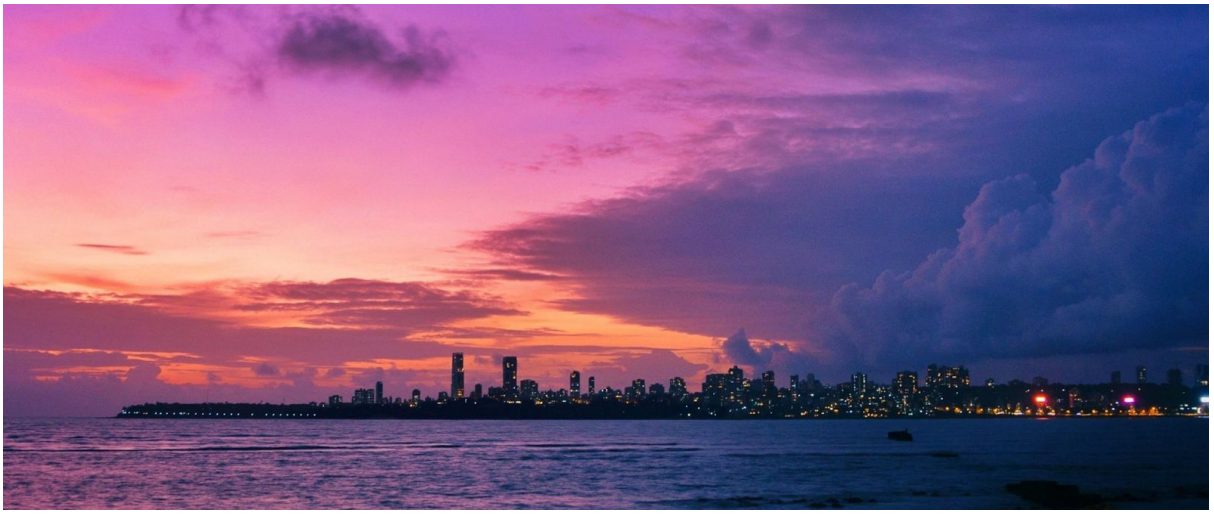
Coursera Capstone

IBM Applied Data Science Capstone

Let's Set Up a Restaurant in South Mumbai

Author: Joel Mathew Cherian

September 2020



If anything is good for pounding humility into you permanently, it's the restaurant business.
~ Anthony Bourdain

1. Introduction

Competition can be an important agent in deciding which type of restaurant should be set up in a particular locality. Multiple factors come into play such as pre-existing restaurants in the area, their ratings, location, prices, etc. Now, for anyone who wants to switch towards the hotel business and want to begin a food franchise in South Mumbai, India, data science can help us analyze this topic. The target of this project is to suggest types of food franchises that can be set up in South Mumbai based on the historic data about the existing restaurants. Using FourSquare API and Zomato API, it is possible to extract data of existing food businesses around the locality and form a conclusion about what can be the best type of restaurant that can be set up which has a higher probability of being profitable.

Based on the data collected from the two API about the existing business in that particular area we can then move towards cleaning and finding a possible correlation between the business, its opinion based on the rating, location, price, etc. Once curated, the user can be prompted with a suggestion regarding what type of restaurant is lacking in what part of South Mumbai and this can help the user come closer towards making a decision.

1.1. Business Problem

The objective of this capstone project is to analyse and select the best spot to set up a restaurant in the South part of Mumbai. South Mumbai is classified to be one of the most expensive spaces in Mumbai and India too. For anyone or any enterprise who plans to start a food joint or restaurant must analyze every spot in this region and understand which region is best for each cuisine that is available. Competition is an important factor while setting up a restaurant and this can affect the sales in a huge way. Also review regarding the existing restaurant will be essential towards deciding the place to set up. So the question i want to answer is, if i have a cuisine i want to sell in a restaurant i wish to set up, then where is south mumbai should i sell?

1.2. Target Audience of this project

This project is particularly useful for any entrepreneur who wants to get in the restaurant business line and plans to get up a restaurant in South Mumbai. Also this project will be help for any franchise which wants to bring a particular cuisine into this region. So in order to have a good understanding about the current situation of the restaurant locations and cuisines provided, this project will help visual and gather meaningful data from various source and help the targeted audience make a meaningful and aware decision.

South Bombay or South Mumbai is the Mumbai City district which is the southernmost precinct of Greater Mumbai. It extends from Colaba to Mahim. It comprises the city's main business localities, making it the wealthiest urban precinct in India.

2. About Data

The target of this project is to suggest types of food franchises that can be set up in South Mumbai based on the historic data about the existing restaurants. Using FourSquare API and Zomato API, it is possible to extract data of existing food businesses around the locality and form a conclusion about what can be the best type of restaurant that can be set up which has a higher probability of being profitable.

Based on the data collected from the two API about the existing business in that particular area we can then move towards cleaning and finding a possible correlation between the business, its opinion based on the rating, location, price, etc. Once curated, the user can be prompted with a suggestion regarding what type of restaurant is lacking in what part of South Mumbai and this can help the user come closer towards making a decision.

2.1. Data Collection and Description

Two main APIs will be required to satisfy the data required for this project and they are :

FourSquare API: This API can help collect all the venues up to a radius specified. We are going to analyze a radius of up to 8km-10km depending on the data requirement in conjugate to the API calls that can be made.

Zomato REST API: Fetched venues from the above API can be used as input for this API which in return gives a rating of the venue, price ranges, etc

Now we can begin fetching the venues in the region of south Mumbai with a radius of 8km wrt to the target latitude and longitude as the center. The Foursquare API has the explore API which allows us to find venue recommendations within a given radius from the given coordinates. We will use this API to find all the venues we need.

FourSquare sends in venues of all kinds and not only restaurants. So providing the categoryId will give us a curated list of venues that are all restaurants. But providing a categoryId will give us only 50 items for one location (100 if no categoryId is provided). So to maximize the number of restaurants we 3receive in that location coordinates, we change the offset value in the API request after every call. Once we have collected a good set of list of venues, the loop is terminated and all the duplicate venue values we receive by changing the offset is removed. This gives us a total of 109 restaurants in a 4 km radius around the target location coordinates. (Number and the types of venues change depending on the time of the day the API is called. That is how Foursquare functions under the hood. But the number will be close to 105–109.)

Zomato's list of APIs provides us with various types of data regarding restaurants. Depending on what is needed for that particular project like cuisine, daily menu, review, etc, all of these data can be accessed through Zomato's REST API.

Accessing the data through the API requires a user access key which is to be accepted from the developer website page, by submitting the necessary information. For this analysis, Zomato's

search API will be used, which will help to search for any particular venue based on its name, latitude, longitude, etc. Since we have 3 major values for search already ready from the previous API used, it will be helpful in determining the other values of the venue itself. Out of the 109 rows fetched through the foursquare API, only 2 of the venues have no data regarding it registered in the Zomato Database and hence the API returns no info. The index of the unknown venue is stored in the error_list and will be removed in the next Stage.

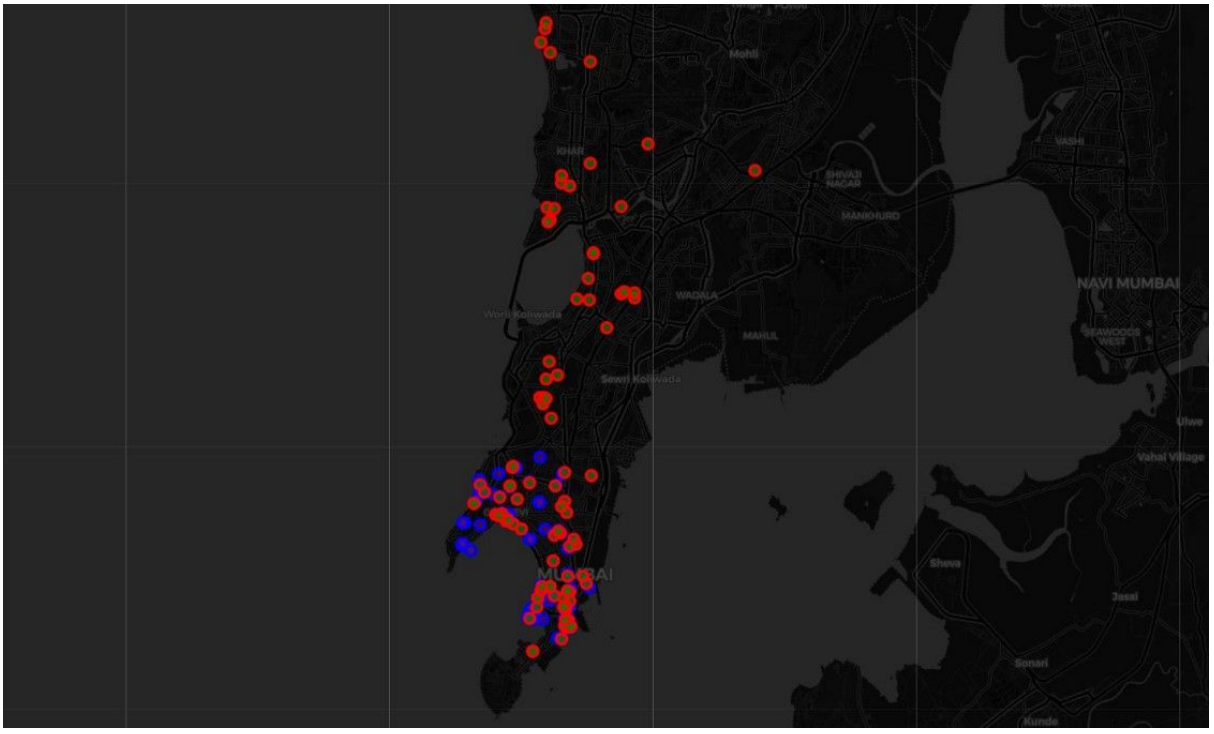
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.colors as colors
import requests
from pandas.io.json import json_normalize
offset = 0
while True:
    url =
('https://api.foursquare.com/v2/venues/explore?categoryId={}&client_id={
}'

'&client_secret={}&v={}&ll={},{}&radius={}&limit={}&offset={}').format(c
ategoryId,
FOURSQUARE_CLIENT_ID,
FOURSQUARE_CLIENT_SECRET,
FOURSQUARE_VERSION,
TARGET_LATITUDE,
TARGET_LONGITUDE,
radius,
TOTAL_VENUES,
offset)
result = requests.get(url).json()
feteched_venues = len(result['response']['groups'][0]['items'])
# changing the offset can introduce many duplicate values. a simple
duplicate drop can remove them
foursquare_venues.drop_duplicates(inplace=True)
foursquare_venues.shape
```

3. Methodology

3.1. Data Preprocessing and Visualization

This stage includes combining data from multiple sources and filtering out the un-relevant data. Starting with removing all the venues from the data frames that have no local data from Zomato to back it up based on the list error_list derived from API calls made during Zomato API calls.
Once the above processing has taken place we have to combine both the data frames as one based on one or more common columns that can be name, latitude, or longitude. It will be explained below after cleaning the data frame.
Now the next processing task is to merge both the data frames as one thereby dealing with one common data frame. But Merging two data frames effectively and accurately required one or more common columns between the data frames. Initially, Let us try mapping both the Zomato and foursquare data onto a map and see how close the overlapping is. Mapping the coordinates from the dataset in a map created using Folium maps library in python.



🔵-> FourSquare API 🔴-> Zomato API

Here we observe that Foursquare API provides us with the right set of venues that are based in the south of Mumbai. But of the corresponding venue, Zomato API gives us an inaccurate venue and this is known by the above map. There are location coordinates that are scattered all around the region and thus it is not essential for our analysis. But we also see that there are location coordinates from both the APIs that overlap or are pretty close to each other. These are locations under consideration for this project.

Here we observe that Foursquare API provides us with the right set of venues that are based in the south of Mumbai. But of the corresponding venue, Zomato API gives us an inaccurate venue and this is known by the above map. There are location coordinates that are scattered all around the region and thus it is not essential for our analysis. But we also see that there are location coordinates from both the APIs that overlap or are pretty close to each other. These are locations under consideration for this project.

Through trial, we can observe that reducing the digits after the decimal point does not induce much change in the overall position of the markers on the map. And hence the above step might not be as necessary. Sorry for the diversion.

But in order to combine both the dataset, we require a unique identifier between them. We understand that both the data frames are arranged similarly and their indexes are similar i.e every venue received from the FourSquare API has its corresponding information at the same index in the Zomato data frame. But the Zomato API does not always give us the same latitude, longitude, and even the same venue information(the venue itself is someplace else). As seen in the above map it is understood that many of the same venues from both the data frames overlap with each other. But some of them do not. Also, names for venues provided by both data frames mostly do not perfectly match with each other.v(There can be a difference in a letter or there are special characters in the names). So also for a particular venue provided by the foursquare API a completely new venue will be provided by Zomato API. Therefore, in order to curate a final data frame for evaluation, we need to combine both the venues from both the data frames by eliminating the duplicates

3.2. Process of Merging and Elimination

Merging can be based on two factors:

1. Initially, a column consisting of the difference between the latitude and longitude of both the venues based on an index. If the difference in the latitude or longitude between both the venues is less than or equal to 0.004 then we go on to analyze the next feature i.e. the venue name
2. Names in both the data frames can be the same and different in some cases. It is obvious that if the venue data received from the Zomato API is completely different then the names won't match and the possibility of venues with the same name in extremely close proximity is very low
3. For every name in each row (name_x, name_y) we split the name into individual words that can compare each word in both data frames now combined. If there is just one word as the name of the venue we compare the name and the location coordinates. But when the name exceeds 2 or more, we compare words, and if words in the venue names have half more words similar + have latitude and longitude similar then it is combined in the final dataset.

```
def checkName(a, b):
    list_A = [x.lower() for x in a.split()]
    list_B = [x.lower() for x in b.split()]
    min_list = min(list_A, list_B)
    max_list = max(list_A, list_B)
    x = len(min_list)
    y = len(max_list)
    if x == 1 and y == 1:
        if min_list[0] == max_list[0]:
            return True
        else:
            return False
    else:
        res = 0
        for i in range(0, x):
            if min_list[i] in max_list:
                res += 1
        if res == x:
            return True
        else:
            return False
```

Now finally, let's merge both the data frames.

```
result = []
for index, row in final.iterrows():
    val = (abs(row['lat_diff']) <= 0.0004) & (abs(row['lng_diff']) <=
0.0004) | checkName(row['name_x'], row['name_y'])
    if val == True:
        result.append(index)
print(len(result))
# result consists of all the indexes in the final dataframe that is to
be a part of the final dataframe to
# be visualized and evaluated
final_df = final.loc[result, :]
final_df.reset_index(inplace=True, drop=True)
final_df.drop(columns=['Unnamed: 0_y', 'Unnamed: 0.1_y', 'name_y',
'lat_y', 'lng_y', 'price_range', 'votes' ,'lat_diff', 'lng_diff'],
inplace=True)
final_df.columns = ['name', 'categories', 'lat', 'lng',
'avg_cost_for_2', 'agg_rating', 'rating_test', 'address',
'review_count', 'cuisines']
```

	name	categories	lat	lng	avg_cost_for_2	agg_rating	rating_test	address	review_c
0	Food for Thought	Café	18.9320	72.8317	1000	4.3	Very Good	45/47, Kitabkhana, Somaiya Bhavan, Mahatma Gan...	472
1	Royal China	Chinese Restaurant	18.9387	72.8329	2500	4.4	Very Good	Hazarimal Somani Marg, Near Sterling Cinema, F...	513
2	Shree Thaker Bhojnalay	Indian Restaurant	18.9512	72.8283	1200	4.9	Excellent	31, Dadisheth Agyari Lane, Off Kalbadevi Road,...	852
3	The Oriental Blossom, Marine Plaza	Asian Restaurant	18.9316	72.8231	3200	3.8	Good	Hotel Marine Plaza, 29, Marine Drive, Churchga...	206
4	Pizza By The Bay	Pizza Place	18.9335	72.8239	2000	4.3	Very Good	143, Soona Mahal, Marine Drive, Churchgate, Mu...	3163
...	...	...	...	...	...	...	...	...	...
72	The Sun	Vegetarian / Vegan Restaurant	18.9550	72.7985	1000	2.9	Average	93, Beach View, Bhulabhai Desai Road, Breach C...	66
73	Santosh sagar	Indian Restaurant	18.9550	72.7977	550	4.1	Very Good	Shop 6, Napeansea Road, Matru Ashish, Malabar ...	153
74	Delifresh	Bakery	18.9650	72.8040	650	3.3	Average	Dhunabad, 106 Bhulabhai Desai Road, Kemps Corn...	3
75	Mafco	Fast Food Restaurant	18.9659	72.8037	550	2.9	Average	GM Road, Chembur, Mumbai	37
76	Howra	Café	18.9663	72.8041	600	4.0	Very Good	Shop 6, Dunhill Apartment, Waroda Road, Hill R...	697

77 rows × 10 columns

Let’s view the map using the final data frame we created. You will be surprised to see this one.

Now we have a total of 77 restaurants scrapped from the entire data frame collected from the API which had a total of 117. Using only the location(lat, lng) difference of 0.0004 we only get 45 restaurant venues. But including the variable name gives us more out of the data from the API. Also, we see that all the venues are part of the South Mumbai region and there will be no venues outside the radius which will surely give us more accurate result

3.3. Data Visualization

Using matplotlib and seaborn as the visualization tools it can be made possible to view different aspects of various restaurant venues in South Mumbai. This can be related to the types of cuisine offered or the types and the frequency of the categories of restaurants etc.
One of the many jobs pertaining to data science is to understand the data you are working with. Visual representation of data makes it very easy and understandable to even others who are new to the data or someone who needs to get insight regarding the data for other tasks. That is what we are going to do now

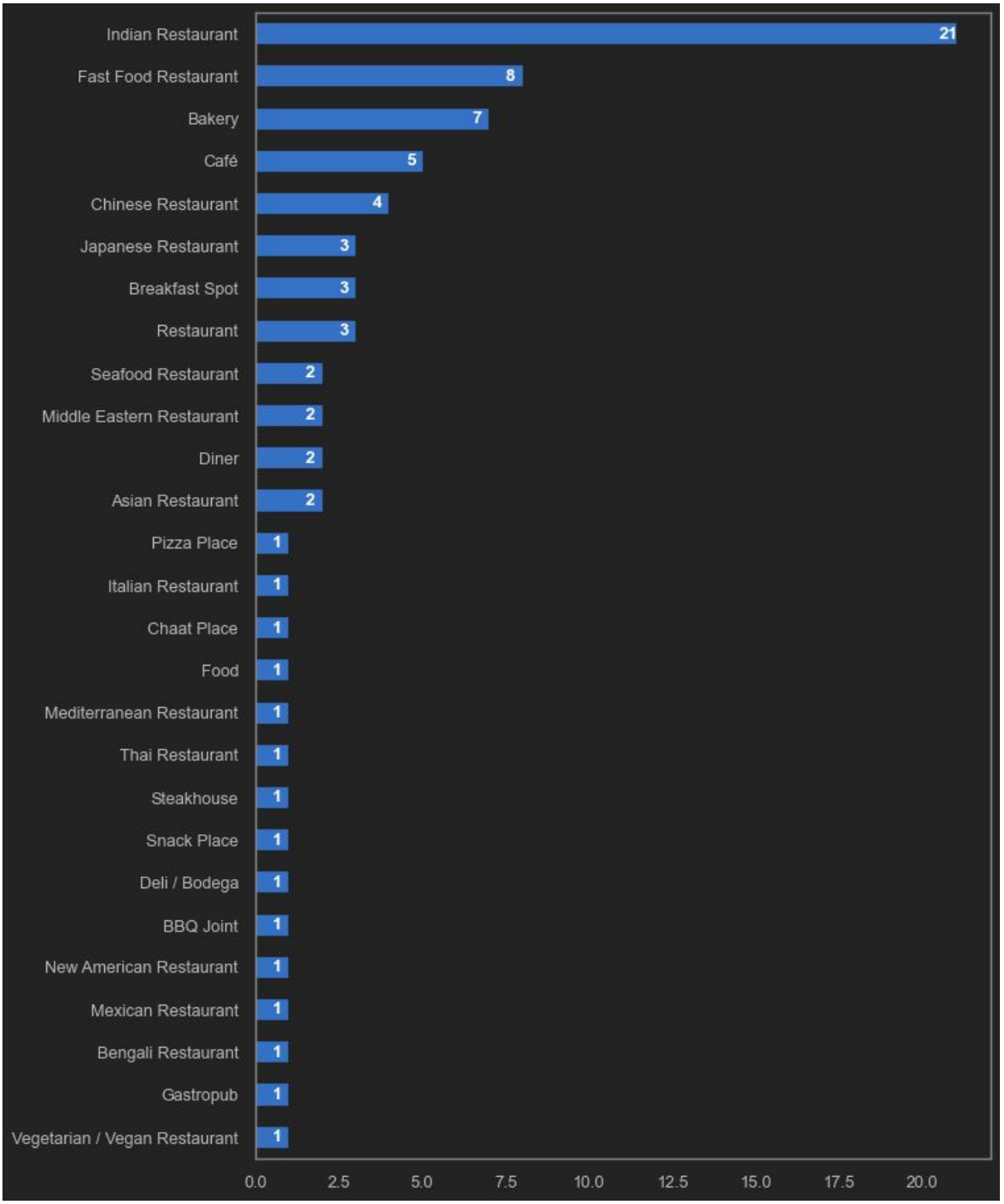


Fig 1: Restaurant Categories and the Count of Each Categories

Here we see that there is a high number of Indian Restaurants, followed by Fast Food Restaurants. But there are many categories which consist of only one restaurant like a pizza place, Mexican restaurant, etc.
But we also have to look at the cuisine the venue serves to understand what kinds of food is abundant over in the south Mumbai region
Let's look at the cuisines offered in South Mumbai Cuisine columns have cuisines of that particular restaurant together and we need to count them individually. And we thus use a dictionary to represent them.

Fig 2: Cuisine Based Word Cloud

Having over 40 cuisines makes it difficult to represent its frequency using a common plotting technique. So word cloud is a unique way of placing the total cuisines in these restaurants we fetched. I used the word cloud API which was introduced during the IBM Data Visualization course in Coursera. The word cloud above gives us an idea about the type of cuisines available in South Mumbai. Bigger the word more is the cuisine available here. No shortage of Chinese Food over here .

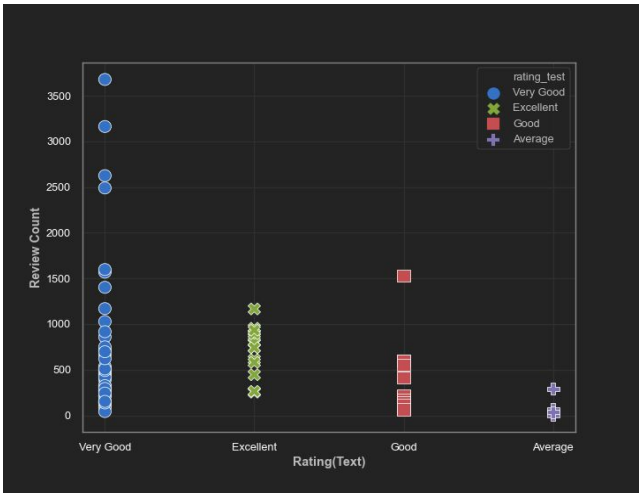


Fig 4: Review of Each restaurant

In the above visual representation, we can see that The three types of restaurants fall into different categories based on the ratings and the number of reviews. Let us start with Excellent restaurants. We see that there is always a review of excellent restaurants. This makes these restaurants well know and well established.
For Restaurants that are termed Very Good, it is seen that there are venues with the high review, and thus this falls under the same category. But there are few restaurants that have lower reviews(lower than 300_) and the cuisines they serve and their location are the kind of restaurants we can target to understand whether placing restaurants adjacent to them will be profitable or no. The same goes for restaurants that have ratings _Good. Many have reviews higher than 500 but there are few with lower review and these fall under the same category as before. Now for restaurants rated Average, we observe that reviews are lower than 500 for almost all the venues and also they are rated average.
So providing a better version of these restaurants in terms of cuisine and location will have a higher impact in terms of growth restaurants that we will establish.

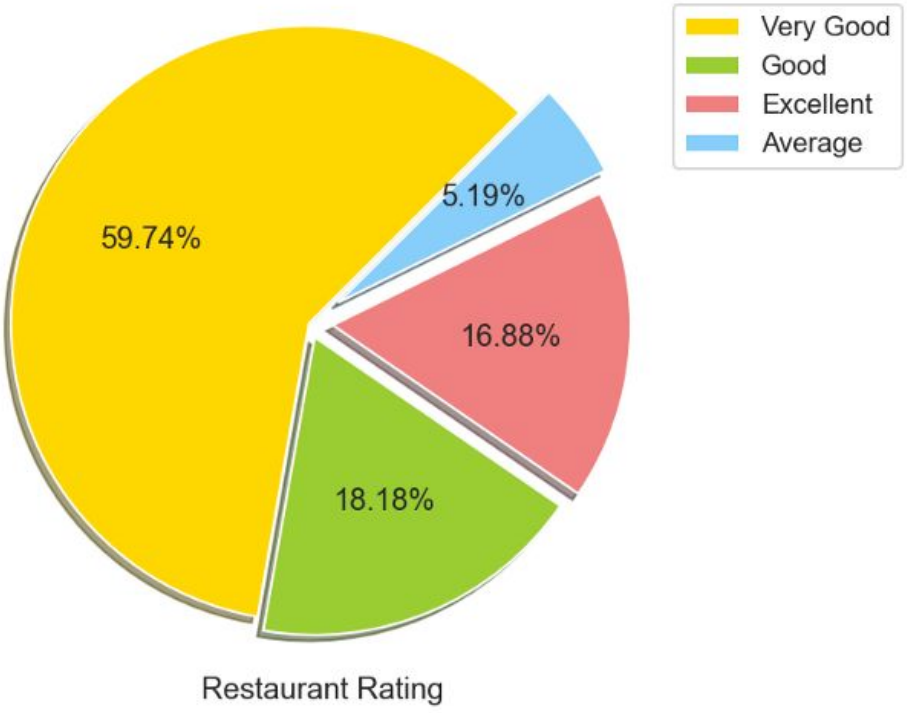


Fig 3: Pie Chart Based on Rating of Restaurants

So, we have amazing restaurants in South Mumbai and very few restaurants are 'Average'.
Let us view how rating count and review count are related together. It is obvious for the fact that higher ratings and higher reviews denote the restaurant is more favorable. higher review and lower rating mean the venue is not a favorable place for customers. Let's try to plot this concept below.

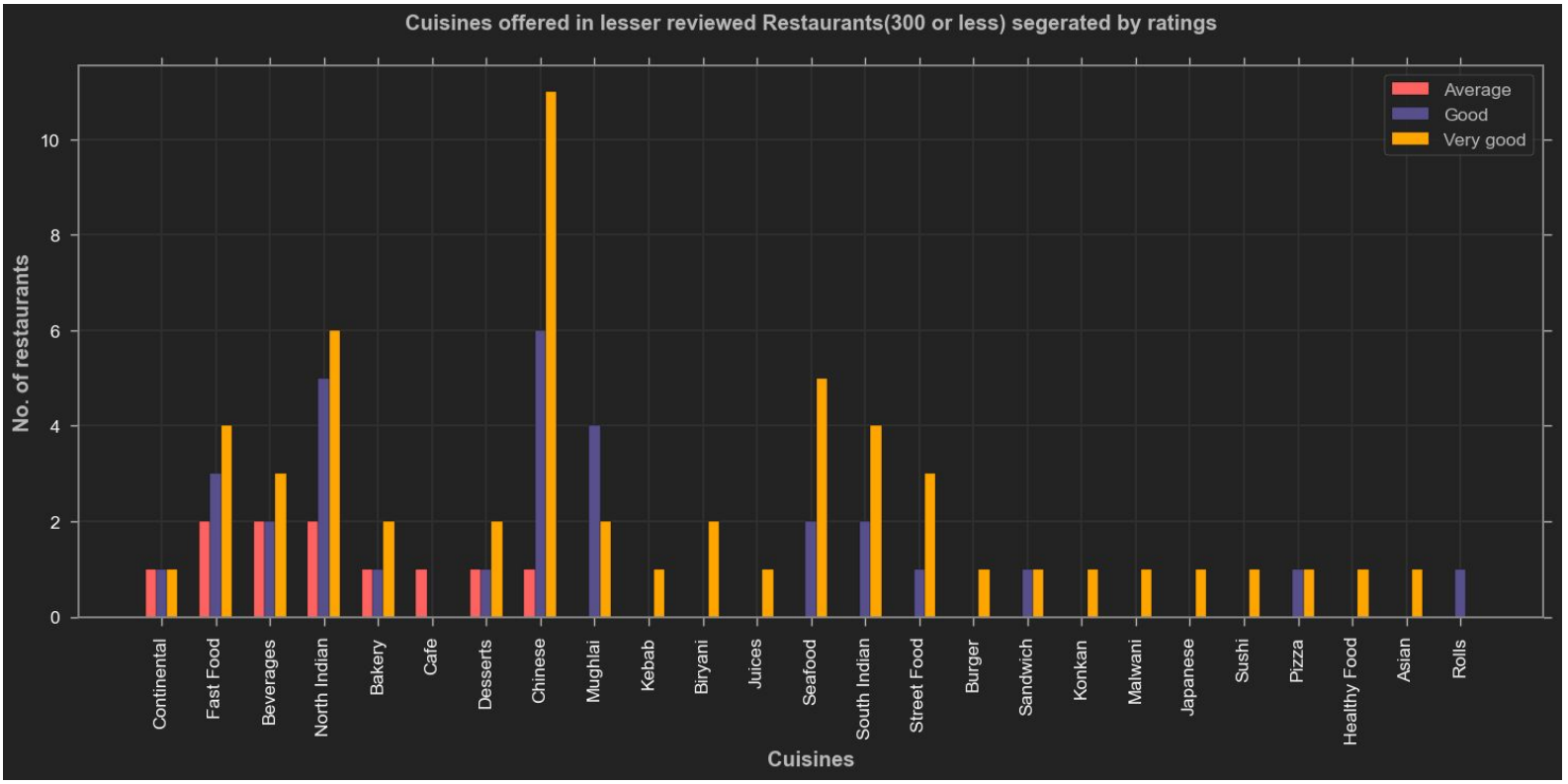


Fig 5

Chinese cuisines are pretty abundant amongst these restaurants with lower reviews. Also, we see that Continental, Beverages, Bakery, cafe, and Chinese have atleast one restaurant with a review that is average, and also number of restaurants rated 'Good' is also pretty few. So this makes it a risk to serve cuisines like these as chances are that there will be not widely accepted by people here. So only if the cuisine to be served is special or unique in terms of the already existing stuff then the chances of getting accepted amongst the public are higher.
we see that Cuisine such as Konkan, Japanese, Sushi, Asian have lesser reviews but the review are all 'Very Good' or 'Good'. This shows us a possibility that restaurants with cuisines like Continental, FastFood, Beverage, North Indian, Bakery, Cafe, Chinese, Mughlai have shown reviews that are average, but the 'Good' and 'Very Good' rating are higher compared to the 'Average' Rating. So we can consider bringing in a restaurant that has a higher rate of being accepted since getting rated 'Average' is low.

3.4. Machine Learning

Initially, my plan was to cluster restaurants based on the cuisine it offered. So the initial step was to tweak the data frame to add a one-hot encoding scheme based on the cuisine offered. Since the clusters can be of arbitrary shape we Density-Based Clustering was the way to go. The data frame will consist of columns like the cuisines, rating, and cost for two.
But looking at the map after clustering the venues gives us no good clusters. This shows us the cuisines offered in different venues are spread randomly with no particular location providing specific cuisine.

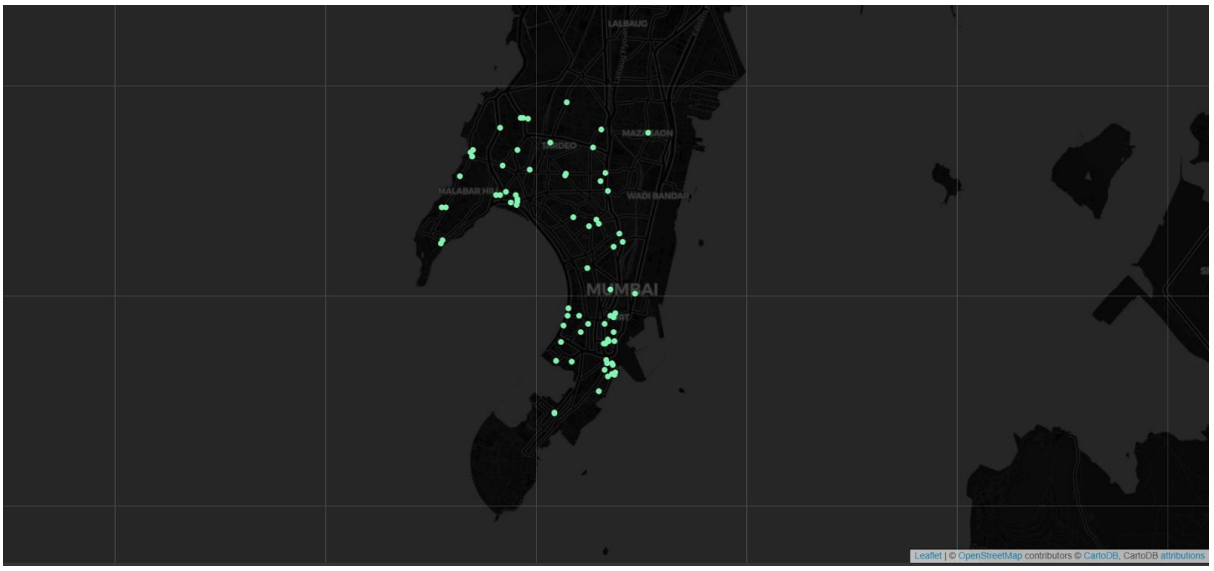


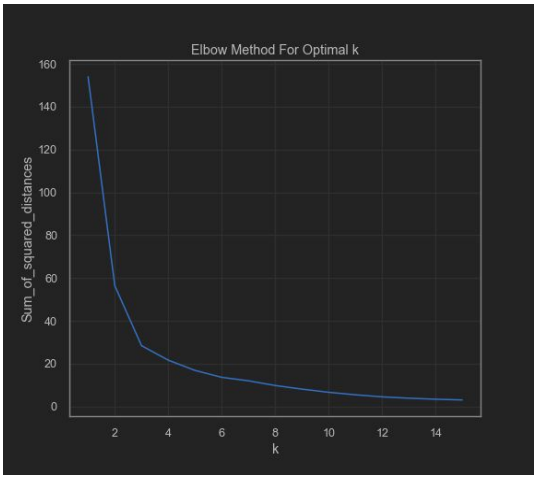
Fig 6: partition resulting from DBSCAN

Now that we understand this fact our next strategy is to cluster based on the Latitude and Longitude of the venue as well as the aggerate rating and cost for two. Since we involve latitude and longitude as a clustering factor we have a higher rate of finding cluster and thus we use KMeans as it can give us the uniform separation between two.
But there is a question of how many clusters do we need to set the value as. Well, I don't know so we have to check by setting cluster num to various values.

```
from sklearn.cluster import KMeans
test4 = final_df[['lat', 'lng']]
# well scaling isn't essential since lat and lng are only used and no
reviews or ratings are used.
test3 = StandardScaler().fit_transform(test4)
sum_of_squared_distance = []
k_range = range(1, 16)
for i in k_range:
    cluster_KM = KMeans(init = "k-means++", n_clusters = i, n_init =
12).fit(test3)
    sum_of_squared_distance.append(cluster_KM.inertia_)
```

Out of all the K values used in the algorithm, which one is the optimal one?

```
plt.plot(k_range, sum_of_squared_distance, 'bx-')
plt.xlabel('k')
plt.ylabel('Sum_of_squared_distances')
plt.title('Elbow Method For Optimal k')
plt.grid(True)
plt.savefig('elbow.png')
plt.show()
```



The above method is called the 'Elbow Method' that is used to depict the Optimal K for a dataset. Using the sum of squared distance for each value of K, we plot it on the graph. It is seen that as the value of k increases the sum of squared distance value tends to zero. Now if the plot looks like an arm the optimal value of k will be the elbow of the arm, i.e. in this case the value of k is 3.

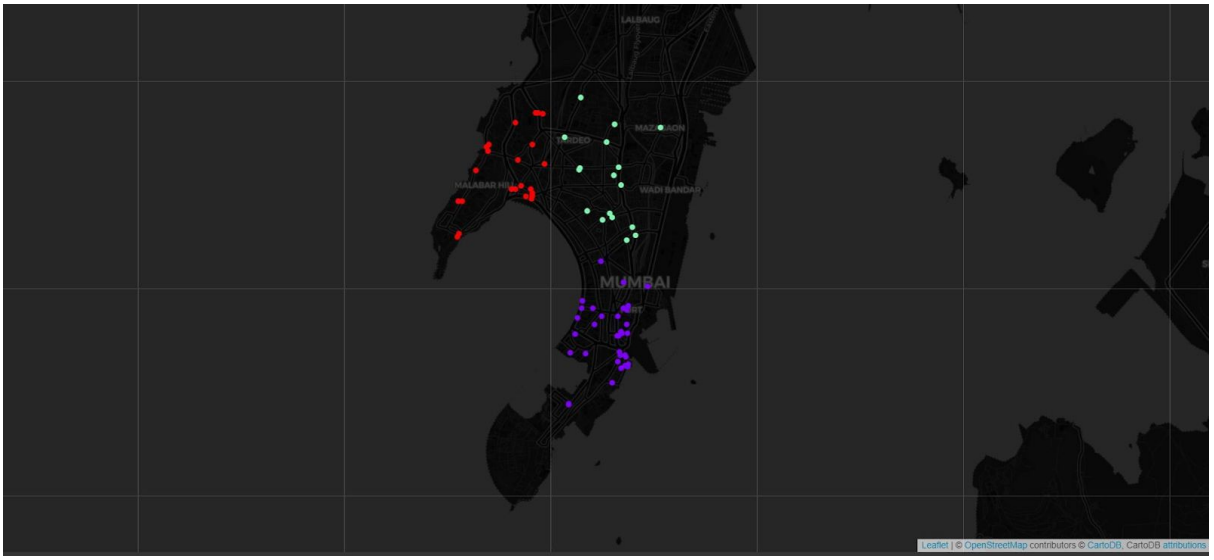


Fig 7: Cluster based on KMeans

Clustering based on the location gives us an accurate depiction of how the restaurant is placed in south Mumbai. With the number of clusters being equal to 3 we assign each and every venue in south Mumbai to a cluster. Now each cluster can have different characteristics based on the reviews or the cuisines etc. We will now explore all three clusters and extract their characteristic factors, thereby determining what each cluster is good at. And in the end, come to a conclusion on which restaurant should be set up in which cluster in South Mumbai.

The above map represents the mean rating across all the venues in the clusters in South Bombay. Since there is a limited amount of venues to judge the cluster, we still see a small difference in the average rating.

- The red cluster as we see has the highest rating of the overall 3 clusters. This denotes that the restaurants over in this cluster have higher ratings and thus the people will usually go towards restaurants in this area. Making it difficult to set up a new restaurant. But for better exposure to the restaurant, we can set up our new restaurant based on the types of cuisine served. and the rating
- For the blue and purple clusters, the ratings are pretty close. The purple cluster has a smaller rating of the three. Thus starting a restaurant with a good menu and right cuisine for that cluster will tend to have a higher value. But again, the cuisine served and the quality of the restaurant will be the judging factor as lesser people will deviate towards restaurants in the purple and blue clusters compared to venues in the red cluster.

```
                        CLUSTER INFORMATION
************************************************************************
                  Cluster Red/ Cluster 0
      Total Number of Categories:   14
      Rating range:                 2.9 ---> 4.9
      Rating in words
            Good:       3
            Excellent:  5
            Average:    4
            Very Good:  12
      Total Reviews in the Cluster: 9774
      Total Cuisines Served         25
            Total Restaurants in the cluster: 24
************************************************************************
************************************************************************
                  Cluster Purple/ Cluster 1
      Total Number of Categories:   17
      Rating range:                 3.7 ---> 4.8
      Rating in words
            Very Good:  26
            Good:       4
            Excellent:  6
      Total Reviews in the Cluster: 28541
      Total Cuisines Served         35
            Total Restaurants in the cluster: 36
************************************************************************
************************************************************************
                  Cluster Blue/ Cluster 2
      Total Number of Categories:   7
      Rating range:                 3.6 ---> 4.9
      Rating in words
            Excellent:  2
            Very Good:  8
            Good:       7
      Total Reviews in the Cluster: 8960
      Total Cuisines Served         14
            Total Restaurants in the cluster: 17
************************************************************************
```

Results

- **Red Cluster**

  Let us start by looking at cluster 0 or the red cluster. We see that this is one of the most popular clusters with the highest number of categories of restaurants and also with the highest amount of reviews received. (28541). Even the range of rating is from 3.7 to 4.8, i.e. there will be close to no bad or even average restaurants in this cluster. Also, the number of cuisines served in the red cluster is very high(35) out of the total 40 cuisines found in the whole of South Mumbai. This makes the cluster A very vivid and unique in terms of options to dine for the customers. But, it makes the cluster A have a higher competition with other already set restaurants.

  The red cluster is therefore very risky for new restaurants to start due to the already existing and well-accepted restaurants. However, it also for a fact that if there is a good plan of execution to start up a restaurant that provides unique or a new variant of cuisine, or even a cuisine that is widely accepted and liked in some other region, then there is a higher probability of the restaurants turning in a profit. This makes cluster A a risky site to set a restaurant, but with the right set of plan there can be a chance to generate a high output since profits in the restaurants in cluster A is high.

- **Purple Cluster**

  There are 14 categories of restaurants in the purple cluster. Compared to the red cluster we see that ratings can range from 2.9 all the way to 4.9, i.e. there are good restaurants, excellent ones, and also average ones. Compared to the red cluster there are less than half reviews we can see over here in this cluster(9774). Out of the total 40 cuisines served in South Mumbai, you can find only(24) types of cuisine in this cluster. But the restaurants rated average in South Bombay are also concentrated in this cluster. Out of the 17 venues, there are 3 average restaurants, 2 Good, 11 Very Good, and 4 Excellent restaurants.

  Total Number of restaurants is 24 and the number of cuisines server are 25. This shows us for every cuisine we can find only one or up to 2 types of restaurants. Rating range from 2.9 to 4.9. This makes cuisine served in a few restaurants in this cluster not widely accepted by the customers. Thus, for a well-decided menu for particular cuisines, we can expect a large number of people trying out the cuisine in the restaurant we setup here. But compared to the red cluster, the number of people coming into restaurants in this cluster is less than half of what we see in the red cluster. The level of risk in this cluster is comparatively lower but turnover profit will take higher time.

- **Blue Cluster**

  Out of all the 3 clusters, we observe that this particular cluster has the lowest footfall(based on the reviews). This can be because the types of cuisines offered over in the restaurants in this cluster are low (14 out of the total 40 cuisines found in South Mumbai). Rating for each restaurant are good but their categories many restaurants tend to be the same. The overall density of restaurants in this cluster is very low as compared to other clusters.

Therefore starting up a new restaurant in this cluster has a higher probability of gaining traction as we see that there are not so many variations in cuisine available. Reviews available for restaurants has no considerable difference compared to the purple cluster. This shows us that their restaurants in the blue cluster do have customers coming in (as ratings in this cluster are also high compared to the purple cluster). But the number of restaurants in this cluster is low and the categories of this many restaurants are also the same. Introducing variations in cuisines offered in this cluster can have a positive impact on the sales as this cluster lacks the variations in the restaurants and cuisines.

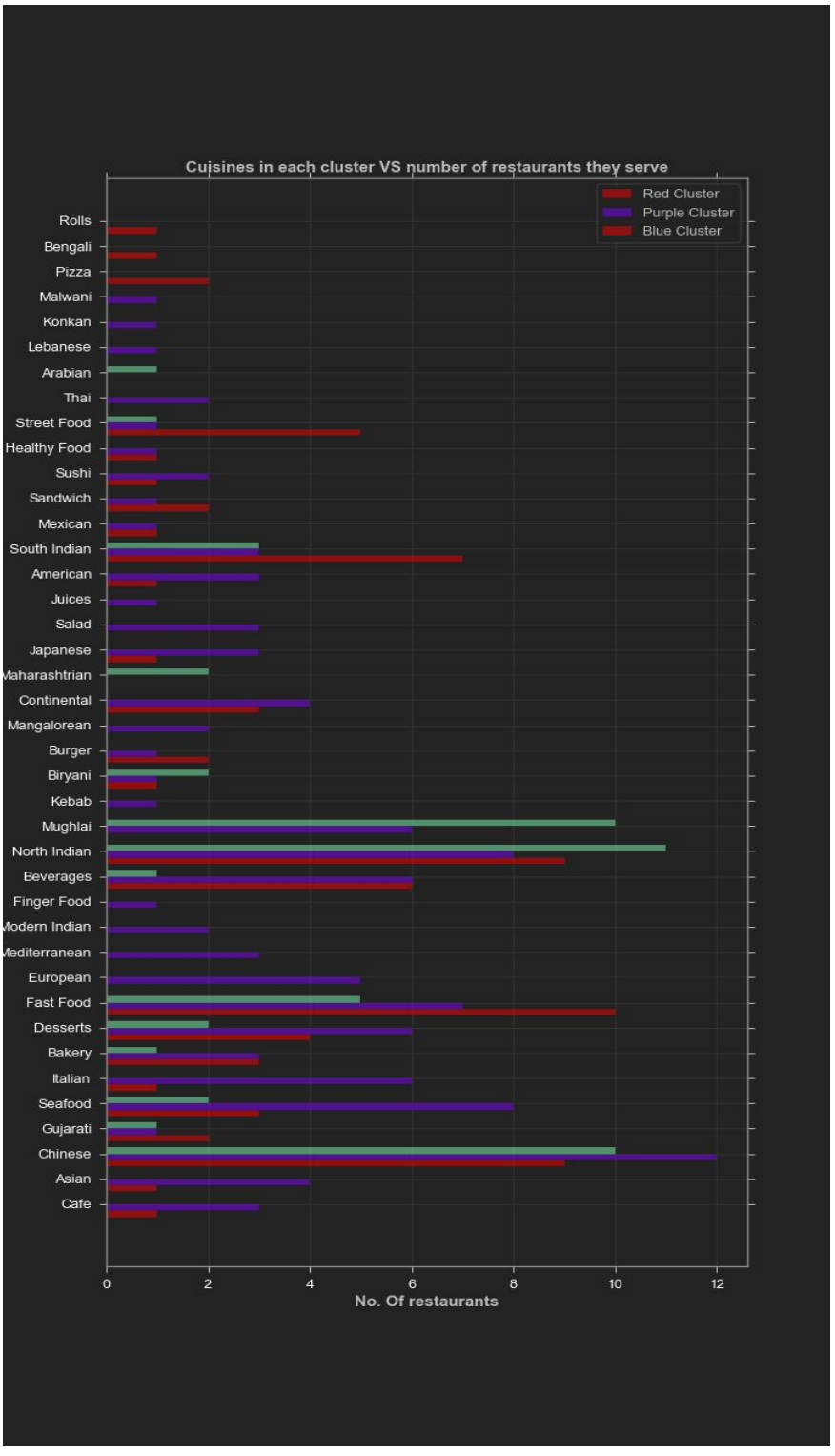


Fig 8: Cuisines in cluster

Limitations and Suggestions for Future Research

In this project, we only consider few factors we considered while coming to a conclusion when deciding about the clusters and the good and the average cuisines served in each clusters. But, the limitation to this is the data that is available as open source to use. Zomato does not provide a refined API and is a bit faulty. Also they do not provide with entire data as they are a privately owned company. Gathering other information such as footfall everyday of the week, types of dishes served will highly help towards creating an accurate representation of what is the best food to eat in South Mumbai and where to find it. Also it will help towards gathering creating a model that helps us identify about the food that customers want in a region but is not available. This can be part of the future research and has the potential to go main stream towards helping individuals and organizations who plan on setting up restaurants in South Mumbai.

Conclusion

So where will I set up my restaurant and what will I sell?

The risk seems to be pretty high in the red cluster. I feel that it will be difficult to generate profits for a new restaurant as the number of restaurants is very high and also these restaurants are widely accepted and liked by people. Setting up a branch for an already established chain of restaurants will have a greater probability to succeed in this cluster.

But analyzing restaurants' red clusters can help us understand people like in that cluster like the cuisine or a particular category of restaurant. Since 35 of the 40 types of cuisines are available in the restaurants in the red cluster, depending on the inclination of people towards a particular category or cuisine, we can introduce them in the other two clusters.

Competition in the purple cluster is not staggering compared to the blue cluster, but we have seen many more different types of cuisines served in the purple cluster as compared to the blue one. Purple cluster is a little better exposed to the citizen compared to the blue cluster, I will most probably set up a restaurant with a unique type of cuisine that is not available in the purple cluster or is not widely accepted. The risk seems to be high compared to setting up a restaurant in the blue cluster but the return on investment will be high. Pizza restaurant, Cafe, Asian Cuisine are some of the major cuisines we see in the red cluster that are not much popular in the purple cluster. Fast food won’t be a viable restaurant to set up because much of these joints already exist. Even Chinese food is very abundant in the entire South Mumbai. Even Meditarian, European, Konkan cuisines are less to be found in the purple clusters and thus can be started here. But Mughlai is a type of cuisine abundant in the red and blue cluster but is not present in the purple cluster, and this is what I will start. Also, I love Mughlai food!

References

KMeans Algorithm implementation
https://blog.cambridgespark.com/how-to-determine-the-optimal-number-of-clusters-for-k-means-clustering-14f27070048f

Foursquare Developers Documentation. *Foursquare*. Retrieved from
https://developer.foursquare.com/docs

Dataset Cleaning and Merging technique
https://towardsdatascience.com/exploring-the-tokyo-neighborhoods-data-science-in-real-life-8b6c2454ca16