IBM Data Science Professional Certificate

Capstone Project

Choosing the Best Location to Open a Gym in Toronto, Canada

Presented by: Zhu Moran

Nov 2020

Project introduction

Nowadays, personal health is becoming a major problem which is ignored by many in the fast-paced society. People are putting a lot of attention in work, study and entertainment while neglecting the importance of having a fit body. Gyms, are becoming more and more popular for people who want to exercise and maintain good personal health. We can see that there are more and more gyms open in CBD areas of big cities. Opening a gym is also considered a good idea for people who want to start their own business. However, the location of gym needs to be carefully examined and it could be very crucial to success of the business.

Business Problem Briefing

The objective of the project is to analyse the neighbourhoods in Toronto and select a best location to open a new gym. The relevant data science methodologies like data cleaning, data sorting and data analysing will be applied. Machine learning techniques like clustering analysis will also be performed. The question to be answered is: if someone or some group wishes to open a new gym in Toronto, where is the best location?

Targeted people

The targeted people or targeted audience will be any individual or groups of people who are interested in opening a new gym in Toronto but having difficulties in selecting the appropriate location.

Data

Firstly, the neighbourhood's data of Toronto could be obtained from Wikipedia page:

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

A data frame will be created based on this data set and provide necessary information about neighbourhoods' info of Toronto.

Next, Foursquare API will be used to get the venue data about these neighbourhoods. Among all the available data, we will be interested in data with relevance to gym. Relevant data analysis techniques will be performed to clean and work with the data. Machine learning using K-means clustering will be performed to group the clusters

Lastly, a map with identified clusters will be provided and indicate the suitable locations to open a gym.

Methodology

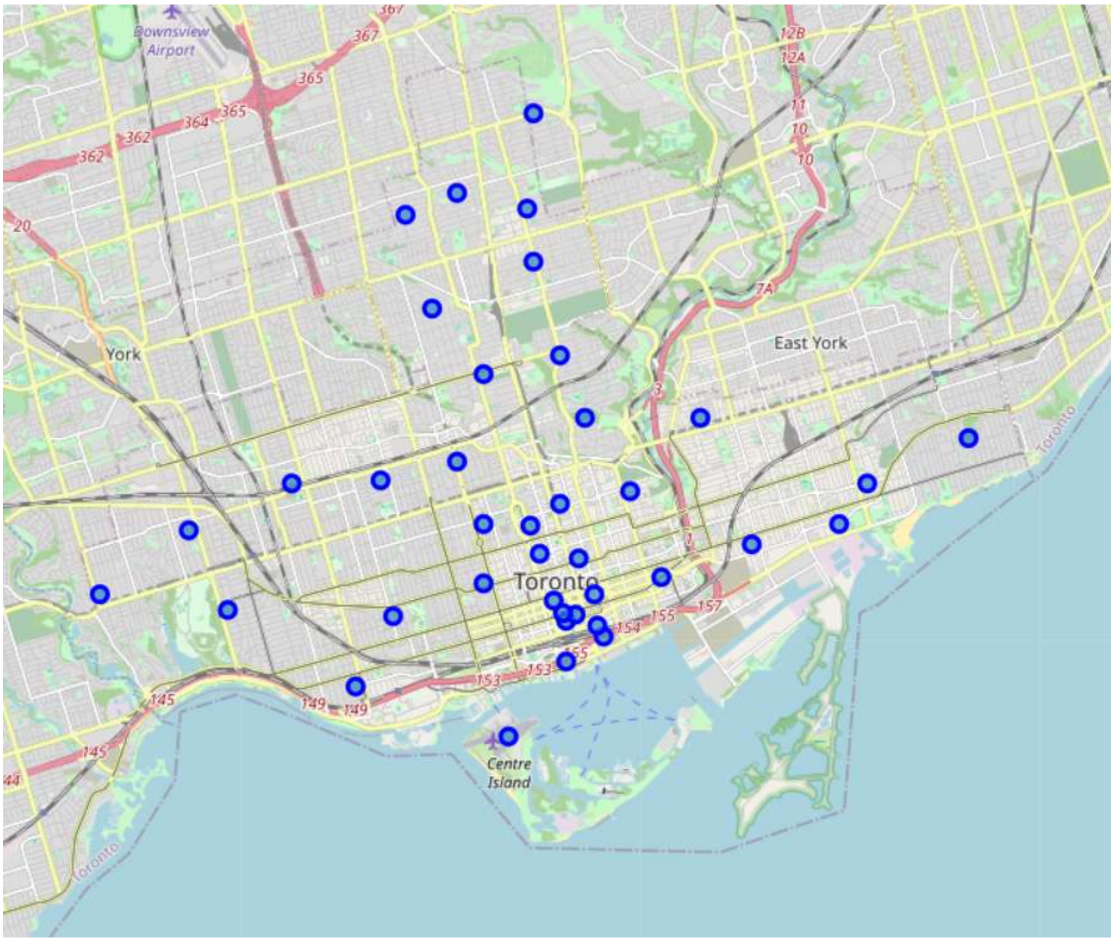
The list of neighbourhoods in Toronto could be obtained from the Wikipedia page

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

Next, we will attach the latitude and longitude of these neighbourhoods using the csv file from this link http://cocl.us/Geospatial_data.

Once these 2 steps are done, a pandas dataframe is created to include all the information about neighbourhoods in Toronto. Here, we perform a quick check using Folium. This is to ensure that the data we obtained from the sources are correct and make sense.

The folium map looks like this:



We will then make use of Foursquare API to get the top 100 venues that are within 1500 meters of the neighbourhood. It is reasonable that someone is willing to travel 1500 meters to hit a gym. A longer travelling distance may result in no motivation to go to the gym. Calls make via Foursquare API will return the venue data that we are interested in JSON file. We then can extract the information about these venues including name, category and location data i.e. latitude and longitude.

Next, we will analyse the data by taking the mean of the frequency of occurrences of each venue's category. This is also the data preparation for clustering analysis.

Since the category that we are interested in is Gym, we will only focus on the gym category.

Lastly, we will perform clustering analysis on the data by using k-means clustering. The neighbourhoods are clustered into 3 clusters based on their frequency of occurrence of Gym. The results from the analysis will give us the idea of which neighbourhood has high concentration of gyms and which are not. This will also help us to answer the question of where to open a new gym. It is ideal to open a gym at clusters with low occurrence of gym.

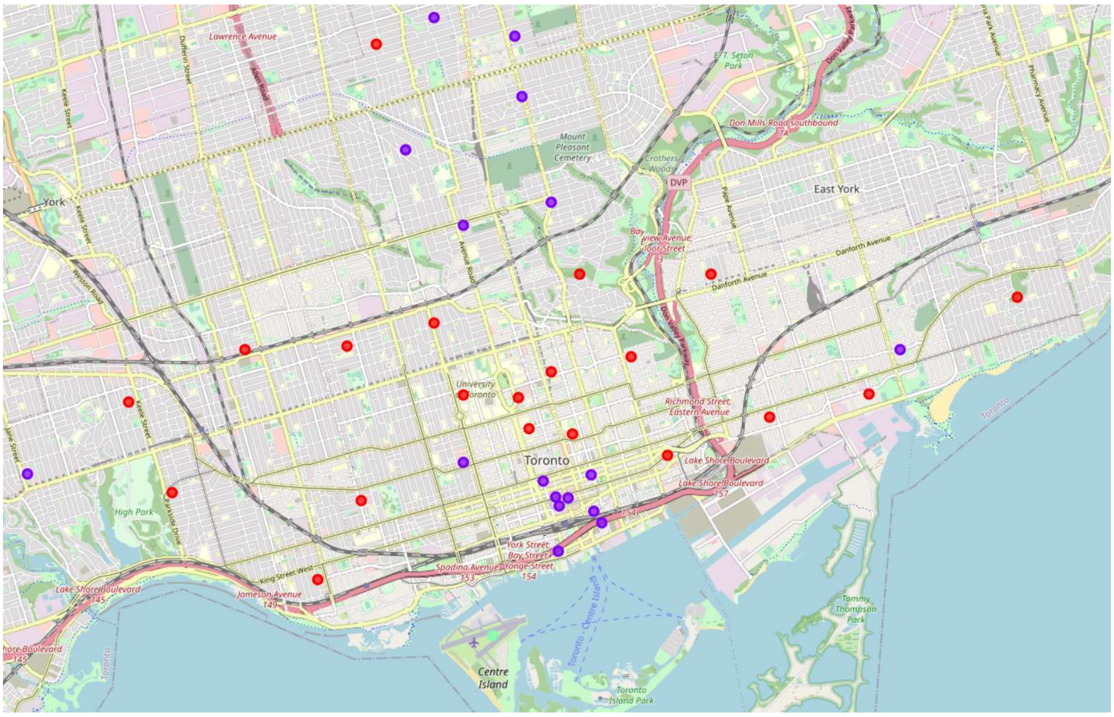
Results Presentation

The results from k-means clustering is plotted on Toronto map based on the occurrence of "Gym".

The 3 clusters representing the following information:

- ➢ Cluster 0: Neighbourhoods with low or no gyms (Red dots)
- ➢ Cluster 1: Neighbourhoods with moderate number of gyms (Purple dots)
- ➢ Cluster 2: Neighbourhoods with relatively a large number of gyms (Green dots)

Visualisation:



Results Discussion

Based on the clustering map, it can be concluded that neighbourhoods in the downtown area have moderate number of gyms while at areas at a distance away from downtown, not many gyms exist. To answer the question raised by this project, we will recommend people to open new gyms at cluster 0 neighbourhoods to avoid competition. Also, it is noticeable that opening a gym in cluster 2 neighbourhood is a very unwise decision.

Limitations and Future Studies

The limitation of this project is very obvious. We only focused on one aspect which is the occurrence of gyms at each neighbourhood. There are many other factors that we did not take into account such as the population, income of people as well as rent. Future study could introduce more factors to be used in clustering algorithm to determine the neighbourhood to open a gym more accurately.

Conclusion

To conclude, we have identified the business problem, collected relevant data, prepared data and performed machine learning by using clustering techniques. To answer the question raised in the project, new business adventurers could open a new gym in the cluster 0 neighbourhoods. This will help them potentially avoid competition from other gyms and dominate the market.