Serving Coffee in Canada

A Descriptive analysis of Cafes in Toronto

A report by:

Souparno Bhattacharya

Introduction:

Business Problem:

Identifying the best neighbourhoods in Toronto, Ontario, Canada for opening a coffee house(Cafe).

Target Audience:

This descriptive exploratory analysis is aimed at aiding anyone, from individual entrepreneurs to global franchises, aspiring to establish a Coffee house in Toronto by identifying the locations with high probability of success in the aforementioned venture.

Data:

Socio-Economic Data:

For the analysis, we use the following features of a given neighbourhood in Toronto, Ontario:

- Location (Latitude, Longitude)
- Potential customers between the age of 15 and 64
- Potential customers who are employed
- Average income of a household
- No. of permits issued for Parks, Forests and Recreation
- No. of crimes committed in a year
- No. of Hot Brewery

All the features except for the "No. of Hot Breweries in each neighbourhood" was obtained from here which contained the relevant data for the city of Toronto for the year 2016 and the "No. of Hot Breweries data was collected using the Foursquare API. Also, the location data of the neighbourhoods was retrieved using the Geopy library in python.

In this analysis, I have expanded out scope of the target variable by including several other commercial eateries that are very similar to coffee shops, namely, Tea shops and Bakeries. All of the concerned eateries are hereafter implied to be included in the canopy term "Hot Brewery" and will be referred to as "Cafes" in this analysis.

Initially, the demographic and socioeconomic data about each neighbourhood was obtained from the link mentioned above. Refer below (Exhibit 1) for the features included in the originally retrieved dataset. This dataset was further processed to obtain the actual dataset over which the analysis had been done.

The transformations done to obtain the actual dataset were:

- Clubbing of all types of crimes into a single "Crimes per week" feature.
- Subtraction of "Low Income Population" from the "Pop 15 – 64 years" and "In Labour Force".
  Note: This was done as it had been assumed that people with low income are generally unable to go to cafes frequently and will be inconsequential for the analysis.

Exhibit 3 shows the measures of central tendencies for the dataset. It can be observed that both the "Potential Customer" features and the "Crimes per week" feature have means which are considerably greater than their medians, indicating heavy skewness due to presence of outliers. Thus, for any modelling purposes, the data has to be first standardised.

Exibit 1: *The original demographic and socio-economic dataset:*

	Neighbourhood	Neighbourhood Id	Combined Indicators	Pop 15 - 64 years	In Labour Force	After-Tax Household Income	Low Income Population	PFR Permits Issued	PFR Community Space Use	Assaults	Sexual Assaults	Break & Enters	Robberies	Thefts	Hazardous Incidents
0	West Humber-Clairville	1.0	NaN	23285.0	17845.0	59703.0	7590.0	1385.0	923.0	259.0	31.0	131.0	82.0	38.0	213.0
1	Mount Olive-Silverstone-Jamestown	2.0	NaN	22300.0	14765.0	46986.0	11540.0	1799.0	2716.0	213.0	16.0	34.0	81.0	3.0	173.0
2	Thistletown-Beaumond Heights	3.0	NaN	6760.0	5060.0	57522.0	2350.0	1191.0	1716.0	35.0	3.0	23.0	12.0	1.0	52.0
3	Rexdale-Kipling	4.0	NaN	7165.0	5480.0	51194.0	2170.0	88.0	3.0	57.0	5.0	16.0	15.0	0.0	46.0
4	Elms-Old Rexdale	5.0	NaN	6370.0	4635.0	49425.0	2790.0	2388.0	242.0	53.0	2.0	9.0	14.0	0.0	64.0
5	Kingsview Village-The Westway	6.0	NaN	14175.0	10265.0	50714.0	6760.0	2.0	0.0	110.0	6.0	34.0	22.0	5.0	118.0
6	Willowridge-Martingrove-Richview	7.0	NaN	13690.0	10870.0	57048.0	3490.0	2711.0	159.0	88.0	6.0	32.0	38.0	4.0	123.0

Exhibit 2: *The dataset for Analysis:*

	Neighbourhood	Neighbourhood Id	After-Tax Household Income	PFR Permits Issued	Crimes per week	Potential customers: 15 - 64 years	Potential customers: Employed
0	West Humber-Clairville	1.0	59703.0	1385.0	14.500000	15695.0	10255.0
1	Mount Olive-Silverstone-Jamestown	2.0	46986.0	1799.0	10.000000	10760.0	3225.0
2	Thistletown-Beaumond Heights	3.0	57522.0	1191.0	2.423077	4410.0	2710.0
3	Rexdale-Kipling	4.0	51194.0	88.0	2.673077	4995.0	3310.0
4	Elms-Old Rexdale	5.0	49425.0	2388.0	2.730769	3580.0	1845.0
5	Kingsview Village-The Westway	6.0	50714.0	2.0	5.673077	7415.0	3505.0
6	Willowridge-Martingrove-Richview	7.0	57048.0	2711.0	5.596154	10200.0	7380.0

Exhibit 3: *Describing the dataset for analysis:*

	Neighbourhood	Neighbourhood Id	After-Tax Household Income	PFR Permits Issued	Crimes per week	Potential customers: 15 - 64 years	Potential customers: Employed
count	140	140.0000	140.000000	140.000000	140.000000	140.000000	140.000000
unique	140	NaN	NaN	NaN	NaN	NaN	NaN
top	Agincourt South-Malvern West	NaN	NaN	NaN	NaN	NaN	NaN
freq	1	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	70.5000	55426.500000	1547.514286	5.488187	9588.764286	6568.764286
std	NaN	40.5586	16118.155356	1436.513828	3.990230	5778.681395	4812.860533
min	NaN	1.0000	30794.000000	1.000000	1.576923	2570.000000	260.000000
25%	NaN	35.7500	46689.500000	370.750000	2.822115	6128.750000	3856.250000
50%	NaN	70.5000	52660.000000	1315.000000	4.519231	8247.500000	5587.500000
75%	NaN	105.2500	59963.000000	2264.500000	6.653846	11125.000000	7623.750000
max	NaN	140.0000	161448.000000	7770.000000	28.057692	50645.000000	44110.000000

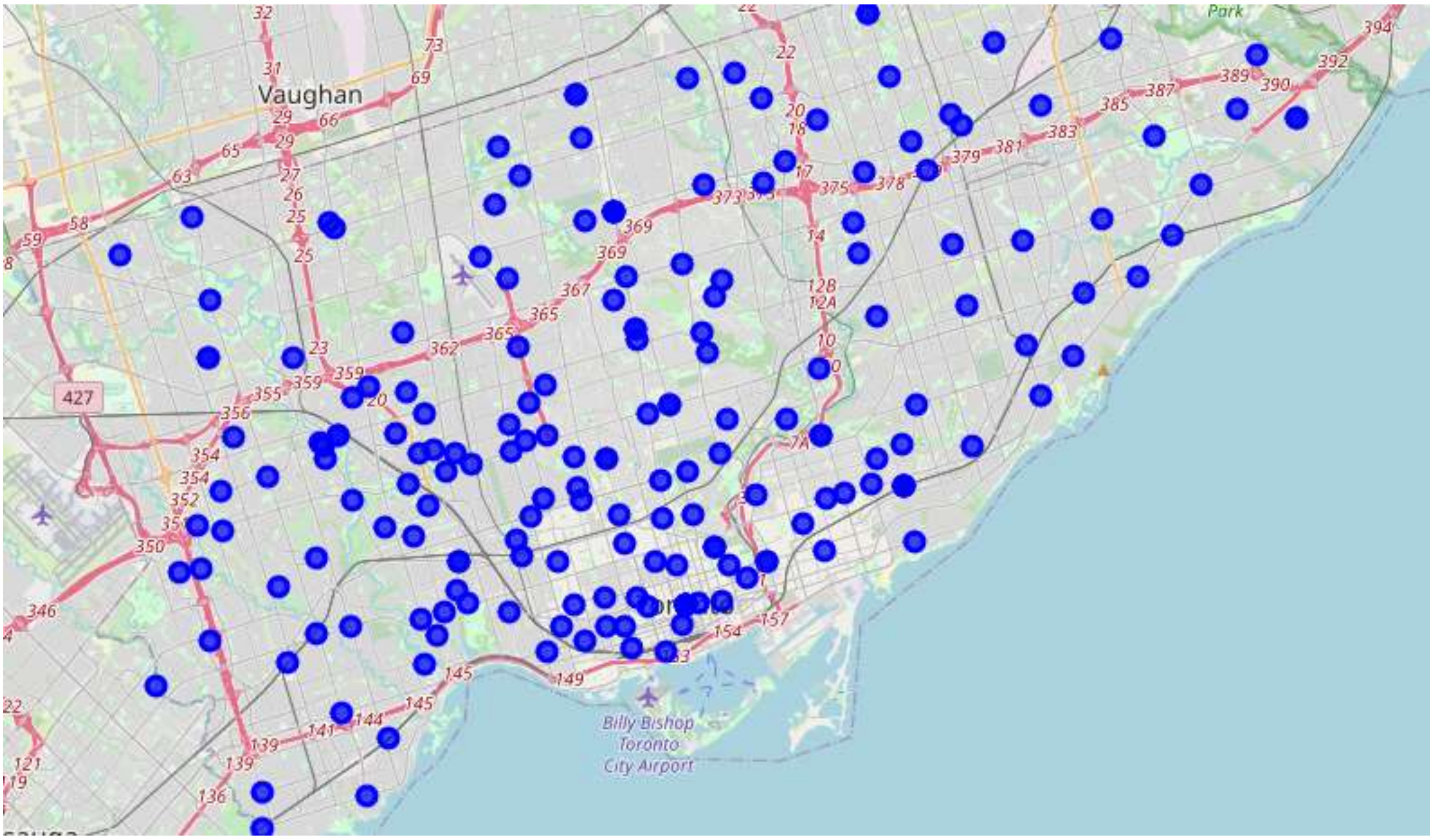
Geographical Data:

The geographical coordinates of the Neighbourhoods were obtained through the Geopy library of Python. The neighbourhoods were obtained from the dataset in Exhibit 2 and separated, with the Neighbourhood ID used for establishing integrity between the two datasets.

Exhibit 4: *Geographical Location of the Neighborhoods:*

	Neighborhood	Neighborhood ID	Latitude	Longitude
0	West Humber	1	43.680604	-79.482074
1	Mount Olive	2	43.653482	-79.383935
2	Silverstone	2	43.749751	-79.599116
3	Jamestown	2	43.653482	-79.383935
4	Thistletown	3	43.737266	-79.565317
5	Rexdale	4	43.721362	-79.565513
6	Kipling	4	43.637593	-79.535494

Exhibit 5: *Mapping the Neighbourhoods of Toronto:*



The list of venues in each of the Neighbourhoods were obtained using the Foursquare API (Exhibit 6). Once retrieved, the venues which classified under the "Hot Brewery" category were kept and the rest were dropped from the table. Later, these were clubbed together to obtain the "No. of Hot Breweries" per Neighbourhood. Refer to Exhibit 7 for the dataset.

NOTE: The “No. of Hot Breweries” dataset contained some Null values which were accordingly addressed during the analysis.

Venues Data:

Exhibit 6: *Venues and their geographical coordinates in each Neighbourhood:*

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	West Humber	43.680604	-79.482074	Tim Hortons	43.680476	-79.475961	Coffee Shop
1	West Humber	43.680604	-79.482074	Pizza Tazza	43.683500	-79.481535	Pizza Place
2	West Humber	43.680604	-79.482074	Modern Sensibility	43.678092	-79.484764	Furniture / Home Store
3	West Humber	43.680604	-79.482074	Her's Lingerie	43.680810	-79.477010	Lingerie Store
4	West Humber	43.680604	-79.482074	Coronation Park	43.684380	-79.480921	Park
5	Mount Olive	43.653482	-79.383935	Downtown Toronto	43.653232	-79.385296	Neighborhood
6	Mount Olive	43.653482	-79.383935	Nathan Phillips Square	43.652270	-79.383516	Plaza

Exhibit 7: *No. of Hot Breweries in each of the Neighbourhood:*

	Neighborhood	Neighborhood ID	Latitude	Longitude	No. of Hot Breweries
0	West Humber	1	43.680604	-79.482074	1.0
1	Mount Olive	2	43.653482	-79.383935	18.0
2	Silverstone	2	43.749751	-79.599116	NaN
3	Jamestown	2	43.653482	-79.383935	18.0
4	Thistletown	3	43.737266	-79.565317	1.0
5	Rexdale	4	43.721362	-79.565513	NaN
6	Kipling	4	43.637593	-79.535494	4.0

Exhibit 8: *Final Dataset for Analysis:*

	Neighbourhood	Neighbourhood Id	After-Tax Household Income	PFR Permits Issued	Crimes per week	Potential customers: 15 - 64 years	Potential customers: Employed	No. of Hot Breweries
0	West Humber-Clairville	1.0	59703.0	1385.0	14.500000	15695.0	10255.0	1.0
1	Mount Olive-Silverstone-Jamestown	2.0	46986.0	1799.0	10.000000	10760.0	3225.0	18.0
2	Thistletown-Beaumond Heights	3.0	57522.0	1191.0	2.423077	4410.0	2710.0	1.0
3	Rexdale-Kipling	4.0	51194.0	88.0	2.673077	4995.0	3310.0	4.0
4	Elms-Old Rexdale	5.0	49425.0	2388.0	2.730769	3580.0	1845.0	0.0
5	Kingsview Village-The Westway	6.0	50714.0	2.0	5.673077	7415.0	3505.0	1.0
6	Willowridge-Martingrove-Richview	7.0	57048.0	2711.0	5.596154	10200.0	7380.0	2.0

The final Dataset for Analysis was obtained by joining the dataset in Exhibit 2 with the one in Exhibit 7 using Outer (Full) join, using the “Neighbourhood ID” as the key.

All further analysis were conducted on the dataset shown in Exhibit 8.

NOTE: All datasets shown in exhibits are just a sample of the original datasets that were used.

Methodology:

Exploratory Analysis:

For the exploratory analysis, we use the dataset in Exhibit 8 and consider the “No. of Hot Breweries” as the target variable.

In the Boxplot for “No. of Hot Breweries” (Exhibit 9), extensive presence of outliers can be seen. The median value and the maximum values are shown to be far apart, indicating that the data is sharply skewed. Another point to be noted is that most of the Neighbourhoods have less that 13 cafes, even though a few neighbourhoods have more than 20 cafes.

Exhibit 9: *Distribution of Cafes by Neighborhood:*

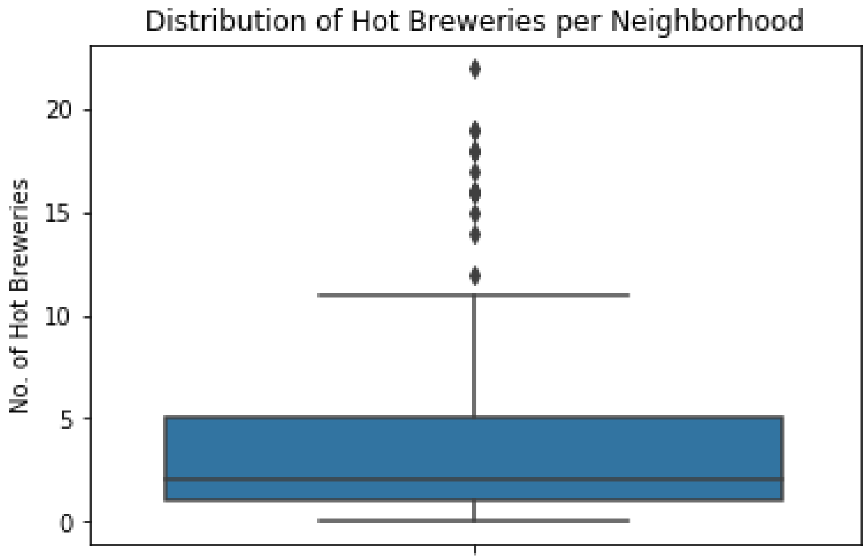
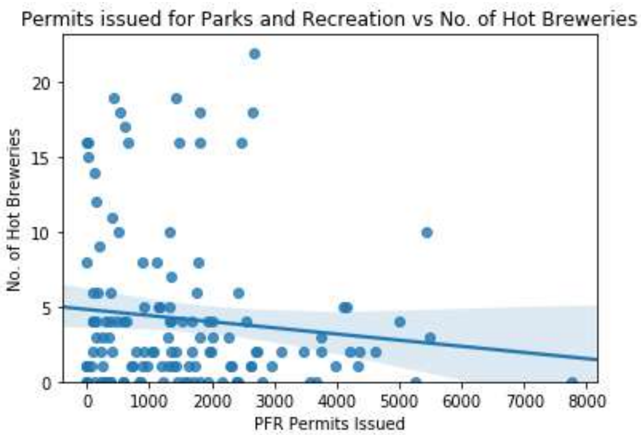


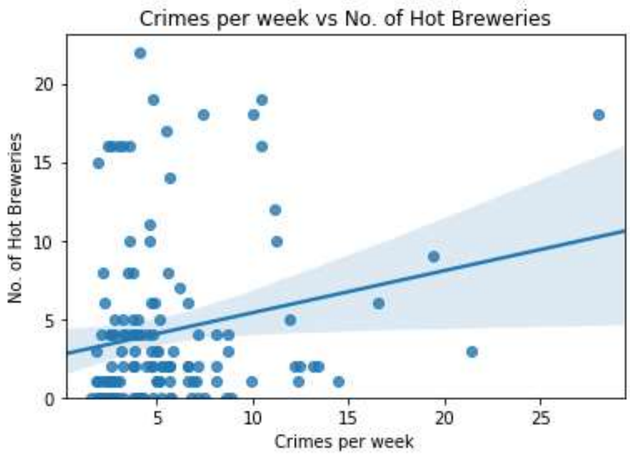
Exhibit 10: *PFR Permits Issued v/s No. of Cafes per Neighbourhood*



The fewer the no. of Permits the municipality issues for the construction of parks, forests and recreational centres, the higher the no. of cafes.

This inverse relation can be justified, since the people in these neighbourhoods have fewer options for places to go and relax, they will naturally have a higher tendency to visit cafes.

Exhibit 11: *Crimes per week v/s No. of Cafes per Neighbourhood*

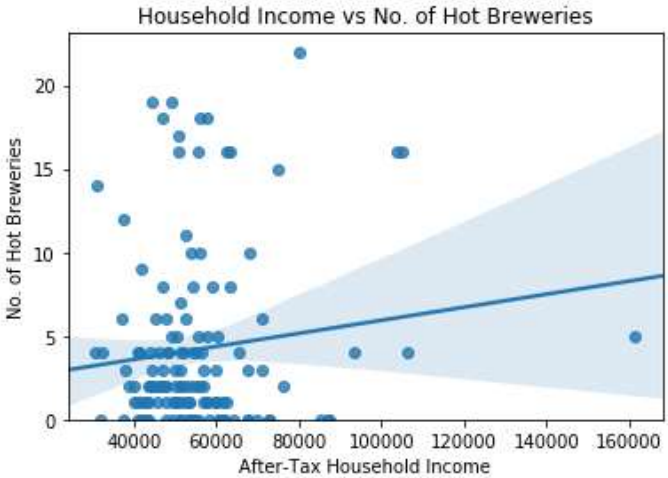


The more the number of cafes in a neighbourhood, the higher the crime rate.

This relation is actually very interesting, and a plausible explanation for this might be that people are usually more laid back and have their guards down when visiting cafes, thus making them more vulnerable for getting victimised.

Another probable explanation might be the fact that cafes are places of interaction for a lot of people, and these interactions have chances of turning ugly, resulting and a crime.

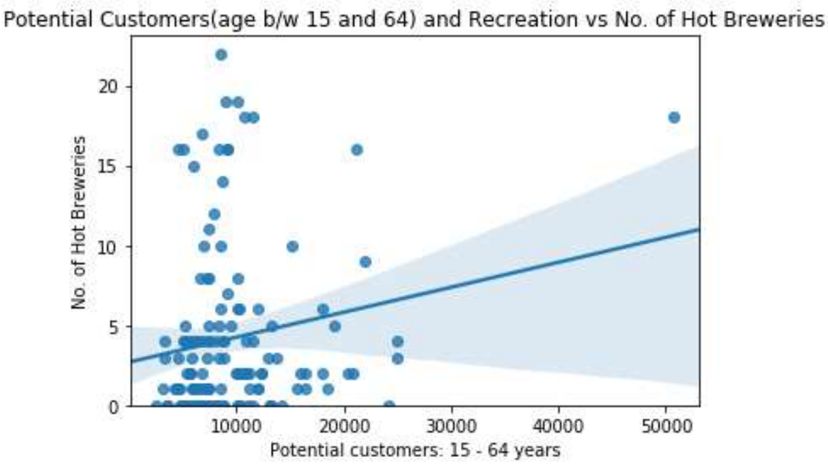
Exhibit 12: *Income v/s No. of Cafes per Neighbourhood.*



Higher the average income of the neighbourhood, more the number of cafes.

This is quite intuitive because coffee houses are usually moderately expensive places of recreation, attracting the affluent.

Exhibit 13: *Age v/s No. of Cafes in the Neighbourhood*



Since people below 15 are not generally attracted to coffee and people above 64 should not be drinking too much coffee (that is even if they can manage to make a trip down to the cafe), only people aged in between were considered and the results are as expected.

Higher the number of people in the particular demography, more the number of cafes in the neighbourhood.

Exhibit 14: *Correlation between the variables:*

	Neighbourhood Id	After-Tax Household Income	PFR Permits Issued	Crimes per week	Potential customers: 15 - 64 years	Potential customers: Employed	No. of Hot Breweries
Neighbourhood Id	1.000000	-0.063650	-0.024293	0.059116	0.097547	0.064290	-0.040797
After-Tax Household Income	-0.063650	1.000000	0.204600	-0.260229	-0.005475	0.104423	0.121016
PFR Permits Issued	-0.024293	0.204600	1.000000	0.126513	0.319577	0.272574	-0.112529
Crimes per week	0.059116	-0.260229	0.126513	1.000000	0.716731	0.595883	0.204758
Potential customers: 15 - 64 years	0.097547	-0.005475	0.319577	0.716731	1.000000	0.951749	0.172462
Potential customers: Employed	0.064290	0.104423	0.272574	0.595883	0.951749	1.000000	0.194100
No. of Hot Breweries	-0.040797	0.121016	-0.112529	0.204758	0.172462	0.194100	1.000000

It can be seen from Exhibit 14 that the target variable, “No. of Hot Breweries” has the strongest positive correlations with “Crimes per week” (0.204), “Potential Customers Employed” (0.194) and “Potential Customers: 15 – 64 years” (0.172). Strongest negative correlation is between the target variable and “PFR Permits Issued” (-112).

Modelling:

In this study, a Decision Tree Classifier will be used to classify each neighbourhood as “GOOD”, “AVERAGE” and “BAD” based in its Neighbourhood ID. Since the dataset being used does not has a categorical variable to classify the data, a new variable has to be conceptualised to train the Decision Tree model

For creating the categorical variable, a new feature, “Normalised Decision Metric” has been created as follows:

- ➔ NDS=(I+P-C)/Prm

NDS: Normalised Decision Metric

I: Avg. Household Income (Directly proportional to target, higher is better)

P: Potential customers between 15 and 64 years of age (Directly proportional to target, higher is better)

C: Crimes per week (Directly proportional to target, lower is better)

Prm: PRM permits issued (Inversely proportional to target)

Once the values of the “Normalised Decision Metric” were obtained, they were binned into 3 categories to classify each of the neighbourhoods as “GOOD”, “AVERAGE” and “BAD”. The result of the classification has been stored in the feature “Verdict”

Exhibit 15: *Final Dataset for Modelling:*

	Neighbourhood	Neighbourhood Id	After-Tax Household Income	PFR Permits Issued	Crimes per week	Potential customers: 15 - 64 years	Potential customers: Employed	No. of Hot Breweries	Normalized Decision Metric	Verdict
0	West Humber-Clairville	1.0	59703.0	1385.0	14.500000	15695.0	10255.0	1.0	0.077503	AVERAGE
1	Mount Olive-Silverstone-Jamestown	2.0	46986.0	1799.0	10.000000	10760.0	3225.0	18.0	0.038302	BAD
2	Thistletown-Beaumond Heights	3.0	57522.0	1191.0	2.423077	4410.0	2710.0	1.0	0.073237	AVERAGE
3	Rexdale-Kipling	4.0	51194.0	88.0	2.673077	4995.0	3310.0	4.0	1.102573	GOOD
4	Elms-Old Rexdale	5.0	49425.0	2388.0	2.730769	3580.0	1845.0	0.0	0.020930	BAD
5	Kingsview Village-The Westway	6.0	50714.0	2.0	5.873077	7415.0	3505.0	1.0	50.988087	GOOD
6	Willowridge-Martingrove-Richview	7.0	57048.0	2711.0	5.596154	10200.0	7380.0	2.0	0.025508	BAD

The features mentioned in Exhibit 16 were used to train the Decision Tree classifier. The entropy level of the Decision Tree was limited to 4 and 70% of the data was used for training and the remaining 30% was used for testing the dataset.

Exhibit 16: *Feature set for Training the Decision Tree Model:*

	After-Tax Household Income	PFR Permits Issued	Crimes per week	Potential customers: 15 - 64 years	No. of Hot Breweries
0	59703.0	1385.0	14.500000	15695.0	1.0
1	46986.0	1799.0	10.000000	10760.0	18.0
2	57522.0	1191.0	2.423077	4410.0	1.0
3	51194.0	88.0	2.673077	4995.0	4.0
4	49425.0	2388.0	2.730769	3580.0	0.0

Evaluation:

The resultant model had the following out-of-sample evaluation metric scores based on the values predicted using the test dataset:

- F1 score: 0.853
- Jaccard Similarity Index: 60.975%

Results:

Classification Results:

Exhibit 17 shows the predicted classification results as well as the original verdict (based on the Normalised Decision Metrics).

Exhibit 17: *Classification Results:*

	Neighbourhood	Neighbourhood Id	After-Tax Household Income	PFR Permits Issued	Crimes per week	Potential customers: 15 - 64 years	Potential customers: Employed	No. of Hot Breweries	Normalized Decision Metric	Verdict	Classified As
0	West Humber-Clairville	1.0	59703.0	1385.0	14.500000	15695.0	10255.0	1.0	0.077503	AVERAGE	AVERAGE
1	Mount Olive-Silverstone-Jamestown	2.0	46986.0	1799.0	10.000000	10760.0	3225.0	18.0	0.038302	BAD	BAD
2	Thistletown-Beaumond Heights	3.0	57522.0	1191.0	2.423077	4410.0	2710.0	1.0	0.073237	AVERAGE	AVERAGE
3	Rexdale-Kipling	4.0	51194.0	88.0	2.673077	4995.0	3310.0	4.0	1.102573	GOOD	GOOD
4	Elms-Old Rexdale	5.0	49425.0	2388.0	2.730769	3580.0	1845.0	0.0	0.020930	BAD	BAD
5	Kingsview Village-The Westway	6.0	50714.0	2.0	5.673077	7415.0	3505.0	1.0	50.988087	GOOD	GOOD
6	Willowridge-Martingrove-Richview	7.0	57048.0	2711.0	5.596154	10200.0	7380.0	2.0	0.025508	BAD	BAD

The following exhibits shows the measure of central tendencies and basic statistical descriptions of the three categories, GOOD, AVERAGE and BAD.

Exhibit 18: *Characteristics of Neighbourhoods that are GOOD for opening a Café:*

	Neighbourhood Id	After-Tax Household Income	PFR Permits Issued	Crimes per week	Potential customers: 15 - 64 years	Potential customers: Employed	No. of Hot Breweries	Normalized Decision Metric
count	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000
mean	72.530612	52524.959184	258.326531	4.978022	7601.734694	5375.000000	5.061224	8.943042
std	35.443089	12052.593720	206.019658	4.311999	3568.920610	2982.324841	5.921037	21.151447
min	4.000000	31304.000000	1.000000	1.615385	2570.000000	260.000000	0.000000	0.113664
25%	57.000000	45058.000000	90.000000	2.423077	5420.000000	3675.000000	0.000000	0.215106
50%	76.000000	51381.000000	243.000000	3.557692	6850.000000	4925.000000	3.000000	0.442106
75%	97.000000	55536.000000	374.000000	5.096154	8790.000000	6700.000000	8.000000	1.102573
max	138.000000	93391.000000	713.000000	21.442308	21920.000000	16695.000000	19.000000	100.000000

Exhibit 19: *Characteristics of Neighbourhoods that are AVERAGE for opening a Café:*

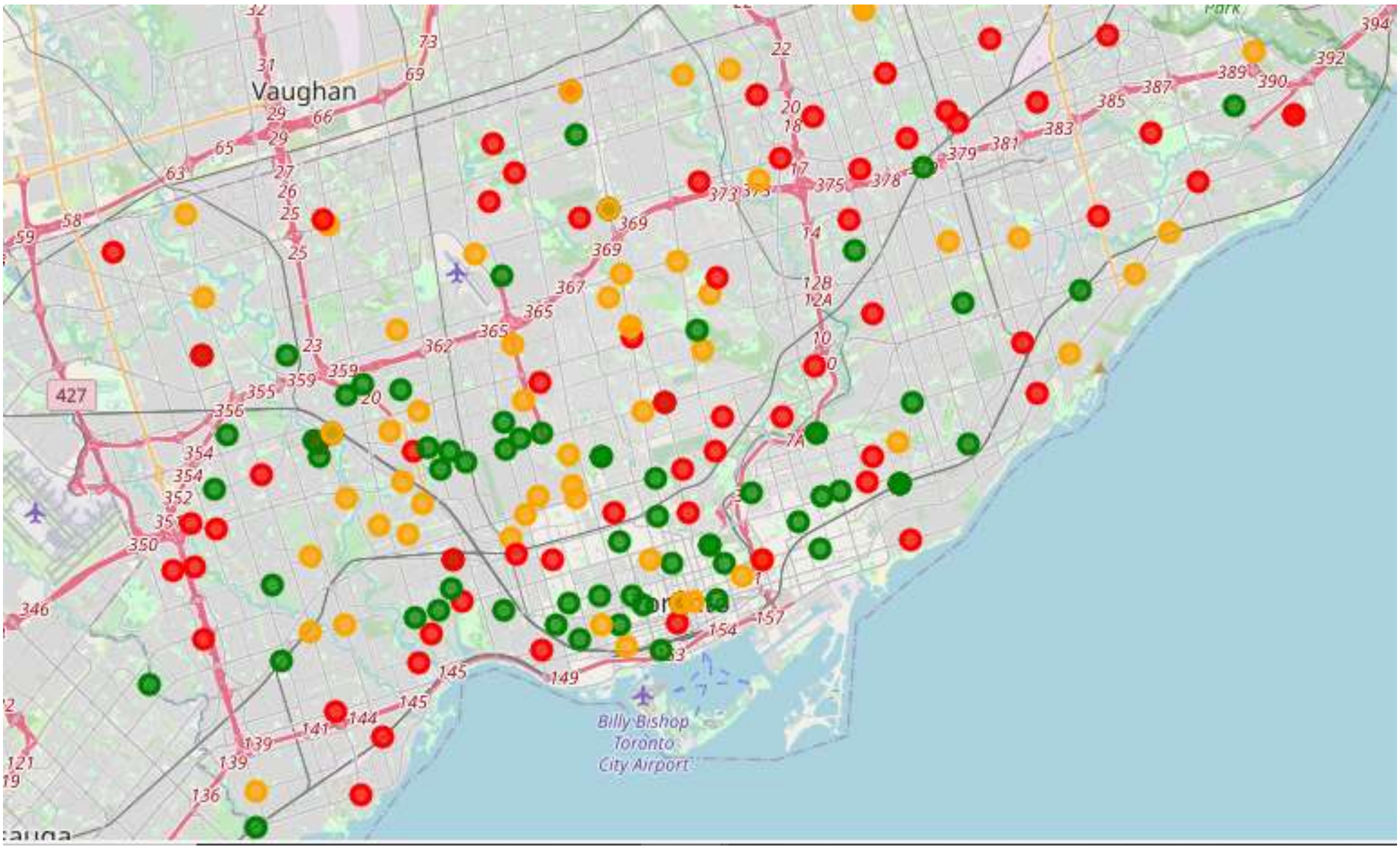
	Neighbourhood Id	After-Tax Household Income	PFR Permits Issued	Crimes per week	Potential customers: 15 - 64 years	Potential customers: Employed	No. of Hot Breweries	Normalized Decision Metric
count	42.000000	42.000000	42.000000	42.000000	42.000000	42.000000	42.000000	42.000000
mean	72.500000	59925.023810	1391.214286	5.803114	10478.571429	7211.666667	4.166667	0.076294
std	43.002694	22838.181135	598.281585	4.689394	8297.452980	7152.246080	4.948105	0.026478
min	1.000000	30794.000000	590.000000	2.019231	3290.000000	510.000000	0.000000	0.033136
25%	33.000000	47637.250000	1056.500000	2.846154	6270.000000	3736.250000	1.000000	0.053884
50%	74.500000	55609.500000	1323.500000	4.144231	8050.000000	5757.500000	2.000000	0.074189
75%	110.000000	61637.500000	1547.000000	7.139423	10305.000000	7317.500000	5.000000	0.092806
max	140.000000	161448.000000	4105.000000	28.057692	50645.000000	44110.000000	19.000000	0.131044

Exhibit 20: *Characteristics of Neighbourhoods that are BAD for opening a Café:*

	Neighbourhood Id	After-Tax Household Income	PFR Permits Issued	Crimes per week	Potential customers: 15 - 64 years	Potential customers: Employed	No. of Hot Breweries	Normalized Decision Metric
count	45.000000	45.000000	45.000000	45.000000	45.000000	45.000000	45.000000	45.000000
mean	65.600000	54338.400000	3072.955556	5.850427	11089.266667	7391.155556	3.266667	0.024090
std	43.809712	11984.226371	1321.992091	2.907904	4304.028754	3531.268391	4.750120	0.014485
min	2.000000	32539.000000	1342.000000	1.826923	3580.000000	1360.000000	0.000000	0.000000
25%	34.000000	46803.000000	2025.000000	4.000000	7495.000000	5400.000000	1.000000	0.013861
50%	52.000000	51247.000000	2636.000000	5.173077	10830.000000	6865.000000	2.000000	0.023618
75%	104.000000	60065.000000	3748.000000	6.846154	13035.000000	9750.000000	4.000000	0.031317
max	137.000000	86816.000000	7770.000000	13.173077	21235.000000	16770.000000	22.000000	0.066989

The above exhibits show that the metrics for 'After-Tax Household Income' is maximum for AVERAGE followed by GOOD and then BAD. GOOD neighbourhoods have the least 'Crimes per week' statistics, followed by AVERAGE and BAD which are very close to each other in this respect. 'Potential Customers' are best located in AVERAGE neighbourhoods, while GOOD neighbourhoods generally have the least of them. BAD locations have the least 'No. of Hot Breweries' while GOOD locations have the highest of this feature. GOOD Neighbourhoods have the least no. of 'PFR Permits Issues' while the BAD ones have the least.

Exhibit 21: *Plotting the categories on the map:*



Discussions:

Observations:

From the results of the above analysis, it can be seen that the best places to open a Café or a Coffee house in Toronto ideally have a low crime rate, so that the patrons visiting the place have a sense of security. The Café could be located in a neighbourhood, whose residents are moderately affluent, not too rich, neither too poor and most importantly, the Coffee house has to be in a place where there are very few parks and other types of recreational centres for the residents to go to. Also, the Café should not be located in a place which is too crowded.

Future Scope:

This study could be further expanded by including a plethora of other features and using various other classification and clustering techniques. For instances, diving deeper into the Age group of the potential customers can provide some really valuable insights, while factors such as commercial property rates in the neighbourhood can greatly enhance the comprehensiveness of the study.

The analysis done here can also be replicated for showrooms, restaurants and other types of commercial establishments. A similar methodology can be used for analysing other geographical places.

Conclusion:

An analysis involving various socio-economic and geographical features of the neighbourhoods of Toronto was carried out and a metric was evaluated to rate the respective neighbourhoods for its suitability to open a Café. The Neighbourhoods were then classified into GOOD, AVERAGE and BAD based on the metric.