Quantitative Analysis of New York City Traffic Congestion

How is the pause of issuing FHV licenses
improving traffic congestion and earnings in Manhattan Borough

Haochen Qin
Student ID: 899179

August 17, 2021

1 Introduction

Before COVID, the traffic congestion in NYC has always been a problem. According to the report done by Taxi and Limousine Commission (TLC) and Department of transportation (DOT), the average speed for a vehicle that travelled in Manhattan core has decreased form 6.1mph in 2010 to 4.3mph in 2018 [1]. There were many factors influencing traffic in NYC. For example, increasing use of for-hire vehicle, weather, time of day.
It is straightforward when understanding the correlation between traffic with time of day and weather. However, it is difficult to understand why the traffic is serious with more hire vehicles made available. In fact, the market has saturated with the introduction of app-based services like Uber, Lyft, Juno, and Via. To cope with this staggering situation, a law was passed in August 2018 to stop issuing any new FHV licenses for a year. It was passed for the purpose of limiting cruising vehicles and increasing efficiency. TLC predicted the law will take effect in August 2019.
To analyse the effect of law, choosing the data period between July to August in 2018 and 2019. Yellow and Green taxi are the type of license studied at, and provided by TLC [2]. Besides the attributes already in the TLC data sets, this project added attributes of average travel speed in Manhattan, weather related attributes of central park in Manhattan for analysis.
Since the study goal is to evaluate whether the law improved the traffic congestion in Manhattan, the target audience would be the policy maker, taxi drivers, FHV drivers.

External Dataset

- Weather Dataset
  The data was derived from National Center for Environmental information (NOAA) [3].The dataset contains attributes of dates, the highest and lowest temperature, the observation of rain or snow in quantitative numbers. It was measured in the station inside central park. Central park is located nearly at the center of Manhattan which is good for a representation of the whole borough. The data set was exported based on the timeline need to be studied which contained 4 months.

- Traffic Speed Dataset
  The New York City DOT provides open real-time data of traffic information measured using sensor within the five boroughs of New York City [4]. The attributes include the average speed

of a vehicle traveled between a certain start point to another, the total travel time, dates, the description of the location, borough, owner of the sensor and some trivial attributes. DOT also provide visualisation of the data, which showed Manhattan has more serious traffic congestion on average.

2 Preprocessing

2.1 NYC TLC Dataset

The data released before 2016 was considered to be dangerous for privacy leakage as the locations of the pick-up and drop-off were too precise. Later, the data has been modified to protect individual privacy by abstracting location information. Since the data was chosen after 2016, the location are refereed to taxi zones.

- Removed any data with a NA present.
- Filtered cash payment, since only card payment has record of tips [5].
- Filtered any instances which were not start nor end in Manhattan.
- Filtered any questionable instances. Trips that has 0 travel distance. Trips with same pickup and drop-off time. Instances that has 0 as Passenger-count need to be removed.
- Filtered any instances that don't have rate code 1, the standard rate.
- Create a new data frame with attributes of dates, average daily tolls-amount, trip-distance, trip count for both yellow and green taxi.

2.2 Weather NOAA Dataset

The weather dataset was logged into csv file first, as the data was small and the time cost is low. The dataset includes the date, precipitation, snow, highest temperature, lowest temperature.

- The data was clean and there is no need for much preprocessing.
- Select the subset of wanted studied time period.

2.3 Traffic Speed DOT Dataset

The dataset has too many attributes that needed to be filtered first before export. First choosing Manhattan as the Borough specifically, then the dates. Filtered other owners and kept NYC-DOT-LIC as the source of data. Last but not the least filtered the data with a 0 speed, status -101, the sensor might be interfered or some other technical issues. The dataset has about 440,000 rows combined.

- Data Cleaning has done using the filter provided on the web source.
- The date attribute need to be separate into dates and time. The dates attribute should be in the same format with TLC dataset.
- Attributes like ENCODED-POLY-LINE-LVLS, ID, LINK-ID, LINK-POINTS, ENCODED-POLY-LINE, OWNER, TRANSCOM-ID, BOROUGH, LINK-NAME, STATUS, ID are all irrelevant to this study. They can all be deleted.
- Create new attributes that separate the date time object into dates and time for further analysis.

- Removed outliers of unusal low and high speed by calculating a IQR. Filtered instances higher or lower than 1.5IQR.
- Create a new attribute which compute the average daily traffic speed of Manhattan.

3 Analysis and Geospatial Visualisation

3.1 Preliminary Analysis

First, to have have a geospatial overview of the data. In Figure 1, the color is extent is determined by the fare-amount depending on different taxi zones of NYC. Green taxi is the only allowed to hail in Manhattan. From the graph, there is not many useful information.

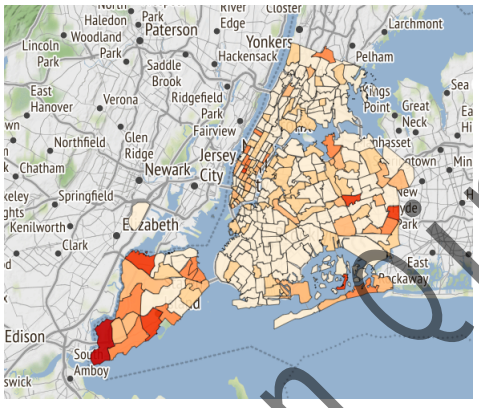

Figure 1: Green taxi: fare amount overview for the whole dataset

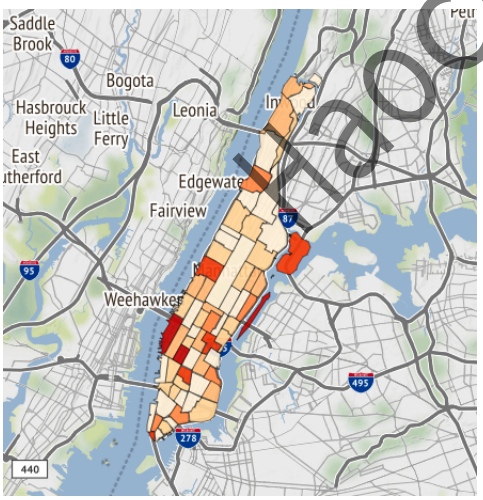

Figure 2: Green taxi: fare amount overview for Manhattan

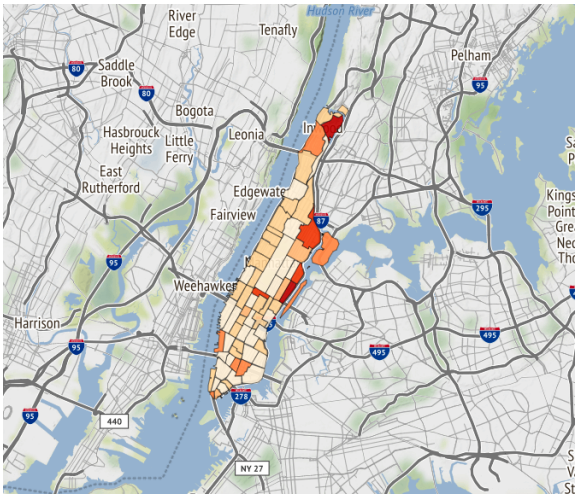

Figure 3: Yellow taxi: fare amount overview for Manhattan

From here, figure 2 and 3 are plotted with the same extent of fare amount. However, these two plots only included trips started and ended in Manhattan. In comparison, green taxi tends to pick-up and receive higher fare amount in Manhattan core (southern Manhattan), and yellow taxi more in the Northern part of Manhattan. This could be the result that green taxi are only allowed to respond to street hail in Manhattan. Green taxi has less data than yellow taxi. This may be the result that

yellow taxi has more hail allowed boroughs and can use e-hail app like Curb or Arro.

Then visualised the external datasets of weather and speed with daily fare-amount to see if they are correlated.

3.2 Attribute Analysis

Plotting heatmaps separately for green taxi and yellow taxi to observe any correlation between the attribute.

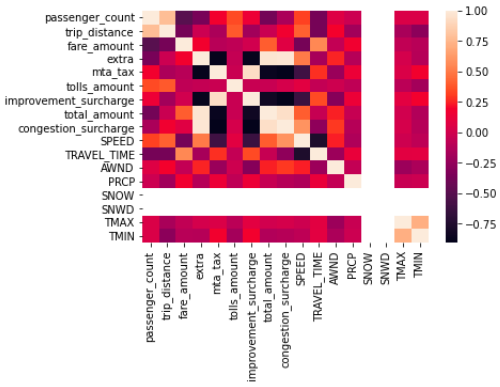

Figure 4: Heatmap of yellow taxi dataset attributes

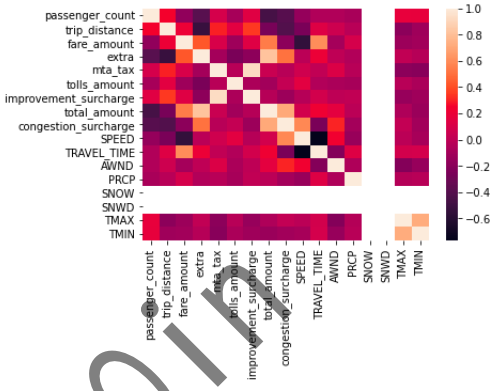

Figure 5: Heatmap of green taxi dataset attributes

1. Firstly, figure 4 and 5 show a common trait that the snow and snow related weather attributes have all zero values. This is because the time period is between July and August, and NYC's weather was sunny and hot. Then these two attributes can be removed from the dataset.
2. In reference to the speed attribute, the average daily fare amount and the travel distance have strong negative relationship for both types of taxi. It is understandable, since the less travel distance and money earned are directed related to the traffic speed. Also, the congestion charge has a positive relationship with the traffic speed. It is not strongly related because the dataset combines 2018 and 2019 data together where congestion fee wasn't charged in 2018. It was introduced later to resolve traffic congestion.
3. There are no extreme weathers in these months. The weather attributes don't contains many useful information.
4. Green taxi is different with yellow taxi as the travel distance is less correlated with fare amount. In figure 2 and 3, green taxi have higher fare amount in Manhattan core where people may be more generous with tips.

Figure 7 plots the correlation between daily speed and fare amount of yellow taxi. The law passed in August 2018 that aim to reduce congestion has taken effect in 2019.
In figure 6, it can be directly observed that the traffic speed has increased in 2019. Figure 8 and Figure 9 take closer look at the speed trend in 2018 and 2019. It was not hard to see a similar trend of speed in both year. This tend could be the result of annual events, 4th of July holiday, curfews, protests and many reasons.

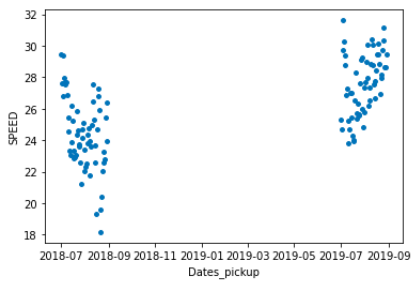

Figure 6: Overview of traffic speed and date

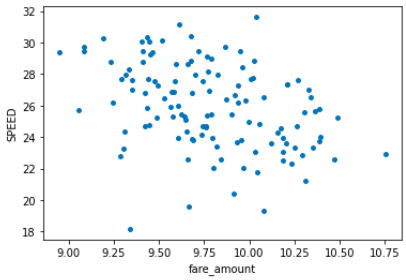

Figure 7: Yellow taxi fare amount vs speed

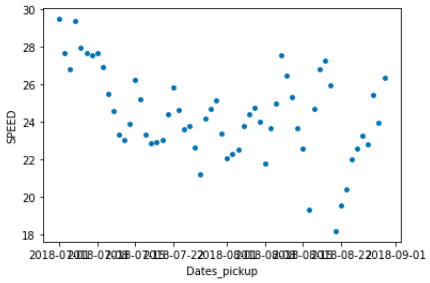

Figure 8: 2018 Yellow taxi traffic speed and date

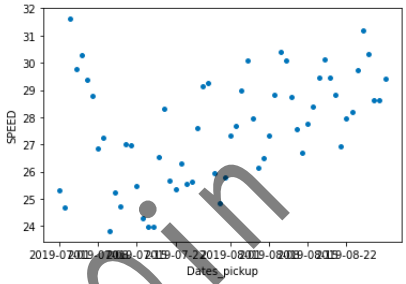

Figure 9: 2019 Yellow taxi traffic speed and date

4 Statistical Modelling

4.1 Model

$$Y = X\beta + \epsilon \tag{1}$$

Use linear model statistics, to predict traffic speed. LASSO (Least Absolute Shrinkage and Selection Operator) is used for best model selection method. LASSO shrinks the attributes that are irrelevant to 0. This method need to find the best shrinking parameter.

4.2 Results

It can be shown in figure 10 and 11, green taxi and yellow taxi are very different from each other. For the green taxi dataset, the congestion charge and fare amount are included in the model. This selection matches what have observed figure 5, the heatmap. For yellow taxi, the model is more complicated, attributes like passenger count, fare amount, tax, improvement surcharge are included.

4.3 Discussion

For the green taxi, the traffic speed is slower with a lower congestion surcharge. Also, the higher fare amount indicate a slower speed of traffic. Yellow taxi is more effected with the high efficiency of the trip, having more people are riding taxi.

	Coefficient
Intercept	25.9594
passenger_count	0.0000
fare_amount	-0.7978
trip_distance	0.0000
tolls_amount	0.0000
extra	0.0000
mta_tax	0.0000
trip_distance	0.0000
improvement_surcharge	0.0000
total_amount	0.0000
congestion_surcharge	0.9000
AWND	0.0000
PRCP	0.0000
TMAX	0.0000
TMIN	0.0000

Figure 10: Green taxi LASSO linear model

	Coefficient
Intercept	25.9594
passenger_count	0.3474
fare_amount	-0.3947
trip_distance	0.0000
tolls_amount	0.0000
extra	0.0000
mta_tax	-0.5975
trip_distance	0.0000
improvement_surcharge	-0.7107
total_amount	0.0000
congestion_surcharge	0.0000
AWND	0.0000
PRCP	0.0000
TMAX	0.0000
TMIN	0.0000

Figure 11: Yellow taxi LASSO linear model

5 Recommendations

For the green taxi, the congestion surcharge worked, as the speed increase and congestion surcharge increase. Policy makers could consider a more effective charging method. For example, once enter the Manhattan core charges are made. Yellow taxi brings insight in relieving congestion issue with increasing passenger count. Allowing shared trips for yellow taxi. In this study, weather isn't a strongly correlated factor to speed. However, it is wrong to conclude weather isn't important. It is trivial only in the study period.

6 Conclusion

Using the dataset from TLC and external datasets, this report studied the relationship between taxi trip and traffic speed in Manhattan. There are more factors that needed to be considered for a more comprehensive report, like the daily average earned amount of FHV, the percentage of vehicles on street. The expected finding would be the observation of higher traffic speed, more green and yellow taxi trips. This project is expected to find average the traffic speed given the weather, taxi data. Weather is a fundamental factor in measuring traffic speed, but observed irrelevant here.
The market for taxi and FHV is saturated. Drivers spend too much time cruising around with empty

car. This way the traffic is bad and having unpredictable bad consequences. Policy makers need to ensure the income for the drivers and manage traffic. More research need to be taken to create optimal solution.

References

[1] "Improving Efficiency and Managing Growth in New York's For-Hire Vehicle Sector." Final Report - New York City Taxi and Limousine Commission and Department of Transportation. Accessed August 15, 2021. https://www1.nyc.gov/assets/tlc/downloads/pdf/fhv-congestion-study-report.pdf

[2] NYC Taxi and Limousine Commission. (2018, 2019). TLC Trip Record Data. Accessed August 15, 2021. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

[3] "Record of Climatological Observations." NOAA .Accessed August 15, 2021. https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00094728/detail

[4] "New York City Traffic Speed Detectors." NYCDOT.Accessed August 15, 2021. https://data.cityofnewyork.us/Transportation/DOT-Traffic-Speeds-NBE/i4gi-tjb9/data

[5] "Taxi Fare." Taxi Fare - TLC. Accessed September 7, 2019. https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page.