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.

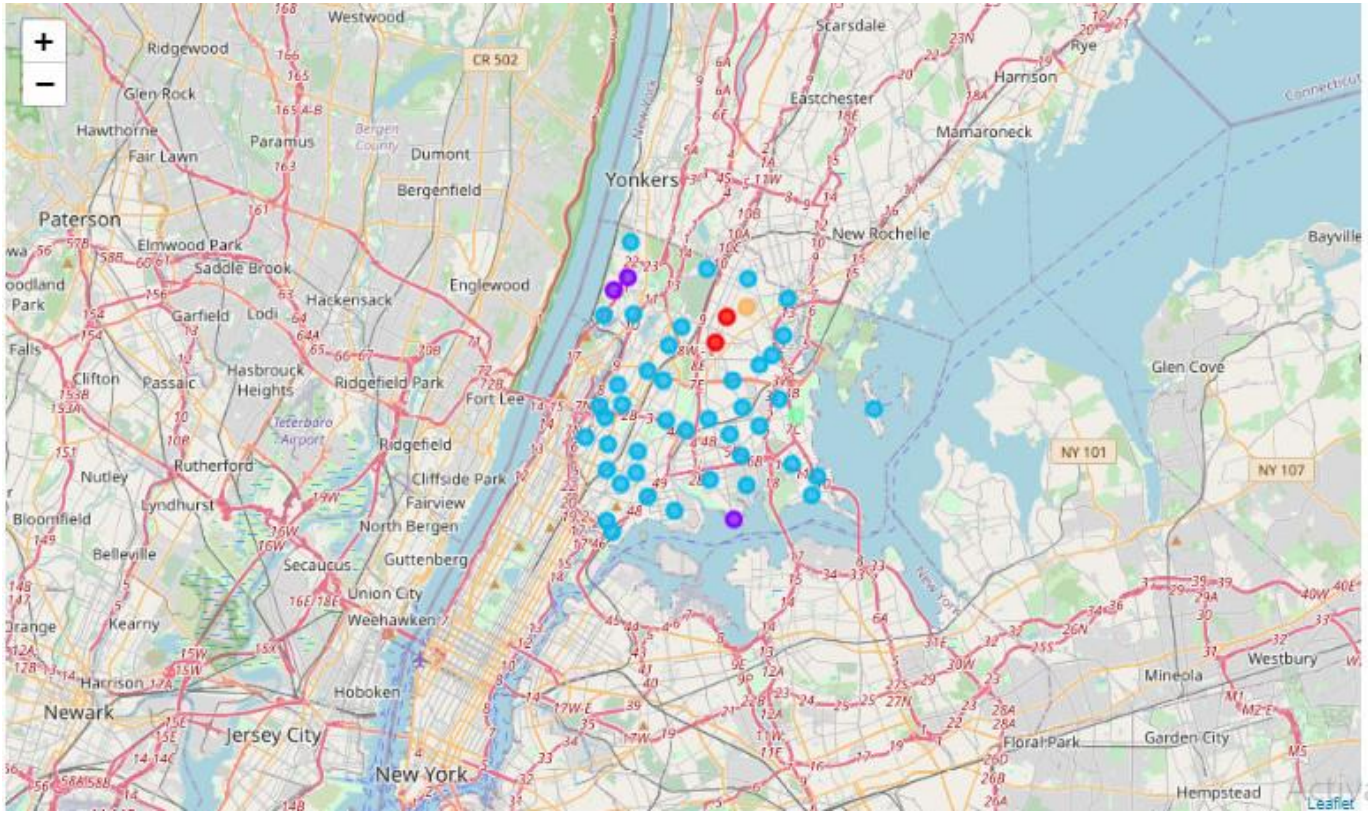
3. Methodology

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.

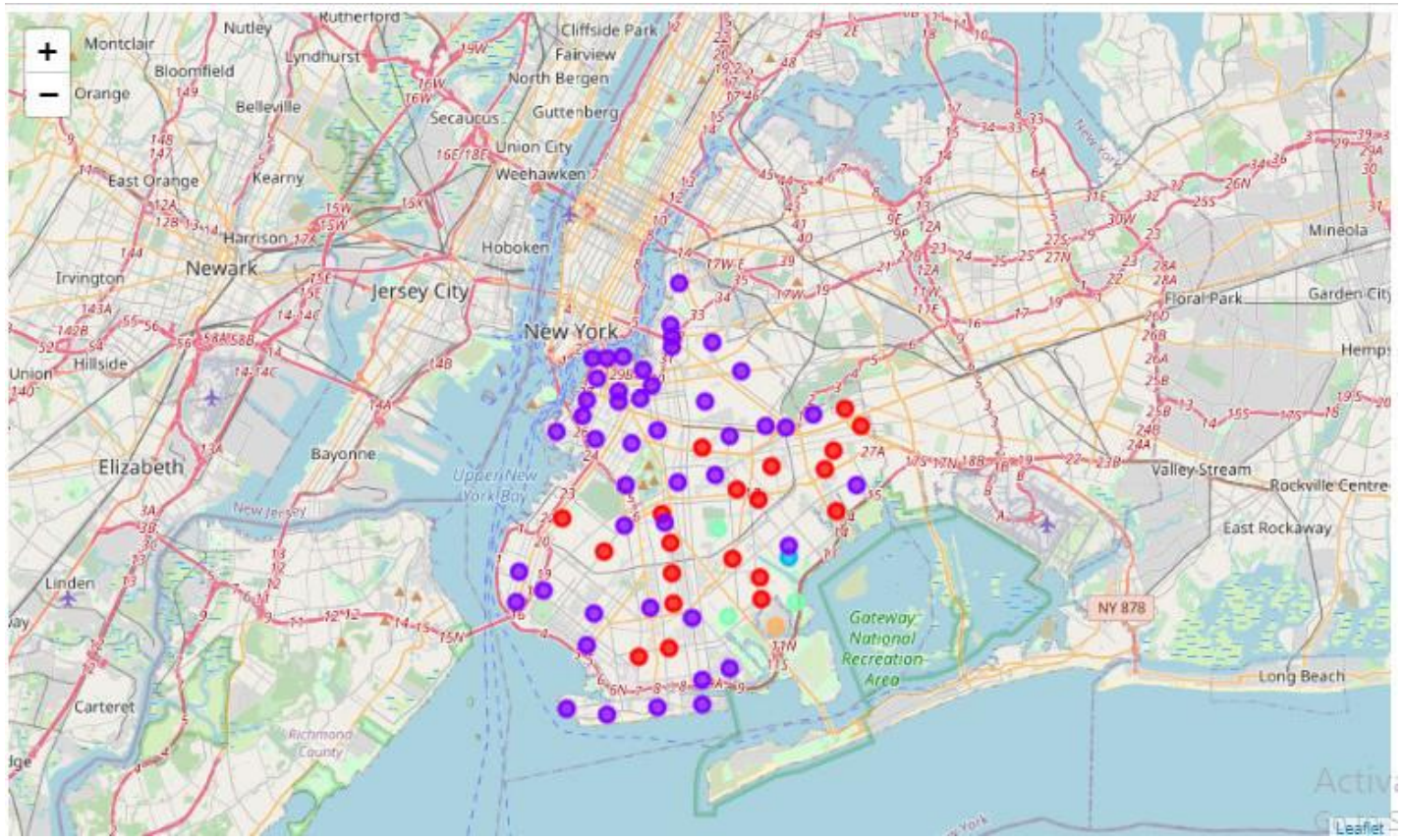
3.2. Bronx Analysis

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.



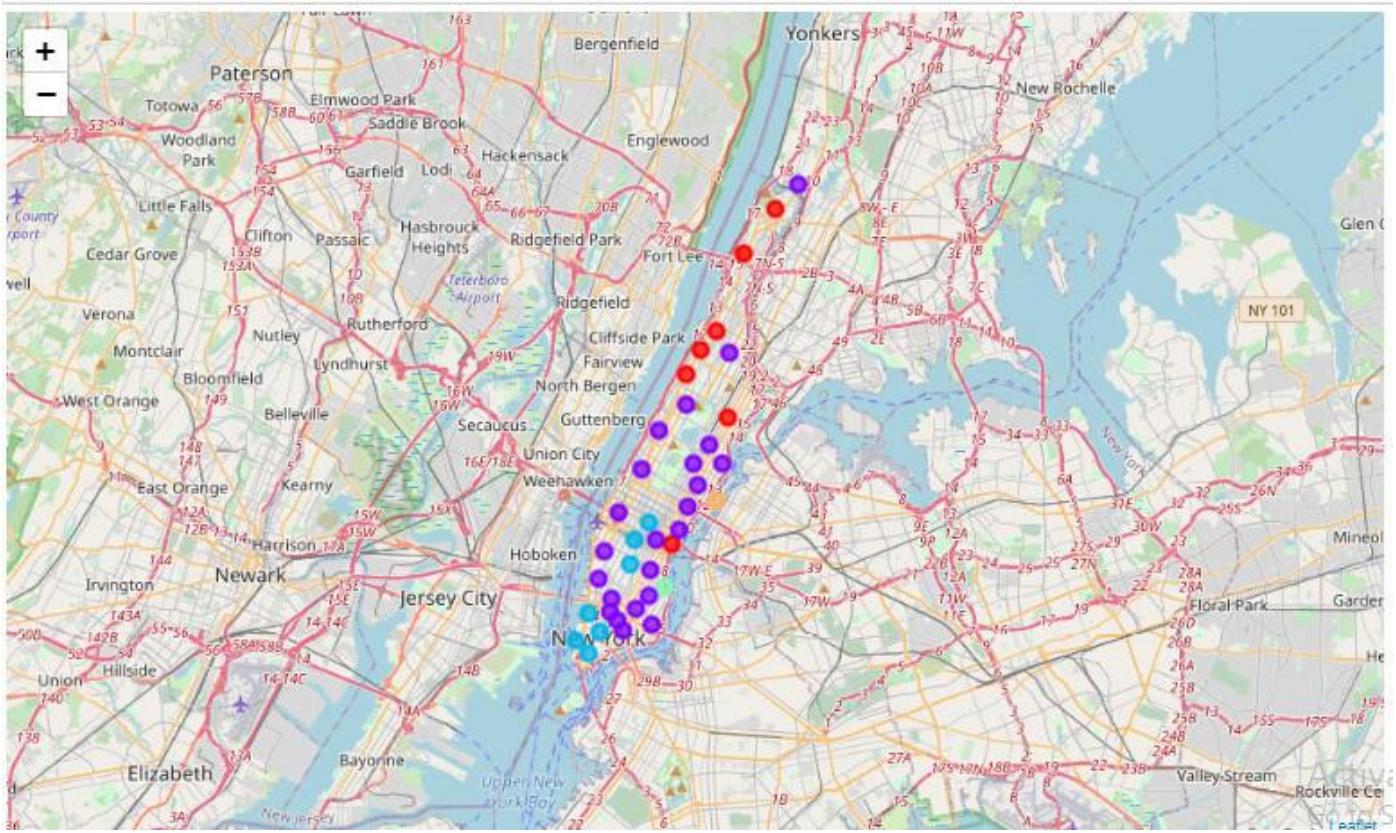
3.3. Brooklyn Analysis

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



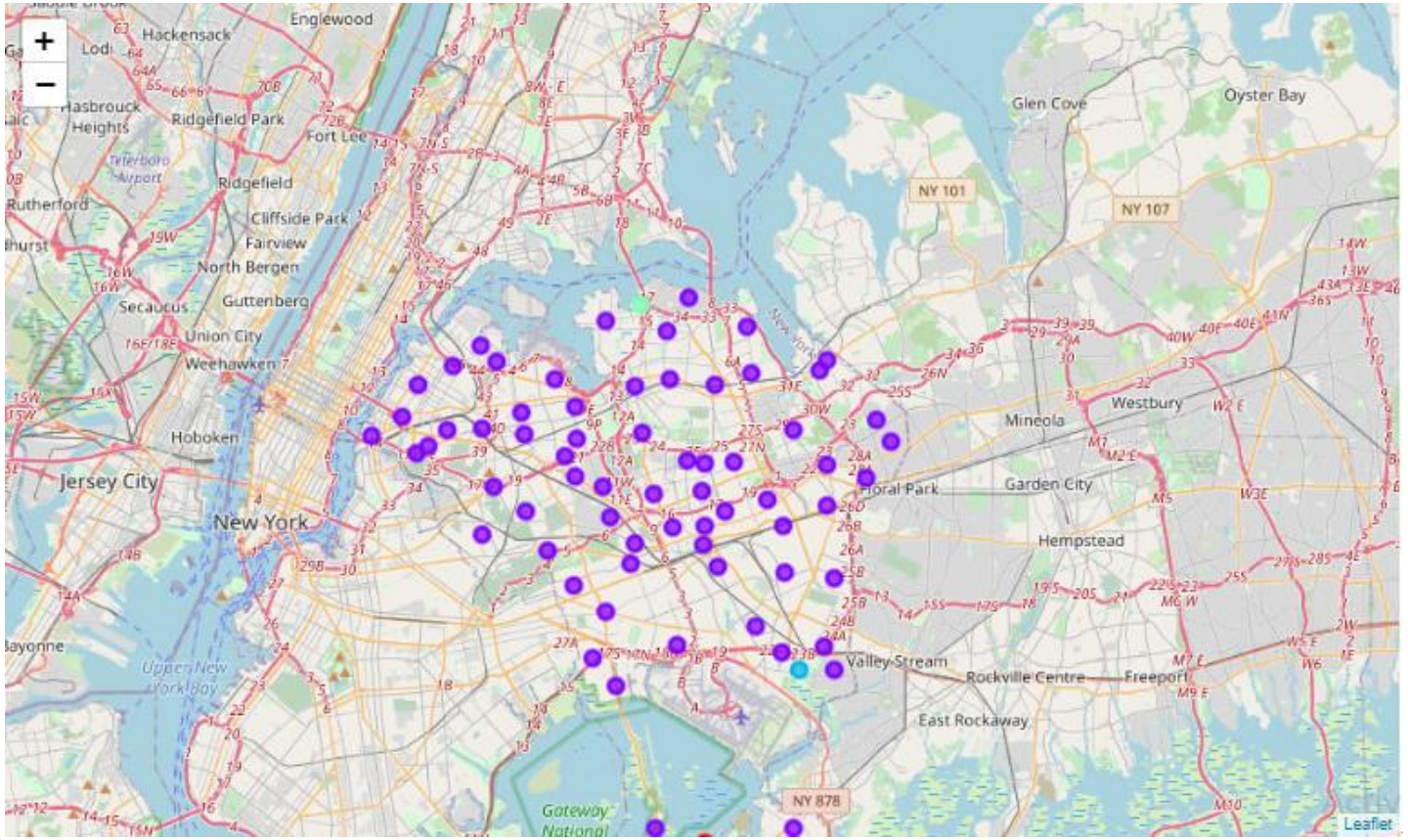
3.4. Manhattan Analysis

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.



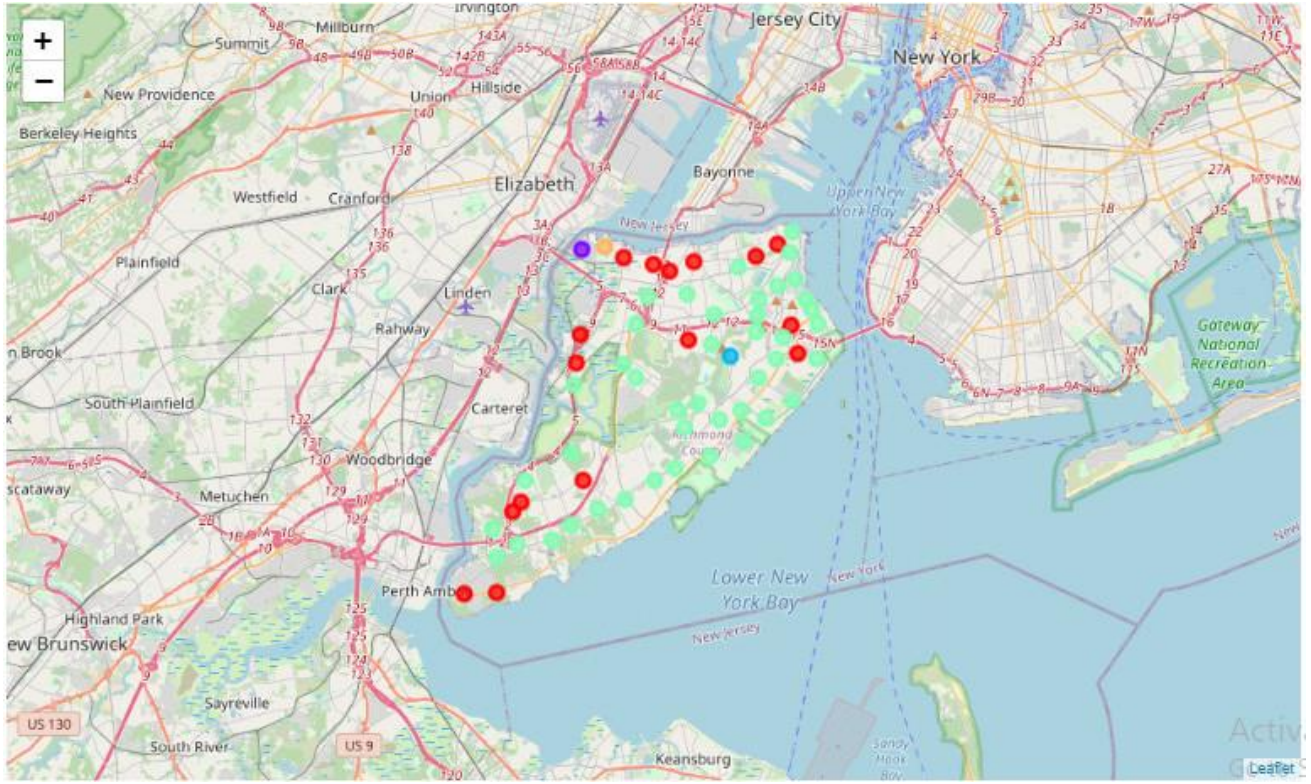
3.5. Queens Analysis

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.



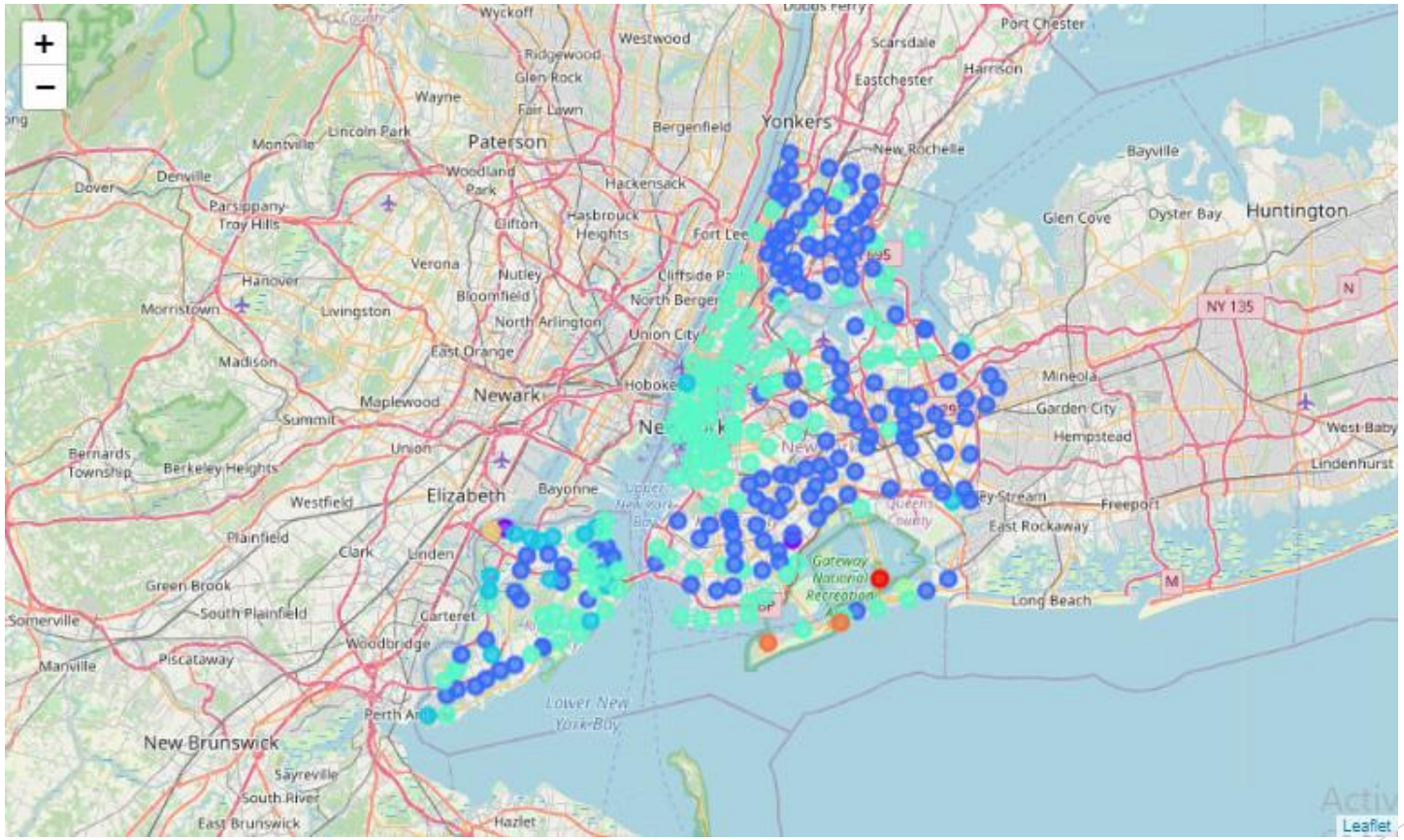
3.6. Staten Island Analysis

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.



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.



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.