

Foundation of Data Science

Final Project Report

EXPLORATORY ANALYSIS OF NYC UBER DATA
TO PREDICT PICK UP SURGES ON A US HOLIDAY.

Megha Sharma : ms11342

Rahul Manjunath Ashlesh: rma460

New York University

Problem Statement

Predict locations in New York City with the best potential to get a ride request for an Uber cab driver for a given hour of the day, on a US holiday.

Motivation:

During the Uber data analysis we noticed random anomalies in the number of rides and ride frequencies, upon investigation we noticed that these were the US holidays. This made us reflect on an encounter with a New York City cab driver who claimed that his intuition to reach places with a high probability to get a ride request on normal weekdays and weekends did not hold good for US holidays. We hope that this prediction can help Uber drivers to know the best locations in New York City where they where they may have a high probability to get a ride request on a US holiday. This could also be extended to estimate the number of available cars required during the holiday season, which in turn will help to optimal decisions regarding supply and demand of Uber cars.

Why predict locations with pick up surges?

First rule every rideshare driver learns is to not chase the surge but predict them.

Data:

We have used New York City Uber data set for the year 2014 from:

- <https://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city>

Features:

1. Timestamp (Year, Month, Date, Hour, Minute, Second)
2. Location latitude
3. Location longitude

Target Variable:

High Density Cluster locations such that:

- Inter cluster distance is approximately 1 block
- Number of rides in 1 hour is 10 or more
- Thus on average 1 ride is requested every 6 min in a radius of 800 meters = (approx) 1 block distance in Manhattan.

Data Exploration Pipeline:

Our data exploration pipeline can be visualized in as per the flowchart below. We have explain each step in detail in this report.

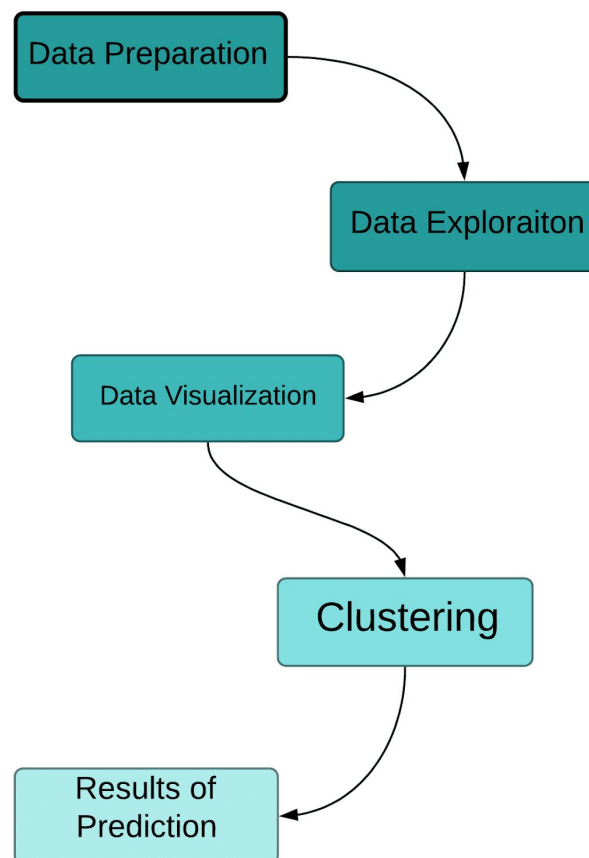


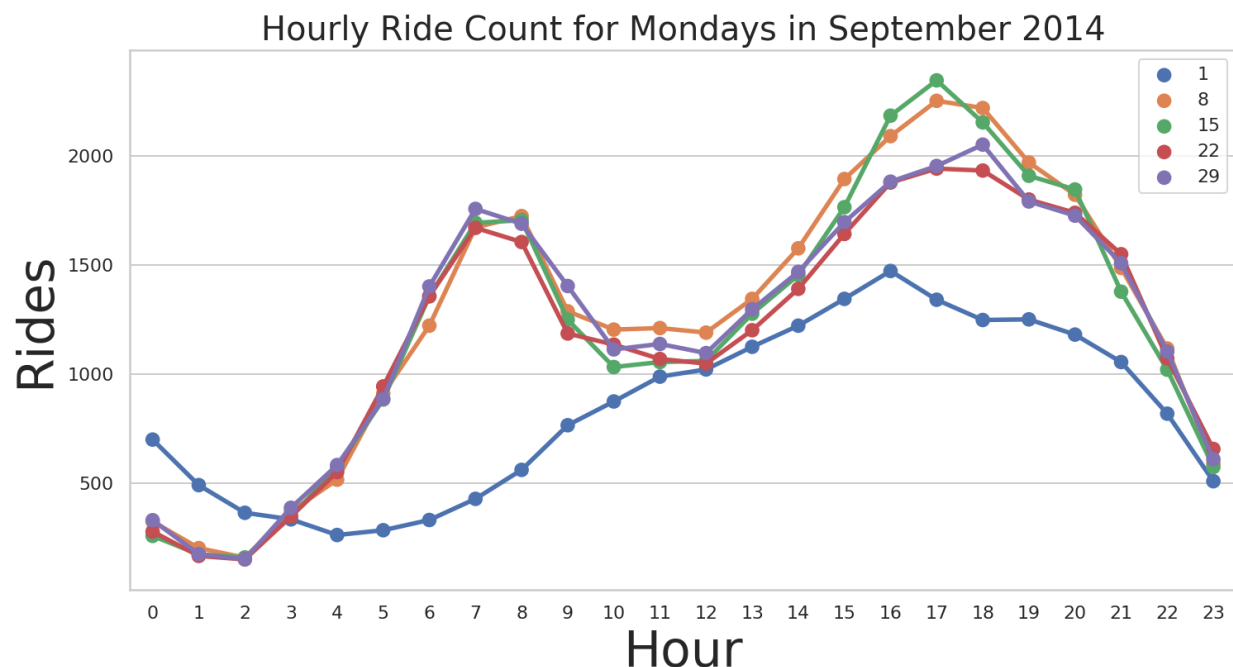
Chart 1. Data exploration Pipeline

Step 1: DATA PREPARATION:

1. The 'uber-raw-data-sep14.csv' data was loaded to a Pandas dataframe.
2. Dropped columns which were not used in this analysis, to optimize data size.
3. Checked and removed NAN values in the dataframe.
4. Created new columns Month, day, hour and min using the timestamp column.
5. Converted the data types as per the model we used in the subsequent steps.

Step 2: DATA EXPLORATION:

Data Exploration was done by aggregating data, and studying the data, which helped us identify the patterns in the data. So initial observations were that the number of rides during weekdays and weekends had a distinct patterns. The average rides on weekdays were similar when compared to average rides on weekends. The prominent difference between weekday and weekend can be noticed in the early morning hours(12AM - 4AM) and late night hours(10PM - 12AM) average rides. A simple line graph shows this pattern.



Graph 1: Weekly average rides per hour for september 2014

Explored the data by Studying:

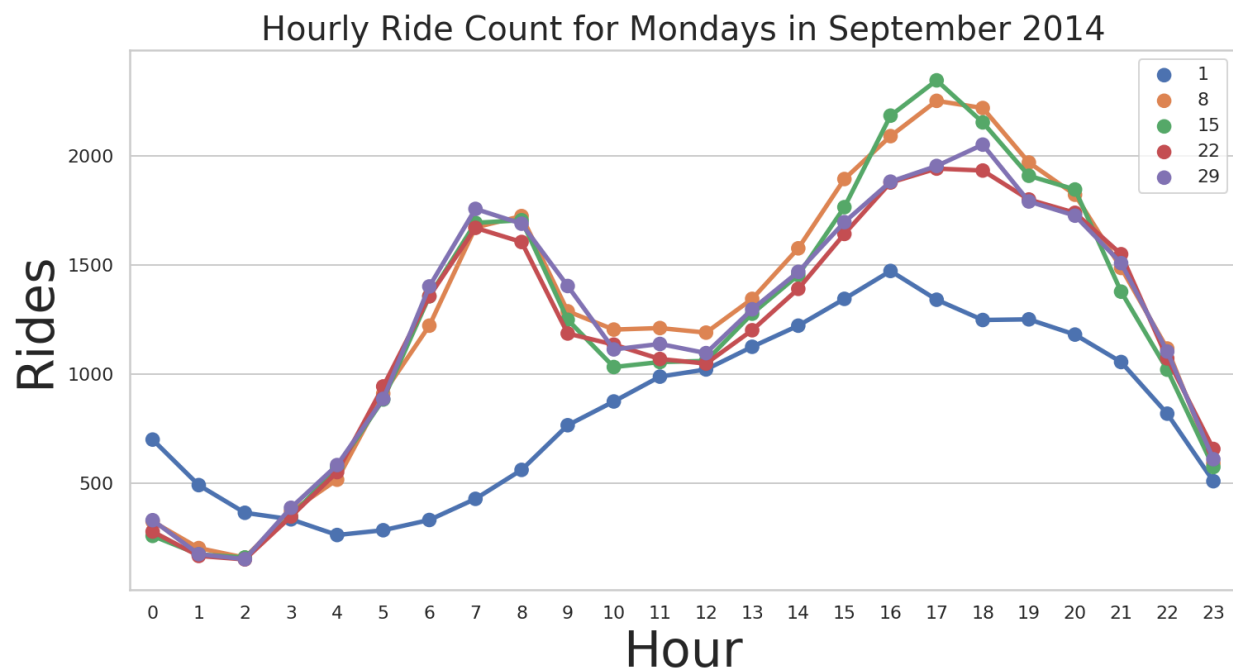
1. Average rides per hour throughout the month.
2. Average rides per hour per weekday & weekday & US Holidays.

By now we wanted to select one US holiday as a sample to test our approach, and we found that the most variation in the number of rides among all the 10 US holiday was found to be on Labour day.

Step 3: DATA VISUALIZATION

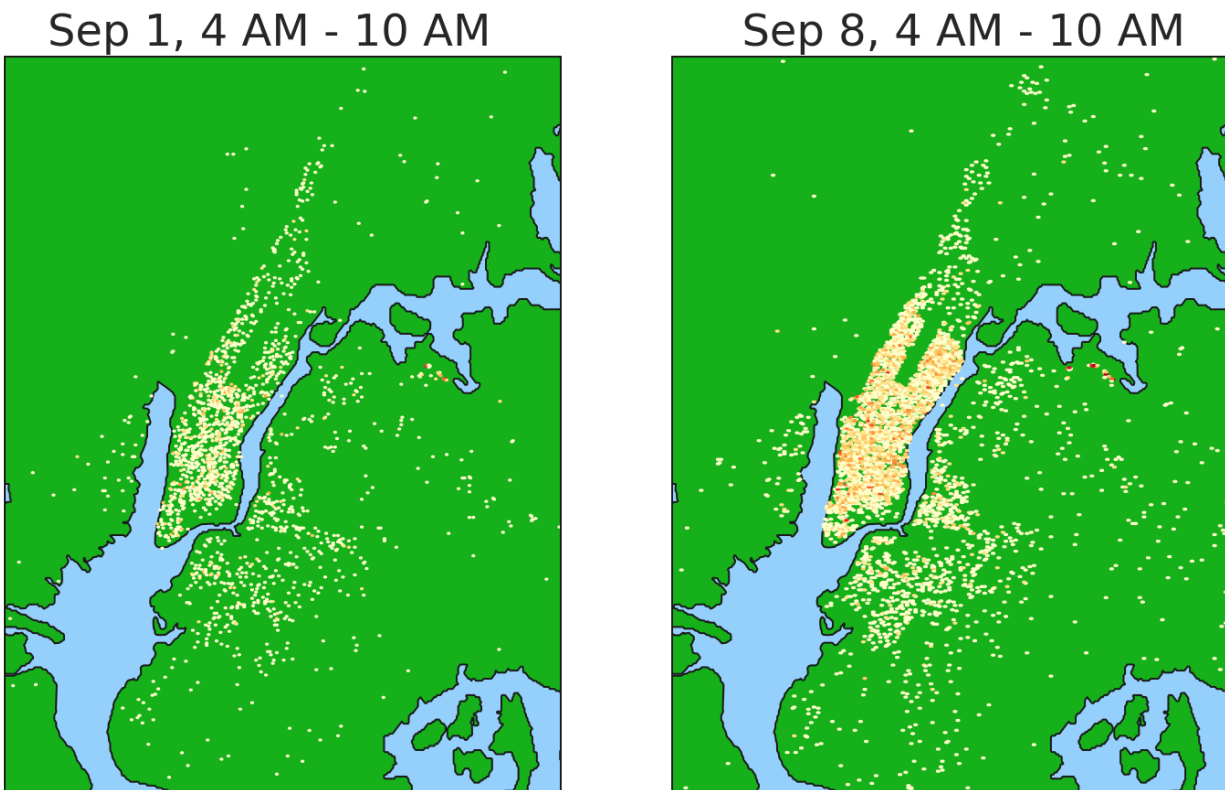
We have used Seaborn, HeatMap and mplleaflet to visualize the data.

Visualizing the Labour day September 1, 2014 we noticed a very distinct variation in the number and time of rides when compared to other Mondays in the same month. We then plotted the below graph 2 visualize the differences.



Graph 2: Hourly Ride count for all Mondays in September 2014, The labels 1,8,15,22 and 29 correspond to the Mondays in that month. Mondays in September: On Mondays (other than the labour day): High pickups 6am to 9am and 4pm to 8pm. However, on the labour day: the overall trend of pickups is unusual and some high pickups are seen from 2pm - 6pm.

We then wanted to visualize this data on a geospatial plot to study the location with high rider turnout.



Graph 3: Heat Map for all ride pickup location between 4 AM and 10 AM on Sep 1st and Sep 8th 2014.

We have plotted a heat map for two consecutive Mondays, 1 September (labour day) and 8 September and made the following observations.

HeatMap Visualisation findings:-

1. There are very few rides booked in the Upper East, Upper West and Midtown Manhattan region. This could be a clear indicator of the holiday, where the usual corporate employees in Manhattan may not needing a ride. But other mondays large number of rides were booked in Midtown Manhattan and around Central Park from at 4am-10am.
2. Thus we can show that on Labour day there were very sparse pickups around central park, proving out initial assumption in this project.

Step 4: CLUSTERING

We decided to use Clustering on the data to get a valuable insight on the patterns in which the people who booked a ride on uber were distributed around in NYU and identify the locations with high density in the number of rides being requested.

We chose clustering as it can be used to predict the high density pick up clusters. It is an unsupervised learning algorithm. We used the algorithm because after visualization we found the pattern of pickups forms clusters of varying densities and shapes.

Therefore to predict locations of high density pick ups, it will be the right approach to use clustering algorithm and not predict a certain latitude longitude. This will give the uber drivers a broad idea about the most in demand locations as clusters.

Model Selection:

We investigated a few clustering algorithms and decided to use Density-based spatial clustering of applications with noise (DBSCAN)

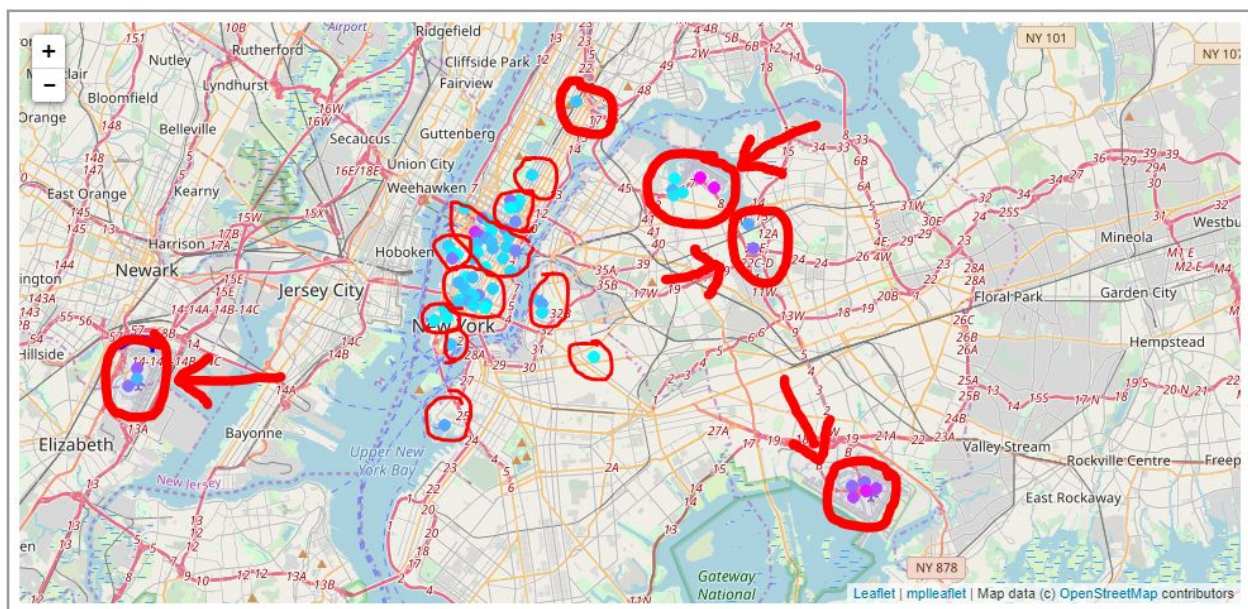
We decided not to use **K-Means & Mean shift** algorithms because of the following reasons:

- K -Means clustering algorithm forms spherical clusters - this might not be the best cluster shape formation in our use case.
- K-Means requires the number of clusters to be specified - we did not have any prior to the implementation of the algorithm thus we could not always decide the best number of clusters. This also makes the cluster forming less natural and could introduce bias in the model.
- K-Means & Mean Shift are not consistent - for the same input, they may generate a different cluster every time, we needed a consistent model.
- They do not handle noise well.

We decided to use DBSCAN clustering algorithm because of the following reasons:-

- Real life data may contain irregularities and therefore the clusters can be of irregular shapes. DBSCAN can form clusters of arbitrary shapes.
- DBSCAN can work with noise in data.
- DBSCAN is consistent with the order of the input data.
- DBSCAN needs only 2 parameters. : eps: maximum distance between the 2 points to be neighbors in a cluster. If the distance is greater than eps then the points are outliers. MinPts: It is the minimum number of data points required to form a dense region/cluster.
- It performs well in separating high density clusters with low density clusters.
- Handles outliers.
- Does not require a prior specification of the number of clusters.

The below plot shows the DBSCAN clusters for September 1st from 4 PM to 8 PM.



Graph 4: Clusters formed for labour day input data with conditions- minimum 10 rides between 4PM and 8PM.

Observations

We used DBSCAN with eps (Maximum distance) of 1 block and MinPts(Number of pickups in a cluster) 10. For a sample observation we Plotted a cluster graph for labour day between 4PM and 8PM and we found that:

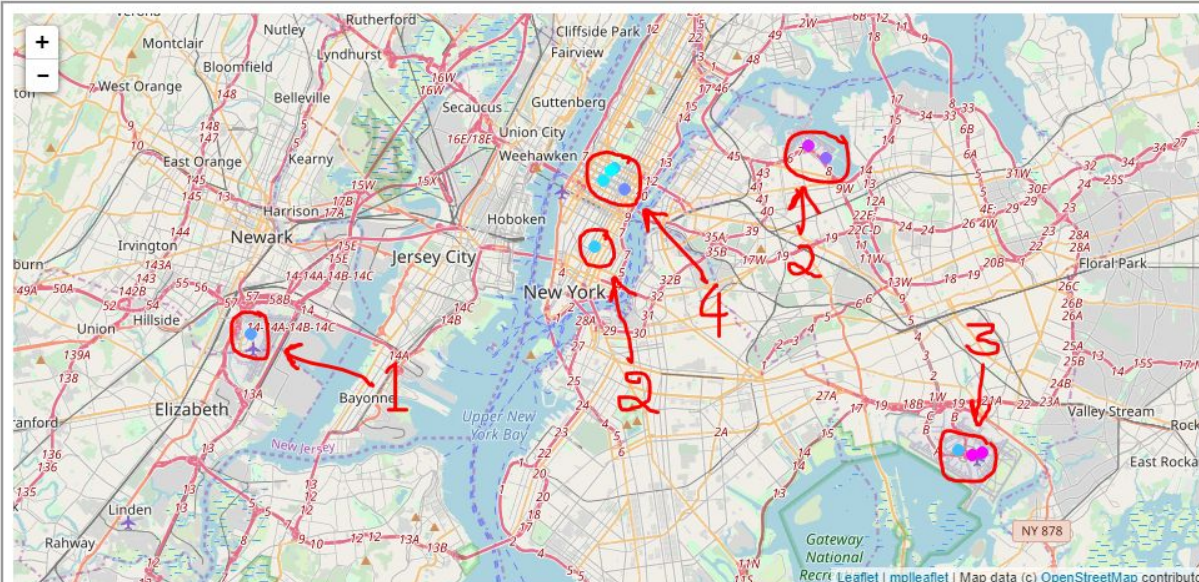
Number of clusters formed: 71

nodes were eliminated as noise: 4938

Some of the noticeable regions with number of clusters were :

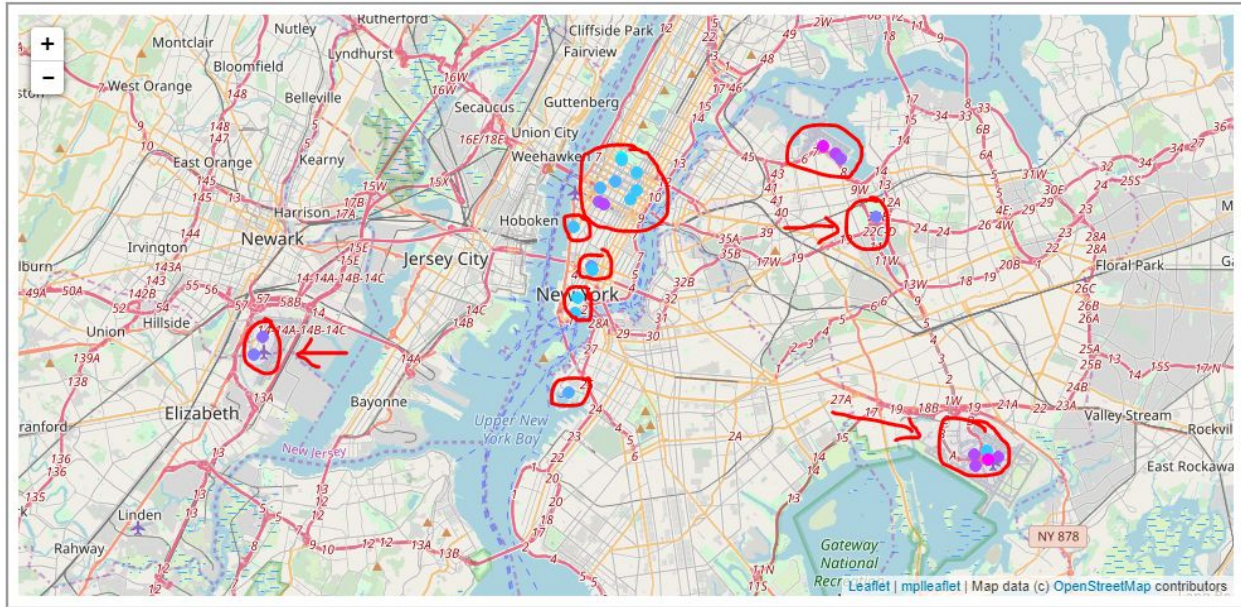
Area/Location	Number of clusters during 4PM to 8PM
JFK Airport	5
LaGuardia Airport	6
Newark Liberty Airport	3
Penn Station, Manhattan(in and around)	9
Downtown Manhattan (highest count along Broadway-Lafayette St)	20
MidTown Manhattan (Best locations: 34St Herald Sq, 42 StTime Sq, 59St, Madison Av -77 St)	9
Bedford Avenue, Kent Avenue	3
Ikea, 1 Beard St, Brooklyn	1
New York State Pavilion	1

Number of clusters: 12
Noise: 2325



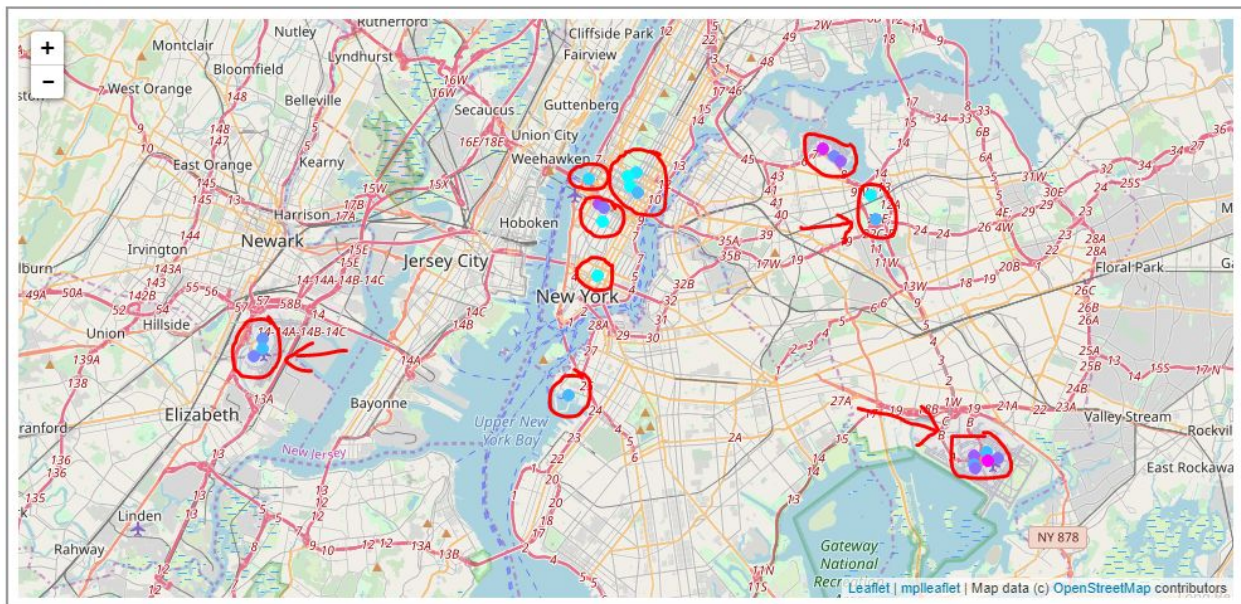
MODEL ACCURACY & INFERENCE:

We found the results to be very consistent with 27 clusters in the training set and 25 clusters formed in the test set. All the clusters formed in the training set were in the same region as the clusters formed in the test set. The Graph 6 and 7 show the training and test set plots.



Graph 6: Clusters formed for labour day input using the training set. 27 clusters were formed.

Number of clusters: 25
Noise: 3325



Graph 7: Clusters formed for labour day input using the training set. 25 clusters were formed.

RESULTS:

The Training and the test set yielded similar results with also formed clusters in the same regions. We can infer that the dataset used to perform this test has a good distribution. The training set having 3 more clusters than the test set can be found in JFK, Newark airports and New York State Pavilion. There might be a slight bias during the distribution which is causing this difference.

EVALUATION APPROACH:

DBSCAN uses Silhouette score as a metric. We tried to evaluate our model using Silhouette score but we were unable to implement it. The Silhouette score is a homogeneity score function which takes as a parameter `y_labels` which are true labels to compare the ground truth labels prior to clustering to the predicted clusters. Since our DBSCAN model is a clustering model, and not classification model, based on real data therefore we do not have the `true_labels`. Thus we could not evaluate our model using Silhouette score metric.

ASSUMPTIONS & LIMITATIONS:

Assumptions:

Our model is based on the assumption that the number of pickups in any location in NYC vary similarly in the following time slots:

12am - 3am : night

4am - 9am : early morning

10am - 4pm: day time and noon

4pm - 9pm: evening

10pm - 11 pm: night

We have assumed other US Holidays would have similar patterns

Limitations: The model works only for US Holidays.

What did you change from your original proposal and why?

We did not include the weather data and other metric in predicting the locations because it was making the analysis complicated.

Foundations of Data Science Project

Team Evaluation:

Team member Net ID	Score
Ms11342	4
Rma460	4