The Battle of Neighbourhoods

Rahul Raj

August 10, 2021

1. Introduction

1.1 Background

With a population just short of 3 million people, the city of Toronto is the largest in Canada, and one of the largest in North America (behind only Mexico City, New York and Los Angeles). Toronto is also one of the most multicultural cities in the world, making life in Toronto a wonderful multicultural experience for all. More than 140 languages and dialects are spoken in the city, and almost half the population Toronto were born outside Canada. 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

With a population just short of 3 million people, the city of Toronto is the largest in Canada, and one of the largest in North America (behind only Mexico City, New York and Los Angeles). Toronto is also one of the most multicultural cities in the world, making life in Toronto a wonderful multicultural experience for all. More than 140 languages and dialects are spoken in the city, and almost half the population Toronto were born outside Canada. 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 find the best neighbourhood in Toronto to open a restaurant using Foursquare location data. In this project we'll go through the solution for this problem for avoiding or considering low risk criteria and high success rate.

1.3 Target Audience

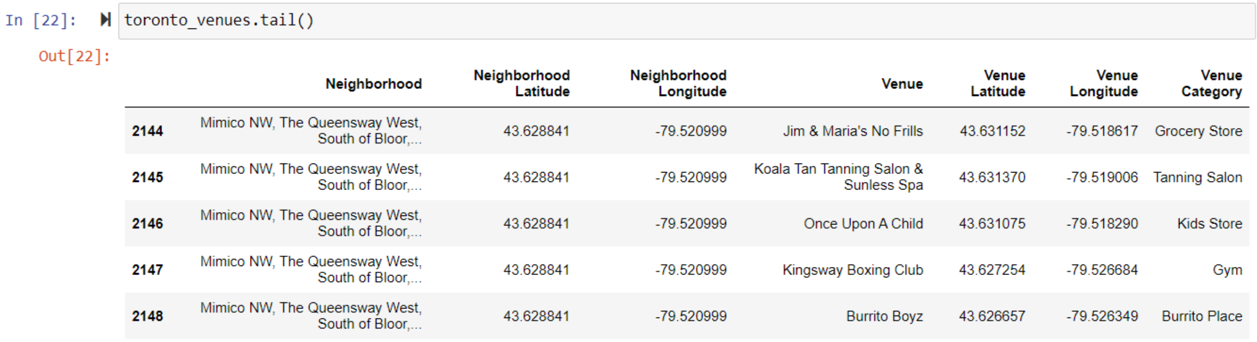
- Business personnel who wants to invest or open a restaurant.
- The freelancer who loves to have their own restaurant as a side business.
- Tourists who wants to eat Italian food.

2. Data acquisition and cleaning

2.1 Data sources

For this project we need the following data:

1. Toronto City data that contains Borough, Neighbourhoods along with their latitudes and longitudes.
    - Source: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
    - Using Beautiful Soup package, we scrap the data in tabular form.
2. Geographical Location data using Geocoder Package.
    - Source: https://cocl.us/Geospatial_data
    - 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
    - Source: https://foursquare.com/developers/apps
    - From Foursquare API we can get the name, category, latitude, longitude for each venue.



```
In [22]:  toronto_venues.tail()
Out[22]:
```

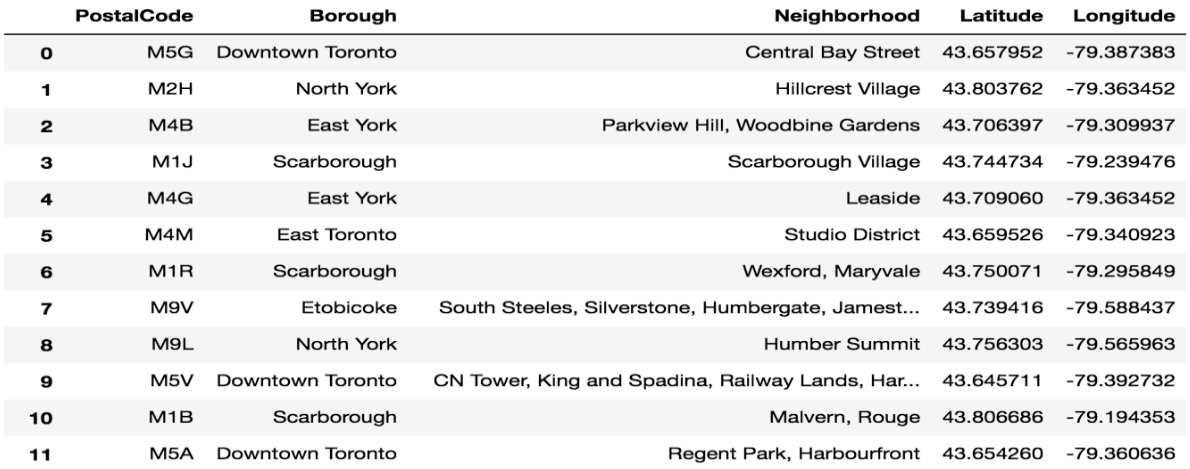
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
2144	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Jim & Maria's No Frills	43.631152	-79.518617	Grocery Store
2145	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Koala Tan Tanning Salon & Sunless Spa	43.631370	-79.519006	Tanning Salon
2146	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Once Upon A Child	43.631075	-79.518290	Kids Store
2147	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Kingsway Boxing Club	43.627254	-79.526684	Gym
2148	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Burrito Boyz	43.626657	-79.526349	Burrito Place

2.2 Methodology

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

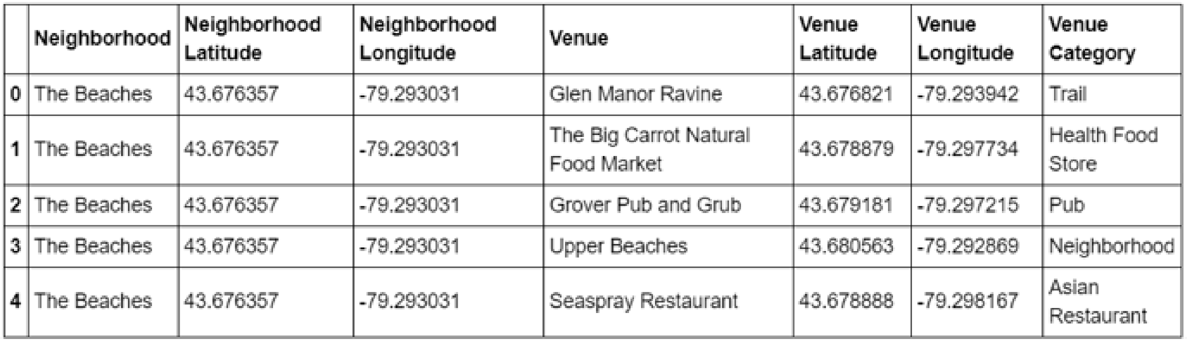
- Only process the cells that have an assigned borough. Ignore cells with a borough that is Not assigned.
- More than one neighbourhood can exist in one postal code area. For example, in the table on the Wikipedia page, you will notice that M5A is listed twice and has two neighbourhoods: Harbourfront 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 borough but a Not assigned neighbourhood, then the neighbourhood will be the same as the borough.

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



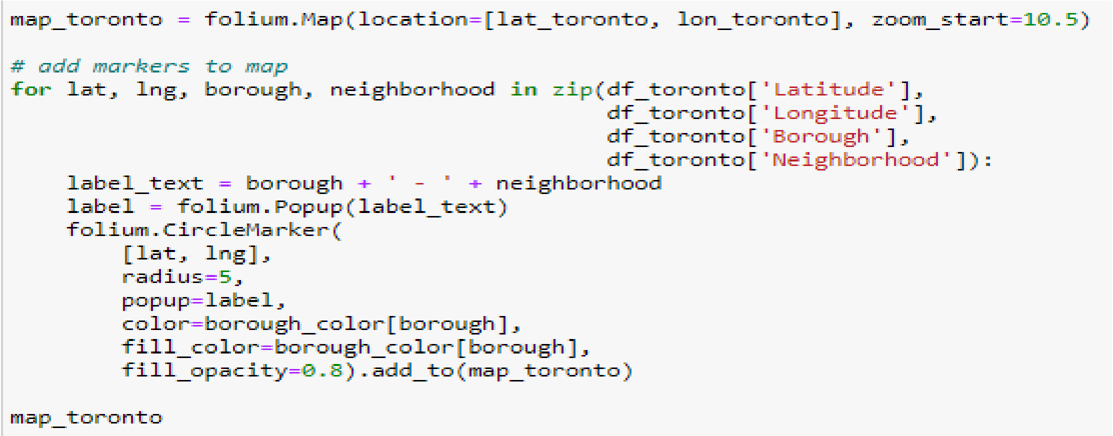
	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383
1	M2H	North York	Hillcrest Village	43.803762	-79.363452
2	M4B	East York	Parkview Hill, Woodbine Gardens	43.706397	-79.309937
3	M1J	Scarborough	Scarborough Village	43.744734	-79.239476
4	M4G	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

Now we will retrieve the venue data present within 500-meter radius of each neighbourhood using Foursquare API and merge with the above table.



	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

Now we need to visualise all neighbourhoods in a map using Folium and colour-coded each. The below bunch of code needed to do so.

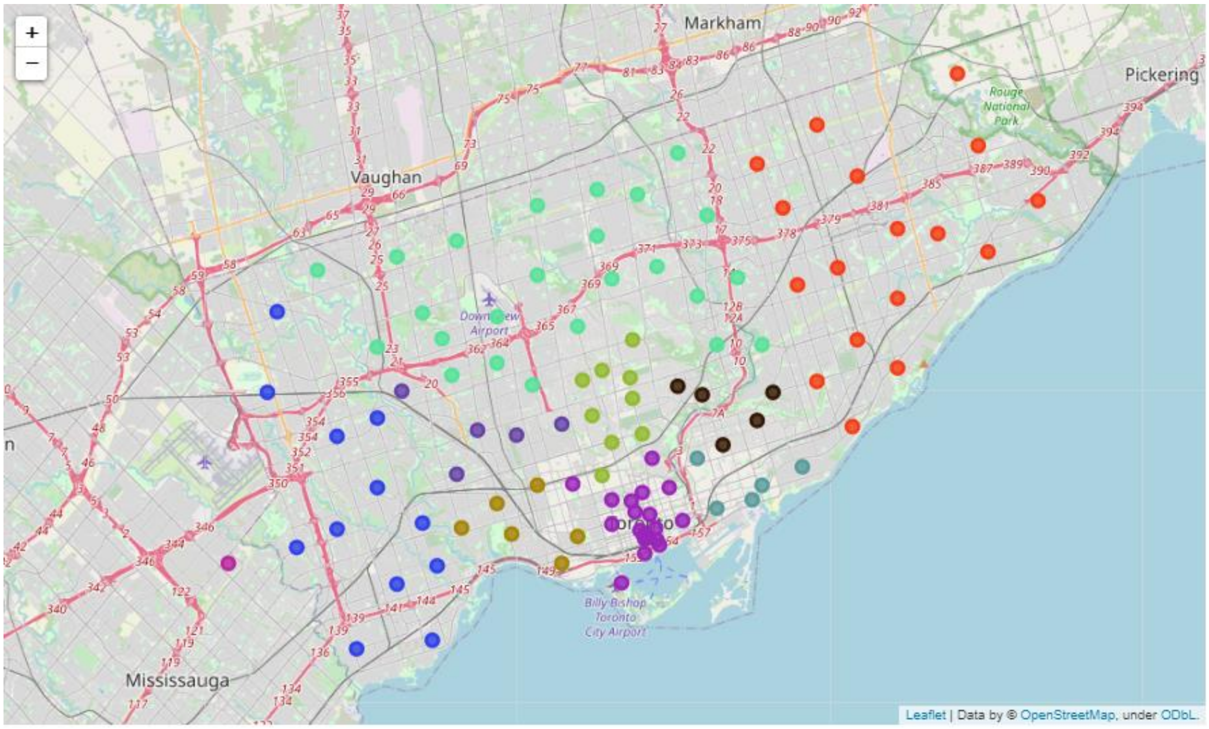


```
map_toronto = folium.Map(location=[lat_toronto, lon_toronto], zoom_start=10.5)

# add markers to map
for lat, lng, borough, neighborhood in zip(df_toronto['Latitude'],
                                           df_toronto['Longitude'],
                                           df_toronto['Borough'],
                                           df_toronto['Neighborhood']):
    label_text = borough + ' - ' + neighborhood
    label = folium.Popup(label_text)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color=borough_color[borough],
        fill_color=borough_color[borough],
        fill_opacity=0.8).add_to(map_toronto)

map_toronto
```

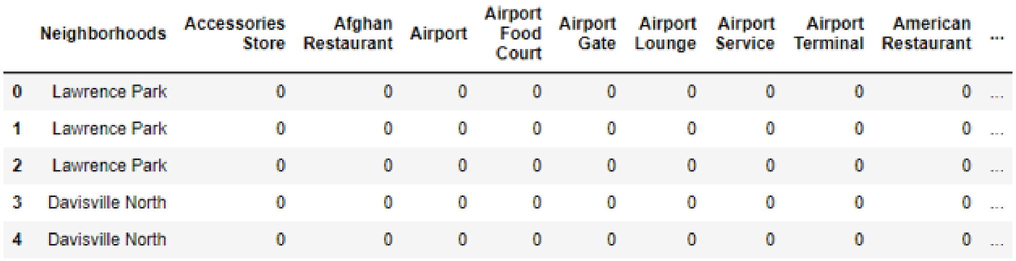
This snippet of code provided us with the map below:


+
−
Markham
Pickering
Rouge National Park
Vaughan
Downsview Airport
Toronto
Billy Bishop Toronto City Airport
Mississauga
Leaflet | Data by © OpenStreetMap, under ODbL.

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 analysing the number of Italian Restaurants all over Toronto. There was a total of 45 Italian Restaurants in Toronto. We then merged the Foursquare Venue data with the Neighbourhood data which then gave us the nearest Venue for each of the Neighbourhoods.

2.3 Data Pre-processing

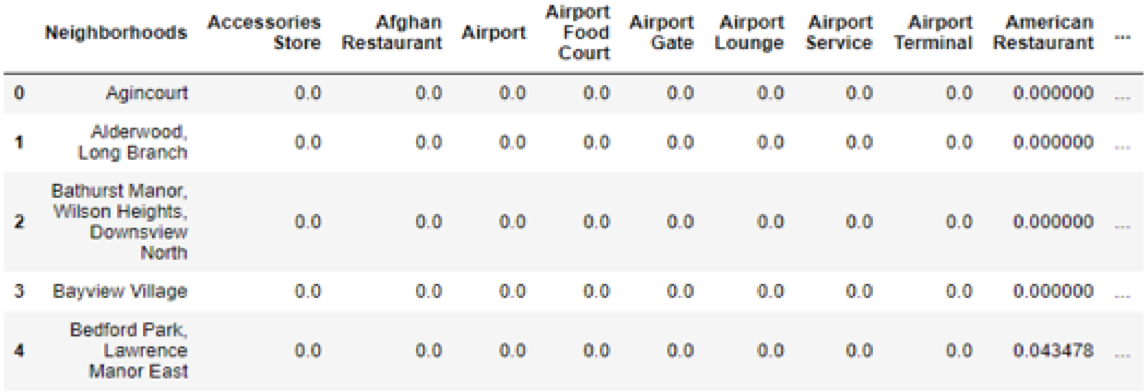
To analyse the respective Italian restaurant, present in that neighbourhood or not, we'll use One hot encoding technique. For each of the neighbourhoods, individual venues were turned into the frequency at how many of those Venues were located in each neighbourhood.



	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	...

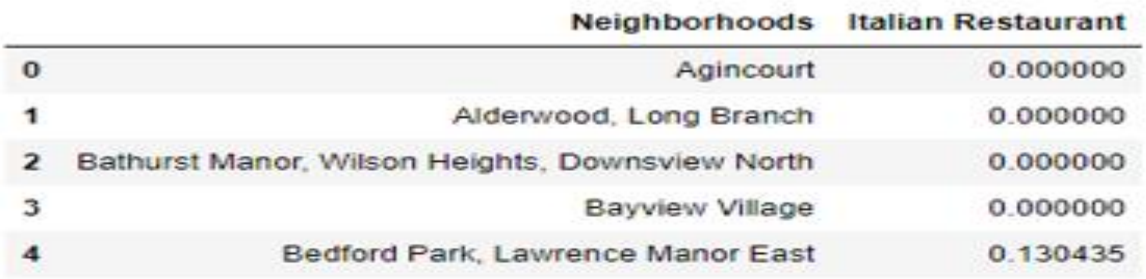
Figure 11: One Hot Encoding

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	...

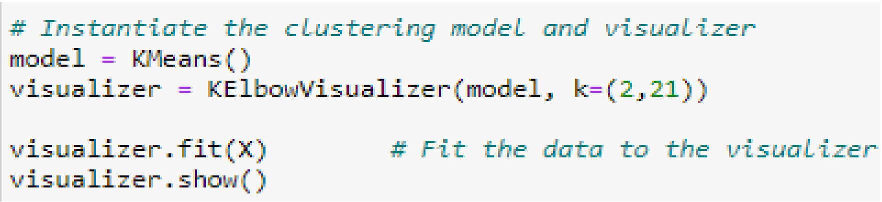
After, we created a new data frame that only stored the Neighbourhood names as well as the mean frequency of Italian Restaurants in that Neighbourhood. This allowed the data to be summarized based on each individual Neighbourhood and made the data much simpler to analyse.



	Neighborhoods	Italian 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	0.130435

3. K-Means Clustering

Now we'll cluster these neighbourhoods based on the frequency of Italian restaurants present. To do this we apply k-means clustering algorithm. To avoid the overfitting and underfitting of the model we need a optimum value of "k". There are many techniques like Elbow method, Silhouette score method to get the best "k" value. Here we're going to use Elbow method to get best "k" value. We'll import 'K Elbow Visualizer' from the yellow brick package. Then we fit our K-Means model above to the Elbow visualizer.



```
# Instantiate the clustering model and visualizer
model = KMeans()
visualizer = KElbowVisualizer(model, k=(2,21))

visualizer.fit(X)        # Fit the data to the visualizer
visualizer.show()
```

This bunch of code will give this below graph

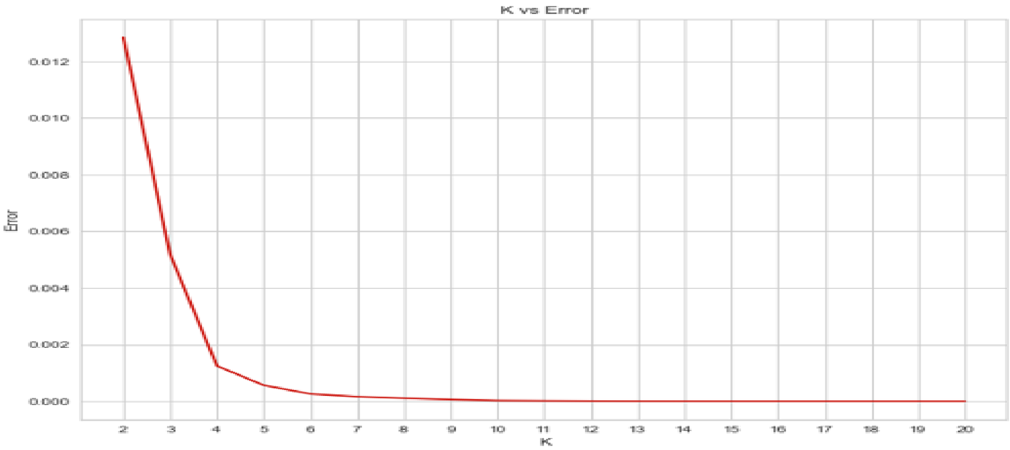
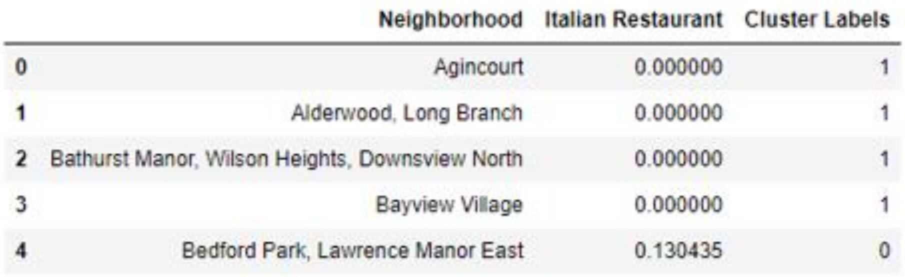


Figure 14: Finding the K vs Error Values

Here, we can see that the best k value for our dataset is 4. That means we will cluster the dataset into 4 cluster. Each of these clusters was labelled from 0 to 3 as the indexing of labels begins with 0 instead of 1.

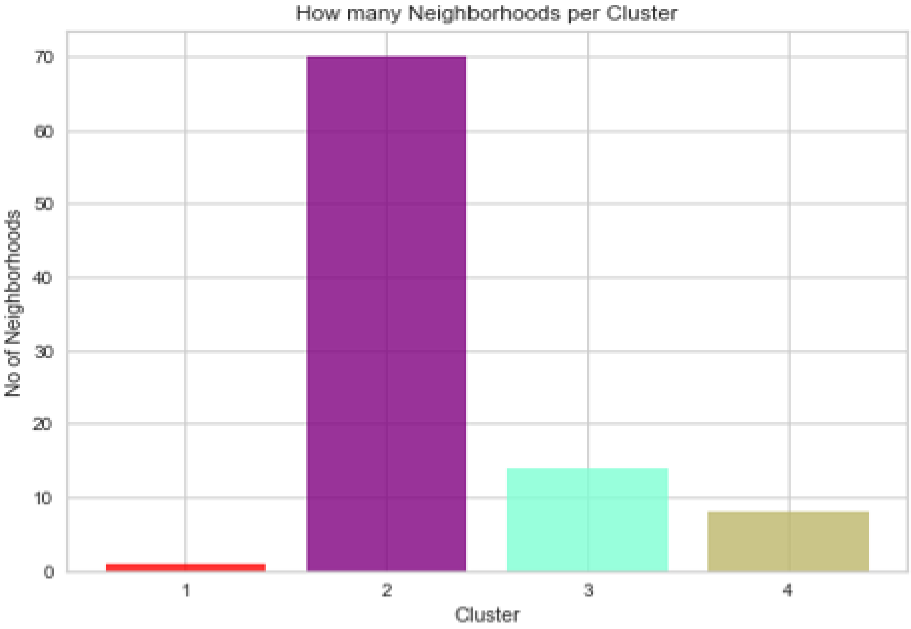


	Neighborhood	Italian Restaurant	Cluster Labels
0	Agincourt	0.000000	1
1	Alderwood, Long Branch	0.000000	1
2	Bathurst Manor, Wilson Heights, Downsview North	0.000000	1
3	Bayview Village	0.000000	1
4	Bedford Park, Lawrence Manor East	0.130435	0

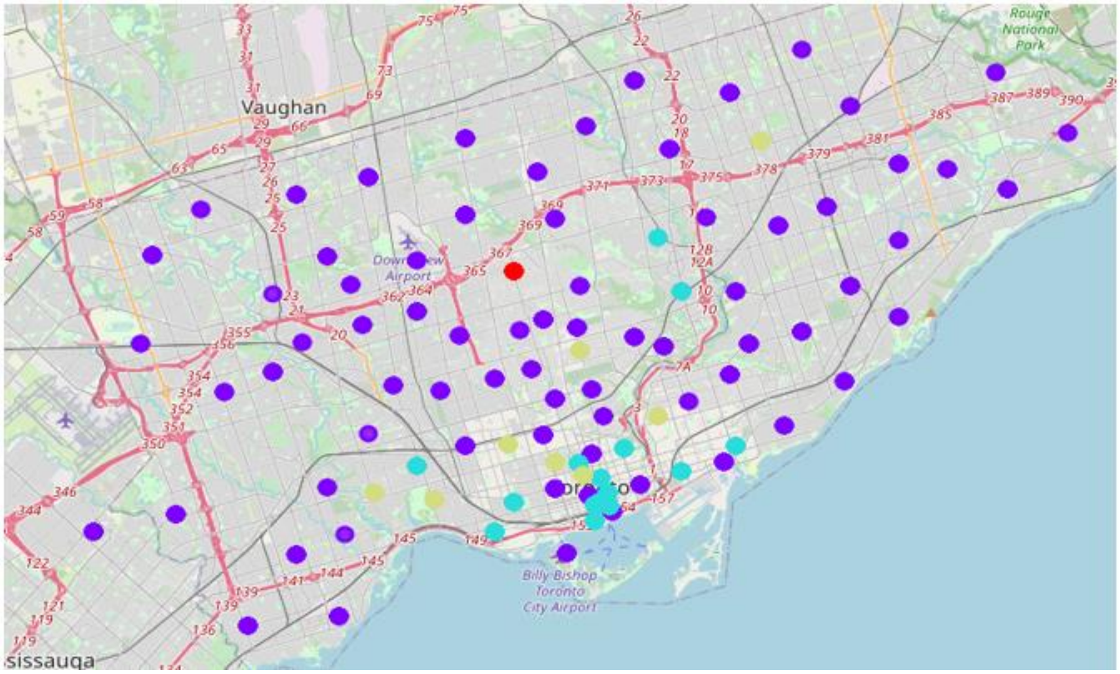
Figure 16: Appropriate Cluster Labels were added

4. Result & Outcomes

The below bar chart shows how many neighbourhoods present in each cluster.



The map below shows the different clusters that had a similar mean frequency of Italian restaurants.



5. 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.