**ZOMATO ANALYSIS**

1.Objective

2. Details about the data

3. Loading dataset and required libraries

4. Data Cleaning

* Deleting redundant columns.
* Renaming the columns.
* Dropping duplicates.
* Cleaning individual columns.
* Remove the NaN values from the dataset
* Some Transformations

5. Data Visualization: Using plots to find relations between the features.

* Plot the Occurrence of Rating
* Restaurants delivering Online or not
* Restaurants allowing table booking or not
* Table booking Rate vs Rate
* Best Location
* Top Foodie Areas in Bangalore
* Most Famous Cuisines
* Types of Services
* Relation between Type and Rating
* Cost of the Restaurant
* People mostly buying food rate
* Budget Friendly Hotel in Bangalore
* Costly Hotel in Bangalore
* Most famous restaurant chains in Bengaluru

6. Regression Analysis

* Correlation matrix
* Linear Regression
* Decision Tree Regression
* Random Forest Regression
* Extra Tree Regression

**OBJECTIVE:**

The basic idea is analysing the Business Problem of Zomato to get a fair idea about the factors affecting the establishment of different types of restaurants at different places in Bengaluru, aggregate rating of each restaurant and many more. Find the Best model for Zomato dataset.

**Details About the Data:**

Structure of the data is 17 columns and 51717 rows

URL: contains the URL of the restaurant on the zomato website

Address=contains the address of the restaurant in Bengaluru

Name: contains the name of the restaurant

Online\_order=whether online ordering is available in the restaurant or not

Book\_table=table book option available or not

Rate=contains the overall rating of the restaurant out of 5

Votes=contains total number of ratings for the restaurant as of the above-mentioned date

Phone=contains the phone number of the restaurant

Location=contains the neighborhood in which the restaurant is located

Rest\_type=restaurant type

Dish\_liked=dishes people liked in the restaurant

Cuisines=food styles, separated by a comma

Approx\_cost(for two people)=contains the approximate cost for a meal for two people

Reviews\_list=list of tuples containing reviews for the restaurant, each tuple consists of two values

menu\_item = items they ordered

listed\_in(type) = type they choose

listed\_in(city) = correct location

**Loading Dataset and Required Libraries:**

Import the required libraries such as Numpy, Pandas, Seaborn, Matplotlib for Data manipulation and visualization.

Import the required libraries such as Linear Regression, Decision tree, Random forest for Model analysis.

**Data Cleaning**

**Deleting Redundant Columns**

Find the redundant column by using heatmap



By this heatmap, we clearly understand dish\_liked column is mostly null values. Url, Address column also no use. So, drop those three columns.

**Renaming the Columns**

approx\_cost(for two people) to cost, listed\_in(type) to type, listed\_in(city) to city for easy access.

**Remove the NaN Values from the Dataset**

Drop the row with the null values by using dropna() function. After this alteration we get 43499 rows.

**Some Transformation in Columns**

**Cost**

The cost column in string format and also the data with the commas. So we need some alteration in that column. Eg: (‘1,300’) changed to (1300)

Replace the commas and change the type to int

**Rate**

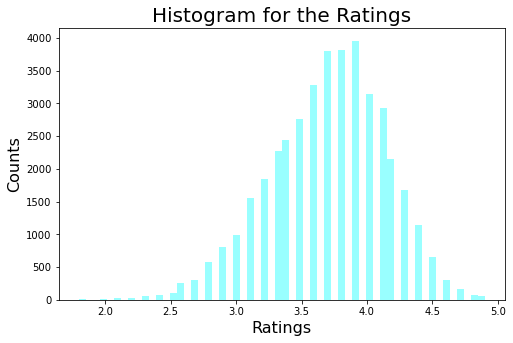
The rate in the column in string format and also with the rating out of five. So, we need some alteration in that column. Eg: (‘4.1/5’) to (4.1)

Replace the decimal operator and change the type to int

**Data Visualization**

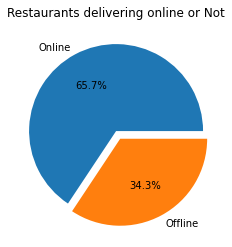
**Using plots to find relations between the features**

**Plot the Occurrence of Rating**

****

**Highest Rating count is 3.9 and Maximum Rating is 4.9**

**Restaurants Delivering Online or Not**

****

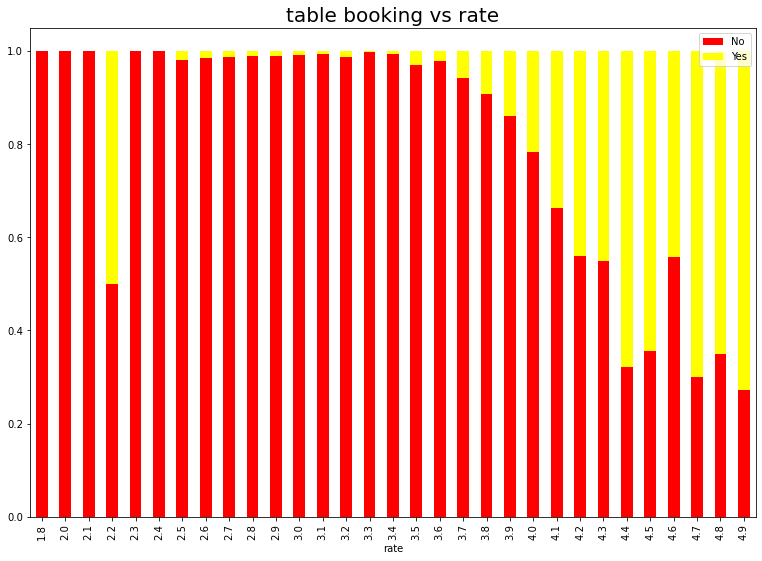
Most of the restaurants delivering online.

**Restaurants Allowing Table Booking or Not**

****

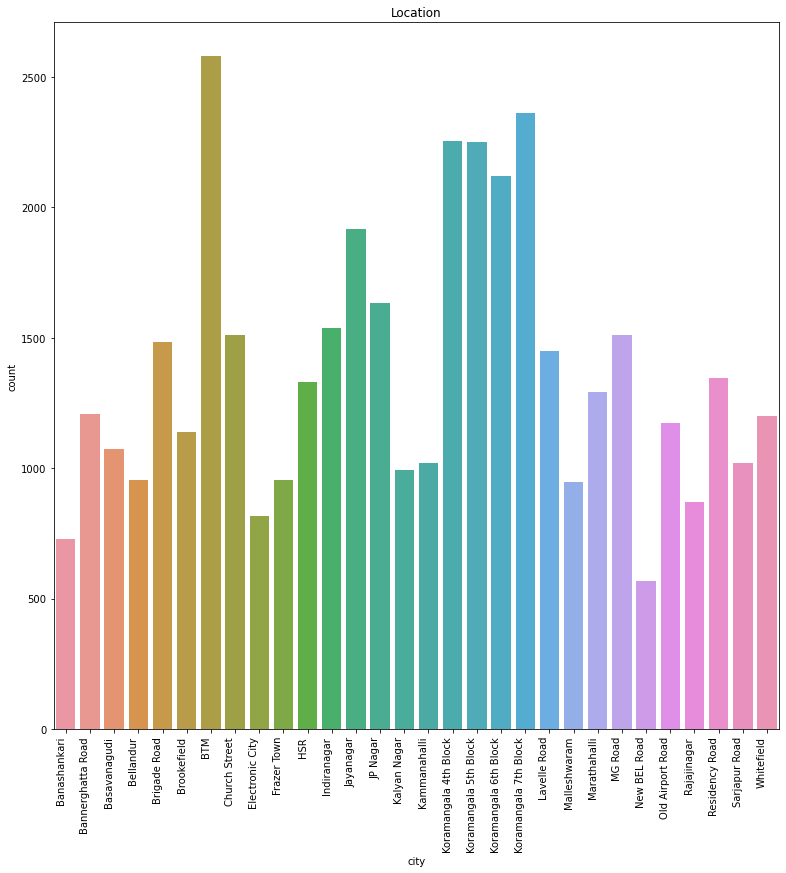
Most of the Restaurant in Bangalore does not allow booking table.

**Table booking Rate vs Rate**

****

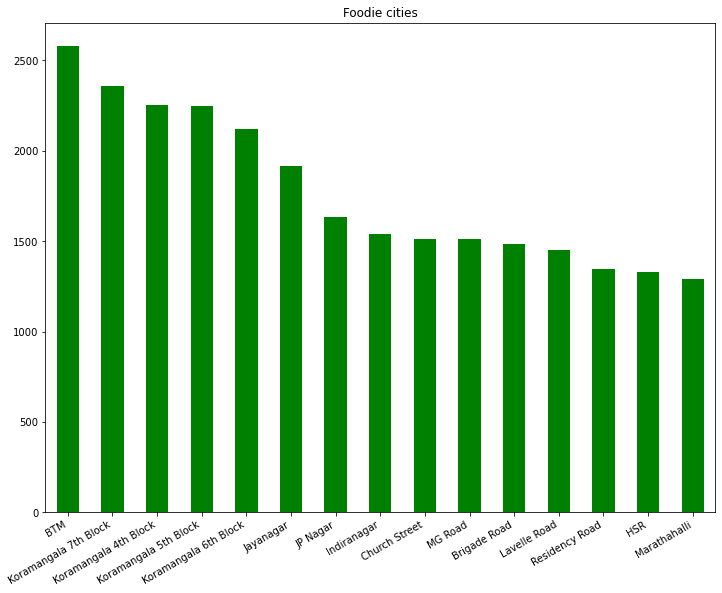
By this plot, we clearly understand that high rating restaurants allow table booking

**Best Locations**



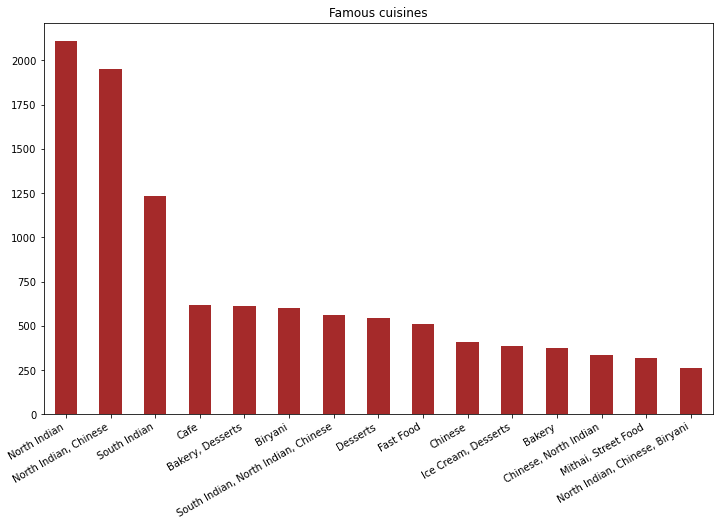
Cities with their counts

**Top Foodie Areas in Bangalore**

****

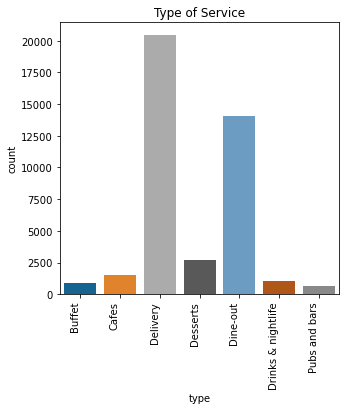
BTM get the first places.

**Most Famous Cuisines**

****

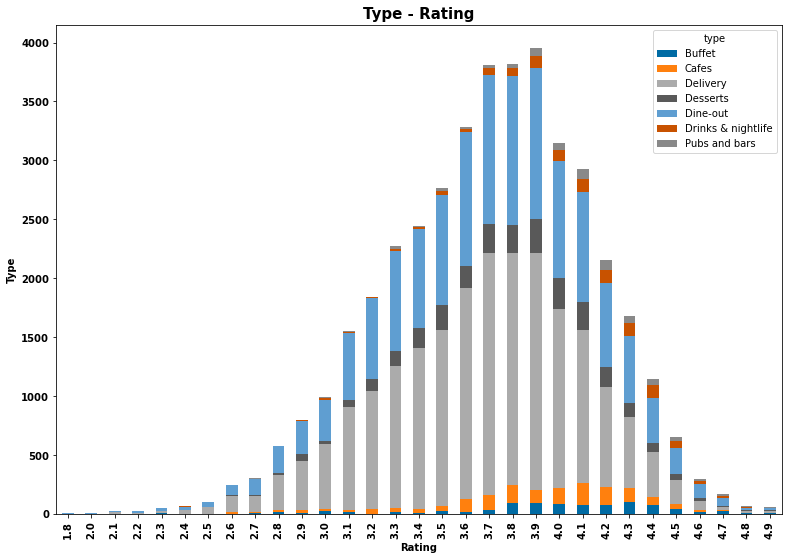
Bangalore is actually located in South India. But people mostly liked North Indian Cuisines.

**Types of Services**

****

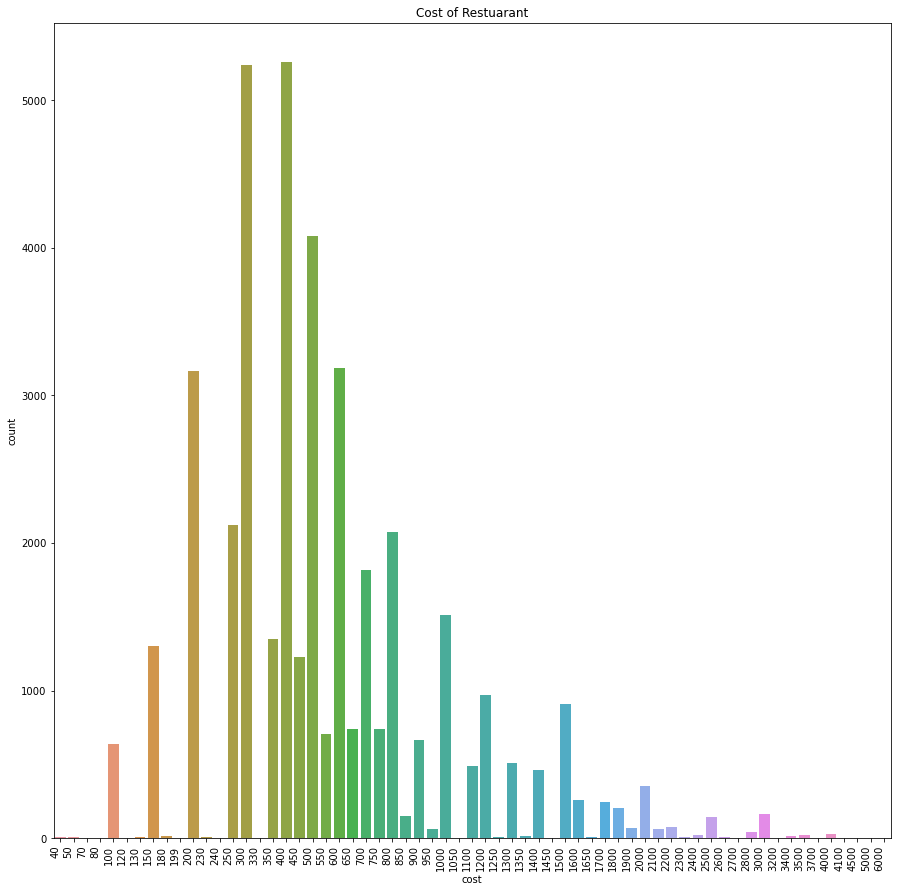
Nowadays, people mostly like Delivery food because Bangalore is most busy city in India.

**Relation between Type and Rating**



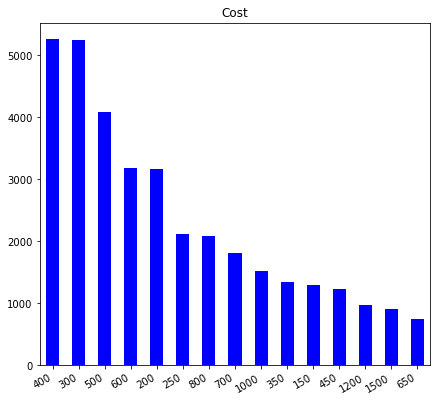
Delivery and Dine-out mostly get high ratings.

**Cost of Restaurant**



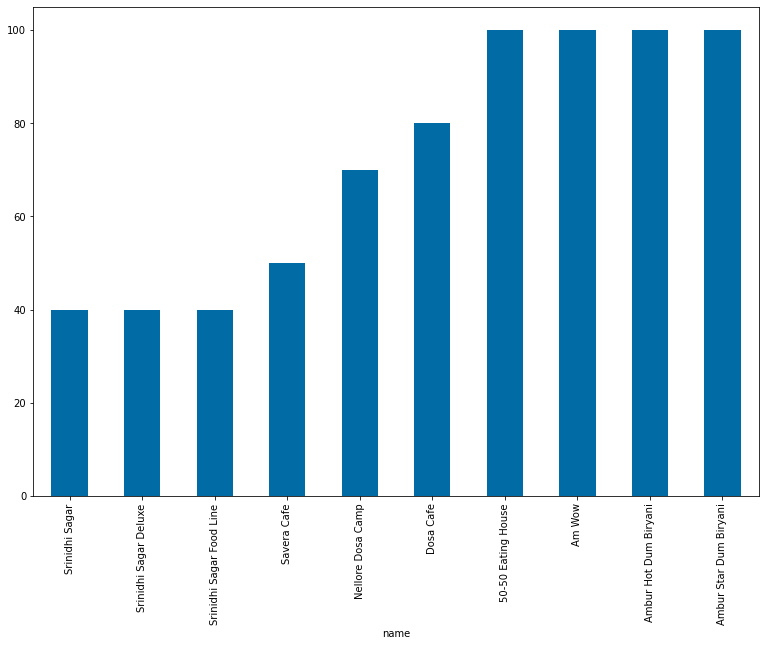
Cost range of the food start from 40 to 6000 rupees.

**People Mostly Buying Food rate**

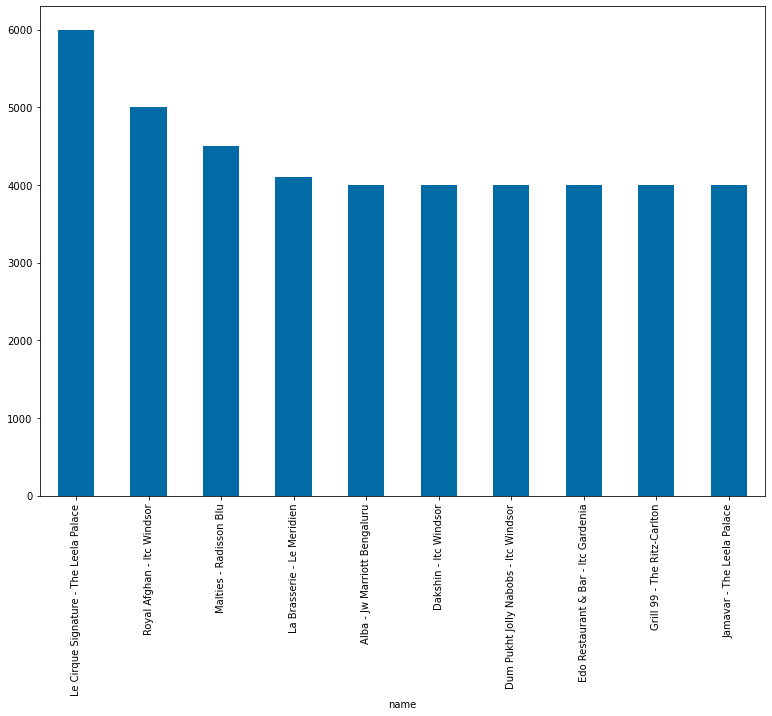


400 rupees in the top place. Most of the people buying food range is between 300-600 rupees.

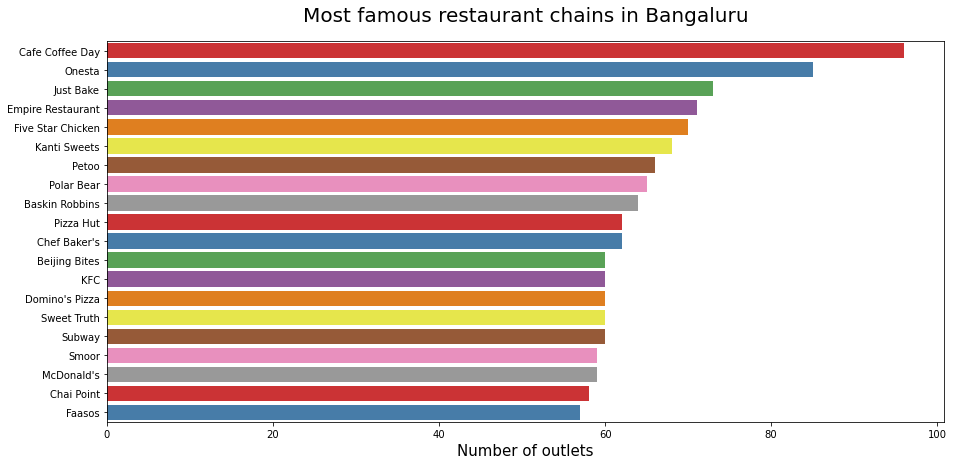
**Budget Friendly Hotel in Bangalore**

****

**Costly Hotel in Bangalore**

****

**Most famous restaurant chains in Bengaluru**

****

**Regression Analysis**

**Correlation Matrix**

****

The highest correlation is between name and address which is 0.62 which is not of very much concern

**Model Analysis**

After the visualization select only required columns for the model selection. The columns areonline\_order, book\_table, votes, location, rest\_type, cuisines, cost, menu\_item for X. rate as the target column.

For this Zomato dataset tried four models’ Linear regression, Decision tree, Random Forest, Extra tree regression. Accuracy is calculated by the r2 score.

**Accuracy for those models**

Linear Regression: 0.28439494620022565

Decision Tree: 0.8492698287817787

Random Forest: 0.8764169255617765

Extra Tree Regression: 0.9386999323276803

**Predicting value for the test data:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| online\_order | book\_table | votes | location | rest\_type | cuisines | cost | menu\_item | rate |
| 0 | 1 | 42 | 20 | 9 | 43 | 250 | 0 | 3.7 |

Linear Regression: 3.57738406

Decision Tree: 3.7

Random Forest: 3.70127637

Extra Tree Regression: 3.7

Decision Tree and Extra tree Regression predict most accurate.

Finally, we get clear idea that **Extra tree Regression** model predict the output with high accuracy **0.9386999323276803**.