# **Title of the Project: Zomato Restaurant**

# **Problem Definition**

The main aim of this project is to analyze data of Zomato's restaurant for assisting those who want to find the value for money restaurants in various parts of the country for the cuisines. By leveraging the detailed restaurant data, the project aims to predict two critical factors:

1. **Average Cost for Two:** The average cost for two people dining at the restaurant.
2. **Price Range:** The price range classification of the restaurant.

# **Objectives:**

1. Firstly, we analyze the dataset to uncover various patterns related to restaurant ratings, costs, and cuisines.
2. Next to determine which restaurants provide the best value for money based on customer ratings and cost.
3. Then, we need to build ML models to predict the 'Average Cost for Two' and 'Price Range' of restaurants based on attributes such as location, cuisine, Has Table booking and Rating color.

**Need of the Project**

Customers can decide and make more choices about where to dine, while restaurant owners can better understand market trends and customer choices, where ML model further enhances these insights by enabling cost predictions, which is important for both budgeting and strategies of the business to grow properly.

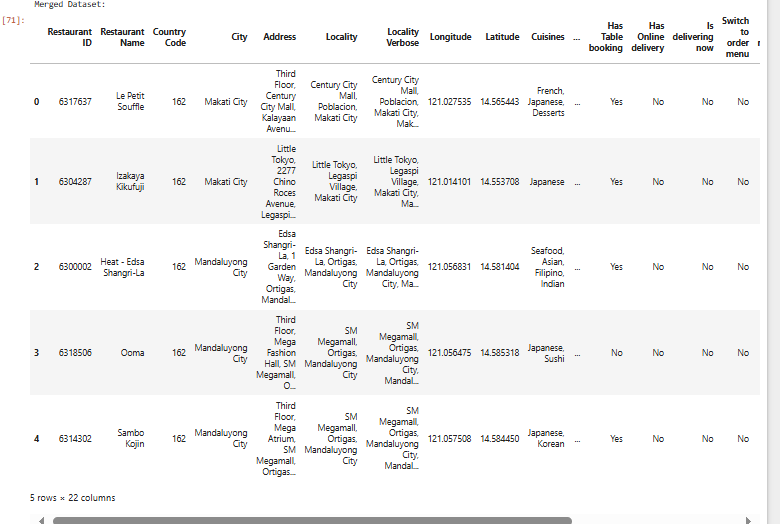
Here, we are using two datasets, 1st is Zomato.csv and other one is country\_code.csv, where,

* country\_code.csv contains all the Country code along with Country Name variable. While, Zomato.csv contains complete information about different restaurants, with each restaurant uniquely identified by its Restaurant Id. The dataset includes the below features:
  + **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's location.
  + **Latitude:** Latitude coordinate of the restaurant'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:** Indicates whether the restaurant accepts table bookings (yes/no).
  + **Has Online Delivery:** Indicates whether the restaurant provides online delivery (yes/no).
  + **Is Delivering:** Indicates whether the restaurant is delivering at the moment (yes/no).
  + **Switch to Order Menu:** Indicates whether the restaurant has switched to an order menu (yes/no).
  + **Price Range:** Range of price of food.
  + **Aggregate Rating:** Average rating out of 5.
  + **Rating Color:** Color representation of the average rating.
  + **Rating Text:** Text description based on the rating.
  + **Votes:** Number of ratings cast by people.

**The main goal is to utilize this data to predict the 'Average Cost for Two' and 'Price Range' of restaurants, aiding users to take the dining choices.**

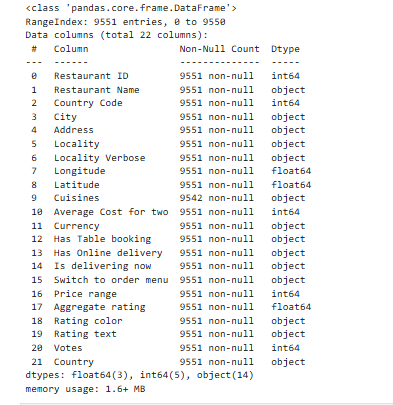
# **Data Analysis – Data Cleaning, Statistics, EDA, Pre-processing and Feature Engineering**

# **First 5 Rows**



The above output displays a merged dataset containing the information about various features such as Restaurant ID, Name, Country Code, City, Address, Locality, Longitude, Latitude, Cuisines, Average Cost for Two, and service availability like Table Booking and Online Delivery and displaying top 5 rows.

# **Info Details of the Dataset**



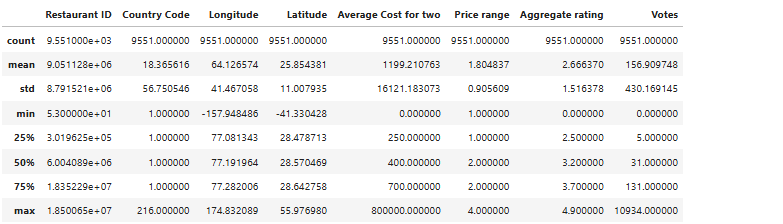
The above outcome shows the structure of the dataset, which contains 9,551 entries and 22 columns. Each column is listed with its name, non-null count, and data type. The features include Restaurant ID, Restaurant Name, Country Code, City, Cuisines, Average Cost for Two, and Price Range, etc.

# **Duplicated Value**

The dataset has 0 duplicate values.



# **Statistical Description**



The statistical summary of the dataset shows important metrices for several columns, including Restaurant ID, Country Code, Longitude, Latitude, Average Cost for Two, Price Range, Aggregate Rating, and Votes which will provide the values for counts, means, standard deviations, and percentiles providing an overview of the central tendencies and dispersion of these variables. Here, the Average Cost for Two has a high standard deviation, indicating variability in dining costs.

# **Missing Values**

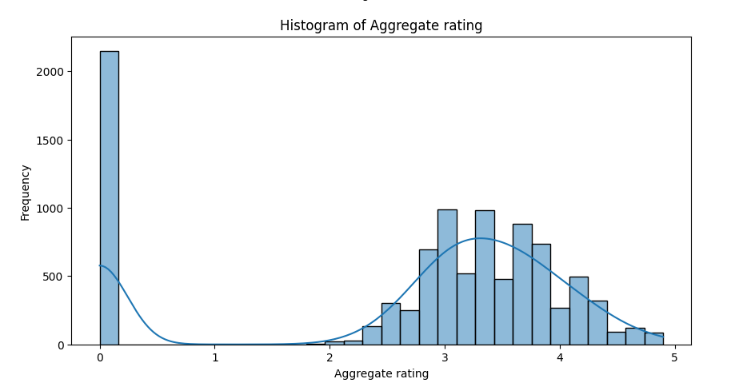


The above outcome shows the count of missing values for each column in the dataset. Most columns have zero missing values, except for the Cuisines column, which has 9 missing values which need to be addressed during preprocessing.

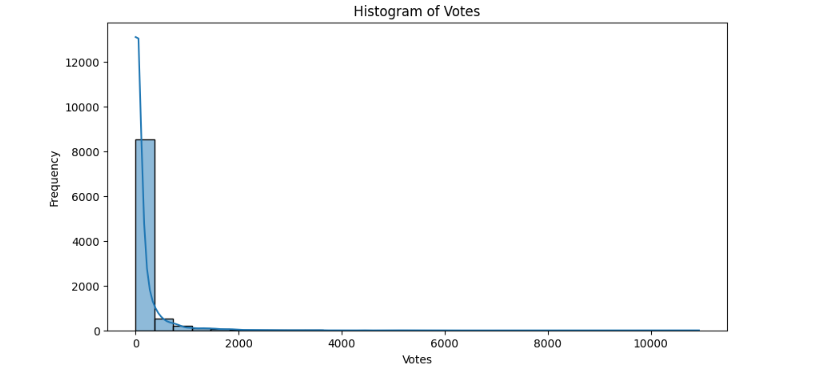
# **Univariate Analysis**

# 

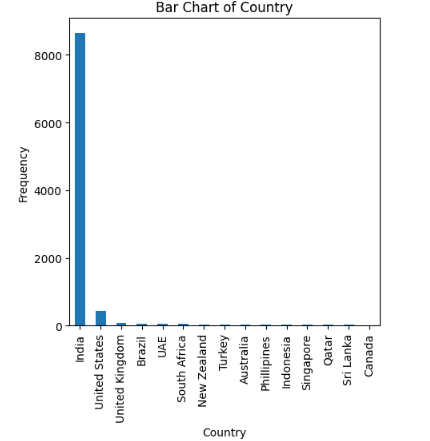
As we are performing Univariate analysis, in the above histogram of "Average Cost for Two" shows that the majority of restaurants have a low average cost for two people, where the distribution is highly right-skewed, indicating most data points are clustered at the lower end of the graph.



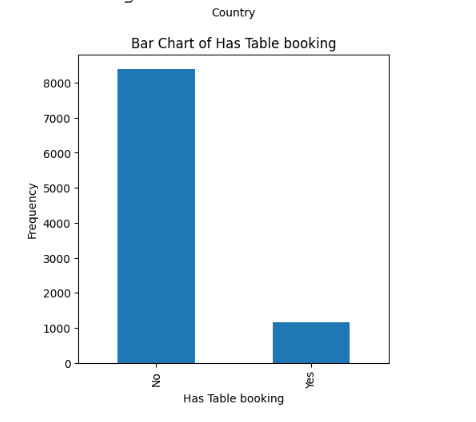
The above histogram of "Aggregate Rating" shows the distribution of ratings for the restaurants. There is a significant peak at 0 rating, which shows that many restaurants have no ratings. Most ratings are spreading between 2.5 and 4.0 which suggests the mostly restaurants gets moderate to good ratings.



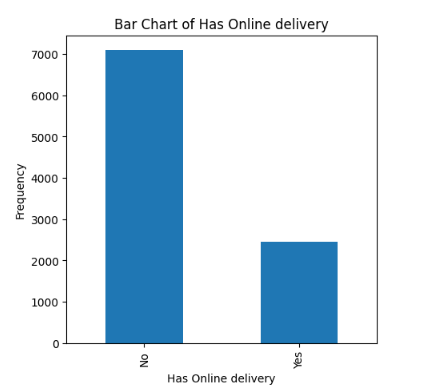
The above histogram of "Votes" variables indicates that most restaurants have received a low number i.e. 0 of votes, and the steep drop-off as the number of votes increases, which shows highly right-skewed distribution.



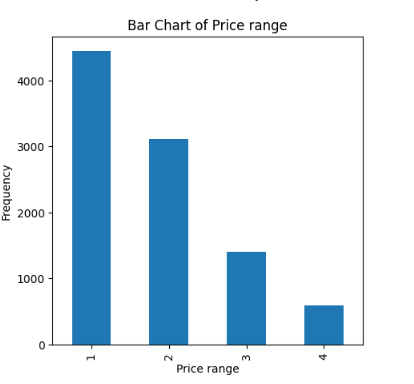
The above bar chart of "Country" variable shows the distribution of restaurant entries by country, where India has the highest number of entries i.e. approx. 9k, more than any other country. The United States follows with a much lesser count, and other countries like the United Kingdom, Brazil, South Africa and UAE have relatively low entries, indicating the dataset is heavily skewed towards Indian restaurants.



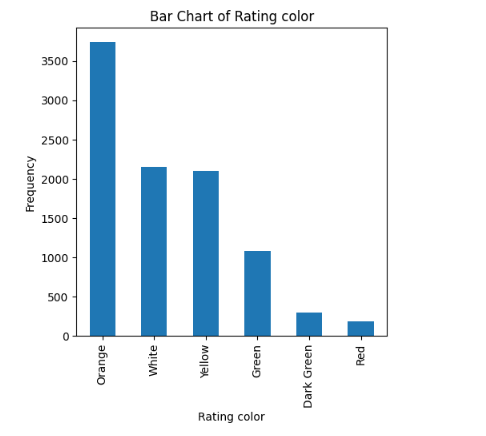
The above bar chart shows “Has Table booking" variable which shows that the majority of restaurants in the dataset do not offer table bookings. The frequency of restaurants without table booking is very high compared to the restaurants which have table booking, which is approx.9k.



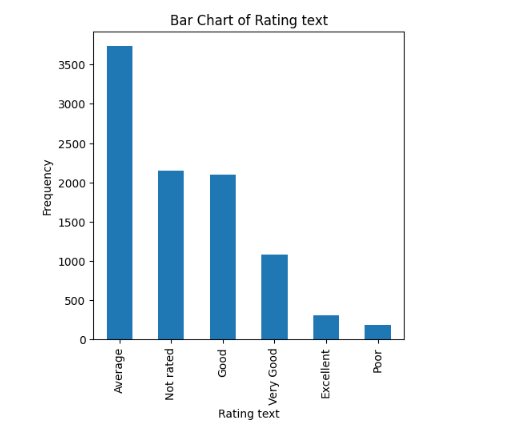
The above bar chart of "Has Online Delivery" variable shows that the maximum restaurants in the dataset do not offer online delivery services. The frequency of restaurants without online delivery is way much higher than those that do which is approx. 7k, indicating that online delivery is less common among the listed restaurants.

****

The above bar chart of "Price Range" shows the distribution of restaurants across different price ranges, where maximum restaurants fall into the lowest price range which is 1, followed by a decreasing number of restaurants in higher price ranges 2, 3, and 4, that indicates that most restaurants in the dataset are not so expensive.

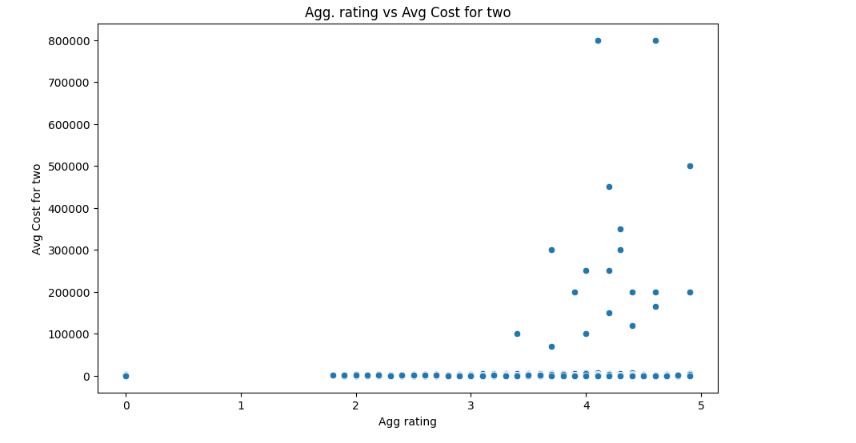
****

The given bar chart of "Rating Color" variable shows the distribution of restaurants by their rating colors and maximum restaurants have an "Orange" rating which is approx. 4k, followed by "White" and "Yellow” rating. Few restaurants have "Green," "Dark Green," or "Red" ratings, which means majority of restaurants receive mid-level ratings.

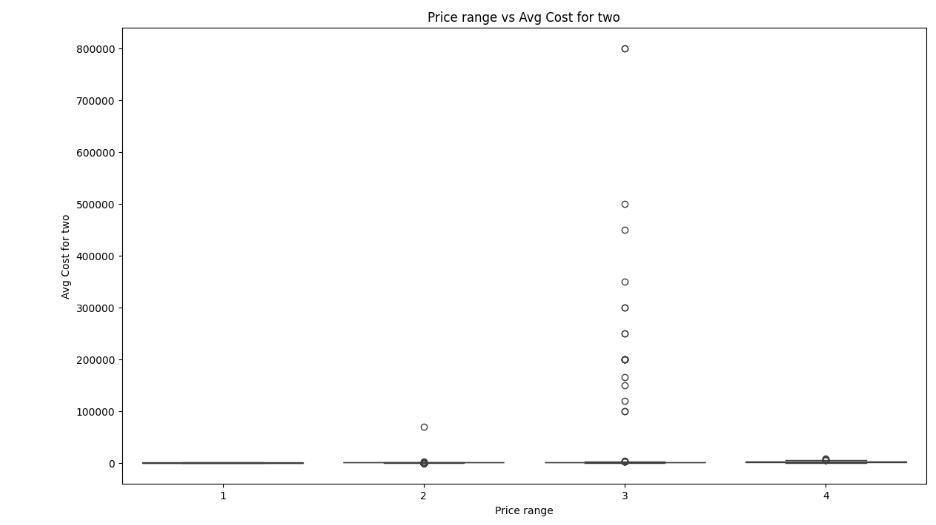
****

The above bar chart of "Rating Text" variable shows the frequency of different rating descriptions which is targeted to restaurants. The most common rating is "Average," i.e. 4k followed by "Not rated" and "Good." Which stats that most restaurants receive mid-level ratings, with fewer receiving extreme high ratings.

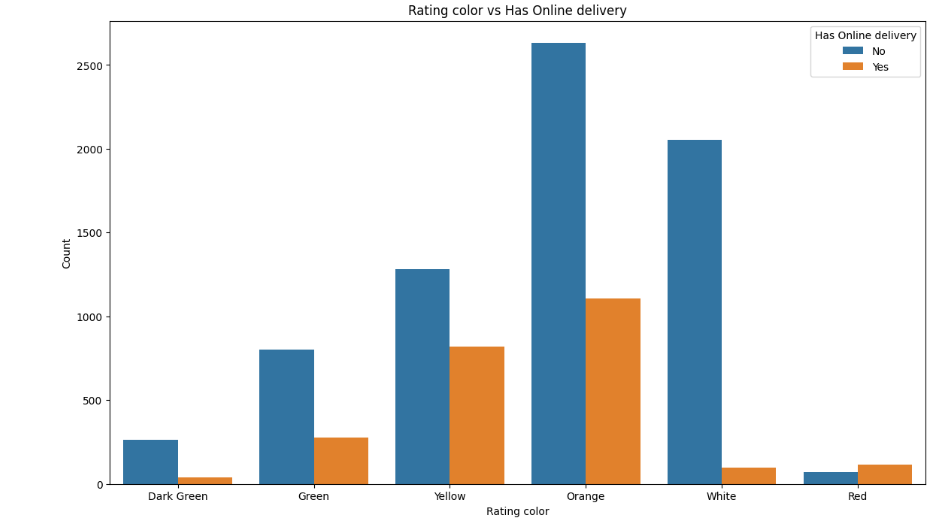
# **Bivariate Analysis**

****

The above scatter plot between "Aggregate Rating vs. Average Cost for Two" shows the relationship between restaurant ratings and their average cost for two people. There is a general trend indicating that higher-rated restaurants tend to have higher average costs. But, somewhere it is suggesting that cost alone does not determine ratings.

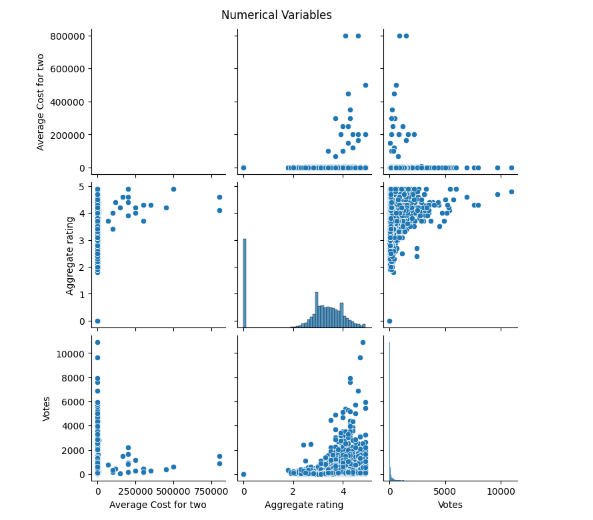
****

The above box plot is between the "Price Range vs. Average Cost for Two" variables which shows the distribution of average costs within each price range. It indicates that higher price ranges generally relates to higher average costs. But, there are few outliers with higher costs, in the lower price ranges, suggesting variability in pricing even within the same price range category.

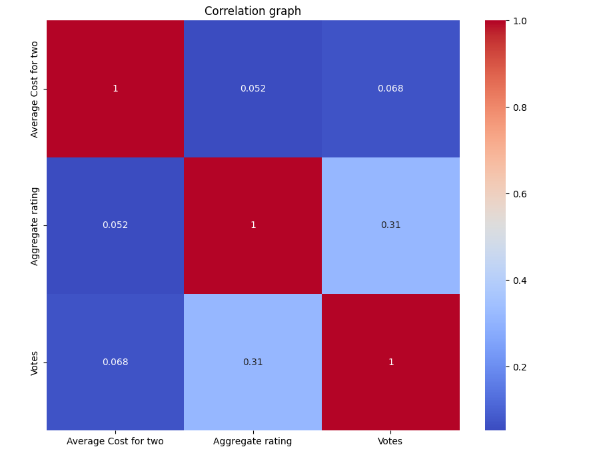
****

The above bar chart is between the "Rating Color vs. Has Online Delivery" variable shows the distribution of restaurant ratings with and without online delivery. Restaurants with online delivery which has range bars are present across all rating colors but are particularly common in the "Orange" and "Yellow" categories. Without online delivery restaurants which has blue bars also fall into these categories, showing that online delivery availability does not correlate with very high or very low ratings.

# **Multivariate Analysis**

****

The given pair plot visualizes relationships between Average Cost for Two, Aggregate Rating, and Votes, where we have observed the positive correlation between Votes and Aggregate Rating, indicating that higher-rated restaurants tend to receive more votes. Secondly, Average Cost for Two has a broad range with many outliers, especially at higher costs.

****

The correlation heatmap showing the relationships between Average Cost for Two, Aggregate Rating, and Votes, where we have observed the moderate positive correlation between Aggregate Rating and Votes which is 0.31, indicating that higher ratings are associated with more votes and also showing weak correlations between Average Cost for Two and Votes, and aggregate ratings, which shows that cost has little impact on ratings or votes.

# **Conclusion after EDA**

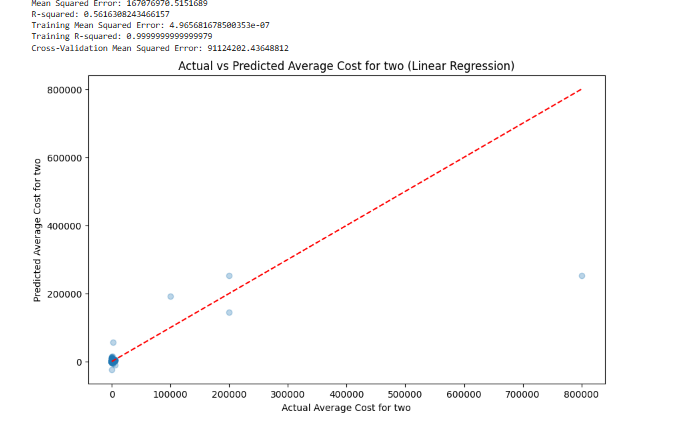
The EDA reveals that most restaurants have low costs and mid-level ratings, where we have positive correlations between ratings and votes. India dominates the dataset, with most restaurants not offering table bookings or online delivery. Price range and cost show expected trends, with variability in higher costs.

# **Data Preprocessing**

**During the Data Pre-processing,** we encoded categorical variables using Label Encoding and One-Hot Encoding, filled missing values, and split the dataset into training and testing sets. The target variable is "Average Cost for Two," which we aim to predict using the other features.

# **ML Models for predicting the Actual Average Cost Feature**

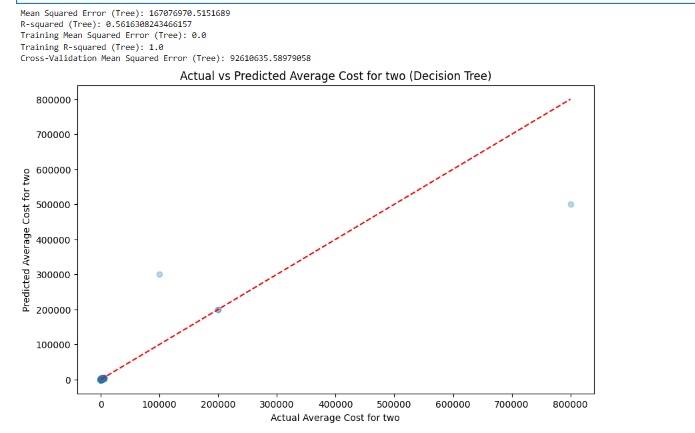
# **Linear Regression**

****

The above scatter plot shows the actual vs. predicted average cost for two people using Linear Regression model. The red line represents perfect predictions. The results indicate a moderate fit where R-squared is 0.56, with some predicted values significantly deviating from actual values, indicating the model's limitations in predicting higher costs accurately.

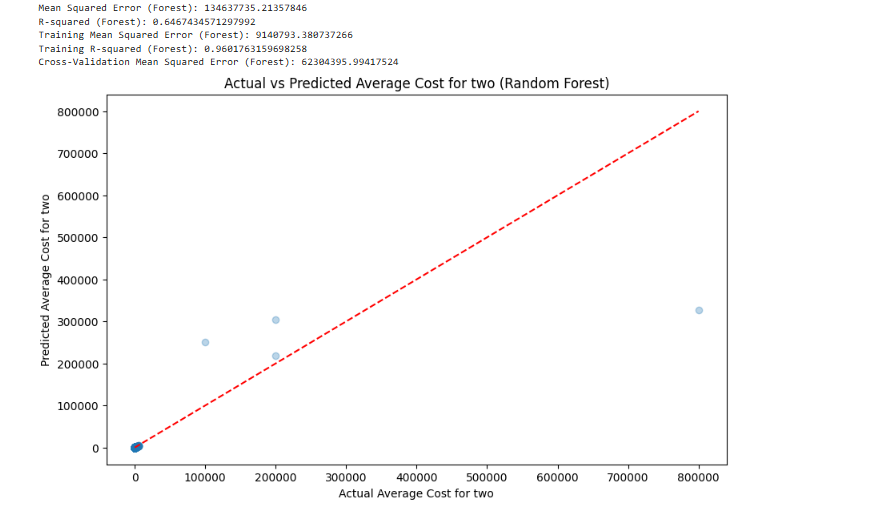
The MSE is approx. 167076970.5151689, with an R-squared value of 0.56. The training mean squared error is very low, but cross-validation error is 91124202.43648812 which suggests overfitting.

# **Decision Tree Regressor**

****

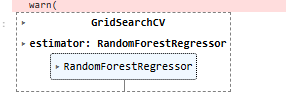
The above plot displays the actual vs predicted average cost for two using a **Decision Tree model**. The mean squared error is approx 167,076,970, with an R-squared value of 0.56. The training mean squared error is 0, indicating perfect fit on training data, but the cross-validation error showing 92,610,635 approx. that suggests overfitting.

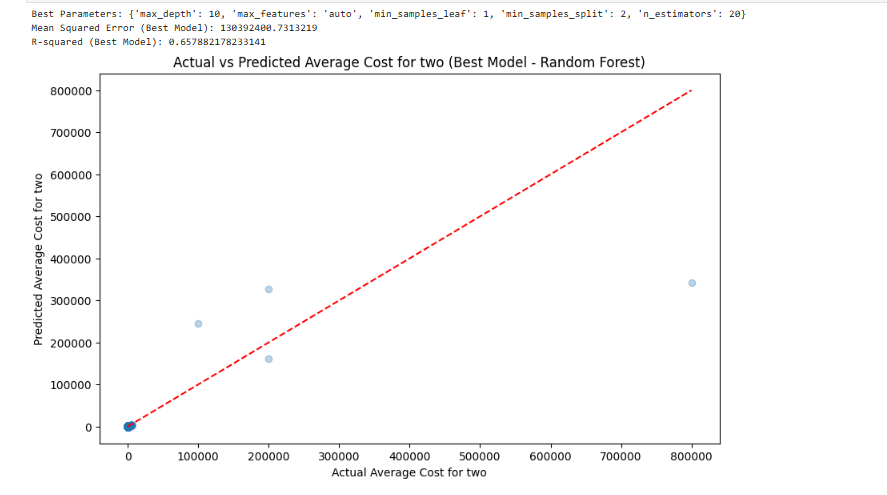
# **Random Forest Regressor**

****

The above scatter plot displays the actual vs predicted avg. cost for two using a **Random Forest model.** The MSE is approx. 134,637,735, with an R-squared value of 0.65. The training MSE is 9,140,793, showing a good fit on training data. The cross-validation error is 62,340,395 that shows the model performs better than others models .

# **RandomForest With GridSearchCV**

****

****

The above scatter plot shows the actual vs. predicted avg cost for two using the best Random Forest model with optimized parameters: max\_depth: 10, max\_features: auto, min\_samples\_leaf: 1, min\_samples\_split: 2, and n\_estimators: 20. The mean squared error is approx.. 108,394,200, with an R-squared value of 0.66, indicating a good fit.

# **Final Conclusion Based on all the above ML Models**

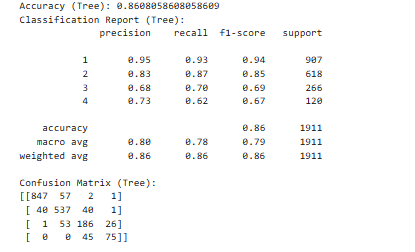
Based on the model evaluations, the Random Forest model with optimized parameters provided the best performance for predicting the average cost for two, achieving an R-squared value of 0.66. While this model performed better than Linear Regression and Decision Tree models. We can perform further feature engineering and gathering data to improve accuracy. Overall, the Random Forest model is recommended for its balance of bias and variance in predicting the target variable.

# **Future Business Recommendations**

* Gather detailed customer feedback to gain insights into what drives high ratings and satisfaction and also include data from rivals to understand market strength to identify areas for improvement.
* We can use feedback, suggestions and ratings to identify areas where service quality can be improved.
* We should improve the online presence using better website design, active social media engagement, and online delivery services.
* We should continuously monitor and analyze data to stay updated with changing customer preferences and market trends.

# **ML model prediction on Price Range**

# **Decision Tree**

****

****

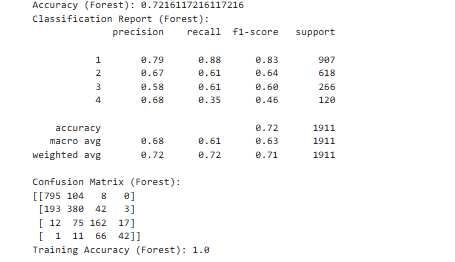
The Decision Tree model for predicting Price Range achieved an accuracy of 86%. The classification report indicates high precision, recall, F1 score which is 0.95,, 0.93, and 0.94 resp. for the lowest price range with decreasing performance for higher ranges. The confusion matrix shows most misclassifications occur between adjacent price ranges. The model giving perfect training accuracy, but cross-validation accuracy is 83.8% suggests overfitting.

# **Cross-validation on Decision Tree**

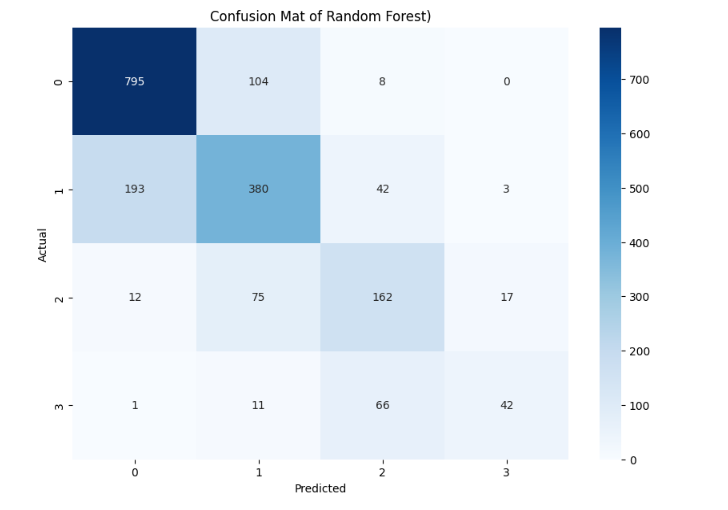
****

The cross-validation accuracy is 83.8%, suggesting that the model may be overfitting. To improve generalization, consider techniques like pruning, regularization, or using ensemble methods such as Random Forest.

# **Random Forest Model**

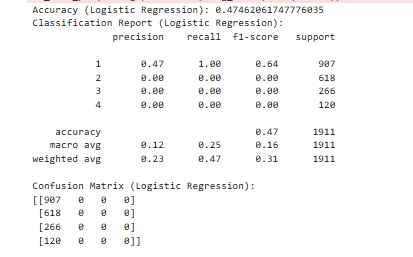
****

Next is Random Forest model for predicting Price Range achieved an accuracy of 72%. The classification report shows high precision and recall for the lowest price range, but performance decreases for higher ranges. The confusion matrix indicates misclassifications, particularly between adjacent price ranges. The model's training accuracy is 100 percent which suggests overfitting. To improve, consider tuning hyperparameters, etc.

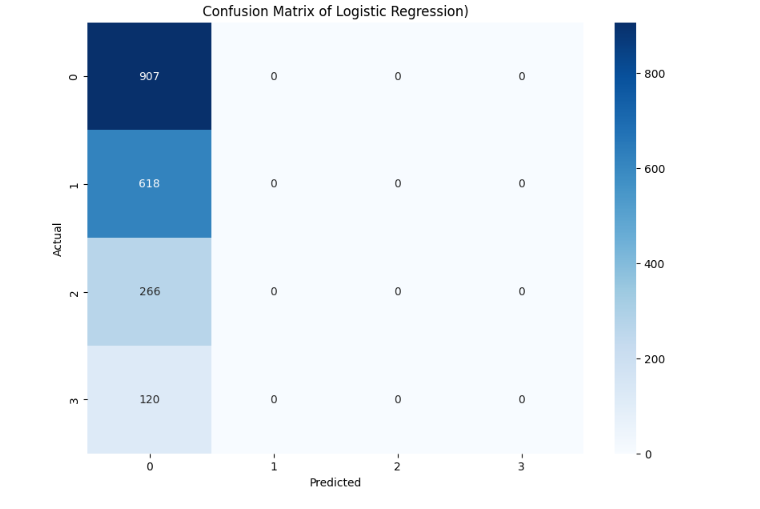
****

Here, the confusion matrix for the Random Forest model shows the distribution of actual vs. predicted price ranges. The model correctly predicts most of the lowest price range which is 0 but has continuously misclassifications between adjacent ranges. There are 104 false positives for range 1 predicted as range 0 and 193 false negatives for range 1 predicted as range 2, showing the chance for improvement in differentiating between closely related price ranges.

# **Logistic Regression**

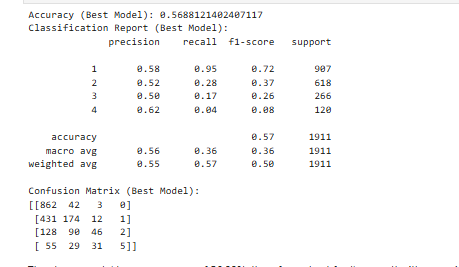
****

The Logistic Regression model is used here for predicting Price Range shows poor performance, with an accuracy of 47% only. The classification report showing that the model only predicts the lowest price range which is 1 correctly, with zero precision, recall, and f1-score for other ranges which is 2,3, and 4. The confusion matrix confirms that the model fails to distinguish between higher price ranges, as all predictions are classified as 0 which means that the Logistic Regression is not suitable for this multiclass classification task, and more complex models like Random Forest should be considered for better outcomes.

****

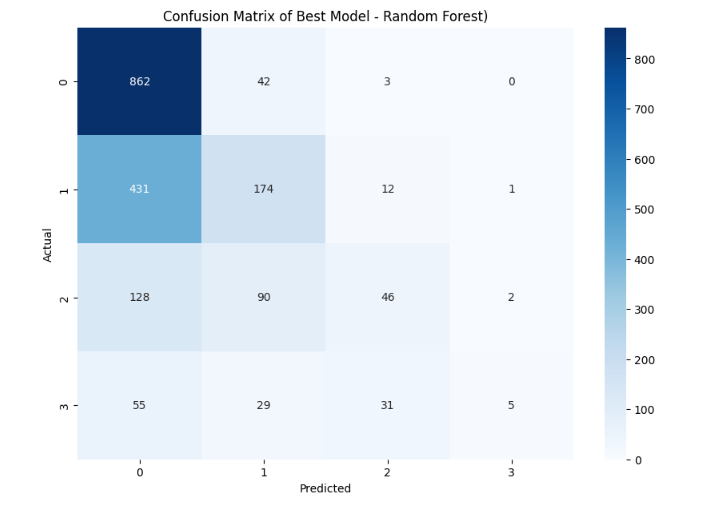
The confusion matrix for the Logistic Regression model showing its poor performance, as it only predicts the lowest price range correctly which is 0. Rest higher price ranges are not classified properly, which means the model fails to differentiate between various price ranges, that means Logistic Regression is not suitable for this task and indicating the need for more complex models.

# **GridSearchCV on Random Forest**

****

The Random Forest model with GridSearchCV achieved an accuracy of 57% comparatively better than logistic regression model.

The classification report shows moderate precision and recall for the lowest price range, which is1 but lower performance for higher ranges which are 2,3, and 4. The confusion matrix indicates continue misclassifications, particularly for adjacent ranges. This model, while better than Logistic Regression, still requires improvement, suggesting further tuning.

****

The confusion matrix for the best Random Forest model with GridSearchCV showing correct predictions for the lowest price range at 862. But it significant misclassifies for higher ranges, particularly 431 instances of price range 1 predicted as range 0.

# **Conclusion based on Various ML Model for Price Rane variable**

The various models applied for predicting the Price Range showing that while some models like Random Forest with GridSearchCV provided moderate accuracy, overall performance across models indicated challenges in distinguishing between price ranges, particularly higher ones.

The Decision Tree model achieved better classification metrics compared to Logistic Regression, which performed poorly.

# **Future Business Insights and Recommendations**

* We should proposed personalized deals and promotions to attract different customer segments and also regularly keep constant eye on customer feedback.
* We should optimize the menu, where we should introduce seasonal menus to keep offerings fresh.
* Enhance the social media engagement to promote special offers and new items and also do the marketing campaign.
* We should create a mobile app for easy reservations, online ordering, and keeping customer recommendations on priority.
* We should provide better training to the staff based on insights to improve the service efficiency and quality.
* We should work more on supply chain operations to reduce costs and ensure availability.