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.
- The Capstone Project Explores Suitable Locations for New Sushi Restaurants in Toronto, Canada using Data Science
- A Business Problem, data source & data used in the project are defined. Machine learning tools that were used to analyze the data and predict solutions are discussed.
- Recommendations are made to the target audience.

Business Problem

- 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 is a multicultural city offering many opportunities for entrepreneurs and Business owners in the restaurant business domain.
- 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

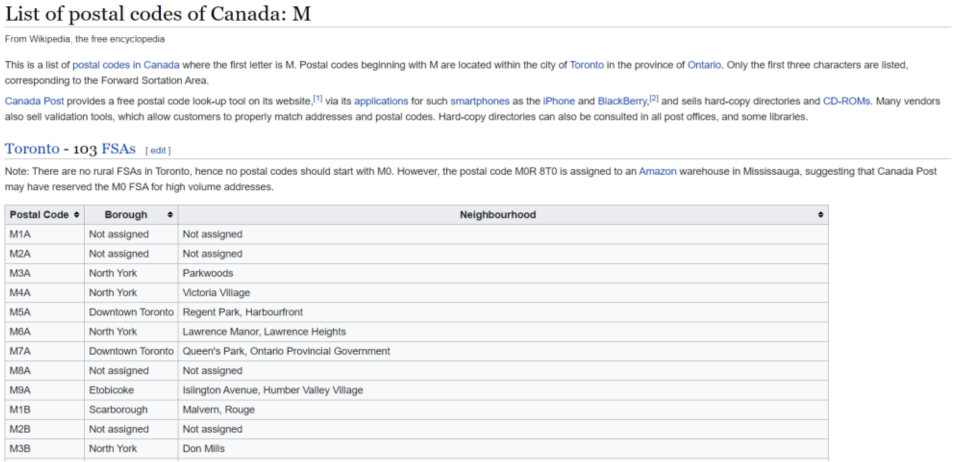
- The target audience are Entrepreneurs and Business owners who want to open new Sushi Restaurants or expand their current business. The analysis will provide key information, which can be used by the target audience.

Overview of Data

- The data required for the analysis was obtained from multiple sources.
- The list of neighbourhoods in Toronto was sourced from Wikipedia.
- The Geographical location of the neighbourhoods was obtained from a csv file.
- 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.

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”).
- 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.



List of postal codes of Canada: M

From Wikipedia, the free encyclopedia

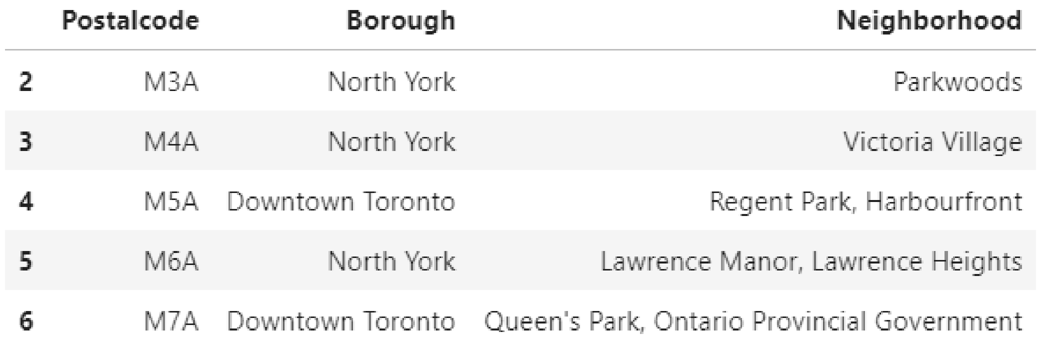
This is a list of postal codes in Canada where the first letter is M. Postal codes beginning with M are located within the city of Toronto in the province of Ontario. Only the first three characters are listed, corresponding to the Forward Sortation Area.

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.

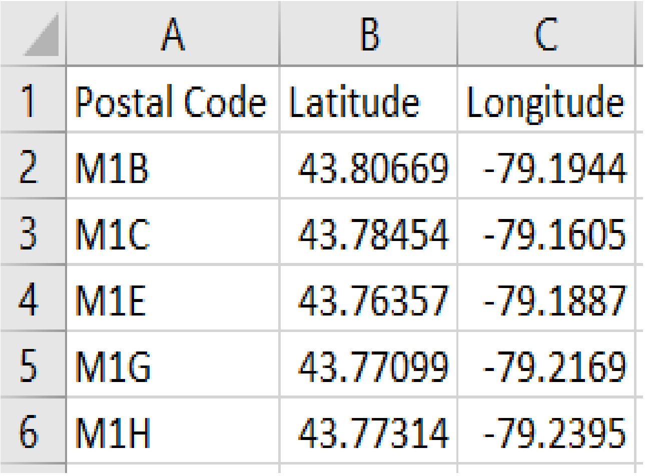
Postal Code	Borough	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Regent Park, Harbourfront
M6A	North York	Lawrence Manor, Lawrence Heights
M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
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



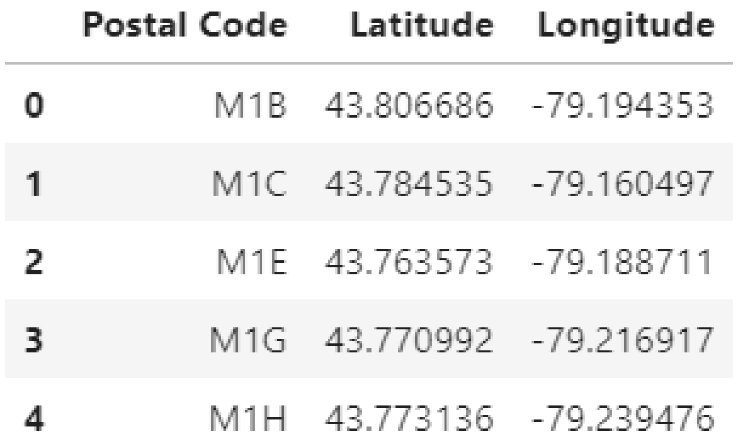
	Postalcode	Borough	Neighborhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
5	M6A	North York	Lawrence Manor, Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

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. The data was converted to Pandas data frame.



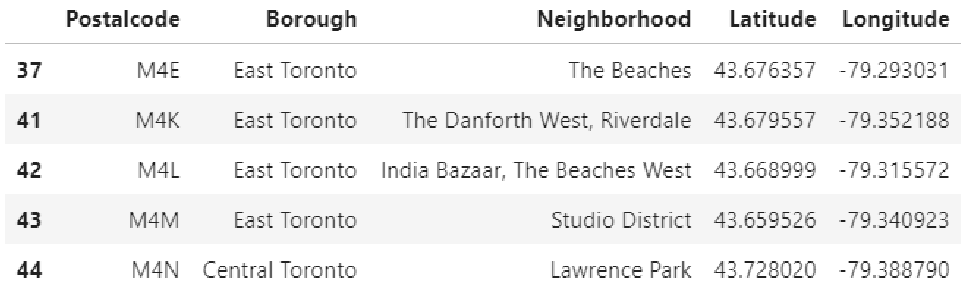
	A	B	C
1	Postal Code	Latitude	Longitude
2	M1B	43.80669	-79.1944
3	M1C	43.78454	-79.1605
4	M1E	43.76357	-79.1887
5	M1G	43.77099	-79.2169
6	M1H	43.77314	-79.2395



	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

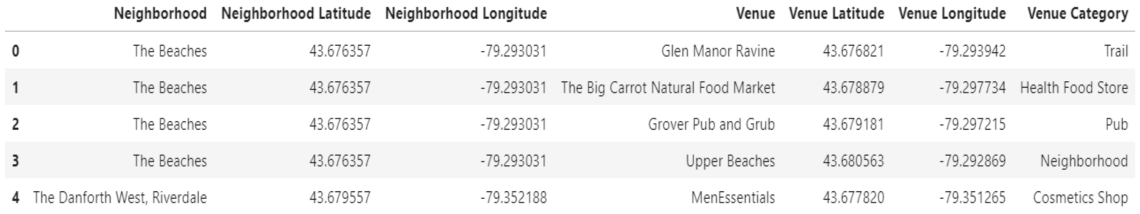
Venue Data using Foursquare

- The Neighborhood data frame and geospatial data frame were merged to get a new data frame.



	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

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

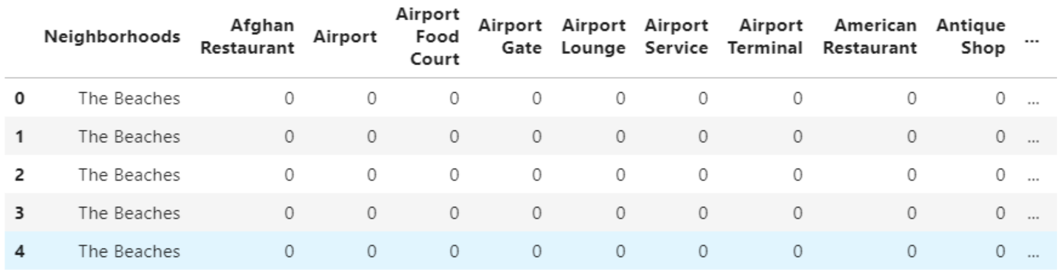


	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	The Beaches	43.676357	-79.293031	Glen Manor Ravine	43.676821	-79.293942	Trail
1	The Beaches	43.676357	-79.293031	The Big Carrot Natural Food Market	43.678879	-79.297734	Health Food Store
2	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

- This Venue data is used for further analysis.

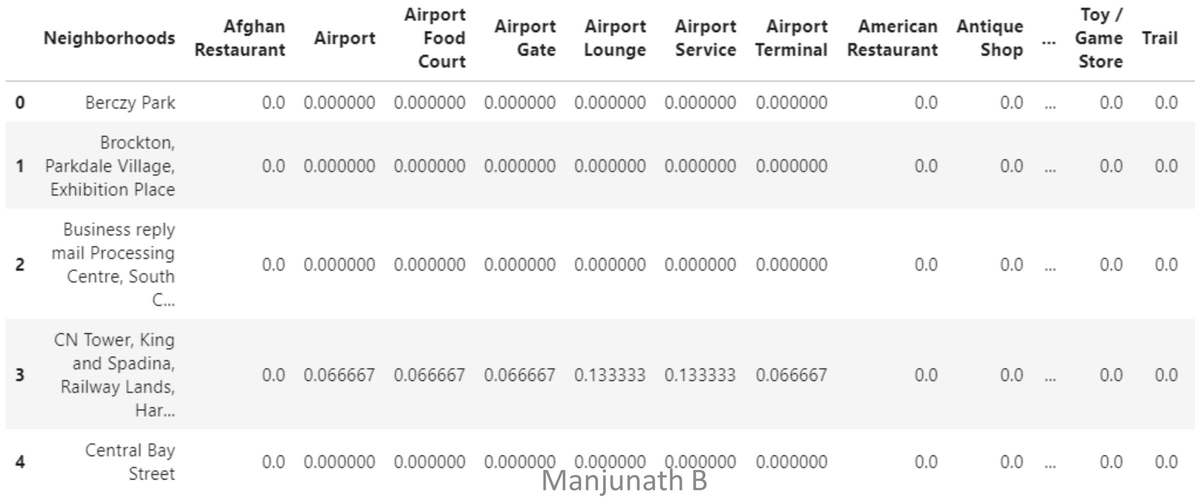
Methodology – One Hot Encoding

- 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.



	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	...

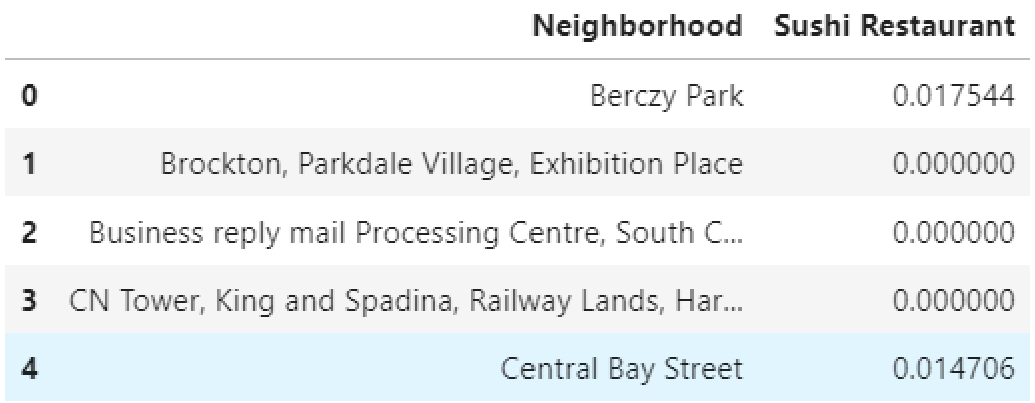
- The rows were grouped by Neighborhood and Average of the frequency of occurrence of each Venue Category was taken.



	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

Methodology – One Hot Encoding

- 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. This will allow the data to be summarized based on each individual Neighborhood and is simpler to analyze.

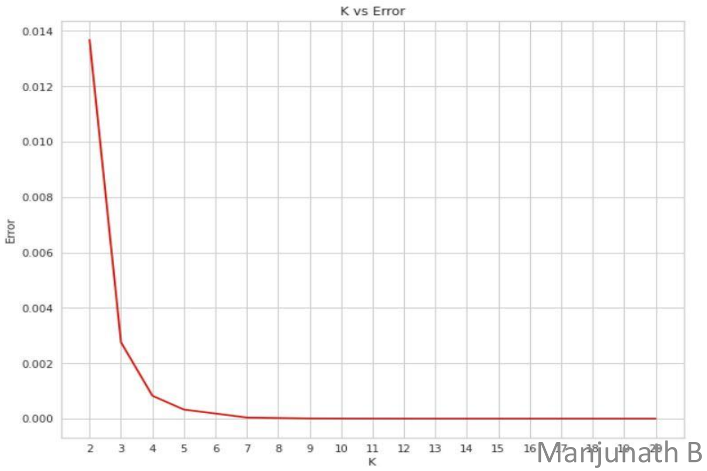


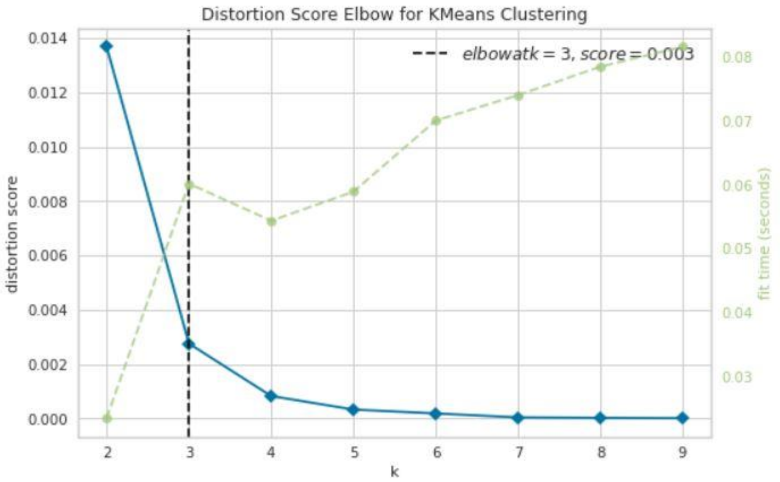
	Neighborhood	Sushi Restaurant
0	Berczy Park	0.017544
1	Brockton, Parkdale Village, Exhibition Place	0.000000
2	Business reply mail Processing Centre, South C...	0.000000
3	CN Tower, King and Spadina, Railway Lands, Har...	0.000000
4	Central Bay Street	0.014706

Methodology – Kmeans Clustering

- 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.

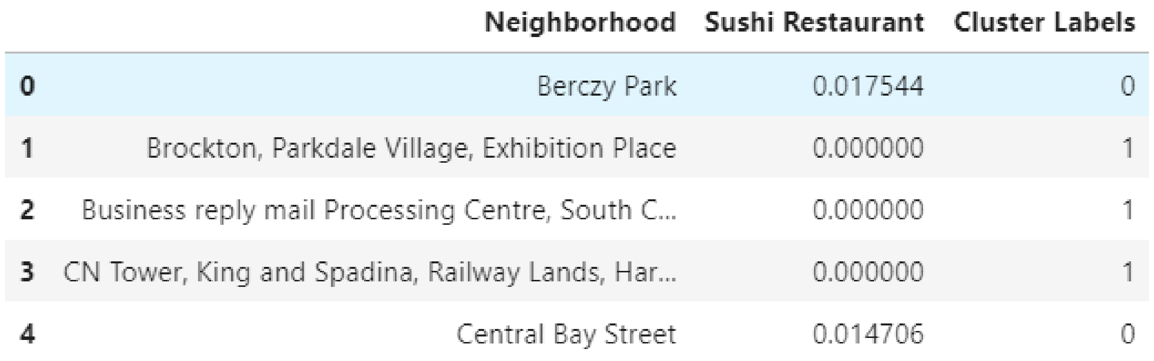
- In this technique, a test was conducted with different number of K values and measured the accuracy and then chose the best K value. The best K value is chosen at the point in which the line has a sharpest turn. In this case, the Elbow Point was at K = 3. That means, the analysis will involve a total of 3 clusters. 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.





Methodology – Kmeans Clustering

- 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.

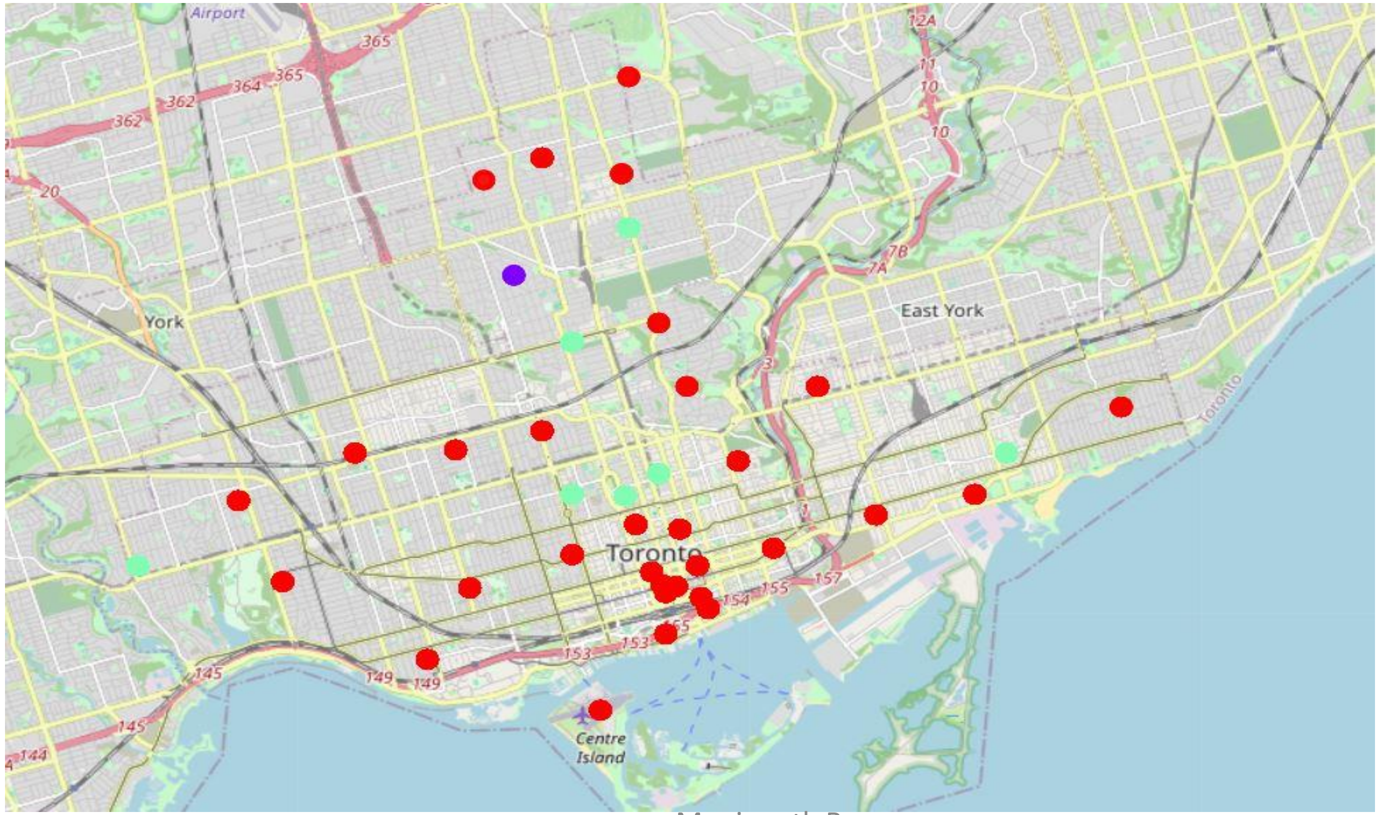


	Neighborhood	Sushi Restaurant	Cluster Labels
0	Berczy Park	0.017544	0
1	Brockton, Parkdale Village, Exhibition Place	0.000000	1
2	Business reply mail Processing Centre, South C...	0.000000	1
3	CN Tower, King and Spadina, Railway Lands, Har...	0.000000	1
4	Central Bay Street	0.014706	0

- 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.

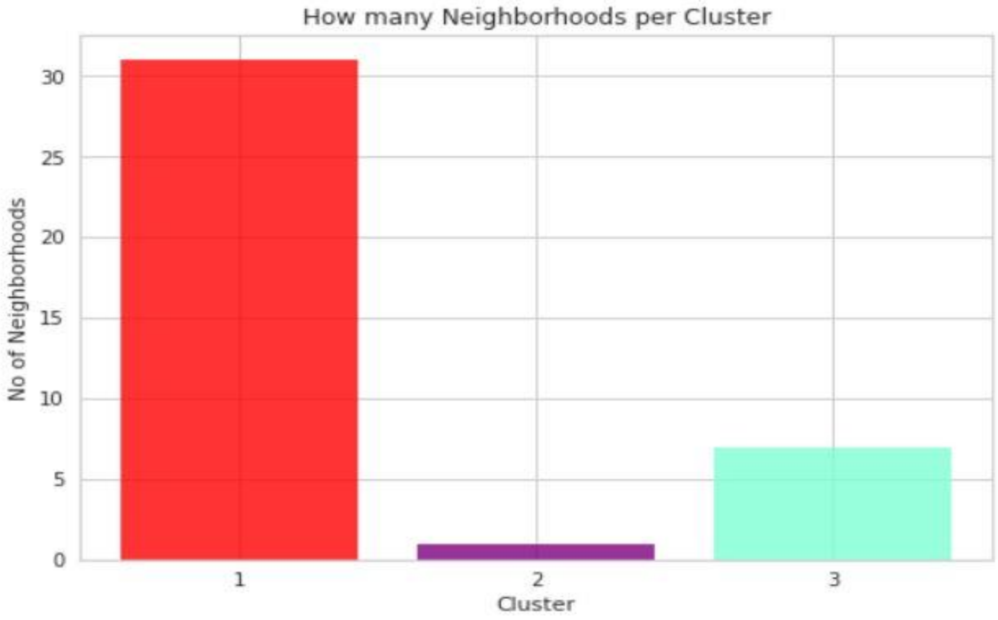
Methodology – Kmeans Clustering

- 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 shows the different clusters that had similar mean frequency of Sushi restaurants.



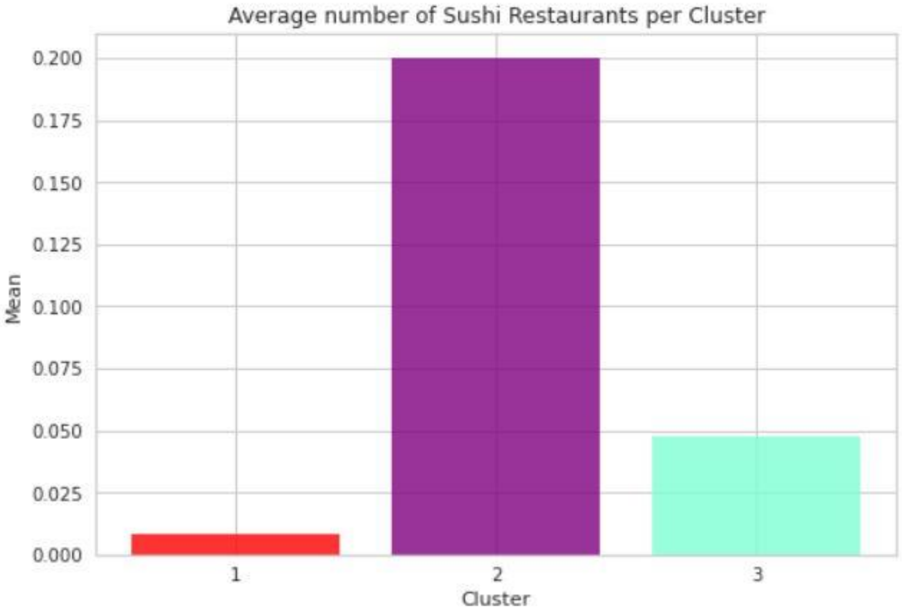
Results

- From the bar graph plotted using Matplotlib 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.



Results

- The Average Sushi Restaurants in each Toronto Neighborhood is then compared.



- 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).

Results

- 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**.

Discussion

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.