IBM Data Science Professional Certificate Capstone Project

Analyzing And Clustering Toronto Neighborhoods To Find The Best Place Gym/Fitness Center

By: Mohammad Odeh

Introduction

Toronto is the largest city in Canada, the capital of Ontario, and the home to more than five million people, this gives it a big advantage for businessmen and investors who want to establish a profitable business.

A business man is interested in opening a gym in the city of Toronto and he is looking for the best place for the business that will lead to more customers, so the objective of this project is do full analysis of the neighborhoods in Toronto by clustering them based on similar businesses (venues), this analysis will help locate areas with less gyms which indicate a great opportunity to open a gym and provide such a service for residents of that area.

Data

To find the solution and perform detailed analysis on the topic, we need data that contains neighborhoods and locations coordinates of the city of Toronto, and we need data about the available gyms centers (venues) in each neighborhood, there for we must:

1- Web scrape the lists of neighborhoods from Wikipedia page : https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

	PostalCode	Borough	Neighbourhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Harbourfront
5	M5A	Downtown Toronto	Regent Park
6	M6A	North York	Lawrence Heights

2- Get the coordinates of each neighborhood using a csv file contains all coordinates of Toronto neighborhoods (http://cocl.us/Geospatial_data)

	PostalCode	Borough	Neighbourhood	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	Harbourfront	43.654260	-79.360636
3	M5A	Downtown Toronto	Regent Park	43.654260	-79.360636
4	M6A	North York	Lawrence Heights	43.718518	-79.464763

3- Get the list of available gyms in each neighborhood by using Foursquare API

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	freq
0	Harbourfront	43.65426	-79.360636	Toronto Cooper Koo Family Cherry St YMCA Centre	43.653191	-79.357947	Gym / Fitness Center	4
1	Harbourfront	43.65426	-79.360636	The Extension Room	43.653313	-79.359725	Gym / Fitness Center	4
2	Harbourfront	43.65426	-79.360636	The Yoga Lounge	43.655515	-79.364955	Yoga Studio	4
3	Harbourfront	43.65426	-79.360636	Corktown District Lofts Gym	43.655652	-79.358125	Gym	4
4	Regent Park	43.65426	-79.360636	Toronto Cooper Koo Family Cherry St YMCA Centre	43.653191	-79.357947	Gym / Fitness Center	4

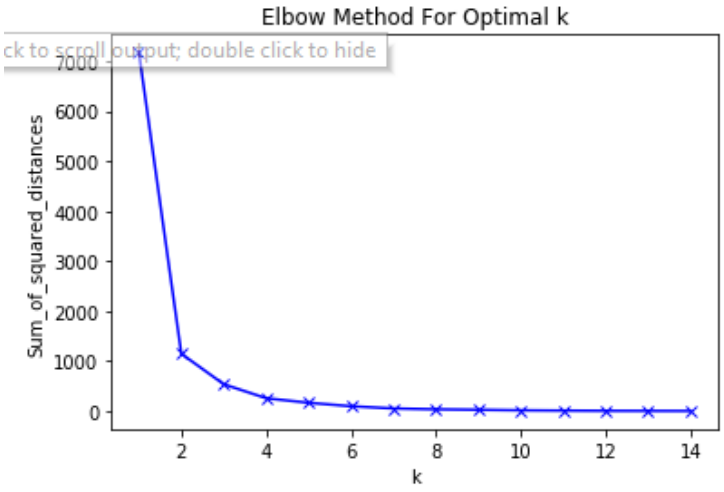
Methodology

For Finding the best location for opening a new gym, a clustering algorithm (K-Means clustering) is applied which will cluster the neighborhoods into clusters depending on the number of gyms in that neighborhood.

The data that will be used for that is the cleaned/edited list of neighborhoods with their corresponding coordinates and number of gyms on each one :

	Neighborhood	Neighborhood Longitude	Neighborhood Latitude	freq
0	Harbourfront	-79.360636	43.654260	4
1	Regent Park	-79.360636	43.654260	4
2	Lawrence Heights	-79.464763	43.718518	4
3	Lawrence Manor	-79.464763	43.718518	4
4	Queen's Park	-79.389494	43.662301	11

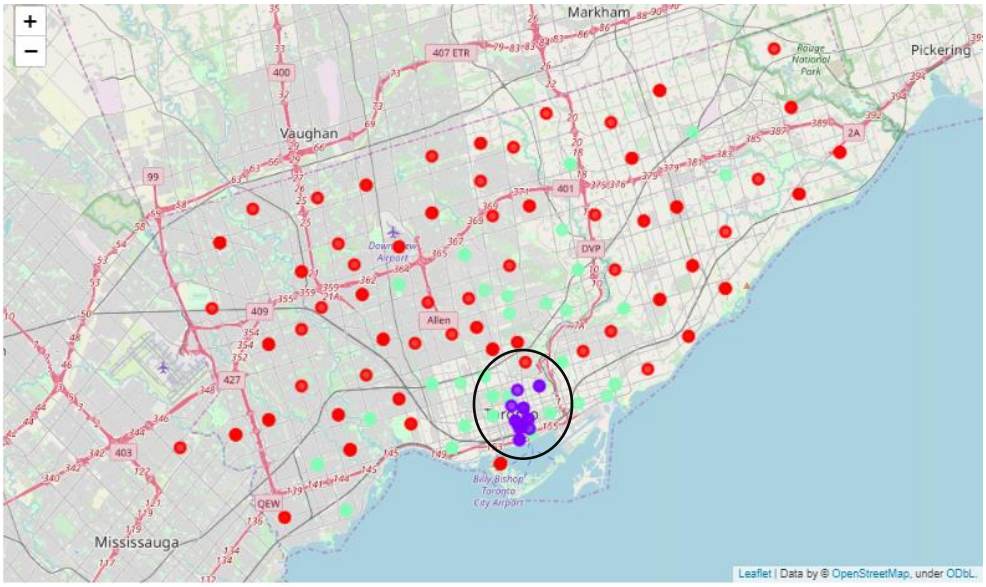
Next in order to find the optimal K (number of clusters), elbow method is used by applying k means for different sets of " K ", then the sum of square distance (Inertia) is calculated for each " K "



So, K=3 is the optimal number of clusters, there for we apply the algorithm to the data with k=3

Result

The resulted clusters showed in the map below, show that there is a concentration of gyms on a specific area while there is a lot of neighborhoods that could be a potential place for opening a gym on.



Conclusion

The results showed that the heart of the gym business is concentrated near the coast and the more we go to the edge of the city the more opportunity there will be to open a gym that will serve the community and the people there.

With the results that shows the neighborhoods and places where it has no gyms, and with further analysis using the population and demographic distribution data among the neighborhoods, we can determine which of these areas are more profitable and will serve more people.