2 Data acquisition and cleaning

2.1 Data sources

The following Wikipedia page is used to get the information about neighbourhoods in Toronto: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. This defines the scope of this project which is the city of Toronto in Canada.

Also we use the following csv file to extract the geographical coordinates of different postal codes (neighborhoods): http://cocl.us/Geospatial_data. This is required to get the venue data and plot the map.

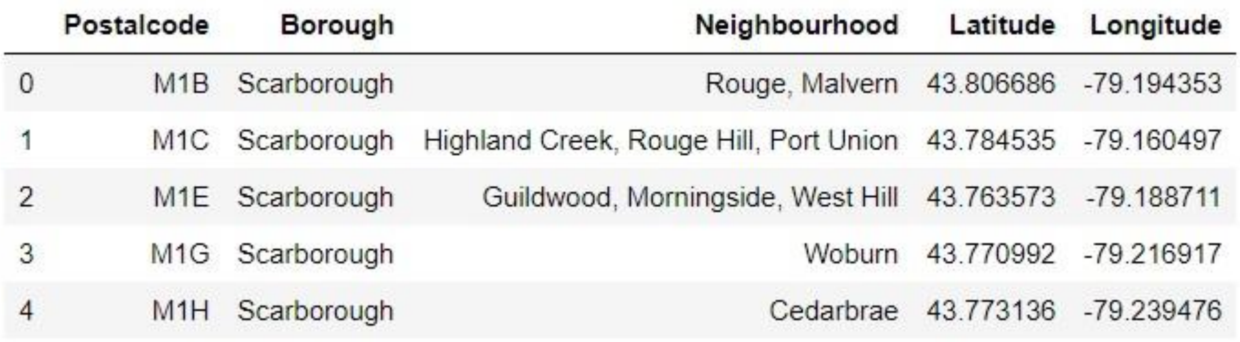
Finally, we request the venue data for each neighborhood from Foursquare API. This data is used to execute clustering on the neighborhoods.

2.2 Data cleaning

We combine the data downloaded from multiple sources into one table. After transforming the data into Pandas data frame, we ignore the rows with ‘Not assigned’ label in Borough column. Then we merge the neighborhoods with the same postal code. Finally, if a neighborhood has ‘Not assigned’ name, we consider the name of their borough as their neighbourhood’s name.

2.3 Feature Selection

After all the merging and cleaning data which we mentioned above, we consider postal code, borough, neighbourhood’s name, latitude, and longitude of each neighbourhood as shown in the following table (there is 103 rows and 5 columns). Note that in the methodology section, we will discuss about how to consider and inserting different events for each neighbourhood as a new data frame.



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