Capstone Project - The Battle of Neighborhoods - Locate ideal neighborhood

Business Problem

In this modern age, Canada is a very popular immigration destination. Lots of people are migrating to various states of Canada and needed lots of information on good housing prices and reputed schools for their children.

This scope of this Project is to help Indians migrating to Toronto to identify a best neighborhood as a comparative analysis between neighborhoods.

The features used in analysis will be to identify neighborhoods where following facilities are mostly available nearby like school, shops, transport facilities, Indian restaurants, medical facilities etc

Thus our business problem to solve is, which neighborhood is ideal to move in for a Indian immigrant.

Approach

This project will utilize publicly available data from Wikipedia and Foursquare.

Specifically, all Toronto neighborhood details along with their postal codes are available here: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

The focus of this project will be to acquire data by using web scraping method of python and then clean the data, populate the data then use foursquare API to collect the data of the all the neighborhood places then use k-means clustering method to find the best fit.

Data and EDA

Get the data from wiki portal - https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M and prepare it.

Get the location coordinates from - https://cocl.us/Geospatial_data

Merge both the data to arrive at our source of data.

Following data sources will be needed to extract/generate the required information:

- Group the neighborhoods based on postal code and use K-means clustering to analyze and visualize the data.
- Location details like latitude and longitude details are fetched from remote location, alternatively they can also be retrieved from Google Maps API.
- Fetch the facilities like Restaurants, Bus Stations, Schools, Coffee shops, Grocery Shops etc in every neighborhood using Foursquare API
- Identify the neighborhood which matches our requirement - having most of preferred facilities, and recommend it to the customer.

Prepare a single source of data (.csv file) and use it in the project.

When looking at the data, I found that there were some columns which had 'Not Assigned' values. Before proceeding, I need to clean them up.

I set all 'Not Assigned' valued neighborhoods with same value as Boroughs field, to standardize the data.

Use the IsNan() to find and clean up empty data.

Sorted and indexed the data, for using with FourSquare APIs.

Have encoding for the Venue Categories to process the data.

The neighborhood details data was merged with the geospatial data and a consolidated file was created for further use.

Folio Map was used to visualize the data.

	Postal Code	Borough	Neighbourhood
count	180	180	180
unique	180	11	100
top	M4P	Not assigned	Not assigned
freq	1	77	77

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

	key_0	Postal Code_x	Borough	Neighbourhood	Postal Code_y	Latitude	Longitude
0	M1B	M1B	Scarborough	Malvern, Rouge	M1B	43.806686	-79.194353
1	M1C	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	M1C	43.784535	-79.160497
2	M1E	M1E	Scarborough	Guildwood, Morningside, West Hill	M1E	43.763573	-79.188711
3	M1G	M1G	Scarborough	Woburn	M1G	43.770992	-79.216917
4	M1H	M1H	Scarborough	Cedarbrae	M1H	43.773136	-79.239476

Map of place of interest – Toronto

Folium mapping feature was used to plot the maps.

Getting Required Info - Solution

I analyze the data, pick identify the features and work on those column data, like grouping and sorting them, and joining the similar fields.

Then I use the Folium's mapping methods to plot the data on to a map for visualization.

K-means clustering is employed to arrive at the required result of identifying the neighborhood which matches to the customers requirement – that is a neighborhood which has the most number of facilities in a set of expected/required facilities.

These facilities (venue details) are fetched for each postal code (location) using Four Square APIs.

So first we create a new FourSquare account and use it to register and consume the api.

A typical call to the API will look like

```
'https://api.foursquare.com/v2/venues/explore?&client_id={}&cli..'
```

Sample result from the API

```
toronto_venues.head()
```

```
(1338, 7)
```

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Malvern, Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
2	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	SEBS Engineering Inc. (Sustainable Energy and ...	43.782371	-79.156820	Construction & Landscaping
3	Guildwood, Morningside, West Hill	43.763573	-79.188711	RBC Royal Bank	43.766790	-79.191151	Bank
4	Guildwood, Morningside, West Hill	43.763573	-79.188711	G & G Electronics	43.765309	-79.191537	Electronics Store

For our business case – we will identify neighborhood which has below facilties (categories)

Bank, Medical Center, Gas Station, Bakery, Department Store, Train Station, Bus Line, Metro Station, Bus Station, Bus Line, Indian Restaurant, Shopping Mall, Pharmacy, Convenience Store, Supermarket, Fruit & Vegetable Store, Farmers Market.

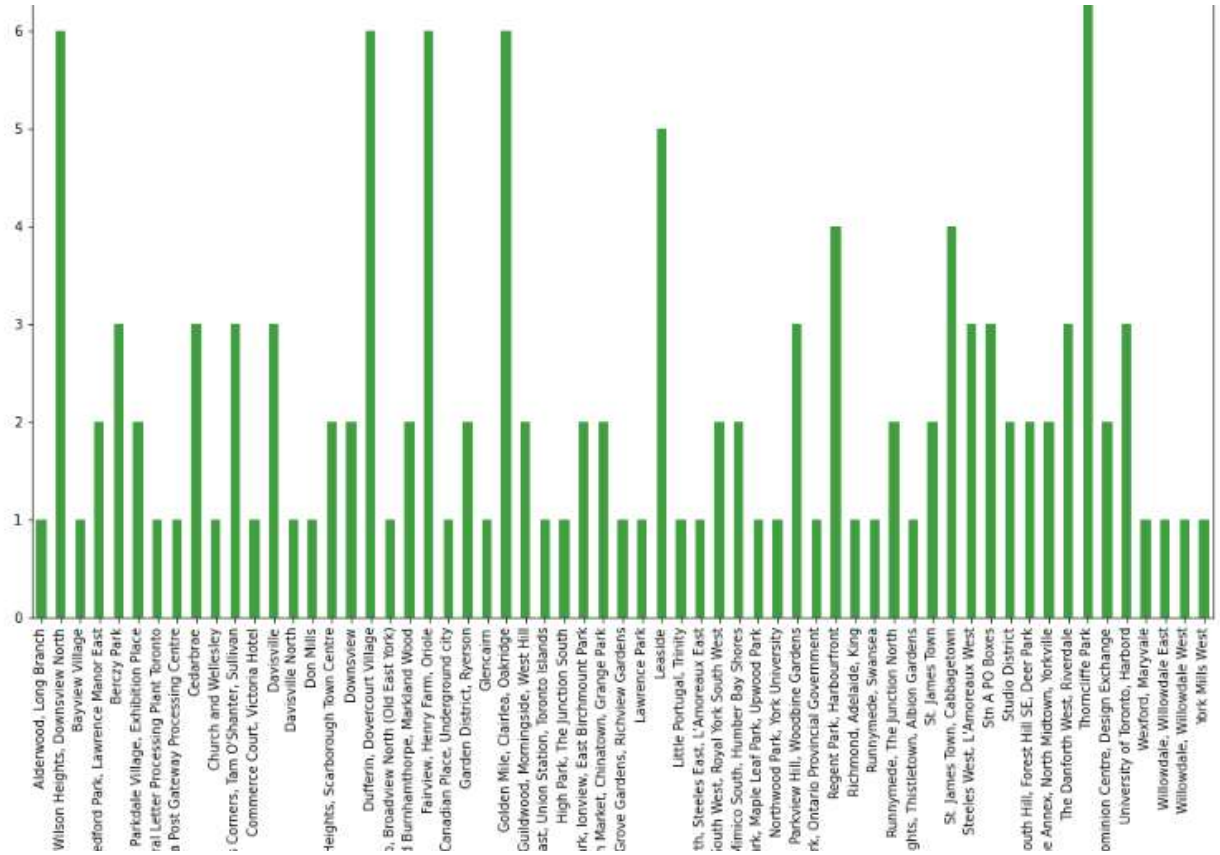
Once we are able to identify a neighborhood area which has the most of these venues, we will mark it as our preferred location for our customer to relocate.

Note: We can enhance this further, by checking each type of venues not only the count. But for our convenience we will currently only bother on the count of these facilities in a neighborhood.

Sample output

Neighbourhood	Venue Category
Alderwood, Long Branch	1
Bathurst Manor, Wilson Heights, Downsview North	6
Bayview Village	1
Bedford Park, Lawrence Manor East	2
Berczy Park	3

We can visualize the data using a histogram as below;



With the result data we can identify which neighborhood had the maximum count of facilities (venue-categories) and recommend that neighborhood to the customer.

Thus we will be able to recommend an ideal location for new Indian immigrants to settle in Toronto.

Conclusion

Finally, by leveraging the Foursquare API and K-Means clustering, we are able to pick a neighborhood which can be recommended for Indian Immigrants to Toronto.