MAST30034 Project1

"How to increase the income of yellow taxi drivers"

Yimeng Liu
Student ID: 1074206

August 15, 2021

1 Introduction

The rise of online ride-hailing has inevitably impacted the traditional taxi industry, and every taxi driver is concerned about how they can increase their income in the face of this competition. This project investigates how taxi drivers can improve their incomes. The timeline of this study occurred during the first half of 2019 because the data collected during this time was not affected by the COVID-19 pandemic, which has influenced all walks of life at varying levels [1]. As a daily transportation tool, the half-year analysis was not much different from the whole year analysis, so the half year data proved to be enough for the study. Yellow taxis were chosen as the type of Licensed Taxi studied in this project because they are the most popular taxi in New York City. The target audience of this study were all yellow taxi drivers, as they were the group that cares most about the income.

The raw dataset used was obtained from the TLC taxi website[1] with a data shape of 44,459,136 rows and 18 columns ($\approx 4GB$). Taxi drivers' income might have been affected by pick-up location, tip amount, fare amount, and the time of the day. Weather[4] as an external dataset was included, as weather was considered as a factor that might affect passengers' ability and desire to tip for each trip. The date used in this project was mainly processed by Python.

2 Preprocessing

2.1 Cleaning

2.1.1 NYC TLC Dataset

Cleaning steps for yellow taxi:

- Removed "store and fwd flag" column since it was not relevant to the analysis.
- Removed "extra", "tolls amount" and "congestion surcharge" columns, since these variables were listed under standard fares on the TLC data website[2, 3]. The assumption was made that they had little impact on fare and tip amounts.
- Converted type of "pickup datetime" and "drop-off datetime" to datetime type and restricted the data in the valid period from 2019-01-01 to 2019-06-30.
- Filtered vendorID to 1 and 2 as well as the ratecodeID from 1 to 6, which was specified in the data dictionary[2]. There were some instances with vendorID equaling to 4 and ratecodeID equaling to 99.

- Filtered "fare amount" to be greater and equal to $2.50, as well as "total amount". Filtered "mta tax" and "improvement surcharge" to be equal to $0.50 and $0.30, respectively. It was in line with the TLC data website regarding standard fares[2, 3].
- Filtered "passenger count" in the range of 1 to 6. "Passenger count" greater than 6 was illegal, which was in line with the TLC data website[5].
- Filtered "payment type" to 1, which it is the only payment type that includes tips as specified in the data dictionary[2].
- Removed all the missing values.
- Filtered "trip distance" to be greater than 0 and less than or equal to 50 after examining the map (Figure 1). The assumption was made that the farthest distance is 50 miles.
- Filtered "tip amount" to be greater than 0. As an assumption, passengers who had paid by card must also pay the tip, which is in line with tripsavvy website[6].
- Removed outliers from 'fare amount' and 'tip percent' with 6 IQR after visualising the boxplots. "Tip percent" reflects a measure of tip over the fare amount. It is restricted within 6 IQR, in order to ensure that the "fare amount" and "tip percent" are reasonable and feasible, which would not be impacted by the extreme and irrational values (Figure 2 to 5)

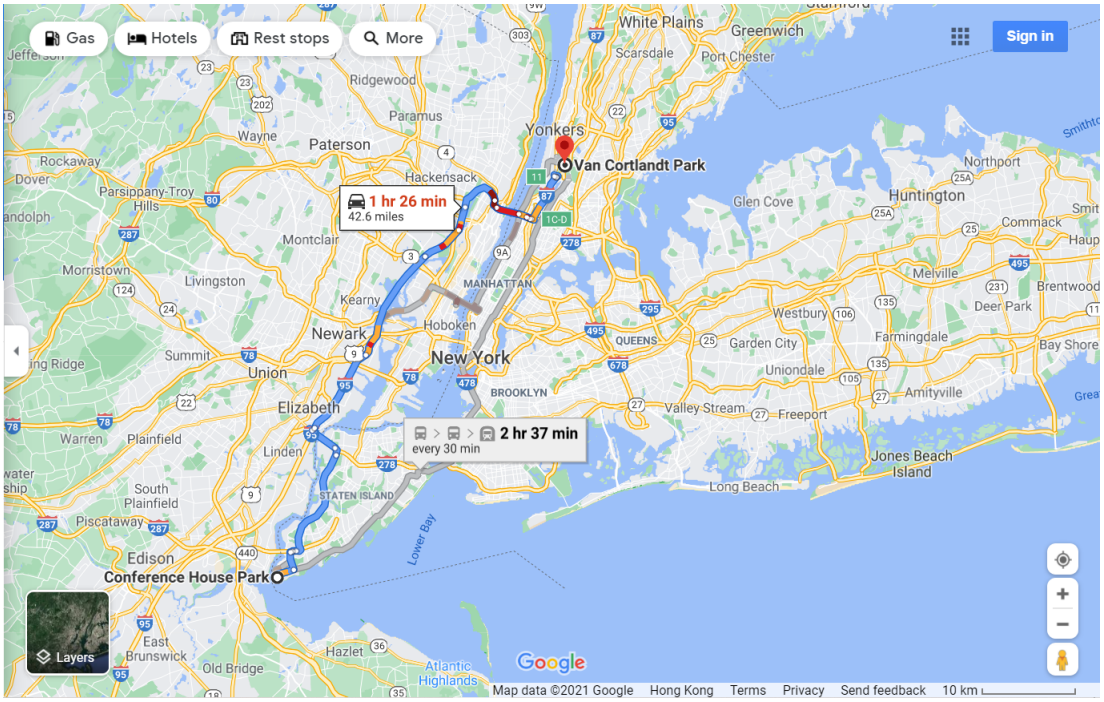

Figure 1: Examining the maximum trip distance

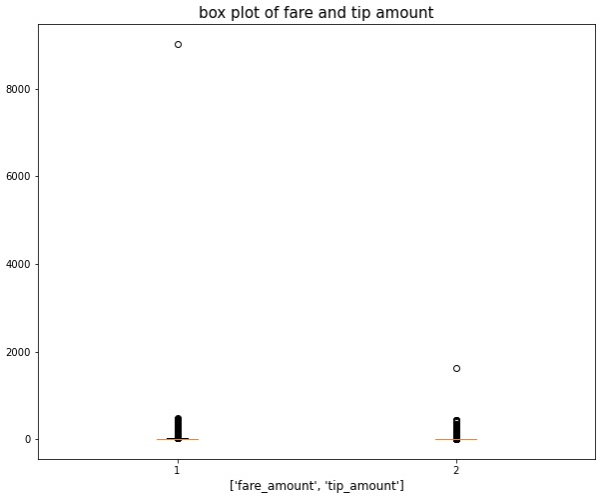

Figure 2: Box plot of fare and tip amount

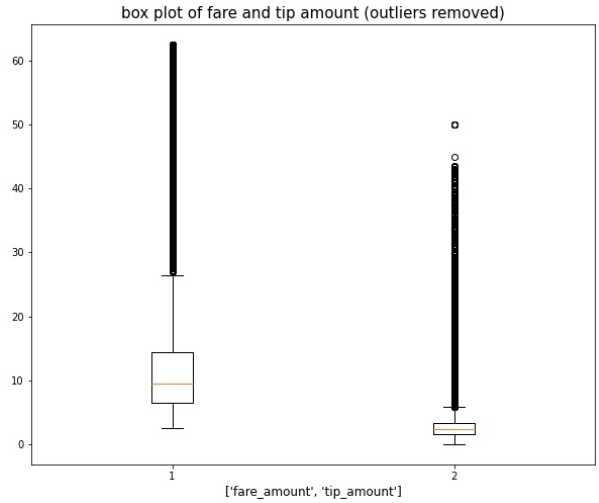

Figure 3: Box plot of fare and tip amount (outliers removed)

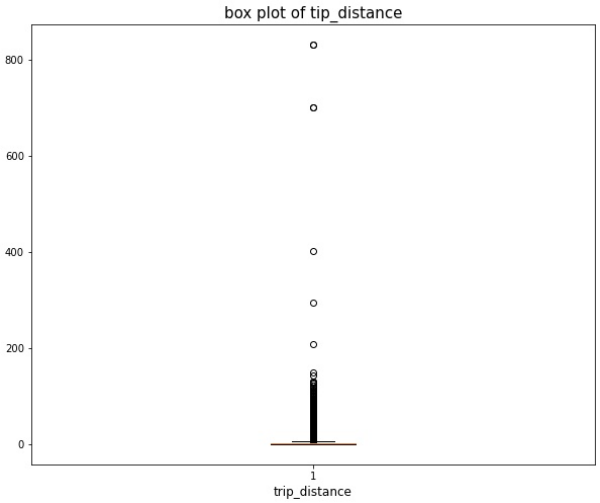

Figure 4: Box plot of trip distance

Figure 5: Box plot of trip distance (outliers removed)

2.1.2 External Dataset 1 - Weather

Cleaning steps for weather dataset[4]:

- Removed "station", "name", "latitude", "longitude", "evaluation", and "TSUN" columns as they were not relevant to the analysis.
- Converted the units of "TMAX" and "TMIN".
- Enumerated the average temperature through "TMAX" and "TMIN".
- Assigned the conditions of weather (e.g., "Bad" and "Good") for each date by aggregating the weather type (e.g., "WT01").
- Removed useless columns, which are as follows "WT01", "WT02", "WT03", "WT04", "WT06", and "WT08".

The cleaned weather dataset of 2019 was saved as a CSV file that can be used for further analysis with cleaned yellow taxi trip data.

2.2 Feature Engineering

Several steps were applied:

- Added a "date" column which was used to extract the exact date of pick-up.
- Added a "duration"(minutes) column. This measured of the time cost for a trip. Filtered valid durations that were greater than 0 minutes.
- Added a "date type" column which was used to classify "workday", "weekend" and "holiday".
- Added a "fare per minute" column to measure of the fare per minute, means how much fare amount can be made per minute for taxi drivers.
- Merged with the cleaned weather dataset.
- Added a "time of day" column, which was represented the time period of 24 hours.

Finally, the cleaned dataset of 2019 yellow taxi data was 67.10% of the raw dataset approximately. It was saved as a CSV file and can be used for further analysis on the topic.

3 Preliminary Analysis

There were several attributes related to the tip amount. By observing the heat map below (Figure 6), it can be concluded that tip amount, trip distance, fare amount and total amount were highly correlated with each other. There was also a slight correlation between tip amount and tip percent. However, fare per minute is not well-correlated with the standard metrics like tip and fare amount.

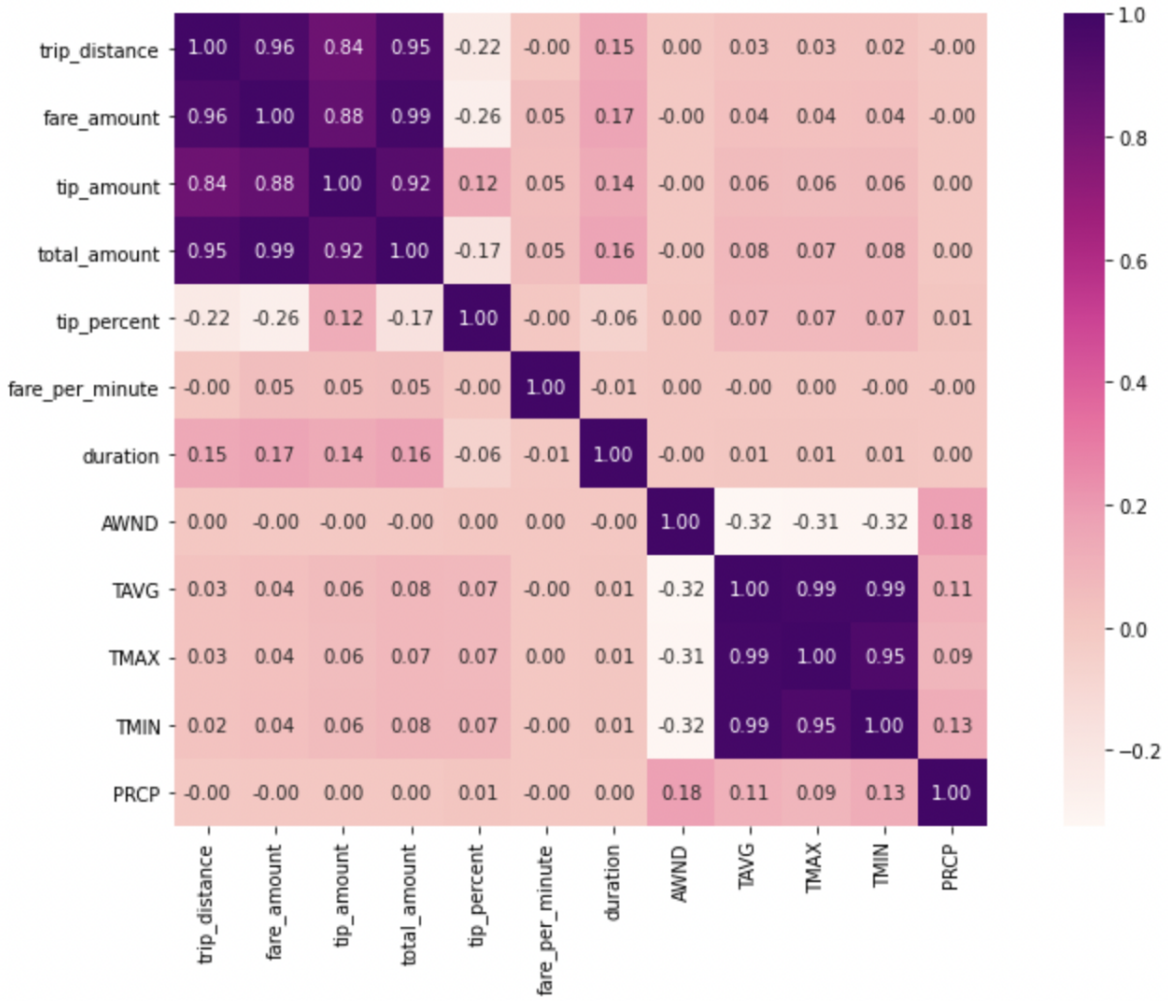

Figure 6: Heat map

3.1 Geospatial Visualisation

The geospatial graph (Figure 7) represents the average tip amount in each zone in NYC and aligns with its corresponding bar plot (Figure 8) by showing the top 15 zones that have the highest average tip amount. The darker color on the map shows areas where higher than average tips occurred. The average tips in most parts of the NYC were between \$0 and \$3. This might be due to the shorter trip distances or shorter durations, leading to fewer tips. On the map, the darkest color appears by Newark Airport (blue marker), which had the highest average tips of \$16. The markers representing three airports in NYC (black: LaGuardia Airport, green: JFK Airport) show higher than average tips: more than \$8. This might because the airports were far from the city, so passengers might have given more of a tip.

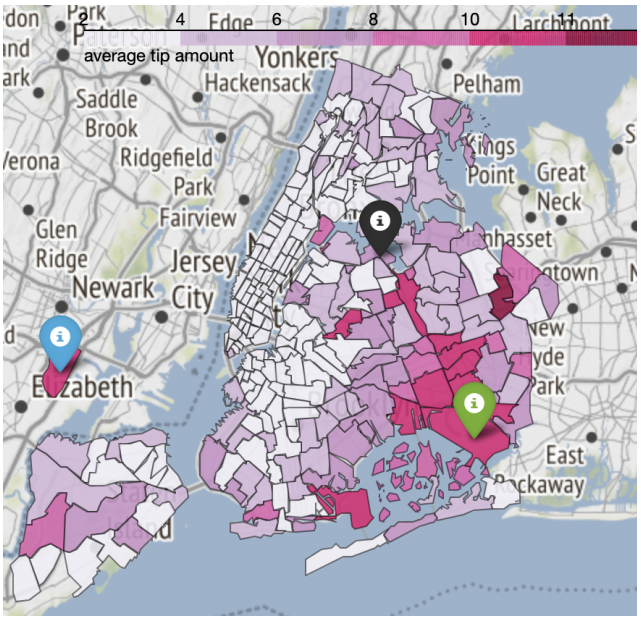

Figure 7: Geo map of average tip amount for each zone

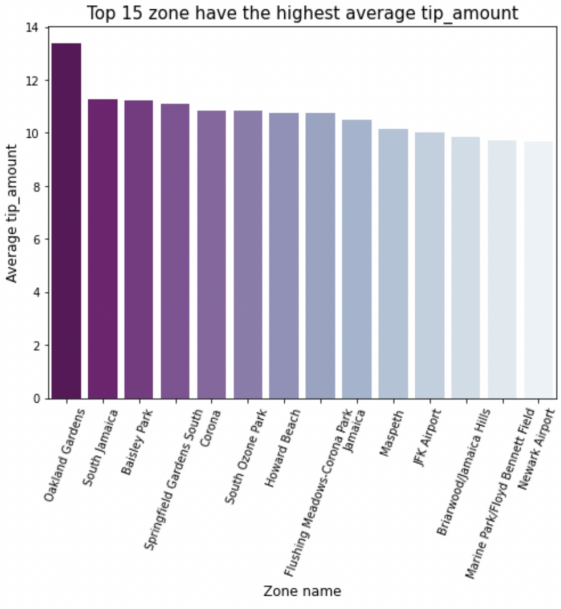

Figure 8: Top 15 zones that have the highest average tip amount

The geospatial graph (Figure 9) represents the total number of transactions for each zone in NYC. It is clear that JFK Airport, LaGuardia Airport, and the bottom part of Manhattan have higher transactions than other zones, reflecting that such zones are busy in case of high demand for taxis. In comparison to the map of the average tip amount, the map indicates that the average trip amount was less in the Manhattan area. Since people may use taxis for short trips in Manhattan that results in a lower tip amount.

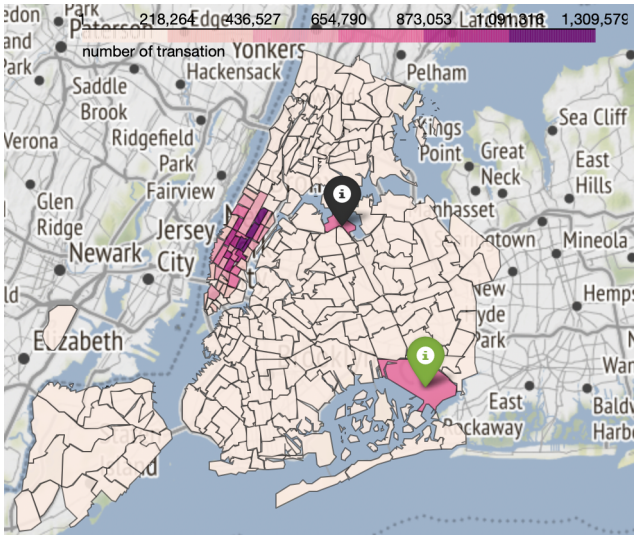

Figure 9: Geo map of total number of transactions for each zone

3.2 Attribute Analysis

Firstly, two bar plots were utilized to observe the relationship between date type, the average number of transactions, and the average fare per minute. Figure 10 makes it apparent that the average number of transactions was lowest during holidays and highest on workdays, however, the average fare per minute do the opposite for holidays and workdays. The reasons might be that people are most likely to travel by car during the holidays and weekends, reducing the number of people taking taxis. Furthermore, since people are travelling out of town for the holidays, the traffic is not too congested, resulting in the fare per minute being higher and more profitable than other dates. Even though most people might go to work by taxi on workdays, the fare per minute would be low due to traffic flow since the traffic flow on workdays is more congested than on weekends and holidays.

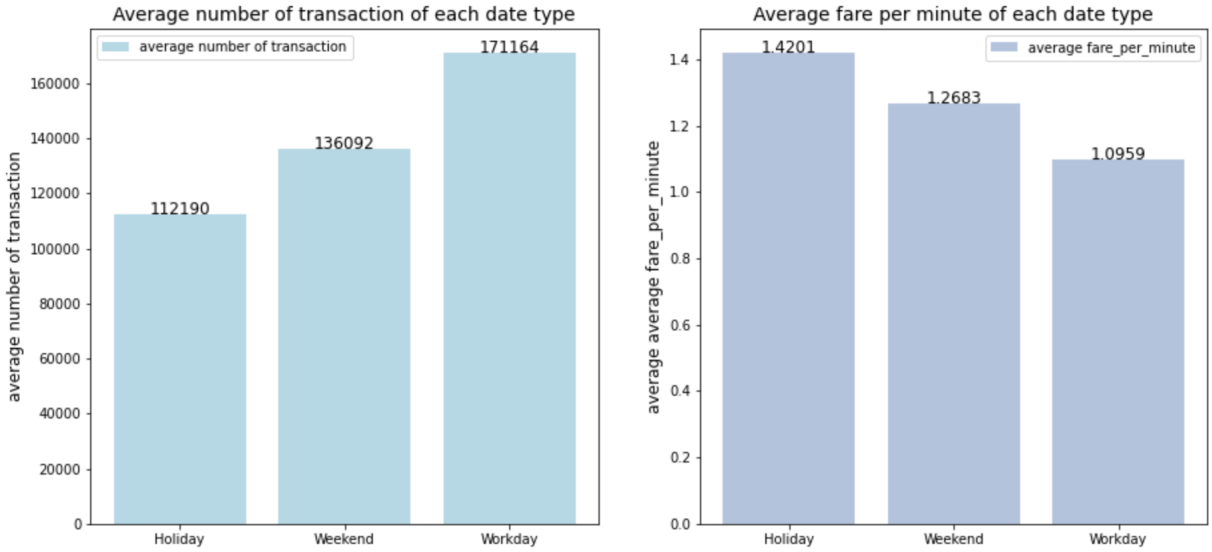

Figure 10: Relationship between date type, the average number of transactions, and the average fare per minute

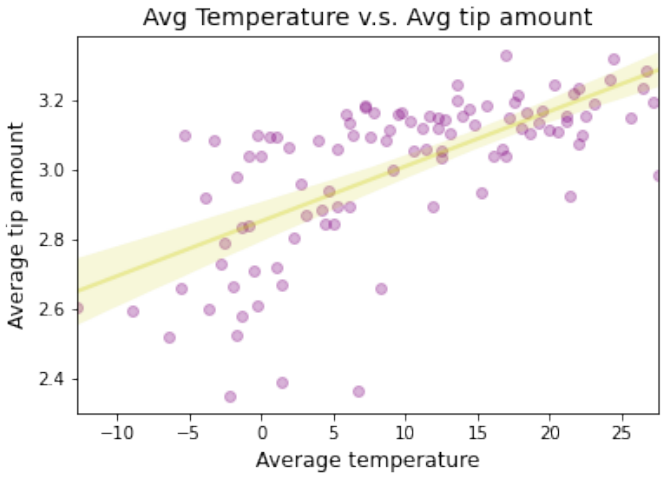

Figure 11: Relationship between average temperature and average tip amount

Moreover, a positive linear relationship can be noted through the scatter plot of average temperature and the average tip amount (Figure 11). The plot illustrates that when the average temperature increases, average tips do as well. In other words, the hotter the temperature, the higher the tip. As a result, taxi drivers drove more passengers in hotter weather, resulting in more income for this greater attendance. Conversely, the colder the temperature, the lower the tip. The plot also points a diminishing return in lower temperatures. People may not want to go out when the weather is cold, staying indoors is obviously more popular. However, if the drivers still maintain high attendance in cold days, it is more difficult to increase income, resulting in less tip.

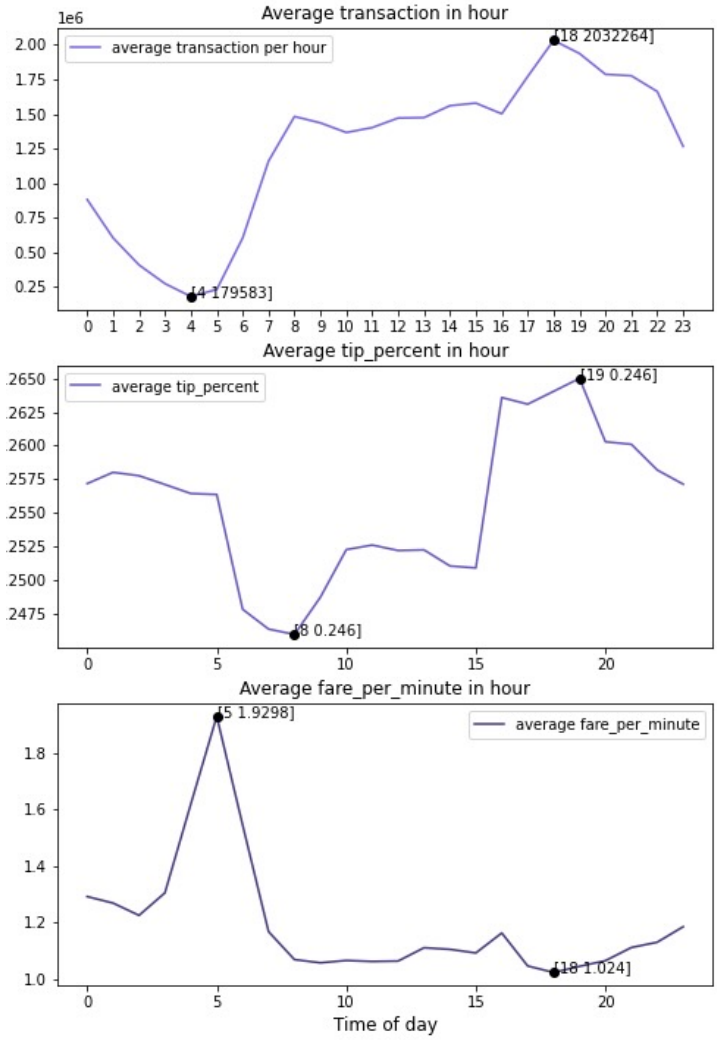

Figure 12: Relationship between average transactions, tip percentage, and fare per minute with time of day

Furthermore, the following three-line plots clearly visualize the relationship between average transactions, tip percentage, and fare per minute regarding the time of day (Figure 12). From the second plot, it is clear that the average tip percentage was lowest from 8 A.M. to 9 A.M. According to the first plot, the number of transactions during this early morning period is high since people taking taxis for short trips on their way to work with a lower fare amount, resulting in a fewer tip amount and tip percentage. Average tip percentage peaked at its highest from 7 P.M. to 8 P.M., demonstrating that

average tip percentage was generally higher at night. This implies that the frequency of people taking taxis at night is high, supported by the result in the first plot. This is probably because people finish working at this time period and more desire having activity and going out at night.

From the perspective of the first and the third plot, the lower engagement from passengers occurred from 4 A.M. to 5 A.M, which is evident in the first plot. Although there are fewer people taking taxis early in the morning, the fare is less, there is no traffic jam yet which means that the fare per minute is more profitable. However, other than the time period from 7 A.M to 4 A.M of the next day, the average fare per minute was generally less than $1.4. This phenomenon shows that the traffic congestion was serious during most of the day and night and the trip distance for passengers of each trip might be short.

4 Statistical Modelling

4.1 Model

Muti-class Logistic regression was used to predict the class of tip amount per trip as it is a classification algorithm. Data were sampled to 1,000,000 rows' worth of information, which was used to ensure the validity and authenticity of the model. Three classes, "low", "medium", and "high", were implemented after viewing the summary of tip amount. Since the number of instances was much larger than the number of attributes, only a few attributes could be used in the prediction about the class of tips, including "date type", "time of day", "TAVG", "SNOW", "SNWD", "total amount", "fare amount", and "trip distance". Therefore, attributes were directly extracted from the sample dataset to form a training set. Moreover, "date type" and "time of day" were converted to dummy variables and treated as quantitative variables. "StandardScaler" was applied to the numerical attributes. The training dataset was split into training and development set in order to build a multi-class logistic model, with tuning the hyperparameters of "max iteration", "multi-class", and "solver". The prediction data came from the dataset[1] for April to June 2020, which was collected in the first half of 2020.

4.2 Results

	Accuracy
Model	0.834
Prediction	0.774

(a) Accuarcy

f1-score	low	medium	high
Model	0.80	0.83	0.90
Prediction	0.80	0.76	0.77

(b) f1-score

Table 1: Results of modelling and prediction

The confusion matrix of the model (Figure 13) shows that the highest proportion of true positives where the model correctly predicted the "medium" class is 40.536%, following by the "low" and "high" classes, which accounted for 27.087% and 15.730%, respectively. The confusion matrix of the 2020 prediction (Figure 14) indicated a similar trend with the model. The proportion of correctly predicting classes of "low", "medium", and "high" are 27.816%, 35.313% and 14.310%, respectively. The accuracy score of model and prediction are 0.834 and 0.774 respectively. Form the classification report of model, the f1-score of the three classes are normally high with respect to 0.80, 0.83, and 0.90. Similarly, the f1-score of prediction's classes of "low", "medium", and "high" are corresponding to 0.80, 0.76 and 0.77 (Table 1).

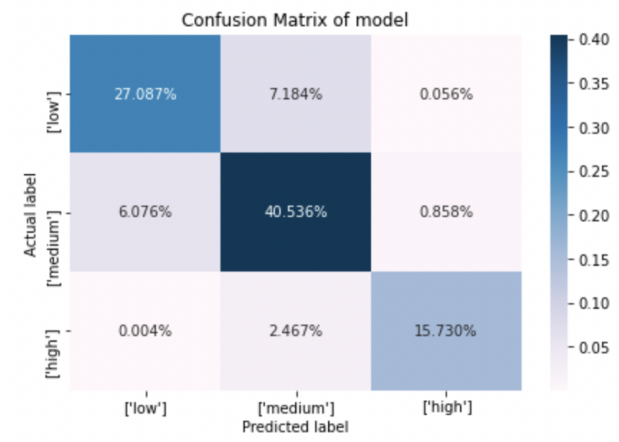

Figure 13: Confusion matrix of model

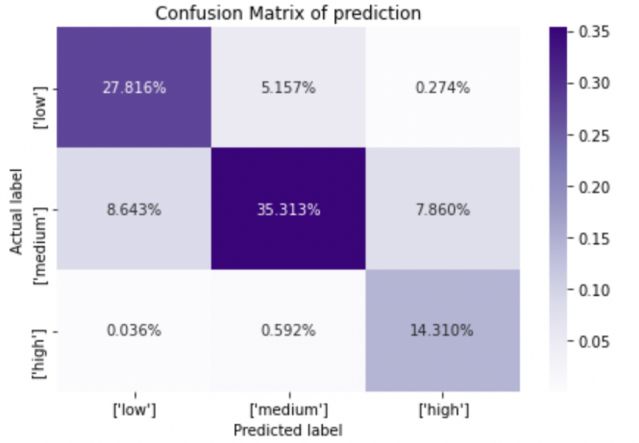

Figure 14: Confusion matrix of prediction

4.3 Discussion

Taking a closer look at the dataset, the number of "medium" classes was the highest, and the number of "high" classes was the lowest, which was highly accorded with the model results. Focusing on the distribution (Figure 15); the distribution is positively skewed. The data points are collected within 10 dollars. The classes of tip amounts were simply divided by an equal width of 2 with no upper limit after observing the mean, median, first quartile, and third quartile of the data. To improve the model, an attempt would be made to divide the classes equally by the equal-frequency or knn method. Moreover, for the "high" class, some data points would be considered as outliers.

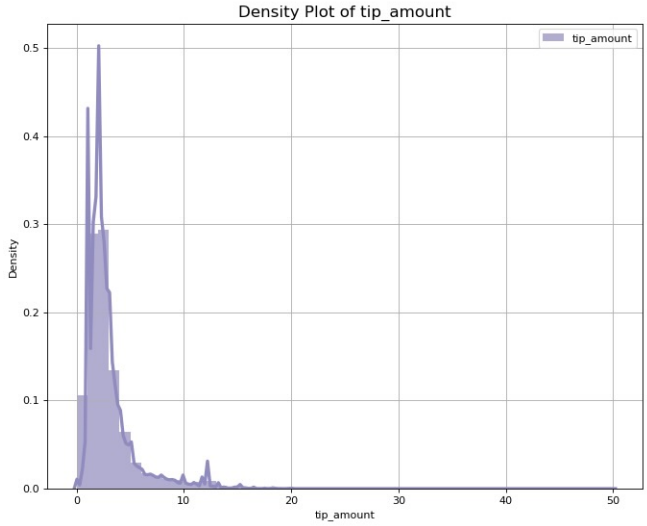

Figure 15: Distribution of tip amount

The accuracy of the 2020 prediction was slightly lower than that of the model. The dataset used to build the model did not affect covid-19. Therefore, covid-19 could be considered to impact the tip amount and the number of taxi transactions in 2020. Overall, the model performed well due to its high accuracy and f1-score for each class.

5 Recommendations

After discovering and investigating the relationship between attributes and the logistic regression, some recommendations can be proposed to help taxi drivers improve their income.

- Drives can work on the zones, where the highest tips are being paid. Zones were highlighted on the map and the figure (Figure 7 and Figure 8).
- Drivers can spend more time driving during hotter days, since there is a higher chance of obtaining more tips which is evident in Figure 10.
- Drivers can increase their work hours at night, specifically from around 7 P.M. to 12 A.M. This is because the demand for taxi is relatively greater during this time frame than others (Figure 12).

6 Conclusion

In conclusion, this project aimed to investigate the topic "How to improve the income of yellow taxi drivers". Through observing the relationships and interactions between the variables and building a logistics regression model exploring the tip amount class. The results suggested that the most relevant attributes affecting the amount of income were tip amount, time of day, temperature, pick-up location and fare per minute. This project's results are summarized and included several recommendations for improvement the income of yellow taxi drivers.

References

[1] "TLC Trip Record Data." TLC Trip Record Data - TLC. Accessed 2021.
https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

[2] "Data Dictionary - Yellow Taxi Trip Records" - TLC. Accessed 1 May, 2018.
https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf

[3] "Taxi Fare." Taxi Fare - TLC. Accessed September 7, 2019.
https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page

[4] "Daily Summaries Station Details" - NOAA. Unknown.
https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00094728/detail

[5] "Passenger Frequently Asked Questions." - TLC. Accessed 2021.
https://www1.nyc.gov/site/tlc/passengers/passenger-frequently-asked-questions.page

[6] Heather Cross: "A Guide to Tipping in New York City" - tripsavvy. Accessed 19 May 2020.
https://www.tripsavvy.com/guide-to-tipping-in-new-york-city-4177115