1. **Data Acquisition and Cleaning**

* 1. **Data Source**

The data source used for this project is web source data, where the postal codes for boroughs in the city of Toronto and their respective neighborhoods in the province of Ontario Canada, were extracted using the python built-in library BeautifulSoup. These data were found on [Wikipedia](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M), with the use of the python library BeautifulSoup, which enables the data to be downloaded into a python data frame that was used for further analysis.

* 1. **Data Wrangling and Cleaning**

Data downloaded using the BeautifulSoup library from the web have lots of special characters and unassigned values. The BeautifulSoup enables the downloaded data to be grouped in columns. The columns were split into postal codes and Boroughs, with the boroughs and neighborhoods on the same column. The first approach was to write a python script to separate the boroughs from the neighborhoods such that they are different columns. Postal that are not assigned to the borough and neighborhood are dropped from the data frame.

Further data cleaning was carried out because it was observed from the data that there are boroughs with multiple neighborhoods, which were separated by special characters. A comma character was used to separate multiple neighborhoods while other characters were dropped.

* 1. **Data Description**

Before cleaning the data there were 180 indexes and 2 columns, after cleaning there are 103 indexes and 3 columns. For the sake of analysis, the location of each neighborhood will be needed to provide further information about venues that are close to each neighborhood. A python library named GEOCODER was used to download the location latitude and longitudinal distances for each neighborhood, this will be discussed in the methodology.

1. **Methodology** 
   1. **Exploratory Analysis**

The data extracted from the web and saved in a python data frame contains the postal code, borough, neighborhood, and the latitude and longitude distance.

**GEOCODER**

The geocoder is a python library that estimates the geographical coordinate of each neighborhood on the map. Considering the large data of neighborhoods for Toronto, the first 12 geographical coordinates values of the neighborhoods (Table 1) were printed, the latitudinal values are seen to be positive values while the longitudinal values are in negative values depending on their location on the map. For this project, we decided to consider the borough that contains a high number of neighborhoods to be used for further analysis (Figure 1). There are three boroughs with high neighborhoods which are; North York, Downtown Toronto, and Scarborough. These three boroughs will be compared to their nearby venues using Foursquare API that will be discussed later, to further assist in choosing the best location to start a business for SMEs.



Table 1: Geographical coordinates (Latitutde and Longitude) for the first 12 neighborhoods using Geocoder

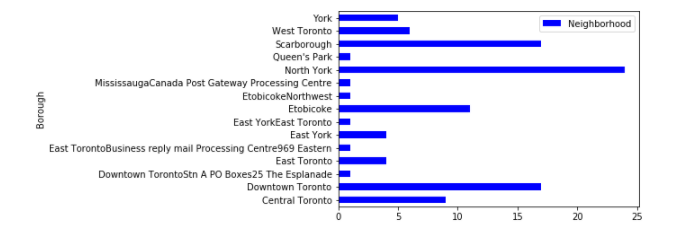


Figure 1: Neighborhood for each Borough

**FOLIUM**

Folium is a python library used to visualize the geographical details for each borough in Toronto city. The details display the borough and neighborhood with each point containing various locations on the map (figure 2).

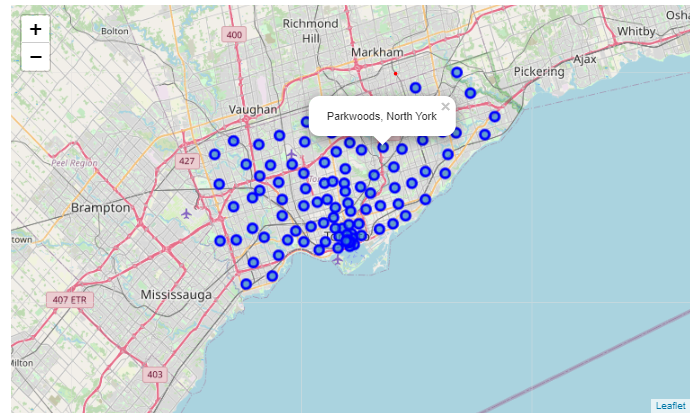


Figure 2: Location of boroughs and their Neighborhoods in Toronto city

The boroughs were extracted individually from the whole Toronto borough to visualize their locations on the map using folium. Figures 3, 4, and 5 show the Neighborhood map of Downtown Toronto, North York, and Scarborough respectively.

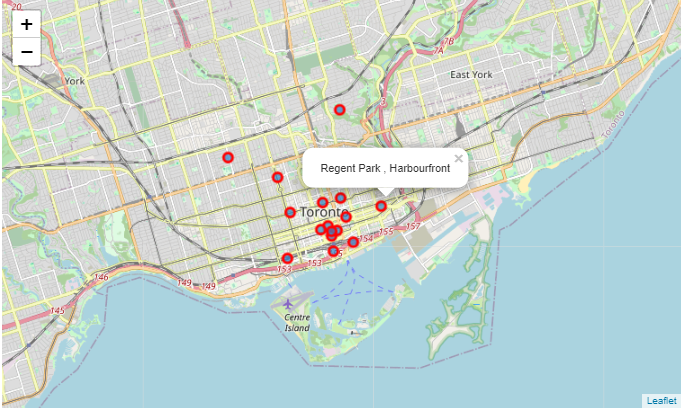


Figure 3: location of Downtown Toronto Neighborhoods

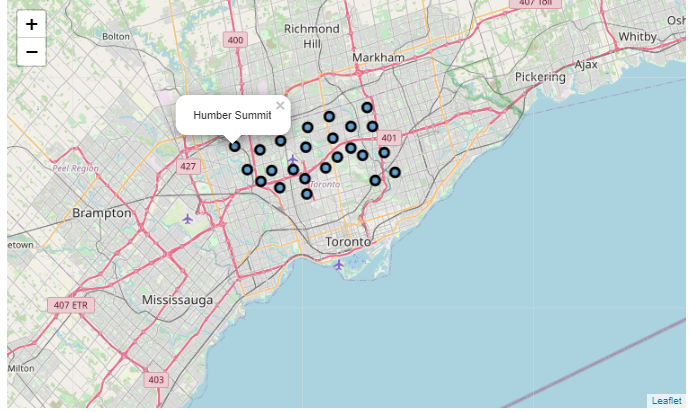


Figure 4: Location of North York Neighbrhoods

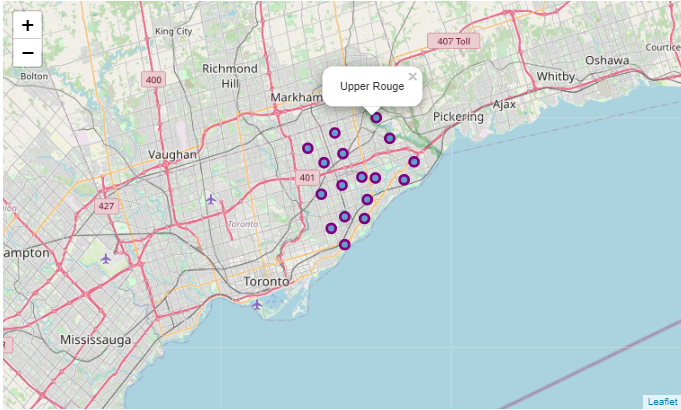


Figure 5: Location of Scarborough Neigborhoods

After visualizing the three different boroughs, I further explore the data by getting the first 100 nearby businesses venues that are within 700 meters radius. To achieve this task I used a foursquare web application programmable interface (API).

* 1. **Foursquare API**

To use the foursquare API, an account was created which makes it possible to generate a client ID and SECRET. These credentials were used to generate an access code that enables the connection of the geographical coordinates of neighborhoods to generate their nearby business venues. The results of the venues are in JSON format, to print the result into a data frame a function was written that shows the company's name, categories, and coordinates (Table 2,3, &4).

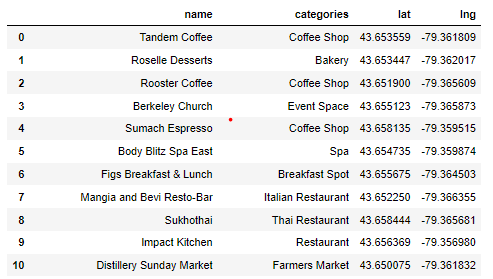


Table 2: First 10 result for Downtown Toronto nearby venues with total 88 venues

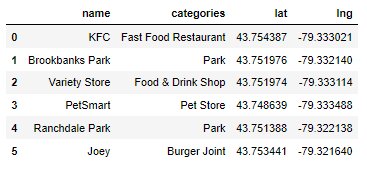


Table 3: North York Nearby venues with total 6 venues

Table 4: Nearby Venues for Scarborough with total of 6 venues



The nearby venues returned from each borough show that Downtown Toronto has more venues compared to North York and Scarborough. Further analysis was carried out on Downtown Toronto using foursquare to see the most common venue and the clusters using the unsupervised K-Means algorithm.