The Visualisation and Analysis of Yellow Taxi data of New York

(For the period of 2017 January)

Bohan Yang (814642)

I. Introduction

This is a report of basic visualisations and analysis on which factors influence the profitability for yellow taxi drivers in New York City. Specifically, which areas of New York City are more profitable for yellow taxi drivers are investigated. Time period of pickups and weather are also taken as possible factors influencing the income of yellow taxi drivers to investigate and analyse. The sample data is of the time period of 2017 January. The project is done with Python 3, mainly using library Pandas, Geopandas and Matplotlib. The data is collected from NYC Taxi and Limousine Commission (TLC) (https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page).

II. Data Period and Attributes Selection

The sample data used in this project is of 2017 January. The year 2017 is chosen because the datasets after 2015 has attributes “taxi zone” instead of “longitude” and “latitude”. This saves computing power when loading in the dataset since it is a much smaller file. Furthermore, the taxi zone ID’s are convenient for plotting a choropleth with a shapefile of taxi zones of New York City.

The attributes chosen for analysis and visualisation are locations of pickups, amount of tips paid by card and pickup times for each trip. Weather data gathered from www.wunderground.com is also taken as an attribute. The amount of tips is considered rather than total fare amount because tips is the most flexible part of the taxi drivers' income composition.

III. Data Cleansing

Firstly, only the attributes "tpep_pickup_datetime", "PULocationID", "DOLocationID", "payment_type" and "tip_amount" are kept and renamed for the reduction of dataset size and convenience of plotting. Secondly, for the purpose of dealing with missing values and avoiding future errors, the instances that contain 0 in any of the attributes are deleted. When these 0's are not missing values, they are not valuable for the visualizations as they will not increase the counts. The unit of temperature for the weather dataset is converted from Fahrenheit to Celsius for easier understanding.

IV. Pre-processing

Several separate summary data frames are created for the efficiency and clarity of plotting different graphs, and then converted into csv files. In order to create geospatial visualizations, all data instances are grouped by pick up location ID, which represents different taxi zones on the map of New York. The amount of tips (received by card only) received by each trip are then aggregated over different zones or time range, for various purpose of plotting. Average amount of tips on each day is calculated to investigate the relations with weather. Average amount of tips received over a time period in each taxi zone is also calculated and visualized.

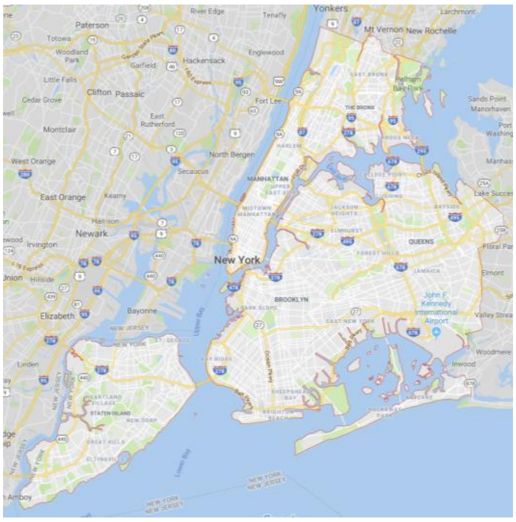
One problem with this dataset is that there are only 251 instances after grouping by taxi zones, due to the problem that there were not any trips in 14 taxi zones in 2017 January. Furthermore, there are 263 taxi zones according to the shape file provided by NYC Taxi and Limousine Commission (TLC), but the taxi dataset also has location ID

264 and 265, whose borough and zone names are unknown. Therefore 14 rows with missing location ID's are added into the data frame by a for loop, with attribute values being 0. Since the location ID 264 and 265 are not included on the map, the instances associated with these two location ID's are deleted for the sake of choropleth plotting.

V. Visualisation and Analysis

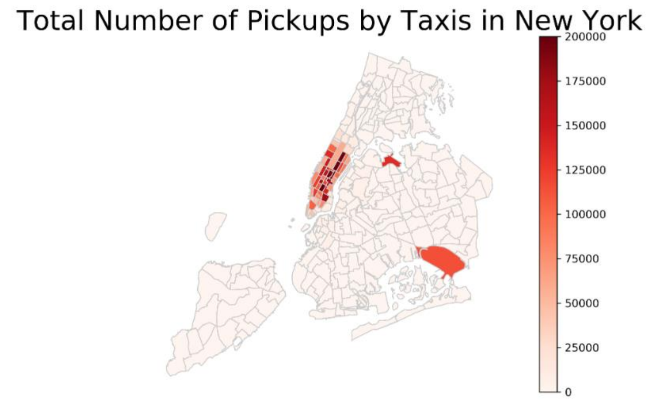
a. Number of Trips

Borough Map of New York



Source: https://goo.gl/maps/aCXKRCSsJRyyQCS2A

Figure 2

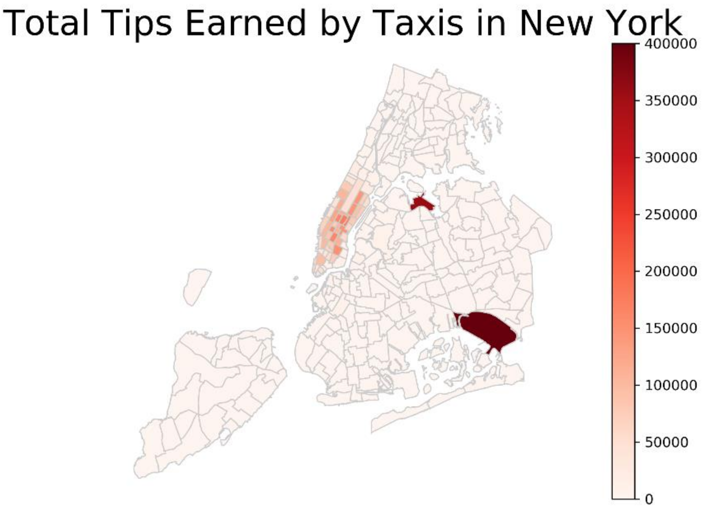


Source: NYC Taxi & Limousine Commission, 2017

Figure 1

As shown from the choropleth above (Figure 2) with assistance of Figure 1, the majority of taxi trips are gathered at Manhattan area. Surprisingly, even though allowed to pick up from all five boroughs, most yellow taxis seem to be picking up from mostly the Manhattan area and the two airport areas.

b. Total Amount of Tips Paid by Card of All Trips



Source: NYC Taxi & Limousine Commission, 2017

Figure 3

Since the number of pickups does not represent the profitability of taxis, and the most flexible and therefore improvable part of taxi driver income is tips, the total amount of tips paid in each taxi zone is plotted on the map of New York City (Figure 3). As shown from the choropleth, even though some zones of the Manhattan have the highest number of pickups, the two airport areas have the highest amount of tips collected. Although the data for tips here excluded tips paid by cash and other methods, the difference between Manhattan area and airport area is significant enough for yellow taxi drivers to consider.

c. The Correlation between Weather and Tips Amount

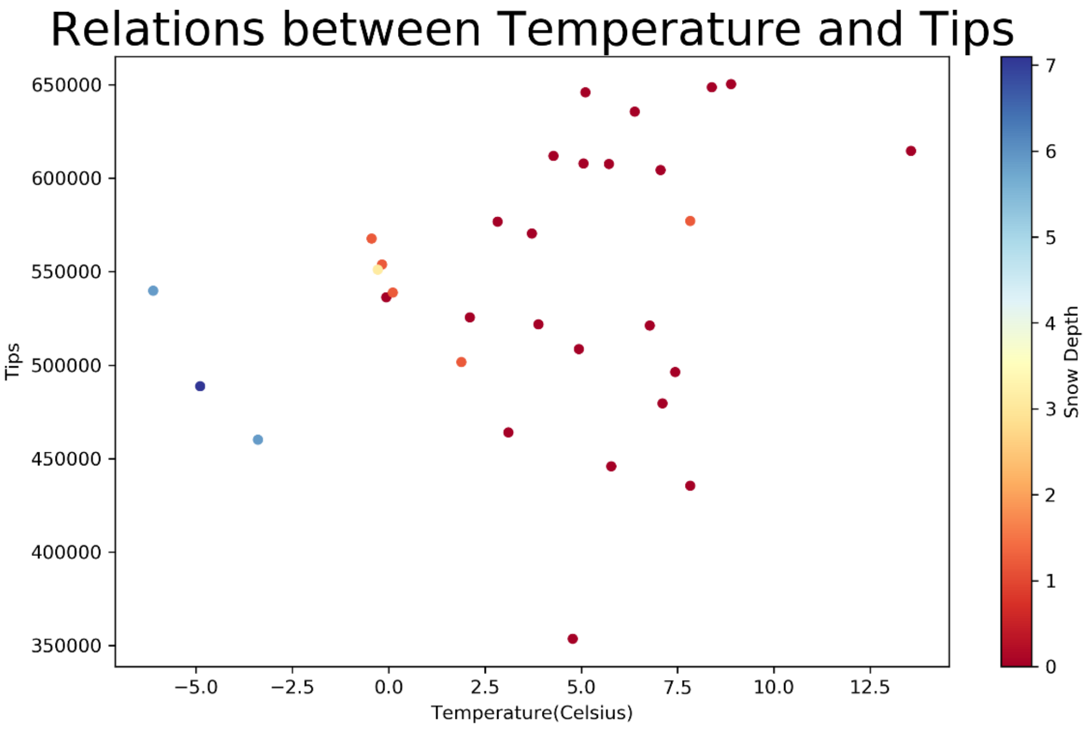


Figure 4

The weather data of New York City in 2017 January is extracted from *https://www.visualcrossing.com/weather/weather-data-services*. Each data point represents the amount of tips accumulated for each day for all yellow taxis in New York City in 2017 January. As shown from the scatter plot (Figure 4), The temperature has no obvious relation to the amount of tips. On the days when the temperature drops to the lowest (around -2.5℃ to -5℃) with the heaviest snow, the amount of tips is only around the median number. This motivates the investigation of total number of trips in relations to the weather, since this can alter the total amount of tips.

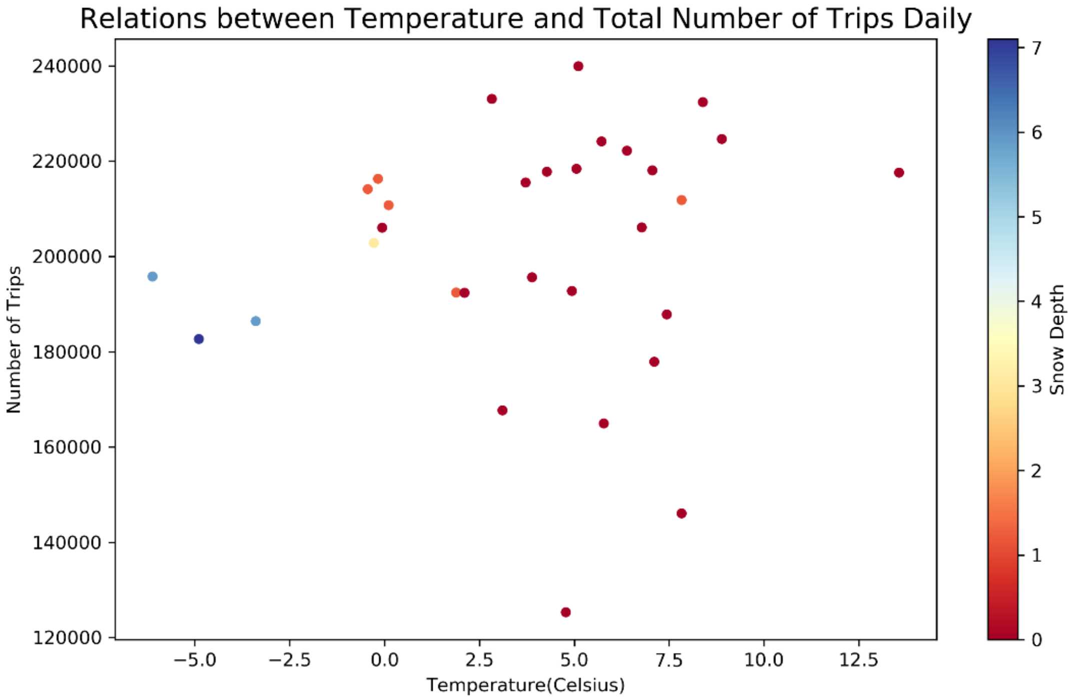


Figure 5

The scatter plot (Figure 5) of total number of trips for all yellow taxis on each day and temperature in January 2017 also shows no obvious relation.

However, the average amount of tips for each trip and the temperature seem to have a positive relation (Figure 6).

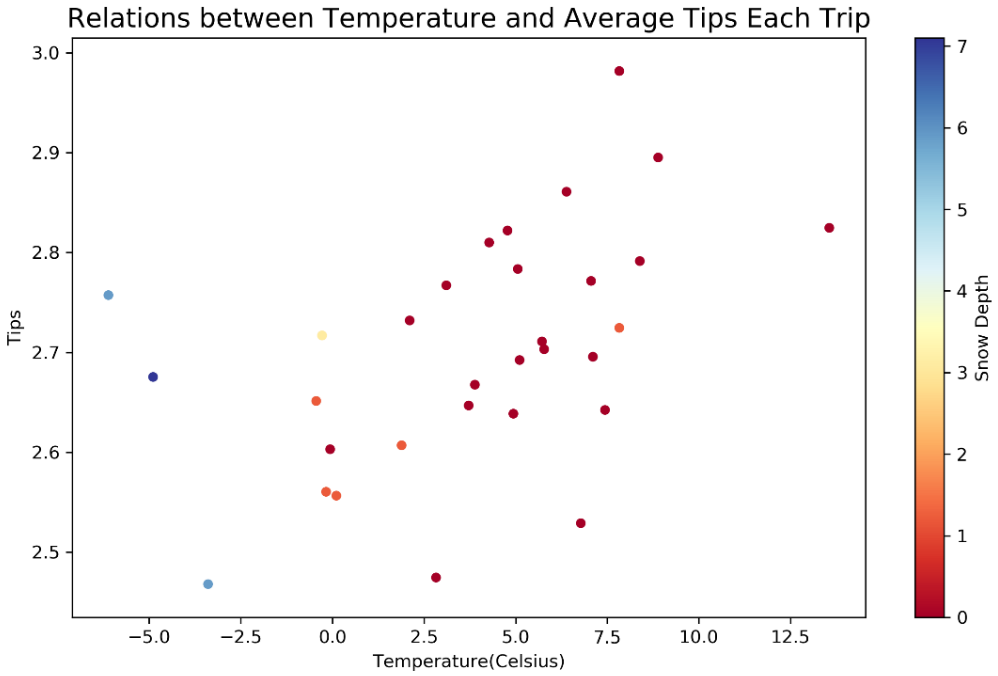


Figure 6

Even though people do not take taxis more often on warmer days, they appear to be more generous with the tips. This could be because people are in a better mood on warmer days in the winter in New York City, and therefore are more generous with tips on those days. When taking snow depth into consideration, it shows that the average tip amount is also higher on snowy days, even though the temperature is lower. It is possible that this is to compensate the inconvenience that snow caused for taxi drivers.

d. The Time Period in a Day and Tips Amount

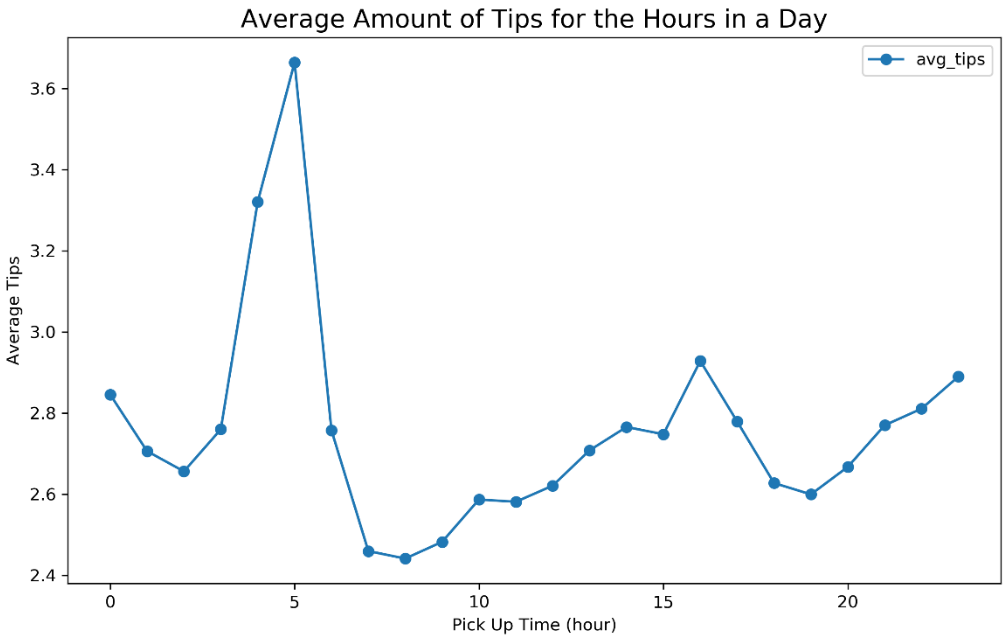


Figure 7

As shown from Figure 7, the peak time of the tip amount yellow taxis received is at 5 a.m. in New York City in January 2017. Also, there is another local maximum around 4 p.m. , which is the beginning of rush hours (4 p.m. to 7 p.m.). Therefore, the yellow taxi drives in New York City should search for potential passengers around 4-6 a.m. and afternoon rush hours to maximize the tips received.

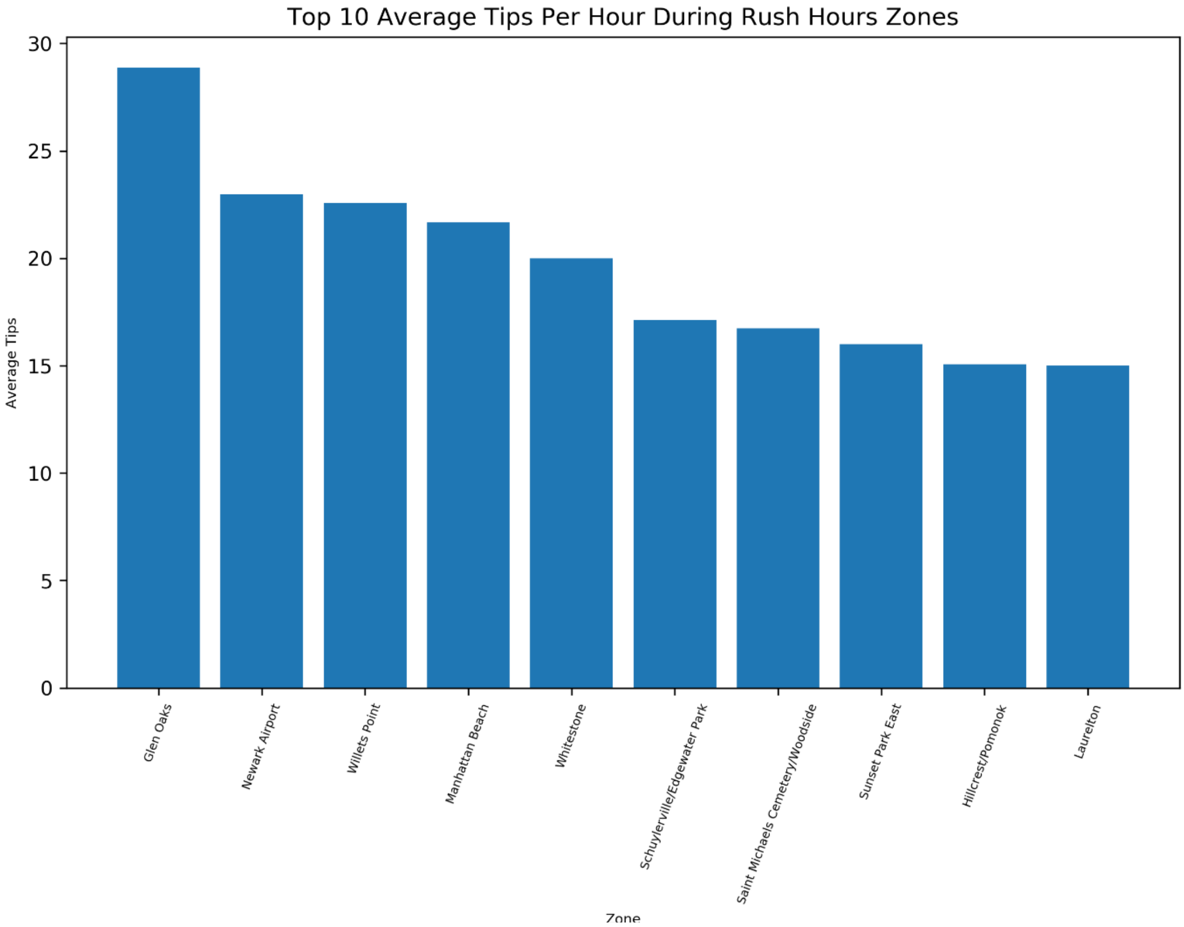
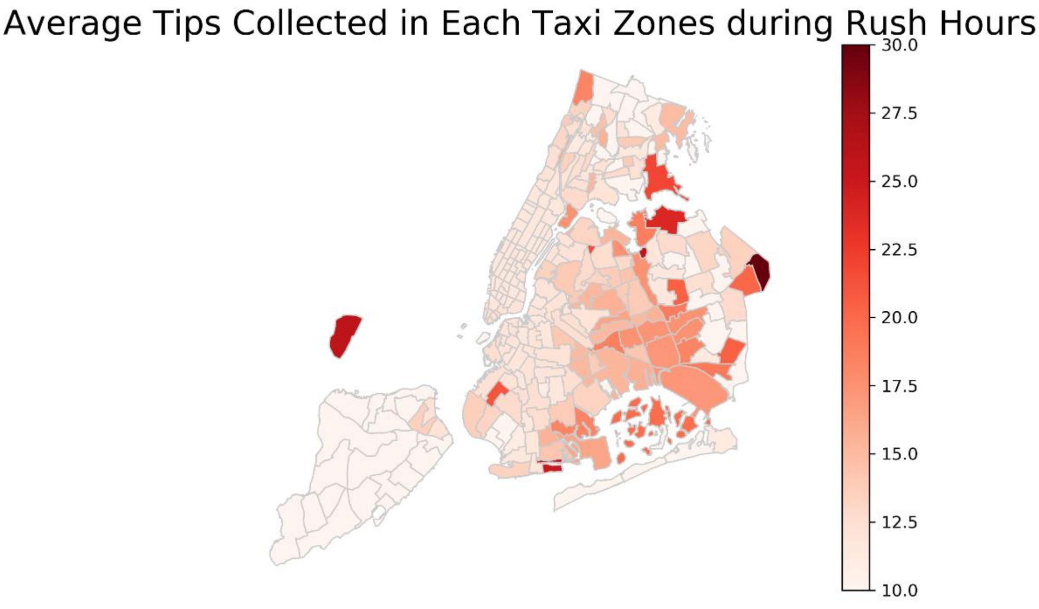


Figure 8

Since the average tips amount during rush hours is a local maximum and the variation is relatively small, this time period is of interest. The histogram (Figure 8) is plotted to show the top 10 taxi zones of pickups where yellow taxi drivers collected the most tips per hour in average during rush hours(4 p.m. to 7 p.m.). These areas could be considered by yellow taxi drivers in New York while picking areas to go for a pickup, in order to maximize profits for the same amount of time.



Source: NYC Taxi & Limousine Commission, 2017

Figure 9

Geographically, Figure 9 is plotted to provide a reference of which areas in New York City can yellow taxi drivers make most tips during the rush hours. It is more practical than a histogram as the drivers can compare the GPS map and this choropleth to decide where to go, considering all factors such as their current locations.

VI. Conclusion

In summary, the amount of tips paid by credit cards for yellow taxi trips in New York City in January 2017 is related to the location, weather and time period of pickups. In general, JFK airport is the best place to go to collect most amount of tips, although the number of pickups is not maximum. In winter, yellow taxi drivers can collect more tips on warmer or snowy days. Timewise, 5 a.m. and rush hours are the most profitable time of a day to pick up passengers. During rush hours, the best ten taxi zones to pick up passengers from in order to have the higher amount of tips are Glen Oaks, Newark Airport, Willets Point, Manhattan Beach, Whitestone, Schuylerville/Edgewater Park, Saint Michaels, Cemetery/Woodside, Sunset Park East, Hillcrest/Pomonok and Laurelton.