Analysis of yellow taxi demand and prediction of tip amount of a yellow taxi trip in New York

Yuhao Li
Student ID: 1054617

August 15, 2021

1 Introduction

1.1 Background and motivation

New York is one of the most popular destinations for tourists around the world, which had 66.6 million tourists in 2018[1]. Among all twelve months in a year, June, July, and August are regarded as the peak season for tourism. Moreover, taxis are widely used for the majority part of tourists that visited New York, which leads to the biggest concern of taxi drivers that how to earn more during peak season.

The objective of this project is to obtain a better strategy for yellow taxi drivers to find the trip that maximizes tip amount by using a linear regression model to estimate the expected tip amount of a taxi trip and obtain the most profitable time, whether to pick passengers during limited peak season. Moreover, the potential differences in taxi demand among numerous factors would also be analyzed.

1.2 Assumption

Some of the assumption includes:

- Assume passenger would not provide additional fee without recording
- Assume each instances in dataset are independent to each other
- Assume there is no seasonality effect to affect the demand of yellow taxi
- Assume there is no extra cost during working hour
- Assume passenger count won't significantly impact driver's preference
- Assume maximum passenger count in a trip is 5
- Assume toll fee won't affect the passenger's preference for paying tips

1.3 Dataset Shape

The original dataset has around 19 million instances with 18 columns. After filtering and combination with the external dataset, the approximate dataset shape is described as 5.5 roughly million rows with 12 columns for pickup date between 2019-06 and 2019-07, and around 2 million rows with 12 columns for pickup date in 2019-08

2 Preprocessing

2.1 Dataset selection

2.1.1 Yellow Taxi dataset from 2019-06 to 2019-08

The dataset is chosen for yellow taxis from June 2019 to August 2019 due to integrity of seasonality, the exclusion for the significant decreasing impact to the number of taxi trips due to COVID-19 in 2020, informative attributes such as "fare-amount" and sufficient data size.

2.1.2 NYC weather dataset[2]

This dataset is collected from Weather Underground website [2], which recorded the weather conditions in New York between 2019-06-01 to 2019-08-31. At the 51st minute of each hour, one weather condition such as "Mostly cloudy", "Light rain" etc. would be recorded. Instead, to simplify various weather scenarios, three types of weather such as "Sunny", "Rainy" and "Other" are used. "Sunny" indicates current weather condition is fair and cloudy. Rainy refers to there is currently a drizzle or light rain. "Other" refers to the weather that does not belong to "sunny" or "rainy" such as 'T-storm'. The main purpose of this dataset is to generate weather features with each taxi record.

2.2 Preprocessing steps

Preprocessing would be mainly divided up into 6 steps. Firstly, dropping all instances with missing values. Secondly, select instances with valid column values for outlier removal. After that, some new features would be appended to the dataset. Furthermore, redundant features would be dropped. Finally, combine the weather dataset and taxi dataset to add weather features. Furthermore, the three-month dataset is separated to 2019-06 and 2019-07 as training set and 2019-08 as a test set while maintaining the same preprocessing procedures.

2.3 Outliers investigation and removal

2.3.1 Discrete feature

- Vendor ID is set to be only equal to 1 or 2 as specified from data dictionary
- Both pickup_datetime and droppff_datetime are set to be strictly in the range between 2019-06-01 and 2019-08-31 for integrity.
- Number of passengers should be greater than 0 and less than or equal to 5 [3]
- Credit card payment would be considered only since only credit card tips would be recorded.
- Valid LocationID should be in the range between 1 and 263[4]
- As specified in data dictionary, a valid Store_and_fwd_flag is either 'Y' or 'N'
- As specified in data dictionary, improvement surcharge would be 0.3 as long as the trip start.
- As specified in data dictionary, mta_tax should be 0.5 if the trip start
- Congestion surcharge should be either 0 USD or 2.5 USD[5]
- Filtered valid extra fee to be 0 USD, 0.5 USD or 1 USD as specified in data dictionary

2.3.2 Continuous attribute

Following table display the valid range of continuous attributes respectively, which are based on boxplots of continuous attributes(See notebook 1 for boxplot) to avoid extremely skewed distribution. Also, some of the upper bound are set based on real-life experience.

Table 1: Outlier removal summary for continous features

Attribute name	range	Attribute name	range
trip-distance(mile)	(0,20]	Fare amount(USD)	[2.5,100]
tolls amount(USD)	[0,60]	tip amount(USD)	[0,10]
total amount(USD)	[2.5,100]	duration(hour)	[1/60,5]

2.4 Feature engineering

Table 2: Feature engineering attribute summary

Attribute name	Description
Duration	dropoff-datetime minus pickupdatetime (hour)
peak-time	Equals 1 if pickup time is between 7 to 11 am or between 4 to 7pm otherwise equals 0
weekday	a categorical variable equals 1 if pickup date is a weekday, otherwise, it equals to 0
weather	a categorical variable indicates 3 weather condition such as sunny, rainy, other
start-date	indicate the pickup date of a trip
hour	indicate the hour of a pickup time from 0 to 23

2.5 Feature selection

Our research goal is to predict expected tips of that trip, which means following features might be irrelevant to analysis

- Since "VendorID" only indicate the identity of provider, which won't affect the amount of tips that passenger pay.
- Since a pickup date time and locationID is more indicative for taxi driver to know where to pickup, which makes dropoff time and locationID less important. Moreover, pick-up date time could be replaced by "start-day" and "hour" generated in feature engineering part.
- "Store-an-fwd-flag" only refers to trip record are held in memory.
- Since all payment type, RateCode ID are set to be only use credit card with standard rate so "payment type" and "RatecodeID" column is useless
- By assumption, we assume passenger count won't affect driver's preference for pickup the passenger
- "Extra", "mta-tax", "improvement-surcharge","congestion surcharge" have fixed set of value and is irrelevant with tip amount
- Toll fee won't affect the passenger's preference for paying tips significantly as specified in assumption.
- Since tip amount instead of total amount is the primary goal, so "total-amount" is useless

3 Preliminary Analysis

3.1 Single attribute analysis

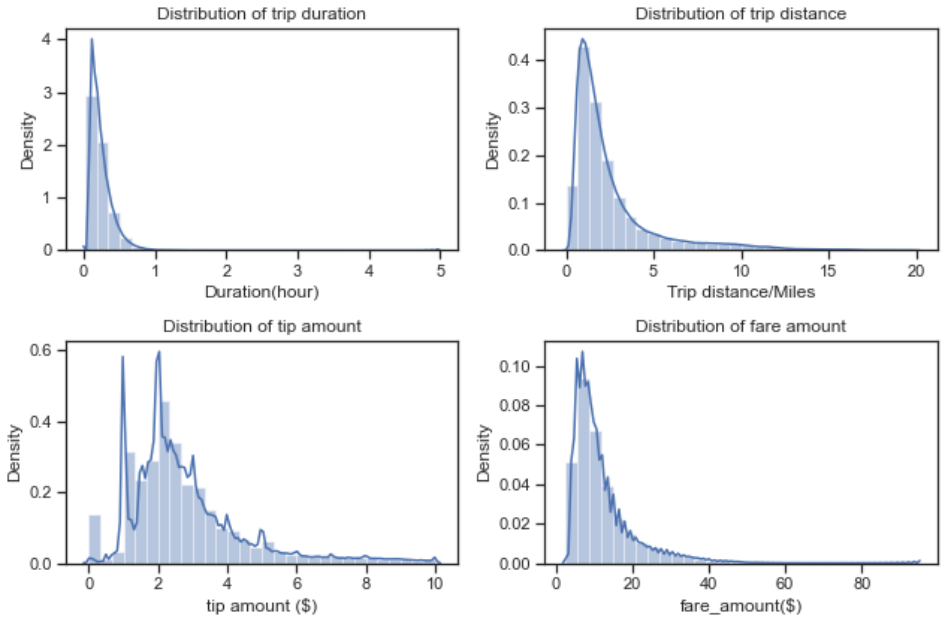

Figure 1: Distribution of 4 continuous attribute and proportion of 3 categorical attributes

Figure 1 illustrates **a strictly positively skewed distribution** away from the center that could be seen from all of the four plots after removing a substantial number of outliers.
As shown from the first plot that describes trip duration, most of the trip duration is between 0 and 1 hour while the majority part of the trip distance is less than 10 miles. **This behavior indicates short-range taxi trip is preferred by most passengers**. Apart from this, a major proportion of the tip amount is between 0 and 6 USD, while there is a major occupation for fare amount in the range from 0 USD to 40USD

3.2 Attribute relationship analysis

Two aspects of interest would be considered, which are demand of yellow taxi and tip amount.

3.2.1 Demand of taxi trips

Firstly, from the upper right of figure 2 describes the impact of weather, it displays that a yellow taxi trip is slightly preferred on a rainy day rather than the other 2 types of weather conditions, with an average demand achieve 4000 trips per hour. Secondly, the lower right plot displays the fluctuation of taxi demand between 2019-06-01 and 2019-7-31, weekend demand is highlighted by green area, where a huge decrease of the number of taxi trips was a weekend for most of the time.
From the left part of figure 2, demand for yellow taxis plummeted at 6 am and reached the peak at 7 PM, which indicates the peak time might lead to higher demand.
From the visualization, it could be inferred that **demand of taxi is positively related to rainy weather, the peak time of a day, and weekday of the week**.
It is also interesting to note the difference between weather conditions is small, which is a highly likely result of an extremely high proportion of sunny weather hours.

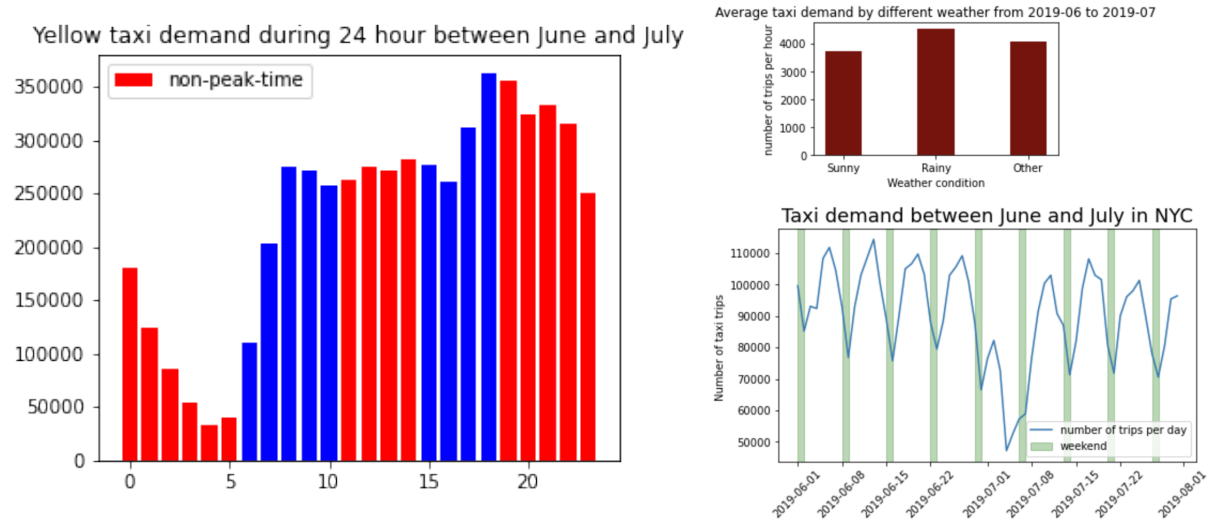

Figure 2: Demand that affected by weather, weekday

3.2.2 Tip amount of a taxi trip

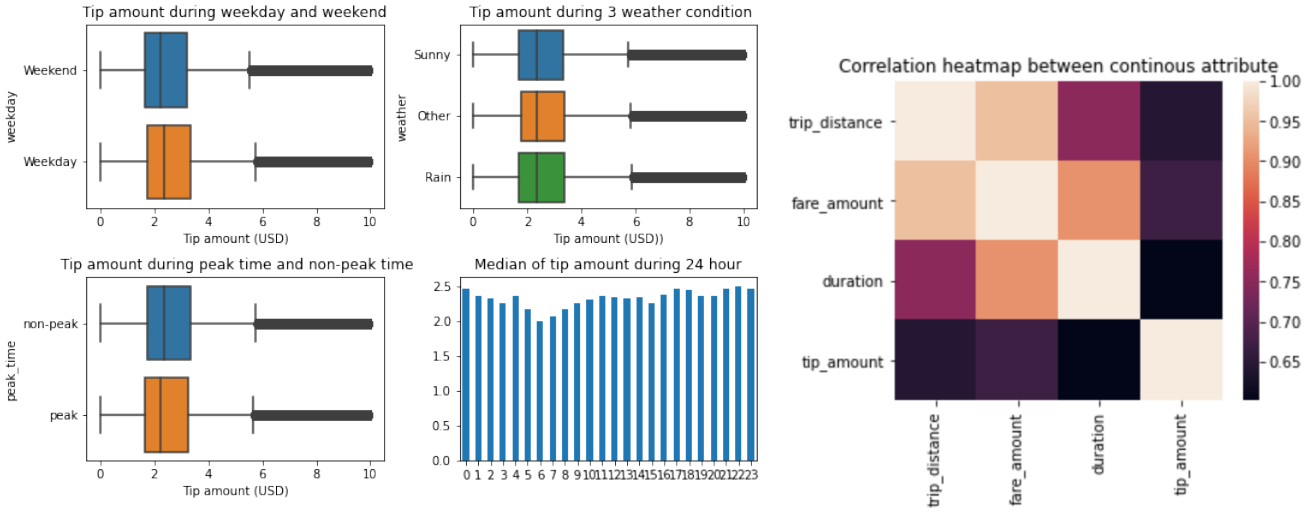

Figure 3: Plots for tip amount caused by discrete factors and correlation between continuous attributes

As shown from figure 3, 4 plots are describing for distribution of tip amount among all factors, tip amount at the weekend, peak time, and "other" weather condition approximately has a higher median. From the last plot, the effect of weather is relatively small.
From the bar part of figure 3, the average tip amount reaches a peak at 6 am, while remaining steady and lowest point between 9 am to 5 pm. Also from the right part, a positive relationship between tip amount and trip distance could be observed. Therefore, there might be **a linear and positive relationship between trip distance and tip amount,**
Surprisingly, the median of tip amount is almost insensitive to any factors visualized above, detailed analysis to confirm if those factors are significant would be further discussed below.

3.3 Geospatial Visualisation

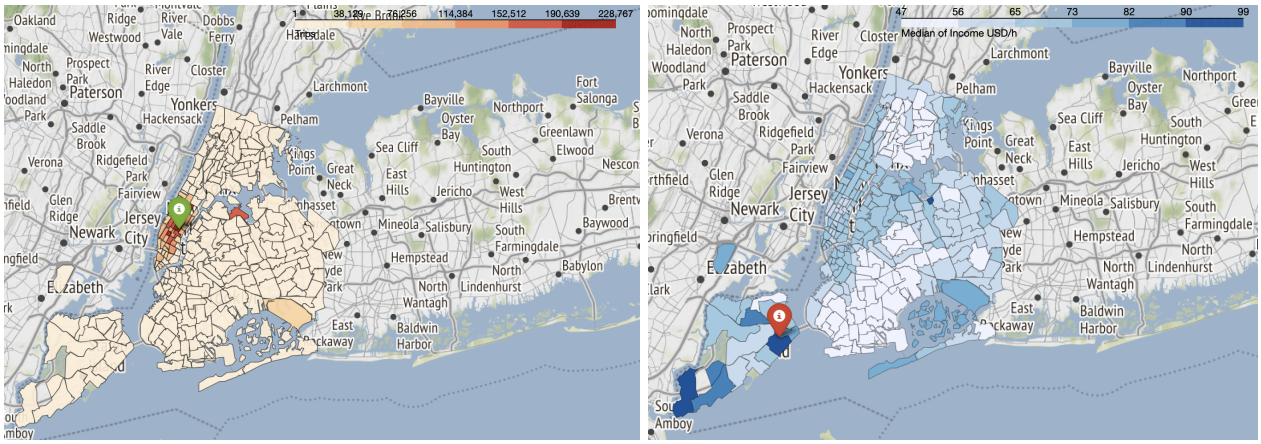

Figure 4: Geospatial Visualization of demand and median of tip amount

From the left part of Figure 4, **Manhattan area shared the highest trip frequency, while yellow taxis are less required for other districts** From the right part of Figure 4, **the highest median for salary per hour at Staten Island**.
It is unsurprising to notice that Manhattan is the busiest place in New York as it is the most crucial financial center and is around with many tourist attractions. Also, this phenomenon might be because of restrictions to the green taxi that could only pick up passengers outside Manhattan.
By contrast, it is surprising that median of tips amount is highest outside Manhattan. This might be due to far distance away from New York centre, which might cause larger trip distance of a trip that indirectly cause the increase of tip amount.

4 Statistical Modelling

Since correlation doesn't imply causation, a statistical model might be required to show the actual relationship between attributes of interest and factors. Therefore, a linear regression model to predict the expected tip amount of a driver would be produced next. The choice of linear regression is mainly because of evident quantitative interpretation of the relationship between factors and attributes of interest. Moreover, there might be a linear relationship between tip amount and other continuous predictors by correlation heatmap in figure 3.

4.1 Model specification

Response variable is set as tip-amount, and predictors involve, trip distance, weather condition, weekday, and peak-time. The interaction effect would also be tested statistically.

4.2 Model refinement

1: "start-date","hour","fare-amount","trip duration" are dropped to simplify model due to irrelevance,and dependency with attribute "peak-time", and "fare-amount","trip duration" would not be known until the journey stop.
2: Step-wise selection procedure would be used for this model, since the size of predictors is large, especially for the interaction model. Bayesian Information Criterion (BIC) is implemented to further

reduce the size of predictors. 3: Numerous evaluation metrics such as mean absolute error(MAE), R-square, and mean squared error(MSE) for interaction model and non-interaction model would be compared and discussed

4.3 Assumption check

By observation of diagnostic plot in R notebook, model fits well with linear model assumption, except for the third plot that shows a large heteroskedasticity that might lead to imprecise parameter estimation.
Moreover, tip amount is assumed to be normally distributed despite a little positive skewness.

4.4 Results

Firstly,interaction model is significant at 5% significance level, which suggest interaction model should be chosen.
Summary of the Analysis of variance(ANOVA) model fit is included in R notebook.

Table 3: Performance for different models

Evaluation metric	Interaction model	No interaction model
R square	0.4183022	0.4188917
MSE	1.575275	1.578386
MAE	0.8377811	0.8381928

4.5 Discussion

The summary table shows interaction model is superior to the non-interaction model from all metrics. Nevertheless, the result is worse than expected, which could be mainly due to the 1: heteroskedasticity of the model. 2: Imprecise assumption of normal distribution for tip amount.
In general, by model summary, at 5% significance level, all factors have an effect on tip amount, except for the sunny weather.
Finally, 1-mile increase of trip distance could lead to $0.438 increase of tip amount, holding all other variables fixed, suggesting **larger trip distance could lead to larger tip amount**

5 Recommendations

- It is recommended that taxi drivers could have the highest probability of picking up more passengers by driving at a weekday, rainy day, and peak time in the Manhattan area as described from statistical modeling result and upper plot of figure 5.
- It is recommended that taxi driver could drive to "Broad Channel zone" to earn the most amount of tips with a median of tip amount at 10 USD as described from the lower plot of figure 5
- It is recommended that taxi drivers should pickup more trips with long trip distances. Since my model summary, a 1-mile increase of trip distance could lead to a 0.438 USD increase of tip amount holding all other variables fixed, suggesting **larger trip distance could lead to larger tip amount**.

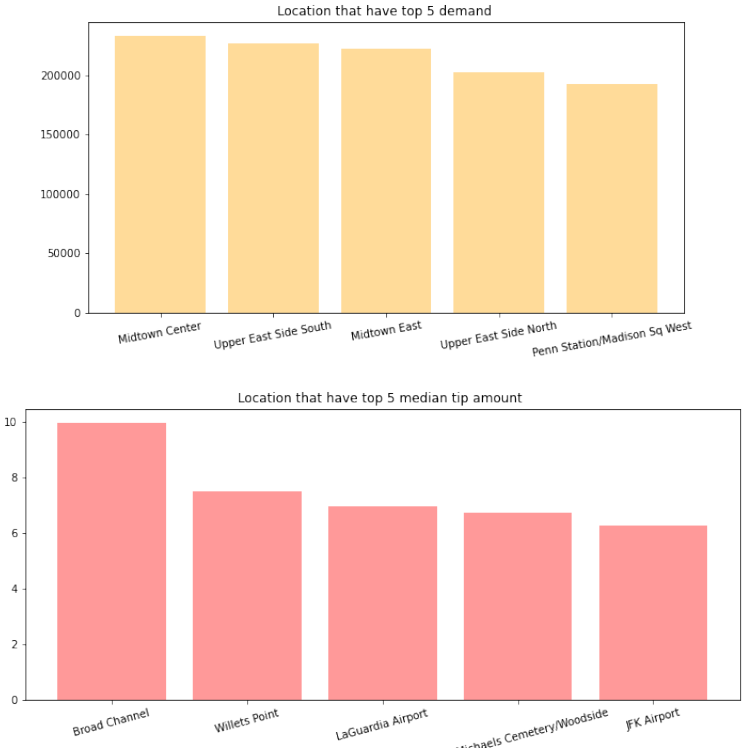

Figure 5: Top 5 busiest district and pay highest tip

6 Conclusion

In this task, two aspects of increasing revenue of taxi driver are considered, namely demand of taxi at time and tip amount from passengers. Even though a comprehensive analysis has been offered to yellow taxi drivers, improvement still exists. For example, trips from certain locations might have high tip amounts but less demand. It is also suggested that next year's data could be tested to attain a more precise result.

References

[1] “The Tourism Industry in New York City” Tourism Industry - Office of the New York . Accessed August 1, 2021.
https://www.osc.state.ny.us/reports/osdc/tourism-industry-new-york-city

[2] ”New York weather daily history ” Weather History - Weather underground. Accessed August 4, 2021.
https://www.wunderground.com/history/daily/us/ny/new-york- city/KLGA

[3] ”Passenger Frequently Asked Questions ” Passenger FAQ - TLC. Accessed August 2 2021.
https://www1.nyc.gov/site/tlc/passengers/passenger-frequently-asked-questions.page

[4] ”NYC Taxi Zones” Taxi Zones- Data.gov. Accessed August 2 2021
https://catalog.data.gov/dataset/nyc-taxi-zones

[5] ”NYC Taxi Fare” Taxi Fare - TLC. Accessed August 2 2021.
https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page