

IBM Data Science Professional Certificate
Capstone Project - The Battle of Neighborhoods
“Where To Locate Your Restaurant”

Background:

Toronto is the largest city in Canada and one of the most ethnically diverse cities in the world. Many immigrant cultures have brought their traditions, languages, and music to Toronto. The city features many distinctive neighborhoods bustling with activity and vitality. These include the largest collection of Victorian-era industrial architecture in North America, the largest urban car-free community in North America, and the bohemian heart of the city, Kensington Market. Toronto is also well known for its multi-cuisine dishes. Downtown Toronto has many opportunities for entrepreneurs to start their business.

Problem Statement:

In the city of Toronto, an entrepreneur is looking to open an Italian restaurant. As a data scientist, where would we recommend him to open it for his maximum profit?

To solve this problem we will have to look at all the neighborhoods of Toronto (in all aspects) and form different clusters depending on the number of Italian restaurants they have and then choose the cluster with the minimum number of Italian restaurants.

Target Audience: Entrepreneurs who want to set up an Italian Restaurant business in the city of Toronto

Data Overview:

The data that we will be using for this project comes from different sources:

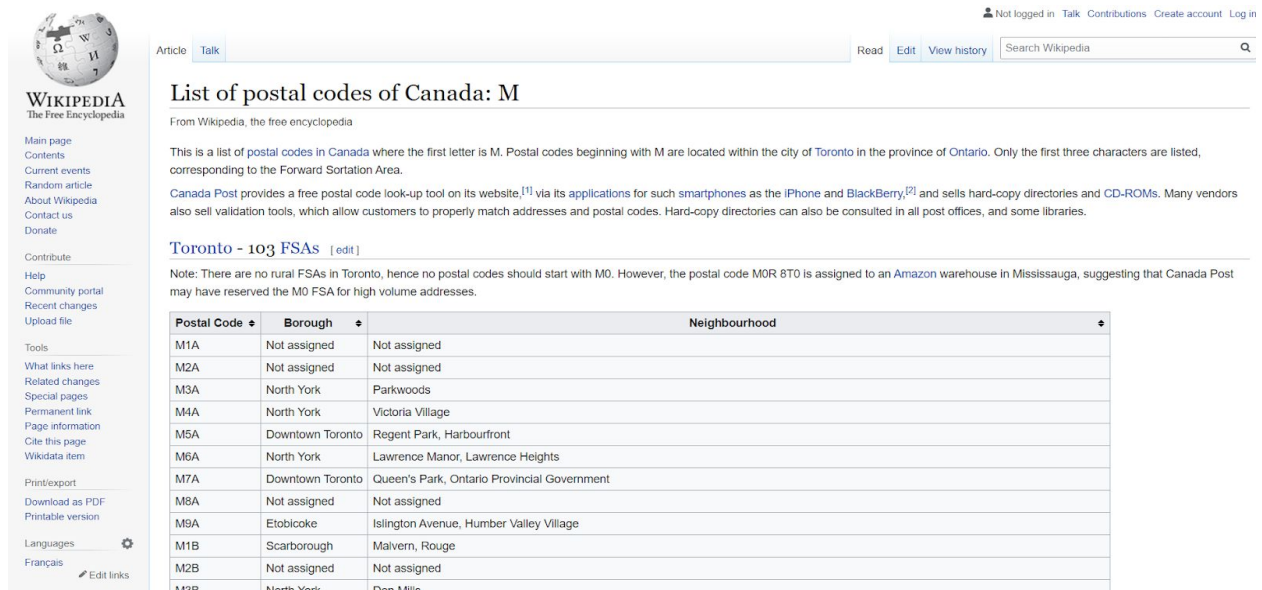
1. Using the Wiki page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M), we will get all neighborhoods of Toronto.
2. Next, we add the geographical locations (in terms of longitude and latitude) of all neighborhoods using the geocoder package.

3. Next using the Foursquare api, we will get venue details of all Italian restaurants in Toronto city. All this info will help us to locate the most suitable location for opening an Italian restaurant.

Methodology:

Data Sources:

1. Web-scraping Toronto neighborhoods data from the wiki page



The screenshot shows the Wikipedia page titled "List of postal codes of Canada: M". The page includes a sidebar with navigation links and a main content area with a table of postal codes. The table has three columns: Postal Code, Borough, and Neighbourhood. The data is as follows:

Postal Code	Borough	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Regent Park, Harbourfront
M6A	North York	Lawrence Manor, Lawrence Heights
M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
M8A	Not assigned	Not assigned
M9A	Etobicoke	Islington Avenue, Humber Valley Village
M1B	Scarborough	Malvern, Rouge
M2B	Not assigned	Not assigned
M3B	North York	Don Mills

Fig1: Wiki Page having List of Neighborhoods in Toronto along with Postal Codes

The Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) provides us the information about the neighborhoods of Toronto city. It includes the postal code, borough, and the name of the neighborhoods. The given page is scrapped using beautiful soup to get the data in the desired format for analysis (fig 2).

	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

Fig2: Pandas Dataframe from the scrapped wiki page

2. The geographical coordinates of the neighborhoods with the respective Postal Codes are present in a CSV file (source: https://cocl.us/Geospatial_data). This CSV file is converted to a pandas data frame and then merged with the scrapped data frame.

	A	B	C
1	Postal Code	Latitude	Longitude
2	M1B	43.8066863	-79.1943534
3	M1C	43.7845351	-79.1604971
4	M1E	43.7635726	-79.1887115
5	M1G	43.7709921	-79.2169174
6	M1H	43.773136	-79.2394761
7	M1J	43.7447342	-79.2394761

Fig3: Geographical data of Neighborhoods in Toronto

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Fig4: Reading CSV file into a data frame

3. Using the Foursquare api, the location, name, and category of different venues in Toronto were retrieved.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	The Beaches	43.676357	-79.293031	Glen Manor Ravine	43.676821	-79.293942	Trail
1	The Beaches	43.676357	-79.293031	The Big Carrot Natural Food Market	43.678879	-79.297734	Health Food Store
2	The Beaches	43.676357	-79.293031	Grover Pub and Grub	43.679181	-79.297215	Pub

Fig5: Venues data of Toronto neighborhoods retrieved using Foursquare api

Data Preparation:

The data frame from the scrapped wiki page will consist of three columns: PostalCode, Borough, and Neighborhood.

1. We only process the cells that have an assigned borough. Cells with a borough Not assigned were dropped.
2. More than one neighborhood can exist in one postal code area. For example, in the table on the Wikipedia page, we will notice that M5A is listed twice and has two

neighborhoods: Harbourfront and Regent Park. These two rows will be combined into one row with the neighborhoods separated with a comma.

3. If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough.

After applying all the above three processes, we got the following data frame (fig6).

	Postcode	Borough	Neighbourhood
0	M1B	Scarborough	Rouge, Malvern
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

Fig6: Processed wiki data frame

Using the geographical location collected from the CSV file, we merged it with the above data frame based on Postal Code.

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

Fig7: Neighborhood information along with the geographical location

Then we filtered the above data frame by picking only those rows which have Toronto word in their Borough column.

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M4E	East Toronto	The Beaches	43.676357	-79.293031
1	M4K	East Toronto	The Danforth West,Riverdale	43.679557	-79.352188
2	M4L	East Toronto	The Beaches West,India Bazaar	43.668999	-79.315572
3	M4M	East Toronto	Studio District	43.659526	-79.340923
4	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790

Fig8: Filtered data frame containing only those rows having Toronto keyword in Borough

To get the nearest Venue for each of the Neighborhoods, we merged the Foursquare Venue data with the Neighborhood data (fig9).

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	The Beaches	43.676357	-79.293031	Glen Manor Ravine	43.676821	-79.293942	Trail
1	The Beaches	43.676357	-79.293031	The Big Carrot Natural Food Market	43.678879	-79.297734	Health Food Store
2	The Beaches	43.676357	-79.293031	Grover Pub and Grub	43.679181	-79.297215	Pub
3	The Beaches	43.676357	-79.293031	Upper Beaches	43.680563	-79.292869	Neighborhood
4	The Beaches	43.676357	-79.293031	Seaspray Restaurant	43.678888	-79.298167	Asian Restaurant

Fig9: Data frame after merging venue data with neighborhood data

Using folium we created a map and color-coded each neighborhood depending on what Borough it was situated in.

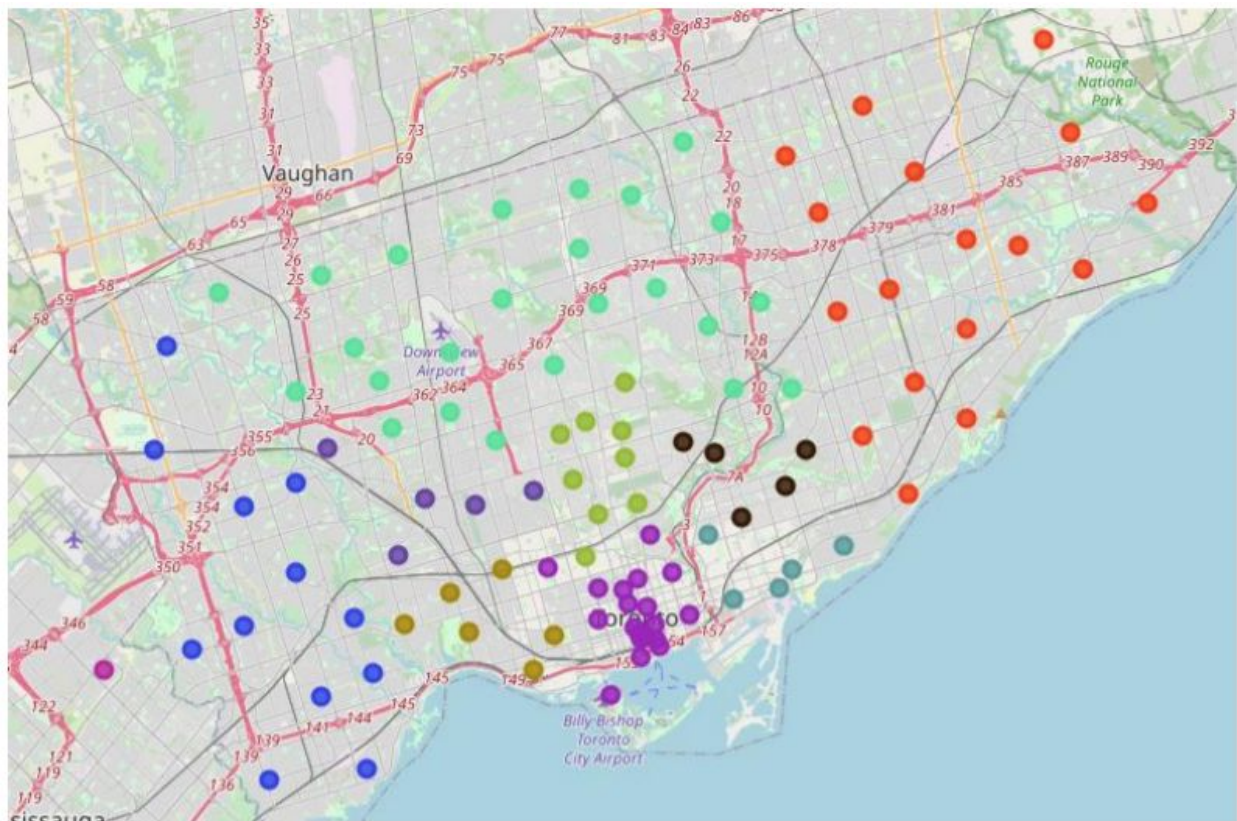


Fig10: Neighborhood map

For every neighborhood, individual venues were one hot encoded to know how many of those Venues were located in each neighborhood?

	Neighborhoods	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	...
0	Lawrence Park	0	0	0	0	0	0	0	0	0	...
1	Lawrence Park	0	0	0	0	0	0	0	0	0	...
2	Lawrence Park	0	0	0	0	0	0	0	0	0	...
3	Davisville North	0	0	0	0	0	0	0	0	0	...
4	Davisville North	0	0	0	0	0	0	0	0	0	...

Fig11: One Hot Encoding Venues in each neighborhood

	Neighborhoods	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	...
0	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...
1	Alderwood, Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...
2	Bathurst Manor, Wilson Heights, Downsview North	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...
3	Bayview Village	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...
4	Bedford Park, Lawrence Manor East	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.043478	...

Fig12: Neighborhoods grouped by the average of the frequency of each Venue

	Neighborhoods	Italian Restaurant
0	Agincourt	0.000000
1	Alderwood, Long Branch	0.000000
2	Bathurst Manor, Wilson Heights, Downsview North	0.000000
3	Bayview Village	0.000000
4	Bedford Park, Lawrence Manor East	0.130435

Fig13: Neighborhoods along with the average Italian Restaurant in that Neighborhood

For analyzing analysis, we clustered the neighborhoods based on the neighborhoods that had a similar average of Italian Restaurants in that neighborhood. We used K-Means clustering algorithm to carry out this task. We used the Elbow Point Technique to get optimum K value. For calculating optimum k value, we tried with different K values and measured the accuracy. The best K value is found at the elbow point (in this case $K = 4$).

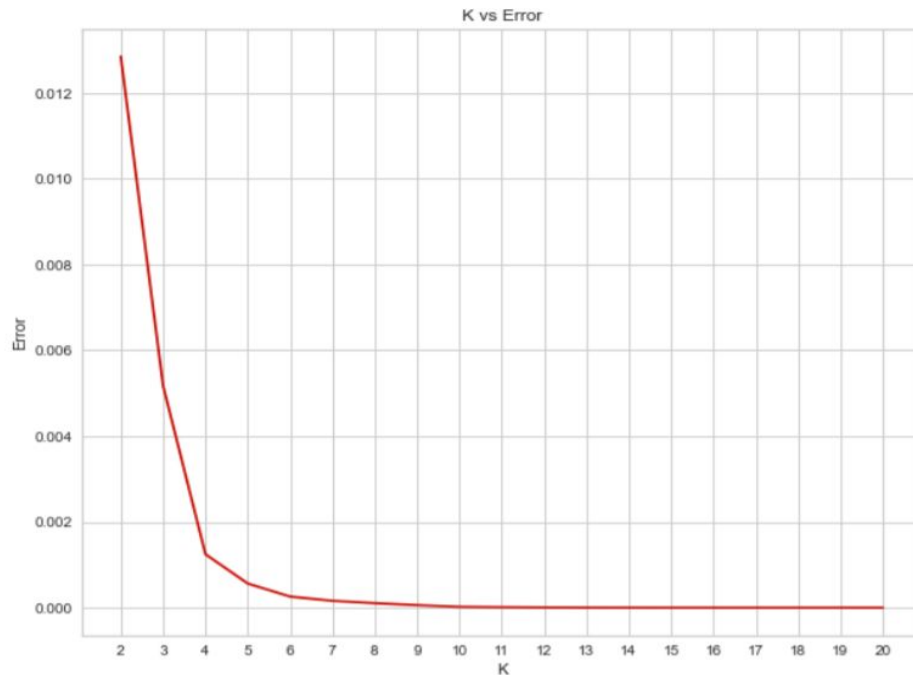


Fig14: Finding the K vs Error Values

Now the neighborhoods were divided into 4 clusters based on the similar mean frequency of Italian Restaurants. Each of these clusters was labeled from 0 to 3.

	Neighborhood	Italian Restaurant	Cluster Labels
0	Agincourt	0.000000	1
1	Alderwood, Long Branch	0.000000	1
2	Bathurst Manor, Wilson Heights, Downsview North	0.000000	1
3	Bayview Village	0.000000	1
4	Bedford Park, Lawrence Manor East	0.130435	0

Fig15: Assigning cluster label to similar neighborhoods

Results & Interpretation:

Before analyzing each cluster individually, first, let's check the total amount of neighborhoods in each cluster and the average Italian Restaurants in that cluster. We see that Cluster 2 has the most neighborhoods (70) while cluster 1 has the least (1). Next, we compared the average Italian Restaurants per cluster. Even though there is only 1 neighborhood in Cluster 1, it has the highest number of Italian Restaurants (0.1304) while Cluster 2 has the most neighborhoods but has the least average of Italian Restaurants (0.0009).

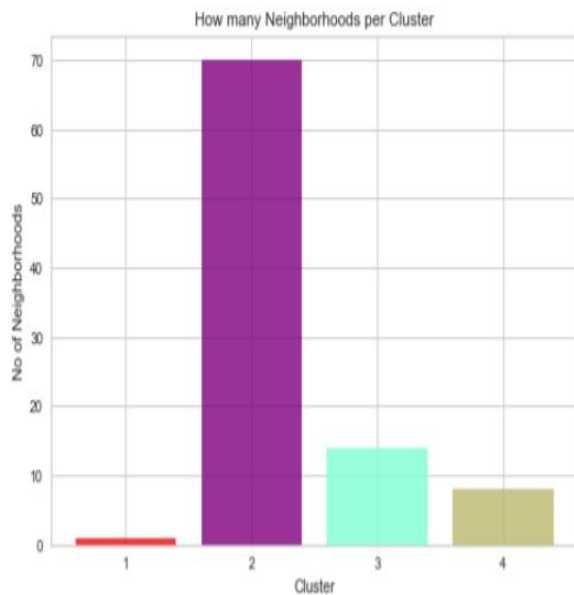


Fig16: Number of Neighborhoods per cluster

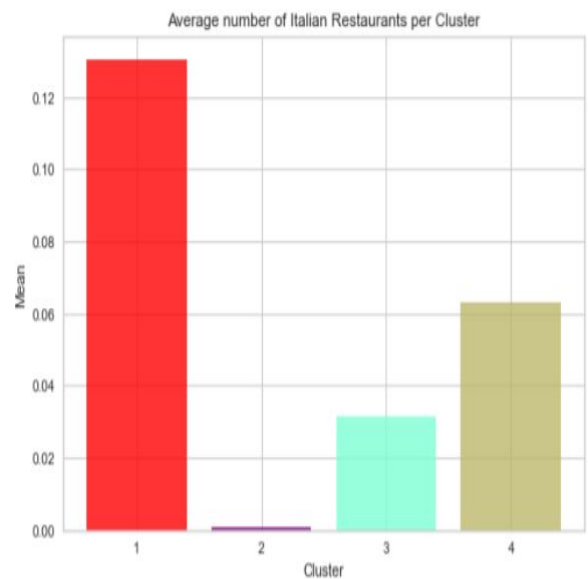


Fig17: Avg Italian restaurant in each neighborhood

Now let's analyze each cluster individually.

Cluster1 Red:

Cluster 1 was in the North York area. Lawrence Manor East and Bedford and were the two neighborhoods in that cluster1. Cluster 1 had 19 unique Venue locations and only 3 were Italian Restaurants. Cluster 1 had the highest average of Italian Restaurants (0.130435).

	Borough	Neighborhood	Italian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	North York	Bedford Park, Lawrence Manor East	0.130435	0	43.733283	-79.41975	LCBO	43.731065	-79.419237	Liquor Store
1	North York	Bedford Park, Lawrence Manor East	0.130435	0	43.733283	-79.41975	Aroma Espresso Bar	43.735975	-79.420391	Café
2	North York	Bedford Park, Lawrence Manor East	0.130435	0	43.733283	-79.41975	Darbar Persian Grill	43.735484	-79.420006	Restaurant
3	North York	Bedford Park, Lawrence Manor East	0.130435	0	43.733283	-79.41975	Satay on the Road	43.735310	-79.419783	Thai Restaurant
4	North York	Bedford Park, Lawrence Manor East	0.130435	0	43.733283	-79.41975	The Copper Chimney	43.736195	-79.420271	Indian Restaurant
5	North York	Bedford Park, Lawrence Manor East	0.130435	0	43.733283	-79.41975	Francobollo	43.734557	-79.419549	Italian Restaurant
6	North York	Bedford Park, Lawrence Manor East	0.130435	0	43.733283	-79.41975	Sakura Garden	43.733398	-79.419491	Sushi Restaurant
7	North York	Bedford Park, Lawrence Manor East	0.130435	0	43.733283	-79.41975	Tim Hortons	43.735356	-79.419605	Coffee Shop
8	North York	Bedford Park, Lawrence Manor East	0.130435	0	43.733283	-79.41975	Pheasant & Firkin	43.735173	-79.419702	Pub
9	North York	Bedford Park, Lawrence Manor East	0.130435	0	43.733283	-79.41975	Freshii	43.731582	-79.419109	Juice Bar

Cluster 2 (Blue) :

	Borough	Neighborhood	Italian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
450	Downtown Toronto	First Canadian Place, Underground city	0.01	1	43.648429	-79.382280	Mercatto	43.650243	-79.380820	Italian Restaurant
480	Downtown Toronto	First Canadian Place, Underground city	0.01	1	43.648429	-79.382280	Pumpnickel's Deli	43.648832	-79.381970	Deli / Bodega
488	Downtown Toronto	First Canadian Place, Underground city	0.01	1	43.648429	-79.382280	Olly Fresco's	43.646912	-79.379597	Deli / Bodega
487	Downtown Toronto	First Canadian Place, Underground city	0.01	1	43.648429	-79.382280	iQ Food Co. (First Canadian Place)	43.648357	-79.382192	Salad Place
486	Downtown Toronto	First Canadian Place, Underground city	0.01	1	43.648429	-79.382280	The Fairmont Royal York	43.645449	-79.381508	Hotel

In cluster 2, there are 70 neighborhoods, 229 different venues, and 1 Italian Restaurant. The average number of Italian Restaurants in Cluster 2 is the lowest (0.01).

Cluster 3 (Turquoise):

	Borough	Neighborhood	Italian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Downtown Toronto	St. James Town, Cabbagetown	0.045455	2	43.667967	-79.367675	Park Snacks	43.666979	-79.363115	Snack Place
1	Downtown Toronto	St. James Town, Cabbagetown	0.045455	2	43.667967	-79.367675	Rosedale Ravine	43.672152	-79.367150	Park
2	Downtown Toronto	St. James Town, Cabbagetown	0.045455	2	43.667967	-79.367675	No Frills	43.663515	-79.367166	Grocery Store
3	Downtown Toronto	St. James Town, Cabbagetown	0.045455	2	43.667967	-79.367675	Wellesley Parliament Square	43.668589	-79.370169	Plaza
4	Downtown Toronto	St. James Town, Cabbagetown	0.045455	2	43.667967	-79.367675	Tender Trap Restaurant	43.667724	-79.369485	Chinese Restaurant
...

Cluster 3 has the second-lowest average of Italian Restaurants. The neighborhood in Cluster 3 belongs to the Downtown Area, East Toronto, West Toronto, and North York. Neighborhoods

such as Toronto Dominion Center, Ryerson, Don Mills, Queen's Park, Garden District were included in this cluster. There area total of 176 unique venues out of which 27 were Italian Restaurants.

Cluster 4 (Dark Khaki):

	Borough	Neighborhood	Italian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Central Toronto	Davisville	0.057143	3	43.704324	-79.38879	Pizza Nova	43.707524	-79.389863	Pizza Place
1	Central Toronto	Davisville	0.057143	3	43.704324	-79.38879	Petro-Canada	43.702269	-79.387955	Gas Station
2	Central Toronto	Davisville	0.057143	3	43.704324	-79.38879	Apple Tree Farmer's Market	43.700326	-79.389760	Farmers Market
3	Central Toronto	Davisville	0.057143	3	43.704324	-79.38879	Starving Artist	43.701538	-79.387240	Restaurant
4	Central Toronto	Davisville	0.057143	3	43.704324	-79.38879	Meow Cat Cafe	43.702927	-79.388190	Café

The venues in Cluster 4 were located in the Downtown, East, West, and Central Toronto areas as well as Scarborough. Some of the neighborhoods in this cluster are Central Bay Street, University of Toronto, Central Bay Street, and Riverdale. There are 91 unique Venues in Cluster 4 with 16 Italian Restaurants. This cluster has the second-highest average of Italian Restaurants (0.063).

Discussion:

Cluste1 (Red) has the highest number of Italian Restaurants are in cluster 1. Bedford Park and Lawrence Manor East in the North York area have the highest average of Italian Restaurants. In cluster2 there is little to no Italian Restaurant as compared to the number of neighborhoods. The Downtown Toronto area in cluster 3 has the second least average of Italian Restaurants. The best place to put a new Italian Restaurant is Downtown Toronto as there are many neighborhoods in the area but no Italian Restaurants so, eliminating any competition. The second-best place that has a good opportunity would be the areas in Fairview, Adelaide, and King, etc. (Cluster 2). These boroughs have 70 neighborhoods with no Italian Restaurants which gives a better opportunity for opening a new Italian restaurant.

So the above places would be recommended to the entrepreneur for opening an Italian restaurant in Toronto city.

Limitations:

The analysis is done using clustering which completely based on data obtained from Foursquare API. The analysis also doesn't consider the Italian population across neighborhoods as this can be an important factor while choosing a place to open a new Italian restaurant.

Conclusion:

In conclusion, to finish off this project, I used the knowledge gained from this course to tackle a business problem as a data scientist. I utilized various libraries such as NumPy, pandas, folium, etc to fetch the information, break down it into useful datasets, and then visualize them. I have also used Foursquare API to find out all the venues in the neighborhoods of Toronto. I also used plotting libraries such as seaborn and matplotlib utilizing different plots to create visual graphs. The limitations of this project can be tackled by gaining more information and using other Machine Learning strategies.

This project will surely pave a path to tackle more complex real-world problems using data science and machine learning.