**2. Data acquisition and cleaning**

Here we will see what data we need, what sources we got the data from and how data was extracted and cleaned and made ready for analysis**.**

**2.1 Data Required**

To solve the problem, we will need the following data:

* List of neighborhoods in Toronto.
* Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
* Venue data, particularly data related to areas with restaurants and shopping places.

**2.2 Data sources**

The Wikipedia page (https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M) contains a list of neighborhoods in and around Toronto, along with its corresponding boroughs and postal codes. We will use the web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and ‘beautifulsoup’packages. Then we will get the geographical coordinates of the neighborhoods using https://cocl.us/Geospatial\_data csv file which will give us the latitude and longitude coordinates of the neighborhoods.

After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Restaurants and shops category in order to help us to solve the business problem put forward.

**2.3 Data extraction and cleaning**

Web scraping of wiki page is done using the BeautifulSoup package and dataframe of three columns: PostalCode, Borough, and Neighborhood is created.

We then only process the cells that have an assigned borough. Ignore cells with a borough that is ‘Notassigned**’.**

Since more than one neighborhood can exist in one postal code area we do a groupby of Postal code and Borough and group neighborhoods in same borough.

If a neighborhood is ‘Not assigned’ neighborhood, then the neighborhood will be the same as the borough.

We will then get the geographical coordinates of the neighborhoods using https://cocl.us/Geospatial\_data csv file which will give us the latitude and longitude coordinates of the neighborhoods. Which we will then merge with our original dataframe based on postalcode.

Additionaly we can get our venue data using Foursquare API.

First the data related to shopping venues in each neighborhood is extracted using foursquare to check the neighborhoods with large scope for shopping.

Next, we analyzed each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. Since we are analyzing the “Restaurants” data, we will filter the “Restaurants” as venue category for the neighbourhoods.

On this we will run kmean cluster to get restaurant concentration in each neighborhood..