Finding the best area in Toronto for a Thai Restaurant

Sun Techamahachai

June 17, 2020

1. Introduction

1.1 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. Multiculturalism is seen through the various neighborhoods including; Chinatown, Corso Italia, Little India, Kensington Market, Little Italy, Koreatown, Little Thai and many more. Downtown Toronto being the hub of interactions between ethnicities brings many opportunities for entrepreneurs to start or grow their business. It is a place where people can try the best of each culture, either while they work or just passing through. Toronto is well known for its great food.

1.2 Problem

The objective of this project is to use Foursquare location data and regional clustering of venue information to determine what might be the 'best' neighborhood in Toronto to open a restaurant. Pad Thai and Pad Gra Prao are one of the most bought dishes in Toronto originating from Thailand. Toronto is one of the largest cities in the world, there are numerous opportunities to open a new Thai restaurant. Through this project, we will find the most suitable location for an entrepreneur to open a new Thai restaurant in Toronto, Canada.

1.3 Interest

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

2. Data acquisition and cleaning

2.1 Data sources

The Wikipedia site shown below 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).

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

	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

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 and 4).

https://cocl.us/Geospatial_data

	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

Figure 3: Conversion of file into Pandas data frame

	A	B	C
1	Postal Code	Latitude	Longitude
2	M1B	43.8066863	-79.1943534
3	M1C	43.7845351	-79.1604971
4	M1E	43.7635726	-79.1887115
5	M1G	43.7709921	-79.2169174
6	M1H	43.773136	-79.2394761
7	M1J	43.7447342	-79.2394761

Figure 4: Geographical data of Neighborhoods in Toronto

2.2 Data Cleansing

We performed a bit of data cleansing. It is seen through figure 5 (below) that the neighborhoods are grouped by the name of the neighborhood, so data clustering is made easier later on.

	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

Figure 5: Venue data pulled from Foursquare explore API

After all the data was collected and put into data frames, cleansing and merging of the data was required to start the process of analysis. When getting the data 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. Borough's that were not assigned get ignored.

2. More than one neighborhood can exist in one postal code area. For example, in the table on the Wikipedia page, you will notice that M5A is listed twice and has two neighborhood: Harbourfront and Regent Park. These two rows will be combined into one row with the neighborhoods separated with a comma as shown in Figure2 row 4.

3. If a cell has a borough but a Not assigned 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

Figure 6: Rows grouped together based on 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

Figure 7: Merging tables together 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 shown below.

	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 Beaches	43.676357	-79.293031	Seaspray Restaurant	43.678888	-79.298167	Asian Restaurant

Figure 8: Local Venues near the respective Neighborhood

3. Exploratory Data Analysis

3.1 Calculation of target variable

Now after cleansing the data, the next step was to analyze it. We then created a map using Folium and color-coded each Neighborhood depending on what Borough it was located in.

This snippet of code provided us with the map below:

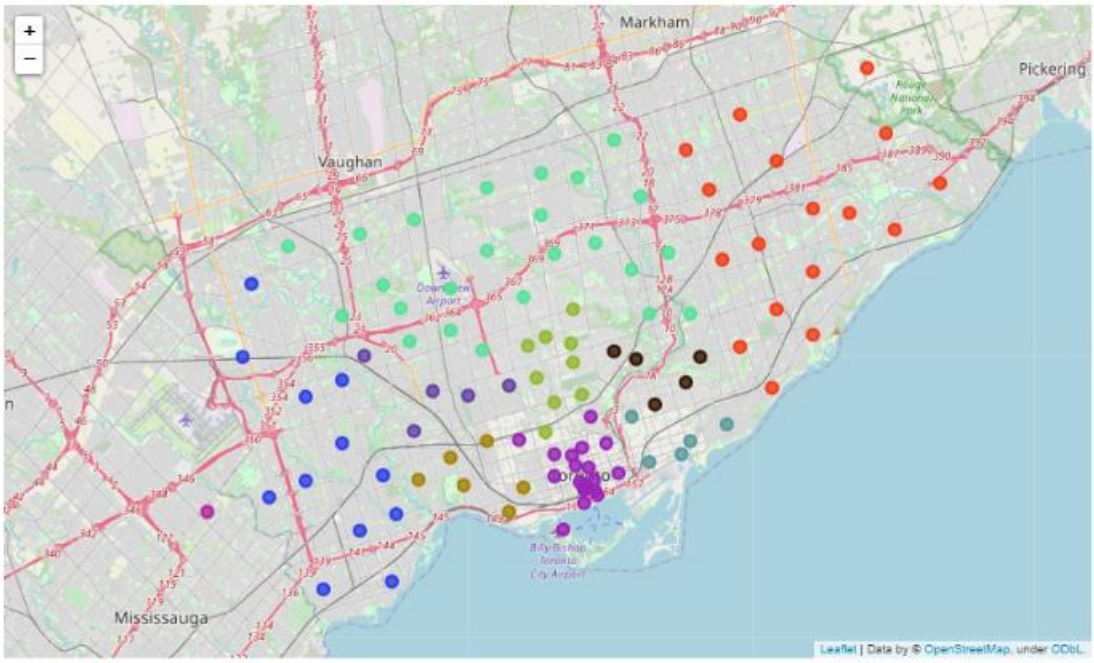

Figure 9: Map used the Foursquare API

Next, we used the Foursquare API to get a list of all the Venues in Toronto which included Parks, Schools, Café Shops, European Restaurants etc. Getting this data was crucial to analyzing the number of Thai Restaurants all over Toronto. There was a total of 21 Thai 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.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Lawrence Park	43.728020	-79.388790	Lawrence Park Ravine	43.726963	-79.394382	Park
1	Lawrence Park	43.728020	-79.388790	Zodiac Swim School	43.728532	-79.382860	Swim School
2	Lawrence Park	43.728020	-79.388790	TTC Bus #162 - Lawrence-Donway	43.728026	-79.382805	Bus Line
3	Davisville North	43.712751	-79.390197	Sherwood Park	43.716551	-79.387776	Park
4	Davisville North	43.712751	-79.390197	Homeway Restaurant & Brunch	43.712641	-79.391557	Breakfast Spot

Figure 10: Venue table merged with Neighborhood data

3.2 Category of Neighborhoods

To analyze the data we performed a technique in which Categorical Data is transformed into Numerical Data. This technique is called One hot encoding. For each of the neighborhoods, individual venues were turned into the frequency at how many of those Venues were located in each neighborhood.

	Neighborhoods	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	...
0	Lawrence Park	0	0	0	0	0	0	0	0	0	...
1	Lawrence Park	0	0	0	0	0	0	0	0	0	...
2	Lawrence Park	0	0	0	0	0	0	0	0	0	...
3	Davisville North	0	0	0	0	0	0	0	0	0	...
4	Davisville North	0	0	0	0	0	0	0	0	0	...

Figure 11: One Hot Encoding

Then we grouped those rows by Neighborhood and by taking the average of the frequency of occurrence of each Venue Category.

	Neighborhoods	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	...
0	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...
1	Alderwood, Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...
2	Bathurst Manor, Wilson Heights, Downsview North	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...
3	Bayview Village	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...
4	Bedford Park, Lawrence Manor East	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.043478	...

Figure 12: Grouped Neighborhoods by the average of the frequency of each Venue

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

	Neighborhoods	Thai Restaurant
0	Agincourt	0.000000
1	Alderwood, Long Branch	0.000000
2	Bathurst Manor, Wilson Heights, Downsview North	0.000000
3	Bayview Village	0.000000
4	Bedford Park, Lawrence Manor East	0.043478

Figure 13: New data frame storing Neighborhoods and the average Thai Restaurant in that Neighborhood

4. Predicting Model

To make the analysis more interesting, we wanted to cluster the neighborhoods based on the neighborhoods that had similar averages of Thai 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.

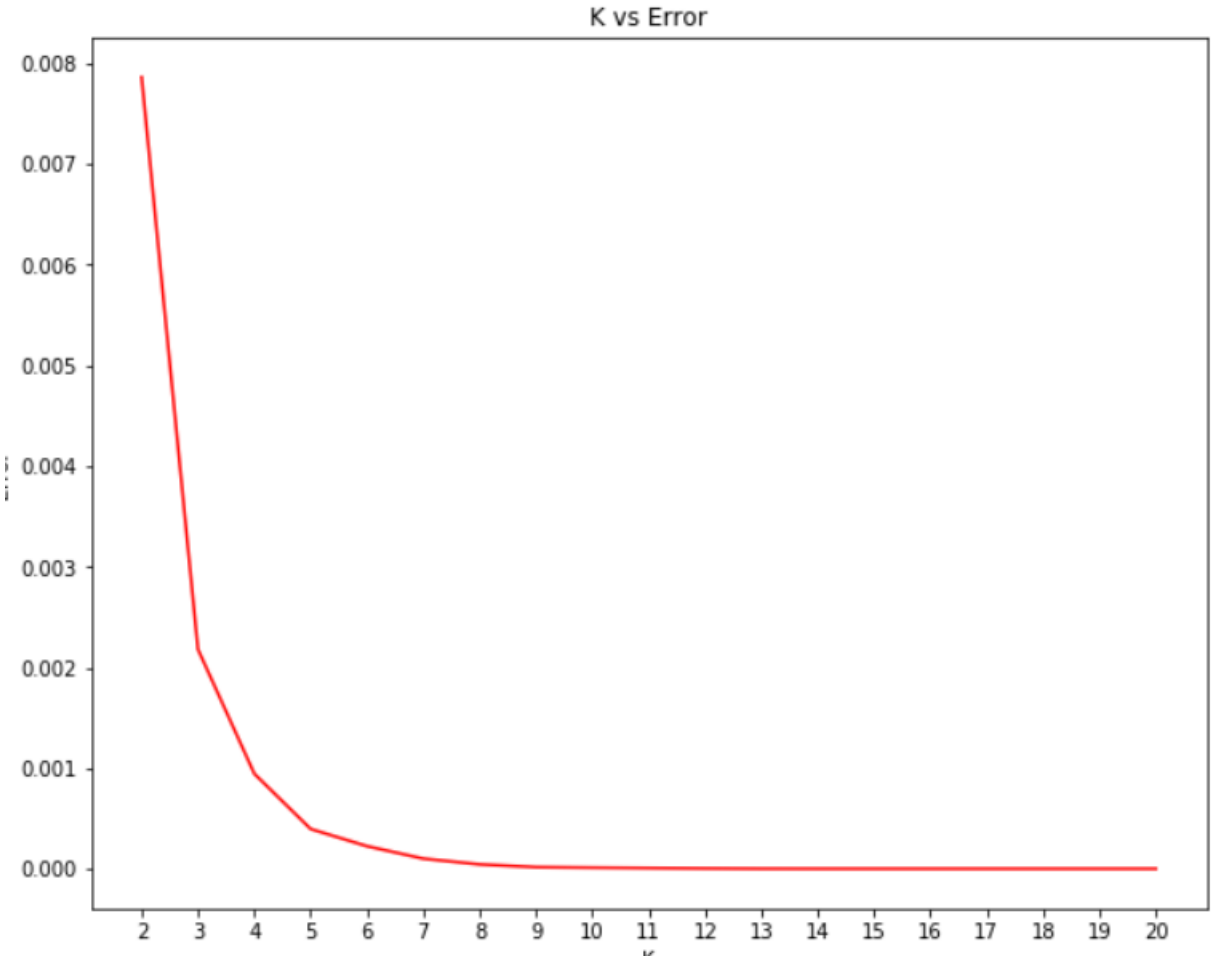

Figure 14: Finding the K vs Error Values

Neighborhoods that had a similar mean frequency of Thai 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.

	Neighborhood	Thai Restaurant	Cluster Labels
0	Agincourt	0.000000	0
1	Alderwood, Long Branch	0.000000	0
2	Bathurst Manor, Wilson Heights, Downsview North	0.000000	0
3	Bayview Village	0.000000	0
4	Bedford Park, Lawrence Manor East	0.043478	2

Figure 15: Appropriate Cluster Labels were added

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 Thai Restaurant in Toronto. Then we created each neighborhood was colored based on the cluster label.

- Cluster 1 — Red
- Cluster 2 — Purple
- Cluster 3 — Turquoise
- Cluster 4 — Dark Khaki

We have a total of 4 clusters (1,2,3,4). Before we analyze them one by one let's check the total amount of neighborhoods in each cluster and the average Thai Restaurants in that cluster. From the bar graph that was made using Matplotlib (figure 16), we can compare the number of Neighborhoods per Cluster. We see that Cluster 4 has the least neighborhoods (1) while cluster 1 has the most (85). Cluster 3 has 9 neighborhoods and cluster 2 has only 2. Then we compared the average Thai Restaurants per cluster.

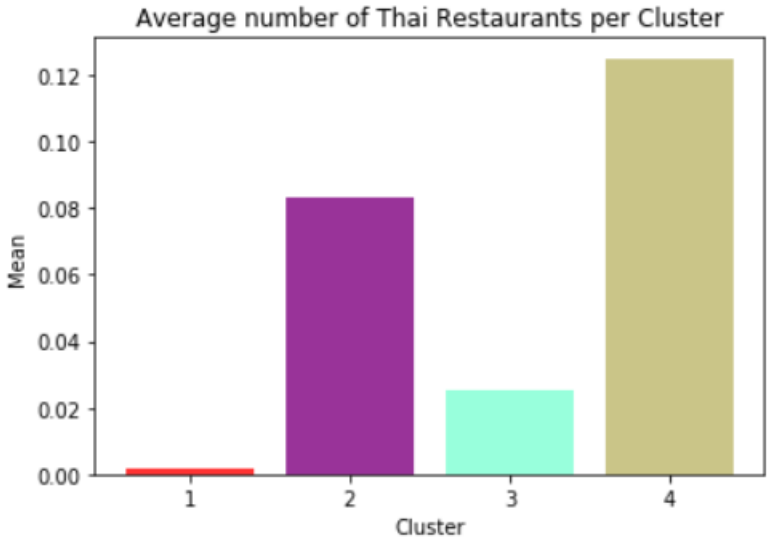

Figure 16: Average Thai restaurant in each neighborhood

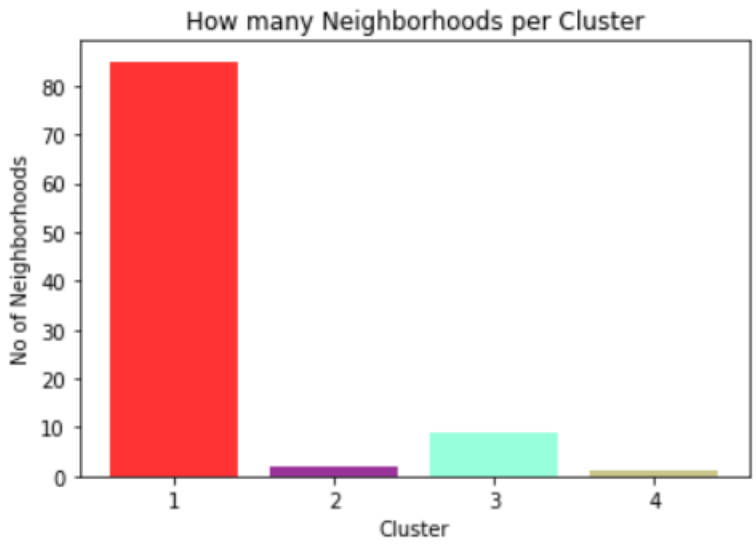

Figure 17: Number of neighborhoods per cluster

Cluster Analysis

This information is crucial as we can see that even though there is only 1 neighborhood in Cluster 4, it has the highest number of Thai Restaurants (0.1250) while Cluster 1 has the most neighborhoods but has the least average of Thai Restaurants (0.0000). The average of the average Thai Restaurant made up the data for Figure 16. We can also see that neighborhoods in Cluster 1 are the most sparsely populated. Now let's analyze the Clusters individually (Note: these are just snippets of the data).

Cluster 1(Red)

	Borough	Neighborhood	Thai Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Central Toronto	Lawrence Park	0.000000	0	43.728020	-79.388790	TTC Bus #162 - Lawrence-Donway	43.728026	-79.382805	Bus Line
1	Central Toronto	Lawrence Park	0.000000	0	43.728020	-79.388790	Zodiac Swim School	43.728532	-79.382860	Swim School
2	Central Toronto	Lawrence Park	0.000000	0	43.728020	-79.388790	Lawrence Park Ravine	43.726963	-79.394382	Park
3	Central Toronto	Davisville North	0.000000	0	43.712751	-79.390197	Gym	43.713126	-79.393537	Gym
4	Central Toronto	Davisville North	0.000000	0	43.712751	-79.390197	Subway	43.708474	-79.390674	Sandwich Place
5	Central Toronto	Davisville North	0.000000	0	43.712751	-79.390197	Best Western Roehampton Hotel & Suites	43.708878	-79.390880	Hotel
6	Central Toronto	Davisville North	0.000000	0	43.712751	-79.390197	Winners	43.713236	-79.393873	Department Store

Cluster 1 was in the Central Toronto area. Lawrence Park, Davisville North, North Toronto West, Davisville Moore Park and Summerhill East were the five Neighborhoods that were in that cluster.

Cluster 1 had 50 unique Venue locations and out of those was no Thai Restaurant. Cluster 1 had the lowest average of Thai Restaurants equating to 0.0000. The reason why the average of Thai Restaurants is the lowest is that no Thai Restaurant is in five neighborhoods, Lawrence Park, Davisville North, North Toronto West, Davisville Moore Park and Summerhill East.

Cluster 2 (Purple)

	Borough	Neighborhood	Thai Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
18	West Toronto	High Park, The Junction South	0.086957	1	43.661608	-79.464763	Isaan Der	43.665311	-79.468078	Thai Restaurant
27	West Toronto	High Park, The Junction South	0.086957	1	43.661608	-79.464763	Junction Flea	43.665258	-79.462868	Flea Market
21	West Toronto	High Park, The Junction South	0.086957	1	43.661608	-79.464763	The Beet Organic Café	43.665340	-79.467137	Café
22	West Toronto	High Park, The Junction South	0.086957	1	43.661608	-79.464763	Junction City Music Hall	43.665334	-79.466253	Music Venue
23	West Toronto	High Park, The Junction South	0.086957	1	43.661608	-79.464763	Mjölk	43.665432	-79.467962	Furniture / Home Store
24	West Toronto	High Park, The Junction South	0.086957	1	43.661608	-79.464763	famous last words	43.665181	-79.468471	Speakeasy
25	West Toronto	High Park, The Junction South	0.086957	1	43.661608	-79.464763	ARTiculations	43.665550	-79.467194	Arts & Crafts Store
26	West Toronto	High Park, The Junction South	0.086957	1	43.661608	-79.464763	SMASH	43.665496	-79.465537	Antique Shop
28	West Toronto	High Park, The Junction South	0.086957	1	43.661608	-79.464763	Pascal's Baguette & Bagels	43.665426	-79.466176	Bakery
19	West Toronto	High Park, The Junction South	0.086957	1	43.661608	-79.464763	Tim & Sue's No Frills	43.664243	-79.468643	Grocery Store

There was a total of 2 neighborhoods and only 2 Thai Restaurant. Therefore, the average amount of Thai Restaurants that were near the venues in Cluster 2 is 0.0869. We can see that nodes of Cluster 3 were dispersed all throughout Toronto making it one of the most sparsely populated clusters.

Cluster 3 (Turquoise)

	Borough	Neighborhood	Thai Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Central Toronto	Davisville	0.029412	2	43.704324	-79.388790	Jules Cafe Patisserie	43.704138	-79.388413	Dessert Shop
1	Central Toronto	Davisville	0.029412	2	43.704324	-79.388790	souvlaki express	43.707378	-79.389848	Greek Restaurant
2	Central Toronto	Davisville	0.029412	2	43.704324	-79.388790	Second Cup	43.704344	-79.388659	Coffee Shop
3	Central Toronto	Davisville	0.029412	2	43.704324	-79.388790	Petro-Canada	43.702269	-79.387955	Gas Station
4	Central Toronto	Davisville	0.029412	2	43.704324	-79.388790	Apple Tree Farmer's Market	43.700326	-79.389760	Farmers Market
5	Central Toronto	Davisville	0.029412	2	43.704324	-79.388790	Starving Artist	43.701538	-79.387240	Restaurant
6	Central Toronto	Davisville	0.029412	2	43.704324	-79.388790	Marcheleo's Gourmet Marketplace	43.708041	-79.392195	Gourmet Shop
7	Central Toronto	Davisville	0.029412	2	43.704324	-79.388790	Hazel's Diner	43.702103	-79.387618	Diner
8	Central Toronto	Davisville	0.029412	2	43.704324	-79.388790	Meow Cat Cafe	43.702927	-79.388190	Café
9	Central Toronto	Davisville	0.029412	2	43.704324	-79.388790	Shoppers Drug Mart	43.707806	-79.389893	Pharmacy
10	Central Toronto	Davisville	0.029412	2	43.704324	-79.388790	Pizza Pizza	43.706138	-79.389292	Pizza Place
11	Central Toronto	Davisville	0.029412	2	43.704324	-79.388790	Cafe Pleiade	43.703026	-79.388018	French Restaurant

Cluster 3 had the second to lowest average of Thai Restaurants. Cluster 3 was mainly located in the Downtown area but also had some neighborhoods in West Toronto, East Toronto and in

North York. Neighborhoods 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 129 unique venues and out of those 14 were Thai Restaurants.

Cluster 4 (Dark Khaki)

	Borough	Neighborhood	Thai Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Scarborough	Cedarbrae	0.125	3	43.773136	-79.239476	Popeyes Louisiana Kitchen	43.775930	-79.235328	Fried Chicken Joint
1	Scarborough	Cedarbrae	0.125	3	43.773136	-79.239476	B&A Bakery	43.774391	-79.243877	Bakery
2	Scarborough	Cedarbrae	0.125	3	43.773136	-79.239476	TD Canada Trust	43.774830	-79.241251	Bank
3	Scarborough	Cedarbrae	0.125	3	43.773136	-79.239476	Centennial Recreation Centre	43.774593	-79.236500	Athletics & Sports
4	Scarborough	Cedarbrae	0.125	3	43.773136	-79.239476	Thai One On	43.774468	-79.241268	Thai Restaurant
5	Scarborough	Cedarbrae	0.125	3	43.773136	-79.239476	Drupati's Roti & Doubles	43.775222	-79.241678	Caribbean Restaurant
6	Scarborough	Cedarbrae	0.125	3	43.773136	-79.239476	Petro-Canada	43.774106	-79.243097	Gas Station
7	Scarborough	Cedarbrae	0.125	3	43.773136	-79.239476	Federick Restaurant	43.774697	-79.241142	Hakka Restaurant

Cluster 4 venues were located in the Scarborough areas. Cedarbrae neighborhood was some of the neighborhood that made up this cluster. There were a total of 8 unique Venues in Cluster 4 with 1 Thai Restaurants. This made up the highest average of Thai Restaurants in that cluster which was approximately 0.1250.

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

1. Cluster 4 (0.1250)
2. Cluster 2 (0.0869)
3. Cluster 3 (0.0434)
4. Cluster 1 (0.0000)

5. Discussion

Most of the Thai Restaurants are in cluster 4 represented by the dark khaki clusters. The Neighborhoods located in the Scarborough area that have the highest average of Thai Restaurants are Cedarbrae. Even though there is a huge amount of Neighborhoods in cluster 1, there is little to no Thai Restaurant. We see that in the West Toronto area (cluster 2) has the second last average of Thai Restaurants. Looking at the nearby venues, the optimum place to put a new Thai Restaurant is in North York as theirs are many Neighborhoods in the area but little to no Thai Restaurants therefore, eliminating any competition. The second best Neighborhoods that have a great opportunity would be in areas such as St. James Town and Cabbagetown, etc which is in Cluster 3. Having 9 neighborhoods in the area with a little bit Thai Restaurants gives a good opportunity for opening up a new restaurant. This concludes the optimal findings for this project and recommends the entrepreneur to open an authentic Thai restaurant in these locations with little to no competition. Nonetheless, if the food is authentic, affordable and good taste, I am confident that it will have great following everywhere.

6. 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.