BATTLE OF THE NEIGHBORHOODS

Exploring Suitable Locations for New Sushi Restaurants in Toronto, Canada using Data Science

Introduction:

This Capstone project work aims to utilize all Data Science Concepts learnt in the IBM Data Science Professional Course offered by Coursera.

To begin with, a Business Problem, data source & data to be used in the project are defined. Machine learning tools will be used to analyze the data and predict solutions.

Recommendations based on the analysis is provided which can be utilized by the business stakeholders to make their decisions.

Table of Contents

Business Problem:

Toronto is a Canadian city with a taste for international cuisine. With a population of approximately 6 million which includes diverse ethnic groups, Toronto is a multicultural city offering many opportunities for entrepreneurs and Business owners in the restaurant business domain.

In this project, Toronto Neighborhood data, Foursquare location data and regional clustering of venue information are used to determine 'best' neighbourhoods in Toronto to open a Sushi restaurant. Sushi is one of the most bought dishes in Toronto originating from Japan. Toronto is home to many Sushi patrons comprising of varied ethnicities which include Chinese, Koreans, Filipinos, Japanese and people from South East Asia. Their combined population is approximately 20% of Toronto.

Toronto receives approximately 22 million international visitors annually. A sizeable chunk of these visitors, explore the exotic food delights on offer in Toronto. As Sushi restaurants are exotic, they will also cater to international visitors.

The presence of many Asians in Toronto will also provide Chefs and labour for the Sushi Restaurants.

Thus, opening of new Sushi restaurants in Toronto is a good business proposition.

Target Audience:

The target audience are Entrepreneurs and Business owners who want to open new Sushi Restaurants or expand their current business. The analysis will provide key information, which can be used by the target audience.

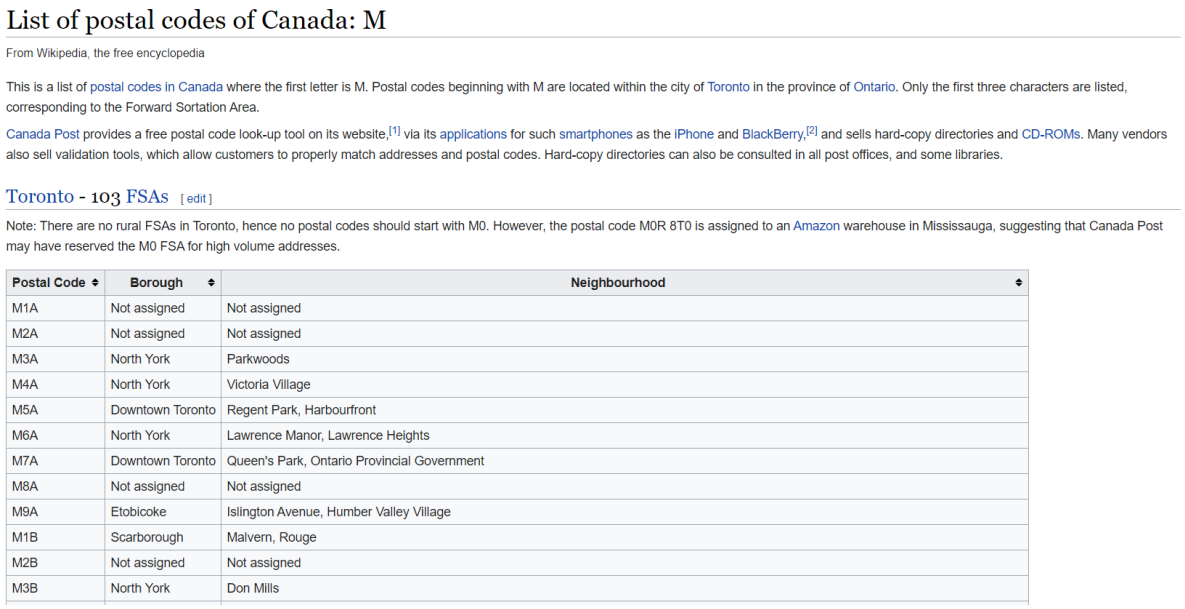
Overview of Data:

The data required for the analysis will be obtained from multiple sources. The list of neighbourhoods in Toronto will be sourced from Wikipedia, the Geographical location of the neighbourhoods from a csv file and Venue data of Sushi restaurants from Foursquare. The

Venue data will help find which neighbourhood is best suitable to open a Sushi restaurant in Toronto.

1. Toronto Neighborhood Data

The list of Toronto neighborhoods is sourced from Wikipedia (Fig.1). ("https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M").



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 Forward Sortation Area.

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 tools, which allow customers to properly 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 Canada Post may have reserved the M0 FSA for high volume addresses.

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

Fig.1 A screen grab of the Wikipedia page containing Toronto Neighborhood data

The list contains Postal Codes, Name of Boroughs and Neighbourhoods. The data is available in a format which is not suitable for the analysis. Therefore, the data is scraped from the Wikipedia page. Data scraping is done from the website as it is suitable for the analysis. The scraped data is then wrangled, cleaned and read into Pandas data frame so that it is in a structured format (Fig.2).

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

Fig.2 Scraped Data in Pandas Data Frame

2. Geographical Location data

The Geographical coordinates of the Toronto neighbourhoods with the respective Postal Codes was sourced from the website https://cocl.us/Geospatial_data. The data is in csv format (Fig.3a). The data was converted to Pandas data frame (Fig 3b).

	A	B	C
1	Postal Code	Latitude	Longitude
2	M1B	43.80669	-79.1944
3	M1C	43.78454	-79.1605
4	M1E	43.76357	-79.1887
5	M1G	43.77099	-79.2169
6	M1H	43.77314	-79.2395

Fig.3a Geospatial data in csv format

	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.3b Geospatial Data in Pandas Data Frame

3. Venue Data using Foursquare

The Neighborhood data frame and geospatial data frame were merged to get a new data frame (Fig.4).

	Postalcode	Borough	Neighborhood	Latitude	Longitude
37	M4E	East Toronto	The Beaches	43.676357	-79.293031
41	M4K	East Toronto	The Danforth West, Riverdale	43.679557	-79.352188
42	M4L	East Toronto	India Bazaar, The Beaches West	43.668999	-79.315572
43	M4M	East Toronto	Studio District	43.659526	-79.340923
44	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790

Fig.4 Merged Data Frame

Then using Foursquare credentials (client ID, client secret and version) and the data in the merged data frame, the venue data is extracted (Fig.5)

	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 Danforth West, Riverdale	43.679557	-79.352188	MenEssentials	43.677820	-79.351265	Cosmetics Shop

Fig.5 Venue Data extracted from Foursquare API

This Venue data is used for further analysis.

Methodology:

After venue data extraction, for machine learning algorithms, the categorical data was transformed to numerical data by a technique called **One Hot Encoding**. Individual venues were turned into frequency, at how many of those Venues were located in each neighborhood (Fig.6).

	Neighborhoods	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	...
0	The Beaches	0	0	0	0	0	0	0	0	0	...
1	The Beaches	0	0	0	0	0	0	0	0	0	...
2	The Beaches	0	0	0	0	0	0	0	0	0	...
3	The Beaches	0	0	0	0	0	0	0	0	0	...
4	The Beaches	0	0	0	0	0	0	0	0	0	...

Fig.6 One Hot Encoding

The rows were grouped by Neighborhood and Average of the frequency of occurrence of each Venue Category was taken (Fig.7).

	Neighborhoods	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	...	Toy / Game Store	Trail
0	Berczy Park	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.0	...	0.0	0.0
1	Brockton, Parkdale Village, Exhibition Place	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.0	...	0.0	0.0
2	Business reply mail Processing Centre, South C...	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.0	...	0.0	0.0
3	CN Tower, King and Spadina, Railway Lands, Har...	0.0	0.066667	0.066667	0.066667	0.133333	0.133333	0.066667	0.0	0.0	...	0.0	0.0
4	Central Bay Street	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.0	...	0.0	0.0

Fig.7 Grouped Neighborhoods by the averaging of the frequency of each Venue

A new data frame was then created which only stored the Neighborhood names as well as the average frequency of Sushi Restaurants in that Neighborhood (Fig.8). This will allow the data to be summarized based on each individual Neighborhood and is simpler to analyze.

	Neighborhood	Sushi Restaurant
0	Berczy Park	0.017544
1	Brockton, Parkdale Village, Exhibition Place	0.000000
2	Business reply mail Processing Centre, South C...	0.000000
3	CN Tower, King and Spadina, Railway Lands, Har...	0.000000
4	Central Bay Street	0.014706

Fig. 8 A data frame showing the Neighborhood names as well as the average frequency of Sushi Restaurants in that Neighborhood

K-Means clustering was used to cluster the neighborhoods based on the neighborhoods that had similar averages of Sushi Restaurants in that Neighborhood. To get our optimum K value that was neither overfitting or underfitting the model, the Elbow Point Technique was used. In this technique, a test was conducted 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 (Fig.9). In this case, the Elbow Point was at K = 3 (Fig.10). That means, the analysis will involve a total of 3 clusters.

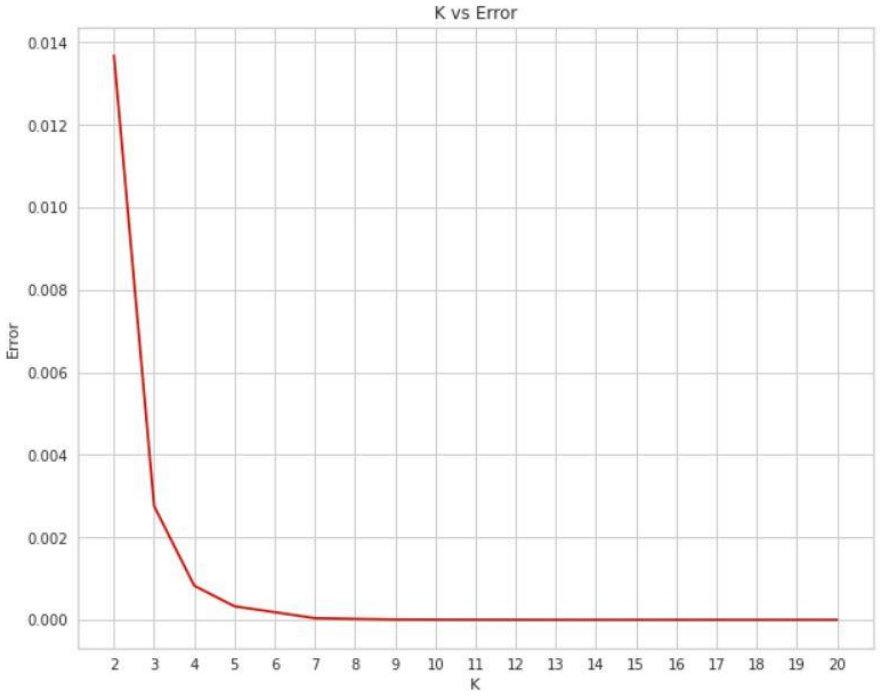


Fig.9 K versus Error

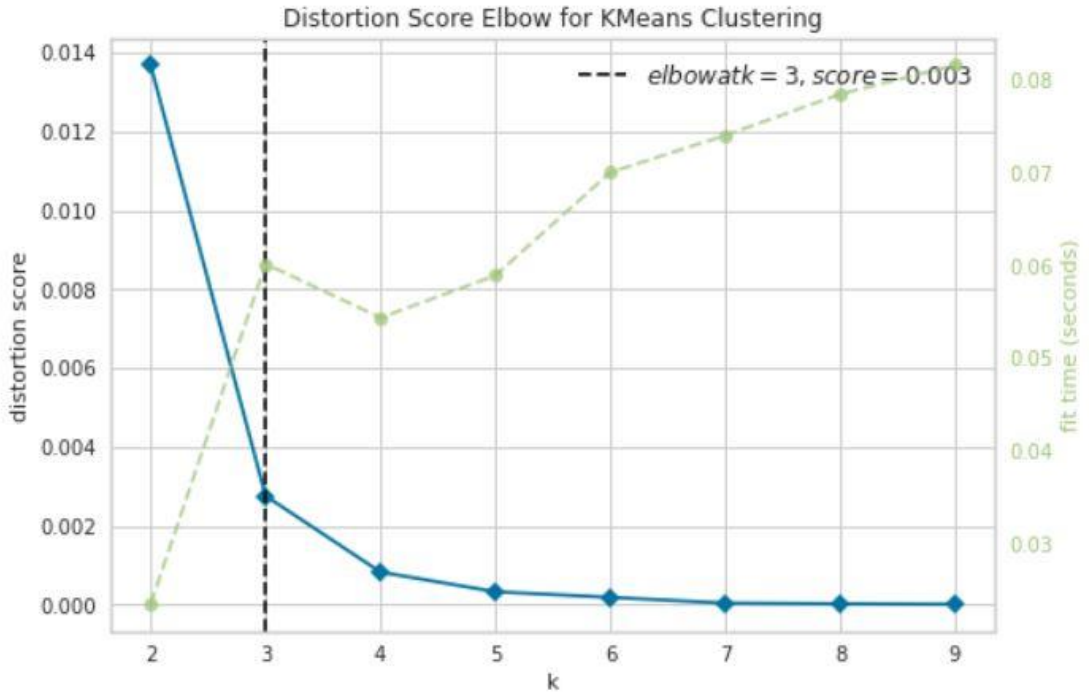


Fig.10 Optimum K at Elbow point

A model was integrated which would fit the error and calculate the distortion score. From the dotted line, we see that the Elbow is at K=3. Moreover, in K-Means clustering, similar objects based on a certain variable are put into the same cluster. Neighborhoods that had similar mean frequency of Sushi Restaurants were divided into 3 clusters. Each of these clusters were labelled from 0 to 2 as the indexing of labels begin with 0 instead of 1.

	Neighborhood	Sushi Restaurant	Cluster Labels
0	Berczy Park	0.017544	0
1	Brockton, Parkdale Village, Exhibition Place	0.000000	1
2	Business reply mail Processing Centre, South C...	0.000000	1
3	CN Tower, King and Spadina, Railway Lands, Har...	0.000000	1
4	Central Bay Street	0.014706	0

Fig.11 Cluster Labels

The venue data was then merged with the table above creating a new table which would be the basis for analyzing opportunities for opening new Sushi Restaurants in Toronto. A map using the Folium package in Python was created and each Toronto neighborhood was marked

with colors based on the cluster label. Cluster 1 was Red, cluster 2 was Purple and cluster 3 was Aquamarine. The map below (Fig.12) shows the different clusters that had similar mean frequency of Sushi restaurants.

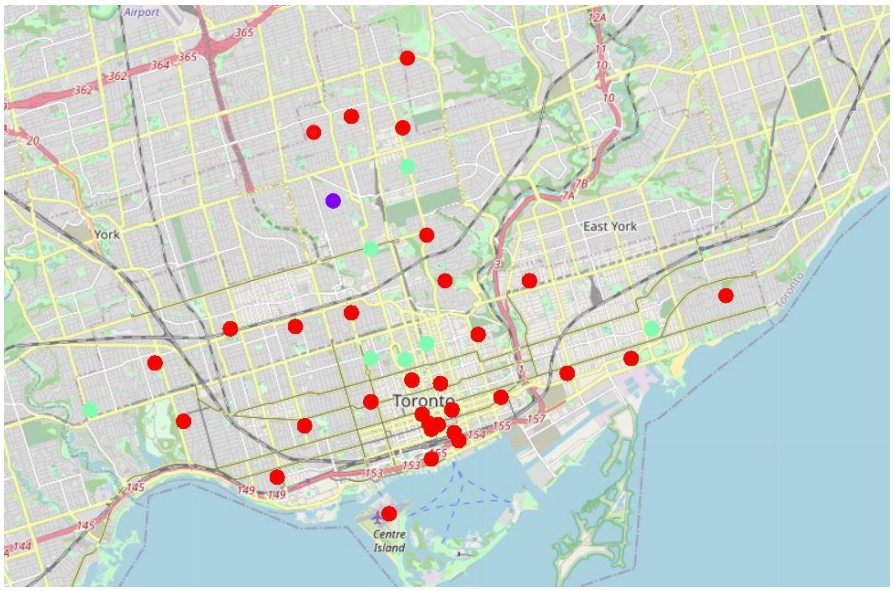


Fig.12 Toronto Map showing the Clusters

Results:

From the bar graph plotted using Matplotlib (Fig. 13), the number of Toronto Neighborhoods per cluster can be visualized. Cluster 2 has the least neighborhoods (1) while cluster 1 has the most (31). Cluster 3 has 7 neighborhoods.

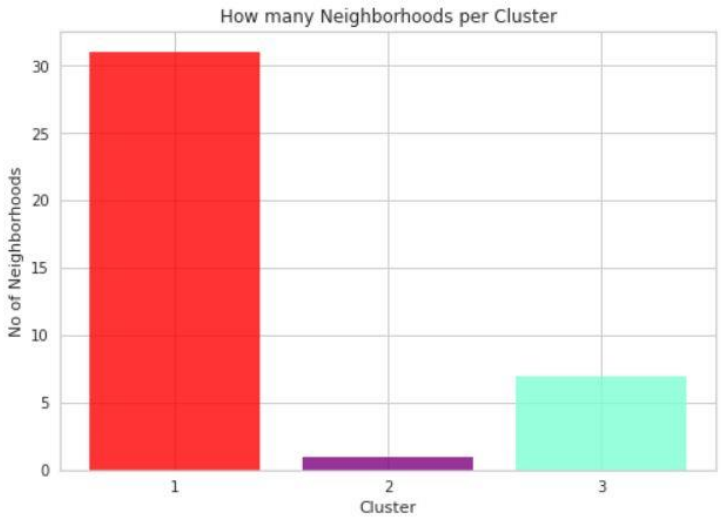


Fig.13 Number of Toronto Neighborhoods per cluster

The Average Sushi Restaurants in each Toronto Neighborhood (Fig.14) is then compared.

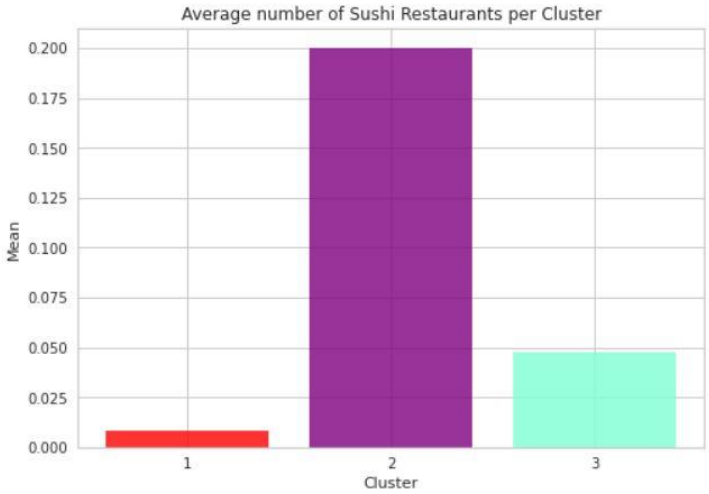


Fig.14 Average Sushi Restaurants in each Toronto Neighborhood

Though there is only 1 neighborhood in Cluster 2, it has the highest average of Sushi Restaurants (0.2) while Cluster 1 has the most neighborhoods (31) but has the least average of Sushi Restaurants (0.0086).

The Cluster wise analysis is done as follows:

Cluster 1 (Red Color)

	Borough	Neighborhood	Sushi Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
651	East Toronto	The Danforth West, Riverdale	0.02381	0	43.679557	-79.352188	The Auld Spot Pub	43.677335	-79.353130	Pub
676	East Toronto	The Danforth West, Riverdale	0.02381	0	43.679557	-79.352188	Marvel Coffee Co.	43.678630	-79.347460	Coffee Shop
668	East Toronto	The Danforth West, Riverdale	0.02381	0	43.679557	-79.352188	Bar Oak	43.677931	-79.348724	Lounge
662	East Toronto	The Danforth West, Riverdale	0.02381	0	43.679557	-79.352188	Simone's Caribbean Restaurant	43.678655	-79.346582	Caribbean Restaurant
670	East Toronto	The Danforth West, Riverdale	0.02381	0	43.679557	-79.352188	Leonidas Chocolates Cafe	43.678118	-79.349485	Café
...	...	...	...	...	...	...	...	...	...	...
902	Downtown Toronto	Commerce Court, Victoria Hotel	0.00000	0	43.648198	-79.379817	Cactus Club Cafe	43.649552	-79.381671	American Restaurant
903	Downtown Toronto	Commerce Court, Victoria Hotel	0.00000	0	43.648198	-79.379817	Ki Modern Japanese + Bar	43.647223	-79.379374	Japanese Restaurant
904	East Toronto	Studio District	0.00000	0	43.659526	-79.340923	McQueens Pub	43.661483	-79.338072	Gastropub
905	East Toronto	Studio District	0.00000	0	43.659526	-79.340923	Saulter Street Brewery	43.658412	-79.346392	Brewery

Fig.15 Cluster1

Cluster 1 comprises of 31 neighborhoods and only 1 Sushi restaurant. Therefore, the average amount of Sushi Restaurants that were near the venues in Cluster 1 is the lowest (~0.0086). In the map, the Cluster 1 nodes were dispersed all over Toronto, making it one of the most sparsely populated cluster.

Cluster 2 (Purple Color)

	Borough	Neighborhood	Sushi Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Central Toronto	Forest Hill North & West, Forest Hill Road Park	0.2	1	43.696948	-79.411307	Oliver jewelry	43.700374	-79.407644	Jewelry Store
1	Central Toronto	Forest Hill North & West, Forest Hill Road Park	0.2	1	43.696948	-79.411307	Kay Gardner Beltline Trail	43.698446	-79.406873	Trail
2	Central Toronto	Forest Hill North & West, Forest Hill Road Park	0.2	1	43.696948	-79.411307	TTC Bus #14 Glencairn	43.700221	-79.410274	Bus Line
3	Central Toronto	Forest Hill North & West, Forest Hill Road Park	0.2	1	43.696948	-79.411307	Forest Hill Road Park	43.697945	-79.406605	Park
4	Central Toronto	Forest Hill North & West, Forest Hill Road Park	0.2	1	43.696948	-79.411307	Nikko Sushi Japenese Restaurant	43.700443	-79.407957	Sushi Restaurant

Fig.16 Cluster2

Cluster 2 comprises of only 1 neighborhood and only 1 Sushi restaurant. Therefore, the average amount of Sushi Restaurants that were near the venues in Cluster 2 is the highest (~0.2). The reason for the highest average is because the Sushi restaurants in this cluster are located only in two neighborhoods - Forest Hill North & West and Forest Hill Road Park in Central Toronto. In the map, the Cluster 2 node depicts a densely populated cluster.

Cluster 3 (Aquamarine Color)

	Borough	Neighborhood	Sushi Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
164	Central Toronto	Summerhill West, Rathnelly, South Hill, Forest...	0.066667	2	43.686412	-79.400049	TTC Stop #	43.685826	-79.404981	Light Rail Station
169	Central Toronto	Summerhill West, Rathnelly, South Hill, Forest...	0.066667	2	43.686412	-79.400049	Mary Be Kitchen	43.687708	-79.395062	Restaurant
159	Central Toronto	Summerhill West, Rathnelly, South Hill, Forest...	0.066667	2	43.686412	-79.400049	Popeyes Louisiana Kitchen	43.689300	-79.395302	Fried Chicken Joint
160	Central Toronto	Summerhill West, Rathnelly, South Hill, Forest...	0.066667	2	43.686412	-79.400049	Pizzaiolo	43.687991	-79.394634	Pizza Place
161	Central Toronto	Summerhill West, Rathnelly, South Hill, Forest...	0.066667	2	43.686412	-79.400049	RBC Royal Bank	43.688058	-79.394478	Bank
...	...	...	...	...	...	...	...	...	...	...
94	Downtown Toronto	University of Toronto, Harbord	0.030303	2	43.662696	-79.400049	Her Father's Cider Bar + Kitchen	43.662448	-79.404703	Beer Bar
93	Downtown Toronto	University of Toronto, Harbord	0.030303	2	43.662696	-79.400049	Magic Noodle	43.662728	-79.403602	Noodle House
91	Downtown Toronto	University of Toronto, Harbord	0.030303	2	43.662696	-79.400049	East of Brunswick	43.665609	-79.403324	Pub
88	Downtown Toronto	University of Toronto, Harbord	0.030303	2	43.662696	-79.400049	Subway	43.664489	-79.399118	Sandwich Place
92	Downtown Toronto	University of Toronto, Harbord	0.030303	2	43.662696	-79.400049	DT Bistro	43.662375	-79.405734	Café

Fig.17 Cluster3

Cluster 3 comprises of 7 neighborhoods and only 12 Sushi restaurants. Therefore, the average amount of Sushi Restaurants that were near the venues in Cluster 3 is the second lowest (~0.048). In the map, the Cluster 3 nodes are located in Summerhill West, Rathnelly, South Hill, etc., in Central Toronto and University of Toronto & Harbord in Downtown Toronto.

Thus, the ordering of the average Sushi Restaurants in each cluster goes as follows:

1. Cluster 2 (≈0.2)

2. Cluster 3 (≈0.048)

3. Cluster 1 (≈0.0086)

Discussion:

Most of the Sushi restaurants are in cluster 2 represented by the Purple node. The neighborhoods located in the Central Toronto area, that have the highest average of Sushi Restaurants are Forest Hill North & West and Forest Hill Road Park.

Though there are a large number of neighborhoods (31) in cluster 1, there is little to no Sushi restaurant. Therefore, opening Sushi restaurants in neighborhoods of Danforth West, Riverdale, Studio District, etc in East Toronto and Commerce Court & Victoria Hotel in Downtown Toronto **is recommended.**

The Central and Downtown Toronto area (cluster 3) has the second last average of Sushi restaurants. Also, there is lesser competition. Therefore, opening Sushi restaurants in Summerhill West, Rathnelly, South Hill, etc in Central Toronto and University of Toronto & Harbord in Downtown Toronto **is recommended**.

Some of the drawbacks of this analysis are:

- Clustering is completely based on data obtained from Foursquare API.
- Also, the analysis does not take into consideration, the Sushi patron population (primarily Asian) which is scattered across the neighbourhoods.

Conclusion:

To conclude, this project handled the process of identifying the business problem; specifying, extracting and preparing the data; performing the machine learning by utilizing k-means clustering and providing recommendations to the target audience.