Assignment: Final project Step 1

Name: Saurabh Shrestha

Data: 5/16/2021

Project: Analysis of Uber pickup and distribution in New York City

Introduction

Uber as we all have heard of is a way of modern transportation. With easy-to-use app, anyone can call up for a ride anywhere with a touch of a button. The ride is especially popular in large cities such as New York city which is arguably known as taxi capital of America. Since the popularity of Uber, it has taken over taxi rides and fewer yellow-cab taxi are seen as compared to a decade ago. Since, transactional activity for the ride is done only via online (mobile app, websites), it was easy to gather large datasets, thanks to Uber. There are millions of ride pick-ups done by Uber within a given year. I am addressing the frequency of Uber pick ups in New York city. It is interesting to learn from this project because we can view how sky rocketing has Uber taken over the For-Hire Vehicle sector in New York city. Data science is in the center of Uber since all the activity is taken place through data usage (customer id, Uber drive id, credit card, location, ratings)

Research Questions

My research questions are about: what time period is Uber highly used in New York city? Other potential research I would be working on would be which location in New York city has high

concentration of Uber rides, high frequency trips taken by Uber in 5 boroughs of the city, trips taken during every day of the month and days when the pick-up happen regularly.

Approach:

There are datasets I will be using where first I will clean up the data using tidyr. Other requirements for the approach would be to find essential predicting variables. As I learned from this week I will work with R2, adjusted statistics, p-value to come to conclusive evidence for my objective. Additional calculation I will be performing would be to find standardized residuals, the leverage, cooks distance, confidence intervals to see if the created model is likely to represent general population.

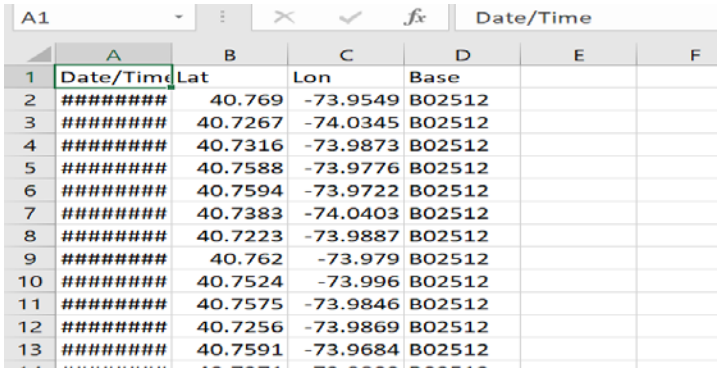
How your proposed approach will address (fully or partially) this problem

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.

Datasets

- *Uber Pickups in New York City*. (2019, November 13). Kaggle. https://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city

  Above dataset is a subset of Uber NY city record. Variables include longitude, latitude, data/time of pick up and Base.



	A	B	C	D	E	F
1	Date/Time	Lat	Lon	Base		
2	#########	40.769	-73.9549	B02512		
3	#########	40.7267	-74.0345	B02512		
4	#########	40.7316	-73.9873	B02512		
5	#########	40.7588	-73.9776	B02512		
6	#########	40.7594	-73.9722	B02512		
7	#########	40.7383	-74.0403	B02512		
8	#########	40.7223	-73.9887	B02512		
9	#########	40.762	-73.979	B02512		
10	#########	40.7524	-73.996	B02512		
11	#########	40.7575	-73.9846	B02512		
12	#########	40.7256	-73.9869	B02512		
13	#########	40.7591	-73.9684	B02512		

- F. (n.d.). *fivethirtyeight/uber-tlc-foil-response*. GitHub. https://github.com/fivethirtyeight/uber-tlc-foil-response/tree/master/uber-trip-data

	LocationID	Borough	Zone
1	1	EWR	Newark Airport
2	2	Queens	Jamaica Bay
3	3	Bronx	Allerton/Pelham Gardens
4	4	Manhattan	Alphabet City
5	5	Staten Island	Arden Heights
6	6	Staten Island	Arrochar/Fort Wadsworth
7	7	Queens	Astoria
8	8	Queens	Astoria Park
9	9	Queens	Auburndale
10	10	Queens	Baisley Park
11	11	Brooklyn	Bath Beach
12	12	Manhattan	Battery Park
13	13	Manhattan	Battery Park City
14	14	Brooklyn	Bay Ridge

Above is dataset which provides location id for Uber pick up, borough of New York city for the pick up and zone of New York city.

- https://github.com/fivethirtyeight/uber-tlc-foil-response/blob/master/uber-trip-data/uber-raw-data-janjune-15.csv.zip

	A	B	C	D
1	Dispatchir	Pickup_da	Affiliated_	locationID
2	B02617	########	B02617	141
3	B02617	########	B02617	65
4	B02617	########	B02617	100
5	B02617	########	B02774	80
6	B02617	########	B02617	90
7	B02617	########	B02617	228
8	B02617	########	B02617	7
9	B02617	########	B02764	74
10	B02617	########	B02617	249
11	B02617	########	B02764	22
12	B02617	########	B02617	263
13	B02617	########	B02617	61
14	B02617	########	B02617	229
15	B02617	########	B02617	164
16	B02617	########	B02617	237
17	B02617	########	B02617	142
18	B02617	########	B02617	188
19	B02617	########	B02617	237
20	B02617	########	B02617	224
21	B02617	########	B02617	238
22	B02617	########	B02682	242

Above dataset shows pick up times , location ID, pick up day.

Required Packages

- ggplot2
- ggthemes
- lubridate
- tidyr
- dplyr
- plyr
- outliers
- readxl
- car
- lmtest

Plot and Table Needs:

The plots used for the project will include the following:

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

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. Will my findings bear any business use if proper model and results are to be found?
3. Are there any other packages I haven't learned so far which could be of use for this project?
4. Is the above datasets void of any inaccuracy, redundancy or of errors for me to analyze?