# Battle Of Neighborhoods

## Explore and Cluster the Neighborhoods of Toronto

**Pooja Bhateley**

**12th April, 2020**

## *1. Introduction*

### 1.1 Background

Toronto is the provincial capital of Ontario and the most populous city in Canada, with a population of 2,731,571 in 2016. Current to 2016, the Toronto census metropolitan area (CMA), of which the majority is within the Greater Toronto Area (GTA), held a population of 5,928,040, making it Canada's most populous CMA. The city is the anchor of the Golden Horseshoe, an urban agglomeration of 9,245,438 people (as of 2016) surrounding the western end of Lake Ontario.

Toronto is an international center of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world.

Brief history about Toronto: People have travelled through and inhabited the Toronto area, located on a broad sloping plateau interspersed with rivers, deep ravines, and urban forest, for more than 10,000 years. After the broadly disputed Toronto Purchase, when the Mississauga surrendered the area to the British Crown, the British established the town of York in 1793 and later designated it as the capital of Upper Canada. During the War of 1812, the town was the site of the Battle of York and suffered heavy damage by United States troops. York was renamed and incorporated in 1834, as the city of Toronto. It was designated as the capital of the province of Ontario in 1867 during Canadian Confederation.

### 1.2 Business Problem

Picture a scenario wherein you live in a perfect neighborhood with all the amenities in close proximity. Then you get a job opportunity in the other part of the city, where you have not been often and are not aware about the facilities and infrastructure available. All your favourite eating joints, coffee shops are in your present neighborhood. But you can't decline the job opportunity and you can't commute everyday for 4 hours from your present residence. The only choice is to move to the new neighborhood. Wouldn't it be great if you can find a house in the locality which is exactly like your present neighborhood, with the same amenities and similar food joints, and which is also closer to your new job?

### 1.3 Objective

The objective of this project is to analyse the various neighborhoods of Toronto, explore the different amenities available in the city, and cluster these neighbouhoods into with similar characteristics. Thus using this clustered data to find neighborhoods which are similar to the current neighborhood of the user.

# *2. Data Acquisition*

## 2.1 Data Sources

To consider the objective stated above, we can list the below data sources used for the analysis.

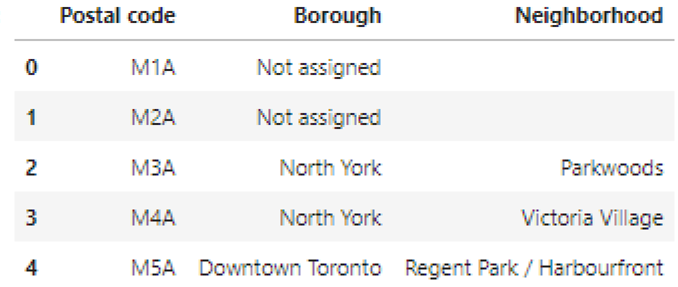
a) Toronto Neighborhood Data: The following Wikipedia page was scraped to pull out the necessary information: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

The information obtained i.e. the table of postal codes was transformed into a pandas data frame for further analysis.

b) Coordinate data for each Neighborhood in Toronto: The following csv file gave us the geographical coordinates of each postal code: http://cocl.us/Geospatial_data

## 2.2 Data Scraping

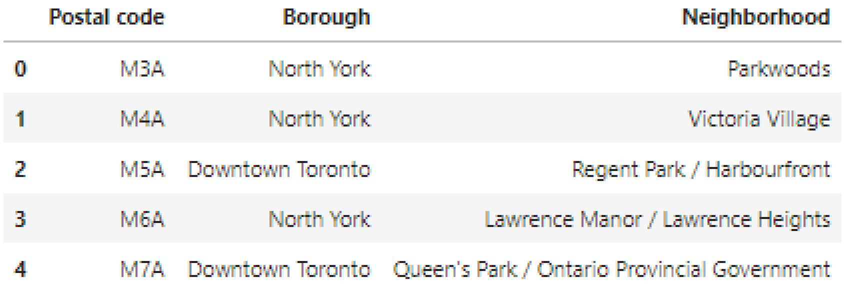
Data is scraped from the Wikipedia page using **Beautiful Soup** library. Using soup object, iterating the .wikitable to get the data from the HTML page and storing it into a list. The list is then converted into **Pandas** Data Frame, which appears as below.



| | Postal code | Borough | Neighborhood |
|---|---|---|---|
| **0** | M1A | Not assigned | |
| **1** | M2A | Not assigned | |
| **2** | M3A | North York | Parkwoods |
| **3** | M4A | North York | Victoria Village |
| **4** | M5A | Downtown Toronto | Regent Park / Harbourfront |

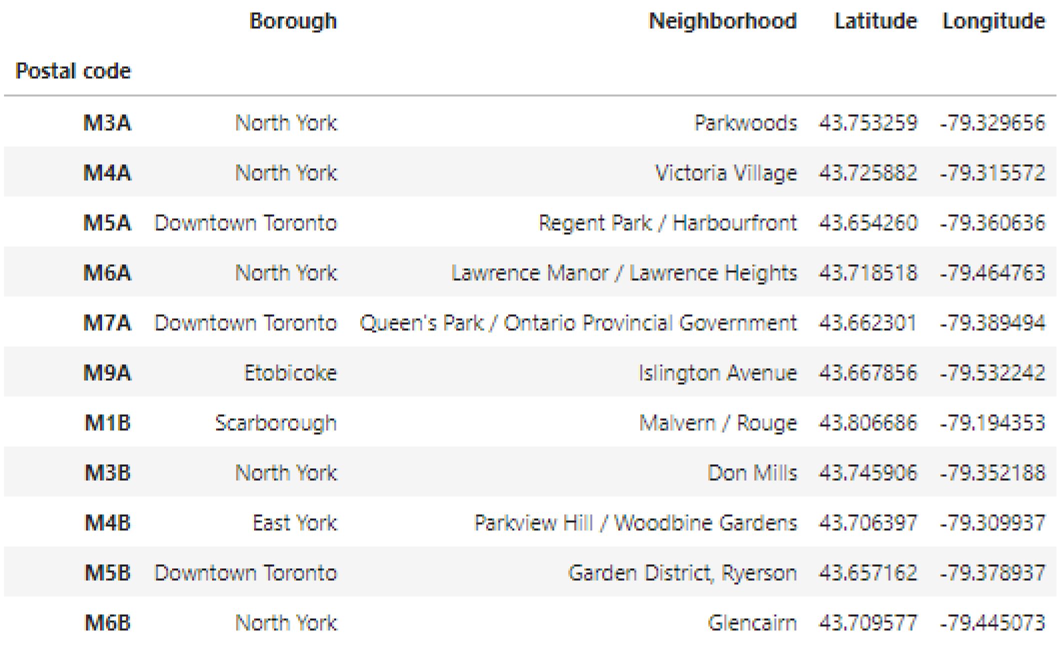
## 2.3 Data Cleaning

Preprocessing the data, by removing rows with "Not Assigned" values, removing duplicating and grouping the neighborhoods with the same postal code, the data frame appears as below.



| | Postal code | 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 | Downtown Toronto | Queen's Park / Ontario Provincial Government |

Now, getting the Geospacial data from the CSV file, and combining it with the data frame with Canada data. The combined data frames appears as below.

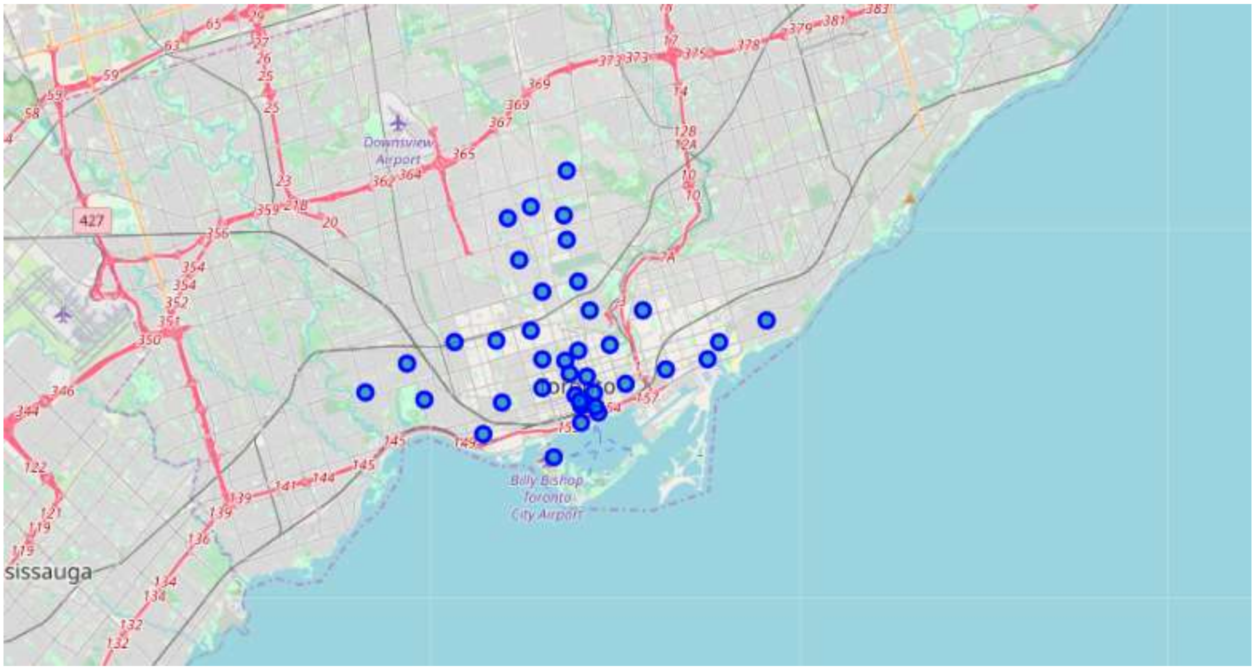


| Postal code | Borough | Neighborhood | Latitude | Longitude |
|---|---|---|---|---|
| M3A | North York | Parkwoods | 43.753259 | -79.329656 |
| M4A | North York | Victoria Village | 43.725882 | -79.315572 |
| M5A | Downtown Toronto | Regent Park / Harbourfront | 43.654260 | -79.360636 |
| M6A | North York | Lawrence Manor / Lawrence Heights | 43.718518 | -79.464763 |
| M7A | Downtown Toronto | Queen's Park / Ontario Provincial Government | 43.662301 | -79.389494 |
| M9A | Etobicoke | Islington Avenue | 43.667856 | -79.532242 |
| M1B | Scarborough | Malvern / Rouge | 43.806686 | -79.194353 |
| M3B | North York | Don Mills | 43.745906 | -79.352188 |
| M4B | East York | Parkview Hill / Woodbine Gardens | 43.706397 | -79.309937 |
| M5B | Downtown Toronto | Garden District, Ryerson | 43.657162 | -79.378937 |
| M6B | North York | Glencairn | 43.709577 | -79.445073 |

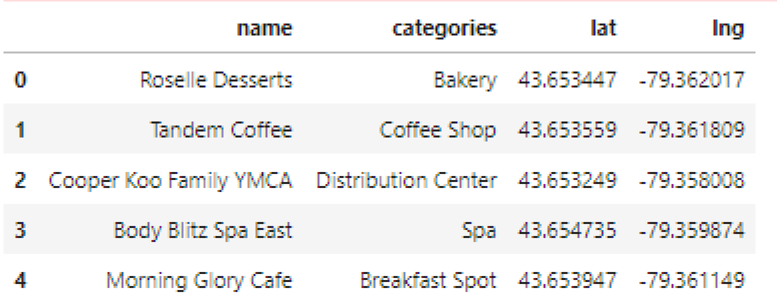
Final step is to filter the data to use only Boroughs in Toronto

## *3.Methodology*

Exploring an initial map of the neighborhoods in Toronto, using **Matplotlib** and **Folium** libraries and the latitudes and longitudes.

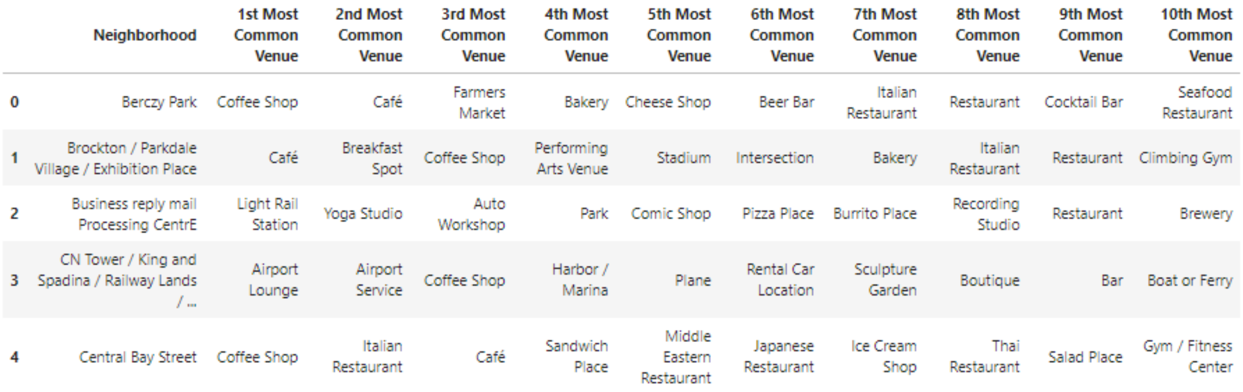


Now, taking the first neighborhood, and getting exploring the venues nearby using Foursquare API, which is the Location Data Application. The data for the first neighborhood, i.e, Regent Park/ Harbourfront appears as below.|



| | name | categories | lat | lng |
|---|---|---|---|---|
| 0 | Roselle Desserts | Bakery | 43.653447 | -79.362017 |
| 1 | Tandem Coffee | Coffee Shop | 43.653559 | -79.361809 |
| 2 | Cooper Koo Family YMCA | Distribution Center | 43.653249 | -79.358008 |
| 3 | Body Blitz Spa East | Spa | 43.654735 | -79.359874 |
| 4 | Morning Glory Cafe | Breakfast Spot | 43.653947 | -79.361149 |

Now, in the similar way, creating a function, to explore and extract the nearby venues in each neighborhood and storing them in a data frame. The category of these venues are then analysed. Using one-hot encoding method, the top 10 venues of each neighbourhood are identified and depicted in a data frame as shown below.

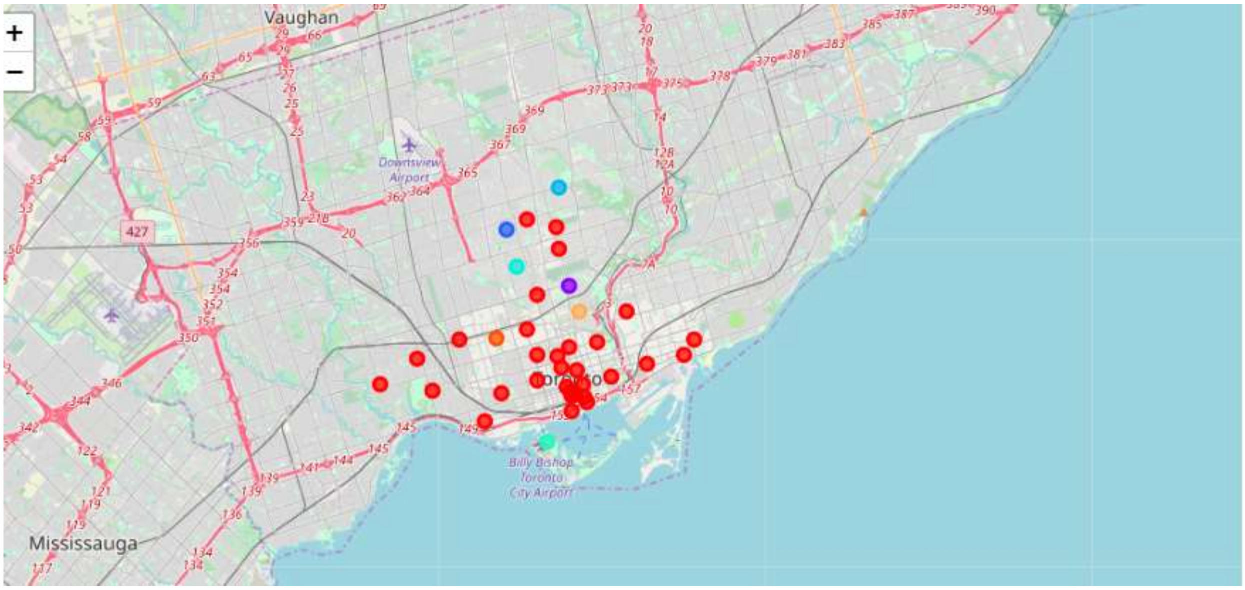


| | Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Berczy Park | Coffee Shop | Café | Farmers Market | Bakery | Cheese Shop | Beer Bar | Italian Restaurant | Restaurant | Cocktail Bar | Seafood Restaurant |
| 1 | Brockton / Parkdale Village / Exhibition Place | Café | Breakfast Spot | Coffee Shop | Performing Arts Venue | Stadium | Intersection | Bakery | Italian Restaurant | Restaurant | Climbing Gym |
| 2 | Business reply mail Processing CentrE | Light Rail Station | Yoga Studio | Auto Workshop | Park | Comic Shop | Pizza Place | Burrito Place | Recording Studio | Restaurant | Brewery |
| 3 | CN Tower / King and Spadina / Railway Lands / ... | Airport Lounge | Airport Service | Coffee Shop | Harbor / Marina | Plane | Rental Car Location | Sculpture Garden | Boutique | Bar | Boat or Ferry |
| 4 | Central Bay Street | Coffee Shop | Italian Restaurant | Café | Sandwich Place | Middle Eastern Restaurant | Japanese Restaurant | Ice Cream Shop | Thai Restaurant | Salad Place | Gym / Fitness Center |

Now that we have the top 10 categories in each neighborhood, we can group the neighborhoods based on their similarities. The algorithm used here will be K-Mean Clustering using **Scikit Learn** Library.

## *4. Results*

The clusters are then depicted on the Toronto map and are color coded,
using **Matplotlib** and **Folium** libraries.



## 4. Discussion and Conclusion

Thus, we see that data which is constantly updated can speak volumes. The postal code data of a city and the location data from the Foursquare API can be used to generate an unsupervised clustered model which can be used to make real life decisions. In a fast moving world, there are many real life problems or scenarios where data can be used to find solutions to those problems.

## 5. References

1. Wikipedia content: https://en.wikipedia.org/wiki/Toronto
2. CSV for Coordinate data: http://cocl.us/Geospatial_data
3. Foursquare API

**--Thank You--**