STAT 19000 Project 3

Topics: Getting Familiar with R

Motivation: In this project, we go deeper into our learning about the R platform. We use vectorized operations, which work on a whole data set all at once. We will gradually get into this frame of mind, for thinking about data sets. This is one of the key assets of R, namely, that once we are used to R's functions, we can do a lot of things with just a little bit of effort. Once we get used to it, this makes R much more versatile and powerful than many other kinds of platforms for data analysis. R is also good at being the common glue that enables us to use several tools together, in tandem, for our analysis. For now, however, we can just focus on getting familiar with the kinds of things that we can do very easily with R.

Context: All of our learning about R will be driven by example data sets. We want to get familiar with importing data sets and working with data that is larger than what our eyes can see. With just a handful of functions, we can quickly understand the types of data in our data set. We do some basic summaries that enable us to begin to understand our data.

Scope: Most of what we do in R is vectorized, i.e., we write just (say) one function and it works on a whole vector of data. This happens automatically, without us needing to know how everything works in the underlying software. We also do not need to write loops in R. Instead, we get familiar with working with a whole vector of data all at once. We will see this by examples that use grocery store transaction data, taxi cab data, and airport data.

Question 1: Introduction to Grocery Store Customer Insights

Try the following *inside* the Scholar environment. This question deals with data from 84.51. They use data science to provide customer insights to Kroger.

Open the file `/class/datamine/data/examples/project3examples.R` and run through the commands in R. As we did last week, you can open the file inside R and type Control-Return on each line to run the line. You do not need to highlight the lines, and you do not need to be at the end of the line to run it. In fact, you can just type Control-Return over and over, to quickly execute each line, one after another. Please be sure to also read the comments in the file as well. It is OK (indeed, it is encouraged) to experiment with the code give there.

1a. Make a table of the number of transactions that occur in each of the four stores in the 84.51 data set called `5000_transactions.csv`, found in the folder `The_Complete_Journey_2_Master`.

Question 2: Introduction to New York City Yellow Taxi Cab Rides

2a. Load the data about yellow taxi cab rides in New York City in June 2019. This data is already stored on Scholar. The data dictionary is located here

https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf

and is also already located on Scholar for your convenience, in the file

`data_dictionary_trip_records_yellow.pdf`

in this directory:

`/class/datamine/data/taxi/dictionary/`

2b. How many passengers (altogether) rode in yellow taxi cab rides in New York City in June 2019?

Question 3: Creating Your Own Map of Airports in the (Continental) United States

3a. Modify the code from Project 2 to visualize the locations of all airports in the (continental) USA. The latitudes and longitudes for this data are given on this webpage:

http://stat-computing.org/dataexpo/2009/supplemental-data.html

You can download the data directly into R, using a `read.csv` command directly on this URL:

http://stat-computing.org/dataexpo/2009/airports.csv

Project Submission:

Submit your solutions for the project at this URL: https://classroom.github.com/a/foV0xkYT using the instructions found in the GitHub Classroom instructions folder on Blackboard.