A VIETNAMESE RESTAURANT IN TORONTO

BACH KHOI VO

June 18, 2021

1. INTRODUCTION

a. Background

Toronto is one of the most densely populated areas in Canada. Being the land of opportunity, it brings in a variety of people from different ethnic backgrounds to the core city of Canada, Toronto. Being the largest city in Canada with an estimated population of over 6 million, there is no doubt about the diversity of the population. Moreover, Toronto is the most destination in Canada for international students from multi cultures such as Vietnam, China, Japanese, Korea and many more. Not only cross-culture in population but also in cuisine field. By the way, Vietnamese cuisine is one of the favorite cuisines in Toronto. In addition, It has become better known in the world over the last couple of decades, with international chefs and prestigious food magazines praising several national dishes.

b. Problem

Through this project, we will find the most suitable location for an entrepreneur to open a new Vietnamese restaurant in Toronto, Canada.

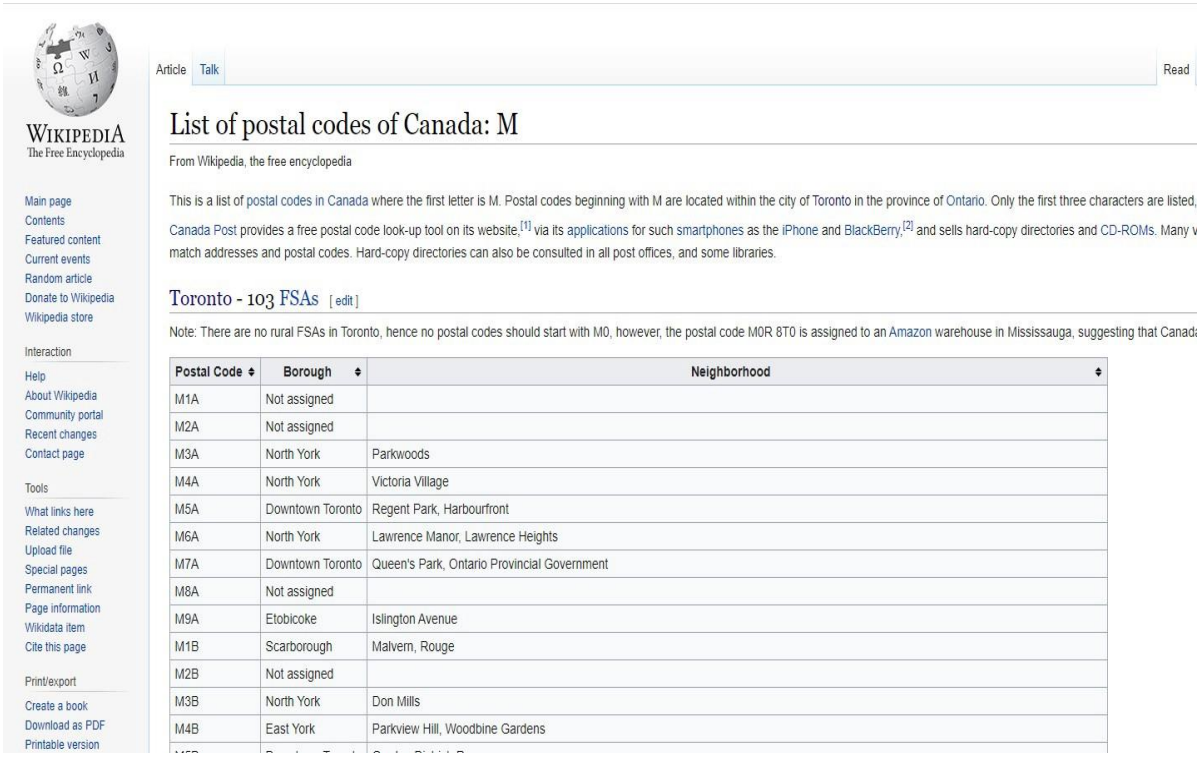
c. Interest

This project is aimed towards Entrepreneurs or Business owners who want to open a new Vietnamese Restaurant or grow their current business. The analysis will provide vital information that can be used by the target audience.

2. DATA OVERVIEW

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 Asia restaurants (via *Foursquare*). The Venue data will help find which neighborhood is best suitable to open an Vietnamese restaurant.

a. 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

This source contains information about all boroughs and neighborhoods with corresponding postal code via the table. I need to implement some web scraping techniques to scrap the data then process and store it into *DataFrame*.

	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

b. Geocoder package

I don't use the package, but a *CSV* file that provided in the previous lab will be used. Then, merging it with Neighborhoods *DataFrame* to get the fully table.

	PostalCode	Borough	Neighborhood	Postal Code	Latitude	Longitude
0	M3A	North York	Parkwoods	M3A	43.753259	-79.329656
1	M4A	North York	Victoria Village	M4A	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	M5A	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	M6A	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	M7A	43.662301	-79.389494

c. Foursquare API

This source is used for getting nearby venues at each neighborhood. After that, cleansing it a little bit and merging to get the final *DataFrame.*

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	KFC	43.754387	-79.333021	Fast Food Restaurant
1	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
2	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
3	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
4	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant

3. METHODOLOGY

In this section, I only focus on 4 phases: *Data Preprocessing*, *Data Exploration, Machine Learning and Data Analyst*.

a. Data Preprocessing / Data Cleansing

- Toronto Neighborhoods

There is only one data need to be preprocessed which was scraped from Wikipedia. Because it was scraped from a website, it actually is in HTML format. So, I use *BeautifulSoup* package to normalize the data and transform it into *DataFrame*.

	PostalCode	Borough	Neighborhood
0	None	None	None
1	M1A\n	Not assigned\n	Not assigned\n
2	M2A\n	Not assigned\n	Not assigned\n
3	M3A\n	North York\n	Parkwoods\n
4	M4A\n	North York\n	Victoria Village\n

After having data in tabular format, I need to drop all rows that contain `Not assigned\n` value.

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

Last step, stripping the last '\n' character in each value to get the cleaned table.

	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

Finally, merging with Geospatial DataFrame and Foursquare nearby Venues DataFrame to get the final table.

s

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	KFC	43.754387	-79.333021	Fast Food Restaurant
1	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
2	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
3	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
4	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant

Toronto Venues One-hot

For feeding to the Machine Learning, we need to group the table by `Neighborhood`, then encode all the categories of venues by One-hot Encoder and drop the other columns except `Neighborhood`.

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	...	Vegetarian / Vegan Restaurant	Video Game Store
0	Agincourt	0	0	0	0	0	0	0	0	0	...	0	0
1	Alderwood, Long Branch	0	0	0	0	0	0	0	0	0	...	0	0
2	Bathurst Manor, Wilson Heights, Downsview North	0	0	0	0	0	0	0	0	0	...	0	0
3	Bayview Village	0	0	0	0	0	0	0	0	0	...	0	0
4	Bedford Park, Lawrence Manor East	0	0	0	0	0	0	0	0	0	...	0	0

▪ Vietnamese Restaurants Neighborhoods

From the `origin` Toronto Neighborhoods DataFrame and the Toronto Venues One-hot DataFrame, we group the table 'origin' by `Neighborhoods` and filter out it into sub DataFrame which Neighborhood have at least one Vietnamese Restaurants.

	Neighborhood	PostalCode	Borough	Latitude	Longitude
0	Little Portugal, Trinity	M6J	West Toronto	43.647927	-79.419750
1	Kensington Market, Chinatown, Grange Park	M5T	Downtown Toronto	43.653206	-79.400049
2	Dorset Park, Wexford Heights, Scarborough Town...	M1P	Scarborough	43.757410	-79.273304
3	Willowdale, Willowdale East	M2N	North York	43.770120	-79.408493
4	Lawrence Manor, Lawrence Heights	M6A	North York	43.718518	-79.464763
5	Garden District, Ryerson	M5B	Downtown Toronto	43.657162	-79.378937
6	Summerhill West, Rathnelly, South Hill, Forest...	M4V	Central Toronto	43.686412	-79.400049

▪ Asia Restaurants Freq/Sum Neighborhoods

This is a sub DataFrame from Toronto Venues One-hot Dataframe that filter out Asia Restaurants. After that, I use aggregation functions such as mean and sum to get the 2 final tables. In fact, Asia Restaurants Freq DataFrame is the main `food` for Machine Learning.

	Neighborhood	Vietnamese Restaurant	Thai Restaurant	Ramen Restaurant	Asian Restaurant	Chinese Restaurant	Japanese Restaurant	Korean Restaurant	Korean BBQ Restaurant	Taiwanese Restaurant	Dim S Restau
0	Agincourt	0	0	0	0	0	0	0	0	0	
1	Alderwood, Long Branch	0	0	0	0	0	0	0	0	0	
2	Bathurst Manor, Wilson Heights, Downsview North	0	0	0	0	0	0	0	0	0	
3	Bayview Village	0	0	0	0	1	1	0	0	0	
4	Bedford Park, Lawrence Manor East	0	1	0	0	0	0	0	0	0	

b. Data Exploration

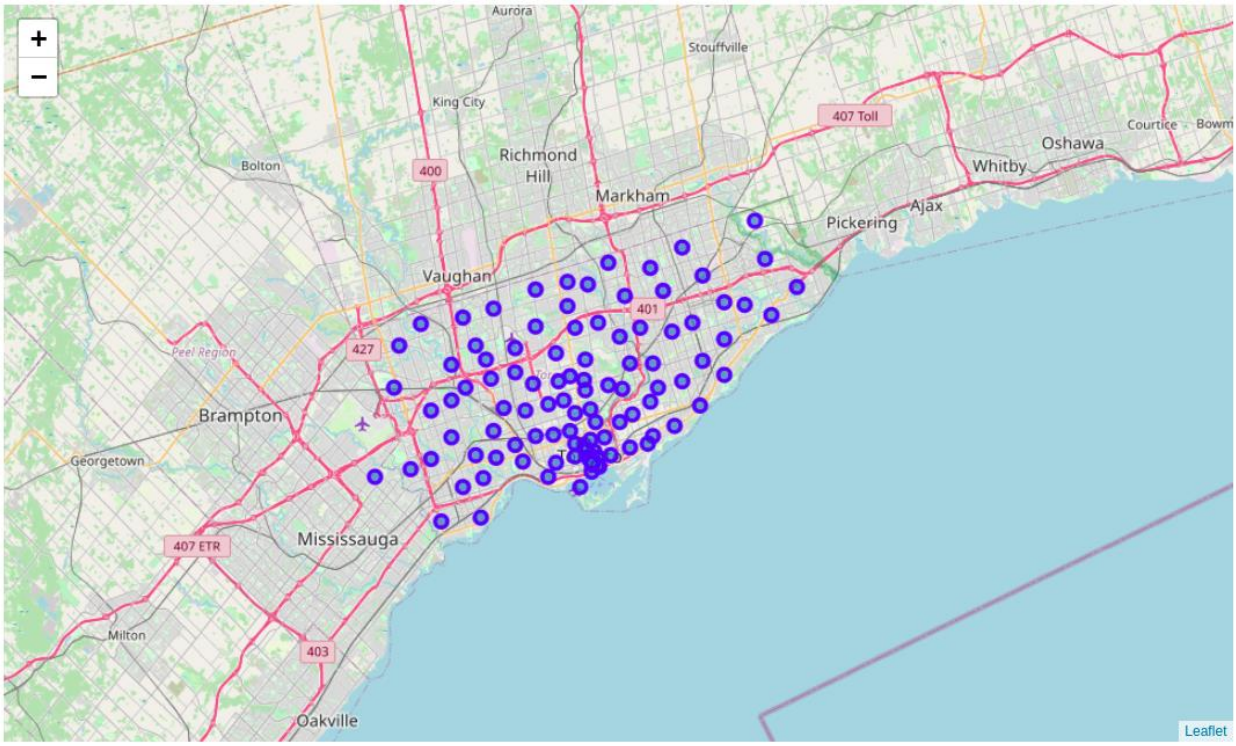
In this phase, I would do two things:

- Drawing Toronto map and marking all neighborhoods on the map.
- Drawing Toronto map and marking all neighborhoods have at least one Vietnamese restaurant.

Before diving into those things, I would like to answer you guys a question: How do I draw the map? And the answer is: I use the Folium package to draw the map.

Toronto Neighborhoods

To drawing this map, I need to loop through all neighborhoods in Toronto Neighborhoods DataFrame, and feed latitude, longitude each of them for Folium to draw the map.



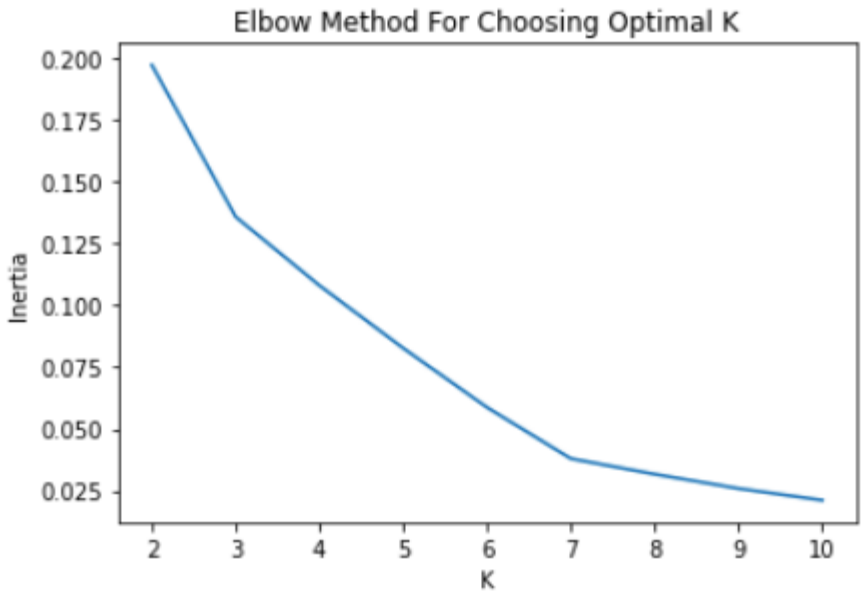
Neighborhoods have Vietnamese Restaurants

Just as drawing Toronto Neighborhoods, I loop through Vietnamese Restaurants Neighborhoods DataFrame, continually

feed latitude and longitude each of Neighborhoods in the table to Folium.

c. Machine Learning

As mentioned above, I would use Asia Restaurants Freq DataFrame for Machine Learning. To make the analysis more interesting, we wanted to cluster the neighborhoods based on the neighborhoods that had similar averages of Asia 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 the sharpest turn. In our case, we had the Elbow Point at K = **7**. That means we will have a total of **7** clusters.

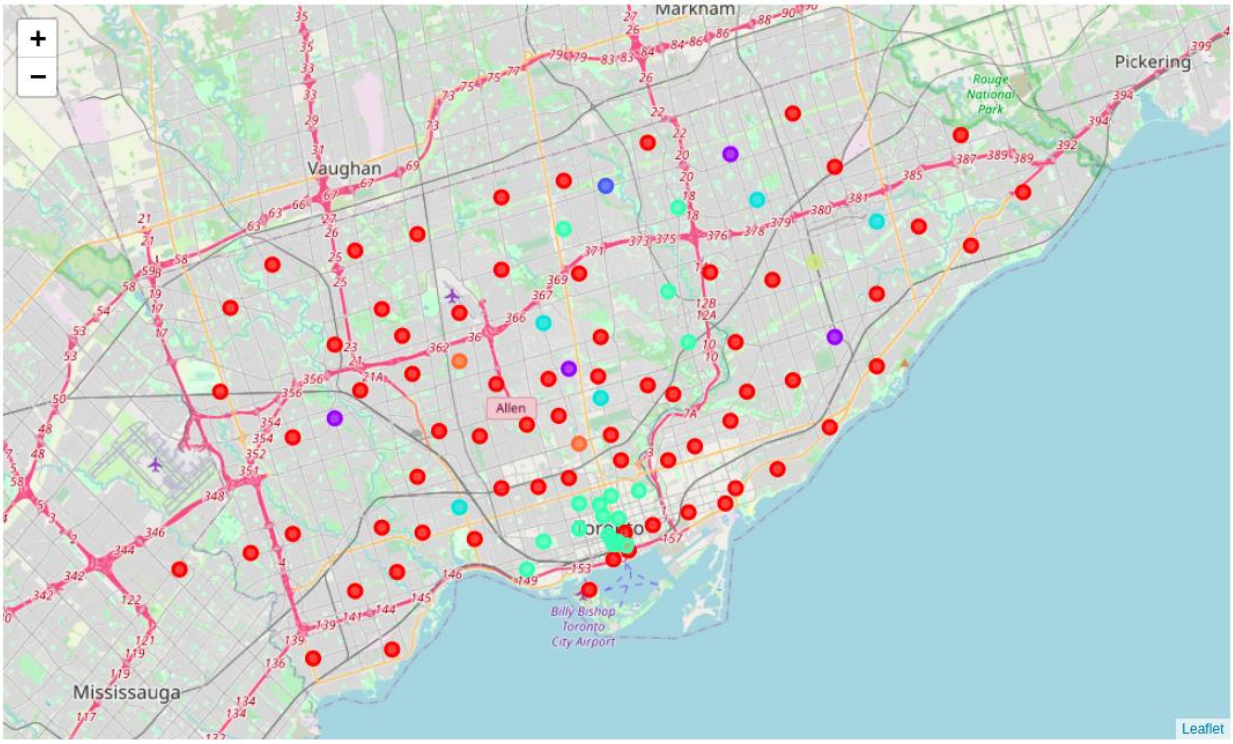


Neighborhoods that had a similar mean frequency of Asia Restaurants were divided into 7 clusters. Each of these clusters was labelled from 0 to 6 as the indexing of labels begins with 0 instead of 1.

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels
0	M3A	North York	Parkwoods	43.753259	-79.329656	0.0
1	M4A	North York	Victoria Village	43.725882	-79.315572	0.0
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	0.0
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	6.0
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	4.0

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 Vietnamese Restaurant in Toronto. Then we created a map using the Folium package in Python and each neighborhood was colored based on the cluster label.



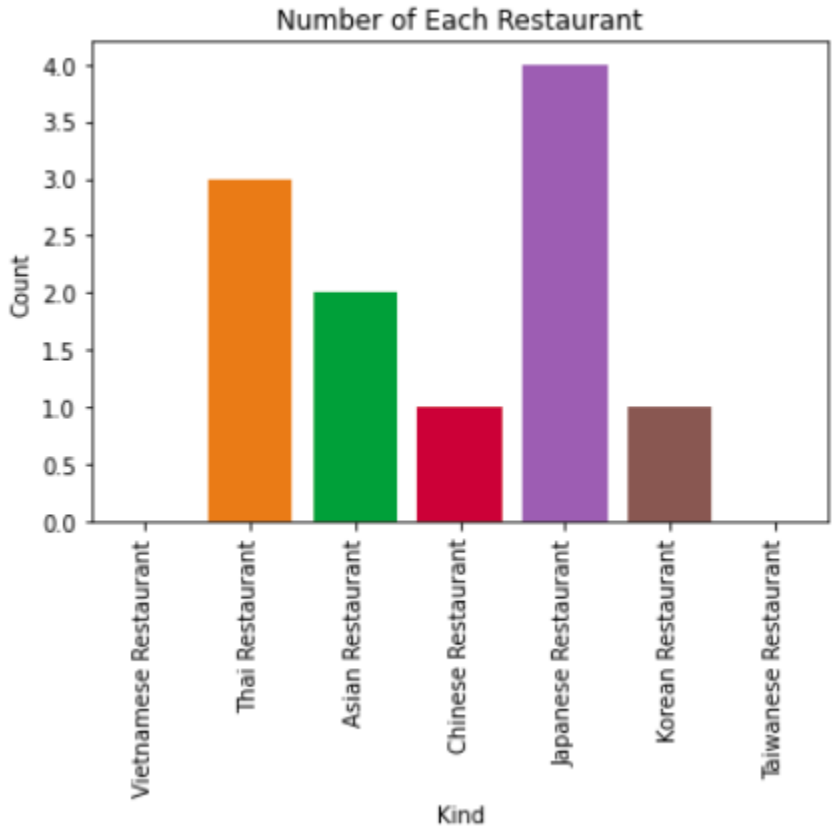
d. Data Analyst

I eliminate all clusters which only have either 1 or 2 Neighborhoods. After cleansing a little bit, there are 4 clusters remain: 0, 1, 3, 4. If you want to take a look at the others clusters, try to see my notebook.

Cluster 0

	PostalCode	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
0	M3A	North York	Parkwoods	43.753259	-79.329656	0	Vietnamese Restaurant	Thai Restaurant	Asian Restaurant	Chinese Restaurant	Japanese Restaurant
1	M4A	North York	Victoria Village	43.725882	-79.315572	0	Vietnamese Restaurant	Thai Restaurant	Asian Restaurant	Chinese Restaurant	Japanese Restaurant
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	0	Asian Restaurant	Vietnamese Restaurant	Thai Restaurant	Chinese Restaurant	Japanese Restaurant
6	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353	0	Vietnamese Restaurant	Thai Restaurant	Asian Restaurant	Chinese Restaurant	Japanese Restaurant
8	M4B	East York	Parkview Hill, Woodbine Gardens	43.706397	-79.309937	0	Vietnamese Restaurant	Thai Restaurant	Asian Restaurant	Chinese Restaurant	Japanese Restaurant

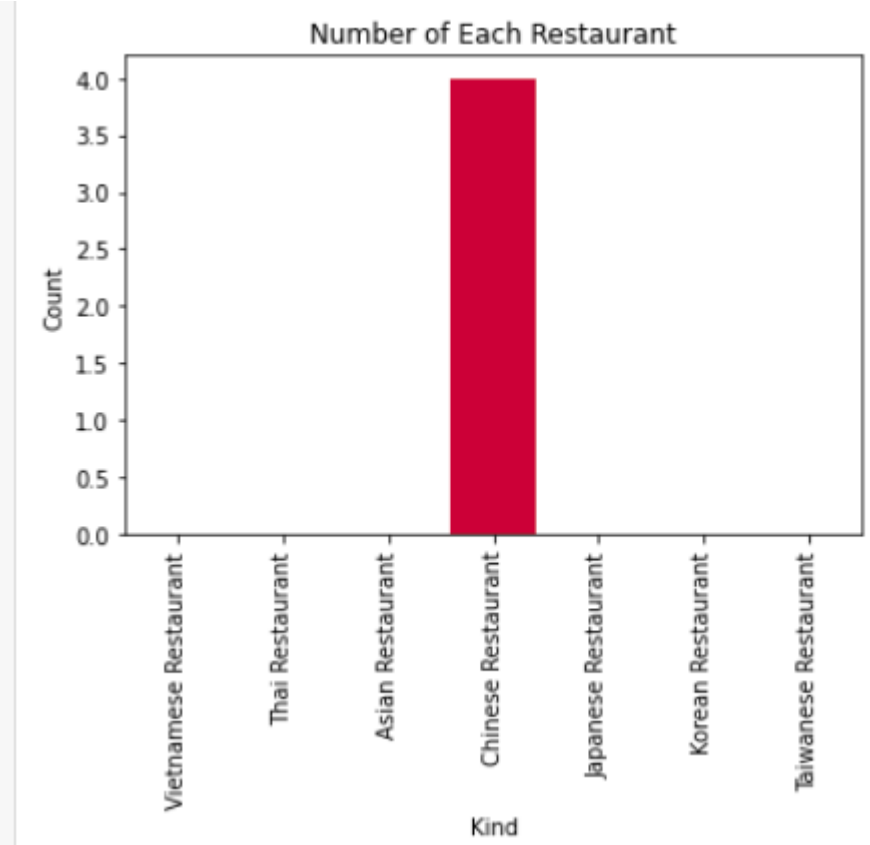
There are 69 Neighborhoods in this cluster that are scattered distribution over Toronto. To get more insights, I try to plot the bar chart that shows number of each Restaurant's types in this cluster.



As the bar chart showing, I can clearly see that there is no Vietnamese Restaurant in this cluster, and this cluster has diverse types of restaurants. So, it may be a desirable choice for our new business.

Cluster 1

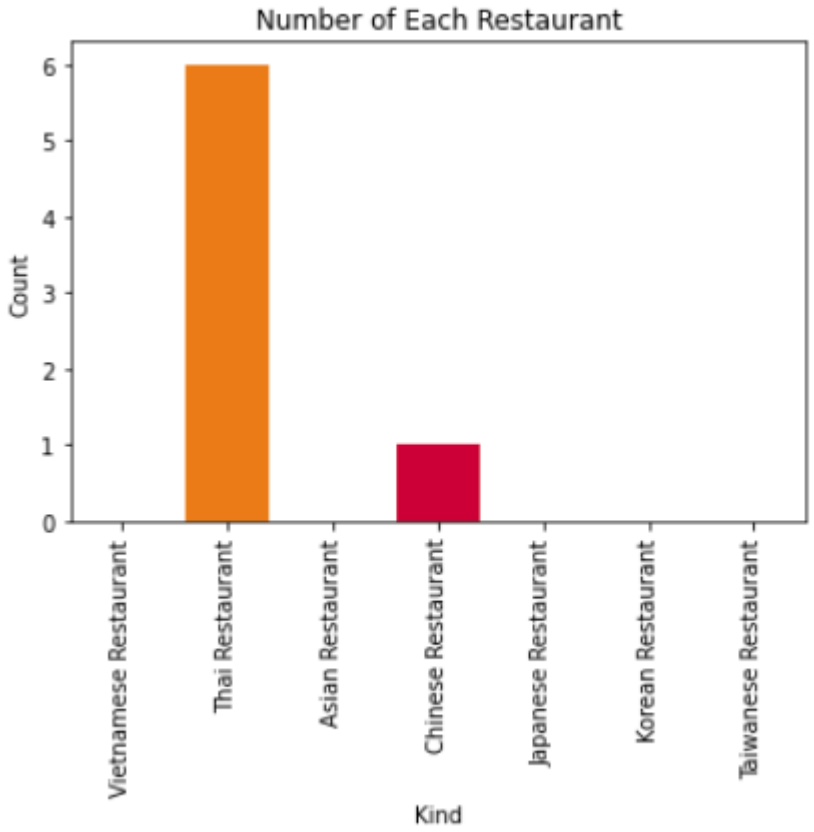
	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
38	M1K	Scarborough	Kennedy Park, Ionview, East Birchmount Park	43.727929	-79.262029	1	Chinese Restaurant	Vietnamese Restaurant	Thai Restaurant	Asian Restaurant	Japanese Restaurant
70	M9P	Etobicoke	Westmount	43.696319	-79.532242	1	Chinese Restaurant	Vietnamese Restaurant	Thai Restaurant	Asian Restaurant	Japanese Restaurant
73	M4R	Central Toronto	North Toronto West, Lawrence Park	43.715383	-79.405678	1	Chinese Restaurant	Vietnamese Restaurant	Thai Restaurant	Asian Restaurant	Japanese Restaurant
90	M1W	Scarborough	Steeles West, L'Amoreaux West	43.799525	-79.318389	1	Chinese Restaurant	Vietnamese Restaurant	Thai Restaurant	Asian Restaurant	Japanese Restaurant



In this cluster 1, it only has 4 neighborhoods that have only one type of restaurants: Chinese Restaurant. So, there is no doubt about this cluster can be called Chinese Cluster. In addition, there are no Vietnamese restaurants here and similar in Vietnamese and China cuisine make this cluster becomes a quite desirable choice for our business.

Cluster 3

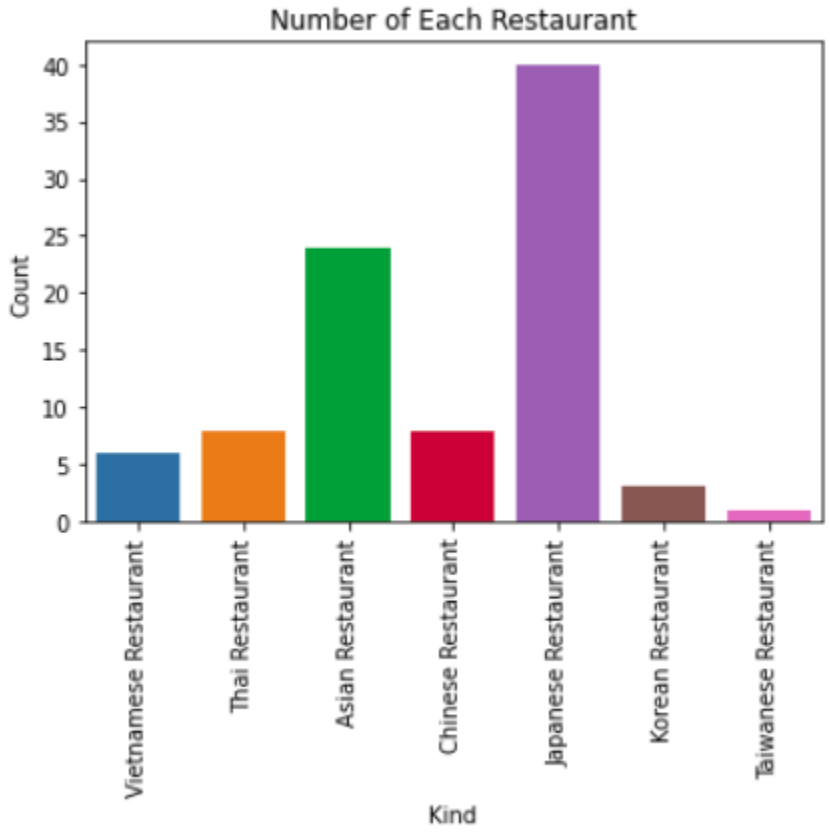
	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
26	M1H	Scarborough	Cedarbrae	43.773136	-79.239476	3	Thai Restaurant	Vietnamese Restaurant	Asian Restaurant	Chinese Restaurant	Japanese Restaurant
55	M5M	North York	Bedford Park, Lawrence Manor East	43.733283	-79.419750	3	Thai Restaurant	Vietnamese Restaurant	Asian Restaurant	Chinese Restaurant	Japanese Restaurant
69	M6P	West Toronto	High Park, The Junction South	43.661608	-79.464763	3	Thai Restaurant	Vietnamese Restaurant	Asian Restaurant	Chinese Restaurant	Japanese Restaurant
79	M4S	Central Toronto	Davisville	43.704324	-79.388790	3	Thai Restaurant	Vietnamese Restaurant	Asian Restaurant	Chinese Restaurant	Japanese Restaurant
82	M1T	Scarborough	Clarks Corners, Tam O'Shanter, Sullivan	43.781638	-79.304302	3	Thai Restaurant	Chinese Restaurant	Vietnamese Restaurant	Asian Restaurant	Japanese Restaurant



Just like the Chinese cluster, this cluster can be called the Thai cluster because it has Thai restaurants at most.

Cluster 4

	PostalCode	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
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	4	Japanese Restaurant	Vietnamese Restaurant	Thai Restaurant	Asian Restaurant	Chinese Restaurant
7	M3B	North York	Don Mills	43.745906	-79.352188	4	Asian Restaurant	Chinese Restaurant	Japanese Restaurant	Vietnamese Restaurant	Thai Restaurant
9	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937	4	Thai Restaurant	Japanese Restaurant	Vietnamese Restaurant	Chinese Restaurant	Asian Restaurant
13	M3C	North York	Don Mills	43.725900	-79.340923	4	Asian Restaurant	Chinese Restaurant	Japanese Restaurant	Vietnamese Restaurant	Thai Restaurant
24	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383	4	Japanese Restaurant	Korean Restaurant	Vietnamese Restaurant	Thai Restaurant	Asian Restaurant
30	M5H	Downtown Toronto	Richmond, Adelaide, King	43.650571	-79.384568	4	Asian Restaurant	Japanese Restaurant	Thai Restaurant	Chinese Restaurant	Vietnamese Restaurant
33	M2J	North York	Fairview, Henry Farm, Oriole	43.778517	-79.346556	4	Japanese Restaurant	Asian Restaurant	Vietnamese Restaurant	Thai Restaurant	Chinese Restaurant



This cluster rounds up Downtown Toronto, and it has 18 neighborhoods in total. These number of each type of restaurants seem to be high because these neighborhoods near the other.

4. CONCLUSION

In conclusion, to end off this project, we had an opportunity on a business problem, and it was tackled in a way that it was similar to how a genuine data scientist would do. We utilized numerous Python libraries to fetch the information, control the content and break down and visualize those datasets. We have utilized Foursquare API to investigate the settings in neighborhoods of Toronto, get a great measure of data from Wikipedia which we scraped with the BeautifulSoup Web scraping Library. We also visualized utilizing different plots present in seaborn and Matplotlib libraries. Similarly, we applied AI strategy to anticipate the error given the information and utilized Folium to picture it on a map.

Places that have room for improvement or certain drawbacks give us that this project can be additionally improved with the assistance of more information and distinctive Machine Learning strategies. Additionally, we can utilize this venture to investigate any situation, for example, opening an alternate cuisine or opening of a Movie Theater and so forth. Ideally, this task acts as an initial direction to tackle more complex real-life problems using data science.