CAPSTONE PROJECT

TABLE OF CONTENTS

1. INTRODUCCION (BUSINESS PROBLEM)

Sport is health, both physical and emotional. It also allows one to look good. There are many options to do it, they can be running, cycling, weights or the increasingly popular crossfit. Every year more and more people start an exercise plan. Thus, it seems to be a good idea to open a gym and as the business grows, why not a chain of them.

Toronto is the most populous city in Canada as well as the financial center of the country, and therefore it seems to be a good option to open a gym in that city.

Thus, in this project, we will see to find the best area to open a gym.

Interested parties are therefore those gym chains seeking to expand in Torontó or entrepreneurs who decide to open a gym there.

2. DATA

We need to know where the gyms are located in Toronto. For this we must know the neighborhoods of Toronto, their respective geographical locations and the gyms that are in them.

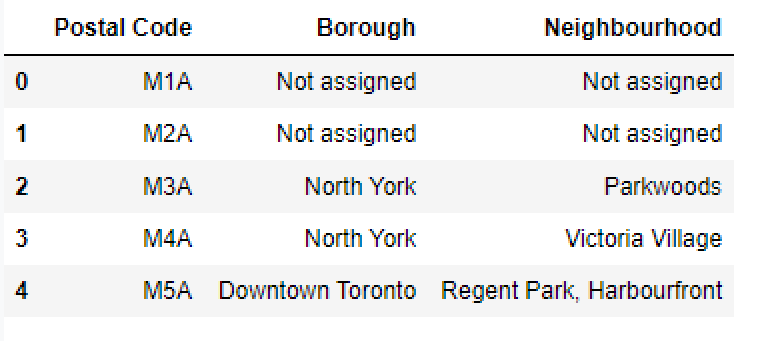
The necessary data as well as from where it will be obtained are detailed below:

- Neighborhoods in Canada
  - Data for boroughs and neighborhoods in Canada.
  - They are obtained by scrapping the page https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- Geospatial location of neighborhoods in Canada
  - Latitude and longitude of neighborhoods in Canada.
  - Obtained from http://cocl.us/Geospatial_data.
- Venue data.
  - Event venues in each neighborhood.
  - Obtained through the Foursquare API.

3. METHODOLOGY

3.1. Data-preprocessing

We start by scrapping the wikipedia page, obtaining the following dataframe:



	Postal Code	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront

Figure 1. Postal codes dataframe.

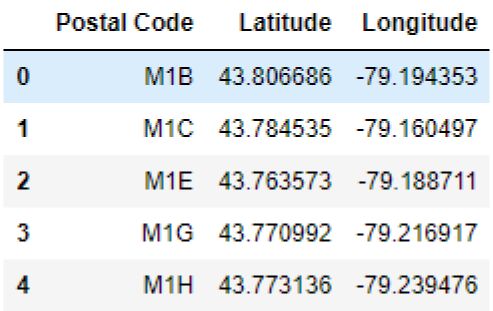
It consists of Canada's zip code, district, and neighborhoods by district.

We process it in the following way:

- If a district has the value "Not assigned" the row will be removed from the df.
- If a neighborhood has the value "Not assigned", its value is set equal to that of the district.

It is allowed to have more than one neighborhood per district.

Next we obtain the geospatial data of the neighborhoods of Canada:

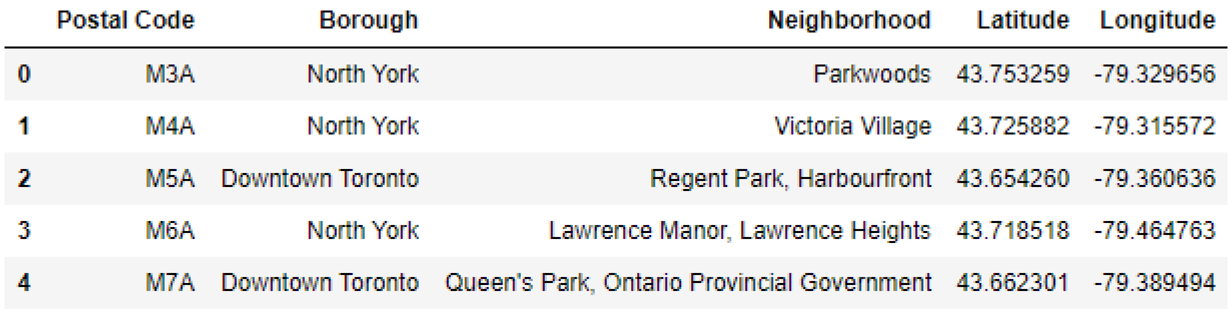


	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Figure 2. Latitude and longitude per postal codes.

It contains the postal code, latitude and longitude.

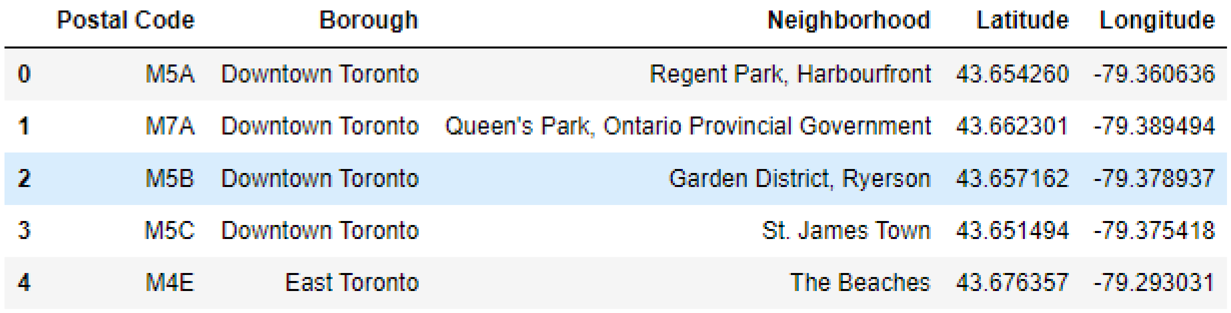
We combine both dataframes into one using the postal code:



	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

Figure 3. Combining dataframes.

As in this project we are only interested in Toronto data, we eliminate all districts that do not contain the word Toronto, leaving our final dataframe as follows:



	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
1	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
2	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937
3	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418
4	M4E	East Toronto	The Beaches	43.676357	-79.293031

Figure 4. We are left with only the Toronto data.

3.2. Exploratory analysis

We start by creating a map to see Toronto's neighborhoods:

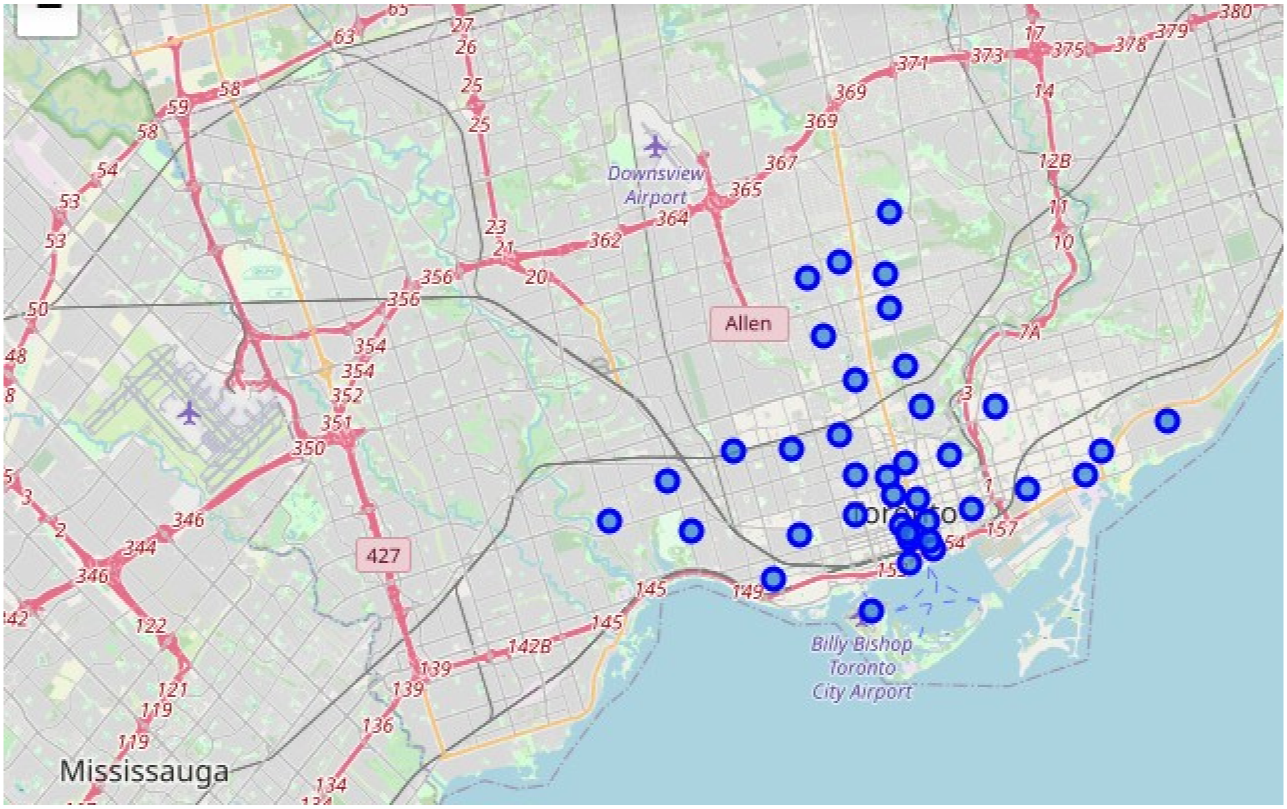
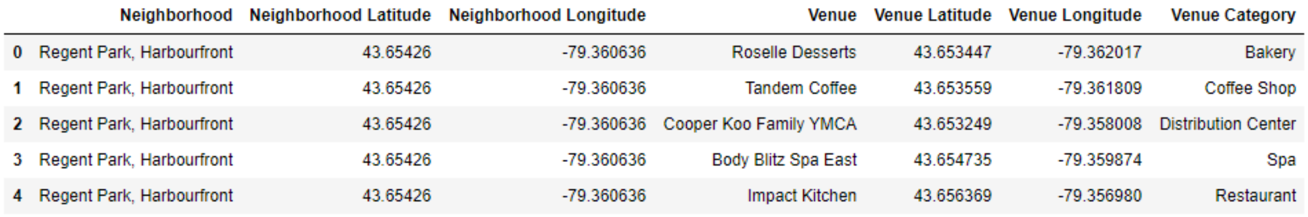


Figure 5. Toronto's neighborhoods.

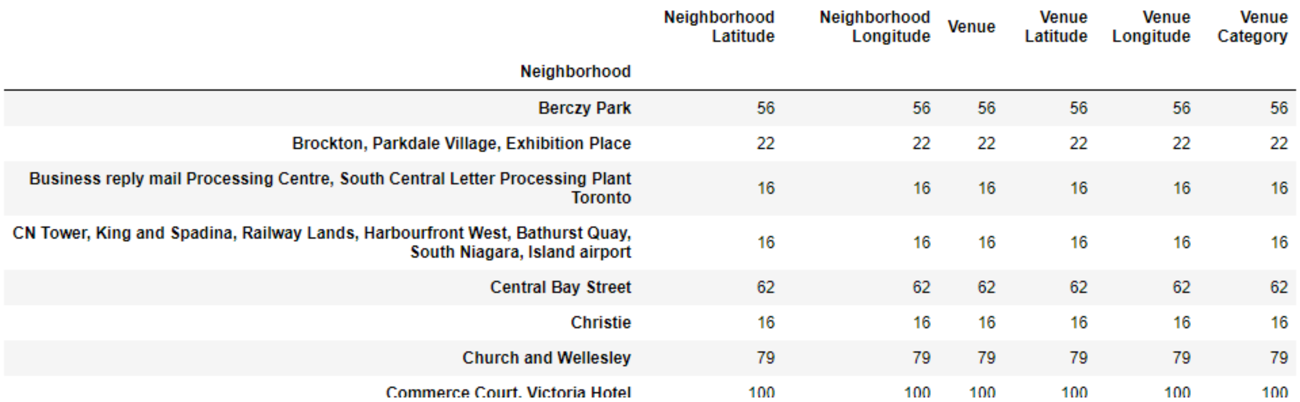
We then use the Foursquare API to see the event venues, their location and type in the various neighborhoods:



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
2	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center
3	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa
4	Regent Park, Harbourfront	43.65426	-79.360636	Impact Kitchen	43.656369	-79.356980	Restaurant

Figure 6. Using the Foursquare API.

and we count how many venues were returned per neighborhood:

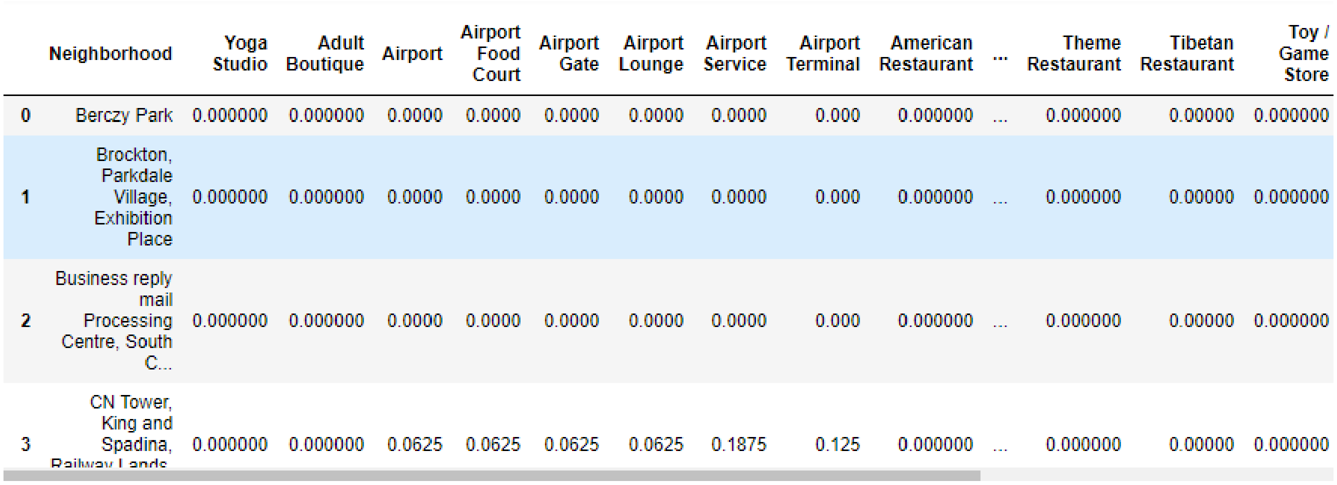


Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Berczy Park	56	56	56	56	56	56
Brockton, Parkdale Village, Exhibition Place	22	22	22	22	22	22
Business reply mail Processing Centre, South Central Letter Processing Plant Toronto	16	16	16	16	16	16
CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport	16	16	16	16	16	16
Central Bay Street	62	62	62	62	62	62
Christie	16	16	16	16	16	16
Church and Wellesley	79	79	79	79	79	79
Commerce Court, Victoria Hotel	100	100	100	100	100	100

Figure 7.Amount of venues return per neighborhood.

Now we go on to analyze each neighborhood.

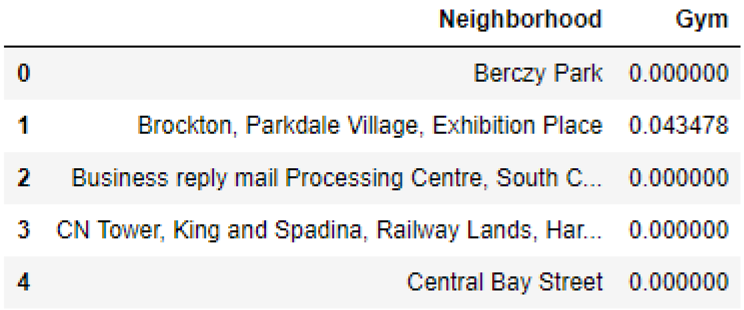
First we use the one-hot encoding technique and then we group by neighborhood and taking the average of each category.



	Neighborhood	Yoga Studio	Adult Boutique	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	...	Theme Restaurant	Tibetan Restaurant	Toy / Game Store
0	Berczy Park	0.000000	0.000000	0.0000	0.0000	0.0000	0.0000	0.0000	0.000	0.000000	...	0.000000	0.00000	0.000000
1	Brockton, Parkdale Village, Exhibition Place	0.000000	0.000000	0.0000	0.0000	0.0000	0.0000	0.0000	0.000	0.000000	...	0.000000	0.00000	0.000000
2	Business reply mail Processing Centre, South C...	0.000000	0.000000	0.0000	0.0000	0.0000	0.0000	0.0000	0.000	0.000000	...	0.000000	0.00000	0.000000
3	CN Tower, King and Spadina, Railway Lands	0.000000	0.000000	0.0625	0.0625	0.0625	0.0625	0.1875	0.125	0.000000	...	0.000000	0.00000	0.000000

Figure 8. One-hot encoding

In this way we obtain the mean of each type of place grouped by neighborhood. As in this project we are only interested in gyms, we will only keep this column.



	Neighborhood	Gym
0	Berczy Park	0.000000
1	Brockton, Parkdale Village, Exhibition Place	0.043478
2	Business reply mail Processing Centre, South C...	0.000000
3	CN Tower, King and Spadina, Railway Lands, Har...	0.000000
4	Central Bay Street	0.000000

Figure 9. We only keep the data of the gyms.

3.3. Machine Learning

Our problem is one of clustering, we work with unlabeled data and we seek to find a structure in it.

Among the clustering algorithms, one that stands out for its efficiency and simplicity is K-means.

So then, let's use K-means to find the clusters of gym locations in Toronto.

K-means requires the value of K to be provided, for this we use the elbow technique, implemented through the KelbowVisualizer tool:

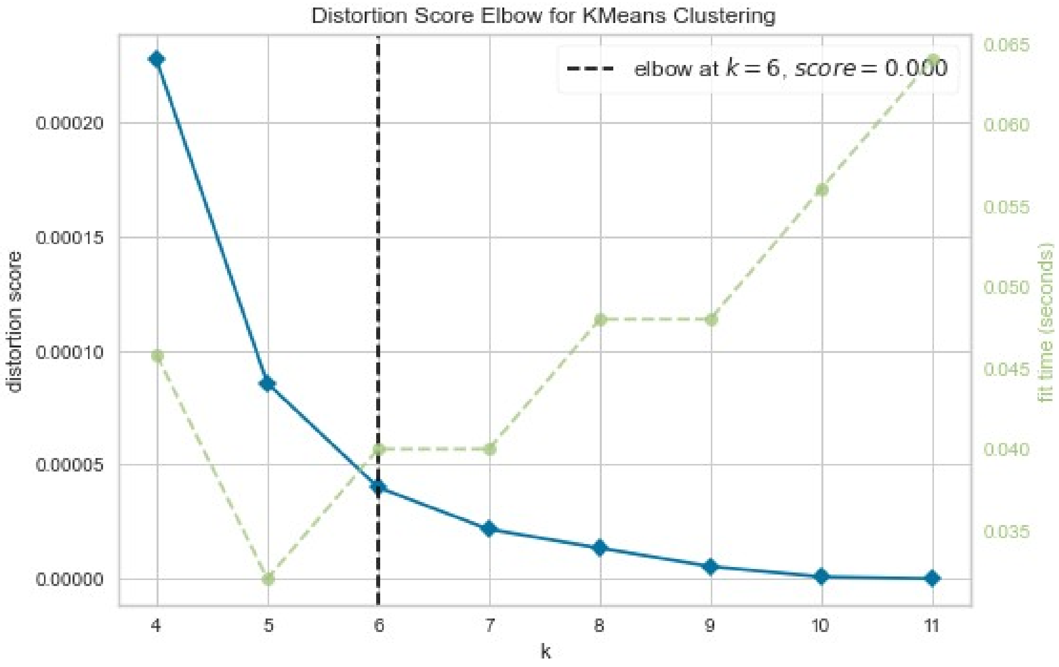


Figure 10. Elbow method.

From the graph above we see that the optimal value of k is 6. Now we apply K-means with k = 6:

```
# import
from sklearn.cluster import KMeans
# number of clusters
kclusters = 6
# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(gym_df_clustering)
gym_df.insert(0, 'Cluster Labels', kmeans.labels_)
```

We visualize the clusters:

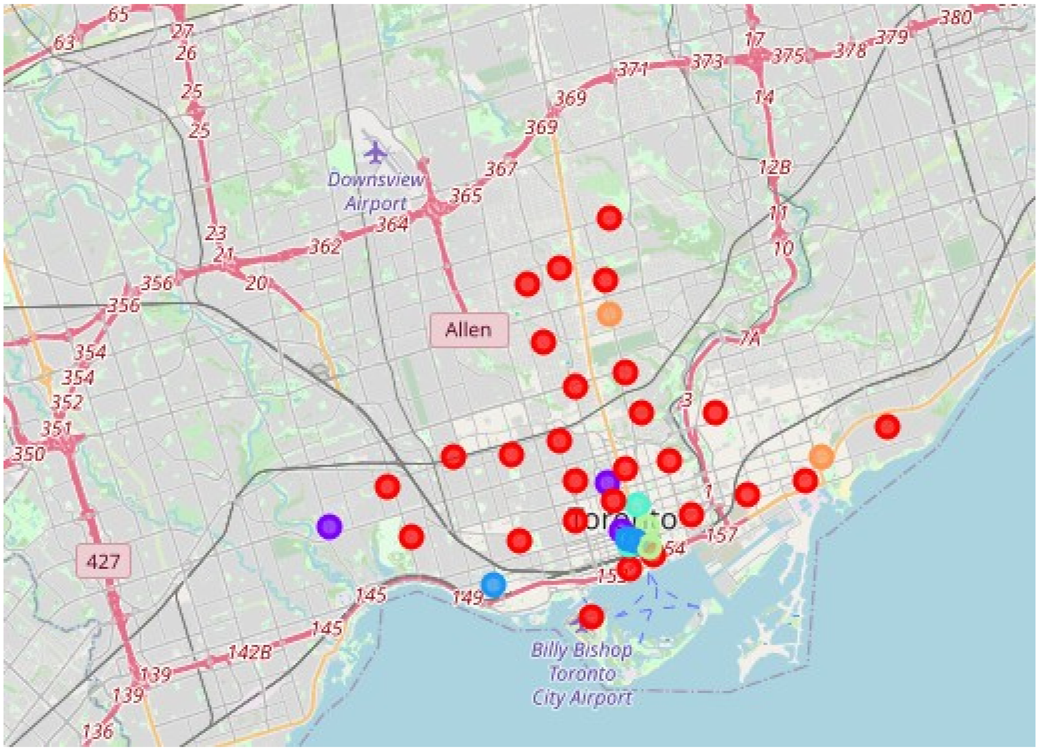


Figure 11. Clusters.

4. RESULTS

The following graph shows the average number of gyms per cluster:

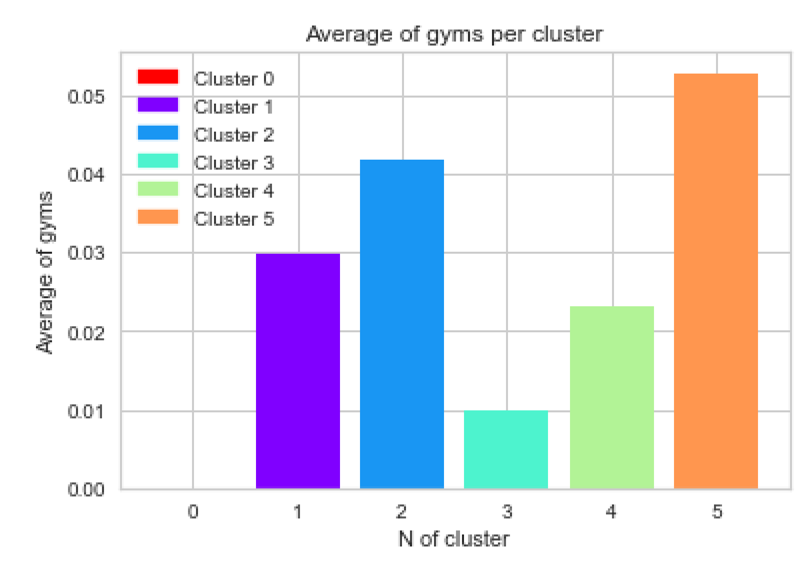


Figure 12. Average of gyms per cluster.

The following figure shows the number (in percentage) of neighborhoods per cluster:

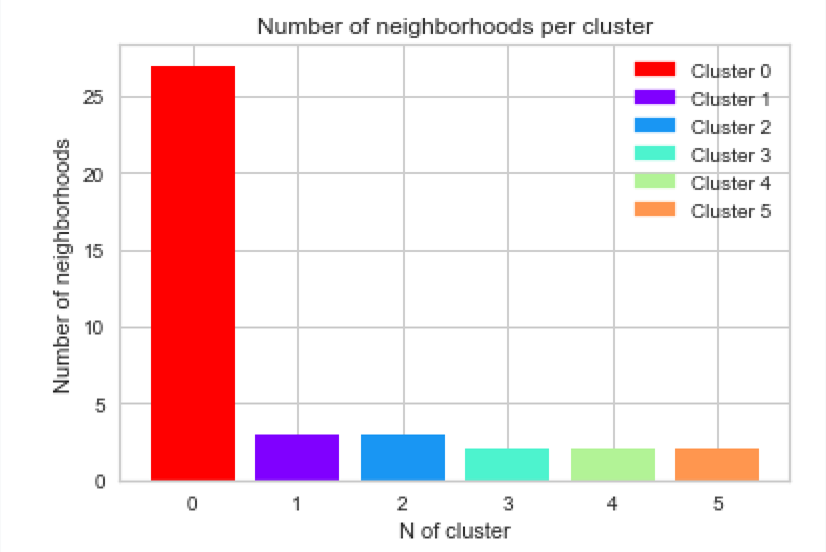


Figure 13. Number of neighborhoods per cluster.

It can be seen that the third cluster is where there are more gyms, while it is observed that there are no gyms for cluster 0. In turn, cluster 0 is the one with the largest number of members.

This tells us at first that a new gym could be installed in the neighborhoods of cluster 0 since it is not something that is available today.

5. DISCUSION

It can be seen that there is a shortage of gyms in Toronto in general, so installing one might be a good idea. In particular, there are none in cluster 0, so placing one there could be something innovative for the neighborhoods in it. However, the limited number of gyms is not a sufficient reason to install one there, there are many other factors to take into account, some more easily quantifiable, such as the costs of the area, as well as others more subjective, such as the fact that how interested local people might be in attending a gym and how much they would be willing to pay for it. To include this in the analysis you need the corresponding data.

6. CONCLUSIONS

In this task, the possibility of opening a gym in the city of Torontó was analyzed. Data from different sources (such as wikipedia and the Foursquare API) were combined to obtain the necessary information, as well as unsupervised learning ML techniques, obtaining clusters regarding the location of the gyms. An initial recommendation is given regarding where a new gym can be opened, but keeping in mind that for a more precise analysis it is necessary to have more information, such as the costs of each neighborhood as well as the interest of the population in this type of activity.