Tip Amount Analysis of Taxis in NYC
along with the Effects of Weather on Taxi Tips

Hanjoo Kim
Student ID: 1078371

August 15, 2021

1 Introduction

Tips contribute immensely in the incomes of various jobs and the New York Taxi drivers are no exceptions. Various attributes such as trip distances, trip duration or even the weather[1] may affect the amounts of tips. This report aims to explore the factors that contribute to the amount of tips a yellow NYC taxi makes, which will hopefully aid in the income earnings of the NYC taxi drivers.

2 Data

2.1 Yellow Taxi Dataset

The dataset used was the yellow from January, April, July and October in 2019, provided by the Taxi and Limousine Commission (TLC) [5]. These months are in each quarter of the year, representing each season. This allows for the data to be in various timeframes of the year which will have different environments that could affect the tip amounts of taxis and give more variability in the data such as weather conditions.

The reason the yellow taxis were chosen was because they are the most prominent types of taxis found in NYC, covering all its areas including Manhattan, which again provides more variability. Moreover, the dataset contained a variety of different features that may provide a deeper insight into the reason behind certain tip amounts.

Data Shape:

- Before Preprocessing: (23688581, 19)
- After Preprocessing: (22750982, 18)

2.2 Weather Dataset

Weather was also assumed to influence the tip amounts in NYC [1]. Therefore, an external dataset containing weather history in NYC was included. The Central Park dataset was used as it was one of the few weather stations located in NYC. Hence, it was assumed that the weather in Central Park will be constant with the overall weather in NYC. The dataset included various features, but precipitation and average temperature seemed most representative of the features and was used.

2.3 Data Cleaning / Preprocessing

1. Yellow Taxi Data: Some cleaning was required as the data had abnormalities and had data that did not follow the guidelines of the TLC.

   - Distances that were 0 were eliminated.
   - Trip durations that were 0 were also eliminated.
   - A RatecodeID of 99 was detected, which was not outlined in the data dictionary [5] provided. Hence it was removed.
   - The TLC states that there is an initial fare charge of $2.5 [3], which should be the minimum fare. Hence, values below that were removed.
   - Tip amounts were also filtered to be at least $0.
   - Some dates were found that did not belong to the months, hence was removed.

2. Weather Data:
   Since there wasn't much variability in the distribution of the precipitation levels or temperature, they were categorized into whether it was raining and whether it was hot or cold.

   - Precipitation:
     - 0 or 'T' (trace of rain) was categorized as no rain (value 0), else it was given 1.
   - Average Temperature of the day:
     - Above 46 Fahrenheit was identified as hot, below was cold.

3 Preliminary Analysis

3.1 Geospatial Visualisation

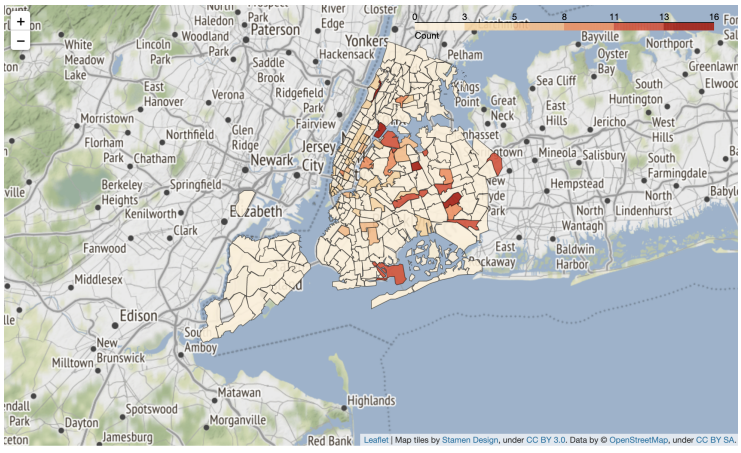

Image 1:A geospatial visualization of the average tip amount a taxi trip makes in the various locations of NYC (by pickup location) for the Yellow Cab trips from June to August.

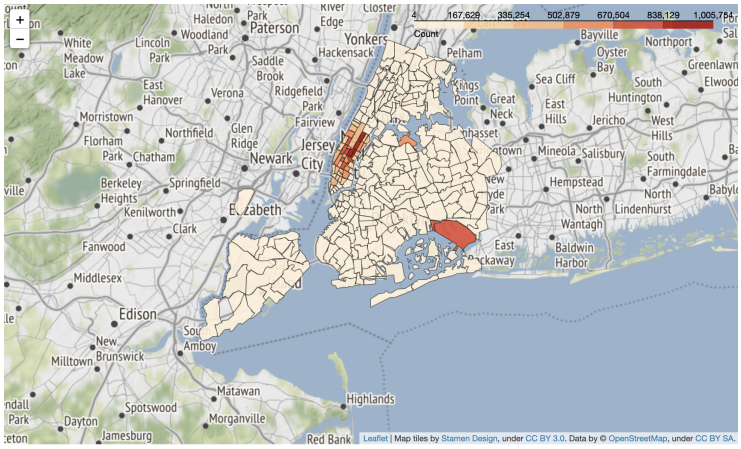

Image 2: A geospatial visualization of the frequencies of taxi trips in NYC areas (by pickup location) for the Yellow Cab trips from June to August.

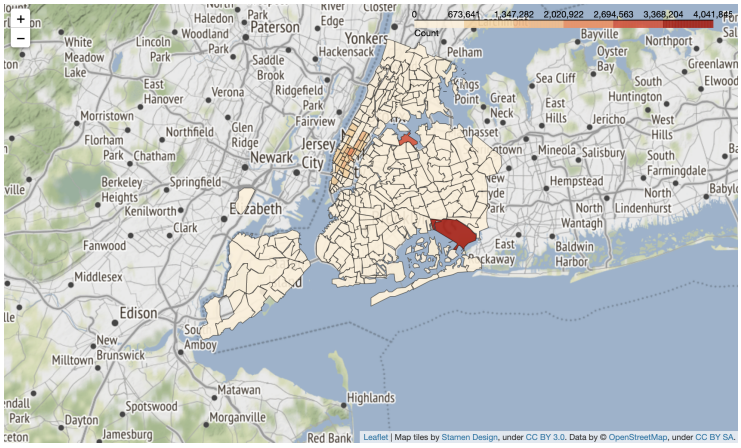

Image 3: Distribution of tip sum of taxi trips in NYC for the Yellow Cab trips from June to August.

These images are all based on the pickup location of the taxi. From image 1, we can observe that the average tip amount of a taxi trip varies around the areas of NYC. It can also be observed that the tips of trips beginning in Manhattan are not the highest and ranges from to $0 to around $10. In contrast, trips starting from outside of Manhattan are quite higher, with some areas ranging from $10 to $15 dollars. This may be because since Manhattan is one of the busiest part of New York and many must travel to Manhattan[2]. Tips are usually known to correlate with the amount of distance traveled. Hence, traveling to Manhattan rather than within it would induce a higher tip, due to more distance being covered and time taken.

The evidence that Manhattan is one of the busiest places in NYC is portrayed in Image 2, where the frequency of taxi trips is the highest out of all the areas. Furthermore, this is also recognized in the John F. Kennedy airport, which is the area with the second-most taxi trips in NYC. Since the frequency of taxi trips are high in the two areas, it can be observed in Image 3 that the total sum of tips earned are the highest in both areas. Additionally, although the JFK airport seemed to have a very low average tip amount, the sheer frequency of taxi trips from the airport allowed a high sum of tips amounts.

3.2 Attribute Analysis

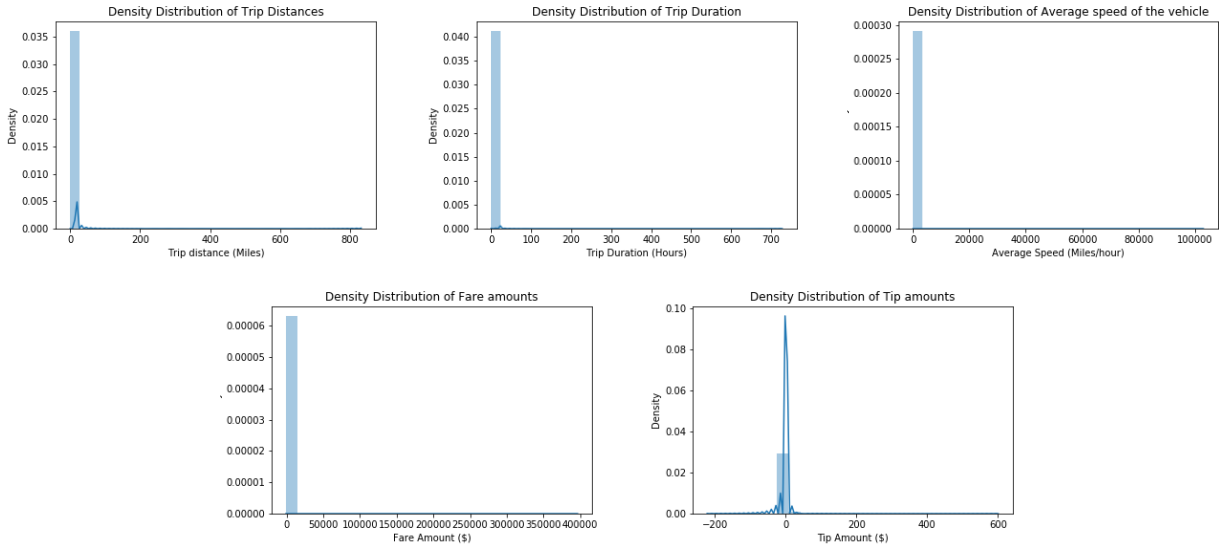

Figure 1: The density distribution plot of continuous attributes as raw data values

Figure 1 shows the distribution of the raw values of the whole dataset of June to August. Many outliers can be seen from the plots, where every feature has extreme values that are not possible. These outliers skew the distribution; hence the values were logified to get a better visualization of the data.

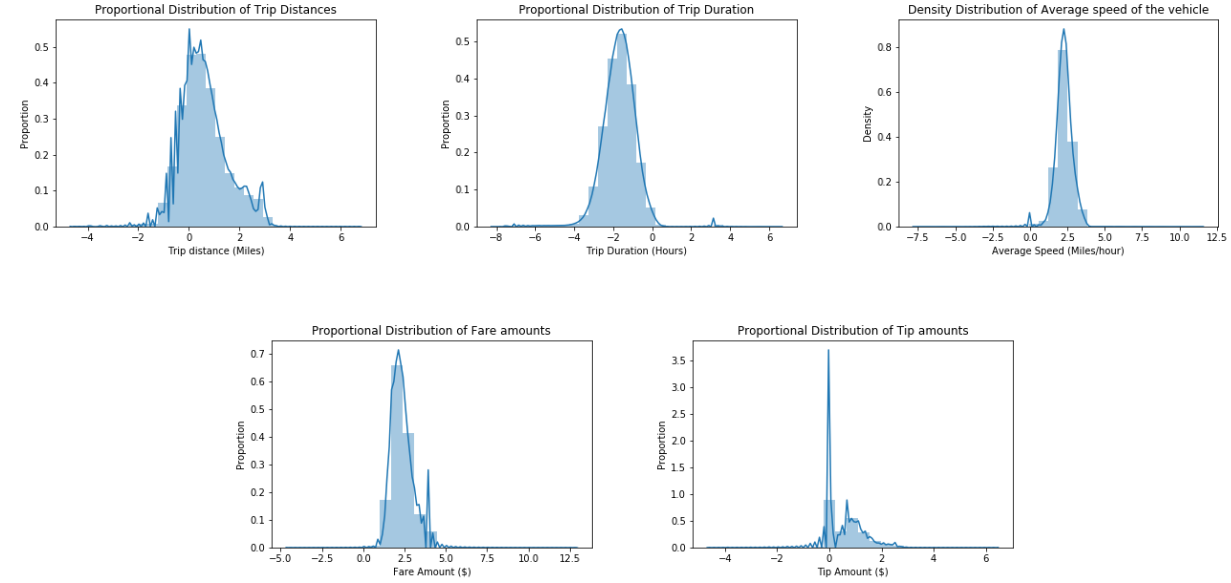

Figure 2: The distribution plot of continuous attributes, but with logified values

From Figure 2, the distributions of the logified values are now able to show where the values are mainly distributed for each feature, allowing identification of approximate limits as boundaries for outliers.

- Trip distances: More than 0 and less than exp(4)

- Tip amounts: From 0 to $exp(4)
  - (tips can be $0)
- Fare amounts: From $2.5 to $exp(5)
  - Initial charge is $2.5
- Trip duration: From 0 to exp(4)
  - Since there is a spike in between exp(3) to exp(4), do not exclude it.

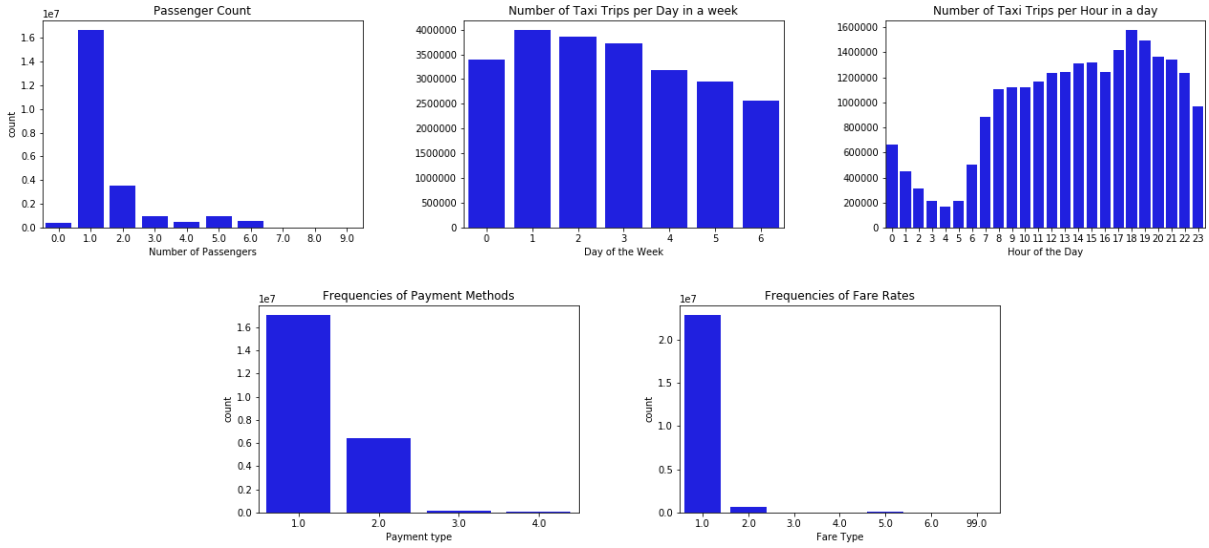

Figure 3: The countplot of categorical attributes

Figure 3 depicts that there are data that are extremely small in count and should be ignored. This includes:

- Passenger count 7, 8 and 9 should be ignored.
- Payment type 3 (No charge) and 4 (Dispute) should be ignored.
  - Payment type 2 (Cash) is also later found to have almost no tips and was ignored.
- Fare Type 3 (Newark), 4 (Nassau or Westchester), 5 (Negotiated Fare) and 6 (group ride) should be ignored.

	tip_amount	total_amount	fare_amount	trip_distance	trip_duration	tolls_amount	average_speed
tip_amount	1.000000	0.709508	0.578576	0.558084	0.075321	0.449201	0.315724
total_amount	0.709508	1.000000	0.976512	0.926656	0.145837	0.686011	0.505526
fare_amount	0.578576	0.976512	1.000000	0.941520	0.152974	0.608196	0.509692
trip_distance	0.558084	0.926656	0.941520	1.000000	0.137616	0.622693	0.663360
trip_duration	0.075321	0.145837	0.152974	0.137616	1.000000	0.078940	-0.057193
tolls_amount	0.449201	0.686011	0.608196	0.622693	0.078940	1.000000	0.406787
average_speed	0.315724	0.505526	0.509692	0.663360	-0.057193	0.406787	1.000000

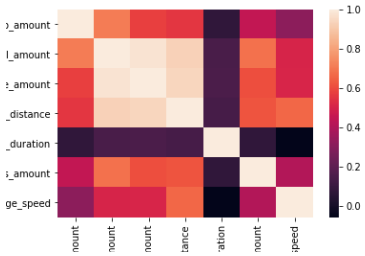

Figure 4: Correlation table and heatmaps of continuous attributes vs tip amounts

The correlation table in Figure 4 portrays that every chosen attribute has a positive correlation with tip amounts: fare amounts, total amount, trip distance, trip duration and tolls amount. Therefore, these features will be used for the Regression models for predictions.

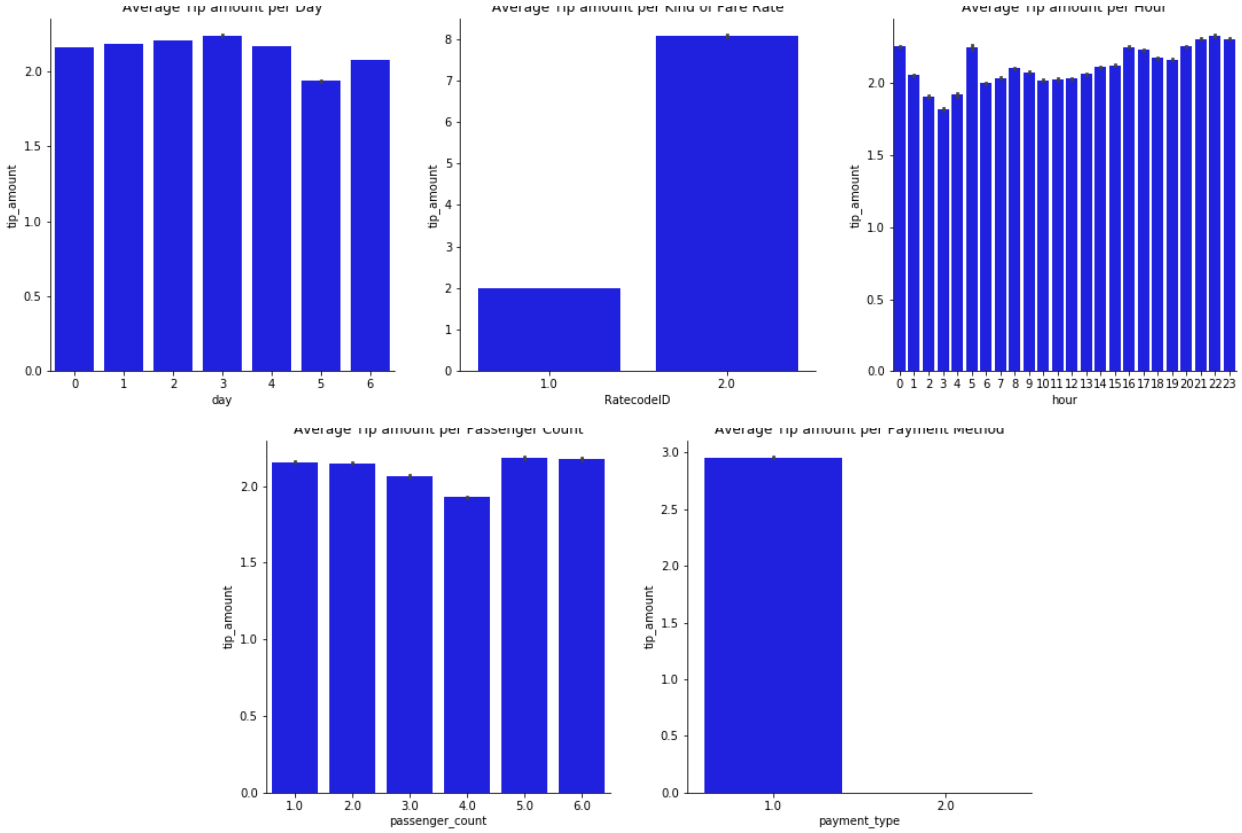

Figure 5: The Average tip amounts against the categorical attributes

Figure 5 depicts the relationships of the tip amounts to categorical attributes of interest. From figure 5, we can observe that tips are mostly given when the type of payment are credit cards (payment type 1). Hence the data was additionally processed to only use credit card transactions for the modellings. Another notable characteristic is that the RatecodeID = 2 (JFK airport) has a much higher average tip amount than RatecodeID = 1 (Standard rate). Finally, It was observed that the highest average tip amount is on Thursday and gets higher later in the day.

4 Statistical Modelling

4.1 Model

Models were made to predict the 2020 tax tip amounts using the 2019 data that has been analyzed. The 2020 data was preprocessed in the same manner as the 2019 data.

In order to choose between the continuous features, a backwards selection method was to be implemented. However, when using all the continuous attributes as a predictor in the model, it was shown that the R-squared value was 0.866 and no attributes had values less than the significance level of 0.05. Therefore, all features were retained and used as predictors. One thing to note was that the categorical features were not included in this backward selection method. However, these categorical features will still be used in the final model as they may have relationships to tip amounts.

Dep. Variable:	tip_amount	**R-squared:**	0.866
Model:	OLS	**Adj. R-squared:**	0.866
Method:	Least Squares	**F-statistic:**	2.455e+07
Date:	Sun, 15 Aug 2021	**Prob (F-statistic):**	0.00
Time:	19:50:30	**Log-Likelihood:**	-2.9599e+07
No. Observations:	22750982	**AIC:**	5.920e+07
Df Residuals:	22750975	**BIC:**	5.920e+07
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.4428	0.001	-3376.820	0.000	-2.444	-2.441
trip_distance	-0.0165	0.000	-83.864	0.000	-0.017	-0.016
fare_amount	-0.7082	0.000	-6734.783	0.000	-0.708	-0.708
tolls_amount	-0.7190	0.000	-3699.252	0.000	-0.719	-0.719
total_amount	0.7425	7.93e-05	9366.116	0.000	0.742	0.743
trip_duration	0.0026	0.000	16.297	0.000	0.002	0.003
average_speed	0.0199	4.78e-05	417.553	0.000	0.020	0.020

Omnibus:	1858295.281	**Durbin-Watson:**	0.998
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	7715834.185
Skew:	0.321	**Prob(JB):**	0.00
Kurtosis:	5.780	**Cond. No.**	121.

Figure 6: Summary of the fitted Regression Model

Before the model selection, the continuous attributes were all standardized into a single scale, as that is required for Regression models. Regression Models were thought to be fit for tip amount predictions as it is a continuous variable and there were attributes of interest that linearly correlated with it. Four types of regression models were examined: Multiple Linear Regression, Lasso Regression, Ridge Regression and an Elastic net Regression. The Multiple Linear Regression method was expected to be the baseline model and the final model was to be chosen out of the three penalized regression models.

4.2 Results

Linear Regression Model (Baseline):

- RMSE value: 1.3177765106749273
- Accuracy Score: 0.6511544806206744

Ridge Regression:

- RMSE value: 1.3182461783242958
- Accuracy Score 0.6509057728146819

Lasso Regression:

- RMSE value: 1.806082901269438
- Accuracy Score 0.3447229054215727

Elastic net Regression:

- RMSE value: 1.7686069549266916
- Accuracy Score 0.3716345725050276

The Ridge Regression model was concluded to be the optimum model out of the three regression models as it had the lowest RMSE. However, the model prediction accuracies was similar to the baseline model accuracy, which was unexpected when the continuous features showed high correlations with the tip amounts. There may be various reasons to this:

- Penalized Regression models were expected to do further feature selection as it is embedded in their algorithm. However, the accuracies output was still extremely low, (especially in Lasso and Elastic net). Perhaps changing the weight of features to 0 may have ruined the model predictions.
- Secondly, it was recognized later in the study that 2020 was the year of the Covid-19 pandemic (first case of NYC on March 1st[8]). Therefore, this would have introduced many unknown effects on the taxi datasets provided after March. Hence, predicting the tip amounts in 2020 would not be applicable.

4.3 Discussion

From the preliminary analysis of the data, it could be observed geographically that the places gaining the most tips are trips beginning in Manhattan. Manhattan is one of the busiest tourist attractions of NYC[2] and many international visitors have to move by Taxi from airports. This is backed up by the high amount of taxi trips from JFK, the 6th most busiest airport in the US[4].
Strong positive correlations could be found in the total charge, fare amounts, trip distances and trip duration with tips. This means that the more distance or time the trip takes, the more likely it is to get tipped.
The model prediction accuracies for 2020 tip amounts were very low, which could have been due to the Covid-19 pandemic. Less tourists would have entered the States, which resulted in the smaller dataset in 2020. A smaller dataset would result in more variance in the data, which will cause it more harder to predict. Hence, the model may be more useful for data of when the Covid-19 pandemic is over.

5 Recommendations

Recommendations for earning more tips would be to:

- Go to places where there is a high frequency of pickups, which was found to be within the Manhattan borough or the JFK airport.
- Airport trips tip higher in average.
- Trips later at night tip slightly higher.
- Longer trips or distances results in a higher tip amount
- The driver should offer the customer to pay in credit card, as it was observed that cash does not tip.

6 Conclusion

This report explored the changes in the tip amounts of trips exposed in different environments. Tips are not always given in taxi trips and is often affected by various attributes of the trip. In the case of New York City taxi trips, it was observed that the JFK airport seemed to be the best place for earning tips and taxi drivers are recommended to pickup passengers from there.

References

[1] Devaraj, S., & Patel, P. (2017). Taxicab tipping and sunlight. PLOS ONE, 12(6), e0179193. https://doi.org/10.1371/journal.pone.0179193

[2] Law, L. (2021). 20 Top-Rated Tourist Attractions in New York City — PlanetWare. Planetware.com. Retrieved 15 August 2021, from https://www.planetware.com/tourist-attractions-/new-york-city-us-ny-nyc.htm.

[3] Taxi Fare - TLC.(2021) Www1.nyc.gov. (2021). Retrieved August 14, 2021, from https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page.

[4] The Busiest Airports in the US - Top 10 by Passenger Numbers. MyFunkyTravel. (2021). Retrieved 15 August 2021, from https://myfunkytravel.com/the-busiest-airports-in-the-us.html.

[5] TLC Trip Record Data - TLC.(2021) Www1.nyc.gov. (2021). Retrieved 14 August 2021, from https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

[6] Wang A. (2021). Applied Data Science MAST30034 Lab 2 Semester 2 - 2021 https://github.com/akiratwang/MAST30034_Python/tree/main/tutorials

[7] Wang A. (2021). Applied Data Science MAST30034 Lab 3 Semester 2 - 2021 https://github.com/akiratwang/MAST30034_Python/tree/main/tutorials

[8] West, M. (2021). First Case of Coronavirus Confirmed in New York State. WSJ. Retrieved 15 August 2021, from https://www.wsj.com/articles/first-case-of-coronavirus-confirmed-in-new-york-state-11583111692.