STAT 19000 Project 4

Topics: Vectorized Operations in R

Motivation: One of the reasons that R is *powerful* is that we can accomplish a lot, with just a *handful of simple functions*. The data sets that we are working with are huge. Some of them have millions of data points. It would take us a very, very long time to accomplish the things in this project by hand. If we learn just a little bit of R, however, we can (very quickly) accomplish a lot of analysis, using just a few commands in R.

Have fun with this project! Learn to ask some questions yourself! Challenge yourself as you analyze the data, and see what you can do! You are already learning the skills of data analysis and data science. Put your new knowledge to work, as you explore the data. You can already do some very cool things!

Context: We do not use any sophisticated statistical analysis in this project. We simply start with basics:

- how to find the average of a vector of data (just add up a sum, and divide by the number of entries; R can do this very easily),
- how to paste two vectors together into a new vector,
- how to sort a vector of data,
- how to make a table that gives the number of entries of each type of data, etc.

We can do all of these things very easily with R. We can do a lot, with only a little bit of familiarity.

Scope: We usually import data in R into a data frame. A data frame is analogous to a spreadsheet in Excel.

There are columns of data, which must have the same type. We often work with just one column of data, which we sometimes call a vector of data.

Many operations in R are very fast and simple to use, when we apply them to a vector of data, in other words, to a column of a data frame. We just need to practice a little bit. We will quickly become familiar with some of R's basic operations.

We encourage you to experiment with these functions. Play around a little bit with the data. With just a little bit of experimentation, we think you will quickly learn to ask questions about the data that you can easily answer with R.

Question 1: Which cities have the most airports?

Before you start question 1, please consider the examples given here:

`/class/datamine/data/examples/project4examples.R`

These will help you to better understand some of the functions that we will use for this project.

1a. Import the data about the airports from the file that we used in Question 3 of Project 3, namely,

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

1b. Paste together the columns for the city and state, using the `paste` command.

1c. Use the `table`, `sort`, and `tail` commands to find the 4 cities that have the most airports. (Ignore the missing data `NA NA`.)

Question 2: What Does a Typical New York City Yellow Taxi Cab Ride Look Like?

2a. What is the mean total fare ("total amount") for a yellow taxi cab ride in June 2019?

2b. What is the mean total number of passengers in a New York City yellow taxi cab ride in June 2019?

2c. Which pickup location ID was the most popular for yellow taxi cab rides in June 2019? Which location does that correspond to in New York?

It might be helpful to remember that the data dictionary is given here:

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

and the list of all pickup locations is given here:

https://s3.amazonaws.com/nyc-tlc/misc/taxi+_zone_lookup.csv

so once you find the PULocationID, you can find the name of the location in this table. A moment's thought should show that this location is the correct one!

Question 3: Which Cities are Contributing to the 2020 Presidential Election Campaign Contributions?

Import the data about the federal campaign contributions that have been given (so far) during the 2020 election season. The data is described here:

https://www.fec.gov/campaign-finance-data/contributions-individuals-file-description/

When you import the data, use the `read.delim` command, with the argument `sep="|"` because the file is not comma-separated. It has the symbol `|` between the elements of the data.

```
myDF <- read.delim("/class/datamine/data/election/itcont2020.txt", sep="|")
```

3a. Why do you think that the data is not comma-separated? What is a reason that they chose to use `|` instead of a comma, as a delimiter for the data? Hint: Consider, for instance, these names of donors: `myDF$NAME[9001:9050]`

3b. Paste together the columns about the city and state where the donors live. Make a new column of the data frame called `location` which contains the city and state (together) as one new column. Hint: You can use `names(myDF)` to see the names of all of the columns of the data frame.

3c. Which city has the largest number of donations so far, in the 2020 election season? (Do not worry about the dollar amount. Only consider the number of donations.)

Project Submission:

Submit your solutions for the project at this URL: https://classroom.github.com/a/9qvDFep3 using the instructions found in the GitHub Classroom instructions folder on Blackboard.