The Battle of Neighborhoods

Capstone Project:

Coffee Shop in Toronto

———

1. Introduction

1.1. Background

Toronto is the provincial capital 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 Greater Toronto Area (GTA) had a 2016 population of 6,417,516. The city covers an area of 630.20 square kilometers (243.32 sq mi) and comprises six districts – East York, Etobicoke, North York, Old Toronto, Scarborough and York – which, were amalgamated to form Toronto's present boundaries in 1998. 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. Toronto is an international center of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world.

The diverse population of Toronto reflects its current and historical role as an important destination for immigrants to Canada. More than 50 percent of residents belong to a visible minority population group, and over 200 distinct ethnic origins are represented among its inhabitants. While the majority of Torontonians speak English as their primary language, over 160 languages are spoken in the city.

Toronto is a prominent center for music, theatre, motion picture production, and television production, and is home to the headquarters of Canada's major national broadcast networks and media outlets. Its varied cultural institutions, which include numerous museums and galleries, festivals and public events, entertainment districts, national historic sites, and sports activities, attract over 43 million tourists each year. Toronto is known for its many skyscrapers and high-rise buildings, in particular the tallest free-standing structure in the Western Hemisphere, the CN Tower.

The city is home to the Toronto Stock Exchange, the headquarters of Canada's five largest banks, and the headquarters of many large Canadian and multinational corporations. Its economy is highly diversified with strengths in technology, design, financial services, life sciences, education, arts, fashion, aerospace, environmental innovation, food services, and tourism.

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 Coffee Shop. This city is home to many coffee shops where coffee addicts find their favorite drink. Whether you're on your way to work looking for a quick cup or a tourist fueling up while visiting the best attractions and parks or you're simply looking for a spot to sit down and map out the best things to do during your local adventures, Toronto is one of the best options to drink a very good coffee.
But...
will we find the most suitable location for an entrepreneur to open a new Coffee Shop in Toronto?

1.3. Interest

This project is aimed towards entrepreneurs or business owners who want to open a Coffee Shop in Toronto.

The success criteria of the project will be a good recommendation of neighborhood choice to open a new Coffee shop based on the lack of this type of establishment in that location, near of neuralgic centers and with little to no competition.

2. Data acquisition and cleaning

2.1. Data sources

For the Toronto neighborhood data, a Wikipedia page exists that has all the information you need to explore and cluster the neighborhoods in Toronto. You will be required to scrape the Wikipedia page and wrangle the data, clean it, and then read it into a pandas dataframe so that it is in a structured format.

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

	Postal Code	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
5	M9A	Etobicoke	Islington Avenue, Humber Valley Village
6	M1B	Scarborough	Malvern, Rouge
7	M3B	North York	Don Mills

Figure 1: Postal Code List of Toronto

Now that you have built a dataframe of the postal code of each neighborhood along with the borough name and neighborhood name, in order to utilize the Foursquare location data, you need to get the latitude and the longitude coordinates of each neighborhood. You will use the Geocoder Python package.

URL: https://geocoder.readthedocs.io/index.html

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

Figure 2: Geographical coordinates of neighborhoods of Toronto

Toronto city geographical coordinates data will be utilized as input for the Foursquare API that will be leveraged to provision venues information for each neighborhood.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
2	Regent Park, Harbourfront	43.65426	-79.360636	Morning Glory Cafe	43.653947	-79.361149	Breakfast Spot
3	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center
4	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa

Figure 3: Example of Foursquare API
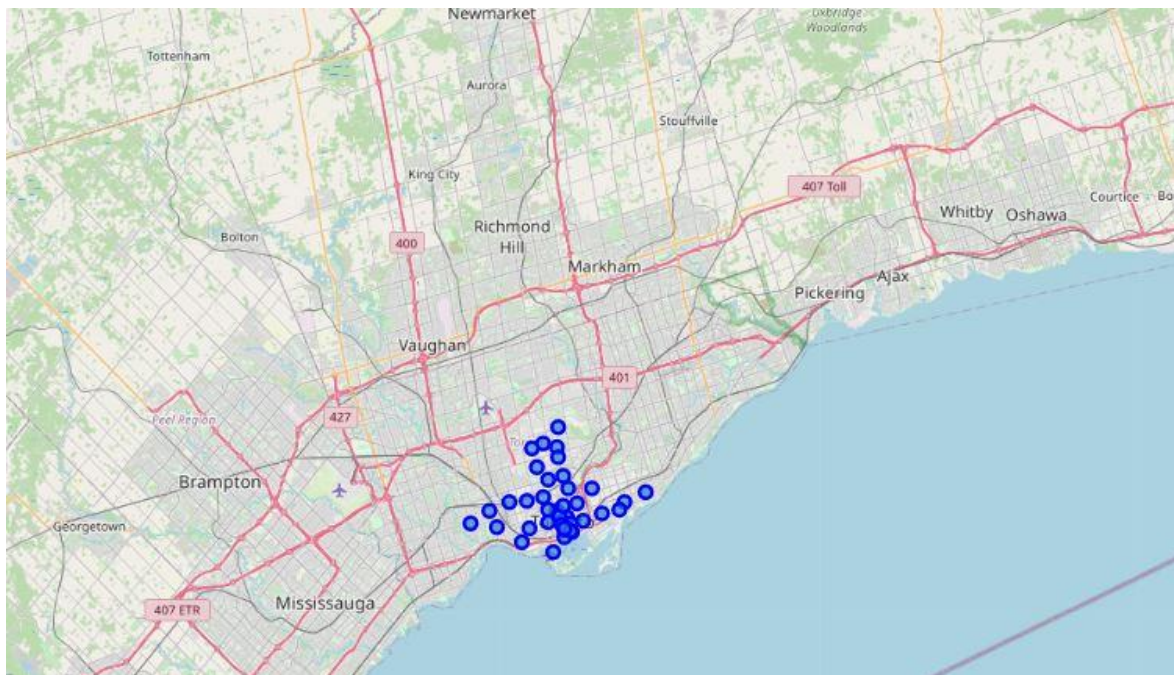
Figure 4: Example of visualization

4

2.2. Data Cleaning

Data downloaded from Wikipedia shows that more than one neighborhood can exist in one postal code area and there are some postal codes that are not associated to one neighborhood and it is marked as "Not Assigned".

Figure 5: List of Postal Code NOT ASSIGNED in Toronto

	Postal Code	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
5	M9A	Etobicoke	Islington Avenue, Humber Valley Village
6	M1B	Scarborough	Malvern, Rouge
7	M3B	North York	Don Mills

Figure 6: Postal Code Area assigned to multiple neighborhoods

Firstly, it is necessary to process the cells that have an assigned borough and combining the neighborhoods with same Postal Code.

	Postal Code	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
5	M9A	Etobicoke	Islington Avenue, Humber Valley Village
6	M1B	Scarborough	Malvern, Rouge
7	M3B	North York	Don Mills

Figure 7: Postal Code List Cleaned

Secondly, in order to utilize the Foursquare location data, it is needed to get the latitude and the longitude coordinates of each neighborhood.

	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
5	M1J	43.744734	-79.239476
6	M1K	43.727929	-79.262029
7	M1L	43.711112	-79.284577
8	M1M	43.716316	-79.239476
9	M1N	43.692657	-79.264848
10	M1P	43.757410	-79.273304

Figure 8: The geographical coordinates for each Postal Code

And then, it is merged both dataframe in one.

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
5	M9A	Etobicoke	Islington Avenue, Humber Valley Village	43.667856	-79.532242
6	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
7	M3B	North York	Don Mills	43.745906	-79.352188
8	M4B	East York	Parkview Hill, Woodbine Gardens	43.706397	-79.309937
9	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937
10	M6B	North York	Glencairn	43.709577	-79.445073

Figure 9: Geographical coordinates for each neighborhood

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.
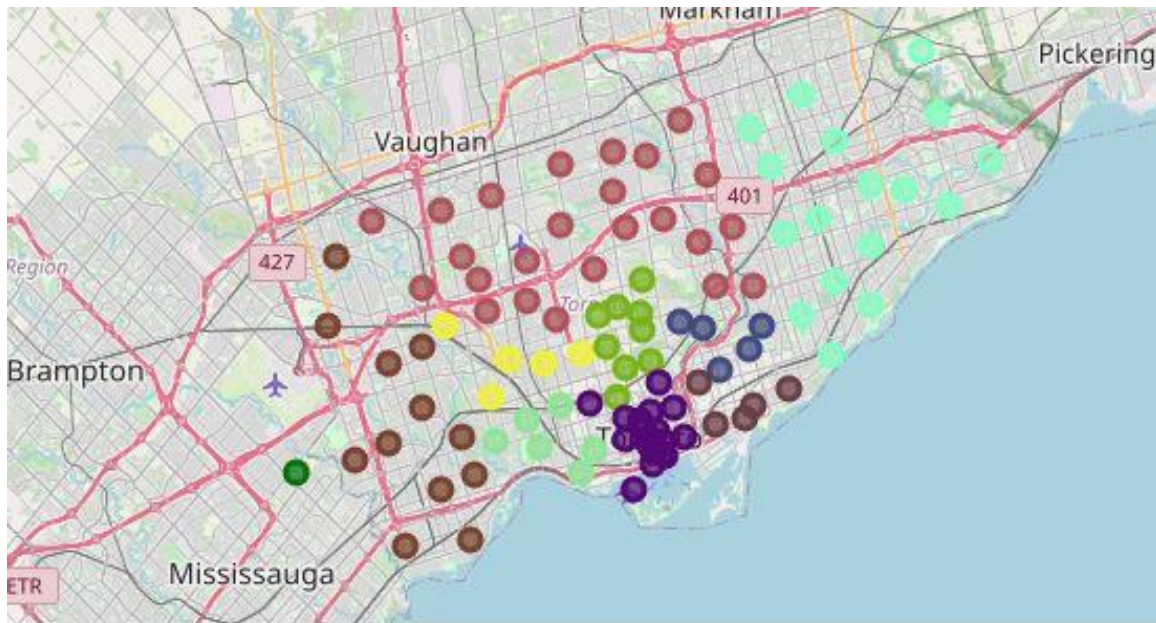
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
2	Regent Park, Harbourfront	43.65426	-79.360636	Morning Glory Cafe	43.653947	-79.361149	Breakfast Spot
3	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center
4	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa

Figure 10: The venue data from the Foursquare API

3. Methodology

3.1. Analytic Approach

When the data is merged with the Foursquare API information, the next step is to analyze the data. First, it is created a map using Folium color coded each Neighborhood depending on borough it is located.

Figure 11: Neighborhoods in Toronto

Next, we used the Foursquare API to get a list of all the Venues in Toronto which included Spa, Bus Line, Coffee Shops, Italian Restaurants etc. Getting this data is crucial to analyzing the number of Coffee Shops are all over Toronto. There is a total of 179 Coffee Shops in Toronto.

Then to analyze the data you perform 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 neighborhoods, individual venues were turned into the frequency at how many of those Venues were located in each neighborhood.

	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	...	Vegetarian / Vegan Restaurant	Video Game Store	Video Store	Vietnamese Restaurant	Warehouse Store	Wine Bar	Wine Shop	Wings Joint	Wome Sto
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	

Figure 12: One Hot Encoding

Then you group those rows by Neighborhood and by taking the Average of the frequency of occurrence of each Venue Category.

	Neighborhood	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	...	Vegetarian / Vegan Restaurant	Video Game Store	Video Store	Vietnamese Restaurant	Warehouse Store	Wine Bar	Wi Sh
0	Agincourt	0.0	0.0000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000000	...	0.000000	0.000000	0.000	0.000000	0.000000	0.000000	0.00
1	Alderwood, Long Branch	0.0	0.0000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000000	...	0.000000	0.000000	0.000	0.000000	0.000000	0.000000	0.00
2	Bathurst Manor, Wilson Heights, Downsview North	0.0	0.0000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000000	...	0.000000	0.000000	0.000	0.000000	0.000000	0.000000	0.00
3	Bayview Village	0.0	0.0000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000000	...	0.000000	0.000000	0.000	0.000000	0.000000	0.000000	0.00
4	Bedford Park, Lawrence Manor East	0.0	0.0000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.043478	...	0.000000	0.000000	0.000	0.000000	0.000000	0.000000	0.00
5	Berczy Park	0.0	0.0000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000000	...	0.017241	0.000000	0.000	0.000000	0.000000	0.000000	0.00
6	Birch Cliff, Cliffside West	0.0	0.0000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000000	...	0.000000	0.000000	0.000	0.000000	0.000000	0.000000	0.00

Figure 13: Grouped Neighborhoods by the average of the frequency of each Venue

After, you create a new data frame which only stored the Neighborhood names as well as the mean frequency of Coffee Shops in that Neighborhood. This allowed the data to be summarized based on each individual Neighborhood and made the data much simpler to analyze.
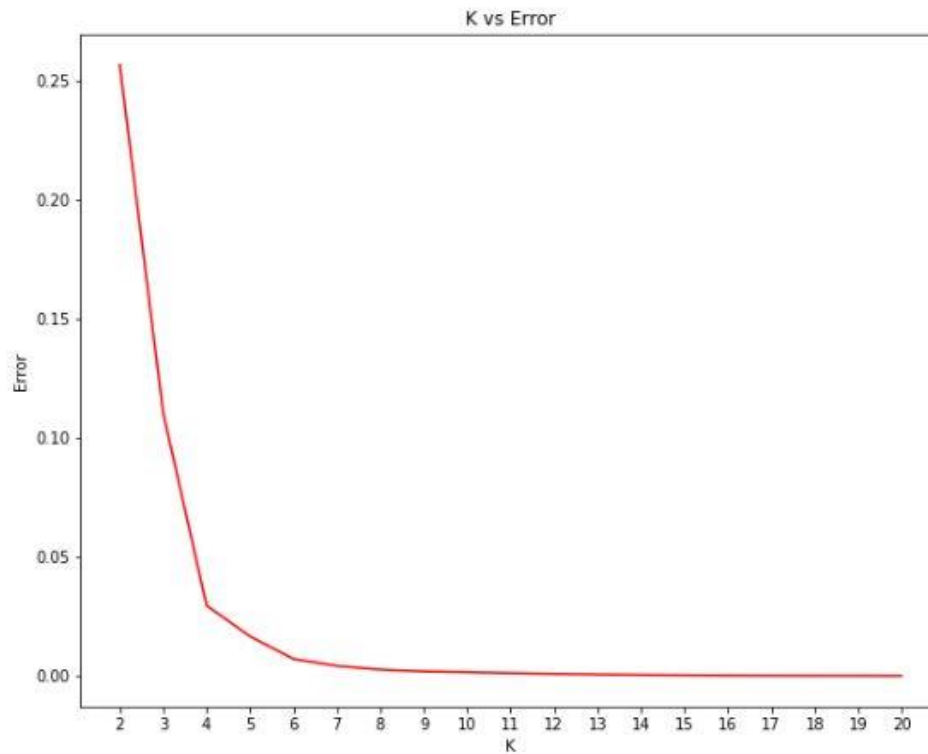
	Neighborhood	Coffee Shop
0	Agincourt	0.000000
1	Alderwood, Long Branch	0.100000
2	Bathurst Manor, Wilson Heights, Downsview North	0.095238
3	Bayview Village	0.000000
4	Bedford Park, Lawrence Manor East	0.086957

Figure 14: The average Coffee Shops in a neighborhood

To make the analysis more interesting, we wanted to cluster the neighborhoods based on the neighborhoods that had similar averages of Coffee Shops 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. 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 15: 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.

	Neighborhood	Coffee Shop	Cluster Labels
0	Agincourt	0.000000	0
1	Alderwood, Long Branch	0.100000	2
2	Bathurst Manor, Wilson Heights, Downsview North	0.095238	2
3	Bayview Village	0.000000	0
4	Bedford Park, Lawrence Manor East	0.086957	2

Figure 16: Appropriate Cluster Labels

4. Results

After, it is 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 Coffee Shop in Toronto. Then we created a map using the Folium package in Python and each neighborhood was colored based on the cluster label.
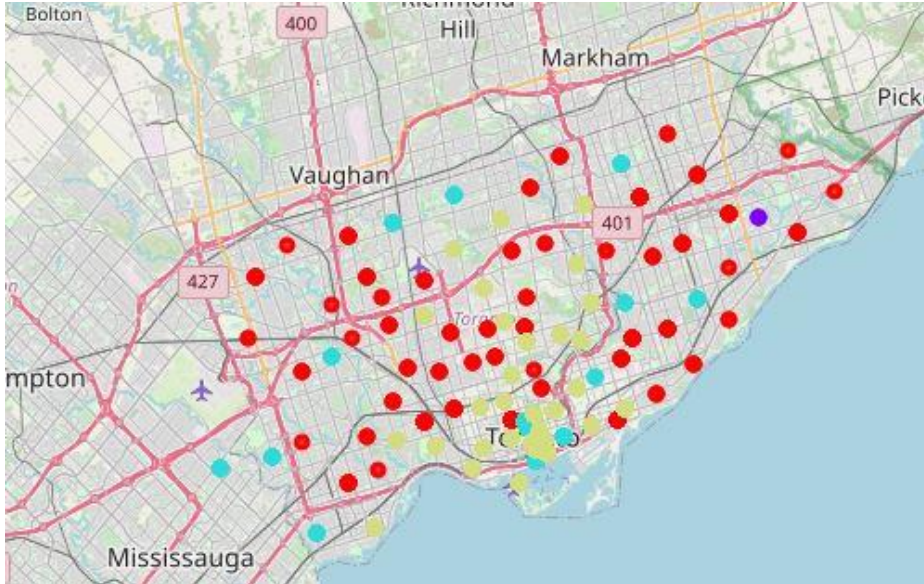

Figure 17: Map with different Clusters

We have a total of 4 clusters (0,1,2,3). Before it has analyzed them one by one lets check the total amount of neighborhoods in each cluster and the average Coffee Shops in that cluster. From the bar graph that was made using Matplotlib (figure 18), we can compare the number of Neighborhoods per Cluster. We see that Cluster 1 has the most neighborhoods (48) while cluster 2 has the least (1). Cluster 4 has 33 neighborhoods and cluster 3 has only 14.
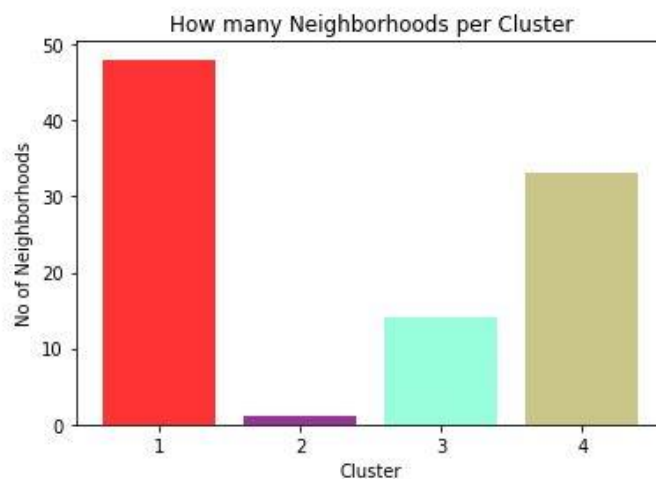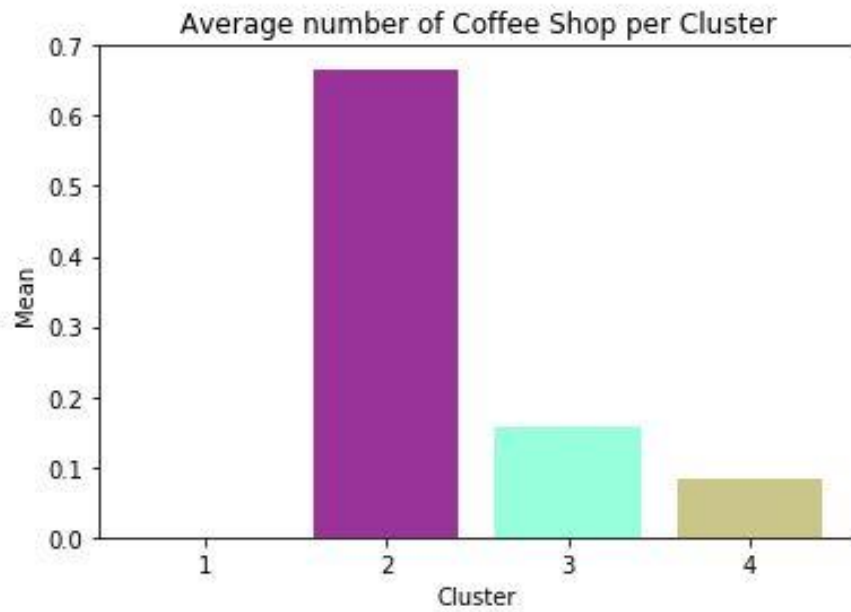

Figure 18:Number of Neighborhoods per cluster

Then we compared the average Coffee Shops per cluster.

Figure 19: Average of Coffee Shops in each neighborhood

This information is crucial as we can see that even through there is only 1 neighborhood in Cluster 2, it has the highest number of Coffee Shops (0.66) while Cluster 1 has the most neighborhoods but has the least average of Coffee Shops (0.0). The average of Coffee Shops made up the data for Figure 19.

Now let's analyze the Clusters individually.

Cluster 1

	Borough	Neighborhood	Coffee Shop	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	North York	Parkwoods	0.0	0	43.753259	-79.329656	Bella Vita Catering & Private Chef Service	43.756651	-79.331524	BBQ Joint
1	North York	Parkwoods	0.0	0	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	North York	Parkwoods	0.0	0	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
3	Scarborough	Malvern, Rouge	0.0	0	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
4	East York	Parkview Hill, Woodbine Gardens	0.0	0	43.706397	-79.309937	Jawny Bakers	43.705783	-79.312913	Gastropub
5	East York	Parkview Hill, Woodbine Gardens	0.0	0	43.706397	-79.309937	East York Gymnastics	43.710654	-79.309279	Gym / Fitness Center
6	East York	Parkview Hill, Woodbine Gardens	0.0	0	43.706397	-79.309937	TD Canada Trust	43.705740	-79.312270	Bank
7	East York	Parkview Hill, Woodbine Gardens	0.0	0	43.706397	-79.309937	Pizza Pizza	43.705159	-79.313130	Pizza Place
8	East York	Parkview Hill, Woodbine Gardens	0.0	0	43.706397	-79.309937	Shoppers Drug Mart	43.705933	-79.312825	Pharmacy

Cluster 1 was in the North York area, Scarborugh, East York… up to 48 neighborhoods. Cluster 1 had 126 unique Venue locations but there was not Coffee Shops. Cluster 1 had the least average of Coffee Shop to 0.0.

Cluster 1 is the most populate cluster, with 48 neighborhoods but there are not Coffee Shops in it.

Cluster 2

	Borough	Neighborhood	Coffee Shop	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Scarborough	Woburn	0.666667	1	43.770992	-79.216917	Starbucks	43.770037	-79.221156	Coffee Shop
1	Scarborough	Woburn	0.666667	1	43.770992	-79.216917	Korean Grill House	43.770812	-79.214502	Korean Restaurant
2	Scarborough	Woburn	0.666667	1	43.770992	-79.216917	Tim Hortons	43.770827	-79.223078	Coffee Shop

Cluster 2 was in the Scarborugh. Cluster 2 had 2 unique Venue locations and there was 2 Coffee Shops. Cluster 2 had the least average of Coffee Shop to 0.66 because there are enough Coffee Shops in the cluster.

Cluster 2 is the least populate cluster and there are sufficient Coffee Shops in it.

Cluster 3

	Borough	Neighborhood	Coffee Shop	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	North York	Victoria Village	0.200000	2	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
1	North York	Victoria Village	0.200000	2	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant
2	North York	Victoria Village	0.200000	2	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop
3	North York	Victoria Village	0.200000	2	43.725882	-79.315572	Eglinton Ave E & Sloane Ave/Bermondsey Rd	43.726086	-79.313620	Intersection
4	North York	Victoria Village	0.200000	2	43.725882	-79.315572	The Frig	43.727051	-79.317418	French Restaurant
5	Downtown Toronto	Regent Park, Harbourfront	0.159091	2	43.654260	-79.360636	GW General	43.650495	-79.357538	Antique Shop
6	Downtown Toronto	Regent Park, Harbourfront	0.159091	2	43.654260	-79.360636	The Healthy Road	43.656265	-79.357119	Health Food Store

Cluster 3 was in North York, Downtown Toronto… up to 14 neighborhoods. Cluster 3 had 116 unique Venue locations and there was 50 Coffee Shops. Cluster 3 had the least average of Coffee Shop to 0.2. Cluster 3 is not one of the most populate cluster and it has a good result of Coffee Shops in it.

Cluster 4

	Borough	Neighborhood	Coffee Shop	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	North York	Lawrence Manor, Lawrence Heights	0.083333	3	43.718518	-79.464763	Tim Hortons	43.719427	-79.467995	Coffee Shop
1	North York	Lawrence Manor, Lawrence Heights	0.083333	3	43.718518	-79.464763	Orfus Road Shopping Outlets	43.719045	-79.460849	Clothing Store
2	North York	Lawrence Manor, Lawrence Heights	0.083333	3	43.718518	-79.464763	Lac Vien Vietnamese Restaurant	43.721259	-79.468472	Vietnamese Restaurant
3	North York	Lawrence Manor, Lawrence Heights	0.083333	3	43.718518	-79.464763	Kitchen Stuff Plus (Clearance Outlet)	43.719096	-79.462675	Furniture / Home Store
4	North York	Lawrence Manor, Lawrence Heights	0.083333	3	43.718518	-79.464763	Roots	43.718214	-79.463893	Boutique
5	North York	Lawrence Manor, Lawrence Heights	0.083333	3	43.718518	-79.464763	Ardene Shoes Outlet	43.718892	-79.461344	Accessories Store

Cluster 4 was in Garden District, St. James Town… up to 33 neighborhoods. Cluster 4 had 228 unique Venue locations and there were 129 Coffee Shops. Cluster 4 had one of the least averages of Coffee Shop to 0.0833.

Cluster 4 is one of the most populate cluster and there are 129 Coffee Shops in it.

5. Discussion

Most of the Coffee Shops are in cluster 4 represented by the turquoise clusters. Even though there is a huge number of Neighborhoods in cluster 1, there is little to no Coffee Shops. We see that in the cluster 3 has the second last average of Coffee Shops but it is not one of the most populate cluster.

Looking at the nearby venues, the optimum place to put a new Coffee Shop is in East York as there are many Neighborhoods in the area but little to no Coffee Shops therefore, eliminating any competition. The second-best Neighborhoods that have a great opportunity would be in areas such as Downtown Toronto which is in Cluster 4. Having 33 neighborhoods in the area with 129 Coffee Shops gives a good opportunity for opening a new one in the Church and Wellesley zone because the average rate in lower than the rest of the neighborhoods of the cluster.

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 population across neighborhoods as this can play a huge factor while choosing which place to open a new Coffee Shop.

This concludes the optimal findings for this project and recommends the entrepreneur to open an authentic Canadian Coffee Shop in these locations with little to no competition.

6. 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 similar to 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 neighborhoods of Toronto, get 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 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 Theater and so forth. Ideally, this task acts as an initial direction to tackle more complex real-life problems using data-science.