**Restaurant Food Cost Prediction**



**1. Business Problem**

**1.1 Problem Description**

Who doesn’t love food? All of us must have craving for at least a few favorite food items, we may also have a few places where we like to get them, a restaurant which serves our favorite food the way we want it to be. But there is one factor that will make us reconsider having our favorite food from our favorite restaurant, the **cost**.

Here in this problem, we will be predicting the cost of the food served by the restaurants across different cities in India. We will use our Data Science skills to investigate the factors that really affect the cost.

*For complete code, please refer to this* [***GitHub repo***](https://github.com/vigneshrajancse/DataScience-Practice/blob/main/Evaluation%20Project%20Week%203/Restaurant%20Food%20Cost%20.ipynb)***.***

**1.2 Problem Statement**

We will try to predict food cost for particular restaurant and the dataset contains below features

**TITLE:** The feature of the restaurant which can help identify what and for whom it is suitable for.

**RESTAURANT\_ID:** A unique ID for each restaurant.

**CUISINES:** The variety of cuisines that the restaurant offers.

**TIME:** The open hours of the restaurant.

**CITY:** The city in which the restaurant is located.

**LOCALITY:** The locality of the restaurant.

**RATING:** The average rating of the restaurant by customers.

**VOTES:** The overall votes received by the restaurant.

**COST:** The average cost of a two-person meal.

**1.3 Machine Learning Formulation**

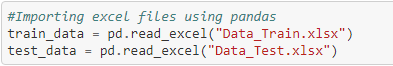
Here we supposed to predicted food cost, so it is basically Regression problem.

**2. Data Analysis**

In this problem, we will use [this](https://github.com/vigneshrajancse/Data-Science-Datasets/blob/main/Restaurant_food_cost.rar) dataset to predict food cost of a restaurant by using features like City, Rating, Votes etc.

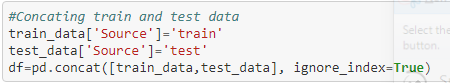
**2.1 Reading CSV file:**

Reading train and test CSV file using pandas, where train dataset have **12,690** rows and **9** columns and test dataset have **4,231** rows and **8** columns. Test CSV has 8 columns because we have to find cost variable.



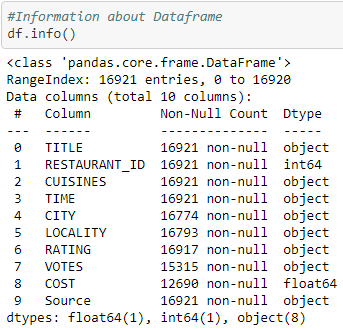
Concatenate train and test data using pandas concatenation method. We are adding one more column as “Source” to identify train and test data, later we will separate train and test data using this “Source” column.

Now we have **16921** rows and **10** columns in new data frame.

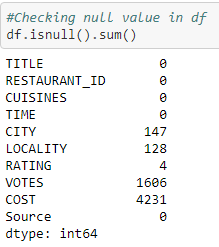


**2.2 Information about train & test dataset:**

In this new dataset there are 10 columns which has 2 numerical (int64 and float64) and 8 categorical (object) data type.

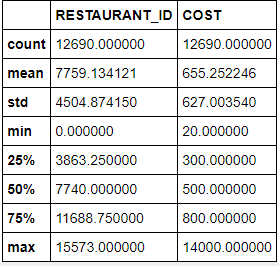


Dataset have missing values in City, Locality, Rating, and Votes. So, we have to fill the missing values using mean, median, or mode.



**2.3 Summary Statistics:**

We have only 2 columns because other than Restaurant ID and cost all are object data type variable. It shows count, mean, median, standard deviation (std), quartile data.



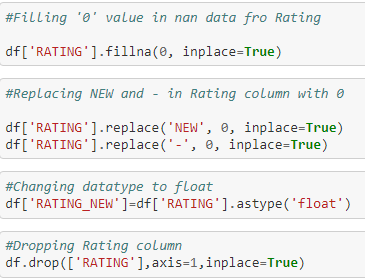
**2.4 Fill Missing Value**

As we already discussed, City, Locality, Rating, and Votes have missing values. So, we are going to fill these missing values.

**2.4.1 Rating:**

Some new restaurants do not have ratings, so we are filling 0 value in missing places. Then we have “New” and “—” in rating column which is inappropriate for our problem so, we are filling 0 value.

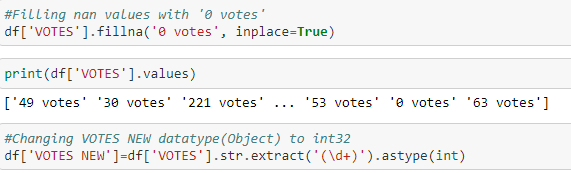
After that we are saving this rating in new column “RATING\_NEW” & converting data type to float and dropping “RATING” column.



**2.4.2 Votes:**

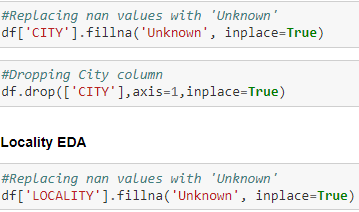
Some new restaurants do not have votes as well as rating, so we are filling 0 value in place of Nan. Then saving those VOTES data into new column VOTES\_NEW and converting data type to int.

And dropping original VOTES column.



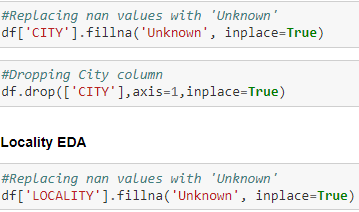
**2.4.3 City:**

There are some missing values in City Column, we cannot add city name using mode because there are 147 missing value. So, we are filling “Unknown” value.



**2.4.4 Locality:**

There are some missing values in Locality Column same like City Column, we cannot add Locality name using mode because there are 128 missing value. So, we are filling “Unknown” value.



We have filled all null values.

**2.5 Dummy Variables**

**2.5.1** **Title:**

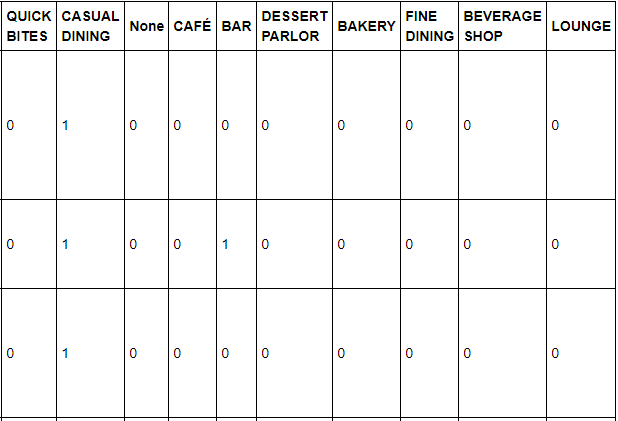
We are going to split frequent title values into a separate column, first we are splitting title using comma then saving in title variable.



After that we are using those titles and converting it to dummies so that we will have different columns with data. And dropping Title column.



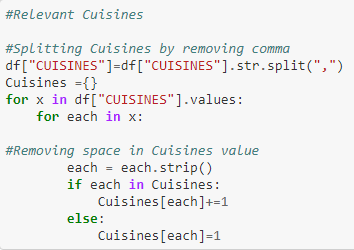
Below are the new title data columns:



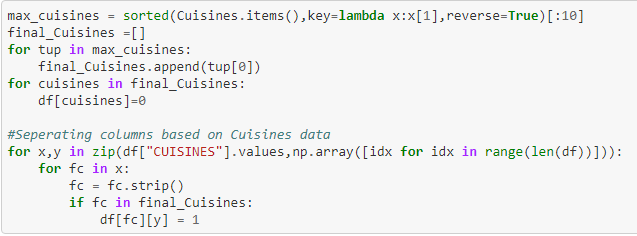
**2.5.2 Cuisines:**

Cuisines column also have frequent data, so we are going to follow the same step that we did for title column.

We are going to split the data in columns using comma and storing it in cuisines variable.

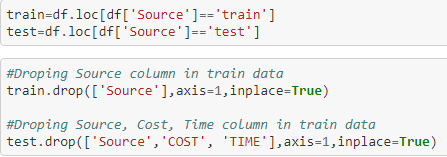


After that we are going to create columns using frequent cuisines data and finally dropping Cuisines column.



**3. Splitting Data into Train and Test dataset:**

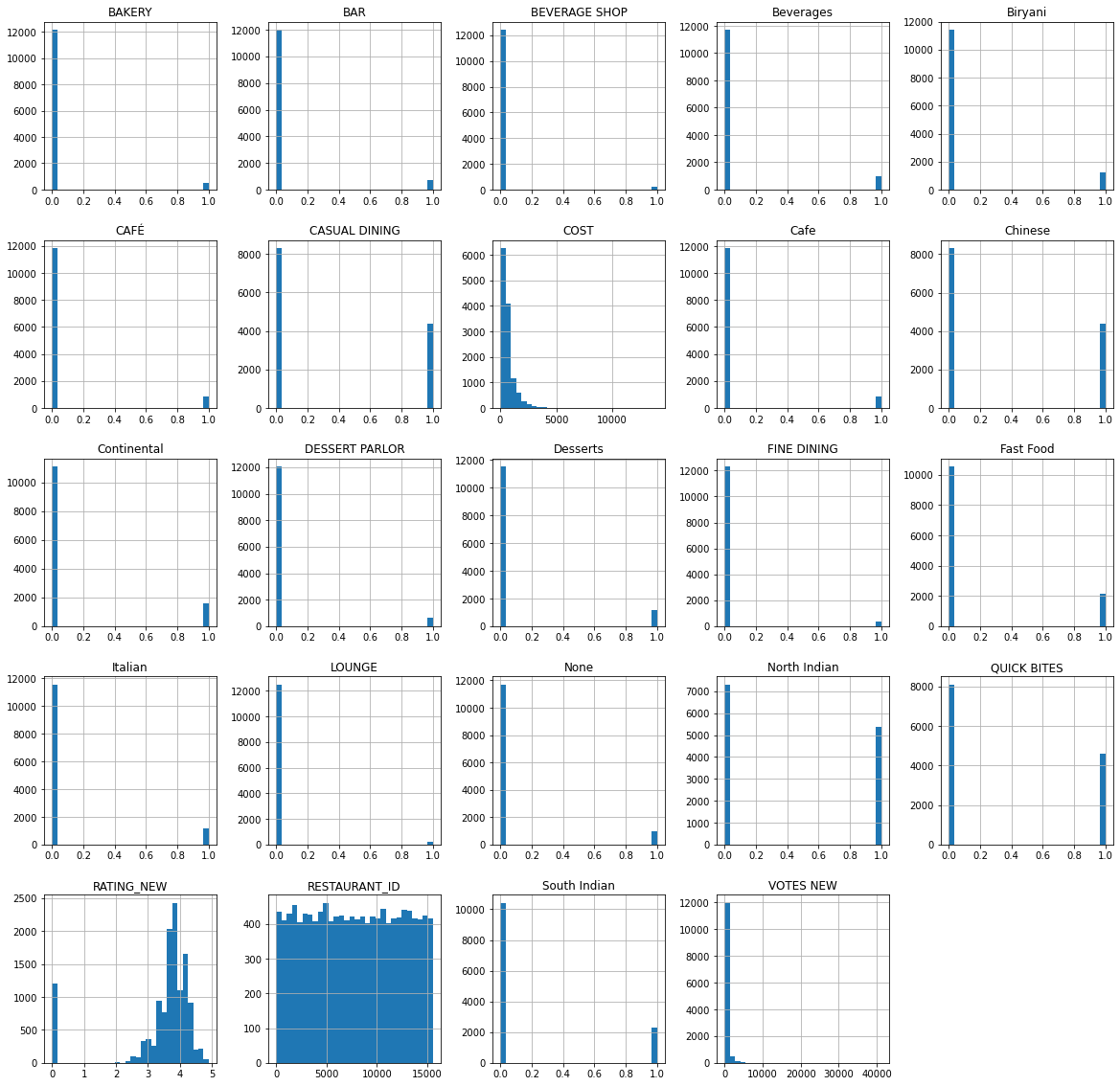
Splitting train and test dataset using Source column and dropping Source from train dataset and Time, cost from test dataset.

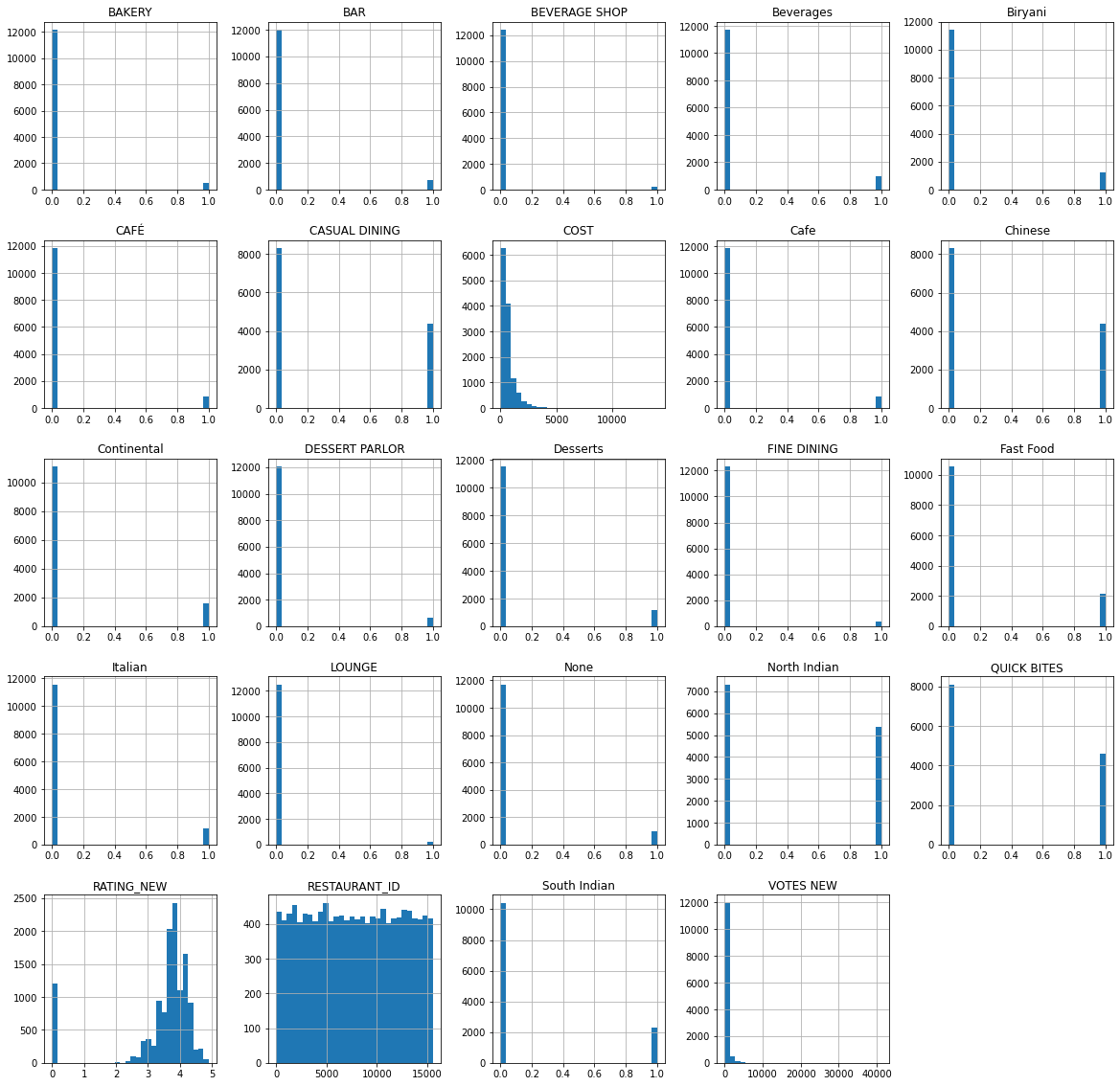


**4. Visualization**

**4.1** **Univariate Analysis**

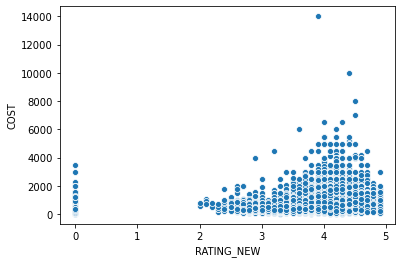
* Cost data is right skewed.
* Rating data is slightly left skewed and has some zero ratings.
* Restaurant ID is unique for all restaurant, so they are largely distributed.
* Other than Cost, Rating, Restaurant ID all are discrete values.





**4.2 Bivariate Analysis**

* We used scatterplot to visualize correlation between RATING\_NEW and COST.
* Scatterplot shows when rating increases, cost of the food also increases.
* When rating is zero, food cost is low.

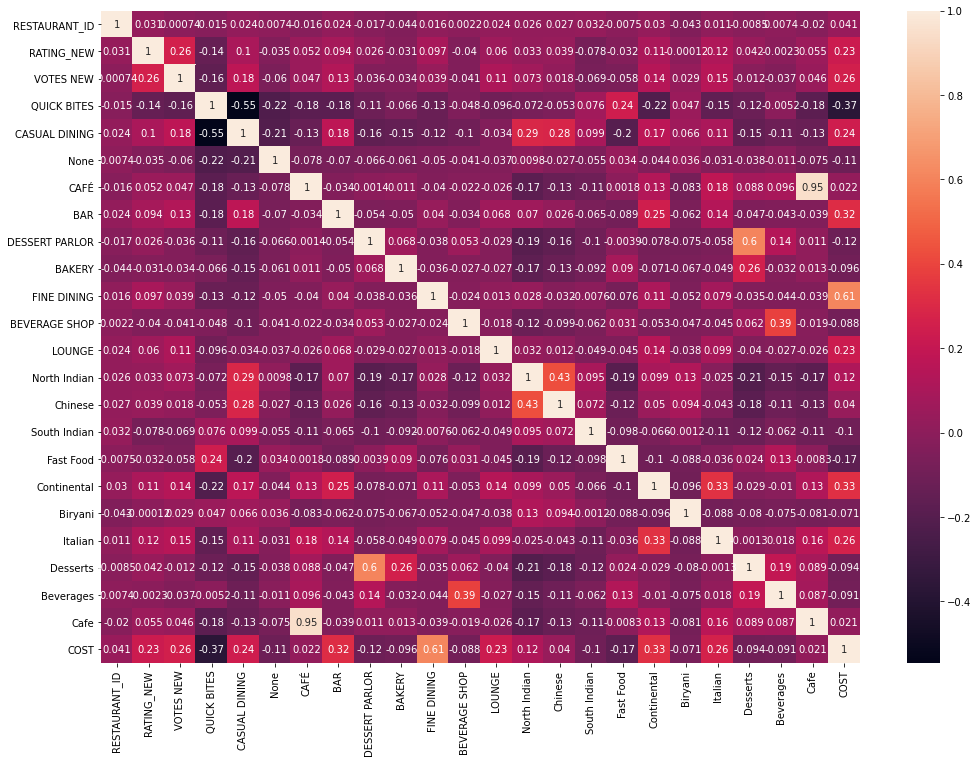


**5. Correlation:**

* Correlation explains how one or more variables are related to each other. These variables can be input data features which have been used to forecast our target variable.
* It’s a bivariate analysis measure which describes the association between different variables.

Mostly all the columns have good correlation with Cost data, except Quick Bites. Fine Dining has high positive correlation. Bar and Continental have good positive correlation.

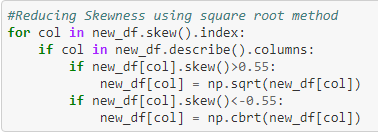
South Indian, Fast Food, Bakery, Dessert Parlor are slightly negative correlation.

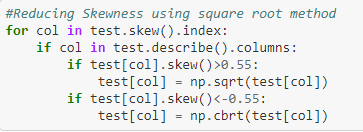


**6. Skewness:**

skewness is the measure of how much the probability distribution of a random variable deviates from the normal distribution.

We are reducing skewness in this dataset using Square root and Cube root method whenever a column reaches 0.55 and –0.55 we will use Square root and Cube root method to reduce skewness from data.





### **7. Saving File to Excel:**

After reducing skewness, we are saving train and test dataset to excel file.

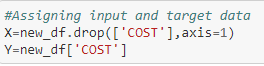


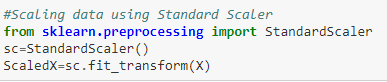
**8. Feature Scaling:**

We are using **StandardScaler** to performs the task of Standardization and it follows Standard Normal Distribution. It Standardize features by removing the mean and scaling to unit variance.

Standardization of a dataset is a common requirement for many machine learning estimators.

We are standardizing input data and storing it in a new variable *ScaledX.*





## **9. Building Machine Learning Model:**

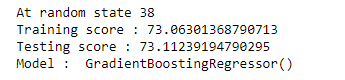
It’s time to build a Machine Learning model with our final dataset.

We are splitting data into train and test dataset because to predict the model we should have test dataset and train dataset is used to train the model.

Test size 0.25 means 25 percent of the data will be split to test dataset and remaining 75 percent data is for train dataset.

We have best random state which is 38.

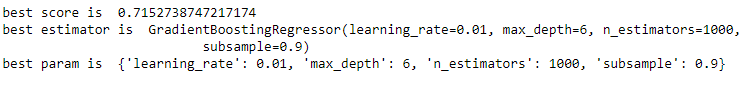
**Random state**: Random state is basically used for reproducing your problem the same every time it is run. If you do not use a random state in train\_test\_split, every time you make the split you might get a different set of train and test data points.



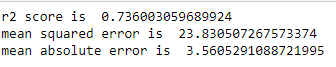
After splitting, we are checking which model gives good accuracy score. Gradient Boosting Regressor is the best model with accuracy score of 0.73 at random state 38.

**10. Hyperparameter Tuning:**

Now we are performing hyperparameter tuning using GridSearchCV for L Gradient Boosting Regressor model to get best parameter. And we got GradientBoostingRegressor(learning\_rate=0.01, max\_depth=6, n\_estimators=1000, subsample=0.9)as the best estimator.



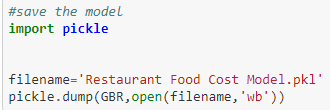
Then using best estimator, we are finding r2 score, cross validation score, mean squared error and mean absolute error.





**11.** **Saving Model:**

Finally, we are saving the model using pickle.



# **12. Conclusion**

* We did exploratory data analysis by removing columns, separated Title & Cuisines column with most common data, calculating missing recovered data, label encoding.
* Visualized data by univariate and bivariate analysis
* Checked correlation and visualized using heatmap and bar plot.
* Scaled the data using standard scaler method
* Gradient Boosting Regressor is the best model for Restaurant Food Cost Prediction.
* Performed Hyperparameter tuning for Gradient Boosting Regressor.
* Final score for Restaurant Food Cost prediction:
* Best r2 Score is 0.74
* Best CV Score is 0.71