The Battle of Neighborhoods

Applied Data Science Capstone Project

Abdel A. Elkhadiri

abdel.Elkhadiri@gmail.com

Table of Contents

Introduction – Problem Statement

1. **Background:**
   - A Greek entrepreneur restauranteur seeking to open a Mediterranean Restaurant in New York City
   - The entrepreneur wants to add Data Scientist to his quest to identify the best Borough/Neighborhood to open the restaurant

2. **Problem Statement:**

What's the best New York City Neighborhood to Open a new Mediterranean Restaurant?

Methodology - Data Selection

- Dataset to be used for the project is the New York neighborhood dataset to be downloaded from IBM Cloud (https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json)

- We will use Foursquare API, one of most popular location based social networks (LBSN), to explore venues in the New York City neighborhoods. (https://foursquare.com/)

Methodology - Data Pre-Processing

- After successful data download we process our obtained json format dataset into pandas data frame.
- Explore our dataset using pandas data frame python package
- Using Foursquare Venue API, we can obtained all the venues using the corresponding latitude and longitude obtained from our dataset
- We can visualize all the five boroughs (Bronx, Manhattan, Brooklyn, Queens, Staten Island) and corresponding neighborhoods using folium python library on an interactive map using the Latitude and Longitude of each neighborhood and display them as markers on the map as illustrated in Figure 1

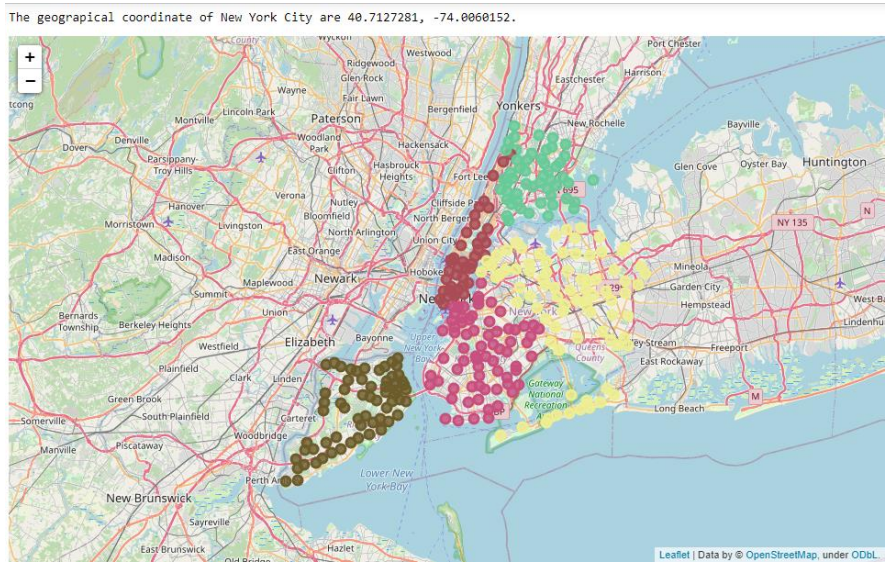


Figure 1: NY city map displaying all five boroughs and corresponding neighborhoods

Methodology - Data Transformation

- We use Machine Learning (ML) method **one hot encoding** to convert categorical variables into a form which we can later use in our ML model
- The categorical value represents the numerical value of the entry in the dataset.
- Using pandas library will transform all venues categories into categorical values. After one encoding, then we group all the neighborhoods by taking the mean of the frequency of occurrence of each category as illustrated in below table (Figure 2)

	Neighborhoods	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Terminal	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	...	Warehouse Store	Waste Facility	Waterfront
0	Allerton	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0
1	Annadale	0.0	0.0	0.0	0.0	0.0	0.181818	0.0	0.0	0.0	...	0.0	0.0	0.0
2	Arden Heights	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0
3	Arlington	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0
4	Arrochar	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0
5	Arverne	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0
6	Astoria	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0
7	Astoria Heights	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0
8	Auburndale	0.0	0.0	0.0	0.0	0.0	0.047619	0.0	0.0	0.0	...	0.0	0.0	0.0
9	Bath Beach	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0

Figure 2: Mean of the frequency of occurrence of each category in New York neighborhoods

Methodology - Data Transformation (Cont'd)

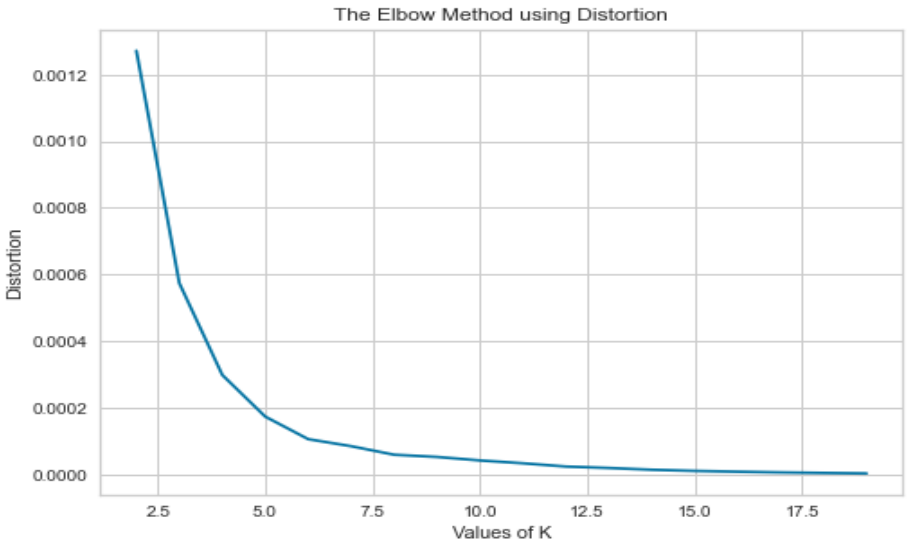
- We finally obtain our final dataset that has the mean of the frequency of all the Mediterranean Restaurant in New York neighborhoods

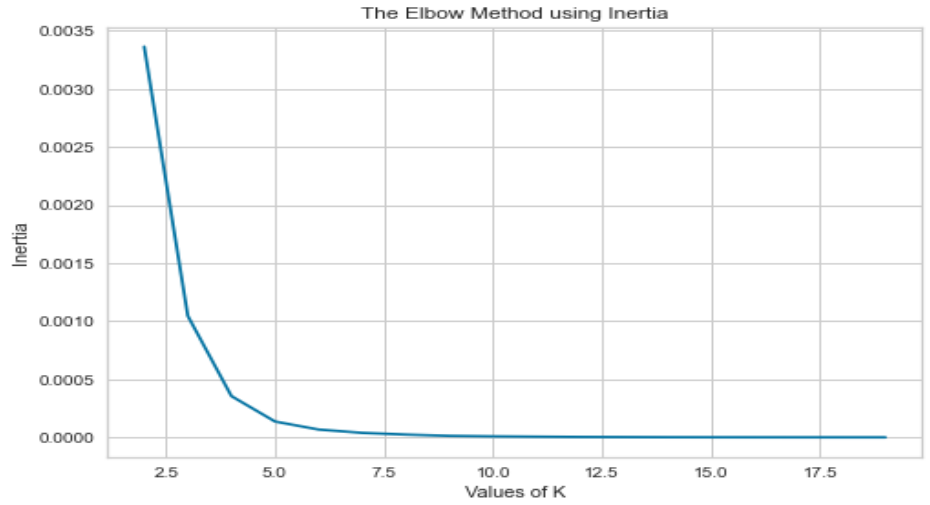
	Neighborhoods	Mediterranean Restaurant
0	Allerton	0.000000
1	Annadale	0.000000
2	Arden Heights	0.000000
3	Arlington	0.000000
4	Arrochar	0.045455
5	Arverne	0.000000
6	Astoria	0.030303
7	Astoria Heights	0.000000
8	Auburndale	0.000000
9	Bath Beach	0.000000

Figure 3: Mean of the frequency of occurrence of Mediterranean Restaurant in New York neighborhoods

Data Mining – k-mean Clustering

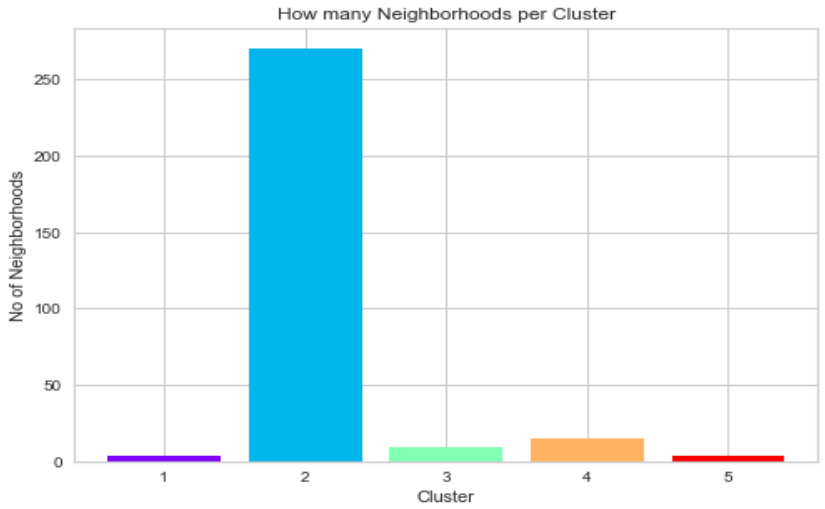
- We will use unsupervised learning algorithm, **k-means clustering** to cluster New York neighborhoods based on the neighborhoods that have similar averages of Mediterranean Restaurant in that neighborhood.
- We use the **Elbow Method** technique to find the optimal number of clusters
- using the Elbow method technique in the sklearn library of python, we define the following:
  - **Distortion:** It is calculated as the average of the squared distances from the cluster centers of the respective clusters. Typically, the Euclidean distance metric is used.
  - **Inertia:** It is the sum of squared distances of samples to their closest cluster center.
- We iterate the values of k from 2 to 20 and calculate the values of distortions for each value of k and calculate the distortion and inertia for each value of k in the given range.
- we imported **KElbowVisualizer** from the **Yellowbrick** package. Then we fit our K-Means model above to the Elbow visualizer. Graphs illustrate the Elbow method using distortion and Inertia.
- **Using k-means Clustering and Elbow method, we concluded that the optimum K value is 5**.

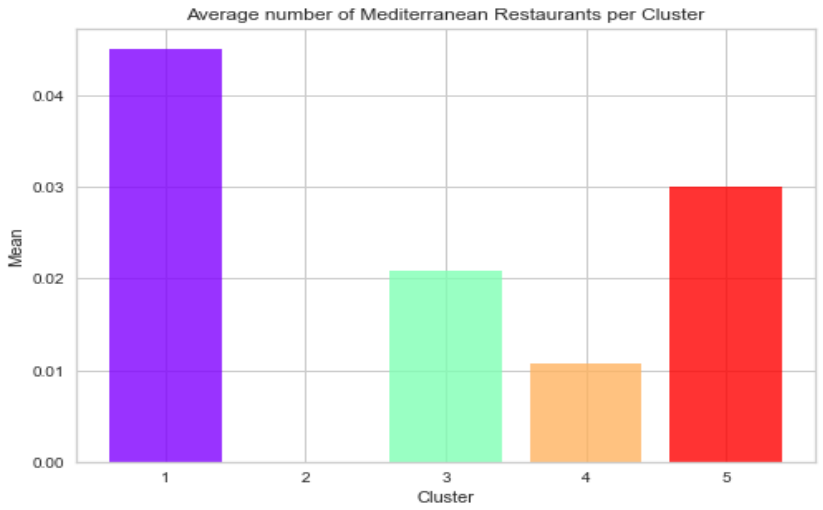




Results Evaluation

- Now and after we obtained a total of 5 clusters, we can start evaluating each cluster by checking the number of neighborhoods in each cluster and the average Mediterranean restaurant in that cluster.
- First, let's find out how many neighborhoods exist per Cluster: **Cluster 1 = 4, Cluster 2 = 270, Cluster 3 = 9, Cluster 4= 15 and Cluster 5= 4** as illustrated in the first graph.
- Now, let's examine the average number of Mediterranean restaurants per Cluster; as shown in the second graph, the average of Mediterranean Restaurants per Cluster are distributed as follows: : **Cluster 1 = 0.047619, Cluster 2 = 0, Cluster 3 = 0.023256, Cluster 4= 0.015873 and Cluster 5= 0.03** as illustrated in the second graph.
- These two graphs give us some critical information to our cluster's analysis. As illustrated, although **Cluster 1 and Cluster 5 both have only 4 neighborhoods, they both have the highest average number of Mediterranean restaurant 0.047619 and 0.03 respectively while Cluster 2 which has 270 neighborhoods has no Mediterranean restaurant.**
- Next will examine each cluster in our Results Evaluation Findings section





Results Evaluation Findings

Next will examine each cluster:

- **Cluster 1** is in Queens Borough and contains 4 neighborhoods with 104 unique venues in which there are five Mediterranean restaurants equating to 0.047619 average as the restaurants are concentrated in two neighborhoods Rego Park and Rockway Park.
- **Cluster 2** contains four boroughs Bronx, Queens, Brooklyn and Staten Island and has the highest number of neighborhoods at 270 with 6001 unique venues with no Mediterranean restaurants.
- **Cluster 3** contains only 2 boroughs Brooklyn and Manhattan and has only nine neighborhoods with 15 Mediterranean restaurants out of 681 unique venues spread out through four neighborhoods Brighton Beach, Morningside Heights, Midtown and Sutton Place.
- **Cluster 4** is in Manhattan, Brooklyn and Queens Boroughs and contains 15 neighborhoods with 1208 unique venues in which there are 16 Mediterranean restaurants equating to 0.015873 average.
- **Cluster 5** covers Queens and Manhattan Boroughs and contains 15 neighborhoods with 346 unique venues in which there are 12 Mediterranean restaurants equating to the second concentration of Mediterranean restaurants after Cluster 1.

Discussion

Based on the analysis of all our optimum clusters, we can conclude that The concentration of Mediterranean restaurants in New York City are in Clusters 1 and 5 with the neighborhoods Rego Park, Rockway Park, Astoria, Soho and Flatiron have the highest average of Mediterranean restaurants followed by Clusters 3 and 4. We can conclude that both boroughs Queens and Manhattan share the highest number of Mediterranean restaurants in followed by Brooklynn while Bronx and Staten Island have little to no Mediterranean restaurants. In addition, Cluster 2 has the greatest number of neighborhoods at 270, but no Mediterranean restaurant.

Therefore, we think that the optimum place to put a new Mediterranean Restaurant is in Staten Island as there are 63 Neighborhoods in the area but no Mediterranean Restaurants, hence, eliminating any competition. The second-best Borough is the Bronx, having 70 neighborhoods in the area with little to no Mediterranean Restaurants gives a good opportunity for opening a new restaurant.

Finally, note that some of the drawbacks of this analysis are that the clustering is completely based on data obtained from the Foursquare API. Also, the analysis does not take into consideration of additional features for example menu cost, menu price average of other restaurants, neighborhood traffic, etc...

Conclusion

This concludes the optimal findings for our Capstone project and recommendation to our Greek restaurateur entrepreneur to open the next healthy Mediterranean restaurant in New York City. We started with a problem statement that required us to search for the right dataset to do our analysis and apply machine learning algorithms to recommend to our Greek restauranteur entrepreneur the best location to open the next healthy Mediterranean restaurant in New York City. Finally, we wrapped the findings with our results evaluation and final recommendation based on the data we have and assumptions made for the project.