Battle of the Neighborhoods

Applied Data Science Capstone - Final Report

Introduction

In this project we will try to find an optimal location for a restaurant. Specifically, this report will be targeted to stakeholders interested in opening an Italian restaurant in Toronto.

Since there are lots of restaurants in Toronto we will try to get locations that are not already crowded with restaurants. We are also particularly interested in areas with no Italian restaurants in vicinity.

We will use our data science powers to generate a few most promising neighborhoods based on this criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

Interest

This project would be of great interest to any entrepreneur looking to open up a taco restaurant in Toronto. The same methods could also be used for entrepreneurs looking to open up businesses of any variety, by making just a few necessary tweaks to fit the category.

Datas

Data Sources

Given the business problem stated above, the factors that would determine where an ideal place to open up a taco restaurant would be the number/location of existing taco restaurants in 2 of 10

Toronto, as well as the number/location of existing office buildings in Toronto. Therefore, the following data sources will be used for this project:

- Postal code data will be sourced from the following Wikipedia page:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

- Geographical coordinates will be sourced from the following .csv link:
http://cocl.us/Geospatial_data

- taco restaurant and office building data will be sourced from Foursquare API

Data Cleaning

The data obtained from the multiple sources listed above was combined into a single table. It was necessary to clean the data as well, in order to get an accurate analysis.

First, the neighborhoods were grouped together by postal code, and any rows lacking a neighborhood entry used the borough in its place. This ensured that the project would have a reasonable amount of locations to work with, and the location data obtained for the postal codes ensured that it would be possible to plot their locations on a map.

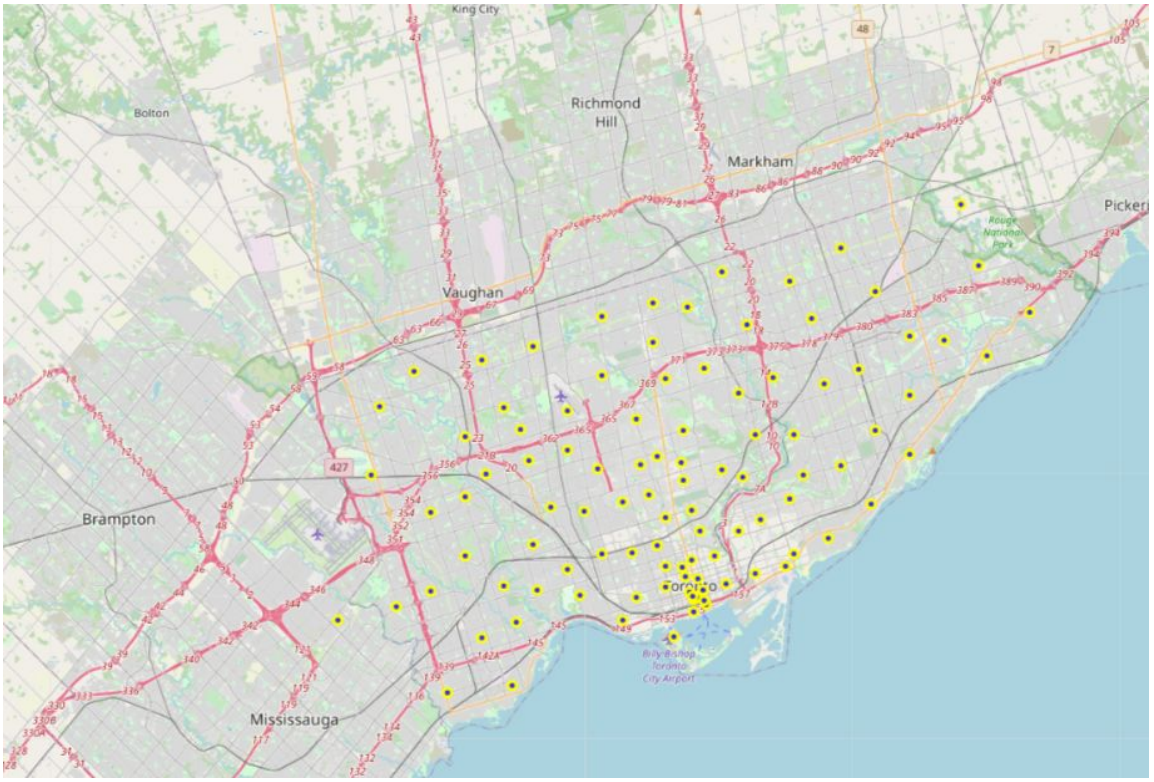
Next, the venues that appeared upon searching for “taco” needed to be filtered down and remove fast food restaurants like KFC.

After getting restaurants I prepared another list for the count of offices and schools per location and marged and cleaned the list from null values. At finally add success rate filed to show customer count per restaurant and normalize it.

Methodology

Coordinates of possible locations in Toronto

First, it was necessary to get possible locations in Toronto and their geographical coordinates. The postal coded data was taken from the Wikipedia link and the geographical coordinates were taken from the .csv file.



Count of taco restaurants for the locations.

The data for existing taco restaurants and office buildings was obtained using the Foursquare API. For each location, a simple search was run for “taco,” with the venues filtered by category “taco restaurant” to ensure that any other business types were not included.

Neighborhood	Restaurant Count
Central Bay Street	4
Church and Wellesley	3
Commerce Court, Victoria Hotel	3
Dufferin, Dovercourt Village	1
East Toronto, Broadview North (Old East York)	1
Fairview, Henry Farm, Oriole	1
First Canadian Place, Underground city	3
Garden District, Ryerson	4
Harbourfront East, Union Station, Toronto Islands	2
Kensington Market, Chinatown, Grange Park	5
Queen's Park, Ontario Provincial Government	1
Richmond, Adelaide, King	4
South Steeles, Silverstone, Humbergate, Jamestown, Mount Olive, Beaumond Heights, Thistletown, Albion Gardens	1
St. James Town	3
Stn A PO Boxes	1
Studio District	1
Toronto Dominion Centre, Design Exchange	1

Count of offices and schools for the locations.

Likewise, a simple search was run for “office,” with the venues filtered by category “office” and “school” to ensure that any other business types were not included. The important data in this case was the number of taco restaurants and/or offices in each location, not the specific venues themselves.

Neighborhood	Office Count
Alderwood, Long Branch	4
Bathurst Manor, Wilson Heights, Downsview North	3
Bedford Park, Lawrence Manor East	4
Berczy Park	50
Birch Cliff, Cliffside West	1
...	...
Weston	1
Wexford, Maryvale	1
Willowdale, Willowdale East	13
Woodbine Heights	2
York Mills West	1

Neighborhood	School Count
Agincourt	1
Alderwood, Long Branch	2
Bathurst Manor, Wilson Heights, Downsview North	1
Bayview Village	2
Bedford Park, Lawrence Manor East	2
...	...
Willowdale, Willowdale West	2
Woburn	1
Woodbine Heights	3
York Mills West	2
York Mills, Silver Hills	4

Merged and cleaned datas.

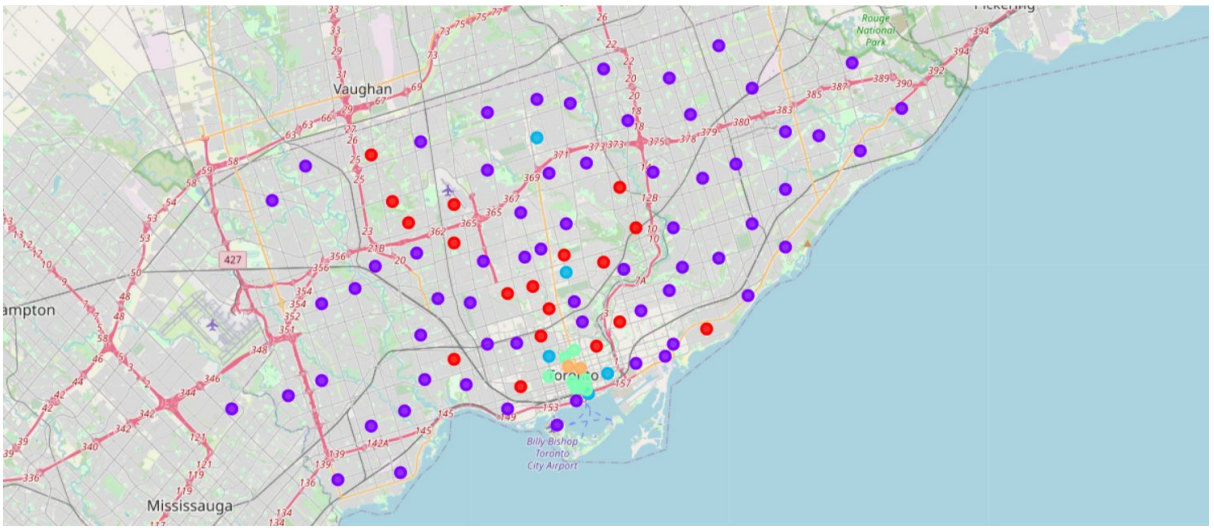
Therefore, the data frames were altered to include only the number of existing coffee shops and offices in each location. After the result I add two column population and success rate per restaurant .These data frames were then combined into a single table, a portion of which is shown in the following figure:

	Restaurant Count	Office Count	School Count	Population	success rate	Neighborhood	Postal Code	Borough	Neighbourhood	Latitude	Longitud
0	0.0	23.0	13.0	283.0	283.0	University of Toronto, Harbord	M3A	North York	Parkwoods	43.753259	-79.32965
1	0.0	10.0	12.0	250.0	250.0	Regent Park, Harbourfront	M4A	North York	Victoria Village	43.725882	-79.31557
2	0.0	13.0	10.0	213.0	213.0	Willowdale, Willowdale East	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.36063
3	0.0	15.0	9.0	195.0	195.0	Davisville	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.46476
4	0.0	50.0	7.0	190.0	190.0	Berczy Park	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.38949
...	...	...	...	...	...	...	...	...	...	...	...
89	0.0	1.0	0.0	1.0	1.0	Dorset Park, Wexford Heights, Scarborough Town...	M9V	Etobicoke	South Steeles, Silverstone, Humbergate, Jamest...	43.739416	-79.58843
90	0.0	1.0	0.0	1.0	1.0	Business reply mail Processing Centre, South C...	M1W	Scarborough	Steeles West, L'Amoreaux West	43.799525	-79.31838
91	0.0	1.0	0.0	1.0	1.0	Glencairn	M4W	Downtown Toronto	Rosedale	43.679563	-79.37752
92	0.0	1.0	0.0	1.0	1.0	Rouge Hill, Port Union, Highland Creek	M5W	Downtown Toronto	Stn A PO Boxes	43.646435	-79.37484
93	0.0	1.0	0.0	1.0	1.0	Wexford, Maryvale	M8W	Etobicoke	Alderwood, Long Branch	43.602414	-79.54348

94 rows × 11 columns

Find the suitable places with help of k-means Clustering

K-means clustering was used to group the locations into five clusters, based on only the number of coffee shops and office buildings. The different clusters were displayed on a map of Toronto, as shown in the following figure:



Each of these clusters were then examined to see how they were differentiated from one another, based on the number of coffee shops and office buildings.

The 1’st Cluster

	Cluster Labels	Restaurant Count	Office Count	School Count	Population	success rate	Neighbourhood	Postal Code	Borough	Latitude	Longitude
5	0	0.0	5.0	8.0	165.0	165.0	Little Portugal, Trinity	M6J	West Toronto	43.647927	-79.419750
6	0	0.0	13.0	7.0	153.0	153.0	Don Mills	M3B	North York	43.745906	-79.352188
7	0	0.0	13.0	7.0	153.0	153.0	Don Mills	M3C	North York	43.725900	-79.340923
8	0	0.0	8.0	7.0	148.0	148.0	The Danforth West, Riverdale	M4K	East Toronto	43.679557	-79.352188
9	0	0.0	3.0	7.0	143.0	143.0	The Annex, North Midtown, Yorkville	M5R	Central Toronto	43.672710	-79.405678
10	0	0.0	1.0	7.0	141.0	141.0	The Beaches	M4E	East Toronto	43.676357	-79.293031
11	0	0.0	0.0	7.0	140.0	140.0	Downsview	M3K	North York	43.737473	-79.464763
12	0	0.0	0.0	7.0	140.0	140.0	Downsview	M3L	North York	43.739015	-79.506944
13	0	0.0	0.0	7.0	140.0	140.0	Downsview	M3M	North York	43.728496	-79.495697
14	0	0.0	0.0	7.0	140.0	140.0	Downsview	M3N	North York	43.761631	-79.520999
15	0	0.0	9.0	6.0	129.0	129.0	St. James Town, Cabbagetown	M4X	Downtown Toronto	43.667967	-79.367675
17	0	0.0	5.0	6.0	125.0	125.0	Leaside	M4G	East York	43.709060	-79.363452
21	0	0.0	13.0	5.0	113.0	113.0	Davisville North	M4P	Central Toronto	43.712751	-79.390197
22	0	0.0	3.0	5.0	103.0	103.0	High Park, The Junction South	M6P	West Toronto	43.661608	-79.464763
24	0	0.0	1.0	5.0	101.0	101.0	Humewood-Cedarvale	M6C	York	43.693781	-79.428191
25	0	0.0	13.0	4.0	93.0	93.0	Summerhill West, Rathnelly, South Hill, Forest...	M4V	Central Toronto	43.686412	-79.400049
26	0	0.0	5.0	4.0	85.0	85.0	Forest Hill North & West, Forest Hill Road Park	M5P	Central Toronto	43.696948	-79.411307
27	0	0.0	5.0	4.0	85.0	85.0	Lawrence Manor, Lawrence Heights	M6A	North York	43.718518	-79.464763

The 2’th Cluster

	Cluster Labels	Restaurant Count	Office Count	School Count	Population	success rate	Neighbourhood	Postal Code	Borough	Latitude	Longitude
29	1	0.0	0.0	4.0	80.0	80.0	Lawrence Park	M4N	Central Toronto	43.728020	-79.388790
30	1	0.0	0.0	4.0	80.0	80.0	York Mills, Silver Hills	M2L	North York	43.757490	-79.374714
31	1	0.0	11.0	3.0	71.0	71.0	Brockton, Parkdale Village, Exhibition Place	M6K	West Toronto	43.636847	-79.428191
32	1	0.0	5.0	3.0	65.0	65.0	Christie	M6G	Downtown Toronto	43.669542	-79.422564
33	1	0.0	4.0	3.0	64.0	64.0	Runnymede, Swansea	M6S	West Toronto	43.651571	-79.484450
...	...	...	...	...	...	...	...	...	...	...	...
93	1	0.0	1.0	0.0	1.0	1.0	Dorset Park, Wexford Heights, Scarborough Town...	M1P	Scarborough	43.757410	-79.273304
94	1	0.0	1.0	0.0	1.0	1.0	Business reply mail Processing Centre, South C...	M7Y	East Toronto	43.662744	-79.321558
95	1	0.0	1.0	0.0	1.0	1.0	Glencairn	M6B	North York	43.709577	-79.445073
96	1	0.0	1.0	0.0	1.0	1.0	Rouge Hill, Port Union, Highland Creek	M1C	Scarborough	43.784535	-79.160497
97	1	0.0	1.0	0.0	1.0	1.0	Wexford, Maryvale	M1R	Scarborough	43.750072	-79.295849

64 rows × 11 columns

The 3'th Cluster

	Cluster Labels	Restaurant Count	Office Count	School Count	Population	success rate	Neighbourhood	Postal Code	Borough	Latitude	Longitude
0	2	0.0	23.0	13.0	283.0	283.0	University of Toronto, Harbord	M5S	Downtown Toronto	43.662696	-79.400049
1	2	0.0	10.0	12.0	250.0	250.0	Regent Park, Harbourfront	M5A	Downtown Toronto	43.654260	-79.360636
2	2	0.0	13.0	10.0	213.0	213.0	Willowdale, Willowdale East	M2N	North York	43.770120	-79.408493
3	2	0.0	15.0	9.0	195.0	195.0	Davisville	M4S	Central Toronto	43.704324	-79.388790
4	2	0.0	50.0	7.0	190.0	190.0	Berczy Park	M5E	Downtown Toronto	43.644771	-79.373306

The 4'th Cluster

	Cluster Labels	Restaurant Count	Office Count	School Count	Population	success rate	Neighbourhood	Postal Code	Borough	Latitude	Longitude
16	3	1.0	50.0	10.0	250.0	125.000000	Stn A PO Boxes	M5W	Downtown Toronto	43.646435	-79.374846
19	3	1.0	50.0	9.0	230.0	115.000000	Toronto Dominion Centre, Design Exchange	M5K	Downtown Toronto	43.647177	-79.381576
20	3	1.0	28.0	10.0	228.0	114.000000	Queen's Park, Ontario Provincial Government	M7A	Downtown Toronto	43.662301	-79.389494
28	3	3.0	50.0	14.0	330.0	82.500000	St. James Town	M5C	Downtown Toronto	43.651494	-79.375418
36	3	3.0	47.0	10.0	247.0	61.750000	Church and Wellesley	M4Y	Downtown Toronto	43.665860	-79.383160
42	3	3.0	50.0	9.0	230.0	57.500000	Commerce Court, Victoria Hotel	M5L	Downtown Toronto	43.648198	-79.379817
43	3	3.0	50.0	9.0	230.0	57.500000	First Canadian Place, Underground city	M5X	Downtown Toronto	43.648429	-79.382280
44	3	4.0	50.0	11.0	270.0	54.000000	Richmond, Adelaide, King	M5H	Downtown Toronto	43.650571	-79.384568
68	3	5.0	21.0	10.0	221.0	36.833333	Kensington Market, Chinatown, Grange Park	M5T	Downtown Toronto	43.653206	-79.400049

The 5'th Cluster

	Cluster Labels	Restaurant Count	Office Count	School Count	Population	success rate	Neighbourhood	Postal Code	Borough	Latitude	Longitude
18	4	4.0	50.0	27.0	590.0	118.0	Garden District, Ryerson	M5B	Downtown Toronto	43.657162	-79.378937
23	4	4.0	50.0	23.0	510.0	102.0	Central Bay Street	M5G	Downtown Toronto	43.657952	-79.387383

Discussion

As seen in the results section, four of the five clusters generated in this project had a ratio of taco restaurants to office buildings of roughly 1:1, 2:1, or 1:2, while only one cluster (cluster 4) had a drastically different ratio of 1:7. This means that for every taco restaurant in these locations, there was an average of seven office buildings. Overall, the data makes it clear that the locations in cluster 4 are in the most need of a new taco restaurant, as the ratio of taco restaurants to office buildings is by far the lowest (or, the ratio of office buildings to taco restaurants is the highest). This means that if an entrepreneur were to open up a new taco restaurant in any of the areas grouped into cluster 4, he or she would be much more likely to have a higher number of customers than if the same shop had opened up in clusters 1, 2, 3, or 5

Conclusion

This project analyzed data from the Toronto area to determine where the best place for an entrepreneur to open up a taco restaurant would be. In addition to the number of taco restaurants in each area, the number of office buildings was considered, due to the idea that taco restaurants located close to office buildings will gain a portion of those employees as customers, as coffee is a very common item to grab on the way into work or throughout the workday. By using Kmeans clustering, it was possible to determine the best locations for a new taco restaurant, based on the fact the ideal cluster contained—by far—the lowest ratio of taco restaurants to office buildings. Therefore, an entrepreneur opening up a new taco restaurant in any of the locations grouped into this cluster would be likely to have a much higher number of potential customers.