

IBM Data Science Professional Certificate Course

Capstone Project - The Battle of the Neighborhoods

Finding out the best neighborhoods to open up a new restaurant in New York City, NY

Mohammed H. Adi

6 March 2021

## **Introduction**

This project is about using the data science skills and techniques that I have learned and developed by taking the IBM Data Science Professional Certificate course. Throughout the project I will be implementing various data science tools and methods and applying them on a real world business problem. This is the capstone project for the IBM Data Science Professional Certificate course, which marks the end of the course and the beginning of my data science path.

## **Business Problem**

This notebook will try to solve a theoretical business problem using Data Science and Machine Learning, specifically using k-means Clustering. The Business problem is finding the best neighborhoods for existing and new F&B businesses to open up a restaurant in all of New York City's (NYC) 5 boroughs. This notebook will focus on New York City as it is a metropolitan city, where the F&B sector is highly competitive. Many F&B businesses have a very difficult time studying and researching the best location to open up their restaurant. This notebook should give a brief insight on the best neighborhoods to open up a new restaurant in and what type of cuisine would most suit these neighborhoods.

## **The Data**

The data that will be used in the report will be from the following sources:

1. The list of Boroughs and Neighborhoods from the 'newyork\_data.json' file from the IBM server.
2. Geo-coordinates of New York City's Neighborhoods using the Nominatim API from Python's Geopy client.
3. Venues data using the Foursquare API.

## **Methodology**

We will begin our analysis by installing/importing the necessary libraries namely:

1. Numpy, which is a library used to handle data in a vectorized manner.
2. Pandas, which is a library used for data manipulation and analysis.
3. Json, which is a library used to handle JSON files.
4. Geopy, which is a library used to convert an address into geo-coordinates.
5. Matplotlib, which is a library used for plotting graphs.
6. Scikit-learn, which is a library used for Machine Learning.
7. Folium, which is a library used for creating maps.

The data will be acquired from a json file and will be cleaned up and rearranged using Python's Pandas library

The Foursquare API, will be used to acquire the top venues of the neighborhoods, and results will be shown in a pandas dataframe and then will be merged with original data and will be cleaned up to be used for machine learning.

The methodology that will be implemented in the analysis will depend on an unsupervised machine learning clustering algorithm, and using the k-means method to create a k number of clusters of NYC's neighborhoods. To find the optimal number of clusters we will begin by using the Elbow Method, and eventually the Silhouette Score for a more accurate choice.

Importing the libraries:

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

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

import json # library to handle JSON files

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

import requests # library to handle requests
from pandas.io.json import json_normalize # tranform JSON file into a pandas dataframe

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

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

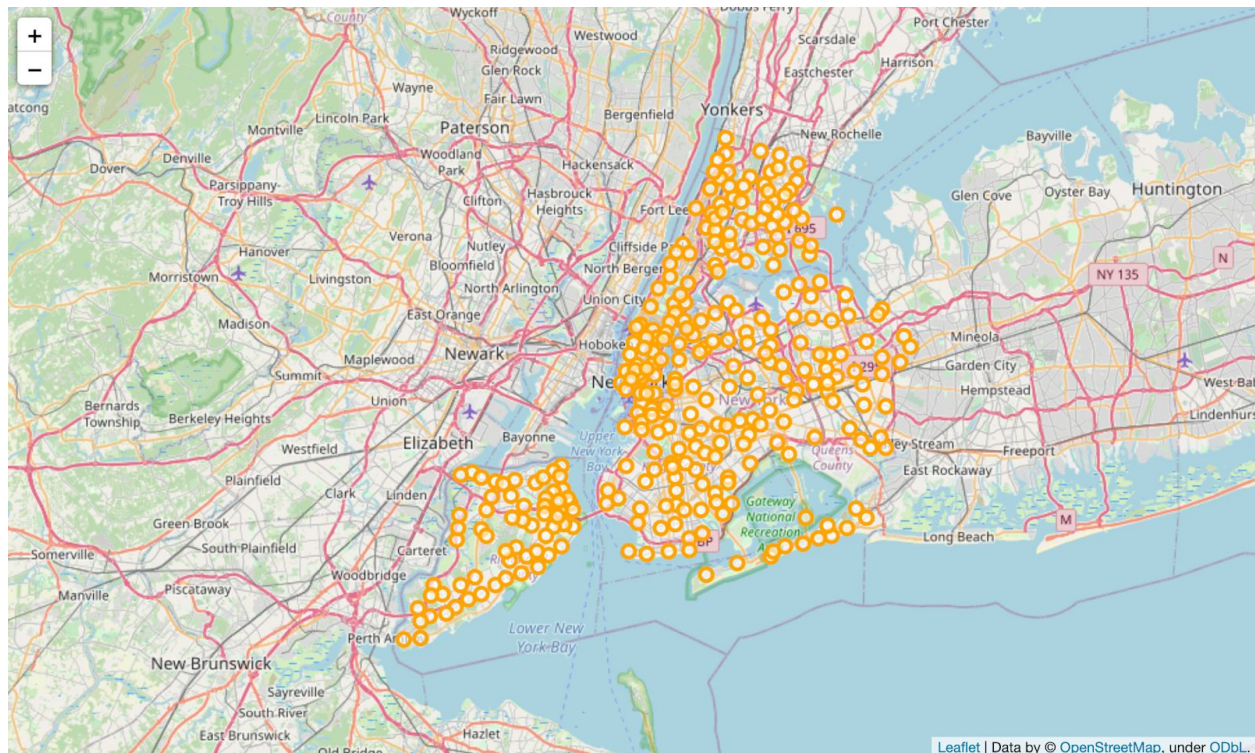
!pip install folium
import folium # map rendering library

print('Libraries imported.')
```

The data will be acquired from a json file and will be cleaned up and rearranged using Python's Pandas library with a shape of (5, 306). The top 10 rows are shown below:

	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
5	Bronx	Kingsbridge	40.881687	-73.902818
6	Manhattan	Marble Hill	40.876551	-73.910660
7	Bronx	Woodlawn	40.898273	-73.867315
8	Bronx	Norwood	40.877224	-73.879391
9	Bronx	Williamsbridge	40.881039	-73.857446

A map of NYC's neighborhoods is created using Folium.



The Foursquare API, will be used to acquire the top venues of the neighborhoods, a function will be defined that will extract the category of each venue retrieved from foursquare, and results will be shown in a pandas dataframe and then will be merged with original data and will be cleaned up to be used for machine learning. The top 5 results for one neighborhood are shown in the dataframe below.

	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	Ripe Kitchen & Bar	40.898152	-73.838875	Caribbean Restaurant
2	Wakefield	40.894705	-73.847201	Ali's Roti Shop	40.894036	-73.856935	Caribbean Restaurant
3	Wakefield	40.894705	-73.847201	Jackie's West Indian Bakery	40.889283	-73.843310	Caribbean Restaurant
4	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy

One Hot Encoding will be used to analyze the neighborhoods and the data will be grouped and mean of occurrence frequency of each category will be calculated:

	Neighborhood	Zoo Exhibit	ATM	Accessories Store	Adult Boutique	Adult Education Center	Afghan Restaurant	African Restaurant	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Amphitheater	Animal Shelter
0	Allerton	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.016667	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	0.117647	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	0.000000	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	0.043478	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	0.000000	0.0	0.0

Now we find the top 5 most common venues in each neighborhood, a sample of two neighborhoods is shown below:

```

----Erasmus----
              venue  freq
0  Caribbean Restaurant  0.11
1             Deli / Bodega  0.07
2          Discount Store  0.07
3    Mobile Phone Shop  0.05
4           Pizza Place  0.05

----Far Rockaway----
              venue  freq
0           Pizza Place  0.09
1    Chinese Restaurant  0.09
2             Deli / Bodega  0.06
3        Sandwich Place  0.06
4    Caribbean Restaurant  0.03

```

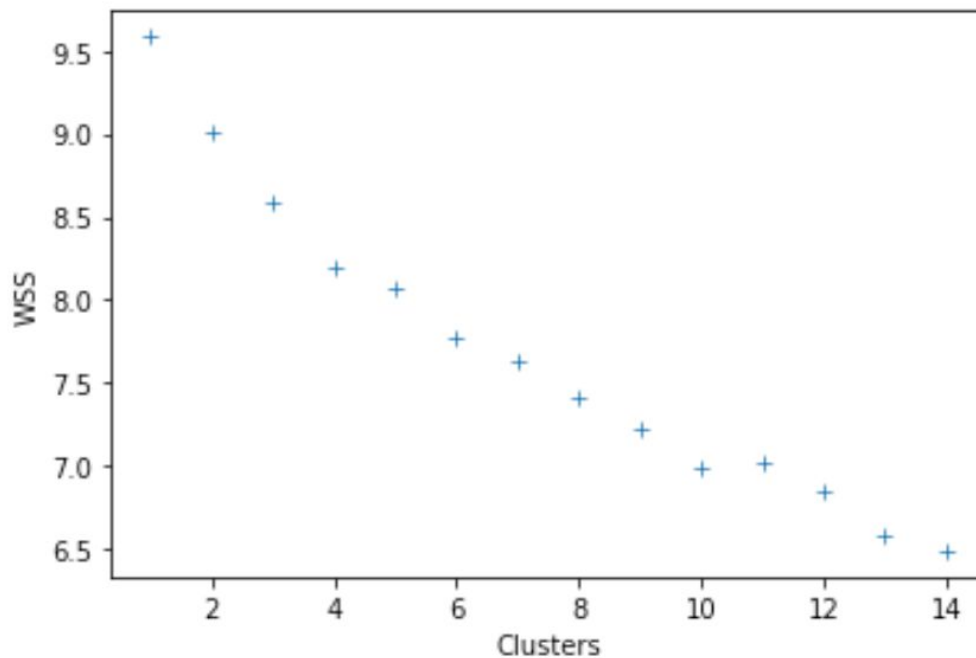
A function that returns the most common venues is defined and will be applied to all neighborhoods to get the top 10 most common venues. The head of the dataframe is shown below:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Allerton	Pizza Place	Donut Shop	Sandwich Place	Deli / Bodega	Caribbean Restaurant	Fast Food Restaurant	Mexican Restaurant	Pharmacy	Fried Chicken Joint	Food
1	Annadale	American Restaurant	Trail	Home Service	Restaurant	Pizza Place	Diner	Train Station	Park	Pub	Liquor Store
2	Arden Heights	Park	Italian Restaurant	Dog Run	Sushi Restaurant	Optical Shop	Food	Bus Stop	Spa	Gift Shop	Mexican Restaurant
3	Arlington	Bus Stop	Deli / Bodega	Hardware Store	Discount Store	American Restaurant	Check Cashing Service	Snack Place	Fast Food Restaurant	Coffee Shop	Donut Shop
4	Arrochar	Bus Stop	Italian Restaurant	Deli / Bodega	Baseball Field	Beach	Bagel Shop	Mediterranean Restaurant	Middle Eastern Restaurant	Nail Salon	Outdoors & Recreation

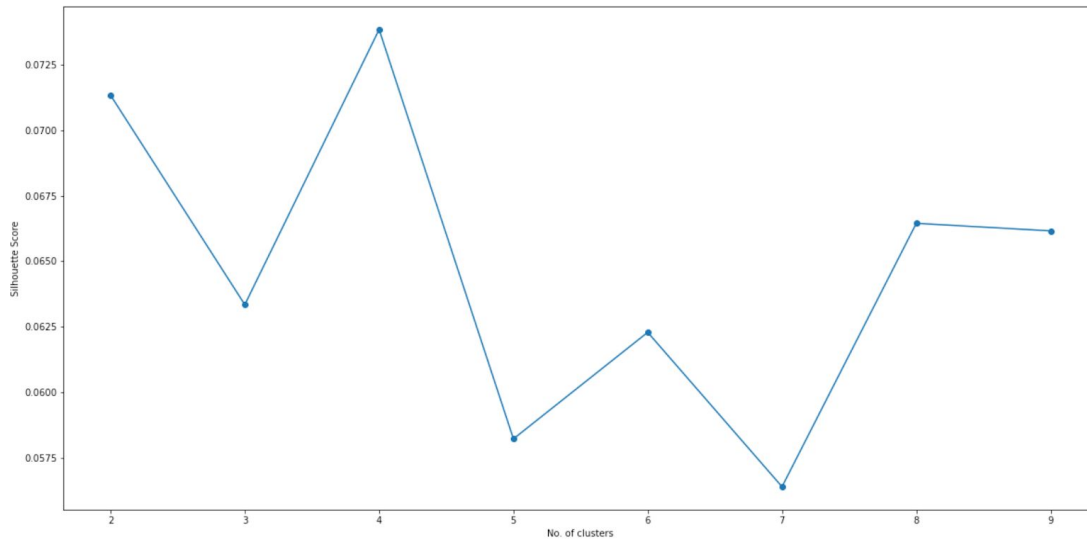
## Machine Learning:

We will be using a Clustering algorithm using the k-means technique to cluster the most common venues in NYC's neighborhoods. To find the optimal number of clusters we will first use the elbow method and create a graph of it.

The elbow method:



As we can see from the above elbow graph, it is not clear for us which is the best number of clusters, and this could happen with the elbow method. An alternative method would be the Silhouette Scores, which should give a more accurate way to select the number of clusters. The silhouette score gives a score between 0 and 1, the higher the more accurate.

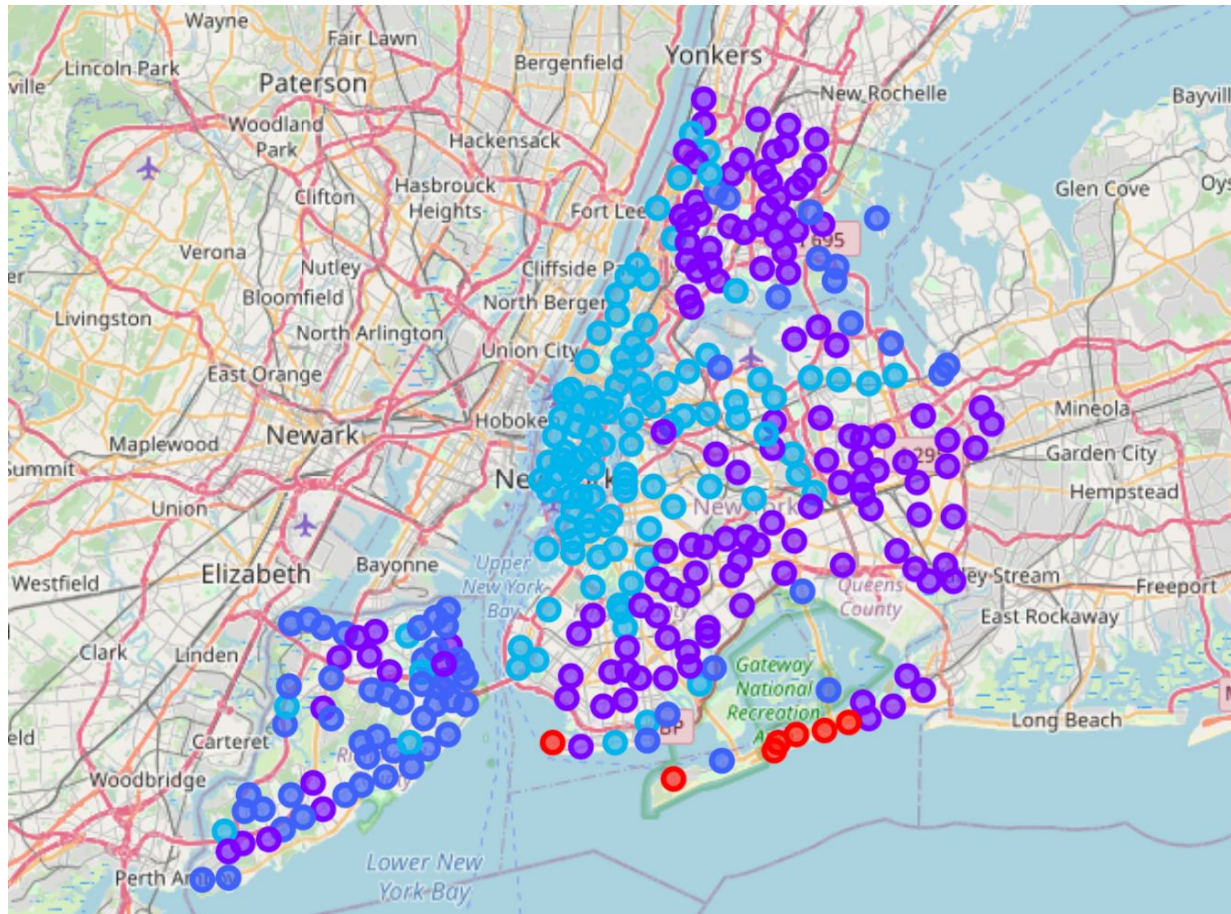


Based on the graph shown above, the optimal number of clusters is 4, and we run the K-means clustering with 4 as the number of clusters. Now, we create a dataframe that contains the cluster labels as well as the top 10 venues for each neighborhood.

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9 C
0	Bronx	Wakefield	40.894705	-73.847201	1	Pharmacy	Caribbean Restaurant	Supermarket	Donut Shop	Fast Food Restaurant	Fried Chicken Joint	Bank	Pizza Place	
1	Bronx	Co-op City	40.874294	-73.829939	1	Pizza Place	Department Store	Mobile Phone Shop	Shopping Mall	Shoe Store	Pharmacy	Clothing Store	Bakery	
2	Bronx	Eastchester	40.887556	-73.827806	1	Caribbean Restaurant	Shopping Mall	Diner	Fast Food Restaurant	Gas Station	Supplement Shop	Grocery Store	Pharmacy	
3	Bronx	Fieldston	40.895437	-73.905643	1	Deli / Bodega	Bus Station	Pizza Place	Bar	Plaza	Sandwich Place	Park	Mexican Restaurant	
4	Bronx	Riverdale	40.890834	-73.912585	3	Bank	Pizza Place	Bar	Park	Diner	Deli / Bodega	Japanese Restaurant	Coffee Shop	



Map to visualize the clusters is created:



## Clusters Examination:

Cluster 1:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
85	Sea Gate	Beach	Supermarket	American Restaurant	Pharmacy	Park	Donut Shop	Chinese Restaurant	Bus Stop	Video Store	Spa
172	Breezy Point	Surf Spot	American Restaurant	Beach	Monument / Landmark	Park	Trail	Deli / Bodega	Exhibit	Factory	Falafel Restaurant
178	Rockaway Beach	Beach	Bar	Pharmacy	Boat or Ferry	Latin American Restaurant	Ice Cream Shop	Pizza Place	Eastern European Restaurant	Arepa Restaurant	Supermarket
179	Neponsit	Beach	Spa	Deli / Bodega	Pub	Mexican Restaurant	Italian Restaurant	Restaurant	Harbor / Marina	Bakery	Boutique
190	Belle Harbor	Beach	Pub	Donut Shop	Bakery	Mexican Restaurant	Bagel Shop	Chinese Restaurant	Harbor / Marina	Pharmacy	Smoke Shop
191	Rockaway Park	Beach	Pizza Place	Donut Shop	Liquor Store	Bagel Shop	Bar	Pharmacy	Deli / Bodega	Pub	Supermarket
302	Hammels	Beach	Surf Spot	Donut Shop	Fast Food Restaurant	Coffee Shop	Pizza Place	Taco Place	Bar	Supermarket	Wine Bar

## Cluster 2: (first 7 rows shown)

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Wakefield	Pharmacy	Caribbean Restaurant	Supermarket	Donut Shop	Fast Food Restaurant	Fried Chicken Joint	Bank	Pizza Place	Bakery	Bus Station
1	Co-op City	Pizza Place	Department Store	Mobile Phone Shop	Shopping Mall	Shoe Store	Pharmacy	Clothing Store	Bakery	Liquor Store	Mattress Store
2	Eastchester	Caribbean Restaurant	Shopping Mall	Diner	Fast Food Restaurant	Gas Station	Supplement Shop	Grocery Store	Pharmacy	Donut Shop	Discount Store
3	Fieldston	Deli / Bodega	Bus Station	Pizza Place	Bar	Plaza	Sandwich Place	Park	Mexican Restaurant	Bank	Chinese Restaurant
6	Marble Hill	Park	Pizza Place	Mexican Restaurant	Café	Donut Shop	Spanish Restaurant	Sandwich Place	Grocery Store	Scenic Lookout	Supermarket
7	Woodlawn	Pub	Bar	Deli / Bodega	Pizza Place	Bakery	Discount Store	Donut Shop	Bank	Baseball Field	Fast Food Restaurant

## Cluster 3: (first 7 rows shown)

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
12	City Island	Harbor / Marina	Seafood Restaurant	Boat or Ferry	Italian Restaurant	Bar	Thrift / Vintage Store	American Restaurant	Music Venue	French Restaurant	Bank
16	Fordham	Italian Restaurant	Pizza Place	Spanish Restaurant	Deli / Bodega	Shoe Store	Mobile Phone Shop	Diner	Coffee Shop	Supplement Shop	Dessert Shop
27	Clason Point	Pool	Boat or Ferry	Event Space	Discount Store	Gym / Fitness Center	Park	Wings Joint	Bus Stop	Farm	Falafel Restaurant
28	Throgs Neck	Italian Restaurant	Deli / Bodega	Donut Shop	Pizza Place	Harbor / Marina	American Restaurant	Asian Restaurant	Coffee Shop	Trail	Bookstore
34	Belmont	Italian Restaurant	Pizza Place	Deli / Bodega	Bakery	Shoe Store	Zoo	Dessert Shop	Latin American Restaurant	Café	Garden
37	Pelham Bay	Italian Restaurant	Deli / Bodega	Fast Food Restaurant	Sandwich Place	Pizza Place	Bank	Bakery	Convenience Store	Donut Shop	Rental Car Location

## Cluster 4: (first 7 rows shown)

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
4	Riverdale	Bank	Pizza Place	Bar	Park	Diner	Deli / Bodega	Japanese Restaurant	Coffee Shop	Health & Beauty Service	Bagel Shop
5	Kingsbridge	Pizza Place	Mexican Restaurant	Sandwich Place	Bar	Diner	Donut Shop	Coffee Shop	Pharmacy	Bank	Park
19	High Bridge	Lounge	Baseball Stadium	Park	Plaza	Burger Joint	Bar	BBQ Joint	Beer Bar	Caribbean Restaurant	Liquor Store
24	Hunts Point	Park	Bakery	Market	Paintball Field	Travel & Transport	Gourmet Shop	Farmers Market	Seafood Restaurant	Mexican Restaurant	Home Service
46	Bay Ridge	Pizza Place	Spa	Italian Restaurant	Coffee Shop	Bagel Shop	Chinese Restaurant	Cosmetics Shop	Greek Restaurant	American Restaurant	Café
48	Sunset Park	Bakery	Mexican Restaurant	Pizza Place	Chinese Restaurant	Bank	Latin American Restaurant	Ice Cream Shop	Asian Restaurant	Sandwich Place	Dumpling Restaurant

## **Conclusion & Recommendations**

After implementing the k-means clustering algorithm, we can see the following for each cluster:

### **Cluster 1:**

In the first cluster we can see that there aren't many restaurants in the top spots of common venues in these neighborhoods, and also there isn't a wide variety of cuisine options in the top spots with the most common options being: American restaurants, Donut shops and Pizza places. Furthermore, we can also see that beaches are quite common in these neighborhoods, making it an ideal option for some types of cuisine like seafood, casual dining, and other types of laid back cuisines like lounges.

### **Cluster 2:**

In cluster 2 we can see that there are many restaurants in the top 3 spots of most common venues in the neighborhoods, like Caribbean restaurants, Donut shops, Chinese restaurants, etc... with Pizza places being quite common. Which can also mean that competition could be high in these neighborhoods. We can also see from the map, that these neighborhoods are mainly in Brooklyn, Queen and The Bronx, and that it is one of the largest clusters where F&B businesses are very common and diverse. Furthermore, we can see that parks occupy a top spot in a couple of neighborhoods, which could give us an idea that these neighborhoods are more family oriented, where Family Diners could do well. The neighborhoods in this cluster could be ideal for fast food outlets, family diners, as well as foreign cuisines given the varse diversity of the neighborhoods.

**Cluster 3:**

In this cluster, we can note that Italian restaurants and pizza places are quite common in these neighborhoods. This could also mean that other types of cuisines could have a good chance there, but this could also mean that the demographic in these neighborhoods prefers these types of cuisines, and that is why we see many of those in these areas. When we look at the map of clusters, we see most of the points are in Staten Island, where the majority of residents are Italian Americans. This cluster could be ideal for a lot of different cuisines and specifically American cuisines given that it has some museums, hotels and other touristy attractions, and also casual or family diners since its neighborhoods have common parks and other family attractions, and finally, it can be ideal for seafood concepts as some of the neighborhoods have Marinas and Harbors as the most common venues.

**Cluster 4:**

If we look at the top most common venues in this cluster, we can find out that it has quite a variety of famous venues from Italian, to Latin American, to American and also Asian cuisines. And looking at the clusters map, we can see that the points of this cluster are mainly in Manhattan, which has a very diverse and competitive market. In this cluster, we can see that coffee shops, bars, cafes and parks are some of the more common venues in the neighborhoods. Making these neighborhoods ideal for such types of cuisines.