The Battle of Neighbourhoods in Toronto – Capstone Project

Find the best place to open a gym in downtown Toronto

Table of contents:

1. Introduction

- Goal: to find out the best neighbourhoods in downtown Toronto to open a profitable gym venue.

- Aimed to: entrepreneurs or business owners wanting to open their own gym business or grow their business in a profitable location where little or no competition exists today.

2. Data

The following data sources will be needed for this analysis:

1. Wikipedia page with the Neighbourhoods and postal codes of Toronto city (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)

2. The geographical coordinates of each postal code through reading the CSV file (https://cocl.us/Geospatial_data) and transforming it into a Pandas data frame.

3. Explore and get the venues data using Foursquare API.

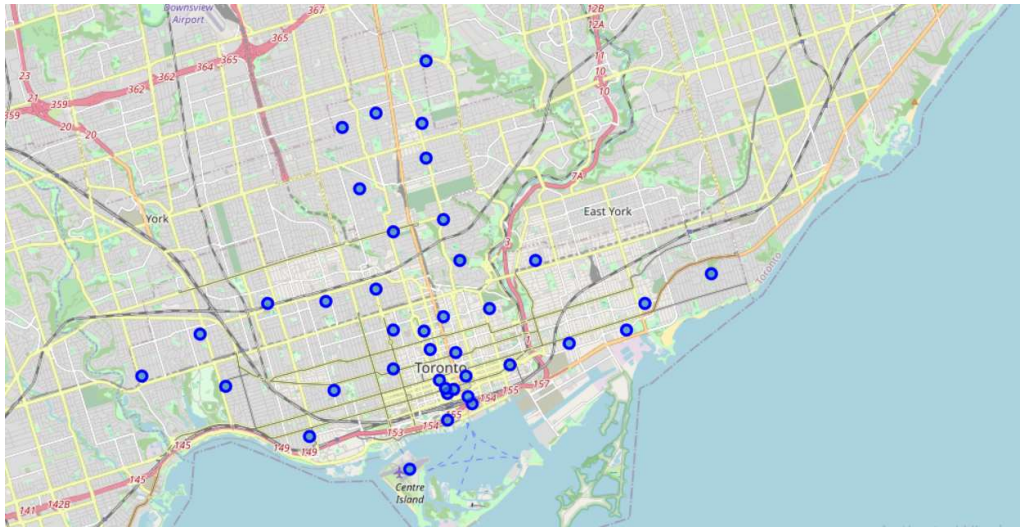
3. Methodology

3.1 Exploratory Data Analysis

First, we transform the data from the Wikipedia source, that is rename features, drop cells (those with Boroughs equal to ‘not assigned’) and group the data.

Then get each neighbourhood’s coordinates using the Geocoder package.

To explore the data, we use the Folium library that can create interactive maps of the area of our interest: downtown Toronto



Next it is the moment to use the Foursquare API to extract the venues of each neighbourhood in downtown Toronto.

```
# Create the GET request url
radius = 500

url = 'https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
    CLIENT_ID,
    CLIENT_SECRET,
    VERSION,
    neighborhood_latitude,
    neighborhood_longitude,
    radius,
    LIMIT)

url
```

After some coding we get to know the number of unique categories of venues in downtown Toronto

```
print(toronto_central_venues.shape)
print('There are {} uniques categories.'.format(len(toronto_central_venues['Venue Category'].unique())))
toronto_central_venues
```

```
(1624, 7)
There are 235 uniques categories.
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	The Beaches	43.676357	-79.293031	Glen Manor Ravine	43.676821	-79.293942	Trail
1	The Beaches	43.676357	-79.293031	The Big Carrot Natural Food Market	43.678879	-79.297734	Health Food Store
2	The Beaches	43.676357	-79.293031	Grover Pub and Grub	43.679181	-79.297215	Pub
3	The Beaches	43.676357	-79.293031	Upper Beaches	43.680563	-79.292869	Neighborhood
4	The Danforth West, Riverdale	43.679557	-79.352188	Pantheon	43.677621	-79.351434	Greek Restaurant
...	...	...	...	...	...	...	...
1619	Business reply mail Processing Centre, South C...	43.662744	-79.321558	TTC Russell Division	43.664908	-79.322560	Light Rail Station
1620	Business reply mail Processing Centre, South C...	43.662744	-79.321558	Jonathan Ashbridge Park	43.664702	-79.319898	Park
1621	Business reply mail Processing Centre, South C...	43.662744	-79.321558	Olliffe On Queen	43.664503	-79.324768	Butcher
1622	Business reply mail Processing Centre, South C...	43.662744	-79.321558	ONE Academy	43.662253	-79.326911	Gym / Fitness Center
1623	Business reply mail Processing Centre, South C...	43.662744	-79.321558	Revolution Recording	43.662561	-79.326940	Recording Studio

3.2 Clustering

To analyze which neighbourhood of downtown Toronto is more suitable to open a new gym, we use the K-Means clustering model: a type of unsupervised learning which is used when you have unlabeled data.

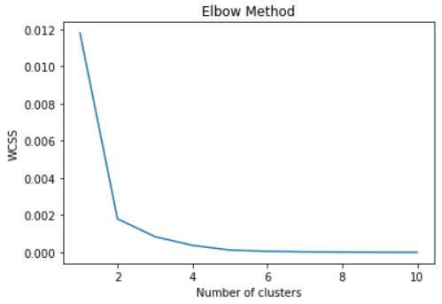
As we want to know which would be the optimal number of clusters, the best K, we will use the Elbow Method. It is an analytical approach that consists of training multiple models using different number of clusters and storing the value of inertia_property (wcss) every time.

The best K is 5, so we run the K-Means clustering with this number.

```
wcss = []

for i in range(1, 11):
    kmeans = KMeans(n_clusters=i, init='k-means++', max_iter=100, n_init=10, random_state=0)
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)

plt.plot(range(1, 11), wcss)
plt.title('Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```

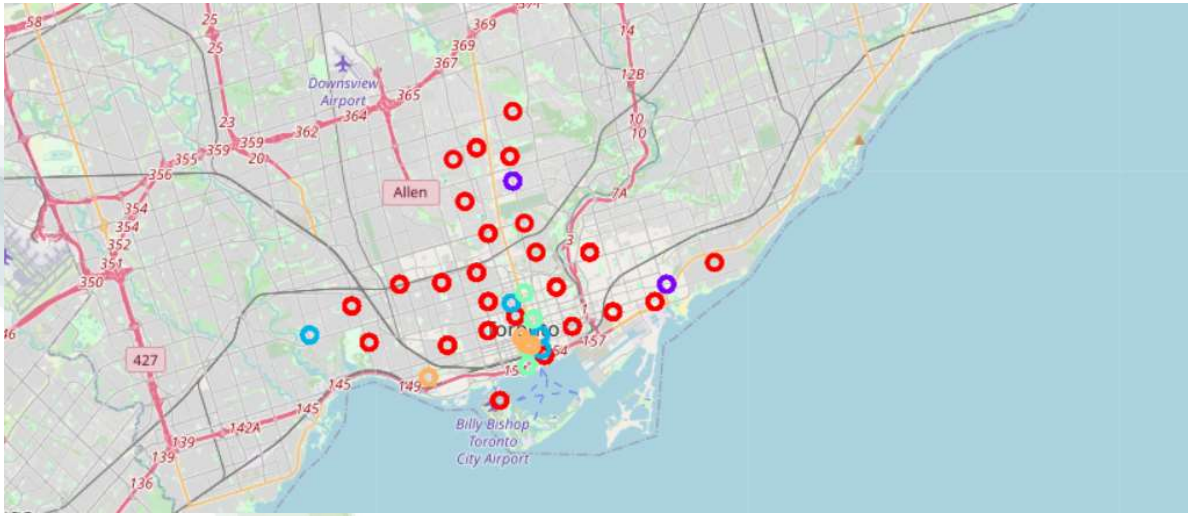


We can conclude that the optimum K value is 5, so we will have 5 clusters.

Finally we merge both data frames, the one with the venues of downtown Toronto and the othe with the clusters

	Neighborhood	Gym	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Berczy Park	0.00	0	43.644771	-79.373306	The Keg Steakhouse + Bar - Esplanade	43.646712	-79.374768	Restaurant
19	Little Portugal, Trinity	0.00	0	43.647927	-79.419750	BYOB Cocktail Emporium	43.644447	-79.417757	Miscellaneous Shop
19	Little Portugal, Trinity	0.00	0	43.647927	-79.419750	Le Dolci	43.650377	-79.415959	Cupcake Shop
19	Little Portugal, Trinity	0.00	0	43.647927	-79.419750	Pilot Coffee Roasters	43.646610	-79.419606	Coffee Shop
19	Little Portugal, Trinity	0.00	0	43.647927	-79.419750	Trinity Bellwoods Park	43.647072	-79.413756	Park
...	...	...	...	...	...	...	...	...	...

And we plot the Folium map with the different clusters

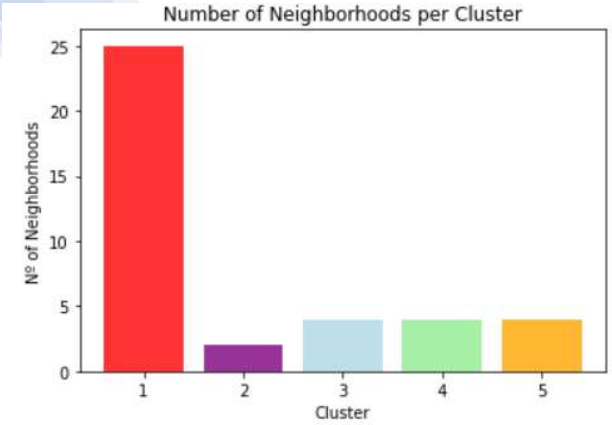


4. Results and discussion

We have a total of five clusters and we get to know how many neighbourhoods there are in each of them.

Then we make an analysis of each cluster and the average number of gym venues on them.:

- Cluster 1 (red) is the bigger one in terms of number of neighbourhoods and there are no gym venues in this cluster.
- Cluster 2 (purple) is the smallest cluster with only two neighbourhoods and has the highest average of gyms.
- Cluster 3 (lightblue) encompasses 4 neighbourhoods and has a total of 6 gyms.
- Cluster 4 (lightgreen) has 4 neighbourhoods as well but a lower average of gym venues than cluster 3.
- Cluster 5 (Orange) comprises also 4 neighbourhoods and has the second higher average number of gyms after cluster 2.



4. Conclusion

Most of the gym venues in downtown Toronto city are in cluster2 represented by the purple colour. The Neighbourhoods located in Central Toronto that have the highest average of gyms are Davisville, India Bazaar and The Beaches West.

Even though cluster1 has the larger number of Neighbourhoods (25), there is no gym within it, so this is a great opportunity to open a new one, with no competition. So Neighbourhoods like The Beaches, Kensington Market, Chinatown, Grange Park, Regent Park, Harbourfront would be the ideal location for the opening of a gym.

We must also point out that there is at least one drawback in this analysis and it is the fact of only using the data that come from Foursqare API, without looking for richer information like population of different Neighbourhoods to have one more element to make the decision.