New York Taxi Visualization Report

Name: Yichuan Zhang

University of Melbourne

Applied Data Science

Contents：

1. Abstract

2. Introduction

3. Project Specification

3.1 Data Characteristics

3.2 Feature Specification

3.3 Software Choice

4 Project Plan

4.1 Development process

5 Implementation

5.1 Data Pre-processing

5.2 Methodology

6 Visualization analysis and attribute analysis

7 Discussion

7.1 Weakness

7.2 Improvement

8 Conclusion

9 Reference

1 Introduction

1.1Background Statement:

Public transportation has become a major travel habit for both local people and tourists, especially taxi takes the major role for local population to commute between home and workplace. The taxi market in New York City is one of the largest and busiest market in this world, since the New York City (below we will call NYC) is the largest financial center world widely. The fast-tempo society endowed the taxi market in NYC with a special role, therefore a typical analysis to this taxi market by using big data analysis is important and insightful.

The project will mainly focus on analysis the taxi market in NYC, and this paper presents partial part of the whole project, which is about visualization analysis to the NYC million size data for the year 2015 and year 2020. More precisely, the dataset for year 2015 will be mainly used for visualization analysis, and dataset for year 2020 will be used for the statistical analysis.

2 Project specification

2.1 Data Characteristic

The data set we are using in this project will be the New York City Taxi and Limousine Service Trip Record Data. The data itself records the specific features from different types of licensed taxi and limousine services in the NYC area. The data set was downloaded from the official website. It is worth to mention that the data set from New York City Taxi & Limousine Commission (TLC) does not include the specific location (latitude and longitude) anymore after Year 2015, according to the NYC Department of City Planning's Neighborhood Tabulation Areas (NTAs). Therefore, shape file is needed to realize the geographic visualization.

Table 1 shows the basic information for the file of yellow taxi data and shape file.

Dataset	Dataset Size	Number of records
Yellow taxi 2020-01	594MB	6405008 row index
Yellow taxi 2015-01	1.99GB	12748986 row index
Shape file	2MB	260 entries

Table 1. Basic information for files

2.2 Feature Specification

The sub-section will introduce the basic attributes included in the NYC data set (Yellow taxi).
Table 2 shows the basic information for each attribute in Yellow Taxi dataset.

Attribute Name	Data Type	Description	Sample
VerdorID	float64	TEPE provider code	1
tpep_pickup_datetime	object	Data and time of pickup	2020-01-18 17:35:00
tpep_dropoff_datetime	object	Data and time of dropoff	2020-01-18 18:37:00
Trip_distance	float64	The trip distance (km)	22.50
PULocationID	float64	Taxi zone of pickup	27
DOLocationID	float64	Taxi zone of dropoff	121

RateCodeID	object	The final rate code	1
Store_and_fwd_flag	int64	Whether the trip record was held in Vehicle	1
Payment_type	int64	How passenger paid for the trip	1
Fare_amount	float64	The time/distance fare	16.55
MTA_tax	float64	Miscellaneous extras and surcharges	0.5
Improvement_surcharge	float64	Surcharge assessed trio at the flag drop	0.3
Tip_amount	float64	Tip amount for credit card only	3.5
Tolls_amount	float64	Total amount of all tolls paid in trip	5.30
Total_amount	float64	Total amount charged to passengers	20.55

Table 2. Basic information for each attribute

Figure 1 is the screenshots of Yellow taxi dataset from 2015-01.

	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	pickup_longitude	pickup_latitude	RateCodeID
0	2	2015-01-15 19:05:39	2015-01-15 19:23:42	1	1.59	-73.993896	40.750111	1
1	1	2015-01-10 20:33:38	2015-01-10 20:53:28	1	3.30	-74.001648	40.724243	1

Figure 1. Example for the yellow taxi 2015-01 dataset

Figure 2 shows the attributes in shapefile.

	OBJECTID	Shape_Leng	Shape_Area	zone	LocationID	borough	geometry
261	262	0.049064	0.000122	Yorkville East	262	Manhattan	MULTIPOLYGON (((-73.94383 40.78286, -73.94376 ...
262	263	0.037017	0.000066	Yorkville West	263	Manhattan	POLYGON ((-73.95219 40.77302, -73.95269 40.772...

Figure 2. Example for the shapefile

2.3 Software Choices

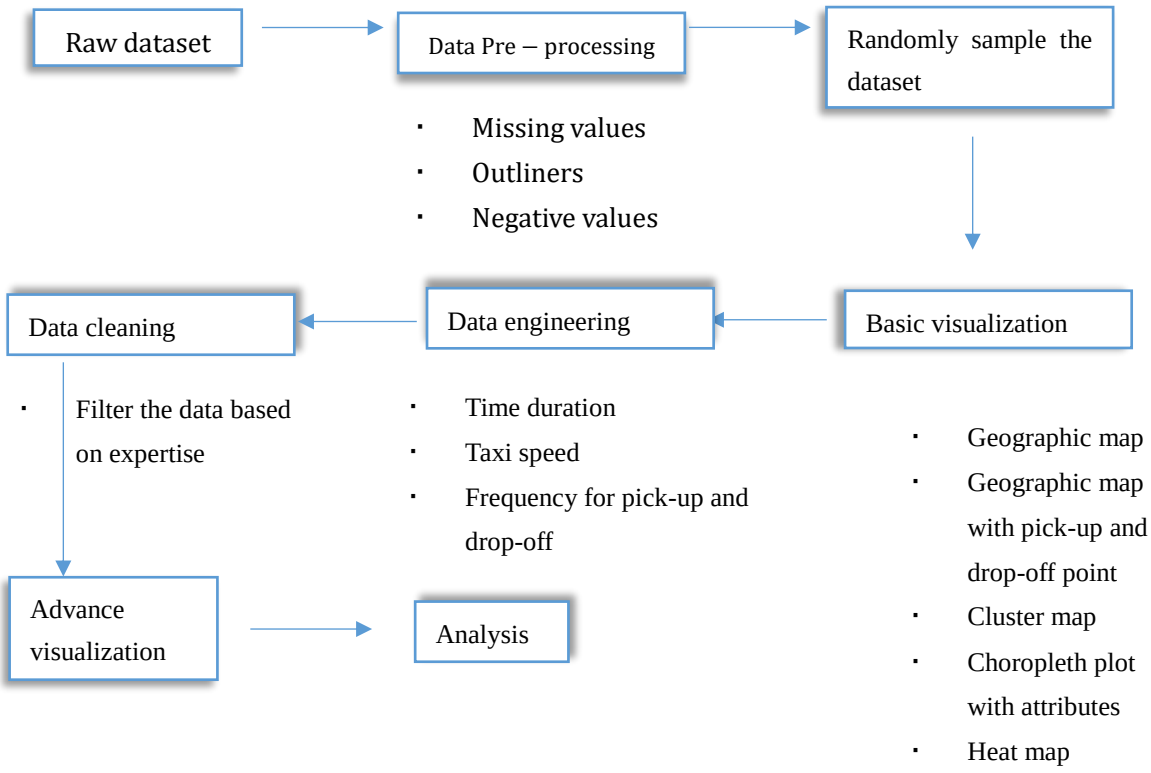
This section introduces the fundamental technologies used, different technologies are possible to conduct this project as well.

The primary tool used in this project is python(版本), which was conducted on the virtual machine (Jupyterhub). For the statistical part of this project, R studio will be used later in the project.

3 Project Plan

3.1 Development Process

In the section, the development process for this project is presented.



For this project, the major interest will be the analysis of visualization to at least one attribute. Thus, the latitude and longitude will be helpful in visualization. However, since the taxi dataset after year 2015 has no specific location, the PULocation and DOLocation should work with the shape file

Moreover, the visualization should be accompanied by other attributes. According to our interest and future plan for this project, this project will try to focus on analyzing the quantitative aspects of the NYC taxi market, therefore the original attribute such as tip amount, total mount will be used in visualization.

Besides just using the original attributes, the feature engineering is necessary for driving the new feature, such as time duration for the trip and speed of taxi, etc

Before reaching the step of visualization, the data pre-processing is always necessary for excluding the strange data point and reducing the size of data set.

3.2 Basic Visualization of the NYC map

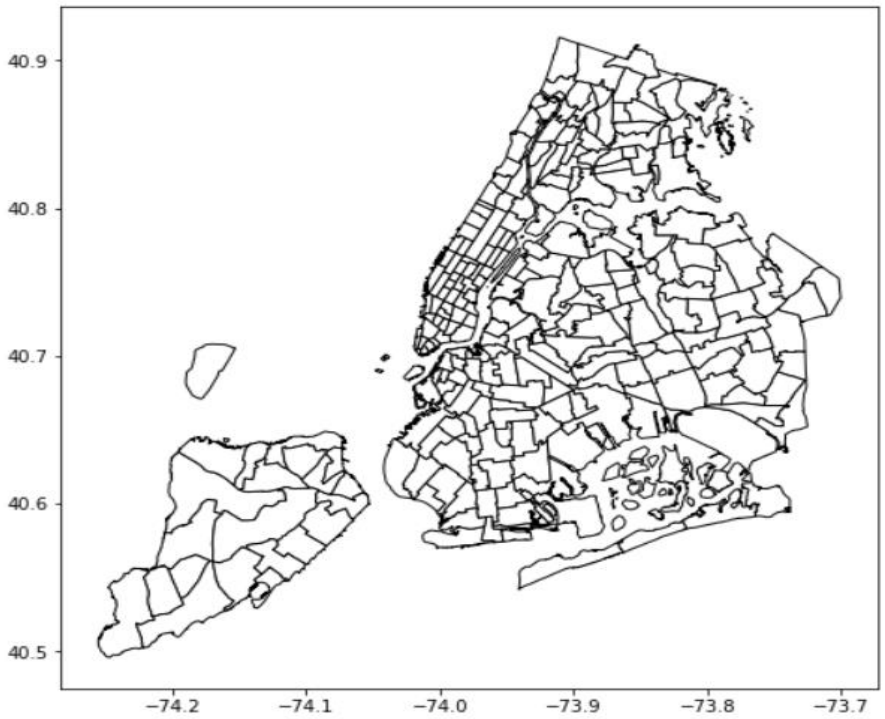


Figure 3. The direct view of NYC map with black and white colors

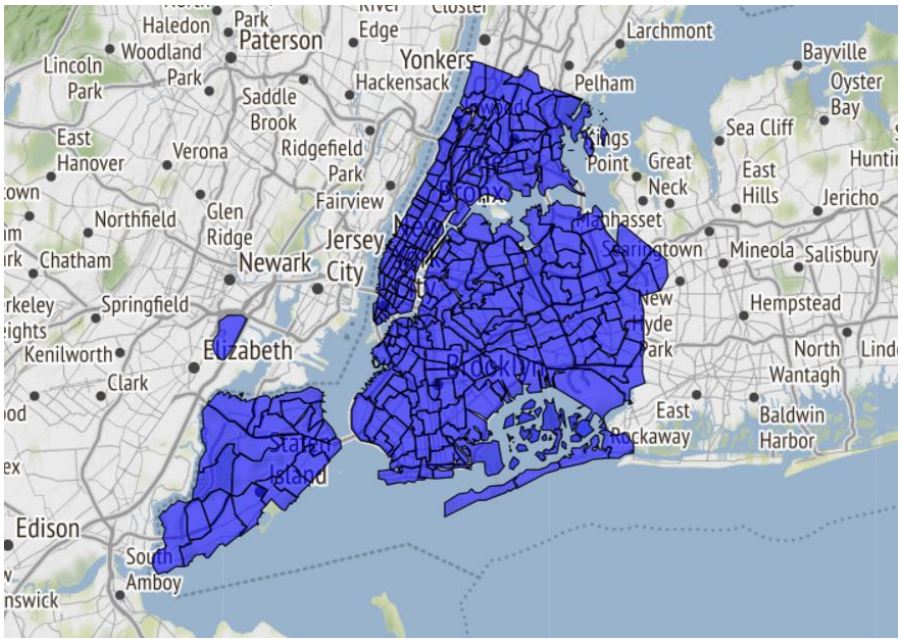
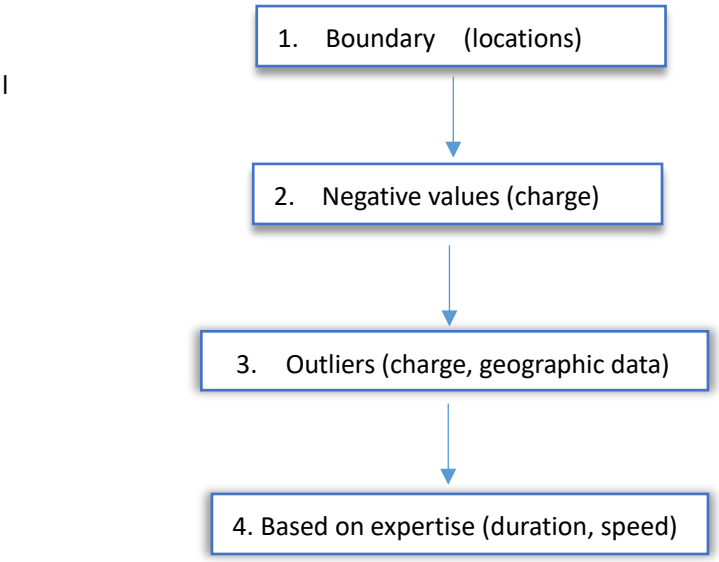


Figure 4. The direct view of NYC map with black and white colors

The figure 3 and figure 4 give basic insight for the area that the project will analyze.

4 Implementation

4.1 Data Pre-processing



Data pre-processing is an indispensable part of data analysis, especially for the million size large data set. From the official data dictionary from TLC, we noticed that the records for each trip share no connection with the server. Most of the record come from the memory of the taxi and still small amount of data was potentially recorded manually (具体数据). Therefore, without a doubt, the data set may include varies of errors and outliers, those strange data points will has a negative impact on the further statistical analysis and visualization plots.

Firstly, we have noticed that the data set is consist of three parts, time data (pick-up time and drop-off time), geographic data (latitude and longitude), charge (tip amount, tax, trip fare), and other attributes (VendorID, payment type). Based on those data, we will do the data pre-processing according to our need.

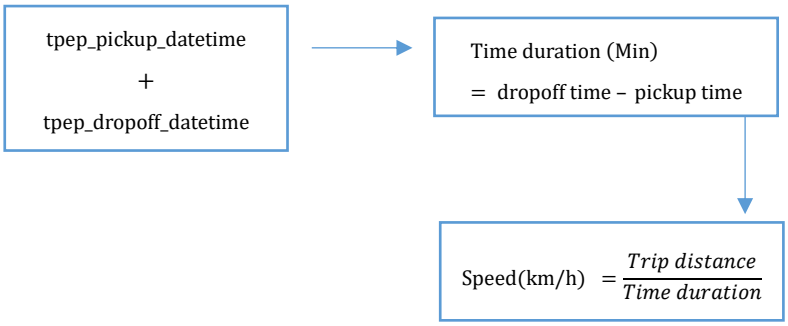
Next, a boundary has to be set and the geographic data has to be filtered according to the NYC region. In this case, we found the NYC boundary data by googling, and after applying the boundary we found *266,958* data are out of the range and some of them located at the sea area, which is not reliable. Therefore we removed all the rows of out-rangers. Eventually, *266,958* data were deducted and *12,482,028* data left.

Moreover, either negatives or extreme values exist in the data set. By calculating the maximum and minimum range of each continuous data columns, we found strange data points exist in every continuous attribute. Therefore, two functions were designed for excluding the negative values and outliers, especially we set the z-score to be *2.58*, with which *0.99* quantiles range was applied. Finally, after cleaning the data set, *833,790* number of data were deducted and *11,648,238* data left.

Finally, we dropped the "other attributes", and randomly select approximate 20% data size from the cleaned dataset for reducing the complexity. Now 2,300,000 data left.

4.2 Data Engineering

The original data set gives no further insight to our project, new attributes are needed for comprehended analysis. In this project, based on the original attributes, new attribute such as trip duration and speed can be derived.



It is worth to mention that there are many unreasonable data points exist in the data set. It is really difficult to filter those data points just according to the pick-up time and drop-off time. For example, some data show that many customers only take the taxi for less than 1 minute, which we think it is reasonable to have this value.

Figure 5 shows the implausible data for trip duration. As we can see from the figure 5, many records indicate the trip only take few seconds from pick-up to drop-off.

	Index	Pickup_time	Dropoff_time
0	4799930	2015-01-28 17:06:02	2015-01-28 17:06:23
1	1146164	2015-01-23 13:25:46	2015-01-23 13:25:56
2	189162	2015-01-06 17:00:16	2015-01-06 17:00:30
3	6019539	2015-01-22 12:18:15	2015-01-22 12:18:29
4	10815187	2015-01-19 19:27:00	2015-01-19 19:27:03
...	...	...	...
5578	1966469	2015-01-17 13:18:36	2015-01-17 13:18:52
5579	1269037	2015-01-21 01:00:54	2015-01-21 01:00:58
5580	8324431	2015-01-28 20:57:26	2015-01-28 20:57:33
5581	2774183	2015-01-09 13:55:43	2015-01-09 13:56:12
5582	9015759	2015-01-07 13:22:46	2015-01-07 13:23:01

5583 rows × 3 columns

Figure 5. Example of improper trip time records

Therefore, this problem will further cause the longest duration time to be 1400 minutes, and largest speed over 1000 (km/h). In this case, we will do the data-preprocessing based on our expertise.

Figure 6 shows the example of implausible data for taxi speed. From the screenshot, it shows somehow, it only take the taxi 0.01hours to drive incredibly far away.

	Index	trip_distance	time_duration	speed
0	6052345	26.5	0.01	2650.0

Figure 6. Example of improper speed records

According to our distribution plot, the passenger is less likely to spend more than 4 hours on taxi,

but we do allow some exceptions to exist, therefore we set the time duration to be 5 hours, which equivalent to 300 minutes. Moreover, based on our expertise, the speed limit in NYC will be 70 KM/H, thus the limit is set to be 70. Further data cleaning will be processed based on those standard.

Figure 7 shows the distribution plot of "Time duration" attribute before cleaning

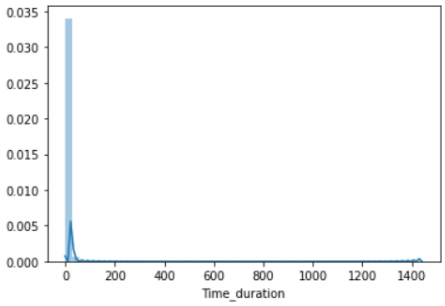


Figure 7. Distribution plot of Time duration attribute (before cleaning)

Figure 8 shows the distribution plot of "Time duration" attribute after cleaning
Figure 9 shows the distribution plot of "Speed" attribute after cleaning

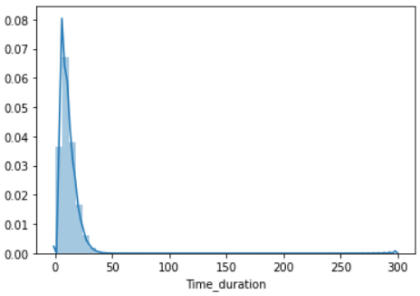


Figure 8. Distribution plot of Time duration attribute (after cleaning)

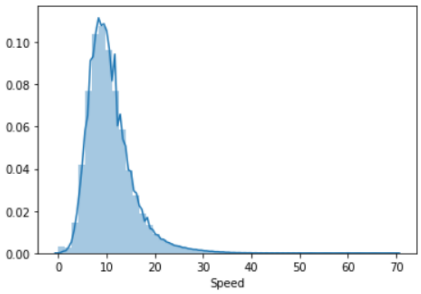


Figure 9. Distribution plot of Speed attribute (after cleaning)

Now, total *2,778* data were removed, and the distribution is pretty clear right now, which is good. Notably, more feature engineering and further data cleaning might be necessary for further statistical analysis.

5 Visualization and Plot Analysis

To help understand the map, the project provides several visualization plots which combine with clear introduction and inference.

5.1 Geographic plot

5.1.1 Pick-up location

Yellow taxi in NYC provides transportation exclusively through street-hails. As the figure shows the pick-up points spread across five boroughs. However, the distribution is not even, since most of the pick-up points are located at Manhattan.

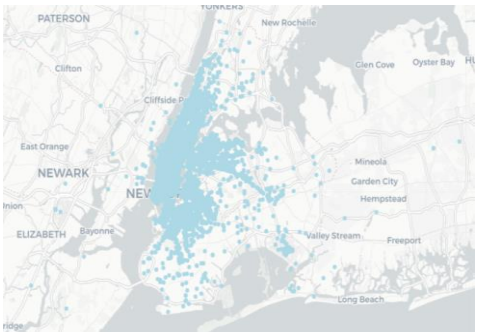


Figure 10. Pick-up locations with partial dataset

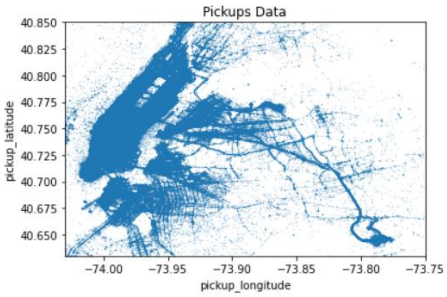


Figure 11. Pick-up location with full dataset

5.1.2 Drop-off location

The Drop-off locations are relatively even, even though Manhattan is still the location with largest amount of drop-off.

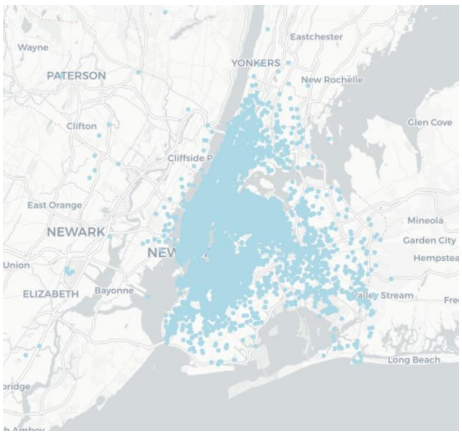


Figure 12. Pick-up locations with partial dataset

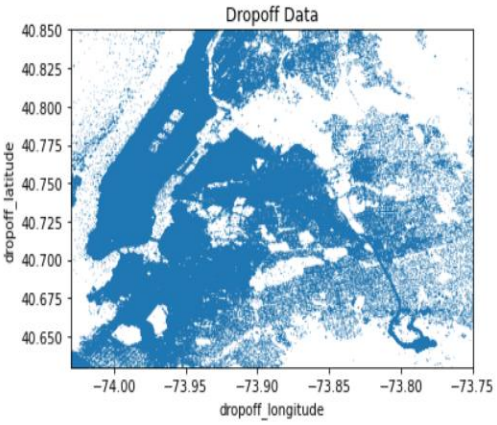


Figure 13. Pick-up location with full dataset

5.1.3 Heat Map

Pick-up:

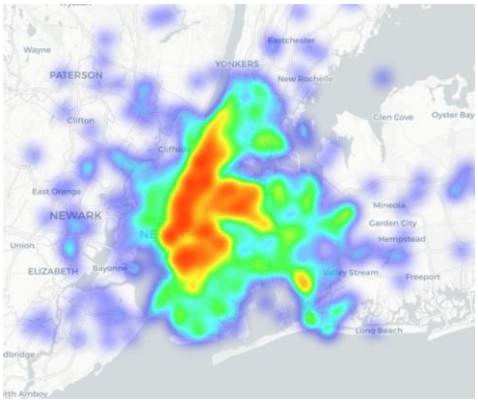


Figure 14. Heat map of Pick-up locations

The heatmap show that which part of the area has the highest pick-up frequency and it reconfirms that Manhattan represets as the busiest pick-up location. However, it is worth to mention that it will be a difficult task to analyze the NYC taxi market as a whole, since more than 80% of data were recorded in Manhattan.

Drop-off:

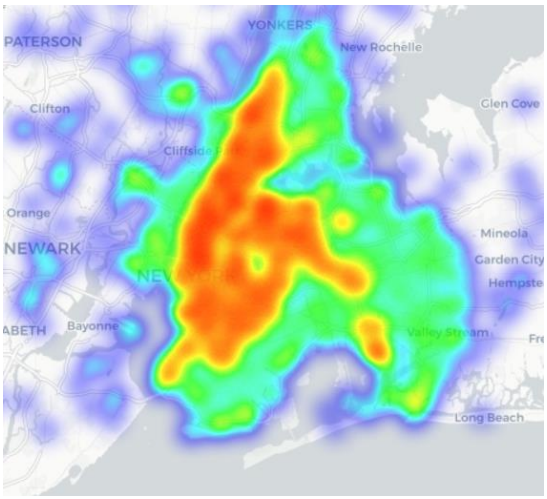


Figure 15. Heat map of drop-off locations

Comparing with the heat map for pick-up, the drop-off heat map also indicate that the Manhattan has the highest drop-off frequency, whereas the drop-off locations are relatively more sporadic. It could mean people who work at Manhattan are more likely to call the yellow taxi for a short business trip, and only a few amount of passengers are willing to take the yellow taxi for a long distance.

5.1.4 Cluster plot

Pick-up

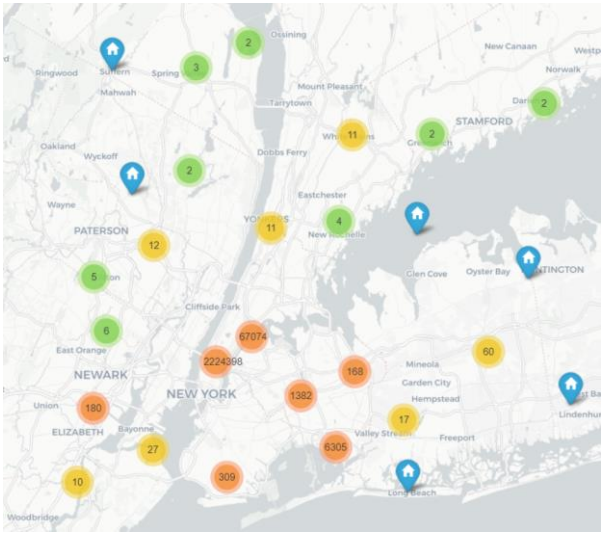


Figure 17. Cluster map of pick-up locations

The cluster map gives us a direct view of the frequency of pick-up in a specific number. As we can see from the figure, Manhattan area occpuys the greatest number of pick-ups with more than 80% from total pick-ups. In this case, it might be interesting to have a look that what makes this location special.

Drop-off:

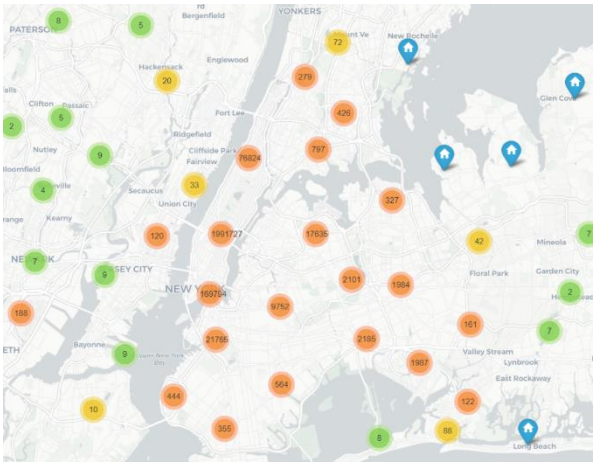


Figure 17. Cluster map of drop-off locations

The Figure xx gives us a view about the drop-off points. Same as before, Manhattan area has greatest number of drop-offs. The more far away, the less drop-off number, expect the airport area.

Basically, those maps above demonstrate a general information for NYC yellow taxi data, which include an overview to the pick-up/drop-off frequency and specific number of them. In the next section, we might want to analyze some specific aspect for NYC yellow taxi, such as charge and duration.

5.1.5 Choropleth plot with trip distance

Choropleth map uses differences in shading to represent the particular quantity.

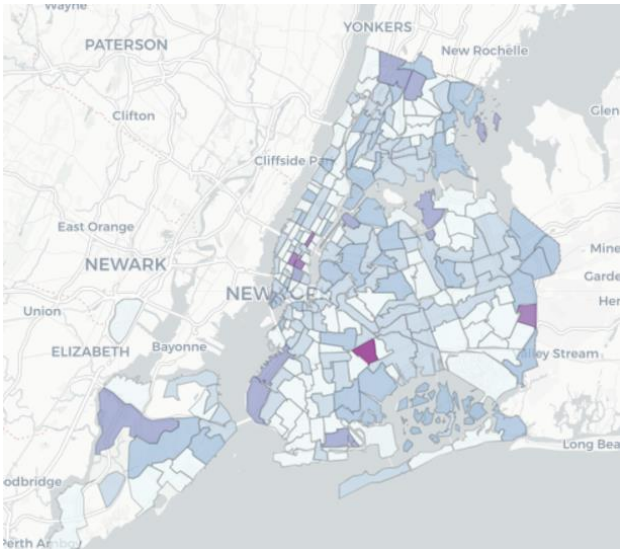


Figure 18. Choropleth map for trip_distance

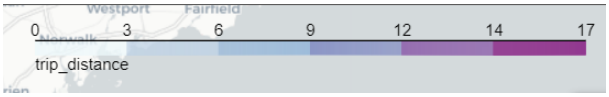


Figure 19. Index for trip_distance

The figure xx shows the trip distance across each area. The area with darker area means that the yellow taxi drives more far away from the pick-up points. A rough overview without more statistical analysis to the figure shows only few trips exceeds 12 KM from the pick-up points, and most trips are 0-9 KM from the pick-up points. It might elucidate that people in NYC is less likely to take the yellow taxi for a long trip and we are not sure whether it is related with the traditional purpose of the yellow taxi.

5.1.6 Choropleth plot with total amount

Total amount does not only represent the earning of yellow taxi drivers, but also indicate the demand and supply in the NYC taxi market. It is an important standard to measure the regional wealth, therefore the project provide the analysis based on total amount.

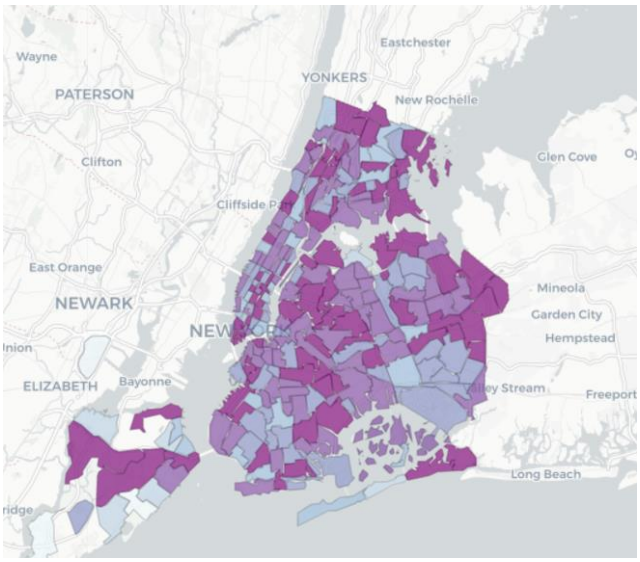


Figure 20. Choropleth map for total_amount

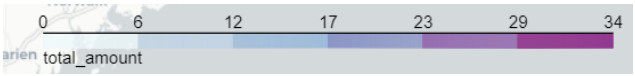


Figure 21. Index range for total_amount

Figure xx shows the total amount that passenger need to be charged, the darker the color, the higher total amount. From the figure, it shows the total charge of the yellow taxi is relatively high, most of areas exceeds 23 US dollars. Combining with the figure of trip distance (figure xx), these two figures obviously indicate that there are some association between fare and trip distance, since the dark areas in the figure of trip distance are normally dark in total amount as well. Therefore, fare amount per kilometer is showing to be expensive in NYC, and this might cause yellow taxi is only for short trip purpose.

5.1.7 Choropleth plot with trip duration & speed

Duration and speed are not the original attributes, but they might be useful, when we want to analyze the efficiency of traffic circumstance in NYC.

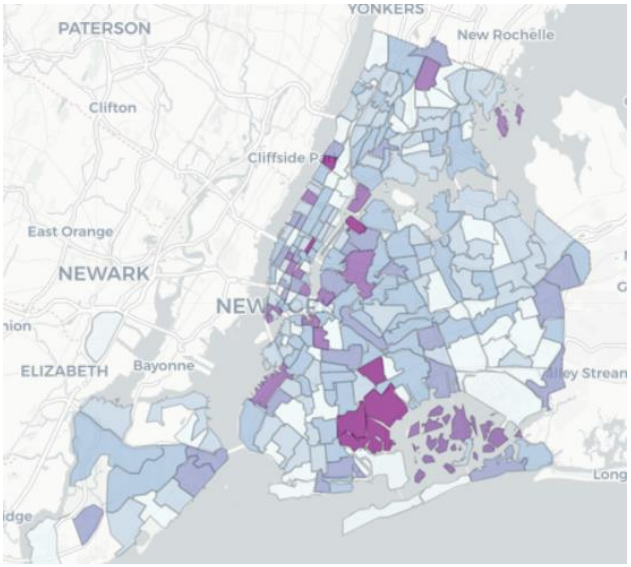


Figure 22. Choropleth map for time_duration

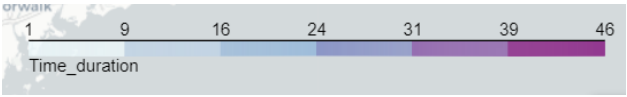


Figure 23. Index range for timr_duration

This figure shows the trip duration for each trip, which means how long it takes between pick-up location and drop-off location. From this figure, it demonstrates that trip duration from most areas are below 24 minutes and only few areas take above 40 minutes. This result might depend on the speed limit and traffic flow in each area, which means beside the trip distance, there are many other dependent variables to influence the trip duration. Therefore, trip duration might be hard to predict and it also share not much similarity with the choropleth map of trip distance.

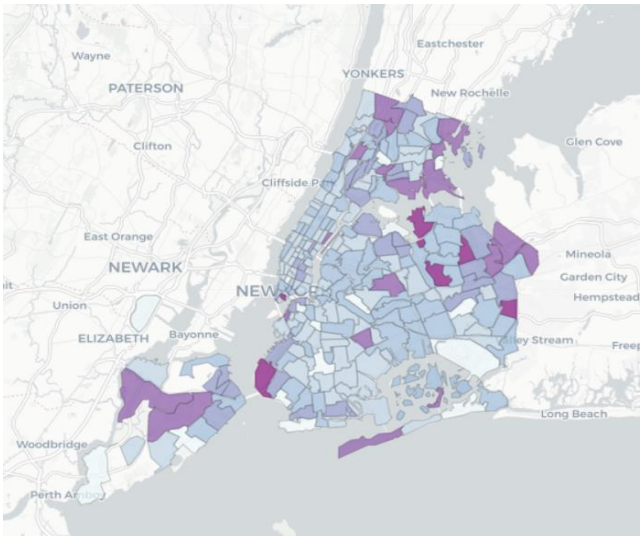


Figure 24. Choropleth map for speed

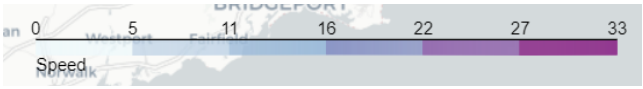


Figure 25. Index range for speed

Figure xx is the choropleth plot in terms of speed. The darker the color, the higher speed.

Combing with choropleth plot in terms of speed, it is reasonable to make the inference that trip duration might be correlated with speed, as longer trip duration is associated with lower speed, and two plots about speed versus trip duration present below.

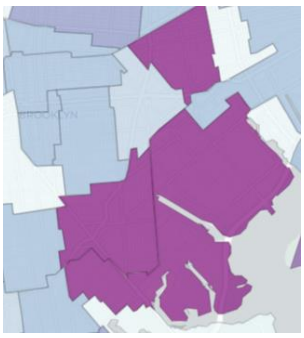


Figure 26. Partial map for time_duration

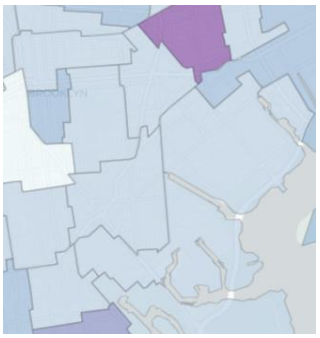


Figure 27. Partial map for speed

6 Discussion

6.1 Weakness

Although this report has made a comprehensive analysis on the geographic attribute of NYC taxi market, there are still many aspects need to be improved. Firstly, according to the NYC Department of City Planning's Neighborhood Tabulation Areas (NTAs), TLC is no longer provide the specific locations for pick-up and drop-off. Therefore, it will be a tough work to visualize the dispersive pick-up/drop-off point based on a rough location. In this case, this project adapts the method to use separate trip datasets from 2015 and 2020. However, the value of visualization plots are questionable, because dependence of two datasets from different years are uncertainly and hard to quantify.

Secondly, randomly sample the original dataset will lose the information in an unpredicted way. The original dataset contains 20 million sizes data, and it is a heavy burden for our laptop to process all at once. For avoiding the memory error and increasing the processing speed, the project randomly selected 20% of the data from the original dataset. It is a not huge problem to do the geographic visualization, and patterns are still able to be observed from the reduced dataset. Nevertheless, the approach might not be a proper solution for statistical analysis, since much of information will lose if only 20% of the data are selected. Thus, bias will inevitable occurred.

6.2 Improvement

To make an improvement for further use, two solutions are targetedly come out. Firstly, to deal with the large dataset, we can use Spark or install pysprak in advanced. Secondly, to avoid the bias from randomly sampling the dataset, we can design a function to process the whole dataset progressively.

7 Conclusion

To sum up, the project presents with a spatial big data analysis and it mainly focuses on the geographic visualization and plot analysis. Specifically, this project took the data from TLC official website and visualized some attributes according to our need. In this project, data pre-processing and data engineering have been done in advanced and then several analyses and inferences were well demonstrated based on attribute visualizations. Following are some interesting findings.

For the Yellow taxi in NYC, most of the records appears at Manhattan City. This bias might be associated with the traditional purpose of yellow taxi, for which yellow taxi are more likely to take as a short business trip. Moreover, even without a fundamental knowledge, a conclusion that it will not a low-cost choice to take the yellow taxi for a long trip in NYC. Furthermore, just according to choropleth plots, it not difficult to find total charge shares some connections with trip distance, and trip duration is associated with speed.

Ultimately, to avoid the further bias, randomly sampling the data are not suggested as a proper way to reduce the complexity. A progressive function and pyspark library can the solutions. Moreover, to justify the correlation between the attributes, more data engineering and data cleaning should be applied and statistical test are supposed to be done in further project.

8 Reference:

New York City Taxi & Limousine Commission. (2015). Yellow Taxi Trip Record [CSV file]. Retrieved from https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2015-01.csv

New York City Taxi & Limousine Commission. (2020). Yellow Taxi Trip Record [CSV file]. Retrieved from https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2020-01.csv

New York City Taxi & Limousine Commission. (2020). Taxi Zone Shapefile [CSV file]. Retrieved from https://s3.amazonaws.com/nyc-tlc/misc/taxi_zones.zip

New York City Taxi & Limousine Commission. (2020). Yellow Trip Data Dictionary [PDF file]. Retrieved from
https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf

Yu. W, MAST30034Repo, (2020), GitHub repository,
https://github.com/weichangyu10/MAST30034Repo/blob/master/workbook/lab1.ipynb

Yu. W, MAST30034Repo, (2020), GitHub repository,
https://github.com/weichangyu10/MAST30034Repo/blob/master/workbook/lab2.ipynb