Data-driven approach to find the best neighborhoods to build gyms in Toronto

Hamid Zaeimi

1 Abstract

The project's goal was to select the best neighborhoods to build gyms in Toronto by using free online data. The web scraping method was used to obtain information about each neighborhood's geographical coordinates, average income, and population. I then used K-means clustering (7) to find locations more suitable for investing in gyms. The neighborhoods were clustered into six groups, two of which had better conditions for this investment.

2 Introduction

The market size of the gym, health, and fitness club industry in Canada was predicted to be more than 4.3 billion U.S. dollars in 2020, and it was expected to increase to 4.59 billion U.S. dollars in 2021 [1]. Moreover, consumer trends and the proliferation of public health campaigns have increased the strength of Canada's Gym, Health and Fitness Clubs industry in recent years. Furthermore, the adult obesity rate is expected to rise. Therefore, the Public Health Agency of Canada has increasingly stressed the merits of fitness regimens and healthy lifestyle choices. Besides, health insurance companies and businesses seeking to improve workforce productivity and lower healthcare costs have provided incentives for health club memberships as a means of preventative care, buoying industry revenue growth [2].

The growing industry of fitness has attracted many investors. Although there is a great increase in fitness apps that helps people to exercise on their own, buying fitness or body-building equipment is not affordable for most people. The best option is purchasing a gym membership and exercise under the supervision of professionals. Therefore, investors are interested in building gyms for this growing demand, especially in cities like Toronto as the financial capital of a developed country because there is a positive relationship between the income of individuals and their attendance at the gyms. People in wealthier states get more exercise as they have more disposable income to buy a gym membership, sports clothes, and running shoes [3].

The location of a gym and other factors such as the professional instructors and gym equipment are essential for its success. The population of people living in an area (the more people, the more potential gym membership) and their income - due to the positive relationship between income and gym membership - are two factors affecting choosing the location of a gym. Although there are other factors like real estate price, the focus is on

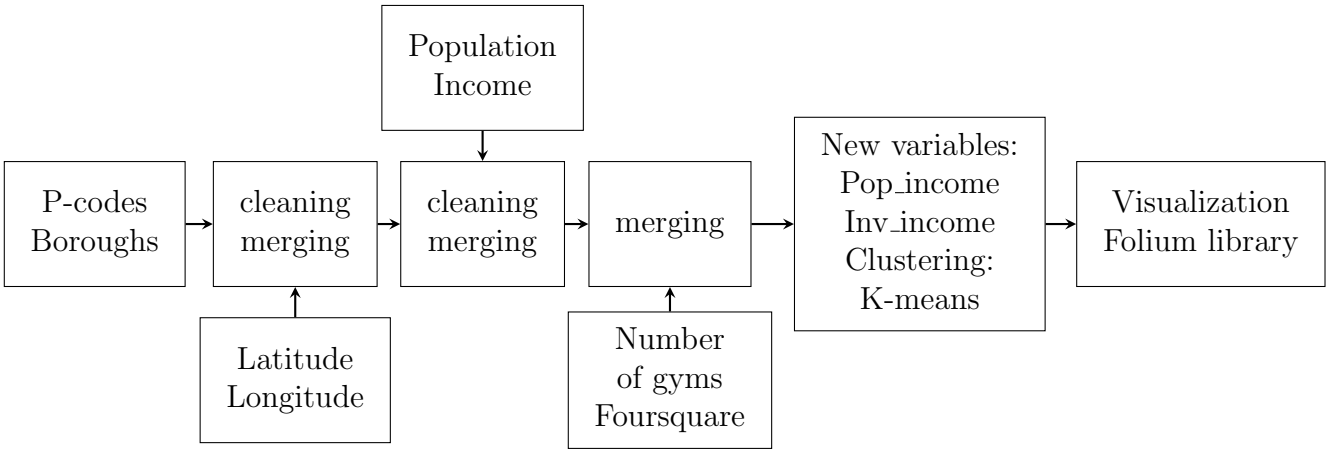

Figure 1. the schematic diagram of the project's methodology

the two mentioned factors in this project, and suggestions for a better location will be made based on the related information. In addition to these two factors, we must get the data about the number of existing gyms in each area because the greater number of existing gyms in an area makes it less attractive to invest in building a new gym there.

3 Data

The data such as information about postal codes of Canada, their latitude and longitude, the population of neighborhoods, and the average income of the neighborhood households - as a sign of neighborhood wealth - were needed. Besides, the data showing the number of existing gyms in each neighborhood is required since the more existing gyms, the less attractive the place for a new gym.

Four different sources were used to collect data. The first part was taken from a Wikipedia page, including Toronto's postal codes, Boroughs, and neighborhoods [4]. The second part of data was obtained from a downloaded CSV file that contained the latitudes and longitudes of 103 postal codes of Toronto [5]. Scraping the Canada Statistics website was used to get the population and average income of each neighborhood [6]. Then, exploring the gyms in each neighborhood (in an area with a defined radius around the geolocation of related postal codes) on the Foursquare website resulted in having the number of gyms in each one [7].

4 Methodology

The data containing postal codes, boroughs, and neighborhoods were collected from Wikipedia [4]. Then, the information got from the CSV file - the latitudes and longitudes of 103 postal codes of Toronto - was downloaded, checked, cleaned, and merged with the first data [5]. (Table 1)

The next step was to get the population (15 – 64 years old) – those more likely to use the gym - and the household average income for each neighborhood. For this purpose, the Canada Statistics website was scraped using the Beautiful Soup package [6]. Cleaning the

Table 1. Toronto data: neighborhoods; geographical coordinates

	Postal Code	Borough	Neighborhood	Lat.	Long.
0	M3A	North York	Parkwoods	43.753	-79.329
1	M4A	North York	Victoria Village	43.725	-79.315
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654	-79.360
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718	-79.464
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.667	-79.532
...	...	...	...	...	...
98	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North	43.653	-79.506
99	M4Y	Downtown Toronto	Church and YoWellesleyrk	43.665	-79.383
100	M7Y	East Toronto	Business reply mail Processing centre	43.662	-79.321
101	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea	43.636	-79.498
102	M8Z	Etobicoke	Mimico NW, The Queensway West	43.628	-79.520

Table 2. Toronto data: population and average income added

	Postal Code	Borough	Neighborhood	Lat.	Long.	Pop.	Average income
0	M3A	North York	Parkwoods	43.753	-79.329	23575	71966
1	M4A	North York	Victoria Village	43.725	-79.315	9405	60491
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654	-79.360	32905	63542
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718	-79.464	13490	62881
4	M9A	Etibicoke	Islington Avenue, Humber Valley Village	43.667	-79.532	24065	117148

Table 3. Toronto data: number of gyms added

	Postal Code	Borough	Neighborhood	Lat.	Long.	Pop.	Average income	Number of gyms
0	M3A	North York	Parkwoods	43.753	-79.329	23575	71966	0
1	M4A	North York	Victoria Village	43.725	-79.315	9405	60491	2
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654	-79.360	32905	63542	19
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718	-79.464	13490	62881	2
4	M9A	Etibicoke	Islington Avenue, Humber Valley Village	43.667	-79.532	24065	117148	0

data and merging the new information with our past data resulted in a new data frame, the first five rows of which are shown in Table 2.

By using the Foursquare API, the number of gyms in an area with a radius of 1000 meters around the location of each postal code were collected and saved [7]. Then, this new data were added to the previous one. The first five rows of the result are presented in Table 3. Preparing the data and clustering the neighborhoods was the next step. Before preparing the data, it was checked again. (Table 4)

As mentioned in the introduction part, the population of people living in a neighborhood and also the average income of its households can be factors affecting the suitability of the neighborhood for building a gym. Besides, the number of existing gyms in the neighborhood has a negative effect on the intention of an investor to provide a budget for a new gym. Therefore, two new variables were made, the first one: the product of population and average income (Pop_income) and the second one: the inverse of the number of gyms (Inv_gym) in the neighborhood. To reduce the effect of having zero or one gym, I added the average number of gyms to each neighborhood number of gyms and then calculated its inverse. These two

Table 4. Description of Toronto data

	Lat.	Long.	Population	Average income	Number of gyms
count	96	96	96	96	96
mean	43.708	-79.396	19863.489	88925.020	8.947
std	0.052	0.097	10075.901	41918.062	17.164
min	43.602	-79.594	1835	47960	0
25%	43.667	-79.464	12486.25	63854	0
50%	43.706	-79.392	18367.5	73658.5	2
75%	43.750	-79.340	25522.5	91918.25	7.25
max	43.836	-79.160	56475	263796	90

new variables were normalized and then used for clustering the neighborhoods. (Table 5)

Table 5. First 5 rows; Pop_income, Inv_gym

	Pop_income	Inv_gym
0	0.396403	1
1	0.096138	0.8
2	0.501380	0.253571
3	0.170519	0.8
4	0.695308	1

Before using the K-means algorithm to cluster neighborhoods, checking for the elbow point was tried, and finally, the K = 6 were chosen for the number of clusters. Then, the cluster labels were added to the data frame. A part of the result is provided in Table 6. Then, the results were visualized using the Folium library. (Figure 2)

Table 6. Toronto data with clusters

Postal Code	Cluster Label	Borough	Neighborhood	Latitude	Longitude	Population	Average income	Number of gyms	Pop_income	Inv_gym
M3A	1	North York	Parkwoods	43.753	-79.329	23575	71966	0	0.396	1
M4A	3	North York	Victoria Village	43.725	-79.315	9405	60491	2	0.096	0.8
M5A	5	Downtown Toronto	Regent Park Harbourfront	43.654	-79.360	32905	63542	19	0.501	0.25
M6A	3	North York	Lawrence Manor Lawrence Heights	43.718	-79.464	13490	62881	2	0.170	0.8
M9A	1	Etibicoke	Islington Avenue Humber Valley Village	43.667	-79.532	24065	117148	0	0.695	1

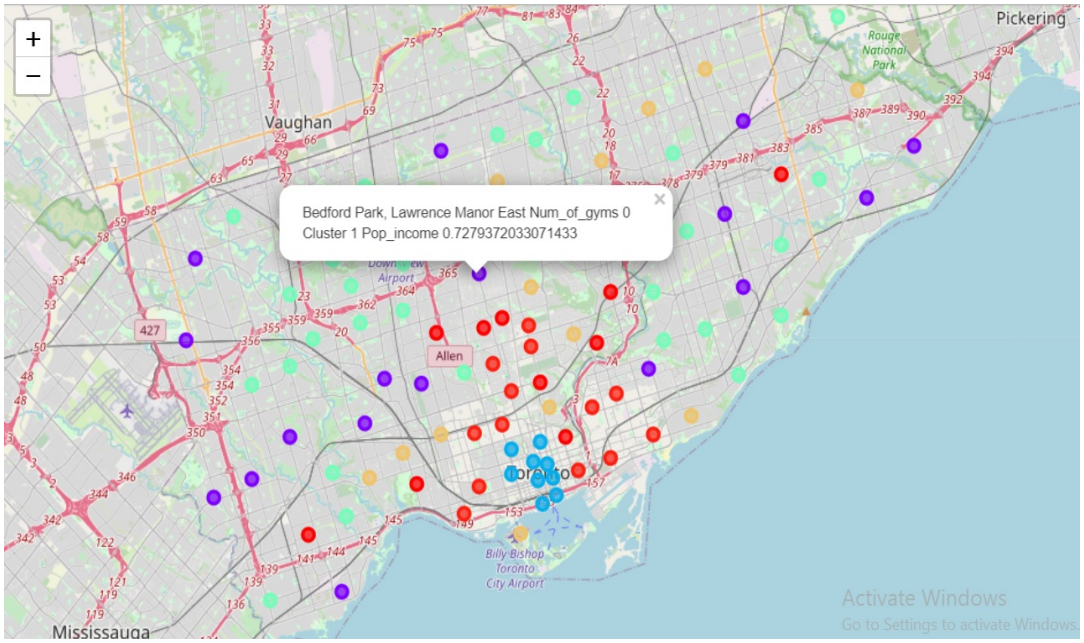

Figure 2. Toronto neighborhoods' clusters shown with different colors

5 Results

Visualizing the neighborhoods within each cluster by using the Python Folium library made distinguishing better ones easier. The following figures present each cluster with a particular color on the Toronto city map. A sample of every cluster is clicked on to show its characteristics such as cluster label, number of gyms, the product of population and average income, and the neighborhood name. Figures 3, 4, 5, 6, 7, and 8 represent examples belonged to clusters 0, 1, 2, 3, 4 and 5 respectively.

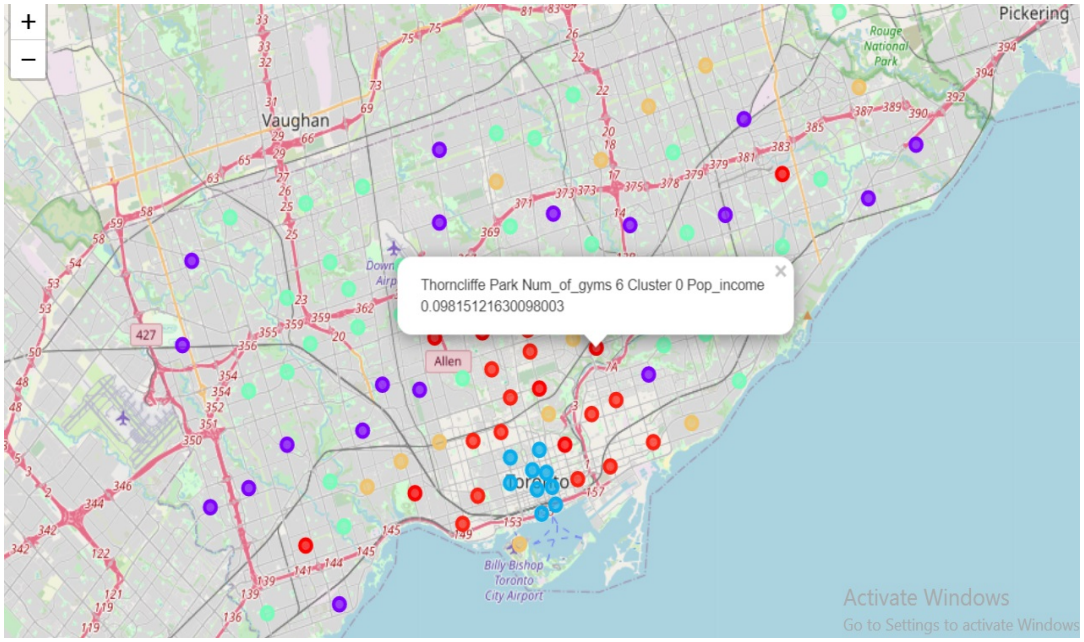

Figure 3. Cluster label 0: a moderate number of gyms; low Pop_income value

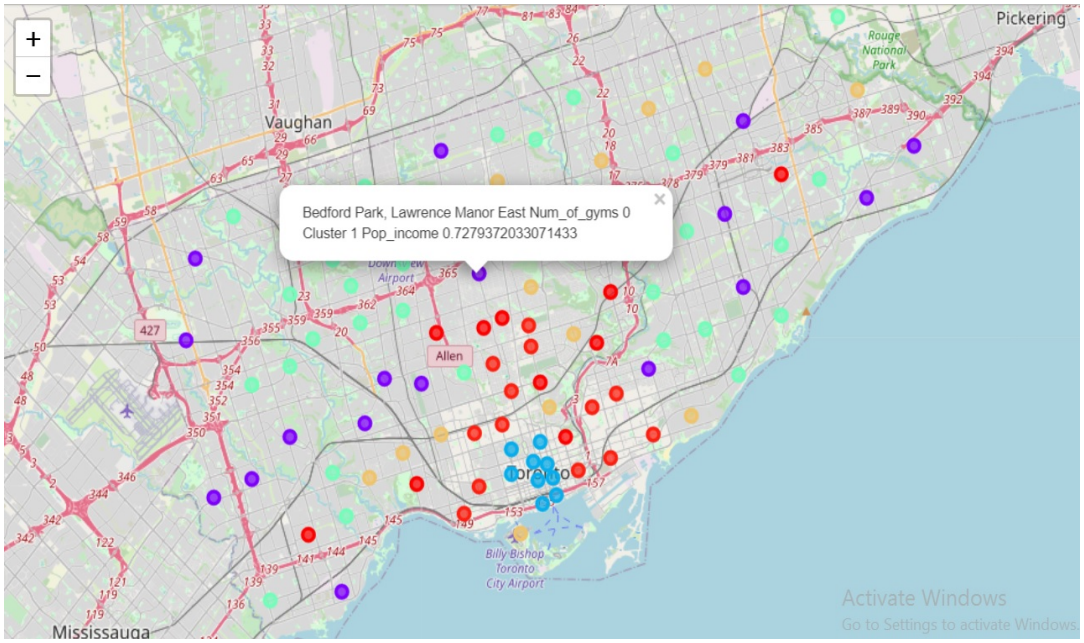

Figure 4. Cluster label 1: zero or a very low number of gyms; a moderate or high Pop_income value

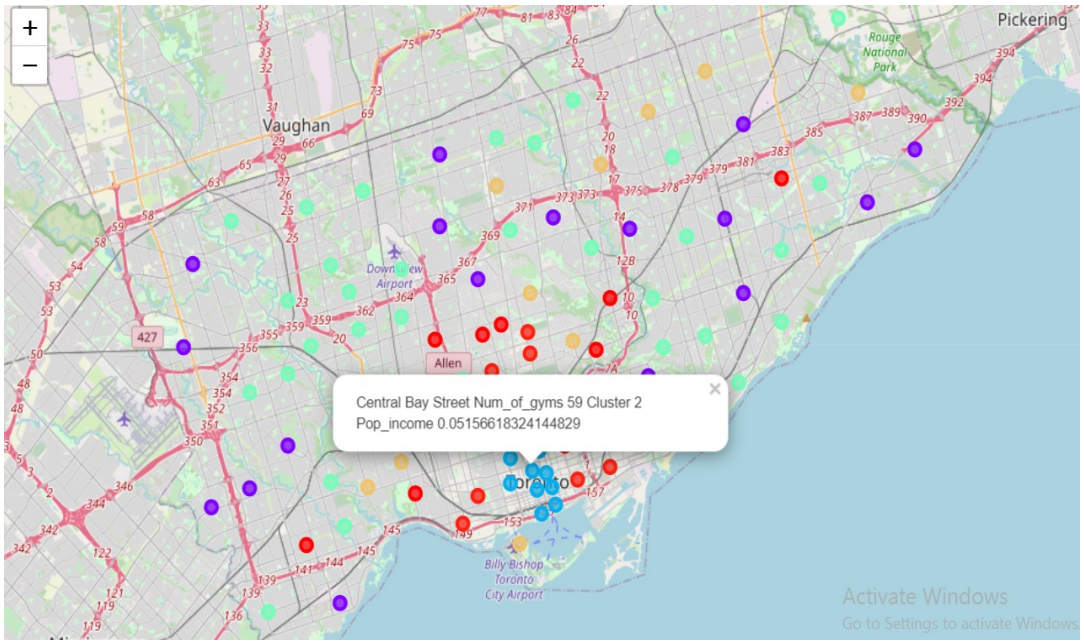

Figure 5. Cluster label 2: a very high number of gyms; low Pop_income value

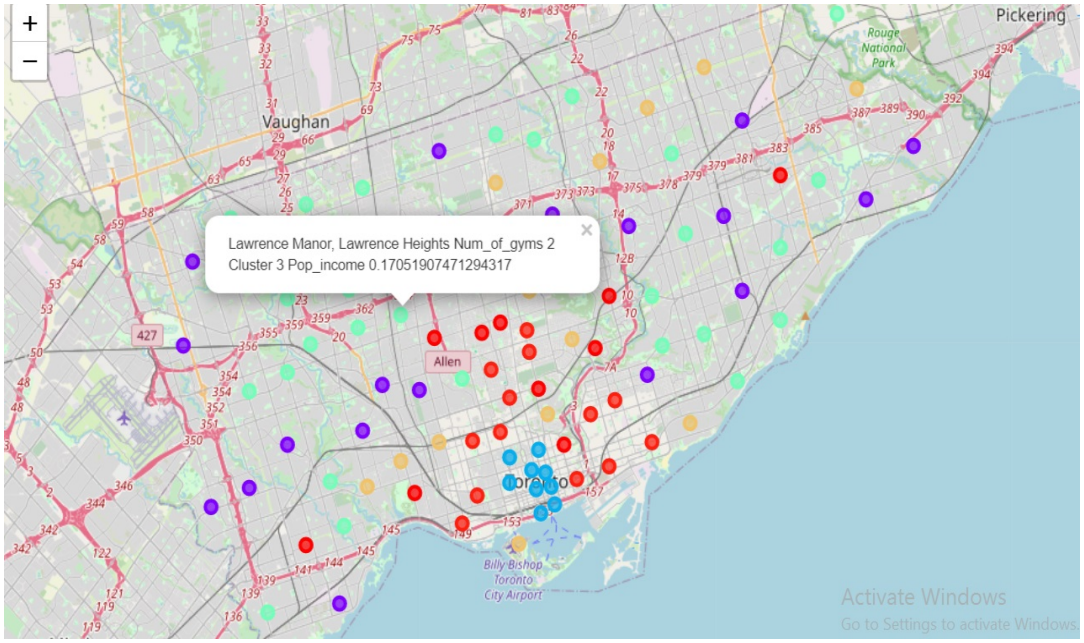

Figure 6. luster label 3: a low number of gyms; low Pop_income value

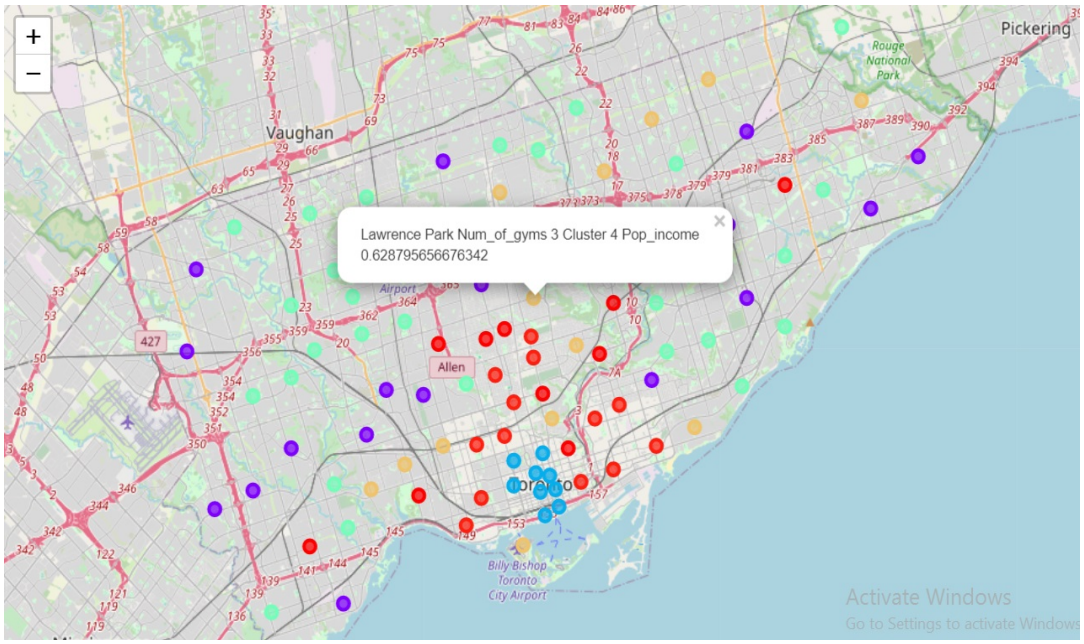

Figure 7. Cluster label 4: a low or moderate number of gyms; high or very high Pop_income value

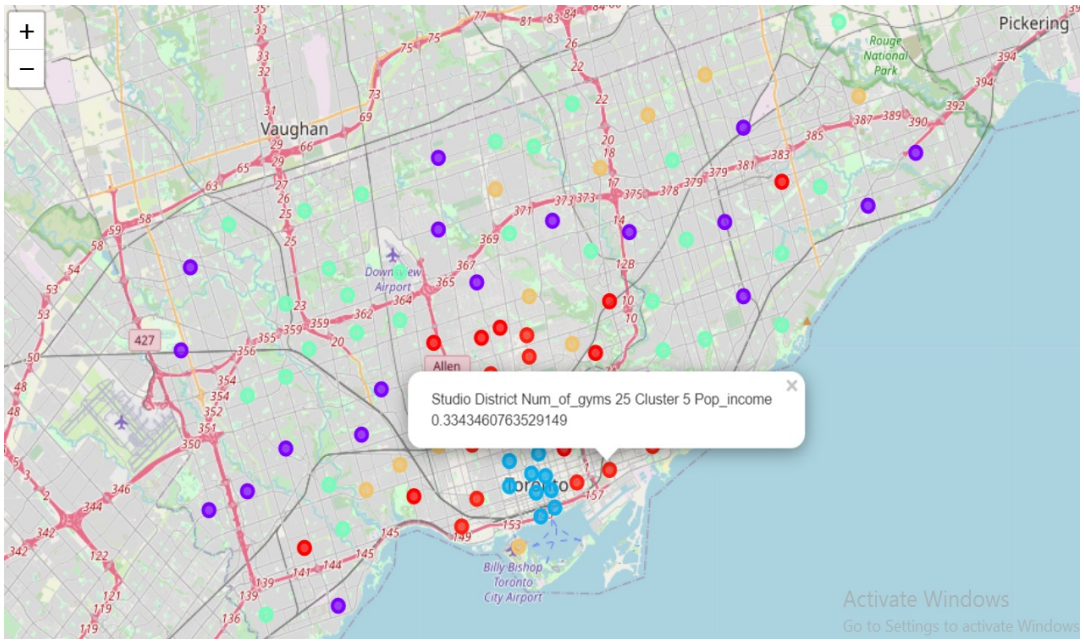

Figure 8. Cluster 5: above moderate number of gyms; a moderate or high Pop_income value

6 Discussion

Checking the final data frame and the map made by Folium, we can conclude that the neighborhoods with the clusters label 1 (dark blue circles) are more appropriate for building gyms. Neighborhoods in cluster 1 have zero or a very low number of gyms but moderate or high Pop_income value, the normalized product of population and average household income. A sample of cluster 1 neighborhoods can be seen in table 7.

Table 7. Cluster label 1

Postal Code	Cluster Label	Borough	Neighborhood	Latitude	Longitude	Population	Average income	Number of gyms	Pop_income	Inv-gym
M3A	1	North York	Parkwoods	43.753	-79.329	23575	71966	0	0.396	1
M9A	1	Etibicoke	Islington Avenue Humber Valley	43.667	-79.532	24065	117148	0	0.695	1
M9B	1	Etobicoke	west Dean park Princess Garden	43.650	-79.554	21765	97197	1	0.508	0.89
M1C	1	Scarborough	Rouge Hill Port Union	43.784	-79.160	23980	103716	0	0.607	1
M4C	1	East York	Woodbine heights	43.695	-79.318	32785	66633	0	0.526	1

The second best choice can be neighborhoods with cluster label 4, a sample of which can be seen in table 8. Neighborhoods in cluster 4 have a low or moderate number of gyms but a high or very high Pop_income value.

Therefore, neighborhoods in these two clusters are more suitable for building a new gym. However, there are other factors like real estate price that can affect the decision.

Table 8. Cluster label 4

Postal Code	Cluster Label	Borough	Neighborhood	Latitude	Longitude	Population	Average income	Number of gyms	Pop_income	Inv_gym
M1B	4	Scarborough	Malvern, Rouge	43.806	-79.194	45135	70929	2	0.797	0.8
M4E	4	East Toronto	The Beaches	43.676	-79.293	17295	107144	3	0.438	0.725
M4G	4	East York	Leaside	43.709	-79.363	12575	156052	5	0.467	0.607
M6H	4	West Toronto	Dufferin Dovercourt	43.669	-79.442	33880	68678	5	0.564	0.607
M2J	4	North York	Fairview Henry Farm	43.778	-79.346	40240	68092	3	0.674	0.725

7 Conclusion

This report used information about postal codes, their latitudes and longitudes, neighborhoods' population, and neighborhoods' average household income, alongside the data taken from the Foursquare website - location technology platform - to choose more suitable locations for investing in building new gyms. This report shows that by using just online resources - without the need to buy any data - accompanied by data science techniques, we can extract much helpful information that can help us make better decisions.

A K_means

K_means is a kind of unsupervised machine learning algorithm aiming at grouping data points based on the similarity of their features. The number of clusters - K - is predefined or chosen by methods like "Elbow point." The first step in this algorithm is initializing K centroids – randomly generated or selected from the provided data. The algorithm then iterates between two steps, which are data assignment and centroid updating. In the data assignment step, each data point is allocated to a group so that its distance from its centroid is minimum. Then, the K centroids are updated to be the mean of all points assigned to their clusters. The iterating between these two steps will continue until some stopping criteria such as having no change in each data point cluster or reaching the maximum iteration number are met.

References

[1] "fitness market size." [Online]. Available: https://bit.ly/3myxqpA

[2] "Gym, health & fitness clubs in canada - industry market research report," 2021. [Online]. Available: https://bit.ly/3iDqeaQ

[3] "wealth, fitness." [Online]. Available: https://wapo.st/3Blg8T4

[4] "Postal codes." [Online]. Available: https://bit.ly/3BmuYZQ

[5] "Geospatial coordinates." [Online]. Available: http://cocl.us/Geospatial_data

[6] "Canada statistics." [Online]. Available: https://bit.ly/3Dl9EEG

[7] "location technology platform." [Online]. Available: https://foursquare.com/