Recommending the best neighbourhoods for opening a coffee shop in Toronto City.

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?

Target Audience of this project:

This project is particularly useful for budding entrepreneurs who would like to start a new coffee shop outlet. Canadians consume more coffee than tap water, according to the Coffee Association of Canada (CAC). There are several thousand coffee franchise locations throughout Canada artfully brewing the Canadian drink of choice. The Canadian coffee industry is a $6.2 billion industry. Coffee creates an estimated 160,000 jobs in Canadian cafes and coffee shops every year. Retail sales of coffee are forecast to grow from $2.6 billion in 2017 to $3.8 billion by 2021, according to Euro monitor's Coffee in Canada report. As part of the quick-service restaurant (QSR) franchise industry, coffee franchises have great appeal to investors and consumers alike. Many Canadians frequent coffee establishments on their way to work in the morning or as a quick pick-me-up during the day. Coffee franchises are indeed part of a profitable industry.

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 208 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.

Methodology

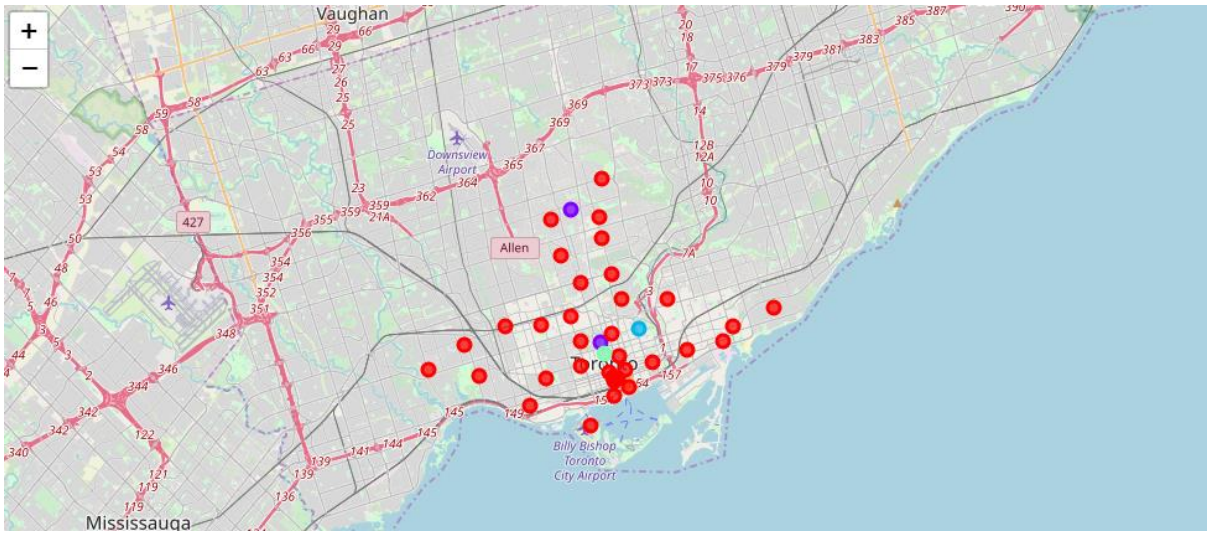
Methodology Firstly, we need to get the list of neighbourhoods in the city of Toronto. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). We will do web scraping using read_html from pandas.io.html library using python to extract the list of neighbourhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use 'Geospatial_Coordinates.csv' file to get the latitude and longitude details of the neighbourhood. After gathering the data, we will populate the data into a pandas DataFrame.Then we will extract the data of Toronto city and then visualize the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned are correctly plotted in the city of Toronto. Next, we will use Foursquare API to get the top 100 venues that are within a radius of 500 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analysing the "Coffee Shop" data, we will filter the "Coffee Shop" as venue category for the neighbourhoods. Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 5 clusters based on their frequency of occurrence for "Coffee Shop". The results will allow us to

identify which neighbourhoods have higher concentration of Coffee Shops while which neighbourhoods have fewer number of Coffee Shops. Based on the occurrence of Coffee Shops in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new Coffee Shops.

Results:

The results from the k-means clustering show that we can categorize the neighbourhood clusters based on the frequency of occurrence for 'Coffee Shop':

- Cluster 0: Neighbourhoods with moderate number of coffee Shops.
- Cluster 1,2,3,4: Neighbourhoods with no existence of Coffee Shops.



Discussion:

Most of the coffee shops are concentrated in Downtown area of Toronto, with moderate number in cluster0 and there are no coffee shops in clusters 1, 2, 3, 4.So this clusters represents a great opportunity and are high potential areas to open a new coffee shop outlet as there is no competition from existing coffee shops. Meanwhile the coffee shops in cluster0 are facing moderate competition from the existing coffee shops. Therefore this project recommends the budding entrepreneurs to open coffee shops in the clusters 1,2,3,4 where there is a high chance of growing business as there are no coffee shops. The entrepreneurs are not advised not to start in cluster0 as it has moderate concentration of coffee shops and

the competition will be there from the existing shops. Entrepreneurs can even setup in cluster0 with some innovative thoughts and can try to sustain the competition.

Conclusion:

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 5 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. budding entrepreneurs and investors regarding the best locations to open a new coffee shop. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 0 are the most preferred locations to open a new coffee shop. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new coffee shop.