Comparing the Cities of Toronto and New York city using neighborhood venues characteristics

Halimatou B. A.

July 26, 2020

1 Introduction

New York City and Toronto are two similar cities in the sense that they are both in the same continent, not very far from each other, multicultural cities and financial hubs in their respective countries. In this project our aim is to compare the two cities' neighborhoods using the Foursquare API to see how similar/ dissimilar they are. We will compare the two cities using their neighborhoods venues to cluster them and then analyse the results to see similarities that brought different neighborhoods to be clustered together or not. Since the two cities are very different in size (New York City is more than twice the size of Toronto), we will mainly focus on the proportion of neighborhoods that end up in the same clusters rather than the numbers. This project could be of interest to people who are from either of the two cities and looking into relocating to the other city, for instance, businesses present in one of the cities and looking into opening into the other city etc...

2 Data

For this project we will use two main sources besides the foursquare API:

- For the city of Toronto, we will use the data from the Wikipedia page: `https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M` and a file containing the location of each of the neighborhoods (latitude and longitudes) `http://cocl.us/Geospatial_data`

- For the City of New York we will use the Json file from the website: `https://geo.nyu.edu/catalog/nyu_2451_34572` Using these two soursces we will extract the Borough and neighborhoods of each of the two cities and merge them into one dataframe, keeping the name of the city in one column.

The foursquare API will then be used to obtain 100 venues within a radius of 500 miles of each of the neighborhoods. These will then be added to the dataframe to then be used for the clustering of the neighborhoods.

3 Methodology

3.1 Data preparation

Before going into the analysis, we first cleaned and prepared the data, to obtain a dataframe, with the columns: Neighborhood, Borough, City, Latitude, Longitude.
The following are the maps of the city of New York and Toronto with the different neighborhoods we are going to consider for the analysis:

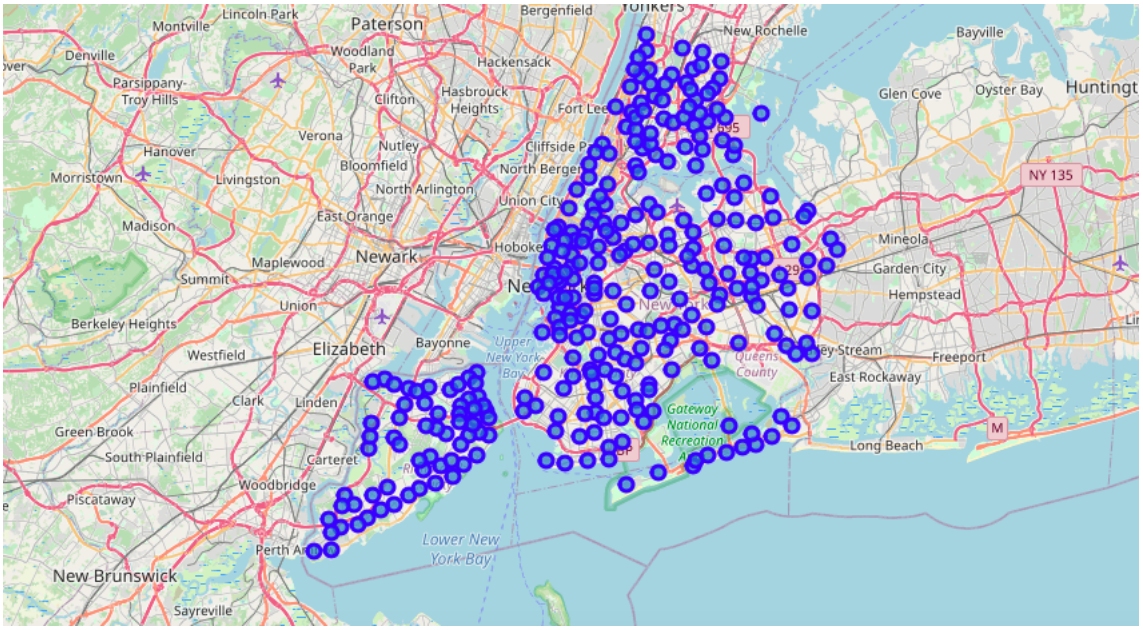

Figure 1: Neighborhoods in the city of New York

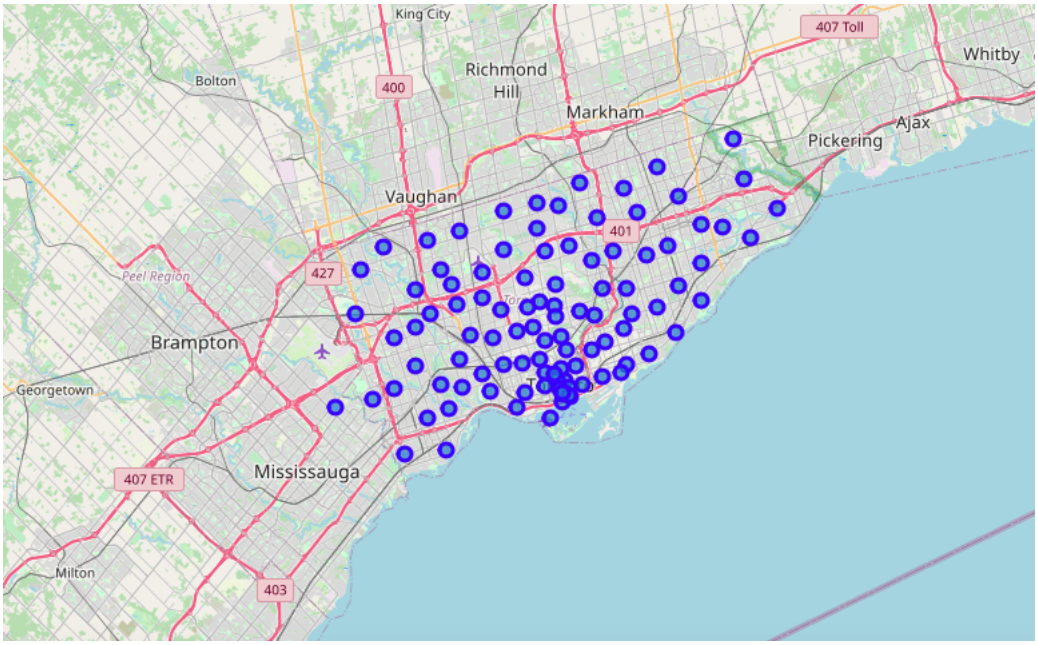

Figure 2: Neighborhoods in the city of Toronto

The following is the initial dataframe, with both New York City and Toronto Neighborhoods, of which about 3 quarters (306) are from the city of New York and a quarter (103) is from Toronto :

```
df_NY_toronto.head()
```

]:

	Borough	City	Latitude	Longitude	Neighborhood
0	North York	Toronto	43.806686	-79.194353	Parkwoods
1	North York	Toronto	43.784535	-79.160497	Victoria Village
2	Downtown Toronto	Toronto	43.763573	-79.188711	Regent Park, Harbourfront
3	North York	Toronto	43.770992	-79.216917	Lawrence Manor, Lawrence Heights
4	Downtown Toronto	Toronto	43.773136	-79.239476	Queen's Park, Ontario Provincial Government

```
df_NY_toronto.tail()
```

]:

	Borough	City	Latitude	Longitude	Neighborhood
404	Manhattan	New York	40.756658	-74.000111	Hudson Yards
405	Queens	New York	40.587338	-73.805530	Hammels
406	Queens	New York	40.611322	-73.765968	Bayswater
407	Queens	New York	40.756091	-73.945631	Queensbridge
408	Staten Island	New York	40.617311	-74.081740	Fox Hills

Figure 3: NY City and Toronto dataframes merged

The foursquare API was then used to create a get request for venues in each of the neighborhoods (limited to100 venues within a radius of 500 miles). A total of 12254 venues were retrieved and 5 attributes were selected from the request to be added onto the dataframe: venue, venue latitude, venue longitude, venue category.
A total 459 Unique categories were retrieved from the request, no venues were found in 13 neighborhoods out of the 409, so the analysis was carried out with 396 neighborhoods.
Those categories were transformed into dummy variables using one hot encoding to obtain the following dataframe:

	Neighborhood	Yoga Studio	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquariun
0	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.
1	Allerton	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.
2	Annadale	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.076923	0.0	0.
3	Arden Heights	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.
4	Arlington	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.125000	0.0	0.

Confirming size:

```
ny_t_grouped.shape
```

```
]: (396, 459)
```

Figure 4: Dataframe with categorical dummy variables

Before clustering the neighborhoods, we looked to see the most common venues in each of the neighborhoods:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Agincourt	Café	Breakfast Spot	Coffee Shop	Bakery	Nightclub
1	Allerton	Deli / Bodega	Pizza Place	Supermarket	Department Store	Breakfast Spot
2	Annadale	American Restaurant	Pizza Place	Cosmetics Shop	Pharmacy	Train Station
3	Arden Heights	Deli / Bodega	Bus Stop	Coffee Shop	Pizza Place	Pharmacy
4	Arlington	Grocery Store	Intersection	Deli / Bodega	American Restaurant	Event Service

Figure 5: Dataframe with most common venues

3.2 Clustering

To cluster the different neighborhoods, we used an unsupervised machine learning algorithm, Kmeans method from the package Scikit-Learn. This method fitted our problem in the sense that Kmeans tries to find groups based maximizing the similarity within each group.
Here we tried to get an optimal number of clusters using the elbow method for a cluster number from 1 to 15, but no clear elbow structure showed, as shown in the picture below:

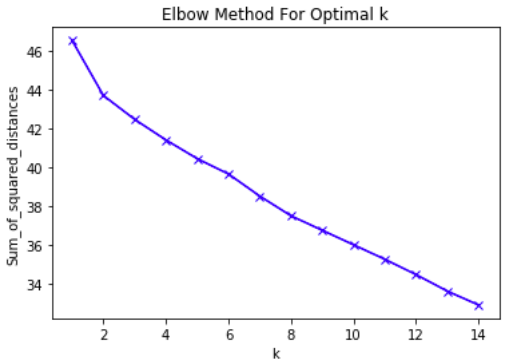

Figure 6: Elbow method for optimal k

So we settled for k=5.

4 Results

After clustering the neighborhoods into 5 groups, we obtained one large group (292 neighborhoods), followed by one medium group consisting of 96 neighborhoods, and then 3 very small groups (less than a dozen of neighborhoods each). We notice that, out of the 5 groups, one contains only neighborhoods from New York, as shown in the following table:

City	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
New York	3	84	2	215	1
Toronto	8	12	0	77	3

Figure 7: Distribution of the neighborhoods among the clusters

Below is the graph representing the proportion of neighborhoods of each city in each of the clusters:

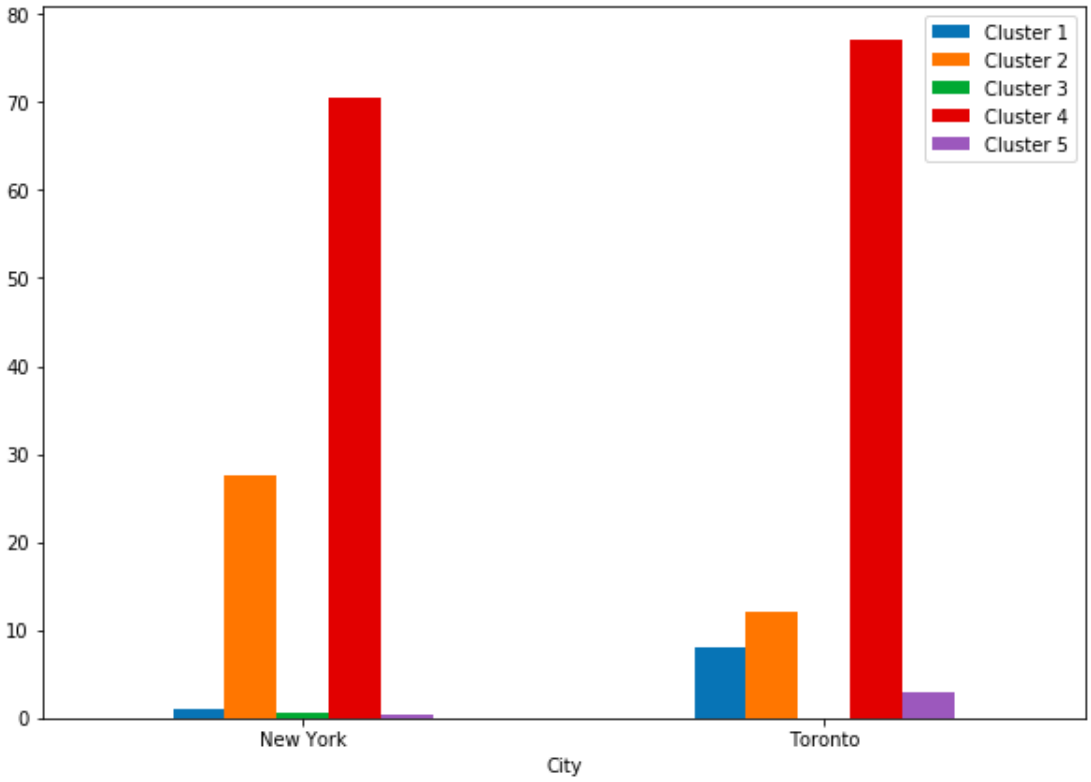

Figure 8: Distribution of the neighborhoods among the clusters

We notice that the highest majority of the neighborhoods in New York and Toronto belong to the cluster number 4 and the next highest proportion of the neighborhoods belong to the cluster 2, then clusters 1 and 5. However although in both cities the order in terms of proportion is the same, the size differs in both cities. For instance 77% of the neighborhoods in Toronto belong to the cluster 4 compared to 70% of neighborhoods in New York City. In contrast only 12% of the neighborhoods in Toronto belong to the cluster 4 compared to 27.5% of neighborhoods in New York City.

City	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
New York	1.0	27.5	0.7	70.5	0.3
Toronto	8.0	12.0	0.0	77.0	3.0

Figure 9: Percentage distribution of neighborhood per city

The following 2 figures shows the maps representing the clustered neighborhoods in New York City and Toronto:

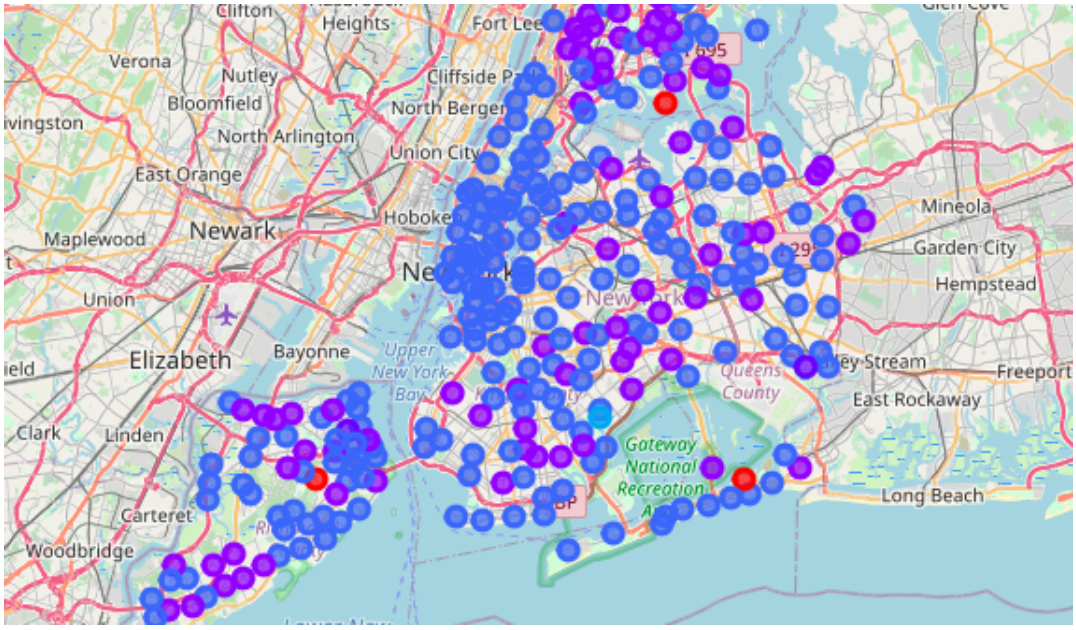

Figure 10: Clustered neighborhoods in New York City

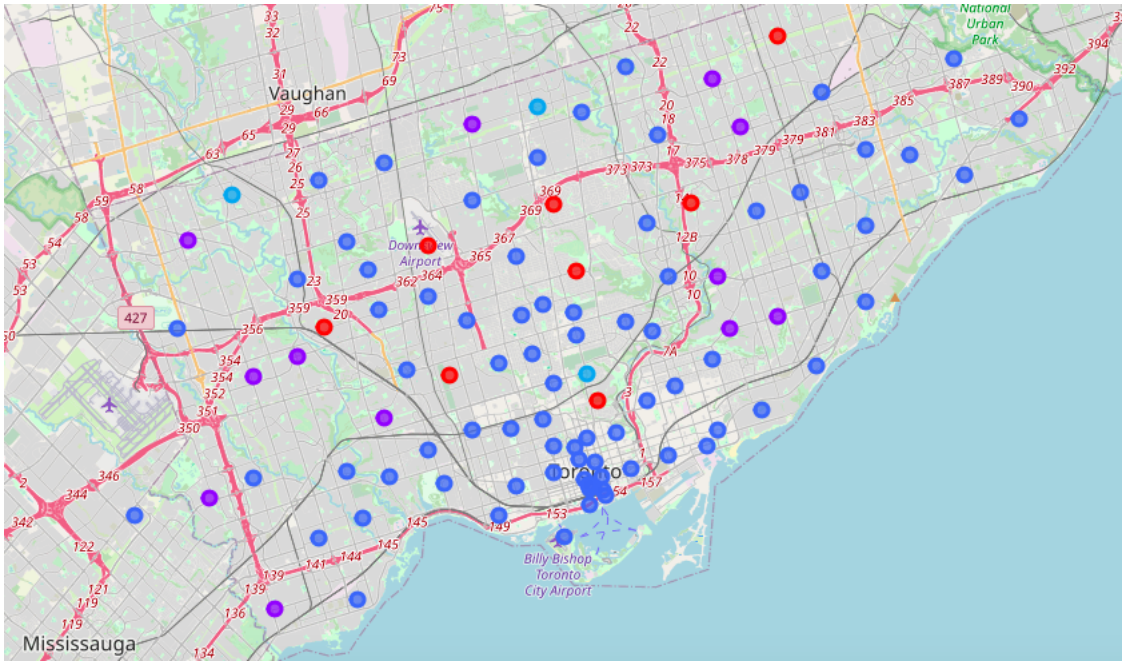

Figure 11: Clustered neighborhoods in Toronto

5 Discussion

From the results presented, we can say that Toronto and New York city are relatively similar. Out of the 5 identified clusters of neighborhoods found in New York City, 4 types can be found in Toronto, with a similar orders in terms of proportion. Moreover, most of the neighborhoods in each of the cities fall in the same category and the only category of neighborhoods that is not present in Toronto, is very small (2 neighborhoods), so these could almost be considered as outliers.

On the other hand one, these results need to be interpreted with caution, New York City is almost 3 times the size of Toronto, so although there are similar features, New York city would obviously have more diversity in terms of its neighborhood characteristics. But despite the big difference in size, the fact that we were able to find most neighborhoods fall into similar categories reinforces our hypothesis that the two cities are similar.

6 Conclusion

In conclusion, based on our analysis, New York City and Toronto are relatively similar. We can find almost the same categories of neighborhoods in both cities, despite the fact that the proportions are different and the size of the two cities are very different too. So a person moving between the cities is almost sure to find a neighborhood with similar characteristics in one of the other cities.

References

Foursquare API
Wikipedia: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
http://cocl.us/Geospatial_data1