Final Report for
Week 4 Battle of Neighborhood

Mauro Mariotto, June 2020

1. Introduction

Toronto is a very large city with several attractions like shops, restaurants, parks, coffee bar, cinemas, Pizzas etc. The city has a lot of neighborhoods in each district called boroughs and several districts are very wide. The goal of this work is to identify some neighborhoods having special characteristics required by the stakeholders in order to simplify the selection of some zones of the city for renting an apartment for a temporary period.

1.1 Problem

The problem to solve is to identify some zones of the city for a business man coming from New York City to Toronto for a temporary period. He would like to receive some information about the structure of the city for renting an apartment in a neighborhood with some features. He asks for the following neighborhood characteristics:

  i.   The neighborhoods must be at maximum 4 Km far from the city center of Toronto,
  ii.   The neighborhoods must have at least 30 Venues within 500 meters, so to a neighborhood with some activities,
  iii.   The neighborhood must have Restaurants and/or Pizza available as he does not like to eat fast food in the street.

We discovered during our analysis that the borough or districts of Toronto are too wide for the purpose of this project and so we decided to analyze neighborhoods and not restrict our view to the districts.

2. Data acquisition and cleaning

2.1 Data source

We have several datasources available. The first one used for this selection comes from a wikipedia page https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M where are listed many boroughs with postacode with their neighborhoods. The second information comes from a GEO spatial coordinates csv file with latitude and longitude of the Postalcodes of Toronto. And finally the most important information comes directly invoking the Foursquare Database API for the city of Toronto asking for venues for each neighborhood. We used the Foursquare API after having captured for each neighborhood its latitude and its longitude using the geopy library available for Python. The library works well and delivers the exact location of the neighborhood of Toronto in GPS coordinates.

As recap, the following datasources were used:

- Wikipedia Page
- Foursquare API
- Geospatial_coordinates.csv file
- Geopy API for GPS Latitude and Longitude of each neighborhood of Toronto

2.2 Data cleaning

As the data coming from the Wikipedia page was not available as csv file we had to parse the html file in order to extract the required information for the city of Toronto. For this purpose we used the BeautifoulSoap API successfully obtaining the following clean dataset after removing all information with N/A or Unassigned Neighborhood identification. The data obtained from the wikipedia page and from the GEO-Spatial coordinates csv file for all borough of Toronto were unfortunately not sufficient for the scope of our analysis and so we decided to enrich the dataset with further information for each neighborhood with its Latitude, Longitude and distance from the city center in meters. For this purpose we used the **geopy API** in order to exactly identify the GPS center of each neighborhood of all available districts. This extra information was necessary for calculating the distance in meters from the city center.

Dataset from Wikipedia page after parsing and removing the Unassigned data (top 10)

	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	Downtown Toronto	Queen's Park, Ontario Provincial Government
5	M9A	Etobicoke	Islington Avenue, Humber Valley Village
6	M1B	Scarborough	Malvern, Rouge
7	M3B	North York	Don Mills
8	M4B	East York	Parkview Hill, Woodbine Gardens
9	M5B	Downtown Toronto	Garden District, Ryerson

*Dataset with Geospatial Coordinates for each Borough (top 10)*

	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	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
5	M9A	Etobicoke	Islington Avenue, Humber Valley Village	43.667856	-79.532242
6	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
7	M3B	North York	Don Mills	43.745906	-79.352188
8	M4B	East York	Parkview Hill, Woodbine Gardens	43.706397	-79.309937
9	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937

As mentioned above we used the **Foursquare API and the Geopy API** to enrich the dataset with information for every neighborhood as the boroughs of Toronto are too large and wide for the purpose of this selection. The solution was to analyze the city of Toronto in each neighborhood in order to achieve a better selection model as required by the stakeholders. In order to calculate the 2D metrics for each neighborhood we also used the **Pyproj API** permitting us to calculate the distance in meters from the city center of Toronto.

Dataset with Geospatial Coordinates for each Neighborhood (top 5) with distance from the city center

	PostalCode	Borough	Neighborhood	PostalCode_lat	PostalCode_lng	Neigh_lat	Neigh_lng	Neigh_UTM_x	Neigh_UTM_y	Distance_from_centre
0	M3A	North York	Parkwoods	43.753259	-79.329656	43.758800	-79.320197	635222.011341	4.846455e+06	12611.614627
1	M4A	North York	Victoria Village	43.725882	-79.315572	43.732658	-79.311189	636006.297239	4.843566e+06	10364.971928
2	M5A	Downtown Toronto	Regent Park	43.654260	-79.360636	43.660706	-79.360457	632196.554750	4.835495e+06	1767.465727
3	M5A	Downtown Toronto	Harbourfront	43.654260	-79.360636	43.640080	-79.380150	630653.432519	4.833173e+06	1544.342400
4	M6A	North York	Lawrence Manor	43.718518	-79.464763	43.722079	-79.437507	625855.506783	4.842192e+06	8858.449803

Toronto Map with all neighborhoods available from the wikipedia page
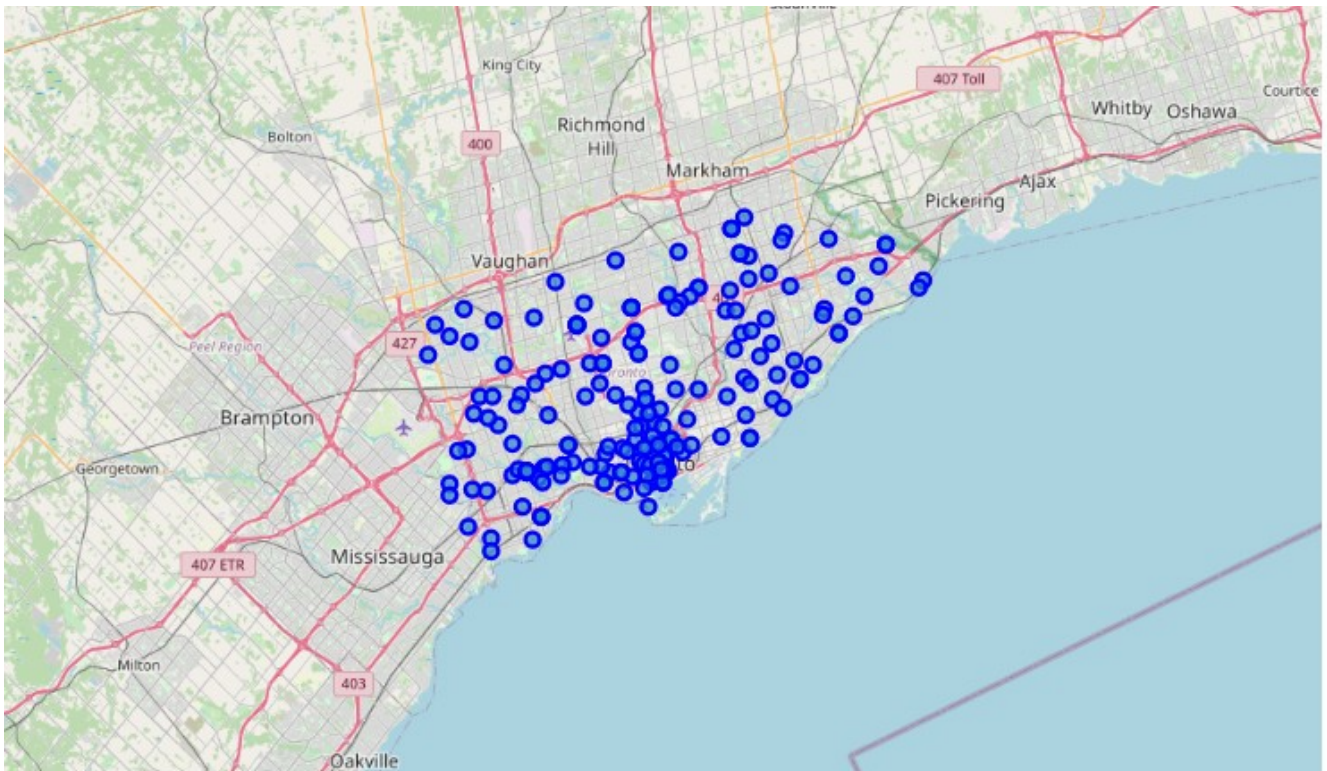
2.3 Foursquare API for venues information for all neighborhoods of Toronto

The foursquare API was executed for each neighborhood of Toronto with the following limitations: 500 meters distance from the GPS coordinates

Here an example of the API during execution

```
Calling foursquare API for neighborhood: Parkwoods
Calling foursquare API for neighborhood: Victoria Village
Calling foursquare API for neighborhood: Regent Park
Calling foursquare API for neighborhood: Harbourfront
Calling foursquare API for neighborhood: Lawrence Manor
Calling foursquare API for neighborhood: Lawrence Heights
Calling foursquare API for neighborhood: Queen's Park
Calling foursquare API for neighborhood: Islington Avenue
Calling foursquare API for neighborhood: Humber Valley Village
Calling foursquare API for neighborhood: Malvern
Calling foursquare API for neighborhood: Rouge
Calling foursquare API for neighborhood: Don Mills
Calling foursquare API for neighborhood: Parkview Hill
Calling foursquare API for neighborhood: Woodbine Gardens
Calling foursquare API for neighborhood: Garden District
Calling foursquare API for neighborhood: Ryerson
Calling foursquare API for neighborhood: Glencairn
Calling foursquare API for neighborhood: West Deane Park
Calling foursquare API for neighborhood: Princess Gardens
```

After calling the Foursquare API we obtained a rich dataset of information about each neighborhood with venue name, GPS coordinate and category.

We found more than 300 unique categories in the city of Toronto

|	Neighborhood	Neigh_lat	Neigh_lng	Venue_name	Venue_lat	Venue_lng	Venue_cat
0	Parkwoods	43.758800	-79.320197	Allwyn's Bakery	43.759840	-79.324719	Caribbean Restaurant
1	Parkwoods	43.758800	-79.320197	LCBO	43.757774	-79.314257	Liquor Store
2	Parkwoods	43.758800	-79.320197	Petro-Canada	43.757950	-79.315187	Gas Station
3	Parkwoods	43.758800	-79.320197	Shoppers Drug Mart	43.760857	-79.324961	Pharmacy
4	Parkwoods	43.758800	-79.320197	Pizza Pizza	43.760231	-79.325666	Pizza Place
...	...	...	...	...	...	...	...
5932	Royal York South West	43.648183	-79.511296	The Old Sod	43.648297	-79.507642	Pub
5933	Royal York South West	43.648183	-79.511296	Tim Hortons	43.646678	-79.513700	Coffee Shop
5934	Royal York South West	43.648183	-79.511296	Rogers	43.647080	-79.511550	Mobile Phone Shop
5935	Royal York South West	43.648183	-79.511296	Gabby's Grill & Taps	43.648452	-79.506482	Bar
5936	Royal York South West	43.648183	-79.511296	Vince's Market	43.646193	-79.515962	Grocery Store

5937 rows × 7 columns

2.4 Feature selection and row filtering

The dataset was filtered using the rules indicated by the stackeholders, namely the neighborhood must be max 4 Km far from the city center and have at least 30 venues around with a Restaurant or a Pizza House available.

Dataset with Geospatial Coordinates for each Neighborhood (top 5) after filtering

	PostalCode	Borough	Neighborhood	PostalCode_lat	PostalCode_lng	Neigh_lat	Neigh_lng	Neigh_UTM_x	Neigh_UTM_y	Distance_from_centre	Neigh_venues_count
3	M5A	Downtown Toronto	Harbourfront	43.654260	-79.360636	43.640080	-79.380150	630653.432519	4.833173e+06	1544.342400	100
6	M7A	Downtown Toronto	Queen's Park	43.662301	-79.389494	43.659659	-79.390340	629789.309083	4.835332e+06	1025.123558	72
13	M4B	East York	Parkview Hill	43.706397	-79.309937	43.653482	-79.383935	630319.138862	4.834656e+06	296.875492	77
15	M5B	Downtown Toronto	Garden District	43.657162	-79.378937	43.656500	-79.377114	630862.602373	4.835002e+06	380.818539	65
16	M5B	Downtown Toronto	Ryerson	43.657162	-79.378937	43.658469	-79.378993	630706.797594	4.835218e+06	509.824145	100
18	M9B	Etobicoke	Princess Gardens	43.650943	-79.554724	43.640466	-79.391224	629759.369750	4.833199e+06	1740.068889	80

Toronto Map with the neighborhoods after filtering
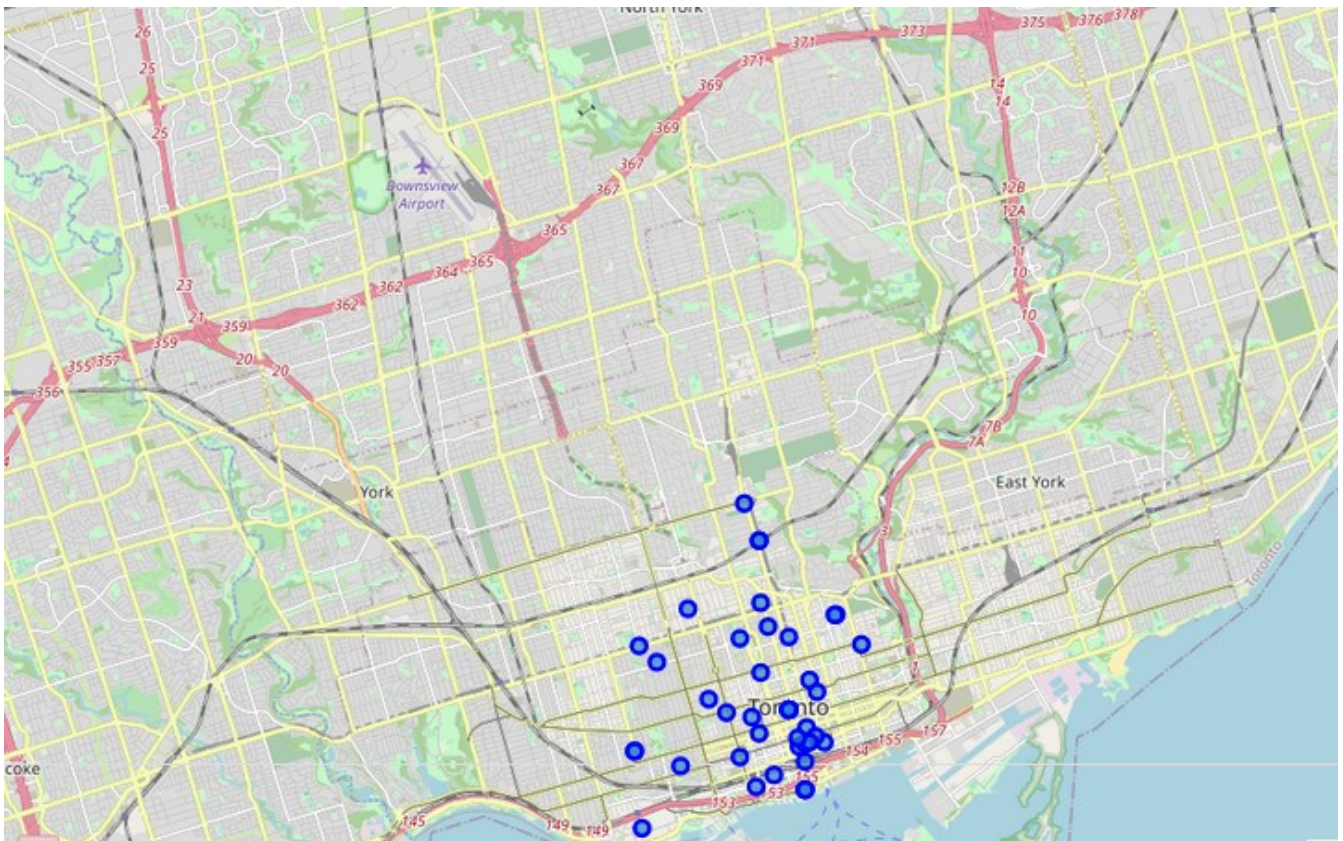
After the filters were applied remained about **40 neighborhoods** available and on these neighborhoods we created a clustering model using Kmeans.

The feature selected for Kmeans were the X,Y 2D Coordinate of each remaining neighborhood.