Capstone Project -The Battle of Neighborhoods

Mohamed Mostafa Fathy Ahmed Awad

February 20, 2021

1. Business problem

Finding the suitable location is a main step to start any business. In this project, I will try to explore the data of Toronto city using the data of the districts and the Foursquare's 'Places API' to help finding the best location to build a shopping mall.

2. Data acquisition and cleaning

2.1 Data sources

In order to answer the above question, data on Toronto City neighborhoods, boroughs to include boundaries, latitude, longitude, restaurants, and restaurant ratings and tips are required. Below is the list of datasets used in this project:

- Toronto City data containing the neighborhoods and boroughs, latitudes, and longitudes will be obtained from the data source:
  https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

- Geospatial coordinates for each postal code to get the related longitude and latitude.

- Foursquare API to explore the venues related to the districts of Toronto city.

2.2 Data preparation and cleansing

During this step, several approaches and tools have been used to prepare the datasets and make it useful for the analysis and applying the machine learning algorithms.

First, I used web scrapping API to download the postal code data from Wikipedia. This dataset contains the postal code, Borough and Neighborhood only.

Second, this requires using additional geospatial coordinates dataset to get the longitude and latitude for each postal code. The format of this dataset was in CSV file so, I load it into data frame.

Finally, I combined the postal code data with the related geospatial coordinates.

2.3 Foursquare API

I used Foursquare API to explore the venues related to the districts of Toronto city as shown in Figure 2.1.

	District	Population	Code	Latitude	Longitude	VenueName	VenueId	VenueLatitude	VenueLongitude	VenueCategory
0	North York	Parkwoods	M3A	43.753259	-79.329656	Allwyn's Bakery	4b8991cbf964a520814232e3	43.759840	-79.324719	Caribbean Restaurant
1	North York	Parkwoods	M3A	43.753259	-79.329656	Donalda Golf & Country Club	4bd4846a6798ef3bd0c5618d	43.752816	-79.342741	Golf Course
2	North York	Parkwoods	M3A	43.753259	-79.329656	Galleria Supermarket	4ccec87654f0b1f7f32824ca	43.753520	-79.349518	Supermarket
3	North York	Parkwoods	M3A	43.753259	-79.329656	Tim Hortons	57e286f2498e43d84d92d34a	43.760668	-79.326368	Café

Figure 2.1 the Foursquare API

3. Exploratory data analysis

3.1. Display districts on map

After combining the district and the related coordinates, it became possible to display the districts on the map as shown in Figure 3.1.

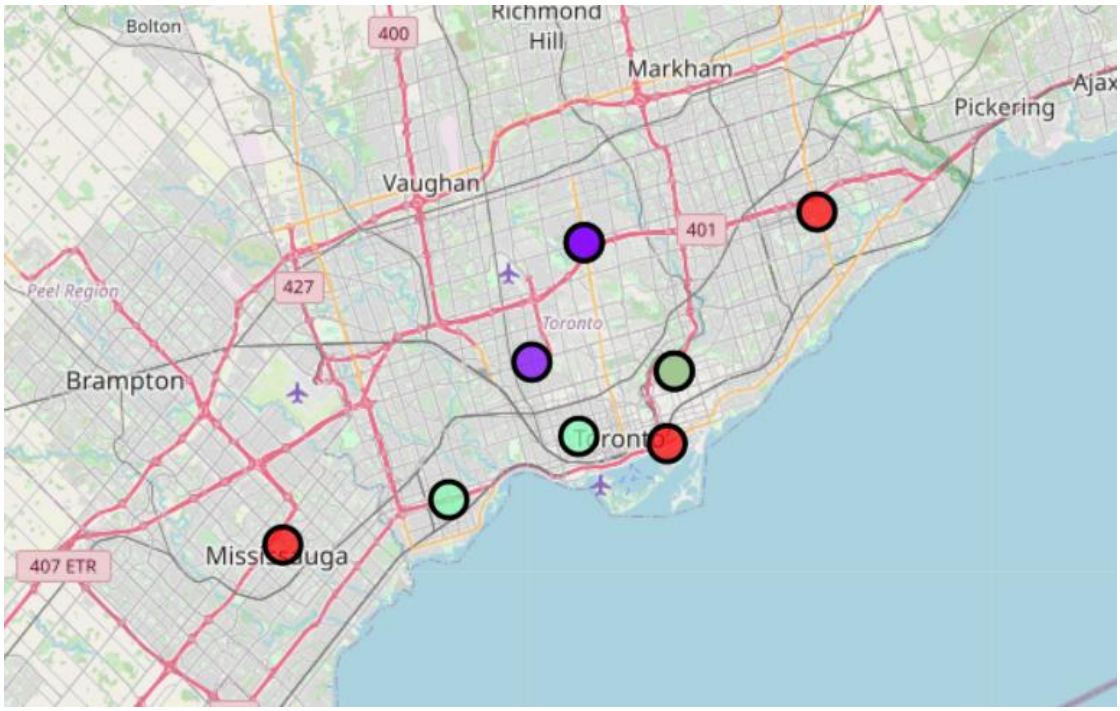

Figure 3.1 the districts on map

3.2. Calculation of number of malls in districts

Using the data retrieved from Foursquare API, I was able to know the category of each venue.

Since the target is to know how many shopping malls exist in each district, I filtered the data by selecting the category with the name “Shopping Mall”.

Now I can group the venues based on the districts and get the total number of malls in each district as shown in the Figure 3.2.

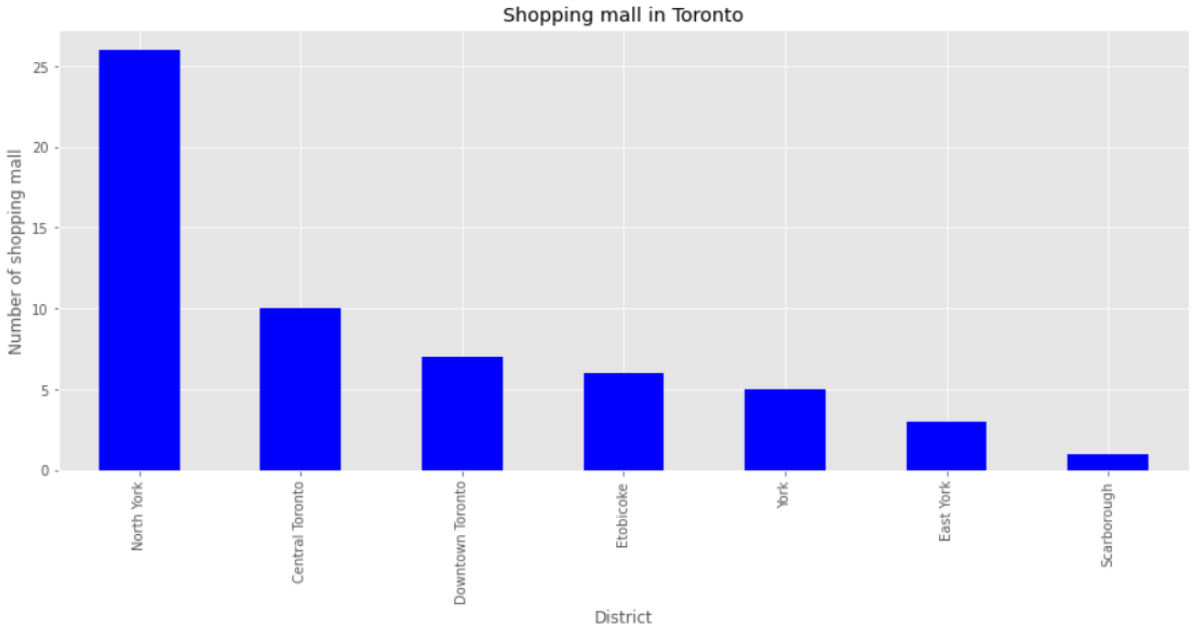

Figure 3.2 the number of shopping malls per district

4. Machine learning model

Using the K-Mean algorithm, we will cluster the district into 3 clusters based on their frequency of occurrence for “Shopping Mall”.

Before using K-Mean, we must transform data using one-hot encoder. The results will allow us to identify which districts have a higher concentration of shopping malls while which districts have a fewer number of shopping malls by using k-means clustering. Based on the occurrence of shopping malls in a different district, it will help us to plan which district we should acquire land for developing new shopping mall.

5. Results

From the results, we found that the shopping malls were clustered using number of shopping malls in each district, so it can be classified to high, medium and low number of shopping malls in each area.

However, there are other factors which can be used to cluster such as number of tourists, income of people around that area, number of people traveled around these areas. This information can make clustering more accurate and will generate more information for decision making.

Figure 5.1 shows the results of the clusters:

```
df_clustered.loc[df_clustered['Cluster'] == 0].head()
```

	District	Shopping Mall	Cluster	Latitude	Longitude
2	East Toronto	0.000000	0	43.691200	-79.341667
5	Mississauga	0.000000	0	43.595310	-79.640579
7	Scarborough	0.000592	0	43.777702	-79.233238
8	West Toronto	0.000000	0	43.651070	-79.347015

```
df_clustered.loc[df_clustered['Cluster'] == 1].head()
```

	District	Shopping Mall	Cluster	Latitude	Longitude
0	Central Toronto	0.011111	1	43.761539	-79.411079
6	North York	0.010888	1	43.761539	-79.411079
9	York	0.010000	1	43.695683	-79.450279

```
df_clustered.loc[df_clustered['Cluster'] == 2].head()
```

	District	Shopping Mall	Cluster	Latitude	Longitude
1	Downtown Toronto	0.003684	2	43.654753	-79.414187
3	East York	0.006000	2	43.691200	-79.341667
4	Etobicoke	0.005076	2	43.620495	-79.513199

Figure 5.1 The cluster results

6. Conclusion

The objective of this project is to support decision making of building a new shopping mall in Toronto. Using the following process:

- Identifying the business problem.
- Specifying the required data.
- Extract and prepare the data.
- Performing machine learning by clustering the data into 3 clusters based on their similarities
- Finally providing recommendations.