Finding the Best Neighborhood in NYC to Open a Korean Restaurant

Soomin Kim
June 27th, 2021

1. Introduction

1.1 Background

New York City is the most populated city in the United States. It is also one of the most ethnically, linguistically, and culturally diverse cities in the world. About 36% of the city's population are foreign-born; 800 languages are spoken in the city; dozens of ethnic enclaves have developed throughout its neighborhoods - from Little Guyana, Koreatown, Little India, Little Odessa, Chinatown, Little Australia, Little Poland to Little Italy.

More specifically, the City is home to more than 100,000 ethnic Koreans, which contributes to the New York metropolitan area containing the second-largest population of Koreans outside of Korea. Given its dense Korean population and a growing interest in K-food business worldwide, there are ample opportunities to expand Korean restaurant business in New York City.

1.2 Aim

This project aims to determine the best location (borough, neighborhood) in New York City to open a Korean restaurant. My criteria for choosing the best location is one that is not already crowded with Korean Restaurants and is as close to the city center as possible.

1.3 Target Audience

This project will provide useful insight for anyone - entrepreneurs, business owners, investors - who are looking to start or expand their Korean restaurant business in New York City.

2. Data Acquisition and Cleaning

2.1 Data Sources:

Data 1: New York City Neighborhood Data:

I collected the New York City Neighborhoods Data from https://cocl.us/new_york_dataset. This data contains a total of 5 boroughs and 306 neighborhoods as well as the latitude and longitude coordinates of each neighborhood in New York City.

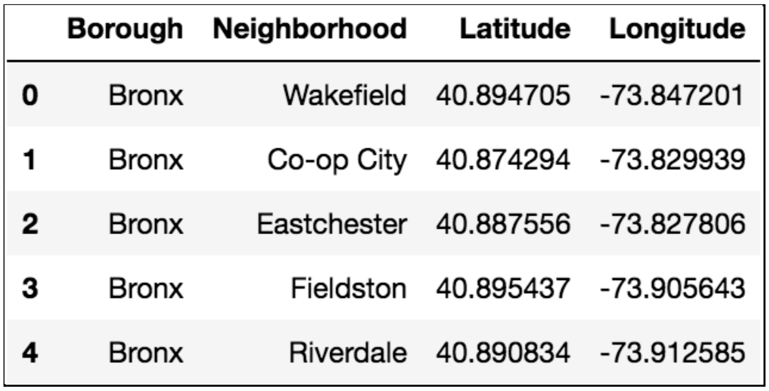
Data 2: Foursquare Location Data:

The Foursquare API provides comprehensive location data, including data based on location, keyword, user, amongst others. I used the Foursquare API to get all the venues in the neighborhoods of New York City.

2.2 Data Cleaning

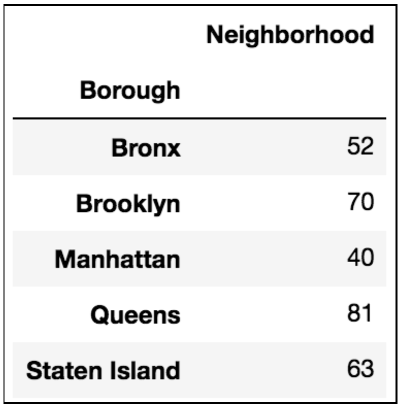
Data 1:

After opening the JSON file from https://cocl.us/new_york_dataset, I discovered that all the relevant data - borough, neighborhood, latitude and longitude coordinates of each neighborhood - were stored in the *features* key. I loaded these features into a pandas dataframe *neighborhoods* (Table 1), and confirmed that this new dataframe had 5 boroughs and 306 neighborhoods (Table 2).



	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

Table 1: first five rows of *neighborhoods* dataframe that contains NYC's borough, neighborhood, latitude, longitude values

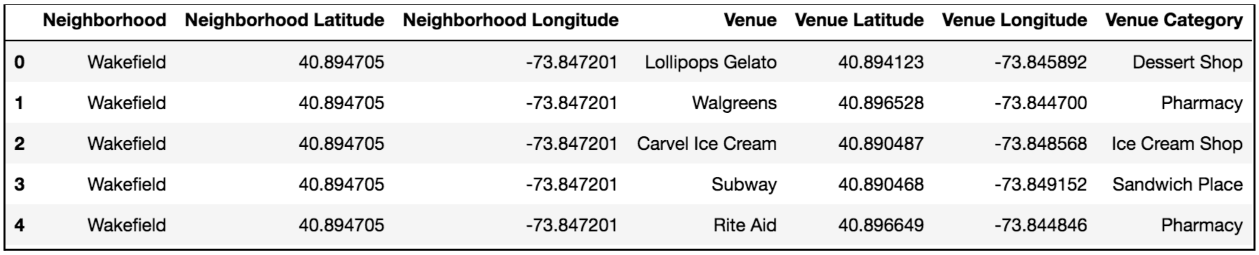


	Neighborhood
Borough	
Bronx	52
Brooklyn	70
Manhattan	40
Queens	81
Staten Island	63

Table 2: neighborhood count for each of 5 boroughs in NYC

Data 2:

I utilized the Foursquare API to retrieve information about the venue, venue category, longitudes and latitudes. First, I defined the Foursquare credentials and version. Then using this information and *neighborhoods*, I made a *getNearbyVenues* call that returns a JSON containing all the relevant information for top 100 venues that are within a radius of 500 meters from NYC. Then, I converted this JSON file into a dataframe named *nyc_venues*. Table 3 shows the first five rows of the resulting dataframe.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
2	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
3	Wakefield	40.894705	-73.847201	Subway	40.890468	-73.849152	Sandwich Place
4	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy

Table 3: first five rows of *nyc_venues* dataframe that includes venues info pulled from Foursquare API

3. Exploratory Data Analysis

3.1 Visualization of the map of New York City by Neighborhood

Using the folium package, I visualized a map of New York with neighborhoods superimposed on top on a leaflet map. More specifically, the CircleMarker() function was used to color-code boroughs, providing a clearer understanding of where each borough is located.

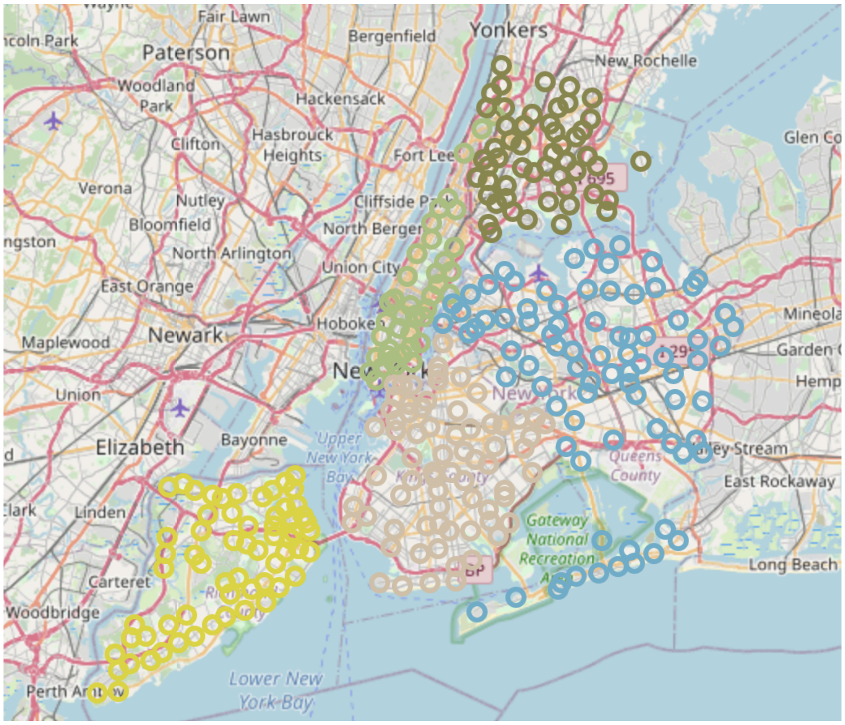


Figure 1: Map of NYC color-coded by neighborhood

3.2 Number of Unique venues and Korean Restaurants in New York City.

After grouping *nyc_venues* by neighborhoods, I found that there are 434 unique venue categories, including Korean Restaurants. And there are 55 Korean Restaurants in New York.

3.3 Top 10 New York Neighborhoods with the Highest Number of Korean Restaurants

Then I conducted a brief data visualization to discover where these 55 Korean Restaurants were mostly located. The Top 10 New York neighborhoods with the highest number of Korean Restaurants are shown in the bar graph in Figure 2. Results showed that Murray Hill has the highest number (15), closely followed by Midtown South (14). There was a significant dip in number from the third highest neighborhood, Oakland Gardens, which has 4. This was followed by East Village and Flushing, each with 3.

Sunnyside Gardens and Little Neck each has 2, and the last three - Utopia, Park Slope, Civic Center - each has 1 Korean Restaurant.

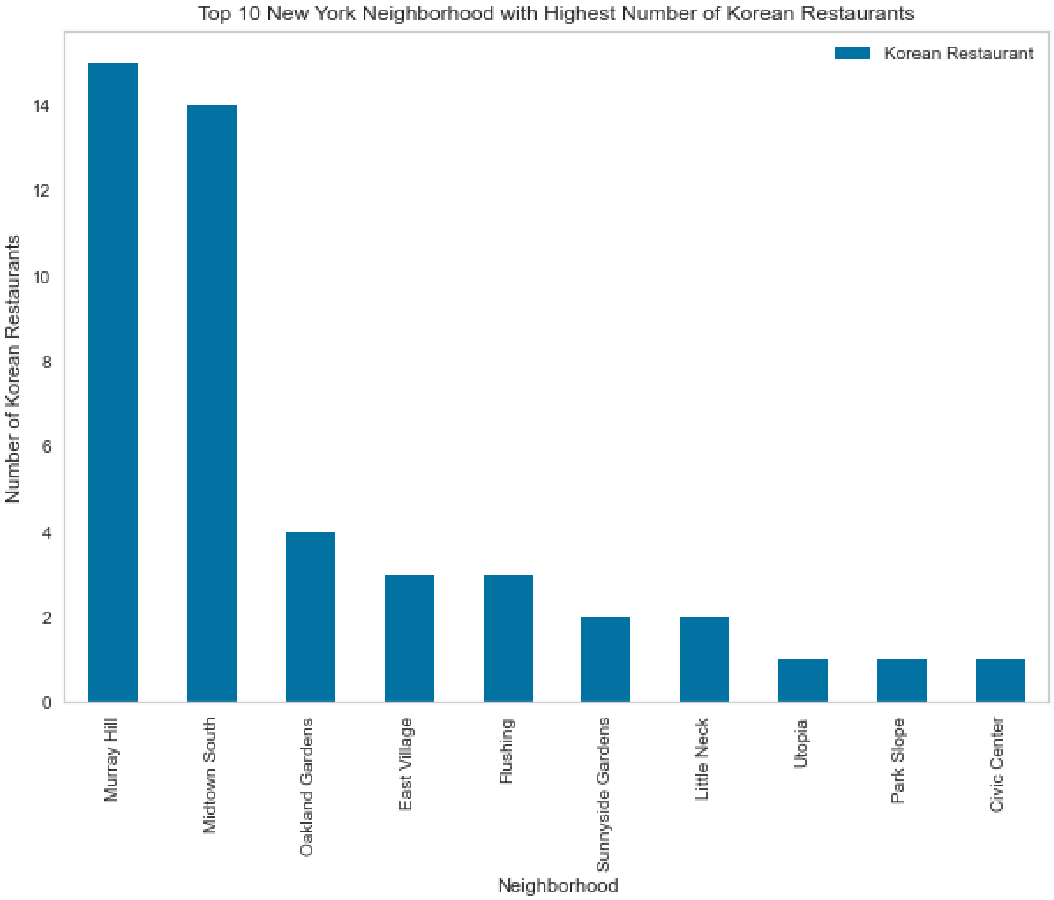
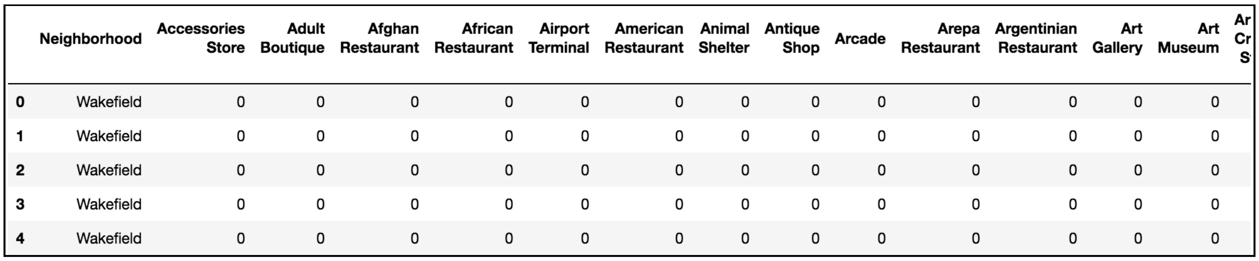


Figure 2: Murray Hill has the highest number of Korean Restaurants, closely followed by Midtown South. There is a significant drop in the number of Korean Restaurants from the third highest neighborhood as shown in this Top 10 figure.

4 Machine Learning

4.1 One Hot Encoding

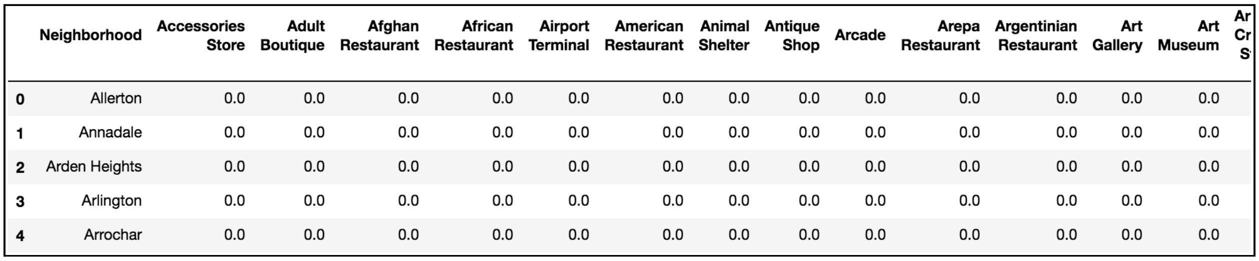
To divide New York City into clusters of neighborhoods with a similar number of Korean Restaurants, I first needed to transform the categorical variables within *nyc_venues* to numerical ones (*nyc_onehot)*. To do so, I used the One Hot Encoding technique, using the pd.get_dummies function within the Pandas package.



	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Terminal	American Restaurant	Animal Shelter	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Ar Cr S
0	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0	0	

Table 4: first five rows of *nyc_onehot*, a dataframe that underwent One Hot Encoding Technique which transformed all categorical variables into numerical ones

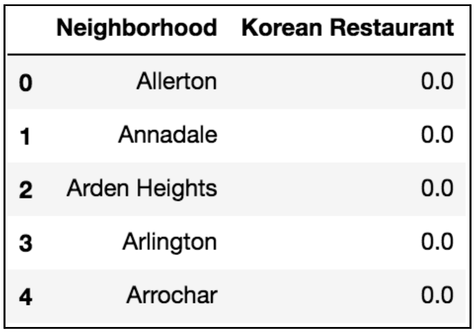
Then, I calculated the mean of all venue categories by neighborhoods to analyze the data by neighborhood.



	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Terminal	American Restaurant	Animal Shelter	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Ar Cr S
0	Allerton	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	Annadale	0.0	0.0	0.0	0.0	0.0	0.0	0.0	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	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3	Arlington	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	Arrochar	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Table 5: first five rows of *nyc_grouped*, a dataframe containing mean values of all venues by neighborhood.

After this, I extracted Neighborhood and Korean Restaurant columns into a dataframe *korean* for further analysis.



	Neighborhood	Korean Restaurant
0	Allerton	0.0
1	Annadale	0.0
2	Arden Heights	0.0
3	Arlington	0.0
4	Arrochar	0.0

Table 6: first five rows of *korean*, a dataframe containing mean number of Korean Restaurants by neighborhood

4.2 K-Means Clustering

Using *korean*, I used the K-Means clustering method to cluster the neighborhoods based on neighborhoods with similar mean frequencies of Korean Restaurants.

First, using the KElbowVisualizer, I identified the optimal number of clusters (best K value) by fitting the model with a range of K values from 1 to 10. The resulting line graph resembled an arm with an "elbow" (the point of inflection on the curve), annotated with a dashed line, which is where the model fits best. In the graph below, the line annotating the "elbow" is at k = 3. This meant that the optimum number of clusters is 3.

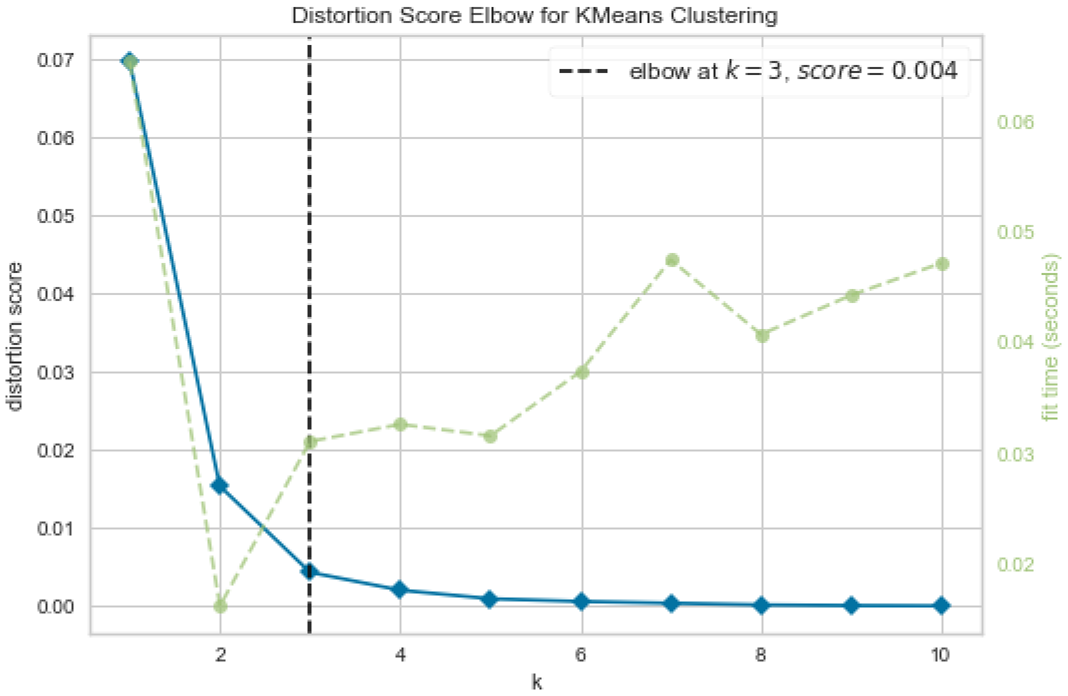
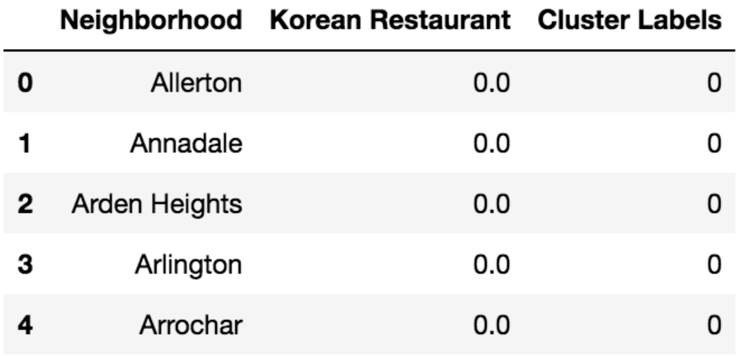


Figure 3: KElbowVisualizer revealed an elbow at k=3, with a score of 0.004, indicating that the optimal number of clusters is 3.

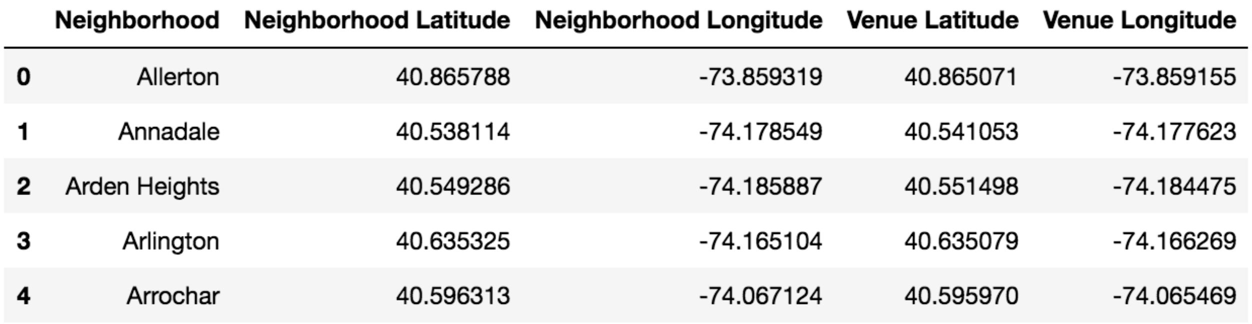
I then ran K-Means clustering method that created 3 clusters of neighborhoods based on the similar mean frequencies of korean Restaurants. Each cluster was labelled as 0, 1, or 2. These were compiled to a new column, ‘Cluster Labels’, and merged with the *korean* dataframe, resulting in *kor_merged* dataframe shown below:



	Neighborhood	Korean Restaurant	Cluster Labels
0	Allerton	0.0	0
1	Annadale	0.0	0
2	Arden Heights	0.0	0
3	Arlington	0.0	0
4	Arrochar	0.0	0

Table 7: first five rows of *kor_merged* dataframe, containing mean frequency of Korean Restaurants and Cluster Label (0, 1, 2) for each neighborhood

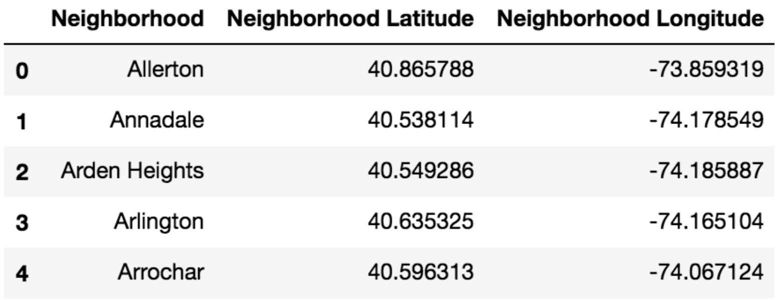
Before joining this cluster label and mean frequency of Korean Restaurant data with location data, I wanted to ensure that the latter was in good shape with only the necessary columns. Initially the *nyc_venues* data was as shown in Table 3. To ensure that we have one location data (longitude and latitude values) for each neighborhood, I created a new dataframe *nyc_venues_grouped* that had mean longitude and latitude values for each neighborhood as shown in Table 8 below:



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue Latitude	Venue Longitude
0	Allerton	40.865788	-73.859319	40.865071	-73.859155
1	Annadale	40.538114	-74.178549	40.541053	-74.177623
2	Arden Heights	40.549286	-74.185887	40.551498	-74.184475
3	Arlington	40.635325	-74.165104	40.635079	-74.166269
4	Arrochar	40.596313	-74.067124	40.595970	-74.065469

Table 8: first five rows of *nyc_venues_grouped* dataframe, containing the average longitude and latitude values of each neighborhood

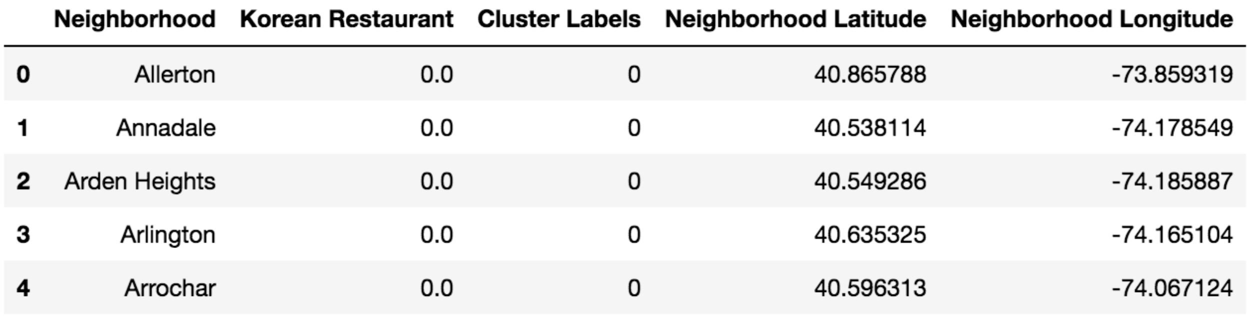
Realizing that I won't need the average venue location data anymore for neighborhood analysis, I dropped the last two columns, resulting in the dataframe below:



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude
0	Allerton	40.865788	-73.859319
1	Annadale	40.538114	-74.178549
2	Arden Heights	40.549286	-74.185887
3	Arlington	40.635325	-74.165104
4	Arrochar	40.596313	-74.067124

Table 9: first five rows of updated *nyc_venues_grouped* dataframe, without the dropped venue location data columns

Then, I joined *kor_merged* (cluster labelled data) with *nyc_venues_grouped* (neighborhood location data) to add latitude and longitude values for each neighborhood.

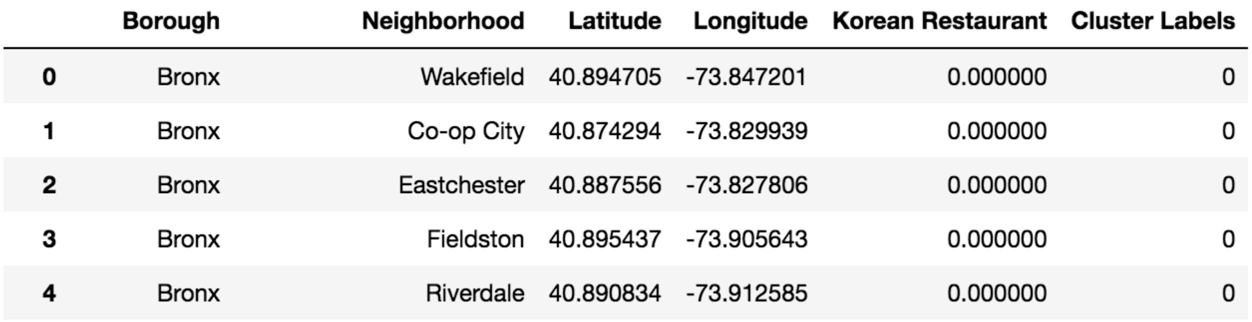


	Neighborhood	Korean Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude
0	Allerton	0.0	0	40.865788	-73.859319
1	Annadale	0.0	0	40.538114	-74.178549
2	Arden Heights	0.0	0	40.549286	-74.185887
3	Arlington	0.0	0	40.635325	-74.165104
4	Arrochar	0.0	0	40.596313	-74.067124

Table 10: first five rows of updated *kor_merged* dataframe with the addition of neighborhood latitude and longitude values for each neighborhood

While this dataframe has the neighborhood data, it doesn't have any boroughs information. This led to problems when analyzing same-named neighborhoods that are actually in different boroughs, leading to inaccurate analysis. In fact, upon evaluation, I realized that the *kor_merged* dataframe contained the same location information for neighborhoods with same names, despite them being of different boroughs. Hence, I realized the importance of integrating boroughs data with more accurate location data for each neighborhood.

As such, I joined the *kor_merged* with *neighborhoods* data that contains boroughs information. Then, I dropped the latitude and longitude columns of *kor_merged* data, which were inaccurate for same-named neighborhoods, and kept the location data obtained from *neighborhoods* data to ensure all neighborhoods, regardless of identical names, are in accurate representation of their location. This resulted in an updated *kor_merged* dataframe shown in Table 11 below:



	Borough	Neighborhood	Latitude	Longitude	Korean Restaurant	Cluster Labels
0	Bronx	Wakefield	40.894705	-73.847201	0.000000	0
1	Bronx	Co-op City	40.874294	-73.829939	0.000000	0
2	Bronx	Eastchester	40.887556	-73.827806	0.000000	0
3	Bronx	Fieldston	40.895437	-73.905643	0.000000	0
4	Bronx	Riverdale	40.890834	-73.912585	0.000000	0

Table 11: first five rows of final *kor_merged* dataframe, containing boroughs information and accurate latitude and longitude values for each neighborhood.

With this final *kor_merged* dataframe, I visualized the clusters by cluster labels to get a better understanding of where the clusters reside. To do so, I used the folium package to create a colored map displaying 3 clusters in NYC.

- Cluster 0: Red
- Cluster 1: Purple
- Cluster 2: Aquamarine

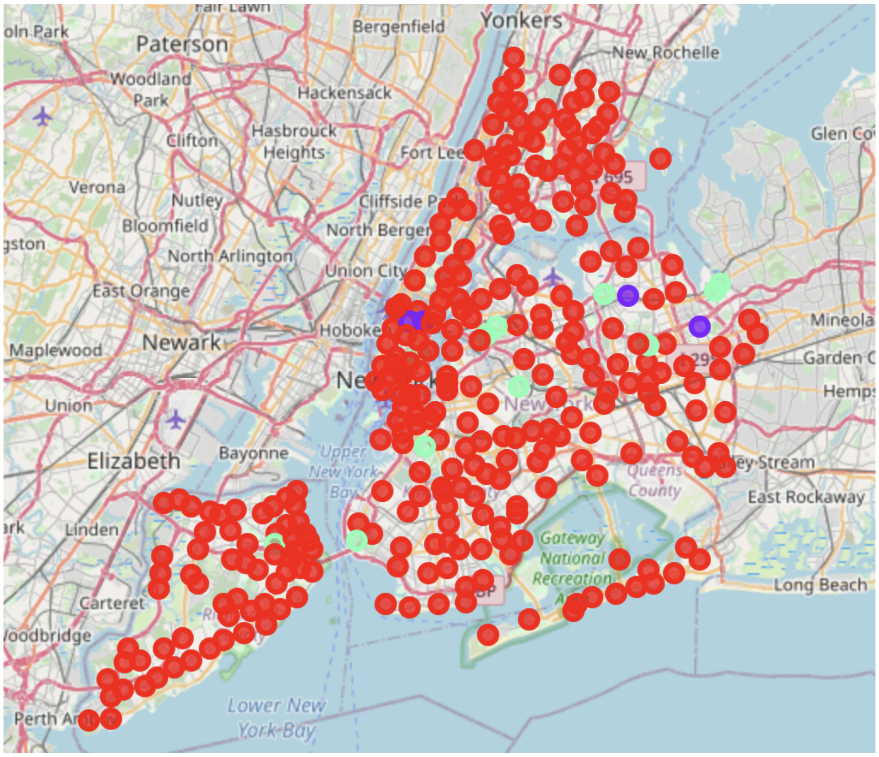
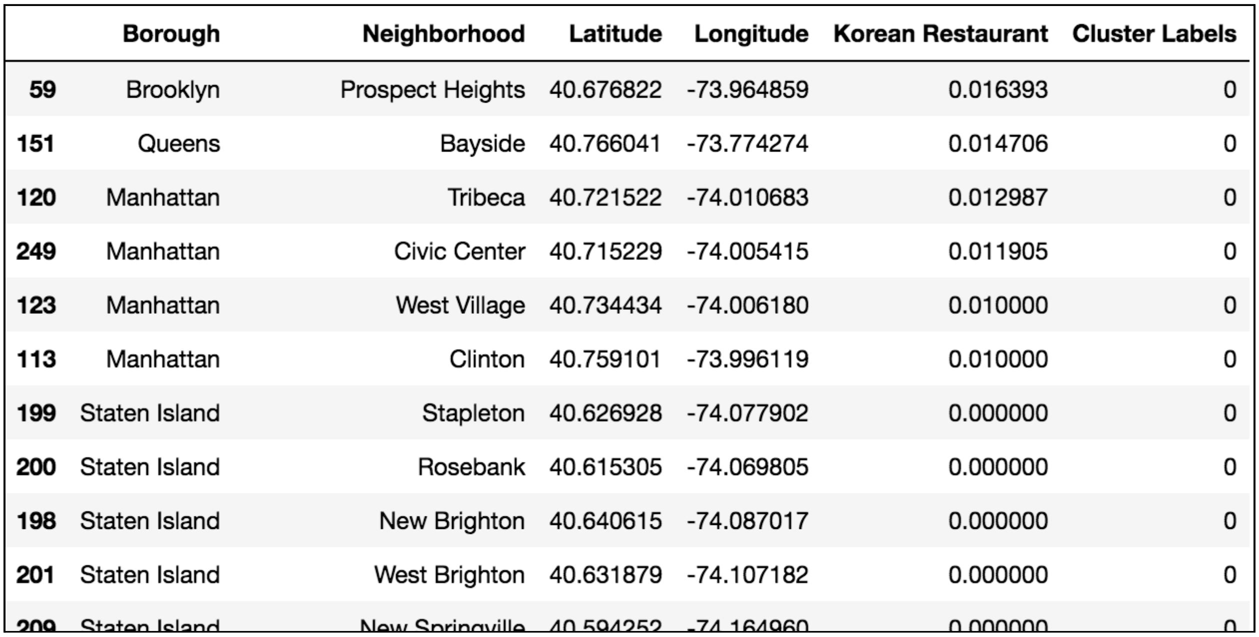


Figure 4: Colored-map showing location of 3 distinct clusters, generated by k-means clustering method based on similar Korean Restaurant mean frequency of occurrence

4.3 Examine the Clusters

Cluster 0 (Red)

Cluster 0 contains neighborhoods with the smallest density of Korean Restaurants, with a mean of 0.0003 (4dp) Korean Restaurant average occurrences across all selected neighborhoods in this cluster. These neighborhoods are all spread out across all five boroughs. Of the 291 neighborhoods in this cluster, however, there are only 6 neighborhoods with non-zero mean occurrence of Korean Restaurants. These are neighborhoods of Brooklyn, Queens, Manhattan. Consequently, the other two boroughs, Staten Island and Bronx, in this cluster had zero mean occurrences of Korean Restaurants.

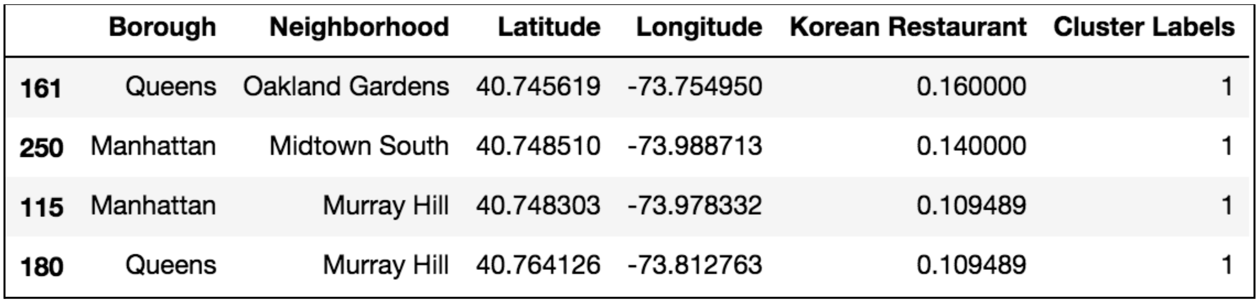


	Borough	Neighborhood	Latitude	Longitude	Korean Restaurant	Cluster Labels
59	Brooklyn	Prospect Heights	40.676822	-73.964859	0.016393	0
151	Queens	Bayside	40.766041	-73.774274	0.014706	0
120	Manhattan	Tribeca	40.721522	-74.010683	0.012987	0
249	Manhattan	Civic Center	40.715229	-74.005415	0.011905	0
123	Manhattan	West Village	40.734434	-74.006180	0.010000	0
113	Manhattan	Clinton	40.759101	-73.996119	0.010000	0
199	Staten Island	Stapleton	40.626928	-74.077902	0.000000	0
200	Staten Island	Rosebank	40.615305	-74.069805	0.000000	0
198	Staten Island	New Brighton	40.640615	-74.087017	0.000000	0
201	Staten Island	West Brighton	40.631879	-74.107182	0.000000	0
209	Staten Island	New Springville	40.594252	-74.164960	0.000000	0

Table 12: first 10 rows of 291, sorted by descending order of ‘Korean Restaurant’ column value, showing neighborhoods in Cluster 0 (red)

Cluster 1 (Purple)

Cluster 1 contains neighborhoods with the highest density of Korean Restaurants, with a mean of 0.1297 (4dp) Korean Restaurant average occurrences. There are a total of 4 neighborhoods in this cluster, all from either Queens or Manhattan. The neighborhood with the highest frequency of Korean Restaurant was Oakland Gardens of Queens with 0.1600 (4dp), followed by Midtown South of Manhattan with 0.1400 (4dp). The third and fourth neighborhoods were Murray Hill of Manhattan and that of Queens, both with 0.1095 (4dp) Korean Restaurant average occurrence.

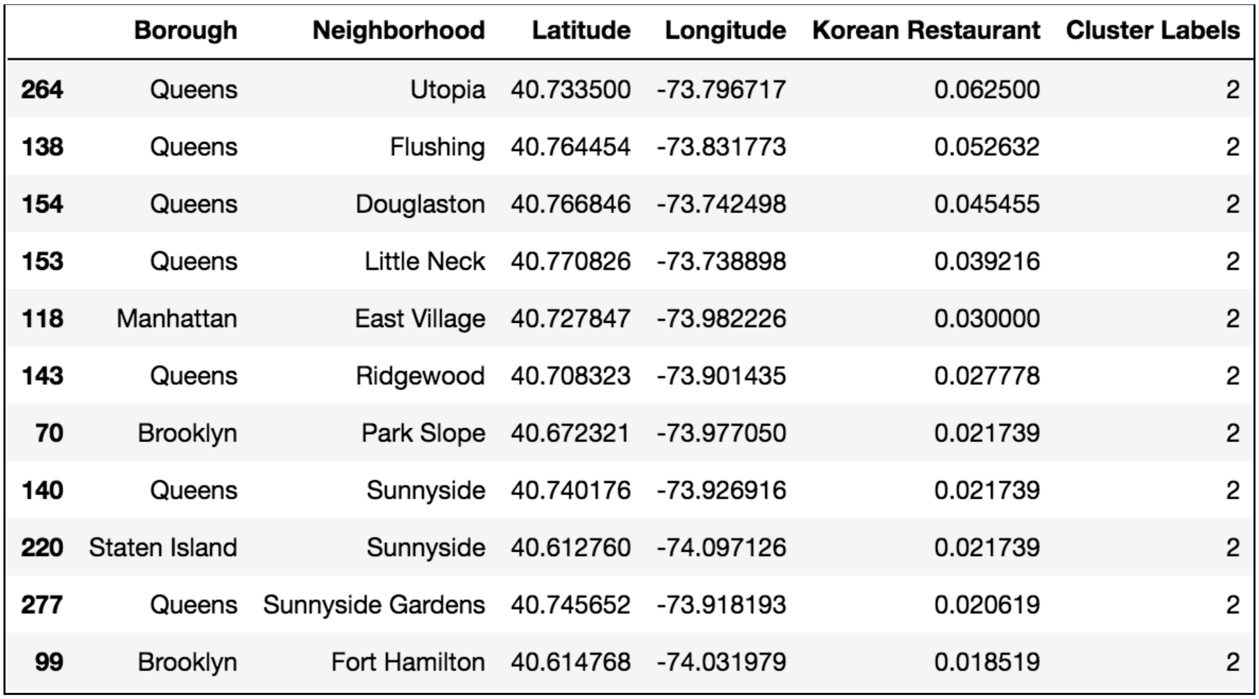


	Borough	Neighborhood	Latitude	Longitude	Korean Restaurant	Cluster Labels
161	Queens	Oakland Gardens	40.745619	-73.754950	0.160000	1
250	Manhattan	Midtown South	40.748510	-73.988713	0.140000	1
115	Manhattan	Murray Hill	40.748303	-73.978332	0.109489	1
180	Queens	Murray Hill	40.764126	-73.812763	0.109489	1

Table 13: Showing 4 neighborhoods in Cluster 1 (purple), a cluster with the highest density of Korean Restaurants, sorted in descending order of ‘Korean Restaurant’ column value.

Cluster 2 (Aquamarine)

Cluster 2 contained neighborhoods with the second highest density of Korean Restaurants, with a mean of 0.0329 (4dp). Of the 11 neighborhoods selected in this cluster, there were 7 from Queens, 2 from Brooklyn, 1 from Manhattan, and 1 from Staten Island. In particular, Queens stood out as the borough with the highest density of Korean Restaurants, with its 4 neighborhoods having the top 4 mean occurrence scores in this cluster.



	Borough	Neighborhood	Latitude	Longitude	Korean Restaurant	Cluster Labels
264	Queens	Utopia	40.733500	-73.796717	0.062500	2
138	Queens	Flushing	40.764454	-73.831773	0.052632	2
154	Queens	Douglaston	40.766846	-73.742498	0.045455	2
153	Queens	Little Neck	40.770826	-73.738898	0.039216	2
118	Manhattan	East Village	40.727847	-73.982226	0.030000	2
143	Queens	Ridgewood	40.708323	-73.901435	0.027778	2
70	Brooklyn	Park Slope	40.672321	-73.977050	0.021739	2
140	Queens	Sunnyside	40.740176	-73.926916	0.021739	2
220	Staten Island	Sunnyside	40.612760	-74.097126	0.021739	2
277	Queens	Sunnyside Gardens	40.745652	-73.918193	0.020619	2
99	Brooklyn	Fort Hamilton	40.614768	-74.031979	0.018519	2

Table 14: Showing 11 neighborhoods in Cluster 2 (aquamarine), a cluster with the second highest density of Korean Restaurants, sorted in descending order of ‘Korean Restaurant’ column value.

4.4 Cluster Analysis using Data Visualization

I also visualized these outcomes to aid understanding. I created two independent bar graphs, the first showing the number of neighborhoods per cluster and the second showing the mean number of Korean Restaurants per cluster.

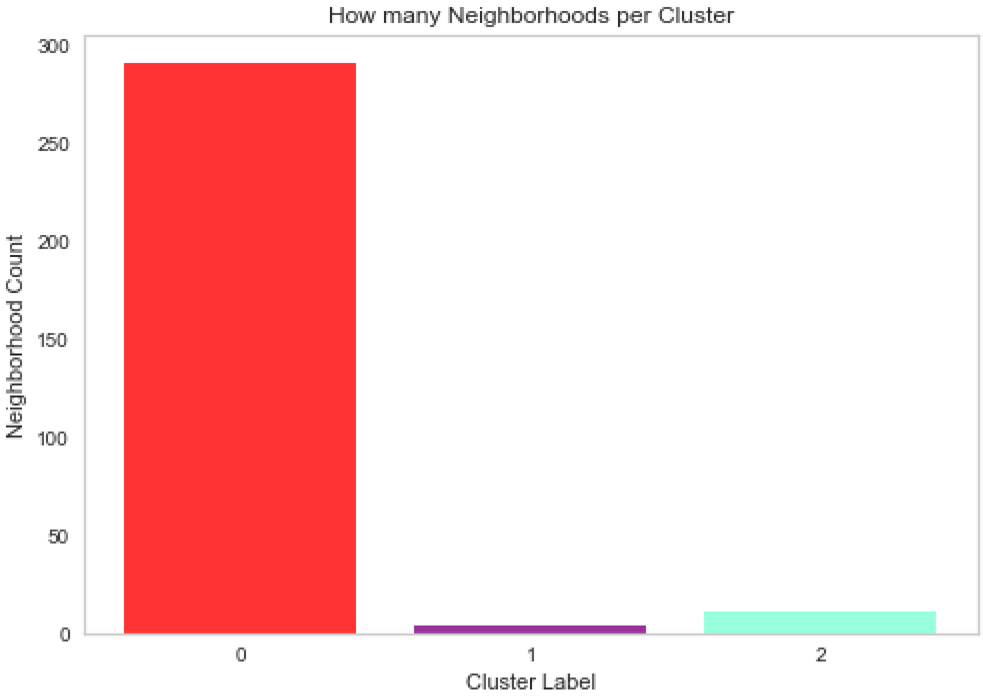


Figure 5: Bar graph showing number of neighborhoods by clusters. As in our previous analysis, Cluster 0 had an incomparably high neighborhood count of 291, while Cluster 1 and 2 had extremely low neighborhood counts at 4 and 11, respectively.

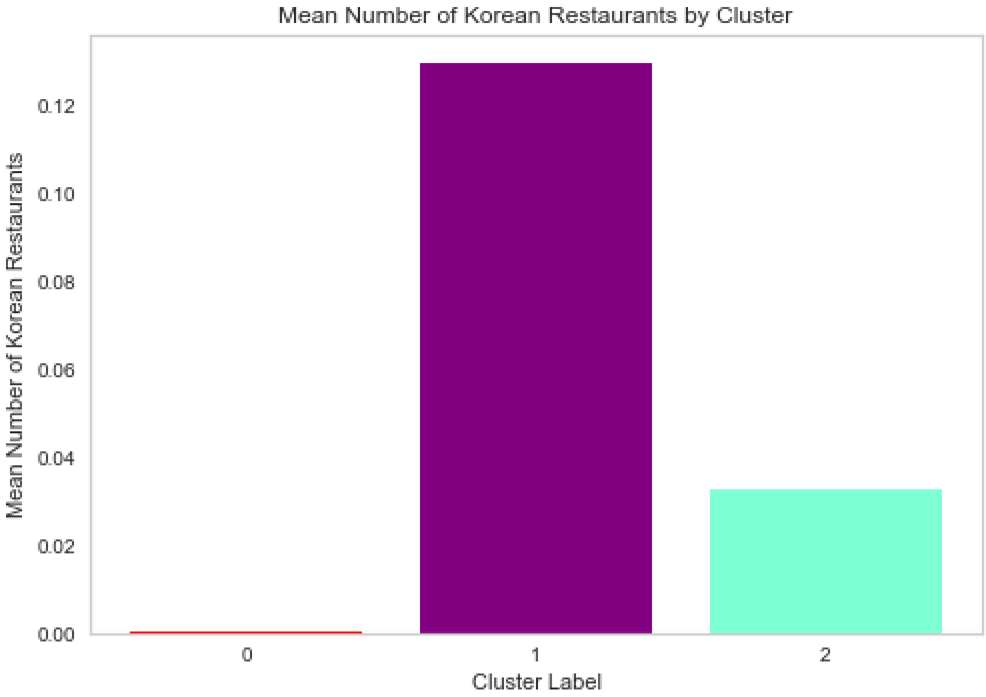


Figure 6: Bar graph showing the mean number of Korean Restaurants in each cluster. Cluster 0 has close to 0 mean occurrences of Korean Restaurants. Cluster 1 has a significantly high bar, close to 0.13 (2dp) mean occurrences of Korean Restaurants. Cluster 2 has a moderate mean occurrence of Korean Restaurants at around 0.03 (2dp).

5. Discussion

Cluster analysis revealed that of the five boroughs in NYC, Manhattan and Queens have the highest density of Korean Restaurants. This is first evident in Cluster 1, where all the neighborhoods selected for high density of Korean Restaurants are either from Manhattan or Queens. Further, Cluster 2, containing neighborhoods with the second highest density of Korean Restaurants, has 8 or 11 neighborhoods selected from Queens or Manhattan, with Queens topping the first four on the list. In Cluster 0, containing neighborhoods with the lowest density of Korean Restaurants, the 6 neighborhoods with non-zero mean occurrences of Korean Restaurants are 83.33% (5 out of 6) from Queens or Manhattan. Looking at the neighborhood-level, exploratory data analysis on the number of Korean Restaurants across New York neighborhoods (3.3) revealed that Murray Hill and Midtown South, of Manhattan, contain incomparably high numbers of Korean Restaurants. Hence, Queens and Manhattan are boroughs containing neighborhoods, in particular, Murray Hill and Midtown South, that are already populated with Korean Restaurants, indicating much competition and generally harder survival rate of new restaurants. Hence, I excluded these areas for the purposes of this project.

Let’s now examine the other three boroughs of NYC - Brooklyn, Staten Island, and Bronx. Cluster 2 includes 7 neighborhoods from Queens, 1 from Manhattan, 2 from Brooklyn, and 1 from Staten Island. Of the 6 neighborhoods in Cluster 0 that have non-zero mean occurrences of Korean Restaurants, none of them are from Staten Island or Bronx. Therefore, neighborhoods in Brooklyn are more populated Korean Restaurants, compared to those in Staten Island and Bronx.

Lastly, as clearly demonstrated in the two bar graphs in 4.4, while there innumerable neighborhoods in Cluster 0, there is a close-to-zero mean occurrence of Korean Restaurants in that Cluster. On the other hand, while there are a fractional number of neighborhoods in Cluster 1, there are a large number of Korean Restaurants in that cluster. Lastly, while there are a small number of neighborhoods in Cluster 2, there is not a huge number of korean Restaurants in that cluster, though not as many as that of Cluster 0. Therefore, I selected among the neighborhoods in Cluster 0, excluding those with non-zero occurrences of Korean Restaurants.

From these findings, I can conclude that the best neighborhood to open a new Korean Restaurant is one within Staten Island or Bronx from Cluster 0. Since there are 62 neighborhoods within Staten Island and 52 neighborhoods within Bronx from Cluster 0, I narrowed it down using the following measures.

When considering neighborhoods, it is also important to consider safety, given the significance of quality of life on the success rate of businesses, including restaurants; the safer the neighborhoods are, the more likely people will venture out to dine in, and thus the more likely restaurant businesses will thrive. Though Bronx is known for its high crime rates, it still has neighborhoods that are relatively safe, all of which are in Cluster 0 and have zero Korean Restaurants in their area: Riverdale, Bedford Park, Fordham, Morris Park, Pelham Bay.

Since Staten Island is relatively safe, I've decided to recommend the most urban and populated part of the island to ensure business success. This is the North Shore, especially the neighborhoods of St. George, Tompkinsville, Clifton, and Stapleton Heights.

6. Conclusion

6.1 Summary

This project aimed to select the best neighborhood to open a new Korean Restaurant in NYC. To do so, I identified areas that are not already crowded with Korean Restaurants. Of these neighborhoods, I selected ones that are relatively safe and urban to ensure best customer attraction and business growth. They are the following:

- Riverdale, Bedford Park, Fordham, Morris park, Pelham Bay of Bronx or
- St. George, Tompkinsville, Clifton, Stapleton Heights of Staten Island

6.2 Limitations and Future Studies

Finding the best neighborhood to open a new Korean Restaurant can depend on so many other factors, besides from density. For example, businesses may thrive or decay depending on the presence of a demanding population for Korean food (e.g. Korean population). Future studies could integrate population and Korean Restaurant interest data to better estimate the neighborhood to open a Korean Restaurant.

Overall, this project has provided useful information regarding relatively safe and urban neighborhoods with less crowded Korean Restaurants for opening a Korean Restaurant in NYC. It will be beneficial for anyone interested in investing in or opening a new Korean Restaurant in NYC.