The Modelling and Analysis of Yellow Taxi data in New York

(For the period of summer and winter, 2017 and 2018)

Bohan Yang (814642)

1. Introduction

This report aims to analyse the factors that affect the profitability of yellow taxi drivers in New York city. Specifically, pick up hours of the day, average driving speed for the trips and temperatures are considered the most while selecting variables and building model. The sample data for this project covers the summer and winter months of year 2017 and 2018. The project is mostly done with Python 3 and R. The taxi data is collected from NYC Taxi and Limousine Commission (TLC) (*https://www1.nyc.gov/site/tlc/about/tlctrip-record-data.page*, 2019). The weather data is collected from *Visual Crossing (https://www.visualcrossing.com/weather/weather-data-services, 2019).*

2. Data Period and Sample Size Determination

In the recent years, as ride-hailing apps rose, the business of taxis in New York city have been deteriorating. As shown from figure 1 (Schneider, 2018), in year 2017, the pickup number for ride-hailing apps had caught up to taxis. Since then, the pickup number for ride-hailing apps has been rapidly raising as it drops for taxis. Since the profitability for taxi drivers under the influence of ride-hailing apps in New York City is of interest, the data for year 2017 and 2018 is selected for this study. Based on the trend for this data, taxis will most likely be continually effected by ride-hailing apps in the following years, therefore year 2017 and 2018 are not only the most recent years, but also the years under the most similar circumstances as future years, therefore are of more value.

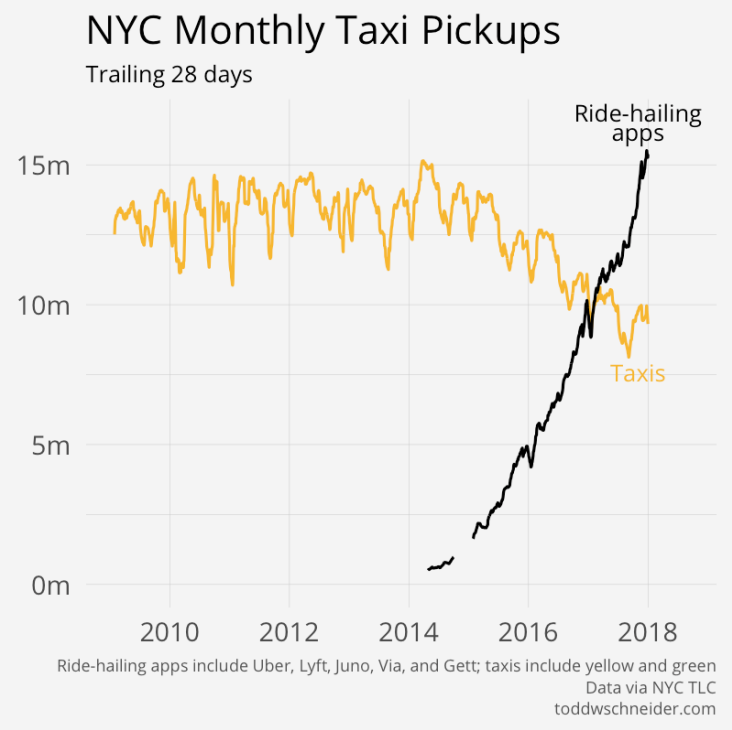

Figure 1 Trend of Pickup Numbers for Ride-hailing Apps and Taxis

From the initial visualizing of the first project, temperature is discovered to be relevant to the amount of tips yellow taxi drivers received. As shown from figure 2 below, since the temperatures are more extreme in winter and summer, December, January and

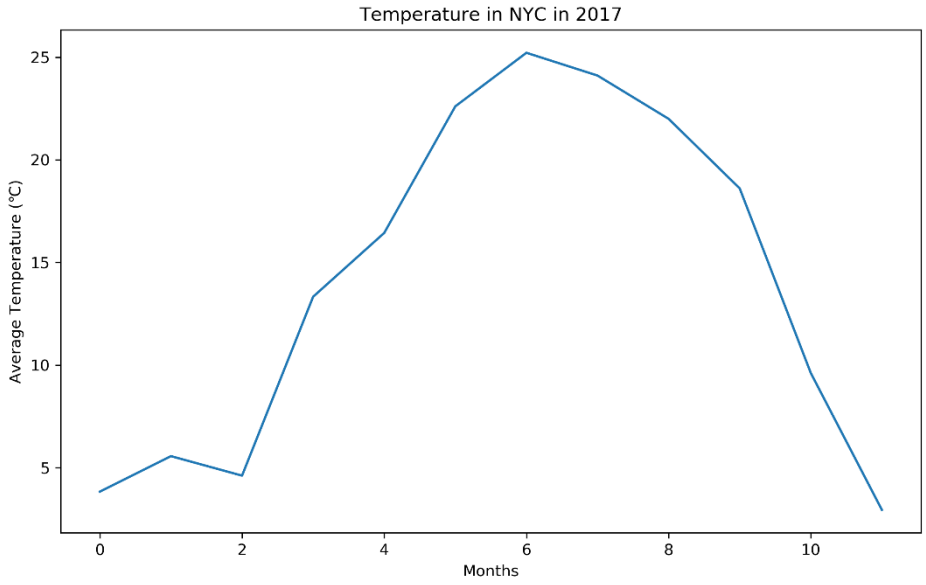

Figure 2 All Year Temperature in NYC in 2017

February, as well as June, July and August are of the most interest among the twelve months.

For each month of yellow taxi data, there are approximately 9 million instances, which is too large and will be too time consuming for training data. In order to cover a larger range of data, a random sample for each month is taken. The necessary sample size is calculated by the formula *Necessary Sample Size = (Z-score* standard deviation/ margin of error)*^{2}. With a 95% confidence interval, standard deviation of tips amount for each month, and margin of error of +/- 0.1, the necessary sample size is calculated separately for each month data. In the end, 30875 instances are chosen randomly over 6 months in 2017 and 2018(12 months in total) for this project.

3. Attribute Selection

Since the aim of the project is the profitability for yellow taxi drivers, the fee-related columns are considered to be the response variable. Among Fare amount, Extra, MTA tax, Improvement surcharge, Tip amount and Tolls amount, Tip amount is the most beneficial to the drivers' income. It is also the only part of the total charge which the drivers can possibly alter for their income. Therefore, Tip amount is chosen as the response variable.

As for possible predictors, pickup datetime and dropoff datetime are chosen for calculations of trip durations and more benefiting pickup time period. Trip distance is chosen for calculation of average driving speed. A slow average driving speed could imply a traffic jam, which could possibly influence the mood of passengers, hence the amount of tips. Furthermore, speed is an element which taxi drivers can control, therefore is of interest. Pickup locationID is kept in the dataset but may not be a useful predictor. However, it could be useful for identifying and explaining the possible outliers or special cases. Weather data for 2017 and 2018 is collected (www.visualcrossing.com, 2019) and maximum, minimum and average temperatures are to be considered as possible predictors as well.

Since the drivers cannot choose to pick up or refuse the trip based on the destination of passengers or the number of passengers, drop off location ID and passenger count are not useful for this project. VendorID, RateCodeID and Store and forward flag are not relevant for this project. Since the Tip amount recorded in the data is only for credit card tips, the payment type can only be Credit card in order for Tip amount to be non-zero. Since all instances then have the same value for the attribute Payment type, this attribute is also not useful anymore.

4. Data Cleaning and Pre-processing

January, February and December datasets are joined to be “summer” datasets, while June, July and August datasets are joined to be “winter” datasets for 2017 and 2018 respectively. Since the discomforting temperatures in summer and winter are opposite of each other, they are originally considered separately while considering the relation of temperature and tip amount(which later proven to be pointless, therefore both datasets are combined to be a single one in the modelling stage).

Checking the four datasets shows that there are not any missing values in the datasets(Figure3).

```
print(df_2017_winter.isnull().values.any())
print(df_2017_summer.isnull().values.any())
print(df_2018_winter.isnull().values.any())
print(df_2018_summer.isnull().values.any())
```

```
False
False
False
False
```

Figure 3

There are a number of instances containing 0 values for trip distance and tip amount. For the instances has 0 value for trip distance, the pickup time and drop off time are the same. These trips could be recorded because of technical issues of the meter. The driver may have started the meter by accident, or a customer got in the car but decided not to take the trip in the end. Either way, since the response variable of the research is tip amount, these invalid trips do not affect the model therefore are discarded. After discarding these instances, 19946 instances are left for further study. About one third of the total amount of instances are removed for containing missing values.

	tpep_pickup_datetime	tpep_dropoff_datetime	trip_distance	PULocationID	tip_amount
0	2017-12-09 11:14:24	2017-12-09 11:41:01	3.35	90	2.00
1	2017-12-22 21:47:10	2017-12-22 21:53:45	1.63	166	0.00
2	2017-12-09 14:56:10	2017-12-09 14:58:38	0.63	24	0.72
3	2017-12-29 23:29:48	2017-12-29 23:34:25	0.76	142	0.00
4	2017-12-02 13:57:34	2017-12-02 14:00:24	0.50	43	1.00

Figure 4. Taxi data before cleaning and pre-processing

Figure 4 shows an example of how the original datasets look like after attribute selection and before data cleaning and pre-processing.

	avg_speed	pickup_hour	Minimum Temperature	Maximum Temperature	Temperature	tip_amount
0	14.045184	12	16.61	26.56	21.44	2.00
1	12.874752	14	16.61	26.56	21.44	1.00
2	17.059046	9	16.61	26.56	21.44	1.06
3	20.189952	23	16.61	26.56	21.44	2.80
4	6.061862	14	16.61	26.56	21.44	1.28

Figure 5. Taxi data ready for modelling and analysis

Figure 5 presents the final datasets after cleaning and pre-processing, used for modelling and analysing. Each row represents data of one trip.

The average speed is calculated by trip distance divided by the duration for the trip. The trip distance in the original dataset is of unit miles, it is converted into kilometre for easier understanding. The unit for average speed in the final datasets is kilometre per hour. The pickup hour is the hour of the day when the trip started. For example, 2 means 2 a.m., and 22 means 10 p.m. . The minimum, maximum and average temperature are picked and merged from the weather dataset (https://www.visualcrossing.com, 2019). The original unit for the temperatures is Fahrenheit, and it is now converted into Celsius. The datatype for Date attribute in the original weather datasets are converted to Datetime in order to merge with the taxi datasets. As the response variable, the tip amount is of American dollars.

While checking for outliers of the datasets, it is found that some trips have unrealistically large average speeds. Since this is the average driving speed in a busy modern city, it should not be able to reach above 100 km/h. Furthermore, the durations for the trips are only a few seconds. These data points are considered as collection error and therefore removed from the datasets.

	Pickup_Date	trip_distance_km	duration_seconds	avg_speed	pickup_hour	Minimum.Temperature	Maximum.Temperature	Temperature	tip_amount	PULocationID
3435	2018-07-30	0.6437376	1	2317.45536	15	20.06	25.67	23.00	10.16	129
2798	2018-07-19	1.2874752	3	1544.97024	16	19.33	27.83	23.22	1.50	264
4144	2018-08-12	0.1126541	3	135.18490	21	23.33	30.56	25.78	17.79	264
1301	2018-06-21	0.4345229	14	111.73445	22	18.83	30.00	23.94	10.00	238

Figure 6 Examples of error instances

5. Basic Statistics

```
tip_amount              1.000000
avg_speed               0.504963
pickup_hour             0.014958
Maximum Temperature     0.006814
Temperature             0.006611
Minimum Temperature     0.003456
Name: tip_amount, dtype: float64
```

Figure 7 Summer dataset Correlations with Tip amount

```
tip_amount              1.000000
avg_speed               0.475685
pickup_hour             0.018606
Maximum Temperature     0.001419
Temperature             0.000638
Minimum Temperature    -0.004317
Name: tip_amount, dtype: float64
```

Figure 8 Winter dataset Correlations with Tip amount

```
tip_amount              1.000000
avg_speed               0.489514
pickup_hour             0.016924
Maximum Temperature     0.015788
Temperature             0.015712
Minimum Temperature     0.014285
Name: tip_amount, dtype: float64
```

Figure 9 Total dataset Correlations to Tip amount

Correlations between predictors and the tip amount for summer, winter and total datasets are produced(Figure 7-9). The correlation coefficient has the range between -1 and 1, and a larger absolute value implies stronger correlation. It appears that the dataset contains both summer and winter data during 2017 and 2018 has more attributes correlated to the tip amount, namely the temperature-related attributes. This could be caused by the wider range of temperature attributes. The datasets are split originally by season to better investigate the temperature attributes, but the statistics show that it is better to combine those, so the total dataset is used from here. The average speed attribute has the strongest positive relation with tip amount.

	avg_speed	pickup_hour	Minimum Temperature	Maximum Temperature	Temperature	tip_amount
count	19932.000000	19932.000000	19932.000000	19932.000000	19932.000000	19932.000000
mean	18.658531	13.807144	11.132528	18.212428	14.419585	2.803980
std	9.787132	6.223783	10.707046	11.438420	10.931711	2.460445
min	0.040520	0.000000	-14.390000	-6.500000	-10.720000	0.010000
25%	12.133274	9.000000	1.670000	7.220000	4.330000	1.450000
50%	16.419974	14.000000	14.330000	21.110000	17.610000	2.000000
75%	22.402652	19.000000	21.110000	28.780000	24.720000	3.060000
max	84.784952	23.000000	26.670000	35.610000	30.610000	32.850000

Figure 10 Descriptive Statistics

From the table(figure 10) shown above, the range of attributes varies. The standard deviations for temperature attributes and tip amount are fairly large considering the means. Data transformation(standardization) before model training is required.

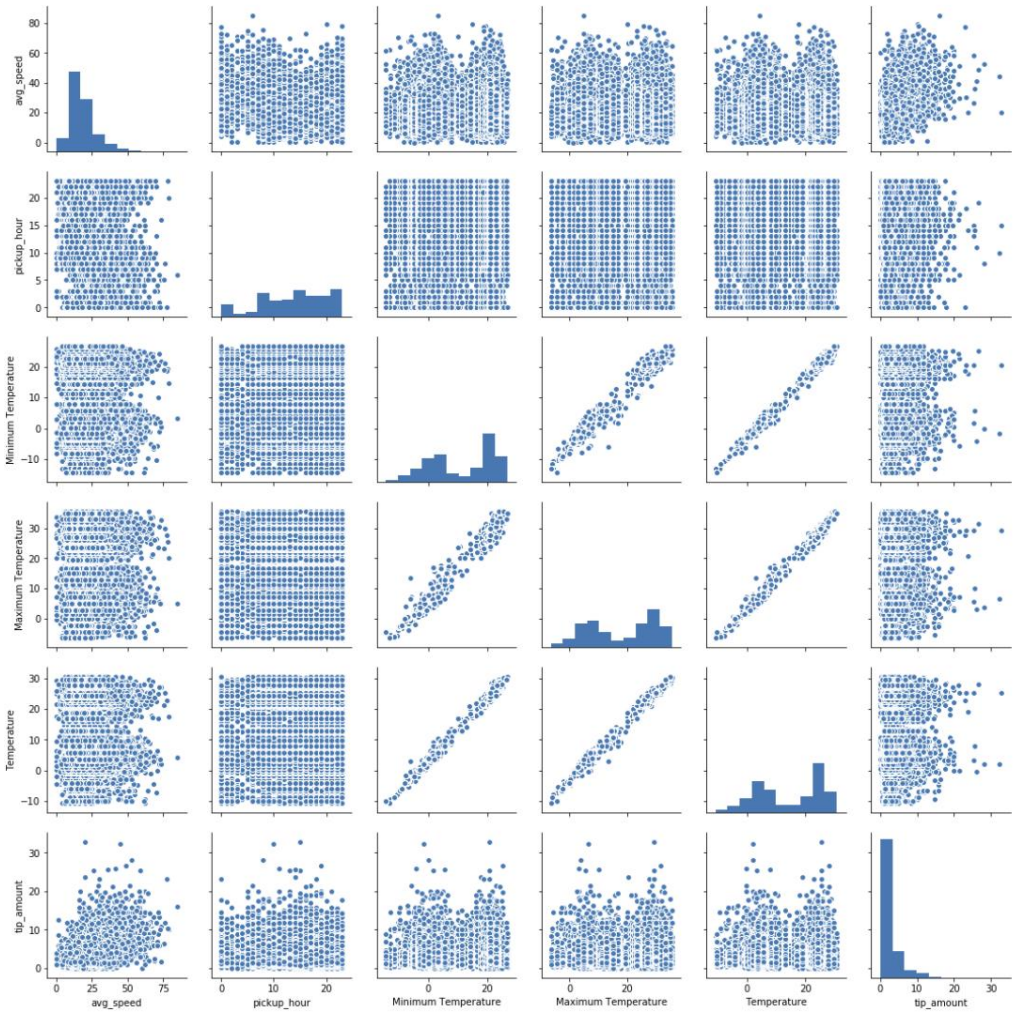

Figure 11 Pairwise Scatter Matrix

From the initial pairwise visualisation, five data points look like they do not belong to the same distribution as the majority of data. After removing these outliers from the data, the relation is better presented (Figure 11). The tip amount and average speed seems to have an obvious relation, while their histograms are also both heavily right skewed. The scatter plot of tip amount versus average speed shows heteroscedasticity, which is a sign of requirement for data transformation(take log). Three temperature attributes look bimodal, which implies the data is from two different populations. In this case it is from summer and winter.

6. Model Initialization, Refinement and Selection

The dataset that combines both winter and summer is further randomly split into training set and test set. To minimize bias, stratified sample method is used. That is, the dataset is split by the proportion of winter and summer data. The training set contains 80% of the dataset instances, and the test set contains 20%.

```
#summer
s_train_set, s_test_set = train_test_split(df.iloc[0:10314, : ], test_size=0.2, random_state=42)

#winter
w_train_set, w_test_set = train_test_split(df.iloc[10314:19932, : ], test_size=0.2, random_state=42)

train = s_train_set.append(w_train_set,ignore_index=True)
test = s_test_set.append(w_test_set,ignore_index=True)
```

Figure 12 Splitting training and test sets

Firstly, using Stepwise Selection method, an additive model is built(Figure 13).

```
train <- read.csv('train.csv')

t_model0 <- lm(tip_amount ~ 1, data = train)
t_model1 <- step(t_model0, scope = ~ . + avg_speed + pickup_hour + Mini
mum.Temperature + Maximum.Temperature + Temperature)
```

Figure 13 Stepwise Selection

```
## Step:  AIC=24314.63
## tip_amount ~ avg_speed + pickup_hour + Maximum.Temperature
##
##                        Df Sum of Sq   RSS   AIC
## <none>                              73226 24315
## + Minimum.Temperature   1       2.7 73223 24316
## + Temperature           1       0.4 73226 24317
## - Maximum.Temperature   1      19.6 73246 24317
## - pickup_hour           1     318.8 73545 24382
## - avg_speed             1   22620.0 95846 28605
```

Figure 14 Additive model Built by Stepwise Selection

Stepwise Selection starts from any initial model (in this case from null model) and compares the Akaike information criterion (AIC) of models built with either adding or subtracting a predictor. As shown above (Figure 14), the additive model using stepwise selection contains three predictors: average speed, pick up hour and maximum temperature. The final AIC value is 24314.63.

However, the interactions between predictors are not considered in an additive model. A model that contains all the pairwise interaction terms is then produced and compared with the additive model(Figure 15).

```
t_model2 <- lm(tip_amount ~ (avg_speed + pickup_hour + Maximum.Temperat
ure)^2, data = train)


anova(t_model1, t_model2)

## Analysis of Variance Table
##
## Model 1: tip_amount ~ avg_speed + pickup_hour + Maximum.Temperature
## Model 2: tip_amount ~ (avg_speed + pickup_hour + Maximum.Temperature
)^2
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1  15941 73226
## 2  15938 73092  3    134.33 9.7638 1.969e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 15 Comparison of Additive Model and Interaction Model

From Figure 15, the 'Pr(>F)' term in the ANOVA output is the p-value for F test. Considering with a 95% confidence interval, the p-value is far less than 5% which is the critical region. Therefore the null hypothesis all interaction terms are irrelevant is rejected. It shows that there are some significant interaction terms.

To investigate which terms have a significant interaction with each other, a Stepwise Selection is further applied to the interaction model. As shown in Figure 16 below, among all interaction terms, only the average speed and the pickup hour has an interaction that's significant to the model. The AIC is also reduced from 24314.63 from the additive model to 24289.44. If the Stepwise Selection starts from the pairwise interaction model that contains all the terms (including minimum temperature and average temperature), the final model appears to be exactly the same as Figure 16.

```
## Step:  AIC=24289.44
## tip_amount ~ avg_speed + pickup_hour + Maximum.Temperature +
##     avg_speed:pickup_hour
##
##                          Df Sum of Sq   RSS   AIC
## <none>                                 73101 24289
## - Maximum.Temperature    1    20.119 73121 24292
## - avg_speed:pickup_hour  1   124.760 73226 24315
```

Figure 16 Final Model

Overall, the final model selected contains predictors average speed, pickup hour, maximum temperature and the interaction term between average speed and pickup hour.

7. Analysis of the Final Model

```
summary(t_model3)

##
## Call:
## lm(formula = tip_amount ~ avg_speed + pickup_hour + Maximum.Temperat
ure +
##     avg_speed:pickup_hour, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6401 -1.2043 -0.3768  0.6563 29.7792
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.5438963  0.0965799   5.632 1.82e-08 ***
## avg_speed               0.1048986  0.0037786  27.761  < 2e-16 ***
## pickup_hour            -0.0065915  0.0062738  -1.051   0.2934
## Maximum.Temperature     0.0031051  0.0014825   2.095   0.0362 *
## avg_speed:pickup_hour   0.0013558  0.0002599   5.216 1.85e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.141 on 15940 degrees of freedom
## Multiple R-squared:  0.2377, Adjusted R-squared:  0.2375
## F-statistic:  1242 on 4 and 15940 DF,  p-value: < 2.2e-16
```

Figure 17 Summary of the Final Model

As shown above in Figure 17, average speed has the smallest p-value for t test, therefore is the most significant and relevant predictor. The estimate for average speed is a positive number, which implies a positive relation. The pickup hour itself is not significant to the model(p-value larger than 0.05), but the interaction between average speed and pickup hour is important. This could be explained by the different traffic load in different hours. For example, in rush hours there are normally traffic jams in the city area, thus the speed is very low. But at midnight cars can drive close to the limit speed of the streets. To gain more tips, the taxi drivers is suggested to consider taking roads with less traffic and boost the driving speed. Pick up in the hours with less traffic can also be a good strategy to earn more tips per trip, but the pickup numbers could possibly be limited. Maximum temperature also has a small positive relation with tip amount. Passengers seem to be more generous on warmer days in both summer and winter in New York city. The final linear model is tip amount = 0.544 + 0.105 * average speed + 0.003 * maximum temperature + 0.001 * average speed * pickup hour.

```
deviance(t_model3) # residual sum of squares

## [1] 73101.16

deviance(t_model3)/t_model3$df.residual # sample variance

## [1] 4.58602

confint(t_model3)

##                                 2.5 %      97.5 %
## (Intercept)             0.3545888303 0.733203815
## avg_speed               0.0974921072 0.112305081
## pickup_hour            -0.0188887912 0.005705771
## Maximum.Temperature     0.0001992483 0.006011039
## avg_speed:pickup_hour   0.0008462842 0.001865314
```

Figure 18 Deviance and Confidence Intervals of the Final Model

Figure 18 shows the residual sum of squares and sample variance for this model, as well as the 95% confidence interval for the parameters relative to the predictors. The sample variance is large, implying room for improvement.

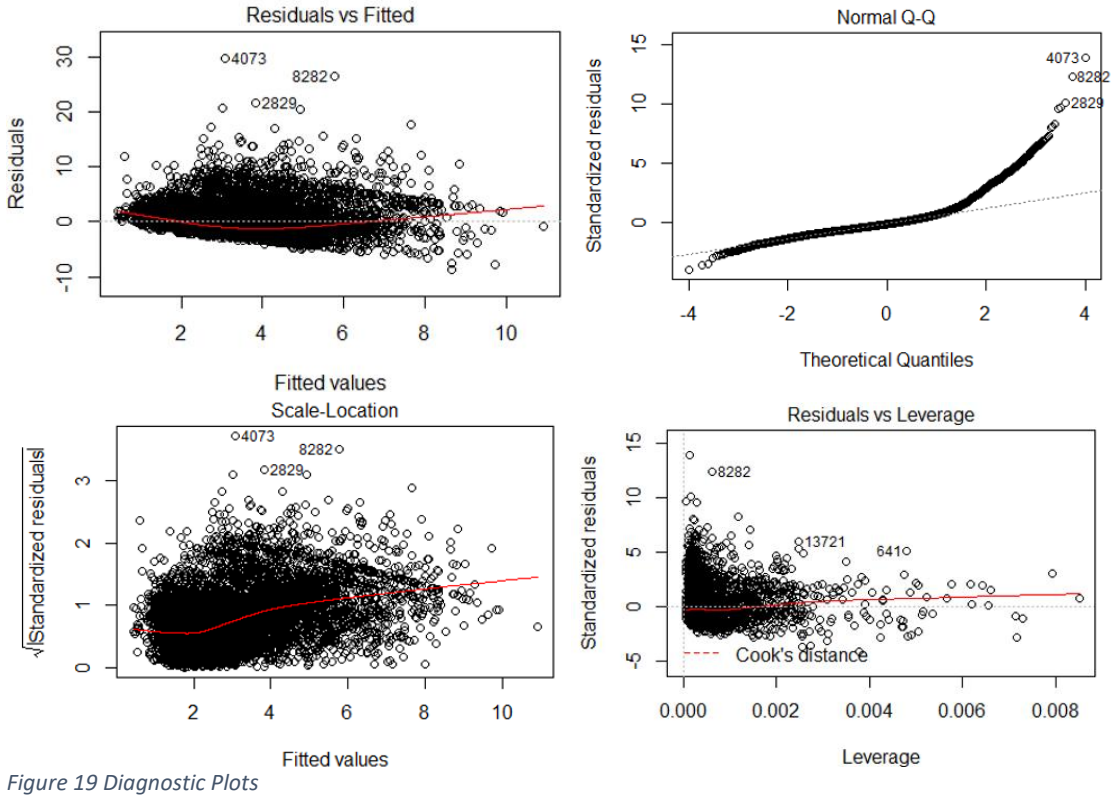

Figure 19 Diagnostic Plots

The diagnostic plots are shown in Figure 19. The Residuals vs. Fitted plot shows a trend of heteroscedasticity. So does the Residuals vs. Leverage plot. For future improvement of the model, data transformation should be considered. Normal Q-Q plot has most of data points following the line, but the right tail is underestimated. There are also 3 outliers, which could be the rare cases when tips are unusually high. If a larger sample is collected, these outliers may not stand out, but rather help model the right tail of the model.

8. Conclusion

In conclusion, the amount of tips paid by credit cards for yellow taxi trips in New York City in summer and winter months in 2017 and 2018 is most (positively) related to the average driving speed, maximum temperature for the day and the interaction between average driving speed and pick up hour. The final linear model is tip amount = 0.544 + 0.105 * average speed + 0.003 * maximum temperature + 0.001 * average speed * pickup hour. There is not any data transformation involved in this modelling process, but it is strongly recommended to perform data transformation for the improvement of this model in the future. It is suggested for the yellow taxi drivers to take roads with less traffic and boost the driving speed in order to gain more tips. Pick up in the hours with less traffic can also be a good strategy to earn more tips per trip, but the pickup numbers could possibly be limited.

Bibliography

Weather Data Services | Visual Crossing. (2019). Retrieved 8 September 2019, from https://www.visualcrossing.com/weather/weather-data-services

Schneider, T. (2019). Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance. Retrieved 8 September 2019, from https://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance/

About TLC - TLC. (2019). Retrieved 8 September 2019, from https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Gron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (1st ed., pp. 65-237). O'Reilly Media, Inc. ©2017.