IBM APPLIED DATA SCIENCE

CAPSTONE PROJECT

OPENING A NEW INDIAN RESTAURANT IN NEW YORK CITY, USA

By: Karpagameenakshi Gurumoorthy

MAY 2020 – JUNE 2020

1. Introduction and Background of the Business Problem

1.1 INTRODUCTION:

New York City (NYC) is the most popular, densely populated major city in the United States of America. It is composed of five boroughs namely Brooklyn, Queens, Manhattan, the Bronx, and Staten Island which were consolidated into a single city in the year 1898. Presently, NYC has become an America's city of immigration as it welcomed people from all over the world. As a result of immigration, NYC has become the hub of diverse culture. Consequently, the construction of several restaurants, shopping malls have increased. Taking account of the people population from India in NYC, this Capstone Project explores the best locations for the construction of Indian Cuisine Restaurants.

1.2 BUSINESS PROBLEM:

Commencing up a new restaurant is not a straightforward job. The initial and foremost step involved in opening a new restaurant is deciding the best location, as it will determine whether a restaurant will be prosperous in the business or it would be a failure. Thus, the main object of this Capstone Project is to examine and select the best locations in NYC, to begin a new Indian restaurant with the help of data science and the data obtained from Foursquare API.

TARGET AUDIENCE:

- The main objective of this project is to help the business person who wants to invest or open a new Indian Restaurant in New York City.
- The other objective is to find the perfect location for setting up the restaurant that helps the investor to have more income.

2. Data section

2.1 DATA SOURCES:

The data sources used for the Capstone Project are,

- In this project to determine the highest Indian Population in the United States of American, I utilized the dataset from the site WorldAtlas "https://www.worldatlas.com/articles/top-10-us-metropolitan-areas-with-the-highest-population-of-indians.html". From the dataset, we can conclude that NYC stands first with highest number of Indian population.

- Along with the population dataset, I am using the site "https://www.worldatlas.com/articles/the-world-s-most-popular-tourist-attractions.html" to obtain the most popular tourist attraction spots in New York City. Thus utilizing the

above two dataset, I chose Times Square in New York City as a target destination for opening the restaurant.

- Foursquare API, a location data provider is utilized for exploring the Indian Restaurants present around Times Square. Also, the same dataset is used for obtaining the data about the Indian Restaurants present in a particular postal code or zip code.

- With the above datasets, depending upon the total amount of restaurants in a particular postal code, a perfect location for setting up the Indian Restaurant around Times Square can be determined. To know the neighborhoods in which a new Indian Restaurant is about to open, I am using the dataset from the site "https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm".

So, from the above data sources we can decide the City, Neighborhood in which an Indian Restaurant could be opened with high profit.

2.2 DATA FILTERING:

The dataset obtained from the site "https://www.worldatlas.com/articles/top-10-us-metropolitan-areas-with-the-highest-population-of-indians.html" consisted of columns such as Rank, City, and Indian Population as shown in Figure.1

Rank	City	Indian Population
1	New York	526,133
2	Chicago	171,901
3	Washington, DC	127,963
4	Los Angeles	119,901
5	San Francisco	119,854
6	San Jose	117,711
7	Dallas	100,386
8	Houston	91,637
9	Philadelphia	90,286
10	Atlanta	78,980
11	Boston	62,598

Figure.1

To obtain the dataset as shown above, web scraping techniques which are offered by the python package Beautiful soup is utilized. Same technique is used to obtain the dataset from all the other sites mentioned in the data source heading. In addition to the web scraping techniques, inbuilt functions present in the Pandas Python library such as drop, sort_values were used to obtain the meaningful information from the dataset. For example in the Figure.1, Rank column can be dropped from the table using drop function. The result obtained is shown in Figure.1.1. Thus this constitutes data filtering.

	City	Indian Population
0	New York	526133
1	Chicago	171901
2	Washington, DC	127963
3	Los Angeles	119901
4	San Francisco	119854
5	San Jose	117711
6	Dallas	100386
7	Houston	91637
8	Philadelphia	90286
9	Atlanta	78980
10	Boston	62598

Figure.1.1 Dataset without the column 'Rank'

3. Methodology

In this capstone project, the dataset having the information about Indian Population along with Highest Tourist Attraction spot in USA dataset is used to determine the best location (i.e. City, Neighborhood) for setting up a new Indian Restaurant. First, Indian Population in USA dataset is obtained using the Beautiful soup python library. The obtained data from the site is converted into dataframe with the help of Pandas library. The resulted dataframe shows that the population of Indians is more in New York City when compared to other cities in USA as shown in Figure.2.

	City	Indian Population
0	New York	526133
1	Chicago	171901
2	Washington, DC	127963
3	Los Angeles	119901
4	San Francisco	119854
5	San Jose	117711
6	Dallas	100386
7	Houston	91637
8	Philadelphia	90286
9	Atlanta	78980
10	Boston	62598

Figure.2 Highest population of Indians in USA

Thus, the New York City is the good location to start an Indian Restaurant. The result can also be visualized in the form of Bar Chart using the artist layer present in the matplotlib library as shown in Figure.2.1.

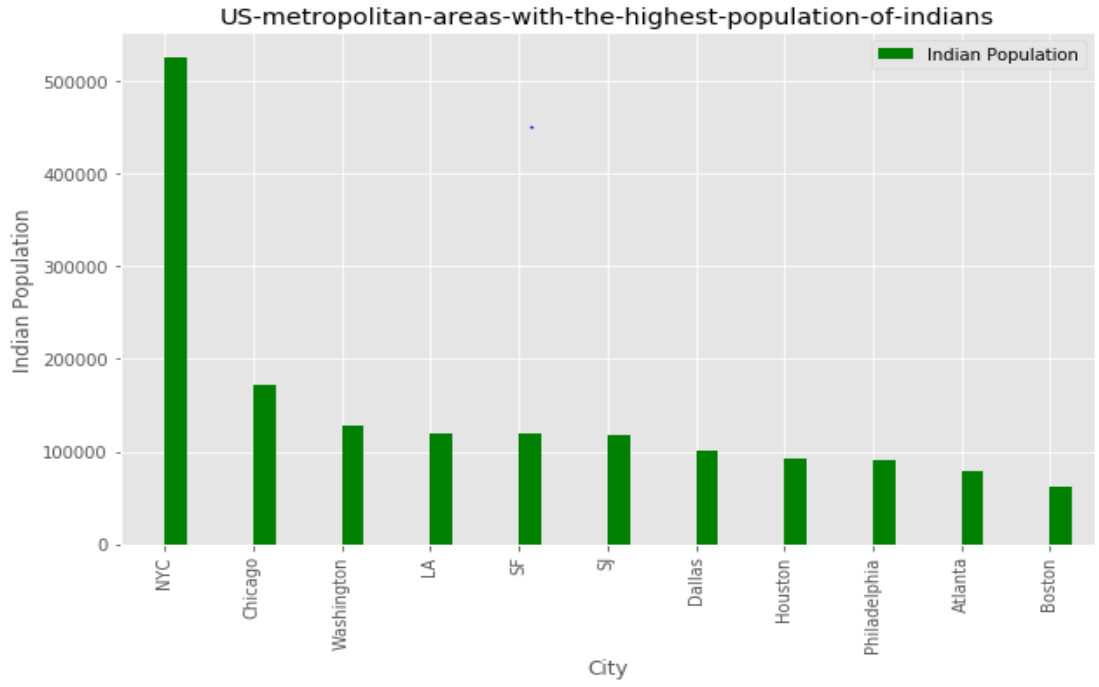

Figure.2.1 Bar chart of US metropolitan areas with the highest population of Indians

Secondly, the data from the site "https://www.worldatlas.com/articles/the-world-s-most-popular-tourist-attractions.html" is used for scraping the information about the popular tourist spot in New York City. Similarly the data is converted into dataframe using the Pandas data Library as shown in Figure.2.2

	Rank	Tourist Attraction	Location	No. of Annual Visitors
0	1	Las Vegas Strip	Las Vegas, US	39668221
1	2	Times Square	New York City, US	39200000
2	3	Central Park	New York City, US	37500000
3	4	Union Station	Washington D.C., USA	32850000
4	5	Niagara Falls	Niagara, US/Canada	22500000
5	6	Grand Central Terminal	New York City, US	21600000
6	7	Faneuil Hall Marketplace	Boston, US	18000000
7	8	Disneyworld's Magic Kingdom	Orlando, US	17536000
8	9	Disneyland Park	Anaheim, US	15963000

Figure.2.2 Popular tourist attraction in NYC

From the Indian population dataset, a conclusion has been made that NYC is a city in which the Indian Restaurant could be constructed. So, considering NYC alone, from Figure.2.2 Times Square is a best tourist attraction spot in NYC which is situated in Manhattan Borough. Hence, with 2 dataset obtained so far, an inference has been made that a new Indian Restaurant can be situated in Times Square, NYC.

3.1 FOURSQUARE API:

Though the city, borough in USA has been identified to start up a new restaurant, the exact neighborhood is yet to be identified. For that purpose, we need the data for the existing Indian Restaurants around Times Square. The details of Indian Restaurant around Times Square can be obtained using the location data provider Foursquare API. Using Foursquare API along with the geocoder, the latitude and longitude of Times Square is initially obtained. After receiving the latitude, longitude, I extracted the data of top 30 Indian Restaurants within the radius of 2000 meters. The resulting data was of a json file, so it is modified into pandas dataframe. Further, I filtered the dataset in such a way that only Indian Restaurants are displayed. The acquired data is shown in the Figure.2.3.

```
pc=dataframe_filtered[dataframe_filtered['categories']=='Indian Restaurant']
pc.head(25)
```

	name	categories	address	cc	city	country	crossStreet	distance	formattedAddress	labeledLatLngs	lat	lng	neighborhood	postalCode	state	id
0	Dakbar Fine Indian Cuisine	Indian Restaurant	152 E 46th St	US	New York	United States	Lexington	1103	[152 E 46th St (Lexington), New York, NY 10017...	[{'label': 'display', 'lat': 40.75372510094207...	40.753725	-73.973640	NaN	10017	NY	4abce9d5f964a5209e221e3
1	2 Dakbar Grill Fine Indian Cuisine	Indian Restaurant	157 E 55th St	US	New York	United States	btwn Lexington & 3rd Ave.	1444	[157 E 55th St (btwn Lexington & 3rd Ave.), Ne...	[{'label': 'display', 'lat': 40.75912154324992...	40.759122	-73.968898	NaN	10022	NY	4bbf143cba9776b09439fb8
2	Basmati Indian Cuisine	Indian Restaurant	764 9th Ave	US	New York	United States	NaN	788	[764 9th Ave, New York, NY 10019, United States]	[{'label': 'display', 'lat': 40.764156, 'lng'...	40.764156	-73.986087	NaN	10019	NY	52012efe498efd4438e30a06
3	Yatra Indian Cuisine	Indian Restaurant	153 E 33rd St	US	New York	United States	NaN	1395	[153 E 33rd St, New York, NY 10016, United Sta...	[{'label': 'display', 'lat': 40.745594, 'lng'...	40.745594	-73.979878	Rose Hill	10016	NY	4f32b64019638c91c7f24707
4	Baluchi's Indian Cuisine	Indian Restaurant	224 E 53rd St	US	New York	United States	NaN	1461	[224 E 53rd St, New York, NY 10022, United Sta...	[{'label': 'display', 'lat': 40.75707199999999...	40.757072	-73.968537	NaN	10022	NY	4f439e3819634bc91f57e8eb
5	Jaipur's Indian Cuisine	Indian Restaurant	1007 2nd Ave	US	New York	United States	NaN	1584	[1007 2nd Ave, New York, NY 10022, United States]	[{'label': 'display', 'lat': 40.756868, 'lng'...	40.756868	-73.967077	NaN	10022	NY	4e4ce1dcbd413c4cc66ced81
7	Chola Eclectic Indian Cuisine	Indian Restaurant	232 E 58th St	US	New York	United States	btw 2nd & 3rd Ave	1726	[232 E 58th St (btw 2nd & 3rd Ave), New York, ...	[{'label': 'display', 'lat': 40.76033870460651...	40.760339	-73.965785	NaN	10022	NY	4a63bfb4f964a520b3c51fe3

Figure.2.3. Filtered Dataframe

3.2 DATA VISUALIZATION:

The Dataframe which consist of the information about the top 30 Indian Restaurants around Times Square can be visualized with the help of Folium library. The result obtained is shown in the Figure.2.4.

Figure.2.4. Map of Indian Restaurants around Times Square

Lastly, the number of restaurants present in the particular postal code is determined using value_counts function. The value_counts produces the result as shown in Figure.2.5.

```
Count=pc['postalCode'].value_counts()
Count
```

```
]: 10022    4
   10019    2
   10036    2
   10018    2
   10016    1
   10017    1
   Name: postalCode, dtype: int64
```

Figure.2.5 No. of Indian Restaurants present in a particular Zip code

The above data can be easily visualized in the form of Bar Chart using matplotlib library and it is shown in Figure.2.6.

Figure.2.6.Bar chart of Indian Restaurants present in a particular area

4. Results

From the bar chart shown in Figure.2.6, a new Indian Restaurant can be opened in the neighborhood of Manhattan having the Postal codes or Zip code as 10017,10016. This is because, in those particular postal codes the number of Indian Restaurants are less when compared to the Restaurants in other postal codes. But, the neighborhood having the postal codes as 10017, 10016 are yet to be identified. For the identification purpose we are using the dataset from site"https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm". The Dataframe obtained after filtering the according to the Borough of NYC can be seen in Figure.3.

```
df3=df2[df2['Borough']=='Manhattan']
df3
```

	Borough	Neighborhood	ZIP Codes
18	Manhattan	Central Harlem	10026, 10027, 10030, 10037, 10039
19	Manhattan	Chelsea and Clinton	10001, 10011, 10018, 10019, 10020, 10036
20	Manhattan	East Harlem	10029, 10035
21	Manhattan	Gramercy Park and Murray Hill	10010, 10016, 10017, 10022
22	Manhattan	Greenwich Village and Soho	10012, 10013, 10014
23	Manhattan	Lower Manhattan	10004, 10005, 10006, 10007, 10038, 10280
24	Manhattan	Lower East Side	10002, 10003, 10009
25	Manhattan	Upper East Side	10021, 10028, 10044, 10065, 10075, 10128
26	Manhattan	Upper West Side	10023, 10024, 10025
27	Manhattan	Inwood and Washington Heights	10031, 10032, 10033, 10034, 10040

Figure.3. Dataframe of Manhattan Borough along with ZIP codes

Thus, by comparing the dataframe as shown in Figure.3 with that of the Postal Code obtained in the Figure.2.5, the perfect location to start a new Indian Restaurant around Times Square is in the Gramercy Park and Murray Hill neighborhood of Manhattan Borough.

5. Discussion and Recommendations

Though I have found out the best location for starting up a new Indian Restaurants with the help of two dimensional data such as highest number of Indian Population, Top tourist attraction spots, I am totally relying on the dataset extracted with the Foursquare API. In addition to that, I only used the

postal codes containing least number of Indian restaurants for setting up a new restaurant. So, we can't always rely on the same API as data may not be always correct. Thus, future research can be done by extracting data from different API, also by using different sets of data other than Indian Population, Top Tourist Attraction spots.

6. Conclusion

Finally, to conclude this project, I used Data Science methodology to come up with the solutions needed for the investors who are likely to start up a new Indian Restaurant in NYC only by considering the data like Indian Population and popular spots. Furthermore, the above data analyses mainly depend on the accuracy of foursquare data. So, the future research can be made from the data extracted from the other databases as it could be useful for obtaining the information with high accuracy.