Capstone Project - Battle of Neighborhoods

Tiago Jorge Loureiro de Freitas

1 Introduction and Business problem

New York City (NYC) is not only one of the most crowded urban areas in USA but also one of the most etymologically different city on the planet. Similarly to NYC, Toronto is the most crowded city in Canada and it's also one of the most multicultural and cosmopolitan urban communities on the planet. A shareholder of a food chain company with restaurants in New York is interested in opening a new restaurant in Toronto but wants to better understand the possible similarities between neighborhoods in NYC and neighborhoods in Toronto.

For this, we will cluster both cities into various groups depending on the classification (kind) of venues in these areas (using the Forusquares, similarly to the previous project). It is expected that different venues categories will be more common in certain groups than others and the most common venue categories should also contrast from one group to another. Using this information we can find out which neighborhoods in Toronto compare to a certain neighborhood in NYC and make a more informed decision on where to open the restaurant in Toronto.

2 Data

The data used for the clustering analysis will have: each neighborhood and their coordinates, as well as the venues and their corresponding coordinates, category and neighborhood in order to find out the most common venues in a specific neighborhood. The NYC data will be obtained from the json file (from `https://geo.nyu.edu/catalog/nyu_2451_34572`) and the Toronto data will be obtained from scraping the Wikipedia link (`https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M`) and the csv file with the coordinates (`https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs_v1/Geospatial_Coordinates.csv`). The corresponding venues data will be obtained using Foursquares.

3 Methodology

3.1 Data Processing

As stated above, the data pertaining the city of Toronto will be obtained from two files: a wikipedia page (webscraping) for the boroughs names and a csv file with the coordinates. We will use the `BeautifulSoup` package to obtain the table of the boroughs along with their postal codes and neighborhoods and the pandas `pd.read_csv` to obtain the coordinates table. The tables will be joined into a table like the one presented below.

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Queen's Park	Ontario Provincial Government	43.662301	-79.389494

Figure 1: Example of the Toronto neighborhood table (first five rows).

The New York data is obtained from the json file, just like in the Week 3 Lab " Segmenting and Clustering Neighborhoods in New York City". This data will be processed into a table similar to the one for the Toronto neighborhoods.

The corresponding venues data will be obtained by importing the nearby venues (using Foursqeares) of each neighborhood, grouping them by category and using one-hot encoding to obtain the most common categories of venues for each neighborhood, like in Figure 2.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Agincourt	Lounge	Skating Rink	Latin American Restaurant	Breakfast Spot	Donut Shop	Diner	Discount Store	Distribution Center	Dog Run	Doner Restaurant
1	Alderwood, Long Branch	Pizza Place	Gym	Coffee Shop	Pub	Sandwich Place	Distribution Center	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store
2	Bathurst Manor, Wilson Heights, Downsview North	Coffee Shop	Bank	Park	Fried Chicken Joint	Sandwich Place	Bridal Shop	Diner	Restaurant	Deli / Bodega	Intersection
3	Bayview Village	Café	Bank	Japanese Restaurant	Chinese Restaurant	Women's Store	Doner Restaurant	Discount Store	Distribution Center	Dog Run	Donut Shop
4	Bedford Park, Lawrence Manor East	Sandwich Place	Italian Restaurant	Coffee Shop	Comfort Food Restaurant	Thai Restaurant	Juice Bar	Restaurant	Fast Food Restaurant	Butcher	Pub

Figure 2: Example of the venues data for Toronto (first five rows).

3.2 Data Visualization

We will use the folium package in order to represent each neighborhood in a map of their respective city, as shown in Figures 3 and 4.

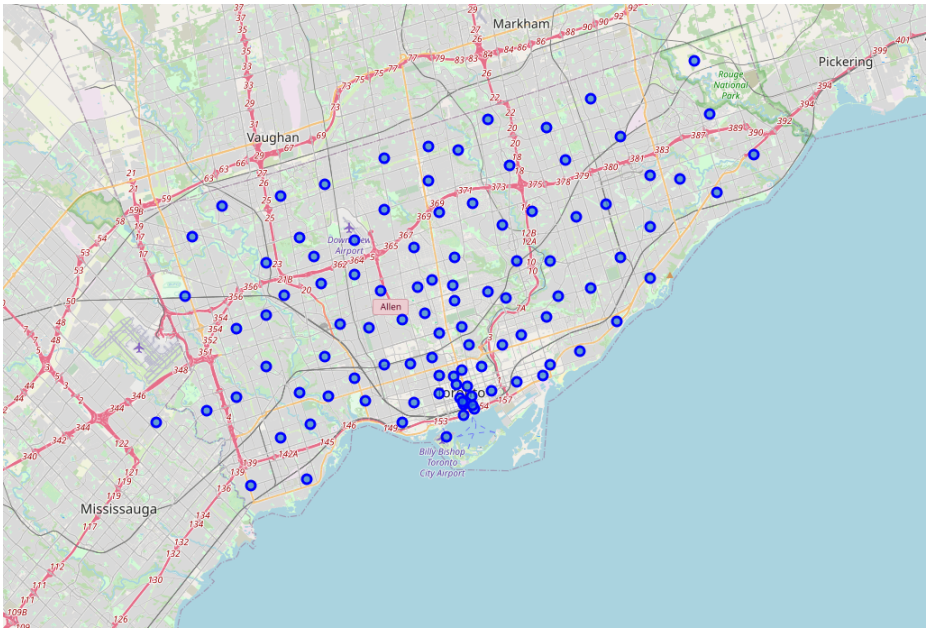

Figure 3: Folium map for the neighborhoods in Toronto.

3.3 Clustering

For this project we will focus in aggregating the different neighborhoods into different clusters using K-Means clustering based on their most common venues. We chose to use $K = 5$ for both cities and will represent the clusters with different colors using the folium package.

4 Results

After clustering the neighborhoods in both cities, we represented them again in a folium map with each cluster as a different color in Figures 10 and 11. We can also take a closer look at each cluster and their neighborhoods with corresponding most common venues as exemplified for the Toronto clusters in Figures 5, 6, 7, 8 and 9.

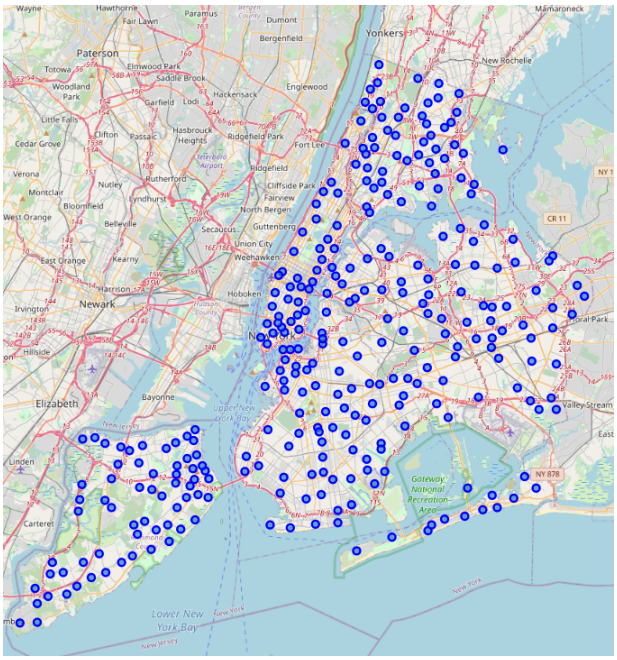

Figure 4: Folium map for the neighborhoods in New York City.

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	North York	0.0	Food & Drink Shop	Park	Fast Food Restaurant	Women's Store	Diner
16	York	0.0	Park	Field	Trail	Hockey Arena	Escape Room
21	York	0.0	Park	Women's Store	Pool	Doner Restaurant	Dessert Shop
40	North York	0.0	Park	Bus Stop	Airport	Donut Shop	Diner
52	North York	0.0	Park	Women's Store	Drugstore	Diner	Discount Store

Figure 5: Toronto - Cluster 1

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
57	North York	1.0	Baseball Field	Women's Store	Diner	Discount Store	Distribution Center
101	Etobicoke	1.0	Baseball Field	Women's Store	Diner	Discount Store	Distribution Center

Figure 6: Toronto - Cluster 2

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	North York	2.0	Intersection	French Restaurant	Coffee Shop	Pizza Place	Hockey Arena
2	Downtown Toronto	2.0	Coffee Shop	Park	Bakery	Pub	Café
3	North York	2.0	Clothing Store	Accessories Store	Furniture / Home Store	Vietnamese Restaurant	Boutique
4	Queen's Park	2.0	Coffee Shop	Sushi Restaurant	Yoga Studio	Bar	Spa
6	Scarborough	2.0	Fast Food Restaurant	Women's Store	Donut Shop	Dim Sum Restaurant	Diner

Figure 7: Toronto - Cluster 3

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
12	Scarborough	3.0	Home Service	Bar	Dessert Shop	Diner	Discount Store
53	North York	3.0	Home Service	Baseball Field	Food Truck	Dim Sum Restaurant	Discount Store
62	Central Toronto	3.0	Garden	Home Service	Event Space	Ethiopian Restaurant	Escape Room

Figure 8: Toronto - Cluster 4

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
32	Scarborough	4.0	Grocery Store	Playground	Women's Store	Donut Shop	Dim Sum Restaurant
46	North York	4.0	Grocery Store	Park	Shopping Mall	Bank	Doner Restaurant

Figure 9: Toronto - Cluster 5

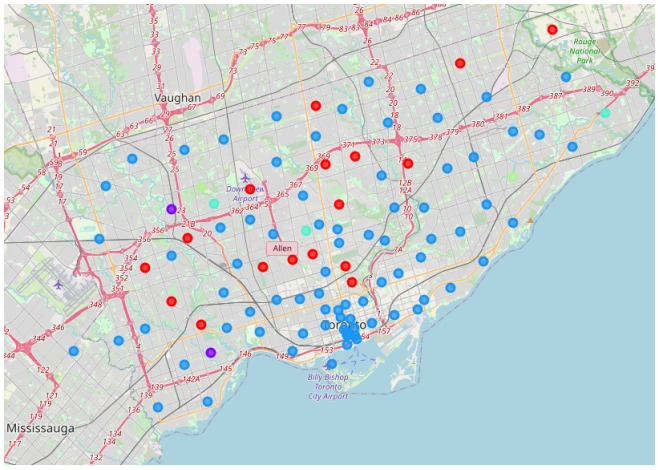

Figure 10: Folium map for the neighborhoods in Toronto with clusters.

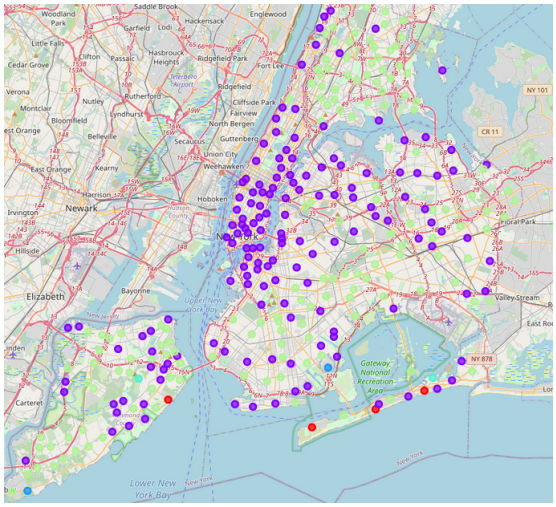

Figure 11: Folium map for the neighborhoods in New York City with clusters.

5 Discussion

It's now time to use the information obtained from the clustering analysis and apply it to our particular problem. Let's say the shareholder of the restaurant chain has a very popular and profitable restaurant in one of the neighborhoods in New York, which neighborhoods in Toronto are similar to the one in NYC that should be considered an option to open the new restaurant?

Let's say the restaurant is at the neighborhood corresponding to index 27 (random choice). This corresponds to Clason Point in the Bronx with the corresponding most common venues presented below.

```
Borough                                 Bronx
Neighborhood                      Clason Point
Latitude                               40.8066
Longitude                             -73.8541
Cluster Labels                               1
1st Most Common Venue                     Park
2nd Most Common Venue     South American Restaurant
3rd Most Common Venue               Scenic Lookout
4th Most Common Venue                    Bus Stop
5th Most Common Venue                Grocery Store
6th Most Common Venue                        Pool
7th Most Common Venue                Boat or Ferry
8th Most Common Venue                   Playground
9th Most Common Venue           Convenience Store
10th Most Common Venue          Filipino Restaurant
Name: 27, dtype: object
```

Figure 12: Clason Point's most common venues (neighborhood with the shareholder's restaurant)

We can also see the cluster where this neighborhood is located. To compare this cluster to another cluster in Toronto let's see what are the 10 1st most common venues in this cluster's neighborhood, meaning what venue category is usually the 1st most common.

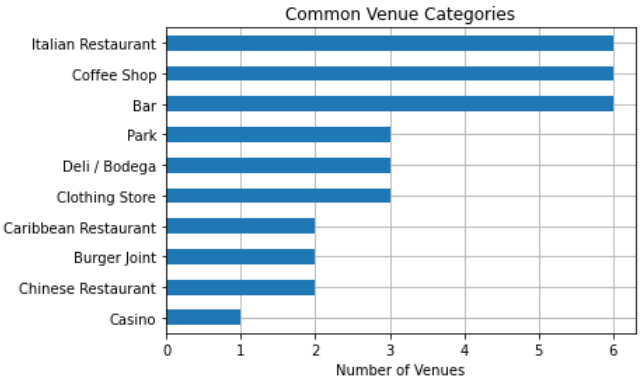

Figure 13: 1st most common venues in cluster containing Clason Point

We can now use this information to compare to the clusters in Toronto. Clusters 2, 4 and 5 are composed of only 2 or 3 neighborhoods so we can exclude just by looking at the Figures 6, 8 and 9. Then, let´s repeat the process used for the NYC cluster to clusters 1 and 3 of Toronto.

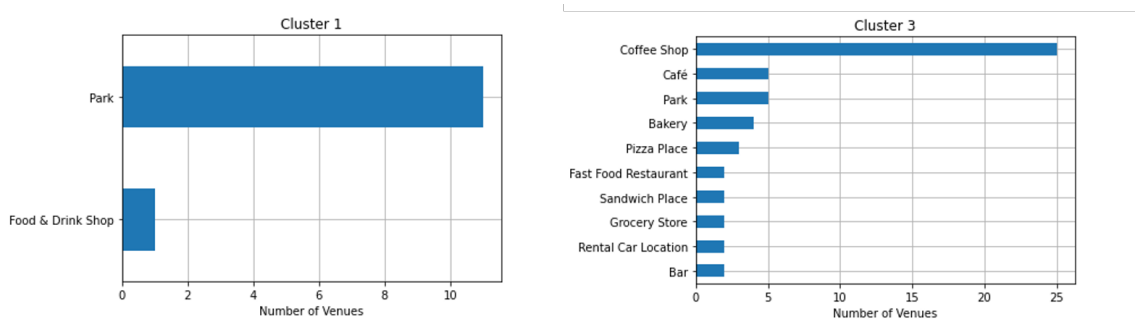

Figure 14: Comparison of 1st most common venues in Toronto's clusters 1 and 3

With the information form Figure 14 we can suggest that the new restaurant should be opened in one of the neighborhoods belonging to Toronto's cluster 3.

6 Conclusion

With this project we were able to compare the neighborhoods in New York City with the neighborhoods in Toronto based on their most common venues. This way, we have important additional information to make business decisions like selecting which neighborhood would be the best to open a new restaurant in a new city.