IBM Data Science Course- Capstone Coursework

The Battle of the Neighborhoods

Introduction

I want to find out whether there are enough gyms located in a particular neighborhood in Toronto. Hence, creating more gyms gives more room for people to use the facilities. Thus, prevent the gyms from being overcrowded, as people have more choices for gyms.

Gyms are a facility that improves your health, helps individuals to stay fit and relives stress. This will be the idea that I would be using in this project to find the most suitable location for a gym. The target audience would be for the general public, as anyone who needs a gym would have another one to choose from.

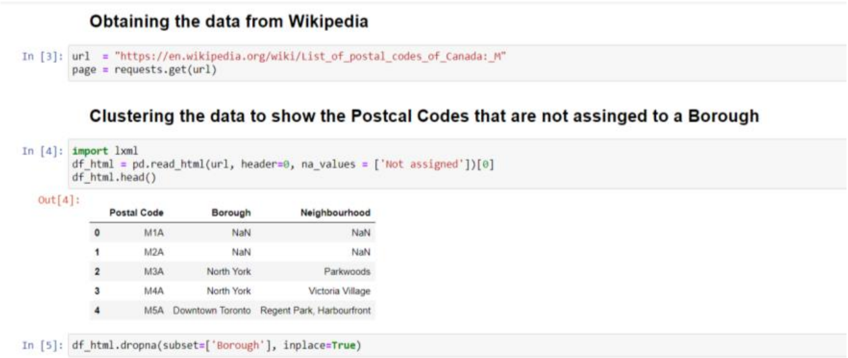
The main aim of this project is to find a location where a gym can be built in Toronto. This will be done by using methods of data science, machine learning, clustering, Foursquare API and IBM Watson Studio, will help to find a solution to the project, as to where can a gym be built?

Data

The data required will be the neighborhoods in Toronto, Canada, which will be obtained from Wikipedia, and will be scrapped to clean the data on IBM Watson Studio. Latitude and longitude are also important factors in this project. Also, using Foursquare API to find other nearby gyms to the area desired. Foursquare API and Geocoder package will be needed to find the data of the areas in Toronto and data of the venues related to the neighborhoods.

Methodology

Pandas HTML was needed to the scrape the data, to make it easier to read and put into a data frame. The data for the list of post codes are obtained from Wikipedia. Second, a table was created with the post codes without a Neighborhood or Borough assigned.



Obtaining the data from Wikipedia

```
In [3]: url  = "https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M"
        page = requests.get(url)
```

Clustering the data to show the Postcal Codes that are not assinged to a Borough

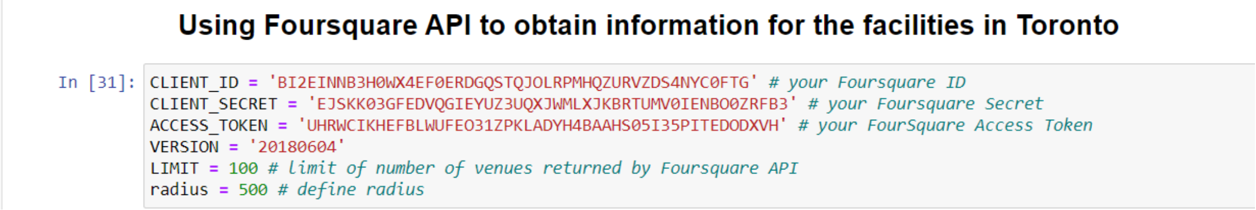
```
In [4]: import lxml
        df_html = pd.read_html(url, header=0, na_values = ['Not assigned'])[0]
        df_html.head()
```

Out[4]:

	Postal Code	Borough	Neighbourhood
0	M1A	NaN	NaN
1	M2A	NaN	NaN
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront

```
In [5]: df_html.dropna(subset=['Borough'], inplace=True)
```

Foursquare API and Geocoder Package were utilized to pull out the coordinates, I.e. Latitude and longitude of the post codes. CSV file was needed to make it work, hence to link the coordinates and make it easier to work. Client ID, Client Secret, and the access token were obtained from Foursquare API, which helped to obtain data about the coordinates and the different categories of the venues in Toronto.

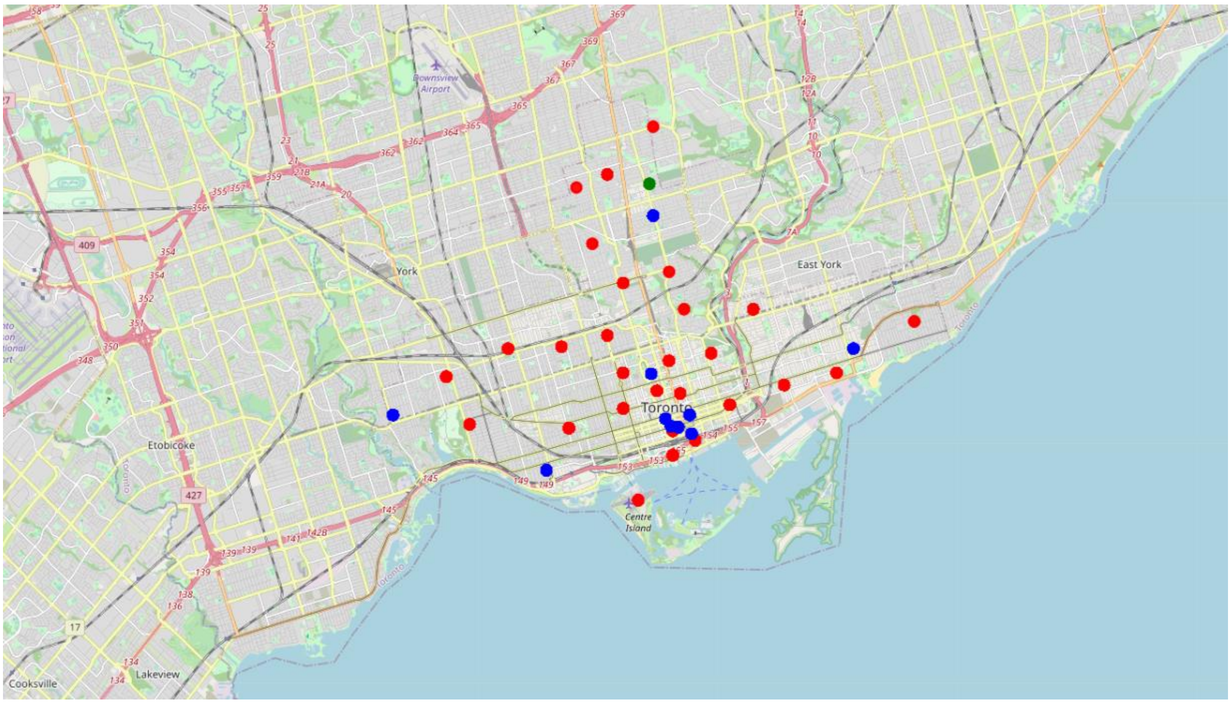


Using Foursquare API to obtain information for the facilities in Toronto

```
In [31]: CLIENT_ID = 'BI2EINNB3H0WX4EF0ERDGQSTQJOLRPMHQZURVZDS4NYC0FTG' # your Foursquare ID
         CLIENT_SECRET = 'EJSKK03GFEDVQGIEYUZ3UQXJWMLXJKBRTUMV0IENBO0ZRFB3' # your Foursquare Secret
         ACCESS_TOKEN = 'UHRWCIKHEFBLWUFEO31ZPKLADYH4BAAHS05I35PITEDODXVH' # your FourSquare Access Token
         VERSION = '20180604'
         LIMIT = 100 # limit of number of venues returned by Foursquare API
         radius = 500 # define radius
```

Folium package was used to create a visual image of the map of Toronto alongside its coordinates. Mean of the frequency of occurrence for different categories of the venues were calculated to cluster and analyse the neighborhoods. K-means clustering was the unsupervised machine learning method to perform the algorithm that finds the k number of the centroids, and locates the data points to its nearest cluster, without overlapping with other data points. The data was clustered into the frequency of the word Gym that occurred. This displayed the preferred locations to build a gym.

Results & Discussion



K-means clustering was used cluster the data in different categories. The Red dots represent the post codes with the lowest number of gyms nearby. The green dot has the highest number of gyms nearby. Finally, the blue dots are the second highest number of gyms nearby.

As the list of the gym venues were discovered that most of the gyms in Toronto are around the center of Toronto and few in the other areas. When creating the k-means machine learning cluster, it can be grouped into two clusters, cluster 1 with <5 which represents the red dots, and cluster 2 with >5, which represents the green dot and the blue dots. Therefore, it would be ideal to build a gym outside of the central of Toronto. This may give more ease to any fanatics of the gym who have travel from other areas of Toronto or outside Toronto to the center of Toronto. In particular areas such as, Church and Wellesley, Garden District and Toronto Dominion Centre.

Conclusion

A business model was built to cluster the neighborhoods in Toronto, in which the data was extratced from Wikipedia. Foursquare API and Geocoder Package was essential to extract the coordinates and the different venues around each neighborhood in Toronto. In conclusion, a gym that was built outisde the center of Toronto would be ideal.