A Qualitative Analysis of the Influences of Tipping

Patrick Tjahjadi

The University of Melbourne

September 2019

1. Introduction

The culture of tipping has been present in some, if not, most people's lives and is not a recent phenomenon. Simply put, throughout many years of history, many people have been introduced to the notion of tipping for goods and services rendered. There are many factors that may influence not only the tip amount but also one's choice whether to tip (Lynn, Zinkhan & Harris, 1993).

This report is an extension and generalisation of Patrick Tjahjadi's report "An Exploration of the Influences of Taxi Tips in Various Locations Throughout New York City", which described possible attributes that contribute towards tipping amount in New York City taxi trips (Tjahjadi, 2019). The aim of this report is to conduct a deeper analysis of the factors that affect tipping. Specifically, this report goes a step further by providing descriptive statistics and additional attributes that may be of interest. As a result, the scope of this report can be extended not only for taxi trips in New York City but for tipping in general, as attributes from datasets outside of taxi trips are considered.

In order to generalise the scope and to obtain more data, the data for this report will be assessed for the entire year of 2015, compared to only December 2015 in the previous report. It is believed that using the data for an entire year would be more representative and allows more variability and insights. The year of 2015 was chosen because the dataset given before July 2016 contains more detailed attributes. Since there is an annual growth for daily use of for-hire vehicles, the latest year that is feasible for this report, which is 2015, is chosen (Conway, Salon & King, 2018).

2. Datasets

This report uses the following datasets:

1. **"Yellow Taxi Trip Records" 2015:** Contains data about taxi trips in New York City per month by the NYC Taxi and Limousine Commission (TLC). While green taxis are restricted from picking up passengers in southern Manhattan, yellow taxis can pick-up and drop-off anywhere within the city. Hence, yellow taxis are chosen since it allows more location coverage. Link: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
2. **"Borough Labor Force Data":** Contains data about the number of people employed and the unemployment rate between 2014 to 2018 recorded by the New York State Department of Labour. More specifically, it provides unemployment rates per month, separated by each borough of New York City. Link: https://www.labor.ny.gov/stats/nyc/

These datasets were chosen because they are both provided by credible and reliable sources. Moreover, the datasets were first assessed to be deemed suitable for this study. Both datasets are in CSV format.

3. Problem Formulation

This report addresses the following problem: "What are some influences that affect tipping?" Several attributes from the datasets will be used to assess the correlation between them and the tip amount. From initial speculation and Tjahjadi's previous report, it is suggested that location, fare amount and trip distance might play a role in tip amounts in taxi trips. The Borough Labour Force Data was specifically chosen because addressing monetary values would require unemployment rates and current economic conditions to be considered.

Although this report mainly addresses factors that may affect tip amount in taxi trips, by adding a dataset compared to the previous report, it is hoped that this report can gain more information regarding influences of tip amounts, which can be used to generalise the scope of the report to apply to all services. The impact of this report should be twofold. Firstly, the results of this report should provide more insights for service-oriented companies on which factors would affect the tip amount. Moreover, they would be able to focus and improve on certain factors which play a huge role in tip amounts which they still might have lacked.

4. Data Pre-processing

The pre-processing phase is done so that only relevant data for this report will be extracted. The pre-processing phase is done in Microsoft Excel and Python and involves several filtering methods. This phase also includes integrating the two datasets into one so further processing can be done.

As the datasets are in CSV format, they are opened in Excel and screened for any missing data, outliers and formatting used for highlighting or beautification (different text colour, font size, etc.). Outliers are left unchanged since they may provide interesting information. These outliers are not deemed erroneous since the datasets are extracted from credible sources. There are, however, outliers that may be removed in the cleansing phase, mostly because they are illogical or inaccurate.

One drawback of Excel worth mentioning is that it is unable to display all the data due to the dataset's large size. Pre-processing data solely using Excel would incur losses in relevant data. Therefore, Python is used to pre-process the data beyond checking for highlighting or beautification.

In order to increase the scope of the data, sampling was done using Python for all 12 months at random. Random sampling would eliminate bias and allow all entries equal chance to be chosen. One million entries of each dataset are randomly extracted. These samples are then combined into one single dataset to be analysed. Due to the limits of available computational power, sampling exceedingly more entries would cause a MemoryError in Python.

Since the tip_amount attribute only contains tips for credit card payments, this report only extracts entries where taxi fare payments are made by credit card (payment type 1).

The Borough Labour Force dataset contains the number of unemployed people per month. In order to integrate it with the Yellow Taxi Trip Records dataset, the unemployment rates of each borough are averaged by month. These averages are calculated by dividing the number of unemployed people with the number of labour force of the 5 boroughs combined. These unemployment rates are then extracted and then put into the Trip Records dataset.

All in all, the pre-processing phase left the amount of data entries from 146112989 to 7520102 entries.

Note that the reason why data sampling is performed before cleaning is so that the cleansing method can be tailored to the samples. This report also assumes due to the many entries being sampled; the samples taken after the cleansing method would also be sufficiently representative of the data.

5. Data Cleansing

The cleansing method is done in Microsoft Excel and Python and involves removing data entries that are considered inaccurate or corrupted. The cleansing method is similar to Tjahjadi's previous report.

Firstly, data entries that do not provide pickup or drop-off points (missing or zero latitude and/or altitude) are removed because the credibility of these trips is unclear.

Also, there are data entries where there are zero passengers on a taxi trip. These entries will be removed since it is unclear whether it is a valid trip with an undetermined number of passengers or falsely recording free-roaming as a taxi trip.

Moreover, taxi trips with distances less than 0.1 miles or duration of 0 (pickup and drop-off at the same minute) are excluded since they would provide little to no value regarding tipping behaviour among passengers. Taxi trips with negative duration are also removed. Notably, the trip records in November has drop-off times that are earlier than their pickup time.

Finally, entries with zero base fare amounts are removed since it is illogical and would only skew the analysis.

According to the TLC ("Taxi Fare", 2019), there exist a 50 cents MTA State Surcharge for trips that end in New York City along with a 30 cents Improvement Surcharge. Since this report only considers taxi trips within the scope of New York City, taxi trips that do not impose these surcharges are removed. More information of the fares is provided in Figure 1.

- **$2.50** initial charge.
- Plus **50 cents** per 1/5 mile when traveling above 12mph or per 60 seconds in slow traffic or when the vehicle is stopped.
- Plus **50 cents** MTA State Surcharge for all trips that end in New York City or Nassau, Suffolk, Westchester, Rockland, Dutchess, Orange or Putnam Counties.
- Plus **30 cents** Improvement Surcharge.
- Plus **50 cents** overnight surcharge 8pm to 6am.
- Plus **$1.00** rush hour surcharge from 4pm to 8pm on weekdays, excluding holidays.
- Plus New York State Congestion Surcharge of **$2.50** (Yellow Taxi) or **$2.75** (Green Taxi and FHV) or **75 cents** (any shared ride) for all trips that begin, end or pass through Manhattan south of 96th Street.
- Plus tips and any tolls.
- There is no charge for extra passengers, luggage or bags, or paying by credit card.
- The on-screen rate message should read: "Rate #01 – Standard City Rate."
- Make sure to always take your receipt.

Figure 1: Official taxi fare breakdown imposed by the TLC. This excludes tolls, airport trips and destinations beyond New York City ("Taxi Fare", 2019).

All aspects considered; the cleansing phase left the amount of data entries from 7520102 to 7362809 entries.

6. Analysis

6.1 Procedures

The pre-processing phase and cleansing phase transform the datasets into a single dataset that is ready to be analysed. The Labour Force dataset is put into the Trip Records dataset. Afterwards, the datasets for each month is merged into one dataset using Python. The following dataset attributes will be considered:

1. fare_amount: The fare calculated by the taxi meter. This is used to assess whether the passenger tips a considerable amount for a particular trip.
2. tip_amount: The amount of tip given by the passenger at the end of the trip. Cash tips are not included in this data.
3. trip_distance: The distance reported by the taximeter in miles.
4. passenger_count: The number of passengers recorded within a taxi trip.
5. Duration: The difference in time between pickup and dropoff time in minutes.
6. Unemployment_rate: The rate of unemployment recorded from the Borough Labour Force dataset.

The first four attributes are originally found in the Trip Records dataset. The Duration attribute is created by subtracting drop-off and pickup time while the unemployment rate is the average rate of unemployment in New York City for a particular month found in the Labour Force dataset. Note that in this report, the term variables and attributes can be used interchangeably.

At first glance, these attributes are speculated to have a relation towards the amount of tip given in a taxi trip. Moreover, all the variables considered for the model are numerical, which makes it easier to perform regression.

6.2 Preliminary Analysis

It is necessary before assessing the relation between attributes and creating a mathematical model that a summary of the data is provided. This report presents brief descriptive coefficients that summarises the dataset using descriptive statistics.

```
> summary(data$tip_amount)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    1.35    2.00    2.69    3.06  850.00
> summary(data$fare_amount)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.04    7.00   10.00   13.30   15.50  810.10
> summary(data$Unemployment_rate)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.980   5.230   5.440   5.548   5.780   6.620
> summary(data$passenger_count)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.664   2.000   9.000
> mean(data$trip_distance)
[1] 7.114197

> summary(data$Duration)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      1       7      12      16      19  329040
>
```

Figure 2: Descriptive statistics of attributes.

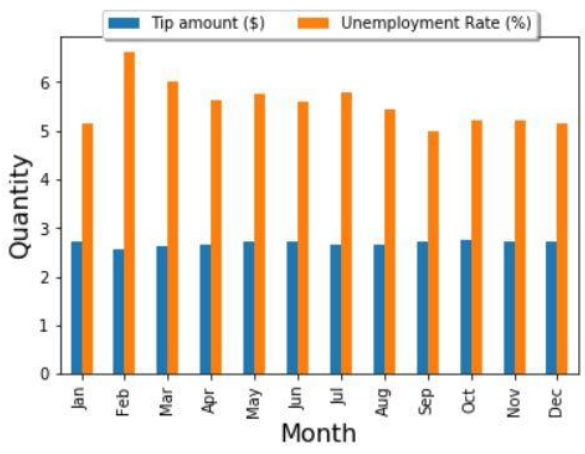
The descriptive statistics in Figure 2 shows the minimum value, 1st quartile, median, mean, 3rd quartile and maximum value of the attributes that are considered for analysis. The descriptive statistics was calculated in R. However, there are some limitations of this analysis:

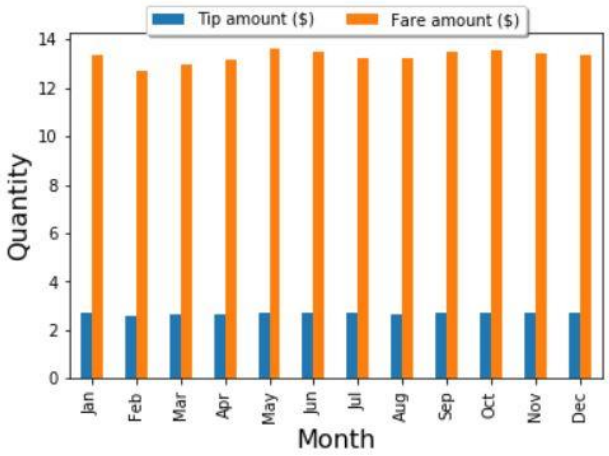
1. The summary function is only able to show trip distance in the form of integers. Hence, it is unable to present trip distance accurately.
2. Measurements aren't presented in detail in the R summary. For example, tip amount, fare amounts are in dollars, trip distance is in miles while unemployment rate is in percentages.
3. Outliers exist in the data, such as a trip giving $850 for tip amount.

```
> var(data$tip_amount)
[1] 6.517884
```

Figure 3: Variance for tip amount.

Figure 3 describes the distribution of tip amount in detail by introducing its variance. It shows how spread out the data is from the mean (Wilson & Kalla, 2009). Although there is no fixed standard on what amount of variance is high, it can be concluded that the variance for tip amount is moderate as it is 2-3 times the mean. This indicates that the distribution of tip amount is moderately spread out.





Figures 4 and 5: Bar charts of average tip amounts against unemployment rate and fare amount per month.

Figures 4 and 5 shows the means of tip amount against unemployment rate and fare amount by month in 2015 from the combined dataset, generated by Python. In particular, there is no trend or major changes in the tip amount over the months. However, higher unemployment rates suggest lower tipping amounts while a higher base fare amount indicates a higher amount of tip. A more thorough discussion of the relationship of the attributes are presented in the correlation analysis section.

6.3 Correlation Analysis

In order to verify the relationship of the 5 predictor variables against tip amount, a correlation analysis is performed. The aim of the correlation analysis is to identify relationships among the attributes and the most important attributes to consider, that is, the predictor variables that contribute the most to tip amount.

In this report, the Pearson correlation coefficient (r) formula is applied:

$$r = \frac{\sum(x - \bar{x})\sum(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

Equation 1: Pearson's Correlation Coefficient

In Equation 1, x and y are two vectors of length n, with $\bar{x}$ and $\bar{y}$ being the means of x and y, respectively. The Pearson correlation coefficient measures the relationship between two variables, ranging from -1 (negative correlation) to 1 (positive correlation) (Bolboaca & Jäntschi, 2006).

The Pearson correlation coefficient is calculated in R for all 5 predictor variables as follows:

```
> library("ggpubr")
> cor(data$tip_amount, data$fare_amount, method = "pearson")
[1] 0.7550995
> cor(data$tip_amount, data$Duration, method = "pearson")
[1] 0.05403317
> cor(data$tip_amount, data$Unemployment_rate, method = "pearson")
[1] -0.01779473
> cor(data$tip_amount, data$trip_distance, method = "pearson")
[1] 0.0008123353
> cor(data$tip_amount, data$passenger_count, method = "pearson")
[1] 0.008196365
```

Figure 5: Pearson correlation coefficient for predictor variables against tip amount in R

Figure 5 shows that the strongest and most obvious predictor variable against tip amount is the base fare amount, showing a strong positive relationship. The trip duration, distance and number of passengers also exhibit a positive relationship against tip amount, albeit being relatively weak relationships. On the contrary, the rate of unemployment shows a negative relationship against tip amount.

To identify and rank the most important attributes, however, it is imperative to assess the causal effects not found from a simple correlation analysis. To address this issue, a linear model was created, and a regression analysis is performed. The model analysis will determine the significance of each variable in modelling the tip amount.

6.4 Model Relevance

Firstly, a linear model is created in R in order to determine which variables are deemed as good fits to model tip amount. Since there are 7362809 samples, it is safe to assume by the central limit theorem that the data is normally distributed (Aberson, Berger, Healy, Kyle & Romero, 2000). Hence, the mathematical model is formed with the linear equation:

$$y = X\boldsymbol{\beta} + \varepsilon$$

Equation 2: Linear regression equation

In Equation 2, y is the dependent variable and X is the explanatory variable(s). $\boldsymbol{\beta}$ indicates the slope coefficient while ε shows the associated error term.

Without loss of generality, a linear model is created with the assumption that the error term ε in the model has a multivariate normal distribution with mean vector $\mathbf{0}$ and the identity covariance matrix (Peña & Slate, 2006). Apparently, the relationship between the variables are linear and should show signs of homoskedasticity.

```
Call:
lm(formula = data$tip_amount ~ data$fare_amount + data$passenger_count +
    data$Duration + data$Unemployment_rate + data$trip_distance)

Residuals:
   Min     1Q Median     3Q    Max
-89.86  -0.29   0.13   0.33 840.05

Coefficients:
                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)             2.938e-01  7.838e-03   37.480   <2e-16 ***
data$fare_amount        1.876e-01  6.020e-05 3116.310   <2e-16 ***
data$passenger_count   -7.427e-04  4.639e-04   -1.601    0.109
data$Duration          -9.088e-02  6.740e-03  -13.483   <2e-16 ***
data$Unemployment_rate -1.758e-02  1.392e-03  -12.634   <2e-16 ***
data$trip_distance      1.878e-08  1.234e-07    0.152    0.879
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.674 on 7362803 degrees of freedom
Multiple R-squared:  0.5702,    Adjusted R-squared:  0.5702
F-statistic: 1.954e+06 on 5 and 7362803 DF,  p-value: < 2.2e-16
```

Figure 6: Summary of the linear model for tip amount in R.

The model suggests that the base fare, passenger count, trip duration and unemployment rate contribute to the amount of tip given by passengers in a taxi trip. There are several aspects to note, however. Firstly, the trip duration, passenger count and unemployment rate have negative coefficients as shown by their t-value (i.e. the higher these three variables, the lesser the tip amount). Also, although the number of passengers in a taxi trip contributes to the tip amount, its effects are not as significant as shown by its relatively small t-value. Moreover, the trip distance does not seem to affect the tip amount, with its very small t-value and high p-value, denoting insignificance. Since a higher t-value indicates a better fit to the model and the null hypotheses for relevance can be rejected, those 4 aforementioned attributes (i.e. base fare, passenger count, trip duration and unemployment rate) are considered.

It should be mentioned that although passenger count and duration exhibit positive correlation against the tip amount, they have negative regression coefficients. This is because regression gives coefficients while controlling for other variables, and the marginal effects of passenger count and duration are surpassed by base fare amount. A mere correlation analysis does not take into account other variables and would not be sufficient to conclude the relationships of the model.

While the model in Figure 6 measures tip amount, it is not parsimonious. The model should be refined so that only significant variables are considered to model tip amount. This can be done by performing stepwise selection and determine which attributes to remove from the model.

```
> model2 = step(model)
Start:  AIC=7584628
data$tip_amount ~ data$fare_amount + data$passenger_count + data$Duration +
    data$Unemployment_rate + data$trip_distance

                        Df Sum of Sq      RSS      AIC
- data$trip_distance     1         0 20626299  7584626
<none>                               20626299  7584628
- data$passenger_count   1         7 20626306  7584629
- data$Unemployment_rate 1       447 20626746  7584786
- data$Duration          1       509 20626808  7584808
- data$fare_amount       1  27205679 47831978 13777688

Step:  AIC=7584626
data$tip_amount ~ data$fare_amount + data$passenger_count + data$Duration +
    data$Unemployment_rate

                        Df Sum of Sq      RSS      AIC
<none>                               20626299  7584626
- data$passenger_count   1         7 20626306  7584627
- data$Unemployment_rate 1       447 20626746  7584784
- data$Duration          1       509 20626808  7584806
- data$fare_amount       1  27205711 47832010 13777691
```

Figure 7: Stepwise selection to create a parsimonious model

```
Call:
lm(formula = data$tip_amount ~ data$fare_amount + data$passenger_count +
    data$Duration + data$Unemployment_rate)

Residuals:
   Min     1Q Median     3Q    Max
-89.86  -0.29   0.13   0.33 840.05

Coefficients:
                        Estimate Std. Error  t value Pr(>|t|)
(Intercept)             0.2937827  0.0078385   37.480   <2e-16 ***
data$fare_amount        0.1876080  0.0000602 3116.313   <2e-16 ***
data$passenger_count   -0.0007427  0.0004639   -1.601    0.109
data$Duration          -0.0908791  0.0067405  -13.483   <2e-16 ***
data$Unemployment_rate -0.0175811  0.0013915  -12.634   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.674 on 7362804 degrees of freedom
Multiple R-squared:  0.5702,    Adjusted R-squared:  0.5702
F-statistic: 2.442e+06 on 4 and 7362804 DF,  p-value: < 2.2e-16
```

Figure 8: Summary of the parsimonious linear model for tip amount in R.

A parsimonious model has the optimal number of predictor variables to model the response variable (Vandekerckhove, Matzke & Wagenmakers, 2015). Here, since trip distance is insignificant, it is removed from the model.

Ultimately, base fare is the most important attribute to model tip amount, followed by duration, unemployment rate, passenger count and lastly, trip distance.

Figure 6 suggests that the equation to model the tip amount can be found by the following equation:

$$Tip\ Amount = 0.29378 + 0.187608 * Fare\ Amount - 0.0908791 * Duration - 0.0175811 * Unemployment\ Rate$$

Equation 3: Linear regression equation to model tip amount

Equation 3 is modelled from estimating the slope coefficients and intercept in Figure 6. This is modelled with a significance level of 5%, meaning that only variables with p-values less than 0.05 are considered statistically significant and put into the equation. If the equation is modelled with a significance level of 11% instead, then passenger count would also be put into the equation.

6.5 Interaction Model

In order to expand the analysis for relationships among the variables in the model, interaction variables are introduced. Interaction suggests that the effect of one predictor variable also depends on another predictor variable (Altman & Bland, 2003).

```
Call:
lm(formula = data$tip_amount ~ data$fare_amount + data$passenger_count +
    data$Duration + data$Unemployment_rate + data$trip_distance +
    data$fare_amount:data$passenger_count + data$fare_amount:data$Duration +
    data$fare_amount:data$Unemployment_rate + data$fare_amount:data$trip_distance +
    data$passenger_count:data$Duration + data$passenger_count:data$Unemployment_rate +
    data$Duration:data$Unemployment_rate + data$Unemployment_rate:data$trip_distance)

Residuals:
   Min     1Q Median     3Q    Max
-90.50  -0.29   0.13   0.33 840.07

Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  2.423e-01  1.597e-02  15.171  < 2e-16 ***
data$fare_amount                             1.913e-01  7.926e-04 241.353  < 2e-16 ***
data$passenger_count                         6.307e-03  5.861e-03   1.076   0.2819
data$Duration                               -1.634e+00  2.331e-01  -7.012 2.35e-12 ***
data$Unemployment_rate                      -6.069e-03  2.861e-03  -2.121   0.0339 *
data$trip_distance                           1.213e-05  7.404e-06   1.638   0.1014
data$fare_amount:data$passenger_count        3.890e-04  4.604e-05   8.450  < 2e-16 ***
data$fare_amount:data$Duration               2.128e-02  1.072e-03  19.838  < 2e-16 ***
data$fare_amount:data$Unemployment_rate     -8.279e-04  1.421e-04  -5.828 5.61e-09 ***
data$fare_amount:data$trip_distance          3.527e-08  1.941e-08   1.817   0.0692 .
data$passenger_count:data$Duration           -1.560e-01  1.195e-02 -13.048  < 2e-16 ***
data$passenger_count:data$Unemployment_rate -1.857e-03  1.045e-03  -1.778   0.0754 .
data$Duration:data$Unemployment_rate         1.893e-01  4.161e-02   4.549 5.38e-06 ***
data$Unemployment_rate:data$trip_distance   -2.238e-06  1.335e-06  -1.676   0.0937 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.674 on 7362795 degrees of freedom
Multiple R-squared:  0.5703,    Adjusted R-squared:  0.5703
F-statistic: 7.516e+05 on 13 and 7362795 DF,  p-value: < 2.2e-16
```

Figure 9: Summary of the parsimonious interaction model for tip amount in R.

By introducing interaction between predictor variables, the model can provide better estimates for the tip amount. This is considered as the final model. The equation for this model is similar to Equation 3, by adding the estimates of the intercept, slope coefficients and interaction between predictor variables.

7. Reflections

There are several unanticipated findings or problems encountered in this study:

1. There exists a notable number of missing or inaccurate values in the dataset that makes it difficult to detect outliers or provide a more accurate representation of the visualisation and the report as a whole. Moreover, it is hard to determine the quantitative limits in which outliers should be removed so they wouldn't skew the analysis drastically.
2. Other factors may exist that affect the number of taxi trips with considerable tip amounts such as age or driver friendliness beyond the scope of the datasets and hence, cannot be inferred.
3. The trip distance variable was significant in Tjahjadi's previous report, but in this report, it does not contribute to the model without interaction significantly. This may be due to more sample sizes, a larger scope of data by including the rest of the months in 2015 or adding more variables into consideration.
4. Continuing from point 3, the trip distance variable is significant when interaction between predictor variables are introduced. It becomes more significant than passenger count. This may be due to the effects trip distance provides to other predictor variables that are significant to model tip amount.

However, some notable improvements compared to Tjahjadi's previous report along with its previous issues were resolved in this report:

1. Adding a new dataset with external variables may contribute to a better model, as shown by considering the unemployment rate of New York City to model the tip amount.
2. Since the pre-processing phase is done in Python and not in Excel, this report bypasses the restriction of size limits only found in Excel. As a result, more samples can be analysed.
3. The previous data pre-processing and cleansing phase were too aggressive, leading to the removal of many valid trips. This report has laxer pre-processing and cleansing conditions. This will result in more valid trips to be included.
4. Instead of taking the first n-rows of the dataset, this report performs sampling on all datasets. The analysis should then be more representative of the actual data.
5. Introducing an interaction model would provide better estimates for tip amount.

8. Conclusion and Future Work

Analysing a model through regression may provide insights towards which predictor variables contribute to the response variable and what extent. Although there is no single factor that becomes the end-all in measuring tip amount, variables such as fare amount, duration and unemployment rate might play a role. Such causal effects would not be sufficient from measuring correlation alone. This report has successfully answered the question: “What are some influences that affect tipping?”

Compared to Tjahjadi’s previous report, there are several improvements and unexpected results that can be found, as noted in the reflections section. Unanticipated differences such as the difference in regression coefficient signs of trip distance are explained why. The descriptive statistics, correlation analysis and mathematical model with variables from another dataset explains in more detail about the factors modelling tip amount.

It is hoped that the results of this report may be used by service-oriented companies to gain insights into the reasons why tips are given that much. The rate of unemployment or the base fare, for instance, may explain the amount of tip given.

Future work would hypothetically demand more computing power to process datasets with a larger scope, such as 10 years of taxi trip records, or more external datasets that may contribute to model tip amount, or other response variables at once so more inferences can be made. There should be a hard limit in which outliers are to be removed for every variable. These outliers and blank entries can then be predicted through machine learning methods. All these methods, however, do not guarantee better performance and goodness of fit. They would need to be experimented hands-on to find significant predictor variables that can model response variables more accurately.

References

- Aberson, C., Berger, D., Healy, M., Kyle, D., & Romero, V. (2000). Evaluation of an Interactive Tutorial for Teaching the Central Limit Theorem. Teaching Of Psychology, 27(4), 289-291. doi: 10.1207/s15328023top2704_08
- Altman, D. G., & Bland, J. M. (2003). Interaction revisited: the difference between two estimates. Bmj, 326(7382), 219.
- Bolboaca, S. D., & Jäntschi, L. (2006). Pearson versus Spearman, Kendall's tau correlation analysis on structure-activity relationships of biologic active compounds. Leonardo Journal of Sciences, 5(9), 179-200.
- Lynn, M., Zinkhan, G., & Harris, J. (1993). Consumer Tipping: A Cross-Country Study. Journal Of Consumer Research, 20(3), 478. doi: 10.1086/209363
- Peña, E., & Slate, E. (2006). Global Validation of Linear Model Assumptions. Journal Of The American Statistical Association, 101(473), 341-354. doi: 10.1198/016214505000000637
- Taxi Fare. (2019). Retrieved 7 August 2019, from https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page
- Tjahjadi, P. (2019). An Exploration of the Influences of Taxi Tips in Various Locations Throughout New York City, 1-9.
- Vandekerckhove, J., Matzke, D., & Wagenmakers, E. J. (2015). Model comparison and the principle of parsimony. Oxford handbook of computational and mathematical psychology, 300-319.
- Wilson, L., & Kalla, S. (2009). Statistical Variance.