DATA ACQUISITION AND CLEANING

1. Data Acquisition

The data acquired for this project is a combination of data from two sources.

The first source of data is scraped from a Wikipedia page that contains the list of Toronto boroughs. This page contains additional information about the boroughs, the following are the columns:

- Postal Code: Postal code of Neighbourhoods
- Borough: Name of Borough
- Neighbourhood: Name of Neighbourhoods
- Link: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

The second data source of the project uses a Toronto latitude and longitude according to the Postal Code. The dataset contains the following columns:

- Postal Code: Postal code of Neighbourhoods
- Latitude: Latitude of Neighbourhoods
- Longitude: Longitude of Neighbourhood
- Link: http://cocl.us/Geospatial_data

2. Data Cleaning

The data preparation for each of the two sources of data is done separately. Neighbourhoods are merged according to their Postal Code. (See fig 2.1).

Out[112]:

	PostalCode	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
5	M9A	Etobicoke	Islington Avenue, Humber Valley Village
6	M1B	Scarborough	Malvern, Rouge
7	M3B	North York	Don Mills
8	M4B	East York	Parkview Hill, Woodbine Gardens
9	M5B	Downtown Toronto	Garden District, Ryerson

Fig 2.1 Borough with Postal Code

The second data is scraped from a Wikipedia page using the Beautiful Soup library in python. Using this library we can extract the data in the tabular format as shown in the website. After the web scraping, string manipulation is required to get the names of the boroughs in the correct form (see fig 2.2). This is important because we will be merging the two datasets together using Postal Code.

Out[117]:

	PostalCode	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
5	M9A	Etobicoke	Islington Avenue, Humber Valley Village	43.667856	-79.532242
6	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
7	M3B	North York	Don Mills	43.745906	-79.352188
8	M4B	East York	Parkview Hill, Woodbine Gardens	43.706397	-79.309937
9	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937

Fig 2.2 List of Toronto Boroughs

The two datasets are merged on the Postal Code to form a new dataset that combines the necessary information in one dataset (see fig 2.3). The purpose of this dataset is to select the neighbourhoods only in Toronto.

Out[119]:

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
1	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
2	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937
3	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418
4	M4E	East Toronto	The Beaches	43.676357	-79.293031
5	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306
6	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383
7	M6G	Downtown Toronto	Christie	43.669542	-79.422564
8	M5H	Downtown Toronto	Richmond, Adelaide, King	43.650571	-79.384568
9	M6H	West Toronto	Dufferin, Dovercourt Village	43.669005	-79.442259

Fig 2.3 Neighbourhoods in Toronto

As dataframe is sorted only for Toronto, now we can explore Toronto. Folium is used visualise neighbourhood of Toronto on map. (see fig 2.4).

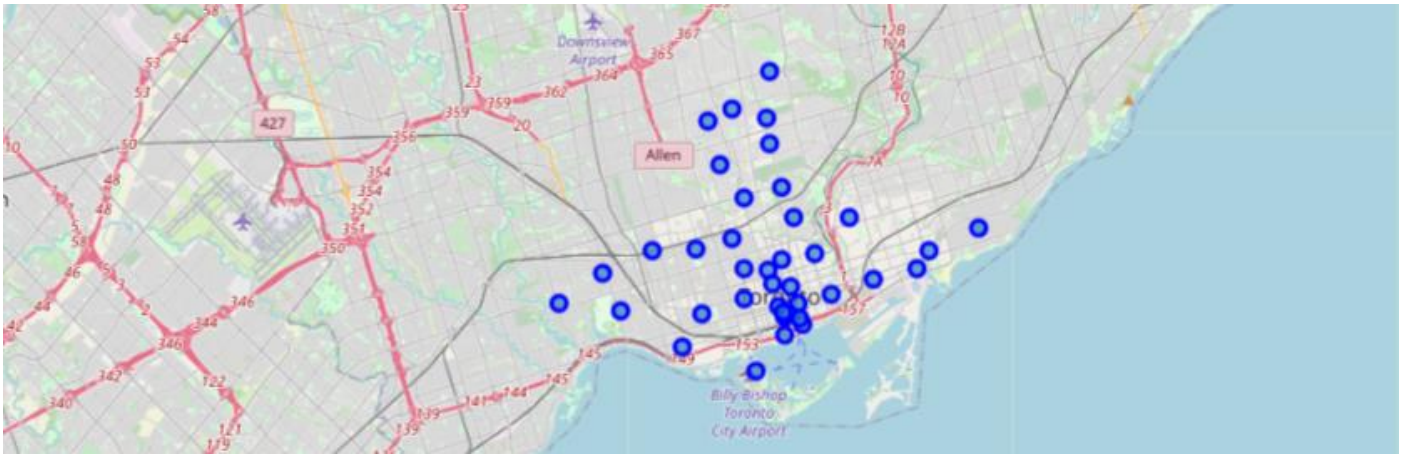

Fig 2.4 Map of neighbourhoods of Toronto

Foursquare API is used get venues and venue category of all neighbourhoods of Toronto. Two new columns with venue and its category is added. (See Fig 2.5).

Out[159]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
2	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center
3	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa
4	Regent Park, Harbourfront	43.65426	-79.360636	Impact Kitchen	43.656369	-79.356980	Restaurant
5	Regent Park, Harbourfront	43.65426	-79.360636	Corktown Common	43.655618	-79.356211	Park
6	Regent Park, Harbourfront	43.65426	-79.360636	The Distillery Historic District	43.650244	-79.359323	Historic Site
7	Regent Park, Harbourfront	43.65426	-79.360636	Morning Glory Cafe	43.653947	-79.361149	Breakfast Spot
8	Regent Park, Harbourfront	43.65426	-79.360636	The Extension Room	43.653313	-79.359725	Gym / Fitness Center
9	Regent Park, Harbourfront	43.65426	-79.360636	Dominion Pub and Kitchen	43.656919	-79.358967	Pub

Fig 2.5 Neighbourhoods with it venues

Using above dataframe, venues category is sorted which are present in high number. It is done to obtain the trend of venue category in these neighbourhood. (See Fig 2.6).

Out[131]:

Venue Category	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
Coffee Shop	159	159	159	159	159	159
Café	89	89	89	89	89	89
Restaurant	55	55	55	55	55	55
Hotel	37	37	37	37	37	37
Italian Restaurant	37	37	37	37	37	37
Park	35	35	35	35	35	35
Japanese Restaurant	32	32	32	32	32	32
Bakery	30	30	30	30	30	30
Pizza Place	29	29	29	29	29	29
Bar	28	28	28	28	28	28

Fig 2.6 Venue category with maximum number

Now new dataset is formed with only restaurant and neighbourhood. It is sorted according to neighbourhood with maximum number of restaurant. Along with these type of restaurant which are in high number in each neighbourhood is obtained. (See Fig 2.7)

Out[133]:

Neighborhood	Restaurant	Japanese Restaurant	Seafood Restaurant	American Restaurant	Italian Restaurant	Sushi Restaurant	Asian Restaurant	Total
Commerce Court, Victoria Hotel	7	2	3	4	2	0	2	20
First Canadian Place, Underground city	4	4	3	3	1	2	3	20
Toronto Dominion Centre, Design Exchange	4	3	3	3	2	2	2	19
Richmond, Adelaide, King	4	1	1	2	0	2	1	11
Harbourfront East, Union Station, Toronto Islands	3	1	1	0	3	1	0	9
Garden District, Ryerson	1	3	1	0	2	1	0	8

Fig 2.7 Neighbourhoods with maximum number of restaurant

The new dataset is used to generate the 10 most common venues for each neighbourhood using the Foursquare API, finally using k means clustering algorithm to cluster similar neighbourhoods together.