L01 Introduction

Data Science I (STAT 301-1)

Shay Lebovitz

Contents

Overview

The goal of this lab is to ensure that all relevant software for this course is properly installed and functional. Students will also download and install important packages. We will demonstrate the basic workflow for future labs which can be generally applied to any data analytic project. We will also work together to construct and run some basic code.

Don't worry if you cannot do everything here by yourself. You are just getting started and the learning curve is steep, but we and your classmates will be there to provide support. Persevere and put forth an honest effort and this course will payoff.

Datasets

We will use the `catsvdogs.txt` dataset. See the codebook, `catsvdogs_codebook.txt` for a detailed description of the variables.

Exercises/Tasks

Complete the following exercises/tasks. For many of these you'll need to simply indicate that you have completed the task. In others, you'll need to run some R code and/or supply a sentence or two.

Exercise 1

Download and install R Software.

Exercise 2

Download and install RStudio.

Exercise 3

Install the following packages:

- `tidyverse`
- `skimr`
- `janitor`
- `nycflights13`
- `gapminder`
- `Lahman`

Exercise 4

Create a project folder, open and save an R script in it, save this Rmd file there, and make sure to have a */data* directory within the project folder that will hold our data and its associated codebook. Appropriately rename the R and Rmd files for submission (e.g. *Coburn_Katie_LO1.Rmd*).

Exercise 5

Suppose a random variable X has finite variance, then as we take larger random samples (i.e. as n increases) we have that

$$\bar{X} \sim N\left(\mu_{\bar{X}} = \mu_X, \sigma^2_{\bar{X}} = \frac{\sigma^2_X}{n}\right)$$

This is an informal statement of which important statistical theorem?

Answer: central limit Theorem

Exercise 6

If the mathematical notation in Exercise 5 does not compile correctly you will need to download and install either MikTex for windows machines or MacTex for Mac OS machines.

Even if the equation in Exercise 5 does compile, you may need to download MikTex/MacTex to compile pdfs. We highly suggest being able to compile pdf's, but it is not a requirement for undergraduates in this course.

Graduate students are required to be able to compile pdf's. It is more for your own good going forward, than being directly beneficial for this course. Please include a complied pdf of this assignment in addition to the standard submission files (.R, .Rmd, & .html).

Exercise 7

It is always handy to have a versatile text editor. Enter the name of the text editor you have below. We suggest downloading Sublime Text. It is free.

Exercise 8

Read the codebook for the `catsvdogs.txt` dataset and upload it using `readr::read_delim()` function. The `readr::` tells you that the function `read_delim()` function is from the `readr` package which is part of the `tidyverse`.

```
catdog<- read_delim('catsvdogs.txt', '|')
```

What was the percentage of dog owners for Illinois in 2012? Answer: 32.4%

Exercise 9

Apply the `skim()` function from the `skimr` package to the `catdog` dataset. What does the `skim()` function return?

skim(catdog) #returns a summary of the data

Exercise 10

Calculate the mean of `percent_dog_owners`. Do you think this is a reasonable estimate for the percent of US dog owners? Why or why not?

mean(catdog$percent_dog_owners) #36.97347, seems about right