Introduction:

Toronto is the provincial capital of Ontario and the most populous city in Canada, with a population of 2,954,024 as of July 2018.Generally most of the people like to spend time with friends by hanging out at place, and few people like to meet new people to discuss their business ideas and some informal meetings. The ideal place for this type of gatherings will be a coffee shop.

Business Problem:

The objective of this capstone project is to select the best location for opening a coffee shop in Toronto. Using the data science methodology and machine learning techniques like KMeans Clustering this project aims to answer the business problem. If a person is looking to open a new coffee shop in Toronto, where will you recommend him to open a coffee shop?

Data:

For solving this business problem we need the following data:

- List of neighbourhoods in the city of Toronto.
- Latitude and Longitude of those neighbourhoods in order to plot the map of those neighbourhoods and also to explore the venues.
- Venue data mainly pertaining to coffee shops is required to perform the clustering on neighbourhoods.

Sources of Data:

The Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) contains a list of neighbourhoods in Toronto, with a total of 203 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page with the help of python. Then we will get the geographical coordinates from the file 'Geospatial_Coordinates.csv'.

After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database 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 Coffee shop category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.