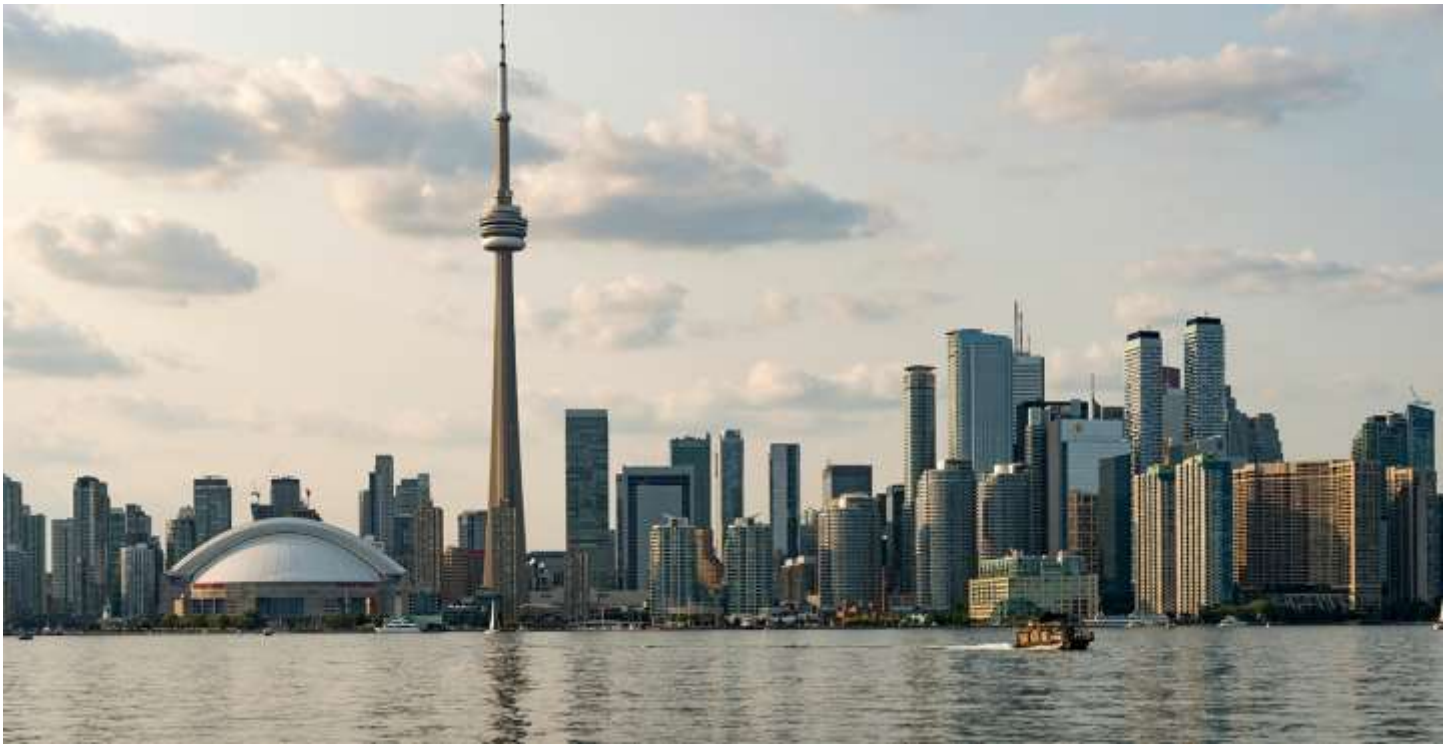
# Coursera Capstone

# IBM Applied Data Science Capstone

# Opening a Gym in Toronto, Canada



**By: Abhijith V**

**August 2020**

## 1. Identifying the Business Problem

Toronto is one of the most densely populated areas in Canada. Toronto is the financial center of Canada and it brings in a plethora of people from different countries. With an estimated population of over 6 million, Toronto is one of the largest cities in Canada with a diverse population. Downtown Toronto is a place where people can view the best of each culture, either while they work or just passing through.

The objective of this project is to use Foursquare location data and regional clustering of venue information to determine what might be the 'best' neighborhood in Toronto to open a Gym. Physical activity and exercises are important for everyone. Regular exercise and physical activity promote strong muscles and bones. It improves respiratory, cardiovascular health, and overall health. Staying active can also help you maintain a healthy weight, reduce your risk for type 2 diabetes, heart disease, and reduce your risk for some cancers. Through this project, we will find the most suitable location for an entrepreneur to open a new Gym in Toronto, Canada.

## 2. Target Audience

The project is mainly focused for Entrepreneurs or Business owners who want to open a new Gym or expand gym outlets to new locations. The analysis will provide vital information that can be used by the target audience.

## 3. Data

One city will be analyzed in this project: Toronto, ON. We will be using the below datasets for analyzing Toronto.

**Data 1**: First dataset is from Wikipedia. The Wikipedia site shown above provided almost all the information about the neighborhoods. It included the Postal code, Borough and the name of the Neighborhoods present in Toronto.

Link to the dataset: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

**Data 2**: Second dataset is Geographical Location data using Geocoder Package which provided us with the Latitudes and Longitudes of the neighborhoods with the respective Postal Codes.

Link to the dataset: https://cocl.us/Geospatial_data

**Data 3**: Venue Data which was pulled using Foursquare API.

## 4. Methodology

4.1 —Cleaning of Data

After all the data was collected and put into data frames, cleansing and merging of the data was required to start the process of analysis. When getting the data from Wikipedia, there were Boroughs that were not assigned to any neighborhood therefore, the following assumptions were made:

1. Cells that have an assigned borough were processed. We ignored Borough's that had values 'Not assigned'.

2. More than one neighborhood for one postal code area was combined into one row separated with a comma.

3. If a cell has a Borough but neighborhood as 'Not assigned', then the neighborhood will have the same as Borough.

| Postal Code | Borough | Neighborhood |
|---|---|---|
| M3A | North York | Parkwoods |
| M4A | North York | Victoria Village |
| M5A | Downtown Toronto | Regent Park, Harbourfront |
| M6A | North York | Lawrence Manor, Lawrence Heights |
| M7A | Downtown Toronto | Queen's Park, Ontario Provincial Government |
| M9A | Etobicoke | Islington Avenue, Humber Valley Village |
| M1B | Scarborough | Malvern, Rouge |
| M3B | North York | Don Mills |
| M4B | East York | Parkview Hill, Woodbine Gardens |
| M5B | Downtown Toronto | Garden District, Ryerson |

We merged the two tables using the Latitude and Longitude collected from the Geocoder package, together based on Postal Code.

| | Postal Code | Latitude | Longitude |
|---|---|---|---|
| 0 | M1B | 43.806686 | -79.194353 |
| 1 | M1C | 43.784535 | -79.160497 |
| 2 | M1E | 43.763573 | -79.188711 |
| 3 | M1G | 43.770992 | -79.216917 |
| 4 | M1H | 43.773136 | -79.239476 |
| 5 | M1J | 43.744734 | -79.239476 |
| 6 | M1K | 43.727929 | -79.262029 |
| 7 | M1L | 43.711112 | -79.284577 |
| 8 | M1M | 43.716316 | -79.239476 |
| 9 | M1N | 43.692657 | -79.264848 |

The venue data pulled from the Foursquare API was merged with the table, providing us with the local venue within a 500-meter radius shown below.

| | 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 | Parkwoods | 43.753259 | -79.329656 | Corrosion Service Company Limited | 43.752432 | -79.334661 | Construction & Landscaping |
| 3 | Victoria Village | 43.725882 | -79.315572 | Victoria Village Arena | 43.723481 | -79.315635 | Hockey Arena |
| 4 | Victoria Village | 43.725882 | -79.315572 | Portugril | 43.725619 | -79.312785 | Portuguese Restaurant |
| 5 | Victoria Village | 43.725882 | -79.315572 | Tim Hortons | 43.725517 | -79.313103 | Coffee Shop |
| 6 | Victoria Village | 43.725882 | -79.315572 | The Fig | 43.727051 | -79.317418 | French Restaurant |
| 7 | Victoria Village | 43.725882 | -79.315572 | Pizza Nova | 43.725824 | -79.312860 | Pizza Place |
| 8 | Regent Park, Harbourfront | 43.654260 | -79.360636 | Roselle Desserts | 43.653447 | -79.362017 | Bakery |
| 9 | Regent Park, Harbourfront | 43.654260 | -79.360636 | Tandem Coffee | 43.653559 | -79.361809 | Coffee Shop |

4.2 —Exploration of data

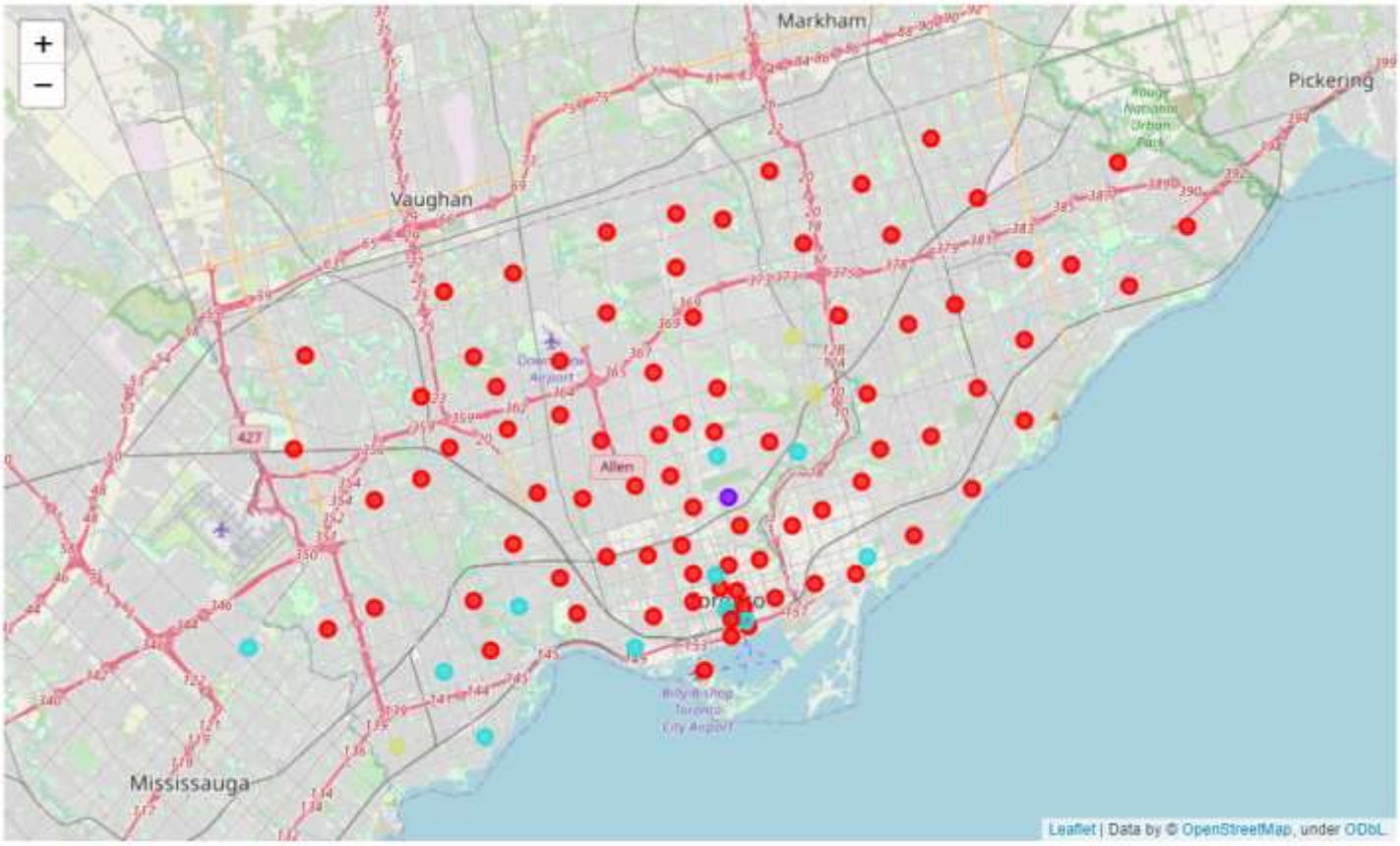
Now after cleansing the data, the next step was to analyze it. We then created a map using Folium and color-coded each Neighborhood depending on what Borough it was located in. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods along with our ID and secret key. The returned venue data is in JSON format and venue name, venue category, venue latitude and longitude are ready to be extracted. We prepare our data for clustering by analyzing each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. Since we are analyzing the “Gym” data, we will filter the “Gym” as venue category for the neighborhoods.

4.3-- K-Means Clustering

We will perform clustering on the data by using k-means clustering. K means clustering is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. The algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for “Gym”. The results will allow us to identify which neighborhoods have higher concentration of Gym while which neighborhoods have fewer number of Gym. Based on the occurrence of Gym in different neighborhoods, it will help us to answer the question, to which neighborhoods are most suitable to open new Gym or to expand.

## 5.Results

The results from the k-means clustering show that we can categorize the neighborhoods into 4 clusters based on the frequency of occurrence for “Gym”: The information is important as we can see that the highest number of Gyms are in Neighborhood in Cluster 1, while Cluster 4 has less number Gym in the neighborhood. The second greatest number of Gyms are formed around Cluster 3 and the least number of Gyms were found in Cluster 2 with only 1 in that neighborhood. The results of the clustering are visualized in the map below with cluster 0 in red, cluster 1 in purple color, and cluster 3 in blue and cluster 4 in green color respectively.



## 6. Conclusion

In conclusion, I had an opportunity to work a business problem, and it was tackled in a way that it was similar to how a genuine data scientist would do. The project utilized numerous Python libraries to fetch the information, control the content and break down and visualize those datasets. Foursquare API was utilized to investigate the settings in neighborhoods of Toronto, get a great measure of data from Wikipedia which were scraped with the Beautifulsoup Web scraping Library. Different plots were also visualized using seaborn and Matplotlib libraries. Similarly, AI strategy was applied to anticipate the error given the information and utilized Folium to picture it on a map. I hope the findings of this project will definitely help the business owners and entrepreneurs to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new gym.