Exploring Toronto Neighborhoods to open a Pizzeria

Using Machine Learning in Python to shortlist potential Neighbourhoods

Table of Contents

Introduction (Description of the Business Problem)

- As a part of the IBM Data Science professional program Capstone Project, I worked on the real datasets to get an experience of what a data scientist goes through in real life. Toronto is the capital city of the Canadian province of Ontario. With a recorded population of 2,731,571 in 2016, it is the most populous city in Canada and the fourth most populous city in North America.

- In this project I will go through step by step process to make a decision whether it is a good idea to open an Pizzeria. I analyze the neighborhoods in Toronto to identify the most profitable area since the success of the restaurant depends on availability of a niche in the market in terms of competition

- Target Audience

  - Business personnel who wants to invest or open Pizzeria in Toronto. This analysis will be a comprehensive guide to start or expand restaurants targeting this space.

  - Freelancer who loves to have their own restaurant as a side business. This analysis will give an idea, how beneficial it is to open a restaurant and what are the pros and cons of this business.

  - Business Analyst or Data Scientists, who wish to analyze the neighborhoods of Toronto using Exploratory Data Analysis and other statistical & machine learning techniques to obtain all the necessary data, perform some operations on it and, finally be able to tell a story out of it.

Data acquisition and cleaning : Data Sources

a) I'm using "List of Postal code of Canada: M" (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) wiki page to get all the information about the neighborhoods present in Toronto. This page has the postal code, borough & the name of all the neighborhoods present in Toronto.

b) Then I'm using "https://cocl.us/Geospatial_data" csv file to get all the geographical coordinates of the neighborhoods.

c) To get location and other information about various venues in Toronto I'm using Foursquare's explore API. Using the Foursquare's explore API (which gives venues recommendations), I'm fetching details about the venues up present in Toronto and collected their names, categories and locations (latitude and longitude).

d) Identify neighbourhoods with lesser competition
Using the foursquare data, we will be taking a count of venue locations marked as "Pizza Place", "Italian Restaurant". This will help us identify in which neighbourhoods there is currently a niche in the market that we can take advantage of. The neighbourhoods where the count is zero for these types of venues will be identified as potential candidates for us.

Data acquisition and cleaning : Data Sources

a) I'm using "List of Postal code of Canada: M" (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) wiki page to get all the information about the neighborhoods present in Toronto. This page has the postal code, borough & the name of all the neighborhoods present in Toronto.

b) Then I'm using "https://cocl.us/Geospatial_data" csv file to get all the geographical coordinates of the neighborhoods.

c) To get location and other information about various venues in Toronto I'm using Foursquare's explore API. Using the Foursquare's explore API (which gives venues recommendations), I'm fetching details about the venues up present in Toronto and collected their names, categories and locations (latitude and longitude).

d) Identify neighbourhoods with lesser competition
Using the foursquare data, we will be taking a count of venue locations marked as "Pizza Place", "Italian Restaurant". This will help us identify in which neighbourhoods there is currently a niche in the market that we can take advantage of.

e) Highlight the potential candidates for Neighbourhoods with several eating options, highlighting that it could be a popular destination

Data acquisition and cleaning : Final Cleaned Data

	Postal Code	Borough	Neighbourhood	Latitude	Longitude	Count of Competit	Count of Eateries
0	M3A	North York	Parkwoods	43.75326	-79.32966	0	1
1	M4A	North York	Victoria Village	43.72588	-79.31557	0	2
2	M5A	Downtown Toronto	Regent Park, Har...	43.65426	-79.36064	0	25
3	M6A	North York	Lawrence Manor,...	43.71852	-79.46476	0	2
4	M7A	Downtown Toronto	Queen's Park, O...	43.6623	-79.38949	1	22
5	M9A	Etobicoke	Islington Avenue,...	43.66786	-79.53224	1	1
6	M1B	Scarborough	Malvern, Rouge	43.80669	-79.19435	0	1
7	M3B	North York	Don Mills	43.74591	-79.35219	1	13
8	M4B	East York	Parkview Hill, W...	43.7064	-79.30994	2	3
9	M5B	Downtown Toronto	Garden District, ...	43.65716	-79.37894	4	48
10	M6B	North York	Glencairn	43.70958	-79.44507	1	5
11	M9B	Etobicoke	West Deane Par...	43.65094	-79.55472	0	0
12	M1C	Scarborough	Rouge Hill, Port ...	43.78454	-79.1605	0	1
13	M3C	North York	Don Mills	43.7259	-79.34092	1	13
14	M4C	East York	Woodbine Heights	43.69534	-79.31839	0	0

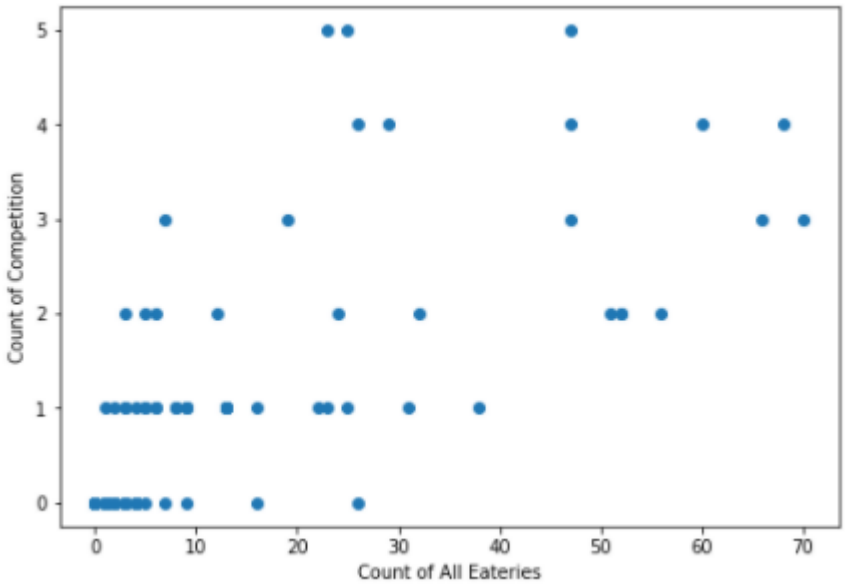
Exploratory Data Analysis

- Implications : Most of the venues have very low number of eateries in total, which may be perhaps be reflective of the “sparseness” of the population levels living in those neighbourhoods.

- Also, we see that on average, each location contains 1 competition outlet and 12 eateries. Which means that Pizzerias / Italian restaurants represent 8% of all Total eateries.

- We generally see that locations with a higher number of Competition outlets [Pizzerias / Italian Restaurants] generally have a larger number of eateries. The suitable locations for our Pizzerias will be ones that have a higher presence of eating outlets but those that still do not have as many Competition outlets i.e. a sweetspot

	Latitude	Longitude	Count of Competition	Count of Eateries
count	103.000000	103.000000	103.000000	103.000000
mean	43.704608	-79.397153	0.922330	12.077670
std	0.052463	0.097146	1.318714	17.723746
min	43.602414	-79.615819	0.000000	0.000000
25%	43.660567	-79.464763	0.000000	1.000000
50%	43.696948	-79.388790	0.000000	4.000000
75%	43.745320	-79.340923	1.000000	14.500000
max	43.836125	-79.160497	5.000000	70.000000



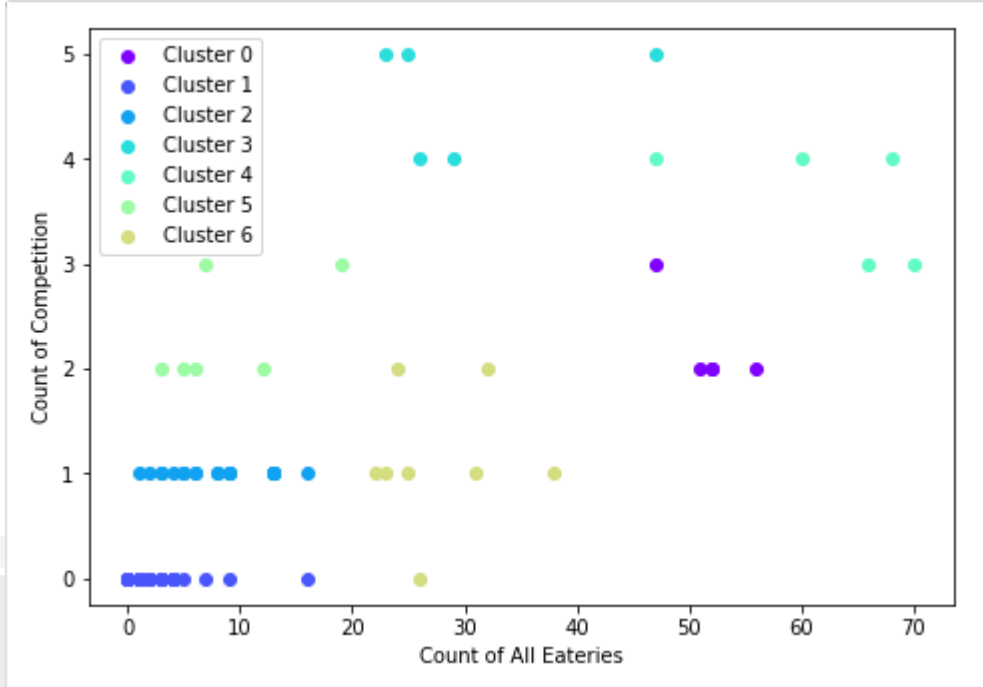
Predictive Modelling

- Neighbourhoods with low number of eateries : Cluster 1,2,5
  Cluster 1: Contains the most number of locations in Toronto, but the average number of eateries in those locations is very low. Likely reflective of the size of the populations in those locations. While cluster 2 and 8 are better off, on average they contain less than 10 eateries, which again represents a low business opportunity.

- Neighbourhoods with higher level of competition : Cluster 3,4.
  Cluster 3 particularly his is the cluster that contains the most number of competition outliers – with an average of 4.5 per location. While Cluster 4 has a better ratio of competition outlets to eateries (5.8%), each location already contains a minimum number of 3 competition outlets. Which is not ideal for us.

- Optimal Neighbourhoods with high level of eateries and low competition : Cluster 0,6
  In both of these clusters, the Avg # of Competition outlets / Avg # of Total Eateries is 4%. Which is half the overall average of 8%. So quite optimal.

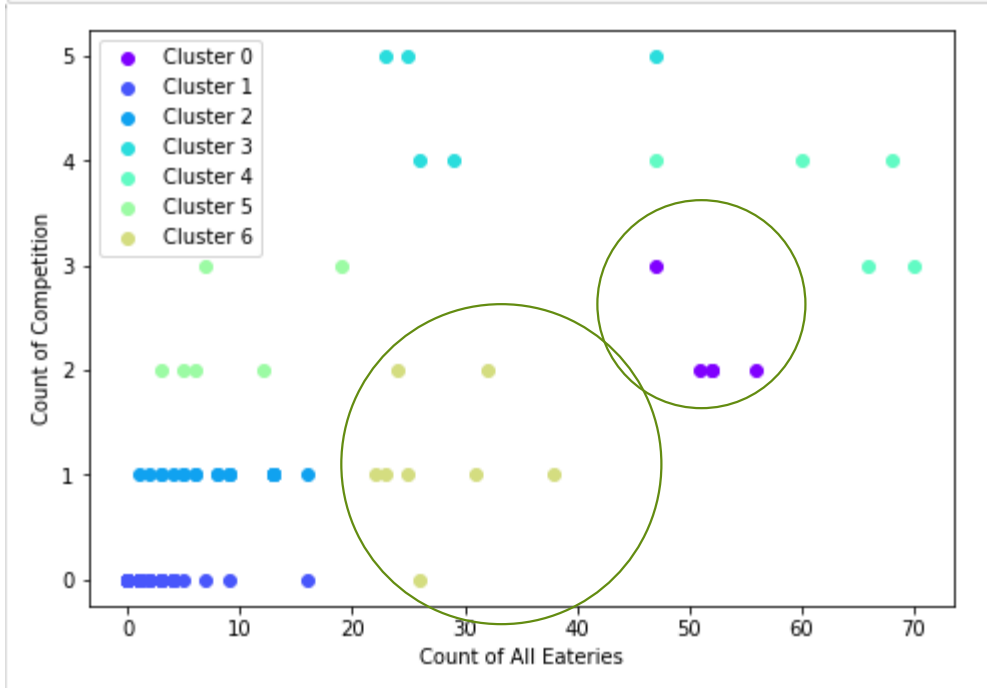
	Count of Competition				Count of Eateries			
	mean	count	min	max	mean	count	min	max
Cluster Labels								
0	2.200000	5	2.0	3.0	51.600000	5	47.0	56.0
1	0.000000	54	0.0	0.0	1.722222	54	0.0	16.0
2	1.000000	20	1.0	1.0	7.950000	20	1.0	16.0
3	4.600000	5	4.0	5.0	30.000000	5	23.0	47.0
4	3.600000	5	3.0	4.0	62.200000	5	47.0	70.0
5	2.333333	6	2.0	3.0	8.666667	6	3.0	19.0
6	1.125000	8	0.0	2.0	27.625000	8	22.0	38.0



Predictive Modelling : Optimal Clusters

- Optimal Neighbourhoods with high level of eateries and low competition : Cluster 0,6
  In both of these clusters, the Avg # of Competition outlets / Avg # of Total Eateries is 4%. Which is half the overall average of 8%. So quite optimal.

- Both these clusters represent good potential. The major difference in both these clusters is the Total number of overall eateries present in those sets of clusters. Cluster 0 avg is 52 eateries and Cluster 6 average is 28 eateries. So the average for both of these is much higher than overall which is quite good.



Predictive Modelling : Optimal Clusters

The Optimum Locations Selected to Open the Pizzeria

```
Toronto_merged3.loc[Toronto_merged3['Cluster Labels'] == 0]
```

	Postal Code	Borough	Neighbourhood	Latitude	Longitude	Count of Competition	Count of Eateries	Cluster Labels
15	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418	2.0	52.0	0
24	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383	3.0	47.0	0
30	M5H	Downtown Toronto	Richmond, Adelaide, King	43.650571	-79.384568	2.0	56.0	0
84	M5T	Downtown Toronto	Kensington Market, Chinatown, Grange Park	43.653206	-79.400049	2.0	51.0	0
99	M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160	2.0	52.0	0

```
Toronto_merged3.loc[Toronto_merged3['Cluster Labels'] == 6]
```

	Postal Code	Borough	Neighbourhood	Latitude	Longitude	Count of Competition	Count of Eateries	Cluster Labels
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	0.0	26.0	6
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	1.0	22.0	6
20	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306	1.0	38.0	6
33	M2J	North York	Fairview, Henry Farm, Oriole	43.778517	-79.346556	1.0	31.0	6
37	M6J	West Toronto	Little Portugal, Trinity	43.647927	-79.419750	2.0	32.0	6
54	M4M	East Toronto	Studio District	43.659526	-79.340923	1.0	23.0	6
59	M2N	North York	Willowdale, Willowdale East	43.770120	-79.408493	2.0	24.0	6
80	M5S	Downtown Toronto	University of Toronto, Harbord	43.662696	-79.400049	1.0	25.0	6

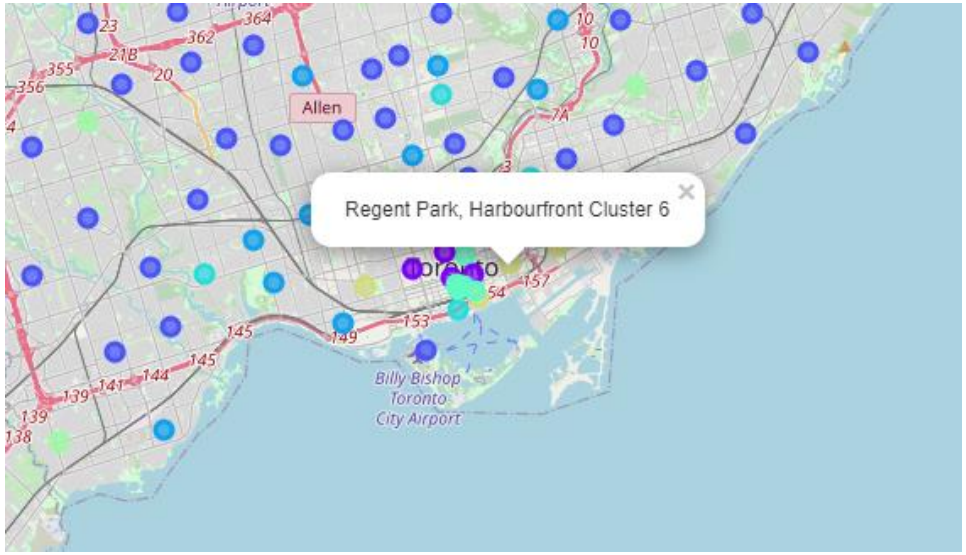
Results and Discussion

- Assuming that the number of eateries in a location is indicative of population, the lower number of eateries might suggest a smaller size of a outlet that we might have to open. Thus also reflective of a lower investment cost. So cluster 6 might be better in that regard.

- After careful consideration of Cluster 6 locations, it maybe a good idea to open a new Pizzeria in Regent Park, Harbourfront since it has high number of Eateries and not a single competition outlet, which likely gives it a greater probability of success. We can then potentially expand into other neighbourhoods mentioned as potential within the optimal cluster selections.

- The drawbacks of this analysis are — the clustering is completely based only on data obtained from Foursquare API and the data about the restaurants in each neighborhood which may not be not up-to date.

- We are also not looking into several other dimensions of determining the selection of location i.e. availability of sourcing, local legislation, land cost etc. Which can be incorporated after discussions with Stakeholders



	Postal Code	Borough	Neighbourhood	Latitude	Longitude	Count of Competition	Count of Eateries	Cluster Labels
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	0.0	26.0	6