Capstone Project
The Battle of Neighborhoods

RYAN COOLEY

02.26.2020

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:

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

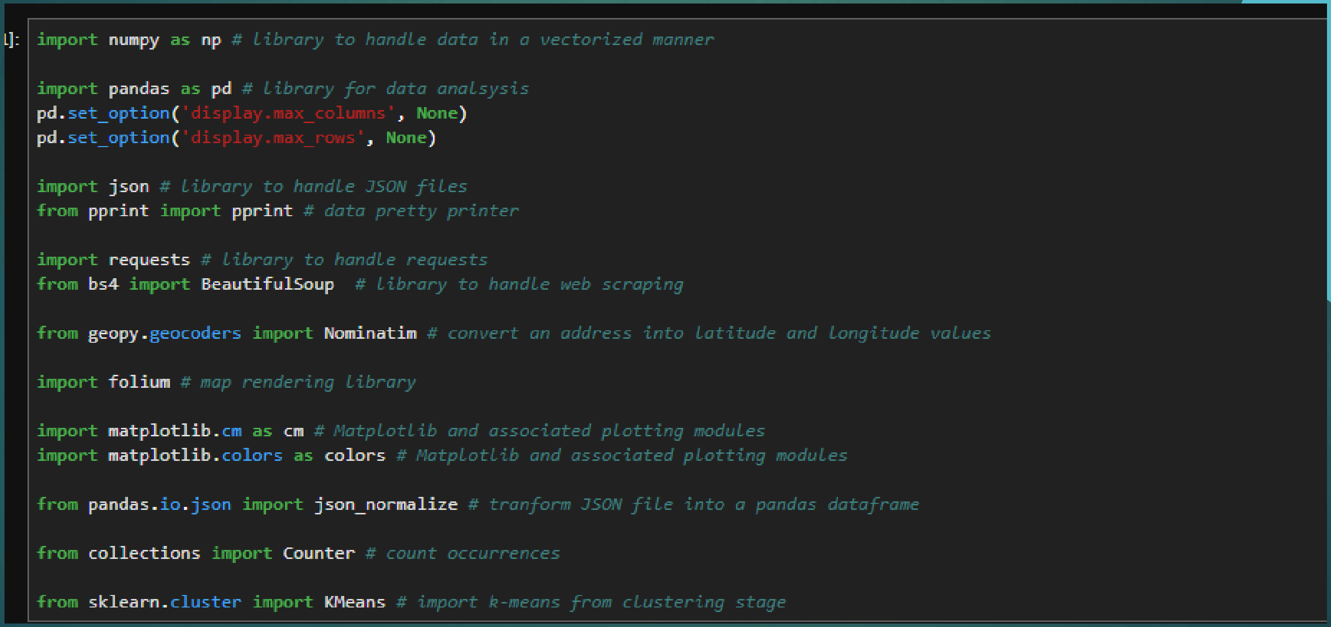
  The data set will be our base neighborhood data set to cross reference against the Foursquare API venue data

- **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.

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

Methodology: Loading Dependencies

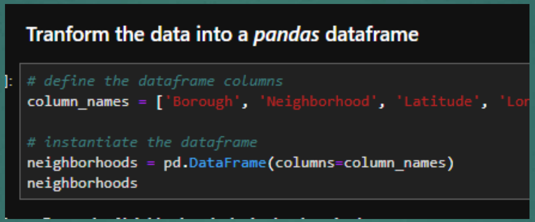
- We will first download all the dependencies:



```
1]: 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
```

Methodology: Transforming and Exploring the Data Set

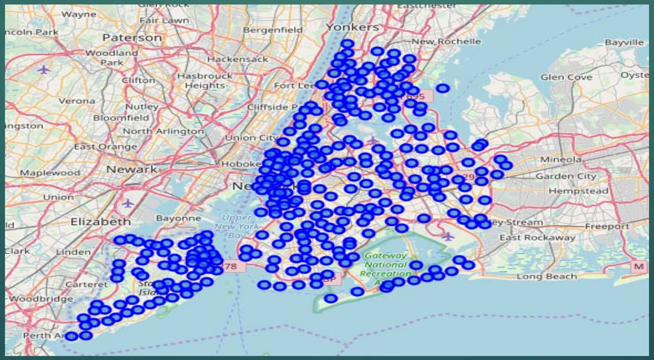
1. We upload the JSON file and transform it in a Pandas Data Frame



Tranform the data into a *pandas* dataframe

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

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

2. When then use the Geopy library to get the Lat and Long and create map:



Methodology: Appending Foursquare data to the NYC Data set

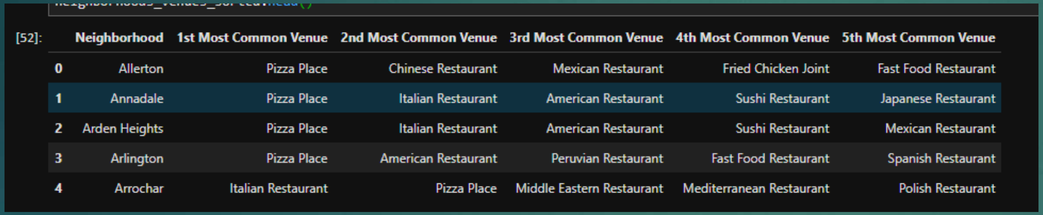
We will take the following steps to append the data set:

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

Methodology: K-Means Clustering

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

1. We will first explore and group the data set:

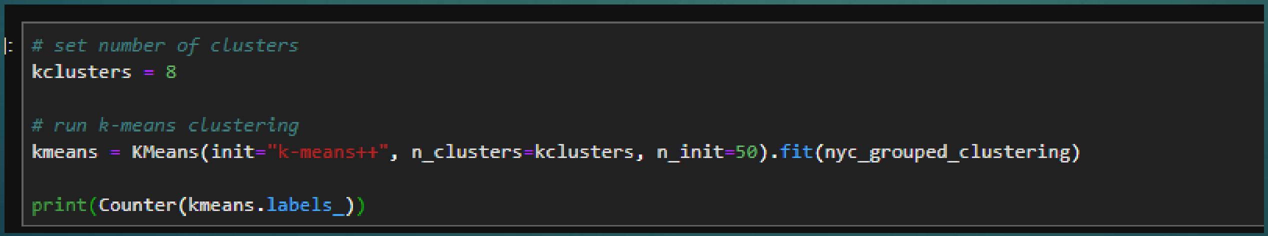


[52]:

	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

2. We will then use two different methods to evaluate how much clusters we need (Elbow and Silhouette Methods for Optimal k (see next slide for detail):

3. Once we pick a method for the optimal number K, we will run the model:

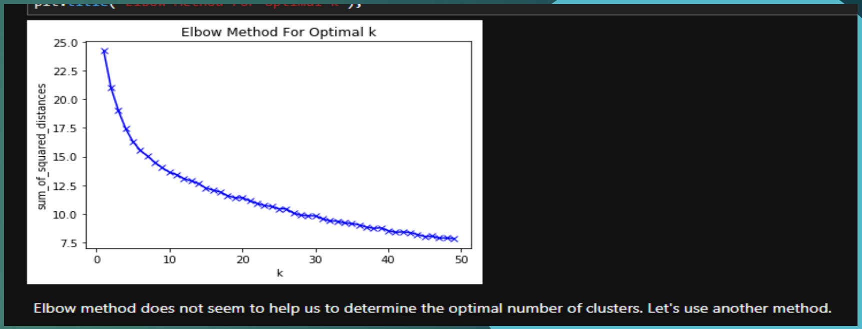


```
# set number of clusters
kclusters = 8

# run k-means clustering
kmeans = KMeans(init="k-means++", n_clusters=kclusters, n_init=50).fit(nyc_grouped_clustering)

print(Counter(kmeans.labels_))
```

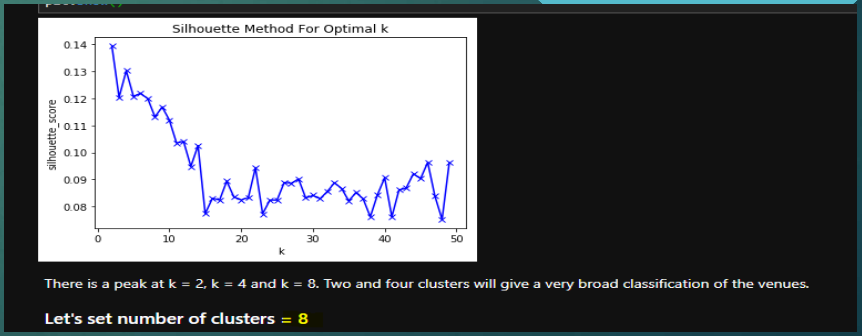
Methodology: Custer Evaluation

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

We will set the cluster number to 8 based on Silhouette Method base on the data set



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



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

Results: High Level Clusters (Segments)

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

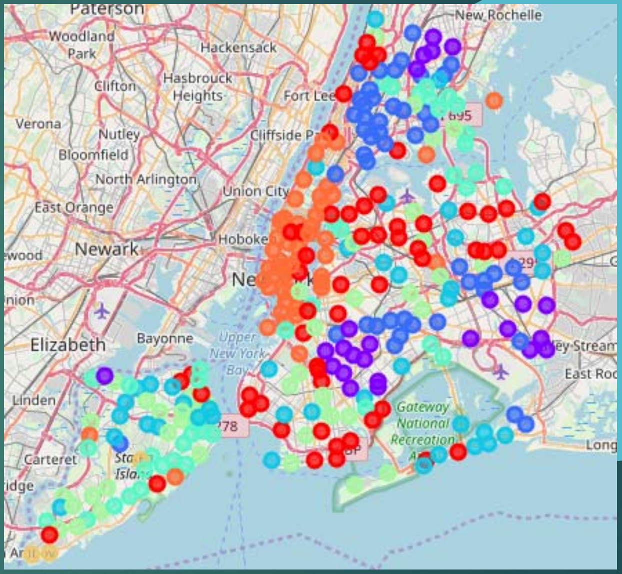
3 – Italian/Pizza/American – Manhattan, Brooklyn, & Queens

4 – Pizza/Italian – Staten Island & The Bronx

5 – Italian/Vietnamese - Staten Island

6 – Mix of Cuisines – Staten Island

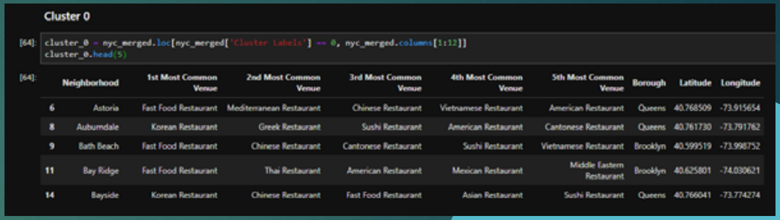
7 – American – Manhattan *& Brooklyn



The Next Slides will break down the clusters or segments we created

Results: 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

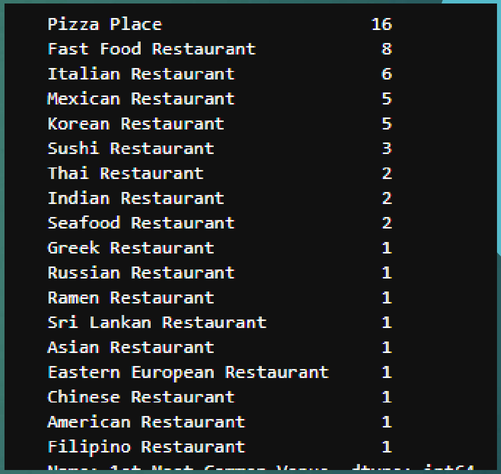


Cluster 0

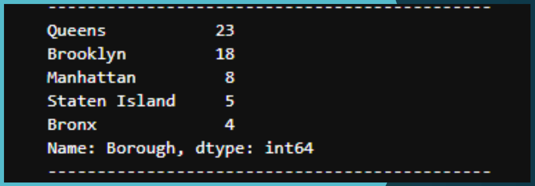
```
[64]: cluster_0 = nyc_merged.loc[nyc_merged['Cluster Labels'] == 0, nyc_merged.columns[1:12]]
      cluster_0.head(5)
```

[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



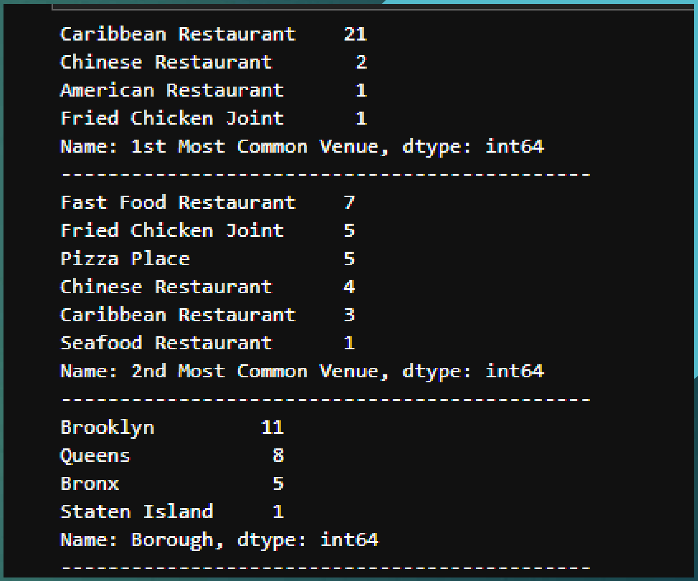
```
Pizza Place                     16
Fast Food Restaurant             8
Italian Restaurant               6
Mexican Restaurant               5
Korean Restaurant                5
Sushi Restaurant                 3
Thai Restaurant                  2
Indian Restaurant                2
Seafood Restaurant               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
-------------------------------------------
```

Results: Cluster 1

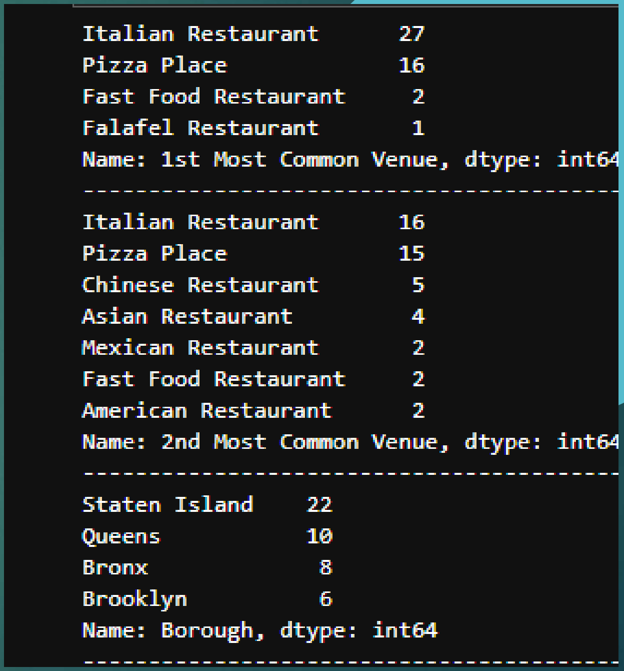
- Segment 1 is a mostly neighborhoods that are Caribbean.
- Most of these neighborhoods reside in Brooklyn and Queen



```
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
---------------------------------------------
```

Results: Cluster 2

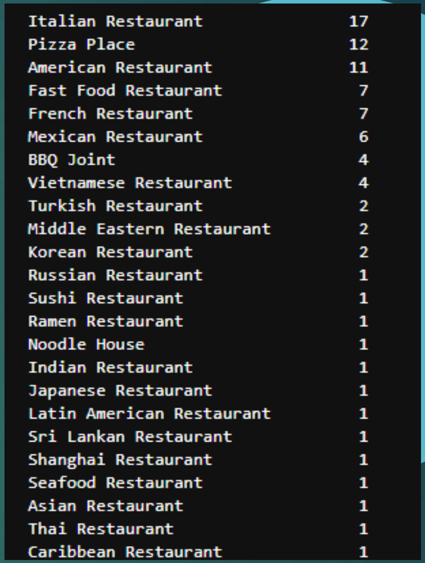
- Segment 2 are mostly a mix of Italian/Pizza
- Most reside in Staten Island



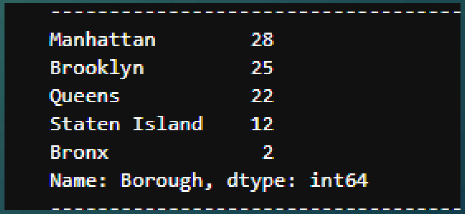
```
Italian Restaurant      27
Pizza Place             16
Fast Food Restaurant     2
Falafel Restaurant       1
Name: 1st Most Common Venue, dtype: int64
----------------------------------------
Italian Restaurant      16
Pizza Place             15
Chinese Restaurant       5
Asian Restaurant         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
----------------------------------------
```

Results: 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.



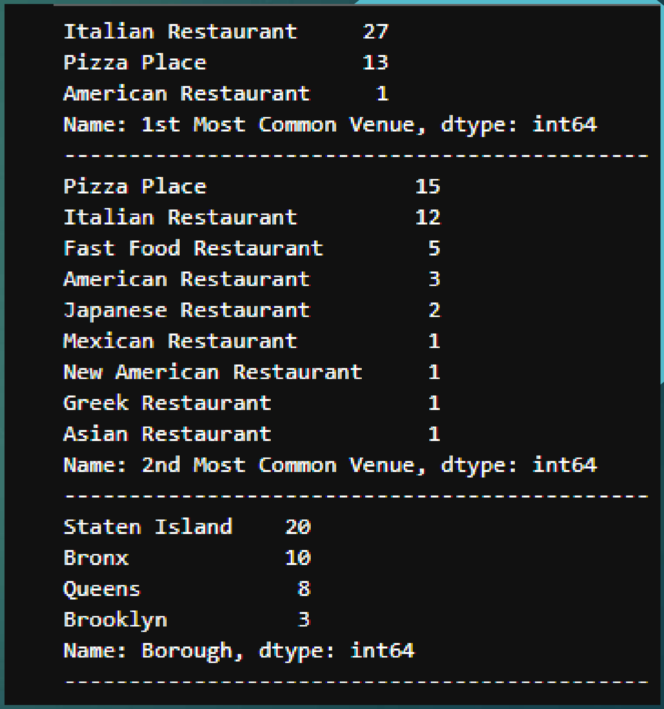
```
Italian Restaurant              17
Pizza Place                     12
American Restaurant             11
Fast Food Restaurant             7
French Restaurant                7
Mexican Restaurant               6
BBQ Joint                        4
Vietnamese Restaurant            4
Turkish Restaurant               2
Middle Eastern Restaurant        2
Korean Restaurant                2
Russian Restaurant               1
Sushi Restaurant                 1
Ramen Restaurant                 1
Noodle House                     1
Indian Restaurant                1
Japanese Restaurant              1
Latin American Restaurant        1
Sri Lankan Restaurant            1
Shanghai Restaurant              1
Seafood Restaurant               1
Asian Restaurant                 1
Thai Restaurant                  1
Caribbean Restaurant             1
```



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

Results: Cluster 4

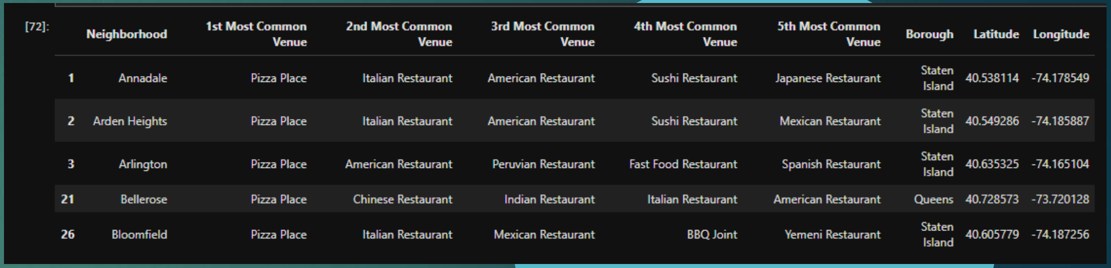
- Segment 4 are neighborhoods that are heavy Italian Restaurants and Pizza Places
- Most are located in Staten Island and the Bronx



```
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
---------------------------------------------
```

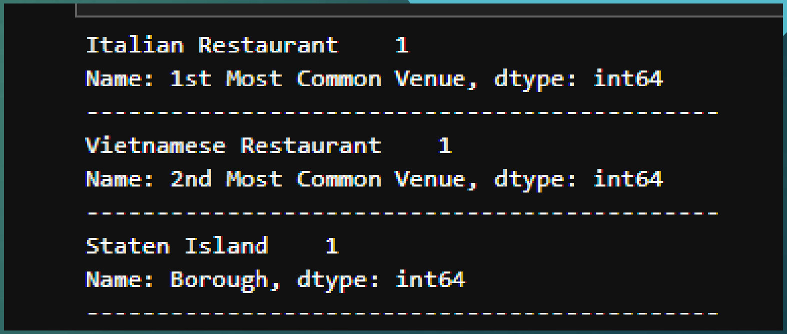
Results: Cluster 5

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



[72]:

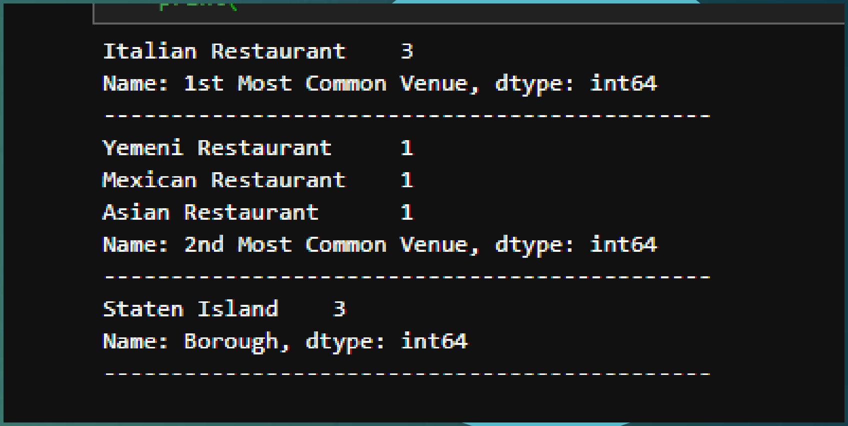
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
1	Annadale	Pizza Place	Italian Restaurant	American Restaurant	Sushi Restaurant	Japanese Restaurant	Staten Island	40.538114	-74.178549
2	Arden Heights	Pizza 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



```
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
-----------------------------------------------
```

Results: Cluster 6

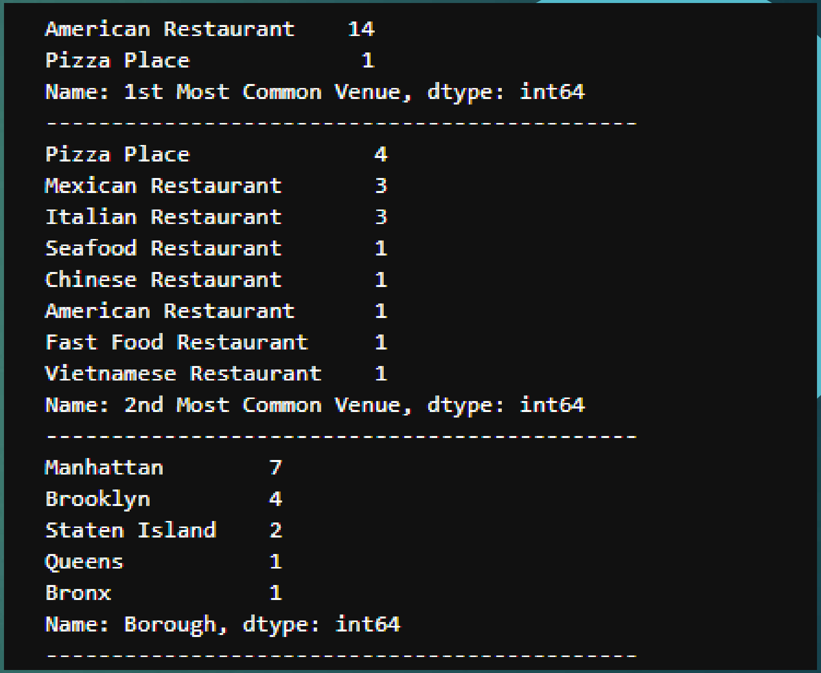
- Segment 6 are neighborhoods on Staten Island that are primary Italian Restaurants



```
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
---------------------------------------------
```

Results: Cluster 7

- Segment 7 are neighborhoods that a majority of restaurants that are American
- Manhattan has the most at 7



```
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
---------------------------------------------
```

Results: High Level

We segments the neighborhoods into 8 different segments depending on what type of cuisine was most common:

0 - Caribbean

1 - Chinese

2 - Italian

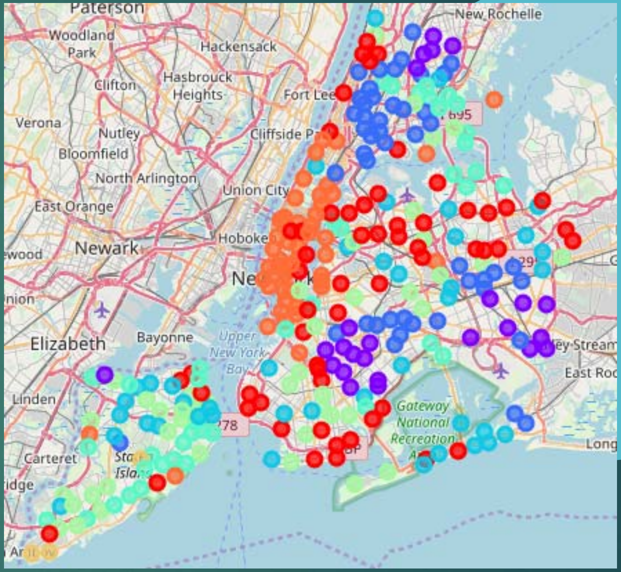
3 – Italian American

4 – Pizza

5 – Mix of Cuisines

6 – Fast Food

7 - American



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.

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.