The Neighborhoods of New York

Shibbu Joseph

March 14, 2021

1. Introduction

1.1. Background

I chose to analyze New York City. It is the largest and most influential American metropolis. New York City is in reality a collection of many neighborhoods scattered among the city’s five boroughs: The Bronx, Brooklyn, Manhattan, Queens, and Staten Island, each exhibiting its own characteristics and ways of life. They say that moving from one neighborhood to another one may be like moving out to a different country. Therefore, it is advantageous to know how similar a neighborhood is to another one in each borough.

1.2. Problem

This project aims to compare by similarity each neighborhood inside each borough, making clusters of neighborhoods, in order to learn which neighborhoods are similar, and which ones are substantially different. Having this information, I will also compare the results by analyzing the clusters distribution between each borough. Finally, I will give a general analysis of the complete New York City, comparing all the neighborhoods in the city.

1.3. Interest

The following set of analyses may be useful to those moving to NYC, or moving from one neighborhood to another one within NYC. This would be especially helpful for those looking to move closer to an area with venues in their desired line of work. Real estate agents looking to improve their suggestions to clients may also find these analyses useful. In order to provide more tailored recommendations, real estate agents may use the venue information to match clients to areas that fit their desired job title and lifestyle.

2. Data acquisition and cleaning process

2.1. Data sources

The data was acquired through the city of New York Open Data team, published in the following site: https://opendata.cityofnewyork.us/. I selected the dataset named Neighborhood Names GIS. The raw data can be found in the following link: https://data.cityofnewyork.us/City-Government/Neighborhood-Names-GIS/99bc-9p23. The data contains the The_geom, object id, name, stacked, borough, Annoline1, Annoline2, Annoline3, AnnoAngle columns of each neighborhood. The_geom column contains the geolocation data, the borough is the name of the borough where the neighborhood belongs to, the Annolines columns are the names of the neighborhoods word by word and the stacked column is the amount of words in the neighborhood name. I will use this data to classify each of the neighborhoods. I will also use the foursquare API to retrieve a list of venues nearby each neighborhoods to make the analysis.

2.2. Data cleaning

The data was downloaded but to work with the dataset I had to make a few changes in the dataset. First of all, I had to eliminate some columns that wouldn't contribute at all with the analysis like stacked, Annoline1, Annoline2, Annoline3, and AnnoAngle, they information that this columns give are redundant for the analysis. After dropping each of these columns there was another problem: the_geom (the geolocation column) had a format of <POINT (Longitude, Latitude)>. To use this data point properly, we needed it in two columns: one for latitude and another one for longitude. To fix this data

point I:

1. Eliminated all the POINT word and the parentheses of each data point.
2. Separate the longitude and the latitude in different columns for each row.
3. After doing all this the data frame was ready.

Having cleaned the unnecessary parts of the data, I divided the data frame in six different frames, one for each borough and the final one for the whole list of neighborhoods in New York.

I had to remove some neighborhoods in the data frame: One in Staten Island and one in New York called “Chelsea”. The first one didn’t have any nearby venues, so it couldn’t be compared to other neighborhoods. The second one was duplicated in the dataset, so I had to remove one instance.

3. Methodology

3.1. Clustering

I will use the k-means method to cluster each of the neighborhoods. First, I used the foursquare API to obtain the venues. Having the list of venues, I applied an onehot coding procedure, that consist of putting in binary information the data we are analyzing, in which I classified what type of venues each were. Following this, I now had to group each venue in the list with the neighborhood it belonged to and with this, I was able to calculate in what frequency each type of venue appeared in each neighborhood. This frequency of each type of venue can be used to classify each neighborhoods top ten most common venues.

Now that I have a data frame with the top type of venues in each neighborhood I was able to proceed to the clustering process. For this type of process the number of cluster will determine the number of different groups that each neighborhood will associate with. So for each borough I decide to create 5 cluster and for the full clustering of the neighborhoods of New York I decided to work with 8 cluster, the reason being is the amount of neighborhoods is much larger, and we can assume we will need more groups to classify the neighborhoods.

Finally plotting in a map the k-means clustering we can show which neighborhoods and their top picks are in each cluster. The analysis of each cluster and look at their properties and comment which cluster is the most common to find in each borough and in New York.

3.2. Bronx Analysis

In the map clustering we can see that two clusters are found more commonly that the others. The cluster number 3 is the one that we can see the most, this cluster contains Italian restaurants and food places. The Bronx could be a good place for family with Italian roots or that love this kind of food.

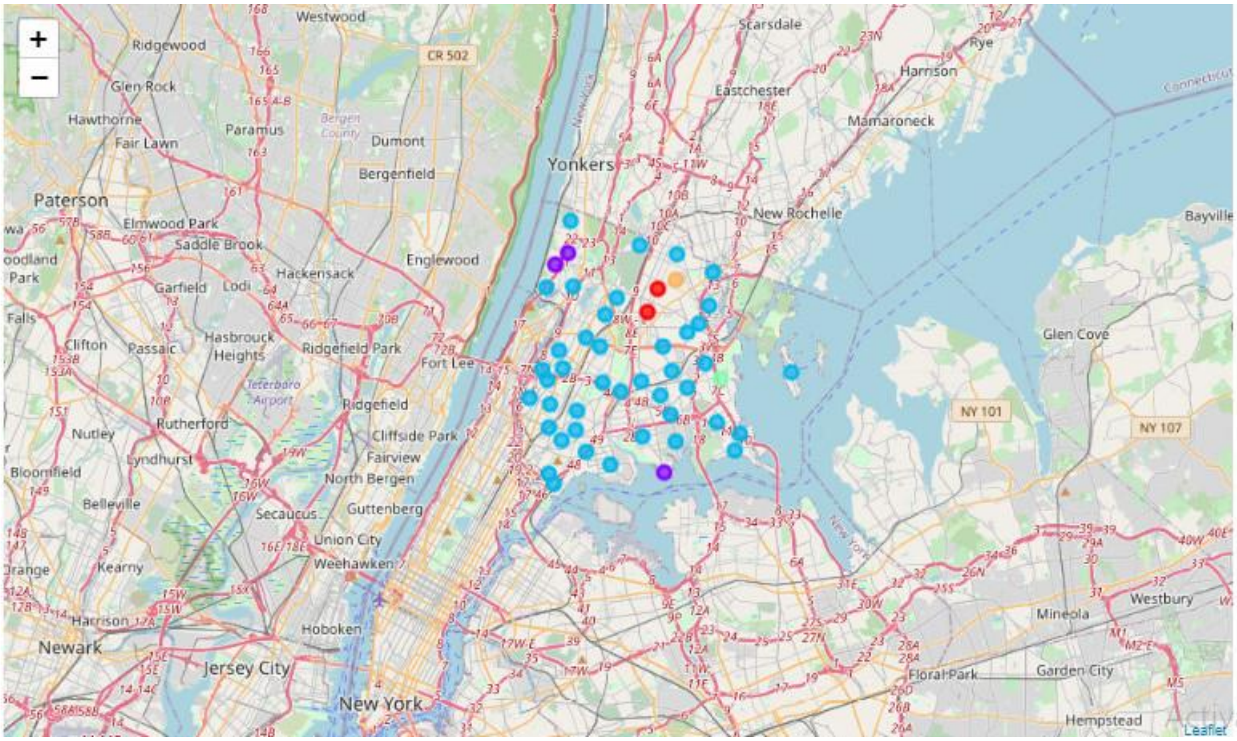
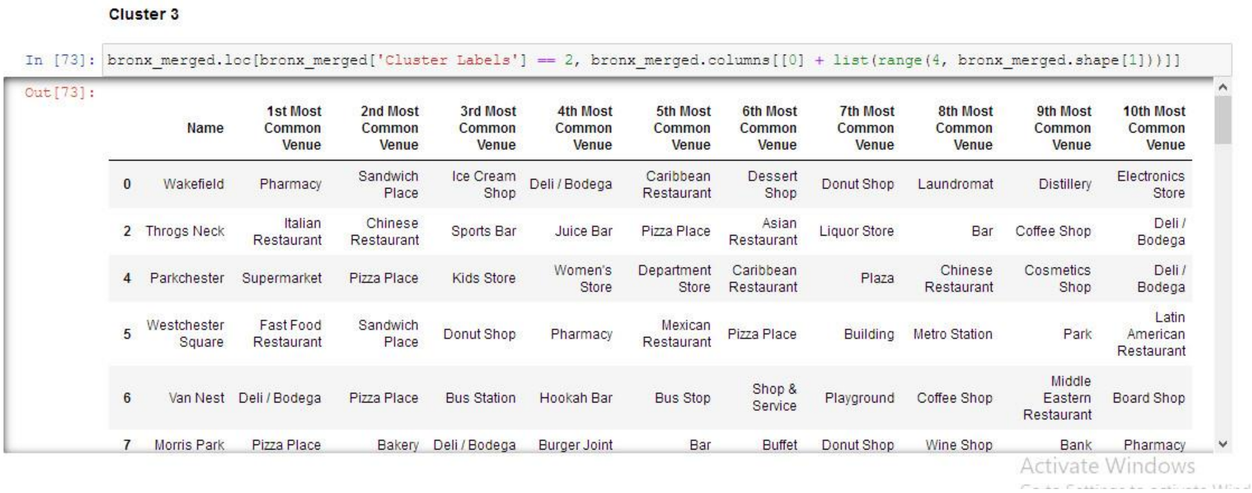


Figure 1 Bronx K-means Clustering



Cluster 3

```
In [73]: bronx_merged.loc[bronx_merged['Cluster Labels'] == 2, bronx_merged.columns[[0] + list(range(4, bronx_merged.shape[1]))]]
```

Out[73]:

	Name	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	Sandwich Place	Ice Cream Shop	Deli / Bodega	Caribbean Restaurant	Dessert Shop	Donut Shop	Laundromat	Distillery	Electronics Store
2	Throgs Neck	Italian Restaurant	Chinese Restaurant	Sports Bar	Juice Bar	Pizza Place	Asian Restaurant	Liquor Store	Bar	Coffee Shop	Deli / Bodega
4	Parkchester	Supermarket	Pizza Place	Kids Store	Women's Store	Department Store	Caribbean Restaurant	Plaza	Chinese Restaurant	Cosmetics Shop	Deli / Bodega
5	Westchester Square	Fast Food Restaurant	Sandwich Place	Donut Shop	Pharmacy	Mexican Restaurant	Pizza Place	Building	Metro Station	Park	Latin American Restaurant
6	Van Nest	Deli / Bodega	Pizza Place	Bus Station	Hookah Bar	Bus Stop	Shop & Service	Playground	Coffee Shop	Middle Eastern Restaurant	Board Shop
7	Morris Park	Pizza Place	Bakery	Deli / Bodega	Burger Joint	Bar	Buffet	Donut Shop	Wine Shop	Bank	Pharmacy

Figure 2 Bronx Most Common Cluster (Number 3) (Not full list of neighborhoods)

3.3. Brooklyn Analysis

In the map we can see two cluster again like the Brooklyn clustering, this time the cluster 4 is the most common. With restaurants, banks and grocery stores are common urban area venues. People looking to live close to convenience venues will have a lot of options in Brooklyn

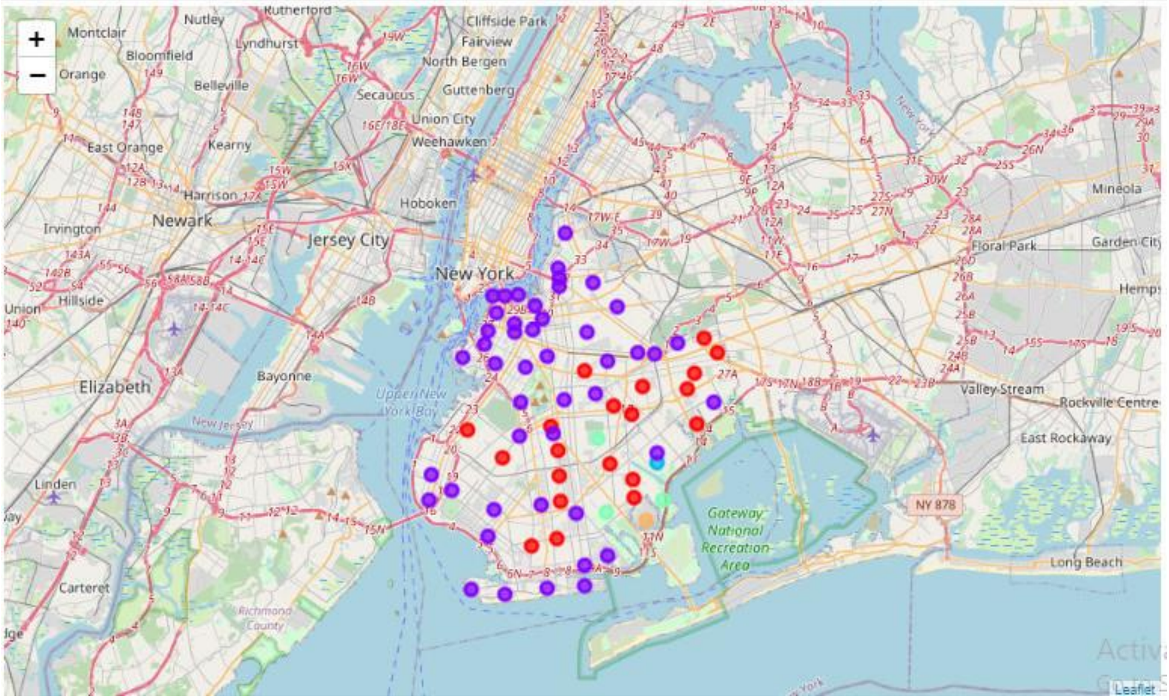
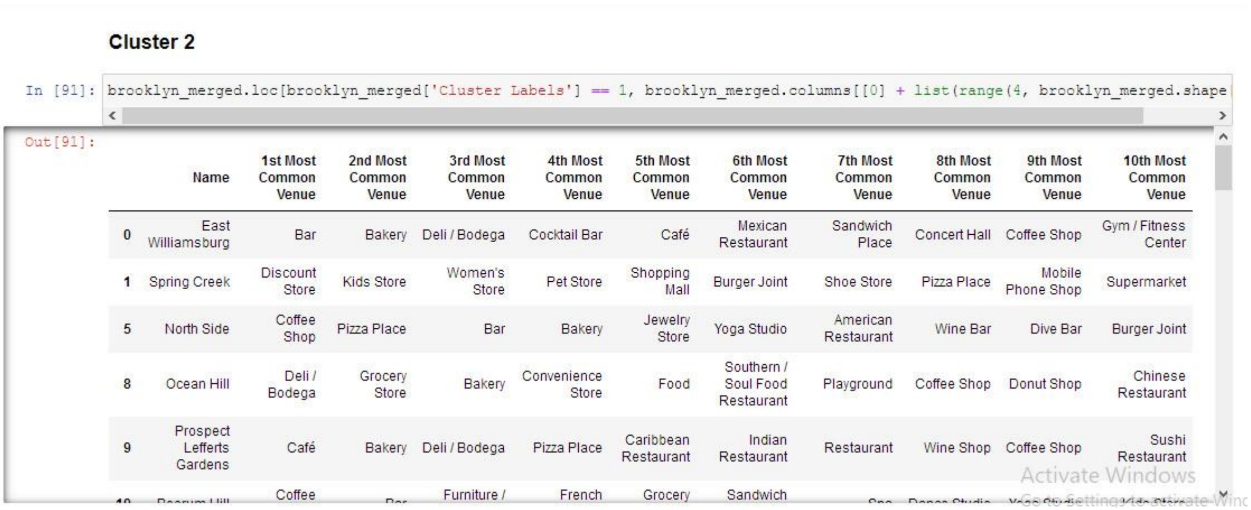


Figure 3 Brooklyn K-means Clustering


Cluster 2

```
In [91]: brooklyn_merged.loc[brooklyn_merged['Cluster Labels'] == 1, brooklyn_merged.columns[[0] + list(range(4, brooklyn_merged.shape
```

Out[91]:

	Name	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	East Williamsburg	Bar	Bakery	Deli / Bodega	Cocktail Bar	Café	Mexican Restaurant	Sandwich Place	Concert Hall	Coffee Shop	Gym / Fitness Center
1	Spring Creek	Discount Store	Kids Store	Women's Store	Pet Store	Shopping Mall	Burger Joint	Shoe Store	Pizza Place	Mobile Phone Shop	Supermarket
5	North Side	Coffee Shop	Pizza Place	Bar	Bakery	Jewelry Store	Yoga Studio	American Restaurant	Wine Bar	Dive Bar	Burger Joint
8	Ocean Hill	Deli / Bodega	Grocery Store	Bakery	Convenience Store	Food	Southern / Soul Food Restaurant	Playground	Coffee Shop	Donut Shop	Chinese Restaurant
9	Prospect Lefferts Gardens	Café	Bakery	Deli / Bodega	Pizza Place	Caribbean Restaurant	Indian Restaurant	Restaurant	Wine Shop	Coffee Shop	Sushi Restaurant
10	Boerum Hill	Coffee	Bar	Furniture /	French	Grocery	Sandwich	Spa	Dance Studio	Yoga Studio	Kids Store

Figure 4 Brooklyn Most Common Cluster (Number 2) (Not full list of neighborhoods)

3.4. Manhattan Analysis

In this clustering we can see a dominance of neighborhood cluster again the number 2 consists of restaurants, coffee shops, hotels and gyms. This type of neighborhood can be attractive for young families and teenagers looking for places with a lot of entertainment and different persons thanks to the hotels and bars.

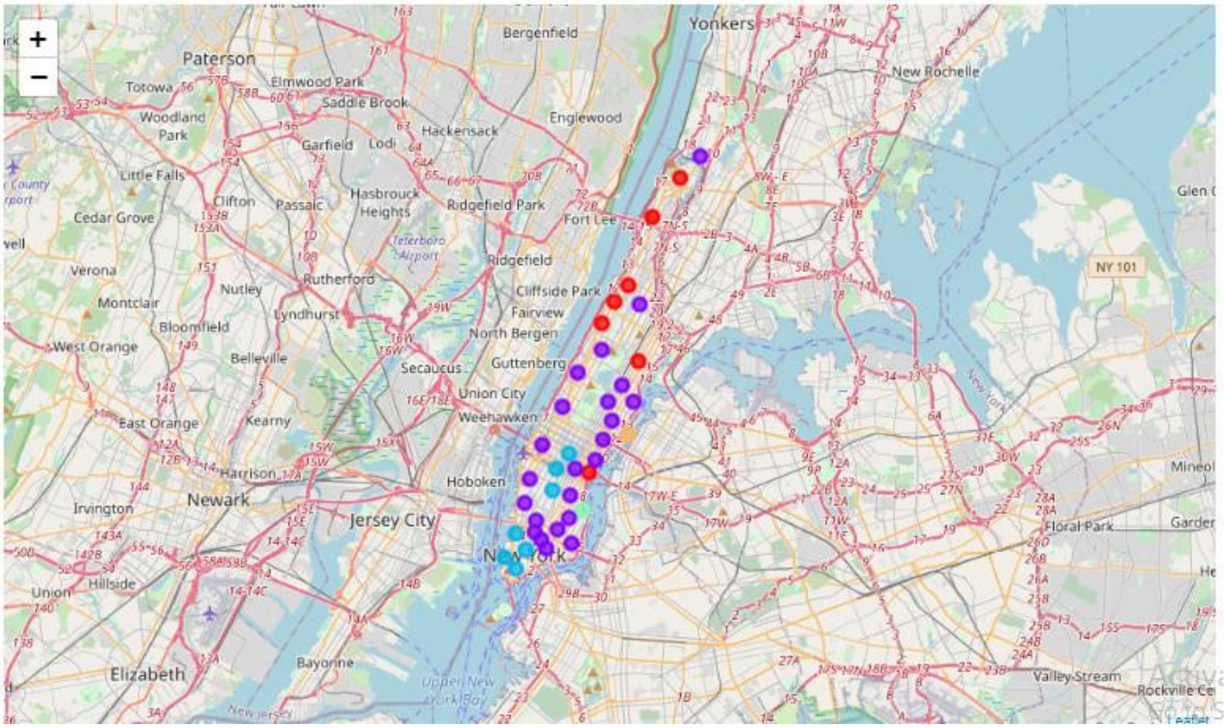
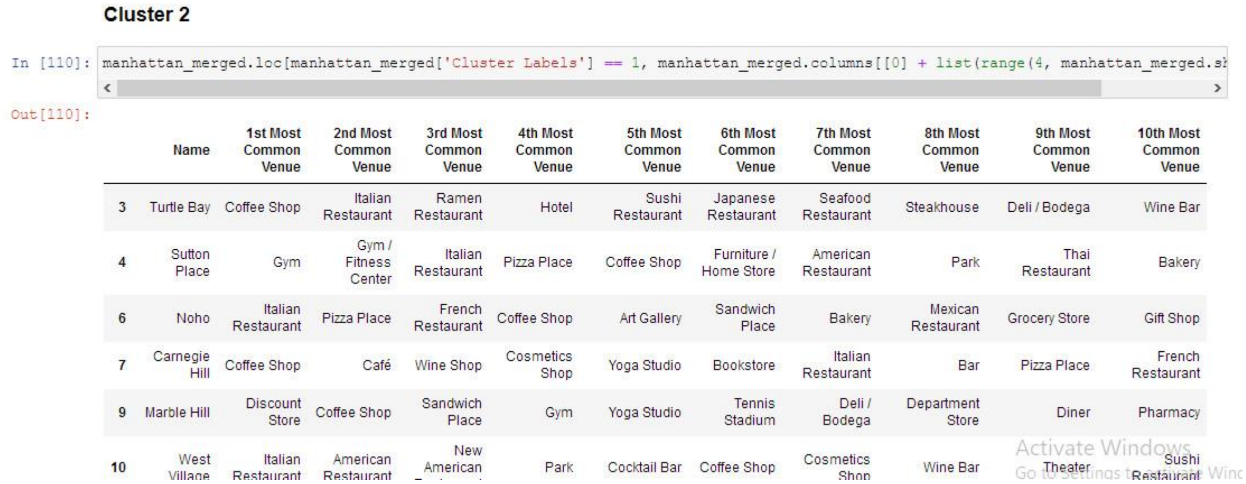


Figure 5 Manhattan K-means Clustering



Cluster 2

```
In [110]: manhattan_merged.loc[manhattan_merged['Cluster Labels'] == 1, manhattan_merged.columns[[0] + list(range(4, manhattan_merged.sh
```

Out[110]:

	Name	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
3	Turtle Bay	Coffee Shop	Italian Restaurant	Ramen Restaurant	Hotel	Sushi Restaurant	Japanese Restaurant	Seafood Restaurant	Steakhouse	Deli / Bodega	Wine Bar
4	Sutton Place	Gym	Gym / Fitness Center	Italian Restaurant	Pizza Place	Coffee Shop	Furniture / Home Store	American Restaurant	Park	Thai Restaurant	Bakery
6	Noho	Italian Restaurant	Pizza Place	French Restaurant	Coffee Shop	Art Gallery	Sandwich Place	Bakery	Mexican Restaurant	Grocery Store	Gift Shop
7	Carnegie Hill	Coffee Shop	Café	Wine Shop	Cosmetics Shop	Yoga Studio	Bookstore	Italian Restaurant	Bar	Pizza Place	French Restaurant
9	Marble Hill	Discount Store	Coffee Shop	Sandwich Place	Gym	Yoga Studio	Tennis Stadium	Deli / Bodega	Department Store	Diner	Pharmacy
10	West Village	Italian Restaurant	American Restaurant	New American	Park	Cocktail Bar	Coffee Shop	Cosmetics Shop	Wine Bar	Theater	Sushi Restaurant

Figure 6 Manhattan Most Common Cluster (Number 2) (Not full list of neighborhoods)

3.5. Queens Analysis

This time we can see a clear dominance of the cluster number 7 that consist of donut shops, Chinese restaurants and delis and bodegas, but this cluster doesn't have a clear pattern we can look at. We can see a bit of Asian influence and can be attractive to people looking for this kind of venues, or by the sea side looking for site near to beaches.

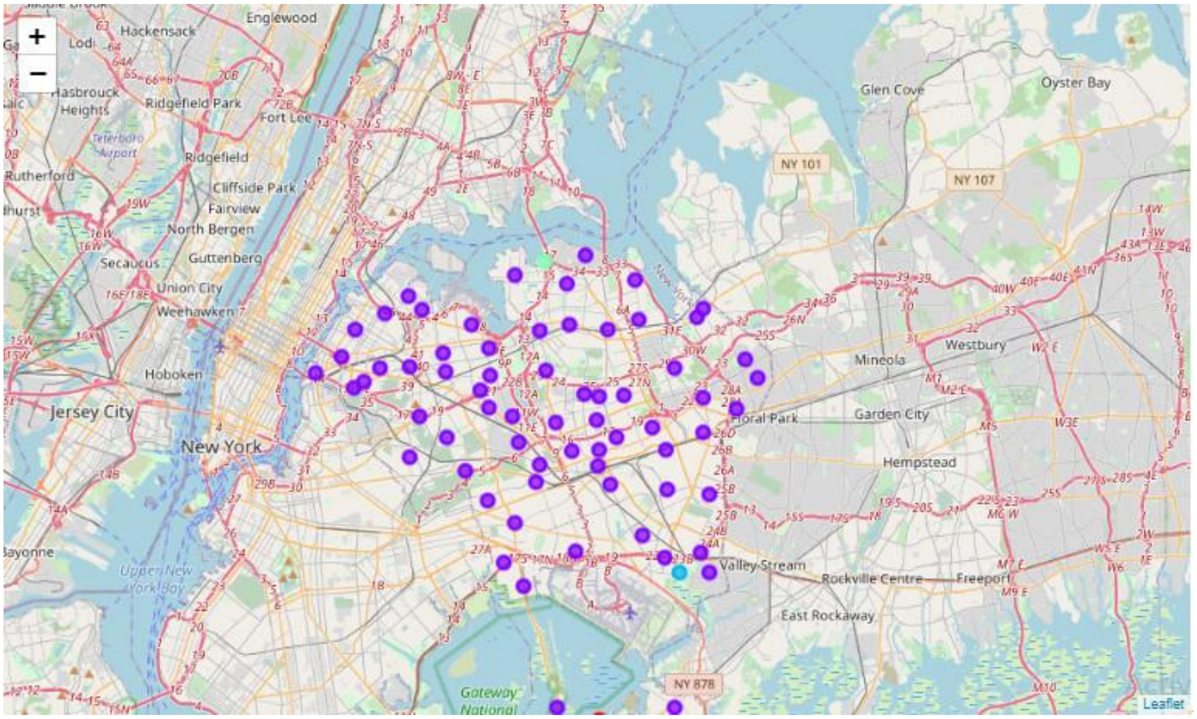
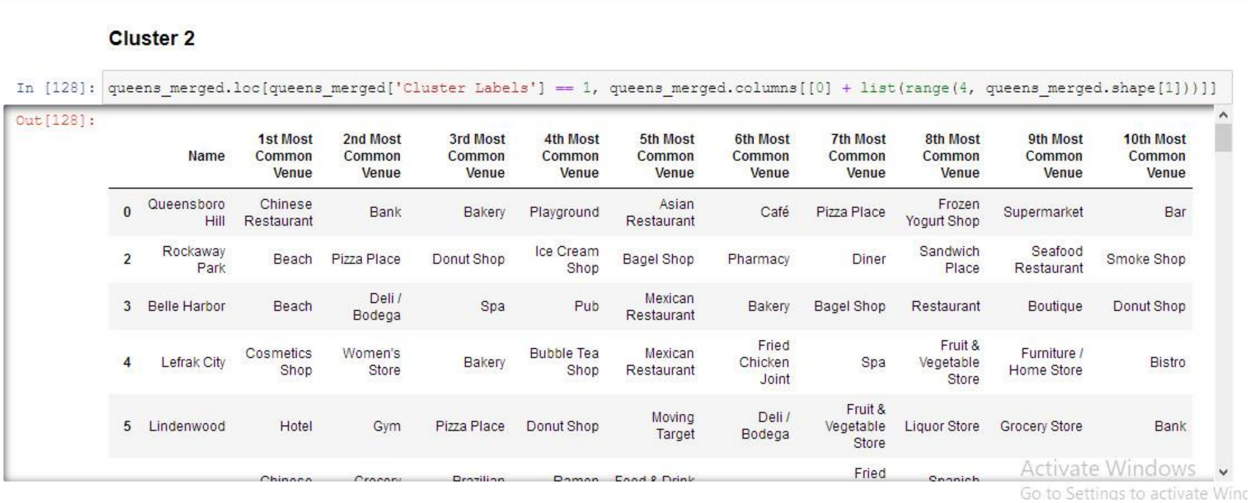


Figure 7 Queens K-means Clustering



Cluster 2

```
In [128]: queens_merged.loc[queens_merged['Cluster Labels'] == 1, queens_merged.columns[[0] + list(range(4, queens_merged.shape[1]))]]
```

Out[128]:

	Name	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	Queensboro Hill	Chinese Restaurant	Bank	Bakery	Playground	Asian Restaurant	Café	Pizza Place	Frozen Yogurt Shop	Supermarket	Bar
2	Rockaway Park	Beach	Pizza Place	Donut Shop	Ice Cream Shop	Bagel Shop	Pharmacy	Diner	Sandwich Place	Seafood Restaurant	Smoke Shop
3	Belle Harbor	Beach	Deli / Bodega	Spa	Pub	Mexican Restaurant	Bakery	Bagel Shop	Restaurant	Boutique	Donut Shop
4	Lefrak City	Cosmetics Shop	Women's Store	Bakery	Bubble Tea Shop	Mexican Restaurant	Fried Chicken Joint	Spa	Fruit & Vegetable Store	Furniture / Home Store	Bistro
5	Lindenwood	Hotel	Gym	Pizza Place	Donut Shop	Moving Target	Deli / Bodega	Fruit & Vegetable Store	Liquor Store	Grocery Store	Bank
		Chinese	Grocery	Brazilian	Ramen	Food & Drink		Fried	Spanish		

Figure 8 Queens Most Common Cluster (Number 2) (Not full list of neighborhoods)

3.6. Staten Island Analysis

In Figure 9, we can see that there is clearly a cluster that is more common that the others. In this specific cluster that is the number 1, we can see that most of this venues are coffee shops, some restaurants, mostly Chinese and American ones. Common people that like family restaurants and Chinese food can look forward to living in Staten Island that most of the neighborhood have this characteristics.

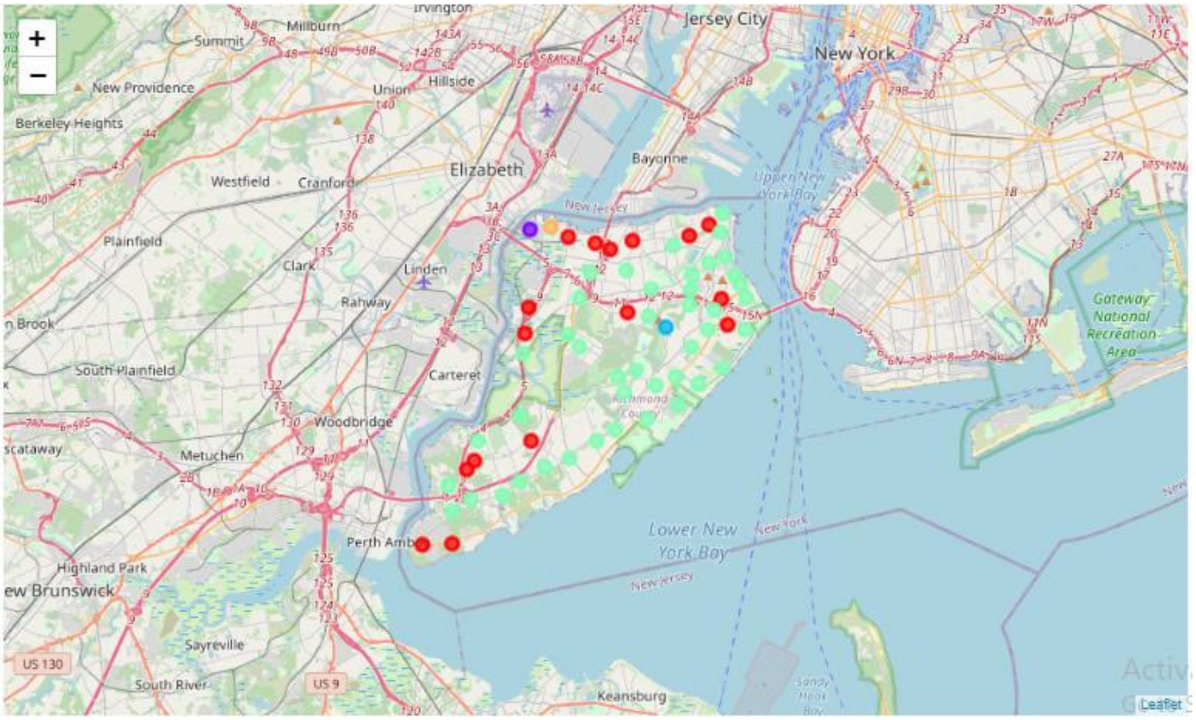
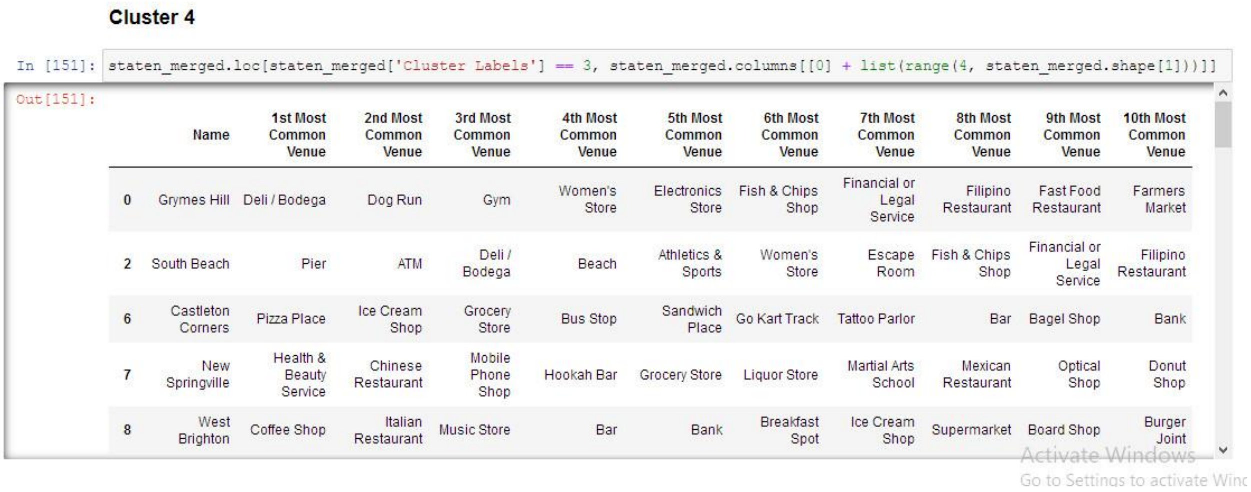


Figure 9 Staten Island K-means Clustering



Cluster 4

```
In [151]: staten_merged.loc[staten_merged['Cluster Labels'] == 3, staten_merged.columns[[0] + list(range(4, staten_merged.shape[1]))]]
```

Out[151]:

	Name	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	Grymes Hill	Deli / Bodega	Dog Run	Gym	Women's Store	Electronics Store	Fish & Chips Shop	Financial or Legal Service	Filipino Restaurant	Fast Food Restaurant	Farmers Market
2	South Beach	Pier	ATM	Deli / Bodega	Beach	Athletics & Sports	Women's Store	Escape Room	Fish & Chips Shop	Financial or Legal Service	Filipino Restaurant
6	Castleton Corners	Pizza Place	Ice Cream Shop	Grocery Store	Bus Stop	Sandwich Place	Go Kart Track	Tattoo Parlor	Bar	Bagel Shop	Bank
7	New Springville	Health & Beauty Service	Chinese Restaurant	Mobile Phone Shop	Hookah Bar	Grocery Store	Liquor Store	Martial Arts School	Mexican Restaurant	Optical Shop	Donut Shop
8	West Brighton	Coffee Shop	Italian Restaurant	Music Store	Bar	Bank	Breakfast Spot	Ice Cream Shop	Supermarket	Board Shop	Burger Joint

Figure 10 Staten Island Most Common Cluster (Number 4) (Not full list of neighborhoods)

3.7. New York Analysis

In the map of New York we can see that there is clearly a cluster that is the most common one, the number 3 in our k-means procedure, we can see that most of them have as common place restaurants be it American or Italian and pizza places. In the first ten we can’t see a clear pattern and we can’t expect to see any clear pattern around this type of cluster.

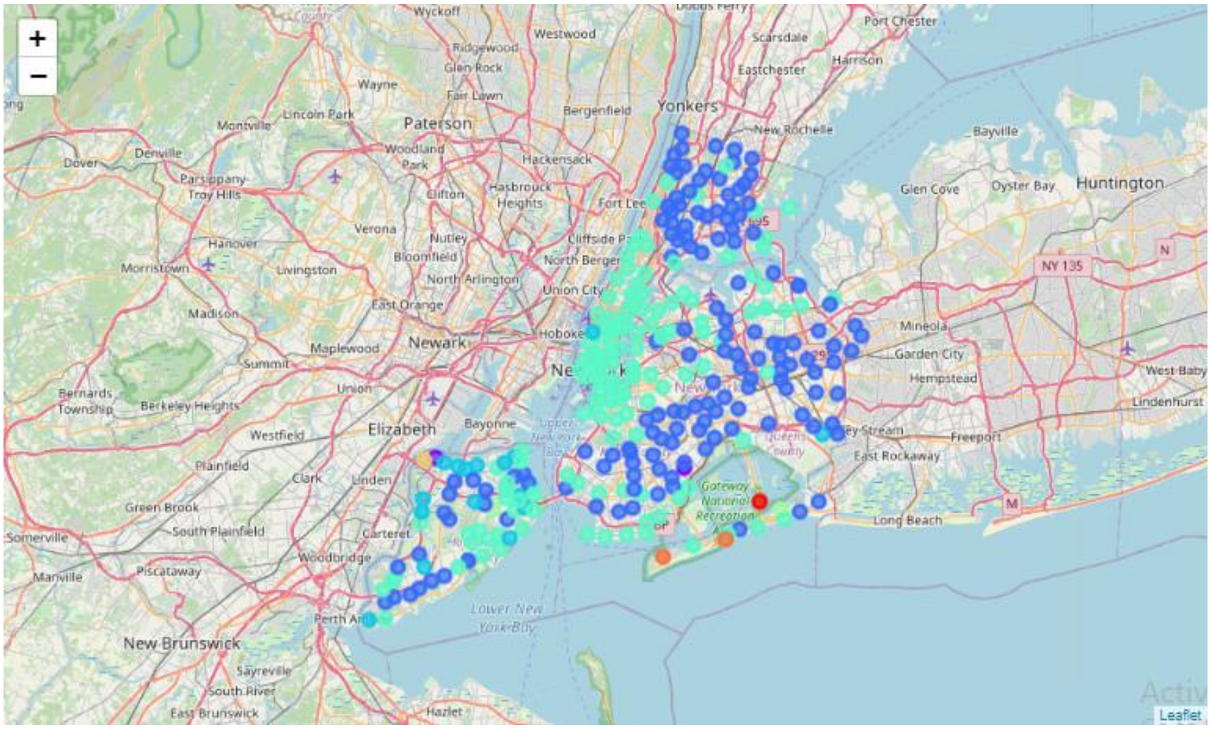
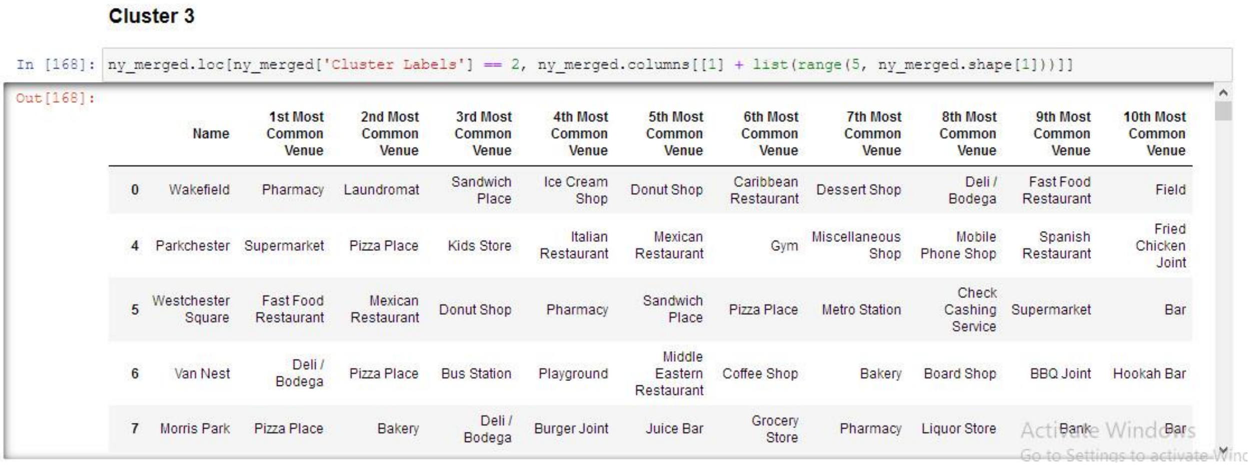


Figure 11 New York K-means Clustering



Cluster 3

```
In [168]: ny_merged.loc[ny_merged['Cluster Labels'] == 2, ny_merged.columns[[1] + list(range(5, ny_merged.shape[1]))]]
```

Out[168]:

	Name	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	Laundromat	Sandwich Place	Ice Cream Shop	Donut Shop	Caribbean Restaurant	Dessert Shop	Deli / Bodega	Fast Food Restaurant	Field
4	Parkchester	Supermarket	Pizza Place	Kids Store	Italian Restaurant	Mexican Restaurant	Gym	Miscellaneous Shop	Mobile Phone Shop	Spanish Restaurant	Fried Chicken Joint
5	Westchester Square	Fast Food Restaurant	Mexican Restaurant	Donut Shop	Pharmacy	Sandwich Place	Pizza Place	Metro Station	Check Cashing Service	Supermarket	Bar
6	Van Nest	Deli / Bodega	Pizza Place	Bus Station	Playground	Middle Eastern Restaurant	Coffee Shop	Bakery	Board Shop	BBQ Joint	Hookah Bar
7	Morris Park	Pizza Place	Bakery	Deli / Bodega	Burger Joint	Juice Bar	Grocery Store	Pharmacy	Liquor Store	Bank	Bar

Figure 12 New York Most Common Cluster (Number 3) (Not full list of neighborhoods)

4. Results

With the k-means clustering it was a success in classifying the different neighborhoods in the boroughs. The clusters in each of the borough most of the neighborhoods fall into one of the labels and the others are left like outlier neighborhoods, so if you are looking for a similar neighborhood in the same borough you will likely be in the common cluster and find a similar neighborhood without a problem, for the outliers it will be a tough search.

Now analyzing the New York City clustering, the analysis wasn’t as effective as the other clustering. This kind of process won’t be as effective with data that is so broad. Starting with New York having a lot of stores and restaurants repeated in a lot of neighborhoods, it can be hard to look in an analytic way which neighborhoods are different.

5. Discussion

Finding a clean and organized dataset is of vital importance to have a smooth work, so it’s important to remark the great work of the open data of New York. Having the right tools it’s what makes this work possible. Even if the Foursquare API worked this time, it has some minor problems and it would be best to avoid this kind of situations when doing a work, so it’s better to study which tools you are going to use before working in the code and use another type of venues API if you have access or knowledge of one.

6. Conclusion

I was able to achieve the aim of the project in doing a successful clustering of the boroughs. The city of New York have a lot of different features, this study didn’t include things like urban areas, atmospheres and lifestyle of the people living in it. So even if the neighborhood similitude is of vital importance for this analysis, to improve this study it might be helpful to add more variables that could be helpful to generate more insights.