New York City and Toronto City Neighborhood and Venues Analysis

Chiara Mattei

July 27, 2021

1 Introduction

1.1 Background

Nowadays, people and workers are moving constantly from one location to another depending on their business, financial and family needs that are continuously changing and evolving.
Due to the Corona pandemic, employees, also supported by companies, are increasingly working from home, within the same country or in a country other than their workplace. This allowed the worker to be given more freedom than in the past, that is to travel to different places and then choose to live and explore new places while remaining in a neighborhood that best suits their interests. This puts the worker in front of a choice on which place, or rather neighborhood of a city, it might be more congenial to move to. Anyone who has moved has asked themselves the following questions: which neighborhood should I move to? What facilities would I like to have near my residence? I am happy in my neighborhood in New York City, what could be the most similar in the city of Toronto?

Figure 1. Landscape of Toronto City

Figure 2. Landscape of New YorkCity

1.2 Problem

This project aims to create clusters of neighborhood of different city based upon most common venues similarity.
To solve the problem I have decided to apply the Data Science methodology to analyze the similarities of neighborhood of two big multinational and multicultural city, New York City and the Toronto City, which are the financial capitals of their respective countries, US and Canada. To solve the problem, we can create a map to compare the different cities with neighborhood superimposed on top and clustered based upon similarities, facilities and venues that each neighborhood offer.

1.3 Interest

Employees, people and families would be very interested in accurate prediction of neighborhood similarities, to avoid incurring additional costs due to a poor choice of destination and having to relocate after a short time.

2 Data

Geographic data containing Postal Code Id, Borough and Neighborhood located within the city of Toronto are taken from two datasets: Neighborhood and Borough name here, while latitude and longitude associated to each postal code gathered from the previously mentioned table here.

	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Queen's Park	Ontario Provincial Government

Table 1. dataset1 of Toronto

	PostalCode	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Table 2. dataset2 of Toronto

Geographic data containing Borough and Neighborhood located within the city of NewYork are found at the link 'https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701ENSkillsNetwork/labs/newyork_data.json'.

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Table 3. dataset of New York

The last dataset that has been used is the Foursquare API to get the most popular venues for the aforementioned cities.

3 Methodology

The first step of the project is to load all the data that has been mentioned in the previous chapter removing all inconsistent values. After I have decided to select and therefore filter NewYork Data taking Borough of Manhattan and for Toronto dataset filtered on Borough containing the string "Tor", with the purpose to focus the analysis on the most central Borough.

I used python folium library to visualize the map of New York and Toronto city with their neighborhood superimposed on top. I used latitude and longitude values to get the visual as below:

Figure 3. Neighborhood of Toronto City

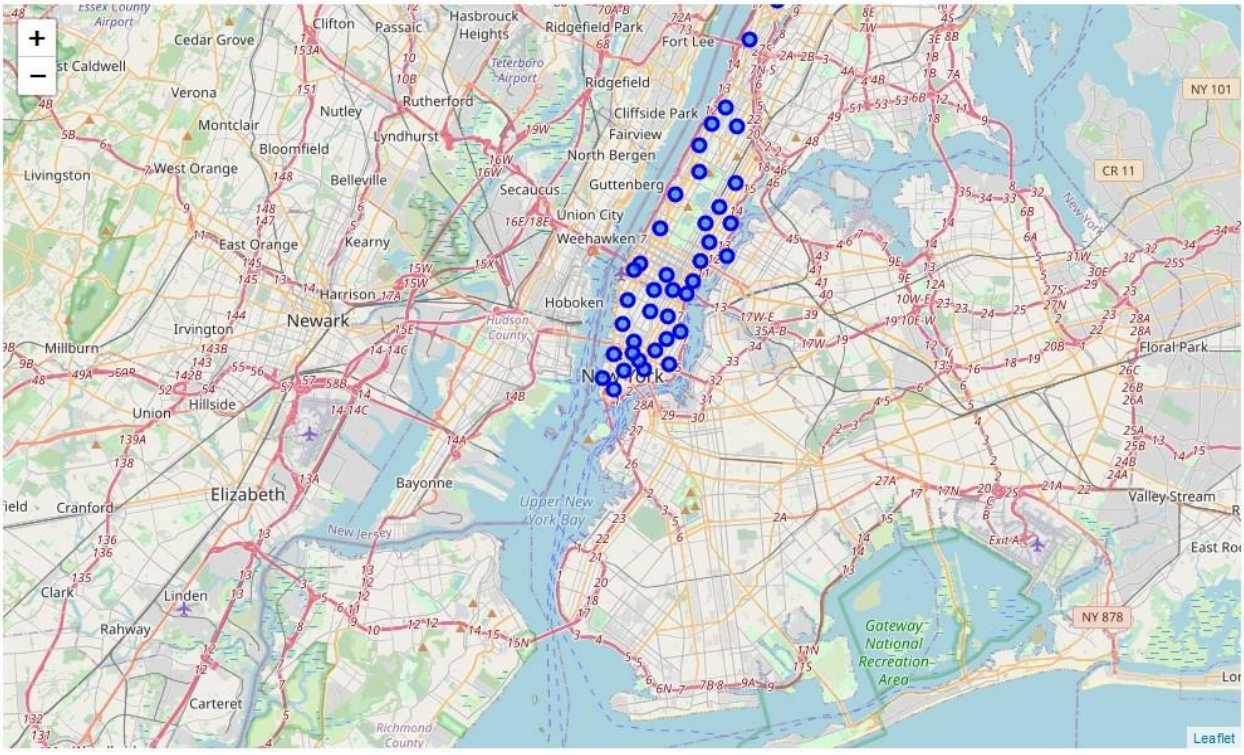

Figure 4. Neighborhood of NewYork City

I utilized the Foursquare API to explore the neighboroughs and segment them. This dataset contains the results of Most 100 common Venues for each latitude and longitude location within a radius of 500 meter. The Feature (*Table 4*) coming from the joint of all the gathered data are: Borogh, Neighborhood, Latitude, Longitude, City, Neighborhood Latitude, Neighborhood Longitude, Venue, Venue Latitude, Venue Longitude and venue Category.

	Borough	Neighborhood	Latitude	Longitude	City	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
2	Downtown Toronto	Regent Park, Harbourfront	43.65426	-79.360636	Toronto	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
2	Downtown Toronto	Regent Park, Harbourfront	43.65426	-79.360636	Toronto	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
2	Downtown Toronto	Regent Park, Harbourfront	43.65426	-79.360636	Toronto	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center
2	Downtown Toronto	Regent Park, Harbourfront	43.65426	-79.360636	Toronto	43.65426	-79.360636	Impact Kitchen	43.656369	-79.356980	Restaurant
2	Downtown Toronto	Regent Park, Harbourfront	43.65426	-79.360636	Toronto	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa

Table 4. Neighborhood of NewYork City

Number of most common venues extracted per Neighborhood within a radius 500 meter can be seen in *Figure 5* and *Figure 6.*

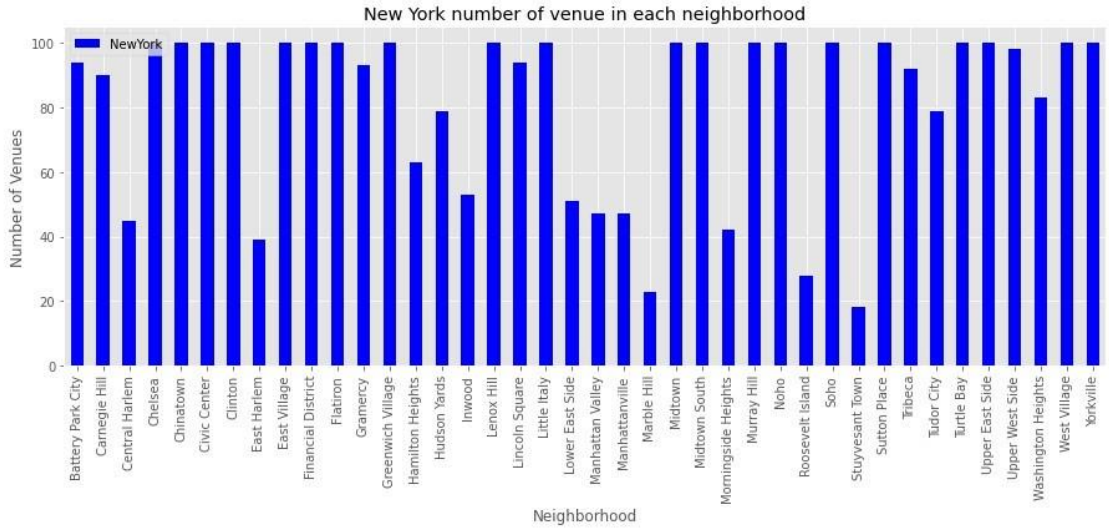

Figure 5. Number of most common Venues extracted from Foursquare in the Neighborhood of NewYork City

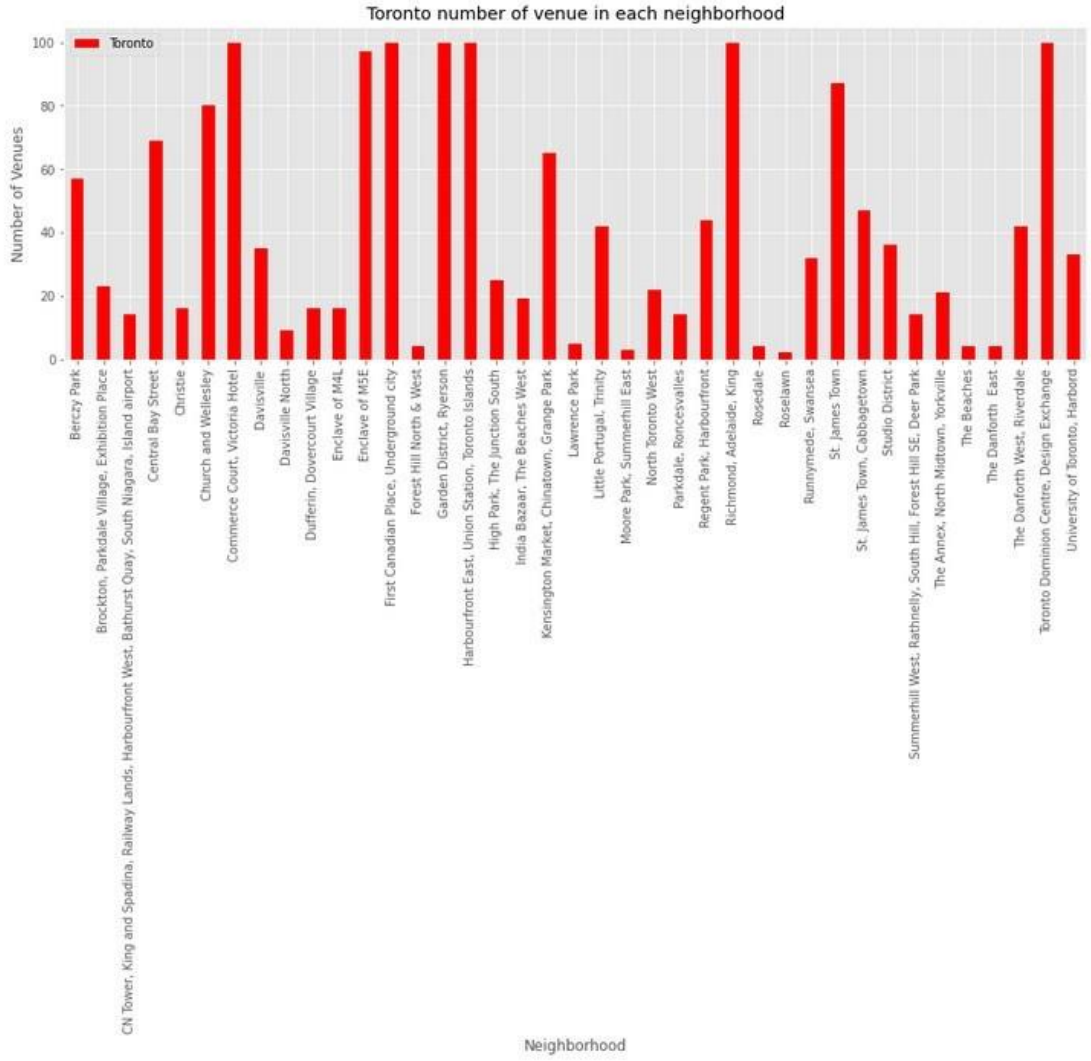

Figure 6. Number of most common Venues extracted from Foursquare in the Neighborhood of Toronto City

Further exploring the dataset, we can see the pareto of the top Venues Category extracted from Foursquare of Toronto and NewYork City in *Figure7*.

Most popular venues come up to be Coffee Shop, Italian Restaurant, Pizza Place, American Restaurant, Cafè, Bakery, Park, Gym, Bar, Mexican Restaurant, which are common Venues located in city center of big Cities.

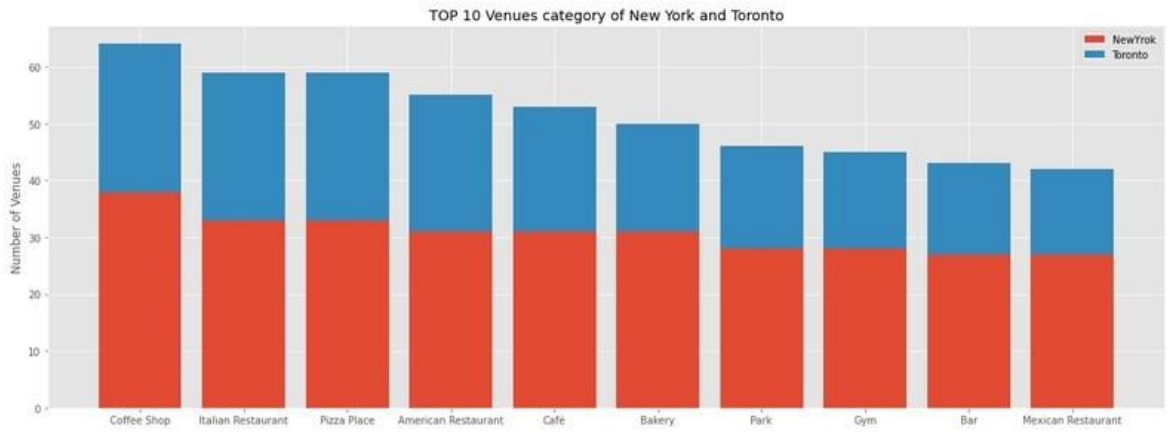

Figure 7. TOP 10 most common Venues Category

The algorithm applied on unlabeled data to solve the problem is the *K*-Means algorithm. K-means is one of the simplest unsupervised learning algorithms that solve the clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. Depending on the value that is set, a different result is obtain.
For the analysis I have decided to set k= 10 number of cluster. [2]

Before applying the algorithm I have perfomed data preparation: I have grouped the dataset based on the number of Venues category in each Neighborhood, and dropped all feature not relevant for the clustering, which are Borogh, Neighborhood, Latitude, Longitude, City, Neighborhood Latitude, Neighborhood Longitude, Venue, Venue Latitude, Venue Longitude.

	Neighborhood	Yoga Studio	Accessories Store	Acupuncturist	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Lounge	...	Video Game Store	Video Store	Vietnamese Restaurant	Volleyball Court	Waterfront	Whisky Bar	Wine Bar	Wine Shop	Wings Joint	Women's Store
0	Battery Park City	0.000000	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	...	0.000000	0.00	0.000000	0.0	0.0	0.0	0.000000	0.021277	0.0	0.021277
1	Berczy Park	0.000000	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	...	0.000000	0.00	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.000000
2	Brockton, Parkdale Village, Exhibition Place	0.000000	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	...	0.000000	0.00	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.000000
3	CN Tower, King and Spadina, Railway Lands, Har...	0.000000	0.000000	0.0	0.0	0.0	0.0	0.071429	0.071429	0.142857	...	0.000000	0.00	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.000000
4	Carnegie Hill	0.033333	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	...	0.000000	0.00	0.011111	0.0	0.0	0.0	0.011111	0.044444	0.0	0.011111
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
74	Upper East Side	0.030000	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	...	0.000000	0.00	0.000000	0.0	0.0	0.0	0.000000	0.020000	0.0	0.010000
75	Upper West Side	0.010204	0.010204	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	...	0.000000	0.00	0.010204	0.0	0.0	0.0	0.030612	0.010204	0.0	0.000000
76	Washington Heights	0.000000	0.012048	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	...	0.012048	0.00	0.000000	0.0	0.0	0.0	0.012048	0.024096	0.0	0.012048
77	West Village	0.000000	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	...	0.000000	0.00	0.000000	0.0	0.0	0.0	0.030000	0.010000	0.0	0.000000
78	Yorkville	0.000000	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	...	0.000000	0.01	0.020000	0.0	0.0	0.0	0.010000	0.030000	0.0	0.000000

Table5. K-means dataset

4 Results

Ten different number of clusters are therefore created and displayed using python folium library (*Figure8* and *Figure9*) and each Neighborhood cluster is identified with a different color.

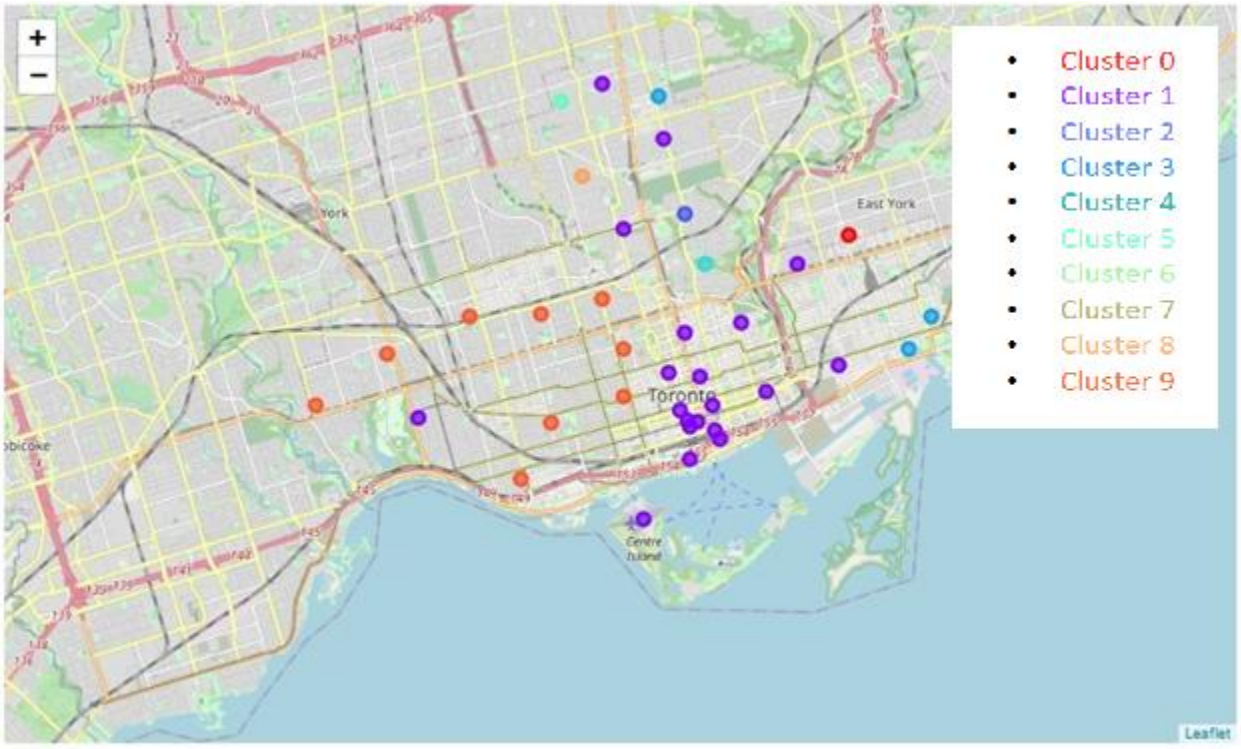

Figure 8. Neighborhood of Toronto City color labeled based on cluster category

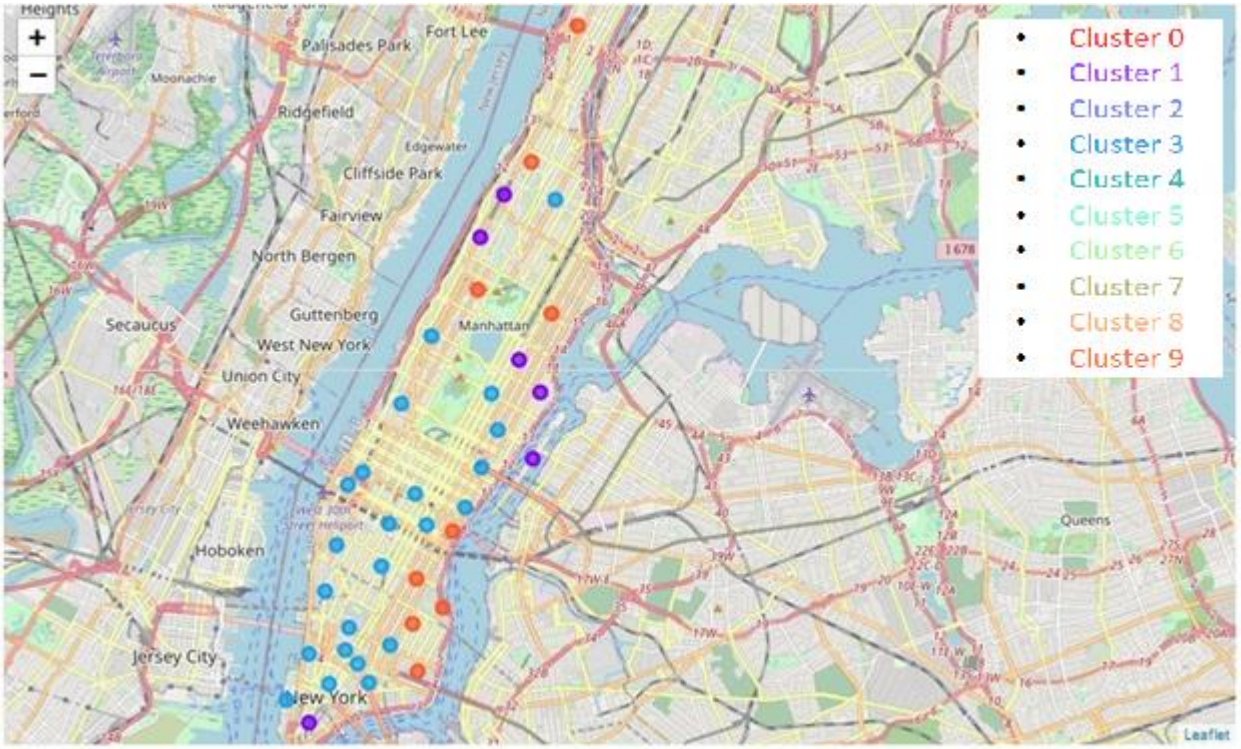

Figure 9. Neighborhood of New York City color labeled based on cluster category

5 Discussion section

The algorithm divided the different neighborhoods into ten different clusters. Cluster 0, Cluster 2, Cluster 4, Cluster 5, Cluster 6, Cluster 7, Cluster 8 have only one neighborhoods belonging to each class.

Number of Neighborhhod belonging to each cluster:

1. Cluster 0 -> number of Neighborhood: 1
2. Cluster 1 -> number of Neighborhood: 27
3. Cluster 2 -> number of Neighborhood: 1
4. Cluster 3 -> number of Neighborhood: 26
5. Cluster 4 -> number of Neighborhood: 1
6. Cluster 5 -> number of Neighborhood: 1
7. Cluster 6 -> number of Neighborhood: 1
8. Cluster 7 -> number of Neighborhood: 1
9. Cluster 8 -> number of Neighborhood: 1
10. Cluster 9 -> number of Neighborhood: 19

The clusters involving more neighborhood of both different cities are therefore cluster 1, cluster 3 and cluster 9.

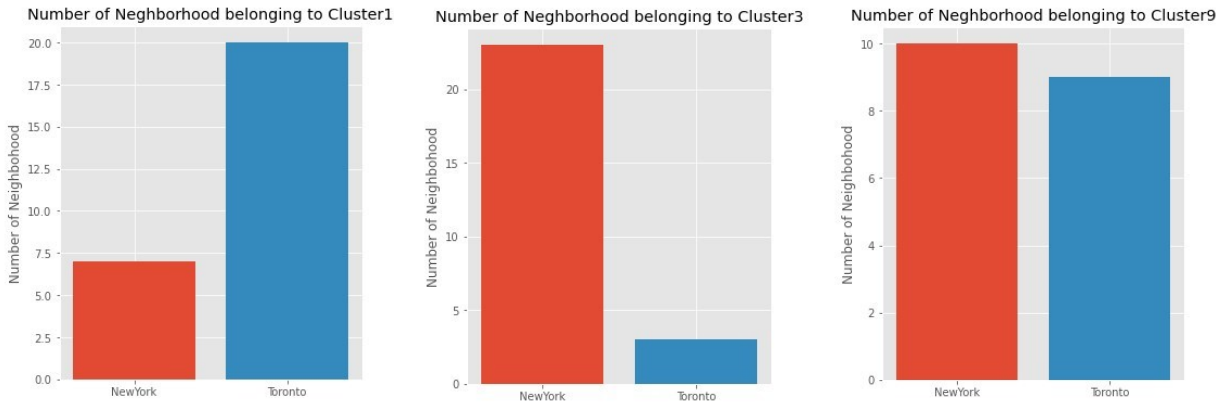

Figure 10. Number of Neighborhood belonging to the Most populated clusters

6 Conclusion section

The purpose of this project was to explore the cities of New York and Toronto and see how similar or dissimilar their neighborhood are. We explored both the cities based on their postal codes, Neighborhood and Borough and then extrapolated the common venues and facilities that each neighborhood offers within a radius of 500 meter from each given latitude and longitude associated to each neighborhood. Finally we have applied K-means algorithm to create clusters of similar neighborhood belonging to both cities. Three clusters with high number of neighborhoods belonging to each cluster were detected, which suggests that the cultural similarity is quite remarkable.
While dealing with unlabeled data and unsupervised technique it is was not possible to divide the dataset into train and test set in order to evaluate most accurate algoritm and therefore apply different technique.

Next step would be to include an additional dataset to include people satisfaction of the suggested similarities among the different neighborhood to evaluate and afterword tune the model to increase its accuracy.

7 References

- [1] Forsquare API
- [2] k-means clustering algorithm