Capstone Project - The Battle of Neighborhoods

Opening Chinese Restaurant in Toronto

1. Introduction

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 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.

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. There are 631,050 Chinese in the Greater Toronto Area as of the 2016 census, second only to New York City for largest Chinese community in North America. So there are numerous opportunities to open a new Chinese restaurant. Through this project, we will find the most suitable location for an entrepreneur to open a new Chinese restaurant in Toronto, Canada.

2. Target Audience

Entrepreneur or business owner who wants to open Chinese restaurant in Toronto but is uncertain about which neighborhood.

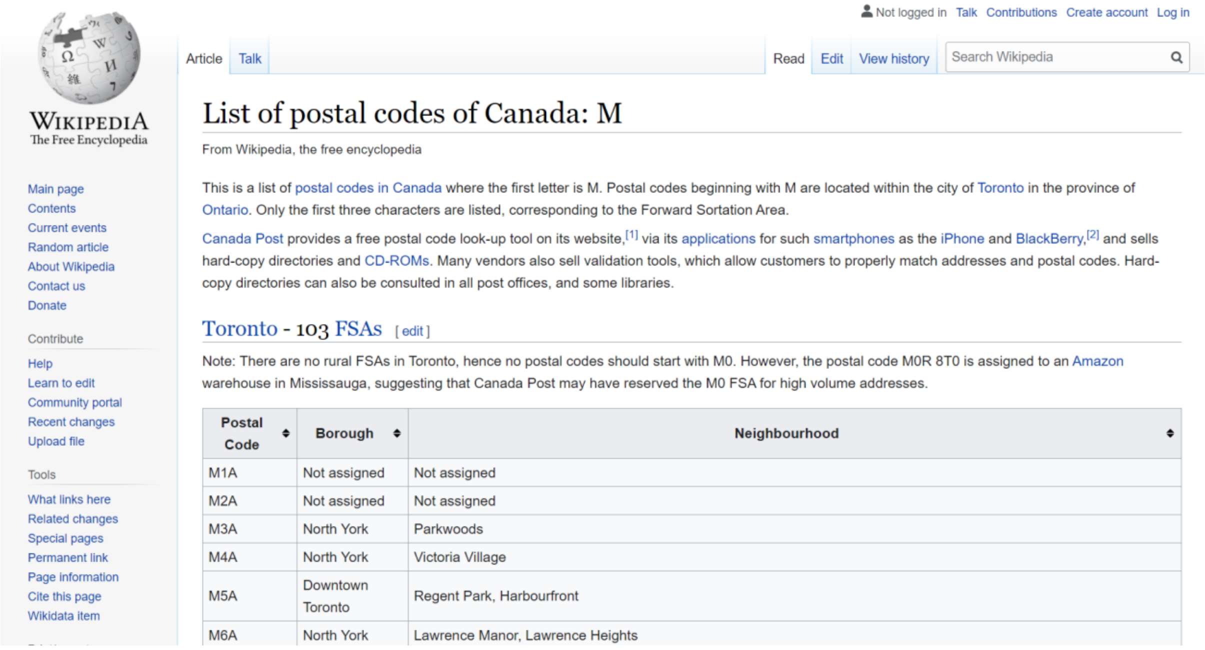
3. Data Overview

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

4. Methodology

We will need to extract the data from the data sources:

Source 1: Toronto Neighborhoods via Wikipedia



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.

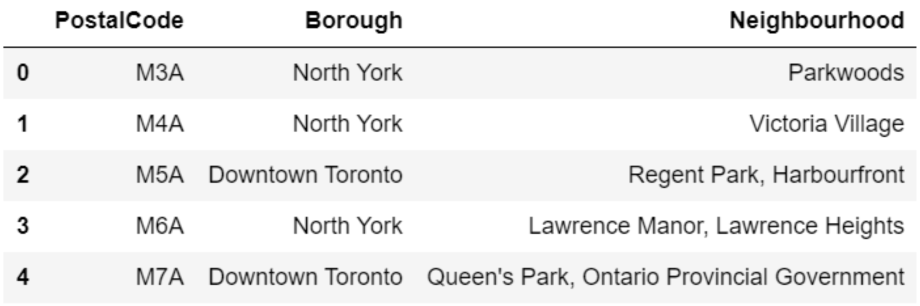
Toronto - 103 FSAs [edit]

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 Neighborhoods 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 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 Figure 2).

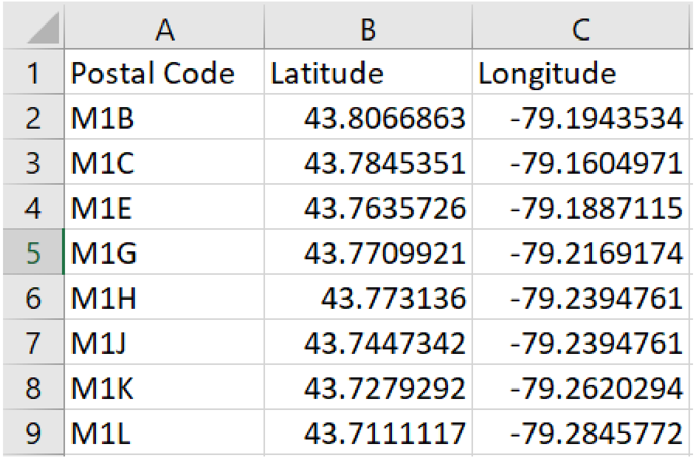


	PostalCode	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

Figure 2: *Data that was scraped from Wikipedia site and put into Pandas data frame*

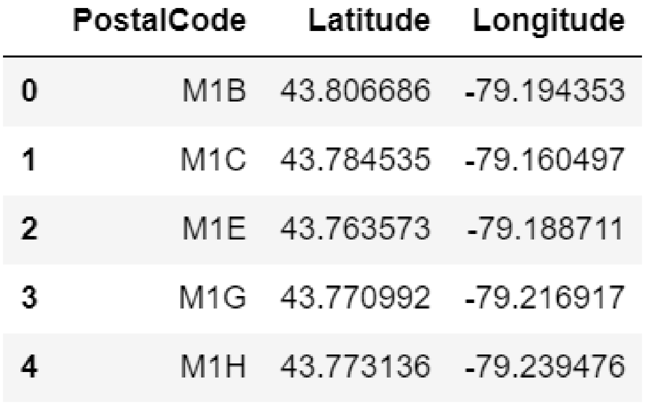
Source 2: Geographical Location data using Geocoder Package

The second source of data provided (https://cocl.us/Geospatial_data) us with the Geographical coordinates of the neighborhoods with the respective Postal Codes (Figure 3). The file was in CSV format, so attaching it to a Pandas data frame was simple (shown in Figure 4).



	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
8	M1K	43.7279292	-79.2620294
9	M1L	43.7111117	-79.2845772

Figure 3: *Geographical data of Neighborhoods in Toronto*

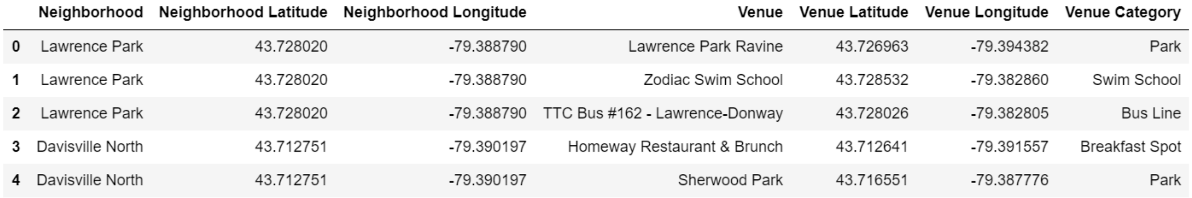


	PostalCode	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Figure 4: *Conversion of file into Pandas data frame*

Source 3: Venue Data using Foursquare

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 me to pull any data.



	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

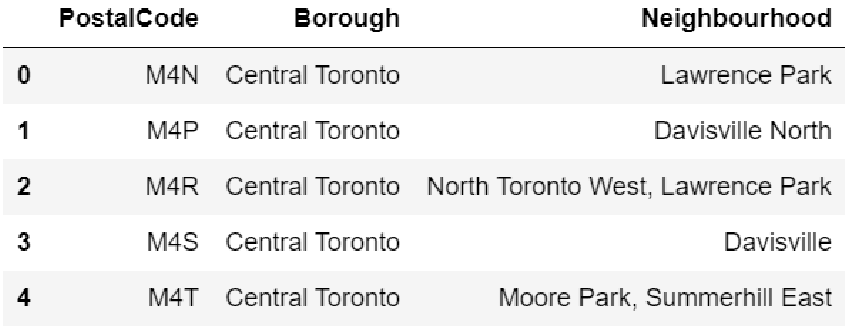
Figure 5: *Venue data pulled from Foursquare explore API*

It is seen through figure 5 (above) that the neighborhoods are grouped by the neighborhood, so data clustering is made easier later on.

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 that is not assigned are 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 neighborhoods: 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.

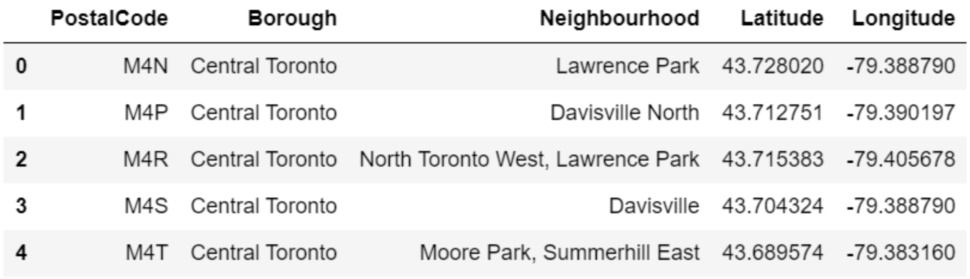
After the implementation of the following assumptions, the rows were grouped based on borough as shown below.



	PostalCode	Borough	Neighbourhood
0	M4N	Central Toronto	Lawrence Park
1	M4P	Central Toronto	Davisville North
2	M4R	Central Toronto	North Toronto West, Lawrence Park
3	M4S	Central Toronto	Davisville
4	M4T	Central Toronto	Moore Park, Summerhill East

Figure 6: *Rows grouped together based on Borough*

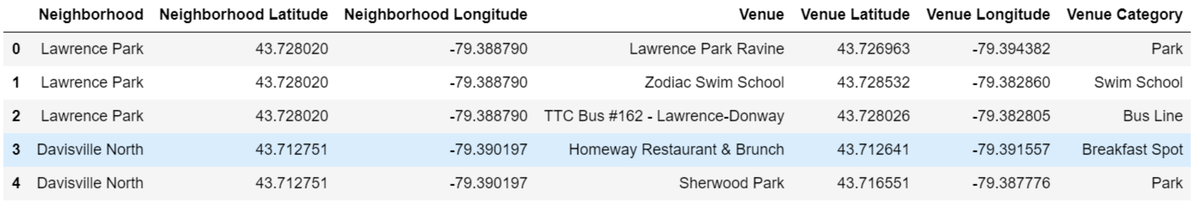
Using the Latitude and Longitude collected from the Geocoder package, we merged the two tables together based on Postal Code.



	PostalCode	Borough	Neighbourhood	Latitude	Longitude
0	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790
1	M4P	Central Toronto	Davisville North	43.712751	-79.390197
2	M4R	Central Toronto	North Toronto West, Lawrence Park	43.715383	-79.405678
3	M4S	Central Toronto	Davisville	43.704324	-79.388790
4	M4T	Central Toronto	Moore Park, Summerhill East	43.689574	-79.383160

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	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 8: *Local Venues near the respective Neighborhood*

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.

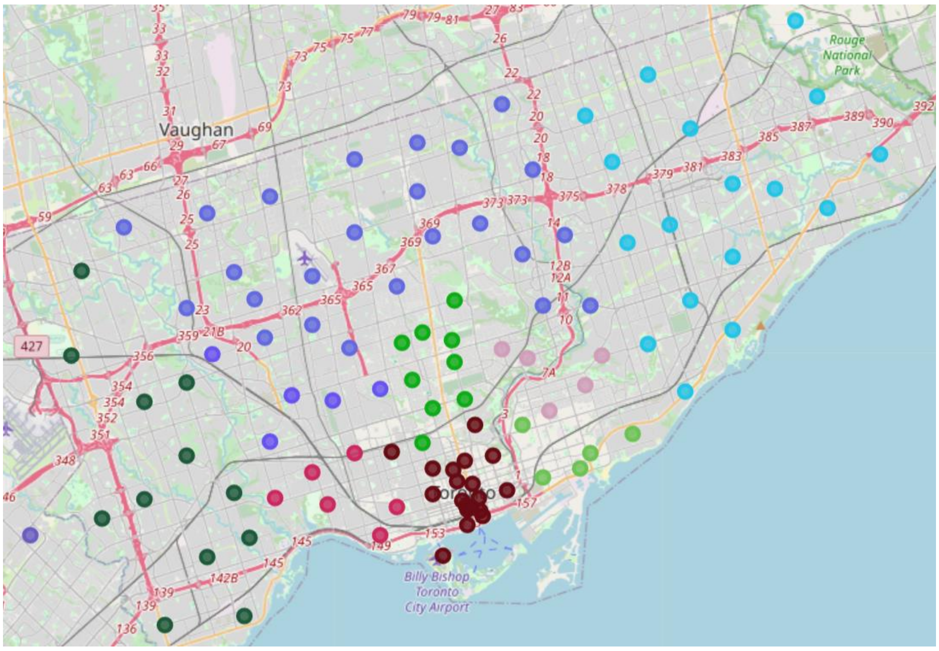
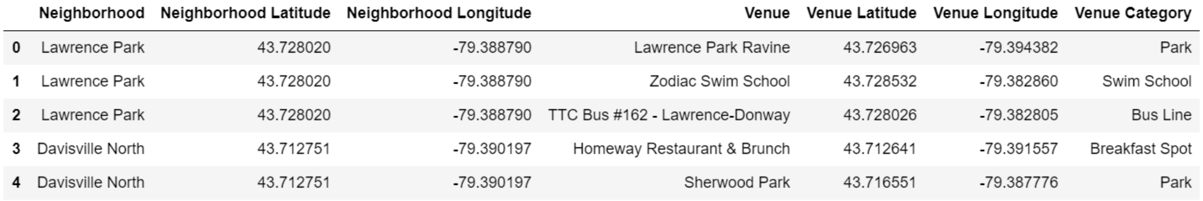


Figure 9: *Toronto Neighborhoods*

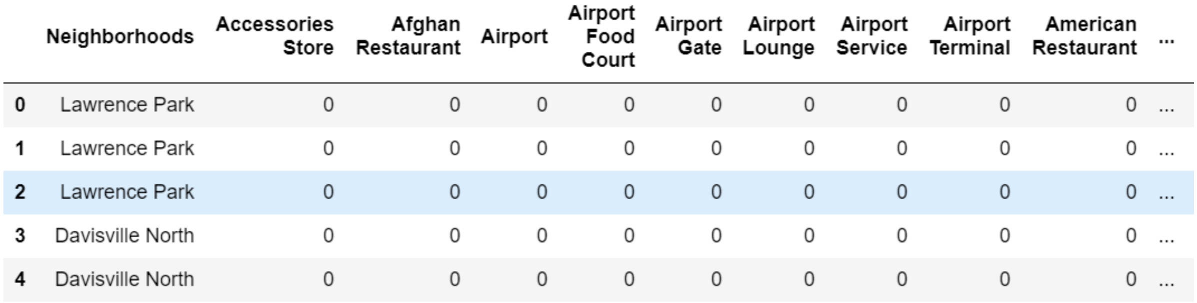
Next, we used the Foursquare API to get a list of all the Venues in Toronto which included Parks, Schools, Café Shops, Asian Restaurants etc. Getting this data was crucial to analyzing the number of Chinese Restaurants all over Toronto. There was a total of 16 Chinese Restaurants in Toronto. We then 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	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 Neighborhood data*

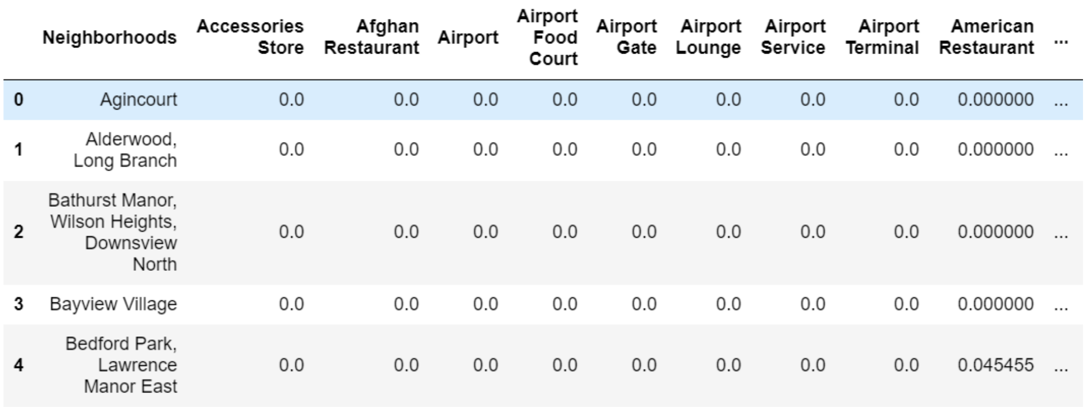
Then to analyze the data we performed a technique in which Categorical Data is transformed into Numerical Data for Machine Learning algorithms. This technique is called One hot encoding. For each of the neighborhoods, individual venues were turned into the frequency at how many of those Venues were located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

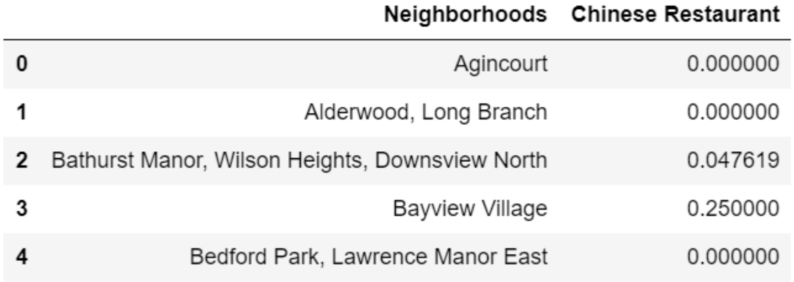
Then we grouped those rows by Neighborhood and by taking the Average of the frequency of occurrence of each Venue Category.



	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	0.045455	...

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

After, we created a new data frame which only stored the Neighborhood names as well as the mean frequency of Chinese 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	Chinese Restaurant
0	Agincourt	0.000000
1	Alderwood, Long Branch	0.000000
2	Bathurst Manor, Wilson Heights, Downsview North	0.047619
3	Bayview Village	0.250000
4	Bedford Park, Lawrence Manor East	0.000000

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

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

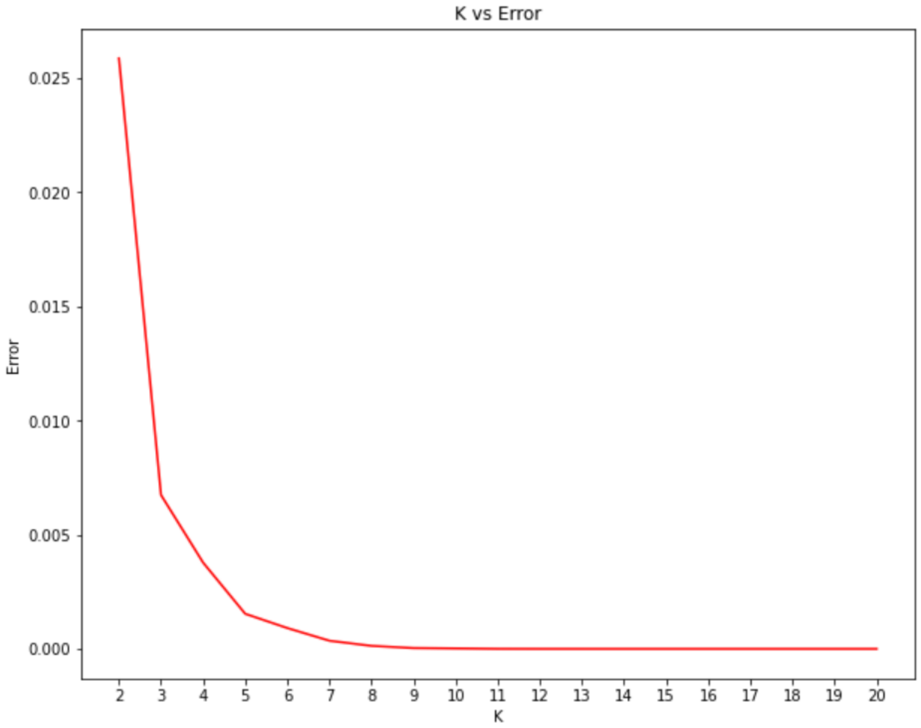
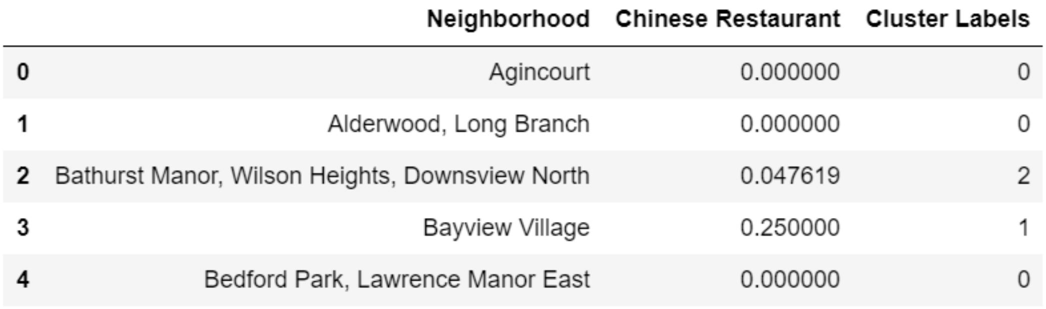


Figure 14: *Finding the K vs Error Values*

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. Neighborhoods that had similar mean frequency of Chinese Restaurants 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.



	Neighborhood	Chinese Restaurant	Cluster Labels
0	Agincourt	0.000000	0
1	Alderwood, Long Branch	0.000000	0
2	Bathurst Manor, Wilson Heights, Downsview North	0.047619	2
3	Bayview Village	0.250000	1
4	Bedford Park, Lawrence Manor East	0.000000	0

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 Chinese Restaurant in Toronto. Then we created a map using the Folium package in Python and each neighborhood was colored based on the cluster label. For example, cluster 2 was purple and cluster 3 was blue.

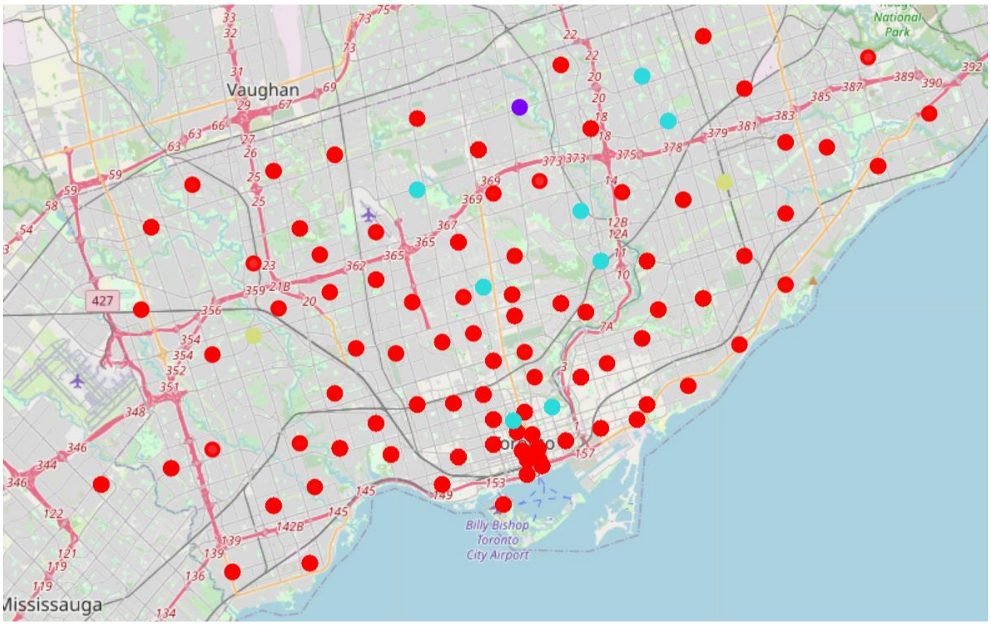


Figure 16: *Map with different Clusters*

The map above shows the different clusters that had similar mean frequency of Chinese restaurants.

5. Analysis

We have a total of 4 clusters (0,1,2,3). Before we analyze them one by one let's check the total amount of neighborhoods in each cluster and the average Chinese Restaurants in that cluster. From the bar graph that was made using Matplotlib (figure 17), we can compare the number of Neighborhoods per Cluster. We see that Cluster 2 has the least neighborhoods (1) while cluster 1 has the most (86). Cluster 3 has 7 neighborhoods and cluster 4 has only 2. Then we compared the average Chinese Restaurants per cluster.

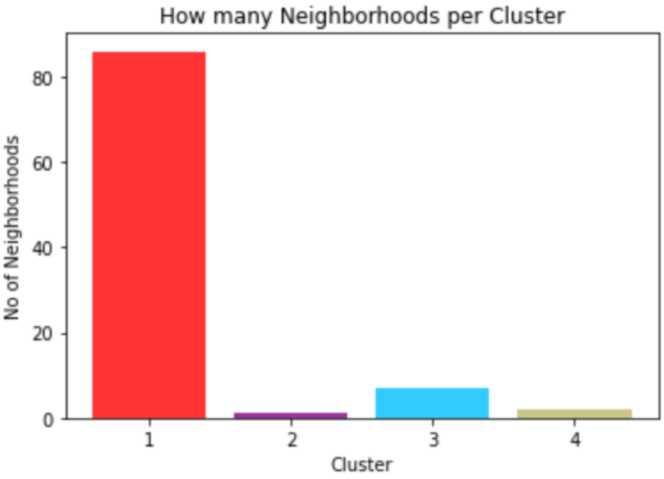


Figure 17: *Number of Neighborhoods per cluster*

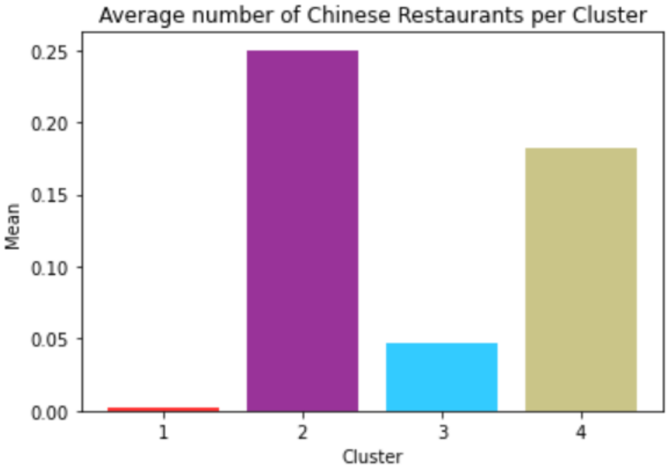
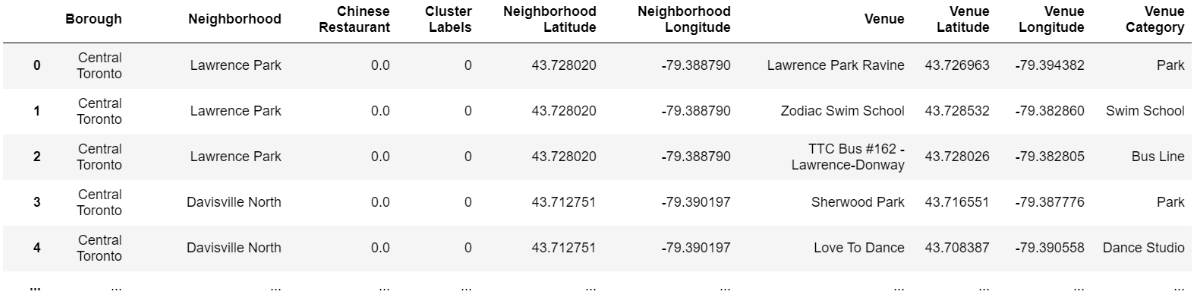


Figure 18: *Average Chinese restaurant in each neighborhood*

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 Chinese Restaurants (0.25) while Cluster 1 has the most neighborhoods but has the least average of Chinese Restaurants (0.0025). The average of the average Chinese Restaurant made up the data for Figure 18. Also, from the map, we can 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).

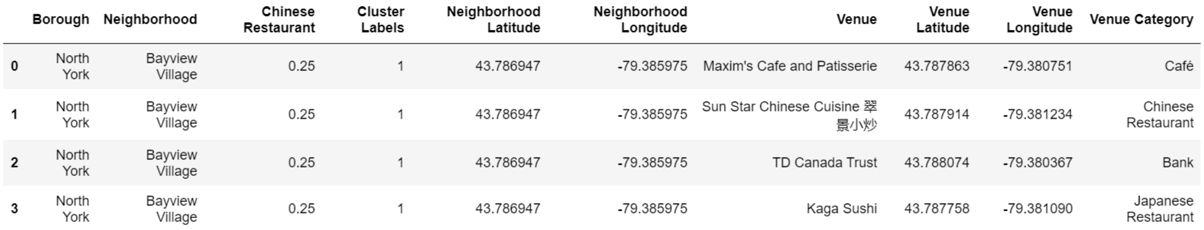
Cluster 1 (Red):



	Borough	Neighborhood	Chinese Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Central Toronto	Lawrence Park	0.0	0	43.728020	-79.388790	Lawrence Park Ravine	43.726963	-79.394382	Park
1	Central Toronto	Lawrence Park	0.0	0	43.728020	-79.388790	Zodiac Swim School	43.728532	-79.382860	Swim School
2	Central Toronto	Lawrence Park	0.0	0	43.728020	-79.388790	TTC Bus #162 - Lawrence-Donway	43.728026	-79.382805	Bus Line
3	Central Toronto	Davisville North	0.0	0	43.712751	-79.390197	Sherwood Park	43.716551	-79.387776	Park
4	Central Toronto	Davisville North	0.0	0	43.712751	-79.390197	Love To Dance	43.708387	-79.390558	Dance Studio
...	...	...	...	...	...	...	...	...	...	...

There was a total of 86 neighborhoods, 268 different venues and only 5 Chinese Restaurant. Therefore, the average amount of Chinese Restaurants that were near the venues in Cluster 1 is the lowest being 0.0025. In the map we can see that nodes of Cluster 1 were dispersed all throughout Toronto making it one of the most sparsely populated cluster.

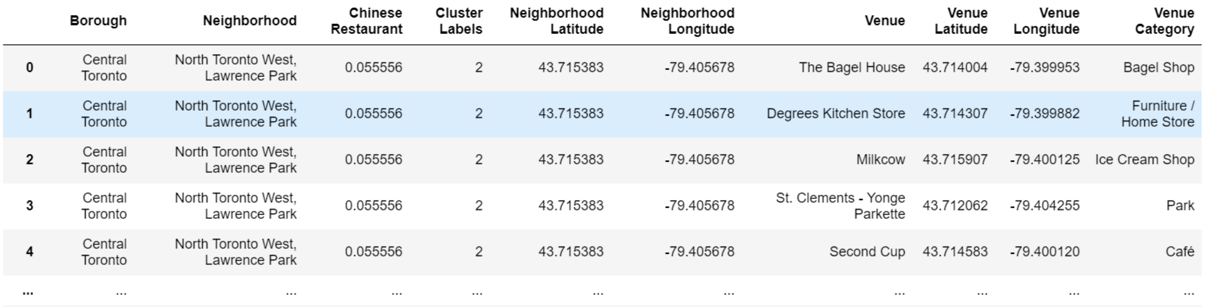
Cluster 2 (Purple):



	Borough	Neighborhood	Chinese Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	North York	Bayview Village	0.25	1	43.786947	-79.385975	Maxim's Cafe and Patisserie	43.787863	-79.380751	Café
1	North York	Bayview Village	0.25	1	43.786947	-79.385975	Sun Star Chinese Cuisine 翠景小炒	43.787914	-79.381234	Chinese Restaurant
2	North York	Bayview Village	0.25	1	43.786947	-79.385975	TD Canada Trust	43.788074	-79.380367	Bank
3	North York	Bayview Village	0.25	1	43.786947	-79.385975	Kaga Sushi	43.787758	-79.381090	Japanese Restaurant

Cluster 2 was in the North York area. Bayview village was the Neighborhood that was in that cluster. Cluster 2 had only 4 unique Venue locations and out of those only 1 was Chinese Restaurant. Cluster 2 had the highest average of Chinese Restaurants equating to 0.25. The reason why the average of Chinese Restaurants is the highest is because all these Restaurants are in this neighborhood, Bayview Village.

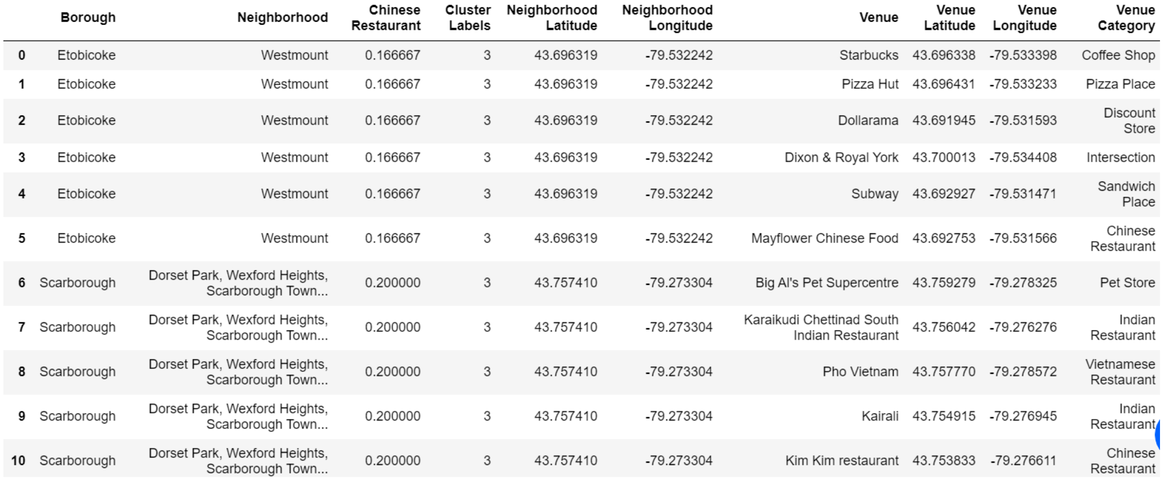
Cluster 3 (Blue):



	Borough	Neighborhood	Chinese Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Central Toronto	North Toronto West, Lawrence Park	0.055556	2	43.715383	-79.405678	The Bagel House	43.714004	-79.399953	Bagel Shop
1	Central Toronto	North Toronto West, Lawrence Park	0.055556	2	43.715383	-79.405678	Degrees Kitchen Store	43.714307	-79.399882	Furniture / Home Store
2	Central Toronto	North Toronto West, Lawrence Park	0.055556	2	43.715383	-79.405678	Milkcow	43.715907	-79.400125	Ice Cream Shop
3	Central Toronto	North Toronto West, Lawrence Park	0.055556	2	43.715383	-79.405678	St. Clements - Yonge Parkette	43.712062	-79.404255	Park
4	Central Toronto	North Toronto West, Lawrence Park	0.055556	2	43.715383	-79.405678	Second Cup	43.714583	-79.400120	Café
...	...	...	...	...	...	...	...	...	...	...

Cluster 3 had the second to lowest average of Chinese Restaurants. Cluster 3 was mainly located in the Downtown Toronto but also had some neighborhoods in North York, Scarborough, and in Central Toronto. Neighborhoods such as St. James Town, Cabbagetown, Don Mills and many more were included in this cluster. There was a total of 72 unique venues and out of those only 1 was Chinese Restaurants.

Cluster 4 (Dark Khaki):



	Borough	Neighborhood	Chinese Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Etobicoke	Westmount	0.166667	3	43.696319	-79.532242	Starbucks	43.696338	-79.533398	Coffee Shop
1	Etobicoke	Westmount	0.166667	3	43.696319	-79.532242	Pizza Hut	43.696431	-79.533233	Pizza Place
2	Etobicoke	Westmount	0.166667	3	43.696319	-79.532242	Dollarama	43.691945	-79.531593	Discount Store
3	Etobicoke	Westmount	0.166667	3	43.696319	-79.532242	Dixon & Royal York	43.700013	-79.534408	Intersection
4	Etobicoke	Westmount	0.166667	3	43.696319	-79.532242	Subway	43.692927	-79.531471	Sandwich Place
5	Etobicoke	Westmount	0.166667	3	43.696319	-79.532242	Mayflower Chinese Food	43.692753	-79.531566	Chinese Restaurant
6	Scarborough	Dorset Park, Wexford Heights, Scarborough Town...	0.200000	3	43.757410	-79.273304	Big Al's Pet Supercentre	43.759279	-79.278325	Pet Store
7	Scarborough	Dorset Park, Wexford Heights, Scarborough Town...	0.200000	3	43.757410	-79.273304	Karaikudi Chettinad South Indian Restaurant	43.756042	-79.276276	Indian Restaurant
8	Scarborough	Dorset Park, Wexford Heights, Scarborough Town...	0.200000	3	43.757410	-79.273304	Pho Vietnam	43.757770	-79.278572	Vietnamese Restaurant
9	Scarborough	Dorset Park, Wexford Heights, Scarborough Town...	0.200000	3	43.757410	-79.273304	Kairali	43.754915	-79.276945	Indian Restaurant
10	Scarborough	Dorset Park, Wexford Heights, Scarborough Town...	0.200000	3	43.757410	-79.273304	Kim Kim restaurant	43.753833	-79.276611	Chinese Restaurant

Cluster 4 venues were located in the Etobicoke and Scarborough. Neighborhoods such as Westmount, Dorset Park, Wexford Heights, and Scarborough Town Centre. There were a total of 9 unique Venues in Cluster 4 with 2 Chinese Restaurants. This made up the second highest average of Chinese Restaurants in that cluster which was approximately 0.182.

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

1. Cluster 2 (≈ 0.25)
2. Cluster 4 (≈ 0.182)
3. Cluster 3 (≈ 0.047)
4. Cluster 1 (≈ 0.0025)

6. Discussion

Most of the Chinese Restaurants are in cluster 2 represented by the purple clusters. The Neighborhood located in the North York area that has the highest average of Chinese Restaurants is Bayview Village. Even though there is a huge number of Neighborhoods in cluster 1, there is little to no Chinese Restaurant. We see that in the Downtown Toronto area (cluster 3) has the second least average of Chinese Restaurants. Looking at the nearby venues, the optimum place to put a new Chinese Restaurant is in Downtown Toronto as there are many Neighborhoods in the area but little to no Chinese Restaurants therefore, eliminating any competition. The second-best Neighborhoods that have a great opportunity would be in areas such as Lawrence Park, Davisville North, etc. which is in Cluster 1. Having 81 neighborhoods in the area with no Chinese Restaurants 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 Chinese population across neighborhoods as this can play a huge factor while choosing which place to open a new Chinese restaurant. This concludes the optimal findings for this project and recommends the entrepreneur to open an authentic Chinese restaurant in these locations with little to no competition.

7. 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.