**Project: EDA on Zomato Restaurant Rating & Predict**

**the Zomato Restaurant Ratings using Machine Learning**

**Submitted By ADAIKKAPPAN A (EBEON0522601032)**



# ABSTRACT

Zomato was launched in 2010. Zomato's technology platform connects customers, restaurant partners and delivery partners, serving their multiple needs. Customers use the Zomato platform to search and discover restaurants, read and write customer-generated reviews and view and upload photos, order food delivery, book a table and make payments while dining out at restaurants. On the other hand, Zomato provides restaurant partners with industry-specific marketing tools which enable them to engage and acquire customers to grow their business while also providing a reliable and efficient last-mile delivery service. Zomato also operates a one-stop procurement solution, Hyperpure, which supplies high quality ingredients and kitchen products to restaurant partners. Zomato also provides our delivery partners with transparent and flexible earning opportunities. Now I am going to analyze the Bangalore Zomato Restaurants Dataset and predict the ratings of a restaurants by using machine learning algorithms.

# CONTENTS

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| Chapter 1 | 1. Introduction 2. Problem Statement (iii). Study of Existing Systems   (iv). Identification of gaps in existing systems (v). Discussed on proposed solution  (vi). Tools/Technology used to implement proposed solution | 4 |
| Chapter 2 | Features & Predictor | 7 |
| Chapter 3 | Methodology:   1. Data Cleaning Pre- processing 2. Machine Learning Algorithms 3. Implementation Steps | 8 |
| Chapter 4 | Analysis of the Result | 42 |
| Chapter 5 | Conclusion | 19 |
|  | References | 20 |

# Chapter 1

## Introduction

Zomato is one of the best online food delivery apps which gives the users the ratings and the reviews on restaurants all over India. These ratings and the Reviews are considered as one of the most important deciding factors which determine how good a restaurant is. We will therefore use the real time Data set with various features a user would look into regarding a restaurant. We will be considering Bangalore City in this analysis.

## Problem Statement:

The basic idea of analysing the Zomato dataset is 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, Bengaluru 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. In spite of increasing demand, it however has become difficult for new restaurants to compete with established restaurants. Most of them serving the same food. Bengaluru being an IT capital of India. Most of the people here are dependent mainly on the restaurant food as they don’t have time to cook for themselves. With such an overwhelming demand of restaurants it has therefore become important to study the demography of a location. What kind of a food is more popular in a locality. Do the entire locality loves vegetarian food. If yes then is that locality populated by a particular sect of people for e.g., Jain, Marwaris, Gujaratis who are mostly vegetarian. This kind of analysis can be done using the data.

## Study of Existing Systems:

1. EDA: Bangalore restaurants:
   * Author: **FF RANKUSHA -** In this project, she did the analysis part in three different stages, that is Univariate Analysis, Bivariate Analysis & Automated Analysis.
2. Zomato [EDA - FE - Model Building]:
   * Author-**ADITYA RAWAT -** In this project, he did the analysis part and applied the machine learning algorithms to predict the ratings of a restaurant.

## Identification of gaps in existing systems:

1. EDA: Bangalore restaurants:
   * Author: **FF RANKUSHA -** Analysis Part should be more in detail.
2. Zomato [EDA - FE - Model Building]:
   * Author-**ADITYA RAWAT -** In this project, he applied only two machine learning algorithms to predict the ratings of a restaurant.

## Discussed on proposed solution:

1. EDA: Bangalore restaurants:
   * Author: **FF RANKUSHA -** Analysis Part should be more in detail like which service types are more popular in Bangalore and which is the most popular cuisines available in Bangalore and check the Top Rating Restaurants in Bangalore having online order or not and check table booking option available or not and check most liked dishes available or not.
2. Zomato [EDA - FE - Model Building]:
   * Author: **ADITYA RAWAT –** We can apply additional two more Machine learning algorithms Random Forest Regressor and Linear Regression and compare the best model by using the accuracy of each model and we can generate the Feature Importance to understand which feature is used to make key decisions.

## Tools/Technology used to implement proposed solution:

* + Python
  + Pandas
  + Numpy
  + Matplotlib
  + Seaborn
  + Plotly
  + Sklearn
  + Jupyter Notebook

# Chapter 2

## Features & Predictor

* url - object
* address - object
* name - object
* online\_order - object
* book\_table - object
* rate - int64
* votes - int64
* phone - object
* location - object
* rest\_type - object
* dish\_liked - object
* cuisines - object
* approx\_cost(for two people) - int64
* reviews\_list - object
* menu\_item - object
* listed\_in(type) - object
* listed\_in(city) - object

## Note

* + Total - 17 columns
  + Numerical – 3 - Continuous: Which is quantitative data that can be measured.
  + String – 14 - Ordinal Data: Categorical data that has an order to it.

# Chapter 3

## Methodology

### Data Cleaning and Pre-processing:

The datasets which were collected from Zomato Restaurant Rating Dataset from Kaggle website contain unfiltered data which must be filtered before the final data set can be used to do analysis. Also, data has some categorical variables which must be modified into numerical values for which we used Panda’s library of Python. In data cleaning step, first we checked whether there are any missing or junk values in the dataset for which we used the is null () function.

### Machine Learning Algorithms:

1. **ExtraTree Regressor:**

An extra-trees regressor, this class implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub- samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

1. **Random Forest Regressor:**

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max\_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree.

1. **Decision Tree Regressor:**

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

1. **Linear Regression:**

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used. There are many names for a regression’s dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

### Implementation Steps:

As we already discussed in the methodology section about some of the implementation details. So, the language used in this project is Python programming. We’re running python code in anaconda navigator’s Jupyter notebook. Jupyter notebook is much faster than Python IDE tools like PyCharm or Visual studio for implementing ML algorithms. The advantage of Jupyter notebook is that while writing code, it’s really helpful for Data visualization and plotting some graphs like histogram and heatmap of correlated matrices. Let’s revise implementation steps: a) Dataset collection. b) Importing Libraries: NumPy, Pandas, Matplotlib, Seaborn and Sklearn libraries were used. c) Exploratory data analysis: For getting more insights about data. d) Data cleaning and pre-processing: Checked for null and junk values using isnull() and isna().sum() functions of python. In Pre-processing phase, we did feature engineering on our dataset. As we converted categorical variables into numerical variables using function of Pandas library. All our datasets contains some categorical variables.

### LIBRARIES:

import pandas as pd import numpy as np

from matplotlib import pyplot as plt import seaborn as sns

import plotly.express as px

from sklearn.preprocessing import LabelEncoder import re

from sklearn.model\_selection import train\_test\_split from sklearn.metrics import r2\_score

from sklearn.linear\_model import LinearRegression from sklearn.ensemble import RandomForestRegressor from sklearn.ensemble import ExtraTreesRegressor from sklearn.tree import DecisionTreeRegressor

import warnings warnings.filterwarnings('ignore')

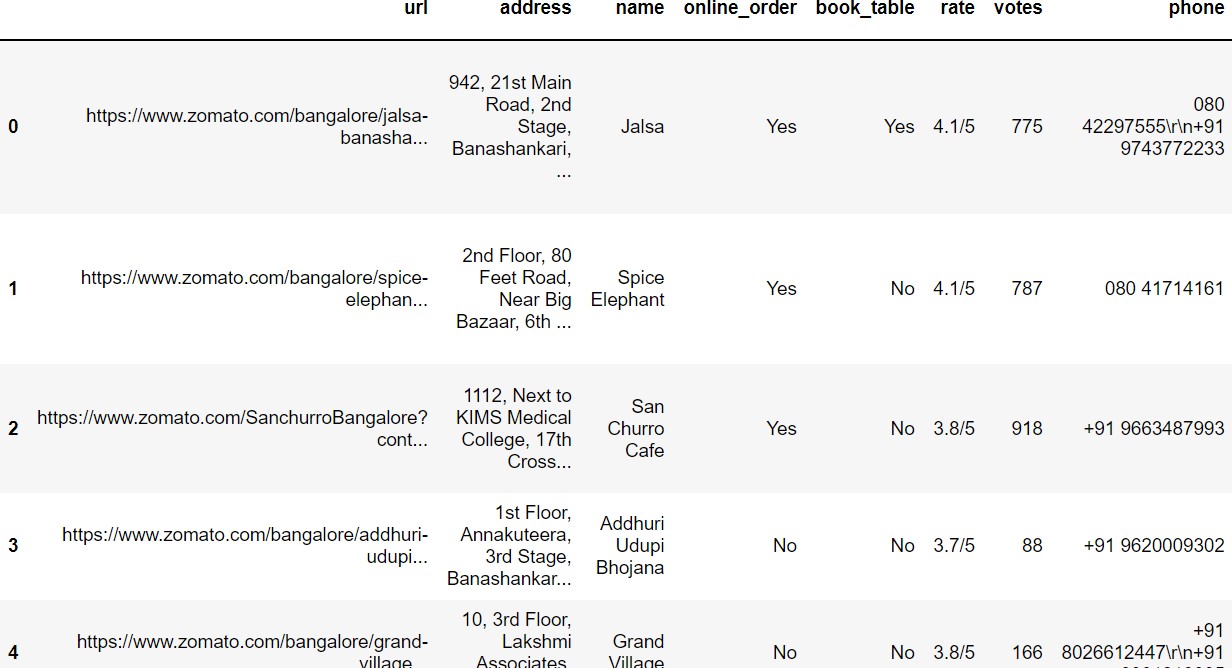
%matplotlib inline

### Reading the Dataset and assigning that to the variable df:

df=pd.read\_csv("zomato.csv")

### Access the first 5 rows of a dataframe:

df.head()



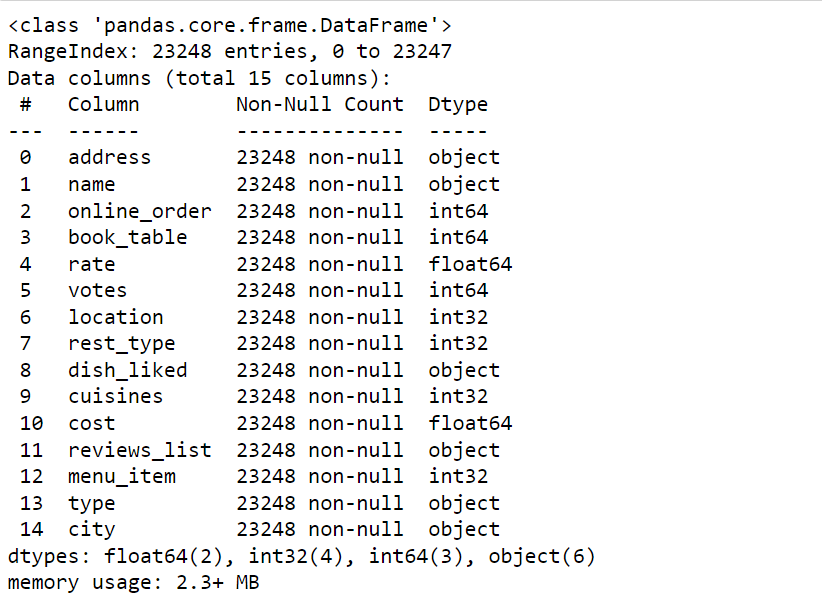
### Access the last 5 rows of a dataframe:

df.tail()

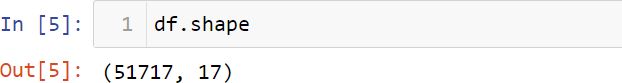


### Prints information about the DataFrame:

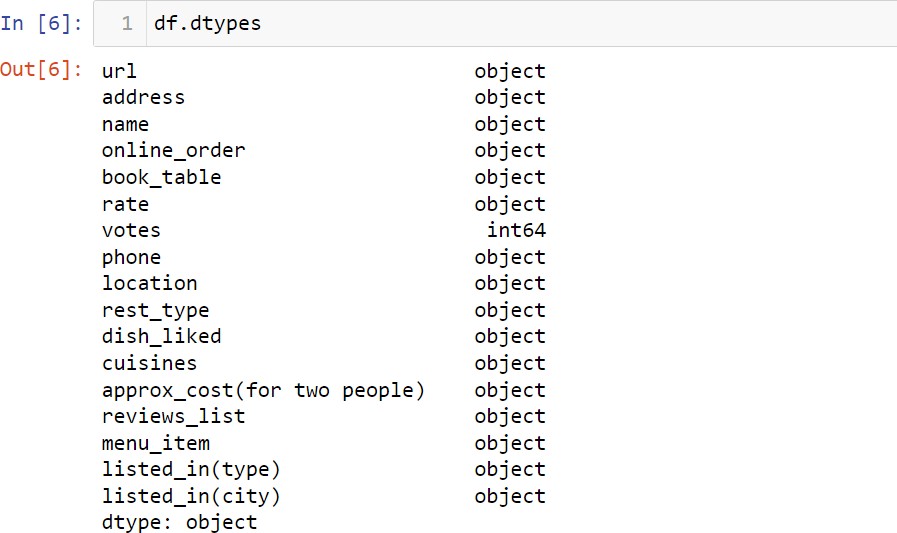
df.info()



### Dimension of the Dataset:



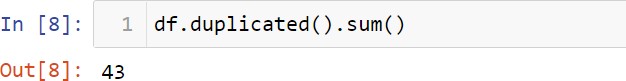
**Checking the data type for each column:**



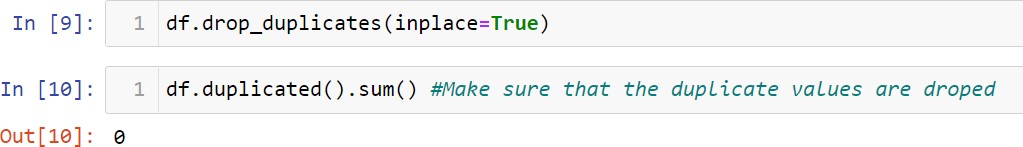
### Delete the Unnnecessary Columns:

df=df.drop(['url','phone'],axis=1)

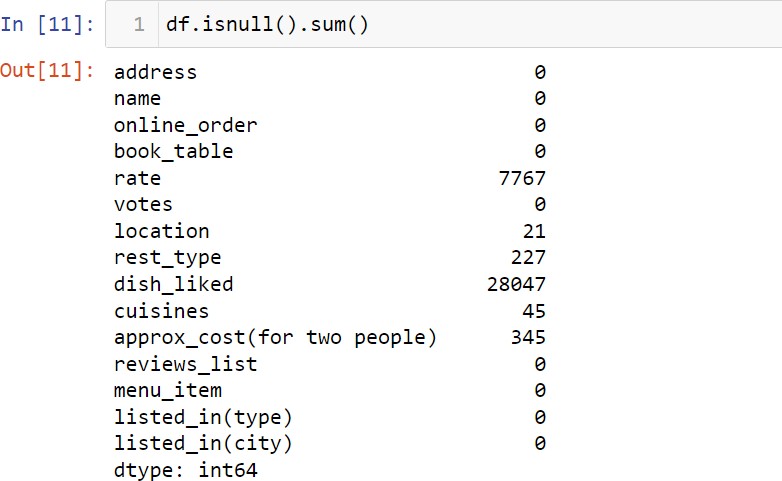
### Checking for duplicate values:



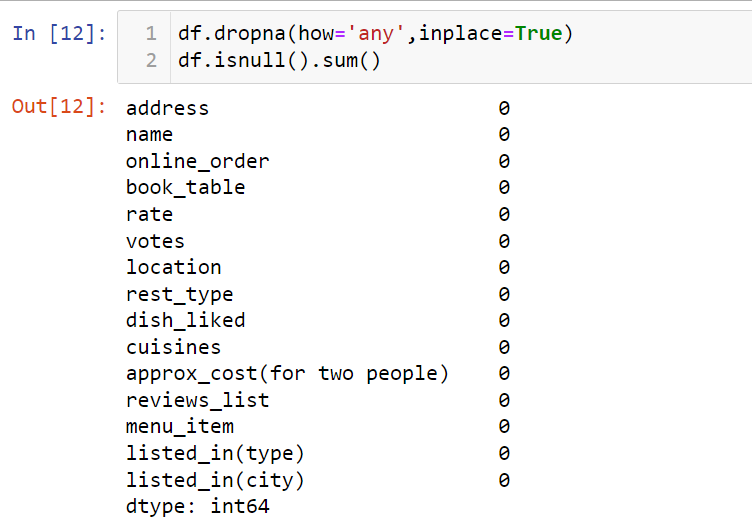
**Drop the duplicate values:**



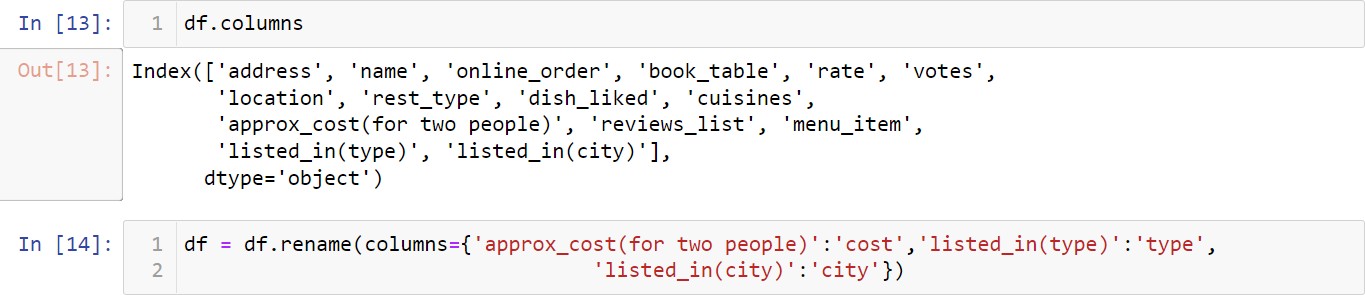
### Checking for null values:



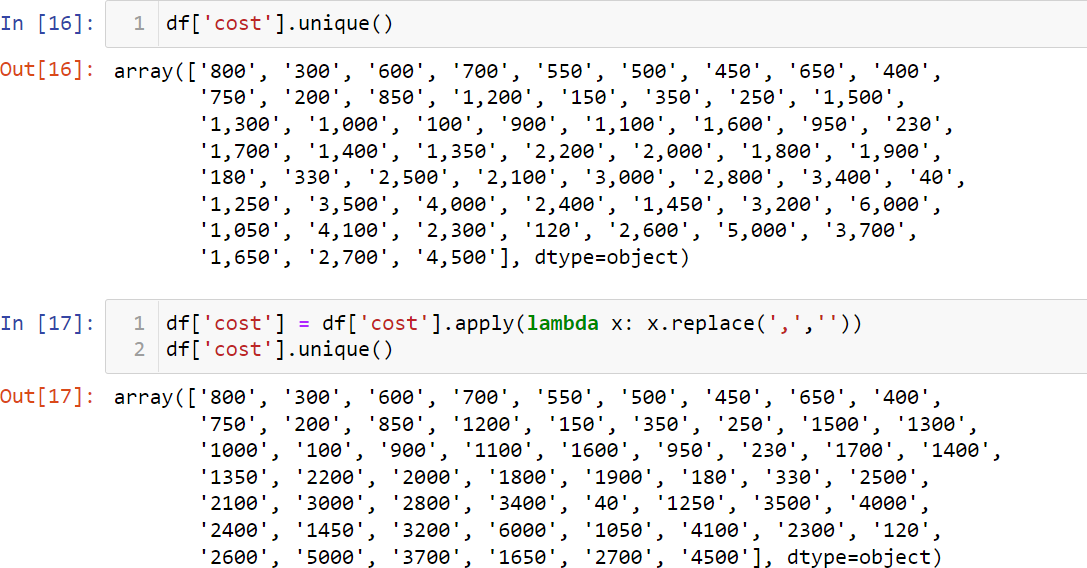
**Drop the null values:**



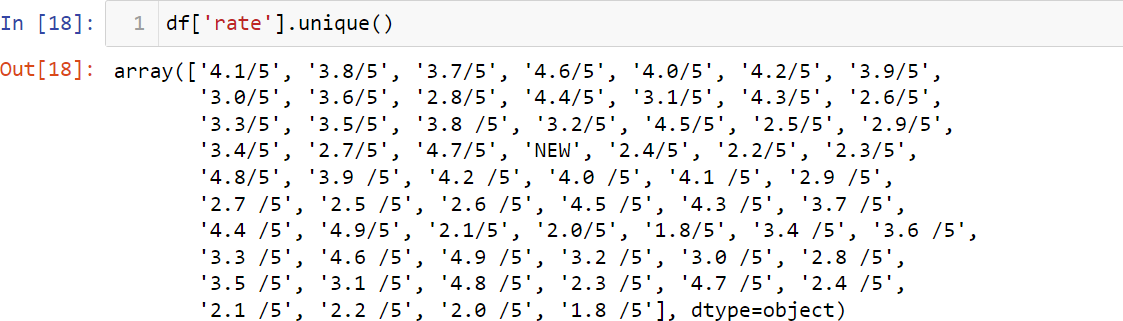
### Renaming the columns appropriately:

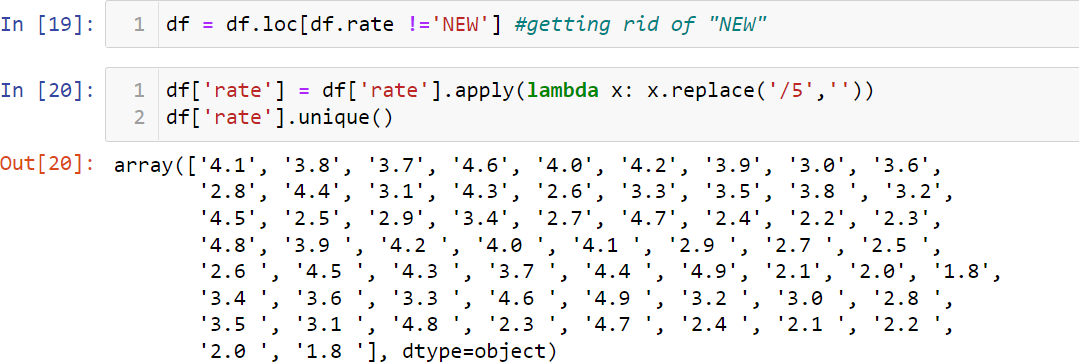


**Replace the comma (',') from cost:**



### Removing '/5' from Rates:

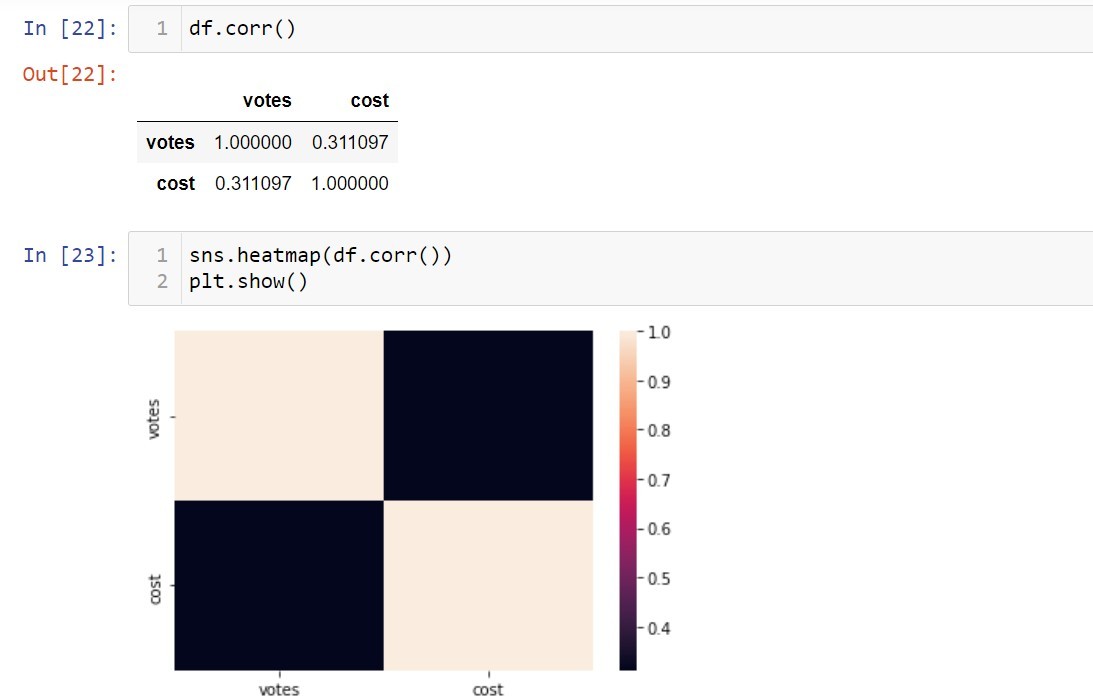




**Convert the cost column datatype to float:**



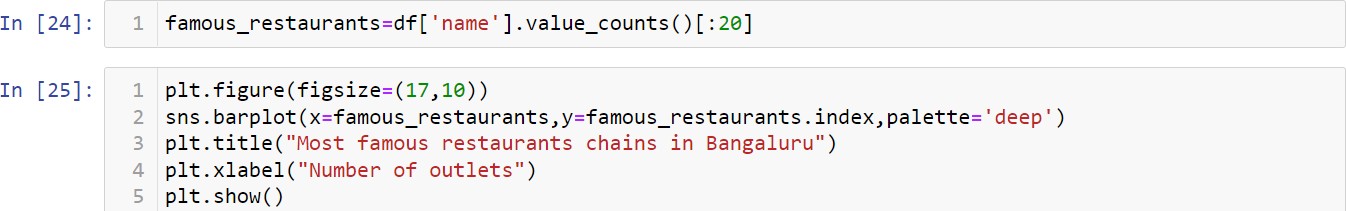
### Correlation:



**Observation:**

* Votes are highly correlated with cost
* Cost is highly correlated with votes

### Most famous restaurants chains in Bangaluru:



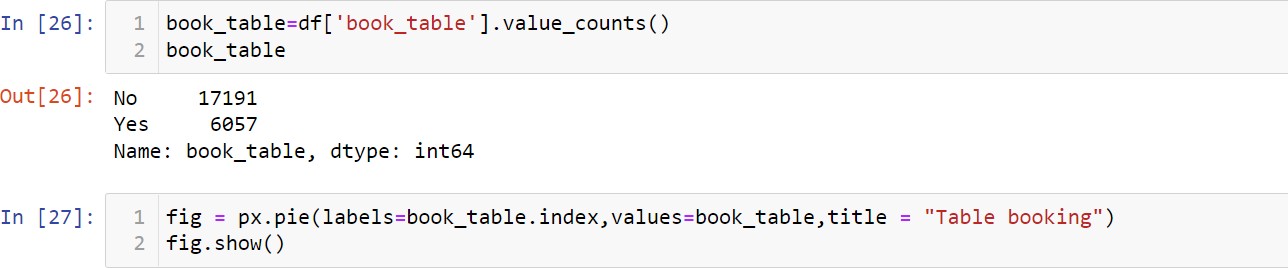


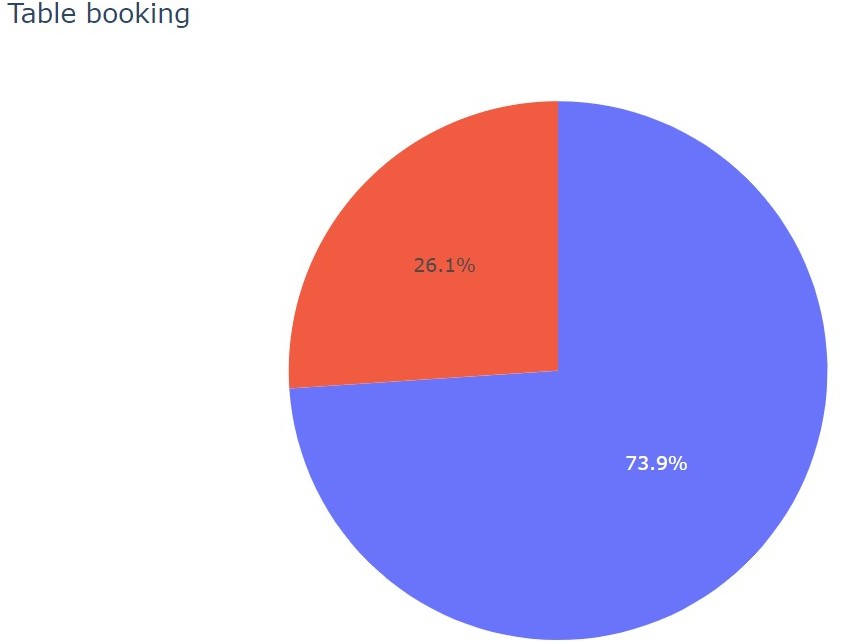
**Observation**

The Top five famous restaurants in Bangalore are

* Onesta
* Empire Restaurant
* KFC
* McDonald's
* Pizza Hut

### Whether restaurant offer Table booking or not:

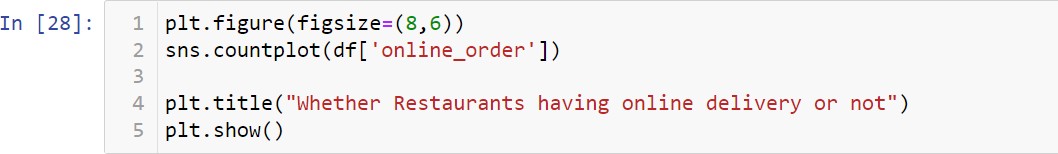


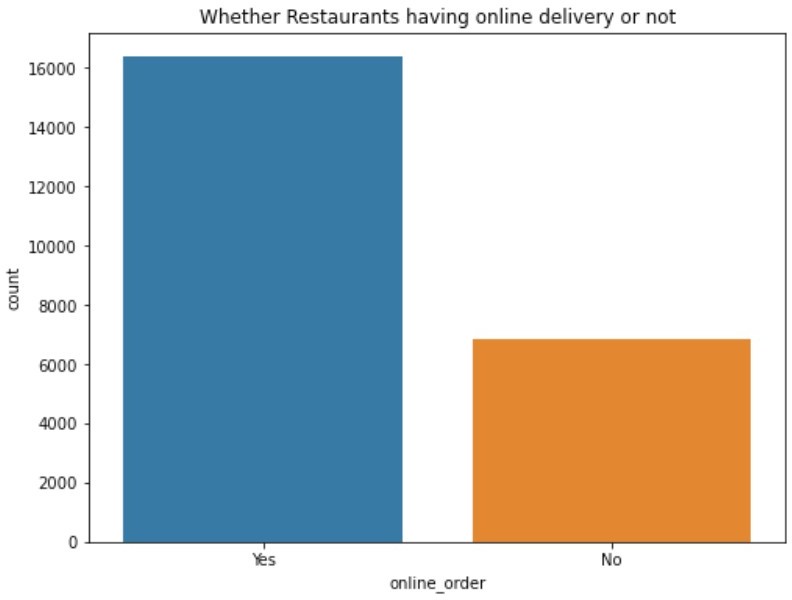


**Observation**

* 73.9% of the Restaurants do not offer table booking
* 26.1% of the Restaurants offer table booking

### Whether Restaurants having online delivery or not:



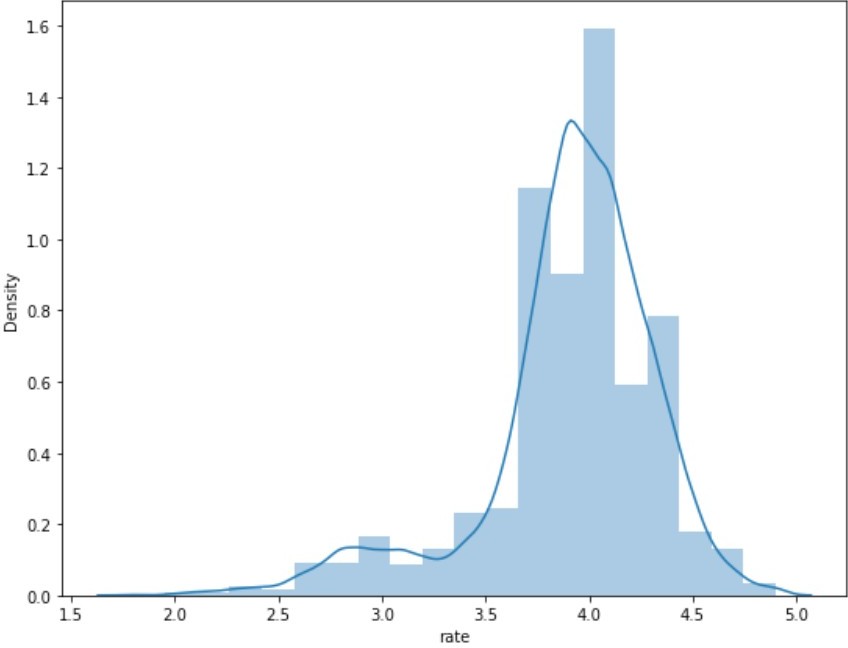


**Observation:**

* Most of the Restaurants offer option for online order and delivery

### Rating Distributions:

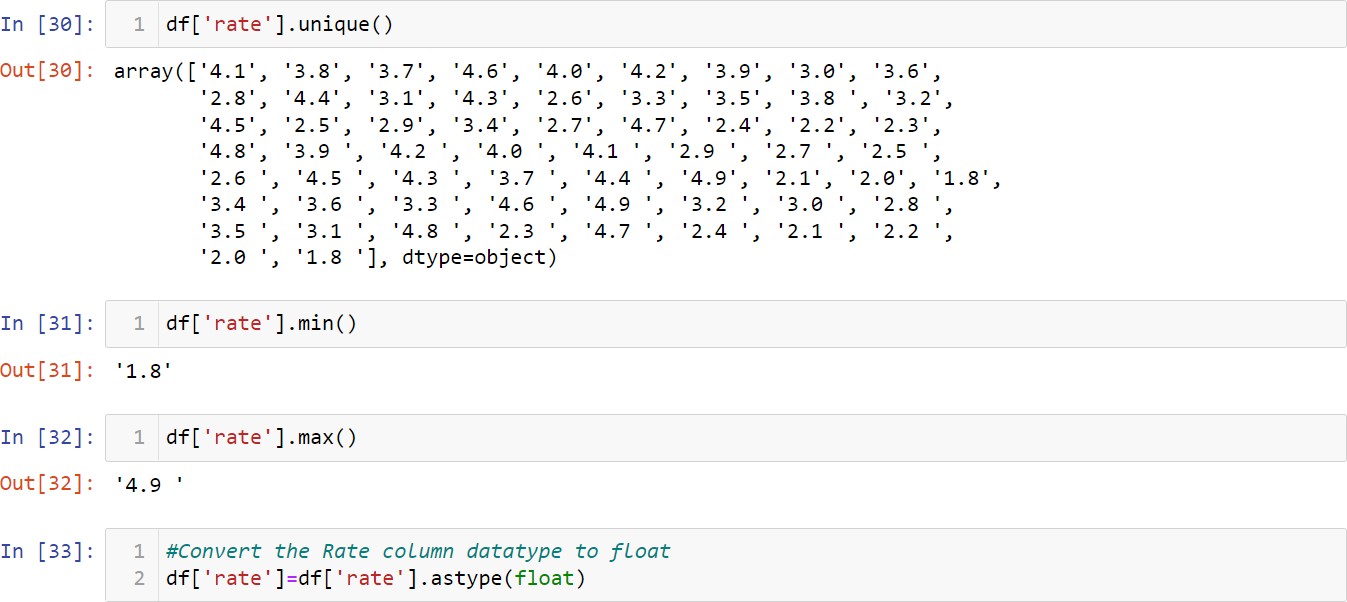




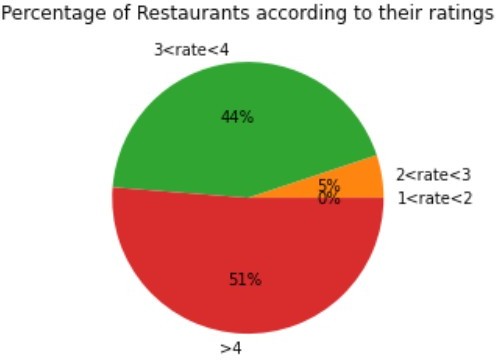
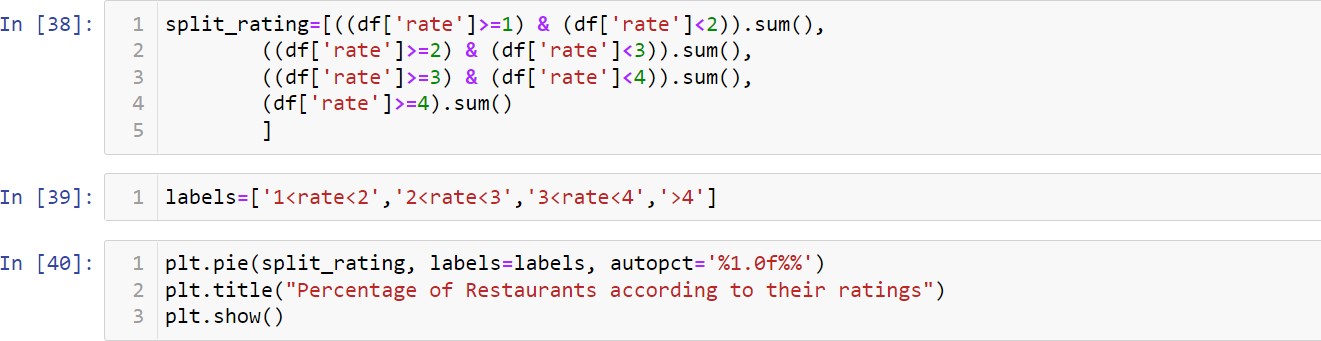
**Observation:**

* We can infer from above that most of the ratings are within 3.5 and 4.5

### Checking the count of ratings as between "1 and 2", "2 and 3", "3 and 4", and "4 and 5":



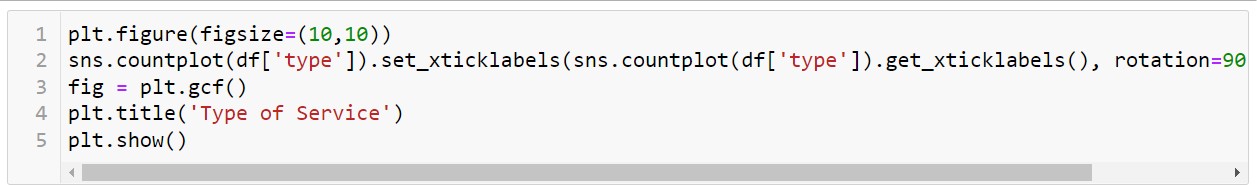
**Plotting the counts with the help of pie chart:**

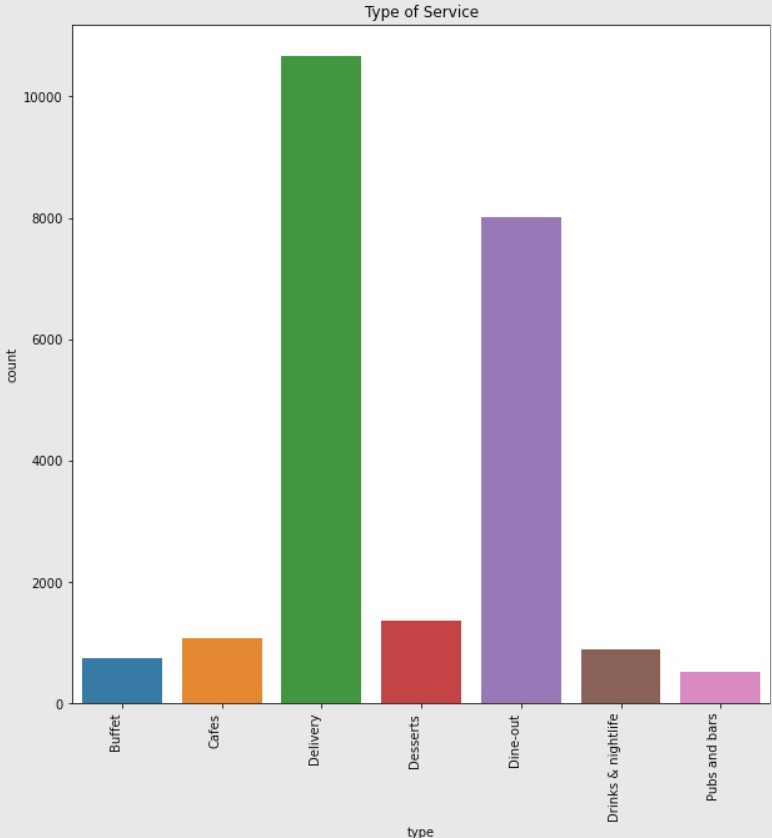


### Observation:

* 51% of Restaurants ratings are greater than 4
* 44% of Restaurants ratings between 3 to 4
* 5% of Restaurants ratings between 2 to 3

### Services Types:

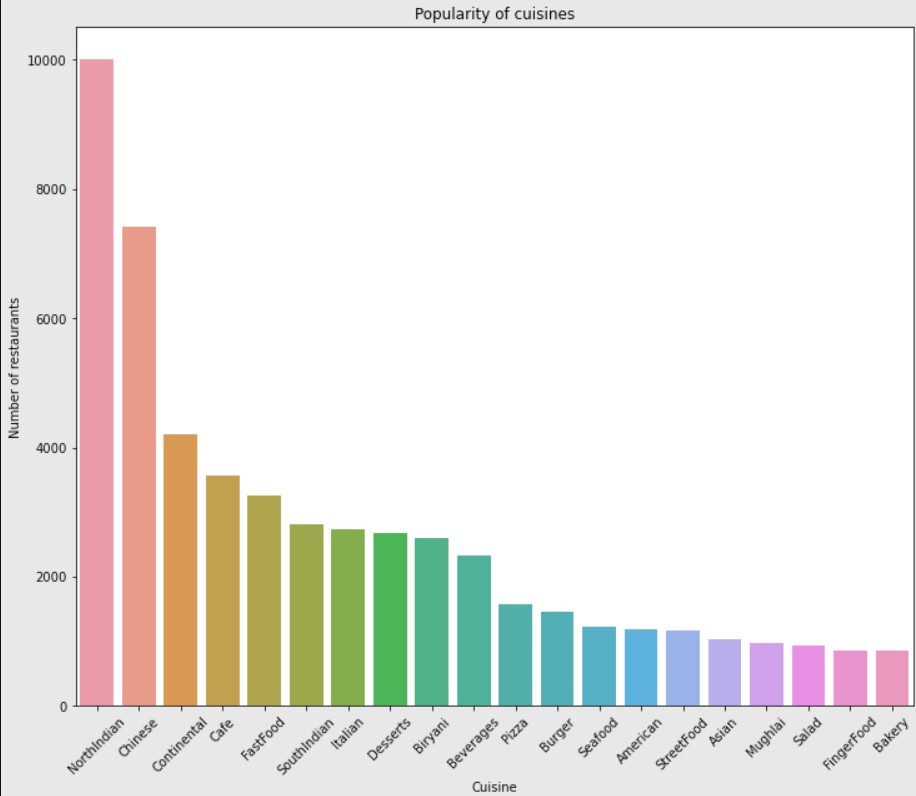
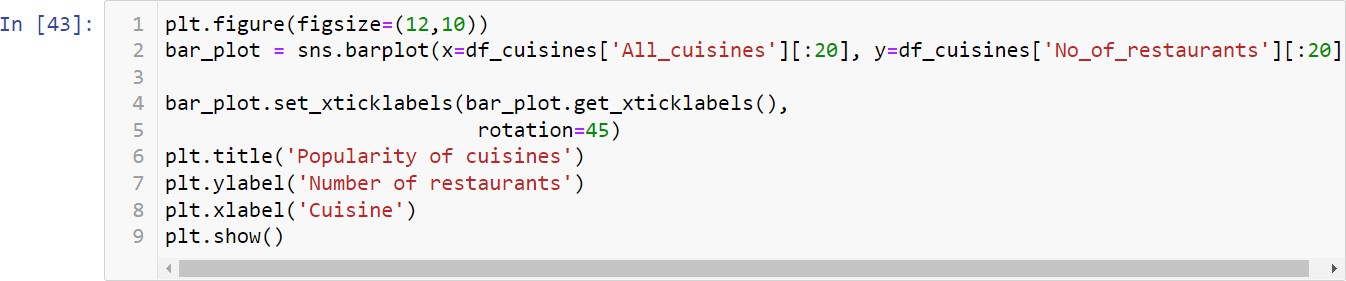
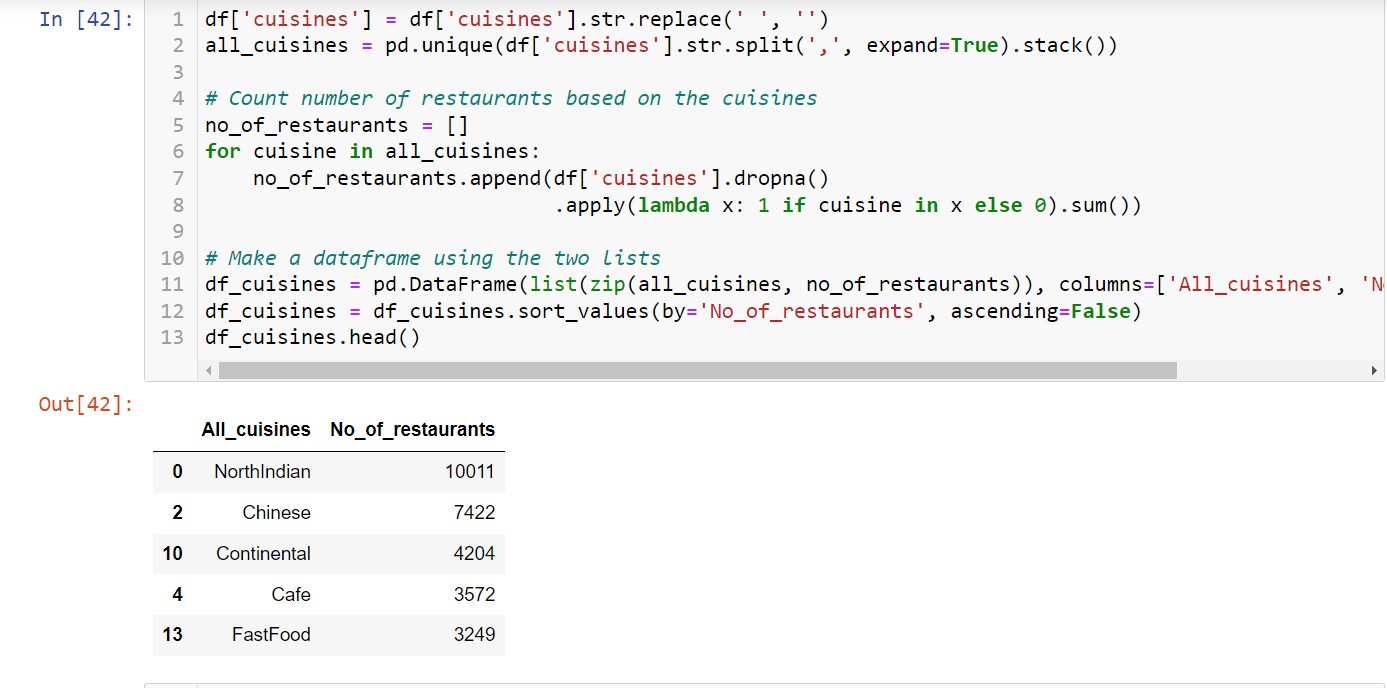




**Observation:**

* Here the two main service types are Delivery and Dine-out

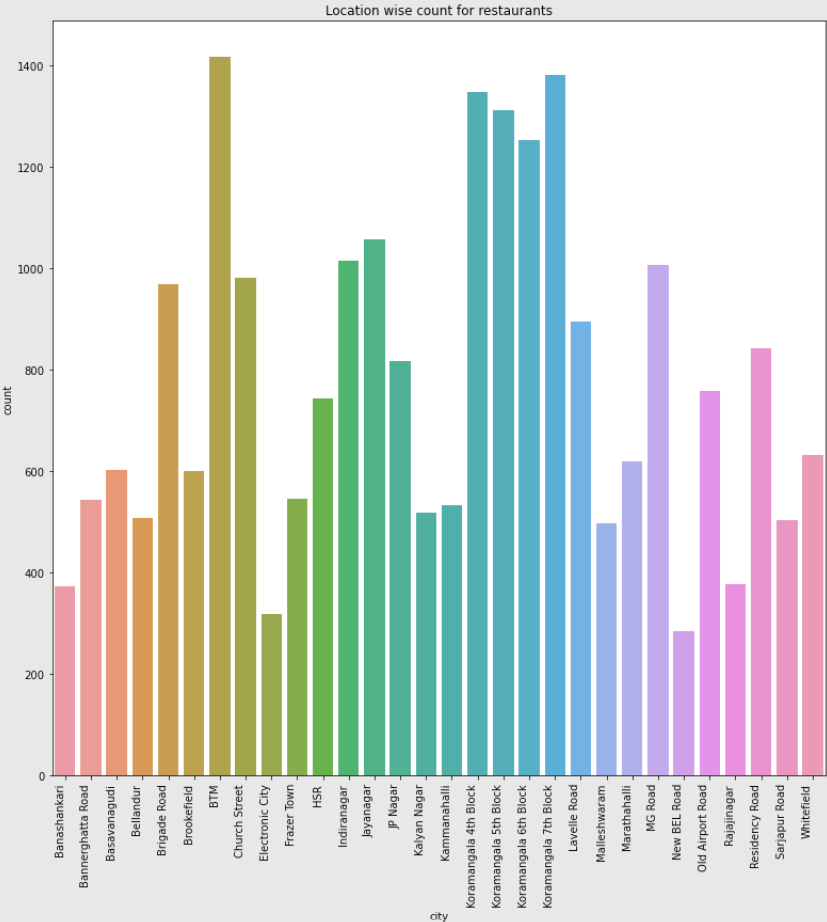
### Cuisines:



**Observation:**

* Here the three popularity cuisines are NorthIndian,Chinese and Continental

### Location wise count for restaurants:



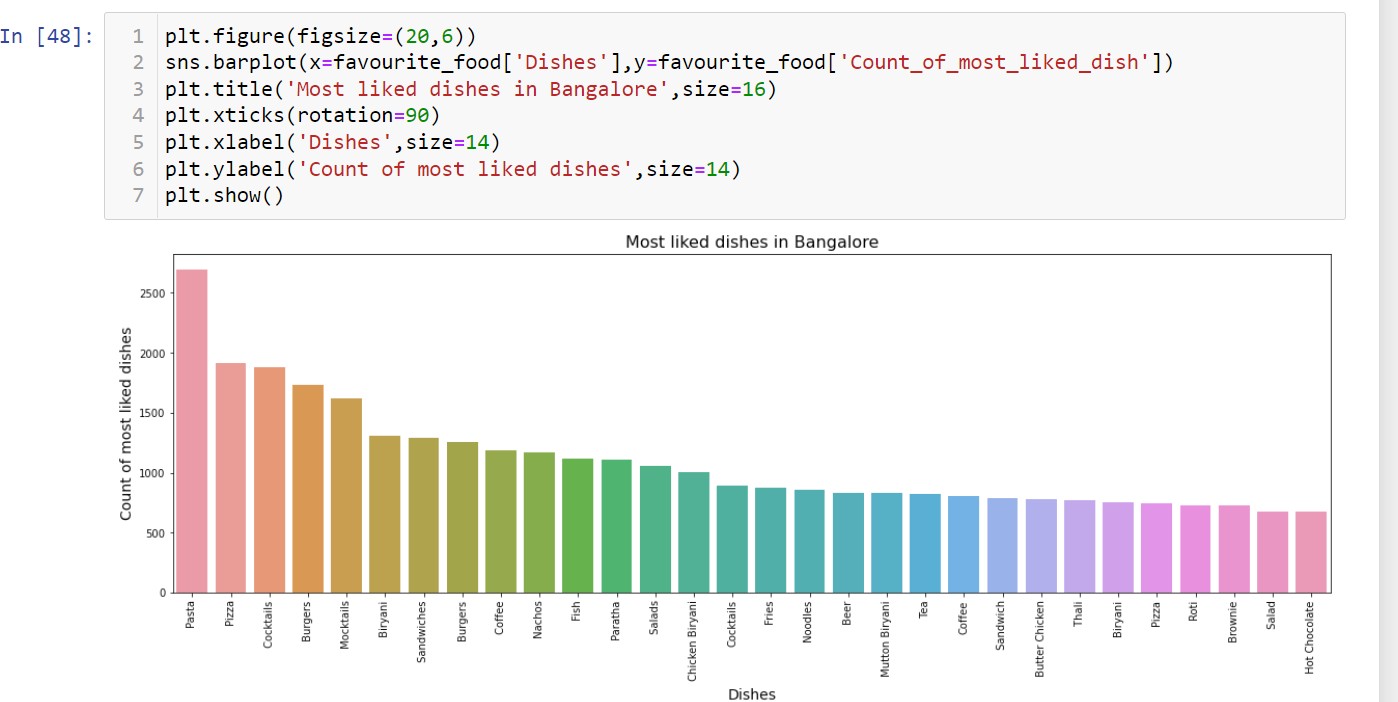
**Observation:**

* We can infer from the analysis that the top 3 Locations are BTM, Koramangala 7th Block and Koramangala 4th Block

### Most Liked Dishes:



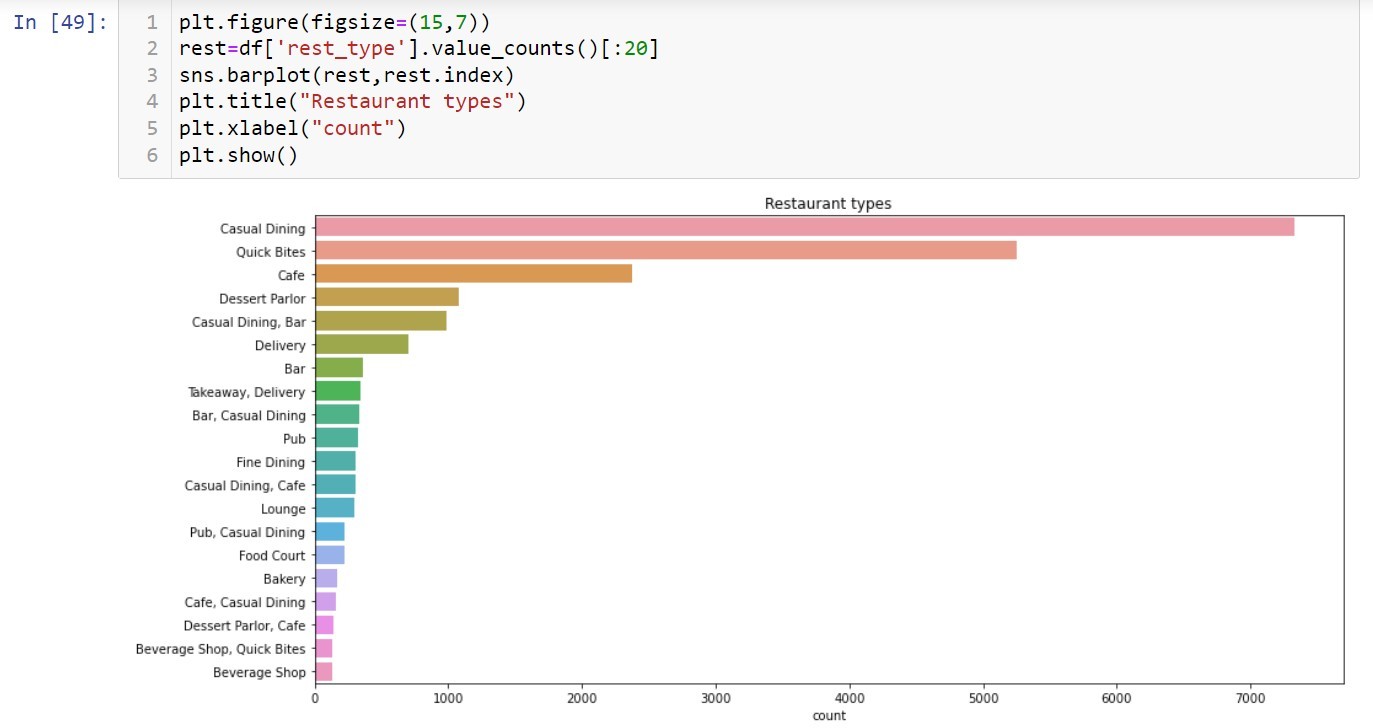




**Observation:**

* We can infer from the analysis that the 5 most liked dishes are Pasta,Pizza,Cocktails,Burgers,and Mocktails

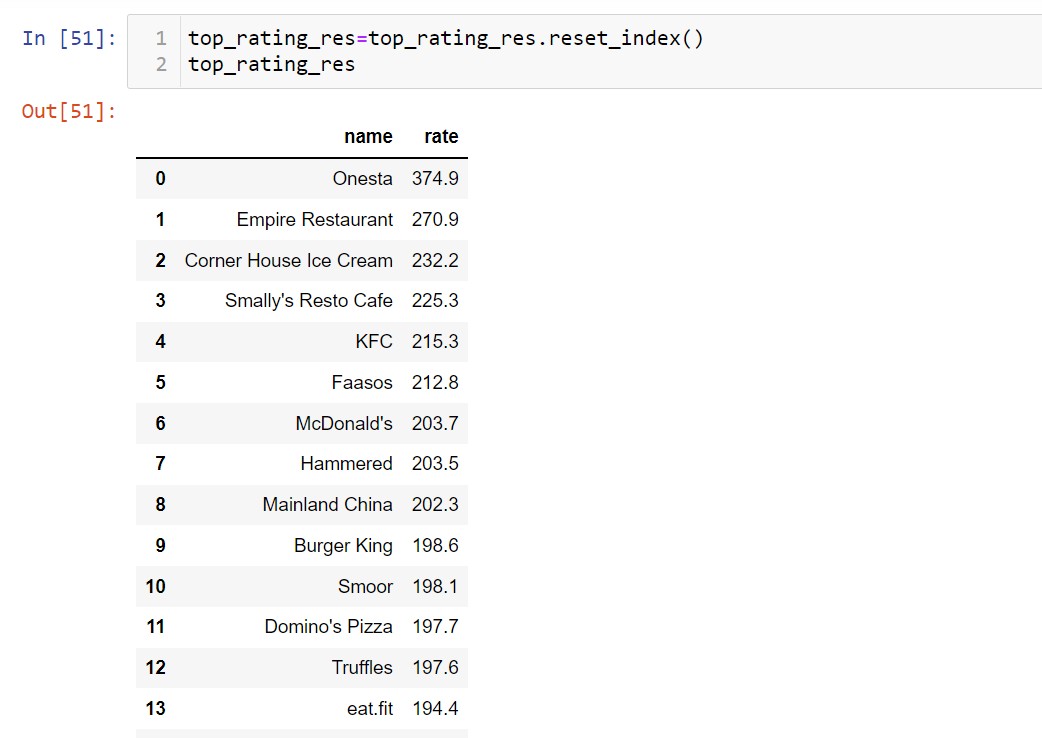
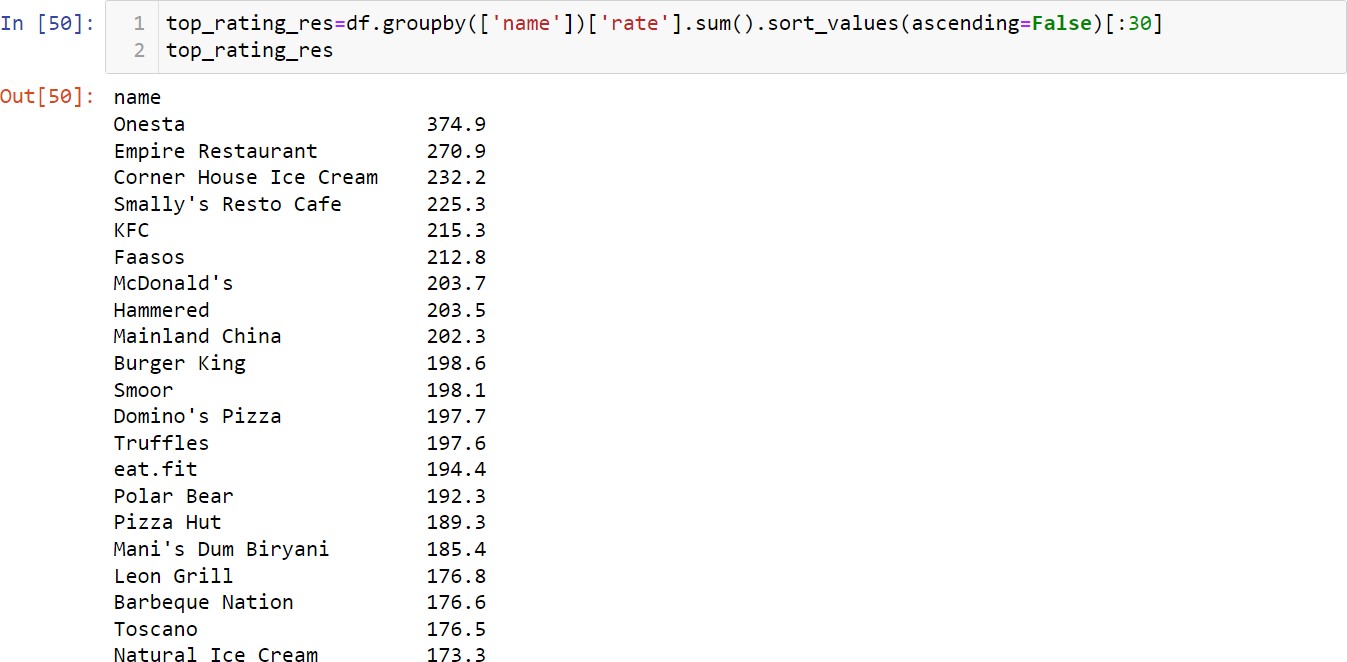
### Restaurants types and their counts:



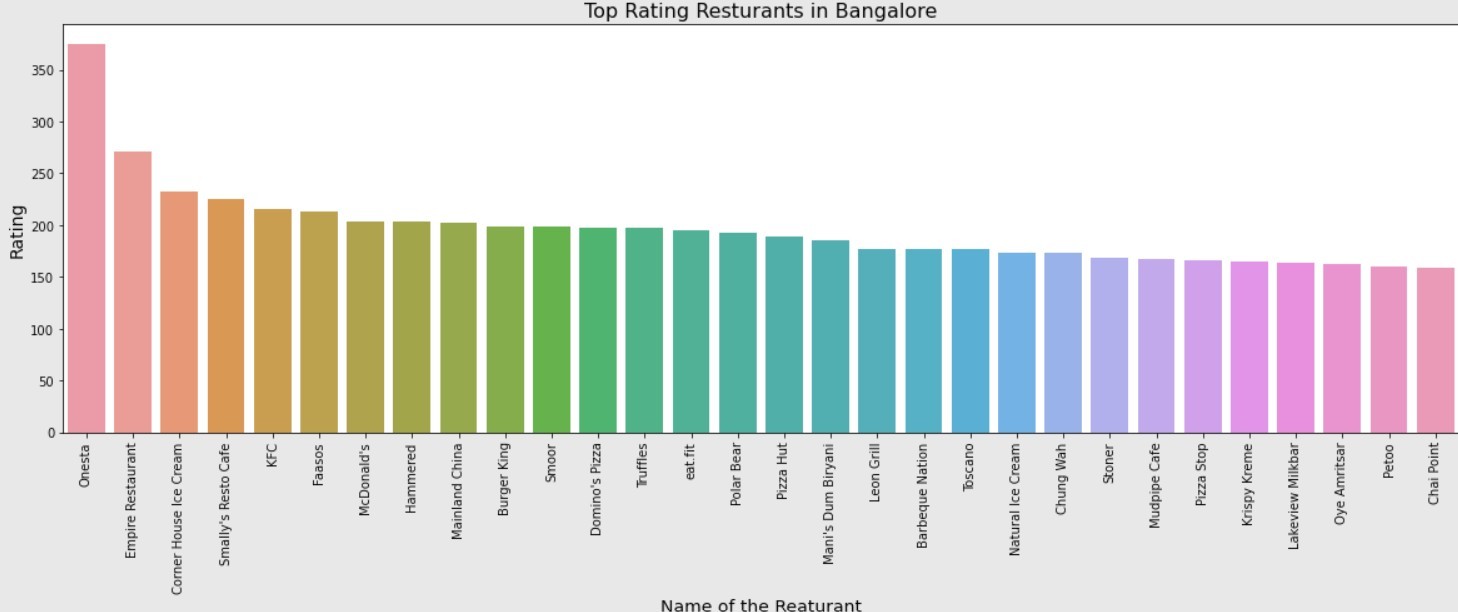
**Observation:**

* Casual Dining, Quick Bites and Cafe are the 3 most common types of Restaurants in Bangalore

### Top Rating Restuarants in Bangalore:



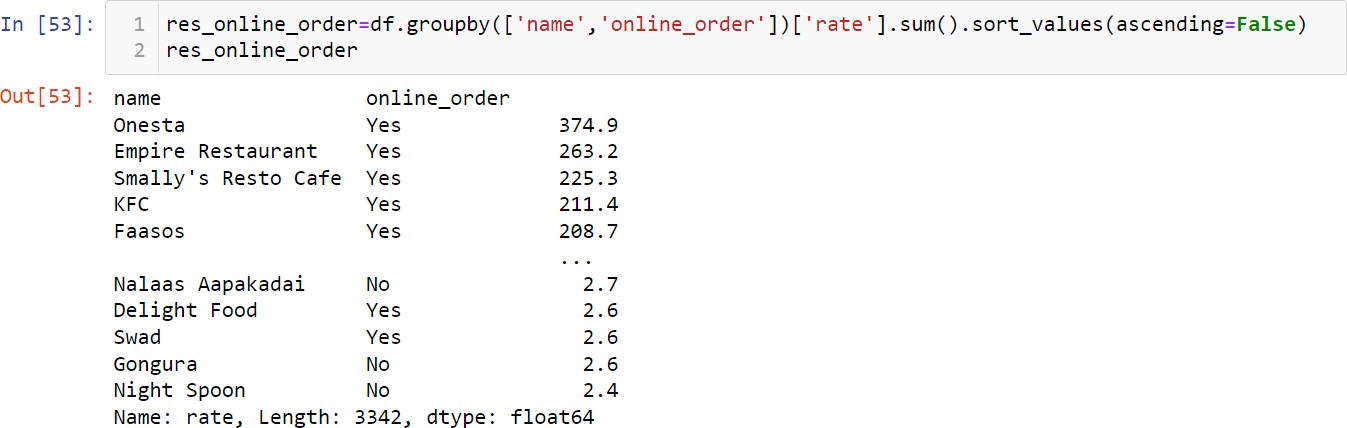




**Observation:**

* We can infer from the analysis that the 5 most top restuarants are Onesta, Empire Restaurant, Corner House Ice Cream, Smally's Resto Cafe, KFC

### Check Top Rating Restaurants having online order:



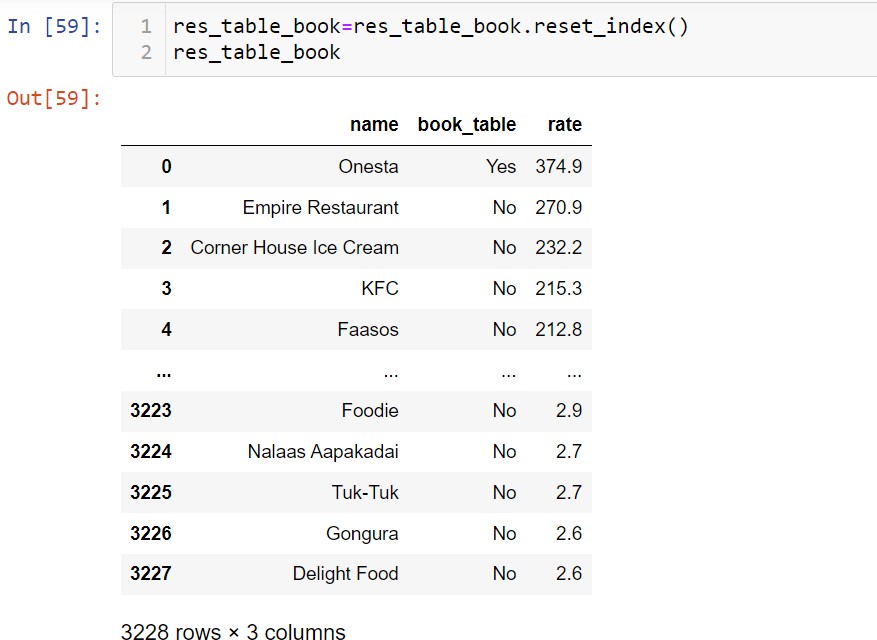
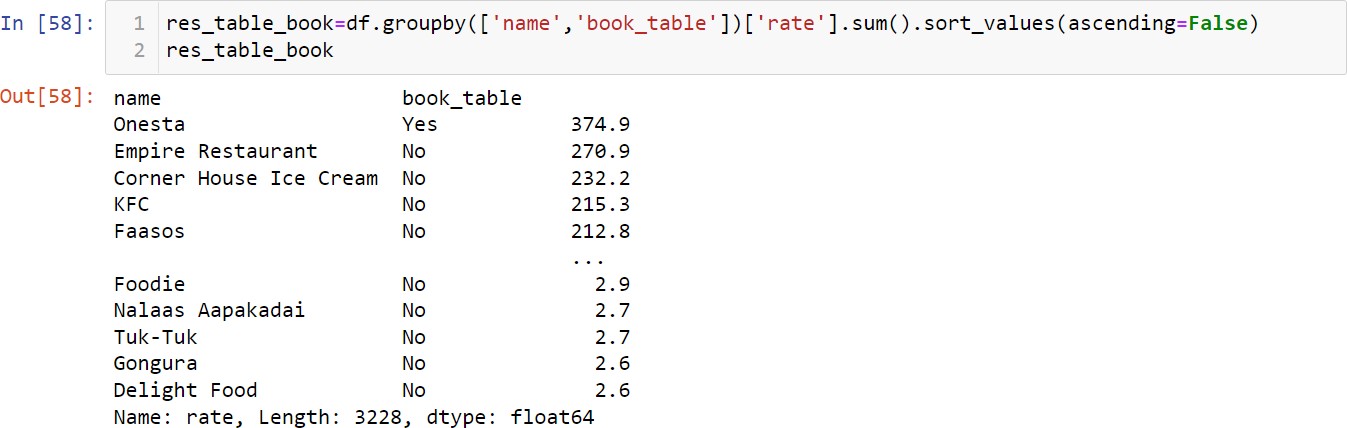


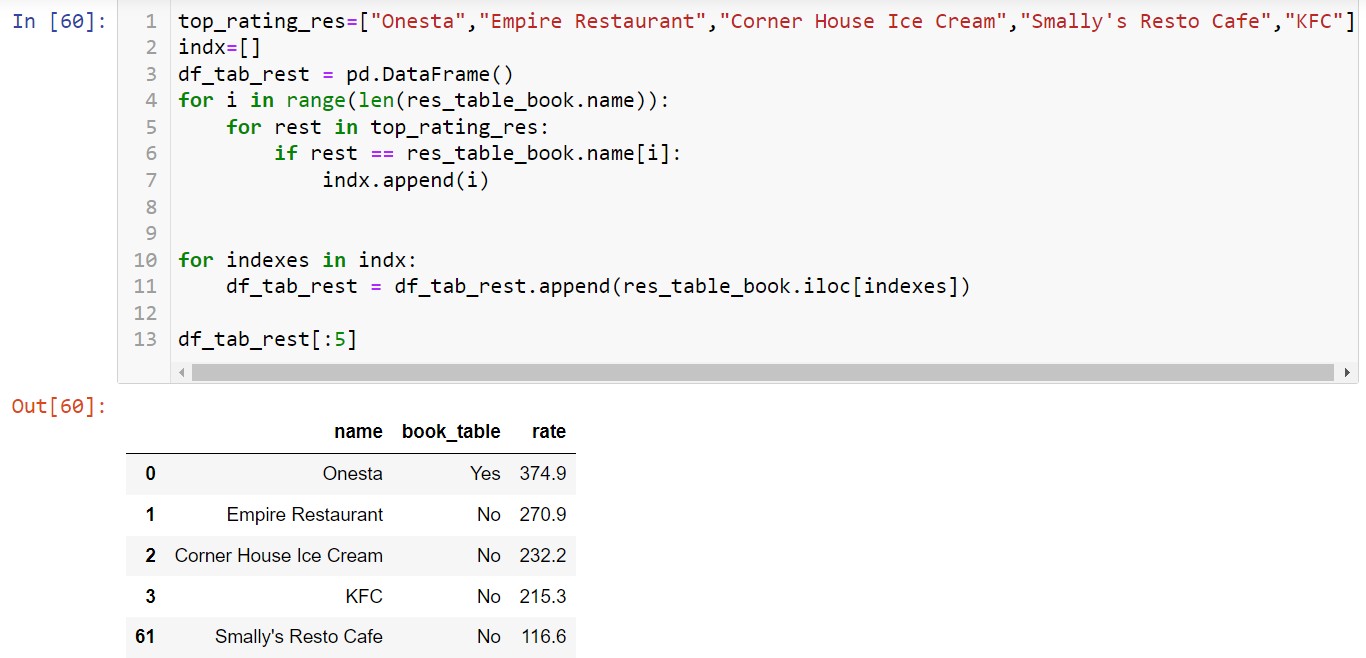


**Observation:**

* We can infer from the analysis that the 5 most top restuarants are Onesta, Empire Restaurant, Corner House Ice Cream, Smally's Resto Cafe, KFC having online order.

### Check Top Rating Restaurants having table booking option:





**Observation:**

* We can infer from the analysis that the top most restaurant Onesta having Table booking option.

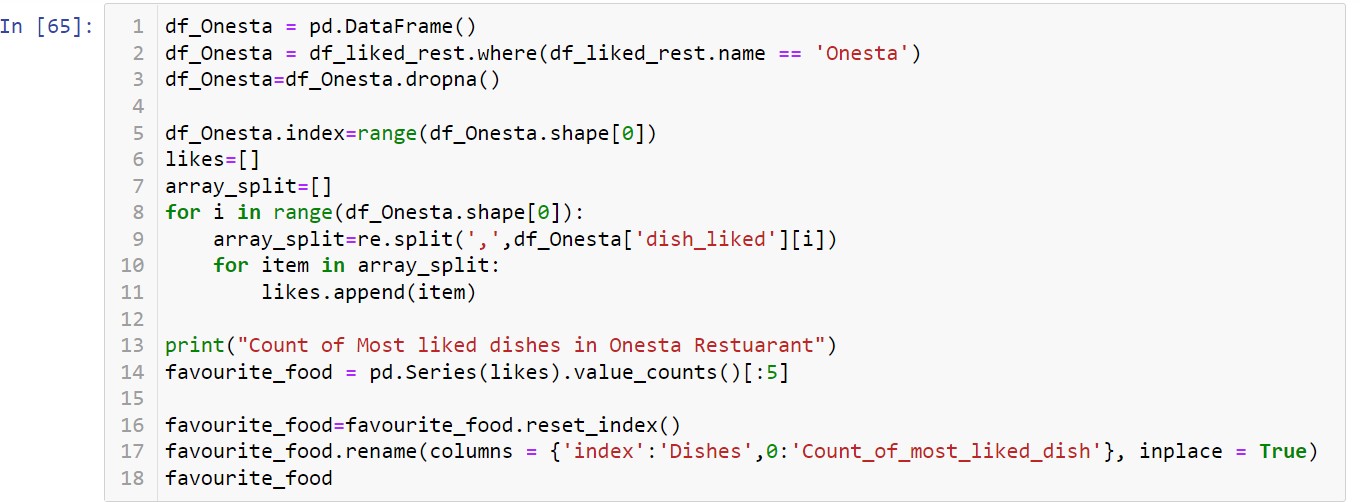
### Check Top Rating Restaurants having most liked dishes or not:

For this Analysis part, I have taken each Top Rating Restaurants separately and then check whether the most liked dishes are available or not

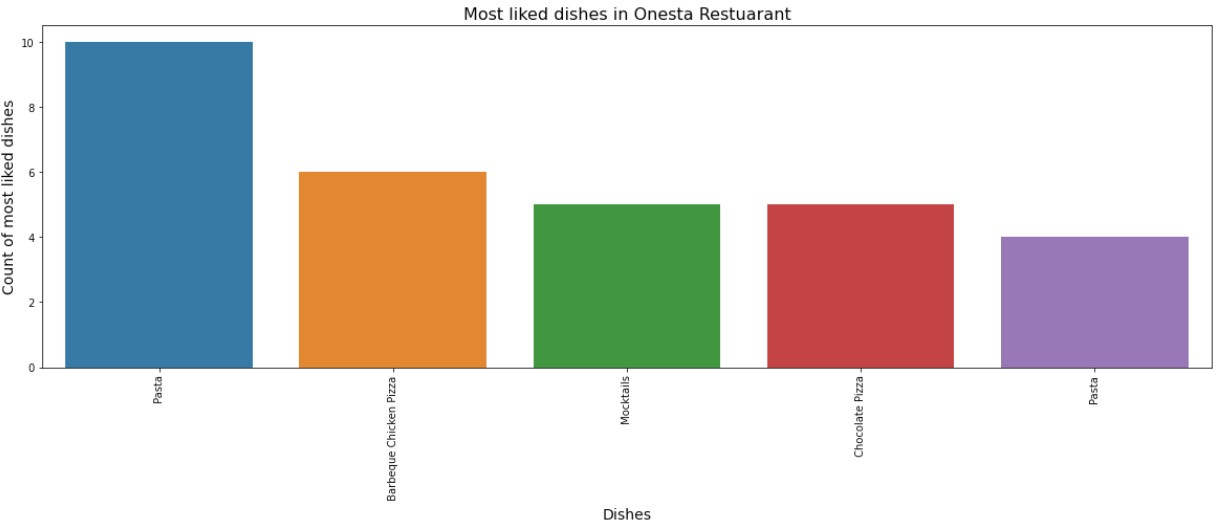




### Onesta Restaurant:



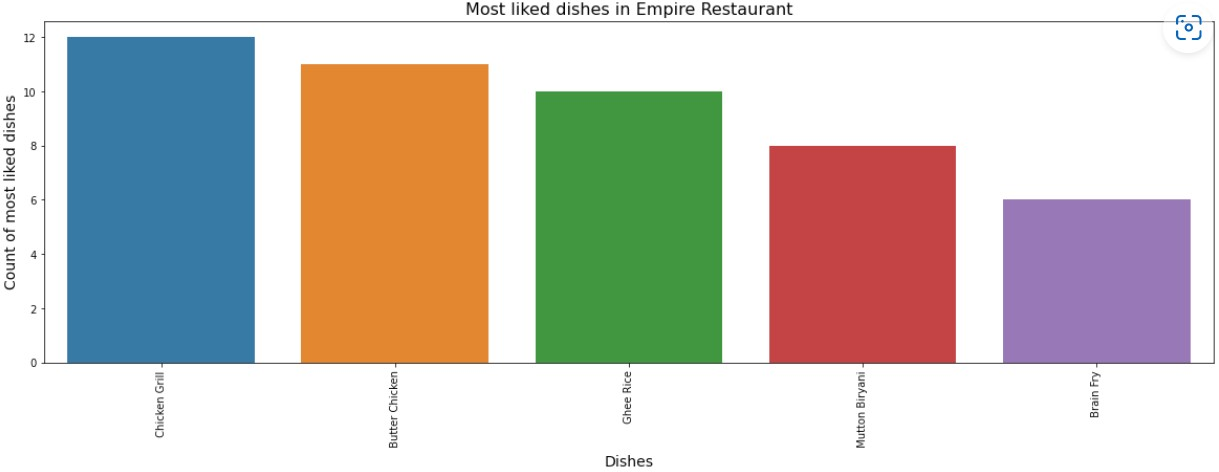




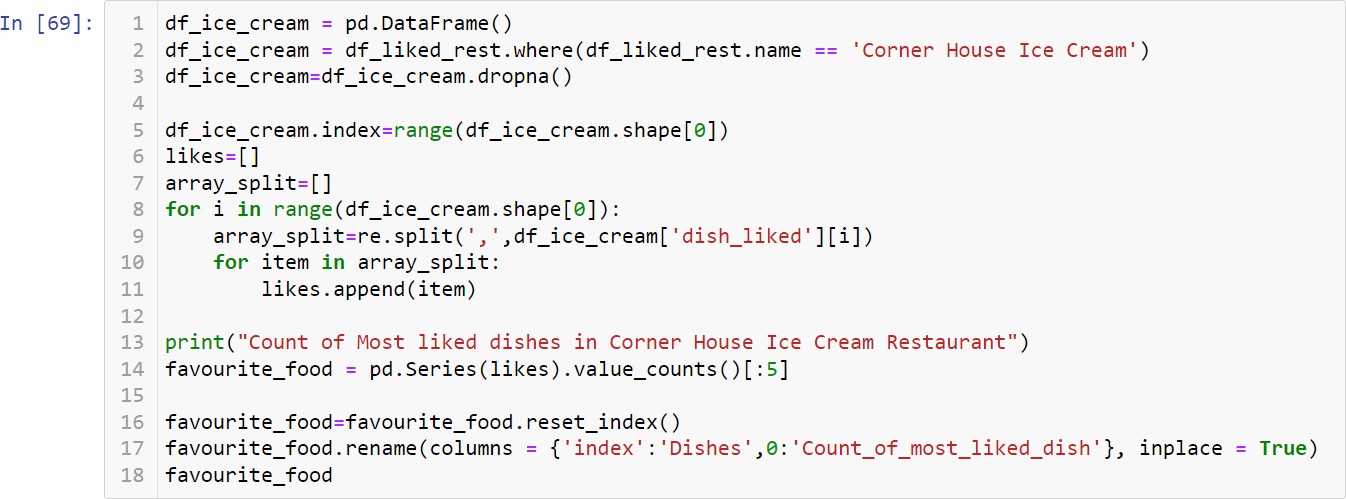
**Empire Restaurant:**



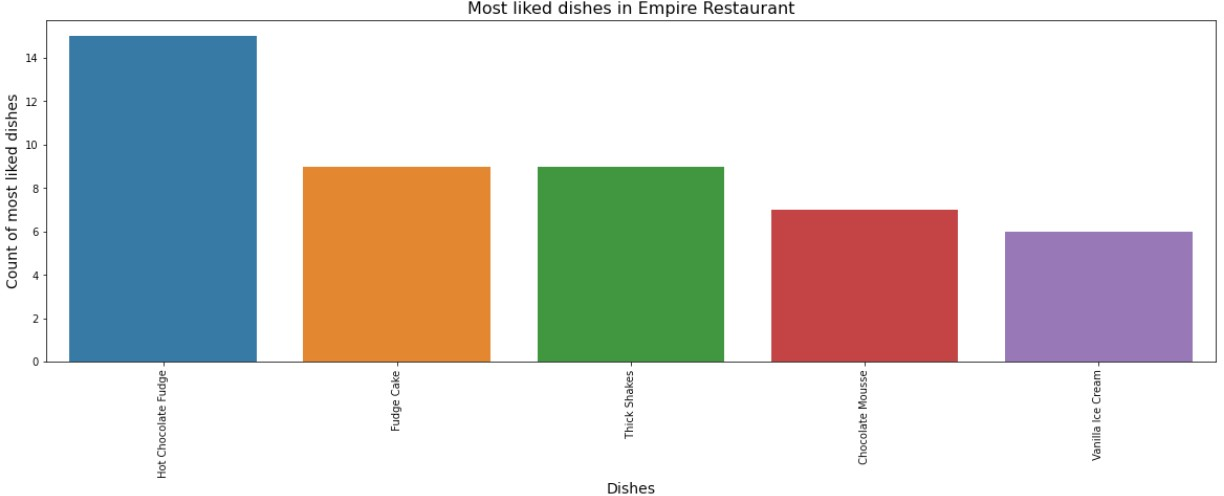




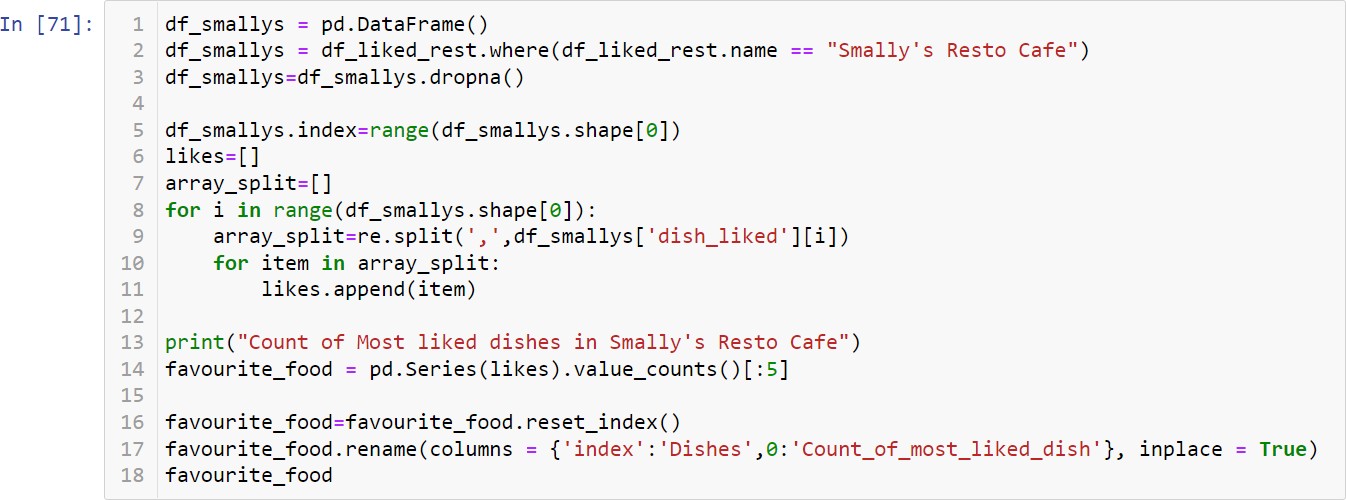
### Corner House Ice Cream Restaurant:



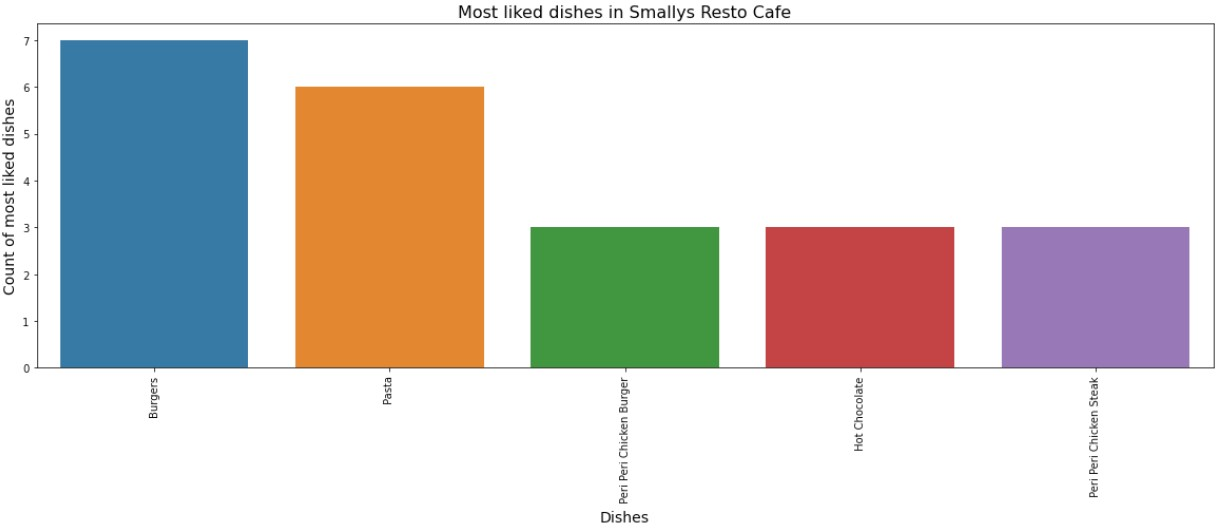




**Smally's Resto Cafe Cream Restaurant:**



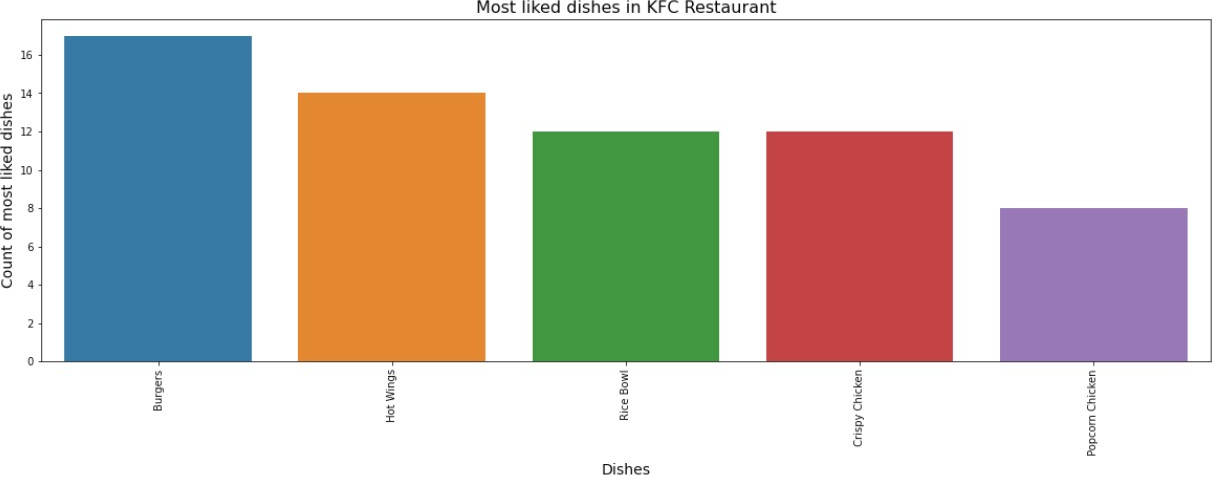




### KFC Restaurant:





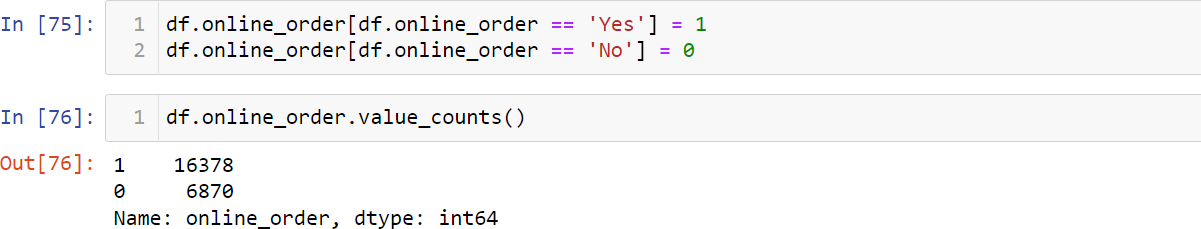


**Observation:**

* We can infer from the analysis and compare with that previously created barplot, "Count of Most Liked Dishes in Bangalore," that the top most restaurants have the favourite dishes.

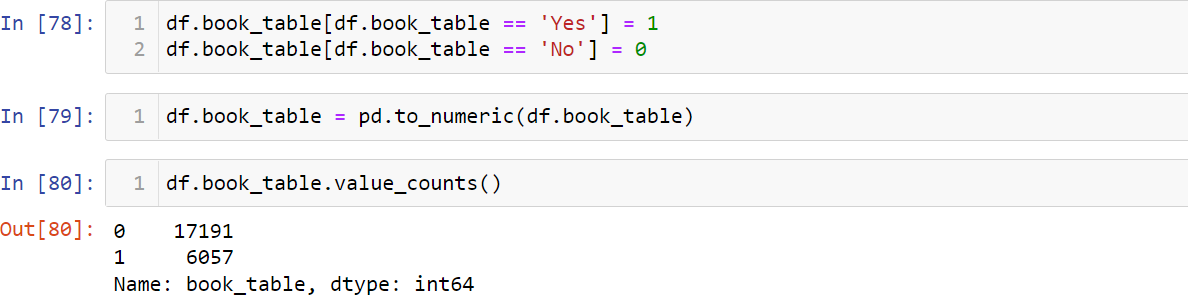
### Creation of a Model:

**Convert the online order column categorical variables into a numeric format:**

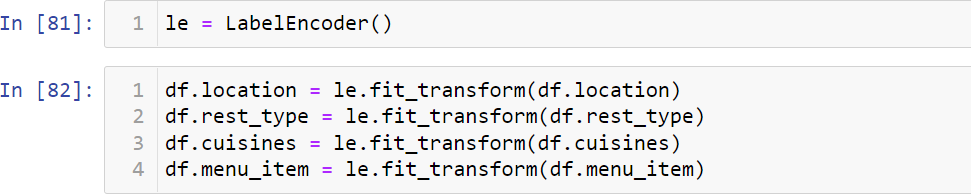




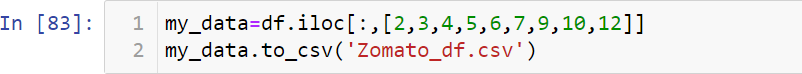
### Convert the book\_table column categorical variables into a numeric format:



**Convert the Object Datatypes to int using label encoder:**



### Create a copy of Dataset:



**Take X and Y:**

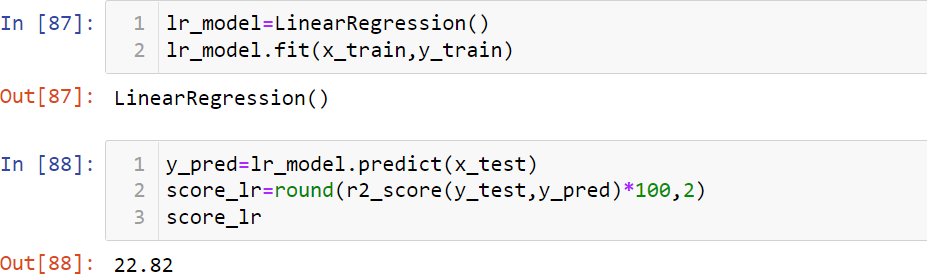




### Split the train and test part:



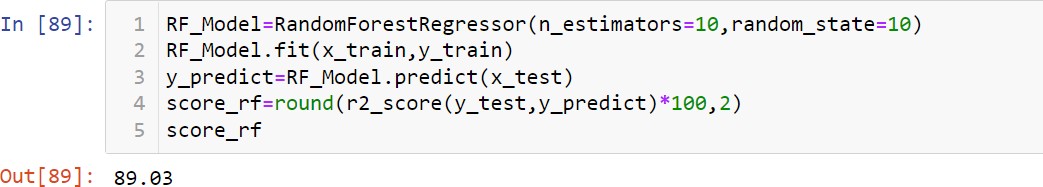
**Model 1:- Linear Regression:**



### Observation:

* 22.82% of Accuracy in Linear Regression Model.

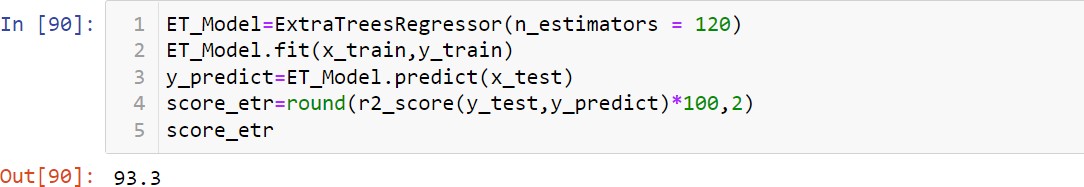
### Model 2:- RandomForest Regressor:



**Observation:**

* 89.03% of Accuracy in Random Forest Regressor Model.

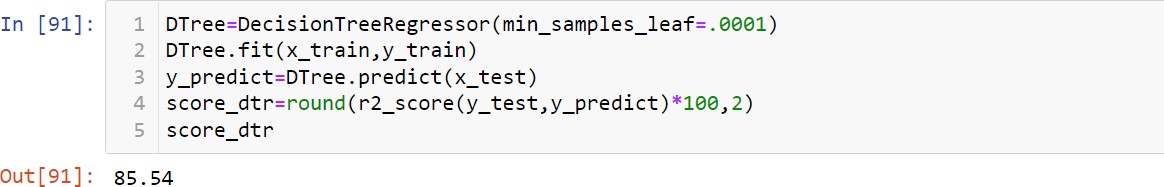
### Model 3:- ExtraTree Regressor:



**Observation:**

* 93.3% of Accuracy in ExtraTree Regressor Model.

### Model 4:- Decision Tree Regressor:

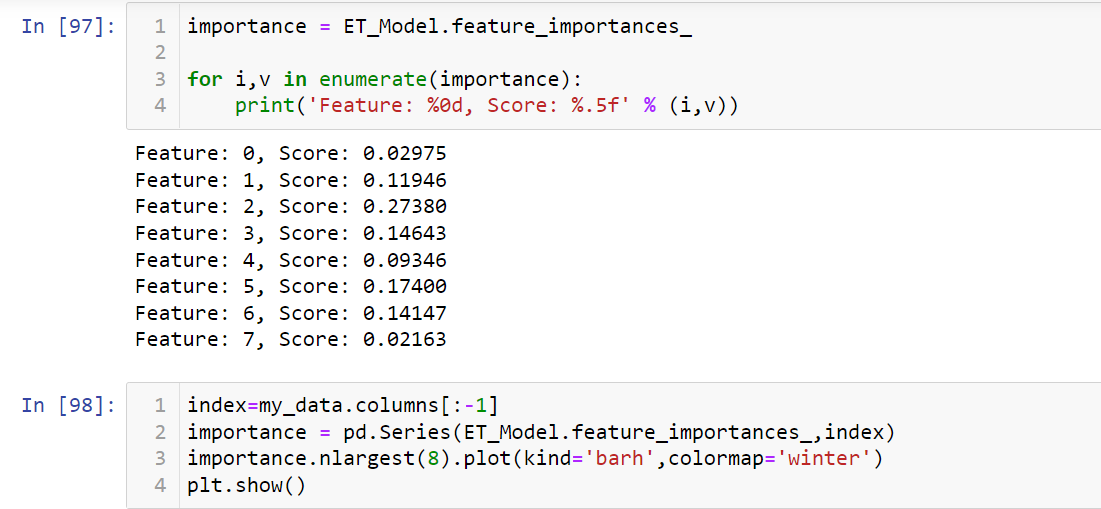


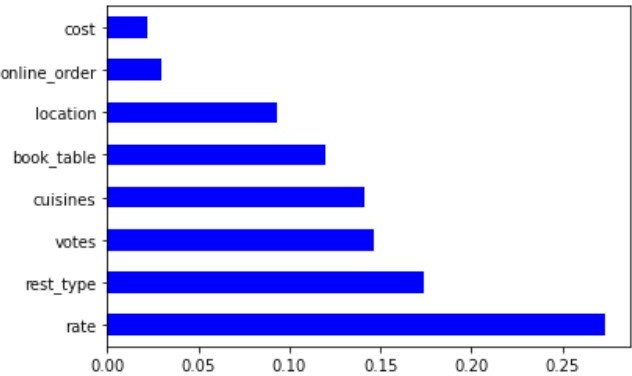
**Observation:**

* 85.54% of Accuracy in Decision Tree Regressor Model.

### Feature Importance For a ExtraTree Regressor Model:

* Feature Importance provides a score that indicates how helpful each feature was in our model.
* The higher the Feature Score, the more that feature is used to make key decisions & thus the more important it is.





### Observation:

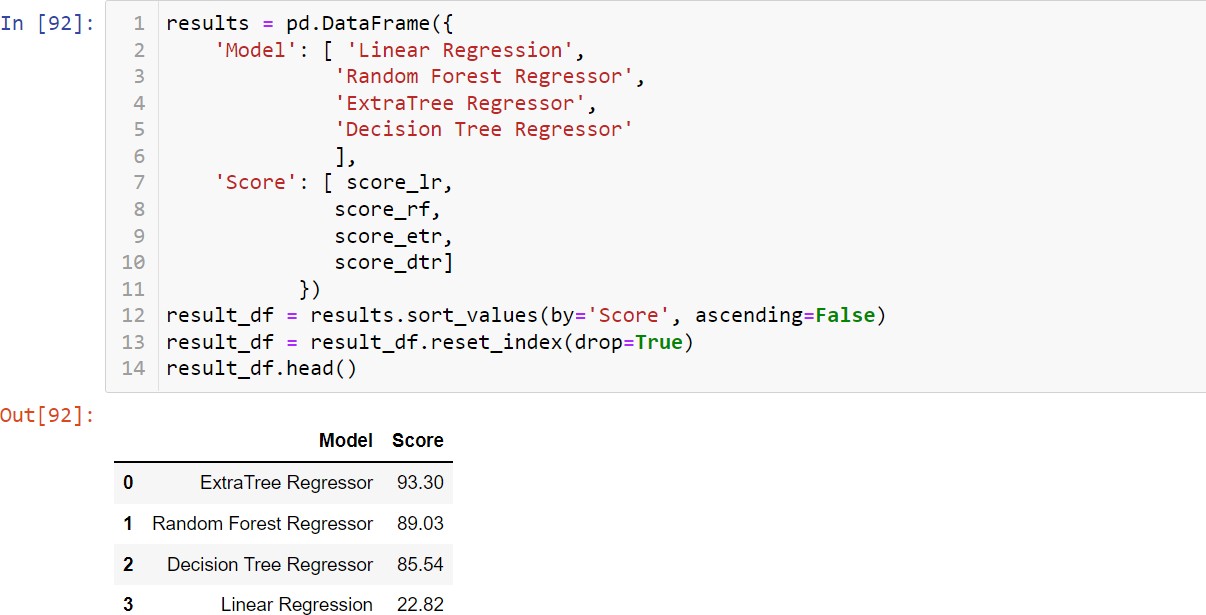
* From the above feature importance graph, we can conclude that the top 5 significant features were rate, rest\_type, votes, cuisines, book\_table.

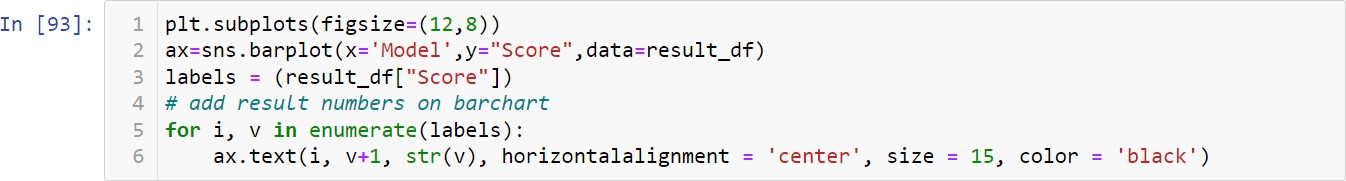
# Chapter 4

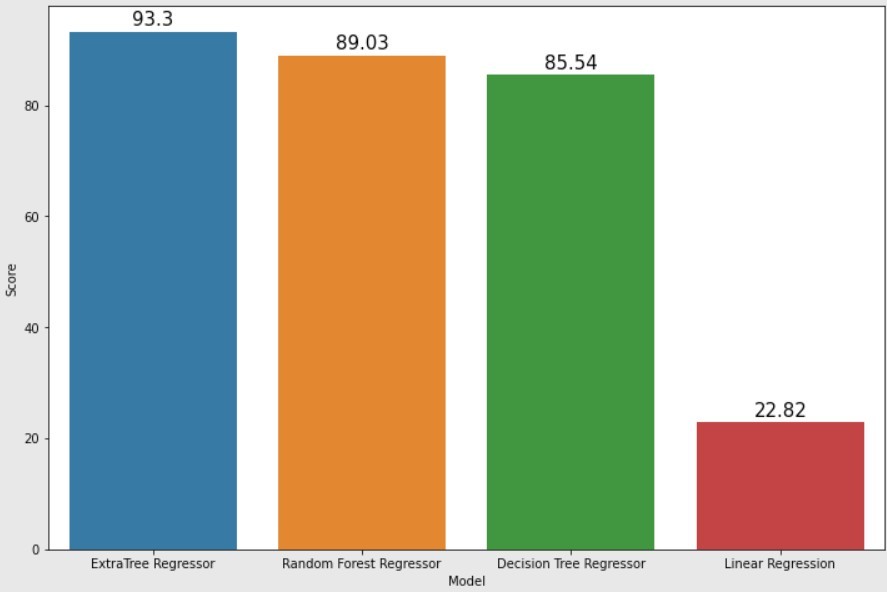
## Analysis of the Result

1. We can infer from the analysis that the top 5 restaurants are Onesta, Empire Restaurant, Corner House Ice Cream, Smally's Resto Cafe, KFC.
2. The two main service types are Delivery and Dine-out.
3. The three popularity cuisines are NorthIndian, Chinese and Continental.
4. We can infer from the analysis that the top 3 famous locations in Bangalore are BTM, Koramangala 7th Block and Koramangala 4th Block.
5. We can infer from the analysis that the 5 most liked dishes are Pasta, Pizza, Cocktails,Burgers,and Mocktails.
6. Casual Dining, Quick Bites and Cafe are the 3 most common types of Restaurants in Bangalore.

### Find the Best Model:







**Observation:**

* + ExtraTree Regressor Model having the highest Accuracy Score 93.28%
  + Random Forest Regressor Model having the Second Highest Accuracy Score 89.03%
  + Decision Tree Regressor Model having the Third Highest Accuracy Score 85.26%
  + Linear Regression Model Provide the lowest Accuracy Score 22.82%

# Chapter 5

## Conclusion

* + Our ExtraTree Regressor algorithm yields the highest accuracy, 93.28%. Any accuracy above 70% is considered good, but be careful because if your accuracy is extremely high, it may be too good to be true (an example of Overfitting). Thus, 80% is the ideal accuracy!
  + Out of the 8 features we examined, the top 5 significant features that helped us to predict the rating (i.e.) rate, rest\_type, votes, cuisines and

book\_table.

* + Our machine learning algorithm will able to predict the restaurant rating.
  + Based upon the analysis of the result, we can suggest the restaurant owners to improve the service type and what type of cuisines need to be provided and what type of dishes need to be included in the menu.
  + The two main service types are Delivery and Dine-out
  + The three popularity cuisines are NorthIndian, Chinese and Continental
  + The Below Dishes need to be included in the Menu

Pasta, Pizza, Cocktails, Burgers, Mocktails, Biryani, Sandwiches, Coffee,

Nachos, Barbeque Chicken Pizza, Chocolate Pizza, Chicken Grill, Butter Chicken, Ghee Rice, Mutton Biryani, Brain Fry, Hot Chocolate Fudge, Fudge Cake, Thick Shakes, Chocolate Mousse, Vanilla Ice Cream, Peri Peri Chicken Burger, Hot Chocolate, Peri Peri Chicken Steak, Hot Wings, Rice Bowl, Crispy Chicken, Popcorn Chicken

Here is access to the data set & code from my GitHub page:

https://github.com/Adaikkkappan/ZomatoRestaurantRating-EDA-and- MachineLearning

## References

1. Titanic dataset from Kaggle Created by **HIMANSHU PODDAR.**
2. Titanic dataset from Kaggle, Author-**FFRANKUSHA.**

https://[www.kaggle.com/code/anatosly/eda-bangalore-restaurants](http://www.kaggle.com/code/anatosly/eda-bangalore-restaurants)

1. Titanic dataset from Kaggle, Author-**ADITYA RAWAT.**

[https://www.kaggle.com/code/adityarawat10/zomato-eda-fe-model-](https://www.kaggle.com/code/adityarawat10/zomato-eda-fe-model-building) [building](https://www.kaggle.com/code/adityarawat10/zomato-eda-fe-model-building)

1. Zomato official website: https:[//www.](http://www.zomato.com/)zo[mato.com/](http://www.zomato.com/)