Find The Top-Rated Restaurants using Classification Techniques

SUBMITTED TO: SAVITA SEHARAWAT

SUBMITTED BY: AMANDEEP KAUR

STUDENT ID: 0773293

Acknowledgement

I would like to thank our instructor Mrs. SAVITA SEHARAWAT for her tremendous direction and assistance in the completion of my capstone project. Your useful advice and suggestions were helpful to me during the project's completion. I would not have been able to complete this project without your help and cooperation. In this aspect, I am eternally grateful to you.

Overview

Bangalore is the capital and biggest city of the Indian territory of Karnataka. Basically, Bangalore has an interesting food culture. Eateries from everywhere, the world can be found here in Bengaluru, with different sorts of foods. Overall, it is possible that Bangalore is the best spot for foodies.

The food business is always at a rise in Bangalore, with 12,000 or more eateries presently active in the city, the number is yet expanding. The developing number of cafés and dishes in Bangalore draws in me to assess the information to get a few experiences, some interesting facts, and figures. So, there is a big challenge for a new business to find out location which would be always crowded and in demand.

So, that’s why I choose the Zomato Bengaluru restaurant dataset for my capstone project. There is a huge problem for opening a business in Bengaluru. I will try to find out top rated restaurant area in the Bangalore city. As, all formulation and data visualization are done by using Python Programming Language.

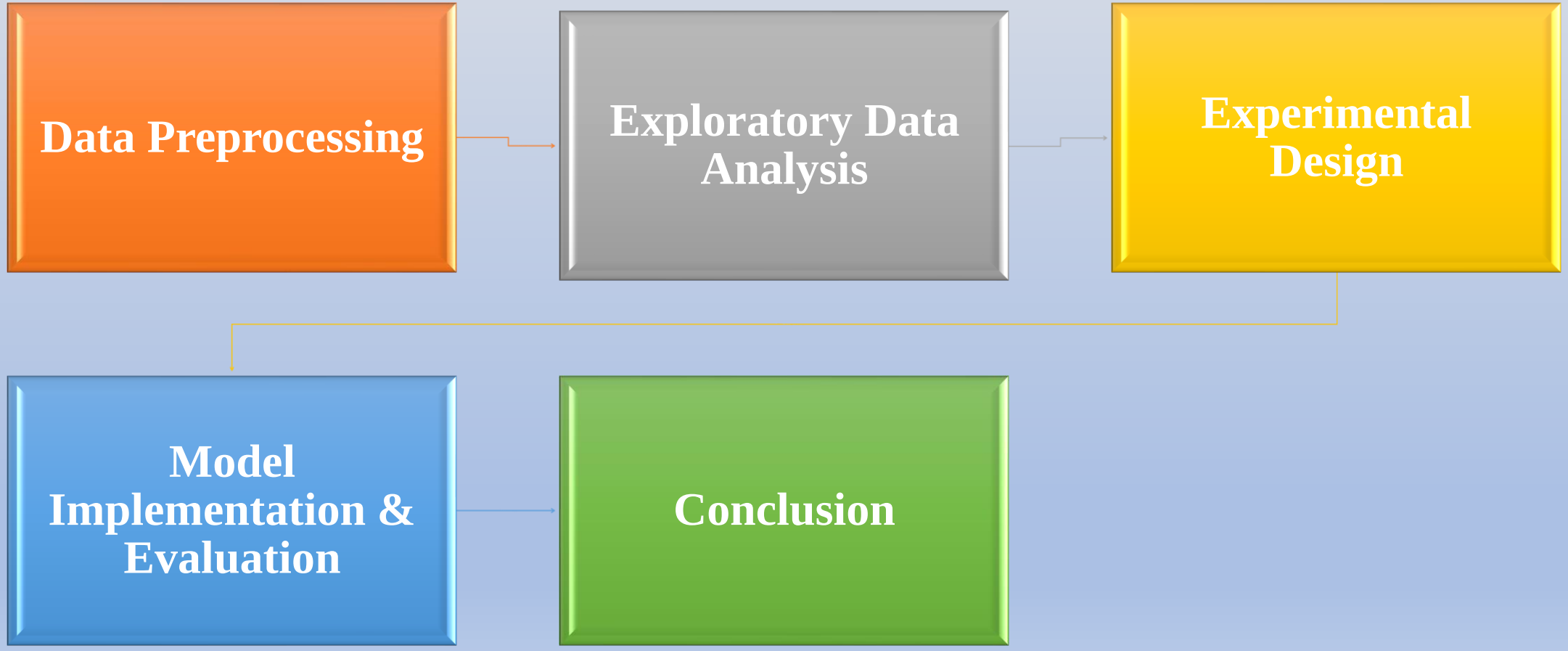
Introduction

My dataset is based on the restaurants in Bengaluru. This dataset contains 51717 records and 17 attributes. This dataset basically contains the information regarding the restaurants dine in, takeout and online order options, reviews, and type of restaurants like casual dining, pubs, bars and café, type of cuisine and all.

The main goal of this project to find best top-rated restaurants in Bengaluru city and does the cost of food affect the rating of restaurants. I will try to figure out which cuisines are famous, about dine-in and takeout option in Bengaluru. These days, mostly people do not have time to cook food at home, so they are preferring restaurant food. With such an overwhelming demand of restaurants, it has become important to study the demography of a location. As, its is a big challenge for a new business to find out location which would be always crowded and in demand.

This dataset also contains the reviews for each of the restaurant and cost, from which I will find out the overall rating for the area that would be helpful to find out the top-rated location for setting up a new business. Moreover, I can also find which cuisine is popular in the area and many more.

Methodology



Data Detail

- I have contained 51717 records and 17 attributes from which I had 16 object attributes and 1 numeric attribute.
- In this dataset I have all the objects at the starting so, I must assign the appropriate datatype to the attributes for the better visualization.
- After assigning the appropriate data type to the attribute, I have 2 objects, 5 categorical attributes, and 5 numeric data types. At the end I have total 12 attributes left for my project.
- As, I dropped 5 attributes, because that attribute was not giving useful information. The dropped attributes are URL, Phone, address, Dish liked and Menu item.

Information about objects, categorical & numeric attributes

Attribute	Description	No. of Levels
Restaurant Type	It contains the information on type of restaurant.	93
Location	It carries information area in which a restaurant is situated. I have 3 best locations.	93
Cuisines	It represents the type of cuisine in the restaurants.	2723
Service Type	It represents the type of meal in the restaurant.	7
City	It carries the information of the area restaurant is listed. We have top 3 cities.	30

Attribute	Description
Name	It represents the name of the restaurant in Bengaluru.
Reviews	It has the list of tuples, which are containing the reviews for the restaurant, and each tuple carries two values, rating and review by the customer who visited restaurant.

Attribute	Description
Online order	It means that restaurant is accepting online order or not.
Book Table	It carries the table booking options available in the restaurant or not.
Votes	It contains the total number of ratings for the restaurant with actual date and time.
Rate	It contains the rating between 0 to 5.
Cost	It carries the information of cost of food for 2 persons.

Data Preprocessing

Attribute	Missing Values
Rate	7757
Location	21
Restaurant Type	227
Cuisines	45
Cost	345

You can see that total 5 attributes (rate, location, restaurant type, cuisines, cost) have missing data which I have to remove because it imbalanced our dataset. I replaced the Na values for these attributes with the help of mean and mode.

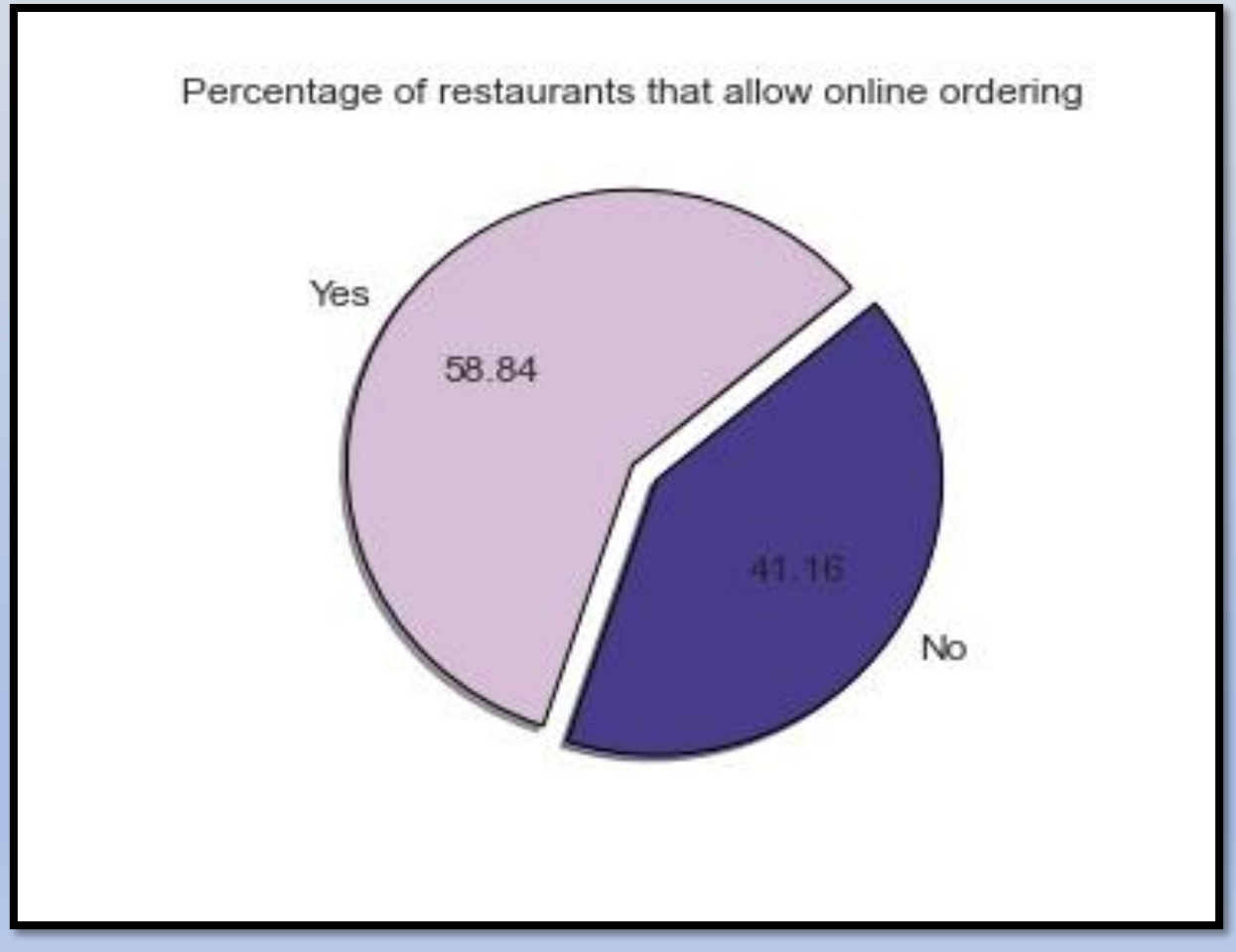
Rate, Cost: In these attributes, I replaced missing value with mean.

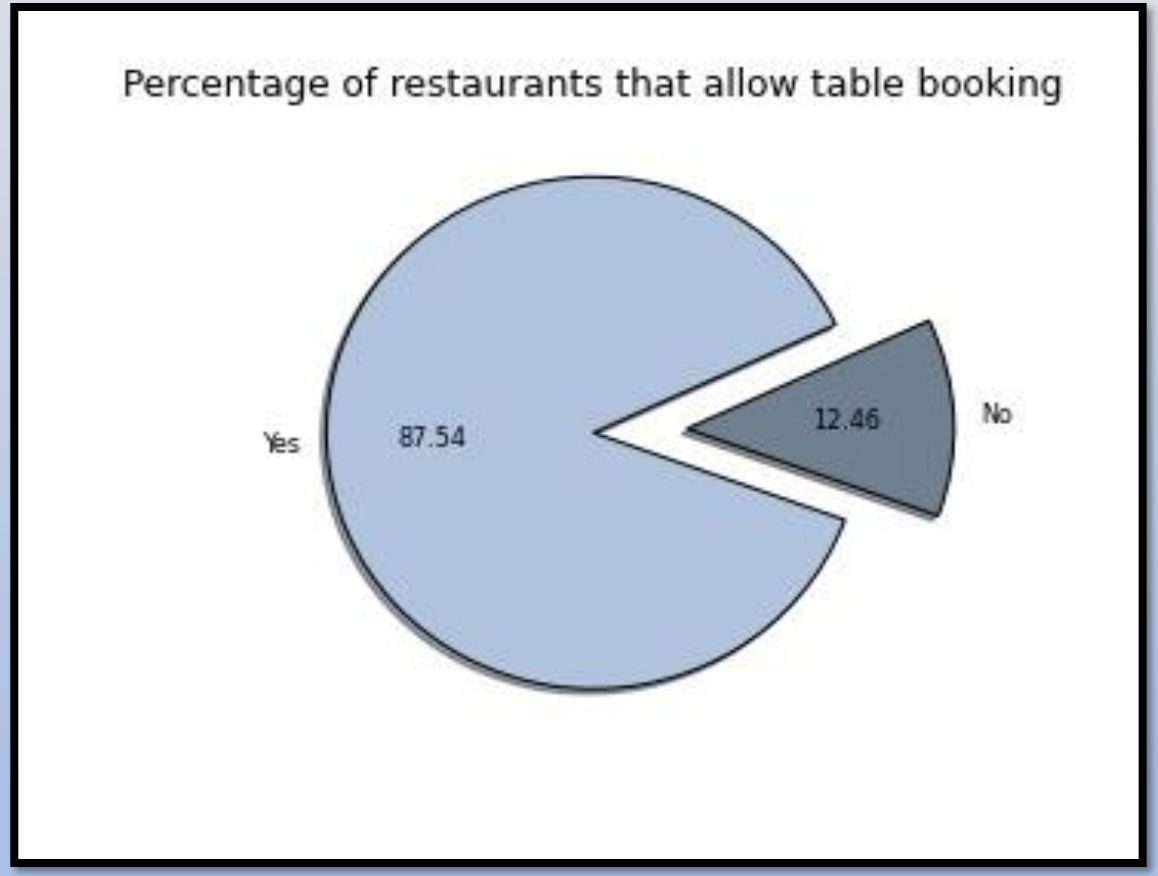
Location, Restaurant type, Cuisines: For these attributes, I fill the missing value with mode.

Exploratory Data Analysis

How many restaurants have online order options?

This pie chart gives the information about the restaurants doing online order or not. More than half (58.84%) of the restaurants in the Bangalore city are providing option for online order. However, 41.16% restaurants are not having option of online order.



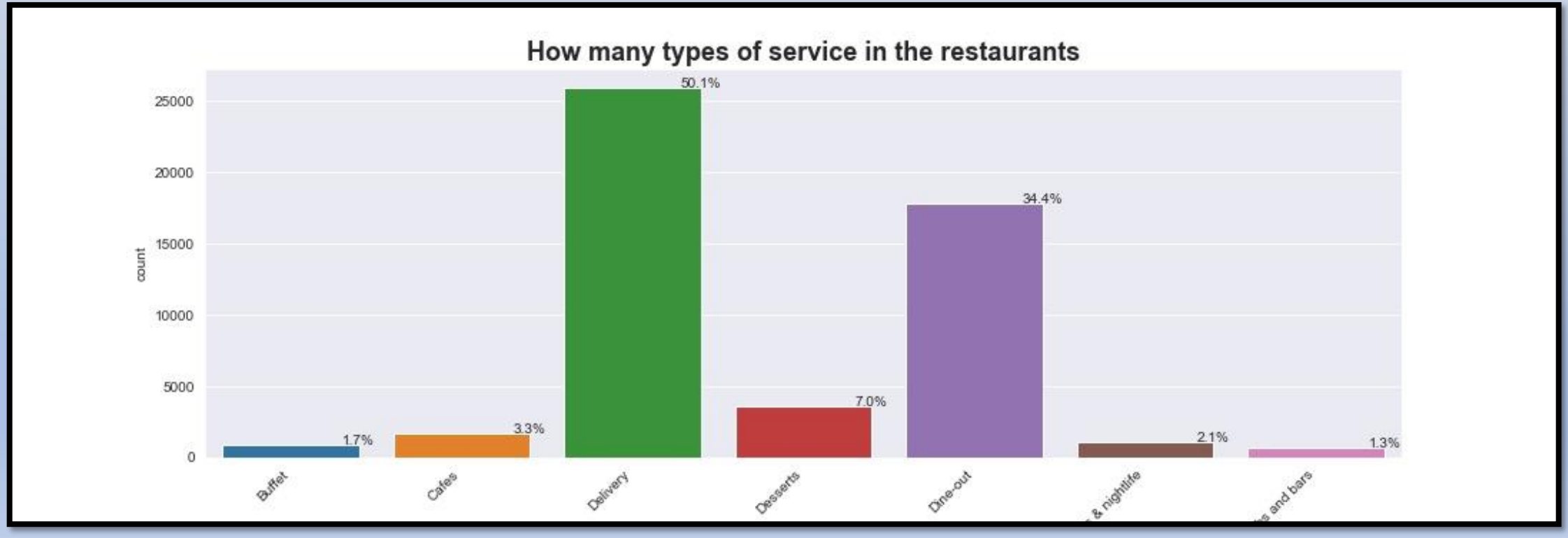


How many restaurants have table-booking option?

It shows how much restaurants are allowing to pre booking of the table or not. More than half (87.54%) of the restaurants providing table booking option and only 12.46% of the restaurants are not having table-booking option.

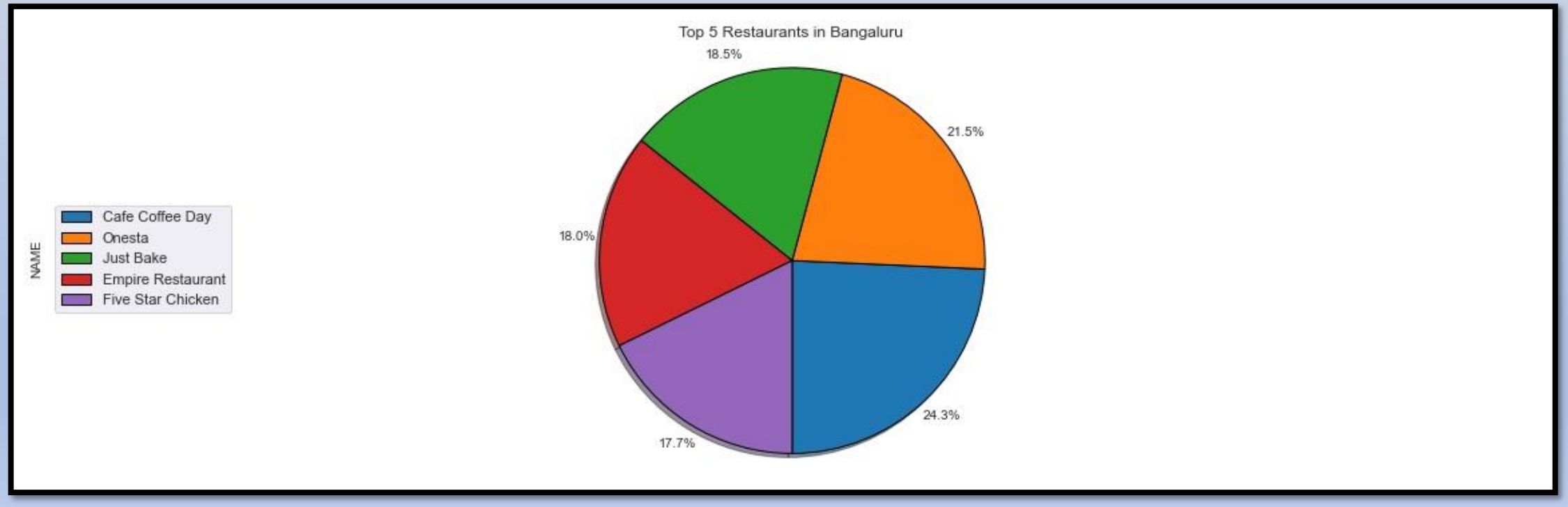
How many types of service are present in the restaurants?

There are total 7 type of services in the restaurants such as buffet, cafes, delivery, desserts, dine-out, clubs & nightlife, pubs and bars. Top-notch service is the delivery option. From these services, customers like to Dine-out and delivery type of service, as compared to the other types like buffet, cafes and bars etcetera.



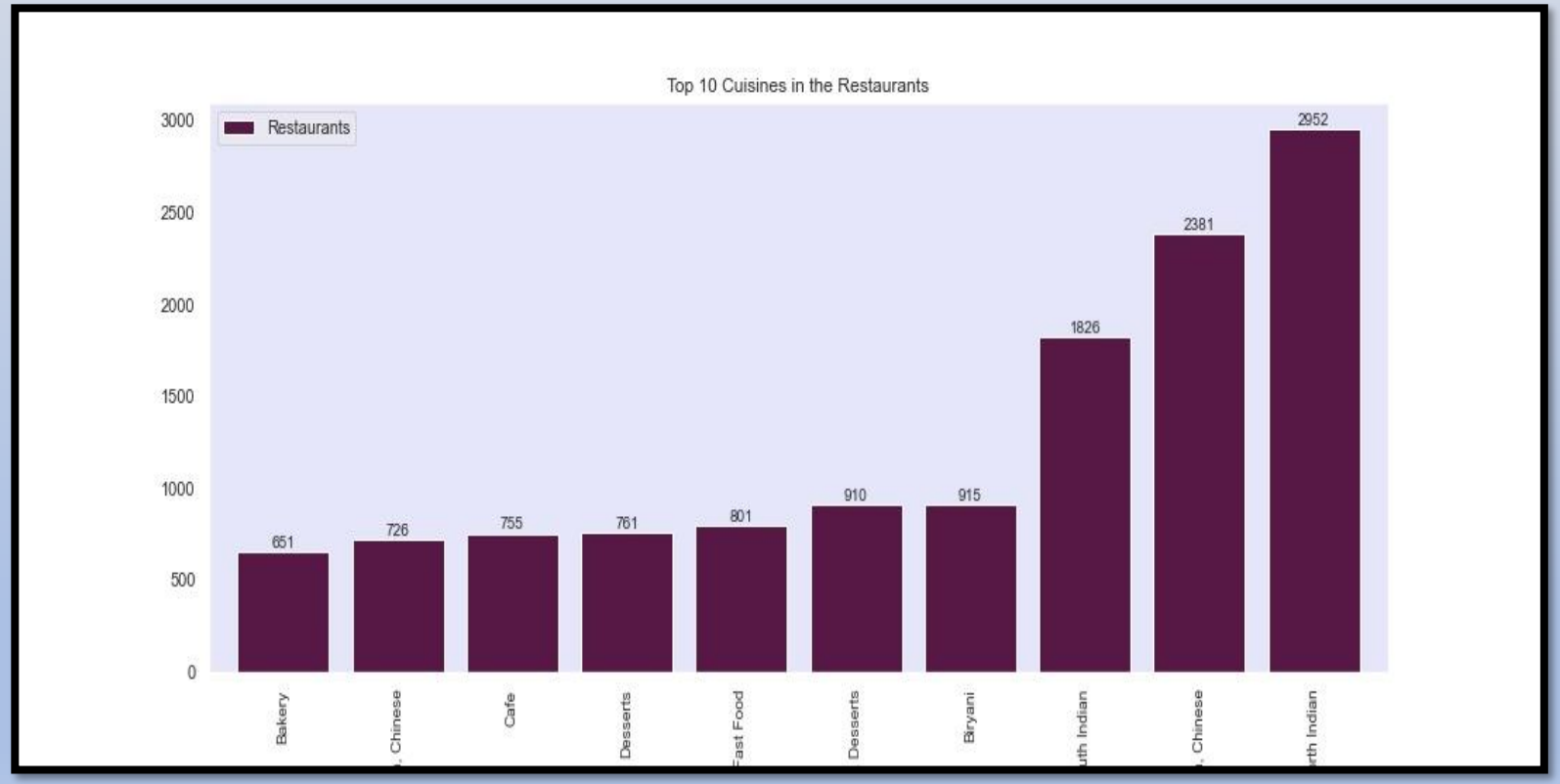
Find the top 5 restaurants name in Bengaluru

This pie chart gives the information about the top five restaurants in Bangalore. These restaurants are namely, Café coffee day, Onesta, just bake, empire restaurant and five-star chicken and the top first restaurant is Café Coffee Day (24.3%); second one Onesta (21.5%), third one is Just Bake (18.5%).



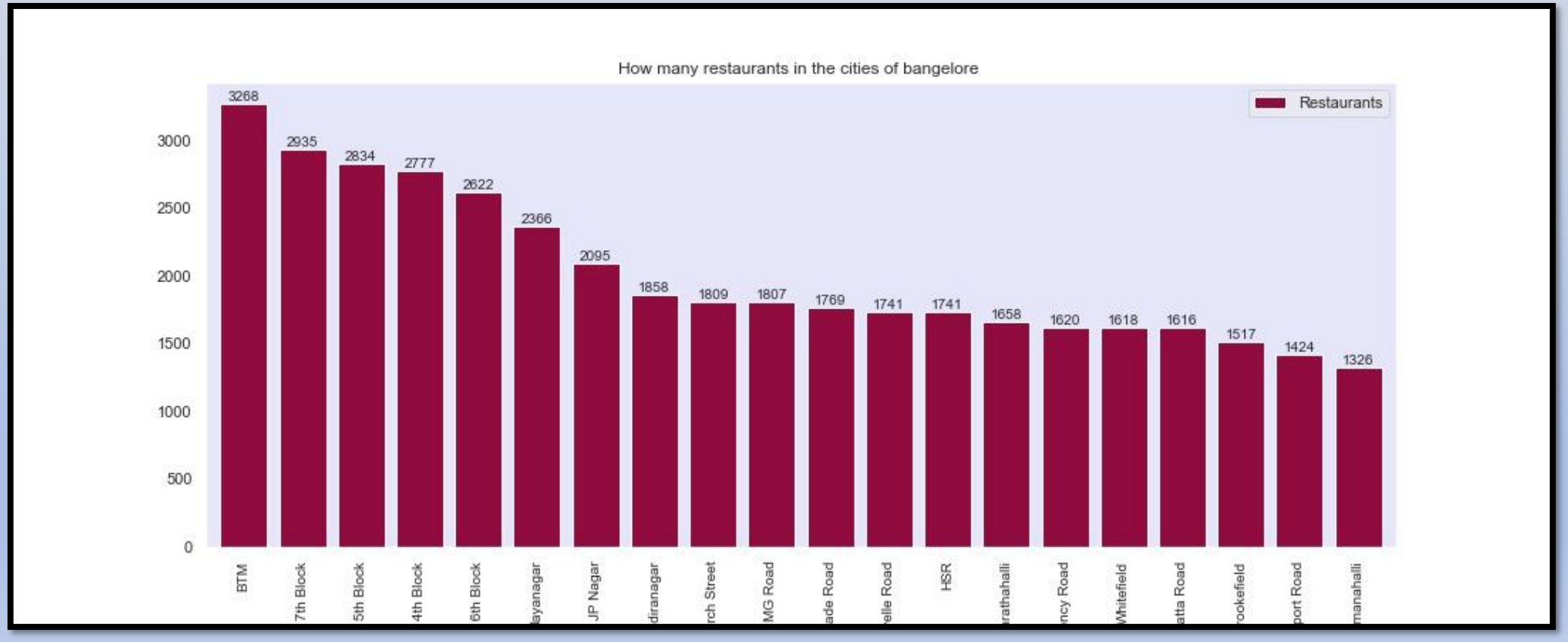
What kind of cuisine is most popular in the locality?

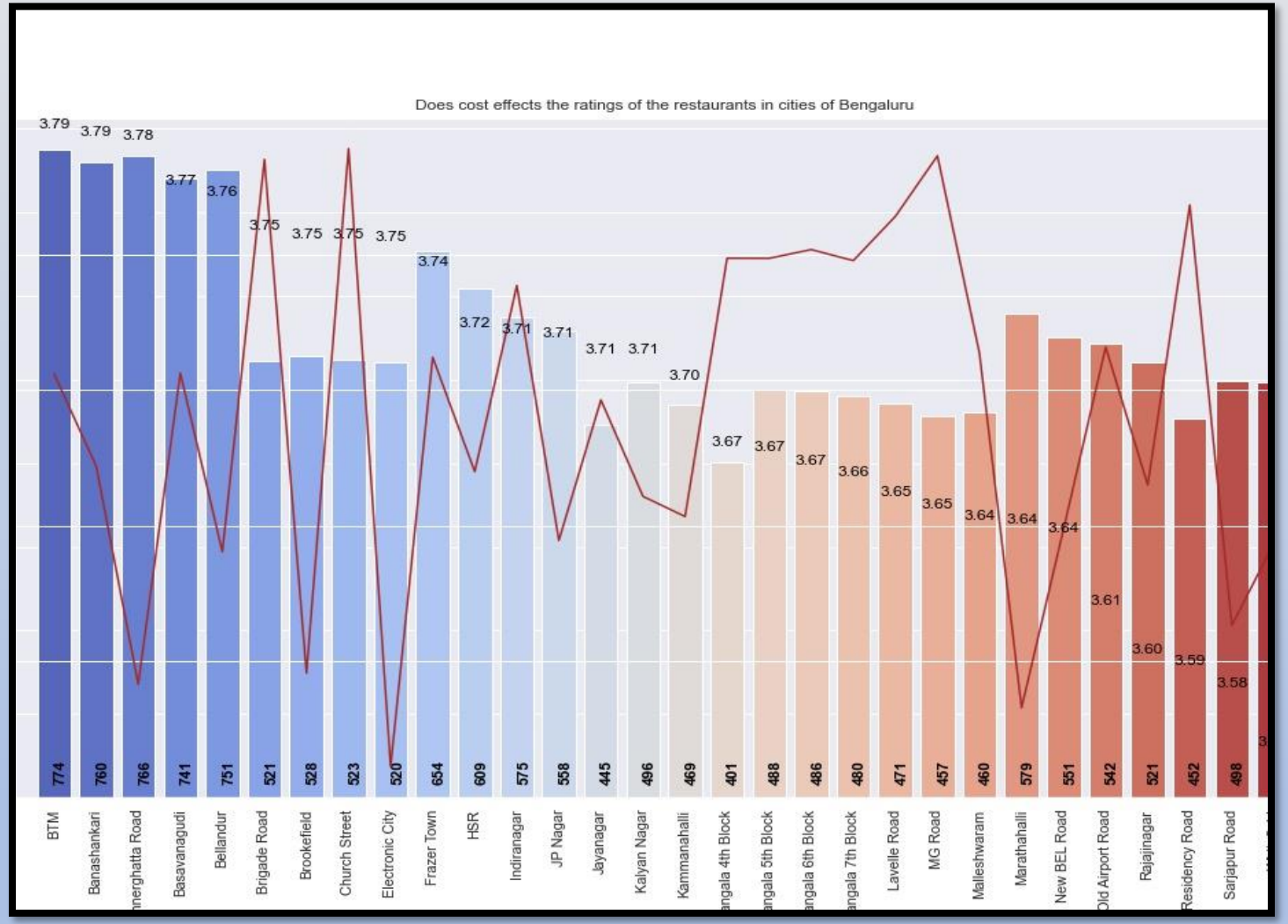
The bar chart shows the top 10 cuisines which are mostly liked by the customer. The very first cuisine, which is popular, is the North Indian cuisine. Second one is Chinese and the third one is south Indian. However, among top 10 cuisines, Bakery is least famous cuisines in people of Bangalore.



How many numbers of restaurants in each neighborhood?

This bar chart represents the count of restaurants in each neighborhood of the Bangalore city. BTM (3268) is having highest number of restaurants. On the contrary, old airport road and Kammanahalli locations are having lowest number of restaurants.





Does cost effects the ratings of the restaurants in cities of Bengaluru?

Above are the top 30 cities in which customers likes to dine in and takeout. In the mentioned cities, the cost of 2 person lies in between 774 to 448 and rating is varied between 3.79 to 3.52. By analysing the above graph, we can say that cost does not affect the rating of the restaurants in Bengaluru. Moreover, we find out top rated restaurants are present in the BTM city and the minimum rating of all these restaurants are lies under between 3.79 to 3.52.

Outliers Removed by IQR Score Method

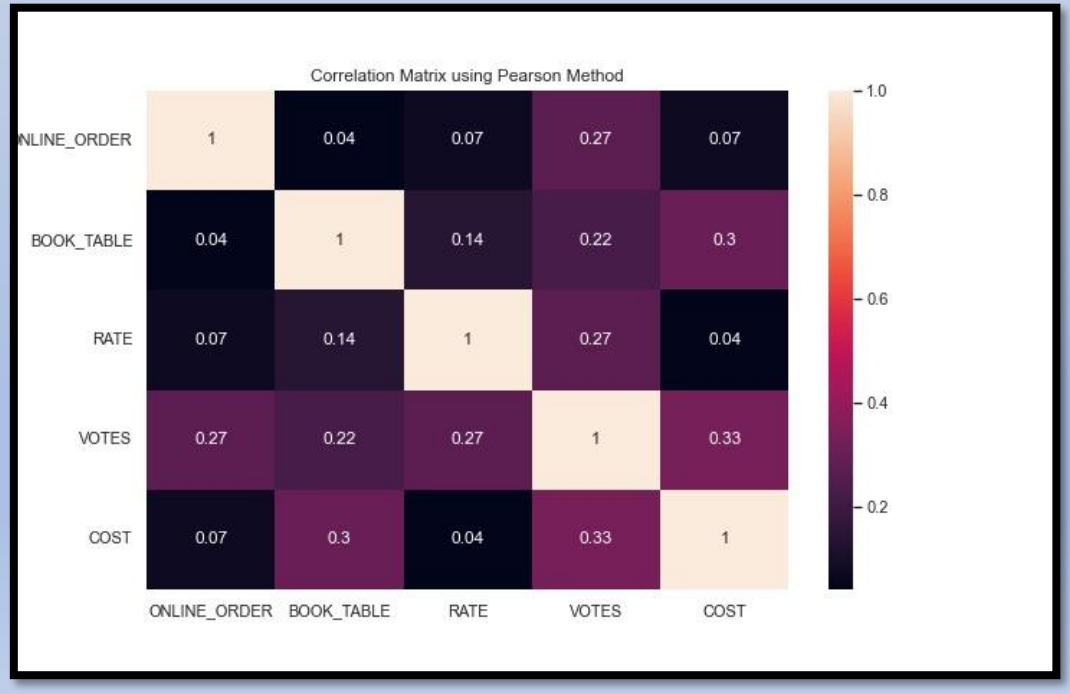
There were some outliers in the rate, cost and votes attribute. So, I decided to used IQR method to display data and outliers (shape of the data) but in order to be get a list of identified outlier, we will need to use the mathematical formula and retrieve the outlier data.

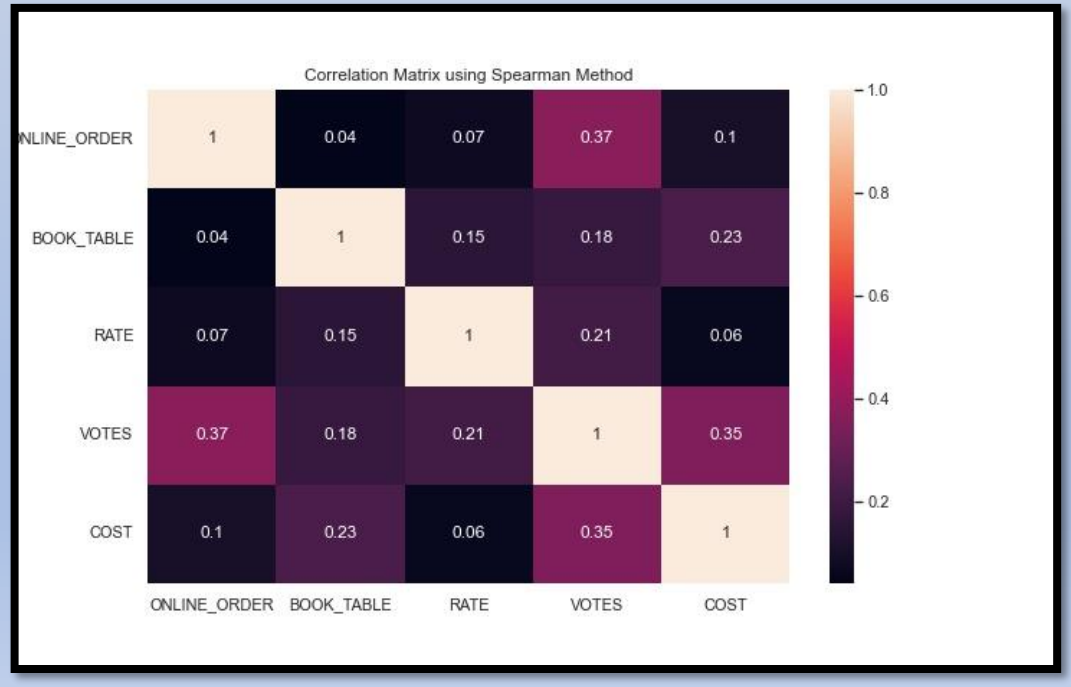
That's why we use IQR score method to remove the outliers from our dataset.

Correlation matrix using Pearson and Spearman method

These two methods are very popular. Generally, we can use Pearson's correlation when we think the variables relationship is linear and Spearman's correlation when we think the relationship is monotonic.

The below charts illustrates the relation between two variables using two different methods of correlation matrix, which shows how the change in one variable affects another variable. The value varies between -1 to 1. There is no strong correlation has been observed between these all the variables.





One-Hot Encoding & Min-Max Scaling

One-hot encoding

One-hot encoding technique represents the categorical data into binary vectors. It is a common process before performing classification techniques. I performed one hot encoding technique on our categorical attributes- Online order, Book table, Location, Restaurant type, Cuisines and Service type.

Min-Max scaling

Min-Max scaling is a normalization technique that enables us to scale data in a dataset to a specific range using each feature's minimum and maximum value. In this we subtract the Minimum from all values, thereby marking a scale from Min to Max. Then divide it by the difference between Min and Max. The result is that our values will go from zero to 1.

Experimental Design

Train-Test-Split

The train-test-split function is for splitting a single dataset for two different purposes: training and testing. The testing subset is for building your model. The testing subset is for using the model on unknown data to evaluate the performance of the model. In this 70% of the data is used by training set and 30% by testing.

Under sampling strategy

This strategy refers to a group of techniques to balance our dataset. It removes the example from the training dataset, which belongs to the majority class. I performed the under-sampling technique on my dataset, because before that my data was not stable. After performing this technique, I made my data stable.

Model Implementation & Evaluation

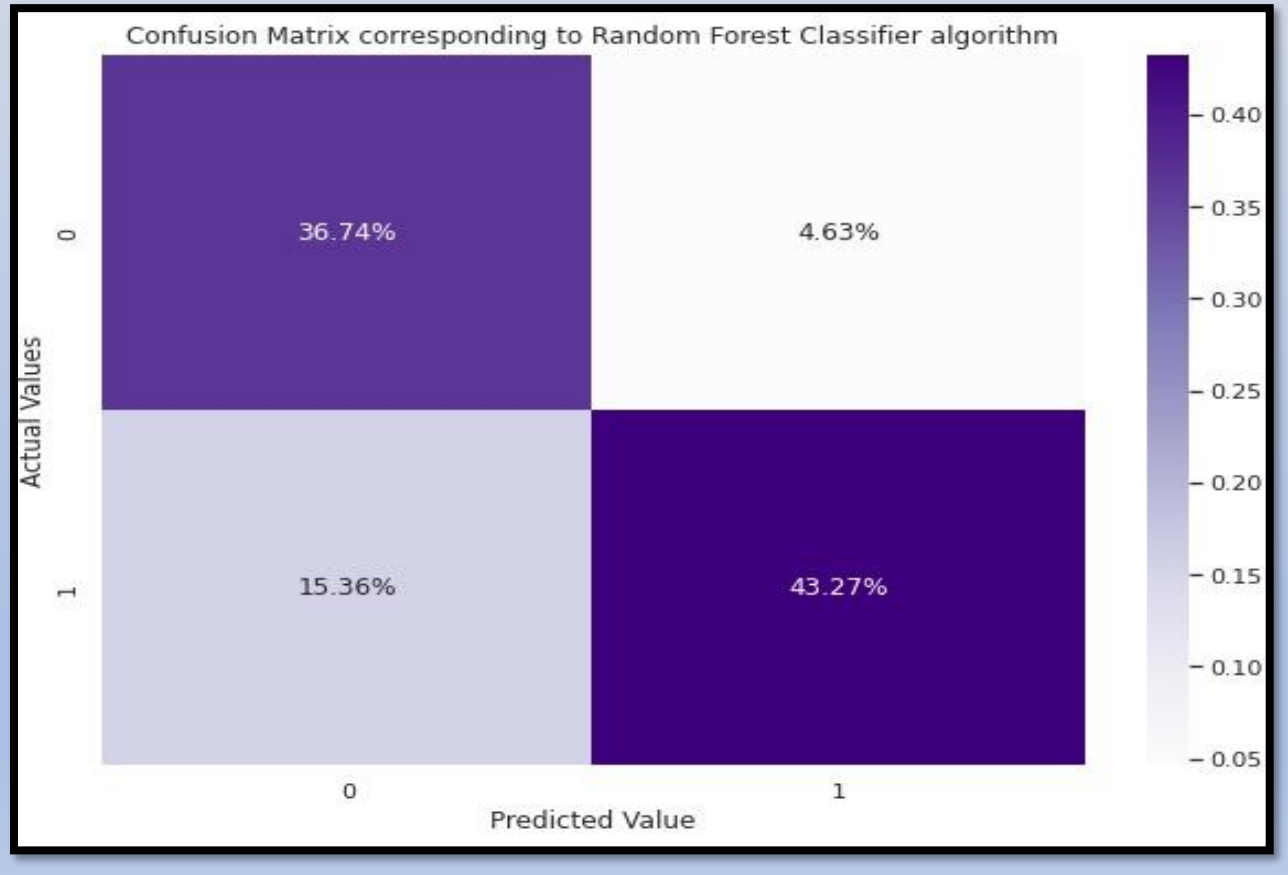
Model	Train-Test-Split
Naive Bayes	55.012322
Logistic Regression	60.847444
KNN	72.986724
Decision Tree	79.569123
Random Forest	80.411797

This is comparison of accuracies of all the models. I use linear regression, naïve bayes, knn, decision tree and random forest classifier for our dataset. Out of all models, Random Forest Classifier is the best-fit model on our dataset which gives the 80% accuracy for our dataset.

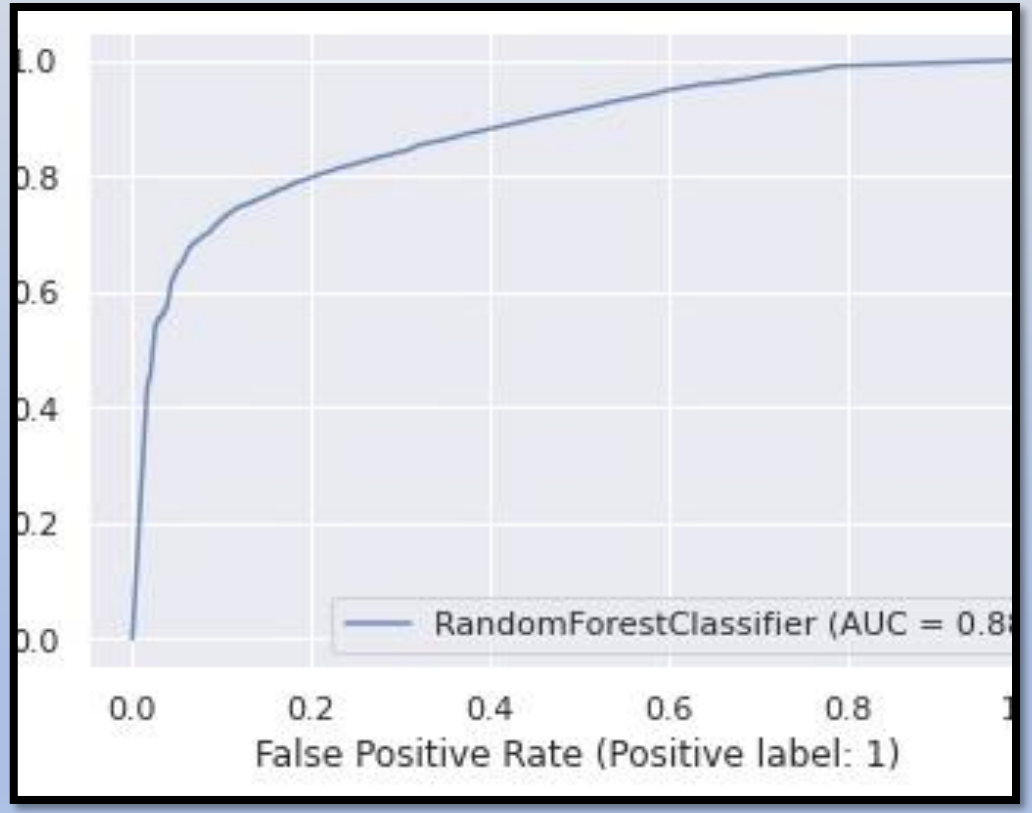
- Logistic Regression: It perform to predict the dependent variable value based on the given independent variable. Therefore, this technique shows the linear relationship between input and output variable.
- K Nearest Neighbors: It is a supervised machine-learning algorithm. This algorithm represents the k nearest neighbor. It is used for classification and regression both.
- Gaussian Naive Bayes: It is a special type of NB algorithm. It is used when the features having continuous values
- Random Forest: A supervised Machine Learning Algorithm is used widely in Classification and Regression problems. It makes decision trees on different samples.
- Decision Tree Classifier: It is a non-parametric supervised learning algorithm used for classification and regression. The mail goal is to create a model that predicts the value of a target variable.

Confusion matrix Corresponding to Random Forest Classifier

The confusion matrix shows the accuracy of the optimal model. It predicted that correspond to 0, only 4.63% is wrong and 36.74% is right. Whereas corresponds to 1 is 15.36% are wrong values and 43.27% is right. So almost 20% is predicted wrong and 80% is right.



ROC curve of the optimal model Random Forest Classifier Algorithm



ROC curve is short of Receiver Operating Characteristics. Generally, if the AUC value lies between 0.5 to 1, where 0.5-0.6 auc denotes a bad classifier, 0.7 to 0.8 is considered acceptable, and 1 denotes an excellent classifier and our model gives the auc 0.8, which is much closed to 1. It means our roc curve gives the fair auc score.

Conclusion

My dataset is all about the restaurants in the cities of Bangalore and my dataset is available on the Kaggle. I collected data from CSV file from the Kaggle, from which half of values were missing, and I did not throw up all values, instead of removing NULL values, I tried to fill the values with mean and mode. I have done exploratory data analysis to answer all the research questions. I used one-hot encoded features; I also normalized our data with the help of min-max scaling, I also do train-test-split and do under sampling strategy to stable my dataset and tried different classification models.

Random Forest classification is the best fit model of our dataset, so we made confusion matrix corresponding to RFC and made the ROC curve with our optimal model. Moreover, we find out top rated restaurants are present in the BTM city and the minimum rating of all these restaurants are lies under between 3.79 to 3.52. Also, we can say that the average rating of new restaurants in Bangalore would be lies in between the 3.

Thankyou