The Battle of Neighborhoods

Exploring Toronto for Italian Restaurant

Introduction:

As a country of immigrants, when families came to Canada, they typically brought with them a taste of home. These traditional recipes merged with Canadian customs and ingredients to create cuisines that can now be found in all manner of restaurants in Canada **.** 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 or pizzeria in Toronto and are looking for a suitable neighbourhood.

Data Overview:

The data set required for the following project was acquired from three different data sources.

The three data sources are listed below,

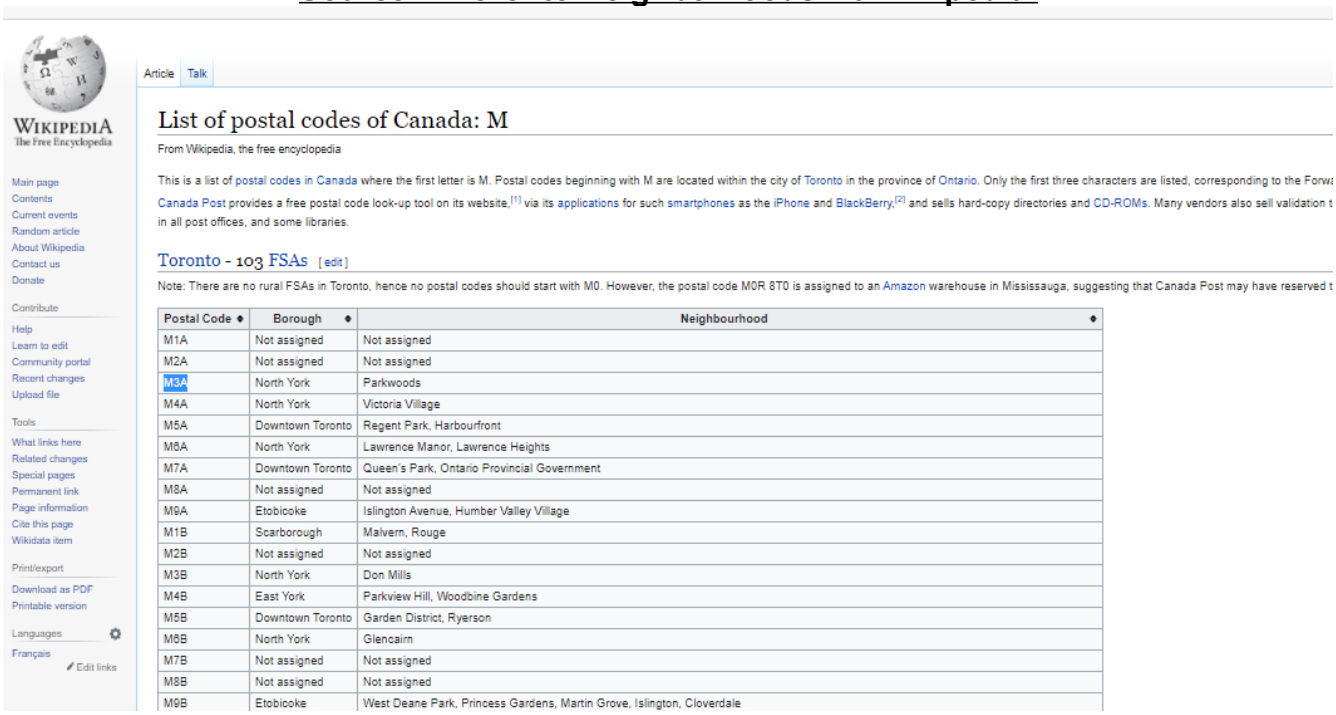
1. A Wikipedia Page to fetch boroughs and neighborhoods of Toronto city.
2. Geographical Location data using Geocoder Package
3. The foursquare api to fetch different public venues in the vicinity of the neighborhood.

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



WIKIPEDIA
The Free Encyclopedia

Article Talk

List of postal codes of Canada: M

From Wikipedia, the free encyclopedia

This is a list of postal codes in Canada where the first letter is M. Postal codes beginning with M are located within the city of Toronto in the province of Ontario. Only the first three characters are listed, corresponding to the Forwa
Canada Post provides a free postal code look-up tool on its website,[1] via its applications for such smartphones as the iPhone and BlackBerry,[2] and sells hard-copy directories and CD-ROMs. Many vendors also sell validation t
in all post offices, and some libraries.

Toronto - 103 FSAs [edit]

Note: There are no rural FSAs in Toronto, hence no postal codes should start with M0. However, the postal code M0R 8T0 is assigned to an Amazon warehouse in Mississauga, suggesting that Canada Post may have reserved t

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
M4B	East York	Parkview Hill, Woodbine Gardens
M5B	Downtown Toronto	Garden District, Ryerson
M6B	North York	Glencairn
M7B	Not assigned	Not assigned
M8B	Not assigned	Not assigned
M9B	Etobicoke	West Deane Park, Princess Gardens, Martin Grove, Islington, Cloverdale

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	Neighbourhood
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 3: Geographical data of Neighborhoods in Toronto

The second source of data provided (https://cocl.us/Geospatial_data) us with the Geographical coordinates of the neighborhoods with the respective Postal Codes which were converted into data frame

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 ‘Client ID’ which will allow me to pull any data

Data Pre-processing

Below are the steps performed

- The first step I performed was to scrape data from the Wikipedia page that consisted of all the boroughs and neighborhoods along with their postal codes and convert them into data frame so that we can do analysis using visualization techniques.
- Dropped rows having missing values in the data frame as missing values can cause discrepancy
- Importing data from a Geospatial_Coordinates.csv file. The .csv file consisted of latitude and longitude coordinates of each postal code. This .csv file was imported into a data frame for ease of analysis in the later stage. Using the Latitude and Longitude

collected from the Geocoder package, we merged the two tables together based on Postal Code.

```
In [19]: # Merging the Data
         df = pd.merge(df, geo_df, on='PostalCode')
         df.head()
```

Out[19]:

	PostalCode	Borough	Neighbourhood	Latitude	Longitude
0	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790
1	M4P	Central Toronto	Davisville North	43.712751	-79.390197
2	M4R	Central Toronto	North Toronto West, Lawrence Park	43.715383	-79.405678
3	M4S	Central Toronto	Davisville	43.704324	-79.388790
4	M4T	Central Toronto	Moore Park, Summerhill East	43.689574	-79.383160

Figure 4: 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.

```
#Get venues for all neighborhoods in our dataset
toronto_venues = getNearbyVenues(names=df_toronto['Neighbourhood'],
                                 latitudes=df_toronto['Latitude'],
                                 longitudes=df_toronto['Longitude'])
```

	Neighbourhood	Neighbourhood Latitude	Neighbourhood 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 5: Local Venues near the respective Neighborhood

Data Exploration

Now after cleansing the data, the next step was to analyze the data .We then created a map using folium and color coded each Neighborhood depending on what Borough it was located in. 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. We then merged the Foursquare Venue data with the Neighborhood data which then gave us the nearest Venue for each of the Neighborhoods.

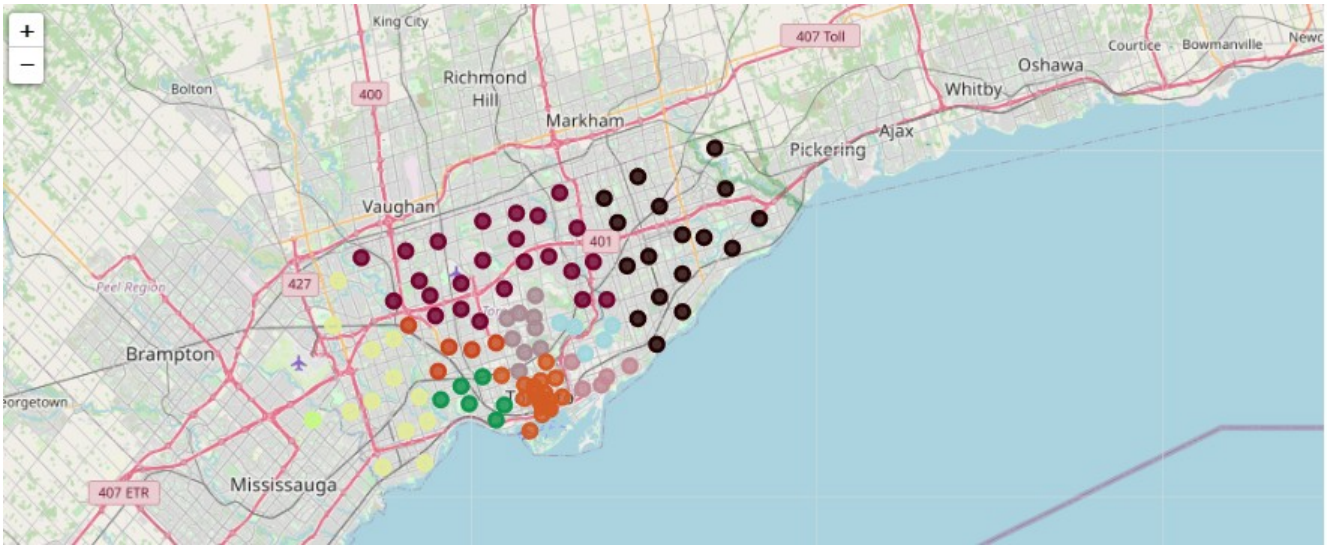


Figure 9: Toronto Neighborhoods

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.

Figure 10: Venue table merged with Neighborhood data

```
In [94]: toronto_venues.head()
```

Out[94]:

	Neighbourhood	Neighbourhood Latitude	Neighbourhood 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

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.

	Neighbourhoods	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.

	Neighbourhoods	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	0.045455

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 simpler to analyze

	Neighbourhoods	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	0.090909

Figure 12: 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 4 no of clusters.

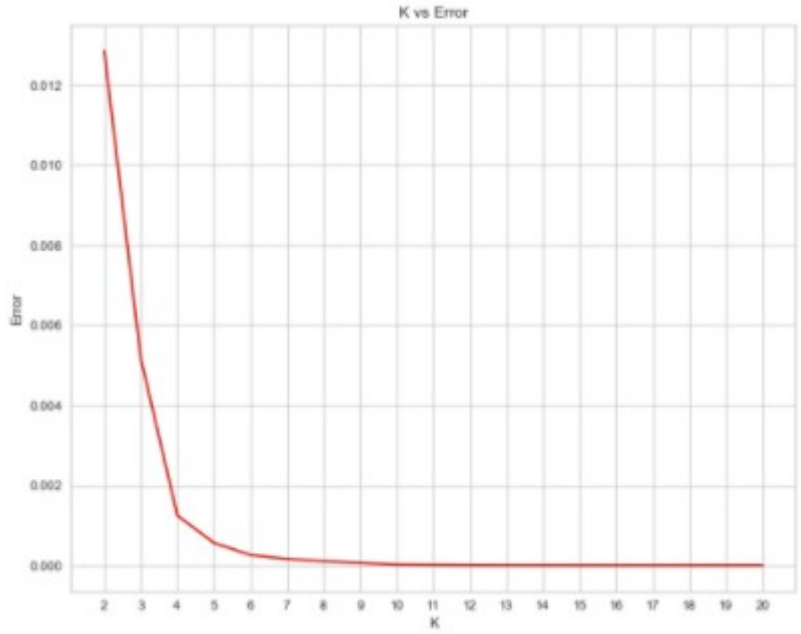


Figure 13: Finding the K vs Error Values

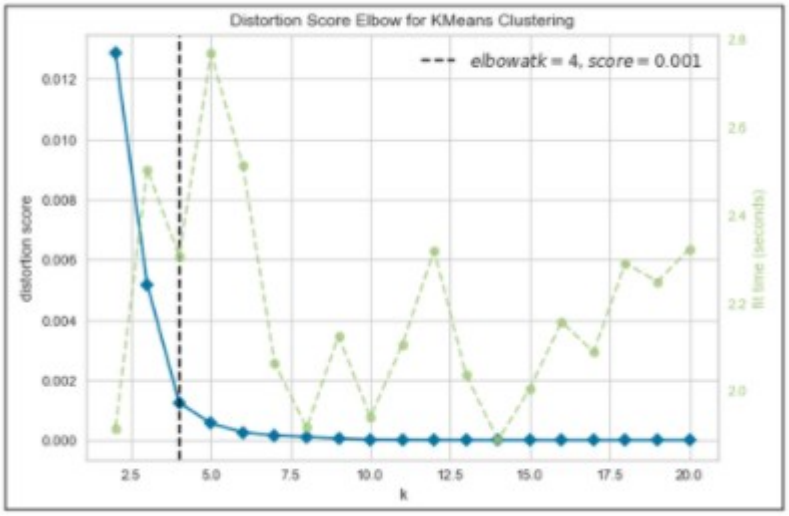


Figure 14: Finding the right K using the Elbow Point

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.

	Neighbourhood	Italian Restaurant	Cluster Labels
0	Agincourt	0.000000	0
1	Alderwood, Long Branch	0.000000	0
2	Bathurst Manor, Wilson Heights, Downsview North	0.000000	0
3	Bayview Village	0.000000	0
4	Bedford Park, Lawrence Manor East	0.090909	3

Figure 15: 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 turquoise blue.

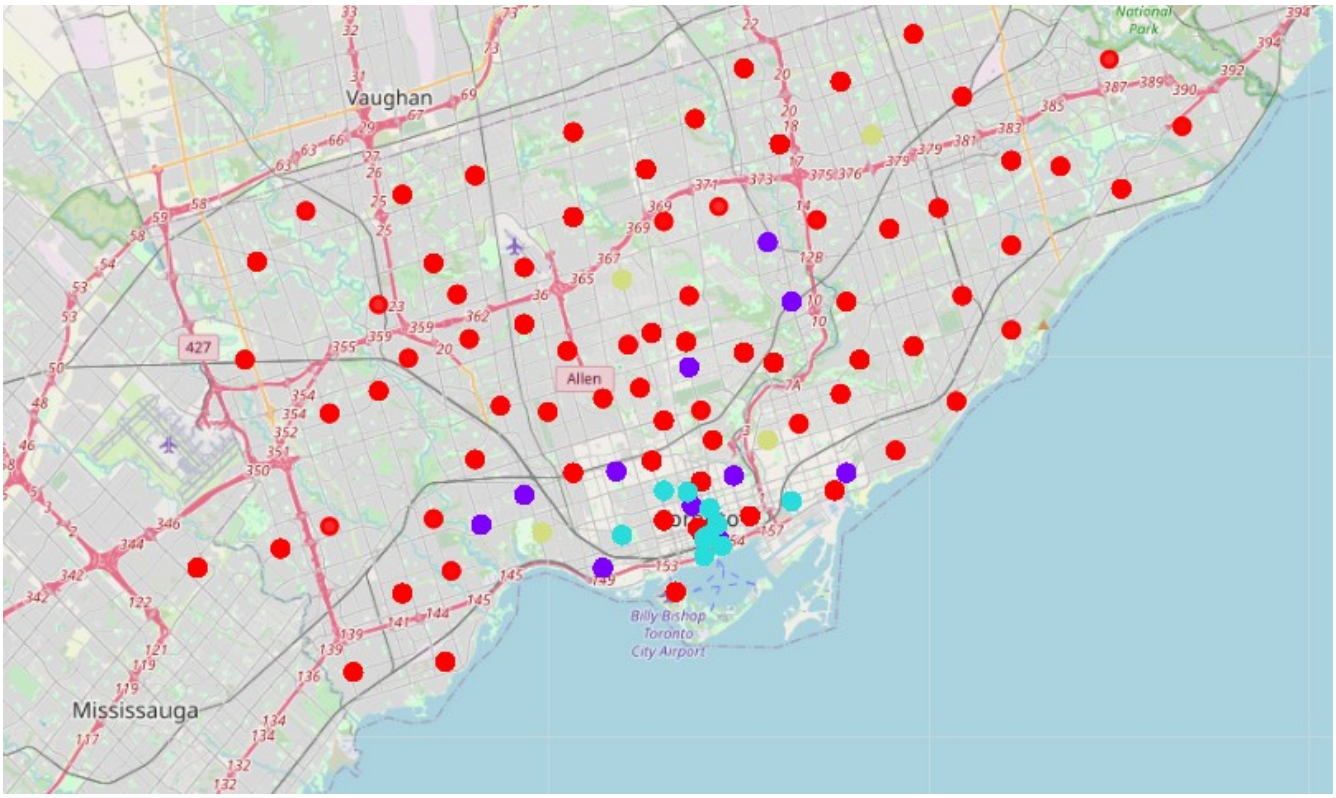


Figure 16: Map with different Clusters

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 lets 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 17) , we can compare the number of Neighborhoods per Cluster. We see that Cluster 4 has the least neighborhoods while cluster 1 has the most (72). Cluster 2 and 3 has 10 neighborhoods each Then we compared the average Italian Restaurants per cluster.

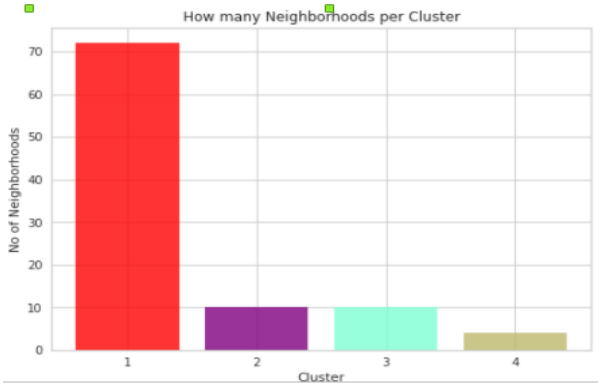


Figure 17: Number of Neighborhoods per cluster

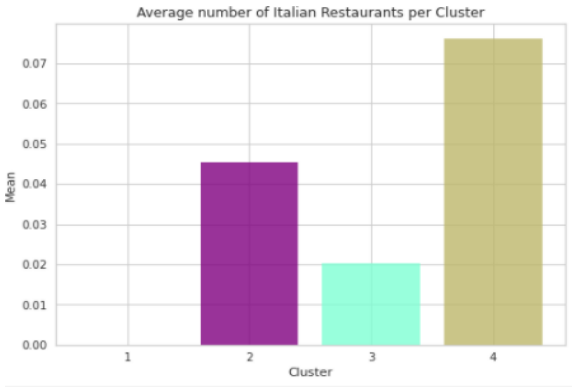


Figure 18 : Average Italian restaurant in each neighborhood

This information is crucial as we can see that even through there is only 4 neighborhood in Cluster 4, it has the highest number of Italian Restaurants while Cluster 1 has the most neighborhoods but has the least average of Italian Restaurants . Also, from the map, we can see that neighborhoods in Cluster 4 is the most sparsely populated. Now let's analyze the Clusters individually (Note: these are just snippets of the data)

Cluster 1 (Red):

	Borough	Neighbourhood	Italian Restaurant	Cluster Labels	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Central Toronto	Lawrence Park	0.0	0	43.728020	-79.388790	TTC Bus #162 - Lawrence-Donway	43.728026	-79.382805	Bus Line
1	Central Toronto	Lawrence Park	0.0	0	43.728020	-79.388790	Zodiac Swim School	43.728532	-79.382860	Swim School
2	Central Toronto	Lawrence Park	0.0	0	43.728020	-79.388790	Lawrence Park Ravine	43.726963	-79.394382	Park
3	Central Toronto	Davisville North	0.0	0	43.712751	-79.390197	900 Mount Pleasant - Residents Gym	43.711671	-79.391767	Gym / Fitness Center
4	Central Toronto	Davisville North	0.0	0	43.712751	-79.390197	Sherwood Off-leash Dog Park	43.715711	-79.390118	Dog Run

Cluster 1 is the Central Toronto ,YorK and other areas. Lawrence Park, Davisville North , Runnymede, The Junction North etc are the Neighborhoods that were in that cluster. Cluster 1 had the lowest average of Italian Restaurants equating with 72 neighbourhood as seen in the Figure 17. As seen in Figure 16, cluster 1 is spread across Toronto and has the maximum neighbours . This seems like a good opportunity given the maximum neighbours and minimum existing Italian restaurant.

Cluster 2 (Purple) :

	Borough	Neighbourhood	Italian Restaurant	Cluster Labels	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
248	Downtown Toronto	Christie	0.062500	1	43.669542	-79.422564	Marian Engel Park	43.673754	-79.423988	Park
256	Downtown Toronto	Christie	0.062500	1	43.669542	-79.422564	Actinolite	43.667858	-79.428054	Restaurant
263	Downtown Toronto	Christie	0.062500	1	43.669542	-79.422564	Foto Grocery	43.667979	-79.428217	Grocery Store
262	Downtown Toronto	Christie	0.062500	1	43.669542	-79.422564	Garrison Creek Park	43.671690	-79.427805	Park
261	Downtown Toronto	Christie	0.062500	1	43.669542	-79.422564	Fiesta Farms	43.668471	-79.420485	Grocery Store

There was a total of 10 neighborhoods of 19 venue category . It is concentrated around Downtown Toronto, West Toronto etc. This cluster has the 2nd highest no of Avg. number of Italian restaurant at ≈ 0.045

Cluster 3 (Turquoise) :

	Borough	Neighbourhood	Italian Restaurant	Cluster Labels	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Downtown Toronto	Garden District, Ryerson	0.020000	2	43.657162	-79.378937	JOEY Eaton Centre	43.656094	-79.381878	New American Restaurant
1	Downtown Toronto	Garden District, Ryerson	0.020000	2	43.657162	-79.378937	BMV Books	43.657047	-79.381661	Bookstore
2	Downtown Toronto	Garden District, Ryerson	0.020000	2	43.657162	-79.378937	lululemon athletica	43.653286	-79.380764	Clothing Store
3	Downtown Toronto	Garden District, Ryerson	0.020000	2	43.657162	-79.378937	Ryerson Square	43.656988	-79.376896	Other Great Outdoors
4	Downtown Toronto	Garden District, Ryerson	0.020000	2	43.657162	-79.378937	Disney Store	43.654248	-79.381232	Toy / Game Store

Cluster 3 has the same no. of neighbourhood as Cluster 2 (10) . Cluster 3 was mainly located in the Downtown and West Toronto like Garden District, Little Portugal etc. The average of Italian Restaurants in this cluster which was ≈ 0.020. There seems to be a good opportunity to grow in this cluster as well

Cluster 4 (Dark Khaki):

	Borough	Neighbourhood	Italian Restaurant	Cluster Labels	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	East Toronto	The Danforth West, Riverdale	0.068182	3	43.679557	-79.352188	Kitchen Stuff Plus	43.678613	-79.346422	Furniture / Home Store
1	East Toronto	The Danforth West, Riverdale	0.068182	3	43.679557	-79.352188	Factory Girl	43.676693	-79.356299	American Restaurant
2	East Toronto	The Danforth West, Riverdale	0.068182	3	43.679557	-79.352188	Katsu Japanese Restaurant	43.678619	-79.347024	Sushi Restaurant
3	East Toronto	The Danforth West, Riverdale	0.068182	3	43.679557	-79.352188	Carrot Commons	43.677485	-79.353076	Restaurant
4	East Toronto	The Danforth West, Riverdale	0.068182	3	43.679557	-79.352188	Second Cup	43.677232	-79.352898	Coffee Shop

Cluster 4 venues were located mainly in East and West Toronto area were some of the neighborhoods that made up this cluster. This cluster has the highest average of Italian Restaurants which is ≈ 0.075 . Based on the analysis , opportunity to grow in this cluster is minimimum as there are only 4 neighbourhood and high average of Italian restaurant. However we can look into this cluster using population demographics data to understand if there are any reasons why there are so many Italian restaurants here and make our decision

Therefore, the ordering of the average Italian Restaurant in each cluster as mentioned is as below

1. Cluster 1 (Red) ≈ 0.00
2. Cluster 3 (Turquoise) ≈ 0.045
3. Cluster 2 Purple) ≈ 0.020
4. Cluster 4 (Dark Kaki) ≈ 0.075

Discussion- Final Recommendation :

Even though there is a huge no. of Neighborhoods in cluster 1 (70+), there is little to no Italian Restaurant, eliminating any competition. Therefore this will be our first recommendation for opening a new Italian restaurant

The second best Neighborhoods that have a great opportunity would be in areas such as Garden street and Little portugal which is in Cluster 3. Having 10 neighborhoods in the area with only few Italian Restaurants gives a good opportunity as well

Cluster 4 has highest no of italian restaurant with only 4 neighbhourhood. However we can further analyze the population type here to understand the reasons high average of Italian restaurants in this cluster which could be high density of Italian population / affluent population who may have preferences for Italian restaurant

Our current analysis does not take into consideration of the Italian population or other population demographics across neighborhoods however this his can play a huge factor in selecting the neighbourhood.

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. Nonetheless, if the food is authentic, affordable and good taste, it will have good prospects anywhere

Conclusion:

During the course of this capstone project, I was able to apply different data science techniques and tools that I learned in the IBM Data Science course. This helped me unearth meaningful insights from the data analysis that I did on the Toronto data set.
Finally to conclude this project, I have got a chance to work on a business problem like how a real like data scientists would do. I have used many python libraries to fetch the data , to manipulate the contents & to analyze and visualize those datasets. I have made use of Foursquare API to explore the venues in neighborhoods of Toronto, then get data from Wikipedia which was scraped with help of Wikipedia python library and visualized using various plots present in seaborn & matplotlib. I also applied machine learning technique to predict the output given the data and used Folium to visualize it on a map.

Some of the drawbacks or areas of improvements shows us that this analysis can be further improved with the help of more data and different machine learning technique. Taking account of Italian population data and other life style choice data (population demographics) can also help improve the prediction. Additionally this data can be used for other scenarios for other entrepreneur ventures in Toronto with few tweeks in the code. Even with these limitations , this project helps acts as initial guidance to take more complex real-life challenges using data-science.

Thank you !