Convenient Sites to Set up a Coffee Shop

Finding strategic locations in the neighborhoods of Toronto, Canada to establish a Coffee Shop using K-Means Clustering

Introduction

Background

One of the biggest factors when you are trying to realize your dream of opening a coffee shop business is the location that you choose. The location of a coffee shop, or just about any type of business for that matter, plays a huge role in the success of the shop. Choosing the right location is key to any good business endeavour.

Business Problem

A client seeks to establish a franchised Coffee Shop, with a niche in Southeast Asian cuisine, in a Toronto neighborhood. Which neighbourhood would appear to be the optimal and most strategic location for the business operations?

The objective of this capstone project is to locate the optimal neighborhood for operation. Our foundation of reasoning would be based on spending power, population, and competition, across each neighbourhood. We will mainly be utilizing the Foursquare API and the extensive geographical and census data from Toronto's Open Data Portal.

Data collection and cleaning

Data Sources

1. Neighborhood names along with their Postal Code data was scraped off of Wikipedia (https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&oldid=945633050).

2. Coordinates were collected from https://cocl.us/Geospatial_data.

3. Demographic data on population, household income, unemployment, youth was collected from http://map.toronto.ca/wellbeing.

4. Coffee shop venue data was collected using Foursquare API (https://developer.foursquare.com/).

Data Cleaning

- From the data scrapped from Wikipedia we removed all the neighborhoods without an assigned borough.
- Neighborhood and borough names were corrected using string functions.
- The Foursquare API data on venue was collected using the postal codes from the Wikipedia.

- Once the Foursquare data was collected the data was converted using one hot encoding, then the frequency of each venue was calculated out of which only Coffee Shop and Café' columns were kept. The Coffee Shop and Café' tables were then added together.
- Finally, the Foursquare data, the coordinate data, demographic data and the postal codes data were merged, the resultant dataset can be seen in the following image:

	Postal Code	Neighborhood	Latitude	Longitude	Total Population	Income	Unemployment	Youth	Coffee Shop
0	M4A	Victoria Village	43.725882	-79.315572	17180.0	47529.0	5.325960	11.059371	0.200000
1	M1B	Rouge	43.806686	-79.194353	45905.0	81553.0	5.391570	15.118179	0.000000
2	M1B	Malvern	43.806686	-79.194353	45085.0	57528.0	6.443385	15.437507	0.000000
3	M1C	Highland Creek	43.784535	-79.160497	13100.0	98087.0	4.923664	15.725191	0.000000
4	M3C	Flemingdon Park	43.725900	-79.340923	22165.0	46554.0	7.150914	13.918340	0.095238

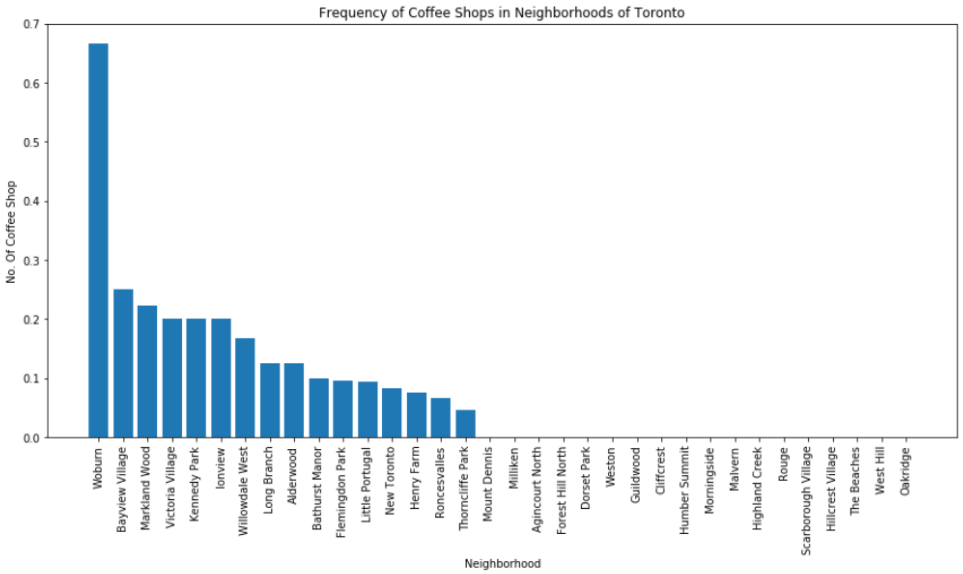
Exploratory Data Analysis

Folium Mapping

Visualizations

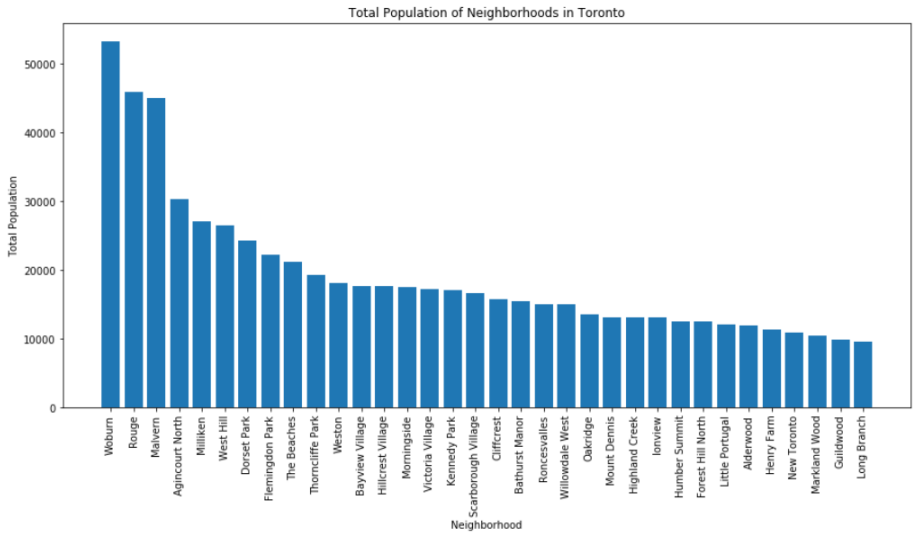
- ## Frequency Distribution of Coffee Shops

The Foursquare API returned the frequencies of coffee shops in the neighborhoods in Toronto. Using matplotlib we could visualize the frequencies of coffee shops in the neighborhoods.
Frequency would directly indicate the amount of competition in a neighborhood. Choosing a neighborhood with a fair amount of competition is important as it provides the demand for such stores. But choosing a neighborhood with very high amount of competition is not advisable.



- ## Distribution of Population

Population is an important factor in choosing a location for your Coffee Shop as it will decide the amount of foot and car traffic the shop is going face. Highly populated areas always have high foot traffic. Foot and car traffic are key, so choosing a location in the vicinity of a business district, shopping mall, or university is always a good option. Such areas are going to increase the visibility of your store and hence increase customers.



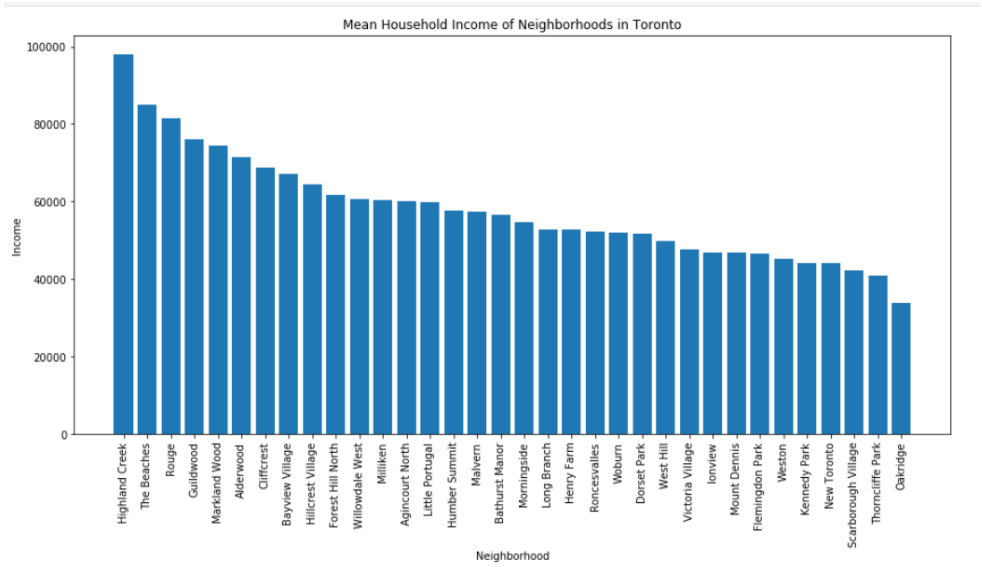
- Distribution of Median Household Income

When you are choosing a location, you must think about your budget also. Some areas will be more expensive than others. The cost of living in an area will also help to dictate how much you can charge for the items on your menu. These are location considerations that you will need to decide between because of how they will affect you financially, from your initial investment to your ongoing operations, and to your profit margins.

High Median Household Income Neighborhood – you are looking at a higher cost for entry into the market, higher operating costs, and higher rent or mortgage costs. If you do not have enough money to keep up with what the area will require a business to be successful, then try looking at a different market. However, if you can hang in this type of market, there is major potential for high profits.

Low Median Household Income Neighborhood – Such areas can be very profitable for coffee shop businesses without having to spend a ton of money to get started, but not without proper assessment of the wants and needs of the community.

Average Median Household Income Neighborhood – If you want to start a coffee shop in an area that has lots of traffic but does not cost as much to open, then look for a location in such neighborhoods. Typically, you will find lower real estate prices and lower lease payments. You can still offer a style and menu that is appreciated in the urban setting. Select your location wisely though, by taking into consideration all the other location suggestions we have already talked about.



Predictive Modelling

Data Pre-processing

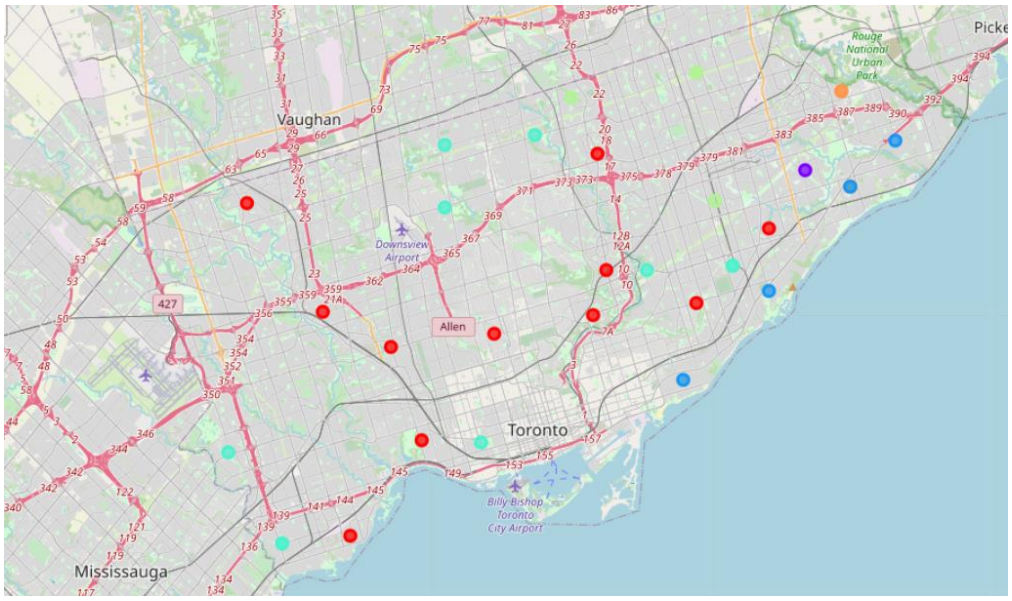
To help mathematical-based algorithms — like our k-Means algorithm in this case — to interpret features with different magnitudes and distributions equally, we will have to normalize our data; as these feature columns are different in scale, we will standardize the values to a common scale. One approach of data normalization is *StandardScaler*.

k-Means Clustering

Before we fit the feature values into our model, we have to pre-assign the number of clusters the algorithm should label. In order to identify the optimal number clusters to use, a range of 3 to 10 clusters were used, then the squared error calculated respectively were used as metrics of their performances.

An analysis using K Elbow Visualizer and Squared error for each k value evident shows that k = 6 would the best value. After identifying the number of clusters, we will fit the standardized feature values into our k-Means algorithm; resulting in 6 clustered neighbourhoods of similar characteristics.

Cluster Labels



Cluster 0

[91]:

	Cluster Label	Postal Code	Neighborhood	Latitude	Longitude	Total Population_x	Income_x	Coffee Shop
7	0	M1E	Morningside	43.763573	-79.188711	17585.0	54746.0	0.000000
14	0	M1J	Scarborough Village	43.744734	-79.239476	16615.0	42131.0	0.000000
23	0	M6M	Mount Dennis	43.691116	-79.476013	13150.0	46811.0	0.000000
4	0	M3C	Flemingdon Park	43.725900	-79.340923	22165.0	46554.0	0.095238
21	0	M9L	Humber Summit	43.756303	-79.565963	12530.0	57786.0	0.000000
15	0	M2J	Henry Farm	43.778517	-79.346556	11340.0	52675.0	0.075758
24	0	M9N	Weston	43.706876	-79.518188	18170.0	45119.0	0.000000
31	0	M8V	New Toronto	43.605647	-79.501321	10905.0	44141.0	0.083333
20	0	M1L	Oakridge	43.711112	-79.284577	13505.0	33980.0	0.000000
26	0	M5P	Forest Hill North	43.696948	-79.411307	12475.0	61596.0	0.000000
13	0	M4H	Thorncliffe Park	43.705369	-79.349372	19225.0	40795.0	0.045455
28	0	M6R	Roncesvalles	43.648960	-79.456325	15050.0	52154.0	0.066667

Cluster 1

86]:

	Cluster Label	Postal Code	Neighborhood	Latitude	Longitude	Total Population_x	Income_x	Coffee Shop
10	1	M1G	Woburn	43.770992	-79.216917	53350.0	52018.0	0.666667

Cluster 2

[87]:

	Cluster Label	Postal Code	Neighborhood	Latitude	Longitude	Total Population_x	Income_x	Coffee Shop
3	2	M1C	Highland Creek	43.784535	-79.160497	13100.0	98087.0	0.0
22	2	M1M	Cliffcrest	43.716316	-79.239476	15700.0	68647.0	0.0
6	2	M1E	Guildwood	43.763573	-79.188711	9815.0	76055.0	0.0
9	2	M4E	The Beaches	43.676357	-79.293031	21135.0	85028.0	0.0

Cluster 3

93]:

	Cluster Label	Postal Code	Neighborhood	Latitude	Longitude	Total Population_x	Income_x	Coffee Shop
17	3	M1K	Ionview	43.727929	-79.262029	13095.0	46955.0	0.200000
18	3	M1K	Kennedy Park	43.727929	-79.262029	17050.0	44226.0	0.200000
19	3	M2K	Bayview Village	43.786947	-79.385975	17675.0	67186.0	0.250000
33	3	M8W	Long Branch	43.602414	-79.543484	9630.0	52771.0	0.125000
12	3	M3H	Bathurst Manor	43.754328	-79.442259	15435.0	56563.0	0.100000
27	3	M2R	Willowdale West	43.782736	-79.442259	15005.0	60537.0	0.166667
32	3	M8W	Alderwood	43.602414	-79.543484	11900.0	71585.0	0.125000
5	3	M9C	Markland Wood	43.643515	-79.577201	10435.0	74376.0	0.222222
0	3	M4A	Victoria Village	43.725882	-79.315572	17180.0	47529.0	0.200000
16	3	M6J	Little Portugal	43.647927	-79.419750	12055.0	59886.0	0.093023

Cluster 4

[92]:

	Cluster Label	Postal Code	Neighborhood	Latitude	Longitude	Total Population_x	Income_x	Coffee Shop
8	4	M1E	West Hill	43.763573	-79.188711	26550.0	49713.0	0.0
11	4	M2H	Hillcrest Village	43.803762	-79.363452	17650.0	64522.0	0.0
29	4	M1V	Agincourt North	43.815252	-79.284577	30280.0	60162.0	0.0
30	4	M1V	Milliken	43.815252	-79.284577	27160.0	60262.0	0.0
25	4	M1P	Dorset Park	43.757410	-79.273304	24360.0	51724.0	0.0

Cluster 5

[90]:

	Cluster Label	Postal Code	Neighborhood	Latitude	Longitude	Total Population_x	Income_x	Coffee Shop
2	5	M1B	Malvern	43.806686	-79.194353	45085.0	57528.0	0.0
1	5	M1B	Rouge	43.806686	-79.194353	45905.0	81553.0	0.0

Clusters	Population level	Household income level	Competition
0	Low	Low	Low
1	High	Low	High
2	Low	High	Low
3	Low	Average	Average
4	Average	Average	Low
5	High	High	Low

Cluster 0:

Beneficial, because of the presence of low competition and low household income. The low household income would mean the budget required would be less. But the lack of high foot traffic could be harmful for the business.

Cluster 1:

Not beneficial, because of the presence of high competition. The low household income would mean the budget required would be less. But presence of high competition in a highly populated area would be very harmful for a new business.

Cluster 2:

Not Beneficial, because of the presence of high household income and low foot traffic, which would mean the budget required would be very high and the neighborhood would yield very few customers.

Cluster 3:

Beneficial, because of average household income with average competition levels, which would mean the client can provide a high budget menu with a low budget investment because of the average real estate costs. The neighborhoods have an average competition level this suggests that there is a moderate amount of foot traffic.

Cluster 4:

Beneficial, because of average household income levels, which would mean the client can provide a high budget menu with a low budget investment because of the average real estate costs. The neighborhoods have an average foot traffic which would mean a moderate number of customers in a low competition area.

Cluster 5:

Beneficial with good investments, because of the presence of high household income and high foot traffic, which would mean the budget required would be very high and the neighborhood would yield a good number of customers.

Conclusion

The safest and the most reliable neighborhoods for a new business are in Cluster 4. Since, the area has an average mean household income, this allows our client to provide a menu which would require a hefty budget but with a reasonable investment, as the area has an average real estate cost. The area also has low competition which would be safe for the business. There is also an average level of population which would provide for a good number of customers.

Other favourable neighborhoods are Cluster 5 and Cluster 3. Cluster 5 would be beneficial but with a good investment because of the area's high real estate prices. The presence of high-density population would yield a good number of customers. Cluster 3 is also beneficial, because of its average household income with average competition levels, which would mean the client can provide a high budget menu with a low budget investment because of the average real estate costs. The neighborhoods have an average competition level this suggests that there would be moderate amount of foot traffic.