Opening of a grocery store in Toronto

Introduction

Toronto is a large city with a population of about 6 million in the greater Toronto area and about 3 million in the Toronto city area [1]. This population will continue to grow and there is a room for opening new stores to fulfill this continues growth. However, to gain competitive advantage companies should consider the neighborhood and analyze which has the greatest potential. Factors to be analyzed area, population in the neighborhood, competition in the neighborhood, and potential to open more stores in similar areas.

Business problem

A grocery store company named BestGrocery want to open their first store in Toronto. They want to gain first entrant advantage by being the first grocery store company in the neighborhood. This advantage is going to be used to leverage themselves into similar neighborhoods and gain a large total market share fast. However, they are not sure where their first, and most important, store should be opened. Therefore they want to do analysis and base their decision on that analysis.

Aspects to consider are population size, average income, and buying power.

Data

The data we need is data on the neighbourhoods and what stores they have and the population in each neighborhood. The neighborhood data can be accessed through the Foursquare API, however the population data is hard to access in a good format. Therefore, we will get the data by scraping information from wikipedia.

The data from the Foursquare API would consist of information about nearby venues and contains the:

- Venue name
- Venue category
- Venue location

The data that can be scraped contains:

- Neighborhood
- Population
- Average income
- Latitude
- Longitude

We want to combine this data into a single, cleaned, and formatted data frame such that we can explore the dataframe.

Methodology

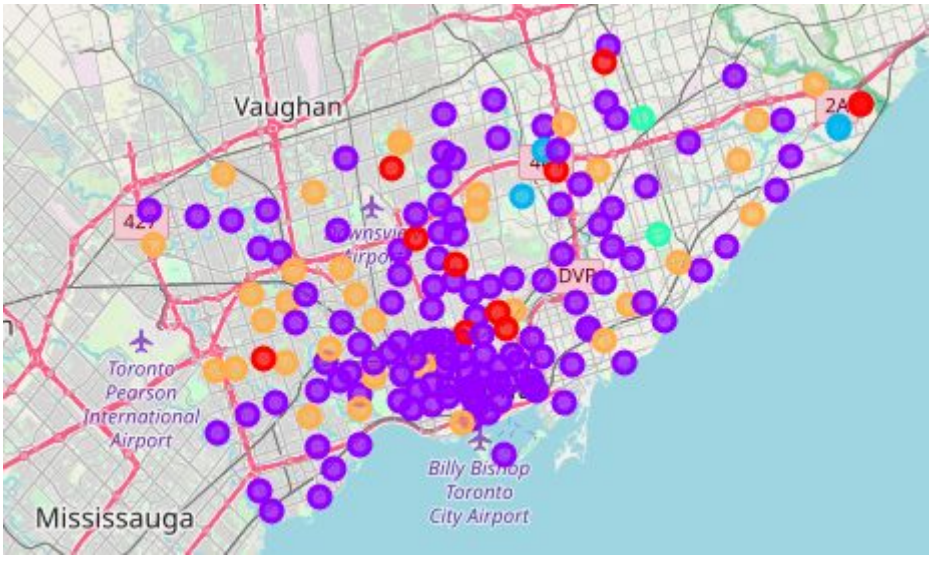
Through webscarping we were able to get scrape the data for the neighborhoods in Toronto and the we scraped the Demographics of Toronto wikipedia page [2]. This information is crucial to determine which neighborhood to target as it contains the population, and the income in that neighborhood. We used the library geocoder to gain acces about the location of each neighborhood. This data also contained information about the ethnicity of the neighborhood and the population growth in that neighborhood. By using this data we add another feature, the "buying power" of the neighborhoods. The buying power is the sum of all the income over the population in that neighborhood.

To get access to data about the neighborhood, such as what venues they have, we used the foursquare API which contains all the necessary information, including venue name, location, and category.

To solve this problem we used kmeans, an unsupervised machine learning algorithm. We set the number of clusters to 5 and then used the nearby venues as the input for this model. This would result in clusters that are grouped by the similarity of the neighborhoods venues, which type of restaurants are their, are their grocery stores, etc.

Results

With the kmeans clustering we got the following clusters, denoted by color, ie each color represent an individual cluster.



Each cluster contained different amount of neighborhoods and we can see that cluster 2 and 3 lack sufficient number of cluster to be considered. To few clusters

would results in to few possible expansion neighborhoods. As it is necessary to have a sufficiently high number of similar neighborhoods to be able to expand later on, cluster 1 and 4 are the clear winner in this regard.

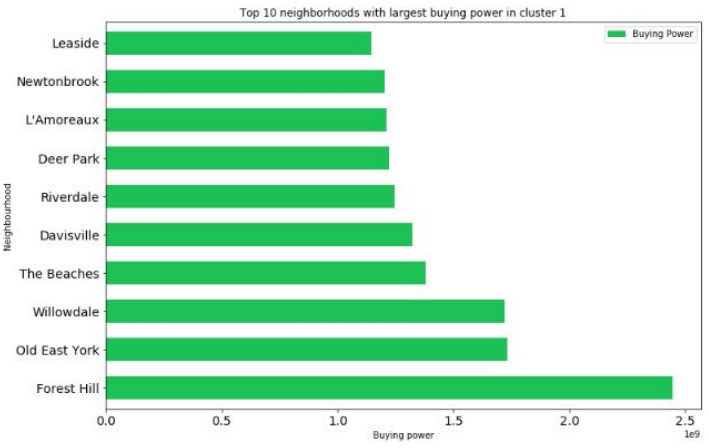
- size of cluster 0: 11
- size of cluster 1: 121
- size of cluster 2: 4
- size of cluster 3: 2
- size of cluster 4: 33

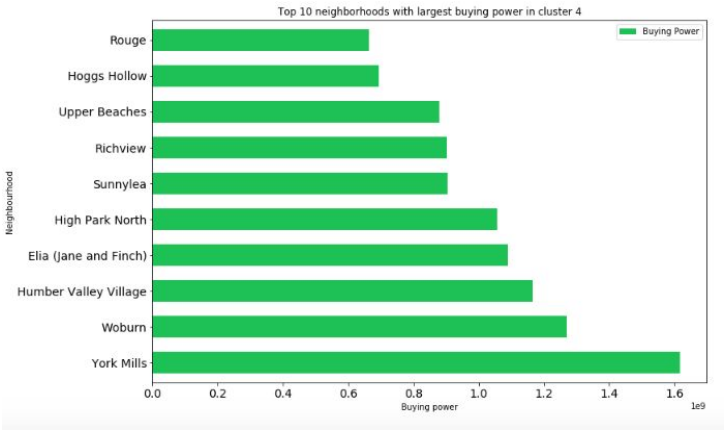
This is also shown when analyzing the sum of the population, income, and buying power over the cluster. Cluster 2 and 3 have significantly less in all features compared to the other clusters.

	Population	Income	Buying Power
0	112718.0	895185.0	7.219585e+09
1	1679089.0	5375666.0	6.522200e+10
2	34446.0	231698.0	1.807593e+09
3	57602.0	50828.0	1.474499e+09
4	494781.0	1472788.0	1.809783e+10

We then analyze the remaining clusters. The number one neighborhood based on buying power in cluster0 is Rosedale, however Rosedale have a high frequency of Grocery stores already. By removing all the neighbourhood with a high frequency of grocery stores we can analyze further.

Cluster 1 has the greatest neighborhood based on buying power, namely Forest Hill. It also have the higher sum of buying power when summing the top 10 neighbourhoods in the clusters (based on buying power). In this regards Forest Hill is the best neighborhood, and cluster 1 is the best cluster of neighborhoods.





We also look at the ethnicity of the population within the neighbourhood to see if BestGrocery can target a certain ethinicity (based on 2nd language) with certain cousine in their stores. However, looking at the top 10 neighborhoods in cluster 1 and 4, we can see little to none similairty between the neighborhoods ethnicty.

Cluster 1:

Neighbourhood	Population	Land Area	Density	Population %	Income	2nd Language	2nd Language %	Buying Power
Forest Hill	24056	4.35	5530	-0.2	101631.0	Russian (2.4%)	02.4% Russian	2.444835e+09
Old East York	52220	7.94	6577	-4.6	33172.0	Greek (4.3%)	04.3% Greek	1.732242e+09
Willowdale	43144	7.68	5618	62.3	39895.0	Cantonese (7.9%)	07.9% Cantonese	1.721230e+09
The Beaches	20416	3.57	5719	7.8	67536.0	Cantonese (0.7%)	00.7% Cantonese	1.378815e+09
Davisville	23727	3.14	7556	4.5	55735.0	Persian (1.5%)	01.5% Persian	1.322424e+09
Riverdale	31007	3.99	7771	-5.5	40139.0	Cantonese (6.7%)	06.7% Cantonese	1.244590e+09
Deer Park	15165	1.46	10,387	5.2	80704.0	Russian (1.1%)	01.1% Russian	1.223876e+09
L'Amoreaux	45862	7.15	6414	0.9	26375.0	Unspecified Chinese (13.9%)	13.9% Unspecified Chinese	1.209610e+09
Newtonbrook	36046	8.77	4110	0.3	33428.0	Russian (8.8%)	08.8% Russian	1.204946e+09
Leaside	13876	2.81	4938	3.0	82670.0	Bulgarian (0.4%)	00.4% Bulgarian	1.147129e+09

Cluster 4:

Neighbourhood	Population	Land Area	Density	Population %	Income	2nd Language	2nd Language %	Buying Power
York Mills	17564	7.29	2409	2.0	92099.0	Korean (4.0%)	04.0% Korean	1.617627e+09
Woburn	48507	13.34	3636	-1.5	26190.0	Gujarati (9.1%)	09.1% Gujarati	1.270398e+09
Humber Valley Village	14453	5.45	2652	-0.1	80618.0	Ukrainian (3.9%)	03.9% Ukrainian	1.165172e+09
Elia (Jane and Finch)	48003	7.66	6267	-10.0	22691.0	Vietnamese (6.9%)	06.9% Vietnamese	1.089236e+09
High Park North	22746	2.18	10,434	-1.6	46437.0	Polish (3.0%)	03.0% Polish	1.056256e+09
Sunnylea	17602	5.23	3366	-1.1	51398.0	Polish (5.2%)	05.2% Polish	9.047076e+08
Richview	26053	6.51	4002	-4.0	34579.0	Italian (4.2%)	04.2% Italian	9.008867e+08
Upper Beaches	19830	2.92	6791	0.5	44346.0	Cantonese (0.7%)	00.7% Cantonese	8.793812e+08
Hoggs Hollow	3123	2.76	1132	2.0	222560.0	Unspecified Chinese (2.4%)	02.4% Unspecified Chinese	6.950549e+08
Rouge	22724	28.72	791	175.0	29230.0	Tamil (15.6%)	15.6% Tamil	6.642225e+08

To end the results, we analyze the population growth in each of the top 10 neighborhoods in both clusters. In cluster 1, a majority is increasing their population. While in cluster 4 a majority of the neighborhoods have a decrease in population. However, the greatest population growth is found in cluster 4, in neighborhood Rogue. In willowdale, cluster 1, we can also see a noteworthy population growth of 62.3%. In this regards cluster 1 wins, but the greatest neighborhood is found in cluster 4.

Summary of results:

	Cluster 1	Cluster 4
Size of cluster	121	33
Buying power (top 10)	1.462970e+10	1.024294e+10
Similarity of cousine (top 10)	Not found	Not found
Number of neighborhoods with positive population growth (top 10)	7	4

Discussion and conclusion

Based on the results, it is clear that cluster one is the best one to enter. This is since,there are a large amount of similar neighborhoods that one could expand into, it has the largest buying power, and its neighborhoods are growing in population.

Within cluster 1, neighborhood Forest Hill has great potential. Since Willowdale is in the top 3 of the greatest buying power within cluster 1, and had a high population growth, BestGrocery should look into expanding there as well. Further analysis of these two neighborhoods should be conducted, analyzing aspects such as:

- Property and rent costs
- Labor market

References:

[1] http://worldpopulationreview.com/world-cities/toronto-population/

[2] https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods