Capstone Project Report

Introduction/Business problem

New York City attracts almost one-third of all foreign visitors to the United States. It attracts almost 47 million foreign and American tourists each year. For tourists, the most fun activity is to eat and finding the right place to eat can be a challenge. Many restaurants aim to attract most tourists to their restaurants so if someone is looking to open a restaurant in NYC, the location plays a vital role. Therefore, the purpose of this project is to determine which neighborhood of New York is best for opening a particular type of restaurant.

Description of the data

Neighborhood Data

New York City has a total of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood. Luckily, this dataset exists for free on the web. Here is the link to the dataset: https://geo.nyu.edu/catalog/nyu_2451_34572

Foursquare API Data

As we need data about different venues in different neighborhoods therefore after finding the list of neighborhoods, we then connect to the Foursquare API to gather information about venues inside each and every neighborhood. For each neighborhood, we have chosen the radius to be 100 meter thus we will use the Foursquare API to explore neighborhoods in New York City. We will use the explore function to get the most common venue categories in each neighborhood.

Methodology

In this section, I will describe the data analysis and how I used the data to yield the results. The table of contents are,

1. Installing and Importing Python Libraries and Dependencies
2. Download and Explore Dataset
3. Map of Brooklyn
4. Explore Neighborhoods in Brooklyn
5. Analyze Each Neighborhood
6. Cluster Neighborhoods
7. Examine Clusters

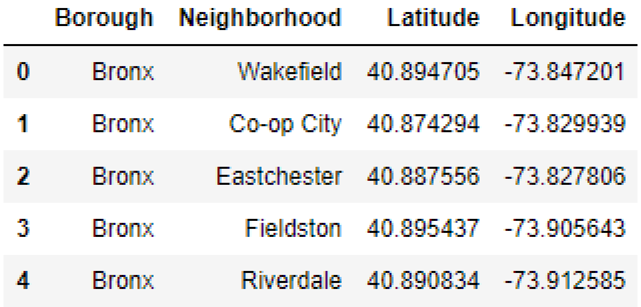
1. Installing and Importing Python Libraries and Dependencies

I started by importing all the libraries and dependencies which I would use in my project.
The three most important libraries were Numpy, Pandas and Folium

.

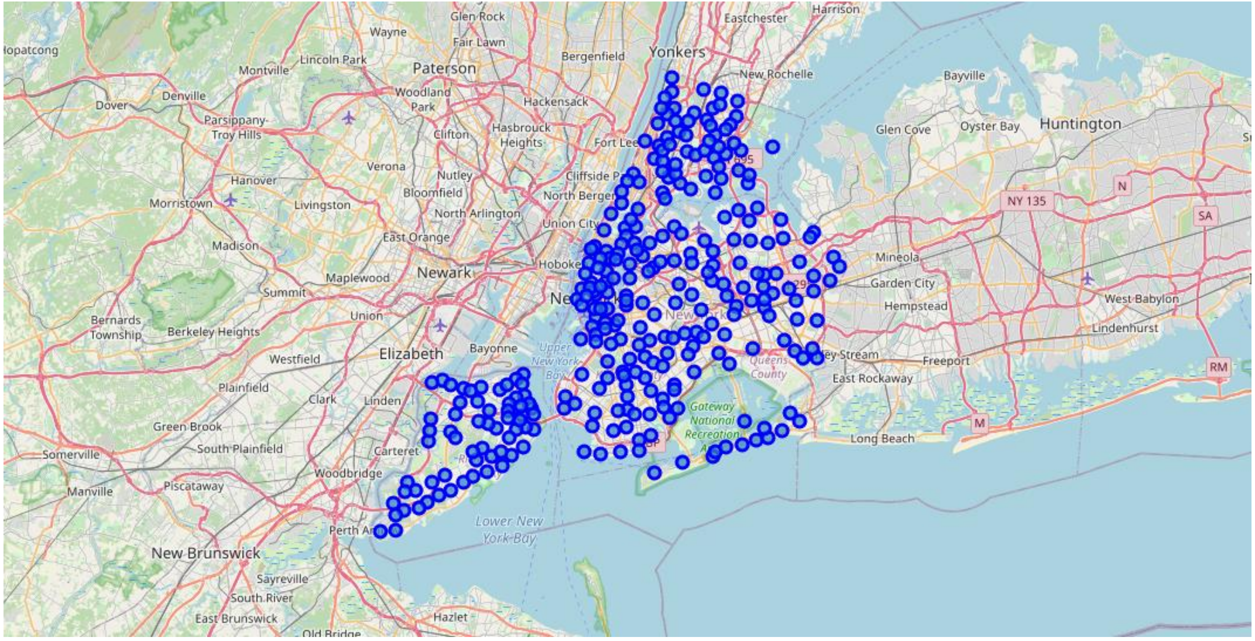
2. Download and Explore Dataset

After this, I scraped data from Wikipedia to create a data frame with the neighborhoods of New York, For this, I used the pandas read function. I had to clean the resulting data frame in terms of unnecessary information or data that could not be handled in a data frame, such as picture data of the coat of arms of each district. The result is a nice data frame:



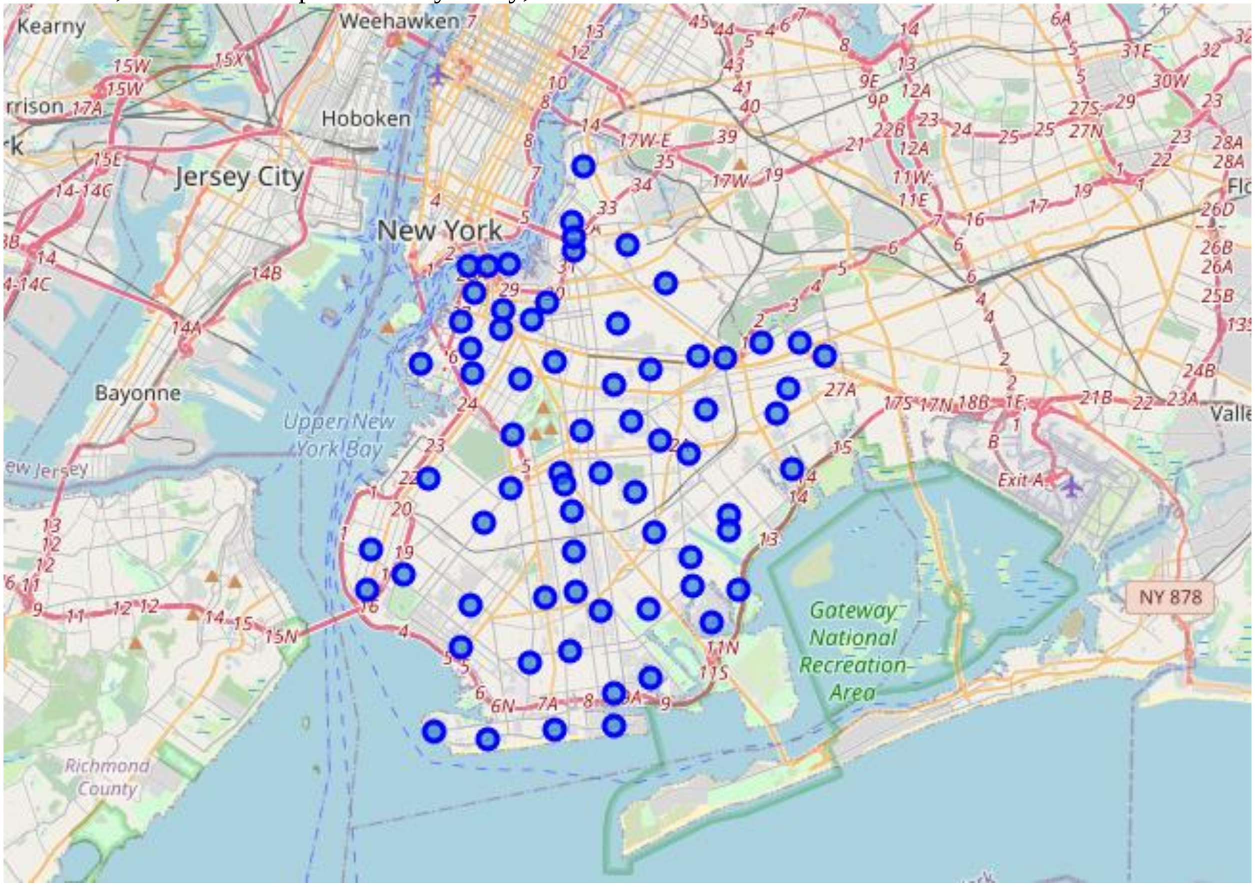
	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Using the folium package and my data frame, I then created a map with all neighborhoods of NY,



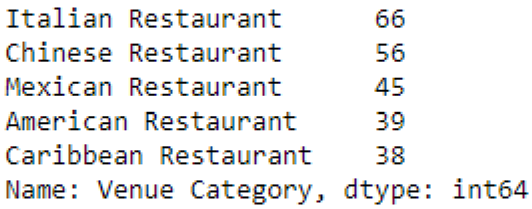
3. Map of Brooklyn

After this, I created a Map of Brooklyn only,



4. Explore Neighborhoods in Brooklyn

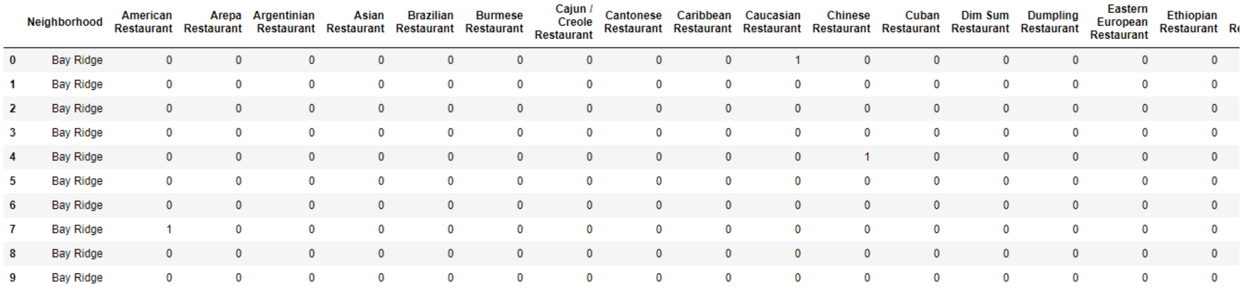
Then, I retrieved the foursquare data for all venues on foursquare with a distance of less than 500 meters from each center of each neighborhood, as indicated as blue dots in the map above. There are 624 restaurants in Brooklyn with 58 unique categories. such as Italian, Chinese, American etc.



```
Italian Restaurant      66
Chinese Restaurant      56
Mexican Restaurant      45
American Restaurant     39
Caribbean Restaurant    38
Name: Venue Category, dtype: int64
```

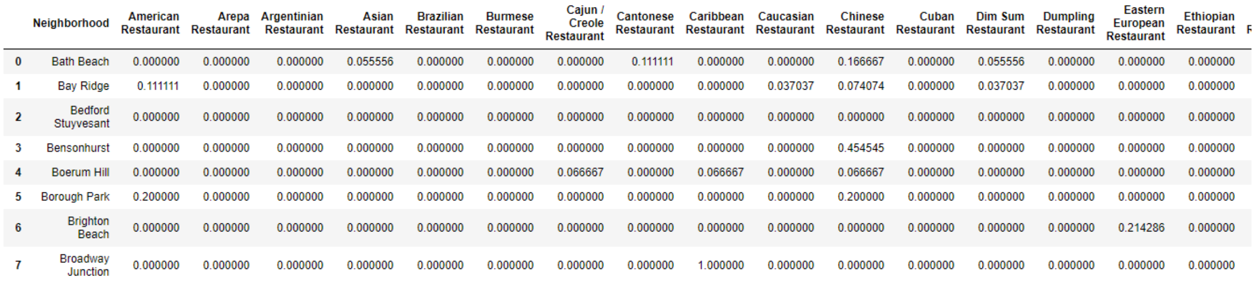
5. Analyze Neighborhoods

To find clusters of restaurant types in the different city districts, I first transformed the data frame with the restaurant venues, associated to city districts, by one-hot encoding (0/1), as seen in the picture below.



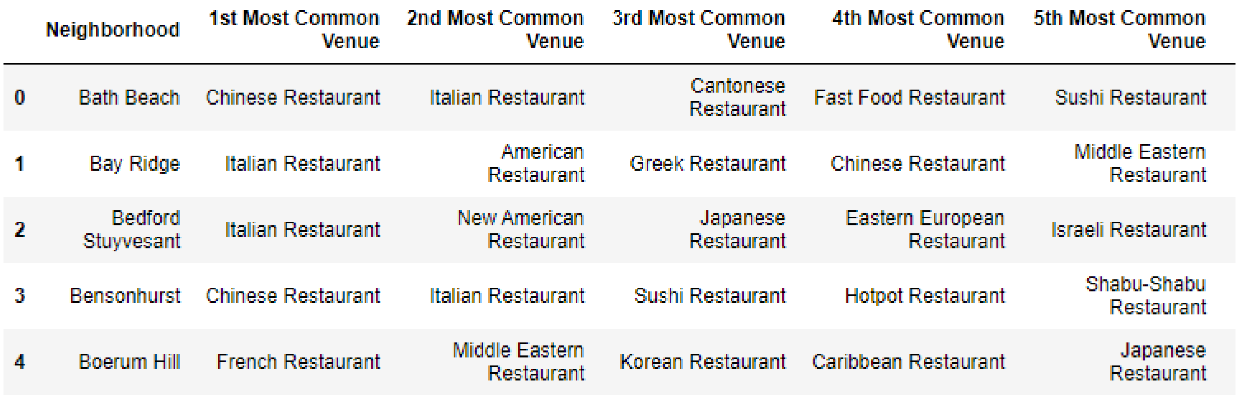
	Neighborhood	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Brazilian Restaurant	Burmese Restaurant	Cajun / Creole Restaurant	Cantonese Restaurant	Caribbean Restaurant	Caucasian Restaurant	Chinese Restaurant	Cuban Restaurant	Dim Sum Restaurant	Dumpling Restaurant	Eastern European Restaurant	Ethiopian Restaurant
0	Bay Ridge	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1	Bay Ridge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Bay Ridge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Bay Ridge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Bay Ridge	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
5	Bay Ridge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	Bay Ridge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	Bay Ridge	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	Bay Ridge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	Bay Ridge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Next, I used grouping to show the frequency of each category of restaurants in each city district.



	Neighborhood	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Brazilian Restaurant	Burmese Restaurant	Cajun / Creole Restaurant	Cantonese Restaurant	Caribbean Restaurant	Caucasian Restaurant	Chinese Restaurant	Cuban Restaurant	Dim Sum Restaurant	Dumpling Restaurant	Eastern European Restaurant	Ethiopian Restaurant
0	Bath Beach	0.000000	0.000000	0.000000	0.055556	0.000000	0.000000	0.000000	0.111111	0.000000	0.000000	0.166667	0.000000	0.055556	0.000000	0.000000	0.000000
1	Bay Ridge	0.111111	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.037037	0.074074	0.000000	0.037037	0.000000	0.000000	0.000000
2	Bedford Stuyvesant	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3	Bensonhurst	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.454545	0.000000	0.000000	0.000000	0.000000	0.000000
4	Boerum Hill	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.066667	0.000000	0.066667	0.000000	0.066667	0.000000	0.000000	0.000000	0.000000	0.000000
5	Borough Park	0.200000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.200000	0.000000	0.000000	0.000000	0.000000	0.000000
6	Brighton Beach	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.214286	0.000000
7	Broadway Junction	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

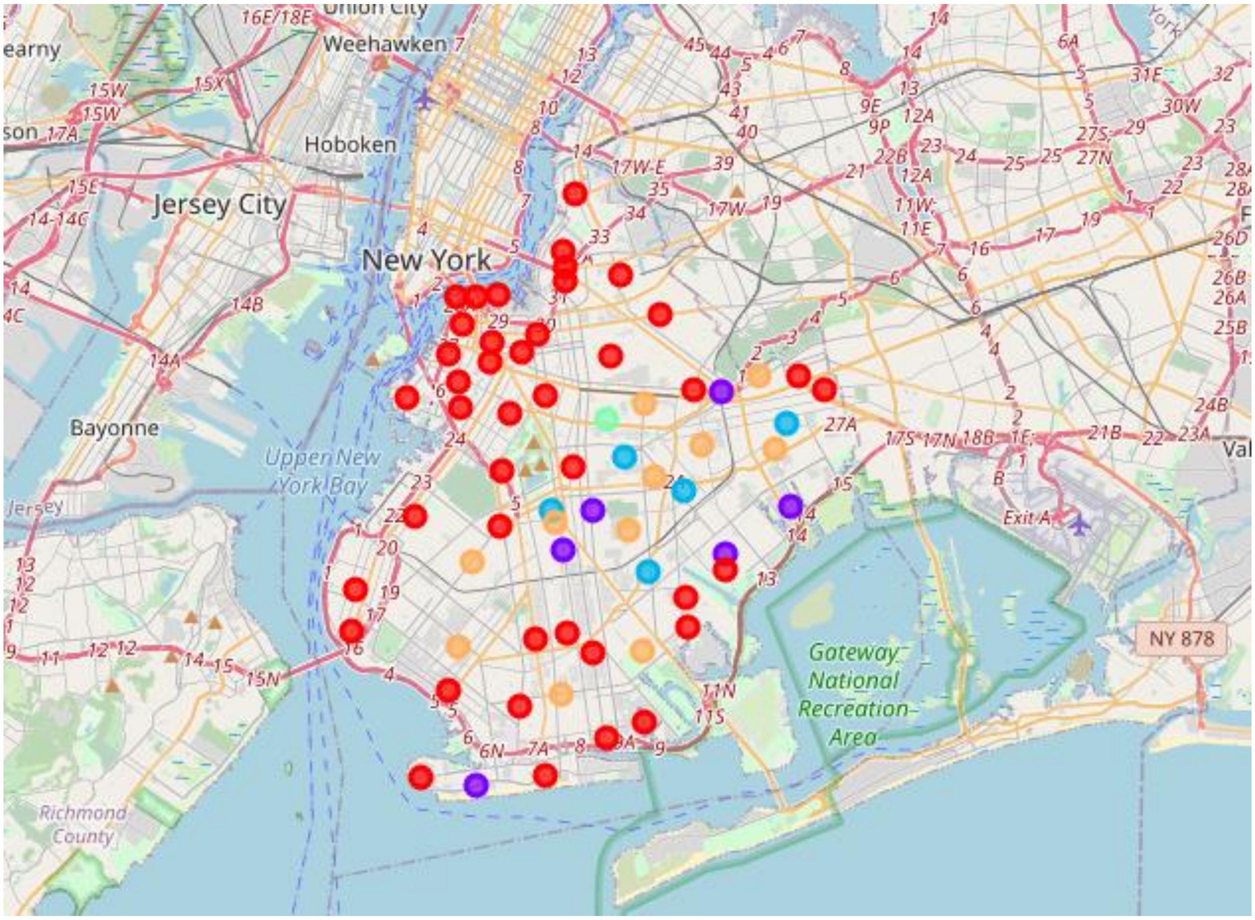
I used this information to create a data frame in which you can see the most common restaurant venue types for each city district.



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Bath Beach	Chinese Restaurant	Italian Restaurant	Cantonese Restaurant	Fast Food Restaurant	Sushi Restaurant
1	Bay Ridge	Italian Restaurant	American Restaurant	Greek Restaurant	Chinese Restaurant	Middle Eastern Restaurant
2	Bedford Stuyvesant	Italian Restaurant	New American Restaurant	Japanese Restaurant	Eastern European Restaurant	Israeli Restaurant
3	Bensonhurst	Chinese Restaurant	Italian Restaurant	Sushi Restaurant	Hotpot Restaurant	Shabu-Shabu Restaurant
4	Boerum Hill	French Restaurant	Middle Eastern Restaurant	Korean Restaurant	Caribbean Restaurant	Japanese Restaurant

6. Cluster Neighborhoods

Now, with all this data, I could finally run an unsupervised machine learning algorithm, more specifically, a k-means clustering algorithm from the scikit-learn package. One could use the elbow method to systematically define the k value, but I simply chose k to be 5, having been inspired by one of the coursera courses to do so.



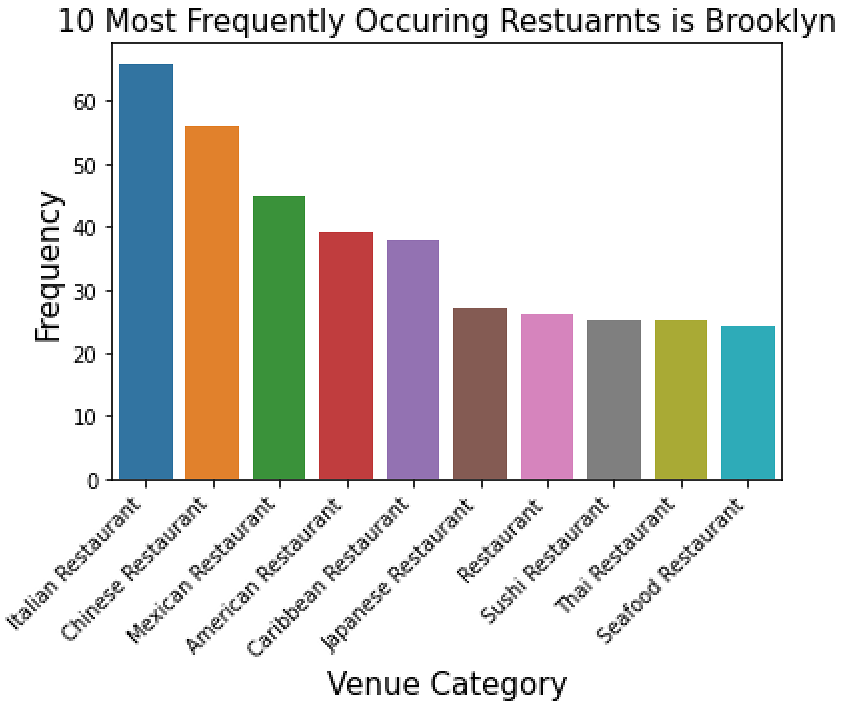
7. Examine Clusters

Now, we can examine each cluster and determine the discriminating venue categories that distinguish each cluster. What we see in the table are the neighborhoods and their most common venues, and they now have been assigned five different cluster labels from 0 to 4.

Cluster 1 could be called Italian Cluster.
Cluster 2 could be called American Caribbean and Yemeni Cluster.
Cluster 3 could be called Caribbean Fast Food Cluster.
Cluster 4 could be called Diverse Cluster.
Cluster 5 could be called Chinese Asian Cluster.

Conclusion

There are more than 60 Italian Restaurants in Brooklyn which means that most people love to eat Italian but this category will also have the most competition after this category we also have Chinese and Mexican Restaurants. We need to find those clusters which have lowest frequency of these categories so we will have less competition and more customers due to unique category. Thus, Cluster 1 is best for opening a Caribbean Restaurant. Cluster 2 is best for opening a Chinese Restaurant. Cluster 5 is best for opening an Italian Restaurant.


10 Most Frequently Occuring Restuarnts is Brooklyn
Frequency
Italian Restaurant
Chinese Restaurant
Mexican Restaurant
American Restaurant
Caribbean Restaurant
Japanese Restaurant
Restaurant
Sushi Restaurant
Thai Restaurant
Seafood Restaurant
Venue Category