Similarities Between Neighborhoods in New York and Toronto

Nicolas Renaud

April 6, 2021

1. Introduction

The aim of this project is to compare the cities of New York, NY, USA and Toronto, ON, Canada. Specifically by comparing neighborhoods between two cities to see which are most similar to each other.

This can be done by clustering the neighborhoods in the respective cities, New York and Toronto, showing which neighborhoods are most alike and which are most dissimilar. Such analysis could be used by businesses, such as restaurants, that see success in one city and are looking to expand to the other city. For example, if an Italian restaurant in Toronto was looking to open a new location in New York the knowledge of which neighborhoods in New York are most similar to those in Toronto would improve the chances of the new location succeeding.

By implementing machine learning methods we will measure the similarities between neighborhoods. By doing so the most similar neighborhoods in both cities will be grouped and clearly displayed.

1 Data

1.1 Data Sources

In order to solve the problem previously put forward, we will need the following for New York and Toronto:

- Geographic data of the neighborhoods
- The number of different types of venues in each neighborhood

The sources of data to be used are the following:

- https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M (A table of postal codes and neighborhoods in Toronto)
- Geospatial_Coordinates.csv (A csv of the latitude and longitude of postal codes in Toronto)
- newyork_data.json (A json file of New York neighborhood latitude and longitude data)
- Foursquare API (To obtain the number of venues in each neighborhood)

1.2 Data Cleaning

Before any analysis can be done the Foursquare API was used to gather the data on the number of each type of venue in each neighborhood for both cities. Then the venue data had to be combined with the neighborhood data. Next the data was standardized using the sklearn standard scaler. Finally, a few rows from each city had to be dropped because they had no data on venues.

2 Methodology

The goal of this project is to identify which New York neighborhoods are the most similar to Toronto neighborhoods. By doing this a successful business in one city

looking to expand to the other can choose similar neighborhoods to the one it has already proven successful in.

The first step, shown in the data section, was to gather data on the neighborhoods of the two cities such as the neighborhoods and what businesses and attractions are in each neighborhood.

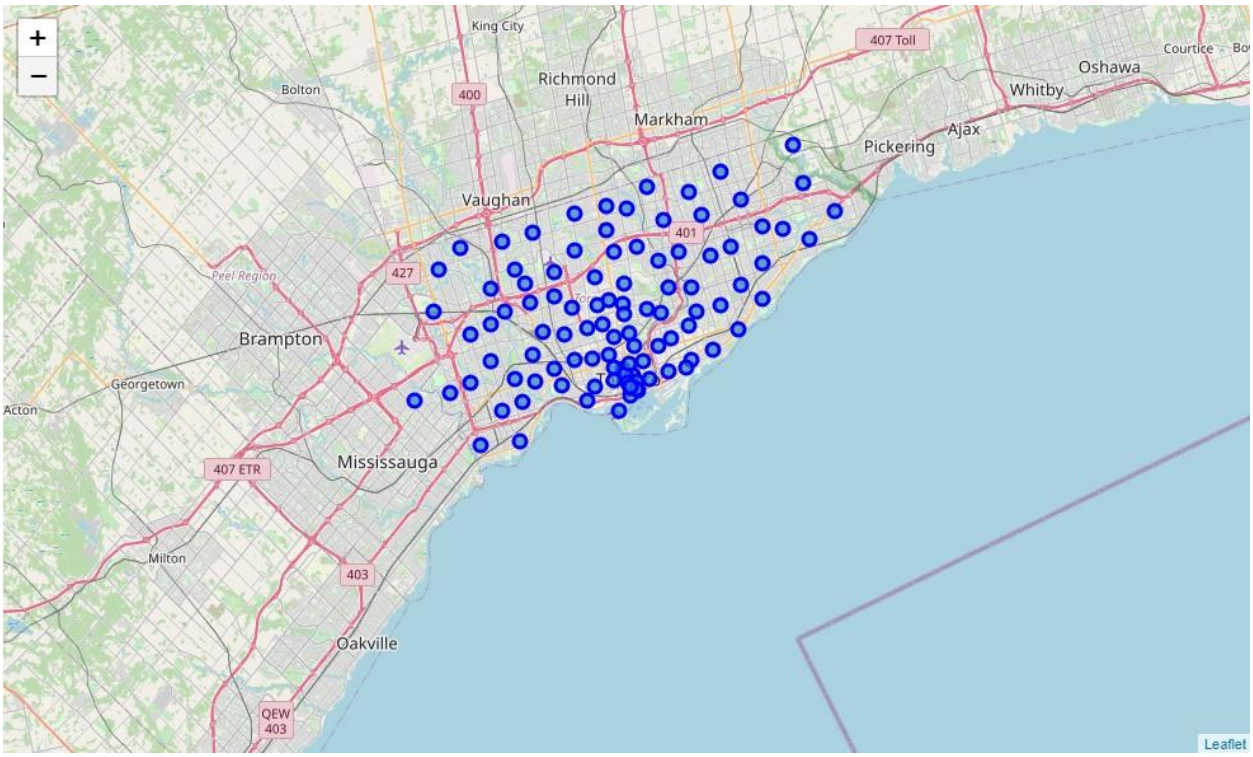

Figure 1. Map of Toronto Neighborhoods

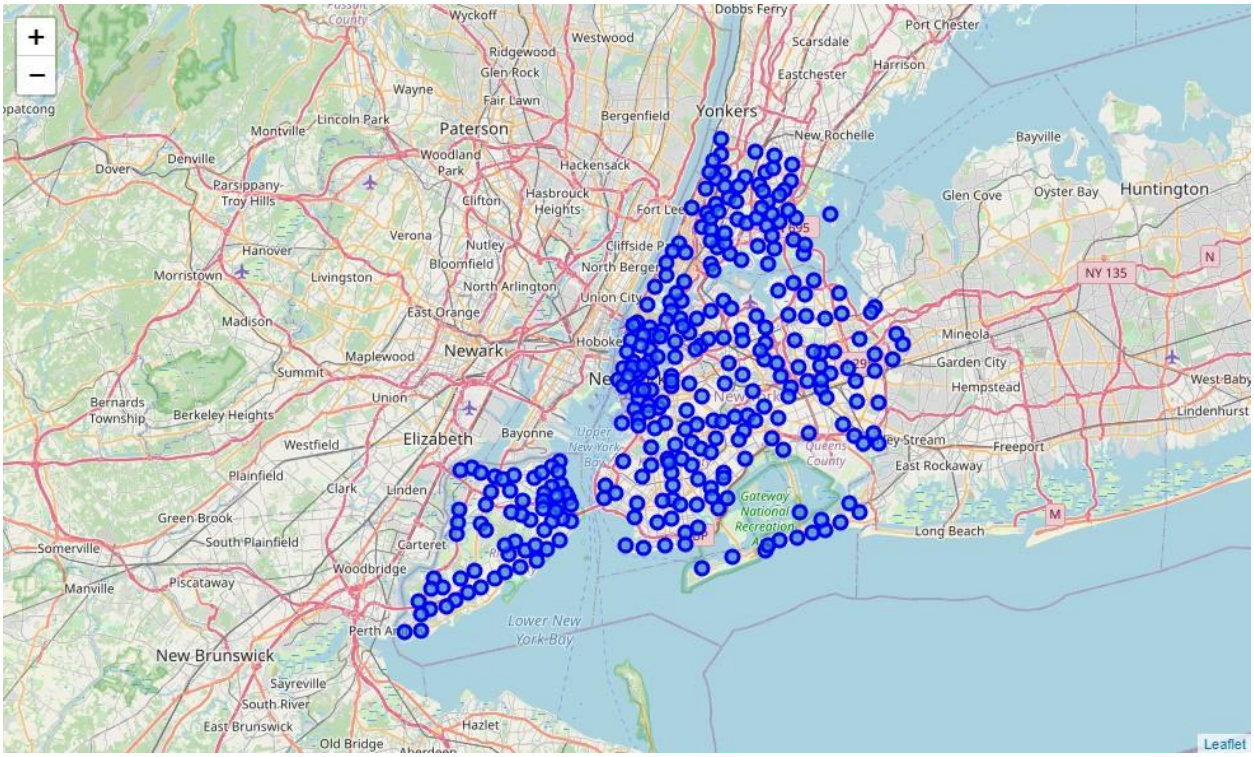

Figure 2. Map of New York Neighborhoods

The second step is to conduct a visual analysis using techniques like correlation and heatmaps to gain an intuitive idea of which neighborhoods are likely candidates for a business expansion.

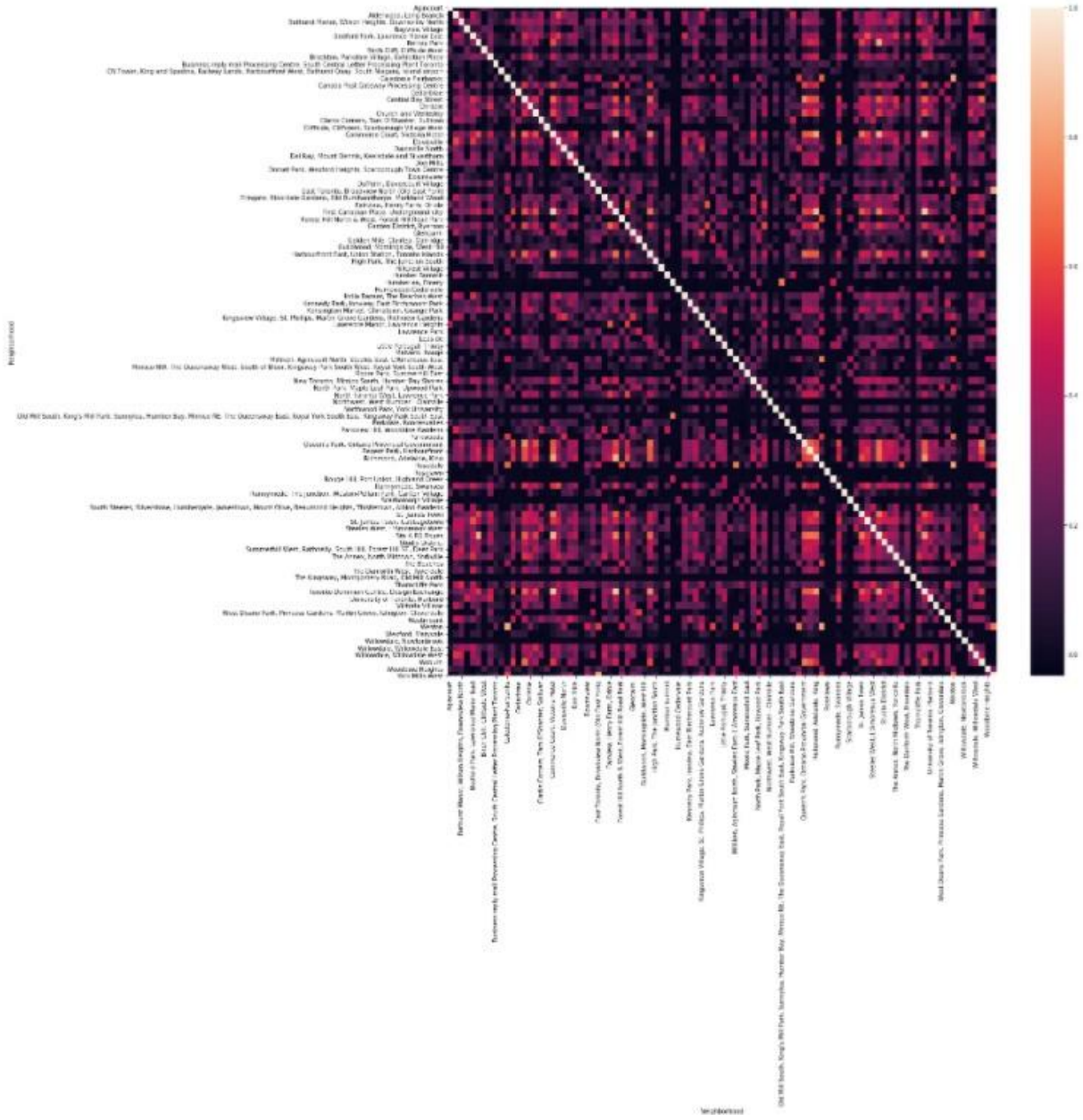

Figure 3. Toronto Neighborhood Heatmap

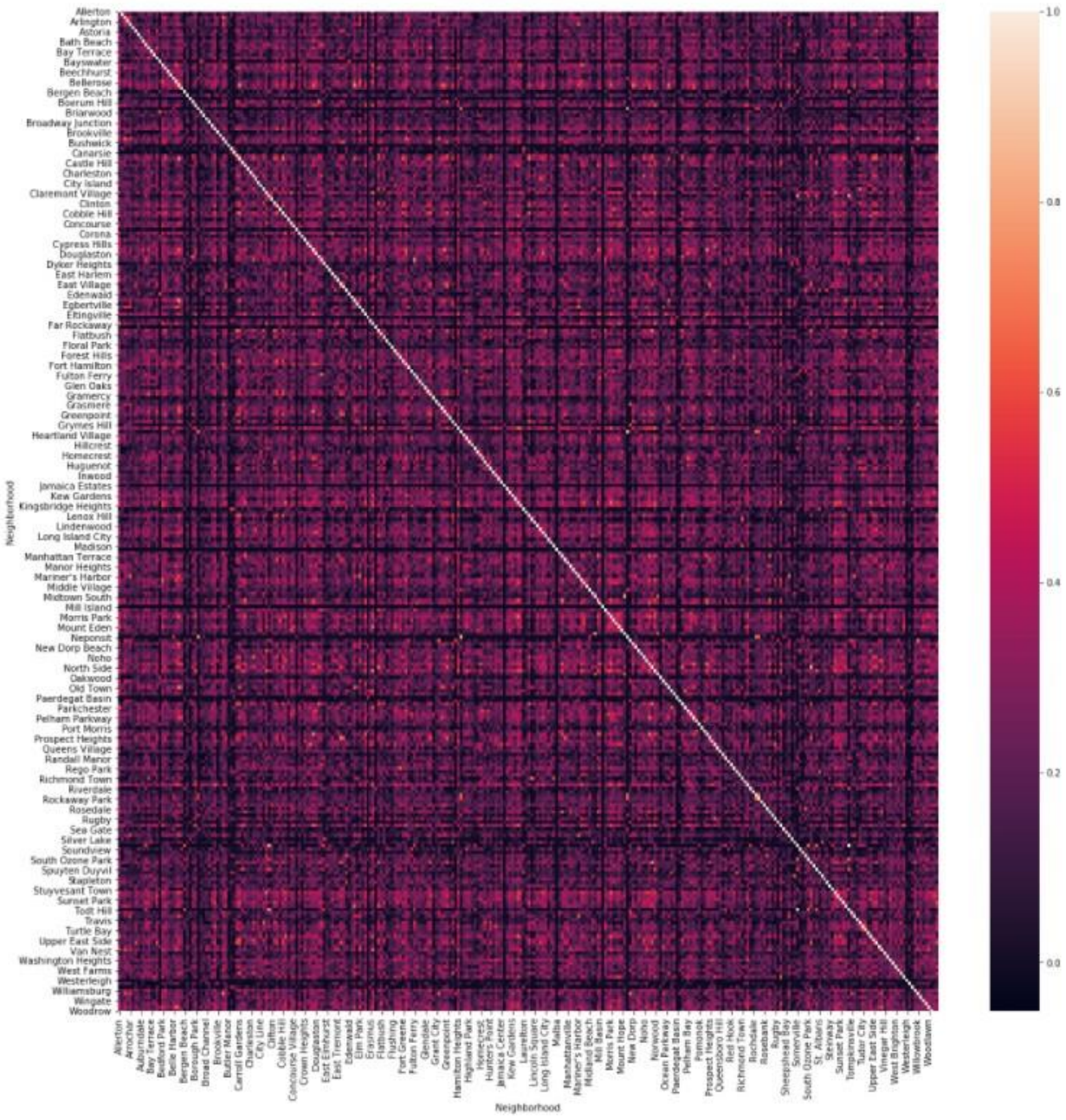

Figure 4. New York Neighborhood Heatmap

The heatmap visualization is to try and see correlation between neighborhoods within New York and Toronto neighborhoods. There do appear to be some neighborhoods with a degree of correlation; however, there are so many neighborhoods that it is hard to tell and neither city had many neighborhoods with a correlation over

0.5. In fact most of the neighborhoods had very similar levels of correlation so it is clear that further analysis is needed.

The third and final step is to use clustering techniques, specifically hierarchical clustering, in order to group the most similar neighborhoods within and in between the two cities. Using this information will present the clearest picture of which neighborhoods are the most similar and therefore which neighborhoods a business should consider when looking to expand to the new city.

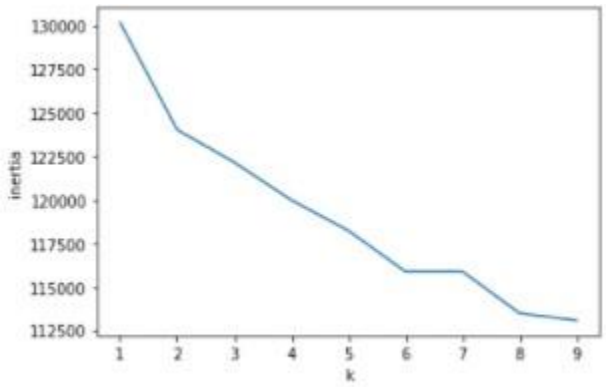

Figure 5. Result of Different ks for KMeans

Before implementing the final KMeans model we looked at different ks from 1 to 9 in order to see which would be best used in the final model. A k of 6 was chosen as there is minimal gain to be had from a larger k value and that may cause overfitting. The final model was first fit on the New York data and then the Toronto data.

3 Analysis

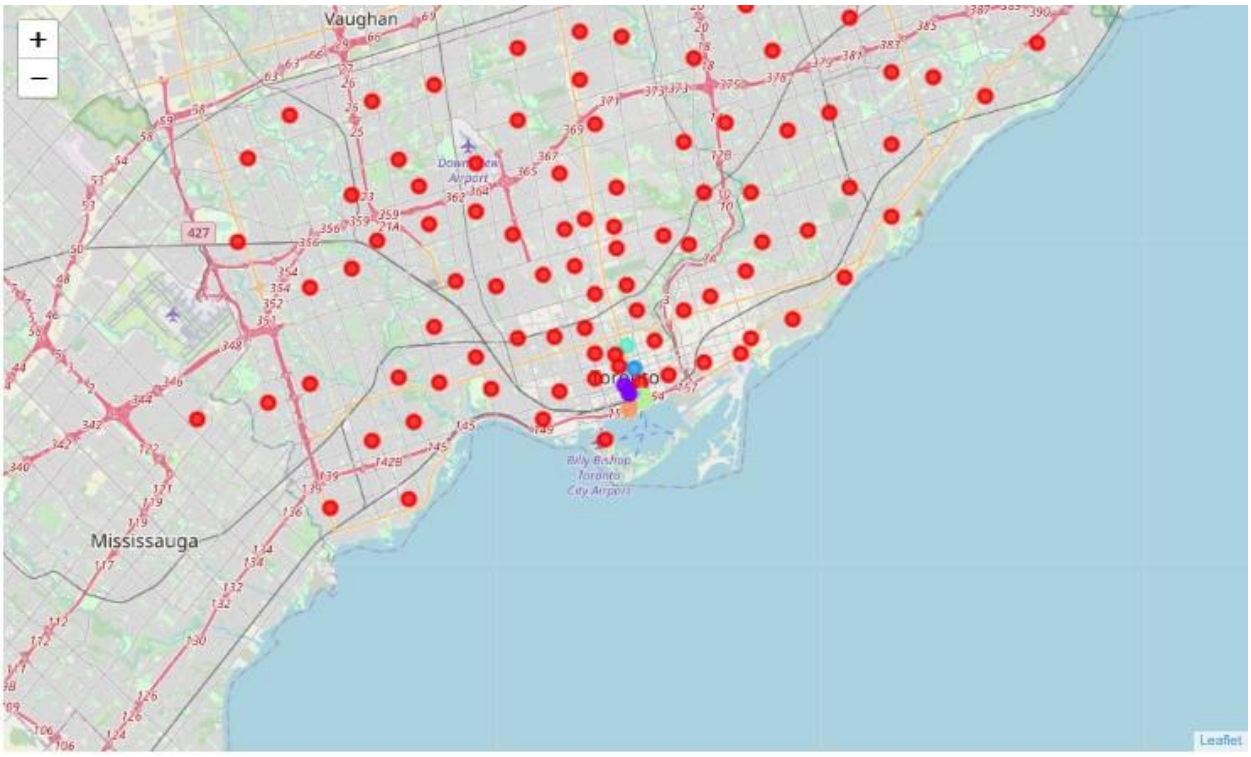

Figure 6. Toronto Neighborhoods by Cluster

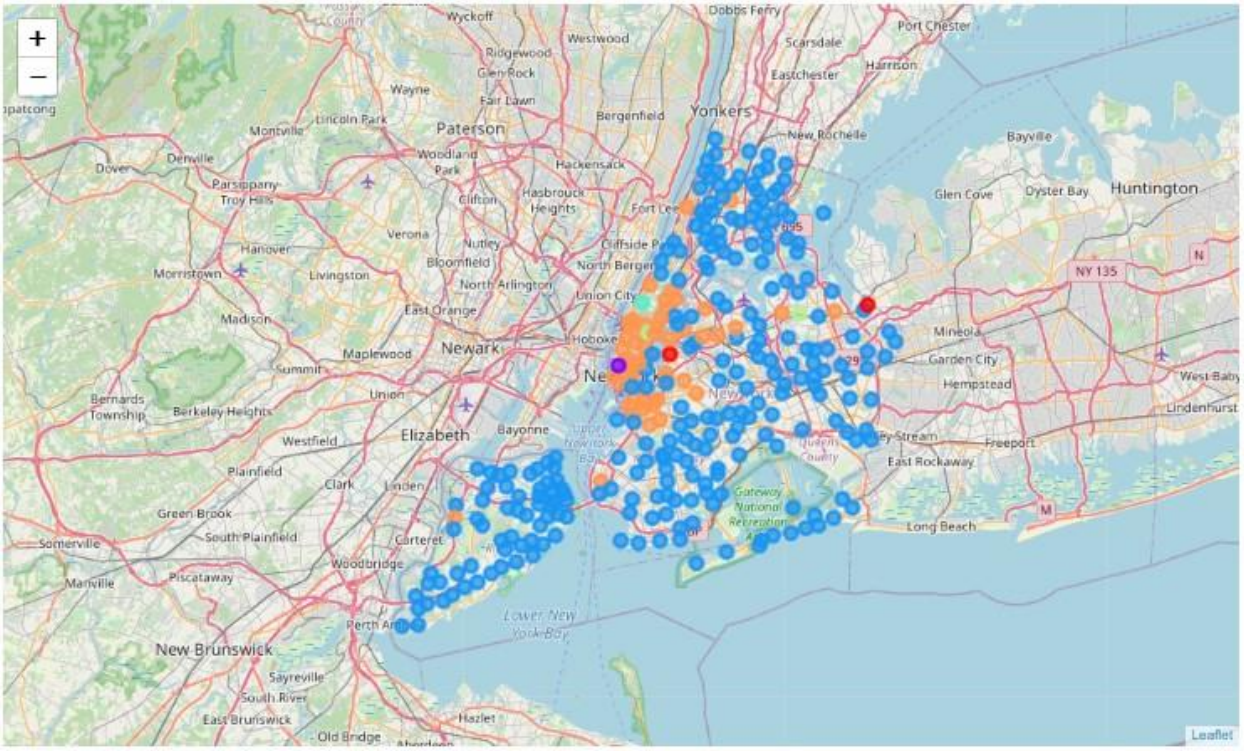

Figure 7. New York Neighborhoods by Cluster

The above analysis shows that for the most part neighborhoods in New York are not very similar to neighborhoods in Toronto; however, there are some neighborhoods that at least according to our cluster analysis share similarities to a degree. In particular most Toronto neighborhoods belong to cluster 0 (shown in red); whereas, most neighborhoods in New York belong to cluster 2 (shown in blue). Despite New York and Toronto neighborhoods overall not being similar there are specific areas that are closer related.

The largest cluster in Toronto also contains Greenpoint and Little Neck from New York, the largest New York cluster contains Commerce Court, Victoria Hotel, First Canadian Place, Underground city, Richmond, Adelaide, King, Toronto Dominion Centre, and Design Exchange from Toronto. This shows that our analysis can be used situationally for a business looking to move from New York to Toronto and can allow them to rule out most neighborhoods for expansion.

Additionally, this clustering analysis finds that the outlying areas of both Toronto and New York belong to the largest clusters and are most similar. New York in particular is interesting as Manhattan appears to have its own cluster as well (shown in orange).

4 Conclusions

This project is a good starting point to advising businesses looking to move from New York to Toronto, or vice versa, which neighborhoods to consider. However, more questions and further analysis should be done. Additional variables may prove helpful for analysis such as crime rates, household income, property value, average education level, and access to public transportation. Other clustering methods may prove to be

more useful like hierarchical clustering or density based clustering. Finally, additionally cities should be considered in the analysis and it would be interesting to see if other cities are more similar and follow the trend of outlying areas being predominantly of one cluster.