Data Science Report On Ideal Locations To Establish a Restaurant within The City of New York

Steve Tambo

February 9th, 2021

1. Introduction

1.1 Background

New York City is the most populous city in the United States and the largest metropolitan area in the world by urban landmass with almost **20 million** people in it's metropolitan statistical area. It has been described as the cultural, financial, and media capital of the world, significantly influencing entertainment, research, technology, education, politics, tourism, art, fashion, and sports. All these provide for great opportunities as far as business ventures go especially for restauranteurs with New York being the country's leading restaurant city. Given the diverse nature of its residents, one must be critical in deciding where to locate their business so they can maximize on profits. This can be done through carefully analyzing the distribution of existing venues within the city and clustering them into distinct sections ; singling out the areas you are most likely to find the highest potential for customer inflow.

1.2 Problem definition

Piacci's is an Italian owned restaurant franchise that has proven highly successful since inception. The establishment was formed in Montreal, Canada in the year 2009 and they specialize in gourmet Italian delicacies. Given their current success they wish to expand to other cities in America more specifically within the neighborhoods of New York.

Data about different venues in New York can be accessed via the FourSquare API. This data shows the names , locations and ratings of each venue as assessed by customers who have already visited them. Having spoken to numerous opportunity assessors they have been advised to seek the services of a data scientist to access and analyze this data so they may determine what would be the ideal locations to set up new restaurants within the city through the use of machine learning algorithms.

2. Data acquisition and preparation

2.1 Data sources

To begin the analysis a data set of all the boroughs and neighborhoods in New York highlighting their names and geographic locations (in terms of latitude and longitude) was needed. This was downloaded and stored in the form of a JSON file named 'newyork_data'.

The second dataset was obtained by connecting to the FourSquare API and downloading data on all the venues around New York from their servers.

2.2 Data preparation

The first dataset containing information about New York neighborhoods could not be used in its current format . I had to use the pandas library on Jupyter notebook to open this JSON file and extract the relevant information into a pandas dataframe. The attributes I was looking for were the name of each borough, the names of the neighborhoods within that borough as well as their geographic locations. All of this was stored within the features key of the 'newyork_data' dictionary. I created a new empty dataframe named 'neighborhoods' and decided to loop through the JSON file storing each new feature within it. After successfully extracting the data I identified **5 boroughs** and **306 neighborhoods** in total.

	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

With the New York neighborhoods data frame prepared it was time to connect to the FourSquare API and obtain information about their venues. The attributes I was interested in were the neighborhood name , latitude and longitude. In relation to the venue I was interested in the name, latitude , longitude and category of each venue within a 500m radius of each neighborhood. After successfully completing the request and receiving the data I identified **10034** venues in total divided uniquely across **429** categories.

3. Methodology : Clustering the neighborhoods

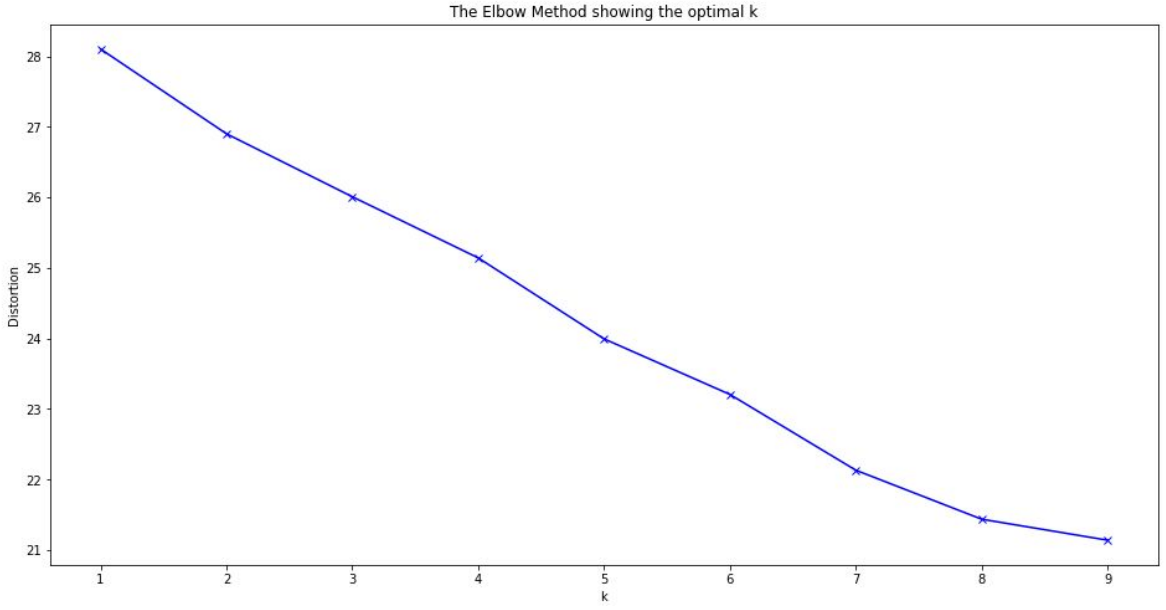
3.1 Feature selection

Quantitative data about each venue was needed in order to fit a machine learning model. I had to create a one hot vector for the mean frequency of each category in each neighborhood. The resulting feature set was stored in a new dataframe 'newyork_onehot' which would be grouped by neighborhood and clustered according to frequency.

	Neighborhood	Yoga Studio	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Terminal	American Restaurant	Antique Shop	...	Volleyball Court	Warehouse Store	Waste Facility	Waterfror
0	Allerton	0.0	0.000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.
1	Annadale	0.0	0.000	0.0	0.0	0.0	0.0	0.0	0.100000	0.0	...	0.0	0.0	0.0	0.
2	Arden Heights	0.0	0.000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.
3	Arlington	0.0	0.125	0.0	0.0	0.0	0.0	0.0	0.125000	0.0	...	0.0	0.0	0.0	0.
4	Arrochar	0.0	0.000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	.
297	Woodhaven	0.0	0.000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.
298	Woodlawn	0.0	0.000	0.0	0.0	0.0	0.0	0.0	0.038462	0.0	...	0.0	0.0	0.0	0.
299	Woodrow	0.0	0.000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.
300	Woodside	0.0	0.000	0.0	0.0	0.0	0.0	0.0	0.038462	0.0	...	0.0	0.0	0.0	0.
301	Yorkville	0.0	0.000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.

3.2 Fitting the model - K Means

Since I was dealing with an unsupervised learning situation, I decided to use the K-means machine learning algorithm to cluster the model mainly because it works well when segmenting large consumer oriented datasets. In order to determine the appropriate K-value (number of clusters) I applied the elbow method as shown below.



The best value of K would be where the 'elbow' of the function occurred; in this case where the value of k was 8. I went ahead and fit the model using the grouped set of features and set the number of clusters to 8.

3.4 Obtaining the most common venues

In order to determine the most suitable locations for setting up a restaurant a list of the most common venues around each neighborhood was needed. I used the numpy library to sort and arrange the top categories in descending order within each neighborhood and stored the resulting features in a new data frame 'neighborhood_venues_sorted'. Only the top 10 most common venues per neighborhood were selected.

	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	Bakery	Bus Station	Chinese Restaurant	Deli / Bodega	Spa	Supermarket	Martial Arts School	Fast Food Restaurant	Gas Station
1	Annadale	Pizza Place	Food	Pharmacy	Train Station	Diner	Restaurant	American Restaurant	Bakery	Cosmetics Shop	Food & Drink Shop
2	Arden Heights	Home Service	Pharmacy	Deli / Bodega	Coffee Shop	Pizza Place	Women's Store	Ethiopian Restaurant	Event Service	Event Space	Exhibit
3	Arlington	Bus Stop	American Restaurant	Grocery Store	ATM	Intersection	Deli / Bodega	Coffee Shop	Farmers Market	Farm	Falafel Restaurant
4	Arrochar	Pizza Place	Deli / Bodega	Italian Restaurant	Bagel Shop	Bus Stop	Nail Salon	Outdoors & Recreation	Middle Eastern Restaurant	Sandwich Place	Liquor Store

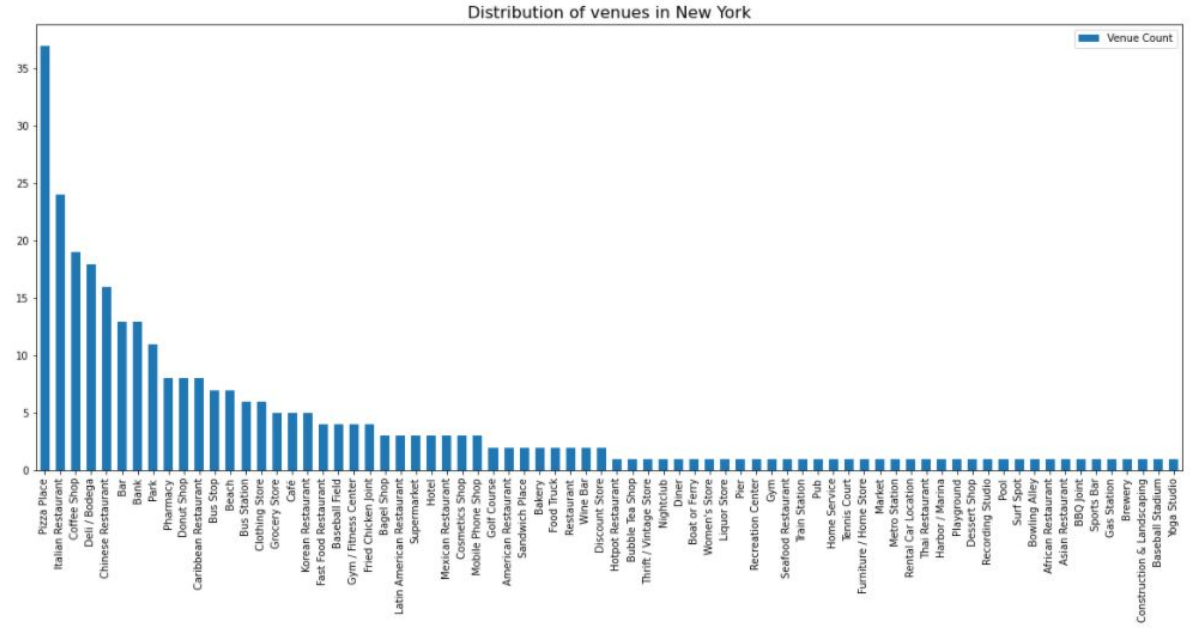
3.5 Adding cluster labels and merging datasets

After sorting the venue data it was important to merge it with the New York dataset so as to have a complete set of features to decipher from. I merged the ‘neighborhood_venues_sorted’ data frame with my initial ‘neighborhoods’ data frame using the neighborhood field and added in the cluster labels which would be useful during analysis and visualization. The resulting data frame was named ‘newyork_merged’ and would serve as the main point of reference for all further analysis.

	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	9th Mc Comm Ven
0	Bronx	Wakefield	40.894705	-73.847201	3	Pharmacy	Deli / Bodega	Laundromat	Sandwich Place	Dessert Shop	Food	Donut Shop	Ice Cream Shop	Caribbe Restaur
1	Bronx	Co-op City	40.874294	-73.829939	3	Restaurant	Bus Station	Fried Chicken Joint	Park	Baseball Field	Grocery Store	Bagel Shop	Pharmacy	Fast Fo Restaur
2	Bronx	Eastchester	40.887556	-73.827806	2	Deli / Bodega	Caribbean Restaurant	Bus Station	Diner	Chinese Restaurant	Automotive Shop	Donut Shop	Bowling Alley	Busine Servi
3	Bronx	Fieldston	40.895437	-73.905643	0	Bus Station	River	Plaza	Business Service	Women's Store	Field	Ethiopian Restaurant	Event Service	Eve Spa
4	Bronx	Riverdale	40.890834	-73.912585	7	Bus Station	Park	Bank	Medical Supply Store	Gym	Plaza	Baseball Field	Food Truck	Fooc Dri Sh

4. Results : Cluster analysis

Before analyzing each cluster I decided to get a general overview of New York. From the graph below one can see the overall distribution of venues around New York . The most common venue within the city was pizza places which explained the famed New York pizza. This was followed by Italian restaurants and coffee shops.



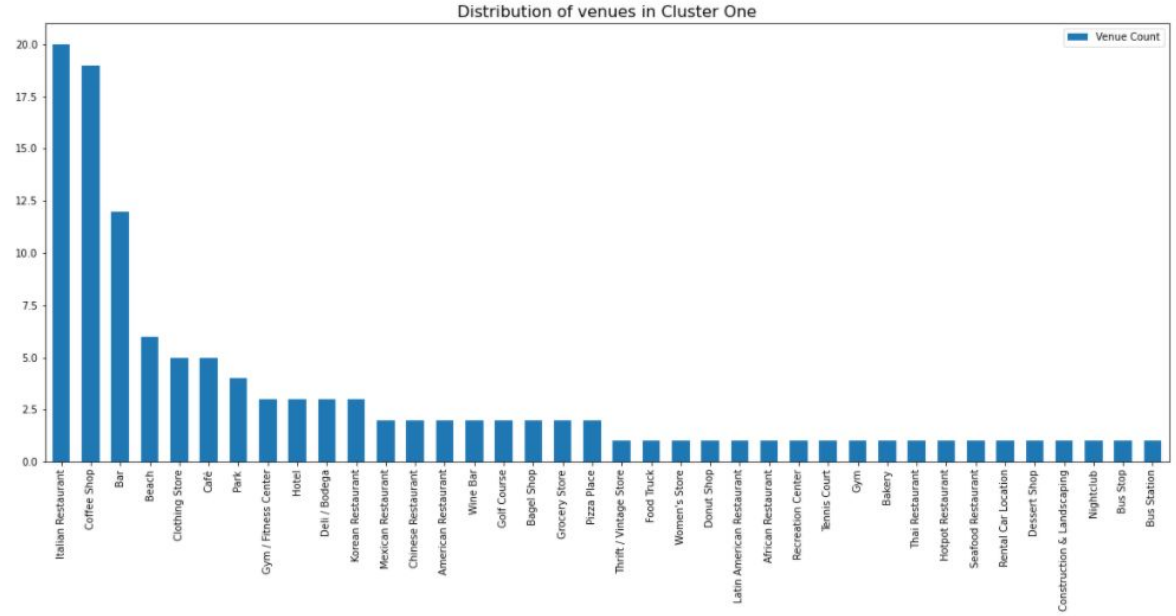
Upon counting the number of neighborhoods within each cluster I realized that clusters two, five , seven and six had less than 3 neighborhoods. I decided to consider them as outliers and ignored them during analysis.

4.1 Cluster Analysis

A. Cluster One

I filtered the newyork_merged data frame and selected only the rows containing a cluster label of 0 (cluster one). The results were as 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
count	118	118	118	118	118	118	118	118	118	118	118
unique	115	38	59	55	61	63	64	74	72	77	69
top	Bay Terrace	Italian Restaurant	Coffee Shop	Pizza Place	Cocktail Bar	American Restaurant	Italian Restaurant	Bar	Event Service	Event Space	Exhibit
freq	2	20	11	12	8	7	7	4	7	8	8



The cluster was densely populated with over 100 neighborhoods. I concluded that this is an eatery cluster because :

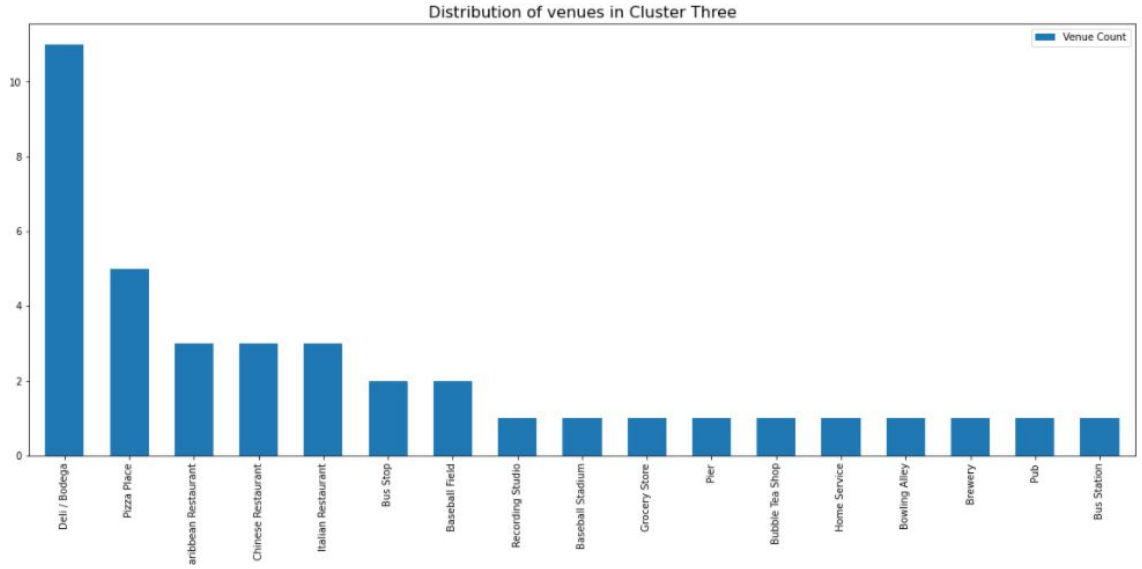
- Majority of the most common venues were restaurants and coffee shops (the top 2 venues)
- The most common venues within the cluster were Italian restaurants
- 7 out of the top 10 venues served food and beverages

It is most probably an area for the high income working class.

B. Cluster Three

I filtered the newyork_merged data frame and selected only the rows containing a cluster label of 2 (cluster three).

	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
count	39	39	39	39	39	39	39	39	39	39	39
unique	39	17	22	25	28	34	30	32	27	29	27
top	Briarwood	Deli / Bodega	Deli / Bodega	Deli / Bodega	Bus Stop	Bus Station	Women's Store	Field	Event Space	Exhibit	Eye Doctor
freq	1	11	13	6	3	2	5	4	5	6	6



The cluster was fairly populated with 39 neighborhoods. I concluded that this was a retail cluster because :

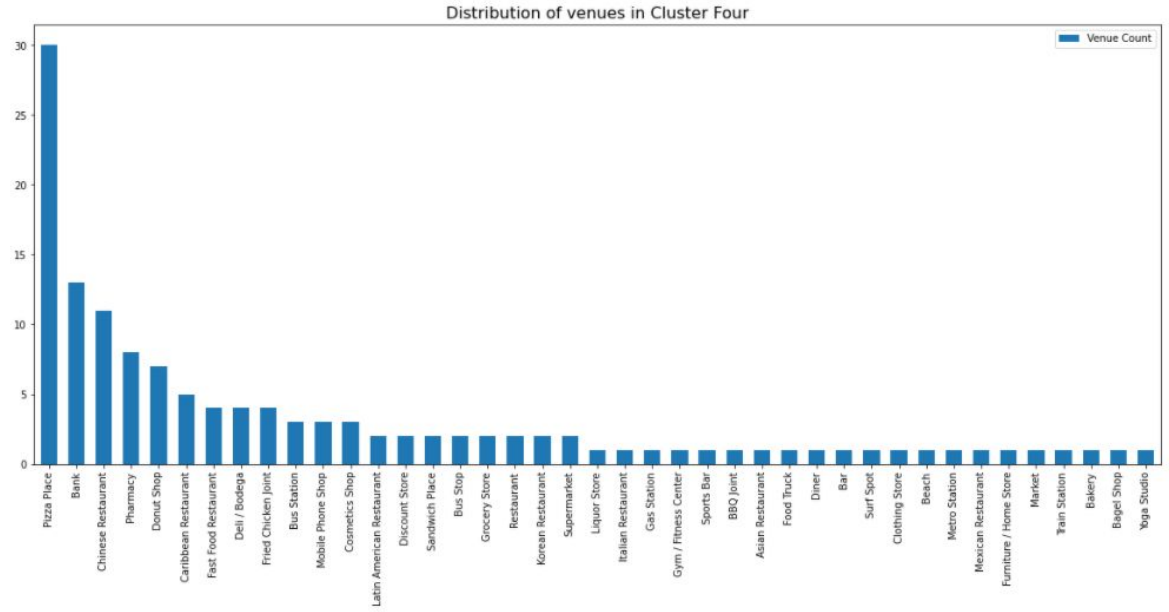
- Majority of the most common venues were deli's/bodegas (top 3 venues)

It could also be an eatery cluster because aside from being just grocery stores, most delis serve food to their customers as well.

C. Cluster Four

I filtered the newyork_merged data frame and selected only the rows containing a cluster label of 3 (cluster four). The results were as 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
count	132	132	132	132	132	132	132	132	132	132	132
unique	131	41	59	59	59	68	78	66	63	81	62
top	Sunnyside	Pizza Place	Grocery Store	Pizza Place	Sandwich Place	Pizza Place	Donut Shop	Bank	Bakery	Pizza Place	Pharmacy
freq	2	30	8	12	10	8	9	8	7	6	7



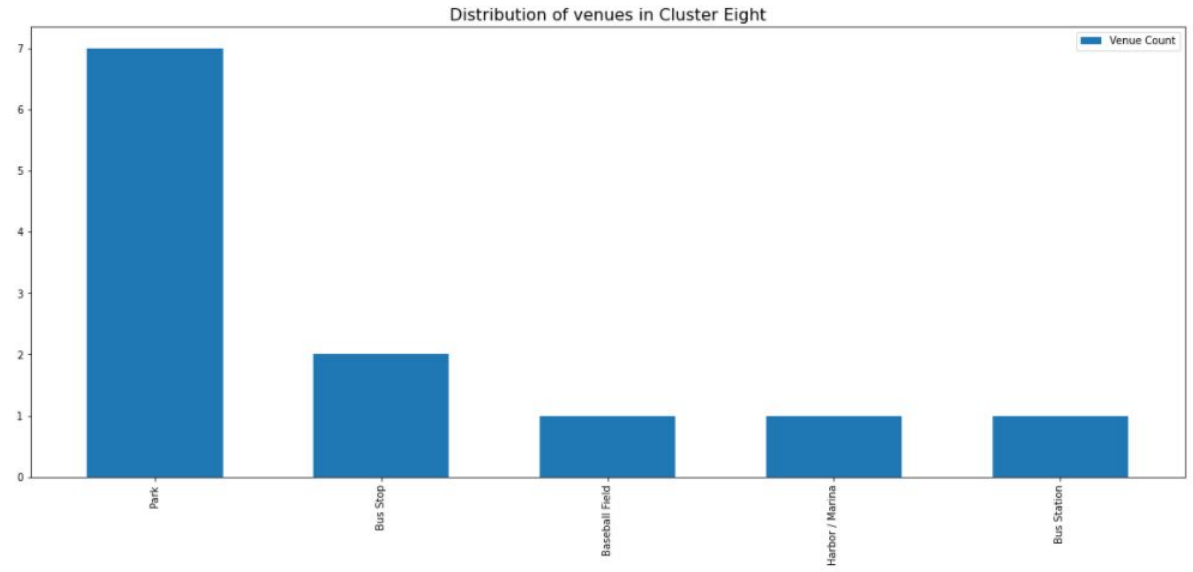
The cluster was densely populated with over 100 neighborhoods. I concluded it was an eatery cluster because :

- The most common venues within the area were pizza places
- 7 out of the top 10 venues sold ready made food items.

D. Cluster Eight

I filtered the newyork_merged data frame and selected only the rows containing a cluster label of 7 (cluster eight). The results were as shown below.

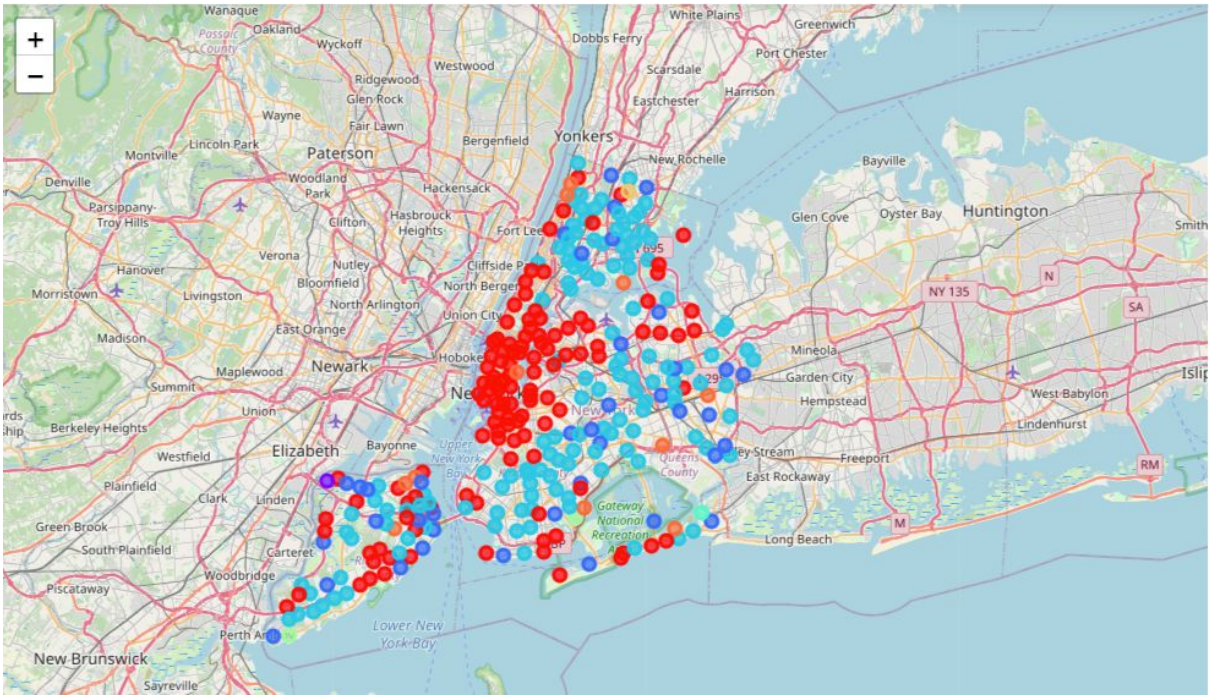
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
count	12	12	12	12	12	12	12	12	12	12	12
unique	12	5	8	10	11	12	11	11	11	11	9
top	Somerville	Park	Park	Deli / Bodega	Boat or Ferry	Bus Stop	Fast Food Restaurant	Discount Store	Event Service	Event Space	Farm
freq	1	7	4	3	2	1	2	2	2	2	2



The cluster was sparsely populated with only 12 neighborhoods. It seemed to be an activity cluster for those who like outdoor leisure activities with parks being the most common venue.

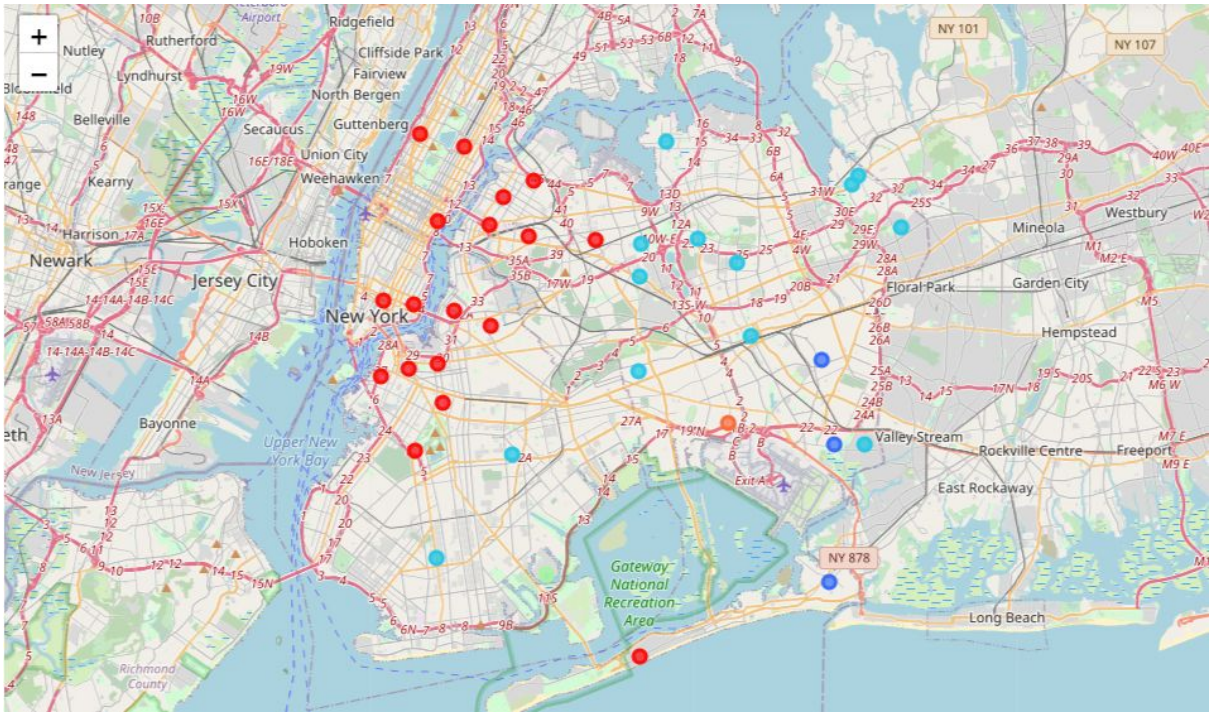
4.2 Cluster Distribution

The map below shows the overall distribution of the clusters within New York. Cluster one (red dots) is located around the CBD . Cluster four (light blue dots) is located around the outskirts of the CBD and is surrounded with evenly separated folds of cluster three (dark blue dots). Cluster eight is represented by the orange dots and can be seen scattered across the city.



5. Discussion : Locating ideal locations to open a restaurant

After identifying each cluster and its general characteristics it was time to locate the ideal neighborhoods for Piacci's to consider . The first step was to filter through the 'newyork_merged' dataframe and select only the neighborhoods in which the least most common venue was a restaurant. In the first filtered data frame, 63 neighborhoods were identified . The next step was to filter out any neighborhoods where the most common venue happened to be an Italian restaurant. In the second filtered data frame, 57 neighborhoods were identified. The final step was to narrow down on the geographics and select the neighborhoods that happened to be close to the CBD. I used the coordinates of Manhattan as a reference point . This means that the selected neighborhoods had to have a latitude of between **40° N** and **40.8° N** and a longitude of between **-72° W** and **-74° W** respectively. The final cohort is visible below with **36** potential neighborhoods identified for consideration.



From the map figure above, the majority of the neighborhoods identified fall within cluster one and four . These were essentially eatery clusters so the information provided can be termed as accurate. Potential for customers looking for Italian restaurants within these neighborhoods would be high. A few of the neighborhoods fall within cluster three and eight. These can be ignored as they are not primarily eatery clusters and customer potential would be relatively low.

It can be deduced that there is great opportunity in New York as far as the restaurant business is concerned given the high number of potential locations (a total of 36 neighborhoods). Demand for good food must be very high in the city.

Conclusion

In this study I segmented New York neighborhoods using the different categories of venues located within them by use of the K means clustering model. I managed to identify four distinct clusters and used this knowledge to locate the best possible place for one to set up a restaurant given the distribution of venues within a neighborhood. This model would prove very useful to anyone looking to open up an Italian restaurant in New York and can be adjusted depending on the venue category one may be interested in.