

PProject Report- Predictive Modeling



June 19, 2022

**Sharjil Shah**

Great Learning, G11 Jan\_22A

Table of Contents

[Data Dictionary: 3](#_Toc106695461)

[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. 4](#_Toc106695462)

[Head of data: 4](#_Toc106695463)

[Tail of data: 4](#_Toc106695464)

[Shape: 4](#_Toc106695465)

[Description of data: 5](#_Toc106695466)

[Data Info: Dataset has int, float and object data types 5](#_Toc106695467)

[Univariate and Bivariate Analysis 5](#_Toc106695468)

[Multivariate Analysis 8](#_Toc106695469)

[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 an ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning. 9](#_Toc106695470)

[Checking if there is value that is “0” 10](#_Toc106695471)

[SCALING 10](#_Toc106695472)

[CHECKING THE OUTLIERS IN THE DATA 10](#_Toc106695473)

[1.2 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. 13](#_Toc106695474)

[R square on train data: 16](#_Toc106695475)

[R square on test data: 17](#_Toc106695476)

[RMSE on Training data: 17](#_Toc106695477)

[1.4 Inference: Basis on these predictions, what are the business insights and recommendations. Please explain and summarize the various steps performed in this project. There should be proper business interpretation and actionable insights present. 17](#_Toc106695478)

[Recommendations 18](#_Toc106695479)

[Data Dictionary 18](#_Toc106695480)

[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. 19](#_Toc106695481)

[Head of the data 19](#_Toc106695482)

[Info 19](#_Toc106695483)

[Describe 20](#_Toc106695484)

[Null Value Check 20](#_Toc106695485)

[Check for Duplicate values 20](#_Toc106695486)

[Univariate/Bivariate Analysis 21](#_Toc106695487)

[Multivariate Analysis 23](#_Toc106695488)

[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). 24](#_Toc106695489)

[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 24](#_Toc106695490)

[Confusion Matrix Train Variables for Logistic regression 25](#_Toc106695491)

[Logistic regression Classification report 25](#_Toc106695492)

[ROC curve for the model 26](#_Toc106695493)

[Confusion matrix train variables for LDA 26](#_Toc106695494)

[LDA classification report 27](#_Toc106695495)

[Roc curve for the model 27](#_Toc106695496)

[2.4 Inference: Basis on these predictions, what are the insights and recommendations. 29](#_Toc106695497)

[Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present. 29](#_Toc106695498)

[Recommendations : 29](#_Toc106695499)

Problem 1: Linear Regression

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

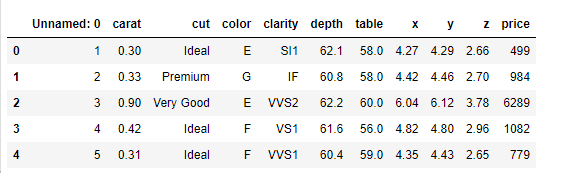
# Data Dictionary:

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

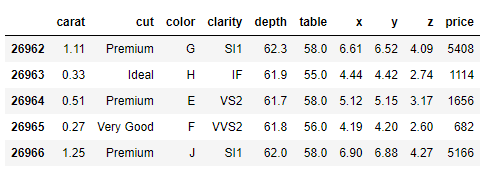
## Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA, duplicate values). Perform Univariate and Bivariate Analysis.

Ans: Checking if data has flown in properly:

### Head of data:



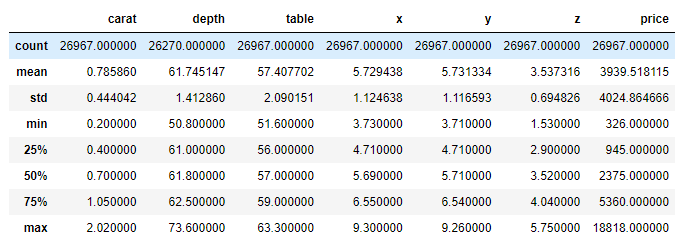
### Tail of data:



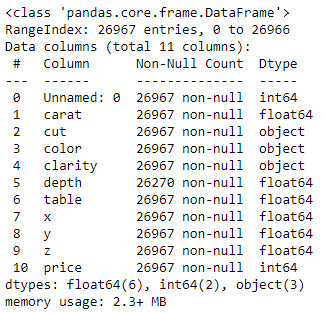
### Shape:

(26967, 10)

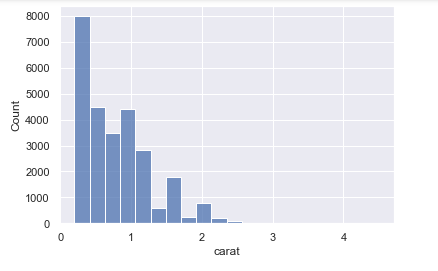
### Description of data:

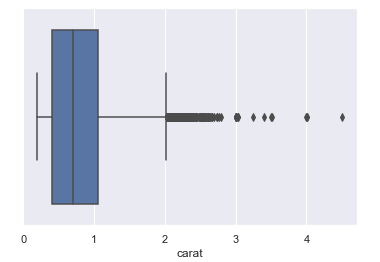


### Data Info: Dataset has int, float and object data types



### Univariate and Bivariate Analysis

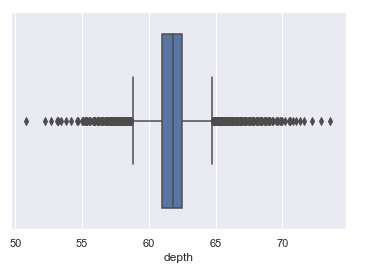
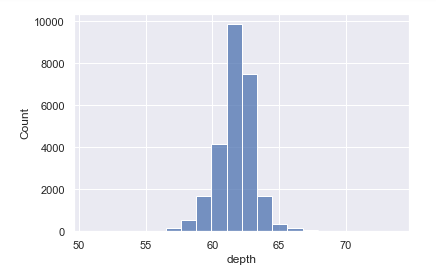




The distribution of data in carat seems to positively skewed, as there are multiple peaks points in

the distribution there could multimode and the box plot of carat seems to have large number of outliers.

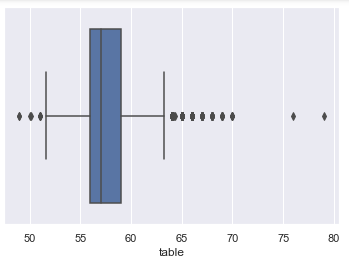
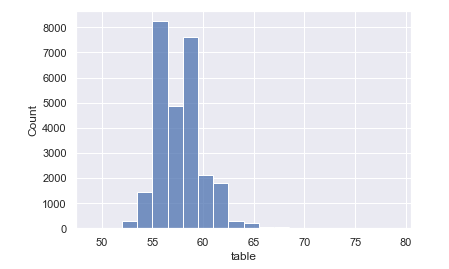
In the range of 0 to 1 where majority of data lies.



The distribution of depth seems to be normal distribution,

The depth ranges from 55 to 65

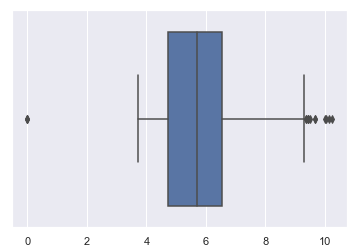
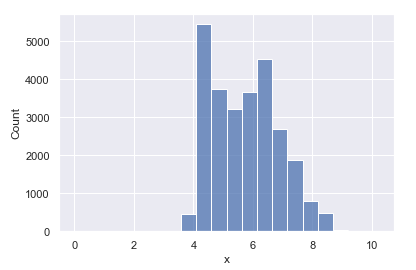
The box plot of the depth distribution holds many outliers.



The distribution of table also seems to be positively skewed

The box plot of table has outliers

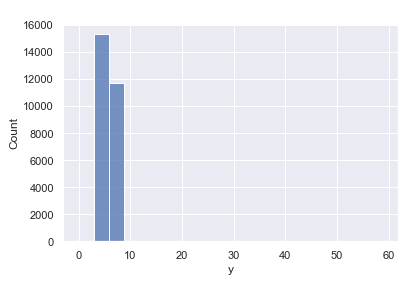
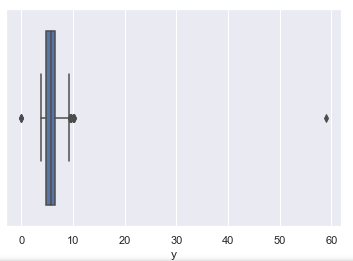
The data distribution where there is maximum distribution is between 55 to 65



The distribution of x (Length of the cubic zirconia in mm.) is positively skewed

The box plot of the data consists of many outliers

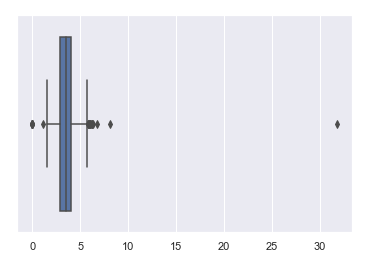
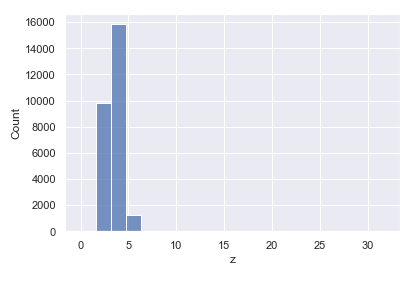
The distribution rages from 4 to 8

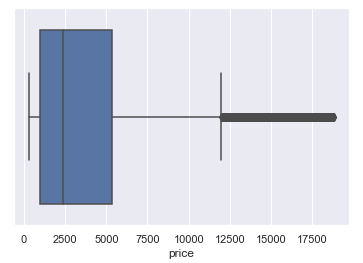
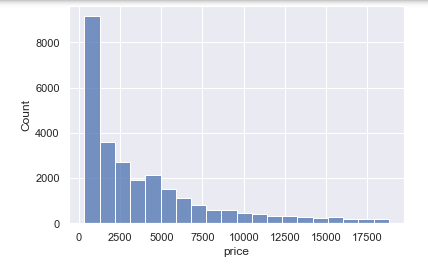
The distribution of Y (Width of the cubic zirconia in mm.) is positively skewed

The box plot also consists of outliers

The distribution too much positively skewed. The skewness may be due to the diamonds are always made in specific shape. There might not be too much sizes in the market

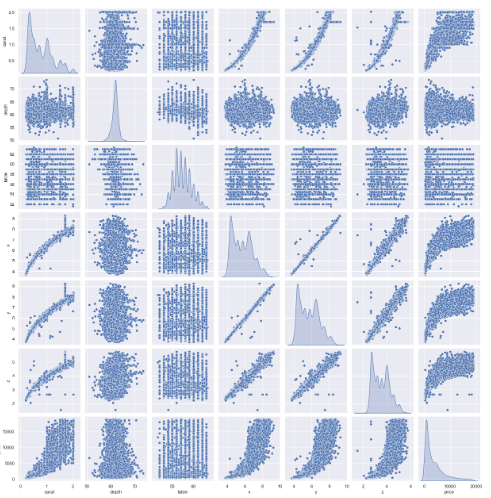


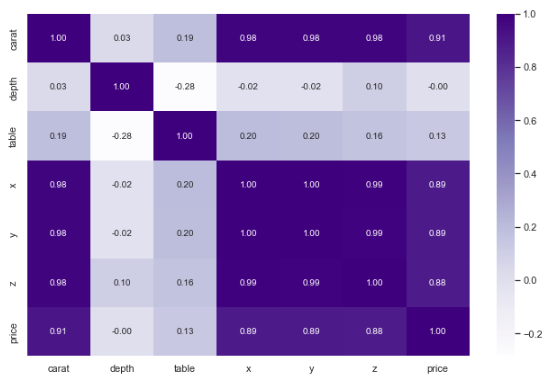
The distribution of z (Height of the cubic zirconia in mm.) is positively skewed The box plot also consists of outliers The distribution too much positively skewed. The skewness may be due to the diamonds are always made in specific shape. There might not be too much sizes in the market



The price has seems to be positively skewed. The skew is positive The price has outliers in the data The price distribution is from rs 100 to 8000.

### Multivariate Analysis

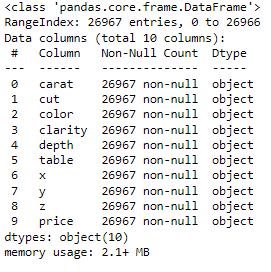




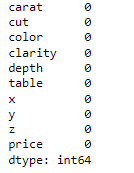
This matrix clearly shows the presence of multi collinearity in the dataset.

## 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 an ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning.

Ans: Based on the below, all columns except for depth has no null values.



Yes we have Null values in depth, since depth being continuous variable mean or median imputation can be done. The percentage of Null values is less than 5%, we can also drop these if we want. After median imputation, we don’t have any null values in the dataset.



### Checking if there is value that is “0”

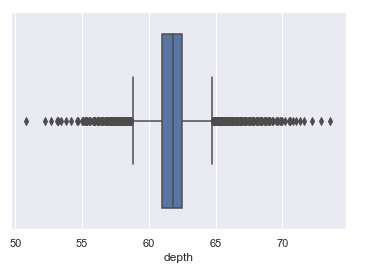
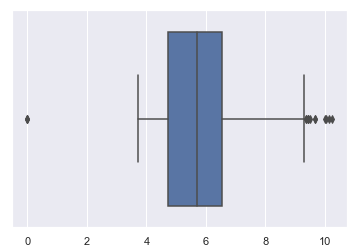
We have certain rows having values zero, the x, y, z are the dimensions of a diamond so this can’t take into model. As there are very less rows.

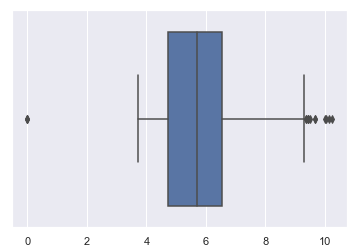
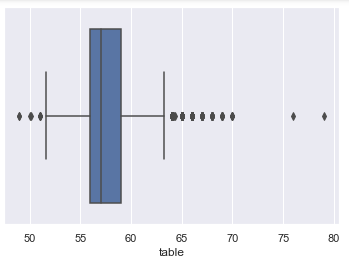
We can drop these rows as don’t have any meaning in model building

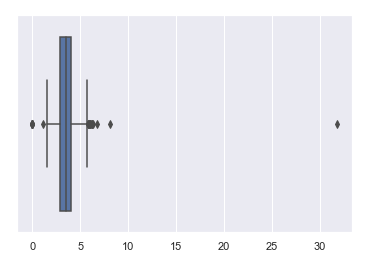
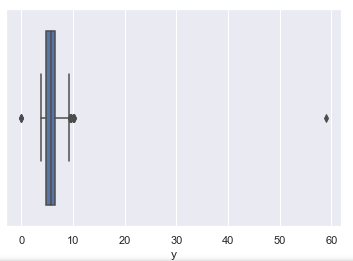
### SCALING

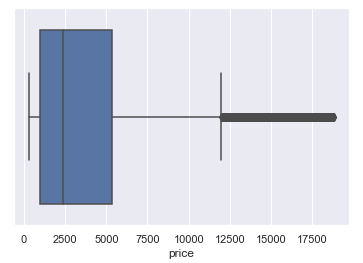
Scaling can be useful to reduce or check the multi collinearity in the data, so if scaling is not applied I find the VIF – variance inflation factor values very high. Which indicates presence of multi collinearity. These values are calculated after building the model of linear regression. To understand the multi collinearity in the model. The scaling had no impact in model score or coefficients of attributes nor the intercept.

### CHECKING THE OUTLIERS IN THE DATA

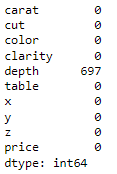
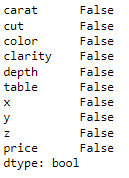


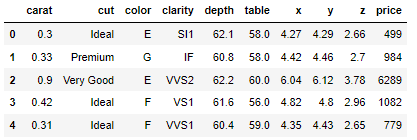


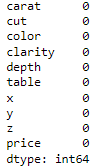
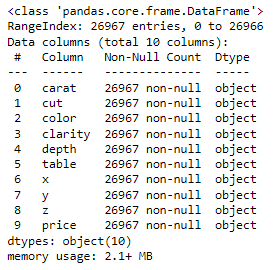


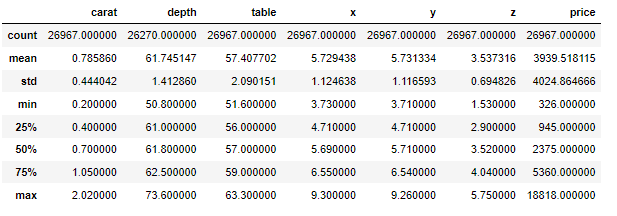


After imputation is done, we see that there are no null values present

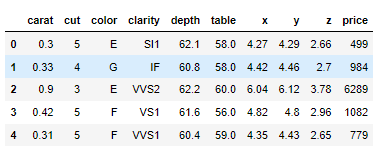


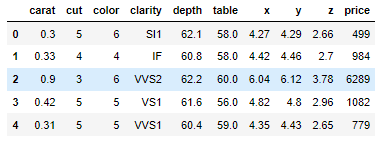


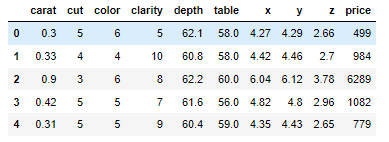
 

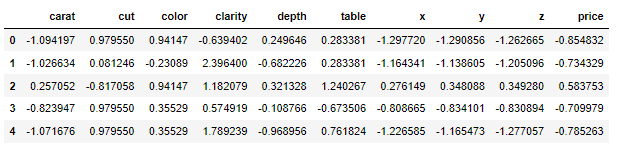


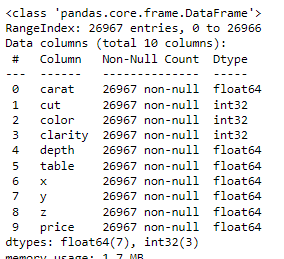
## Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train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

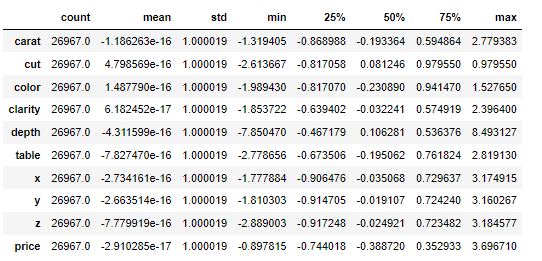


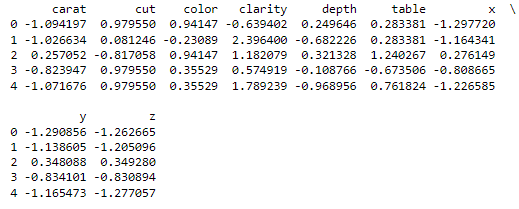


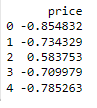


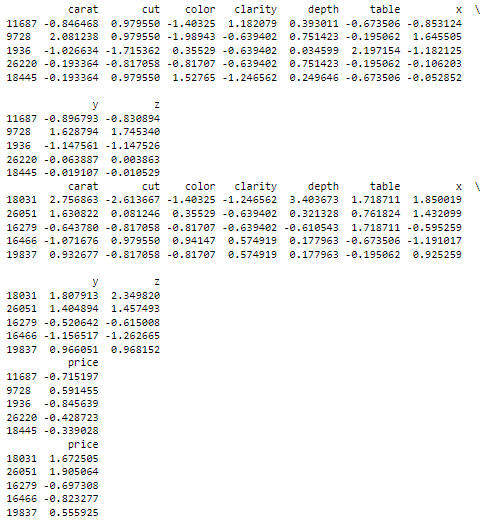












The coefficient for carat is 1.2801213328224776

The coefficient for cut is 0.04406130649318646

The coefficient for color is 0.1233528534141011

The coefficient for clarity is 0.1924067541374261

The coefficient for depth is -0.0038329577996184796

The coefficient for table is -0.015416741736580713

The coefficient for x is -0.5361037488818727

The coefficient for y is 0.44081340476733066

The coefficient for z is -0.16420841159037242

The intercept for our model is 0.0015672526389941363

### R square on train data:



### R square on test data:



### RMSE on Training data:



RMSE on Test data:



We still find we have multi collinearity in the dataset, to drop these values to lower level we can drop columns after doing stats model.

From stats model we can understand the features that do not contribute to the Model

We can remove those features after that the Vif Values will be reduced

Ideal value of VIF is less than 5%.

To ideally bring down the values to lower levels we can drop one of the variable that is highly correlated.

Dropping variables would bring down the multi collinearity level down.

## 1.4 Inference: Basis on these predictions, what are the business insights and recommendations. Please explain and summarize the various steps performed in this project. There should be proper business interpretation and actionable insights present.

We had a business problem to predict the price of the stone and provide insights for the company on the profits on different prize slots. From the EDA analysis we could understand the cut, ideal cut had number profits to the company. The colours H, I, J have bought profits for the company. In clarity if we could see there were no flawless stones and they were no profits coming from l1, l2, l3 stones. The ideal, premium and very good types of cut were bringing profits where as fair and good are not bringing profits. The predictions were able to capture 95% variations in the price and it is explained by the predictors in the training set.

Using stats model if we could run the model again we can have P values and coefficients which will give us better understanding of the relationship, so that values more 0.05 we can drop those variables and re run the model again for better results.

For better accuracy dropping depth column in iteration for better results.

### Recommendations

1. The ideal, premium, very good cut types are the one which are bringing profits so that we could use marketing for these to bring in more profits.

2. The clarity of the diamond is the next important attributes the more the clear is the stone the profits are more

PROBLEM 2: LINEAR REGRESSION

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

### Data Dictionary

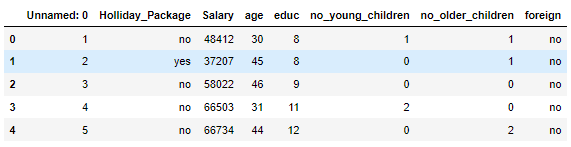
|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Holiday\_Package | Opted for Holiday Package yes/no? |
| Salary | Employee salary |
| age | Age in years |
| edu | Years of formal education |
| no\_young\_children | The number of young children (younger than 7 years) |
| no\_older\_children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

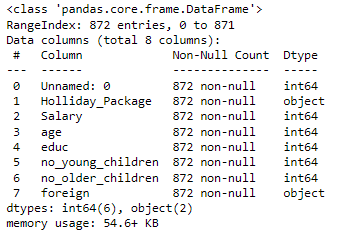
Loading all the necessary library for the model building.

Now, reading the head and tail of the dataset to check whether data has been properly fed.

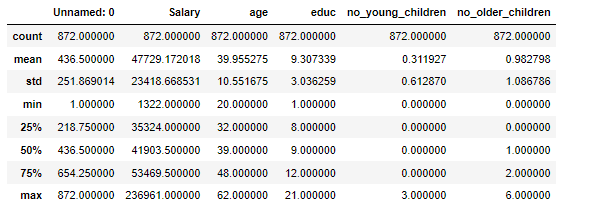
### Head of the data



### Info



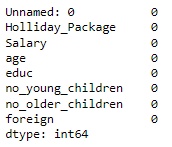
### Describe



We have integer and continuous data, Holiday package is our target variable Salary, age, educ and number young children, number older children of employee have the went to foreign, these are the attributes we have to cross examine and help the company predict weather the person will opt for holiday package or not.

1. No null values in the dataset,
2. We have integer and object data

### Null Value Check

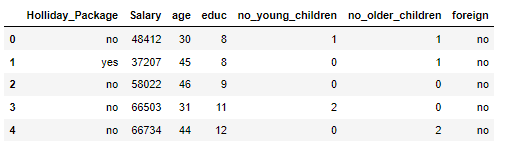


There are no NULL values found in the Data

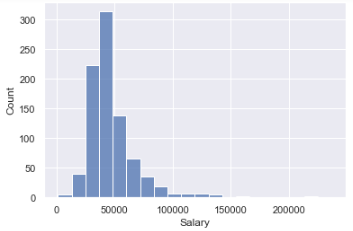
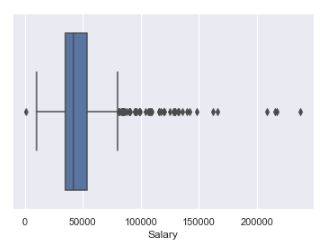
### Check for Duplicate values

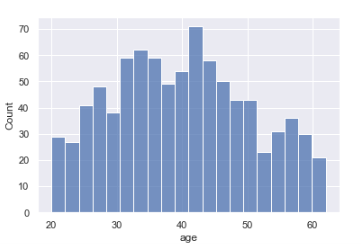
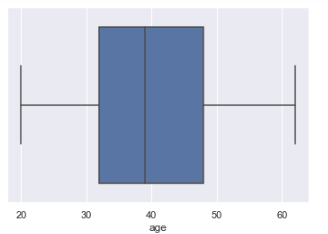
Number of duplicate rows = 0

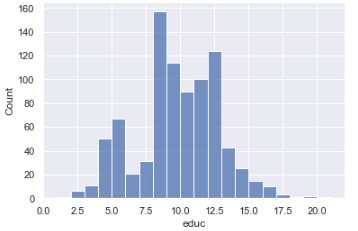
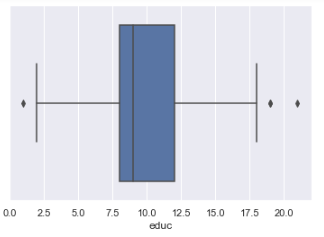
Drop unnamed column

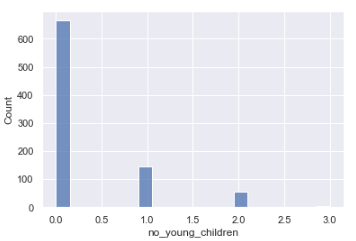
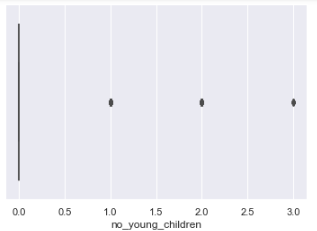


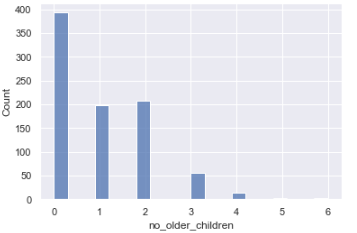
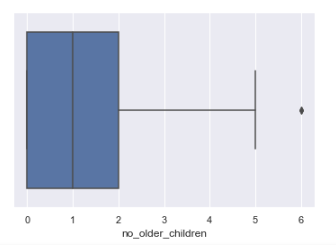
### Univariate/Bivariate Analysis

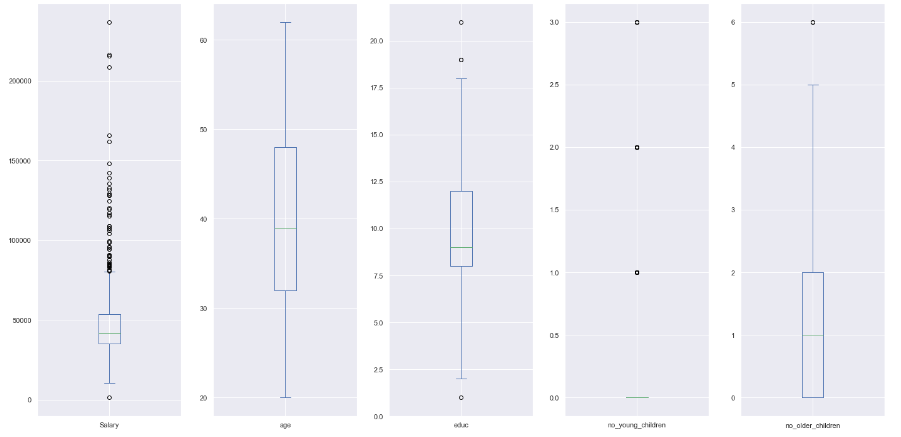
 

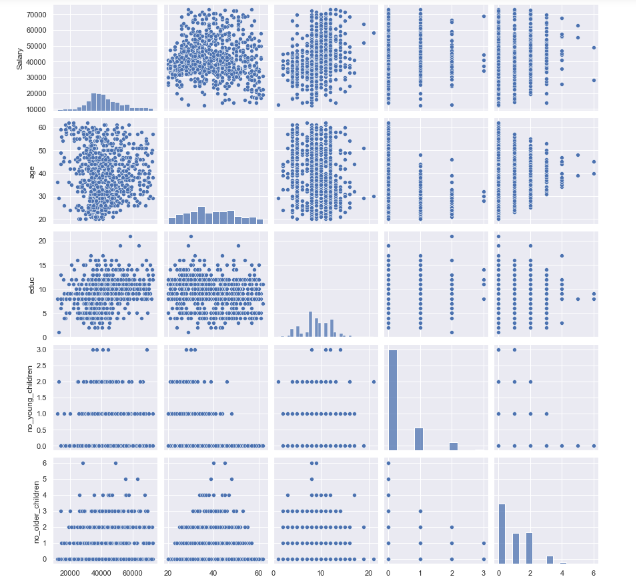
 



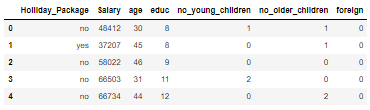
* From Holiday v/s Salary, We can see employee below salary 150000 have always opted for holiday package.
* From Age v/s Salary, Employee age over 50 to 60 have seems to be not taking the holiday package, whereas in the age 30 to 50 and salary less than 50000 people have opted more for holiday package.
* Based on the analysis, it looks like only 45% people are interested in holiday package

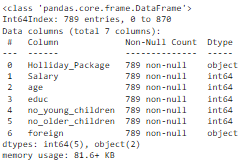
### Multivariate Analysis



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

## 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 score for Logistic regression train variables

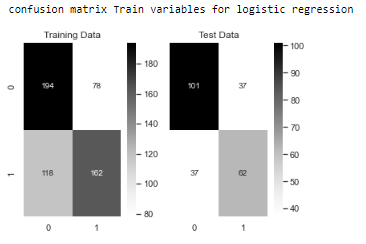
0.644927536231884



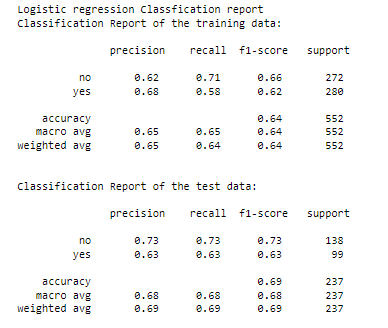
Accuracy score for Logistic regression test variables

0.6877637130801688

### Confusion Matrix Train Variables for Logistic regression



### Logistic regression Classification report

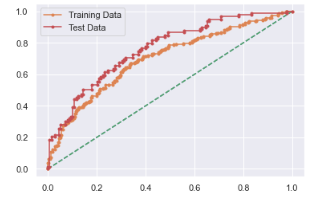


AUC and ROC FOR Logistic regression

AUC for the Training Data: 0.701

AUC for the Test Data: 0.763

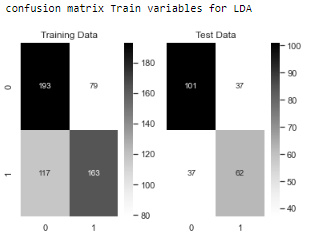
### ROC curve for the model



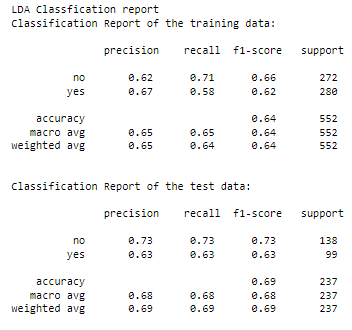
Accuracy score for LDA test variables

0.6877637130801688

### Confusion matrix train variables for LDA



### LDA classification report



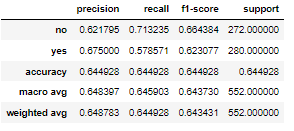
AUC and ROC FOR LDA

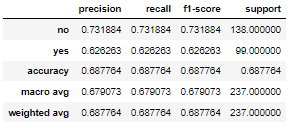
AUC for the Training Data: 0.700

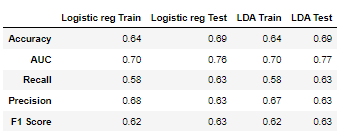
AUC for the Test Data: 0.767

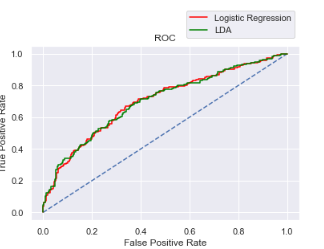
### Roc curve for the model

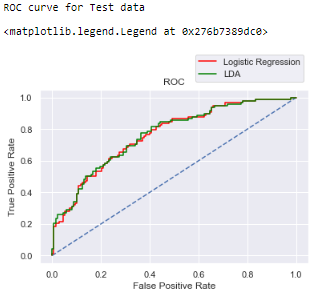












## 2.4 Inference: Basis on these predictions, what are the insights and recommendations.

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

We had a business problem where we need to predict whether an employee would opt for a holiday package or not.

For this problem we had done predictions both using logistic regression and linear discriminant analysis. Since both the techniques are giving the same results.

The EDA analysis clearly indicates certain criteria that where we could find people aged above 50 are less interested in holiday packages.

This is one of the observations we found that the aged people are less interested in holiday packages.

People ranging from the age 30 to 50 generally opt for holiday packages.

Employee age between 50 and 60 usually less interested to opt a holiday package, whereas employee aged 30 to 50 and salary less than 50000 people are comparatively opt more holiday packages.

The important factors deciding the predictions are salary, age and education

### Recommendations :

1. To improve holiday packages over the age above 50 we can provide religious destination places.

2. For people earning more than 150000 can be provided vacation holiday packages.

3. For employee having more number of older children can be provided with packages in holiday vacation places.