STA 444/5 - Introductory Data Science using R

Derek L. Sonderegger

August 12, 2020

Contents

Preface

This book is intended to provide students with a resource for learning R while using it during an introductory statistics course. The *Introduction* section covers common issues that students in a typical statistics course will encounter and provides a simple examples and does not attempt to be exhaustive. The *Deeper Details* section addresses issues that commonly arise in many data wrangling situations and is intended to give students a deep enough understanding of R that they will be able to use it as their primary computing resource to manipulate, graph and model data.

The pdf version of this book isn't quite as good as the on-line version because I've had to remove some of the animated gifs as well as remove chapters that show how to create html output.

Other Resources

There are a great number of very good online and physical resources for learning R.

- Hadley Wickham and Garrett Grolemund's free online book R for Data Science. This is a wonderful introduction to the `tidyverse` and is free. If there is any book I'd recommend buying, this would be it. Many of the topics my book covers are perhaps better covered in Hadley and Garrett's book. However, I think it is better to triangulate on a concept utilizing multiple sources so I've presented my taking on teaching these concepts.
- Hadley Wickham and Jenny Bryan have a whole book on R packages to effectively manage large projects.
- Hadley Wickham also has a book about Advanced R programming and is quite helpful in understanding deeper issues relating to Object Oriented program in R, Environments, Namespaces, and function evaluation.

Non-Hadley books:

- Michael Freeman's book Programming Skills for Data Science. This book covers much of what we'll do in this class and is quite readable.

Acknowledgments

These online books are a huge amount of work and without the support of my wife Aubrey, this book would not be possible.

Introduction

Chapter 1

Familiarization

Placeholder

1.1 Working within an Rmarkdown File

1.2 R file Types

1.2.1 R Scripts (.R files)

1.2.2 R Markdown (.Rmd files)

1.2.3 R Notebooks (.Rmd files)

1.3 R as a simple calculator

1.4 Assignment

1.5 Vectors

1.6 Packages

1.7 Finding Help

1.7.1 How does this function work?

1.7.2 How does this package work?

1.7.3 How do I do XXX?

1.8 Exercises

Chapter 2

Data Frames

Placeholder

2.1 Introduction to Importing Data

2.1.1 From a Package

2.1.2 Import from `.csv` or `.xls` files

2.2 Data Types

2.3 Basic Manipulation

2.4 Exercises

Chapter 3

Graphing

Placeholder

3.1 Basic Graphs

3.1.1 Scatterplots

3.1.2 Box Plots

3.2 Faceting

3.3 Annotation

3.3.1 Axis Labels and Titles

3.3.2 Text Labels

3.3.2.1 Using a `data.frame`

3.3.2.2 Setting attributes in-line

3.4 Exercises

Chapter 4

Data Wrangling

Placeholder

4.1 Verbs

4.1.1 `add_row`

4.1.2 `bind_rows`

4.1.3 Subsetting

4.1.3.1 `select()`

4.1.3.2 `filter()`

4.1.3.3 `slice()`

4.1.4 `arrange()`

4.1.5 mutate()

4.1.6 summarise()

4.2 Split, apply, combine

4.3 Exercises

Chapter 5

Statistical Models

Placeholder

5.1 Formula Notation

5.2 Basic Models

5.2.1 t-tests

5.2.1.1 Two Sample t-tests

5.2.1.2 Paired t-tests

5.2.2 lm objects

5.3 Accessor function

5.4 Exercises

Chapter 6

Flow Control

Placeholder

6.1 Logical Expressions

6.2 Decision statements

6.2.1 In `dplyr` wrangling

6.2.2 General `if else`

6.3 Loops

6.3.1 `while` Loops

6.3.2 `for` Loops

6.3.3 `mosaic::do()` loops

6.4 Functions

6.5 Exercises

Chapter 7

Factors

Placeholder

Edit Factor Labels

Reorder Levels

Add or Subtract Levels

7.1 Creation and Structure

7.2 Change Labels

7.3 Reorder Levels

7.4 Add or substract Levels

7.5 Exercises

Miscellaneous

Placeholder

Example Distributions

`mosaic::plotDist()` function

Base R functions

d-function

p-function

q-function

r-function

Exercises

Rmarkdown Tricks

Placeholder

Chunk Options

Verbatim & List Environments

7.6 [1] 5

Mathematical expressions

Tables

7.7 Girth Height Volume

7.8 1 8.3 70 10.3

7.9 2 8.6 65 10.3

7.10 3 8.8 63 10.2

7.11 4 10.5 72 16.4

7.11.1 Simple Tables

7.11.2 Grid Tables

7.11.3 Pipe Tables

R functions to produce table code.

7.11.4 `knitr::kable`

7.11.5 Package `pander`

Code Appendix

7.11.6 Code Appendix

Data Wrangling Process

Placeholder

7.12 Introduction

7.13 Import

7.14 Tidying

7.15 Cleaning

7.16 Use

Deeper Details

Chapter 8

Data Structures

Placeholder

8.1 Vectors

8.1.1 Accessing Vector Elements

8.1.2 Scalar Functions Applied to Vectors

8.1.3 Vector Algebra

8.1.4 Commonly Used Vector Functions

8.2 Matrices

8.3 Data Frames

8.3.1 `data.frames` vs `tibbles`

8.4 Lists

8.5 Exercises

Chapter 9

Importing Data

Placeholder

9.1 Working directory

9.2 Comma Separated Data

9.3 MS Excel

9.4 Multiple files

9.5 Exercises

Chapter 10

Functions

Placeholder

10.1 Basic function definition

10.2 Parameter Defaults

10.3 Ellipses

10.4 Function Overloading

10.5 Debugging

10.5.1 Rmarkdown Recommendations

10.5.2 Step-wise Execution

10.5.3 Print Statements

10.5.4 `browser`

10.6 Scope

10.7 Exercises

Chapter 11

String Manipulation

Placeholder

11.1 Base function

11.2 `stringr`: Basic operations

11.2.1 Concatenating with `str_c()` or `str_join()`

11.2.2 Calculating string length with `str_length()`

11.2.3 Extracting substrings with `str_sub()`

11.2.4 Pad a string with `str_pad()`

11.2.5 Trim a string with `str_trim()`

11.3 `stringr`: Pattern Matching Tools

11.3.1 Detecting a pattern using str__detect()

11.3.2 Locating a pattern using str__locate()

11.3.3 Replacing sub-strings using `str_replace()`

11.3.4 Splitting into sub-strings using `str_split()`

11.4 Regular Expressions

11.4.1 Regular Expression Ingredients

11.4.2 Matching a specific string

11.4.3 Matching arbitrary numbers

11.4.4 Greedy matching

11.5 Fuzzy Pattern Matching

11.5.1 Key Collision Merge

11.5.2 String Distances

11.5.3 N-gram Merge

11.6 Exercises

Chapter 12

Dates and Times

Placeholder

12.1 Creating Date and Time objects

12.2 Extracting information

12.3 Arithmetic on Dates

12.4 Exercises

Chapter 13

Data Reshaping

Placeholder

13.1 `data.frames` vs `tibbles`

13.2 `cbind` & `rbind`

13.3 `tidyr`

13.3.1 Verbs

13.4 Storing Data in Multiple Tables

13.5 Table Joins

13.6 Row summations

13.7 Exercises

Chapter 14

R Packages

Placeholder

14.1 Introduction

14.1.1 Useful packages and books

14.2 Package Structure

14.2.1 Minimal files and directories

14.2.2 Optional Files and Directories

14.3 Documenting

14.3.1 Data Documentation

14.3.2 Documenting Functions

14.4 Testing

14.5 The DESCRIPTION file

14.6 Sharing your Package

14.7 An Example Package

14.8 Exercises

Chapter 15

Data Scraping

Placeholder

15.1 Web Pages

15.1.1 Example Wikipedia Table

15.1.2 Lists

15.2 Scraping .pdf files

15.3 Exercises

Chapter 16

API Data Queries

Placeholder

16.1 Introduction

16.2 Census Bureau API

16.3 Package `censusapi`

16.3.1 Population Estimates

16.4 Package `tidycensus`

16.5 Exercises

Chapter 17

Databases

Placeholder

17.1 Tutorial Set-Up

17.2 SQL

17.3 `dbplyr`

17.4 Exercises