Comparing Neighborhoods of New York and Toronto

By : Qamar Zia

1 Introduction

1.1 Background

New York city is the most densely populated city in United States with an estimated population of 8,336,817 people living in the city. New York is also known as the cultural, financial, and media capital of world. It is known for the diversity in culture in its all Five boroughs including Brooklyn, Queens, Manhattan the Bronx and Staten Island. The city is so much diverse in culture that as much as 800 different languages are spoken in it. On the other hand Toronto is the provincial capital of Ontario and the most densely populated city in Canada with an estimated population of 5,928,040 . It is the 4th most populated city in North America. Toronto like New York is also the business hub and is very diverse in culture. People have travelled from all around the world and make Toronto their home . Toronto encompasses of several municipalities including East York, Etobicoke, Forest Hill, Mimico, North York, Parkdale, Scarborough, Swansea, Weston and York. Comparing the neighborhoods of both cities will give us great deal of insight like similarity and differences between two cities. And on the bases of these insights we can help people in making a choice of moving in a different city yet in similar surroundings.

1.2 Problem Statement

In todays world people are often moving from one place to other for work or for change of life and they also want to be in a place that is much like home. For instance a person comes too you saying that he is moving from New York to Toronto and he is so much confused that where should he find a home in Toronto. By Clustering the two cities on the bases of features we can help the person in his confusion and help him make a decision.

2 Data Gathering and Data Wrangling

2.1 Data Acquisition

For this project I will need two data sets one containing the data of boroughs of New York and the other one containing Data for Toronto. Luckily the data for New York was available on internet with its longitude and latitude values at Here in a geo json file. For the data of Toronto it was not directly available and the names and postal codes were available on Wikipedia Here.

2.2 Data Wrangling

For the data of New York which was already available on Internet in a Geo Json file consisted of different features which I did not need for this project after exploratory analysis of the file using I came to know that the important values are under “features” category so I extracted the data and made a Data frame using Python Pandas library the final Data frame Consisted of 4 columns “Borough, Neighborhood, Latitude and Longitude.

As for the data of Toronto which was not directly available as a file I had to work for data accusation, Searching for thee internet I came across the Wikipedia page containing a table of Postal codes of Toronto as well as borough and neighborhoods. I started web scrapping for data and done it by using Python Beautiful Soup Library , After extraction of table and converting it into data frame I saw a problem of missing values as many of the borough were not assigned with any value so I dropped the rows which consisted of “Not Assigned” values . Further analysis revealed that a single neighborhood consisted of multiple postal codes with unassigned values so I combined those values in a single row. The new data frame consisted of three columns Postal Code , Borough and Neighborhood. I still needed two more columns like Longitude and Latitude for completing the data. For those two columns I assigned Each column with respective longitude and latitude.

2.3 Feature Selection

After successfully gathering and cleaning data on both Toronto and New York with their neighborhoods and their Longitude Latitude values we now need some key features by which we can compare the two cities .

For this purpose I used Foursquare API . Foursquare API provides with an access to an enormous database consisting of venues from all around the world including variety of information such as Venues, Common places , and other tips etc. Foursquare API is a popular API used by big companies like Apple Maps, Snapchat, Twitter which is a proof that using this API is a good way to explore neighborhoods in both cities Toronto and

New York. Having signed up for a Foursquare developer, using the Client ID and Client Secret, it is possible to make API requests in order in order to retrieve venue information.

By calling the Explore function for both locations in Foursquare API I got different venues that were in the neighborhoods of Toronto and New York. Combining the two neighborhoods in single data frame I got to see that there are 236 common places in both cities and now on bases of these I can compare both cities.

2.4 Dimensionality Reduction:

2.4.1 Missing value ratio:

For selecting the right features first I have done a missing value Ratio test for seeing that if there is any missing value in data frame I should drop that corresponding feature . I see the results that there were no missing values so moved to next test.

2.4.2 Low Variance Filter

The variance is a statistical measure of the amount of variation in each variable. If the variance is too low, it means that it does not change much and hence it can be ignored. As we want to lower the dimensions we need to find if any variable is of low variance than others for this purpose I used a Low Variance Filter test and found out that no feature was to be ignore .

2.4.3 High Correlation Filter

The last test is to find if any multi collinearity exist between the features for this I have used a High Correlation Filter. If there is a very high correlation between two input variables, we can safely drop one of them.

2.4.4 Principal Component Analysis

As we have 236 features in the Merged dataset which can be a large number for cluster analysis I performed a Statistical method called Principal Component Analysis to see that if we can reduce the number of features and yet not affecting the overall Variance of dataset and I was able to shrink down number of features to 170 . This will improve the speed of our Machine Learning clustering algorithim.

Fig 2.1(a) PCA Before dimensionality reduction

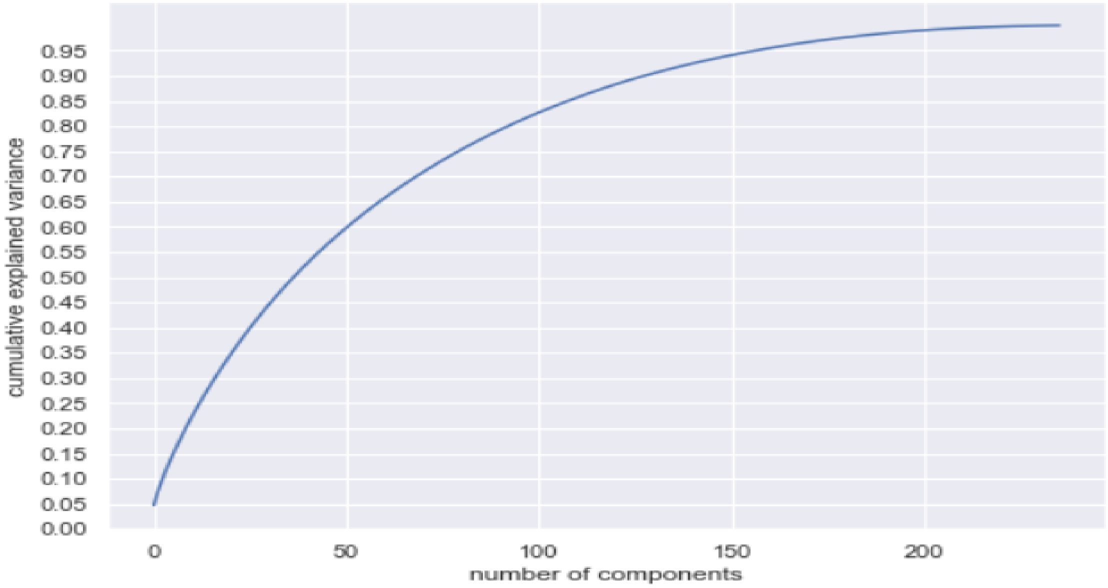
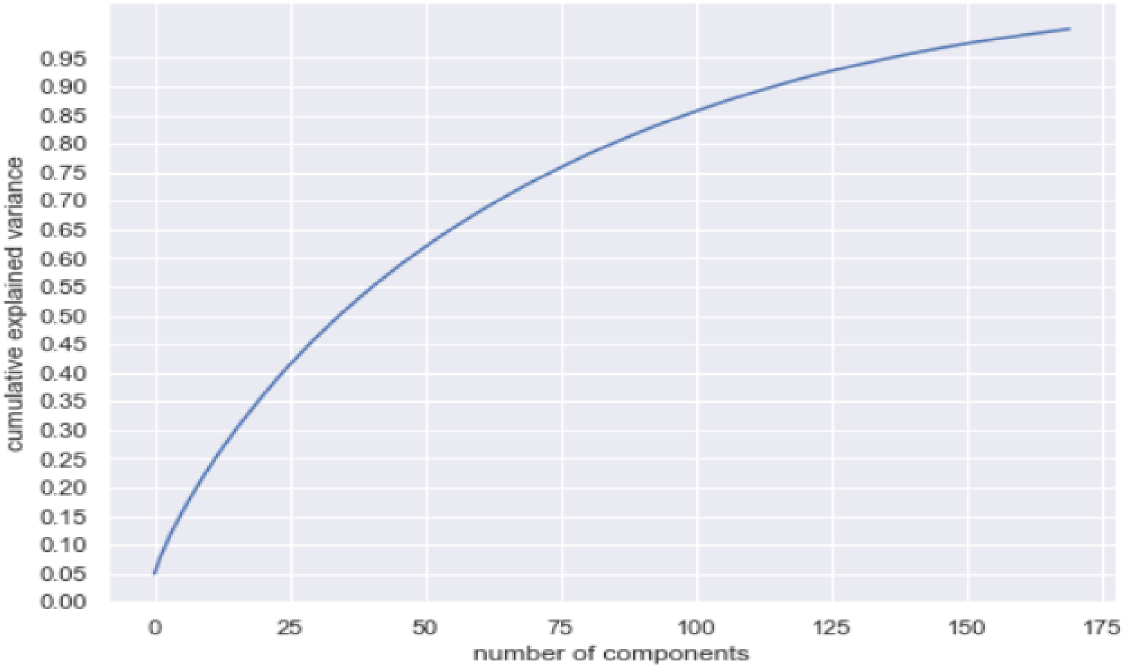


Fig 2.1(b) PCA After dimensionality reduction



3 Methodology

As we are dealing with a dataset that is not labeled a therefore I am going to use a Machine Learning technique called Clustering as this is a type of Unsupervised Learning , more specifically I am going to use K-Mean clustering .

3.1 Finding Optimal Number of Clusters:

While dealing with cluster analysis the biggest problem one has to face is to find optimal number of clusters , As giving value of K more will loose the meaning of clusters and similarly if we give smaller clusters we will also loose the meaning behind clustering .there are multiple methods to find optimal number of clusters and the ones that I followed are as follows

3.1.1 The Elbow Method:

In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use.

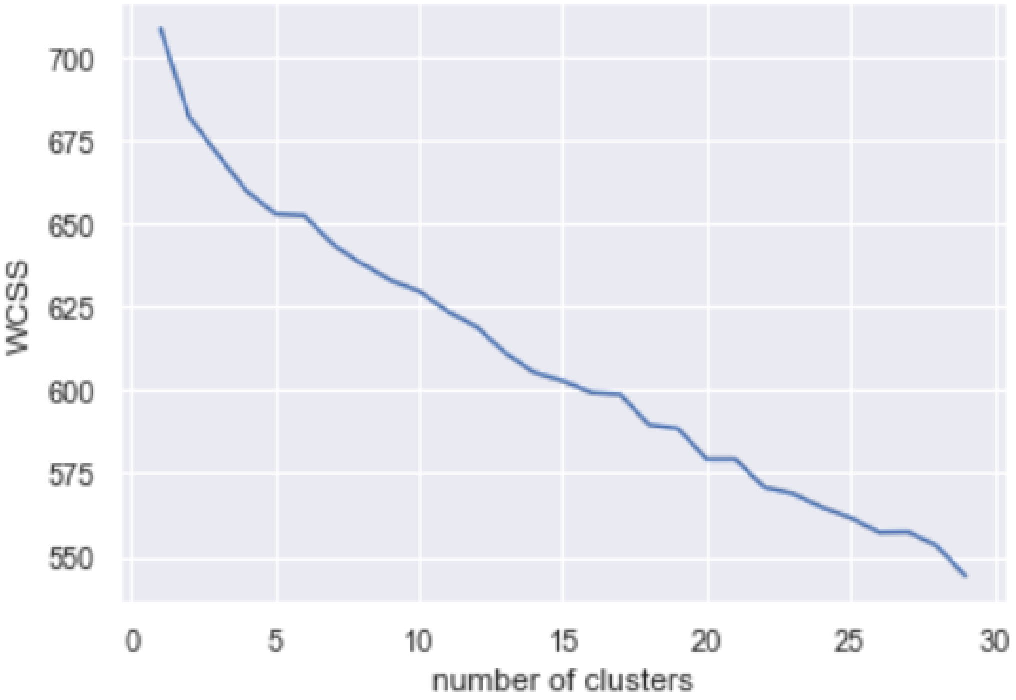
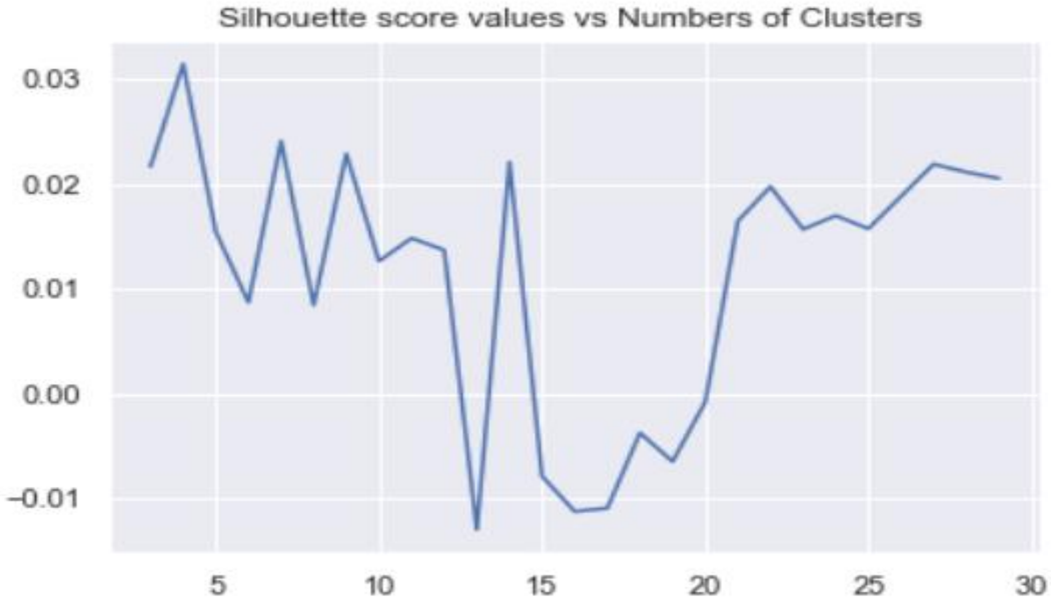
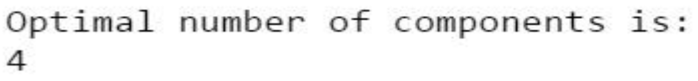


Fig 3.1 Elbow at number 5

3.1.2 The Silhouette Method:

Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1].

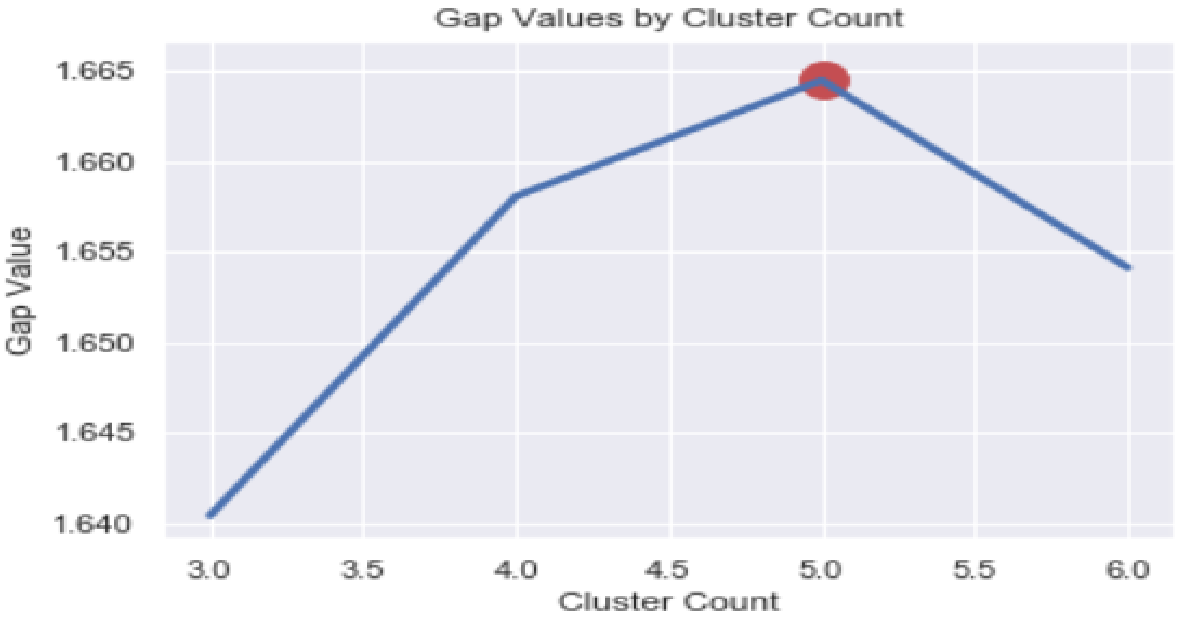




```
Optimal number of components is:
4
```

3.1.2 The Gap Statistic Method:

The gap statistic is a method for approximating the “correct” number of clusters, k, for an unsupervised clustering. We do this by assessing a metric of error (the within cluster sum of squares) with regard to our choice of k.



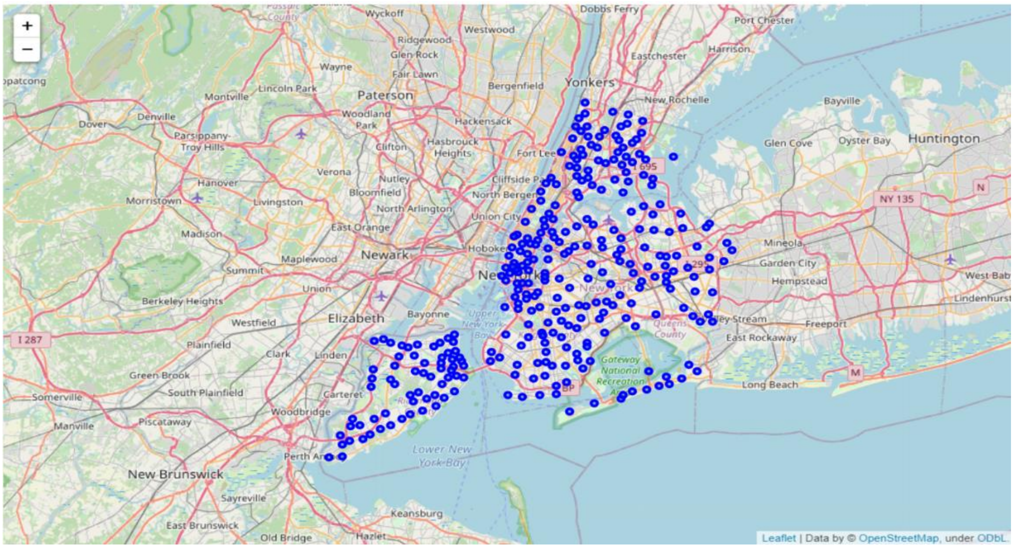
3.2 K-Mean Clustering:

After finding the optimal number of clusters that is 5 I initialized the k-mean algorithm to find the cluster of neighborhoods with similar features as k-mean find clusters up to n-1 clusters where n being the maximum number of features we had to find the optimal k after that it was an easy task of finding the sum of square distance between each point and cluster similar values together .

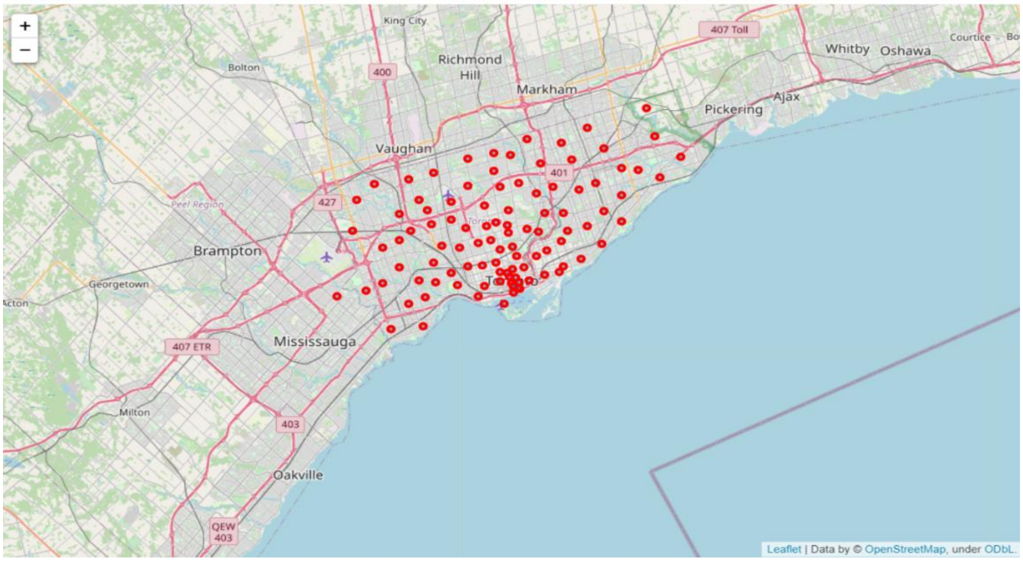
After the algorithm completes merge the clusters with original cities data and display it on map.

4 Visualization of Clusters on Map:

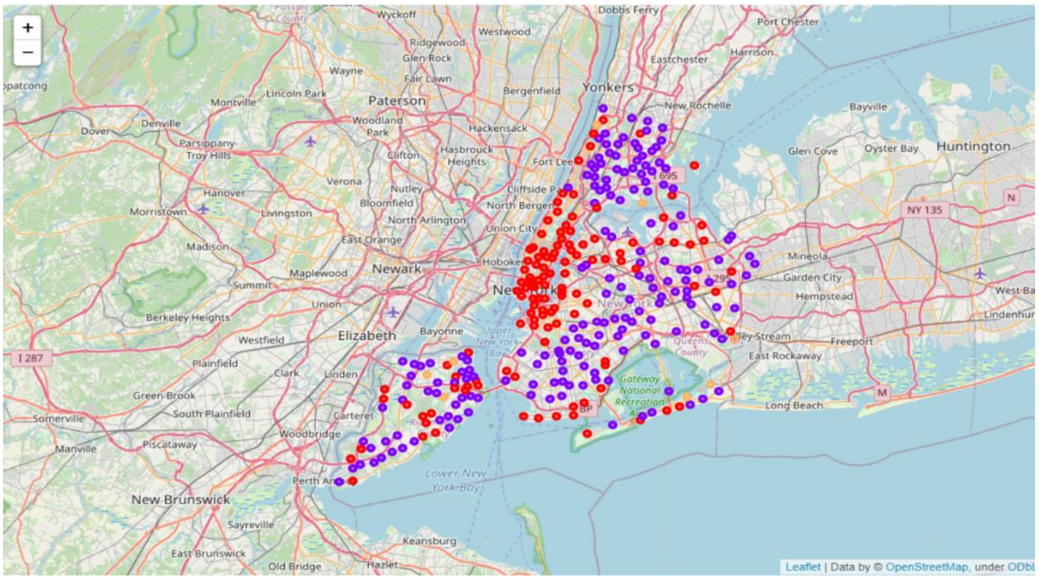
For visualizing clusters on map I used a very strong visualization library called Folium which can display any location on map of earth given its attributes like geographical location latitude and longitude .



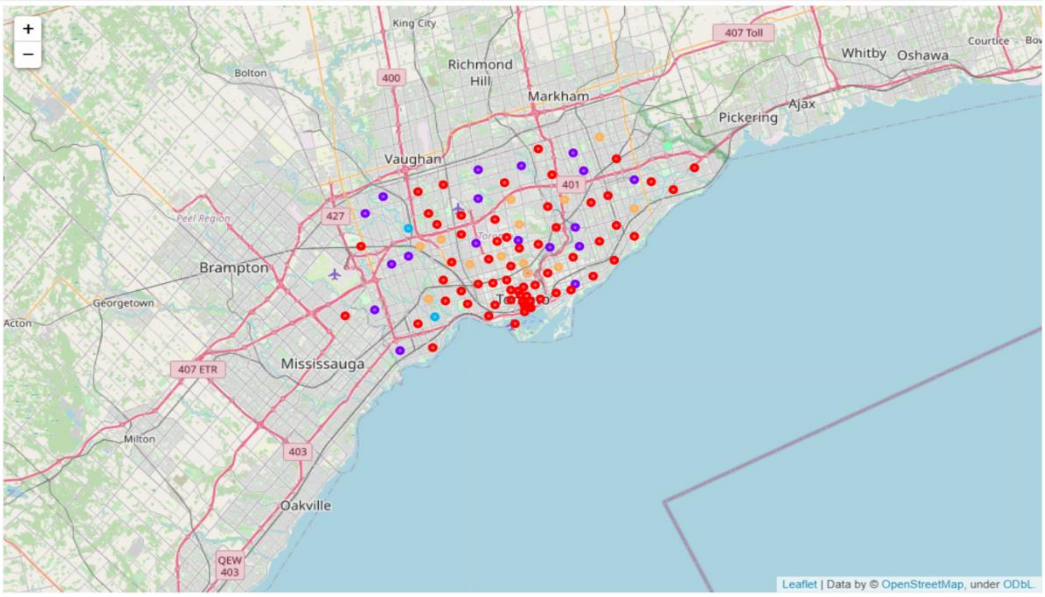
Map of New York With Extracted neighborhoods without Clustering.



Map of Toronto With Extracted neighborhoods without Clustering.



Map of New York With Extracted neighborhoods and Clusters.



Map of Toronto With Extracted neighborhoods with Clusters.

5 Conclusion:

After the completion of the whole process and presenting clusters on the map I can safely give the opinion and suggestion to the person moving from New York to Toronto that he can move to the neighborhood of his own choice with the surroundings that he like and he will somewhat feel like home even if he is in another city .

This is just a single example of what we can achieve from Data Science there can be other problems which we can solve using this methodology .