

## Introduction

As we enter a new year, and as we emerge from a global pandemic where citizens all over the world were confined to their homes as a safety precaution, there are many people who are making personal goals to get fit and exercise more. Therefore, this would be an optimal time for an entrepreneur to open up a gym. Finding the ideal location to open a gym in Toronto would benefit anyone who would like to capitalize on this opportunity.

## Business Problem

The objective of this project is to find the ideal location for an entrepreneur to open up a new gym in Toronto. To accomplish this objective, multiple data science methods will be used such as data scraping, machine learning algorithms, and geographic data visualization to determine the neighborhoods within Toronto that are best-suited for a new gym.

## Data

To accomplish this objective, the following data will be used:

- List of neighborhoods in Toronto
- Geographic coordinates of these neighborhoods
- Foursquare API data on current locations of gyms in Toronto

## Methodology

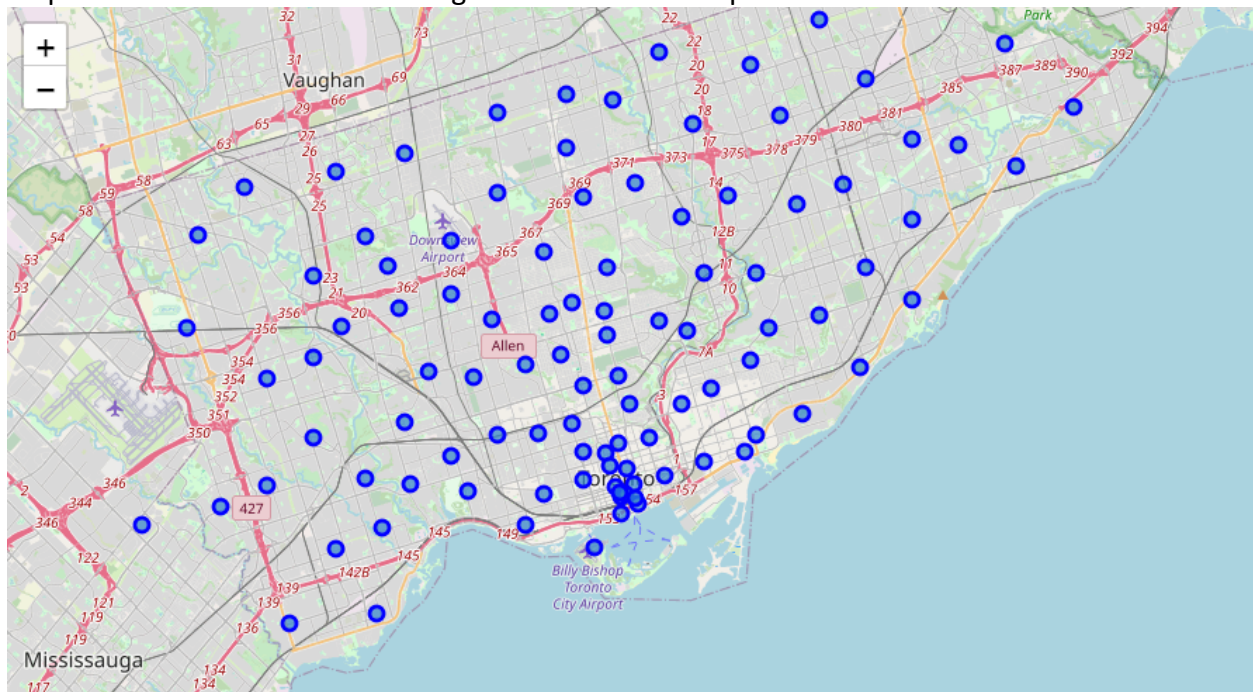
I first extracted a list of neighborhoods in Toronto from Wikipedia ([https://en.wikipedia.org/wiki/List\\_of\\_postal\\_codes\\_of\\_Canada:\\_M](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)). I used beautiful soup to scrape this page and pull the data into an html format, and use it to create a pandas dataframe. After editing the table to eliminate incomplete data and prepare it for analysis, I was ready to go with a dataframe of Toronto boroughs and neighborhoods:

	Postal Code	Borough	Neighbourhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
5	M6A	North York	Lawrence Manor, Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

Then I used a CSV file from a previous IBM Data Science program assignment to grab the geographic coordinates of the Toronto Postal Codes and corresponding neighborhoods, and merge the data with the dataframe I had already established:

	Postal Code	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	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

After establishing the geographic locations of each neighborhood, I used folium to generate a map and add markers for each neighborhood to the map:

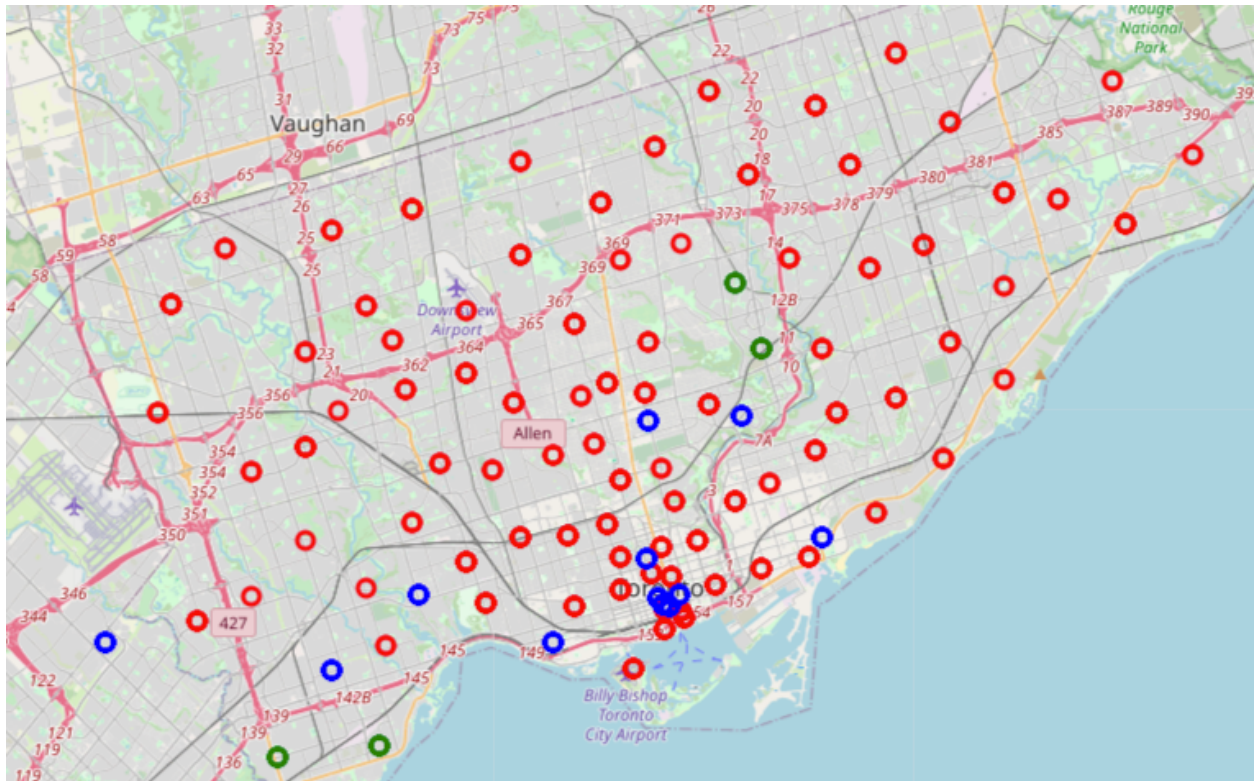


Needing to access current data about number of gyms in each neighborhood, I used API data from Foursquare to pull the list of the top 100 venues within 500 meters of each neighborhood's assigned geographic coordinate point. With this data I can check how many unique categories I can get from the venues that were returned from my query. After verifying that "gym" was one of the unique categories in the Foursquare data, I began analyzing the data to find which neighborhoods that had gyms included in their respective top 100 venues. I grouped the rows of data by neighborhood and getting the average number of occurrence for each category, then specifically the occurrence of gyms.

	Neighborhoods	Gym
0	Agincourt	0.000000
1	Alderwood, Long Branch	0.125000
2	Bathurst Manor, Wilson Heights, Downsview North	0.000000
3	Bayview Village	0.000000
4	Bedford Park, Lawrence Manor East	0.000000
5	Berczy Park	0.000000
6	Birch Cliff, Cliffside West	0.000000
7	Brockton, Parkdale Village, Exhibition Place	0.045455
8	Business reply mail Processing Centre, South C...	0.000000

The final piece of my analysis was to use k-means clustering. The goal of K-means clustering is to have similar samples grouped together in clusters, and dissimilar samples into different clusters. I used this method to group the neighborhoods of Toronto into three clusters based on their frequency of occurrence for "gyms".

## Results



The results from the k-means clustering are shown visually above, and the clusters are color-coded based on their frequency of occurrence of “gyms” in the Foursquare data:

Cluster 0 (green): Neighborhoods with a small occurrence of gyms.

Cluster 1 (blue): Neighborhoods with an average number of gyms.

Cluster 2 (red): Neighborhoods with a large occurrence of gyms.

## Recommendations

Most of the neighborhoods of Toronto have a high number of gyms that exist in the Foursquare top-100 data. There are some Cluster 1 neighborhoods across the southern and central parts of Toronto where there are some gyms, but could possibly have a market ready for more gym options. There are three neighborhoods in cluster 0 where there is a low occurrence of gyms, and an entrepreneur may find opportunities to open a gym and be met with a high demand. Those neighborhoods are:

- Toronto Dominion Centre, Design Exchange
- Station A
- Garden District, Ryerson

## Conclusion

The new year and post-pandemic world lends itself to a growing market for people joining gyms, and Toronto has some neighborhoods that seem to have a demand for more of them. The analysis in this report indicates that there are areas of Toronto where new gyms could be successful. Whether it's in a neighborhood with very few gyms or a neighborhood with some gyms, but could possibly be receptive to more, there are opportunities in Toronto to expand this market.