Where to open an Indian Restaurant in New York City

IBM Data Science Capstone Project

TABLE OF CONTENTS

1. INTRODUCTION

As hundreds of thousands of Indian people are living in New York, so there is a good chance of popularity of Indian foods. So there are huge Indian restaurants all around New York. In this capstone project we would like to know where these restaurants are located and how a new opportunity can be unlocked to open a new restaurant where there is low competition but high demand, specially where there are high Indian population with low number of Indian restaurants.

It may introduce a decent open door for an Indian American previously living in NYC and are knowledgeable with the Spots and the Areas. As Indian food is very popular with Americans along with Indian Americans, there are as of now numerous cafés the vast majority of which are an Establishment or a family possessed business.

The New York City district is home to the biggest Indian American populace among metropolitan regions by a huge edge and speaks to the second-biggest metropolitan Asian public diaspora both outside of Asia and inside the New York City metropolitan territory.

Feel, menu, cleanliness and obviously taste are extremely significant components to be remembered prior to getting into the Neighborliness Business yet these are generally issues that can be handled inside by the person(s) in control. The area of a café is likewise of most extreme significance paying little heed to the historical backdrop of a business or the flavor of the food. On the off chance that individuals don't come in to eat, at that point none of the arrangements matter. That is the difficulty I am handling in this undertaking.

2. PROBLEM STATEMENT

The goal is to find out a reasonable location(s) to open an Indian Café in New York City, USA. This undertaking utilizes different Information Science and AI techniques (k-means clustering) to give an Answer for the customer. The venture means to give an Answer for the Inquiry : 'Where would it be a good idea for you to think about opening an Indian Restaurant in New York City?'

3. DATA

3.1 Data

I have used the following Data for the completion of the project :
● List of Boroughs and Neighborhoods in NYC - This gives the coordinates of all the neighborhoods and is used to call the Foursquare API.
● List of Places and Venues in NYC - This contains data about all the nearby venues like
Restaurants, Bars, Gym etc.
● Demographics of American Indians in New York City - Vital to understand the distribution of the target audience in NYC.
● Latitude and Longitude Data of the neighborhood(s) - To plot and visualize our data.

3.2 Data Sources

● New York City Neighborhoods Data from NYU website [1].
● Nearby Venues Data created using Foursquare API [2].
● The Demographics Data is scraped from Wikipedia [3].
● Latitude and Longitude values are obtained using the Geocoder package in python.

4. METHODOLOGY

4.1 Boroughs

The information segment above plainly portrays that our NYC information comprises Borough and Neighborhoods in these Precincts. The information contains 5 Wards - Queens, Brooklyn, Bronx, Manhattan and Staten Island and more than 300 neighborhoods altogether. So before we start our examination of the Neighborhoods we select a proper Precinct. This includes investigating each of them 5. The information is separated for every Borough and is utilized to settle on the decision to the Foursquare API.

	Borough	Count
0	Queens	81
1	Brooklyn	70
2	Staten Island	63
3	Bronx	52
4	Manhattan	40

Fig: Count of Neighborhoods in each Borough

4.2 Foursquare API

The focal piece of this task includes utilizing the Foursquare API to get different

subtleties of close by scenes, as - the Category (Pizza Place, Monument and so on), The directions of the spot (in Latitude and Longitude) and the Name of the Venue. We need to proclaim our Foursquare credentials like the Client ID and Client Secret. We accept a span estimation of 500, which returns settings inside a sweep of a large portion of a kilometer. To forestall an excessive number of records being returned by the capacity call a restriction of 100 is set.

The url is built with our proclaimed credentials and a solicitation call is made to the API. The information returned is as a json payload. The pandas dataframe is then built by perusing portions of this information. Along these lines 5 dataframes are made - one for every Borough.

Now that the data has been structured for the preprocessing, we to decide on a Borough for the analysis and so we look into 2 aspects -
1. Pre-existing Indian Restaurants
2. Demographics of the Indian American Population

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	St. George	40.644982	-74.079353	A&S Pizzeria	40.643940	-74.077626	Pizza Place
1	St. George	40.644982	-74.079353	Beso	40.643306	-74.076508	Tapas Restaurant
2	St. George	40.644982	-74.079353	Staten Island September 11 Memorial	40.646767	-74.076510	Monument / Landmark
3	St. George	40.644982	-74.079353	Richmond County Bank Ballpark	40.645056	-74.076864	Baseball Stadium
4	St. George	40.644982	-74.079353	Shake Shack	40.643660	-74.075891	Burger Joint
5	St. George	40.644982	-74.079353	Ruddy & Dean	40.644074	-74.076683	Bar
6	St. George	40.644982	-74.079353	Enoteca Maria	40.641941	-74.077320	Italian Restaurant
7	St. George	40.644982	-74.079353	St. George Theatre	40.642253	-74.077496	Theater
8	St. George	40.644982	-74.079353	Marie's 2	40.642176	-74.076669	Italian Restaurant
9	St. George	40.644982	-74.079353	The Gavel Grill	40.642157	-74.076674	American Restaurant

Fig: Nearby venues at Staten Island

4.3 Pre-existing Indian Restaurants

Since we wish to open a new Indian Restaurant, it assists with investigating ones that are now present. So we get the tally of Indian Restaurants (from the Venue Category) in every Borough and consolidate them to get a thought of the dispersion or grouping of them. Consistently, to keep away from rivalry it would bode well to choose a Borough with few Indian Restaurants. It tends to be seen that Manhattan and Queens have the most number of Restaurants and Staten Island with the least.

	Borough	Indian Restaurant
0	Manhattan	24
1	Queens	18
2	Brooklyn	16
3	Bronx	3
4	Staten Island	1

Fig: Total number of restaurants in each Borough

4.4 Demographics of Indian Americans

An Indian Restaurant would basically oblige the Indian American populace and Indian

travelers. So we investigate the Indian American populace in NYC. The information for the equivalent was scratched from Wikipedia and is from a 2014 American Community Survey (that assembles registration information including identity). This causes us to limit our area for the objective populace. The crude information scratched contains some arranging and superfluous segments that should be cleaned before it tends to be utilized. When finished, it would appear that this -

	Rank	Borough	City	Indian Americans	Density of Indian Americans per square mile	Percentage of Indian Americans in municipality's population
0	1.0	Queens (2014)[32]	New York City	144896	1326.5	6.2
1	2.0	Brooklyn (2012)	New York City	25270	357.9	1.0
2	3.0	Manhattan (2012)	New York City	24359	1060.9	1.5
3	4.0	The Bronx (2012)	New York City	16748	398.6	1.2
4	5.0	Staten Island (2012)	New York City	6646	113.6	1.4

Fig: Demographic of Indians in each Borough at NY

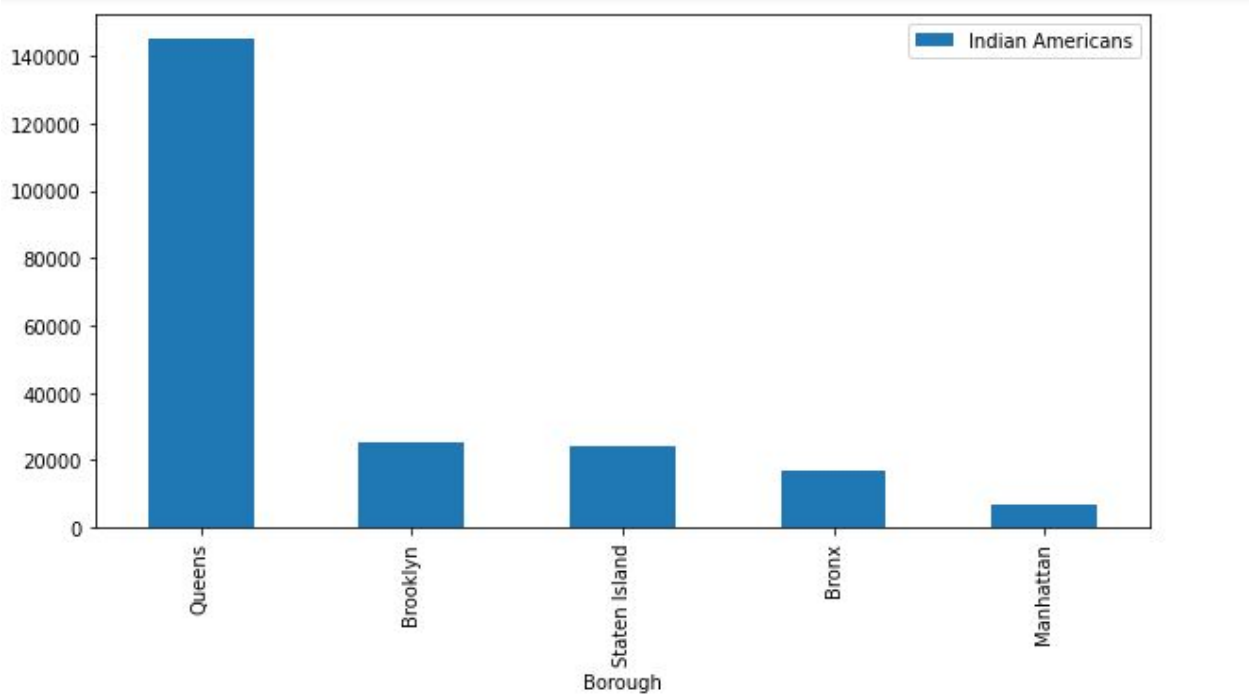

Fig: Indians Americans in each Borough, NY

4.5 Initial Analysis

Although Queens has the highest population of Indian Americans and the highest % population, we don't consider it as there are already numerous pre-existing restaurants.

Manhattan has very few Indian Americans with a low % and also has the most no. of Indian Restaurants, so we eliminate it.

Staten Island seems like a good first choice to begin our analysis as it does not have too many restaurants with a decent Indian Population.

	Borough	Indian Restaurant	Indian Americans	Density of Indian Americans per square mile	% Population
0	Queens	18	144896	1326.5	6.2
1	Brooklyn	16	25270	357.9	1.0
2	Staten Island	1	24359	1060.9	1.5
3	Bronx	3	16748	398.6	1.2
4	Manhattan	24	6646	113.6	1.4

Fig: # of Population and # of Indian Restaurants in each Borough, NY

4.6 Preprocessing

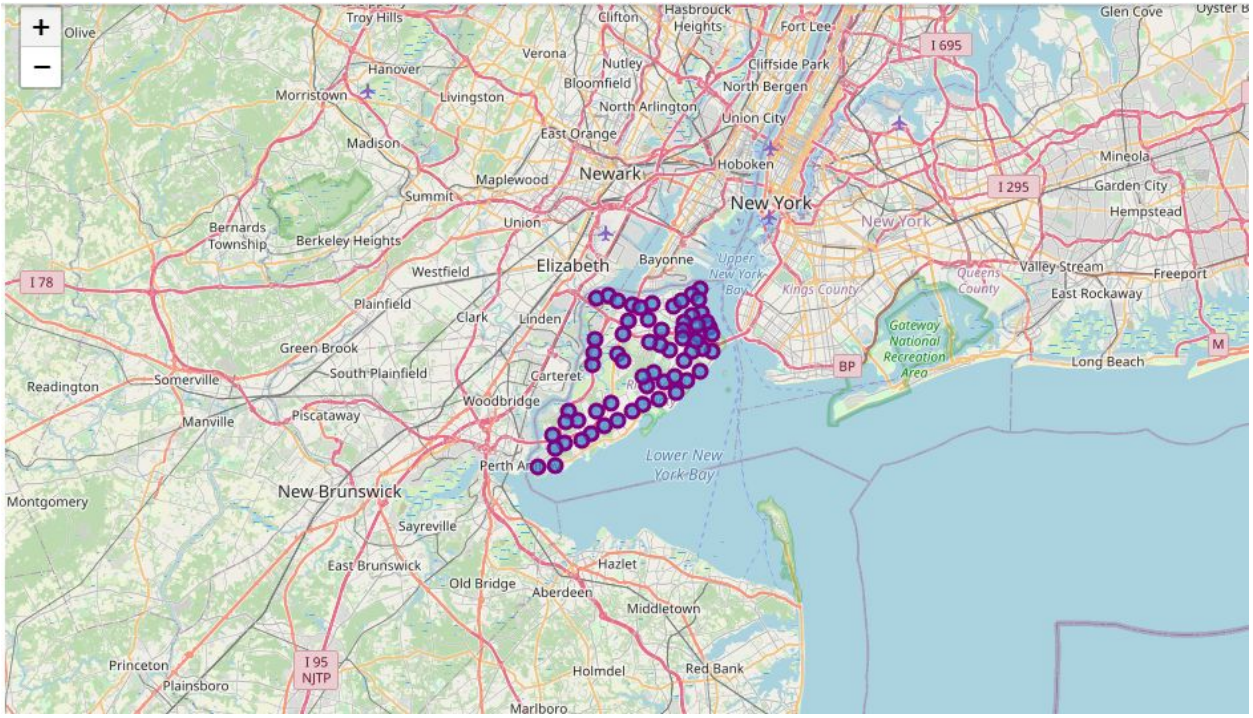

Fig: Map of Staten Island

4.6.1 One Hot Encoding

The data as mentioned above contains details of the nearby venues - Location, Category etc. This data needs to be transformed into a suitable format prior to Clustering.

One Hot Encoding is first performed on the ‘Venue Category’ attribute. This is done using the pandas get_dummies() function. Encoding assigns a Nominal Value to our Categorical data so the model does not interpret any numbers as importance or weight.

4.6.2 Grouping the Categories

The new dataframe is now grouped by Neighborhood and the mean for each Category is taken. This gives an average estimate for each Category in the neighborhood.

Once this is done, we then select only the Indian Restaurants and Neighborhoods as the other Attributes are not of concern to us. This dataframe is used to cluster the data points.

4.7 Clustering

4.7.1 Selecting K Value

The ‘k’ stands for - number of clusters. It’s value in k-means Clustering is selected by the “Elbow Method”. The Elbow Method involves plotting the Cost vs k-value; where k is an integer > 1. The point where the curve makes a transition is generally chosen as the k-value. I have used a k value of 3 for the Analysis, although there was a transition at k=2, the cost decreased further at k=3 and this would give us more diverse Clusters to examine. The aim is to minimize the Within-Cluster-Sum-of-Squares - Cost by using the Inertia criteria in the sklearn library.

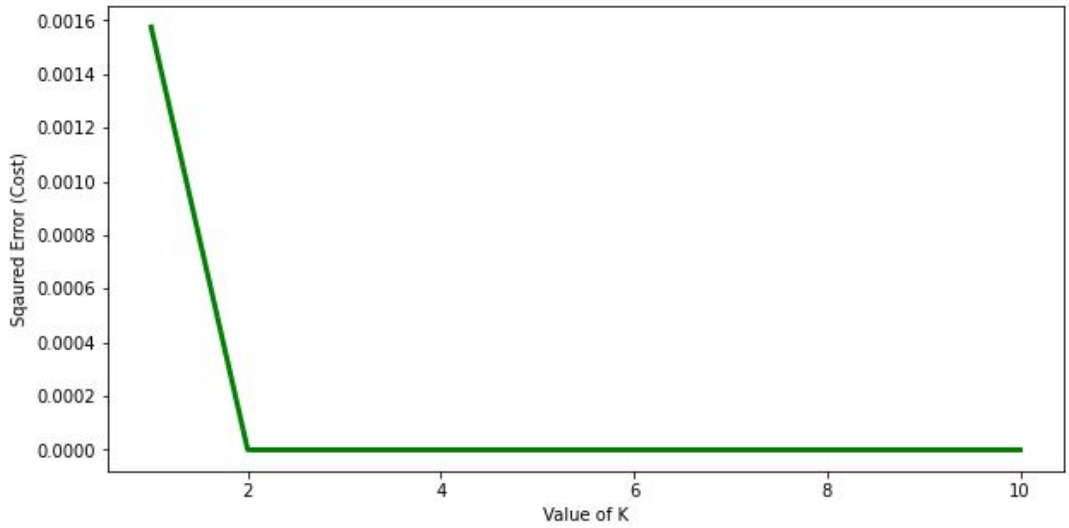

Fig: Elbow Plot of Cost Vs k

4.7.2 Cluster Labels

Next Clustering is performed and the Cluster Labels are saved. The Cluster Labels are merged with the previous dataframe containing only Indian Restaurants.

5. RESULTS

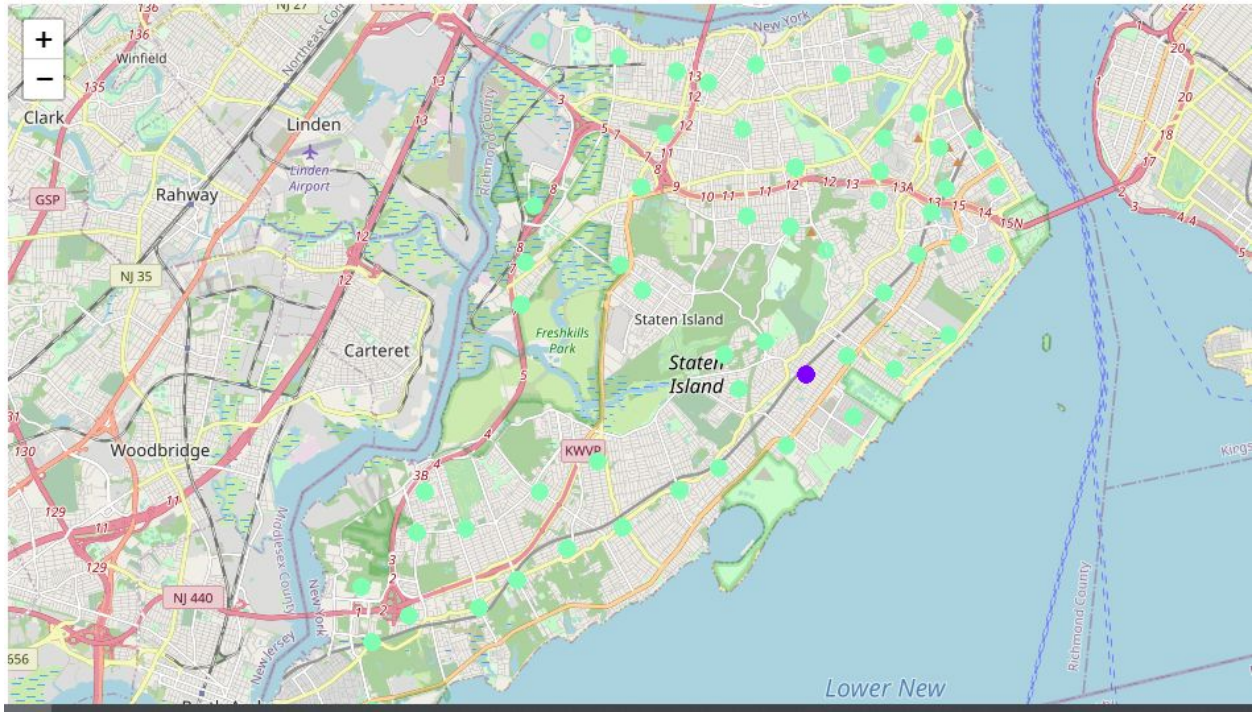

Fig: Plot of all the Clusters

- Cluster 1 has the one Indian Restaurants and is therefore not considered.
- Cluster 0 and 2 is ideal as no restaurants are present. Therefore we can look into the places in these Clusters.

	Neighborhood	Venue Category	Count
0	Eltingville	Sushi Restaurant	4
1	Old Town	Italian Restaurant	4
2	Clifton	Mexican Restaurant	3
3	Shore Acres	Italian Restaurant	3
4	Bulls Head	Chinese Restaurant	3
5	Dongan Hills	Italian Restaurant	3

Fig: Most common Neighborhoods in the Cluster

Looking at nearby venues, it seems Cluster 0 and 2 might be a good location as there are not a lot of Indian restaurants in these areas. There are 60 odd neighborhoods present in the Cluster and the most common ones being Old Town, Clifton, Shore Acres, Bulls Head, Eltingville, Dongan Hills.

Therefore our Indian Restaurant can be opened in any of these neighborhoods with little to no competition. Nonetheless, if the food is affordable, authentic and has good taste, I am confident that it will have a great following everywhere.

6. DISCUSSION

In light of the investigation Old Town, Clifton, Shore Acres, Bulls Head, Eltingville, Dongan Hills are a portion of the areas to think about opening our café. Since there is very low competition but high Indian population density, these places are best fit to open a new restaurant immediately. Manhattan has the most number of Indian Restaurants however the most un-number of Indian Americans, something that may be intriguing to investigate.

Another point is that there are other country type restaurants like Chinese restaurant, Italian restaurant, Mexican restaurant, but there is no Indian restaurant but there is high Indian Population over these locations. So this indicates that a new Indian restaurant can be established immediately to attract all these residents.

A portion of the disadvantages of this examination are — the grouping is totally founded distinctly on information acquired from Foursquare API and the information about the Indian populace dissemination in every area is likewise founded on the 2014 evaluation which isn't cutting-edge. Along these lines there is a colossal hole in the populace dissemination information. Despite the fact that there are heaps of regions where it very well may be improved at this point this investigation has unquestionably furnished us with some great experiences, starter data on conceivable outcomes and a head begin this business issue by setting the progression stones appropriately.

7. CONCLUSION

We have worked on a business problem like how a real data scientist would. We used python libraries to fetch the data (json, requests etc), to manipulate the contents (pandas) & to analyze and visualize(matplotlib, Folium) those datasets. We have made

use of the Foursquare API to explore the venues in neighborhoods of New York, then get data from Wikipedia which we scraped using the pandas library. We also applied machine learning techniques (Clustering) to predict the output given the data and used Folium to visualize it on a map.

8. CITATIONS

[1] https://geo.nyu.edu/catalog/nyu_2451_34572
[2] https://foursquare.com/developers/apps
[3] https://en.wikipedia.org/wiki/Indians_in_the_New_York_City_metropolitan_region