Identifying Franchise Locations in Toronto.

Parth Thakurdesai

November 11, 2019

Introduction:

1.1 Background

Coffee Chain market is a fiercely competitive industry and opening locations is key to growth and wither competition. Locating potential neighbourhoods for Coffee Shops is key part of the role of the management team. Identifying potential neighbourhoods to open shops is a tricky task which can be made easier by availability of data.

1.2 Problem

You are Director of Western Operations for a Canadian Coffee company. Your task is to identify neighbourhoods that have lower density of Coffee Shops. It would be ideal if the coffee shops in the area are rated at satisfactory or below. You are confident that you can beat the nearest competitor on quality but you have a competing price. One way to wage a quality war would be to identify neighbourhoods with lower density of coffee shops that are relatively close to downtown and are rated below satisfactory.

1.3 Interest

Most likely beneficiaries of this project would be the Coffee Chain that has employed you. Or third-party Coffee Chains interested in analysis of potential Neighbourhoods.

Data Acquisition and Cleaning:

2.1 Data sources

For this project, we will be scraping data from Wikipedia page found here:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
This will provide us with a data frame that consists of three columns: **Postal Code, Borough** and **Neighbourhoods.** The dataset covers all neighbourhoods in Toronto.

2.2 Data Cleaning

After generating a pandas data frame from the Wikipedia URL, we decide to covert all the null values in the Borough column to np.NaN. This made dropping the bull values in a column easier. We then used .dropna() function to drop all the np.NaN values from the data frame.

Then we decide to group the data based on the postal codes by using .groupby() command.

Next we needed to Aggregate the values of the Neighbourhood column based on the PostalCode values. We updated the original data frame and used .agg(',').join) command to aggregate the values of Neighbourhood and Borough column.
Now the data frame has columns that are grouped by Postal Codes.

After building the data frame of the postal code of each neighbourhood, we needed to convert the postal codes into latitude and longitude values. These will become the centroids of the communities we explore (assumption). We require the latitude and longitude values to work with foursquare API.

We imported pgeocoder package and created a Nominatim object with country code 'ca' for Canada. Then we posted a query to retrieve information by inputting the postal code.

The query returned the following values: postal_code, country_code, place_name, state_name, state_code, country_code, community_name, community_code, latitude, longitude. We then decide to extract latitude and longitude values. Using a for loop the program updated the dataframe to include latitude and longitude values along with postal codes.

2.3Feature selection (FourSquare API calls)

Now that we had a completed data frame consisting of the following:

- **Postal Code**
- **Borough**
- **Neighbourhood**
- **Latitude**
- **Longitude**

We will now create a for loop to make API calls with search endpoint with "coffee" as query search. We will limit the output to 10 within a 1 km radius area.
JSON response from the server is expected and is filtered and transformed into a pandas dataframe. The length of this data frame will be equal to the number of coffee shops in the area. Within each of these for loops we will embed another for loop that will make API calls with stats endpoint with the VENUE_ID of the place retrieved from the data frame. The venue stats provides us with a rating of the place. These ratings are averaged over each neighbourhood and averaged out.

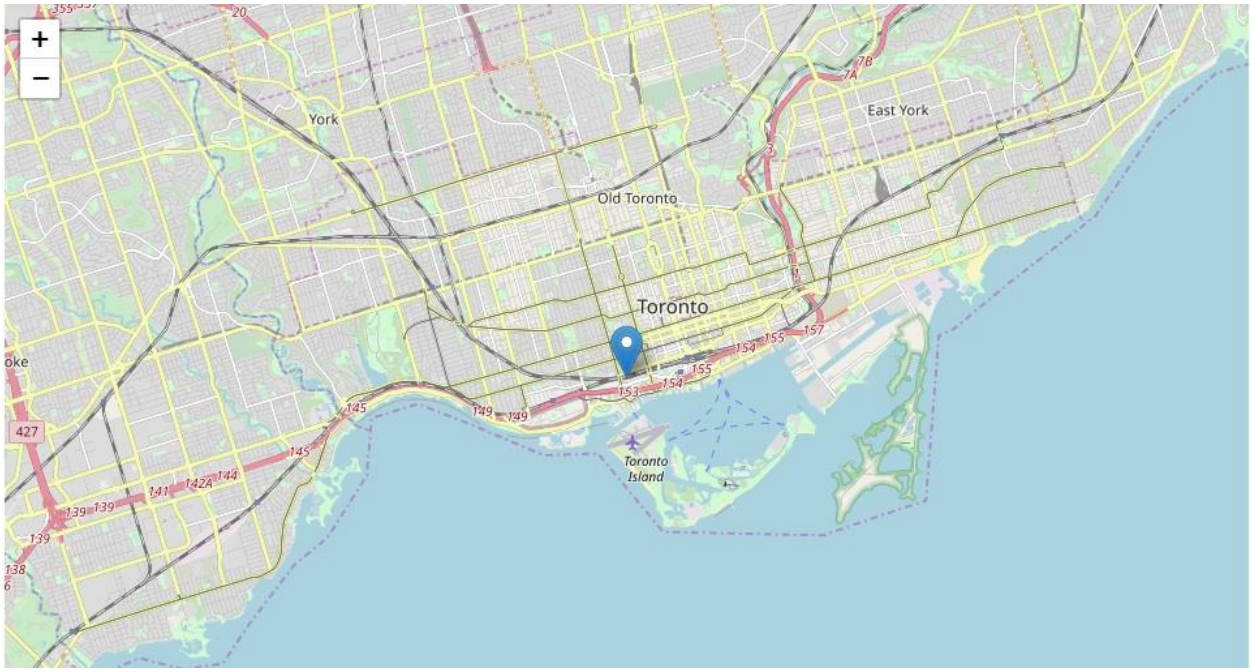
(Unfortunately lot of coffee shops don't have rating, this severely handicaps our scope)
Meaningful observations can be made by making premium API calls which require money.

The final data frame consists of the following values:

	Postcode	Borough	Neighbourhood	Latitude	Longitude	Distance (km)	Number of Coffee Shops	Average Rating
35	M4B	East York	Woodbine Gardens,Parkview Hill	43.7063	-79.3094	10.315336	3	Not Available
38	M4G	East York	Leaside	43.7124	-79.3644	8.485618	2	Not Available
39	M4H	East York	Thorncliffe Park	43.7059	-79.3464	8.443861	0	Not Available
44	M4N	Central Toronto	Lawrence Park	43.7301	-79.3935	9.977952	1	Not Available
48	M4T	Central Toronto	Moore Park,Summerhill East	43.6899	-79.3853	5.617722	2	Not Available
50	M4W	Downtown Toronto	Rosedale	43.6827	-79.373	5.163019	2	Not Available
59	M5J	Downtown Toronto	Harbourfront East,Toronto Islands,Union Station	43.623	-79.3936	1.990998	1	Not Available

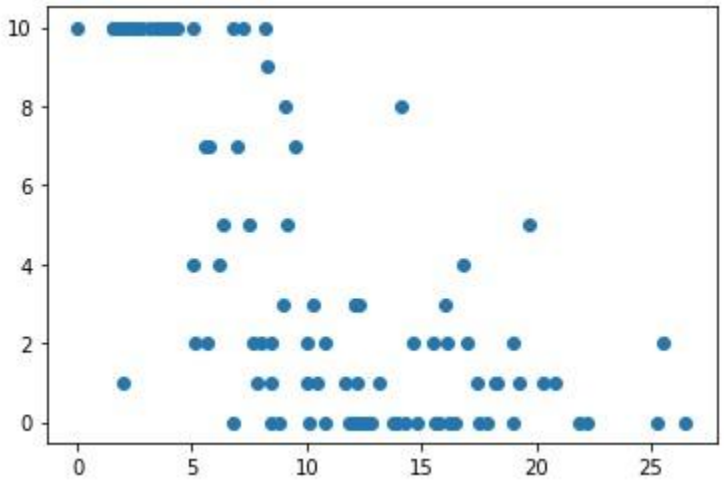
Exploratory Data Analysis:

3.1 City Center (CN Tower) on Tornoto Map



3.2 Neighbourhoods marked in red (Folium Maps)

3.3 Scatter Plot of Distance vs. number of coffee shops



4 Conclusion:

Based on API call limit and limited inferences we identified potential neighbourhoods that are relatively underserved, relatively close to downtown.

```
{'Bedford Park,Lawrence Manor East',
 'Caledonia-Fairbanks',
 'Glencairn',
 'Harbourfront East,Toronto Islands,Union Station',
 'Humber Bay Shores,Mimico South,New Toronto',
 "Humber Bay,King's Mill Park,Kingsway Park South East,Mimico
NE,Old Mill South,The Queensway East,Royal York South East,Sun
nylea",
 'Kingsway Park South West,Mimico NW,The Queensway West,Royal
York South West,South of Bloor',
 'Lawrence Heights,Lawrence Manor',
 'Lawrence Park',
 'Leaside',
 'Moore Park,Summerhill East',
 'Rosedale',
 'Roselawn',
 'The Junction North,Runnymede',
 'The Kingsway,Montgomery Road,Old Mill North',
 'Thorncliffe Park',
 'Woodbine Gardens,Parkview Hill'}
```

5 Recommendation:

Premium API calls can provide more insight and have more relevant data. Then machine learning can be applied.