INFORMATICS FOR

ENGINEERING MANAGEMENT

EM -624 FINAL PROJECT

NYC Taxi Trip Data Analysis

Submitted by: AARTHI SHUNMUGAM

Guided by: Dr. Carlo Lipizzi

Research Question:

This project is designed to analyze the statistical data of Taxi trips in New York City. The analysis includes comparing two different type of taxis in New York, Yellow cabs Vs Green cabs. The main aim of this research is to answer the following questions.

1.Which cab is most commonly used by the people of NYC, Yellow cab or Green cab?

2. What is the average distance people take for both Green and Yellow cabs?

3.How much people tip on average when compared to Uber?

4. What is the difference between Yellow and Green cab? Also, to compare the average speed for Yellow cabs vs Green cabs.

Research motivation:

These days in major cities like New York, people often use Uber and Lyft for their travel. Uber is serving New York's outer boroughs more than taxis due to the increase in the public transportation, which became Uber's best friend. So, there arises a question, Are Taxis like Yellow and Green cabs are still in use by public in New York city? The answer to this question can be explained by the following study.

Recently, I read an article from fivethirtyeight.com about a debate between Uber and New York City Mayor over whether the ride-for-hire company was exacerbating Manhattan congestion was fueled by incomplete, misleading data. There was no way of knowing exactly where Uber cars and taxis pick up passengers, and so the city agreed to a study of Uber's effects from the year 2015 as part of its detente with the company.

So, data was collected for 93 million trips taken by Uber and conventional taxis over a six-month period from April to September for the year 2015, including date, time and coordinates of the pickups. And while we can't yet say whether Uber has exacerbated Manhattan congestion, the data we've analyzed shows that Uber has a point when it claims that it is doing a better job than taxis in serving

the boroughs of New York City outside of Manhattan. Of the 4.4 million Uber rides for which the data shows a pickup location, 22 percent started outside of Manhattan, compared with just 14 percent of the 88.4 million yellow and green taxi rides.

Dataset Description:

URL: http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

Source of Data:

New York City Taxi and Limousine Commission.

On browsing through the websites, I found this interesting data from New York City Taxi and Limousine commission. The data describes about the taxi yellow and green cab services in New York.

It contains Trip sheet data from the year 2009 to 2017 for each month from January to December (includes the records of total trips for both yellow and green cabs).

The yellow and green taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data used in the attached datasets were collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP).

Types of Data Items:

There are totally 17 attributes is the dataset. The dataset consists of mostly numerical values. That is the attributes in the dataset are mix of continuous and categorical data. The list and description of the data items are below.

- VendorID - A code indicating the TPEP provider that provided the record. 1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.
- tpep_pickup_datetime - The date and time when the meter was engaged.

- tpep_dropoff_datetime - The date and time when the meter was disengaged.
- Passenger_count - The number of passengers in the vehicle. This is a driver-entered value.
- Trip_distance - The elapsed trip distance in miles reported by the taximeter.
- Pickup_longitude - Longitude where the meter was engaged.
- Pickup_latitude - Latitude where the meter was engaged.
- RateCodeID - The final rate code in effect at the end of the trip. 1= Standard rate 2=JFK 3=Newark 4=Nassau or Westchester 5=Negotiated fare 6=Group ride
- Store_and_fwd_flag - This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka “store and forward,” because the vehicle did not have a connection to the server. Y= store and forward trip N= not a store and forward trip
- Dropoff_longitude - Longitude where the meter was disengaged.
- Dropoff_ latitude Latitude - where the meter was disengaged.
- Payment_type - A numeric code signifying how the passenger paid for the trip. 1= Credit card 2= Cash 3= No charge 4= Dispute 5= Unknown 6= Voided trip
- Fare_amount - The time-and-distance fare calculated by the meter. Extra Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges.
- MTA_tax- $0.50 MTA tax that is automatically triggered based on the metered rate in use. Improvement_surcharge - $0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.
- Tip_amount – This field is automatically populated for credit card tips. Cash tips are not included.
- Tolls_amount - Total amount of all tolls paid in trip.
- Total_amount - The total amount charged to passengers. Does not include cash tips.

Data Preparation:

Step 1:

Both the datasets (the dataset for yellow cab and the dataset for blue cab) was collected for the month of June for the year 2017.

Step 2:

It is required to merge these two datasets, so that it will be easy to compare the trips of both yellow and green taxis, and do the analysis.

Step 3:

The data set is cleaned by removing the missing values that is the NA values from the concatenated final dataset.

Also, not much cleaning is required for this data as it was simple and easy to handle.

Methodology:

During the exploration, each variable was carefully analyzed and compared to other variables and eventually the target variable, percentage tip. The following was the procedure that I followed to carry out the data analysis.

1. Initially all the required python packages that is required to do the analysis are imported, such as pandas, matplotlib, Numpy, seaborn and datetime.
2. Both the csv files for yellow and green cabs are read using pandas. In order, to merge the datasets for comparing both the cabs, it is necessary to rename the columns into the same title.
3. Identify the missing values or NA's to clean the data.
4. Then the analysis is done by plotting graphs for different variables, which helps us to identify the average number of trips that taxis take for each hour throughout the day, peak time when people take taxis and to find out the tips percentage and fares for booking taxis to NYC airports.
5. During the analysis there was issues while plotting the graphs, as the size of data was too large, and it took a lot of time to obtain the results.

Data Analysis:

The below figure represents the bar plot, which shows the average number of trips taken by the taxis on the y-axis and the pickup date for the entire month of June 2017 on the x axis. The plot shows the graph for both yellow cabs and green cabs.

So, from the below graph it is observed that yellow cabs take more trips when compared to green cabs. This proves that green cabs cover only short distance inside Manhattan, whereas people prefer yellow cabs for large distances, hence higher number of trips for yellow cabs.

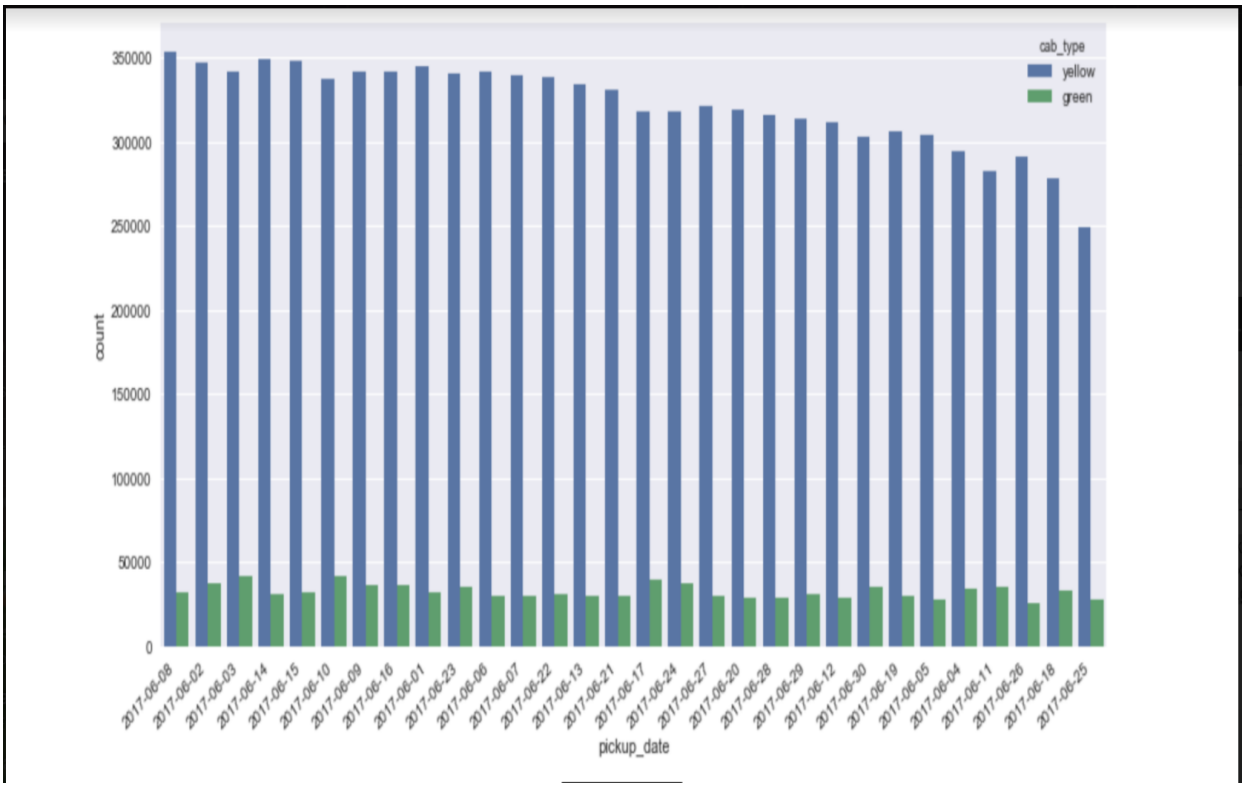

Figure 1. Average number of trips covered by both yellow and green cabs for a day throughout June 2017.

The below figure represents the bar plot, which shows the average distance covered by both yellow and green cabs for a day. It is plotted by considering the data for June 2017. The x-axis represents the type of cab (yellow or green) and the y-axis represents the average miles travelled by the taxis on that day.

So, from the below figure it is clear, that the average distance of yellow cabs is high when compared to green, which implies that yellow cabs travel more distance than green cabs. From this analysis, we can conclude that people prefer to take yellow cabs for long distance than green cabs. This proves the practically reality that yellow cabs cover the areas outside Manhattan and Green cabs predominantly serve inside Manhattan.

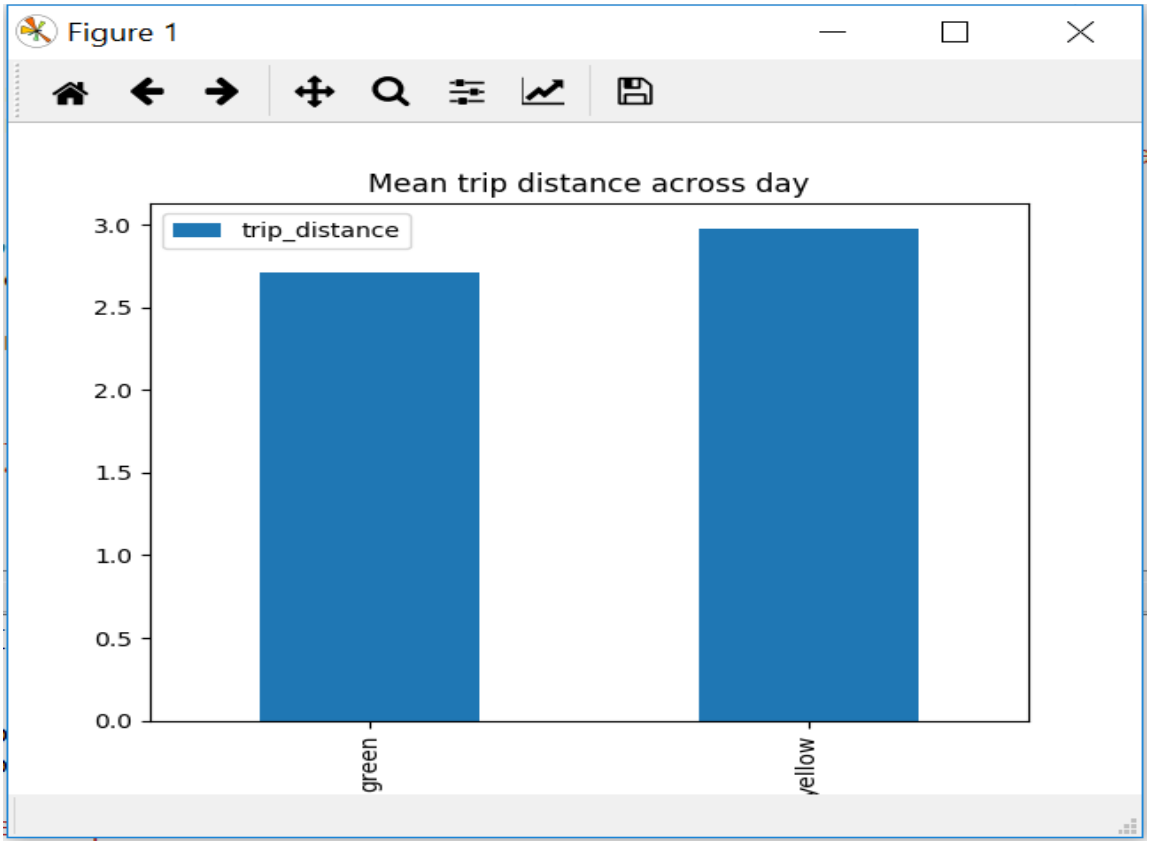

Figure 1: Average trip distance travelled by both yellow and green cabs for a day.

Identifying the peak time during the day. From the below figure it is clear, that evening 6pm to 7pm are the peak hours when people take cabs frequently.

We observe that, the taxis take more number of trips in the evening than morning. I would hypothesize that the people prefer to take cabs back from work and that would be considered as the peak time, when the taxis are busy.

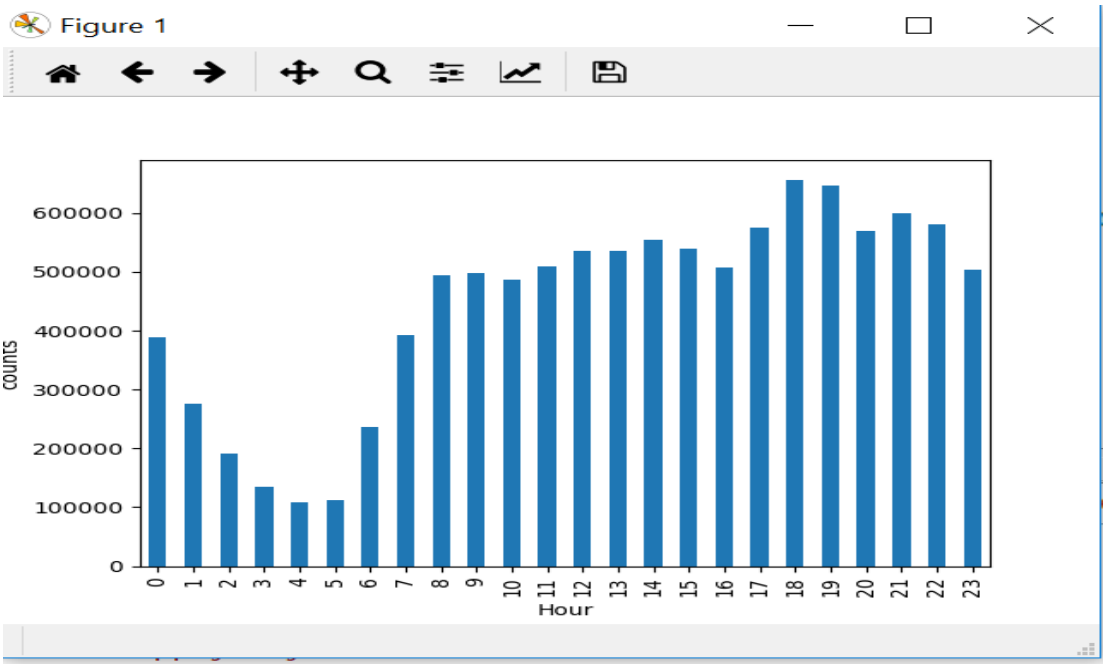

Figure 2.a) Average number of trips taken per hour for June 2017.

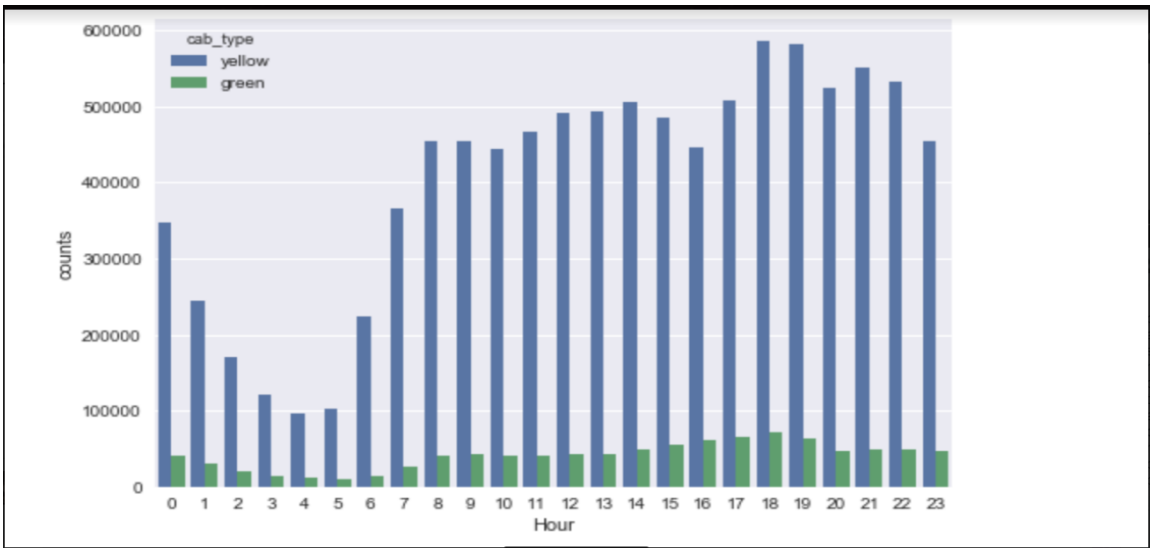

Figure 2.b) Average number of trips taken per hour by both yellow and green cabs for June 2017

Let's also compare trips that originate (or terminate) from (at) one of the NYC airports. We can look at how many they are, the average fair,

Reading through the dictionary of variables, I found that the variable RateCodeID contains values indicating the final rate that was applied. Among those values, I realized that there is Newark and JFK which are the major airports in New York.

In this part, I will use this knowledge and group data with RateCodeID 2 (JFK) and 3 (Newark). - An alternative (which I didn't due to time constraint) is to (1) get coordinates of airports from google map or http://transtats.bts.gov

The below figure shows the total number of trips that the Taxis take to /from NYC airports. Also, the average fare of trips (calculated by the meter) to/from NYC airports is approximately 53$ per trip and the average total amount charged excluding the tip is approximately 67$ per trip. The average total amount charged is high when compared to the amount charged by the meter. This is because the final total amount will include Tolls and MTA taxes. Hence, the final amount is higher than the meter reading.

```
]: airports_trips = df1[(df1.RatecodeID==2) | (df1.RatecodeID==3)]
   print ("Number of trips to/from NYC airports: ", airports_trips.shape[0])
   print ("Average fare (calculated by the meter) of trips to/from NYC airports: $", airports_trips.fare_amount.mean(),"per trip")
   print ("Average total charged amount (before tip) of trips to/from NYC airports: $", airports_trips.total_amount.mean(),"per trip

   Number of trips to/from NYC airports:  246165
   Average fare (calculated by the meter) of trips to/from NYC airports: $ 53.16590762293584 per trip
   Average total charged amount (before tip) of trips to/from NYC airports: $ 67.33707631062713 per trip
```

Figure 3. Number of trips and average fare charged to/from the NYC airports.

This was the key phase of my analysis. A look at the distribution of the target variable, "Tip_percentage" showed that 60% of all transactions did not give tip (see Figure below, left). A second tip at 18% corresponds to the usual NYC customary gratuity rate which fluctuates between 18% and 25%.

Based on tip_percent value counts, it seems substantial number of users have tipped 0 %. The users who tip usually seems to fall between 18 to 25 % which falls in the average gratuity percentage range.

A check on Google map shows that people generally tip more while taking yellow or green cabs than Uber, which is proved by the above analysis.

```
In [8]: df1['tip_percent'] = 100 * df1['tip_amount']/df1['fare_amount']
        print((df1['tip_percent']).value_counts().head())

0.000000     3973714
22.000000     107385
20.000000      86549
16.666667      82778
24.000000      75370
Name: tip_percent, dtype: int64
```

Figure 4. Tip percentage based on fare amount.

Conclusion:

While observing the results of the analysis carefully, it can be inferred that people take yellow cabs frequently than green cabs. And the peak time when the taxi drivers pick up the passengers is in the evening between 6 to 8 pm, whereas the morning pickups seems to be less. While, comparing the taxis with Uber, people tip high for taxis when compared to Uber. The tip percentage is around 18% to 25% for taxis. Also, people prefer taxis when compared to Ubers, in busy areas like Manhattan in New York city, green cabs for short and yellow taxis for long distances. People who do not use mobile apps also take taxis instead of Uber.