Data acquisition & preparation

1.1 Data source

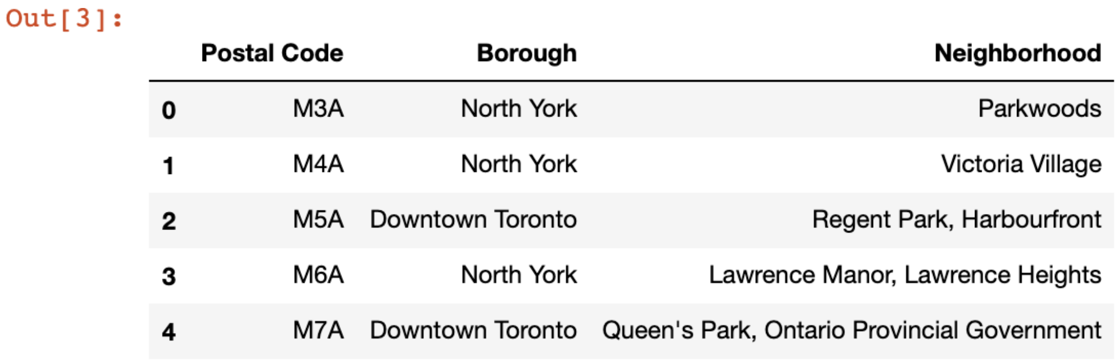
- “List of Postal code of Canada: M” (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) to get Postal code, borough & the name of all the neighborhoods about the neighborhoods in Toronto (use only assigned boroughs).
- “https://cocl.us/Geospatial_data.csv” file to get all the geographical coordinates of the neighborhoods.
- “Demographics of Toronto” (https://en.m.wikipedia.org/wiki/Demographics_of_Toronto#Ethnic_diversity) wiki page to get the demographic distribution of the population by ethnicity, which will be helpful in identifying the suitable neighborhood to open a new Chinese restaurant.
- Geographic data using Foursquare’s API explore-mode to fetch details about the venues in Toronto and with regard to venues' names, categories, and locations (latitude and longitude).

Using Foursquare API (https://developer.foursquare.com/docs), the following information is to be retrieved for each venue:

- Name: The name of the venue.
- Category: The category type as defined by the API.
- Latitude: The latitude value of the venue.
- Longitude: The longitude value of the venue.

1.2 Data acquisition & cleaning

a) Scraping Toronto Neighborhoods Table from Wikipedia\ Scraped the “List of Postal code of Canada: M” in order to obtain the data about the Toronto & the Neighborhoods.

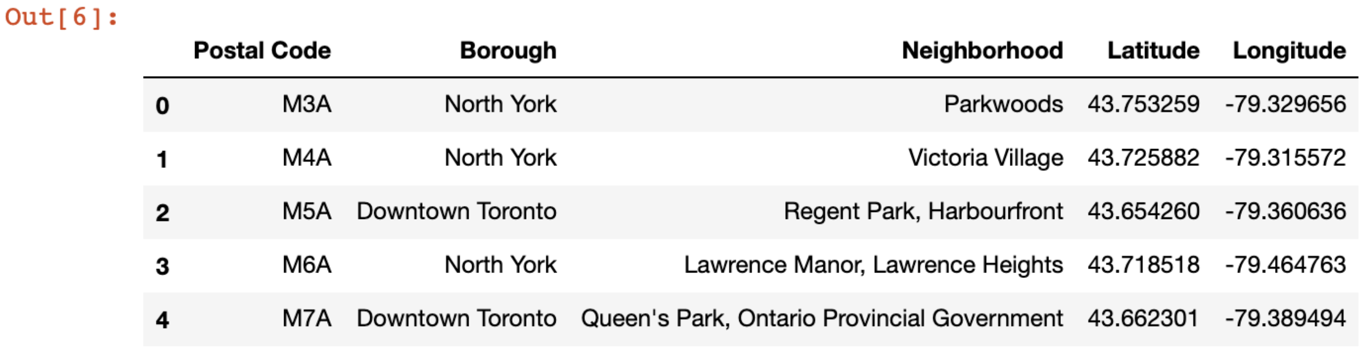

Out[3]:

	Postal Code	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

Assumptions made to attain the below DataFrame:

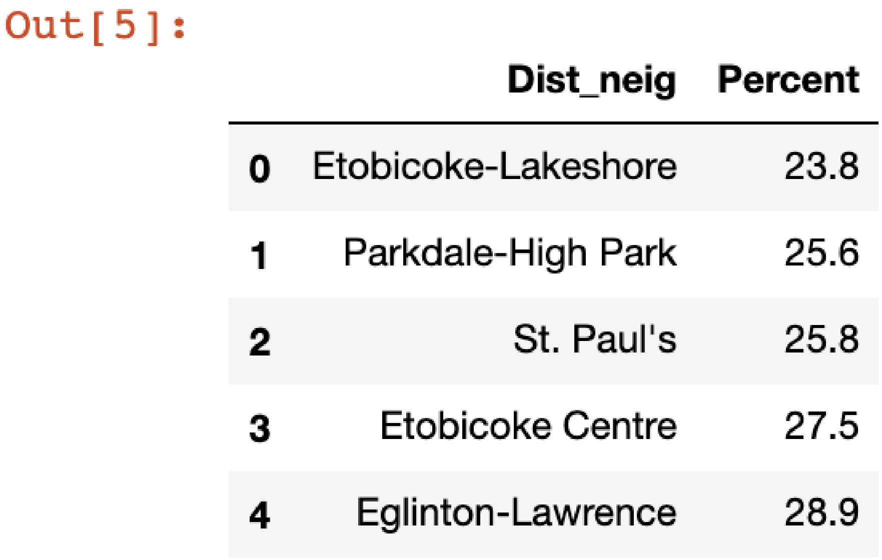
- Dataframe will consist of three columns: PostalCode, Borough, and Neighborhood
- Only the cells that have an assigned borough will be processed.
- More than one neighborhood can exist in one postal code area. For example, in the table on the Wikipedia page, you will notice that M5A is listed twice and has two neighborhoods: Harbourfront and Regent Park. These two rows will be combined into one row with the neighborhoods separated with a comma as shown in row 11 in the above table.
- If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough.

b) Adding geographical coordinates to the neighborhoods. Next important step is adding the geographical coordinates to these neighborhoods. To do so I'm going to be using the Geospatial Data CSV file provided above and combining it with the existing neighborhood data frame by merging them on the postal code column.


Out[6]:

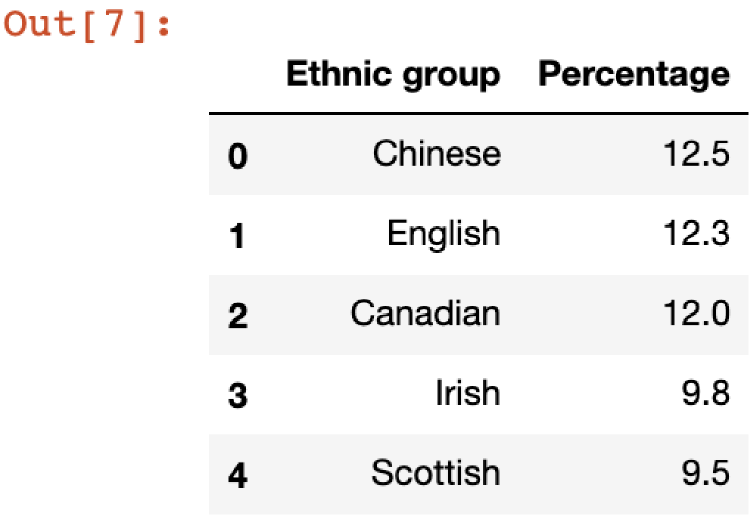
	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

c) Scrap the distribution of the population from Wikipedia. I scraped “Demographics of Toronto” to obtain the data about the Toronto & the neighborhoods in it. Compared to all the neighborhoods in Toronto below given neighborhoods only had a considerable amount of Chinese crowd. We are examining those neighborhood’s population to identify the densely populated neighborhoods with the Chinese population.



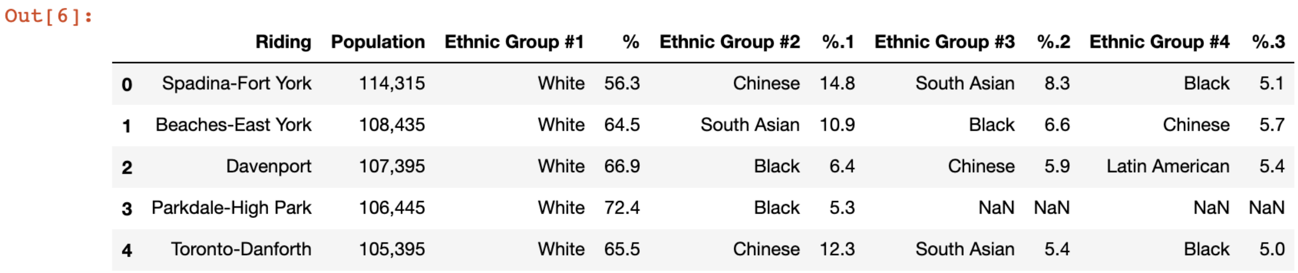
Out[5]:

	Dist_neig	Percent
0	Etobicoke-Lakeshore	23.8
1	Parkdale-High Park	25.6
2	St. Paul's	25.8
3	Etobicoke Centre	27.5
4	Eglinton-Lawrence	28.9



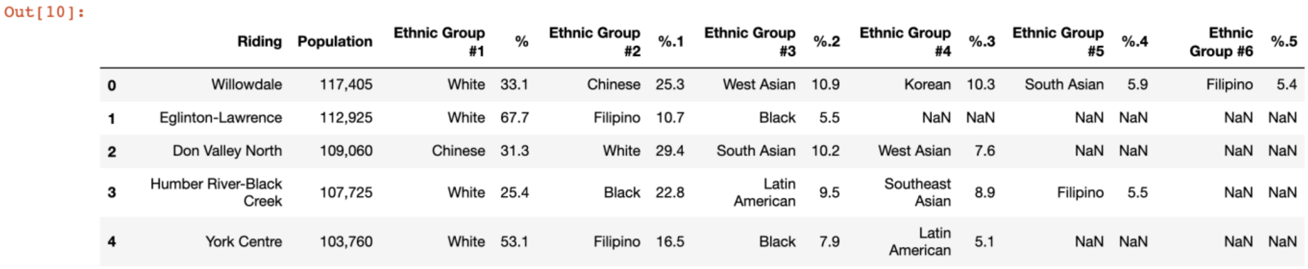
Out[7]:

	Ethnic group	Percentage
0	Chinese	12.5
1	English	12.3
2	Canadian	12.0
3	Irish	9.8
4	Scottish	9.5



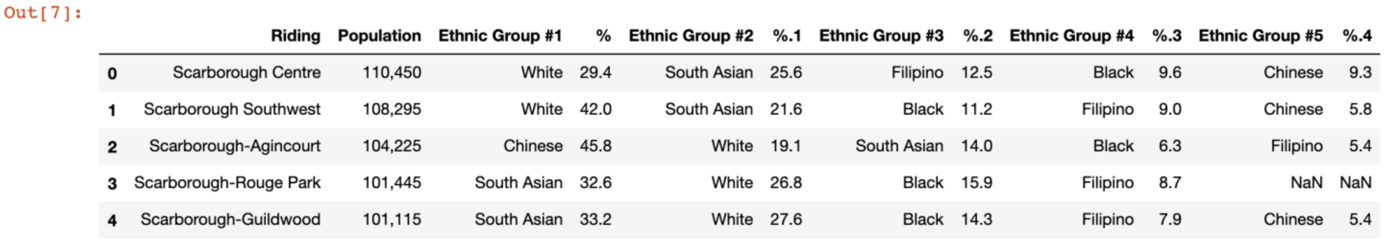
Out[6]:

	Riding	Population	Ethnic Group #1	%	Ethnic Group #2	%.1	Ethnic Group #3	%.2	Ethnic Group #4	%.3
0	Spadina-Fort York	114,315	White	56.3	Chinese	14.8	South Asian	8.3	Black	5.1
1	Beaches-East York	108,435	White	64.5	South Asian	10.9	Black	6.6	Chinese	5.7
2	Davenport	107,395	White	66.9	Black	6.4	Chinese	5.9	Latin American	5.4
3	Parkdale-High Park	106,445	White	72.4	Black	5.3	NaN	NaN	NaN	NaN
4	Toronto-Danforth	105,395	White	65.5	Chinese	12.3	South Asian	5.4	Black	5.0



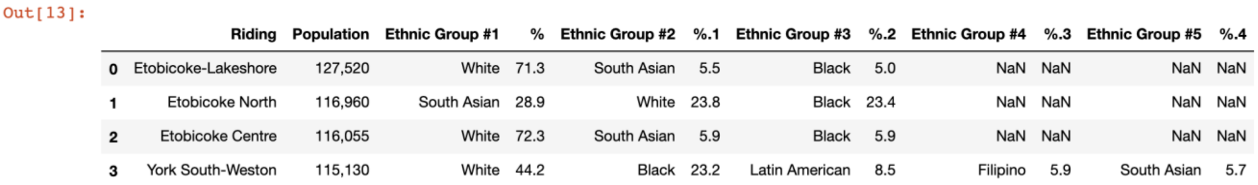
Out[10]:

	Riding	Population	Ethnic Group #1	%	Ethnic Group #2	%.1	Ethnic Group #3	%.2	Ethnic Group #4	%.3	Ethnic Group #5	%.4	Ethnic Group #6	%.5
0	Willowdale	117,405	White	33.1	Chinese	25.3	West Asian	10.9	Korean	10.3	South Asian	5.9	Filipino	5.4
1	Eglinton-Lawrence	112,925	White	67.7	Filipino	10.7	Black	5.5	NaN	NaN	NaN	NaN	NaN	NaN
2	Don Valley North	109,060	Chinese	31.3	White	29.4	South Asian	10.2	West Asian	7.6	NaN	NaN	NaN	NaN
3	Humber River-Black Creek	107,725	White	25.4	Black	22.8	Latin American	9.5	Southeast Asian	8.9	Filipino	5.5	NaN	NaN
4	York Centre	103,760	White	53.1	Filipino	16.5	Black	7.9	Latin American	5.1	NaN	NaN	NaN	NaN



Out[7]:

	Riding	Population	Ethnic Group #1	%	Ethnic Group #2	%.1	Ethnic Group #3	%.2	Ethnic Group #4	%.3	Ethnic Group #5	%.4
0	Scarborough Centre	110,450	White	29.4	South Asian	25.6	Filipino	12.5	Black	9.6	Chinese	9.3
1	Scarborough Southwest	108,295	White	42.0	South Asian	21.6	Black	11.2	Filipino	9.0	Chinese	5.8
2	Scarborough-Agincourt	104,225	Chinese	45.8	White	19.1	South Asian	14.0	Black	6.3	Filipino	5.4
3	Scarborough-Rouge Park	101,445	South Asian	32.6	White	26.8	Black	15.9	Filipino	8.7	NaN	NaN
4	Scarborough-Guildwood	101,115	South Asian	33.2	White	27.6	Black	14.3	Filipino	7.9	Chinese	5.4



Out[13]:

	Riding	Population	Ethnic Group #1	%	Ethnic Group #2	%.1	Ethnic Group #3	%.2	Ethnic Group #4	%.3	Ethnic Group #5	%.4
0	Etobicoke-Lakeshore	127,520	White	71.3	South Asian	5.5	Black	5.0	NaN	NaN	NaN	NaN
1	Etobicoke North	116,960	South Asian	28.9	White	23.8	Black	23.4	NaN	NaN	NaN	NaN
2	Etobicoke Centre	116,055	White	72.3	South Asian	5.9	Black	5.9	NaN	NaN	NaN	NaN
3	York South-Weston	115,130	White	44.2	Black	23.2	Latin American	8.5	Filipino	5.9	South Asian	5.7

d) Get location data using Foursquare. Foursquare API is a very useful online application used by many developers & other applications like Uber etc. In this project I have used it to retrieve information about the places present in the neighborhoods of Toronto. The API returns a JSON file and we need to turn that into a data-frame. Here I’ve chosen 100 popular spots for each neighborhood within a radius of 1.0 km.

Out[8]:

	Neighborhood	Asian Restaurant	Chinese Restaurant
22	Don Mills	0.074074	0.037037
48	Little Portugal, Trinity	0.069767	0.000000
78	The Annex, North Midtown, Yorkville	0.041667	0.000000
29	First Canadian Place, Underground city	0.030000	0.000000
63	Regent Park, Harbourfront	0.021277	0.000000
28	Fairview, Henry Farm, Oriole	0.015385	0.000000
72	St. James Town	0.013158	0.000000
64	Richmond, Adelaide, King	0.010753	0.000000
83	Toronto Dominion Centre, Design Exchange	0.010000	0.010000
18	Commerce Court, Victoria Hotel	0.010000	0.000000
3	Bayview Village	0.000000	0.250000
23	Dorset Park, Wexford Heights, Scarborough Town...	0.000000	0.166667
86	Westmount	0.000000	0.142857
74	Steeles West, L'Amoreaux West	0.000000	0.133333
16	Clarks Corners, Tam O'Shanter, Sullivan	0.000000	0.071429
73	St. James Town, Cabbagetown	0.000000	0.048780
55	North Toronto West	0.000000	0.047619
84	University of Toronto, Harbord	0.000000	0.029412
31	Garden District, Ryerson	0.000000	0.010000
35	Harbourfront East, Union Station, Toronto Islands	0.000000	0.010000