A Recommender System for expats looking for a family friendly neighbourhood in Downtown Toronto.

IBM - Final Capstone Project

Anushka Patil

28th July 2020

1.INTRODUCTION

1.1 Background knowledge

According to The Economist, Toronto is placed in the world's top ten "most liveable cities". This is one of the reasons why Toronto is full of non-Canadians(expats). Toronto is known to have build a knowledge-based economy by hiring skilled professionals from overseas. This has also led to an extremely diverse population combined with the multi-culturalism that stems from Toronto's location. According to the census of 2016, around 46.1% of residents in Toronto were born outside Canada.

1.2 Business Problem

Since Toronto is full of expats, it is implied that many families shift to Toronto every year and seek for a family-friendly neighbourhood to live in.

Family-friendly neighbourhoods are characterized by:

1. good entertainment venues nearby (theatres, spa, malls, bookstores etc.)
2. close to daily amenities(grocery store, gym, parlour, bank etc.)
3. having a good transport network (bus station, metro, airport etc.)
4. amenities for children (school, park, foodcourts, university etc.)
5. far from nightlife places (bars, pubs etc.)

There is an agent in the downtown borough of Toronto that helps expats find family-friendly neighbourhoods using this recommender system. This system finds neighbourhoods that have the maximum number of family friendly venues.

Features(or categories) that an expat requiers in their neighbourhood can be filtered out and customized as per the needs.

1.3 Target Audience

This project will interest expats moving to Downtown Toronto borough of Canada. It can help them identify a suitable location for leasing/buying their house.

2. DATA ACQUISITION AND CLEANING

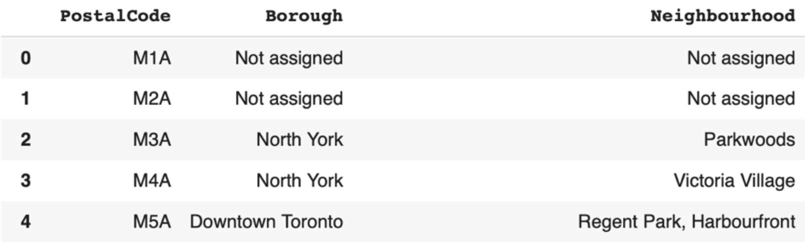
For this project, we will be needing three datasets:

1. Dataset that contains a list of all neighbourhoods, their boroughs and postal code of Canada.
2. Geospatial data that contains geo-locational information (latitudes and longitudes) about different boruoghs and their neighbouhoods.
3. Data obtained using 'Foursquare' to get details about different venues present in the neighbouhood of the specific borough (Downtown-in this project)

2.1 Dataset 1: Containing a list of boruoghs of Canada

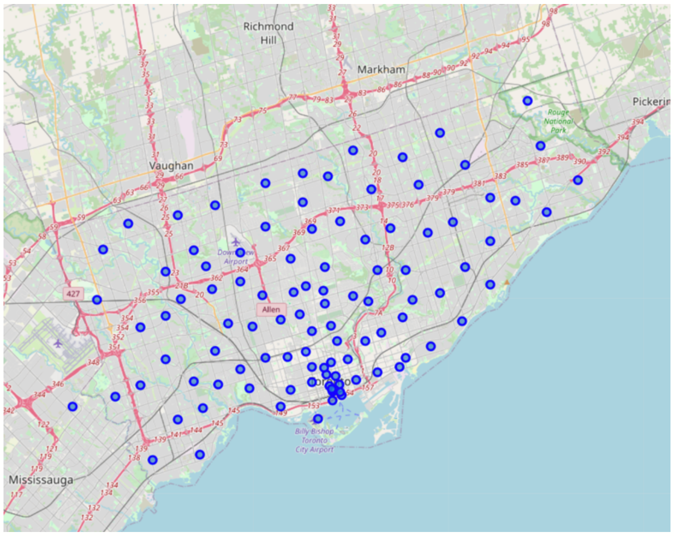
This data will be scrapped from a wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) using BeautifulSoup. Beautiful Soup is a package in Python which is used to extract text and numbers off of HTML and XML files, which are the more prevalent type of files in the internet. This dataset will contain the following feauters:

- Postal Code
- Borough
- Neighbouhood



	PostalCode	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront

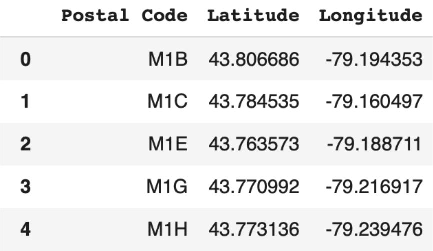
This dataset contains 180 rows. This is cleaned to remove the neighbouhoods not falling in any boruoghs and finally we are left with 103 rows. It is mainly used to identify all the neigbouhoods and their respective boruoghs of Canada. Postal Code will be used as an identifier key to merge this dataset with their geo-locational information.



2.2 Dataset 2: Containing latitudes and longitudes corresponding to different postal codes

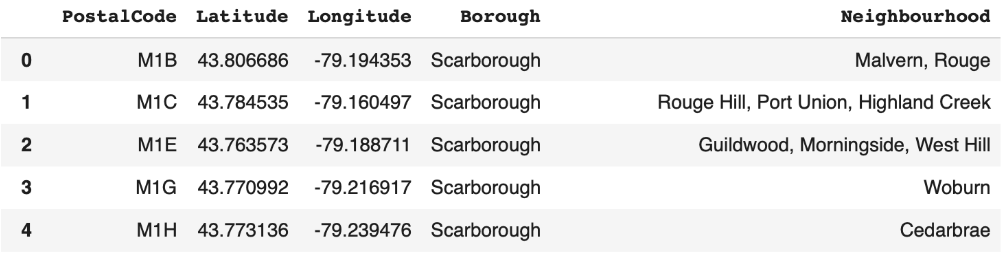
This dataset is obtained from https://http://cocl.us/Geospatial_data. This dataset contains the following features:

- Postal Code
- Latitude
- Longitude



	Postal Code	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

Now, this is merged with the first dataset to form a complete dataset, from which we can retirve the neighbourhoods of the particular borough the expat wants to live in. In this project, we have used Downotown Toronto as the desired borough.



	PostalCode	Latitude	Longitude	Borough	Neighbourhood
0	M1B	43.806686	-79.194353	Scarborough	Malvern, Rouge
1	M1C	43.784535	-79.160497	Scarborough	Rouge Hill, Port Union, Highland Creek
2	M1E	43.763573	-79.188711	Scarborough	Guildwood, Morningside, West Hill
3	M1G	43.770992	-79.216917	Scarborough	Woburn
4	M1H	43.773136	-79.239476	Scarborough	Cedarbrae

2.3 Dataset 3: Obtained using Foursquare

To proceed with the project, we need information and details about all the venues present in the neighbourhood of our required borough, which is Downtown Toronto. For this purpose, we will use the locational information provided by Foursquare (https://foursquare.com), a local search-and-discovery service application.

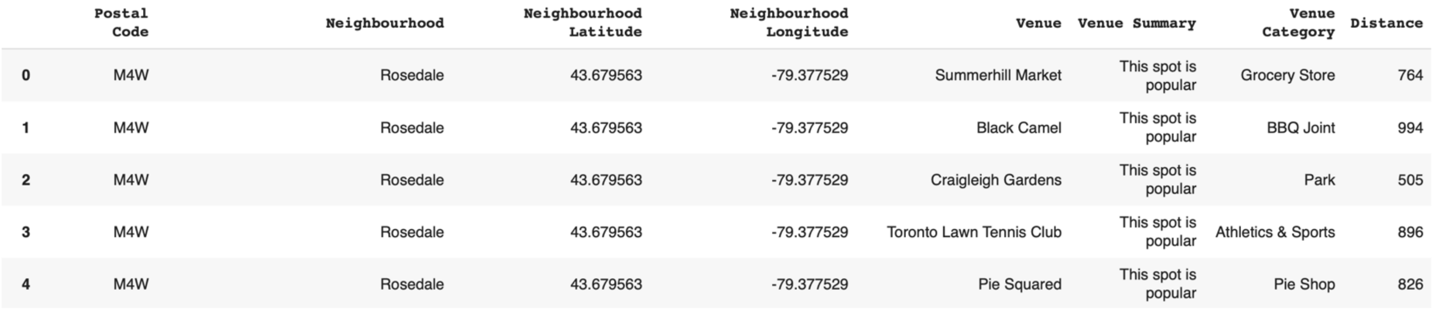
We need two types of information from foursquare:

1. **Basic**: latitude, longitude of the venue to correctly identify the neighbourhood
2. **Advance**: Category in which the venue falls (for ex: Grocery Store, Park, Theatre etc.)

We will use the 'explore' endpoint in foursquare to obtain a list of all the venues near to the location(latitude, longitude) of downtown toronto.

The data retrived contains the following information:

- Postal Code
- Neighbourhood
- Neighbourhood latitude
- Neighbourhood longitude
- Venue
- Venue Summary
- Venue Category
- Distance



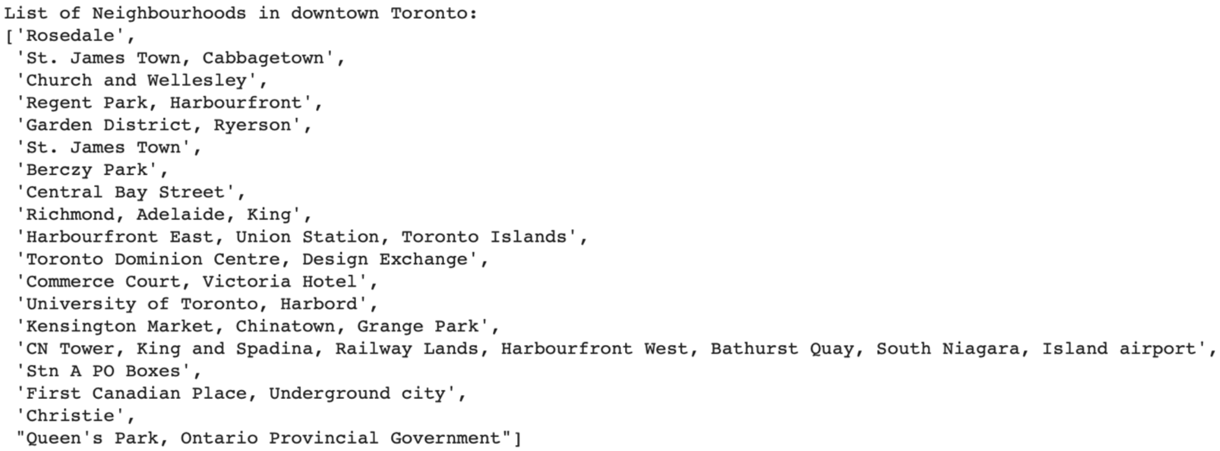
	Postal Code	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Summary	Venue Category	Distance
0	M4W	Rosedale	43.679563	-79.377529	Summerhill Market	This spot is popular	Grocery Store	764
1	M4W	Rosedale	43.679563	-79.377529	Black Camel	This spot is popular	BBQ Joint	994
2	M4W	Rosedale	43.679563	-79.377529	Craigleigh Gardens	This spot is popular	Park	505
3	M4W	Rosedale	43.679563	-79.377529	Toronto Lawn Tennis Club	This spot is popular	Athletics & Sports	896
4	M4W	Rosedale	43.679563	-79.377529	Pie Squared	This spot is popular	Pie Shop	826

The dataset was found to contain 203 unique categories. These categories were filtered out to be those desired in a family-friendly neighbourhood, and k-means clustering was performed to find the best set of neighbourhoods to live in.

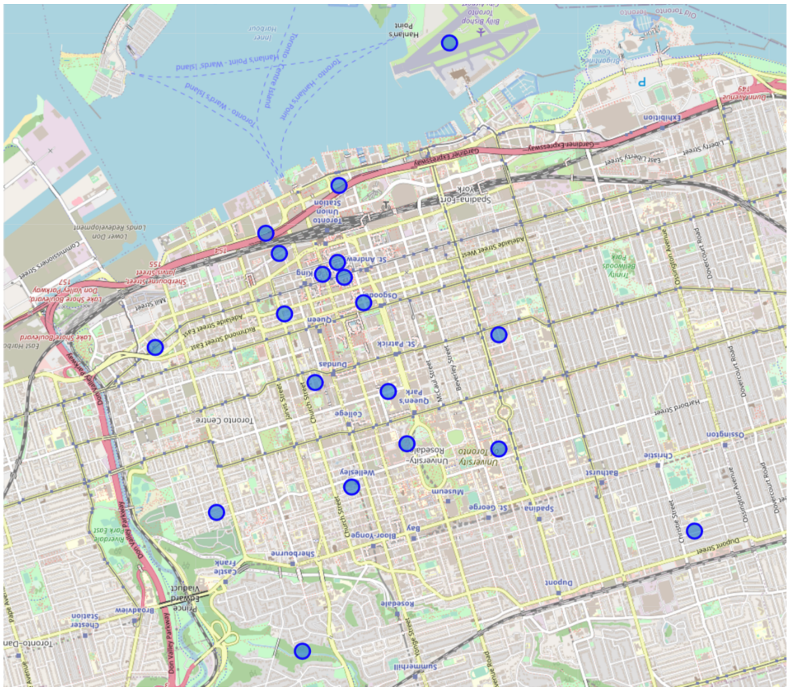
3. METHODOLOGY

3.1 Finding Neighborhoods of Downtown Toronto

Using the complete dataset created by combining the first two datasets, we will now extract the neighborhoods of just Downtown Toronto. 19 neighborhoods were identified and mapped using folium library.



```
List of Neighbourhoods in downtown Toronto:
['Rosedale',
 'St. James Town, Cabbagetown',
 'Church and Wellesley',
 'Regent Park, Harbourfront',
 'Garden District, Ryerson',
 'St. James Town',
 'Berczy Park',
 'Central Bay Street',
 'Richmond, Adelaide, King',
 'Harbourfront East, Union Station, Toronto Islands',
 'Toronto Dominion Centre, Design Exchange',
 'Commerce Court, Victoria Hotel',
 'University of Toronto, Harbord',
 'Kensington Market, Chinatown, Grange Park',
 'CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport',
 'Stn A PO Boxes',
 'First Canadian Place, Underground city',
 'Christie',
 "Queen's Park, Ontario Provincial Government"]
```



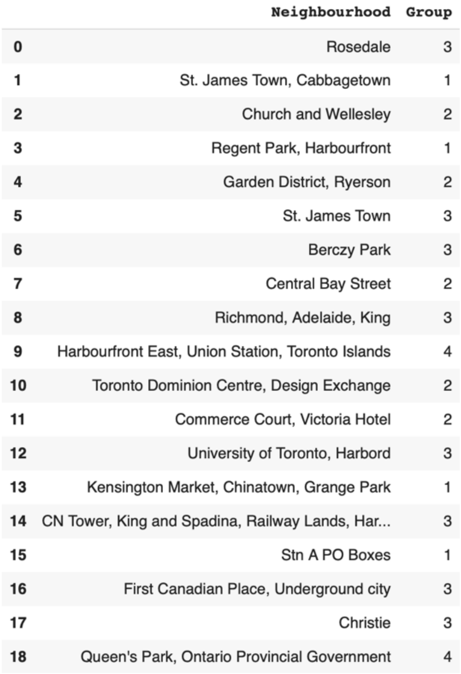
3.2 Extraction of unique Categories and One Hot Encoder

Using the third dataset obtained using FourSquare API, 200 unique categories were identified.
This category column is then one hot encoded. What one hot encoding does is, it takes a column which has categorical data, which has been label encoded and then splits the column into multiple columns each corresponding to a particular type of category. The numbers are replaced by 1s and 0s, depending on which column has what value. In our case, we’ll get 200 new columns, one for each unique category.
After this, a list is created manually to select categories considered relevant to a family-friendly neighbourhood. The categories are chosen based on the characteristics mentioned in the business problem section. The selected categories are as follows:

'Grocery Store','Park','Bank','Playground','Sandwich Place','Candy Store','Metro Station','Diner','Restaurant','Bakery','Farm','Pet Store','Gift Shop','Garden','Dance Studio','Pool','Theater','Performing Arts Venue','Bookstore','Salon / Barbershop', 'Arts & Crafts Store','Ice Cream Shop','Historic Site','Supermarket','Yoga Studio','Health & Beauty Service','Furniture / Home Store','Video Store','Comic Shop','Clothing Store','Shopping Mall','Cosmetics Shop','Gym','Dog Run','Museum', 'Farmers Market','Chocolate Shop','Dessert Shop','Spa','Gym / Fitness Center','Shoe Store','Event Space','Food Truck','Gym Pool','Electronics Store','Skating Rink', 'Pharmacy','Music Venue','Department Store','Monument / Landmark','Art Museum',’Poutine Place','Concert Hall','Church','Fountain','Tailor Shop','Basketball Stadium','Sporting Goods Shop','Beach','Lake','Train Station','University','Movie Theater','Aquarium','Baseball Stadium','Indie Movie Theater','Organic Grocery','Health Food Store','Music Store',’Women's Store’,'Food Court','Optical Shop','Airport','Harbor / Marina','Sculpture Garden','Cupcake Shop','Rock Climbing Spot'.

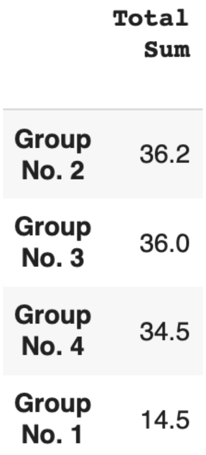
3.3 K-Means Clustering – A machine learning technique

K-means clustering is one of the simplest and a popular unsupervised machine learning algorithms. ‘k’ refers to the number of centroids and cluster refers to collection of the datapoints are aggregated together based on their similarities. In this project, k=4 and the datapoints are nothing but the neighborhoods. Since k=4, the 19 neighborhoods are divided into 4 groups and the result where as follows.



	Neighbourhood	Group
0	Rosedale	3
1	St. James Town, Cabbagetown	1
2	Church and Wellesley	2
3	Regent Park, Harbourfront	1
4	Garden District, Ryerson	2
5	St. James Town	3
6	Berczy Park	3
7	Central Bay Street	2
8	Richmond, Adelaide, King	3
9	Harbourfront East, Union Station, Toronto Islands	4
10	Toronto Dominion Centre, Design Exchange	2
11	Commerce Court, Victoria Hotel	2
12	University of Toronto, Harbord	3
13	Kensington Market, Chinatown, Grange Park	1
14	CN Tower, King and Spadina, Railway Lands, Har...	3
15	Stn A PO Boxes	1
16	First Canadian Place, Underground city	3
17	Christie	3
18	Queen's Park, Ontario Provincial Government	4

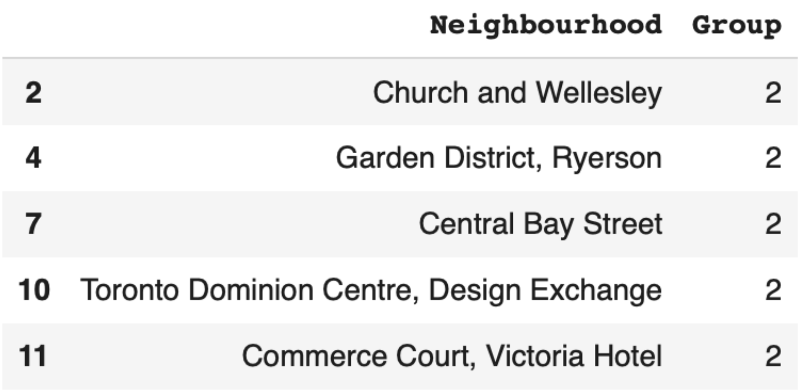
After performing k-means clustering, the groups are printed out in the order of having maximum sum. Here, maximum sum corresponds to the group of neighborhoods having maximum number of family friendly venues nearby.



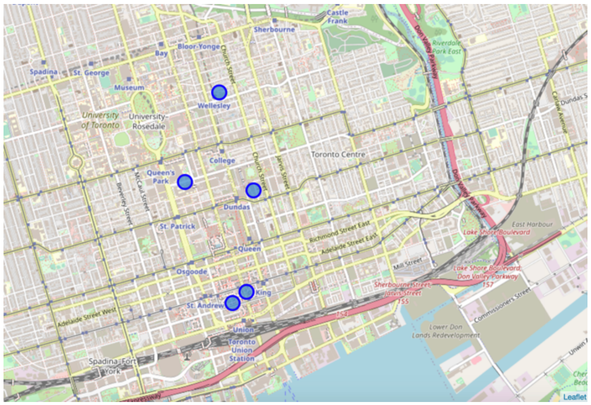
	Total Sum
Group No. 2	36.2
Group No. 3	36.0
Group No. 4	34.5
Group No. 1	14.5

4. CONCLUSION

Group 2 is the set of most family friendly neighborhoods in the Downtown borough of Canada.



	Neighbourhood	Group
2	Church and Wellesley	2
4	Garden District, Ryerson	2
7	Central Bay Street	2
10	Toronto Dominion Centre, Design Exchange	2
11	Commerce Court, Victoria Hotel	2



5. FUTURE SCOPE

In the future, the same project can be extended to accommodate a factor of safety using the crime dataset made available on Kaggle.