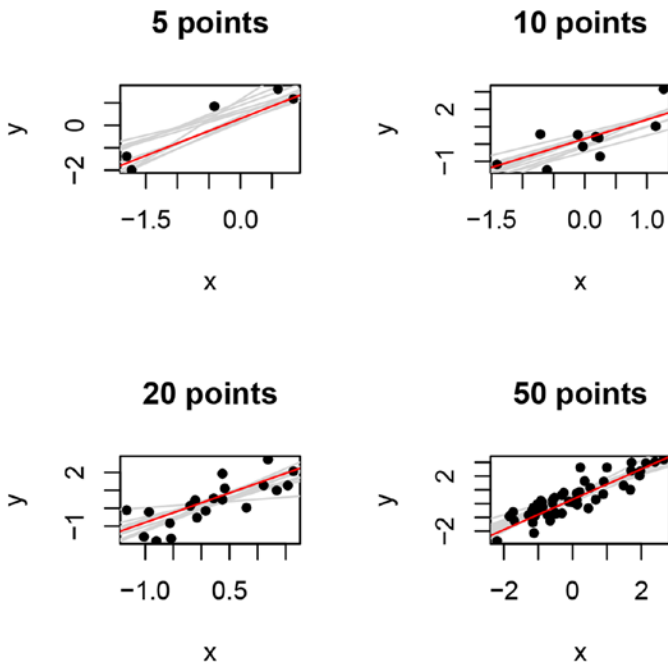
Figure 14-11. *Lines drawn from the posterior. The true line is shown in red.*

327

Predicting Target Variables for New Predictor Values

Given a new value x, we want to make predictions about the corresponding y. For a fixed **w**, again, we have $p(y \mid x, \mathbf{w}, \beta) = \mathrm{N}(\mathbf{w}^T\phi(x), 1/\beta)$, but since we don't know **w** we have to integrate over all **w**. The way the training data improves our prediction is that we integrate over **w** weighted by the posterior distribution of **w** rather than the prior:

$$p(y \mid x, \boldsymbol{x}, \boldsymbol{y}, \alpha, \beta) = \int p(y \mid x, \boldsymbol{w}, \beta) p(\boldsymbol{w} \mid \boldsymbol{x}, \boldsymbol{y}, \alpha, \beta) d\boldsymbol{w}$$

This kind of integral over the product of two normal distributions gives us another normal distribution, and one can show that it is:

$$p(y \mid x, \boldsymbol{x}, \boldsymbol{y}, \alpha, \beta) = N\left(m_{x,y}^T \phi(x), \sigma_{x,y}^2(x)\right)$$

where $m_{x,y}$ is the mean from the posterior distribution of **w** and where

$$\sigma_{x,y}^2(x) = \frac{1}{\beta} + \phi(x)^T \boldsymbol{S}_{x,y} \phi(x)$$

where $\boldsymbol{S}_{x,y}$ is the covariance matrix from the posterior distribution of **w**.

With the full distribution for the target value, y, given the predictor value, x, we can, of course, make predictions. The point prediction for y is, of course, the mean of this normal distribution, so $m_{x,y}^T\phi(x)$. But we can do more than just predict the most likely value, we can of course also get confidence values because we know the distribution.

Write a function that predicts the most likely y value for a given x value. Use it to plot the inferred model against (x, y) points. See Figure 14-12.
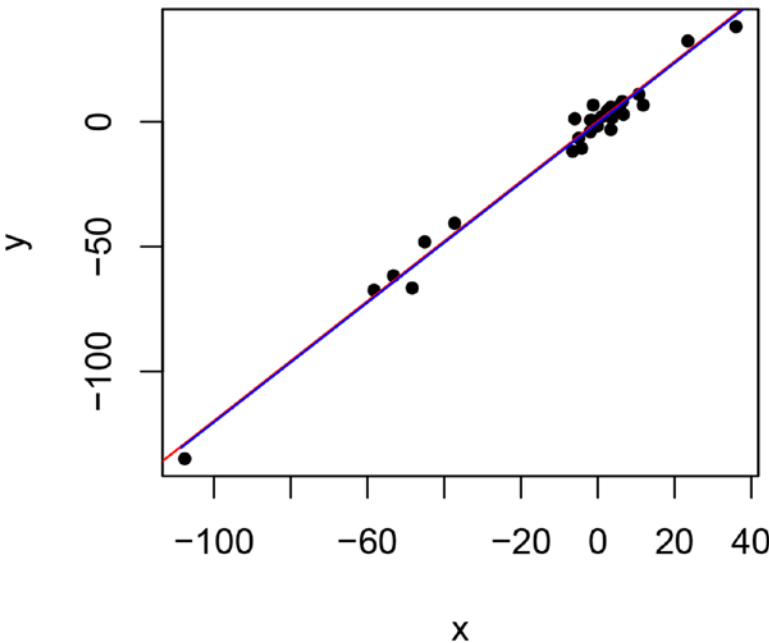
Figure 14-12. *True linear model in red and predicted values in blue*

Use the fact that you also have the predicted distribution for y to write a function that gives you quantiles for this distribution and uses it to plot 95% intervals around the predictions. See Figure 14-13.
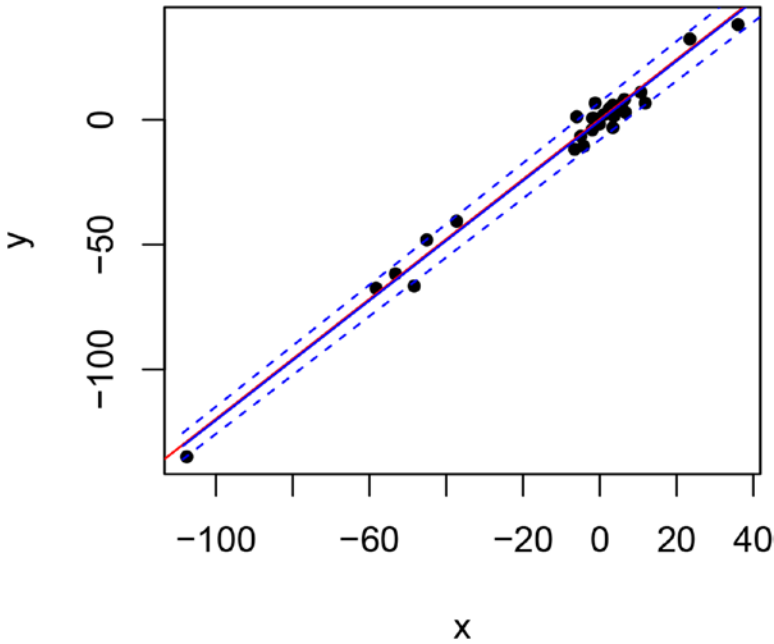
Figure 14-13. *Prediction with 95% support interval*

Just plotting the lines with 95% support intervals doesn't directly show how the uncertainty around a point depends on the uncertainty in the weights of the model if we just plot the line around the points used to train the model. There the support is roughly the same for all the points. We will see a difference, though, if we are far away from the points used for training. There small uncertainties in the weights—the lines through the points—magnifies and spreads the interval. See Figure 14-14.
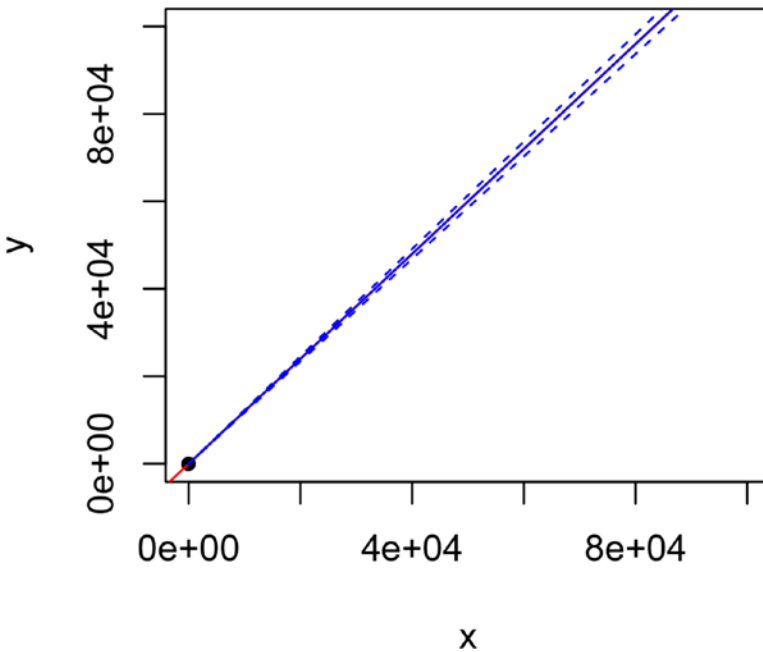
Figure 14-14. *Prediction with 95% support interval, wider range*

Formulas and Their Model Matrix

We continuing working with our Bayesian linear regression and we will generalize the kind of formulas we can fit.

Recall from Chapter 6, that when fitting our models to data we did this using a so-called model matrix (or design matrix) of this form:

$$\Phi_x = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Row i in this matrix contains the vector $(1, x_i) = \phi(x_i)^T$ capturing the predictor variable for the i'th observation, x_i. It actually has two predictor variables, it is just that one is a constant 1. So what it captures is the input we need to predict where the target parameter will be for at a given point for the predictor variable and because we have a model with both an y-axis intercept and an incline it needs two variables. To predict the target, we use the inner product of this row and the weights of our fitted model:

$$y_i = \boldsymbol{w}^T \phi(x_i) + \in = w_0 \cdot 1 + w_1 \cdot x_i + \in .$$

We call $\phi(x)$ the *feature vector* for a point x and the *model matrix* contains a row for each data point we want to fit or make predictions on, such that row i is $\phi(x_i)^T$. With the feature vector on the form we have used here, $\phi(x)^T = (1, x)$, we are fitting a line, but the feature doesn't have to have this form. We can make more complicated feature vectors.

330

If we instead used the feature vector $\phi(x)^T = (1, x, x^2)$ and added another weight to **w** so it now had three dimensions, (w_0, w_1, w_2), we could be predicting the target in the same way, $y = \mathbf{w}^T\phi(x_i) + \varepsilon$, except now of course $\mathbf{w}^T\phi(x_i) = w_0 + w_1 x + w_2 x^2$, so we would be fitting a quadratic polynomial.

If you are thinking now "hey, that is no longer a linear model!" you are wrong. The linearity in the model was never actually related to the linearity in x. It is a linearity in **w** that makes the model linear, and as long as we are getting the predicted value as the inner product of a weight vector like this and some feature vector, it is a linear model we are working with. You can make the feature vector $\phi(x)$ as crazy as you like.

If you construct the model matrix the same way:

$$\Phi_x = \begin{bmatrix} \phi(x_1)^T \\ \phi(x_2)^T \\ \phi(x_3)^T \\ \vdots \\ \phi(x_n)^T \end{bmatrix}$$

the mathematics for fitting weights and predicting new target values works the same, except of course that the weight vector w has the number of dimensions that reflects the dimensions in the feature vectors.

The feature vector doesn't have to be a function of a single variable, x, either. If you want to fit a linear model in two variables—such as a plane—then you can just let the feature vector depend on two variables: $\phi(x, z)^T = (1, x, z)$. The linear combination with the weight vector would be $\mathbf{w}^T\phi(x, z) = w_0 + w_1 x + w_2 z$, which would exactly be a linear model in two variables.

Working with Model Matrices in R

The way we specify both feature vectors and model matrices in R is a *formula*. A Formula is created as an expression containing the tilde symbol, ~, and the target variable should be put to the left and the explanatory variables on the right.

R has quite a rich syntax for specifying formula, and if you are interested, you should read the documentation by writing this in the R shell:

```
?formula
```

For the linear model, we would write y ~ x. The intercept variable is implicitly there; you don't need to tell R that you want the feature vector to include the "-1", instead, you would have to remove it explicitly. You can also specify polynomial feature vectors, but R interprets multiplication, *, as something involving interaction between variables.[1] To specify that you want the second order polynomial of x, you need to write y ~ I(x^2) + x. The function I is the identity function and using it here makes R interpret the x^2 as squaring the number x instead of trying to interpret it as part of the formula specification. If you *only* want to fit the square of x, you would just write y ~ I(x^2). For a general n degree polynomial you can use y ~ poly(x,n, raw=TRUE).

To fit our linear model we need data for two things. In the model we have already implemented we had vectors **x** and **y**, but in the general case the prediction variable **x** should be replaced with the model matrix **Φ**. From **Φ** and **y,** we can fit the model.

---

[1] In formulas, x*z means x + z + x:z whereas x:z is the interaction between x and z—in practice the product of their numbers—so y ~ x*z means $\phi(x, z) = (1, x, z, x Ð z)$).

R has functions for getting both from a formula and data. It isn't *quite* straightforward, though, because of scoping rules. If you write a formula somewhere in your code, you want the variables in the formula to refer to the variables in the scope where you are. Not somewhere else where the code *might* look at the formula. So the formula needs to capture the current scope—similar to how a closure captures the scope around it. On the other hand, you also want to be able to provide data directly to models via data frames. Quite often, the data you want to fit is found as columns in a data frame, not as individual variables in scope. Sometimes it is even a mix.

The function model.frame lets you capture what you need for collecting data relevant for a formula. It will know about the scope of the formula, but you can add data through a data frame as well. Think of it as a data.frame, just with a bit more information about the data that it gets from analyzing the formula.

We can see all of this in action in this small example:

```
predictors <- data.frame(x = rnorm(5), z = rnorm(5))
y <- with(predictors, rnorm(5, mean = 3*x + 5*z + 2))

model <- y ~ x + z

model.frame(model, data = predictors)
##            y          x           z
## 1  0.9532092  0.02035145 -0.11778756
## 2 -2.7934758 -1.06963074  0.03347605
## 3  3.3926486  1.11133916 -0.43176349
## 4 10.8303077  1.68233471  1.27280461
## 5 -2.0380722 -0.32349845 -0.59252024
```

Here we have two predictor variables, x and z, in a data frame, and we simulated the response variable, y, in the global scope. The model we create using the formula y ~ x + z (which means $\phi(x, z)^T = (1, x, z)$) and we construct a model frame from this that contains the data for all the variables used in the formula.

The way the model frame gets created, R first looks in the data frame it gets for a variable, and if it is there it uses that data, if it is not, it uses the data it can find in the scope of the formula. If it cannot find it at all, it will, of course, report an error.

The data frame is also used to construct expressions from variables. In the scope you might have the variable x but not the variable x^2, where the latter is needed for constructing a model matrix. The model.frame function will construct it for you.

```
x <- runif(10)
model.frame(~ x + I(x^2))
##            x         I(x^2)
## 1  0.35846214 0.128495....
## 2  0.76297492 0.582130....
## 3  0.42375496 0.179568....
## 4  0.09579368 0.009176....
## 5  0.48314622 0.233430....
## 6  0.56738521 0.321925....
## 7  0.89860842 0.807497....
## 8  0.51414054 0.264340....
## 9  0.15684623 0.024600....
## 10 0.30248554 0.091497....
```

In this example, we don't have a response variable for the formula; you don't necessarily need one. You need it to be able to extract the vector **y** of course, so we do need one for our linear model fitting, but R doesn't necessarily need one.

Once you have a model frame, you can get the model matrix using the function model.matrix. It needs to know the formula and the model frame (the former to know the feature function ϕ and the latter to know the data we are fitting).

Next, we build two models, one where we fit a line that goes through $y = 0$ and the second where we allow the line to intersect the *y*-axis at an arbitrary point.

Notice how the data frames are the same—the variables used in both models are the same—but the model matrices differ.

```
x <- runif(10)
y <- rnorm(10, mean=x)

model.no.intercept <- y ~ x + 0
(frame.no.intercept <- model.frame(model.no.intercept))
##            y         x
## 1  -0.6475994 0.2170210
## 2   2.3601909 0.7161212
## 3   0.2708529 0.7415493
## 4   1.3094623 0.5982522
## 5   1.6820729 0.7725481
## 6  -0.3574741 0.3912436
## 7   0.5509808 0.4675246
## 8   0.8465933 0.6210651
## 9   1.6873893 0.7360315
## 10  1.3658199 0.8070293
model.matrix(model.no.intercept, frame.no.intercept)
##            x
## 1  0.2170210
## 2  0.7161212
## 3  0.7415493
## 4  0.5982522
## 5  0.7725481
## 6  0.3912436
## 7  0.4675246
## 8  0.6210651
## 9  0.7360315
## 10 0.8070293
## attr(,"assign")
## [1] 1
model.with.intercept <- y ~ x
(frame.with.intercept <- model.frame(model.with.intercept))
##            y         x
## 1  -0.6475994 0.2170210
## 2   2.3601909 0.7161212
## 3   0.2708529 0.7415493
## 4   1.3094623 0.5982522
## 5   1.6820729 0.7725481
## 6  -0.3574741 0.3912436
## 7   0.5509808 0.4675246
```

```
## 8   0.8465933 0.6210651
## 9   1.6873893 0.7360315
## 10  1.3658199 0.8070293
model.matrix(model.with.intercept, frame.with.intercept)
##    (Intercept)         x
## 1            1 0.2170210
## 2            1 0.7161212
## 3            1 0.7415493
## 4            1 0.5982522
## 5            1 0.7725481
## 6            1 0.3912436
## 7            1 0.4675246
## 8            1 0.6210651
## 9            1 0.7360315
## 10           1 0.8070293
## attr(,"assign")
## [1] 0 1
```

The target vector, or response variable, **y**, can be extracted from the data frame as well. You don't need the formula this time because the data frame actually remembers which variable is the response variable. You can get it from the model frame using the function `model.response`:

```
model.response(frame.with.intercept)
##          1          2          3          4
## -0.6475994  2.3601909  0.2708529  1.3094623
##          5          6          7          8
##  1.6820729 -0.3574741  0.5509808  0.8465933
##          9         10
##  1.6873893  1.3658199
```

Exercises

Try the following exercises to become more comfortable with the concepts discussed in this chapter.

Building Model Matrices

Build a function that takes as input a formula and optionally, through the … variable, a data frame and build the model matrix from the formula and optional data.

Fitting General Models

Extend the function you wrote earlier for fitting lines to a function that can fit any formula.

Model Matrices Without Response Variables

Building model matrices this way is all good and well when you have all the variables needed for the model frame, but what happens when you don't have the target value? You need the target value to fit the parameters of your model, of course, but later on, you want to predict targets for new data points where you do *not* know the target, so how do you build the model matrix then?

With some obviously fake data, the situation could look like this:

```
training.data <- data.frame(x = runif(5), y = runif(5))
frame <- model.frame(y ~ x, training.data)
model.matrix(y ~ x, frame)
##   (Intercept)        x
## 1           1 0.6983229
## 2           1 0.2849977
## 3           1 0.1836589
## 4           1 0.2277518
## 5           1 0.2773418
## attr(,"assign")
## [1] 0 1
predict.data <- data.frame(x = runif(5))
frame <- model.frame(y ~ x, predict.data)
## Error in model.frame.default(y ~ x, predict.data): variable lengths differ (found for
'x')
```

Of course, we get a problem when trying to build the frame without knowing the target variable y.

If only there were a way to remove the response variable from the formula! And there is.

The function delete.response does just that. You cannot call it directly on a formula. R first needs to collect some information for this function to work, unlike the other functions you've seen. But you can combine it with the function terms to get a formula without the response variable that you can then use to build a model matrix for data where you don't know the target values.

```
responseless.formula <- delete.response(terms(y ~ x))
frame <- model.frame(responseless.formula, predict.data)
model.matrix(responseless.formula, frame)
##   (Intercept)          x
## 1           1 0.05530272
## 2           1 0.34011728
## 3           1 0.23095021
## 4           1 0.29074418
## 5           1 0.37240380
## attr(,"assign")
## [1] 0 1
```

Exercises

Try the following exercises to become more comfortable with the concepts discussed in this chapter.

Model Matrices for New Data

Write a function that takes as input a formula and a data frame as input that does *not* contain the response variable and build the model matrix for that.

Predicting New Targets

Update the function you wrote earlier for predicting the values for new variables to work on models fitted to general formula. If it doesn't already permit this, you should also extend it so it can take more than one such data point. Make the input for new data points come in the form of a data frame.

Interface to a blm Class

By now we have an implementation of Bayesian linear regression but not necessarily in a form that makes it easy to reuse. Wrapping the data relevant for a fitted model into a class and providing various methods to access it is what makes it easy to reuse a model/class.

Generally, you want to access objects through functions as much as you can. If you know which $fields the class has it is easy to write code that just accesses this, but that makes it hard to change the implementation of the class later. A lot of code that makes assumptions about how objects look like will break. It will also make it hard at some later point to change the model/class in an analysis because different classes generally do not look the same in their internals.

To make it easier for others—and your future self—to use the Bayesian linear regression model, we will make a class for it and provide functions for working with it.

This involves both writing functions specific to your own class and writing polymorphic functions that people, in general, expect a fitted model to implement. It is the latter that will make it possible to replace another fitted model with your blm class.

How you go about designing your class and implementing the functions—and choosing which functions to implement, in general—is up to you. Except, of course, when you implement blm specific versions of already existing polymorphic functions; in that case you need to obey the existing interface.

How you choose to represent objects of your class and which functions you choose to implement for it is generally up to you. There is a general convention in R, though, that you create objects of a given class by calling a function with the same name as the class. So I suggest that you write a constructor called blm.

There isn't really any obvious classes to inherit from, so the class of blm objects should probably only be "blm" and not a vector of classes. If you want to make a class hierarchy in your implementation or implement more than one class to deal with different aspects of your model interface, you should knock yourself out.

Constructor

A *constructor* is what we call a function that creates an object of a given class. In some programming languages, there is a distinction between *creating* and *initializing* an object. This is mostly relevant when you have to worry about memory management, and such and can get quite complicated and it is not something we worry about in R. It is the reason, though, that in Python the constructor is called __init__ and it is actually the initialization it handles. The same is the case for Java, which enforces the rule that the constructor must have the same name as the class, where for R it is just a convention. In Java you have a special syntax for creating new objects: new ClassName(). In Python you have to use the name of the class to create the object—ClassName()—but the syntax looks just like a function call. In R it is only *convention* that says that the class name and the constructor should be the same. The syntax for creating an object looks like a function call because it *is* a function call and nothing special is going on in that function except that it returns an object where we have set the class attribute.

So you should write a function called blm that returns an object where you have set the class attribute to "blm". You can do this with the class<- replacement function or the structure function when you create the object. The object is a list—that is the only way you have of storing complex data, after all—and what you put in it depends on what you need for the functions that will be the interface of your class. You might have to go back and change what data is stored in the object from time to time as you develop the interface for your function. That is okay. Try to use functions to access the internals of the object as much as you can, though, since that tends to minimize how much code you need to rewrite when you change the data stored in the object.

Updating Distributions: An Example Interface

Let's consider a case of something we could have as an interface to Bayesian linear models. This is not something you *have* to implement, but it is a good exercise to try.

The thing we do when we fit models in Bayesian statistics is that we take a prior distribution of our model parameters, $P(\theta)$, and update them to a posterior distribution, $P(\theta \mid D)$, when observing data D. Think of it this way: the prior distribution is what we just know about the parameters. Okay, typically we just make the prior up based on mathematical convenience, but you should think about it as what we know about the parameters from our understanding of how the universe works and what prior experience has taught us. Then when you observe more, you add information about the world that changes the conditional probability of how the parameters look given the observations you have made.

There is nothing really magical about what we call prior and posterior here. Both are just distributions for our model parameters. If the prior is based on previous experience, then it is really a posterior for those experiences. We just haven't modeled it that way.

Let's say we have observed data D_1 and obtained a posterior $P(\theta \mid D_1)$. If we then later observe more data, D_2, we obtain even more information about our parameters and can update the distribution for them to $P(\theta \mid D_1, D_2)$.

We can of course always compute this distribution by taking all the old data and all the new and push it through our fitting code. But if we have chosen the prior distribution carefully with respect to the likelihood of the model—and by carefully I mean that we have a so-called *conjugate* prior—then we can just fit the new data but with a different prior: the old posterior.

What a conjugate prior is, is a prior distribution that is chosen such that both prior and posterior are from the same class of distributions (just with different parameters). In our Bayesian linear model, both prior and posterior are normal distributions so we have a conjugate prior. This means that we can, in principle, update our fitted model with more observations just by using the same fitting code but with a different prior.

I hinted a bit at this in the exercises earlier, but now you can deal with it more formally. You need a way of representing multivariate normal distributions—but you need this anyway to represent your `blm` objects—and a way of getting to a fitted one inside your `blm` objects to extract a posterior.

There are many ways to implement this feature so you have something to experiment with. You can have an `update` function that takes prior and new observations as parameters and outputs the (updated) posterior. Here you need to include the formula as well somehow, to build the model matrix. Or you can let `update` take a fitted object together with new data and get the formula and prior info from the fitted object. Of course, if you do this you need to treat the prior *without* any observations as a special case—and that prior will not know anything about formulas or model matrices.

We can try with an interface like this:

```
update <- function(model, data, prior) { ... }
```

where `model` is the formula, `data` a new dataset and `prior` the prior to use for fitting. This is roughly the interface you have for the constructor, except there you don't necessarily have `data` as an explicit parameter (you want to be able to fit models without data in a data frame, after all), and you don't have `prior` as a parameter at all.

Thinking about it a few seconds and realizing that whatever model fitting we put in here is going to be exactly the same as in `blm`, we can change the interface to get rid of the explicit `data` parameter. If we let that parameter go through `...` instead, we can use exactly the same code as in `blm` (and later remove the code from `blm` by calling `update` there instead).

```
update <- function(model, prior, ...) { ... }
blm <- function(model, ...) {
    # some code here...
    prior <- make_a_prior_distribution_somehow()
    posterior <- update(model, prior, ...)
    # some code that returns an object here...
}
```

To get this version of `blm` to work you need to get the prior in a form you can pass along to `update` but if you did the exercises earlier you should already have a function that does this (although you might want to create a class for these distributions and return them as such so you can manipulate them through an interface if you want to take it a bit further).

Of course, instead of getting rid of the model fitting code in the body of `blm` you could also get rid of `update` and put that functionality in `blm` by letting that function take a prior parameter. If you do that, though, you want to give it a default so you can use the original one if it isn't specified.

```
blm <- function(model, prior = NULL, ...) {
    # some code here...
    if (is.null(prior)) {
        prior <- make_a_prior_distribution_somehow()
    }
    posterior <- update(model, prior, ...)
    # some code that returns an object here...
}
```

Let's stick with having `update` for now, though. How would we use `update` with a fitted model?

```
fit1 <- blm(y ~ x)
fit2 <- update(y ~ x, new_data, fit1)
```

This doesn't work because `fit1` is a `blm` object and not a normal distribution. You need to extract the distribution from the fitted model.

If you have stored the distribution in the object—and you should because otherwise, you cannot use the object for anything since the fit *is* the posterior distribution—you should be able to get at it. What you don't want to do, however, is access the posterior directly from the object as `fit1$posterior` or something like that. It would work, yes, accessing the internals of the object makes it harder to change the representation later. I know I am repeating myself here, but it bears repeating. You don't want to access the internals of an object more than you have to because it makes it harder to change the representation.

Instead, write a function `posterior` that gives you the posterior.

```
posterior <- function(fit) fit$posterior
```

This function has to access the internals—eventually you will have to to get the information, after all—but if this is the only function that does it, and every other function uses this function, then you only need to change this one function if you change the representation of the object.

With that function in hand you can do this:

```
fit2 <- update(y ~ x, new_data, posterior(fit1))
```

You can also write `update` such that it can take both fitted models as well as distributions as its input. Then you just need a way of getting to the prior object (that might be a distribution or might be a fitted model's posterior distribution) that works either way.

One approach is to test the class of the prior parameter directly.

```
update <- function(model, prior, ...) {
    if (class(prior) == "blm") {
        prior <- posterior(prior)
    }
    # fitting code here
}
```

This is a terrible solution, though, for various reasons. First of all, it only works if you either get a prior distribution or an object with class "blm". What if someone, later on, writes a class that extends your blm? Their class attribute might be c("myblm","blm"), which is different from "blm" and so this test will fail, and so will the following code because there you assume that you have a distribution but what you have is an object of a very different class.

To get around that problem, you can use the function inherits. It tests if a given class name is in the class attribute, so it would work if someone gives your update function a class that specializes your blm class.

```
update <- function(model, prior, ...) {
    if (inherits(prior, "blm")) {
        prior <- posterior(prior)
    }
    # fitting code here
}
```

This is a decent solution—and one you will see in a lot of code if you start reading object oriented code—but it still has some drawbacks. It assumes that the only objects that can provide a distribution you can use as a prior is either the way you have implemented priors by default (and you are not testing that above) or an object of class "blm" (or specializations thereof).

You could, of course, make a test for whether the prior, if it isn't a fitted object, is of a class you define for your distributions, which would solve the first problem. But how do you deal with other kinds of objects that might also be able to give you a prior/posterior distribution?

Whenever you write such a class that can provide it, you can also update your update function, but other people cannot provide a distribution for you this way (unless they change your code). Explicitly testing for the type of an object in this way is not a good code design. The solution to fixing it is the same as for accessing object internals: you access stuff through functions.

If we require that any object we give to update as the prior parameter can give us a distribution if we ask for it, we can update the code to be just

```
update <- function(model, prior, ...) {
    prior <- posterior(prior)
    # fitting code here
}
```

This requires that we make a polymorphic function for posterior and possibly that we write a version for distribution objects as well. I will take a shortcut here and make the default implementation the identity function.

```
posterior <- function(x) UseMethod("posterior")
posterior.default <- function(x) x
posterior.blm <- function(x) x$posterior
```

The only annoyance now is that we call it posterior. It is the posterior distribution when we have a fitted object but it isn't really otherwise. Let's change it to distribution:

```
distribution <- function(x) UseMethod("distribution")
distribution.default <- function(x) x
distribution.blm <- function(x) x$posterior
```

Then update update accordingly:

```
update <- function(model, prior, ...) {
    prior <- distribution(prior)
    # fitting code here
}
```

This way it looks even nicer in the update function.

Designing Your blm Class

As you play around with implementing your blm class, think about the interface you are creating, how various functions fit together, and how you think other people will be able to reuse your model. Keep in mind that "future you" is also "other people," so you are helping yourself when you do this.

The update function we developed is an example of what functionality we could put in the class design and how we made it reusable. You should think about other functions for accessing your objects and design them.

One example could be extracting the distribution for a given input point. You implemented a function for predicting the response variable from predictor variables already, and next you will do it in the predict function again, but if you want to gain the full benefits of having a distribution for the response at a given input you want to have the distribution. How would you provide that to users? How could you use this functionality in your own functions?

Play around with it as you develop your class. Whenever you change something, think about whether this could make other functions simpler or if things could be generalized to make your code more reusable.

Model Methods

There are some polymorphic functions that are generally provided by classes that represent fitted models. Not all models implement all of them, but the more you implement, the more existing code can manipulate your new class; another reason for providing interfaces to objects through functions only.

The following sections include a list of functions that I think your blm class should implement. The functions are listed in alphabetical order, but many of them are easier to implement by using one or more of the others. So read through the list before you start programming. If you think that one function can be implemented simpler by calling one of the others, then implement it that way.

In all cases, read the R documentation for the generic function first. You need the documentation to implement the right interface for each function anyway so you might at least read the whole thing. The description in this note is just an overview of what the functions should do.

coefficients

This function should return fitted parameters of the model. It is not entirely straightforward to interpret what that means with our Bayesian models where a fitted model is a distribution and not a single point parameter. We could let the function return the fitted distribution, but the way this function is typically used that would make it useless for existing code to access the fitted parameters for this model as a drop in replacement for the corresponding parameters from a lm model, for example. Instead, it is probably better to return the point estimates of the parameters which would be the mean of the posterior you compute when fitting.

Return the result as a numeric vector with the parameters named. That would fit what you get from lm.

confint

The function `confint` gives you confidence intervals for the fitted parameters. Here we have the same issue as with `coefficients`: we infer an entire distribution and not a parameter (and in any case, our parameters do not have confidence intervals; they have a joint distribution). Nevertheless, we can compute the analogue to confidence intervals from the distribution we have inferred.

If our posterior is distributed as $\mathbf{w} \sim N(\mathbf{m}, \mathbf{S})$, then component i of the weight vector is distributed as $w_i \sim N(m_i, \mathbf{S}_{i,i})$. From this, and the desired fraction of density you want, you can pull out the thresholds that match the quantiles you need.

You take the `level` parameter of the function and get the threshold quantiles by exploiting that a normal distribution is symmetric. So you want the quantiles to be `c(level/2, 1-level/2)`. From that, you can get the thresholds using the function `qnorm`.

deviance

This function just computes the sum of squared distances from the predicted response variables to the observed. This should be easy enough to compute if you could get the squared distances, or even if you only had the distances and had to square them yourself. Perhaps there is a function that gives you that?

fitted

This function should give you the fitted response variables. This is *not* the response variables in the data you fitted the model to, but instead the predictions that the model makes.

plot

This function plots your model. You are pretty free to decide how you want to plot it, but I could imagine that it would be useful to see an x-y plot with a line going through it for the fit. If there are more than one predictor variable, though, I am not sure what would be a good way to visualize the fitted model. There are no explicit rules for what the `plot` function should do, except for plotting something so you can use your imagination.

predict

This function should make predictions based on the fitted model. Its interface is

predict(object, ...)

but the convention is that you give it new data in a variable `newdata`. If you do not provide new data, it instead gives you the predictions on the data used to fit the model.

print

This function is what gets called if you explicitly print an object or if you just write an expression that evaluates to an object of the class in the R terminal. Typically it prints a very short description of the object.

For fitted objects, it customarily prints how the fitting function was called and perhaps what the fitted coefficients were or how good the fit was. You can check out how `lm` objects are printed to see an example.

If you want to print how the fitting function was called you need to get that from when you fit the object in the blm constructor. It is how the constructor was called that is of interest, after all. Inside that function, you can get the way it was called by using the function sys.call.

residuals

This function returns the residuals of the fit. That is the difference between predicted values and observed values for the response variable.

summary

This function is usually used as a longer version of print. It gives you more information about the fitted model.

It does more than this, however. It returns an object with summary information. What that actually means is up to the model implementation so do what you like here.

Building an R Package for blm

We have most of the pieces put together now for our Bayesian linear regression software, and it is the time we collect it in an R package. That is the next step in our project.

You already have an implementation of Bayesian linear regression with a class, blm, and various functions for accessing objects of this type. Now it is time to collect these functions in a package.

Deciding on the Package Interface

When you designed your class functionality and interface, you had to decide on what functionality should be available for objects of your class and how all your functions would fit together to make the code easy to extend and use. There is a similar process of design involved with making a package.

Everything you did for designing the class, of course, is the same for a package but for the package you have to decide on which functions should be exported and which should be kept internal.

Only exported functions can be used by someone else who loads your package so you might be tempted to export everything you can. This, however, is a poor choice. The interface of your package is the exported functions, and if you export too much, you have a huge interface that you need to maintain. If you make changes to the interface of a package, then everyone using your package will have to update his or her code to adapt to the changing interface. You want to keep changes to the package interface at a minimum.

You should figure out which functionality you consider essential parts of the package functionality and what you consider internal helper functions and only export the functions that are part of the package interface.

Organization of Source Files

R doesn't really care how many files you use to have your source code in or how the source code is organized, but you might. At some point in the future, you will need to be able to find relevant functions to fix bugs or extend the functionality of your package.

Decide how you want to organize your source code. Do you want one function per file? Is there instead some logical way of splitting the functionality of your code into categories where you can have a file per category?

Document Your Package Interface Well

At the very least, the functions you export from your package should be documented. Without documentation a user (and that could be you in the future) won't know how a function is supposed to be used.

This documentation is mostly useful for online help—the kind of help you get using ?—so it shouldn't be too long but should give you a good idea of how a function is supposed to be used.

To give an overall description of the entire package and how various functions fit together and how they should be used, you can write documentation for the package as a whole.

Like with package data, there isn't a place for doing this, really, but you can use the same trick as for data. Put the documentation in a source code file in the R/ directory.

Here is my documentation for the admixturegraph package:

```
#' admixturegraph: Visualising and analysing admixture graphs.
#'
#' The package provides functionality to analyse and test admixture graphs
#' against the \eqn{f} statistics described in the paper
#' \href{http://tinyurl.com/o5a4kr4}{Ancient Admixture in Human History},
#' Patternson \emph{et al.}, Genetics, Vol. 192, 1065--1093, 2012.
#'
#' The \eqn{f} statistics -- \eqn{f_2}, \eqn{f_3}, and \eqn{f_4} -- extract
#' information about correlations between gene frequencies in different
#' populations (or single diploid genome samples), which can be informative
#' about patterns of gene flow between these populations in form of admixture
#' events. If a graph is constructed as a hypothesis for the relationship
#' between the populations, equations for the expected values of the \eqn{f}
#' statistics can be extracted, as functions of edge lengths -- representing
#' genetic drift -- and admixture proportions.
#'
#' This package provides functions for extracting these equations and for
#' fitting them against computed \eqn{f} statistics. It does not currently
#' provide functions for computing the \eqn{f} statistics -- for that we refer
#' to  the \href{https://github.com/DReichLab/AdmixTools}{ADMIXTOOLS} software
#' package.
#'
#' @docType package
#' @name admixturegraph
NULL
```

The @docType and @name tags tell Roxygen that I am writing documentation for the entire package.

Adding README and NEWS Files to Your Package

It is customary to also have a README and a NEWS file in your package. The README file describes what your package does and how and can be thought of as a short advertisement for the package, while the NEWS file describes which changes you have made to your package over time.

Many developers prefer to use "markdown" as the format for these files—in which case they are typically named README.md and NEWS.md—and especially if you put your package on GitHub, it is a good idea to have the README.md file since it will be prominently displayed when people go to the package home page on GitHub.

README

What you write in your README file is up to you, but it is customary to have it briefly describe what the package does and maybe give an example or two on how it is used.

If you write it in markdown—in a file called README.md—it will be the home page if you put your package on GitHub.

You might want to write it in R markdown instead to get all the benefits of knitr to go with the file. In that case, you should just name the file README.Rmd and put this in the header:

```
---
output:
  md_document:
    variant: markdown_github
---
```

This tells knitr that it should make a markdown file as output—it will be called README.md.

NEWS

This file should simply contain a list of changes you have made to your package over time. To make it easier for people to see which changes go with which versions of the package, you can split it into sections with each section corresponding to a version.

Testing

In the package, we should now make sure that all of our functions are tested by at least one unit test and that our package can make it through a package test.

GitHub

Sign up to GitHub and create a repository for the project. Move the code there.

Conclusions

Well, this is the end of the book but hopefully not the end of your data science career. I have said all I wanted to say in this book. There are many things I have left out. Text processing for instance. R is not my favorite language for processing text so I don't use it, but it does have functionality for it. It just goes beyond the kind of data we have looked at here. If you want to process text, like genomes or natural languages, you need different tools than the ones I have covered in this book. I have assumed that you are just working on data frames. It made the book easier to write. But it doesn't cover all that data science is about. For more specialized data analysis you will need to look elsewhere. There are many good books, and I might even write about it at some later time. It just wasn't within the scope of this book.

It is the end of this book, but I would like to leave you with some pointers for where to learn more about data science and about R. There are different directions you might want to go in depending on whether you are more interested in analyzing data or more about developing methods. R is a good choice for either. In the long run, you probably will want to do both. The books listed next will get you started in the direction you want to go.

Data Science

- *The art of data science* by Roger Peng and Elizabeth Matsui

This is a general overview of the steps and philosophies underlying data science. It describes the various stages a project goes through—exploratory analysis, fitting models, etc.—and while it doesn't cover any technical details, it is a good overview.

Machine Learning

- *Pattern matching and machine learning* by Christopher Bishop

This is a book I have been using to teach a machine learning class for many years now. It covers a lot of different algorithms for both supervised and unsupervised learning—also types of analysis not covered in this book. It is rather mathematical and focused on methods, but if you are interested in the underlying machine learning, it is a great introduction to that.

Data Analysis

- *Linear models in R* by Julian J. Faraway

- *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models* by Julian J. Faraway

Linear models and generalized linear models are the first things I try. Pretty much always. These great books for seeing how those models are used in R.

- *R graphics* by Paul Murrell

- *ggplot2: Elegant Graphics for Data Analysis* by Hadley Wickham

The first book describes the basic `graphics` package and the `grid` system that underlies `ggplot2`. The second book, obviously, is the go-to book for learning more about `ggplot2`.

R Programming

- *Advanced R* by Hadley Wickham

- *R Packages* by Hadley Wickham

These are great books if you want to learn more about more advanced R programming and package development.

- *Seamless R and C++ Integration with Rcpp* by Dirk Eddelbuettel

If you are interested in integrating C++ and R, then Rcpp is the way to go, and this is an excellent introduction to Rcpp.

The End

This is where I leave you. I hope you have found the book useful, and if you want to leave me any comments and criticism, please do. It will help me improve it for future versions. If there are things you think should be added, let me know, and I will add a chapter or two to cover it. And definitely, let me know if you find any mistakes in the book. I will be particularly grateful if you spot any mistakes in the code included in the book.

Acknowledgements

I would like to thank Asger Hobolth for many useful comments on earlier versions of this manuscript that helped me improve the writing and the presentation of the material.

Index