Best location to open a gym in Toronto

Chouvattanak Eng

Table of Contents

1. Introduction

1.1 Background

People in these day and age incorporate fitness into their lifestyle. Thus, we can see the surge of gyms and fitness that answer to these demands. But opening a gym or other business is a tough decision, it involves making many difficult decisions such as: Who is our targeted customers, how much should be cost for the gym membership cost? Are there any competitions in the region? And one of the most important question that needs thorough answer would be what is the best location for customers to come and exercise and in turn will optimize the profitability.

1.2 Business Problem

Imagine a client that want to open a gym in Toronto and want our service to help find the optimum location that will benefit the business in the long run. Which location in Toronto is the optimum point of interest? We first need to think about the factors that contribute to this. It would be based on income, competition and density of people in the neighborhood can also play an important factor as well. So, to solve this problem, we will mainly use Foursquare API to get the venues location, Neighborhoods in Toronto from Wikipedia and census data from Toronto's Open Data Portal.

1.3 Interest

The targeted audiences of this project would be the business people who want to open a new gym or expand their franchised. Through this study, they will have a clear overview of the locations in Toronto and can confidently target their specific clients, which will give them competitive advantage and a head start in the gym business.

2. Data

2.1 Data Sources

We mainly focus on 4 data sources in this instance.

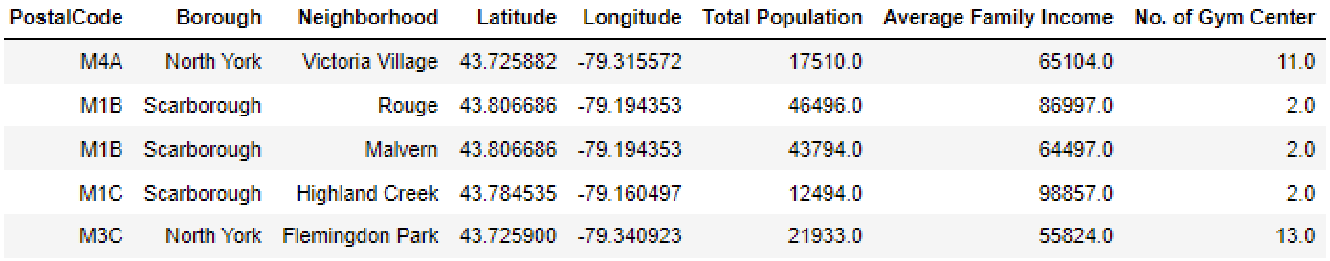
- Wikipedia: We will extract the postal code, borough and neighborhoods in Toronto.
- Geospatial Data: A geospatial data of Toronto that contains the Postal code along with latitude and longitude of neighborhoods in Toronto.
- Foursquare API: An API call to get the locations and information of venues in Toronto. (Foursquare API requires a developer account in order to log in)
- Toronto Census data: List of total population, household income and other info in the neighborhoods in Toronto.

2.2 Data Cleaning

Below procedures are summary the data cleaning and wrangling process:

1. Pulling data from data sources.
2. Drop row and column based on data quality
3. Mapping all data into one table
4. Prepare data by selecting and applying standard scaling to the features that will be used in the K-mean clustering model

Below is the expected table that will be used in modeling**:**



PostalCode	Borough	Neighborhood	Latitude	Longitude	Total Population	Average Family Income	No. of Gym Center
M4A	North York	Victoria Village	43.725882	-79.315572	17510.0	65104.0	11.0
M1B	Scarborough	Rouge	43.806686	-79.194353	46496.0	86997.0	2.0
M1B	Scarborough	Malvern	43.806686	-79.194353	43794.0	64497.0	2.0
M1C	Scarborough	Highland Creek	43.784535	-79.160497	12494.0	98857.0	2.0
M3C	North York	Flemingdon Park	43.725900	-79.340923	21933.0	55824.0	13.0

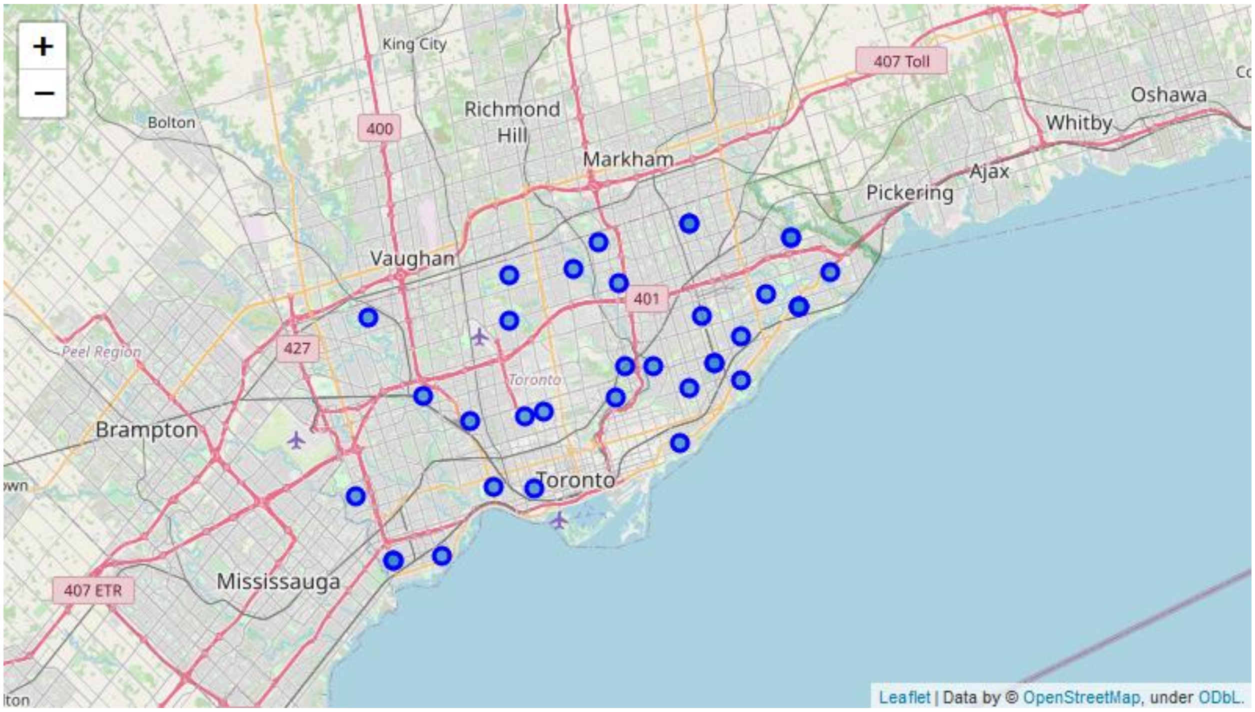
3. Methodology

The above data will be combined together into a single table using pandas library and will further apply standard scaling operation to further help with our model that will be used in this study. the samples are roughly 40 samples and 3 to 4 features will be selected based on quality and its relation to the study.

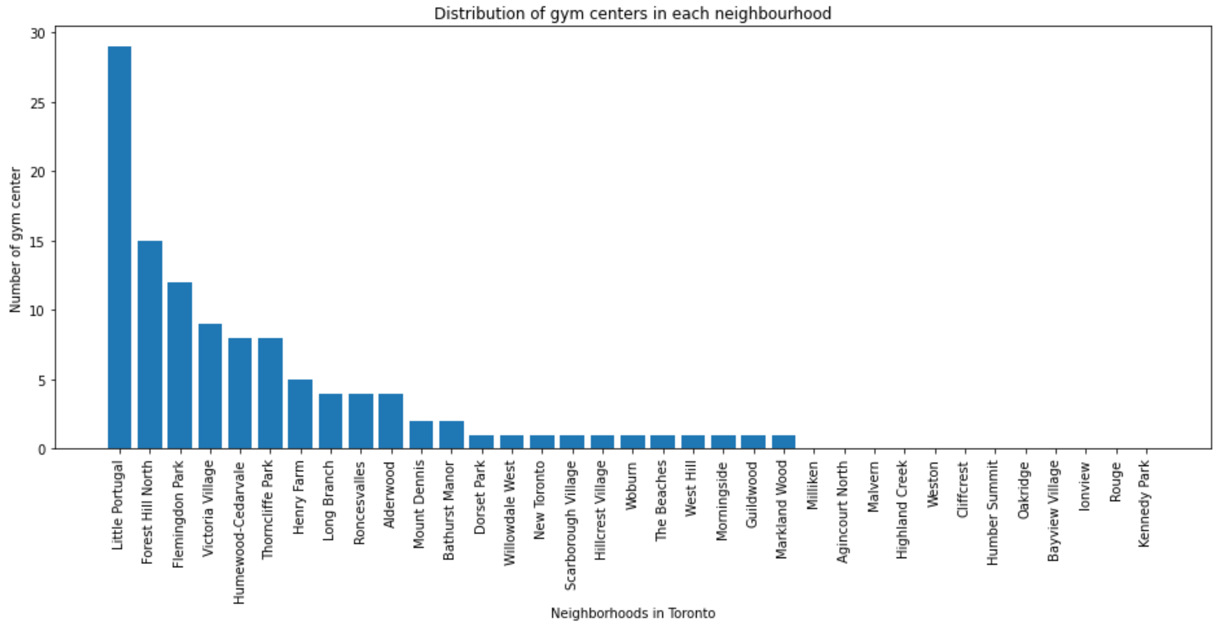
- For Wikipedia, the data will be scraped from table and loop through each row and table data using BeautifySoup library, which will give us the list of all neighborhoods along with its borough and postal code in Toronto. Some row needs to be drop due to NA information in both borough and neighborhood columns.
- We also need geo spatial file, which contains the postal code, latitude and longitude of the neighborhoods, in order to visualize the Toronto maps and help us better understand when we cluster the neighborhoods using k-mean clustering model.
- After combine the geospatial data with neighborhoods data using postal code as its primary key, we will use the Foursquare API to query the gym center in the area and count the gym centers in each neighborhood. Thus, we add the number of gym center into our table and use it as a feature in the upcoming model.
- Two more features can be found in Toronto census data, which will list down all the neighborhoods along with option to specify the information that will want, in this case, it would be the total population and average household income in each neighborhood.
- Overall, 3 features will be selected in this study. Total population, average household income and frequency of gym centers in the area.
- After the required information is gathered and put into a table, we then can select the features in the table for our k-mean clustering model.
- Standard scaling technique will be used on the 3 features to scale down data into similar units to facilitate the model training.
- The output of the training will be the cluster labels that will be used to map back into the toronto_data table, thus, we can used the labeled dataset to visualize in folium map.
- Finally, we will discuss and conclude the result of the clustering.

4. Exploratory Data Analysis

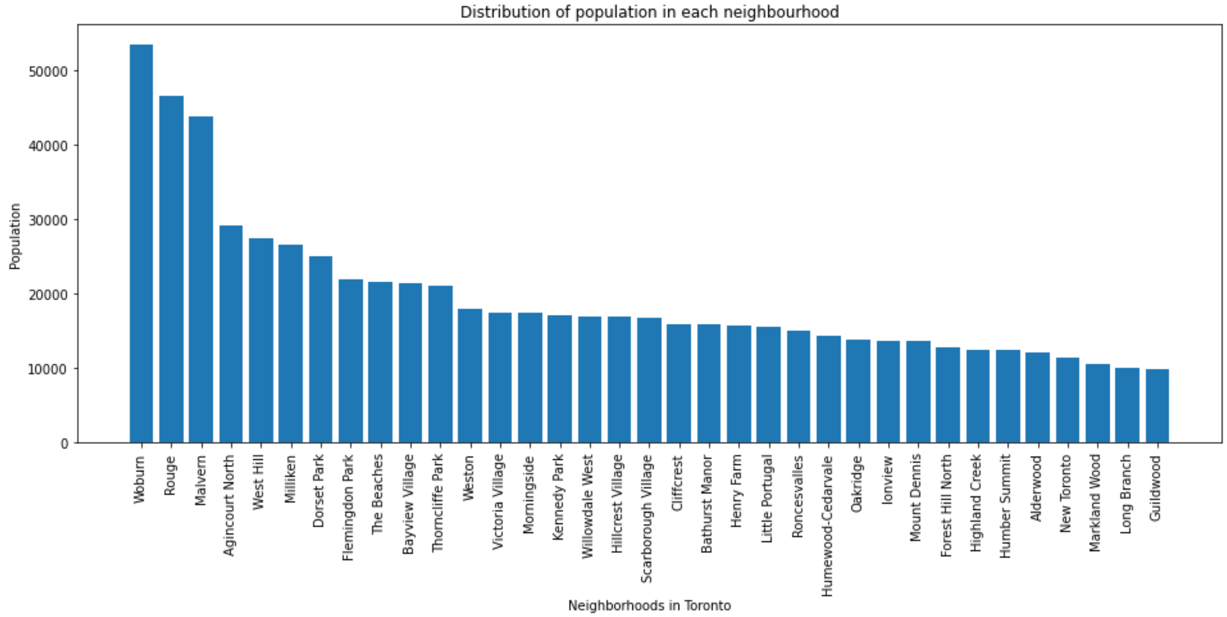
4.1 Let's view Toronto's neighborhoods superimposed on top.



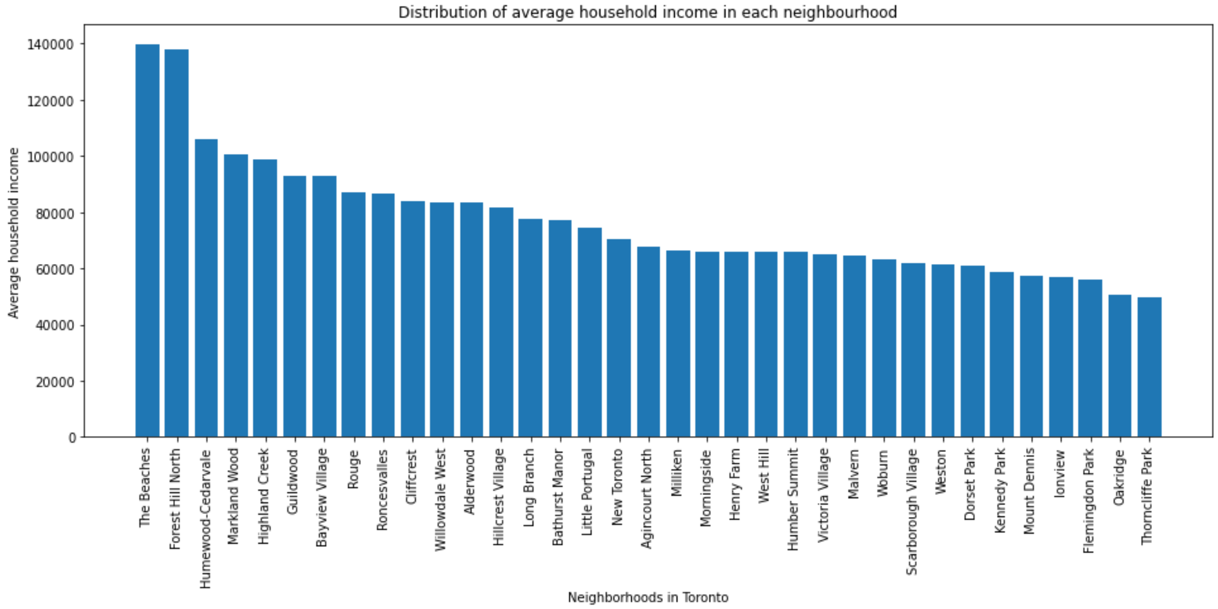
4.2 Distribution of Gyms in the neighborhoods



4.3 Distribution of population in the neighborhoods



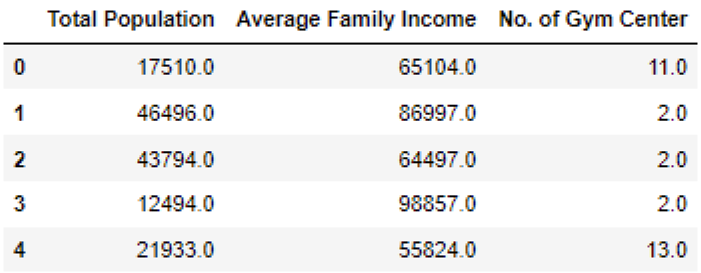
4.4 Average household income in the neighborhoods



4.5. Model selection

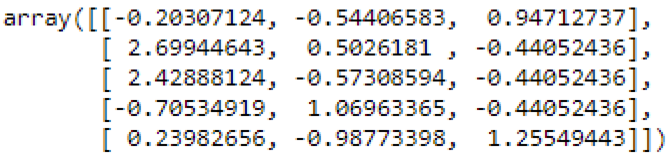
4.5.1 Data Preprocessing

Features selection for K-mean modeling.



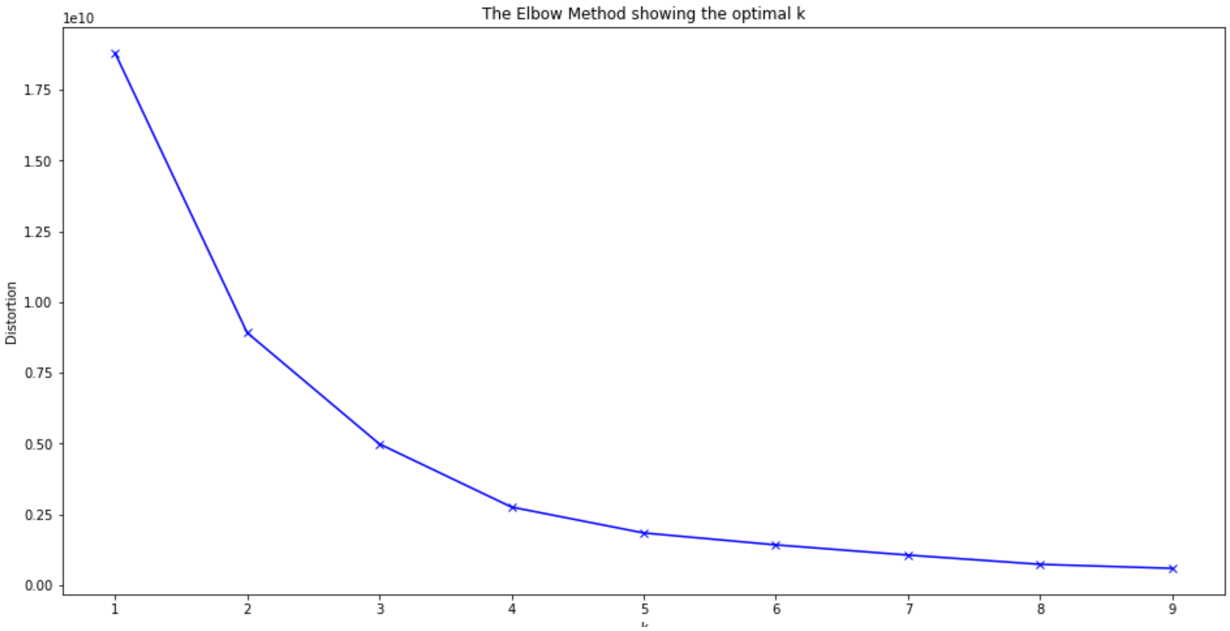
	Total Population	Average Family Income	No. of Gym Center
0	17510.0	65104.0	11.0
1	46496.0	86997.0	2.0
2	43794.0	64497.0	2.0
3	12494.0	98857.0	2.0
4	21933.0	55824.0	13.0

Using standard scaling technique, we will be able to scale all the data to facilitate the modeling algorithm.

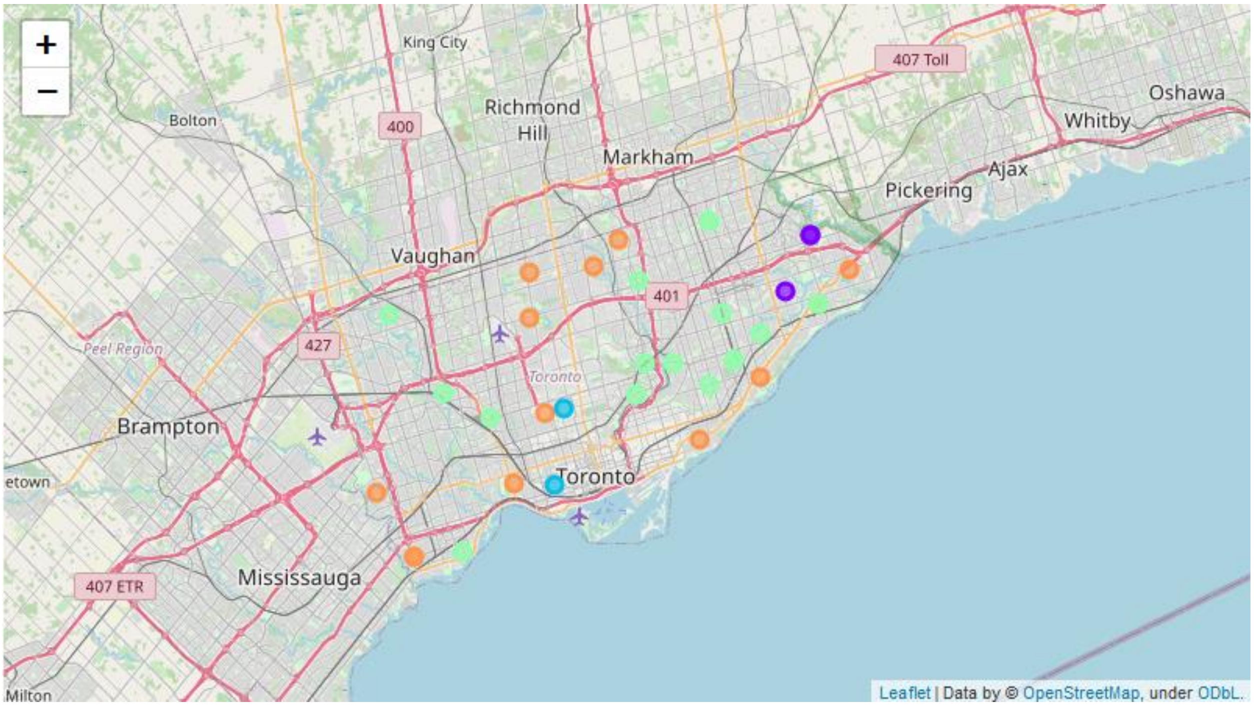


```
array([[-0.20307124, -0.54406583,  0.94712737],
       [ 2.69944643,  0.5026181 , -0.44052436],
       [ 2.42888124, -0.57308594, -0.44052436],
       [-0.70534919,  1.06963365, -0.44052436],
       [ 0.23982656, -0.98773398,  1.25549443]])
```

4.5.2 k-Means Cluster Modeling



4.5.3 Visualize labels after modeling



5. Results and Discussion

5.1 Summary Clusters

In summary:

- **Cluster 1:** With high to mid population and spending power, with few competitors in the area signifies the lack of interest from population. But with good marketing campaign of healthy lifestyle, there could be a lot of potential customers in these neighborhoods.
- **Cluster 2:** Low population in the area could be the concern of return of investment. Thus, investment in these neighborhoods wouldn't be a good choice.
- **Cluster 3:** High competitors and spending power shows that these neighborhoods are premium customers even if the population is relatively low, because of the spending power, 1 customer could be a member of multiple gym centers.
- **Cluster 4:** poses high barrier of entry with high competitors in the neighborhood and low population. Thus, customers in these neighborhoods would be cautious when choosing the gym of their choice.

6. Conclusion:

Throughout this study, 4 cluster labels are chose based on K-mean elbow method. Within these 4 labels, the most promising label would be in Cluster 1, as it has high population, mid spending power and fewer competitions give us opportunity to penetrate these neighborhoods. With more research and study into these neighborhoods on how to persuade potential customers into our business. There are a lot of potentials in these areas.

Another interest cluster would be cluster 3, as it has a lot of premium customers. if we could franchise the best brand, these high spending customer would surely choose our gym from competitors.

In conclusion, this study is relatively useful as location selection is very important in the success of the business. We can also use these study as the blueprint for select the best location for other business as well.