Assignment: Final Project Step 3

Name: Saurabh Shrestha

Date: June 5, 2021

Project Name: Uber pick up analysis 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 pickups 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)

Problem Statement (Questions incurred)

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.

Address problem statement (data used, method employed, recommendation for a model)

Firstly, I collected the dataset from Kaggle titled “Uber Pickup in New York City”. The data set had trip data for over 20 million Uber (and other for-hire vehicle) trips in NYC. Files included were from April to September of 2014.

Second I worked on missing value with omitting rows to omit all rows that contain NA value with na.omit() function. I also had date and time which were unreadable to human eye. For that I converted Date. Time to type date time using mutate function by extracting day of the month, extracting day of the week and extracting the hour of the day such as hour_of_day = hour (Date. Time)

Few example I used for methodology such as mutate, analyze average count of trip per week day, comparison of days with each month, the average of number of rides considering every dame date of every month, trips by month. Additionally, used heat map to see Uber pickups for days’ vs months. It showed me which day of the week taken the number of rides in the month and another heat map showcased which base has taken the number of rides in the month and for which I used group by and summary functions.

Analysis/Interesting insight from analysis

From the dataset analysis what I found interesting were (from April to September 2014 datasets):

1. September was the highest Uber pick-ups month in New York City and April the lowest.

2. Trips by day and month analysis showed that Sundays had the lowest Uber pick-ups in New York city compared to other days.
3. Numbers changed for other days' pickups. Such as in September, Saturday had the highest pick-ups whereas in July, Tuesday and Wednesday had the highest pickups in the overall month.
4. Analysis on number of trip with hour of the day showed that highest peak was during 17th till 19th hour – during 5pm to 7pm. Lowest number of trip were during 2- 4 am in the morning.
5. Within New York City (Manhattan, Queens, Brooklyn, Staten Island, Bronx), highest number of pick-ups were from Manhattan borough- specially from lower central park till lower Manhattan. (downtown area)

Implications to the consumers

In my earlier step 1 final project, I had mention, my research paper if analyzed adequately could have benefits for business people and or regular customers. From my analysis, I think it could benefit for-hire vehicle owners, Uber driver and even Uber riders. Which time of the day, hour, weekday vs weekend, borough has the highest pick up could benefit the Uber driver and owner of the business. Other benefit for Uber driver would be knowing which month had the highest pickups (such as September) and which had the lowest (April) For Uber customers, it could be beneficial in knowing when to avoid taking the Uber ride. Such as during peak hour from 5:00 till 7:00 pm range showed highest amount of demand. And Uber algorithm works on higher demand equal higher price for fare. So, it's better to avoid.

Limitations of the analysis

1. The datasets I found were limited to New York City thus the analysis were restricted within the city.
2. Timeline of the Uber pickup dataset were limited in nature (only for certain months of certain years) so the analysis was for limited time period.
3. The dataset only had very minimal columns such as data, time of Uber pick up, location of pickup. So, analysis mostly worked on were time, location, rides and soon.
4. Other information/data which could have improved analysis would be customer demographic, expense incurred in each borough of New York City.
5. Others could improve by either posting Uber data or by implementing new research question and solving on it.

Concluding Remarks

Based on the research I was able to understand how to work with dataset in R. Analyzing them based on function from different packages such as ggplot2, ggmap, dplyr, tidyverse, repr, lubricate helped to slice, clean, group, summarize, mutate, arrange, manipulate dataset and visualize the analyzed data with varied plots such as scatterplot, Heat map, bar chart, diagram. With analysis and visualization of the dataset it became easier to understand from the data. From my dataset I was able to grasp knowledge on how Uber rides are getting popular in New York City. Also, while working on the data, my research showed me statistics of how popular yellow cab was diminishing.