Battle of Neighborhoods: Mumbai

Analytical Approach To Indian Cuisine

Dhaval Modi

24 February 2021

1. Introduction

i. Background

Mumbai, a diverse and vibrant city of the South East Asia is located on the west coast of India. It is the place from where the western world learned about the miracles of spices. It is also termed as the financial capital of India and is the seventh most populous city in the world with close to 20 Million people calling it a home. Along with its population also comes cultural diversity as it represents almost every ethnicity of India. Hence, it would be very fitting to explore Mumbai for opening an "Indian Cuisine" restaurant. But based on the complexities of the city, selecting the best location for the investment is an uphill task in itself. Hence, this project aims to simplify it with the help of data analytics and Machine Learning.

ii. Problem

Mumbai being a vast and complex urban landscape, this project aims to select optimum neighborhoods for opening an "Indian Cuisine" Restaurant through application of Machine Learning on FourSquare data on popularity of venues and the rental values across neighborhoods in Mumbai.

iii. Interest

This project would be of interest to entrepreneurs and international/local restaurant chains who are looking to open a new restaurant or add to their existing chain of restaurants in Mumbai. The analysis presented here is for "Indian Cuisine" restaurants only, however it can be modified to analyze any type of restaurant / cuisine or venue category.

2. Data Acquisition & Cleaning

i. Data Sources

Neighborhood wise location data is not readily available for Mumbai, hence the approach taken was to get data of Post Offices located in the city as these are spread over the entire city and can be quite an accurate representation of neighborhoods.

Rent for 1 BHK apartment was taken as the representative value of rents in the areas. Rental data for neighborhoods, while available through different property management portals, discretion was needed to select the right match for Area (PO Name) vs Rent and hence, it was decided to manually compile the rent data from various property management portals. Sources of Data were:

- pincode.india-server.com
- magicbricks.com
- 99acres.com
- makaan.com
- foursqaure.com

ii. Data Cleaning

The primary data was scrapped using BeautifulSoup from "india-server.com" which provided a list of 239 Post Offices located in 89 Pin Codes. As mentioned above, we will use

Post Office (PO) names to represent Neighborhoods / Areas in Mumbai. The resultant DataFrame looked like:

	Post Office	Pin Code
1	A I Staff Colony	400029
2	Aareymilk Colony	400065
3	Agripada	400011
4	Airport	400099
5	Ambewadi	400004

Image 1: Neighborhoods of Mumbai with PIN codes

Next, Geocode function was used to obtain the latitude and longitude coordinates of the Post Offices. While the function returned values for most of the data points, coordinates were not available for 63 of the Post Office locations. This can be because of abbreviated name, locations being name of a street or simply due to incomplete data on geocode database. Upon grouping the Post Offices by Pin Codes it was found that all the Pin Codes were covered through the location coordinates available, hence it was decided to drop these 63 Post Offices / Neighborhoods which would not have any major impact on the end result. The resultant DataFrame included 176 Neighborhoods (PO Names) with their location coordinates and looked like:

	Post Office	Pin Code	Latitude	Longitude
0	Agripada	400011	18.9753	72.8249
1	Airport	400099	19.0901	72.8637
2	Ambewadi	400004	18.9907	72.8413
3	Andheri East	400069	19.1159	72.8542
4	Andheri	400053	19.1197	72.8464

Image 2: Neighborhoods with Coordinates

3. Methodology

To address our problem of selecting the optimum neighborhood for opening an Indian Cuisine Restaurant in Mumbai, the approach taken can be summarized in three steps:

i) Data Exploration:

In this step we explore FourSquare data on top ten most popular venues in each of the neighborhoods in Mumbai. Then we look at the distribution of rent across the neighborhoods.

ii) Clustering:

This is where the things get interesting and we build intelligence into our analysis by clustering the neighborhoods based on popular venues related to Indian Cuisine and their rent. We use K Means clustering algorithm for this approach.

iii) Arriving at Results:

This is where we get to see the fruit of our analysis; the selected neighborhoods are plotted on the map of Mumbai using folium library of python.

Let's look at each of the steps mentioned above in more detail.

i. Data Exploration:

We start off with the DataFrame of neighborhoods & the location coordinates. Upon exploring one of the neighborhoods "Agripada", for the availability popular venues within the vicinity of 1 Km, we find that FourSquare returns 30 venues. Hence, we take the same approach for all the neighborhoods and through coding in python we get the below DataFrame consisting of 4065 different venues from 220 venue categories across neighborhoods of Mumbai.

Area	Area Latitude	Area Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Agripada	30	30	30	30	30	30
Airport	24	24	24	24	24	24
Ambewadi	18	18	18	18	18	18

Image 3: Number of Venue Categories per Neighborhood

(4065, 7)

	Area	Area Latitude	Area Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Agripada	18.975302	72.824898	Celejor	18.975844	72.823679	Bakery
1	Agripada	18.975302	72.824898	Tote On The Turf	18.980266	72.820294	Nightclub
2	Agripada	18.975302	72.824898	Mahalaxmi Race Course (Royal Western India Tur...	18.980535	72.818588	Club House
3	Agripada	18.975302	72.824898	cafe coffee day	18.976988	72.824051	Coffee Shop
4	Agripada	18.975302	72.824898	Neel	18.980407	72.820403	Indian Restaurant

Image 4: Neighborhoods with Popular Venues

Next, upon applying code in python we transform the DataFrame to get top 10 venue categories in each of the Neighborhoods.

	Area	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Agripada	Indian Restaurant	Fast Food Restaurant	Bakery	Department Store	Coffee Shop	Gym	Pizza Place	Diner	Nightclub	Racetrack
1	Airport	Coffee Shop	Airport	Café	Bar	Donut Shop	Fast Food Restaurant	Lounge	Cocktail Bar	Cupcake Shop	Indian Restaurant
2	Ambewadi	Train Station	Plaza	Hotel	Bar	Bakery	Chinese Restaurant	Lounge	Luggage Store	Maharashtrian Restaurant	Restaurant
3	Andheri	Indian Restaurant	Sandwich Place	Vegetarian / Vegan Restaurant	Fast Food Restaurant	Hotel	Chinese Restaurant	Shoe Store	Food Court	Falafel Restaurant	Electronics Store

Image 5: Top 10 Venue Categories per Neighborhood

Now, let's look at the distribution of rent for 1 BHK apartment across Neighborhoods of Mumbai. For visualization, a bar graph was plotted using "plot" method in python with Neighborhoods (Area) on X axis and the Rent values on Y axis as shown below.

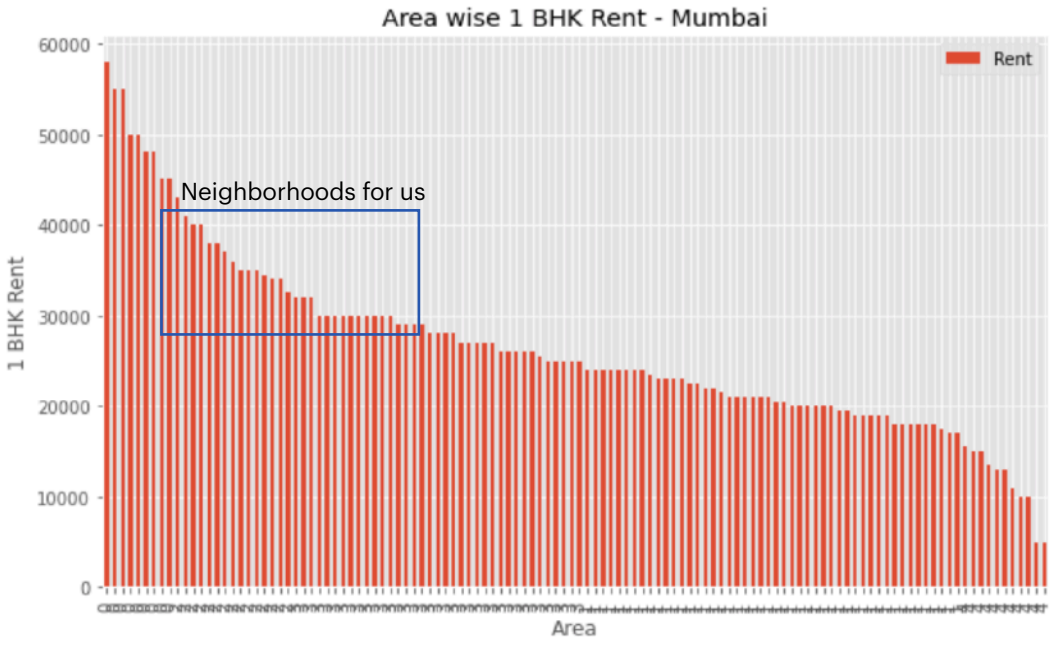

Image 6: Distribution of Rent in Neighborhoods

We see that rents ranged from a minimum of INR 6K to a maximum of 58K. However, for our problem, we are more interested in opening the restaurant in an "upper middle class neighborhood" where we expect the best return on our investment. These would be neighborhoods with rents fall in the range of INR 30K to 40K as shown in the graph.

Now, let's move on to the second step.

ii. Clustering:

We have two objectives here:

a. To identify similar neighborhoods based on popularity of venues related to "Indian Cuisine."
b. To identify similar neighborhoods based on rent values.

To achieve these objectives, we use K Means clustering algorithm as it helps us to find the optimum number of clusters and is relatively easy to handle.

Firstly, we use One-Hot encoding to prepare the DataFrame for clustering and group the data by Neighborhood based on the mean of frequency of venue categories for each Neighborhood. We get the resulting DataFrame as below:

Area	ATM	Airport	Airport Food Court	Airport Lounge	Airport Service	American Restaurant	Antique Shop	Arcade	...	Turkish Restaurant	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	Waterfront	Whisky Bar	Win Ba
Agripada	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.000000	0.
Airport	0.0	0.125000	0.0	0.0	0.041667	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.041667	0.
Ambewadi	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.055556	0.0	0.0	0.000000	0.
Andheri	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.076923	0.0	0.0	0.000000	0.
Andheri East	0.0	0.033333	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.066667	0.0	0.0	0.000000	0.

Image 7: One Hot Encoding

Next, we refine the DataFrame to include only the Neighborhoods where venue categories related to “Indian Cuisine” feature as top 3 most popular venues. To achieve this, first we create a DataFrame to include the venue categories that we are interested in. Then, we compare this DataFrame with the DataFrame showing Neighborhood-wise top 3 venue categories using “isin” method in python and retain neighborhoods where “Indian Cuisine” related venue categories feature as top 3. This does require a few lines of code, however, we get the resulting DataFrame with 115 neighborhoods as below.

0	Bengali Restaurant
1	Chaat Place
2	Goan Restaurant
3	Indian Chinese Restaurant
4	Indian Restaurant
5	Maharashtrian Restaurant
6	Mughlai Restaurant
7	Multicuisine Indian Restaurant
8	North Indian Restaurant
9	Punjabi Restaurant
10	Rajasthani Restaurant
11	Seafood Restaurant
12	South Indian Restaurant
13	Tea Room
14	Vegetarian / Vegan Restaurant

Image 8: Venue categories of Interest

	Area	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	True Count
0	Agripada	Indian Restaurant	Fast Food Restaurant	Bakery	1
3	Andheri	Indian Restaurant	Sandwich Place	Vegetarian / Vegan Restaurant	2
4	Andheri East	Fast Food Restaurant	Indian Restaurant	Hotel	1
5	Andheri Railway Station	Indian Restaurant	Sandwich Place	Vegetarian / Vegan Restaurant	2
7	Anushakti Nagar	Food	Indian Restaurant	Fast Food Restaurant	1

Image 9: DataFrame with Neighborhoods where Indian Cuisine related Venue Categories Feature in Top 3

And lastly, we refine the one-hot encoded DataFrame to retain the 115 neighborhoods as arrived at above. The head of the DataFrame looks like:

6]:

	Area	Bengali Restaurant	Chaat Place	Goan Restaurant	Indian Chinese Restaurant	Indian Restaurant	Maharashtrian Restaurant	Mughlai Restaurant	Multicuisine Indian Restaurant	North Indian Restaurant	Punjabi Restaurant	Rajasthani Restaurant	Se Resta
0	Agripada	0.0	0.0	0.0	0.0	0.166667	0.0	0.0	0.0	0.0	0.0	0.0	
3	Andheri	0.0	0.0	0.0	0.0	0.153846	0.0	0.0	0.0	0.0	0.0	0.0	
4	Andheri East	0.0	0.0	0.0	0.0	0.166667	0.0	0.0	0.0	0.0	0.0	0.0	

Image 10: One Hot Encoded DataFrame for Clustering

Now we are ready for running the K Means clustering algorithm. Upon dropping the “Area” column from the above DataFrame, we run K Means for values of k (number of clusters) ranging from 1 to 10 and determine the optimum value of k based on the elbow plot.

The plot uses distortion (euclidean distance of k from cluster data points) on Y axis and the number of clusters on X axis.

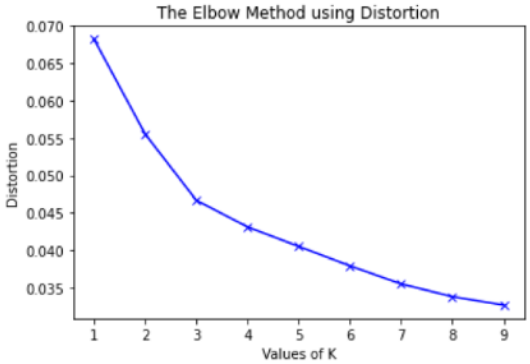

Image 11: Elbow Plot for Clustering based on Venue Categories

Based on the plot we select number of clusters as 4. Hence, we segment the selected Neighborhoods into four different clusters and plot them on the map of Mumbai using folium library of Python.

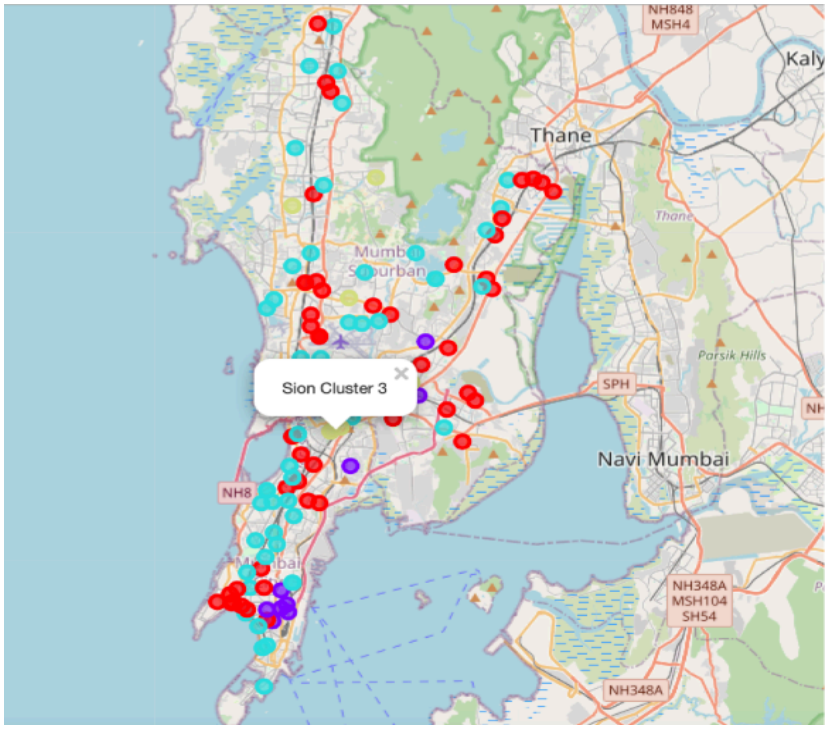

Image 12: Map of Mumbai with showing Clusters based on Venue Categories

Now let's focus on objective "b" which is identifying similar Neighborhoods based on their rent value. We again resort to the K Means algorithm but run it this time for the range of k values between 1 and 20 (we really want the best k possible). The DataFrame used and the elbow plot look like:

	Area	Rent
0	Agripada	28000
1	Andheri	30000
2	Andheri East	23000
3	Andheri Railway Station	13000
4	Anushakti Nagar	30000

Image 13: Neighborhoods with Rent Values

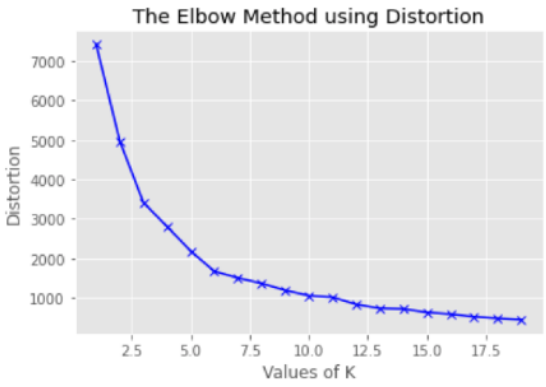

Image 14: Elbow Plot for Clustering based on Rent Values

Based on the plot above, we select number of clusters (k) as 5, hence segmenting the Neighborhoods into 5 different clusters based on their rent value.

However, we still haven't solved the problem of where actually should the restaurant be put up. For that, let's move on to the next section.

iii. Arriving at the results:

From the previous two sections we have:

a. A total of 115 Neighborhoods segmented into 4 clusters where "Indian Cuisine" related venue categories feature in top 3 most popular venues.
b. These same neighborhoods segmented into 5 clusters based on their rent.

Now, as mentioned in the "Data Exploration" section we are interested in "upper middle class" neighborhoods only with rents ranging from approximately INR 30K to 40K per month. To arrive at these neighborhoods, let's have a look at the median rent in each of the 5 clusters.

Cluster Labels_rent	Rent
0	50000
1	20750
2	36000
3	28000
4	13000

Image 15: Median Rent in each Cluster

From the DataFrame we see that, the Neighborhoods we would be interested in fall under the cluster 2. Hence, we select all the neighborhoods from cluster 2 and add other information to these neighborhoods from our earlier analysis. And finally, we a get a total of 14 neighborhoods that would suit our requirement and the resulting DataFrame is as below.

	Area	Rent	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	Latitude	Longitude
0	Bhawani Shankar	40000	0	Indian Restaurant	Coffee Shop	Fast Food Restaurant	19.015929	72.837460
1	Colaba	34500	2	Indian Restaurant	Diner	Hotel	18.915091	72.825969
2	Gokhale Road	43000	2	Lounge	Indian Restaurant	Bakery	19.008791	72.830730
3	Grant Road	40000	0	Indian Restaurant	Bakery	Snack Place	18.964447	72.813573
4	Haines Road	35000	2	Indian Restaurant	Hotel	Shopping Mall	18.989172	72.822429
5	Hanuman Road	36000	1	Indian Restaurant	Bakery	Fast Food Restaurant	18.948282	72.830243
6	Juhu	38000	2	Hotel	Indian Restaurant	Seafood Restaurant	19.107021	72.827528
8	Mahim	34000	0	Indian Restaurant	Fast Food Restaurant	Bakery	19.042314	72.839834
9	Mahim East	34000	0	Indian Restaurant	Fast Food Restaurant	Bakery	19.042314	72.839834
10	Matunga	35000	0	Indian Restaurant	Fast Food Restaurant	South Indian Restaurant	19.027436	72.850147
11	Powai Iit	32500	0	Indian Restaurant	Café	Coffee Shop	19.129055	72.918227
12	Santacruz(West)	38000	0	Indian Restaurant	Hotel	Coffee Shop	19.092336	72.853201
13	Shivaji Park	37000	2	Chinese Restaurant	Indian Restaurant	Movie Theater	19.027236	72.838348

Image 16: Selected Neighborhoods with Rent Value, Category Custer & Coordinates

Let’s also plot these neighborhoods along with their rent values on the map of Mumbai and we are ready to open our restaurant.

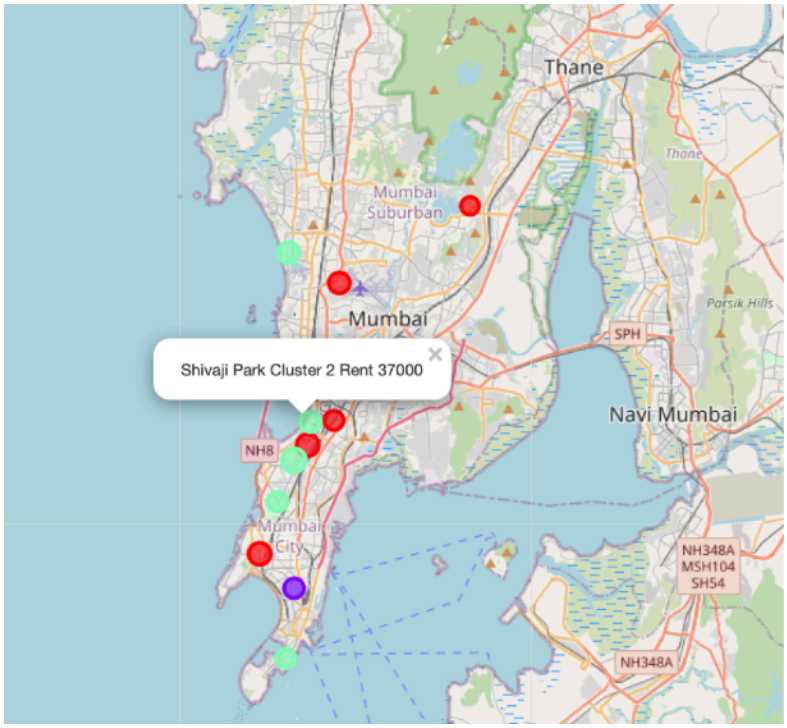

Image 17: Map of Mumbai with selected Neighborhoods

4. Conclusion:

With the use of unsupervised Machine Learning approach (K Means) in python we were able to analyze and segment different neighborhoods of Mumbai based on popularity of venue categories and their rent value. Through the analysis optimum neighborhoods were selected within specific rental budget for opening an "Indian Cuisine restaurant” in Mumbai.

This can be very useful for any new entrepreneur or a restaurant chain looking to enter into Mumbai market.

5. Discussion & Future Direction:

While I was able to solve the problem at hand, the analysis can be expanded and further refined. Firstly, there are other approaches (like hierarchical clustering & DBSCAN) available for clustering as against K Means used in the analysis which might influence the selection of Neighborhoods. Secondly, the analysis was conducted keeping in mind “Indian Cuisine” related venue categories only, however, it can be expanded to analyze any type of venue category and different rent thresholds. And lastly, based on the availability of data, the scope can also be increased to include that data on number of footfalls per venue category which could highly improve the selection of Neighborhoods.