Indian Restaurant Startup Recommendation in Toronto Neighbourhood

Coursera Capstone Project

Anand Sukumar

December 2019

Indian Restaurant Startup Recommendation in Toronto Neighbourhood

1. Introduction

1.1. Background

Toronto, the capital of the province of Ontario, is a major Canadian city along Lake Ontario's northwestern shore. It's a dynamic metropolis with a core of soaring skyscrapers, all dwarfed by the iconic, free-standing CN Tower. Its richness in harbouring and welcoming people from various countries and different ethnic backgrounds especially India, China, Italy etc., speaks its diversity. Year over year, the population density is spiking across multi cultural backgrounds and thereby invites more business opportunities to establish ethnic specific amenities for people to make use of. A good restaurant established on top of a specific ethnic style would be a favourable option to go for in such environment/society.

1.2. Problem

This project focuses on recommending a favourable venue for starting a good Indian restaurant in Toronto neighborhoods. Toronto has quite a good percentage of Indian community, and drilling down in detail to identify the specific location neighborhoods based on density / origin will provide better confidence in establishing the restaurant in order to improve the customer base. Another important point to consider here is to explore the already established venues around these neighborhoods and choose a close proximity location to keep up the customers inflow consistent.

1.3. Target Audience

Who will be benefited by this Project findings?

Any Entrepreneur who is looking to startup an Indian restaurant in Toronto neighborhoods. He/She can make use of this project findings to decide upon the favourable locations

Indian community people or any foodies who are interested to have good Indian food at their proximity locations can make use of this project findings

2. Data acquisition and cleaning

2.1. Data sources

A. Postal Codes of Toronto - Am using the wikipedia portal (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) to find the Neighbourhood information of Toronto location. This lookup consists of Postal Codes, Boroughs and Neighbourhood details of Toronto.

B. Geospatial Data - Am extracting the geospatial data from http://cocl.us/Geospatial_data which provides the Latitude - Longitude coordinate details of requested neighborhoods.

C. Demographics of Toronto - Am using the wikipedia portal link (https://en.wikipedia.org/wiki/Demographics_of_Toronto) to get the demographics details of Toronto whereby we can find the Indian ethnic origin neighbourhood details.

D. To get location and other information about various venues in Toronto I'm using Foursquare's explore API. Using the Foursquare's explore API (which gives venues recommendations), I'm fetching details about the venues up

present in Toronto and collected their names, categories and locations (latitude and longitude). From Foursquare API (https://developer.foursquare.com/docs), I retrieved the following for each venue:

- Name: The name of the venue.
- Category: The category type as defined by the API.
- Latitude: The latitude value of the venue.
- Longitude: The longitude value of the venue.

2.2 Data Cleaning

A. Web Scraping Toronto Postal Codes and Neighbourhood

I web scraped the wikipedia link https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M using BeautifulSoup library and narrowed down the result data to seek to the postal code and neighbourhood content.

The drill down result table had 3 data elements:

- Postal Code
- Borough
- Neighbourhood

I read the data as html table and converted the data to Pandas dataframe with the Column names same as listed above.

	PostalCode	Borough	Neighborhood
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	Harbourfront

Then made below cleansing to the Dataframe:

- Dropped the rows that had Boroughs listed as Not Assigned as they are not going to help our analysis
- Grouped the Neighbourhoods belonging to the same Postal code and Borough and stuffed them in single row as comma separated instead of multiple rows

	PostalCode	Borough	Neighborhood
0	M1B	Scarborough	Rouge, Malvern
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

B. Appending Geographical Coordinates using Geospatial Data

Extracted the Geospatial Coordinates (.CSV file) data from the link http://cocl.us/Geospatial_data using Pandas library and loaded them in a Dataframe.

The CSV file and hence the Dataframe had the below data element columns:
- Postal Code
- Latitude
- Longitude

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Now we had both Neighbourhood dataframe as well as Coordinates dataframe w.r.t corresponding Postal Codes, so I inner joined and merged both the data frames and created a new dataframe with help of Pandas Merge method which

resulted in combined dataframe that will be helpful for easy reference and next step of analysis and exploration.

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	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

C. Web scraping Demographics of Toronto

I web scraped the wikipedia link https://en.wikipedia.org/wiki/Demographics_of_Toronto using BeautifulSoup library and narrowed down the result data to seek to the section that listed the Neighbourhoods and their top ethnic origins. Applied few data processing steps to the data and processed them to a Dataframe with 2 columns (Neighbourhood, Top Ethnic Origin).

	Neighborhood	Top_Ethnic_Origin
0	Milliken	Chinese
1	Steeles	Chinese
2	Agincourt North	Chinese
3	Malvern	Indian
4	Clairville	Indian
5	Jamestown	Indian
6	Malvern West	Chinese
7	Black Creek	Vietnamese
8	Rouge	Indian
9	Flemingdon Park	Indian
10	Thorncliffe Park	Pakistani

Our focus was on Indian origin locations - so narrowed down the dataframe for Indian origin rows and applied the below merging process :

- With help of Postal codes dataframe, I identified and merged the corresponding Postal codes and Boroughs corresponding to the neighbourhoods

- With help of Geospatial coordinates dataframe, I identified and merged the corresponding Latitude and Longitude coordinates

Both were processed using pandas merge methods.
Now the dataframe looked like:

```
Indian_neighborhood
```

]:

	Neighborhood	Top_Ethnic_Origin	PostalCode	Borough	Latitude	Longitude
0	Malvern	Indian	M1B	Scarborough	43.806686	-79.194353
1	Northwest	Indian	M9W	Etobicoke	43.706748	-79.594054
2	Jamestown	Indian	M9V	Etobicoke	43.739416	-79.588437
3	Rouge	Indian	M1B	Scarborough	43.806686	-79.194353
4	Flemingdon Park	Indian	M3C	North York	43.725900	-79.340923

There were 5 neighbourhoods where we could see to have harboured a good percentage of Indian ethnic origin and our exploration will be based on this filtered dataset.

D. Explore venues using Foursquare Location API

Foursquare is a social location service that allows users to explore the world around them. We will make use of the location APIs to fetch the venues surrounding the Toronto neighbourhoods and see to identify the optimal location for the restaurant.

In order to consume the location API, we need to possess a developer account whereby we will be given a Client ID, Client Secret codes which we need to use while invoking the APIs. Our focus here is on fetching the nearby venues of the specific neighbourhoods - so we will be using the below API and corresponding parameters.

```
# create the API request URL
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)

# make the GET request
results = requests.get(url).json()["response"]['groups'][0]['items']
```

Constituents of the API:

- Client ID - Your Foursquare developer account's Client ID
- Client Secret - Your Foursquare developer account's Client Secret
- Version - Foursquare API version
- Latitude and Longitude - Coordinates of the Neighbourhood
- Radius - Surrounding area threshold to be considered for Venue lookup
- Limit - Maximum number of Venues

Information fetched from the API:

The API returns as JSON data and we need to convert them into a dataframe. We need to extract/split/cleanse the Venue output such that we have the below as separate data elements per neighbourhood:

- Venue Name
- Venue Category
- Venue Latitude
- Venue Longitude

A new data frame is created by merging the Neighbourhoods with corresponding Venues data as 1 x N format as 1 neighbourhood can contain multiple venues.

The dataframe looks like :

```
Indian_neighborhood_venues
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Malvern	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Northwest	43.706748	-79.594054	Economy Rent A Car	43.708471	-79.589943	Rental Car Location
2	Northwest	43.706748	-79.594054	Saand Rexdale	43.705072	-79.598725	Drugstore
3	Jamestown	43.739416	-79.588437	Subway	43.742421	-79.589471	Sandwich Place
4	Jamestown	43.739416	-79.588437	Shoppers Drug Mart	43.741685	-79.584487	Pharmacy
5	Jamestown	43.739416	-79.588437	Popeyes Louisiana Kitchen	43.741202	-79.584545	Fried Chicken Joint
6	Jamestown	43.739416	-79.588437	The Beer Store	43.741694	-79.584373	Beer Store
7	Jamestown	43.739416	-79.588437	Sheriff's No Frills	43.741696	-79.584379	Grocery Store
8	Jamestown	43.739416	-79.588437	Pizza Pizza	43.741569	-79.584489	Pizza Place
9	Jamestown	43.739416	-79.588437	McDonald's	43.741757	-79.584230	Fast Food Restaurant
10	Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
11	Flemingdon Park	43.725900	-79.340923	Sorento Restaurant	43.726575	-79.341989	Italian Restaurant
12	Flemingdon Park	43.725900	-79.340923	Oomomo	43.726429	-79.343283	Discount Store
13	Flemingdon Park	43.725900	-79.340923	Fitness Connection	43.727473	-79.341707	Gym

3. Data Exploration and Visualization

3.1 Map view

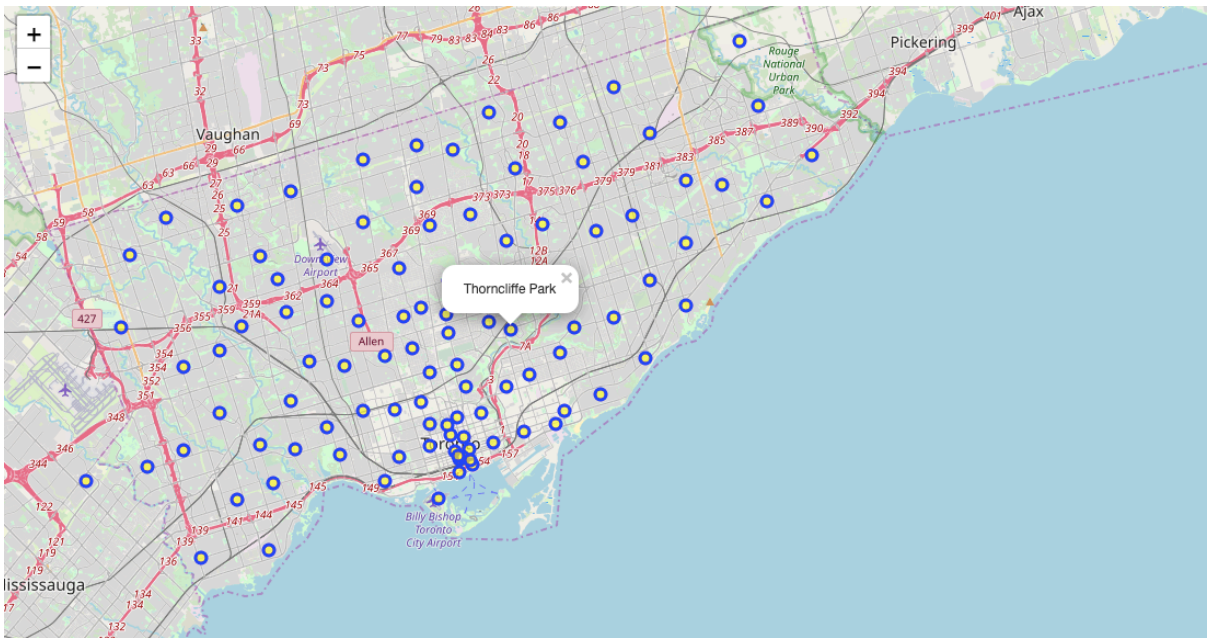
In order to get a visual experience of where these neighbourhoods are situated and how far they are spread across the Toronto location, I have made use of Folium python library to visualise the neighbourhoods in a Leaflet Map view.

Below is the Map view of Toronto Neighbourhoods:

Toronto Neighborhood map view

```
map_toronto = folium.Map(location=[Latitude, Longitude], zoom_start = 12)

for lat, lng, lbl in zip(toronto_postal_coord['Latitude'], toronto_postal_coord['Longitude'], toronto_postal_coord['Neighborhood']):
    label = folium.Popup(lbl, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius = 5,
        popup = label,
        fill = True,
        color = 'blue',
        fill_color = 'yellow',
        fill_opacity = 0.7,
        parse_html = False
    ).add_to(map_toronto)
map_toronto
```



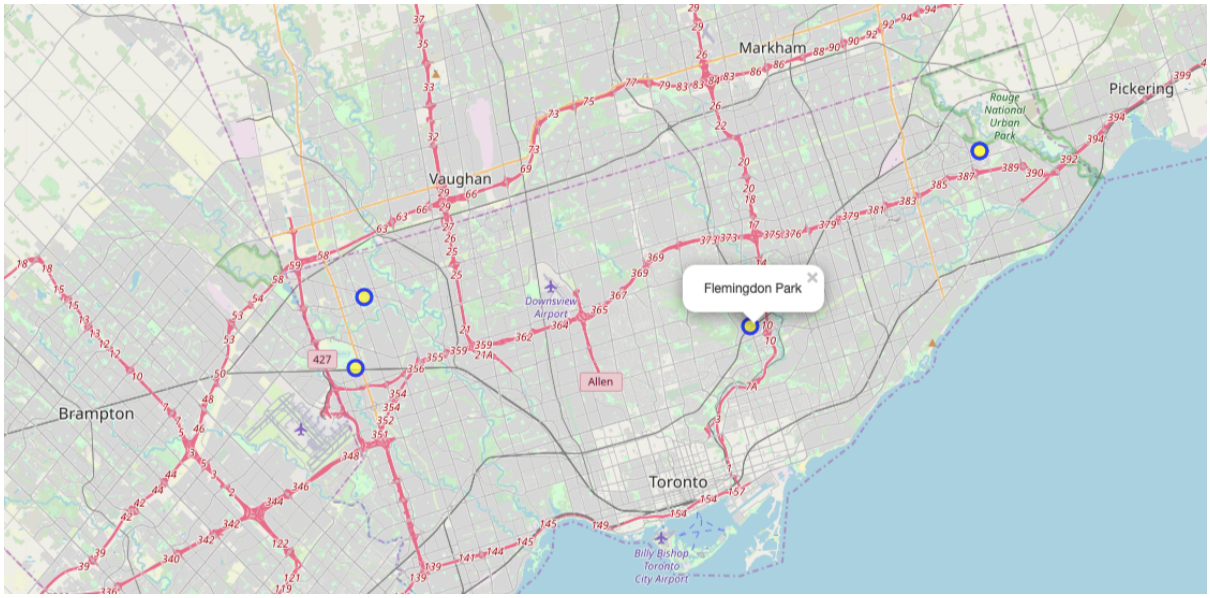
Have used Circle marker and label popup features to mark the locations and pop-out the location name as user clicks the circle marker.

Below is the Map view of Indian origin neighbourhood in Toronto location:

Toronto Indian Neighborhood map view

```
map_indian_toronto = folium.Map(location=[Latitude, Longitude], zoom_start = 11)

for lat, lng, lbl in zip(Indian_neighborhood['Latitude'], Indian_neighborhood['Longitude'], Indian_neighborhood['Neighborhood']):
    label = folium.Popup(lbl, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius = 7,
        popup = label,
        fill = True,
        color = 'Blue',
        fill_color = 'Yellow',
        fill_opacity = 0.7,
        parse_html = False
    ).add_to(map_indian_toronto)
map_indian_toronto
```



3.2 Neighbourhood Venues

I used the Foursquare Location API to fetch the venues of the 5 Indian Neighbourhood in Toronto namely Malvern, Northwest, Jamestown, Flemingdon Park and Rouge. I extracted the Venues data by setting max venues limit per neighbourhood as 20 and a Radius surrounding of 1000 Meters and did some splitting and cleansing to end up in having the Venue Name, Category, Venue Coordinates and appended the same against the corresponding Neighbourhoods as mentioned in detail in Data Cleaning section.

I was able to derive a group by stat dataframe where I grouped the Venues count by neighbourhood to see the spread of Venues per neighbourhood.

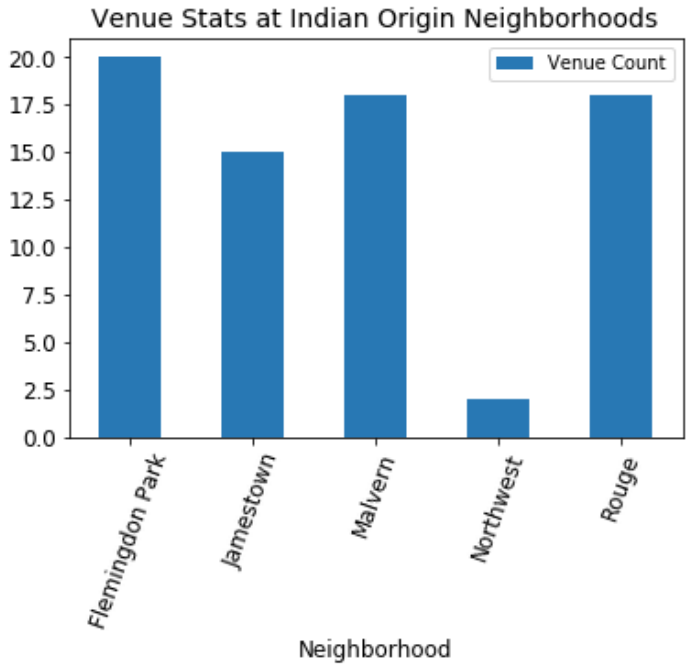
Below is the stat snapshot:

```
Indian_neighborhood_stat
```

:

	Neighborhood	Venue
0	Flemingdon Park	20
1	Jamestown	15
2	Malvern	18
3	Northwest	2
4	Rouge	18

Visualising the same as below:



We see that Flemingdon Park has predominantly quite a good number of venues within our radius threshold search followed by Malvern, Rouge and Jamestown while there are only very few venues in Northwest neighbourhood.

3.3 Restaurant Venues

In Earlier visualization since we observed the venue count stats in general, now I wanted to derive the restaurant stats per neighbourhood in order to see the spread of restaurants across these neighbourhoods. Additionally was also looking to know the spread of any Indian restaurants present at these locations. So filtered the dataframe in terms of Restaurant category and grouped them by Neighbourhood.

```
Indian_neighborhood_restaurant_venues
```

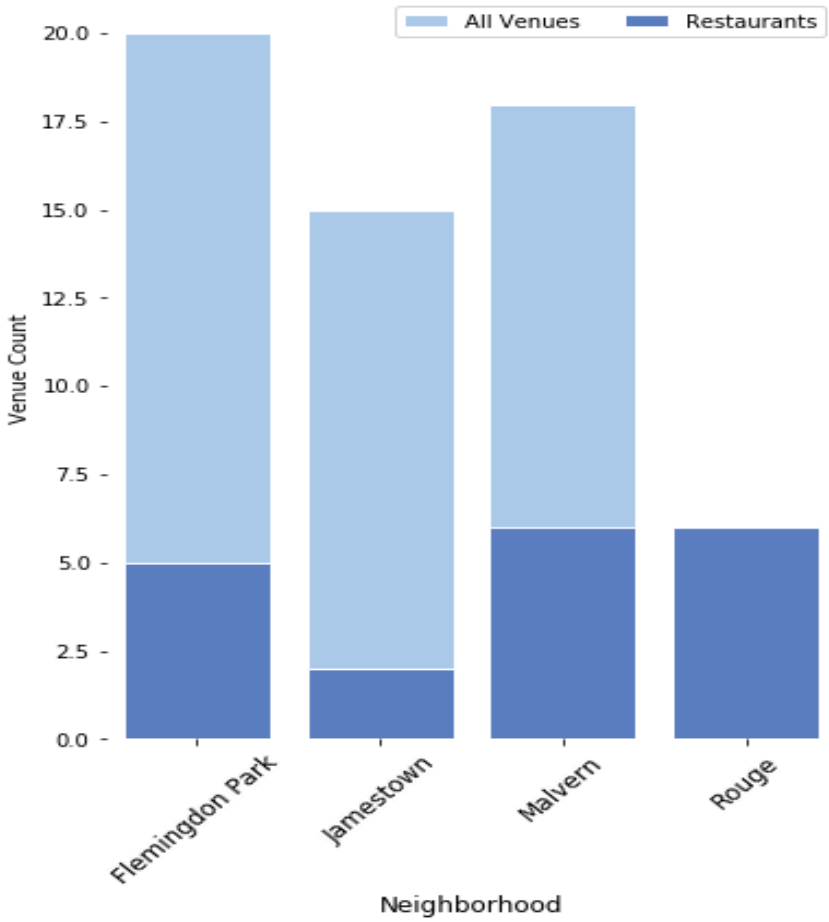
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Flemingdon Park	43.725900	-79.340923	Sorento Restaurant	43.726575	-79.341989	Italian Restaurant
6	Flemingdon Park	43.725900	-79.340923	Kaiseki Yu-Zen Hashimoto	43.720931	-79.337185	Japanese Restaurant
7	Flemingdon Park	43.725900	-79.340923	Swiss Chalet Rotisserie & Grill	43.726737	-79.341403	Restaurant
8	Flemingdon Park	43.725900	-79.340923	Diwan	43.725002	-79.332019	Middle Eastern Restaurant
9	Flemingdon Park	43.725900	-79.340923	JOEY Don Mills	43.733990	-79.344734	Restaurant
30	Jamestown	43.739416	-79.588437	McDonald's	43.741757	-79.584230	Fast Food Restaurant
32	Jamestown	43.739416	-79.588437	NORI SUSHI	43.742775	-79.586985	Japanese Restaurant
36	Malvern	43.806686	-79.194353	Caribbean Wave	43.798558	-79.195777	Caribbean Restaurant
37	Malvern	43.806686	-79.194353	Wendy's	43.802008	-79.198080	Fast Food Restaurant
38	Malvern	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
40	Malvern	43.806686	-79.194353	Harvey's	43.800106	-79.198258	Fast Food Restaurant
46	Malvern	43.806686	-79.194353	Mr Jerk	43.801262	-79.199758	African Restaurant
48	Malvern	43.806686	-79.194353	Charley's Exotic Cuisine	43.800982	-79.200233	Chinese Restaurant
56	Rouge	43.806686	-79.194353	Caribbean Wave	43.798558	-79.195777	Caribbean Restaurant
57	Rouge	43.806686	-79.194353	Wendy's	43.802008	-79.198080	Fast Food Restaurant
58	Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
60	Rouge	43.806686	-79.194353	Harvey's	43.800106	-79.198258	Fast Food Restaurant

I used Seaborn python library to visualize the spread - PFB snapshot:

Restaurant venues per Indian Neighborhood

```
import seaborn as sns

f, ax = plt.subplots(figsize = (6,8))
sns.set_color_codes('pastel')
sns.barplot(x = 'Neighborhood', y = 'Venue', data = Indian_neighborhood_stat,
            label = 'All Venues', color = 'b', edgecolor = 'w')
sns.set_color_codes('muted')
sns.barplot(x = 'Neighborhood', y = 'Venue', data = Indian_neighborhood_restaurants_stat,
            label = 'Restaurants', color = 'b', edgecolor = 'w')
ax.legend(ncol = 2, loc = 'upper right', fontsize=10)
plt.xticks(rotation=45, fontsize=12)
plt.yticks(fontsize=10)
plt.xlabel('Neighborhood',fontsize=12)
plt.ylabel('Venue Count', fontsize=10)
sns.despine(left = True, bottom = True)
plt.show()
```



We see that Malvern has better restaurant spread among its venues followed by Flemingdon Park. Jamestown has less than 20% of the constituents as restaurants. Rouge seems to possess only restaurants as its venues. This statement could change when we expand the radius threshold of venues exploration.

4. Predictive Modelling

4.1 Clustering Indian Neighbourhoods of Toronto

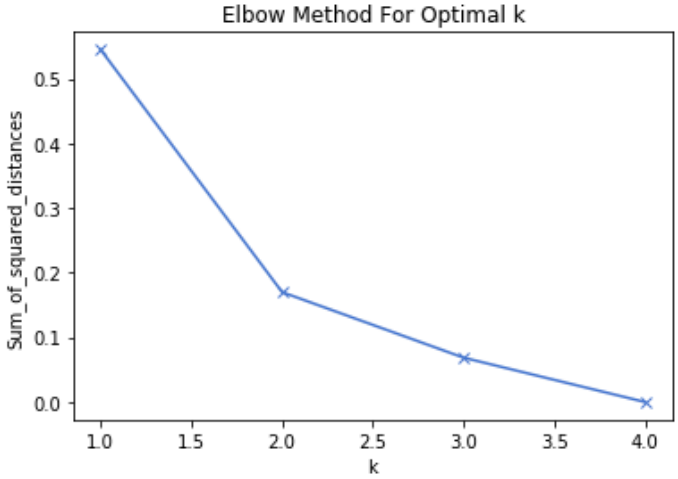
Now that we have done quite a handful of visual exploration analysis, we next proceed in clustering the neighbourhoods in order to derive our prediction. So I

found KMeans clustering algorithm to be suitable one to apply here. As a first step of algorithm, we need to derive the optimal k value to apply. This value can be determined by using the Elbow method as visualised below:

Finding Optimal k using Elbow visualizer and Applying KMeans Algorithm

```
Sum_of_squared_distances = []
K = range(1,5)
for k in K:
    km = KMeans(n_clusters=k)
    km = km.fit(Indian_neighborhood_grouped)
    Sum_of_squared_distances.append(km.inertia_)
```

```
plt.plot(K, Sum_of_squared_distances, 'bx-')
plt.xlabel('k')
plt.ylabel('Sum_of_squared_distances')
plt.title('Elbow Method For Optimal k')
plt.show()
```



Based on Elbow method we see k = 3 as the best optimal value to apply.

```
k = 3

model = KMeans(n_clusters=k, random_state = 0)
model.fit(Indian_neighborhood_grouped)
model.labels_
```

```
]: array([0, 0, 2, 1, 2], dtype=int32)
```

After clustering the Indian neighbourhoods dataframe, I merged the Cluster value against the corresponding neighbourhoods.

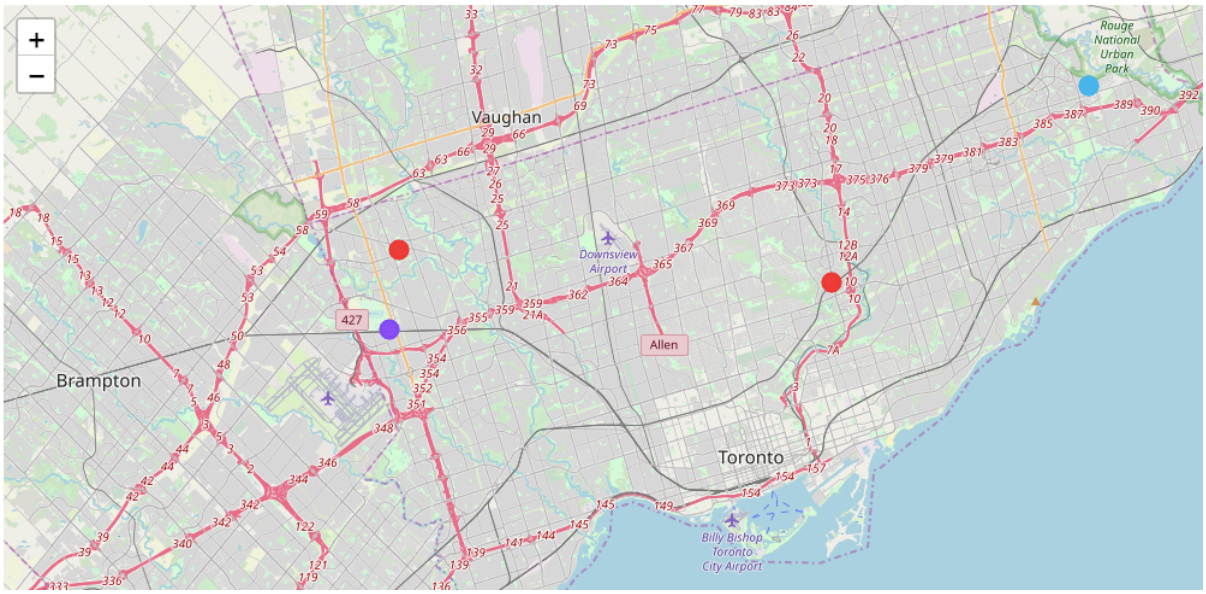
```
Indian_neighborhood = Indian_neighborhood.assign(Cluster = model.labels_)
```

```
Indian_neighborhood
```

]:

	Neighborhood	Top_Ethnic_Origin	PostalCode	Borough	Latitude	Longitude	Cluster
0	Flemingdon Park	Indian	M3C	North York	43.725900	-79.340923	0
1	Jamestown	Indian	M9V	Etobicoke	43.739416	-79.588437	0
2	Malvern	Indian	M1B	Scarborough	43.806686	-79.194353	2
3	Northwest	Indian	M9W	Etobicoke	43.706748	-79.594054	1
4	Rouge	Indian	M1B	Scarborough	43.806686	-79.194353	2

Visualizing the clusters in Map view,



Cluster 0 - Shown in Red colour in the map, comprising Flemingdon Park and Jamestown has quite a good number of venues attracting more customers.

Cluster 1 - Shown in Purple colour in the map, comprising Northwest doesn't seem to have much presence in terms of venues.

Cluster 2 - Shown in Aqua green colour in the map, comprising Malvern and Rouge shares very close proximity location points and as well has good presence of venues especially Rouge having more restaurants venue % compared to other neighbourhoods.

5. Results and Discussion

5.1 Results

Arriving at the end of our analysis where we have traversed right from identifying the Business problem of finding a favourable location to startup a good Indian restaurant and we went over the Neighbourhood analysis where where we analysed the Toronto neighbourhoods and narrowed down to Indian neighbourhoods based on Top ethnic origin percentage and then did few visualization analysis and Clustering to understand the data and derive the favourable location. We made use of Wikipedia data to set up a very realistic data-analysis scenario.

Based on the data exploration on Indian neighborhoods in Toronto, we can observe Flemingdon Park and Jamestown Cluster to be the best suitable locations to setup Indian restaurant.

Favourable points to justify the observation:

- Flemingdon Park and Jamestown has high percentage of Indian origin
- Being Indian Origin location, we don't see any Indian Restaurant venues at these locations, so the competition should be less
- Jamestown and Flemingdon Park Cluster has more venues which will attract more people gatherings and thereby will attract them to the restaurant as well. Example - Groceries, Concert hall, Clothing stores. So having Indian restaurant at close proximities near these venues should give more advantage of pulling more customers.

Malvern and Rouge cluster looks second best optimal location to setup Indian restaurant.

5.2 Discussion

On observing the analysis, in terms of drawbacks - I have done the clustering completely based only on data obtained from Foursquare API and the data about the Indian Ethnic origin neighbourhood which is also based on the 2016 census which is not up-to date. So there is a gap of 3 years in the population distribution data. Also the radius threshold and venues limits considered in Foursquare Api plays a pivotal role as well as the results might change if we consider a very broader limits of radius and venues counts. Even though there are lots of areas where it can be improved yet this analysis has certainly provided us with some good insights, preliminary information on possibilities & a head start into this business problem by setting the step stones properly.

6. Conclusion

Coming to the conclusion statement of this project analysis, I have made use of quite a many python libraries for Data ingestion, Cleansing, Preprocessing, Exploring and Visualizing the datasets. Further I have made use of Foursquare API to explore the venues in neighborhoods of Toronto, then got to use Web scraping methods with help of Beautiful soup python library and visualised using matplotlib and seaborn plots. I further did applied Clustering machine learning technique to predict the output and used Folium to visualize on a map.

In terms of areas of improvements I would like to convey that my analysis can be further improved with the help of more data and different machine learning technique. Further we can tweak this project to analyse and apply for other scenario such as opening a different cuisine restaurant or opening a book store etc., Overall this project analysis and findings gave me good confidence in my learning expertise. Thanks.

7. References

Project Notebook Code Github Link:
https://github.com/anand-sukumar/Coursera_Capstone/blob/master/Indian%20Restaurant%20at%20Toronto%20Recommendation.ipynb

List of neighborhoods in Toronto:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Demographics of Toronto:
https://en.wikipedia.org/wiki/Demographics_of_Toronto

Geographical Coordinates data: http://cocl.us/Geospatial_data

Foursquare Developer Documentation: https://developer.foursquare.com/docs