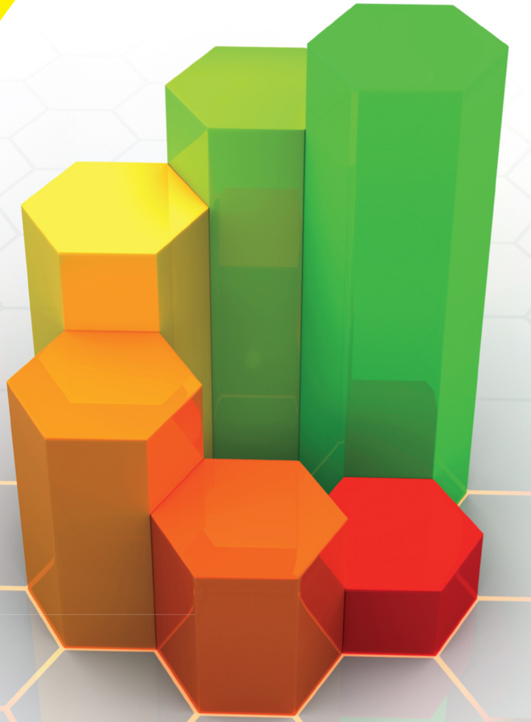
Making Everything Easier!™

R FOR DUMMIES®

Learn to:

- Use R for data analysis and processing
- Write functions and scripts for repeatable analysis
- Create high-quality charts and graphics
- Perform statistical analysis and build models



Andrie de Vries

Joris Meys

R FOR DUMMIES®

by Andrie de Vries and Joris Meys

WILEY
A John Wiley and Sons, Ltd, Publication

R For Dummies®

Published by
John Wiley & Sons, Ltd.
The Atrium
Southern Gate
Chichester
West Sussex
PO19 8SQ
England

Email (for orders and customer service enquires): cs-books@wiley.co.uk
Visit our home page on www.wiley.com

LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY: THE PUBLISHER, THE AUTHOR, AND ANYONE ELSE IN PREPARING THIS WORK MAKE NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS WORK AND SPECIFICALLY DISCLAIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION WARRANTIES OF FITNESS FOR A PARTICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES OR PROMOTIONAL MATERIALS. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR EVERY SITUATION. THIS WORK IS SOLD WITH THE UNDERSTANDING THAT THE PUBLISHER IS NOT ENGAGED IN RENDERING LEGAL, ACCOUNTING, OR OTHER PROFESSIONAL SERVICES. IF PROFESSIONAL ASSISTANCE IS REQUIRED, THE SERVICES OF A COMPETENT PROFESSIONAL PERSON SHOULD BE SOUGHT. NEITHER THE PUBLISHER NOR THE AUTHOR SHALL BE LIABLE FOR DAMAGES ARISING HEREFROM. THE FACT THAT AN ORGANIZATION OR WEBSITE IS REFERRED TO IN THIS WORK AS A CITATION AND/OR A POTENTIAL SOURCE OF FURTHER INFORMATION DOES NOT MEAN THAT THE AUTHOR OR THE PUBLISHER ENDORSES THE INFORMATION THE ORGANIZATION OR WEBSITE MAY PROVIDE OR RECOMMENDATIONS IT MAY MAKE. FURTHER, READERS SHOULD BE AWARE THAT INTERNET WEBSITES LISTED IN THIS WORK MAY HAVE CHANGED OR DISAPPEARED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ.

For general information on our other products and services, please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993, or fax 317-572-4002.

For technical support, please visit www.wiley.com/techsupport.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats. For more information about Wiley products, visit us at www.wiley.com.

British Library Cataloguing in Publication Data: A catalogue record for this book is available from the British Library.

ISBN 978-1-119-96284-7 (paperback); ISBN 978-1-119-96312-7 (ebook); ISBN 978-1-119-96313-4 (ebook); ISBN 978-1-119-96314-1 (ebook)

Printed and bound in the United States by Bind-Rite

10 9 8 7 6 5 4 3 2 1

WILEY

About the Authors

Andrie de Vries: Andrie started to use R in 2009 to analyze survey data, and he has been continually impressed by the ability of the open-source community to innovate and create phenomenal software. Andrie is director of PentaLibra Limited, a boutique market-research firm specializing in surveys and statistical analysis. He has contributed two R packages to CRAN and is developing several packages to make the analysis and reporting of survey data easier. He also is actively involved in the development of LimeSurvey, the open-source survey-management system. To maintain equilibrium in his life, Andrie is working toward a yoga teacher diploma at the Krishnamacharya Healing and Yoga Foundation.

Joris Meys, MSc: Joris is a statistical consultant and R programmer in the Department of Mathematical Modeling, Statistics, and Bioinformatics at Ghent University (Belgium). After earning a master's degree in biology, he worked for six years in environmental research and management before starting an advanced master's degree in statistical data analysis. Joris writes packages for both specific projects and general implementation of methods developed in his department, and he is the maintainer of several packages on R-Forge. He has co-authored a number of scientific papers as a statistical expert. To balance science with culture, Joris spends most of his spare time playing saxophone in a couple of local bands.

Dedication

This book is for my wife, Annemarie, because of her encouragement, support, and patience. And for my 9-year-old niece, Tanya, who is really good at math and kept reminding me that the deadline for this book was approaching!

—Andrie de Vries

For my mother, because she made me the man I am. And for Granny, because she rocks!

—Joris Meys

Authors' Acknowledgments

This book is possible only because of the tremendous support we had from our editorial team at Wiley. In particular, thank you to Elizabeth Kuball for her patient and detailed editing and gentle cajoling, and to Sara Shlaer for pretending not to hear the sound of missed deadlines. Thank you to Kathy Simpson for teaching us how to write in Dummies style, and to Chris Katsaropoulos for getting us started.

Thank you to our technical editor, Gavin Simpson, for his thorough reading and many helpful comments.

Thank you to the R core team for creating R, for maintaining CRAN, and for your dedication to the R community in the form of mailing lists, documentation, and seminars. And thank you to the R community for creating thousands of useful packages, writing blogs, and answering questions.

In this book, we use several packages by Hadley Wickham, whose contribution of ggplot graphics and other helpful packages like plyr continues to be an inspiration.

While writing this book we had very helpful support from a large number of contributors to the R tag at Stack Overflow. Thank you to James (JD) Long, David Winsemius, Ben Bolker, Joshua Ulrich, Kohske Takahashi, Barry Rowlingson, Roman Luštrik, David Purdy, Nick Sabbe, Joran Elias, Brandon Bertelsen, Michael Sumner, Ian Fellows, Dirk Eddelbuettel, Simon Urbanek, Gabor Grotendieck, and everybody else who continue to make Stack Overflow a great resource for the R community.

From Andrie: Thank you to the large number of people who have influenced my choices. Professor Bruce Hardie at London Business School made a throwaway comment during a lecture in 2004 that first made me aware of R; his pragmatic approach to making statistics relevant to marketing decision makers was inspirational. Gary Bennett at Logit Research has been very helpful with his advice and a joy to work with. Paul Hutchings at Kindle Research was very supportive in the early stages of setting up my business.

From Joris: Thank you to the professors and my colleagues at the Department of Mathematical Modeling, Statistics, and Bioinformatics at Ghent University (Belgium) for the insightful discussions and ongoing support I received while writing this book.

Publisher's Acknowledgments

We're proud of this book; please send us your comments at `http://dummies.custhelp.com`. For other comments, please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993, or fax 317-572-4002.

Some of the people who helped bring this book to market include the following:

Acquisitions, Editorial

Project Editor: Elizabeth Kuball

Executive Commissioning Editor: Birgit Gruber

Assistant Editor: Ellie Scott

Copy Editor: Elizabeth Kuball

Technical Editor: Gavin Simpson

Editorial Manager: Jodi Jensen

Senior Project Editor: Sara Shlaer

Editorial Assistant: Leslie Saxman

Special Editorial Help: Kathy Simpson

Cover Photo: ©iStockphoto.com/Jacobo Cortés

Cartoons: Rich Tennant (`www.the5thwave.com`)

Marketing

Associate Marketing Director: Louise Breinholt

Senior Marketing Executive: Kate Parrett

Composition Services

Sr. Senior Project Coordinator: Kristie Rees

Layout and Graphics: Sennett Vaughan Johnson, Lavonne Roberts

Proofreaders: Lindsay Amones, Susan Hobbs

Indexer: Sharon Shock

UK Tech Publishing

Michelle Leete, VP Consumer and Technology Publishing Director

Martin Tribe, Associate Director–Book Content Management

Chris Webb, Associate Publisher

Publishing and Editorial for Technology Dummies

Richard Swadley, Vice President and Executive Group Publisher

Andy Cummings, Vice President and Publisher

Mary Bednarek, Executive Acquisitions Director

Mary C. Corder, Editorial Director

Publishing for Consumer Dummies

Kathleen Nebenhaus, Vice President and Executive Publisher

Composition Services

Debbie Stailey, Director of Composition Services

Contents at a Glance

Table of Contents

Chapter 13: Manipulating and Processing Data 219

Chapter 14: Summarizing Data 253

Introduction

Welcome to *R For Dummies,* the book that takes the steepness out of the learning curve for using R.

We can't guarantee that you'll be a guru if you read this book, but you should be able to do the following:

- Perform data analysis by using a variety of powerful tools
- Use the power of R to do statistical analysis and other data-processing tasks
- Appreciate the beauty of using vector-based operations rather than loops to do speedy calculations
- Appreciate the meaning of the following line of code:

  ```
  knowledge <- apply(theory, 1, sum)
  ```
- Know how to find, download, and use code that has been contributed to R by its very active community of developers
- Know where to find extra help and resources to take your R coding skills to the next level
- Create beautiful graphs and visualizations of your data

About This Book

R For Dummies is an introduction to the statistical programming language known as R. We start by introducing the interface and work our way from the very basic concepts of the language through more sophisticated data manipulation and analysis.

We illustrate every step with easy-to-follow examples. This book contains numerous code snippets, several write-it-yourself functions you can use later on, and complete analysis scripts. All these are for you to try out yourself.

We don't attempt to give a technical description of how R is programmed internally, but we do focus as much on the why as on the how. R doesn't function as your average scripting language, and it has plenty of unique features that may seem surprising at first. Instead of just telling you how you have to talk to R, we believe it's important for us to explain how the R engine reads what you tell it to do. After reading this book, you should be able to manipulate your data in the form you want and understand how to use functions we *didn't* cover in the book (as well as the ones we do cover).

R and RStudio

R For Dummies can be used with any operating system that R runs on. Whether you use Mac, Linux, or Windows, this book will get you on your way with R.

R is more a programming language than an application. When you download R, you automatically download a console application that's suitable for your operating system. However, this application has only basic functionality, and it differs to some extent from one operating system to the next.

RStudio is a cross-platform application with some very neat features to support R. In this book, we don't assume you use any specific console application. But because RStudio provides a common user interface across the major operating systems, we think that you'll understand how to run it quite quickly. For this reason, we use RStudio to demonstrate some of the concepts rather than operating-system-specific editor.

This book is a reference. You don't have to read it from beginning to end. Instead, you can use the table of contents and index to find the information you need. In each chapter, we cross-reference other chapters where you can find more information.

Conventions Used in This Book

Code snippets appear like this example, where we simulate 1 million throws of two six-sided dice:

```
> set.seed(42)
> throws <- 1e6
> dice <- sapply(1:2,
+     function(x)sample(1:6, throws, replace=TRUE)
+ )
> table(rowSums(dice))

     2      3      4      5      6      7      8 
 28007  55443  83382 110359 138801 167130 138808 
     9     10     11     12 
110920  83389  55816  27945 
```

Each line of R code in this example is preceded by one of two symbols:

- **>:** The prompt symbol, >, is not part of your code, and you should not type this when you try the code yourself.
- **+:** The continuation symbol, +, indicates that this line of code still belongs to the previous line of code. In fact, you don't have to break a line of code into two, but we do this frequently, because it improves the readability of code and helps it fit into the pages of a book.

The lines that don't start with either the prompt symbol or the continuation symbol are the output produced by R. In this case, you get the total number of throws where the dice added up to the numbers 2 through 12. For example, out of 1 million throws of the dice, on 28,007 occasions, the numbers on the dice added to 2.

You can copy these code snippets and run them in R, but you have to type them exactly as shown. There are only three exceptions:

- Don't type the prompt symbol, >.
- Don't type the continuation symbol, +.
- Where you put spaces or tabs isn't critical, as long as it isn't in the middle of a keyword. Pay attention to new lines, though.

You get to write some R code in every chapter of the book. Because much of your interaction with the R interpreter is most likely going to be in interactive mode, you need a way of distinguishing between your code and the results of your code. When there is an instruction to type some text into the R console, you'll see a little > symbol to the left of the text, like this:

```
> print("Hello world!")
```

If you type this into a console and press Enter, R responds with the following:

```
[1] "Hello world!"
```

For convenience, we collapse these two events into a single block, like this:

```
> print("Hello world!")
[1] "Hello world!"
```

This indicates that you type some code (`print("Hello world!")`) into the console and R responds with `[1] "Hello world!"`.

Finally, many R words are directly derived from English words. To avoid confusion in the text of this book, R functions, arguments, and keywords appear in `monofont`. For example, to create a plot, you use the `plot()` function in R. When talking about functions, the function name will always be followed by open and closed parentheses — for example, `plot()`. We refrain from adding arguments to the function names mentioned in the text, unless it's really important.

On some occasions we talk about menu commands, such as File⇨Save. This just means that you open the File menu and choose the Save option.

What You're Not to Read

You can use this book however works best for you, but if you're pressed for time (or just not interested in the nitty-gritty details), you can safely skip anything marked with a Technical Stuff icon. You also can skip sidebars (text in gray boxes); they contain interesting information, but nothing critical to your understanding of the subject at hand.

Foolish Assumptions

This book makes the following assumptions about you and your computer:

- **You know your way around a computer.** You know how to download and install software. You know how to find information on the Internet and you have Internet access.
- **You're not necessarily a programmer.** If you are a programmer, and you're used to coding in other languages, you may want to read the notes marked by the Technical Stuff icon — there, we fill you in on how R is similar to, or different from, other common languages.
- **You're not a statistician, but you understand the very basics of statistics.** *R For Dummies* isn't a statistics book, although we do show you how to do some basic statistics using R. If you want to understand the statistical stuff in more depth, we recommend *Statistics For Dummies,* 2nd Edition, by Deborah J. Rumsey, PhD (Wiley).
- **You want to explore new stuff.** You like to solve problems and aren't afraid of trying things out in the R console.

How This Book Is Organized

The book is organized in six parts. Here's what each of the six parts covers.

Part I: R You Ready?

In this part, we introduce you to R and show you how to write your first script. You get to use the very powerful concept of vectors to make simultaneous calculations on many variables at once. You get to work with the R workspace (in other words, how to create, modify, or remove variables). You find out how save your work and retrieve and modify script files that you wrote in previous sessions. We also introduce some fundamentals of R (for example, how to extend functionality by installing packages).

Part II: Getting Down to Work in R

In this part, we fill you in on the three R's: reading, 'riting, and 'rithmetic — in other words, working with text and number (and dates for good measure).

You also get to use the very important data structures of lists and data frames.

Part III: Coding in R

R is a programming language, so you need to know how to write and understand functions. In this part, we show you how to do this, as well as how to control the logic flow of your scripts by making choices using `if` statements, as well as looping through your code to perform repetitive actions. We explain how to make sense of and deal with warnings and errors that you may experience in your code. Finally, we show you some tools to debug any issues that you may experience.

Part IV: Making the Data Talk

In this part, we introduce the different data structures that you can use in R, such as lists and data frames. You find out how to get your data in and out of R (for example, by reading data from files or the Clipboard). You also see how to interact with other applications, such as Microsoft Excel.

Then you discover how easy it is to do some advanced data reshaping and manipulation in R. We show you how to select a subset of your data and how to sort and order it. We explain how to merge different datasets based on columns they may have in common. Finally, we show you a very powerful generic strategy of splitting and combining data and applying functions over subsets of your data. When you understand this strategy, you can use it over and over again to do sophisticated data analyses in only a few small steps.

We're just itching to show you how to do some statistical analysis using R. This is the heritage of R, after all. But we promise to keep it simple. After reading this part, you'll know how to describe and summarize your variables and data using R. You'll be able to do some classical tests (for example, calculating a t-test). And you'll know how to use random numbers to simulate some distributions.

Finally, we show you some of the basics of using linear models (for example, linear regression and analysis of variance). We also show you how to use R to predict the values of new data using some models that you've fitted to your data.

Part V: Working with Graphics

They say that a picture is worth a thousand words. This is certainly the case when you want to share your results with other people. In this part, you discover how to create basic and more sophisticated plots to visualize your data. We move on from bar charts and line charts, and show you how to present cuts of your data using facets.

Part VI: The Part of Tens

In this part, we show you how to do ten things in R that you probably use Microsoft Excel for at the moment (for example, how to do the equivalent of pivot tables and lookup tables). We also give you ten tips for working with packages that are not part of base R.

Icons Used in This Book

As you read this book, you'll find little pictures in the margins. These pictures, or *icons,* mark certain types of text:

TIP

When you see the Tip icon, you can be sure to find a way to do something more easily or quickly.

REMEMBER

You don't have to memorize this book, but the Remember icon points out some useful things that you really should remember. Usually this indicates a design pattern or idiom that you'll encounter in more than one chapter.

WARNING!

When you see the Warning icon, listen up. It points out something you definitely don't want to do. Although it's really unlikely that using R will cause something disastrous to happen, we use the Warning icon to alert you if something is bound to lead to confusion.

TECHNICAL STUFF

The Technical Stuff icon indicates technical information you can merrily skip over. We do our best to make this information as interesting and relevant as possible, but if you're short on time or you just want the information you absolutely *need* to know, you can move on by.

Where to Go from Here

There's only one way to learn R: Use it! In this book, we try to make you familiar with the usage of R, but you'll have to sit down at your PC and start playing around with it yourself. Crack the book open so the pages don't flip by themselves, and start hitting the keyboard!

Part I

R You Ready?

The 5^{th} Wave By Rich Tennant

"Okay, ma'am, I'm going to ask you to walk a straight line; then I'm going to ask you to bisect that line with a perpendicular line that slopes to the equation $y = 3x + 5$."

In this part . . .

From financial headquarters to the dark cellars of small universities, people use R for data manipulation and statistical analysis. With R, you can extract stock prices and predict profits, discover beginning diseases in small blood samples, analyze the behavior of customers, or describe how the gray wolf recolonized European forests.

In this part, you discover the power hidden behind the 18th letter of the alphabet.

Chapter 1

Introducing R: The Big Picture

In This Chapter

- Discovering the benefits of R
- Identifying some programming concepts that make R special

With an estimated worldwide user base of more than 2 million people, the R language has rapidly grown and extended since its origin as an academic demonstration language in the 1990s.

Some people would argue — and we think they're right — that R is much more than a statistical programming language. It's also:

- A very powerful tool for all kinds of data processing and manipulation
- A community of programmers, users, academics, and practitioners
- A tool that makes all kinds of publication-quality graphics and data visualizations
- A collection of freely distributed add-on packages
- A toolbox with tremendous versatility

In this chapter, we fill you in on the benefits of R, as well as its unique features and quirks.

TIP

You can download R at `www.r-project.org`. This website also provides more information on R and links to the online manuals, mailing lists, conferences and publications.

Tracing the history of R

Ross Ihaka and Robert Gentleman developed R as a free software environment for their teaching classes when they were colleagues at the University of Auckland in New Zealand. Because they were both familiar with S, a commercial programming language for statistics, it seemed natural to use similar syntax in their own work. After Ihaka and Gentleman announced their software on the S-news mailing list, several people became interested and started to collaborate with them, notably Martin Mächler.

Currently, a group of 18 people has rights to modify the central archive of source code. This group is referred to as the R Development Core Team. In addition, many other people have contributed new code and bug fixes to the project.

Here are some milestone dates in the development of R:

- **Early 1990s:** The development of R began.
- **August 1993:** The software was announced on the S-news mailing list. Since then, a set of active R mailing lists has been created. The web page at `www.r-project.org/mail.html` provides descriptions of these lists and instructions for subscribing. (For more information, turn to "It provides an engaged community," later in this chapter.)
- **June 1995:** After some persuasive arguments by Martin Mächler (among others) to make the code available as "free software," the code was made available under the Free Software Foundation's GNU General Public License (GPL), Version 2.
- **Mid-1997:** The initial R Development Core Team was formed (although, at the time, it was simply known as the core group).
- **February 2000:** The first version of R, version 1.0.0, was released.

Ross Ihaka wrote a comprehensive overview of the development of R. The web page `http://cran.r-project.org/doc/html/interface98-paper/paper.html` provides a fascinating history.

Recognizing the Benefits of Using R

Of the many attractive benefits of R, a few stand out: It's actively maintained, it has good connectivity to various types of data and other systems, and it's versatile enough to solve problems in many domains. Possibly best of all, it's available for free, in more than one sense of the word.

It comes as free, open-source code

R is available under an open-source license, which means that anyone can download and modify the code. This freedom is often referred to as "free as in speech." R is also available free of charge — a second kind of freedom, sometimes referred to as "free as in beer." In practical terms, this means that you can download and use R free of charge.