# Machine Learning Project on Zomato Restaurant

The restaurant industry is a dynamic and competitive space where customer preferences, dining trends, and service quality play crucial roles. In this blog post, we will analyze Zomato restaurant data to uncover insights about the dining scene. We'll explore various aspects such as restaurant types, cuisines, ratings, and cost for two. By leveraging Python and its data analysis libraries, we can derive meaningful conclusions that can help restaurant owners, food enthusiasts, and market analysts.

Introduction :-

Zomato is a popular social mobile application which connects food lovers and restaurant. Now Zomato doesn’t need an introduction as it has spread to 24 countries already. Zomato is an app which allows you to find nearby restaurants, check its menu and prices, check reviews and ratings given by users, view images and much more. Zomato was launched in 2008 by  Deepinder Goyal and Pankaj Chaddah and now it has become one of the most popular restaurant app. While Zomato grew in features and size but its core features remain the same, i.e. Discovery of restaurants and user reviews and ratings. Using the huge user base Zomato also provides sponsored content across the platform. Now Zomato has spread to over 10,000 cities across the world and is currently being used by more than 50 million people.



Dataset Link-

The datasets have been downloaded from the Github link provided below.

<https://github.com/FlipRoboTechnologies/ML_-Datasets/blob/main/Z_Restaurant/Country-Code.xlsx>

<https://raw.githubusercontent.com/FlipRoboTechnologies/ML_-Datasets/main/Z_Restaurant/zomato.csv>

Data Storage:

This problem statement contains two datasets- Zomato.csv and country\_code.csv.

Country\_code.csv contains two variables:

• Country code

• Country name

The collected data has been stored in the Comma Separated Value file Zomato.csv. Each

Restaurant in the dataset is uniquely identified by its Restaurant Id.

Every Restaurant contains the following variables:

• Restaurant Id: Unique id of every restaurant across various cities of the world

• Restaurant Name: Name of the restaurant

• Country Code: Country in which restaurant is located

• City: City in which restaurant is located

• Address: Address of the restaurant

• Locality: Location in the city

• Locality Verbose: Detailed description of the locality

• Longitude: Longitude coordinate of the restaurant&#39;s location

• Latitude: Latitude coordinate of the restaurant&#39;s location

• Cuisines: Cuisines offered by the restaurant

• Average Cost for two: Cost for two people in different currencies ��

• Currency: Currency of the country

• Has Table booking: yes/no

• Has Online delivery: yes/ no

• Is delivering: yes/ no

• Switch to order menu: yes/no

• Price range: range of price of food

• Aggregate Rating: Average rating out of 5

• Rating color: depending upon the average rating color

• Rating text: text on the basis of rating of rating

• Votes: Number of ratings casted by people

Problem statement

**In this dataset predict 2 things –**

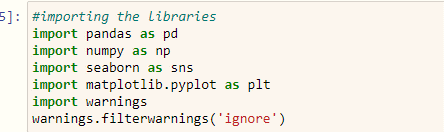
1) Average Cost for two

2) Price range

The price range spans from one to four, with four indicating premium-priced restaurants.

Data Analysis : Load, Clean, Format :

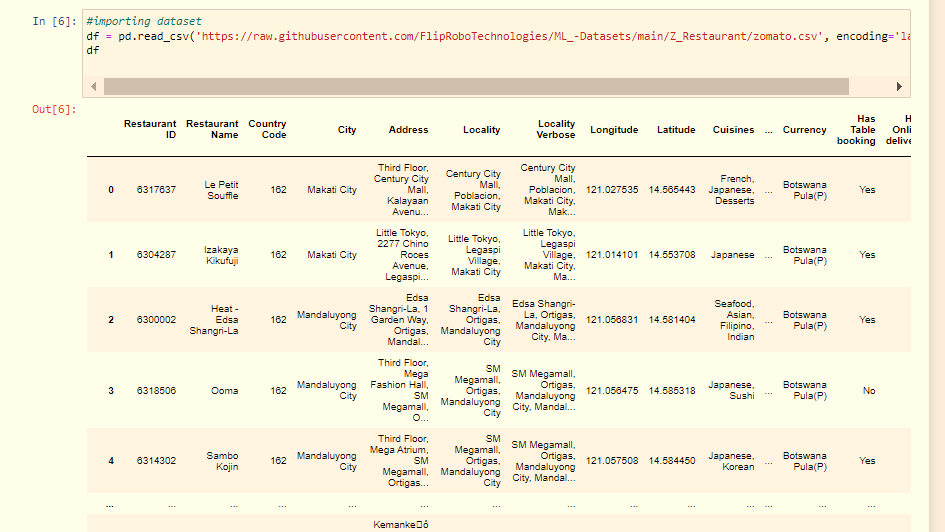
Importing necessary libraries.



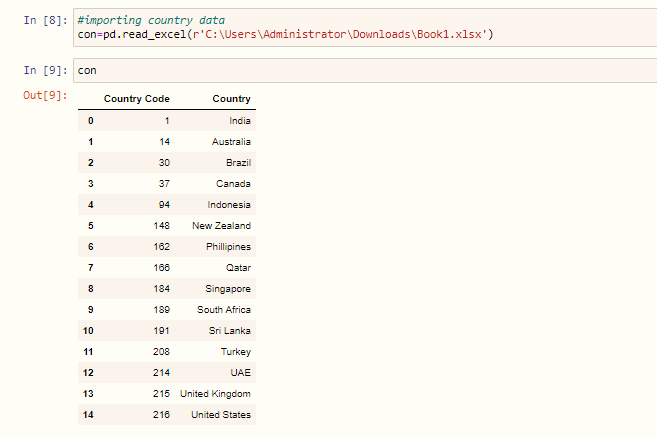
Pandas is utilized for data reading and analysis, while Numpy assists with numerical analysis. Matplotlib and Seaborn are employed for data visualization.

Loading dataset of Zomato :

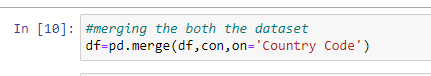
The dataset has been successfully loaded utilizing pandas.read\_csv.



Similarly we load the Country dataset :-

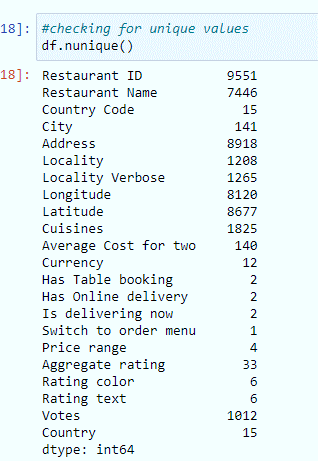
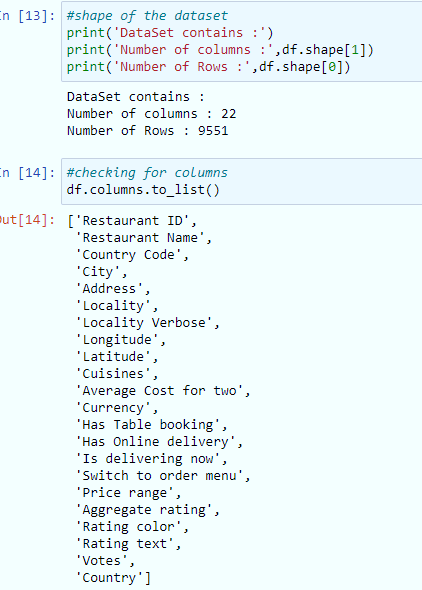


Afterward, the merge function from pandas is utilized to combine both datasets based on the common 'Country Code' column.



Exploring the data :

Conducting a comprehensive analysis of the dataset structure by examining the columns, identifying unique values, and assessing the data types for each attribute.

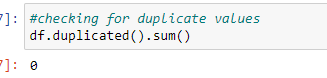


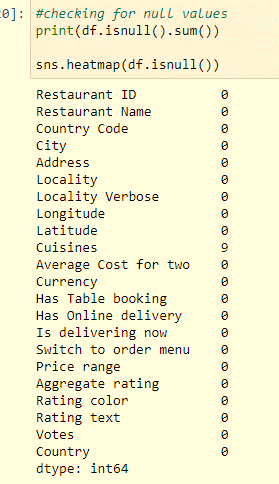
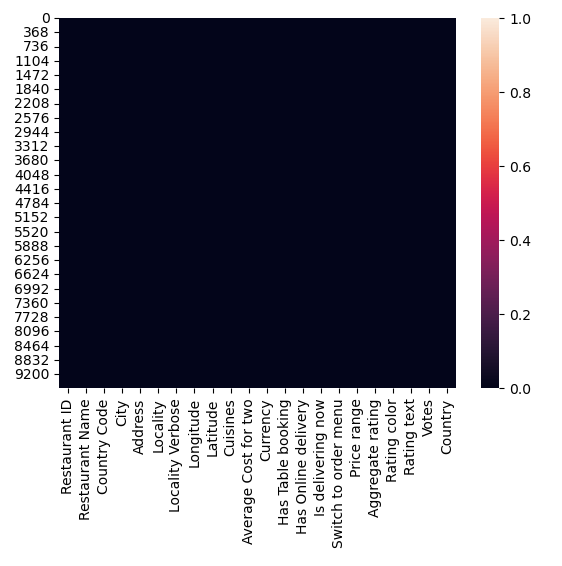
Insights:-

* The dataset comprises 9,551 rows and 22 columns. Among these, 3 columns are of float data type, 5 columns are of integer type, and 14 columns are of object data type.
* Our target features, 'Average cost for two' and 'Price range', both belong to the integer data type.
* 'Price range' exhibits 4 distinct values, while 'Average cost for two' shows 140 distinct values.
* Uniform counts across all features suggest the absence of null values.

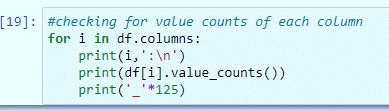
Data Preprocessing :

Proceeding with our analysis of the data, our next step involves data cleaning. This encompasses identifying and addressing null values, duplicate entries, whitespace, and other similar issues.



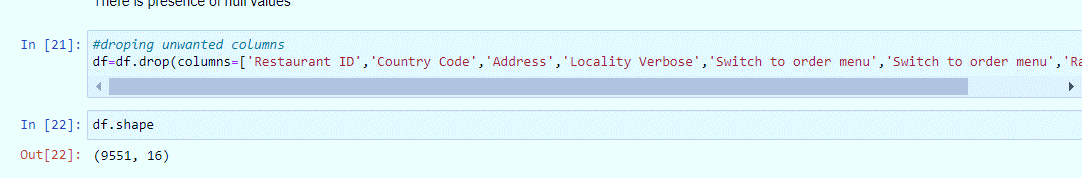
 

Fortunately, the dataset contains no null or duplicate values, which simplifies the data analysis process.



This provides the value counts for each feature, allowing us to review all unique values and identify any potential invalid data. Fortunately, this dataset does not contain any such issues.

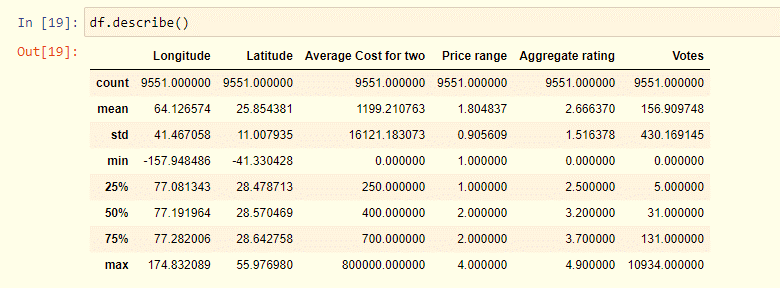
Additionally, we removed some unnecessary columns that did not provide any useful information.



Statistical Analysis : -

In this section, we will review key statistical parameters such as percentiles, minimum and maximum values, standard deviation, counts, mean, and median. Analyzing these metrics will provide us with valuable insights into the dataset, including its skewness and potential outliers.

Let's review the statistical analysis of our Zomato restaurant dataset :



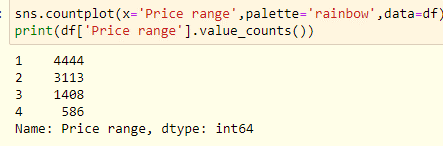
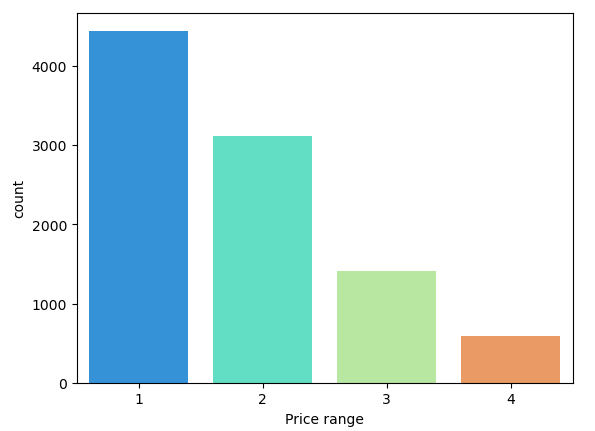
Key insights derived from the above table are :-

* All counts are same there are no null values.
* There is a difference between mean value and median (50%) value which indicates that datasets are skewed.
* There is a large gap between maximum value and 75% percentile which means there maybe the presence of outliers.
* Minimum vote is 15.90 and maximum is 10934
* Minimum aggregate rating is 0 and maximum is 4.9
* Minimum Price range is 1 and maximum price range is 4.

Exploratory Data Analysis (EDA) :-

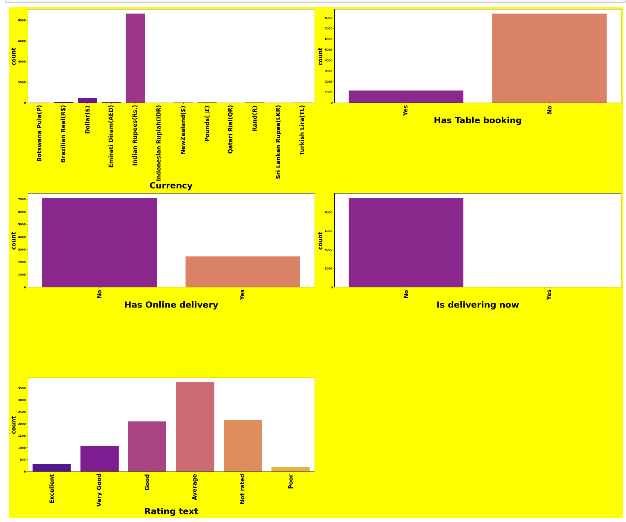
In this section, we will conduct a comprehensive analysis of all features and target variables.

Let's initiate our analysis by exploring the 'Price range', our target variable, using a count plot.

The distribution shows a higher frequency in price ranges 1 and 2, with a lower occurrence in range 4. This indicates that a greater number of cuisines are associated with lower price ranges, specifically 1 and 2, while fewer cuisines fall within the higher price range category of 4.

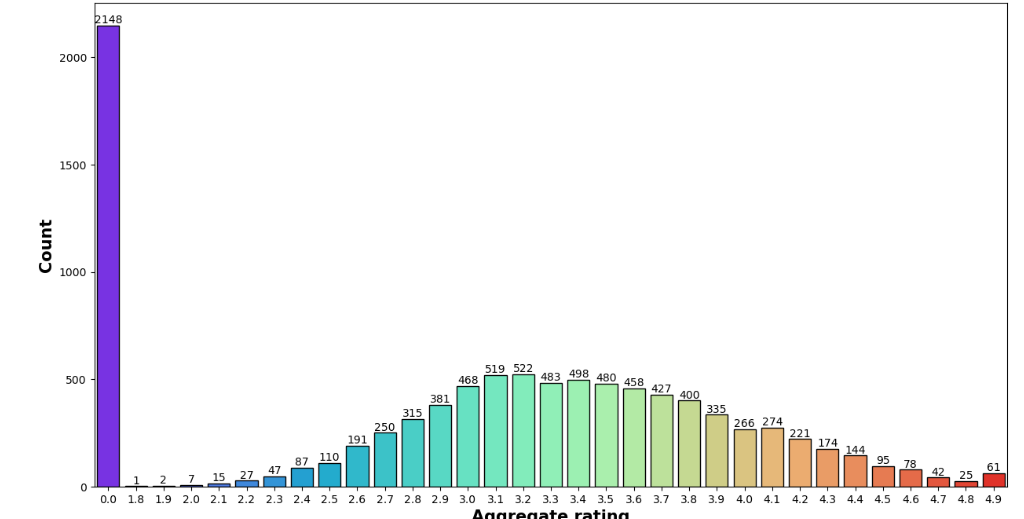
Visualizes the distributions of various feature values through count plots:



Observations:-

* There is limited availability for table bookings, suggesting that a significant portion of restaurants do not offer this service to their customers.
* Similarly, online delivery options are also limited, indicating that a substantial number of restaurants do not provide this convenience to their patrons.
* The majority of restaurant ratings tend to cluster around categories such as 'Average', 'Not Rated', and 'Good'. This suggests that most diners have experiences that are generally satisfactory or better.
* Conversely, ratings categorized as 'Poor' receive the least frequency, indicating that fewer restaurants are perceived negatively by customers, based on the available data.

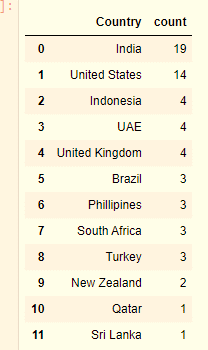
Visualization of Aggregate Ratings Using Count plot:-



Observations:-

* **Dominant Rating Trend**: The prevailing trend in restaurant ratings reveals that a significant number of establishments have garnered a rating of 0, underscoring a broad base of feedback.
* **Minimal Ratings**: Conversely, a minority of restaurants have received the minimum rating of 1.8, indicating a sparse distribution in lower ratings.
* **Highest Acclaim**: Impressively, 61 restaurants have achieved the pinnacle rating of 4.9, highlighting exceptional customer satisfaction and acclaim within the dining landscape.

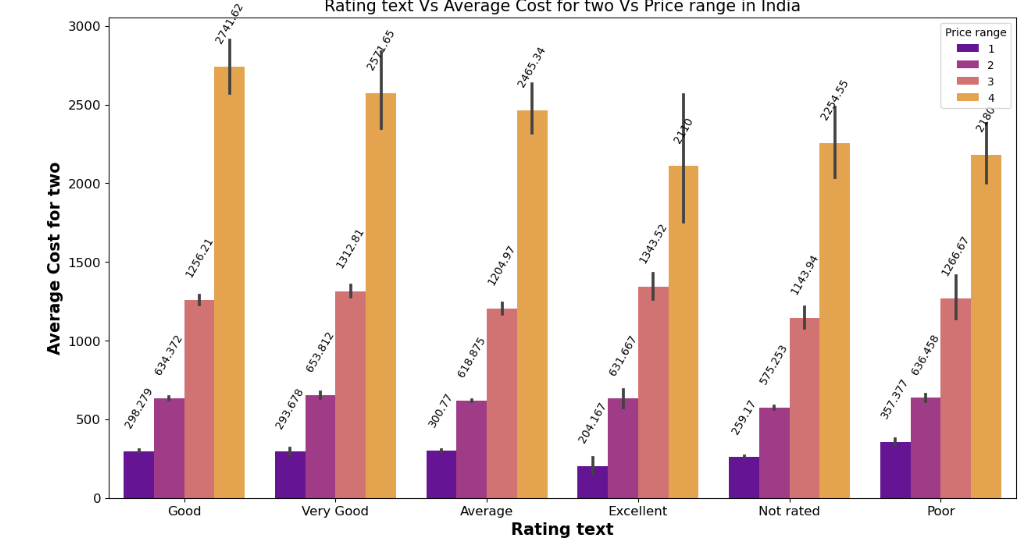
Aggregate Ratings across Countries:-



Restaurants based in India and the United States consistently attained the highest ratings, frequently earning an impressive score of 4.9.

Let’s go through these countries.

Rating text vs. Average Cost for two vs. Price range in India:

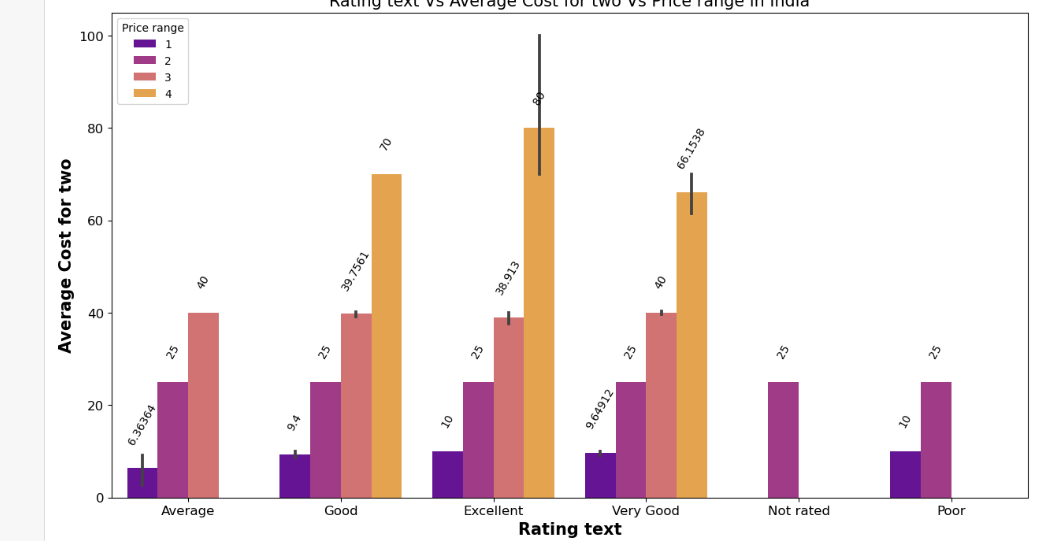


Observations:-

In the Indian market:

* Price range 4 garnered the highest number of "Good" ratings.
* Price range 3 received the most "Excellent" ratings.
* Price range 2 achieved the highest number of "Very Good" ratings.
* Price range 1 garnered the most "Poor" ratings.

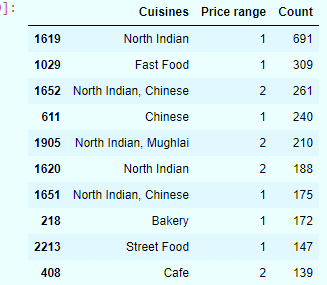
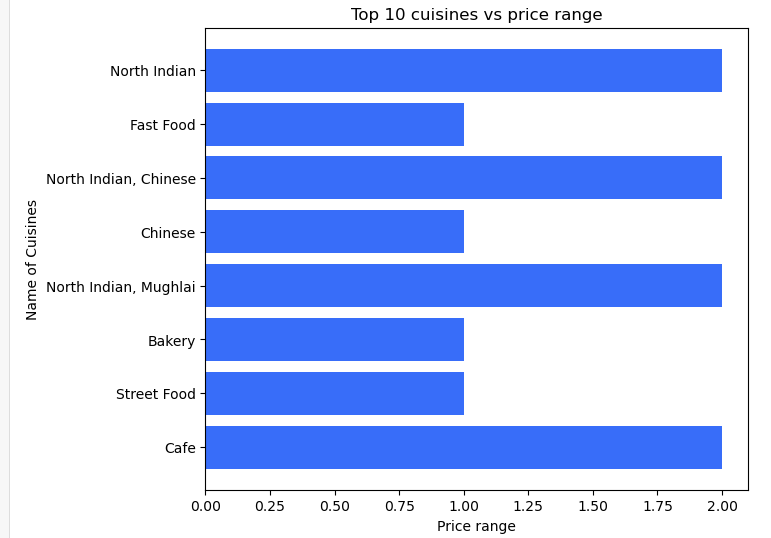
Rating text vs. Average Cost for two vs. Price range in United States:



Observations:-

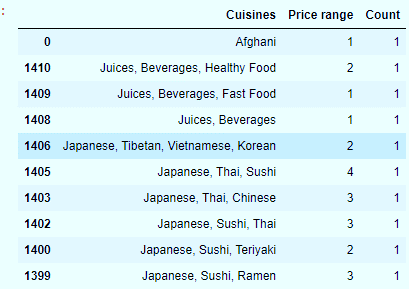
* Price range 4 garnered the highest number of "Excellent" ratings.
* Price range 3 received the most "Average" and "Very Good" ratings.
* Price range 2 received a diverse distribution across various ratings.
* Price range 1 received the highest number of "Poor" and "Excellent" ratings.

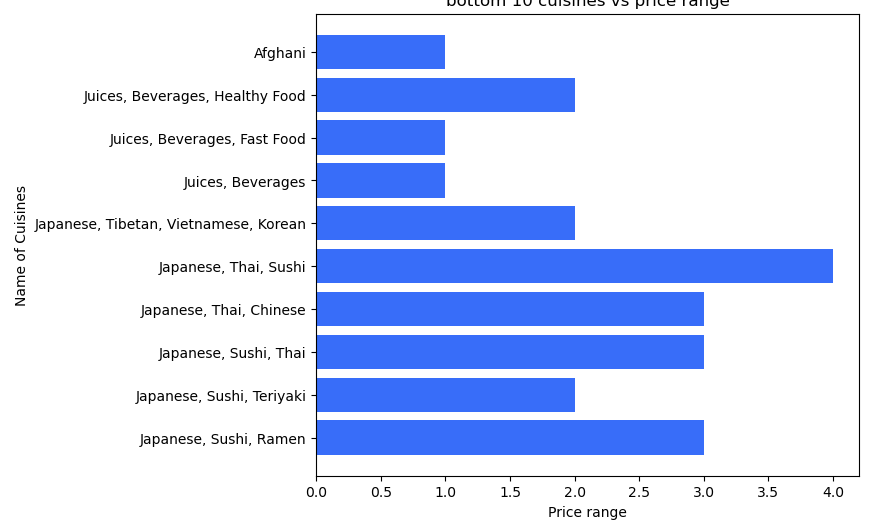
Top 10 Cuisine-Price Range Counts:-

Observations:-

* **North Indian Cuisine (Price Range 1)**: Dominates the lower price range, indicating widespread popularity among budget-conscious diners.
* **Fast Food (Price Range 1)**: Also prevalent in the lower price range, showcasing its appeal as a convenient and affordable dining option.
* **North Indian, Chinese Cuisine (Price Range 2)**: Found frequently in the moderate price range, offering a mix of North Indian and Chinese flavors at slightly higher prices.
* **Chinese Cuisine (Price Range 1)**: High popularity in the lower price range, highlighting its consistent demand for economical dining experiences.
* **North Indian, Mughlai Cuisine (Price Range 2)**: Popular in the moderate price range, known for its rich and distinctive dishes that cater to varied tastes.
* **Bakery Items (Price Range 1)**: Favored in the lower price range, indicating a preference for bakery products as quick snacks or light meals.
* **Cafes (Price Range 2)**: Prominent in the moderate price range, cafes are popular for offering moderately priced meals and beverages in a relaxed setting.

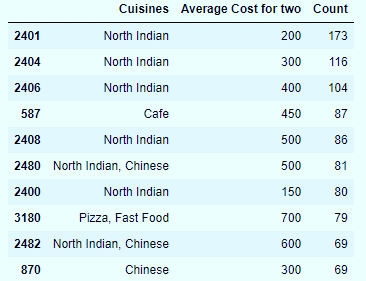
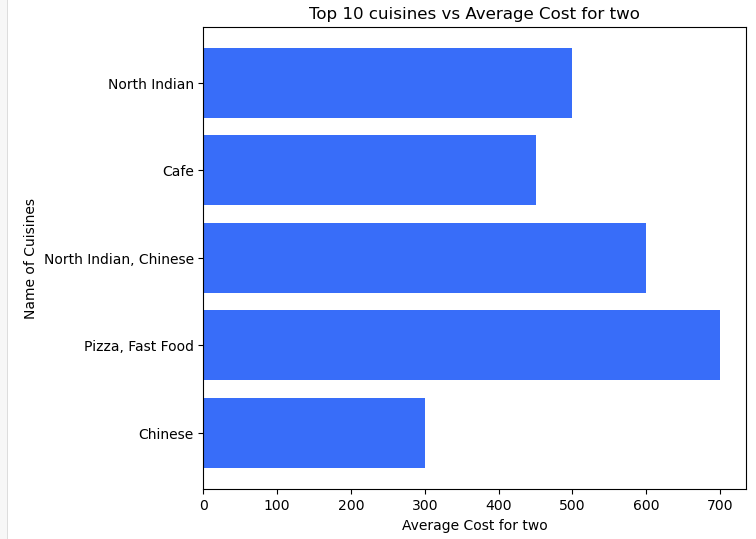
Top 10 Least Popular Cuisine-Price Range Counts:-



Observations:-

* **Afghani Cuisine (Price range 1)**:
* There is only one restaurant offering Afghani cuisine in the lowest price range. This suggests limited availability or demand for Afghani cuisine in this pricing category.
* **Juices, Beverages, Healthy Food (Price range 2)**:
* There is one restaurant offering a combination of juices, beverages, and healthy food in the moderate price range. This indicates a niche market catering to health-conscious consumers who are willing to pay slightly more for nutritious options.
* **Japanese, Thai, Sushi (Price range 4)**:
* A single restaurant offering Japanese, Thai, and Sushi dishes in the highest price range suggests a premium dining experience combining diverse Asian cuisines. This may appeal to customers seeking high-quality, authentic Japanese and Thai culinary experiences.

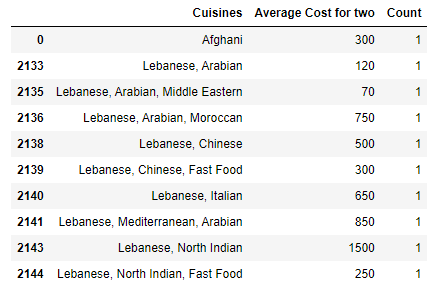
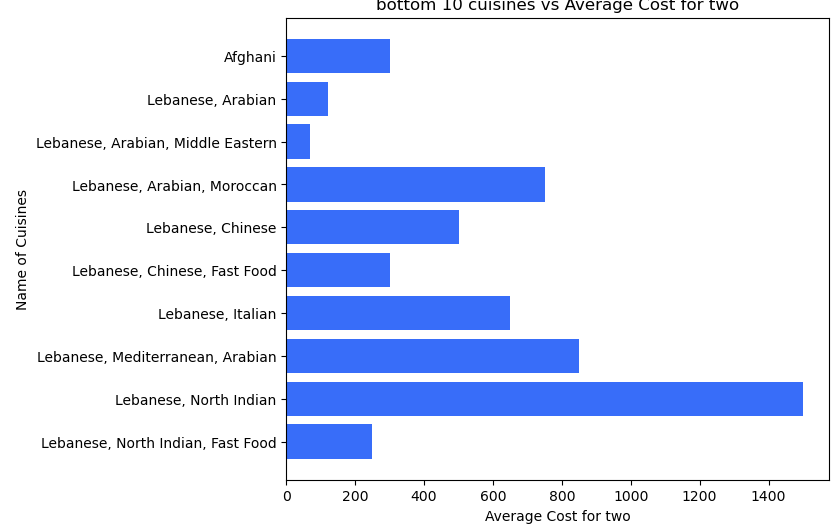
Top 10 Cuisine Average Cost for two Counts:-

Observations:-

* **North Indian Cuisine**:
* There is a significant presence across various price ranges (200, 300, 400, 500), with 200 and 300 being the most common.
* This suggests that North Indian cuisine is popular across different budget preferences, catering to a wide range of dining needs.
* **Cafe Cuisine**:
* Typically priced around 450, with a substantial count of 87.
* Cafes often offer a moderate cost option, appealing to customers looking for a relaxed dining experience without high expenditures.
* **Pizza and Fast Food**:
* Priced at 700, with a count of 79.
* Fast food and pizza joints at this price point indicate a preference for quick, casual dining with higher average spending.
* **Chinese Cuisine**:
* Offers options at 300 and 600, both with a count of 69.
* Chinese cuisine caters to both budget-conscious diners and those willing to spend more for a fuller dining experience.

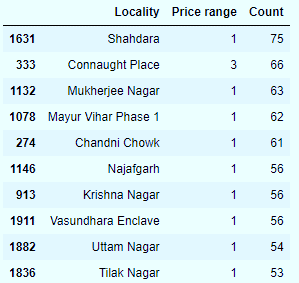
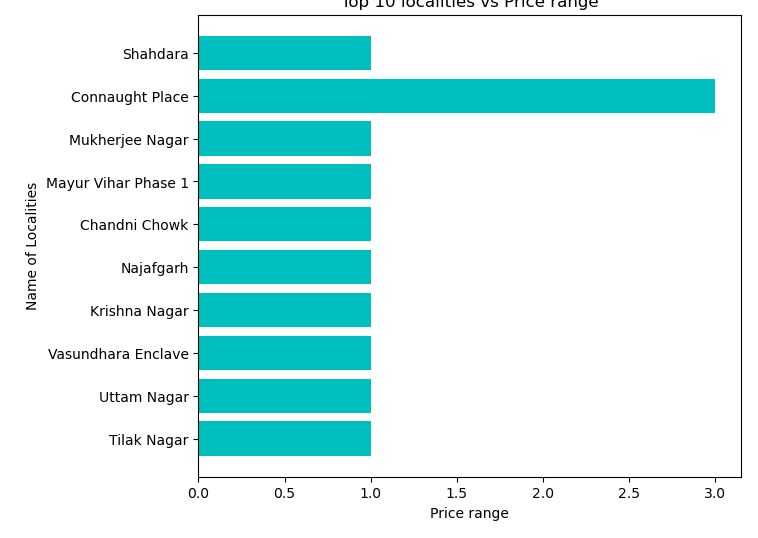
Top 10 Least Popular Cuisines Average Cost For Two Counts:

Observations:-

* Afghani cuisine is priced moderately at $300.
* Lebanese and Arabian cuisines are available at lower costs of $120 and $70 respectively, appealing to budget-conscious diners.
* Higher-priced options include Lebanese, Moroccan, and Mediterranean blends, reflecting a focus on premium ingredients and culinary experiences.

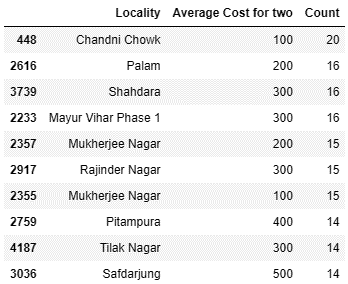
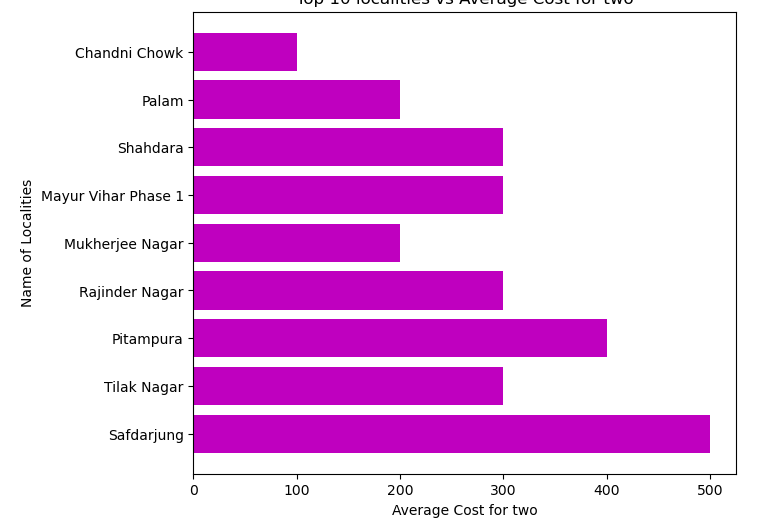
Top 10 Localities by Price Range Counts:-

Observations:-

* **Shahdara**:
* **Price Range 1**: Shahdara has 75 restaurants in this price range, indicating a significant number of affordable dining options in the area.
* **Connaught Place** :
* **Price Range 3**: With 66 restaurants offering higher-priced options, Connaught Place is known for its diverse dining scene catering to various tastes and budgets.
* **Mukherjee Nagar**:
* **Price Range 1**: Mukherjee Nagar follows closely with 63 restaurants in the lower price range, suggesting it also offers a range of budget-friendly dining choices.
* **Mayur Vihar Phase 1**, **Chandni Chowk**, **Najafgarh**, **Krishna Nagar**, **Vasundhara Enclave**, **Uttam Nagar**, **Tilak Nagar**:
* These localities predominantly have restaurants in **Price Range 1**, indicating a preference or market demand for more affordable dining options in these areas.

Top 10 Localities by Restaurant Count and Average Cost:-

Observations:-

* Chandni Chowk has 20 restaurants with an average cost of 100, indicating a concentration of budget-friendly options.
* Palam, Shahdara, Mayur Vihar Phase 1, and Rajinder Nagar each host 16 restaurants priced at 200 or 300, reflecting moderate pricing trends.
* Mukherjee Nagar offers 15 restaurants at both 100 and 200 average costs, providing diverse dining options.
* Pitampura, Tilak Nagar, and Safdarjung feature restaurants with higher average costs of 300 to 500, suggesting a mix of mid-range to upscale dining choices in these areas.

EDA Concluding Remarks :-

**From all above analysis we can conclude that:**

* Restaurant ratings vary widely, with many achieving high ratings like 4.9, indicating strong customer satisfaction.
* North Indian cuisine is popular across different price ranges, from budget-friendly to upscale dining.
* Cafes are consistent in offering moderately priced meals, making them attractive for relaxed dining.
* Fast food and pizza places priced higher at $700 cater to those looking for quick meals with a bit more expense.
* Localities such as Shahdara and Mukherjee Nagar offer a mix of dining options across various price points, reflecting diverse culinary preferences and local demands.

Data Pre-processing ,Feature engineering :-

Data pre-processing is an essential phase in data analysis that significantly influences the performance of machine learning models. The key to accurate model predictions lies in the careful selection and preparation of features used for training. Below are the critical steps involved in this process:

**1. Imputing Missing Values** Addressing missing data is crucial to maintain dataset integrity. Common methods include filling missing values with the mean, median, or mode, or using advanced techniques like k-nearest neighbors (KNN) imputation to predict missing values based on similar data points.

**2. Scaling Numeric Features** Scaling ensures that numerical features are on a similar scale, which is important for algorithms sensitive to feature magnitude. Techniques like standardization (z-score normalization) or min-max scaling (rescaling to a range) are often used to normalize data.

**3. Finding and Removing Outliers** Outliers can skew model training and reduce performance. Identifying outliers through statistical methods (e.g., z-scores, IQR) or visualizations (e.g., box plots) and removing or transforming them helps in creating a more robust model.

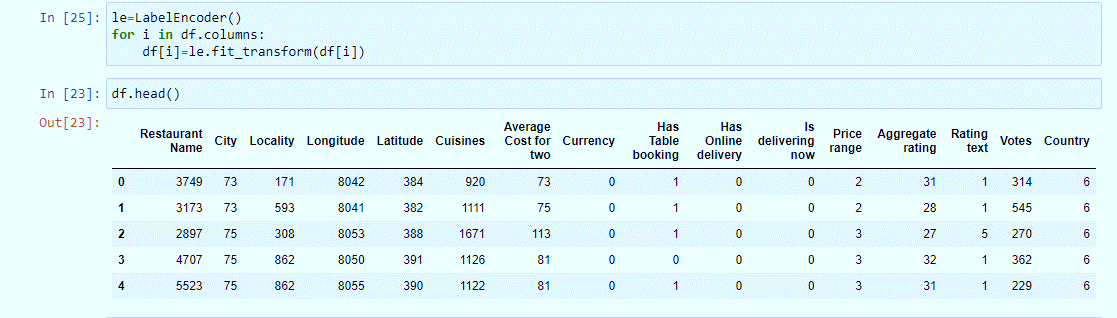
**4. Encoding Categorical Variables** Machine learning models require numerical input, so categorical variables need to be converted. Techniques like one-hot encoding, label encoding, or target encoding are used to transform categorical data into a numerical format that models can interpret.

We will now proceed with the necessary steps for our subsequent analysis.:-

1. Encoding categorical data :-

Machine learning models require numerical input data. Therefore, we encode categorical data into numeric forms using techniques such as one-hot encoding, label encoding, and target encoding.

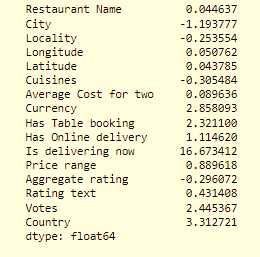
For this case I have used label encoder to convert categorical data.



After encoding, we will proceed with addressing the skewness and outliers in the dataset.

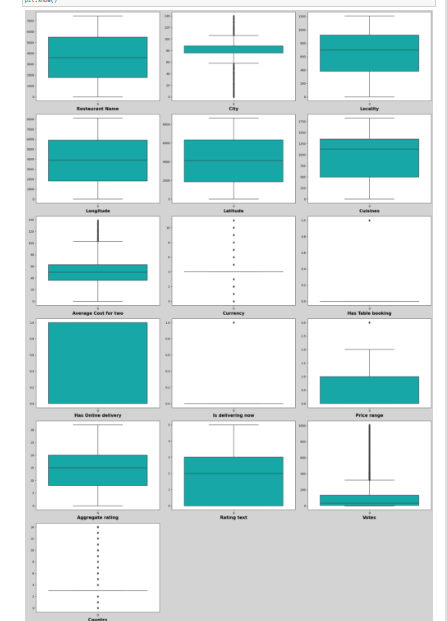
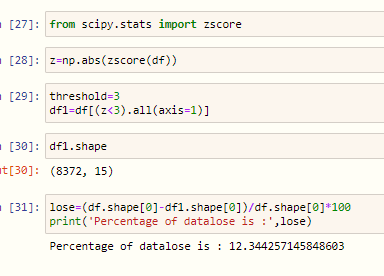
1. Skewness and Outlier Detection and Removal:-

Skewness in model building refers to the imbalance or asymmetry in the distribution of data. It can affect how well machine learning algorithms perform, especially those that assume data is normally distributed. Addressing skewness through data transformations like logarithmic or square root transformations helps ensure that models can better understand and utilize the data, leading to more accurate predictions and insights.

The data spread is depicted in the figure, with skewness values provided in a corresponding table. Given that a skewness value of 7 is considered standard, the column 'is delivering now' exhibits notably high skewness, leading us to decide to remove this column from furtheranalysis. 

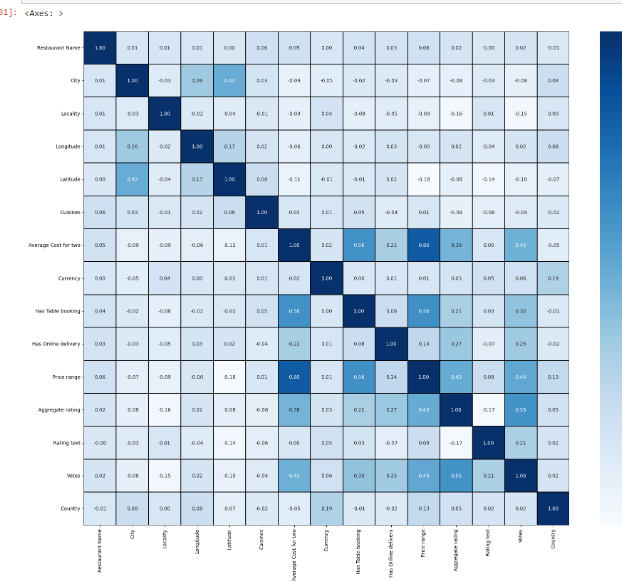
The presence of outliers can significantly decrease model accuracy and potentially mislead the results. Techniques such as box plots, z-scores, and the Interquartile Range (IQR) method are effective in identifying and managing outliers.

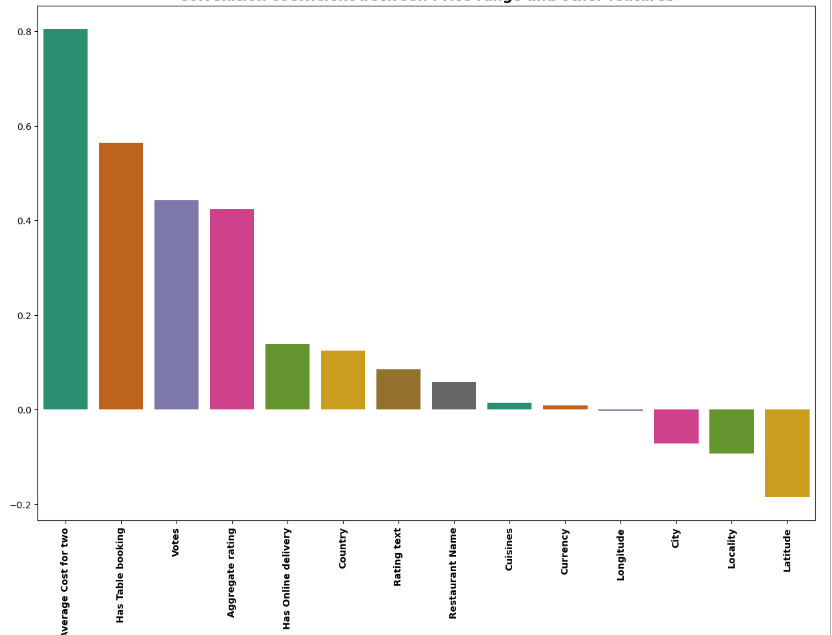
Outliers in certain columns are identified and managed using the z-score method.

1. Correlation :-

Correlation refers to the statistical measure that describes the strength and direction of a relationship between two numerical variables. It quantifies how much one variable changes with respect to changes in another variable. A correlation coefficient close to +1 indicates a strong positive correlation, -1 indicates a strong negative correlation, and 0 indicates no correlation. Correlation is often visualized using tools like scatter plots or heatmaps to understand the interdependencies between variables in a dataset.



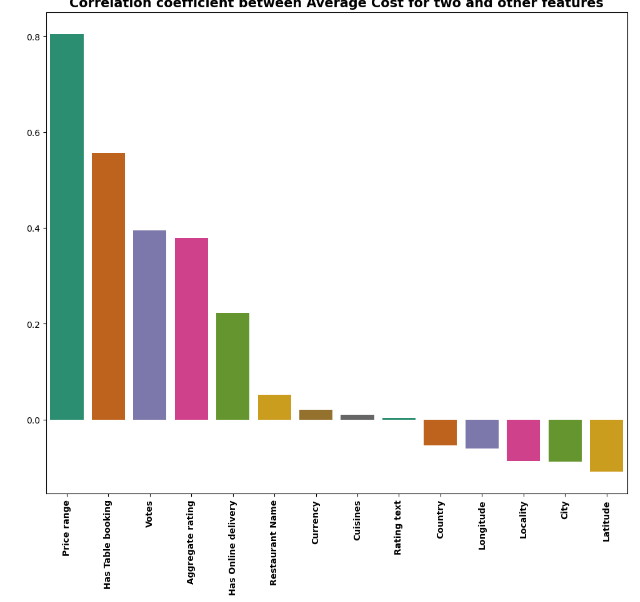
**Correlation coefficient between Price range and other features**



Price range is:

* positively correlated with "Average cost for two","Has Table Booking","Votes","Aggregate rating"
* negatively correlated with "Latitude","Locality"

**Correlation coefficient between Average Cost for two and other features**

****

Average cost for two:

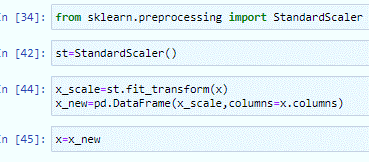
positively correlated with "Price rating","Has Table booking","Votes","Agrgregate rating"

* Negatively correlated with "Latitude","City".

1. Scaling the data:

Scaling the data refers to the process of standardizing the range of numerical data to ensure all features contribute equally to the analysis. It involves transforming variables so that they fall within a specific range, typically between 0 and 1 or centered around zero with a standard deviation of 1. Scaling is important in machine learning to prevent features with larger numerical ranges from dominating those with smaller ranges, ensuring fair comparison and more effective model training. Common scaling techniques include min-max scaling (normalization) and standardization (z-score normalization).





Firstly, we divided the data into the target variable (y) and the remaining features (X). Next, we applied the StandardScaler method to standardize the features in X.

Machine Learning Model :-

In our problem statement we are asked to build the model for two features 1)price range and 2)average cost for two . Let's begin by focusing on developing the model for predicting the price range.

# Model building to predict “Price range”

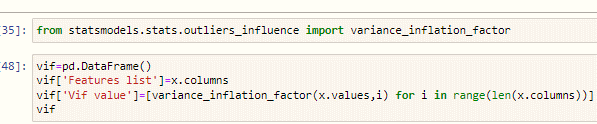
**Multicollinearity:-**

Multicollinearity happens when two or more predictors in a model are strongly correlated, which can confuse the model about their individual effects and make it less reliable. Detecting it involves checking how much predictors correlate. Fixing it can mean removing redundant predictors or combining correlated ones.

# Varience Inflation Factor

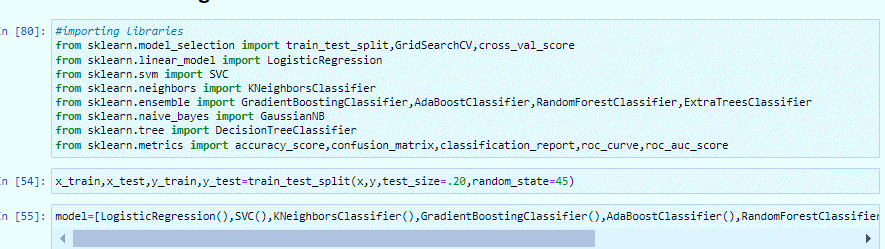
Varience\_inflation\_factor is used to check for multicollinearity





Here, there is no occurrence of multicollinearity

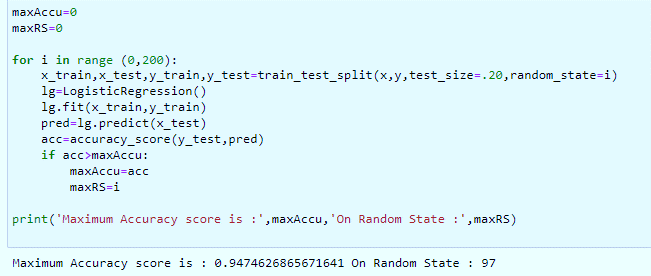
Among the features.

We will develop a supervised machine learning model using a classification algorithm. We'll begin by importing the necessary libraries, including those specific to our chosen classifier for classification tasks.

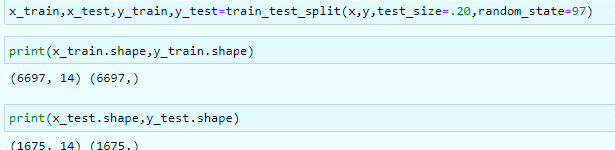
Here we imported the libraries required for model building, evaluation matrics addition to that I have created a pipeline which includes all different algorithms.

# Finding best Random State:

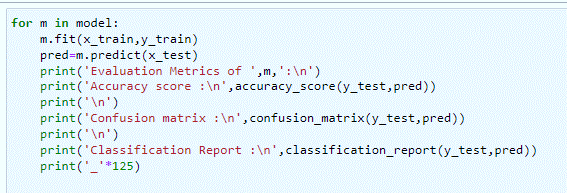
# Initially we find the best random state by building base model using LogisticRegression model using loop for range (0,200)

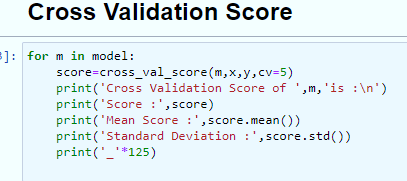
****

It gives accuracy score is .94 for random state 97



Later we will split x and y into train and Test For random state 97





.

**Here is the values of all algorithms**

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | Accuracy Score | Mean Cross Validation Score | Standard Deviation |
| LogisticRegression | 0.947 | 0.917 | 0.015 |
| SVC | 0.933 | 0.894 | 0.053 |
| GradientBoostingClassifier | 0.989 | 0.888 | 0.163 |
| KNeighborsClassifier | 0.784 | 0.696 | 0.083 |
| AdaBoostClassifier | 0.860 | 0.849 | 0.035 |
| RandomForestClassifier | 0.983 | 0.955 | 0.048 |
| ExtraTreesClassifier | 0.950 | 0.877 | 0.122 |
| GaussianNB | 0.660 | 0.594 | 0.059 |
| DecisionTreeClassifier | 0.987 | 0.883 | 0.160 |

We are going with RandomForestClassifier is giving good accuracy score 0.983 and mean cross validation score 0.955

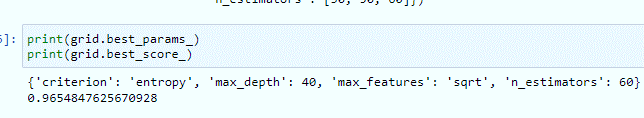
# 

# Hyper Parameter Tuning :-

# Hyperparameter tuning is the process of finding the best set of hyperparameters for a machine learning algorithm to optimize its performance. Hyperparameters, such as the learning rate, number of trees in a random forest, or number of hidden layers in a neural network, are set before training and significantly impact the model's results. Techniques like grid search, random search, and Bayesian optimization are commonly used to identify the optimal values, aiming to improve the model's accuracy and generalization to new data.

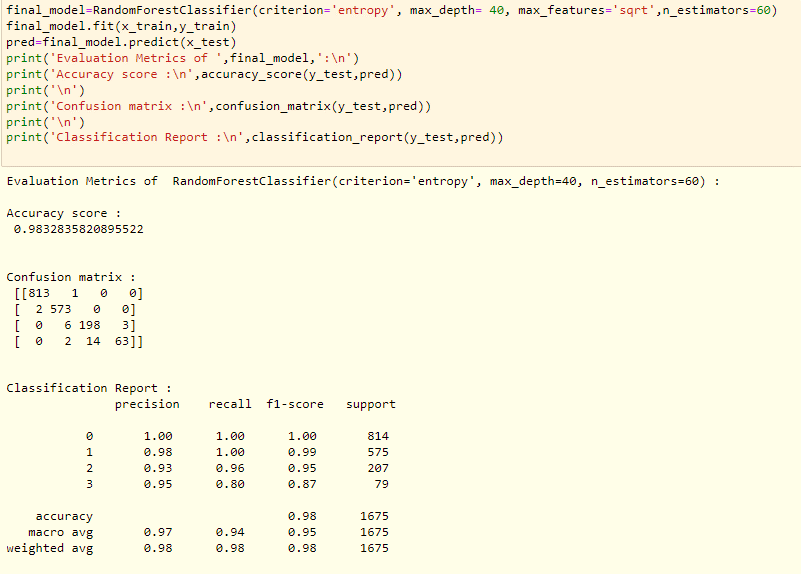
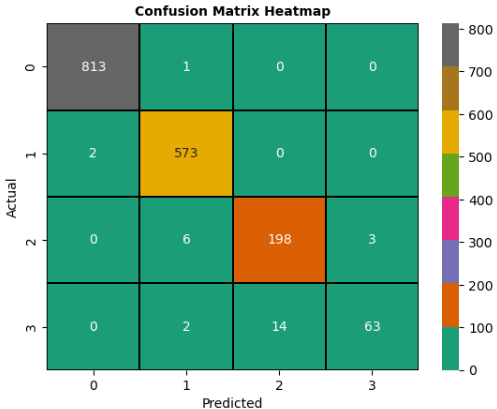
# 

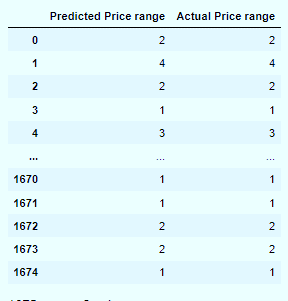
Here Gridsearchcv is used for hyper parameter tuning.



* Using these parameters we build our final model.

**Evaluation Metrics and confusion matrix of final model Random Forest Classifier**



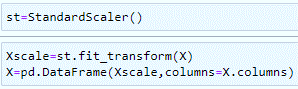


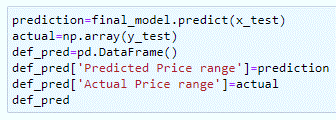
**Final prediction using built model**:

Here is the final prediction generated by the latest model using the available data. This evaluation was conducted to assess the model's performance, and the results indicate that the model is functioning effectively.

# Model building to predict "Average Cost for two"

Initially we split and scale the data:



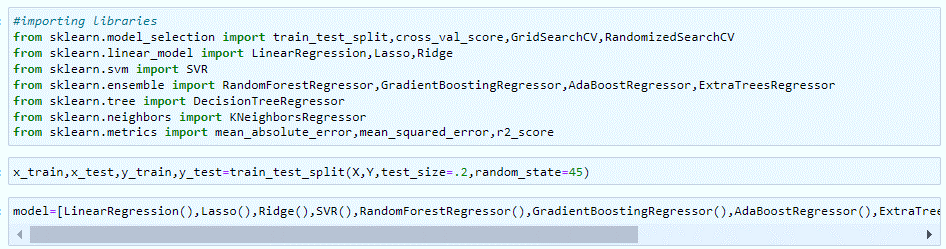
****

Further we will proceed with checking for multicollinearity :-



There is no presence of multicollinearity among the Features.

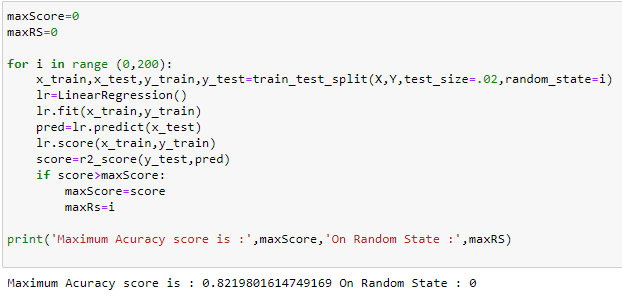
We will develop a supervised machine learning model using a Regression algorithm. We'll begin by importing the necessary libraries, including those specific to our chosen regressor for regression tasks.



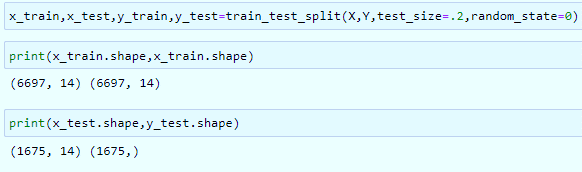
Here we imported the libraries required for model building, evaluation matrics addition to that I have created a pipeline which includes all different algorithms.

# Finding best Random State:

# Initially we find the best random state by building base model using LinearRegression model using loop for range (0,200)

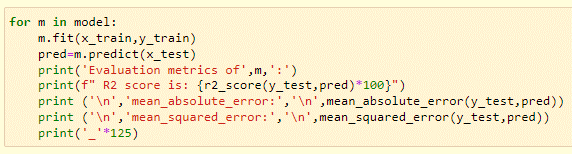
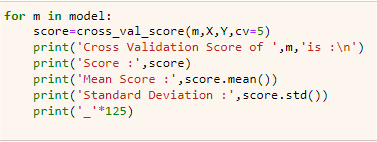


Here the best random state for Linear regression is 0 with accuracy score 0.82



We split the data into train and test for random state 0

The code evaluation metrics and cross validation score for various algorithms are computed using this framework



**Here is the values of all algorithms**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | **R2 Score** | **Mean Absolute Error** | **Mean Squared Error** | **Mean CV Score** | CV Score Std Dev |
| LinearRegression() | 71.45 | 6.98 | 94.71 | 0.32 | 0.55 |
| Lasso() | 68.51 | 7.20 | 104.45 | 0.39 | 0.44 |
| Ridge() | 71.45 | 6.98 | 94.71 | 0.32 | 0.55 |
| SVR() | 72.79 | 6.59 | 90.26 | 0.66 | 0.12 |
| RandomForestRegressor() | 85.67 | 5.08 | 47.53 | 0.78 | 0.09 |
| GradientBoostingRegressor() | 85.54 | 5.40 | 47.96 | 0.77 | 0.08 |
| AdaBoostRegressor() | 73.31 | 7.95 | 88.51 | 0.61 | 0.13 |
| ExtraTreesRegressor() | 85.27 | 5.21 | 48.86 | 0.74 | 0.14 |
| DecisionTreeRegressor() | 73.68 | 6.47 | 87.29 | 0.58 | 0.17 |
| KNeighborsRegressor() | 75.51 | 6.88 | 81.24 | 0.63 | 0.12 |

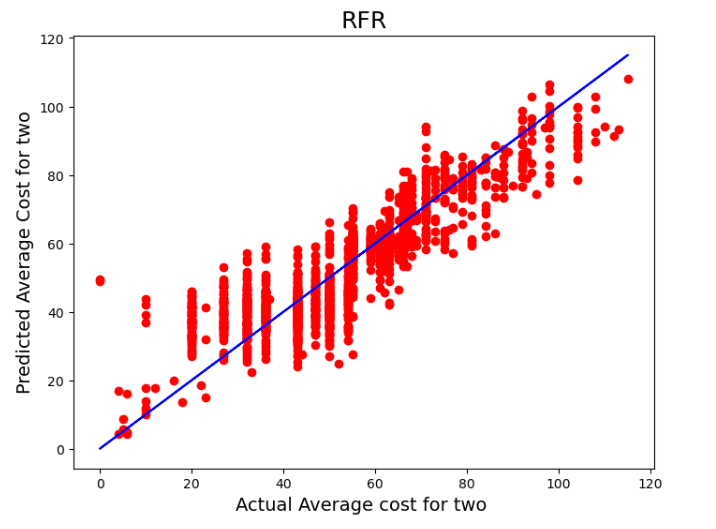
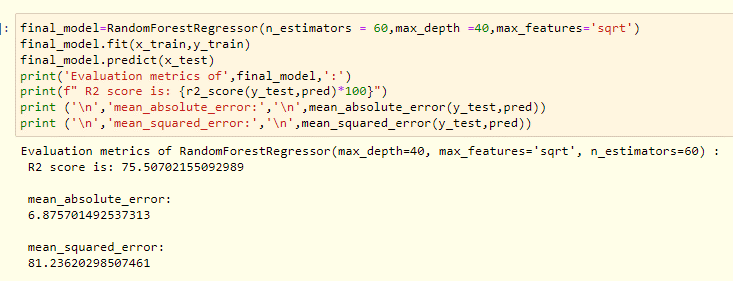
Here we proceed with RandomForestRegressor algorithm which is giving r2 score 85.67 and it’s cv score is .78

# Hyper Parameter Tuning :-

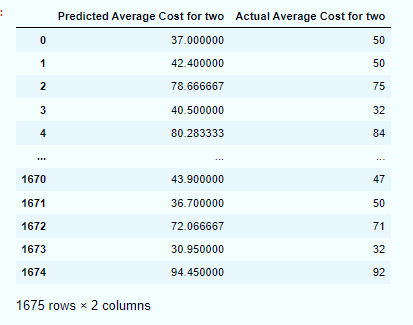
# 

* Using these parameter we will build our final predictive model

**Evaluation Metrics and Predicted vs. Actual: Visualizing Accuracy with RFR**



**Final prediction using built model**



* Here is the final prediction generated by the latest model using the available data. This evaluation was conducted to assess the model's performance.

Concluding Remarks:-

* North Indian and North Indian Chinese cuisines are popular across different price ranges.
* Fast foods like pizza tend to be more expensive.
* Localities such as Mayur Vihar Phase 1, Chandni Chowk, Najafgarh, Krishna Nagar, Vasundhara Enclave, Uttam Nagar, and Tilak Nagar prefer affordable dining options (Price Range 1).
* Shahdara has many affordable dining options with 75 restaurants in Price Range 1.
* Higher-priced cuisines like Lebanese, Moroccan, and Mediterranean focus on premium ingredients.
* Lebanese and Arabian cuisines can be enjoyed at lower costs, around $120 and $70, respectively.
* Price range and average price for two people are positively related.
* Our analysis included scaling, outlier removal, and model building steps.

If you're interested in exploring the coding aspect further, you can find more information on GitHub. [Click](https://github.com/Radhakulk/Evaluation-projects/blob/main/Evaluation%20project%20phase%203/Zomato%20Restaurant.ipynb)

**Thank You**