**Business Report**

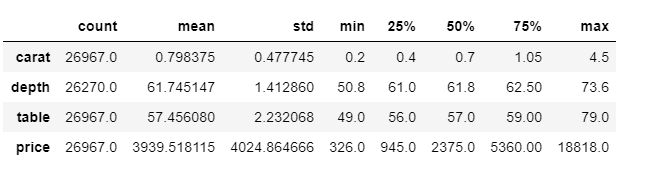
**Predictive Modeling Project**

**Problem 1: Linear Regression**

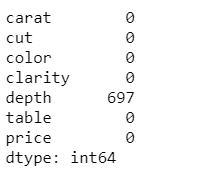
You are hired by a company Gem Stones co ltd, which is a cubic zirconia manufacturer. You are provided with the dataset containing the prices and other attributes of almost 27,000 cubic zirconia (which is an inexpensive diamond alternative with many of the same qualities as a diamond). The company is 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.

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

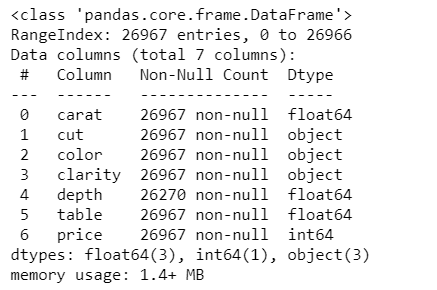
Data description –



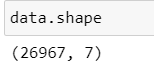
Null values –



Data types –

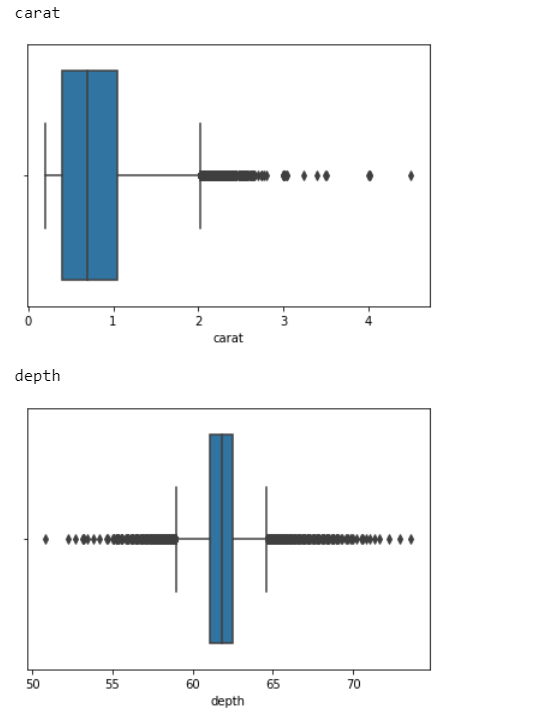


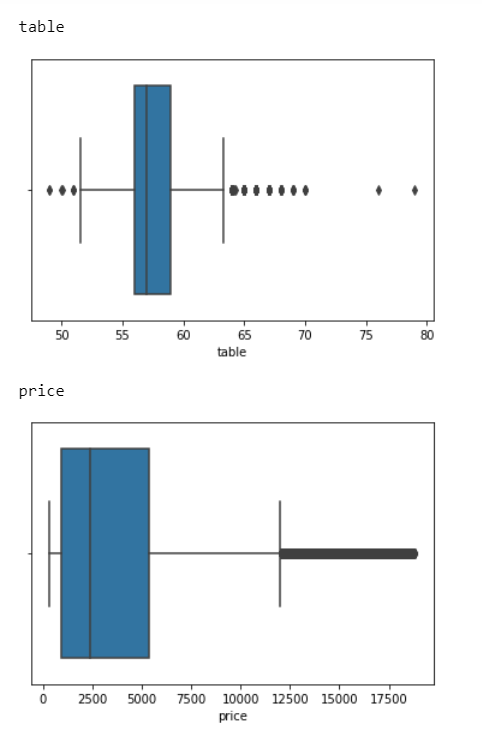
Shape –



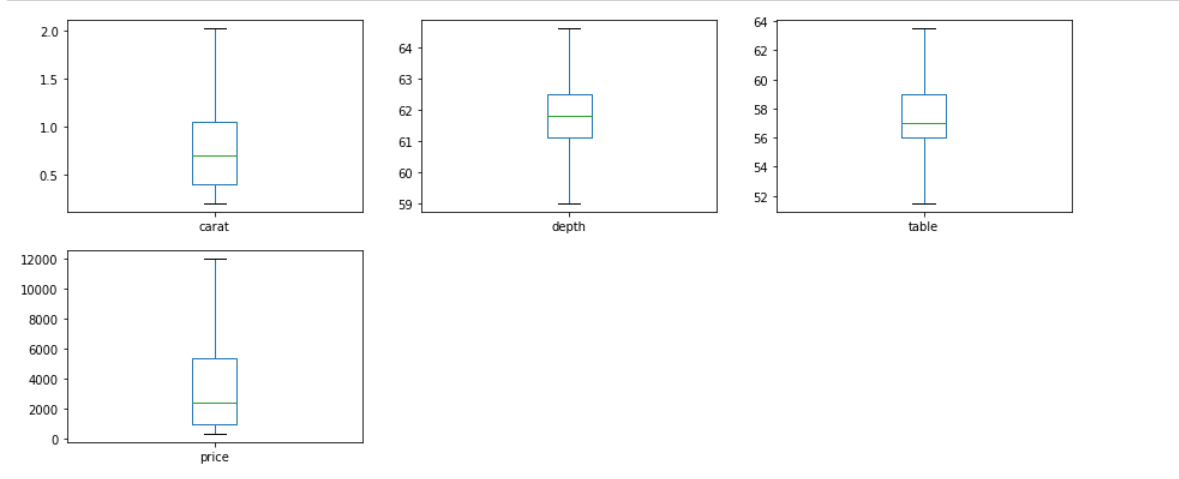
Checking for Outliers -

Before treating outliers

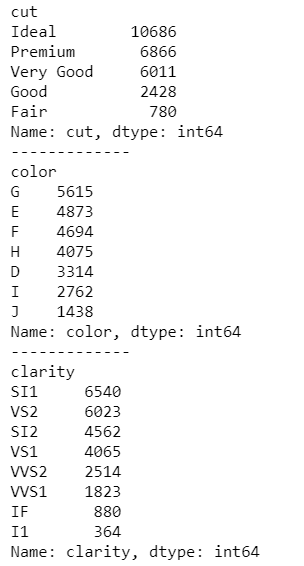




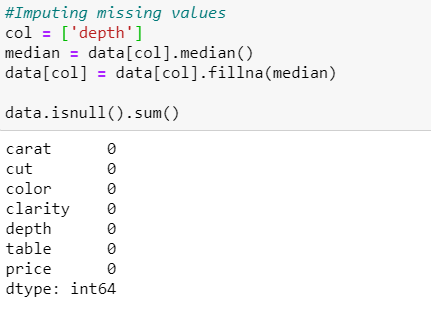
After treating outliers



Value counts of object variable

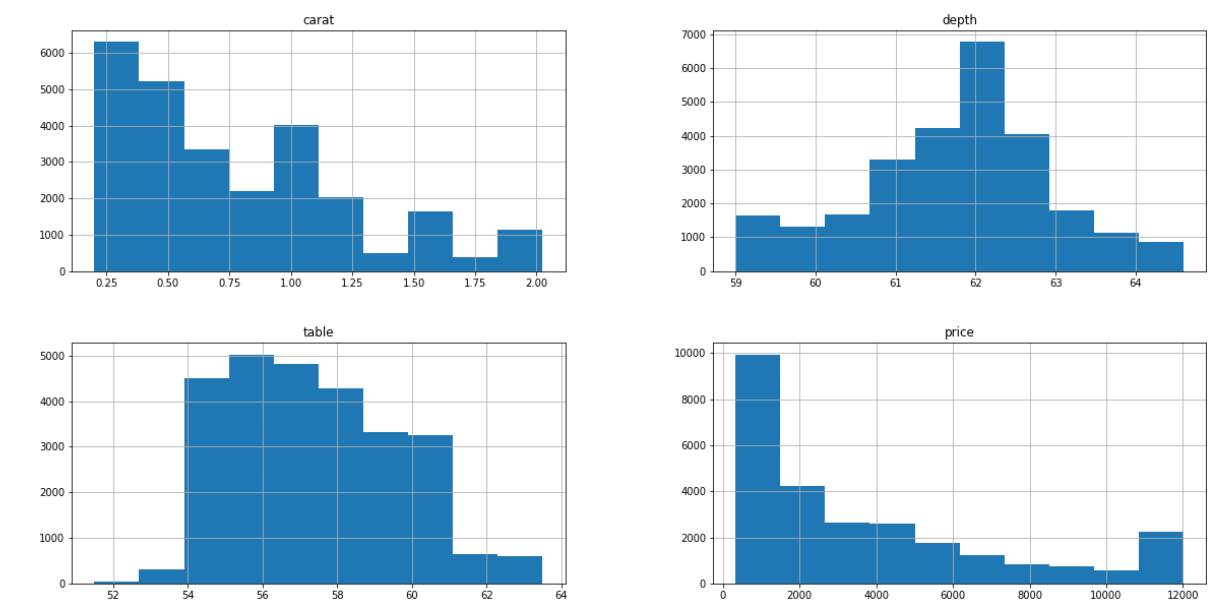


**1.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Do you think scaling is necessary in this case?**

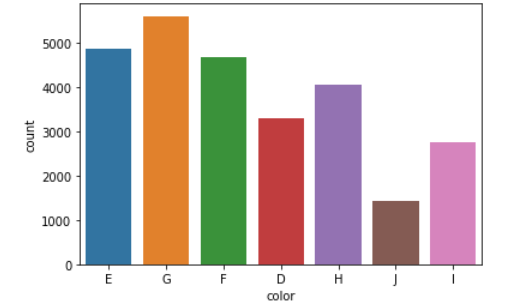


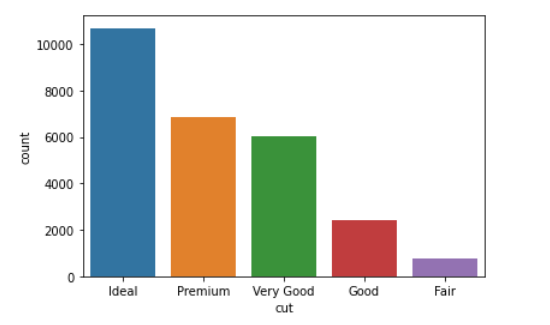
Univariate analysis –

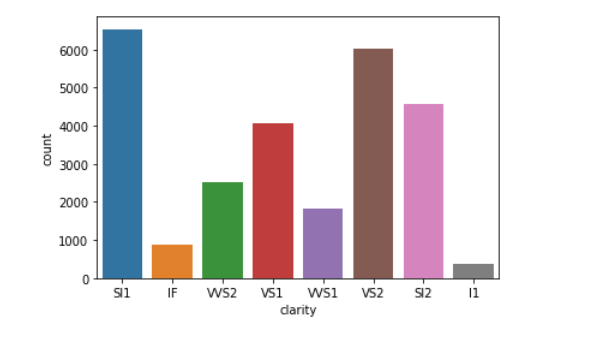
Numerical variables –



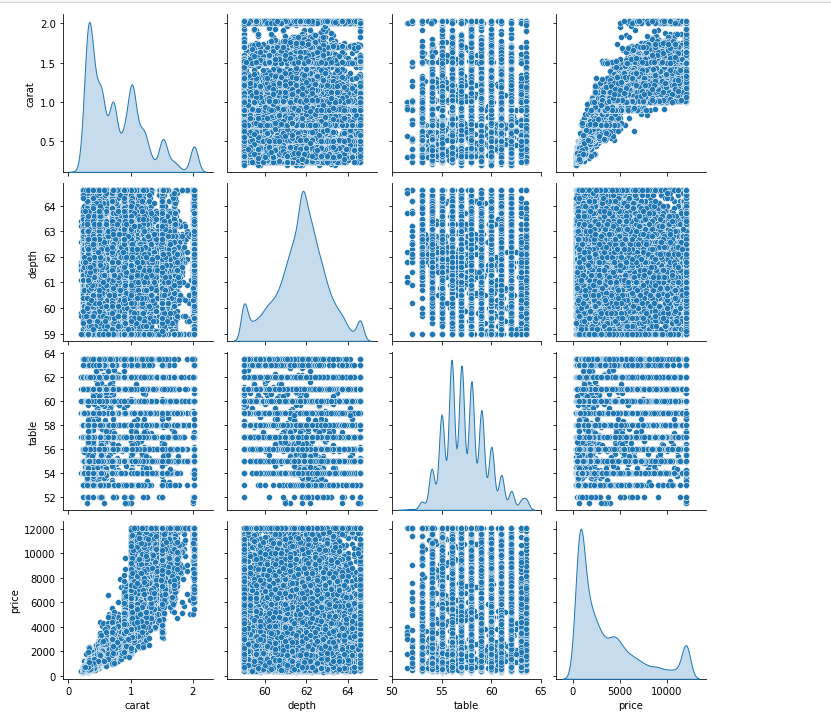
Categorical variables –







Bivariate Analysis –

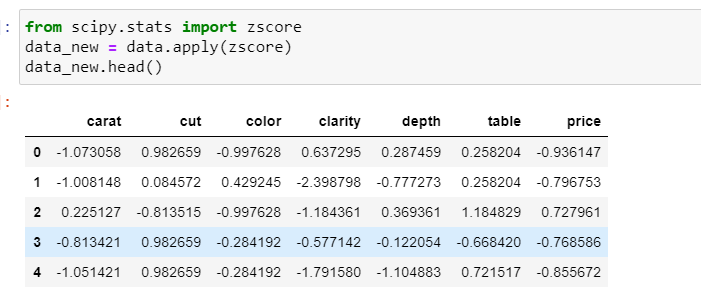


Heat map –

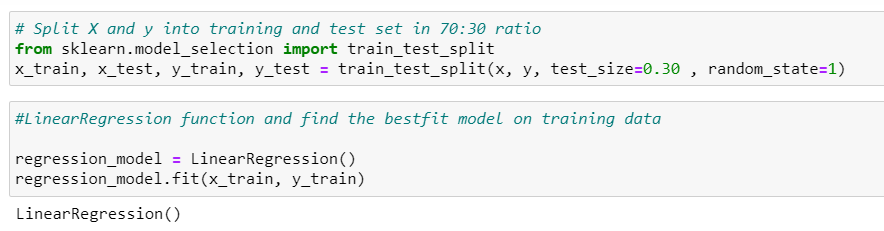


Carat is highly correlated to price

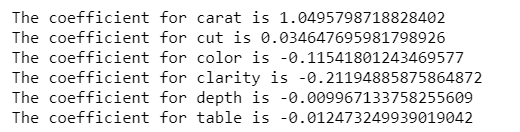
Yes scaling is necessary in this data



**1.3 Encode the data (having string values) for Modelling. Data Split: Split the data into test and train (70:30). Apply Linear regression. Performance Metrics: Check the performance of Predictions on Train and Test sets using Rsquare, RMSE.**



Coeefficient of the independent attributes



The intercept for our model is -7.640063035874738e-05

R-Square value –

Training data: 0.9267507660474659

Testing data: 0.9256625149886726

RMSE Value -

Training data: 0.2710629878500264

Testing data: 0.27166632333949225

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

**Carat is the most significant variable for predicting price.**

VIF of all the variables are below 5

carat ---> 1.2822159939702893

cut ---> 1.4915904035963454

color ---> 1.1020010636387867

clarity ---> 1.2012771184065307

depth ---> 1.3228997907234574

table ---> 1.577822870641595

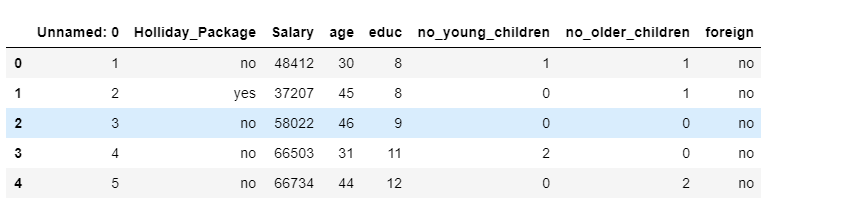
Below is the linear equation-

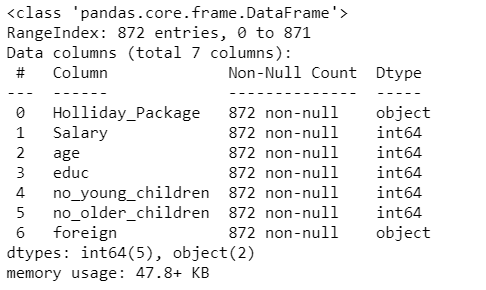
(0.0) \* Intercept + (1.01) \* carat + (0.03) \* cut + (-0.19) \* clarity + (-0.02) \* depth + (-0.01) \* table

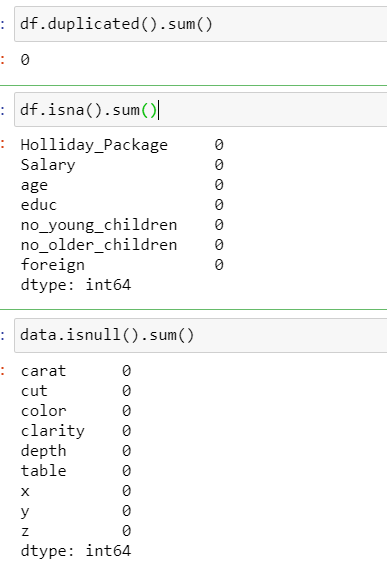
**Problem 2: Logistic Regression and LDA**

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.

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.

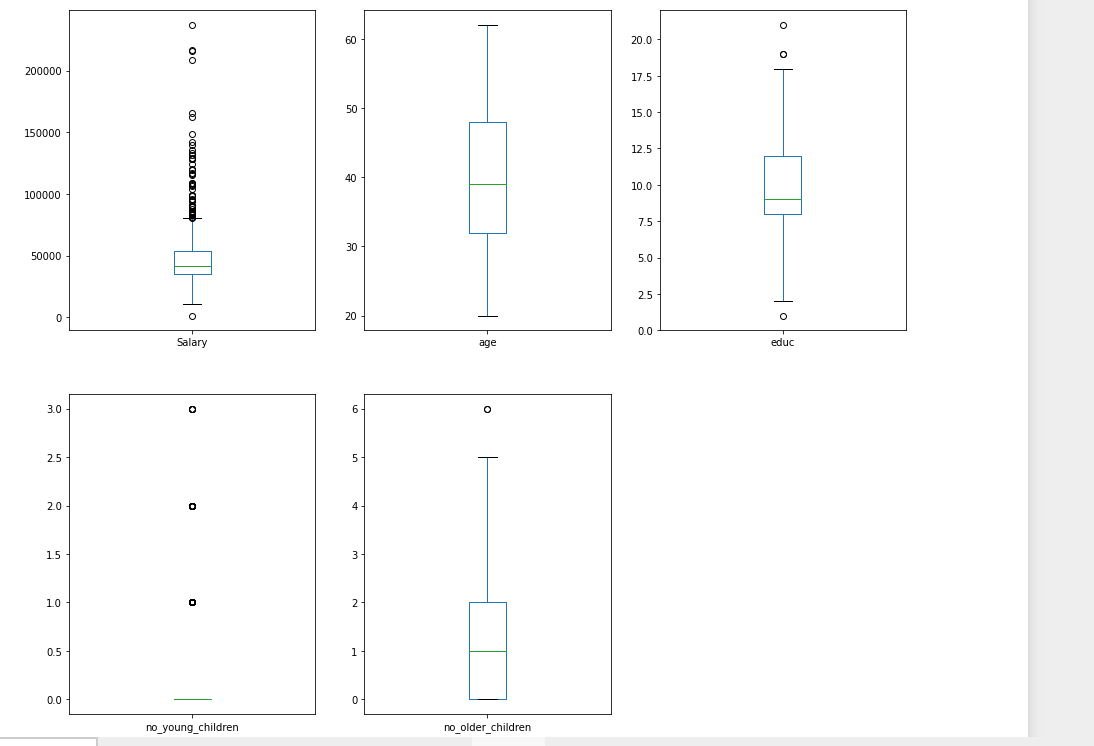




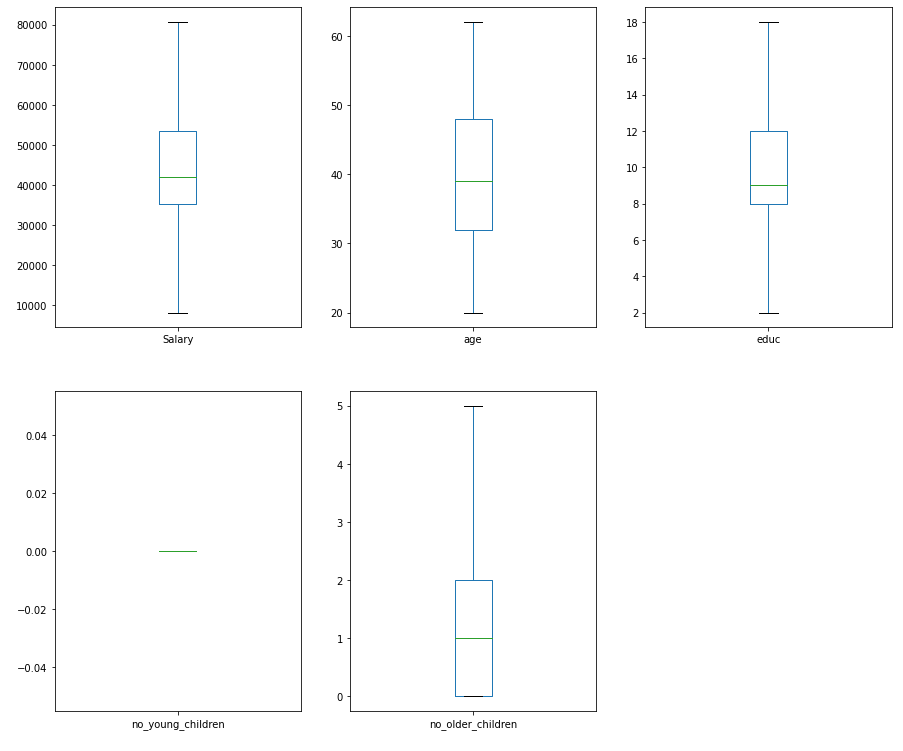


No duplicate or null values

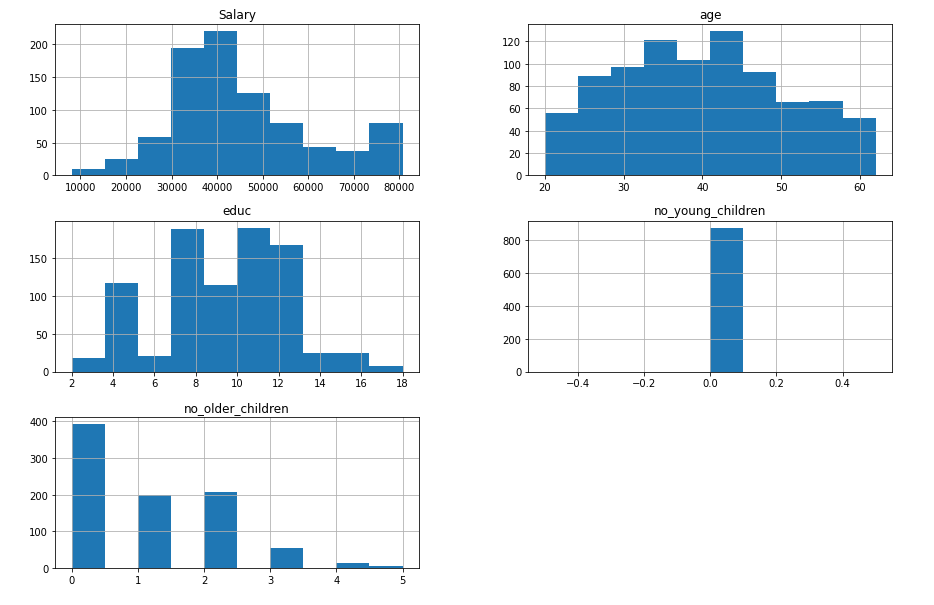
Before outlier treatment



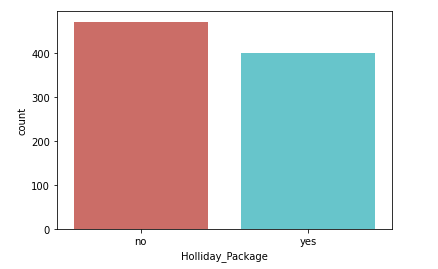
After outlier treatment

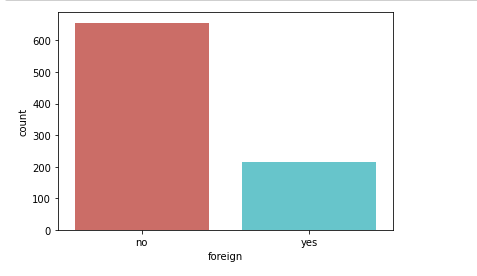


Univariate analysis



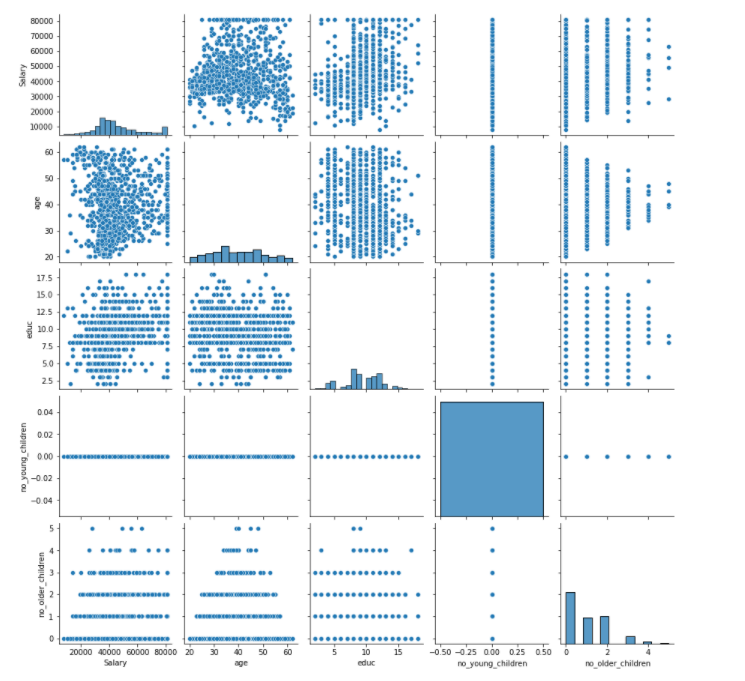
Categorical variables



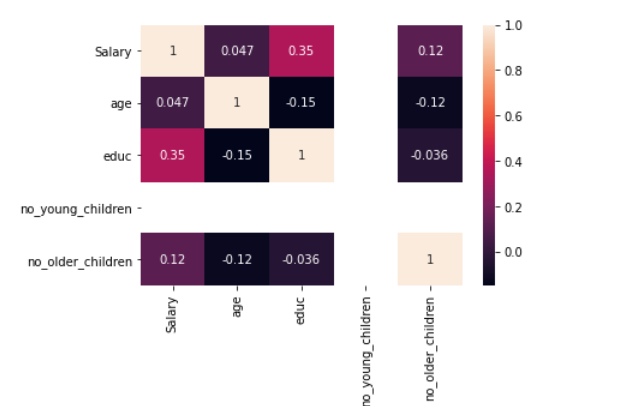


Bivariate analysis

Pair plot

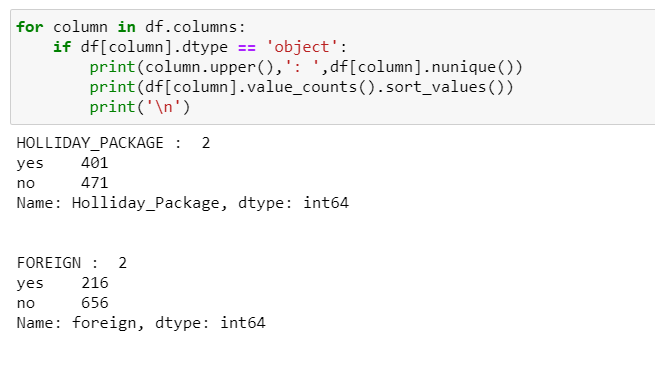


Heat map

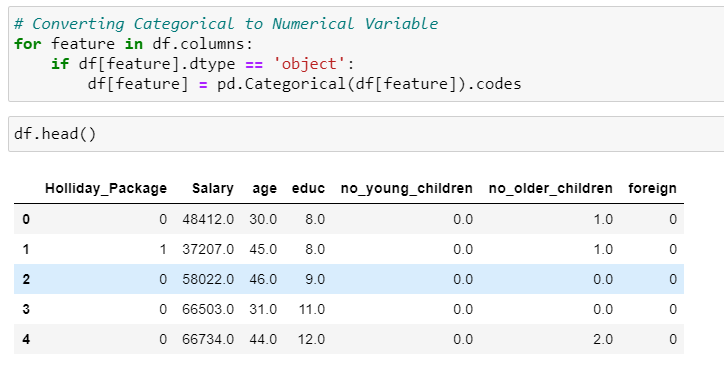


None of the variables are highly correlated

Value counts of Categorical variables



Converting Categorical variables into Numerical



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).



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.

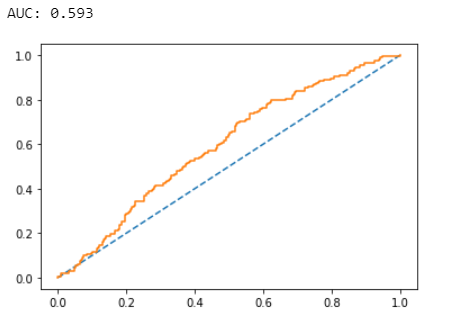
Accuracy –

Training Data: 0.5393442622950819

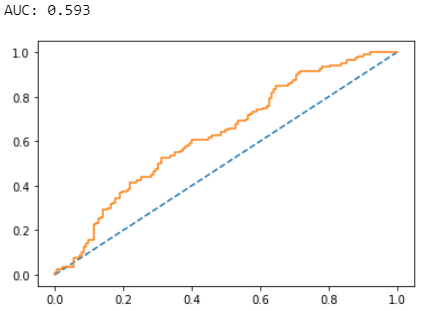
Testing Data: 0.5419847328244275

AUC and ROC:

Training data:

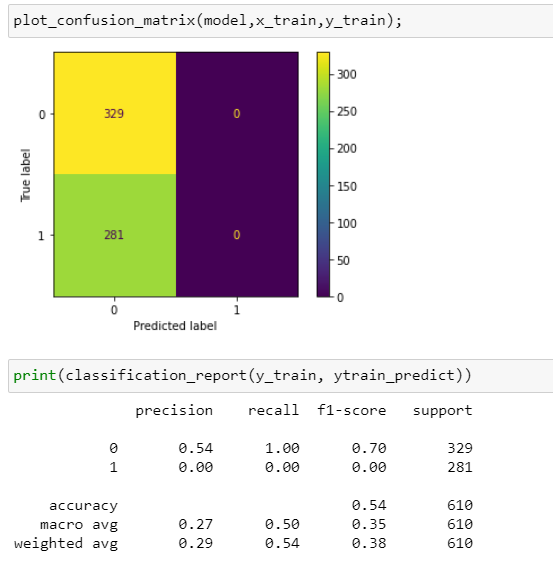


Testing data:

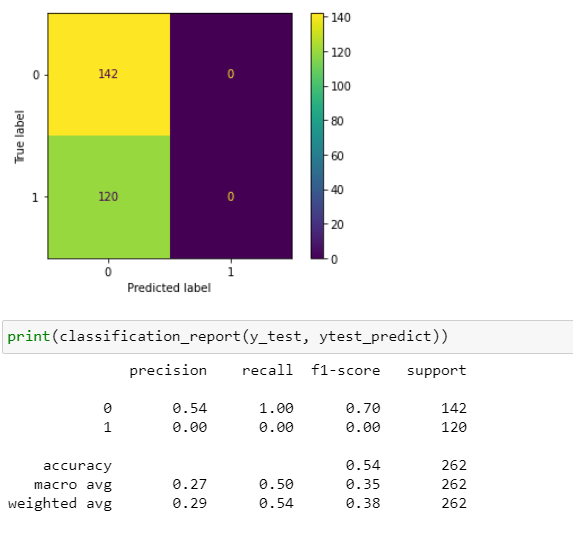


Confusion Matrix and Classification Report

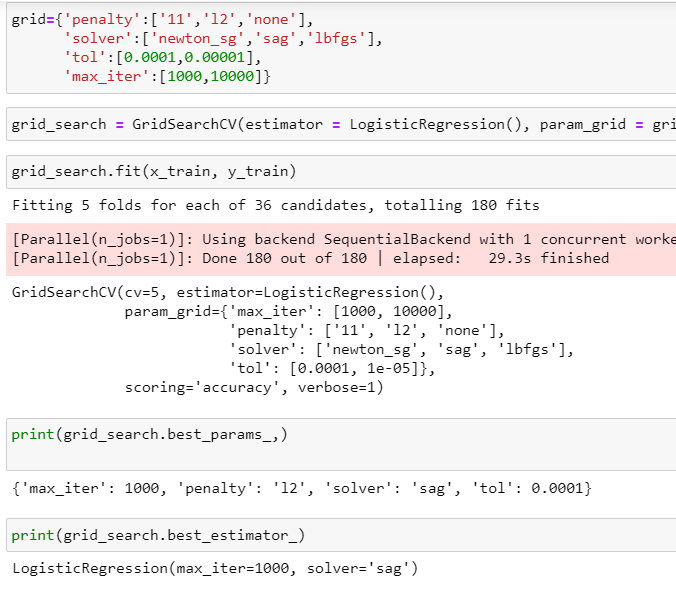
Training data:



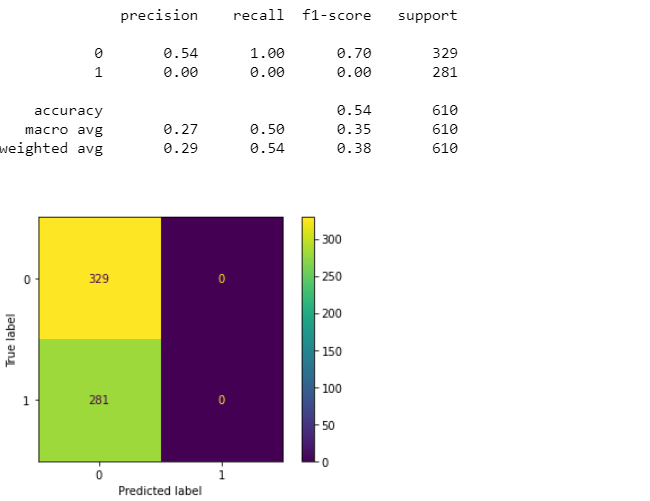
Testing data:



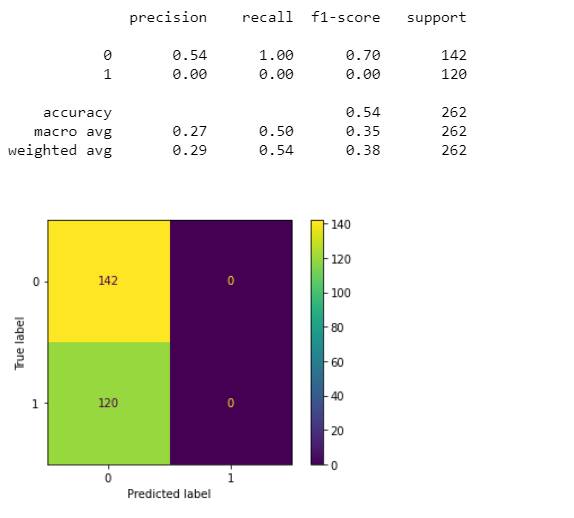
Applying GridSearchCV for Logistic Regression



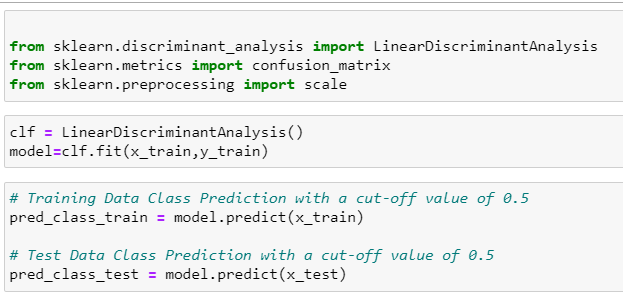
Confusion matrix and Classification report on the training data



Confusion matrix and Classification report on the testing data



Linear Discriminant Analysis



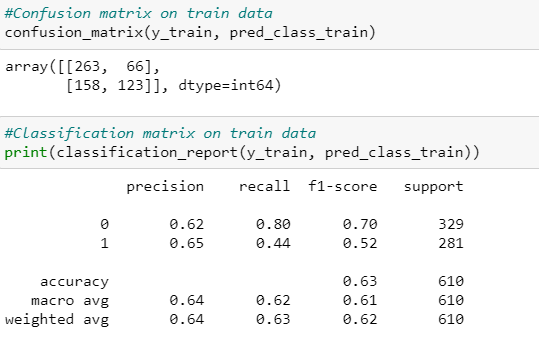
Accuracy –

Training data: 0.6327868852459017

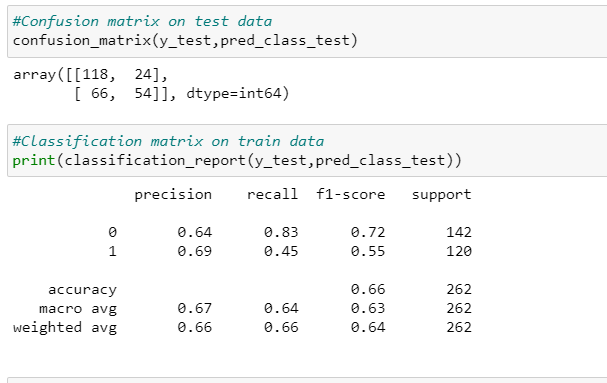
Testing data: 0.6564885496183206

Confusion matrix and Classification report -

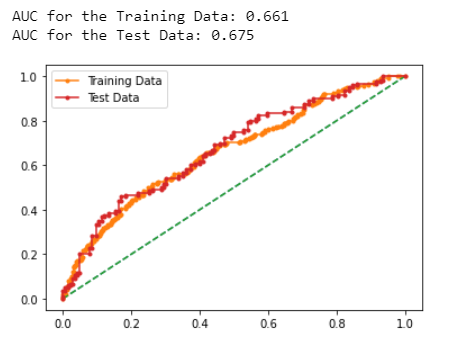
Training data



Testing data



AUC and ROC for training and testing data –



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

Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

Since accuracy score of LDA is better than Logistics regression , therefore LDA is a better model in this case