Chester Ismay and Albert Y. Kim

A MODERN DIVE into Data with R

Contents

1

Prerequisites

This book was written using the **bookdown** R package from Yihui Xie. In order to follow along and run the code in this book on your own, you'll need to have access to R and RStudio. You can find more information on both of these with a simple Google search for "R" and for "RStudio." An introduction to using R, RStudio, and R Markdown is also available in a free book here (**?**). It is recommended that you refer back to this book frequently as it has GIF screen recordings that you can follow along with as you learn.

We will keep a running list of R packages you will need to have installed to complete the analysis as well here in the `needed_pkgs` character vector. You can check if you have all of the needed packages installed by running all of the lines below. The last lines including the `if` will install them as needed (i.e., download their needed files from the internet to your hard drive).

You can run the `library` function on them to load them into your current analysis. Prior to each analysis where a package is needed, you will see the corresponding `library` function in the text. Make sure to check the top of the chapter to see if a package was loaded there.

```
needed_pkgs <- c("nycflights13", "dplyr", "ggplot2", "knitr",
  "ggplot2movies", "dygraphs", "rmarkdown", "mosaic", "tibble")

new.pkgs <- needed_pkgs[!(needed_pkgs %in% installed.packages())]

if(length(new.pkgs)) {
  install.packages(new.pkgs, repos = "http://cran.rstudio.com")
}
```

Book was last updated:

```
## [1] "By Chester on Wednesday, November 09, 2016 11:56:30 PST"
```

Colophon

The source of the book is available here and was built with versions of R packages (and their dependent packages) given below. This may not be of importance for initial readers of this

book, but the hope is you can reproduce a duplicate of this book by installing these versions of the packages.

package	*	version	date	source
assertthat		0.1	2013-12-06	CRAN (R 3.3.0)
base64enc		0.1-3	2015-07-28	CRAN (R 3.3.0)
BH		1.60.0-2	2016-05-07	CRAN (R 3.3.0)
bitops		1.0-6	2013-08-17	CRAN (R 3.3.0)
caTools		1.17.1	2014-09-10	CRAN (R 3.3.0)
colorspace		1.2-6	2015-03-11	CRAN (R 3.3.0)
curl		1.2	2016-08-13	CRAN (R 3.3.0)
DBI		0.5	2016-08-11	CRAN (R 3.3.0)
dichromat		2.0-0	2013-01-24	CRAN (R 3.3.0)
digest		0.6.10	2016-08-02	CRAN (R 3.3.0)
dplyr		0.5.0	2016-06-24	CRAN (R 3.3.0)
dygraphs		1.1.1-1	2016-08-06	CRAN (R 3.3.0)
evaluate		0.9	2016-04-29	CRAN (R 3.3.0)
formatR		1.4	2016-05-09	CRAN (R 3.3.0)
ggdendro		0.1-20	2016-04-27	CRAN (R 3.3.0)
ggplot2		2.1.0	2016-03-01	CRAN (R 3.3.0)
ggplot2movies		0.0.1	2015-08-25	CRAN (R 3.3.0)
gridExtra		2.2.1	2016-02-29	CRAN (R 3.3.0)
gtable		0.2.0	2016-02-26	CRAN (R 3.3.0)
highr		0.6	2016-05-09	CRAN (R 3.3.0)
hms		0.2	2016-06-17	CRAN (R 3.3.0)
htmltools		0.3.5	2016-03-21	CRAN (R 3.3.0)
htmlwidgets		0.7	2016-08-02	CRAN (R 3.3.0)
jsonlite		1.0	2016-07-01	CRAN (R 3.3.0)
knitr		1.14	2016-08-13	CRAN (R 3.3.0)
labeling		0.3	2014-08-23	CRAN (R 3.3.0)
lattice		0.20-33	2015-07-14	CRAN (R 3.3.1)
latticeExtra		0.6-28	2016-02-09	CRAN (R 3.3.0)
lazyeval		0.2.0	2016-06-12	CRAN (R 3.3.0)
magrittr		1.5	2014-11-22	CRAN (R 3.3.0)
markdown		0.7.7	2015-04-22	CRAN (R 3.3.0)
MASS		7.3-45	2016-04-21	CRAN (R 3.3.1)
Matrix		1.2-6	2016-05-02	CRAN (R 3.3.1)
mime		0.5	2016-07-07	CRAN (R 3.3.0)
mosaic		0.14.4	2016-07-29	CRAN (R 3.3.0)
mosaicData		0.14.0	2016-06-17	CRAN (R 3.3.0)
munsell		0.4.3	2016-02-13	CRAN (R 3.3.0)

nycflights13	0.2.0	2016-04-30	CRAN (R 3.3.0)
plyr	1.8.4	2016-06-08	CRAN (R 3.3.0)
R6	2.1.3	2016-08-19	CRAN (R 3.3.0)
RColorBrewer	1.1-2	2014-12-07	CRAN (R 3.3.0)
Rcpp	0.12.6	2016-07-19	CRAN (R 3.3.0)
readr	1.0.0	2016-08-03	CRAN (R 3.3.0)
reshape2	1.4.1	2014-12-06	CRAN (R 3.3.0)
rmarkdown	1.0.9013	2016-09-14	Github (rstudio/rmarkdown@b66d11b)
scales	0.4.0	2016-02-26	CRAN (R 3.3.0)
stringi	1.1.1	2016-05-27	CRAN (R 3.3.0)
stringr	1.1.0	2016-08-19	CRAN (R 3.3.0)
tibble	1.1	2016-07-04	CRAN (R 3.3.0)
tidyr	0.6.0	2016-08-12	CRAN (R 3.3.0)
xts	0.9-7	2014-01-02	CRAN (R 3.3.0)
yaml	2.1.13	2014-06-12	CRAN (R 3.3.0)
zoo	1.7-13	2016-05-03	CRAN (R 3.3.0)

2

Introduction

2.1 Preamble

This book is inspired by three books:

- "Mathematical Statistics with Resampling and R" (**?**),
- "Intro Stat with Randomization and Simulation" (**?**), and
- "R for Data Science" (**?**).

The first book, while designed for upper-level undergraduates and graduate students, provides an excellent resource on how to use resampling to build statistical concepts like normal distributions using computers instead of focusing on memorization of formulas. The last two books also provide a path towards free alternatives to the traditionally expensive introductory statistics textbook. When looking over the vast number of introductory statistics textbooks we found that there wasn't one that incorporated many of the new R packages directly into the text. Additionally, there wasn't an open-source, free textbook available that showed new learners all of the following

1. how to use R to explore and visualize data
2. how to use randomization and simulation to build inferential ideas
3. how to effectively create stories using these ideas to convey information to a lay audience.

We will introduce sometimes difficult statistics concepts through the medium of data visualization. In today's world, we are bombarded with graphics that attempt to convey ideas. We will explore what makes a good graphic and what the standard ways are to convey relationships with data. You'll also see the use of visualization to introduce concepts like mean, median, standard deviation, distributions, etc. In general, we'll use visualization as a way of building almost all of the ideas in this book.

Additionally, this book will focus on the triad of computational thinking, data thinking, and inferential thinking. We'll see throughout the book how these three modes of thinking can build effective ways to work with, describe, and convey statistical knowledge. In order to do so, you'll see the importance of literate programming to develop literate data science. In other