Taylor R. Brown

An Introduction to R and Python For Data Analysis: A Side By Side Approach

To Clare

Contents

List of Tables

List of Figures

Welcome

Teaching a Course With This Textbook

You will notice that some of the exercise questions are unusually specific. For example, they will ask the student to assign the answer in a certain form to a variable with a very specific name. This is because they are written with automatic grading in mind.

All of the exercises in this text have been very generously "battle tested" by the Fall 2021 STAT 5430 class at the University of Virginia.

License(s)

Preface

About this book

This book is written to be used in a computing class that teaches both R and Python to graduate students in a statistics or data science department. This book is written for students that do not necessarily possess any previous familiarity with writing code.

- If you are using them for analyzing data, R and Python do a lot of the same things in pretty similar ways, so it does not always make sense to either a.) teach one language after the other, or b.) choose one over the other. The side-by-side approach helps to teach/learn more, save time, and reinforce shared concepts and key differences.
- This text does not describe statistical modeling techniques in detail, but many exercises will ask students to implement them. These exercises will not assume mathematical familiarity. The objective is to test programming ability, and an added benefit is that the work might be useful later in their careers.
- This book is written for aspiring data scientists, not necessarily aspiring software developers. Why do I draw the distinction? When discussing different types, for example, I do not discuss data structures in any depth. Rather, I discuss examples of applications where different types would be most useful.
- Generally speaking, chapters should be read in order, but occasionally skipping ahead can be beneficial. Later chapters are more may assume some familiarity with previous chapters. Also, within a chapter's section, sometimes a discussion for a particular topic in Python, say, might follow a discussion about the same topic in R. In this case, the later section will often assume that the previous section has been read first.

Conventions

Sometimes R and Python code look very similar, or even identical. This is why I usually separate R and Python code into separate sections. However, sometimes I do not, so whenever it is necessary to prevent confusion, I will remind you what language is being used in comments (more about comments in **??**).

```
# in python
print('hello world')
## hello world
```

```
# in R
print('hello world')
## [1] "hello world"
```

Installing the Required Software

To get started, you must install both R and Python. The installation process depends on what kind of machine you have (e.g. what type of operating system your machine is running, is your processor 32 or 64 bit, etc.).

Below, I suggest running R with RStudio, and Python with Anaconda, and I provide some helpful links. I suggest downloading these two bundles separately; however, I should note that the recommendation below is not the only installation method. For example: - one can run R and Python without downloading RStudio or Anaconda, - one can install RStudio with Anaconda, - one can run Python from within Rstudio, - one can run Python from within Rstudio that is managed by Anaconda, etc., and - options and procedures are very likely to change over time.

Instructors can prefer alternative strategies, if they wish. If they do, they should verify that Python's version is `>=3.6`, and R's is `>=4.0.0`. If so, all the code in this book should run.

Installing R (and RStudio)

It is recommended that you install R and *RStudio Desktop*. *RStudio Desktop* is a graphical user interface with many tools that making writing R easier and more fun.

Install R from the Comprehensive R Archive Network (CRAN). You can access instructions for your specific machine by clicking here.[1]

You can get RStudio Desktop directly from the company's website[2].

Installing Python by Installing Anaconda

It is recommended that you install *Anaconda*, which is a package manager, environment manager, and Python distribution with many third party open source packages. It provides a graphical user interface for us, too, just as RStudio does. You can access instructions for your specific machine and OS by clicking here[3].

[1] `https://cran.r-project.org/`

[2] `https://www.rstudio.com/products/rstudio/download/#download`

[3] `https://docs.anaconda.com/anaconda/install/#`

Part I

Introducing the Basics