Maximising the Profit Gained by the Taxi Drivers of the New York City

Wanxuan Zhang
Student ID: 1079686

August 16, 2021

1 Introduction

This project aims to make a quantitative analysis of the New York City Taxi and Limousine Service Trip Record Data. The data set chosen were the yellow taxi trip recorded for the months January until June in 2018. The purpose is to investigate the data set of yellow taxis from January to June in 2018, which aims to maximise the profit gained by the taxi drivers.

1.1 Yellow Taxi Data

Yellow taxi data were chosen because it includes the attributes that describe the component of the profit for each trip. This kind of taxi is also allowed to have trips in the Manhattan area which allows for an integrated investigation about the profit gain of the drivers in New York City. This data set has approximately 5.3 million rows and 17 columns. With a target audience of the taxi drivers, attributes in interest include the pick-up and drop-off time, passenger counts, the location where each trip starts and ends, trip distance, fare amount, tip amount and total amount. The time and location of the trips help the drivers to identify when and where the most demands occur. While also get an insight into under what circumstances will they be getting the most tips.

1.2 Weather Data

The weather data from NOAA is included as an additional data set to investigate the relationship between the weather condition and the tip amount. This data set has approximately 181 rows and 44 columns. It describes the weather recorded by the station at the New York City Central Park from January until June 2018. Here it assumes that all the regions in New York City have the same weather condition as recorded for this station.

2 Processing the data

2.1 Cleaning

- All the records with zero passengers count or zero trip distance were removed.
- Remove the records that have a pick-up time and drop off time outside the range of the start of January to the end of June.
- Only trips with credit card payment were kept as the tip amount is only recorded for this type of payment.

- Trips with fare amount or tolls amount less than $2.5 were filtered since it is the start-up amount.

2.2 Pre-processing and Feature Engineering

Before the data is ready for analysis, some new features are integrated out of existing ones. For the Yellow Taxi data, by plotting the box plot of trip distance and amounts, it is clear that many outliers were present in the data set. The outliers are all present at the right of the data set, since both distance and amounts cannot be less than zero. There should not be any outlier that is less than the minimum value of zero. Those trips with a trip distance or amounts that are more than three times of its IQR is removed to avoid these extreme cases affecting the later evaluation. After removing the outliers, re-draw the box plot, it shows that the amounts are normal distributed with a slight right skew.

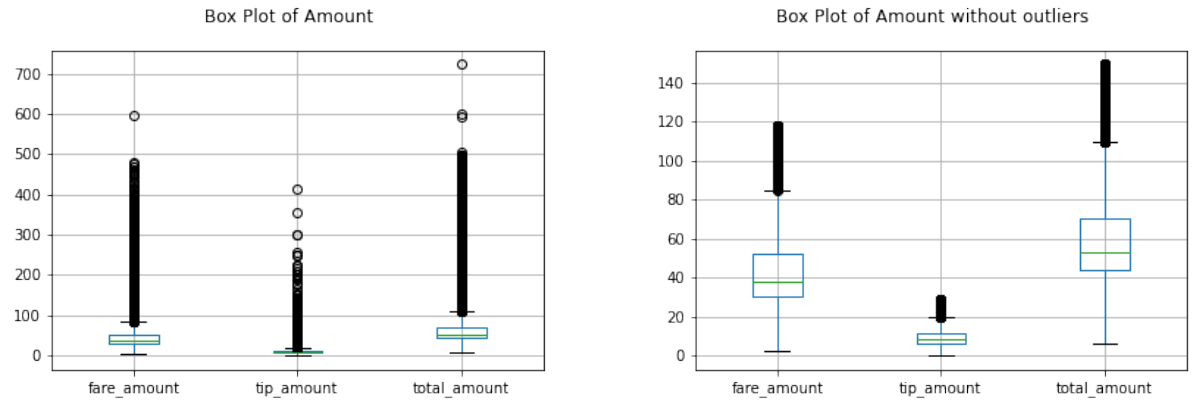

Figure 1: Box Plot of Amounts with and without Outliers

Moreover, the attribute describing the date and time for pick up has been split into two attributes each indicating the pick-up time and pick-up date accordingly. A new feature "time" is generated from the time spent between the start and end of the trip. Another feature "type" is formed indicating the type of a day, whether it is a holiday, weekday or weekend.

For the weather data, the feature of average temperature is generated by calculating the mean of the maximum and minimum temperature each day since most of its values are missing. Remove the features of maximum and minimum temperature. The type of weather that indicating bad weather such as whether it is snowing or raining on that day, have been concatenated into a new feature call "bad weather" with any of the bad weather is satisfied for this new feature to be shown as "Y", otherwise "N".

3 Analysis and Geo-spatial Visualisation

To maximise the profitability, the amount gained by the driver each day should also be maximised. We should consider the factor that where the driver is most likely to get a customer. Furthermore, there is a positive linear relationship between the total amount and the trip distance. It means that drivers will spend more time on a long trip to gain more profit. Hence, the most efficient factor to maximise the profit for each trip is to have a higher tip amount.

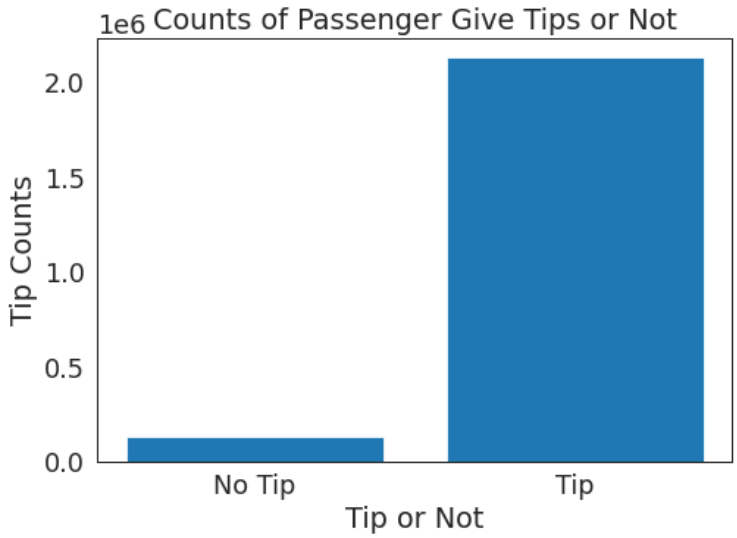

Figure 2: Counts to Whether the Passenger Give Tips

3.1 Distribution of Tip Amount

From this bar plot, it shows that most of the passengers will pay some tips. It is practical to investigate the factors that affect the tip amount and how drivers can maximise tips in order to increase their profitability.

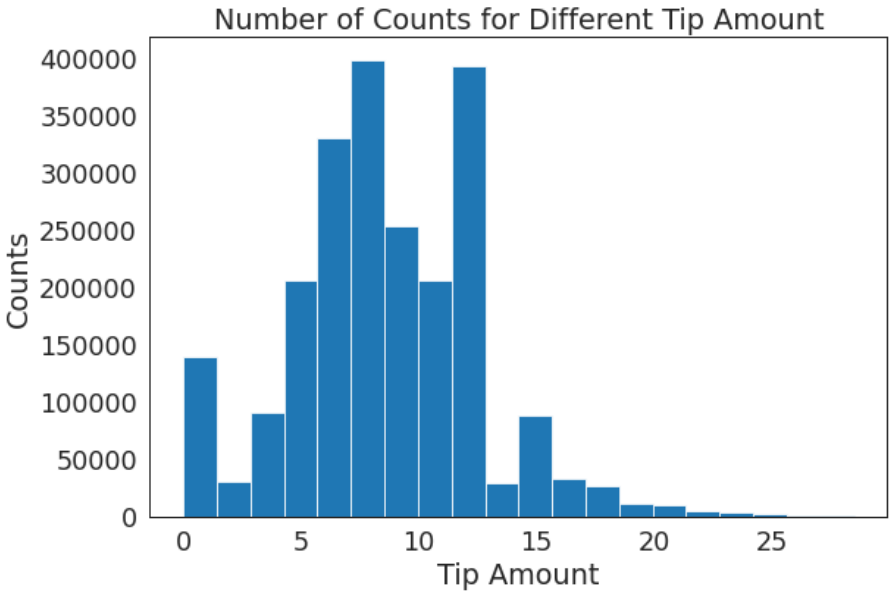

Figure 3: Number of Counts for Different Tip Amount

From the histogram, out of the passengers who are paying tips, it depicts that the tip amount is slightly right-skewed, with a mean around 8 dollars, which is $8.51 by calculation. However, since the tip amount is restricted to be greater than or equal to zero, it can be actually normal distributed but with its left tail cut off.

3.2 Correlation Heat-map

By plotting the correlation heat-map, all the features of weather have a low correlation with the tip amount, which is less than 0.2. Whereas total amount, fare amount, tolls amount and trip distance have correlations approximately higher than 0.5. Those features with high correlation with tip amount can be used to build models in the tip prediction.

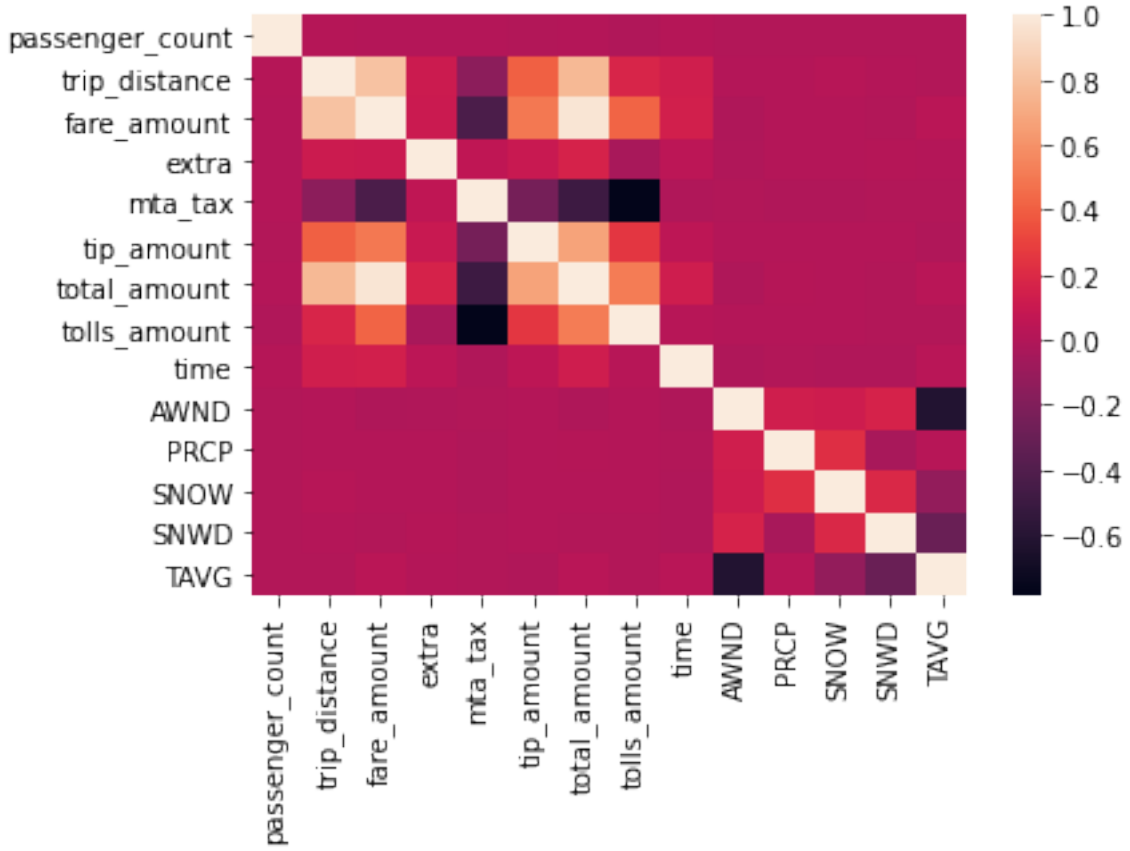

Figure 4: Correlation Heat-map

3.3 Type of a day and the Tip Amount

By classifying days into weekdays, weekends, holidays, it shows that the average number of trip frequency is similar for different day types, with a daily trip frequency of approximately 11900, 12600 and 11300 for holiday, weekday and weekend accordingly. Here no matter a holiday is on weekdays or weekends, it is all classified as a holiday. It means that the probability of a driver get passengers is similar on the different types of days. Moreover, shown by the plot below, the average tip per trip for different type of day is also closed. It suggests that the tip amount gained does not have a correlation with the type of day.

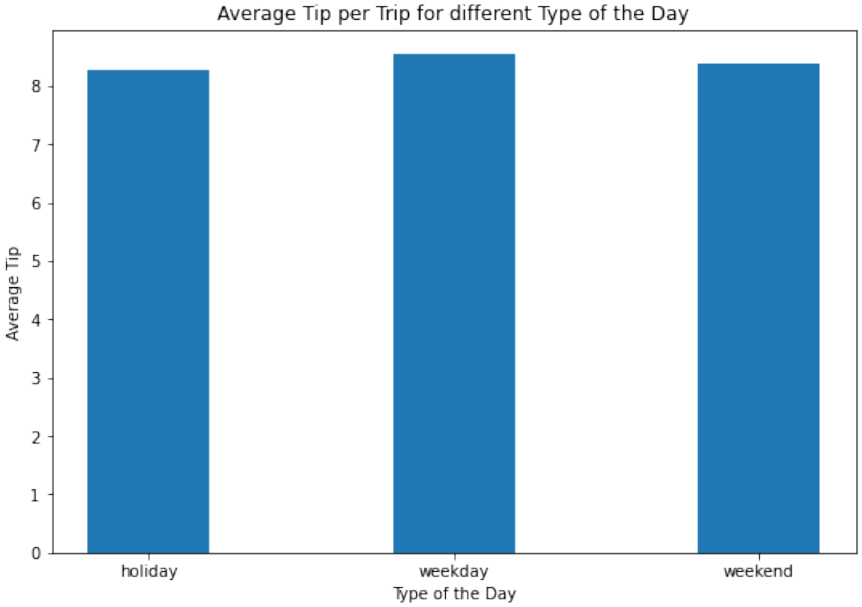

Figure 5: Average Tip per Trip for different Type of the Day

3.4 Time of a Day and Trip Frequency

By inspecting the relationship between the time of a day and the trip frequency, trips occur more frequently around 12:00 to 18:00 in the afternoon. Even there can be more taxis presents at this time frame, it is still a valid indication of when the most taxi demand occurs and helps the driver to decide whether to work at that time.

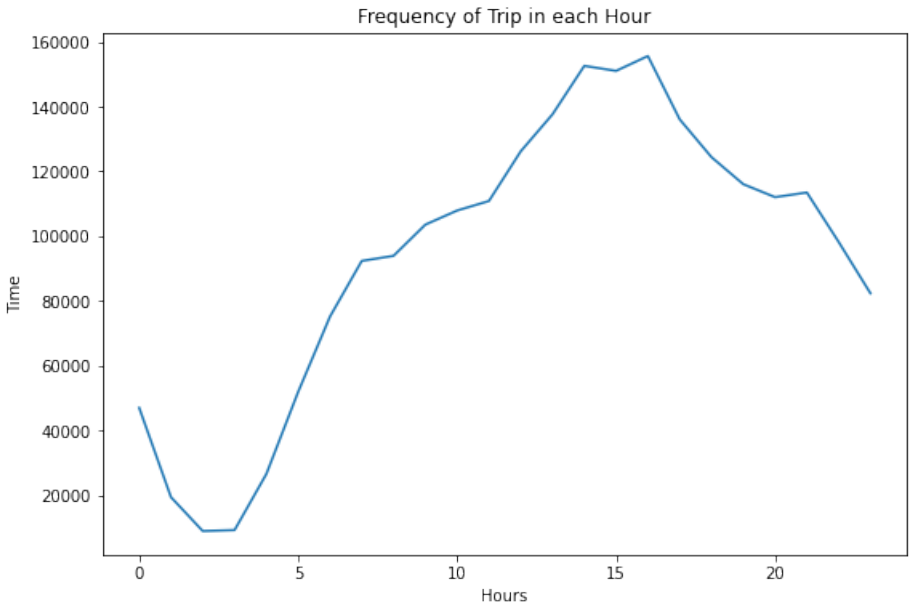

Figure 6: Frequency of Trip in each Hour

3.5 Top Tipped Region

Plotting the map of the New York City showing the average tip amount in each region, it illustrates that regions in the Manhattan and airport area have a relatively high tip amount per trip.

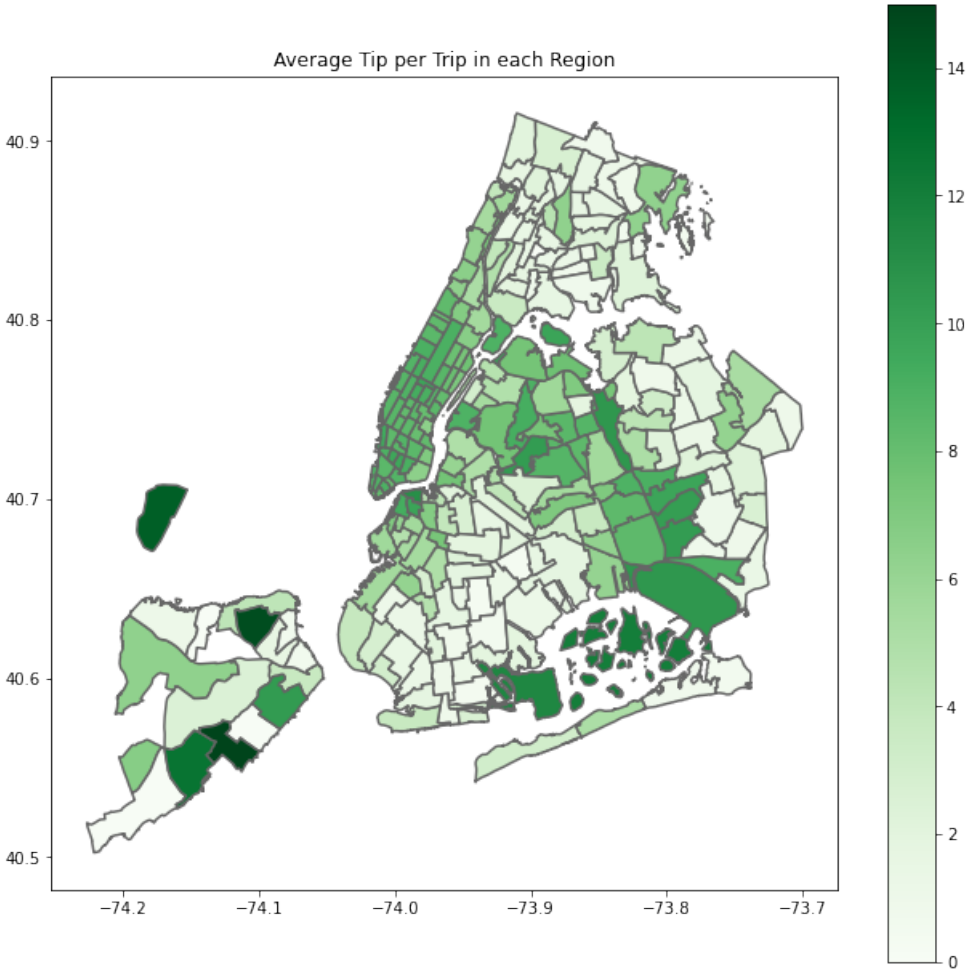

Figure 7: Average Tip per Trip in each Region

The plot below shows the top 10 tipped regions per trip. These are some areas that drivers can consider to pick up their passengers, which could possibly help them to gain the most tip amount. For further analysis, We can also calculate the possibility of one region go to another. It can help the drivers to choose a pick-up location of the "high tip" region, which its passengers have a higher chance to be dropped off in another "high tip" region. Hence, it is easier for the driver to find the next passengers from that region.

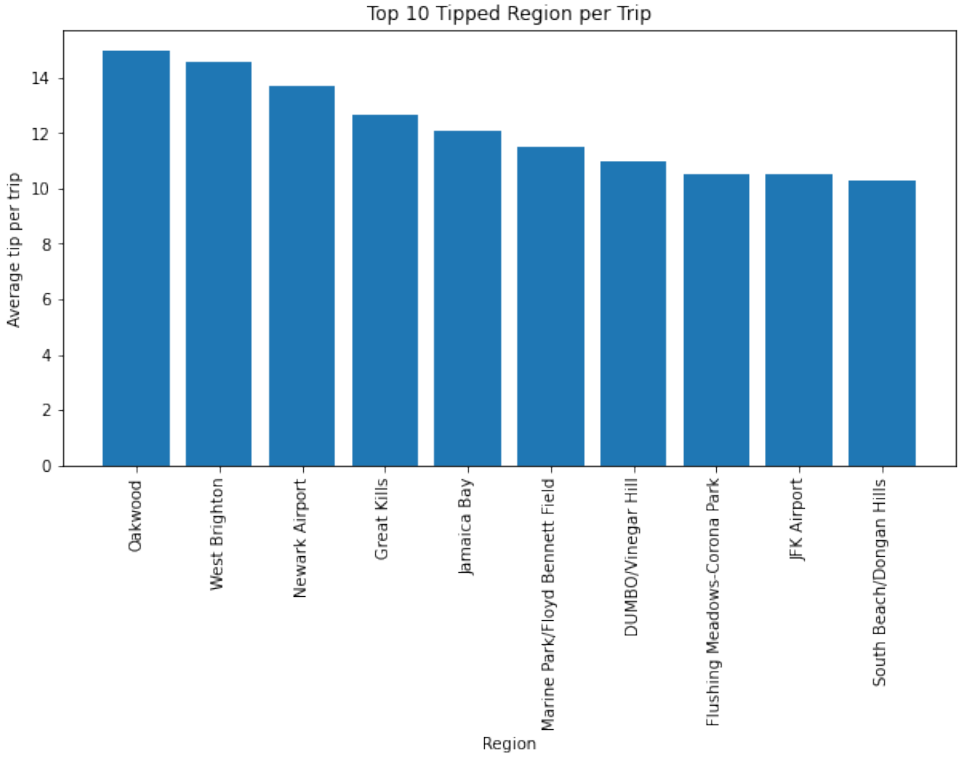

Figure 8: Top 10 Tipped Region per Trip

4 Statistical Modelling

4.1 Feature Selection

To build a regression model to predict the tip amount, first sample 1% of the records to build the model, removing features that have low correlations with tip amount such as weathers. Use the ordinary least square method to estimate the features in a linear regression model. By testing the strength of the relationship for passenger count, trip distance, fare amount, total amount, tolls amount and the trip duration to the tip amount. The feature passenger count has p-values of 0.181 which is greater than 0.05 hence remove it. Now calculating the ordinary least square again, the R^2 value stays the same as 0.94 which indicates a potential problem of over-fitting. Moreover, by inspecting the AIC and BIC value for these two models, it shows that both AIC and BIC has been reduced after the attribute passenger count been removed. Hence, the null hypothesis is rejected and we can conclude that the intercept parameters are non-zero.

	Full	Reduced
AIC	62740.74	62744.12
BIC	62796.90	62784.23

Table 1: AIC and BIC of Full vs Reduced model

4.2 Model and Results

Ridge regression is chosen because it is suitable to deal with scenarios where multicollinearity exists, as the features are not completely independent. Trip distance, fare amount and total amount are highly correlated with each other. The features used are the trip distance, fare amount, total amount, tolls amount and the trip duration. Standardise the features and fit to the train set Use grid search to adjust the value of the parameter alpha, with a 10-fold cross-validation repeat three times, to get a optimise value of 0.005. Fit the alpha value into the ridge regression model to obtain the coefficients of the features(trip distance, fare amount, total amount, tolls amount, time) and intercept, we get the coefficients to be -0.04288, -11.8218, 15.0420, -1.6943, -0.0268 and 8.5119 respectively.

By calculating the root mean square error of both training and testing data with and random selected alpha 5 and the tuned alpha 0.005, we find that the RMSE decreases for using the tuned alpha 0.005 for the training but slightly increase for the test set. Both RMSEs are higher for the test set compared to the training set which is reasonable since the training set is used to train the model. Since the RMSE is high for this model, it performs worse than expected. This might be because the data do not fit the model well. I should consider other models in future investigations.

RMSE		
alpha	Train	Test
5	0.96578072	0.96965420
0.005	0.96578066	0.96965498

Table 2: RMSE of Train vs Test Data with Different Alphas

5 Recommendations and Discussion

In general, the model has a relatively high RMSE which can be further improved in future modelling. The speed of the trip can be a potential factor that affects the tip amount, while it can only be calculated by the distance divide by time. It does not consider the possibility of having heavy traffic and the stopping time. As the trip takes longer time than what the passengers were expected due to these circumstances and the as a result that the fare amount increases, the passengers were likely to give fewer tips. Moreover, some other external data set such as the oil price can be also considered in maximising the profitability of drivers.
Some recommendations for the taxi drivers are that in order to increase their profitability, they can go to those top tipped regions as shown in the previous graphs. The weather and the type of a day (weekday, weekend, holiday) do not have much effect on the tip amount that they might get for a trip. If we assume that the trip frequency is equal to the taxi demand, the drivers should go out between 12:00 to 18:00 in order to increase the possibility of having passengers and hence increase their profitability.

6 Conclusion

Overall, the model built does not perform as good as expected, further improvements can be done to obtain better results. Since I chose to build a regression model, some categorical features is not used when building the model. Other classification models such as Logistic Regression or Linear SVC can be considered in future investigation in order to build a better model.

References

[1] "TLC Trip Record Data." TLC Trip Record Data - TLC. (2021). Accessed August 01, 2021. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

[2] "Daily Summaries Station Details." Daily Summaries Station Details. (2021). Accessed 5 August 2021
https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00094728/detail