The Battle of Neighborhoods

Finding the best neighborhood in Toronto for a start-up of coffee shop which is specialized in Kopi Luwak and this is one of the most expensive coffee in the world due to its complex process.

Figure 1: City of Toronto.

This report is part of the IBM Data Science capstone project which is designed to challenge the proficiency level of the writer in the understanding of data science and knowing how to translate data into a storytelling report through data visualization. The aim of this report is to showcase how exactly can we use the data to solve a business problem where we use Foursquare location data.

Contents

1. Business Problem

With a population of 6.2 million people (recorded as per June 20, 2021), Toronto, the capital of the province of Ontario, is a major Canadian city along Lake Ontario’s northwestern shore. Toronto is also one of the most multicultural cities in the world and living in Toronto is a wonderful multicultural experience for almost everyone. More than 140 languages and dialects are spoken in the city, and almost half the population in Toronto was born outside Canada. In addition, places such as Canada's Wonderland, Ripley's aquarium, Downtown Chinatown are some of the well-known tourist attentions.

The objective of this project to find the best neighbourhood in Toronto to open this Kopi Luwak coffee shop in a very strategized location where the competitive is kept to the minimum and has a high volume of neighbour clustering. Thus, the solution must have sustainable criteria such as volume of people, ease of accessing to public transportation and probability of becoming a coffee shop moat in Toronto.

2. Target Audience

This project will be relevant for business owners who wish to find a place to start up their business , coffee loves , coffee manufacturers and food delivery services.

3. Overview of the Data

For this project we need the following data:

1. Toronto City data that contains Borough, Neighborhoods along with there latitudes and longitudes

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

Description: This Wikipedia page contain all the information we need to explore and cluster the neighbourhoods in Toronto. We will be required to scrape the Wikipedia page and wrangle the data, clean it, and then read it into a pandas data frame so that it is in a structured format like the Toronto dataset.

2. Geographical Location data using Geocoder Package

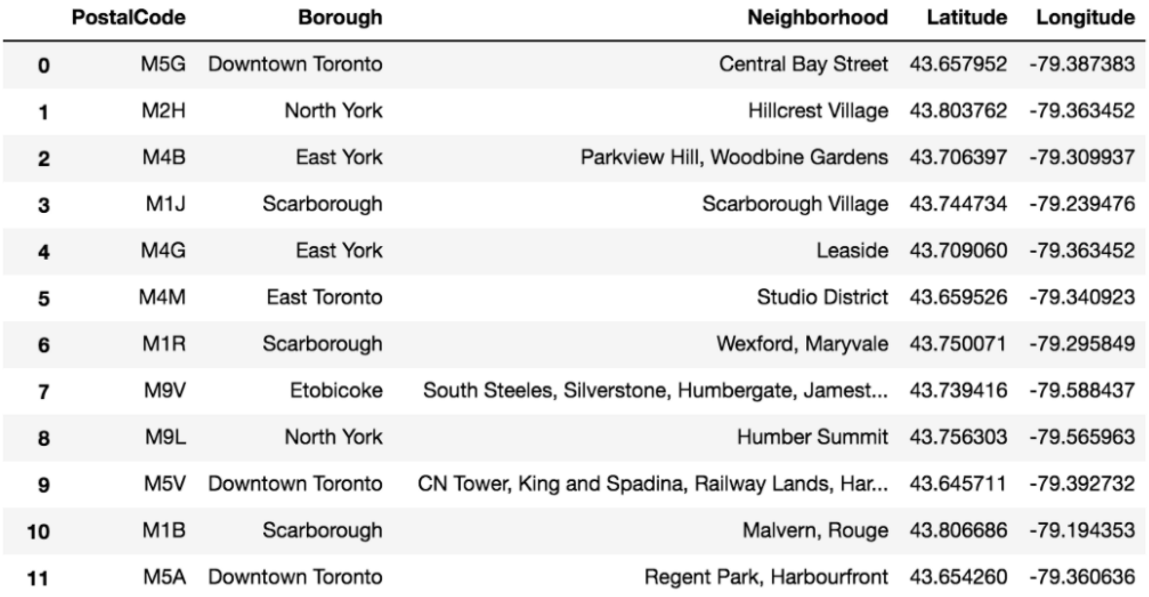
Data Source: https://cocl.us/Geospatial_data

Description: The second source of data provided us with the Geographical coordinates of the neighbourhoods with the respective Postal Codes.

3. Venue Data using Foursquare API

Data Source: https://foursquare.com/developers/apps

Description: From Foursquare API we can get the name, category ,latitude ,longitude for each venue.



	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383
1	M2H	North York	Hillcrest Village	43.803762	-79.363452
2	M4B	East York	Parkview Hill, Woodbine Gardens	43.706397	-79.309937
3	M1J	Scarborough	Scarborough Village	43.744734	-79.239476
4	M4G	East York	Leaside	43.709060	-79.363452
5	M4M	East Toronto	Studio District	43.659526	-79.340923
6	M1R	Scarborough	Wexford, Maryvale	43.750071	-79.295849
7	M9V	Etobicoke	South Steeles, Silverstone, Humbergate, Jamest...	43.739416	-79.588437
8	M9L	North York	Humber Summit	43.756303	-79.565963
9	M5V	Downtown Toronto	CN Tower, King and Spadina, Railway Lands, Har...	43.645711	-79.392732
10	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
11	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636

Figure 2: Neighbours in Toronto.

4. Methodology

In Wikipedia, most of the data is not as organized and structure as the one which we used for this project. Nevertheless, web scraping through the use Beautiful Soup is still necessary to make this project possible.

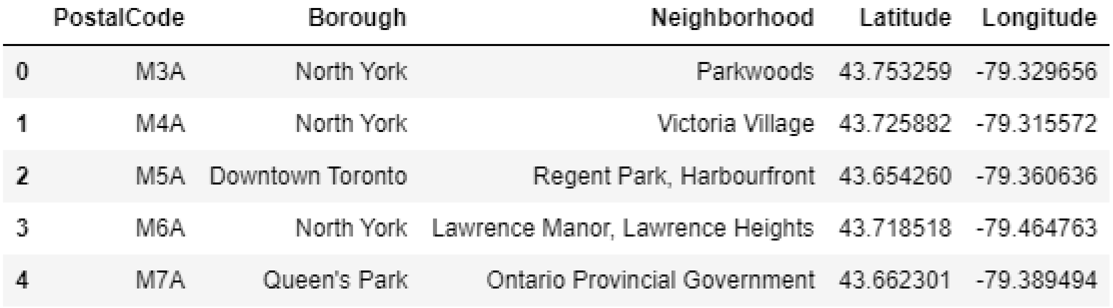
After scraping the data from Wikipedia there were Boroughs that were not assigned to any neighbourhood therefore, the following assumptions were made:

Only process the cells that have an assigned borough. Ignore cells with a borough that is Not assigned.

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 row 11 in the above table.

If a cell has a Boroughs but a Not assigned to a neighbourhood, then the neighbourhood must be the same as Boroughs.

We merged the two tables together based on Postal Code using the Latitude and Longitude collected from the Geocoder package.

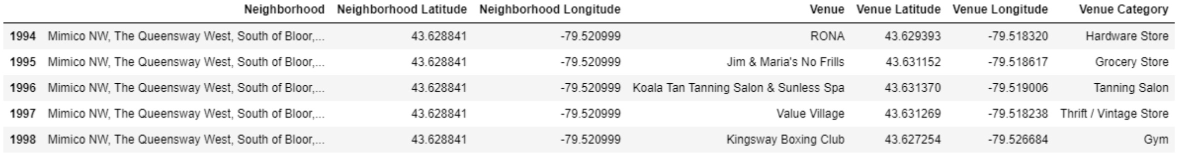


	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A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

Figure 3: A list of the first 5 Postal Code in Toronto.

4.1 Use of Foursquare API

Using the Foursquare API, we can retrieve the venue data which present the radius of 500 m of each neighbourhood and merge it with Figure 3.

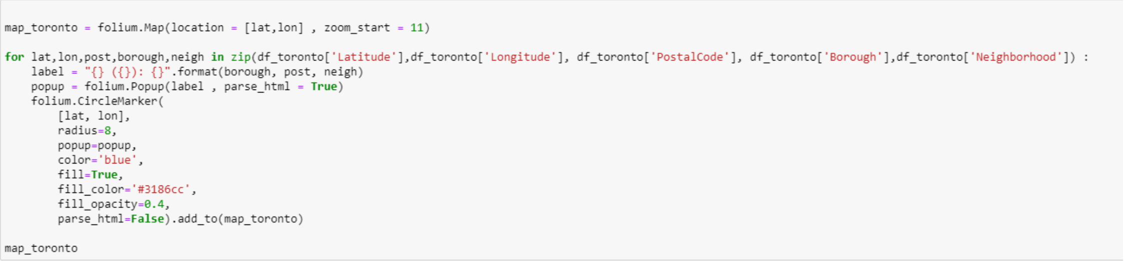


	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
1994	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	RONA	43.629393	-79.518320	Hardware Store
1995	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Jim & Maria's No Frills	43.631152	-79.518617	Grocery Store
1996	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Koala Tan Tanning Salon & Sunless Spa	43.631370	-79.519006	Tanning Salon
1997	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Value Village	43.631269	-79.518238	Thrift / Vintage Store
1998	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Kingsway Boxing Club	43.627254	-79.526684	Gym

Figure 4: Cluster all the neighbourhoods within a radius of 500 m together.

4.2 Using Folium for Data Visualization

To visualize the data in a map for easy reference, the use of Folium is required. Hence, applying the following codes to showcase all the coffee shops in Toronto.



```
map_toronto = folium.Map(location = [lat,lon] , zoom_start = 11)

for lat,lon,post,borough,neigh in zip(df_toronto['Latitude'],df_toronto['Longitude'], df_toronto['PostalCode'], df_toronto['Borough'],df_toronto['Neighborhood']) :
    label = "{} ({}): {}".format(borough, post, neigh)
    popup = folium.Popup(label , parse_html = True)
    folium.CircleMarker(
        [lat, lon],
        radius=8,
        popup=popup,
        color='blue',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.4,
        parse_html=False).add_to(map_toronto)

map_toronto
```

Figure 5: List of codes to generate all the coffee shops in Toronto.

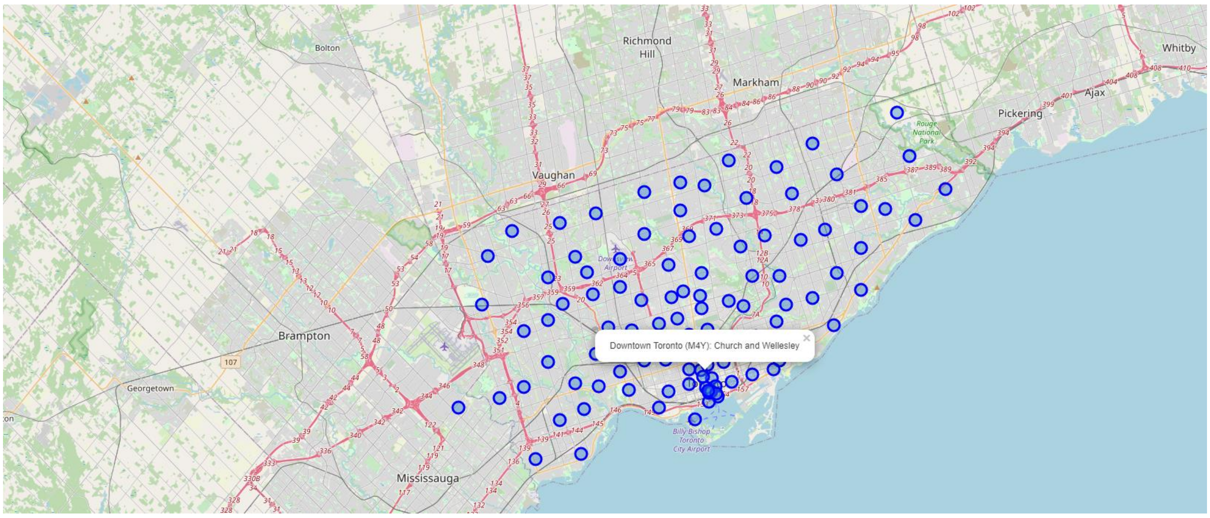


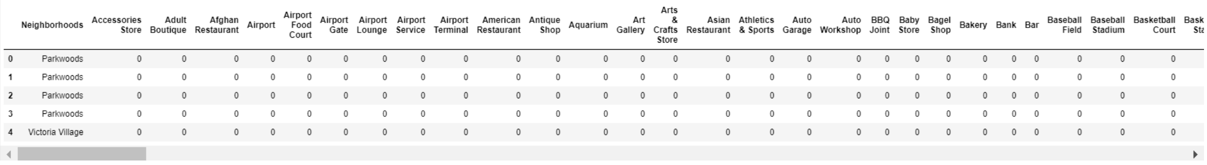
Figure 6: A snippet of the code which generates this map.

Foursquare API was used to extract a list of all the Venues in Toronto which included Parks, Schools, Café Shops, Asian Restaurants and etc.

As shown in Figure 6, the extraction of the data was important because the coffee shop owner to analyse the number of Coffee Shop in Toronto. There was a total of 201 Coffee Shop in Toronto(with blue dots as the indication). Next, we merged the Foursquare Venue data with the Neighbourhood data so that the nearest venue for each neighbourhood can be found.

4.3 Data Pre-processing

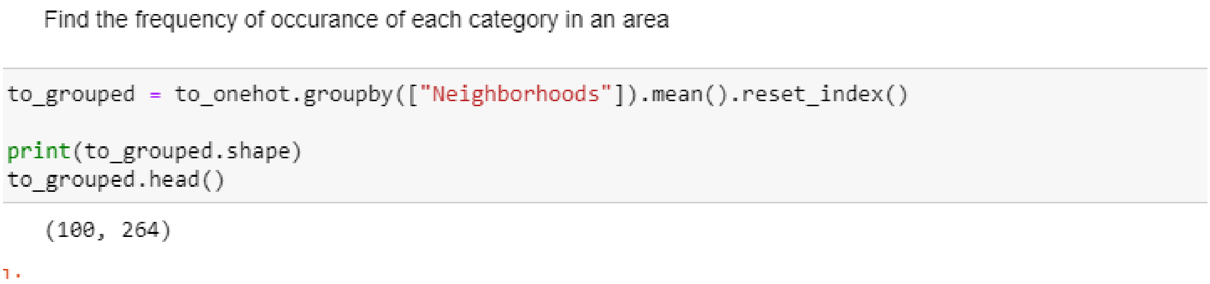
To better understand whether the set-up of a Coffee Shop in Toronto is feasible, the use of One Hot Encoding approach is required. The frequency for each neighbourhood and venues were shown.



	Neighborhoods	Accessories Store	Adult Boutique	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Garage	Auto Workshop	BBQ Joint	Baby Store	Bagel Shop	Bakery	Bank	Bar	Baseball Field	Baseball Stadium	Basketball Court	Bask Sta
0	Parkwoods	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	Parkwoods	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	Parkwoods	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	Parkwoods	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	Victoria Village	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Figure 7:One Hot Encoding Visualization.

Furthermore, we clustered all those rows for the neighbours through the mean of occurrence for each venue category.



Find the frequency of occurance of each category in an area

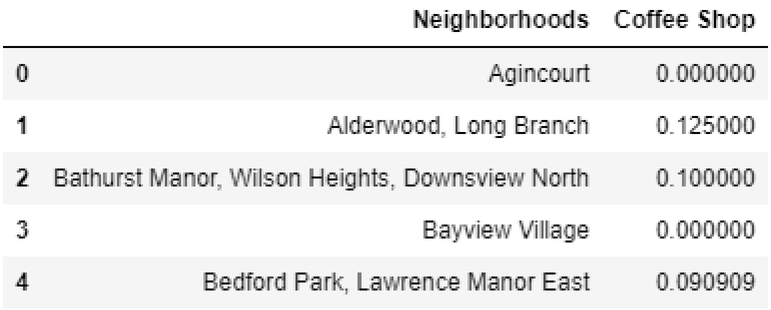
```
to_grouped = to_onehot.groupby(["Neighborhoods"]).mean().reset_index()

print(to_grouped.shape)
to_grouped.head()
```

```
(100, 264)
```

].

Figure 8: A list of coding for showing the mean of occurrence for each venue category.

With a clean data frame which only stored the neighbourhood names and the mean of occurrence of the Coffee Shop in the neighbourhood. The summarized data for each category in the neighbourhood can be visualized in a more structured and simplified way for analysing .



	Neighborhoods	Coffee Shop
0	Agincourt	0.000000
1	Alderwood, Long Branch	0.125000
2	Bathurst Manor, Wilson Heights, Downsview North	0.100000
3	Bayview Village	0.000000
4	Bedford Park, Lawrence Manor East	0.090909

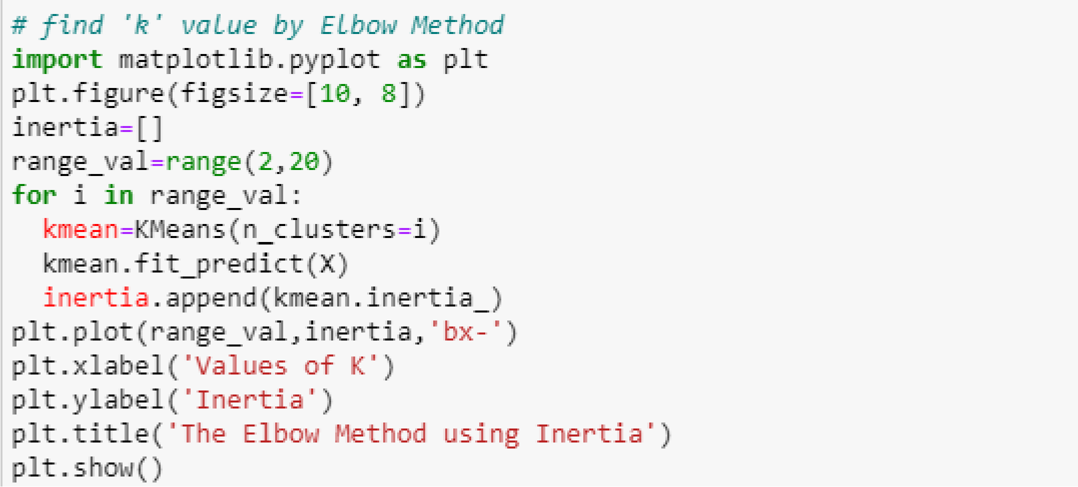
Figure 9: A clean data frame with only neighbourhoods and Coffee Shop as the column names.

4.4 K-Means Clustering

Based on the frequency of the Coffee Shop, we can cluster all these neighbourhoods together. Using the K-Means clustering Algorithm ,we can find the most optimized K value for the purpose of finding the right number of clustering. In addition, other alternatives such as Silhouette coefficient ,BIC score and Elbow methods can be used to find the most optimized K value. However, in this project, we can focus on using the Elbow method for the simplicity of this project.

4.5 Elbow method

To obtain the most optimized K value, we had to import “KElbowVisualizer” from the Yellowbrick Package. Hence, we can fit the K-Mens model to the Elbow visualizer.



```
# find 'k' value by Elbow Method
import matplotlib.pyplot as plt
plt.figure(figsize=[10, 8])
inertia=[]
range_val=range(2,20)
for i in range_val:
  kmean=KMeans(n_clusters=i)
  kmean.fit_predict(X)
  inertia.append(kmean.inertia_)
plt.plot(range_val,inertia,'bx-')
plt.xlabel('Values of K')
plt.ylabel('Inertia')
plt.title('The Elbow Method using Inertia')
plt.show()
```

Figure 10: The list of codes to generate Elbow visualization.

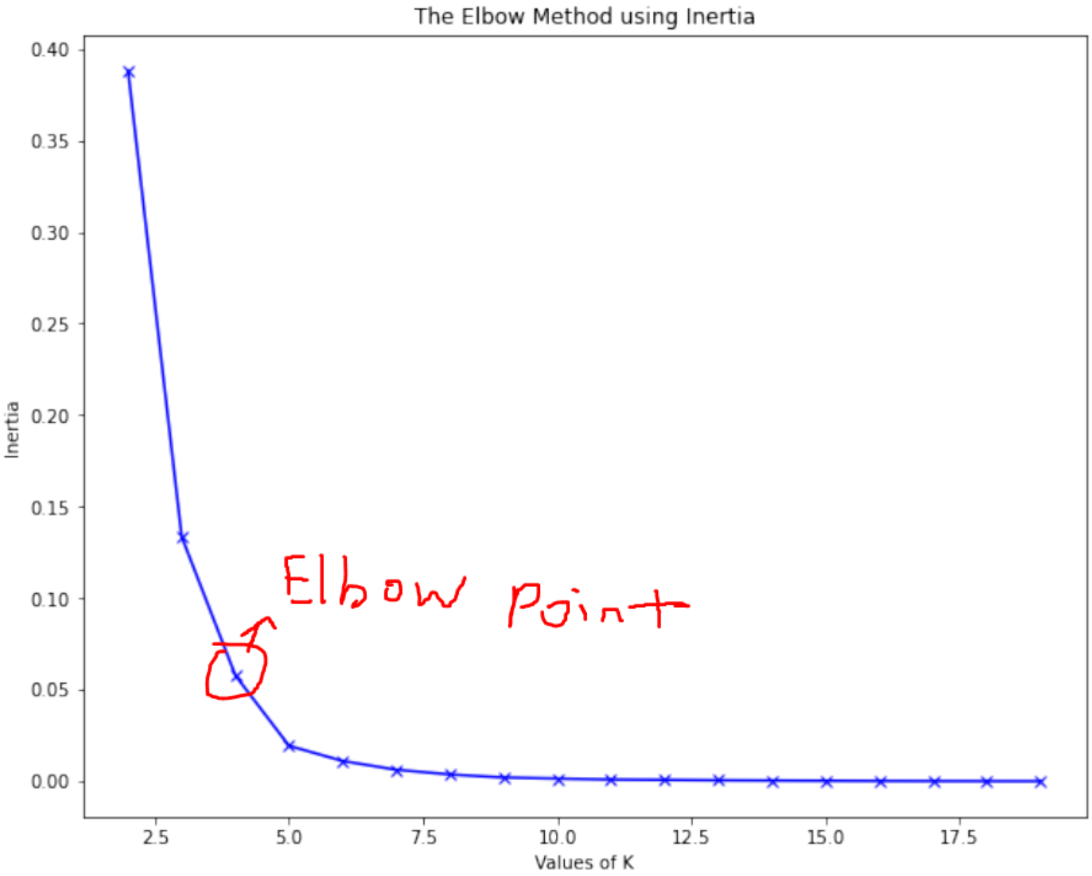
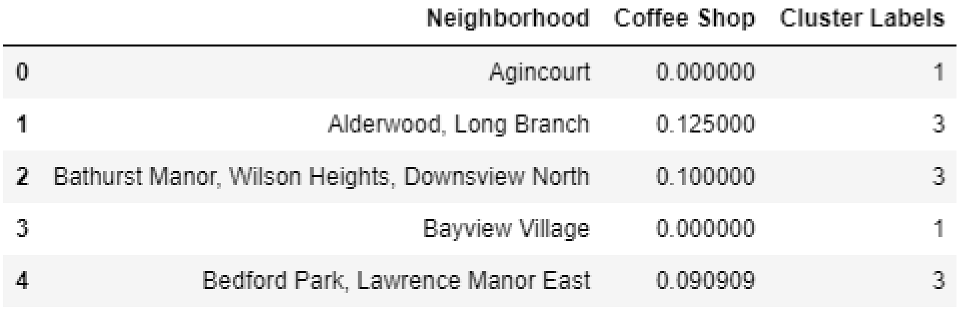


Figure 11: Elbow method visualization.

Based on Figure 12,the turning point of this plot showed that the best k value for this dataset is 4. In other words, there are four different cluster existed in this dataset. The index of these labels starts from 0 to 3; index[0] represents first cluster and index[3] represents fourth cluster.



	Neighborhood	Coffee Shop	Cluster Labels
0	Agincourt	0.000000	1
1	Alderwood, Long Branch	0.125000	3
2	Bathurst Manor, Wilson Heights, Downsview North	0.100000	3
3	Bayview Village	0.000000	1
4	Bedford Park, Lawrence Manor East	0.090909	3

Figure 12: A list of neighbourhoods with clustering labels.

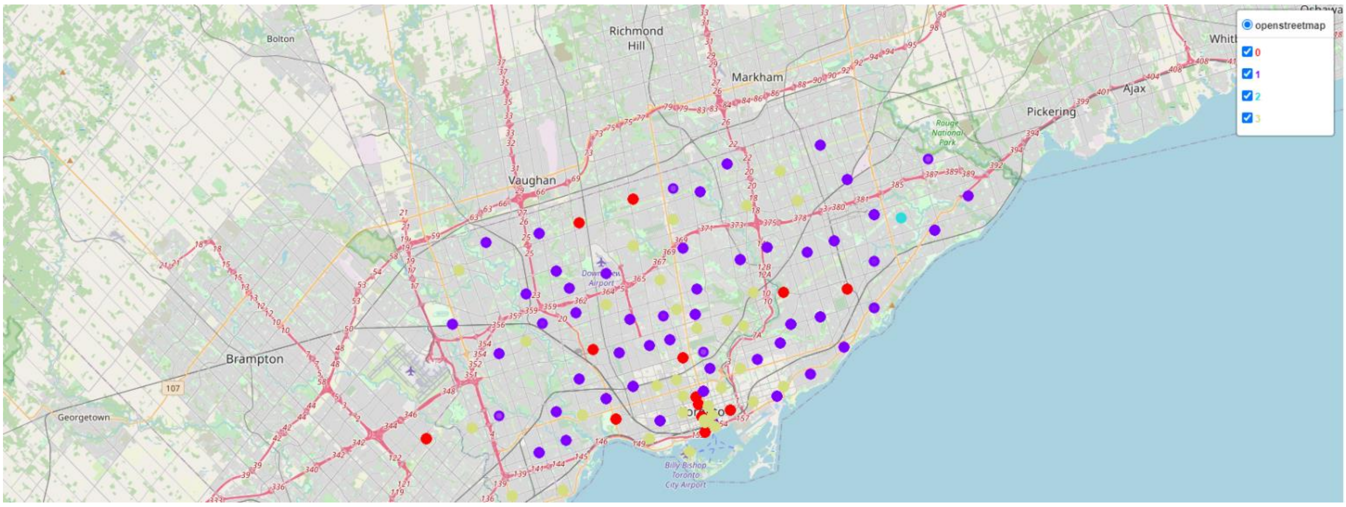


Figure 13: Using Folium API to show the clustering in the neighbourhoods.

5. Results

Based on , these two bar charts show that the coffee shop (Kopi Luwak) is best to set-up in cluster 2 (North York) due to the high number of neighbourhood and least number of existing coffeeshop in the area. However, if the coffee shop owner is confident that his coffee has the competitive edge over other competitors in term of delivering the top-notch coffee breed and exclusive to those who feel appreciate for the quality, he may choose to opt for cluster 3 (Scarborough).

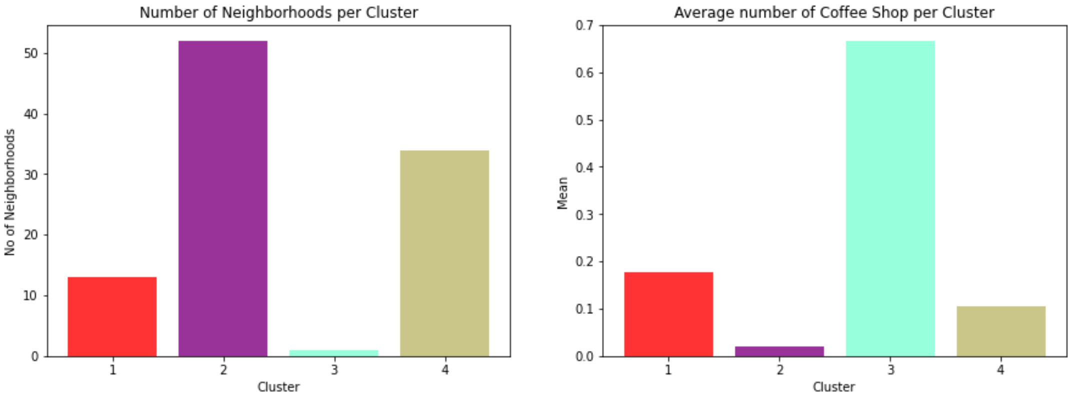


Figure 14: Bar charts between the number of neighbours and coffee shop per cluster.

6. Discussion

Using data visualization ,cleaning and extracting of data (information of Tornto) which we can showcase to our coffeeshop owner that we managed to find the highest volume of coffee shop in the Tornto and neighbourhoods through clustering. The purpose of clustering helps to determine a region where places which are nearby can be categorized together so that the data is more appealing and easier to understand as compared to several number of clusters with different names. With the presentation of this project, the coffee shop will gain a better understanding of where he can find the most strategized location to set up his store and ward off his competitors.

7. Future Work

These are some ideas which can improve the quality of this existing project:

1) Apply different techniques of clustering algorithms for the neighbourhoods.
2) Consider other features such as the food, market area and hotspot.
3) Include more venues in a neighbourhood for better analysis through the use of Foursquare API.