2. Data Acquisition and Cleaning

2.1 Data Acquisition

The data acquired for this project is a combination of data from three sources. The first data source of the project uses a List of postal codes of Canada: M that shows the neighbours per borough in Toronto. The dataset contains the following columns:

- **Post_Code** : Post Code for all regions in Toronto.
- **Borough** : Common name for London borough.
- **Neighbourhood** : All the neighbours in that Borough.

The second source of dataset is created from scratch using the list of neighbourhood available on the site Latitudes and Longitudes . This page contains additional information about the boroughs, the following are the columns:

- **Post_Code** : Post Code for all regions in Toronto.
- **Latitudes** : Latitudes of all regions of each Borough in Toronto.
- **Longitudes** : Longitudes of all regions of each Borough in Toronto.
- **Neighbourhood:** Name of the neighbourhood in the Borough.

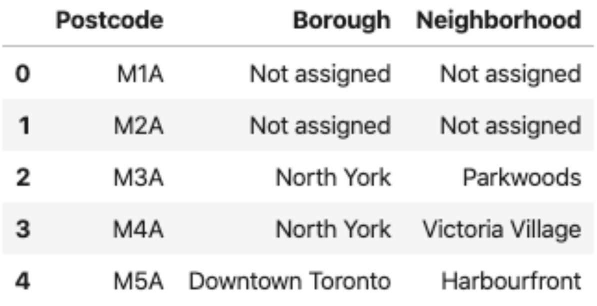
The third data source is the Foursquare API as found on the given link. This dataset is responsible for information of all neighbours latitude and longitude by requesting url using Foursquare API. This contains:

- **CLIENT_ID** = # your Foursquare ID
- **CLIENT_SECRET** =# your Foursquare Secret
- **VERSION** = # Foursquare API version

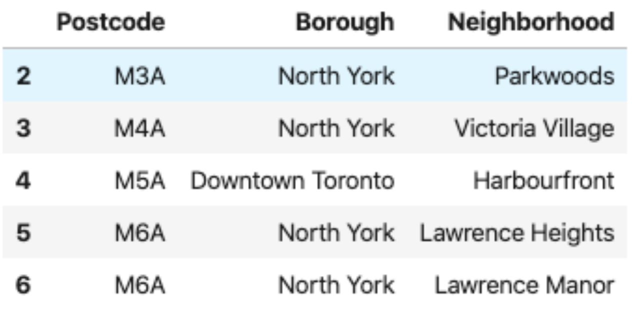
2.1 Data Cleaning

The data preparation for each of the three sources of data is done separately. From the Toronto data, the Borough post_code and their neighbourhood are present in our datasets.

The part A data is scraped directly from wikipedia which had ‘Not assigned’ values. After cleaning data of part A to part B we can see good form of dataset having no such stuffs.



	Postcode	Borough	Neighborhood
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	Harbourfront

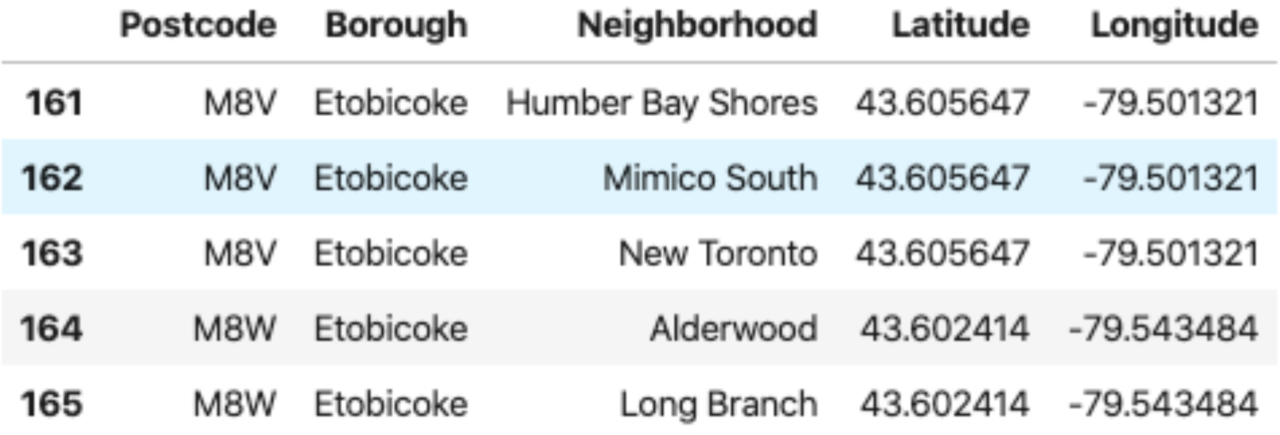


	Postcode	Borough	Neighborhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Harbourfront
5	M6A	North York	Lawrence Heights
6	M6A	North York	Lawrence Manor

part A **part B**

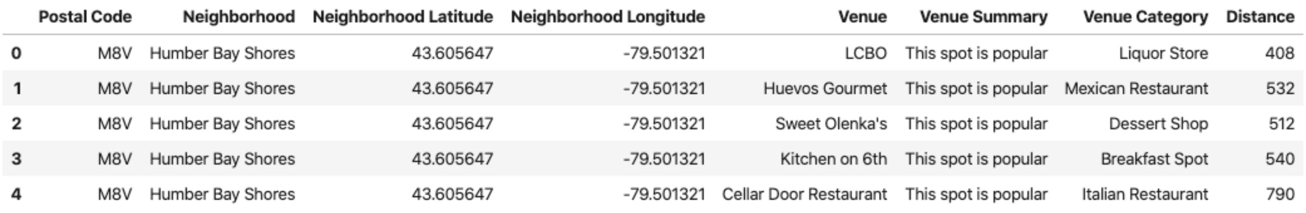
***Figure-2.1.1*:** Data from Wikipedia

Now we will use Geospatial_data to get Latitude and longitude of our neighbours on ‘postcode’ and finally we merge them to get a new dataframe as shown in below figure:



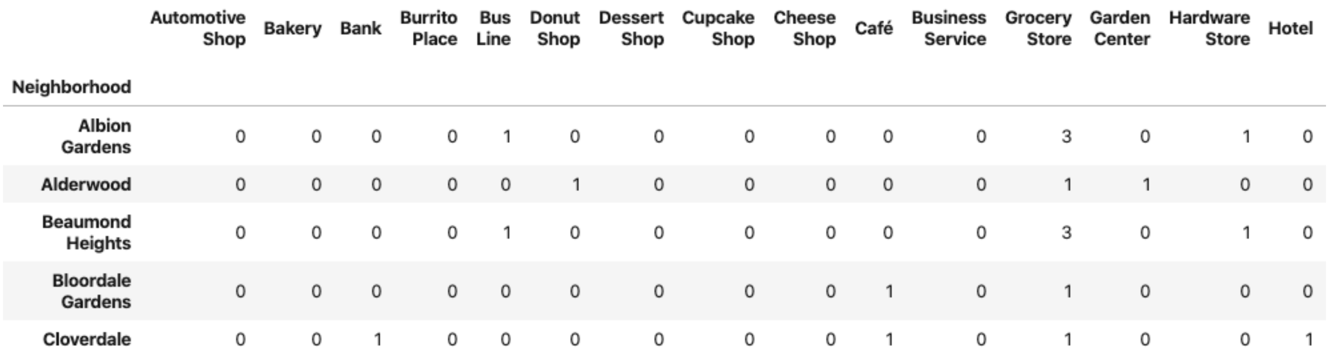
	Postcode	Borough	Neighborhood	Latitude	Longitude
161	M8V	Etobicoke	Humber Bay Shores	43.605647	-79.501321
162	M8V	Etobicoke	Mimico South	43.605647	-79.501321
163	M8V	Etobicoke	New Toronto	43.605647	-79.501321
164	M8W	Etobicoke	Alderwood	43.602414	-79.543484
165	M8W	Etobicoke	Long Branch	43.602414	-79.543484

We needed the venues and vanue_Category for manipulation to get the result and conclusion. For that we will use Foursquare API to collect all the relevant data to reach at conclusion of our problem'answer.



	Postal Code	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Summary	Venue Category	Distance
0	M8V	Humber Bay Shores	43.605647	-79.501321	LCBO	This spot is popular	Liquor Store	408
1	M8V	Humber Bay Shores	43.605647	-79.501321	Huevos Gourmet	This spot is popular	Mexican Restaurant	532
2	M8V	Humber Bay Shores	43.605647	-79.501321	Sweet Olenka's	This spot is popular	Dessert Shop	512
3	M8V	Humber Bay Shores	43.605647	-79.501321	Kitchen on 6th	This spot is popular	Breakfast Spot	540
4	M8V	Humber Bay Shores	43.605647	-79.501321	Cellar Door Restaurant	This spot is popular	Italian Restaurant	790

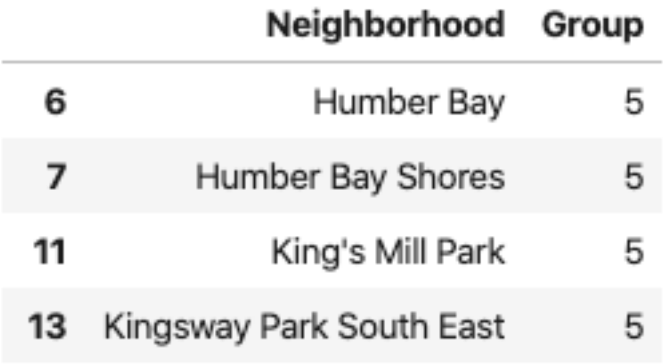
fig 2.1.2: Data using Foursquare API



	Automotive Shop	Bakery	Bank	Burrito Place	Bus Line	Donut Shop	Dessert Shop	Cupcake Shop	Cheese Shop	Café	Business Service	Grocery Store	Garden Center	Hardware Store	Hotel
Neighborhood															
Albion Gardens	0	0	0	0	1	0	0	0	0	0	0	3	0	1	0
Alderwood	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0
Beaumond Heights	0	0	0	0	1	0	0	0	0	0	0	3	0	1	0
Bloordale Gardens	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
Cloverdale	0	0	1	0	0	0	0	0	0	1	0	1	0	0	1

fig 2.1.2: Neighbour-category Data using Foursquare API

Now we will use **k-means** to cluster neighbours into k=5 clusters finally to get best result output for our result and conclusion. Here we are showing some of our best venue as output using this algorithm.



	Neighborhood	Group
6	Humber Bay	5
7	Humber Bay Shores	5
11	King's Mill Park	5
13	Kingsway Park South East	5

fig 2.1.2: k-means output