FIT 5147

Data Exploration Project

NYC Taxi Trip Statistics

By - Pradnya Alchetti
Student ID - 29595916

Table Of Contents:

1. Introduction

Exploratory data analysis is the process of entire investigation of data. EDA provides a rough base for further data analysis. Various tools such as R, Tableau, Python can be used to perform these tasks.

This report focuses on performing EDA on New York Green Taxi Trip Data for the year 2016 combining it with the NYC Point Of Interest data and providing insights on the observed trends. The document will first illustrate the problem statement followed by data checking and preprocessing. It will then subsequently unravel the problem statement using different explorations and graphs and conclude with the insights found.

2. Problem Statement:

Taxi is always the first option preferred by any person when he/she wants to travel.
The taxi service provides flexibility, ease and comfort. May the travel be to office, site seeing, visiting a friend etc. the taxi service is always the best friend.

I am a hard core explorer when it comes to traveling, visiting new and interesting places, exploring nature and different lifestyle. When I travel, I see to it that I'll have all the native experiences of that place with the minimal travel cost. Therefore, I explore all the major services provided in a city before visiting that place. In a city like New York, I was amazed to know how efficiently the taxi service works in such a busy state.

In New York there are two kinds of Taxi Services known as Green Taxi and Yellow Taxi, the difference between them is that Yellow taxi are finite in number and provides street hail service while the Green taxi provides street hail and pre-arrangement service.

Fascinated by this service, I was curious to know which locations do people visit the most? Is the place one of the Point of Interest? When do people visit such Point of Interests? How the fare rates differ to reach that place? Do they differ with time or distance?

Motivated to find answers to such questions I chose NYC Green Taxi Trip Data and NYC Point Of Interest data to answer the following questions:

1. When taxi is street hailed the most visited Point Of Interest for the month of June
2. How number of visits to the same location as in Q1 affects the fare rates or the month of June
3. How do the fare rates differ with respect to time of the day

3. Data Wrangling:

Initially I was planning to use only NYC Green Taxi Trip Data for the entire year of 2016 but as the data had more than 16M records, I used the data only for the month of June. To increase the complexity of exploration I combined the data with the Point Of Interest in New York for the year 2016.

Following are the datasets used:

- NYC Green Taxi trip data for 2016, which has been reported by different LPEP providers

  Link: https://data.cityofnewyork.us/Transportation/2016-Green-Taxi-Trip-Data/hvrh-b6nb

  Tabular data(CSV format): 16.4M Rows x 23 Columns. It has both spatial and temporal attributes along with simple text and integer attributes.

  Data Dictionary: https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_green.pdf
- NYC Taxi zone shape files

  Link: https://s3.amazonaws.com/nyc-tlc/misc/taxi_zones.zip
- NYC Point of Interest data which is a compilation of what the different city agencies consider to be a Common Place or Place/Point of Interest.

  Link: https://data.cityofnewyork.us/City-Government/Points-Of-Interest/rxuy-2muj

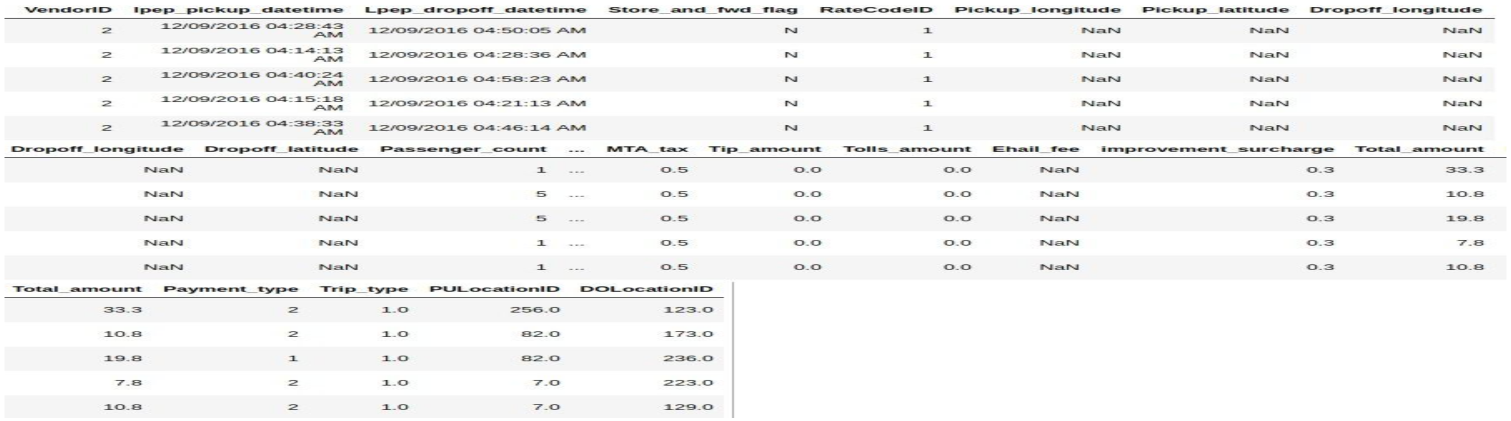
  Tabular data(CSV format): 19898 Rows x 16 columns. It has both spatial and temporal attributes along with simple text and integer attributes.

  Data dictionary:

https://data.cityofnewyork.us/api/views/rxuy-2muj/files/ebabcf1d-c6e7-43ca-a031-5168036b2fbb?download=true&filename=PointOfInterest.pdf

- Data Cleaning:

The above csv files were read into two separate dataframe using Python. The given data was formatted using coding methods in Python.
The dataframe after reading the NYC Green Taxi data with 16M records is as follows:

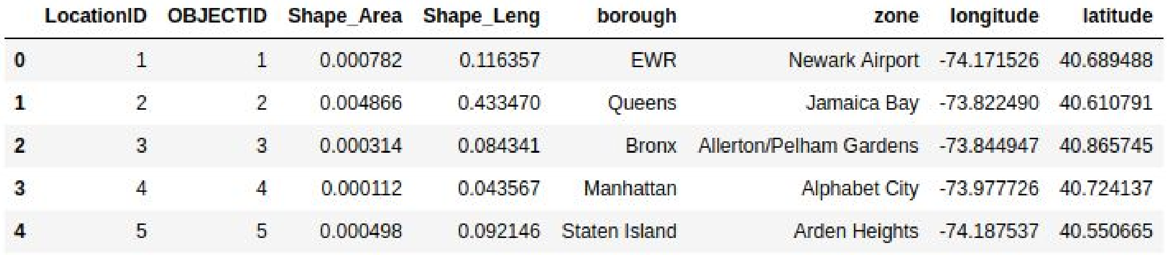


VendorID	lpep_pickup_datetime	Lpep_dropoff_datetime	Store_and_fwd_flag	RateCodeID	Pickup_longitude	Pickup_latitude	Dropoff_longitude
2	12/09/2016 04:28:43 AM	12/09/2016 04:50:05 AM	N	1	NaN	NaN	NaN
2	12/09/2016 04:14:13 AM	12/09/2016 04:28:36 AM	N	1	NaN	NaN	NaN
2	12/09/2016 04:40:24 AM	12/09/2016 04:58:23 AM	N	1	NaN	NaN	NaN
2	12/09/2016 04:15:18 AM	12/09/2016 04:21:13 AM	N	1	NaN	NaN	NaN
2	12/09/2016 04:38:33 AM	12/09/2016 04:46:14 AM	N	1	NaN	NaN	NaN

Dropoff_longitude	Dropoff_latitude	Passenger_count	...	MTA_tax	Tip_amount	Tolls_amount	Ehail_fee	improvement_surcharge	Total_amount
NaN	NaN	1	...	0.5	0.0	0.0	NaN	0.3	33.3
NaN	NaN	5	...	0.5	0.0	0.0	NaN	0.3	10.8
NaN	NaN	5	...	0.5	0.0	0.0	NaN	0.3	19.8
NaN	NaN	1	...	0.5	0.0	0.0	NaN	0.3	7.8
NaN	NaN	1	...	0.5	0.0	0.0	NaN	0.3	10.8

Total_amount	Payment_type	Trip_type	PULocationID	DOLocationID
33.3	2	1.0	256.0	123.0
10.8	2	1.0	82.0	173.0
19.8	1	1.0	82.0	236.0
7.8	2	1.0	7.0	223.0
10.8	2	1.0	7.0	129.0

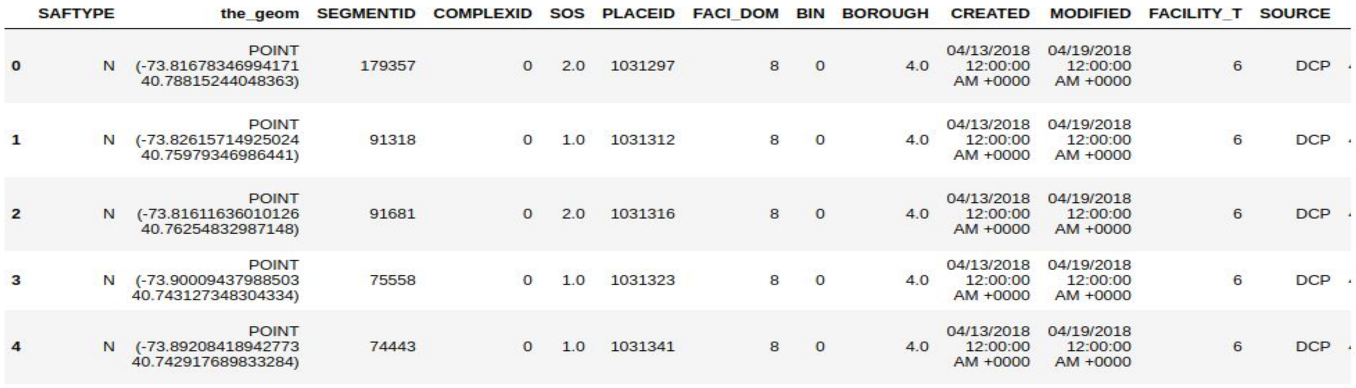
From the above dataframe we can see that most of the Pickup_longitude, Pickup_latitude, Dropoff_longitude, Dropoff_latitude are missing which are the important fields for further analysis. In this case the Pickup Location ID and DropOff Location ID is given.
These location IDs are used to lookup in the **taxi_zone.shp** file which provides the latitude and longitude corresponding to the location id.
The taxi_zone.shp was then read using shape file libraries in python and extracted all the fields as below:



	LocationID	OBJECTID	Shape_Area	Shape_Leng	borough	zone	longitude	latitude
0	1	1	0.000782	0.116357	EWR	Newark Airport	-74.171526	40.689488
1	2	2	0.004866	0.433470	Queens	Jamaica Bay	-73.822490	40.610791
2	3	3	0.000314	0.084341	Bronx	Allerton/Pelham Gardens	-73.844947	40.865745
3	4	4	0.000112	0.043567	Manhattan	Alphabet City	-73.977726	40.724137
4	5	5	0.000498	0.092146	Staten Island	Arden Heights	-74.187537	40.550665

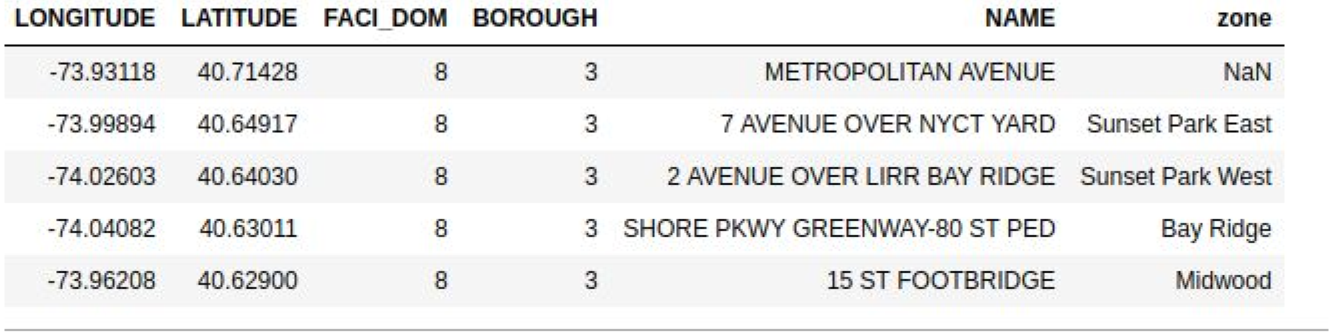
The missing fields for latitude and longitude in green taxi data were then filled up using taxi shape file.
The precision of the latitude and longitude was too long upto 16 decimals which was then rounded up to 5 decimals.

The original POI dataframe was as follows:



	SAFTYPE	the_geom	SEGMENTID	COMPLEXID	SOS	PLACEID	FACI_DOM	BIN	BOROUGH	CREATED	MODIFIED	FACILITY_T	SOURCE
0	N	POINT (-73.81678346994171 40.78815244048363)	179357	0	2.0	1031297	8	0	4.0	04/13/2018 12:00:00 AM +0000	04/19/2018 12:00:00 AM +0000	6	DCP
1	N	POINT (-73.82615714925024 40.75979346986441)	91318	0	1.0	1031312	8	0	4.0	04/13/2018 12:00:00 AM +0000	04/19/2018 12:00:00 AM +0000	6	DCP
2	N	POINT (-73.81611636010126 40.76254832987148)	91681	0	2.0	1031316	8	0	4.0	04/13/2018 12:00:00 AM +0000	04/19/2018 12:00:00 AM +0000	6	DCP
3	N	POINT (-73.90009437988503 40.743127348304334)	75558	0	1.0	1031323	8	0	4.0	04/13/2018 12:00:00 AM +0000	04/19/2018 12:00:00 AM +0000	6	DCP
4	N	POINT (-73.89208418942773 40.742917689833284)	74443	0	1.0	1031341	8	0	4.0	04/13/2018 12:00:00 AM +0000	04/19/2018 12:00:00 AM +0000	6	DCP

The above dataframe was reformatted to convert the geom point to proper latitude and longitude and precision of the latitude and longitude was set to 5 decimals.



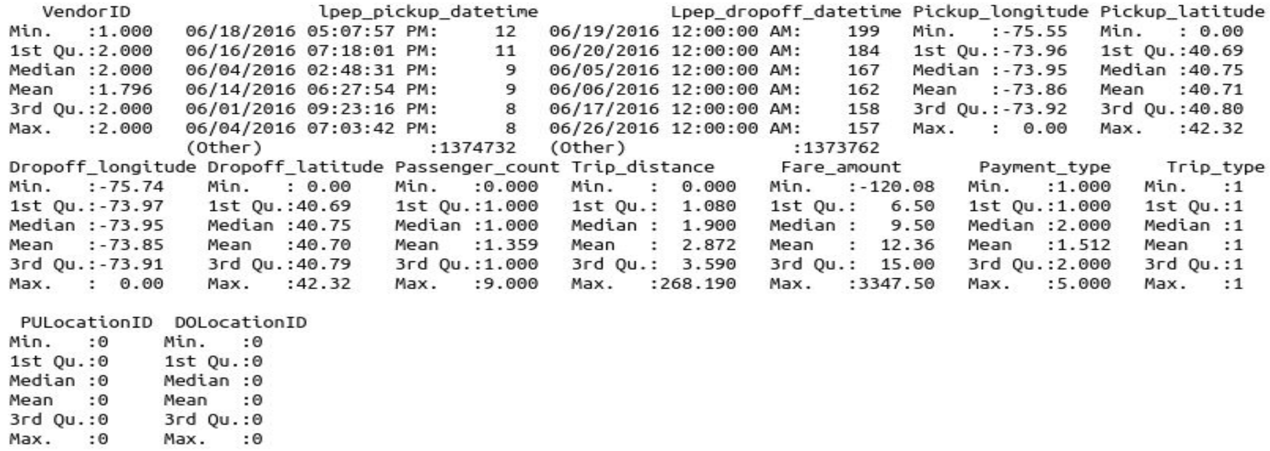
LONGITUDE	LATITUDE	FACI_DOM	BOROUGH	NAME	zone
-73.93118	40.71428	8	3	METROPOLITAN AVENUE	NaN
-73.99894	40.64917	8	3	7 AVENUE OVER NYCT YARD	Sunset Park East
-74.02603	40.64030	8	3	2 AVENUE OVER LIRR BAY RIDGE	Sunset Park West
-74.04082	40.63011	8	3	SHORE PKWY GREENWAY-80 ST PED	Bay Ridge
-73.96208	40.62900	8	3	15 ST FOOTBRIDGE	Midwood

Only the columns such as latitude, longitude, borough and name were retrieved. The zone name was retrieved using the taxi zone shape file.

In order to answer the questions, instead of using the entire Green Taxi data of 16M records I segregated the data based on TripType and extracted the trip data only for TripType=1 that is Street Hail. This data was further divided to one month data that is the month of June and columns such as MTA Tax, Tolls amount, surcharge were dropped which rendered me with more than 1M records.
Hence, for all the further data analysis I am using **Street Hail trip data** for the month of **June** 2016.

4. Data Checking:

The green taxi data was further checked for errors by analysing the summary statistics.



```
    VendorID                lpep_pickup_datetime             Lpep_dropoff_datetime Pickup_longitude Pickup_latitude
 Min.   :1.000   06/18/2016 05:07:57 PM:      12   06/19/2016 12:00:00 AM:     199   Min.   :-75.55   Min.   : 0.00
 1st Qu.:2.000   06/16/2016 07:18:01 PM:      11   06/20/2016 12:00:00 AM:     184   1st Qu.:-73.96   1st Qu.:40.69
 Median :2.000   06/04/2016 02:48:31 PM:       9   06/05/2016 12:00:00 AM:     167   Median :-73.95   Median :40.75
 Mean   :1.796   06/14/2016 06:27:54 PM:       9   06/06/2016 12:00:00 AM:     162   Mean   :-73.86   Mean   :40.71
 3rd Qu.:2.000   06/01/2016 09:23:16 PM:       8   06/17/2016 12:00:00 AM:     158   3rd Qu.:-73.92   3rd Qu.:40.80
 Max.   :2.000   06/04/2016 07:03:42 PM:       8   06/26/2016 12:00:00 AM:     157   Max.   :  0.00   Max.   :42.32
                 (Other)               :1374732   (Other)               :1373762
 Dropoff_longitude Dropoff_latitude Passenger_count Trip_distance      Fare_amount        Payment_type     Trip_type
 Min.   :-75.74    Min.   : 0.00    Min.   :0.000   Min.   :  0.000   Min.   :-120.08   Min.   :1.000   Min.   :1
 1st Qu.:-73.97    1st Qu.:40.69    1st Qu.:1.000   1st Qu.:  1.080   1st Qu.:   6.50   1st Qu.:1.000   1st Qu.:1
 Median :-73.95    Median :40.75    Median :1.000   Median :  1.900   Median :   9.50   Median :2.000   Median :1
 Mean   :-73.85    Mean   :40.70    Mean   :1.359   Mean   :  2.872   Mean   :  12.36   Mean   :1.512   Mean   :1
 3rd Qu.:-73.91    3rd Qu.:40.79    3rd Qu.:1.000   3rd Qu.:  3.590   3rd Qu.:  15.00   3rd Qu.:2.000   3rd Qu.:1
 Max.   :  0.00    Max.   :42.32    Max.   :9.000   Max.   :268.190   Max.   :3347.50   Max.   :5.000   Max.   :1

  PULocationID  DOLocationID
 Min.   :0     Min.   :0
 1st Qu.:0     1st Qu.:0
 Median :0     Median :0
 Mean   :0     Mean   :0
 3rd Qu.:0     3rd Qu.:0
 Max.   :0     Max.   :0
```

The above table provides the summary statistics of all columns in taxi data.
From the above table we can observe the following errors:

- Latitude Longitude errors

Observation: The minimum and maximum Pickup longitude are -75.55 and 0.00 respectively. Similarly the minimum and maximum Pickup latitude are 0.00 and 42.32
Subsequently a similar case is also observed for Dropoff_longitude and Dropoff_latitude.
Correction: By using the data dictionary provided for Green Taxi data, the range of latitude and longitude for New York were determined.

Min Latitude: 40.45326	Max Latitude: 40.94788
Min Longitude: -74.2748	Max Longitude: -73.70741

Those records whose Pickup and Dropoff coordinates fall in this range are considered for further analysis.

- Trip distance is zero

Observation: From the above summary it is observed that the minimum distance is 0.00 miles.
Correction: Using the Haversine distance method we calculated the trip distance in miles between pickup coordinates and dropoff coordinates. There were some cases where the pickup and dropoff coordinates were same. This indicates that the trip distance is metered as 0 miles but the customer might have taken a round trip.

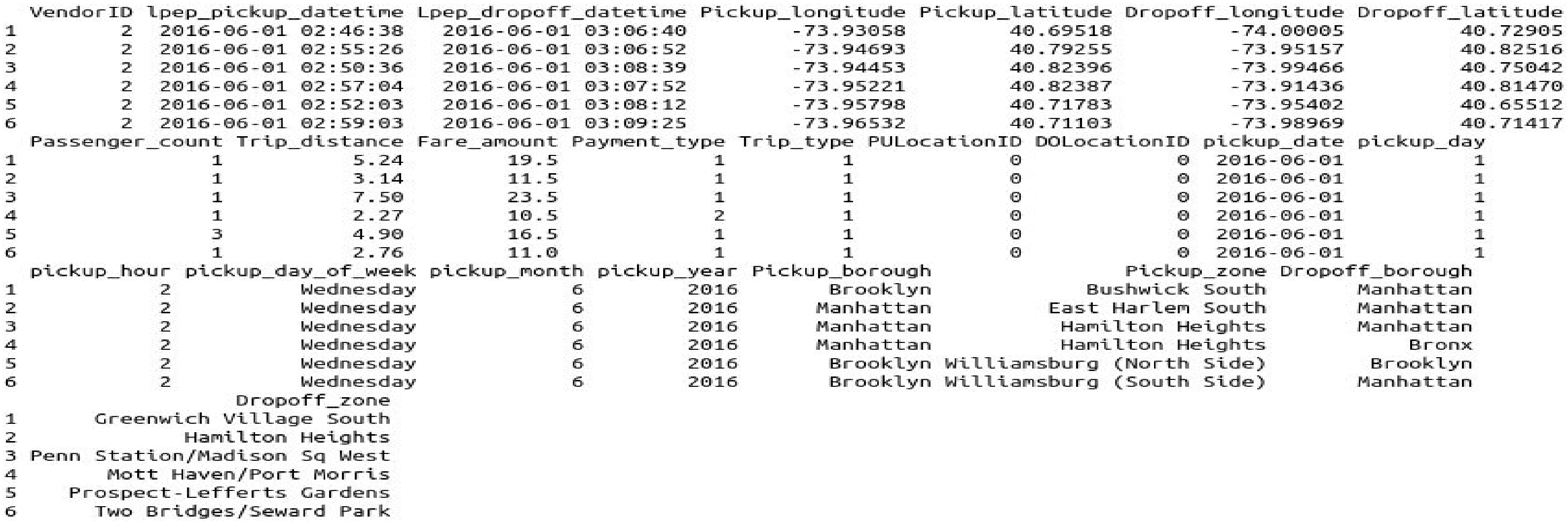
- Fare Amount

Observation: In the above table the minimum fare amount observed -120.08. It is known that the fare amount can never be negative.
Correction: Retrieved all the records with negative fare amount and retrieved all the similar records from the data with non negative fare by comparing the latitude, longitude and trip distance. The negative fare amount was then replaced with the mean of fare amount of all similar non-negative records.

Feature Engineering: The pickup and dropoff date and time were used to determine the following:

- pickup_date, pickup_day, pickup_hour, pickup_day_of_week, pickup_month, pickup_year
- dropoff_date, dropoff_day, dropoff_hour, dropoff_day_of_week, dropoff_month, dropoff_year
- pickup_borough and pickup_zone, dropOff borough and dropOff zone were determined using the taxi zone shape file.

Green Taxi trip data used for exploration is as follows:



```
  VendorID lpep_pickup_datetime Lpep_dropoff_datetime Pickup_longitude Pickup_latitude Dropoff_longitude Dropoff_latitude
1        2  2016-06-01 02:46:38   2016-06-01 03:06:40        -73.93058        40.69518         -74.00005         40.72905
2        2  2016-06-01 02:55:26   2016-06-01 03:06:52        -73.94693        40.79255         -73.95157         40.82516
3        2  2016-06-01 02:50:36   2016-06-01 03:08:39        -73.94453        40.82396         -73.99466         40.75042
4        2  2016-06-01 02:57:04   2016-06-01 03:07:52        -73.95221        40.82387         -73.91436         40.81470
5        2  2016-06-01 02:52:03   2016-06-01 03:08:12        -73.95798        40.71783         -73.95402         40.65512
6        2  2016-06-01 02:59:03   2016-06-01 03:09:25        -73.96532        40.71103         -73.98969         40.71417
  Passenger_count Trip_distance Fare_amount Payment_type Trip_type PULocationID DOLocationID pickup_date pickup_day
1               1          5.24        19.5            1         1            0            0  2016-06-01          1
2               1          3.14        11.5            1         1            0            0  2016-06-01          1
3               1          7.50        23.5            1         1            0            0  2016-06-01          1
4               1          2.27        10.5            2         1            0            0  2016-06-01          1
5               3          4.90        16.5            1         1            0            0  2016-06-01          1
6               1          2.76        11.0            1         1            0            0  2016-06-01          1
  pickup_hour pickup_day_of_week pickup_month pickup_year Pickup_borough              Pickup_zone Dropoff_borough
1           2          Wednesday            6        2016       Brooklyn           Bushwick South       Manhattan
2           2          Wednesday            6        2016      Manhattan         East Harlem South       Manhattan
3           2          Wednesday            6        2016      Manhattan         Hamilton Heights       Manhattan
4           2          Wednesday            6        2016      Manhattan         Hamilton Heights           Bronx
5           2          Wednesday            6        2016       Brooklyn Williamsburg (North Side)        Brooklyn
6           2          Wednesday            6        2016       Brooklyn Williamsburg (South Side)       Manhattan
                   Dropoff_zone
1      Greenwich Village South
2             Hamilton Heights
3 Penn Station/Madison Sq West
4         Mott Haven/Port Morris
5     Prospect-Lefferts Gardens
6        Two Bridges/Seward Park
```

5. Data Exploration:

After fixing all the errors and removing the outliers from the data, it was used for further analysis.

The exploration was performed in R

Methodology:

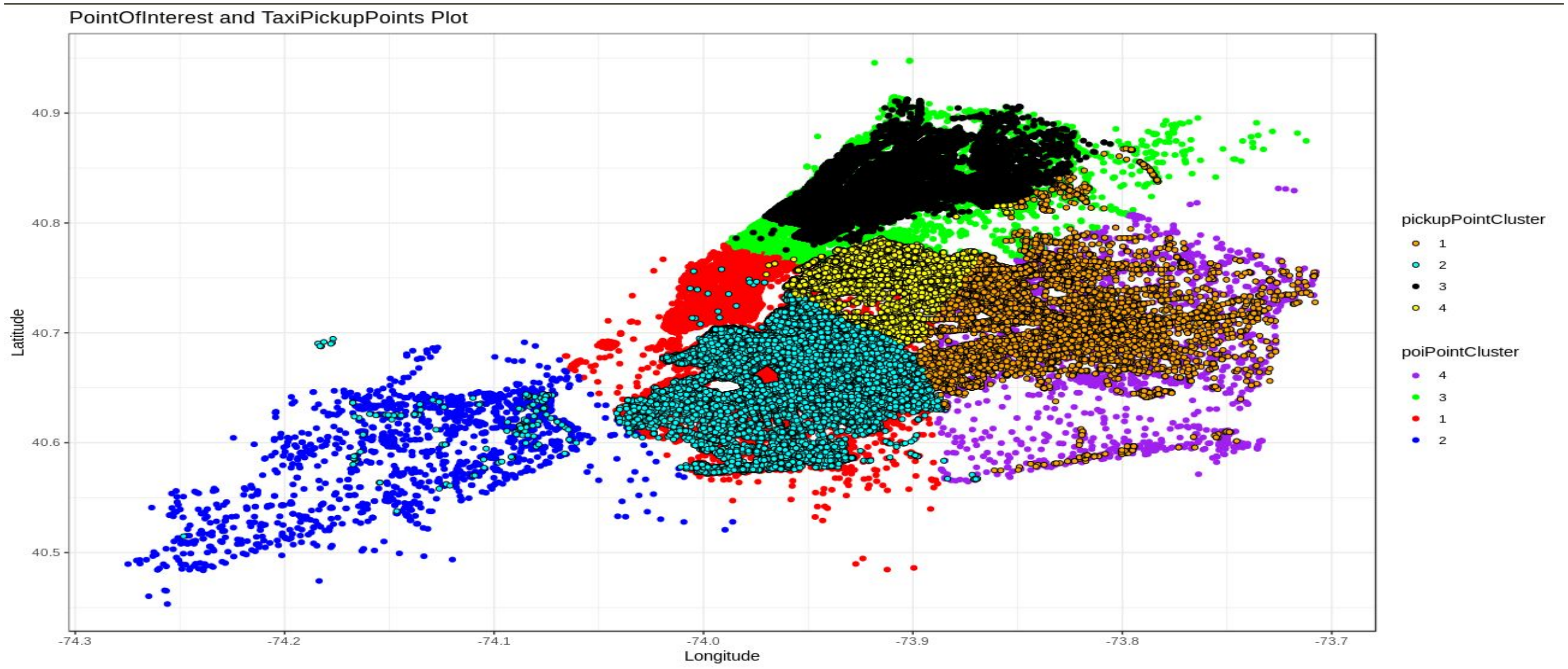
As the data consists of many numeric features it is tempting to directly find correlation between different attributes by plotting random visualisations. This would be a tedious task. In order to avoid this, I have divided my exploration into the following structure:

- Combine two datasets and find correlation between them
- Determine the most dense area and extract that data for further analysis.
- Determine relation between different attributes from the extracted data
- Accordingly with each above sub tasks I will try to answer my main questions.
- Conclude by aggregating the results and explaining the insights found

Combining two datasets and extracting data based on density:
The POI data consists of point coordinates while the Taxi Trip data consists of Pickup and Dropoff coordinates. It is impossible to directly merge the data on latitude and longitude as there may be a case where the pickup or dropoff coordinates are in the vicinity of the Point Of Interest.

In order to address this situation we have formed separate clusters of the both datasets and overlaid them on top of one another to find the correlation between them

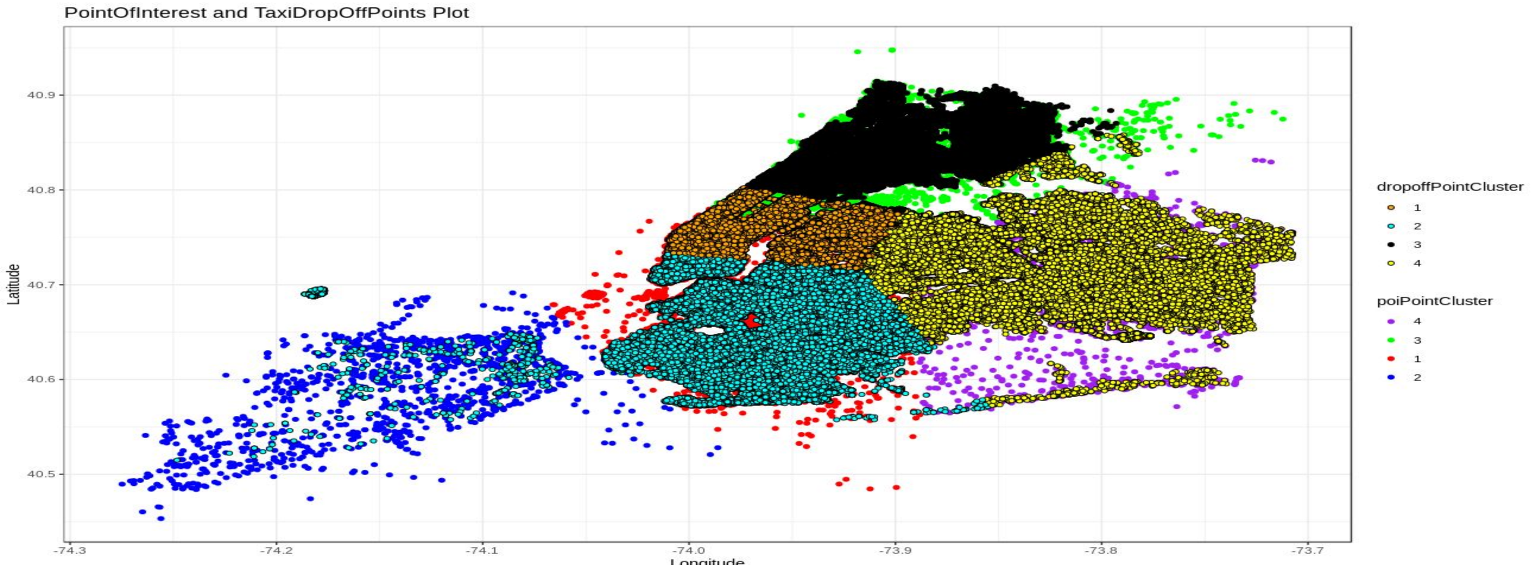
The kmeans clustering divides the data into n equal clusters where each data element belongs to the cluster with nearest means.
We have divided the POI dataset into 4 clusters and the pickups locations in the Taxi Trip dataset into 4 clusters respectively.



In the above plot for clusters of POI and Pickups Points, the X-axis represents the longitude while the Y-axis represents the Latitude. The POI points are represented by Red(Cluster 1), Green(Cluster 3), Blue(Cluster 2), Purple(Cluster 4) clusters. The Pickup points are represented by Black(Cluster 3), Cyan(Cluster 2),Yellow(Cluster 4),Orange(Cluster 1).
The clustering also helps us to determine how is the data distributed across the region.
From the above plot we clearly see that the Pickup points are distributed across five boroughs of New York. It is observed from the plot that Pickups and POI overlaps are dense in the Cluster 1, Cluster 2 and Cluster 4 of Pickup points which also indicates that the POI in these clusters are visited the most.

The Pickup points cluster 1, 2 and 4 helped me to determine the max and min range of latitude and longitude where there are most pickups observed.

In order to observe the highest DropOff points with respect to POI points we have divided the POI locations into 4 clusters and DropOff points into 4 clusters respectively.

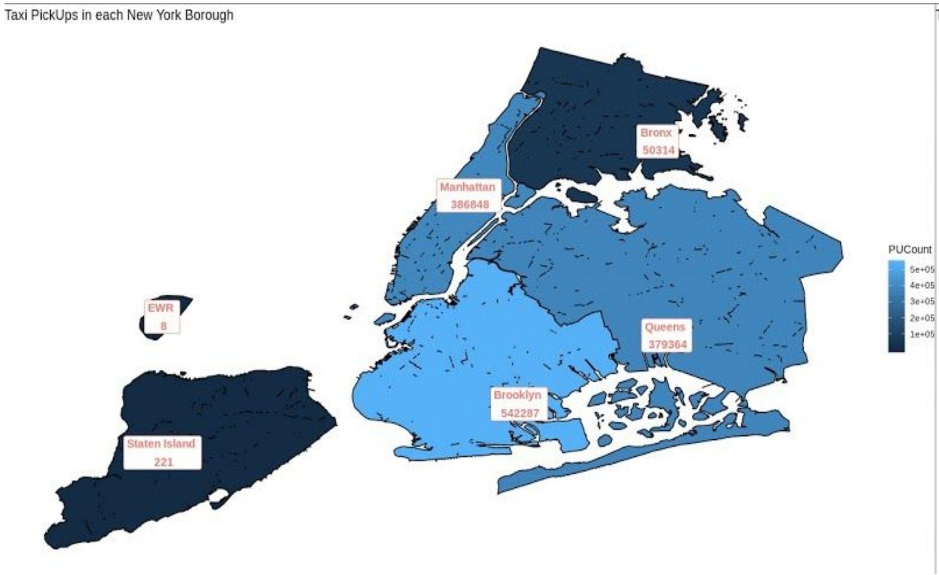


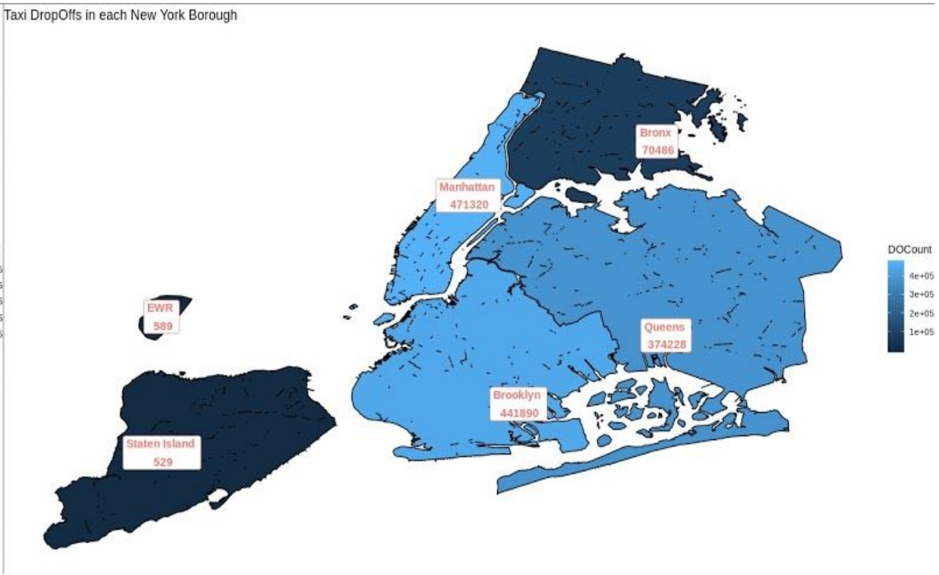
In the above graph it is observed that the POI (Cluster 1) and DropOff Clusters (Cluster 1 and Cluster 2) overlap densely which indicates that as most of the DropOffs are in this region, POIs of that region may be visited the most.

The DropOff points cluster 1 and 2 helped me to determine the max and min range of latitude and longitude where there are most dropoffs observed.

From the cleaned actual Green Trip Dataset I extracted only those records where the pickup latitude and longitude lie in the range observed in the pickup cluster and dropoff latitude and longitude lie in the range observed in the dropoff cluster.
Hence, this provides me with the data where there most pickups and dropOffs were observed.
But our goal is to find the most visited POI in these regions.

Determining the boroughs with most Pickups and DropOffs
New York City is divided into 5 boroughs viz. Manhattan, Brooklyn, Queens, Bronx, Staten Island.
With the above filtered data through clustering we determine the boroughs with the highest Pickups and dropOffs



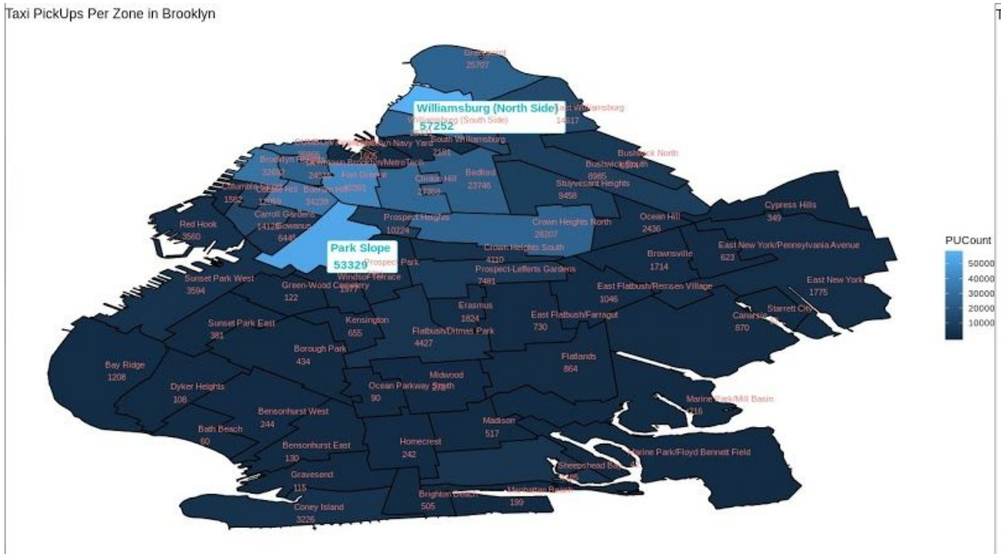


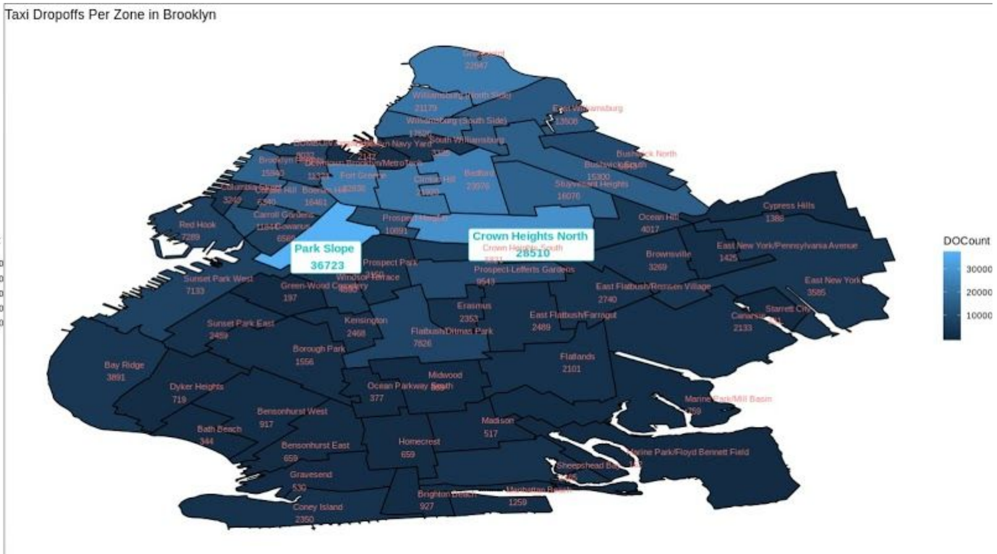
The left map determines different boroughs in New York City and the number of Pickups in each borough in June 2016. The intensity of the color decreases as the number of Pickups in that borough increases. It is observed that **Brooklyn** has the **highest number of Pickups** with 542287 Pickups.

The right map determines the number of DropOffs with respect to each borough. It can be observed that **Manhattan** has the **highest number of DropOffs** with 471320 DropOffs.

As the highest number of **Pickups** are in **Brooklyn** and the highest number of **DropOffs** are in **Manhattan**, we can say that the POI in these boroughs would be visited the most.

Determining the zones with most Pickups and DropOffs Brooklyn
As we determined the highest pickups were from Brooklyn, now we are trying to determine which zone exactly had the highest pickups and dropOffs.



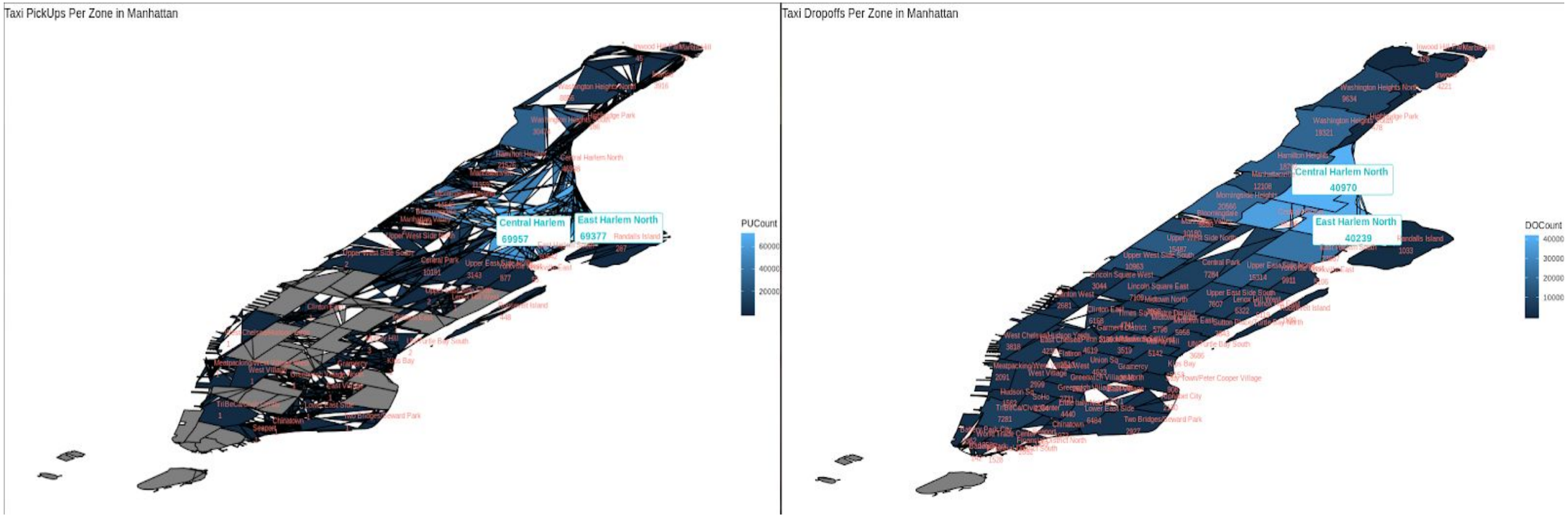


The left map represents the number of Pickups in each zone of Brooklyn. It is observed that most pickups are observed in **Park Slope** (53329) followed by Williamsburg North Side.
The right map represents the number of DropOffs in each zone of Brooklyn. It is observed that most pickups are observed in **Park Slope** (36723).

From the above observations we can say that **Park Slope** is the most visited zone in **Brooklyn**.

Determining the zones with most Pickups and DropOffs Manhattan

As we determined the highest dropOffs were from Manhattan, now we are trying to determine which zone exactly had the highest pickups and dropOffs.

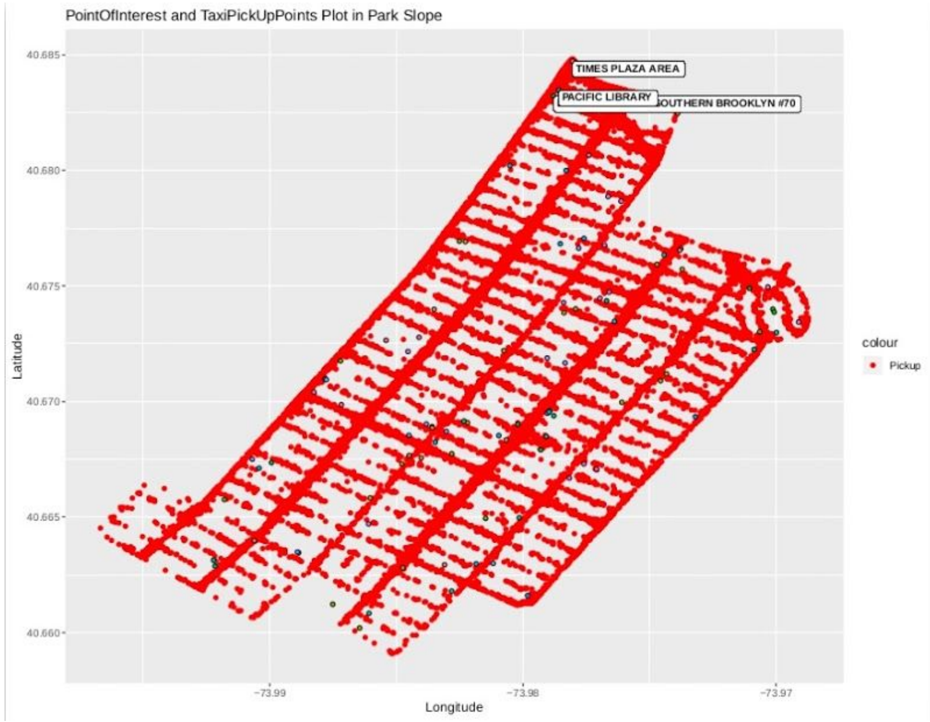


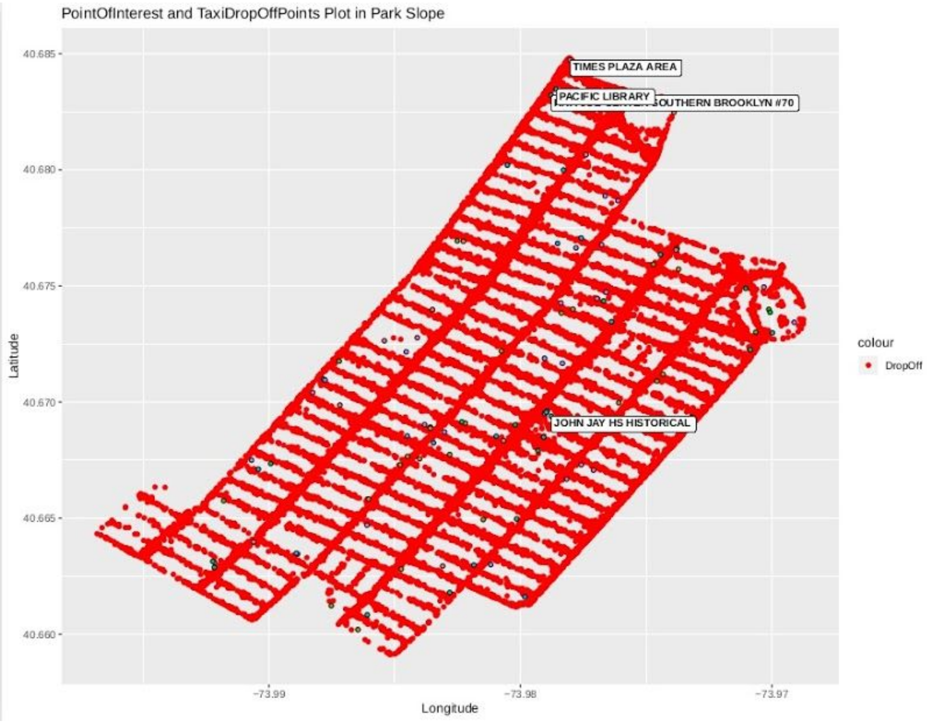
The left map determines the number of Pickups in each zone of Manhattan while the right map determines the number of DropOffs in each zone of Manhattan. It can be observed that the most **Pickups** are from **East Harlem North** and the most **Dropoffs** are to **Central Harlem South** followed by **East Harlem North**.

From the above observations we can say that **East Harlem North** is the most visited zone in **Manhattan**.

Determining which POI were visited the most in Park Slope,Brooklyn:

As we determined most Pickups and DropOffs in Brooklyn are in Park Slope. We now extract the POI in borough Brooklyn and zone Park Slope. We then observe which POI locations were visited the most in Park Slope.





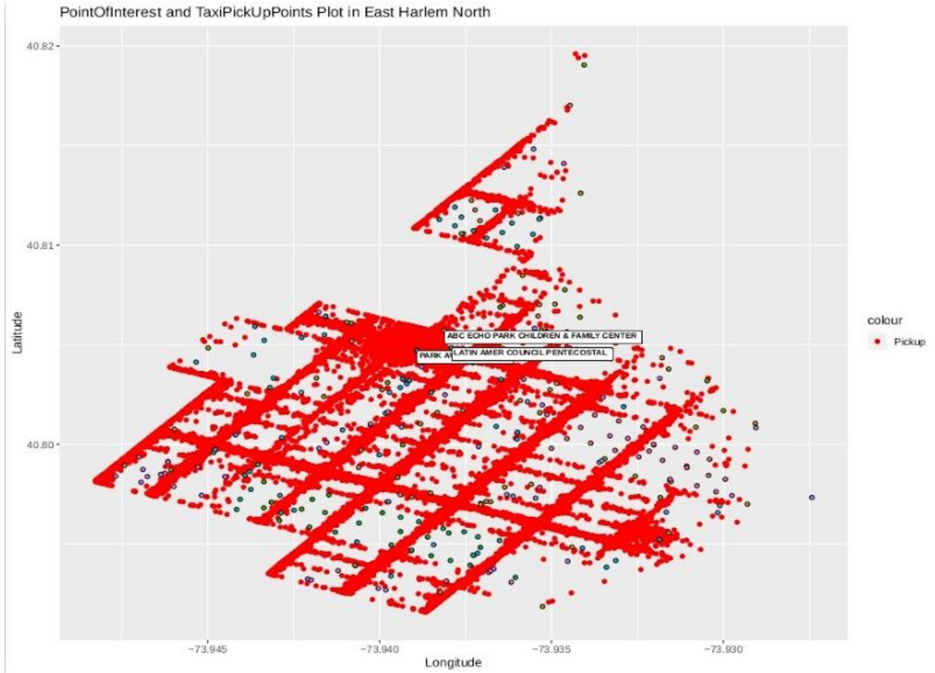
The above plots are scatter plot where in the left plot X-axis determines the Longitude and Y-axis determines the Latitude. The points in red determine the Pickup points and the coloured points determine the POIs in Park slope. It is observed that the pickups are dense near **Times Plaza Area, Pacific Library and HRA Job Center Southern Brooklyn**.

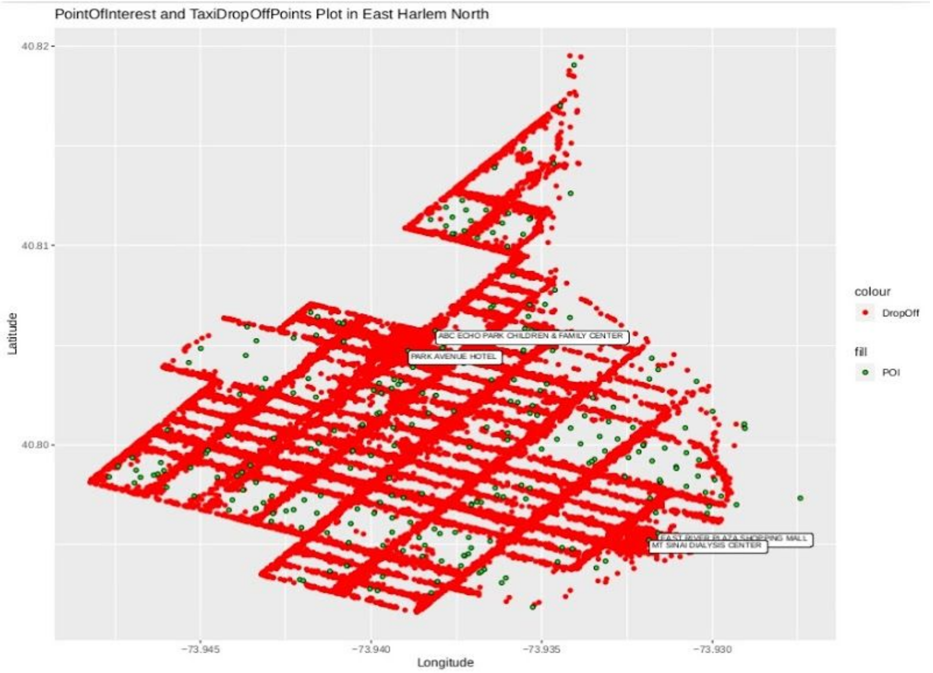
In the right plot X-axis determines the Longitude and Y-axis determines the Latitude. The points in red determine the DropOff points and the coloured points determine the POIs in Park slope. It is observed that the dropOffs are dense near **Times Plaza Area, Pacific Library, HRA Job Center Southern Brooklyn and John Jay HS Historical**.

Thus, we can say that the most visited **Point Of interests** are in **Park Slope, Brooklyn** viz.
Times Plaza Area, Pacific Library, HRA Job Center Southern Brooklyn and John Jay HS Historical

Determining which POI were visited the most in East Harlem North, Manhattan:

As we determined most Pickups and DropOffs in Manhattan are in East Harlem North. We extract POI for the borough Manhattan and zone East Harlem North. We now observe which POI locations were visited the most in East Harlem North.





In the left plot X-axis determines the Longitude and Y-axis determines the Latitude. The points in red determine the Pickup points and the coloured points determine the POIs in East Harlem North. It is observed that the pickups are dense near **ABC Echo Park Children & Family Center, Park Avenue Hotel, Latin Amer Council Pentecostal**

In the right plot X-axis determines the Longitude and Y-axis determines the Latitude. The points in red determine the DropOff points and the coloured points determine the POIs in East Harlem North. It is observed that the dropOffs are dense near **ABC Echo Park Children & Family Center, Park Avenue Hotel**

Thus, we can say that the most visited **Point Of interests** in **East Harlem North, Manhattan** are viz.
ABC Echo Park Children & Family Center, Park Avenue Hotel, Latin Amer Council Pentecostal

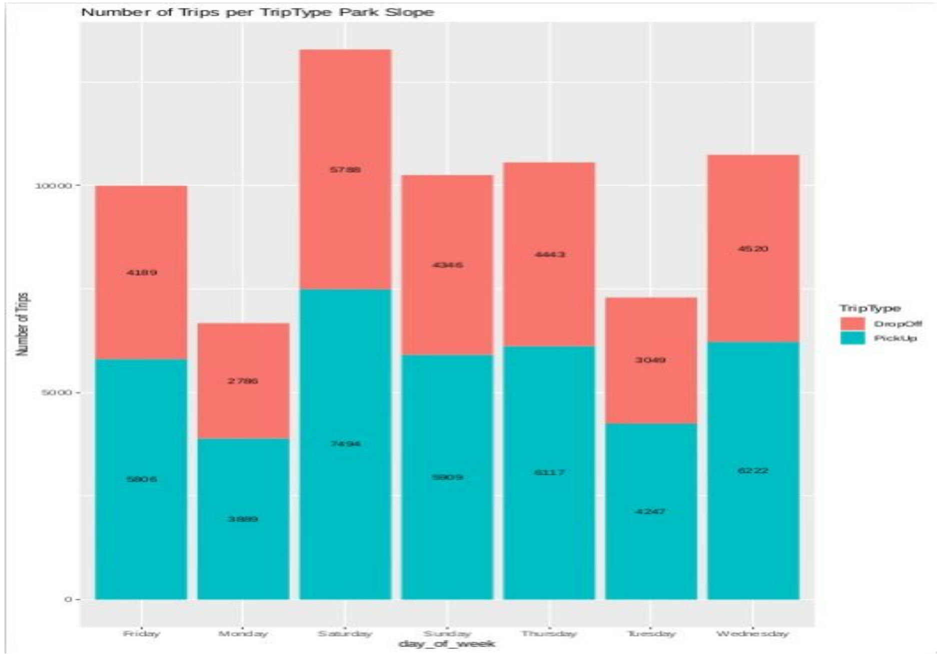
As we determined the most visited Point Of Interests, Times Plaza Area, Pacific Library, HRA Job Center Southern Brooklyn and John Jay HS Historical lie in Park Slope, Brooklyn while ABC Echo Park Children & Family Center, Park Avenue Hotel, Latin Amer Council Pentecostal lie in East Harlem North, Manhattan.
We now extract all the pickups and dropoffs trip data where the distance of these POI from the pickup or dropoff point is less than 0.4 miles. We make use of the Haversine distance formula to find the distance between two coordinates.

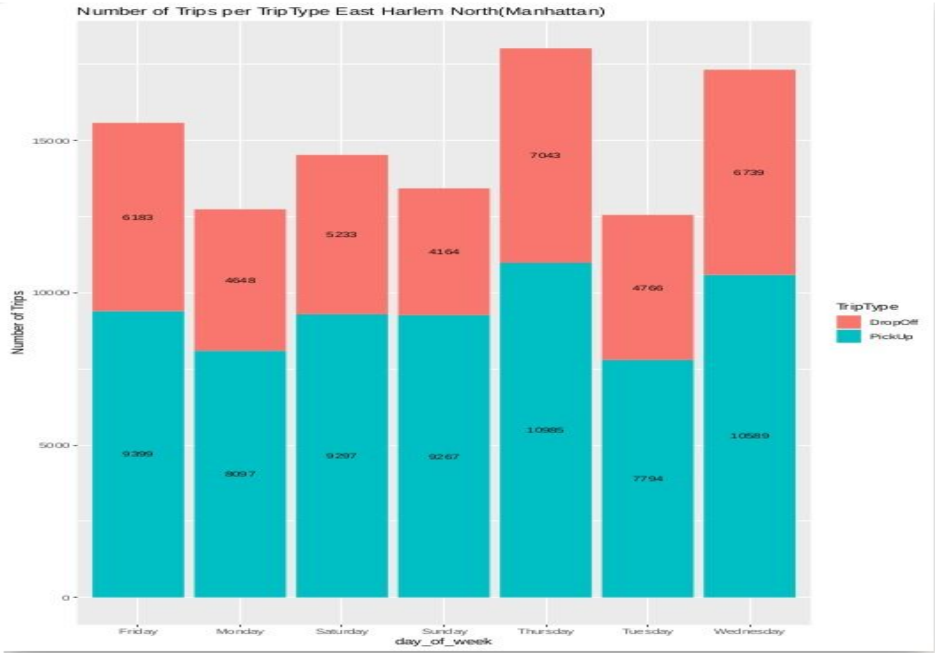
We assume that all the trips with the pickups or dropOff coordinates within 0.4 miles of most visited POI in Park Slope or East Harlem North had come to visit one of the Point of Interest.

On the basis of this assumption, we determine the best time to visit such places and also the fare rates distribution in these zones.

Analysis on trips near POI for Park Slope and East Harlem

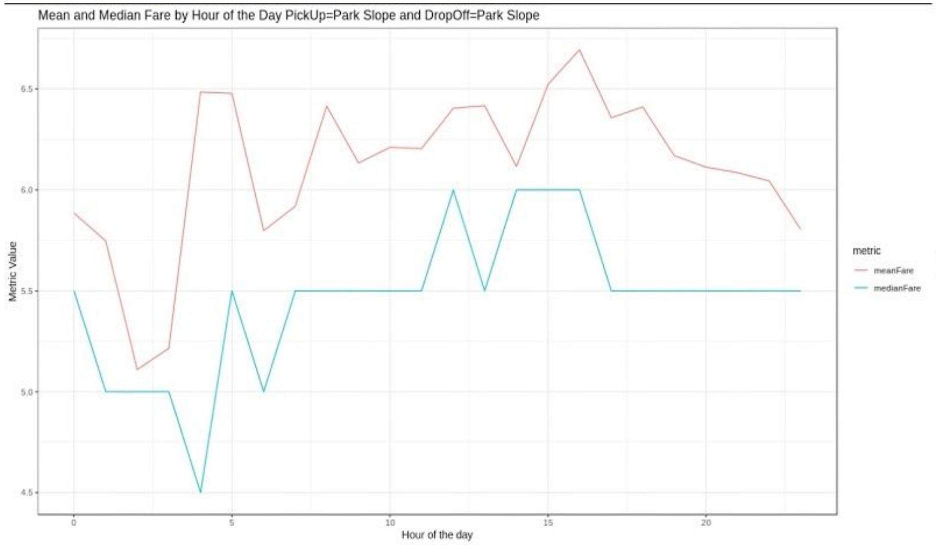
- **Analysis on the number of trips with respect to Time of the week**

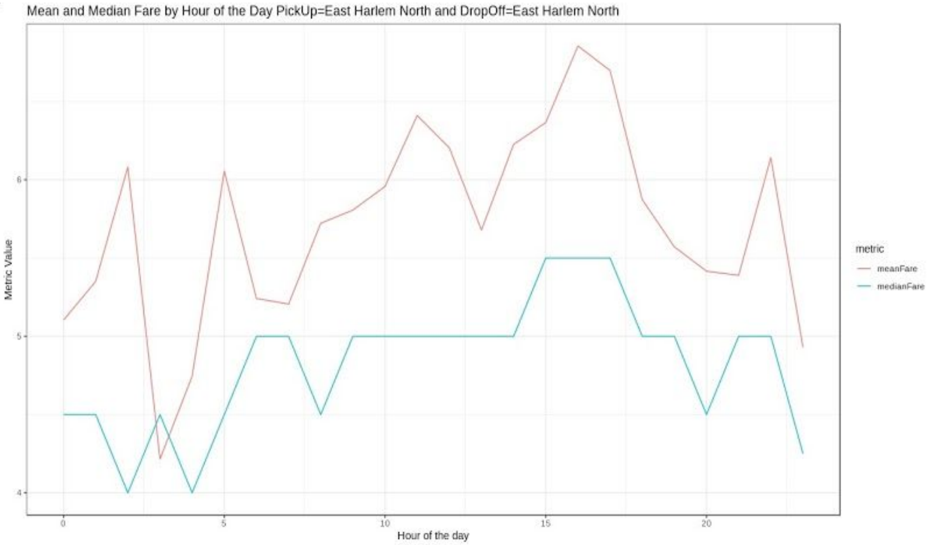




In the above plots the X-axis represents the day of the week while the Y axis represents total number of trips. The left plot represents the number of pickups and dropoffs each day in Park Slope while the right plot represents the number of pickups and dropoffs each day in East Harlem North. It is observed that in Park Slope most the trips are on Saturday which is a weekend. This indicates that people often visit the POI on weekends. On the other side in East Harlem North most Pickups and DropOffs are observed on Thursday which indicates that as the most visited POI in East Harlem North come under commercial category people tend to travel on weekdays.

- **Analysis on Fare rates with respect to time of the day East Harlem North**



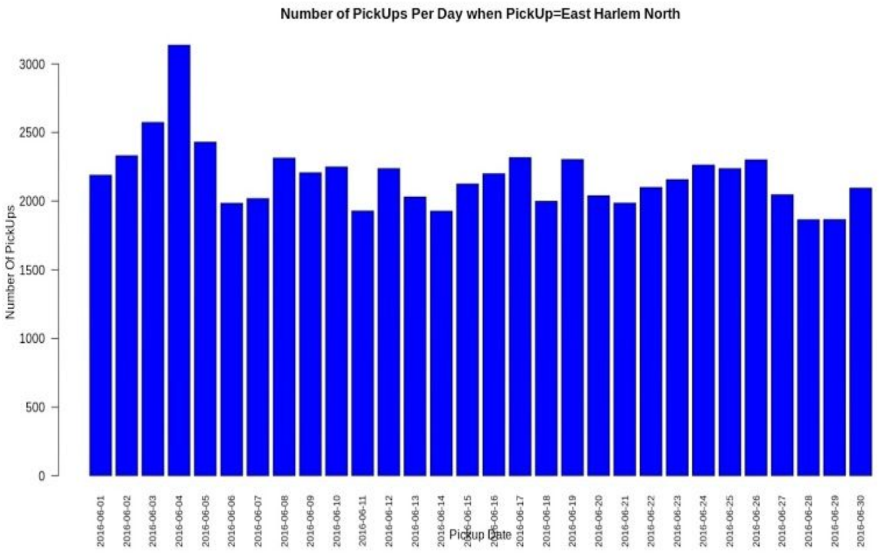


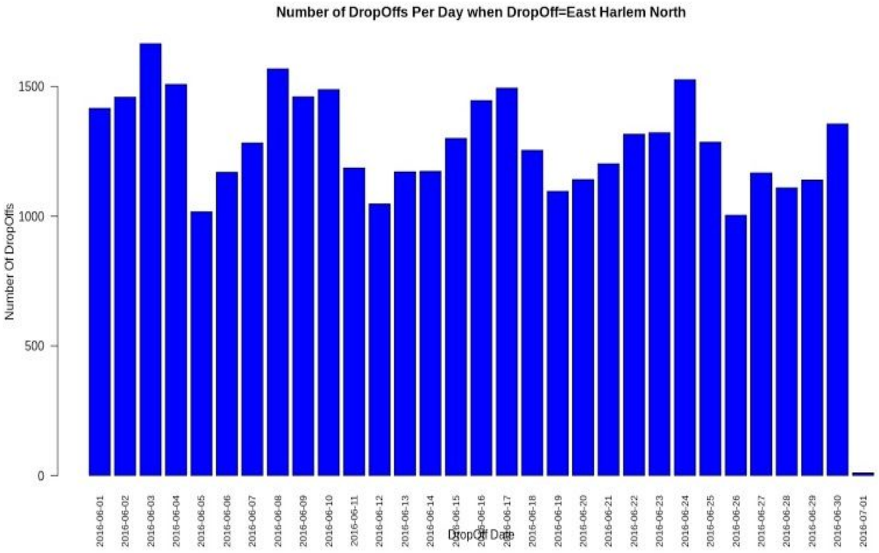
In the above plots the X- axis represents the Hour of the Day and the Y-axis represents the Metric value where the red line indicates the mean of the fare rates while the blue line indicates the median of the fare rates.

The left plot represents the variation in the mean and median fare rates in Park Slope with respect to the hour of the day. Similarly, the right plot represents the variation in mean and median fare rates in East Harlem North.

It is observed that in both the zones the fares are highest in the evening after 3pm. As our POIs come under commercial and recreation category, it indicates that maybe after the office hours people tend to travel which means there is high demand for taxi at this time and thus the fare rates are high.

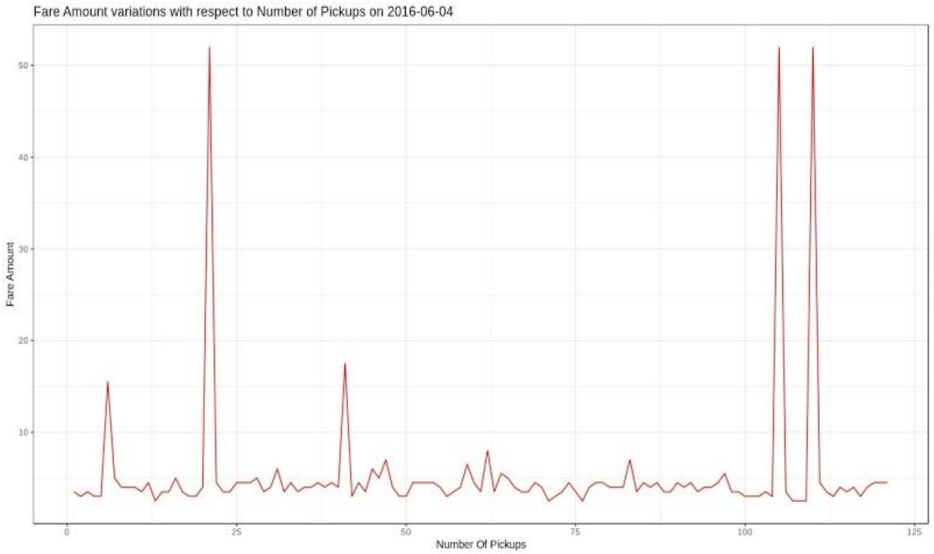
- **Analysis on fare rates with respect to number of Visits to POI on same day - East Harlem North**

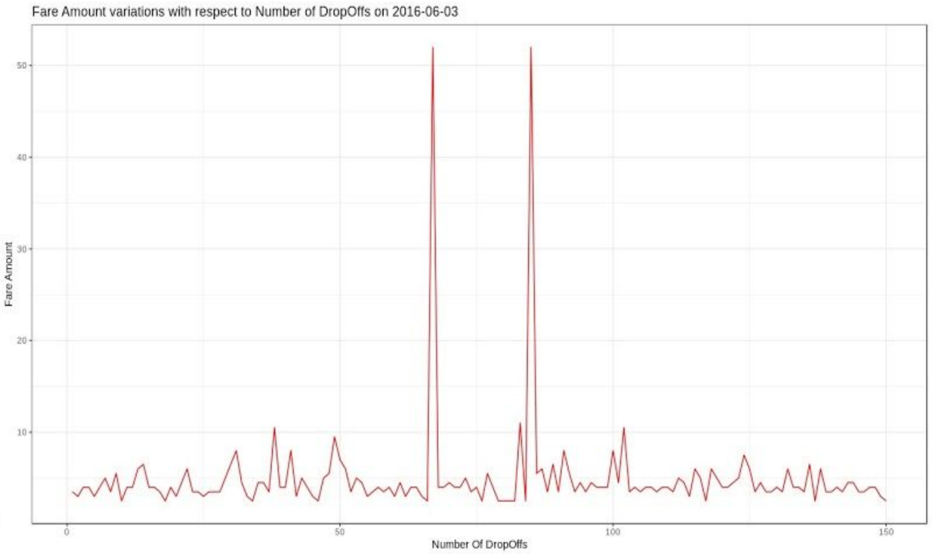




In the left plot the X-axis represents the PickUp date and Y-axis represents the Number Of Pickups in East Harlem North. It is observed from the plot that most number of Pickups were on 2016-06-04 with more than 3000 pickups near the POI.
In the right plot the X-axis represents the DropOff date and Y-axis represents the Number Of DropOffs in East Harlem North. It is observed from the plot that most number DropOffs were on 2016-06-03 with more than 1500 dropOffs near POI.

These dates can then be used to analyse the fare rates depending on the number of visits to that place.





In the above plots the X-axis represents the number of visits and the Y- axis represents the fare rate with respect to the number of visits.
The left plot represents the fare variations when there is high frequency of Pickups from East Harlem North(2016-06-04) while the right plot represents the fare variations when there is high frequency of DropOffs to East Harlem North(2016-0603) near POI.
From both the plots, it is observed that the frequency of visits to a place near POI does not have much effect on the fare rates. There are unusual spikes which might be due to time of travel that is the PickUp or DropOff might be in midnight or in the evening.

6. Conclusion:

The above exploration and analysis of NYC Green Taxi Trip Data with respect to NYC Point of Interest shows that in the month of June 2016, the highest number of the pickups were from Brooklyn which indicates many people visited Times Plaza Area, Pacific Library, HRA Job Center Southern Brooklyn and John Jay HS Historical in Park Slope, Brooklyn.
On the other side the highest number of dropOffs were observed in Manhattan which indicates that many people visited ABC Echo Park Children & Family Center, Park Avenue Hotel, Latin Amer Council Pentecostal in East Harlem North, Manhattan.
Thus we explored both the zones Park Slope and East Harlem North and found the most visited Point Of Interest in the respective zones.

On exploring the trips near most visited POI it is observed that most of the trips are on weekends this indicates that the number of street hail taxi should be increased on weekends in these zones so that people have an easy access and can travel comfortably.
For people who want to travel with cheap rates should travel to these places in the morning or the mid of the day when the taxi fare rates are low.

7. Reflection:

This exploration project proved to be a great learning curve where I learnt exploring the different aspects of data using R. I used Python for data processing and R for explorations. With the help of this project I learnt Kmeans clustering and its applications in different areas. This project helped me to improve my visual analytics skills by making use of different visualisations to answer the questions. I have other ideas that will help me to retrieve better insights such as applying network theory and showing visualisations using networks, when there is a taxi trip from one point to the other. I am aiming to show this in the next visualisation project.

8. References:

About TLC - TLC. (2019). Retrieved from https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

R, E. (2019). Extract shapefile value to point with R. Retrieved from https://stackoverflow.com/questions/34272309/extract-shapefile-value-to-point-with-r

R, C. (2019). Converting geographic coordinate system in R. Retrieved from https://gis.stackexchange.com/questions/45263/converting-geographic-coordinate-system-in-r

K-Means Clustering in R Tutorial. (2019). Retrieved from https://www.datacamp.com/community/tutorials/k-means-clustering-r

distHaversine function | R Documentation. (2019). Retrieved from https://www.rdocumentation.org/packages/geosphere/versions/1.5-5/topics/distHaversine