Taxi Tipping habits in New York City

John Melleuish
Student ID: 1171367
Github repository

August 21, 2022

1 Introduction

Tipping in the United States is often seen as a national passion especially in the restaurant industry. The suggestion of not tipping your server would spark disagreements and judgemental looks amongst the party your with. But is it the same with taxi drivers? Do people see tipping as compulsory? and if so what external factors influence someone level of generosity?

This report will investigate from a taxi drivers perspective what kind of environment can they put themselves in, in order to maximise their chances of receiving a gratuity.

Two datasets were used in this analysis; NYC yellow and green taxis dataset (184397591 rows x 19 columns) as well as weather data from LaGaurdia Airport (28361 rows $\times$ 102 columns). Both Yellow and Green NYC taxis were chosen for this investigation due to the relevant datasets containing almost identical variables captured including tip amount. Ride Share data (Uber, Lyft, private limousines etc.) was available, however they did not make the amount tipped per ride available and thus couldn't be use.

The timeline of analysis was taken from January 2018 through to December 2019. This period was chosen as it is the most recent data not effected by the COVID-19 Pandemic. 2020 on wards despite being newer data, has a whole range of factors that effect the reliability of the conclusions that will be drawn from it. As we move out of the pandemic, taxi drivers will find data without a pandemic asterisk far more valuable as people decisions around tipping would've no doubt changed.

2 Preprocessing

2.1 Taxi Dataset

This report primarily uses data from the NYC Taxi and Limousine Commission [1]. It is comprised of 184,397,591 instances with 19 variables that were measured for every trip taken taken. Due to this report being an analysis into tipping habits, not all of these variables were relevant to the investigation. 9 variables in total were concluded to have a potential to effect someones likelihood to tip which included:

- Pick-up Location
- Drop-off Location
- Passenger Count
- Trip Distance
- Trip Time
- Fare Amount
- Surcharge Amount
- Total Amount
- Time of Day

The dataset was then cleaned up via invalidation of records, for each variables, a common sense test was performed for which values were plausible to have. Passenger count for example had many records

in the hundred which of course is clearly an error and thus we used the legal maximum of 5 occupants as the maximum and 1 occupant as the minimum. Similar rules were put in place for the rest of the variables. Once complete 7,882,213 instance remained (4.27% of the original dataset).

2.2 Weather Dataset

A weather dataset was also used to support this analysis as it is believed weather can have a strong effect on someones mood and potential tip generously. The data was taken from The National Centres for Environmental Information [2] at the weather station located at LaGaurdia Airport. This Airport was chosen as it is the most centrally located airport in NYC and thus would capture the average weather en counted on each day the best.

The dataset came with 28,361 instances with 100 variables of highly precise measurements of which most were considered insignificant to the rate of tipping. 5 of the most common weather variables were chosen as they are widely available on a daily bases and understood to most, these include:

- Temperature
- Dew Point
- Pressure
- Wind Speed
- Wind Direction

The dataset was then cleaned up through invalidating records that fit outside plausible measurements. this was done by finding what the maximum and minimum recorded in NYC history were and setting them to be the upper and lower bounds that the data must fit inside. once complete 4841 instances remained (17.06% of the original dataset)

The two datasets were then joined through an inner join on their mutual date variable, the dataset produced consisted of 3,435,336 instances with the 16 variables discussed above.

3 Variable Analysis

3.1 Pickup Location (Geospatial Visualisation)

In the dataset there were 265 taxi zones across NYC. Figure 1 shows the distribution of the average tips given per taxi ride across the city:

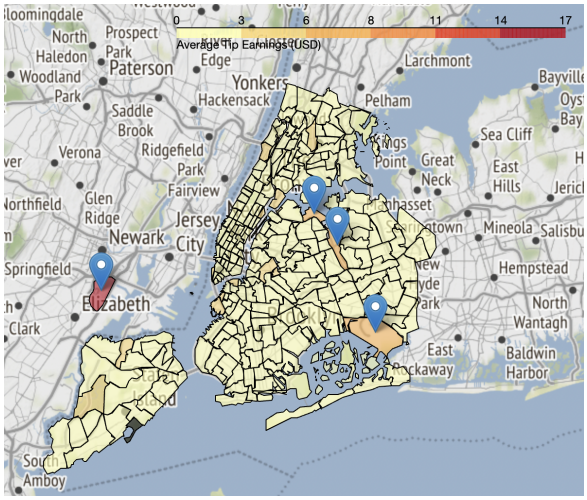

Figure 1: Distribution of Tips across NYC

As visible, there are 4 major outliers in this dataset highlighted with a pop-up icon, with the rest being fairly uniformly distributed under $3 per trip. These outliers are the locations of the 3 major airports

in NYC as well as Flushing Meadow Park (home of the US open and NY Mets baseball team). Of these however Newmark Airport has by far the greatest average tip value. This is due to it being the furtherest airport from the centre of the city and thus people tend to have longer taxi trips (potentially greater tips) when being picked up from this airport.

Flushing Meadows is the interesting result from this analysis due to it being the only outlier not an airport. Due to it being one of NYC most prominent parks, host to a wide range of events and activities, this result fits with our above analysis of the airports as placing people from all over NYC are likely to visit (hence longer trip, greater tips).

3.2 Fare Amount

In popular culture, the most tip are worked out to be a percentage of the raw bill, but is this the same for taxis? Figure 2* shows the relationship between the fare amount and how much someone tipped. 3 common tipping percentages (15%, 20%, 33%) were added to the graph to help visualise potential relationships.

* note: only 10% of the data is displayed in order to see the relationships clearer.

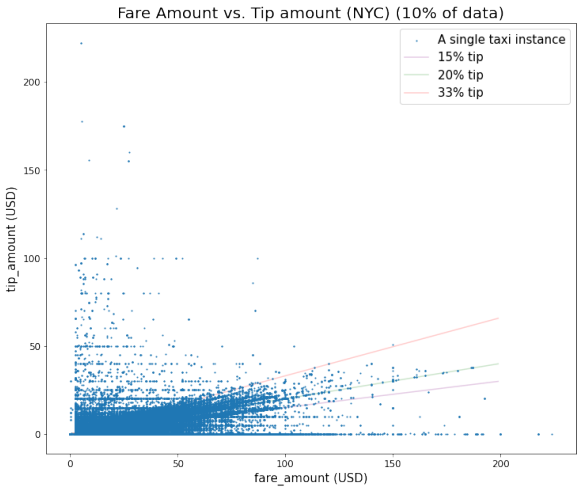

Figure 2: Fare Amount vs. Tips

As displayed, it is evident that the patterns of people tipping are correlated with the fare amount. looking horizontally, it can be seen that it is quite common for someone to tip a whole number amount most notably $0. However, a great percentage of tips (51.99%) fall in the range between 15 and 33% (78.92% if you remove those that didn't tip).

interestingly, there is a greater variance in tipping amount for small fares. it appear as though the the higher the fare amount is, the more likely someone is to either tip between 15 and 33% or not tip at all.

this most interesting part of this figure is $0 tip amount with 34.13% of instances recorded no tips to the driver. Although majority of those that didn't tip had fare amounts less than $50, many people still did not tip the driver well above this threshold. This goes against intuition of America being a culture of compulsory tipping.

3.3 Time of Day

Human moods tend to fluctuate throughout the day with base on our circadian rhythm [3], thus it is no surprise that tipping generosity may fluctuate throughout the day. Figure 3 shows the ratio of people tipping to those that don't throughout the day and figure 4 shows the average amount someone will tip with and without including those that didn't tip.

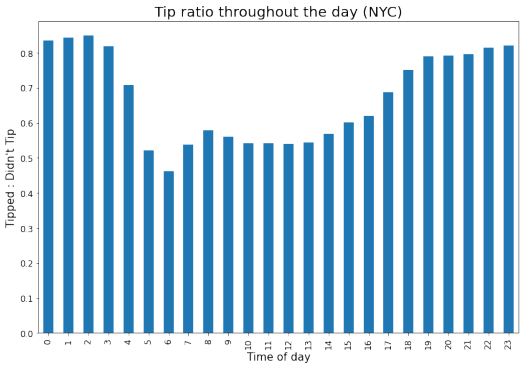

Figure 3: Tipping Ratio

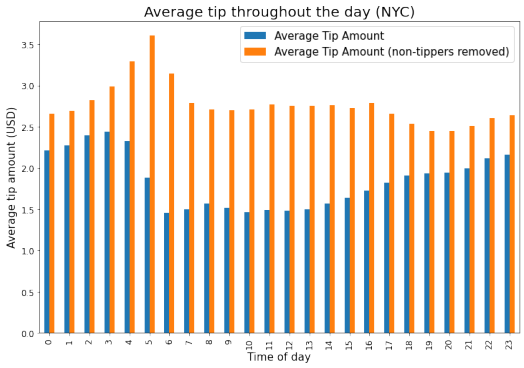

Figure 4: Average Tips

3.4 Temperature

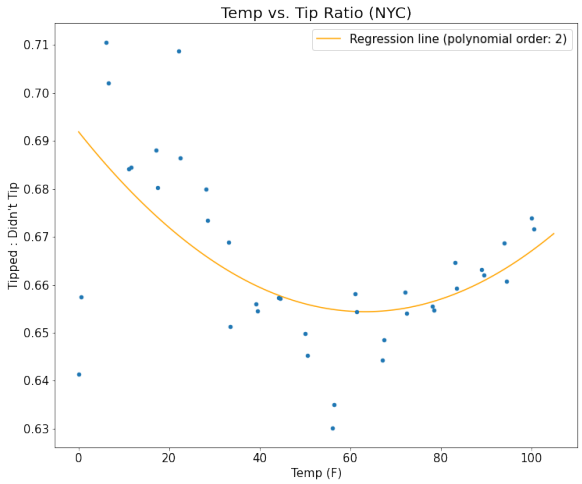

Figure 5: Temp vs. Tips Ratio

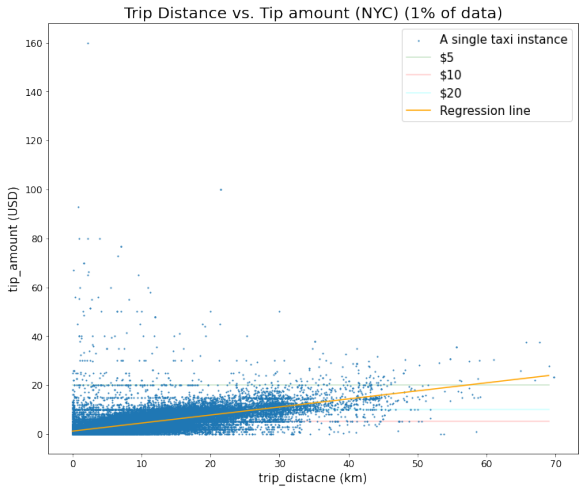

Figure 6: Trip Distance vs. Tips

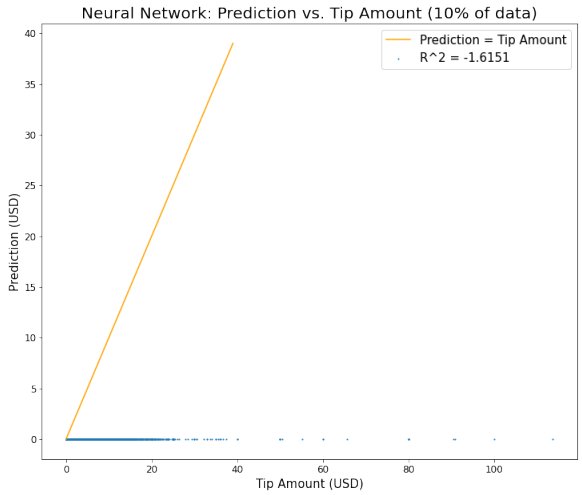

Figure 7: Neural Network

3.5 Trip Distance (optional)

4 Modelling

4.1 Neural Network

4.2 Linear Regression

4.3 Ridge Regression

5 Recommendations

6 Conclusion

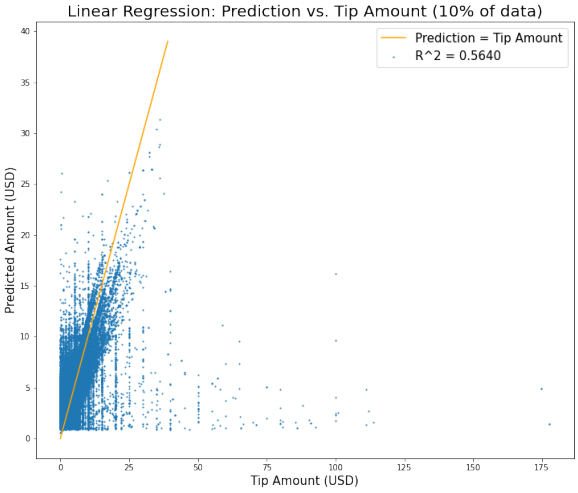

Figure 8: Linear Regression

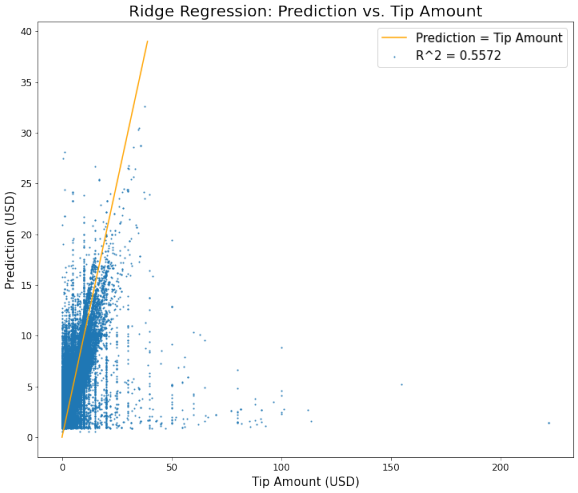

Figure 9: Ridge Regression

https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page [2] [3] Walker, W.H., Walton, J.C., DeVries, A.C. et al. Circadian rhythm disruption and mental health. Transl Psychiatry 10, 28 (2020). https://doi.org/10.1038/s41398-020-0694-0