# 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>

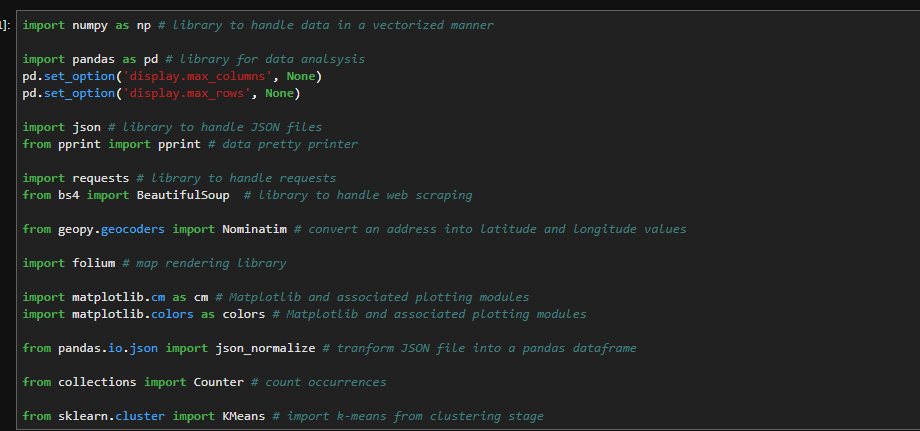
The data set will be our base neighborhood data set to cross reference against the Foursquare API venue data

1. **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.
2. **Geophy** Library in Python: this will help us get the Lat and Long of the NYC data set

## Methodology:

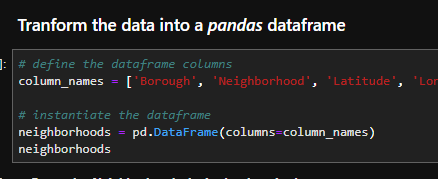
### Loading Dependciens

We first most load the following libraries into Jupyter Notebook



### Transforming and Exploring the NYC Data Set

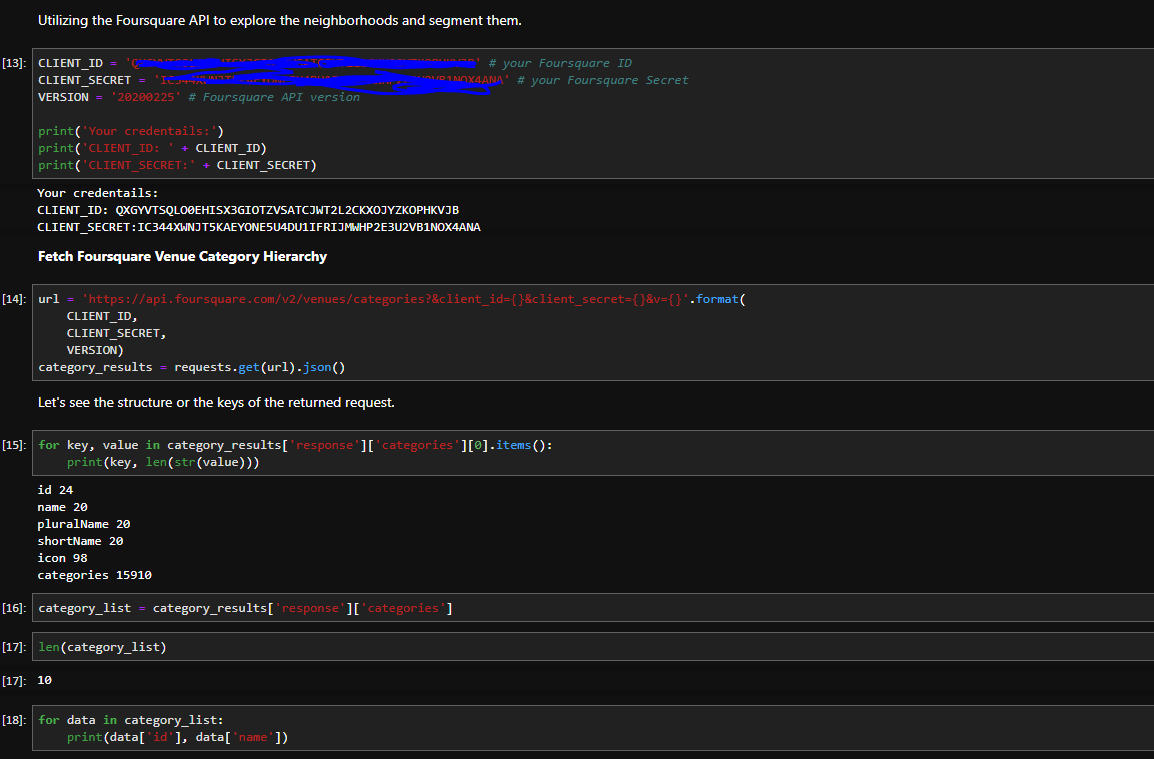
We upload the NYC data set and run a couple lines of code to transform the data. We then use Geopy to get the Latitude and Longitude of each borough and plot on a map:

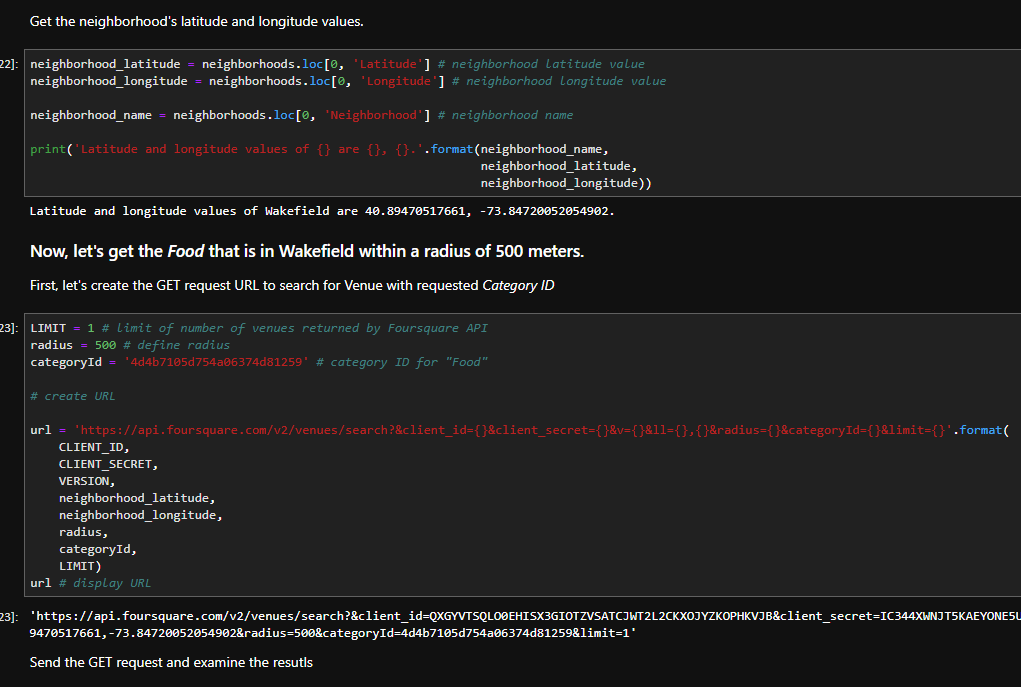


### Appending the Foursquare data to the NYC Data Set

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





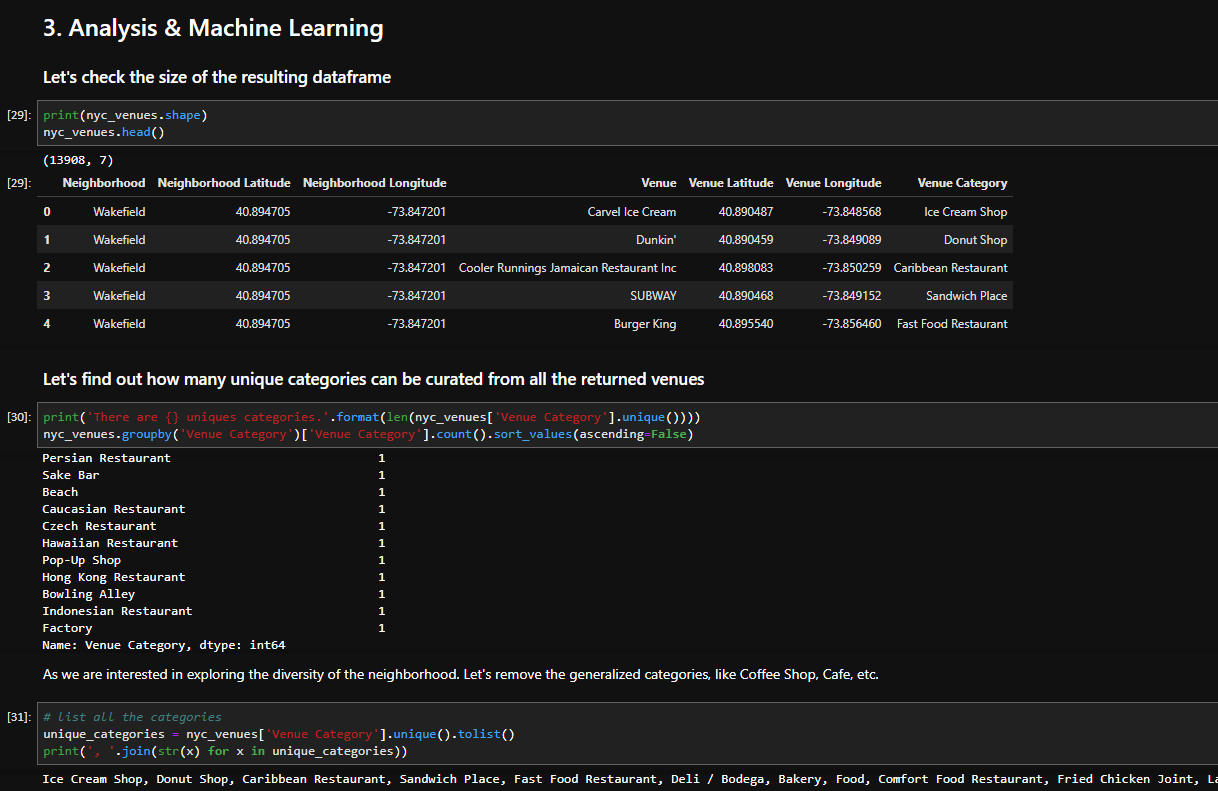
### Model Selection

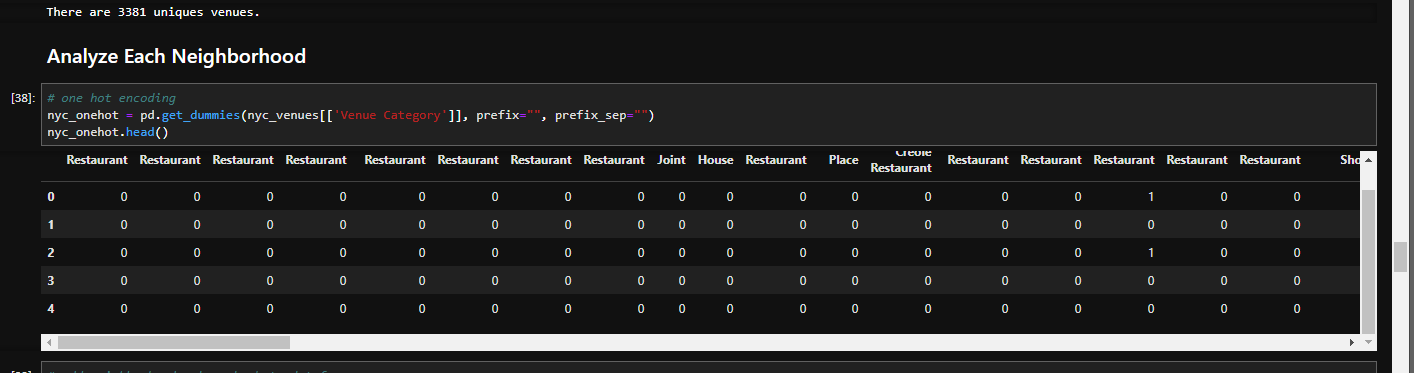
We will chose the K-Means Clustering Algorithm to help build segments for the neighborhoods based on types of cuisines in that particular neighborhood.

**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:

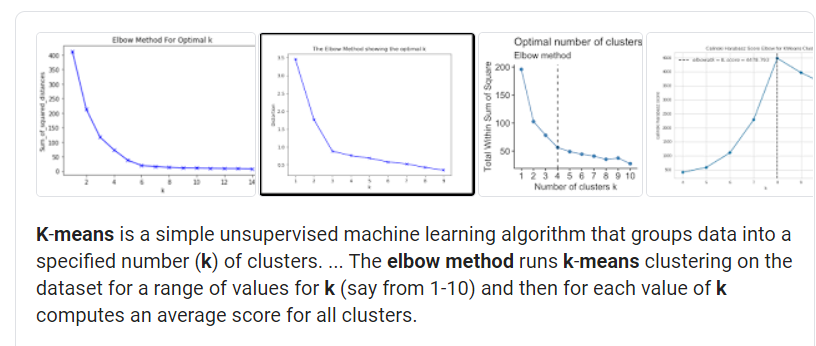


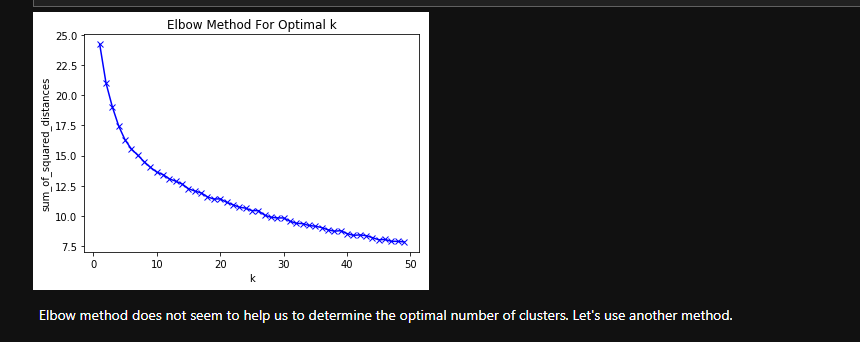




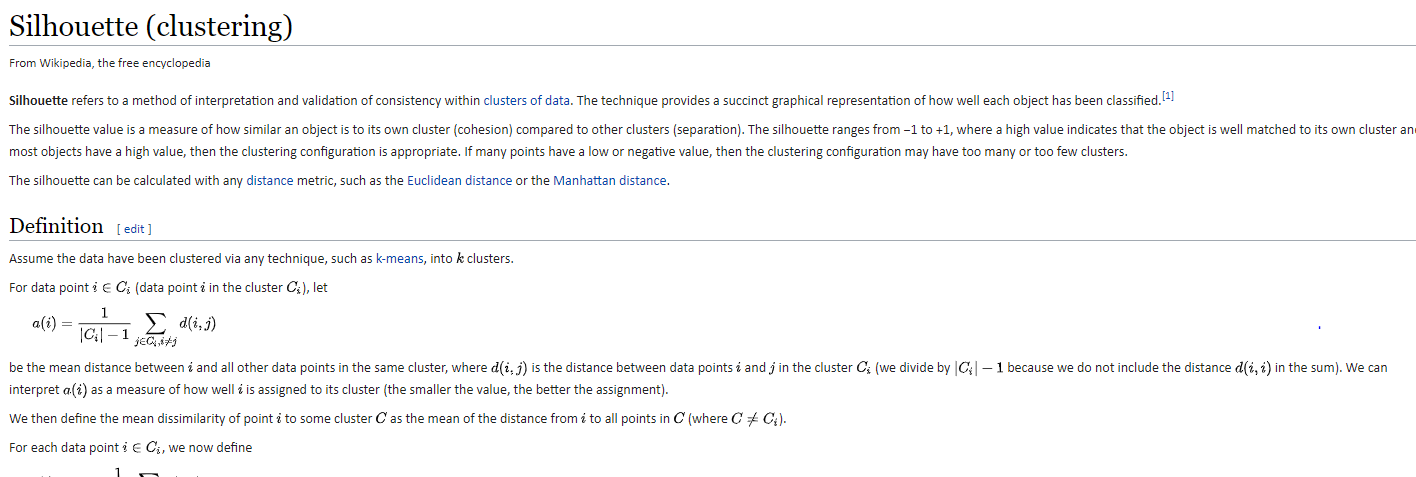
### Cluster Evaluation

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.

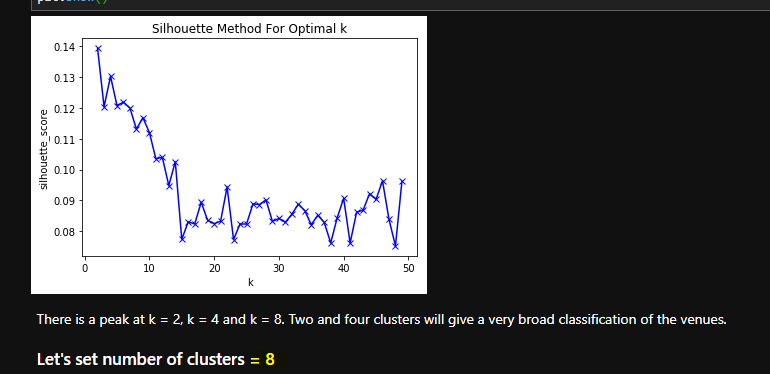




1. **Silhouette Method** - value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation)



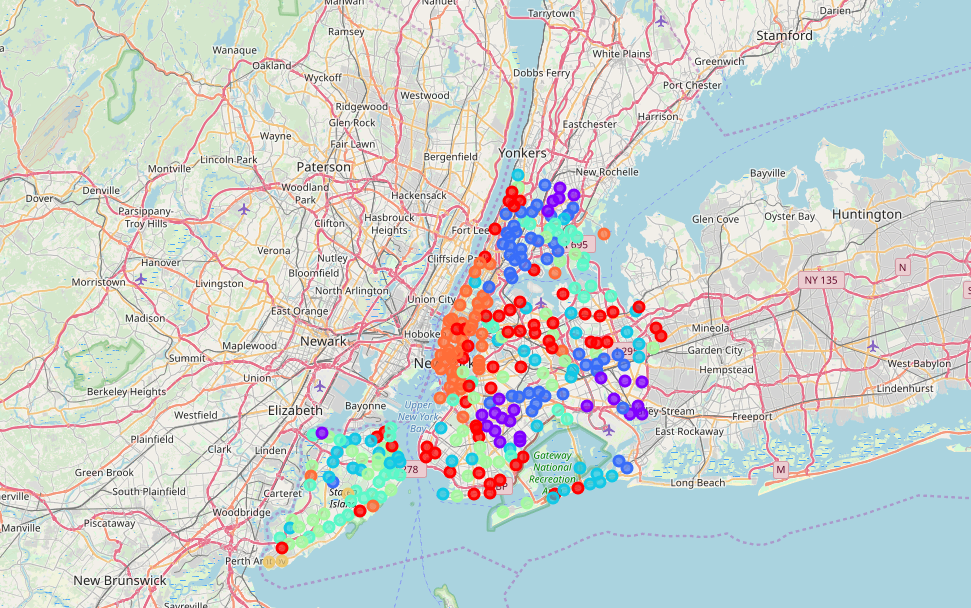
Source: <https://en.wikipedia.org/wiki/Silhouette_(clustering)>



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

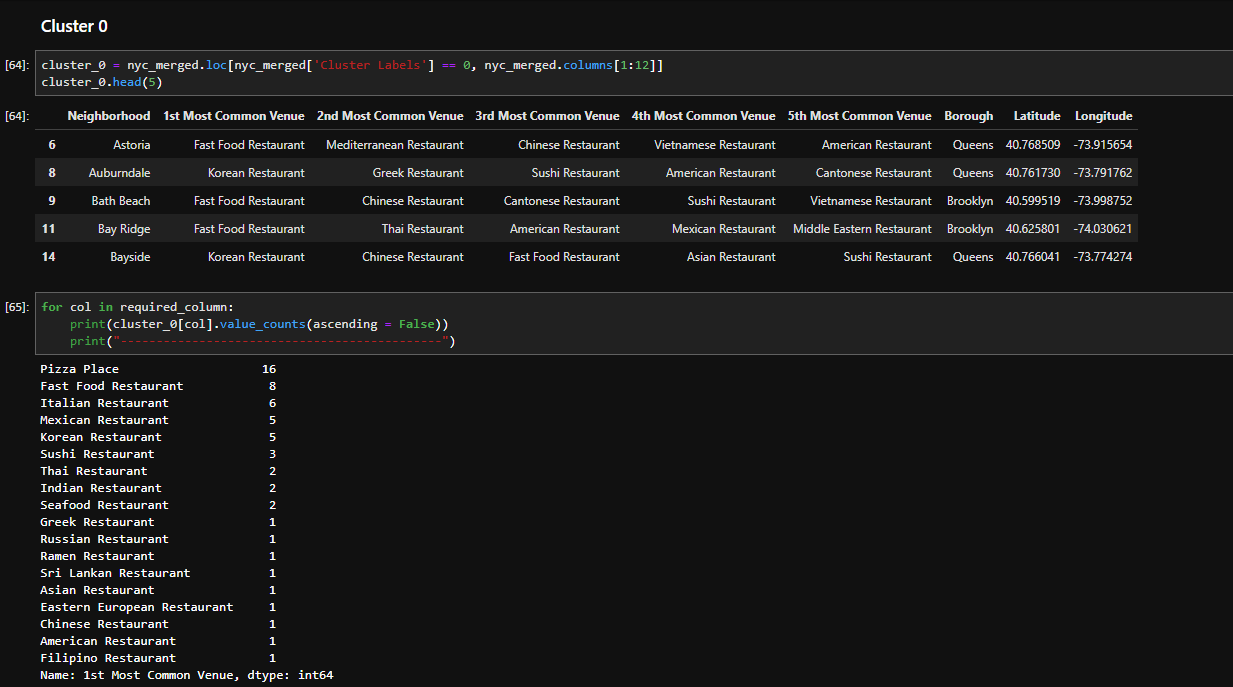
5 – Italian/Vietnamese - Staten Island

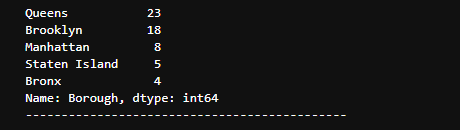
6 – Mix of Cuisines – Staten Island

7 – American – Manhattan \*& Brooklyn

### Cluster 0

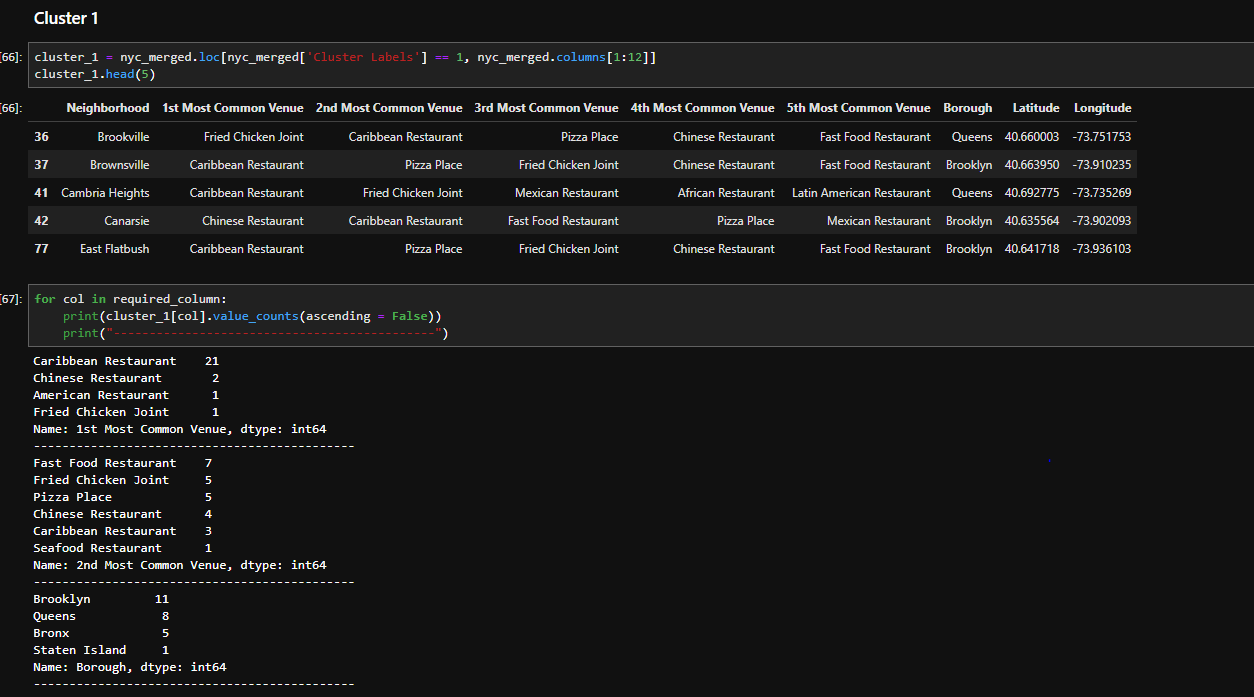
* Segment 0 are neighborhoods that had a major of restaurants that are Pizza Place and Fast Food
* Most of the neighborhoods reside in Brooklyn and Queens





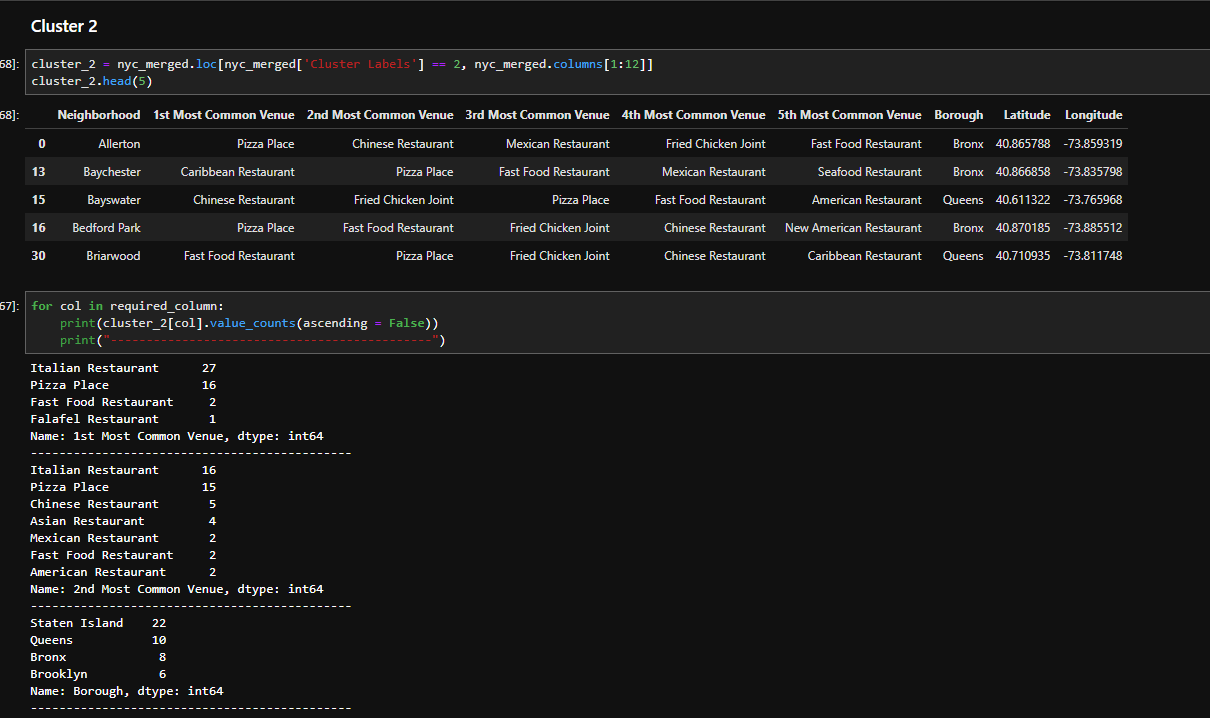
### Cluster 1

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



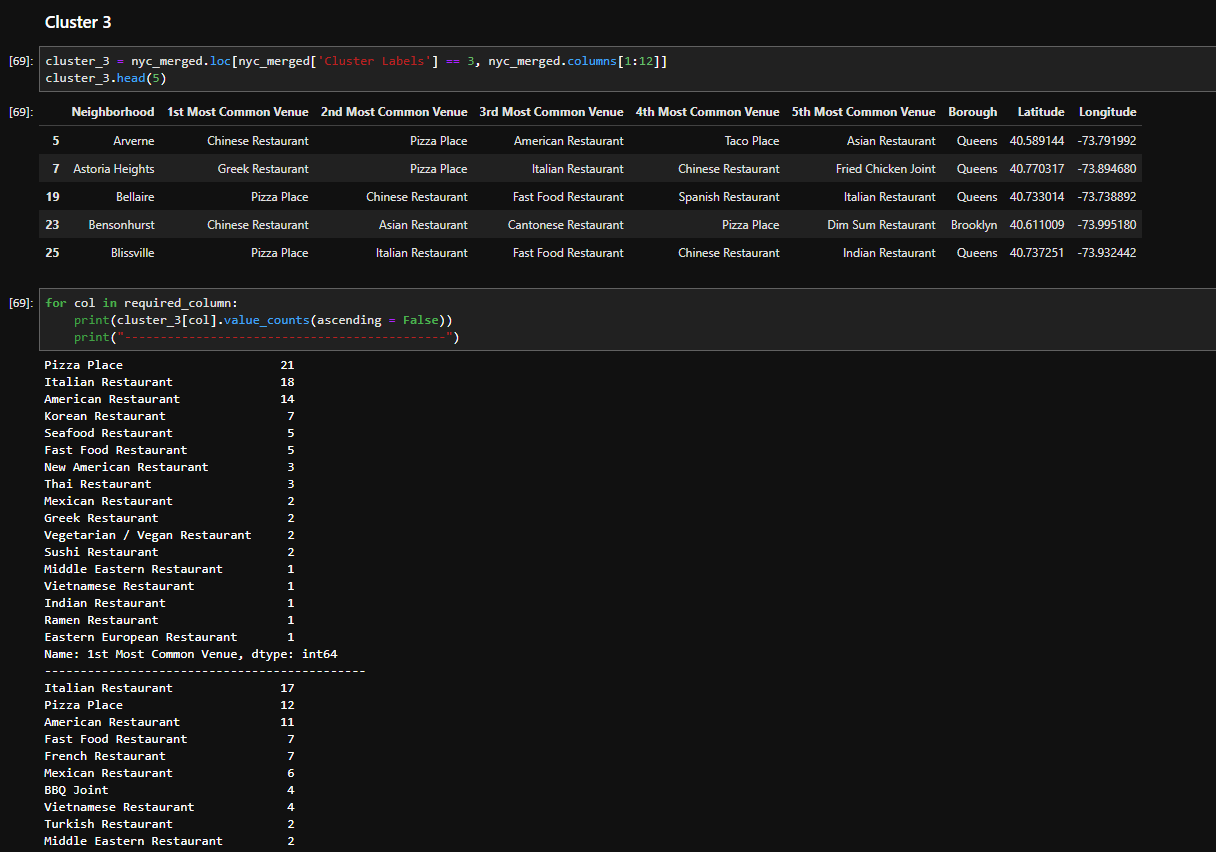
### Cluster 2

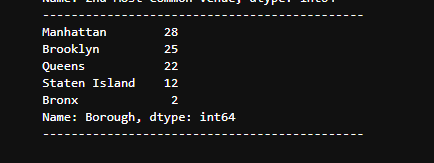
* Segment 2 are mostly a mix of Italian/Pizza
* Most reside in Staten Island



### Cluster 3

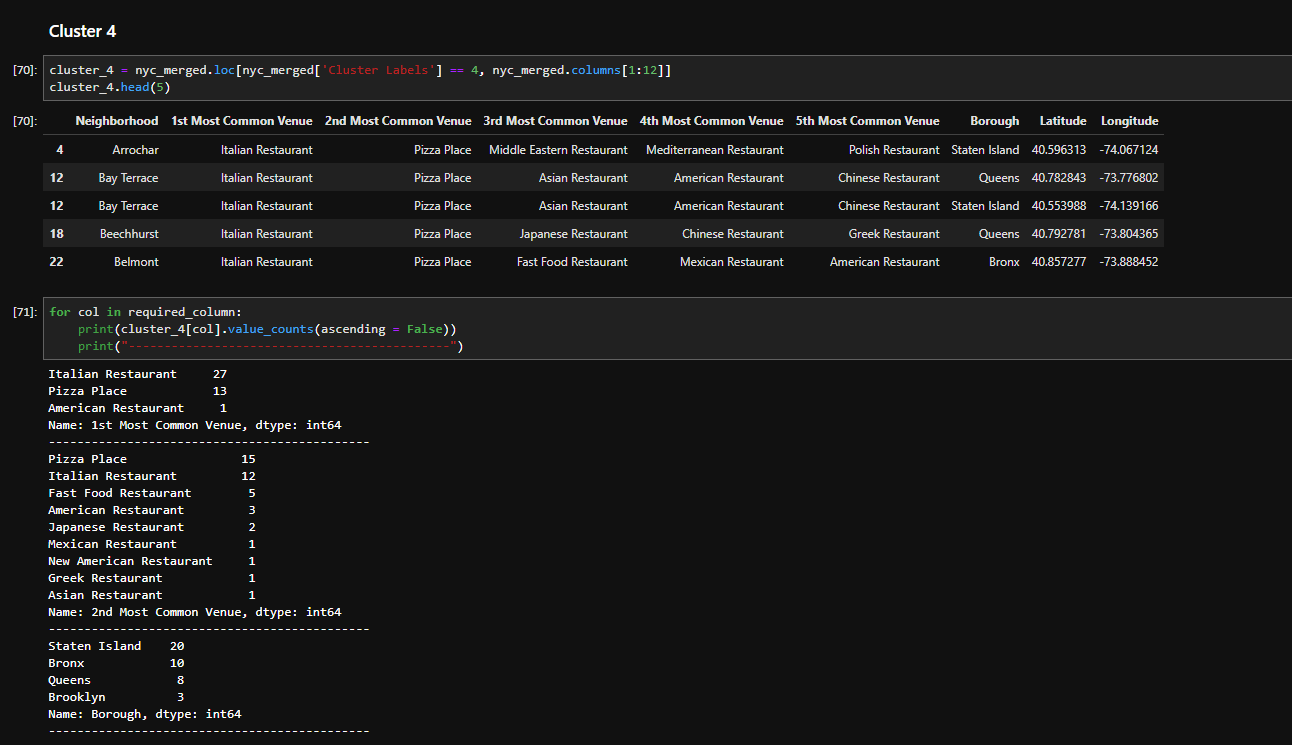
* Segment 3 are heavy Italian, Pizza, and American
* This is our largest segment with a majority of neighborhoods in Manhattan, Brooklyn, and Queens.





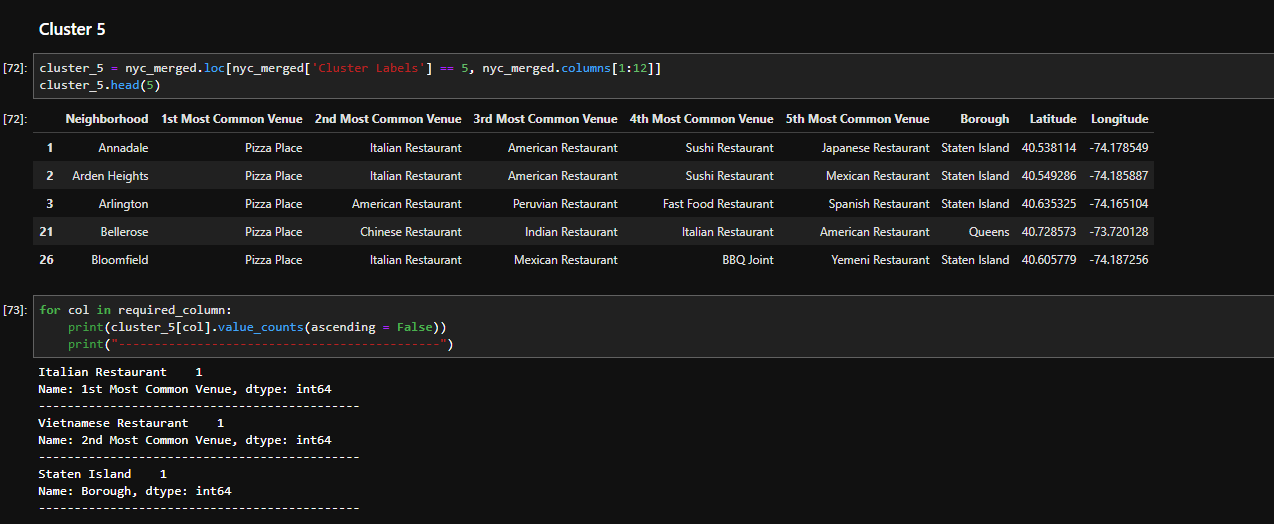
### Cluster 4

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



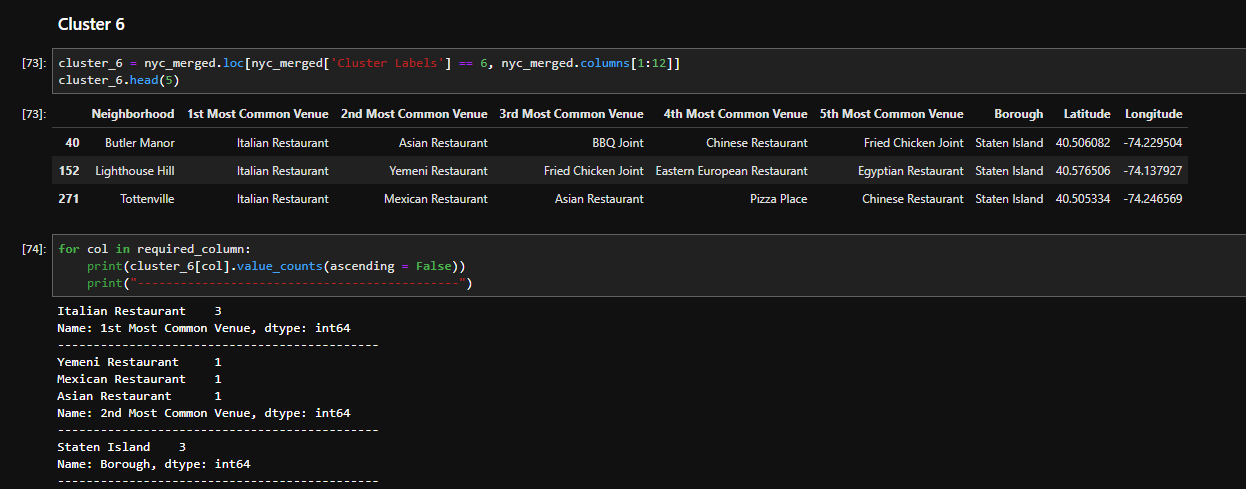
### Cluster 5

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

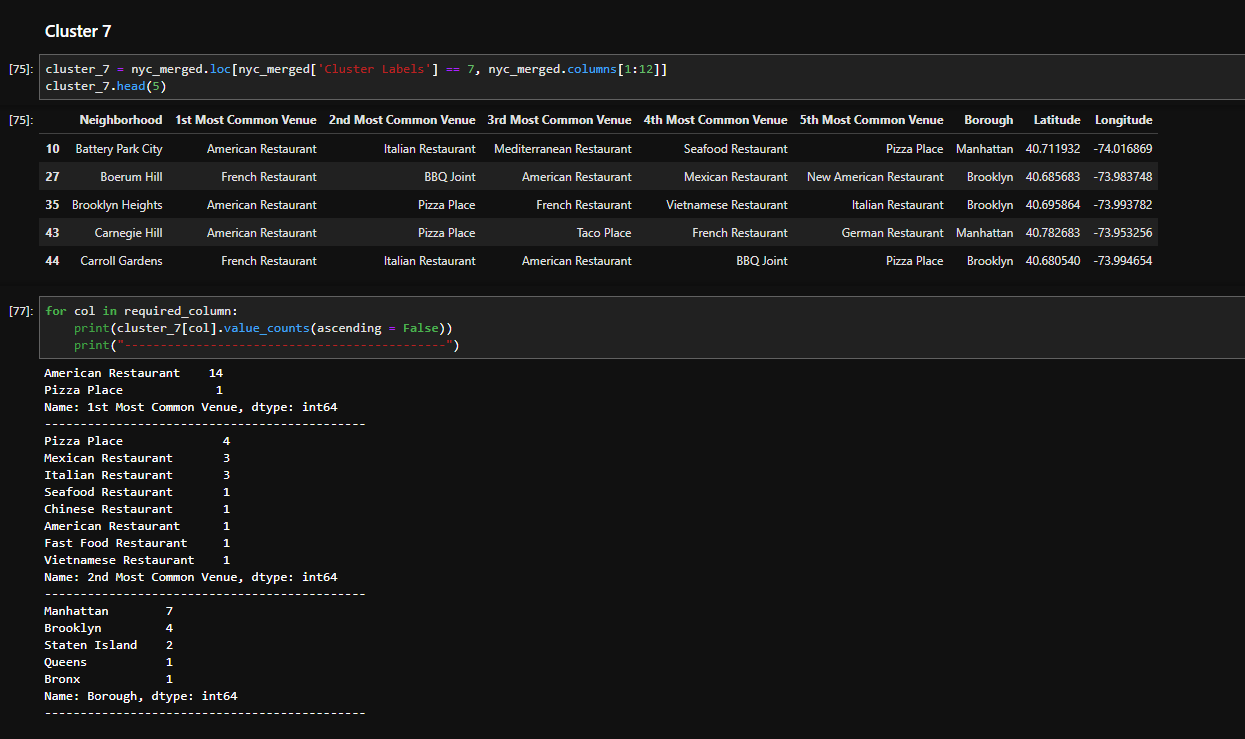


### Cluster 6

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



### Cluster 7



## 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. This data could also be used by Local Government in NYC to figure out which neighborhoods are dominated by what type of culture.