1. Discussion and Background of the Business Problem: Indian Restaurants

➢ Introduction Section:

This final project explores the best locations for Indian restaurants throughout the Queens of New York. New York is a major metropolitan area with more than 8.4 million (Quick Facts, 2018) people living within city limits. New York City is the largest city in the United States with a long history of international immigration. They came from many parts of the world. According to the 2007 American Community Survey estimates, New York City is home to approximately 315,000 people from the Indian subcontinent.

With its diverse culture, comes diverse food items. There are many restaurants in New York City, each belonging to different categories like Chinese, Indian, French, etc.

➢ Target Audience

Business personnel who wants to invest or open a restaurant.

* Freelancer who loves to have their own restaurant as a side business.

* Finding the best location for opening a restaurant.

* Budding Data Scientists, who want to implement some of the most used Exploratory Data Analysis techniques to

obtain necessary data, analyze it and, finally be able to tell a story out of it.

➢ Data Section

For this project we need the following data:

1. New York City data that contains Borough, Neighbourhoods along with there latitudes and longitudes

* Data Source: https://cocl.us/new_york_dataset

* Description: This data set contains the required information. And we will use this data set to explore various neighbourhoods of New York city.

2. Indian restaurants in Queens neighbourhood of New York city.

* Data Source: Foursquare API

* Description: By using this API we will get all the venues in Queens neighbourhood. We can filter these venues to get only Indian restaurants.

➢ Approach

Collect the New York city data from https://cocl.us/new_york_dataset

* Using Foursquare API we will get all venues for each neighbourhood.

* Filter out all venues which are Indian Restaurants.

* Data Visualization and some statistical analysis.

* Analysing using Clustering (Specially K-Means):

1. Find the best value of K

2. Visualize the neighbourhood with number of Indian Restaurants.

* Compare the Neighbourhoods to Find the Best Place for Starting up a Restaurant

* Inference From these Results and related Conclusions

➢ Problem Statement

What is the best location for an Indian restaurant in Queens, New York City?

In what Neighbourhood should I open an Indian restaurant to have the best chance of being successful?

2. Data Preparation:

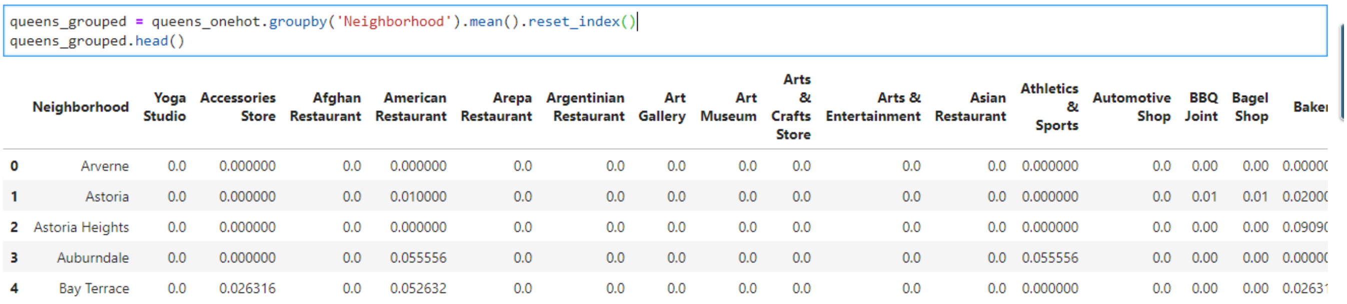
I will use New York City data for this project.

Using Foursquare Location Data:

Foursquare data is very comprehensive and it powers location data for Apple, Uber etc. For this business problem I have used, as a part of the assignment, the Foursquare API to retrieve information about the Venue, Venue category with there longitudes and latitudes. The call returns a JSON file and we need to turn that into a data-frame. Here I’ve chosen 100 popular spots for each neighbourhoods a radius of 500 meters. Below is the data-frame obtained from the JSON file that was returned by Foursquare —

3. Exploratory Data Analysis:

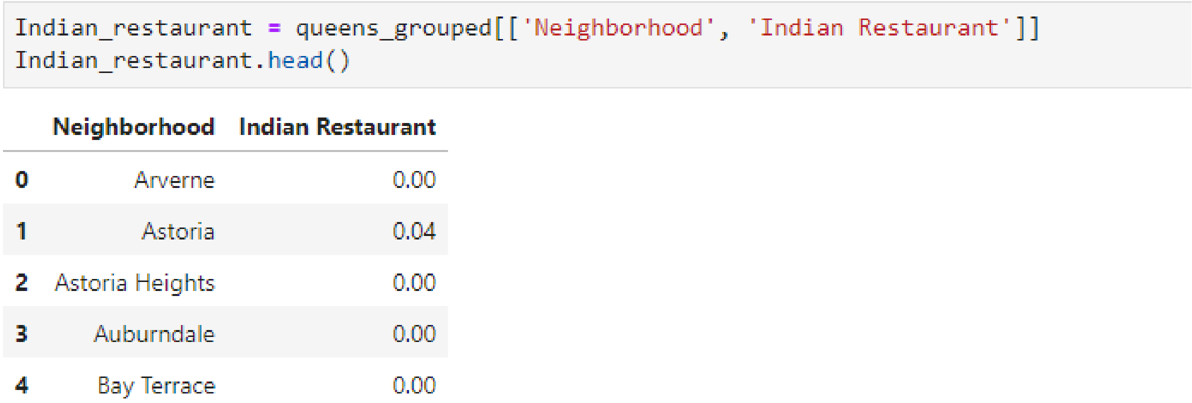
There are 271 unique categories in which Indian Restaurant is one of them. We will do one hot encoding for getting dummies of venue category. So that we will calculate mean of all venue groupby there neighbourhoods.



```
queens_grouped = queens_onehot.groupby('Neighborhood').mean().reset_index()
queens_grouped.head()
```

	Neighborhood	Yoga Studio	Accessories Store	Afghan Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant	Athletics & Sports	Automotive Shop	BBQ Joint	Bagel Shop	Bake
0	Arverne	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.00	0.00	0.00000
1	Astoria	0.0	0.000000	0.0	0.010000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.01	0.01	0.02000
2	Astoria Heights	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.00	0.00	0.09090
3	Auburndale	0.0	0.000000	0.0	0.055556	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.055556	0.0	0.00	0.00	0.00000
4	Bay Terrace	0.0	0.026316	0.0	0.052632	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.00	0.00	0.02631

After this we will extract only Neighborhood and Indian Restaurant column for further analysis:



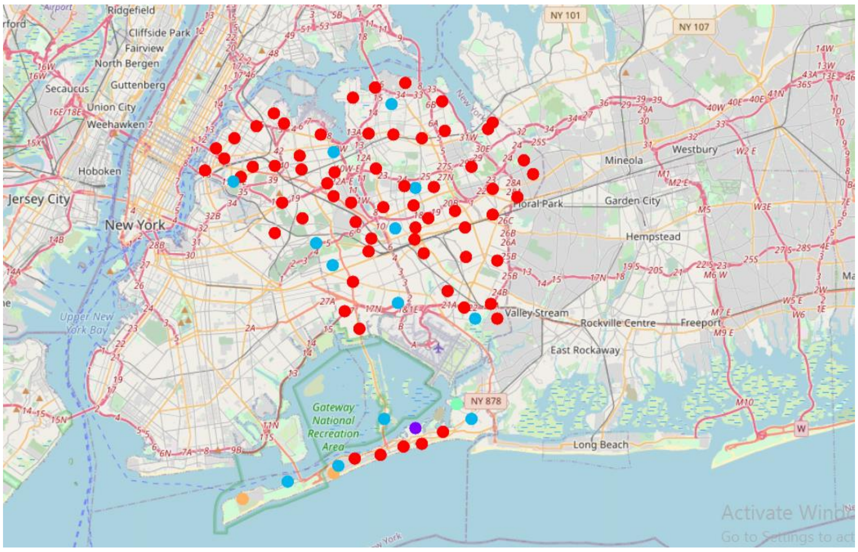
```
Indian_restaurant = queens_grouped[['Neighborhood', 'Indian Restaurant']]
Indian_restaurant.head()
```

	Neighborhood	Indian Restaurant
0	Arverne	0.00
1	Astoria	0.04
2	Astoria Heights	0.00
3	Auburndale	0.00
4	Bay Terrace	0.00

Let’s Examine the Clusters:

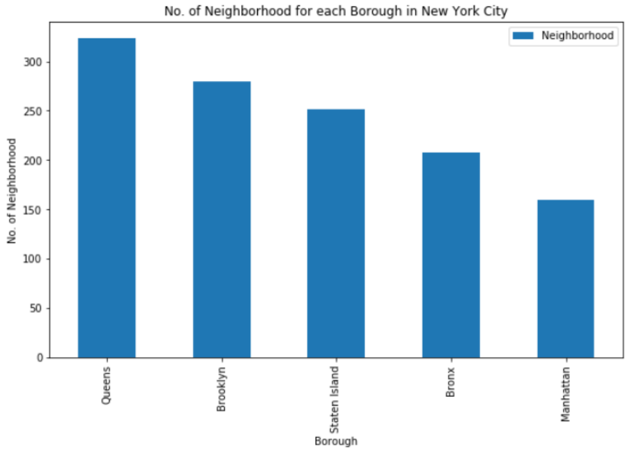
We can see these 4 clusters in the Map using Folium Library.

Here, we have 4 clusters 0,1,2 and 3 respectively. In cluster 0 we have neighborhoods which have least number of Indian Restaurants

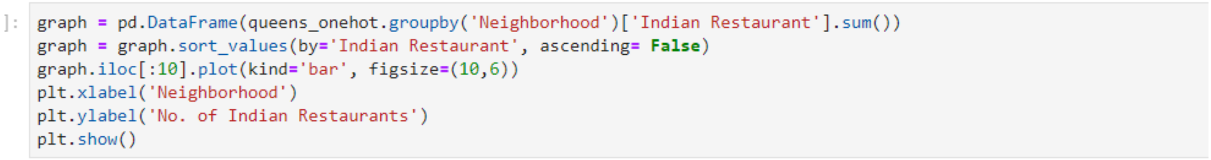


4. Visualization

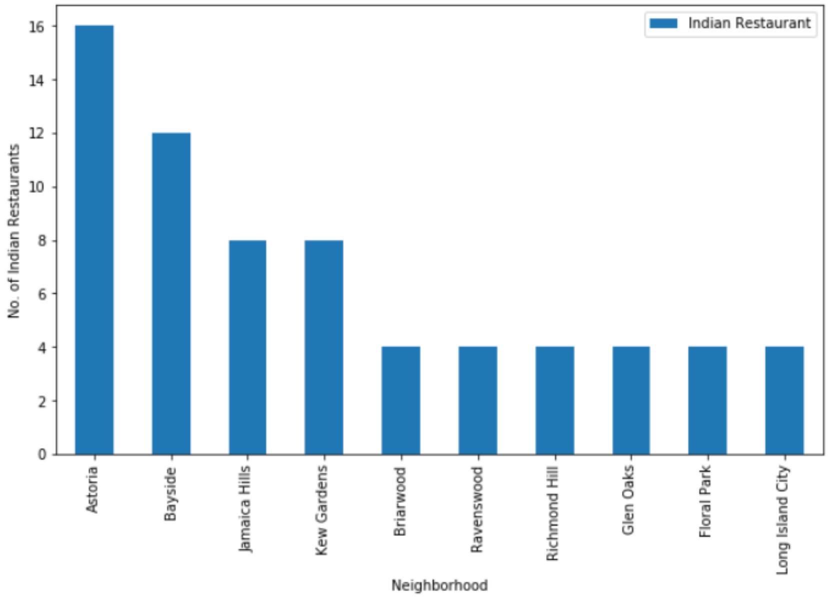
- There are 5 boroughs in the New York City in which Queens has highest number of neighborhoods.



- After that we will see which neighborhood has highest number of Indian restaurants.



```
graph = pd.DataFrame(queens_onehot.groupby('Neighborhood')['Indian Restaurant'].sum())
graph = graph.sort_values(by='Indian Restaurant', ascending= False)
graph.iloc[:10].plot(kind='bar', figsize=(10,6))
plt.xlabel('Neighborhood')
plt.ylabel('No. of Indian Restaurants')
plt.show()
```



In the above image we see that Astoria has the highest number of Indian restaurants.

5. Result

The results of the exploratory data analysis and clustering is summarized below:

Astoria neighbourhood has the highest number of Indian restaurants.

Jamaica Estates neighbourhood has a high density of Indian restaurants.

Cluster 0 neighbourhoods have the least number of Indian restaurants.

I will open my restaurant in **South Ozone Park** neighbourhood because it is near International Airport. Because all immigrants will come to the nearest restaurant. So, the profit will be more.

6. Discussion

According to the analysis, **South Ozone Park** will provide the least competition for an upcoming Indian restaurant as the International Airport is close to this neighbourhood. So, all this is the best place for Indian immigrants for having lunch/dinner and the frequency of Indian restaurants is very low compared to other neighbourhoods.

Astoria has the highest number of Indian restaurant and **Jamaica Estates** is highly dense so, we will not open there.

Some drawbacks of analysis are: the clustering is completely based on the data provided by Foursquare API. Since land price, the distance of venues from the closest station, the number of potential customers, could all play a major role and thus, this analysis is definitely far from being conclusory. However, it definitely gives us some very important preliminary information on the possibilities of opening restaurants in the Queens borough of New York City.

Also, another pitfall of this analysis could be consideration of only one major borough of New York City, taking into account all the areas under the 5 major boroughs would give us an even more realistic picture. Furthermore, these results also could potentially vary if we use some other clustering techniques like DBSCAN.

7. Conclusion

Finally, to conclude this project, we have got a small glimpse fo how real-life Data science project looks like. I have used some frequently used python libraries to handle JSON file, plotting graphs, and other exploratory data analysis. Use Foursquare API to major boroughs of New York City and their neighbourhoods. Potential for this kind of analysis in a real-life business problem is discussed in great detail. Also, some of the drawbacks and chances for improvements to represent even more realistic pictures are mentioned. As a final note, all of the above analyses is depended on the adequacy and accuracy of Four Square data. A more comprehensive analysis and future work would need to incorporate data from other external databases.