PREDICTIVE MODELING PROJECT

BUSINESS REPORT

**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 best and J the worst. |
| Clarity | Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst 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. |

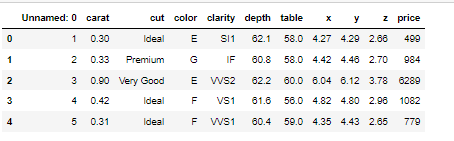
Dataset for Problem 1: [cubic\_zirconia.csv](https://olympus.mygreatlearning.com/courses/78182/files/6689408/download?verifier=LdxYDuUuOXYcQY1zucTtbnktg6sJx4FasrmiGrmp&wrap=1)

* 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.**

**Load the data:**

**EDA:**

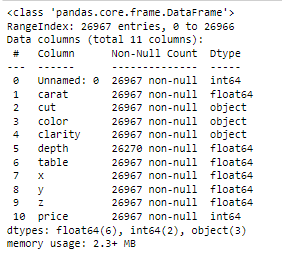
**Head:**

****

**Data shape:**

(26967, 11)

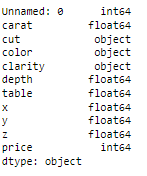
**Data info**

****

Based on above output this data set has 11 variables and 26966 observations.

Based on above output there is missing values are present in depth variable in data set.

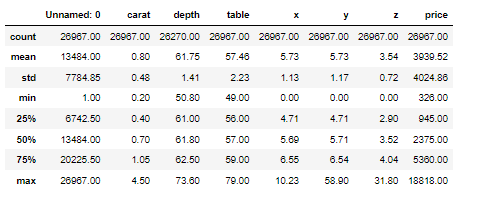
**Data types:**



Most of the columns in the data are numeric in nature (int64 or float 64 type).

The cut, color and clarity are string columns(object type).

**Data describe:**

****

The above table shows 5 point summary of the data.

Based on output the data looks good.

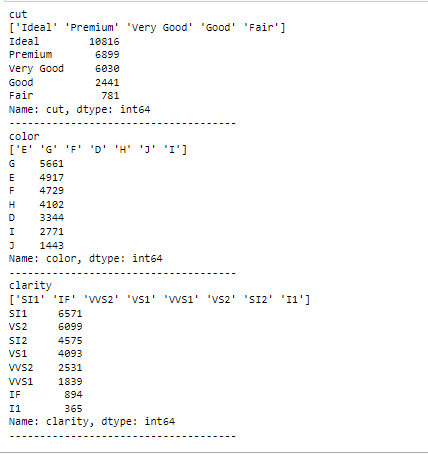
All the variables mean and median are nearly equal.

Before going to the further analysis, we confirmed if we have any duplicate values in the data.

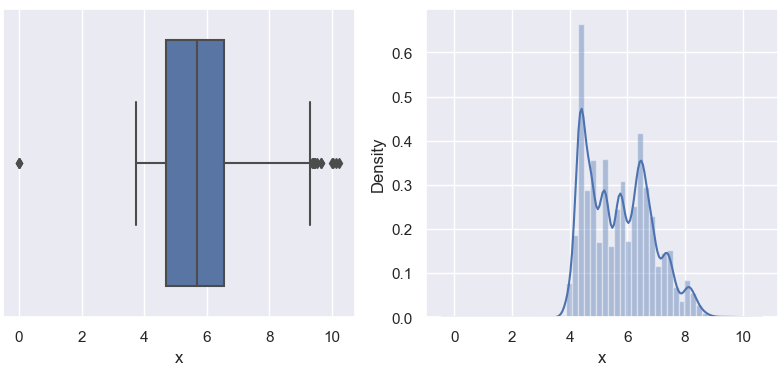
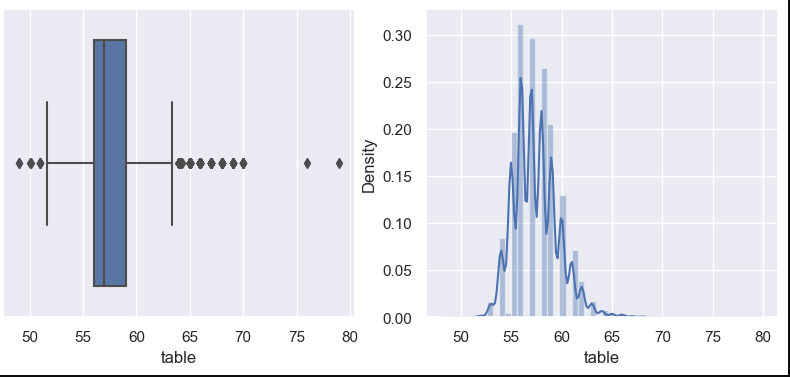
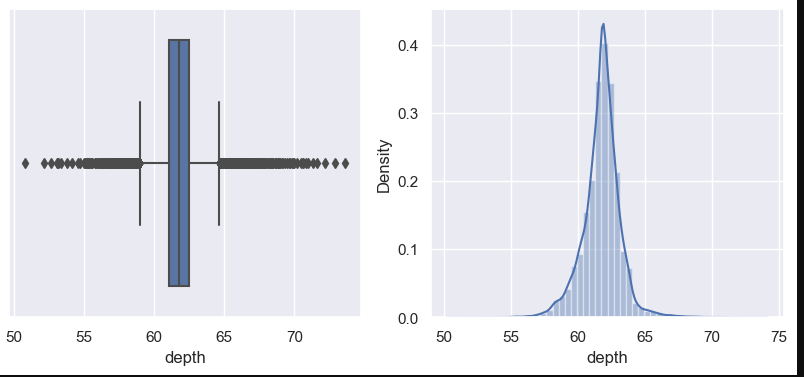
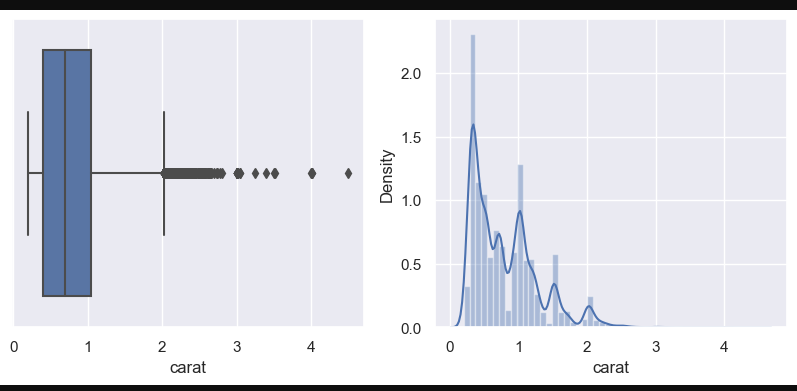
**Duplicates:**

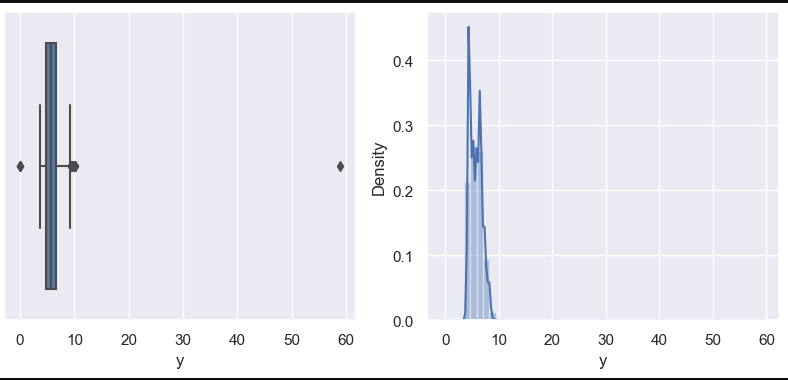
There are no duplicate values in the dataset.

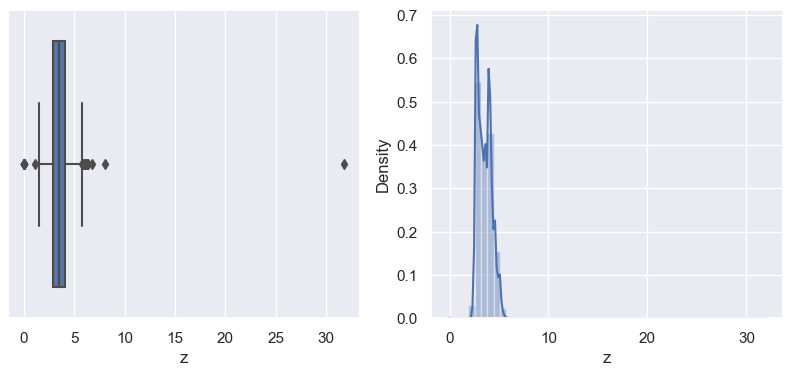
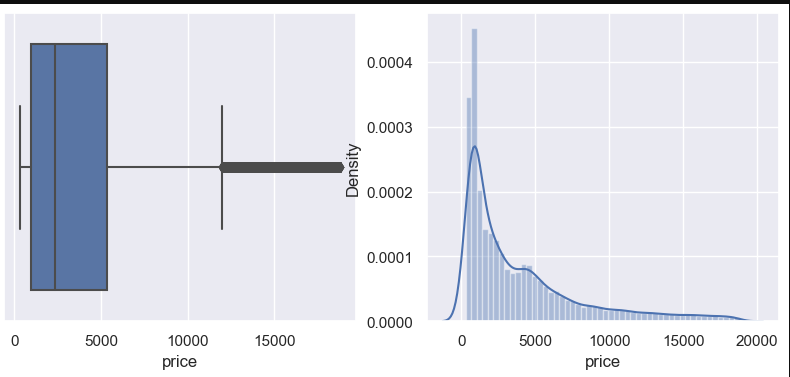
# Univariate Analysis categorical variables

****

# Univariate Analysis of Numeric Variables

****

****

** **

**Based on above output:**

* We can say that the ‘carat’ parameter is right skewed and outliers are present.
* The ‘depth ‘parameter seems to be normal distribution
* we can observe say that the ‘table ‘parameter is right skewed
* The ‘X ’parameter seems to be right skewed.
* The ‘Y’,’Z’ and ‘price’ parameters are highly right skewed.
* There are outliers present in the entire numeric variables.

## Outlier Checks & Treatment:

## The outliers are present in the numeric variables. So, we have to treat the outliers for performing the regression model.

## Here we can do the capping and flooring to treat outliers.

## Before treat the outlier:

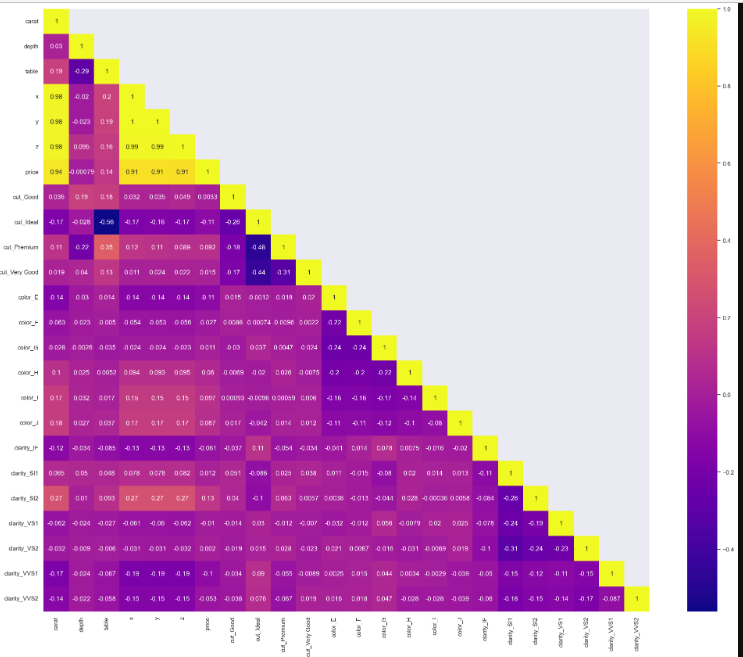
## 

## After outlier treatment:

# 

# Bi-Variate Analysis - Numeric Features

**Correlation Plot**

****

## Pair plot:

## 

* It involves the analysis of two variables , for the purpose of determining the empirical relationship between them.
* It can be inferred that most features correlate with the price of Diamond. The notable exception is "depth" which has a negligible correlation.

**INSIGHTS BASED ON EDA**:

* 'Price' is the target variables while all others are the predictors.
* The data set contains 26967 row, 11 column.
* In the given data set there are 2 Integer type features, 6 Float type features and 3 Object type features. Where 'price' is the target variable and all other are predictor variable.
* The first column is an index ("Unnamed: 0") as this only serial no, we can remove it from data because it is not useful for the prediction.
* On the given data set the mean and median values does not have much difference. We can observe Min value of "x", "y", "z" are zero this indicates that they are faulty values. As we know dimensionless or 2-dimensional diamonds are not possible.
* So we have filter out those as it clearly faulty data entries. There are three object data type 'cut', 'colour' and 'clarity'.
* We can observe there are 697 missing value in the depth column.
* There are significant amount of outlier present in all numeric variables, which will affect the outcome of our regression model. So we have treated the outlier. We can see that the distribution of some quantitative features like "carat" and the target feature "price" are heavily "right-skewed".

**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.**

## Check for Missing Values:

## 

## Imputing missing values:

* Based on above output shows that there are a total of 697 null values in the depth column.
* Followed by which the median is computed for each attribute so that it can be used to replace the null values that are present in the dataset.
* In below given output shows we can see that the null values are replaced by the median that's computed.
* After the removing the null values the shape of the dataset becomes 26925 rows and 10 columns.

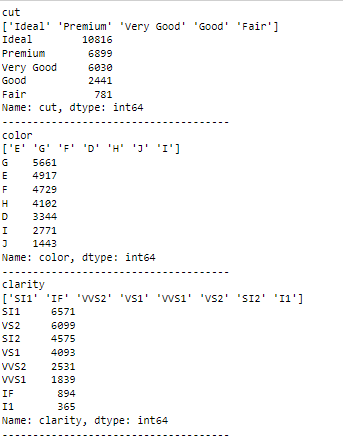
## 

**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.**

**Converting categorical to dummy variables (Encoding:**

In order to performing the liner regression model we need to convert the string values into numeric values. For this, here I used one hot encoding method for object data type variables.

**Before encoding:**



## 

## After encoding:

## 

## 

## 

## TRAIN AND SPLIT:

## Copied all the predictor variables into X data frame and copy target variable into the Y data frame. Using the dependent variable we spilt the X and Y into training set and test set.

## For this we use the sklearn package and then split X and Y in 70:30 ration and then involve the linear regression function and find the best fit model on training data.

The intercept for this model is -2944.74355583

## The Intercept is the expected mean value of y when x=0, and when x is not equal to zero then the intercept has no intrinsic meaning.

## In the present case when the other predictor variable is zero i.e. like carat, cut, colour, clarity then c=-2944,which means that the price is -2944 which doesn’t make any sense so in order to deal with this we have to carry out z score and make it nearly zero.

**R square on training data:** 0.9403068602256325

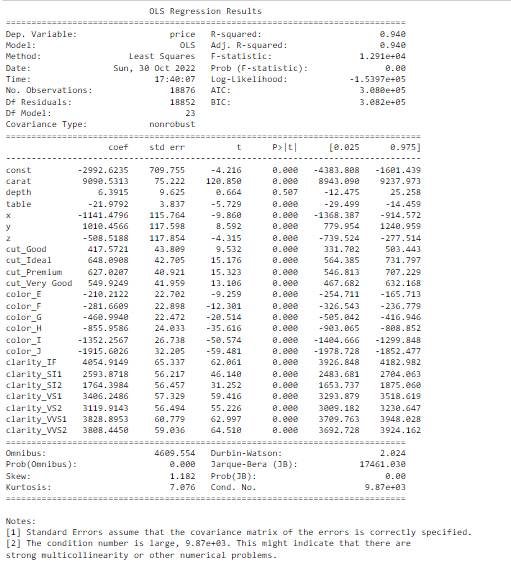
**R square on testing data:** 0.9415168542142924

* R square is always between 0 and 100%, in which 0% indicates that the model explains none of the variability of the response data around its mean and 100% indicates that the model explains all the variability of the response data around its mean.

In the linear regression model we can see the R-square value on training and test data respectively as 0.9403068602256325 and 0.9415168542142924.

**Linear Regression using stats models**

* Assuming the null hypothesis is true, i.e. price from that universe we have drawn co-efficient.
* As we see here the overall P value is less than alpha, so rejecting HO and accepting Ha that at least 1 regression co-efficient is not '0'. Here all regression co-efficient are not '0'.
* For example, below output we can see the p value is showing 0.507 for 'depth' variable, which is much higher than 0.05. That means this dimension is of no use. So we can say that the attribute which are having p value greater than 0.05 are poor predictor for price.



Root Mean Squared Error (Training) ------RMSE: 843.7765260400255

Root Mean Squared Error (lest) ------------RMSE: 848.96331995

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

**Inference:**

* There is a strong correlation between predicted variable and actual value. But there is lots of spread.

**Linear regression Performance Metrics:**

Intercept for the model: -2992.62348525 R square on training data: 0.9403068602256325

R square on testing data0.9415168542142924RMSE on Training data: 843.7765260400255 RMSE on Testing data: 848.96331995 as the training data & testing data score are almost inline, we can conclude this model is a Right-Fit Model.

**From statsmodels:**

* The overall P value is less than alpha.
* Finally we can conclude that Best 5 attributes that are most important are 'Carat', 'Cut', 'colour', clarity' and width i.e. 'y' for predicting the price.
* When 'carat', ‘cut’, 'colour’ , 'clarity' and 'y' increases by 1 unit, diamond price also increases, keeping all other predictors constant.
* We can see that the p value is 0.507 for depth variable, which is much greater than 0.05. That means this attribute is of no use.
* There are also some negative co-efficient values, we can see the 'X' i.e Length of the cubic zirconia in mm. having negative co-efficient -1417.9089. And the p value is less than 0.05, so can conclude that as higher the length of the stone is a lower profitable stones.
* Similarly for the 'z' variable having negative co-efficient i.e. -711.23. And the p value is less than 0.05, so we can conclude that as higher the 'z' of the stone is a lower profitable stones.

**Recommendations:**

* Based on liner regression model the Gem Stones company should consider the features 'Carat', 'Cut', 'colour', 'clarity' and width i.e. 'y' as most important for predicting the price. To distinguish between higher profitable stones and lower profitable stones so as to have better profit share.
* The 'Premium Cut' on Diamonds are the most Expensive, followed by 'Very Good' Cut, these should consider in higher profitable stones.
* The Diamonds clarity with 'VS1' &'VS2' are the most expensive. So these two categories also consider in higher profitable stones.
* As we see for 'X' i.e. Length Of the stone, higher the length of the stone is lower the price.
* Stones with higher 'z' is are lower in profitability.

**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.

**Dataset for Problem 2:** [Holiday\_Package.csv](https://olympus.mygreatlearning.com/courses/78182/files/6689409/download?verifier=HKFo9MOQn4j80q8hGX1OXljdOXmpDyw3nJFGR6eF&wrap=1)

**Data Dictionary:**

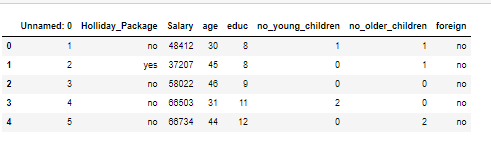
|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Holiday\_Package | Opted for Holiday Package yes/no? |
| Salary | Employee salary |
| age | Age in years |
| edu | Years of formal education |
| no\_young\_children | The number of young children (younger than 7 years) |
| no\_older\_children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

**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.**

**Load the data:**

**EDA:**

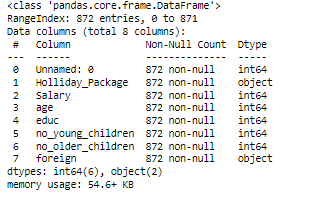
**Head:**

****

**Data shape:**

(872, 8)

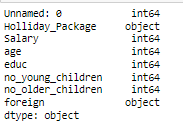
**Data info:**

****

Based on above output this data set has 8 variables and 871 observations.

Based on above output there is no null values in data set.

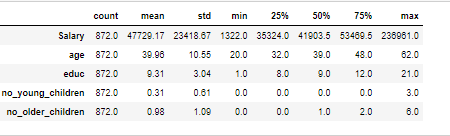
**Data types:**



**Duplicates:**

There are no duplicates in the dataset.

**Statistical Summary of Numeric Features:**

****

The above table shows 5 point summary of the data.

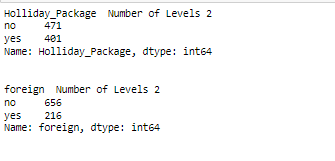
Based on output the data looks good.

All the variables mean and median are nearly equal.

Before going to the further analysis, we confirmed if we have any duplicate values in the data.

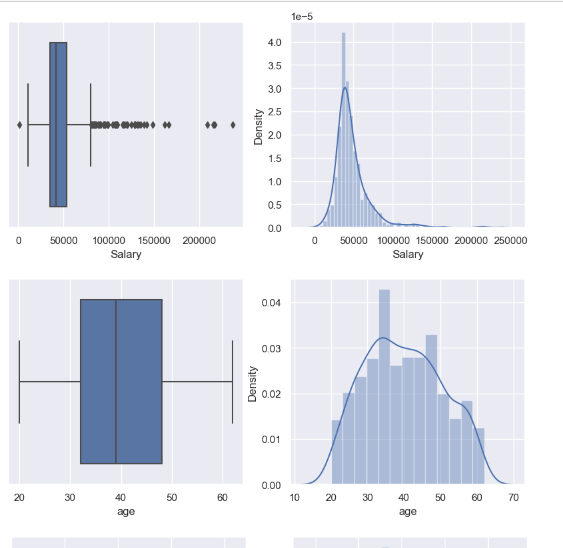
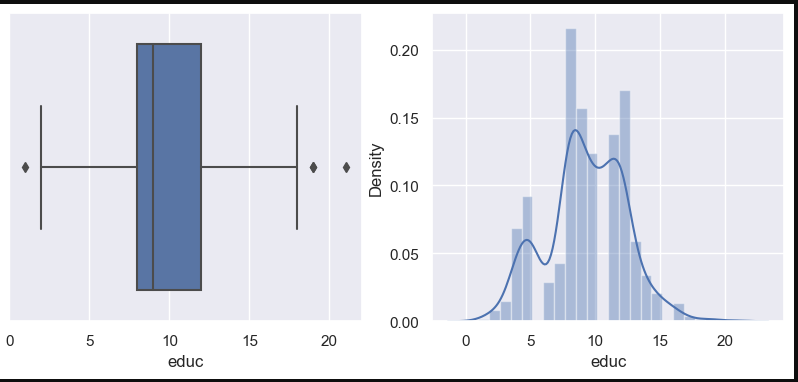
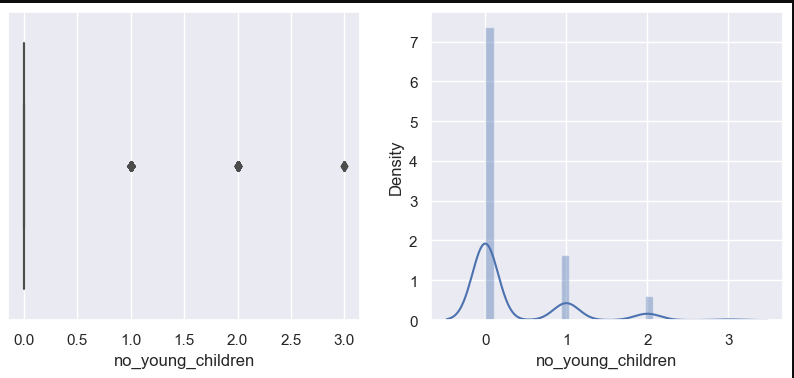
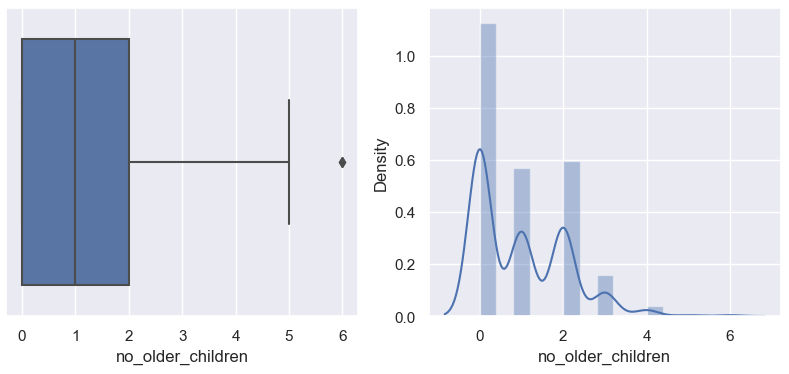
# Univariate Analysis

**Unique values and value counts of the categorical columns:**

****

**Univariate Analysis of Numeric Variables**

**Box plot & density plot of each numerical column as a subplot:**

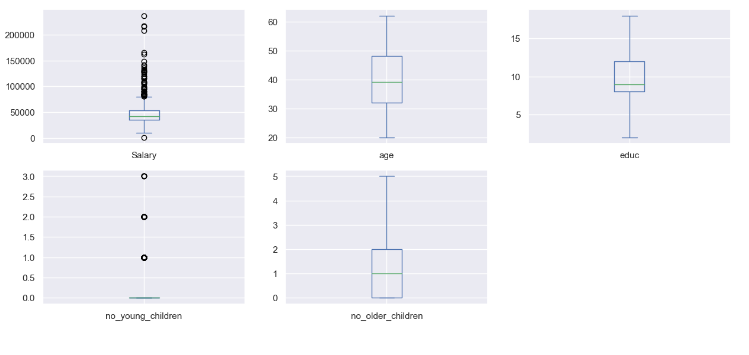
**   **

* We can see that most of the distributions are right skewed except for ‘educ’ variable.
* Salary distribution has the max no of outliers
* There are some outliers in ‘educ’ , ‘no of young children’ and ‘no. of older children’.

**Treating the outliers:**

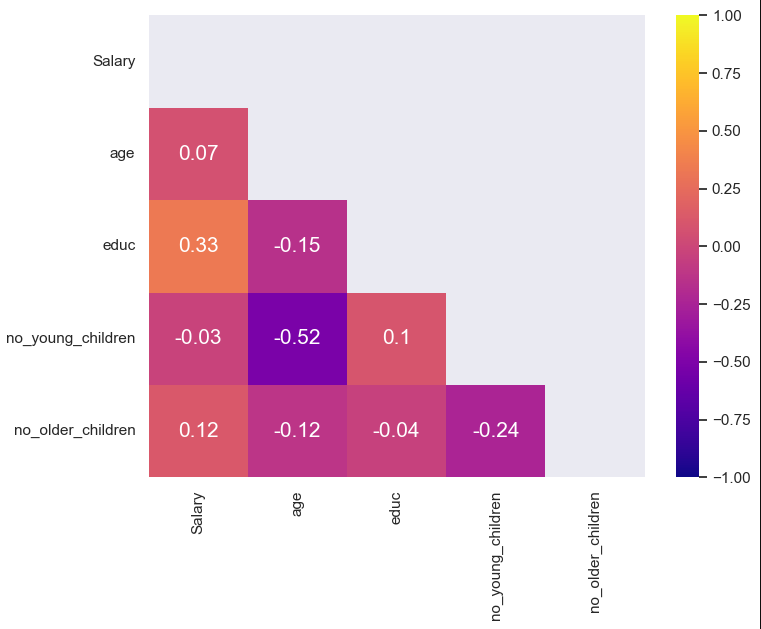
* We can treat Outliers with the following code.
* We will treat the outliers for the 'edu' and 'no\_older\_children' variables only.
* ‘No\_young\_children’ variable has more than 75% values are zero, so capping the outliers would mean converting all non-zero values to zero.
* Also, for salary, these seem to be genuine values and not error values reflecting as outliers.

**After outliers treatment:**

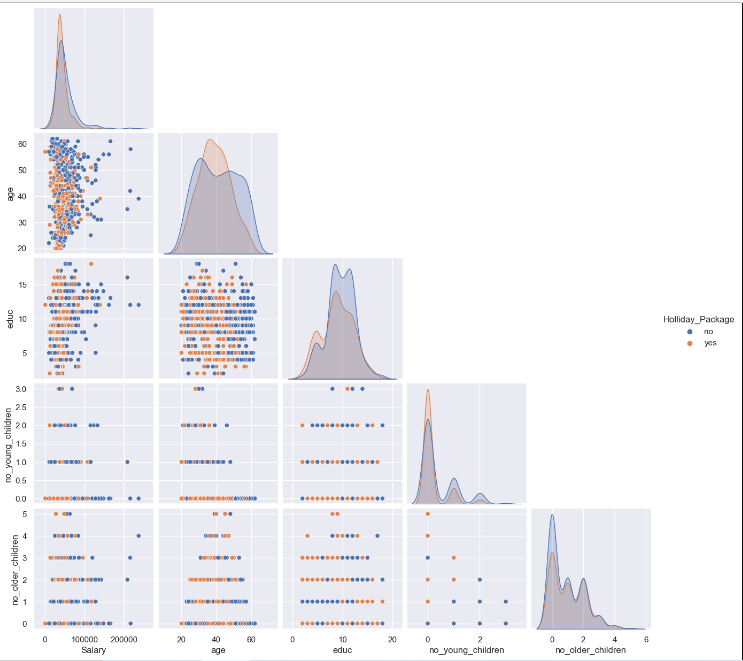
****

# Bivariate Analysis:

**Numeric Features - Checking for Correlations:**

****

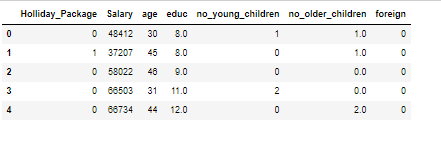
**Pair plot**

****

* There is no correlation between the data, the data seems to be normal.
* There isno huge difference in the data distribution among the holiday package.
* I don’t see any clear two different distributions in the dataset provided.

**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).**

**Encoding Categorical Feature Labels:**

****

Here we have done ONE HOT ENCODING to create dummy variables for the ‘foreign’ variable and I have done label encoding for ‘holiday\_package’ variable.

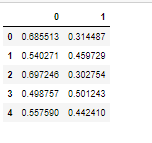
Better results are predicted by logistic regression model if encoding is done.

**Train Test Split:**

We will split the data in 70/30 ratio.

# Logistic Regression Model:

**Class Probability Prediction**

****

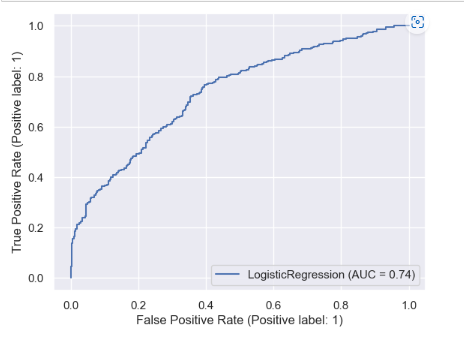
# Linear Discriminant Analysis:

**Accuracy score:** 0.6639344262295082

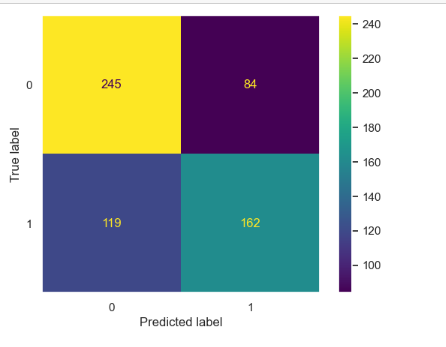
**2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.**

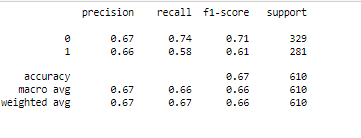
**Logistic regression:**

**Training Data**

****

**AUC SCORE 0.74**

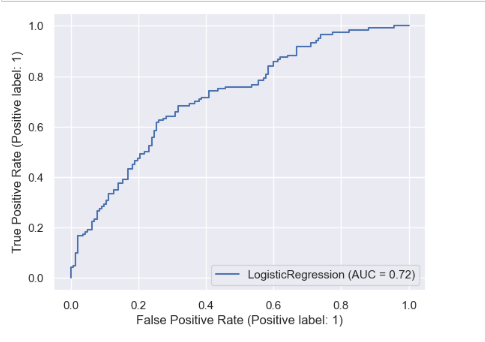
****

****

Here we can see that precision for 1 is 0.66, recall is 0.58 accuracy is 0.67 and f1 score is 0.61

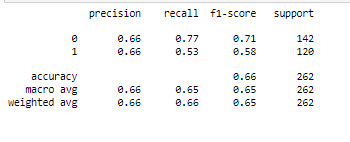
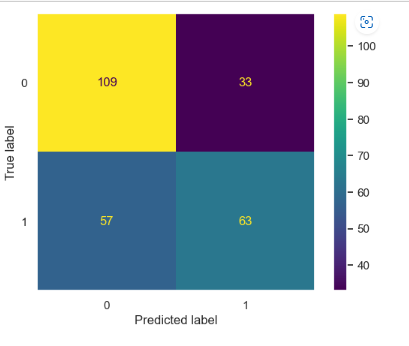
**Test Data:**

**AUC and ROC**

****

**AUC SCORE 0.72**

**Confusion Matrix & Classification Report Metrics**

****

Here we see that precision for 1 is 0.66, recall is 0.53accuracy is 0.66 and f1 score is 0.58.

# Linear Discriminant Analysis:

**Model Evaluation**

## Training Data:

## 

## AUC SCORE 0.73

## 

## 

Here we see that precision for 1 is 0.65 , recall is 0.58 accuracy is 0.66 and f1 score is 0.61

## Test Data:

## 

## AUC score 0.72

## Confusion Matrix & Classification Report Metrics

## 

## 

Here we see that precision for 1 is 0.65 , recall is 0.50 accuracy is 0.65 and f1 score is 0.56.

* Comparing both these models, we find both results are same, but LDA works better when there is category target variable.  
  As we can see the results for AUC/ROC for both the models are almost equivalent to each other.  
  So it is very difficult to differentiate between the two.
* The scores are also almost at par with each other.
* Both the models are working perfectly at par with each other.  
  Since LDA works better with categorical values so we will pick it in this situation.

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

* So we had been given a problem where we had to find out whether the employees will opt for a  
  holiday package or not.  
  We looked in the data using logistic regression and LDA.
* Predictions were done using both the models.  
  Employees who are in the age gap of 30 to 50 opt for holiday packages.
* It seems like young people believe I spending on holiday packages so age here plays a very  
  important role in deciding whether they will opt for package or not.
* Also people who have salary less than 50000 opt for holiday packages. So salary  
  is also a deciding factor for the holiday package.

**Recommendations**

Based on the model predictions, we should focus on the age group of 30 to 50 as well as salary more than 150000. We need to look for good options and attractive schemes in order to opt holiday package.  
We can also implement family package system. In this way we can attract all age group people to buy holiday package.

We can provide special customized packages for people who are earning more than 150000.