Taxi Hotspot Clustering And Proximity Recommendation System

Submitted in partial fulfillment of the requirements of

PG-DIPLOMA IN BIG DATA ANALYSIS

By

Monica Jha	230310125011
Lokesh Sali	230310125009
Shivani Phuke	230310125017

Project Guide

Dr. Priyanka Jain

(AD, C-DAC, ACTS, New Delhi)

CENTER FOR DEVELOPMENT OF ADVANCED COMPUTING

New Delhi.

March 2023 – September 2023

TABLE OF CONTENTS

CERTIFICATE

This is to certify that the project entitled "**Taxi Hotspot Clustering And Proximity Recommendation System**" is a teamwork work of "**Monica Jha (230310125011), Lokesh Sali(230310125009), Shivani Phuke (230310125017).**" Submitted to **CDAC, ACTS, New Delhi** in partial fulfillment of the requirement for the PG- Diploma in Big Data Analysis.

Dr. Priyanka Jain
(AD, CDAC, ACTS, NEW DELHI)
(Poject Guide)

Mr. Anurag Singh
(Faculty Supervisor/Guide)

DECLARATION

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

(Monica Jha)

230310125011

(Lokesh Sali)

230310125009

(Shivani Phuke)

230310125017

ABSTRACT

Through this project, we try to predict the best possible way to find the number of hotspots feasible for taxis in a city. As input, we take the pickup points's latitude and longitude as coordinates. Using these coordinates we try to predict the regions/hotspots, where there could be the highest possibility of getting a taxi. Using this, the business can actually report the waiting spots/hot spots to be made. So that the customers can go to these hotspots to get a taxi booked. Along with this, the project can be further extended as an application, where on the basis of customers current latitude and longitude, the model can recommend him/her the nearest hotspot.

INTRODUCTION

In this Project, we find the number of hotspots feasible for taxis in a city, where there is the highest possibility of getting a taxi. The Taxi Hotspot Prediction Model will enable people to understand and view the various regions in which the taxi count is higher / lower. Using this the business can actually report the waiting hotspots to be made so that customers can go to these hotspots to get a taxi booked. The project can be further extended as an application, where on the basis of customers current latitude and longitude, the model can recommend the nearest hotspot to them.

DESCRIPTION

People find it difficult to get taxis at their location and there are so many of them who wait too long for a taxi possibly cause taxi driver rejects the ride or in city there are certain areas where taxi is not easily accessible. Hence its ideal to find a best service for customers. A Taxi Hotspot Clustering and Proximity Recommendation System is a technology-driven solution designed to enhance the efficiency and convenience of taxi services by identifying popular or strategic locations (hotspots) for pickups and providing real-time recommendations to taxi drivers or passengers regarding the nearest available taxis or optimal pickup points. This idea extends on the basis of customers Location we can recommend the nearest hotspot as a pickup point with the highest Possibility of getting a taxi by using their pickup points (on the basis of their Longitude and latitude) as coordinates.

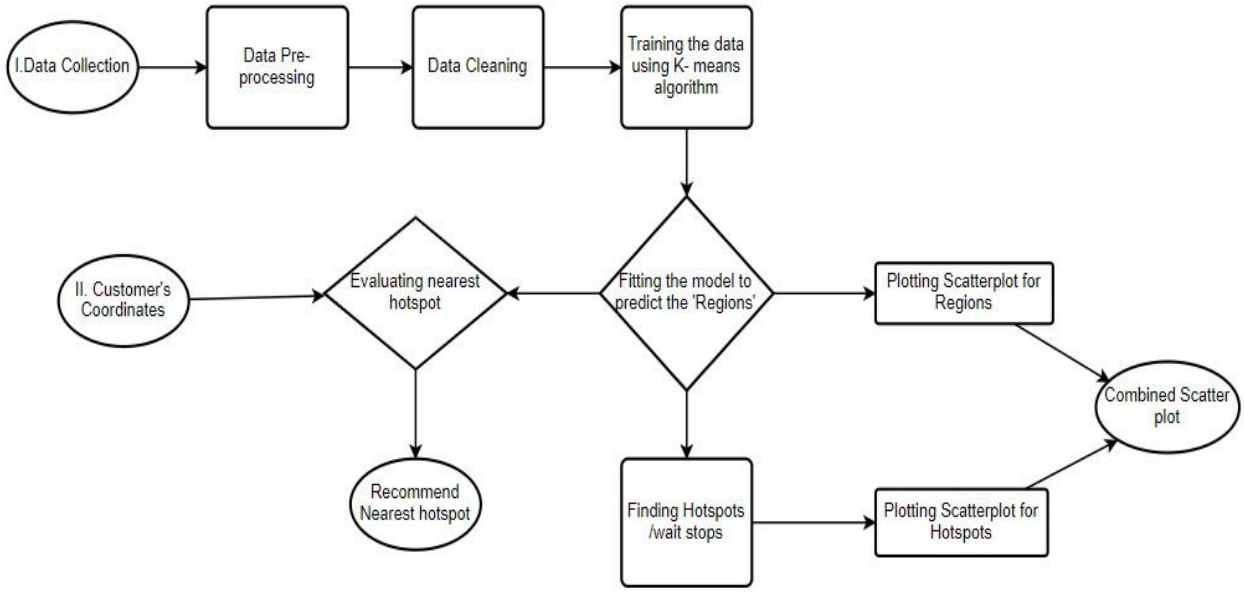
SCOPE OF PROJECT

- The primary goal of this system is to enhance taxi services by reducing waiting times for passengers and increasing ride opportunities for drivers. It achieves this by recommending optimal pickup points based on real-time data analysis.
- Taxi drivers benefit from improved opportunities to find passengers quickly, reducing downtime and increasing their earnings. The system helps drivers make more informed decisions about where to wait for ride requests.
- Taxi companies can use the system to identify underserved areas with high demand for rides, allowing them to expand their services strategically.

TECHNOLOGY USED

- Machine Learning: Google Collab for generating model
- Programming language: Python with libraries: NumPy, Pandas, matplotlib and seaborn for visualisation.
- Algorithm used: K means algorithm
- Dataset source: Kaggle
- Dataset link : https://www.kaggle.com/code/sohaibanwaar1203/taxi-demand-prediction

WORKFLOW DIAGRAM



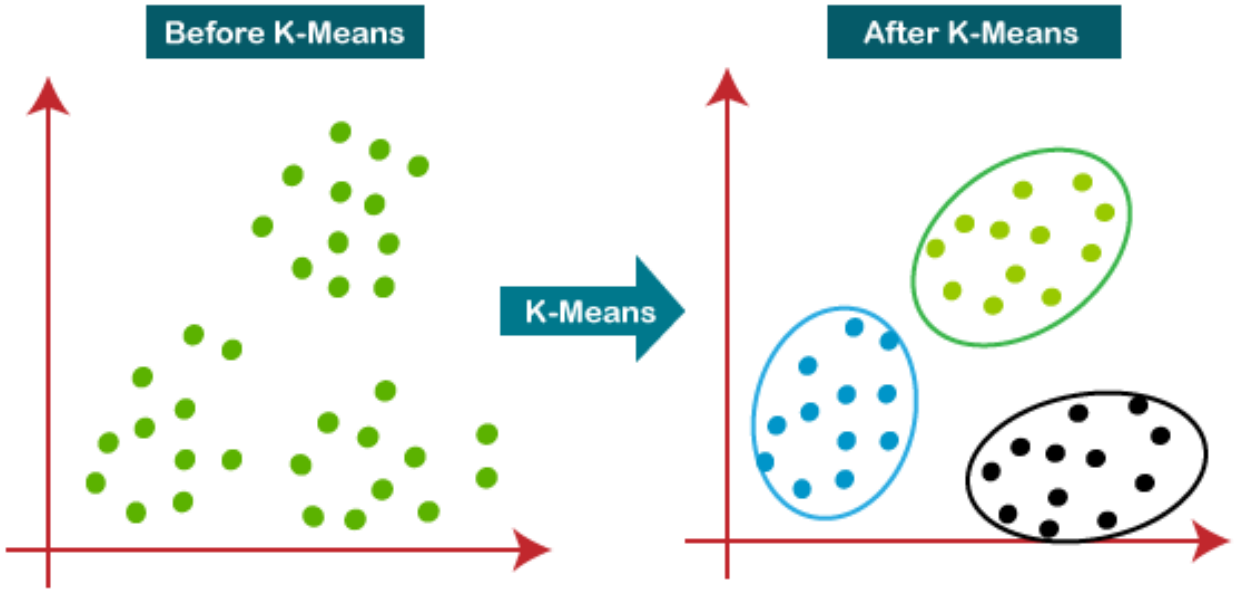
ALGORITHM USED

K-Means Clustering:

K-means clustering is a popular unsupervised machine learning algorithm used for partitioning a dataset into distinct groups or clusters based on similarity or distance metrics. It's commonly used for data analysis, data mining, and various applications in machine learning.

The k-means Clustering algorithm mainly performs two tasks:

- Determines the best value for K centre points or centroids by an iterative process.
- Assigns each data point to its closest k-centre. Those data points which are near to the particular k-centre, create a cluster.



Centroid: Centroid is the mid point of Cluster. In other words Nearest Point of the Number Of Clusters(K).

The working of the K-Means algorithm is explained in the below steps:

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroids. (It can be different from the input dataset).

Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.

Step-4: Calculate the variance and place a new centroid of each cluster.

Step-5: Repeat the third steps, which means assign each datapoint to the new closest centroid of each cluster.

Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.

Step-7: The model is ready.

IMPLEMENTATION OF ALGORITHM IN PROJECT

Regions generation:

Using the above mentioned steps , the clusters are formed. These clusters are nothing but our Regions which have the highest number of pickup points.

So when we have the dataset with pick up coordinates, the algorithm starts with a random centroid and starts locating all the coordinates and eventually forming clusters.

Any new coordinate that would be added in future will work again the same logic and would automatically be assigned a cluster, in our case , Region.

Hot spot/waiting stop:

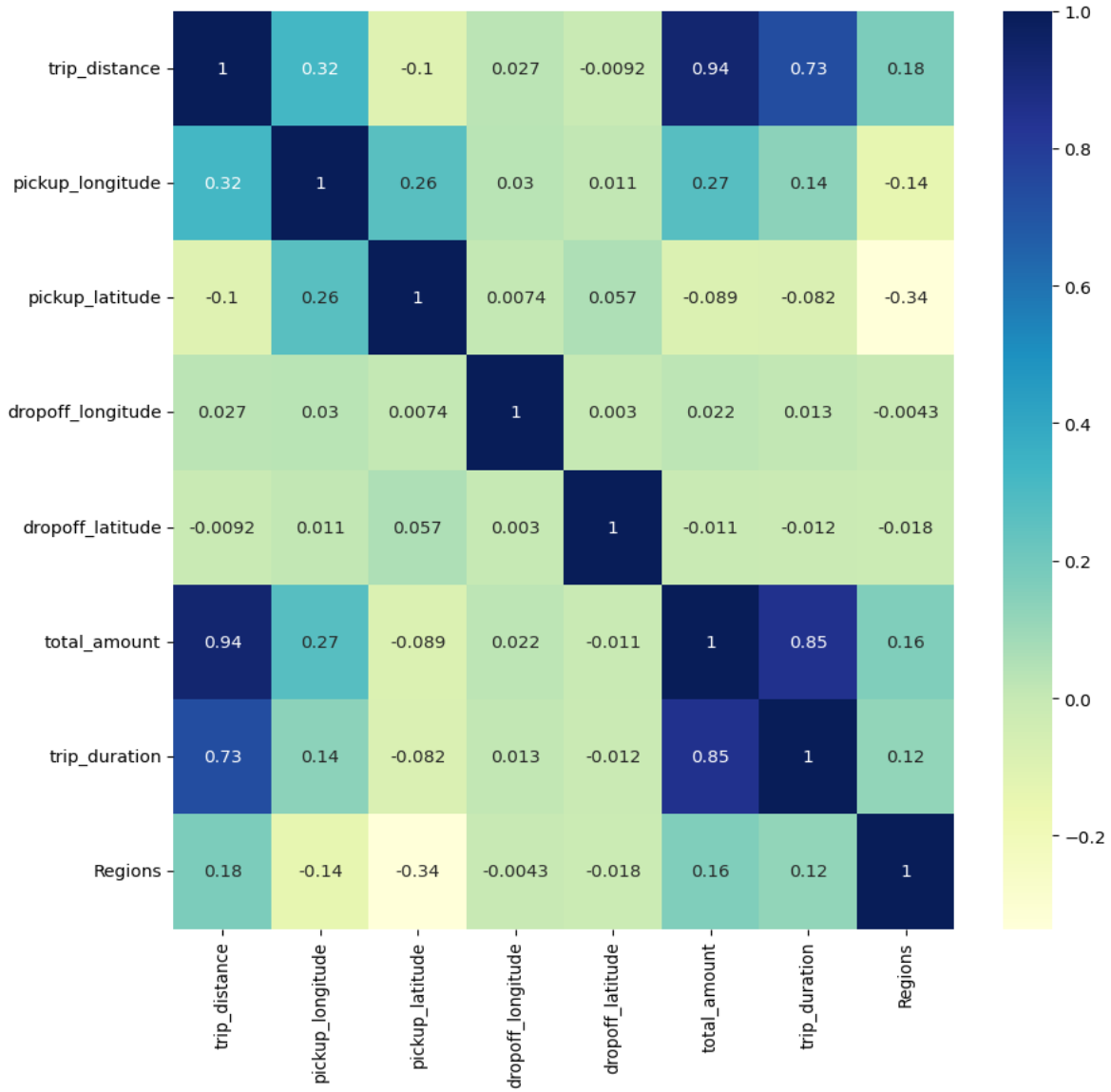
Once the regions are identified, the next step is to identify the waiting stop. Which could be found out on the concept of centroid.

So each region/cluster will have a centroid. This centroid would be nothing but our waiting stop.

VISUALIZATION OF DATA

While performing the exploratory data analysis, following visualizations were performed.

Correlation: This was done to identify the relationships between the variables.



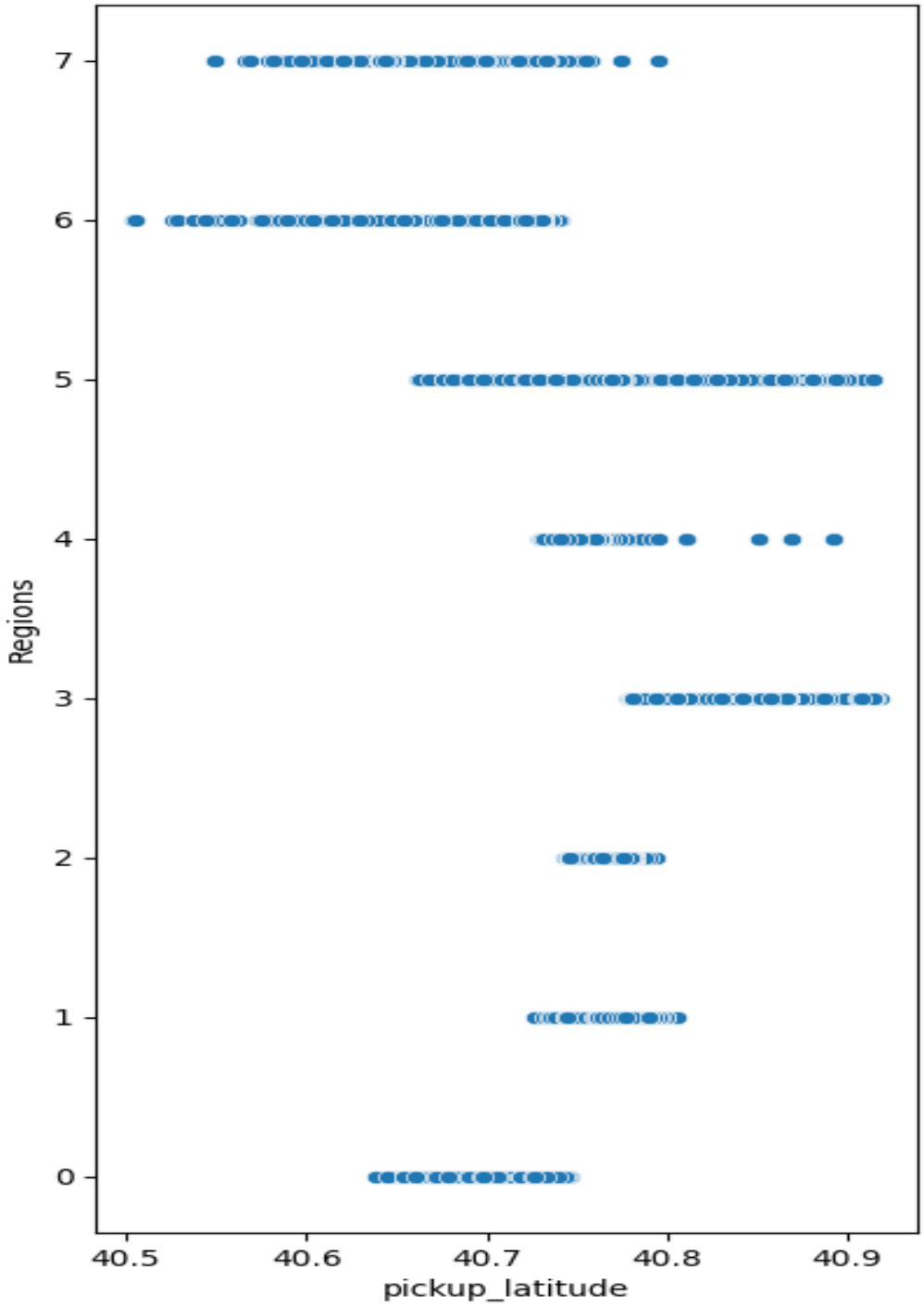
The observation was:

a. trip duration is highly correlated with trip_amount.

b. trip_Amount is highly correlated with trip distance.

Scatter Plot:- -This will shows the relationship between two variables:

I.e: Region vs Pickup latitude and Region vs Pickup longitude



Observation: there was a linear relationship observed between the given variables.

WORKING OF PROJECT

Phase 1: Training the model to predict the regions and hotspots:

1. Fitting the model:

```
pickup_regions = KMeans(init='random').fit(pickup_coordinates)
```

2. Using a variable named "pickup_cluster_column" to predict the regions created:

```
pickup_cluster_column = pickup_regions.predict(df_2015[["pickup_latitude", "pickup_longitude"]])
print(pickup_cluster_column)
```

3. Saving the pickup regions generated into a new column and adding that column to our dataset:

```
df_2015["Regions"] = pickup_cluster_column
```

```
df_2015.head(10)
```

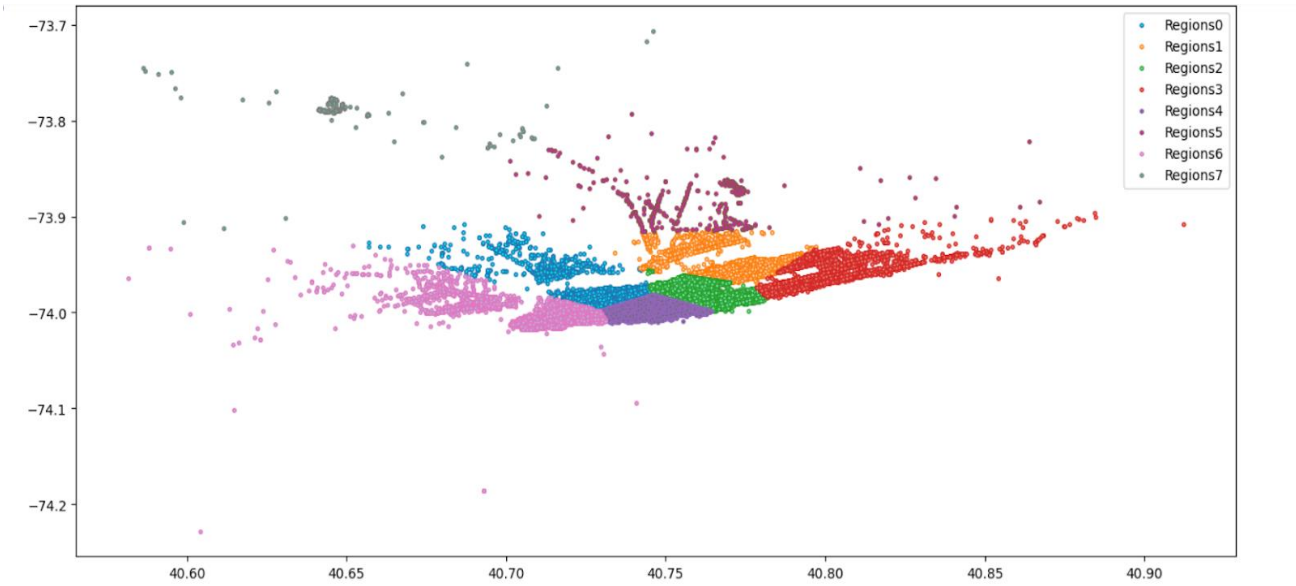
pickup_time	dropoff_time	trip_distance	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	total_amount	trip_duration	Regions
2015-01-15 19:05:39	2015-01-15 19:23:42	1.59	-73.993896	40.750111	-73.974785	40.750618	17.05	18.0	4
2015-01-10 20:33:38	2015-01-10 20:53:28	3.30	-74.001648	40.724243	-73.994415	40.759109	17.80	19.0	6
2015-01-10 20:33:38	2015-01-10 20:43:41	1.80	-73.963341	40.802788	-73.951820	40.824413	10.80	10.0	3
2015-01-10 20:33:39	2015-01-10 20:35:31	0.50	-74.009087	40.713818	-74.004326	40.719986	4.80	1.0	6
2015-01-10 20:33:39	2015-01-10 20:52:58	3.00	-73.971176	40.762428	-74.004181	40.742653	16.30	19.0	2

4. We can now list the number of Regions formed:

```
regions = list(set(df_2015["Regions"]))
print(regions)
```

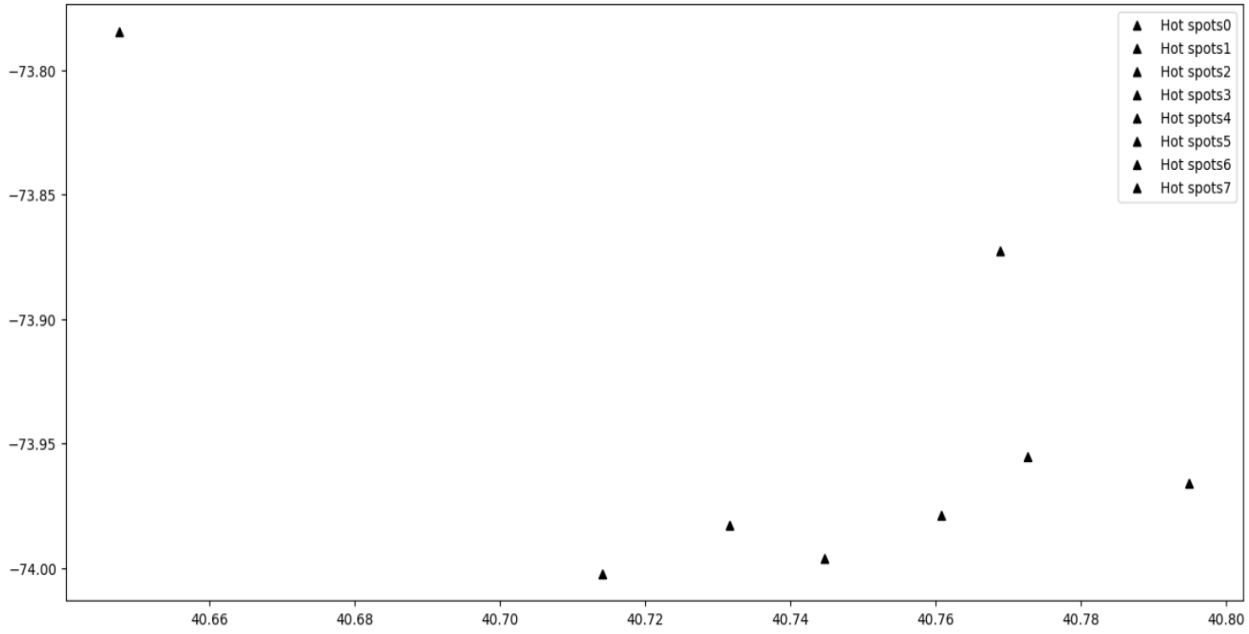
```
[0, 1, 2, 3, 4, 5, 6, 7]
```

5. Once the Regions are determined, we can plot a graph for this. With X - axis as latitude and y-axis as longitude:

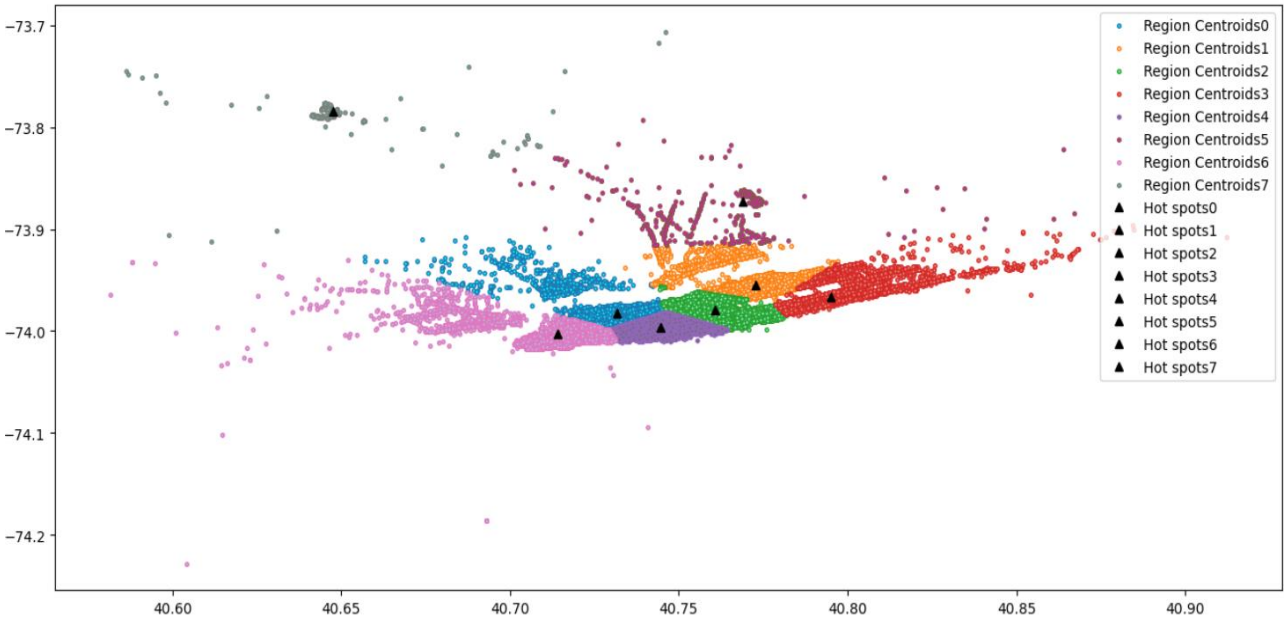


6. Now we can move to find and plot the hot spots/waiting stops:

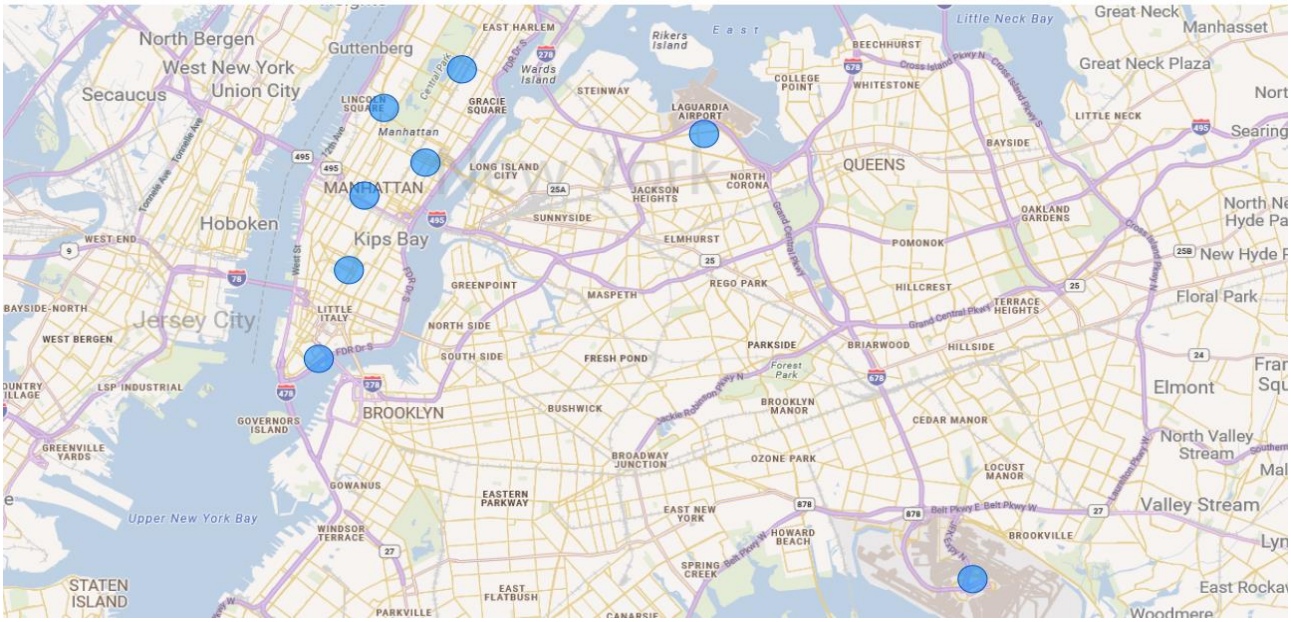
print(pickup_regions.cluster_centers_)



7. On the basis of above two graphs, we can combine the graph to understand better how regions and hotspots would look on a single map:



Similarly, these hot spots can be marked on the map using PowerBI:



Phase 2- Recommending the nearest hotspot:

Here we test our model and recommend customer his/her nearest hotspot:

Recommending the nearest hotspot for the customer on the basis of his/her coordinates

```
# Get a pickup lat-long combination from customer, and predict the cluster number it belongs to.
customer_output = pickup_regions.predict([[40.68, -74.02]])
print("Your nearest hotspot number is :", customer_output)
```

```
Your nearest hotspot number is : [6]
```

FUTURE SCOPE

On The Basis Of This Model:

- We can build a web/ mobile application to locate real time hot spots for customers.
- We can build a fare surcharge model for taxi companies to earn more profit on the basis of these hot spot regions.
- The companies can also reversely use this model to identify the other regions which are not that highly in demand and root causes for it. So that they can increase pickups and install more waiting spots.
- Womens who travel daily or especially at night instead of taking taxi randomly, these hotspots are safe for taking taxi.