CAPSTONE PROJECT – THE BATTLE OF NEIGHBORHOODS

BUSINESS PROBLEM

A restaurateur wants to open an Italian restaurant in Manhattan and Queens Borough in the New York City. He needs to find a suitable neighborhood in these boroughs where an Italian restaurant would flourish and earn him profits.

DATA

The data is obtained using the Foursquare location API that explores the Manhattan and Queens boroughs. This API gathers all information of the various food joints available in the neighborhoods of these boroughs. However, to solve the given business problem, only the top 5 common venues for each neighborhood is considered for the analysis.

METHODOLOGY

The following steps were performed in analyzing and recommending places where the restaurateur could open an Italian restaurant.

1. Download and explore dataset of New York
2. Use Foursquare API to explore the hoods of Manhattan and Queens
3. Analyze each neighborhood
4. Cluster neighborhoods for Manhattan and Queens
5. Examine clusters and recommend hoods

Download and explore dataset of New York

The New York dataset is downloaded from the URL that was used in Week 3 Lab. The New York city has 5 boroughs and 306 neighborhoods which is extracted from the 'features' tag of the downloaded json data.

Since the restaurateur is interested in only Manhattan and Queens, separate dataframes for each borough that contains its neighborhoods and associated geospatial coordinates are created.

The geospatial coordinates of Manhattan and Queens is obtained using the geolocator APIs that will be used to plot the map.

Use Foursquare API to explore the hoods of Manhattan and Queens

The next step involves exploring the hoods of these boroughs to get the top 5 food joints in each hood. The Foursquare venue API gets the geospatial coordinates, venue name and its category for each hood in Manhattan and Queens. It explores each neighborhood within a radius of 2000m and returns a maximum of 100 venues.

Analyze each neighborhood

In order to determine the most appropriate place to open an Italian restaurant, we need to determine which neighborhoods are similar to each other and separate them from dissimilar ones. This can be done using clustering techniques. However, clustering requires the dataset to be numeric. Thus, one hot encoding is employed on the extracted dataset of each borough. Then for each hood, the venues are sorted based on its frequency and the top 5 are considered for clustering.

Cluster neighborhoods for Manhattan and Queens

KMeans clustering is used for the analysis where the hoods are clustered in 4 groups. The resulting clusters are plotted on the map (each for Manhattan and Queens) using Folium Map APIs.

Examine clusters and recommend hoods

Once the clusters are obtained, the number of Italian restaurants in each cluster is determined. The clusters having the highest number of Italian restaurants suggest that those neighborhoods are best to have a restaurant set up. In particular, the hoods that belong to these clusters but don't have an Italian restaurant in the top 5 venues are the most appropriate ones and is recommended to the restaurateur.

RESULT

The clustering analysis can be viewed on the map as follows:
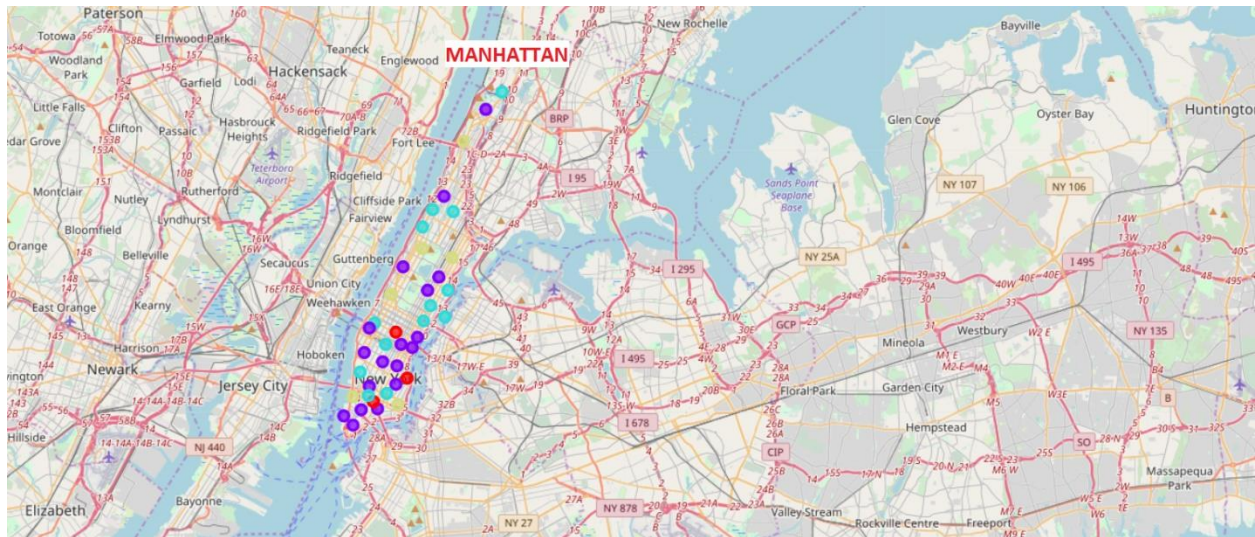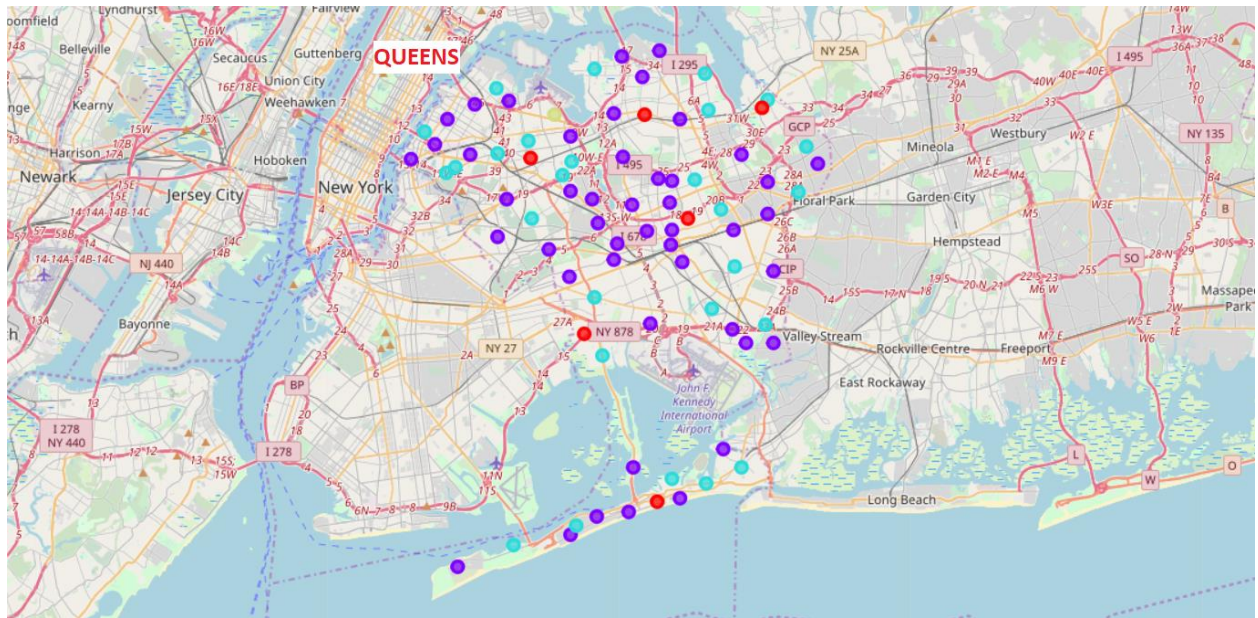
The recommended hoods in Queens are:

	Neighborhood
0	Kew Gardens
1	Richmond Hill
2	Flushing
3	Maspeth
4	Glendale
5	Woodhaven
6	South Ozone Park
7	Kew Gardens Hills
8	Briarwood
9	Jamaica Center
10	Queens Village
11	Hollis
12	South Jamaica
13	Springfield Gardens
14	Cambria Heights
15	Rosedale

	Neighborhood
16	Broad Channel
17	Breezy Point
18	Arverne
19	Rockaway Beach
20	Neponsit
21	Queensboro Hill
22	Hillcrest
23	Rockaway Park
24	Brookville
25	North Corona
26	Jamaica Hills
27	Utopia
28	Pomonok
29	Sunnyside Gardens
30	Bayswater

The recommended hoods in Manhattan are:

	Neighborhood
0	Inwood
1	Murray Hill
2	Gramercy
3	Tudor City
4	Marble Hill
5	Clinton

DISCUSSION AND RECOMMENDATION

Efficient analysis and recommendation can be done if the restaurateur provides additional details required to open a restaurant which includes infrastructure area, budget and proximity to amenities. In such a case, different machine learning algorithms like regression can be applied to estimate a proper location to start a venture.

CONCLUSION

Given the problem statement to open an Italian restaurant in Manhattan and Queens boroughs in the New York city, Foursquare API is used to obtain the most common venues in the neighborhoods of these boroughs. Clustering of these venues groups similar hoods and gives an idea about those hoods that are more likely to have Italian restaurants. Recommendations are made by suggesting hoods that belong to such clusters but haven't had an Italian restaurant.