Assignment: Final Project Step 2

Name: Saurabh Shrestha

Date: May 23, 2021

Project Name: Uber pickup analysis and distribution in New York City

How to import and Clean data

My project will be on Uber pickup and its distribution. The dataset I have in hand mostly consists of date and time of pickup, location of the pick-up in terms of longitude and latitude, base which is TLC base company code affiliated with the Uber pick up. The datasets I will be working are CSV (comma separated value) and XLSX(Excel). For importing in R, I can use readxl library and use read.csv to import the data from csv file.

Final Dataset

One of the data set looks like below where we have date and time and latitude and longitude of the pick-up location in New York City.

Date/Time	Lat	Lon
4/1/2014 0:11	40.769	-73.9549
4/1/2014 0:17	40.7267	-74.0345
4/1/2014 0:21	40.7316	-73.9873
4/1/2014 0:28	40.7588	-73.9776
4/1/2014 0:33	40.7594	-73.9722
4/1/2014 0:33	40.7383	-74.0403
4/1/2014 0:39	40.7223	-73.9887
4/1/2014 0:45	40.762	-73.979
4/1/2014 0:55	40.7524	-73.996
4/1/2014 1:01	40.7575	-73.9846
4/1/2014 1:19	40.7256	-73.9869
4/1/2014 1:48	40.7591	-73.9684
4/1/2014 1:49	40.7271	-73.9803
4/1/2014 2:11	40.6463	-73.7896
4/1/2014 2:25	40.7564	-73.9167
4/1/2014 2:31	40.7666	-73.9531

Another data set which will be using is with variables active vehicles(Uber), the trips it took within a certain time frame (a year).

date	active_vehicles	trips
1/1/2015	190	1132
1/1/2015	225	1765
1/1/2015	3427	29421
1/1/2015	945	7679
1/1/2015	1228	9537
1/1/2015	870	6903
1/2/2015	785	4768
1/2/2015	1137	7065
1/2/2015	175	875
1/2/2015	890	5506
1/2/2015	196	1001
1/2/2015	3147	19974
1/3/2015	201	1526
1/3/2015	1188	10664

Questions for future steps

- Did I delete any column unnecessarily?
- Are my choices of variable in data set good for the model?
- Are there other methods to make my dataset look clean?

What information is not self-evident?

For Uber rides there are other factors which can affect the rides frequency. Not self-evident factor could include weather of the day, demographics of the area of the Uber ride and so on. It is difficult to create model based on little data which I was able to find after long research. Whatever I found on Uber dataset was all that on the Internet regarding the subject matter. Some other information which is not shown/given includes cost of the Uber ride in the given location. This can affect number of rides based on time of the day.

What are different way you could look at this data?

With the collected datasets I will be able to select, mutate variables, join the datasets. Other learned methods from the class I will be able to find correlation between variables such as pick up hours, date of pick up, location of pick-up, frequency of rides in certain location. Based on these variables and using learned methods from the class I will be able to solve my research questions.

How do you plan to slice and dice the data?

I sliced my data based on time of the hour for pick up, total number of rides for the pick-up in each given month and total number of active vehicles based on the date and hours.

How could you summarize your data to answer key questions?

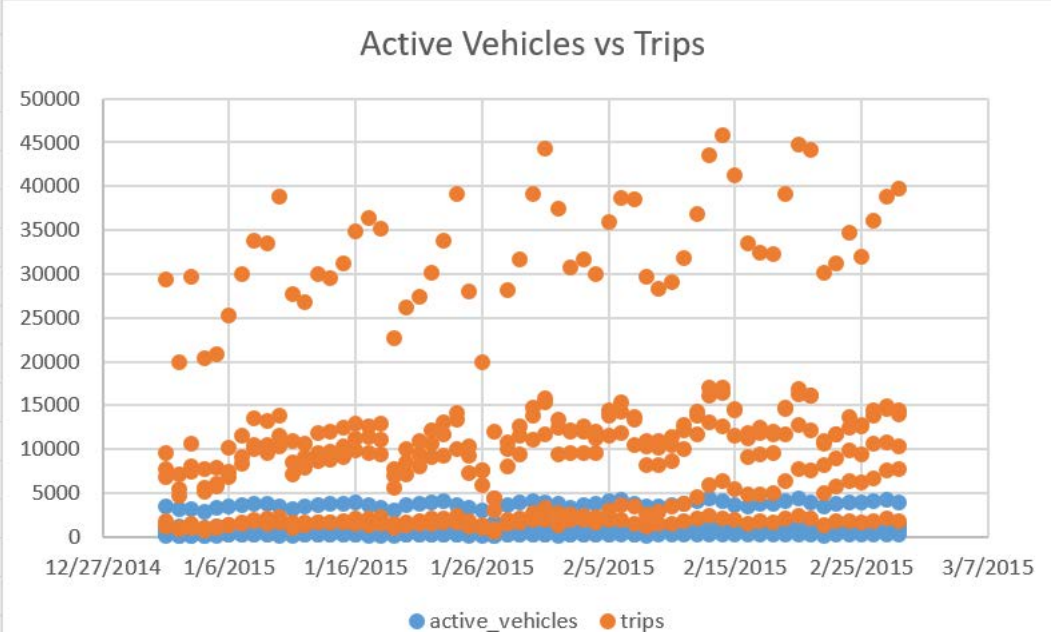
I will use summary function to get the data I need in question. This will be done after slicing and dicing of the data has been completed.

Plot and Table Needs:

The plots used for the project will include the following:

- Scatterplot
- Histogram
- Line
- Bar chart
- Area chart

One of the plot used for this project is as under. It is a scatterplot which shows active Uber vehicles and the frequency of trips it took for pick-up for each day.



Incorporating any machine learning techniques to answer my research questions?

I will be using multiple linear regression model to learn the relationship between time of the day and frequency of Uber pickups.

Questions for future steps:

1. Will the found datasets be enough for finding adequate solution or will additional data could be fruitful for the project?
2. What methods or style can I incorporate to improve on clear analysis suitable for the reader?
3. What further things I need to learn to better my project? i.e. packages, writing technique and so on.
4. Is my datasets void of any inaccuracy, redundancy or of errors for me to analyze?