The Battle of Neighborhoods

This report is the part of capstone project of IBM DATA SCIENCE CERTIFICATE. The main aim of this project is to apply every skill and tools learned from this course and apply for solving a business problem where we can use the Foursquare location data.

**Table of Content**

1. Business Problem
2. Target Audience
3. Data Description
4. Methodology
5. Results & Outcomes
6. Conclusion

**1. Business Problem**

The largest city of Canada Toronto is one of the most multicultural cities in the world, making life in Toronto a wonderful multicultural experience for all. Mostly Immigrants are living in canda so they have given different cuisine, religious place, ethnic stores every one gives the hints of there culture. The aim of this project is to find the best neighborhood 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.

**2. Target Audience**

* People who wants to eat Indian food
* Business professional who wants to invest or open a restaurant.

**3. Data Description**

**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 contains all the information we need to explore and cluster the neighborhoods 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:** It is the Geographical coordinates of the neighborhoods 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.



**4. Methodology**

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

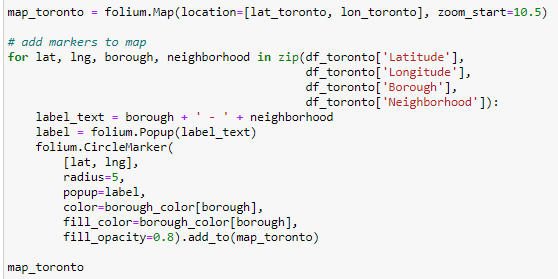
Only process the cells that have an assigned borough. Ignore cells with a borough that is **Not assigned.** More than one neighborhood can exist in one postal code area. For example, in the table on the Wikipedia page, you will notice that **M5A** is listed twice and has two neighborhoods: **Harbourfront**and **Regent Park**. These two rows will be combined into one row with the neighborhoods separated with a comma as shown in **row 11**in the above table. If a cell has a borough but a **Not assigned**neighborhood, then the neighborhood 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.



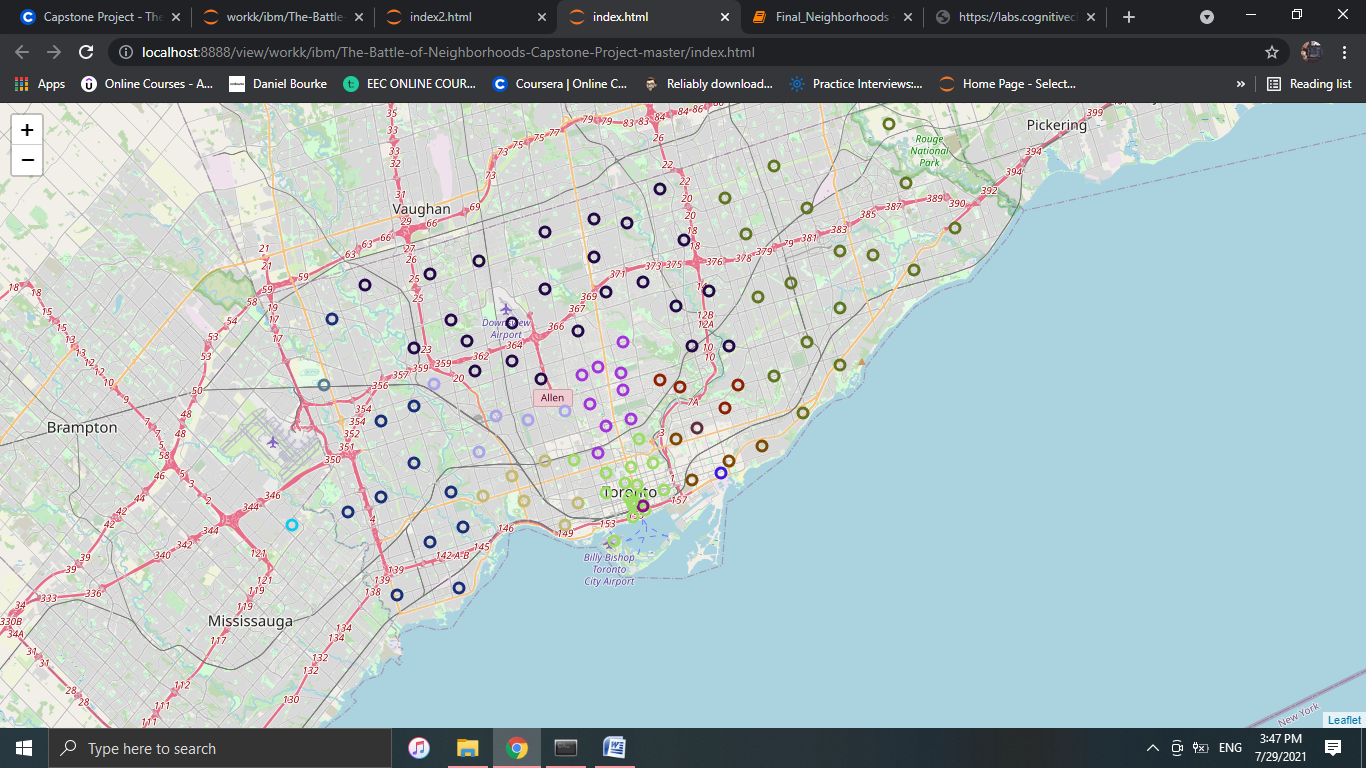
Now we will retrieve the venue data present within 500 meter radius of each neighborhood using Foursquare API and merge with the above table.



Now we need to visualize all neighborhoods in a map using Folium and color-coded each. The below bunch of code needed to do so.



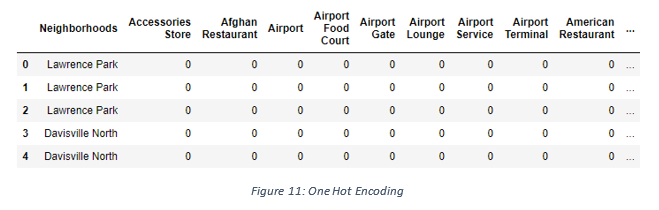
This snippet of code provided us with the map below:



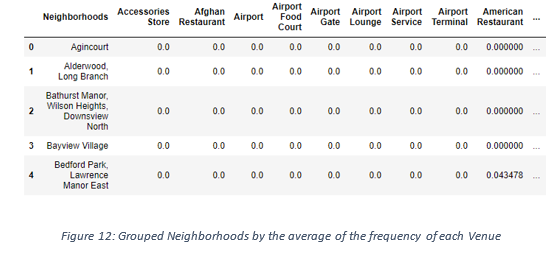
Next, we used the Foursquare API to get a list of all the Venues in Toronto which included Parks, Schools, Café Shops, and Asian Restaurants etc. Getting this data was crucial to analyzing the number of Indian Restaurants all over Toronto. There was a total of 45 Indian Restaurants in Toronto. We then merged the Foursquare Venue data with the Neighborhood data which then gave us the nearest Venue for each of the Neighborhoods.

**Data Preprocessing**

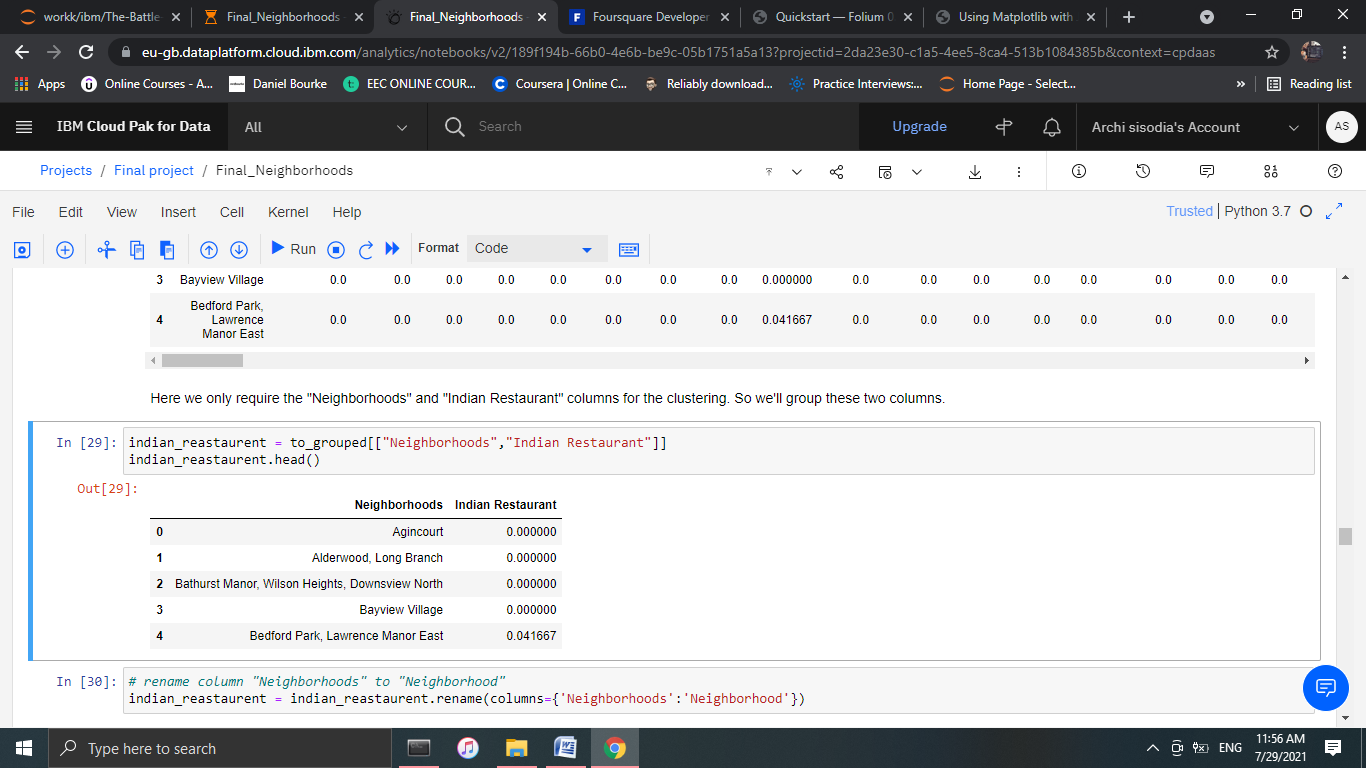
To analyze the respective Indian restaurant present in that neighborhood or not, we’ll use **One hot encoding** technique. For each of the neighborhoods, individual venues were turned into the frequency at how many of those Venues were located in each neighborhood.



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

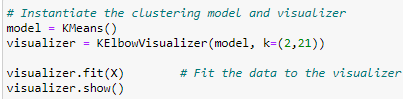


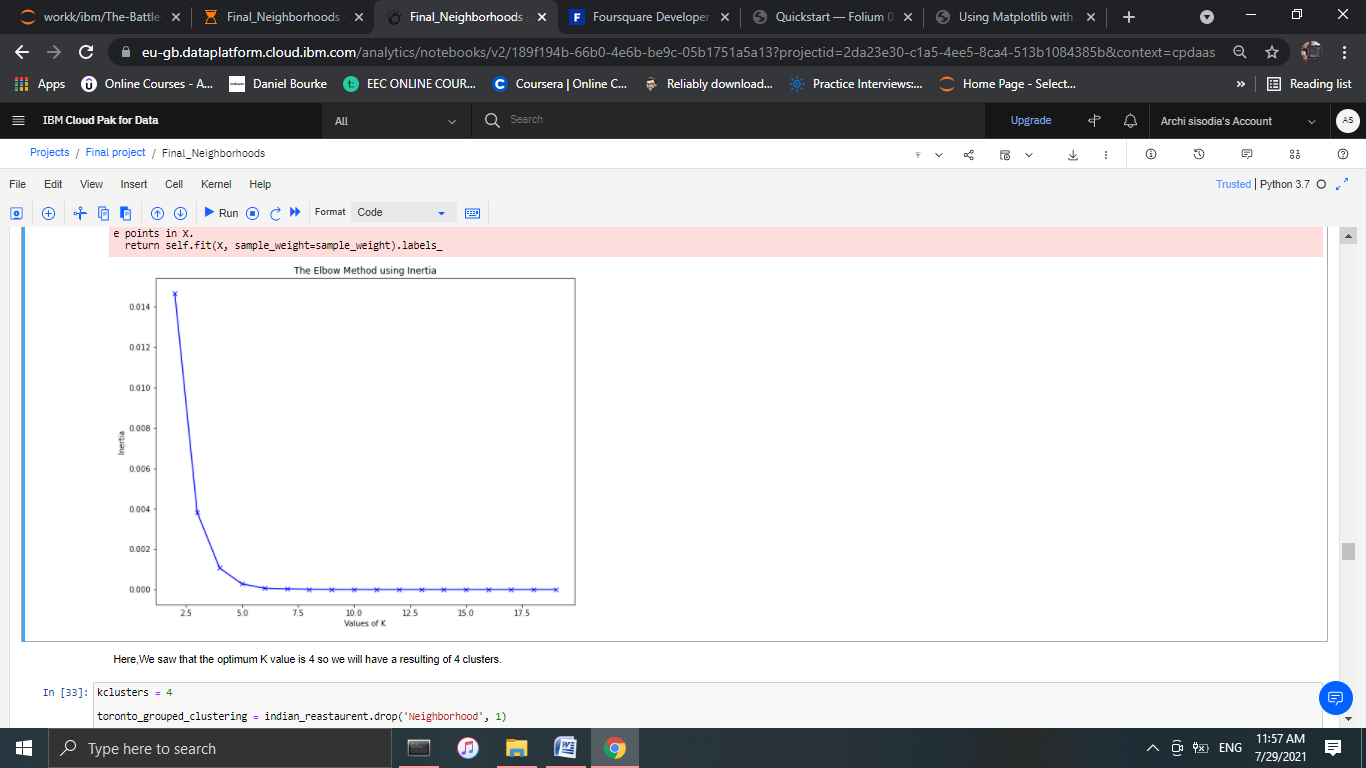
After, we created a new data frame that only stored the Neighborhood names as well as the mean frequency of Indian Restaurants in that Neighborhood. This allowed the data to be summarized based on each individual Neighborhood and made the data much simpler to analyze.

****

**K-Means Clustering**

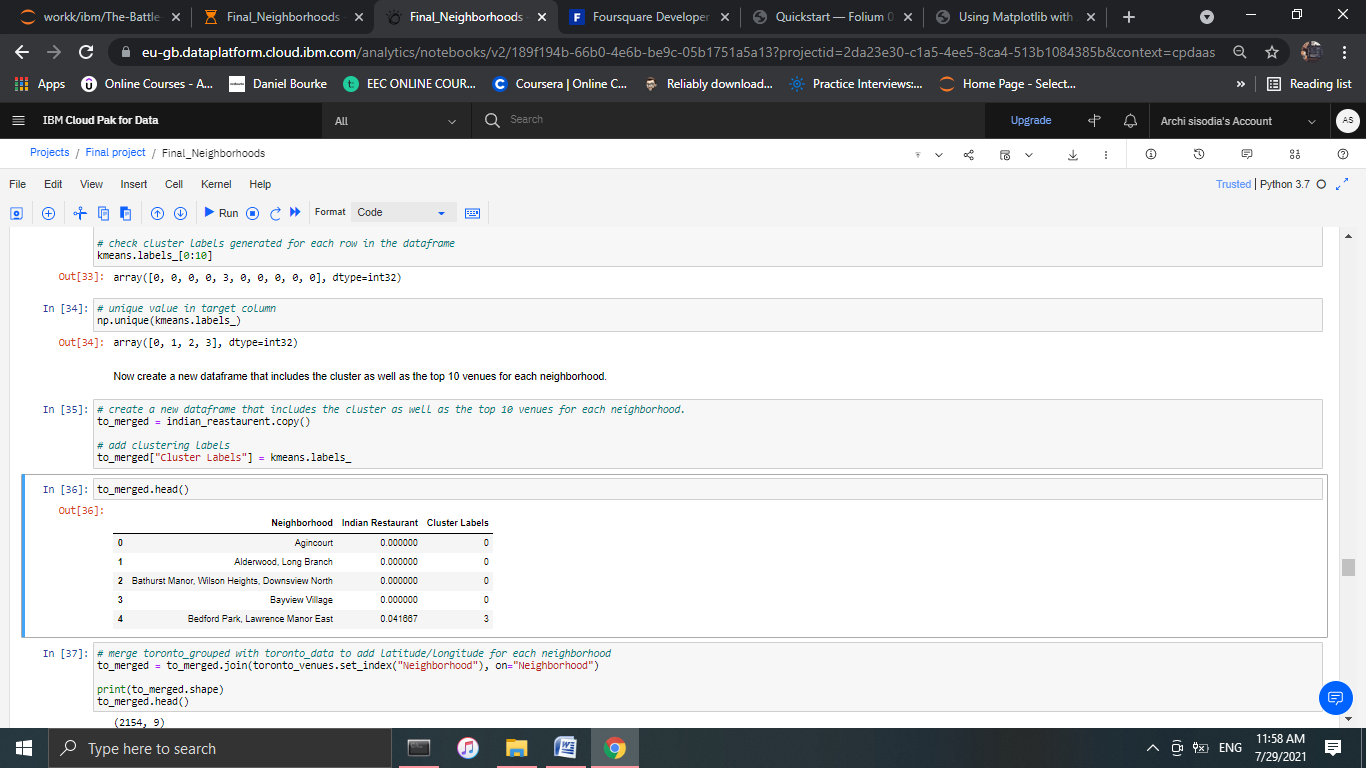
Now we’ll cluster these neighborhoods based on the frequency of Indian restaurants present. To do this we apply k-means clustering algorithm. To avoid the over fitting and under fitting of the model we need a optimum value of **“k”**.There are many techniques like **Elbow method**, **Silhouette score** method to get the best **“k”** value.Herewe’r going to use **Elbow method** to get best “k” value. We’ll import ‘*KElbowVisualizer*’ from the *Yellowbrick package.*Then we fit our K-Means model above to the Elbow visualize.





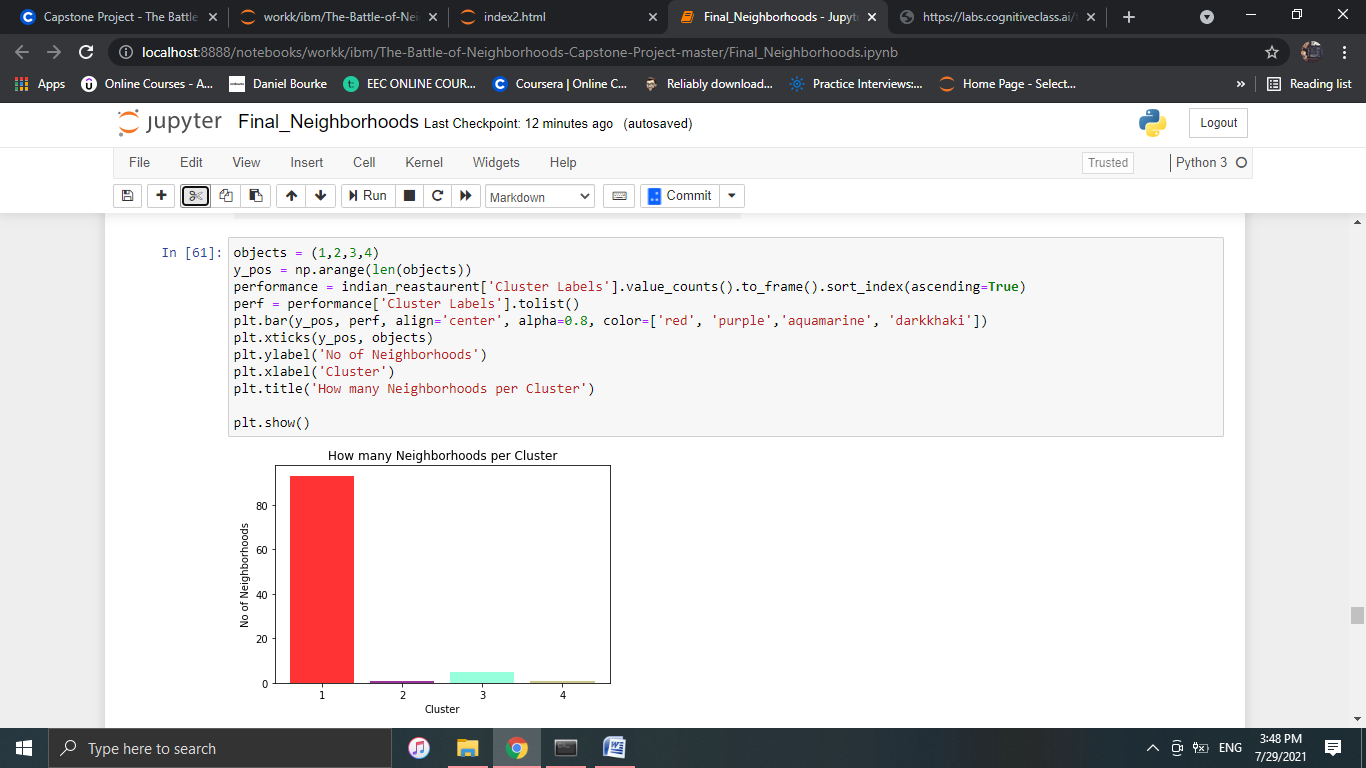
This bunch of code will give this below graph

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.

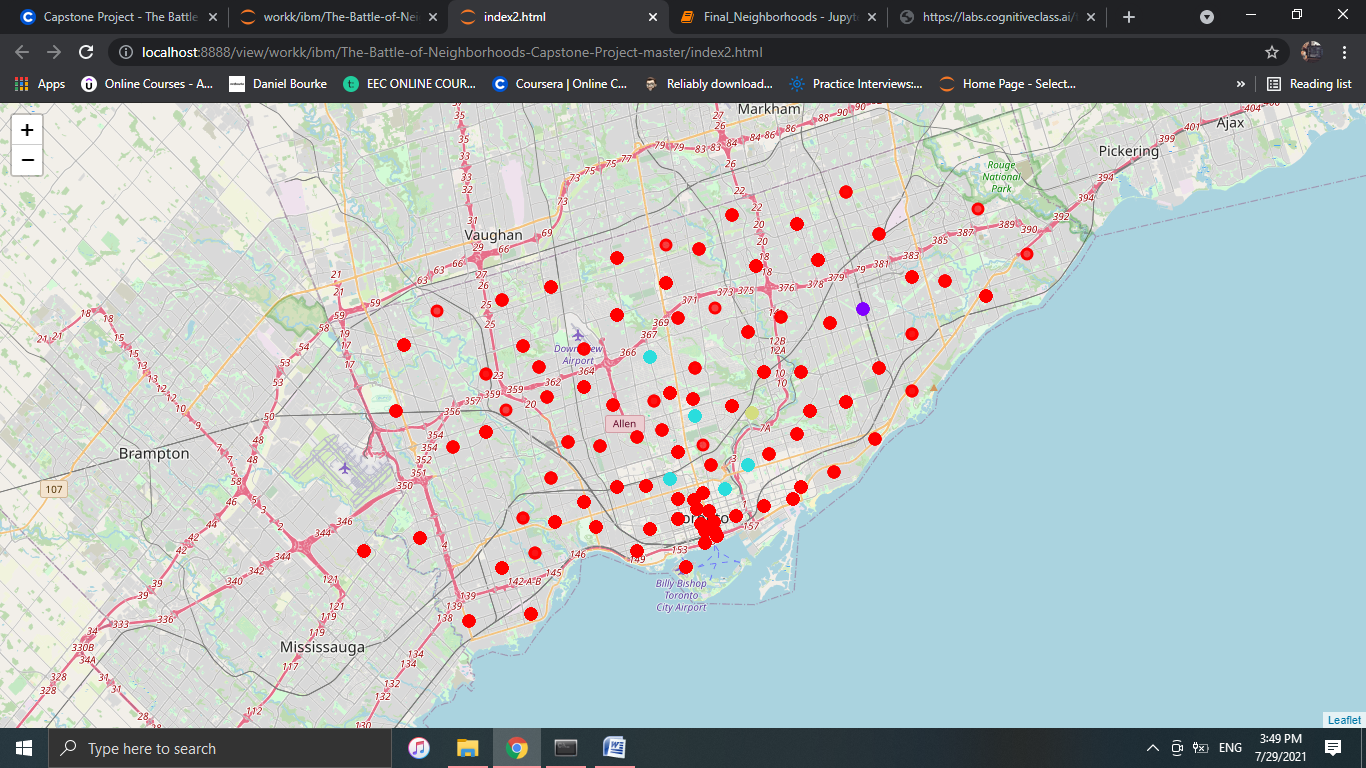


**5. Result & Outcomes**

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



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



**6. 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.