The recommendation for Indian Restaurant in Toronto, Canada

Coursera Capstone Assignment: The battle of Neighborhoods

Author:

Siddharth Sharma

Introduction

When starting a new restaurants, first thing one needs to explore the prospective locations. The location should be chosen in such a way that it attracts the customers and provide less competition with similar restaurants. This can be achieved by segmenting the neighborhood on the basis of the similarities of the locations and clustering the locations on the basis of the competition with other Indian restaurants.

Problem Statement

Given the location data of the neighborhood of the Toronto, Canada, one would like to cluster the similar neighborhoods according to the frequency of Indian restaurants and on the basis of analysis we would like to recommend best neighborhoods to start a new Indian restaurant.

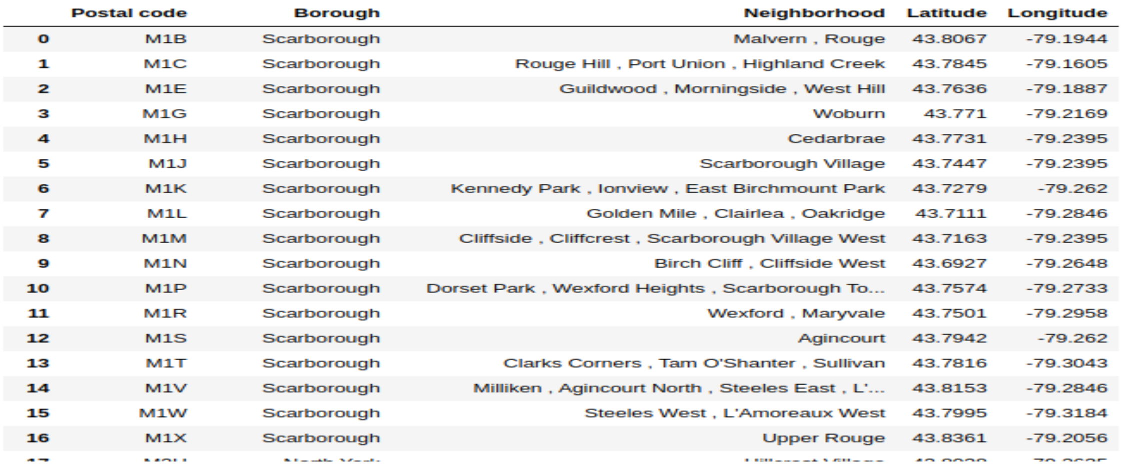
Data

To solve this problem, I will need below data:

•List of neighborhoods in Toronto, Canada.
"https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M"
•Latitude and Longitude of these neighborhoods.
•Venue data related to Indian restaurants. This will help us find the neighborhoods that are most suitable to open a Indian restaurant.

Extracting the data

•Scrapping of Toronto neighborhoods via Wikipedia
•Getting Latitude and Longitude data of these neighborhoods via Geocoder package
•Using Foursquare API to get venue data related to these neighborhoods.



	Postal code	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern , Rouge	43.8067	-79.1944
1	M1C	Scarborough	Rouge Hill , Port Union , Highland Creek	43.7845	-79.1605
2	M1E	Scarborough	Guildwood , Morningside , West Hill	43.7636	-79.1887
3	M1G	Scarborough	Woburn	43.771	-79.2169
4	M1H	Scarborough	Cedarbrae	43.7731	-79.2395
5	M1J	Scarborough	Scarborough Village	43.7447	-79.2395
6	M1K	Scarborough	Kennedy Park , Ionview , East Birchmount Park	43.7279	-79.262
7	M1L	Scarborough	Golden Mile , Clairlea , Oakridge	43.7111	-79.2846
8	M1M	Scarborough	Cliffside , Cliffcrest , Scarborough Village West	43.7163	-79.2395
9	M1N	Scarborough	Birch Cliff , Cliffside West	43.6927	-79.2648
10	M1P	Scarborough	Dorset Park , Wexford Heights , Scarborough To...	43.7574	-79.2733
11	M1R	Scarborough	Wexford , Maryvale	43.7501	-79.2958
12	M1S	Scarborough	Agincourt	43.7942	-79.262
13	M1T	Scarborough	Clarks Corners , Tam O'Shanter , Sullivan	43.7816	-79.3043
14	M1V	Scarborough	Milliken , Agincourt North , Steeles East , L'...	43.8153	-79.2846
15	M1W	Scarborough	Steeles West , L'Amoreaux West	43.7995	-79.3184
16	M1X	Scarborough	Upper Rouge	43.8361	-79.2056
17	M2H	North York	Hillcrest Village	43.8038	-79.3635

Figure 1: Final dataframe after adding postal codes and their respective coordinates.

Methodology

The project starts with scrapping the data from the Wikipedia page (“https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M”) as a tabular data and converting it into a pandas DataFrame. The pandas dataframe provides a wide range of inbuilt functions for cleaning the data. All the null values are taken care. The Borough that were not assigned were dropped and unassigned neighborhood are assigned with respective Borough.

However only the postal code are provided in this Wikipedia page, the latitude and longitude of the neighborhoods are also needed to search for venues using FourSquare API. One can use geocoder library to get coordinates of the neighborhoods, however, Geocoder can fail sometimes and requires a multiple instances. To overcome this redundant work, .csv file containing the coordinates of the neighborhood of Toronto is download and read into dataframe using pandas. Next step is to merge these dataframes into one dataframe containing Borough, Neighborhoods and their respective coordinates. The neighborhoods are plotted on the Toronto map using Folium Library and shown is figure 1.

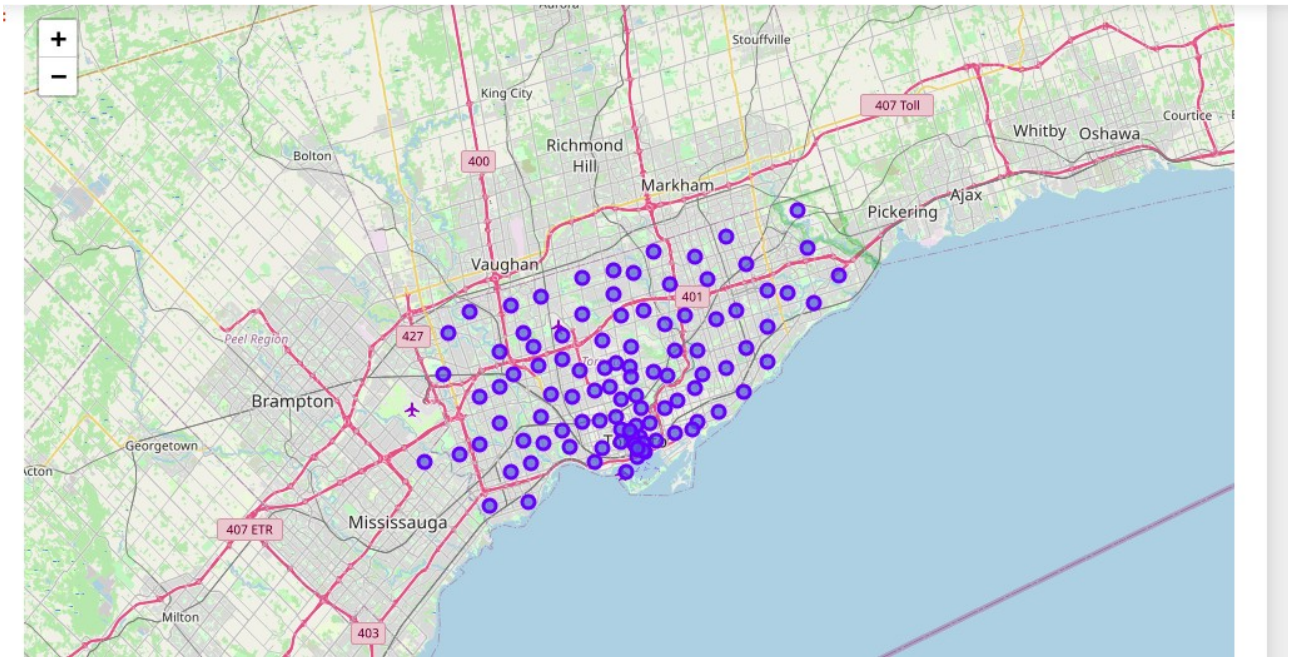
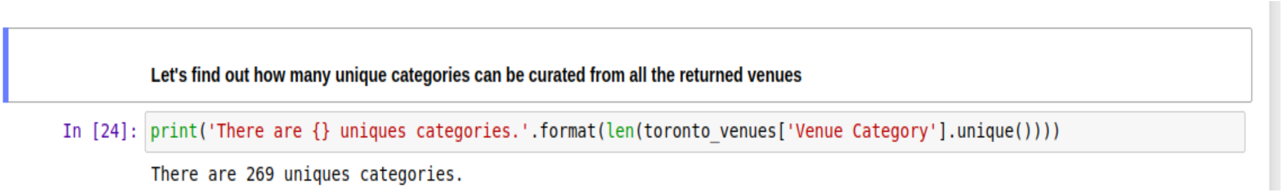


Figure 2: Toronto map with different neighborhoods as pointers.

Next, I use Foursquare API to pull the list of top 100 venues within 500 meters radius. I have created a Foursquare developer account in order to obtain account ID and API key to pull the data. From Foursquare, I am able to pull the names, categories, latitude and longitude of the venues. With this data, I can also check how many unique categories that I can get from these venues as shown in figure 2.



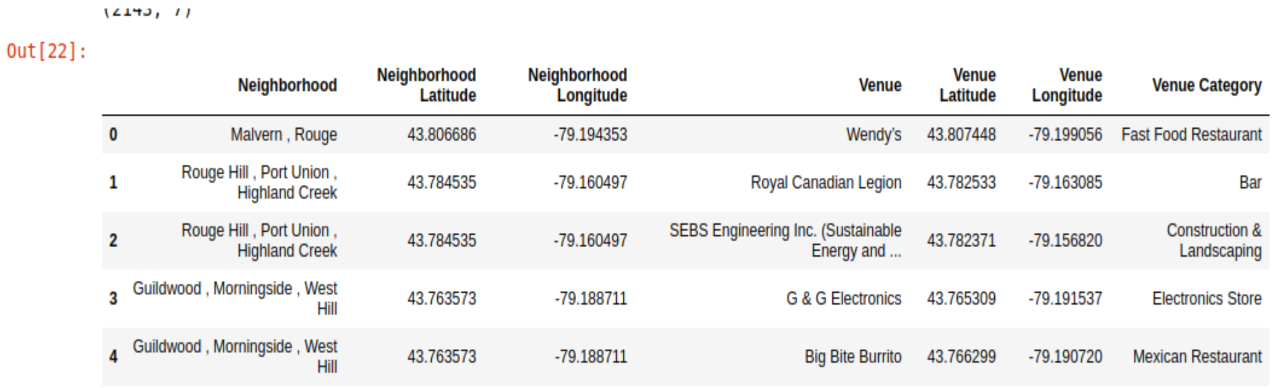
Let's find out how many unique categories can be curated from all the returned venues

```
In [24]: print('There are {} uniques categories.'.format(len(toronto_venues['Venue Category'].unique())))

         There are 269 uniques categories.
```

Figure 3: Unique venues in the neighborhoods

In the next step the neighborhoods are analyzed by grouping them and taking the mean of the frequency of occurrence of each venue category as shown in figure 3.

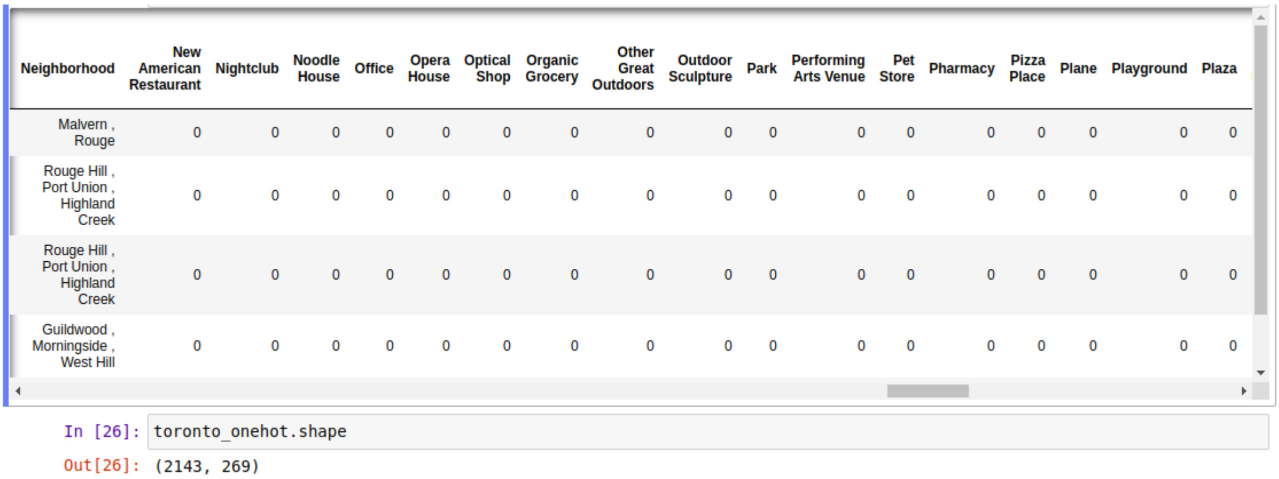


```
(2143, 7)
Out[22]:
```

	Neighborhood	Neighborhood Latitude	Neighborhood 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	G & G Electronics	43.765309	-79.191537	Electronics Store
4	Guildwood , Morningside , West Hill	43.763573	-79.188711	Big Bite Burrito	43.766299	-79.190720	Mexican Restaurant

Figure 4: Neighborhoods with nearby venue and their categories

In the next step the dataframe is updated to one-hot kind of assignments with 0,1 values assigning for absence/presence of a particular venue. The respective results are shown in the figure 4.



Neighborhood	New American Restaurant	Nightclub	Noodle House	Office	Opera House	Optical Shop	Organic Grocery	Other Great Outdoors	Outdoor Sculpture	Park	Performing Arts Venue	Pet Store	Pharmacy	Pizza Place	Plane	Playground	Plaza
Malvern , Rouge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rouge Hill , Port Union , Highland Creek	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rouge Hill , Port Union , Highland Creek	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Guildwood , Morningside , West Hill	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

```
In [26]: toronto_onehot.shape

Out[26]: (2143, 269)
```

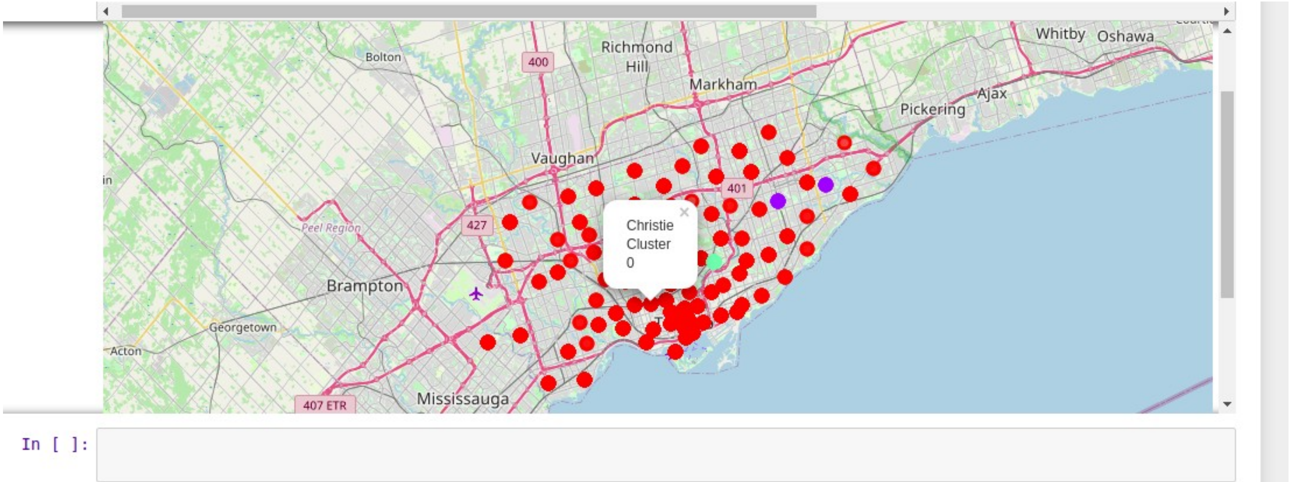
Figure 5: Neighborhoods with one-hot labelling as 0,1

In the next step we made a new dataframe with neighborhoods having indian restaurants as frequency associated with number of restaurants available in the vicinity.

Lastly, I performed the clustering method by using k-means clustering. K-means clustering algorithm identifies k (=3) number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and it is highly suited for this project as well. I have clustered the neighborhoods in Toronto into 3 clusters based on their frequency of occurrence for “Indian restaurants”. Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the restaurant.

Results

Using the k mean clustering algorithm with k=3, the neighborhoods are segmented and clusered around the centroids as shown in figure 5



```
In [ ]:
```

Figure 6: The clustered neighbourhoods with centroids shown. Red as cluster 1, blue as cluster 2, cyan as cluster 3

The three clusters are as follows :

● *cluster 1:* ● *cluster 3:* ● *cluster 2:*

The mean value of frequency of indian restaurants are determined for all three clusters are shown in table 1 and also the shown with the help of bar graph in figure 6.

TABLE 1: The mean of the frequency of indian restaurants in each clustered neighborhoods.

	Cluster 1	Cluster 2	Cluster 3
Mean	0.003	0.3	0.11

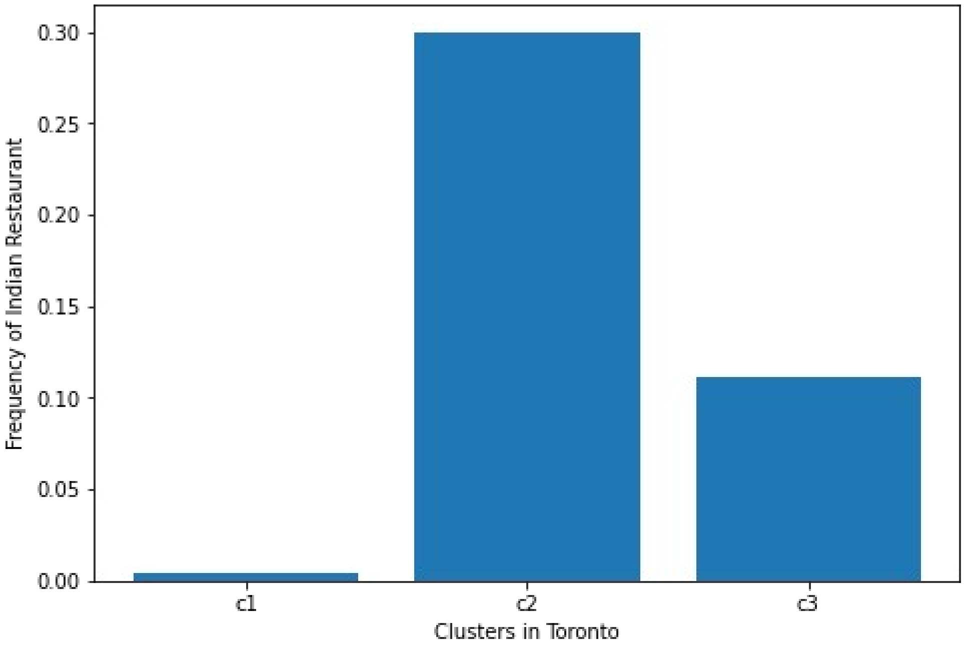


Figure 7: The bar graph showing the mean of the frequency of Indian restaurants in the different clusters.

Discussion

As we can see from the figure 6 and table 1, **cluster 1** has the least amount of indian restaurants, whereas **cluster 2** has the highest amount of indian restaurants. Through this data analysis, we can recommend to choose any neighborhood from the cluster 1 to start a new Indian restaurant, as there is very less competition and a demand for Indian restaurant.

Conclusion

In this project only one criteria was taken into consideration and that is the occurrence or presence of other Indian restaurants. One can analyse other features such as population density, gdp, availability of space and their prices, availability of parking and many more features but that is beyond the scope of this project. With the available data we can recommend to consider neighborhoods in cluster 1 for opening a new indian restaurant.