Capstone Project: The Battle of Neighborhoods

Background & Problem Description

New York City is one of the most diverse and populated cities in the world. It is a melting pot of different cultures and cuisines from around the world. It is also considering a foodie heaven because there are so many options. That means that there are a lot of options to choose from and that selecting the best place can be tough. It should be important to know which places are the best depending upon the neighborhood you are in. This project will help to understand the diversity of a neighborhood by leveraging venue data from Four square's 'Place API' and 'k-means' clustering machine learning algorithm. The audience would be anyone that is interested to use this analysis to understand the distribution of different cultures and cuisines in New York City.

Data Preparation

These are the Data Sources Used for this Analysis:

1. **New York Data Set:** https://geo.nyu.edu/catalog/nyu_2451_34572

The data set will be our base neighborhood data set to cross reference against the Foursquare API venue data

2. **Foursquare API:** to get the most common venues of given Borough of New York City and to get the venues' record of given venues of New York City.

3. **Geophy** Library in Python: this will help us get the Lat and Long of the NYC data set

Methodology:

1. Loading Dependciens

We first most load the following libraries into Jupyter Notebook

```python
import numpy as np # library to handle data in a vectorized manner

import pandas as pd # library for data analsysis
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)

import json # library to handle JSON files
from pprint import pprint # data pretty printer

import requests # library to handle requests
from bs4 import BeautifulSoup  # library to handle web scraping

from geopy.geocoders import Nominatim # convert an address into latitude and longitude values

import folium # map rendering library

import matplotlib.cm as cm # Matplotlib and associated plotting modules
import matplotlib.colors as colors # Matplotlib and associated plotting modules

from pandas.io.json import json_normalize # tranform JSON file into a pandas dataframe

from collections import Counter # count occurrences

from sklearn.cluster import KMeans # import k-means from clustering stage
```

2. Transforming and Exploring the NYC Data Set

We upload the NYC data set and run a couple lines of code to transform the data. We then use Geopy to get the Latitude and Longitude of each borough and plot on a map:

Tranform the data into a *pandas* dataframe

```python
# define the dataframe columns
column_names = ['Borough', 'Neighborhood', 'Latitude', 'Lon

# instantiate the dataframe
neighborhoods = pd.DataFrame(columns=column_names)
neighborhoods
```

3. Appending the Foursquare data to the NYC Data Set

We take the following steps to append the data:

1. Create the API request URL with our Foursquare developer credentials
2. Make the GET request

3. Return only relevant information for each nearby venue within our NYC data set

4. Append all nearby venues to a list

Utilizing the Foursquare API to explore the neighborhoods and segment them.

[13]:
```
CLIENT_ID = '[redacted]' # your Foursquare ID
CLIENT_SECRET = '[redacted]' # your Foursquare Secret
VERSION = '20200225' # Foursquare API version

print('Your credentails:')
print('CLIENT_ID: ' + CLIENT_ID)
print('CLIENT_SECRET:' + CLIENT_SECRET)
```

```
Your credentails:
CLIENT_ID: QXGYVTSQLO0EHISX3GIOTZVSATCJWT2L2CKXOJYZKOPHKVJB
CLIENT_SECRET:IC344XWNJT5KAEYONE5U4DU1IFRIJMWHP2E3U2VB1NOX4ANA
```

Fetch Foursquare Venue Category Hierarchy

[14]:
```
url = 'https://api.foursquare.com/v2/venues/categories?&client_id={}&client_secret={}&v={}'.format(
    CLIENT_ID,
    CLIENT_SECRET,
    VERSION)
category_results = requests.get(url).json()
```

Let's see the structure or the keys of the returned request.

[15]:
```
for key, value in category_results['response']['categories'][0].items():
    print(key, len(str(value)))
```

```
id 24
name 20
pluralName 20
shortName 20
icon 98
categories 15910
```

[16]:
```
category_list = category_results['response']['categories']
```

[17]:
```
len(category_list)
```

[17]:
```
10
```

[18]:
```
for data in category_list:
    print(data['id'], data['name'])
```

Get the neighborhood's latitude and longitude values.

```
22]: neighborhood_latitude = neighborhoods.loc[0, 'Latitude'] # neighborhood latitude value
     neighborhood_longitude = neighborhoods.loc[0, 'Longitude'] # neighborhood longitude value

     neighborhood_name = neighborhoods.loc[0, 'Neighborhood'] # neighborhood name

     print('Latitude and longitude values of {} are {}, {}.'.format(neighborhood_name,
                                                                    neighborhood_latitude,
                                                                    neighborhood_longitude))
```

```
Latitude and longitude values of Wakefield are 40.89470517661, -73.84720052054902.
```

Now, let's get the *Food* that is in Wakefield within a radius of 500 meters.

First, let's create the GET request URL to search for Venue with requested *Category ID*

```
23]: LIMIT = 1 # limit of number of venues returned by Foursquare API
     radius = 500 # define radius
     categoryId = '4d4b7105d754a06374d81259' # category ID for "Food"

     # create URL

     url = 'https://api.foursquare.com/v2/venues/search?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&categoryId={}&limit={}'.format(
         CLIENT_ID,
         CLIENT_SECRET,
         VERSION,
         neighborhood_latitude,
         neighborhood_longitude,
         radius,
         categoryId,
         LIMIT)
     url # display URL
```

```
23]: 'https://api.foursquare.com/v2/venues/search?&client_id=QXGYVTSQLO0EHISX3GIOTZVSATCJWT2L2CKXOJYZKOPHKVJB&client_secret=IC344XWNJT5KAEYONE5U
     9470517661,-73.84720052054902&radius=500&categoryId=4d4b7105d754a06374d81259&limit=1'
```

Send the GET request and examine the resutls

4. Model Selection

We will chose the K-Means Clustering Algorithm to help build segments for the neighborhoods based on types of cuisines in that particular neighborhood.

Definition: k-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. k-Means minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. Better Euclidean solutions can for example be found using k-medians and k-medoids. Source: https://en.wikipedia.org/wiki/K-means_clustering

1. We will first group the data set and perform some analysis to understand the data better:

[52]:	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Allerton	Pizza Place	Chinese Restaurant	Mexican Restaurant	Fried Chicken Joint	Fast Food Restaurant
1	Annadale	Pizza Place	Italian Restaurant	American Restaurant	Sushi Restaurant	Japanese Restaurant
2	Arden Heights	Pizza Place	Italian Restaurant	American Restaurant	Sushi Restaurant	Mexican Restaurant
3	Arlington	Pizza Place	American Restaurant	Peruvian Restaurant	Fast Food Restaurant	Spanish Restaurant
4	Arrochar	Italian Restaurant	Pizza Place	Middle Eastern Restaurant	Mediterranean Restaurant	Polish Restaurant

3. Analysis & Machine Learning

Let's check the size of the resulting dataframe

[29]:
```
print(nyc_venues.shape)
nyc_venues.head()
```

```
(13908, 7)
```

[29]:	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
1	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop
2	Wakefield	40.894705	-73.847201	Cooler Runnings Jamaican Restaurant Inc	40.898083	-73.850259	Caribbean Restaurant
3	Wakefield	40.894705	-73.847201	SUBWAY	40.890468	-73.849152	Sandwich Place
4	Wakefield	40.894705	-73.847201	Burger King	40.895540	-73.856460	Fast Food Restaurant

Let's find out how many unique categories can be curated from all the returned venues

[30]:
```
print('There are {} uniques categories.'.format(len(nyc_venues['Venue Category'].unique())))
nyc_venues.groupby('Venue Category')['Venue Category'].count().sort_values(ascending=False)
```

```
Persian Restaurant          1
Sake Bar                    1
Beach                       1
Caucasian Restaurant        1
Czech Restaurant            1
Hawaiian Restaurant         1
Pop-Up Shop                 1
Hong Kong Restaurant        1
Bowling Alley               1
Indonesian Restaurant       1
Factory                     1
Name: Venue Category, dtype: int64
```

As we are interested in exploring the diversity of the neighborhood. Let's remove the generalized categories, like Coffee Shop, Cafe, etc.

[31]:
```
# list all the categories
unique_categories = nyc_venues['Venue Category'].unique().tolist()
print(', '.join(str(x) for x in unique_categories))
```

```
Ice Cream Shop, Donut Shop, Caribbean Restaurant, Sandwich Place, Fast Food Restaurant, Deli / Bodega, Bakery, Food, Comfort Food Restaurant, Fried Chicken Joint, La
```

```
There are 3381 uniques venues.
```

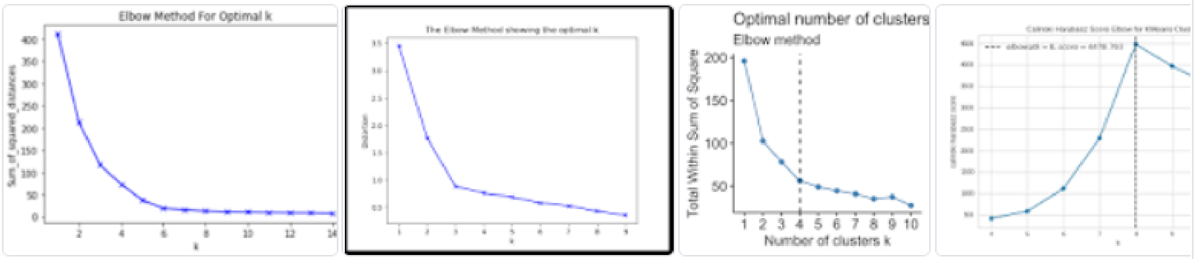
Analyze Each Neighborhood

[38]:
```
# one hot encoding
nyc_onehot = pd.get_dummies(nyc_venues[['Venue Category']], prefix="", prefix_sep="")
nyc_onehot.head()
```

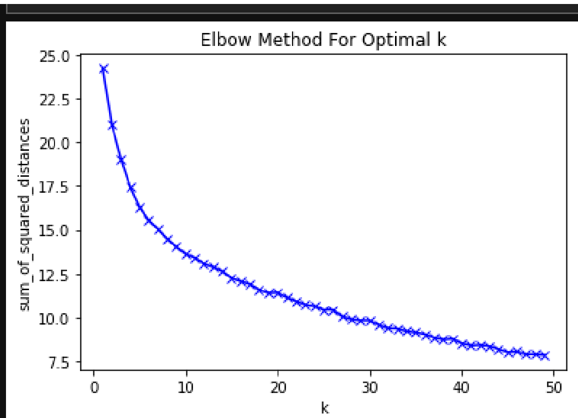
	Restaurant	Restaurant	Restaurant	Restaurant	Restaurant	Restaurant	Restaurant	Restaurant	Joint	House	Restaurant	Place	Creole Restaurant	Restaurant	Restaurant	Restaurant	Restaurant	Restaurant	Sho
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

5. Cluster Evaluation

1. **Elbow Method -** calculate the sum of squared distances of samples to their closest cluster center for different values of k. The value of k after which there is no significant decrease in sum of squared distances is chosen.



K-means is a simple unsupervised machine learning algorithm that groups data into a specified number (**k**) of clusters. ... The **elbow method** runs **k-means** clustering on the dataset for a range of values for **k** (say from 1-10) and then for each value of **k** computes an average score for all clusters.



Elbow method does not seem to help us to determine the optimal number of clusters. Let's use another method.

2. **Silhouette Method** - value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation)

Silhouette (clustering)

From Wikipedia, the free encyclopedia

Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object has been classified.[1]

The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster an most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

The silhouette can be calculated with any distance metric, such as the Euclidean distance or the Manhattan distance.

Definition [edit]

Assume the data have been clustered via any technique, such as k-means, into k clusters.

For data point $i \in C_i$ (data point i in the cluster C_i), let

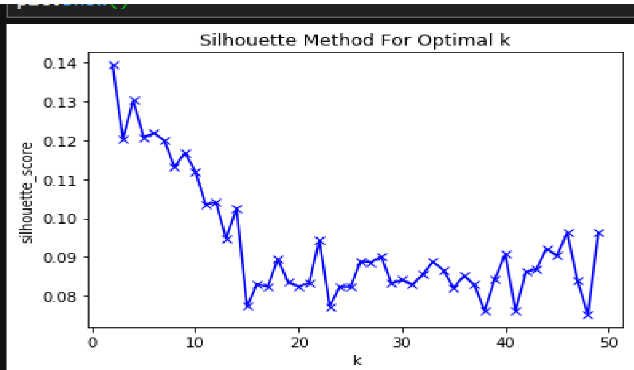
$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

be the mean distance between i and all other data points in the same cluster, where $d(i, j)$ is the distance between data points i and j in the cluster C_i (we divide by $|C_i| - 1$ because we do not include the distance $d(i, i)$ in the sum). We can interpret $a(i)$ as a measure of how well i is assigned to its cluster (the smaller the value, the better the assignment).

We then define the mean dissimilarity of point i to some cluster C as the mean of the distance from i to all points in C (where $C \neq C_i$).

For each data point $i \in C_i$, we now define

Source: https://en.wikipedia.org/wiki/Silhouette_(clustering)



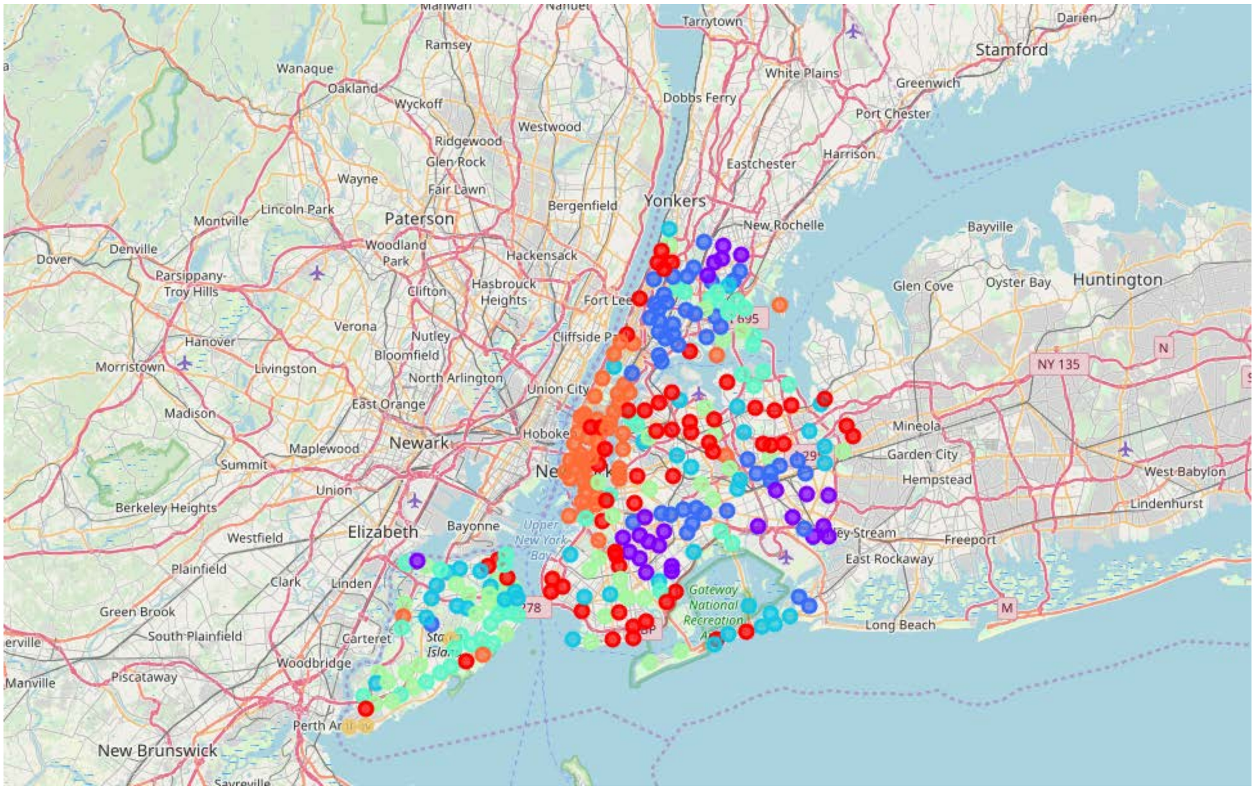
There is a peak at k = 2, k = 4 and k = 8. Two and four clusters will give a very broad classification of the venues.

Let's set number of clusters = 8

Based on this method, the recommendation from our data set is use 8 Clusters.

Results

The model produced 8 segments grouping the neighborhoods by borough and by Cuisines type. The map to the right is a high level view of the clusters created



The model produced 8 segments grouping the neighborhoods by borough and by Cuisines type. The map to the right is a high level view of the clusters created

0 - Pizza/Fast Food – Queens & Brooklyn

1 – Caribbean Cuisines – Brooklyn & Queens

2 – Italian/Pizza – Staten Island

3 – Italian/Pizza/American – Manhattan, Brooklyn, & Queens

4 – Pizza/Italian – Staten Island & The Bronx

5 – Italian/Vietnamese - Staten Island

6 – Mix of Cuisines – Staten Island

7 – American – Manhattan *& Brooklyn

Cluster 0

- Segment 0 are neighborhoods that had a major of restaurants that are Pizza Place and Fast Food

- Most of the neighborhoods reside in Brooklyn and Queens

Cluster 0

```
[64]: cluster_0 = nyc_merged.loc[nyc_merged['Cluster Labels'] == 0, nyc_merged.columns[1:12]]
cluster_0.head(5)
```

[64]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
6	Astoria	Fast Food Restaurant	Mediterranean Restaurant	Chinese Restaurant	Vietnamese Restaurant	American Restaurant	Queens	40.768509	-73.915654
8	Auburndale	Korean Restaurant	Greek Restaurant	Sushi Restaurant	American Restaurant	Cantonese Restaurant	Queens	40.761730	-73.791762
9	Bath Beach	Fast Food Restaurant	Chinese Restaurant	Cantonese Restaurant	Sushi Restaurant	Vietnamese Restaurant	Brooklyn	40.599519	-73.998752
11	Bay Ridge	Fast Food Restaurant	Thai Restaurant	American Restaurant	Mexican Restaurant	Middle Eastern Restaurant	Brooklyn	40.625801	-74.030621
14	Bayside	Korean Restaurant	Chinese Restaurant	Fast Food Restaurant	Asian Restaurant	Sushi Restaurant	Queens	40.766041	-73.774274

```
[65]: for col in required_column:
    print(cluster_0[col].value_counts(ascending = False))
    print("---------------------------------------------")
```

```
Pizza Place                  16
Fast Food Restaurant          8
Italian Restaurant            6
Mexican Restaurant            5
Korean Restaurant             5
Sushi Restaurant              3
Thai Restaurant               2
Indian Restaurant             2
Seafood Restaurant            2
Greek Restaurant              1
Russian Restaurant            1
Ramen Restaurant              1
Sri Lankan Restaurant         1
Asian Restaurant              1
Eastern European Restaurant   1
Chinese Restaurant            1
American Restaurant           1
Filipino Restaurant           1
Name: 1st Most Common Venue, dtype: int64
```

```
Queens           23
Brooklyn         18
Manhattan         8
Staten Island     5
Bronx             4
Name: Borough, dtype: int64
---------------------------------------------
```

Cluster 1

- Segment 1 is a mostly neighborhoods that are Caribbean.
- Most of these neighborhoods reside in Brooklyn and Queen

Cluster 1

[66]:
```
cluster_1 = nyc_merged.loc[nyc_merged['Cluster Labels'] == 1, nyc_merged.columns[1:12]]
cluster_1.head(5)
```

[66]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
36	Brookville	Fried Chicken Joint	Caribbean Restaurant	Pizza Place	Chinese Restaurant	Fast Food Restaurant	Queens	40.660003	-73.751753
37	Brownsville	Caribbean Restaurant	Pizza Place	Fried Chicken Joint	Chinese Restaurant	Fast Food Restaurant	Brooklyn	40.663950	-73.910235
41	Cambria Heights	Caribbean Restaurant	Fried Chicken Joint	Mexican Restaurant	African Restaurant	Latin American Restaurant	Queens	40.692775	-73.735269
42	Canarsie	Chinese Restaurant	Caribbean Restaurant	Fast Food Restaurant	Pizza Place	Mexican Restaurant	Brooklyn	40.635564	-73.902093
77	East Flatbush	Caribbean Restaurant	Pizza Place	Fried Chicken Joint	Chinese Restaurant	Fast Food Restaurant	Brooklyn	40.641718	-73.936103

[67]:
```
for col in required_column:
    print(cluster_1[col].value_counts(ascending = False))
    print("---------------------------------------------")
```

```
Caribbean Restaurant    21
Chinese Restaurant       2
American Restaurant      1
Fried Chicken Joint      1
Name: 1st Most Common Venue, dtype: int64
---------------------------------------------
Fast Food Restaurant    7
Fried Chicken Joint     5
Pizza Place             5
Chinese Restaurant      4
Caribbean Restaurant    3
Seafood Restaurant      1
Name: 2nd Most Common Venue, dtype: int64
---------------------------------------------
Brooklyn         11
Queens            8
Bronx             5
Staten Island     1
Name: Borough, dtype: int64
---------------------------------------------
```

Cluster 2

- Segment 2 are mostly a mix of Italian/Pizza
- Most reside in Staten Island

Cluster 2

[68]:
```
cluster_2 = nyc_merged.loc[nyc_merged['Cluster Labels'] == 2, nyc_merged.columns[1:12]]
cluster_2.head(5)
```

[68]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
0	Allerton	Pizza Place	Chinese Restaurant	Mexican Restaurant	Fried Chicken Joint	Fast Food Restaurant	Bronx	40.865788	-73.859319
13	Baychester	Caribbean Restaurant	Pizza Place	Fast Food Restaurant	Mexican Restaurant	Seafood Restaurant	Bronx	40.866858	-73.835798
15	Bayswater	Chinese Restaurant	Fried Chicken Joint	Pizza Place	Fast Food Restaurant	American Restaurant	Queens	40.611322	-73.765968
16	Bedford Park	Pizza Place	Fast Food Restaurant	Fried Chicken Joint	Chinese Restaurant	New American Restaurant	Bronx	40.870185	-73.885512
30	Briarwood	Fast Food Restaurant	Pizza Place	Fried Chicken Joint	Chinese Restaurant	Caribbean Restaurant	Queens	40.710935	-73.811748

[67]:
```
for col in required_column:
    print(cluster_2[col].value_counts(ascending = False))
    print("---------------------------------------------")
```

```
Italian Restaurant      27
Pizza Place             16
Fast Food Restaurant     2
Falafel Restaurant       1
Name: 1st Most Common Venue, dtype: int64
---------------------------------------------
Italian Restaurant      16
Pizza Place             15
Chinese Restaurant       5
Asian Restaurant         4
Mexican Restaurant       2
Fast Food Restaurant     2
American Restaurant      2
Name: 2nd Most Common Venue, dtype: int64
---------------------------------------------
Staten Island    22
Queens           10
Bronx             8
Brooklyn          6
Name: Borough, dtype: int64
---------------------------------------------
```

Cluster 3

- Segment 3 are heavy Italian, Pizza, and American
- This is our largest segment with a majority of neighborhoods in Manhattan, Brooklyn, and Queens.

Cluster 3

```
[69]: cluster_3 = nyc_merged.loc[nyc_merged['Cluster Labels'] == 3, nyc_merged.columns[1:12]]
      cluster_3.head(5)
```

[69]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
5	Arverne	Chinese Restaurant	Pizza Place	American Restaurant	Taco Place	Asian Restaurant	Queens	40.589144	-73.791992
7	Astoria Heights	Greek Restaurant	Pizza Place	Italian Restaurant	Chinese Restaurant	Fried Chicken Joint	Queens	40.770317	-73.894680
19	Bellaire	Pizza Place	Chinese Restaurant	Fast Food Restaurant	Spanish Restaurant	Italian Restaurant	Queens	40.733014	-73.738892
23	Bensonhurst	Chinese Restaurant	Asian Restaurant	Cantonese Restaurant	Pizza Place	Dim Sum Restaurant	Brooklyn	40.611009	-73.995180
25	Blissville	Pizza Place	Italian Restaurant	Fast Food Restaurant	Chinese Restaurant	Indian Restaurant	Queens	40.737251	-73.932442

```
[69]: for col in required_column:
          print(cluster_3[col].value_counts(ascending = False))
          print("---------------------------------------------")
```

```
Pizza Place                      21
Italian Restaurant               18
American Restaurant              14
Korean Restaurant                 7
Seafood Restaurant                5
Fast Food Restaurant              5
New American Restaurant           3
Thai Restaurant                   3
Mexican Restaurant                2
Greek Restaurant                  2
Vegetarian / Vegan Restaurant     2
Sushi Restaurant                  2
Middle Eastern Restaurant         1
Vietnamese Restaurant             1
Indian Restaurant                 1
Ramen Restaurant                  1
Eastern European Restaurant       1
Name: 1st Most Common Venue, dtype: int64
---------------------------------------------
Italian Restaurant               17
Pizza Place                      12
American Restaurant              11
Fast Food Restaurant              7
French Restaurant                 7
Mexican Restaurant                6
BBQ Joint                         4
Vietnamese Restaurant             4
Turkish Restaurant                2
Middle Eastern Restaurant         2
```

```
Name: 2nd Most Common Venue, dtype: int64
---------------------------------------------
Manhattan        28
Brooklyn         25
Queens           22
Staten Island    12
Bronx             2
Name: Borough, dtype: int64
---------------------------------------------
```

Cluster 4

- Segment 4 are neighborhoods that are heavy Italian Restaurants and Pizza Places
- Most are located in Staten Island and the Bronx

Cluster 4

```
[70]: cluster_4 = nyc_merged.loc[nyc_merged['Cluster Labels'] == 4, nyc_merged.columns[1:12]]
cluster_4.head(5)
```

[70]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
4	Arrochar	Italian Restaurant	Pizza Place	Middle Eastern Restaurant	Mediterranean Restaurant	Polish Restaurant	Staten Island	40.596313	-74.067124
12	Bay Terrace	Italian Restaurant	Pizza Place	Asian Restaurant	American Restaurant	Chinese Restaurant	Queens	40.782843	-73.776802
12	Bay Terrace	Italian Restaurant	Pizza Place	Asian Restaurant	American Restaurant	Chinese Restaurant	Staten Island	40.553988	-74.139166
18	Beechhurst	Italian Restaurant	Pizza Place	Japanese Restaurant	Chinese Restaurant	Greek Restaurant	Queens	40.792781	-73.804365
22	Belmont	Italian Restaurant	Pizza Place	Fast Food Restaurant	Mexican Restaurant	American Restaurant	Bronx	40.857277	-73.888452

```
[71]: for col in required_column:
    print(cluster_4[col].value_counts(ascending = False))
    print("---------------------------------------------")
```

```
Italian Restaurant     27
Pizza Place            13
American Restaurant     1
Name: 1st Most Common Venue, dtype: int64
---------------------------------------------
Pizza Place                15
Italian Restaurant         12
Fast Food Restaurant        5
American Restaurant         3
Japanese Restaurant         2
Mexican Restaurant          1
New American Restaurant     1
Greek Restaurant            1
Asian Restaurant            1
Name: 2nd Most Common Venue, dtype: int64
---------------------------------------------
Staten Island    20
Bronx            10
Queens            8
Brooklyn          3
Name: Borough, dtype: int64
---------------------------------------------
```

Cluster 5

- Segment 5 are neighborhoods that have a variety or “diverse” amount of cuisines mostly in Staten Island

Cluster 5

```
[72]: cluster_5 = nyc_merged.loc[nyc_merged['Cluster Labels'] == 5, nyc_merged.columns[1:12]]
cluster_5.head(5)
```

[72]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
1	Annadale	Pizza Place	Italian Restaurant	American Restaurant	Sushi Restaurant	Japanese Restaurant	Staten Island	40.538114	-74.178549
2	Arden Heights	Pizza Place	Italian Restaurant	American Restaurant	Sushi Restaurant	Mexican Restaurant	Staten Island	40.549286	-74.185887
3	Arlington	Pizza Place	American Restaurant	Peruvian Restaurant	Fast Food Restaurant	Spanish Restaurant	Staten Island	40.635325	-74.165104
21	Bellerose	Pizza Place	Chinese Restaurant	Indian Restaurant	Italian Restaurant	American Restaurant	Queens	40.728573	-73.720128
26	Bloomfield	Pizza Place	Italian Restaurant	Mexican Restaurant	BBQ Joint	Yemeni Restaurant	Staten Island	40.605779	-74.187256

```
[73]: for col in required_column:
    print(cluster_5[col].value_counts(ascending = False))
    print("---------------------------------------------")
```

```
Italian Restaurant    1
Name: 1st Most Common Venue, dtype: int64
---------------------------------------------
Vietnamese Restaurant    1
Name: 2nd Most Common Venue, dtype: int64
---------------------------------------------
Staten Island    1
Name: Borough, dtype: int64
---------------------------------------------
```

Cluster 6

- Segment 6 are neighborhoods on Staten Island that are primary Italian Restaurants

Cluster 6

```
[73]: cluster_6 = nyc_merged.loc[nyc_merged['Cluster Labels'] == 6, nyc_merged.columns[1:12]]
      cluster_6.head(5)
```

[73]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
40	Butler Manor	Italian Restaurant	Asian Restaurant	BBQ Joint	Chinese Restaurant	Fried Chicken Joint	Staten Island	40.506082	-74.229504
152	Lighthouse Hill	Italian Restaurant	Yemeni Restaurant	Fried Chicken Joint	Eastern European Restaurant	Egyptian Restaurant	Staten Island	40.576506	-74.137927
271	Tottenville	Italian Restaurant	Mexican Restaurant	Asian Restaurant	Pizza Place	Chinese Restaurant	Staten Island	40.505334	-74.246569

```
[74]: for col in required_column:
          print(cluster_6[col].value_counts(ascending = False))
          print("--------------------------------------------")
```

```
Italian Restaurant    3
Name: 1st Most Common Venue, dtype: int64
--------------------------------------------
Yemeni Restaurant     1
Mexican Restaurant    1
Asian Restaurant      1
Name: 2nd Most Common Venue, dtype: int64
--------------------------------------------
Staten Island    3
Name: Borough, dtype: int64
--------------------------------------------
```

Cluster 7

Cluster 7

```
[75]: cluster_7 = nyc_merged.loc[nyc_merged['Cluster Labels'] == 7, nyc_merged.columns[1:12]]
      cluster_7.head(5)
```

[75]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
10	Battery Park City	American Restaurant	Italian Restaurant	Mediterranean Restaurant	Seafood Restaurant	Pizza Place	Manhattan	40.711932	-74.016869
27	Boerum Hill	French Restaurant	BBQ Joint	American Restaurant	Mexican Restaurant	New American Restaurant	Brooklyn	40.685683	-73.983748
35	Brooklyn Heights	American Restaurant	Pizza Place	French Restaurant	Vietnamese Restaurant	Italian Restaurant	Brooklyn	40.695864	-73.993782
43	Carnegie Hill	American Restaurant	Pizza Place	Taco Place	French Restaurant	German Restaurant	Manhattan	40.782683	-73.953256
44	Carroll Gardens	French Restaurant	Italian Restaurant	American Restaurant	BBQ Joint	Pizza Place	Brooklyn	40.680540	-73.994654

```
[77]: for col in required_column:
          print(cluster_7[col].value_counts(ascending = False))
          print("--------------------------------------------")
```

```
American Restaurant    14
Pizza Place             1
Name: 1st Most Common Venue, dtype: int64
--------------------------------------------
Pizza Place              4
Mexican Restaurant       3
Italian Restaurant       3
Seafood Restaurant       1
Chinese Restaurant       1
American Restaurant      1
Fast Food Restaurant     1
Vietnamese Restaurant    1
Name: 2nd Most Common Venue, dtype: int64
--------------------------------------------
Manhattan        7
Brooklyn         4
Staten Island    2
Queens           1
Bronx            1
Name: Borough, dtype: int64
--------------------------------------------
```

Discussion

- Three analysis were down to understand the clusters:

1. Count of Borough
2. Count of 1st Mot Common Venue
3. Count of 2nd Most Common Venue

As reference on slide 9, Pizza was the most common venue amongst all of the clusters. We did discover that there seems to be a variety of other venues associated with the clusters with pizza. Staten Island seemed to have the most diverse clusters.

High Level Here is How the Clusters Looked:

- 0 - Pizza/Fast Food – Queens & Brooklyn
- 1 – Caribbean Cuisines – Brooklyn & Queens
- 2 – Italian/Pizza – Staten Island
- 3 – Italian/Pizza/American – Mahanttan, Brooklyn, & Queens
- 4 – Pizza/Italian – Staten Island & The Bronx
- 5 – Italian/Vietnamese - Staten Island
- 6 – Mix of Cuisines – Staton Island
- 7 – American – Manhattan *& Brooklyn

Conclusion

By applying the cluster algorithm, K-means, to a multi-dimensional dataset, a very detail result set can be created to help us understand and visualization the neighborhoods and culture in NYC based on the type of cuisines venues there are. Pizza and Italian were very most dominate in NYC but there were also a lot of Asian and Caribbean venues as well. That speaks to the diversity of the city.

The results from the project could be improved by maybe incorporating an API from Yelp! to get customer feedback and ratings of venues into this dataset. This would help the stakeholders get an idea of how good a place is based on the average customer review and rating. This data could also be used by Local Government in NYC to figure out which neighborhoods are dominated by what type of culture.