Finding the best neighborhoods in Toronto to open a new gym

Youssef Hosni

March 1, 2021

1. Introduction

1.1. Background

Toronto is the capital city of the Canadian province of Ontario. With a recorded population of 2,731,571 in 2016, it is the most populous city in Canada and the fourth most populous city in North America. The city is the anchor of the Golden Horseshoe, an urban agglomeration of 9,245,438 people (as of 2016) surrounding the western end of Lake Ontario, while the Greater Toronto Area (GTA) proper had a 2016 population of 6,417,516. Toronto is an international center of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world.

1.2. Problem

A businessman that owns a series of gyms/Fitness centers would like to expand his series to Toronto city in Canada he would like to know which neighborhoods in Toronto city are the best to open a new Gym/Fitness center in it. So, the main problem is finding the best neighborhood to open a new gym in Toronto city, Canada. The best neighborhood is the one that will lead to the highest profit and more clients when a new gym opens in it. Therefore, it will be important to find the neighborhoods that have a vacant to open new gym and to be sure that at the same time it will be able to attract customers to it. So, it will be important to study the demographics of the neighborhoods and to know which neighborhood had the highest population and the percentage of youth in this population, level of education, and the percentage of employers. The population with these properties is more expected to use the gym more, therefore more profit. Also, studying the venues in each neighborhood as the more activities in the neighborhood the more it will attract customers to the gym/training center as there will be a lot of pf services in this area in addition to the gym to attract customers. The number of gyms also is used as an important feature as it will be better to Focus on the neighborhoods that do not have gyms already existed in it.

1.3. Interest

The analysis would be helpful also for any service that is based on youth, such as cinemas, restaurants, shopping malls but some modifications are required. The analysis would be also helpful to healthcare providers because the neighborhoods that are best for gym/training centers would be the least needed for hospitals.

2. Data collection and cleaning

2.1. Data collection

The data used to solve this problem was obtained from the Toronto open data portal. The demographic dataset was downloaded from here, and the geographical data was downloaded from here and the venue information was obtained from the Foursquare API.

Commented [YH1]:

2.2 Data cleaning and feature extraction

2.2.1. Demographic dataset

The demographics dataset contains 1538 rows and 140 columns. The 1538 represents the demographics information for the 140 neighborhoods. The data that I found related to this problem is the total population, the youth age, the number of an educated population, and the number of employers. These data were selected only from the demographic dataset. The reason for choosing these data only is the relevance to the problem.

2.2.2. Geographical dataset

The geographical dataset was not available alone, but I found it in the crime rate dataset. First, I extracted the geographical information for each neighborhood. The latitudes and longitudes were saved as a list of strings, so they had first to be cleaned from the commas and brackets and then convert the list of strings into floats and finally to merge it with the demographic's dataset.

2.2.3. Venues information from Foursquare API

The Foursquare API was used to obtain information about the venues for each neighborhood. First, the venues for each neighborhood were obtained, the number of venues in each neighborhood was calculated, and the number of gyms in each neighborhood. The first feature is important as it indicates the business activity of each neighborhood so the more venues in the neighborhood the more it is active, and this means that it attracts more customers and is expected to make more profit if I open a new gym in it. The second feature is also important as it indicates the vacant of opening a new gym, so if in a certain neighborhood three or four gyms it will not be probably a good neighborhood, except if the neighborhood is very large with a large population.

The final dataset consists of 140 rows * 10 features, the 10 features are the Neighborhoods, Total population, number of educated people, youth 15-45, number of employers, latitudes and longitudes, number of gyms, and number of venues. The final data is as shown in figure 1.

	Neighborhood	Total population	number of educated people	number of 15-45	number of employers	long_latt	number_gyms	number_venues
0	Agincourt North	30280.0	19805.0	11850.0	13230.0	[-79.2816161258827, 43.797405754163]	0.0	26.0
1	Agincourt South-Malvern West	21990.0	14535.0	8840.0	9860.0	[-79.2891688527481, 43.7851873380096]	0.0	34.0
2	Alderwood	11900.0	7915.0	4520.0	6240.0	[-79.5532040267975, 43.5954996876866]	1.0	17.0
3	Annex	29180.0	23495.0	15095.0	16770.0	[-79.4121466573202, 43.6744312990078]	3.0	63.0
4	Banbury-Don Mills	26910.0	20555.0	9615.0	13030.0	[-79.326504539789, 43.7325704244428]	2.0	14.0

Figure 1. The first 5 rows of the final dataset will be used after cleaning the datasets and merging them.

2.3 Data exploration

First to visualize the neighborhood on the map, the folium library and is as shown in figure 2.

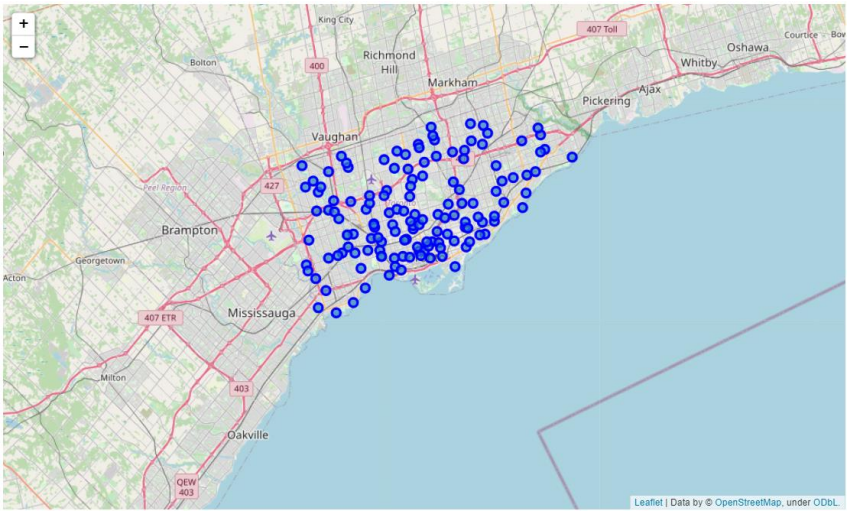


Figure 2. The neighborhoods of the city Toronto on the map.

To calculate the properties of the features, the numerical features were first converted from object type to int or float. The dataset properties are shown in the table1.

Table 1. The dataset properties.

	Total population	number of educated people	number of 15-45	number of employers	number_gyms	number_venues
count	140.000000	140.000000	140.000000	140.000000	140.000000	140.000000
mean	18676.928571	12801.250000	8102.035714	9049.357143	0.914286	40.328571
std	9099.209342	6606.830919	4541.999603	4631.833115	1.322118	29.377582
min	6490.000000	3585.000000	2805.000000	2790.000000	0.000000	4.000000
25%	11851.250000	8052.500000	5060.000000	5916.250000	0.000000	18.000000
50%	16367.500000	11290.000000	6822.500000	7595.000000	0.000000	27.500000
75%	22410.000000	16352.500000	9885.000000	10930.000000	1.000000	59.250000
max	53350.000000	39080.000000	29695.000000	31375.000000	7.000000	100.000000

The distribution of each of the features is shown on the histograms shown in figure 3. From the figure, we can see that the number of gyms is zero in 70 of the neighborhoods which opens an opportunity to open new gyms, especially in the neighborhoods with a large population and venues number.

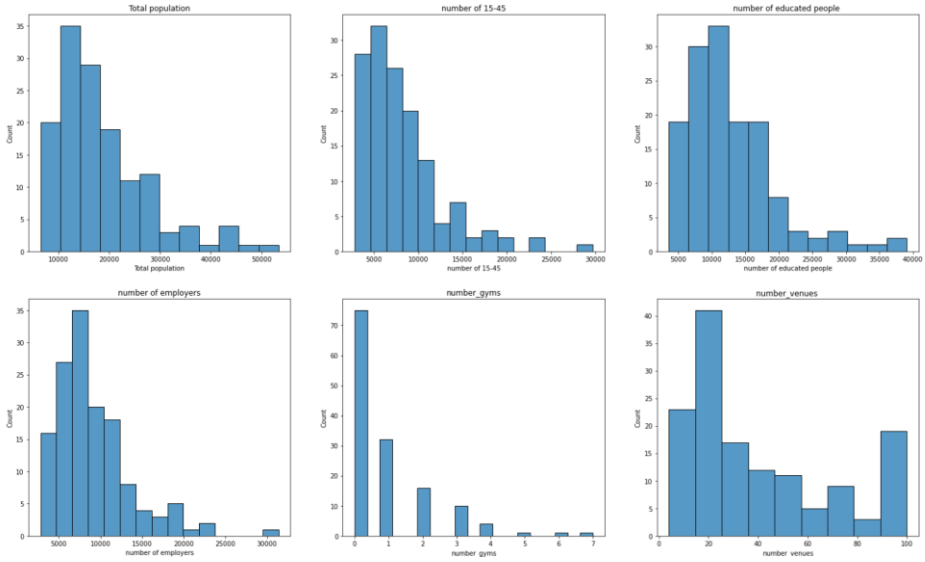
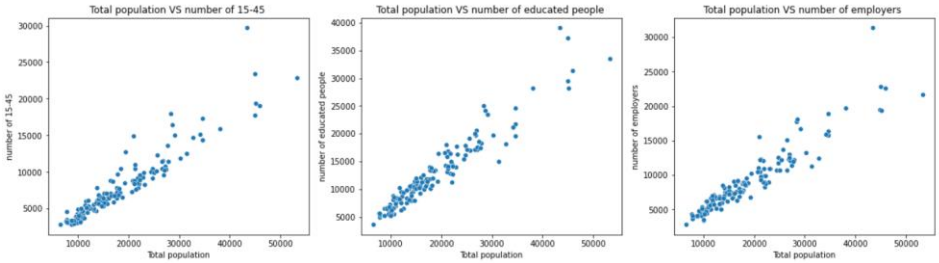


Figure 3. The distribution of each feature of the dataset.

The relation between the features were examined by plotting scatter plots between the different variables, first between the Total population and the demographics features as shown in figure 4.



The people aged 15-45 and the number of educated people and employers are linearly correlated, and there is an increase in all the three features between the values of 40000-50000, so it is important to see the effect of this on the number of venues and number of gyms within this range.

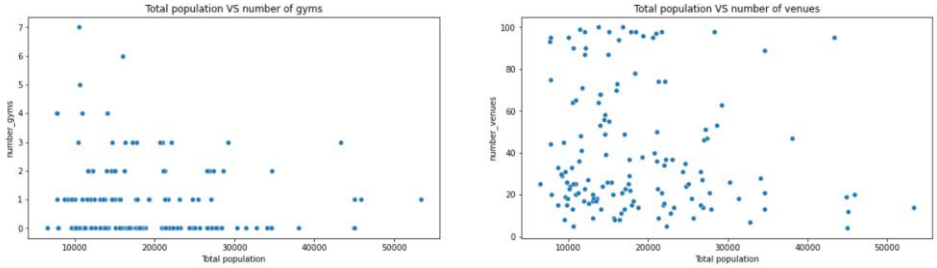


Figure 4. The number of gyms and number of venues changes with the total population.

From figure 4 we can observe the change in the number of venues and number of gyms with the total population. The number of gyms and venues is independent of the total population. Most of the gyms and venues are in the range of 10,000-40,000, as most of the neighborhoods are in this range. The number of gyms and venues increased in the range of 40,000-50,000 and this is as expected.

The relation between the number of gyms and the rest of the features is as shown in figure 5.

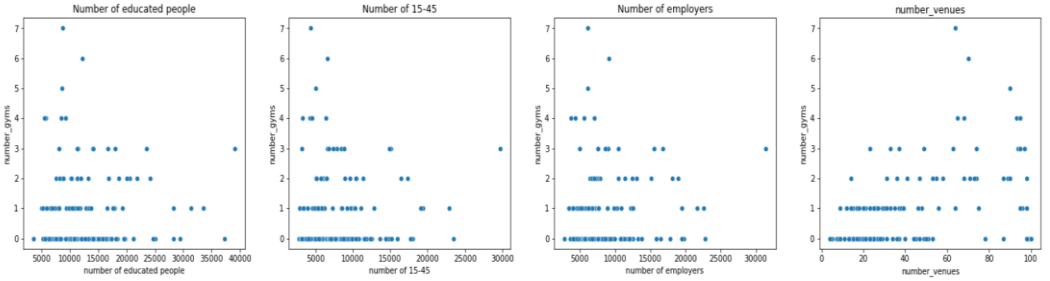


Figure 5. The relation between the number of gyms and the rest of the features.

The number of the gyms is dependent on the number of venues more than the other of the features, the number of gyms increases with the number of venues and is uniformly distributed with the rest of the features.

3. Methodology

Since the main problem is finding the best neighborhoods to open a new gym, so this problem can be seen as an unsupervised learning problem. Unsupervised learning is a type of algorithm that learns patterns from untagged data. The hope is that through mimicry, the machine is forced to build a compact internal representation of its world. Since we try to group the neighborhoods that are the best to open a new gym together this method can be seen as a clustering problem. K-means clustering method is used to cluster the data. k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

3.1 Data preprocessing

The data were normalized using the min-max normalization. This is an important step because the k-means algorithms depend on distance measurement, so it is important that the data used be in a similar scale. The formula of the min-max scaler is as the following: $\frac{feature - \min(feature)}{\max(feature) - \min(feature)}$

The neighborhood and the geographical data were dropped from the data as they will be used by the clustering algorithm.

3.2. K-means clustering

The best k was found using the elbow method, in which the average distance from the clusters is calculated for different values of k and the best k is the k at the elbow as shown in figure 6. The best k was found to be 3.

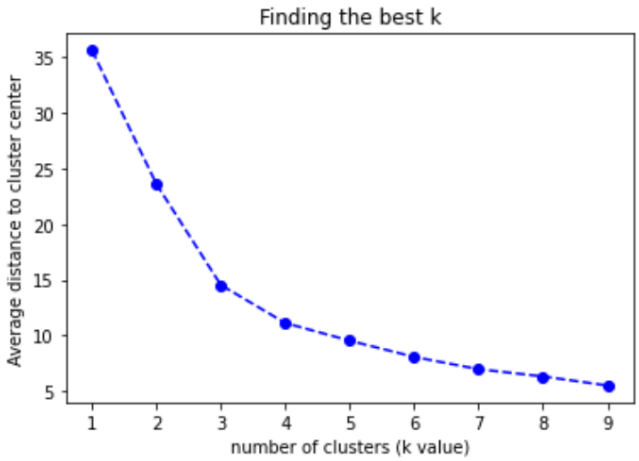


Figure 6. Finding the best k values using the elbow method.

4. Results

The neighborhoods are clustered into three clusters as shown in figure 7. The red color is the first cluster, the violet is the second cluster, green is the third cluster.

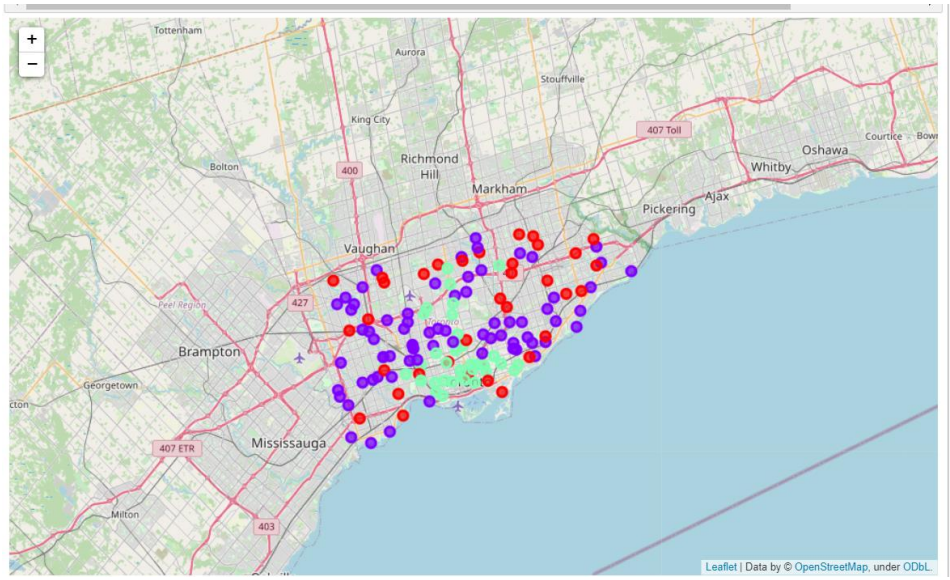


Figure 7. Neighborhoods are clustered into four clusters.

The properties of neighborhoods in each cluster are shown in the tables below.

Table 2. The properties of the first cluster neighborhoods

	Total population	number of educated people	number of 15-45	number of employers	number_gyms	number_venues	Labels
count	34.000000	34.000000	34.000000	34.000000	34.000000	34.000000	34.0
mean	31538.382353	21854.558824	14057.647059	15292.352941	0.676471	32.970588	0.0
std	7958.399672	6294.231651	4770.428676	4651.130888	0.944541	24.274298	0.0
min	21135.000000	15020.000000	8450.000000	10715.000000	0.000000	4.000000	0.0
25%	26550.000000	17437.500000	10406.250000	12125.000000	0.000000	15.750000	0.0
50%	28107.500000	19632.500000	12460.000000	13130.000000	0.000000	25.500000	0.0
75%	34631.250000	24517.500000	16357.500000	18012.500000	1.000000	46.750000	0.0
max	53350.000000	39080.000000	29695.000000	31375.000000	3.000000	98.000000	0.0

Table 3. The properties of the second cluster neighborhoods

	Total population	number of educated people	number of 15-45	number of employers	number_gyms	number_venues	Labels
count	71.000000	71.000000	71.000000	71.000000	71.000000	71.000000	71.0
mean	14469.859155	9503.732394	5832.183099	6679.788732	0.450704	23.309859	1.0
std	4531.205451	3063.223409	1995.527451	2010.455960	0.751926	11.218087	0.0
min	6490.000000	3585.000000	2805.000000	2790.000000	0.000000	5.000000	1.0
25%	10532.500000	6765.000000	4007.500000	4995.000000	0.000000	15.000000	1.0
50%	13535.000000	9600.000000	5550.000000	6655.000000	0.000000	21.000000	1.0
75%	17662.500000	11625.000000	7097.500000	8107.500000	1.000000	29.500000	1.0
max	23185.000000	16330.000000	11035.000000	11045.000000	3.000000	53.000000	1.0

Table 4.The properties of the third cluster neighborhoods

	Total population	number of educated people	number of 15-45	number of employers	number_gyms	number_venues	Labels
count	35.000000	35.000000	35.000000	35.000000	35.000000	35.000000	35.0
mean	14717.285714	10695.857143	6921.142857	7791.571429	2.085714	82.000000	2.0
std	4028.120337	3383.702394	2527.695732	2461.997566	1.788385	15.835552	0.0
min	7655.000000	4980.000000	3240.000000	3450.000000	0.000000	49.000000	2.0
25%	11877.500000	8505.000000	5242.500000	6675.000000	1.000000	69.000000	2.0
50%	14610.000000	10345.000000	6380.000000	7515.000000	2.000000	87.000000	2.0
75%	17315.000000	12092.500000	8145.000000	8950.000000	3.000000	96.500000	2.0
max	22080.000000	18010.000000	14920.000000	15535.000000	7.000000	100.000000	2.0

The feature distribution for the neighborhoods in each cluster is shown in figure 8-12.

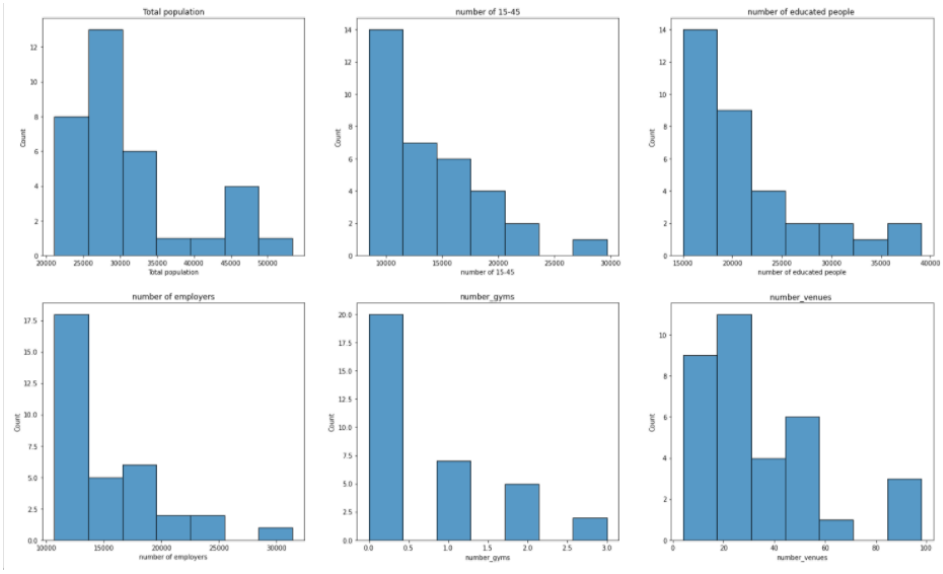


Figure 8. The feature distribution of the neighborhoods in the first cluster.

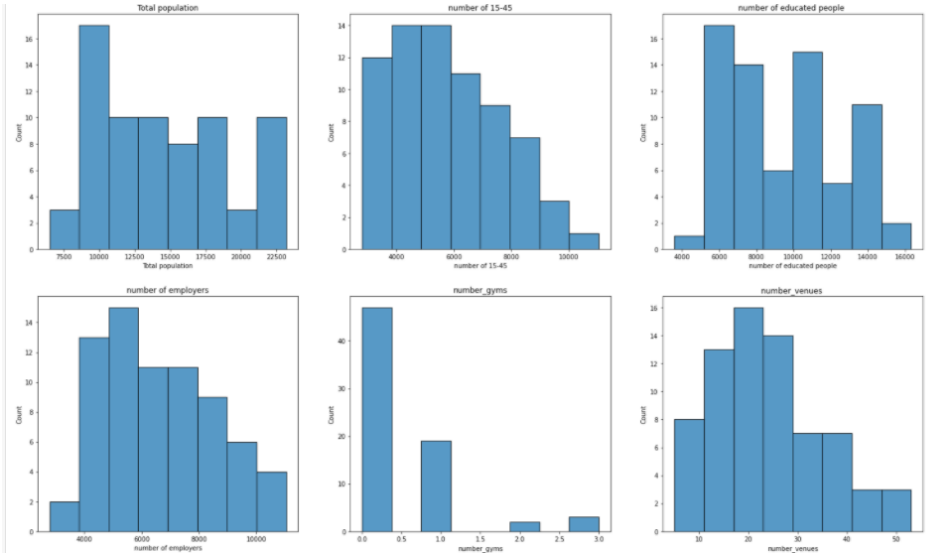


Figure 9. The feature distribution of the neighborhoods in the *second* cluster.

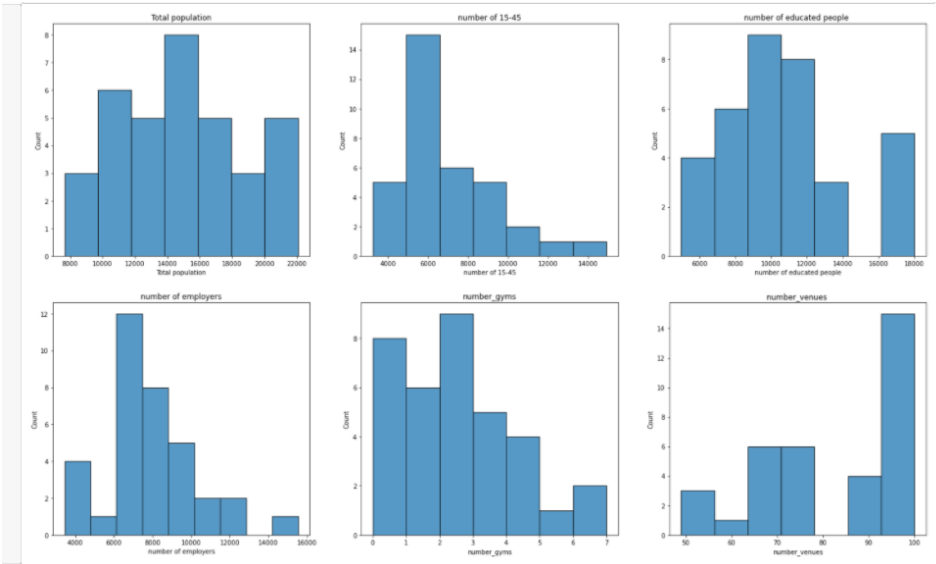


Figure 10. The feature distribution of the neighborhoods in the third cluster.

Observing the properties of the neighborhood of each of the resulted clusters, the thirty-five neighborhoods in the first cluster has large total population compared to the second and third cluster with mean of 31,538 and most of the neighborhood are in the range 20,000 to 35,000 as shown in figure

8. The number of the 15-45 population in the first cluster is also large compared to the rest of the clusters, with an average value of 14,057 and within the range of 10,000 to 20,000. The same is with the number of educated population and employer populations. Regarding the number of gyms 20 neighborhoods have no gyms in it and 6 of them have one gym, and 5 have 2 gyms and 3 have 3 gyms. The number of venues in this cluster neighborhood is in the range of 10 to 40, which is between the second and third cluster neighborhoods.

The second cluster contains 71 neighborhoods, the mean of the total population is 14,500 and most of the neighborhoods fall in the range of 10,000 to 22,500, this is similar to the third cluster, but the numbers are smaller than the first cluster. The educated population in this cluster falls in the range of 5000 to 14000 with the mean value of 9503 which is again similar to the neighborhoods of the third cluster, but smaller than the first cluster. The number of 15-45 population of the neighborhoods of the second cluster also fall in the range of 3000 to 9000 and with the mean value of 5832 which is also similar to the neighborhoods of the third cluster and smaller than that of the neighborhoods of the first cluster. The number of gyms in the neighborhoods is as the following: 47 neighborhoods have 0 gyms and 19 neighborhoods have 1 gym and 2 neighborhoods have 2 gyms and 3 neighborhoods have 3 gyms. The number of venues is in the range of 5 to 40 and a mean of 23.3, which is the least compared to the first and third clusters.

The total population of the thirty-five neighborhoods in the third cluster is in the range of 8000 to 22000, with a mean value of 14,717. This is similar to the second cluster and smaller than the neighborhoods of the first cluster. The population with 15-45, educated and employed population are in a similar range as the neighborhoods of the second cluster. The number of gyms in this cluster neighborhood is the highest compared to the first and second clusters neighborhoods. The number of neighborhoods with the corresponding number of gyms is shown in the table below.

Table 5. The number of neighborhoods and their corresponding number of gyms.

Number of gyms	0 gyms	1 gym	2 gyms	3 gyms	4 gyms	5 gyms
Number of neighborhoods	8	6	9	5	4	1

The number of venues is in the range of 65 to 100, with a mean value of 82, which is quite here compared to the neighborhoods of the first and second clusters. This agrees with what was expected and observed before during data exploration that the number of gyms is correlated with the number of venues.

5. Discussion

Using K-means clustering the neighborhoods are clustered into four clusters based on the similarity of the features. The features used were the total population, the 15–45 population, the educated population, the employed population, the number of gyms, and the number of venues for each neighborhood.

The number of gyms was found to be correlated with the number of venues, and this can be seen obviously with the neighborhoods of the third clusters. Although these neighborhoods do not have the highest population, and youth and adults but they have the highest number of gyms due

to the presence of a lot of venues around. Therefore, it seems that the gym owners were concentrating on the areas with a lot of venues regardless of the population of this neighborhood. The neighborhoods of the first cluster are the ones with the highest population and with moderate numbers of venues, but due to this fact, there are no a lot of gyms in the neighborhoods of these clusters.

The neighborhoods that are the best to open a new gym, can be found by following one of the two following approaches. The first is selecting the neighborhoods from the third cluster that have only one gym or no gyms and select the one with the highest number of venues. Following this approach, the best neighborhood will be the "Trinity-Bellwoods" neighborhood as it contains 100 venues, and the population is higher than the average. Other options will be "Palmerston-Little Italy", "High Park-Swansea" neighborhoods. The best one of them can be determined by some other factors, such as the cost of buying new land or renting a building, the availability of vacant land or building in this neighborhood.

The second approach is to find the neighborhood with a higher number of venues and has no gyms or only one gym in the neighborhoods of the first cluster. "Church-Yonge Corridor**"** has no gyms and has 98 venues, and it has a high population of 28345, which makes it a very good choice, compared with what was found with the first approach. Therefore, the best four neighborhoods are "Church-Yonge Corridor**",** "Trinity-Bellwoods", "Palmerston-Little Italy" and "High Park-Swansea"

6. Conclusion

Using the demographics data and venue information for each neighborhood obtained from Foursquare API, I was able to cluster the neighborhoods into three clusters using the K-means clustering algorithm. The number of gyms was found to be correlated to the number of venues, and the neighborhoods with a large number of venues and gyms are clustered into the third cluster, so the most suitable neighborhood out of this cluster is the "Trinity-Bellwoods" neighborhood. The first cluster contains a neighborhood with a large population and a small number of gyms and a moderate number of venues. The "Church-Yonge Corridor" neighborhood is the best choice out of this cluster as it contains 98 venues and a large population. The number of venues is almost similar to that of the "Trinity-Bellwoods" and the population is double of it, making it the best neighborhood to open a new gym in Toronto city.

It is important in the future to study the type of venues that are related to the number of gyms and also the price of the land and the rents of the commercial building in the neighborhoods in the first and third clusters to be able to select the most suitable neighborhood.