Selecting a Location for a New Business

Coursera Capstone

Jackie Ye

Introduction

Selecting an optimal location for a new business is very crucial for its survival. Assuming a client wanted to open a new coffee shop in New York City, what neighborhood would you put it in? In this project, I will obtain coffee shop locations in New York City and analyze what area to open a coffee shop based on the min and max of the coffee shops in the 5 boroughs.

Data

Data was extracted and processed from various sources.

First, I obtained the neighborhood and ZIP Code data from health.ny.gov website using BeautifulSoup and wget. I then obtained the the ZIP Code's general latitude and longitude and created the ZIP Code's boundaries with estimated Northeast and Southwest latitude and longitude using Geocoding API from Google. I created a GeoJSON like list with this data for plotting purposes. Lastly, I obtained existing coffee shop data within 1-mile radius of a ZIP Code using Foursquare API calls. I processed this data by removing any duplicate stores that may have popped up due to errors and/or overlapping into another search with a different ZIP Code.

Methodology

My analysis consisted of two main focuses: statistical and visual.

Statistical

Using the Pandas and Numpy library, I was able to scrub and combine data into a Pandas DataFrame where I was able to isolate counts of coffee shops per neighborhood. Using this DataFrame, I was easily able to obtain the min and max of the neighborhoods to find most and least competitive areas (See chart below).

Neighborhood	Total
West Queens	132

Southeast Queens	131
Northwest Brooklyn	107
Upper West Side	104
Lower Manhattan	103
Lower East Side	103
Flatbush	98
Inwood and Washington Heights	90
Northwest Queens	90
Bushwick and Williamsburg	81
Sunset Park	79
Southwest Queens	71
Upper East Side	68
North Queens	66
Gramercy Park and Murray Hill	61
Chelsea and Clinton	53
Southern Brooklyn	53
Greenpoint	53
Stapleton and St. George	50
Southeast Bronx	49
West Central Queens	46
Greenwich Village and Soho	44
Northeast Bronx	41
Borough Park	39
East Harlem	39
Rockaways	36
South Shore	35

Central Brooklyn	26
Port Richmond	26
Southwest Brooklyn	16
Central Harlem	12
Hunts Point and Mott Haven	10
Mid-Island	8
Kingsbridge and Riverdale	8
Canarsie and Flatlands	7
Jamaica	3
Bronx Park and Fordham	3
Central Queens	2
East New York and New Lots	1
Northeast Queens	1
High Bridge and Morrisania	1

Using this DataFrame, I combined it with the data obtained from Foursquare API for Sklearns' Cluster Module for K-Means Machine Learning Algorithm. From the elbow curve (shown below), the number of clusters where it peaks off is at n=3. However, I decided not to use n=3 due to the fact that I am analyzing 5 boroughs of New York City and using 3 clusters in my opinion would not be able to accurately represent the neighborhoods.

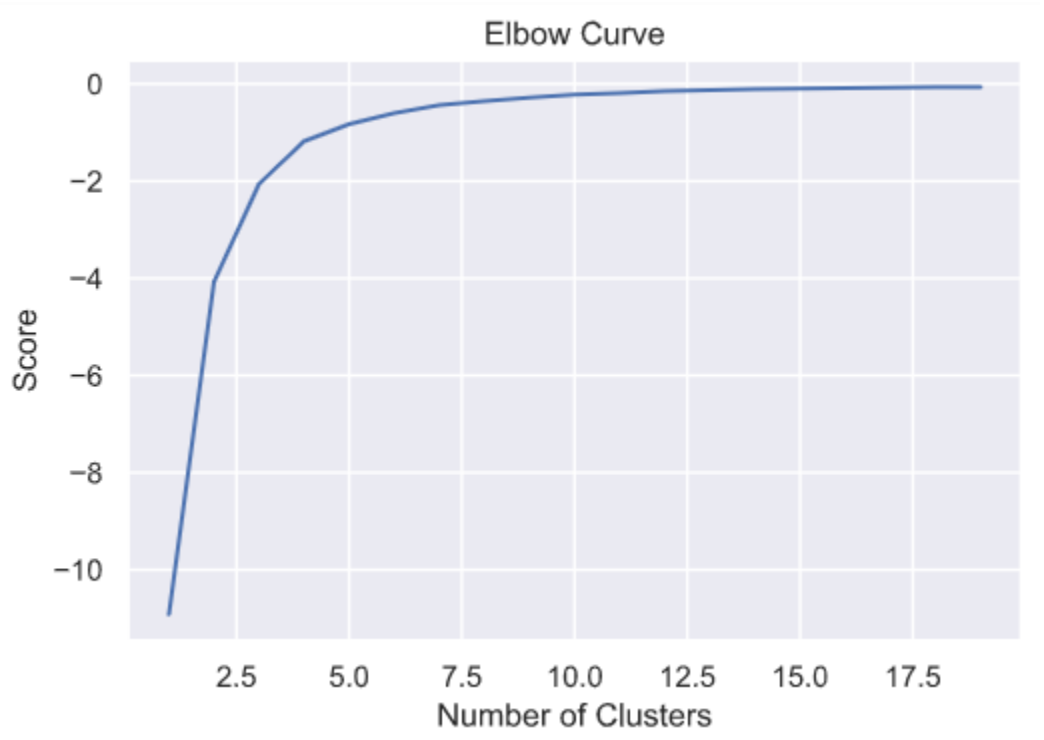

Figure 1 - Elbow Curve

Ignoring this curve, I opted to use n=20 to be able to separate more of the boroughs. After running the K-Means algorithm, the result is shown below in a seaborns plot.

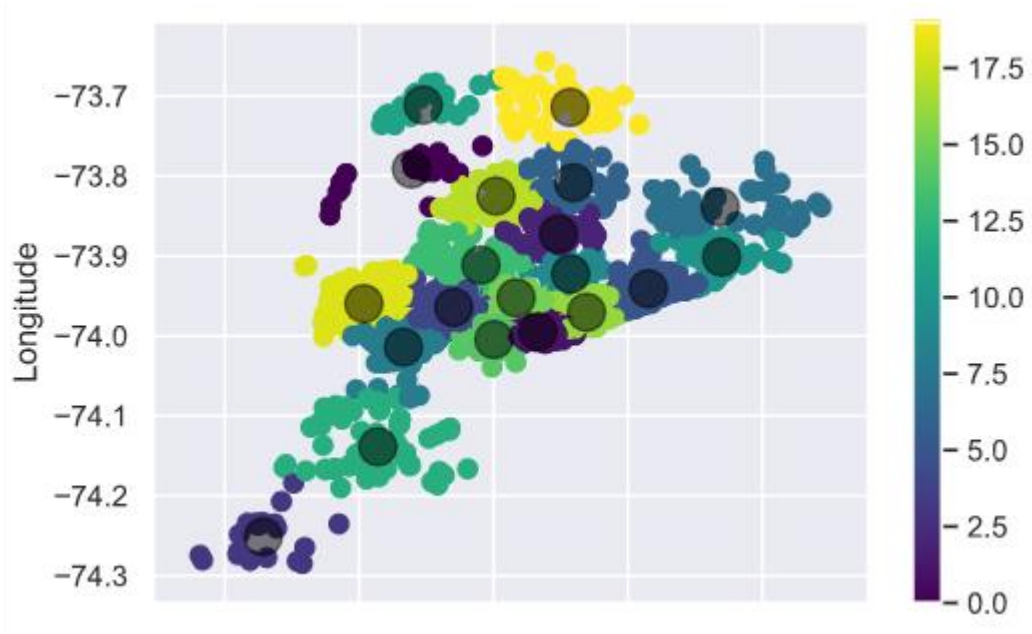

Figure 2 - Seaborns Plot of K-Means

Using this data, I was able to obtain the max and min count for the 20 clusters (shown below).

cluster_label	Total
1	217
16	202
4	166
5	159
14	141
8	140
9	134
18	114
13	103
15	97
6	89
2	86
12	71
10	64
17	58
7	56
19	53
0	43
11	31
3	22

Visual

For visualization, I used Folium and Leaflets to plot the data.

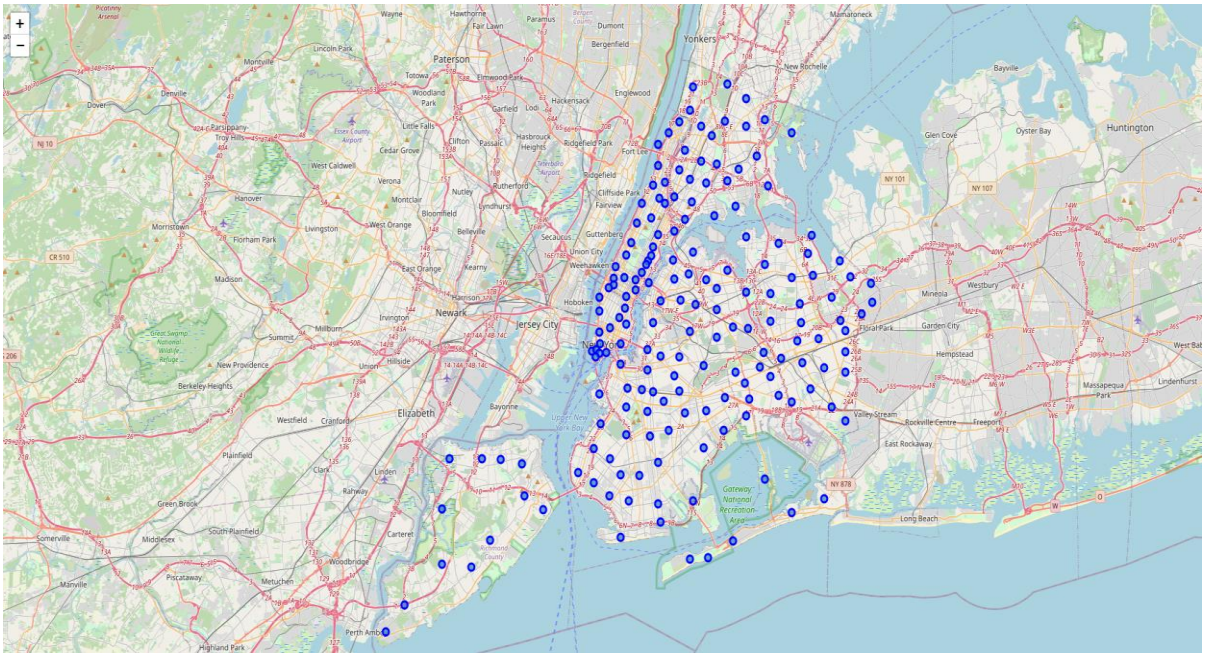

Figure 3 - Latitude and Longitude of NYC ZIP Codes

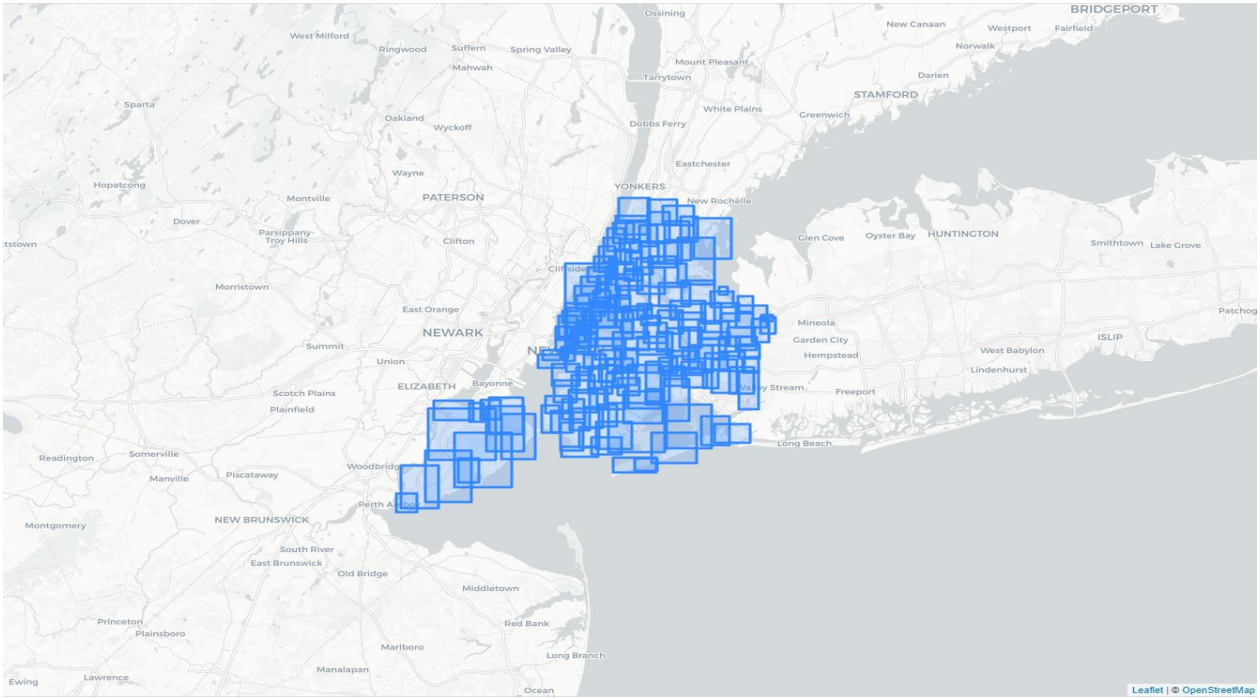

Figure 4 - Boundaries of NYC ZIP Codes

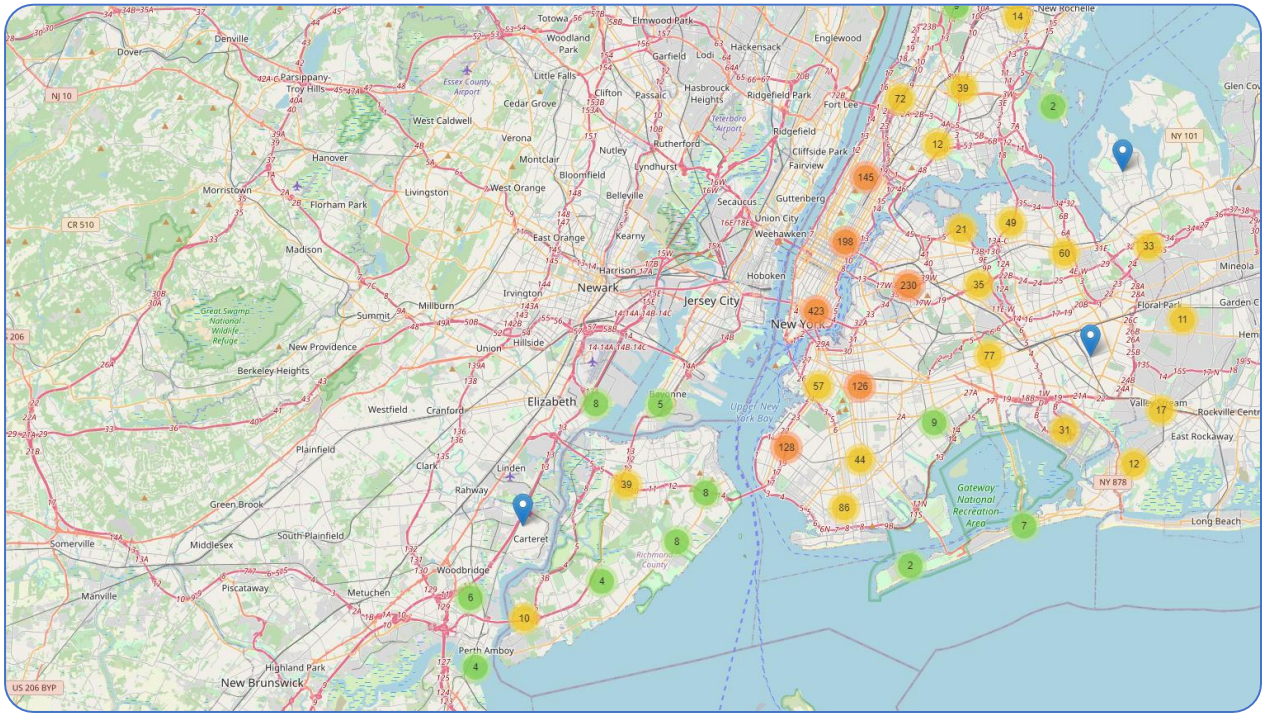

Figure 5 - Coffee Shop Locations in NYC

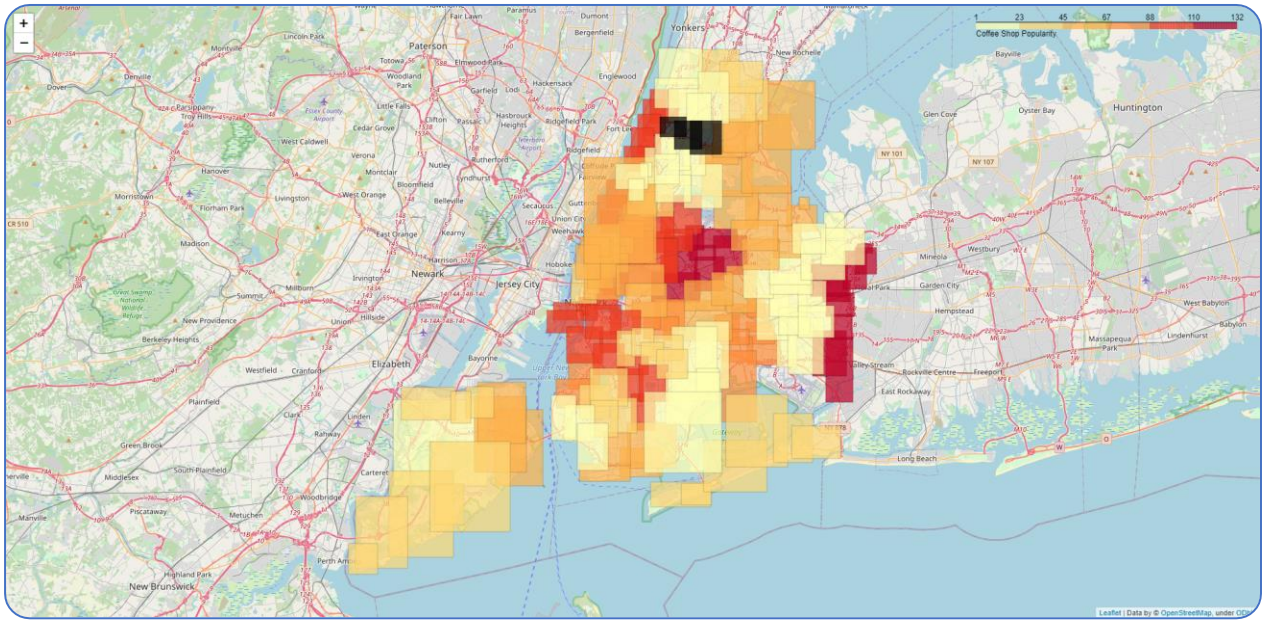

Figure 6 - Heat Map of NYC Coffee Shop

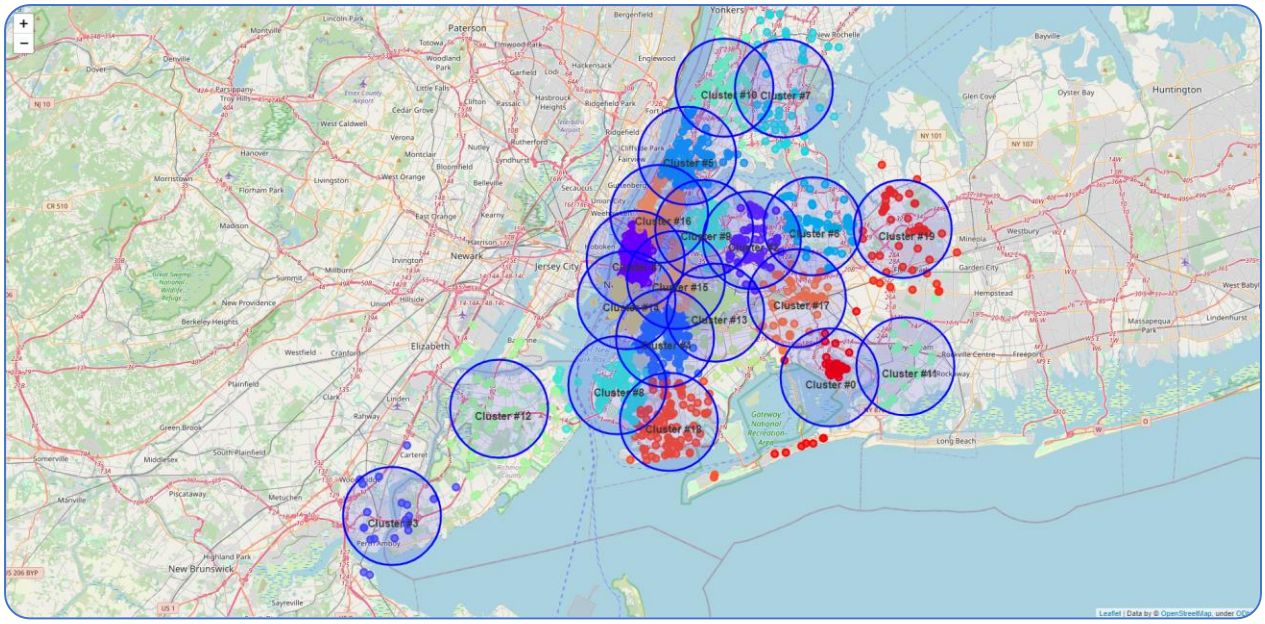

Figure 7 - Rendering of Clusters for Coffee Shops in NYC

Results

The results are as follows:

- West Queens (132) and Southeast Queens (131) had the most coffee shops.
- High Bridge and Morrisania (1), East New York and New Lots (1), and Northeast Queens (1) had the least coffee shops.
- Cluster #1 (217) had the most coffee shops.
- Cluster #3 (22) had the least coffee shops.

Discussion

I feel that the best optimal location for a coffee shop would be in Cluster #1 or Cluster #9. Although Cluster #0 had the fewest coffee shops, I believe the area is not suitable for development since it is near an airport.

For the elbow plot, I feel that if I used more input parameters, I would have been able to display a more optimized cluster parameter, n. More input parameters, such as more

accurate ZIP code boundaries, customer survey data, median income by neighborhoods, and crime rate data, would help refine the data and results that were obtained. We can then use this to score the area as well and provide a better recommendation for the client.

Conclusion

For this Cousera Capstone project I was able to combine both machine learning and data visualization to understand more about Data Science. I learned that with limitted inputs in data, it can result in a limitted output. It is important to gather as much data as possible to refine your results, however due to limited resources I was not able to do so. I would like to thank IBM for creating this Coursera course and introducing me to the world of Data Science.