Welcome to the R Cookbook, 2nd Edition

R is a powerful tool for statistics, graphics, and statistical programming. It is used by tens of thousands of people daily to perform serious statistical analyses. It is a free, open source system whose implementation is the collective accomplishment of many intelligent, hard-working people. There are more than 10,000 available add-on packages, and R is a serious rival to all commercial statistical packages.

But R can be frustrating. It’s not obvious how to accomplish many tasks, even simple ones. The simple tasks are easy once you know how, yet figuring out that “how” can be maddening.

This book is full of how-to recipes, each of which solves a specific problem. Each recipe includes a quick introduction to the solution followed by a discussion that aims to unpack the solution and give you some insight into how it works. We know these recipes are useful and we know they work, because we use them ourselves.

The range of recipes is broad. It starts with basic tasks before moving on to input and output, general statistics, graphics, and linear regression. Any significant work with R will involve most or all of these areas.

If you are a beginner, then this book will get you started faster. If you are an intermediate user, this book will be useful for expanding your horizons and jogging your memory (“How do I do that Kolmogorov–Smirnov test again?”).

The book is not a tutorial on R, although you will learn something by studying the recipes. It is not a reference manual, but it does contain a lot of useful information. It is not a book on programming in R, although many recipes are useful inside R scripts.

Finally, this book is not an introduction to statistics. Many recipes assume that you are familiar with the underlying statistical procedure, if any, and just want to know how it’s done in R.

The Recipes

Most recipes use one or two R functions to solve a specific problem. It’s important to remember that we do not describe the functions in detail; rather, we describe just enough to solve the immediate problem. Nearly every such function has additional capabilities beyond those described here, and some have amazing capabilities. We strongly urge you to read the functions’ help pages. You will likely learn something valuable.

Each recipe presents one way to solve a particular problem. Of course, there are likely several reasonable solutions to each problem. When we knew of multiple solutions, we generally selected the simplest one. For any given task, you can probably discover several alternative solutions yourself. This is a cookbook, not a bible.

In particular, R has literally thousands of downloadable add-on packages, many of which implement alternative algorithms and statistical methods. This book concentrates on the core functionality available through the basic distribution combined with several important packages known collectively as the *tidyverse*.

The most concise definition of the tidyverse comes from Hadley Wickham, its originator and one of its core maintainers:

> *The tidyverse is a set of packages that work in harmony because they share common data representations and API design. The* `tidyverse` *package is designed to make it easy to install and load core packages from the tidyverse in a single command. The best place to learn about all the packages in the tidyverse and how they fit together is R for Data Science.*

A Note on Terminology

The goal of every recipe is to solve a problem and solve it quickly. Rather than laboring in tedious prose, we occasionally streamline the description with terminology that is correct but not precise. A good example is the term *generic function*. We refer to `print(`*`X`*`)` and `plot(`*`X`*`)` as generic functions because they work for many kinds of x, handling each kind appropriately. A computer scientist would wince at our terminology because, strictly speaking, these are not simply “functions”; they are polymorphic methods with dynamic dispatching. But if we carefully unpacked every such technical detail, the essential solutions would be buried in the technicalities. So we just call them functions, which we think is more readable.

Another example, taken from statistics, is the complexity surrounding the semantics of statistical hypothesis testing. Using the strict language of probability theory would obscure the practical application of some tests, so we use more colloquial language when describing each statistical test. See the introduction to Chapter 9 for more about how hypothesis tests are presented in the recipes.

Our goal is to make the power of R available to a wide audience by writing readably, not formally. We hope that experts in their respective fields will understand if our terminology is occasionally informal.

Software and Platform Notes

The base distribution of R has frequent and planned releases, but the language definition and core implementation are stable. The recipes in this book should work with any recent release of the base distribution.

Some recipes have platform-specific considerations, and we have carefully noted them. Those recipes mostly deal with software issues, such as installation and configuration. As far as we know, all other recipes will work on all three major platforms for R: Windows, macOS, and Linux/Unix.

Other Resources

Here are a few suggestions for further reading, if oyu'd like to dig a little deeper:

On the web

The mother ship for all things R is the R project site. From there you can download R for your platform, add-on packages, documentation, and source code as well as many other resources.
Beyond the R project site, we recommend using an R-specific search engine, such as RSeek, created by Sasha Goodman. You can use a generic search engine, such as Google, but the "R" search term brings up too much extraneous stuff. See Recipe 1.11 for more about searching the web.
Reading blogs is a great way to learn about R and stay abreast of leading-edge developments. There are surprisingly many such blogs, so we recommend following two blogs-of-blogs: R-bloggers, created

by Tal Galili, and PlanetR. By subscribing to their RSS feeds, you will be notified of interesting and useful articles from dozens of websites.

R books

There are many, many books about learning and using R. Listed here are a few that we have found useful. Note that the R project site contains an extensive bibliography of books related to R.

R for Data Science, by Hadley Wickham and Garrett Grolemund (O’Reilly), is an excellent introduction to the tidyverse packages, especially for using them in data analysis and statistics. It is also available online.

We find the *R Graphics Cookbook*, 2nd ed., by Winston Chang (O’Reilly), indispensible for creating graphics. The book *ggplot2: Elegant Graphics for Data Analysis* by Hadley Wickham (Springer) is the definitive reference for the graphics package `ggplot2`, which we use in this book. Anyone doing serious graphics work in R will want *R Graphics* by Paul Murrell (Chapman & Hall/CRC).

R in a Nutshell, by Joseph Adler (O’Reilly), is the quick tutorial and reference you’ll keep by your side. It covers many more topics than this cookbook.

New books on programming in R appear regularly. We suggest *Hands On Programming with R* by Garrett Grolemund (O’Reilly) for an introduction, or *The Art of R Programming* by Normal Matloff

(No Starch Press). Hadley Wickham's *Advanced R* (Chapman & Hall/CRC) is available either as a printed book or free online and is a great deeper dive into advanced R topics. *Efficient R Programming*, by Colin Gillespie and Robin Lovelace (O'Reilly), is another good guide to learning the deeper concepts about R programming. *Modern Applied Statistics with S*, 4th ed., by William Venables and Brian Ripley (Springer), uses R to illustrate many advanced statistical techniques. The book's functions and datasets are available in the `MASS` package, which is included in the standard distribution of R.

Serious geeks can download the R Language Definition from the R Core Team. The Definition is a work in progress, but it can answer many of your detailed questions regarding R as a programming language.

Statistics books

For learning statistics, a great choice is *Using R for Introductory Statistics* by John Verzani (Chapman & Hall/CRC). It teaches statistics and R together, giving you the necessary computer skills to apply the statistical methods.

You will need a good statistics textbook or reference book to accurately interpret the statistical tests performed in R. There are many such fine books—far too many for us to recommend any one above the others.

Increasingly, statistics authors are using R to illustrate their methods. If you work in a specialized field, then you will likely find a useful and relevant book in the R project bibliography.

Conventions Used in This Book

The following typographical conventions are used in this book:

Italic

Indicates new terms, URLs, email addresses, filenames, and file extensions.

`Constant width`

Used for program listings, as well as within paragraphs to refer to program elements such as variable or function names, databases, packages, data types, environment variables, statements, and keywords.

`Constant width bold`

Shows commands or other text that should be typed literally by the user.

`Constant width italic`

Shows text that should be replaced with user-supplied values or by values determined by context.

TIP

This element signifies a tip or suggestion.

NOTE

This element signifies a general note.

WARNING

This element indicates a warning or caution.

Using Code Examples

Supplemental material (code examples, source code for the book, exercises, etc.) is available for download at *http://rc2e.com*. The Twitter account for content associated with this book is @R_cookbook.

This book is here to help you get your job done. In general, you may use the code in this book in your programs and documentation. You do not need to contact us for permission unless you're reproducing a significant portion of the code. For example, writing a program that uses several chunks of code from this book does not require permission. Selling or distributing a CD-ROM of examples from O'Reilly books does require permission. Answering a question by citing this book and quoting example code does not require permission. Incorporating a significant amount of example code from this book into your product's documentation does require permission.

We appreciate, but do not require, attribution. An attribution usually includes the title, author, publisher, and ISBN. For example: “*R Cookbook*, 2nd ed., by J.D. Long and Paul Teetor. Copyright 2019 J.D. Long and Paul Teetor, 978-1-492-04068-2.”

If you feel your use of code examples falls outside fair use or the permission just described, feel free to contact us at permissions@oreilly.com.