Introduction to Data Science with R

Ning Lu

2019-11-27

Contents

1 Knowledge is sharing. 5

2 Think about your end before your start 7

3 Documentation - bookdown-styled 9

4 Elastigirl is Here 13

5 Applications 17
5.1 Example one 17
5.2 Example two 17

6 Data cleaning 19
6.1 dplyr 19
6.2 data.table 19

7 Q in Bond movie 21

8 Shortcuts 23
8.1 Code section 23
8.2 Tidyverse 23

9 Introduction to Machine Learning 25
9.1 Supervised learning 25
9.2 Unsupervised learning 26

10 RMarkdown with Sublime 27

Chapter 1

Knowledge is sharing.

This is a *sample* book written in **Markdown**. You can use anything that Pandoc's Markdown supports, e.g., a math equation $a^2 + b^2 = c^2$.

The **bookdown** package can be installed from CRAN or Github:

```
install.packages("bookdown")
# or the development version
# devtools::install_github("rstudio/bookdown")
```

Remember each Rmd file contains one and only one chapter, and a chapter is defined by the first-level heading `#`.

To compile this example to PDF, you need XeLaTeX. You are recommended to install TinyTeX (which includes XeLaTeX): https://yihui.org/tinytex/.

Chapter 2

Think about your end before your start

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter **??**. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter **??**.

Figures and tables with captions will be placed in `figure` and `table` environments, respectively.

```
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

Reference a figure by its code chunk label with the `fig:` prefix, e.g., see Figure 2.1. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table 2.1.

```
knitr::kable(
  head(iris, 20), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

You can write citations, too. For example, we are using the **bookdown** package (Xie, 2019) in this sample book, which was built on top of R Markdown and **knitr** (Xie, 2015).

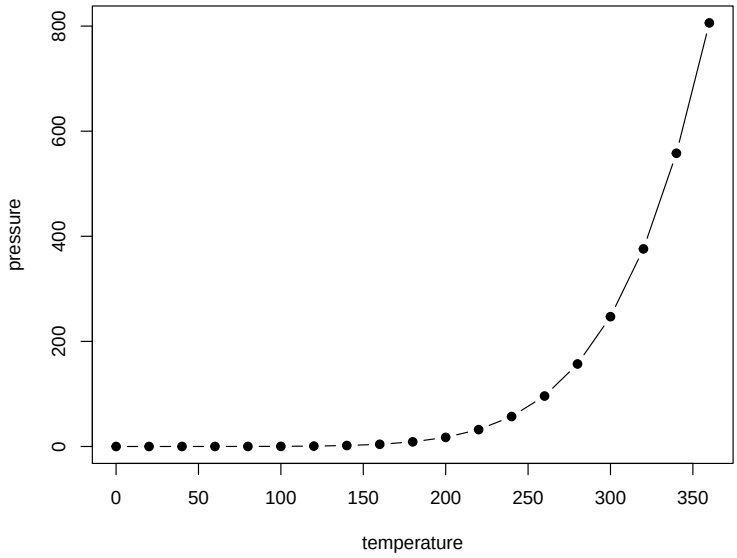

Figure 2.1: Here is a nice figure!

Table 2.1: Here is a nice table!

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa

Chapter 3

Documentation - bookdown-styled

First, choose *New Project* and *New Directiory*.

Second, choose *Book Project using bookdown* and pick a name as well as your preferred directory. RStudio will automatically set up the Rproj as well as the folder skeleton.

Third, tie the existing project with Git through the `usethis` package. It will re-organise the existing project folder and prepare the Git integration.

```
> usethis::use_git()

  Setting active project to '/Users/sushicat/Dropbox/R_Me/R_DE'
  Initialising Git repo
  Adding '.Rhistory', '.RData', '.Rproj.user' to '.gitignore'
There are 15 uncommitted files:
* '_bookdown.yml'
* '_output.yml'
* '.gitignore'
* '01-intro.Rmd'
* '02-literature.Rmd'
* '03-method.Rmd'
* '04-application.Rmd'
* '05-summary.Rmd'
* '06-references.Rmd'
* 'book.bib'
* 'index.Rmd'
* 'preamble.tex'
* 'R_DE.Rproj'
```

```
* 'README.md'
* 'style.css'
Is it ok to commit them?

1: Not now
2: For sure
3: No way

Selection: 2
 Adding files
 Commit with message 'Initial commit'
 A restart of RStudio is required to activate the Git pane
Restart now?

1: Not now
2: Yup
3: Absolutely not

Selection: 2
```

Fourth, create a GitHub repo through the `usethis` package and if the project name is available on the owner's repos. When facing git protocol, choose `https`.

```
> usethis::use_github()

 Setting active project to '/Users/sushicat/Dropbox/R_Me/R_DE'
 Checking that current branch is 'master'
Which git protocol to use? (enter 0 to exit)

1: ssh   <-- presumes that you have set up ssh keys
2: https <-- choose this if you don't have ssh keys (or don't know if you do)

Selection: 2
 Tip: To suppress this menu in future, put
  `options(usethis.protocol = "https")`
  in your script or in a user- or project-level startup file, '.Rprofile'.
  Call `usethis::edit_r_profile()` to open it for editing.
 Check title and description
  Name:        Bradford
  Description:
Are title and description ok?

1: Yeah
2: Not now
3: Absolutely not
```

```
Selection: 1
  Creating GitHub repository
  Setting remote 'origin' to 'https://github.com/dataning/R_DE.git'
  Pushing 'master' branch to GitHub and setting remote tracking branch
  Opening URL 'https://github.com/dataning/R_DE'
```

Fifth, create and save a random R.script in the current project. The commit and push the change of the project to your GitHub repo. You can go to your GitHub repo and check if the R script has been added. This should tell you whether your Rproj and GitHub Repo are fully synced/integrated.

Sixth, go to Netlify and deploy your GitHub repo on Netlify. This will give you the ability to perform continuous deployment as well as deployment to custom domain.

Type in your Rpoj's GitHub repo name.

You need to put down `_book` in *Publish directory*.

Chapter 4

Elastigirl is Here

> When designing the Incredible family, Brad Bird wanted each of their superpowers to be related to their personality. He felt that as a mother, Helen was required by society to be pulled in many different directions, which led to her being given an elastic ability.

The same we can say to all sort of data science projects. We are always required by different stakeholders to be pull in many different directions. For us, we have to nail down where we are and how to initiate a new project first.

First of all, we find where we stand.

```
> here::here()

[1] "/Users/sushicat/Dropbox/R_Me/Hero_book"
```

Second, we find out what we are being surrounded.

```
> fs::dir_ls()

01-think.Rmd           02-pm.Rmd              03-load-data.Rmd       04-tidy-data.Rmd
05-bayesian.Rmd        06-Elastigirl_1.Rmd 06_Elastigirl_1.R   20-references.Rmd
CreditCard             Creditcard_hack.R      Data                   Hero_book.Rproj
Hero_book.log          README.md              _book                  _bookdown.yml
_bookdown_files        _output.yml            book.bib               index.Rmd
packages.bib           preamble.tex           style.css
```

Third, we pick somewhere (in this case - the data folder) to explore further.

```
> fs::dir_ls("Data")
> fs::dir_ls("Data/Subway_delays")

Data/Subway_delays/Subway&SRT_Logs_April_2018.xlsx
```

```
Data/Subway_delays/Subway&SRT_Logs_February_2018.xlsx
Data/Subway_delays/Subway&SRT_Logs_March_2018.xlsx
Data/Subway_delays/Subway&SRT_Logs_May_2018.xlsx
Data/Subway_delays/Subway_&_SRT_Logs_(August_2018).xlsx
Data/Subway_delays/Subway_&_SRT_Logs_(September_2018).xlsx
Data/Subway_delays/Subway_&_SRT_Logs_December_2018.xlsx
Data/Subway_delays/Subway_&_SRT_Logs_November_2018.xlsx
Data/Subway_delays/Subway_SRT_Logs(January 2018).xlsx
Data/Subway_delays/Subway_SRT_Logs(July_2018).xlsx
Data/Subway_delays/Subway_SRT_Logs(June2018).xlsx
Data/Subway_delays/Subway_SRT_Logs(October 2018).xlsx
```

Alternatively, we can use the tree strcuture to show the folder.

```
> fs::dir_tree("Data/Subway_delays")

Data/Subway_delays
   Subway&SRT_Logs_April_2018.xlsx
   Subway&SRT_Logs_February_2018.xlsx
   Subway&SRT_Logs_March_2018.xlsx
   Subway&SRT_Logs_May_2018.xlsx
   Subway_&_SRT_Logs_(August_2018).xlsx
   Subway_&_SRT_Logs_(September_2018).xlsx
   Subway_&_SRT_Logs_December_2018.xlsx
   Subway_&_SRT_Logs_November_2018.xlsx
   Subway_SRT_Logs(January 2018).xlsx
   Subway_SRT_Logs(July_2018).xlsx
   Subway_SRT_Logs(June2018).xlsx
   Subway_SRT_Logs(October 2018).xlsx
```

Fourth, we make a shortcut if this is where we'd like to use or come back later.

```
> fs::dir_tree(here::here("Data", "Subway_delays"))

/Users/goal/Dropbox/R_Me/Hero_book/Data/Subway_delays
   Subway&SRT_Logs_April_2018.xlsx
   Subway&SRT_Logs_February_2018.xlsx
   Subway&SRT_Logs_March_2018.xlsx
   Subway&SRT_Logs_May_2018.xlsx
   Subway_&_SRT_Logs_(August_2018).xlsx
   Subway_&_SRT_Logs_(September_2018).xlsx
   Subway_&_SRT_Logs_December_2018.xlsx
   Subway_&_SRT_Logs_November_2018.xlsx
   Subway_SRT_Logs(January 2018).xlsx
   Subway_SRT_Logs(July_2018).xlsx
   Subway_SRT_Logs(June2018).xlsx
```

```
   Subway_SRT_Logs(October 2018).xlsx
```

Let's chain everything together. We present the folder with the dataset - it's like placing the meat and veggie into an oven tray. We then put the tray to an oven called `purrr` and it would import all the spreadsheet files in this particular folder - it's like an oven. Finally, we use the cleaning wipe from `janitor` and clean up the the column names - the ambiguity bit.

```
delays_clean <- fs::dir_ls(here::here("Data", "Subway_delays")) %>%
  purrr::map_dfr(readxl::read_excel) %>%
  janitor::clean_names()
```

https://sharlagelfand.netlify.com/posts/usethis-for-reporting/

Chapter 5

Applications

Some *significant* applications are demonstrated in this chapter.

5.1 Example one

5.2 Example two

Chapter 6

Data cleaning

Some *significant* applications are demonstrated in this chapter.

6.1 dplyr

6.2 data.table

Chapter 7

Q in Bond movie

> The character Q never appears in the novels by the author Ian Fleming, where only Q and the Q Branch are mentioned;[2] although Q does appear in the novelisations by Christopher Wood, and the later novels by John Gardner and Raymond Benson who adopted Eon's decision to combine the character with Major Boothroyd, the armourer from Dr. No.

There're a number of ways to set up Rproject and GitHub. Here we list two main approaches here.

The first way is the **pull** way where we get both Rproject and git integrated from outside - GitHub. You use the `github` function from `usethis` package and put down ("OWNER/REPO_NAME") and opt for https when you get asked on git protocol.

```
> usethis::create_from_github("dataning/learn_usethis")

Which git protocol to use? (enter 0 to exit)

1: ssh   <-- presumes that you have set up ssh keys
2: https <-- choose this if you don't have ssh keys (or don't know if you do)

Selection: 2
 Tip: To suppress this menu in future, put
  `options(usethis.protocol = "https")`
  in your script or in a user- or project-level startup file, '.Rprofile'.
  Call `usethis::edit_r_profile()` to open it for editing.
 Cloning repo from 'https://github.com/dataning/learn_usethis.git' into '/Users/sushicat/Desktop
 Setting active project to '/Users/sushicat/Desktop/learn_usethis'
 Writing 'learn_usethis.Rproj'
 Adding '.Rproj.user' to '.gitignore'
```

```
  Opening '/Users/sushicat/Desktop/learn_usethis/' in new RStudio session
  Setting active project to '<no active project>'
```

The second way is to imagine you're working in a random folder and you wish to set up the Rproject

```
> library(usethis)
> library(here)

here() starts at /Users/sushicat/Dropbox/R_Me
```

```
> here::here()
[1] "/Users/sushicat/Dropbox/R_Me"
```

```
> path <- file.path(here(), "learn_usethis")
create_project(path)

  Creating '/Users/sushicat/Dropbox/R_Me/learn_usethis/'
  Setting active project to '/Users/sushicat/Dropbox/R_Me/learn_usethis'
  Creating 'R/'
  Writing 'learn_usethis.Rproj'
  Adding '.Rproj.user' to '.gitignore'
  Opening '/Users/sushicat/Dropbox/R_Me/learn_usethis/' in new RStudio session
  Setting active project to '<no active project>'
```

Chapter 8

Shortcuts

8.1 Code section

- Insert Section — Ctrl+Shift+R (Cmd+Shift+R on the Mac)
- Jump To — Shift+Alt+J

RStudio Addin - creates boxed in titles in an Rscript

`devtools::install_github("ThinkR-open/littleboxes")`

https://riptutorial.com/r/example/13622/assignment-with-----

https://style.tidyverse.org/pipes.html

8.2 Tidyverse

8.2.1 "%>%"

- "Insert %>%" inserts "%>%" (Shift-ALT-m or Shift-cmd-m

One is to create a new dataset and the other one is to re-assign the new data elements into an existing dataset.

```
db <- df %>%
      select(1:3) %>%
      filter(mpg > 20, cyl == 6)

df %<>% select(1:3) %>%
        filter(mpg > 20, cyl == 6)

# Good
iris %>%
  group_by(Species) %>%
```

```
  summarize_if(is.numeric, mean) %>%
  ungroup() %>%
  gather(measure, value, -Species) %>%
  arrange(value)

# Bad
iris %>% group_by(Species) %>% summarize_all(mean) %>%
ungroup %>% gather(measure, value, -Species) %>%
arrange(value)
```

Chapter 9

Introduction to Machine Learning

There're two main tribes of machine learning. One tribe is called *supervised learning* and it's focused on problems with predictive nature. The other one is called *unsupervised learning* and it's focused on problems with descriptive nature.

9.1 Supervised learning

The idea of supervised learning is that given a set of attributes or features of the objects or targets, the model is then trained to generate accurate prediction on the objects or targets. The idea of supervision refers to the fact that the target values provide a supervisory role, which means that the learning algorithm is expected to go through a set of cases where the target an its features are both presented. The learning algorithm (such as tree algorithm, neural networks) will then determine the right combination or optimisation among the features (such as feature engineering) to predict the targets.

Some of the common challenges involved with supervised learning are including, - To predict the likelihood of customer churning over a certain period of time through a set of customer attributes; - To predict the likelihood of employee retention of using a range of employee and workplace attributes; - To predict the transaction price of real estates through a range of home attributes; - To predict the re-admission of the patients through a set of patient attributes and symptoms.

Some of the family members are expected to exist in machine learning - Y: Outcome measure (or response, target, dependent variable)

- Y is numeric. When the outcome measure is numeric / continuous outcome, this is regarded as reg

- Y is categorical. When the outcome measure is categorical outcome, this is regarded a

- X: Feature measure (or predictor, attribute, independent variable)
- training data
- testing data

9.2 Unsupervised learning

In essence, unsupervised learning is concerned with identifying groups in a data set. The groups may be defined by the rows (i.e., clustering) or the columns (i.e., dimension reduction). The goal of clustering is to segment observations into similar groups based on the observed variables; for example, to divide consumers into different homogeneous groups, a process known as market segmentation. In dimension reduction, we are often concerned with reducing the number of variables in a data set.

Unsupervised learning is often performed as part of an exploratory data analysis (EDA). However, the exercise tends to be more subjective, and there is no simple goal for the analysis, such as prediction of a response. - Divide consumers into different homogeneous groups so that tailored marketing strategies can be developed and deployed for each segment. - Identify groups of online shoppers with similar browsing and purchase histories, as well as items that are of particular interest to the shoppers within each group. Then an individual shopper can be preferentially shown the items in which he or she is particularly likely to be interested, based on the purchase histories of similar shoppers. - Identify products that have similar purchasing behavior so that managers can manage them as product groups.

Chapter 10

RMarkdown with Sublime

One of the best things for R markdown is that you can write and code on the same surface. However, the interface of RStudio feels like more "brief writing" than "proper writing". For all the writing, I prefer to do it in Sublime because it's fast and code-integrated.

The way to do is to install `R-box` and `knitr` in the Sumline.

- Shift + CMD + P
- Select `R-box` or `knitr` packages
- Once the installation is finished, please make sure the file type is R Markdown from the bottom right corner.
- You should provide a basic YAML setting for your R markdown.

```
---
title: Lens and Movements Piloting (LAMP)
author: Lawrence Ning
date: March 22, 2005
output: pdf_document
---
```

Bibliography

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2019). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.16.