APPLIED DATA SCIENCE CAPSTONE - FINAL REPORT

Introduction / Business Problem

Toronto is the biggest and most populated city in Canada. It is the capital of the province of Ontario and the country's financial center. It is considered a global economic and financial hub and is a high-income city in one of the highest income per capita countries in the world. Toronto is also a remarkably diverse culturally and ethnically. This poses a huge opportunity to open a commercial business to take advantage of this favorable and special features.

A good idea seems to be to open a restaurant in a specific area and neighborhood of the city, particularly in Scarborough which is a popular destination for new immigrants in Canada, and one of the most diverse and multicultural areas, with lots of natural landmarks and venues to visit. In order to analyze the most convenient area to open a restaurant it might be useful to identify highly populated neighborhoods in that area of Toronto and evaluate the income level for those areas. Once we have done that it would be useful as well to understand other social and demographic data of those areas in order to identify the kinds of restaurants that would be most likely successful.

One of the assumptions we use is that, taking that into account Toronto is a well-developed high-income city there will always be opportunities and room for new restaurants. In addition, Toronto is visited every year by millions of people that go there for both, vacation and as immigrants. This dynamism of the city offers always fresh opportunities for almost every kind of new business entering the market.

This project will be of special interest for anyone looking for an opportunity to open a new restaurant in the city of Toronto and who is aiming to obtain high levels of income and returns from his investment in doing so. This project will also be of interest for anyone who already has one or several restaurants in the city of Toronto and who is looking for the best way to relocate any of them in order to increase income and returns. It will also be useful for any kind of supply chain for restaurants that is interested in optimizing logistics, reducing distances for orders deliveries and things like that. Finally, this project will also be useful for anyone trying to understand in a better way de cultural and ethnical diversity and segmentation of Toronto in order to make further demographic or cultural analysis of the city itself.

Data Description

In order to analyze Toronto's neighborhoods, we need specific data that we can obtain from Foursquare.com.

More specifically, we need information about location and demographics that can help us understand which would be the neighborhoods that will increase the likelihood of success in opening a new restaurant.

For this objective, we are going to analyze population, income and second language information to identify highly populated neighborhoods with high mean income to increase the probability of success for a new restaurant, and the second language information will be useful to understand the ethnical and

cultural characteristics of each neighborhood to try to match them with the kind of restaurant that can have potential in the identified neighborhoods.

We are going to use the k-means clustering technique to map, segment and cluster the neighborhoods with similarities in the analyzed characteristics and group them and understand those similarities to help us make the best decision in opening a restaurant.

With all this information we will make a recommendation on opening a specific kind of restaurant in a specific neighborhood or group of neighborhoods (clusters) with objective information in order to increase its probability of success.

Methodology

In order to analyze Toronto's neighborhoods, we need specific data that we can obtain from Foursquare.com.

1. We get the information about location and demographics of the neighborhoods in Scarborough which is one of the most extensive areas in Toronto and can help us understand which would be the neighborhoods that will increase the likelihood of success in opening a new restaurant.

2. We then, get information of population, income and 2nd language for every neighborhood in Scarborough.

3. We analyze the information of these three variables in order to identify relevant ethnic groups by their 2nd language, while keeping an eye in those neighborhoods highly populated and with high income as well.

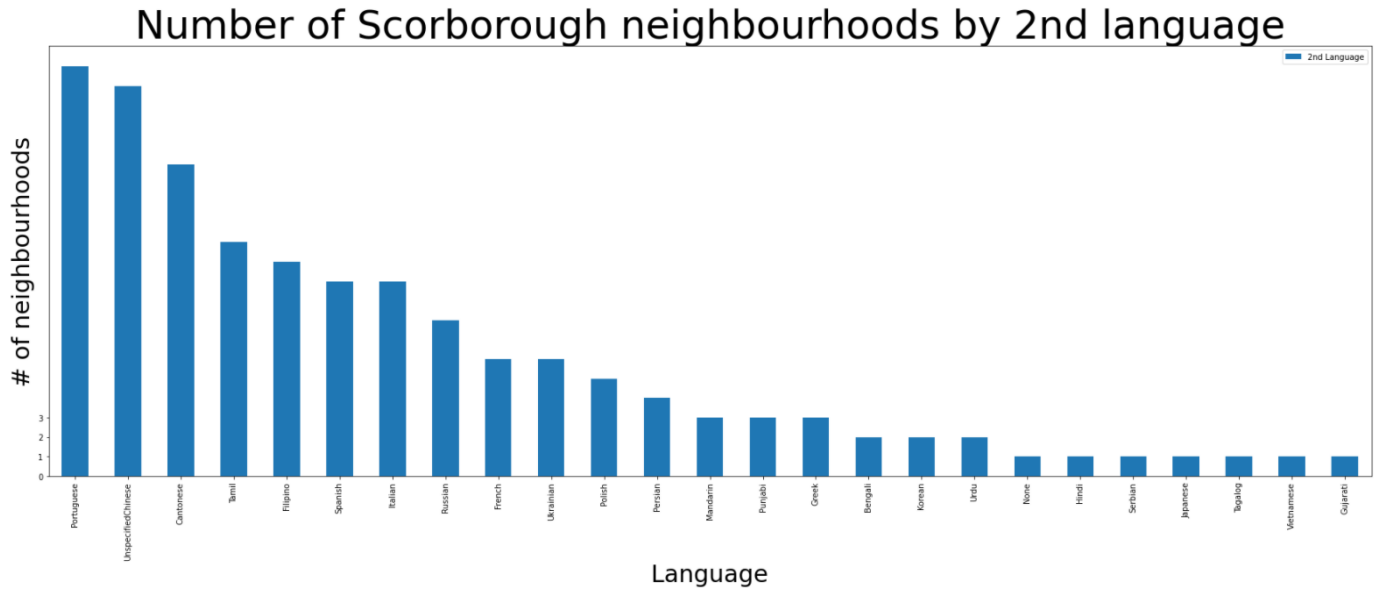
4. We also get the list of the different venues in the neighborhoods of Scarborough to get a glimpse of their location and their categories.

5. We use the k-means clustering technique using the considered variables to find statistical similarities between the neighborhoods in Scarborough and to identify the main criteria of similarity identified in the usage of the technique.

6. We finally consider all the results and gathered information and discuss the main conclusions in order to make a recommendation on opening a specific kind of restaurant in a specific neighborhood or group of neighborhoods (clusters) with objective information in order to increase its probability of success.

Results

The first interesting result is that the 2nd language of most neighborhoods in Scarborough, Toronto is Portuguese, followed by Chinese, Cantonese and Tamil as shown in the bar chart below:



With this information it is important to identify the countries that have these native languages, and we find the following:

Portuguese: Brazil, Portugal, some African countries.

Chinese: China, Taiwan, Singapur

Cantonese: China

Tamil: India, Sri Lanka, Singapur

We then explore the and group them in categories to find which categories are more common: venues in Scarborough

	Count
Coffee Shop	5
Bakery	4
Fast Food Restaurant	4
Bank	4
Intersection	4
Pizza Place	4
Breakfast Spot	3
Chinese Restaurant	3
Pharmacy	3
Fried Chicken Joint	2
Sandwich Place	2
Thai Restaurant	2
Bus Line	2
Electronics Store	2
Indian Restaurant	2

We find that the most common categories of venues are those related with food or restaurants. Though, we do not find a huge amount of any of these categories that could make us think that there is an excess of supply of any of these kinds of venues.

We finally make the k-cluster analysis for the neighborhoods in Scarborough using 5 clusters (k=5) and we find that a relevant criteria of segmentation for clusters 1,2,3 and 4 is the 2nd language. On the other hand, we have cluster 0 that includes a wide variety of 2nd languages, but they are mostly different from those identified in the other clusters as follows:

Cluster 0:

	Cluster	Neighbourhood	Population	Income	2nd Language
3	0	Allenby	2513	245592	Russian
4	0	Amesbury	17318	27546	Spanish
5	0	Armour Heights	4384	116651	Russian
7	0	Bathurst Manor	14945	34169	Russian
8	0	Bay Street Corridor	4787	40598	Mandarin
...	...	...	...	...	...
146	0	Wilson Heights	13732	37978	Filipino
147	0	Woburn	48507	26190	Gujarati
149	0	York Mills	17564	92099	Korean
150	0	York University Heights	26140	24432	Italian
151	0	Yorkville	6045	105239	French

Cluster 1:

	Cluster	Neighbourhood	Population	Income	2nd Language
6	1	Banbury	6641	92319	UnspecifiedChinese
13	1	Birch Cliff	12266	48965	UnspecifiedChinese
19	1	Cabbagetown	11120	50398	UnspecifiedChinese
34	1	Discovery District	7262	41998	UnspecifiedChinese
36	1	Don Mills	21372	47515	UnspecifiedChinese
37	1	Don Valley Village	29740	30442	UnspecifiedChinese
51	1	Garden District	8240	37614	UnspecifiedChinese
53	1	Grange Park	9007	35277	UnspecifiedChinese
57	1	Harbourfront / CityPlace	14368	69232	UnspecifiedChinese
62	1	Hillcrest	18327	33465	UnspecifiedChinese
63	1	Hoggs Hollow	3123	222560	UnspecifiedChinese
71	1	Lambton	9654	30920	Portuguese
91	1	North York City Centre	10427	34330	UnspecifiedChinese
95	1	Parkway Forest	8498	24333	UnspecifiedChinese
96	1	Parkwoods	26533	34811	UnspecifiedChinese
98	1	Playter Estates	3968	44557	UnspecifiedChinese

Cluster 2:

	Cluster	Neighbourhood	Population	Income	2nd Language
12	2	Bendale	28945	29723	Tamil
27	2	Cliffcrest	14531	38182	Tamil
38	2	Dorset Park	14189	26525	Tamil
44	2	Eglinton East	22387	25307	Tamil
61	2	Highland Creek	12853	33640	Tamil
82	2	Malvern	44324	25677	Tamil
87	2	Morningside	11472	27139	Tamil
109	2	Rouge	22724	29230	Tamil
112	2	Scarborough Junction	25780	25405	Tamil
113	2	Scarborough City Centre	16403	26756	Tamil
114	2	Scarborough Village	12796	24413	Tamil
139	2	West Hill	25632	27936	Tamil

Cluster 3:

	Cluster	Neighbourhood	Population	Income	2nd Language
18	3	Brockton	9039	27260	Portuguese
21	3	Carleton Village	6544	23301	Portuguese
24	3	Christie Pits	5124	30556	Portuguese
31	3	Davenport	8781	28335	Portuguese
39	3	Dovercourt Park	8497	28311	Portuguese
41	3	Dufferin Grove	9875	27961	Portuguese
42	3	Earlscourt	17240	26672	Portuguese
46	3	Fashion District	4642	63282	Portuguese
56	3	Harbord Village	5906	45792	Portuguese
58	3	Harwood	3375	22136	Portuguese
69	3	Junction Triangle	6666	28067	Portuguese
72	3	L'Amoreaux	45862	26375	UnspecifiedChinese
79	3	Little Italy	7917	31231	Portuguese
80	3	Little Portugal	5013	29224	Portuguese
90	3	Niagara	6524	44611	Portuguese

Cluster 4:

	Cluster	Neighbourhood	Population	Income	2nd Language
1	4	Agincourt	44577	25750	Cantonese
2	4	Alexandra Park	4355	19687	Cantonese
9	4	Bayview Village	12280	46752	Cantonese
10	4	Bayview Woods – Steeles	13298	41485	Cantonese
17	4	Bridle Path	1540	314107	Cantonese
28	4	Cliffside	9386	32701	Cantonese
43	4	East Danforth	21440	33847	Cantonese
70	4	Kensington Market	3740	23335	Cantonese
78	4	Leslieville	23567	30886	Cantonese
85	4	Milliken	26272	25243	Cantonese
105	4	Riverdale	31007	40139	Cantonese
120	4	Steeles	24696	26660	Cantonese
126	4	The Beaches	20416	67536	Cantonese
135	4	Upper Beaches	19830	44346	Cantonese
140	4	West Rouge	9300	44605	Cantonese

Finally, we calculate the mean income per cluster as follows:

Cluster	Income
0	$53.836
1	$61.929
2	$28.327
3	$33.122
4	$53.560

Being cluster 1 (Chinese) and cluster 4 (Cantonese) the clusters with highest mean income (not considering cluster 0 because of its 2nd language diversity) and also highly populated, and followed by cluster 3 (Portuguese) and cluster 2 (Tamil).

Discussion and Observation

The first interesting observation is that there is a huge number of Portuguese neighborhoods in Scarborough, followed by Chinese and then South Asian communities represented in those neighborhoods with Tamil as 2nd language.

It is also noteworthy that the venue categories are surprisingly diverse, and we see no big concentration in any of the identified categories, which represent opportunities for almost any commercial business planning to enter this market. In addition, most of the venue categories we find in Scarborough are food offer related which will be useful when considering locating supply chain units for this industry in Scarborough.

We found that a natural and effective segmentation criterion for the different neighborhoods in Scarborough, Toronto is the 2nd language. This allows us to identify the clusters with higher concentration in the ethnicities we had already identified as relevant for our analysis.

Finally, we observe that Chinese and Cantonese clusters are the clusters with highest mean income posing a big opportunity for anyone interested in opening a high-end type of restaurant or other commerce business in that area.

Conclusions

1. There should be a lot of demand for Brazilian and Portuguese food restaurants, though a good way to take advantage of this might be to open mid to lower end restaurants in neighborhoods belonging to Cluster 3, taking into account that the mean income of this cluster is much lower than others'.

2. There is a similar opportunity for Indian and South Asian food restaurants in neighborhoods belonging to Cluster 2, even though there are fewer neighborhoods with potential compared with Brazilian and Portuguese food restaurants.

3. There is a huge opportunity for Chinese food restaurants in neighborhoods belonging to Clusters 1 and 4, but this opportunity could be taken opening high-end or even luxury restaurants considering the higher mean income of the neighborhoods belonging to this clusters. This will increase the return on the investment and can help in diversifying the income sources for anyone interested also in opening Brazilian, Portuguese or South Asian food restaurants in clusters 3 and 2.

4. From the point of view of someone in the supply chain for restaurants it is clear that Scarborough is a great area with huge opportunities to locate supply chain facilities for almost every kind of food, but specially for the aforementioned such as Chinese, Brazilian, Portuguese, Indian and South Asian, food.