

**Amrita Jena**

**Predictive Modelling Project**

**PGP - Data Science and Business Analytics. PGPDSBA Online May\_B 2021**

**Problem 1: Linear Regression**

**You are hired by a company Gem Stones co ltd, which is a cubic zirconia manufacturer. You are provided with the dataset containing the prices and other attributes of almost 27,000 cubic zirconia (which is an inexpensive diamond alternative with many of the same qualities as a diamond). The company is earning different profits on different prize slots. You have to help the company in predicting the price for the stone on the bases of the details given in the dataset so it can distinguish between higher profitable stones and lower profitable stones so as to have better profit share. Also, provide them with the best 5 attributes that are most important**.

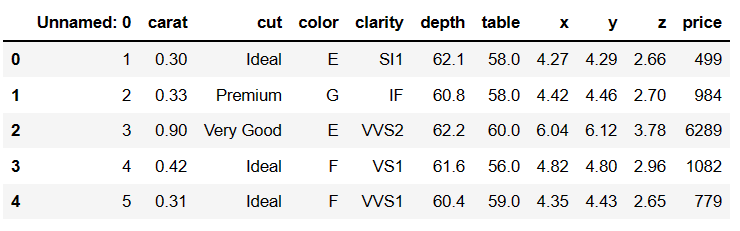
**Data Dictionary:**

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Carat | Carat weight of the cubic zirconia. |
| Cut | Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal. |
| Color | Colour of the cubic zirconia.With D being the worst and J the best. |
| Clarity | Clarity refers to the absence of the Inclusions and Blemishes. (In order from Worst to Best in terms of avg price) IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, l1 |
| Depth | The Height of cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter. |
| Table | The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter. |
| Price | the Price of the cubic zirconia. |
| X | Length of the cubic zirconia in mm. |
| Y | Width of the cubic zirconia in mm. |
| Z | Height of the cubic zirconia in mm. |

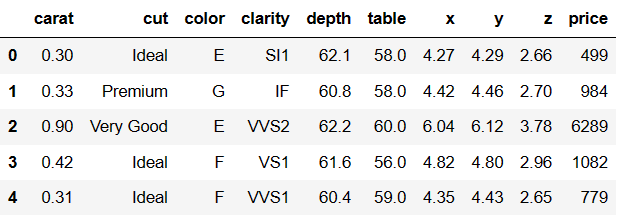
**1.1. Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA, duplicate values). Perform Univariate and Bivariate Analysis.**

We will perform EDA on our dataset. Exploratory Data Analysis or (EDA) is understanding of the data sets by summarizing their main characteristics often plotting them visually. This step is very important especially when we build models from data. Plotting in EDA consists of Histograms, Count plot, Box plot, Pair plot and many more. It often takes much time to explore the data and is a very longwinded process. Through the process of EDA, we can define the problem statement or definition on our data set which is very important.

This is our dataset: -

.

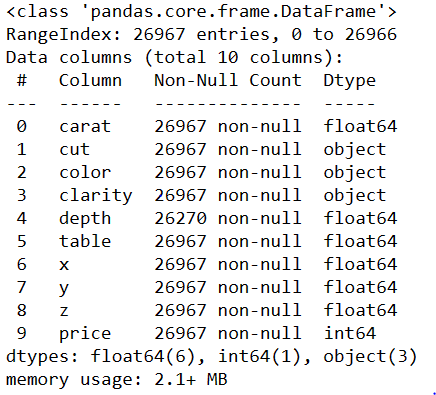
We can see from the above data that there is an unnamed column that provides indexing. Since Pandas itself provides the indexing of the data, the "unnamed" column is unnecessary. Therefore, we can drop that column



**Checking the shape of the data: –**

Previously we were having 26967 rows and 11 columns but after dropping **Unnamed** column now we have 26967 rows and 10 columns.

**Checking the info of the data: –**

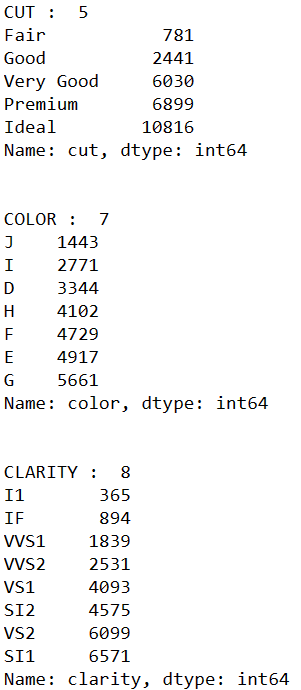


There is total 3 object data types- Cut, Clarity and Color.

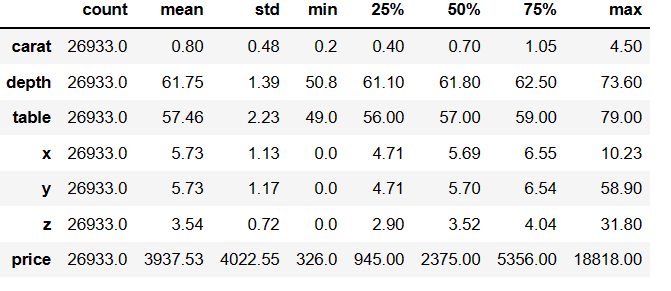
6 float data types- carat, depth, table, x, y and z.

1 int data type ‘Price’. Price is our target variable.

Unique counts of each categorical variables-



**Descriptive Statistics of the data set: -**



Carat: - This is an independent variable, and it ranges from 0.2 to 4.5. mean value is around 0.8 and 75% of the stones are of 1.05 carat value. Standard deviation is around 0.48 which shows that the data is skewed and has a right tailed curve. Which means that majority of the stones are of lower carat. There are very few stones above 1.05 carat.

Depth: - The percentage height of cubic zirconia stones is in the range of 61.75 to 73.60. Average height of the stones is 61.75. 25% of the stones are 61.10 and 75% of the stones are 62.5. Standard deviation of the height of the stones is 1.39. Standard deviation is indicating a normal distribution

Table: - The percentage width of cubic Zirconia is in the range of 57 to 79. Average is around 57. 46. 25% of stones are below 56 and 75% of the stones have a width of less than 59. Standard deviation is 2.23. Thus, the data does not show normal distribution and is similar to carat with most of the stones having less width also this shows outliers are present in the variable.

Price:- Price is the Predicted variable or the Target variable. Prices are in the range of 3938 to 18818. Median price of stones is 2375, while 25% of the stones are priced below 945. 75% of the stones are in the price range of 5356. Standard deviation of the price is 4022. Indicating prices of majority of the stones are in lower range as the distribution is right skewed.

X,Y,Z variables follows a normal distribution with few outliers.

**Now we will check for duplicates:-**

After analysing it is observed that the data frame has 34 no of duplicate rows.



Since we have 34 duplicate records in the data, we will remove this from the data set so that we get only distinct records.

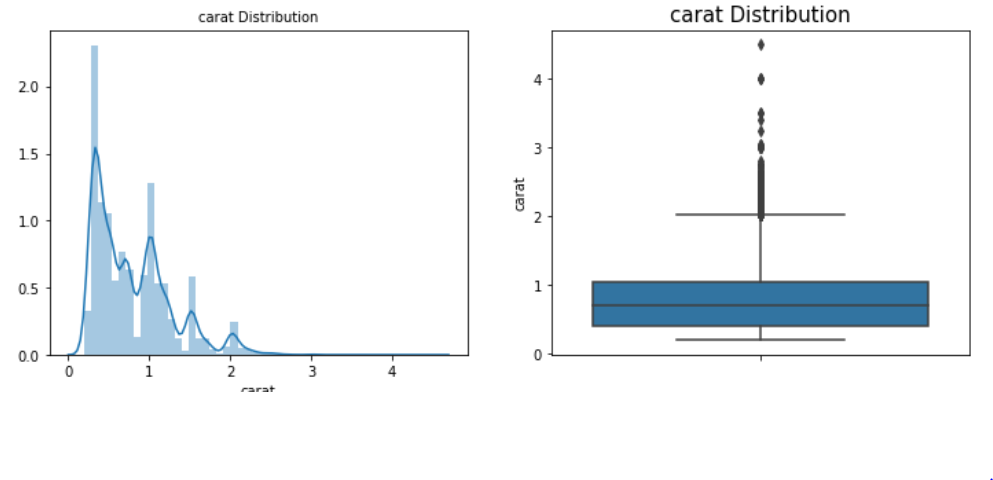
Post removing the duplicate, we will check whether the duplicates has been removed from the data set or not.



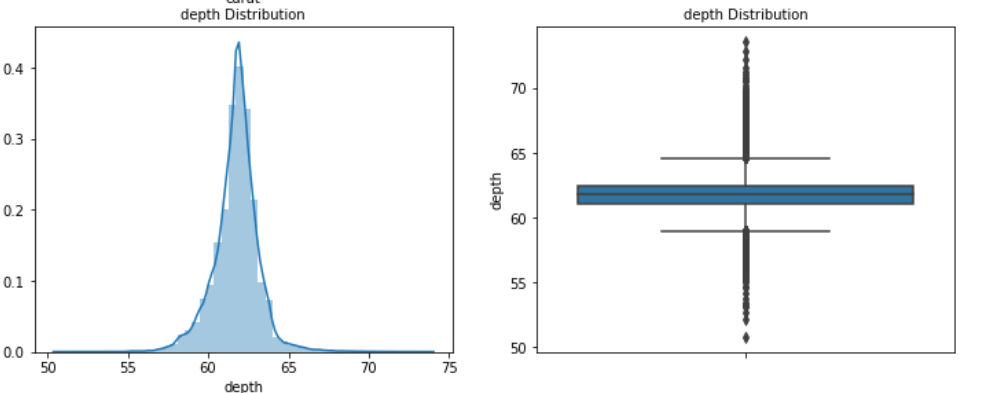


Previously we had 26967 rows and now after removing duplicates we have 26933 rows.

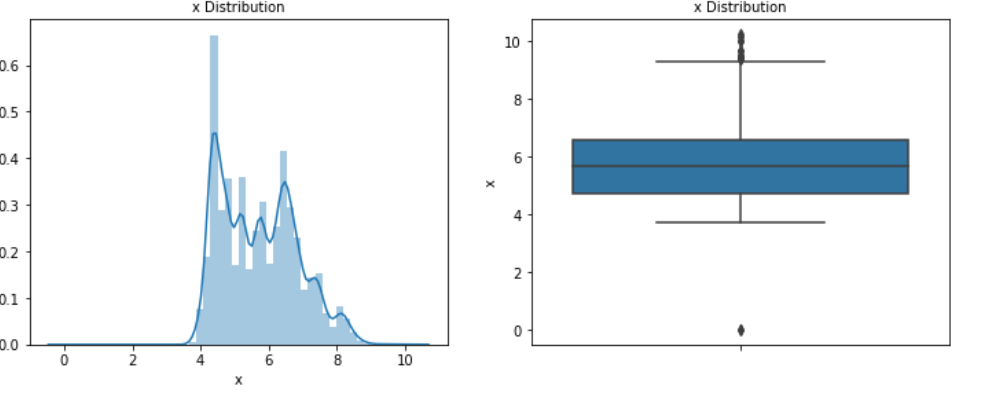
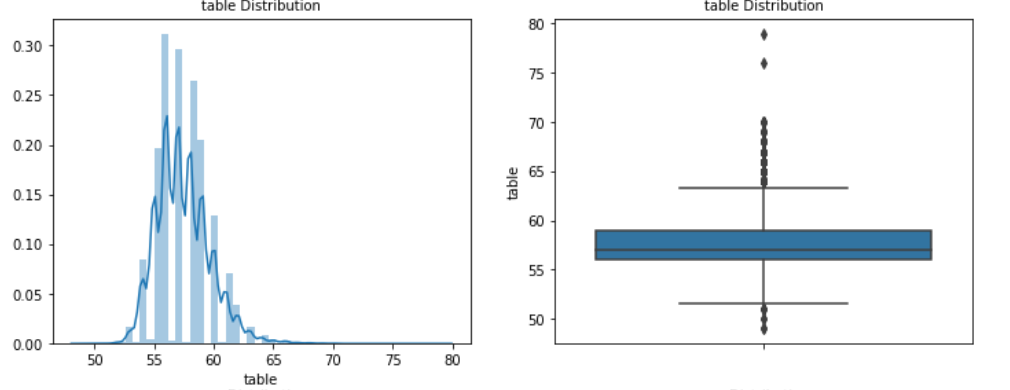
**Univariate \ Bivariate Analysis:-**



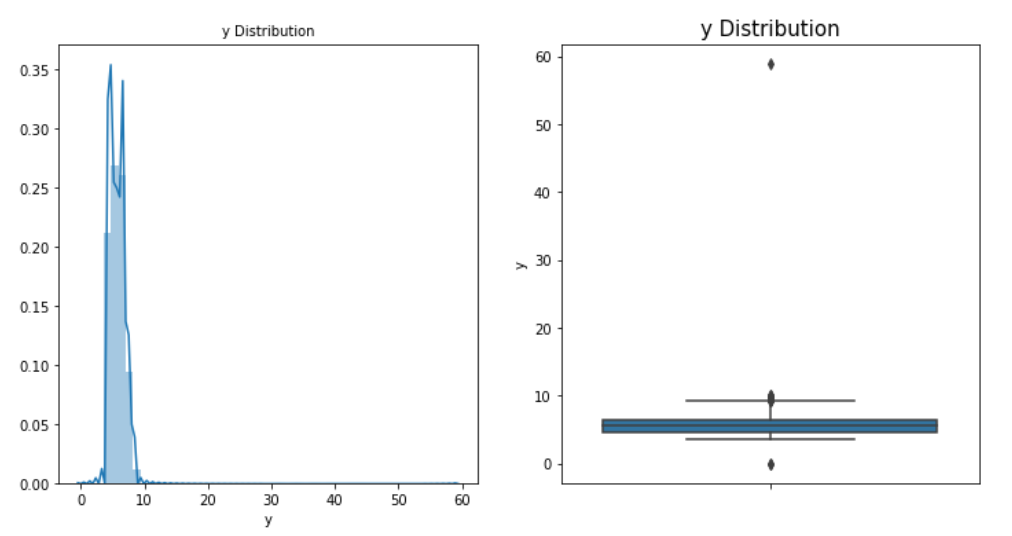
The distribution of data in carat seems to be positively skewed, as there are multiple peaks points in the distribution it could be multimodal and the box plot of carat seems to have large number of outliers. It seems in the range of 0 to 1 maximum data lies.

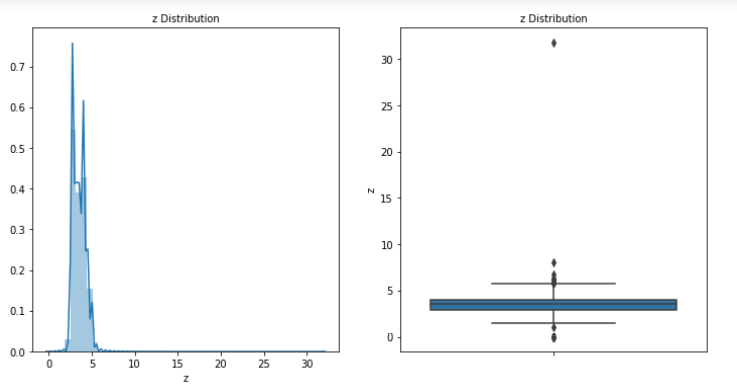


The distribution of depth appears to be normally distributed. The depth ranges from 55 to 65 and holds many outliers.

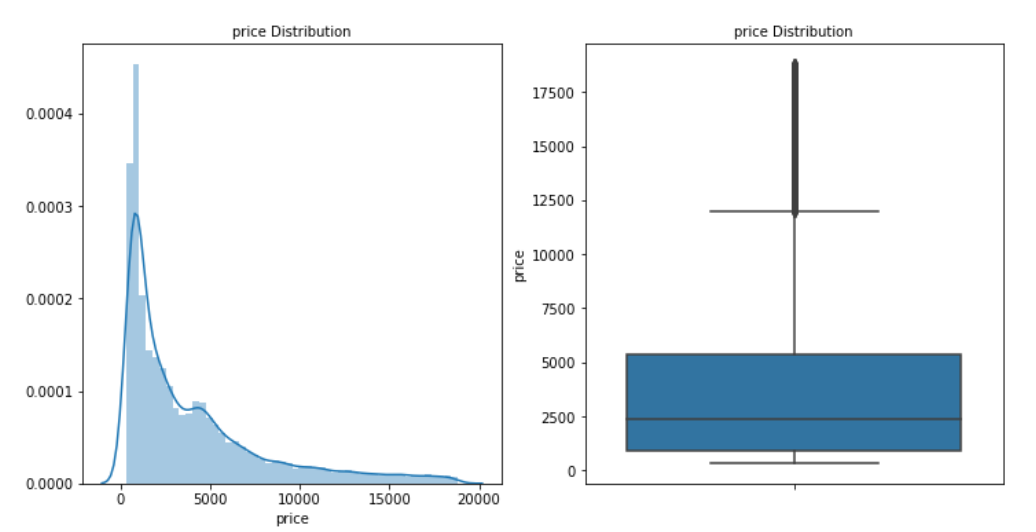


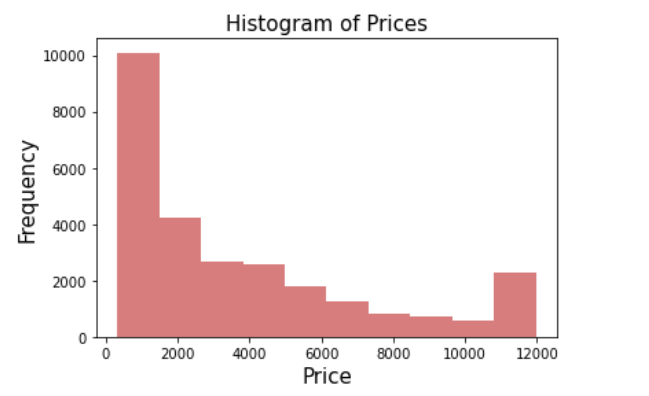
Both table and x variable seem positively skewed with outliers present.

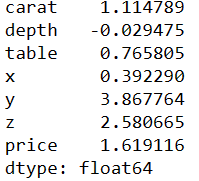




Z (Height of the cubic zirconia in mm.) and Y variable (Width of the cubic zirconia in mm.) is highly positive skewed and also consists of outliers. The distribution is too much positively skewed which may be due to the fact that diamonds are always made in a specific shape and there might not be too many sizes available in the market.

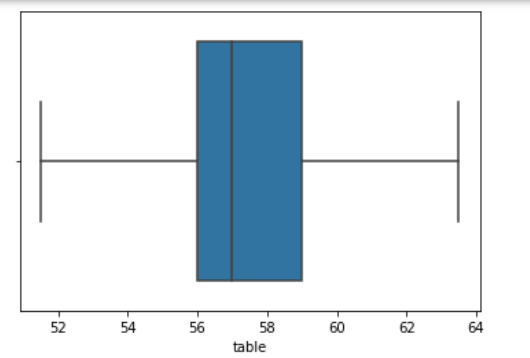
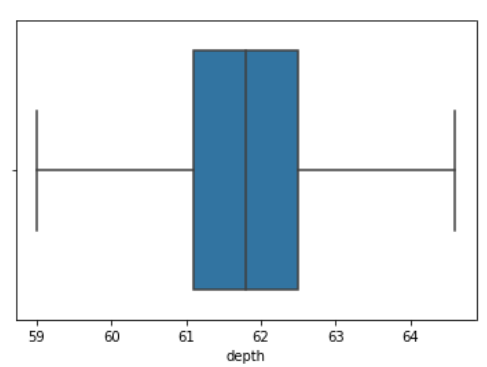
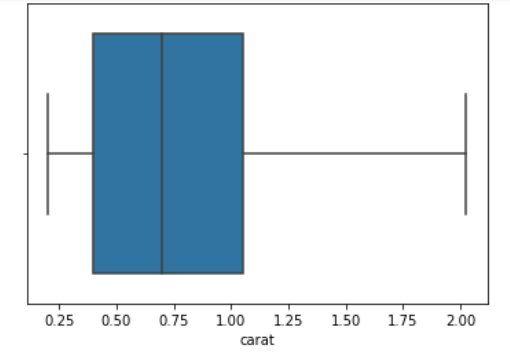


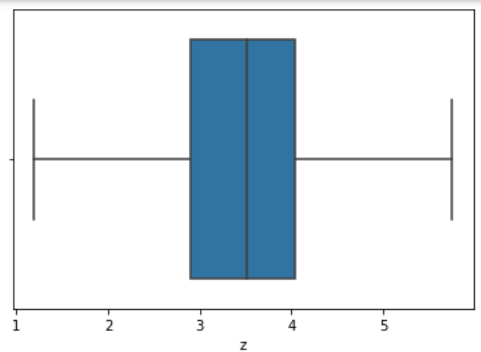
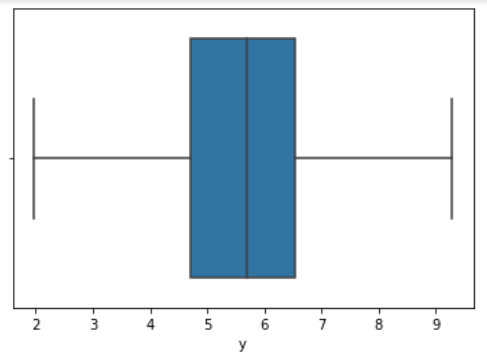
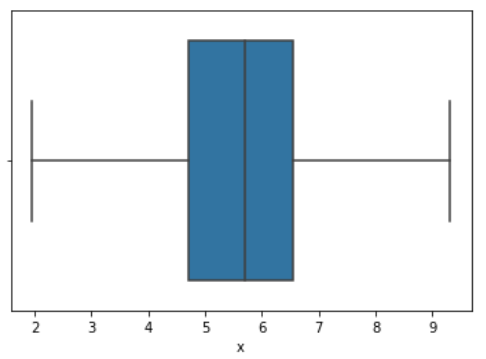


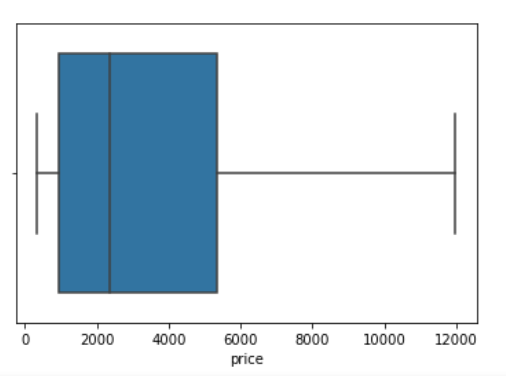


Our target variable is also positively skewed and contains outliers. From the above it appears the distribution of prices is skewed to the right i.e., as the prices increase, the number of diamonds decrease. Our skew is a positive value (greater than 1) indicating a positive skewness which means that our mean is greater than the median.

**The outliers were removed after treating:-**





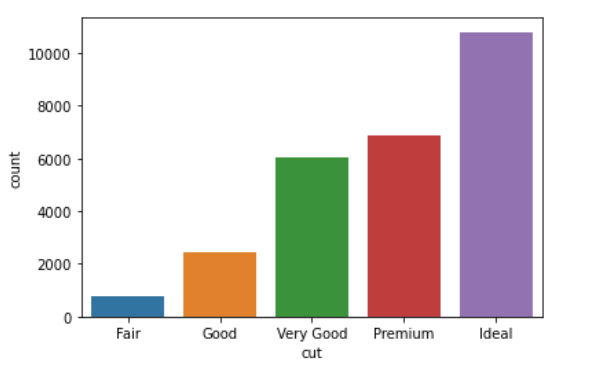
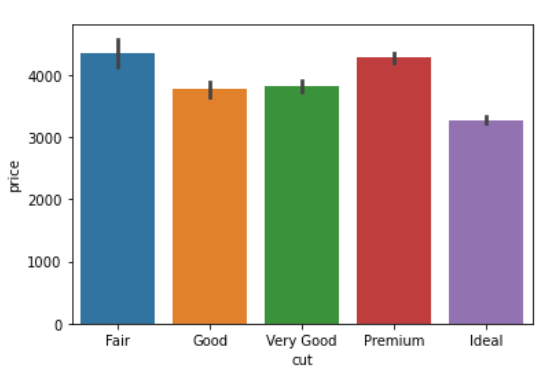


**Categorical Variables-**

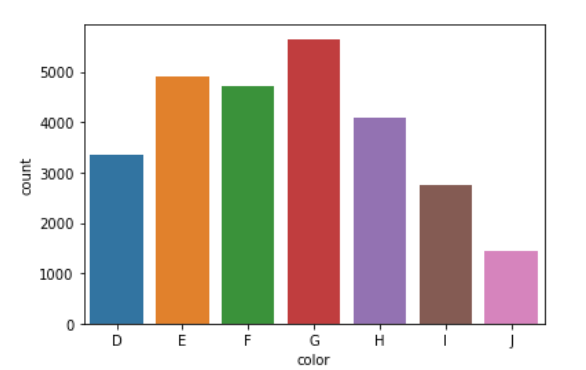
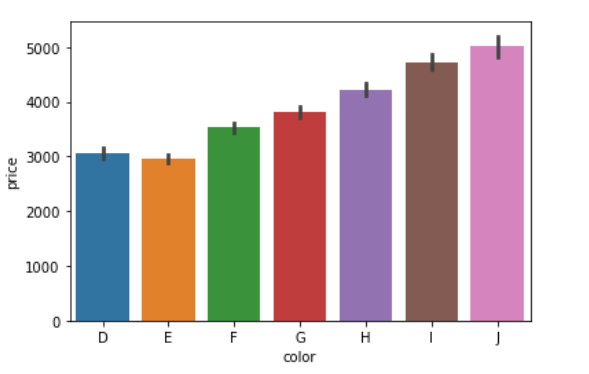
There are three categorical variables - Cut, Colour and Clarity

**CUT-**

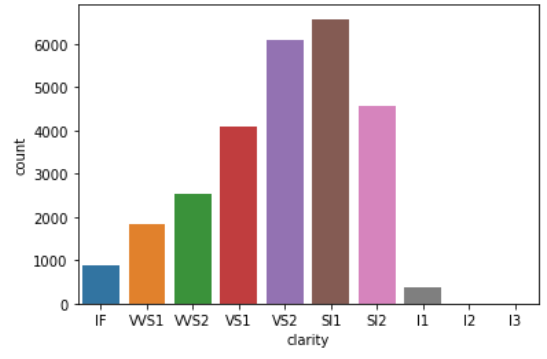
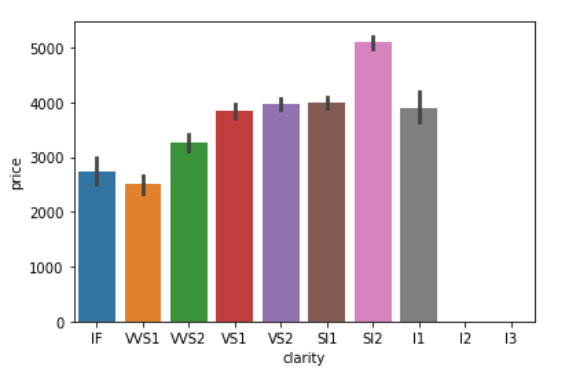
.

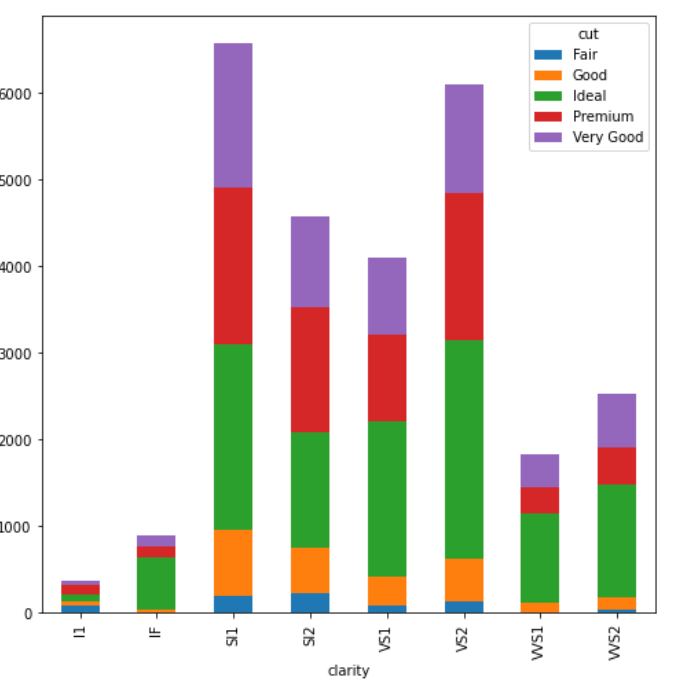
Different varieties of cut are arranged in increasing order. From the above plot it is observed that ‘Ideal’ cut is the most purchased cut as its lower in price.

Frpm the above plot we could see that G is the most preferred colour followed by E,F,H. But J being the most unpreferable colour is highly priced than other colours.

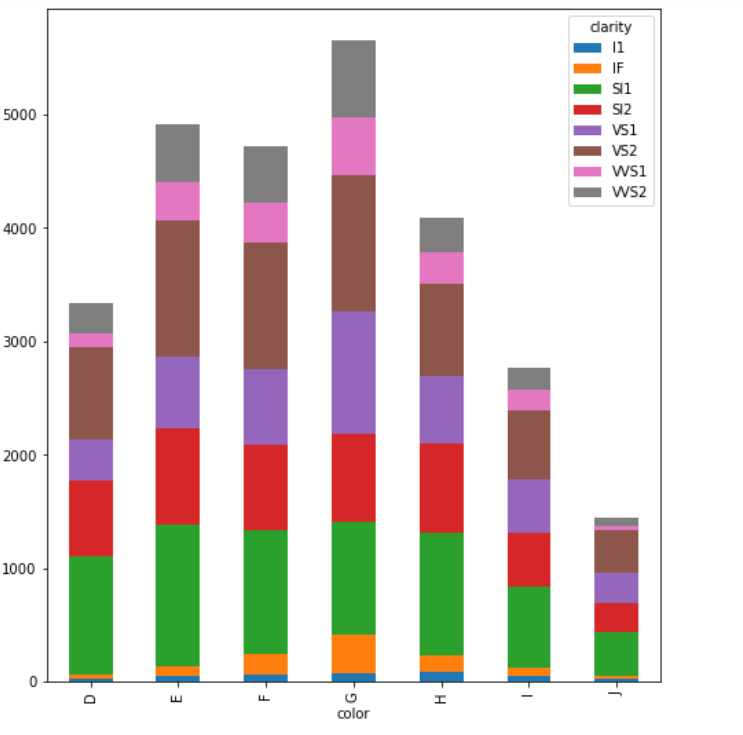
SI1 is the most preffered clarity criteria in diamonds followed by VS2 and SI2. Its also intermideately priced.



We can see from above that most of the buyers prefer to buy diamond of SI1 clarity followed by VS2, SI2, and VS1.In that, the cut they prefer is Ideal, Premium, and very good's diamond cut category. Moreover, we can infer that people are not taking the highest clarity diamonds, such as IF or VVS1 and others and are ready to sacrifice on clarity but are more focusing on the cut of the diamonds.



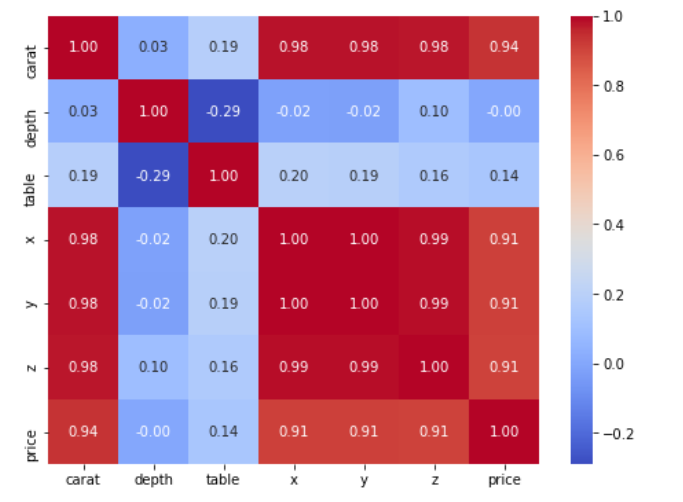
We can see that people prefer Ideal cut over any other cut diamonds followed by Premium and Very Good. It suggests that people are focusing on cut than clarity.



We can see that from above that most of the people prefer G color followed by E, F, and H.In that the clarity they mostly prefer SI1 followed by VS1,VS2 and SI2 category.

# MULTIVARIATE ANALYSIS:-

# 

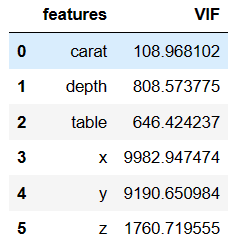
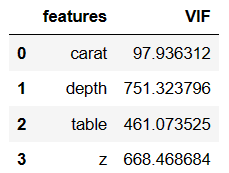
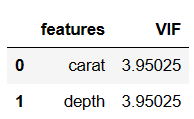
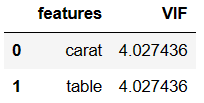
 

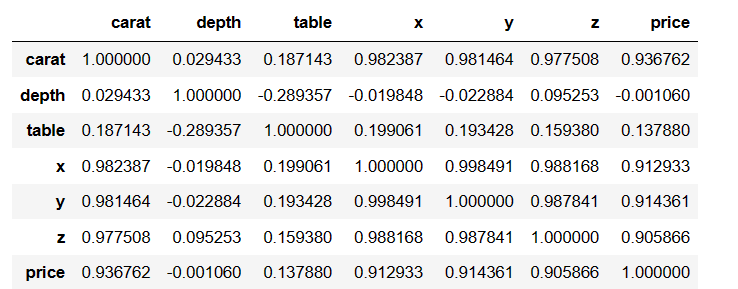
From the heatmap and pair plot the presence of multicollinearity is clearly manifested. Other than depth and table we can state that there exists a strong correlation amongst most of the variables as the majority of the data value in each cell in more than .75, which suggests multicollinearity.

* depth has a negative correlation with all other variables suggesting that this variable is inversely proportional to all other variables.
* Similarly, table has a very weak correlation with other variables but does not an inverse relationship.

The target variable price has a very strong correlation with carat, x, y and z.

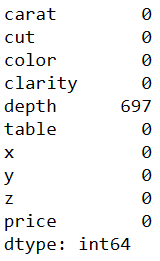
**To reduce the multicollinearity, VIF will be applied and hence remove any feature that has the highest VIF score and return the test until we reach the tolerated limit of five. In this scenario when we retain carat and either depth/table we will be also able to achieve VIF score**.



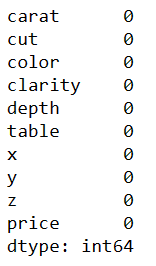
**1.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Check for the possibility of combining the sub levels of a ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning.**

**Checking for null values-**



There are **697** null values i.e. 2.52% in Depth variable and no other variables has null values. There are various ways of treating our missing values in the data set. And which technique to use when is actually dependent on the type of data we are dealing with.

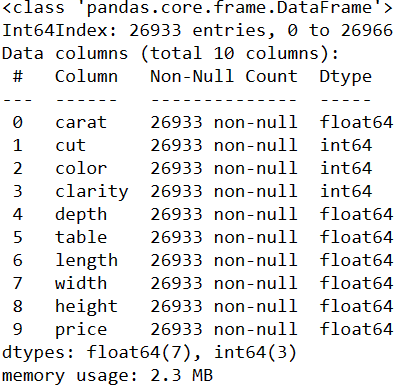
We will use median impute technique to treat null values as our data had outliers. After applying median impute no null values are present in our data set.

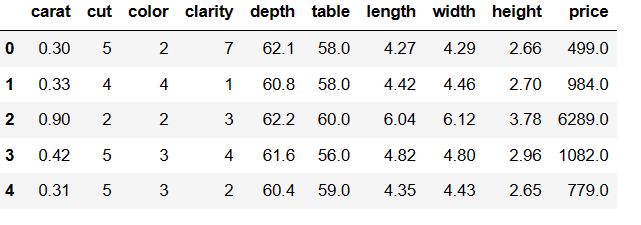


In the data frame we could see the x,y,z column which indicates Length of the cubic zirconia in mm., Width of the cubic zirconia in mm. and Height of the cubic zirconia in mm. X,Y,Z appears to be vague as one may not suddenly know what it stands for and is not clear and are misguiding . Therefore, we could rename the columns as length, width and height.

So I can modify my columns name for better understanding**.**

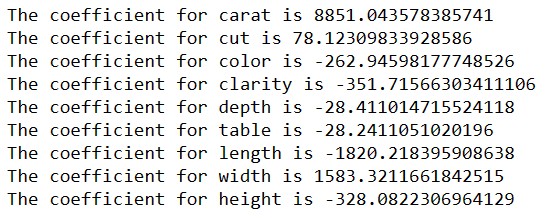
Then all the Ordinal Categorical columns were encoded and converted into numeric data type.





**1.3 Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning.**

* **First we Split X and y into training and test set in 70:30 ratio.**
* **Then we invoke the Linear Regression function to find the best fit model on training data.**
* **Coefficients for each of the independent attributes are as follows**:



* **The intercept for our model is 5028.30032407703**
* **After applying R^ on training data we get value of 0.9205774490301728**
* **After applying R^ on testing data we get value of 0.9216634433853348**
* **By applying RMSE on Training data and Testing data we get values 976.6853506548496 and 972.3231690793775.**

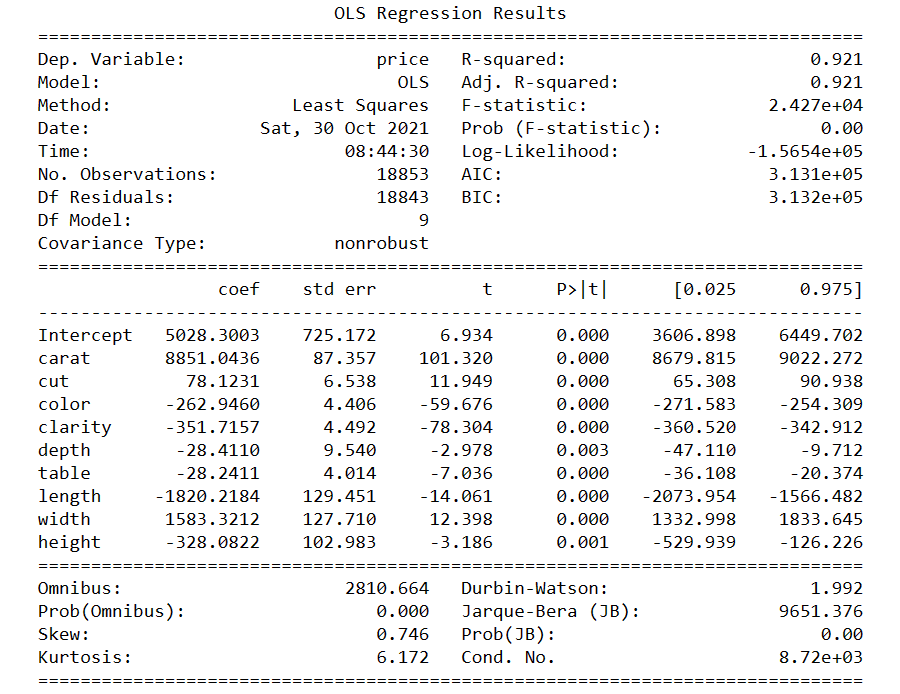
**Using Stats model library to get R type outputs.**

**R^2 is not a reliable metric as it always increases with addition of more attributes even if the attributes have no influence on the predicted variable.**

**Instead, we use adjusted R^2 which removes the statistical chance that improves R^2.**

**Scikit does not provide a facility for adjusted R^2... so we use statsmodel, a library that gives results similar to what we obtain in R language .This library expects the X and Y to be given in one single data frame.**

* **We merged X and Y.**
* **Then we obtained the lml summary**



* **Let us check the sum of squared errors by predicting value of y for test cases and subtracting from the actual y for the test cases. This could be pertaining to low p value of all dependent variables.**
* **We applied Underroot of mean\_sq\_error which is the standard deviation i.e. avg variance between predicted and actual values - 972.3231690793758**
* **We obtained # Model score - R2 or coeff of determinant(formula R^2=1–RSS / TSS)-**

**0.9216634433853348**

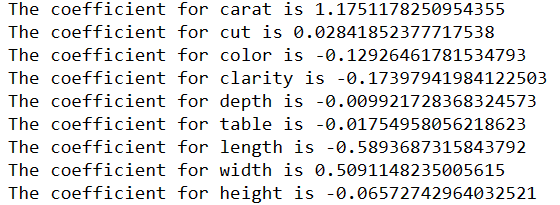
**ITERATION 2-**

**How do we improve the model? the R^2 is .909, how do we improve it.**

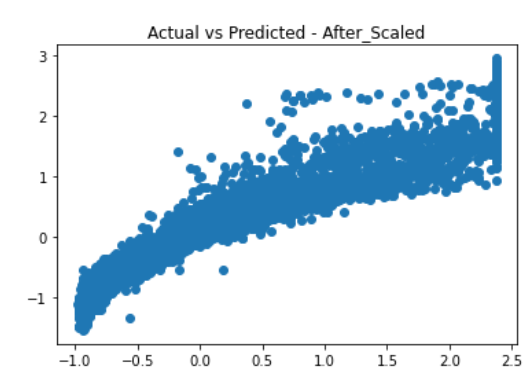
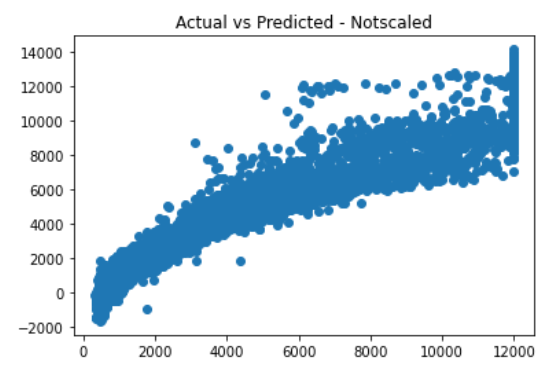
**The indpendent attributes have different units and scales of measurement**

**It is always a good practice to scale all the dimensions using z scores or some other method to address the problem of different scales**

* **Coefficient and intercept after scaling –**

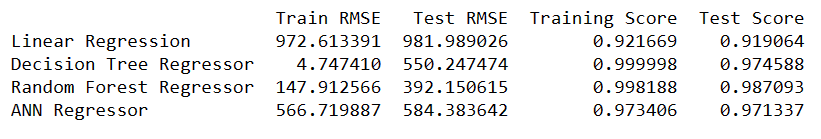


* **The intercept for our model is -1.9111347093590655e-16 after scaling**
* **R^ or coeff of determinant after scaling is 0.9216619580771904**
* **mean\_sq\_error after scaling is 0.2798893387087284**



We could observe a difference in scaled and unscaled data the co-efficient and intercept came down to almost zero in the scaled data compared to the unscaled dataset. Similarly in scatter plot plotted we can clearly see that except for the scale we don’t see any change in the relationship or model performance.

Due to low performance of the model we try to build other models like Linear Regression, Decision Tree , Random Forest an ANN . We used unscaled data for Linear Regression, Decision Tree and Random Forest whereas we used scaled data for ANN model alone.

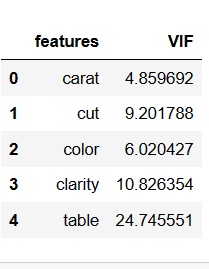


On the basis of RMSE scores we could say that Decision Tree and Random Forest have very large gap between Train and Test RMSE. So we will go for a model with low RMSE. Hence we will compare both ANN regressor and Linear Regressor.

Model accuracy of all four models are all considerable. After comparing them with RMSE and the Test and Train scores of ANN and Linear Regressor we can go for the ANN model first but linear regression model also is not of much difference.

**1.4 Inference: Basis on these predictions, what are the business insights and recommendations.**

The five most promising variables out of the 9 variables which will bring maxuimum profit for the diamond business are based on VIF score and they are **carat, cut, color, clarity and table.**



**Inference:**

• Ideal cut has high demand even though; Premium and Fair cut are the most profit yielding stones and the profit margin for these two cuts is on the higher side.

• Clarity refers to the absence of the Inclusions and Blemishes and has emerged as a strong predictor of price as well.

* Clarity of stone types IF, VVS\_1, VVS\_2 and vs1 are helping the firm put an expensive price cap on the stones.
* Color of the stones such H, I and J won't be helping the firm put an expensive price cap on such stones. T
* The company should instead focus on stones of color D, E and F to command relative higher price points and support sales.
* As Carat, Cut, Color, Clarity and Table are the primary key features attributing in profit more variation should be explored of the same features.
* This also can indicate that company should be looking to come up with new color stones like clear stones or a different color/unique color that helps impact the price positively.
* The company should focus on the stone's carat and clarity so as to increase their prices. Ideal customers will also contribute to more profits. The marketing efforts can make use of educating customers about the importance of a better carat score and importance of clarity index. Post this, the company can make segments, and target the customer based on their income/paying capacity etc, which can be further studied. Recommendation:

# **Problem-2**

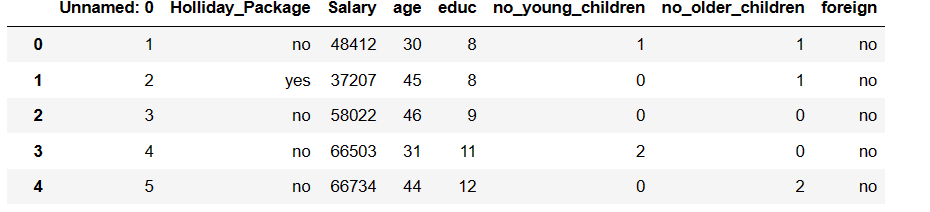
# ****Problem 2: Logistic Regression and LDA-You are hired by a tour and travel agency which deals in selling holiday packages. You are provided details of 872 employees of a company. Among these employees, some opted for the package and some didn't. You have to help the company in predicting whether an employee will opt for the package or not on the basis of the information given in the data set. Also, find out the important factors on the basis of which the company will focus on particular employees to sell their packages.****

## **Data Dictionary**

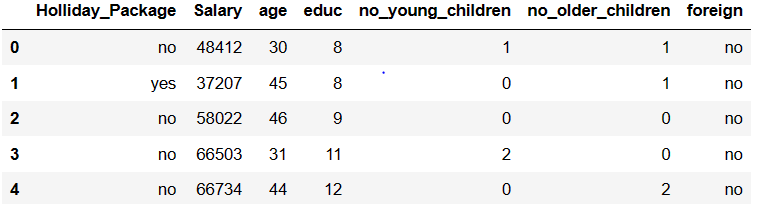
Table 2.1 Data Dictionary for Holiday Package Dataset

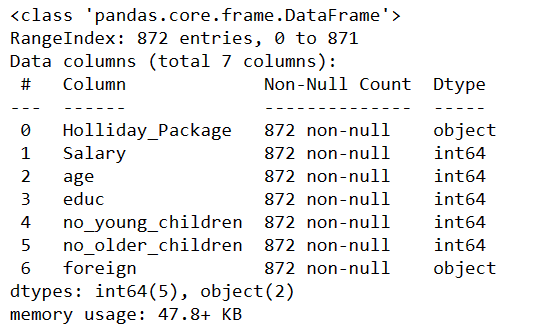
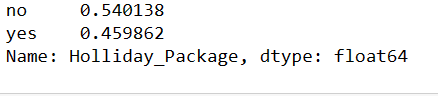
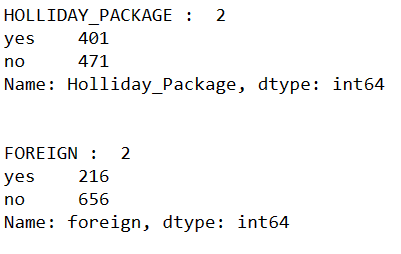
|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| **Holiday\_Package** | Opted for Holiday Package yes/no? |
| **Salary** | Employee salary |
| **age** | Age in years |
| **edu** | Years of formal education |
| **no\_young\_children** | The number of young children (younger than 7 years) |
| **no\_older\_children** | Number of older children |
| **foreign** | foreigner Yes/No |

**Dataset Sample-**



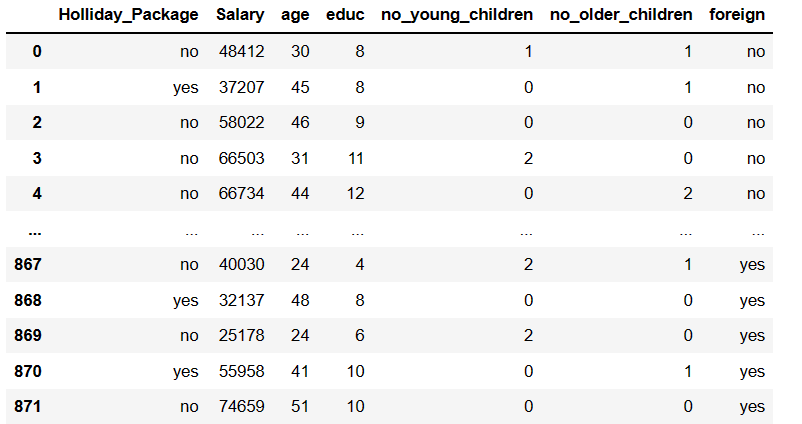
***2.1 Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis.***



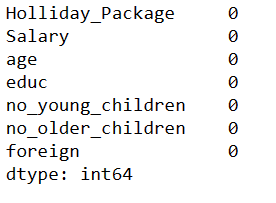
There are 872 rows and 7 columns in the dataset. Out of 7 columns, 5 columns are of Integer data type and 2 columns are of Object datatype. Holliday Package is our target data .

Let us check whether any of the columns has any value other than numeric i.e. data is not corrupted such as a "?" instead of a number.we use np.isreal a numpy function which checks each column for each row and returns a bool array, where True if input element is real. Apply map is pandas data frame function that applies the np.isreal function column wise.Following line selects those rows which have some non-numeric value in any of the columns hence the ~ symbol



There are no corrupted values present in our data set.

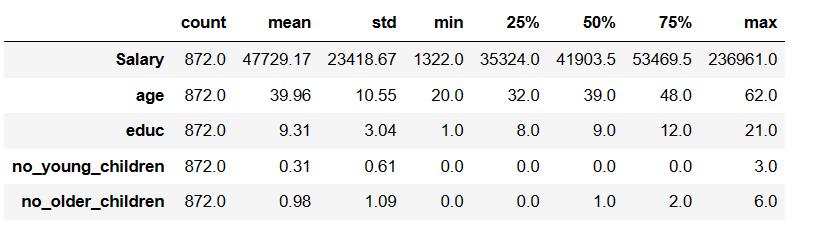
**Checking for null and duplicate values: -**





No null and duplicate values are found in the dataset.

**Descriptive Summary-**



Holiday Package — This variable is a categorical Variable. output with the This will be our Target Variable.

• Salary, age, educ, no\_young\_children, no\_older\_children, variables are numerical or continuous variables.

• Salary ranges from 1322 to 236961. Average salary of employees is around 47729 with a standard deviation of 23418. Standard deviation indicates that the data is not normally distributed. skew of 0.71 indicates that the data is right skewed and there are few employees earning more than an average of 47729. 75% of the employees are earning below 53469 while 25% of the employees are earning 35324.

• Age of the employee ranges from 20 to 62. Median is around 39. 25% of the employees are below 32 and 25% of the employees are above 48. Standard deviation is around 10. Standard deviation indicates almost normal distribution.

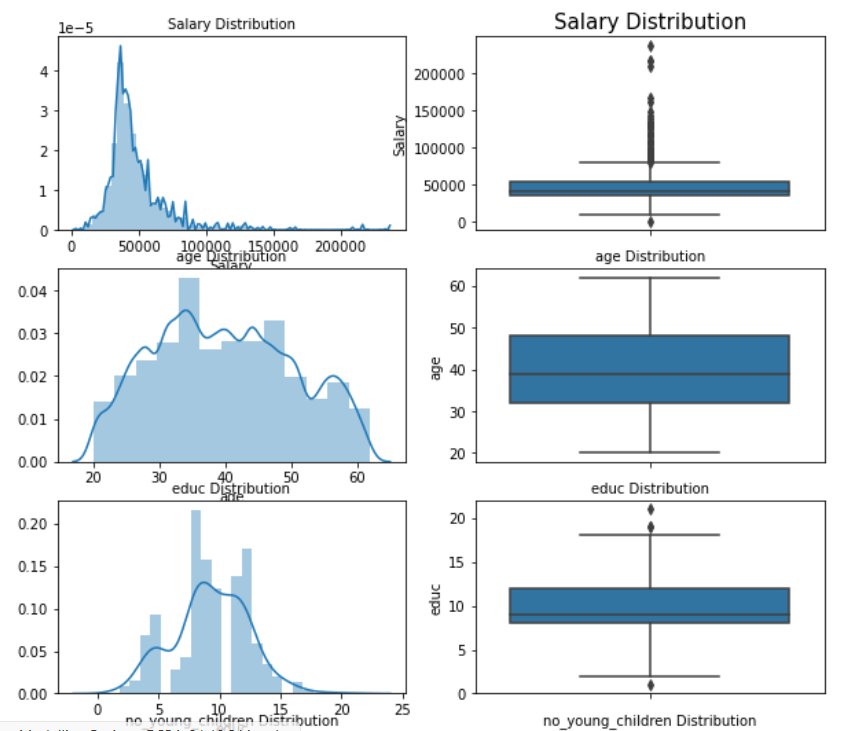
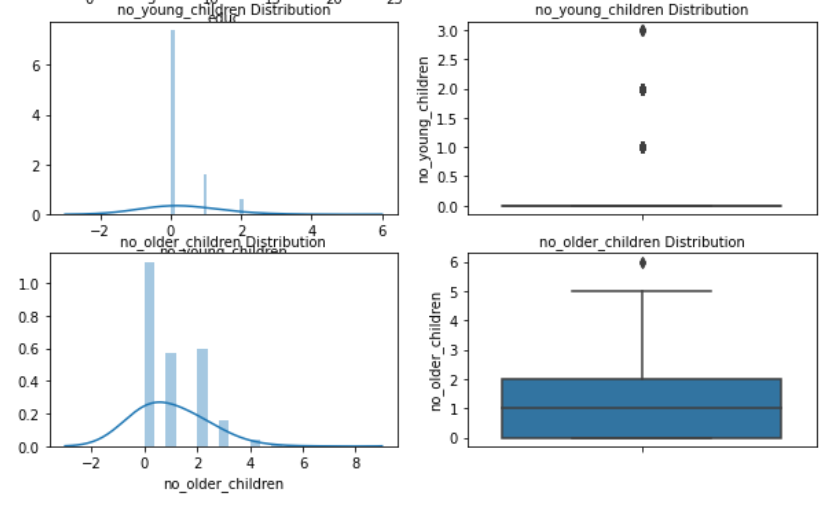
• Years of formal education ranges from 1 to 21 years. 25% of the population has formal education for 8 years, while the median is around 9 years. 75% of the employees have formal education of 12 years. Standard deviation of the education is around 3. This variable is also indicating skewness in the data

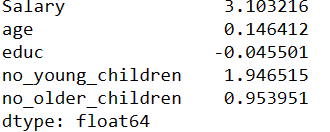
• Foreign is a categorical variable

• We have dropped the first column 'Unnamed: 0' column as this is not important for our study. Unnamed is a variable which has serial numbers so may not be required and thus it can be dropped for further analysisThe shape would be — 872 rows and 7 columns

There are no null values and duplicate values.

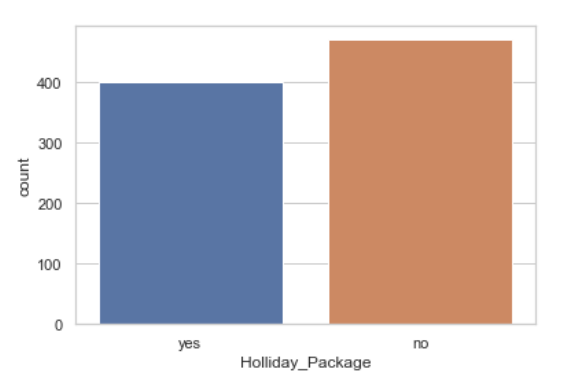
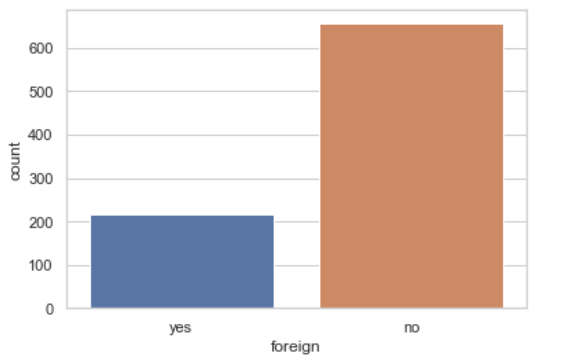
**Univariate/Bivariate Analysis-**





From the plots and from the skewness result we could infer that educ is negatively skewed and rest all variables are positively skewed. Outlier is present in all continuous variable except age variable.

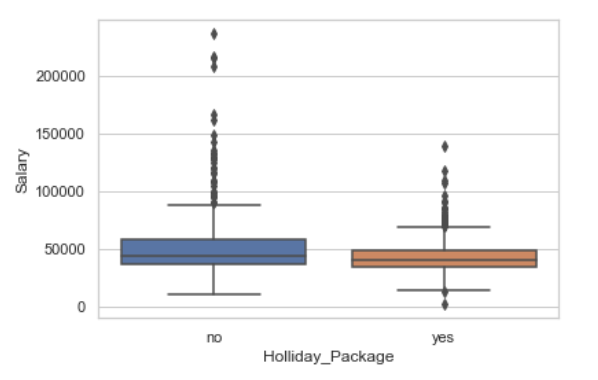
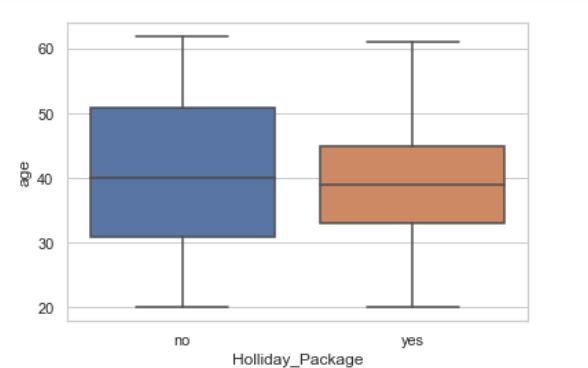
**Checking unique values for target variable 'Holliday \_Package' and 'foreign' both are categorical in nature: -**

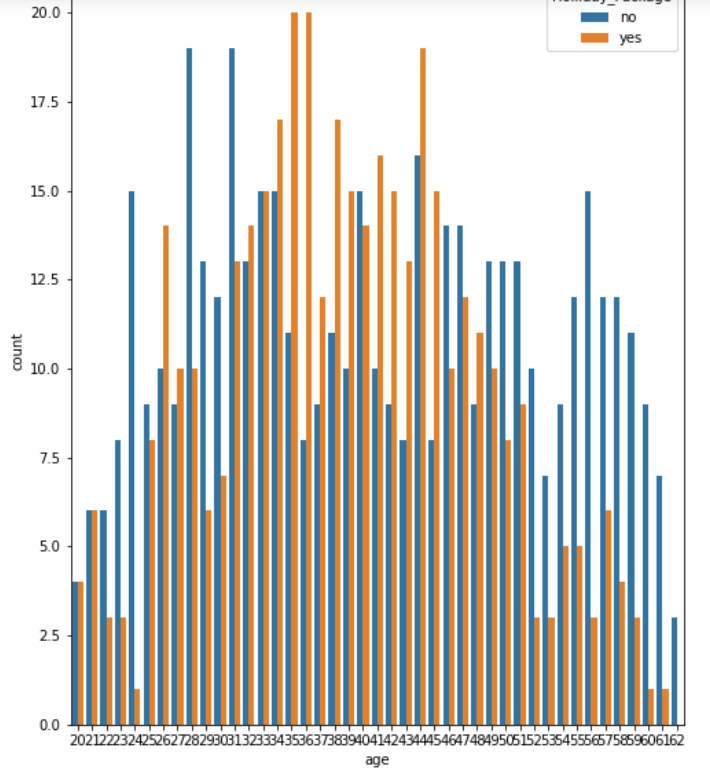
 

We can observe that 54% of the employees are not opting for the holiday package and 46% are interested in the package. This implies we have a dataset which is fairly balanced.

In foreingner distribution we could see that 25% are foreigners and 75% are not foreigners.

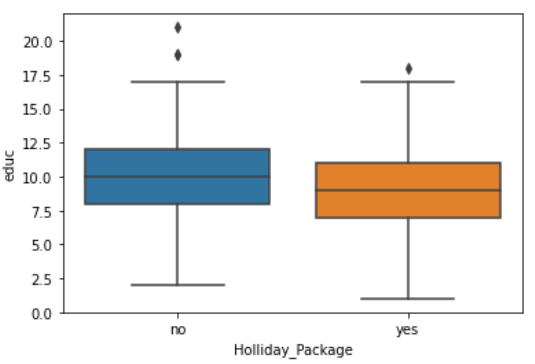
**Bivariate analysis with Target Variables:-**



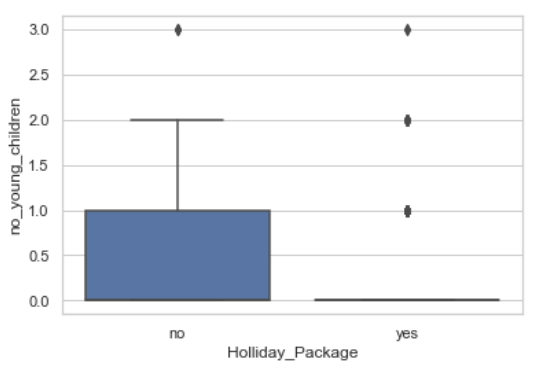
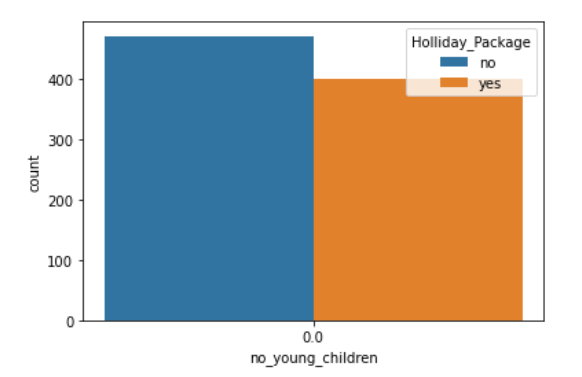


While performing the bivariate analysis we observe that Salary for employees opting for holiday package and for not opting for holiday package is similar in nature. However, the distribution is fairly spread out for people not opting for holiday packages.

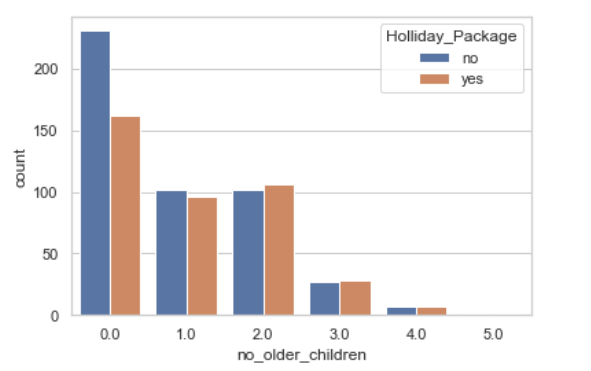
In Age distribution it could be seen that age group between 35-45 i.e the middle age group population are on higher sides for availing holiday pacheges than old or young age people.



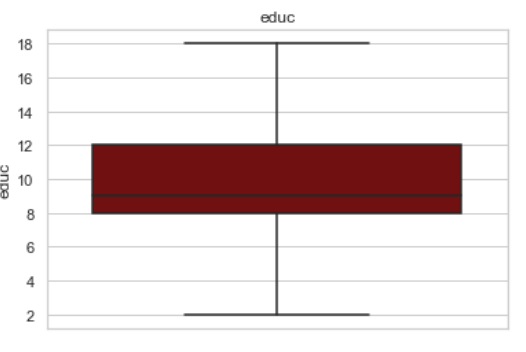
Education variable is not very influential in opting for holiday packages. It could be observed that employees with less years of formal education(1 to 7 years) and higher education are not opting for the Holiday package as compared to employees with formal education of 8 year to 12 years.

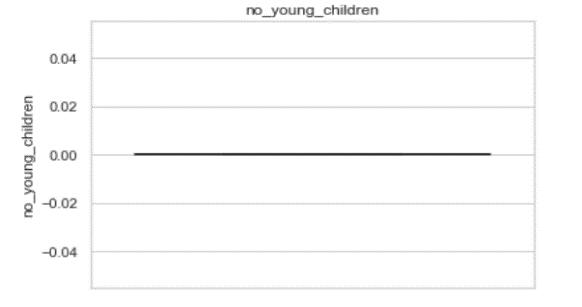


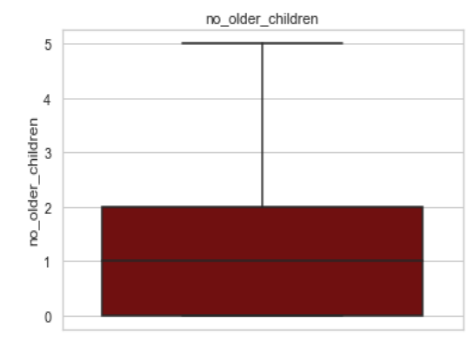
Here is a significant difference in employees with younger children who are opting for holiday package and employees who are not opting for holiday package.



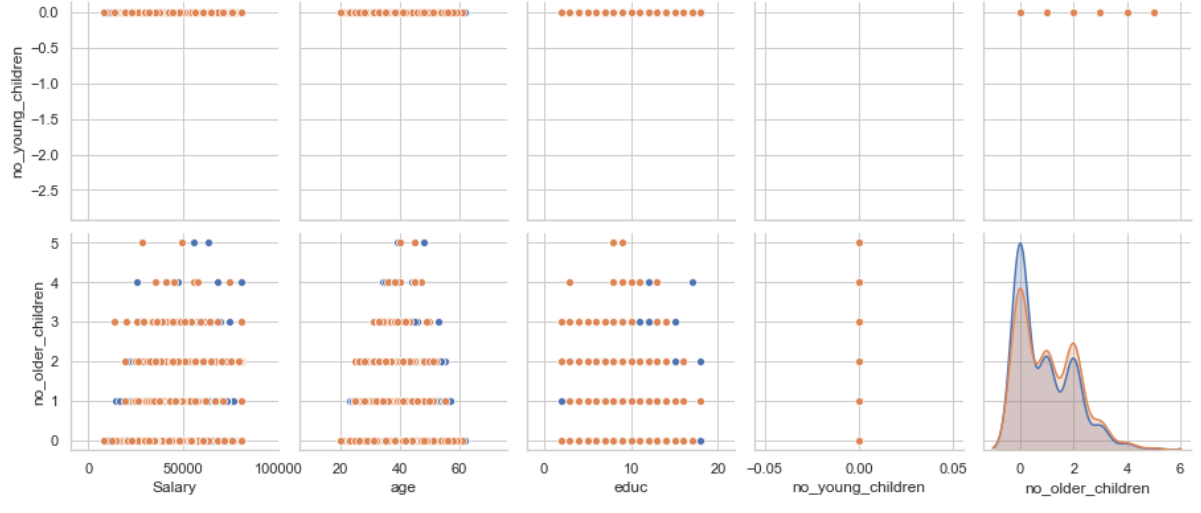
The frequency for opting ot not opting holiday package is appearing to be smilar with employees having older children.

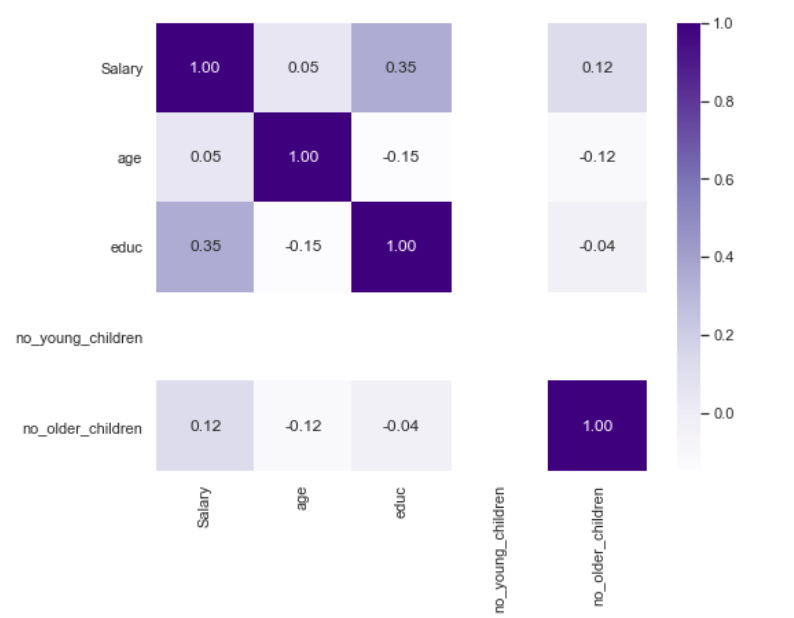
**Visualising After treating outliers:-**





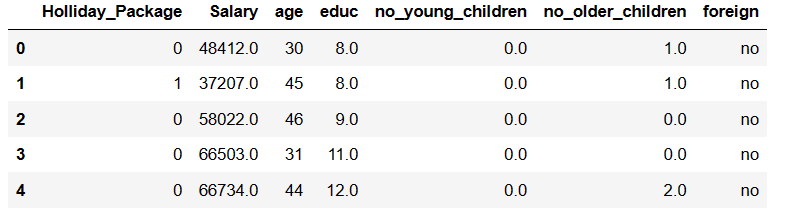


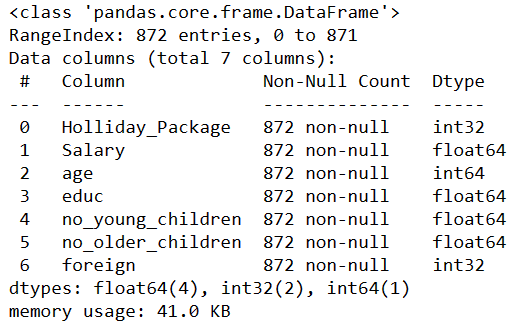




From the pairplot and heatmap we could say that there isn't any strong correlation between any variables. Salary and education display moderate corelation and no\_older\_children is somewhat correlated with salary variable. However, there are no strong correlation in the data set. The target variable Holiday\_Package has a very weak correlation with independent variables.

The categorical variables 'Holliday Package' and ‘Foreign‘ were converted into numeric by using the LabelEncoder functionality inside sklearn.





**2.2 Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis).**

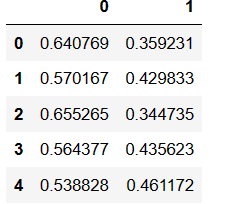
All the predictor variables were copied into X data frame and target variable was copied into the y data frame. X and y were split into training and test set in 70:30 ratio.

Y train value counts is

y\_test.value\_counts is

Now we will build the **LOGISTIC REGRESSION MODEL.**

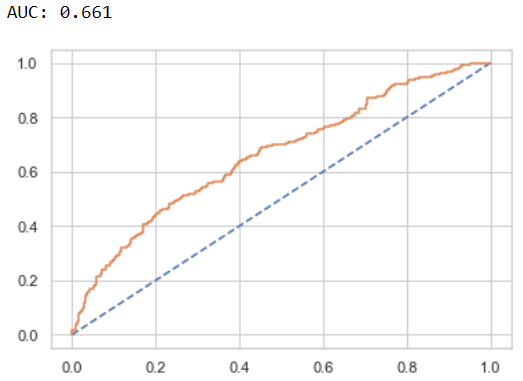
After fitting the Logistic Regression model and predicting on Training and Test dataset we get the Predicted Classes and Probs.



Model Evaluation-

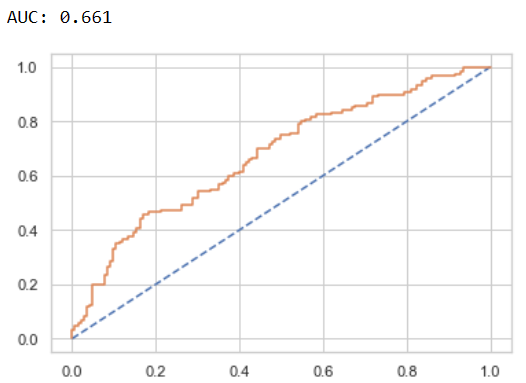
We now check for Accuracy in Training Data. 

Next step is to calculate **Area under curve** and **roc curve** for training date



We now check for Accuracy in Test Data.

Next step is to calculate **Area under curve** and **roc curve** for test date .



**Linear Discriminant Analysis:**

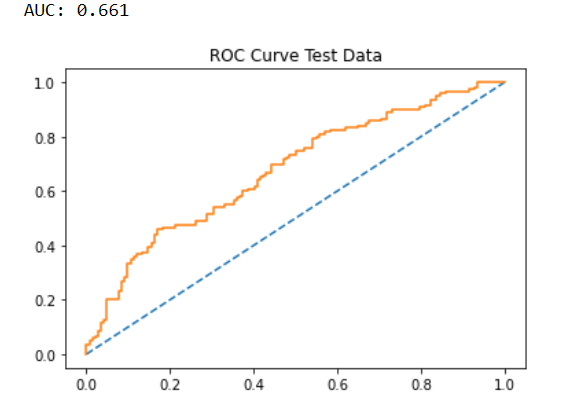
Data was split into Train (70%) - Test (30%) and fit the data in the Linear Discriminant Analysis model into the train data and then try to predict the outcome of using the test data. After that comparison is done between the actual and predicted values to calculate the accuracy of the model.

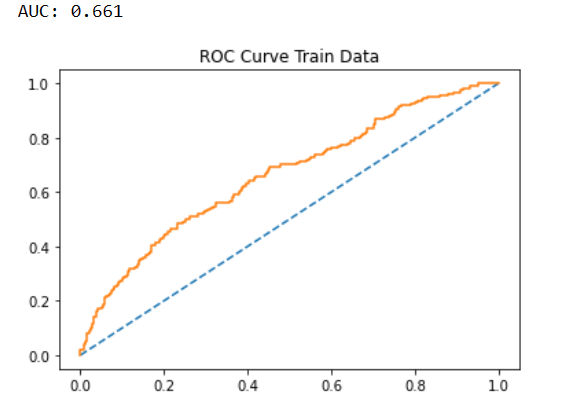


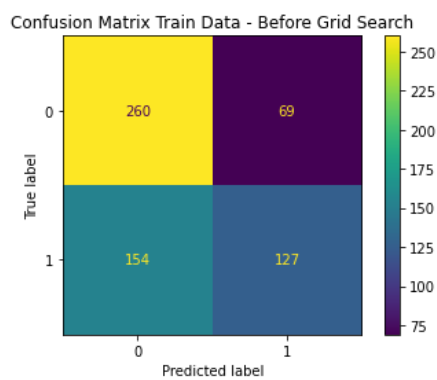
2.3-Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.

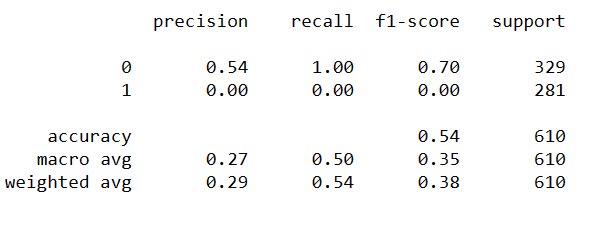
Now we will compare the performance matrix for both the models

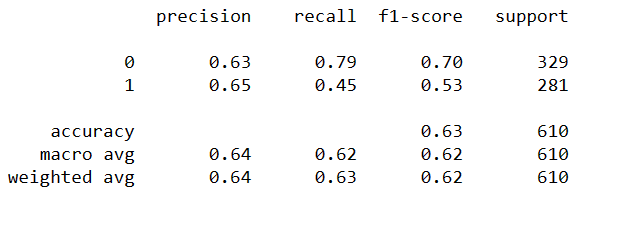
**LOGISTIC REGRESSION BEFORE GRID SEARCH**-



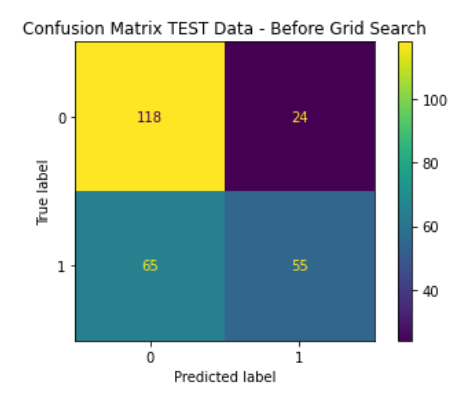
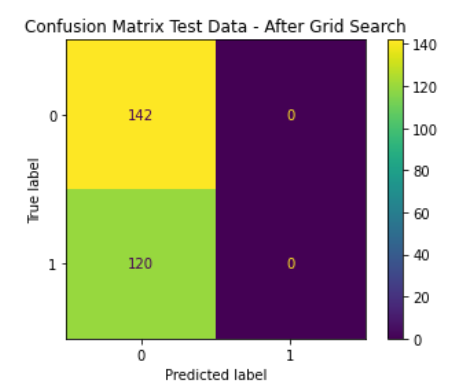


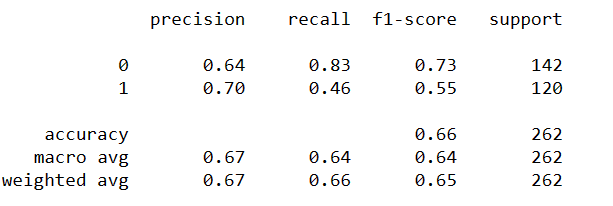
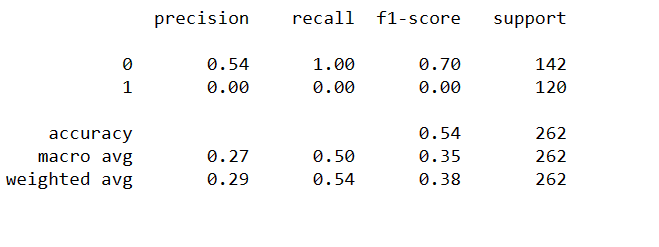






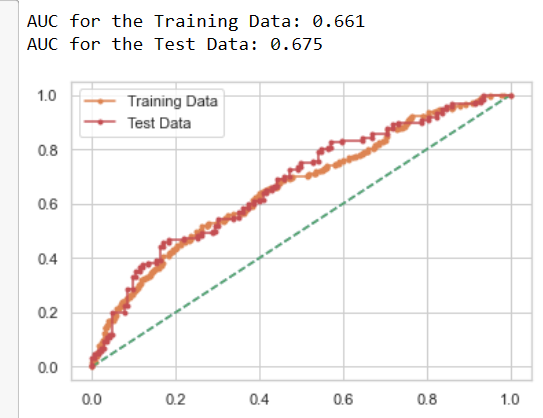
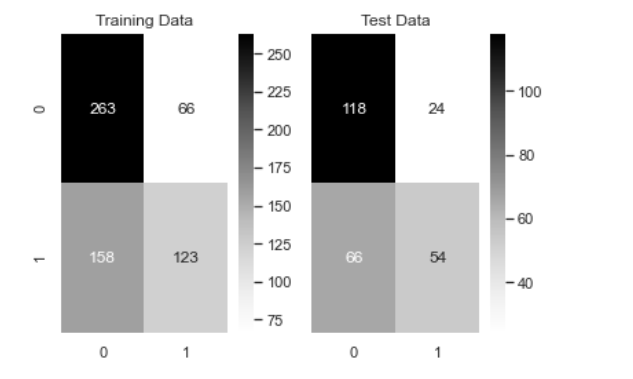
After scaling the training data is of not much impact than the unscaled one.

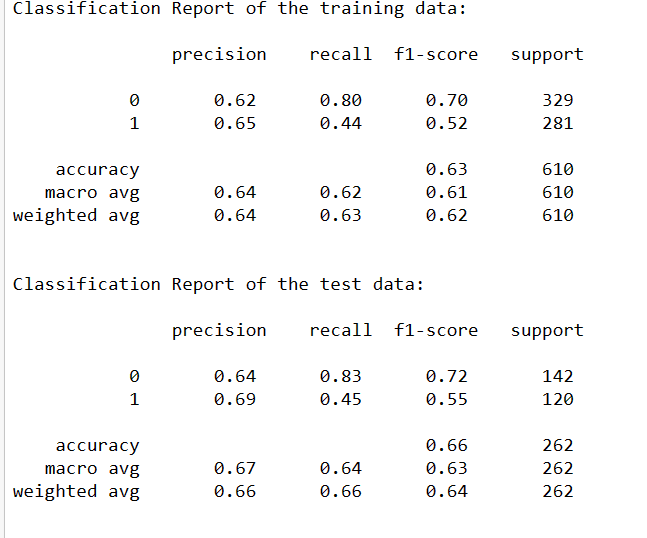




Similarly after scaling the testing not much impact on the model is seen than the unscaled one.

**Linear Discriminant Analysis:**





Accuracy of the model is very low still the LDA model is consistent with both train and test data. So it could be a feasible model.

Assumptions of multivariate normality and equal variance-covariance matrices across groups are required before proceeding with LDA, but such assumptions are not required for LR and hence LR is considered to be much more robust than LDA. But for this case study both LDA and Logistic regression before Grid Search proves to be efficient in predicting the target variable. Grid search CV in Logistic regression failed to provide a better efficient model rather it gave a futile outcome.

The results of Logistic Regression (before grid search) and LDA are compared here. The accuracy precision and recall are all very similar for both LDA and Logistic Regression in both train and test data. So one can choose to go ahead with either of the model.

**2.4 Inference: Basis on these predictions, what are the insights and recommendations.**

Inference-

1) There is no plausible effect of salary, age, and education on the prediction for Holliday\_packages. These variables don't seem to impact the decision to opt for holiday packages as we couldn't establish a strong relation of these variables with the target variable

2) Foreign has emerged as a strong predictor with a positive coefficient value. The log likelihood or likelihood of a foreigner opting for a holiday package is high.

3) no\_young\_children variable is negating the probability for opting for holiday packages, especially for couple with number of young children .

The company can try to bin salary ranges to see if they can derive some more meaningful interpretations out of that variable. May be club the salary or age in different buckets and see if there is some plausible impact on the predictor variable..

Recommendation:

1) The company should really focus on foreigners to drive the sales of their holiday packages as that's where the majority of conversions are going to come in.

2) The company can try to direct their marketing efforts or offers toward foreigners for a better conversion opting for holiday packages

3) The company should also stay away from targeting parents with younger children. The chances of selling to parents with 2 younger children is probably the lowest. This also gels with the fact that parents try and avoid visiting with younger children.

4) If the firm wants to target parents with older children, that still might end up giving favourable return for their marketing efforts then spent on couples with younger children.