

# R IN A NUTSHELL

A Desktop Quick Reference

O'REILLY®

Joseph Adler

# R IN A NUTSHELL

If you're considering R for statistical computing and cata visualization, this book provides a quick and practical guide to just about everything you can do with the open source R language and software environment. You'll learn how to write R functions and use R packages to help you prepare, visualize, and analyze data. Author Joseph Adler illustrates each process with a wealth of examples from medicine, business, and spons.

Updated for R 2.14 and 2.15, this second edition includes new and expanded chapters on R performance, the ggplo.2 data visualization package, and parallel R computing with Hadoop.

- Get started quickly with an R tutorial and hundreds of examples
- Explore R syntax, objects, and other language details
- Find thousands of user-contributed R packages online, including Bioconductor
- Learn how to use R to prepare data for analysis
- Visualize your data with R's graphics, lattice, and gyplot2 packages
- Use R to calculate statistical fests, fit models, and compute probability distributions
- Speed up intensive computations by writing parallel R programs for Hadoop
- Get a complete desktop reference to R

" R is free and powerful, but it's difficult to get started. R in a Nutshell is a critical compassion for learning R and is a great reference manual for every data scientist."

> —DJ Patil Data Scientist in Residence at Greylock

" R is fast becoming the lingua franca of statistics, and R in a Nutshell is the best entry point for people getting started. From data visualization to time series analysis, the book covers all the topics needed to do powerful data science."

> —Anthony Goldbloom Founder and CEO of Kaggle

Joseph Adler, a senior data scientist at Linkedin, has years of experience in data mining and data analysis at companies including Double Click, American Express, and VeriSign. He's the inventor of several patents for computer security and cryptography, and the author of O'Reilly's Baseboll Hacks.

US \$49.99 CAN \$52.99 1SBN: 978-1-449-31208-4







Twitter: @oreillymedia facebook.com/oreilly



# R

# **IN A NUTSHELL**

**Second Edition** 

Joseph Adler



Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

#### R in a Nutshell, Second Edition

by Joseph Adler

Copyright © 2012 Joseph Adler. All rights reserved. Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (<a href="http://my.safaribooksonline.com">http://my.safaribooksonline.com</a>). For more information, contact our corporate/institutional sales department: 800-998-9938 or <a href="mailto:corporate@oreilly.com">corporate@oreilly.com</a>.

**Editors:** Mike Loukides and Meghan Blanchette

**Production Editor:** Holly Bauer **Proofreader:** Julie Van Keuren

**Indexer:** Fred Brown

**Cover Designer:** Karen Montgomery **Interior Designer:** David Futato **Illustrators:** Robert Romano and Re-

becca Demarest

September 2009: First Edition.
October 2012: Second Edition.

#### **Revision History for the Second Edition:**

2012-09-25 First release

See http://oreilly.com/catalog/errata.csp?isbn=9781449312084 for release details.

Nutshell Handbook, the Nutshell Handbook logo, and the O'Reilly logo are registered trademarks of O'Reilly Media, Inc. *R in a Nutshell*, the image of a harpy eagle, and related trade dress are trademarks of O'Reilly Media, Inc.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and O'Reilly Media, Inc., was aware of a trademark claim, the designations have been printed in caps or initial caps.

While every precaution has been taken in the preparation of this book, the publisher and author assume no responsibility for errors or omissions, or for damages resulting from the use of the information contained herein.

ISBN: 978-1-449-31208-4

[LSI]

1348585490

# **Table of Contents**

refa	reface xiii		
Part	I. R Basics		
1.	Getting and Installing R	3	
	R Versions	3	
	Getting and Installing Interactive R Binaries	3	
	Windows	4	
	Mac OS X	5	
	Linux and Unix Systems	5	
2.	The R User Interface	7	
	The R Graphical User Interface	7	
	Windows	8	
	Mac OS X	8	
	Linux and Unix	9	
	The R Console	11	
	Command-Line Editing	13	
	Batch Mode	13	
	Using R Inside Microsoft Excel	14	
	RStudio	15	
	Other Ways to Run R	17	
3.	A Short R Tutorial	19	
	Basic Operations in R	19	
	Functions	21	
	Variables	22	

	Introduction to Data Structures	2
	Objects and Classes	2
	Models and Formulas	2
	Charts and Graphics	30
	Getting Help	3.
4.	R Packages	3
	An Overview of Packages	3
	Listing Packages in Local Libraries	3
	Loading Packages	40
	Loading Packages on Windows and Linux	40
	Loading Packages on Mac OS X	40
	Exploring Package Repositories	4
	Exploring R Package Repositories on the Web	4.
	Finding and Installing Packages Inside R	4.
	Installing Packages From Other Repositories	4.
	Custom Packages	4.
	Creating a Package Directory	4.
	Building the Package	4
5.	An Overview of the R Language	. 5
5.	Expressions Objects Symbols Functions Objects Are Copied in Assignment Statements	5; 5; 5; 5;
5.	Expressions Objects Symbols Functions Objects Are Copied in Assignment Statements Everything in R Is an Object	5; 5; 5; 5; 5; 5;
5.	Expressions Objects Symbols Functions Objects Are Copied in Assignment Statements Everything in R Is an Object Special Values	5; 5; 5; 5; 5; 5;
5.	Expressions Objects Symbols Functions Objects Are Copied in Assignment Statements Everything in R Is an Object Special Values NA	5 52 52 52 53 55 55 55 55
5.	Expressions Objects Symbols Functions Objects Are Copied in Assignment Statements Everything in R Is an Object Special Values NA Inf and -Inf	5 5; 5; 5; 5; 5; 5; 5;
5.	Expressions Objects Symbols Functions Objects Are Copied in Assignment Statements Everything in R Is an Object Special Values NA Inf and -Inf NaN	5; 5; 5; 5; 5; 5; 5; 5; 5;
5.	Expressions Objects Symbols Functions Objects Are Copied in Assignment Statements Everything in R Is an Object Special Values NA Inf and -Inf NaN NULL	5; 5; 5; 5; 5; 5; 5; 5; 5;
5.	Expressions Objects Symbols Functions Objects Are Copied in Assignment Statements Everything in R Is an Object Special Values NA Inf and -Inf NaN NULL Coercion	55.55.55.55.55.55.55.55.55.55.55.55.55.
5.	Expressions Objects Symbols Functions Objects Are Copied in Assignment Statements Everything in R Is an Object Special Values NA Inf and -Inf NaN NULL Coercion The R Interpreter	55 57 57 57 57 57 57 57 57 57 57 57 57 5
5.	Expressions Objects Symbols Functions Objects Are Copied in Assignment Statements Everything in R Is an Object Special Values NA Inf and -Inf NaN NULL Coercion	55.55.55.55.55.55.55.55.55.55.55.55.55.
<ol><li>6.</li></ol>	Expressions Objects Symbols Functions Objects Are Copied in Assignment Statements Everything in R Is an Object Special Values NA Inf and -Inf NaN NULL Coercion The R Interpreter	55 57 57 57 57 57 57 57 57 57 57 57 57 5
	Expressions Objects Symbols Functions Objects Are Copied in Assignment Statements Everything in R Is an Object Special Values NA Inf and -Inf NaN NULL Coercion The R Interpreter Seeing How R Works  R Syntax Constants	5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
	Expressions Objects Symbols Functions Objects Are Copied in Assignment Statements Everything in R Is an Object Special Values NA Inf and -Inf NaN NULL Coercion The R Interpreter Seeing How R Works  R Syntax Constants Numeric Vectors	5 5 5 5 5 5 5 5 5 5 5 5 5 6 6 6 6 6 6 6
	Expressions Objects Symbols Functions Objects Are Copied in Assignment Statements Everything in R Is an Object Special Values NA Inf and -Inf NaN NULL Coercion The R Interpreter Seeing How R Works  R Syntax Constants Numeric Vectors Character Vectors	5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 6
	Expressions Objects Symbols Functions Objects Are Copied in Assignment Statements Everything in R Is an Object Special Values NA Inf and -Inf NaN NULL Coercion The R Interpreter Seeing How R Works  R Syntax Constants Numeric Vectors Character Vectors Symbols	5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
	Expressions Objects Symbols Functions Objects Are Copied in Assignment Statements Everything in R Is an Object Special Values NA Inf and -Inf NaN NULL Coercion The R Interpreter Seeing How R Works  R Syntax Constants Numeric Vectors Character Vectors Symbols Operators	5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 6
	Expressions Objects Symbols Functions Objects Are Copied in Assignment Statements Everything in R Is an Object Special Values NA Inf and -Inf NaN NULL Coercion The R Interpreter Seeing How R Works  R Syntax Constants Numeric Vectors Character Vectors Symbols	5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5

	Assignments	69
	Expressions	69
	Separating Expressions	69
	Parentheses	70
	Curly Braces	70
	Control Structures	71
	Conditional Statements	71
	Loops	72
	Accessing Data Structures	75
	Data Structure Operators	75
	Indexing by Integer Vector	76
	Indexing by Logical Vector	78
	Indexing by Name	79
	R Code Style Standards	80
	2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2	
7.	R Objects	83
•	Primitive Object Types	83
	Vectors	86
	Lists	87
	Other Objects	88
	Matrices	88
		89
	Arrays Factors	89
	Data Frames	91
	Formulas	92
	Time Series	94
	Shingles	95
	Dates and Times	95
	Connections	96
	Attributes	96
	Class	99
0	Complete and Fusing an enter	101
8.	Symbols and Environments	101
	Symbols	101
	Working with Environments	102
	The Global Environment	103
	Environments and Functions	104
	Working with the Call Stack	104
	Evaluating Functions in Different Environments	105
	Adding Objects to an Environment	107
	Exceptions	108
	Signaling Errors	108
	Catching Errors	109
9.	Functions	111
	The Function Keyword	111

	Arguments	11
	Return Values	113
	Functions as Arguments	113
	Anonymous Functions	114
	Properties of Functions	11.
	Argument Order and Named Arguments	117
	Side Effects	118
	Changes to Other Environments	118
	Input/Output	119
	Graphics	119
10.	Object-Oriented Programming	12 <sup>-</sup>
	Overview of Object-Oriented Programming in R	122
	Key Ideas	122
	Implementation Example	123
	Object-Oriented Programming in R: S4 Classes	129
	Defining Classes	129
	New Objects	130
	Accessing Slots	130
	Working with Objects	13
	Creating Coercion Methods	13
	Methods	132
	Managing Methods	133
	Basic Classes	134
	More Help	133
	Old-School OOP in R: S3	133
	S3 Classes	133
	S3 Methods	130
	Using S3 Classes in S4 Classes	13
	Finding Hidden S3 Methods	13
Part	III. Working with Data	
11.	Saving, Loading, and Editing Data	14
	Entering Data Within R	14
	Entering Data Using R Commands	14
	Using the Edit GUI	142
	Saving and Loading R Objects	143
	Saving Objects with save	143
	Importing Data from External Files	140
	Text Files	140
	Other Software	154
	Exporting Data	153
	Importing Data From Databases	150
	Export Then Import	15

	Database Connection Packages RODBC	156 157
	DBI	167
	TSDBI	172
	Getting Data from Hadoop	172
12.	Preparing Data	173
12.	Combining Data Sets	173
	Pasting Together Data Structures	173
	Merging Data by Common Fields	177
	Transformations	177
	Reassigning Variables	179
	The Transform Function	179
	Applying a Function to Each Element of an Object	180
	Binning Data	185
	Shingles	185
	Cut	186
	Combining Objects with a Grouping Variable	187
	Subsets	187
	Bracket Notation subset Function	188
		188
	Random Sampling	189
	Summarizing Functions	190
	tapply, aggregate	190
	Aggregating Tables with rowsum	193
	Counting Values	194
	Reshaping Data	196
	Data Cleaning	205
	Finding and Removing Duplicates	205
	Sorting	206
Part	IV. Data Visualization	
13.	Graphics	213
15.	An Overview of R Graphics	213
	Scatter Plots	214
	Plotting Time Series	220
	Bar Charts	222
	Pie Charts	226
	Plotting Categorical Data	227
	Three-Dimensional Data	232
	Plotting Distributions	232
	Box Plots	242
	Graphics Devices	242
	Customizing Charts	240
	Customizing Charts	Z <del>4</del> /

	Common Arguments to Chart Functions	247
	Graphical Parameters	247
	Basic Graphics Functions	257
14.	Lattice Graphics	. 267
	History	267
	An Overview of the Lattice Package	268
	How Lattice Works	268
	A Simple Example	268
	Using Lattice Functions	270
	Custom Panel Functions	272
	High-Level Lattice Plotting Functions	272
	Univariate Trellis Plots	273
	Bivariate Trellis Plots	297
	Trivariate Plots	305
	Other Plots	310
	Customizing Lattice Graphics	312
	Common Arguments to Lattice Functions	312
	trellis.skeleton	313
	Controlling How Axes Are Drawn	314
	Parameters	315
	plot.trellis	319
	strip.default	320
	simpleKey	321
	Low-Level Functions	322
	Low-Level Graphics Functions	322
	Panel Functions	323
15.	ggplot2	325
	A Short Introduction	325
	The Grammar of Graphics	328
	A More Complex Example: Medicare Data	333
	Quick Plot	342
	Creating Graphics with ggplot2	343
	Learning More	347
Part	V. Statistics with R	
16.	Analyzing Data	351
	Summary Statistics	351
	Correlation and Covariance	354
	Principal Components Analysis	357
	Factor Analysis	360
	Bootstrap Resampling	361

17.	Probability Distributions	363
	Normal Distribution	363
	Common Distribution-Type Arguments	366
	Distribution Function Families	366
18.	Statistical Tests	. 371
	Continuous Data	371
	Normal Distribution-Based Tests	372
	Non-Parametric Tests	385
	Discrete Data	388
	Proportion Tests	388
	Binomial Tests	389
	Tabular Data Tests	390
	Non-Parametric Tabular Data Tests	396
19.	Power Tests	. 397
	Experimental Design Example	397
	t-Test Design	398
	Proportion Test Design	398
	ANOVA Test Design	400
20.	Regression Models	401
	Example: A Simple Linear Model	401
	Fitting a Model	403
	Helper Functions for Specifying the Model	404
	Getting Information About a Model	404
	Refining the Model	410
	Details About the lm Function	410
	Assumptions of Least Squares Regression	412
	Robust and Resistant Regression	414
	Subset Selection and Shrinkage Methods	416
	Stepwise Variable Selection	416
	Ridge Regression	417
	Lasso and Least Angle Regression	418
	elasticnet	419
	Principal Components Regression and Partial Least Squares Regression	420
	Nonlinear Models	420
	Generalized Linear Models	421
	glmnet	424
	Nonlinear Least Squares	427
	Survival Models	428
	Smoothing	433
	Splines	433
	Fitting Polynomial Surfaces	435
	O ,	

	Kernel Smoothing	436
	Machine Learning Algorithms for Regression	437
	Regression Tree Models	439
	MARS	450
	Neural Networks	455
	Project Pursuit Regression	459
	Generalized Additive Models	462
	Support Vector Machines	464
21.	Classification Models	467
	Linear Classification Models	467
	Logistic Regression	467
	Linear Discriminant Analysis	472
	Log-Linear Models	476
	Machine Learning Algorithms for Classification	477
	k Nearest Neighbors	477
	Classification Tree Models	478
	Neural Networks	482
	SVMs	483
	Random Forests	483
22.	Machine Learning	485
	Market Basket Analysis	485
	Clustering	490
	Distance Measures	490
	Clustering Algorithms	491
23.	Time Series Analysis	495
	Autocorrelation Functions	495
	Time Series Models	496
Part	VI. Additional Topics	
24.	Optimizing R Programs	503
	Measuring R Program Performance	503
	Timing	503
	Profiling	504
	Monitor How Much Memory You Are Using	505
	Profiling Memory Usage	506
	Optimizing Your R Code	507
	Using Vector Operations	507
	Lookup Performance in R	509
	Use a Database to Query Large Data Sets	516
	Preallocate Memory	516
	,	

	Cleaning Up Memory	516
	Functions for Big Data Sets	517
	Other Ways to Speed Up R	518
	The R Byte Code Compiler	518
	High-Performance R Binaries	520
25.	Bioconductor	525
	An Example	525
	Loading Raw Expression Data	526
	Loading Data from GEO	530
	Matching Phenotype Data	532
	Analyzing Expression Data	533
	Key Bioconductor Packages	537
	Data Structures	541
	eSet	541
	AssayData	543
	AnnotatedDataFrame	543
	MIAME	544
	Other Classes Used by Bioconductor Packages	545
	Where to Go Next	546
	Resources Outside Bioconductor	546
	Vignettes	546
	Courses	547
	Books	547
26.	R and Hadoop	549
	R and Hadoop	549
	Overview of Hadoop	549
	RHadoop	554
	Hadoop Streaming	568
	Learning More	571
	Other Packages for Parallel Computation with R	571
	Segue	571
	doMC	572
	Where to Learn More	572
Appei	ndix: R Reference	573
Biblio	graphy	673
	<b>.</b>	



# **Preface**

It's been over 10 years since I was first introduced to R. Back then, I was a young product development manager at DoubleClick, a company that sold advertising software for managing online ad sales. I was working on inventory prediction: estimating the number of ad impressions that could be sold for a given search term, web page, or demographic characteristic. I wanted to play with the data myself, but we couldn't afford a piece of expensive software like SAS or MATLAB. I looked around for a little while, trying to find an open-source statistics package, and stumbled on R. Back then, R was a bit rough around the edges and was missing a lot of the features it has today (like fancy graphics and statistics functions). But R was intuitive and easy to use; I was hooked. Since that time, I've used R to do many different things: estimate credit risk, analyze baseball statistics, and look for Internet security threats. I've learned a lot about data and matured a lot as a data analyst.

R, too, has matured a great deal over the past decade. R is used at the world's largest technology companies (including Google, Microsoft, and Facebook), the largest pharmaceutical companies (including Johnson & Johnson, Merck, and Pfizer), and at hundreds of other companies. It's used in statistics classes at universities around the world and by statistics researchers to try new techniques and algorithms.

# Why I Wrote This Book

This book is designed to be a concise guide to R. It's not intended to be a book about statistics or an exhaustive guide to R. In this book, I tried to show all the things that R can do and to give examples showing how to do them. This book is designed to be a good desktop reference.

I wrote this book because I like R. R is fun and intuitive in ways that other solutions are not. You can do things in a few lines of R that could take hours of struggling in a spreadsheet. Similarly, you can do things in a few lines of R that could take pages of Java code (and hours of Java coding). There are some excellent books on R, but

I couldn't find an inexpensive book that gave an overview of everything you could do in R. I hope this book helps you use R.

### When Should You Use R?

I think R is a great piece of software, but it isn't the right tool for every problem. Clearly, it would be ridiculous to write a video game in R, but it's not even the best tool for all data problems.

R is very good at plotting graphics, analyzing data, and fitting statistical models using data that fits in the computer's memory. It's not as good at storing data in complicated structures, efficiently querying data, or working with data that doesn't fit in the computer's memory.

Typically, I use a scripting language like Perl, Python, or Ruby to preprocess files before using them in R. (If the files are really big, I'll use Pig.) It's technically possible to use R for these problems (by reading files one line at a time and using R's regular expression support), but it's pretty awkward. To hold large data files, I usually use Hadoop. Sometimes I use a database like MySQL, PostgreSQL, SQLite, or Oracle (when someone else is paying the license fee).

## What's New in the Second Edition?

This edition isn't a total rewrite of the first book. But I have tried to improve the book in a few significant ways:

- There are new chapters on ggplot2 and using R with Hadoop.
- Formatting changes should make code examples easier to read.
- I've changed the order of the book slightly, grouping the plotting chapters together.
- I've made some minor updates to reflect changes in R 2.14 and R 2.15.
- There are some new sections on useful tools for manipulating data in R, such as plyr and reshape.
- I've corrected dozens of errors.

# **R License Terms**

R is an open-source software package, licensed under the GNU General Public License (GPL). This means that you can install R for free on most desktop and server machines. (Comparable commercial software packages sell for hundreds or thousands of dollars. If R were a poor substitute for the commercial software packages, they might have limited appeal. However, I think R is better than its commercial counterparts in many respects.)

#### Capability

You can find implementations for hundreds (maybe thousands) of statistical and data analysis algorithms in R. No commercial package offers anywhere near the scope of functionality available through the Comprehensive R Archive Network (CRAN).

#### Community

There are now hundreds of thousands (if not millions) of R users worldwide. By using R, you can be sure that you're using the same software your colleagues are using.

#### Performance

R's performance is comparable, or superior, to most commercial analysis packages. R requires you to load data sets into memory before processing. If you have enough memory to hold the data, R can run very quickly. Luckily, memory is cheap. You can buy 32 GB of server RAM for less than the cost of a single desktop license of a comparable piece of commercial statistical software.

# **Examples**

In this book, I have tried to provide many working examples of R code. I deliberately decided to use new and original examples, instead of relying on the data sets included with R. I am not implying that the included examples are not good; they are good. I just wanted to give readers a second set of examples. In most cases, the examples are short and simple and I have not provided them in a downloadable form. However, I have included example data and a few of the longer examples in the nut shell R package, available through CRAN. To install the nutshell package, type the following command on the R console:

#### > install.packages("nutshell")

1. There is some controversy about GPL licensed software and what it means to you as a corporate user. Some users are afraid that any code they write in R will be bound by the GPL. If you are not writing extensions to R, you do not need to worry about this issue. R is an interpreter, and the GPL does not apply to a program just because it is executed on a GPL-licensed interpreter.

If you are writing extensions to R, they might be bound by the GPL. For more information, see the GNU foundation's FAQ on the GPL: http://www.gnu.org/licenses/gplfaq. However, for a definite answer, see an attorney. If you are worried about a specific application, see an attorney.

# **How This Book Is Organized**

I've broken this book into parts:

- Part I, R Basics, covers the basics of getting and running R. It's designed to help get you up and running if you're a new user, including a short tour of the many things you can do with R.
- Part II, The R Language, picks up where the first section leaves off, describing the R language in detail.
- Part III, Working with Data, covers data processing in R: loading data into R, transforming data, and summarizing data.
- Part IV, *Data Visualization*, describes how to plot data with R.
- Part V, *Statistics with R*, covers statistical tests and models in R.
- Part VI, Additional Topics, contains chapters that don't belong elsewhere: tuning R programs, writing parallel R programs, and Bioconductor.
- Finally, I included an Appendix describing functions and data sets included with the base distribution of R.

If you are new to R, install R and start with Chapter 3. Next, take a look at Chapter 5 to learn some of the rules of the R language. If you plan to use R for plotting, statistical tests, or statistical models, take a look at the appropriate chapter. Make sure you look at the first few sections of the chapter, because these provide an overview of how all the related functions work. (For example, don't skip straight to "Random forests for regression" on page 448 without reading "Example: A Simple Linear Model" on page 401.)

### Conventions Used in This Book

The following typographical conventions are used in this book:

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, data types, environment variables, statements, and keywords. (When showing input and output on the R console, I use constant width text to show prompts and other information produced by the R interpreter.)

#### Constant width bold

Shows commands or other text that should be typed literally by the user. (When showing input and output on the R console, I use constant width bold text to show you what I typed, including comments.)

#### Constant width italic

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



This icon indicates a tip, suggestion, or general note.



This icon indicates a warning or a caution.

In this book, I will sometimes show commands that I entered on my operating system prompt (i.e., in a Bash shell on Linux), and sometimes show commands that I entered in the R console. For commands that I entered in the operating system shell, I use a \$ character to show the prompt; for commands entered in the R console, I will use > or + to show the prompt. (In either case, don't type the prompt character.)

# **Using Code Examples**

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 in a Nutshell by Joseph Adler. Copyright 2012 Joseph Adler, 978-1-449-31208-4."

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

## Safari® Books Online



Safari Books Online (www.safaribooksonline.com) is an on-demand digital library that deli digital library that delivers expert content in both book and video form from the world's leading authors in technology and business.

Technology professionals, software developers, web designers, and business and creative professionals use Safari Books Online as their primary resource for research, problem solving, learning, and certification training.

Safari Books Online offers a range of product mixes and pricing programs for organizations, government agencies, and individuals. Subscribers have access to thousands of books, training videos, and prepublication manuscripts in one fully searchable database from publishers like O'Reilly Media, Prentice Hall Professional, Addison-Wesley Professional, Microsoft Press, Sams, Que, Peachpit Press, Focal Press, Cisco Press, John Wiley & Sons, Syngress, Morgan Kaufmann, IBM Redbooks, Packt, Adobe Press, FT Press, Apress, Manning, New Riders, McGraw-Hill, Jones & Bartlett, Course Technology, and dozens more. For more information about Safari Books Online, please visit us online.

## **How to Contact Us**

Please address comments and questions concerning this book to the publisher:

O'Reilly Media, Inc. 1005 Gravenstein Highway North Sebastopol, CA 95472 800-998-9938 (in the United States or Canada) 707-829-0515 (international or local) 707-829-0104 (fax)

We have a web page for this book where we list errata, examples, and any additional information. You can access this page at <a href="http://oreil.ly/r\_in\_a\_nutshell\_2e">http://oreil.ly/r\_in\_a\_nutshell\_2e</a>.

To comment or to ask technical questions about this book, send email to bookquestions@oreilly.com.

For more information about our books, courses, conferences, and news, see our website at http://www.oreilly.com.

Find us on Facebook: http://facebook.com/oreilly Follow us on Twitter: http://twitter.com/oreillymedia

Watch us on YouTube: http://www.youtube.com/oreillymedia

# Acknowledgments

First, I'd like to thank everyone who read the first book. I wrote R in a Nutshell to be useful. I tried to write the book that I wanted to read; I tried my best to share as much useful information as I could about R. That's an ambitious goal, and I wrote an imperfect book. I appreciate all the feedback, suggestions, and corrections that I have received from readers and have tried my best to improve the book in the second edition.

I'd like to thank the team at O'Reilly for their support. Tim O'Reilly has said that he follows three guiding principles: work on something that matters to you more than money, create more value than you capture, and take the long view.<sup>2</sup> I tried to follow these principles when writing this book. As an author, I felt like the team at O'Reilly followed these principles. My goal in writing R in a Nutshell was to write the best book I could write. I hope that when people read this book, they learn something new and use what they learned to solve important problems.

2. See http://radar.oreilly.com/2009/01/work-on-stuff-that-matters-fir.html.

Many people helped support the writing of this book. First, I'd like to thank all of my technical reviewers. These folks check to make sure the examples work, look for technical and mathematical errors, and make many suggestions on writing quality. It's not possible to write a quality technical book without quality technical reviewers: Peter Goldstein, Aaron Mandel, and David Hoaglin are the reason that this book reads as well as it does.

For the past two years, I've worked at LinkedIn, ground zero for the data revolution. I've learned a huge amount working side by side with people like DJ Patil, Monica Rogati, Daniel Tunkelang, Sam Shah, and Jay Kreps. I've had the chance to discover interesting patterns, figure out how to share them with other people, and figure out how to scale my programs to work for hundreds of millions of users. I hope the second edition of this book reflects some of the lessons that I've learned on data, and helps other people learn the same things.

I'd like to thank Randall Munroe, author of the xkcd comic. He kindly allowed us to reprint two of his (excellent) comics in this book. You can find his comics (and assorted merchandise) at http://www.xkcd.com.

Additionally, I'd like to thank everyone who provided or suggested improvements. Aaron Schatz of Football Outsiders provided me with play-by-play data from the 2005 NFL season (the field goal data is from its database). Sandor Szalma of Johnson & Johnson suggested GSE2034 as an example of gene expression data. Jeremy Howard of Kaggle suggested adding glmnet.

Finally, I'd like to thank my wife, Sarah, my daughter, Zoe, and my son, Zeke. Writing a book takes a lot of time, and they were very understanding when I needed to work. They were also very understanding when I dragged them to the San Diego Zoo to look at the harpy eagles.

# **R** Basics

This part of the book covers the basics of R: how to get R, how to install it, and how to use packages in R. It also includes a quick tutorial on R and an overview of the features of R.

# sample content of R in a Nutshell (In a Nutshell (O'Reilly))

- download Great North Road here
- Martin Misunderstood pdf, azw (kindle), epub, doc, mobi
- click Monarch (War of the Princes, Book 3)
- The Everything Store: Jeff Bezos and the Age of Amazon pdf, azw (kindle)
- http://anvilpr.com/library/Great-North-Road.pdf
- http://pittiger.com/lib/Martin-Misunderstood.pdf
- http://schrolf.de/books/Monarch--War-of-the-Princes--Book-3-.pdf
- <a href="http://academialanguagebar.com/?ebooks/A-Highlander-For-Christmas--Draycott-Abbey--Book-7-.pdf">http://academialanguagebar.com/?ebooks/A-Highlander-For-Christmas--Draycott-Abbey--Book-7-.pdf</a>