Analysing Tip Amount in New York City Taxi Services

Visualisation Report

Li Sean Wong
[redacted]

August 16, 2021

1 Introduction

The New York City Taxi and Limousine Commission (TLC)[1] have consistently released detailed historical data recording billions of taxi trips throughout New York City (NYC) since 2009. This report aims to analyse and discover the factors that could impact the amount of tip a taxi driver would receive. This would give both new and existing taxi drivers a guideline to optimise their tip amount based on location, trip distance and other factors.

This report uses data from 2019 and 2020 to analyse the iconic NYC Yellow Taxi as well as the Green Taxi. These taxis are chosen as Green Taxi's can access areas where Yellow Taxis cannot, covering a wider area. 2019 and 2020 are chosen respectively as they are the latest datasets to date. Yearly unemployment data, which was obtained from the 2021 Income and Affordability Study[2], is also considered for each borough, assuming that the unemployment rate is equal everyday of the year.

2 Preprocessing

2.1 NYC TLC Dataset

Various preprocessing steps were done before any analysis. Firstly, the entire year's worth of data is combined into a single dataframe. Then, the dataframe is filtered to only consists instances where payments were made via credit card, as tip amounts were automatically recorded. Lastly, the dataframe is further filtered to contain instances where:

- There is at least 1 passenger
- There is more than 0 miles travelled
- The fare amount is positive
- The tip amount is positive

This is done to both 2019 and 2020 datasets for the Yellow Taxi and Green Taxi respectively, which reduced the number of instances from 120 million to 80 million instances. However, the preprocessing steps assume that the taxi could travel for an impossibly large distance and receive an impossibly large tip, which may result in a bias visualisation.

2.2 Unemployment Dataset

The unemployment dataset is manually created by observing the data in the 2021 Income and Affordability Study by the NYC Rent Guidelines Board. The dataset stores the unemployment rate for each borough for the years 2019 and 2020 respectively. However, the data in the study does not include EWR's unemployment rate. This study assumes that the unemployment rate for EWR is the average of all other boroughs, which may result in a bias visualisation.

3 Analysis and Visualisation

3.1 Yellow Taxi

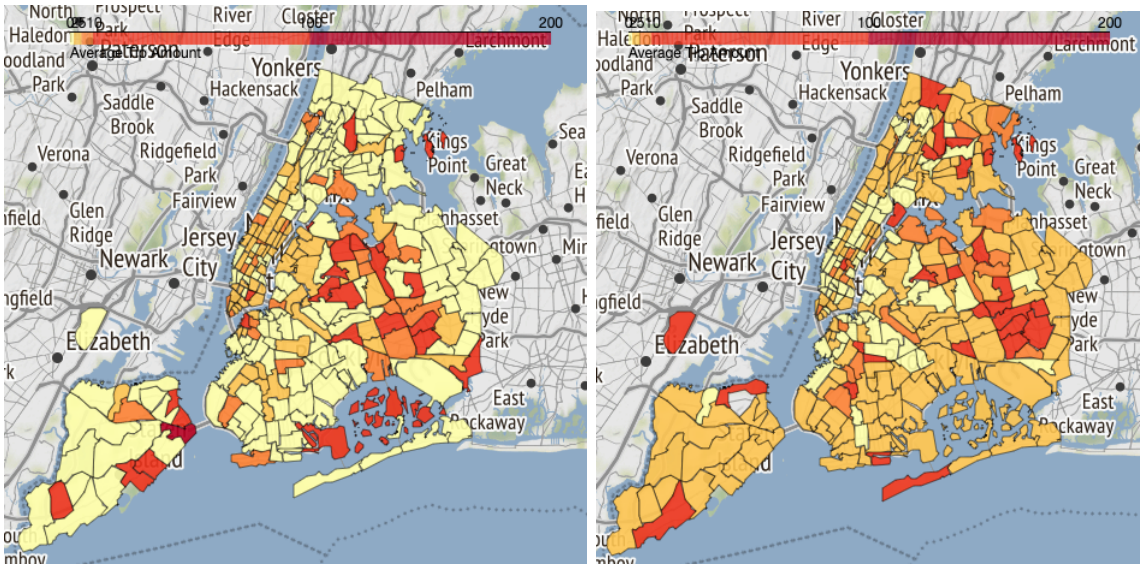


Figure 1: Average Tip Amount per Zone (2019 vs 2020)

From (Figure 1), in 2019 we observe higher average tips surrounding John F. Kennedy (JFK) International Airport as well as around the coastal areas of Staten Island and the general top left area of Queens. This is to no surprise as most tourists travel and pay by cash, which would explain the average low tip at JFK International Airport. As for the top left of Queens, the higher tipping average could be due to The Shops at SkyView Center as well as New World Mall. These popular shopping destinations attract higher income individuals, who would tip more generously. Furthermore, this are of Queens is home to the Queen Zoo.

In 2020, we observe a general increase in average tips, more prominently at EWR and towards the top of Manhattan. EWR's increase as well as the general increase at JFK International Airport could be a result of the Covid 19 pandemic, increasing the number of credit card payments from tourists to avoid as much contact as possible. The top of Manhattan is home to the Yankee Stadium as well as the Bronx Zoo, which could indicate a higher admission to these locations. The most interesting increase can be observed at Breezy Point (left of Rockaway Beach). This could be due to the beach club located there.

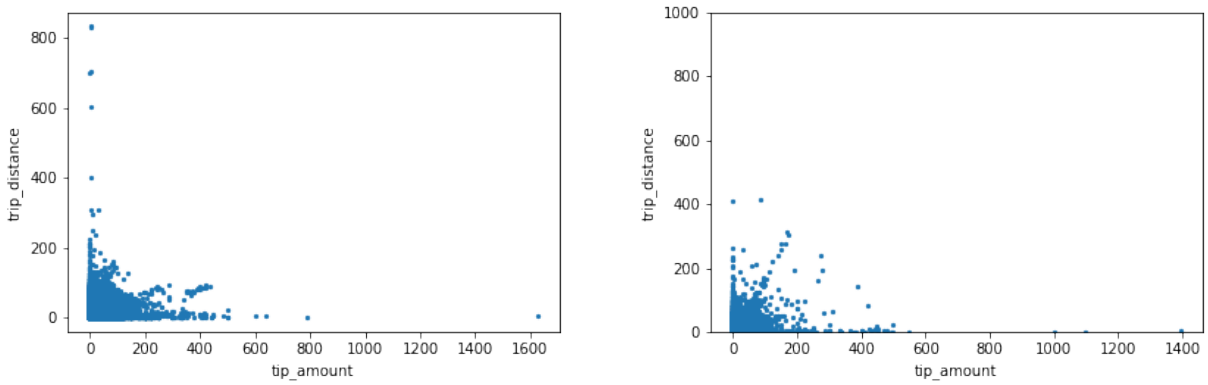


Figure 2: Average Tip Amount per Mile (2019 vs 2020)

From (Figure 2), both 2019 and 2020 graphs do not indicate a clear relationship between travel distance and tip amount. There is no negative or positive correlation nor is there a linear-like relationship between them, excluding a small positive linear trend in 2020 between 0 to 200. However, other points within this region do not reciprocate this relationship.

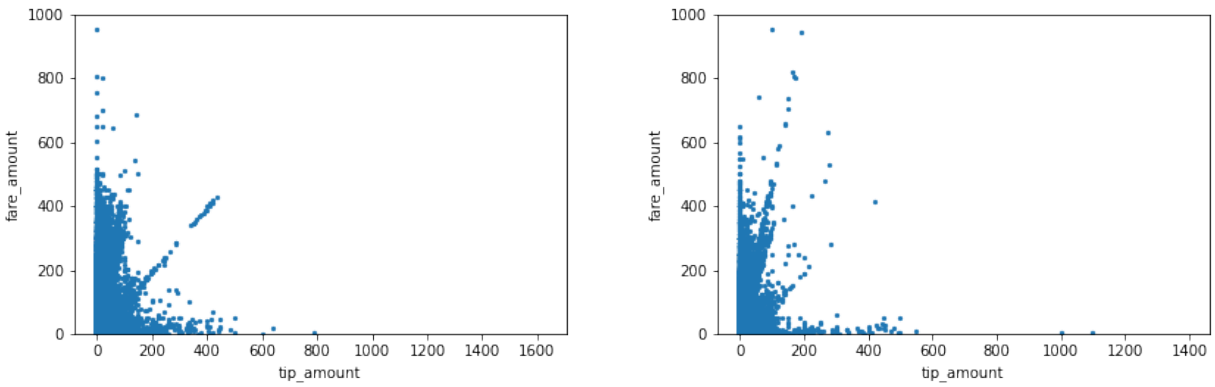


Figure 3: Average Tip Amount Relative to Fare Amount (2019 vs 2020)

From (Figure 3), similar to (Figure 2) both 2019 and 2020 graphs do not have a consistent trend between the two variables. The graph implies that the tip amount is almost random and purely based on the customer. Some linear trends are observed but are not consistent enough to be concluded.

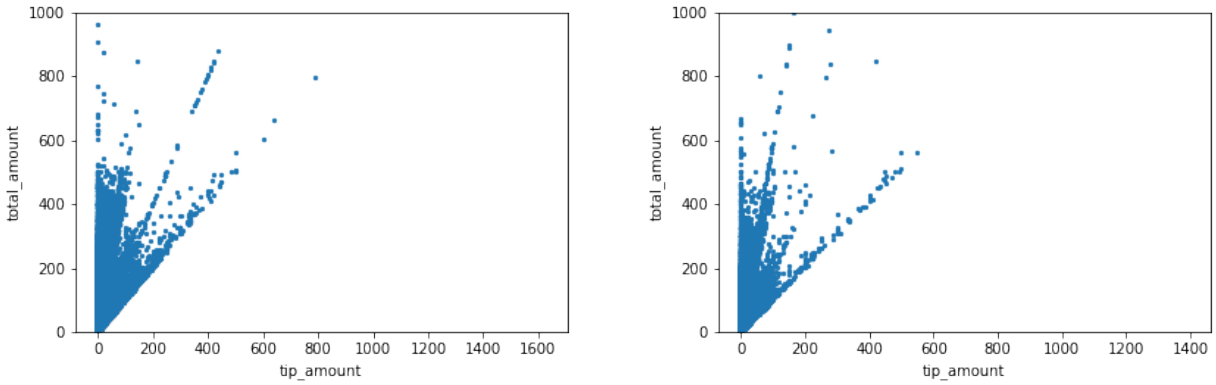


Figure 4: Average Tip Amount Relative to Total Amount (2019 vs 2020)

From (Figure 4), both 2019 and 2020 do not show a clear linear trend. However, an increase in minimum tip can be seen. This indicates that as the total amount increases, the minimum tip the increases as well.

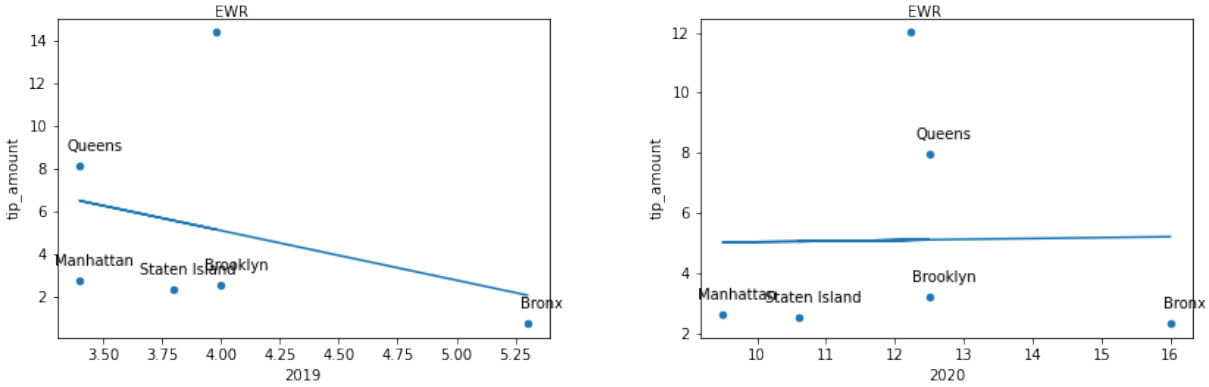


Figure 5: Average Tip Amount Relative to Borough Unemployment Rate (2019 vs 2020)

From (Figure 5), in both 2019 and 2020, we can observe that Queens and EWR act as outliers in the data. This is due to the fact that EWR is skewed by tourism and Queens is skewed by shopping malls and the zoo.

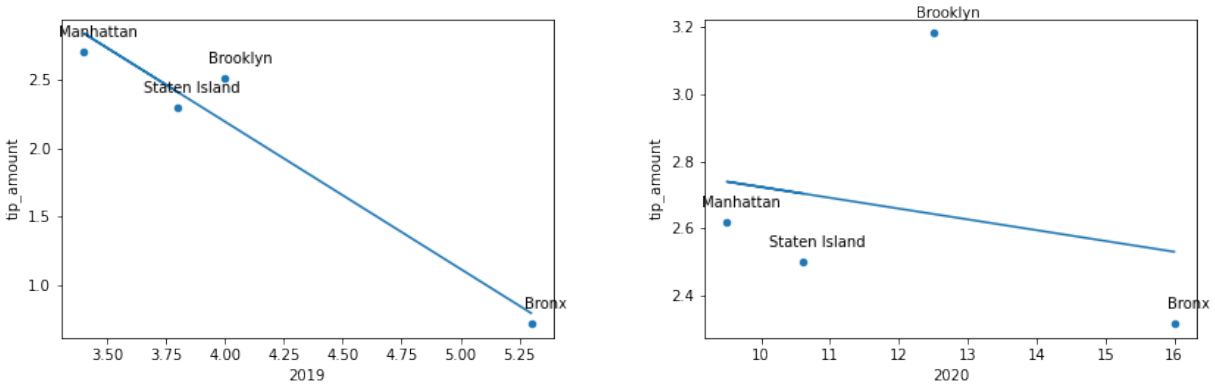


Figure 6: Average Tip Amount Relative to Borough Unemployment Rate Revised (2019 vs 2020)

From (Figure 6), after removing EWR and Queens from (Figure 5), the graphs have a negative trend given by:

- 2019: tip amount = -1.074273*(unemployment rate) + 6.49034
- 2020: tip amount = -0.03217*(unemployment rate) + 3.04465

3.2 Green Taxi

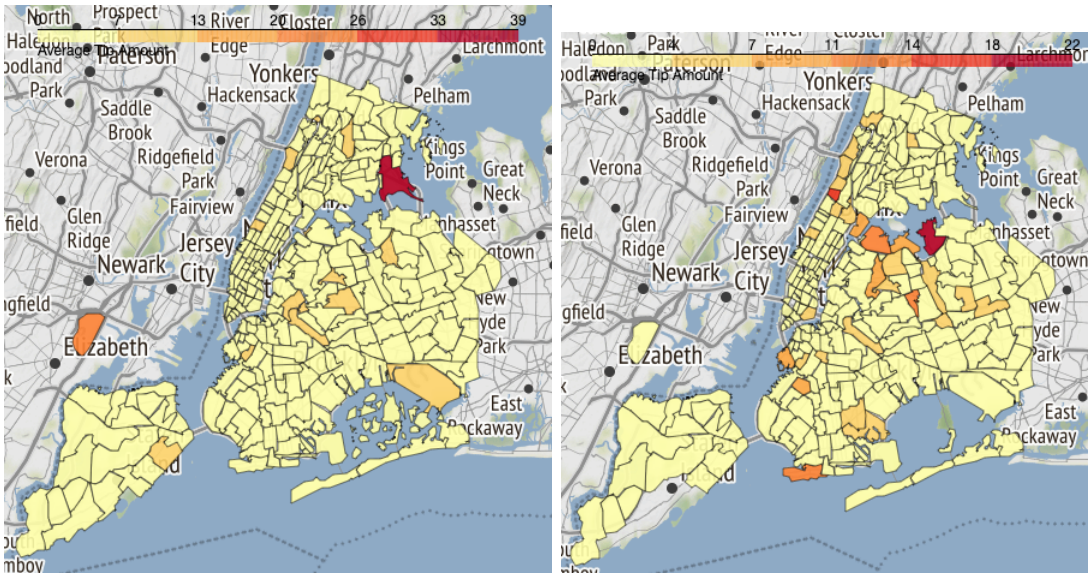


Figure 7: Average Tip Amount per Zone (2019 vs 2020)

From (Figure 7), the average tip amount is relatively even in both 2019 and 2020 with the exception of 2 red points in the top half of the map as well as EWR in 2019. These red patches are placed at rather unusual spots being colleges.

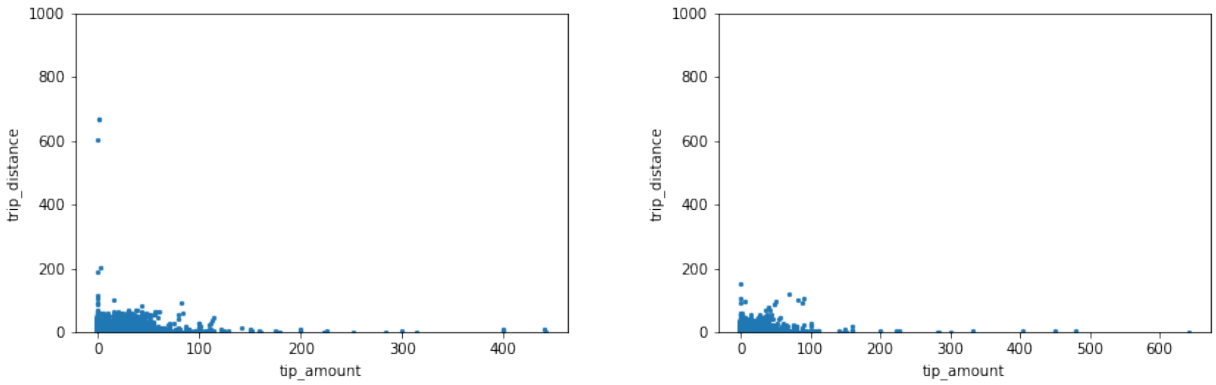


Figure 8: Average Tip Amount per Mile (2019 vs 2020)

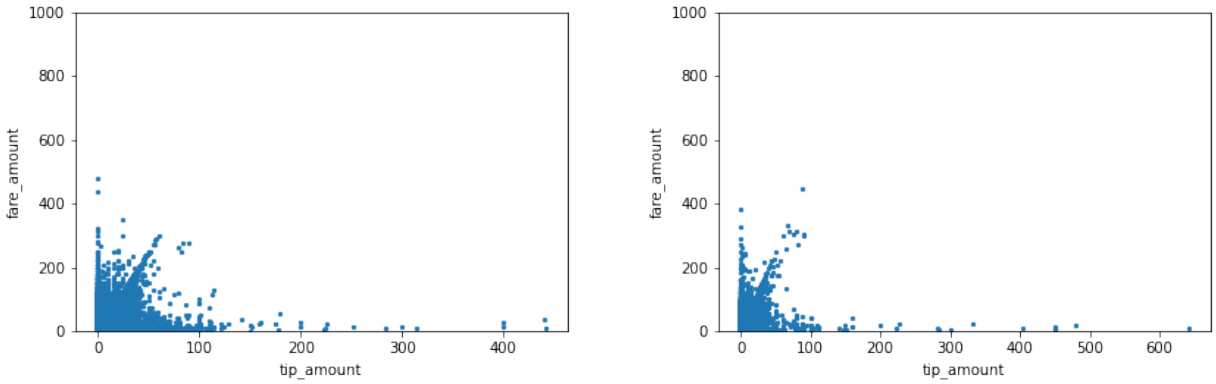


Figure 9: Average Tip Amount Relative to Fare Amount (2019 vs 2020)

From (Figure 8) and (Figure 9), similar to the Yellow Taxi, the trip distance does not seem to impact the amount of tip received. The fare amount also does not seem to impact the tip amount, despite having an inconsistent linear trend.

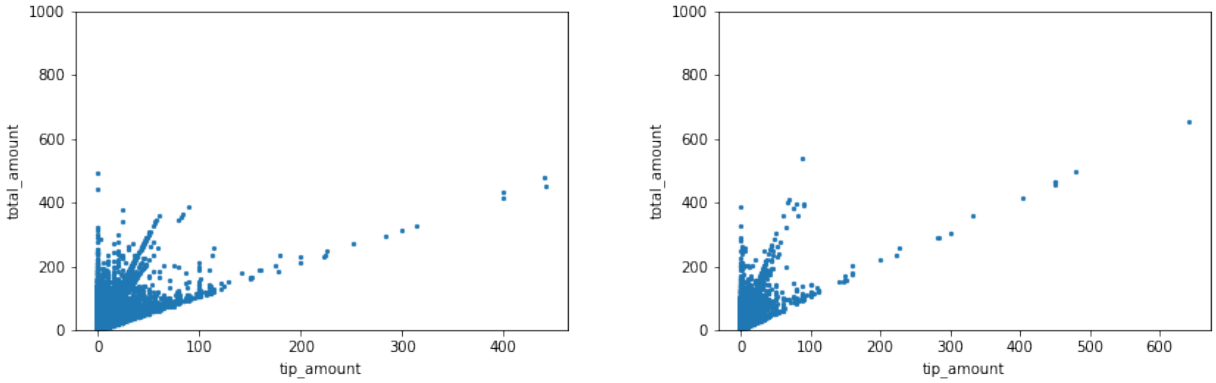


Figure 10: Average Tip Amount Relative to Total Amount (2019 vs 2020)

From (Figure 10), once again we observe no linear trend. However, as observed under Yellow Taxi, we see the minimum tip about increase as total amount increases.

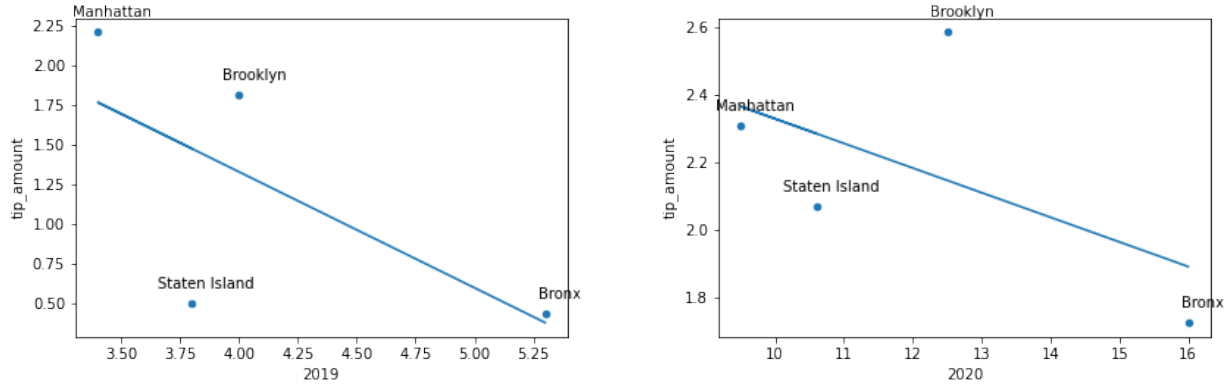


Figure 11: Average Tip Amount Relative to Borough Unemployment Rate (2019 vs 2020)

From (Figure 11), EWR and Queens were removed (similarly to Yellow Taxi), the graphs have a negative trend given by:

- 2019: tip amount = -0.73297*(unemployment rate) + 4.25928
- 2020: tip amount = -0.07309*(unemployment rate) + 3.05930

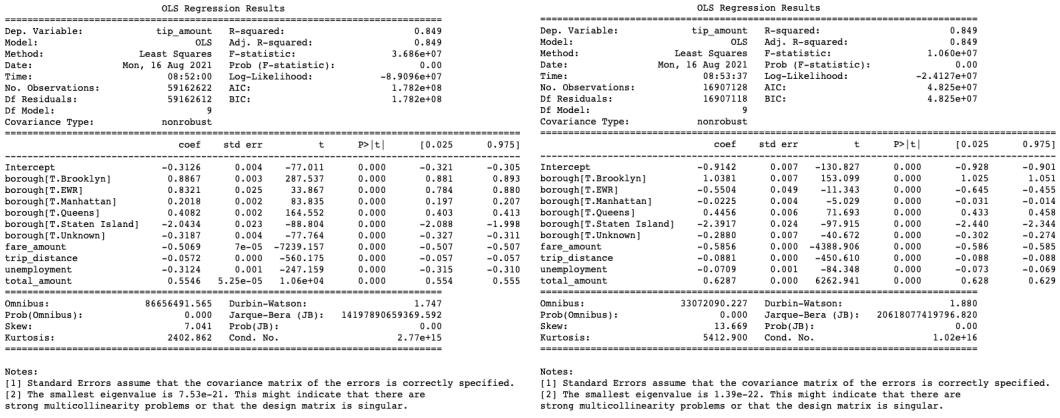
4 Statistical Modelling

4.1 Multilayer Perceptron

Model was covered under COMP30027 Machine Learning.
A prediction model was attempted using a multilayer perceptron model, predicting 2020 data using 2019 data. However, due to the lack of computation power, the process failed to run and therefore no results were obtained.

4.2 Ordinary Least Squares (OLS)

An OLS was carried out for all datasets to predict tip amount. This would assist in confirming the observations under the analysis section.



OLS Regression Results

Dep. Variable:	tip_amount	R-squared:	0.849
Model:	OLS	Adj. R-squared:	0.849
Method:	Least Squares	F-statistic:	3.686e+07
Date:	Mon, 16 Aug 2021	Prob (F-statistic):	0.00
Time:	08:52:00	Log-Likelihood:	-8.9096e+07
No. Observations:	59162622	AIC:	1.782e+08
Df Residuals:	59162612	BIC:	1.782e+08
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.3126	0.004	-77.011	0.000	-0.321	-0.305
borough[T.Brooklyn]	0.8867	0.003	287.537	0.000	0.881	0.893
borough[T.EWR]	0.8321	0.025	33.867	0.000	0.784	0.880
borough[T.Manhattan]	0.2018	0.002	83.835	0.000	0.197	0.207
borough[T.Queens]	0.4082	0.002	164.552	0.000	0.403	0.413
borough[T.Staten Island]	-2.0434	0.023	-88.804	0.000	-2.088	-1.998
borough[T.Unknown]	-0.3187	0.004	-77.764	0.000	-0.327	-0.311
fare_amount	-0.5069	7e-05	-7239.157	0.000	-0.507	-0.507
trip_distance	-0.0572	0.000	-560.175	0.000	-0.057	-0.057
unemployment	-0.3124	0.001	-247.159	0.000	-0.315	-0.310
total_amount	0.5546	5.25e-05	1.06e+04	0.000	0.554	0.555

Omnibus:	86656491.565	Durbin-Watson:	1.747
Prob(Omnibus):	0.000	Jarque-Bera (JB):	14197890659369.592
Skew:	7.041	Prob(JB):	0.00
Kurtosis:	2402.862	Cond. No.	2.77e+15

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 7.53e-21. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

OLS Regression Results

Dep. Variable:	tip_amount	R-squared:	0.849
Model:	OLS	Adj. R-squared:	0.849
Method:	Least Squares	F-statistic:	1.060e+07
Date:	Mon, 16 Aug 2021	Prob (F-statistic):	0.00
Time:	08:53:37	Log-Likelihood:	-2.4127e+07
No. Observations:	16907128	AIC:	4.825e+07
Df Residuals:	16907118	BIC:	4.825e+07
Df Model:	9		
Covariance Type:	nonrobust		

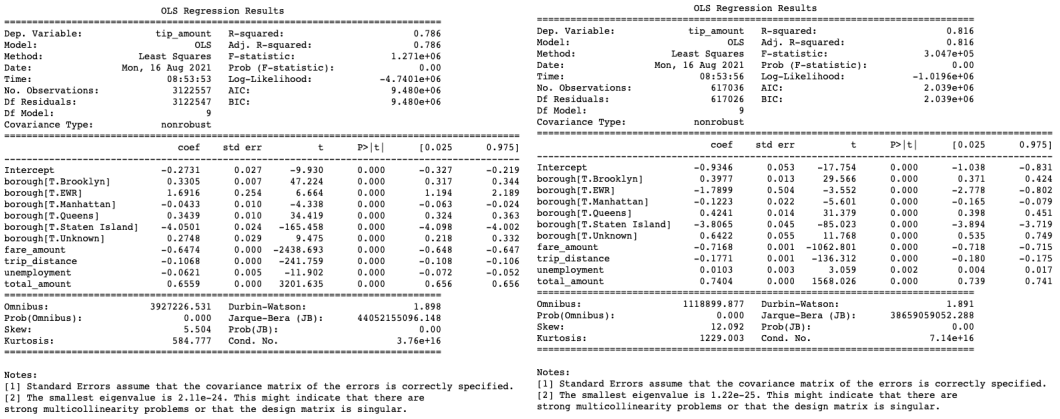
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.9142	0.007	-130.827	0.000	-0.928	-0.901
borough[T.Brooklyn]	1.0381	0.007	153.099	0.000	1.025	1.051
borough[T.EWR]	-0.5504	0.049	-11.343	0.000	-0.645	-0.455
borough[T.Manhattan]	-0.0225	0.004	-5.029	0.000	-0.031	-0.014
borough[T.Queens]	0.4456	0.006	71.693	0.000	0.433	0.458
borough[T.Staten Island]	-2.3917	0.024	-97.915	0.000	-2.440	-2.344
borough[T.Unknown]	-0.2880	0.007	-40.672	0.000	-0.302	-0.274
fare_amount	-0.5856	0.000	-4388.906	0.000	-0.586	-0.585
trip_distance	-0.0881	0.000	-450.610	0.000	-0.088	-0.088
unemployment	-0.0709	0.001	-84.348	0.000	-0.073	-0.069
total_amount	0.6287	0.000	6262.941	0.000	0.628	0.629

Omnibus:	33072090.227	Durbin-Watson:	1.880
Prob(Omnibus):	0.000	Jarque-Bera (JB):	20618077419796.820
Skew:	13.669	Prob(JB):	0.00
Kurtosis:	5412.900	Cond. No.	1.02e+16

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.39e-22. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Figure 12: Ordinary Least Squares for Yellow Taxi (2019 vs 2020)

From (Figure 12),in both 2019 and 2020 pickup zone locations impacts average tip amount in some, but not all, boroughs. Also, fare amount, trip distance and unemployment rate do not seem to impact tip amount as well.



OLS Regression Results

Dep. Variable:	tip_amount	R-squared:	0.786
Model:	OLS	Adj. R-squared:	0.786
Method:	Least Squares	F-statistic:	1.271e+06
Date:	Mon, 16 Aug 2021	Prob (F-statistic):	0.00
Time:	08:53:53	Log-Likelihood:	-4.7401e+06
No. Observations:	3122557	AIC:	9.480e+06
Df Residuals:	3122547	BIC:	9.480e+06
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.2731	0.027	-9.930	0.000	-0.327	-0.219
borough[T.Brooklyn]	0.3305	0.007	47.224	0.000	0.317	0.344
borough[T.EWR]	1.6916	0.254	6.664	0.000	1.194	2.189
borough[T.Manhattan]	-0.0433	0.010	-4.338	0.000	-0.063	-0.024
borough[T.Queens]	0.3439	0.010	34.419	0.000	0.324	0.363
borough[T.Staten Island]	-4.0501	0.024	-165.458	0.000	-4.098	-4.002
borough[T.Unknown]	0.2748	0.029	9.475	0.000	0.218	0.332
fare_amount	-0.6474	0.000	-2438.693	0.000	-0.648	-0.647
trip_distance	-0.1068	0.000	-241.759	0.000	-0.108	-0.106
unemployment	-0.0621	0.005	-11.902	0.000	-0.072	-0.052
total_amount	0.6559	0.000	3201.635	0.000	0.656	0.656

Omnibus:	3927226.531	Durbin-Watson:	1.898
Prob(Omnibus):	0.000	Jarque-Bera (JB):	44052155096.148
Skew:	5.504	Prob(JB):	0.00
Kurtosis:	584.777	Cond. No.	3.76e+16

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 2.11e-24. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

OLS Regression Results

Dep. Variable:	tip_amount	R-squared:	0.816
Model:	OLS	Adj. R-squared:	0.816
Method:	Least Squares	F-statistic:	3.047e+05
Date:	Mon, 16 Aug 2021	Prob (F-statistic):	0.00
Time:	08:53:56	Log-Likelihood:	-1.0196e+06
No. Observations:	617036	AIC:	2.039e+06
Df Residuals:	617026	BIC:	2.039e+06
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.9346	0.053	-17.754	0.000	-1.038	-0.831
borough[T.Brooklyn]	0.3977	0.013	29.566	0.000	0.371	0.424
borough[T.EWR]	-1.7899	0.504	-3.552	0.000	-2.778	-0.802
borough[T.Manhattan]	-0.1223	0.022	-5.601	0.000	-0.165	-0.079
borough[T.Queens]	0.4241	0.014	31.379	0.000	0.398	0.451
borough[T.Staten Island]	-3.8065	0.045	-85.023	0.000	-3.894	-3.719
borough[T.Unknown]	0.6422	0.055	11.768	0.000	0.535	0.749
fare_amount	-0.7168	0.001	-1062.801	0.000	-0.718	-0.715
trip_distance	-0.1771	0.001	-136.312	0.000	-0.180	-0.175
unemployment	0.0103	0.003	3.059	0.002	0.004	0.017
total_amount	0.7404	0.000	1568.026	0.000	0.739	0.741

Omnibus:	1118899.877	Durbin-Watson:	1.891
Prob(Omnibus):	0.000	Jarque-Bera (JB):	38659059052.288
Skew:	12.092	Prob(JB):	0.00
Kurtosis:	1229.003	Cond. No.	7.14e+16

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.22e-25. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Figure 13: Ordinary Least Squares for Green Taxi (2019 vs 2020)

From (Figure 13),in both 2019 and 2020, similarly to the Yellow Taxi, pickup zone locations impacts average tip amount in some, but not all, boroughs. Also, fare amount, trip distance and unemployment rate do not seem to impact tip amount as well.

5 Recommendations

Future studies should try to filter large tips that are incorrect as it might cause a bias visualisation. Furthermore, different external datasets could be implemented to further support claims such as:

- weather data
- covid-19 data
- number of events data
- etc.

Lastly, extremely large datasets could be filtered into smaller datasets through random sampling to reduce computational cost.

6 Conclusion

In conclusion, the following were found in this study:
For Yellow Taxis, in 2019 Brooklyn, EWR, Manhattan and Queens were reliant on pickup locations, where some zones would generate a higher average tip. In 2020, only Brooklyn was reliant on pickup location. In general, the tip amount is heavily reliant on the total amount.
For Green Taxis, similarly to the Yellow Taxis, in 2019 Brooklyn, EWR, Manhattan and Queens were reliant on pickup locations, where some zones would generate a higher average tip. In 2020, only Brooklyn was reliant on pickup location. In general, the tip amount is heavily reliant on the total amount.
However, these conclusion are made based purely on credit card payments. Generally, higher tips would be obtained at JFK International Airport as well as EWR as more data is impacted by tourists.

References

[1] "Taxi Fare." Taxi Fare - TLC. Accessed August 15, 2021. https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page.

[2] "2021 Income and Affordability Study" Appendices - 1. Accessed August 15, 2021. https://rentguidelinesboard.cityofnewyork.us/wp-content/uploads/2021/04/2021-IA.pdf