Capstone Project: Location Data Analysis of Toronto to Open a Fitness Centre

1. Introduction

1.1. Background & Business Problem

According to 2019 ParticipACTION Adult Report Card conducted by Canadian Fitness and Lifestyle Research Institute, 29% of Canadian adults (aged 18 to 79 years old) had low active lifestyle while 18% of them were observed having sedentary lifestyle [1]. Moreover, only 16% of Canadian adults achieved the national guideline of at least 150 minutes of weekly MVPA (Moderate to Vigorous Physical Activity). Engaging sedentary behaviors such as sitting in a reclined position or using computer for extended period of time may increase the risks of obesity, cardiovascular diseases and many more. Reducing sedentary behaviors and increasing physical activity can benefit to many individuals, especially working adults who spend most of their time in the office throughout the day. In addition, a national study investigating work-life balance in Canada revealed that many working Canadians were overloaded by the demands of work and family [2].

Thus, in recent years, the commitment in promoting physical activity to get healthier lifestyle was initiated by the government and different organizations. Health and fitness industry has grown exponentially over the last five years due to increasing health consciousness of Canadians. Hence, in this project, as a data scientist, I would like to explore the neighborhoods in Toronto, Canada to open a yoga studio or fitness centre. Toronto, the capital city of the province of Ontario, is a cosmopolitan and most populous city in Canada. Being the international centre for businesses and crowded city with working professionals, Toronto is a good place to consider opening a fitness centre for those who would like to promote active lifestyle among Canadians.

Problem Statement – To choose one of the best neighborhoods in Toronto to open a yoga studio or fitness centre to promote healthy lifestyle

1.2. Target Audience

This section explains the target audience of this project. In other words, it explains who would be interested or which group of people would be benefited from this project. First of all, entrepreneurs who would like to start a business in health and fitness industry in Toronto will be interested. The location data analysis in this project will be helpful for them to choose the best neighborhoods in Toronto to start their own business. Moreover, the analysis will

guide them in predicting profit and loss of business by locations. Second, the potential customers such as working Canadians in Toronto who would like to engage in active lifestyle can explore neighborhoods with many health and fitness options. Business personnel who would like to invest in health and fitness industry can also be benefited from this project. In addition, this project would be able to tell a story of exploring different neighborhoods by analyzing the datasets and applying machine learning techniques so that other data scientists can explore them as well.

2. Data Processing

2.1. Data Sources

The data sources for this project are listed as below:

- To get the list of all the neighborhoods in Toronto, I will be utilizing web scraping technique to extract the content from Wikipedia page of "List of postal codes of Canada: M" (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). The postal code, borough and neighborhood names will be presented in panda DataFrame.
- To get the geospatial data of all the neighborhoods that includes geographical coordinates, Geospatial Coordinates (https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs_v1/Geospatial_Coordinates.csv) csv file will be used. Next, the two DataFrames will be combined to get the information of postal code, borough, neighborhood names and their coordinates in one DataFrame.
- Foursquare API will be used to explore the most common venues in Toronto. Venue information such as names, categories, latitude, longitude, etc. will be collected. Specifically, I will be exploring the details of neighborhoods that have more offices since it will be good to focus on office workers as potential customers due to exposing more sedentary lifestyle.
- Since our interest is to open a fitness centre, the required category IDs will be retrieved from Foursquare developer website at https://developer.foursquare.com/docs/build-with-foursquare/categories/.

2.2. Data Acquisition

The content from Wikipedia page of "List of postal codes of Canada: M" are scraped in order to get the information of Toronto neighborhoods. The first DataFrame consists of Postal Cdoe, Borough and Neighborhood names (see Figure 1). If the name of neighbourhood is 'not assigned', it is replaced with the name of borough.

	PostalCode	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Queen's Park	Ontario Provincial Government
...	...	...	...
98	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North
99	M4Y	Downtown Toronto	Church and Wellesley
100	M7Y	East Toronto Business	Enclave of M4L
101	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...
102	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...

103 rows × 3 columns

Figure 1: DataFrame scraped from Wikipedia page after cleaning

After getting the clean DataFrame, the geographical coordinates (i.e., latitude and longitude) of each neighborhood in Toronto is extracted from the existing csv file. Since the geographical DataFrame only consists of postal code, latitude and longitude columns, I combined the two DataFrame into one based on the postal code. After the cleaning process, there are a total of 15 boroughs and 103 neighborhoods in final DataFrame (see Figure 2 below).

In [6]:
```
#Merging two dataframe
lat_lng.rename(columns={'Postal Code':'PostalCode'},inplace=True)
df_toronto = pd.merge(df,lat_lng)
df_toronto.head(10)
```

Out[6]:

	PostalCode	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	Queen's Park	Ontario Provincial Government	43.662301	-79.389494
5	M9A	Etobicoke	Islington Avenue	43.667856	-79.532242
6	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
7	M3B	North York	Don Mills North	43.745906	-79.352188
8	M4B	East York	Parkview Hill, Woodbine Gardens	43.706397	-79.309937
9	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937

In [7]:
```
#Checking how many unique boroughs and neighborhoods
print('The dataframe has {} boroughs and {} neighborhoods.'.format(
        len(df_toronto['Borough'].unique()),
        df_toronto.shape[0]
    )
)
```

```
The dataframe has 15 boroughs and 103 neighborhoods.
```

Figure 2: DataFrame after merging

3. Methodology

Based on the business problem of this project, the data will be explored and analyzed in the following steps.

- explore the nearby offices, gyms and yoga studios of all the neighborhood in each borough
- decide which borough(s) have more offices and popular among working professionals
- delve into each neighborhood of selected borough
- cluster the neighborhoods of selected borough by predictive modelling and recommend the location

In addition, the following factors are considered to decide the suitable location or neighborhood.

- number of nearby offices
- number of nearby gym/fitness centres, and
- number of nearby yoga studios

3.1. Exploratory Data Analysis

After scraping the necessary neighborhood data and latitude and longitude of each neighborhood, I will be exploring nearby offices, gyms and fitness centres, and yoga studios in all neighborhoods in Toronto. Specifically, the goal is to open a yoga studio or fitness centre for office workers, the number of nearby offices and gyms in the neighbourhoods is explored.

Exploring Nearby Offices, Gyms and Yoga Studios in all the Neighborhood

For this project, I fetched the category of offices, gym/fitness centres and yoga studios from Foursquare. I set the limit as 100 venues within the radius of 500 metres for each neighborhood. The required category IDs are retrieved from Foursquare developer website at https://developer.foursquare.com/docs/build-with-foursquare/categories/.

	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	No. of Nearby Offices	No. of Nearby Gym/Fitness Centres	No. of Nearby Yoga Studios
0	North York	Parkwoods	43.753259	-79.329656	5	0	0
1	North York	Victoria Village	43.725882	-79.315572	3	0	0
2	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	23	3	2
3	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	6	4	0
4	Queen's Park	Ontario Provincial Government	43.662301	-79.389494	22	9	2

Figure 3: No. of nearby offices, gyms and yoga studios in all neighborhoods

After that, the data is grouped by borough and calculated the total number of nearby offices, gyms, and yoga studios in each borough, so that it will be clear to know which borough should we focus on.

	Borough	No. of Nearby Offices	No. of Nearby Gym/Fitness Centres	No. of Nearby Yoga Studios
0	Central Toronto	39	21	9
1	Downtown Toronto	631	208	39
2	Downtown Toronto Stn A	54	15	2
3	East Toronto	25	19	5
4	East Toronto Business	9	4	1

Figure 4: Total no. of nearby offices, gyms and yoga studios by Borough

Let’s visualize and see which borough has more offices.

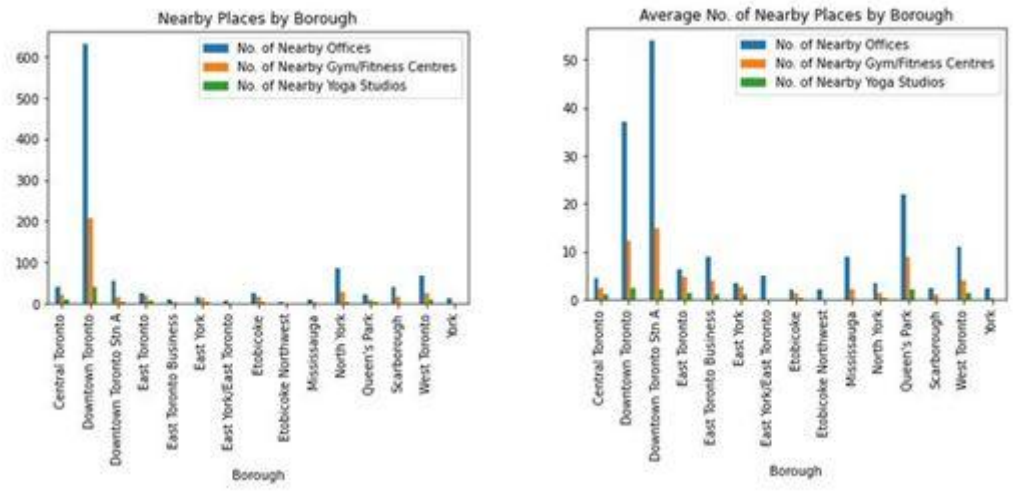

Figure 5: No. of nearby places and average no. of nearby places by Borough

3.2. Location Data Analysis

After exploring the neighborhoods in each borough, as can be seen in tables and bar charts, 'Downtown Toronto' is the most popular among working professional with more than 600 nearby offices. Being the central business district, the area is surrounded by city landmarks and popular places such as retail centres and shops, hundreds of restaurants, hotels, boutiques, etc. So, let's investigate more of this area since this is promising area to start a business.

First, let's see only the neighborhoods in 'Downtown Toronto' borough.

```
In [17]: df_downtown = df_toronto[df_toronto['Borough'] == 'Downtown Toronto'].reset_index(drop=True)
         df_downtown
```

Out[17]:

	PostalCode	Borough	Neighbourhood	Latitude	Longitude
0	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
1	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937
2	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418
3	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306
4	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383
5	M6G	Downtown Toronto	Christie	43.669542	-79.422564
6	M5H	Downtown Toronto	Richmond, Adelaide, King	43.650571	-79.384568
7	M5J	Downtown Toronto	Harbourfront East, Union Station, Toronto Islands	43.640816	-79.381752
8	M5K	Downtown Toronto	Toronto Dominion Centre, Design Exchange	43.647177	-79.381576
9	M5L	Downtown Toronto	Commerce Court, Victoria Hotel	43.648198	-79.379817
10	M5S	Downtown Toronto	University of Toronto, Harbord	43.662696	-79.400049
11	M5T	Downtown Toronto	Kensington Market, Chinatown, Grange Park	43.653206	-79.400049
12	M5V	Downtown Toronto	CN Tower, King and Spadina, Railway Lands, Har...	43.628947	-79.394420
13	M4W	Downtown Toronto	Rosedale	43.679563	-79.377529
14	M4X	Downtown Toronto	St. James Town, Cabbagetown	43.667967	-79.367675
15	M5X	Downtown Toronto	First Canadian Place, Underground city	43.648429	-79.382280
16	M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160

Figure 6: All neighbourhoods in Downtown Toronto

Folium is a python library to visualize the geographic details and I used it to create an interactive map of Downtown Toronto. The coordinate data of latitude and longitude values are used to visualize the neighbourhoods in Downtown Toronto. The code snippet and the map is shown is Figure 7 below.

```
# create map of Downtown Toronto using latitude and longitude values
map_downtown = folium.Map(location=[43.6563221,-79.3809161], zoom_start=11)

# add markers to map
for lat, lng, label in zip(df_downtown['Latitude'], df_downtown['Longitude'], df_downtown['Neighbourhood']):
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='blue',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(map_downtown)

map_downtown
```

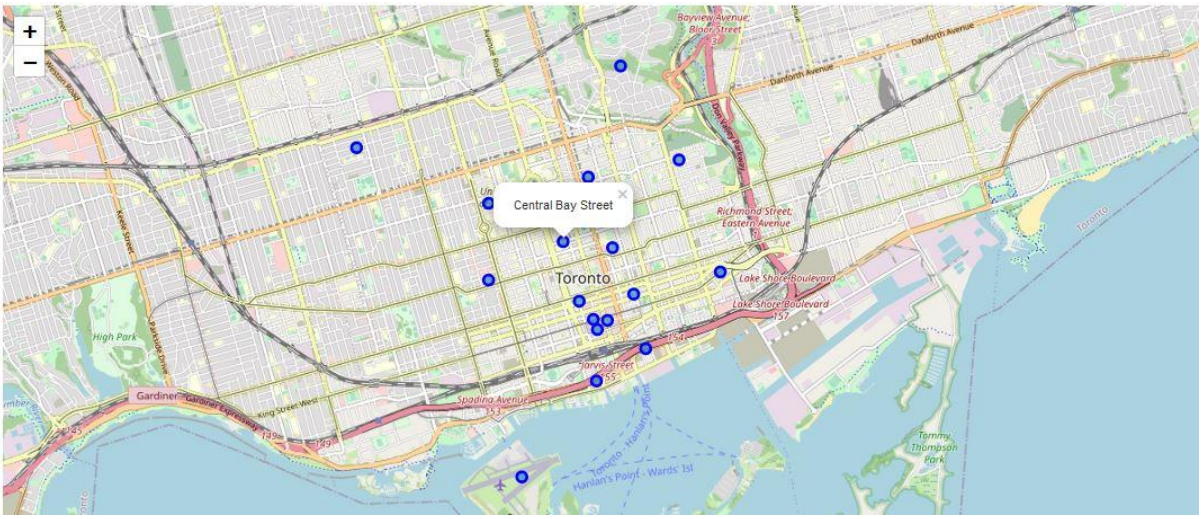

Figure 7: Code snippet and folium map of neighbourhoods in Downtown Toronto

Now, number of nearby offices, gyms and yoga studios in Downtown Toronto is extracted.

	Borough	Neighborhood	No. of Nearby Offices	No. of Nearby Gym/Fitness Centres	No. of Nearby Yoga Studios
0	Downtown Toronto	Regent Park, Harbourfront	23	3	2
1	Downtown Toronto	Garden District, Ryerson	47	18	4
2	Downtown Toronto	St. James Town	63	17	2
3	Downtown Toronto	Berczy Park	53	7	1
4	Downtown Toronto	Central Bay Street	57	19	3
5	Downtown Toronto	Christie	7	4	2
6	Downtown Toronto	Richmond, Adelaide, King	59	24	3
7	Downtown Toronto	Harbourfront East, Union Station, Toronto Islands	43	13	0
8	Downtown Toronto	Toronto Dominion Centre, Design Exchange	70	18	1
9	Downtown Toronto	Commerce Court, Victoria Hotel	75	22	2
10	Downtown Toronto	University of Toronto, Harbord	6	7	4
11	Downtown Toronto	Kensington Market, Chinatown, Grange Park	20	4	8
12	Downtown Toronto	CN Tower, King and Spadina, Railway Lands, Har...	2	0	0
13	Downtown Toronto	Rosedale	4	0	0
14	Downtown Toronto	St. James Town, Cabbagetown	11	9	2
15	Downtown Toronto	First Canadian Place, Underground city	68	22	1
16	Downtown Toronto	Church and Wellesley	23	21	4

Figure 8: No. of nearby offices, gyms and yoga studios in Downtown Toronto

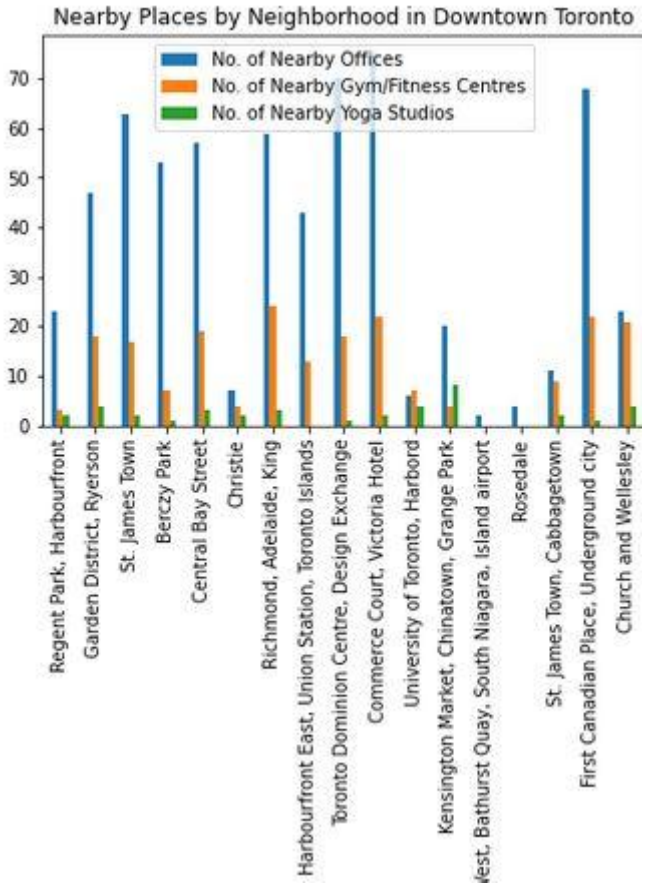

Figure 9: Bar Chart of nearby places in Downtown Toronto

The percentage of nearby gym/fitness centres and yoga studios are also visualized in the following figure.

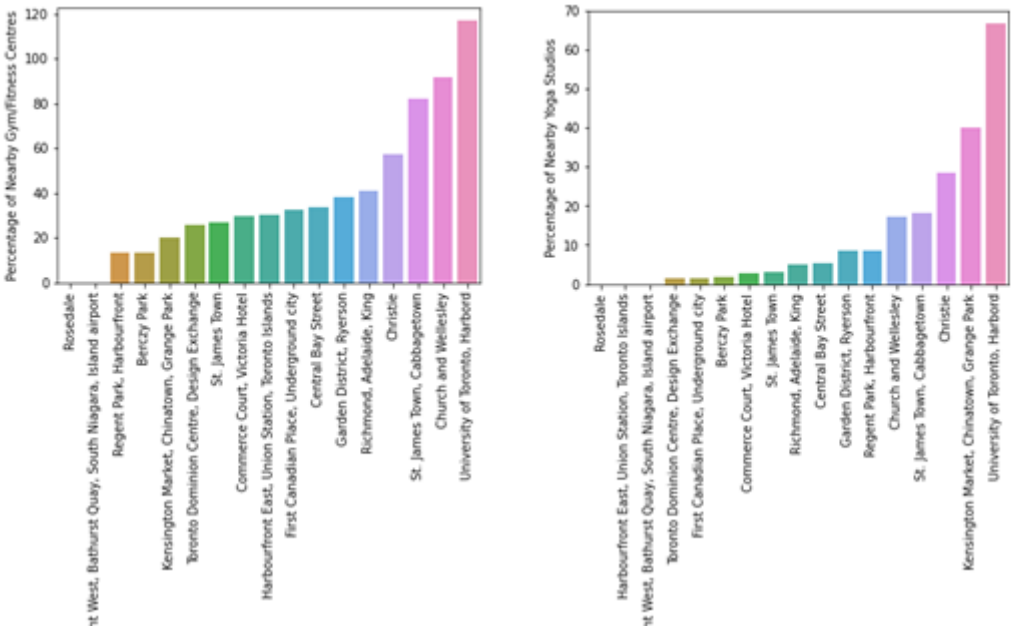

Figure 10: Percentage of nearby gyms and yoga studios in Downtown Toronto

3.3. Predictive Modelling

Clustering Neighborhood by K-means clustering

To cluster the neighborhoods, unsupervised machine learning technique K-means algorithm will be adopted in this project. First, the elbow method is used to identify the optimal k value in a given dataset. As shown in the figure below, the best k value seems to be 3 after analyzing elbow method using distortions.

```
nearby_downtown = nearby_downtown.drop(['Borough', 'Neighborhood'], 1)

distortions=[]
for k in range(1,10):
    km=KMeans(n_clusters=k)
    km.fit(nearby_downtown)
    distortions.append(km.inertia_)

plt.figure(figsize=(16,8))
plt.plot(range(1,10), distortions, 'bx-')
plt.xlabel('k')
plt.ylabel('Distortions')
plt.title('The Elbow Method showing the optimal k')
plt.show()
```

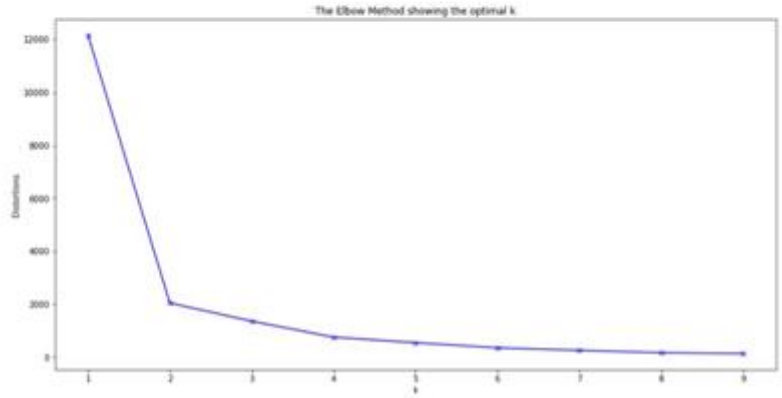

Figure 11: Elbow method to identify best k value

After clustering with k=3, here is the merged DataFrame of combining with cluster label column.

	Neighborhood	Cluster Labels	No. of Nearby Offices	No. of Nearby Gym/Fitness Centres	No. of Nearby Yoga Studios
0	Regent Park, Harbourfront	1	23	3	2
1	Garden District, Ryerson	0	47	18	4
2	St. James Town	2	63	17	2
3	Berczy Park	0	53	7	1
4	Central Bay Street	2	57	19	3
5	Christie	1	7	4	2

Figure 12: DataFrame with cluster labels for each neighborhood in Downtown Toronto

Examine Each Cluster

We have a total of three clusters and let's examine each cluster to identify and recommend the most promising neighborhood to open a yoga studio or a fitness centre.

Figure 13 shows the summary of Cluster 0.

```
In [30]: nearby_downtown.loc[nearby_downtown['Cluster Labels'] == 0]
```

Out[30]:

	Neighborhood	Cluster Labels	No. of Nearby Offices	No. of Nearby Gym/Fitness Centres	No. of Nearby Yoga Studios
1	Garden District, Ryerson	0	47	18	4
3	Berczy Park	0	53	7	1
7	Harbourfront East, Union Station, Toronto Islands	0	43	13	0

```
In [31]: nearby_downtown.loc[nearby_downtown['Cluster Labels'] == 0].mean()
```

```
Out[31]: Cluster Labels                    0.000000
         No. of Nearby Offices            47.666667
         No. of Nearby Gym/Fitness Centres 12.666667
         No. of Nearby Yoga Studios        1.666667
         dtype: float64
```

Figure 13: Code snippet to retrieve Cluster 0

Figure 14 below highlights the summary of Cluster 1.

```
In [32]: nearby_downtown.loc[nearby_downtown['Cluster Labels'] == 1]
```

Out[32]:

	Neighborhood	Cluster Labels	No. of Nearby Offices	No. of Nearby Gym/Fitness Centres	No. of Nearby Yoga Studios
0	Regent Park, Harbourfront	1	23	3	2
5	Christie	1	7	4	2
10	University of Toronto, Harbord	1	6	7	4
11	Kensington Market, Chinatown, Grange Park	1	20	4	8
12	CN Tower, King and Spadina, Railway Lands, Har...	1	2	0	0
13	Rosedale	1	4	0	0
14	St. James Town, Cabbagetown	1	11	9	2
16	Church and Wellesley	1	23	21	4

```
In [33]: nearby_downtown.loc[nearby_downtown['Cluster Labels'] == 1].mean()
```

```
Out[33]: Cluster Labels                    1.00
         No. of Nearby Offices            12.00
         No. of Nearby Gym/Fitness Centres 6.00
         No. of Nearby Yoga Studios        2.75
         dtype: float64
```

Figure 14: Code snippet to retrieve Cluster 1

The code snippet and summary of Cluster 2 is shown in Figure 15.

```
In [34]: nearby_downtown.loc[nearby_downtown['Cluster Labels'] == 2]
```

Out[34]:

	Neighborhood	Cluster Labels	No. of Nearby Offices	No. of Nearby Gym/Fitness Centres	No. of Nearby Yoga Studios
2	St. James Town	2	63	17	2
4	Central Bay Street	2	57	19	3
6	Richmond, Adelaide, King	2	59	24	3
8	Toronto Dominion Centre, Design Exchange	2	70	18	1
9	Commerce Court, Victoria Hotel	2	75	22	2
15	First Canadian Place, Underground city	2	68	22	1

```
In [35]: nearby_downtown.loc[nearby_downtown['Cluster Labels'] == 2].mean()
```

```
Out[35]: Cluster Labels                  2.000000
         No. of Nearby Offices          65.333333
         No. of Nearby Gym/Fitness Centres    20.333333
         No. of Nearby Yoga Studios      2.000000
         dtype: float64
```

Figure 15: Code snippet to retrieve Cluster 2

I have also visualized the bar chart of nearby offices, gyms and yoga studios by clusters to clearly see the differences.

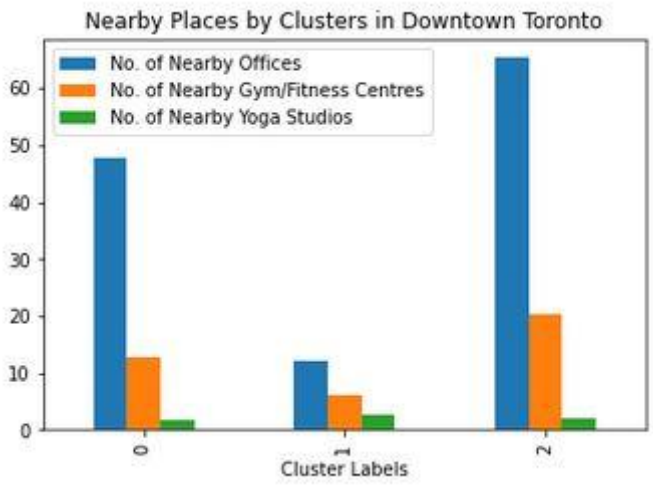

Figure 16: Bar chart of nearby places by clusters

Moreover, the folium map of each cluster in Downtown Toronto is also created.

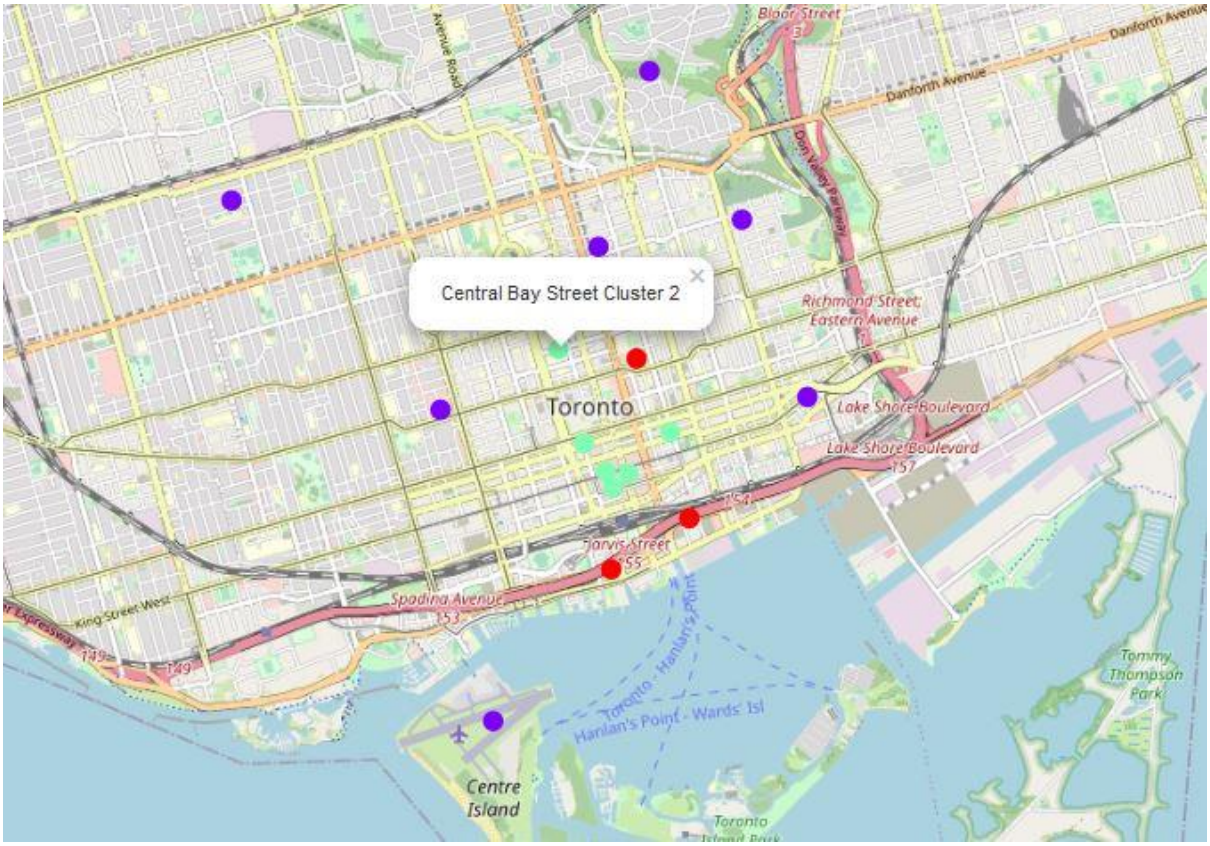

Figure 17: Folium map of each cluster in Downtown Toronto

The analysis and visualization of bar charts and cluster map helps us in identifying the most promising neighborhood to open a yoga studio or fitness centre. Cluster 0 represents the middle cluster with average no. of 48 offices, 13 gym/fitness centres and 2 yoga studios nearby. Cluster 1 has the least average no. of places with 12 offices, 6 gym/fitness centres and 3 yoga studios. Meanwhile, Cluster 3 stands out with the highest average no. of nearby places with 65 offices, 20 gym/fitness centres and 2 yoga studios.

When we look at the cluster map, Cluster 0 (shown in red color in the map) and Cluster 1 (shown in blue color in the map) are sparsed across the Downtown Toronto while neighborhoods in Cluster 2 (shown in green color in the map) are grouped together which seem to be popular places with a lot of offices. Thus, I decided to take a closer look at Cluster 2 and explore the area.

Figure 18: A closer look of Cluster 2 on Folium map

Financial District, Toronto

As shown in the map (Figure 18), three of the six neighborhoods in Cluster 2 are in Financial District. Finacial District, Toronto is considered to be the main business district with many office towers and centre of Canada's financial industry. For this project, I do not want the new business place to be in the secluded area with little number of offices and gyms. Thus, let's explore the neighborhood around financial district. I am curious about the popular spots within 500 metres of the financial district area. So, let's input the latitude and longitude of financial district and url first. I used Foursquare API to get the info about the locations.

Based on the data retrieved from Foursquare API, there are about 30 popular places around Financial District, representing different categories such as restaurants, café, gym, bakery, etc. As our interest is to open a fitness centre, I explore specifically on how many gym or fitness centres and offices near the area. I set the limit of 100 within the 500 metres of the area. After the analysis, the total number of offices near Financial District is 72. The total number of gym/fitness centres and yoga studios are 19 and 1 respectively.

4. Results and Discussion

The main goal of this project is to analyze the neighborhood location to decide where to open a yoga studio or fitness centre to promote healthier lifestyle for Canadian working professionals. Thus, the locations and places in this project mainly refer to offices, gym/fitness centres and yoga studios. After exploring the nearby places in each borough, Downtown Toronto has the most number of offices and gyms nearby, followed by North York and West Toronto. Thus, Downtown Toronto is focused for more in-depth analysis.

The analysis shows that there are many neighborhood candidates in Downtown Toronto. With the help of predictive modelling, specifically K-means clustering helps to identify the different clusters of each neighborhood. After analyzing the elbow method, I used the best value of k=3 to cluster the neighborhoods. Cluster 0 represents the middle cluster with average no. of 48 offices, 13 gym/fitness centres and 2 yoga studios nearby. Cluster 1 has the least average no. of places with 12 offices, 6 gym/fitness centres and 3 yoga studios. Meanwhile, Cluster 2 stands out with the highest average no. of nearby places with 65 offices, 20 gym/fitness centres and 2 yoga studios.

When choosing the best location to run the fitness business in this project, the more popular places with more offices is mainly considered. The decision factor(s), of course, will be different for each individual and business personnel. Thus, the optimal places to open a business will be different based on business problem and decision criteria. Here, I would recommend exploring more on Financial District area in Downtown Toronto. According to data analysis, there are 72 nearby offices within 500 metres of Financial District, in which there are 19 gym/fitness centres and only 1 yoga studios. Thus, starting a yoga studio in the area would not be a bad idea.

Some limitations of the analysis would be the location data obtained in this project which is completely from Foursquare API and the decision criteria is mainly subjective. However, this could be the starting point for more detailed analysis and insightful recommendations of future studies.

5. Conclusion

To conclude the project, due to the sedentary lifestyle and spending most of the time in offices, working Canadians have difficulty in work-life balance and they are overloaded by demand of work and family. So, the idea is to choose one of the best neighborhoods in Toronto to open a yoga studio or fitness centre to promote healthy lifestyle. First, after collecting the necessary data, the nearby offices, gyms and yoga studios of all the neighborhoods in each borough in Toronto was explored. Second, the borough with more offices and popular among working professionals was selected and delved into each neighborhoods of selected borough which was Downtown Toronto. Next, the predictive modelling helped to identify the three clusters of Downtown Toronto and recommended the location.

As mentioned before, the decision criteria to choose the optimal place(s) to run a yoga studio or a fitness centre will be different for each business personnel and there are not only locations but other additional factors to consider as well. However, this project can hopefully act as an initial concept for other business ideas using data science.

References

[1] ParticipACTION, 2019 Adult report card (https://www.participaction.com/en-ca/resources/adult-report-card)
[2] Bulletin 05: Perceived Life Stress (https://cflri.ca/bulletin-05-perceived-life-stress)