1. Problem Statement:

The focal point of our analysis revolves around predicting two critical aspects of dining experiences: the average cost for two individuals and the price range. With data analytics and machine learning, we aim to provide valuable insights for food dining, aiding them in making informed decisions regarding their dining choices.

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

**Country\_code.csv** contains two variables:

* Country code
* Country name

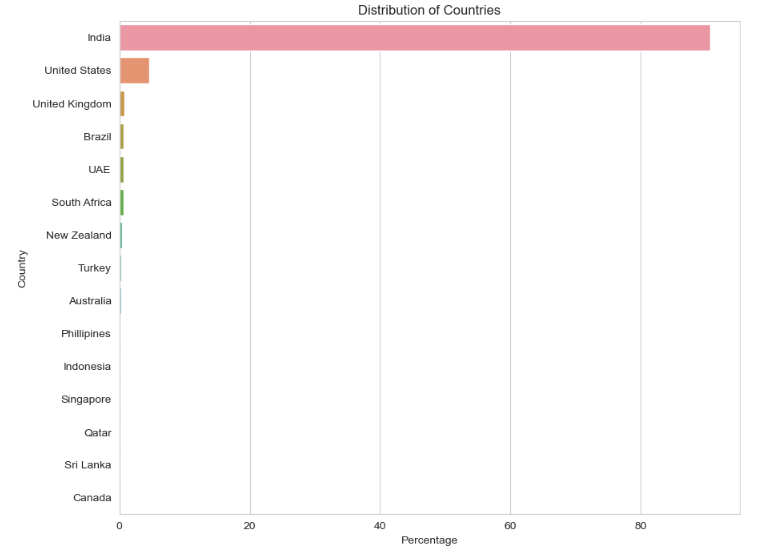
The collected data has been stored in the Comma Separated Value file **Zomato.csv**. Each

restaurant in the dataset is uniquely identified by its Restaurant Id. Every Restaurant contains the following variables:

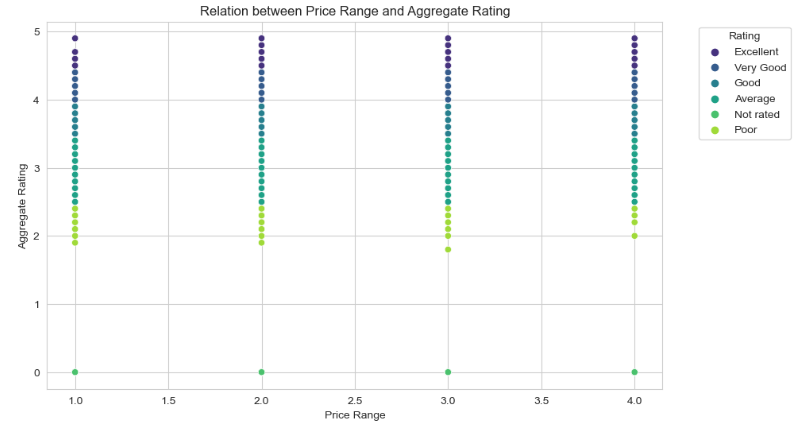
* Restaurant Id: Unique id of every restaurant across various cities of the world
* Restaurant Name: Name of the restaurant
* Country Code: Country in which restaurant is located
* City: City in which restaurant is located
* Address: Address of the restaurant
* Locality: Location in the city
* Locality 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

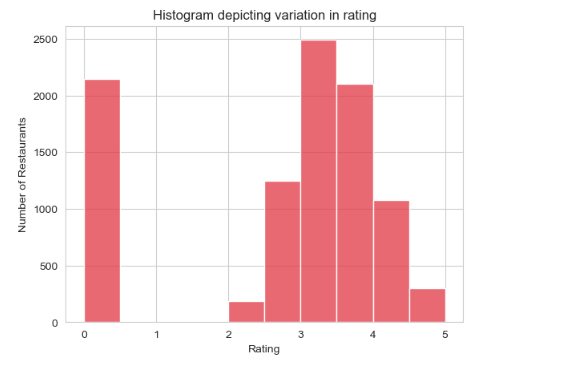
1. Data Analytics:

* There are 9551 Entries with 21 unique columns.
* Zomato is there is multiple countries, India among the top users followed by United Staes and United Kingdom

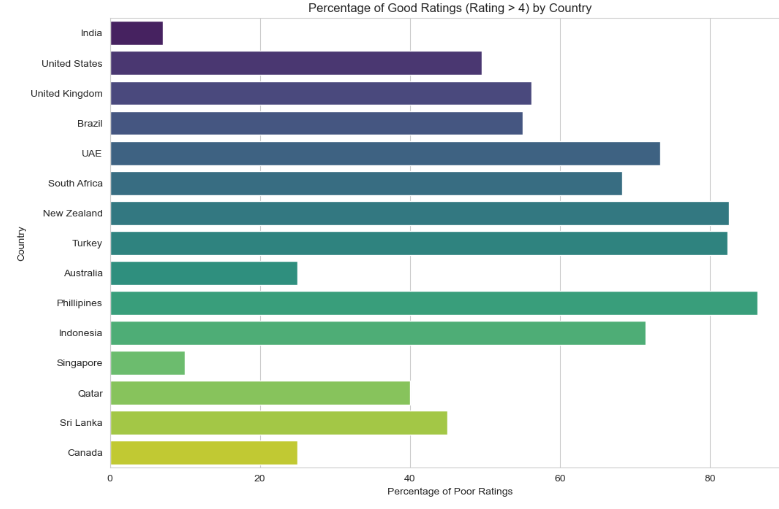


* Price Range, Aggregated Rating and Ratings are interlinked.

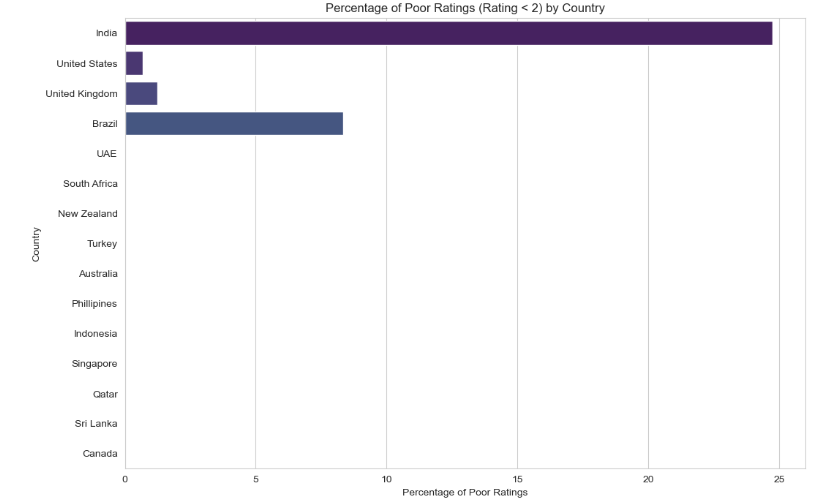


* There are either 0 ratings or above 2. With average rating is between 3 and 4.
* 
* Though India has most of the users from Zomato, but the % of rating > 4 is least among all.

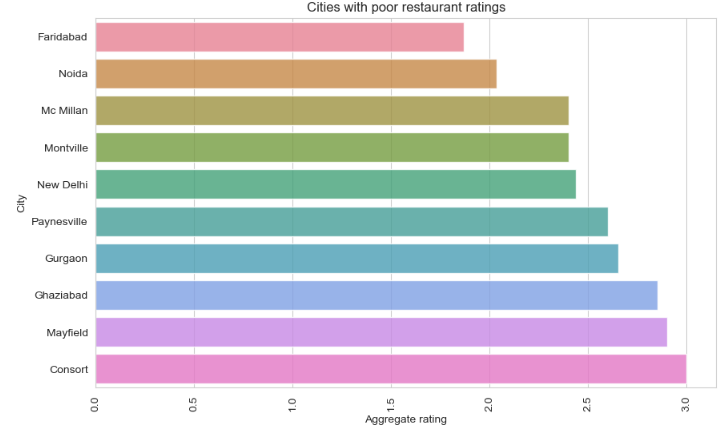
It showcase the quatity of Zomato is not appreciated despite its high usage. % of good rating is highest in Indonesia followed by Turkey and New Zealand



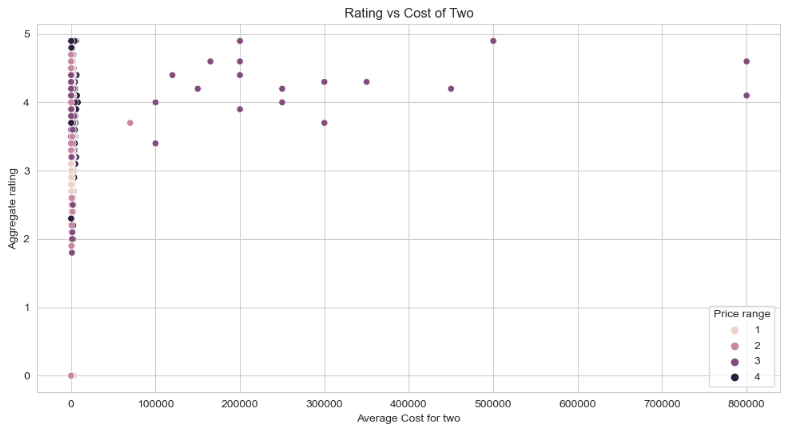
* Though India has most of the users from Zomato, in parallel it has the lowest ratings.



* Among the top 10 poor restaurant ratings, 5 cities are from India. Consort and Mayfield has the top two cities with poor restaurant ratings.



* Comparing the “Cost of two” and aggregated ratings – no linear co-relation is found.



We’ve drawn many inferences from the survey. Here’s a summary of a few of them:

* List of countries the survey is spread across —Philippines, Brazil, United States, Australia, Canada, Singapore, UAE, India, Indonesia, New Zealand, United Kingdom, Qatar, South Africa, Sri Lanka, Turkey.The total number to the country is 15
* Zomato is there is multiple countries, India among the top users followed by United Staes and United Kingdom
* Price Range, Aggregated Rating and Ratings are interlinked. The above information helps us to understand the relation between Aggregate rating, color, and text. We conclude the following color assigned to the ratings:
  + - \*Rating 0 — White — Not rated
    - \*Rating 1.8 to 2.4 — Red — Poor
    - \*Rating 2.5 to 3.4 — Orange — Average
    - \*Rating 3.5 to 3.9 — Yellow — Good
    - \*Rating 4.0 to 4.4 — Green — Very Good
    - \*Rating 4.5 to 4.9 — Dark Green — Excellent
* There are either 0 ratings or above 2. With average rating is between 3 and 4.
* Though India has most of the users from Zomato, but the % of rating > 4 is least among all.

It showcase the quatity of Zomato is not appreciated despite its high usage. % of good rating is highest in Indonesia followed by Turkey and New Zealand

* Among the top 10 poor restaurant ratings, 5 cities are from India. Consort and Mayfield has the top two cities with poor restaurant ratings.
* Though India has most of the users from Zomato, in parallel it has the lowest ratings.
* There is no relation between cost and rating. Some of the best-rated restaurants are low on cost and vice versa.
* Connaught Palace has maximum restaurants listed on Zomato but in terms of online delivery acceptance Defence colony and Malviya Nagar seems to be doing better.

1. EDA Concluding Remarks:

* No Duplicates were observed.
* Null values were there in “Cuisines” [Count - 9], they were replaced with mode of the columns.
* As the information was in 2 separate sheets, they were merged after the removal of NaN / Null values.
* As the dataset has both the strings and integers. Columns were separated as Numerical and Categorical columns for further EDA analysis.
* While doing the descriptive statistics analysis below were the observations:
  + Column count was same: No missing data.
  + There is no negative/valid \_ Dataset is OK [Except for Lat and Long]
  + Median (50%) > Mean - "Price range" --> Left Skewed
  + Median (50%) < Mean - "Votes”, “Average Cost for 2","Country Code”, “Rest ID" -> Right Skewed
  + Standard Dev for "Average Cost for 2" -> Outliers are present.
  + Percentile 75% and max: "Average Cost for 2", "Votes" - outliers are present.
* Above observations were verified using – sns plots, box plots (outliers) and skewness with python functions.
* Skewness was removed from numerical columns using the zscore. Total data lost was 1.8%
* While checking the non-unique values – Restaurant ID has the highest count, followed by Address, longitude and latitudes.
* Categorical Columns were changed using the Label Encoder.
* Very low co-relation observed is observed between the output and among all of the other inputs. (< 0.3)
* High multicollinearity is observed among Locality Verbose and Locality.
* Based on above inputs: 'Address', 'Currency', 'Restaurant ID', 'Switch to order menu' , 'Longitude' , 'Latitude' , 'Locality' columns were dropped.

Model1: Regression Model

1. Pre-processing Pipeline and Building Machine Learning Models:

## Standard Scalarization also known as z-score normalization, is used in machine learning to transform numerical features to have a mean of 0 and a standard deviation of 1. It is applied on the data set using the sklearn pre-processing library.

* As the output is continuous, after the standard scalarization we used Linear Regressor to find the best Random state.
* Data is split into test and train datasets with split of 30%.
* We tested the various models: Linear Regression, RandomForestRegressor, DecisionTreeRegressor, KNN, Lasso, Ridge, SVR, ExtraTreesRegressor.
* These models were first trained, and then they are used for calculating the output of the test data. This generated output is compared with actual output to calculate the Mean Absolute Error, Mean Squared Error, Root Mean squared error, R2 Score, R2 Score of the trained data.
* Model with least errors and highest R2 score is considered.
* In out case KNN was the best suited model.
* To improve the accuracy – Hyper Parameter tuning is done using the range of parameters from KNN model. It is done using the Grid Search CV.
* These optimized parameters are re applied to the KNN model with test and train data. Output is observed if the Errors are further reduced, and accuracy is improved.
* Finally, model is saved as \*.pkl file.

Model2: Classification Model : **Price range**

1. Pre-processing Pipeline and Building Machine Learning Models:

## Standard Scalarization also known as z-score normalization, is used in machine learning to transform numerical features to have a mean of 0 and a standard deviation of 1. It is applied on the data set using the sklearn pre-processing library.

* Oversampling: It is done on output stream to improve anti-aliasing performance, to increase resolution and to reduce noise.
* As the output is continuous, after the standard scalarization we used Random Forest Classifier to find the best Random state.
* Data is split into test and train datasets with split of 30%.
* We tested the various models: Random Forest Classifier ,Extra Trees Classifier, Logistic Regression Linear Regression, SVC, Gradient Boosting Classifier, Ada Boost Classifier , Bagging Classifier, Multinomial NB, Gaussian NB, KNeighbors Classifier and Decision Tree Classifier
* These models were first trained, and then they are used for calculating the output of the test data. This generated output is compared with actual output to calculate the accuracy score, confusion matrix , classification report.
* Model with least errors and highest accuracy is considered.
* In this case Random Forest Classifier is the best suited model.
  + Accuracy – 98%
  + Cross Validation Score – 95%
  + Precision – 99%
  + F1 Score -.99
* To improve the accuracy – Hyper Parameter tuning is done using the range of parameters from Random Forest Classifier model. It is done using the Grid Search CV (cv = 5). The parameter includes 'n\_estimators': [50, 100, 150], 'max\_depth': [None, 10, 20, 30], 'min\_samples\_split': [2, 5, 10], 'min\_samples\_leaf': [1, 2, 4], 'bootstrap': [True, False]
* Below are the optimized parameters:

'bootstrap': False, 'max\_depth': 30, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 100}

* These optimized parameters are re applied to the Random Forest Classifier model with test and train data. Output is observed if the Errors are further reduced, and accuracy is improved. Accuracy is improved to 98.87
* Finally, model is saved as \*.pkl file