Problem Description

In this project we will be analyzing the data of from *foursquare* and the *postal code* data to determine which places have major number of visitors. This project will be beneficial for any individual who is looking forward to opening a restaurant or cafe around Canada area that is in the proximity of the postal codes been used to prepare this project. Therefore, this project will analyze the data and determine which postal code areas are high attractions.

All in all, this project will be an asset for an entrepreneur (to determine which place will attract more customer),tourist and also to law enforcement officers in certain cases to tighten the security in most popular areas.

Data Usage

The data that will be used in this project will mainly be collected from two major sources for the analysis.

1. **Wikipedia postal code:**

The postal code that will be used for the analysis of the certain areas of Canada will be retrieved from the free source (https://en.wikipedia.org/wiki/Postal_codes_in_Canada#Table_of_all_postal_codes).This link has all the postal code related information for Canada, which will play a vital role in analysis of data from foursquare.

2. **Foursquare:**

The data retrieved from foursquare will be used to analyze the information of different venues and events that is happening within the proximity of the postal code data retrieved from Wikipedia. In addition, the location data will be used to segment, target and determine the presence in each postal code vicinity.

The data will be mainly processed through clusters, to segment on the basis of the neighborhood. In this analysis we will build k-means clustering algorithm, use it and analyze the data. Furthermore, the data that will be used for analysis consist of different Borough information on the basis of which we will determine venues name and categories. In addition, we use explore function, but have limited the analysis only to radius of 600 in each Toronto neighborhood and number returned venues by Foursquare API is limited to 50.Since Foursquare is a location data provider about various events within the area of retriever interest, the information such as venues name, location, menus are included. Therefore, Foursquare deemed to be the major data source, as all the information for the analysis will be obtained through the API.

Approach

For the analysis of the information, we have sub categorized the data into boroughs of which 4 most visited places are retrieved which will be a valuable determiner to examine which neighborhood has most customers.

Methodology

- **Data cleaning:**
  Firstly the data obtained from the source file is processed and manipulated to for the analysis. The various information obtained from setting a venues as :

  - Neighborhood, its latitude and longitude
  - Venue, its latitude and longitude
  - Venue type
- **Data preparation**
  After the data is retrieved, the data is filter as per the usage of the data for the analysis.
- **Feature selection**
  Data was then grouped by neighborhood to analyze which neighborhood had the maximum flow of customers; obtaining the mean of each category. In addition, the analysis only retrieves the top 4 location in each neighborhood
- **Clustering**
  For the analysis, the data is clustered in 4 categories using k means algorithm, which in return will help us understand each neighborhood density.