Coursera Capstone

1 – Introduction

Problem Background :
Joining a gym is a popular, and effective way to get your health in check. However, not any neighborhoods of a given city have a gym, and, unfortunately, these are the neighborhoods where physical activity is needed the most. In this project, we focus on the city of Toronto. Highly populated, upper-class areas of the city are in no shortage of places where one could practice a sport or some sort of physical activity. However, there are some neighborhoods where gyms, pools, or parks are rare. This begs the question of finding the optimal location for a gym, taking into account, of course, the supply and demand, but also the need for residents to have some sort of physical activity, by choosing a neighborhood where the residents are sedentary and unhealthy.
Problem Description and Target Audience
This is clearly a problem that a low-cost gym chain (the target audience) should solve in order to avoid being in locations that are too competitive, as well as to have the most societal impact by promoting health and fitness in sedentary neighborhoods. In particular, we will need to find the neighborhoods that have the least amount of locations where one can exercise (gyms, pools, parks) And that are the most likely to be unhealthy. This will require using Foursquare and public health data.
Success Criteria :
The success criteria of this project will be a good recommendation of the neighborhood choice in Toronto for an eventual opening of a new franchise for a low-cost gym chain based on 2 key factors : lack of available gym and other locations where one can exercise a physical activity, and the second factor is the general unhealthiness of the neighborhood.

2– Data

The key data for this project is the Toronto neighborhoods data, and the health data.

For this, we will use Urban HEART data (http://www.torontohealthprofiles.ca/urbanheartattoronto.php). Urban HEART stands for "Urban Health Equity Assessment and Response Tool." Urban HEART is a framework that a variety of organizations with diverse mandates can use together to maximize their collective impact on equity. This framework provides opportunities for collaboration while allowing organizations to continue to focus on their unique roles and mandates. Urban HEART can shed light on where, why and how our diverse initiatives are/aren't converging to produce real change.

More precisely, the metrics we will use are the number of gym, parks, and pools, to assess the availability of sport activities for each neighborhood, and also the Walk Score from the Urban HEART dataset, which is a coefficient that indicates the "walkability" of each neighborhood, and also the diabetes prevalence in each neighborhood, which is a great indicator of the need for physical activity.

We will also need the number of fast food restaurants in each neighborhood to have another indicator of the general unhealthiness of each neighborhood.

The number of gyms,parks,fast food restaurants, and pools in each neighborhood are found by using the Foursquare API, using neighborhood latitude and longitude that was provided by the geopy package.

This data will be used to cluster the neighborhoods in order to find the optimal gym location, ie the unhealthiest neighborhood with the least amount of gyms, parks, and pools.

Neighborhood	Walk score	Diabetes	Latitude	Longitude	Gyms	Parks	FastFood	Pools
Agincourt North	66.0	9.5	43.808038	-79.266439	0	0	4	0
Alderwood	70.0	8.5	43.601717	-79.545232	1	0	1	1
Annex	94.0	5.5	43.670338	-79.407117	3	3	4	1
Banbury-Don Mills	67.0	6.5	43.751672	-79.370169	0	1	0	0
Bathurst Manor	61.0	8.5	43.665519	-79.411937	7	2	5	1
Bay Street Corridor	99.0	5.1	43.672798	-79.390734	21	6	7	8
Bayview Village	71.0	6.0	43.769197	-79.376662	0	0	2	0
Bayview Woods-Steeles	57.0	7.1	43.798127	-79.382973	0	0	0	0
Bendale	64.0	11.5	43.753520	-79.255336	0	2	1	0
Black Creek	62.0	12.7	45.622607	-77.156242	0	0	0	0

3 – Methodology

The first step was to import the Urban HEALTH data from a publicly available dataset (UrbanHeart_MatrixData.xlsx). This dataset contains many inequity indicators between Toronto's neighborhoods, but what we are interested in is the diabetes prevalence and the Walk score for each neighborhood. The walk score indicates the "walkability" of each neighborhood, i.e., the higher the score, the more the neighborhood residents can walk during their daily activities.

	Neighborhood	Walk score	Diabetes
0	Agincourt North	66.0	9.5
1	Agincourt South-Malvern West	66.0	9.5
2	Alderwood	70.0	8.5
3	Annex	94.0	5.5
4	Banbury-Don Mills	67.0	6.5
5	Bathurst Manor	61.0	8.5
6	Bay Street Corridor	99.0	5.1
7	Bayview Village	71.0	6.0
8	Bayview Woods-Steeles	57.0	7.1
9	Bedford Park-Nortown	73.0	5.6

The second step was to get latitude and longitude data for each neighborhood. This was done using geopy.

Getting Localisation data from geopy

```
geolocator = Nominatim(user_agent="toronto_explorer")
Lat = []
Long = []
for neighb in health_df['Neighborhood']:
    address = neighb
    location = geolocator.geocode(address,country_codes='CA')
    if (location is not None):
        location = geolocator.geocode(address,country_codes='CA')
        Lat.append(location.latitude)
        Long.append(location.longitude)
    else:
        Lat.append(float('nan'))
        Long.append(float('nan'))
```

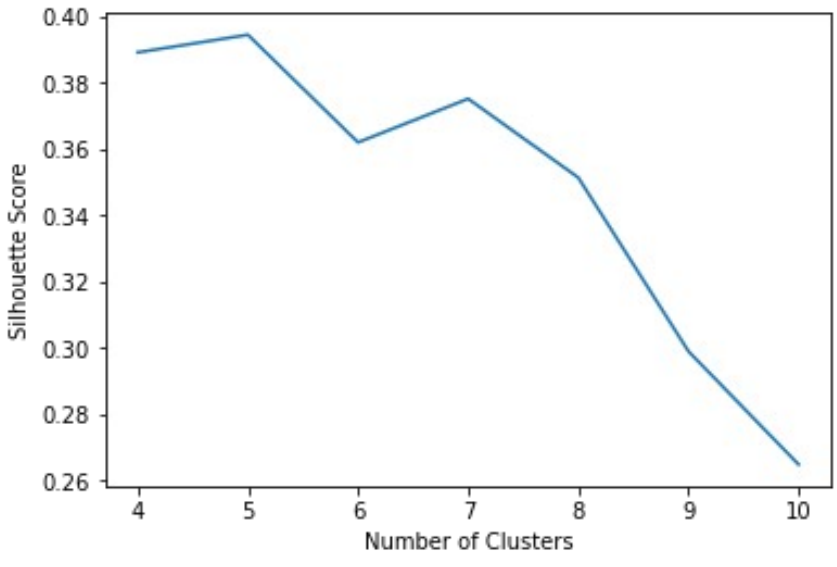
```
health_df['Latitude'] = Lat
health_df['Longitude'] = Long
```

After this, we used the foursquare API to get the number of gym, parks, pools, and fast food restaurants in each neighborhood.

	Neighborhood	Walk score	Diabetes	Latitude	Longitude	Gyms	Parks	FastFood	Pools
0	Agincourt North	66.0	9.5	43.808038	-79.266439	0	0	4	0
2	Alderwood	70.0	8.5	43.601717	-79.545232	1	0	1	1
3	Annex	94.0	5.5	43.670338	-79.407117	3	3	4	1
4	Banbury-Don Mills	67.0	6.5	43.751672	-79.370169	0	1	0	0
5	Bathurst Manor	61.0	8.5	43.665519	-79.411937	7	2	5	1
6	Bay Street Corridor	99.0	5.1	43.672798	-79.390734	21	6	7	8
7	Bayview Village	71.0	6.0	43.769197	-79.376662	0	0	2	0
8	Bayview Woods-Steeles	57.0	7.1	43.798127	-79.382973	0	0	0	0
11	Bendale	64.0	11.5	43.753520	-79.255336	0	2	1	0
13	Black Creek	62.0	12.7	45.622607	-77.156242	0	0	0	0

Now that we have our dataset, we cluster it using a k-means algorithm, using silhouette score to find the optimal number of cluster.

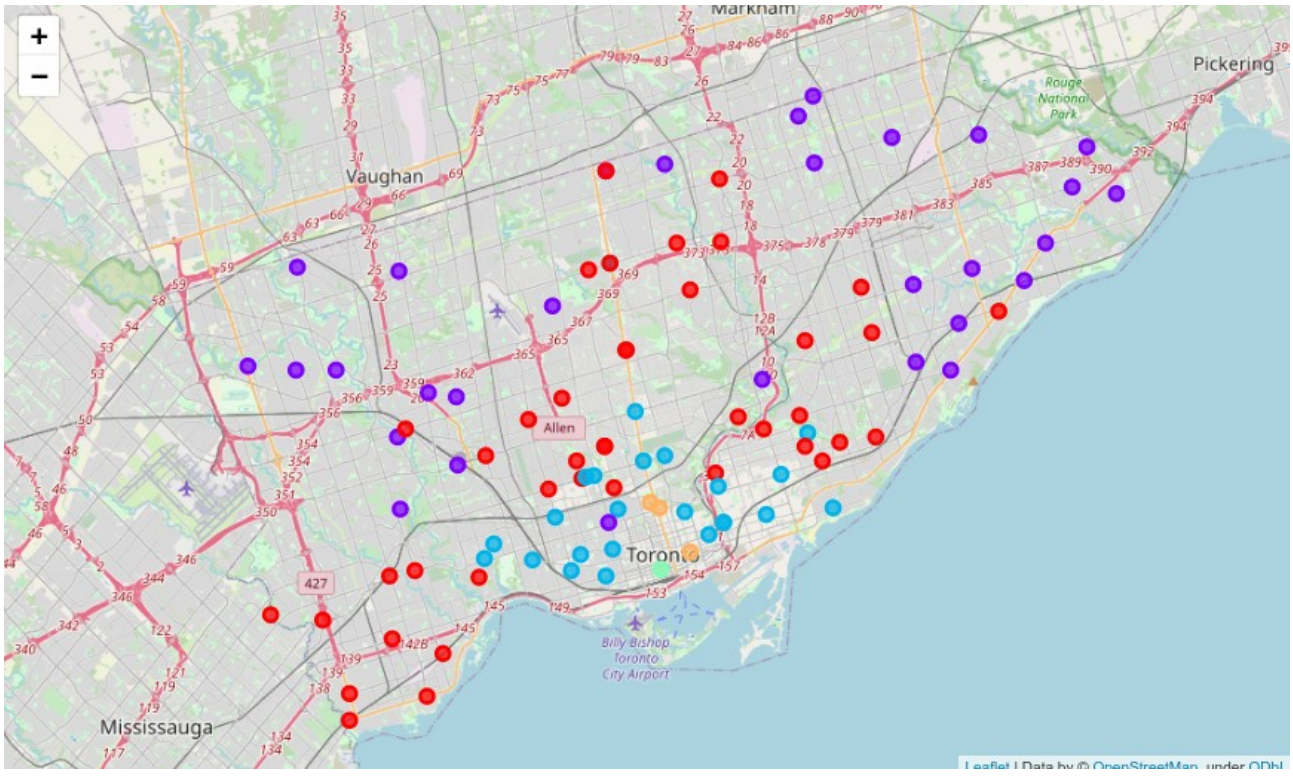
From the silhouette score, we see that the optimal number of clusters is 5.



4 – Results

A label was assigned to each neighborhood. The following map shows the clusters found.

	Cluster Labels	Neighborhood	Walk score	Diabetes	Latitude	Longitude	Gyms	Parks	FastFood	Pools
0	1	Agincourt North	66.0	9.5	43.808038	-79.266439	0	0	4	0
2	0	Alderwood	70.0	8.5	43.601717	-79.545232	1	0	1	1
3	2	Annex	94.0	5.5	43.670338	-79.407117	3	3	4	1
4	0	Banbury-Don Mills	67.0	6.5	43.751672	-79.370169	0	1	0	0
5	1	Bathurst Manor	61.0	8.5	43.665519	-79.411937	7	2	5	1
6	4	Bay Street Corridor	99.0	5.1	43.672798	-79.390734	21	6	7	8
7	0	Bayview Village	71.0	6.0	43.769197	-79.376662	0	0	2	0
8	1	Bayview Woods-Steeles	57.0	7.1	43.798127	-79.382973	0	0	0	0
11	1	Bendale	64.0	11.5	43.753520	-79.255336	0	2	1	0
13	1	Black Creek	62.0	12.7	45.622607	-77.156242	0	0	0	0



In what follows we describe each of the five clusters found. In each cluster we compute the mean number of gyms, parks, pools and fast foods, as well as the mean WalkScore and the mean diabetes prevalence.

Entrée [576]:
```
#Examining Cluster 1
C = health_df.loc[health_df['Cluster Labels'] == 0,health_df.columns[[1,2,3,6,7,8,9]]]
print('Cluster 1 ')
print(C.mean(0,numeric_only=True))
```

```
Cluster 1
Walk score    72.975610
Diabetes       8.287805
Gyms           0.853659
Parks          1.195122
FastFood       1.853659
Pools          0.195122
dtype: float64
```

Entrée [577]:
```
#Examining Cluster 2
C = health_df.loc[health_df['Cluster Labels'] ==1,health_df.columns[[1,2,3,6,7,8,9]]]
print('Cluster 2 ')
print(C.mean(0,numeric_only=True))
```

```
Cluster 2
Walk score    59.617647
Diabetes      18.617647
Gyms           0.529412
Parks          0.647059
FastFood       1.529412
Pools          0.088235
dtype: float64
```

Entrée [578]:
```
#Examining Cluster 3
C = health_df.loc[health_df['Cluster Labels'] ==2,health_df.columns[[1,2,3,6,7,8,9]]]
print('Cluster 3 ')
print(C.mean(0,numeric_only=True))
```

```
Cluster 3
Walk score    88.266667
Diabetes       7.243333
Gyms           2.633333
Parks          1.888000
FastFood       2.888000
Pools          0.688000
dtype: float64
```

Entrée [579]:
```
#Examining Cluster 4
C = health_df.loc[health_df['Cluster Labels'] ==3,health_df.columns[[1,2,3,6,7,8,9]]]
print('Cluster 4 ')
print(C.mean(0,numeric_only=True))
```

```
Cluster 4
Walk score    68.0
Diabetes       5.7
Gyms          22.0
Parks          8.0
FastFood      44.0
Pools         13.0
dtype: float64
```

Entrée [588]:
```
#Examining Cluster 5
C = health_df.loc[health_df['Cluster Labels'] ==4,health_df.columns[[1,2,3,6,7,8,9]]]
print('Cluster 5 ')
print(C.mean(0,numeric_only=True))
```

```
Cluster 5
Walk score    97.333333
Diabetes       6.466667
Gyms          28.000000
Parks          5.666667
FastFood      11.000000
Pools          6.666667
dtype: float64
```

4 – Discussion

We have found Toronto's unhealthiest neighborhoods ! They are the neighborhoods contained in cluster 2 (purple in the map), which, quite sadly, is almost half the city. There are very few gyms to compete with in these neighborhoods (roughly one gym for two neighborhoods) , and the diabetes prevalence is very high, with an abysmal walk score. It seems we have found the perfect locations to open a low-cost gym. This would be a win-win, because the stakeholders are bound to take profit by investing in a untapped market, and this will be the occasion to propose physical activity to people that are in dire need of it.

5 – Conclusion :

With this, we have concluded that cluster 2 is the ideal locations to open low-cost gyms, in order to provide a service that is not accessible in these neighborhoods, and to promote health and fitness in neighborhoods that are largely unhealthy compared to the rest of the city.