An Exploration of the Influences of Taxi Tips in Various Locations Throughout New York City

Patrick Tjahjadi

The University of Melbourne

August 2019

1. Introduction

The culture of tipping is prevalent across the United States and has depended on various factors (Lynn, Zinkhan & Harris, 1993). Although customer satisfaction plays a major role in the amount of tip received, there may be other factors to consider. This report suggests factors that may affect tipping.

The aim of this report is to generate a geospatial visualisation of New York City with a specific focus on the tips of taxi passengers throughout many of the city's locations. Therefore, the aim is to locate points in which part of the city do passengers tend to tip their taxi trips in a considerable amount. With this analysis, this report can determine the degree of factors affecting tip amount, such as the wealth and satisfaction of passengers, overall service of the taxi drivers and the locations where a considerable amount of tipping tends to occur.

With the steady increase of population and tourism in New York City over the past decade, this report studies the location trend of tipping in taxi trips within the city in December 2015 (Kelner, 2008). The year of 2015 was chosen because the dataset given prior to July 2016 contains more detailed attributes, including longitude and latitude of pick-up and drop-off locations, which allows ease of reporting and visualisation. With an annual growth of roughly 0.8% for daily use of for-hire vehicles, the latest detailed data would provide more information regarding taxi trips which is 2015 (Conway, Salon & King, 2018) ("Travel Facts and Figures", 2019). The month of December tends to have more tourists visiting than any other month due to the Christmas and new year season, with a 49.9% share of total overseas arrivals to the United States in December 2015 ("I-94 Arrivals: Monthly-Quarterly-Annual", 2019). This would allow more cognizance to the data due to the tendency of tourists to take taxis.

2. Dataset

This report uses the dataset of the Yellow Taxi Trip Records in December 2015 provided by the New York City Taxi and Limousine Commission (TLC), which is a reliable and credible source. The dataset contains various attributes such as the pickup and drop-off locations and time, trip distance, fare amount and tip amount, which are useful information for the purpose of this report. The link to the dataset can be found from their website:

https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

A full description of the attributes provided by the dataset is available within the data dictionary in the URL:
https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf

Whereas the green taxis aren't allowed to pick up passengers in southern Manhattan, yellow taxis can pick-up and drop-off passengers anywhere within the city (Fischer-Baum, 2015). Therefore, the yellow taxis dataset was chosen since it allows more location coverage.

3. Data Pre-processing

The pre-processing phase is done in Microsoft Excel and involves several filtering methods. Pre-processing is done so only relevant data for this report will be extracted.

Firstly, both CSV files must be opened in Excel and screened for any missing data, outliers and formatting used for highlighting or beautification (different text colour, font size, etc.). As Microsoft Excel is unable to display all the data due to its large size, only the first 5 days of December are used in this report, as it should be fairly representative of the behaviour of taxi trips within the month.

For this report, the following dataset attributes will be considered:

1. dropoff_longitude: The location where the trip ended. Specifically, it refers to the longitude where the taximeter was disengaged.
2. dropoff_latitude: The location where the trip ended. Specifically, it refers to the latitude where the taximeter was disengaged.
3. 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.
4. tip_amount: The amount of tip given by the passenger at the end of the trip. Cash tips are not included in this data.
5. trip_distance: The distance reported by the taximeter in miles.

The other attributes are used for data pre-processing and cleansing methods to filter and remove suspected inaccurate and corrupted data.

The tip_amount attribute only contains tips for credit card payments. As cash tips are not included, this report only extracts entries where taxi fare payments are made by credit card (payment type 1).

Outliers detected in the dataset in terms of the fare and tip amounts are best left unchanged because they may provide interesting information within a particular taxi trip. Moreover, these outliers are assumed to be accurately recorded due to the reputable source of the New York City Taxi and Limousine Commission. Several interesting inferences could then be made due to those outliers. There are, of course, exceptions to this rule, which will be discussed within the data cleansing phase.

A generally agreed standard for tipping in taxi trips in New York City is around 15-20% of the fare amount (“Tipping Guide”, n.d.). In this report, the amount of tip given during a taxi trip is deemed considerable if the tip is greater than or equal to 30% of the fare amount. This amount is chosen as it is considered relatively higher than average and would imply various factors, such as high customer satisfaction or prestigious destinations.

All in all, the pre-processing phase reduced the dataset from 1,048,575 to 62,362 entries.

4. Data Cleansing

The cleansing method is done in Microsoft Excel and Python and involves removing data entries that are considered inaccurate or corrupted.

Firstly, data entries that do not provide pickup or drop-off points (missing or zero latitude and/or altitude) are removed since they are not feasible for visualisation. Also, there are data entries where there are zero passengers within 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 are excluded since they would provide little to no value regarding tipping behaviour among passengers. This is to reduce variance within the dataset. Finally, entries with zero fare amounts are removed.

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. Data cleansing suggests that some taxi trips do not end in New York City and would not be feasible for visualisation and hence, filtered.

Moreover, the TLC imposes a $2.5 flat rate for all taxi fares ("Taxi Fare", 2019). This is used as an indicator whether the fare was reasonable considering the trip distance covered. The fare is considered unreasonably high if it costs more than $4 to cover one mile in a taxi trip overall. For instance, a fare is considered unreasonably high if it costs more than $6.5 for a trip that covered one mile. Trips with unreasonably high fares are excluded from this report for data cleansing purposes.

- **$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).

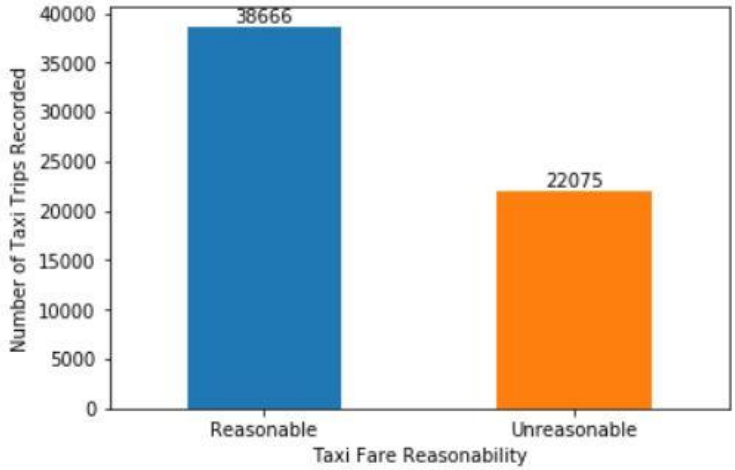

Figure 2: Number of reasonable NYC yellow taxi fares in December 2015.

It should be noted from Figure 2 that the reasonability of the taxi fares in the dataset does not consider airport trips and tolls, the latter due to it being a separate attribute in the dataset.

Overall, the data cleansing phase reduced the dataset from 62,362 to 38,664 entries.

5. Analysis

5.1 Model Relevance

The main purpose of the data pre-processing and cleansing phase was to filter the data in such a way that only data entries that are deemed accurate and reasonable are visualised. After that, the drop-off locations of the remaining entries are visualised.

Firstly, in order to determine which variables are deemed as good fits to model tip amount, a model relevance test was performed in R. The relationship is assumed to be linear and normally distributed with 3 possible attributes. The result is as follows:

```
Call:
lm(formula = mydata$tip_amount ~ mydata$trip_distance + mydata$fare_amount +
    mydata$passenger_count)

Residuals:
    Min      1Q  Median      3Q     Max
 -4.618  -0.379  -0.227  -0.062 217.975

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.367045   0.021696  16.918   <2e-16 ***
mydata$trip_distance   0.149540   0.009097  16.439   <2e-16 ***
mydata$fare_amount     0.288648   0.003130  92.214   <2e-16 ***
mydata$passenger_count -0.001755   0.007369  -0.238    0.812
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.361 on 60737 degrees of freedom
Multiple R-squared:  0.6367,    Adjusted R-squared:  0.6367
F-statistic: 3.548e+04 on 3 and 60737 DF,  p-value: < 2.2e-16
```

Figure 3: Summary of the linear model for tip amount in R.

So **which variables contribute the most to tip amount**? There are 3 numerical attributes that can be considered to fit the model. Since a higher t-value indicates a better fit to the model and the null hypotheses for relevance can be rejected for both trip distance and fare amount, those 2 attributes are considered. Passenger count does not seem to have any relevance against the tip amount.

5.2 Attribute Visualisation

Here, the distance of the taxi trip is considered as a factor. Figure 4 elucidates the importance of taxi distance with regards to the tip given. It shows the proportion of the taxi trips that provided a considerable tip compared to the entire dataset separated by distance. The distance is discretised with equal width, with an interval of 2 miles. The upper bound is inclusive while the lower bound is exclusive. For instance, 0 – 2 miles refer to 0 to 2 miles inclusive, while 2 - 4 miles refer to 2 exclusive to 4 miles inclusive.

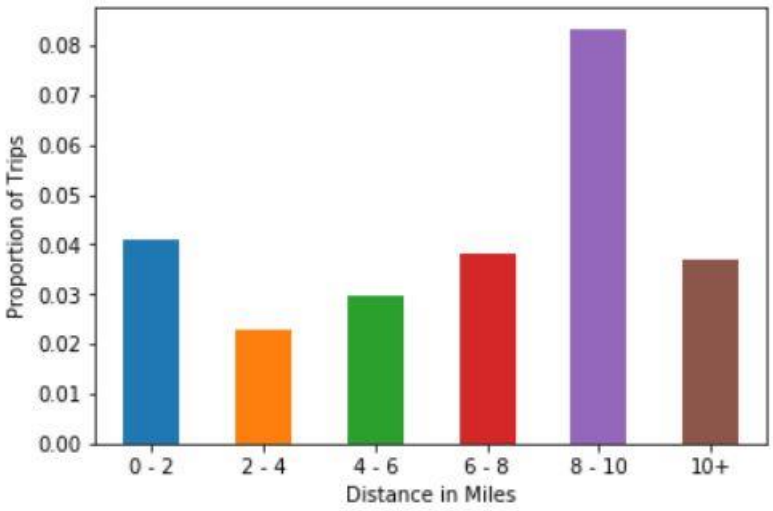

Figure 4: Proportion of NYC taxi trips in December 2015 with a considerable tip amount by distance covered.

Here, roughly 4% of all taxi trips that cover 0 – 2, 6 – 8 and 10+ miles are given a considerable tip amount. The highest proportion, however, is from taxi trips that cover 8 – 10 miles. More than 8% of all taxi trips that cover 8 – 10 miles are given a considerable tip amount.

There is no one possible explanation to describe how trip distance affects taxi tips. Based on speculation, passengers who travel to far distances might tend to tip more due to services rendered by the taxi driver for covering faraway trips, but not a high enough tip for trips that cover more than 10 miles due to the base fare already being expensive. Passengers on short-distance trips might tip more because of the base fare still being affordable.

Moreover, the base fare contributes to the amount of tip given by passengers. Similar to Figure 4, the base fares are discretised with equal width, with an interval of 4 dollars and displayed in Figure 5.

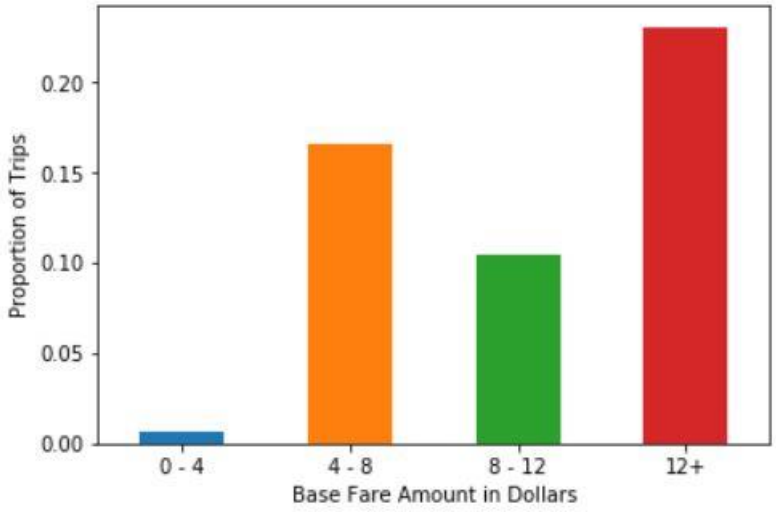

Figure 5: Proportion of NYC taxi trips in December 2015 with a considerable tip amount by fare amount.

There is a positive correlation between taxi fare amount and tip amount. This is evident from the high proportion of trips with a high fare amount that gives a considerable amount of tip. To contrast, less than 1% of trips with a base fare of $0 to $4 give a considerable tip amount while more than 20% of trips with a base fare of more than $12 do.

5.3 Geospatial Visualisation

The visualisation involves a map of New York City and its surrounding areas. The process is done in R using hexagonal binning. Since hexagons are more similar to a circle than squares are, it allows for more efficient data aggregation (Nelli, 2014). The visualisation is split into 100 equally sized hexagonal bins, with 100 being chosen due to the ease of contrast of intensity during visualisation. A smaller number of bins would yield an ambiguous visualisation since the area covered would be too wide to gain any insights. A higher number of bins would be harder to contrast in terms of density.

In the visualisation, lat (latitude) represents the y-axis while lon (longitude) represents the x-axis. The density colours chosen are red for high density and blue for low density for easier contrasts between the two.

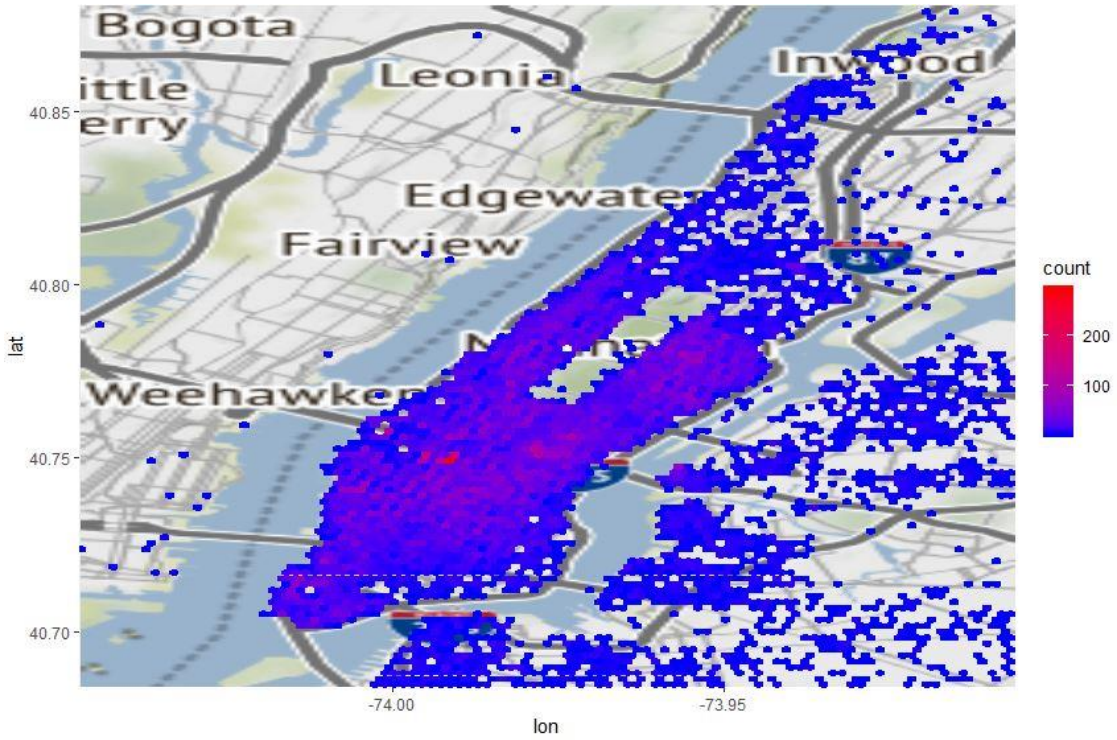

Figure 6: Visualisation of taxi drop-off locations with a considerable amount of tip in New York City and its surroundings on December 2015.

As expected, most of the locations are concentrated in Manhattan, particularly in Midtown Manhattan. It is apparent due to its busyness as the central business district, where tourist attractions such as Times Square and Rockefeller Centre lie along with offices and commuter hubs such as the Grand Central Terminal. (Benjamin, 2019).

Other boroughs of New York City such as Brooklyn, Queens and the Bronx have their proportion of taxi drop-off locations with considerable tips. However, their densities are not as prevalent as Manhattan. Surprisingly, despite the cleansing phase, some taxi trips do still end up outside of New York City, to the state of New Jersey, as shown by some points in the left side of the map around Weehawken. This shows either inconsistency in the dataset or the cleansing phase.

However, although Midtown Manhattan tends to have a higher density of considerable taxi tips, it does not purely indicate passenger behaviour because most taxi trips end up in Manhattan. Various factors, however, can be speculated from the location visualisation, from the general satisfaction of the passenger in the taxi trip, or their wealth, associating Midtown Manhattan with places for the wealthy.

Strangely, there are a few drop-off points which are located on the Hudson River. Although it is illogical, it shows that there exists a margin of error in the dataset in locating the latitude and longitude of taxi drop-off points. Determining the exact drop-off points of these locations is beyond the scope of this report.

6. Reflections

There are several issues encountered within 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.
2. There lacks statistical evidence as to why Midtown Manhattan in particular has the highest density of considerable taxi tips and why people tend to tip more for trips that cover 8 to 10 miles, besides the fact that most taxi trips are present in Midtown Manhattan. Most of the reasons are based on personal speculation and judgment.
3. A separate dataset such as an event happening within New York City on a particular day would explain more clearly why there is a higher density of considerable taxi tips in one location or area.
4. The exact taxi pick-up and drop-off points proved to have a margin of error and should be more accurate in the future for location prediction.
5. Other factors may exist that affect the number of taxi trips with considerable tip amounts such as time of day, age or driver friendliness, with the latter two beyond the scope of the dataset and hence, cannot be inferred.

7. Conclusion

There is no single factor that can be selected to act as means of causation against the frequency of considerable tipping in taxi trips. However, this report suggests that location, fare amount and distance might play a role in tip amounts to a certain extent. The analysis section describes the possible reasons for location and distance as factors. This report has also answered the question: **"What are the influences of tip amounts in taxi trips?"**.

Ultimately, additional context and information of a particular trip is required in order to improve the feasibility of this report. With the given dataset, this report is restricted from attributes that may provide a better fit against tipping. This would also provide a more detailed geospatial visualisation, with better contrasts from one location to another.

In conclusion, assessing the factors of tipping alone would only provide limited insights to the reason why. There are also many considerations of possible factors that affect the amount of tip given by passengers, but not possible to analyse due to the limited context of the dataset. Improvements can be done by solving the issues provided in the reflection section, and future work should include this.

8. References

- Benjamin, E. (2019). 10 Things To Do and See in Midtown Manhattan. Retrieved 9 August 2019.
- Conway, M., Salon, D., & King, D. (2018). Trends in Taxi Use and the Advent of Ridehailing, 1995–2017: Evidence from the US National Household Travel Survey. Urban Science, 2(3), 79. doi: 10.3390/urbansci2030079
- Fischer-Baum, R. (2015). New York's Green Cabs Stay Close To The City Center. Retrieved 13 August 2019, from https://fivethirtyeight.com/features/new-yorks-green-cabs-stay-close-to-the-city-center/
- I-94 Arrivals: Monthly-Quarterly-Annual. (2019). Retrieved 13 August 2019, from https://travel.trade.gov/view/m-2017-I-001/index.asp
- Kelner, S. (2008). Tourism, Ethnic Diversity and the City. *Contemporary Sociology: A Journal Of Reviews*, *37*(1). doi: 10.1177/009430610803700149
- Lynn, M., Zinkhan, G., & Harris, J. (1993). Consumer Tipping: A Cross-Country Study. *Journal Of Consumer Research*, *20*(3), 478. doi: 10.1086/209363
- Nelli, F. (2014). Hexagonal Binning - a new method of visualization for data analysis - Meccanismo Complesso. Retrieved 9 August 2019.
- Taxi Fare. (2019). Retrieved 7 August 2019, from https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page
- Tipping Guide. Retrieved 6 August 2019, from https://archive.nytimes.com/www.nytimes.com/fodors/top/features/travel/destinations/unitedstates/newyork/newyorkcity/fdrs_feat_111_10.html?n=Top%252FFeatures%252FTravel%252FDestinations%252FUnited+States%252FNew+
- Travel Facts and Figures. (2019). Retrieved 13 August 2019, from https://www.ustravel.org/research/travel-facts-and-figures