Capstone Project - The Battle of Neighbourhoods

1. Introduction:

1.1 Background

In the competitive world, it is of utmost importance to know the surroundings and take necessary steps leading to success of a venture. A similar case exists among food industry where a large number of start-ups emerge and may succeed or fail depending on lot of factors.

Such is the case with Cafés, which are currently popular among youths and profitable business foe entrepreneurs. Yet, a descent location and favourable surroundings contribute heavily to the success.

1.2 Problem

Data acquisition of geo-location using foursquare and boroughs data o Wikipedia page to analyse and cluster the existing cafés and acquiring favourable locations to start-up a café.

This project aims to analyse the existing geo-locations of the cafés in Toronto city and clustering the cafés using tools like Python, Jupyter Notebook, Foursquare data and Machine learning algorithms.

1.3 Interest

The entrepreneurs who are willing to venture a café at Toronto City in exploring the geo-locations of existing cafés and acquiring a profitable location.

2. Data Acquisition and Cleaning

2.1 Data Acquisition

The data acquired for this project is a combination of two sources. The first data source being Wikipedia data on boroughs of Ontario and second being Foursquare data.
Wikipedia link to data : Link
Foursquare: Link

2.2 Data Cleaning:

The dataset of boroughs of Ontario can be scraped from Wikipedia page using pandas or beautiful soap library in python. The following is the screenshot of the dataset on boroughs.

	Postal Code	Borough	Neighborhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
5	M6A	North York	Lawrence Manor, Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

Later the geocoder data of longitude and latitudes is acquired for the given location of postal code.

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Merging the datasets to get the location data of the specific borough, results in the following dataset-

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790
1	M5N	Central Toronto	Roselawn	43.711695	-79.416936
2	M4P	Central Toronto	Davisville North	43.712751	-79.390197
3	M5P	Central Toronto	Forest Hill North & West, Forest Hill Road Park	43.696948	-79.411307
4	M4R	Central Toronto	North Toronto West, Lawrence Park	43.715383	-79.405678

The data is later on used to acquire the foursquare data on venues around the region. Thus further leading to clustering of data to analyse the region and acquiring potential geo-location.