EDAV Fall 2019 PSet 5, part A

Foad Khoshouei, Nima Chitsazan

This assignment is designed to help you get started on the final project. Be sure to review the final project instructions (https://edav.info/project.html), in particular the new section on reproducible workflow (https://edav.info/project.html#reproducible-workflow), which summarizes principles that we've discussed in class.

1. The Team

[2 points]

a) Who's on the team? (Include names and UNIs)

Foad Khoshouei fk2377

Nima Chitsazan nc2806

b) How do you plan to divide up the work? (Grading is on a group basis. The point of asking is to encourage you to think about this.)

We plan to work together on the project entirely and cover all parts as a team.

2. The Questions

[6 points]

List three questions that you hope you will be able to answer from your research.

a) How does taxi fare change with regards to trip distance and time of the day and day of the week.

b) What are the popular areas of pickup/drop-off in NYC for green taxis.

c) How does the percent of tip change with regards to total fare and what does it depend on.

3. Which output format do you plan to use to submit the project?

[2 points]

We plan to submit the project in html_document format.

(You don't have to use the same format for this assignment – PSet 5, part A – and the final project itself.)

Choices are:

pdf_document

html_document

bookdown book: https://bookdown.org/yihui/bookdown/

shiny app: https://shiny.rstudio.com/

(Remember that it's ok to have pieces of the project that don't fit into the chosen output format; in those cases you can provide links to the relevant material.)

4. The Data

What is your data source? What is your method for importing data? Please be specific. Provide relevant information such as any obstacles you're encountering and what you plan to do to overcome them.

[5 points]

The data comes from NYC TLC website. We are going to work on the data collected on green taxis in NYC for 2018 in NYC. Since the data is stored monthly, and we plan to use the data for the full year (2018), we are using R to access the links on TLC website which stores the data on Amazon S3 buckets to download and combine them as one data-set. In addition, since the data-set for the full year is huge, we plan to sub-sample it to a workable size. The data source is:

https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

5. Provide a short summary, including 2-3 graphs, of your initial investigations.

[10 points]

Initially looking at the histogram of Fares (outliers removed), the values are distributed anywhere from $3 to about $32. Most of the fares are within the range of $5 to $20 with a maximum frequency at about $7.

Inspecting the percentage of tip (100*tip/fare) shows that values are anywhere from 0% to about 50%. It is interesting to observe that the histogram shows some bi-modality. A big group of passengers paid a tip of 0% and the other big category is around 20%. This shows there is some inconsistency in tipping behavior of green taxi costumers.

Lastly, we plotted the histogram for distance travelled for each trip and the values are distributed between 0 and 10 miles. It is interesting to observe that most of the trips are less than 3 miles in distance.