Most suitable place to open a Bangladeshi restaurant in Toronto

Qazi Mussabbir

9th Dec 2020

Contents

Figures

1. Introduction

1.1 Background

Toronto is one of the most cosmopolitan cities in the world. It has a very diverse population with over 160 difference languages spoken. On the global front, it is a very important business and financial hub. Each year Canada, and in particular Toronto draws an increased number of immigration from across the world due it's open and liberal approach to welcoming immigrants in search for better lives. As of 2016, Canada has 1,963,330 Canadians with South Asian geographical origins, constituting 5.6% of the Canadian population and 32% of Canada's Asian Canadian population. The population of South Asians in Toronto in 2016 stood at 995,125 and continuing to increase every year. In particular, the Bangladeshi community is steadily growing and has a population base of close 40,000. Many of the newly arrived immigrants from Bangladesh are very entrepreneurial and looking at ways to establish a successful business by establishing a South Asian restaurant, serving Bangladeshi cuisine in particular by identifying a particular gap in the market. In this assignment, we use location based data (from Foursquare), to identify the best borough and neighborhood to set up a Bangladeshi restaurant.

1.2 Problem

Setting up any new business comes with its own risk. In a bustling city like Toronto, the food scene can be very competitive, and opening a restaurant is a moderately risky venture. It will be prudent for any budding restaurateur, to seek data driven insights to understand the right location to setup the desired type of restaurant by looking at a numbers of parameters like cuisine concentration, trends, reviews and proximity to other revenues like malls, cinemas, etc. The project aims to develop a clustering model for identifying the right borough and neighborhood for setting up a Bangladeshi restaurant.

1.3 Interested Parties

The output of the model can be used by any entrepreneur for identifying the right location for setting up a Bangladeshi restaurant in Toronto. Additionally, in the future, the data collected could lead to insights that could be used by South Asian communities for lifestyle purposes for finding various venue information like restaurants, cafes, spiritual centres, clothing stores etc.

2. Data acquisition and cleaning

2.1 Data sources

The dataset used for this project is taken from a Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) which contains a table consisting of all the postal codes, boroughs and neighbourhoods in Toronto. The required data is scraped using a web scraper tool called 'Beautiful Soup' and converted into a *pandas* dataframe like the one shown below

	Postal Code	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
...	...	...	...
175	M5Z	Not assigned	Not assigned
176	M6Z	Not assigned	Not assigned
177	M7Z	Not assigned	Not assigned
178	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...
179	M9Z	Not assigned	Not assigned

180 rows × 3 columns

Figure 1: Dataframe composed from the scraped data

The primary objective of the data is to provide a tabular view on all the post codes, boroughs and neighborhood across Toronto. The dataframe has 180 rows and 3 columns as can be seen from the figure above.

2.2 Data cleaning

The datasets in its original form needed cleaning and transformation to be rendered fit for use by a model. To start, there were quite a few 'Not assigned' rows across the 'Borough' and 'Neighbourhood' columns. For each of the 'Not assigned' values for Borough, there are 77 'Not assigned' values for Neighbourhood column too**.** In total there were 77 rows which were dropped.

```
df[df['Borough'] == 'Not assigned'].count()
```

```
Postal Code      77
Borough          77
Neighbourhood    77
dtype: int64
```

```
In [21]: #Dropping rows in Borough column with 'Not assigned' values

         df.drop(df[df['Borough'] == 'Not assigned'].index, inplace = True)
```

```
In [13]: #checking if all the 'Not assigned' rows have indeed been dropped
         df[df['Borough'] == 'Not assigned']
```

Out[13]:

Postal Code	Borough	Neighbourhood

Figure 2: Dropped rows containing 'Not assigned' values

It was mentioned in the assignment questionnaire that If a cell has a borough but a 'Not assigned' neighborhood, then the neighborhood will need to be the same as the borough and in this respects visual checks on the dataframe were done and no values equal to 'Not assigned' were found to

meet this condition. However, the following code was executed to ensure that if there were any 'Not assigned' neighborhoods, then the neighborhood will equal to the name of the borough.

```
#If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough
for a in df_group['Borough']:
    if a != 'Not assigned':
        Borough_na = a
        #print(a)
        for i in df_group['Neighbourhood']:
            #print(i)
            y = 0
            #neighbourhood_na = i
            #print(neighbourhood_na)
            if i == 'Not assigned':
                #print("I am here")
                neighbourhood_na = Borough_na
                #print(i)
                df_group.loc[y]['Neighbourhood'] = neighbourhood_na
                print(df_group.loc[y]['Neighbourhood'])
                y += 1
```

Figure 3: Assigning 'Non assigned' neighbourhoods with valid values

2.2.1Dealing with duplicate data (combining data rows)

As mentioned in the assignment questionnaire more than one neighborhood can exist in one postal code area. For example, it is mentioned that M5A is listed twice and has two neighborhoods: Harbourfront and Regent Park. However, these two rows are already combined into one row with the neighborhoods separated with a comma. In order to ensure that there were no other duplicate values for the 'Postal Code' column, the following test was executed

```
: pd.Series(df['Postal Code']).is_unique
: True
: df[df['Postal Code'] == 'M5A']
:
```

	Postal Code	Borough	Neighbourhood
4	M5A	Downtown Toronto	Regent Park, Harbourfront

Figure 4: Checks for identifying any duplicate values for Postal Code

2.2.2 Geospatial data for Toronto

For the purpose of the getting the latitude and longitude data for all the neighbourhoods in Toronto, a csv file that has the geographical coordinates of each postal code from the link http://cocl.us/Geospatial_data has been used. The data from the csv file is converted to a pandas dataframe, which is then merged using a 'join' operation with the Toronto neighbourhood dataframe as seen below:

```
#Exploring the geospatial dataframe with the grouped dataframe
df_group_merged = pd.merge(df_group,df_geo)
```

```
df_group_merged
```

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476
...	...	...	...	...	...
98	M9N	York	Weston	43.706876	-79.518188
99	M9P	Etobicoke	Westmount	43.696319	-79.532242
100	M9R	Etobicoke	Kingsview Village, St. Phillips, Martin Grove ...	43.688905	-79.554724
101	M9V	Etobicoke	South Steeles, Silverstone, Humbergate, Jamest...	43.739416	-79.588437
102	M9W	Etobicoke	Northwest, West Humber - Clairville	43.706748	-79.594054

103 rows × 5 columns

Figure 5: Merged dataframe consisting for Geospatial data

2.2.3 Neighbourhood Profile

The Toronto population census data containing summarized neighbourhood information like age and sex, families and households, language, immigration and internal migration, ethno-cultural diversity, housing, education, income, etc has been used to understand the South Asian diversity across the neighbourhoods of Toronto. By South Asian, this report specifically considers East Indians, Bangladeshis and Pakistanis due to historical ties and cultural similarities. It is felt that South Asian population of a particular location will have implications for opening up a Bangladeshi restaurant in a that location and through exploratory analysis this report aims to provide insights in this regard. A csv file containing the aforementioned profile data was downloaded from https://open.toronto.ca/dataset/neighbourhood-profiles/ and converted to a pandas dataframe.

2.2.4 Venue data from Four Square

In order to understand the potential for setting up a new restaurant in a particular area, one needs to analyze the different types of venues near a particular neighbourhood like popular restaurants, cafes, bars, social clubs, venue trends etc. For retrieving such venue information, Four Square APIs will be used. In order to use the Four Square API, a developer account has been created with necessary credentials. The response from the API call is a json file which contains the requested venue data.

```
: CLIENT_ID = 'DJ3WT0BU4G01WASVHSNKMJME24MVFZNPXUGULBP30MRNXZAL' # your Foursquare ID
CLIENT_SECRET = 'VHRAWJ3PLFPUSI5HEPHUXQSQU510SFS5PSVG05VCGFXEFRKV' # your Foursquare Secret
VERSION = '20180605' # Foursquare API version
LIMIT = 100 # A default Foursquare API limit value

print('Your credentails:')
print('CLIENT_ID: ' + CLIENT_ID)
print('CLIENT_SECRET:' + CLIENT_SECRET)
```

```
Your credentails:
CLIENT_ID: DJ3WT0BU4G01WASVHSNKMJME24MVFZNPXUGULBP30MRNXZAL
CLIENT_SECRET:VHRAWJ3PLFPUSI5HEPHUXQSQU510SFS5PSVG05VCGFXEFRKV
```

Figure 6: Four Square developer account and credentials

3. Exploratory Analysis

3.1 Folium library and leaflet maps

For the purpose of the geo-spatial visualizations for Folium library has been used to draw interactive leaflet maps.

```
# create map of New York using latitude and longitude values
map_toronto = folium.Map(location=[latitude, longitude], zoom_start=10)

# add markers to map
for lat, lng, borough, neighbourhood in zip(final_df['Latitude'], final_df['Longitude'], final_df['Borough'], final_df['Neighbou
    label = '{}, {}'.format(neighbourhood, borough)
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='blue',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(map_toronto)

map_toronto
```

Figure 7: Folium code snippet

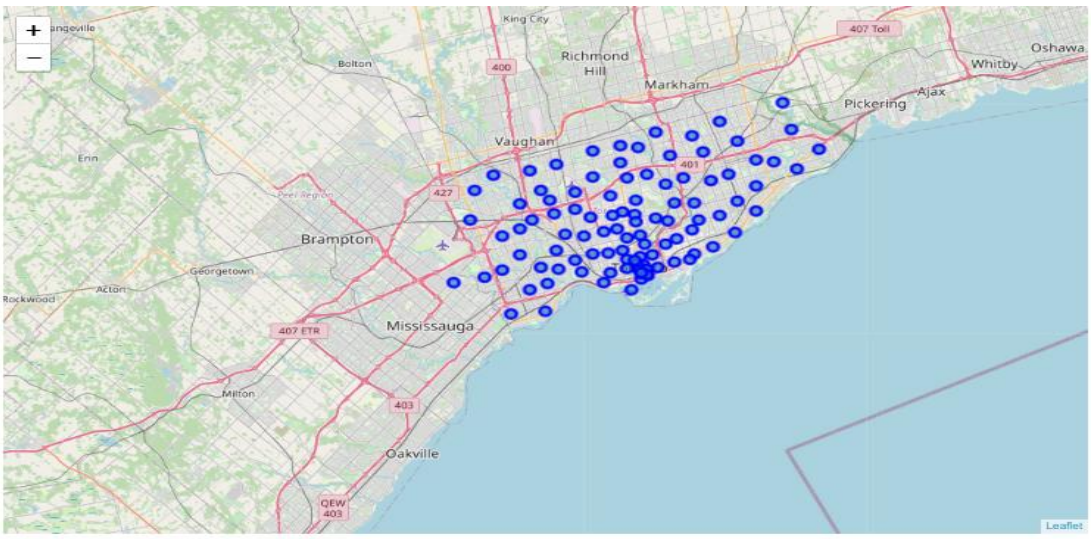

Figure 8: Map of Toronto Boroughs and Neighbourhood

3.2 Number of neighbourhoods across the boroughs

As it can be seen from the figure below, North York, Downtown Toronto and, Central Toronto are the three boroughs with the most neighbourhoods

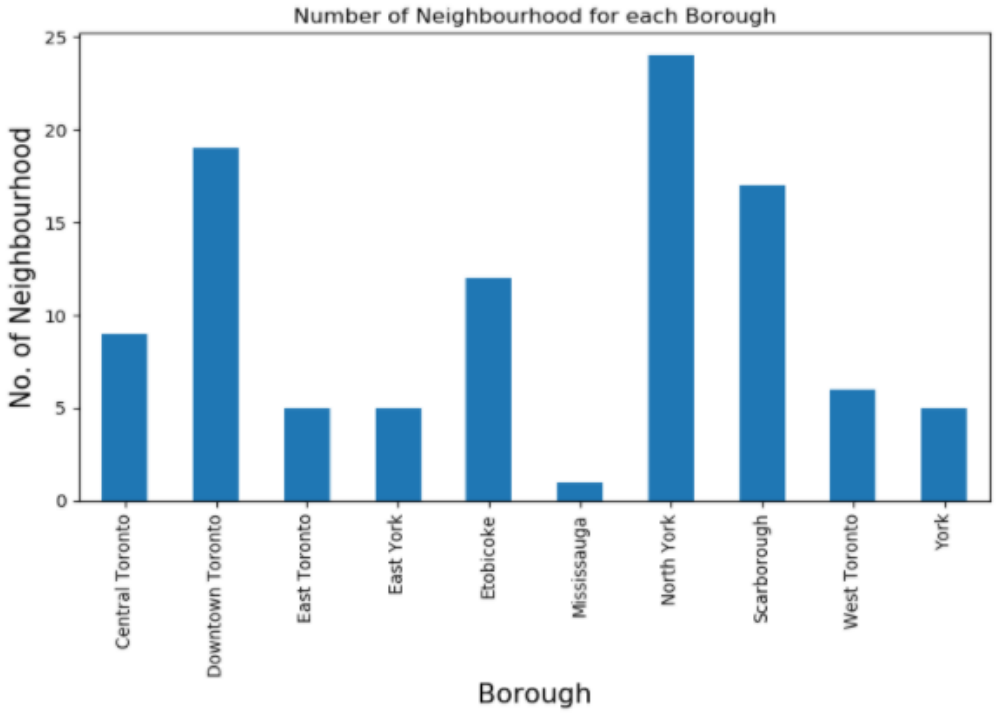

Figure 9: Number of neighbourhoods per borough

3.3 South Asian population diversity

As it can be seen from the figure below, the five most populous neighbourhoods in Toronto for South Asians are Agincount North, are Agincount South-Malvern West, Alderwood, Annex and Banbury-Don Mills. The most populous neighbourhood for Bangladeshis is Agincount South-Malvern West.

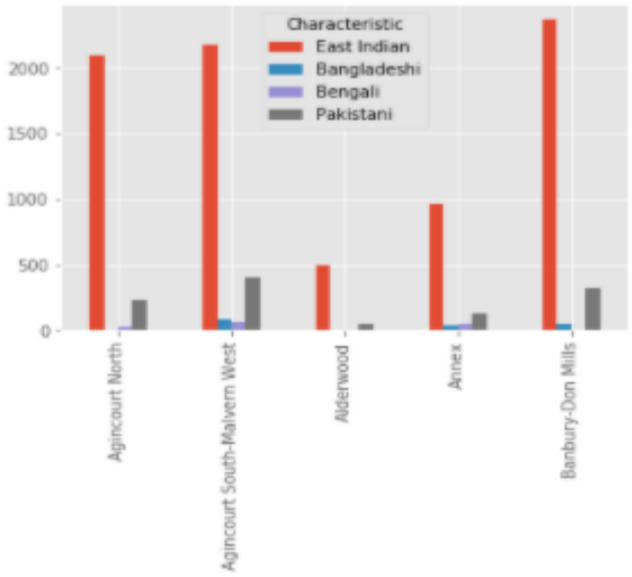

Figure 10: South Asian population split for top five neighbourhoods

3.4 Invoking Four Square API to get venue data

The function presented is invoked to get the venue names, Id and categories

```
#Function to extract the nearby venues name and venue categories for each of the neighhbourhood.
#Function similar as the one above, but this is more specifically created to be used for finding the total number of South Asian


def get_venues(lat,lng):

    #set variables
    radius=1000
    LIMIT=100
   # CLIENT_ID = os.environ['CLIENT_ID'] # your Foursquare ID
    #CLIENT_SECRET = os.environ['CLIENT_SECRET'] # your Foursquare Secret
    VERSION = '20180605' # Foursquare API version

    #url to fetch data from foursquare api
    url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            LIMIT)

    # get all the data
    results = requests.get(url).json()
    venue_data=results["response"]['groups'][0]['items']
    venue_details=[]
    for row in venue_data:
        try:
            venue_id=row['venue']['id']
            venue_name=row['venue']['name']
            venue_category=row['venue']['categories'][0]['name']
            venue_details.append([venue_id,venue_name,venue_category])
        except KeyError:
            pass

    column_names=['ID','Name','Category']
    df = pd.DataFrame(venue_details,columns=column_names)
    return df
```

Figure 11: Code snippet for function for retrieving venue related data

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Malvern, Rouge	43.806686	-79.194353	Images Salon & Spa	43.802283	-79.198565	Spa
1	Malvern, Rouge	43.806686	-79.194353	Harvey's	43.800020	-79.198307	Restaurant
2	Malvern, Rouge	43.806686	-79.194353	Wendy's	43.802008	-79.198080	Fast Food Restaurant
3	Malvern, Rouge	43.806686	-79.194353	RBC Royal Bank	43.798782	-79.197090	Bank
4	Malvern, Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
...	...	...	...	...	...	...	...
4873	South Steeles, Silverstone, Humbergate, Jamest...	43.739416	-79.588437	46 Martingrove North	43.732211	-79.589618	Bus Line
4874	South Steeles, Silverstone, Humbergate, Jamest...	43.739416	-79.588437	Panorama Park	43.747021	-79.583497	Park
4875	Northwest, West Humber - Clairville	43.706748	-79.594054	Tim Hortons	43.714657	-79.593716	Coffee Shop
4876	Northwest, West Humber - Clairville	43.706748	-79.594054	Saand Rexdale	43.705072	-79.598725	Drugstore
4877	Northwest, West Humber - Clairville	43.706748	-79.594054	Pearson Hotel And Conference Center Toronto	43.700043	-79.586059	Hotel

4878 rows × 7 columns

Figure 12: Dataframe containing all venue category

3.5 Most popular cuisines in Toronto

Using the extracted Four Square API venue data, the unique value counts for each of the different restaurant venue categories were identified and then visually represented as shown below. In total, sixty unique types of restaurant categories were found. Italian is the most popular cuisine followed by Japanese and Thai. Indian is the sixth most popular cuisine in Toronto.

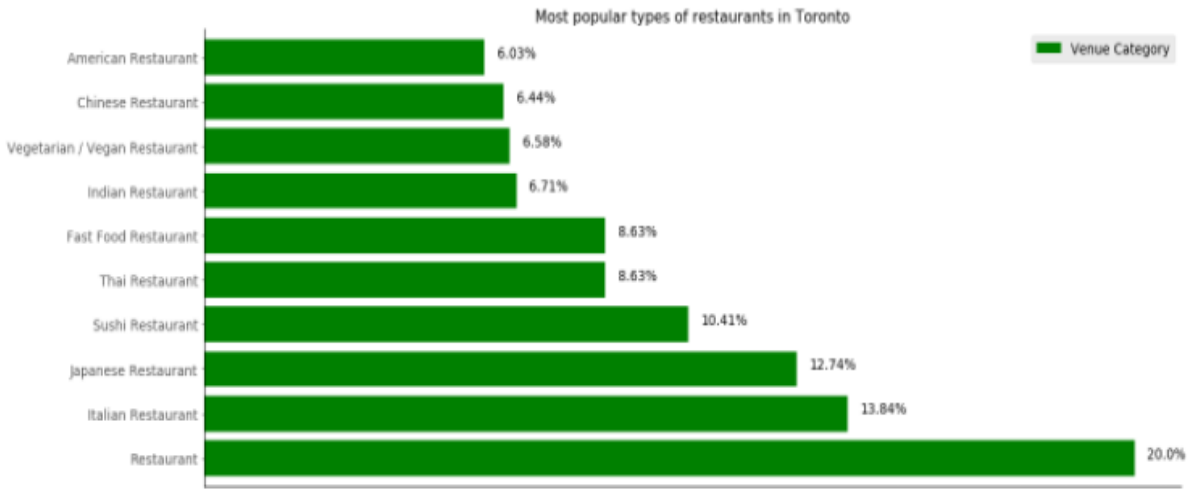

Figure 13: Most popular cuisines in Toronto

3.6 Total South Asian restaurants across the boroughs and neighbourhoods

Downtown Toronto, East Toronto, and Scarborough seems to be the top three boroughs for finding an Indian restaurant. Interestingly our retrieved data from Four Square did not result in finding any Bangladeshi or Pakistani restaurants for any of the neighbourhoods, which is obviously not correct and can be attributed to either Four Square not having the data, or Bangladeshi and Pakistani restaurants being tagged as Indian. The latter is possibly as more reasonable assumption.

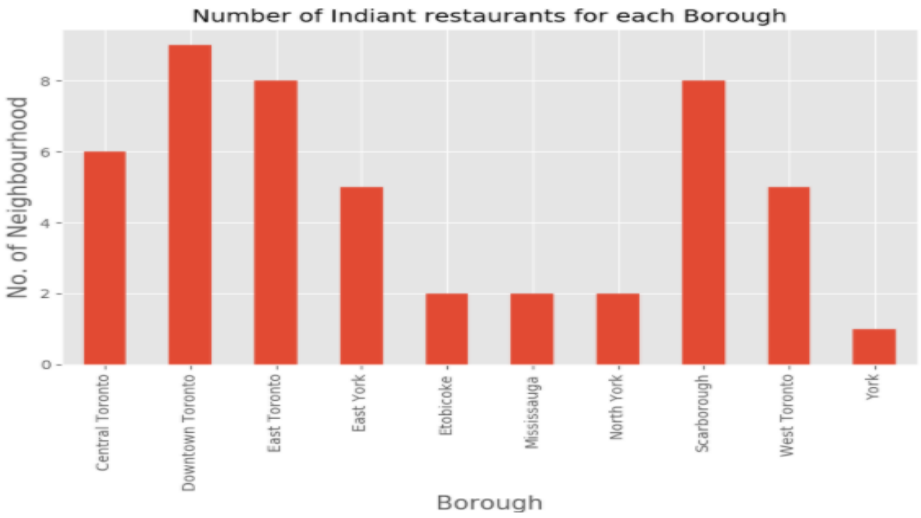

Figure 14: Number of Indian restaurant per borough

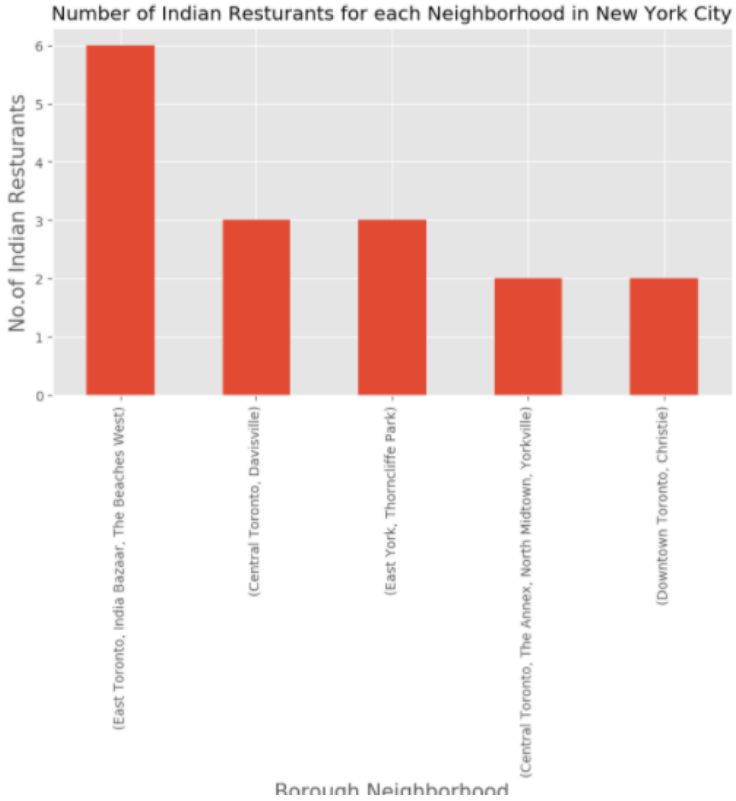

Figure 15: Top five neighbourhoods for Indian restaurants

3.5 Relationship between neighborhood ethic demographic and South Asian restaurants

The names of the neighborhoods in the population data presented in section 3.3 do not completely match with names in the dataframe shown above in section 3.6. However, just comparing the two sets of data, there seems to be no strong association with south population density across with neighbourhoods with the proliferation of South Indian restaurants (i.e. Indian restaurants in this regards). The neighbourhood of 'The Annex' seems to be only neighbourhood to have both a higher population of South Asian along with a higher prevalence for Indian restaurants.

4 Model Outcome

For purpose of this assignment, an unsupervised clustering model need to be built using the K-Means clustering method

4.1 Clustering neighbourhoods of Toronto

The code snippet for executing the model is shown below:

```
# set number of clusters
kclusters = 5

toronto_grouped_clustering = toronto_grouped.drop('Neighbourhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(toronto_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

```
array([3, 3, 3, 3, 1, 1, 1, 1, 1, 1])
```

```
# add clustering labels
neighbourhoods_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)
```

```
neighbourhoods_venues_sorted.shape
```

```
(98, 12)
```

```
kmeans.labels_
```

```
array([3, 3, 3, 3, 1, 1, 1, 1, 1, 1, 3, 1, 3, 1, 1, 1, 3, 0, 1, 1, 1, 3,
       1, 3, 3, 1, 1, 3, 1, 1, 1, 1, 3, 3, 3, 1, 1, 3, 3, 3, 3, 1, 3, 3,
       1, 3, 1, 1, 1, 1, 1, 3, 1, 1, 1, 3, 1, 4, 1, 1, 1, 3, 3, 1, 1, 1,
       1, 1, 1, 1, 3, 3, 3, 1, 1, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3,
       3, 3, 3, 1, 1, 3, 1, 1, 1, 2])
```

```
toronto_grouped.shape
```

Figure 16: Code snippet for K-means clustering model

The clustered model outcome is then visually plotted in folium as seen below:

4.2 Examine the clusters

- Cluster 0 is in Scarborough with the neighbourhoods 'Cliffside, Cliffcrest, Scarborough Village West' . It was not grouped with any other clusters due to the unique venue category. Cluster 0 has a Pizza place as the most common venue category.
- Cluster 1 is the largest cluster of neighborhood postal codes spread across the boroughs in Toronto. Cluster 1 can be categorized having alots of coffee shops, cafes and park. Interestingly

Indian restaurant has been observed to be 1st common venue for a neighbourhood within this cluster.

- Cluster 2 is another unique cluster in North York with parks, restaurants and pool.
- Cluster 3 is the 2nd largest cluster with primarily coffee shops and pizza places, banks and restaurants. This seems like an exciting and happening part of Toronto.
- Cluster 4 is a unique cluster with hotels and coffee shops, drug stores.

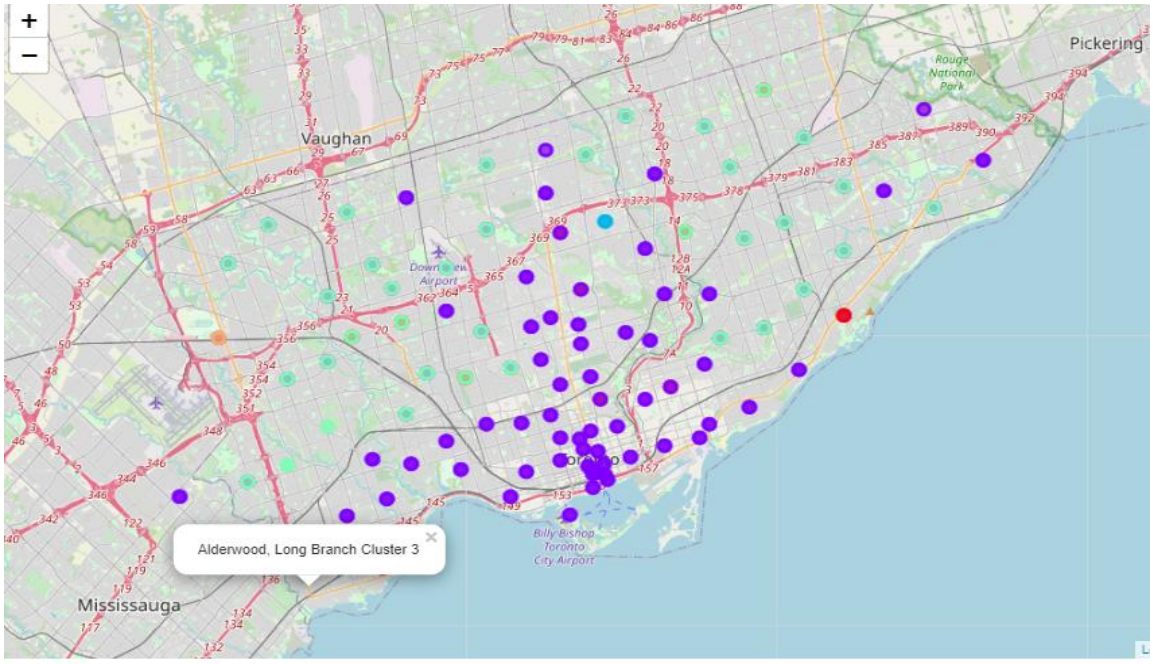

Figure 17: Folium view of the clustered neighbourhoods

4.3 Results and discussions

We have reached the end of the analysis, in this section we will document all the findings from above clustering & visualization of the dataset. In this project, we started off with the business problem of identifying a good neighborhood to open a new Bangladeshi restaurant. To achieve that we looked into all the neighborhoods in Toronto, analysed the South Asian population in each neighborhood & number of South Asian (Indian, Bangladeshi and Pakistani) restaurants in those neighborhoods to come to conclusion about which neighborhood would be a better spot. We have used variety of data sources to set up a very realistic data-analysis scenario. We have found out that

- Of the total boroughs we identified that only Central Toronto, Downtown Toronto, East Toronto, East York, North York & Scarborough boroughs have high amount of Indian restaurants with the help of Bar charts
- The borough/neighbourhood with highest number of India restaurant are East Toroto (India Bazaar, The Beaches West), Central Toronto (Davisville), East York (Thorncliffe Park), Central Toronto (The Annex, North Midtown, Yorkville), Downtown Toronto (Christie)

- Although Scarborough has a high density of South Asians living in it, the number of South Asian (Indian or Bangladeshi) restaurants do seem to as much as the other populous regions for South Asians.
- Cluster 3 seems like the ideal cluster to open a new Bangladeshi restaurant, due to the nature of the cluster where lots restaurants exists already. Moreover, cluster 3 quite a few Scarborough neighbourhood in it and because of the reason in the point above, it makes all the sense to open the Bangladeshi restaurant in Scarborough.

4.4 Conclusion

In order to ensure that Bangladeshi restaurant is opened in a location where there will be minimal competition from similar cuisines (e.g. Indian, Pakistani etc), and taking into account the South population density across the boroughs and neighbourhood, Cluster 3 and more precisely Scarborough will be most logical location to open the intentioned restaurant.