IBM Data Science Capstone Project

Introduction:

A stakeholder is investing in a new Fast Food restaurant in Toronto, CA. As an investor, they want to make sure that the owner of this restaurant is considering many variables when deciding which location to choose for their new restaurant. The end goal, of course, for the stakeholder and the owner is profitability of this new restaurant. Many factors can affect profitability of a restaurant, so I am going to do an analysis of some of those variables so that the owner can make a more informed decision regarding the location of this new restaurant.

The questions and thus variables that I will be trying to address here are limited, but powerful. Through the data described in the next section, I will be able to answer the following questions, which will be very useful in the decision-making process for this owner and stakeholder.

1) How many neighborhoods are there in this postcode area?
   a. This will be a proxy for population of the postcode since more neighborhoods implies higher population.
2) How many restaurants are there overall in each postcode?
   a. This will indicate how crowded the industry is overall in each postcode, which will help the owner decide if there is room in a certain postcode for new competition.
3) How many similar restaurants are there in each postcode?
   a. This will allow the owner to see how many restaurants there are that are in a similar category, which will provide information about direct competition to his new restaurant.

With these questions answered, I will cluster the postcodes using the k-means algorithm, which will allow us to more easily balance total number of restaurants with the number of direct competitors in fast food restaurants.

Description of the data:

I will be using two main data sources. First, I scraped the website https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M to create a pandas data frame that includes the postcode, borough name, and neighborhood names within each postcode. I also used the geopy geocoder library in python to get the latitude and longitude of each postcode as well. This dataset allowed me to answer the first question above. Second, I used the foursquare API to input the geographical information from the first dataset to gather information about restaurants near each postcode. This dataset allowed me to answer questions two and three above.

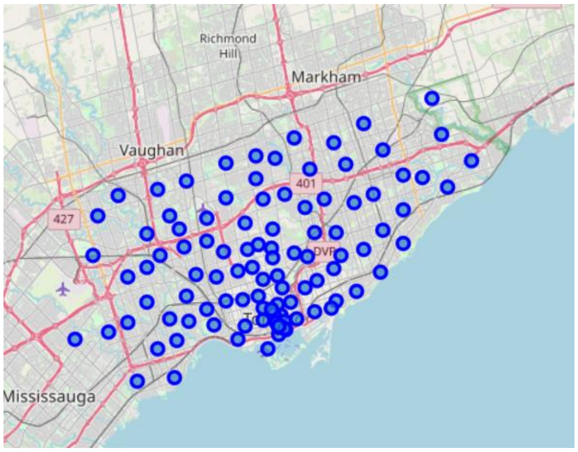
Methodology:

The methodology used involved the steps outlined below.

1) Gather data on the postcodes, their latitudes and longitudes, and neighbourhoods within them for the Toronto area.
   a. This data was gathered by scraping the website listed in the data description above.

	Postcode	Borough	Neighbourhood	Latitude	Longitude	# Neighbourhoods
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353	2
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	3
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711	3
3	M1G	Scarborough	Woburn	43.770992	-79.216917	1
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476	1

b. I also overlaid these postcode locations onto a map of Toronto so that I could get a visual of their locations before the clustering.



2) Use the foursquare API to gather information about venues near each postcode's geospatial coordinates.

a. Here are the first five rows of that information from foursquare.

	Postcode	Postcode Latitude	Postcode Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	M1B	43.806686	-79.194353	Wendy's	43.802008	-79.198080	Fast Food Restaurant
1	M1B	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
2	M1B	43.806686	-79.194353	Staples Morningside	43.800285	-79.196607	Paper / Office Supplies Store
3	M1B	43.806686	-79.194353	Harvey's	43.800020	-79.198307	Restaurant
4	M1B	43.806686	-79.194353	Caribbean Wave	43.798558	-79.195777	Caribbean Restaurant

b. Notice that each postcode has multiple venues and venue categories. This required some additional formatting to get it into a format that could be used for clustering.

3) Use the OneHot encoding method to organize that foursquare data by venue category.

a. I needed a way to sum up the quantities of each venue category to be able to get an understanding of the competition. OneHot encoding helped me set that up. Here are the first five rows.

	Postcode	Accessories Store	Afghan Restaurant	Airport	Airport Lounge	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	Aquarium	...	Video Store	Vietnamese Restaurant	Warehouse Store	Whisky Bar	Wine Bar	Wine Shop	Wings Joint	Women's Store	Yoga Studio	Zoo
0	M1B	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	M1B	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	M1B	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	M1B	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	M1B	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

5 rows × 328 columns

4) Group the OneHot encoded dataframe by postcode to get the sum of each category for each postcode and filter the columns to include only those within the restaurant category.

a. The OneHot encoding allowed me to group by postcode and sum the numbers of venues in each category. Here are the first five rows.

	Postcode	Postcode Latitude	Postcode Longitude	Afghan Restaurant	American Restaurant	Asian Restaurant	Belgian Restaurant	Brazilian Restaurant	Cajun / Creole Restaurant	Cantonese Restaurant	...	Taiwanese Restaurant	Tapas Restaurant	Thai Restaurant	Theme Restaurant	Tibetan Restaurant	Turkish Restaurant	Udon Restaurant	Vegetarian / Vegan Restaurant	Vie Re
0	M1B	43.806686	-79.194353	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
1	M1C	43.784535	-79.160497	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
2	M1E	43.763573	-79.188711	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
3	M1G	43.770992	-79.216917	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
4	M1H	43.773136	-79.239476	0	0	1	0	0	0	0	...	0	0	1	0	0	0	0	0	

5 rows × 67 columns

b. Having the number of each type of restaurant in each postcode was needed to calculate the total number of restaurants in each postcode.

5) Filter that further to include only the venue category of Fast Food Restaurant and calculate the percent of restaurants that are fast food in each postcode area.
    a. Then I needed to filter this further so that I could see exactly the number of specifically Fast Food Restaurants there were in each postcode as well. Here are the first five rows.

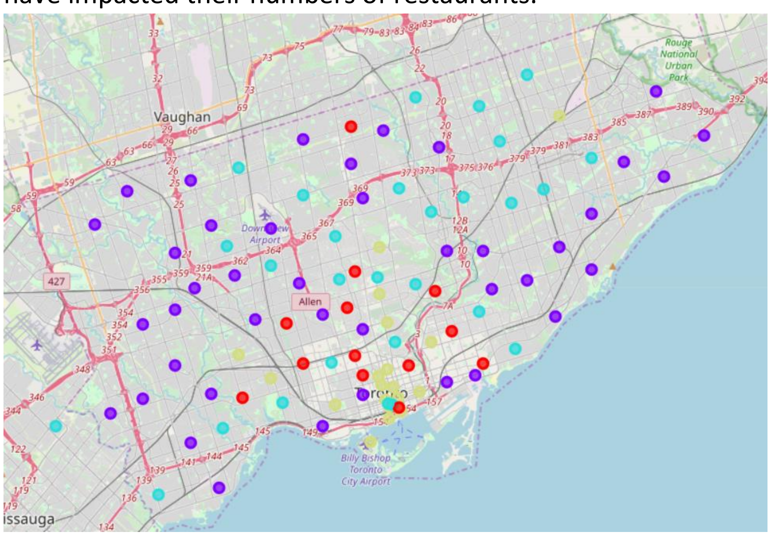
	Postcode	Postcode Latitude	Postcode Longitude	# Neighbourhoods	Fast Food Restaurant	Total Restaurants	Percent Fast Food
0	M1B	43.806686	-79.194353	2	2	6	0.33
1	M1C	43.784535	-79.160497	3	0	1	0.00
2	M1E	43.763573	-79.188711	3	2	3	0.67
3	M1G	43.770992	-79.216917	1	1	3	0.33
4	M1H	43.773136	-79.239476	1	1	8	0.12

    b. This allowed me to also calculate the percent of restaurants that are of the fast food type in each postcode as well.

6) Use the k-means algorithm to cluster the dataframe by total number of restaurants, number of fast food restaurants, number of neighbourhoods, and percent of fast food restaurants in each postcode area.
    a. With this information in hand, I was able to use k-means to cluster them into 4 clusters based on the pertinent data. Here are the first five rows.

	Postcode	Postcode Latitude	Postcode Longitude	# Neighbourhoods	Fast Food Restaurant	Total Restaurants	Percent Fast Food	Cluster Labels
0	M1B	43.806686	-79.194353	2	2	6	0.33	1
1	M1C	43.784535	-79.160497	3	0	1	0.00	1
2	M1E	43.763573	-79.188711	3	2	3	0.67	1
3	M1G	43.770992	-79.216917	1	1	3	0.33	1
4	M1H	43.773136	-79.239476	1	1	8	0.12	2

7) Visualize the clusters on a map of Toronto to gain insights as to their general locations.
    a. Now I can finally map these clusters back onto my Toronto map to gain insights as to how their locations have impacted their numbers of restaurants.



Cluster 0: Red

Cluster 1: Purple

Cluster 2: Teal

Cluster 3: Yellow

    b. While the clusters appear to be relatively spread out, we can see that there are commonalities within each cluster. It appears Cluster 3 (Yellow) is located mostly within the area of Downtown Toronto. Cluster 0 (Red) is generally located close to Downtown, but mostly just outside of it. Cluster 2 (Teal) is generally just a little further out from downtown compared to Cluster 1. Cluster 1 (Purple) is generally the furthest out from Downtown. Below, we will look at the data within each cluster to determine how those general locations effect the total number of restaurants and the Fast Food Restaurants. We will also do a quick analysis of each cluster to help you make a more informed decision about where your new fast food restaurant might be most profitable.

8) Look at and analyze the data within each cluster in a way that gives us important information to make an informed decision about our recommendations for postcode area locations for this new restaurant.

a. Cluster 0: Red

Average Number of Neighbourhoods by post code: 1.69

Average Number of Total Restaurants by postcode: 31

Average percent of restaurants that are fast food by postcode: 3.46

	Postcode	Postcode Latitude	Postcode Longitude	# Neighbourhoods	Fast Food Restaurant	Total Restaurants	Percent Fast Food	Cluster Labels
21	M2N	43.789053	-79.408493	2	5	38	0.13	0
39	M4J	43.705369	-79.349372	1	3	27	0.11	0
40	M4K	43.685347	-79.338106	1	3	32	0.09	0
46	M4S	43.715383	-79.405678	1	2	36	0.06	0
51	M4Y	43.667967	-79.367675	2	1	32	0.03	0
84	M7A	43.651571	-79.484450	2	1	29	0.03	0
42	M4M	43.668999	-79.315572	2	0	29	0.00	0
64	M5R	43.696948	-79.411307	2	0	31	0.00	0
65	M5S	43.672710	-79.405678	3	0	27	0.00	0
66	M5T	43.662696	-79.400049	2	0	30	0.00	0
69	M5X	43.646435	-79.374846	1	0	30	0.00	0
74	M6G	43.689026	-79.453512	1	0	35	0.00	0
76	M6J	43.669005	-79.442259	2	0	30	0.00	0

As we mentioned above, this cluster is generally the second closest to the downtown Toronto area. From this geographical perspective, we think the data makes sense. As you can tell from the data below, these postcodes have a high number of total restaurants and a small percentage of Fast Food restaurants.

This could suggest that these areas have a high population and a higher socio-economic demographic. It could also suggest that the cost to open and operate a restaurant in these areas are relatively inexpensive as well, and thus could provide a good cost-benefit ratio.

Based on this, I think these postcode areas could be an excellent choice for your new fast food restaurant because the high number of total restaurants suggest a lot of people traffic and the lower average percent of fast food suggest lower direct competition than some of the other clusters. I also think this would be a good cluster to choose from because since most of the postcodes are not exactly in downtown, the cost of owning and operating such a restaurant should be lower than those located in the heart of downtown.

b. Cluster 1: Purple

Average Number of Neighbourhoods by post code: 2.33

Average Number of Total Restaurants by postcode: 2.35

Average percent of restaurants that are fast food by postcode: 11.46

	Postcode	Postcode Latitude	Postcode Longitude	# Neighbourhoods	Fast Food Restaurant	Total Restaurants	Percent Fast Food	Cluster Labels
2	M1E	43.763573	-79.188711	3	2	3	0.67	1
34	M4B	43.725882	-79.315572	1	1	2	0.50	1
100	M9V	43.688905	-79.554724	4	1	2	0.50	1
0	M1B	43.806686	-79.194353	2	2	6	0.33	1
7	M1L	43.711112	-79.284577	3	1	3	0.33	1
24	M3A	43.782736	-79.442259	1	1	3	0.33	1
79	M6M	43.713756	-79.490074	3	1	3	0.33	1
6	M1K	43.727929	-79.262029	3	2	6	0.33	1
5	M1J	43.744734	-79.239476	1	1	3	0.33	1
3	M1G	43.770992	-79.216917	1	1	3	0.33	1
16	M2H	0.000000	0.000000	1	1	4	0.25	1
80	M6N	43.691116	-79.476013	4	1	4	0.25	1
87	M8V	43.662744	-79.321558	1	1	5	0.20	1
73	M6E	43.693781	-79.428191	1	1	6	0.17	1
92	M9A	43.628841	-79.520999	5	0	0	0.00	1
88	M8W	43.605647	-79.501321	3	0	1	0.00	1

As we mentioned above, this cluster includes postcodes that are generally furthest away from Downtown. From this geographical perspective, we think the data makes sense. As you can tell from the data below, these postcodes have a small number of total restaurants and a high percentage of Fast Food restaurants, but also have a higher number of neighborhoods as well.

This could suggest a number of things, including that since there are more neighborhoods on average in these postcodes that there is not as much space for businesses to take hold and it also explains why the percent of restaurants that are fast food is higher.

Based on this, I would generally not suggest postcodes in this cluster as there is already a lot of direct competition because the fast food percentage is so high in most of the neighbourhoods.

c. Cluster 2: Teal

Average Number of Neighbourhoods by post code: 1.89

Average Number of Total Restaurants by postcode: 10.75

Average percent of restaurants that are fast food by postcode: 8.96

	Postcode	Postcode Latitude	Postcode Longitude	# Neighbourhoods	Fast Food Restaurant	Total Restaurants	Percent Fast Food	Cluster Labels
71	M6B	43.718518	-79.464763	2	3	9	0.33	2
15	M1W	43.799525	-79.318389	1	2	7	0.29	2
32	M3N	43.728496	-79.495697	1	2	7	0.29	2
14	M1V	43.815252	-79.284577	4	2	11	0.18	2
13	M1T	43.781638	-79.304302	3	2	12	0.17	2
70	M6A	43.648429	-79.382280	2	2	12	0.17	2
61	M5M	43.648198	-79.379817	2	2	13	0.15	2
17	M2J	43.803762	-79.363452	1	1	7	0.14	2
10	M1P	43.757410	-79.273304	3	2	14	0.14	2
4	M1H	43.773136	-79.239476	1	1	8	0.12	2
86	M7Y	43.636966	-79.615819	1	1	9	0.11	2
28	M3J	43.754328	-79.442259	3	1	10	0.10	2
63	M5P	43.711695	-79.416936	1	1	11	0.09	2
45	M4R	43.712751	-79.390197	1	1	11	0.09	2
38	M4H	43.709060	-79.363452	1	1	12	0.08	2
91	M8Z	43.636258	-79.498509	8	1	16	0.06	2

As mentioned above, this cluster is further away from downtown, but not as far as cluster 0. From this geographical perspective, we think this again, makes sense. The data below indicates that this cluster includes postcodes with a medium number of restaurants and a medium percent of restaurants that are fast food.

This could suggest that this is mostly a residential area where there are a good number of people that want quick meals relatively nearby, but is likely not a large tourist area.

Based on this data, this could be a good cluster to choose a postcode area from, but not the ideal choice in my view because many of them already have a higher fast food percentage.

d. Cluster 3: Yellow

Average Number of Neighbourhoods by post code: 1.89

Average Number of Total Restaurants by postcode: 22.21

Average percent of restaurants that are fast food by postcode: 1.84

	Postcode	Postcode Latitude	Postcode Longitude	# Neighbourhoods	Fast Food Restaurant	Total Restaurants	Percent Fast Food	Cluster Labels
44	M4P	43.728020	-79.388790	1	4	25	0.16	3
81	M6P	43.673185	-79.487262	2	1	22	0.05	3
52	M5A	43.665860	-79.383160	1	1	20	0.05	3
41	M4L	43.679557	-79.352188	2	1	19	0.05	3
53	M5B	43.654260	-79.360636	1	1	25	0.04	3
12	M1S	43.794200	-79.262029	1	0	23	0.00	3
59	M5K	43.640816	-79.381752	3	0	26	0.00	3
82	M6R	43.661608	-79.464763	2	0	22	0.00	3
77	M6K	43.647927	-79.419750	2	0	24	0.00	3
68	M5W	43.628947	-79.394420	7	0	22	0.00	3
60	M5L	43.647177	-79.381576	2	0	25	0.00	3
56	M5G	43.644771	-79.373306	1	0	26	0.00	3
58	M5J	43.650571	-79.384568	3	0	19	0.00	3
57	M5H	43.657952	-79.387383	1	0	24	0.00	3
55	M5E	43.651494	-79.375418	1	0	18	0.00	3
54	M5C	43.657162	-79.378937	2	0	18	0.00	3
48	M4V	43.689574	-79.383160	2	0	25	0.00	3
47	M4T	43.704324	-79.388790	1	0	20	0.00	3

As mentioned above, this cluster includes postcodes that are generally in or very near downtown Toronto. The data below indicates that there is a medium-high number of restaurants and a low percentage of fast food restaurants.

This could suggest that there is likely a large population and that it is likely a high tourist destination as well, but that cost of owning and operating a restaurant is costly.

Based on this, this is could be a good cluster to choose your postcode location from, but I fear that since there are a lower number of restaurants in this cluster compared to cluster 0 that this cost in this area is high. More study would be necessary to determine that, but that is beyond the scope of this analysis.

Results:

I discuss the results of the analysis under the data of each cluster above, but I will be a quick summary of those results here. The clusters are mainly organized by Low, Medium, Medium-high, and High numbers of restaurants and percentage of restaurants that are fast food. The number of neighborhoods is also gathered for each postcode, which could provide some helpful information as well. Cluster 0 has postcodes with high number of restaurants and low percentages of fast food restaurants, but low number of neighbourhoods. Cluster 1 has low number of restaurants, high percentages of fast food, and high number of neighbourhoods. Cluster 2 has medium number restaurants, medium number of neighbourhoods, but medium number of percentages of fast food. Finally, cluster 3 has high number of restaurants, medium number of neighbourhoods, and low percentages of fast food.

Discussion:

It is of interest to the stakeholder and owner of a new fast food restaurant to try to total number of restaurants and percentage of those restaurants that are fast food, as well as consider the number of neighbourhoods. So, based on the data, here are my recommendations. Cluster 0 has postcodes with a high number of total restaurants, but a low percentage of restaurants that are of the fast food category. This would be an ideal mix as it would indicate that there is a lot of people traffic, but low number of direct competitors to a fast food restaurant. It is my first recommendation of postcodes to research. My second-choice recommendation would be postcodes from cluster 2. While they do have a lower number of total restaurants and a higher percentage of fast food restaurants, they also have a higher number of neighbourhoods, which would suggest more families looking for a quick meal. It also suggests that the cost of ownership in that area is likely lower as well. Cluster 3 could be a good choice but it is my third recommendation because those postcodes are located in the heart of downtown and thus will likely be expensive to own and operate a fast food restaurant, plus many restaurants would already be established there.

Conclusion:

In conclusion, There are some meaningful insights to be gained from these data and through a k-means clustering, we were able to identify some postcode areas that are likely better candidates for a new fast food restaurant than others. It is important to note, though, that while these are powerful data results, more research could be done regarding cost of ownership in certain postcodes and populations sizes as well. Despite that, I believe we have some excellent information to move forward with seeking a location for your new fast food restaurant.