CAPSTONE PROJECT: BATTLE OF NEIGHBORHOODS

INTRODUCTION

Toronto, the capital of Ontario province is the most populous city in Canada and the 4th most in entire North America with a population of 2.9 million in 2017. Toronto is an international centre of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world, attracting tourists as well as immigrants from various different parts of the world. Of the 2011 population, a whopping 49% of people living in Toronto are immigrants (higher than the Canada-wide 29%) of which the Indian community accounts for 6.3% of the total population. Having a considerable amount of Indian population in Toronto means a good chunk of people will be practicing Indian culture and customs in Toronto and the ethic food is undoubtedly one of the biggest part of Indian lives. Research paper published by Joel Waldfogel found that Indian cuisine was the 4th most popular in the world hence all these factors indicate no shortage in demand of Indian restaurants in Toronto.

PROBLEM STATEMENT- To find the best place in Toronto to open a new Indian Restaurant.

DATA

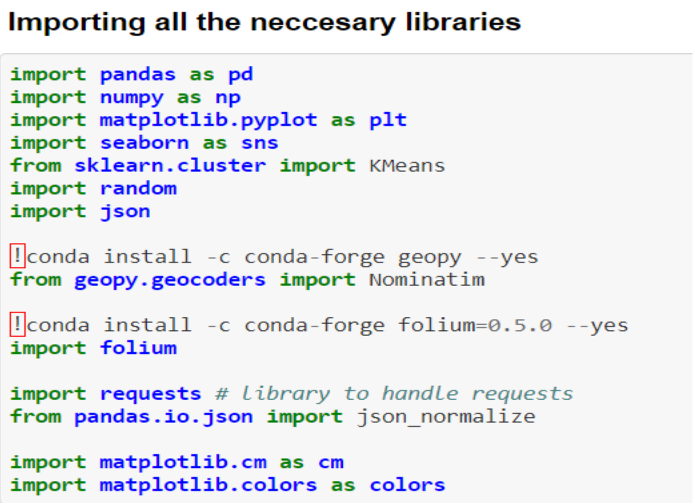
For this project we used the following data:

1) Wikipedia page 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M' to get data containing the postal codes of various regions of Canada.
2) Foursquare API: to get the information about different venues (Indian Restaurants) in Toronto.
3) Beautiful Soap for web page scraping.
4) "https://cocl.us/Geospatial_data": We will get the data of Toronto from here.

METHODOLOGY

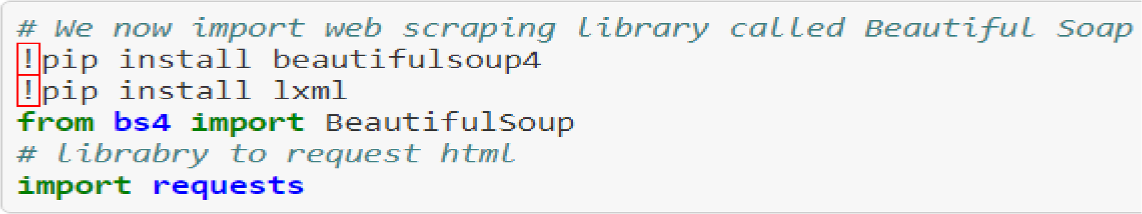
Importing Libraries, Web Scraping and Data Cleaning

At first, we imported all the necessary libraries



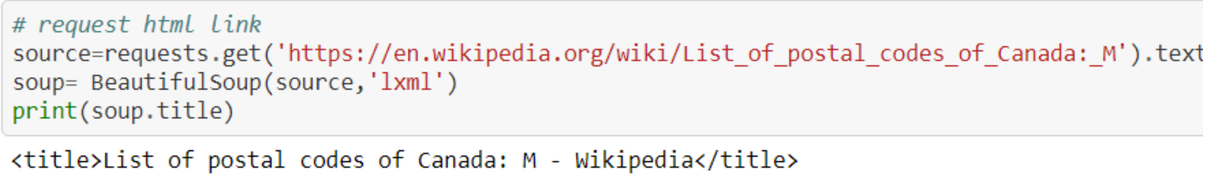
Importing all the neccesary libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
import random
import json

!conda install -c conda-forge geopy --yes
from geopy.geocoders import Nominatim

!conda install -c conda-forge folium=0.5.0 --yes
import folium

import requests # library to handle requests
from pandas.io.json import json_normalize

import matplotlib.cm as cm
import matplotlib.colors as colors
```



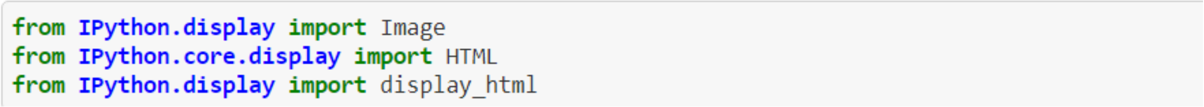
```
# We now import web scraping library called Beautiful Soap
!pip install beautifulsoup4
!pip install lxml
from bs4 import BeautifulSoup
# librabry to request html
import requests
```

We then used web scraping to get data of postal codes of various Boroughs of Canada from the Wikipedia page 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M', we store the information in a new dataset containing names of boroughs, neighbourhoods and postal codes.



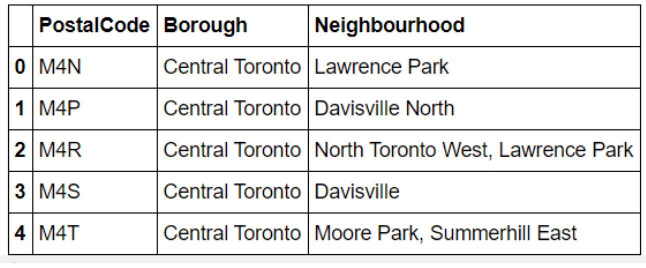
```
# request html link
source=requests.get('https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M').text
soup= BeautifulSoup(source,'lxml')
print(soup.title)
```

```
<title>List of postal codes of Canada: M - Wikipedia</title>
```



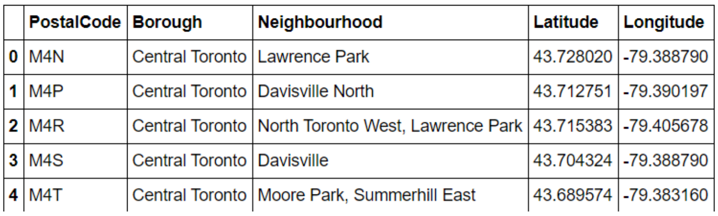
```
from IPython.display import Image
from IPython.core.display import HTML
from IPython.display import display_html
```

Now we see that the dataset contains various ‘Not assigned’ values hence we need to clean the data set by dropping the postal codes having ‘Not Assigned’ values and storing the values in a new dataframe.



	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

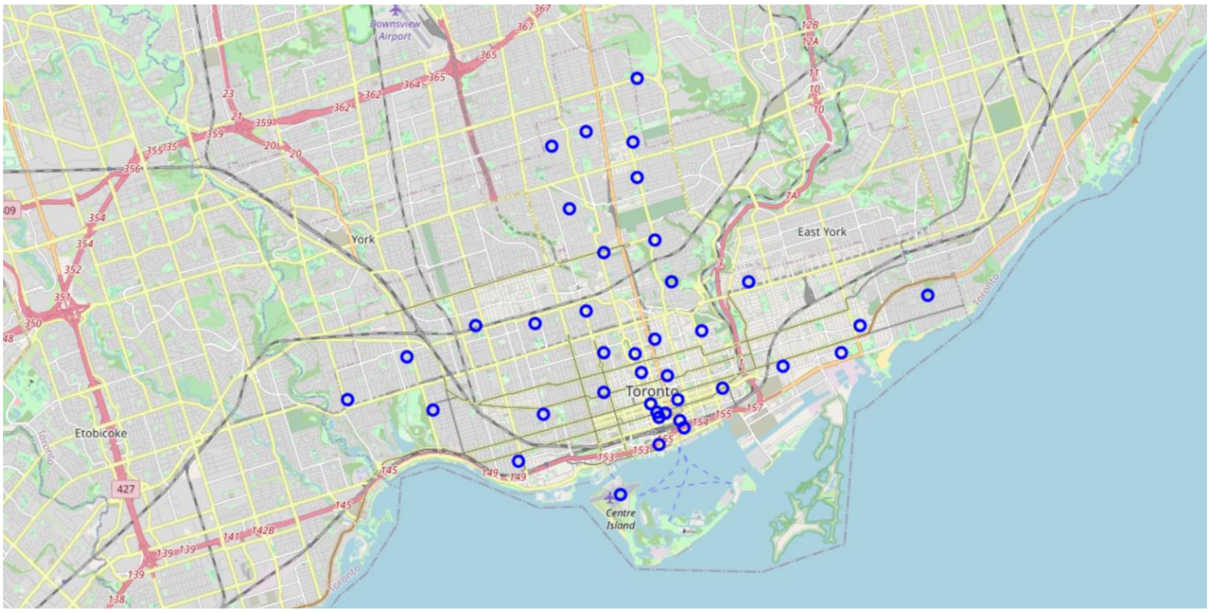
Now create a new table containing the co-ordinates of all the postal codes with help of "https://cocl.us/Geospatial_data" and then merge this table with the table containing names of boroughs and neighbourhoods. We now only look for those neighbourhoods that have word ‘Toronto’ in it as our focus is only on the city of Toronto.



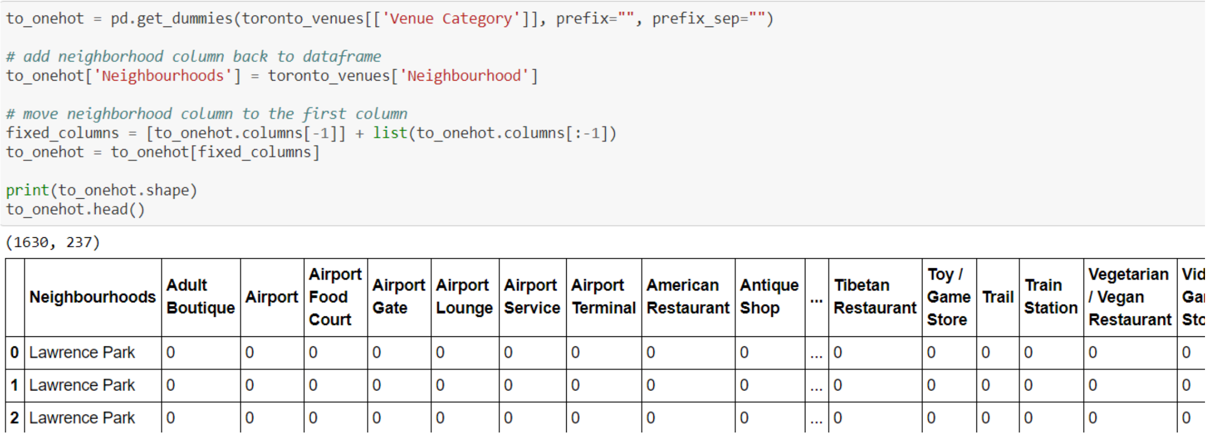
	PostalCode	Borough	Neighbourhood	Latitude	Longitude
0	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790
1	M4P	Central Toronto	Davisville North	43.712751	-79.390197
2	M4R	Central Toronto	North Toronto West, Lawrence Park	43.715383	-79.405678
3	M4S	Central Toronto	Davisville	43.704324	-79.388790
4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

Visualization

We now visualize all the neighbourhoods of Toronto with help of folium map and create popup label showing the neighbourhood’s name.



Then with the help of Foursquare API we create a get request URL to get the top 100 venues within a radius of 500 meters. Then we do one hot encoding for getting dummies of the venue category and group the whole data by Neighbourhoods.



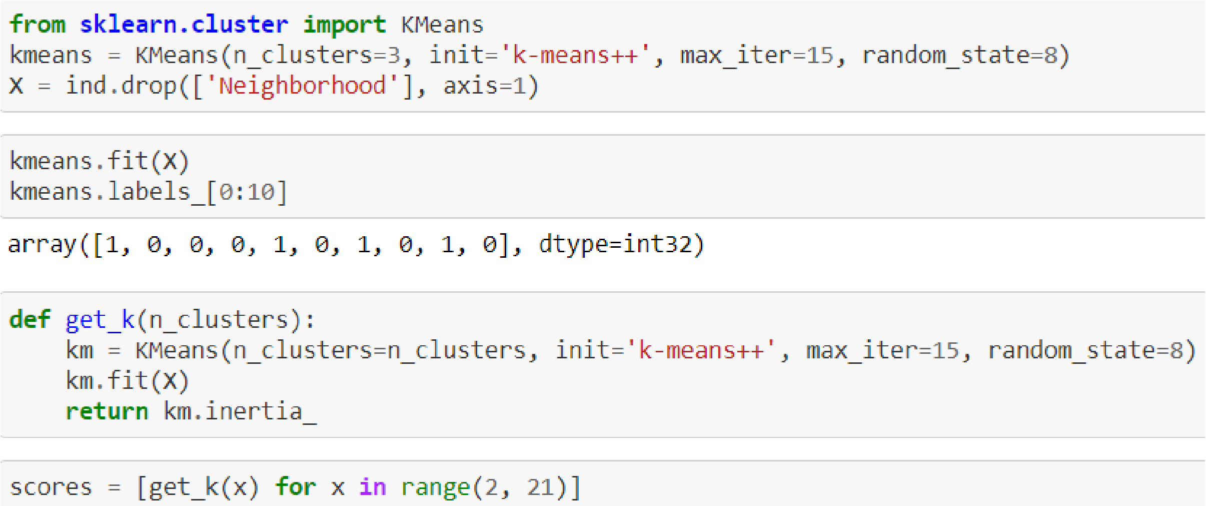
```
to_onehot = pd.get_dummies(toronto_venues[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
to_onehot['Neighbourhoods'] = toronto_venues['Neighbourhood']

# move neighborhood column to the first column
fixed_columns = [to_onehot.columns[-1]] + list(to_onehot.columns[:-1])
to_onehot = to_onehot[fixed_columns]

print(to_onehot.shape)
to_onehot.head()
```

```
(1630, 237)
```

	Neighbourhoods	Adult Boutique	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	...	Tibetan Restaurant	Toy / Game Store	Trail	Train Station	Vegetarian / Vegan Restaurant	Vid Ga Sto
0	Lawrence Park	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
1	Lawrence Park	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
2	Lawrence Park	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0

Machine Learning

We will use k-means clustering to create separate clusters containing different neighbourhoods of Toronto. But for using k-means clustering we must first find the optimal value for k so we create a graph containing different values of k and the corresponding error. By analysing the graph, we see that the elbow point for the graph occurs at k=4, hence we choose k=4. choose k=4.



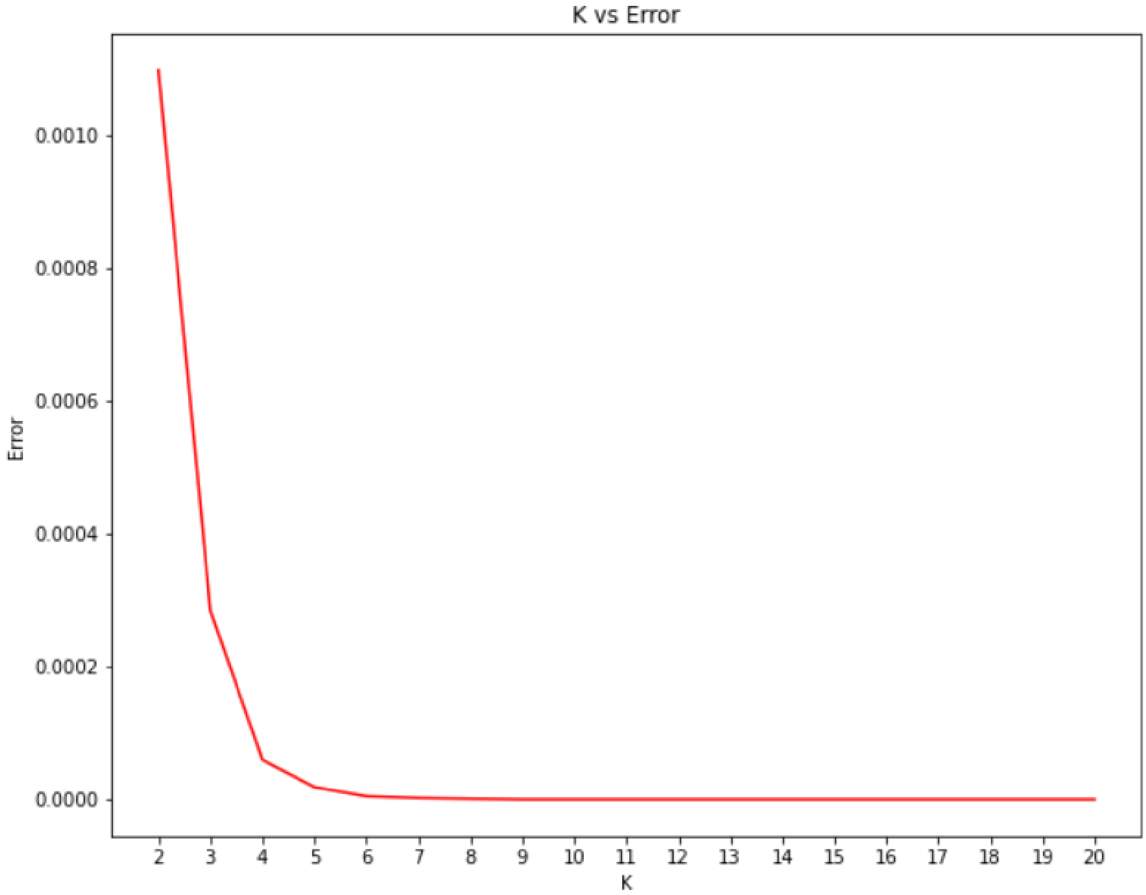
```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=3, init='k-means++', max_iter=15, random_state=8)
X = ind.drop(['Neighborhood'], axis=1)
```

```
kmeans.fit(X)
kmeans.labels_[0:10]
```

```
array([1, 0, 0, 0, 1, 0, 1, 0, 1, 0], dtype=int32)
```

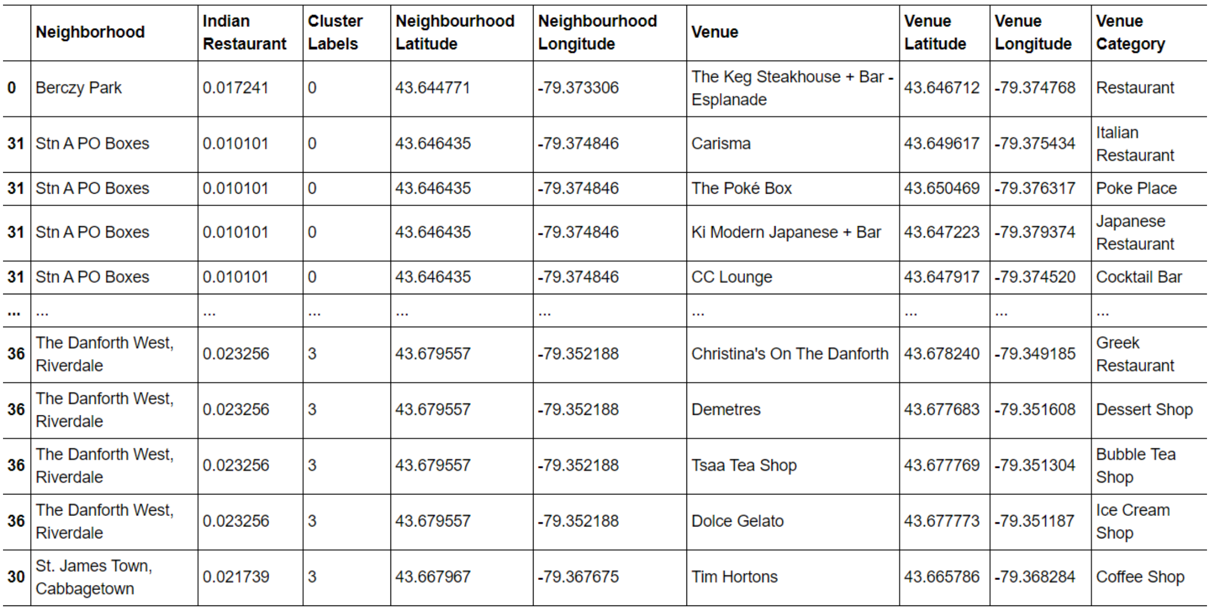
```
def get_k(n_clusters):
    km = KMeans(n_clusters=n_clusters, init='k-means++', max_iter=15, random_state=8)
    km.fit(X)
    return km.inertia_
```

```
scores = [get_k(x) for x in range(2, 21)]
```



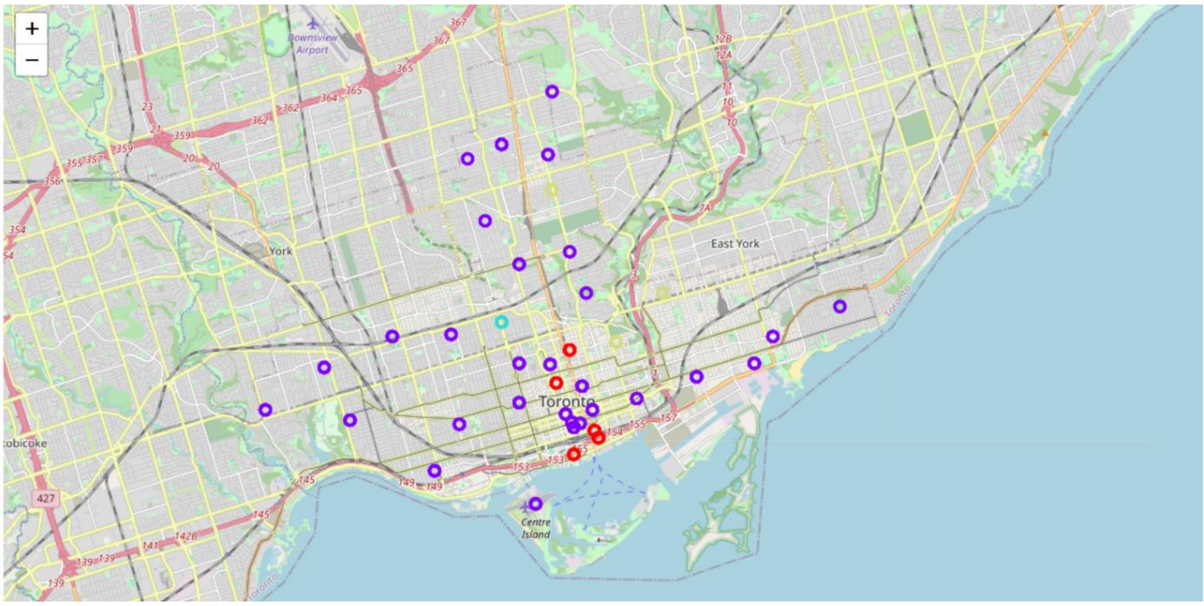
We see that the optimum **K** value is **4** so we will have a resulting of 4 clusters

So now we cluster the neighbourhoods and merge the tables so that we can see the names of neighbourhood, cluster, co-ordinates, venue categories.



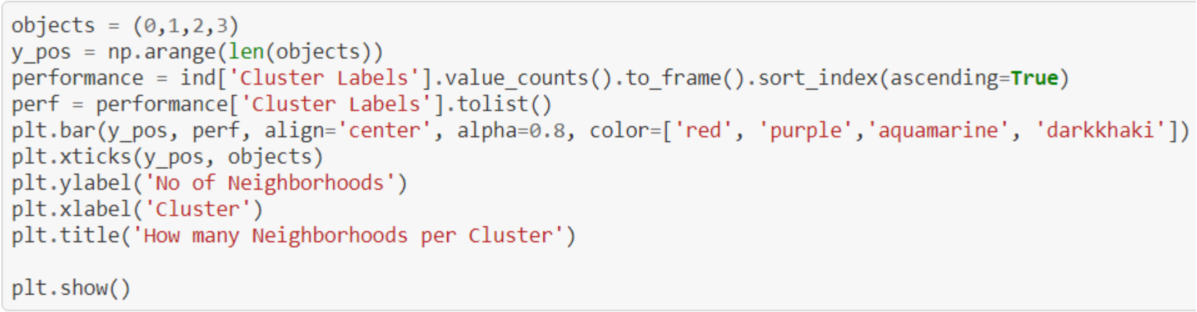
	Neighborhood	Indian Restaurant	Cluster Labels	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Berczy Park	0.017241	0	43.644771	-79.373306	The Keg Steakhouse + Bar - Esplanade	43.646712	-79.374768	Restaurant
31	Stn A PO Boxes	0.010101	0	43.646435	-79.374846	Carisma	43.649617	-79.375434	Italian Restaurant
31	Stn A PO Boxes	0.010101	0	43.646435	-79.374846	The Poké Box	43.650469	-79.376317	Poke Place
31	Stn A PO Boxes	0.010101	0	43.646435	-79.374846	Ki Modern Japanese + Bar	43.647223	-79.379374	Japanese Restaurant
31	Stn A PO Boxes	0.010101	0	43.646435	-79.374846	CC Lounge	43.647917	-79.374520	Cocktail Bar
...	...	...	...	...	...	...	...	...	...
36	The Danforth West, Riverdale	0.023256	3	43.679557	-79.352188	Christina's On The Danforth	43.678240	-79.349185	Greek Restaurant
36	The Danforth West, Riverdale	0.023256	3	43.679557	-79.352188	Demetres	43.677683	-79.351608	Dessert Shop
36	The Danforth West, Riverdale	0.023256	3	43.679557	-79.352188	Tsaa Tea Shop	43.677769	-79.351304	Bubble Tea Shop
36	The Danforth West, Riverdale	0.023256	3	43.679557	-79.352188	Dolce Gelato	43.677773	-79.351187	Ice Cream Shop
30	St. James Town, Cabbagetown	0.021739	3	43.667967	-79.367675	Tim Hortons	43.665786	-79.368284	Coffee Shop

Now we show all the clusters in a Toronto map with help of folium to better visualise the results.

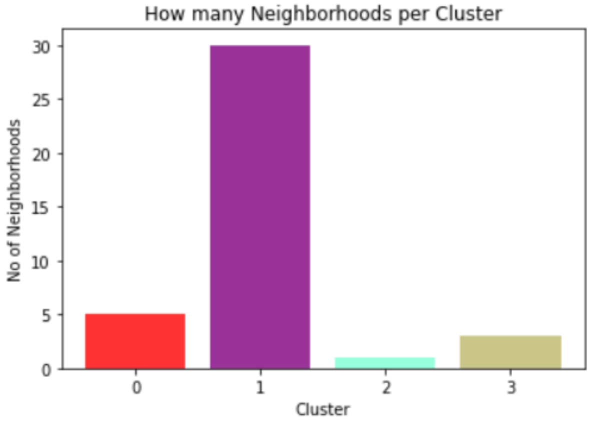


RESULTS

We find out the number of neighbourhoods in each cluster to give us the idea of each cluster. We find out that Cluster 1 has the greatest number of neighbourhoods (30) followed by Cluster 0, Cluster 3 and Cluster 2.

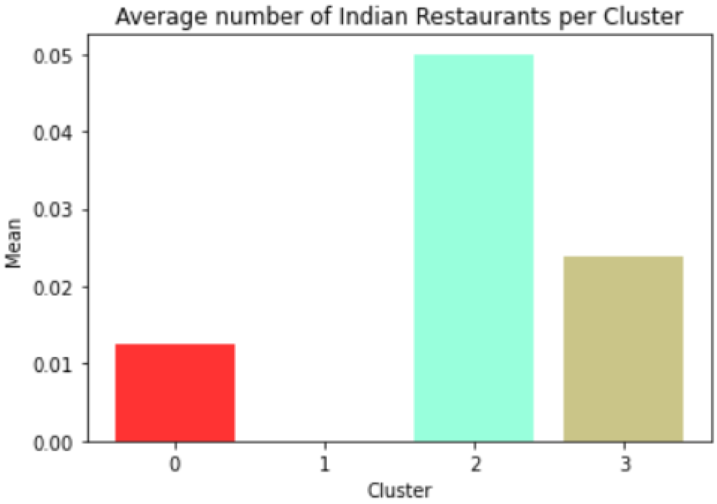


```
objects = (0,1,2,3)
y_pos = np.arange(len(objects))
performance = ind['Cluster Labels'].value_counts().to_frame().sort_index(ascending=True)
perf = performance['Cluster Labels'].tolist()
plt.bar(y_pos, perf, align='center', alpha=0.8, color=['red', 'purple','aquamarine', 'darkkhaki'])
plt.xticks(y_pos, objects)
plt.ylabel('No of Neighborhoods')
plt.xlabel('Cluster')
plt.title('How many Neighborhoods per Cluster')

plt.show()
```



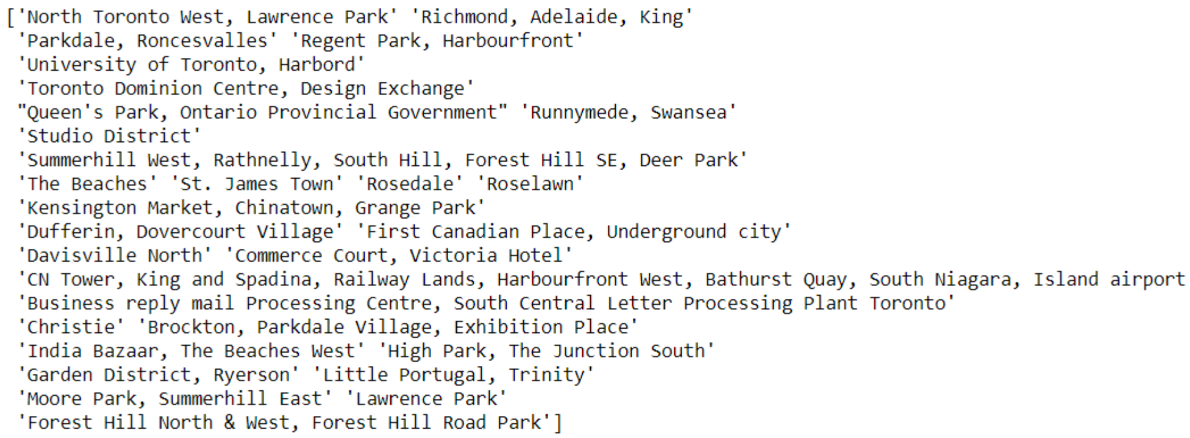
We clearly see that **cluster 1 has most no. of neighborhoods in it (30) meanwhile cluster 3 has the least (<5)**

Then we look at the number of Indian restaurants in each of these clusters and find that Cluster 2 has the highest number of Indian restaurants followed by Cluster 3 and Cluster 0 meanwhile Cluster 1 has little to no Indian restaurants



We can cleary see that **cluster 1 has least no. of Indian resturants and cluster 2 has the most**

Then we find the neighbourhoods in Cluster 2.



```
['North Toronto West, Lawrence Park' 'Richmond, Adelaide, King'
 'Parkdale, Roncesvalles' 'Regent Park, Harbourfront'
 'University of Toronto, Harbord'
 'Toronto Dominion Centre, Design Exchange'
 "Queen's Park, Ontario Provincial Government" 'Runnymede, Swansea'
 'Studio District'
 'Summerhill West, Rathnelly, South Hill, Forest Hill SE, Deer Park'
 'The Beaches' 'St. James Town' 'Rosedale' 'Roselawn'
 'Kensington Market, Chinatown, Grange Park'
 'Dufferin, Dovercourt Village' 'First Canadian Place, Underground city'
 'Davisville North' 'Commerce Court, Victoria Hotel'
 'CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport
 'Business reply mail Processing Centre, South Central Letter Processing Plant Toronto'
 'Christie' 'Brockton, Parkdale Village, Exhibition Place'
 'India Bazaar, The Beaches West' 'High Park, The Junction South'
 'Garden District, Ryerson' 'Little Portugal, Trinity'
 'Moore Park, Summerhill East' 'Lawrence Park'
 'Forest Hill North & West, Forest Hill Road Park']
```

DISCUSSION

As we can see that the Cluster 1 has the most no. of Neighbourhoods but has the least number of Indian restaurants of any cluster of Toronto. This shows that there is **a potential market for opening a new Indian restaurant in Cluster 1 neighbourhoods** like Richmond, Studio District, Rosedale etc. Meanwhile the situation is quite the opposite on Cluster 2 which as the fewest no. of neighbourhoods but the most no. of Indian Restaurants. Cluster 0 and Cluster 3 both have average no. of neighbourhoods and average no. of Indian restaurants hence they are well balanced. In the end I will conclude that if I had to open a new Indian Restaurant in Toronto, I would have opened it in Cluster 1 neighbourhood.

CONCLUSION

With this report we found out a potential region for opening a new Indian restaurant in Toronto with the help of various data science tools. The results of the report can be improved with more comprehensive analysis and with help of other better data sources in the future.