The Battle of the Neighborhoods:

Toronto neighborhoods exploration for Gym opening

Fred Metsma

April 15, 2021

1. Introduction

This data science project is made as a part of the IBM Data Science Professional Certificate program. Project aim for the author is to work with a real dataset(s) and solve a real problem, using data science methodology. Project main objects are:

- Define a problem
- Define a data that is needed to solve the problem
- Search for data
- Use Foursquare location data to solve the problem

1.1 Background

Imagine a situation where Gym chain has a goal to open 5 venues in Toronto, most populated city in Canada. Toronto with 103 neighborhoods is big, so company needs to analyze all the neighborhoods and find out which neighborhoods of the city has the most potential for business.

1.2 Problem

Analyzing population density is easiest way to discover where the people are and it should hedge against making gym in the commercial neighborhood. Population density should also show how much potential customers there are but equally important is to find out where the Gyms are located today on how they are distributed. Population and Gyms distribution data combined should give as an answer.

2. Data

To solve the problem, we need to have different sort of datasets:

- Neighborhoods geographical data
- Neighborhoods population data
- Gyms geographical data

We were able to use neighborhoods latitude-longitude data which was earlier provided by Coursera. Neighborhoods are listed there using postal codes. So we needed also a dataset which connects postal codes with neighborhoods.

So, for solving the problem we needed all together 4 datasets:

- **List of postal codes to define Toronto neighborhoods**. To satisfy this data need, we use Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M), where are listed all Canada postal codes starting with M. Postal Codes starting with M belong to Toronto. We are going to use BeautifulSoup package to scrape the data from the webpage.
- **Foursquare API for Gyms geographical data**. Foursquare database consists of different Venues with its details, including category, longitude and latitude. Foursquare gives access to json files which we structure and wrangle the data into suitable form.
- **Geospatial data about neighborhoods.** We need longitude and latitude for using Foursquare location data. For that we use data from http://cocl.us/Geospatial_data, where all Toronto postal codes are listed with its longitude and latitude. We are going to use Wget to import csv and pandas to turn it into dataframe.
- **Canada population data divided by postal code**. Statistics Canada provide different kind of information about Canada. We are going to use population data csv from following link (https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Tables/CompFile.cfm?Lang=Eng&T=1201&OFT=FULLCSV). Data is from the last Census of Population, which took place in 2016. We are going to use Wget to import csv and pandas to turn it into dataframe.

First i fetched the neighborhoods data with postal codes from Wikipedia, used Beutifulsoup to it wrangle into dataframe. Next i downloaded the Geospatial data and turned it into dataframe and last i fetched the population data from Statcan. When all the data was gathered, i compared

dataframes for missing values and discovered that one postal code was without population value. It was a Canada Post Gateway Processing Centre. I merged three datasets into one toronto_df dataset and then just dropped the Canada Post Gateway Processing Centre post code row.

:

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Population
0	M3A	North York	Parkwoods	43.753259	-79.329656	34615.0
1	M4A	North York	Victoria Village	43.725882	-79.315572	14443.0
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	41078.0
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	21048.0
4	M7A	Queen's Park	Ontario Provincial Government	43.662301	-79.389494	10.0
...	...	...	...	...	...	...

Figure 1. toronto_df dataset

Next i used Foursquare API to fetch Gyms data. By the example of previously used data fetching from Foursquare, i designed the function to get data and added a categoryID for fetching only Gym parent venues. Gym should be in less than 30 minutes walking distance, so i set the radius from Neighborhood (postalcode) to 1500 meters.

We are interested only in Gym and Fitness Center's, so we define a categoryId which is 4bf58dd8d48988d175941735 for Gym and Fitness center parent category.

]:
```
def getNearbyGym(postalcodes, latitudes, longitudes, radius=1500):

    venues_list=[]
    for pcs, lat, lng in zip(postalcodes, latitudes, longitudes):


        # create the API request URL, parent category 4bf58dd8d48988d175941735 is for all the Gyms
        url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}&categoryId=4b
f58dd8d48988d175941735'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            LIMIT)
```

Figure 2. Fetching data from Foursquare

After doing this, i had Gyms dataset divided by postal codes and their Latitude-Longitude information.

	PostalCode	Neighborhood Latitude	Neighborhood Longitude	Gym	Gym ID	Gym Latitude	Gym Longitude	Gym Category
0	M3A	43.753259	-79.329656	LA Fitness	4c0bf0756071a593ae01e232	43.747535	-79.317674	Gym / Fitness Center
1	M4A	43.725882	-79.315572	North Beach Indoor Volleyball Academy	4b4fd4a9f964a520fd1627e3	43.737191	-79.323714	Gym / Fitness Center
2	M4A	43.725882	-79.315572	Fit4Less	50158f3fe4b0e3a6f525a0a1	43.725660	-79.297823	Gym
3	M4A	43.725882	-79.315572	GoodLife Fitness North York Ferrand and Rochefort	4af0dc2af964a520b8df21e3	43.719812	-79.331564	Gym / Fitness Center

Figure 3. List of Gyms

After having all the necessary Gyms data, i merged it with toronto_df dataframe and got final Toronto dataset, so everything was ready for analysis.

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Population	Gyms
0	M3A	North York	Parkwoods	43.753259	-79.329656	34615.0	1.0
1	M4A	North York	Victoria Village	43.725882	-79.315572	14443.0	10.0
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	41078.0	36.0
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	21048.0	6.0
4	M7A	Queen's Park	Ontario Provincial Government	43.662301	-79.389494	10.0	75.0
...	...	...	...	...	...	...	...

Figure 4. Final Toronto dataset: fin_toronto

3. Analysis and methodology

Analysis was started with visualizing population in the neighborhoods. Choropleth maps was used for the task.

For the choropleth map it was also needed to get neighborhoods boundaries geo information. After searching the GeoJson file it was found from here. Down on the map you see neighborhoods from the least to most populated. Toronto most populated neighborhoods tend to be in the northern part of city and least populated is the center of the city.

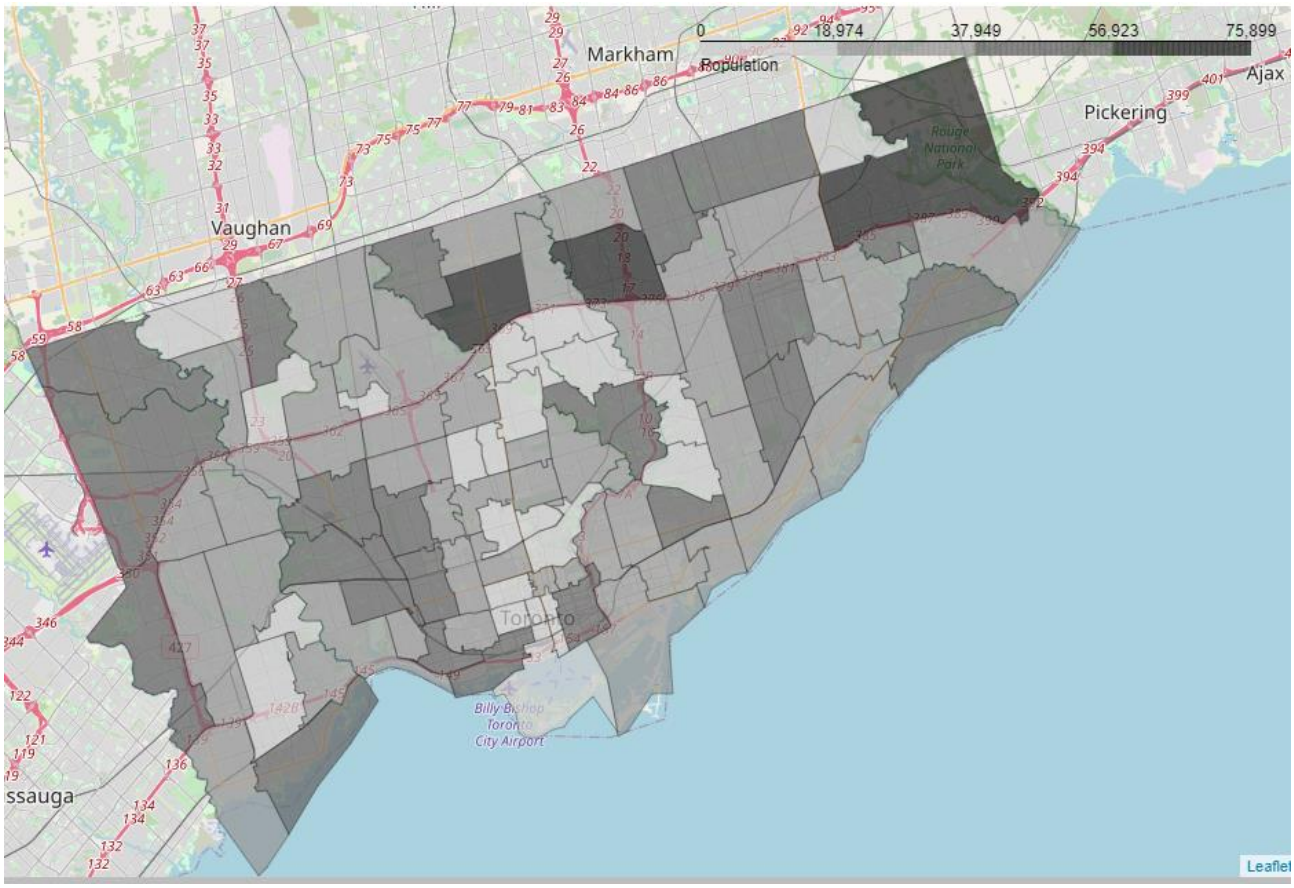

Figure 5. Toronto population in neighborhoods

Next task was to visualize Gyms in the Toronto. Folium maps was used for the task. Down here on the map, you can see that gyms are highly concentrated in the Toronto central area. This is quite opposite where the people live.

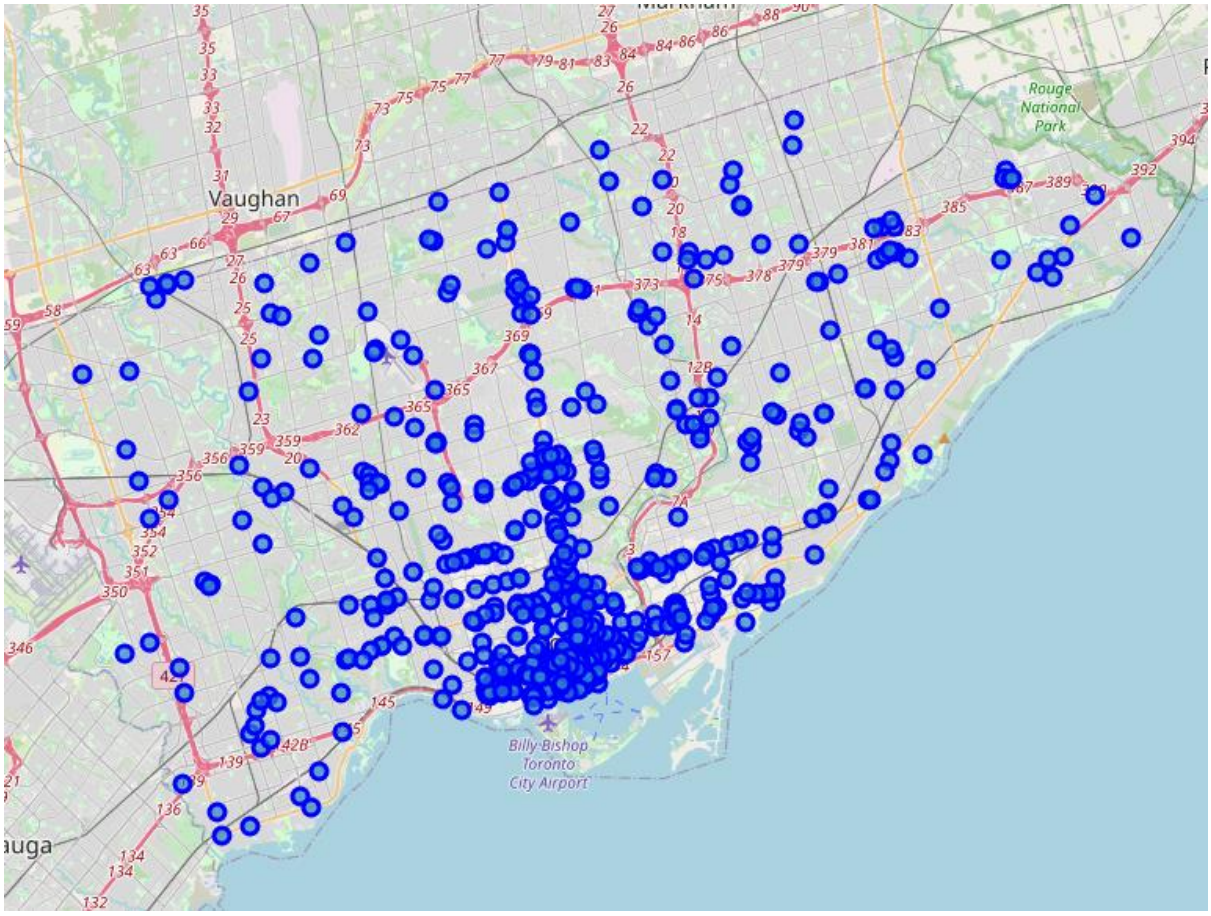

Figure 6. Toronto gyms

To see how the Gyms and neighborhoods population is related, the data was plotted on the scatterplot. From the looking on to the graph, it can be said that there is a negative correlation between population in neighborhood and number of gyms. According to data it can be said that people are more used to train closer to their workplaces, not homes as were expected. It gives an **idea**.

Now, during the COVID crises, more people have used to work from home and many probably stay working from the home also after the crisis. When gyms are located in the commercial area it is high probability that the demand for near-the-home gyms will grow. So, we stay on our path to look for neighborhoods with less gyms per population.

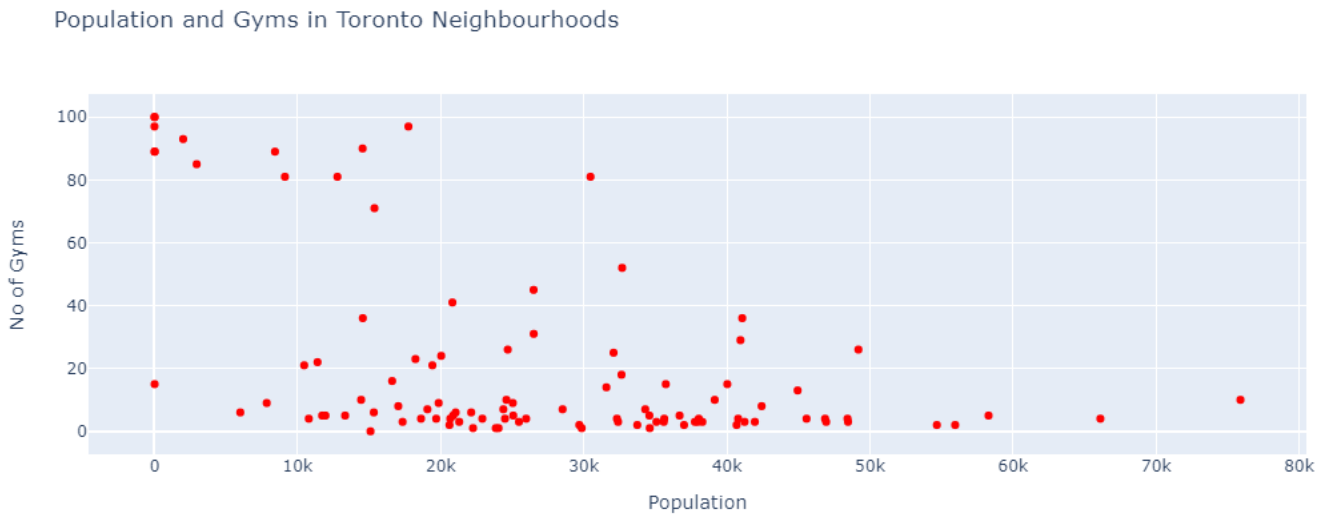

Figure 7 Population and Gyms in Toronto neighborhoods

4. Results

Next was moved on to the clustering. Kmeans was choosed for the clustering and clusters were set to 5. Kmeans optimizes the squared errors (distances from the points), so it should divide scatterplot into 5 most similar clusters. We are looking for neighborhoods with most gyms per population. After the clustering scatterplot is looked again for the clusters. From the clustered scatterplot we can see that the most interesting cluster for us is Cluster 3

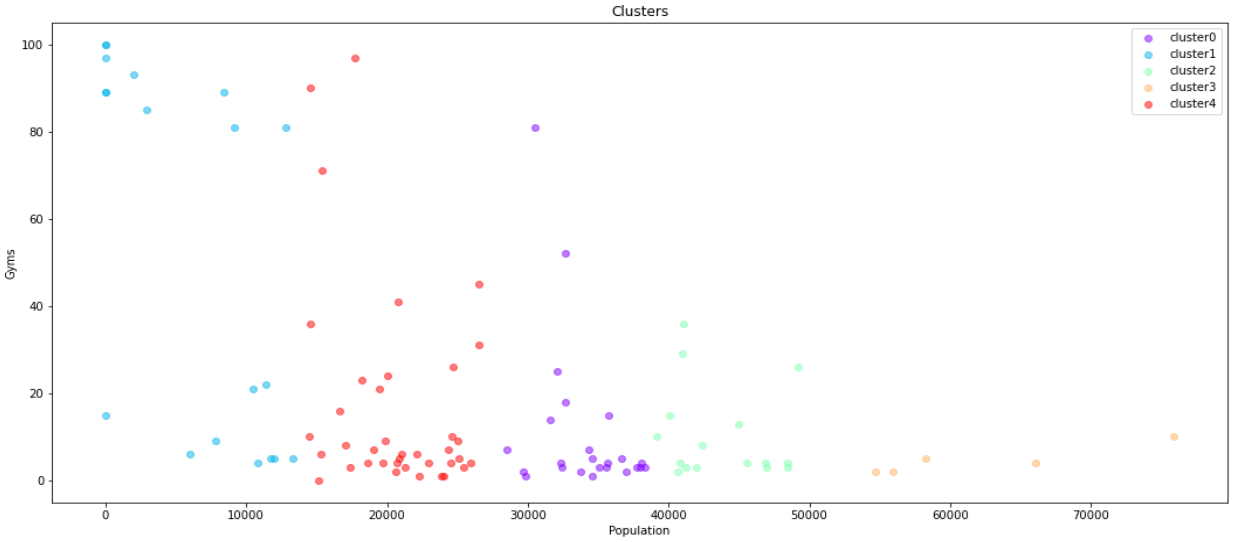

Figure 8 Clustered. Population and Gyms in Toronto neighborhoods

To see how the most interesting neighborhoods are located, the clusters are plotted using choropleth maps. As previously told, the most interesting is Cluster 3, from the map it is medium blue. From the map legend it is between 3 and 4. Second best is Cluster 2. From the map it is light blue.

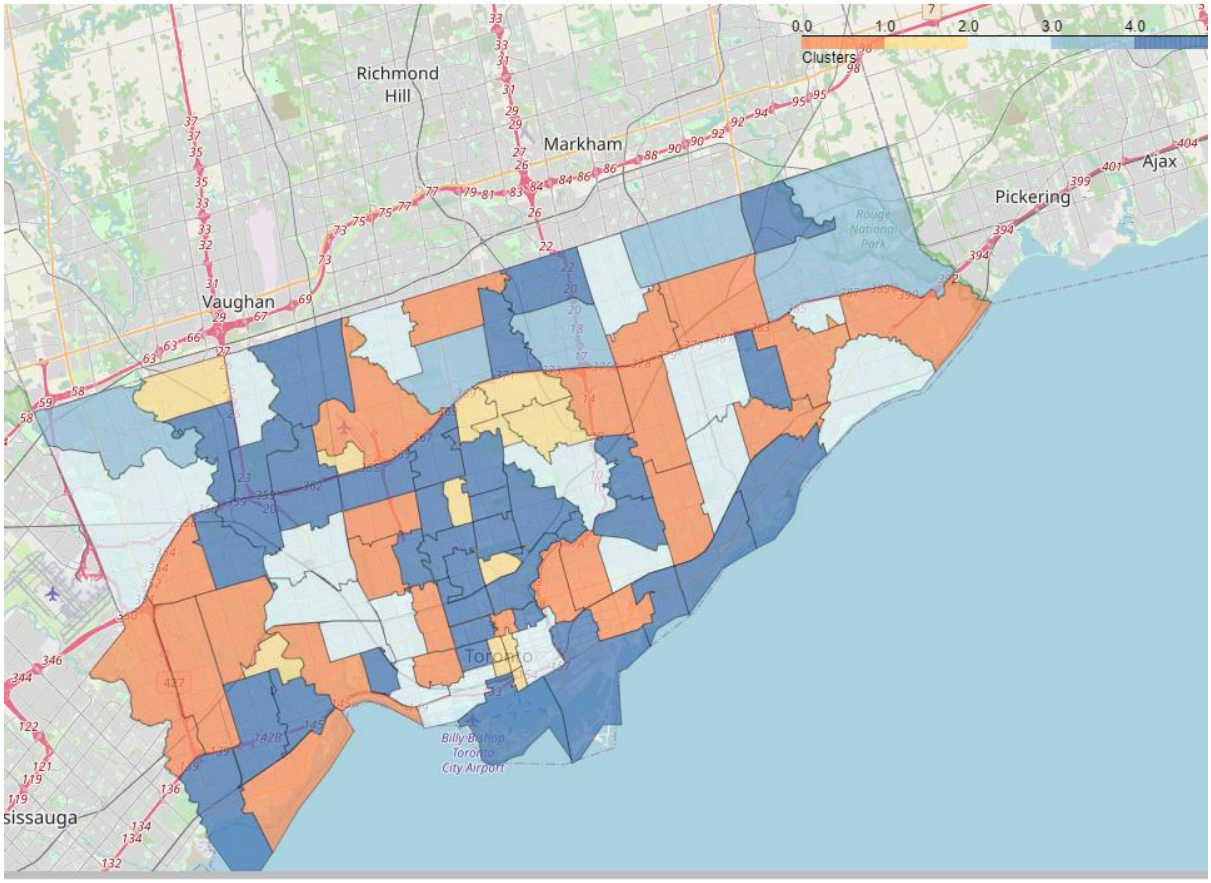

Figure 9. Clustered choropleth map. Medium

5. Conclusion

In this research i analyzed the neighborhoods of Toronto for expanding Gym venues by 5 units. 4 Datasets were used for analysis – Postal Codes, neighborhoods Geospatial data, neighborhoods population data and venue data from Foursquare. One helping dataset was used for choropleth maps. Analyze was carried out by clustering city into 5 cluster by using population and gym numbers.

The results say that least gyms wit maximum populations are in Cluster 3 which consists of 5 neighborhoods.

Neighborhood	Population	Gyms
Malvern, Rouge	66108.0	4.0
Fairview, Henry Farm, Oriole	58293.0	5.0
Willowdale South	75897.0	10.0
Milliken, Agincourt North, Steeles East, L'Amo...	54680.0	2.0
South Steeles, Silverstone, Humbergate, Jamest...	55959.0	2.0

Figure 10. Cluster 3 neighborhoods.

In every one of these there is definitely room for more than one venue for sport.