Analysing NYC Taxi Data

Objective: Thoroughly analyze the Taxi trip and fare datasets and provide useful.

Preliminary Findings.

1. The top 10 busiest spots of pickup and dropoff.
2. Anomaly in the data wherein the the co-ordinates for pickup and dropoff is recorded as (0.0,0.0). Surprisingly there are plenty of medallions with this defect. We are assuming that this is due to faulty operation of GPS. Faulty because there are numerous records which has a pickup location but the dropoff location is recorded as 0.0 and vice versa.

Final Findings:

1. Further analysing on the GPS data we found that about 589 taxis do not have GPS working throughout the year and just these taxis contribute to about 92545 rides per year; and an additional 610 taxis did not have GPS readings for majority of the year; Thus, this incomplete data does not help determine the busiest hub in manhattan wherein the difference in number of rides for Top 10 are fairly close.

   Further we found that around 101,183 had no entry for pickup locations and 220,460 did not have entry for drop-off locations. To better understand this we queried to find multiple taxis with such records and tracked its movement in ascending order of time and with little difference in time we picked the next pickup location to the record with drop off as '0.0'. Shockingly most of these locations were in the manhattan, predominantly lower manhattan and east midtown. On further research we learnt that it is caused by the presence of tall buildings that scatter the GPS signals.
2. We decided to delve further into tip amount in the fare data set. Some interesting findings were:
   a. Total rides in 2013 = 173179771
      Total rides paid by Card = 93334004

Total tip data for rides paid by card = 90450929

Total rides paid in Cash = 79110096

Total tip data for rides paid in cash = 6708

Because there is not enough data for tip paid in cash we decided to perform further analysis only with Card payment records.

b. From tip amount analysis for payment with credit card we found that the percentage of people who do not tip vary greatly between midnight and 5:00 AM which is otherwise constant for the rest of the day. Our assumption for this pattern is; longer the person stayed at a bar lesser the tip. Apart from this major factor other reason that could justify this data is probably frustrated people working late shifts.
c. Further performing boundary analysis around fare amounts that has 10 as multiple (e.g., 68,69,70,71,72) as the average of tip steadily increases with fare amount, we find that people tend to tip less when fare is exactly a multiple of 10 in this case 70. This could be for two reasons, One is the person's tendency to round off to nearest multiple of 5 or 10 to do an easier math and ignore a standard 18% tip, other reason could be that person's thinking that spending $70 or $71 is lot more than $69 or $68.

3. In trip data; on querying to find the average number of trips in a month for each hour of the day, we noticed a sudden drop in number of rides between 4:00 PM and 5:00 PM. To understand this, we further analysed number of trips and total number of taxis on daily basis. And this fact was still evident that there was a drop in total rides just between 4 to 5 PM and it rises again after that. To get a better reasoning we pulled out the total number of taxis(medallions) and the drivers(hack license) for that time interval and found that there were more drivers than taxis (an average of 235). This could mean that there was a **<u>shift of drivers</u>.**

<u>Failed Attempts:</u>

1. We were able to find a rich dataset (NYC Open Data) on the Traffic collisions. That had relatable fields like latitude and latitude of collision and the type of vehicle involved as 'TAXI'. An unbelievable number of around 15,000 major and minor TAXI collisions have occurred. We wanted to find that if any of these taxi's had a pickup entry and if the dropoff was at the collision location also if the passenger paid for

the trip. We failed to find and match because most of the location data in the collision data set are street and junction names. And we also cannot relate dropoff time with collision time because it's not exact. Another hurdle was being unable to relate collisions without passengers to the trip data provided.

2. With the medallion, hack licence and taxi running times we set out to find number of drivers working day and night shift, also how long the taxi was in motion without any passenger. We could not obtain accurate results because the taxi could have been without any passenger for really long hours between the actual shift end and last drop off.

Technologies Used:

- Initially we struggled with the performance of NYU HPC with uploading large files from local machine so moved the game to AWS EMR with hive.
- Exhausting the credits shortly after preliminary findings, after a prolonged period of time we uploaded all the data sets into NYU HPC. The performance was much better than the initial period.
- For analysis we build complex and efficient SQL queries compatible with hive norms, such that the have would run as few mapper functions as possible.
- The resulting tables were exported and used to plot graphs for better visualizations.

Team Contribution:

All tasks including discussions, planning, execution and documentation were always carried out in group ensuring equal contribution from all.

The following link is a ppt for the prelimnary finidings

https://docs.google.com/a/nyu.edu/presentation/d/1bgv0si4ddk0ybhQLapPngREQIxSXB3-wVP7drhaLmMw/edit?usp=sharing