Analysis of Gyms and Sporting venues' distribution across Toronto

Salaheldin Gaffar

Data Scientist

Table of Contents

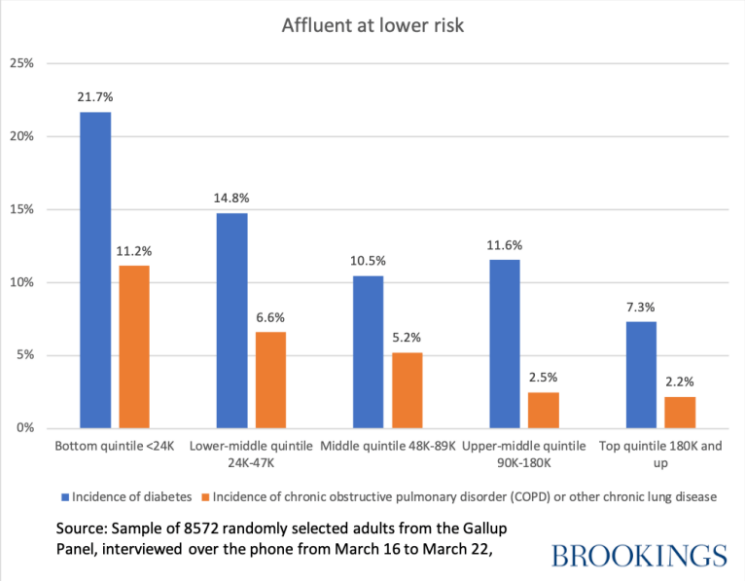
Introduction

Description and Background

During the 2020 COVID-19 pandemic, several issues in the health of the world population were not only unveiled but in fact emphasized. As of today, October 24th 2020, 1.1 million people have died all over the world out of 42 million reported cases, which gives the virus approximately a 2.6% death rate. Out of these 42 million cases and 1.1 million deaths, 200 thousand cases and 9862 deaths occurred in Canada, which shows an alarmingly elevated death rate of around 4.9%.

The higher than global death rate in Canada can be explained through numerous studies that have been conducted to correlate the death rate with age, race, gender and demographic, overall health status, access to quality healthcare. These studies have been supplied with plenty of data has been collected throughout the past 10 months, a feature that was not available to the previous generations that witnessed the 1918 Spanish flu pandemic.

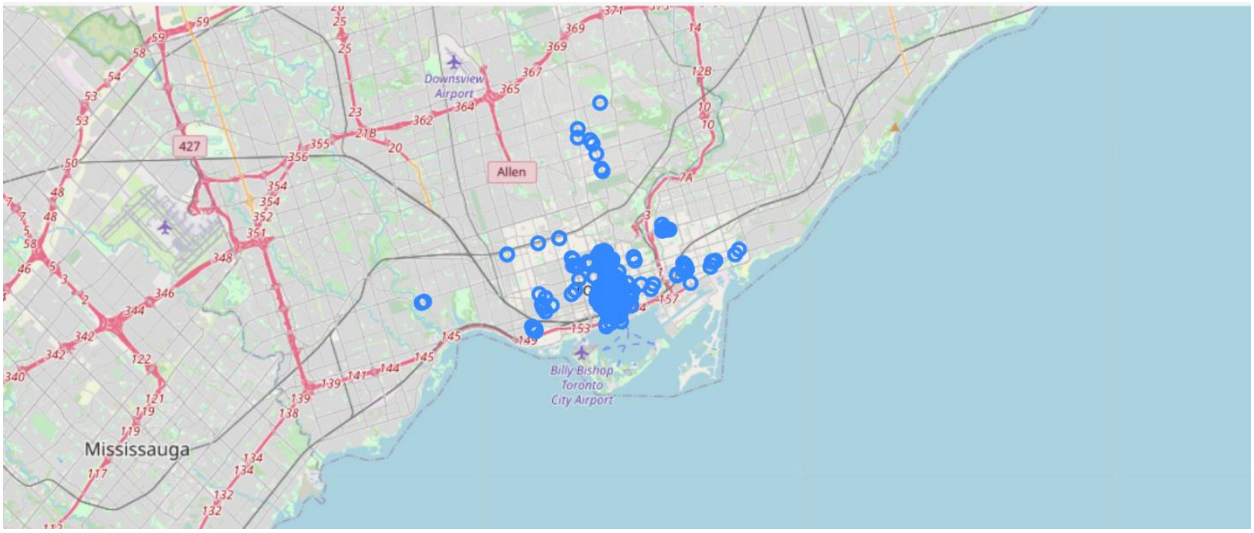
It has been proven that there was a significant disparity between the death rates of upper-class and lower-class individuals, which can be explainable through several factors including: diet and exercise habit differences, pre-existing condition occurrences, nature of work and feasibility of working from home and several other factors. The image below represents one of these factors, as per Gallup's poll conducted in the USA; the plot shows the incidence of pre-existing conditions across different income divisions.



It is no secret that obesity is strongly correlated to both presented pre-existing conditions and Canada places 44th out 190 countries in terms of overall mean BMI scores (refer to WHO data at 2014) with an average BMI of 27.2 (above 25.0 is overweight). It is also worth mentioning that Canada ranks 26th in BMI in males with an average BMI of 26.

For the reasons stated above, Canadian authorities as well as several other organizations such as **Obésité Canada** (OC) are very interested in the reasons behind this increased obesity levels. Several nutrition related initiatives have been in execution for the past 10 years and will continue in the future. However, the second part of the health equation, exercise will the focus of this study.

In particular, the focus of the study will be concerning the availability of gym and other sporting venues for the citizen of Toronto, Canada. The map below shows the registered sporting venues in Toronto and their distribution. It is required to understand whether the distribution below is in fact correlated to the population count for each neighborhood, or whether sporting facilities are poorly distributed and more governmental and NGO incentives should be offered to gyms and their equivalents to open in the less-served neighborhoods.



Target reader for report

This report should prove to be beneficial for several entities; it should serve **government officials** in Toronto identify regions where more work should be done in providing sporting facilities for citizens.

Moreover, charitable and other civil society organizations can benefit from identifying regions where sporting facilities are lacking. Finally, businesses that are looking for attractive locations to open gyms, where there are plenty of target customers can find guidance in this report that can help them with their market research.

Methodology

Data description

The data needed for our study will include the following:

1. Foursquare venue data to determine the locations of the registered gym venues
2. Demographic data on each Toronto neighborhood (includes population, population density, average income) supplied by Wikipedia.
3. Walk score and several other important data for each neighborhood (a score that indicates how much a neighborhood is walking-friendly) provided by **Toronto Wellbeing** web tool.

The venue data coupled with the population details will serve to determine whether a certain neighborhood belongs to a cluster where there are enough sporting facilities or not. Moreover, the walk scores and average income data will determine whether the area shows good promise for sporting businesses to open, or whether more incentives need to be offered by charities and government to make sporting venues profitable at a certain area.

Data preparation

The process starts with obtaining the data for different neighborhoods across Toronto, and then using the neighborhood co-ordinates to extract the gym and sports related venues within a 500 meters radius of each neighborhood using the Foursquare API. Unfortunately, Foursquare sometimes returns some unrelated venue categories in addition to the requested ones.

This meant that the different venue categories needed to be manually inspected and the irrelevant venues were dropped.

Some irrelevant categories that were filtered out were:

1. Residential Building (Apartment / Condo)
2. College Academic Building
3. Student Center
4. High School
5. Bath House
6. Building
7. Bath House
8. Medical Center
9. Hotel

Another important source of data to be used was the census data for different neighborhoods in Toronto, which included the average income and population count. It was also required to match the data obtained for the neighborhood co-ordinates with the data obtained from the census data, since they come from different sources.

The next step was to prepare the data to be used for the machine learning clustering model (k-nearest neighbor clustering).

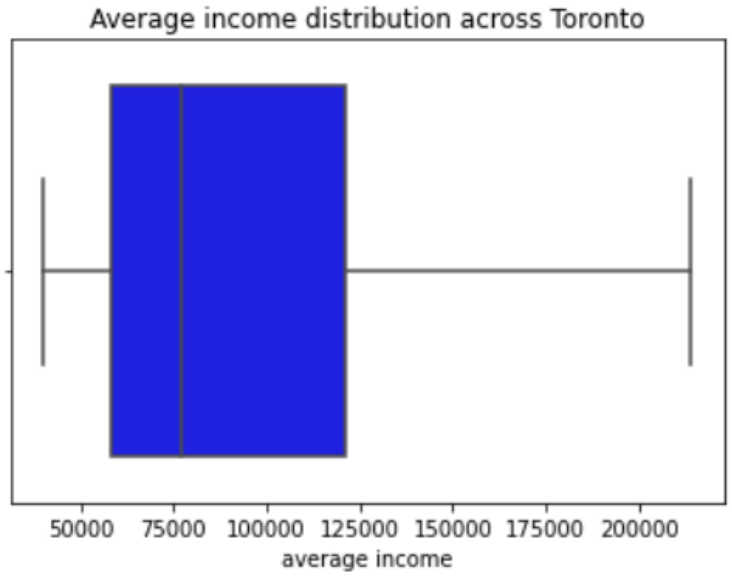
The desired inputs for the model were the following two derived variables:

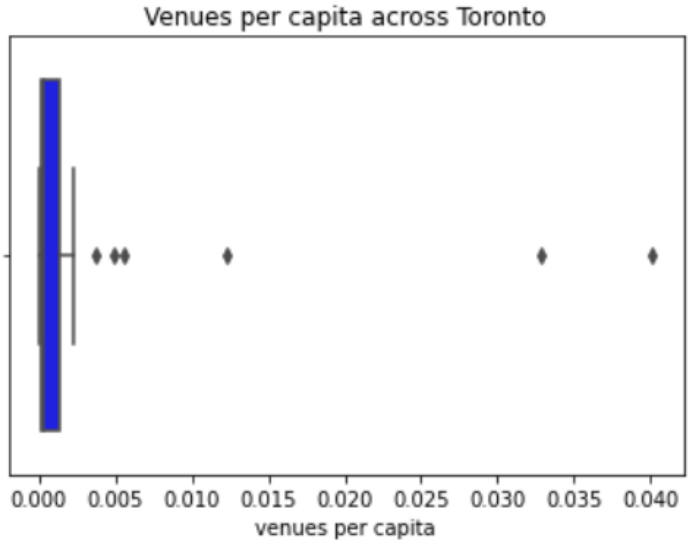
1. Available venues per capita (total sporting venue count divided by the total population)
2. Average income of the citizens of each neighborhood

After adding all the native variables into a single dataframe, it became a simple task to derive the two required variables.

Exploratory data analysis

A very important thing to note while exploring the census data in Toronto is the significant disparity in income between different neighborhoods in Toronto. Moreover, most neighborhoods are underserved in terms of sporting venues except for a few outliers.





K-Clustering modeling

After preparing all the required data for the model, it was necessary to feed the data to the k-clustering algorithm.

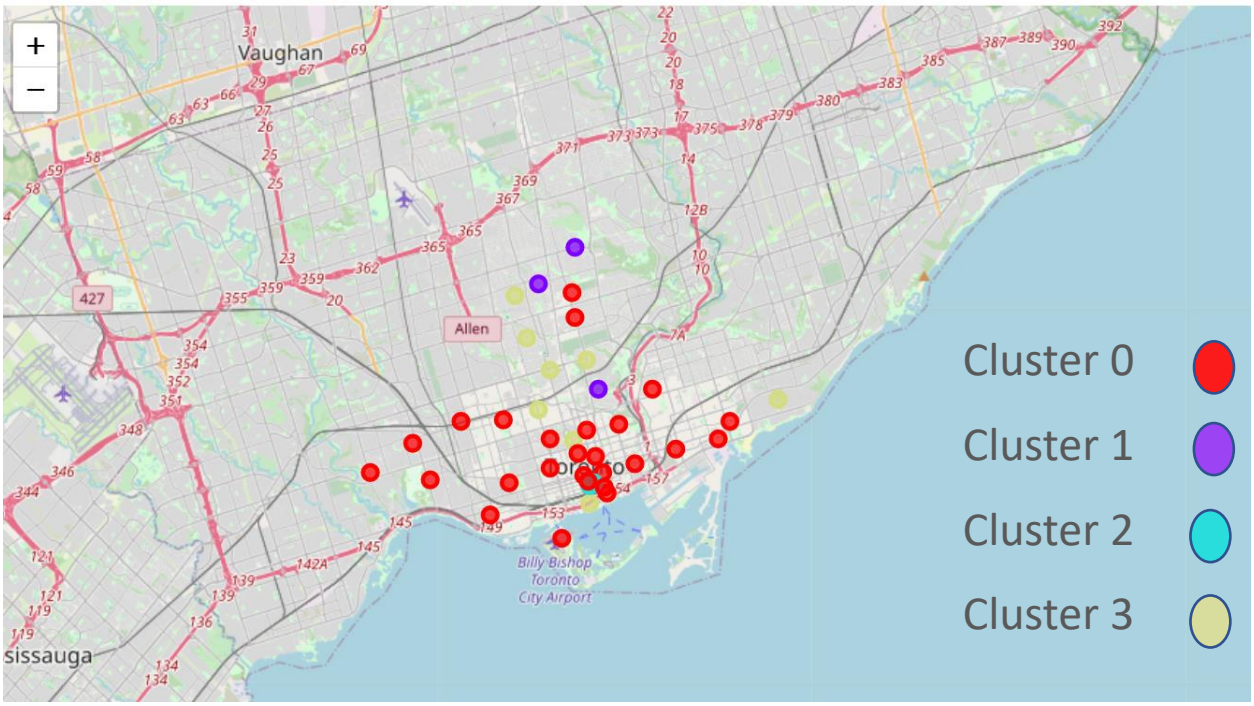
The main target was to segment the neighborhood into the following four categories:

1. Neighborhoods that had enough coverage by sporting venues **(cluster 2)**
2. Neighborhoods that did not have enough coverage and households were of the highest income (useful for corporate level gym businesses to open there, where customers are willing to pay premium prices) **(cluster 1)**
3. Locations where there was not enough coverage and household incomes were low **(cluster 0)**
4. Locations where there was not enough coverage and household incomes were medium **(cluster 3)**
5.

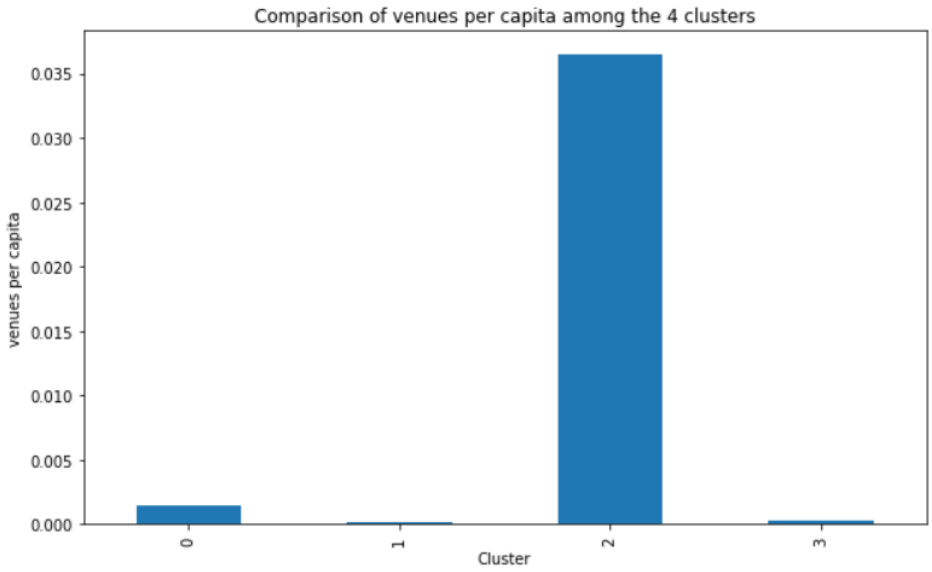
This first and fourth clusters would be the main target for the focus of charitable organizations and NGOs, where incentives for small business owners should be put in place to open sporting venues there.

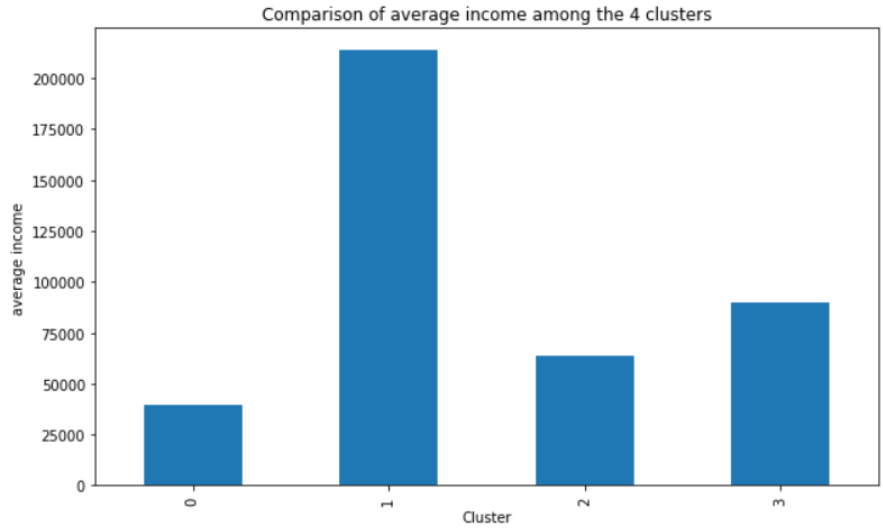
Results section

After the model is utilized, the following 4 clusters are distributed as follows on the Toronto map :



The plots below show the different clusters where one cluster contain neighborhoods with the maximum venues per capita and medium average incomes and another cluster contains neighborhoods with the highest average income and the least venues per capita.





Discussion section

Interpretation of the results

The importance of identifying the different segmentations of neighborhoods in Toronto in terms of availability of sporting venues is to serve as a guide for stakeholders in order to maximize the benefit for citizens as well as the business owners and investors as well as make sure charity funds are allocated where they should be.

It has been shown that cluster 1 (2nd cluster) is most adequate for corporate level sporting venues and clusters 0 and 4 are better served by smaller scale gyms and Yoga studios. Also, there was no need to invest more in neighborhoods belonging to cluster 2 (3rd cluster), since it was currently the most served in terms of sporting venues.

Future improvements

During this study, two important factors were included when dividing the neighborhoods in Toronto into several segments: the average income in the neighborhood and the available sporting venues per capita.

The study can be refined through the following:

1. Classifying venues according to the recommended capacities for each (e.g. a skating rink can occupy more than a Yoga studio, so each venue should be weighted by its capacity to serve customers
2. Parks and social clubs and other places with available sporting facilities can be included in this study
3. Adding the real-estate prices and renting costs paid by businesses in each area and adding this data to the clustering inputs, so as to divide the prospect neighborhoods into different classes in terms of the costs incurred to the business owner who invests there as in addition to the availability of customers, which is already displayed in the current study.
4. Use more recent census data, where the current dataset being used was last updated in 2006.
5. Use more weighting parameters for venues to include how far they are from the neighborhood center, since our current 500m radius can be considered too much for elderly citizens that would like to find a place to exercise

Conclusion section

According to the presented results, neighbors in Toronto can be clustered into three main segments:

1. high-income neighborhoods where not enough sporting venues are available
2. low and middle-income neighborhoods where not enough sporting venues are available
3. Neighborhoods where enough venues exist to cover the current population levels

It is recommended that each segment be treated differently by regulators, investors as well as NGOs and charities. High-income neighborhoods can be approached by large corporate-level gyms and sporting centers where customers will be more likely to pay premium prices for gym subscriptions. As for the low-income neighborhoods, more incentives should be offered by the appropriate stakeholders in order to increase the number of venues available to the public.

References

https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

https://www.brookings.edu/blog/up-front/2020/03/27/class-and-covid-how-the-less-affluent-face-double-risks/