# CAPSTONE PROJECT REPORT

This is a capstone project for IBM Applied data science specialization. In this project a situation is thought about wherein there may be insufficient number of gyms in Toronto. Potential gym operators would care about this problem as this may be a lucrative opportunity for gym operators who may wish to open their gyms in Toronto, due to a shortage of gyms in some locations which may negatively affect the fitness schedule of a lot of city residents, therefore finding appropriate location for opening a gym is quite important, and with this objective in mind, this project has been designed to help potential gym operators identify the most suitable areas.

## BUSINESS PROBLEM

The aim of this project is to find the most suitable location for a potential gym operator to open a gym in Toronto, Canada. By using data science methods and machine learning algorithms like K-means clustering, this project intends to solve the problem: If a gym operator desires to open up a gym in Toronto, what would be the well suited location for opening the gym?

## TARGET AUDIENCE

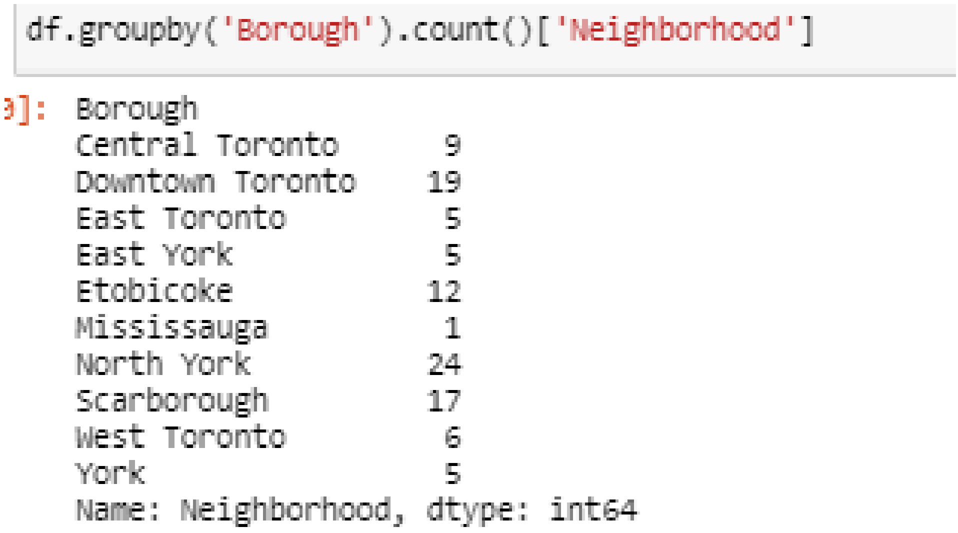
Any gym operator who may be interested to open a new gym in Toronto.

## DATA

To solve this problem, we will need below data:

I. List of Boroughs in Toronto:

This data is required in order to be able to get a sense of the number of boroughs in Toronto, and the number of neighbourhoods corresponding to them.

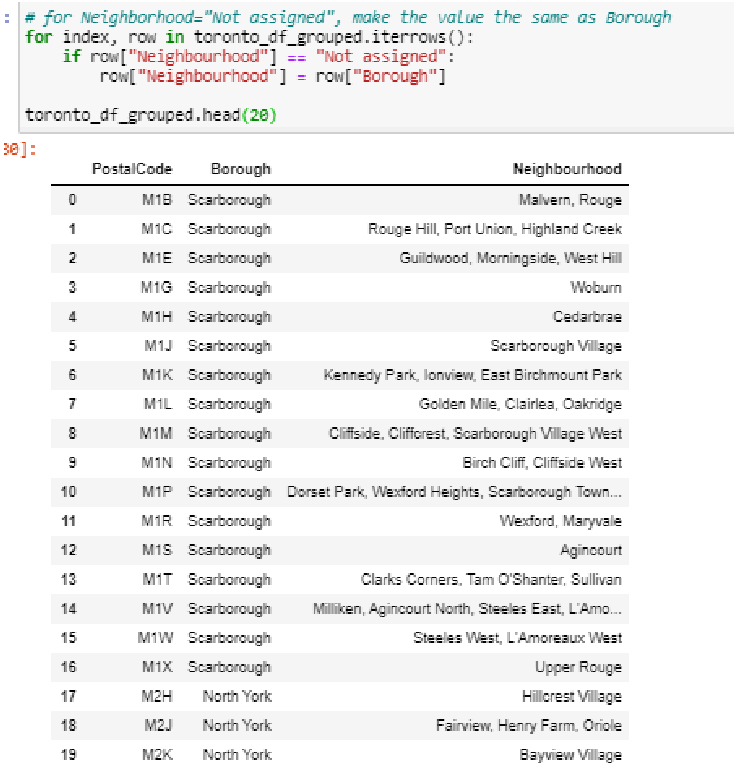


```
df.groupby('Borough').count()['Neighborhood']
```

```
]:  Borough
    Central Toronto      9
    Downtown Toronto    19
    East Toronto         5
    East York            5
    Etobicoke           12
    Mississauga          1
    North York          24
    Scarborough         17
    West Toronto         6
    York                 5
    Name: Neighborhood, dtype: int64
```

## II. List of neighbourhoods in Toronto (displayed 20 of them for example):

This data is required to get the nearby venues, while using the Foursquare API.



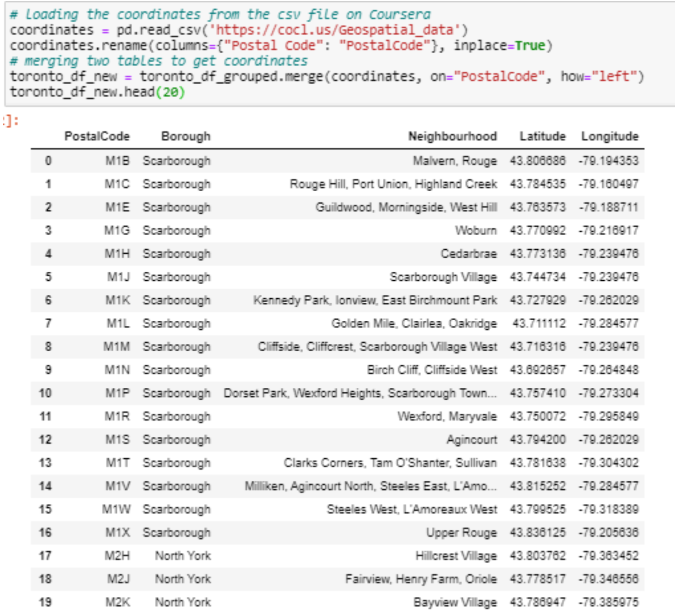
```
# for Neighborhood="Not assigned", make the value the same as Borough
for index, row in toronto_df_grouped.iterrows():
    if row["Neighbourhood"] == "Not assigned":
        row["Neighbourhood"] = row["Borough"]

toronto_df_grouped.head(20)
```

30]:

| | PostalCode | Borough | Neighbourhood |
|---|---|---|---|
| 0 | M1B | Scarborough | Malvern, Rouge |
| 1 | M1C | Scarborough | Rouge Hill, Port Union, Highland Creek |
| 2 | M1E | Scarborough | Guildwood, Morningside, West Hill |
| 3 | M1G | Scarborough | Woburn |
| 4 | M1H | Scarborough | Cedarbrae |
| 5 | M1J | Scarborough | Scarborough Village |
| 6 | M1K | Scarborough | Kennedy Park, Ionview, East Birchmount Park |
| 7 | M1L | Scarborough | Golden Mile, Clairlea, Oakridge |
| 8 | M1M | Scarborough | Cliffside, Cliffcrest, Scarborough Village West |
| 9 | M1N | Scarborough | Birch Cliff, Cliffside West |
| 10 | M1P | Scarborough | Dorset Park, Wexford Heights, Scarborough Town... |
| 11 | M1R | Scarborough | Wexford, Maryvale |
| 12 | M1S | Scarborough | Agincourt |
| 13 | M1T | Scarborough | Clarks Corners, Tam O'Shanter, Sullivan |
| 14 | M1V | Scarborough | Milliken, Agincourt North, Steeles East, L'Amo... |
| 15 | M1W | Scarborough | Steeles West, L'Amoreaux West |
| 16 | M1X | Scarborough | Upper Rouge |
| 17 | M2H | North York | Hillcrest Village |
| 18 | M2J | North York | Fairview, Henry Farm, Oriole |
| 19 | M2K | North York | Bayview Village |

## III. Latitude and Longitude of these neighbourhoods (displayed 20 for example):

This data is required to better understand the location of gyms in neighbourhoods in Toronto and is also needed as an input while using the Foursquare API.



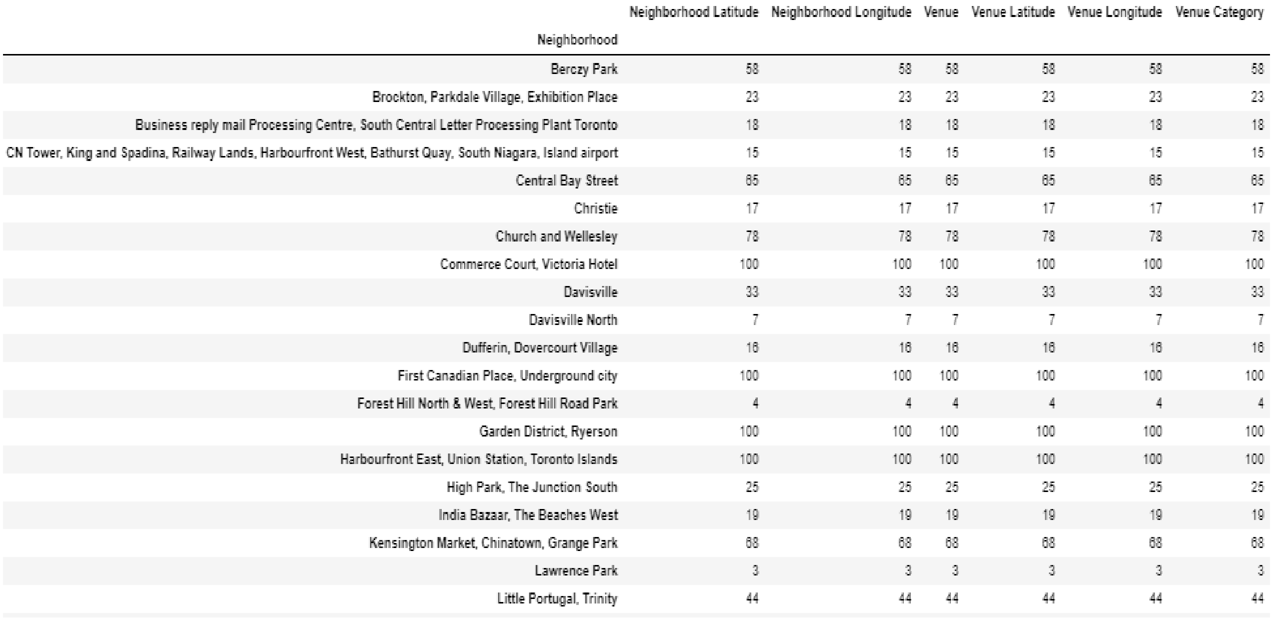
```
# Loading the coordinates from the csv file on Coursera
coordinates = pd.read_csv('https://cocl.us/Geospatial_data')
coordinates.rename(columns={"Postal Code": "PostalCode"}, inplace=True)
# merging two tables to get coordinates
toronto_df_new = toronto_df_grouped.merge(coordinates, on="PostalCode", how="left")
toronto_df_new.head(20)
```

]:

| | PostalCode | Borough | Neighbourhood | Latitude | Longitude |
|---|---|---|---|---|---|
| 0 | M1B | Scarborough | Malvern, Rouge | 43.806686 | -79.194353 |
| 1 | M1C | Scarborough | Rouge Hill, Port Union, Highland Creek | 43.784535 | -79.160497 |
| 2 | M1E | Scarborough | Guildwood, Morningside, West Hill | 43.763573 | -79.188711 |
| 3 | M1G | Scarborough | Woburn | 43.770992 | -79.216917 |
| 4 | M1H | Scarborough | Cedarbrae | 43.773136 | -79.239476 |
| 5 | M1J | Scarborough | Scarborough Village | 43.744734 | -79.239476 |
| 6 | M1K | Scarborough | Kennedy Park, Ionview, East Birchmount Park | 43.727929 | -79.262029 |
| 7 | M1L | Scarborough | Golden Mile, Clairlea, Oakridge | 43.711112 | -79.284577 |
| 8 | M1M | Scarborough | Cliffside, Cliffcrest, Scarborough Village West | 43.716316 | -79.239476 |
| 9 | M1N | Scarborough | Birch Cliff, Cliffside West | 43.692657 | -79.264848 |
| 10 | M1P | Scarborough | Dorset Park, Wexford Heights, Scarborough Town... | 43.757410 | -79.273304 |
| 11 | M1R | Scarborough | Wexford, Maryvale | 43.750072 | -79.295849 |
| 12 | M1S | Scarborough | Agincourt | 43.794200 | -79.262029 |
| 13 | M1T | Scarborough | Clarks Corners, Tam O'Shanter, Sullivan | 43.781638 | -79.304302 |
| 14 | M1V | Scarborough | Milliken, Agincourt North, Steeles East, L'Amo... | 43.815252 | -79.284577 |
| 15 | M1W | Scarborough | Steeles West, L'Amoreaux West | 43.799525 | -79.318389 |
| 16 | M1X | Scarborough | Upper Rouge | 43.836125 | -79.205636 |
| 17 | M2H | North York | Hillcrest Village | 43.803762 | -79.363452 |
| 18 | M2J | North York | Fairview, Henry Farm, Oriole | 43.778517 | -79.346556 |
| 19 | M2K | North York | Bayview Village | 43.786947 | -79.385975 |

IV. Venue data:

This data is required in order to be able to know the number of different venues and to check if gym is present as one of the multiple venues.



| | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|
| Neighborhood | | | | | | |
| Berczy Park | 58 | 58 | 58 | 58 | 58 | 58 |
| Brockton, Parkdale Village, Exhibition Place | 23 | 23 | 23 | 23 | 23 | 23 |
| Business reply mail Processing Centre, South Central Letter Processing Plant Toronto | 18 | 18 | 18 | 18 | 18 | 18 |
| CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport | 15 | 15 | 15 | 15 | 15 | 15 |
| Central Bay Street | 65 | 65 | 65 | 65 | 65 | 65 |
| Christie | 17 | 17 | 17 | 17 | 17 | 17 |
| Church and Wellesley | 78 | 78 | 78 | 78 | 78 | 78 |
| Commerce Court, Victoria Hotel | 100 | 100 | 100 | 100 | 100 | 100 |
| Davisville | 33 | 33 | 33 | 33 | 33 | 33 |
| Davisville North | 7 | 7 | 7 | 7 | 7 | 7 |
| Dufferin, Dovercourt Village | 16 | 16 | 16 | 16 | 16 | 16 |
| First Canadian Place, Underground city | 100 | 100 | 100 | 100 | 100 | 100 |
| Forest Hill North & West, Forest Hill Road Park | 4 | 4 | 4 | 4 | 4 | 4 |
| Garden District, Ryerson | 100 | 100 | 100 | 100 | 100 | 100 |
| Harbourfront East, Union Station, Toronto Islands | 100 | 100 | 100 | 100 | 100 | 100 |
| High Park, The Junction South | 25 | 25 | 25 | 25 | 25 | 25 |
| India Bazaar, The Beaches West | 19 | 19 | 19 | 19 | 19 | 19 |
| Kensington Market, Chinatown, Grange Park | 68 | 68 | 68 | 68 | 68 | 68 |
| Lawrence Park | 3 | 3 | 3 | 3 | 3 | 3 |
| Little Portugal, Trinity | 44 | 44 | 44 | 44 | 44 | 44 |

## DATA EXTRACTION

I. The scrapping of Toronto neighbourhoods from Wikipedia
II. Getting Latitude and Longitude data of neighbourhoods
III. Using Foursquare API to get venue data related to Neighbourhoods

## METHODOLOGY

Firstly, the list of neighbourhoods in Toronto needs to be obtained. The list can be obtained by extracting from the below Wikipedia link:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
The web scraping was done by making use of pandas to pull tabular data directly from a webpage into the data frame.
In order to get the coordinates to be able to use foursquare, the CSV file provided by IBM was utilized, so that the coordinates can be attached next to the Toronto neighbourhoods. After collection of coordinates, map of Toronto is visualized using Folium. Next, the Foursquare API is used
to obtain the list of top 100 venues within 500 meters radius. A Foursquare developer account was made use of  in order to obtain account ID and API key

to extract the required data. The names, categories, latitude, and longitude of the venues were obtained through Foursquare. With this data, the unique categories from these venues can also be checked.

Then, each neighbourhood is analysed by grouping the rows by neighbourhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later.

Here, a condition was made to particularly check for “gym”.

The clustering was done through k-means clustering.

The neighbourhoods in Toronto have been clustered into 3 clusters based on their frequency of occurrence for “Gym”. Based on the results (the concentration of clusters), the optimal location for opening a new gym can be recommended.

## RESULT

### Clusters

The results from k-means clustering can be relied upon to classify Toronto neighbourhoods into 3 clusters, with each cluster having a different number of gyms.

- The red coloured marker represents cluster 0
- The blue coloured marker represents cluster 1
- The green coloured marker represents cluster 2

## RECOMMENDATIONS

- Most of the gyms are in cluster 2, so any potential gym operator should not open their gym in cluster 2, as there will be high competition
- Cluster 1 has one gym, so there will be less competition. Hence, any potential gym operator would be well advised to open a gym here.
- Cluster 0 has 5 gyms, as such any potential gym operator may face low to medium level of competition.
- **Advice:** This project recommends that any potential gym operator should open their gym in cluster 1 as there is low competition due to the presence of only one gym, this will in turn result in high profitability, owing to the low competition