# Quantitative Analysis on New York City Yellow Taxis

Factors Impacting The Tippping Behaviour of Customers

Binh (Brian) Duc Vu
Student ID:1053531

August 15, 2021

## 1 Introduction

The Yellow Taxi is an instantly recognizable symbol of New York City, in part due to its iconic colour, but also due to its role in transporting passengers around in an always-buzzing city. For 2018 alone, it recorded more than 100 million trips. The project will involve examining the effect of factors such as employment, income, trip distance, trip duration and time of trip, and then building a model to predict tipping behaviour of customers based on these factors. The investigation thus aims to provide taxi drivers with recommendations for areas of operation to maximise profitability.

## 2 Data Selection

### 2.1 NYC TLC Dataset

The Yellow Taxi dataset chosen is taken from the New York City TLC data (NYC TLC, n.d.). With the source detailing taxi trips consistently since its inception in 2009, tracking more than 10 attributes for each trip, the entire dataset is too large to work with without a more powerful computer. Instead, the 2018 and 2019 sets for Yellow Taxi were taken - 2018 for training and analysing, and 2019 for predicting.

Given the COVID-19 pandemic and its subsequent travel restrictions affecting taxi travel, the 2020 data was excluded from consideration due to its high irregularity. Instead, data from 2018 and 2019 was chosen - as it would be a more realistic reflection of a post-pandemic New York City. The entire data for 2018 and 2019 was used to explore time as a factor in the analysis - which comes to about 100 million instances and 17 attributes for each year.

As the location data is presented in zones, a shapefile of the zones was also collected from the same source - with 263 unique zones.

The Yellow Taxi was chosen as the vehicle of interest over other available choices, like the Green Taxis or the For-Hire/High Volume For-Hire vehicles (NYC TLC, n.d.), due to their full availablility compared to Green Taxis, and consistent trip recordings compared to FHVs (which had a new category introduced in 2019).

### 2.2 Census Data

The employment statistics data is taken from the Census Bureau. The census data is collected by the Census Bureau in census tracts every year, covering a substantial range of categories. With this, the Bureau website allows for filtering of data by topic and geography. From this, the 2017 and 2018 (United States Census Bureau, n.d.-a, n.d.-b) 5-year estimates for selected economic characteristics were selected. Each dataset was of roughly 500 attributes and 2169 rows - for 2169 census tracts. Each topic includes its number and percentage estimates and errors.

Economic statistics due to its nature are often useful when investigating monetary variables. Certain employment attributes have also been noted to be relevant in predicting taxi demand in zones (Correa et al., 2021). Thus, selected economic statistics were chosen as variables to explain tipping behaviours of passengers.

As the census data is collected annually, it is not realistic to predict the 2019 taxi data using 2019 census data (it is unavailable). At such, the 2017 and 2018 census periods were chosen for analysis and predictions. The 5-year estimates were also chosen for each year. The 5-year estimates were also used instead of 1-year due to improved statistical reliability (United States Census Bureau, 2020).

To support merging with taxi zones, a shapefile was for census tracts was also used (NYC, n.d.).

## 3 Preprocessing

### 3.1 Yellow Taxi data

The taxi data collected was accompanied by a data dictionary, which allowed the entries' integrity to be checked easily (NYC TLC, 2018).

1. In order to handle the size of the datasets well, Apache Arrow was used alongside parquet as a serialization format.

2. All datasets were read in using a schema to ensure type consistency, and then serialized into a parquet file to speed up preprocessing and analysis.

3. According to the dictionary, the tip amounts are only recorded for credit card. As such, all other payment types were removed.

4. A count of passenger counts revealed mostly normal passenger counts - except 0, 96 and 192 (Figure 1). The latter two were removed, but for the former, a compelling explanation is the taxi being used to deliver items - which would have no customers but still the features of a normal ride. Since it does not interfere with the analysis, it is retained.

5. The entries with Rate Code ID inconsistent with the dictionary, as well as 3 (Newark Airport) and 4 (Westchester/Nassau County) were removed, as they lie outside New York City (NYC TLC, 2018).

6. A discrepancy between the taxi data and taxi shapefile was fixed by removing any rows which didn't appear in both.

| passenger_count | count |
|---|---|
| 192 | 1 |
| 1 | 73072141 |
| 6 | 2783068 |
| 3 | 4295075 |
| 96 | 1 |
| 5 | 4602861 |
| 9 | 275 |
| 4 | 2029082 |
| 8 | 313 |
| 7 | 390 |
| 2 | 15087976 |
| 0 | 933067 |

Figure 1: Passenger Counts

7. All negative trip distances were removed without question - there is no explanation for why they are negative.

8. Similarly, a trip duration feature was engineered by subtracting drop-off and pickup times - after which all negative trip durations were removed.

9. All trips with incorrectly labelled years were removed.

10. Finally, as tipping amounts are our focus, all other monetary attributes were removed except for fare amounts.

11. The resulting dataset consisted of pickup & drop-off times, locations, tip and fare amounts, trip duration and distances and passenger counts.

### 3.2 Census Data

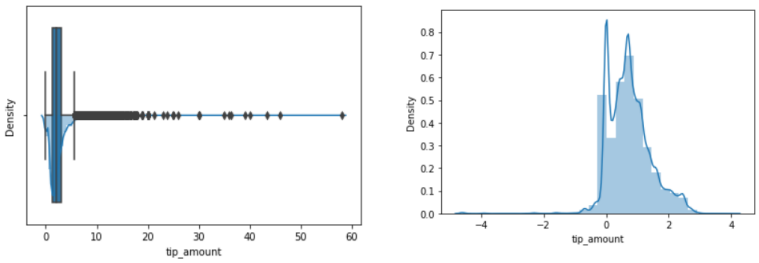
1. Missing/unavailable data was denoted (X) - this was replaced by numpy's NaN to conform to data types.

2. The columns of the census dataset were labelled with codes and incoherent text expressions. As such, the dataset was manually inspected, and all the columns of interest were filtered and renamed. In particular, as the data needs to be aggregated afterwards, number estimates were taken. The attributes collected were that of employment rates and income distribution.

## 4 Exploratory Data Analysis

Other than employment statistics, the time the trip was taken is also an equally interesting variable that may influence tipping behaviour. This stems from common idea like passengers being more generous on the weekend than the working weekday - or that they tip more during holiday seasons. Another detail is that the data will only be grouped by pickup locations and times - as drivers do not know drop-off information ahead of time.

An assumption of independence between attributes will be made - this will allow analysis to be simplified signicantly, given the number of factors considered.

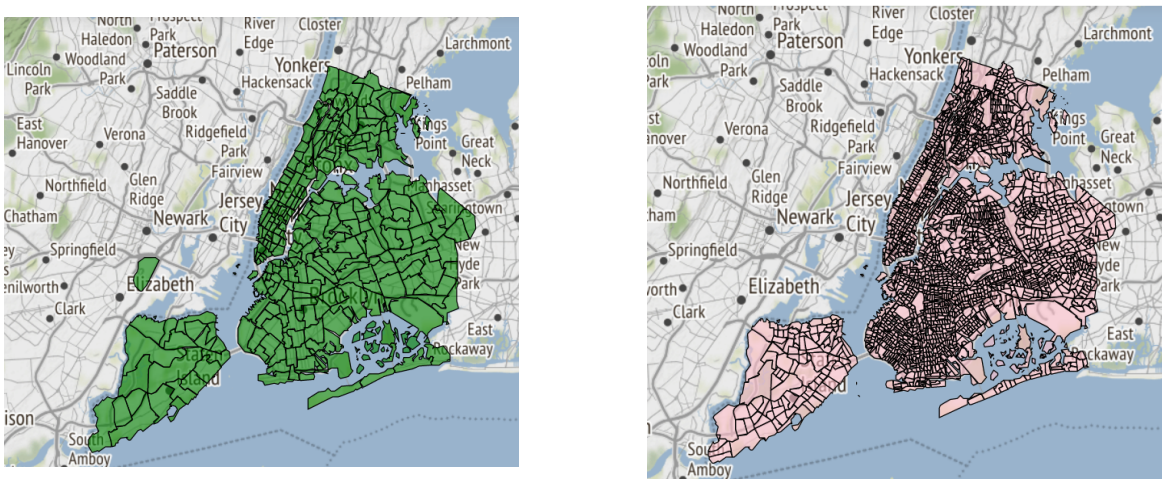



(a) Tip Amount Boxplot (b) log (Tip Amount) Histogram

Figure 2: Tip Amount Visualizations (n = 69938)

### 4.1 Preliminary Analysis

Figures 2a) and 2b) show the distribution of a data sample in a boxplot and a histogram - most tips were concentrated betweeen 0 and 10 dollars - which is expected.
Figure 3 shows the zone availability for the taxi zones and the census tracts, respectively. It can be observed that the census tracts are subsets of the taxi data, and as such the datasets will be merged by merging the census tracts into the taxi zones.




(a) Taxi Zone Availabililty (b) Census Tracts Availability

Figure 3: Zone Availability

Tip amount alone does not provide enough information for the analysis, as it neglects factors like the proportion to fare amount, or the number of trips that go untipped. The untipped trips, in particular, make it difficult to apply a normal approximation to tips (Figure 2b). As such, two formulas describing tipping behvaiour were devised:

1. Non-zero tip amounts as a proportion of the fare:

$$E(\frac{\text{Tip Amount}}{\text{Fare Amount}}) \text{ where Tip Amount} > 0 \tag{1}$$

2. The rate of untipped trips:

$$E(\frac{\text{Untipped Trips}}{\text{Total Trips}}) \tag{2}$$

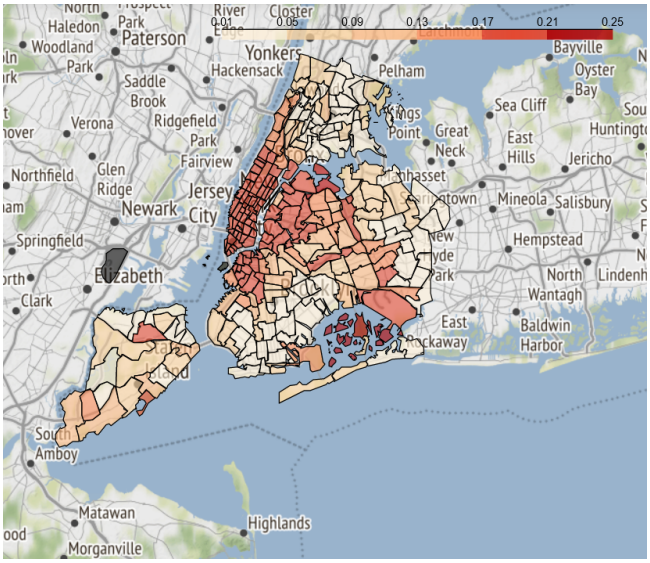
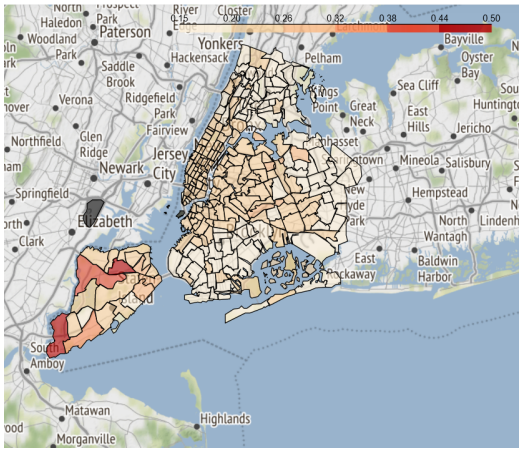
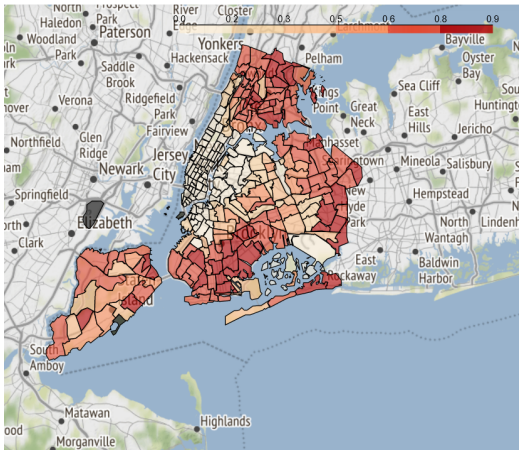



Figure 4: Average Tip Rate including Untipped Trips




(a) Average Tip Rate




(b) Untipped Percentage

Figure 5: Tipping Metrics

Figure 4 represents the average tip proportion for each zone including untipped trips (so the average of (tip amount / fare amount)). Figure 5 represents the metrics described above.

3 small zones unexpectedly missed data - a lookup inspection showed them to be Governor's Island/Ellis Island/Liberty Island labelled together. As the taxi zone shapefile and data came from the same source, errors may be due to faults within the data itself - and thus these zones will simply be ignored.

Some very interesting observations can be made. The most notable of these are that the average tip amount for tipped trips are very consistent across different zones - most of the zones had a tip proportion ranging from 15-25%, which is a standard tipping amount. The anomalies (tip proportion greater than 30%) only had less than 300 trips for the whole year - and thus can be considered outliers.

On the other hand, Figures 4 and 5b) seem to be directly correlated - zones with low rates of untipped trips had correspondingly high average tip rates.

From these observations, it can be seen that whether a trip is tipped or not is actually more important that the specific amount being tipped. Thus, the analysis will be directed towards the proportion of untipped trips as a target variable.

## 4.2 Time as a Factor for Tip Rates

### 4.2.1 Months of a year

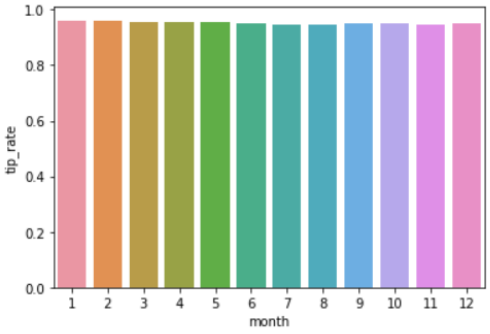



Figure 6: Tip Rate by Month

Figure 6, showing the average tip rate by month, did not report anything interesting - the tip rates were close to each other and consistently high. For this, the variable will not be considered further.

### 4.2.2 Weekends/Weekdays

| | weekend | tip_rate |
|---|---|---|
| 0 | True | 0.947632 |
| 1 | False | 0.954693 |

Figure 7: Tip Rate by Weekend

Figure 7 showed the average for weekends and weekdays, with weekends denoted as 'True'. Similar to Figure 6, it failed to show any trends - the tip rates are almost identical, and thus will not be considered further.

## 4.3 Trip Distance and Duration as a factor for Tip Rate

| | trip_distance | trip_duration | tip_amount | tipped |
|---|---|---|---|---|
| trip_distance | 1.000000 | 0.153955 | 0.763815 | -0.061425 |
| trip_duration | 0.153955 | 1.000000 | 0.124906 | -0.045307 |
| tip_amount | 0.763815 | 0.124906 | 1.000000 | 0.247119 |
| tipped | -0.061425 | -0.045307 | 0.247119 | 1.000000 |

Figure 8: Tip Rate Correlation Matrix

In figure 8, through the correlation matrix, distances and duration appear to have some correlation to tip amounts and tip rates, so they will be added to the prediction model later.

### 4.4 NYC Selected Economic Statistics

In order to merge the census data and the taxi data by zone, several merging operations were done: between census tract and taxi zone shapefiles, between census tract data and shapefile, and finally between census data and taxi data. During these operations, some data was lost due to conflicting dataset entries, even from the same source. This will be seen in some of the plots containing missing values.

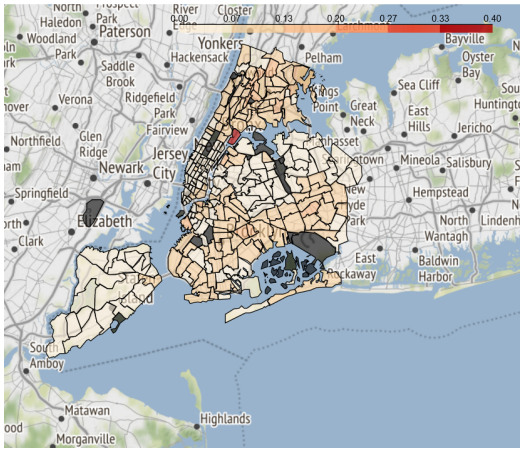



Figure 9: Unemployment Rate by Zone

From figure 9, it can be seen that employment rates are lowest in Manhattan and Staten Island, and highest around the Bronx - showing some similarities to figure 5b). From this visual analysis, these economic factors will be included in the model and its effects more quantitatively explored.

## 5 Statistical Modelling

The model chosen is a Generalized Linear Model, in the binomial family with a logit link. As the output being modelled is the tipping rate, each zone can be assumed to have a Binomial distribution with different parameters. As such, a GLM with Binomial Family is a very suitable choice here.
Before the model was built, predictor variables were normalized to ensure variables were at the same scale. The results of fitting the 2018 taxi data and 2017 census data is shown in figure 10. The x1 and x2 parameters represented trip distance and duration, and the remaining parameters were of employment statistics

Notably, the duration and distance columns had low p-values and thus were statistically significant, whereas for employment data, none of the attributes had any significance. This means that the economic statistics selected do not carry any information regarding the tipping behaviour of passengers.

With these results, a new model was constructed with just the two parameters that had significance.

The new model is detailed in Figure 11. Even though the variables were deemed significant, the R2 value was only at 0.42 - indicating that the variables were not enough to explain the variance of the data.

Afterwards, the model was used to predict the proportion of untipped trips for New York City in 2019 based purely on trip distance and duration - which gave the results of Figure 12. Once again, the

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **x1** | -8.3934 | 2.839 | -2.957 | 0.003 | -13.957 | -2.830 |
| **x2** | 1.9733 | 0.951 | 2.075 | 0.038 | 0.109 | 3.837 |
| **x3** | 0.6435 | 118.879 | 0.005 | 0.996 | -232.355 | 233.642 |
| **x4** | 62.9221 | 191.269 | 0.329 | 0.742 | -311.958 | 437.803 |
| **x5** | -63.3446 | 190.718 | -0.332 | 0.740 | -437.145 | 310.456 |
| **x6** | -62.5114 | 208.028 | -0.300 | 0.764 | -470.238 | 345.215 |
| **x7** | -0.7319 | 199.892 | -0.004 | 0.997 | -392.513 | 391.050 |
| **x8** | -6.8121 | 142.798 | -0.048 | 0.962 | -286.691 | 273.067 |
| **x9** | 7.6449 | 55.149 | 0.139 | 0.890 | -100.444 | 115.734 |
| **x10** | -9.3638 | 119.607 | -0.078 | 0.938 | -243.790 | 225.062 |
| **x11** | 91.4310 | 150.720 | 0.607 | 0.544 | -203.974 | 386.836 |
| **x12** | 125.9136 | 81.305 | 1.549 | 0.121 | -33.441 | 285.268 |
| **x13** | 3.0719 | 116.234 | 0.026 | 0.979 | -224.742 | 230.886 |

Figure 10: Binomial Regression Model Parameters with Y = Tipped Trips

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **x1** | 3.4773 | 0.575 | 6.044 | 0.000 | 2.350 | 4.605 |
| **x2** | -12.9437 | 2.264 | -5.718 | 0.000 | -17.380 | -8.507 |

(a) Parameters of the reduced model

```
r2 score:  0.4253794872119071
mean_squared_error:  0.054743135270763965
```

(b) Performance of the reduced model

Figure 11: Reduced model

scores are unimpressive - it is clear that the taxi data on its own is not enough to explain tipping patterns. However, the performance seems consistent onto predictions, meaning that there was not likely to be any risk of overfitting.

# 6 Evaluation

From the investigation, the most interesting detail noted was the dependence of tip amounts in general on the proportion of untipped trips - and that tipped trips were consistent in quantity across the whole city.

With this, it was observed that operating in Manhattan, parts of Queens and Brooklyn, and JFK Airport, were the best areas for drivers to operate in for maximum tip gain (Figure 4). Other observations were that tip amounts were strongly correlated to distance, but not duration, and also that the times in which people hailed taxis had little effect on tipping behaviour on its own. This is perhaps indicative of duration being an unpredictable variable due to New York traffic.

The results of the model built was disappointing. Despite seemingly following a similar distribution to tip rates, employment rates and other factors had no statistical significance. A possible explanation is taxis mostly being taken by a specific group of people who use it regularly (e.g. office workers), and thus would have been dependent on occupation, not income.

The resulting model's poor performance could either mean that there are more variables that could describe the data better, or that the data was noisy on its own. Given that there were only two

```
r2 score:  0.4400743558240219
mean_squared_error:  0.054247652409150526
```

Figure 12: Accuracy of the Prediction

parameters, the former seems more likely.

# 7 Concluding Remarks

Analysis of several factors affecting the tipping behaviours of New York citizens returned mixed results - as only trip distance emerged as a clear factor, while other factors had little to no effect.

To simplify analysis, the assumption of Independence was made - however, it can be relaxed in further studies for more thorough analysis of the factors.
In further studies, other variables can also be explored - one suggestion made above are the occupation groups by zone - among others.

Finally, prediction involved a fairly simple Binomial Regression model, which seemed suitable for the situation. However, approaches like Ordinal Regression, or Decision Trees can be considered in the future.

## References


[1] Correa, D., Xie, K., & Ozbay, K. (2021). Exploring the Taxi and Uber Demands in New York City - An Empirical Analysis and Spatial Modeling. Figshare.com. https://doi.org/10.6084/m9.figshare.14503002.v4

[2] NYC Open Data. (n.d.). NYC taxi zones. Retrieved August 14, 2021, from https://data.cityofnewyork.us/Transportation/NYC-Taxi-Zones/d3c5-ddgc

[3] NYC. (n.d.). Political and administrative districts - Download and metadata. Retrieved August 16, 2021, from https://www1.nyc.gov/site/planning/data-maps/open-data/districts-download-metadata.page

[4] NYC TLC. (2018, May 1). Data Dictionary - Yellow Taxi Trip Records. Retrieved August 14, 2021, from https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf

[5] NYC TLC. (n.d.). TLC trip record data. Retrieved August 14, 2021, from https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

[6] NYC TLC. (n.d.). Get a vehicle license. Retrieved August 14, 2021, from https://www1.nyc.gov/site/tlc/vehicles/get-a-vehicle-license.page

[7] United Stats Census Bureau. (2020, December 10). American community survey 5-Year data (2009-2019). https://www.census.gov/data/developers/data-sets/acs-5year.html

[8] United States Census Bureau. (n.d.-a). Explore census data. Retrieved August 14, 2021, from https://data.census.gov/cedsci/table?q=employment&t=Employment%20and%20Labor%20Force%20Status%3AOccupation&g=0500000US36005.140000,36047.140000,36061.140000,36081.140000,36085.140000&tid=ACSDP5Y2017.DP03&hidePreview=true

[9] United States Census Bureau. (n.d.-b). Explore census data. Retrieved August 14, 2021, from https://data.census.gov/cedsci/table?q=employment&t=Employment%20and%20Labor%20Force%20Status%3AOccupation&g=0500000US36005.140000,36047.140000,36061.140000,36081.140000,36085.140000&tid=ACSDP5Y2018.DP03&hidePreview=true