Quick Analysis of Tipping Yellow Taxi drivers in New York

Dion D. Tran 994765

Melbourne University

Author Note

MAST30034 – Applied Data Science

Abstract

For many countries world-wide, America included, tipping has become the social norm to the point where it is almost always expected (Cross, 2020) (Tomasini, M., Menezes, R., & Oliveira, M. (2017)). This occurrence is said to date back as far as the 1960s to when the congress of the United States enacted 'regulations around federal minimum wages and ultimately determined that entry-level service jobs receiving at least $30 a month in tips required only a minimum wage of $2.13/hr (Tipping etiquette in the United States by an American. Wanderlustingk. (2020)).' As that was 60 years ago, tipping has now been said to be engrained so much in their culture that people are expected to tip for any personal services no just for minimum wage hospitality jobs (St.Louis, R. (2014)). These personal services include 'hair stylists, tattoo artiss, massage therapists' and last taxi drivers. This report will cover basic analysis over tipping yellow taxi drivers in New York and will explore if any attributes help contribute for increased tipping frequency (Tipping etiquette in the United States by an American. Wanderlustingk. (2020))..
Attributes include, trip distance, total payment and location on where they were picked up or dropped off.

1. Introduction

This report will use the 2020 TLC Trip Record Data of the Yellow Taxis across 3 months, January, February and March. With this data we will try to find which variable impacts the tip amount and frequency the largest. Variables that are being taken into account will be; Pick up Location, Drop off Location, Time of the Day, Trip Distance and Passenger Count. Choropleths will be made of New York to find which locations, pick up or drop off, impact tip amount. Boxplots will be made showing tip amounts per hour in each month. Trip Distance and Passenger count will also have relating plots and statistics to see if there is any relevant positive correlation.

2. Data

To get a large enough understanding of human society and their behavior, the data we use must be large enough and cover almost every detail. Due to the basic nature of this report, we will only be covering the Yellow's Taxis first three month of the year 2020.

1. Preprocessing

The main preprocessing and cleaning of the data would be the removal of NA columns but also the filtering of taxi data in where a tip amount is possible. The only way a tip amount can be recorded is when the consumer pays for the trip with a credit card. Though the customer can still tip in cash, it will not be recorded and thus it can not be analyzed unless we search for alternative data. However, as there is no instant workaround, the data is filtered to only have records in which a credit card was their method of payment.

3. Visual Analysis.

To best represent the relationship between tip amount and other attributes, a wide array of visualizations were chosen. The most notable would be the geospatial choropleth visualization which shows and highlights different location pick up and drop offs for taxis and their corresponding average tip amount in the location. For less complex ideas, such as its relationship with time, weekdays or passenger, we used simple scatter plots, bars and histograms.

1. Basic Statistics of Tip Amounts by itself

	Jan	Feb	Mar
count	4694878.00	3481150.00	1608126.00
mean	2.99	2.97	2.90
std	2.83	2.70	2.78
min	0.00	0.00	0.00
25%	1.70	1.75	1.66
50%	2.32	2.35	2.26
75%	3.26	3.32	3.26
max	1100.00	549.02	800.00

Table 1. Tip Amount Description

After preprocessing and cleaning the data, a table of statistics is made on just the Tip amount.

Notable points about Table 1. Summary between the three months are similar with mean approximately reaching 3. Traits that will be considered would be the overall count in March.

We remove outliers using IQR.

Showing a surprisingly low percentage of not tipping, '~3%'.

Overall counts of the top ten tip amounts shows that tipping one or two dollar is more common than not tipping at all, shown in Table 2.

	Jan Count	Feb Count	Mar Count	Jan %Count	Feb %Count	Mar %Count
1.00	112673	89475.0	129638.0	7.61	7.93	8.75
2.00	73555	57471.0	84416.0	4.96	5.09	5.70
0.00	46500	32278.0	56372.0	3.14	2.86	3.80
2.06	30128	22275.0	30718.0	2.03	1.97	2.07
1.96	29974	21696.0	30458.0	2.02	1.92	2.06
2.16	29053	21391.0	29842.0	1.96	1.90	2.01
1.86	28601	20863.0	29353.0	1.93	1.85	1.98
2.26	28066	20665.0	28609.0	1.89	1.83	1.93
1.50	27577	21550.0	30642.0	1.86	1.91	2.07
1.76	26184	19236.0	26756.0	1.77	1.71	1.81

Table 2. Top Ten most common Tip Amounts and their percentages.

2. *Tip Amount vs Time*

Comparing average tip amount of each weekday and each hour of a day. Each day of the month held not much of interest and thus was chosen not to be reported on.

1. Tips Amount of each Weekday

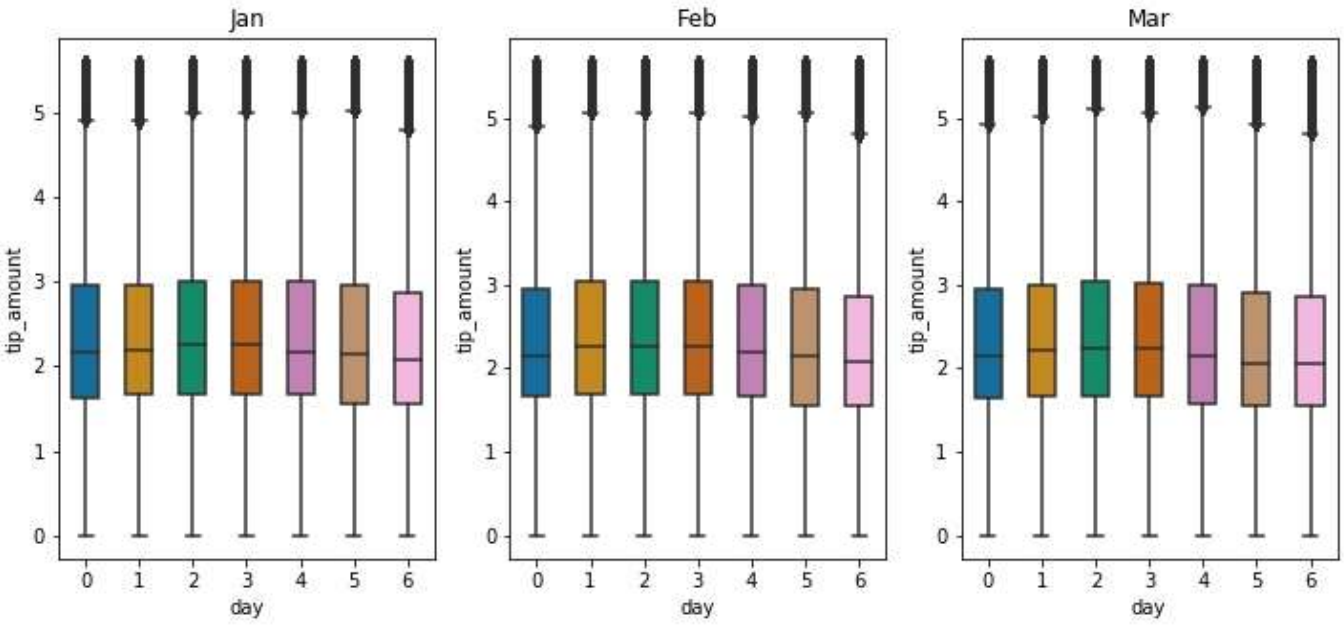

Figure 1. Each months average tip amount per each day of the week. With 0 being Monday, 1 being Tuesday, 2 being Wednesday, 3 being Thursday, 4 being Friday, 5 being Saturday and 6 being Sunday. All months show similar patterns with one another.

Each subplot in Figure 2, reveals a months' weekdays' average tip amount. Though each days' difference is not too large, the same trend can be seen amongst all three months. Tuesday and Wednesday had the highest average tip amount across all months whilst the weekends had the lowest average tip amount. This was surprising, as it was hypothesized that the weekends may have a higher average tip rate as they are usually considered days of rest.

2. Tip Amounts of each Hour of the day.

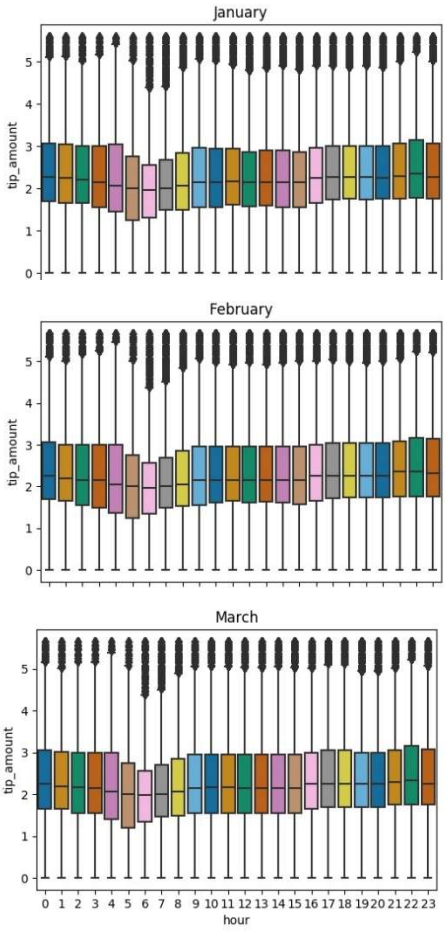

Figure 3. Tip Amount of each hour.

The most notable trait in Figure 3 that was consistent between the three months would be the sloped dip in between hours 3 to 8am. Each month also showed that later through the night you can expect higher tip amounts until 10pm.

3. Correlations

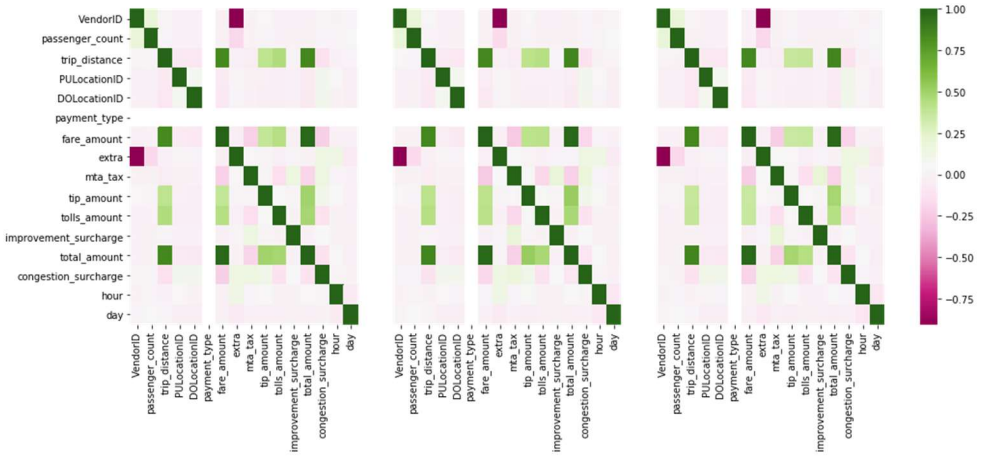

Figure 2. Correlation Heat Maps, Left to Right, January - February - March

Figure 4 shows the similarity between each month, as each month's heat map is almost identical with the other. From the plot we can infer that trip distance, fare amount, toll amount and total amount.

With closer inspection, Figure 4 shows the pearson correlation between said variables, with trip distance, fare amount and total amount being quite close. In general, having a longer trip distance, which in turn increases total and fare amount, will in turn result in higher average tip amount. This seems very likely theoretically as the longer the drive, the more appreciative the passenger may feel for the driver.

	Jan	Feb	Mar
tip_amount	1.00	1.00	1.00
trip_distance	0.40	0.41	0.37
fare_amount	0.38	0.40	0.36
tolls_amount	0.05	0.05	0.03
total_amount	0.50	0.53	0.49

Table 3. Precise Correlation

4. Geospatial Choropleth Visualization

With the use of python libraries such as geo-pandas and folium, these choropleths were created.

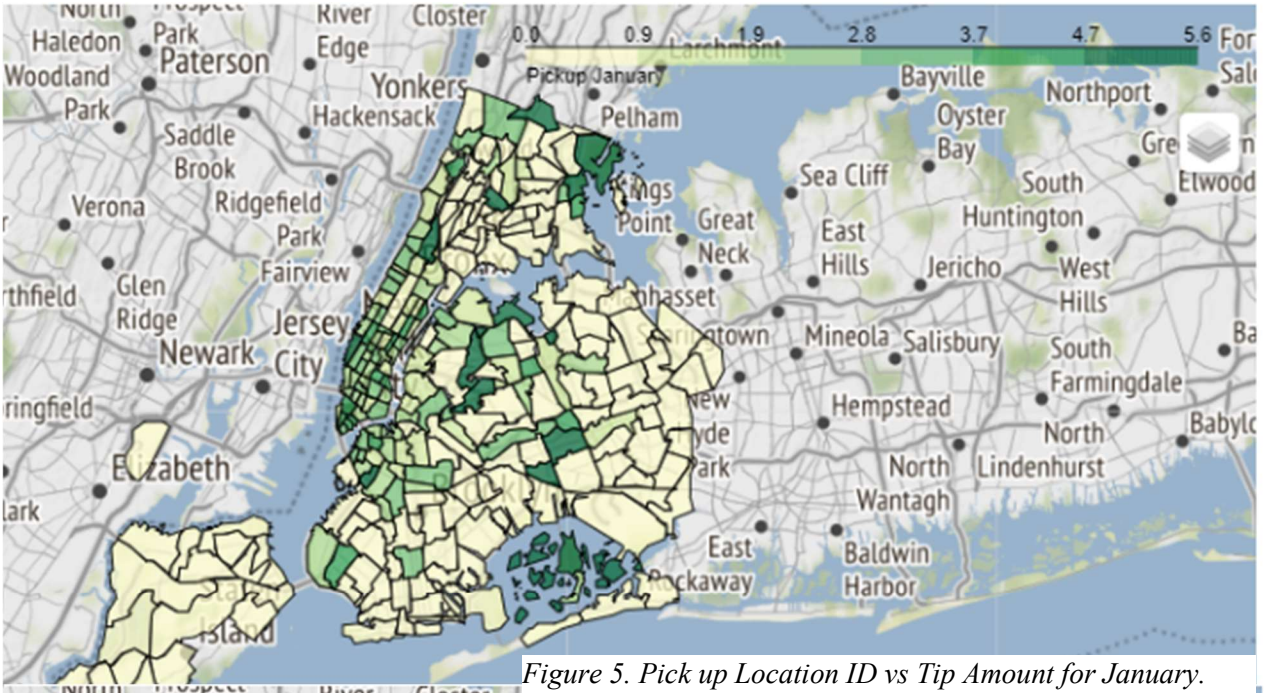

Figure 5. Pick up Location ID vs Tip Amount for January.

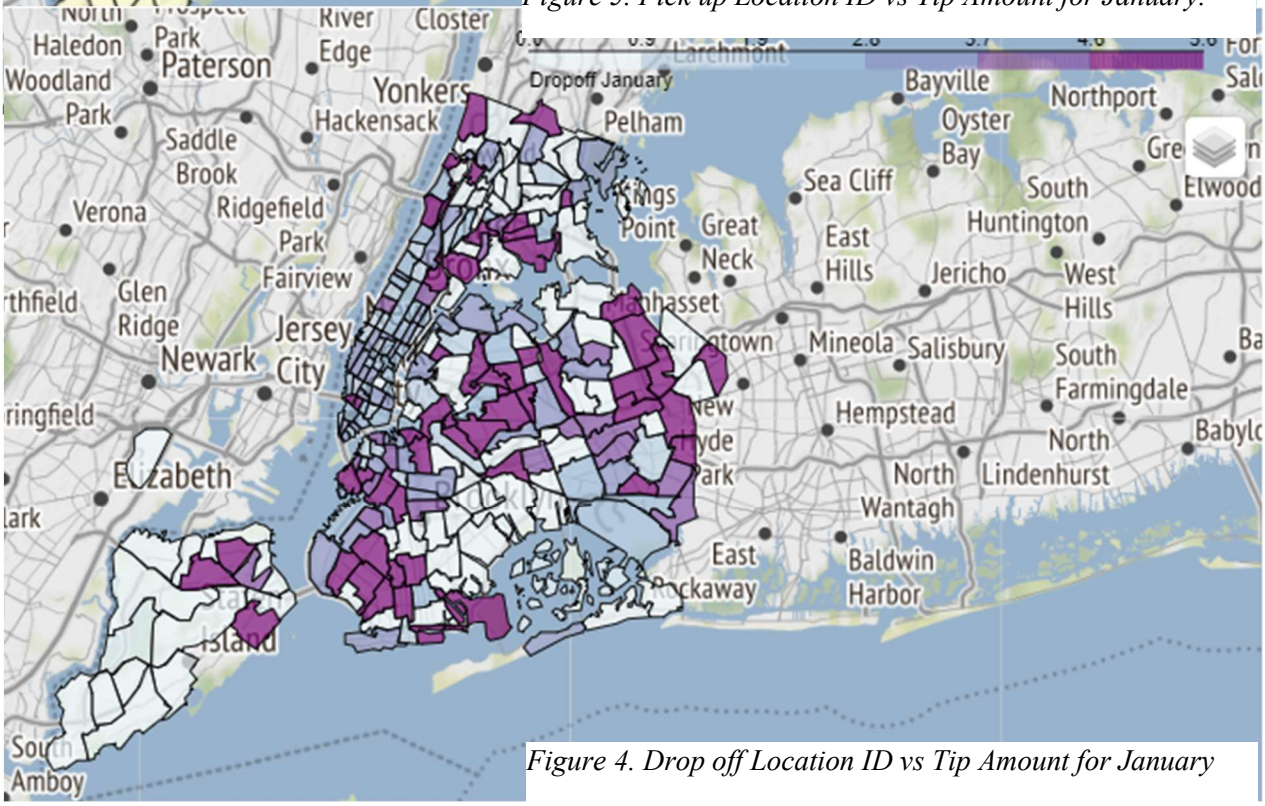

Figure 4. Drop off Location ID vs Tip Amount for January

Each choropleth takes in data from one of the three months and one of its two location IDs, either pickup location or drop off location. Figure 5. shows light green highlighting over many of the areas in New York, depicting an average tip amount of 0-1 dollars. Around the middle however, there is a cluster of areas shaded darker with an average tip amount of almost triple the total mean, 2-3. Of the 13 Areas, there were 8 cemeteries spread throughout, a golf course and a horse racing arena. Figure 6 did not show a favor towards cemeteries, as the tip amounts were more evenly spread. The noteworthy zones did have a cemeteries but also, a large golf course, parks, and a zoo.

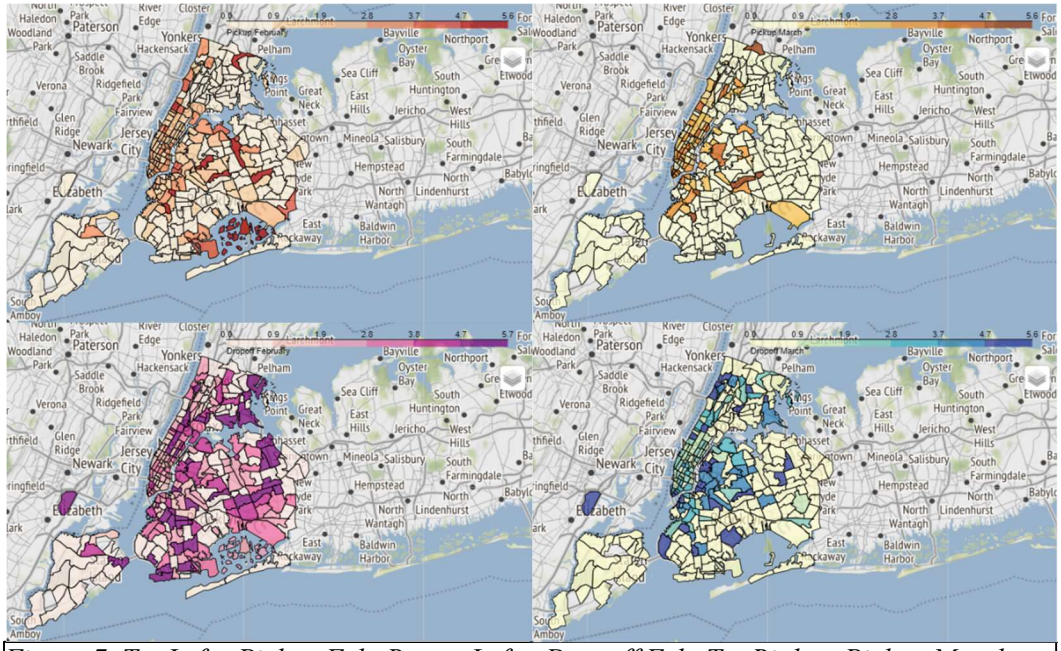

Figure 7. TopLeft - Pickup Feb, BottomLeft – Dropoff Feb, TopRight – Pickup March, BottomRight – Dropoff March.

With the other 2 months shown in figure 7, the areas that had a large tip amount weren't too far out and were much more spread here. The only consistency that was found was the increasing number of cemeteries present in locations with above average tip amounts. This could be due to both a lack of travel to these areas, as cemeteries are not hotspots of interest, combined with human mourning swaying one's judgement. Further inspection, such as relating locations with income and other sources may help bring more solid conclusions.

3. Conclusion

Tipping is unique as though it is not required, it has still persisted on to point of social nomality. This report tried to scratch the bare bones of the tips a yellow taxi driver in New York would get and if there were any variables that could help increase or decrease the tip amount. Tipping seemed particularly common as approximate only 3 percent of people would not tip after their ride. Most common times to tip were Tuesday and Wednesday nights before 10pm, where the worst time for tips would be 3 to 8am weekends. When picking up passengers, locations with cemeteries were more often correlated with higher tip rates, as well as any areas of interests that came with more high end expenses, ie. Golf courses, parks, zoos and docks.

References

Cross, H. (2020). *A Guide to Tipping in New York City*. TripSavvy. Retrieved 4 September 2020, from https://www.tripsavvy.com/guide-to-tipping-in-new-york-city-4177115#:~:text=For%20table%20service%20at%20a,putting%20it%20on%20a%20card.

St.Louis, R. (2014). *How to tip like a New Yorker*. lonelyplanet.com. Retrieved 4 September 2020, from https://www.lonelyplanet.com/articles/how-to-tip-like-a-new-yorker.

Tipping in the USA: Tipping etiquette in the United States by an American. Wanderlustingk. (2020). Retrieved 4 September 2020, from https://www.wanderlustingk.com/travel-blog/tipping-in-the-usa.

Tomasini, M., Menezes, R., & Oliveira, M. (2017). Tippers and Stiffers: an Analysis of Tipping Behavior in Taxi Trips. Retrieved 4 September 2020, from https://www.researchgate.net/publication/317929720_Tippers_and_Stiffers_an_Analysis_of_Tipping_Behavior_in_Taxi_Trips.

Code refrenced from MAST30034 workshops and tutorial's point.