Customer Base Segmentation for Opening new Gym Chain

1. Introduction

1.1 Background

The aim of this project is to help an entrepreneur who is planning to open Cross fit gyms in different areas of Canada. Due to this corona crisis people have become more focused on health issues and recent study has shown that there is a surge in new subscriptions of gyms. CrossFit is a lifestyle characterized by safe, effective exercise and sound nutrition. CrossFit can be used to accomplish any goal, from improved health to weight loss to better performance. The program works for everyone—people who are just starting out and people who have trained for years. So this project will highlight those places where number of gyms are less so that ROI would be more and Investment risk will be less.

1.2 Business Problem

The objective of this capstone project is to find the best locations to open gyms in Toronto. If there are already of gyms available, then people may be less likely to join the new gym because existing gyms already have their own customer base and initially people may reluctant to move to new gym.

1.3 Target Audience

The entrepreneur who wants to open the gym chain.

2. Data

2.1 Data Definition

Link: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

We will extract data from the above mentioned link to get the neighborhood data of Canada.

Foursquare Data Provider:

We will need data about different venues in different neighborhoods of that specific borough. In order to gain that information, we will use "Foursquare" locational information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

After finding the list of neighborhoods, we then connect to the Foursquare API to gather information about venues inside each and every neighborhood. For each neighborhood, we have chosen the radius to be 100 meter.

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

1. Neighborhood
2. Neighborhood Latitude
3. Neighborhood Longitude
4. Venue
5. Name of the venue e.g. the name of a store or restaurant
6. Venue Latitude
7. Venue Longitude
8. Venue Category

3. Methodology Section

Clustering Approach:

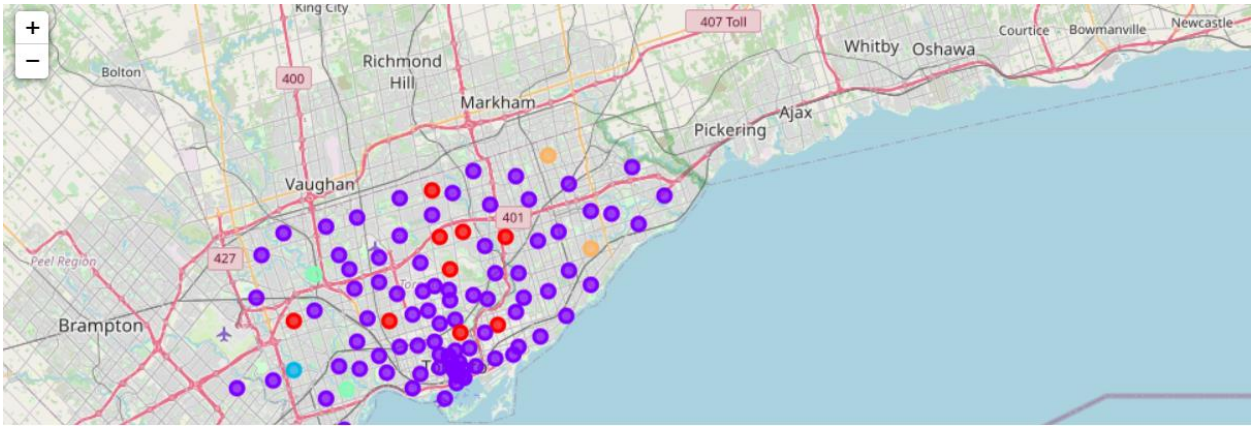
To find the best place to open the gym, we decided to explore neighborhoods, segment them, and group them into clusters to find similar neighborhoods in Toronto. We will be using K-Means clustering algorithm to cluster the location in 5 different clusters.

Work Flow:

Using credentials of Foursquare API features of near-by places of the neighborhoods would be mined. Due to http request limitations the number of places per neighborhood parameter would reasonably be set to 100 and the radius parameter would be set to 500. Then we will cluster our data in 5 clusters and we will avoid clusters having gyms in their most occurring venues, next we will check the proportion of fitness concerned people in clusters having no gyms in their most occurring places. We will select the cluster which has more number of these people. Finally, we will select the borough where fitness oriented venue is most common.

4. Result Section

4.1 Clusters



The clusters are visualized using below color codes

Violet: Cluster 1

Green: Cluster 3

Orange: Cluster 4

Red: Cluster 0

Blue: Cluster 2

Characteristics of Clusters:

Diner, Donut Shop, Park: Cluster 1

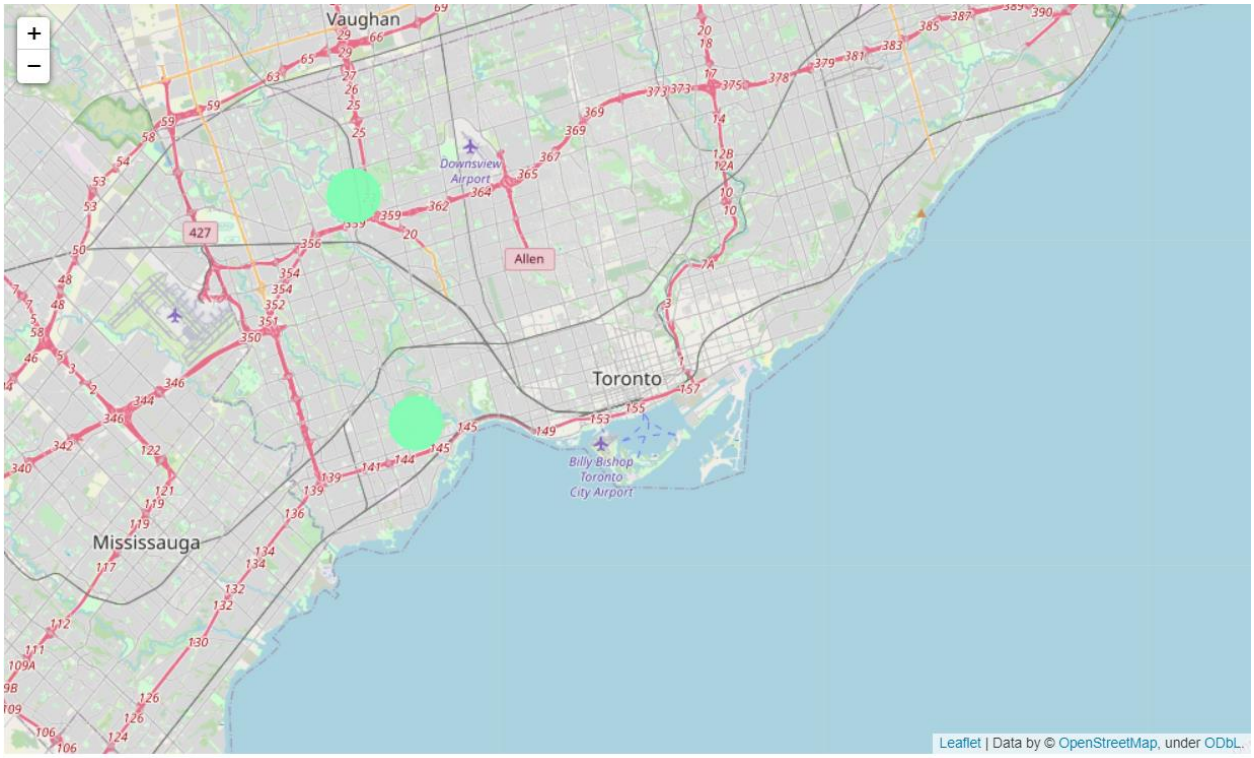
Coffee Shop : Cluster 2

Diner, Discount Store, Distribution Center: Cluster 3

Baseball Field, Dim Sum Restaurent, Discount Store: Cluster 4

Playground: Cluster 5

4.2 Recommended neighborhoods to open gyms



5. Discussion Section:

In our research we have taken mode of every clusters up to 10 venues, this will give us the top 10 most occurring venues in each cluster, by looking at it we can say gyms are not in top occurring venues, but before recommending the best place to open the first gym we need to narrow down our search.
For that we will check which clusters having most health conscious people, we can see cluster 3 and cluster 4 are having Yoga Studio in most occurring venues so it seems people in these clusters are already conscious about their fitness, let's consider these two clusters to select our final place.

6. Conclusion:

In this project, using k-means cluster algorithm I separated the neighborhood into 5 different clusters and for 103 different latitude and longitude from dataset, which have very-similar neighborhoods around them. Using the charts above results presented to a particular neighborhood based on average house prices and school rating have been made.

I feel rewarded with the efforts and believe this course with all the topics covered is well worthy of appreciation. This project has shown me a practical application to resolve a real situation that has impacting personal and financial impact using Data Science tools. The mapping with Folium is a very powerful technique to consolidate information and make the analysis and decision better with confidence.

Future Works:

I have considered few factors to recommend the best place to open new gyms, we can consider other factors as well like proportion of youth in population, gender bias, offers given by competitors.

Python Libraries:

Pandas: For creating and manipulating dataframes.

Folium: Python visualization library would be used to visualize the neighborhoods cluster distribution of using interactive leaflet map.

Scikit Learn: For importing k-means clustering.

JSON: Library to handle JSON files.

Geocoder: To retrieve Location Data.

Requests: To scrap and library to handle http requests