Applied Data Science Capstone Project

1 Introduction

A chain of protein powder stores is looking to open multiple new stores in city of Toronto. They are exploring how many stores they should open and where they should be located. The goal of this project is to give them a proposal of possible locations. The approach is to open stores in neighborhoods that have many gyms.

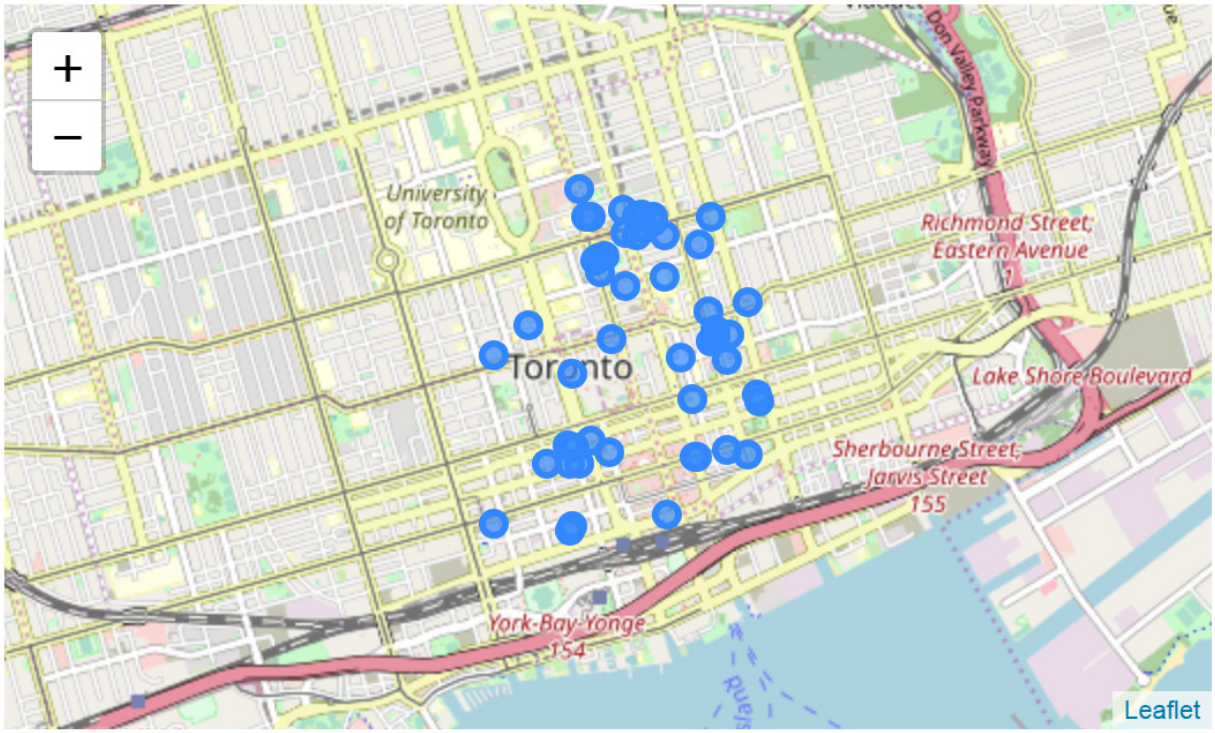
2 Data

The latitude and longitude of the gyms in the city center of Toronto will be extracted from Foursquare. For this use, the explore command will be called to the Foursquare API. The other features of the venues will be ignored for this analysis; we consider only the location to be relevant. An area with a high density of gyms will be a primary choice for opening a new store.

3 Methodology

For this project, an unsupervised technique is used for clustering the gyms in the city center of Toronto based on its spatial data.

The gyms within a radius of 10 km of the center of the city are considered. A call to the Foursquare API results in a dataset of 50 gyms, which are shown in the map below.



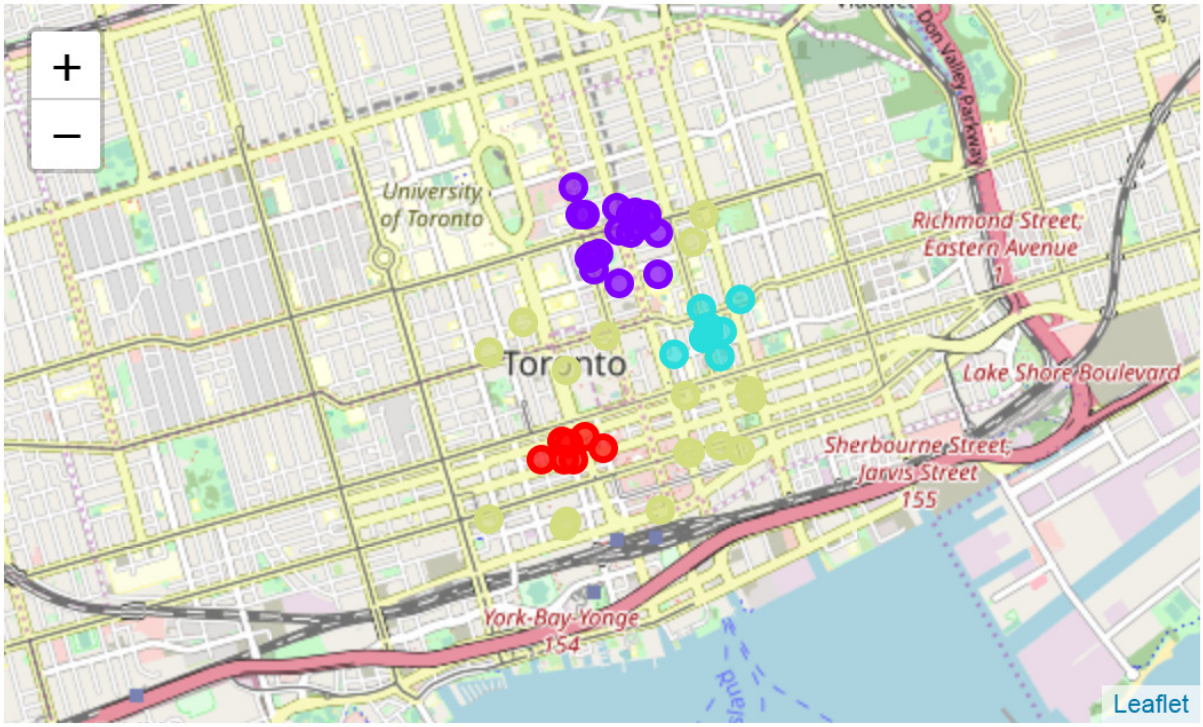
DBSCAN was chosen as a clustering algorithm, mostly for its ability to distinguish outliers from the cluster elements. Another reason for choosing DBSCAN is that the number of clusters does not need to be specified when running the algorithm.

The client is mostly interested in locating its stores in areas that are dense with gyms, in order to be visible to a large number of gym visitors. Since the number of the stores to be opened by the client is

not fixed, we can play around with the radius of the clusters and the minimum number of samples in the clusters. Finally we chose a set of parameters that resulted in 3 geographical clusters and a number of outliers, as presented in Section 4.

4 Results

As described in the Section 3, the DBSCAN clustering algorithm was used to cluster the gyms in the city center of Toronto into 3 dense clusters, as shown in the figure below.



The recommendation is to search for possible store locations within these clusters. The cluster centerpoints are defined in terms of their latitude and longitude values as

Clus_Db	Latitude	Longitude
0	43.649297	-79.386696
1	43.661013	-79.383225
2	43.655706	-79.376297

5 Discussion

It was observed that the gyms in the city center of Toronto can be divided into 3 clusters. To be well represented and maximize the number of visitors to the new stores, it is recommended to open 3 new stores, located within each of these 3 clusters. There are some outliers points, namely gyms that are more spread out over the city and not a part of a concentrated cluster. However, since the goal is to maximize the client base with a limited number of stores, the recommendation is to look for new store locations within the clusters.

6 Conclusion

In conclusion, the client is advised to open 3 new stores with locations inside of the 3 identified clusters.