The Battle of Neighborhood- Best location to open GYM/Fitness Centre in Toronto

Laxmikantha Herle

Introduction

This project aims to utilize all Data Science concepts learned in the Data Science Professional Course. We define a Business Problem, the data that will be utilized and using that data, we are able to analyze it using Machine Learning tools. In this project, we will go through all the processes in a step by step manner from problem designing, data preparation to final analysis and finally will provide a conclusion that can be leveraged by the business stakeholders to make their decisions.

Table of Contents

- Background & Problem Description
- Data Description
- Methodology
- Machine Learning
- Data Analysis
- Discussion and Conclusion
- References

Background & Problem Description

Gone are the days when gyms targeted only a small segment of the population. Nowadays, gyms are attracting more members. Members can choose a low-budget gym and still get a good sweat in or they could opt for a smaller, more specialized boutique studio to meet their specific fitness goals. Plus, there are 24-hour fitness Centre popping up nationwide to make going to the gym is a lot more convenient for today's consumers. Revenue for the Gym, Health and Fitness Clubs industry in Canada has grown over the five years to 2020 as a result of consumer trends and the proliferation of public health campaigns. With an increasing rate of adult obesity expected during the period, the Public Health Agency of Canada (PHAC) has stressed adherence to fitness and healthy lifestyle choices. According to the PHAC and the Canadian Institute for Health Information, obesity is expected to incur more than $5.0 billion in costs each year, which must be covered by the healthcare system.

The objective of this project is to determine "what might be the 'best' neighborhood in Toronto to open a GYM/Fitness Centre". Will use foursquare location data and regional clustering of venue information to determine the 'best' neighborhood in Toronto to open a GYM/Fitness Centre. We will find the most suitable location for an entrepreneur to open a new GYM/Fitness Centre in Toronto, Canada.

Target Audience

Information provided by this report would be useful for People who wants open GYM/Fitness Centre in Toronto, Canada. The Objective is to locate and recommend to People which neighborhood of Toronto will be the best choice to open GYM/Fitness Centre.

Data Description

To consider the objective stated above, we can list the below data sources used for the analysis

I. Toronto Neighborhood Data: The following Wikipedia page was scraped to pull out the necessary information: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
   The information obtained i.e. the table of postal codes was transformed into a pandas data frame for further analysis.

II. Coordinate data for each Neighborhood in Toronto: The following csv file gave us the geographical coordinates of each postal code:http://cocl.us/Geospatial_data

III. Venue Data in Toronto, Canada. Geographical Coordinates data will be utilized as input for the Foursquare API that will be leveraged to provision venues information for each neighborhood. We will use Foursquare API to explore neighborhood in Toronto, Canada.

Methodology

We used the *BeautifulSoup* package to transform the data in the table on the Wikipedia page into the pandas data frame. After all the data was collected and put into data frames, cleansing and merging of the data was required to start the process of analysis. When getting the data from Wikipedia, there were Boroughs that were not assigned to any neighborhood therefore, the following assumptions were made,

- Only the cells that have an assigned borough will be processed.
- Borough's that were not assigned get ignored.
- More than one neighborhood can exist in one postal code area. For example, in the table on the Wikipedia page, will notice that M5A is listed twice and has two neighborhood: Harbourfront and Regent Park. These two rows will be combined into one row with the neighborhood separated with a comma
- If a cell has a borough but a not assigned neighborhood, then the neighborhood will be the same as the borough.

After the implementation of the following assumptions, the rows were grouped based on the borough as shown below Fig.1

	Postal Code	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

Fig.1 Toronto Neighborhood Data

Load & Explore Geo Coordinate Data

The second source of data provided with the Geographical coordinates of the neighborhood with the respective Postal Codes. The file was in CSV format, so we had to attach it to a Pandas data frame, we merged the two tables together based on Postal Code as shown below Fig.2.

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

Fig.2 Toronto Neighborhood Data with Coordinates

We then use the python *Folium* library to visualize geographic details of Toronto and its boroughs. Created a map of Toronto with boroughs superimposed on top using the latitude and longitude values to get the visual as below Fig.3

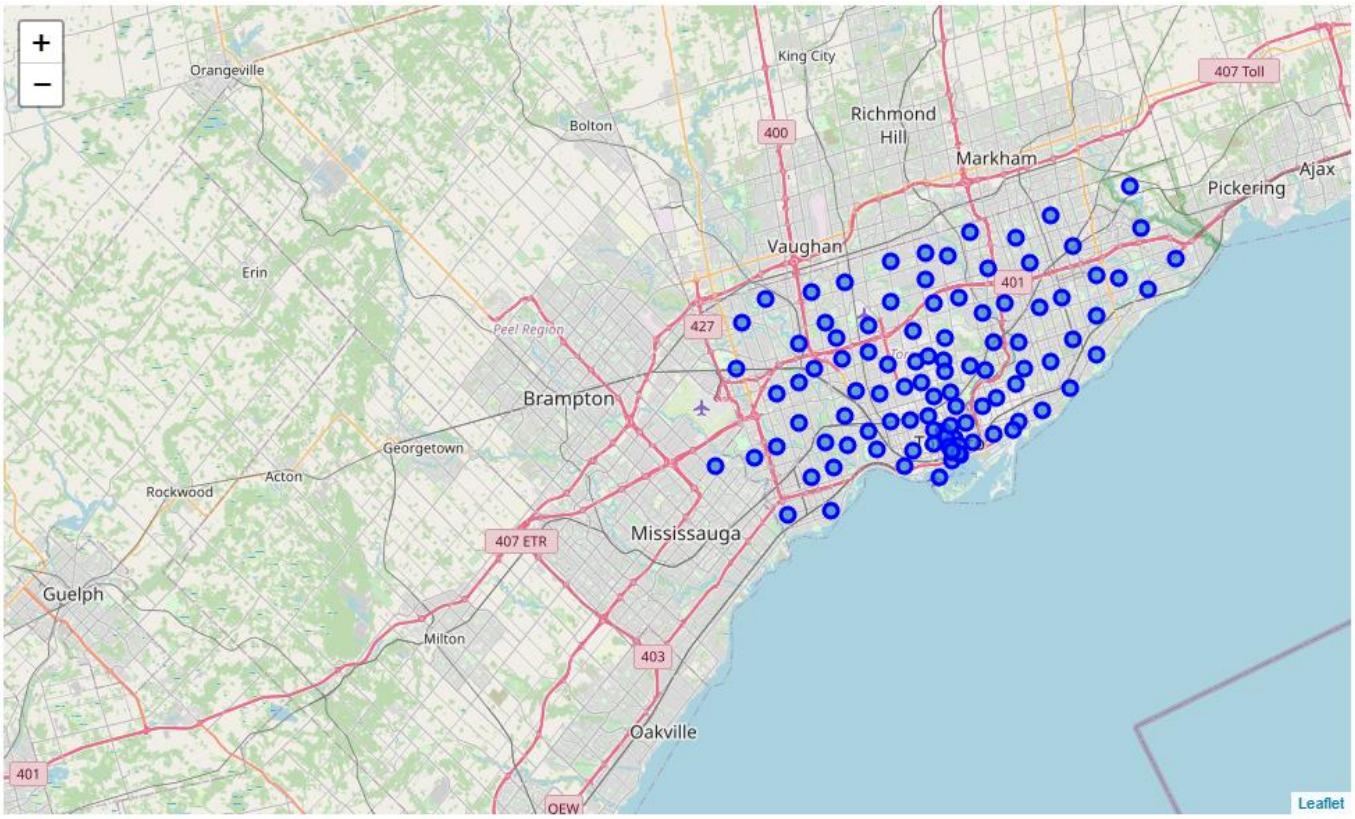

Fig.3 Toronto Neighborhood Data in Map

We are going to start utilizing the Foursquare API to explore the neighborhood venues and segment them. We set the LIMIT parameter to 100, which would limit the number of venues returned by the Foursquare API and the radius of 500 meter. Getting this data was crucial to analyzing the number of GYM/Fitness Centre all over Toronto. Then we merged the Foursquare Venue data with the Neighborhood data which then gave us the nearest Venue for each of the Neighborhoods Fig.4

	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	North York	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	North York	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	North York	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
3	North York	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop
4	North York	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant

Fig.4 Neighborhood Data with Venue details

Machine Learning

Then to analyze the data we performed a technique in which Categorical Data is transformed into Numerical Data for Machine Learning algorithms. This technique is called One hot encoding. For each of the neighborhoods, individual venues were turned into the frequency at how many of those Venues were located in each neighborhood. Then we grouped those rows by Neighborhood and by taking the average of the frequency of occurrence of each Venue Category.

	Neighborhood	Yoga Studio	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	C
0	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	
1	Alderwood, Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	
2	Bathurst Manor, Wilson Heights, Downsview North	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	
3	Bayview Village	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	
4	Bedford Park, Lawrence Manor East	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.04	0.0	0.0	0.0	0.0	

Fig.5 One hot Encoding

After, we created a new data frame that only stored the Neighborhood names as well as the mean frequency of GYM/Fitness Centre in that Neighborhood. This allowed the data to be summarized based on each individual Neighborhood and made the data much simpler to analyze. We combine Both GYM & Fitness Centre Venues for better identification and Analysis.

	Neighborhood	GYM_T
0	Agincourt	0.000000
1	Alderwood, Long Branch	0.111111
2	Bathurst Manor, Wilson Heights, Downsview North	0.000000
3	Bayview Village	0.000000
4	Bedford Park, Lawrence Manor East	0.000000

Fig.6 Data Frame with only GYM and Fitness Centre avg

K-Means Clustering

We used K-Means clustering. To get our optimum K value that was neither overfitting nor under fitting the model, we used the Elbow Point Technique. In this technique, we ran a test with different number of K values and measured the accuracy and then chose the best K value. The best K value is chosen at the point in which the line has the sharpest turn. In our case, we had the Elbow Point at K = 4. That means we will have a total of 5 clusters.

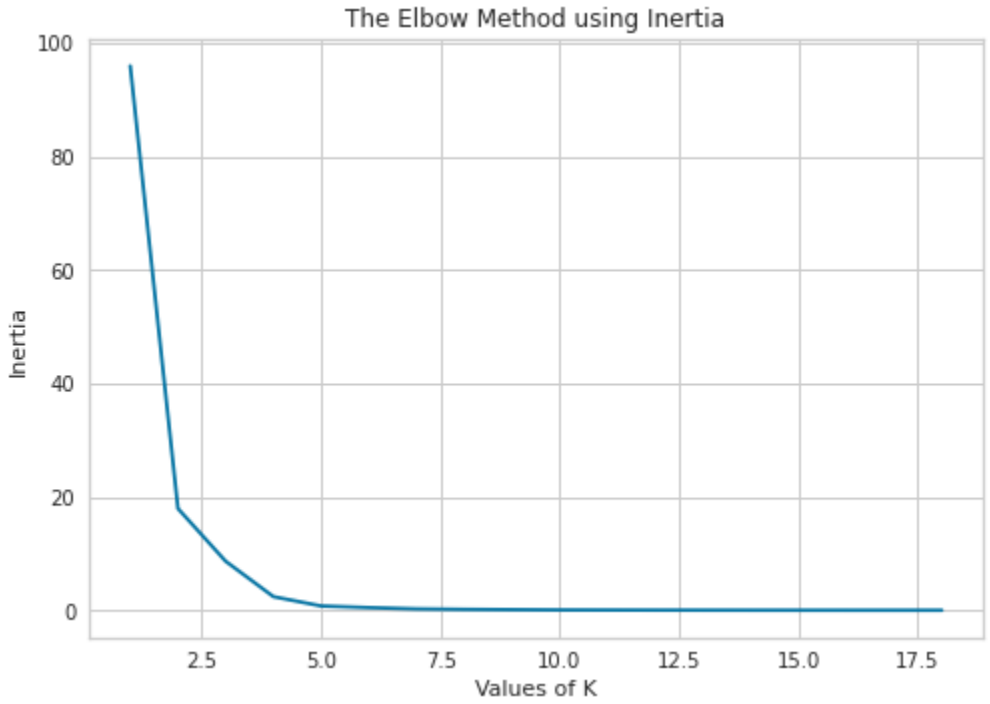

Fig.7 Elbow method using Inertia

Then we used a model that accurately pointed out the optimum K value. We imported '*KElbowVisualizer*' from the *Yellowbrick* package. Then we fit our K-Means model above to the Elbow visualizer.

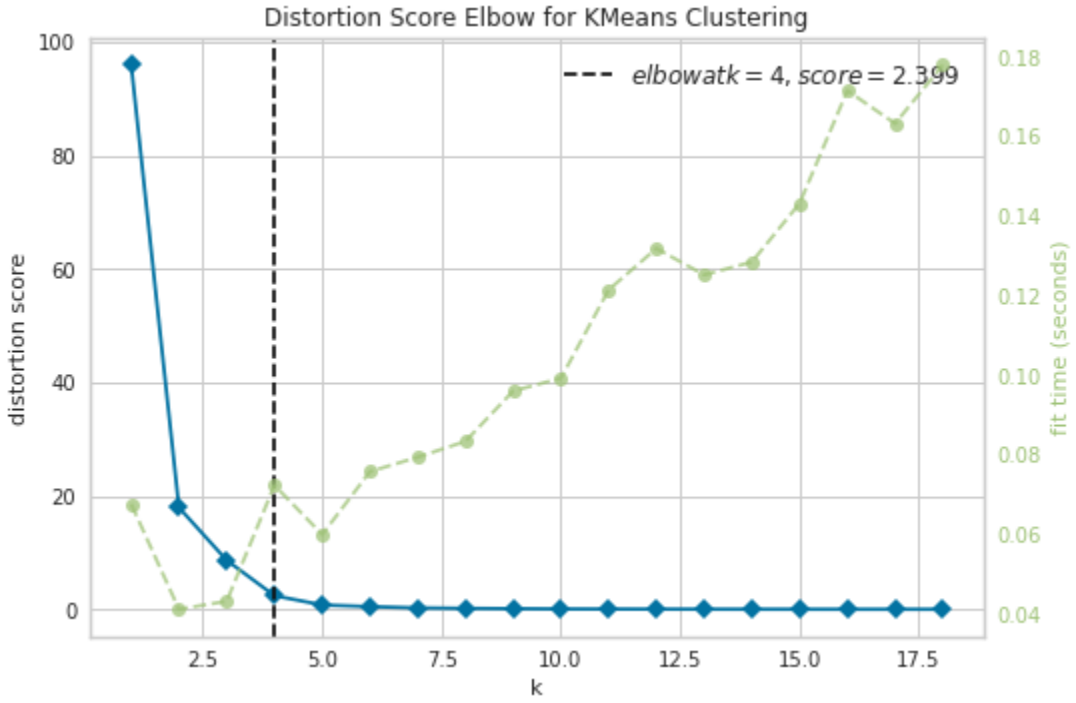

Fig.8 Finding K Value from KElbowVisualizer

We just integrated a model that would fit the error and calculate the distortion score. From the dotted line, we see that the Elbow is at K=4. Moreover, in K-Means clustering, objects that are similar based on a certain variable are put into the same cluster.

Each of these clusters was labelled from 0 to 3 as the indexing of labels begins with 0 instead of 1

After, we merged the venue data with the table above creating a new table which would be the basis for analyzing new opportunities for opening a GYM/Fitness Centre in Toronto. Then we created a map using the *Folium* package in Python and each neighborhood was colored based on the cluster label.

Cluster 1 — Orange

Cluster 2 — Purple

Cluster 3 — Red

Cluster 4 — Green

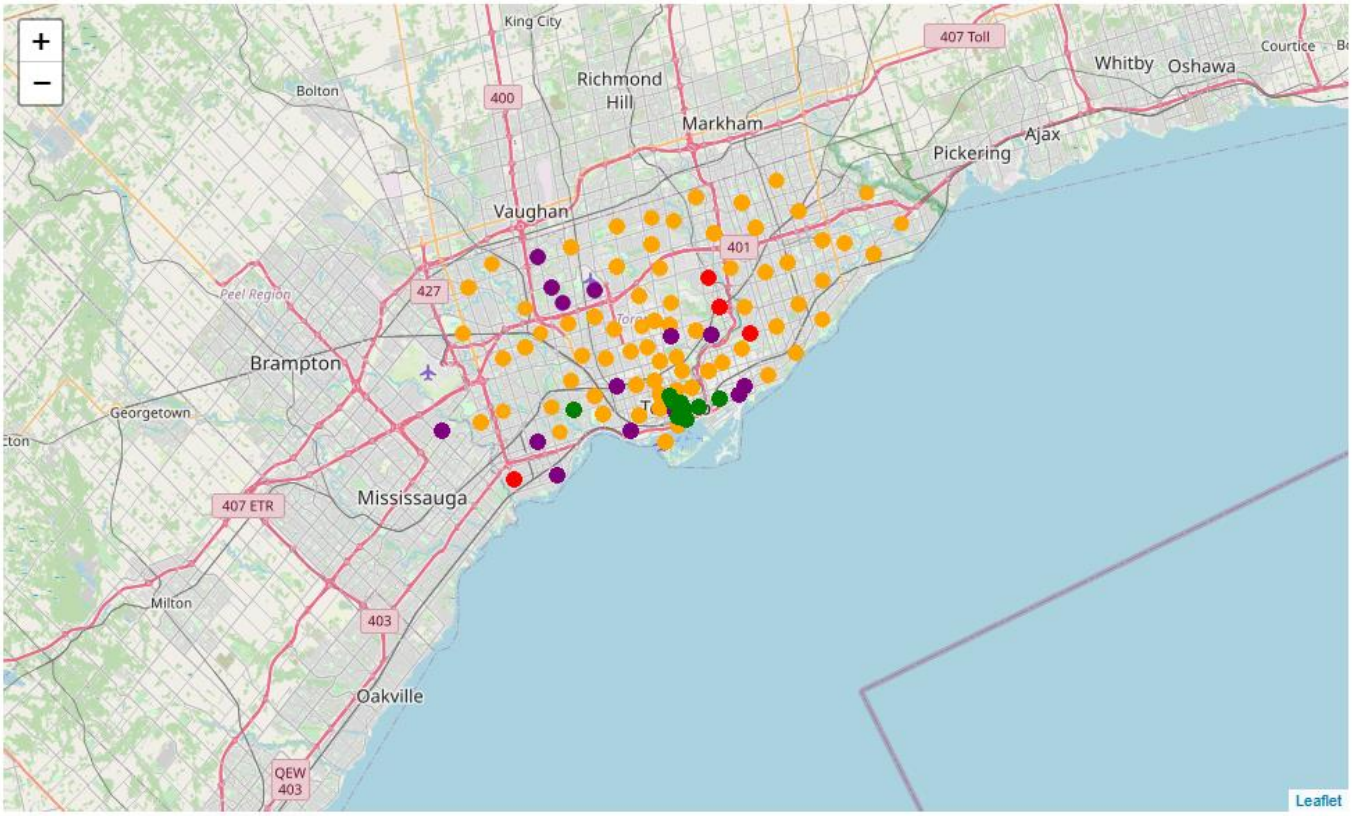

Fig.9 Clusters

Data Analysis

Let's check the total amount of neighborhoods in each cluster and the average GYM/Fitness Centre in that cluster. From the bar graph below, we can compare the number of Neighborhoods per Cluster. We see that Cluster 1 has the highest no neighborhoods 70 while cluster 3 has least 3. Cluster 2 has 13 Neighborhoods and Cluster 4 has 10 Neighborhoods.

Then we compared the average GYM/Fitness Centre per cluster. Cluster 3 is having highest Mean and Cluster 1 is having least.

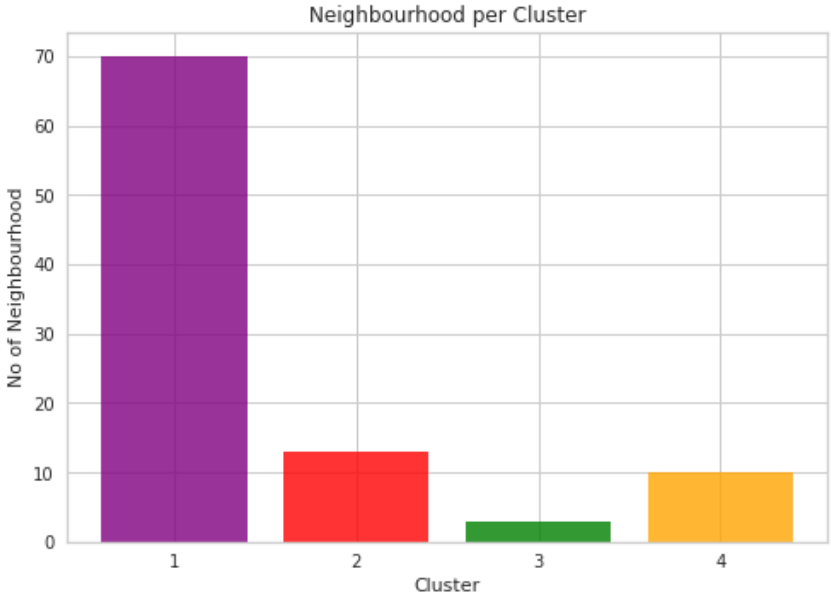

Fig.10 No of Neighborhood per Cluster

Fig.11 Average GYM per Cluster

Analysis of each Cluster

Cluster 1

Cluster-1 had 216 unique Venue Categories. Cluster 1 is not having any GYM or Fitness Center. But Cluster-1 is having highest no of Neighborhood in it. Cluster-1 having 70 Neighborhoods and 9 Boroughs. In that North York is having 17 neighborhoods and Scarborough having 16 neighborhoods.

	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	GYM_T	Cluster Labels
0	North York	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park	0.0	0
2069	Downtown Toronto	Church and Wellesley	43.665860	-79.383160	The Anndore House	43.668801	-79.385413	Hotel	0.0	0
2081	Downtown Toronto	Church and Wellesley	43.665860	-79.383160	Cawthra Square Dog Park	43.666583	-79.380040	Dog Run	0.0	0
58	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	Suzy Shier	43.718846	-79.465906	Clothing Store	0.0	0
2080	Downtown Toronto	Church and Wellesley	43.665860	-79.383160	NC Salon +	43.669406	-79.386748	Health & Beauty Service	0.0	0

Fig.12 Cluster 1

Cluster 2

Cluster-2 had 13 Neighborhoods and 129 unique Venue Categories across 8 different Boroughs.26 venue location in Cluster-2 is having GYM or Fitness Center. Which is highest among all 4 Clusters. Downtown Toronto is having highest of 15 GYM or Fitness Center in it. Cluster-2 had average GYM or Fitness Center rate as 0.05.

	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	GYM_T	Cluster Labels
1498	Mississauga	Canada Post Gateway Processing Centre	43.636966	-79.615819	Anoush	43.636769	-79.620840	Mediterranean Restaurant	0.076923	1
1490	Mississauga	Canada Post Gateway Processing Centre	43.636966	-79.615819	Hilton Garden Inn	43.638519	-79.618721	Hotel	0.076923	1
1745	Etobicoke	New Toronto, Mimico South, Humber Bay Shores	43.605647	-79.501321	Hex-Mex	43.601261	-79.502284	Mexican Restaurant	0.076923	1
1743	Etobicoke	New Toronto, Mimico South, Humber Bay Shores	43.605647	-79.501321	Sense Appeal	43.601729	-79.501063	Café	0.076923	1
1742	Etobicoke	New Toronto, Mimico South, Humber Bay Shores	43.605647	-79.501321	Pet Valu	43.602431	-79.498653	Pet Store	0.076923	1

Fig.13 Cluster 2

Cluster 3

Cluster-3 had 31 unique venue Categories in 3 neighborhoods. Had 5 venue location is having GYM or Fitness Centre in it. With an average rate of 0.111, which is highest among all 4 clusters. Which means Cluster-3 is having highest average rate of GYM or Fitness Center even though it had 3 Neighborhoods.

	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	GYM_T	Cluster Labels
233	North York	Don Mills	43.725900	-79.340923	Harvey's	43.726603	-79.341035	Restaurant	0.111111	2
218	North York	Don Mills	43.725900	-79.340923	Oomomo	43.726429	-79.343283	Discount Store	0.111111	2
1872	Etobicoke	Alderwood, Long Branch	43.602414	-79.543484	Toronto Gymnastics International	43.599832	-79.542924	Gym	0.111111	2
230	North York	Don Mills	43.725900	-79.340923	Genghis Khan Mongolian Grill	43.726906	-79.341216	Asian Restaurant	0.111111	2
237	North York	Don Mills	43.725900	-79.340923	Barber Greene Square	43.727654	-79.340810	Shopping Mall	0.111111	2

Fig.14 Cluster 3

Cluster 4

Cluster-4 had 155 unique Venue Categories in that 15 venue location contains GYM or Fitness Center. Though Average Rate of GYM in Cluster-4 is 0.02 which is having 2nd highest numbers of GYM in it. Downtown Toronto is having 13 GYM and East Toronto and west Toronto contains one GYM or Fitness Center each.

	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	GYM_T	Cluster Labels
1785	Downtown Toronto	Stn A PO Boxes	43.646435	-79.374846	Garrison Bespoke	43.648102	-79.376334	Tailor Shop	0.030928	3
1780	Downtown Toronto	Stn A PO Boxes	43.646435	-79.374846	D.W. Alexander	43.648333	-79.373826	Cocktail Bar	0.030928	3
1778	Downtown Toronto	Stn A PO Boxes	43.646435	-79.374846	Biff's Bistro	43.647085	-79.376342	French Restaurant	0.030928	3
1777	Downtown Toronto	Stn A PO Boxes	43.646435	-79.374846	Berczy Park	43.648048	-79.375172	Park	0.030928	3
1776	Downtown Toronto	Stn A PO Boxes	43.646435	-79.374846	Hockey Hall Of Fame (Hockey Hall of Fame)	43.646974	-79.377323	Museum	0.030928	3

Fig.15 Cluster 3

Discussion and Conclusion

Most of the GYMs are in Cluster-2, Neighborhood located in Downtown Toronto is having highest no of GYM in it. Neighborhood Commerce Court and First Canadian Place having 5 GYM these location. East York, Mississauga and North York is having less GYM in these locations, each contains 1 GYM After Downtown Toronto North York is having 19 Neighborhoods in it. But less No of GYM venues. So North York is suitable for opening New GYM. The second-best Borough that have a great opportunity would be Scarborough. Which is not having any GYM Venues. Cluster-1 will be having more no of Neighborhoods in North York and Scarborough and Cluster-1 is not having GYM venues in it. So Cluster-1 Neighborhoods will be the best location to open a New GYM. Some of the drawbacks of this analysis are — the clustering is completely based on data obtained from the Foursquare API. Also, the analysis does not take into consideration of the population across neighborhoods as this can play a huge factor while choosing which place to open a new GYM or Fitness Center.

This concludes the optimal findings for this project and recommends the entrepreneur to open a GYM or Fitness Center in these locations with little to no competition. To end off this project, we utilized numerous Python libraries to fetch the information, control the content and break down and visualize those datasets. We have utilized Foursquare API to investigate the settings in neighborhoods of Toronto, get a great measure of data from Wikipedia which we scraped with the *Beautifulsoup* Web scraping Library. We also visualized utilizing different plots present in *seaborn* and *Matplotlib* libraries. Similarly, we applied AI strategy to anticipate the error given the information and utilized *Folium* to picture it on a map, We can utilize this venture to investigate any situation, for example, opening an alternate cuisine or opening of a Movie Theater and so forth. Ideally, this task acts as an initial direction to tackle more complex real-life problems using data science.

References

Wikipedia content: https://en.wikipedia.org/wiki/Toronto

CSV for Coordinate data: http://cocl.us/Geospatial_data

Foursquare API

Statistics Data on GYM:

https://www.ibisworld.com/canada/market-research-reports/gym-health-fitness-clubs-industry

https://www.statista.com/outlook/313/108/fitness/canada#market-users

https://www.cfa.ca/blog/new-year-new-franchise-opportunities-with-franchise-canada-2