# The Battle of Neighbourhoods

Raj

## Introduction

Toronto is the capital city of the Canadian province of Ontario. With a recorded population of 2,731,571 in 2016, it is the most populous city in Canada and the fourth most populous city in North America. The city is the anchor of the Golden Horseshoe, an urban agglomeration of 9,245,438 people (as of 2016) surrounding the western end of Lake Ontario,[18] while the Greater Toronto Area (GTA) proper had a 2016 population of 6,417,516. Toronto is an international centre of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world.

The objective of this project is to use Foursquare location data and regional clustering of venue information to determine what might be the 'best' neighbourhoods in Toronto to open a Yoga Studio. In recent years, yoga has become most popular way to stay healthy.

## Target Audience

Yoga  trainers and individuals who is willing to open a yoga studio.

## 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 neighbourhoods in Toronto (via Wikipedia), the Geographical location of the neighbourhoods (via Geocoder package) and Venue data pertaining to Yoga  studio(via Foursquare). The Venue data will help find which neighbourhoods is best suitable to open a Yoga Studio.

## Methodology

**1.** Extract the data from the data sources.

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 |

***Figure 1:Wikipedia Page showing List of Neighbourhoods in Toronto with respective Postal Codes***

The Wikipedia site (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) shown above, provided almost all the information about the neighbourhoods. It included the postal code, borough and the name of the neighbourhoods 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).

| | 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***

| | 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: Geogaphical data of Neighborhoods in Toronto***

| | 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: Conversion of file into Pandas data frame***

The second source of data provided (https://cocl.us/Geospatial_data) us with the Geographical coordinates of the neighborhoods with the respective Postal Codes. The file was in CSV format, so attaching it to a Pandas data frame was simple (shown in figure 3).
The retrieval of the location, name and category about the various venues in Toronto was collected through the Foursquare explore API. To obtain the data, it was required to make an account where it would provide a 'Secret Key' as well as a 'Client ID' which would allow to pull any data.

| | 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 pull from Foursquare explore API Data***

It is seen through figure 5 (above) that the neighborhoods are grouped by the neighborhood, so data clustering is made easier later on.
Subsequently, 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 neighbourhoods therefore, the following assumptions were made:

1. Only the cells that have an assigned borough will be processed. Borough that is not assigned are ignored.

2. More than one neighbourhood 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 neighbourhoods: Harbourfront and Regent Park. These two rows will be combined into one row with the neighbourhoods separated with a comma as shown in Figure2 row 4.

3. If a cell has a borough but a Not assigned neighbourhoods, then the neighbourhoods will be the same as the borough.

After the implementation of the following assumptions, the rows were grouped based on 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 Neighbourhood***

Now after cleansing the data, the next step was to analyse it. We then created a map using folium and colour coded each Neighbourhood depending on what Borough it was located in.

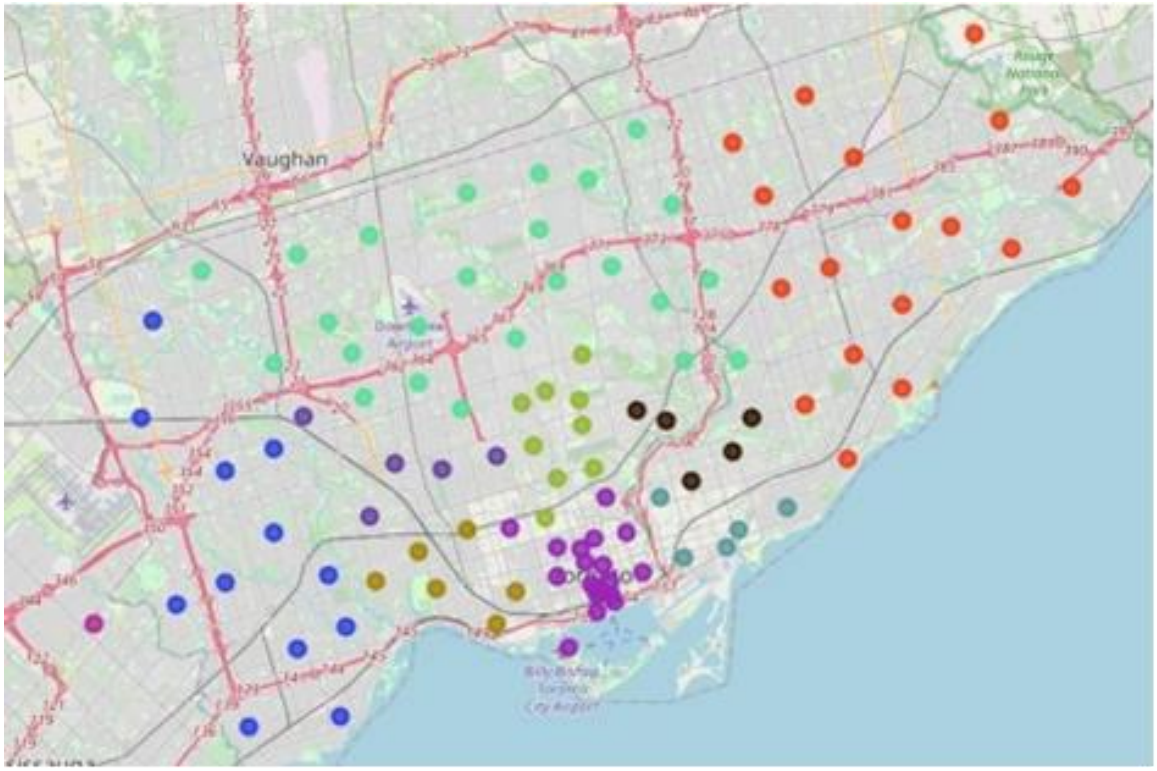


***Figure 9: Toronto Neighborhoods***

Next, we used the Foursquare API to get a list of all the Venues in Toronto which included Parks, Schools, Café Shops, Asian Restaurants etc. Getting this data was crucial to analyzing the number of Yoga Studios all over Toronto. There was not many Yoga Studios 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 | Homeway Restaurant & Brunch | 43.712641 | -79.391557 | Breakfast Spot |
| 4 | Davisville North | 43.712751 | -79.390197 | Sherwood Park | 43.716551 | -79.387776 | Park |

***Figure 10: Venue table merged with neighbourhood data***

Then to analyse the data we performed a technique in which Categorical Data is transformed into Numerical Data for Machine Learning algorithms. This technique is called One hot encoding. For each of the neighbourhoods, individual venues were turned into the frequency at how many of those Venues were located in each neighbourhood.

| | 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 Neighbourhood 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 Neighbourhoods by the average of the frequency of each Venue***

After, we created a new data frame which only stored the Neighbourhood names as well as the mean frequency of Yoga Studios in that Neighbourhood. This allowed the data to be summarized based on each individual Neighbourhood and made the data much simpler to analyse.

| | NeighboUrhoods | Yoga Studio |
|---|---|---|
| 0 | Agincourt | 0.0 |
| 1 | Alderwood, Long Branch | 0.0 |
| 2 | Bathurst Manor, Wilson Heights, Downsview North | 0.0 |
| 3 | Bayview Village | 0.0 |
| 4 | Bedford Park, Lawrence Manor East | 0.0 |

***Figure 13: New data frame storing Neighbourhoods and the average Italian Restaurant in that Neighbourhood***

To make the analysis more interesting, we wanted to cluster the neighbourhoods based on the neighbourhoods that had similar averages of Yoga Studio in that Neighbourhood. 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 a sharpest turn. In our case we had the Elbow Point at K = 4. That means we will have a total of 4 clusters.

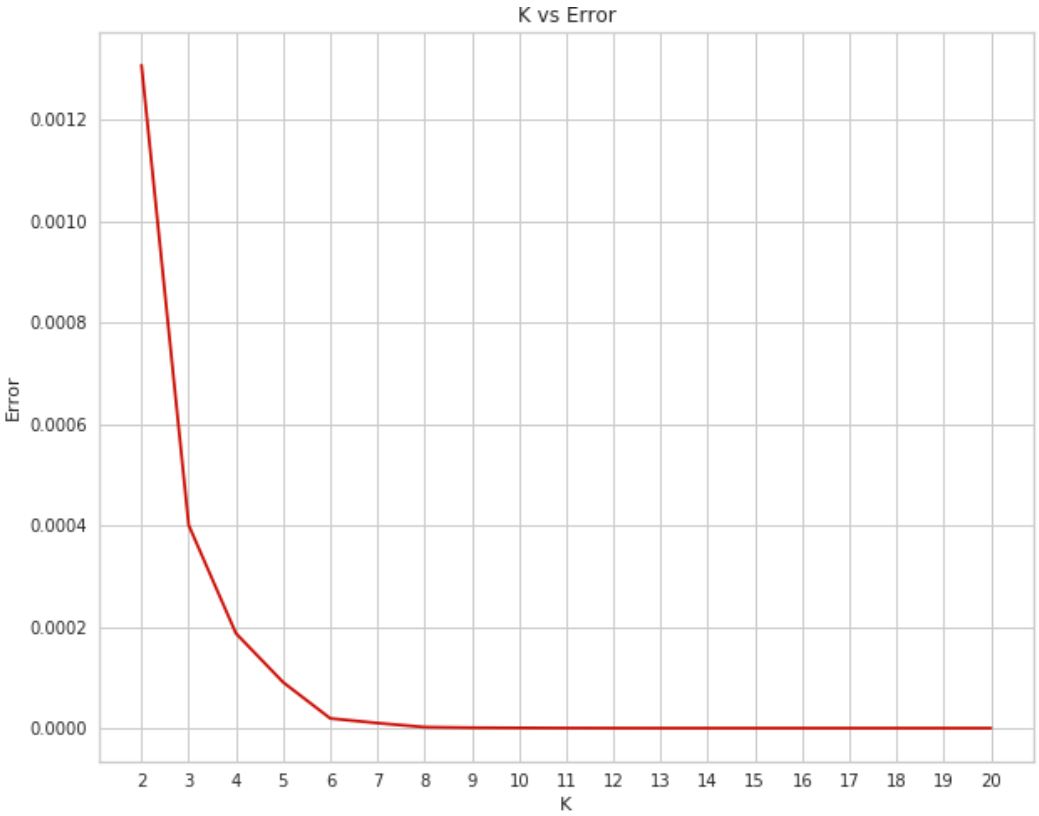


***Figure 14: Finding the K vs Error Values***

We integrated a model which would fit the error and calculate the distortion score. From the dotted line, we see that the Elbow is at K=4. Moreover, in K-Means clustering, objects that are similar based on a certain variable are put into the same cluster. Neighbourhoods that had similar mean frequency of Yoga Studio were divided into 4 clusters. Each of these clusters were labelled from 0 to 3 as the indexing of labels begin with 0 instead of 1.

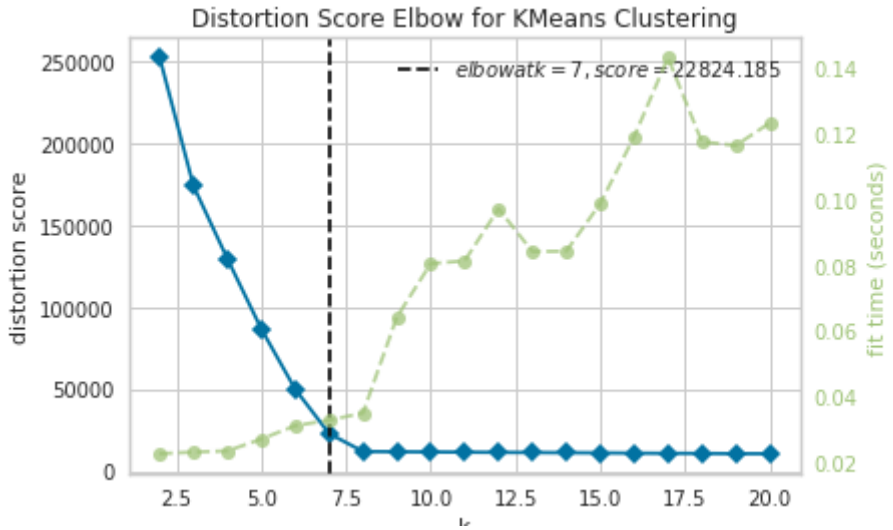


***Figure 15: Finding the right K using Elbow point***

After, we merged the venue data with the table above creating a new table which would be the basis for analysing new opportunities for opening a new Italian Restaurant in Toronto. Then we created a map using the Folium package in Python and each neighborhood was coloured based on the cluster label. For example, cluster 2 was purple and cluster 3 was blue.

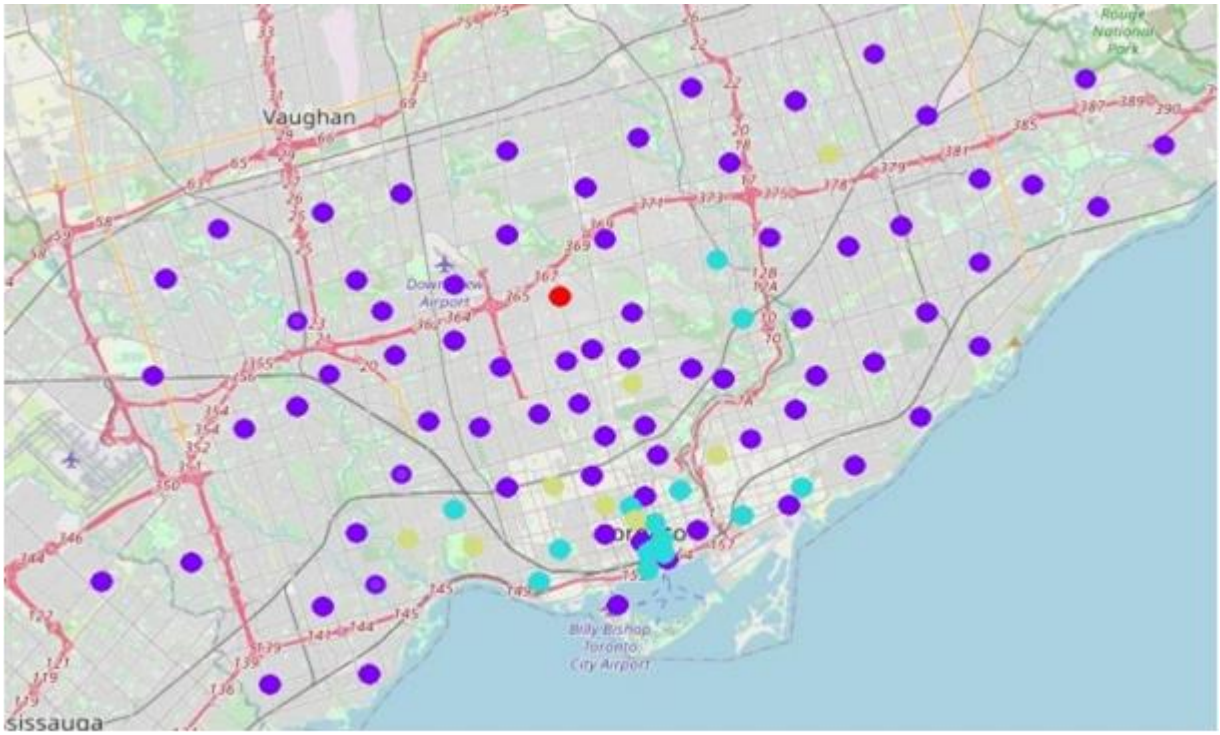


***Figure 16 Map with different clusters***

**Analysis:**

We have a total of 4 clusters (0,1,2,3). Before we analyse them one by one let's check the total amount of neighbourhoods in each cluster and the average Yoga Studio in that cluster. From the bar graph that was made using Matplotlib (figure 18) , we can compare the number of Neighbourhoods per Cluster. We see that Cluster 1 has the least neighbourhood's (1) while cluster 2 has the most (70). Cluster 3 has 14 neighbourhood's and cluster 4 has only 8. Then we compared the average Yoga Studio per cluster.

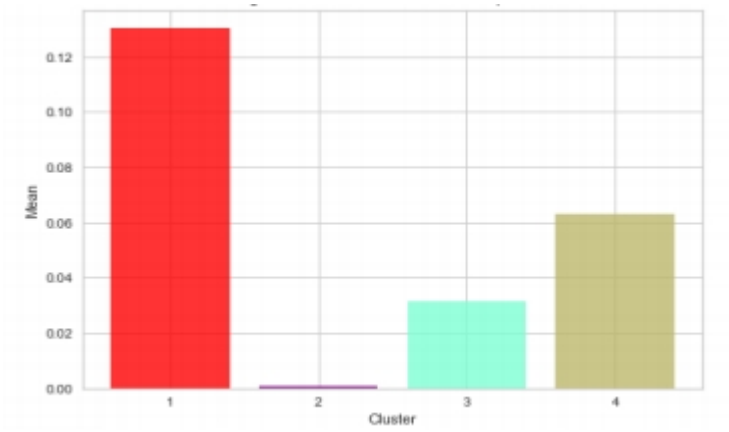


***Figure 19: Average Italian restaurant in each neighborhood***

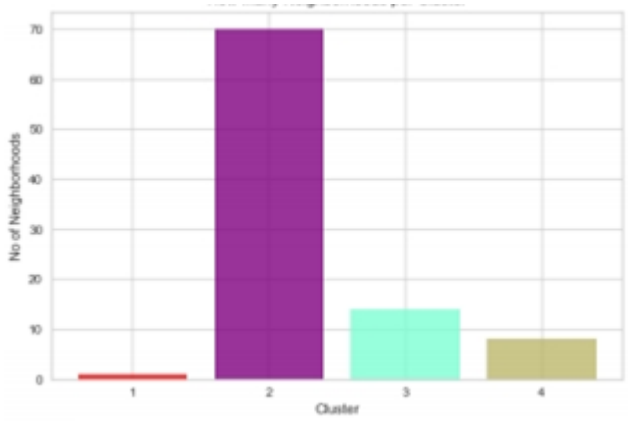


***Figure 20: Number of neighbours per cluster***

This information is crucial as we can see that even through there is only 1 neighbourhoods in Cluster 1, it has the highest number of Yoga Studio (0.1304) while Cluster 2 has the most neighbourhoods but has the least average of Yoga Studio (0.0009). The average of the average Italian Restaurant made up the data for Figure 18. Also, from the map, we can see that neighbourhoods in Cluster 2 are the most sparsely populated. Now let's analyse the Clusters individually (Note: these are just snippets of the data).

| | NeighboUrhood | NeighboUrhood Latitude | NeighboUrhood 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 | Homeway Restaurant & Brunch | 43.712641 | -79.391557 | Breakfast Spot |
| 4 | Davisville North | 43.712751 | -79.390197 | Sherwood Park | 43.716551 | -79.387776 | Park |

| NeighboUrhood | NeighboUrhood Latitude | NeighboUrhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|
| Agincourt | 5 | 5 | 5 | 5 | 5 | 5 |
| Alderwood, Long Branch | 7 | 7 | 7 | 7 | 7 | 7 |
| Bathurst Manor, Wilson Heights, Downsview North | 21 | 21 | 21 | 21 | 21 | 21 |
| Bayview Village | 4 | 4 | 4 | 4 | 4 | 4 |
| Bedford Park, Lawrence Manor East | 22 | 22 | 22 | 22 | 22 | 22 |
| Berczy Park | 55 | 55 | 55 | 55 | 55 | 55 |
| Birch Cliff, Cliffside West | 4 | 4 | 4 | 4 | 4 | 4 |
| Brockton, Parkdale Village, Exhibition Place | 25 | 25 | 25 | 25 | 25 | 25 |
| Business reply mail Processing Centre, South Central Letter Processing Plant Toronto | 16 | 16 | 16 | 16 | 16 | 16 |
| CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport | 17 | 17 | 17 | 17 | 17 | 17 |
| Caledonia-Fairbanks | 4 | 4 | 4 | 4 | 4 | 4 |
| Canada Post Gateway Processing Centre | 14 | 14 | 14 | 14 | 14 | 14 |
| Cedarbrae | 8 | 8 | 8 | 8 | 8 | 8 |
| Central Bay Street | 68 | 68 | 68 | 68 | 68 | 68 |
| Christie | 16 | 16 | 16 | 16 | 16 | 16 |
| Church and Wellesley | 75 | 75 | 75 | 75 | 75 | 75 |
| Clarks Corners, Tam O'Shanter, Sullivan | 12 | 12 | 12 | 12 | 12 | 12 |

| | NeighboUrhoods | Accessories Store | Afghan Restaurant | Airport | Airport Food Court | Airport Gate | Airport Lounge | Airport Service | Airport Terminal | American Restaurant | ... | Train Station | Turkish Restaurant | Vegetaria / Vega Restaurar |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Lawrence Park | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 |
| 1 | Lawrence Park | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 |
| 2 | Lawrence Park | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 |
| 3 | Davisville North | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 |
| 4 | Davisville North | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 |

## Discussion

Most of the Yoga Studio are in cluster 1 represented by the red clusters. The Neighbourhoods located in the North York area that have the highest average of Yoga Studio are Bedford Park and Lawrence Manor East. Even though there is a huge number of Neighbourhoods in cluster 2, there is little to no Italian Restaurant. We see that in the Downtown Toronto area (cluster 3) has the second last average of Italian Restaurants. Looking at the nearby venues, the optimum place to put a new Italian Restaurant is in Downtown Toronto as there are many Neighbourhoods in the area but little to no Yoga Studio therefore, eliminating any competition. The second-best Neighbourhoods that have a great opportunity would be in areas such as Adelaide and King, Fairview, etc. which is in Cluster 2. Having 70 neighbourhoods in the area with no Yoga Studio gives a good opportunity for opening a new restaurant. Some of the drawback of this analysis are – the clustering is completely based on data obtained from Foursquare API. Also, the analysis does not take into consideration of the Italian population across neighbourhoods as this can play a huge factor while choosing which place to open a new Italian restaurant. This concludes the optimal findings for this project and recommends the entrepreneur to open an authentic Italian restaurant in these locations with little to no competition.

## Conclusion

In conclusion, to end off this project, we had an opportunity on a business problem, and it was tackled in way that it was like how a genuine data scientist would do. We utilized numerous Python libraries to fetch the information , to control the content and to break down and visualize those datasets. We have utilized Foursquare API to investigate the settings in neighbourhoods of Toronto, get great measure of data from Wikipedia which we scraped with the beautiful soup 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 gives 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 Theatre and so forth. Ideally, this task acts as an initial direction to tackle more complex real-life problems using data-science.