Battle of Neighborhood

Restaurant Business in Toronto

IBM Data Science Professional Certificate – Capstone Project

SHROWTHI Bharadwaja

1. Introduction for the Business Case:

Toronto is the capital city of the Canadian province of Ontario. It is the most populous city in Canada and the fourth most populous city in North America with an estimated population of ~6 million. The cuisine of Toronto reflects Toronto's size and multicultural diversity. Canadian cuisine varies widely depending on the regions of the nation. The four earliest cuisines of Canada have indigenous, English, Scottish and French roots. The traditional cuisine of English Canada is closely related to British cuisine.

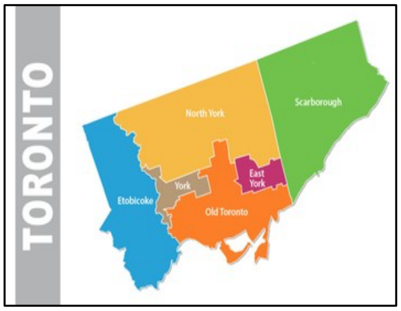


Figure 1: Boroughs map in Toronto, Canada

Overtime, with subsequent waves of immigration in the 19th and 20th centuries, Canadian food has been shaped and impacted by those of indigenous people, settlers, and immigrants. Toronto is well known for its great food. Canadian culinary includes an array of international cuisines influenced by multiculturalism of the town. Different ethnic neighborhoods throughout the city focus on variety of cuisines. Examples: Chinese, Indian, Italian, Japanese, Caribbean, Jewish, Vegetarian/Vegan, American, Mediterranean, Fast Food Centers etc.

Figure 2: One of the food festival emblems in Canada.

A number of culinary festivals take place in Toronto each year. Any trip to Toronto is incomplete without checking out its food tours. Going on any one of these, will without a doubt, leave you satiated.

In this Capstone project, the Boroughs and neighborhoods in the city of Toronto is analyzed for a suitable location to start a new restaurant. For this purpose, the cuisine style for the new restaurant is generalized so that the potential entrepreneurs can have the choice for greater success with consistent return on investments.

Data science is the process of using algorithms, methods and systems to extract knowledge and insights from structured and unstructured data. It applies advanced analytics and machine learning to help users predict and optimize business outcomes. In this project, an unsupervised K-Means Cluster algorithm is used to analyze the existing restaurant businesses for a given neighborhood area in Toronto. Based on this analysis, location recommendations for new restaurant business is suggested. In addition, possible improvements are also suggested to consider for a better data science model.

2. Data Sources and their description:

Three major data sources have been identified for this analysis.

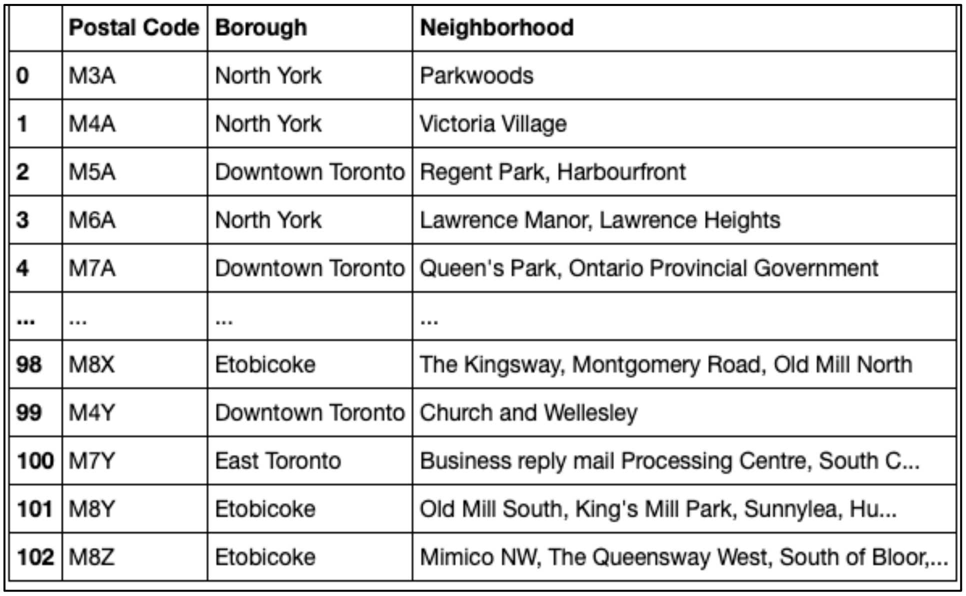
1. Neighborhoods of city of Toronto with Boroughs and venues information. Web scrapping method is implemented on the Toronto's Wikipedia webpage to extract this information.
   https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&oldid=1008658627'
2. Latitude and longitude geospatial data of the neighborhoods are collected from the https://cocl.us/Geospatial_data

3. Venue data, in particular data related to restaurants is extracted using Foursquare API. Foursquare provides a count on the types of cuisine according to a predefined set of categories as documented on its website https://developer.foursquare.com/docs/resources.

Subsequent methodology involves data cleaning, data wrangling to map visualization of Toronto.

3. Data Cleaning, Wrangling and Preprocessing:

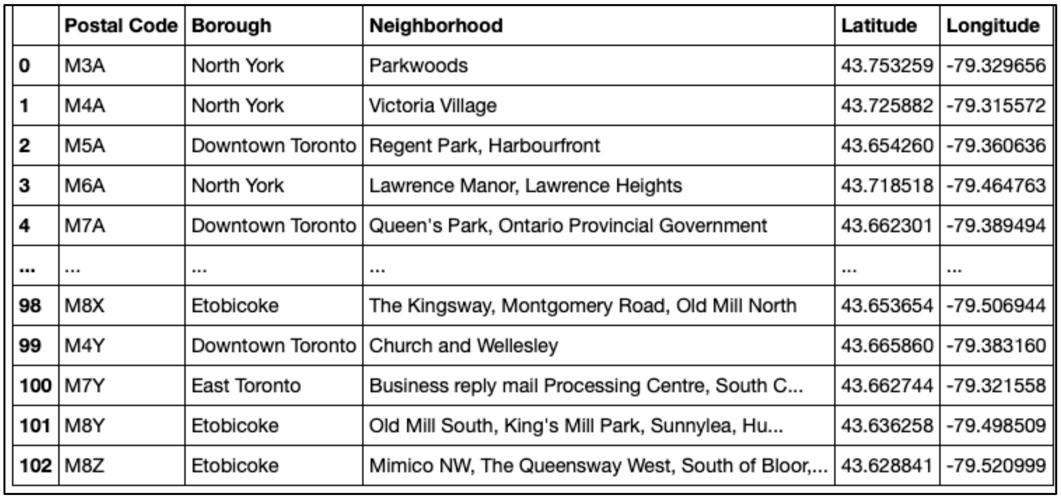
Canadian Postal Codes and Neighborhood are scraped from Wikipedia website using Beautiful soup. All the Neighborhoods that are not assigned are removed. This is followed by combining the neighborhoods with same postal codes. Neighborhoods that are not assigned are replaced with their Borough's name. The resultant Pandas data frame size is 103 rows x 3 columns, as shown in Table 1.



	Postal Code	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
...	...	...	...
98	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North
99	M4Y	Downtown Toronto	Church and Wellesley
100	M7Y	East Toronto	Business reply mail Processing Centre, South C...
101	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...
102	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...

Table 1: Canadian Neighborhoods with respective postal codes

The above data frame is merged with Toronto’s geospatial data as shown below.



	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
...	...	...	...	...	...
98	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North	43.653654	-79.506944
99	M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160
100	M7Y	East Toronto	Business reply mail Processing Centre, South C...	43.662744	-79.321558
101	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...	43.636258	-79.498509
102	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999

Table 2: Canadian Neighborhoods with respective postal codes and geospatial coordinates

Utilizing google geocoder for Toronto’s longitude and latitude, the neighborhoods are visualized with Folium mapping library.

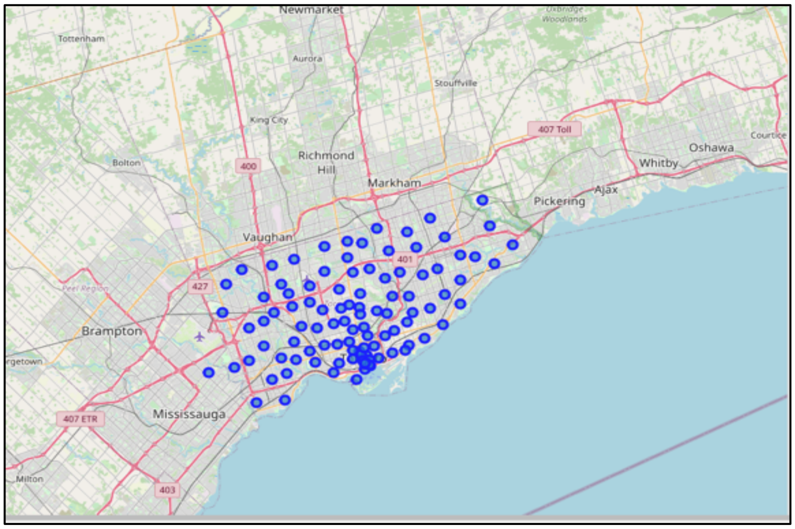


Figure 3: Canadian Neighborhoods map

Foursquare returns the venues' frequency by neighborhoods for a given zip code and their respective latitude and longitude. This information can be used as a rough guide as Foursquare returns the findings based on a specified radius from that given latitude and longitude. The corresponding data is stored into an "URL" for subsequent feature selection and analysis. (Link: 'Top 200 venues within 5000 meters radius')

4. Feature Selection:

The feature selection and subsequent analysis is carried out for "Restaurant" category with Foursquare API free developer account. The total restaurants per neighborhood is shown below:

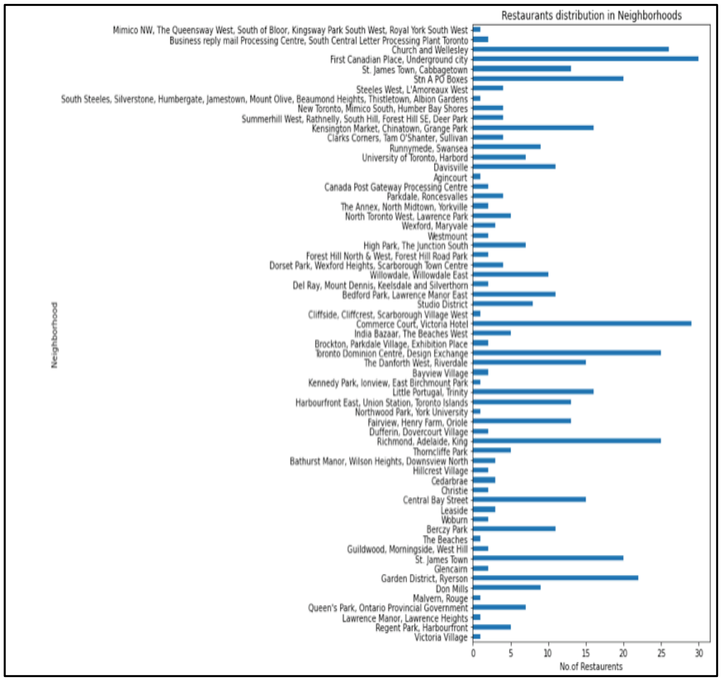
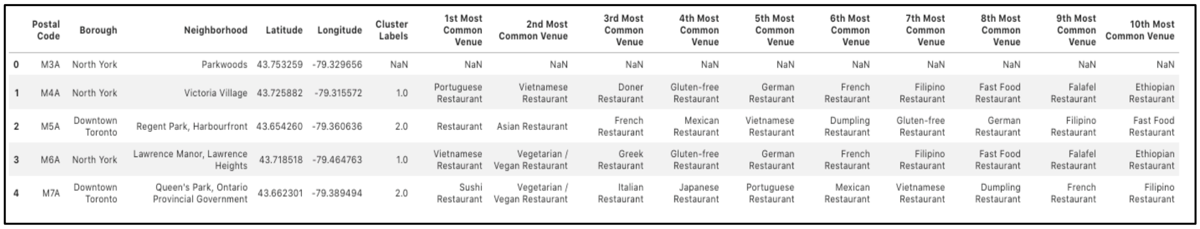


Figure 4: Restaurants frequency per neighborhood

Once the restaurants frequency per neighborhood data is further sorted with one hot encoding method for top 10 restaurant venues per neighborhood and grouped them based on their cuisine style. This data frame is used for the unsupervised K-Means cluster analysis for a suitable location to start a restaurant and its cuisine style.



	Postal Code	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	M3A	North York	Parkwoods	43.753259	-79.329656	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	M4A	North York	Victoria Village	43.725882	-79.315572	1.0	Portuguese Restaurant	Vietnamese Restaurant	Doner Restaurant	Gluten-free Restaurant	German Restaurant	French Restaurant	Filipino Restaurant	Fast Food Restaurant	Falafel Restaurant	Ethiopian Restaurant
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	2.0	Restaurant	Asian Restaurant	French Restaurant	Mexican Restaurant	Vietnamese Restaurant	Dumpling Restaurant	Gluten-free Restaurant	German Restaurant	Filipino Restaurant	Fast Food Restaurant
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	1.0	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Greek Restaurant	Gluten-free Restaurant	German Restaurant	French Restaurant	Filipino Restaurant	Fast Food Restaurant	Falafel Restaurant	Ethiopian Restaurant
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	2.0	Sushi Restaurant	Vegetarian / Vegan Restaurant	Italian Restaurant	Japanese Restaurant	Portuguese Restaurant	Mexican Restaurant	Vietnamese Restaurant	Dumpling Restaurant	French Restaurant	Filipino Restaurant

Table 3: Restaurants frequency per neighborhood

5. Data Modeling using K-Means Cluster Analysis:

Following K-Means cluster analysis, the restaurants distribution is analyzed per neighborhood. A cluster size number 5 has resulted the following distribution of the restaurants on the Toronto's map.

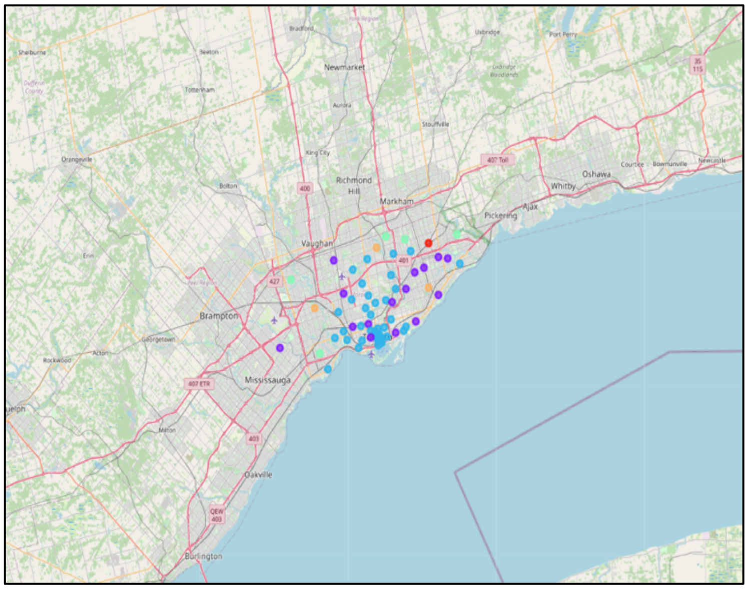


Figure 5: Clusters of restaurants per neighborhood on Toronto's map. Folium library is used for mapping these clusters. Total of 5 clusters are considered.

At this stage, the cuisine style is not yet segregated per neighborhood. For a potential investor, the choice depends on the local restaurants and their cuisine style frequency. Hence, the data is further analyzed for cuisine styles per cluster, as shown below:

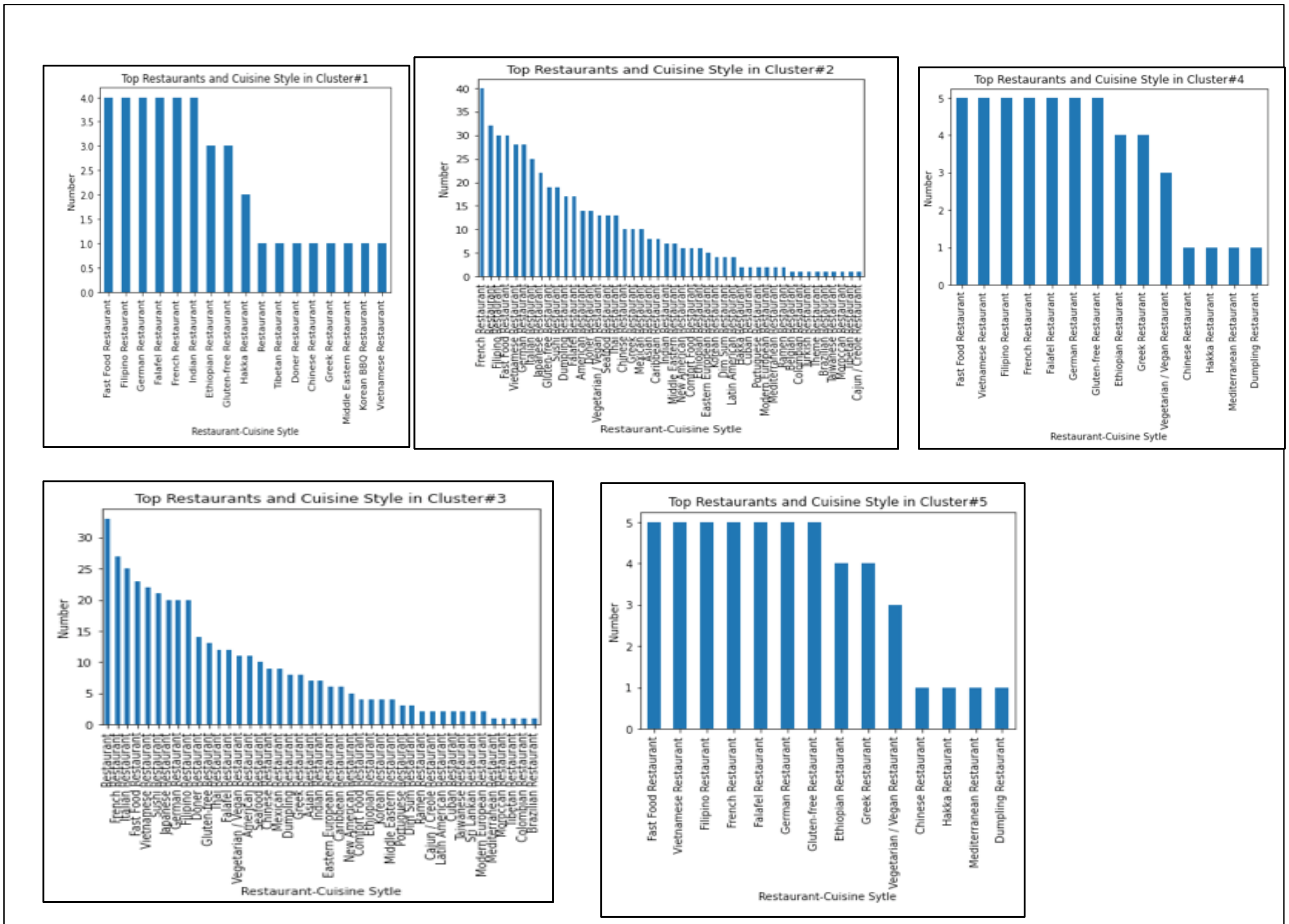


Figure 6: Distribution of restaurants per cuisine style in a given Cluster#.

The resultant K-Means cluster analysis clearly indicates, the cluter#1, #4, and #5 have small number of restaurants with a given cuisine style. These neighborhoods can be good for new restaurants due to low competition. However, for a successful restaurant business, consistent customers visitation is very important. Even though these three clusters indicate low number of existing restaurant businesses, one needs to consider the other businesses, local events frequency, as well as demographic information for consistent business. Then only, the entrepreneur can have better return on investment.

On the other hand, cluster#2 and #3 are high populated with restaurants with a variety of cuisine options. This trend clearly indicates either these two clusters are business centers or tourist spots with more frequent local events. In these neighborhoods, a new restaurant needs to face huge competition from the existing established restaurants. For this reason, investor needs to be extra cautious to attract new customers, if a new restaurant business is started. In addition, entrepreneur has a choice to choose a cuisine style that has low frequency in these neighborhoods. This can give a better advantage of the investor to focus and expand his business in these neighborhoods. Definitely quality, customer service, ambience, and price play a role to attract customers and hence a successful business.

6. Conclusions and future directions:

The primary objective of this analysis is to recommend a suitable neighborhood in Toronto City for a new restaurant business. Using unsupervised K Means cluster algorithm, it is shown that restaurant business and choice of cuisine style is determined by local restaurant competition.

This analysis can give a probable cluster to start a new restaurant with specific cuisine style. Some areas can have number of restaurants with a wide choice of cuisine style. In such areas, choosing a cuisine style can play a good role for success. Of course, this analysis also requires local demographic distribution for regular customers.

Cluster#2 and #3 have large number of restaurants. Here, quality of service and food taste at an affordable price can play a big role. Cluster#1, #4, #5 have small number of restaurants. Even though the local competition is small in these areas, it is important to consider customer visits based on local situation.

At the end it is imperative quality of service and food taste matters for successful restaurant business in addition to customer turn around frequency, local businesses, events, attractions. The analysis can be extended considering these sets of data for better analysis. Complete analysis and report can be found from the following links.

1. The python code for this analysis is given at: Capstone Project: Python Code on Github
2. The presentation file is given at: Capstone Project: Presentation on Github

Further extension of this analysis can be done considering local demographic information, venue assessment with quality rating, as well as local events etc.

References:

1. ***https://en.wikipedia.org/wiki/Cuisine_in_Toronto***
2. ***https://tastytourstoronto.com***
3. **https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&oldid=1008658627**
4. **https://cocl.us/Geospatial_data**