A. Introduction

This project aims to utilize all Data Science Concepts learned in the IBM Data Science Professional Course. I start by defining a Business Problem, then discuss the data that will be utilized. Using that data, I will be able to analyze it using Machine Learning tools. In this project, I will go through all the processes in a step by step manner from problem designing, data preparation to final analysis.

Table of Contents

1. Identifying the Business Problem (Introduction):

Food is a big deal in Toronto, a very big deal. If you are a foodie, Toronto should be a “must-visit” destination on your bucket list.

Toronto is one of the most densely populated areas in Canada. It is located in the province of Ontario, Canada and it is also the capital of Ontario. It has a recorded population of over 6 million people. It is not only the most populous city in Canada, it is also the fourth most populous city in North America. Toronto is an international center of business, finance, arts, and culture, and I would add food. It is recognized as one of the most multicultural and cosmopolitan cities in the world. It is without a doubt a foodie destination.

For over 200 years, Immigrants from South America, Asia, Africa, the Caribbean and Europe have been bringing their culinary cultures to Toronto. Over time, the mixing and blending of these cultures give birth to food that can only be found where cultures harmoniously reside. One such example is Rasta Pasta restaurant, which is collision of Italian and Jamaican flavors.

There are also traditional Canadian foods that are not just cheese, curd and gravy, e.g Pow Wow Café which features the First People of Canada cuisine such as Ojibwe-style fry bread etc. There is also Antler Kitchen and Bar that features seasonal menus of authentic Canadian game such as bison, venison, duck and wild-caught halibut etc.

Multiculturalism is seen through the various neighborhoods including; Chinatown, Corso Italia, Little India, Kensington Market, Little Italy, Koreatown and many more.

2. Target Audience

This project is aimed towards Entrepreneurs or Business owners, who may want to understand the food climate in Toronto and explore their options of investing in the food business in Toronto. The analysis will provide vital information that can be used by the target audience.

3. Data Overview

The data required will be a combination of CSV files that have been prepared for the purposes of the analysis from multiple sources. These files will provide the list of neighborhoods in Toronto (via Wikipedia), the Geographical location of the neighborhoods (via Geocoder package) and Venue data pertaining to restaurants (via Foursquare).

3.1 — Data acquisition:

Source 1: Toronto Neighborhoods via Wikipedia

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

Article Talk Read

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,

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 v match addresses and postal codes. Hard-copy directories can also be consulted 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 Canad

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

1. https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

The Wikipedia site 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

Source 2: Geographical Location data using Geocoder Package

	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

Fig 3 Conversion of the file into Pandas data frame

2. https://cocl.us/Geospatial_data

The second source of data provided us with the Geographical coordinates of the neighborhoods with the respective Postal Codes. The file was in CSV format, so we had to attach it to a Pandas data frame(shown in figure 3).

Source 3: Venue Data using Foursquare

4. Methodology

4.1 — Data Cleansing

After all the data was collected, they were put into data frames. I then cleansed and merged the data so it can be analyzed. In the data retrieved 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.

2. Multiple neighborhoods can share the same postal code, e.g. For example, M5A is listed twice and has two neighborhoods: Harbourfront and Regent Park. These two rows were combined into one row with the neighborhoods separated with a comma.

3. Some borough were not assigned a neighborhood, then the neighborhood will be the same as the borough.

After the implementation of the following assumptions, the rows were grouped based on the 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

Fig. 4 Grouped together by 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

Fig 5 Merged tables 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.

4.2 — Data Exploration

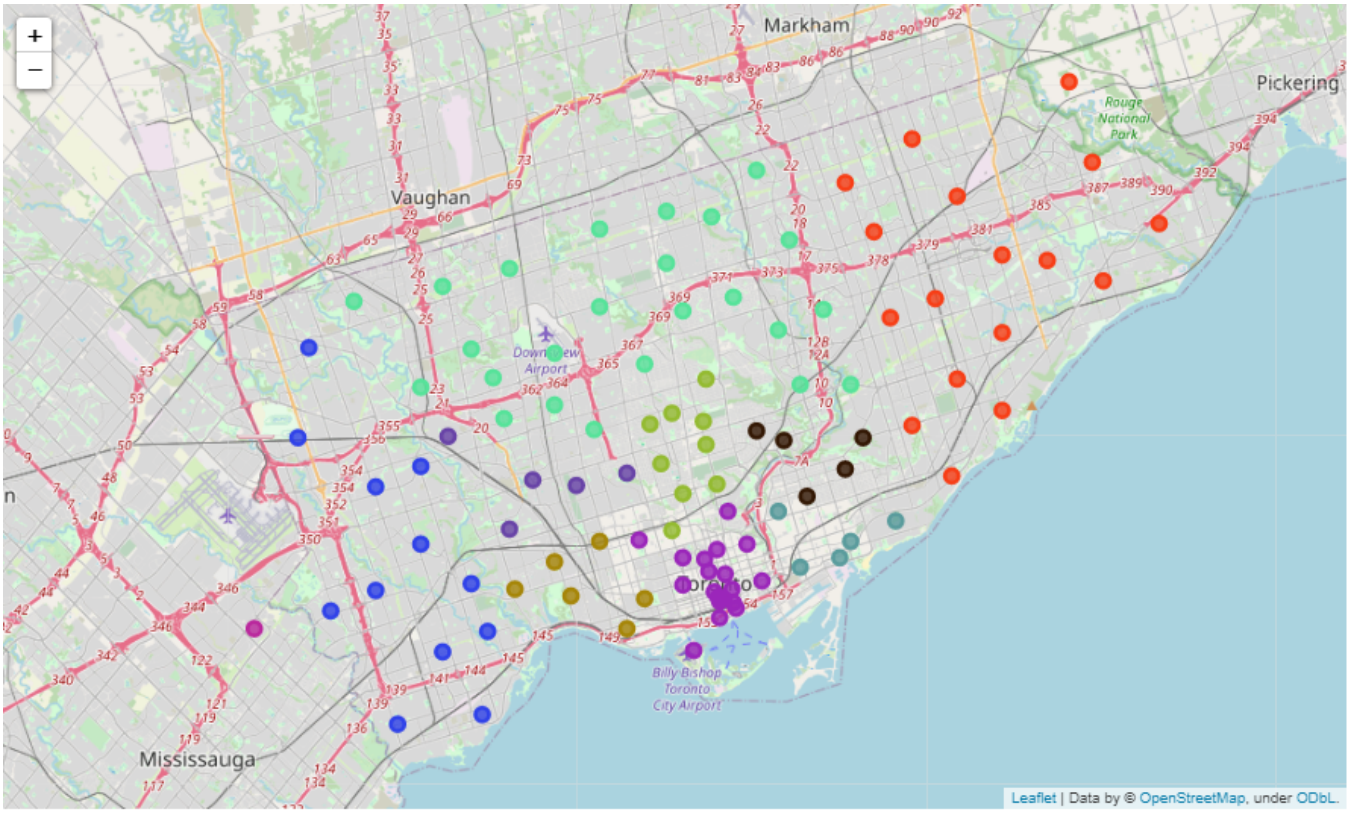
Now after cleansing the data, I analyzed it. I then created a map using Folium and color-coded each Neighborhood depending on what Borough it was located.

Next, I used the Foursquare API to get a list of all the Venues in Toronto which included Parks, Schools, Café Shops, Asian Restaurants etc. This data was crucial to analyze the number and different types Restaurants in Toronto. There was a total of 8 Caribbean 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.

```
In [47]: toronto_venues.head()
```

Out[47]:

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Malvern, Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
2	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	Scarborough Historical Society	43.788755	-79.162438	History Museum
3	Guildwood, Morningside, West Hill	43.763573	-79.188711	RBC Royal Bank	43.766790	-79.191151	Bank
4	Guildwood, Morningside, West Hill	43.763573	-79.188711	G & G Electronics	43.765309	-79.191537	Electronics Store



4.3 — Machine Learning

Then to analyze the data I 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 neighbourhoods, individual venues were turned into the frequency at how many of those Venues were located in each neighbourhood.

Out[48]:

Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Agincourt	4	4	4	4	4	4
Alderwood, Long Branch	8	8	8	8	8	8
Bathurst Manor, Wilson Heights, Downsview North	23	23	23	23	23	23
Bayview Village	4	4	4	4	4	4
Bedford Park, Lawrence Manor East	24	24	24	24	24	24
...	...	...	...	...	...	...
Willowdale, Willowdale West	5	5	5	5	5	5
Woburn	4	4	4	4	4	4
Woodbine Heights	6	6	6	6	6	6
York Mills West	2	2	2	2	2	2
York Mills, Silver Hills	1	1	1	1	1	1

96 rows × 6 columns

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

Out[117]:

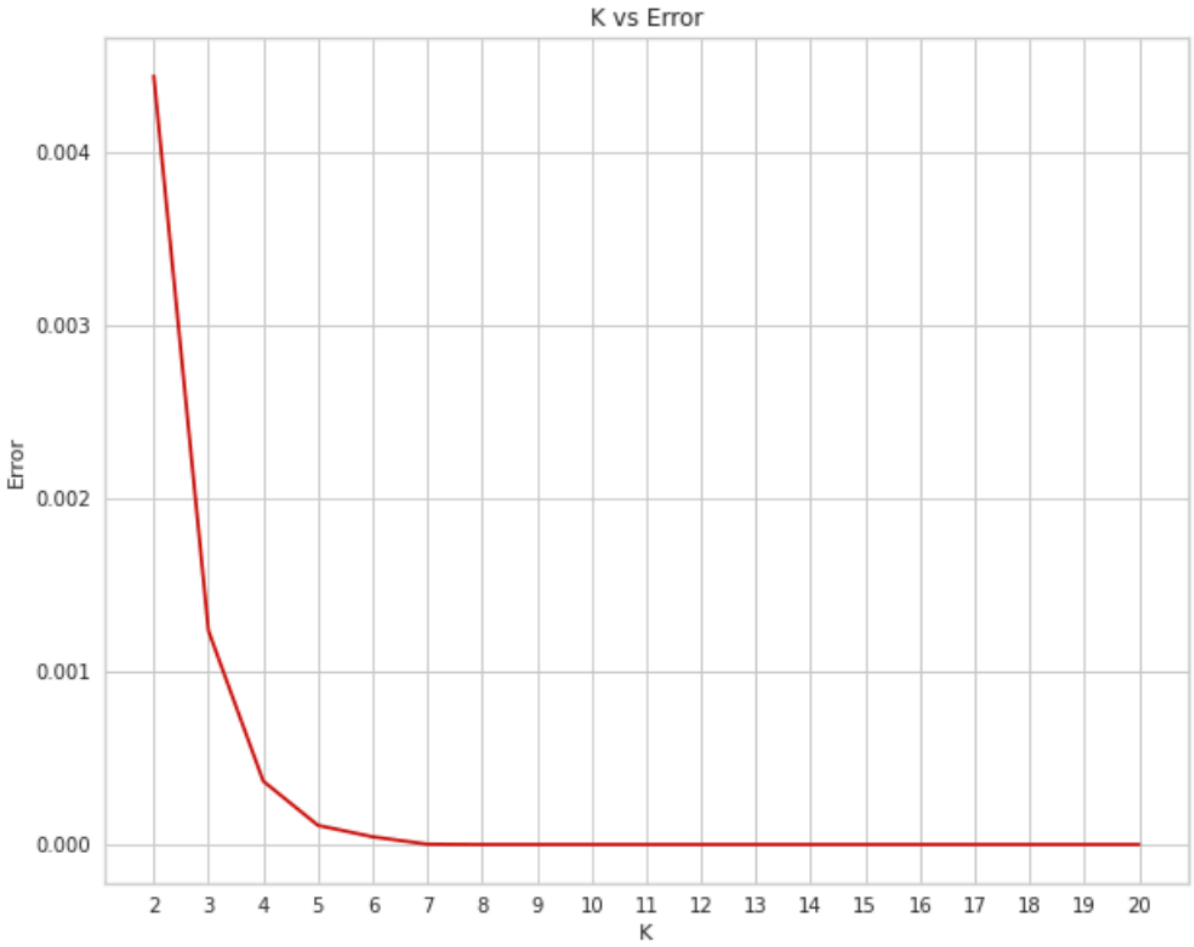
	Neighbourhoods	Accessories Store	Adult Boutique	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	...	Train Station	Truck Stop	Vegetarian / Vegan Restaurant	Videc Game Store
0	Malvern, Rouge	0	0	0	0	0	0	0	0	0	...	0	0	0	0
1	Rouge Hill, Port Union, Highland Creek	0	0	0	0	0	0	0	0	0	...	0	0	0	0
2	Rouge Hill, Port Union, Highland Creek	0	0	0	0	0	0	0	0	0	...	0	0	0	0
3	Guildwood, Morningside, West Hill	0	0	0	0	0	0	0	0	0	...	0	0	0	0
4	Guildwood, Morningside, West Hill	0	0	0	0	0	0	0	0	0	...	0	0	0	0

5 rows × 268 columns

K-Means Clustering

To make the analysis more interesting, I wanted to cluster the neighbourhoods based on the neighbourhoods that had similar averages of Caribbean Restaurants in that Neighborhood. To do this I used **K-Means** clustering. To get our optimum K value that was neither overfitting or underfitting the model, I used the **Elbow Point** Technique. In this technique, I 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 the sharpest turn. In our case, we had the Elbow Point at K = 5. That means we will have a total of 5 clusters.

Fig 5



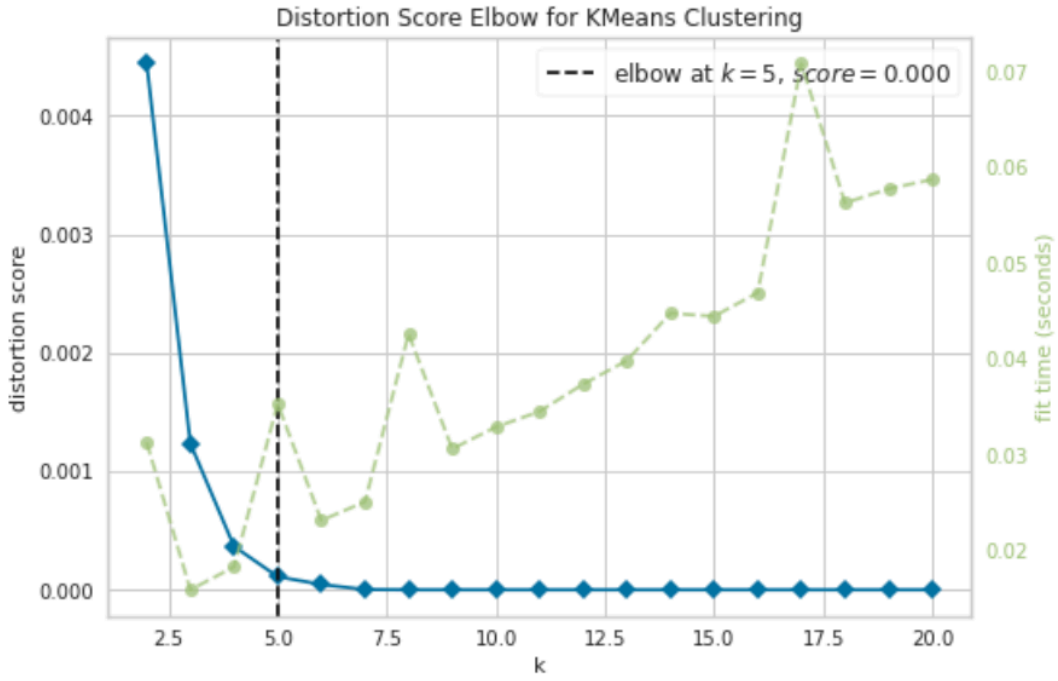
Then we used a model that accurately pointed out the optimum K value. We imported '*KElbowVisualizer*' from the *Yellowbrick package.* Then we fit our K-Means model above to the Elbow visualizer.

```
# Instantiate the clustering model and visualizer
model = KMeans()
visualizer = KElbowVisualizer(model, k=(2,21))

visualizer.fit(X)        # Fit the data to the visualizer
visualizer.show()
```

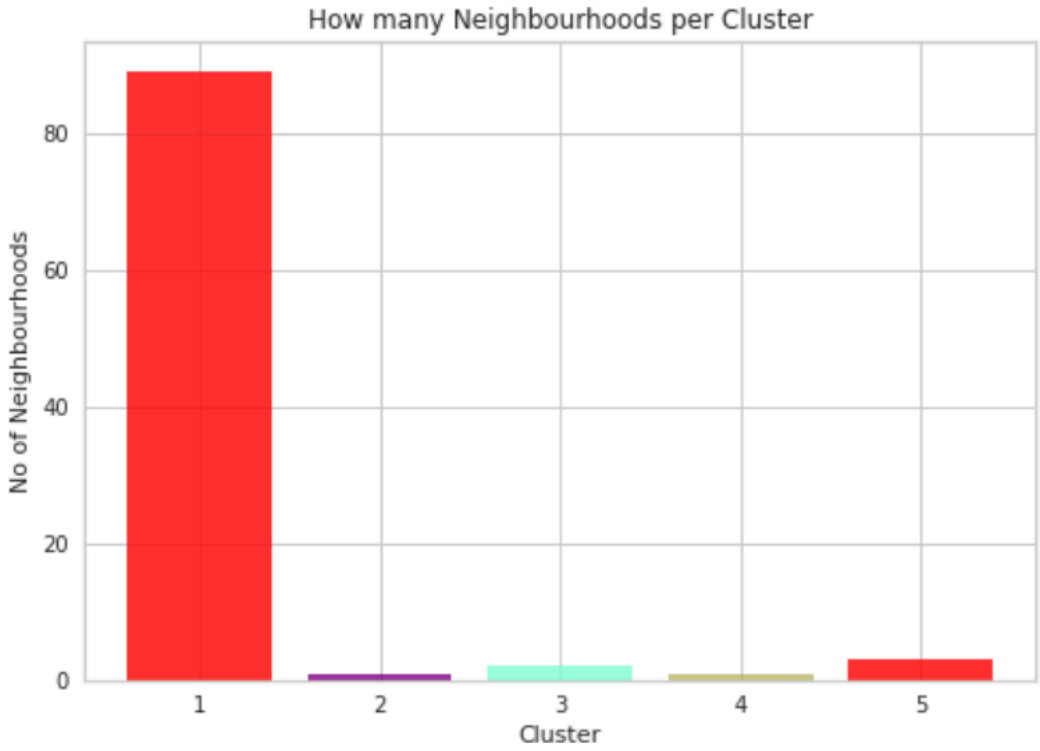
This gave the model below:

```
self.estimator.fit(X, **kwargs)
```



4.4 — Data Analysis

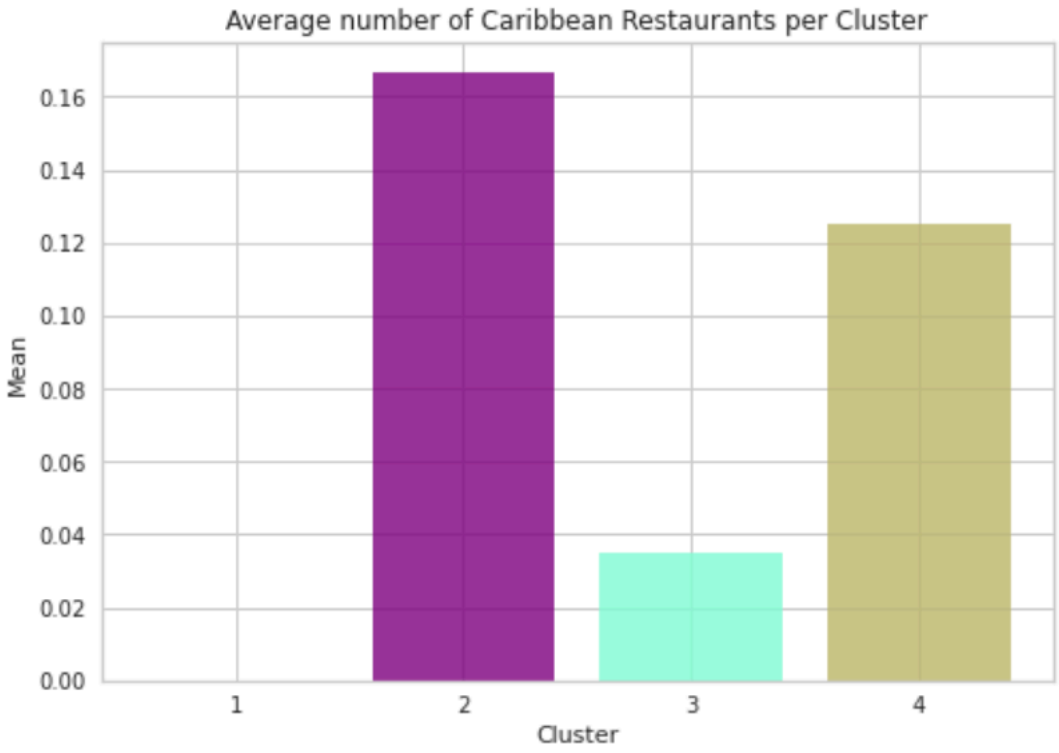
I have a total of 5 clusters (0,1,2,3,4,). Before we analyze them one by one let's check the total amount of neighbourhoods in each cluster and the average Caribbean Restaurants in that cluster. From the bar graph that was made using Matplotlib , we can compare the number of Neighborhoods per Cluster. We see that Cluster 1 has the most neighbourhoods (80) while cluster 2 and 4 has the least(1). Cluster 3 has 14 neighbourhoods and cluster 4 has only 8. Then we compared the average Caribbean Restaurants per cluster.



I integrated a model that would fit the error and calculate the distortion score. From the dotted line, we see that the Elbow is at K=5. Moreover, in K-Means clustering, objects that are similar based on a certain variable are put into the same cluster. Neighbourhoods that had a similar mean frequency of Caribbean Restaurants were divided into 4 clusters. Each of these clusters was labelled from 0 to 3 as the indexing of labels begins with 0 instead of 1.

After, I 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 Caribbean Restaurant in Toronto. Then I created a map using the

Folium package in Python and each neighbourhood was coloured based on the cluster label.



Cluster Analysis

This information is crucial as we can see that Cluster 2 has the most neighbourhoods and also has the highest average of Caribbean Restaurants (0.16).

Cluster 1(Red):

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 there were no Caribbean Restaurants. Cluster 2 had the highest average of Caribbean Restaurants equating to 0.16. The reason why the average of Caribbean Restaurants is the highest is that all these Restaurants are in two neighbourhoods, Bedford and Lawrence Manor East.

Cluster 2 (Purple) :

	Borough	Neighbourhood	Caribbean Restaurant	Cluster Labels	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	North York	Northwood Park, York University	0.166667	1	43.76798	-79.487262	MUSE Massage Spa	43.765686	-79.489318	Massage Studio
1	North York	Northwood Park, York University	0.166667	1	43.76798	-79.487262	Carribean Heat	43.764155	-79.490227	Caribbean Restaurant
2	North York	Northwood Park, York University	0.166667	1	43.76798	-79.487262	Tim Hortons	43.764289	-79.488790	Coffee Shop
3	North York	Northwood Park, York University	0.166667	1	43.76798	-79.487262	Fox & Fiddle	43.763795	-79.488497	Bar
4	North York	Northwood Park, York University	0.166667	1	43.76798	-79.487262	Bad Boy Furniture - North York	43.764314	-79.486588	Furniture / Home Store
5	North York	Northwood Park, York University	0.166667	1	43.76798	-79.487262	Lastman's Bad Boy	43.763878	-79.486435	Furniture / Home Store

There was a total of 70 neighbourhoods, 229 different venues and only 1 Caribbean Restaurant.

Cluster 3 (Turquoise):

Cluster 3 had the second to second lowest average of Caribbean Restaurants. Cluster 3 was mainly located in the Downtown area but also had some neighbourhoods in West Toronto, East Toronto and in North York. Neighbourhoods 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 there was an average of 0.04 Caribbean Restaurants.

Cluster 4 (Dark Khaki):

Cluster 4 venues were located in the Downtown, West, East and Central Toronto areas as well as Scarborough. Neighbourhoods such as Central Bay Street, University of Toronto, Central Bay Street and Riverdale were some of the neighbourhoods that made up this cluster. There were a total of 91 unique Venues in Cluster 4 with an average of 0.12 Caribbean Restaurants.

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

1. Cluster 2 (≈0.16666)

2. Cluster 4 (≈0.125)

3. Cluster 3 (≈0.0400)

4. Cluster 1 (≈0.0000)

5. Discussion:

Most of the Caribbean Restaurants are in cluster 2 represented by the purple clusters. This is probable because of the large number of neighbourhoods located in this cluster.

In general there are not many Caribbean restaurants in Toronto. Looking at the nearby venues, the optimum place to put a new Caribbean Restaurant in the North York area as there are many Neighborhoods in the area but no Caribbean Restaurants, therefore, eliminating any competition. The second-best Neighborhoods that have a great opportunity would be Downtown Toronto since there are little to no Caribbean Restaurants.

Some of the cons of this analysis are — the clustering is completely based on data obtained from the Foursquare API. Also, the analysis does not take into consideration of the Caribbean population across neighbourhoods as this can play a huge factor while choosing which place to open a new Caribbean restaurant. This concludes the optimal findings for this project and recommends the entrepreneur to open an authentic Caribbean restaurant in these locations with little to no competition.

6. Conclusion

To conclude this project, I had an opportunity to utilize numerous Python libraries to fetch the information, control the content and break down and visualize those datasets. I have utilized Foursquare API to investigate the settings in the neighbourhoods of Toronto. I used the Beautifulsoup Web scraping Library to scrape data from Wikipedia . I also visualize the data using different plots present in seaborn and Matplotlib libraries. Similarly, I applied AI strategy to anticipate the error given the information and utilized Folium to picture it on a map.