Exploring venues in Mumbai, India using Foursquare and Zomato API

1. Introduction

The basic idea of analyzing the Zomato dataset is to get a fair idea about the factors affecting the establishment of different types of the restaurant at different places in Mumbai, aggregate rating of each restaurant, Mumbai being one such city has more than 12,000 restaurants with restaurants serving dishes from all over the world. With each day new restaurants opening the industry hasn't been saturated yet and the demand is increasing day by day. Inspite of increasing demand it, however, has become difficult for new restaurants to compete with established restaurants. Most of them serving the same food.

1.1 Background

Whenever a person searches for a venue in a new city, they're highly interested in the best places that the city has to offer. The person might want to know how good a given restaurant is or the price range it falls under. This extra information would help decide which venue to choose amongst the many venues in the city. Combining the location of the venues in the city with their price and rating information would surely help visitors in a city make better informed decisions about the places they should visit.
Mumbai is composed of a number of sectors spread across a total area of 603 sq Km. There are many venues (especially restaurants, hotels and cafes) which can be explored. This project explores various venues in Mumbai and attributes the data based on user ratings and average price. To explore this information, this project involves the juxtaposition of both the Foursquare API and the Zomato API to fetch complete information of various venues (including name, address, category, rating, and price). Further, a map of the venues with specific color attributes will be plotted to highlight their position, and information about these venues. Such plots imbibe bountiful information in the form of their colored representations and location on the map. This enables any visitor to take a quick glance and decide what place to visit.

1.2 Interested audience

The target audience for such a project is twofold. Firstly, any person who is visiting Mumbai, India can use the plots and maps from this project to quickly select places that suit their budget and rating preferences. Secondly, a company can use this information to create a website or a

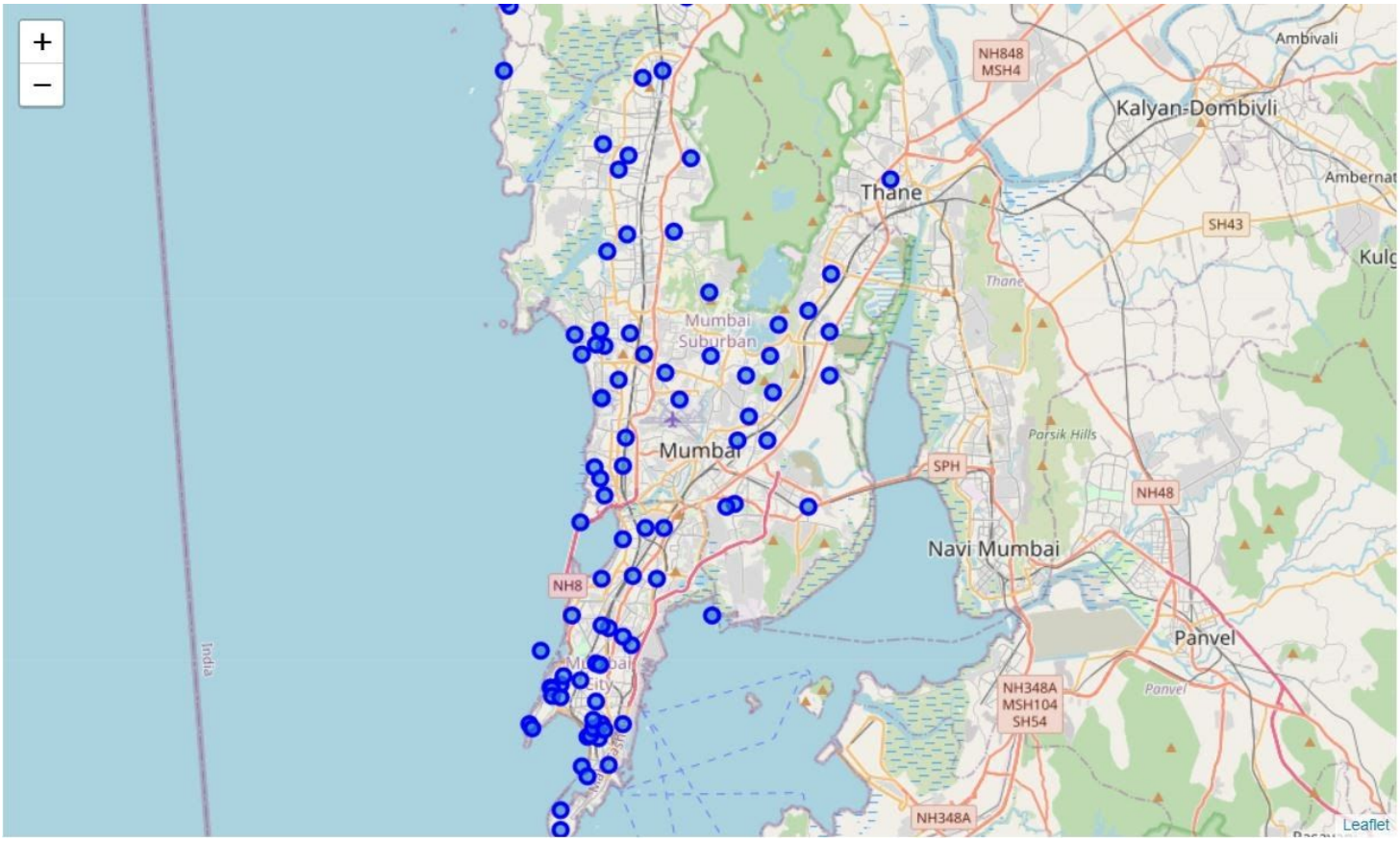
mobile application, which is updated on a regular basis, to allow individuals to the city or even expand the same functionality to other places.

2. Data

2.1 Data Sources

1. To get list of neighbourhoods I have scrapped the data from Wikipedia (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai). And retrieved table

	Area	Location	Latitude	Longitude
0	Amboli	Andheri,Western Suburbs	19.129300	72.843400
1	Chakala, Andheri	Western Suburbs	19.111388	72.860833
2	D.N. Nagar	Andheri,Western Suburbs	19.124085	72.831373
3	Four Bungalows	Andheri,Western Suburbs	19.124714	72.827210
4	Lokhandwala	Andheri,Western Suburbs	19.130815	72.829270

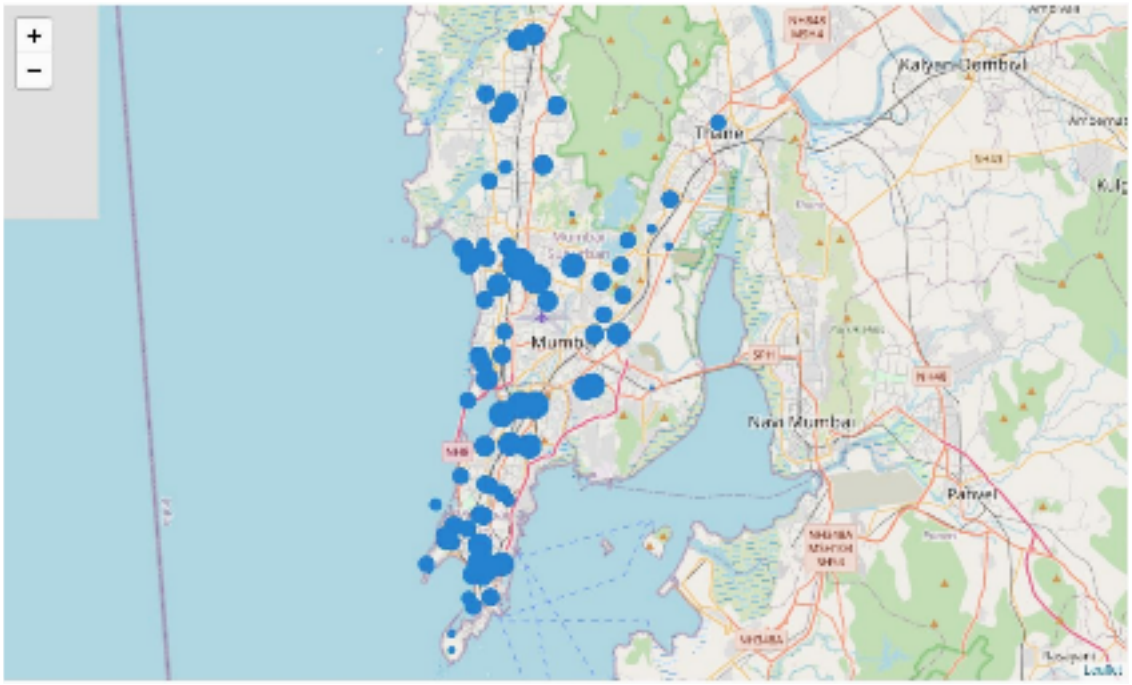


2. To get location and other information about various venues in Mumbai, I used two APIs and decided to combine the data from both of them together.
   Using the Foursquare's explore API (which gives venues recommendations), I fetched venues up to a range of 2 kilometers from the center of Mumbai and collected their names, categories and locations (latitude and longitude).
   Using the name, latitude and longitude values, I used the Zomato search API to fetch venues from its database. This API allows to find venues based on search criteria (usually the name), latitude and longitude values and more. Given that the data from the two APIs did not align completely, I had to use data cleaning to combine the two datasets properly.

From Foursquare API (https://developers.zomato.com/api), I retrieved the following for each venue:

	Neighborhood	Latitude	Longitude	venue_id	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Amboli	19.1293	72.8434	5174e2be498e39cf0d1c20cb	Shawarma Factory	19.124591	72.840398	Falafel Restaurant
1	Amboli	19.1293	72.8434	4b0587e2f964a52095a522e3	Merwans Cake shop	19.119300	72.845418	Bakery
2	Amboli	19.1293	72.8434	55fc3615498e141bd45da525	Jaffer Bhai's Delhi Darbar	19.137714	72.845909	Mughlai Restaurant
3	Amboli	19.1293	72.8434	51fa1f52ccdae6540ad807b4	Hard Rock Cafe Andheri	19.135995	72.835335	American Restaurant
4	Amboli	19.1293	72.8434	4e4eb3c68877402b06b92160	5 Spice , Bandra	19.130421	72.847206	Chinese Restaurant
5	Amboli	19.1293	72.8434	52b3fe27498e991e4fe996cb	Pizza Express	19.131893	72.834668	Pizza Place
6	Amboli	19.1293	72.8434	4cc1d37c3d7fa1cd0de39a5f	Joey's Pizza	19.126762	72.830001	Pizza Place
7	Amboli	19.1293	72.8434	56433c85498ee4d7ac3bca83	Doolally Taproom	19.135917	72.833094	Brewery
8	Amboli	19.1293	72.8434	4b0587d6f964a520bca322e3	Mainland China	19.140391	72.838033	Chinese Restaurant
9	Amboli	19.1293	72.8434	4f4e4c34e4b027c8742327cf	The Little Door	19.139265	72.833180	Pub

Bubble map showing area wise distribution of no of restaurant venues -



3. From Zomato API (https://developers.zomato.com/api), I retrieved the following for each venue:

- Cost_for_two(Rs.) 6526 non-null int64
- Cuisines 6525 non-null object
- Home_Delivery 6526 non-null bool
- Operational_hours 6518 non-null object
- Restaurant_Location 6511 non-null object
- Restaurant_Name 6526 non-null object
- Rating 6526 non-null float64
- Votes 6526 non-null float64
- Rating_Category 6526 non-null object
- Operational_after_Midnight 6526 non-null bool
- Cuisine_count 6526 non-null int64
- Feature_Count 6526 non-null int64
- Res_Type_Count 6526 non-null int64
- Competitors_in_Location 6511 non-null float64
- Score 6526 non-null float64

	Cost_for_two(Rs.)	Cuisines	Operational_hours	Restaurant_Location	Restaurant_Name	Rating	Votes	Rating_Category	Operational_after_Midnight	Cuisi
0	1500	Finger Food, Continental, European, Italian	12noon – 1am (Mon-Sun)	Kamala Mills Compound	Lord of the Drinks	4.9	1326.0	5	1	
1	800	Pizza	11am – 12:30AM (Mon-Sun)	Malad West	Joey's Pizza	4.6	5974.0	5	1	
2	2500	Seafood	Closed (Mon),12noon – 3pm, 7pm – 12midnight...	Bandra West	Bastian	4.5	1438.0	5	0	
3	1800	Finger Food, Continental	12noon – 1am (Mon-Sun)	Lower Parel	Tamasha	4.9	3275.0	5	1	
4	450	North Indian, Street Food, Fast Food, Chinese	12noon – 4pm, 7pm – 11:45pm (Mon-Sun)	Vashi	Bhagat Tarachand	4.1	1422.0	4	0	

2.2 Data Cleaning

From figure 1 and figure 2, we can clearly see that some venues from the two APIs do not align with each other. Thus, I decided to combine them using their latitude and longitude values.

To combine the two datasets, I had to check that the latitude and longitude values of each corresponding venue match. After careful analysis, I decided to drop all corresponding venues from the two datasets that had their latitude and longitude values different by more than 0.0004 from one another. Thus, I rounded both the latitude and longitude values up to 4 decimal places. Then, I calculated the difference between the corresponding latitude and longitude values and saw if the difference was less than 0.0004 which should ideally mean that the two locations are the same. This removed many outliers from the two datasets. Once this was done, I observed that there were still some venues which were not correctly aligned.

They can be categorised as follows:

1. There are venues that have specific restaurants/cafes inside them as provided by Zomato API (Pizza Hut inside Elante Mall).
2. Two locations are so close that they have practically same latitude and longitude values (The Pizza Kitchen and Zara).
3. Some venues have been replaced with new venues (Underdoggs has now been replaced by The Brew Estate).

Venues belonging to category 1 and 3 are perfect to keep. However, the venues that belong to category 2 should be dropped. After careful inspection and removal, the final dataset had a total of 1498 venues with which we can work.

Merged dataset contains following columns -

'Cost_for_two(Rs.)', 'Cuisines', 'Home_Delivery', 'Operational_hours', 'Restaurant_Location', 'Restaurant_Name', 'Rating', 'Votes', 'Rating_Category', 'Operational_after_Midnight', Cuisine_count', 'Feature_Count'	'Res_Type_Count', 'Competitors_in_Location', 'Score', 'Neighborhood', 'Latitude', 'Longitude', 'Venue_id', 'VenueLatitude', 'VenueLongitude', 'VenueCategory'

	Cost_for_two(Rs.)	Cuisines	Home_Delivery	Operational_hours	Restaurant_Location	Restaurant_Name	Rating	Votes	Rating_Category	Operational_after_M
0	800	Pizza	False	11am – 12:30AM (Mon-Sun)	Malad West	joey's pizza	4.6	5974.0	Excellent	
1	800	Pizza	False	11am – 12:30AM (Mon-Sun)	Malad West	joey's pizza	4.6	5974.0	Excellent	
2	800	Pizza	False	11am – 12:30AM (Mon-Sun)	Malad West	joey's pizza	4.6	5974.0	Excellent	
3	800	Pizza	False	11am – 12:30AM (Mon-Sun)	Malad West	joey's pizza	4.6	5974.0	Excellent	
4	800	Pizza	False	11am – 12:30AM (Mon-Sun)	Malad West	joey's pizza	4.6	5974.0	Excellent	

5 rows × 22 columns

3. Methodology and Exploratory Data Analysis

- As a first step, I retrieve the venues in Mumbai from Foursquare and Zomato APIs. I extract the location data from the Foursquare API for all venues up to a distance of 4 kilometers from the center of Mumbai. Using this, I fetch the venue information including price and rating data from Zomato API.

- Using data cleaning, the dataset from the two APIs will be combined based on the venue names, latitude, and longitude values. One to one matching and careful data inspection would be used to remove any remaining outliers such as multiple venues at the same location from the two datasets. The final data will include the venue name, category, address, latitude, longitude, rating, price range, and average cost per person.

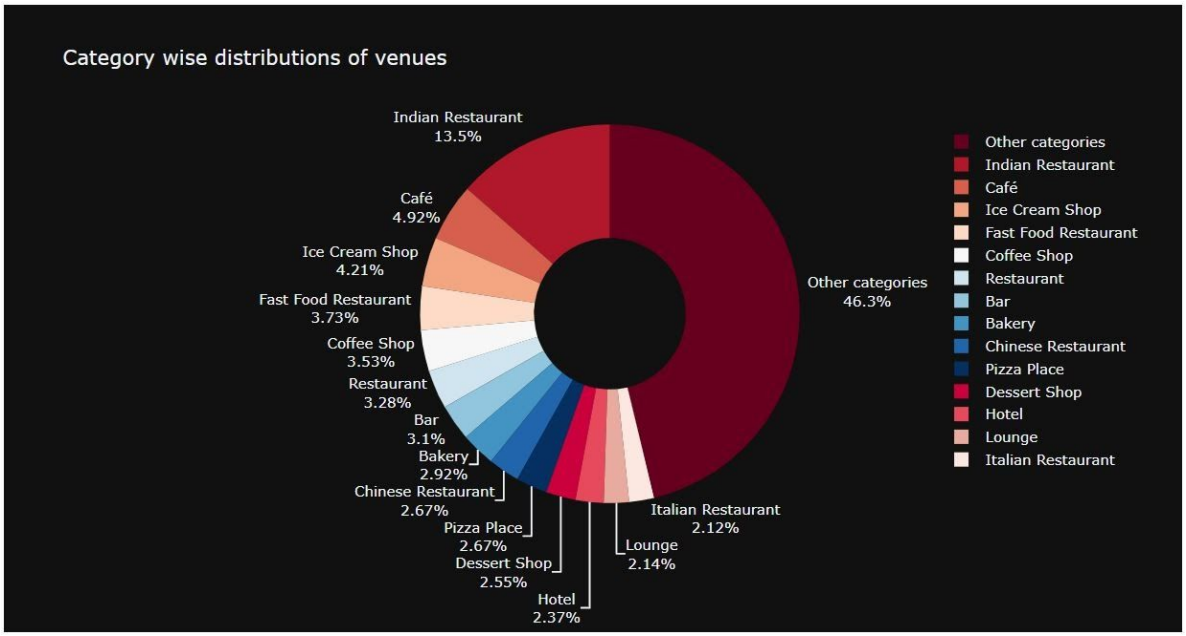
- Using this dataset, I begin by analyzing the top venue types that exist in Mumbai. I will then explore the venues on maps. This will allow us to better understand the location of various venues and the places where many venues co-exist and create place worth visiting. I'll also explore the venues based on the ratings and price range of various venues. The venues will be plot using proper color coding such that a simple glance at the map would reveal the location of the venues as well as give information about them. I aim to identify places which can be recommended to visitors based on their price and rating preferences. I'll also cluster the venues and see if we can draw meaningful information out of what kind of venues exist in Mumbai.

- As a final step, I will analyse these plots and try to draw conclusions on what places can be recommended to visitors. I'll discuss my findings and any inferences I can drawn.

3.1 Categories

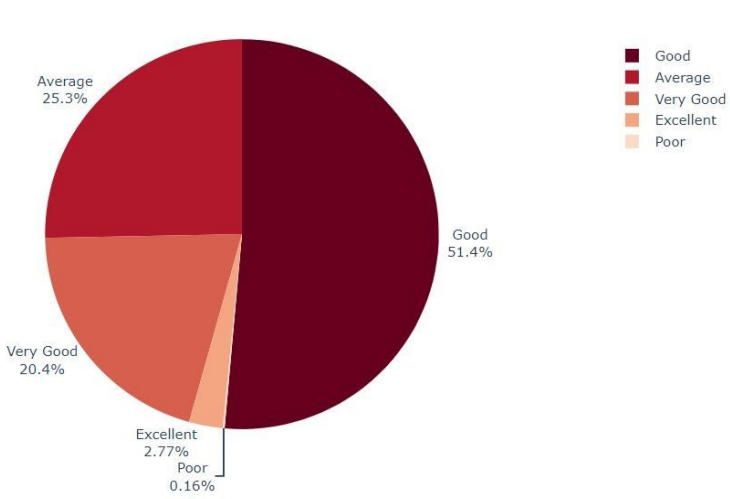
I begin my analysis by taking a look at the various categories of venues that exist in Mumbai. As there are many restaurants, I believe that the majority venues shall include restaurants.

Venue Categories -
Indian Restaurant', 'Café', 'Ice Cream Shop','Fast Food Restaurant', 'Coffee Shop', 'Restaurant', 'Bar' ,'Bakery', 'Chinese Restaurant', 'Pizza Place', 'Dessert Shop','Hotel', 'Lounge', 'Italian Restaurant', 'Pub', 'Seafood Restaurant', 'Juice Bar', 'Multiplex', 'Asian Restaurant','Snack Place', 'Beach', 'Donut Shop', 'Gym / Fitness Center','Vegetarian / Vegan Restaurant', 'Sandwich Place','Scenic Lookout', 'Gym', 'Shopping Mall', 'Deli / Bodega', 'Park', 'Spa', 'Cricket Ground', 'Diner', 'Theater', 'Clothing Store','BBQ Joint', 'Department Store', 'Food Truck' and many more..



We see that the majority venues are actually Indian Restaurants. This is closely followed by Cafe'. For someone who is visiting Mumbai and loves either Cafes or Indian Restaurants, they'd surely love their stay.As we can see food outlet categories have more distributions than any other categories,we will collect all types of food outlet venues and restaurants for analysis.

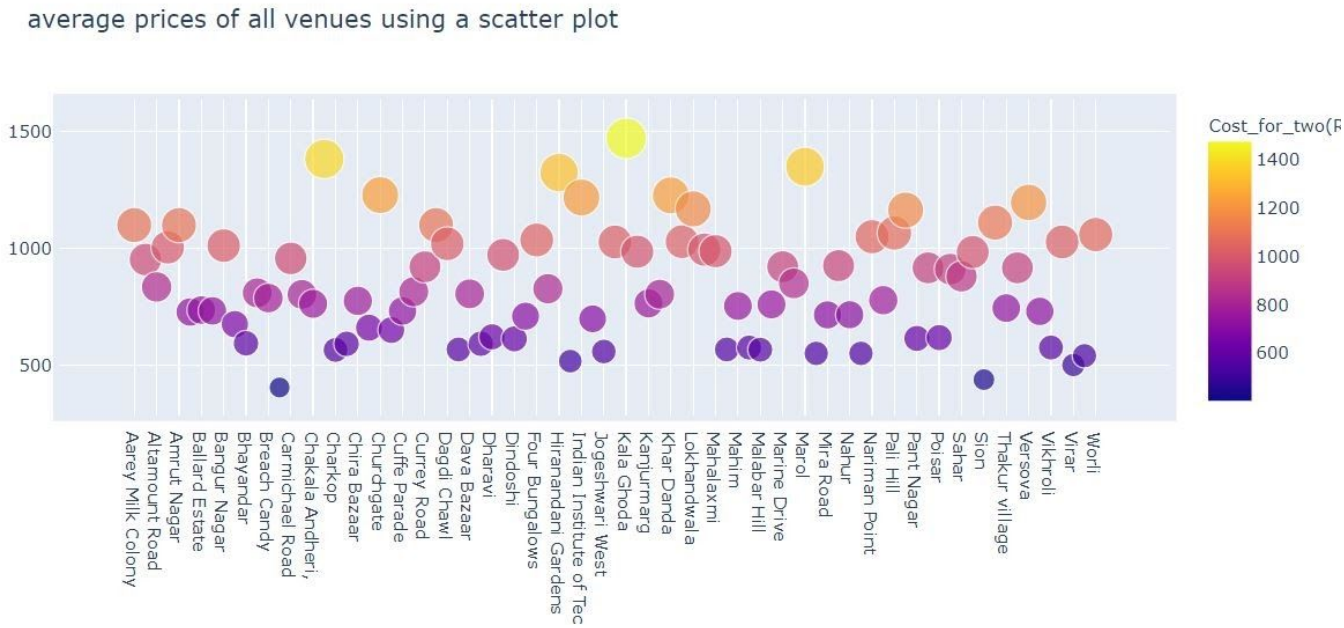
3.2 Rating



Overall, Mumbai on an average has good rating for its venues.

3.3 Price

Plot of venues with different prices -



Taking a look at figure , reveals that the majority venues have an average cost of Rs 800 to Rs 1000 for 2 persons.

Figure includes all the venues where high priced venues are marked by yellow and orange while the low priced venues are marked with blue,violet. From the plot, we observe that venues Marol ,Sion are primarily lower priced. The venues near Kalaghoda, IIT campus have steep prices. Mahalakshmi venue seems to have a mix of both high priced and low priced venues.

3.4 Prediction of Rating -

We have columns/Feature called Online order and Book table which are Categorical variables and for machine learning to model work we should input numerical values to perform. hence use Label Encoding on these 2 Features that encode Yes/No as 0/1.

Drop the unnecessary columns that wont play much in deciding the Rating of the Restaurants. Also do apply Label Encoding on Features like “Location”, “Restaurant type”, and “Cuisines”. After Encoding split the Dataset to X and Y variables and again split to Train and Test sets of 80% and 20%. Apply Standardisation on Dataset as we have different scale ranges for different Features. Hence after applying Standard scaling it will bring all the values to a common range which is easy for model to compute and makes computation fast.
Target Value (Rating) is correlated well with the 'votes', 'feature_count' and 'cost_for_two', with correlation values being 0.51, 0.31 and 0.33 respectively.
After applying Several Regression models such as DecisionTreeRegressor, DecisionTreeClassifier, RandomForestClassifier and SVM, DecisionTreeRegressor has yielded us Best Accuracy compared to all the other models which is of 96.09%.
You can even see the Predicted vs Actual to see how well our model is predicting the ratings of the Restaurants. you can see below how close they are -

Actual	Predicted
2.8	2.8
4.4	4.2
3.2	3.2
2.6	2.6
2.4	2.4
3.0	3.0
4.6	4.6
2.2	2.2
3.0	3.0
3.4	3.4

Model accuracy can be improved by using/applying Hyperparameter optimization, Ensemble methods, Cross validation.

4. Results and Discussion

The basic idea of analyzing the Zomato dataset is to get a fair idea about the factors affecting the establishment of different types of the restaurant at different places in mumbai, aggregate rating of each restaurant, Mumbai being one such city has more than 12,000 restaurants with restaurants serving dishes from all over the world. With each day new restaurants opening the industry hasn't been saturated yet and the demand is increasing day by day. Inspite of increasing demand it, however, has become difficult for new restaurants to compete with established restaurants. Most of them serving the same food. After collecting data from the Foursquare and Zomato APIs, we got a list of 6495 different venues. However, not all venues from the two APIs were identical. Hence, we had to inspect their latitude and longitude values as well as their names to combine them and remove all the outliers. This resulted in a total venue count of 1498.

We identified that from the total set of venues, majority of them were Cafes and Indian Restaurants. A visitor who loves Cafes/Indian Restaurants would surely benefit from coming to Mumbai.

While the ratings range from 1 to 5, majority venues have ratings close to 3. This means that most restaurants provide good quality food which is liked by the people of the city, thus indicating the high rating. These clusters also have very high ratings (more than 3).

When we take a look at the price values of each venue, we explore that many venues have prices which are in the range of Rs 800 to Rs 1000 for one person. However, the variation in prices is very large, given the complete range starts from Rs 200 and goes uptil Rs 2000.

In this model, we have considered various restaurants records with features like the name, average cost, locality, whether it accepts online order, can we book a table, type of restaurant, no od cuisines available.This model will help business owners predict their rating on the parameters considered in our model and improve the customer experience. Different algorithms were used but in the end the final model is selected on which **'DecisionTreeRegressor**' gives the highest accuracy compared to others.

5. Conclusion

Thus we can conclude that , a number of features about existing restaurants of different areas in a city and analyses them to predict rating of the restaurant. This makes it an important aspect to be considered, before making a dining decision. Such analysis is essential part of planning before establishing a venture like that of a restaurant. Lot of researches have been made on factors which affect sales and market in restaurant industry. Various dine-scape factors have been analysed to improve customer satisfaction levels. If the data for other cities is also collected, such predictions could be made for accurate.