Tip Behaviour Analysis for Yellow Taxis in NYC

Arshnoor Kaur
Student ID: 1170894
Github repo with commit

August 21, 2022

1 Introduction

Tipping has become integral part of the New York City etiquette and culture. New Yorker citizens convey their satisfaction or dissatisfaction of the service they received by tipping appropriately. Often service industry workers including drivers and servers rely on tips as an essential section of their earnings to manage the higher cost of living in New York [1].

Thus, this report aims to analyse tip behaviour for Yellow Taxi vehicles in New York City in order to provide valuable information to Yellow Taxi drivers on how they can improve their tip profits by using analysing various features and conducting linear regression and decision tree modelling.

1.1 Data

The Taxi and Limousine Commission publishes data containing important travel statistics to enable the analysis of key trends in the transport industry. Yellow Taxi vehicles contribute to the majority of taxi vehicles in New York City with 13,587 licensed vehicles [2] and are the only taxi vehicles permitted to accept street-hails in any area of New York City. Because of this, yellow taxis are used most frequently for trips and hence, they were selected as the taxi type to investigate.

Following 2020, the COVID-19 pandemic resulted in a drastic change in everyday lifestyles and travel routines. In order to analyse tipping behaviour without the confounding influence of COVID-19, yellow taxi data was chosen from October 2018 and March 2019.

Weather has been known to affect our moods and therefore, can often influence our actions. As a result, weather may be associated with tipping behaviour and thus, weather data from John F. Kennedy International Airport JFK, Queens, NYC published by National Centers for Environmental Information [3] was included. This location was used for weather statistics because it is located near the centre of NYC. The timeline of the weather data was matched to the yellow taxi dataset from October 2018 to March 2019. This range also allows for the analysis of varying weather conditions.

The attributes selection from each dataset are outlined in 2.1.2 and 2.2.1.

2 Pre-processing

Both the yellow taxi and weather dataset were well organised and had little inconsistencies. The steps taken to curate necessary data and remove outliers are stated below.

2.1 Yellow Taxi Dataset

Before any pre-processing steps were implemented, the total number of records was found to equal 47,798,251.

2.1.1 Outlier detection

Outliers were identified by visual techniques, analysing the data and comparing against documentation provided by TLC [2].Through visual analysis as shown in Figure 1, it was found that the taxi data contained records which did not match the values specified in the relevant data dictionary. Thus, following outliers were removed:

- Passenger counts less than or equal to 0.
- Tip amounts less than 0.
- Abnormal trip distances - greater than 246.9 miles or less than or equal to 0.
- Vendor ID not equal to 1 or 2.
- Pick-up and drop-off locations not between 1 and 263.

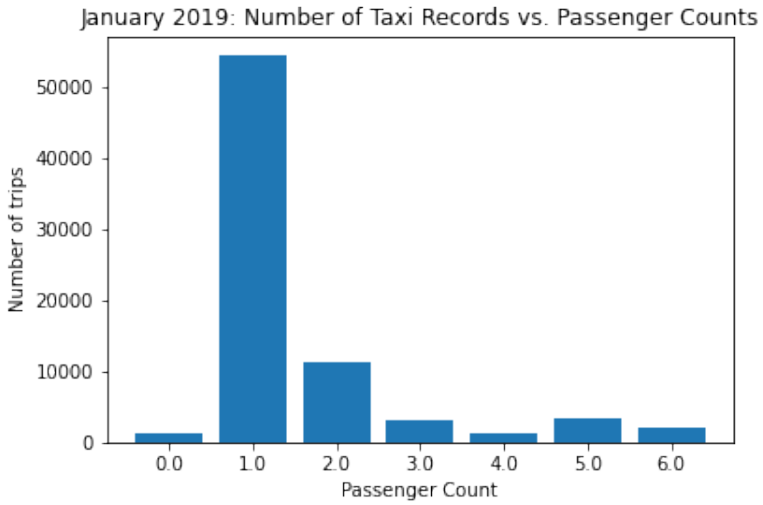

Figure 1: Outlier visualisation where passengers count = 0

2.1.2 Feature Selection and Engineering

Attributes that did could not be associated with tip amounts were filtered out (e.g. VendorID). Retaining too many attributes can result in over-fitting of the model which can limit generalisation abilities. Ultimately, the following attributes were considered to have the greatest impact on tip amounts.

- Records with credit card payments as tip data is only available for this payment type.
- Pick-up and drop-off location IDs.
- Trip distance.
- Date and time.

- Day of the week (added).
- Passenger count.
- Total amount of the trip.
- Rate code IDs.

The total number of records after pre-processing was 32,367,337.

2.2 Weather Dataset

The total number of records before pre-processing was 6695.

2.2.1 Feature Selection and Engineering

1. Missing values were removed.
2. Daily weather statistics were created.
3. Average temperature, wind speed and dew point were selected for analysis as they were expected to influence tip amounts most significantly.

The total number of records after pre-processing was 182.

2.3 Imputation

Following the above steps, imputation of null values was not required as either these values were removed or not selected to be studied. Considering the amount of data available for analysis, it was practical to discard these records.

3 Analysis

3.1 Preliminary Analysis

Initial tip amount distribution analysis (Figure 2) indicates that tip amounts follow a normal distribution centered around $2.5 USD and display a right skew with values ranging between $0-$16 USD.

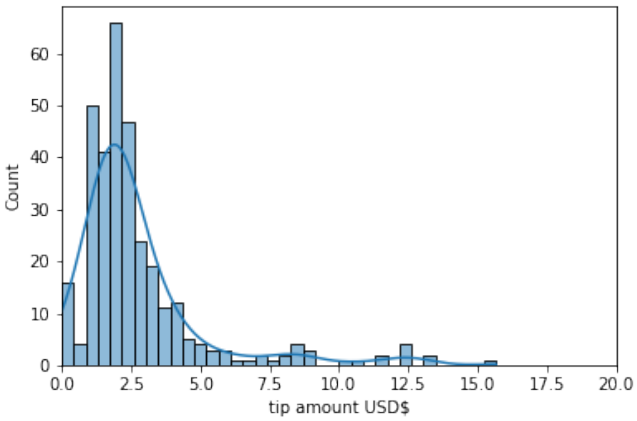

Figure 2: Distribution of tip amount from sampled taxi data

3.2 Analyse the effect of month and day of week on average tips

Surprisingly, the month and day of week had minimal influence on the average tips earned (Figures 3 and 4). This indicates that while holidays periods (Christmas in December) and weekends might have more tourists and therefore more trips, the average tips do not increase. Hence, it was established that these features were not going to prove valuable in predicting tip amounts.

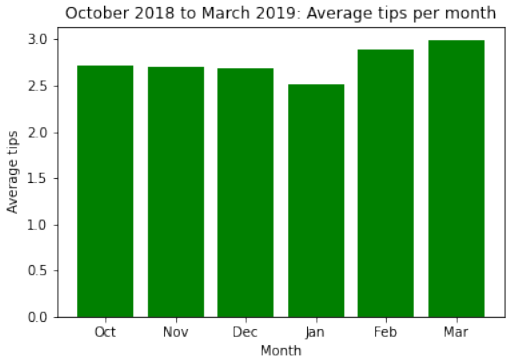

Figure 3: Bar plot of average tips per month

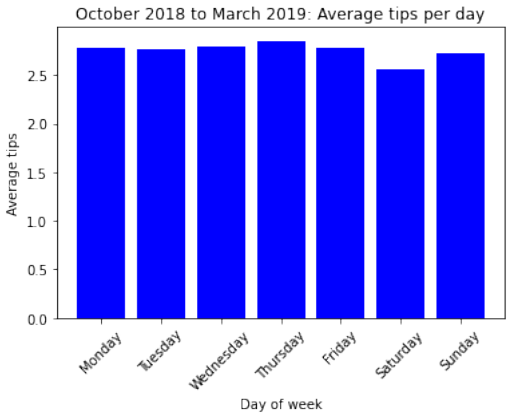

Figure 4: Bar plot of average tips per day of week

3.3 Analyse the effect of pickup hour on average tips

Figure 5 highlights that driving a yellow taxi between 12pm and 6pm results in the highest average tips earned. Contrarily, trips 12am and 5am generated low average tips. This might be attributed to passengers that are travelling home inebriated at night and not behaving as they normally would to even consider tipping.

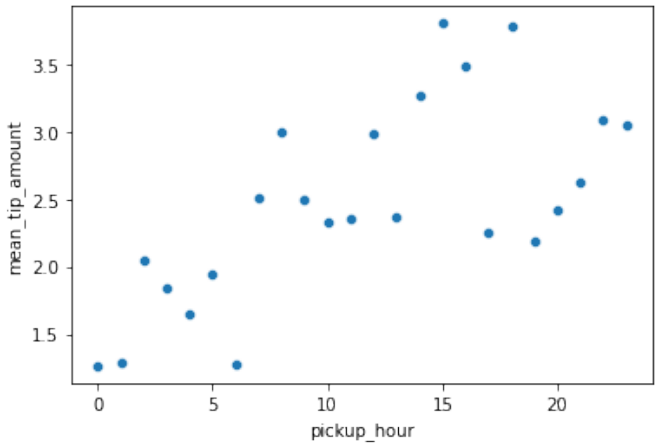

Figure 5: Scatter plot of average tips per hour of day

3.4 Analyse the effect of pick up locations on tips

From Figure 6, the majority of the pick up locations reflect average tips of between $0-$5 USD. Trips originating near Newark Airport and JFK Airport, Queens exhibit much higher average tips.

Interestingly, LaGuardia Airport did not display equally high tips like the other airports. This may be attributed to the fact that this airport was found to have the most dissatisfied travellers in 2018 [4]. A passenger picked up from this airport may be experiencing negative sentiments and thus, be less motivated to tip.

Nevertheless, pick up and drop off location may provide valuable insight on generating higher tips.

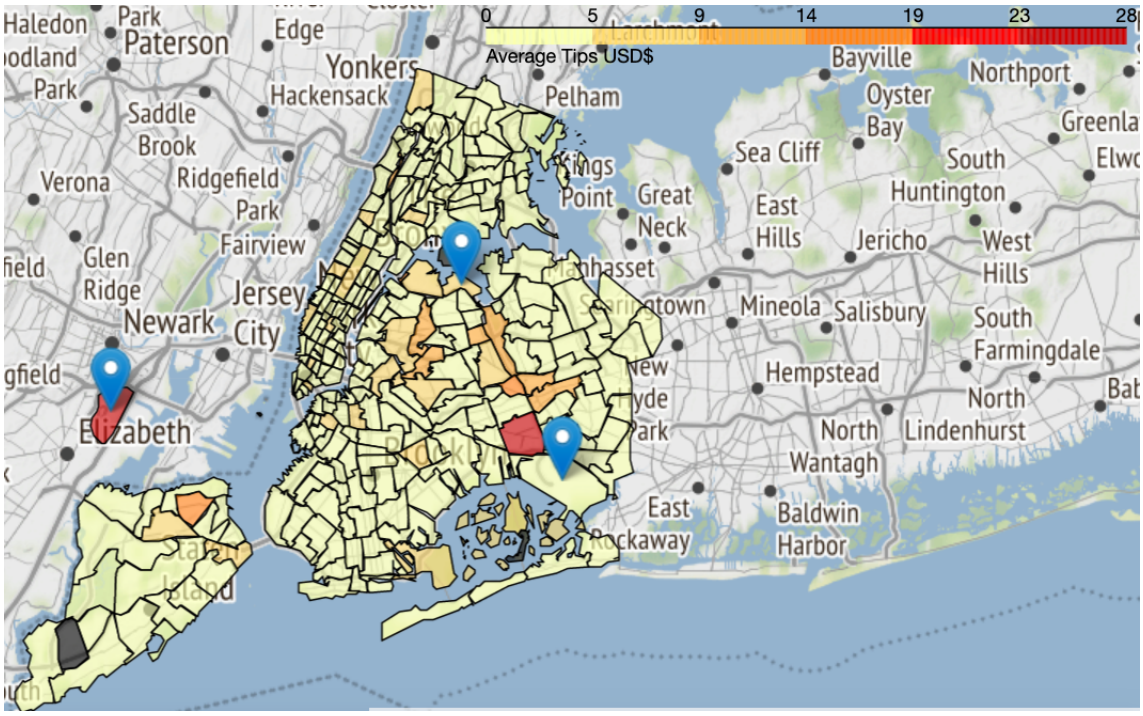

Figure 6: Average tips by pickup location for December 2018

4 Statistical Modelling

4.1 Feature Selection: correlation matrix for tips

Figure 7 highlights that pick up location, trip distance, rate code and total amount are highly correlated with tip amount. Average temperature, dew point and wind speed are correlated with tips to a small extent. Conversely, drop off location and passenger count seem to very weakly correlated. This sets the basis for features to utilise for modelling.

4.2 Model 1: Linear Regression

Linear regression models aim to determine the line of best fit between X (features) and Y (target).

Assumptions: Linear Regression assumes the relationship between X (features) and Y (tip amount) is linear and that for each value of X, Y is normally distributed. The correlation matrix (Figure 7) suggests a linear relationship between selected features and tip amount. The histogram (Figure 2) indicates that the distribution of tip amount is normal. Hence, the assumptions are satisfied. Furthermore, the features used must be continuous and thus, categorical data including rate code ID and pickup location ID must be transformed using one hot encoding.

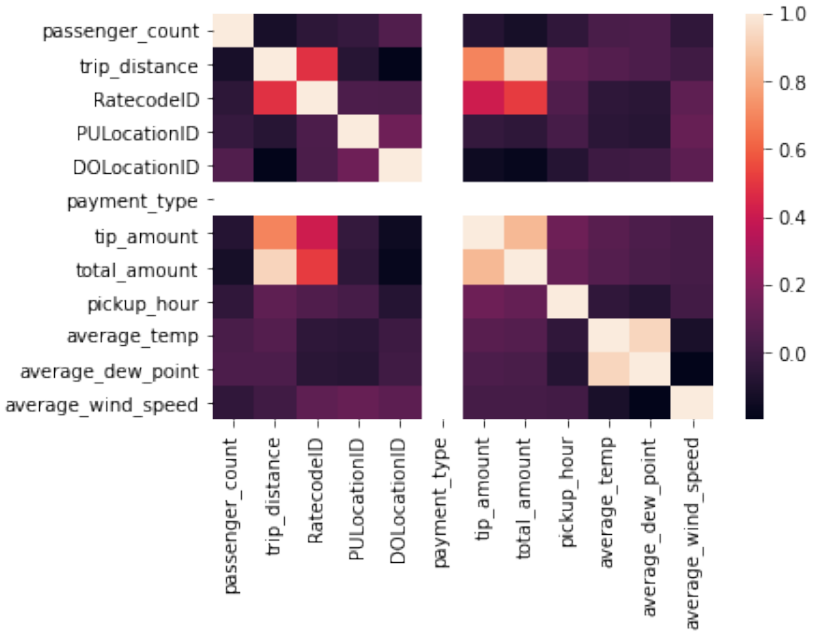

Figure 7: Correlation matrix for features from sampled taxi and weather data

Attributes used: Pick up location, trip distance, rate code, total amount, average temperature, average dew point and average wind speed

4.2.1 Model 1: Evaluation

The linear regression model resulted in a Root Mean Square Error of 1.281. This value is very close to 1 which indicates that the predicted tip amounts and actual tip amounts are very similar on average. Hence, this model will be highly suitable for predicting tips based on different features values accurately.

	coefficient
intercept	0.135099
trip_distance	-0.353345
PULocationID	0.000211
RatecodeID	-0.837902
average_temp	-0.000619
average_dew_point	0.000596
average_wind_speed	0.000549
total_amount	0.249109

Figure 8: Linear regression model: coefficients of features

4.3 Model 2: Decision Tree Regression

Decision Tree Regression is a supervised model that splits data in a way that maximises the predictability of the target (tip amount) and therefore, where entropy is low based on different feature values.

4.3.1 Model 2: Evaluation

The decision tree regression model resulted in a Root Mean Square Error of 1.346. This value is quite close to 1 suggesting that the difference between predicted tip amounts and actual tip amounts is not very large. Thus, this model will be also be useful in predicting tips.

4.4 Comparison of Model 1 and Model 2

The difference in RMSE values of the decision tree regression and the linear regression model can be attributed to decision trees being prone to over-fitting. Because of this, decision trees might be not generalising as effectively as the linear regression model resulting in a slightly larger RMSE value. Decision tree models have the capacity to support non-linearity and can perform better than linear models when features do not display a linear relationship. However, from the RMSE values above, it can be seen that linear model has outperformed. Thus, indicating that linear relationships exist between features.

5 Recommendations

From Figure 8, it was found that rate codes, trip distances, total amount had the greatest correlation with tip amount. However, these attributes are likely to be fixed or determined by the passenger route and hence, it will infeasible for taxi drivers to manipulate these to earn higher tips.

Pickup hours also display a relatively strong positive correlation with tip amount. Taxi drivers may prefer to operate their taxis between 12pm and 6pm in order to generate higher tips on average (Figure 5). These times may be when traffic is the heaviest and passengers may appreciate that it is more difficult to drive in these conditions and so, tip more liberally. Additionally, the travel speeds of taxis when roads are busy during (12pm to 6pm) might be associated with higher or lower average tips. However, further evidence is required to validate these hypothesises.

Pick up locations indicated a positive correlation with tips. Drivers may wish to begin more tips near the zones highlighted in Figure 9 to earn higher tips on average. Two of these fives zones (Figure 9) are attributed to Newark airport and JFK airport. Passengers departing from Newark airport and JFK airport are likely to be tourists or heading home from a pleasant vacation. Thus, these passengers in their rejuvenated spirits, might be more inclined to tip higher amounts.

6 Conclusion

The iconic Yellow taxis date back over a hundred years [5] and are key part of New York City history and culture. Similarly, tips contribute an essential element to the income of taxi drivers who rely greatly on their tips to afford the increasing cost of living. It was found through analysis of features, visual techniques and modelling, that a linear relationship exists between the analysed features (4.2.1) and tip amount. Furthermore, it was discovered that pick up hours and pick up locations could be manipulated by taxi drivers to increase their tip earnings and generate a higher income. Further

	tip_amount	total_trips	average_tip_amount	Zone
0	1976.28	152	13.001842	Newark Airport
5	122.90	12	10.241667	Arrochar/Fort Wadsworth
127	1044107.68	109728	9.515417	JFK Airport
91	6725.85	714	9.419958	Flushing Meadows-Corona Park
189	2035.72	240	8.482167	Randalls Island

Figure 9: Top 5 pickup zones for higher average tips

investigation of external datasets such as analysing New York City sport datasets (e.g. Knicks losing or winning) may also provide valuable insight on factors that affects tipping behaviour.

7 References

[1] Heather Cross. (2020, May 19). A Guide to Tipping in New York City. TripSavvy. https://www.tripsavvy.com/guide-to-tipping-in-new-york-city-4177115#:~:text=A%20simple%20thing%20to%20remember

[2] NYC Taxi and Limousine Commission (TLC). (2018). About TLC - TLC. Nyc.gov. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

[3] National Centers for Environmental Information. (n.d.). Data Access — National Centers for Environmental Information (NCEI). Www.ncei.noaa.gov. Retrieved August 10, 2022, from https://www.ncei.noaa.gov/access/search/data-search/global-hourly

[4] Matousek, M. (2019, August 15). I flew out of the most hated airport in the US, and while it still had issues, I was shocked at how much better it became in less than a year. Business Insider. https://www.businessinsider.com/laguardia-airport-photos-review-2018-12?utm_source=copy-link&utm_medium=referral&utm_content=topbar

[5] Backes, A. D. (2020, January 12). History of New York's Yellow Taxi Cab. ClassicNewYork History.com. https://classicnewyorkhistory.com/history-of-new-yorks-yellow-taxi-cab/