The Battle of the Neighborhoods

IBM Data science Capstone Project Report

1. Introduction

1.1 Problem Background

New York is the largest and one of the most populated cities in the United States. Being an economic and cultural captial of the country, it attract thousands of people seeking better job and education opportunities. It is estimated that 1 in 7 residents in some neighborhoods in Manhattan are people who migrated to NYC within the last year. Residents in these areas tend to subsequently settle deeper into the boroughs. On the other hand, many residents from outskirts of the city may seek for an apartment closer to the center.

Every day, many individuals make decisions on moving to new York City or relocating to new neighborhood. Choosing new neighborhood is very important decision that can significantly impact the quality of one's life. However, only few have sufficient amount of time to investigate in depth the data concering more than 300 neighborhoods to decide which one is the most suitable. The analytic tools that would automate this process might be therefore useful for both individuals and rental or consulting companies.

1.2 Problem Description

Our company offers rental advise and broking services. We help people to find a place to live that would fit their needs. One of our current clients is a family that lives in neighborhood of Bay Terrace, NY. Recently one of the parents received an attractive job offer in a bank located on Wall Street. Since the distance between these two locations is quite big, the family decided to search for a similar neighbourhood in terms of safety, park accessibility, social status of residents etc, that would be closer to the office - no more than 15 km apart. The traffic congestion and commuting time should be not worse than at Bay Terrace. Moreover, since one of the children is disabled, they would like to choose location that has at least one special school in the immediate neighborhood. The family was contend with the accessibility to different venues in Bay Terrace such as supermarkets, playgrounds, resteurants etc., so they would like to live in a neighbourhood that has similar venues, if possible.

1.3. Target Group

The target audiance of this report would be:

a) individuals who need advice on relocation
b) real estate maklers and planners that would seek for the suitable location for new housing investment
c) rental companies, for personalized advertisements and better rental advice for client

2. Data

To solve the above mentioned problem we would need the following data:

1. List of New York city neighborhoods and their coordinates
2. Data on characteristics of the neighborhoods, such as crime rate, parks availability, rent prices, commuting time, cleanliness of the streets etc.
3. Venue data, in particular data regarding the location of special schools in neighborhoods

The list of 306 NYC neighborhoods and coordinates used in the analysis is available here.

We get **the data on the profiles of neighborhoods** from the government site on NYC Community District Profiles. The dataset is provided in csv format and includes 59 observations corresponding to the number of community districts and nearly 200 variables. However, not all of them are useful for the analysis. We decided to use ten of them, namely:

1. qcd_full_title - name of Community District
2. pct_lot_area_open_space - parkland and open space area (as % of total CD area)
3. crime_per_1000 - number of crimes per 1000 people in CD
4. pct_hh_rent_burd' - Percentage of households that spend 35% or more of their income on rent
5. mean_commute - Mean commute time to work for residents
6. unemployment_cd - unemployment in CD
7. neighborhoods_x - neighborhoods included in the CD
8. pct_clean_strts - percentage of streets rated acceptably clean in the CD
9. poverty_rate - poverty rate in CD 10 .pct_lot_area___industrial_manufacturing - percentage of lot area in CD with industrial/manufacturing buildings

To match the coordinates from the first dataset with proper indicators from the second dataset, we firstly transform the values in the column 'neighborhoods' in NYC CD Profiles data into rows. As a result, we obtain a dataframe where each row represents one neighborhood. Then, we match the neighborhoods with the corresonding indicators using variable 'qcd_full_title' (name of Community District). Finally, we match the resulting dataset with the list of 306 NYC neighborhoods with coordinates using the neighborhood name. In the end, the duplicated neighborhoods are removed.

Foursquare API is used to obtain the **venue data for the neighborhoods** in New York. Foursquare API provides access to data of over 100 million places including the information on category of the venue and its accurate location. IT is possible to query for the of most popular venues nerby the specified locations. We will use the previously mentioned list of neighboroods to get the list of the top venues in particular location and compare it with the top venues in the current neighborhood of the client.

3. Methodology

The objective of the analysis is to find a neighborhood that:

1. Would be similar to Bay Terrace in regards of the selected indicators
2. Meets the requiremends of proximity to the Wall St office, proximity to special school and similar accessibility to venues as in Bay Terrace.

In order to achieve the first objective, k-means clustering method will be used. K-means clustering is popular unsupervised machine learning algorithm that splits the data into k groups based on the distance between the observations. In other words, the K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. The target number of groups *k* is pre-defined by the researcher. In this study, we decided to use k = 4, as we might expect four different types of neigborhoods: strict center, center, the inner part and outskirts of the city. Once the neighborhoods are clustered, we will identify the cluster to which Bay Terrace was classified and search for suitable candidate location within this cluster. Bt using geopy python package and Foursquare API we will filter the shortlisted neighborhoods based on the distance from the office and special schools nearby. Finally, we will download the data on the top venues nerby the shortlisted neigborhoods. Again, we will use k-means clustering to find out which neighborhoods provide similar accessibility to the venues as Bay Terrace.

4. Results

4.1 Explanatory data analysis

Before applying the clustering, we explore the dataset and get descriptive statistics. Firstly, we map the neighborhoods featured in the dataset. The map of New York neigborhoods is presented on the next page in Figure 1. Bay Terrace Neighborhood and te 13 km radius around the Wall St. 111 were marked red. The neighborhoods within the circle meet the requirement of proximity to the office. Notice that most of them are located in Manhattan, Brooklyn or Queens.

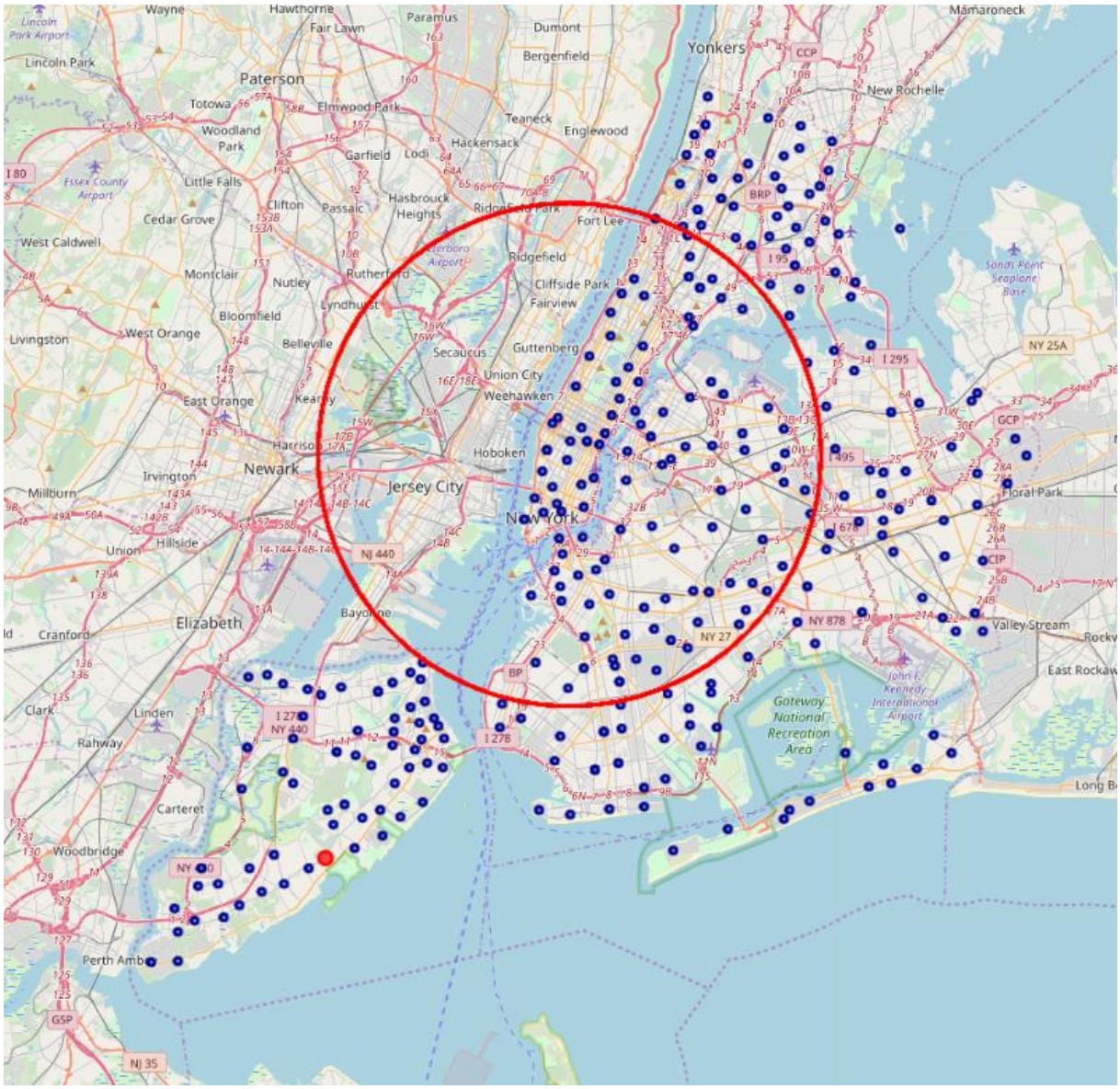
Next, we obtain mean values of indicators for each borough to get better understanding of characteristics of boroughs. The result is presented in the table below.

Table 1 – mean values of indicators for every borough

	Borough	Open_area	Crime	Street_clean	Rent_Burden	Commute_time	Unemployment	Poverty	Industrial_area	Latitude	Longitude
0	Bronx	0.124235	13.605862	95.301887	48.764151	43.830189	7.386792	24.007547	0.040005	40.850820	-73.873735
1	Brooklyn	0.136358	11.366159	93.746377	44.779710	42.710145	5.465217	20.420290	0.051886	40.646101	-73.952079
2	Manhattan	0.122914	19.313509	96.156757	36.575676	30.459459	3.975676	12.951351	0.012707	40.762799	-73.975752
3	Queens	0.167045	9.162469	96.580247	46.992593	43.864198	5.024691	19.344444	0.044659	40.707201	-73.826134
4	Staten Island	0.213780	6.764254	98.239683	45.911111	42.809524	3.676190	16.384127	0.027841	40.589165	-74.136091

The table shows that boroughs of New York share different qualities. Manhattan, which is the central and most densely populated borough, features the highest crime rate, lowest commute time and lowest fraction of industrial area. These characteristics are typical for the centres of big cities.

Figure1 – New York neighborhoods



The relatively low rent burden in Manhattan can be explained by higher wealthiness of inhabitants. On the other end of the spectrum, Staten Island Borough that is furthest from the center has the lowest crime rate, the highest amount of parks and open areas and the cleanest streets. Brooklyn and Bronx that are in the more inner part of NYC have higher crime and poverty rates and lower park area. Based on the table, we see that the borough that share similar characteristics with Staten Island is Queens borough.

Afterwards we explore the relations between variables using scatterplot matrix, presented in Figure 2. The following dependencies can be noticed:

a) Poverty is strongly and positively correlated with rent burden and unemployment and negatively correlated with street cleanliness

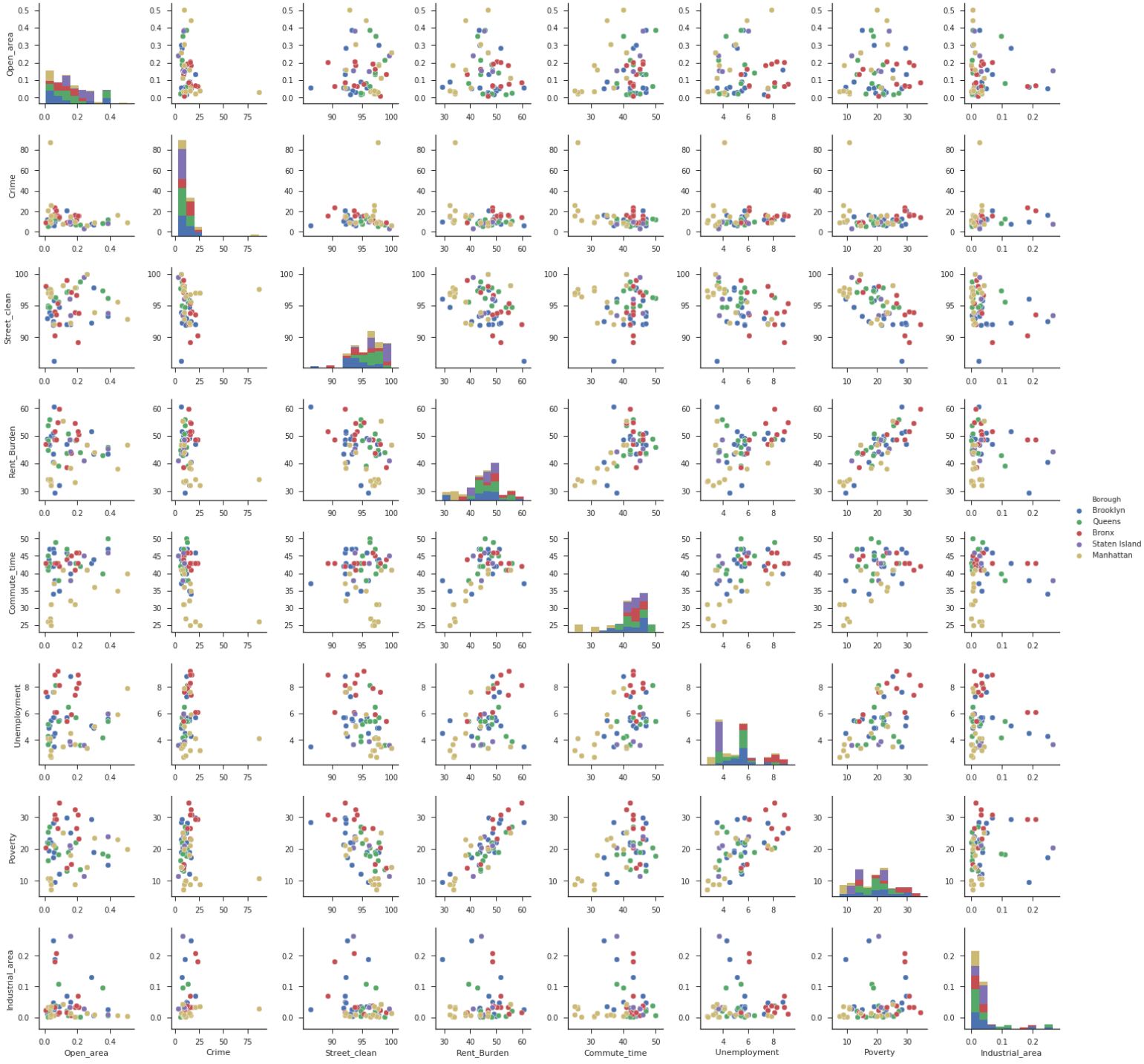
b) One of the observations is an outlier in crime rate. This observation is Flariton neighborhood located in Manhattan. If we ignore the outlier, the following relationship can be implied from the plot: crime is positively correlated with unemployment and negatively correlated with cleanliness. The relation between crime and poverty seems

to be non-linear: crime tends to be higher in neighborhoods with high and low poverty rate, which makes sense: people living in poverty seek other, sometimes illegal sources of income. At the same time, rich districts are more vulnerable to crime, since they might be valuable target for thieves, robbers etc.

c) Rent burden is positively correlated with poverty, unemployment and commuting time.

d) The relations between size industrial and open areas with other variables are not clear

Figure 2 – Scatterplots of selected variables



4.2 Clustering

The clustering analysis is performed based on variables *Open_area, Crime, Street_clean, Rent_Burden, Commute_time, Unemployment, Poverty, Industrial_area*. The number of clusters was set to 4. In order to account for different value ranges of indicators, the data was normalized before applying clustering. The number of neighborhoods assigned to the particular clusters is displayed in the Table 2.

Table 2 – Cluster assignment frequency

	Cluster
0	153
3	108
2	39
1	3

The first two clusters contain the vast majority of neigborhoods. The third cluster includes almost 40 neighborhoods, while the last one - only 3. To get better understanding of the cluster assignment a map is created. Looking at the map, one can notice a pattern in clustering. The smallest cluster 1 contains 3 neighborhoods in the center of Manhattan. Almost all remaining neighborhoods in Manhata and neighborhoods in Queens and Brooklyn that are close to center were assigned to Cluster 2. Cluster 0 (in red) contains majority of neighborhoods in Brooklyn, Queens and Bronx. The last cluster consists of neighborhoods in the outskirts, mainly in Staten Island.

Table 3 –Mean indicator values for each cluster

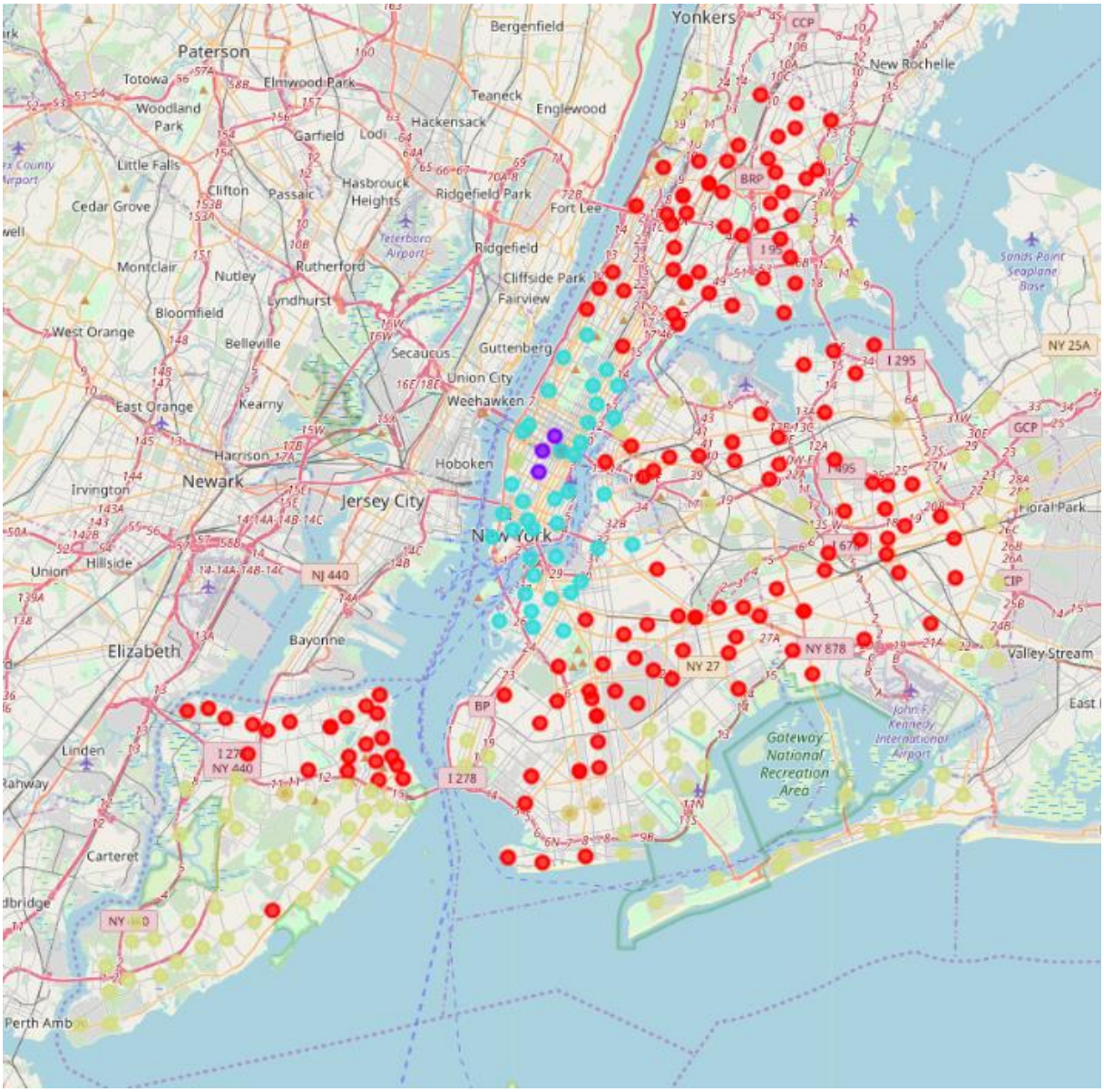
Cluster	Open_area	Crime	Street_clean	Rent_Burden	Commute_time	Unemployment	Poverty	Industrial_area	Latitude	Longitude
0	0.194200	8.449518	96.906952	45.245455	43.743316	4.893048	17.620321	0.031749	40.669172	-73.956558
1	0.031681	87.226379	97.600000	34.200000	26.000000	4.100000	10.800000	0.026816	40.747625	-73.987110
2	0.110000	13.498240	93.717808	52.310959	43.082192	6.426027	27.184932	0.045632	40.763145	-73.897888
3	0.077485	14.029322	95.840000	33.605000	31.100000	3.947500	11.195000	0.054832	40.730274	-73.980283

Using the table and map presented in Figure 3, we can summerize the clusters as follows:

- **Cluster 0 (red)** - contains mostly neighborhoods in the inner part of the city: mostly in Bronx, Queens, Brooklyn and also part of Staten Island that is closer to the center. This cluster has low crime rate and lots of open areas. However, rent burden, poverty and commuting time are relatively high.
- **Cluster 1 (purple)** concentrates three neighborhoods in the heart of Manhattan. It has the lowest poverty rate, lowest share of open area and exeptionally high crime rate.
- The rest of the neighborhoods of Manhattan and neighborhoods that are close to Manhattan were assigned to **Cluster 2 (blue).** This cluster is characterized by the highest poverty rate and unemployment and, consequently, also high rent burden.
- Remaining neighborhoods are grouped in **Cluster 3 (olive green).** Majority of them is loated in the outer part of the city. Cluster 3 has the second highest crime rate, but communting time, unemployment and rent burden are low. It is the most industrialized area.

The current neighborhood if our client, Bay Terrace, was classified to cluster 0. It seems to be a good choice for a family who want to live in a safe, clean location with good access to parks and recreational areas.

Figure 3 – New York neighborhoods clustered. Cluster 0 – red, Cluster 1 –purple, Cluster 2 – blue, Cluster 3 - green



4.3 Including client's requirements

Now that the similar neighborhoods in the city were identified, we can filter the neighborhoods from Cluster 0 to see, whcih of them meet the requirements defined in the first part of the report. Firstly, we apply commuting time criterion: the commute time has to be no higher than the value of 41 Terrace Bay. 33 out of 153 neighborhoods in Cluster 0 meet this requirement.

Secondly, the distance from the office located on Wall St. 111 has to be no larger than 13km. We can get the distance bewteen neighborhoods and the address using geopy package. In

particular, we use functions geolocator and distance. We loop through each element in the dataframe with the shotlisted neighborhoods, calculate the distance and save it to the dataframe. The result is displayed below. Neigborhoods Concourse, Concourse Village, Auburndale, Beechhurst, College Point, Malba and Whitestone neighborhoods do not meet this requirement, therefore they were excluded from the dataset with the shortlisted neighborhoods.

The next requirement is the immediate proximity of special school. To search for such venues nearby neighborhood, Foursquare API was used. The list below shows number of special schools in the proximity of 500 meters.

- Bushwick: 5
- Borough Park: 3
- Kensington: 1
- Ocean Parkway: 2
- East Elmhurst: 3
- Jackson Heights: 5
- North Corona: 0
- Flushing: 1
- Murray Hill: 10

The last request regarded the venues in the candidate neighborhood. The types of the venues in the neighborhood should be similar to those in Bay Terrace. Again, Foursquare API was used to get types of the top 50 venues in the proximity of 500 meters. Then, we cluster the shortlisted nieghborhoods into two groups, based on the venues exclusively. Out of the 9 remaining shortlisted neighborhoods, only East Elmhurst was classified to the same group as Bay Terrace. Therefore, it was selected as recommendation for our client

5. Discussion

The analysis revealed different qualities of neighborhoods in New York City. However, the scope of indicators used was limited and included only the most important and easily accessible ones. Finding and merging datasets with information that might be important for housing decision is challenging. The further analysis should also include the absolute rent, maintenance and parcel prices for different neighborhoods - unfortunately not all this data is easily accessible. Another obstacle is API Foursquare call and queries quota for personal account - because of it we were only able to explore the venues nearby the shortlisted neighborhoods. Including venues in all 300 neighborhoods and including it in primary cluster analysis would provide more valuable insights.

6. Conclusion

New York neighborhoods are very diverse and can differ significantly from each other in terms of the used indicators. The analysis show how we can group neighborhoods based on

their qualities and how different indicators are correlated to each other. In general, research confirmed widely known facts that unemployment and poverty are positively correlated. Crime tends to higher in low and high poverty regions. People in poorer neighborhoods spend higher fraction of their budget on accomodation. The safest neighborhoods are located in the inner part of the city - center and outskirts ten to have higher crime rate. We recommended Cluster 0 with low crime rate and large recreational areas for the family client, but individual preferences may vary: people who value short commuting time should consider neighborhoods in Cluster 1 (strict center). Clusters 1 and 3 may be preffered for those who seek job, as they have very low unemployment rate.