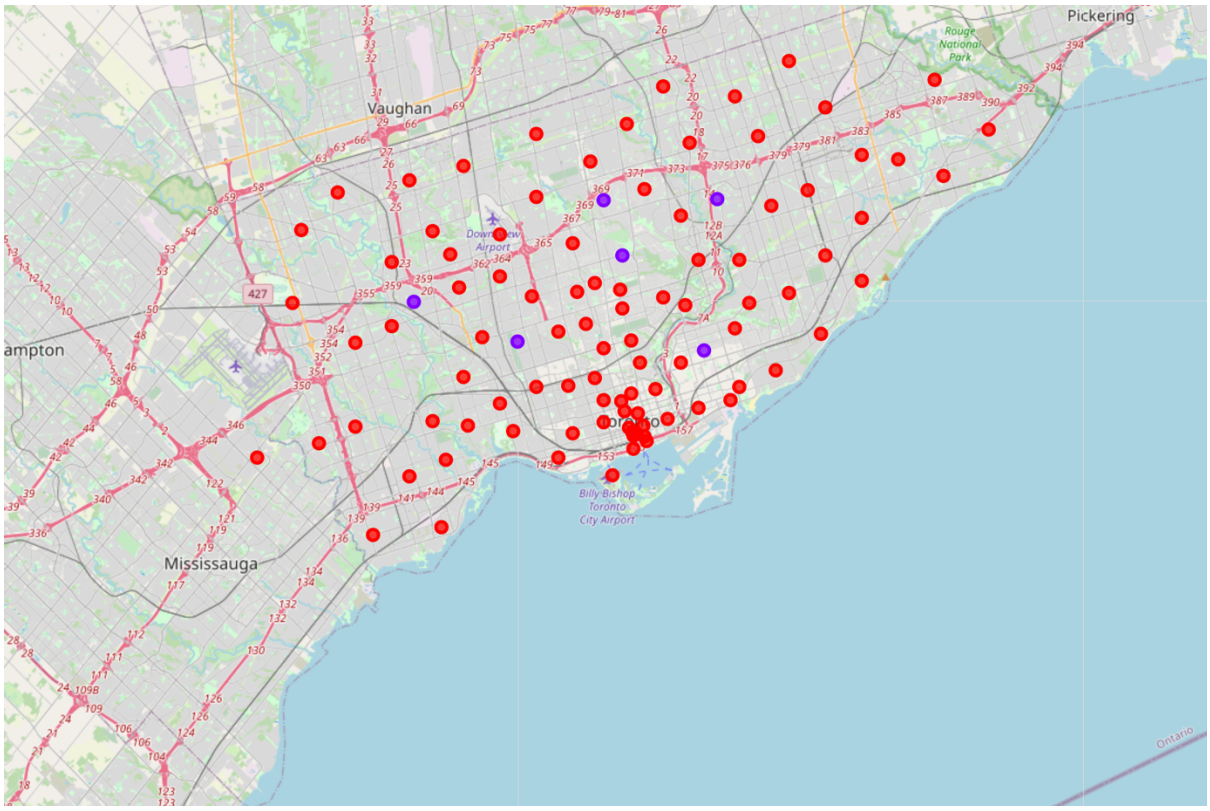
Capstone Project - The Battle of Neighborhoods (Week 2)

New York vs Toronto

Table of Contents



I. Introduction

The problem that I am going to tackle is to compare the neighborhoods of Toronto and New York (restricted to Manhattan) and try to identify clusters of similar ones across both cities.

The idea behind that is for someone moving from one city to the other to be able to find a similar neighborhood to live in.

The audience is therefore someone moving from New York (Manhattan) to Toronto or from Toronto to New York (Manhattan) and looking for a similar neighborhood to live in.

II. Data

The data that will be used for this analysis is the same one used for the New York and Toronto projects, but the idea is to analyze them together.

In detail, the data used will be:

- A dataset of NY that contains the 5 boroughs and the neighborhoods that exist in each borough (306 in total) as well as the latitude and longitude coordinates of each neighborhood.
  This dataset can be found at the following address: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json

- A similar dataset will be constructed for Toronto using the data from the Wikipedia page https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M and a csv file available here http://cocl.us/Geospatial_data for the latitudes and longitudes

- With those datasets, the top 100 venues for each neighborhood will be retrieved using the Foursquare API and the clustering analysis will be done using the type of venues present in each neighborhood and their frequency

III. Methodology

First, I will gather all the data necessary using the NYc dataset that we previously used and the data that we retrieved from Wikipedia for Toronto.
Then, using this data, I will retrieve all the venues in a radius of 500m of the location of the neighborhood using the Foursquare API.

This result data will then be encoded to be used in a k-means clustering technique in order to put together the similar neighborhoods.

The k-means clustering technique will be run 13 with k values from 2 to 14 to determine which value is the best using the Silhouette Score.

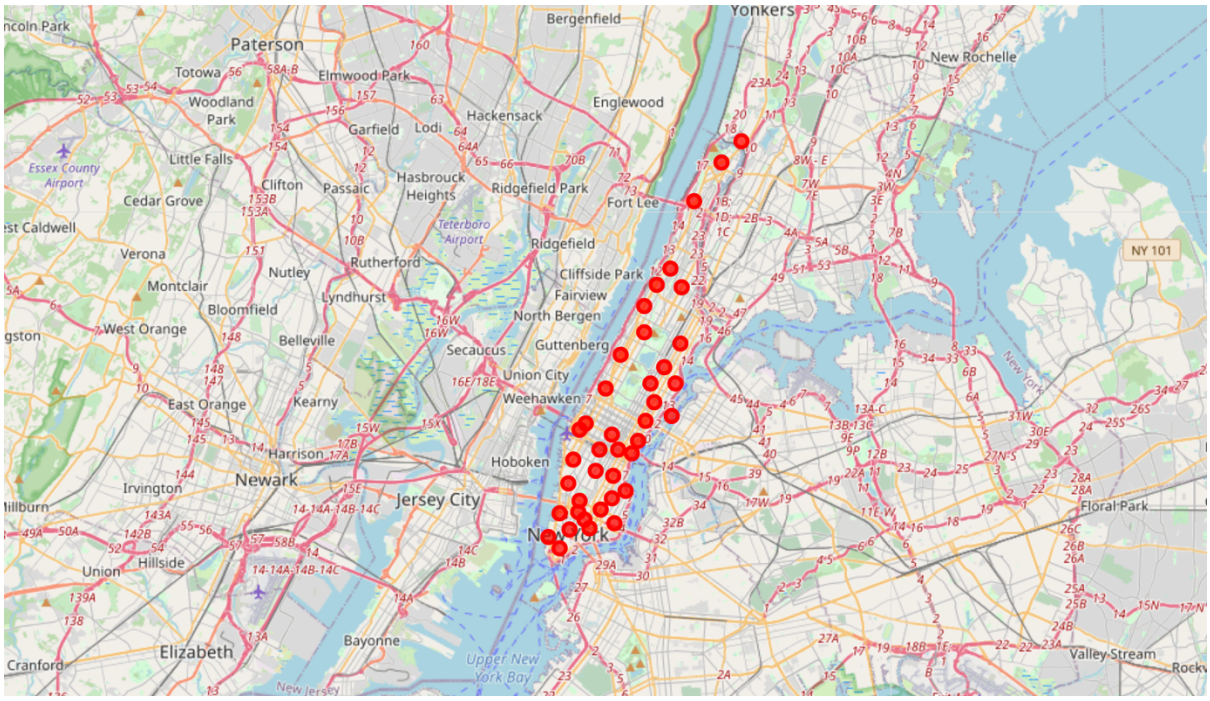
The results of the k-means clustering technique will be displayed using folium maps of Toronto and NYc.

Finally, unlike what was initially planned, this operation will be run 3 times in order to not just compare Manhattan with Toronto center but also all 5 boroughs of NYc with Toronto center and all 5 boroughs of NYc with all of Toronto.

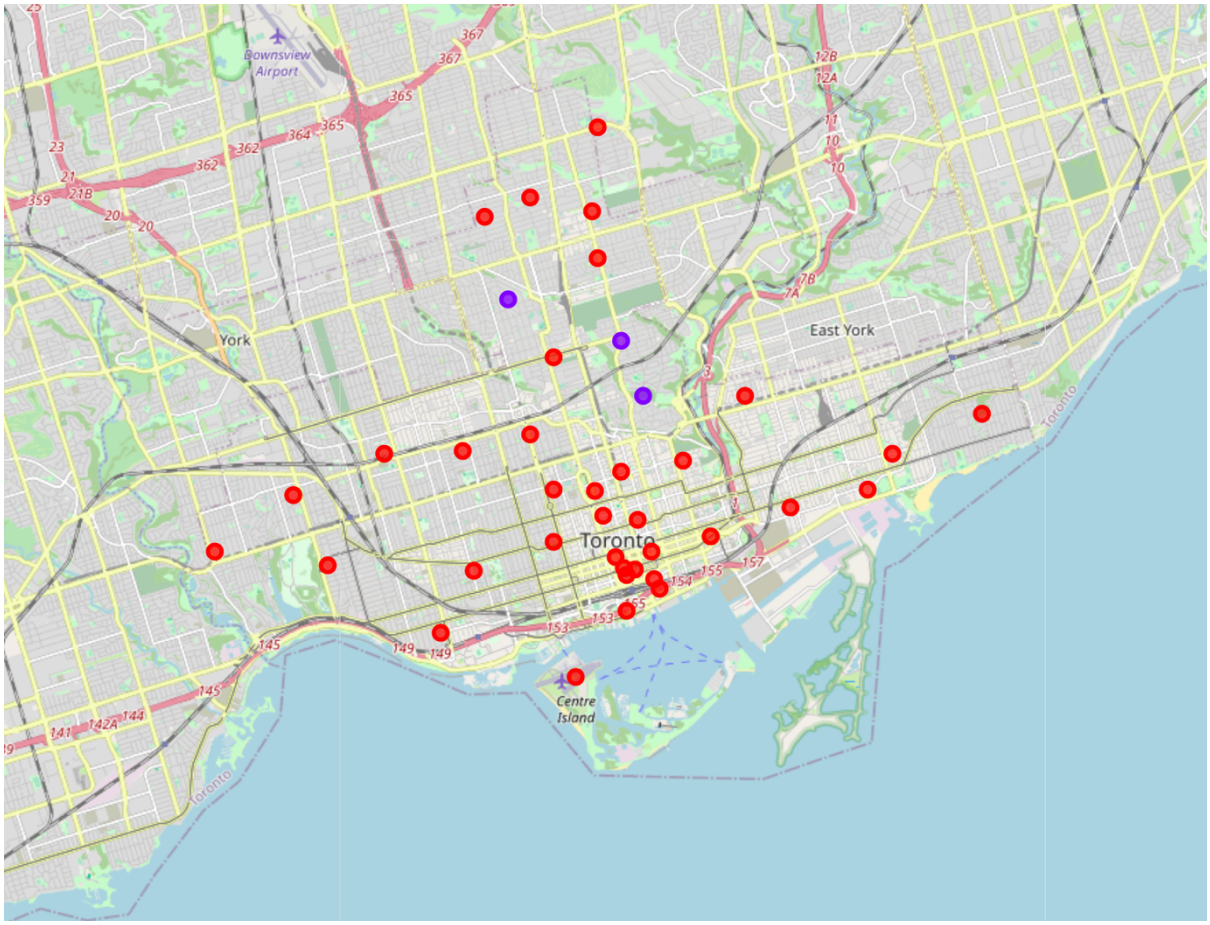
IV. Results

Let's get straight to the point, the results are not very good or at least not very useful for the goal that was set to enable someone moving from Toronto to NY or from NY to Toronto to find a similar neighborhood in the other city.
Indeed, the analysis of Manhattan against the neighborhoods of Toronto center gave only 1 cluster for NY and 2 for Toronto with only 3 neighborhoods in the second cluster for Toronto.

Below are the result maps (the first one for Manhattan and the second one for Toronto)



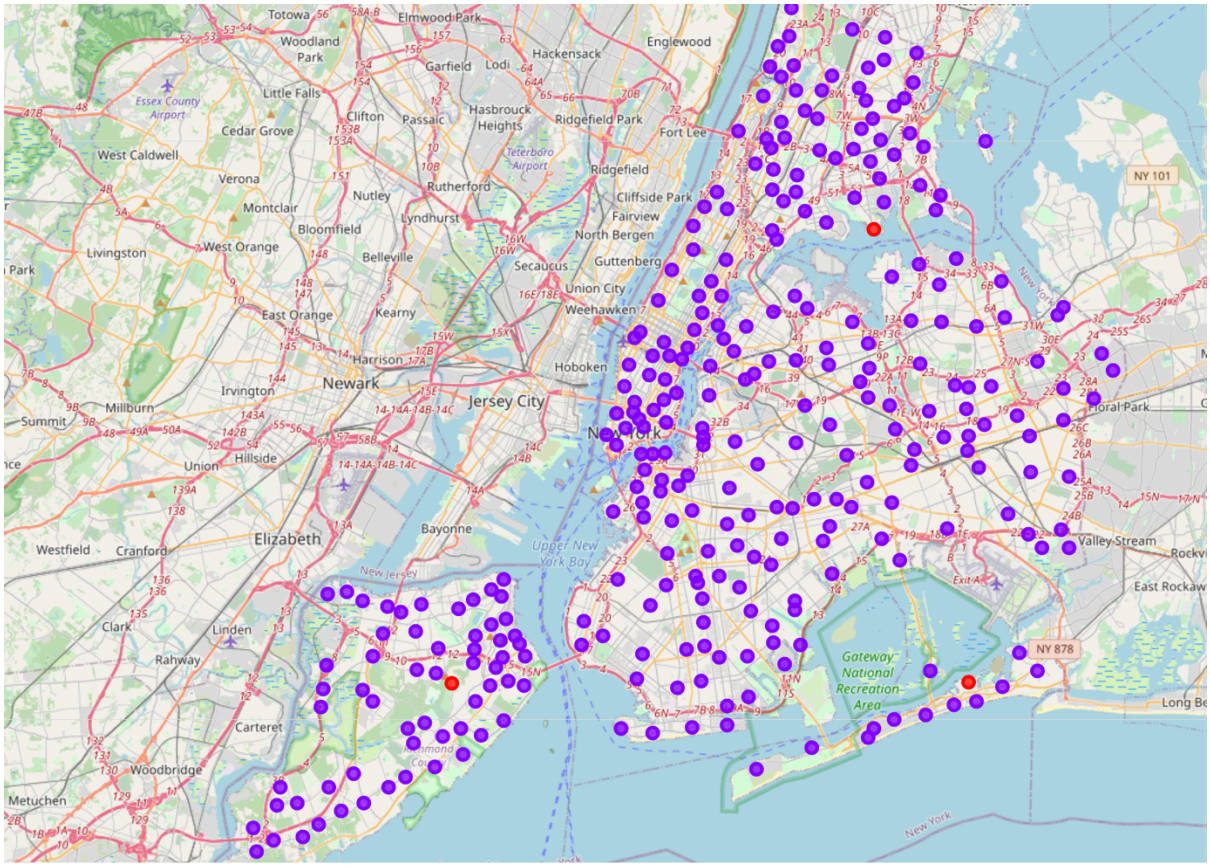
Map of Manhattan with the neighborhoods (cluster 1 in red)



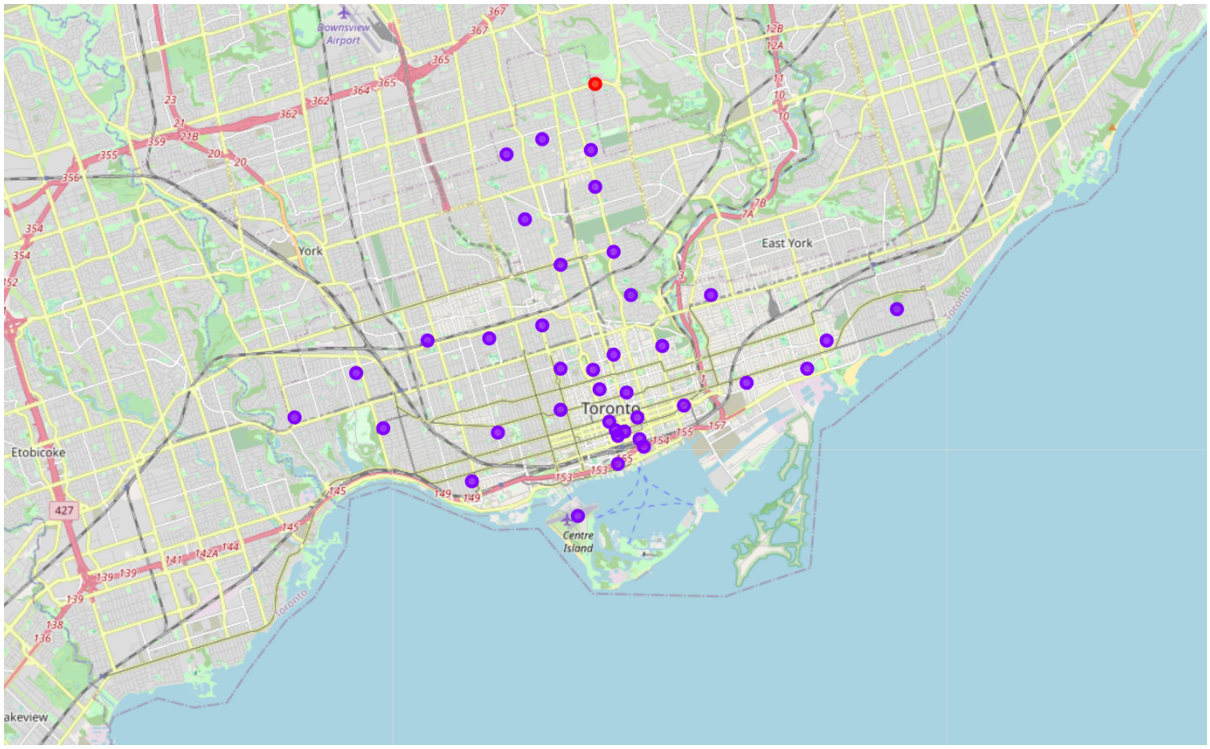
Map of Toronto center with the neighborhoods (cluster 1 in red, cluster 2 in purple)

The analysis was therefore pushed to find more diversity by including all the boroughs of NYc with Toronto center. The results are a little different with NYc showing 3 neighborhoods (Somerville, Todt Hill & Clason Point) in the second cluster (Toronto has only Lawrence Park in this cluster) but in the end, nothing really significant.

Below are the result maps (the first one for NYc and the second one for Toronto)



Map of NYc with the neighborhoods (cluster 1 in purple, cluster 2 in red)



Map of Toronto center with the neighborhoods (cluster 1 in purple, cluster 2 in red)

Finally, the analysis was against pushed to the whole city of Toronto against the all 5 boroughs of NYc.
The results are as follow, in the second cluster for Toronto we find:

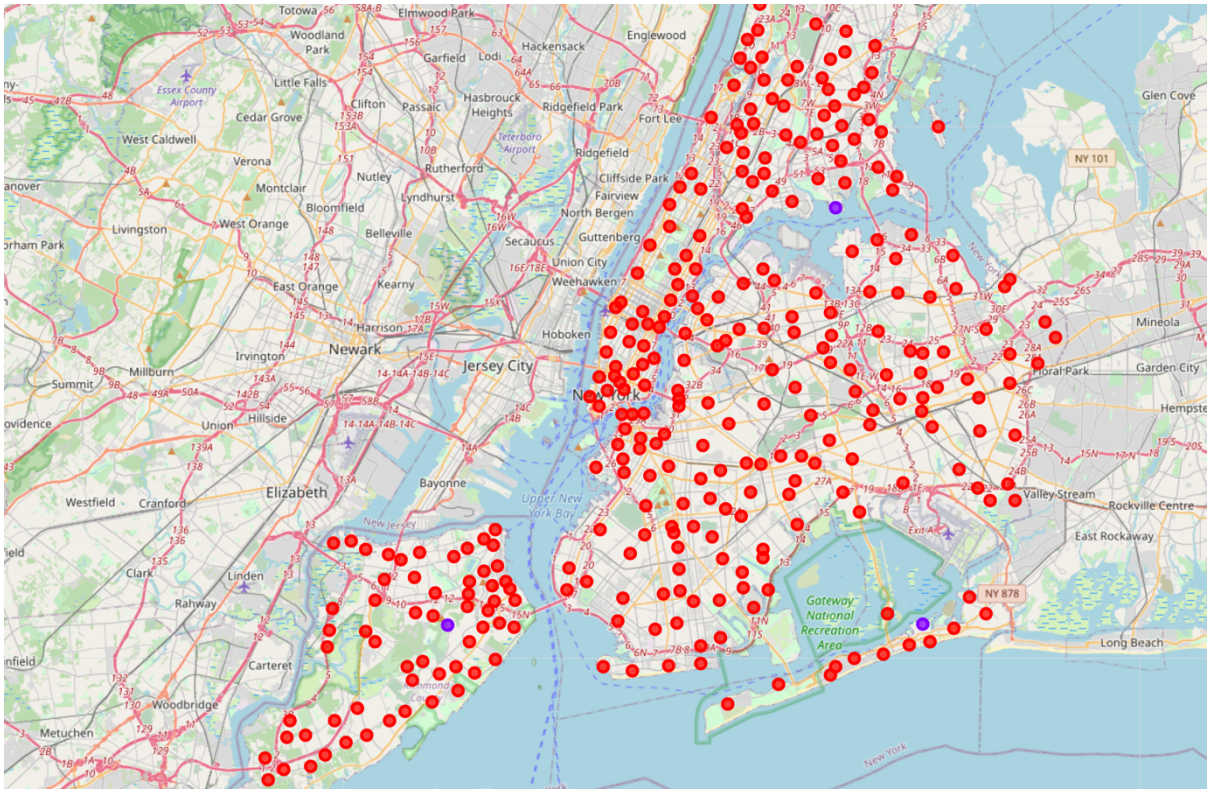
Caledonia-Fairbanks / East Toronto, Broadview North (Old East York) / Weston Cluster / Lawrence Park / Parkwoods / York Mills West

And for NYc we have:

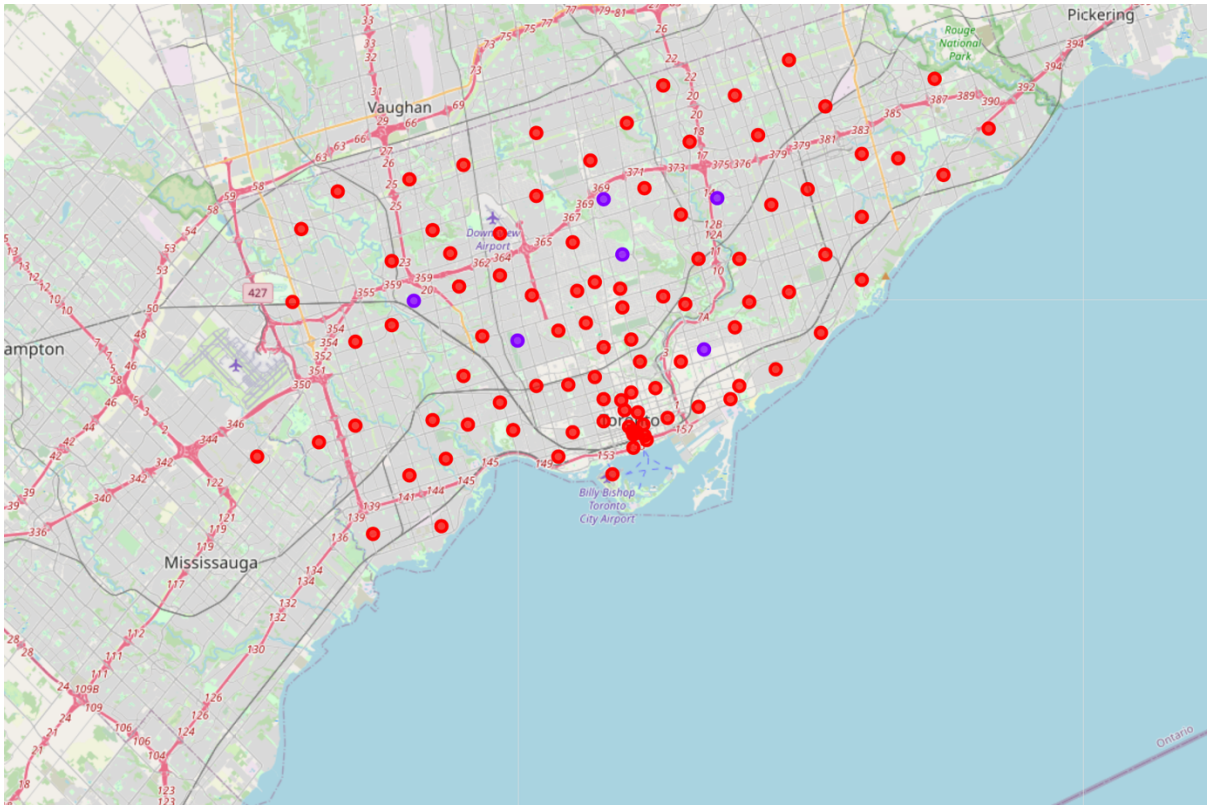
Somerville / Todt Hill / Clason Point

This is again, a little bit better but not really significant.

Below are the result maps (the first one for NYc and the second one for Toronto)



Map of NYc with the neighborhoods (cluster 1 in red, cluster 2 in purple)



Map of Toronto center with the neighborhoods (cluster 1 in red, cluster 2 in purple)

V. Discussion

The results found are not really a surprise, with both cities being in highly developed countries and on the same continent of north America it was expected that they would be very similar.
What is unexpected and a little bit disappointing in regards with our goal is that most neighborhoods are in the same cluster meaning that we cannot really help our public in choosing their new place to stay other than say that most neighborhoods are the same.
As this conclusion is probably false, here are a few ideas to explore in order to follow up on those initials results:

- Use another clustering technique than K-means
- Explore other measures of 'best k' than the Silhouette score that each time gave us 2 clusters. (The analysis of using many more clusters was tried but not included as it seemed to show the same results) (This can be replicated very easily by changing the value of kclusters)
- Redo the analysis by comparing each borough of NYc to Toronto center (the way the notebook was prepared makes this analysis very easy to do)
- Use the analysis on other cities to try to find different results (for example a European or Asian city against NYc)
- Find other markers to define the neighborhoods than the venues

VI. Conclusion

In conclusion, we did not find very good results with our analysis in the way that they are not really helpful for our targeted audience. Indeed, what our results show are that most neighborhoods of NYc and Toronto are very similar by being grouped in the same cluster (aside from a few of them mentioned above).
Nevertheless, many axes were found to try to improve those results such as using other clustering techniques and compare the results or use this analysis on other cities.