Finding the Best Neighborhood in NYC to Open a Korean Restaurant

By Soomin Kim

Introduction

Background

New York City

- Most populated city in the US
- Most ethnically, linguistically, culturally diverse cities in the world
- 36% of city population are foreign-born
- 800 languages are spoken in the City
- Dozens of ethnic enclaves: Little Guyana, Koreatown, Little India, Little Odessa, Chinatown, Little Australia, Little Poland, Little Italy
- > 100,000 ethnic Koreans
- Dense Korean population & Fast growing K-food business
- Opportunities to expand Korean restaurant business in NYC

Aim & Target Audience

Aim:

- To find best location (borough, neighborhood) in NYC to open a Korean restaurant.

Success Criteria:

- Not already crowded with Korean restaurants
- As close to the City center as possible

Target Audience:

- Anyone - entrepreneurs, business owners, investors - looking to start or expand their Korean restaurant business in NYC

Data

Data Sources

Data 1: New York City Neighborhood Data:
- 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.

Data Cleaning - Data 1

	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

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

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

Data Cleaning - Data 2

	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

Exploratory Data Analysis

Visualization of the map of NYC by neighborhood

- Folium package → to visualize a map of New York with neighborhoods superimposed on top on a leaflet map.
- The CircleMarker() function → to color-code boroughs, providing a clearer understanding of where each borough is located.
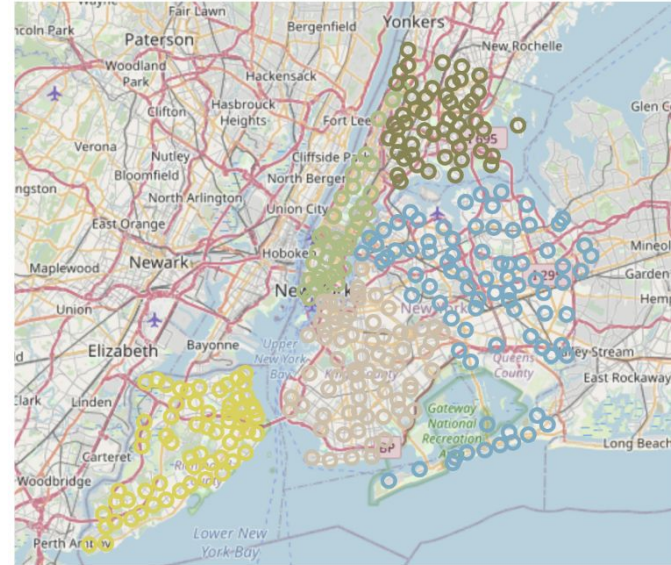

Figure 1: Map of NYC color-coded by neighborhood

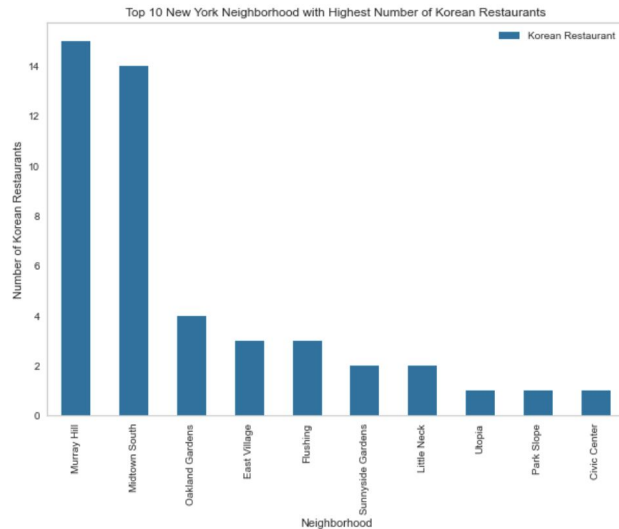# Top 10 NY Neighborhoods with the Highest Number of Korean Restaurants

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.

Machine Learning

One Hot Encoding

- pd.get_dummies function within the Pandas package
- To transform categorical variables (*nyc_venues*) → numerical ones (*nyc_onehot*)
- In order to divide New York City into clusters of neighborhoods with a similar number of Korean Restaurants

	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

Calculate 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.

Extract 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

K-Means Clustering

- K-Means clustering method to cluster the neighborhoods based on neighborhoods with similar mean frequencies of Korean Restaurants.
    - KElbowVisualizer to identify 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 to follow, the line annotating the "elbow" is at k = 3. This meant that the optimum number of clusters is 3.
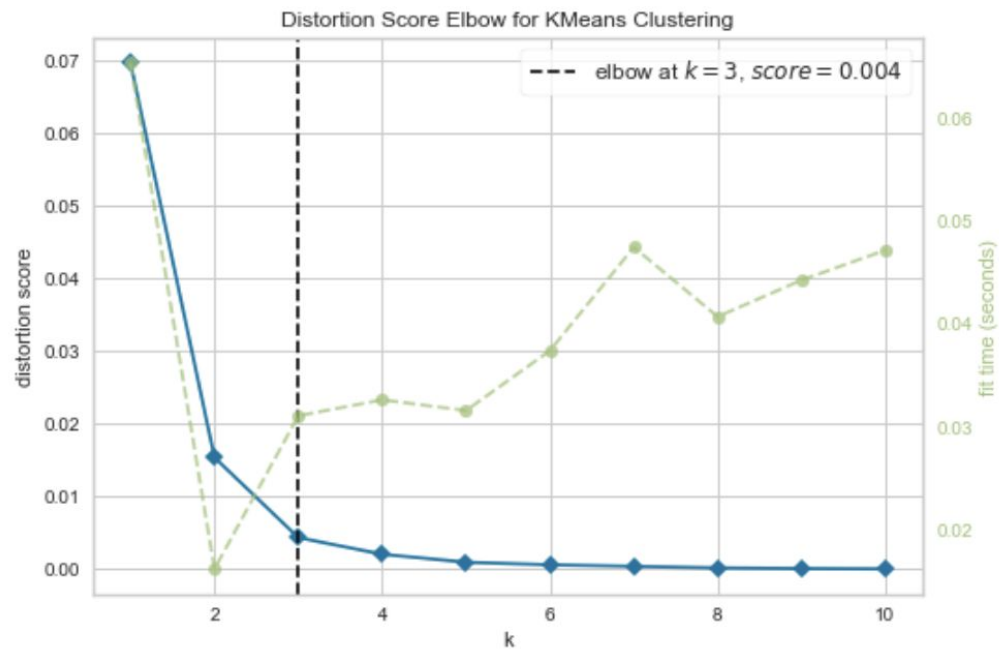
Figure 3: KElbowVisualizer revealed an elbow at k=3, with a score of 0.004, indicating that the optimal number of clusters is 3.

K-Means clustering method to create 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

To ensure that we have one location data (longitude and latitude values) for each neighborhood, created a new dataframe *nyc_venues_grouped* that had mean longitude and latitude values for each neighborhood.

	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.

	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

Same-named Neighborhood Problem

- While this dataframe has the neighborhood data, it doesn't have any boroughs information.
    - → problems when analyzing same-named neighborhoods that are actually in different boroughs, leading to inaccurate analysis.
    - *kor_merged* dataframe actually contained the same location information for neighborhoods with same names, despite them being of different boroughs.
    - So integrated boroughs data with more accurate location data for each neighborhood.

	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.

Cluster Map

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

Results: Examining the Clusters

Cluster 0 (Red)

- 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.
- 291 neighborhoods, all 5 boroughs
- 6 neighborhoods with non-zero mean occurrence of Korean Restaurants
    - → Brooklyn, Queens, Manhattan
- 285 with zero mean occurrences of Korean Restaurants
    - → Staten Island, Bronx

	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)

- Neighborhoods with the highest density of Korean Restaurants, with a mean of 0.1297 (4dp) Korean Restaurant average occurrences.
- 4 neighborhoods
- Queens or Manhattan
    - 1) Oakland Gardens of Queens with 0.1600 (4dp)
    - 2) Midtown South of Manhattan with 0.1400 (4dp).
    - 3, 4) Murray Hill of Manhattan and Murray Hill of Queens, both with 0.1095 (4dp)

	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)

- Neighborhoods with the second highest density of Korean Restaurants, with a mean of 0.0329 (4dp)
- 11 neighborhoods, 4 boroughs
- 7 from Queens, 2 from Brooklyn, 1 from Manhattan, and 1 from Staten Island

	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.

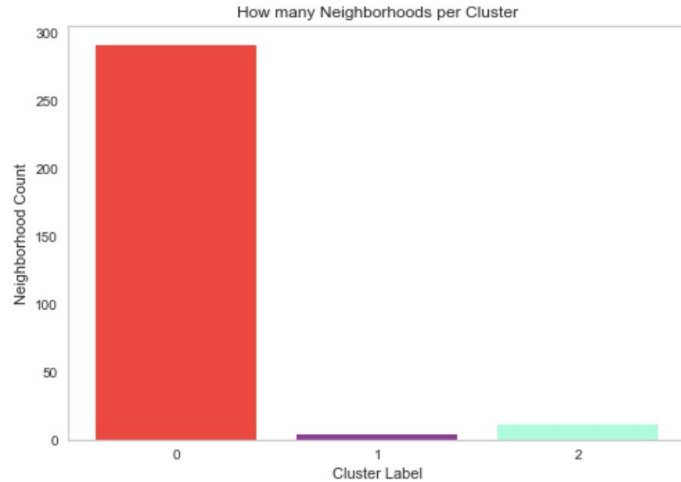# Cluster Analysis using Data Visualization

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.
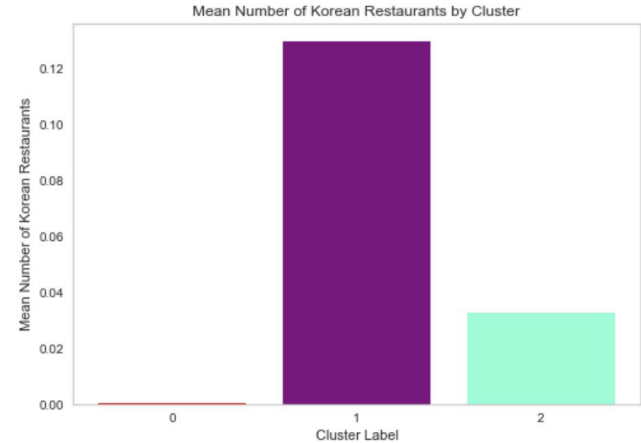

Figure 6: Bar graph showing the mean number of Korean Restaurants in each cluster. Cluster 0 has close to 0 mean occurrences of Korean Restaurants. Cluster 1 has a significantly high bar, close to 0.13 (2dp) mean occurrences of Korean Restaurants. Cluster 2 has a moderate mean occurrence of Korean Restaurants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

Discussion

- Manhattan, Queens → high density of Korean restaurants → high competition → excluded these
- Brooklyn (denser) > Staten Island, Bronx → excluded Brooklyn

Neighborhoods from **Staten Island** or **Bronx**!

- Selected neighborhoods that are relatively safe & urban

Conclusion

Summary

Selected neighborhoods that are:
- not already crowded with Korean Restaurants
- relatively safe and urban

Riverdale, Bedford Park, Fordham, Morris park, Pelham Bay of Bronx

St. George, Tompkinsville, Clifton, Stapleton Heights of Staten Island

Limitations and Future Studies

- <u>Limitation</u>: finding the best neighborhood to open a new Korean Restaurant can depend on so many other factors, besides from density.
    - E.g. Businesses may thrive or decay depending on the presence of a demanding population for Korean food (e.g. Korean population).
- <u>Future studies</u>: 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 new Korean Restaurant in NYC.

It will be beneficial for entrepreneurs or business owners who are interested in investing in or opening a new Korean Restaurant in NYC.