

The Battle of Neighborhoods

Muhammad Umar Khan

Introduction:

Toronto is one of the most densely populated area in Canada. Being the land of Opportunity, it brings in a variety of people from different ethnic backgrounds to the core city of Canada, Toronto. Being the largest city in Canada with an estimated population of over 6 million, there is no doubt about the diversity of the population. The multiculturalism is seen through the various neighborhoods including; Chinatown, Corso Italia, Little India, Kensington Market, Little Italy, Koreatown and many more. Downtown Toronto being the hub of interactions between ethnicities, brings many opportunities for entrepreneurs to start or grow their business. It is a place where people can try the best of each culture, either while they work or just passing through. Toronto is well known for its great food.

The objective of this project is to use Foursquare location data and regional clustering of venue information to determine what might be the 'best' neighborhood in Toronto to open a restaurant. Pizza and Pasta are one of the most bought dishes in Toronto originating from Italy. Toronto being the fourth largest home to Italians with a population over 500k, there are numerous opportunities to open a new Italian restaurant. Through this project we will find the most suitable location for an entrepreneur to open a new Italian restaurant in Toronto, Canada.

Target Audience:

- Entrepreneurs who want to open an Italian Restaurant in Toronto

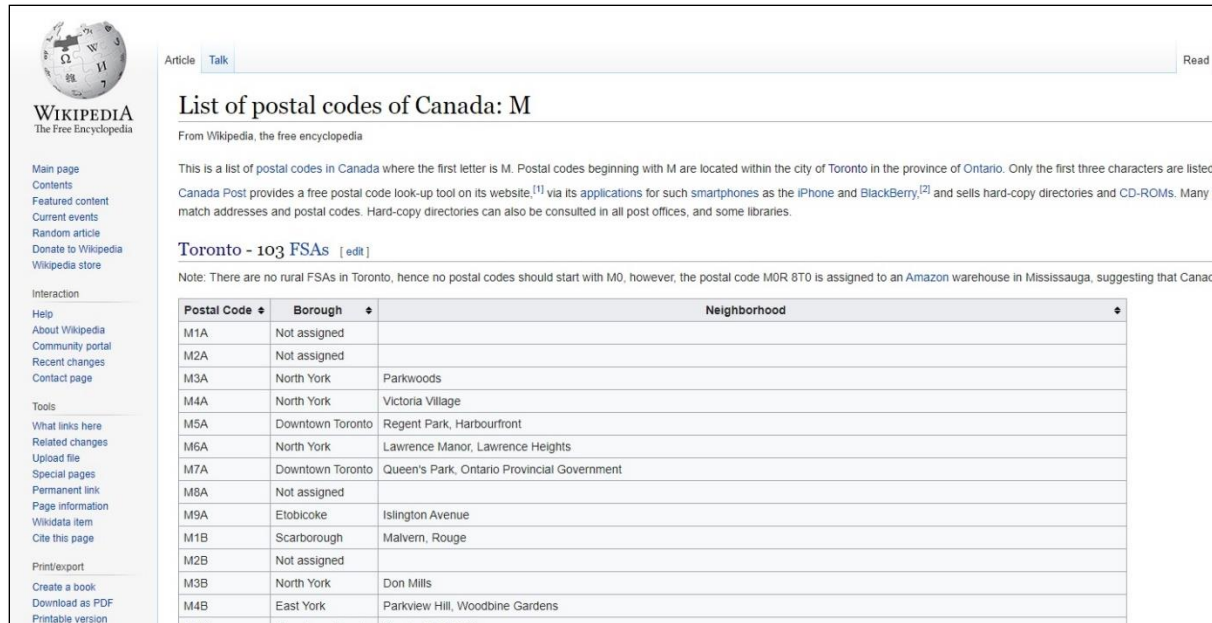
Data Overview:

The data that will be required will be a combination of CSV files that have been prepared for the purposes of the analysis from multiple sources which will provide the list of neighborhoods in Toronto (via Wikipedia), the Geographical location of the neighborhoods (via Geocoder package) and Venue data pertaining to Italian restaurants (via Foursquare). The Venue data will help find which neighborhood is best suitable to open an Italian restaurant.

Methodology:

First, we will need to extract the data from the data sources:

Source 1: Toronto Neighborhoods via Wikipedia



The screenshot shows the Wikipedia page titled "List of postal codes of Canada: M". The page content includes a table with three columns: "Postal Code", "Borough", and "Neighborhood". The table lists various postal codes starting with 'M' and their corresponding boroughs and neighborhoods in Toronto. For example, M3A is North York with the neighborhood Parkwoods, M4A is North York with Victoria Village, M5A is Downtown Toronto with Regent Park, Harbourfront, M6A is North York with Lawrence Manor, Lawrence Heights, and M7A is Downtown Toronto with Queen's Park, Ontario Provincial Government.

Postal Code	Borough	Neighborhood
M1A	Not assigned	
M2A	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	
M9A	Etobicoke	Islington Avenue
M1B	Scarborough	Malvern, Rouge
M2B	Not assigned	
M3B	North York	Don Mills
M4B	East York	Parkview Hill, Woodbine Gardens

Figure 1: Wikipedia Page showing List of Neighborhoods in Toronto with respective Postal Codes

The Wikipedia site (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) shown above, provided almost all the information about the neighborhoods. It included the postal code, borough and the name of the neighborhoods present in Toronto. Since the data is not in a format that is suitable for analysis, scraping of the data was done from this site (shown in figure2).

	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

Figure 2: Data that was scraped from Wikipedia site and put into Pandas data frame

Source2: Geographical Location data using Geocoder Package

	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

Figure 4: 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

Figure 3: Conversion of file into Pandas data frame

The second source of data provided (https://cocl.us/Geospatial_data) us with the Geographical coordinates of the neighborhoods with the respective Postal Codes. The file was in CSV format, so attaching it to a Pandas data frame was simple (shown in figure 3).

Source3: Venue Data using Foursquare

The retrieval of the location, name and category about the various venues in Toronto was collected through the Foursquare explore API. To obtain the data, it was required to make an account where it would provide a 'Secret Key' as well as a 'Client ID' which would allow me to pull any data.

	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

Figure 5: Venue data pulled from Foursquare explore API

It is seen through figure 5 (above) that the neighborhoods are grouped by the neighborhood, so data clustering is made easier later on.

After all the data was collected and put into data frames, cleansing and merging of the data was required to start the process of analysis. When getting the data from Wikipedia, there were Boroughs that were not assigned to any neighborhood therefore, the following assumptions were made:

1. Only the cells that have an assigned borough will be processed. Borough that is not assigned are ignored.
2. More than one neighborhood can exist in one postal code area. For example, in the table on the Wikipedia page, you 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 as shown in *Figure2* row 4.
3. If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough.

After the implementation of the following assumptions, the rows were grouped based on borough as shown below.

	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

Figure 6: Rows grouped together based on Borough

Using the Latitude and Longitude collected from the Geocoder package, we merged the two tables together 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

Figure 7: Merging tables together based on Postal Code

After, the venue data pulled from the Foursquare API was merged with the table above providing us with the local venue within a 500-meter radius shown below.

	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

Figure 8: Local Venues near the respective Neighborhood

Now after cleansing the data, the next step was to analyze it. We then created a map using folium and color coded each Neighborhood depending on what Borough it was located in.

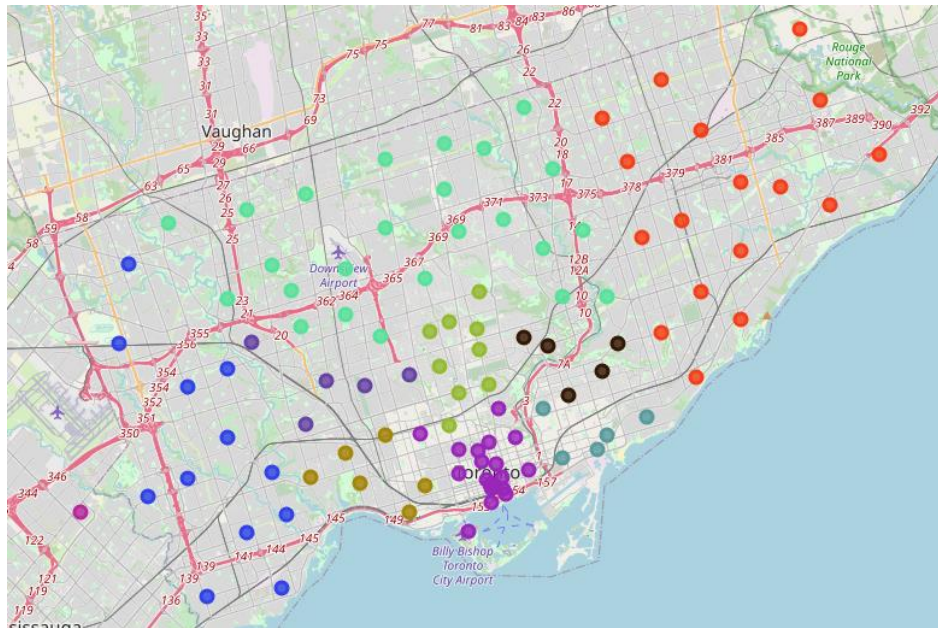


Figure 9: Toronto Neighborhoods

Next, we used the Foursquare API to get a list of all the Venues in Toronto which included Parks, Schools, Café Shops, Asian Restaurants etc. Getting this data was crucial to analyzing the number of Italian Restaurants all over Toronto. There was a total of 45 Italian Restaurants in Toronto. We then merged the Foursquare Venue data with the Neighborhood data which then gave us the nearest Venue for each of the Neighborhoods.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Lawrence Park	43.728020	-79.388790	Lawrence Park Ravine	43.726963	-79.394382	Park
1	Lawrence Park	43.728020	-79.388790	Zodiac Swim School	43.728532	-79.382860	Swim School
2	Lawrence Park	43.728020	-79.388790	TTC Bus #162 - Lawrence-Donway	43.728026	-79.382805	Bus Line
3	Davisville North	43.712751	-79.390197	Homeway Restaurant & Brunch	43.712641	-79.391557	Breakfast Spot
4	Davisville North	43.712751	-79.390197	Sherwood Park	43.716551	-79.387776	Park

Figure 10: Venue table merged with Neighborhood data

Then to analyze the data we performed a technique in which Categorical Data is transformed into Numerical Data for Machine Learning algorithms. This technique is called **One hot encoding**. For each of the neighborhoods, individual venues were turned into the frequency at 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	...

Figure 11: One Hot Encoding

Then we grouped those rows by Neighborhood and by taking the **Average** of the frequency of occurrence of each Venue Category.

	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	...

Figure 12: Grouped Neighborhoods by the average of the frequency of each Venue

After, we created a new data frame which only stored the Neighborhood names as well as the mean frequency of Italian Restaurants in that Neighborhood. This allowed the data to be summarized based on each individual Neighborhood and made the data much simpler to analyze.

	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

Figure 13: New data frame storing Neighborhoods and the average Italian Restaurant in that Neighborhood

To make the analysis more interesting, we wanted to cluster the neighborhoods based on the neighborhoods that had similar averages of Italian Restaurants in that Neighborhood. To do this we used **K-Means** clustering. To get our optimum K value that was neither overfitting or underfitting the model, we used the **Elbow Point** Technique. In this technique we ran a test with different number of K values and measured the accuracy and then chose the best K value. The best K value is chosen at the point in which the line has a sharpest turn. In our case we had the Elbow Point at K = 4. That means we will have a total of 4 clusters.

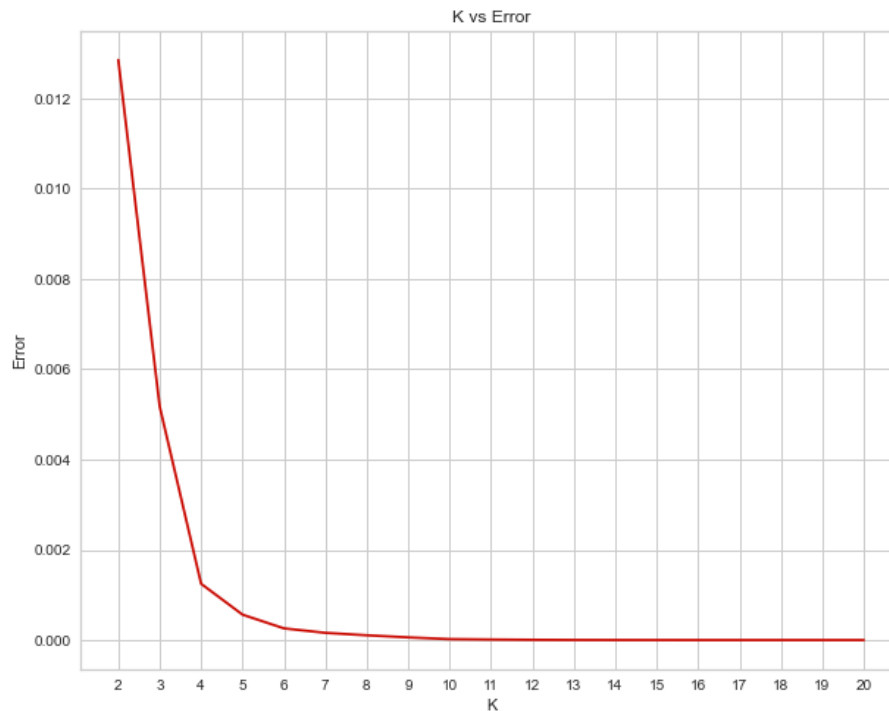


Figure 14: Finding the K vs Error Values

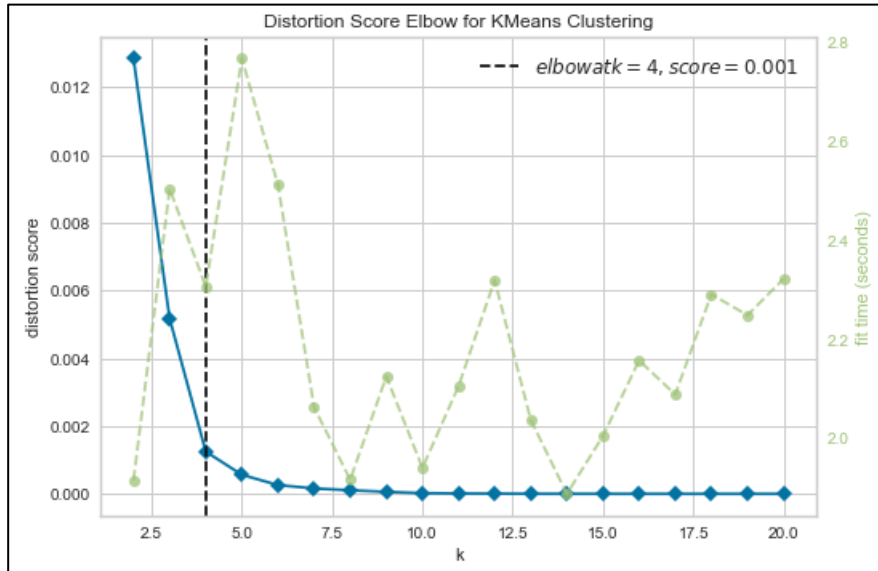


Figure 15: Finding the right K using the Elbow Point

We integrated a model which would fit the error and calculate the distortion score. From the dotted line, we see that the Elbow is at K=4. Moreover, in K-Means clustering, objects that are similar based on a certain variable are put into the same cluster. Neighborhoods that had similar mean frequency of Italian Restaurants were divided into 4 clusters. Each of these clusters were labelled from 0 to 3 as the indexing of labels begin with 0 instead of 1.

	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

Figure 16: Appropriate Cluster Labels were added

After, we merged the venue data with the table above creating a new table which would be the basis for analyzing new opportunities for opening a new Italian Restaurant in Toronto. Then we created a map using the Folium package in Python and each neighborhood was colored based on the cluster label. For example, cluster 2 was purple and cluster 3 was blue.

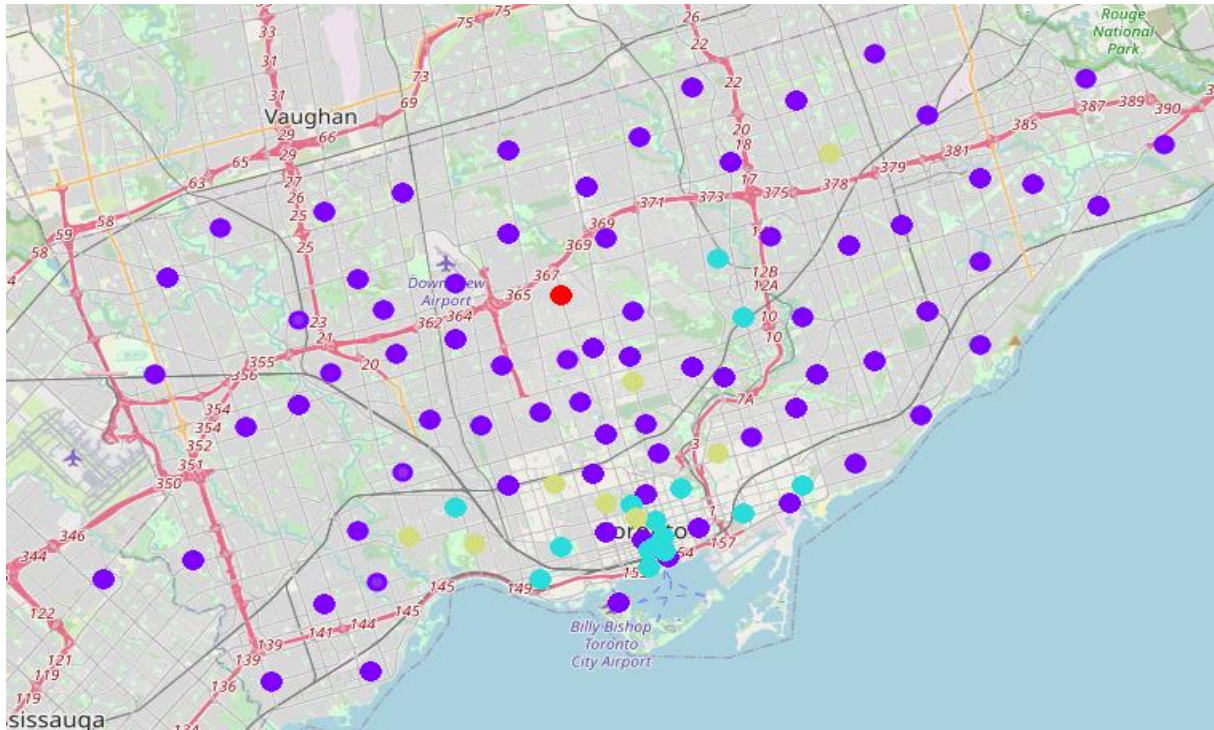


Figure 17: Map with different Clusters

The map above shows the different clusters that had similar mean frequency of Italian restaurants.

Analysis:

We have a total of 4 clusters (0,1,2,3). Before we analyze them one by one let's check the total amount of neighborhoods in each cluster and the average Italian Restaurants in that cluster. From the bar graph that was made using Matplotlib (figure 18), we can compare the number of Neighborhoods per Cluster. We see that Cluster 1 has the least neighborhoods (1) while cluster 2 has the most (70). Cluster 3 has 14 neighborhoods and cluster 4 has only 8. Then we compared the average Italian Restaurants per cluster.

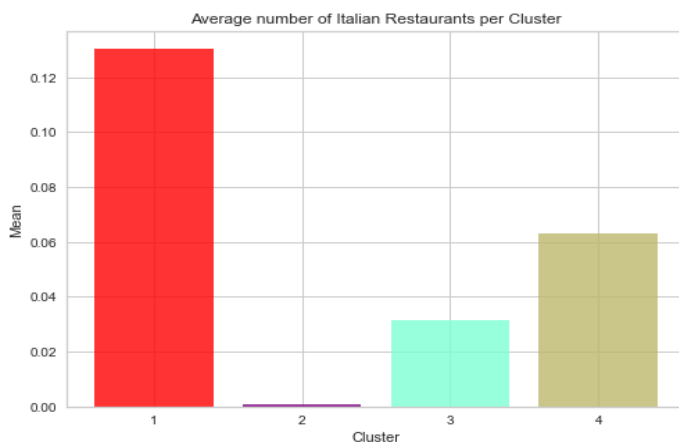


Figure 19: Average Italian restaurant in each neighborhood

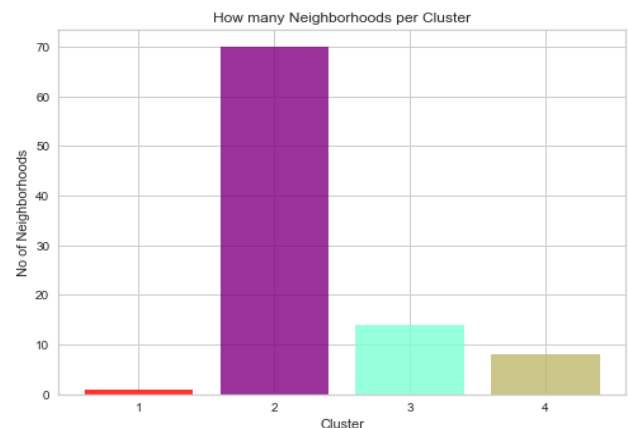


Figure 19: Number of Neighborhoods per cluster

This information is crucial as we can see that even through 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). The average of the average Italian Restaurant made up the data for Figure 18. Also, from the map, we can see that neighborhoods in Cluster 2 are the most sparsely populated. Now let's analyze the Clusters individually (Note: these are just snippets of the data).

Cluster 1 (Red):

	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 1 was in the North York area. Bedford and Lawrence Manor East were the two Neighborhoods that were in that cluster. Cluster 1 had 19 unique Venue locations and out of those only 3 were Italian Restaurants. Cluster 1 had the highest average of Italian Restaurants equating to 0.130435. The reason why the average of Italian Restaurants is the highest is because all these Restaurants are in two neighborhoods, Bedford and Lawrence Manor East.

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	Pumpernickel'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
...

There was a total of 70 neighborhoods, 229 different venues and only 1 Italian Restaurant. Therefore, the average amount of Italian Restaurants that were near the venues in Cluster 2 is the lowest being 0.01. In the map we can see that nodes of Cluster 3 were dispersed all throughout Toronto making it one of the most sparsely populated cluster.

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 had the second to lowest average of Italian Restaurants. Cluster 3 was mainly located in the Downtown Area but also had some neighborhoods in West Toronto, East Toronto and in North York. Neighborhoods such as Ryerson, Toronto Dominion Center, Don Mills, Garden District, Queen's Park and many more were included in this cluster. There was a total of 176 unique venues and out of those 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
...

Cluster 4 venues were located in the Downtown, West, East and Central Toronto areas as well as Scarborough. Neighborhoods such as Central Bay Street, University of Toronto, Central Bay Street and Riverdale were some of the neighborhoods that made up this cluster. There were a total of 91 unique Venues in Cluster 4 with 16 Italian Restaurants. This made up the second highest average of Italian Restaurants in that cluster which was approximately 0.063.

Therefore, the ordering of the average Italian Restaurant in each cluster goes as follows:

1. Cluster 1 (≈ 0.1304)
2. Cluster 4 (≈ 0.0632)
3. Cluster 3 (≈ 0.0317)
4. Cluster 2 (≈ 0.0009)

Discussion:

Most of the Italian Restaurants are in cluster 1 represented by the red clusters. The Neighborhoods located in the North York area that have the highest average of Italian Restaurants are Bedford Park and Lawrence Manor East. Even though there is a huge number of Neighborhoods in cluster 2, there is little to no Italian Restaurant. We see that in the Downtown Toronto area (cluster 3) has the second last average of Italian Restaurants. Looking at the nearby venues, the optimum place to put a new Italian Restaurant is in Downtown Toronto as there are many Neighborhoods in the area but little to no Italian Restaurants therefore, eliminating any competition. The second-best Neighborhoods that have a great opportunity would be in areas such as Adelaide and King, Fairview, etc. which is in Cluster 2. Having 70 neighborhoods in the area with no Italian Restaurants gives a good opportunity for opening a new restaurant. Some of the drawback of this analysis are – the clustering is completely based on data obtained from Foursquare API. Also, the analysis does not take into consideration of the Italian population across neighborhoods as this can play a huge factor while choosing which place to open a new Italian restaurant. This concludes the optimal findings for this project and recommends the entrepreneur to open an authentic Italian restaurant in these locations with little to no competition.

Conclusion:

In conclusion, to end off this project, we had an opportunity on a business problem, and it was tackled in way that it was similar to how a genuine data scientist would do. We utilized numerous Python libraries to fetch the information, to control the content and to break down and visualize those datasets. We have utilized Foursquare API to investigate the settings in neighborhoods of Toronto, get great measure of data from Wikipedia which we scraped with the BeautifulSoup Web scraping Library. We also visualized utilizing different plots present in seaborn and matplotlib libraries. Similarly, we applied AI strategy to anticipate the error given the information and utilized Folium to picture it on a map.

Places that have room for improvement or certain drawbacks gives us that this project can be additionally improved with the assistance of more information and distinctive Machine Learning strategies. Additionally, we can utilize this venture to investigate any situation, for example, opening an alternate cuisine or opening of a Movie Theater and so forth. Ideally, this task acts as an initial direction to tackle more complex real-life problems using data-science.