Gym Data Analytics For Business in Toronto, York Area.

Elena Grishin

November 5, 2020

1. Introduction

1.1 Background

According to Statistics Canada, the population figure for the York Region in 2019 is 1,191,400. It was constantly growing for the last several years averaging 1.9 percent per year since 2011. The idea behind this project is as the population is growing there are not enough gyms in that region. So, this project is designed to help a group of stakeholders to select a good location to open a gym.

1.2 Problem

The York region neighbourhood's data that might be used to find a suitable location to open a gym. Using data science methods such as clustering, this project aims to answers a business question of where in Toronto, York Area stakeholders should open a gym.

2. Data

To accomplish the task we will need two data sets merged into one.

First, is a list of postal codes that can be obtained from the Wikipedia page.

- Wikipedia page data source

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

Second, is the data set of latitude and longitude of each neighbourhood.

- Latitude and Longitude data source

  http://cocl.us/Geospatial_data

Data needs to be cleaned to exclude neighbourhoods that have no assigned values.

Also, data needs to be simplified to include only data for Toronto, York region.

3. Methodology

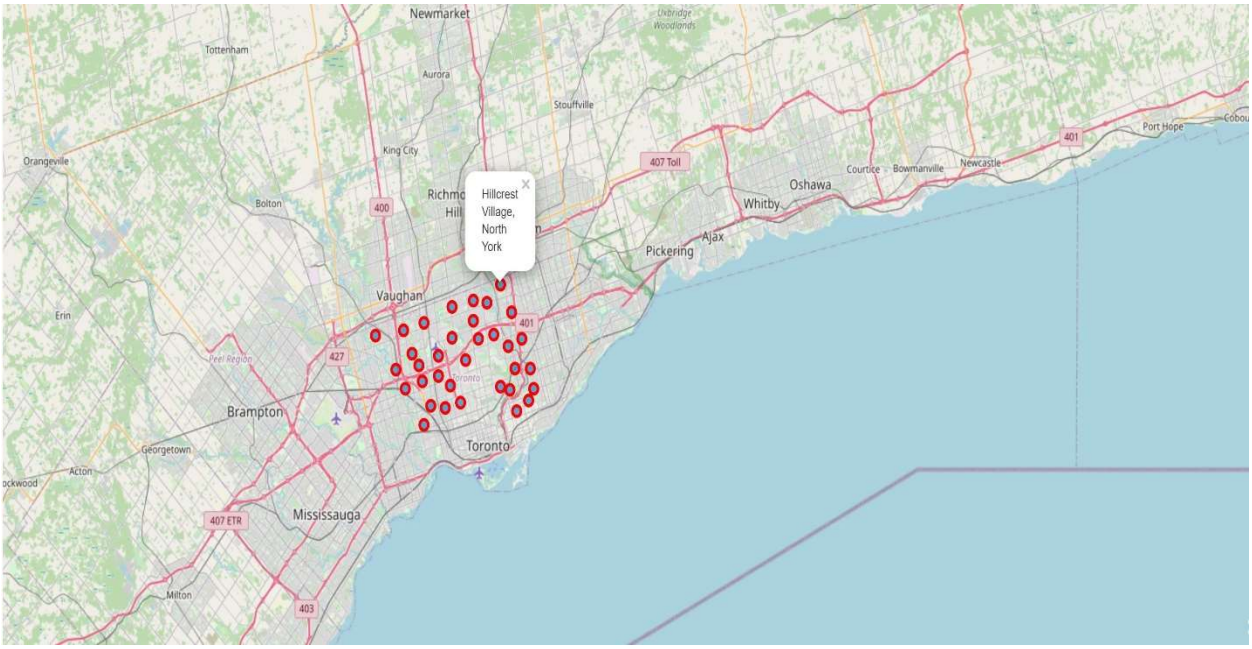
3.1 Creating York area data frame

First, the Toronto data was scraped from the Wikipedia page into a pandas data frame. It was cleaned to ignore cells with a Borough and Neighbourhood that have no assigned value. Second, the latitude and the longitude coordinates of each neighbourhood were retrieved to a separate data frame. Next, both data frames were merged on the Postal Code column. The merged data frame was investigated to show its neighbourhoods, and the size was checked.

The investigation showed that the data frame contains not only neighbourhoods requested for the project. Therefore, the simplification strategy was used to create the final York area data frame that includes only Toronto, York region data. The York area data were grouped by the neighbourhood, and its size was checked.

3.2 Geospatial visual analysis

Geographical coordinates of Toronto were retrieved using geopy Python client. And, Python Folium library has been used to utilize the power of geospatial visual analysis, a map.



3.3 Explore the neighborhoods

The Foursquare developer account was created to obtain the account ID and API key to pull the data from the Foursquare API. Using getNearbyVenues function 336 venues with 120 unique categories within a 500-meter radius were pulled for future analysis.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
3	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant
4	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop

The data was checked specifically to verify it contained the "Gym" venue category.

To analyze each neighbourhood the mean for each Venue Category was calculated, and a new data frame was created to include the only gym category.

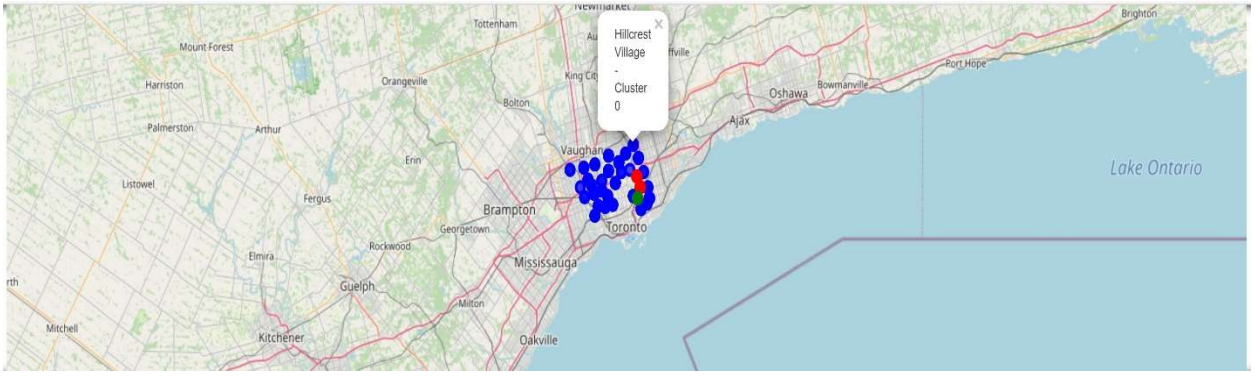
3.4 Cluster Neighborhoods

The unsupervised machine learning algorithm, clustering method by using k-means clustering was performed. The k-means clustering identifies k number of centeriods, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

Based on the frequency of occurrence for Gym, it was decided to cluster the neighbourhoods in York area into 3 clusters.

4. Results

The map was created for visual presentation to show 3 clusters. The below map shows Cluster 0 with red colour, Cluster 1 with green colour, and Cluster 3 with blue colour respectively. In addition, three clusters of data were filtered and represented in tables view for the analysis.



5. Discussion

After examining three clusters it looks like most of the gyms are located in Don Mills and fewer gyms in Thorncliffe Park. Therefore, Parkview Hill, Woodbine Gardens neighbourhood might be a good location to open a gym. However, in this project, we have only one factor, the existence of Gym in each neighbourhood taken into consideration. In reality, there so many different factors such as demographic area, gym accessibility, the income of residents that should be taken into consideration when planning to open a Gym.

6. Conclusion

The location of a gym or fitness facility is very important. In this research, we used the clustering method and only one factor, the existence of Gym, to recommend a good location to open a gym. Future research can take into consideration additional factors and other data science methods to support or disprove the recommendations.