Table of Contents

Executive Summary

In this document, I will walk through the analysis of New York City Green Taxi Data collected by New York City Taxi and Limousine commission. Green taxis are taxis that are not allowed to pick up passengers inside of the densely populated areas of Manhattan. I will use the data from September 2015 for our analysis which is downloaded from the link below.
http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml.

The NYC Green Taxi dataset contains 21 attributes about the trips taken by New Yorkers across the five boroughs of New York city. These columns contain information such as pickup datetime, drop-off datetime, trip distance, amounts charged (including the actual meter charge as well as taxes), pickup latitude-longitude, drop-off latitude-longitude and tip to the driver. The NYC Green Taxi is targeted in areas historically underrepresented by Yellow cabs and majority of the trips are street hails over dispatch.

Through the analysis, below patterns are observed:

- Majority of the trips are short distance trips (within 6 miles) - 88% of our data are short distance trips
- The highest average number of trips are taken between 5am to 7am
- The analysis indicates that majority of the NYC Green Taxi trips are non-airport area trips which confirms the fact that Green Taxis are not allowed in JFK airport which is predominantly served by yellow cabs – only 2.9% of the trips were NYC airport area trips
- Random Forest gives us the best RMSE value for predicting the tip as a percentage of total fare
- The busiest boroughs are Brooklyn and Manhattan followed by Queens. While Brooklyn has the highest number of pickups, Manhattan has the highest number of drop offs.
- It was interesting to find that even though the number of pickups is very less in the yellow pickup turfs, i.e. Manhattan below 110th St. on the West Side, and below 96th St. on the East Side, there are still fair share of drop-offs in this area
- Additionally, the busiest neighborhoods are North Side-South Side Brooklyn, DUMBO – Brooklyn Downtown, East Harlem, Central Harlem and Astoria
- The traffic movement of NYC Green taxi is majorly within the same boroughs rather than inter boroughs. This might be due to the fact that most of the trips are short distance trips
- Majority of the inter borough traffics are from Brooklyn to Manhattan

Data Analytic Tool and Packages Used

Along with the basic libraries needed for data analytics, I have used several other libraries such as matplotlib, folium and geopandas for data visualization and reading shapefiles in Python. I have also used Python Machine Learning packages to build the predictive model for tip percentage.

Problem 1 - Data Download and Data Statistics

Two different dataset has been used for the analysis:

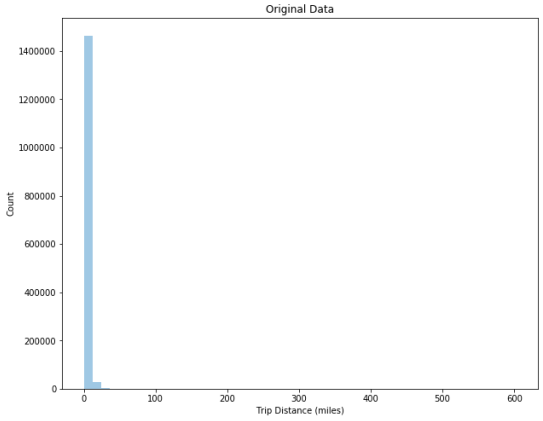
- NYC Green Taxi Trip Report Data for September 2015: The dataset has been downloaded directly into the Python using the link. https://s3.amazonaws.com/nyc-tlc/trip+data/green_tripdata_2015-09.csv
- NYC Neighborhood Tabulation Areas: This data contains information about NYC borough and neighborhood boundaries as created by the NYC Department of City. The corresponding shapefile has been downloaded from below link. https://data.cityofnewyork.us/City-Government/Neighborhood-Tabulation-Areas/cpf4-rkhq

Below table provides more information about these datasets:

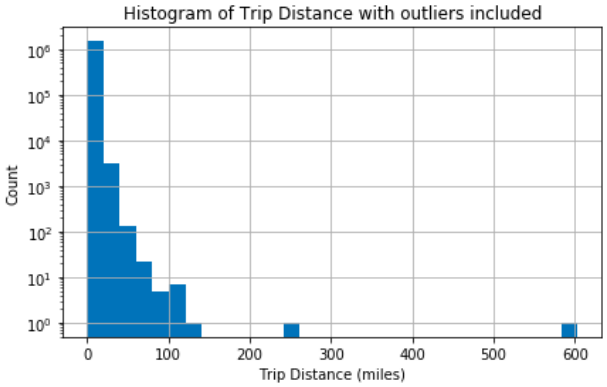
Dataset	No of Rows	No of Columns
Trip Data for September 2015	1494926	21
NYC Neighborhood Data	195	8

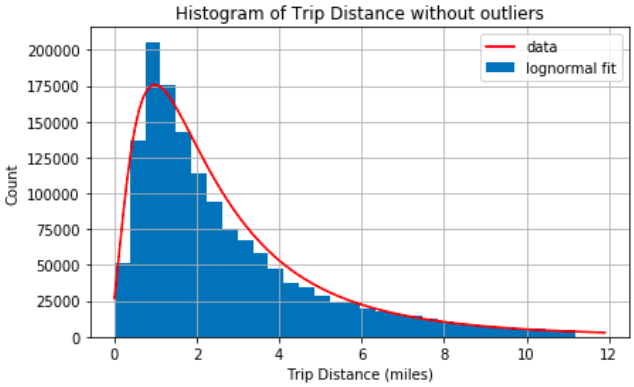
Problem 2- Exploring Trip Distance Field

- Plot a histogram of the number of the trip distance ("Trip Distance").



The above histogram does not illustrate our data properly. Hence, I will plot the raw Trip distance on logarithmic scale as well as remove the outliers from Trip Distance and plot the data points for better visualization. *Here, outliers are defined as data point located outside 3 standard deviations from the median.*





- Report any structure you find and any hypotheses you have about that structure.

  The NYC Green Taxi was commissioned to pick up street hail passengers in areas not commonly served by yellow medallion cabs. The goal of the Green Taxi program is to improve access to street-hail transportation throughout the five boroughs in areas of New York City historically underserved by the yellow taxi industry. Since it is targeted to street hail passengers, I can hypothesize that the majority of the trips will be short distance.

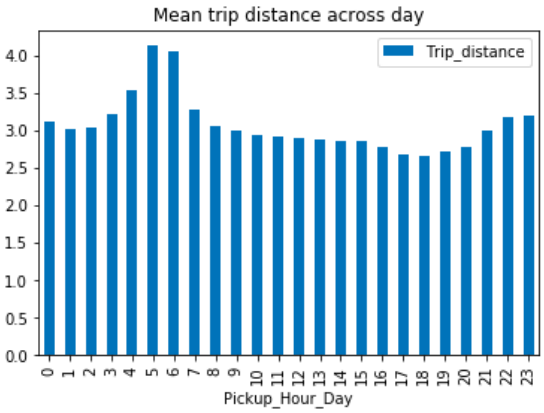
  The above histogram informs that the distribution of Trip distance for NYC Green taxi data is not normally distributed and skewed towards right, i.e. 88% of our data has short distance trips (within 6 miles).

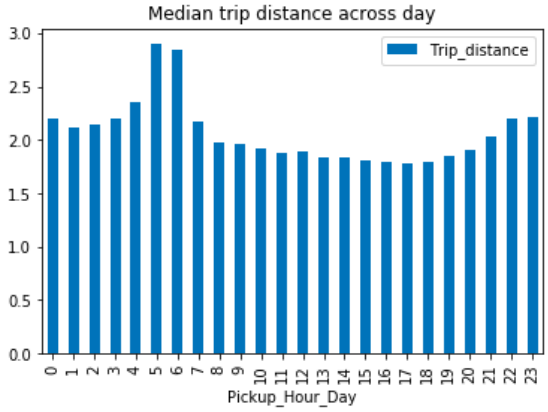
Problem 3.a – Trip Distance Grouped by Hour of the day

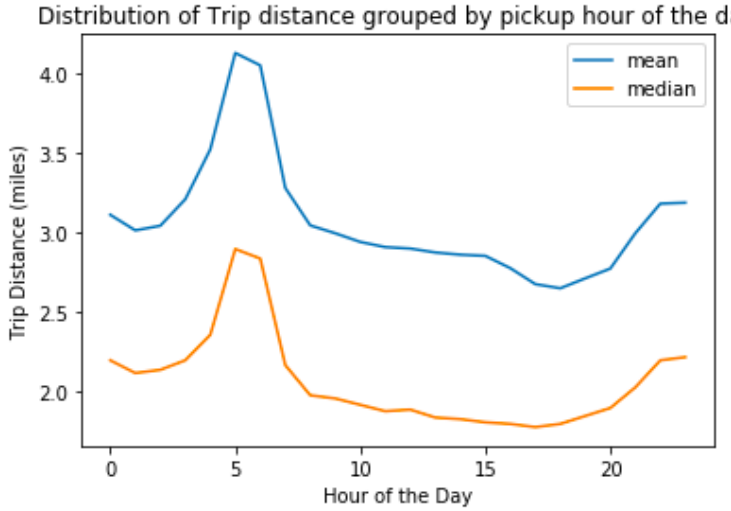
- Report mean and median trip distance grouped by hour of day.

Mean and Median of Trip distance grouped by pickup hour of the day

Pickup_Hour_Day	mean	median
0	3.115276	2.20
1	3.017347	2.12
2	3.046176	2.14
3	3.212945	2.20
4	3.526555	2.36
5	4.133474	2.90
6	4.055149	2.84
7	3.284394	2.17
8	3.048450	1.98
9	2.999105	1.96
10	2.944482	1.92
11	2.912015	1.88
12	2.903065	1.89
13	2.878294	1.84
14	2.864304	1.83
15	2.857040	1.81
16	2.779852	1.80
17	2.679114	1.78
18	2.653222	1.80
19	2.715597	1.85
20	2.777052	1.90
21	2.999189	2.03
22	3.185394	2.20
23	3.191538	2.22







Problem 3.b – NYC Airport Area Trips

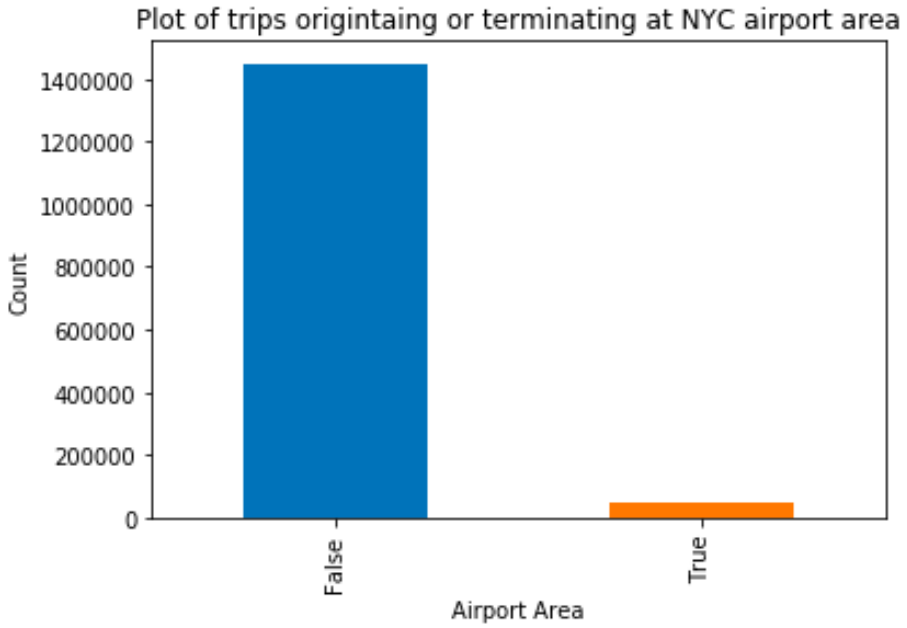
- We'd like to get a rough sense of identifying trips that originate or terminate at one of the NYC area airports. Can you provide a count of how many transactions fit this criterion, the average fare, and any other interesting characteristics of these trips?

  In order to understand the trips originating or terminating at one of the NYC area airports, the column "RateCodeID" has been explored. As per the data dictionary, this column indicates the area of final rate charged. Hence, this column is being used as an indication of NYC airport area trip, 2 = JFK and 3 = Newark. Using this approach, I get below information about the total number of trips that fit into airport area trips, their average fare and average total amounts charged.

```
The number of trips that originate or terminate at one of the NYC area airports is 5552
The average fare of trips originating or terminating in NYC airports area is 48.976945244956774
The average total amounts charged for trips originating or terminating in NYC airports area is 57.20842038904719
```

  However, this was a very brute force approach. The data dictionary indicates that the RateCodeID is the final rate code in effect at the end of the trip and hence might not consider trips originating at the airport. Also, apart from code 2 and code 3 I have code 5 which indicates negotiated fare and code 6 for group fare. Hence, this approach might not give us a good approximation for NYC area airport trip.

As a more sophisticated approach to find out trips originating or terminating in the NYC airports area, the pickup and drop off longitude and latitude information will be used. It is being assumed that any latitude and longitude within the 1-mile radius of any of the three NYC airports (JFK, LaGuardia and Newark) are marked as NYC area airport. I will use python math library to calculate the distance between two given longitude and latitude. The entire source code can be viewed in Jupyter notebook. As per this approach, the total number of trips originating or terminating in the NYC airports area is 43949, i.e. only *2.9% of total number of trips for Green Taxi is for the NYC airport area.* The average fare and average total amounts charged for these trips are shown below.



	Fare_amount					Total_amount				
	mean	median	min	max	std	mean	median	min	max	std
Near_Airport										
False	12.128541	9.5	-475.0	580.5	9.495778	14.525869	11.3	-475.0	581.3	10.767754
True	26.233087	22.5	-45.0	450.0	16.970860	31.746895	28.5	-45.0	450.0	20.601498

Problem 4 – Predicting derived variable Tip Percentage

In this section, I will explore the New Yorkers' tipping habit for NYC Green Taxi data. A derived variable for tip as a percentage of the total fare is built. This section also illustrates the predictive model for predicting the new derived variable for tip as a percentage of the total fare.

Following steps are being taken to build the predictive model:

- Data Cleaning
- Feature Engineering
- Model building

Data Cleaning

Columns of interest have been explored and outliers have been removed from continuous variables. Missing, invalid and NA values have been replaced with median and mode for continuous and categorical variables respectively.

Please note that outliers are defined as data points outside the 3-standard deviation of the median.

Column Name	Missing Value	Invalid Values	Outliers
Trip_distance	No	No	Removed
Tip_Amount	No	Yes, some negative values – replaced with median	Removed
Total_Amount	No	Yes, some negative values – replaced with median	Removed

Feature Engineering
4 new columns have been created using the pickup datetime and drop-off datetime:

Original Column	Derived Column 1	Derived Column 2
Pickup_datetime	pickup_hour	pickup_week
Drop-off _datetime	drop-off _hour	drop-off _week

Model Building

A derived variable for tip as a percentage of the total amount is built using the columns, Tip_amount and Total_amount. I have removed all the data points where the tip amount is smaller than total amount since it indicates an anomaly which might have been due to either New Yorkers being too generous, or some coupon code must have been applied. I have used **Root Mean Square Error (RMSE)** as the performance metric. The following columns are included to build the regression model:

- Trip_distance
- Total_amount
- Pickup_Hour
- Drop-off _Hour
- Pickup_Week
- Drop-off _Week

The Linear Regression model gave a RMSE value of 18.39 and the mean tip percentage is only 17%. This indicates a non-linear relationship between our independent variable and dependent variable. I built a Random Forest model and got a RMSE value of 5.54 which is a big improvement over the linear model.

Additionally, I created another random forest model by adding remaining variables, trip distance, pickup hour, drop off hour, pickup week and drop off week; this model gave a RMSE value of 5.21. This is not a very big improvement over the simple model which means that the additional features are not adding any values into the model.

Hence, the simple random forest with only Total Amount as the independent variable can be deemed the best model so far. I could have further increased the RMSE score by hyper parameter tuning which I didn't do at this point.

Problem 5 - Data Visualization for Intra vs Inter Borough Traffic

This section explores intra-vs. inter borough traffic through data visualization for NYC Green Taxi data. In order to explore inter borough and intra borough traffic, data has been prepared as per below:

- Converted the data frame into Geo dataframe using geopandas and shapely.geometry
- Used spatial join to merge our dataframe and the NYC Neighborhood Tabulation Data

Inter-Borough Traffic

Fig. 1 and Fig. 2 shows the distribution of trips across the NYC city based on the pickup borough and drop-off borough respectively. The NYC Green Taxi is targeted to the areas underrepresented by the NYC Yellow cabs. Hence, the number of pickups in the Manhattan below 110th St. on the West Side, and below 96th St. on the East Side Upper Manhattan area is far less than the number of drop-offs. In fact, a bar plot for the number of pickups and drop-offs for each borough indicates that Manhattan has the highest number of drop-offs. (See Fig. 3)

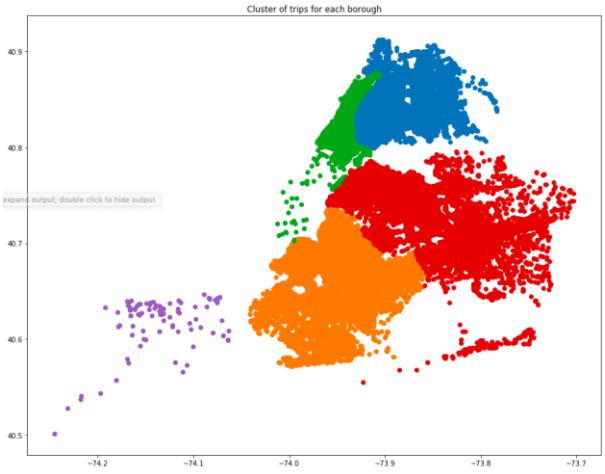


Fig. 1

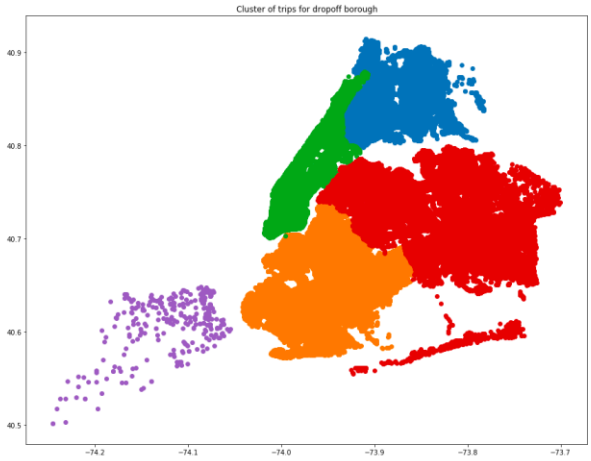


Fig. 2

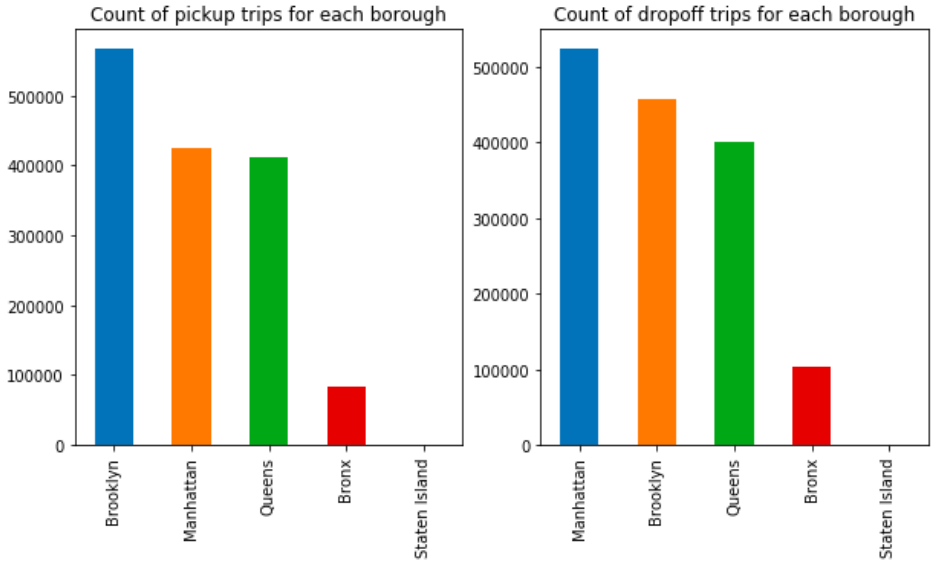


Fig. 3

I also wanted to explore the movement of traffic for our dataset. The heatmap (Fig. 4) shows the movement of traffic among all five boroughs. It is interesting to note that there was no trip that originated at Staten Island and ended in Bronx. It is also clearly visible that since New Yorkers are utilizing the Green Taxi

for short distances, most of the trips are within the same borough and concentrated in highly dense areas of Manhattan, Queens and Brooklyn.

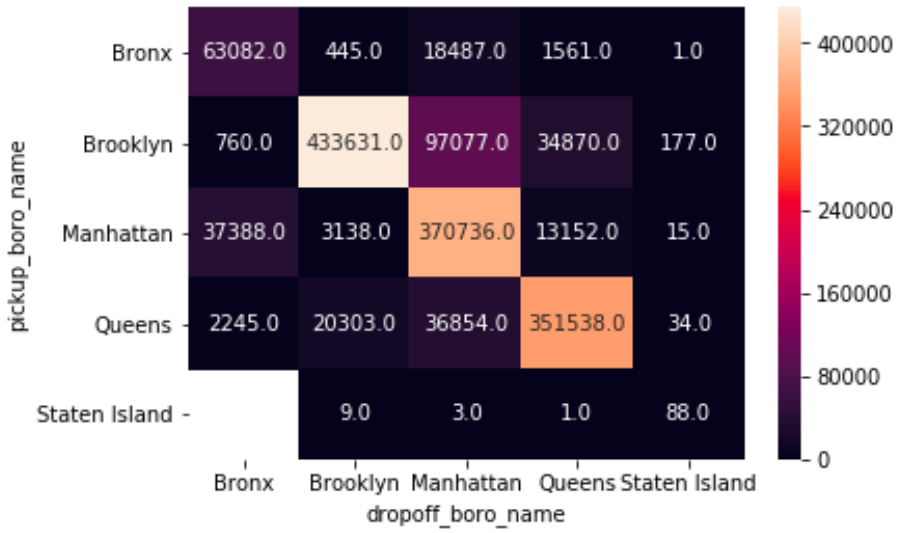


Fig. 4

To further understand the congestion of inter vs intra borough traffic, I calculated the speed of each trips and below table shows the mean and median speed. It's very noticeable that the intra borough trips have lower mean and median speed. This is an indication that there is a high congestion of traffic within the city as compared to inter borough trips since they might go via flyovers.

The mean and median speed for inter vs intra borough trips

	pickup_boro_name	dropoff_boro_name	mean	median
0	Bronx	Bronx	12.9003	11.6814
1	Bronx	Brooklyn	23.4557	22.7611
2	Bronx	Manhattan	14.3113	12.7582
3	Bronx	Queens	25.4898	25.1077
4	Bronx	Staten Island	20.5834	20.5834
5	Brooklyn	Bronx	24.3666	24.5106
6	Brooklyn	Brooklyn	11.676	11.2186
7	Brooklyn	Manhattan	14.0374	13.561
8	Brooklyn	Queens	18.4701	16.9219
9	Brooklyn	Staten Island	25.6982	25.1476
10	Manhattan	Bronx	14.8376	13.2184
11	Manhattan	Brooklyn	19.7144	19.1907
12	Manhattan	Manhattan	12.1242	11.2545
13	Manhattan	Queens	23.1365	22.5238
14	Manhattan	Staten Island	29.0261	26.1222
15	Queens	Bronx	23.7577	23.5897
16	Queens	Brooklyn	17.8731	16.0792
17	Queens	Manhattan	15.7286	14.8164
18	Queens	Queens	12.9246	11.5648
19	Queens	Staten Island	29.1857	30.2185
20	Staten Island	Brooklyn	21.87	21.141
21	Staten Island	Manhattan	31.5597	34.1553
22	Staten Island	Queens	46.1922	46.1922
23	Staten Island	Staten Island	16.6633	17.6769

Since the time of the day and day of the week has a high impact on traffics in the city. I explored how the traffic is distributed throughout the day and through each day of the week based on the pickup and drop off borough information.

Manhattan is busiest during the morning hours traffic, i.e. between 8am to 10 am since it has the highest number of pickups as well as drop offs. However, there is a sharp drop in the number of pickups during the evening rush hours for Manhattan. While the number of pickups increases in Brooklyn and the traffic flows majorly either between Manhattan and Brooklyn or within Brooklyn. Additionally, the traffic is fairly uniform throughout the day in Bronx. (See Fig. 5)

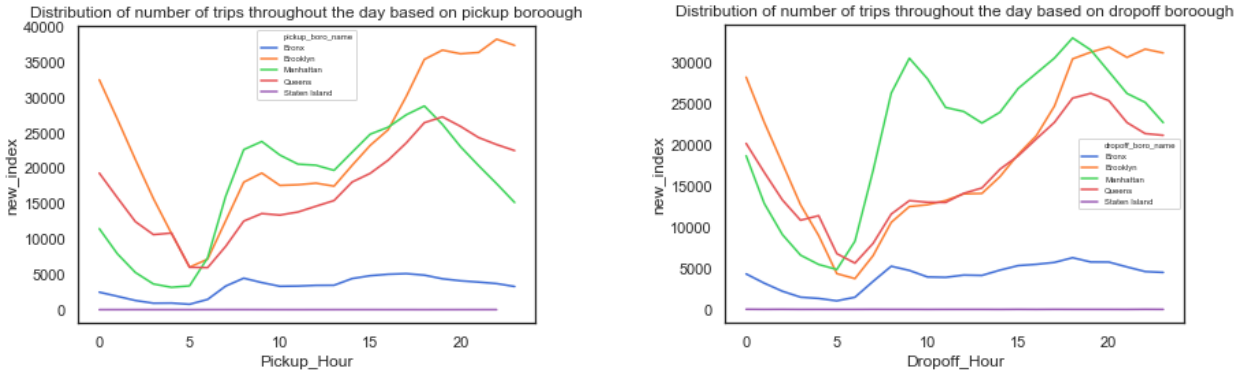


Fig. 5

Exploring the trips data for each week has some interesting observations about New York traffic. Overall, Brooklyn is the busiest borough throughout the week in terms of pickup whereas Manhattan is the busiest in terms of drop off except over the weekends where most of the trips are within Brooklyn. (See Fig. 5)

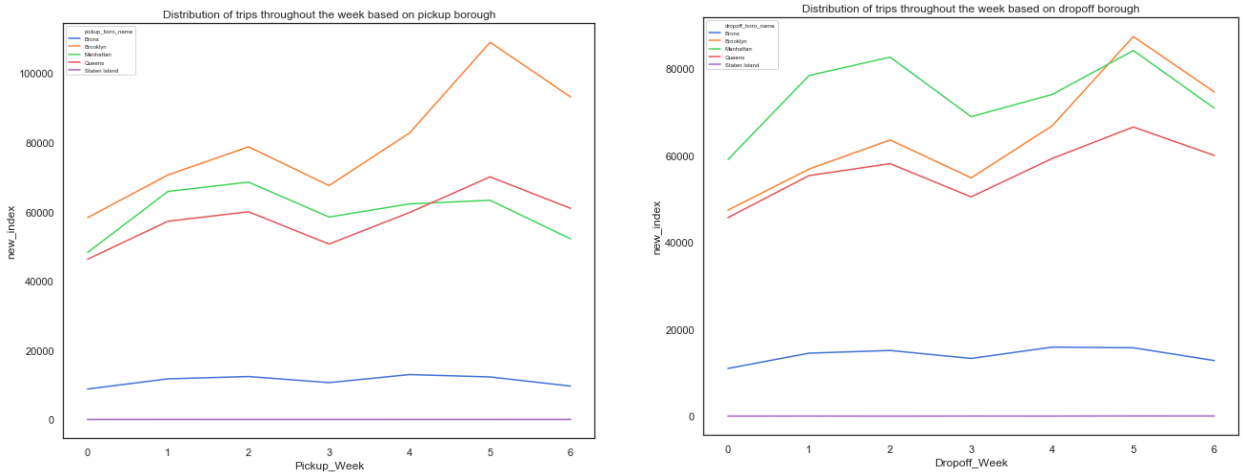


Fig. 6

Intra-Borough Traffic

In order to understand the intra-borough traffic, I have focused on the neighborhood information for each borough. Before digging further, I have looked at the top 5 busiest neighborhoods for each borough. (See Fig. 7, Fig. 7 and Fig. 9)

Borough	**5 Busiest Neighborhoods in descending order**
Brooklyn	North Side- South Side Downtown Brooklyn Park Slope – Gowanus Neighborhood

	Fort Greene Carroll Gardens
Manhattan	East Harlem North Central Harlem South East Harlem South Morningside Heights Central Harlem North
Queens	Astoria Elmhurst Jackson Heights Hunters Point Forest Hills
Bronx	West Concourse Melrose South-Mott Haven North Mott Haven-Port Morris Neighborhood East Concourse-Concourse Village Norwood
Staten Island	Mariner's Harbor New Brighton Port Richmond New Springville Westerleigh

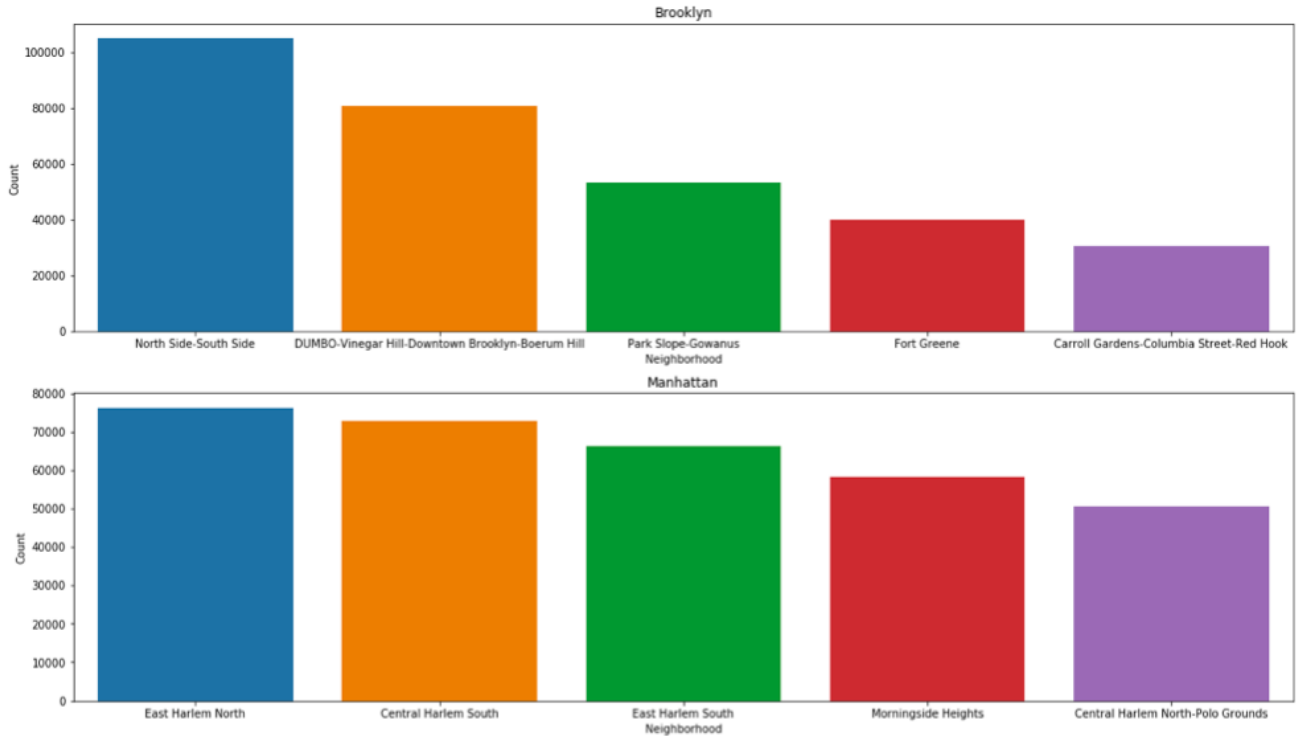


Fig. 7

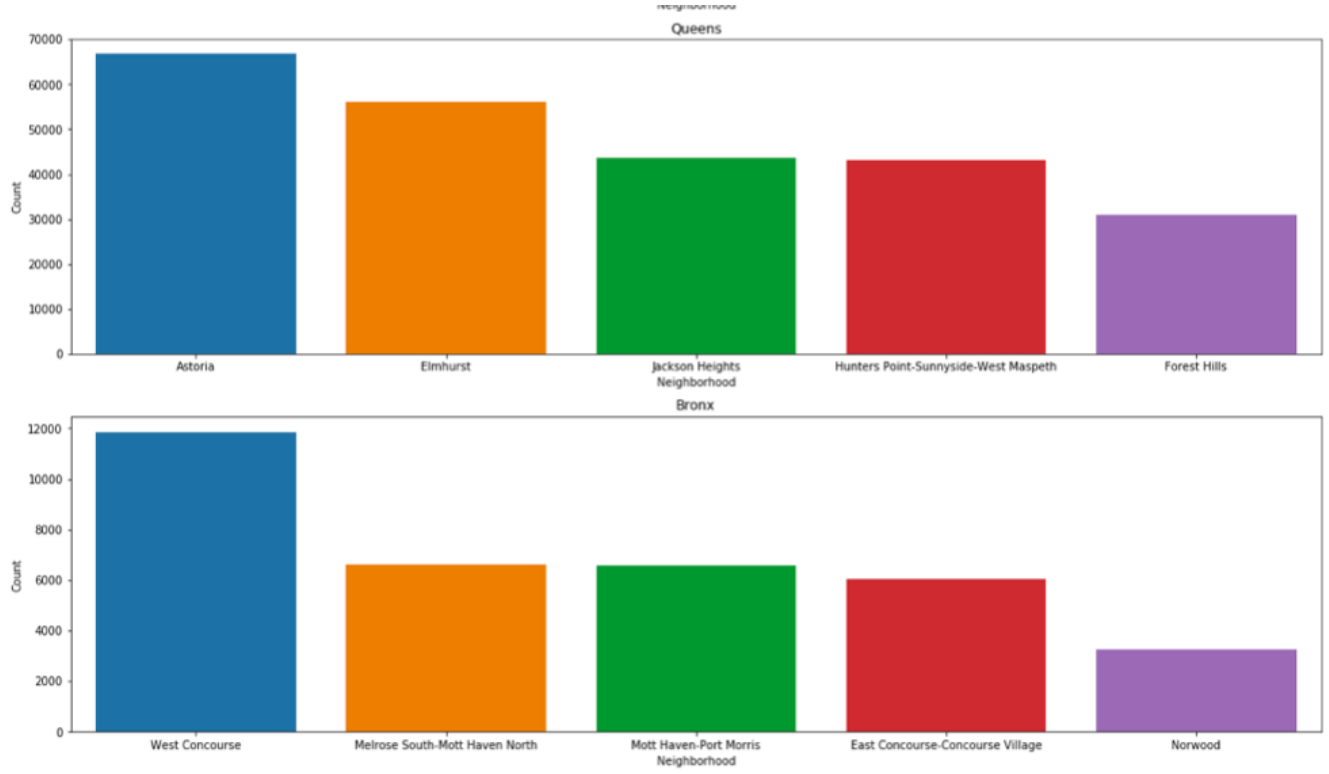


Fig. 8

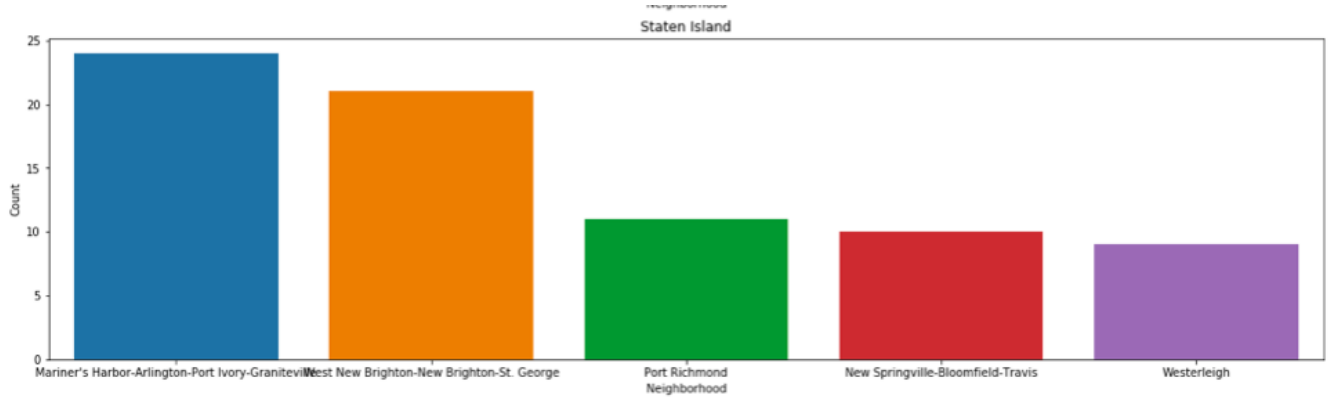
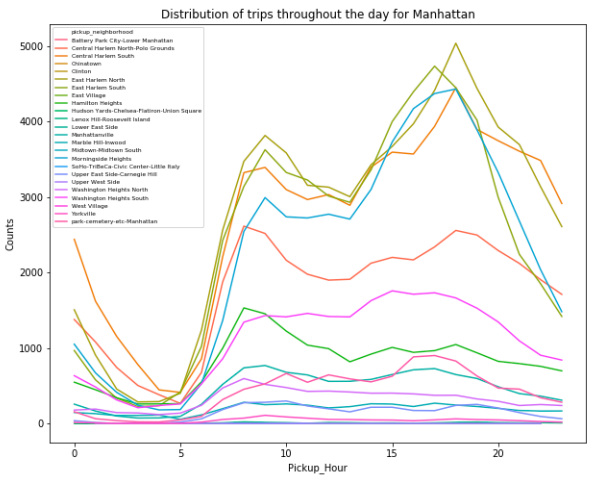
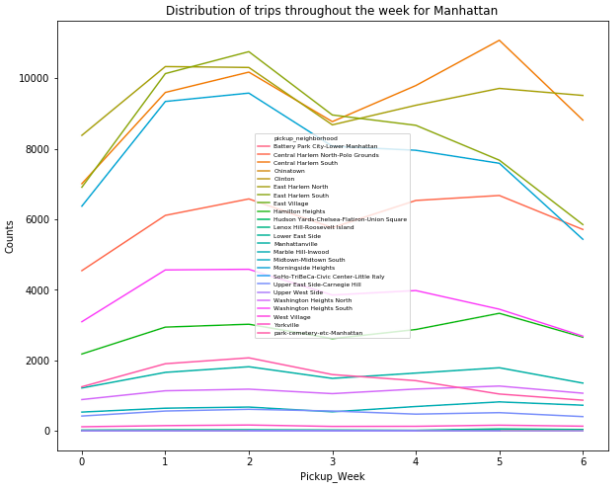


Fig. 9

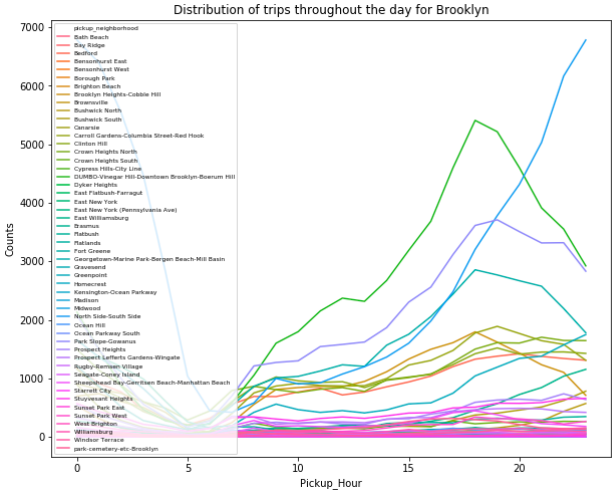
Now, I have explored the data further to look for the patterns for a given borough. The data set has been divided into 5 chunks corresponding to a specific borough. For example, any data point which has either the pickup borough or drop-off borough as Manhattan has been deemed as Manhattan trips. *Please note that the traffic pattern is being explained using the pickup timestamp.*

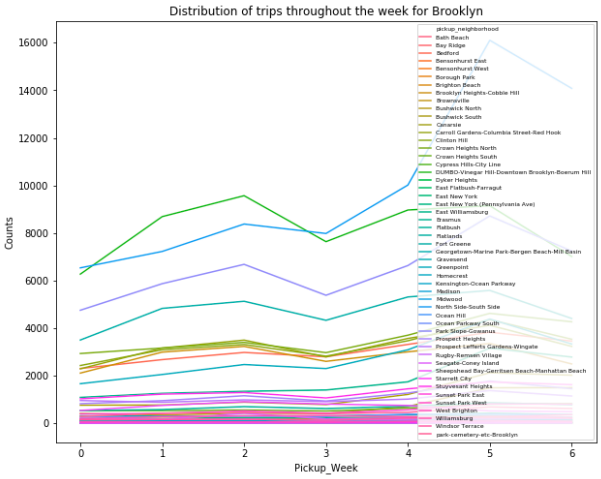
Manhattan: The figure indicates that the peak hours for Manhattan is fairly between 8 am to 9 am in the morning and 6pm to 7 pm in the evening. However, evenings generally have more traffic/trips transactions that mornings. Also, Saturday is the busiest day of the week followed by Tuesday. It is also interesting to note that East Harlem and Central Harlem has the most traffic confirming that these neighborhoods are dominated by working class of New York.



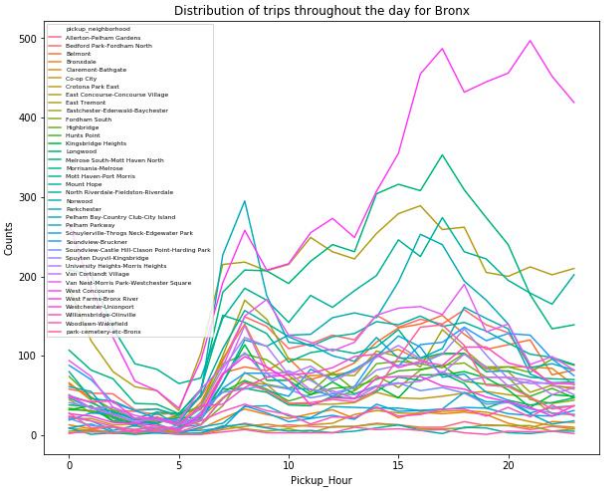


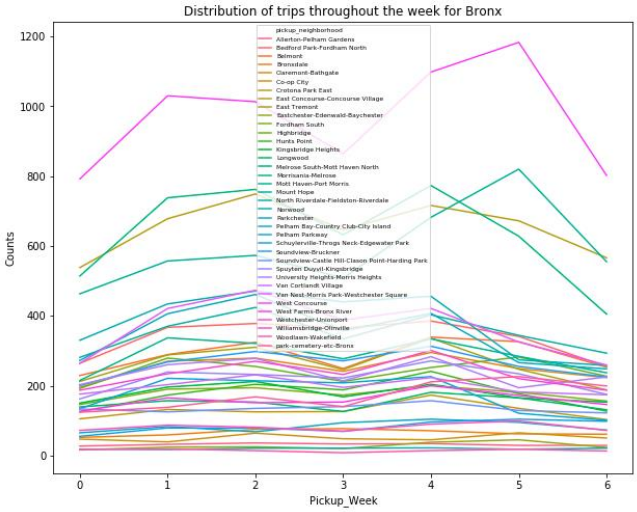
Brooklyn: The distribution of traffic in Brooklyn is very interesting since it does not have a rush hour in the morning. This is indicative of Brooklyn population which is often blanketed under the term hipster. Brooklyn is busiest during late evening hours, 6 pm to 8pm and mostly on Saturdays. The neighborhood has vibrant nightlife and live music centers. Majority of the trips are happening in DUMBO – Downtown Brooklyn and North Side- South side Brooklyn neighborhoods. This distribution of traffic is due to the fact that DUMBO is one of the most visited neighborhoods due to popular dining and recreational spots.



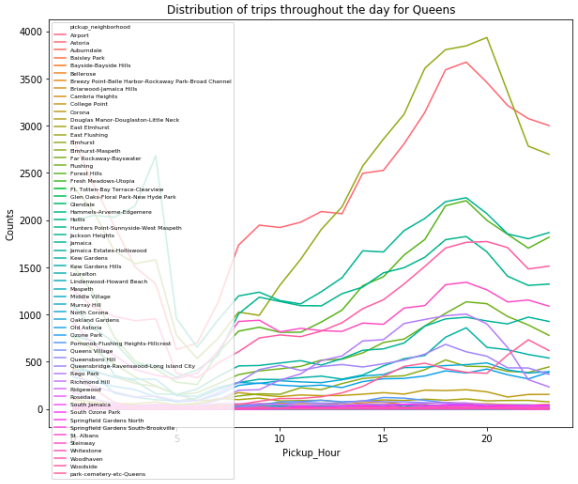


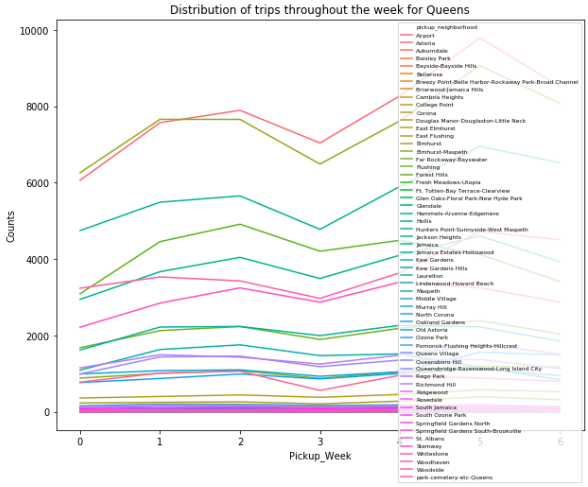
Bronx: Bronx weekly traffic pattern is very similar to Manhattan in terms that both of them has more traffic during the beginning of the week and on Saturdays. It is interesting to see that the majority of the traffic is concentrated in West Concourse area and Melrose-South Mott Haven. However, these two neighborhoods reside two different class of New Yorkers. The graph also indicates that West Concourse has high traffic during evening hours whereas other neighborhoods have similar traffic during the morning and evening rush hours.



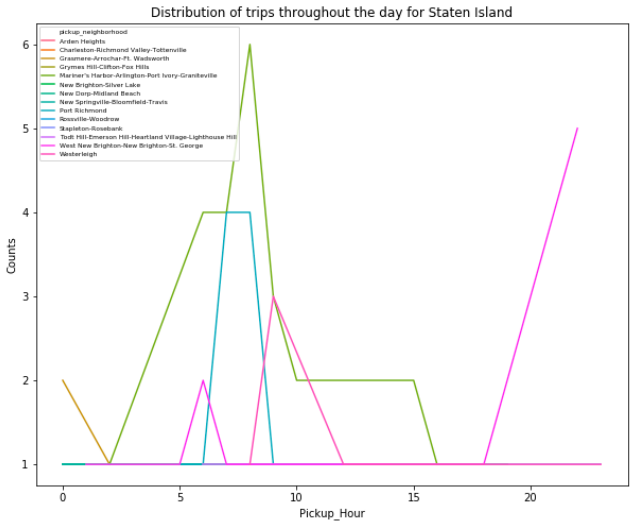


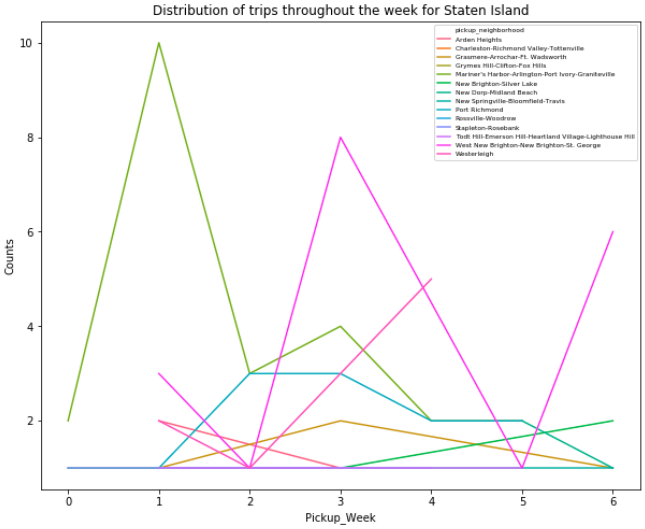
Queens: The NYC Green Taxi's peak hours for Queens is mostly during evenings between 6pm to 8pm. Similar to other boroughs, the weekend traffic is higher than the weekday traffic. The majority of the traffic is concentrated between Astoria and East Elmhurst followed by Forest Hills. Since, these neighborhoods is highly populated by middle class working population, majority of them might tend to take metro to commute to work over taxis.



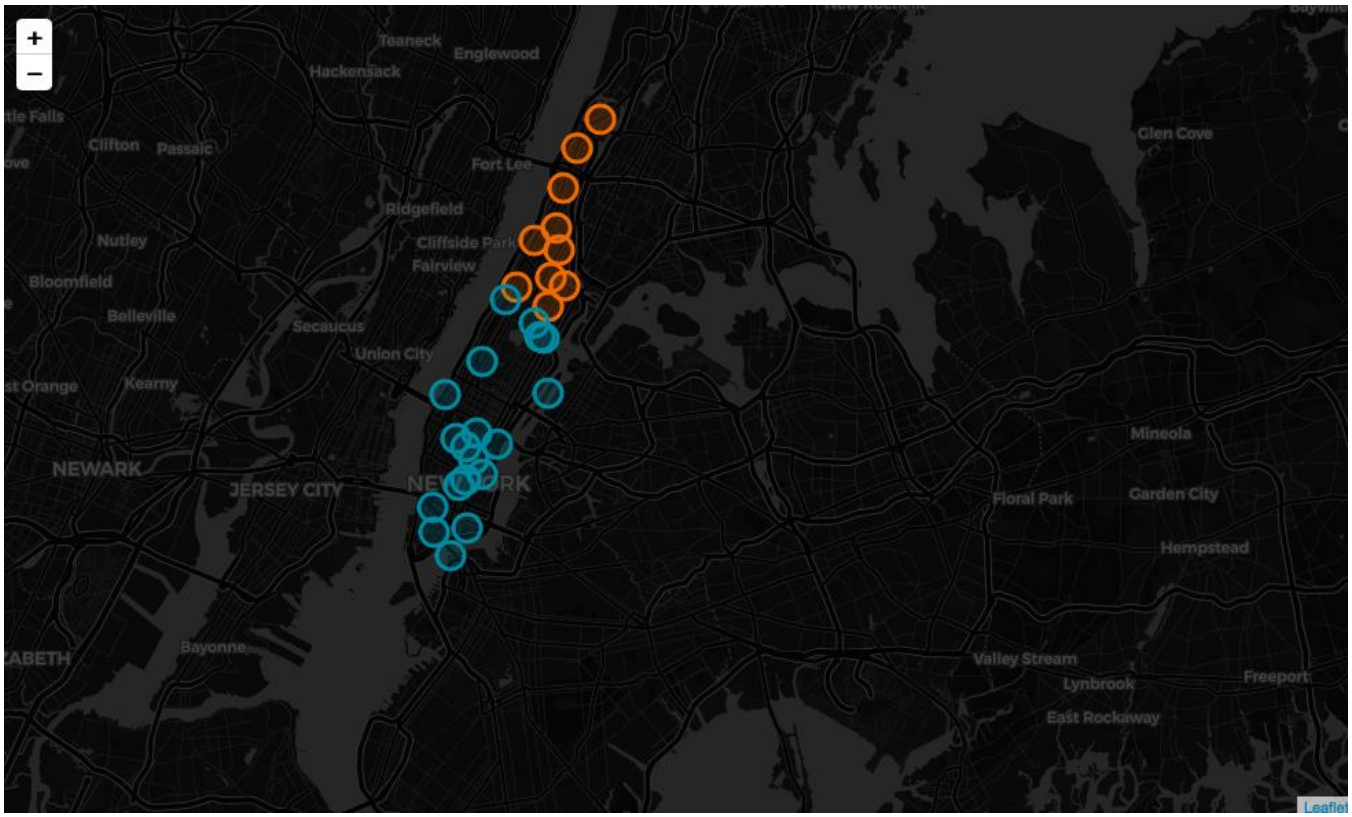


Staten Island: Our dataset had very less no of data points for pickups in Staten Island. Hence, the below graph might not be representative of the traffic in Staten Island. As per our dataset, Staten Island is the least busy borough for NYC Green Taxi. This might be due to the fact that this is the least populated borough of NYC and sometimes called "forgotten borough" by New Yorkers. The other reason might be the alternative transport option available in Staten Island, "The ferry" unlike other boroughs.





Since the traffic pattern in Manhattan appeared most interesting to me, I have created an interactive graph which will show neighborhoods having higher pickups than drop offs in tangerine and vice versa in teal over the NYC city map.



Below Fig. 10 shows the heat map for departure count in neighborhoods of Manhattan and Fig. 11 shows the heat map for arrival count in Manhattan.

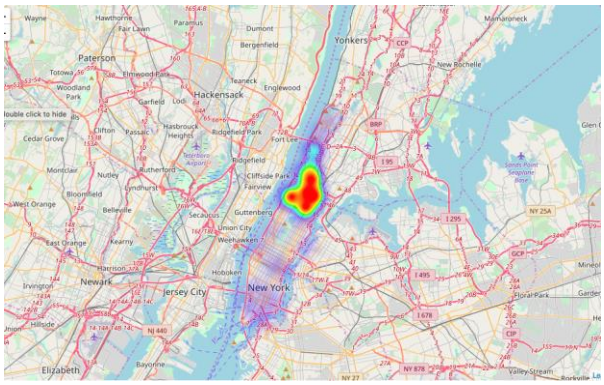


Fig. 10

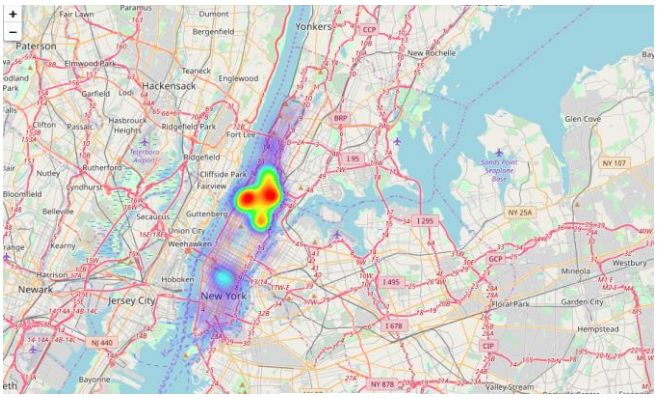


Fig. 11