Optimal Location for Opening a Restaurant
IBM Data Science Capstone Project
Rohit Murakonda

Introduction

While opening a restaurant can be a very lucrative business, a lack of demand causes many restaurants to close within the first year of opening. There are many different factors that can account for a restaurant's success such as location, competition and quality of the food. This is an important question that every business owner must face when choosing whether to open a restaurant or not, as well as location of the business.

To demonstrate the process of picking a location for a client opening a business, the project will focus on answering the following question: "If the client wanted to open an Indian Restaurant in Toronto, what areas are the best options to open the restaurant?" For an Indian Restaurant, the location and competition are both determined by where the restaurant is opened. If there are too many Indian Restaurants in the local vicinity, the profitability of the restaurant will be severely decreased. Additionally, starting a restaurant in a location with higher income would increase the profitability of the business over starting in a poorer area.

Data

To answer the business problem, the following factors have to be extracted from various data sources:

- Population & Ethnic Distribution of Each Neighborhood (Toronto Census)
- Income Distribution of Each Neighborhood (Toronto Census)
- Number of Restaurants in Each Neighborhood (Foursquare API)
- Number of Indian Restaurants in Each Neighborhood (Foursquare API)

The Toronto Census data was extracted from https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/#8c732154-5012-9afe-d0cd-ba3ffc813d5a.

Methodology

The first step of the project was to combine the Toronto dataset, containing the postal code, borough, neighborhood name, latitude and longitude for each postal code in Toronto, and the census dataset. The first four columns of combined dataframe can be seen in Figure 7 of the Appendix.

Using the income distribution for each neighborhood, the spending power of each area was calculated using the median of each category weighted by the number of people in that income category. Thus, the spending power represents the overall capital of each area (i.e. total income of the inhabitants). Since the spending power for each area is considerably large and the relative strength is difficult to visualize, the spending power for each area was standardized.

The next step was to visualize the location of the various postal codes within Toronto to obtain a general understanding the location (Figure 1). As seen from the map, the postal codes are densely clustered near downtown Toronto and spread out as the distance from downtown increases. This is important because while some postal codes might not have many restaurants, if the area is located near downtown, adjacent regions can heavily impact the profitability of the restaurant.

Figure 1: Location of each postal code within Toronto, Canada.

Now that the region has been clearly visualized, the Foursquare API was used to explore each neighborhood and return the top 200 venues within 2,000 meters (1.2 miles) of the longitude and latitude for each postal code. The extracted venue categories were encoded using one-hot encoding and the total restaurants and Indian restaurants in each region were calculated (Figure 2).

	Neighborhood	Total Restaurants	Indian Restaurants
0	Adelaide, King, Richmond	34	0
1	Agincourt	53	3
2	Agincourt North, L'Amoreaux East, Milliken, St...	60	3
3	Albion Gardens, Beaumond Heights, Humbergate, ...	19	3
4	Alderwood, Long Branch	40	0
5	Bathurst Manor, Downsview North, Wilson Heights	24	0
6	Bayview Village	18	0
7	Bedford Park, Lawrence Manor East	44	1
8	Birch Cliff, Cliffside West	13	0
9	Bloordale Gardens, Eringate, Markland Wood, Ol...	6	0

Figure 2: Result of calculating the number of restaurants in every region.

With the resulting data, the Postal Code, Borough name, Latitude, Longitude and Density columns of each region were dropped from the dataFrame. Then, the population, area, spending power, total number of restaurants and the number of Indian restaurants were used to train a k-Means clustering algorithm with 5 clusters (Figure 3). The characteristics of the resulting clusters can be found in Table 1.

Results

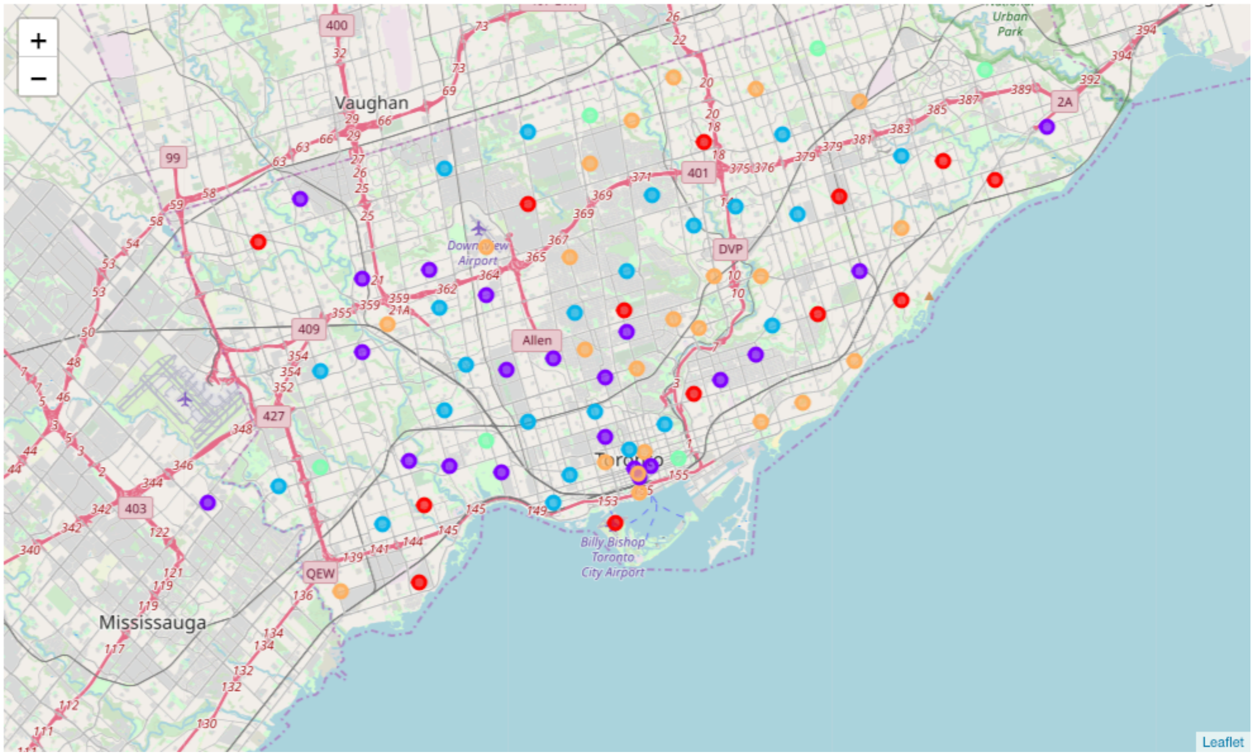

Figure 3: Result of the clustering algorithm. Cluster 0 = **Red** Cluster 1 = **Purple** Cluster 2 = **Blue** Cluster 3 = **Turquoise** Cluster 4 = **Orange**

Cluster	**Characteristics**
Cluster 0	Positive Spending Power (0.3 – 1.8)
Cluster 1	Negative Spending Power (-1.2 -- -0.8)
Cluster 2	Near Zero Spending Power (-0.5 – 0.5)
Cluster 3	High Positive Spending Power (1.7+)
Cluster 4	Negative Spending Power (-0.8 – 0) With Large Number of Restaurants

Table 1: Characteristics of the clusters resulting from k-Means clustering algorithm

	Borough	Cluster Labels	Population	Area	Spending Power	Total Restaurants	Indian Restaurants
0	Scarborough	3	90290.0	45.74	1.756524	12	0
14	Scarborough	3	86468.0	19.96	1.712083	60	3
21	North York	3	90362.0	13.80	2.350813	50	0
46	Downtown Toronto	3	76716.0	8.01	3.838132	37	0
70	West Toronto	3	82712.0	10.51	2.841538	29	1
79	Etobicoke	3	105450.0	26.38	3.748670	17	0

Figure 4: Data from neighborhoods belonging to Cluster 3

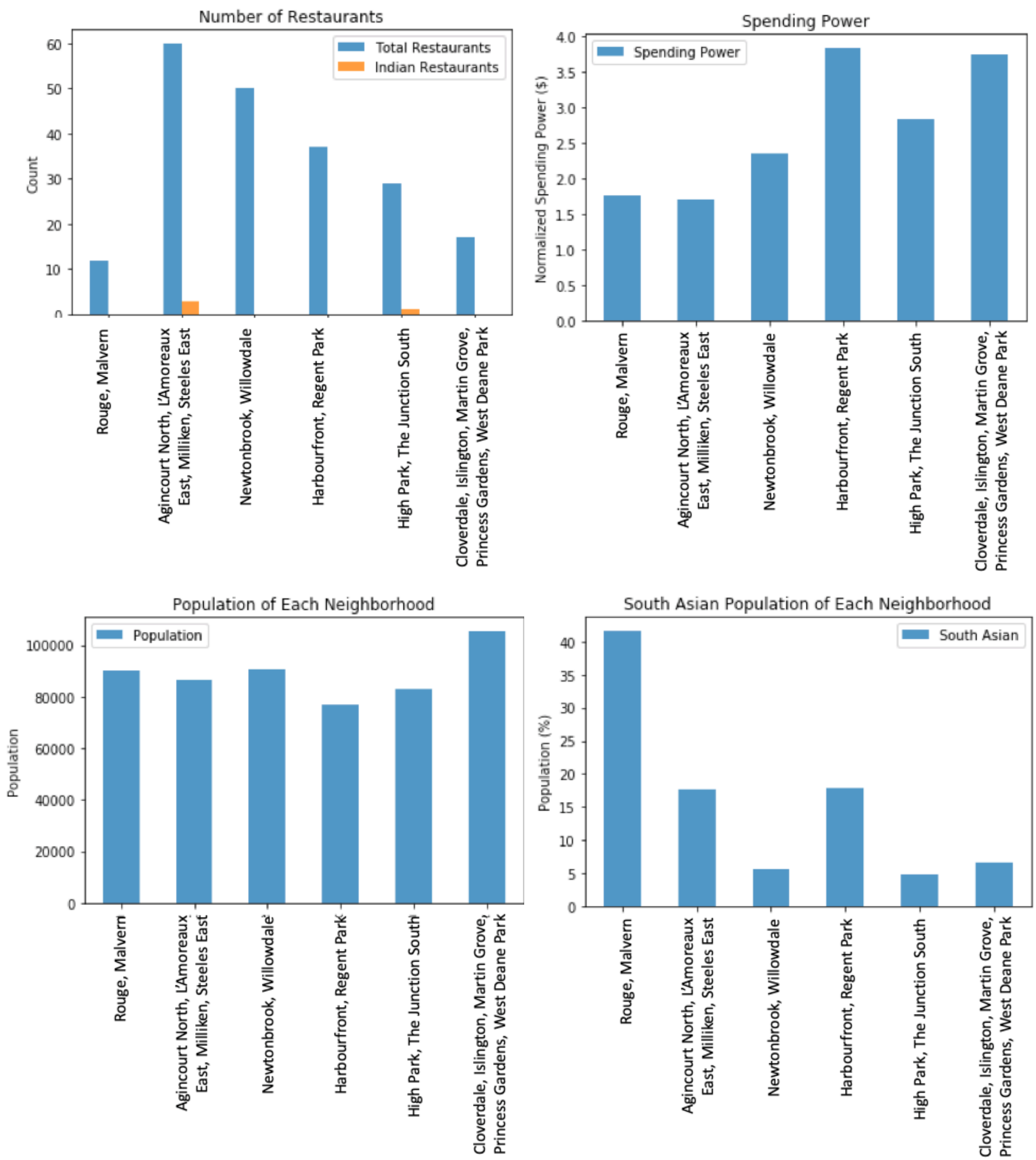

Figure 5: These plots shows the characteristics of neighborhoods belonging to cluster 3.

Discussion

From the results of the clustering algorithm, it was determined that neighborhoods corresponding to cluster 3 were the best choice for opening an Indian restaurant based on the normalized spending power and population. This narrowed down possible locations to six different areas. Using the results in Figure 5, the Agincourt North, L'Amoreaux East, Milliken, Steeles East region the Newtonbrook, Willowdale region and the Harbourfront, Regent Park region were eliminated due to the large number of restaurants in the area.

From the three remaining regions, I would recommend that the client open his/her restaurant in either the Rouge, Malvern region or the Cloverdale, Islington, Martin Grove, Princess Gardens, West Deane Park region. Both regions have very few restaurants and are farther away from the downtown area. While the Cloverdale, Islington, Martin Grove, Princess Gardens, West Deane Park region has a higher spending power and population, the Rouge, Malvern region has a higher percentage of South Asians and thus the optimal region to open the Indian Restaurant.

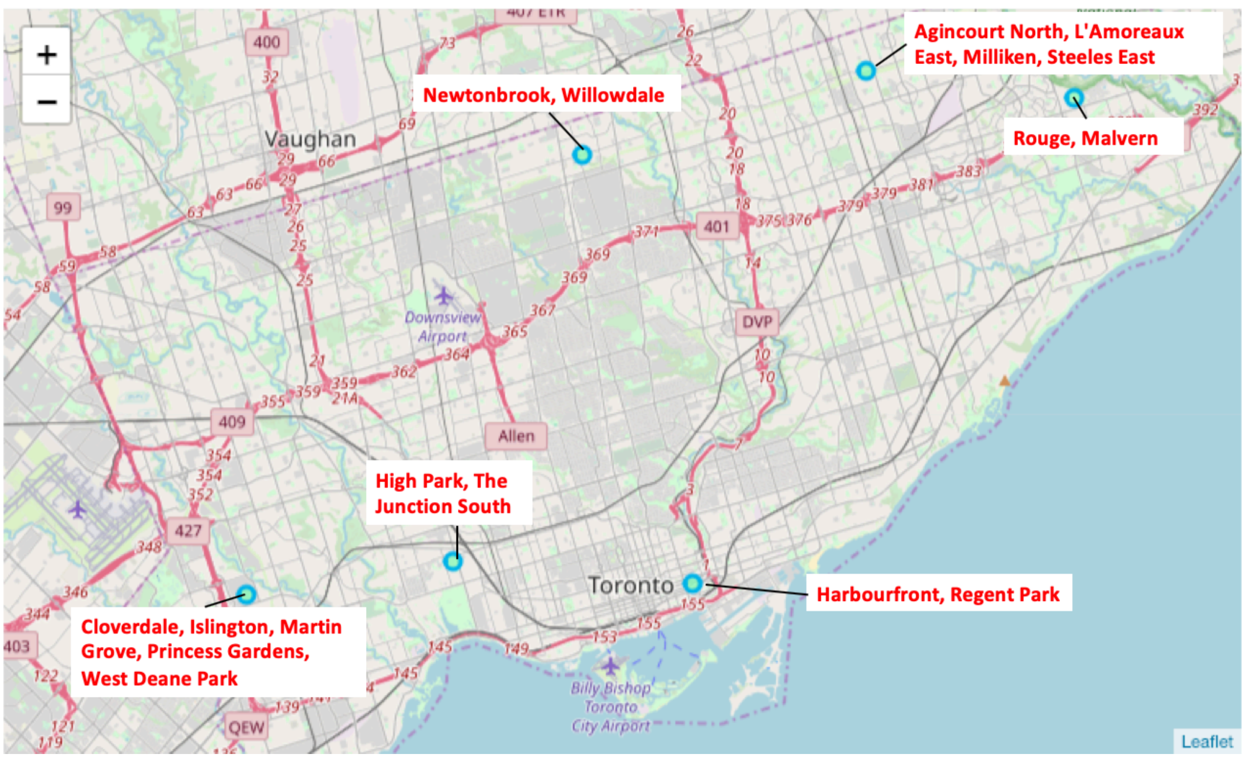

Figure 6: Map of Toronto with the neighborhoods in Cluster 3 labeled.

Conclusion

Opening a restaurant is a complex task that can lead to a large monetary loss if not done properly. Thus, extensive research about the area would greatly increase the likelihood of the restaurant succeeding. From the project above, I demonstrated the workflow necessary for a client to determine what area the restaurant should open. For specifically, I determined that the optimal location to open an Indian restaurant in Toronto should be in the Rouge, Malvern region.

Appendix

	PostCode	Borough	Neighborhood	Latitude	Longitude	Population	Density	Area	< 5k	5k - 10k	10k - 15k	15k - 20k
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353	90290.0	6208.0	45.74	290.0	240.0	420.0	720.0
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	12494.0	2403.0	5.20	60.0	25.0	45.0	60.0
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711	54764.0	8570.0	19.04	315.0	540.0	815.0	970.0
3	M1G	Scarborough	Woburn	43.770992	-79.216917	53485.0	4345.0	12.31	435.0	455.0	685.0	1170.0
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476	29960.0	4011.0	7.47	615.0	220.0	255.0	450.0

...

20k - 25k	25k - 30k	30k - 35k	35k - 40k	40k - 45k	45k - 50k	50k - 60k	60k - 70k	70k - 80k	80k - 90k	90k - 100k	100k - 125k	125k - 150k	150k - 200k	> 200k	South Asian
730.0	925.0	955.0	1090.0	1055.0	1110.0	2330.0	2150.0	1930.0	1845.0	1640.0	3355.0	2315.0	2390.0	1300.0	41.64
70.0	80.0	90.0	120.0	80.0	115.0	230.0	230.0	200.0	195.0	210.0	490.0	410.0	550.0	440.0	36.14
880.0	890.0	905.0	885.0	905.0	815.0	1565.0	1360.0	1255.0	1140.0	1050.0	1970.0	1320.0	1390.0	915.0	18.74
825.0	960.0	910.0	950.0	955.0	815.0	1725.0	1405.0	1240.0	1070.0	865.0	1660.0	1030.0	855.0	430.0	40.28
370.0	475.0	465.0	520.0	495.0	530.0	935.0	845.0	765.0	615.0	575.0	1015.0	700.0	635.0	275.0	27.72

...

Chinese	Black	Filipino	Latin American	Arab	Southeast Asian	West Asian	Korean	Japanese	White	Spending Power
6.00	16.49	9.92	1.41	0.84	0.55	1.32	0.16	0.15	14.64	2.331712e+09
7.64	12.41	6.44	1.64	0.68	0.68	0.80	1.04	0.28	25.49	3.970375e+08
3.44	15.05	8.04	1.74	0.50	0.90	1.29	0.37	0.53	43.03	1.511462e+09
6.95	10.91	7.65	1.39	1.14	0.59	2.47	0.39	0.19	23.36	1.240412e+09
14.69	6.38	9.63	1.77	1.12	1.03	2.72	0.68	0.52	26.77	7.651875e+08

Figure 7: The above image shows the first 5 rows of the imported dataFrame used during the project. The dataFrame contains the postal code, borough name, neighborhood name, latitude and longitude of each postal code in Toronto, Canada. Additionally, the table contains the population in each neighborhood, population density (people per square kilometer), area (squared kilometers) and the income distribution in Canadian dollars. Finally, the ethnicity distribution in % for each neighborhood is also included.