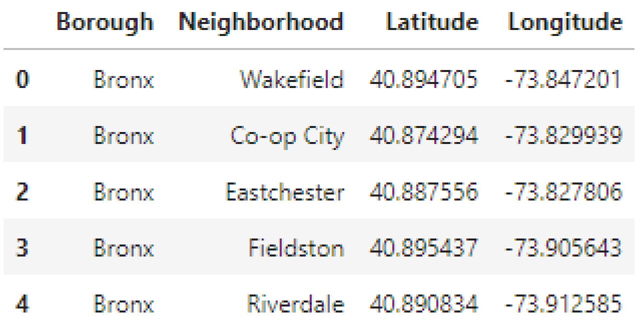
DATA

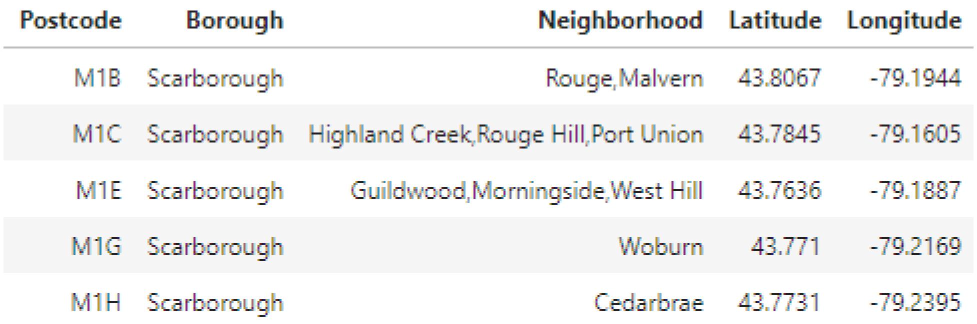
For this project, we are going to need two sets of data:

- First for new York Data we used a the same dataset we have used previously in the 3rd week lab in the course[1] which provided a json file that contains an already well organized data that didn't need much cleaning and looked eventually like this:



	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

- Then for the Toronto data we retrieved the boroughs and neighborhoods from a Wikipedia page[2] using web scraping techniques and then we used the geospatial data file that was provided in the 3rd week assignment in order to add the coordinate for each area and after some cleaning by removing all unassigned boroughs and grouping neighborhoods that belongs to the same postal code we came with the below table:



Postcode	Borough	Neighborhood	Latitude	Longitude
M1B	Scarborough	Rouge,Malvern	43.8067	-79.1944
M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.7845	-79.1605
M1E	Scarborough	Guildwood,Morningside,West Hill	43.7636	-79.1887
M1G	Scarborough	Woburn	43.771	-79.2169
M1H	Scarborough	Cedarbrae	43.7731	-79.2395

- And after that we did some formatting for the Toronto neighborhood data frame to match the same format of the New York Neighborhood data frame in order to concatenate them into one data frame that contains after adding a suffix to each Neighborhood name to indicate the country the neighborhood belongs to, this step can be very useful if we decided to expand the project in the future to include more cities

- Eventually we used foursquare API to find the venues surrounding each neighborhood as within 700 (m) and then we extracted the 7 most common venues around each neighbor hood

Note: all the values of the parameters that we have chosen so far and in later sections are chosen after some trial and error and can possibly be tuned to achieve a more accurate results