**Project 1**

Problem summary: To build a Linear Regression model to predict the price of the cubic zirconia and to highlight the factors that are important:

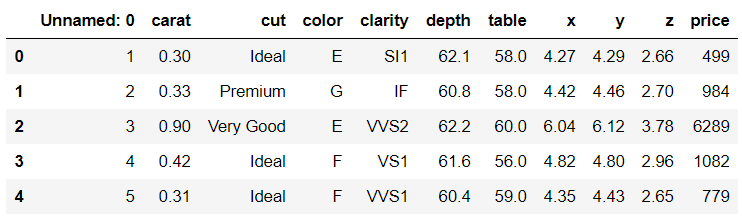
1.1:

The data file: cubic\_zirconia has 26967 entries. The shape of the data looks like this:

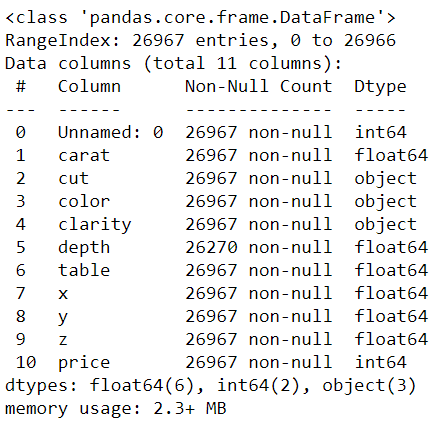


The column names are: Unnamed: 0, carat, cut, color, clarity, depth, table, x, y, z, price.

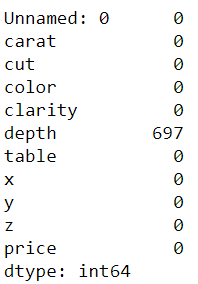
The head of the data looks like this:



We have 3 object data types: cut, color and clarity, 2 integer data types: Unnamed: 0 and price, 6 float data types: carat, depth, table, x, y, z.



There are 697 null entries in the depth column.



We then check for unique values in the columns with the object data type:

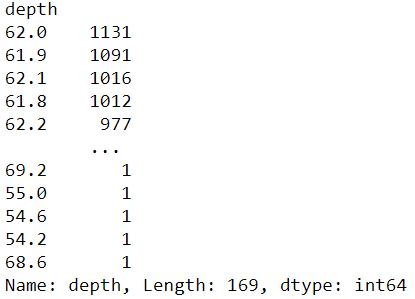


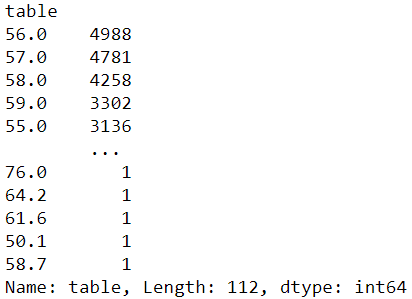
It is to note that the category: FL (Flawless) is not available in the dataset. This has however been mentioned in the data dictionary though.

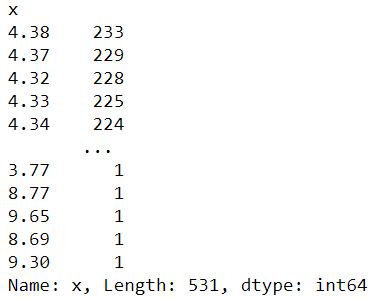
We also look at the value counts to see the presence of any unwanted inclusions in the data:

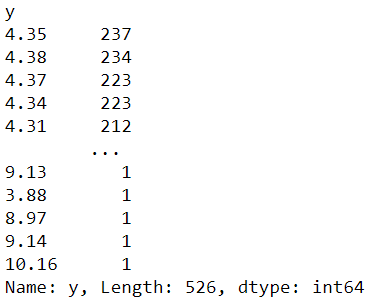
Note: This would also be detected in the data type check and in the describe function check. Nevertheless, we will still go ahead and check the value counts of the numerical variables:

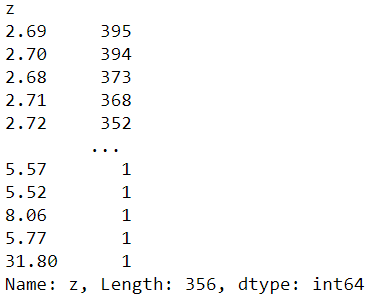


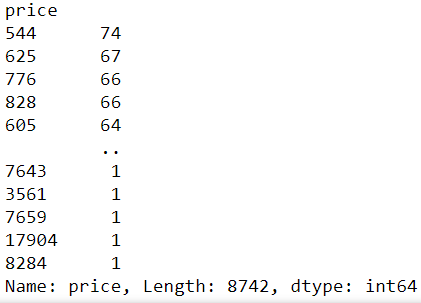




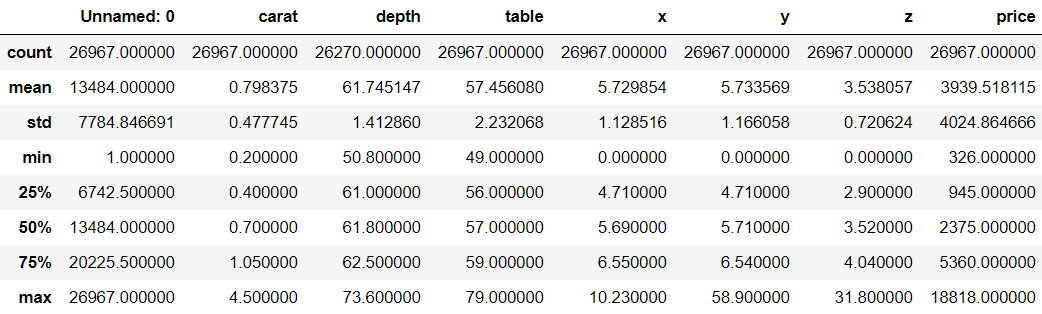








The describe function of the data gives the following output:

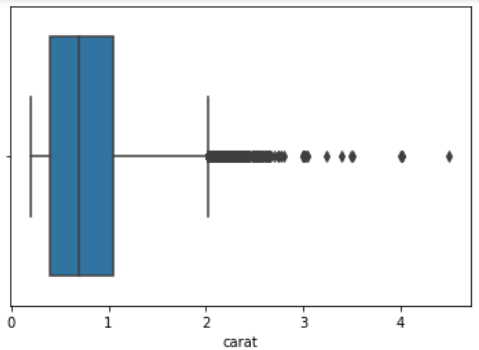


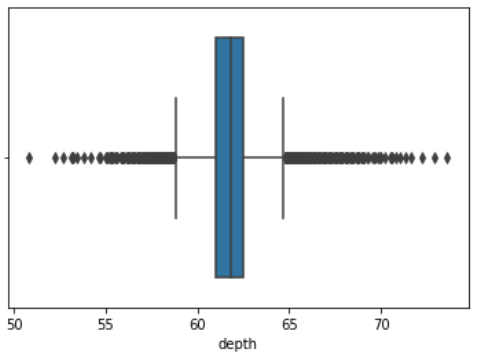
It can be seen that there is slight that there is slight skewness present in two of the variables: price and x.

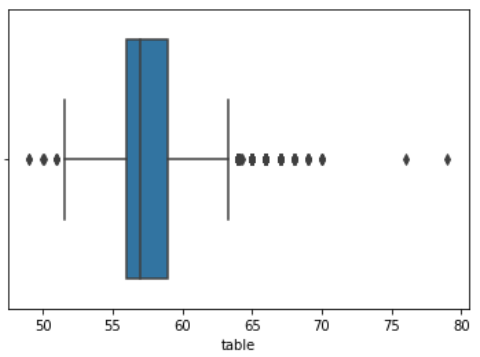
We then drop the column: Unnamed: 0 from our further analysis.

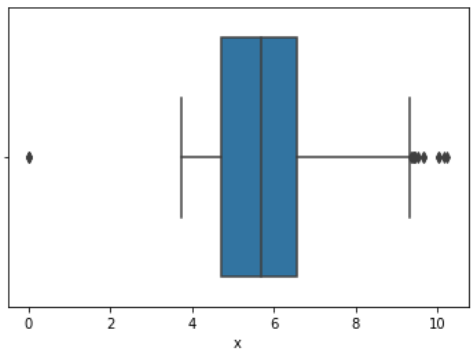
**Univariate Analysis:**

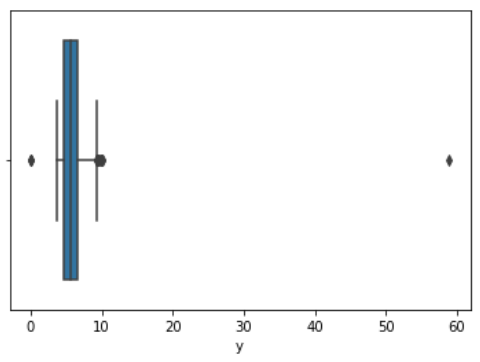
We create a for-loop to show the boxplots of the continuous variables: carat, depth, table, x, y, z, price:

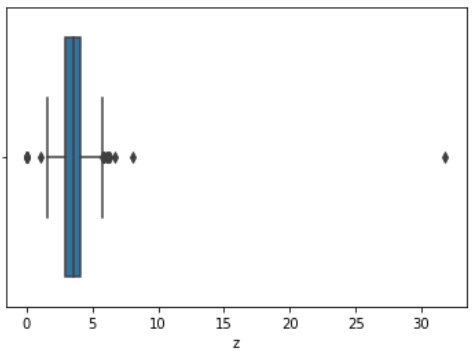


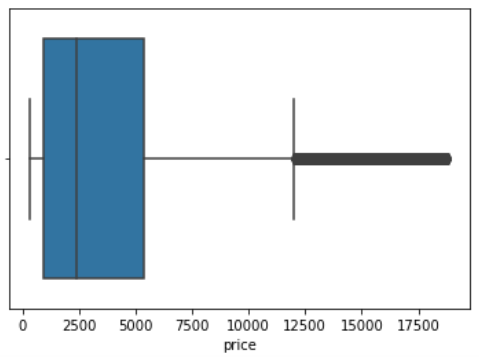








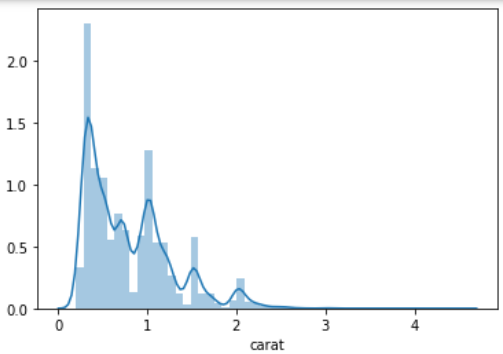


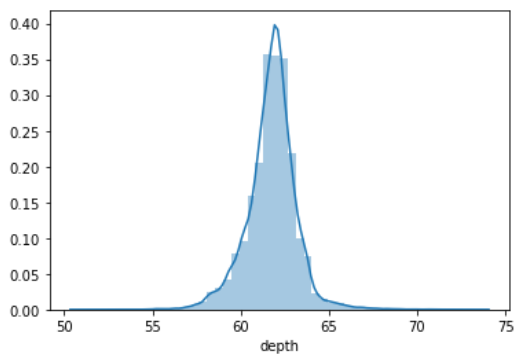


It can be seen that outliers exist in all the continuous variables in either side of the whiskers of the boxplot.

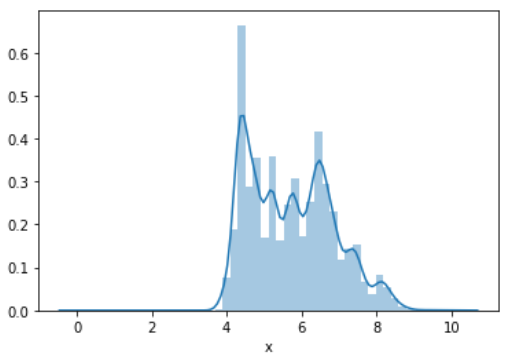
Only for the price variable, the outliers are present towards the upward side.

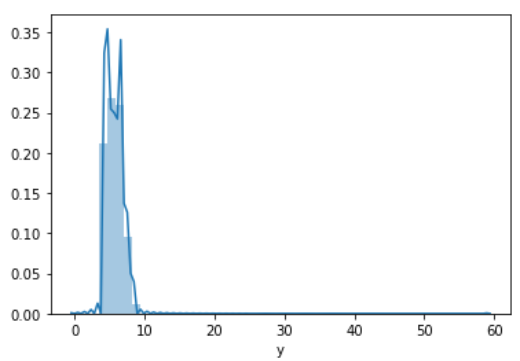
We also plotted histograms for the continuous variables:

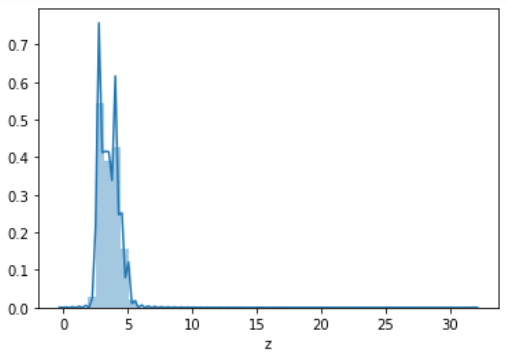


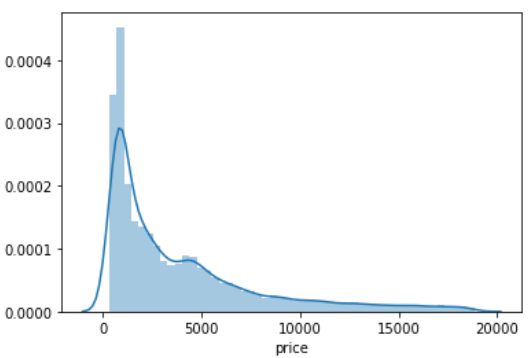












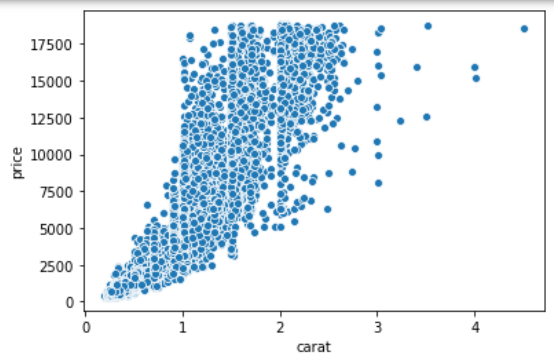
Only the depth variable shows a normal distribution.

The variables y, z, price and carat are highly right skewed.

There are four peaks observed in the carat distribution with each diminishing in size as compared to the previous one.

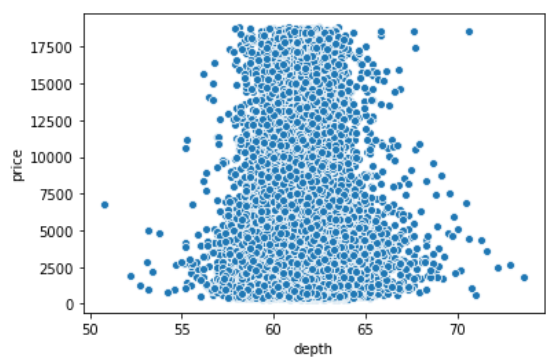
**Bivariate Analysis:**

Scatterplot of carat and price:



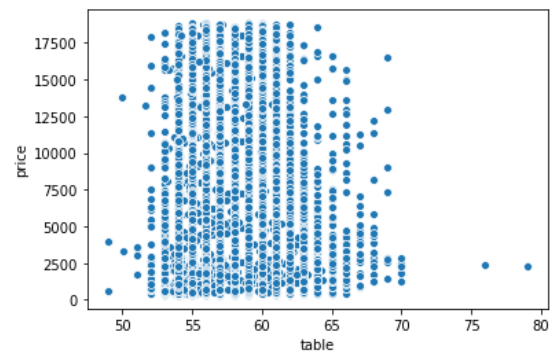
The relationship between these two variables is highly linear in nature with positive linearity.

Scatterplot of depth and price:

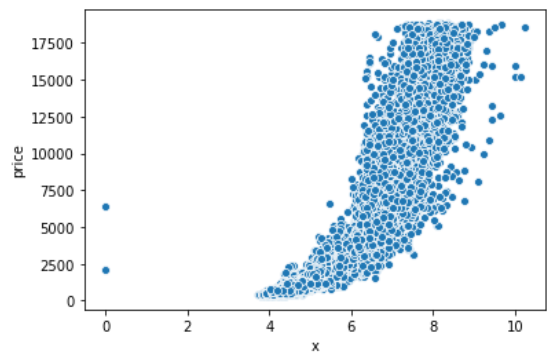


These two variables are not really showing linear correlation.

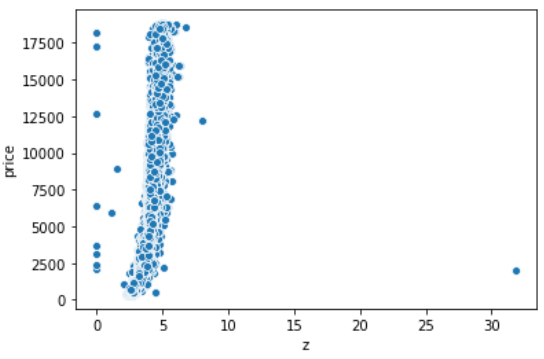
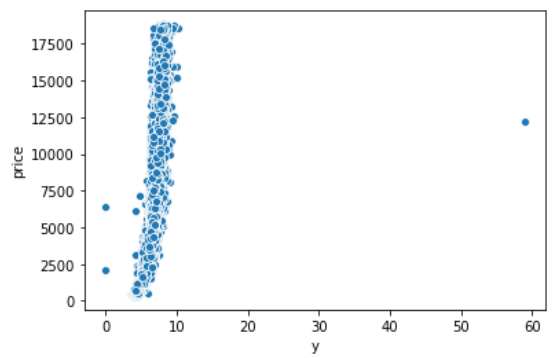
Scatterplot between table and price:



The relationship is non-linear.

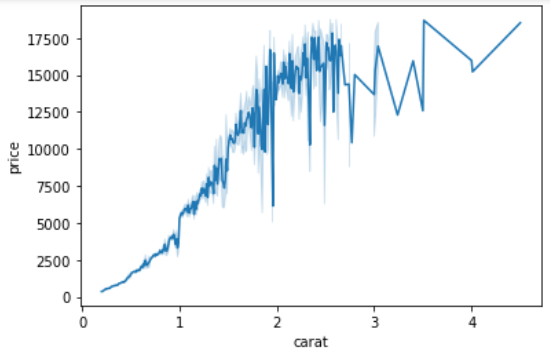


The relationship between x and the target variable is perfectly linear.



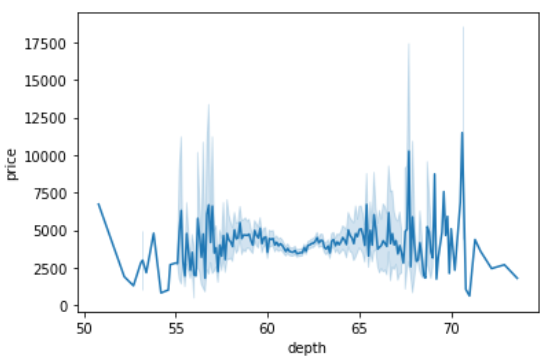
The variables y and z have a positive relationship with the target variable.

Lineplot of the continuous variables:



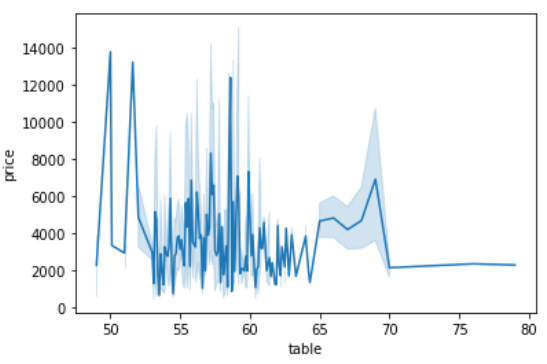
The relationship is disturbed for higher values of carat and price: For carat values of more than 2.75, the price fluctuates.

For lower values of carat and price, the relationship is linear.

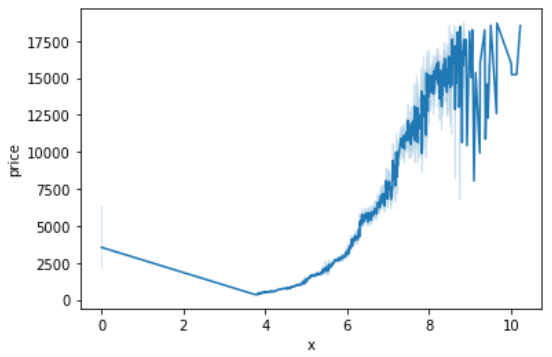


There is distortion in the price between the depth values of 53.5 to 57 and from 66 to 72.

The price is constant between for the depth values between 57 to 66.

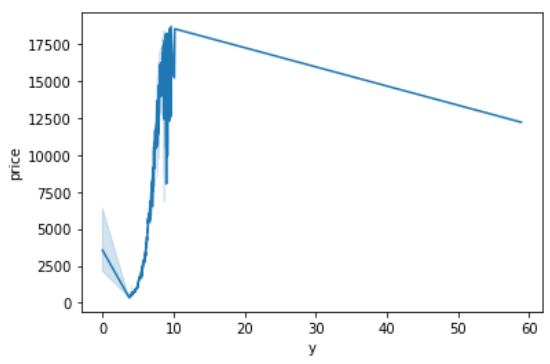


There is no interpretable relationship between the table and price variables.

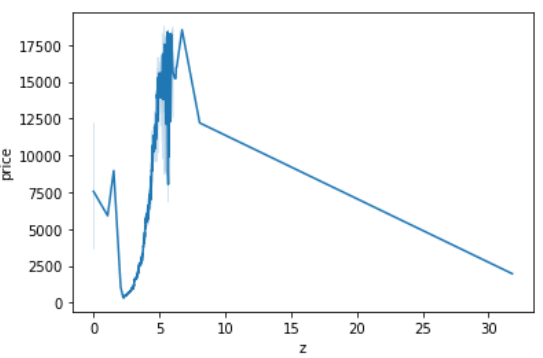


From the x value of 0 to 3.9 the price constantly decreases with an increase in the value of x.

Post this, the price increases from the value of x of 4.2 till 9. Beyond the value of 9 for x the values of price become uninterpretable.



There is a sharp increase in the price from y = 1 till y = 10. Beyond the value of y = 10.5, the price decreases steadily.



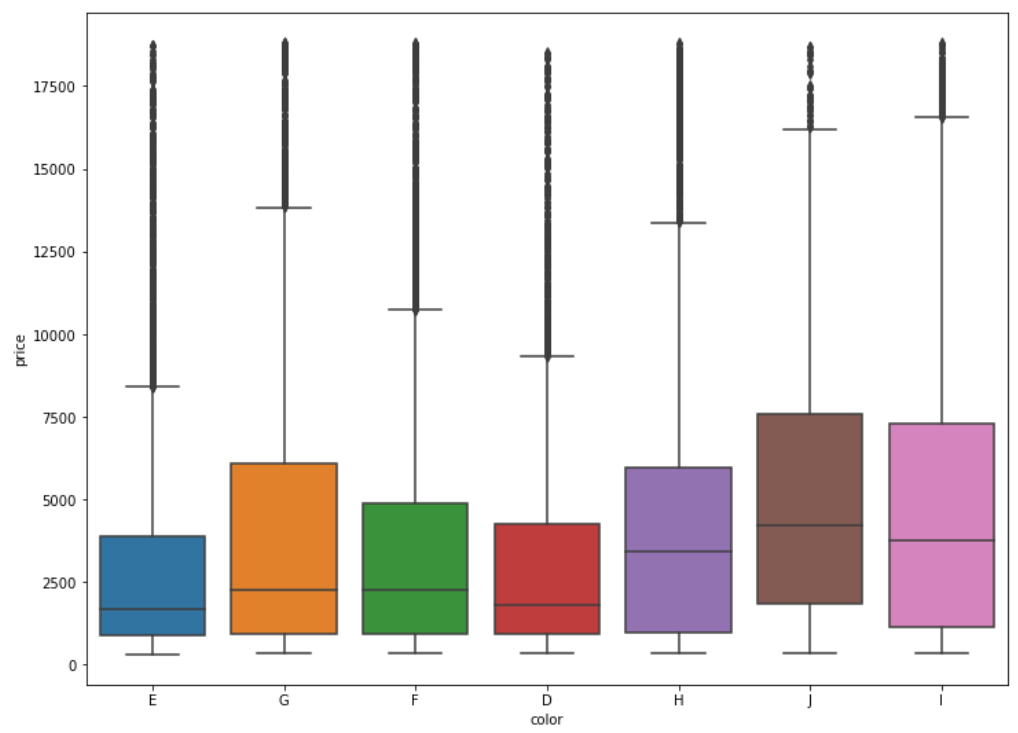
The behaviour of the variable z shows a similar distribution with respect to the target variable like that of the variable y.

**Categorical Boxplot between cut and price:**



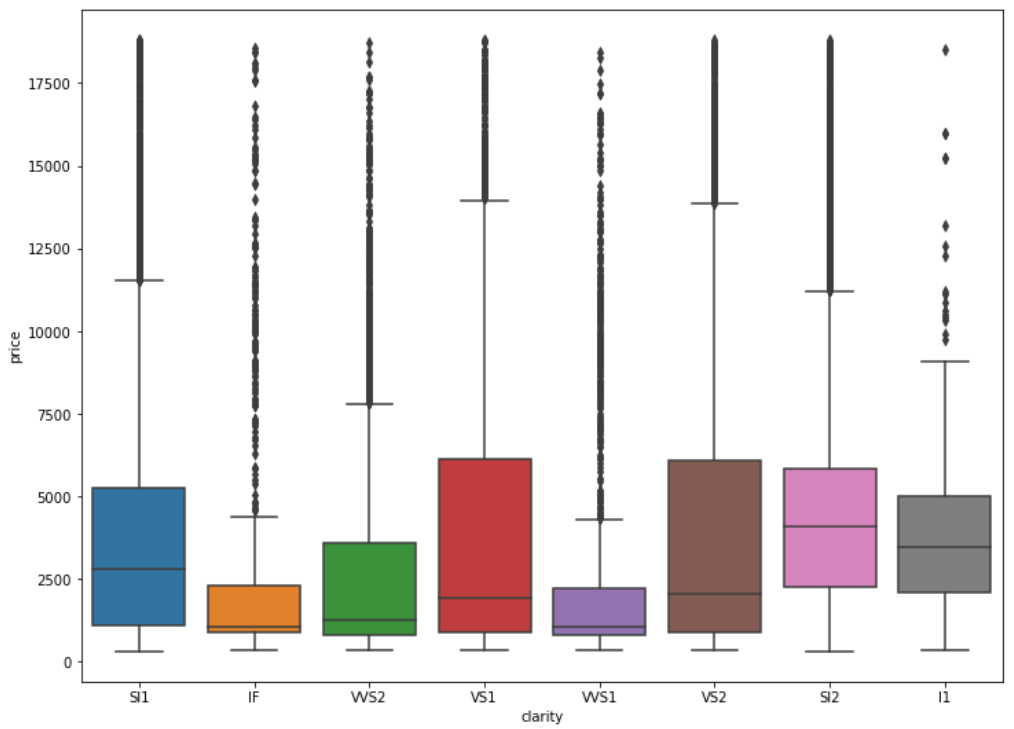
It is interesting to observe that the median price of the cut type: Fair is higher than that of the median price for the cut type: Premium. The median price for the ideal cut is the lowest. The size of the boxplot is the biggest for the cut type: Premium and lowest for the cut type: Fair.

Categorical boxplot between color and price:



The color J has the highest median price as compared to the premium colors like D & E which have the lowest median in the price. The size of the boxplot is the highest for the color J & I followed by G & H indicated that the volume of sales for the former two is the highest. The smallest boxplot available for the color category is E.

Categorical Boxplot between the clarity and the price:

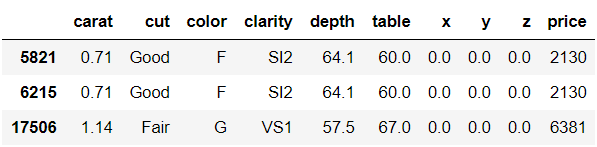


The median values of price for the best clarity: IF is the lowest when compared to the median values of price for the worst clarity I1. The most sold are VS1 and VS2 and the least sold are IF ad VVS1 clarity types.

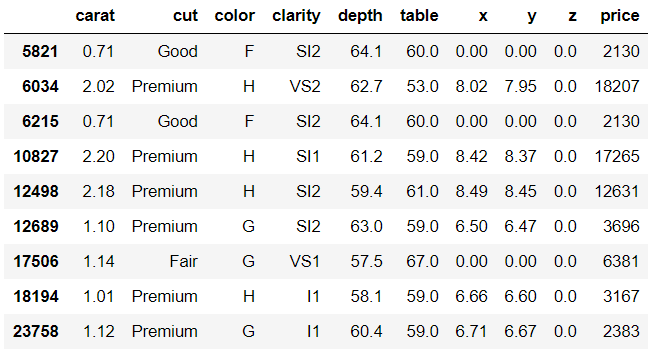
1.2.) The 697 null entries in the depth column are imputed by the median.

We then check that the variables: x, y and z have values equal to 0.

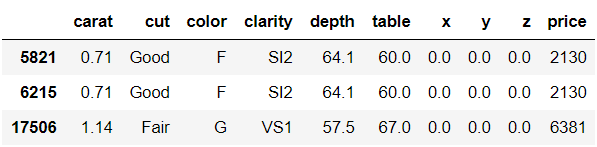
Where x and y is equal to 0:



Where only z is equal to 0:



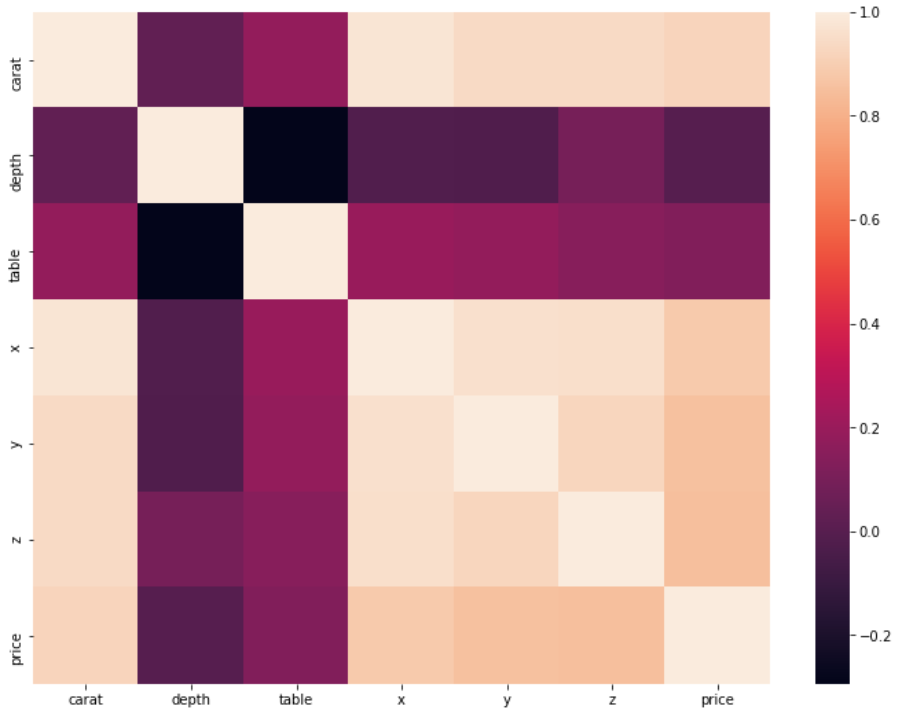
This is practically not possible as they represent the length, width and height respectively. They cannot have 0 values at the same time. Following are the three rows where all the three values are zero and hence will be removed from our analysis.



It is also observed that there are 33 duplicate rows in the dataset and will be removed.

Scaling does not seem to be necessary at first as the variables x,y and z are all in mm. The variables depth, table are values derived out of a computation with the values x,y and z.

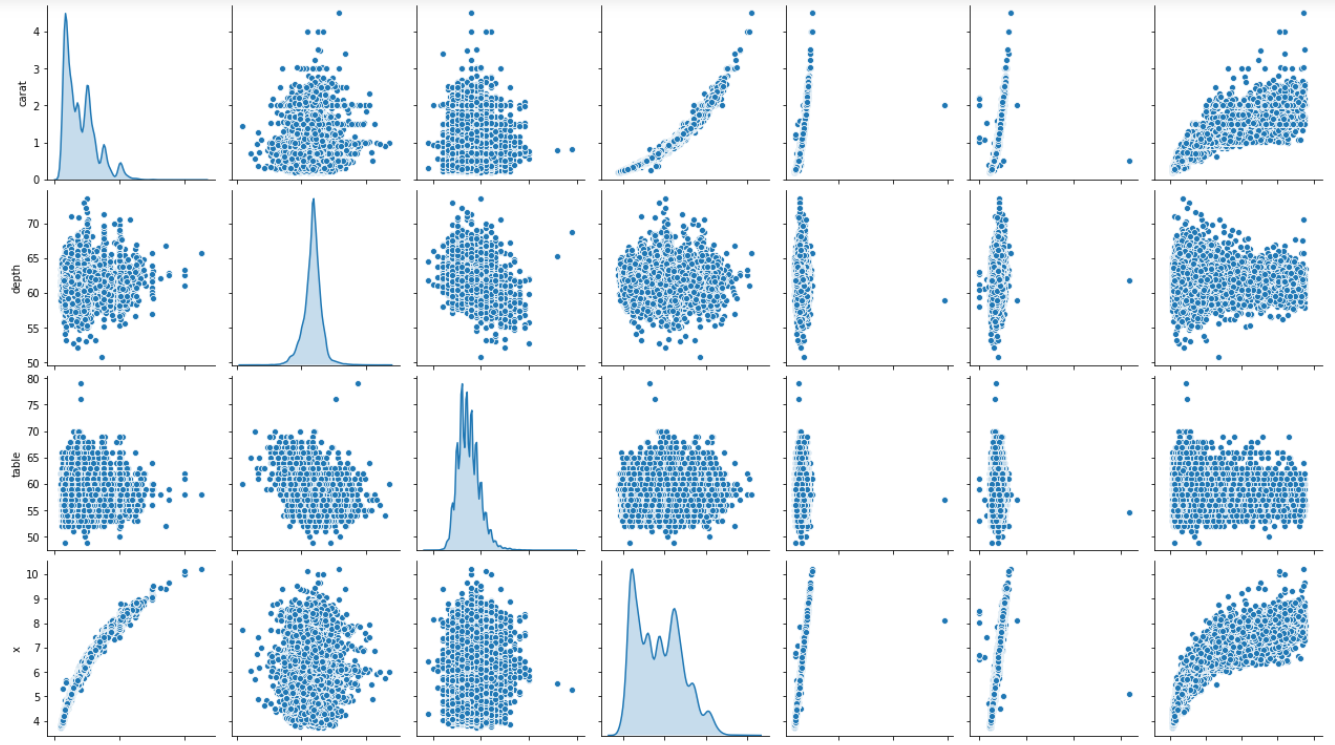
Heatmap of the variables:

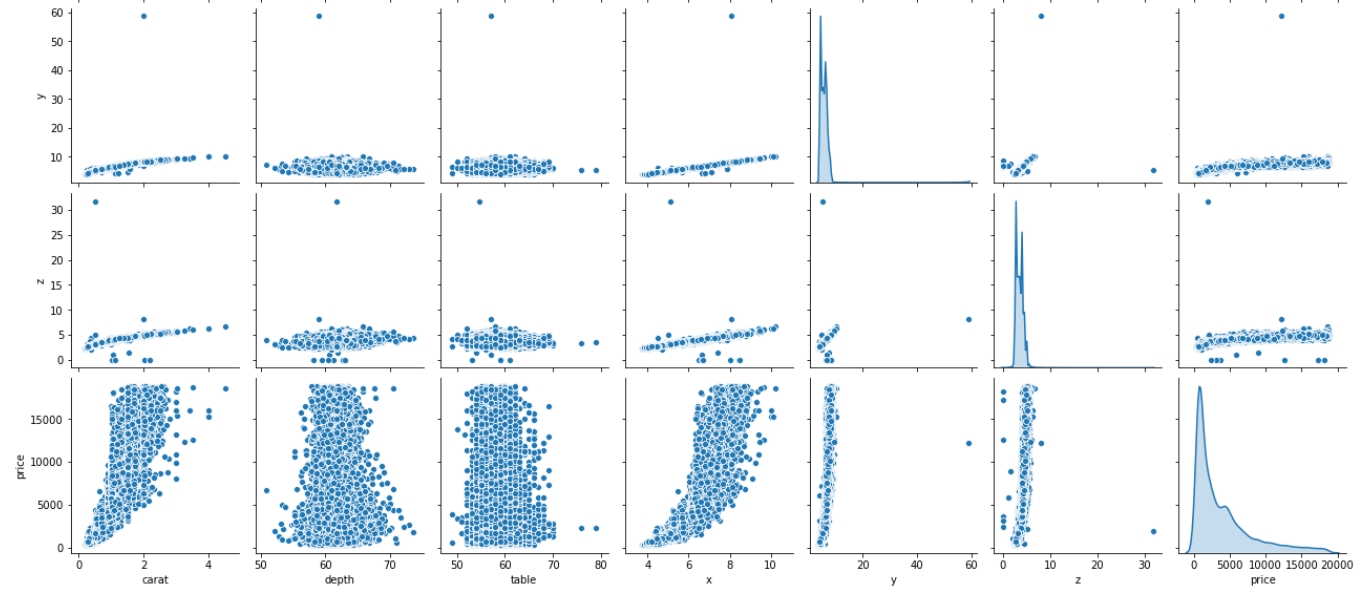


There is structural multicollinearity between the variables.

There is very high positive collinearity and negative collinearity between the variables. The variables x, y, z are highly correlated with each other and also with the carat variable. The depth is negatively correlated with the carat variable and the table variable.

Pairplot:





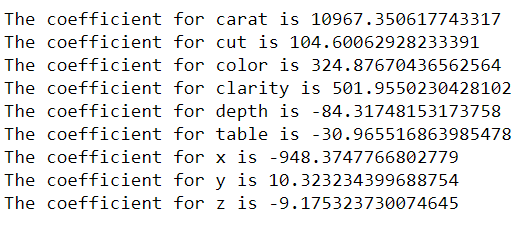
1.3.)

We then encode the categorical data ordinally based on the superiority levels.

We then split the data into X and Y with all the independent variables in the X and the dependent variable: Price in the Y. The ratio for splitting the data is 70% train and 30% test.

We then fit the linear regression model into the train set.

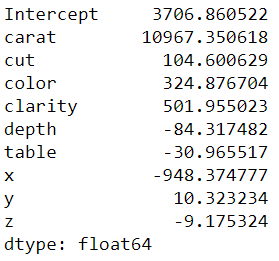
The regression model coefficients are as follows:

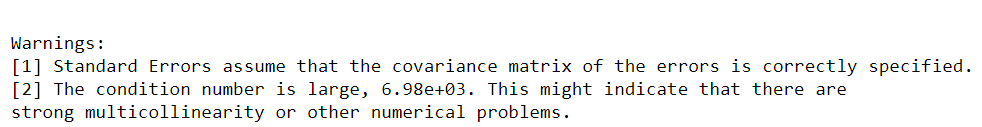




The regression model score for the train set is 90.77 and for the test set is 91.01.

We then use the Stats model approach to fit the linear regression model to get R type output.



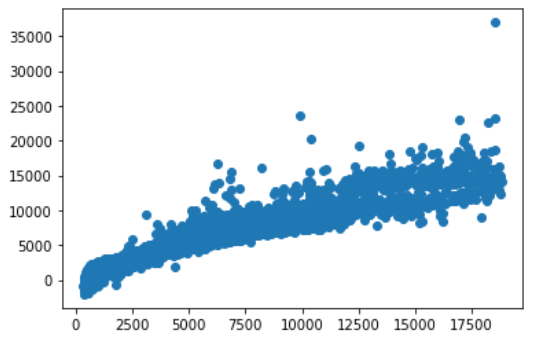
 

The adjusted R squared is 0.908 which is good.

The p\_values for the variables y and z are higher than our cut-off of 0.05 and hence lose their importance in this variation of the model.

The Root Mean Squared Error is 1218.85

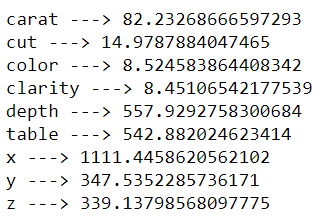
We then plot a scatterplot between the y predicted and the actual y for the test data.



There is a linear correlation between the two.



We then check for the variance inflation factor:

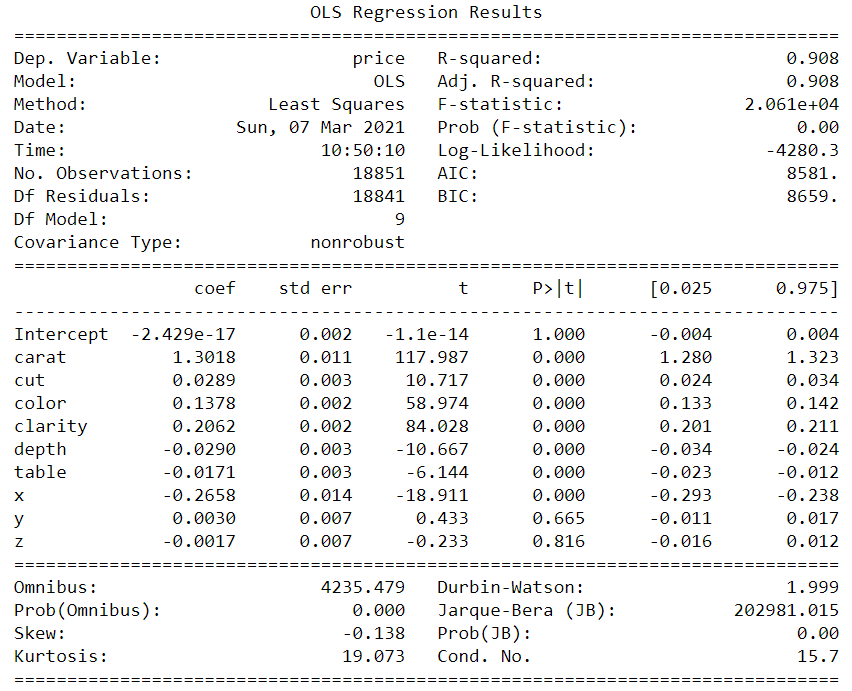


The VIF values are very large indicated very high multicollinearity between the variables. The variables color and clarity have the least VIF values but even those are above 8 which is high.

It can be observed that the intercept value is too large = 3706.86 and this needs to be eliminated for accurate interpretation of our results.

Hence we scale the data using the z score scaling and apply the linear regression model with the following results:

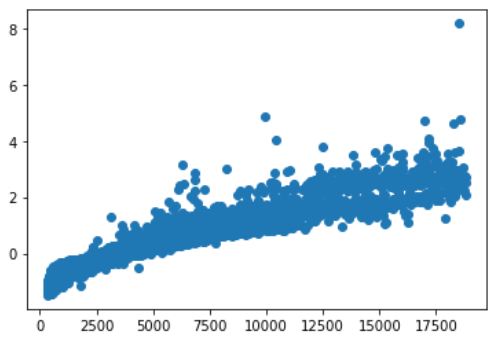




The adjusted R squared is the same as 0.908

The Root Mean squared error is 0.3036

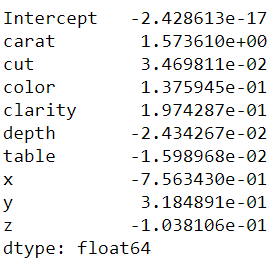
A scatterplot of the model for the predicted y test and the actual y test is:

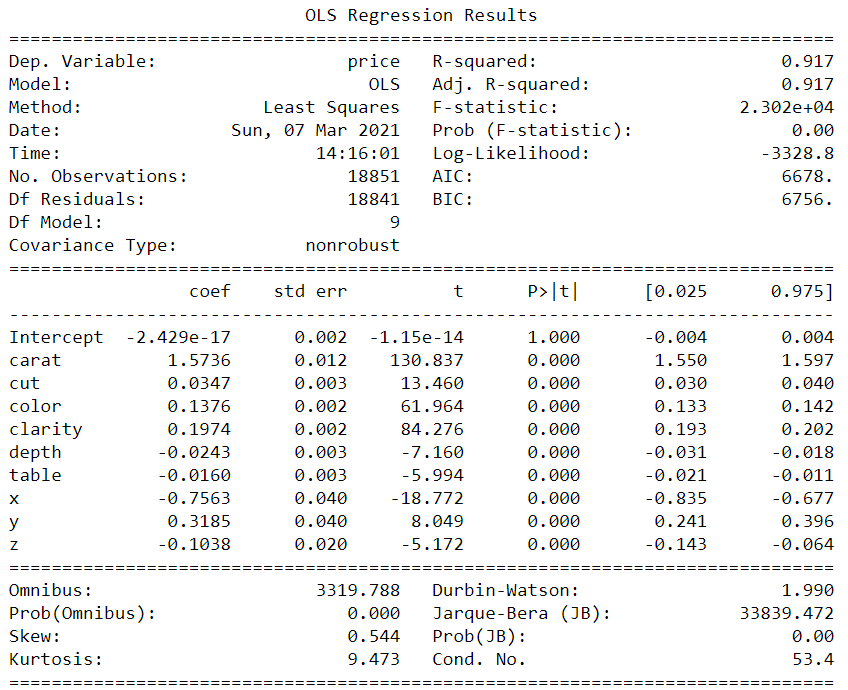




The intercept term has been removed and it can be seen that for every one unit increase in the carat the price increases by 1.3 keeping all the other variables constant. It is to note that the variables are scaled and hence the coefficient values should be interpreted as a relative term and not as an absolute value.

**Building the model after treating the outliers and scaling:**



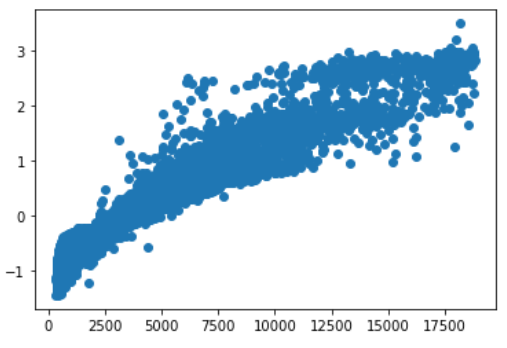


The Adjusted R squared value is now 0.91

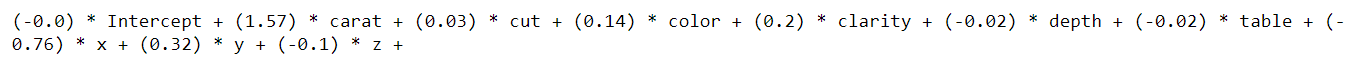
The p\_values of the all the variables have now become important.

The Root Mean Squared Error is 0.288

A scatterplot between the predicted y and the actual y for the test data is as follows:



The linearity has increased.

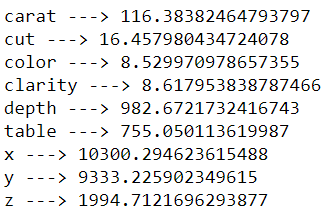


The weight of the carat variable has increased and goes to say that for every one unit increase in carat the price would go up by 1.57 keeping all the other factors constant.

The five most important variables are:

Carat, Length of the cubic zirconia in mm., Width of the cubic zirconia in mm., color and clarity.

The Variance Inflation Factor remains to be high as observed previously.

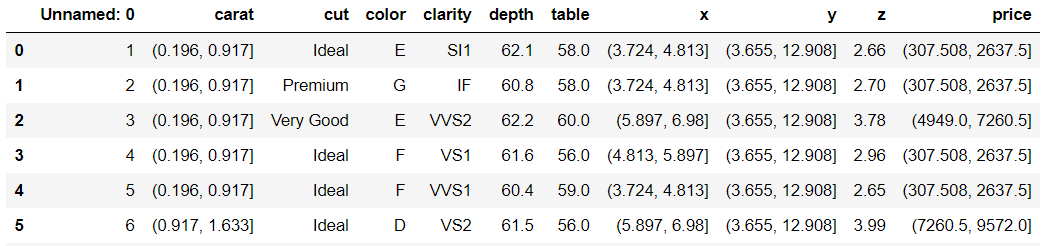


**Business Insights and recommendations:**

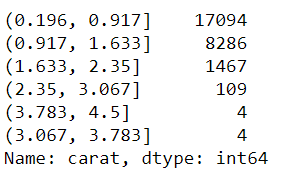
The important variables: carat, x and y have been binned into 6 ranges using the cut function in python.

This has been done to identify segments in the customer and provide insights and recommendations accordingly.

The dataframe looks like this now:

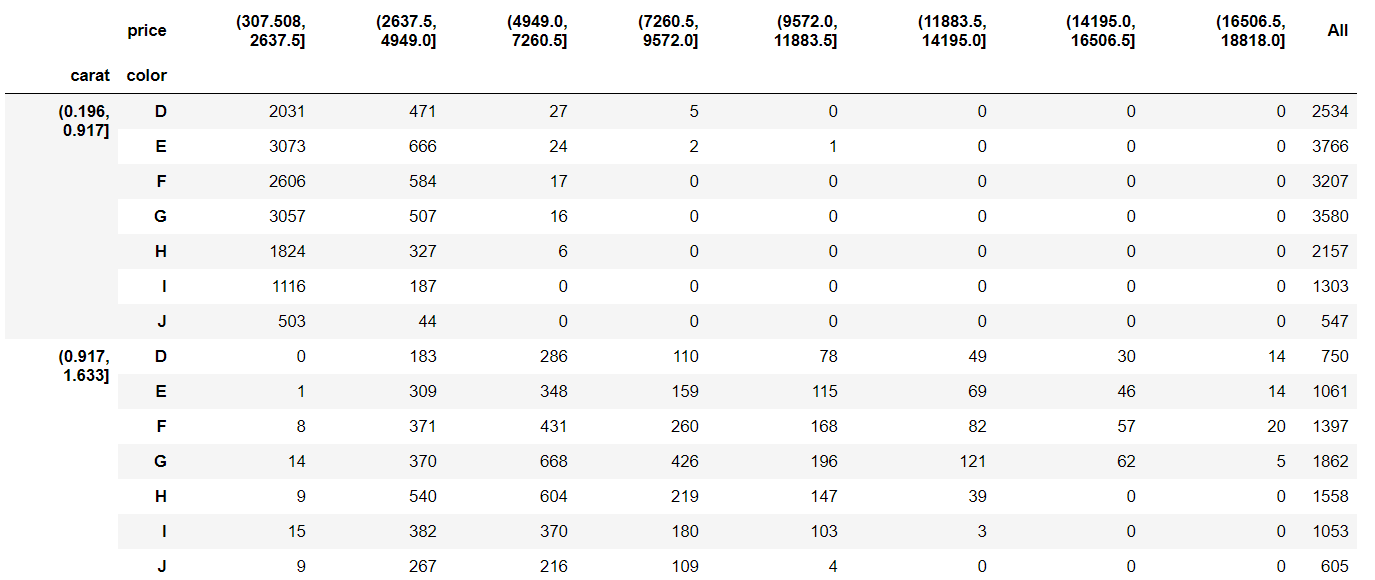


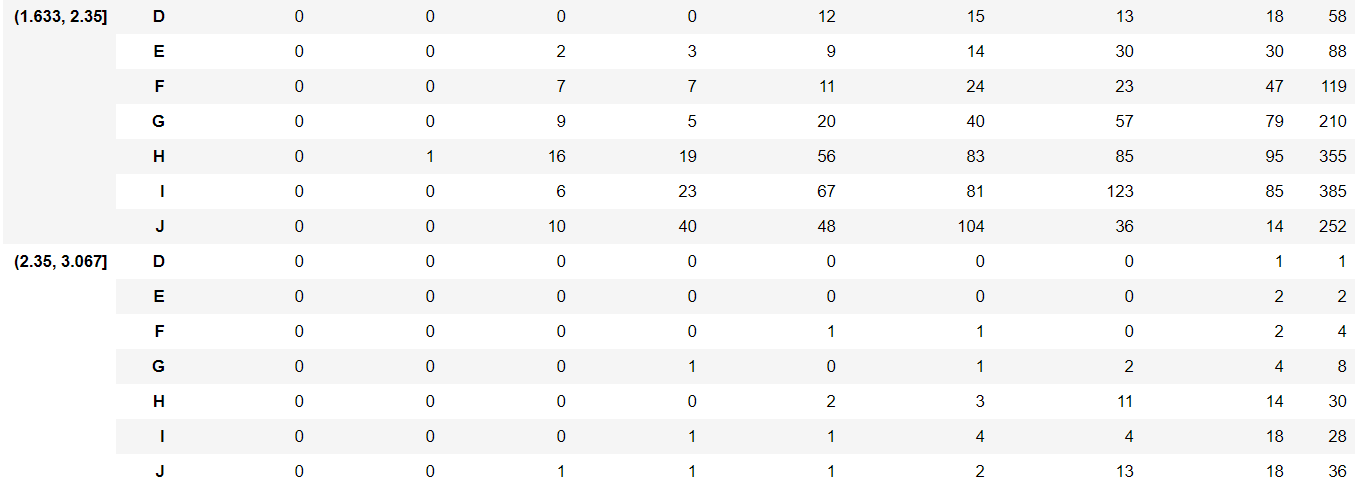
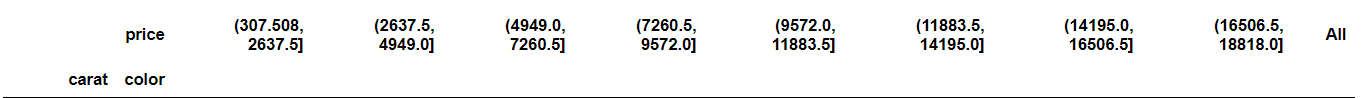
Since carat is the most important variable in determining the price of the cubic zirconia, we look into the different bins of the carat variable:



It can be seen that the maximum number of purchases have taken place in the carat range: 0.196 – 0.917 and this is followed by the high number of purchases in the carat range: 0.917 – 1.633.

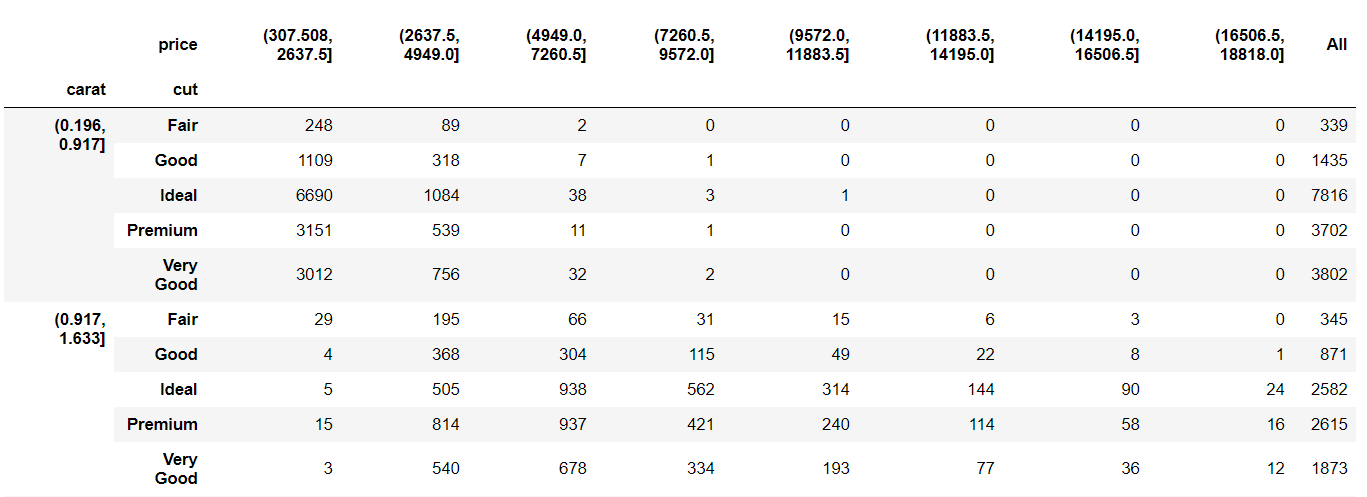
Following is the crosstab of the different carat ranges, the color and the different price ranges:

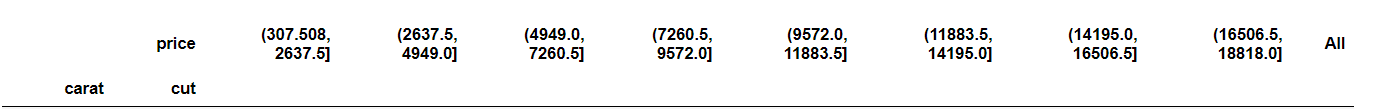
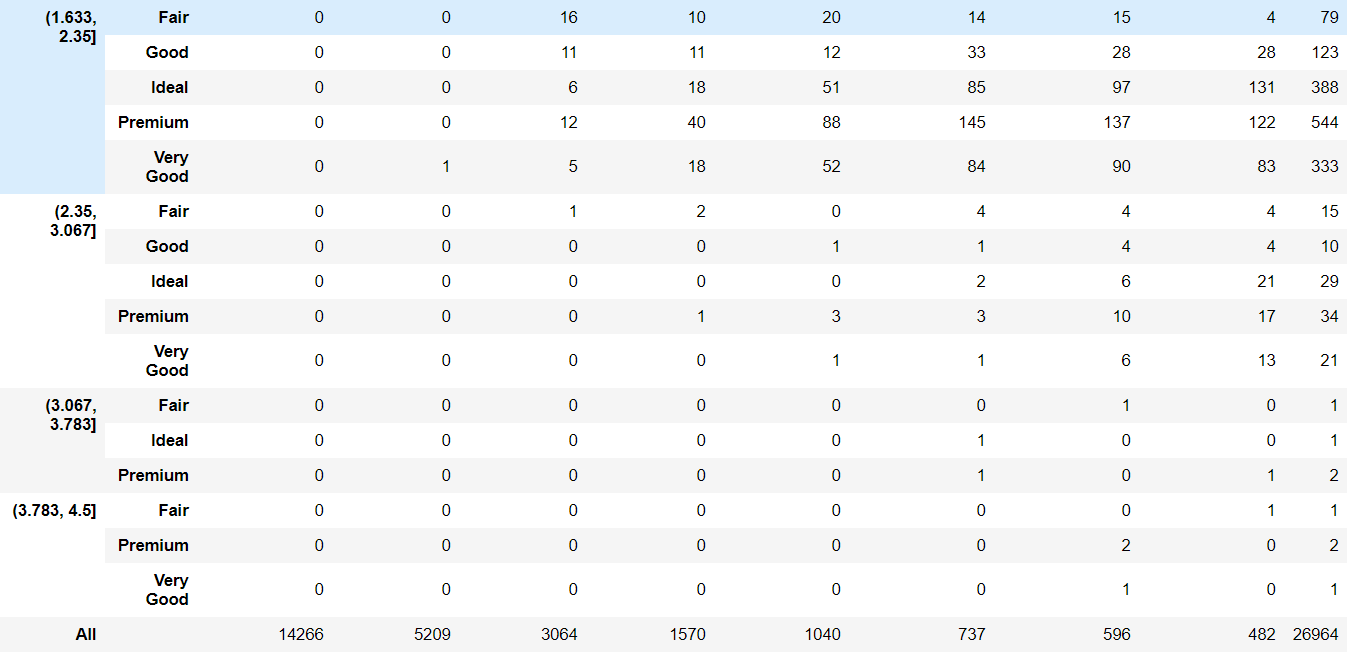




* For the carat range 0.196 – 0.917 there are many stones in the colors D,E,F and G and for the price ranges 307.508 – 4949.0. Colors E,G and F are the top three colors in this price and carat range.
* For the higher carat range: 0.917 – 1.633, the price range shifts slightly from 2637.5 to 9572.0 The stone distribution basis the different colors appears to be more even in this range.
* It is in the carat range of 1.633 – 2.35 is where the high price stones are present. The most number of stones are present in the color I.
* For the carat range 2.35 – 3.067, fewer stones are present but the ones that are present are found in the highest price ranges of 14195 – 16506.5 and 16506.5 – 18817.0 Colors J, H and I are the most favorable in this range.

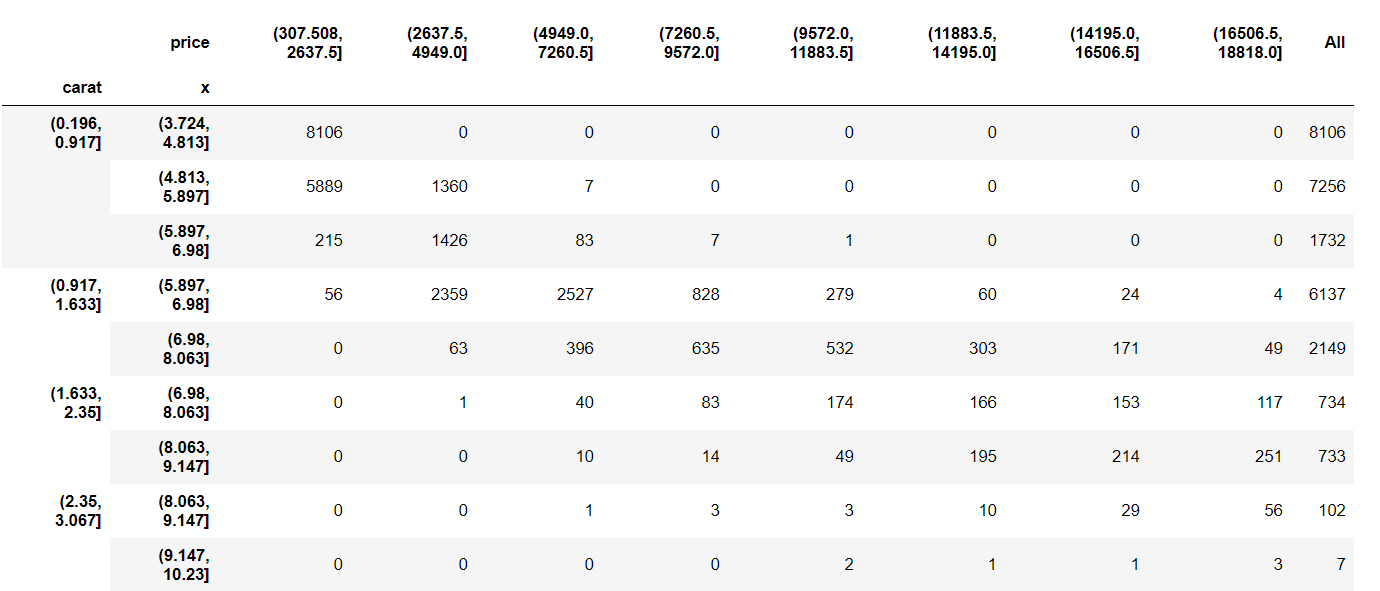
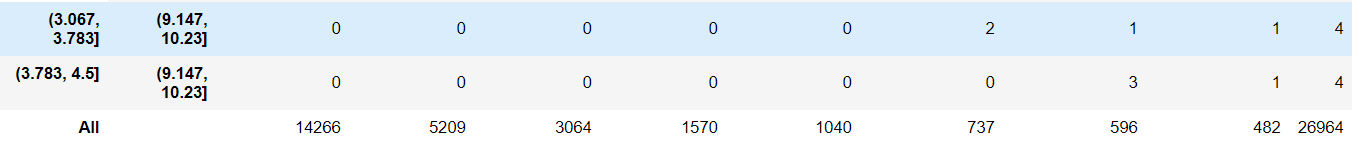
Following is the crosstab of the different carat ranges, the cut and the different price ranges:



* Similarly as for the color, the price ranges for the stone for different cuts are low for the lower carat range of 0.196 - 0.917
* Ideal, Premium and Very good are the different cut variants that are high in numbers in the high price ranges from 11883.5 – 18818.0
* For the carat range 2.35 – 3.067 there are 27 stones with the highest price range of 16506.5 – 18818.0 and they are of the cut Ideal.

Following is the crosstab of the different carat ranges, the x: Length of the cubic zirconia in mm and the different price ranges:

* Lot of stones are present in the price range 307.508 – 2637.5, carat range: 0.196 – 0.917 and for the length range of 3.724 - 4.813 and 4.813 – 5.897.
* High price ranges: 14195.0 – 18818.0 have a high number of stones present in the carat range: 1.633 – 2.35 and for the length range 6.98 – 9.147.
* Scaling becomes essential to eliminate the intercept term.
* After the outlier treatment, the adjusted R square is 0.91 and the Root Mean Square Error for the train data is 0.288 and for the test data it is 0.2872. It can be seen that the model is a right fit model and hence can be termed as a good model.
* The p- value for the intercept term is 1 and hence shows us that the intercept is not to be included. But since the value of the intercept term is close to 0 we will neglect the same.
* The p – values for all the other variables have become important after the outlier treatment.

**Project 2: Logistic Regression and LDA**

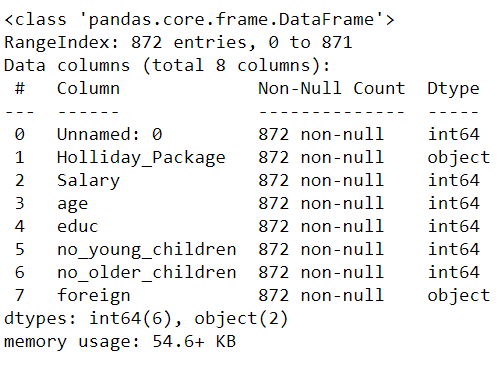
Problem summary: To build a logistic regression model and a linear discriminant analysis model to predict whether an employee will opt for the Holiday Package or not.

1.1)

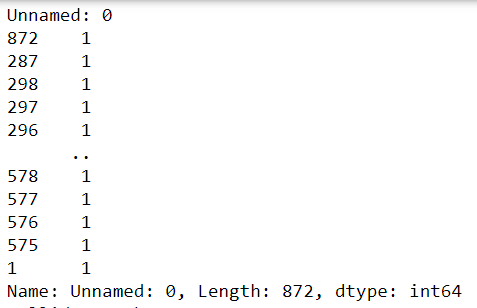
The data file: Holiday\_Package has 872 entries. The shape of the data looks like this:



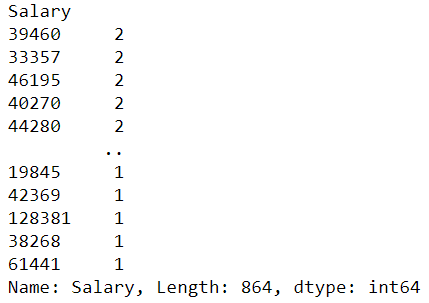
We look at the info of this dataset:

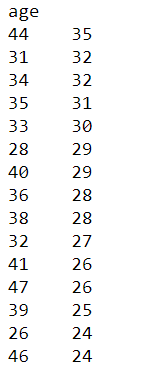


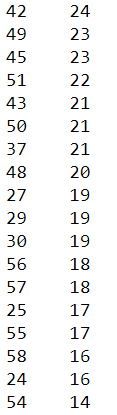
The value counts of the data frame looks like:

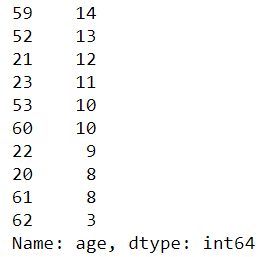


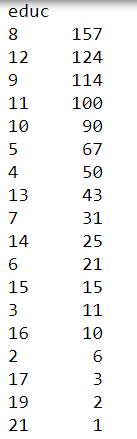


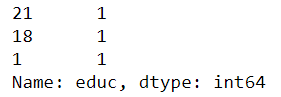


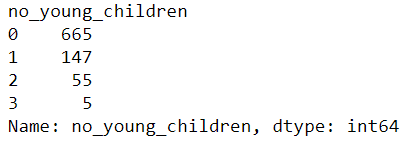


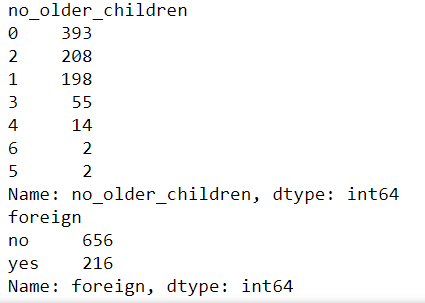




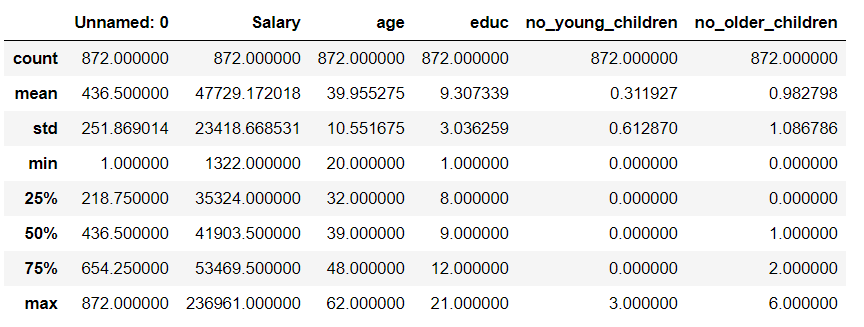






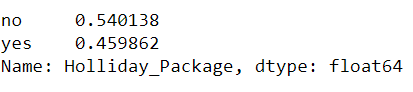


The describe function on the data looks like this:



The variables: Unnamed: 0 seems to be an employee number which will not be used in the model building section. Salary and age and educ are continuous variables. The variables: no\_young\_children and no\_older\_children refer to the number of children younger than 7 years for the former and the number of older children for the latter. These are more of categorical variables.

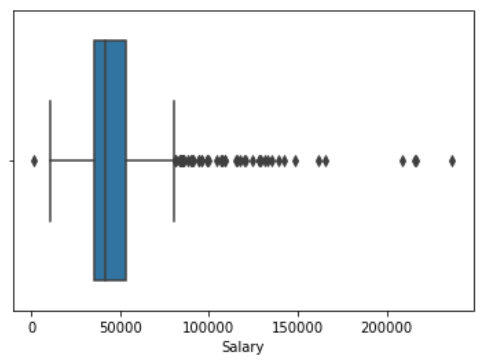
The distribution in the target variable: Holliday\_Package

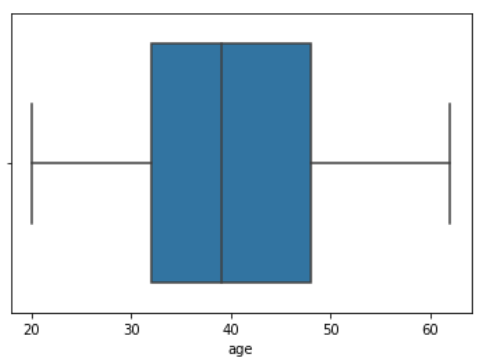


This distribution in the target variable is almost even with 54% for the ‘no’ and 46% for the ‘yes’.

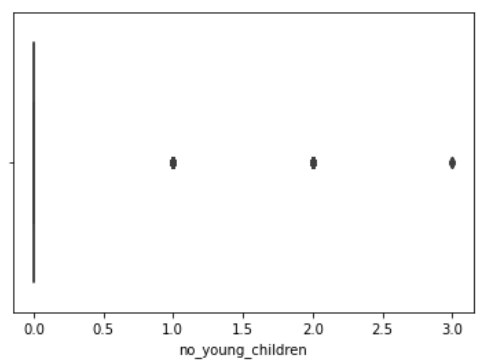
**Univariate and Multivariate Analysis:**

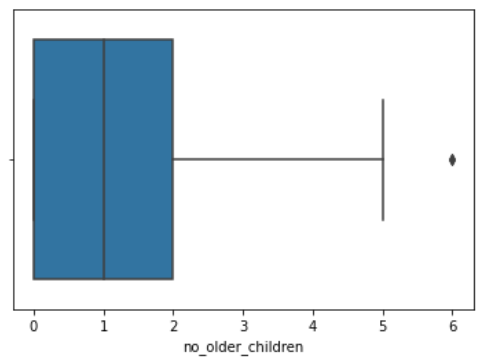
Boxplot:





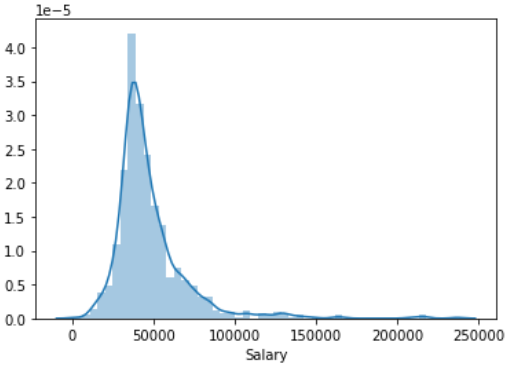


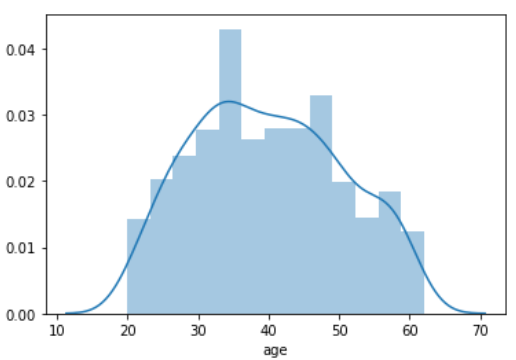


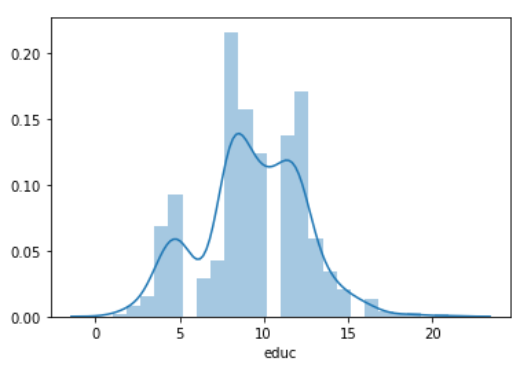


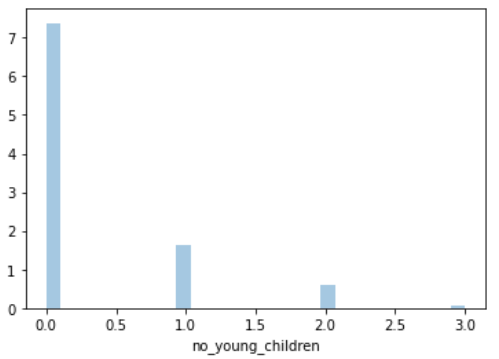
* ‘Salary’ has a lot of outliers.
* The variable ‘age’ does not have any outliers and indicates to be a categorical variable when grouped.
* The variable ‘educ’ has a few outliers on either side of the whiskers but also indicates to be a categorical variable when grouped as the values are definite and small in range.
* ‘no\_older\_children’ is a categorical variable as the numbers lie only within 0, 1, 2 and 3.
* Only one outlier is present in the ‘no\_older\_children’ with 6 children.

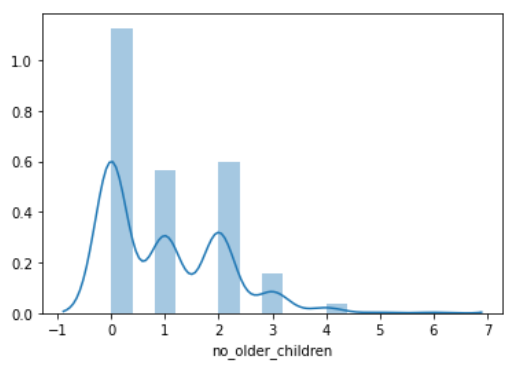
Distribution plot:











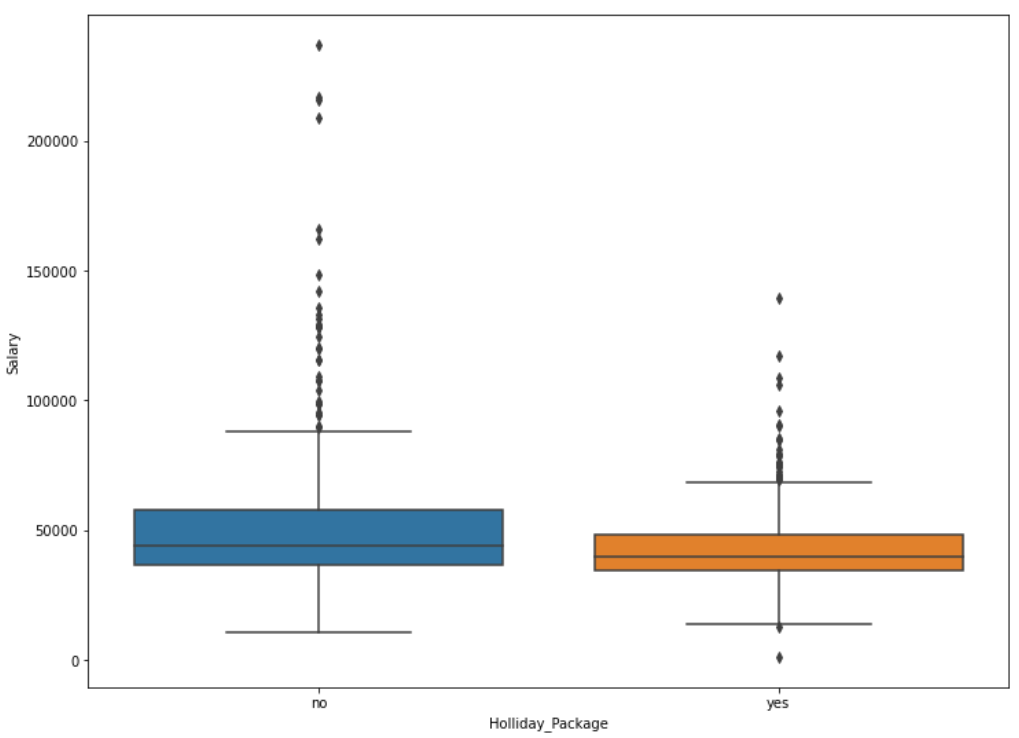
* ‘Salary’ is a fairly normal distribution with slight right skewness.
* ‘age’ and ‘educ’ are also normally distributed.
* ‘no\_ young\_children’ has more than 70 % of 0 children, with close to 15% of employees with 1 child and around 5% of employees with 2 children.
* ‘no\_older\_children’ has many employees with 0 children and is more of a right skewed distribution.
* It is actually inappropriate to view histograms for the variables: ‘no\_ young\_children’ & ‘no\_older\_children’ as they are categorical in nature.

There are no duplicate rows in the dataset.

A scatterplot between age and Salary looks like a cloud:

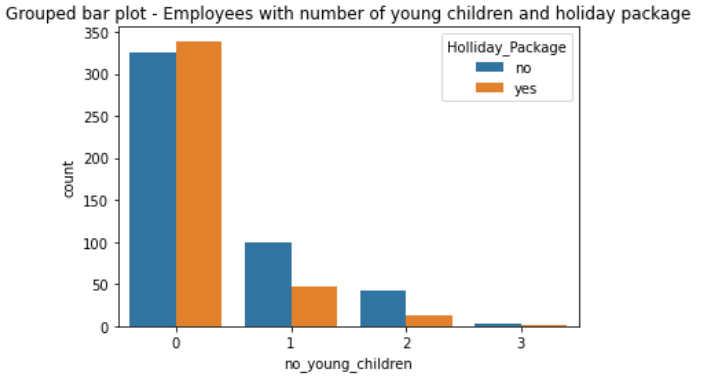


Categorical Boxplot:



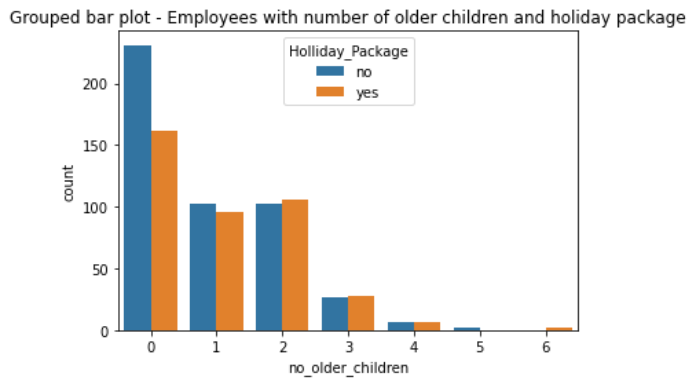
* The median salary of the employees who did not opt for the holiday package is slightly higher than that of people who did opt for it.
* The outliers are more in number and larger in magnitude of salary for the people who did not opt for the holiday package.

The following plot shows us a countplot of the employees who have young children and also the distribution on the ‘Holliday\_Package’ variable:



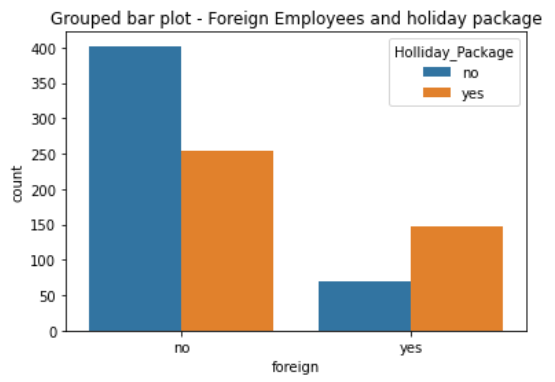
* For employees with 0 young children, the difference in the distribution on the holiday package is not large. It is almost even.
* For employees with 1 child, close to 100 employees did not opt for the holiday package and only 50 opted for the holiday package.
* For employees with 2 children, less than 50 did not opt for the holiday package and around 10 employees opted for the holiday package.

The following plot shows us a countplot of the employees who have older children and also the distribution on the ‘Holliday\_Package’ variable:



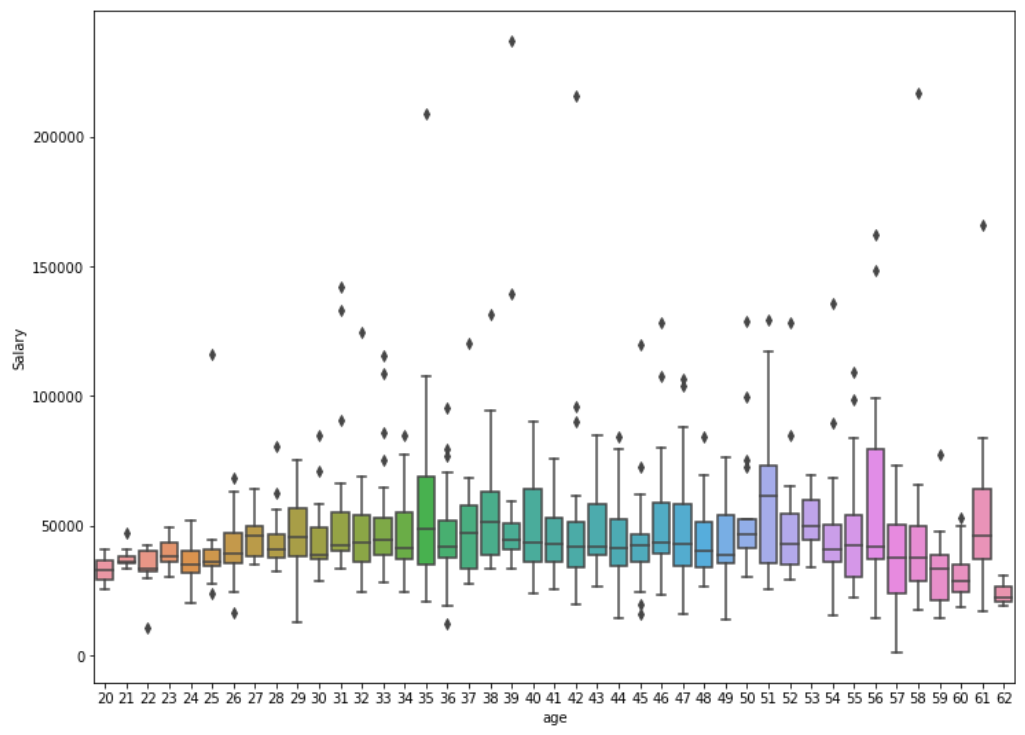
* For employees with 0 children, around 250 employees did not opt for the holiday package and slightly over 150 employees have opted for the same.
* For employees with 1 child, the distribution between the employees for the holiday package is almost even and is around 100.
* It is interesting to note that more number of employees with 2 children have opted for the holiday package than the number employees with 1 child did.
* Around 20 employees have opted for each of the two options in the Holiday Package variable.

Following is the count distribution between the variables: ‘foreign’ and ‘Holliday\_Package’



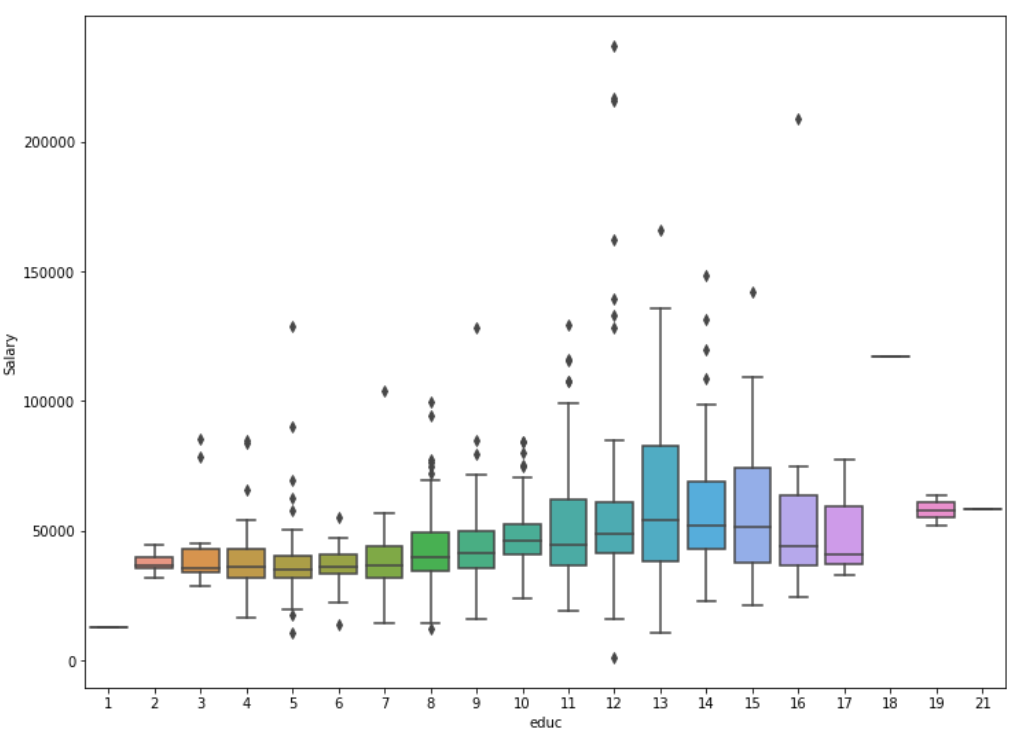
* Around 400 non-foreign employees have opted out of the holiday package whereas 250 have opted for the holiday package.
* The distribution is opposite for the foreign employees as around 150 have opted for the holiday package and only 60+ have opted out of it.

We then look at the boxplot for the age and Salary variables:



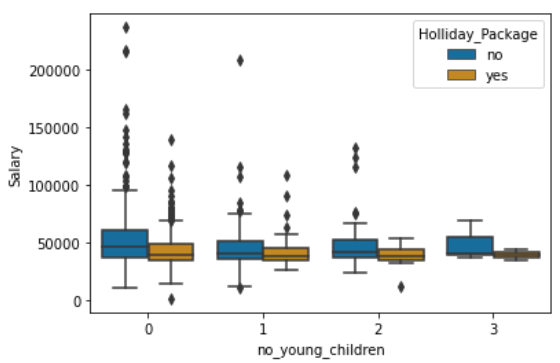
* It is interesting to note the size of the boxplots for the age of 56, 51 and 35. They are considerably large in number.
* The median salary distribution for the age of 51 is the highest and is lowest for the age 62.
* Outliers for salary are present in almost every age group.

**Boxplot between the ‘educ’ and ‘salary’ variables:**



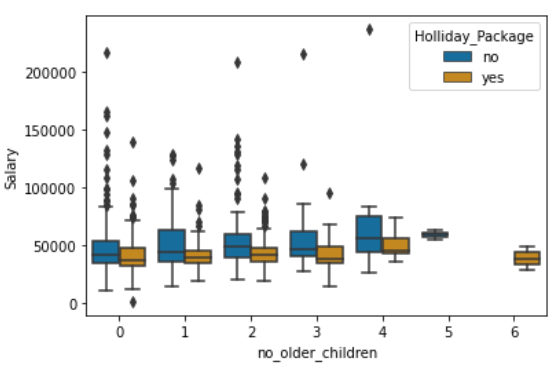
* Many employees are present in the education value of 13. Accordingly the median salary is also the highest for them.
* It is interesting to note that the median salary for employees increases with the years of formal education. Beyond 13 years of formal education the median salaries for the employees decreases with increase in the years of formal education.
* Outliers are also present for almost all the years of formal education.

**Boxplot between the variables: ‘no\_young\_children’ ,‘Salary’ and ‘Holliday\_Package’:**



* Many positive outliers for salary are present for the employees with 0 young children between those who did opt for the holiday package and for those who did not.

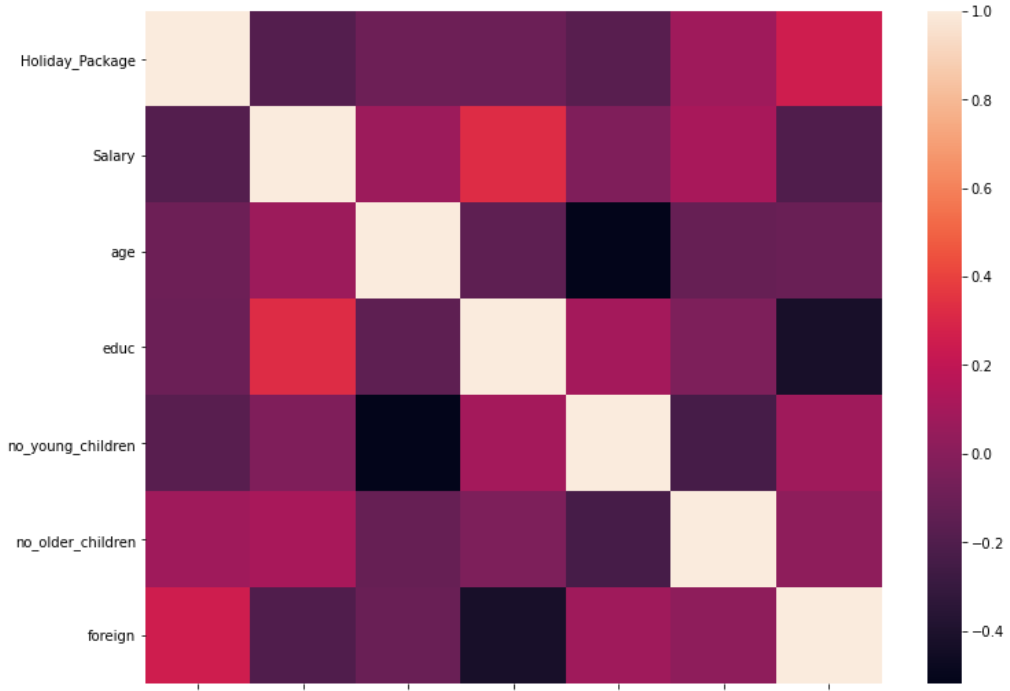
**Boxplot between the variables: ‘no\_older\_children’,’Salary’ and ‘Holliday\_Package’:**



* The median salary of the employees increases with the number of older children until 5 children. It is interesting to note that the salary for the employee with 6 older children is lower than the employees with 5 older children.
* The number of employees (with older children) who have not opted for the holiday package is distinctly more than those who have opted for the same.

**Multicollinearity checks:**

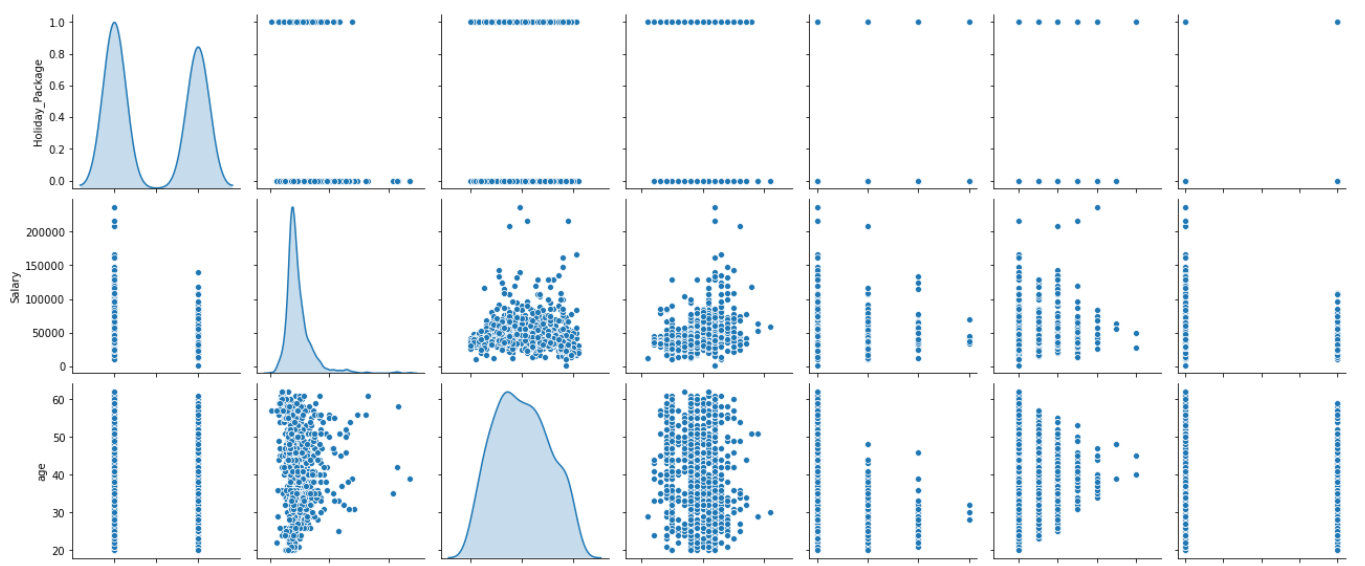
Heatmap:

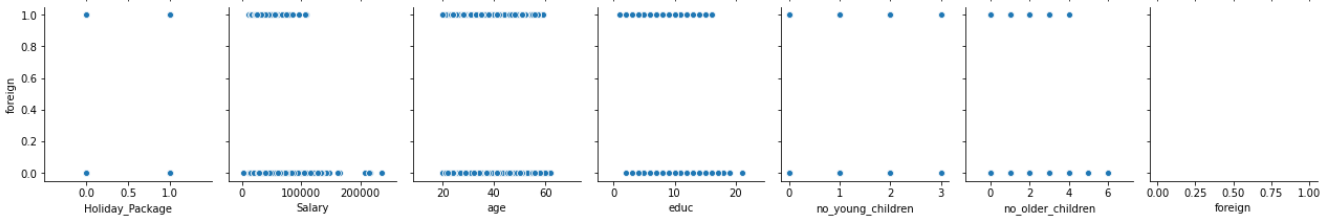




* Salary and educ variables positively correlated which goes to show that higher the educ, higher is the salary. But this correlation is only upto a certain value of educ, beyond which the salary decreases with increase in educ.
* No\_young\_children and age have a negative correlation.
* According to this plot, foreign is found to have some positive correlation with the target variable: Holiday\_Package.
* Most of the variables are weakly correlated with one another.

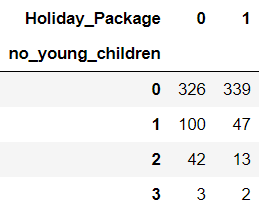
Pairplot:



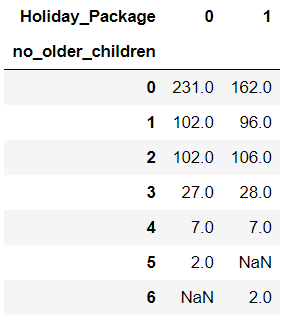
* There is absolutely no correlation between any of the variables according to the pairplot.

**Crosstab of no\_young\_children and Holiday\_Package:**

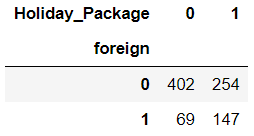


* Mostly the employees with 0 young children have opted for the Holiday\_Package. This is the only section where the proportion of employees who have opted for the Holiday\_Package has been more than those who have not.
* There are few observations for employees with 3 young children.

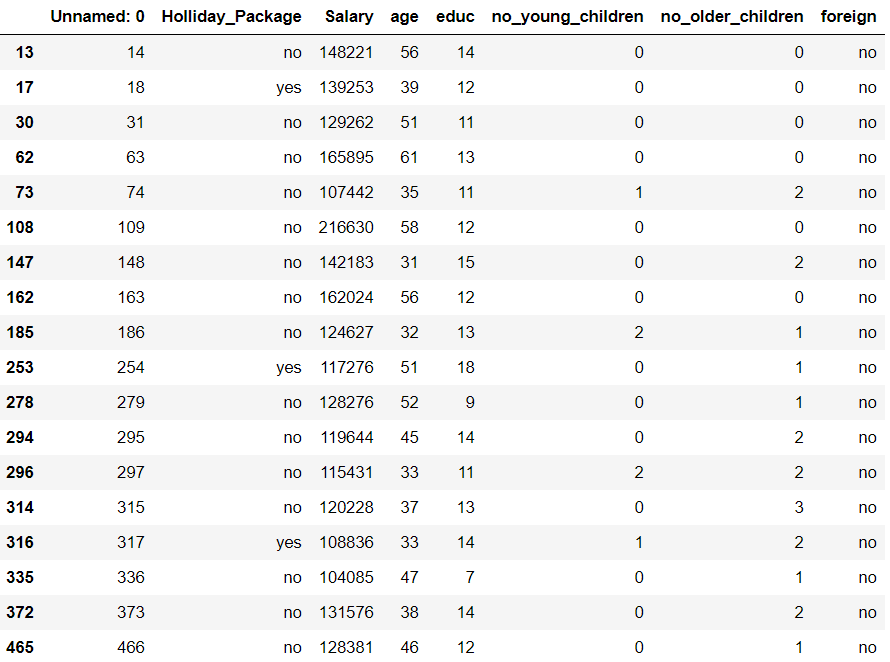
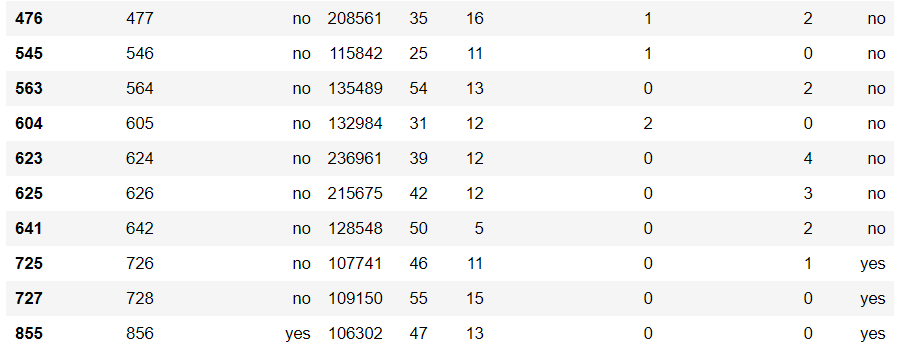
**Crosstab of no\_older\_children and Holiday\_Package:**



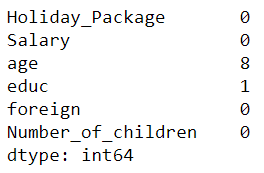
**Crosstab of foreign and Holiday\_Package:**



**Employees with salary more than 100000:**

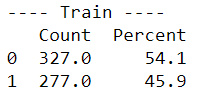
 

* Only 4 employees from this list have actually opted for the Holiday\_Package and hence there is no significant pattern to be declared here.
* We then combine the 'no\_young\_children' & 'no\_older\_children' into one column as: 'Number\_of\_children'.
* The 'no\_young\_children' & 'no\_older\_children' columns are then removed from our data.
* The ‘age’, ‘educ’ and ‘Salary’ variables are binned. The ‘age’ is binned in the following order: 20-30 is binned as 1, 30-40 as 2, 40-50 as 3 and 50-62 as 4. The ‘educ’ variable is binned in the following order: 1-5 as 1, 5-9 as 2, 9-12 as 3, 12-16 as 4, 16-21 as 5. The ‘Salary’ is binned into 10 almost equal bins and are labelled from 1 to 10 in the ascending order of Salary.

It is observed that there are 8 null entries in the age column and 1 null entry in the educ column. 

We will choose to drop the same as it is less than 3% of our data.

We then split the data into train and test with a proportion of 70:30. Following is the distribution of 0 and 1 of the Holiday\_Package in the train and test sets respectively:

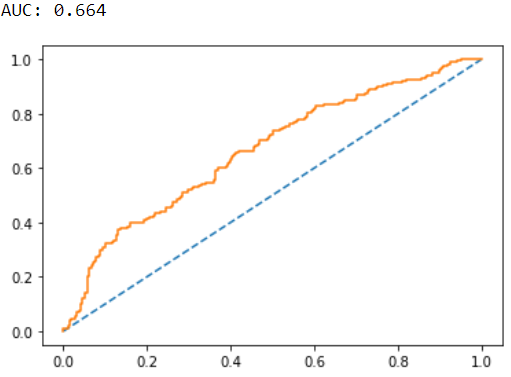




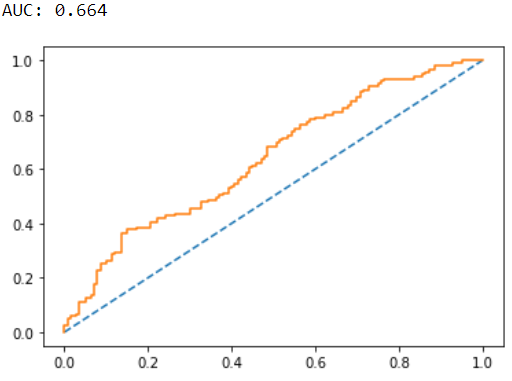
We then fit a Logistic Regression model with the following optimization parameters: ‘newton-cg’ as the solver, maximum iterations of 10000, with penalty as none so the properties of the regression coefficients are not controlled, verbose is set to True for verbosity and n\_jobs is set to 2 so that two CPU cores are used when parallelizing over classes.

|  |  |  |
| --- | --- | --- |
|  | Train | Test |
| Accuracy | 61.25 | 60.23 |

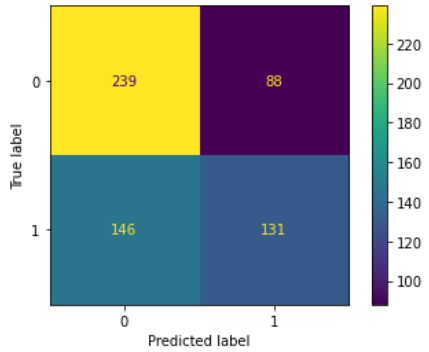
**AUC and ROC for the training data**

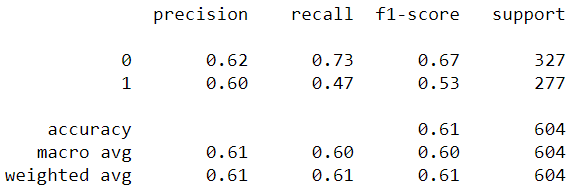


**AUC and ROC for the testing data**

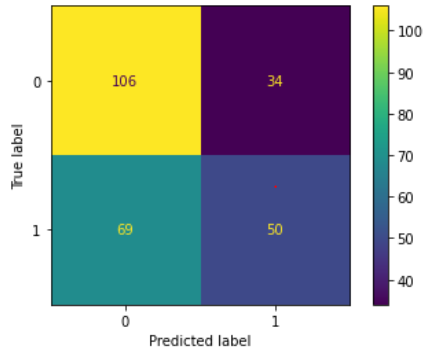


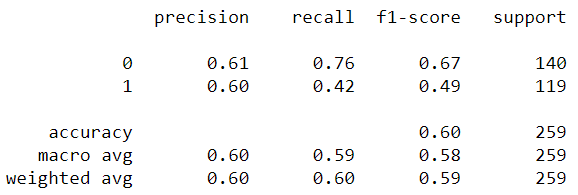
**Confusion Matrix and Classification report for the train data:**





**Confusion Matrix and Classification report for the test data:**





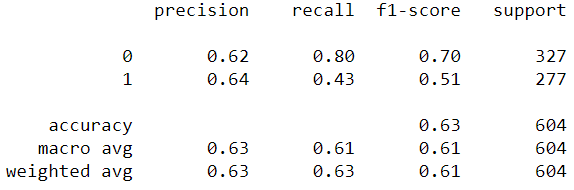
**Applying GridSearchCV for Logistic Regression:**

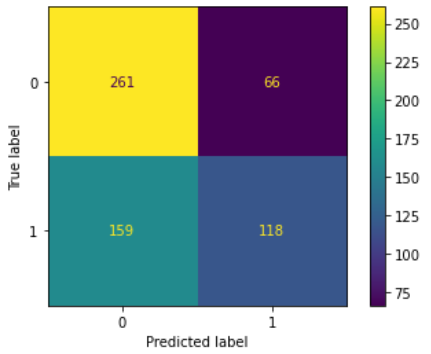
We use the Grid parameters as penalty: ‘l2’ and ‘none’, solver: ‘sag’ and ‘lbfgs’ and tolerance: 0.1,0.01,0.001

Maximum iteration in the model as 10000 and n\_jobs as 2. In the grid search we use cross validation as 3 and n\_jobs as -1 that means using all the core processors. The best grid is then fit on the train and test data.

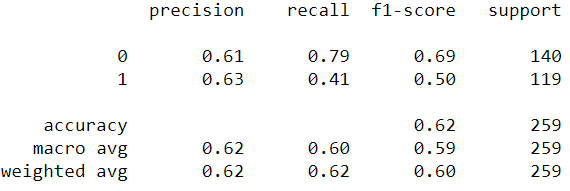
The model accuracy on the train data is 62.74% and on the test it is 61.77%

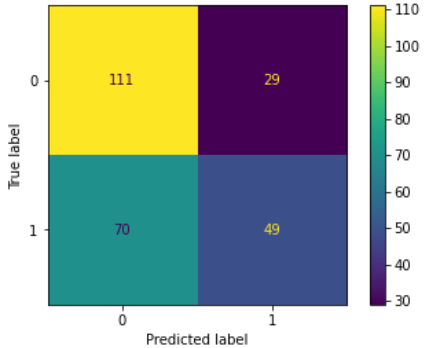
**Confusion Matrix and Classification report on the train data:**





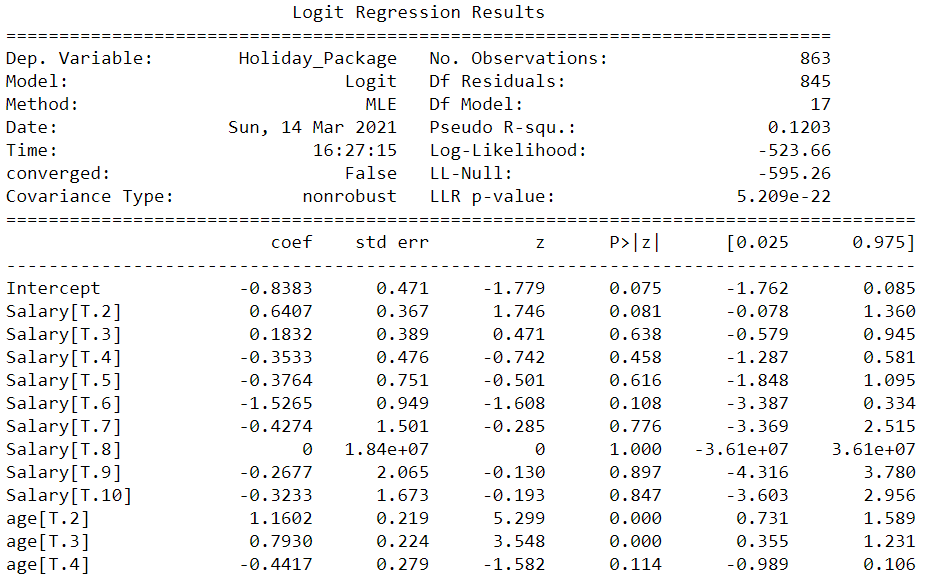
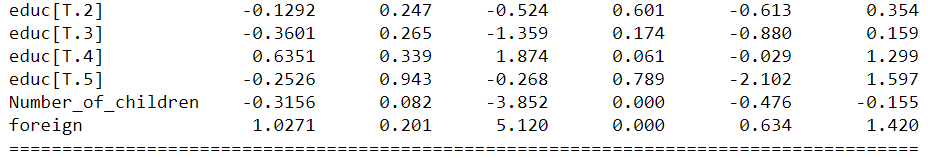
**Confusion Matrix and Classification report on the test data:**





### **Building a logistic regression model using the GLM model approach and Model Evaluation:**

We use the method as ‘bfgs’.



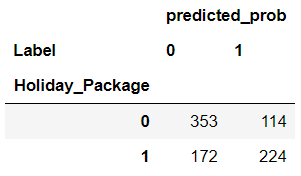


It is to observe that in this model iteration, the Salary variable has shown to be not useful for any of the bins at 95% confidence since the p values are greater than 0.05. The age variable for the bin 4 has shown to be not useful. None of the educ variables have shown to be useful in predicting the target: Holiday\_Package since the p values are greater than 0.05.

The Akaike Information Criteria (AIC) = -2(log Lm) + 2p

= -2(-523.66) + 2 x 21 = -1047.32 + 42 = -1005.32

Confusion Matrix for the model:



If the cut-off is set at 0.5, then misclassification probability is (114+172/863 = 0.3314 = 33%

Therefore, accuracy of the model is 1 – 0.3314 = 66.8%

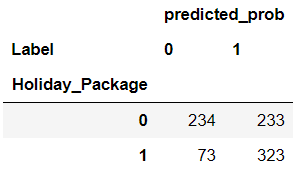
Precision is 224/(172+224) = 56.56%

Specificity is 353/(353+114) = 75.58%

Recall is 224/(224+114) = 66.27%

F1-score = 2 \* 0.5656 \* 0.6627 / (0.5656+0.6627) = 61.02%

Let us change the cut-off value to 0.35 and find the confusion matrix for the model:



Misclassification probability is (233+73)/863 = 0.3545 = 35%

Therefore, the accuracy of the model is 1 – 0.3545 = 64.5%

Precision is 323/(73+323) = 81.5%

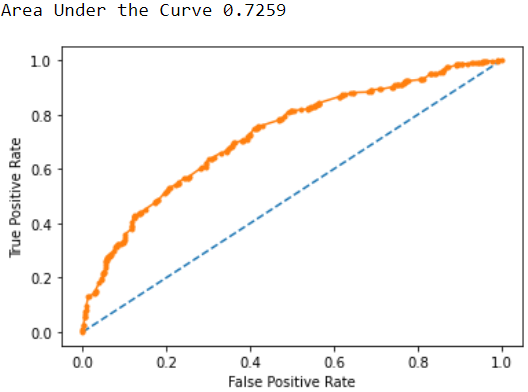
Specificity is 234/(234+233) = 50.10%

Recall is 323/(323+233) = 58.09%

F1-score = 2 \* 0.815 \* 0.5809/(0.815+0.5809) = 67.82%

By changing the cut-off value to 0.35, the accuracy of the model has decreased slightly but the precision and F1 score values have improved a lot. Specificity and Recall have decreased as compared to before.

**ROC curve and AUC:**



The AUC in this model is the highest we have got in any model so far.

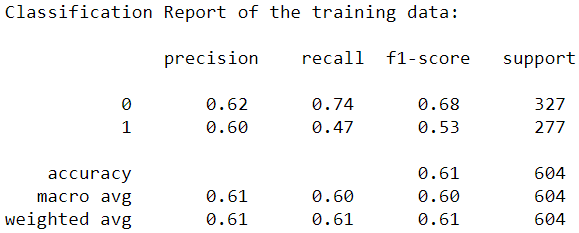
**LDA Model:**

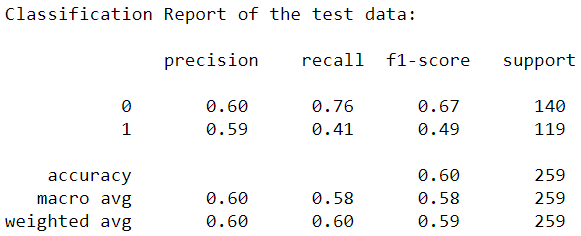
We fit the Linear discriminant Analysis on the train data and generate the model.

### **Confusion Matrix on Training and Test Data:**

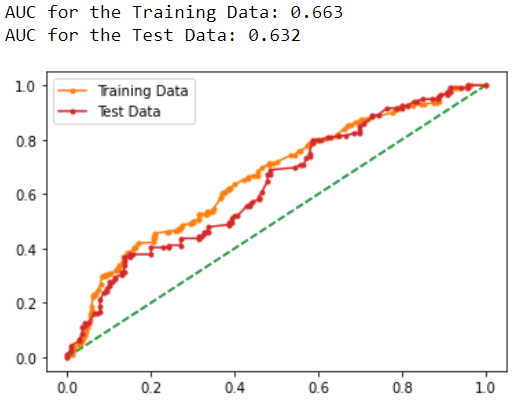


### **Training Data and Test Data Classification Report Comparison:**

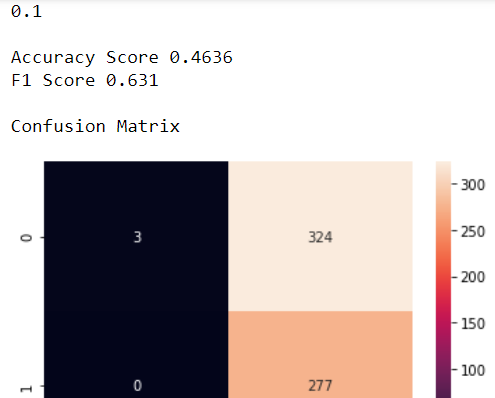
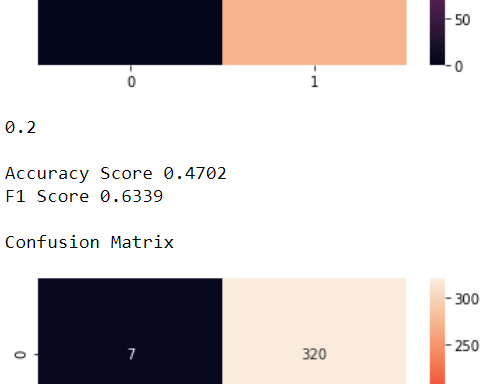
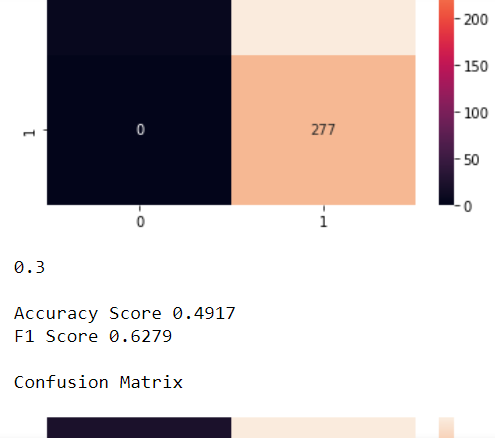
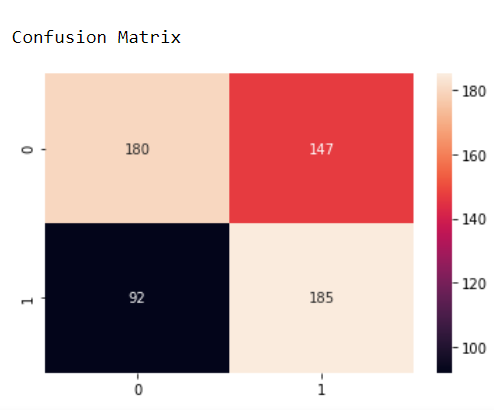
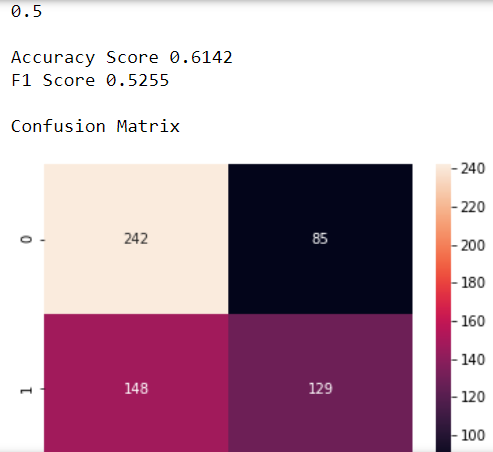
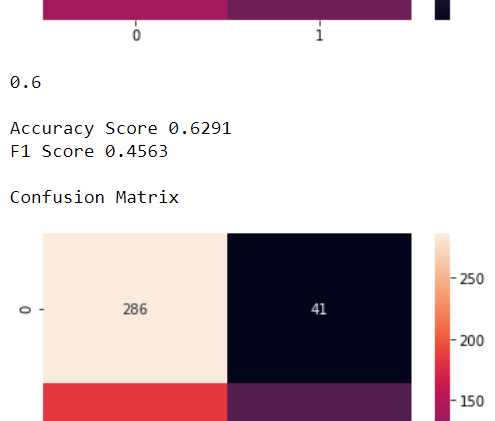
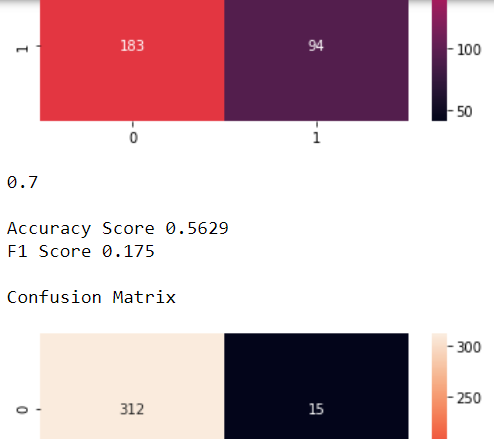
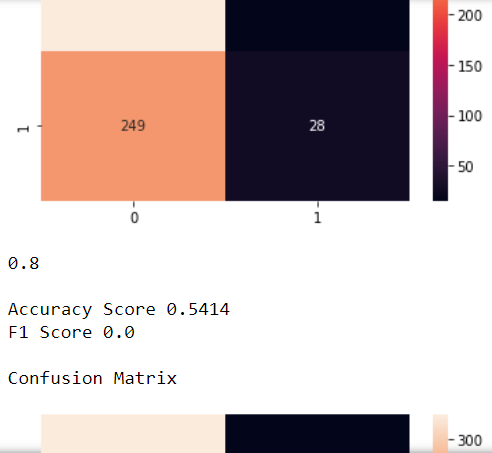
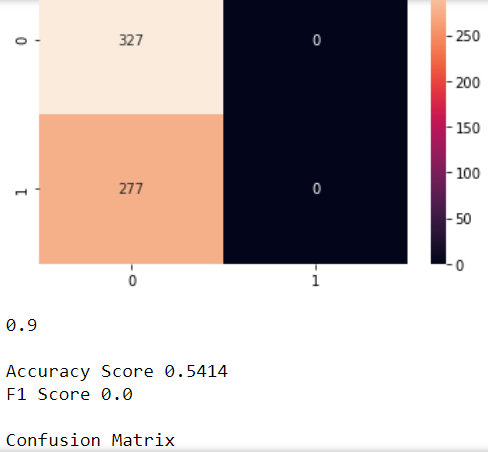
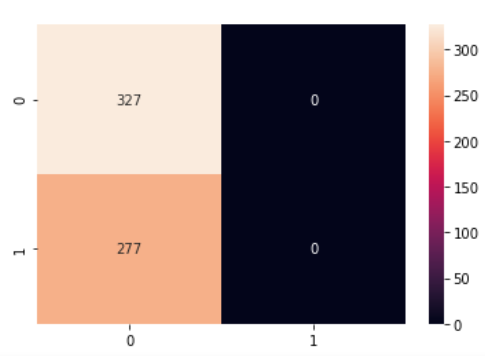




### **AUC and ROC for the train and test data:**



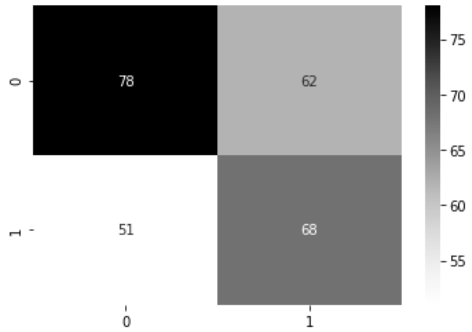
### **Changing Custom cut-off values to improve accuracy:**

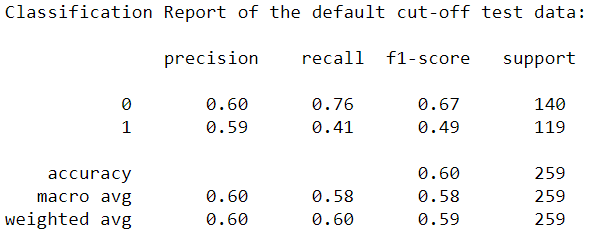
          

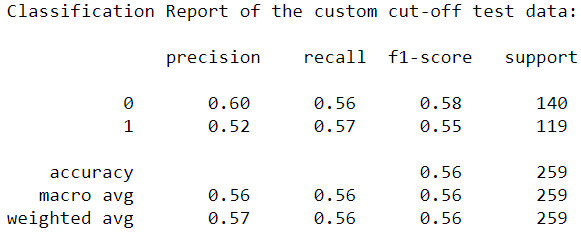
It can be observed that it is a trade off between accuracy and F1 score. When one increases the other decreases and vice-versa.

We take the cut off as 0.4 as both the accuracy and F1 score is 60%

**Confusion matrix and Classification report for custom cut-off:**





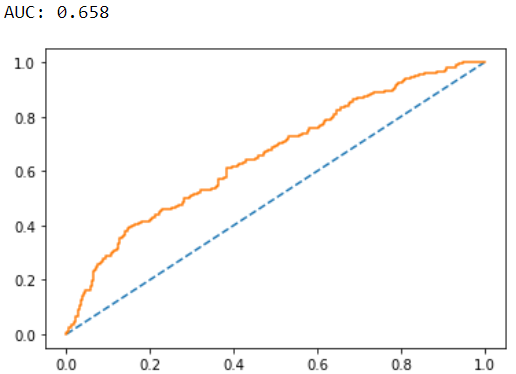


### **Without binning the variables: age and educ and Salary:**

We split the data into train and test and run the logistic regression model with the model tuning parameters as: solver='newton-cg',max\_iter=10000,penalty='none',verbose=True,n\_jobs=2

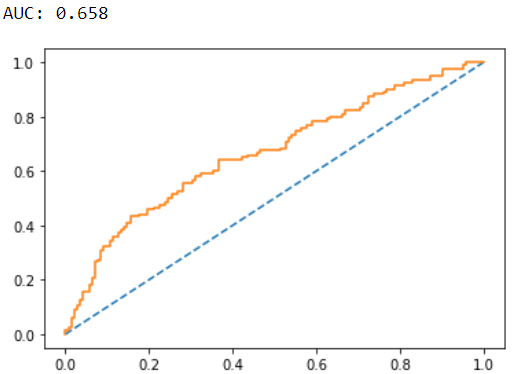
The model accuracy score on the train data is 61.96%

**AUC and ROC for the training data:**

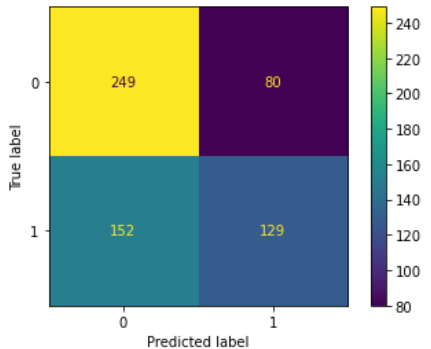


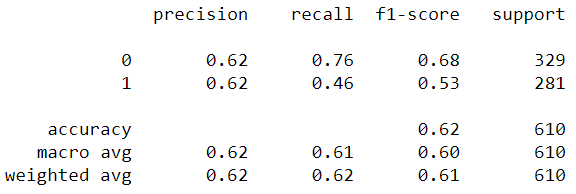
The model accuracy score on the test data is 64.12%

**AUC and ROC for the testing data:**

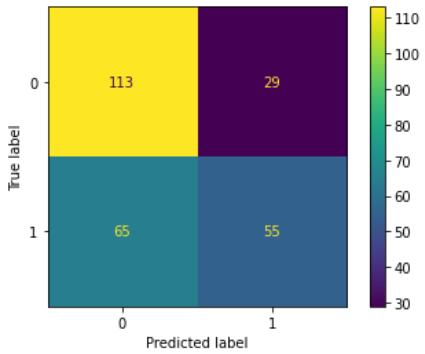


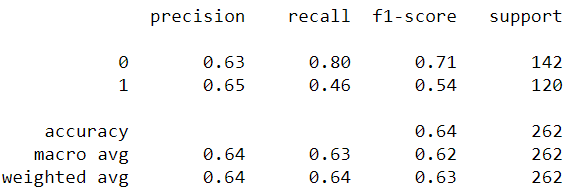
**Confusion Matrix and Classification report on the training data:**





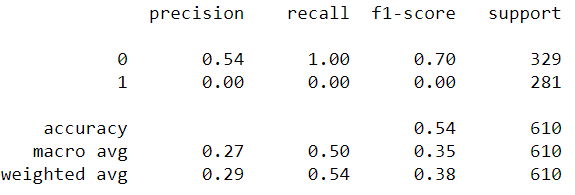
**Confusion Matrix and Classification report on the testing data:**

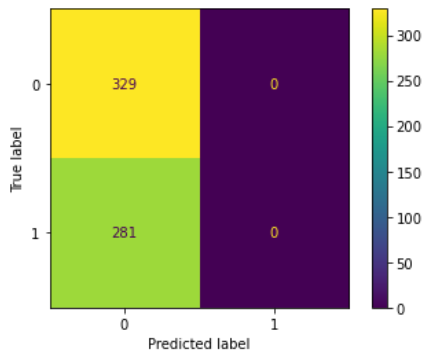




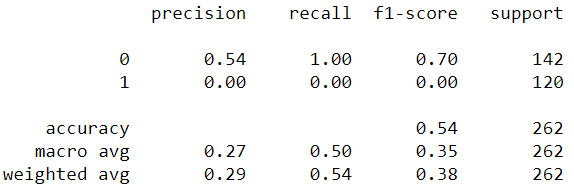
# **Applying GridSearchCV for Logistic Regression:**

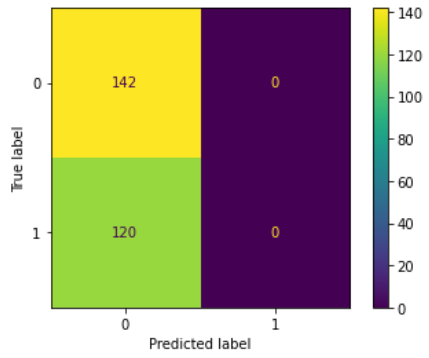
**Classification report and Confusion matrix on the train data:**





**Classification report and Confusion matrix on the test data:**

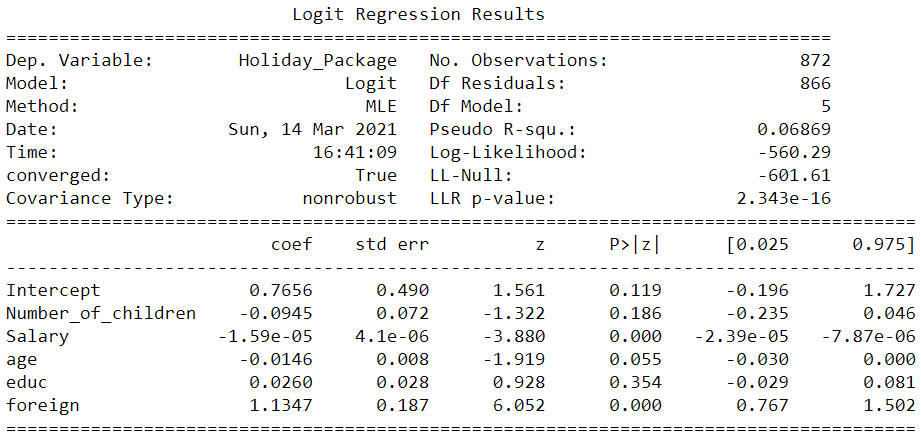




We can clearly see that the model is over-fitting and hence the Grid Search CV does not yield useful results.

**Building a logistic regression model using the GLM model approach:**

The model summary is as follows:





The Nagelkerke R Squared for Model2: Not Treating outliers for Salary and not binning the salary, educ & age variables is 0.12.

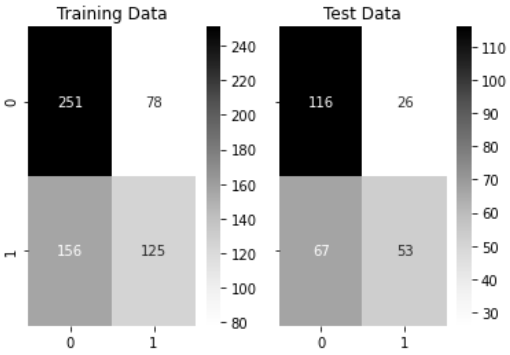
It can be observed that the p-values for all the other variables except Salary and foreign are more than 0.05 and hence indicate to be rejected.

The McFadden Pseudo R Squared and Nagelkerke R Squared for this model is also low.

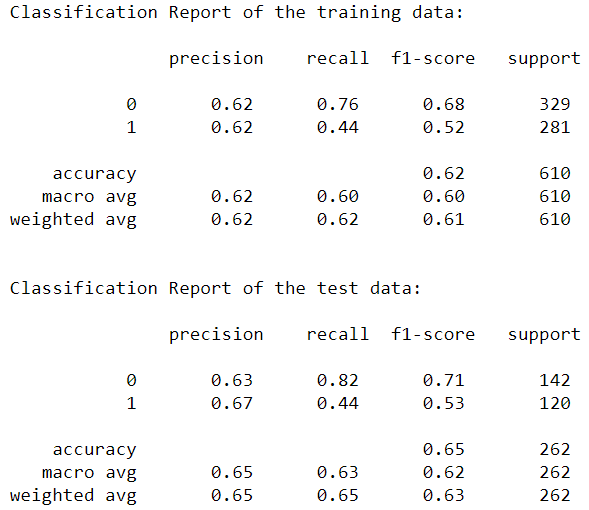
Hence, we will not continue with this model as this model has indicated many of the variables as not useful.

**LDA Model:**

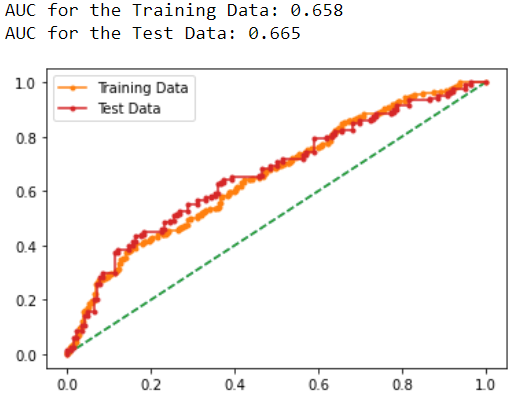
**Confusion Matrix on Training and Test Data:**



**Training Data and Test Data Classification Report Comparison:**

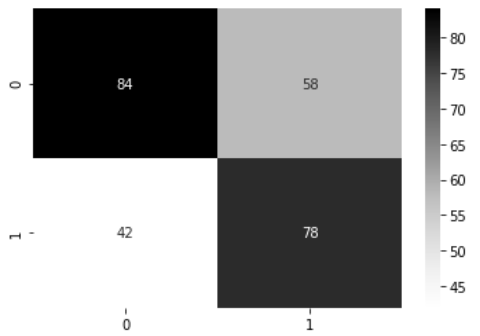


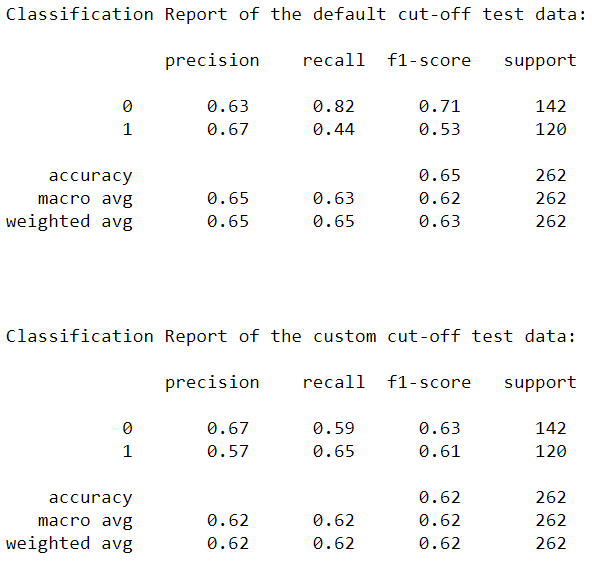
**AUC and ROC for the train and test data:**



**Changing Custom cut-off values to improve accuracy:**

Taking the cut-off value as 0.4 as both the accuracy and F1 score is 60%:





**Treating outliers for Salary and binning the educ & age variables:**

We fit a logistic regression model from sklearn using the following model tuning parameters: solver='newton-cg',max\_iter=10000,penalty='none',verbose=True,n\_jobs=2

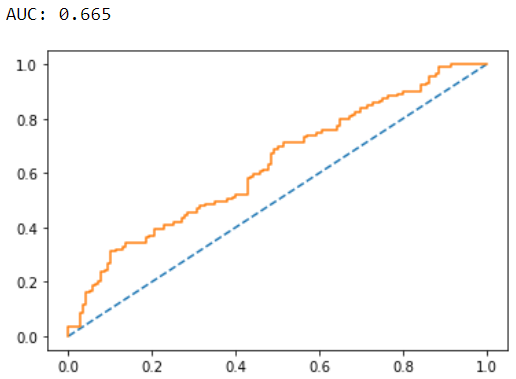
Model accuracy on the train data is 61.92%

**AUC and ROC for the training data:**



Model accuracy on the test data is 58.68%

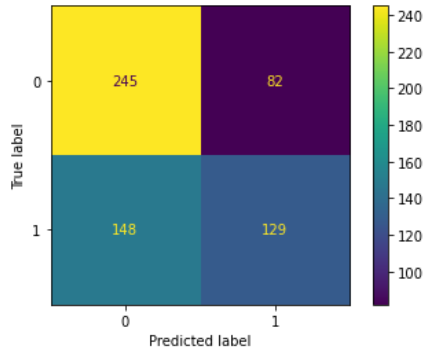
**AUC and ROC for the testing data:**

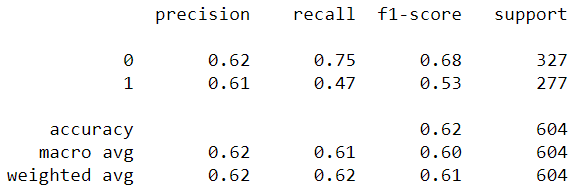


The model is slightly overfitting as the accuracy on the test data is 59% and on the train data it is 62%.

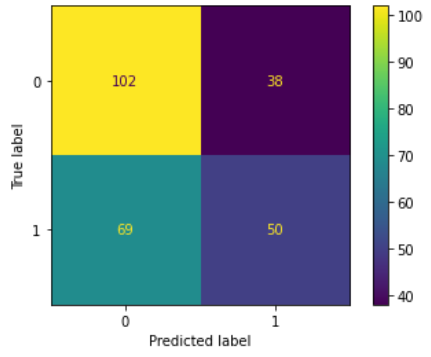
The AUC values are the same for the train and test data, although the ROC curve for the test data is visually lower than that of the train data.

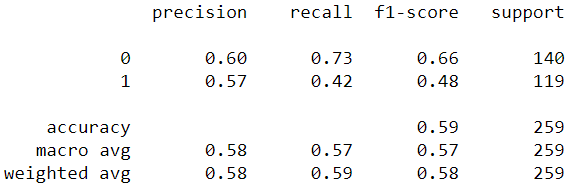
**Confusion Matrix & Classification report for the training data:**





**Confusion Matrix & Classification report for the testing data:**





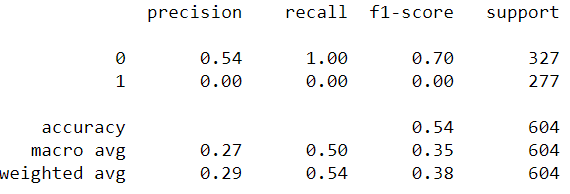
Accuracy and Precision are two important metrics for our problem statement. We can slightly relax the other values as the consequences of misclassification is not much.

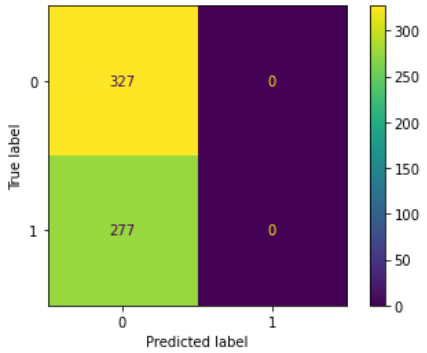
**Applying GridSearchCV for Logistic Regression:**

We apply the following grid parameters with penalty as ‘l2’ and ‘none’, solver as ‘sag’ and ‘lbfgs’, tolerance as ‘0.1’,’0.01’,’0.001’.

In the grid search CV we apply 3 fold cross-validation, number of jobs as -1 (this means that all core processers will be used), scoring as ‘f1’ score. We extract the best grid and plot the confusion matrix and classification report on the train and test data:

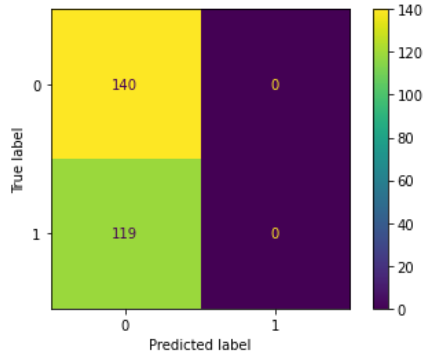
**Train data:**





**Test data:**



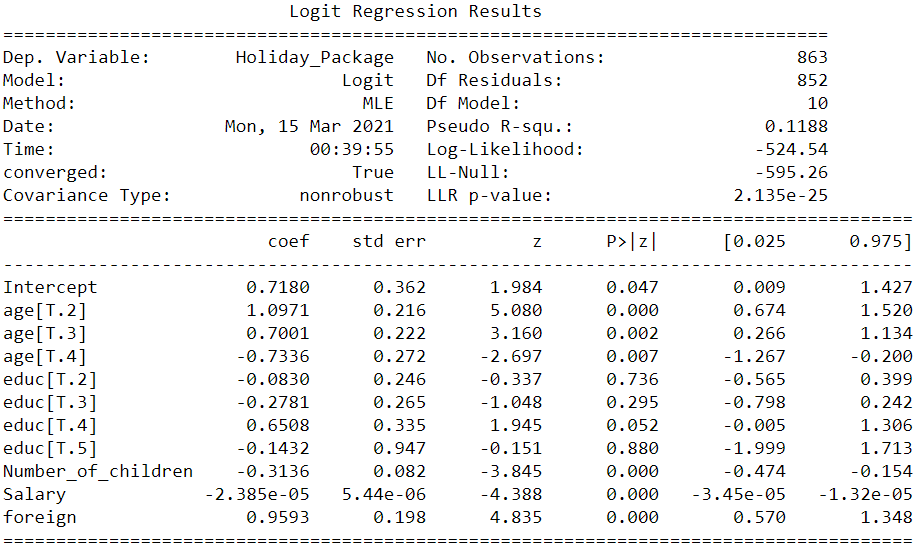


The model is clearly overfitting on both the training and testing data.

Hence will not proceed with further analysis.

**Building a logistic regression model using the GLM model approach:**

The model summary is printed:







The p-values for age, Number\_of\_children, Salary, foreign are less than our cut off 0.05. The educ for the 4th bin (educ: 12 – 16) is also marginal on 0.052 and hence can still be considered. The McFadden Pseudo R Squared and Nagelkerke R Squared values for this model is also good.

The Akaike Information Criteria (AIC) for this model = -2(log Lm) + 2p = -2(-524.54) + 2 \* 12 = 1073.08

For one more age-2nd bin, odds of choosing the Holiday package increases by e^1.0971 = 2.9955.

For one more age-3rd bin, odds of choosing the Holiday package increases by e^0.700 = 2.0138.

For one more age-4th bin, odds of choosing the Holiday package increases by e^(-0.7336) = 0.48

For one more educ-4th bin, odds of choosing the Holiday package increases by e^(0.6508) = 1.91

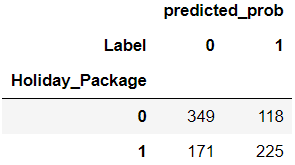
For one more Number\_of\_children, odds of choosing the Holiday package increases by e^(-0.3136) = 0.73

For one more Salary, odds of choosing the Holiday package increases by a very negligible number.

For one more foreign, odds of choosing the Holiday package increases by e^0.9593 = 2.609

Hence let us proceed with further analysis:

**Confusion matrix with threshold as 0.5:**



If the cut-off is set at 0.5, then misclassification probability is (118 + 171)/863 = 33.48%

Therefore, accuracy of the model is 1 – 33.48% = 66.51%

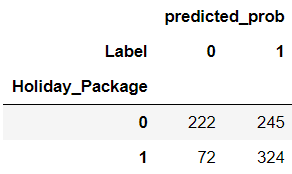
Precision is 225/(171+225) = 56.81%

Specificity is 349/(349+118) = 74.73%

Recall is 225/(225+118) = 65.59%

F1-score = 2 \* (0.5681) \* (0.6559)/(0.5681 + 0.6559) = 0.7452/1.224 = 60.88%

Let us change the cut-off value to 0.35 and find the confusion matrix for the model:



If the cut-off is set at 0.35, then misclassification probability is (245 + 72)/863 = 36.73%

Therefore, accuracy of the model is 1 – 36.73% = 63.27%

Precision is 324/(324 + 72) = 324/396 = 81.81%

Specificity is 222/(222 + 245) = 47.53%

Recall is 324/(324 + 245) = 56.94%

F1-score = 2 \* (0.8181) \* (0.5694) / (0.8181 + 0.5694) = 0.9316/1.3875 = 67.14%

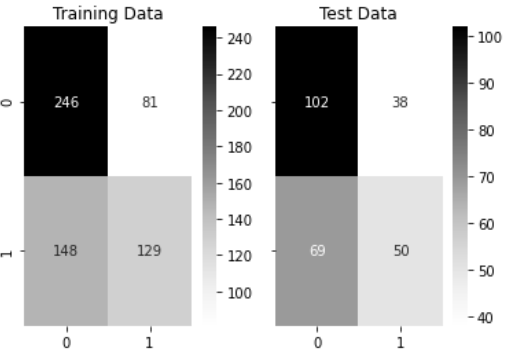
It is to note that the False Positives do not really have any financial impact on the company as. But the False Negatives does as the company will miss out on the opportunity to sell the holiday package to the employee. Precision is hence the most important metric in our data.

**AUC and ROC curve:**

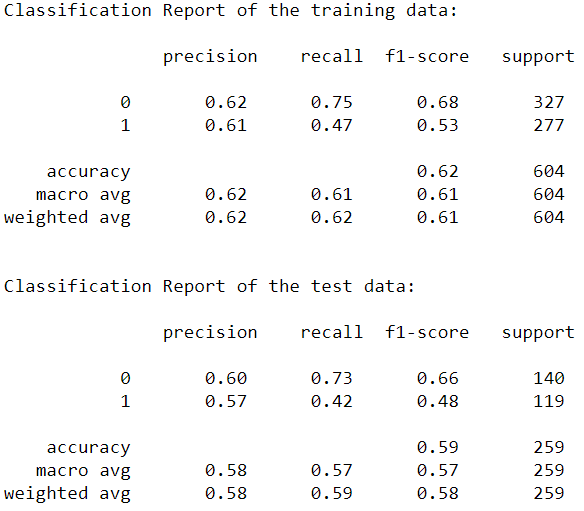


**LDA Model:**

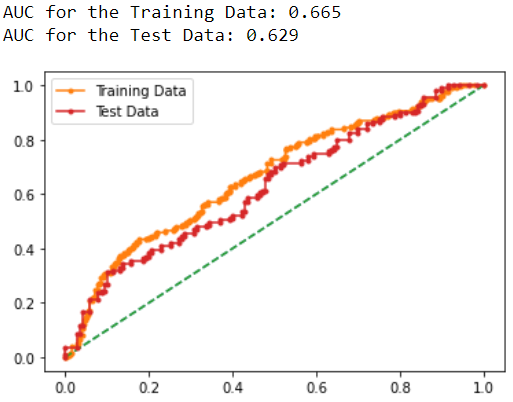
**Confusion Matrix on Training and Test Data:**



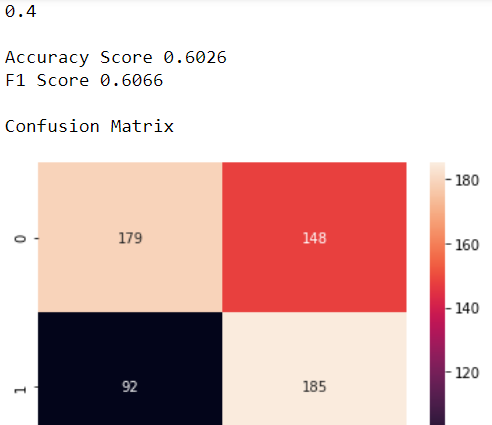
**Training Data and Test Data Classification Report Comparison:**



**AUC and ROC for the train and test data:**

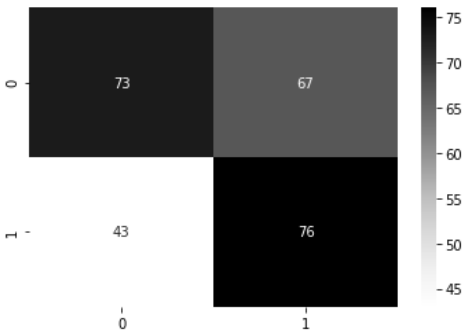


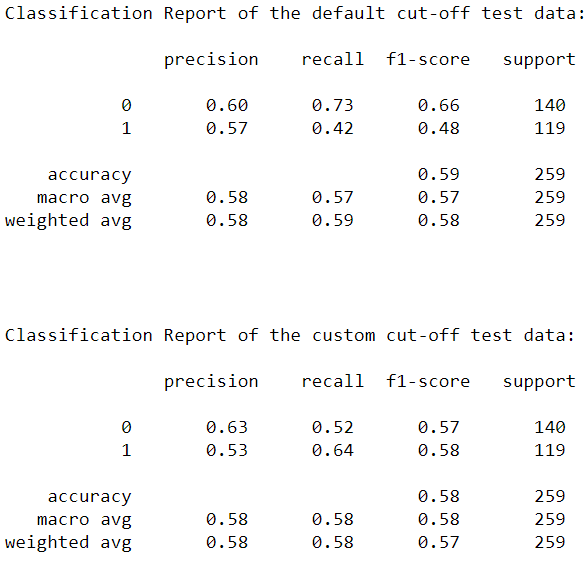
**Changing Custom cut-off values to improve accuracy:**

Taking the cut-off value as 0.4 as both the accuracy and F1 score is 60% on the train data:

Test data results:



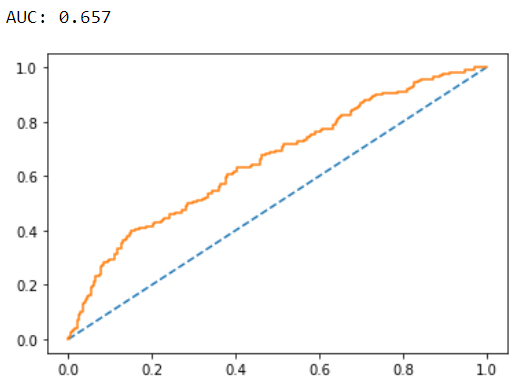


**Treating outliers for Salary and not binning the variables educ and age:**

From Sklearn linear model we run the logistic regression model on the data with the following model tuning parameters: solver='newton-cg', max\_iter=10000, penalty='none', verbose=True, n\_jobs=2

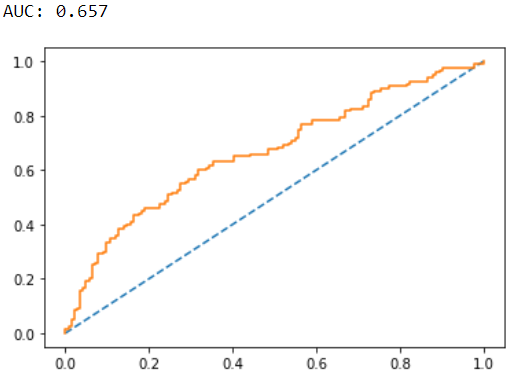
The model accuracy score on the train data is 61.47%

**AUC and ROC for the training data:**

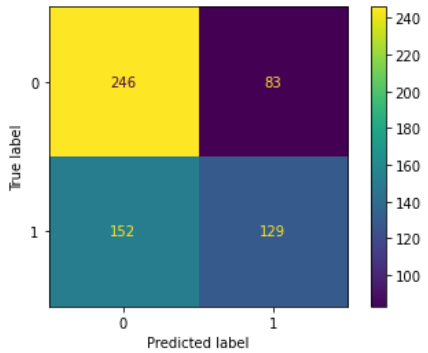


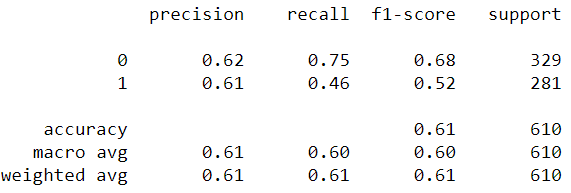
Model accuracy score on the test data: 63.74%

**AUC and ROC for the test data:**

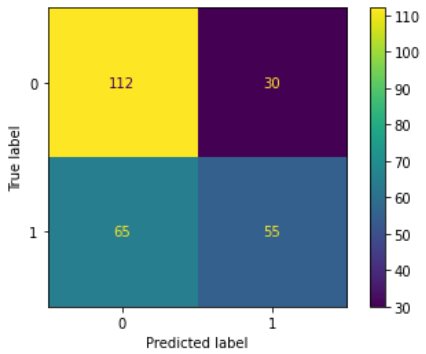


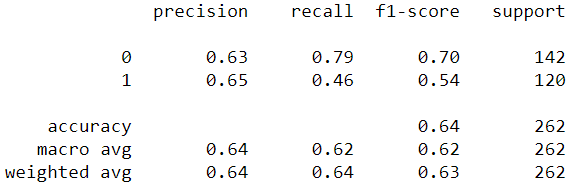
**Confusion Matrix & Classification report for the training data:**





### **Confusion Matrix & Classification report for the testing data:**





**Applying GridSearchCV for Logistic Regression:**

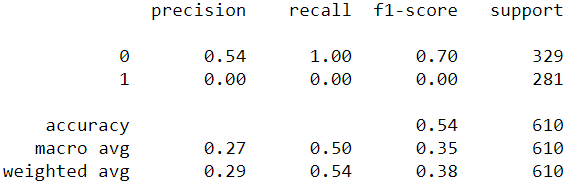
We use the following grid parameters: ‘penalty’ as ‘l2’ and ‘none’, ‘solver’ as ‘sag’ and ‘lbfgs’, ‘tolerance’ as ‘0.1’,’0.01’,’0.001’.

The model is run with maximum iterations as 10000 and n\_jobs as 2. In the grid search we fit an additional cross validation of 3.

The best grid is fit on the data and the results are explained:

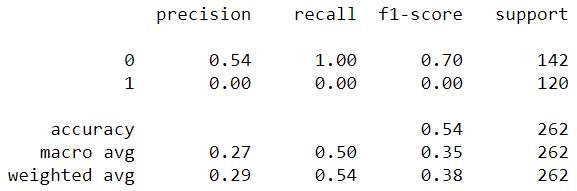
**Confusion matrix and Classification report on the Train data:**





**Confusion matrix and Classification report on the Test data:**



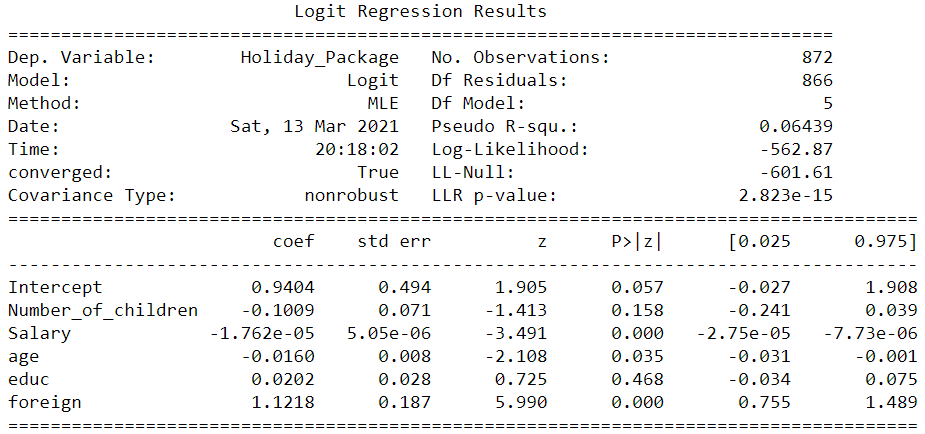


Using the grid search we got an overfitting model. We will not continue with further analysis on this model.

### **Building a logistic regression model using the GLM model approach:**

We build a logistic regression model using the Statsmodel approach.

We print the model summary:







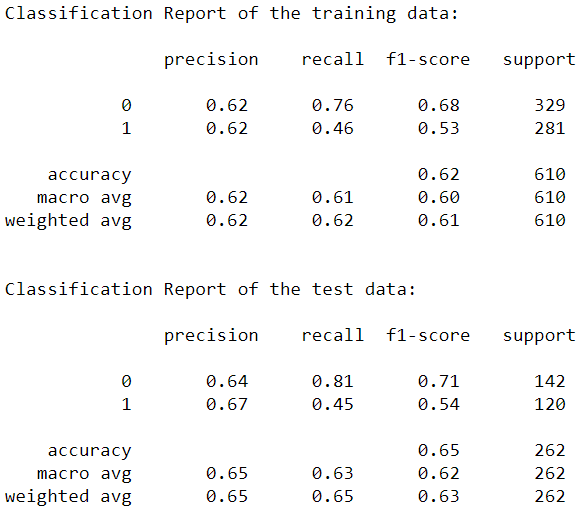
The Number\_of\_children and educ have become not useful variables and the McFadden Pseudo R Squared and Nagelkerke R Squared for this model is also low. Hence we will not continue with further analysis on this model.

### **LDA Model:**

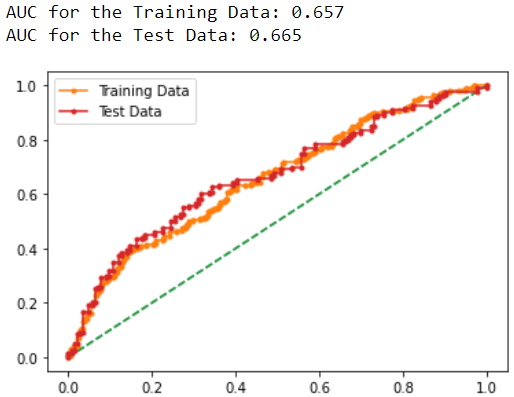
### **Confusion Matrix on Training and Test Data:**



**Training Data and Test Data Classification Report Comparison:**



### **AUC and ROC for the train and test data:**



**Inferences: Insights and Recommendations**

* Employees with 0 number of young children have a high probability 51% chance of choosing the holiday package.

|  |  |
| --- | --- |
| Number of older children | Percentage of choosing the Holiday Package in % |
| 0 | 41.22 |
| 1 | 48.48 |
| 2 | 50.96 |
| 3 | 50.90 |

* For a non-foreign employee there is a 38% probability that he will choose the holiday package.
* For a foreign employee there is a 68.05% probability that he will choose the holiday package.

### ‘Treating outliers for Salary and binning the educ & age variables’ this is the best version of the data that has produced good results. ‘Building a logistic regression model using the GLM model approach:’ - This is the best model.

* The most important variables are foreign, Number of older children, employees between the age 30 – 50, employees with education levels between 12-16.
* An additional variable: ‘Holidaying history’ can be included in the data collection phase and would definitely prove to be an important variable in determining the Holiday Package being taken.
* The model cut off should be set to 0.35 and the precision is 82%. False Negatives are also reduced to 72. The AUC is also the highest with 72%.