IBM Data Science

Capstone Project – Toronto Supermarket

Mustapha Ziade

I. Introduction

Peter just graduated from university and wants to start his own business. In fact, Peter wants to expand his family business of Supermarket. While his family business is operating in his hometown, Peter decides to expand the business in Toronto, where he studied. With his expertise in this industry, in addition to his knowledge of the market, and his relationship with suppliers, Peter is expecting the new business to be very successful. However, he is not sure in which area to open the supermarket. This project is intended to help Peter identify the best location for his business. Each neighbourhood on Toronto will be studied separately. There will be two main criteria to determine the score of neighbourhoods; competition which will be the number the supermarket in the area, and the demand which will be the population of the area. In the end, the neighbourhood with the best score will be the one to be advised for opening the business.

II. Data

As there are two criteria to determine the score (competition and demand), two data sources will be needed:

First, the competition is determined by the number of supermarkets in a neighbourhood. This information will be collected from Foursquare location data. Venues will be searched for the term “Supermarket” while having the latitude & the longitude fixed to the middle of each neighbourhood and having a radius of 1000. This data will tell how many nearby supermarkets each area has.

Second, the demand is determined by the population of the neighbourhoods. This data will be retrieved from Wikipedia. The following link is the page that will be used to get the name of neighbourhoods and their population.
https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods
The below table is part of the main table on the Wikipedia page.

Name	FM	Census Tracts	Population	Land area (km2)	Density (people/km2)	% Change in Population since 2001	Average Income	Transit Commuting %	% Renters	Second most common language (after English) by name	Second most common language (after English) by percentage	Map
Toronto CMA Average		All	**5,113,149**	**5903.63**	**866**	**9.0**	**40,704**	**10.6**	**11.4**			
Agincourt	S	0377.01, 0377.02, 0377.03, 0377.04, 0378.02, 0378.08, 0378.14, 0378.23, 0378.24	44,577	12.45	3580	4.6	25,750	11.1	5.9	Cantonese (19.3%)	19.3% Cantonese	
Alderwood	E	0211.00, 0212.00	11,656	4.94	2360	-4.0	35,239	8.8	8.5	Polish (6.2%)	06.2% Polish	
Alexandra Park	OCoT	0039.00	4,355	0.32	13,609	0.0	19,687	13.8	28.0	Cantonese (17.9%)	17.9% Cantonese	

From the table, the name of neighbourhood and their population is the only data needed. The rest will be dropped.

III. Methodology

- Step 1: Extracting neighbourhood list
  In this step, the table from Wikipedia will be extracted using the BeautifulSoup package.
- Step 2: Cleaning the data
  Most of the columns in the table are not needed. They will be dropped. Next, the column “Name” will be moved out of index and will be renamed “Neighbourhood”. Finally, there is a table with Neighbourhood and population only.
- Step 3: Getting Coordinates
  Using geocoder, Python will be defined to run at each row, get the name of the neighbourhood and get its coordinate. After that, Latitude and Longitude will be added as two new columns.

- Step 4: Extracting Supermarket Data
  Using Foursquare API, a URL will be created for each row to and search the database for supermarket within a radius of 1000. The total number will be counted and added to list. After all rows are done and the list is full, it will be added as a new column representing the number of supermarkets.
- Step 5: Extracting Convenience Store Data
  Same as step 4, the procedure will be repeated, but now searching for convenience stores instead.
- Step 6: Calculating Total Competitors
  Simply, the sum of columns of Supermarkets and Convenience Stores will be calculated and added as a new column representing the total competition.
- Step 7: Calculating the Score
  The score is calculated by dividing the population to the total number of competitors. In the end, the data will be sorted by descending order. The neighbourhood with the highest score will be ranked first and it means it is the most attractive to open the new business.

IV. Results

After cleaning the Wikipedia table, here is how the table looked like:

	Neighbourhood	Population
1	Agincourt	44577
2	Alderwood	11656
3	Alexandra Park	4355
4	Allenby	2513
5	Amesbury	17318

And this is the table after adding the coordinates:

	Neighbourhood	Population	Latitude	Longitude
1	Agincourt	44577	43.786260	-79.280840
2	Alderwood	11656	43.604960	-79.541160
3	Alexandra Park	4355	43.651090	-79.405500
4	Allenby	2513	43.712674	-79.547686
5	Amesbury	17318	43.702833	-79.481727

Adding Supermarkets Count:

	Neighbourhood	Population	Latitude	Longitude	Supermarkets
1	Agincourt	44577	43.786260	-79.280840	1
2	Alderwood	11656	43.604960	-79.541160	1
3	Alexandra Park	4355	43.651090	-79.405500	8
4	Allenby	2513	43.712674	-79.547686	1
5	Amesbury	17318	43.702833	-79.481727	3

Adding Convenience Stores Count:

	Neighbourhood	Population	Latitude	Longitude	Supermarkets	Convenience Stores
1	Agincourt	44577	43.786260	-79.280840	1	2
2	Alderwood	11656	43.604960	-79.541160	1	2
3	Alexandra Park	4355	43.651090	-79.405500	8	20
4	Allenby	2513	43.712674	-79.547686	1	3
5	Amesbury	17318	43.702833	-79.481727	3	3

Computing the Final Score:

	Neighbourhood	Population	Latitude	Longitude	Supermarkets	Convenience Stores	Total Competitors	Score
1	Agincourt	44577	43.786260	-79.280840	1	2	3	14859.000000
98	Malvern	44324	43.810230	-79.220380	2	1	3	14774.666667
168	Willowdale	43144	43.782270	-79.428130	1	2	3	14381.333333
102	Milliken	26272	43.823250	-79.277290	1	1	2	13136.000000
86	L'Amoreaux	45862	43.797300	-79.312220	1	3	4	11465.500000
131	Rouge	22724	43.807660	-79.174050	1	1	2	11362.000000
13	Bendale	28945	43.759630	-79.257390	1	2	3	9648.333333
50	Elia (Jane and Finch)	48003	43.757262	-79.517709	2	3	5	9600.600000
59	Glen Park	18426	43.649560	-79.552270	1	1	2	9213.000000
44	Downsview	36613	43.720188	-79.499920	1	3	4	9153.250000
144	Sunnylea	17602	43.643016	-79.498345	1	1	2	8801.000000
172	York Mills	17564	43.746475	-79.391617	1	1	2	8782.000000
139	Smithfield	34996	43.631590	-79.485776	1	3	4	8749.000000
126	Richview	26053	43.684538	-79.516774	1	2	3	8684.333333
159	Victoria Village	17047	43.731540	-79.314280	1	1	2	8523.500000
75	Humber Valley Village	14453	43.641471	-79.492537	1	1	2	7226.500000
108	Newtonbrook	36046	43.787300	-79.409830	1	4	5	7209.200000
11	Bayview Woods – Steeles	13298	43.794850	-79.382220	1	1	2	6649.000000
113	Old East York	52220	43.696220	-79.332890	1	7	8	6527.500000
69	Highland Creek	12853	43.789480	-79.176140	1	1	2	6426.500000
63	Guildwood	12820	43.749530	-79.189920	1	1	2	6410.000000

Now finally there is a score which can help in making the decision of location for opening the new business. After sorting the result, the top neighbourhoods seem to be a pretty choice for the decision. One might select among the top 6 area as number 5 & 6 are very close in term of score.

V. Discussion

First thing to notice is that making a decision pure on the competitions is not the best solution. As it appears for example, the neighbourhood “Guildwood”, which is in the end of the table above, has only 1 supermarket and 1 convenience store, yet it is not on the top of the table and that is because of the population. There is low demand in this area making it less attractive.

Second, looking at the top of the list, the best locations are “Agincourt”, “Malvern”, “Willowdale”, “Milliken”, “L’Amoreaux”, & “Rouge”. However, “Milliken” and “Rouge” have almost half the population of the others. And so, it is better to drop them. As “L’Amoreaux” has 4 while the others have 3, there is an advantage to pick the other locations. Finally, among the top 3, “Malvern” has access to 2 supermarkets while the others have only 1. Keeping in mind that the business is a supermarket, it is better to pick the other two areas.

In the end, the final winners in this analysis are these 2 neighbourhoods “Agincourt” and “Willowdale”.

VI. Conclusion

This analysis helped a lot in narrowing the choices of the decision. Furthermore, a survey to more understand the people and their needs. The results can help in picking the final location or might help as well in reconsidering another top choice.