**Zomato Data Analysis with Machine Learning**

**Overview:**

In this article, we will be predicting average cost and price range for the restaurant data available in Zomato. We will be using data provided, contains two datasets - Zomato.csv and country\_code.csv. “Average cost for two” prediction is through regression approach and “Price range”predictionis through classification approach.

**Problem statement**:

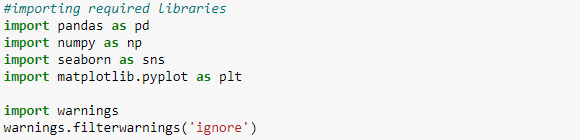
Zomato Data Analysis is one of the most useful analysis for foodies who want to taste the best cuisines of every part of the world which lies in their budget. This analysis is also for those who want to find the value for money restaurants in various parts of the country for the cuisines. Additionally, this analysis caters the needs of people who are striving to get the best cuisine of the country and which locality of that country serves that cuisines with maximum number of restaurants.

Following are the major steps in data science project life cycle:

1. Data analysis: Here we will get to know about how the present in provided data set
2. Exploratory data analysis: EDA is one of the most important steps in the data science project life cycle and here we make inferences from the visualizations and data analysis
3. Feature selection: Based on statistical support important features need to be selected
4. Model building: Here we will be using 5 ML models and then we will choose the best
5. Saving model: Saving the best model using pickle to make the prediction from real data.

There are many intermediate steps along with these above mentioned 5 steps, let us look at one-by-one along with the project flow.

1. **Importing libraries to start with**:



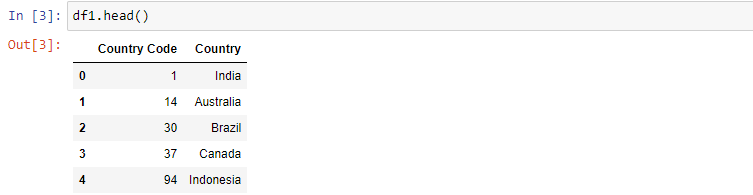
1. **Loading the dataset**:

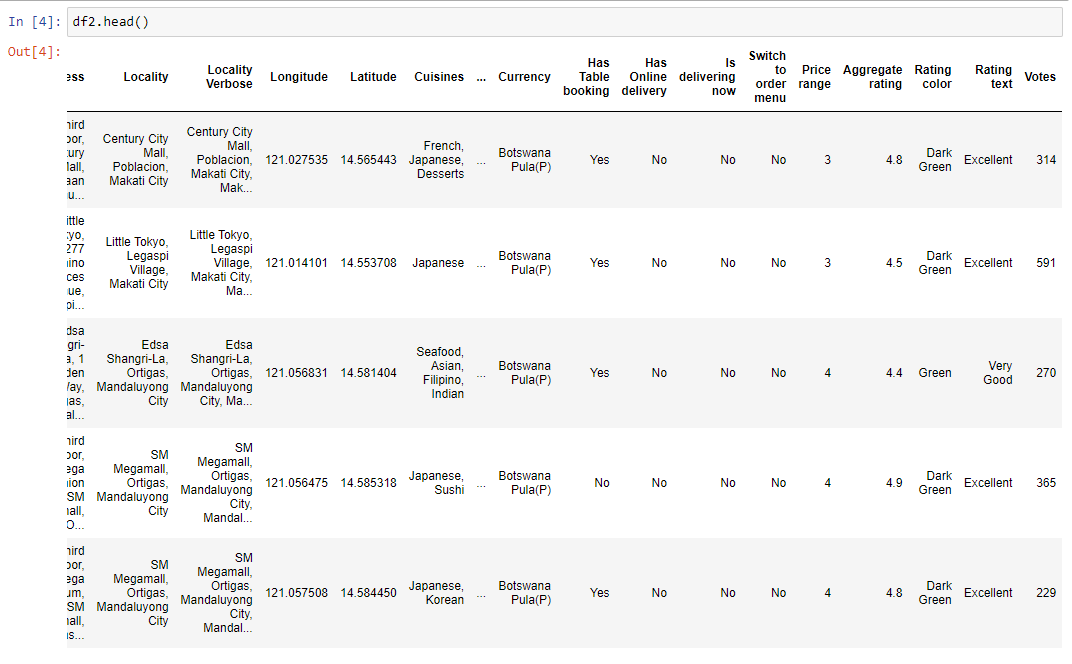


df1 – is of no use in prediction as it has only country code mapped to country name

df2 – is the main data set which we are going to work on

df = df2





1. **Exploratory data analysis**: 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'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: yes/no

• Has Online delivery: yes/ no

• Is delivering: yes/ no

• Switch to order menu: yes/no

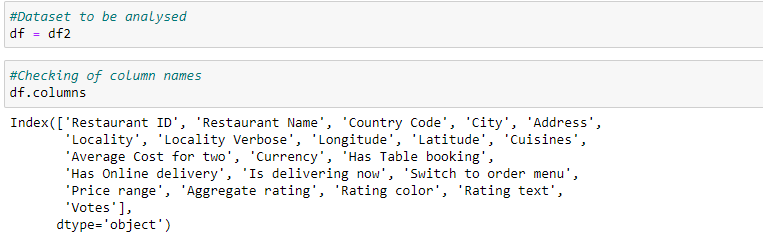
• Price range: range of price of food

• Aggregate Rating: Average rating out of 5

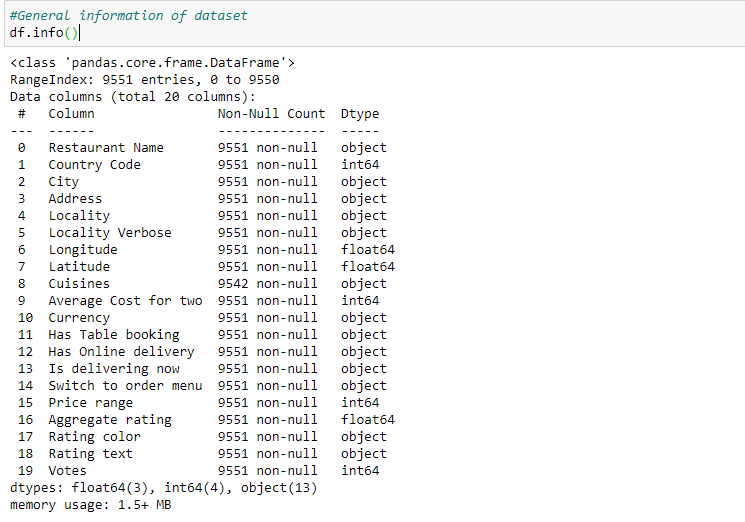
• Rating color: depending upon the average rating color

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

• Votes: Number of ratings casted by people

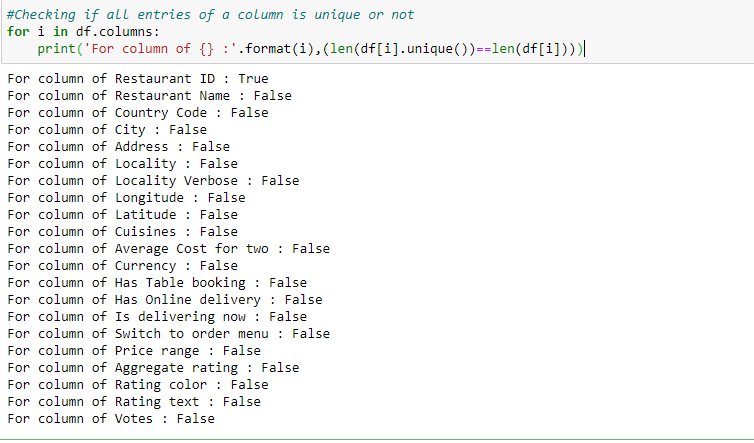


1. **General information of dataset**:



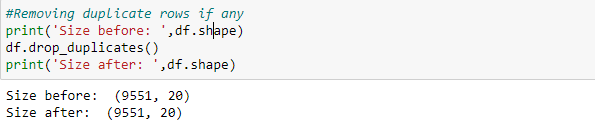
We can observe that cuisines has 9 null (NaN) values and is of obect datatype. We can handle this missing value by mode imputation.

1. **Checking if all entries of a column is unique or not:**



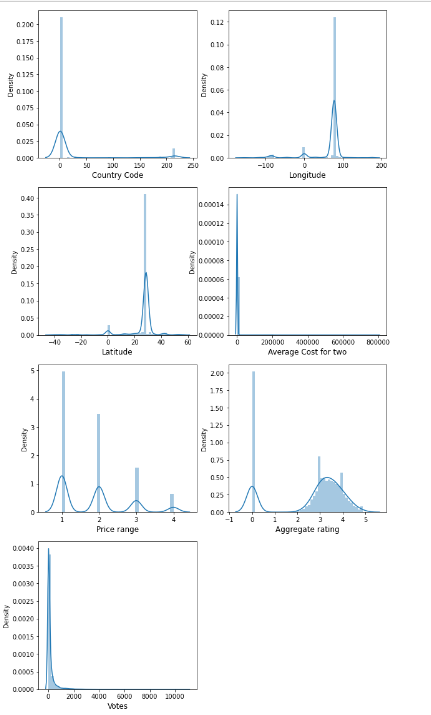
We can observe that Restaurant ID is having all unique entries and act just like indexing column. So let drop this column.

1. **Removing duplicate rows if any:**



Therefore no any duplicate rows present in the dataset.

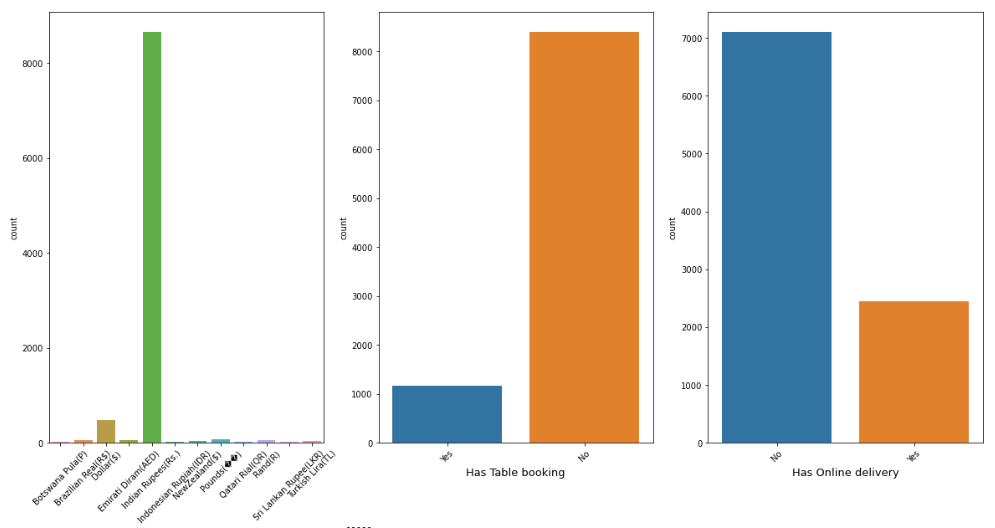
1. **Checking of distribution plot of each columns having numerical data:**

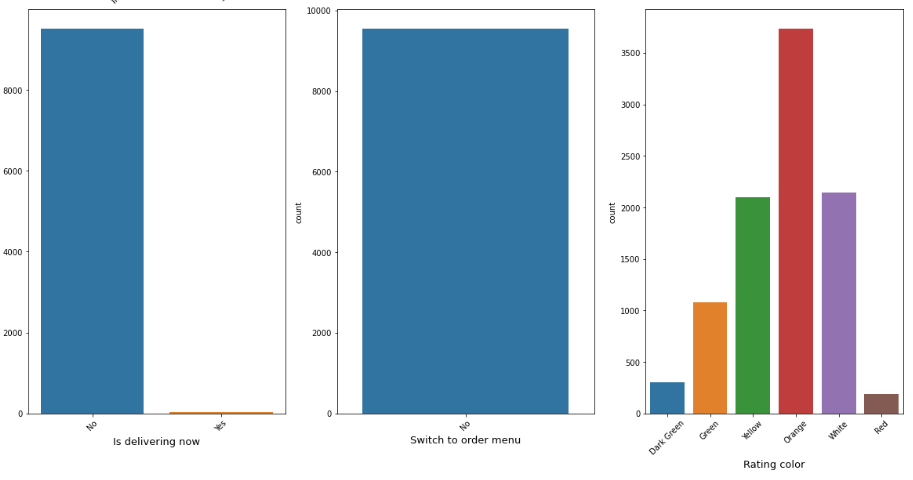


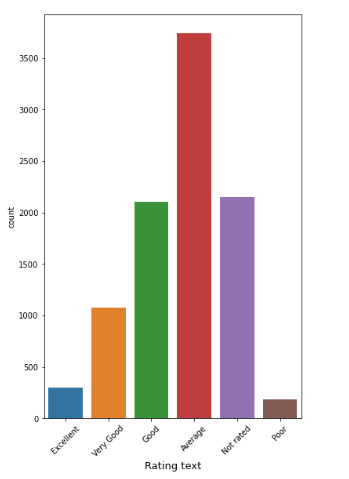
We can observe most of the country code is 1, which is India. Correspondingly Latitude and longitude locations points out Indian restaurants co-ordinates

Majorly average cost for 2 lies within range of 1500 INR correspondingly within price range of 1 & 2 most of restaurants are there. The data are highly skewed, needs outliers handling and proper transformation before training the model

1. **Checking of value counts in each columns having categorical data:**







We can observe that majority of currency is INR.

Majority of restuarants has no table booking and has online delivery

'Switch to order menu' is having all single entries as 'No', this wonts help in predictions

'Is delivering now' is having highly imbalanced class having very less proportion of 'Yes'

So let us drop these columns

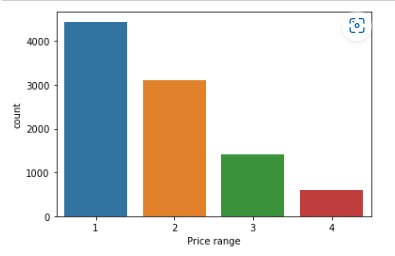
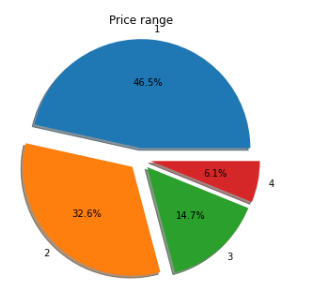
'Address' can be dropped as this carry almost all unique entries and location information can be obtained by co-ordinates also

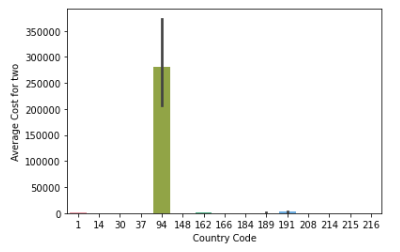
'Locality Verbose' convey almost similar information as that of 'Locality'

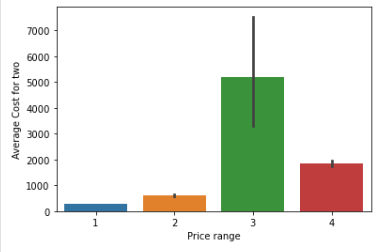
So let us drop these 2 columns

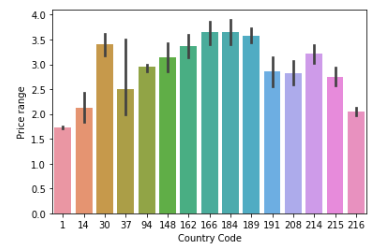
Rating color and text conveying same information, let us drop tone among the column.

1. **Checking of value counts in target variable:**



Let us plot target variables Price range vs Average cost and with different country code and check





We can observe that average cost is more in Price range 3 restuarants

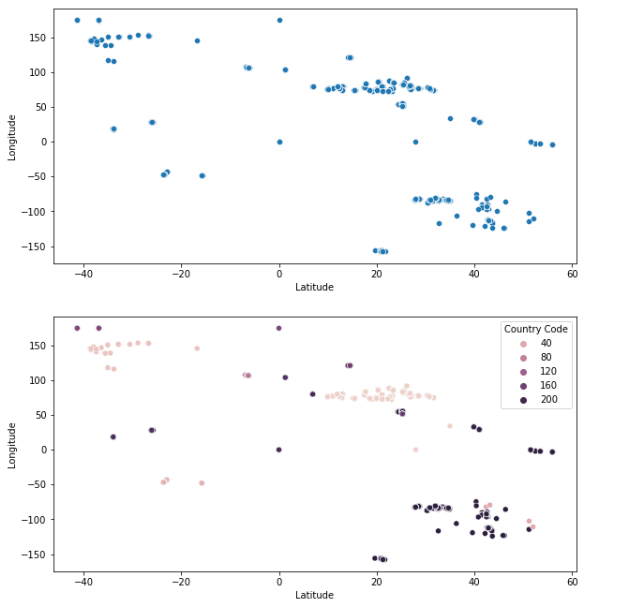
We can observe that average cost is more in country code 94 - Indonesia

From garph 3 we can observe that Indonesia falls under price range 3

Least price range near 1.8 for 1 - India

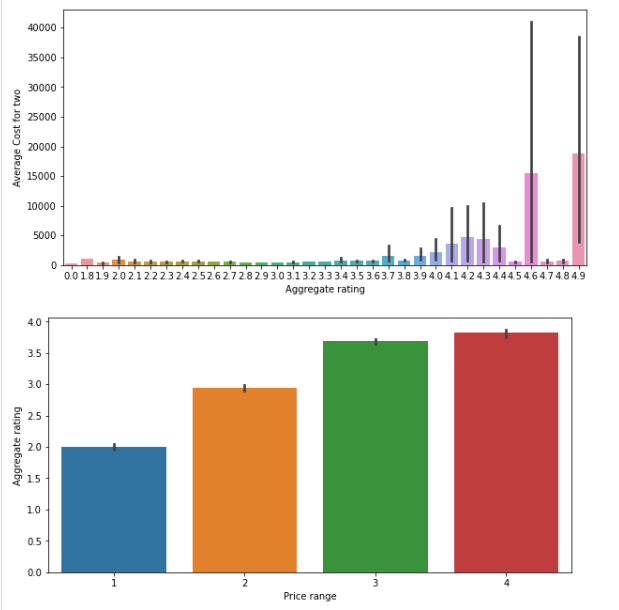
Highest price range of 3.6 for 166 - Qatar and 184 – Singapore.

1. **Latitude vs Longitude for getting co-ordinates:**

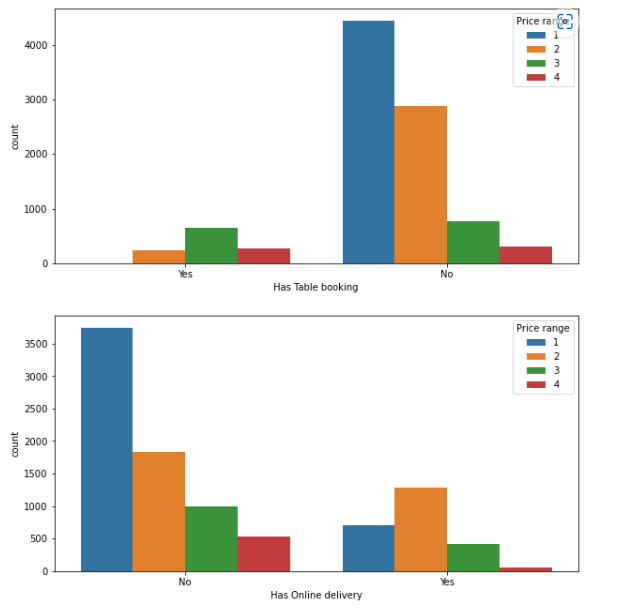


The above points shows geo co-ordinates of restaurants present around the world

1. **Some more plots to visualize:**



We can observe that as price range increases average rating increases and correspondingly there is more Average cost for 2



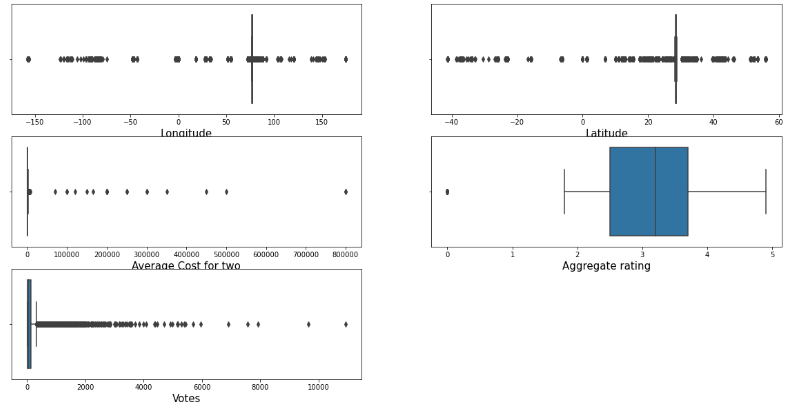
Among the resuarants having no table booking: Price range 1 hotels are most common

Among the resuarants having table booking: Price range 3 hotels are most common

Among the resuarants having no online delivery: Price range 1 hotels are most common

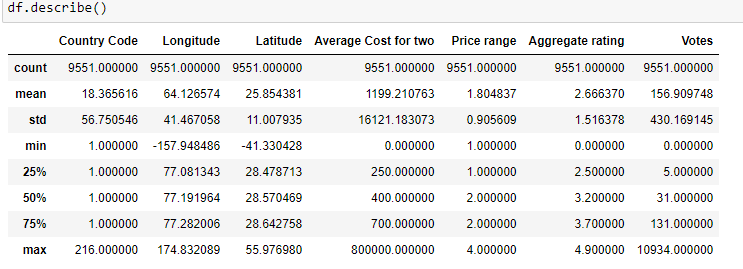
Among the resuarants having online delivery: Price range 2 hotels are most common

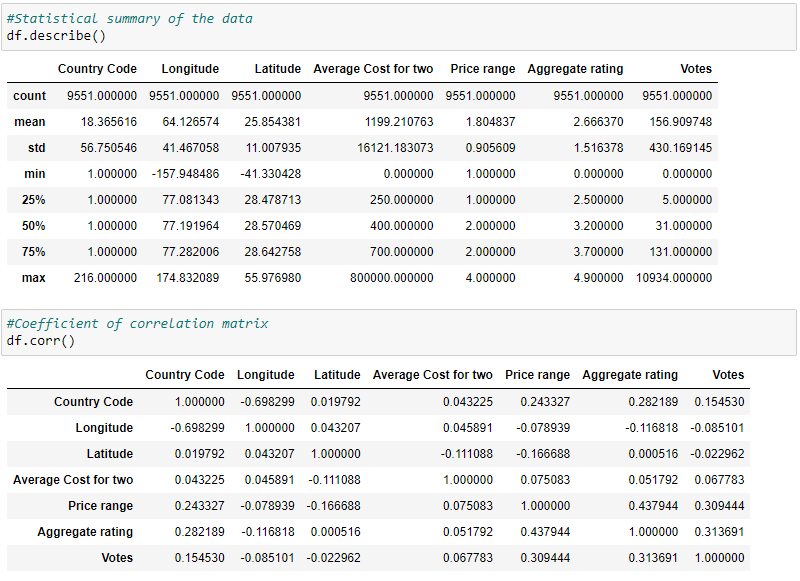
1. **Outliers/noise checking for input numerical feature columns:**



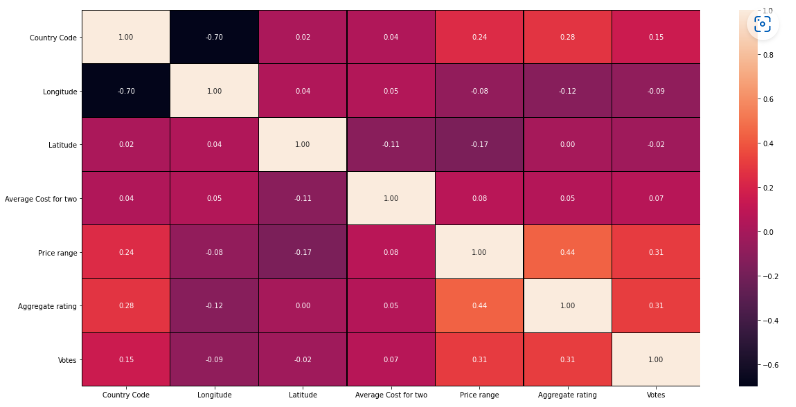
All data has outliers, need to be properly handled

1. **Statistical summary of the data and correlation matrix:**

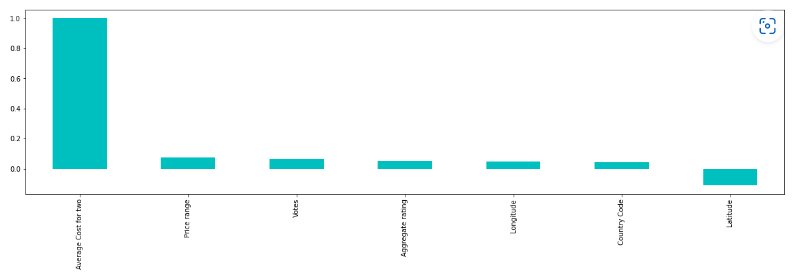




1. **Plotting heat map for proper visualization:**



Sorting descending values of correlation data of target variable **Average Cost for two:**



Average Cost for two 1.000000

Price range 0.075083

Votes 0.067783

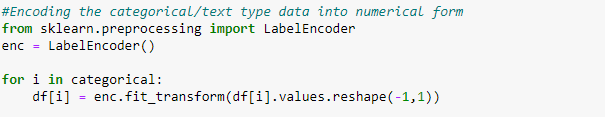
Aggregate rating 0.051792

Longitude 0.045891

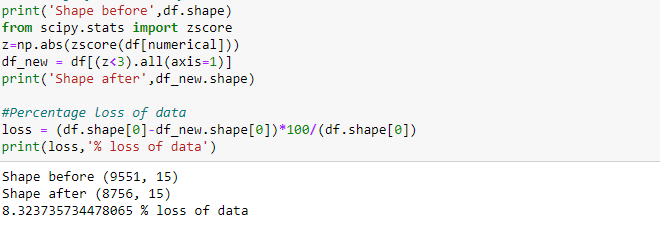
Country Code 0.043225

Latitude -0.111088

1. **Encoding the categorical/text type data into numerical form:**



1. **Removing of outliers by z-score method:**

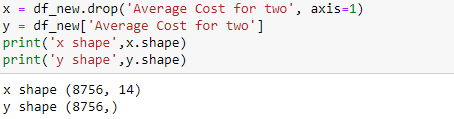


Data loss is less than 10% after removal of outliers.

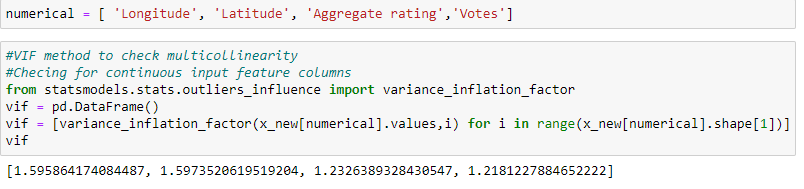
**Model building:**

**Average cost for two prediction (Regression approach)**

**Separating input and target feature columns:**

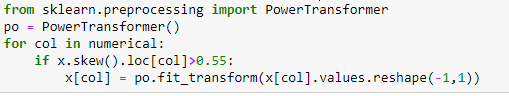


**VIF values checking:**

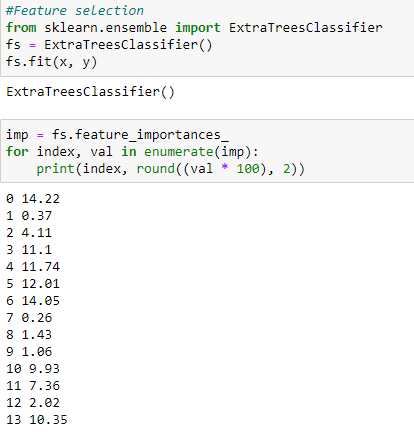


No issue of multicollinearity as the VIF value of columns having continuous data are within the limit of 5

**Apply power transform to reduce skewness to less than 0.55:**



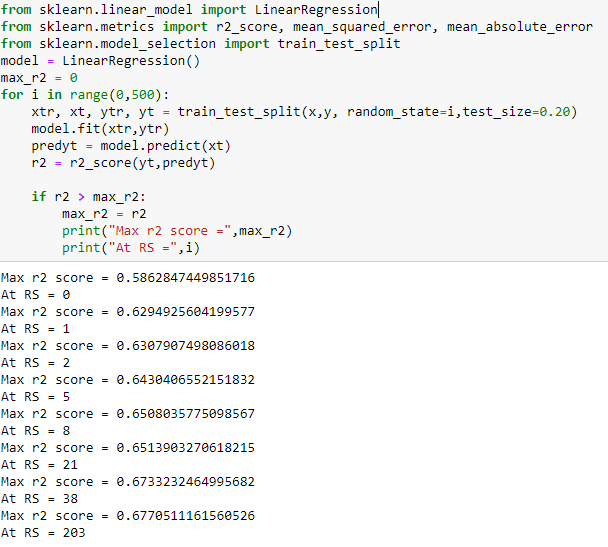
**Most important features selection:**



Let us select important features and drop rest columns. By considering top 70% are important. Dropping off less important columns.



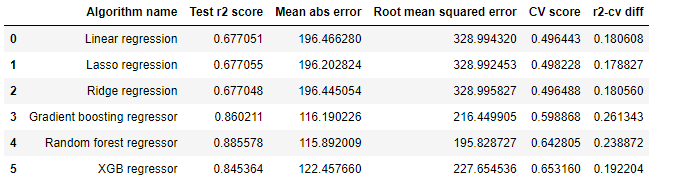
**Selecting of best random state for splitting train-test data:**



**Steps in training ML algorithms:**

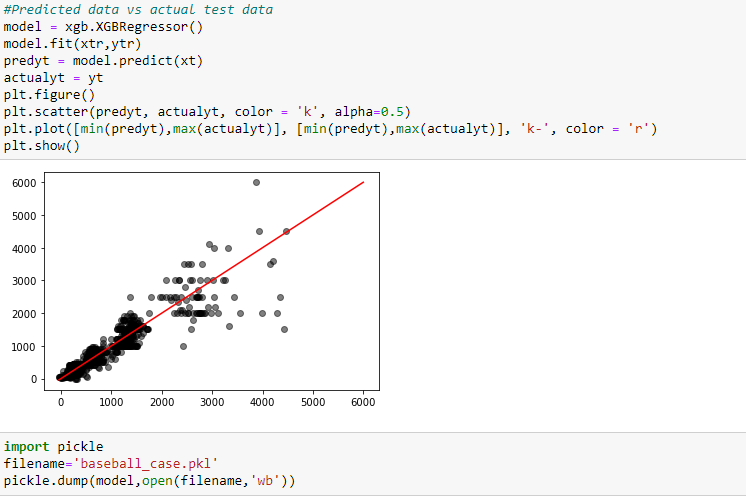
1. Training with best random state
2. Cross validation to check over fitting
3. Hyper parameter tuning and selecting best parameters
4. Training with best parameters
5. Again cross validation to check over fitting
6. Saving the model result into a data frame

The above steps are repeated for following 6 regression algorithms and results compared:



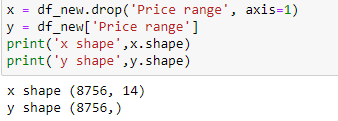
We can observe that XGB regressor giving the best results with max Test R2 score and least difference between CV score and test R2 score.

**Prediction using best selected model and saving the model**:

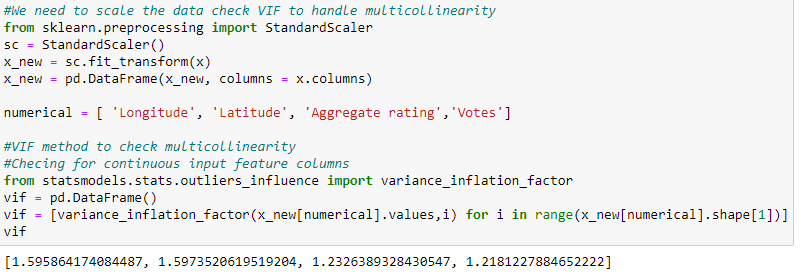


**Price Range prediction (Classification approach)**

**Separating input and target feature columns:**

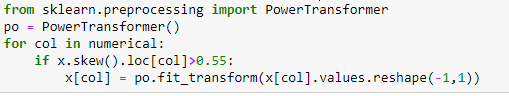


**VIF values checking:**

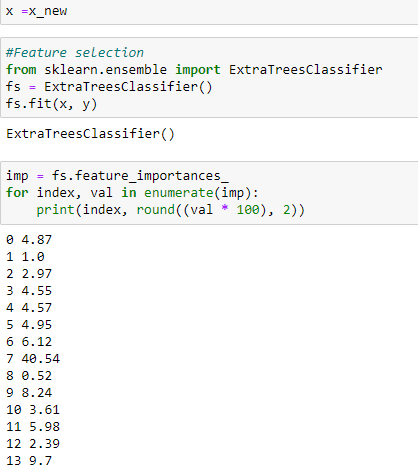


No issue of multicollinearity as the VIF value of columns having continuous data are within the limit of 5.

**Apply power transform to reduce skewness to less than 0.55:**



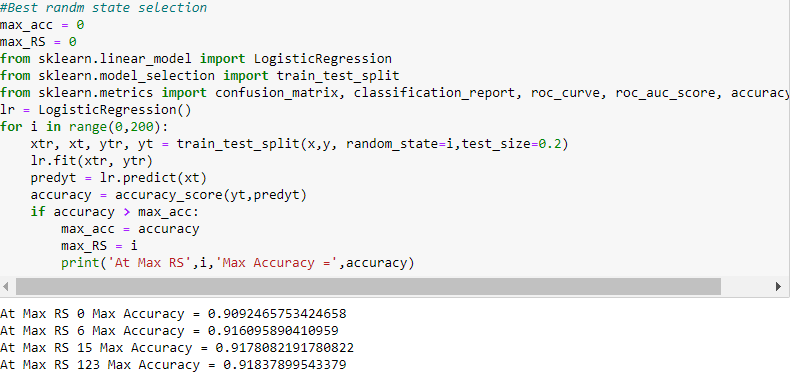
**Most important features selection:**



Let us select important features and drop rest columns. By considering top 70% are important. Dropping off less important columns.



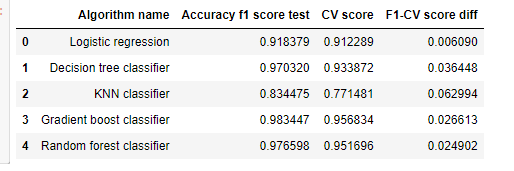
**Selecting of best random state for splitting train-test data:**



**Steps in training ML algorithms:**

1. Training with best random state
2. Cross validation to check over fitting
3. Hyper parameter tuning and selecting best parameters
4. Training with best parameters
5. Again cross validation to check over fitting
6. Saving the model result into a data frame

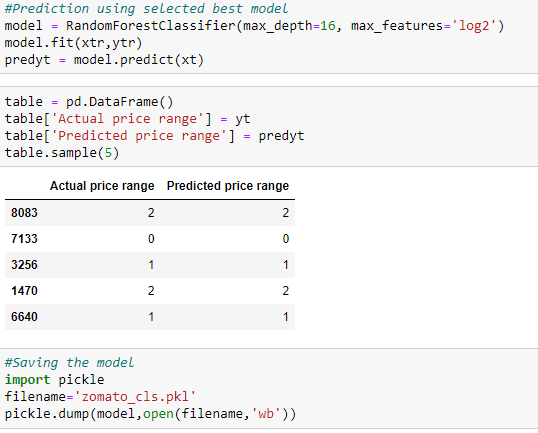
The above steps are repeated for following 5 classification algorithms and results compared:



Both GB classifier and RF classifiers giving best results

Based on lowest CV-F1 test score differnce as minimum RFC can be selected as best model

**Prediction using best selected model and saving the model:**



1. **Conclusion:**

After using all these patient records, we are able to build the machine learning model

1. XGB regression – best one for average cost prediction
2. Random forest classifier – best one for price range prediction

Along with that we were able to draw some insights from the data via data analysis and visualization.