PREDICTIVE MODELING PROJECT REPORT

Vinyas Shreedhar

PGP-DSBA Online

May’21

Date: 20/10/2021

Table of Contents

Problem 1: Linear Regression 6

Cubic Zirconia 6

Introduction 6

Data Dictionary for Cubic Zirconia 6

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. 7

Sample of the Dataset 7

Exploratory Data Analysis 7

Information on the dataset 7

Descriptive Statistics 8

Univariate Analysis 9

Countplots 9

Boxplots 11

Distribution Plots 12

Skewness and Kurtosis 13

Bivariate Analysis 13

Catplots 13

Pairplot 15

Correlation Heatmap 16

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 ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning. 16

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. 17

Linear Regression 17

Assumptions for Linear Regression 17

Encoding Categorical variables 18

Multiple Linear Regression using Scikit Learn 18

The Coefficient of Determination $R2$ 19

Multiple Linear Regression using Stats Model 19

Linear Regression post Outlier Treatment 20

Checking Multicollinearity using Variance Inflation Factor (VIF) 22

1.4 Inference: Basis on these predictions, what are the business insights and recommendations. 22

Linear Regression Formula: 22

Business Insights and Recommendations 23

Problem 2 – Logistic Regression and Linear Discriminant Analysis (LDA) 24

Introduction **24**

Data Dictionary for Holiday Package **24**

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. 24

Sample of the Dataset **24**

Information on the dataset **25**

Descriptive Statistics **25**

Exploratory Data Analysis (EDA) **26**

Univariate Analysis **26**

Countplots **26**

Boxplots **27**

Boxplots of Numerical variables **30**

Distribution Plots of Numerical variables **31**

Skewness and Kurtosis **31**

Pairplot **32**

Correlation Heatmap **33**

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). 33

Encoding Categorical variables **33**

Logistic Regression **34**

Odds, Log Odds and Logit Transform **34**

Model 1 – Logistic Regression using Sklearn **35**

Model 2 – Logistic Regression **36**

Model 3 – Logistic Regression **38**

Logistic Regression using GridSearchCV **39**

Logistic Regression using Stats Model **42**

Linear Discriminant Analysis (LDA) **43**

LDA – removing foreign column **45**

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. 47

Model Performance before Outlier Treatment **47**

Model Performance post Outlier Treatment **47**

Model Performance when "foreign" column removed **47**

Comparison of models and Inferences **47**

List of Figures

Figure 1. Countplot of cut 9

Figure 2. Countplot of color 9

Figure 3. Countplot of clarity 10

Figure 4. Boxplot of each variable 11

Figure 5. Distribution Plot of each variable 12

Figure 6. Catplot between cut and price 13

Figure 7. Catplot between clarity and price 14

Figure 8. Catplot between color and price 14

Figure 9. Pairplot 15

Figure 10. Correlation Heatmap 16

Figure 11. Example of Linear Regression graph 17

Figure 12. Scatter plot of Actual vs Predicted Price 21

Figure 13. Count of Holiday Package 26

Figure 14. Count of foreign 27

Figure 15. Salary w.r.t Age (Holiday_Package) 27

Figure 16. Salary w.r.t educ (Holiday_Package) 28

Figure 17. Salary w.r.t no_young_children (Holiday_Package) 28

Figure 18. Salary w.r.t no_older_children (Holiday_Package) 28

Figure 19. Salary w.r.t age (foreign) 29

Figure 20. Salary w.r.t educ (foreign) 29

Figure 21. Salary w.r.t no_young_children (foreign) 29

Figure 22. Salary w.r.t no_older_children (foreign) 30

Figure 23. Boxplots of numerical variables 30

Figure 24. Distribution plots of numerical variables 31

Figure 25. Pairplot 32

Figure 26. Correlation Heatmap 33

Figure 27. Confusion Matrix and Classification Report – Training Data 35

Figure 28. Confusion Matrix and Classification Report – Testing Data 35

Figure 29. AUC and ROC 36

Figure 30. Confusion Matrix and Classification Report – Train Data 36

Figure 31. Confusion Matrix and Classification Report – Test Data 37

Figure32. AUC and ROC 37

Figure 33. Confusion Matrix and Classification Report – Train Data 38

Figure 34. Confusion Matrix and Classification Report – Test Data 38

Figure35. AUC and ROC 39

Figure 36. Confusion Matrix and Classification Report – Train Data 39

Figure 37. Confusion Matrix and Classification Report – Test Data 40

Figure 38. Boxplots post Outlier Treatment 40

Figure 39. Confusion Matrix and Classification Report – Train Data 41

Figure 40. Confusion Matrix and Classification Report – Test Data 41

Figure 41. AUC and ROC 42

Figure 42. Logit Regression Results 42

Figure 43. Example of LDA 43

Figure 44. Confusion Matrix and Classification Report – Train data 43

Figure 45. Confusion Matrix and Classification Report – Test data 44

Figure 46. AUC and ROC 44

Figure 47. Confusion Matrix and Classification Report – Train data 45

Figure 48. Confusion Matrix and Classification Report – Test data 45

Figure 49. AUC and ROC 46

List of Tables

Table 1. Data Dictionary 6

Table 2. Dataset sample 7

Table 3. Info on the Dataset 7

Table 4. Descriptive Statistics 8

Table 5. Skewness and Kurtosis 13

Table 6. Missing Values 16

Table 7. Data before encoding 18

Table 8. Data post encoding 18

Table 9. Model2 Parameters 19

Table 10. Model2 OLS Regression results 20

Table 11. Model11 Parameters 21

Table 12. Model11 OLS Regression results 21

Table 12. Model Performance 22

Table 13. Data Dictionary 24

Table 14. Sample Dataset 24

Table 15. Info of the dataset 25

Table 16. Missing values 25

Table 17. Descriptive Statistics 25

Table 18. Skewness and Kurtosis 31

Table 19. Dataset after encoding 33

Table 20. Model Scores – Before Outlier Treatment 47

Table 21. Model Scores – Post Outlier Treatment 47

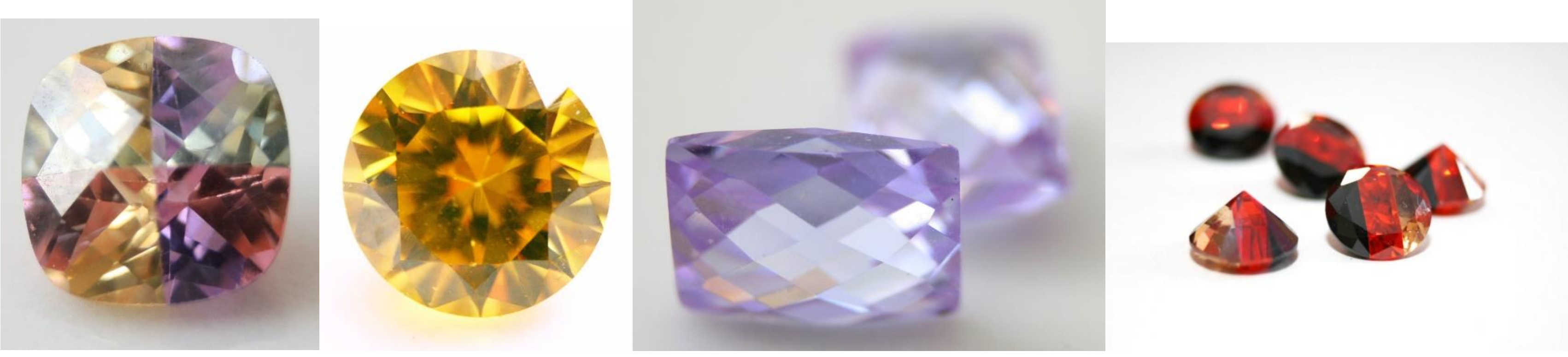
Table 22. Model Scores – foreign column removed 47

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.

Cubic Zirconia

Cubic zirconia (CZ) is the cubic crystalline form of zirconium dioxide (ZrO2). The synthesized material is hard and usually colorless, but may be made in a variety of different colors. Because of its low cost, durability, and close visual likeness to diamond, synthetic cubic zirconia has remained the most geologically and economically important competitor for diamonds since commercial production began in 1976. Its main competitor as a synthetic gemstone is a more recently cultivated material, synthetic moissanite.



Source: https://en.wikipedia.org/wiki/Cubic_zirconia

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. The data consists of 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). We will perform exploratory data analysis to understand what the given data has to say and then use linear regression techniques to predict the price of cubic zirconia based on the independent attributes we have.

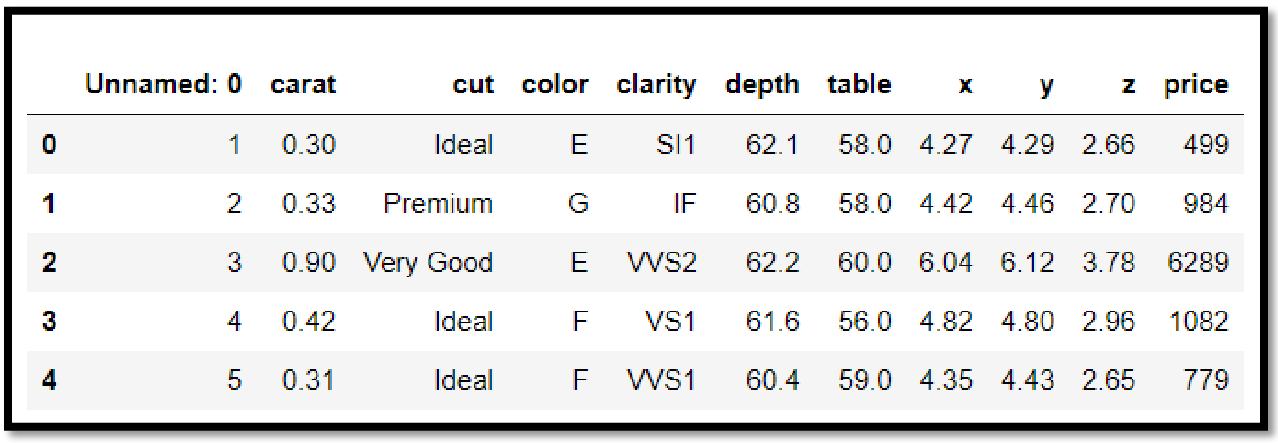
Data Dictionary for Cubic Zirconia

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	Color 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.

Table 1. Data Dictionary

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.

Sample of the Dataset



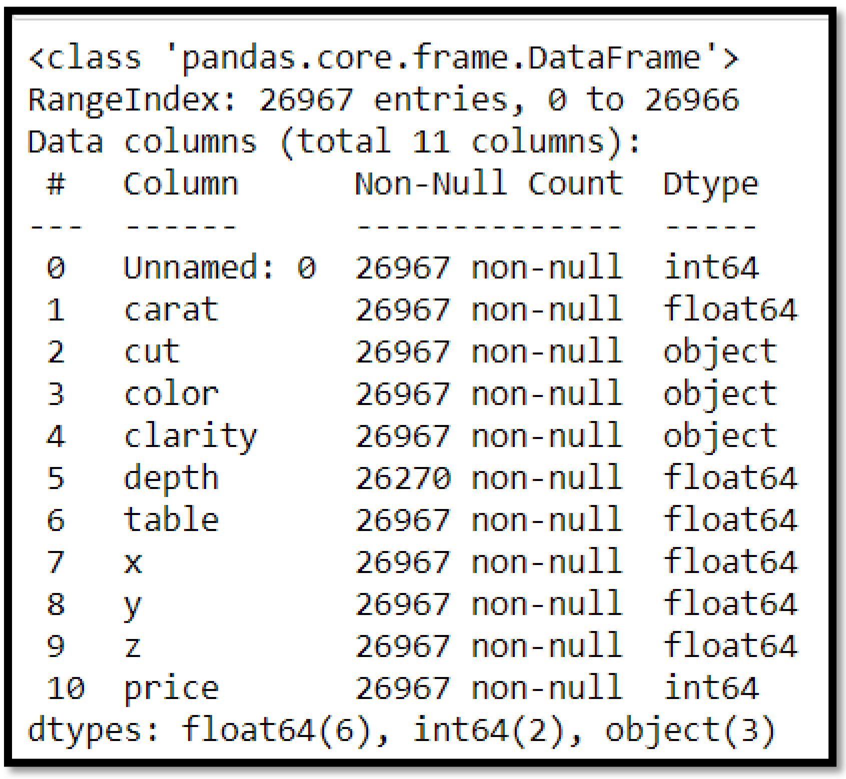
	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
0	1	0.30	Ideal	E	SI1	62.1	58.0	4.27	4.29	2.66	499
1	2	0.33	Premium	G	IF	60.8	58.0	4.42	4.46	2.70	984
2	3	0.90	Very Good	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289
3	4	0.42	Ideal	F	VS1	61.6	56.0	4.82	4.80	2.96	1082
4	5	0.31	Ideal	F	VVS1	60.4	59.0	4.35	4.43	2.65	779

Table 2. Dataset sample

Dataset has 11 variables with different aspects of cubic zirconia. Based on these aspects or information the price is defined.

Exploratory Data Analysis

Information on the dataset



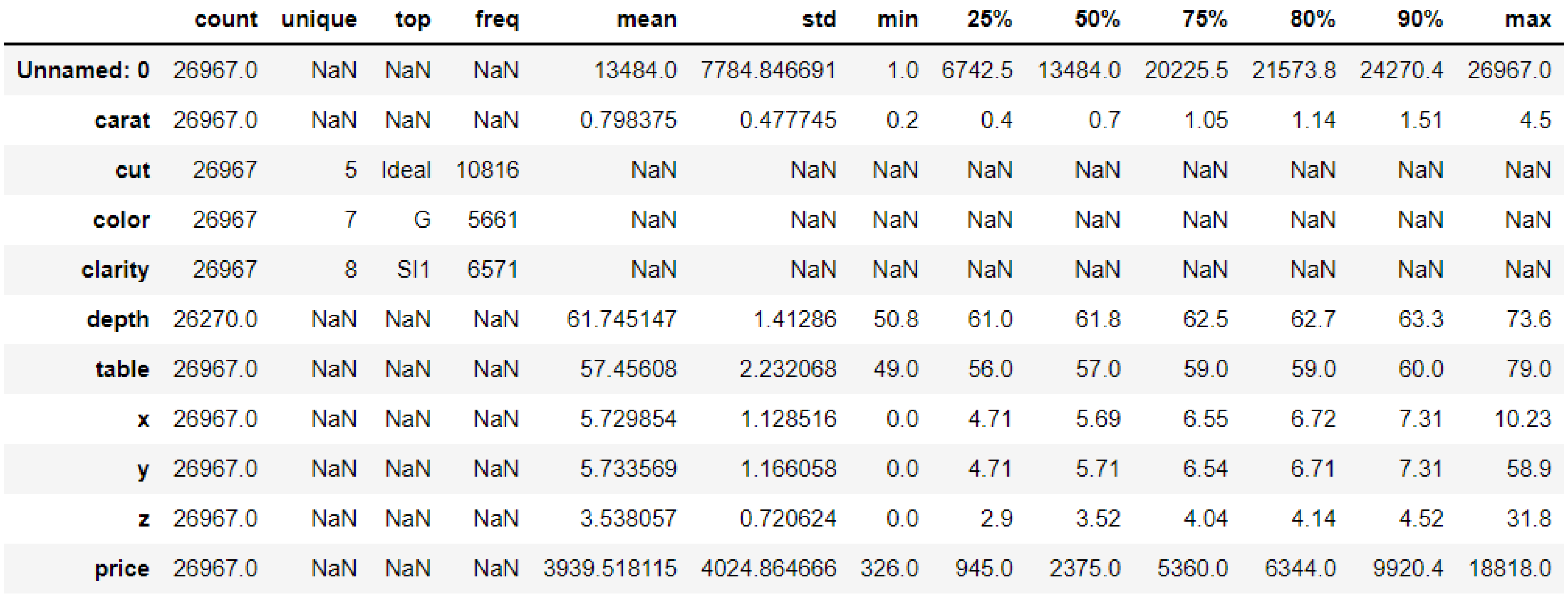
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26967 entries, 0 to 26966
Data columns (total 11 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   Unnamed: 0  26967 non-null  int64  
 1   carat       26967 non-null  float64
 2   cut         26967 non-null  object 
 3   color       26967 non-null  object 
 4   clarity     26967 non-null  object 
 5   depth       26270 non-null  float64
 6   table       26967 non-null  float64
 7   x           26967 non-null  float64
 8   y           26967 non-null  float64
 9   z           26967 non-null  float64
 10  price       26967 non-null  int64  
dtypes: float64(6), int64(2), object(3)
```

Table 3. Info on the Dataset

There are total 26,967 rows and 11 columns in the dataset. Out of 11, 3 columns are of object data type and 8 columns are of either integer or float data types. We can see that there are some missing values in the depth column in the dataset.

Descriptive Statistics

Summary statistics or 5-point summary helps us to understand the Interquartile Range like minimum, maximum, 25th, 50th and 75th percentiles, mean or average, standard deviation and count of data observations etc. The most popular descriptive statistics are Measures of Central Tendency which are Mean, Median and Mode which are fundamental to any data analysis.



	count	unique	top	freq	mean	std	min	25%	50%	75%	80%	90%	max
Unnamed: 0	26967.0	NaN	NaN	NaN	13484.0	7784.846691	1.0	6742.5	13484.0	20225.5	21573.8	24270.4	26967.0
carat	26967.0	NaN	NaN	NaN	0.798375	0.477745	0.2	0.4	0.7	1.05	1.14	1.51	4.5
cut	26967	5	Ideal	10816	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
color	26967	7	G	5661	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
clarity	26967	8	SI1	6571	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
depth	26270.0	NaN	NaN	NaN	61.745147	1.41286	50.8	61.0	61.8	62.5	62.7	63.3	73.6
table	26967.0	NaN	NaN	NaN	57.45608	2.232068	49.0	56.0	57.0	59.0	59.0	60.0	79.0
x	26967.0	NaN	NaN	NaN	5.729854	1.128516	0.0	4.71	5.69	6.55	6.72	7.31	10.23
y	26967.0	NaN	NaN	NaN	5.733569	1.166058	0.0	4.71	5.71	6.54	6.71	7.31	58.9
z	26967.0	NaN	NaN	NaN	3.538057	0.720624	0.0	2.9	3.52	4.04	4.14	4.52	31.8
price	26967.0	NaN	NaN	NaN	3939.518115	4024.864666	326.0	945.0	2375.0	5360.0	6344.0	9920.4	18818.0

Table 4. Descriptive Statistics

From the above table below are the observations.

1. **Unnamed: 0** column can be removed as it does not add value to the dataset.
2. Mean carat weight of the cubic zirconia is **0.79** with a maximum weight up to **4.5**. Seems to be normally distributed.
3. Most of cut quality of cubic zirconia is **"Ideal"**.
4. Color of the cubic zirconia is most of the times **G**.
5. Clarity of cubic zirconia is mostly **SI1**.
6. The average depth of cubic zirconia is **61.74** with a maximum value of **73.6**. This too is likely to be normally distributed.
7. table shows a mean of **57.45** and a median of **57.00** with a maximum value of **79.00**. Seems to be normally distributed with a slight **skewness**.
8. The average length of cubic zirconia is **5.72mm** ranging to a maximum length of **10.23mm**.
9. The average width of cubic zirconia is **5.73mm** to a maximum width of **58.9mm**.
10. The average height of cubic zirconia is **3.53mm** to a maximum height of **31.8mm**.
11. The attributes x, y, z have minimum value as **0.00mm**. Not sure if this is a valid data input. Need to check with business if these are incorrect data entries or valid measurements. For now we leave it as is.
12. price variable which is our dependent variable ranges from a minimum of **326.00** to a maximum of **18818.00** with a mean price of **3939.51** to a median price of **2375.00**. **90%** of the product items are within the price of **9920.40**.

Univariate Analysis

Countplots

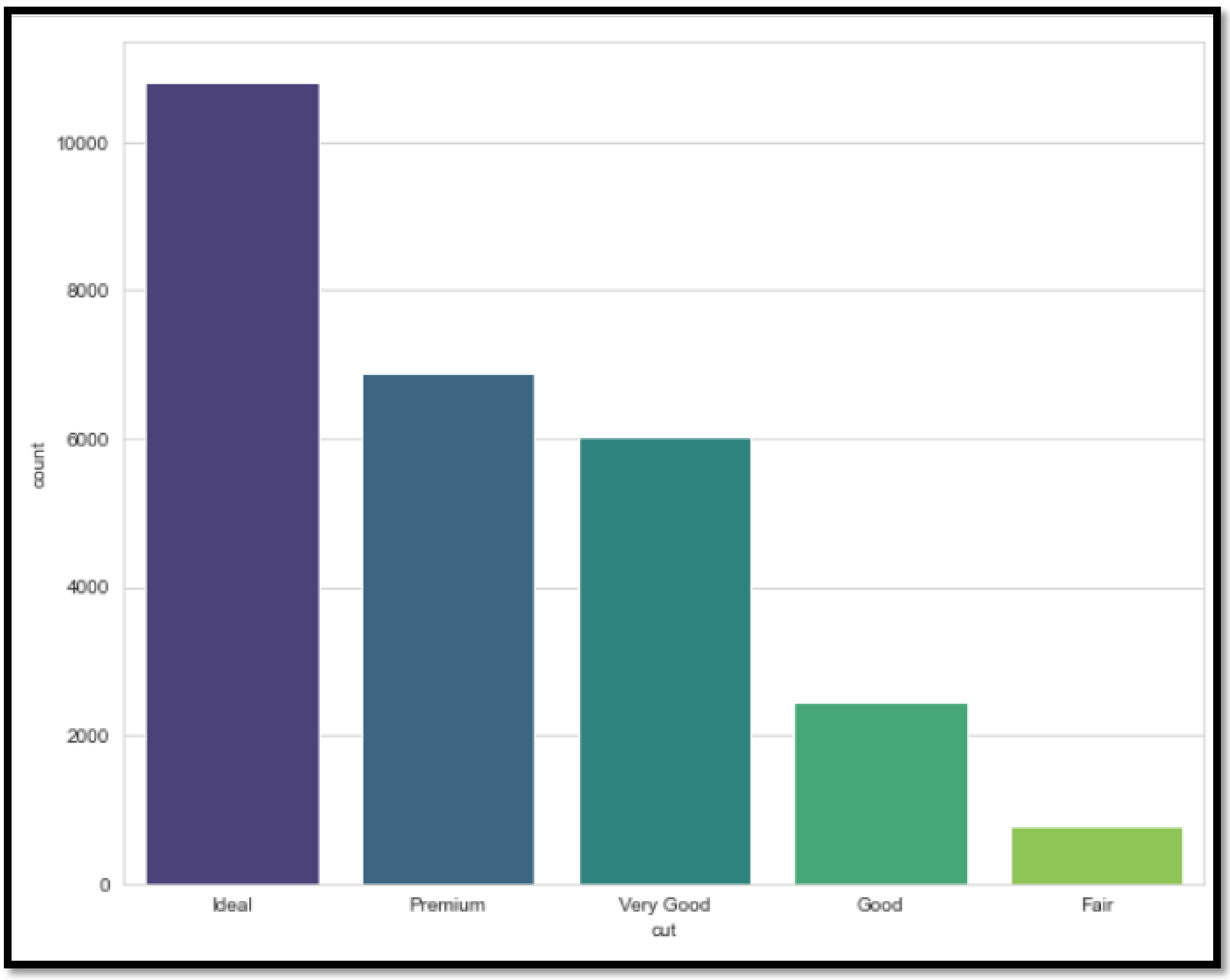


Figure 1. Countplot of cut

The above count plot shows that Ideal cut diamonds are the maximum in the dataset followed by Premium and Very Good cut.

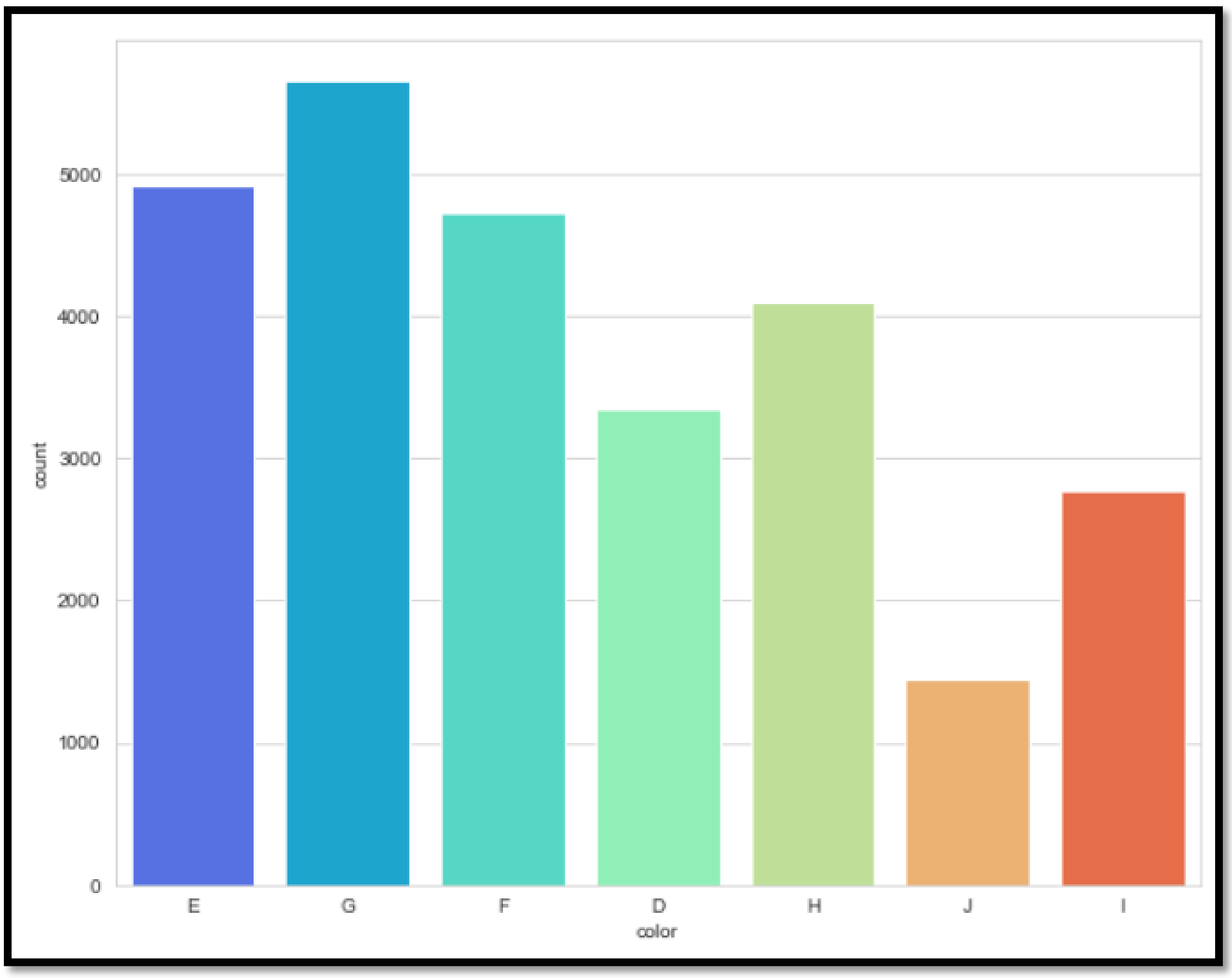


Figure 2. Countplot of color

Among the diamond colors, G and E are the maximum. Close to 4900 diamonds are E color category which is very close to worst color. J being the best color is least among the diamonds in the dataset.

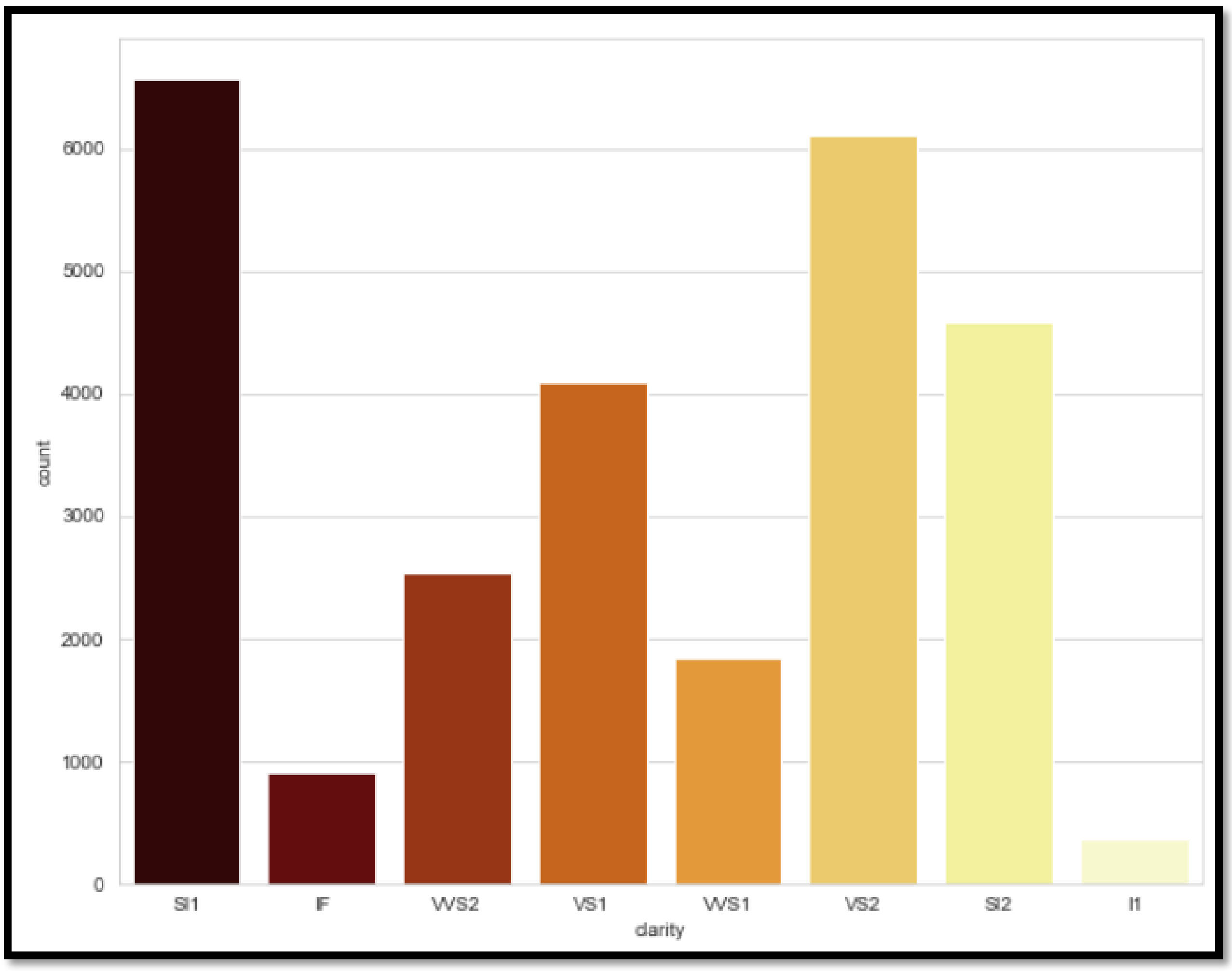


Figure 3. Countplot of clarity

When it comes to clarity, SI1 and VS2 are the most predominant diamonds from the dataset. These two are moderately superior clarity.

Boxplots

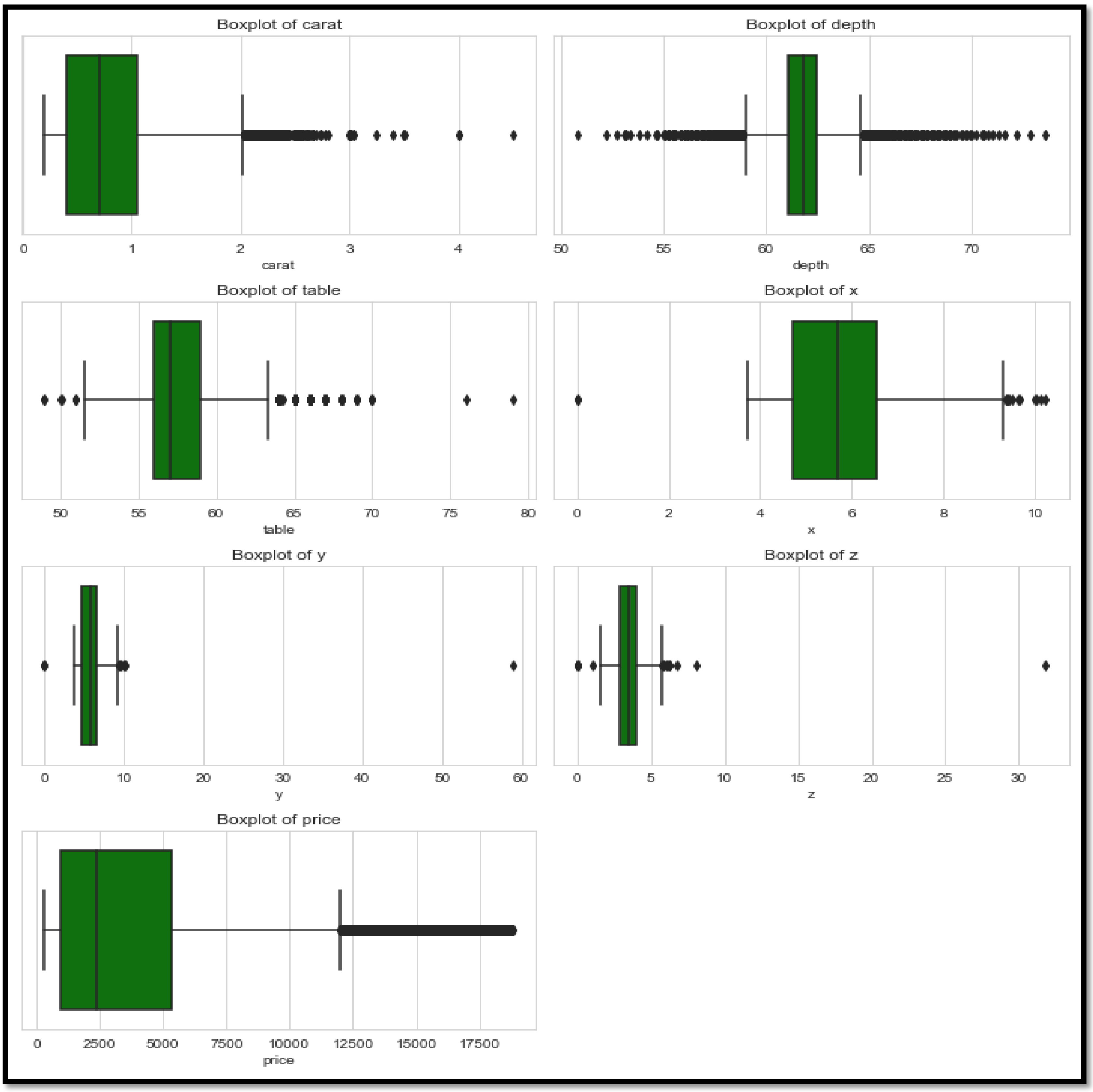


Figure 4. Boxplot of each variable

From the above boxplots, we can see that there are outliers in all the variables.

Distribution Plots

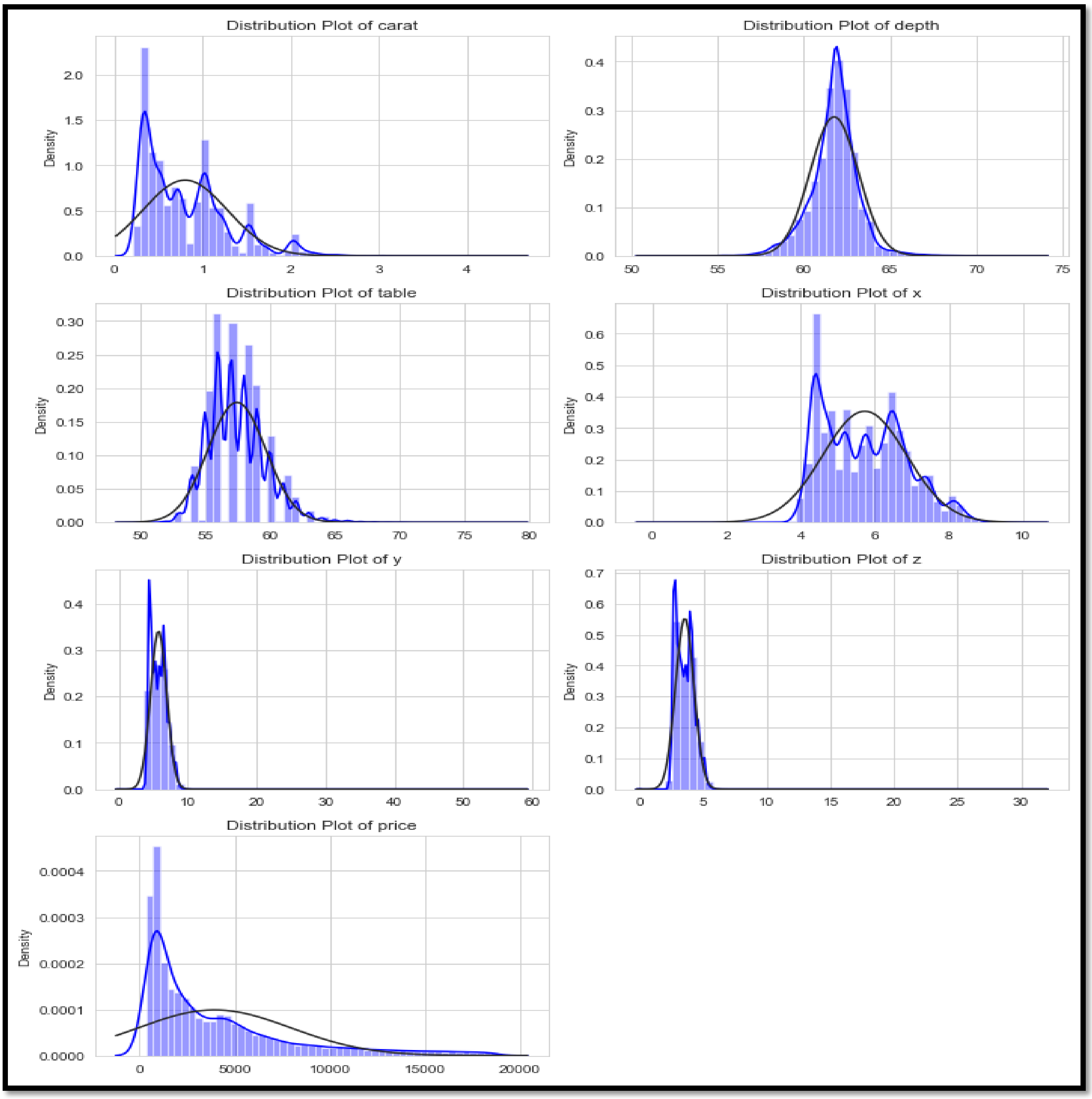


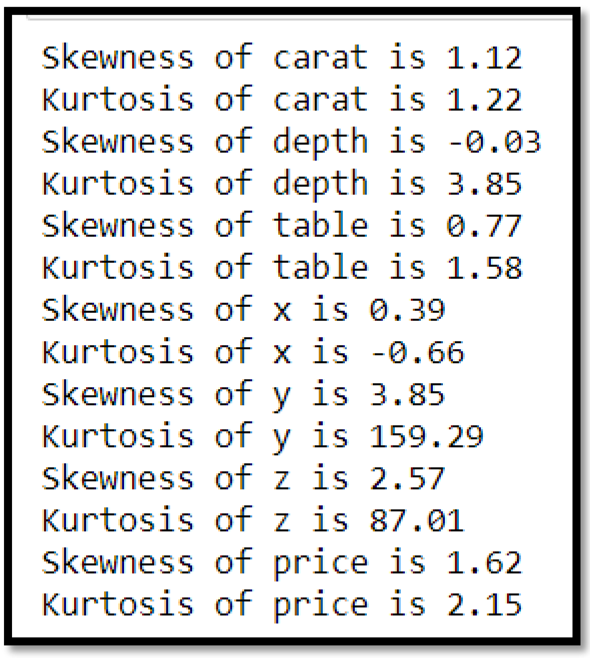
Figure 5. Distribution Plot of each variable

From the above plots, distribution plot of depth shows a normal distribution. carat, table, y, z and price variables show right skewed distributions. Variable x shows somewhat a normal distribution with multi-modal peaks.

Skewness and Kurtosis

According to Wikipedia, "skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive, zero, negative, or undefined." For normally distributed data, the skewness should be about zero. For unimodal continuous distributions, a skewness value greater than 0 means that there is more weight in the right tail of the distribution.

According to Wikipedia, "kurtosis is a measure of the "tailedness" of the probability distribution of a real-valued random variable. Like skewness, kurtosis describes the shape of a probability distribution." In Fisher’s definiton, the kurtosis of the normal distribution is zero. The distribution with a higher kurtosis has a heavier tail.



```
Skewness of carat is 1.12
Kurtosis of carat is 1.22
Skewness of depth is -0.03
Kurtosis of depth is 3.85
Skewness of table is 0.77
Kurtosis of table is 1.58
Skewness of x is 0.39
Kurtosis of x is -0.66
Skewness of y is 3.85
Kurtosis of y is 159.29
Skewness of z is 2.57
Kurtosis of z is 87.01
Skewness of price is 1.62
Kurtosis of price is 2.15
```

Table 5. Skewness and Kurtosis

Bivariate Analysis

Catplots

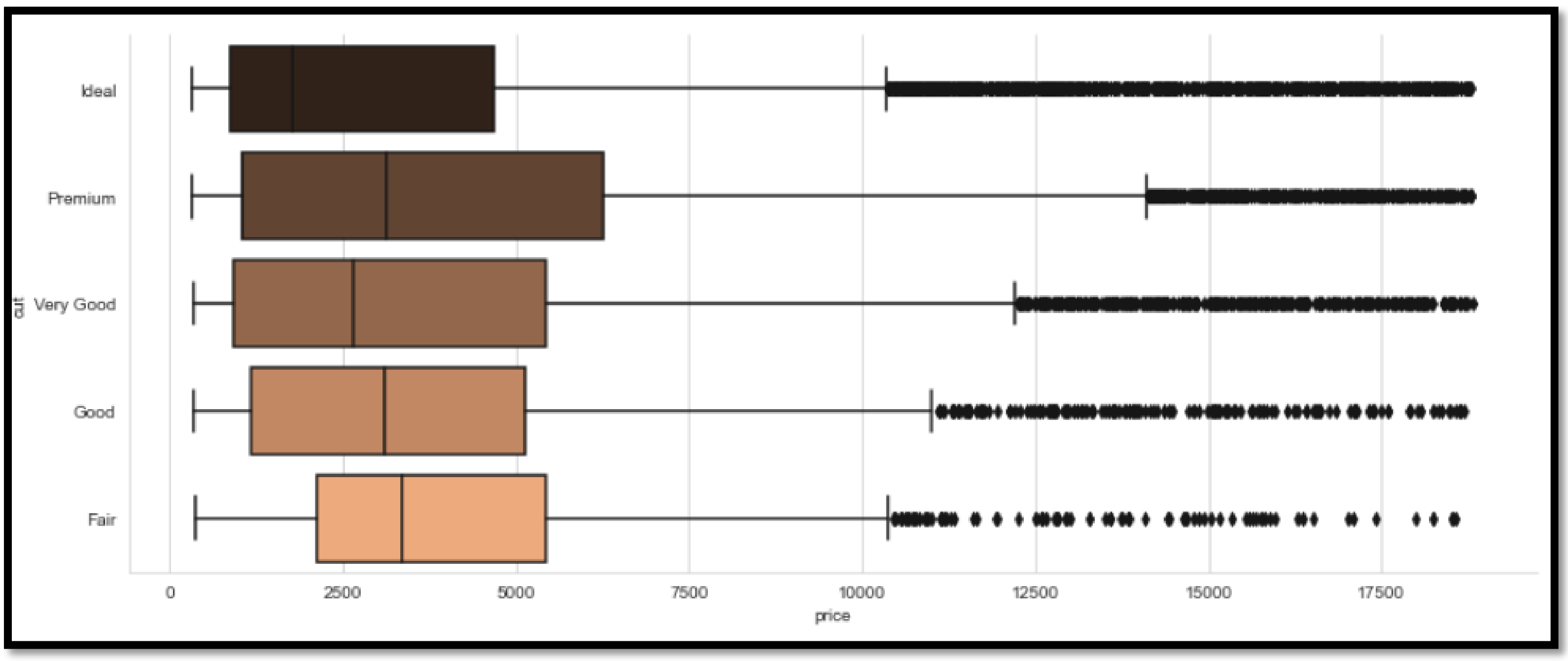


Figure 6. Catplot between cut and price

When compared to price range with diamond cuts, premium cuts are at the higher side however we see a lot of outliers for all types of cuts ranging from Fair to Ideal.

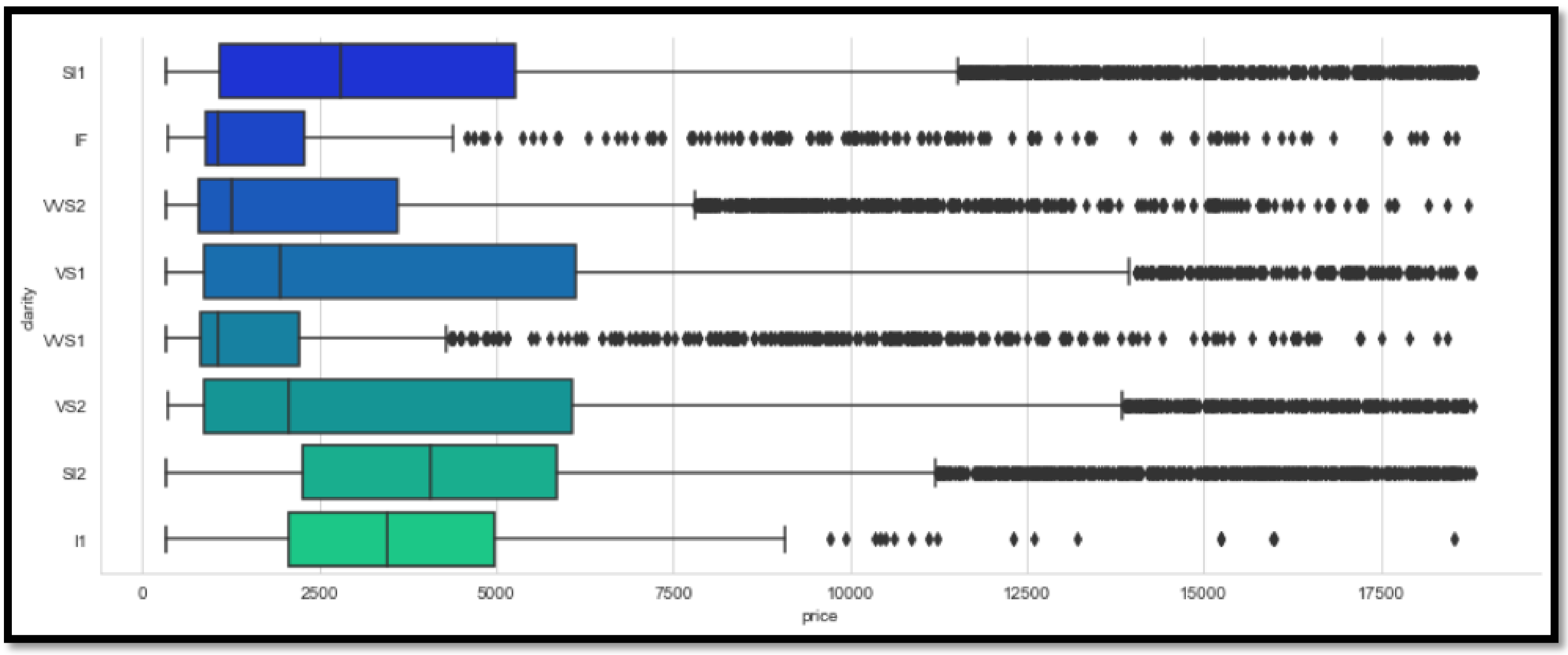


Figure 7. Catplot between clarity and price

With respect to clarity, the price range is higher for VS1 and VS2 however we see a lot of outliers for most of the clarity diamonds. As expected VVS1 which is one of the worst clarity diamonds have the least price range but also lot of outliers.

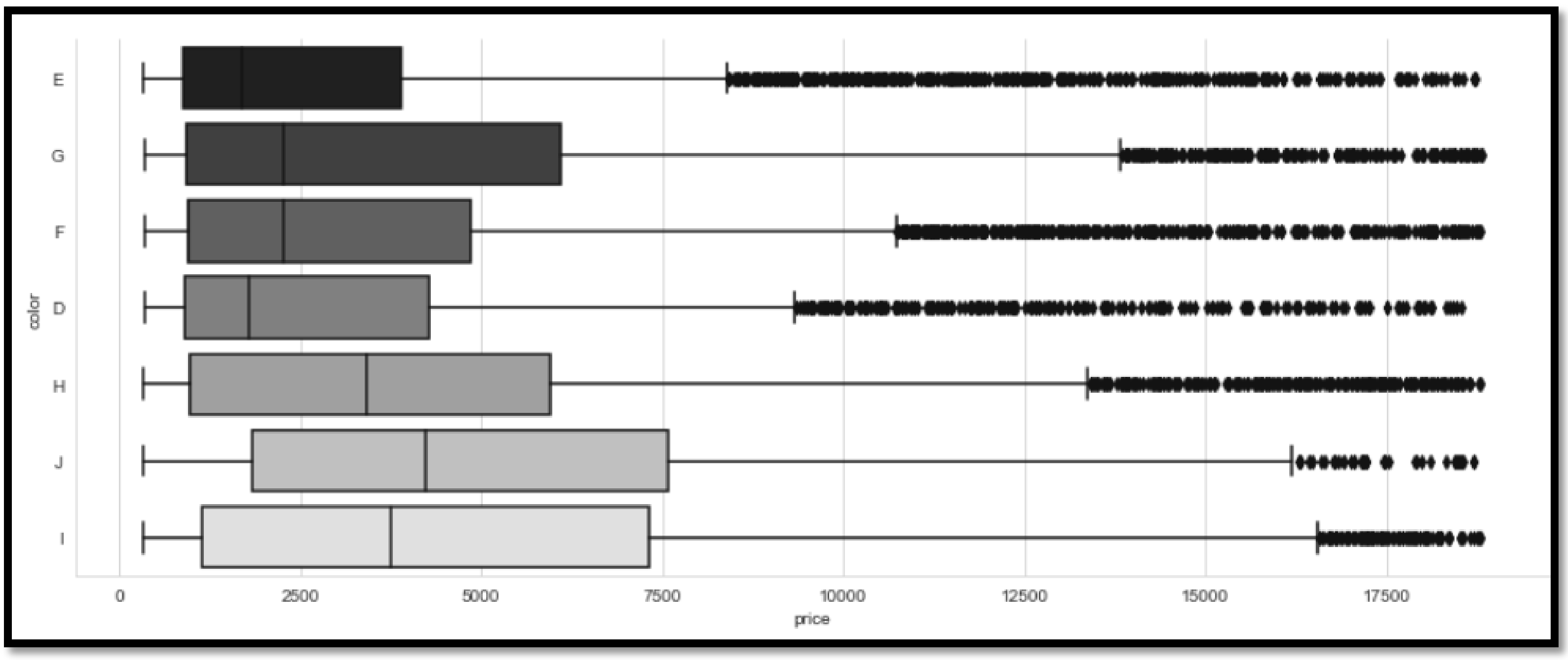


Figure 8. Catplot between color and price

When it comes to color of the diamonds, it very evident that the best color category I and J has higher price range whereas worst colors D and E has lowest price range. Apparently these too have outliers.

Pairplot

Pairplot shows the relationship between the variables in the form of scatter plot and distribution of the variable in the form of histogram.

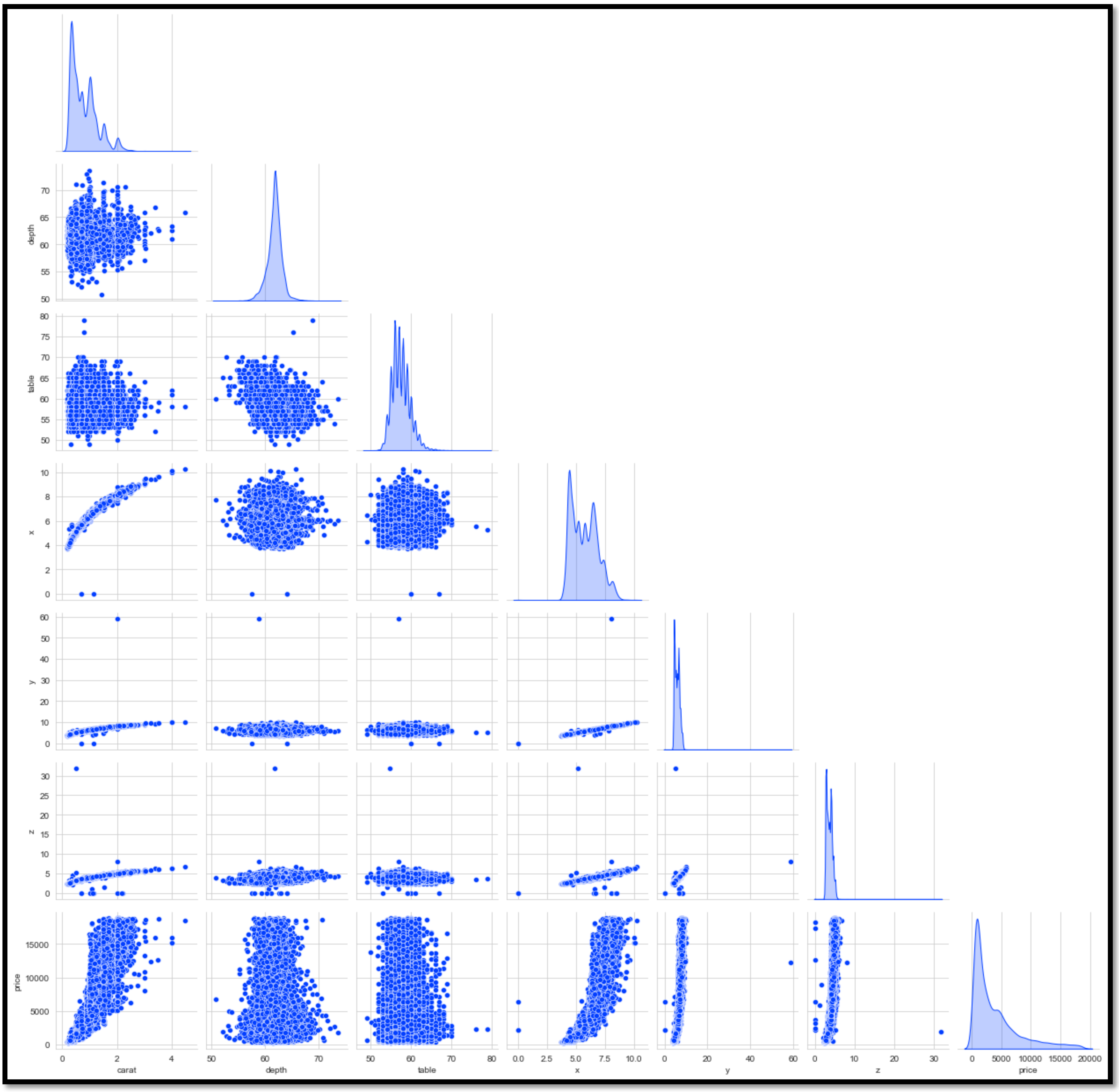


Figure 9. Pairplot

From the above pairplot we can see that there is a strong positive relationship between carat and price. Similarly x, y and z have a moderately positive correlation with carat. Variables x, y and z themselves seem to have high multicollinearity. However we will also see if this is true when we check the correlation heatmap.

Correlation Heatmap

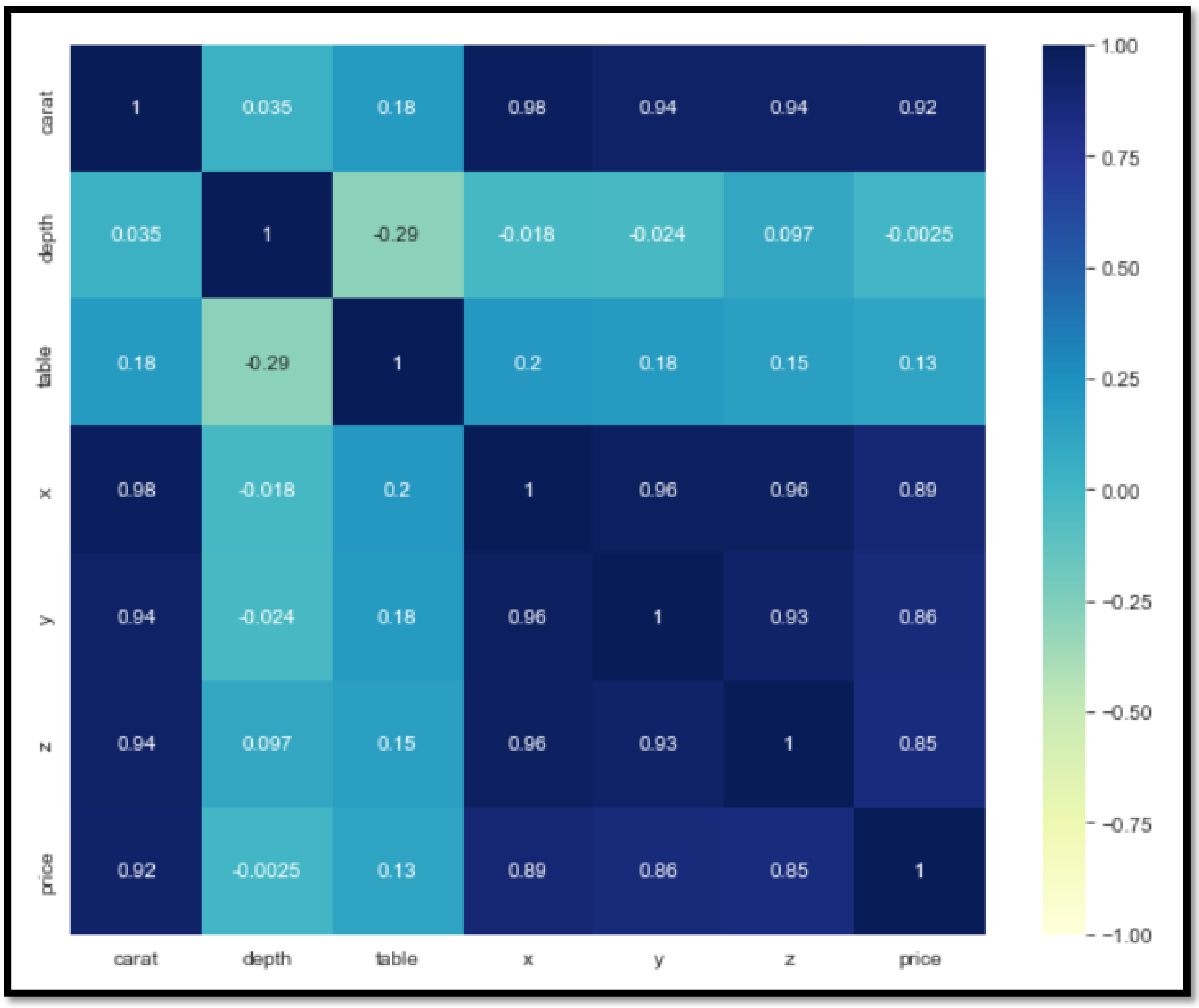


Figure 10. Correlation Heatmap

From the above correlation plot we can see that x, y and z have high positive correlation with carat. There is also Multicollinearity between x, y and z variables. Price and carat have a strong positive correlation with a correlation coefficient of 0.92. Correlation between x, y and z with price is positive but they are moderate. Correlation values are always between 1 and -1. Those which are closer to 1 are positively correlated and those which near -1 are negatively correlated. Values near to 0 have no correlation.

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

From the earlier info table, we saw that there are 697 Null values in depth column. We are not sure why these are kept blank and since we do not want to lose valuable information from the dataset, we will be imputing the Median value for missing values in depth column.

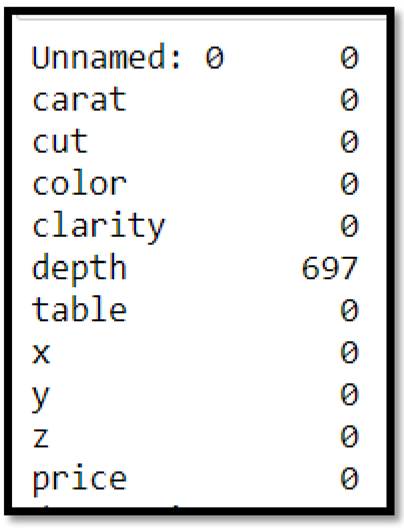

```
Unnamed: 0      0
carat           0
cut             0
color           0
clarity         0
depth         697
table           0
x               0
y               0
z               0
price           0
```

Table 6. Missing Values

From the earlier summary statistics table, the attributes x, y, z have minimum value as 0.00mm. They are clear indicators that they have been incorrectly captured as bad data. It’s a data input error because there cannot be diamonds with 0.00mm length, width and height. Hence we will be dropping those rows as it is not a significant loss of information.

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.

Linear Regression

$$\boldsymbol{y = \alpha + \beta x}$$

Simple definition of Linear Regression is: if there are **n** observations with **y** being the dependent variable and **x** being the independent variable, then variable y takes a linear combination of x to form the best fit line which predicts the y values given the x values.

$y = value\ of\ the\ continuous\ response\ (or\ dependent)\ variable$

$\alpha = intercept$

$\beta_x = slope\ coefficients$

$x = random\ variable\ which\ is\ continuous\ in\ nature$

$$\boldsymbol{y = \alpha + \beta x + \varepsilon}$$

$\varepsilon = error\ term$

ε represents the error term. Note that error here does not indicate any mistake, simply the difference between the expected and observed values of the response. The error terms provide crucial insight into the regression process.

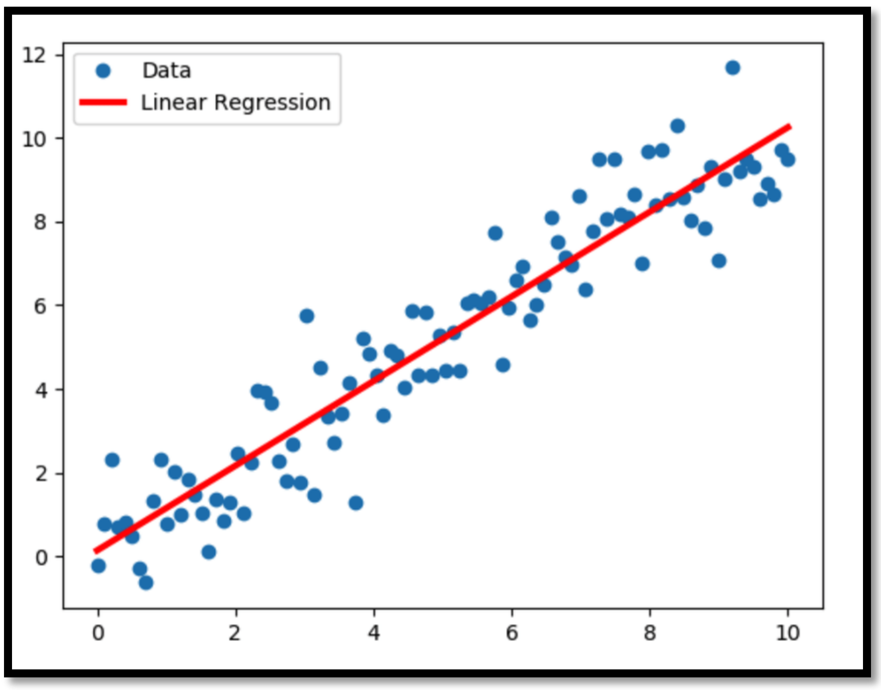


Figure 11. Example of Linear Regression graph

Extended formula for multiple linear regression

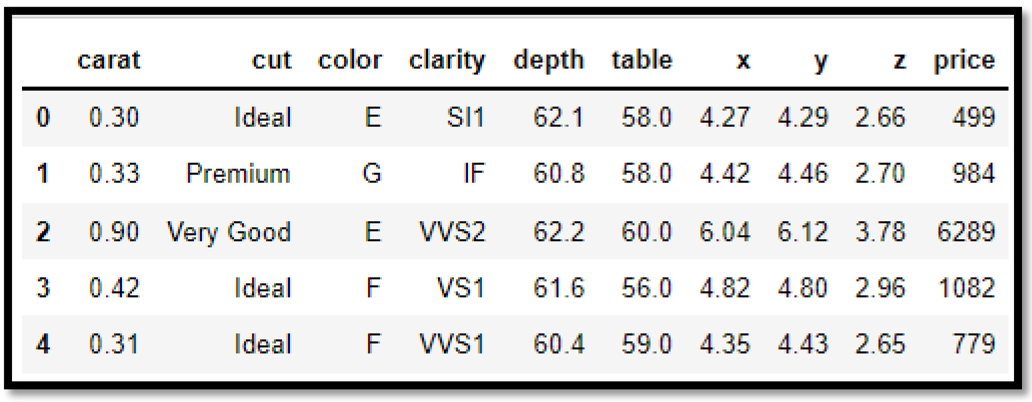
$$y = \alpha + \beta x_1 + \beta x_2 + \cdots \beta x_n + \varepsilon$$

Source: Google and Google Images

Assumptions for Linear Regression

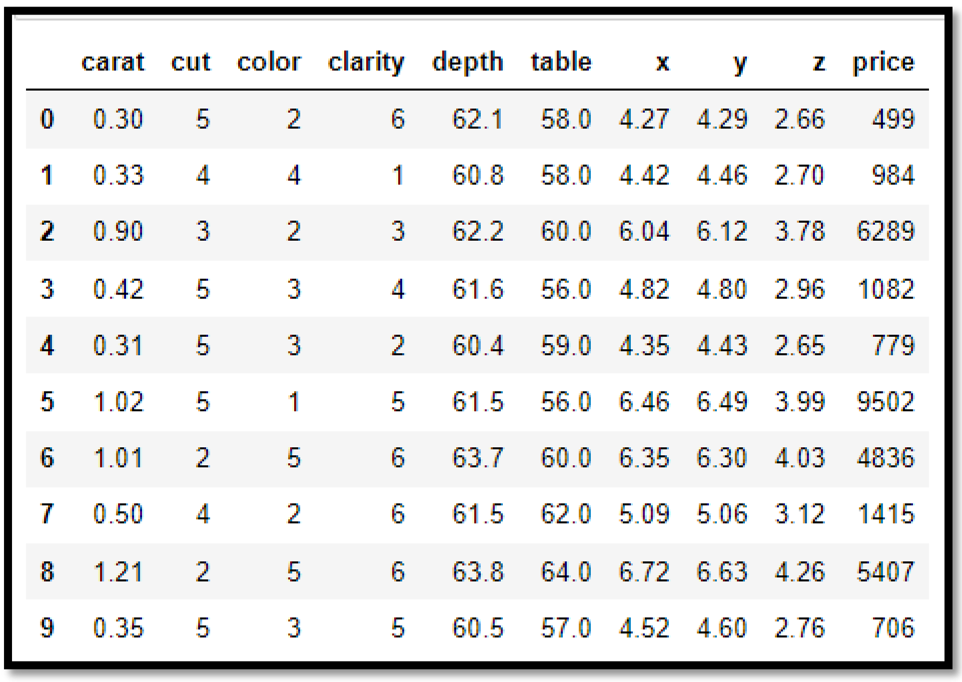
1. Basic assumption is that the dependent variable is **linearly related** to the estimated parameter.
2. The observations Y_i and the error terms ϵ_i are **independent**.
3. The error variables ϵ_i are **normally** distributed.
4. The errors have **no bias** i.e, $\epsilon_i = 0$ and they are ***homoscedastic*** i.e, they have **equal variance**.

Encoding Categorical variables



	carat	cut	color	clarity	depth	table	x	y	z	price
0	0.30	Ideal	E	SI1	62.1	58.0	4.27	4.29	2.66	499
1	0.33	Premium	G	IF	60.8	58.0	4.42	4.46	2.70	984
2	0.90	Very Good	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289
3	0.42	Ideal	F	VS1	61.6	56.0	4.82	4.80	2.96	1082
4	0.31	Ideal	F	VVS1	60.4	59.0	4.35	4.43	2.65	779

Table 7. Data before encoding



	carat	cut	color	clarity	depth	table	x	y	z	price
0	0.30	5	2	6	62.1	58.0	4.27	4.29	2.66	499
1	0.33	4	4	1	60.8	58.0	4.42	4.46	2.70	984
2	0.90	3	2	3	62.2	60.0	6.04	6.12	3.78	6289
3	0.42	5	3	4	61.6	56.0	4.82	4.80	2.96	1082
4	0.31	5	3	2	60.4	59.0	4.35	4.43	2.65	779
5	1.02	5	1	5	61.5	56.0	6.46	6.49	3.99	9502
6	1.01	2	5	6	63.7	60.0	6.35	6.30	4.03	4836
7	0.50	4	2	6	61.5	62.0	5.09	5.06	3.12	1415
8	1.21	2	5	6	63.8	64.0	6.72	6.63	4.26	5407
9	0.35	5	3	5	60.5	57.0	4.52	4.60	2.76	706

Table 8. Data post encoding

We have encoded the categorical variables cut, color and clarity in the ascending order from worst to best since linear regression does not take string variables as parameters into model building.

Below is the encoding for ordinal values:

CUT: Fair = 1, Good = 2, Very Good = 3, Premium = 4 and Ideal = 5

COLOR: D = 1, E = 2, F = 3, G = 4, H = 5, I = 6 and J = 7

CLARITY = IF = 1, VVS1 = 2, VVS2 = 3, VS1 = 4, VS2 = 5, SI1 = 6, SI2 = 7 and I1 = 8

Multiple Linear Regression using Scikit Learn

An observation is considered to be an outlier if that particular observation has been mistakenly captured in the data set which we are not sure of at this moment. Treating outliers sometimes results in the models having better performance but the models lose out on generalization. Hence we will not be treating the outliers. We will stick to keeping the outliers as is.

We will be using the sklearn.model_selection package to use train_test_split, sklearn.linear_model to use LinearRegression and sklearn.metrics to use mean_squared_error and r2_score.

Post fitting the regression model below is the output:

The **coefficients** are **1.13504255e+04, 5.52968799e+01, -2.79682162e+02, 2.92759026e+02, -1.61274335e+02, -9.52483249e+01, -1.24993665e+03, 7.81024366e-01, -3.05416673e+01**
Intercept: **17045.88250984**
Mean Squared Error (MSE) = **1846369.51**
Root Mean Squared Error (RMSE) = **1358.8118001718906**
Coefficient of Determination (r-square) on the train data = **88.87%**
Coefficient of Determination (r-square) on the test data = **88.34%**

The Coefficient of Determination R^2

The coefficient of determination $\boldsymbol{R^2}$ is a summary measure that explains how well the sample regression line fits the data. The rationale and computation of $\boldsymbol{R^2}$ is discussed below:
Let $\widehat{Y_\iota} = \hat{\beta}0 + \hat{\beta}1Xi$ and $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. $\widehat{Y_\iota}$ is the estimated value of the response for a given value of the predictorX_i.

The difference between the observed and the estimated values of the response is called residual. Residual is the estimated value of the unobserved error component in the regression equation.

$\widehat{\epsilon_\iota} = Y_i - \widehat{Y_\iota}$
Residuals are very important part of regression and they have many useful properties. In fact, it can be shown that$\sum_{i=1}^{n} \widehat{\epsilon_\iota} = 0$, if the estimation method is OLS.

$\sum_{i=1}^{n}\left(Y_i^2 - \hat{Y}\right)^2$: is the total variation of the actual Y about their sample mean and termed as the ***total sum of squares (SST)***. This is closely linked to the sample variance of Y.

$\sum_{i=1}^{n}\left(\widehat{Y_\iota} - \bar{Y}\right)^2$: is the sum of squares due to regression and is called ***regression sum of squares (SSR)***

$\sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$: is the sum of squared differences between the observed and the predicted values of the response. This is known as the ***residual sum of squares or the error sum of squares (SSE)***

$$SST=SSR+SSE$$

$1=\frac{SSR}{SST}+\frac{SSE}{SST}$ (on dividing SST on both sides)

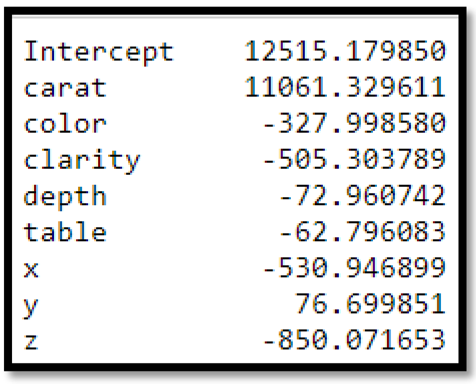
We now define $\boldsymbol{R^2}$ as: $R^2=\frac{SSR}{SST}$ or $R^2=1-\frac{SSE}{SST}$

Source: Google

Multiple Linear Regression using Stats Model

Using the Ordinary Least Squares Method under Stats Model below are the results:

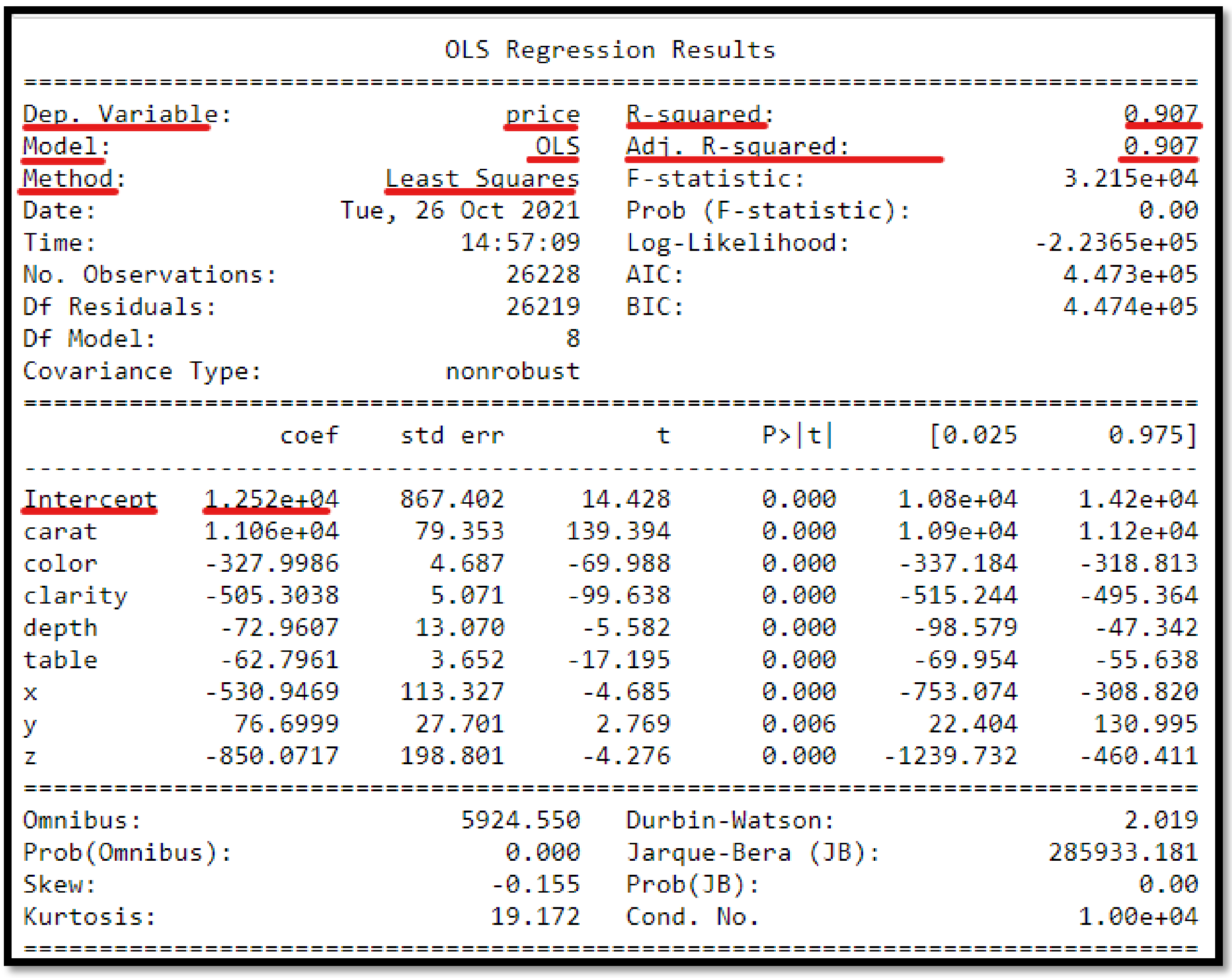
Model2 Parameters



```
Intercept    12515.179850
carat        11061.329611
color         -327.998580
clarity       -505.303789
depth          -72.960742
table          -62.796083
x             -530.946899
y               76.699851
z             -850.071653
```

Table 9. Model2 Parameters

Model2 Summary



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.907
Model:                            OLS   Adj. R-squared:                  0.907
Method:                 Least Squares   F-statistic:                 3.215e+04
Date:                Tue, 26 Oct 2021   Prob (F-statistic):               0.00
Time:                        14:57:09   Log-Likelihood:            -2.2365e+05
No. Observations:               26228   AIC:                         4.473e+05
Df Residuals:                   26219   BIC:                         4.474e+05
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   1.252e+04    867.402     14.428      0.000    1.08e+04    1.42e+04
carat       1.106e+04     79.353    139.394      0.000    1.09e+04    1.12e+04
color       -327.9986      4.687    -69.988      0.000    -337.184    -318.813
clarity     -505.3038      5.071    -99.638      0.000    -515.244    -495.364
depth        -72.9607     13.070     -5.582      0.000     -98.579     -47.342
table        -62.7961      3.652    -17.195      0.000     -69.954     -55.638
x           -530.9469    113.327     -4.685      0.000    -753.074    -308.820
y             76.6999     27.701      2.769      0.006      22.404     130.995
z           -850.0717    198.801     -4.276      0.000   -1239.732    -460.411
==============================================================================
Omnibus:                     5924.550   Durbin-Watson:                   2.019
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           285933.181
Skew:                          -0.155   Prob(JB):                         0.00
Kurtosis:                      19.172   Cond. No.                     1.00e+04
==============================================================================
```

Table 10. Model2 OLS Regression results

1. Dependent variable is "**price**".
2. Model and Method used is **Ordinary Least Square (OLS)** method which uses mathematical algorithm for linear regression.
3. No of observations are **26228**.
4. **R-squared** value is **0.907** which means 90.7% of the outcome variability is explained by the model.
5. **Adj. R-squared** value is **0.907** which is the correct R-square according to the number of dependent variables.
6. **F-statistic** value is used for the calculation of the p-value of the model, Probability (F-statistic) which here is less than **0.05**. This also tells us Python is using an ANOVA test which implies an F-distribution.
7. **Coef** shows the coefficients of each dependent variable.
8. **Std err** shows how accurate our coefficient values are. Std err is inversely related to accuracy. Lower the std err signifies higher the accuracy.
9. **P>|t|** is the p-value. This shows how statistically significant each independent variable is on the price (dependent variable). P-value less than 0.05 means they are statistically quite significant.

Post building 11 models with different parameters to see if the model improves, we have arrived at Model10 and Model11 as our final selection. Hence only those results are provided below.

Linear Regression post Outlier Treatment

Model10 Coefficients:

The coefficient for carat is **8751.300881277873**
The coefficient for cut is **107.97604034695344**
The coefficient for color is **-274.4338929100507**
The coefficient for clarity is **-431.2570095305264**
The coefficient for depth is **91.40613133509609**
The coefficient for table is **-13.411837851864844**
The coefficient for x is **-908.6273289854977**
The coefficient for y is **1710.6519949538217**
The coefficient for z is **-1853.0026678045633**

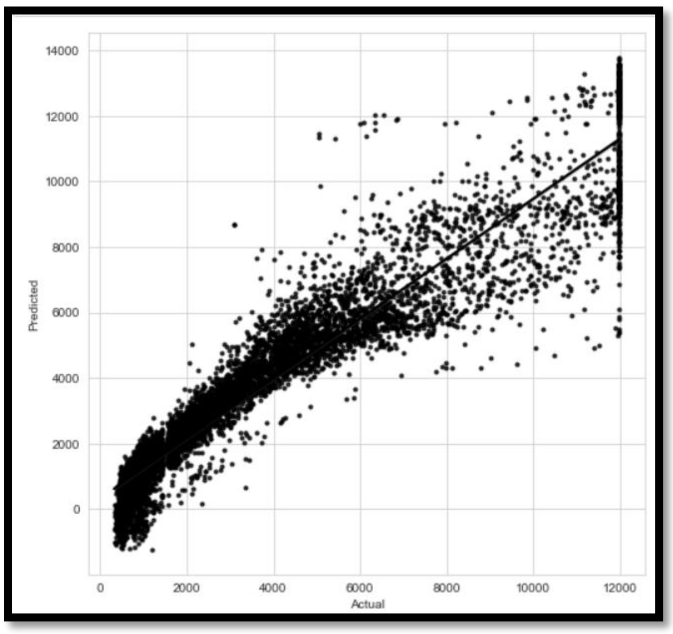
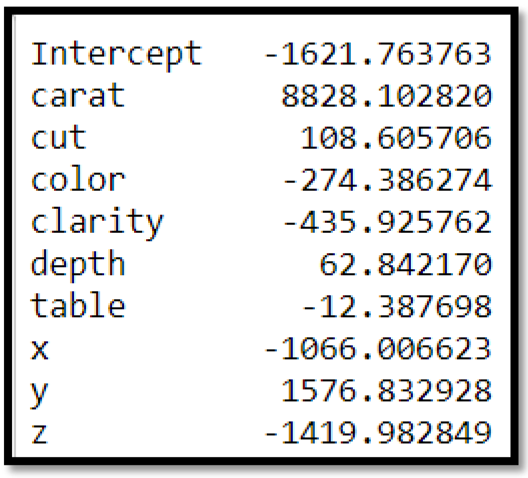


Figure 12. Scatter plot of Actual vs Predicted Price

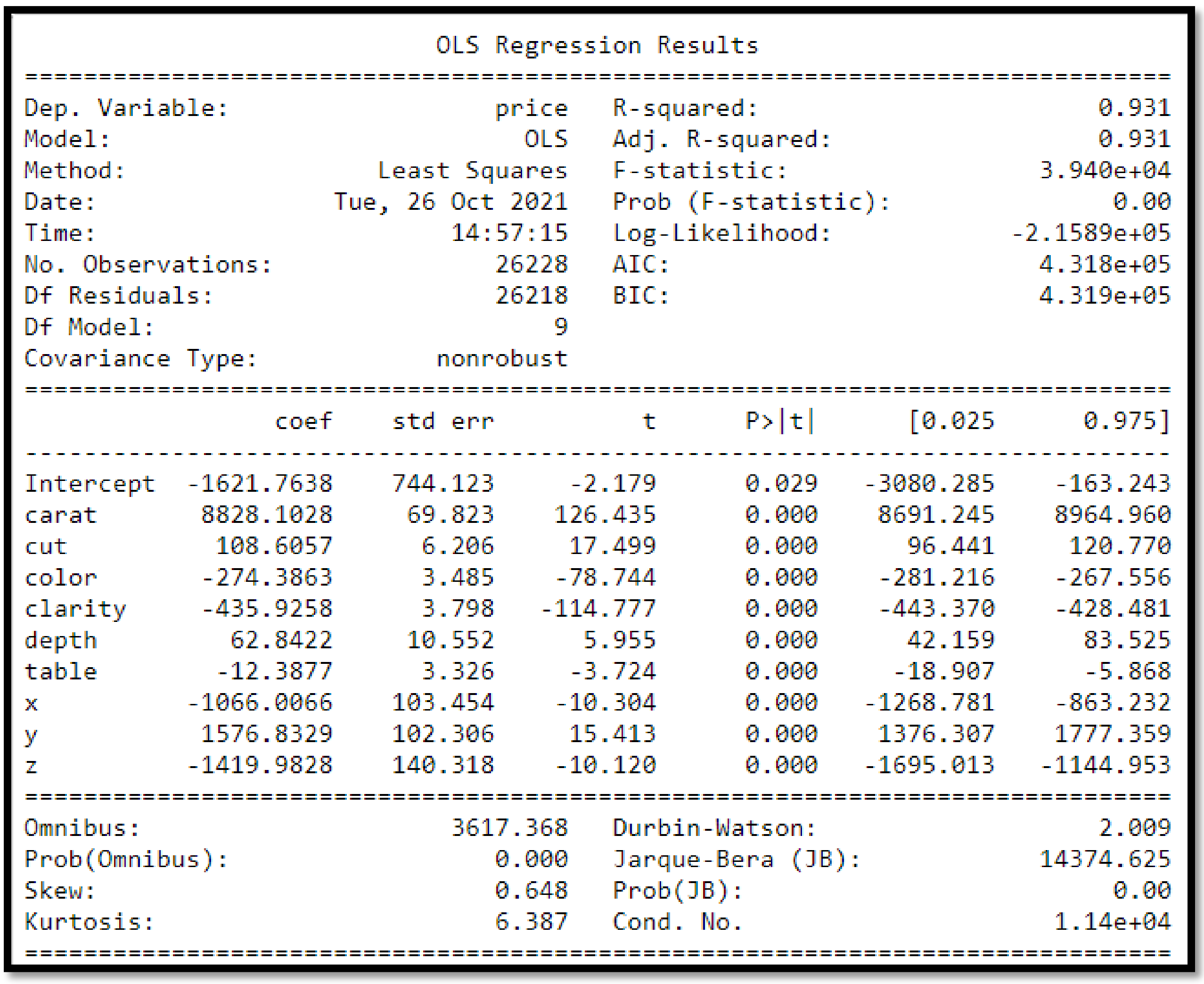
Model11 Parameters



```
Intercept    -1621.763763
carat         8828.102820
cut            108.605706
color         -274.386274
clarity       -435.925762
depth           62.842170
table          -12.387698
x            -1066.006623
y             1576.832928
z            -1419.982849
```

Table 11. Model11 Parameters

Model11 Summary



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.931
Model:                            OLS   Adj. R-squared:                  0.931
Method:                 Least Squares   F-statistic:                 3.940e+04
Date:                Tue, 26 Oct 2021   Prob (F-statistic):               0.00
Time:                        14:57:15   Log-Likelihood:            -2.1589e+05
No. Observations:               26228   AIC:                         4.318e+05
Df Residuals:                   26218   BIC:                         4.319e+05
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -1621.7638    744.123     -2.179      0.029   -3080.285    -163.243
carat       8828.1028     69.823    126.435      0.000    8691.245    8964.960
cut          108.6057      6.206     17.499      0.000      96.441     120.770
color       -274.3863      3.485    -78.744      0.000    -281.216    -267.556
clarity     -435.9258      3.798   -114.777      0.000    -443.370    -428.481
depth         62.8422     10.552      5.955      0.000      42.159      83.525
table        -12.3877      3.326     -3.724      0.000     -18.907      -5.868
x          -1066.0066    103.454    -10.304      0.000   -1268.781    -863.232
y           1576.8329    102.306     15.413      0.000    1376.307    1777.359
z          -1419.9828    140.318    -10.120      0.000   -1695.013   -1144.953
==============================================================================
Omnibus:                     3617.368   Durbin-Watson:                   2.009
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            14374.625
Skew:                           0.648   Prob(JB):                         0.00
Kurtosis:                       6.387   Cond. No.                     1.14e+04
==============================================================================
```

Table 12. Model11 OLS Regression results

Checking Multicollinearity using Variance Inflation Factor (VIF)

```
carat ---> 124.01621195373215
cut ---> 17.58710826236311
color ---> 6.114473645140826
clarity ---> 12.398816270307671
depth ---> 1261.7699031381692
table ---> 887.9258441321977
x ---> 10722.381239794784
y ---> 9354.51456273367
z ---> 3622.677517683246
```

Variance Inflation Factor (VIF) is one of the methods to check if independent variables have correlation between them. If they are correlated, then it is not ideal for linear regression models as they inflate the standard errors which in turn affect the regression parameters. As a result, the regression model becomes non-reliable and lacks interpretability.

General rule of thumb: If VIF values are equal to 1, then that means there is no multicollinearity. If VIF values are equal to 5 or exceedingly more than 5, then there is moderate multicollinearity. If VIF is 10 or more, then that means there is high collinearity.

From the above we can conclude that variables **carat, cut, clarity, depth, table, x, y and z** has high multicollinearity whereas variable **color** has moderate correlation. However variables **cut, color and clarity** are categorical variables which are transformed to numerical using encoding. Hence it is difficult to say whether the VIF values indicate the right results.

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

Below is the model comparison post different combination of dependent variables in influencing price variable.

Model No	Package	Outlier Treatment	Formula	R-squared	Adjusted R-square	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)
model10	sklearn	Yes	price ~ carat+cut+color+clarity+depth+table+x+y+z	0.93	NA	835883.17	914.27
model11	stats model	Yes	price ~ carat+cut+color+clarity+depth+table+x+y+z	0.93	0.93	NA	NA
model2	stats model	No	price ~ carat+color+clarity+depth+table+x+y+z	0.91	0.91	NA	NA
model3	stats model	No	price ~ carat+color+clarity+depth+table+x	0.91	0.91	NA	NA
model9	stats model	No	price ~ carat+cut+color+clarity+depth+table+x+y+z	0.91	0.91	NA	NA
model5	stats model	No	price ~ carat+color+clarity+depth+table	0.90	0.90	NA	NA
model7	stats model	No	price ~ carat+cut+color+clarity+depth+table	0.90	0.90	NA	NA
model8	stats model	No	price ~ carat+cut+clarity+depth	0.89	0.89	NA	NA
model6	stats model	No	price ~ carat+x+y+z	0.86	0.86	NA	NA
model4	stats model	No	price ~ carat+x	0.85	0.85	NA	NA
model1	sklearn	No	price ~ carat+cut+color+clarity+depth+table+x+y+z	0.81	NA	3229430.89	1797.07

Table 12. Model Performance

Linear Regression Formula:

Price (y) = (-1621.76) * Intercept + (8828.1) * carat + (108.61) * cut + (-274.39) * color + (-435.93) * clarity + (62.84) * depth + (-12.39) * table + (-1066.01) * x + (1576.83) * y + (-1419.98) * z

Business Insights and Recommendations

1. **Model11** and **Model10** have performed better after treating outliers.

2. The highest **R-squared** is **93.00%** which we achieved by using **stats_model** post outlier treatment.

3. We can see that there is **high multicollinearity** in the dataset.

4. **Intercept** of the model is **-1621.76**.

5. R-squared **93.00%** shows a good accuracy which means **93%** of the price is explained by the model.

6. **RMSE** on training data is **1168.49** and RMSE on testing data is **914.26**.

7. As per the above graph (Figure12) there is a **strong linear relationship** between the actual and predicted values with some noise to an extent which signifies the unexplained variance.

8. As per our model **carat, cut, color, clarity, depth, table, x, y and z** are the best attributes to influence price. For 1 unit increase in carat the price of cubic zirconia increases by 8828.1 USD provided the other coefficients are constant. Similarly for 1 unit increase in y the price increases by 1576.83 USD keeping the other coefficients constant. The same analogy applies to all the other coefficients.

9. Increase in **carat weight** of the diamond will **increase the price** of the diamond considerably.

10. Width (y) of the diamond in mm also plays an important factor. As the **width** increases the **price** also increases.

11. **Brighter** the color of the diamond, the price **increases**.

12. Gem Stones Ltd should work on **carat**, **color** and **width** of the diamonds which are strong contributors for price.

13. Since **x, y and z had zero values** we have removed them as they are clear indicators that they have been incorrectly captured as **bad data**. It’s a **data input error** because there cannot be diamonds with 0.00mm length, width and height.

14. Gem Stones Ltd can collect more data in future which helps in building more robust models for price prediction. For now model11 and model10 is good for price prediction. If Gem Stones Ltd wants to use the model without outlier treatment, they can as well use the Model2 which predicts the price **91.00%** accurately.

Problem 2 – Logistic Regression and Linear Discriminant Analysis (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.

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. The data consists of employees who either opted for holiday package or not and other attributes of almost 872 observations. We will perform exploratory data analysis to understand what the given data has to say and then use logistic regression and linear discriminant analysis (LDA) techniques to predict whether an employee will opt for the holiday package based on the independent attributes we have.

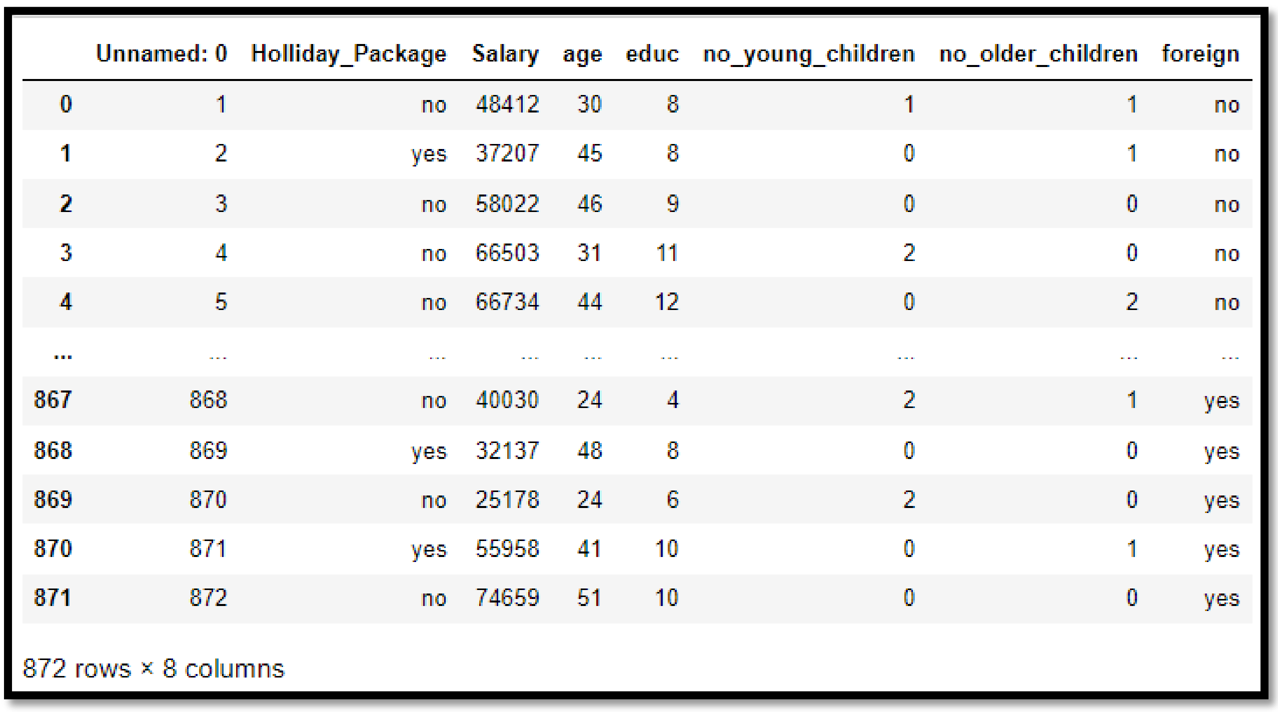
Data Dictionary for Holiday Package

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

Table 13. Data Dictionary

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.

Sample of the Dataset



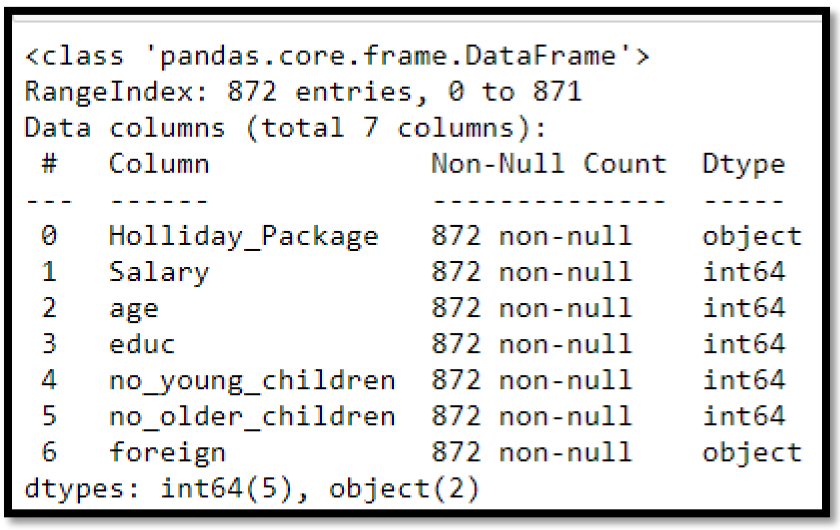
	Unnamed: 0	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	1	no	48412	30	8	1	1	no
1	2	yes	37207	45	8	0	1	no
2	3	no	58022	46	9	0	0	no
3	4	no	66503	31	11	2	0	no
4	5	no	66734	44	12	0	2	no
...	...	...	...	...	...	...	...	...
867	868	no	40030	24	4	2	1	yes
868	869	yes	32137	48	8	0	0	yes
869	870	no	25178	24	6	2	0	yes
870	871	yes	55958	41	10	0	1	yes
871	872	no	74659	51	10	0	0	yes

872 rows × 8 columns

Table 14. Sample Dataset

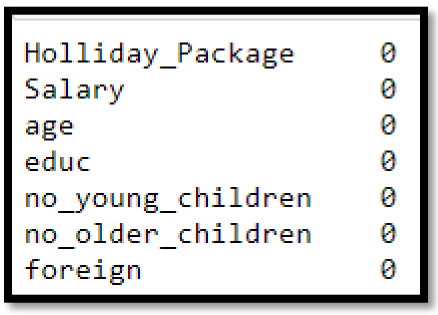
We will be removing Unnamed: 0 column as it adds no value for our analysis.

Information on the dataset



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   Holliday_Package   872 non-null    object
 1   Salary             872 non-null    int64 
 2   age                872 non-null    int64 
 3   educ               872 non-null    int64 
 4   no_young_children  872 non-null    int64 
 5   no_older_children  872 non-null    int64 
 6   foreign            872 non-null    object
dtypes: int64(5), object(2)
```

Table 15. Info of the dataset



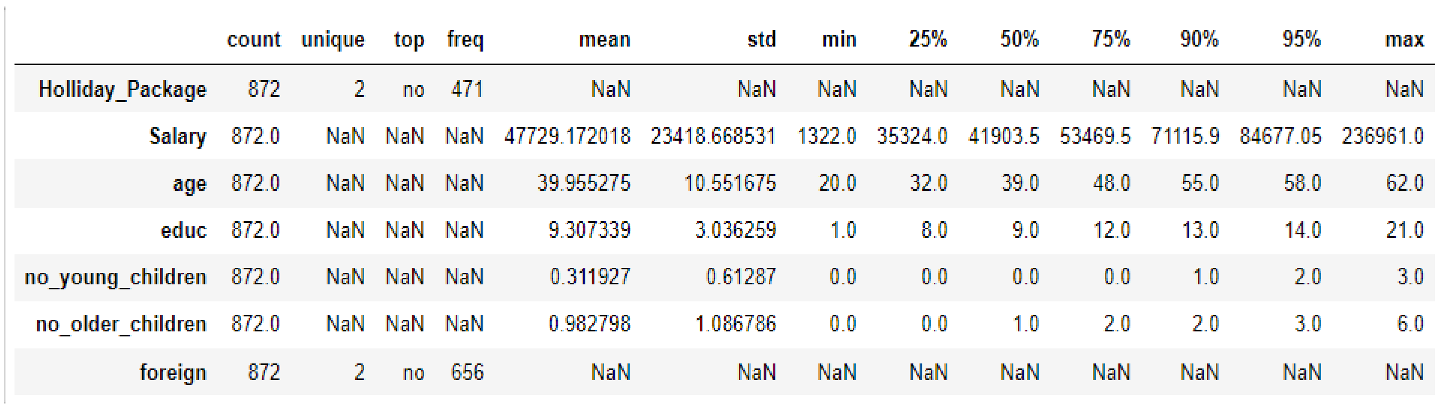
```
Holliday_Package     0
Salary               0
age                  0
educ                 0
no_young_children    0
no_older_children    0
foreign              0
```

Table 16. Missing values

1. The dataset has 872 rows and 8 columns. Since we have removed the first column 'Unnamed: 0' we are now left with 7 columns.
2. As per the info we can see that there are Zero null value counts. There are 5 integer and 2 object variable data types.
3. There are NO missing values in the given dataset.
4. There are NO duplicate values in the dataset.

Descriptive Statistics

Summary statistics or 5-point summary helps us to understand the Interquartile Range like minimum, maximum, 25th, 50th and 75th percentiles, mean or average, standard deviation and count of data observations etc. The most popular descriptive statistics are Measures of Central Tendency which are Mean, Median and Mode which are fundamental to any data analysis.



	count	unique	top	freq	mean	std	min	25%	50%	75%	90%	95%	max
Holliday_Package	872	2	no	471	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Salary	872.0	NaN	NaN	NaN	47729.172018	23418.668531	1322.0	35324.0	41903.5	53469.5	71115.9	84677.05	236961.0
age	872.0	NaN	NaN	NaN	39.955275	10.551675	20.0	32.0	39.0	48.0	55.0	58.0	62.0
educ	872.0	NaN	NaN	NaN	9.307339	3.036259	1.0	8.0	9.0	12.0	13.0	14.0	21.0
no_young_children	872.0	NaN	NaN	NaN	0.311927	0.61287	0.0	0.0	0.0	0.0	1.0	2.0	3.0
no_older_children	872.0	NaN	NaN	NaN	0.982798	1.086786	0.0	0.0	1.0	2.0	2.0	3.0	6.0
foreign	872	2	no	656	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 17. Descriptive Statistics

From the above table below are the observations.

1. Out of **872** employees **471** employees have NOT opted for the holiday package which explains why the company is trying to predict whether an employee will opt or not.
2. The average employee salary is **47729.17** whereas median salary is **41903.5** with a maximum salary of **236961.0**. Looks like this is a **right skewed** distribution. Upto **95%** of the data is below the salary of **84677.05** which suggests that there might be outliers which is again natural. There might be employees who are working for more than **25 yrs** with a high salary which explains the skewness in the data.
3. Employee age ranges from a minimum of **20 yrs** to a maximum of **62 yrs**. The average and the median age is **39 yrs**. Hence this seems to be **normally distributed** which is also naturally true.
4. Years of formal education range from a minimum of **1 yr** to a maximum of **21 yrs**. The mean/median years of education are **9 yrs**. However it raises sufficient suspicions on the minimum 1 yr education. Whether this is incorrect data entry or something correct should be validated by tours and travel agency. However we will keep it as is for now.
5. No of young children (younger than 7 yrs) range from **0 to 3** which suggests that none of the employees have more than 3 children (younger than 7 yrs).
6. No of older children range from **0 to 6**. Adding to it **95%** of the data shows that no of older children are below **3**. Hence we might see outliers in this attribute.
7. Out of **872** employees we have **656** employees who are NOT foreigners. Hence there is a balance of **85:15** ratio of domestic and foreign employees.

Exploratory Data Analysis (EDA)

Univariate Analysis

Countplots

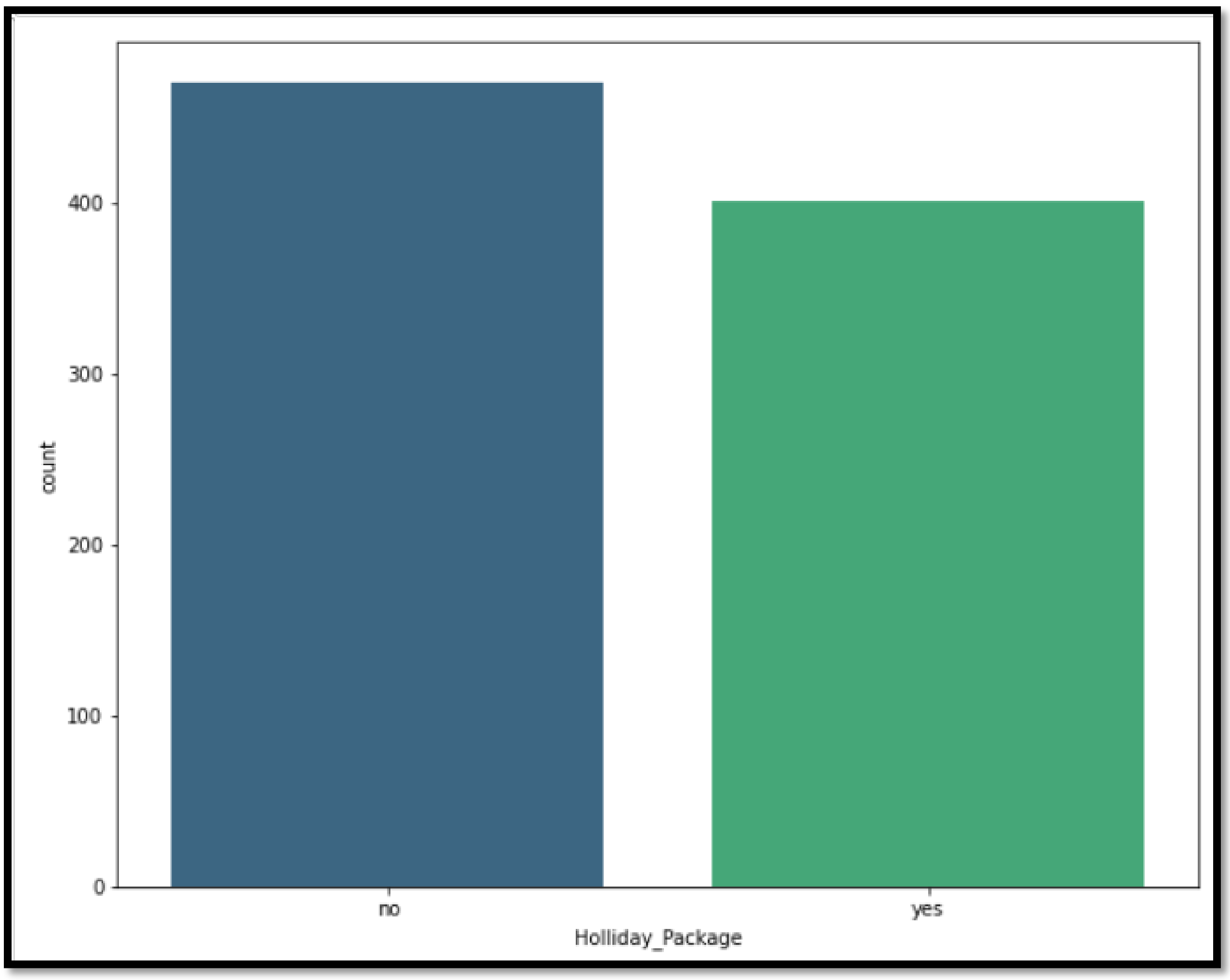


Figure 13. Count of Holiday Package

From the above we can see that majority of the employees have NOT opted for a holiday package.

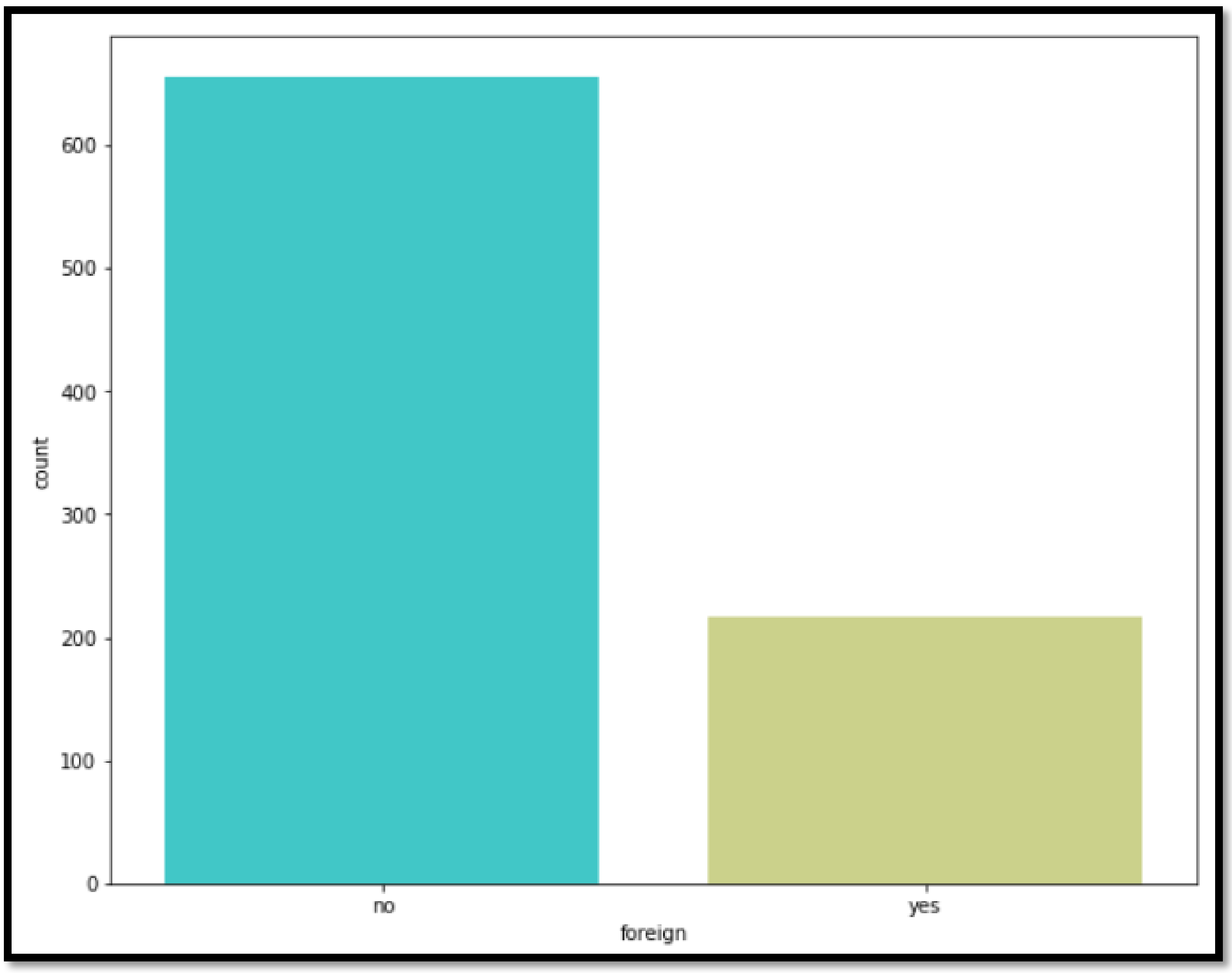


Figure 14. Count of foreign

As explained earlier, majority of the employees are domestic. We have less than half of employees who are foreigners.

Boxplots

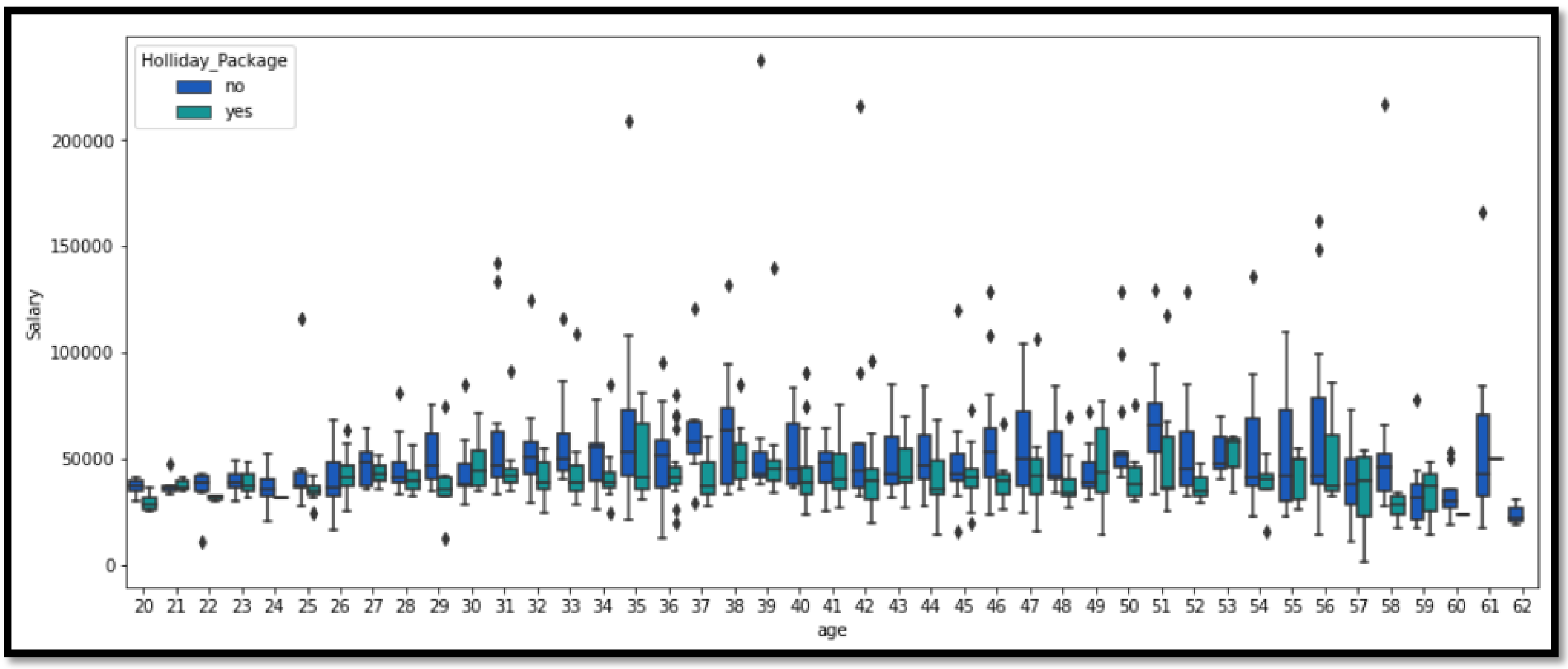


Figure 15. Salary w.r.t Age (Holiday_Package)

From the above plot we can see that salary is centered around the range of 30 yrs to 50 yrs of age. Employees who are more than 50 yrs have opted for holiday package which suggests that they are in the verge of retiring and hence are taking a break from their career.

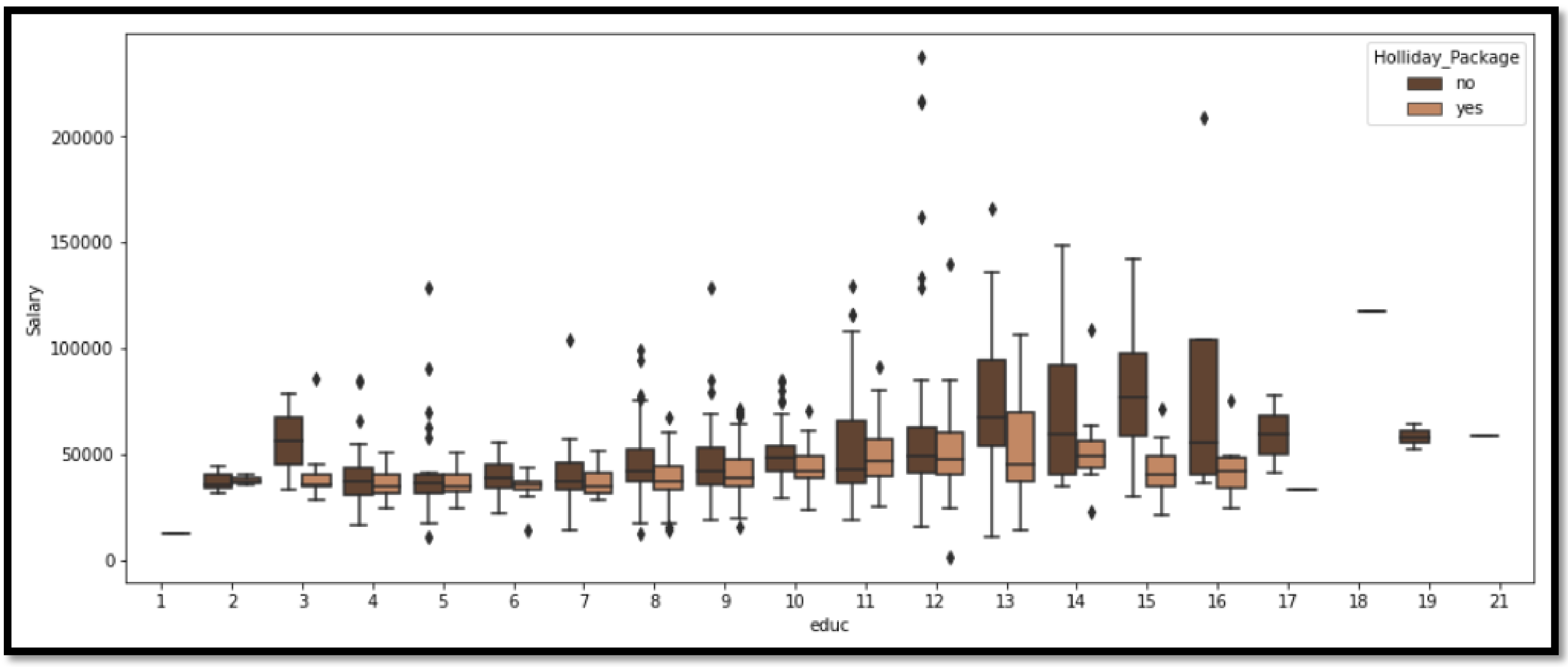


Figure 16. Salary w.r.t educ (Holiday_Package)

Employees who have completed 13yrs to 16yrs are the highest in number. Some of the employees with 12 yrs education are getting exuberant higher salaries which are evident as outliers seen in the above plot.

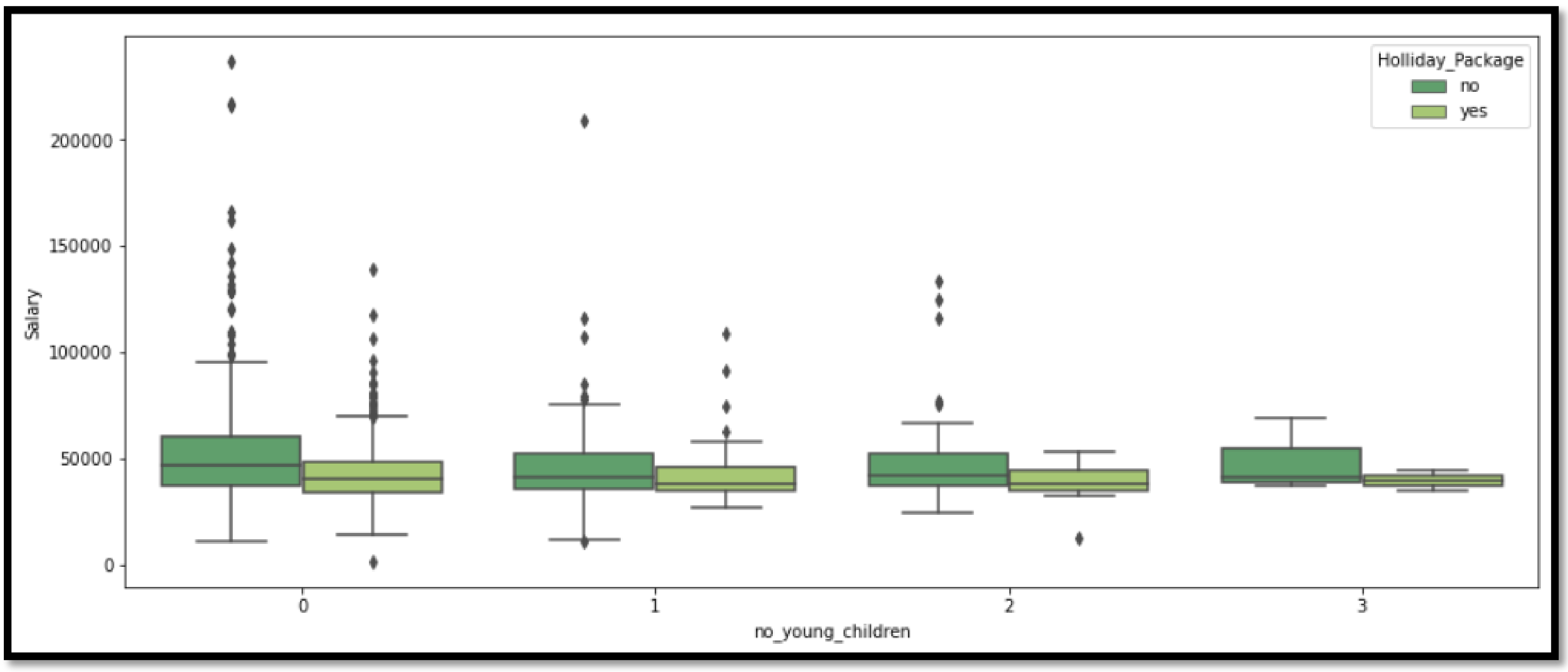


Figure 17. Salary w.r.t no_young_children (Holiday_Package)

Employees who have NO younger children (younger than 7 yrs) have a higher salary range which also extends beyond 2 lakh whereas employees who have 3 children are the least with salary between 50,000 to 100000.

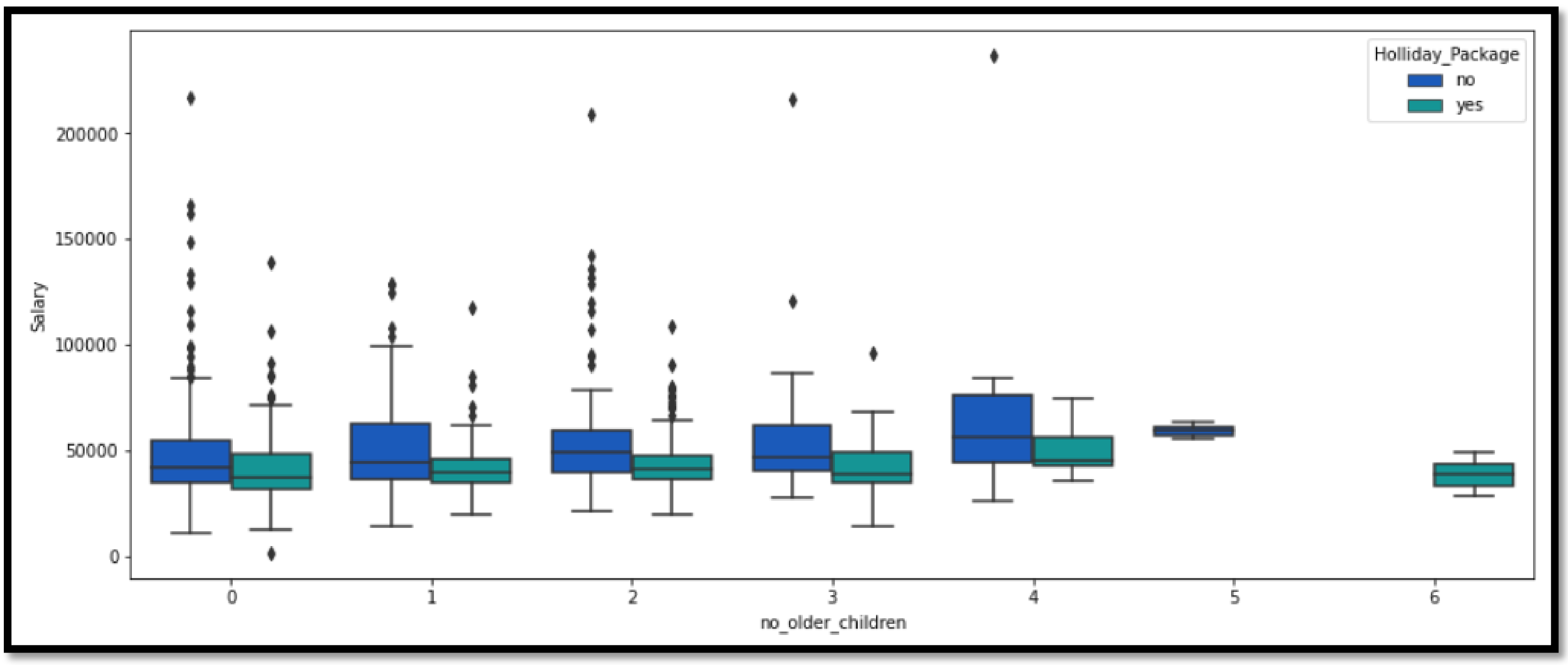


Figure 18. Salary w.r.t no_older_children (Holiday_Package)

Similarly employees who have NO older children have a higher salary range. However the median salary range remains almost same for most of the employees who have older children.

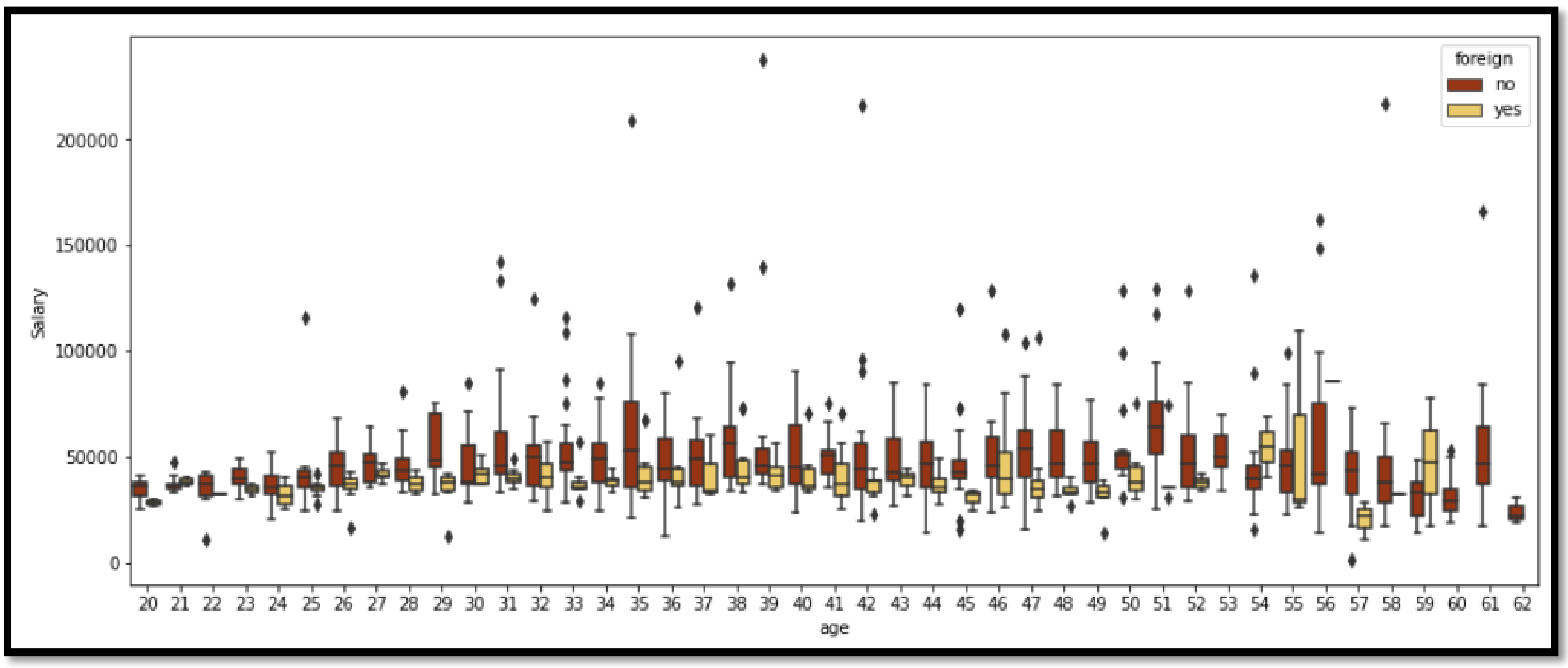


Figure 19. Salary w.r.t age (foreign)

Employees who have not opted for holiday package are more in numbers but we can see outliers across range of age. Non-foreign employees are higher in number.

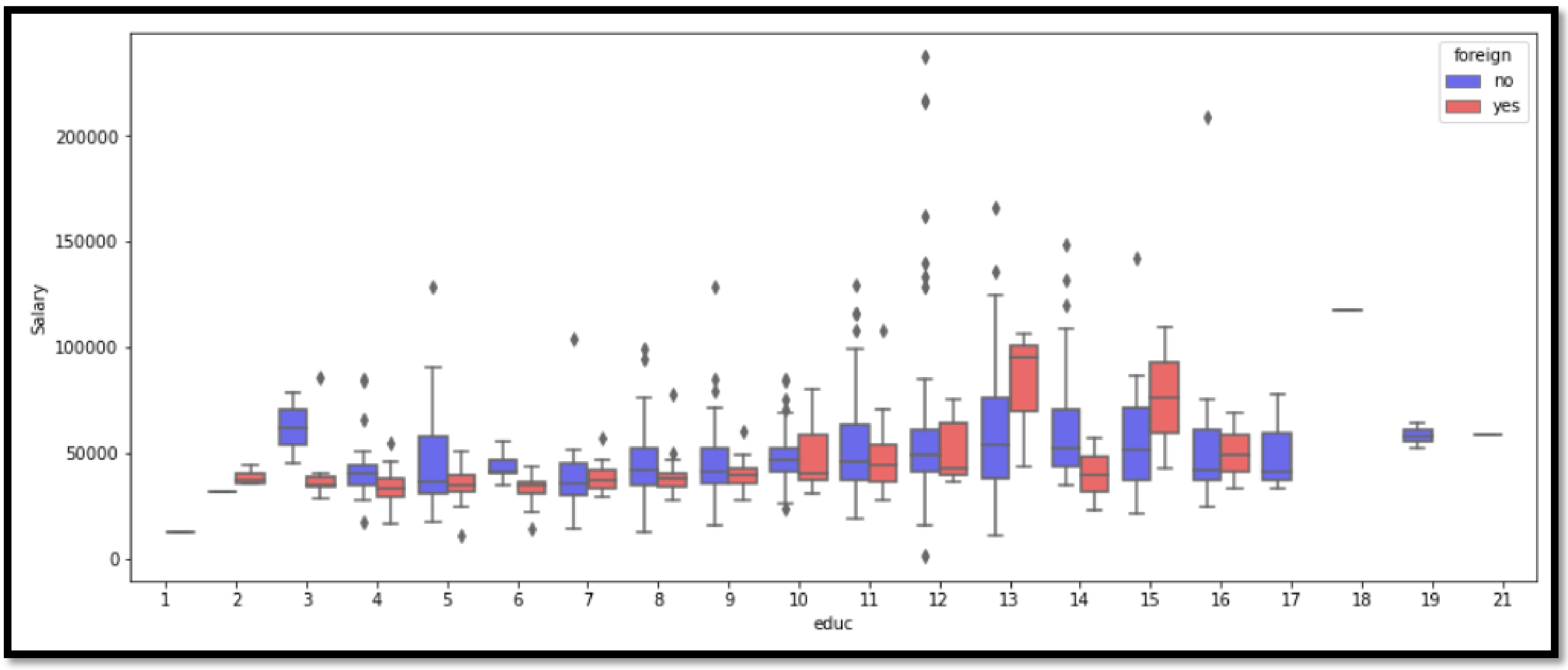


Figure 20. Salary w.r.t educ (foreign)

Employees with 11yrs to 17yrs of education are more in numbers which indicates that they are mostly under graduates or post graduates. Employees who are non-foreign with 12 yrs of education have the highest salary range.

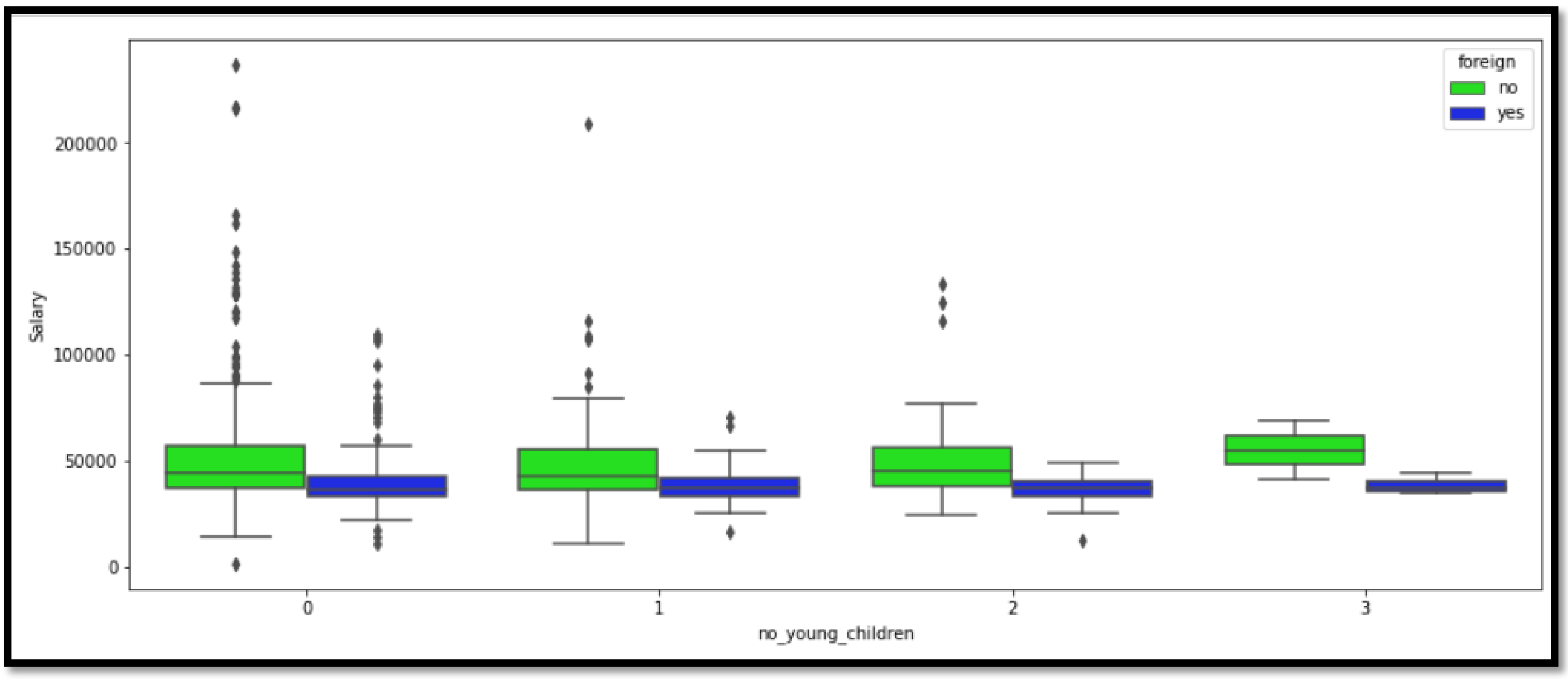


Figure 21. Salary w.r.t no_young_children (foreign)

Employees who have NO younger children have a higher salary range. Non foreign employees are getting more salary compared to foreign employees.

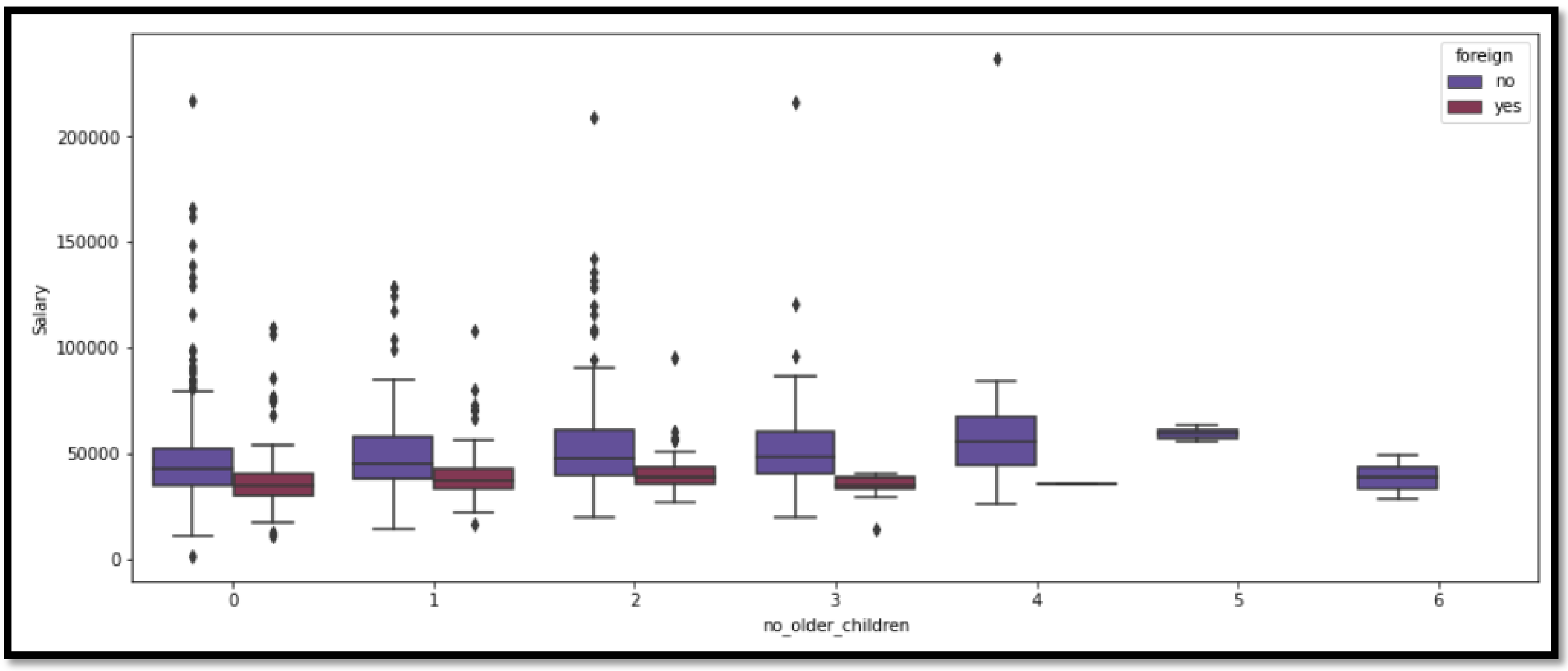


Figure 22. Salary w.r.t no_older_children (foreign)

Employees who have older children between 0 to 3 have a higher salary range whereas those who have more than 3 older children are in the moderate salary range. Again non foreign employees are higher.

Boxplots of Numerical variables

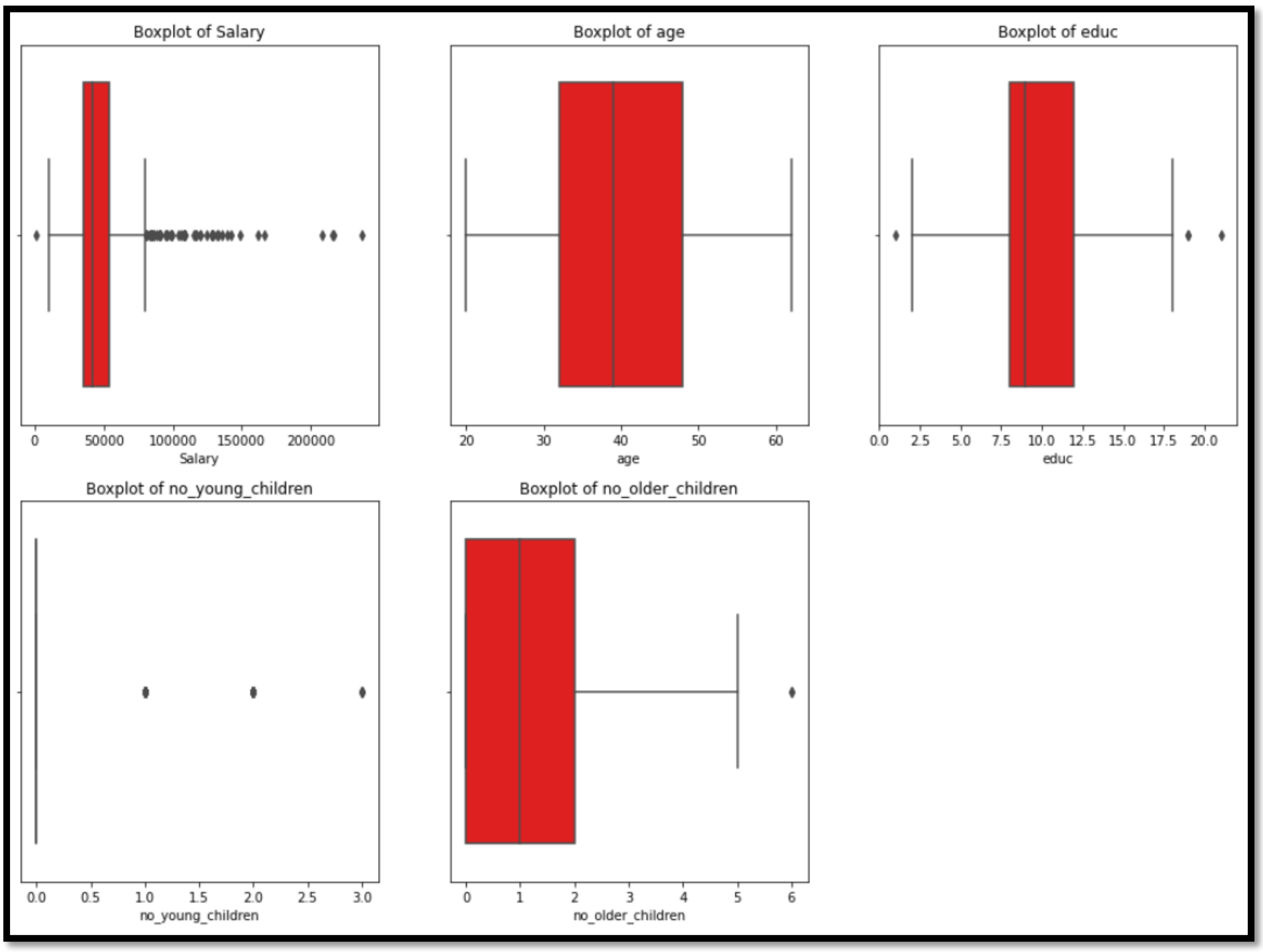


Figure 23. Boxplots of numerical variables

Distribution Plots of Numerical variables

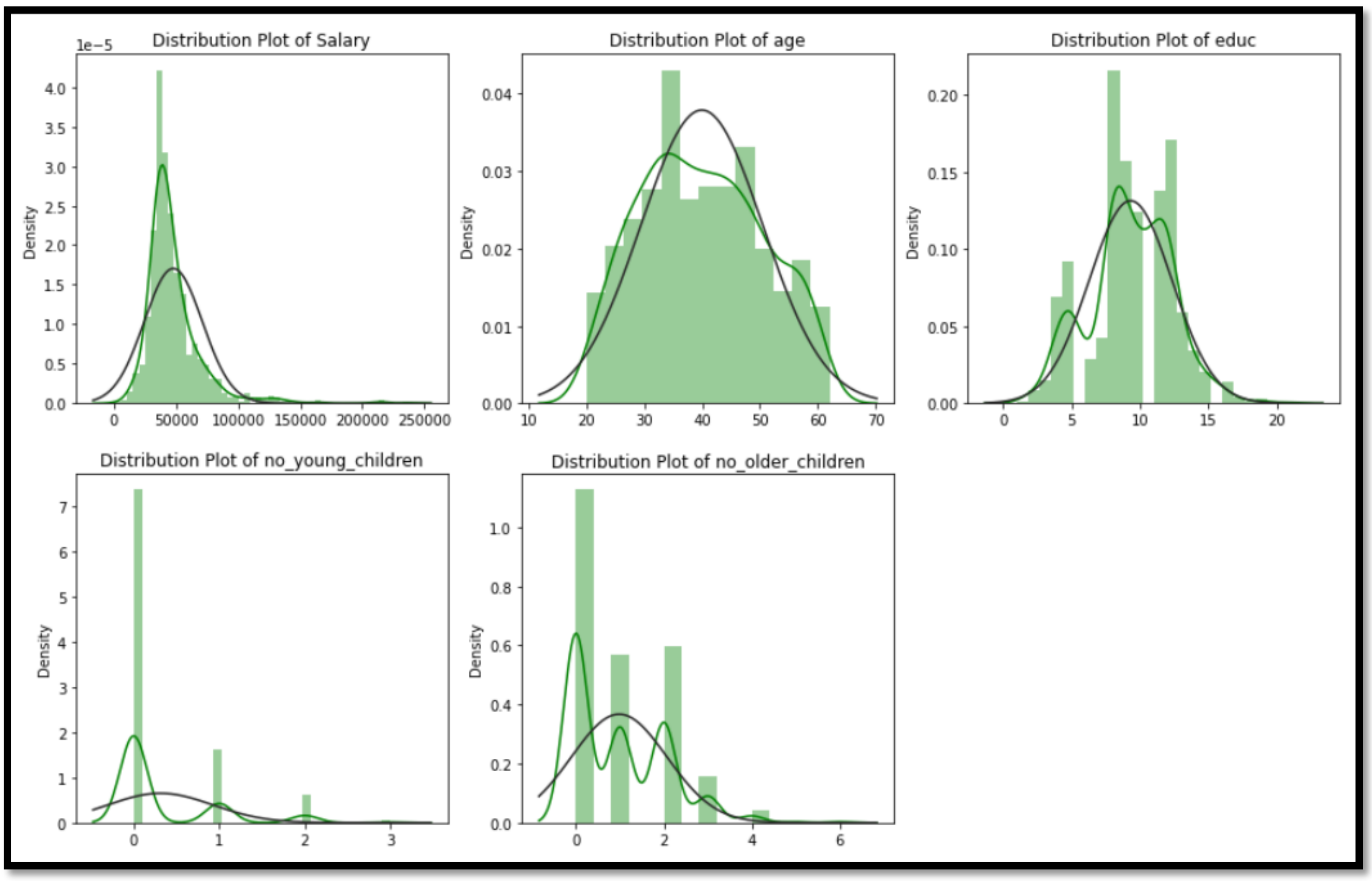
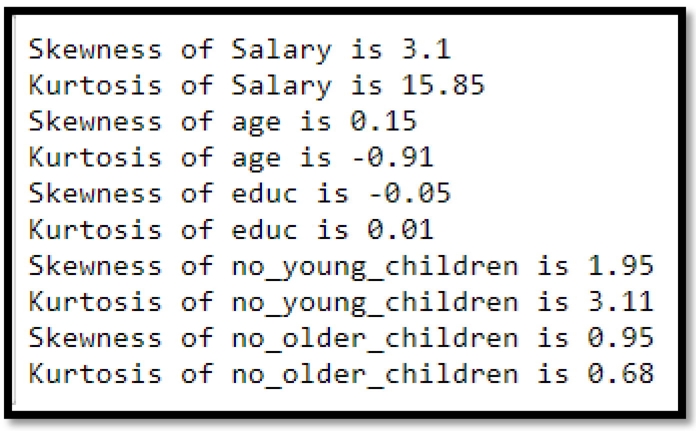


Figure 24. Distribution plots of numerical variables

From the above distribution plots we can infer the below:

1. Age is normally distributed.
2. Education is normally distributed with multiple peaks.
3. Salary is right skewed distribution.
4. No of young children is right skewed distribution with multiple peaks.
5. No of older children is right skewed distribution with multiple peaks.

Skewness and Kurtosis



```
Skewness of Salary is 3.1
Kurtosis of Salary is 15.85
Skewness of age is 0.15
Kurtosis of age is -0.91
Skewness of educ is -0.05
Kurtosis of educ is 0.01
Skewness of no_young_children is 1.95
Kurtosis of no_young_children is 3.11
Skewness of no_older_children is 0.95
Kurtosis of no_older_children is 0.68
```

Table 18. Skewness and Kurtosis

The skewness in Salary and no_young_children is clearly explained from the above values. Skewness values closer to Zero indicate they are normally distributed.

As per Fisher's definiton, the kurtosis of the normal distribution is zero. The distribution with a higher kurtosis has a heavier tail. In this case Salary and no_young_children have a heavier tail which is also evident from the distribution plots. Kurtosis values closer to Zero indicate they are normally distributed.

Pairplot

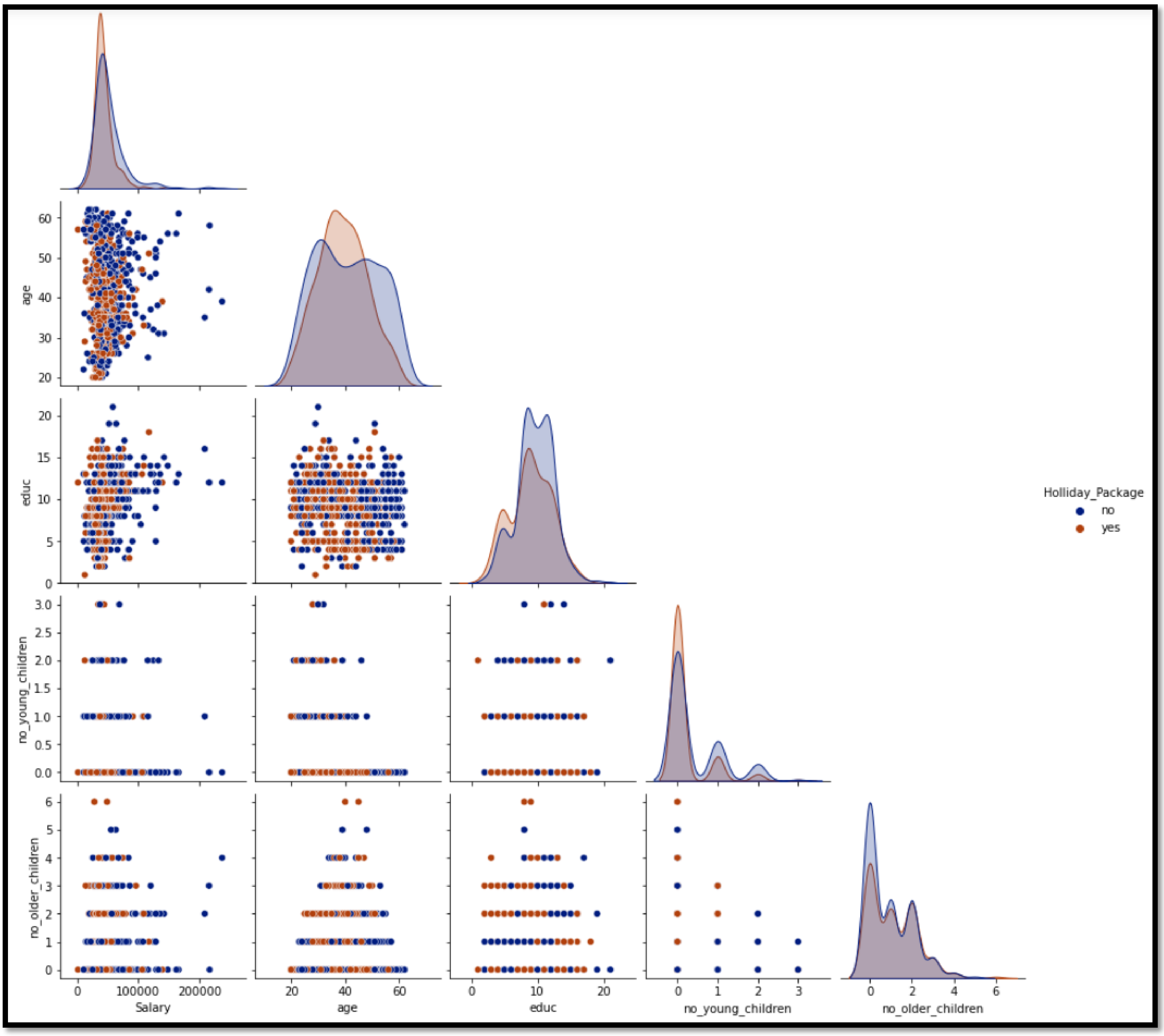


Figure 25. Pairplot

Above pairplot show that none of the variables are correlated. We will also validate the same in the correlation heatmap. The scattered data points clearly indicate that employees who have NOT opted for package are more in numbers.

Correlation Heatmap

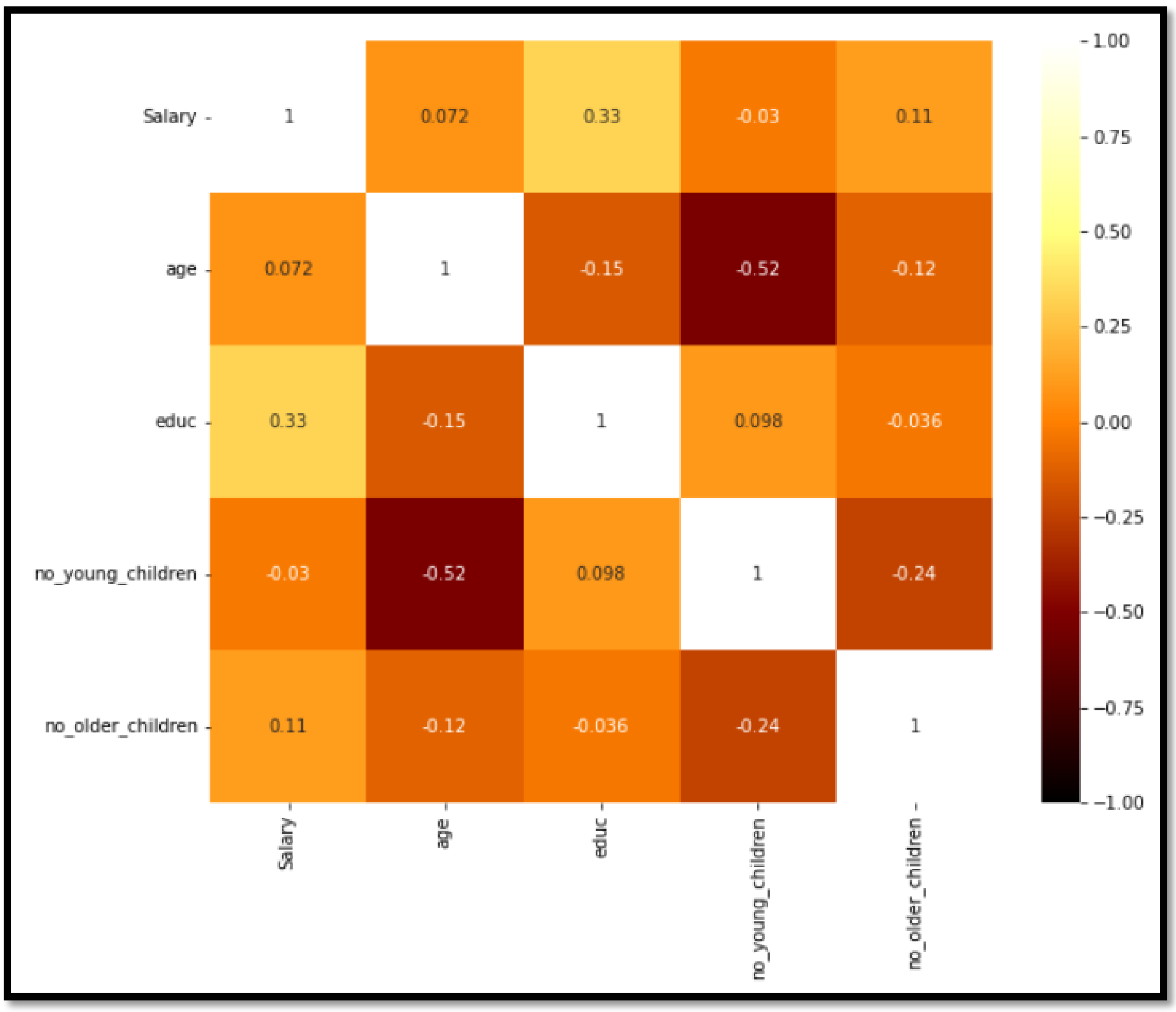
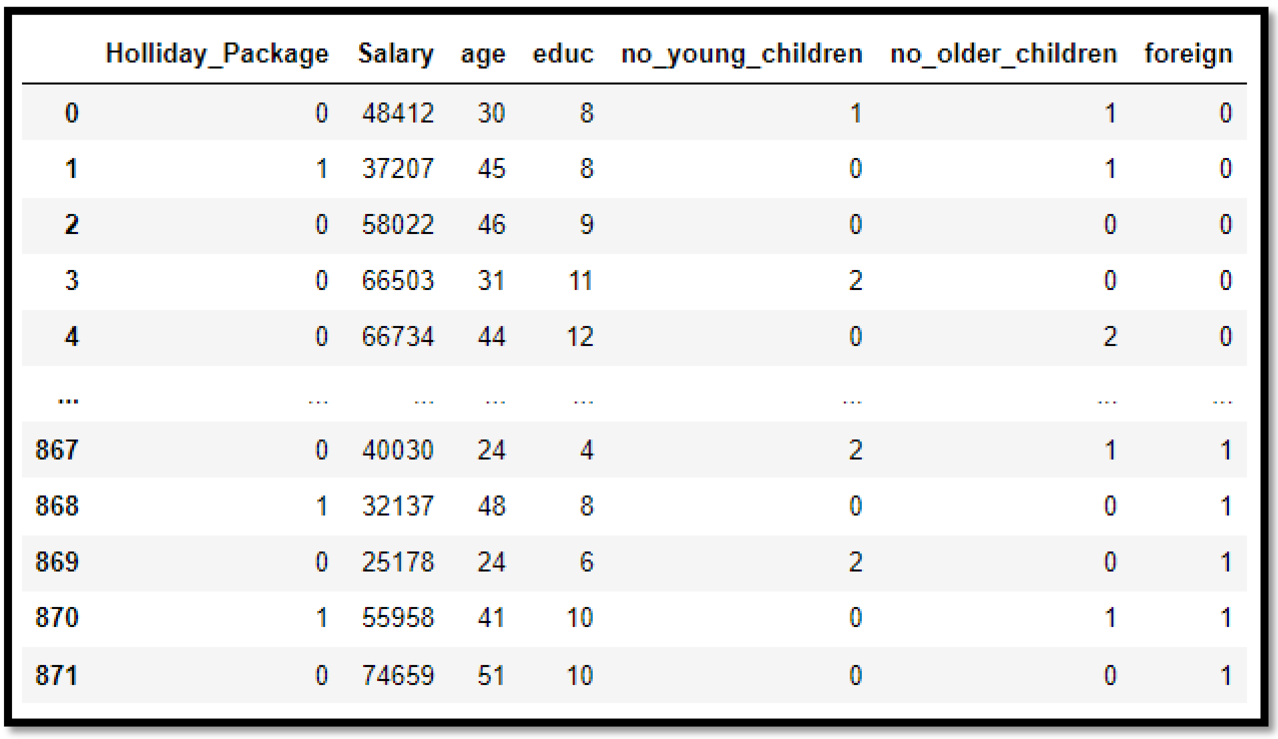


Figure 26. Correlation Heatmap

Correlation heatmap clearly evidences that there exists NO correlation between the attributes. However age and no_young_children have a less than moderate negative relationship with a correlation coefficient of -0.52 which can be considered as trivial. Correlation values are always between 1 and -1. Those which are closer to 1 are positively correlated and those which near -1 are negatively correlated. Values near to 0 have no correlation.

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 variables



	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	0	48412	30	8	1	1	0
1	1	37207	45	8	0	1	0
2	0	58022	46	9	0	0	0
3	0	66503	31	11	2	0	0
4	0	66734	44	12	0	2	0
...	...	...	...	...	...	...	...
867	0	40030	24	4	2	1	1
868	1	32137	48	8	0	0	1
869	0	25178	24	6	2	0	1
870	1	55958	41	10	0	1	1
871	0	74659	51	10	0	0	1

Table 19. Dataset after encoding

Logistic Regression

Also known as Logit , Maximum-Entropy classifier, is a supervised learning method for classification. It establishes relation between dependent class variable and independent variables using regression.

The dependent variable is categorical i.e. it can take only integral values representing different classes.

The probabilities describing the possible outcomes of a query point are modeled using a logistic function.

Belongs to family of discriminative classifiers. They rely on attributes which discriminate the classes well.

There are two broad categories of Logistic Regression algorithms:
a. **Binary** Logistic Regression when dependent variable is strictly binary
b. **Multinomial** Logistic Regression when the dependent variable has multiple categories.

There are two types of Multinomial Logistic Regression
I. **Ordered** Multinomial Logistic Regression (dependent variable has ordered values)
II. **Nominal** Multinomial Logistic Regression (dependent variable has unordered categories)

The main differences between linear and logistic regression are:

I. Correlation between a binary response and a continuous predictor is not defined. Hence scatterplots are not useful in this case.
II. The numerical value assigned to the two levels of a binary response is arbitrary. Consider, for example, the case of loan default. Physically the two levels are defaulters and non-defaulters. It is possible to assign values 0 and 1 respectively to the two levels; it is also possible to assign 1 and 0; or 1 and -1; or any other arbitrary set of numbers. Hence it is not possible to regress Y on the predictors.
III. Therefore, Prob (Y = 1), i.e. probability of Y at a given level, is modelled through a regression.
IV. Since probability of an event is a number between 0 and 1, linear relationship between Prob (Y = 1) and the set of predictors is not possible. Hence a suitable transformation is necessary.

Odds, Log Odds and Logit Transform

The concept of odds ratio is related to probability. Formally

$$\textbf{Odds of Success} = \boldsymbol{Pr(Success)} \,/\, \mathbf{1} - \boldsymbol{Prob(Success)}$$

If success and failure have equal probability, then the numerical value of odds is 1. As success probability increases, odds increases. While probability is a number between 0 and 1, there is no such restriction on odds. It is a positive number, taking a very small value when success probability is small but may have a high numerical value if success probability is large. Theoretically, odds is a number between (0, +∞). Another related quantity is called log odds which is defined as log$e$ $Odds$. Since odds is a positive quantity, log odds is always defined, but its value may be positive or negative. When value of odds is 1, log odds is 0. When the two outcomes, success and failure, are equi-probable, i.e., each probability is 50%, odds is 1 and log odds is 0. When probability of success is less than 50%, odds is less than 1 and log odds is a negative quantity. When probability of success is more than 50%, the value of odds is more than 1 and log odds is a positive quantity. Theoretically, log odds is a number between (-∞, +∞). In logistic regression the log odds of success is modelled as a linear function of the predictors. Log odds of success is also known as logit transformation. Mathematically it may be claimed that a logit transform maps a number between 0 and 1 on a real line; logistic transformation is its reverse where a real number is converted into a probability. This is the genesis of the name logistic regression.

Logistic Regression assigns probabilities to different classes to which a query point is likely to belong. To do so, it learns from the training set vectors of weights and bias. Each weight $(\boldsymbol{w_i})$ is assigned to one input feature$\boldsymbol{X_i}$. The weight assigned to each feature represents how important that feature is for classification decision.

The weights can be positive i.e. direct correlation of the feature with the class of interest while a negative weight indicates inverse relation with the class of interest. To classify a query point, the classifier takes the weight sum of the features and the bias to represent the evidence of the query point belonging to the class of interest.
1. **Z = w.x+ b**

Since the weights are running numbers and so is the bias term, **Z** can take values from –infinity to +infinity. To transform the value of Z into probability (range between 0 and 1). Z is passed through **Sigmoid function** (mathematical transformation).

1. **P(y= 1) = Sigmoid(Z) = 1/(1 +e^{-z})**
2. **P(y= 0) = 1 –P(y =1) = 1 –(1/(1 +e^{-z})) = e^{-z}/ (1 + e^{-z})**
3. **y = 1 if P(y=1|X) > .5, else y = 0**

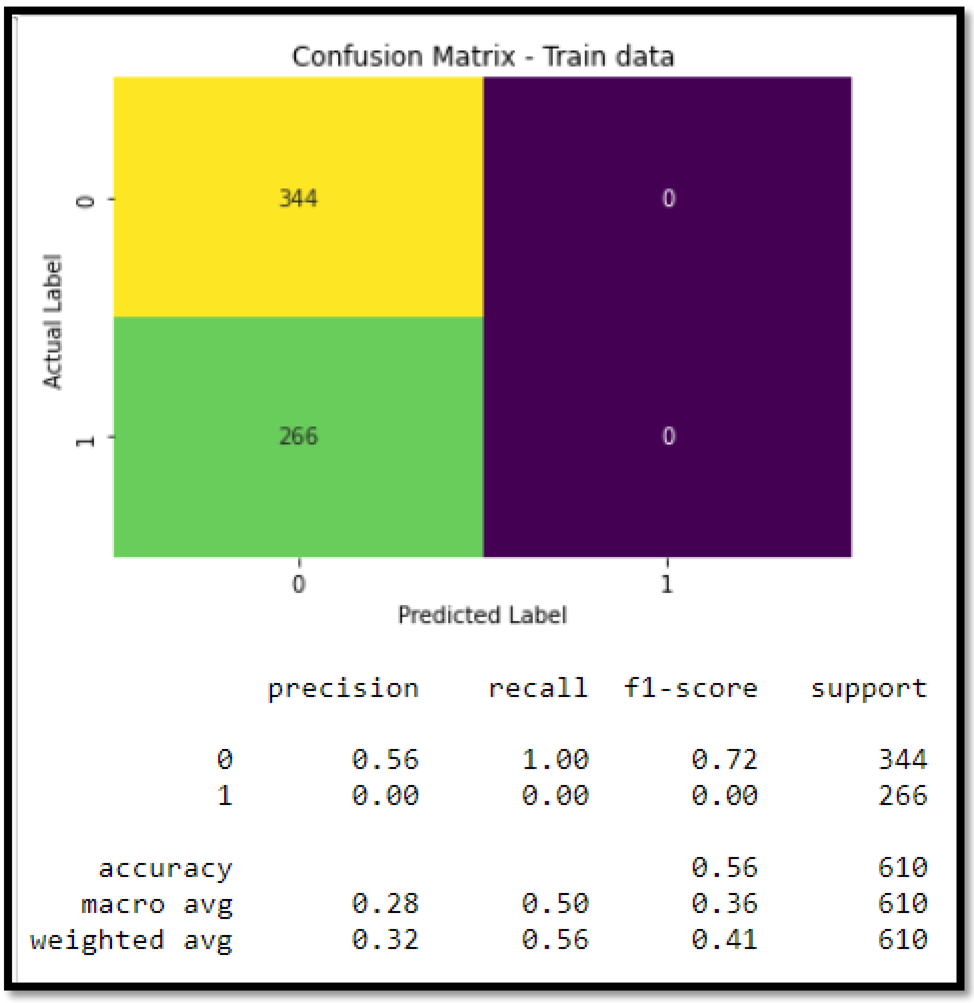
The algorithm uses cross-entropy **loss function** (negative log likelihood loss) to find the most optimal weights and bias across entire data set put together (N records)

$$\boldsymbol{logLoss} = \frac{-\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{N}}\big(\boldsymbol{y_i}(\boldsymbol{logp_i})\big) + \ (\mathbf{1}-\boldsymbol{y_i})\boldsymbol{log}(\mathbf{1}-\boldsymbol{p_i})$$

Most optimal weights and bias would be those that minimize overall all training error i.e. misclassification in the training data.

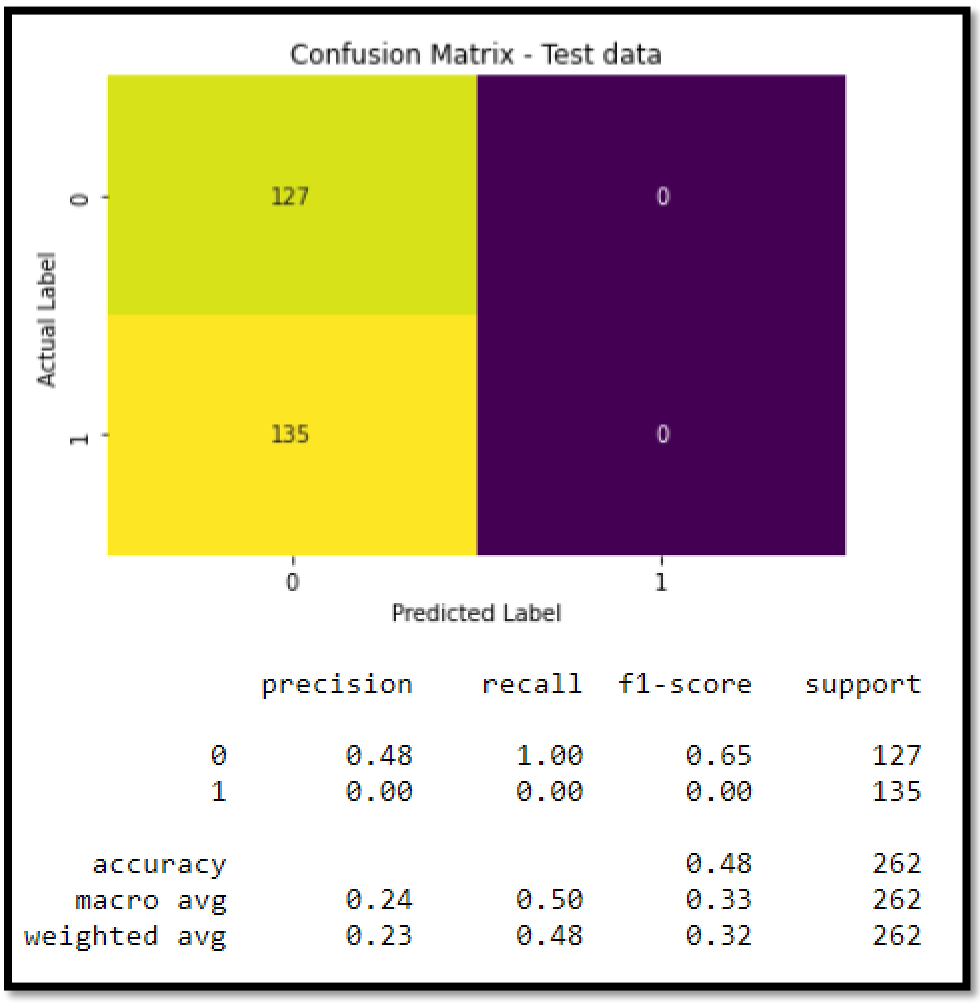
Source: Google

Model 1 – Logistic Regression using Sklearn



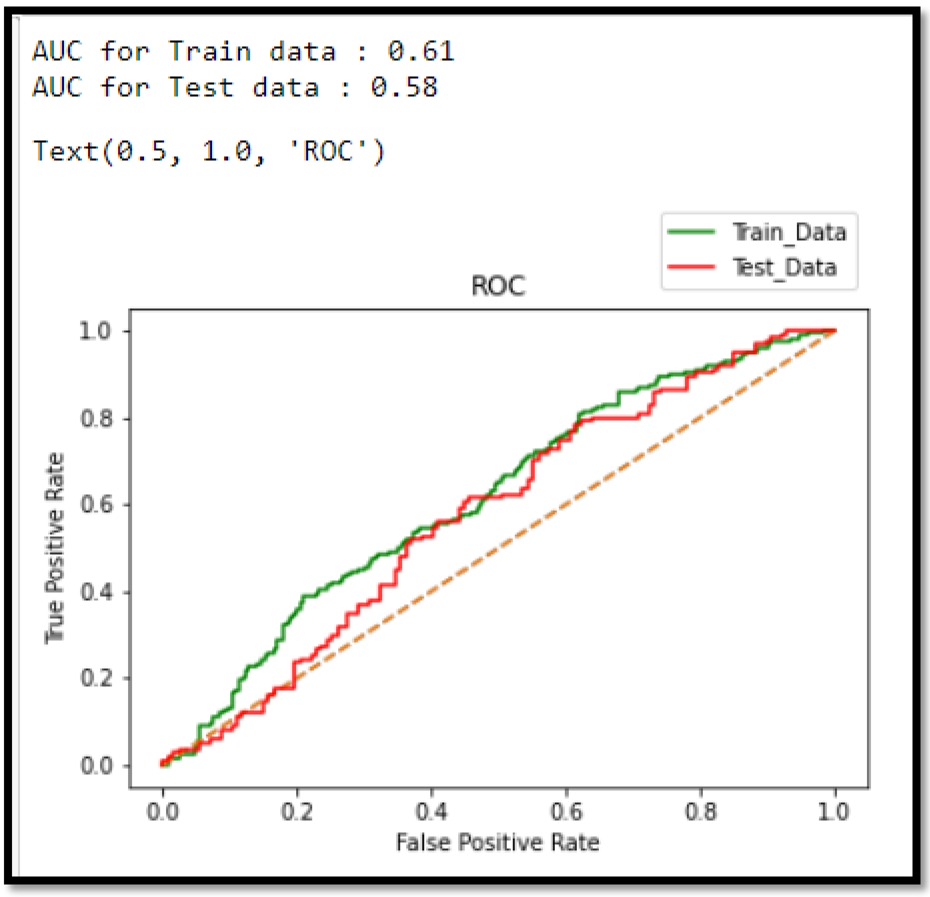
```
              precision    recall  f1-score   support

           0       0.56      1.00      0.72       344
           1       0.00      0.00      0.00       266

    accuracy                           0.56       610
   macro avg       0.28      0.50      0.36       610
weighted avg       0.32      0.56      0.41       610
```

Figure 27. Confusion Matrix and Classification Report - Training Data



```
              precision    recall  f1-score   support

           0       0.48      1.00      0.65       127
           1       0.00      0.00      0.00       135

    accuracy                           0.48       262
   macro avg       0.24      0.50      0.33       262
weighted avg       0.23      0.48      0.32       262
```

Figure 28. Confusion Matrix and Classification Report - Testing Data

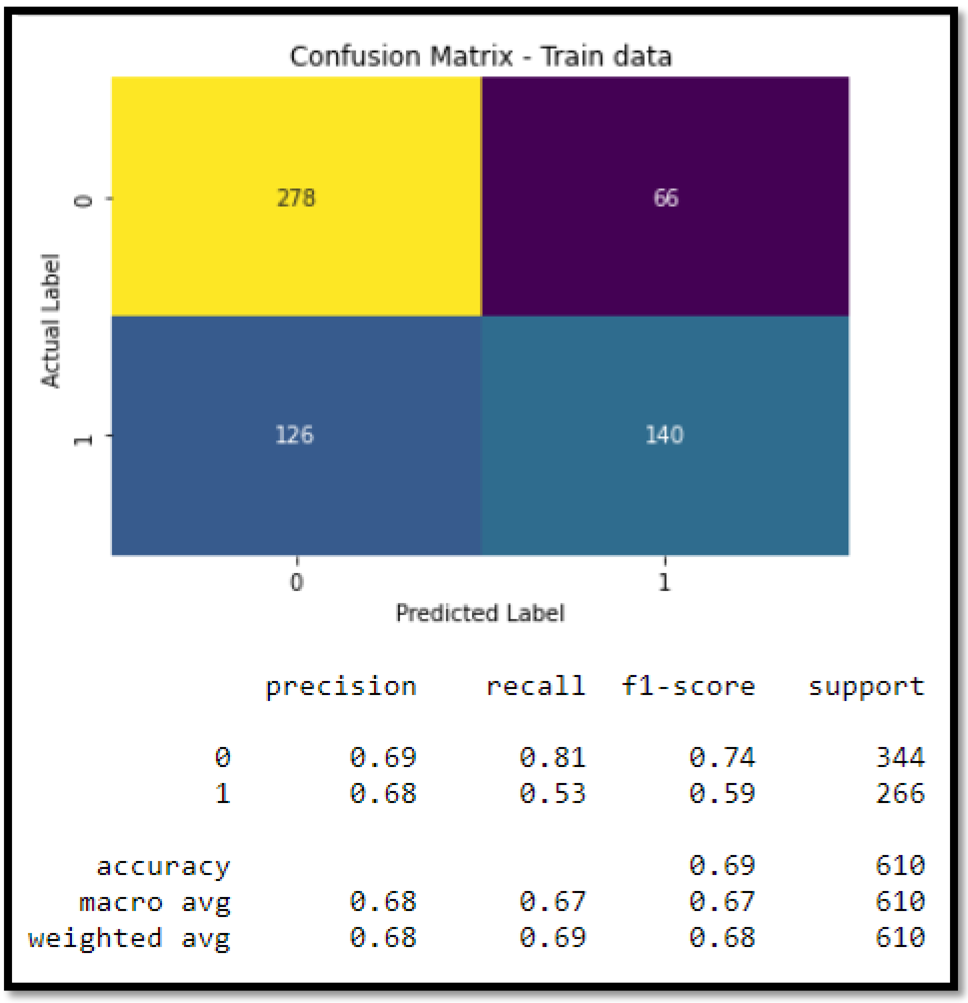


```
AUC for Train data : 0.61
AUC for Test data : 0.58

Text(0.5, 1.0, 'ROC')
```

Figure 29. AUC and ROC

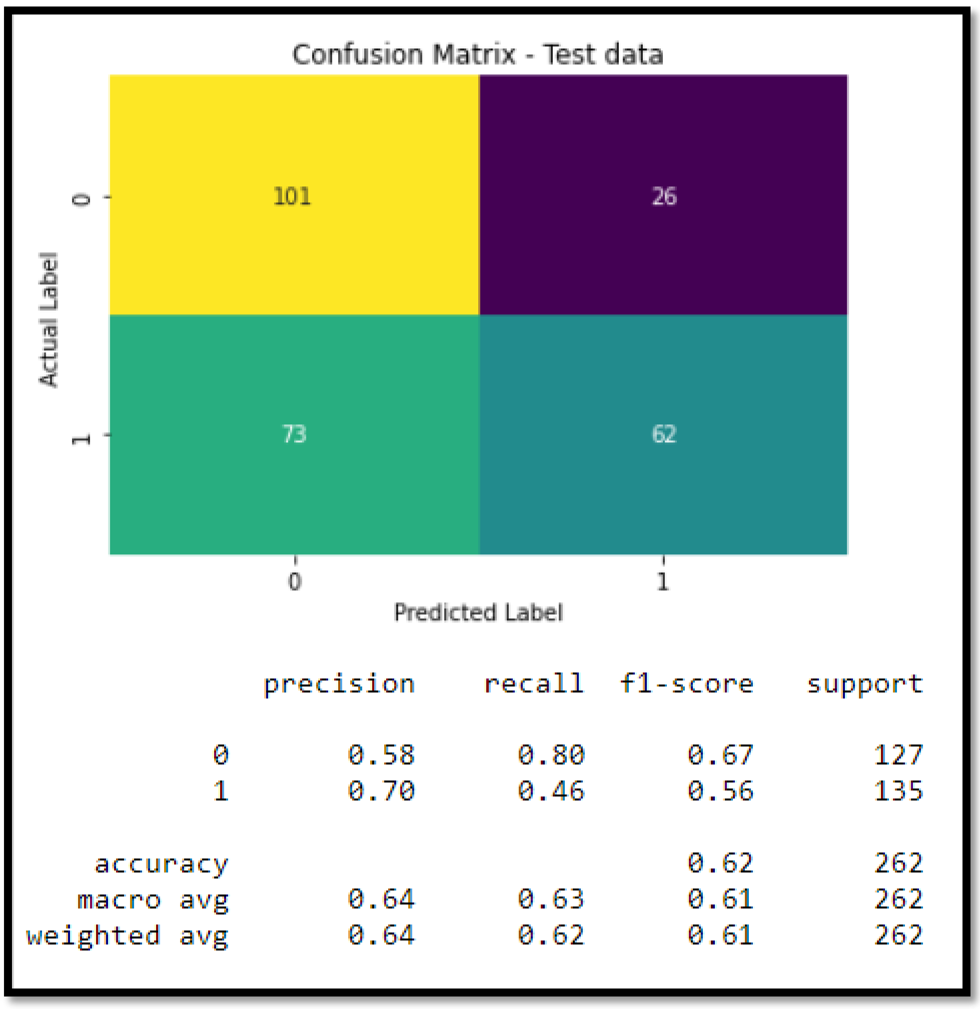
When we used solver=”lbfgs” - The accuracy for train data is 0.56 whereas for test data it is 0.48. AUC for train data is 0.61 whereas AUC for test data is 0.58. However the model can do better.

Model 2 – Logistic Regression



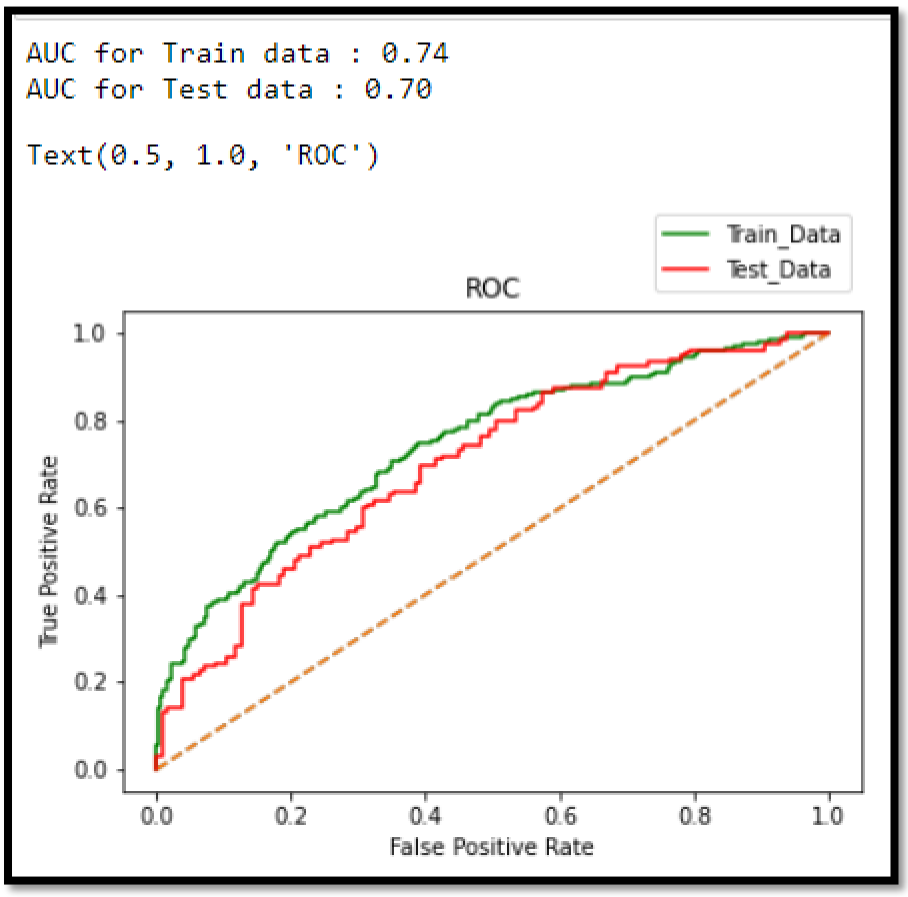
```
              precision    recall  f1-score   support

           0       0.69      0.81      0.74       344
           1       0.68      0.53      0.59       266

    accuracy                           0.69       610
   macro avg       0.68      0.67      0.67       610
weighted avg       0.68      0.69      0.68       610
```

Figure 30. Confusion Matrix and Classification Report - Train Data



```
              precision    recall  f1-score   support

           0       0.58      0.80      0.67       127
           1       0.70      0.46      0.56       135

    accuracy                           0.62       262
   macro avg       0.64      0.63      0.61       262
weighted avg       0.64      0.62      0.61       262
```

Figure 31. Confusion Matrix and Classification Report - Test Data

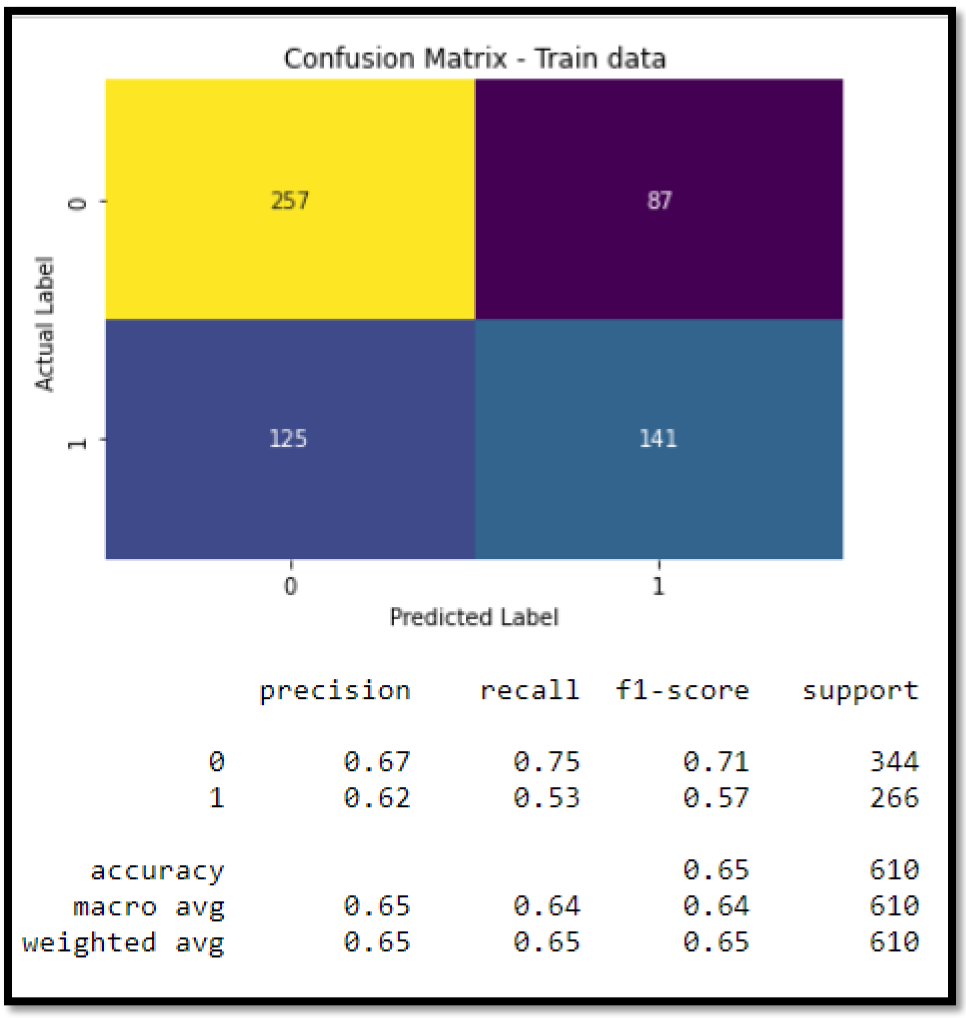


```
AUC for Train data : 0.74
AUC for Test data : 0.70

Text(0.5, 1.0, 'ROC')
```

Figure32. AUC and ROC

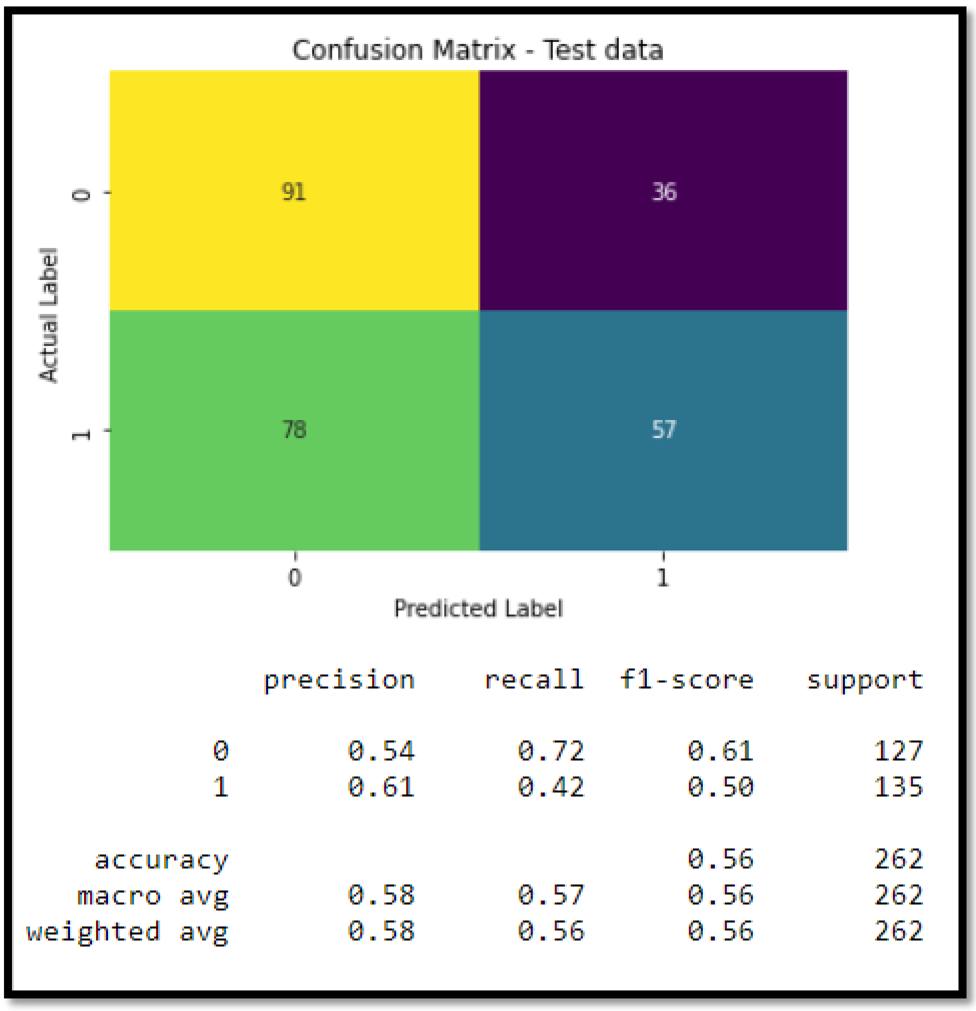
When we use solver = “newton-cg” - The accuracy for train data is 0.69 whereas for test data it is 0.62. AUC for train data is 0.74 whereas AUC for test data is 0.70. The model has performed better with this solver.

Model 3 – Logistic Regression



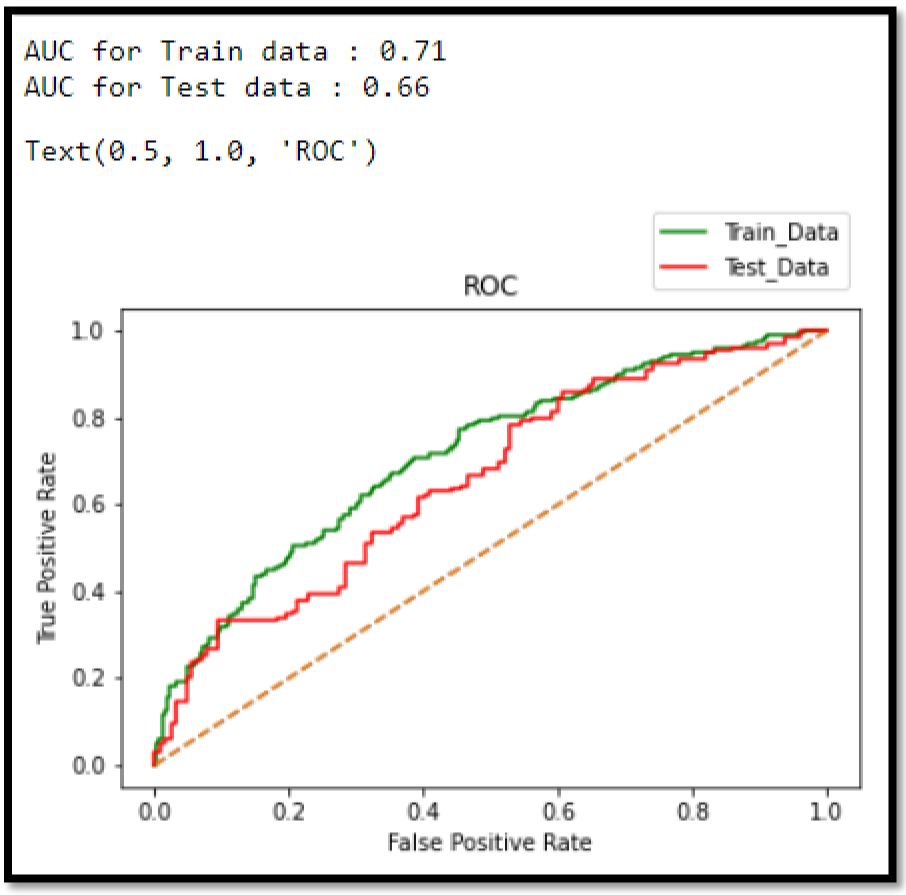
```
              precision    recall  f1-score   support

           0       0.67      0.75      0.71       344
           1       0.62      0.53      0.57       266

    accuracy                           0.65       610
   macro avg       0.65      0.64      0.64       610
weighted avg       0.65      0.65      0.65       610
```

Figure 33. Confusion Matrix and Classification Report - Train Data



```
              precision    recall  f1-score   support

           0       0.54      0.72      0.61       127
           1       0.61      0.42      0.50       135

    accuracy                           0.56       262
   macro avg       0.58      0.57      0.56       262
weighted avg       0.58      0.56      0.56       262
```

Figure 34. Confusion Matrix and Classification Report - Test Data

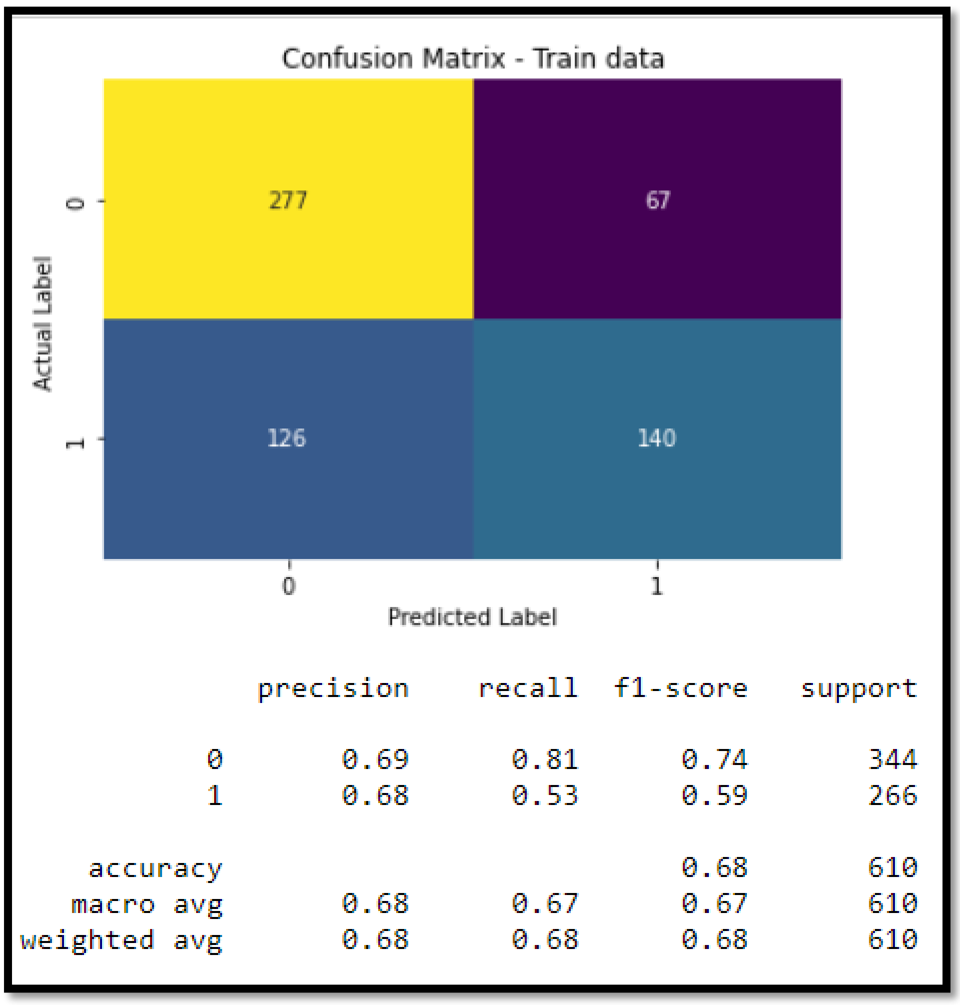


```
AUC for Train data : 0.71
AUC for Test data : 0.66

Text(0.5, 1.0, 'ROC')
```

Figure35. AUC and ROC

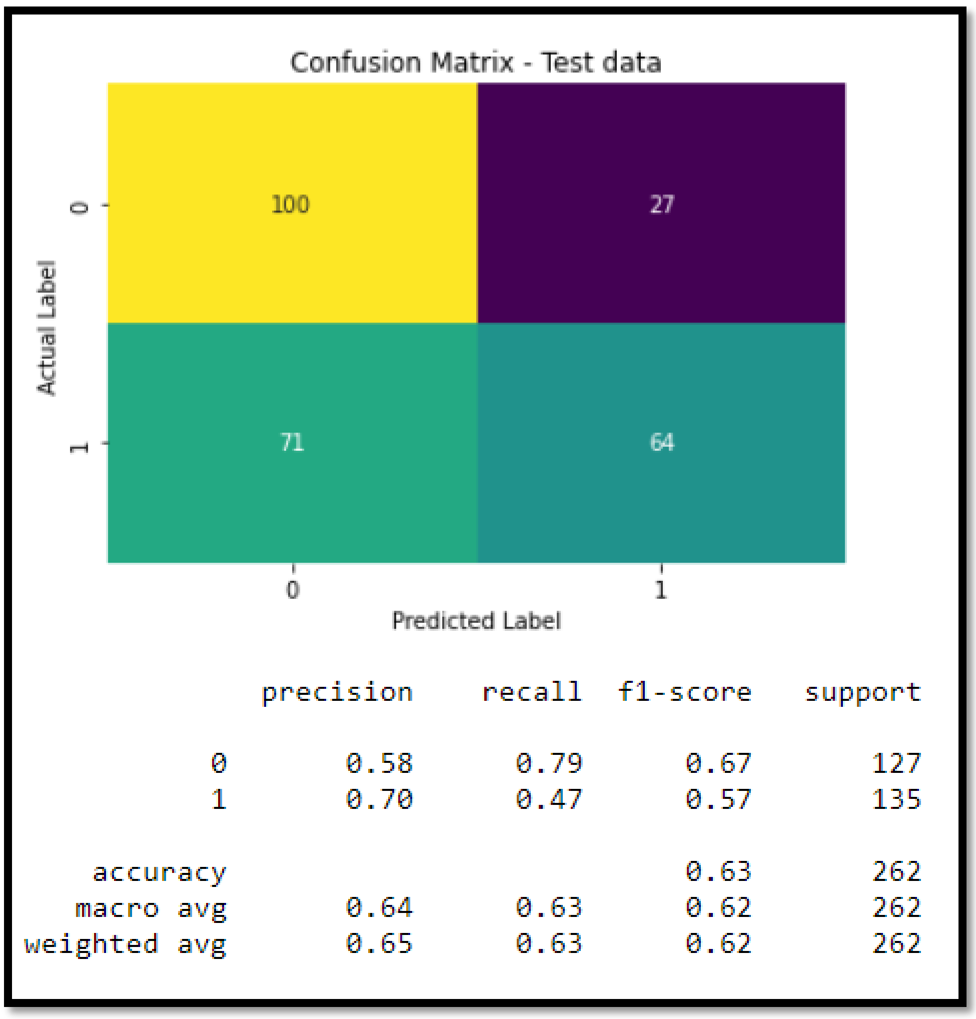
When we use solver = “newton-cg” and remove “Holiday Package” and “foreign” columns- The accuracy for train data is 0.65 whereas for test data it is 0.56. AUC for train data is 0.71 whereas AUC for test data is 0.66. The model has not performed well with these parameters.

Logistic Regression using GridSearchCV



```
              precision    recall  f1-score   support

           0       0.69      0.81      0.74       344
           1       0.68      0.53      0.59       266

    accuracy                           0.68       610
   macro avg       0.68      0.67      0.67       610
weighted avg       0.68      0.68      0.68       610
```

Figure 36. Confusion Matrix and Classification Report - Train Data



```
              precision    recall  f1-score   support

           0       0.58      0.79      0.67       127
           1       0.70      0.47      0.57       135

    accuracy                           0.63       262
   macro avg       0.64      0.63      0.62       262
weighted avg       0.65      0.63      0.62       262
```

Figure 37. Confusion Matrix and Classification Report - Test Data

When we use GridSearchCV - The accuracy for train data is 0.68 whereas for test data it is 0.63. The model has performed slightly well with these parameters compared to the previous model.

Let us now see how the models perform after outlier treatment. We will be capping the data using Inter Quartile Range (IQR) method.

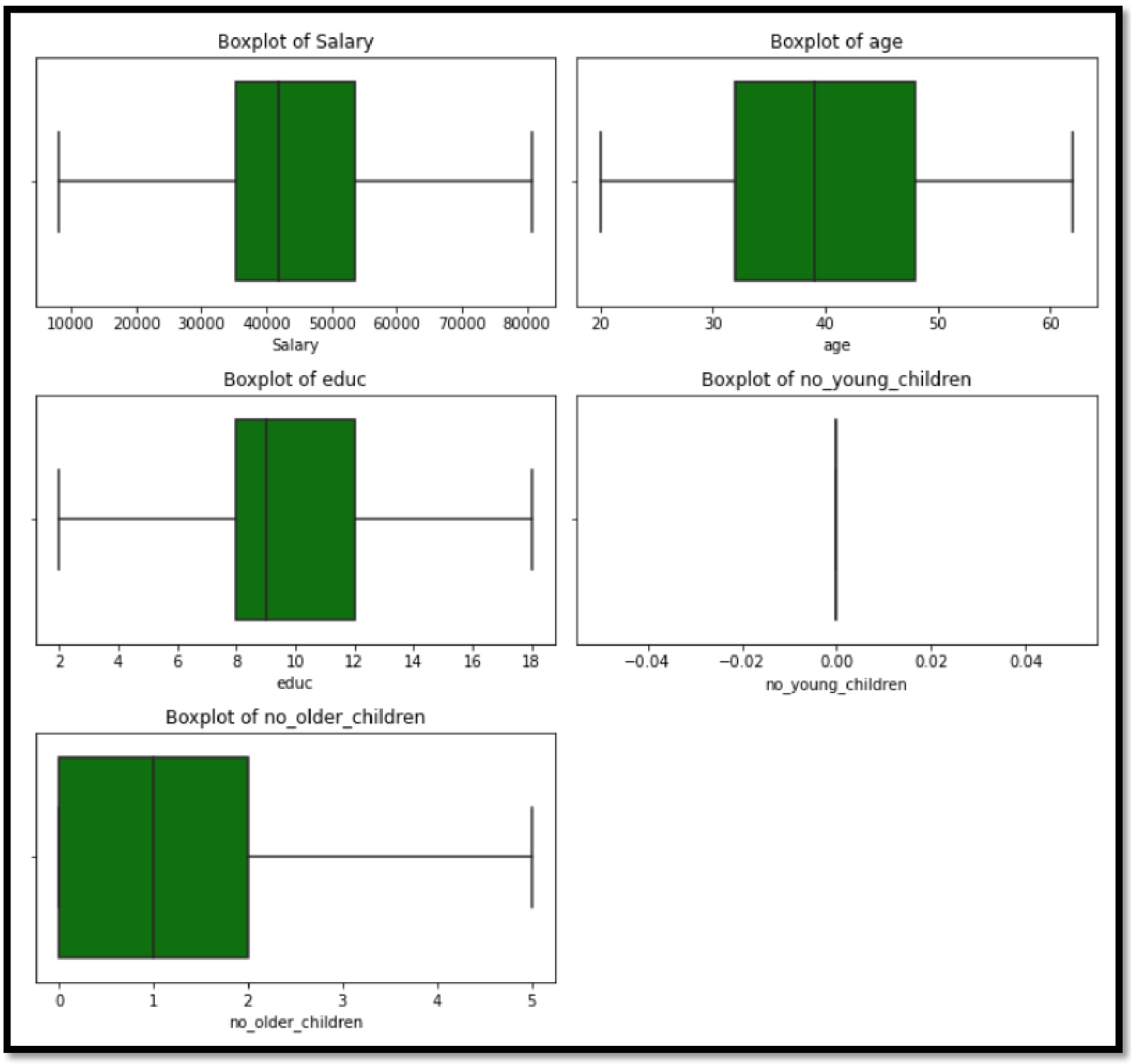
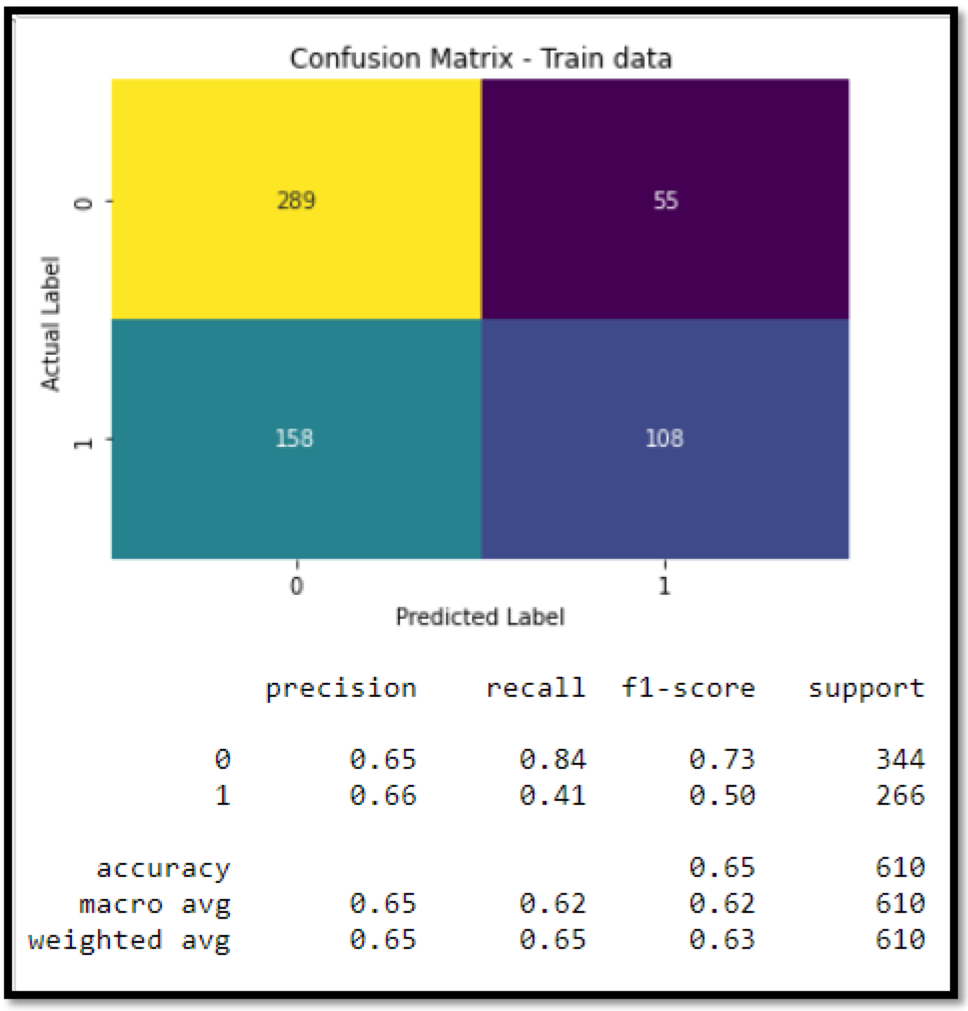
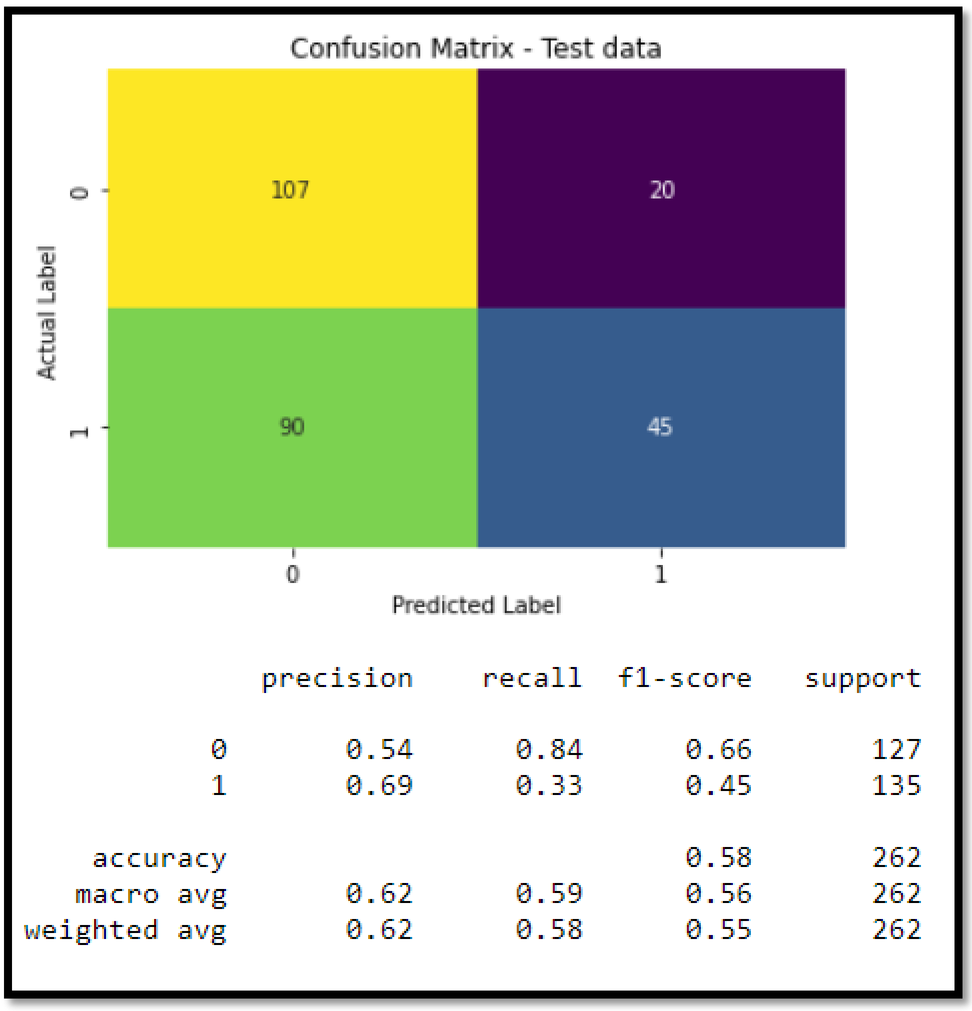


Figure 38. Boxplots post Outlier Treatment



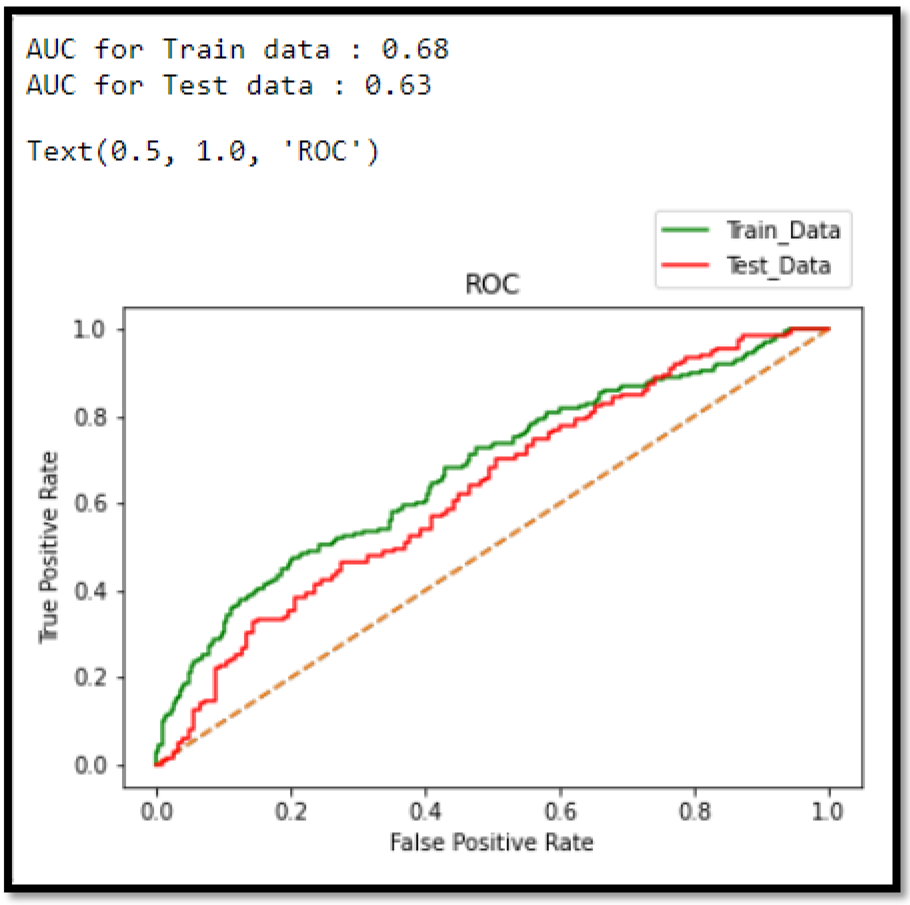
```
              precision    recall  f1-score   support

           0       0.65      0.84      0.73       344
           1       0.66      0.41      0.50       266

    accuracy                           0.65       610
   macro avg       0.65      0.62      0.62       610
weighted avg       0.65      0.65      0.63       610
```

Figure 39. Confusion Matrix and Classification Report - Train Data



```
              precision    recall  f1-score   support

           0       0.54      0.84      0.66       127
           1       0.69      0.33      0.45       135

    accuracy                           0.58       262
   macro avg       0.62      0.59      0.56       262
weighted avg       0.62      0.58      0.55       262
```

Figure 40. Confusion Matrix and Classification Report - Test Data

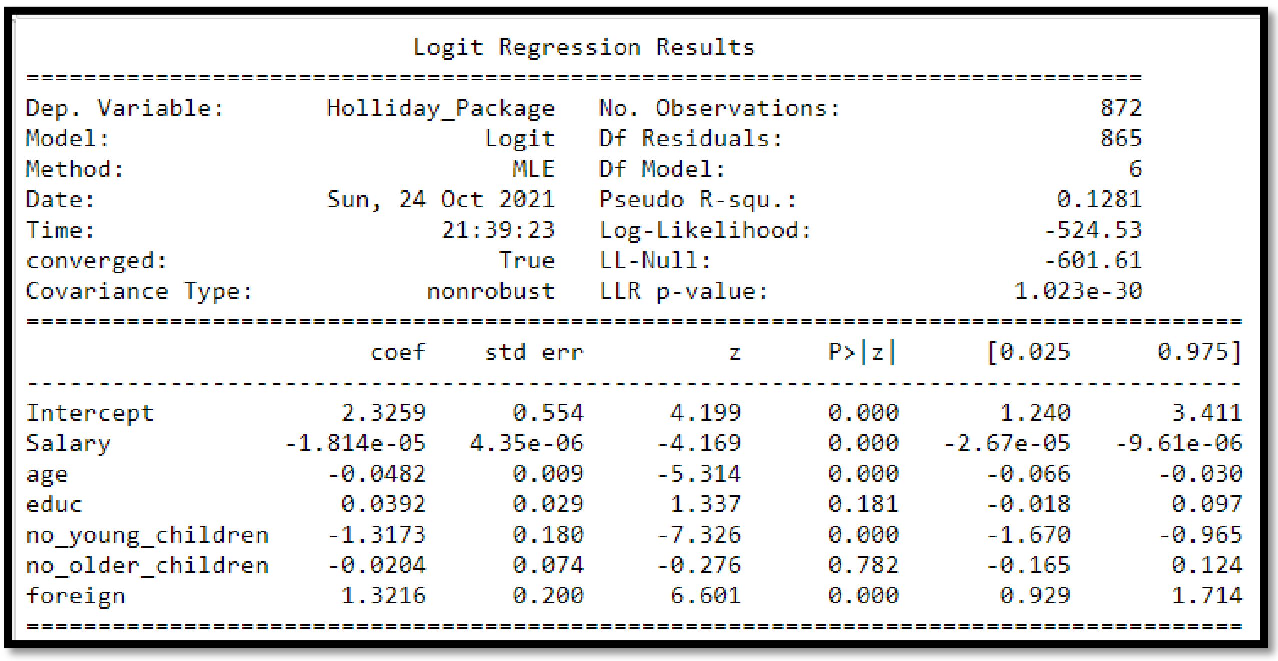


```
AUC for Train data : 0.68
AUC for Test data : 0.63

Text(0.5, 1.0, 'ROC')
```

Figure 41. AUC and ROC

The model has not performed well even after outlier treatment. Let us see how the model performs when we use stats model.

Logistic Regression using Stats Model



```
                           Logit Regression Results
==============================================================================
Dep. Variable:       Holliday_Package   No. Observations:                  872
Model:                          Logit   Df Residuals:                      865
Method:                           MLE   Df Model:                            6
Date:                Sun, 24 Oct 2021   Pseudo R-squ.:                  0.1281
Time:                        21:39:23   Log-Likelihood:                -524.53
converged:                       True   LL-Null:                       -601.61
Covariance Type:            nonrobust   LLR p-value:                 1.023e-30
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept             2.3259      0.554      4.199      0.000       1.240       3.411
Salary            -1.814e-05   4.35e-06     -4.169      0.000   -2.67e-05   -9.61e-06
age                  -0.0482      0.009     -5.314      0.000      -0.066      -0.030
educ                  0.0392      0.029      1.337      0.181      -0.018       0.097
no_young_children    -1.3173      0.180     -7.326      0.000      -1.670      -0.965
no_older_children    -0.0204      0.074     -0.276      0.782      -0.165       0.124
foreign               1.3216      0.200      6.601      0.000       0.929       1.714
=====================================================================================
```

Figure 42. Logit Regression Results

We can see that the model has performed well but there are variables which are not statistically significant to predict the dependent variable. Educ has a p-value of 0.181 and no_older_children has a p-value of 0.782 which are greater than the confidence interval of 0.05.

Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) uses linear combinations of independent variables to predict the class in the response variable. The concept of searching for a linear combination of predictor variables that best separates the classes of the target variable.

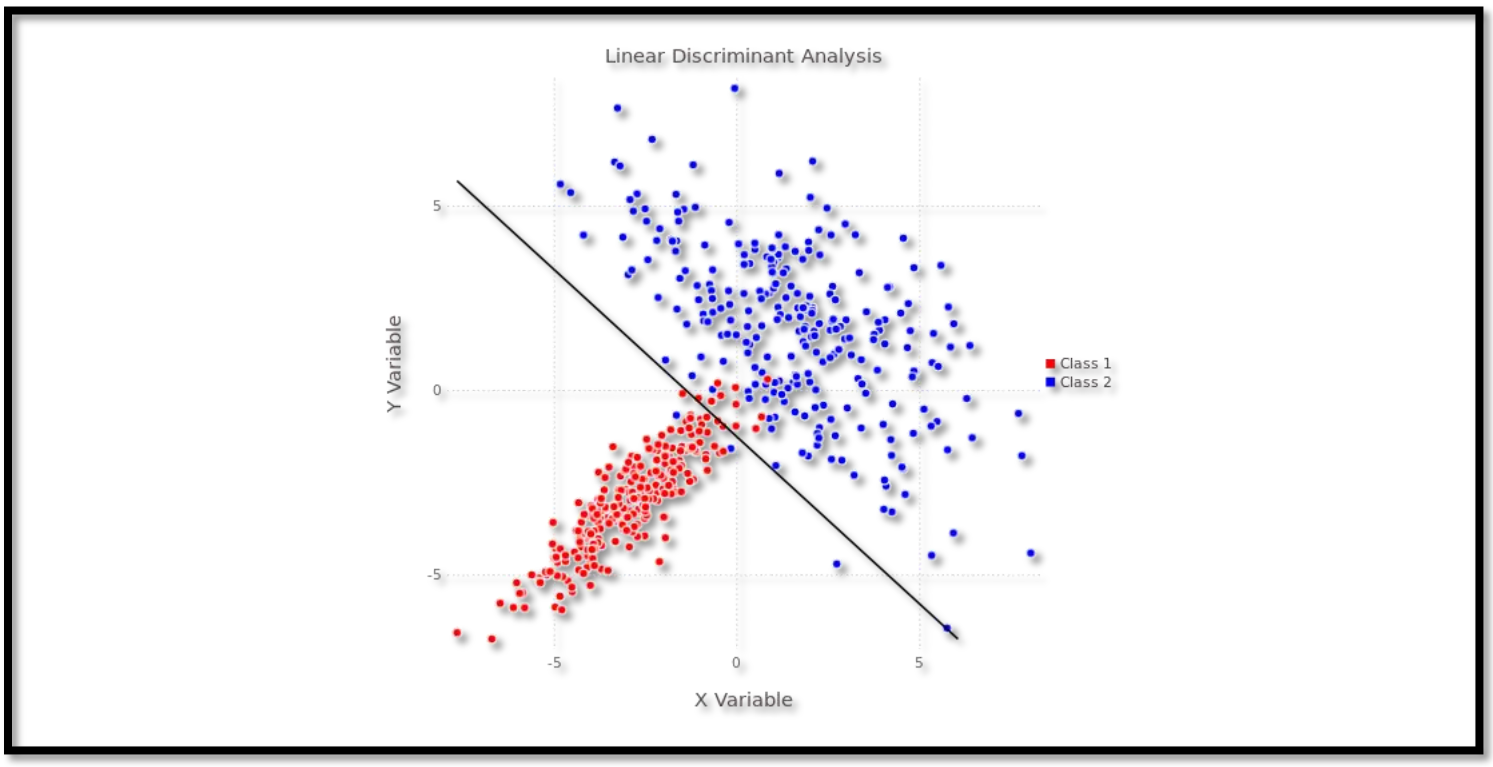
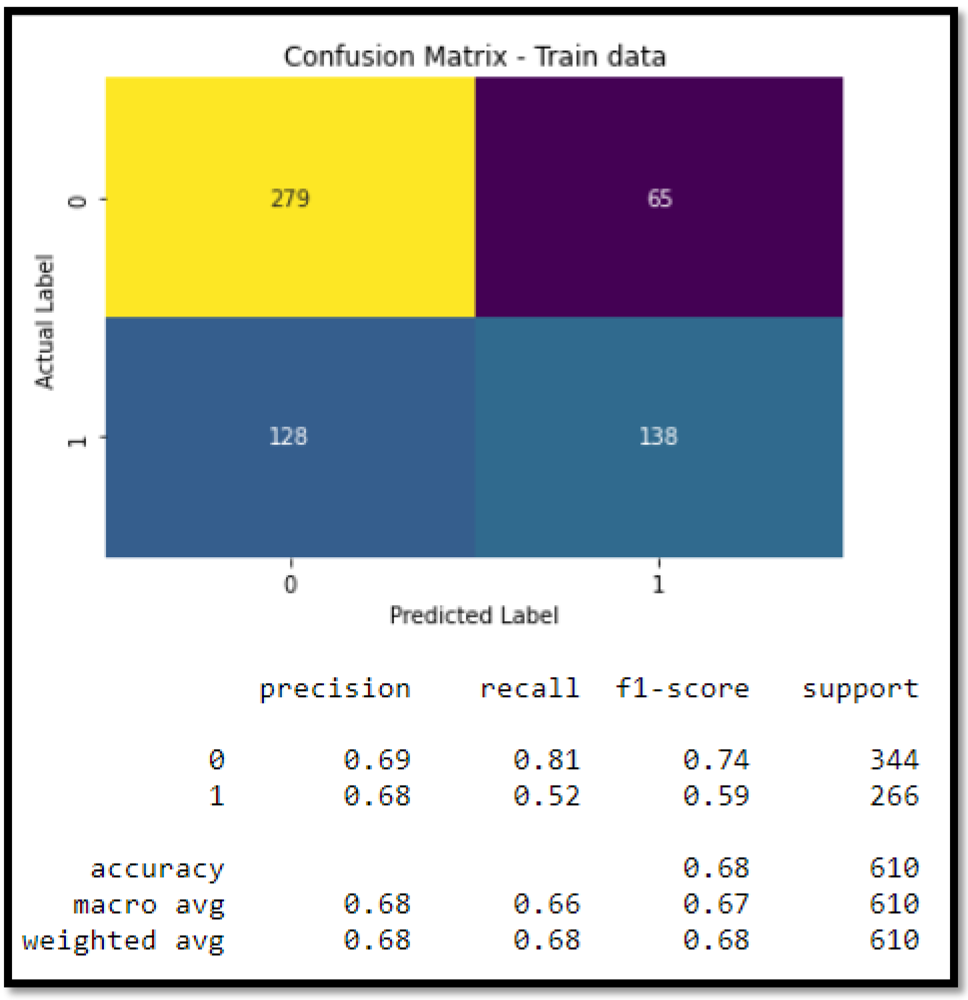


Figure 43. Example of LDA

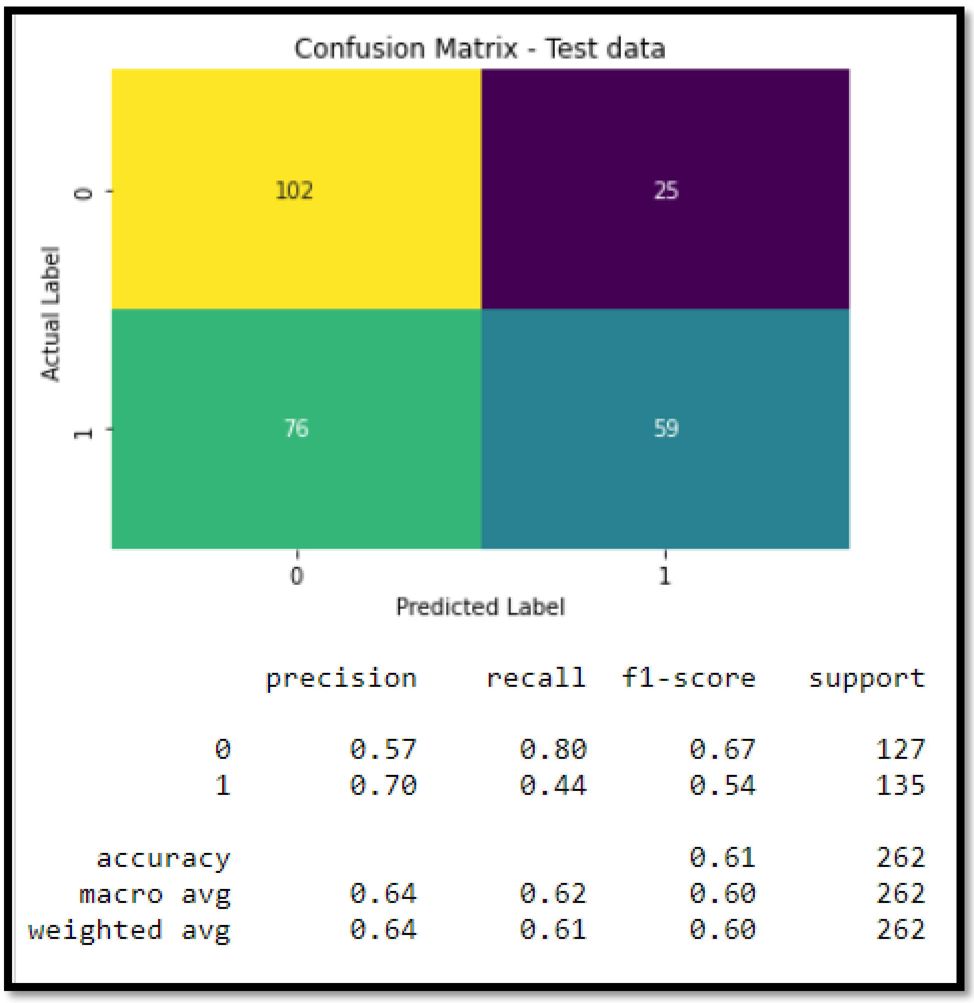
LDA is very similar to PCA (Principal Component Analysis). PCA is an unsupervised technique which is used for dimensionality reduction such that maximum variance is retained. LDA is a supervised learning technique used for classification problems. LDA uses a linear combination of data points such that the two classes are well separated i.e, the mean of class 1 is far away from the mean of class 2 however the variance of class 1 closer to mean of class 1 and similarly variance of class 2 is closer to mean of class 2. This will clearly separate the classes into two groups. LDA is not used for dimensionality reduction.

LDA is sensitive to outliers. LDA does not perform well if there is multicollinearity.



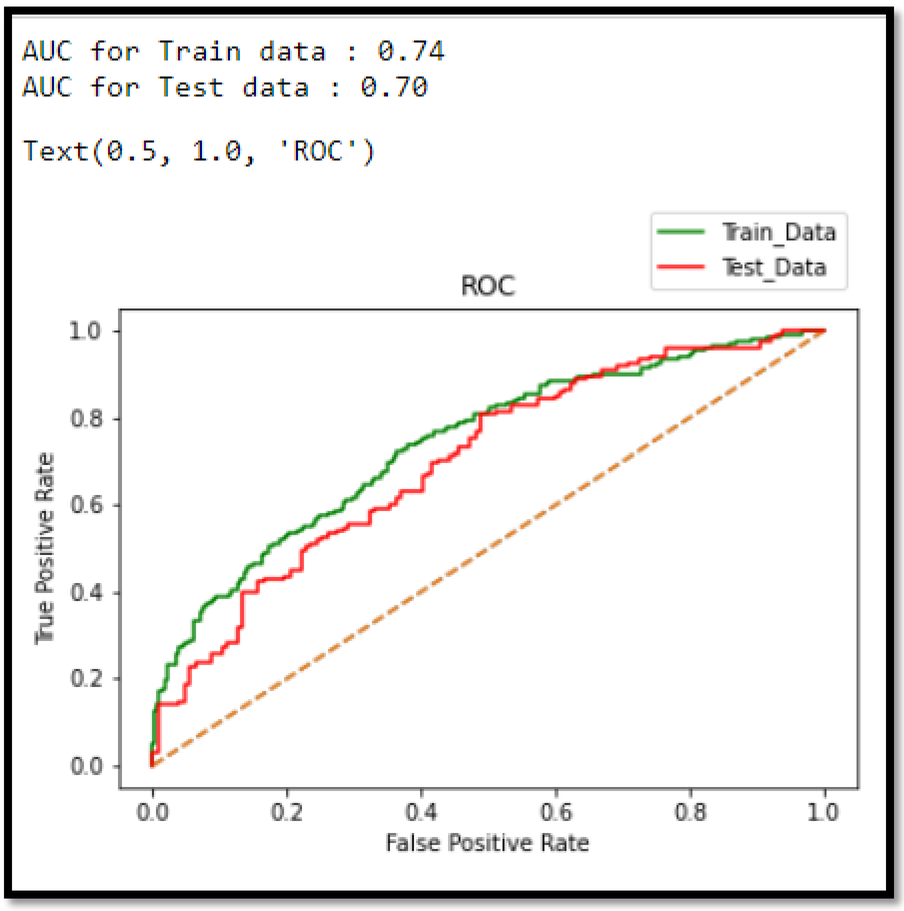
```
              precision    recall  f1-score   support

           0       0.69      0.81      0.74       344
           1       0.68      0.52      0.59       266

    accuracy                           0.68       610
   macro avg       0.68      0.66      0.67       610
weighted avg       0.68      0.68      0.68       610
```

Figure 44. Confusion Matrix and Classification Report - Train data



```
              precision    recall  f1-score   support

           0       0.57      0.80      0.67       127
           1       0.70      0.44      0.54       135

    accuracy                           0.61       262
   macro avg       0.64      0.62      0.60       262
weighted avg       0.64      0.61      0.60       262
```

Figure 45. Confusion Matrix and Classification Report - Test data

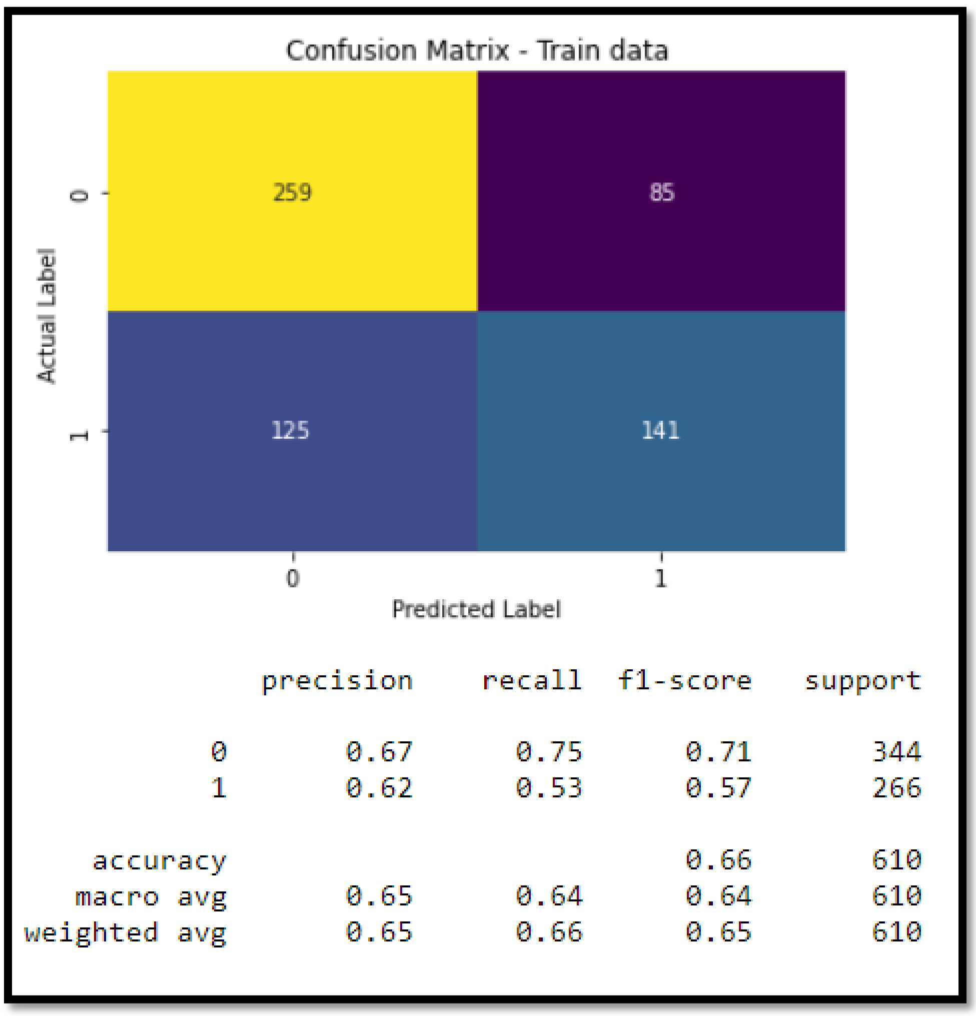


```
AUC for Train data : 0.74
AUC for Test data : 0.70

Text(0.5, 1.0, 'ROC')
```

Figure 46. AUC and ROC

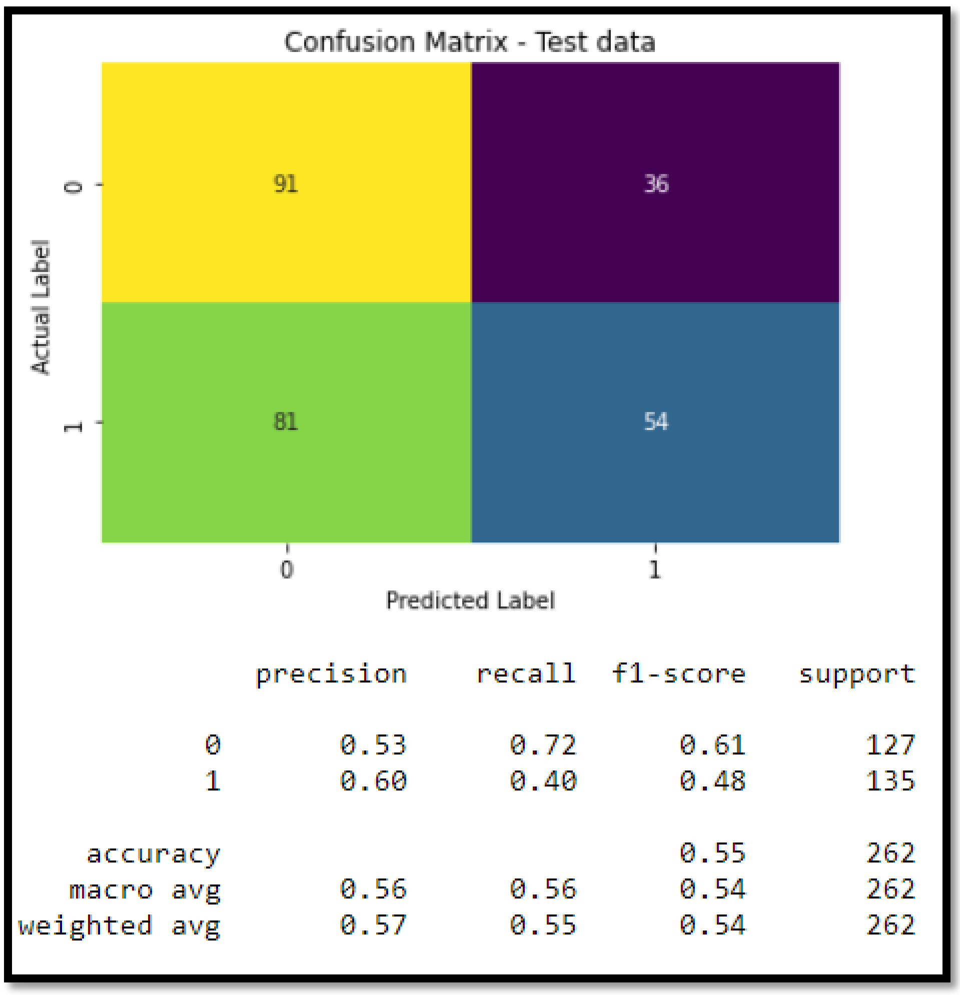
When we perform LDA post outlier treatment - The accuracy for train data is 0.68 whereas for test data it is 0.61. The AUC on train data is 0.74 whereas AUC for test data is 0.70.

LDA – removing foreign column



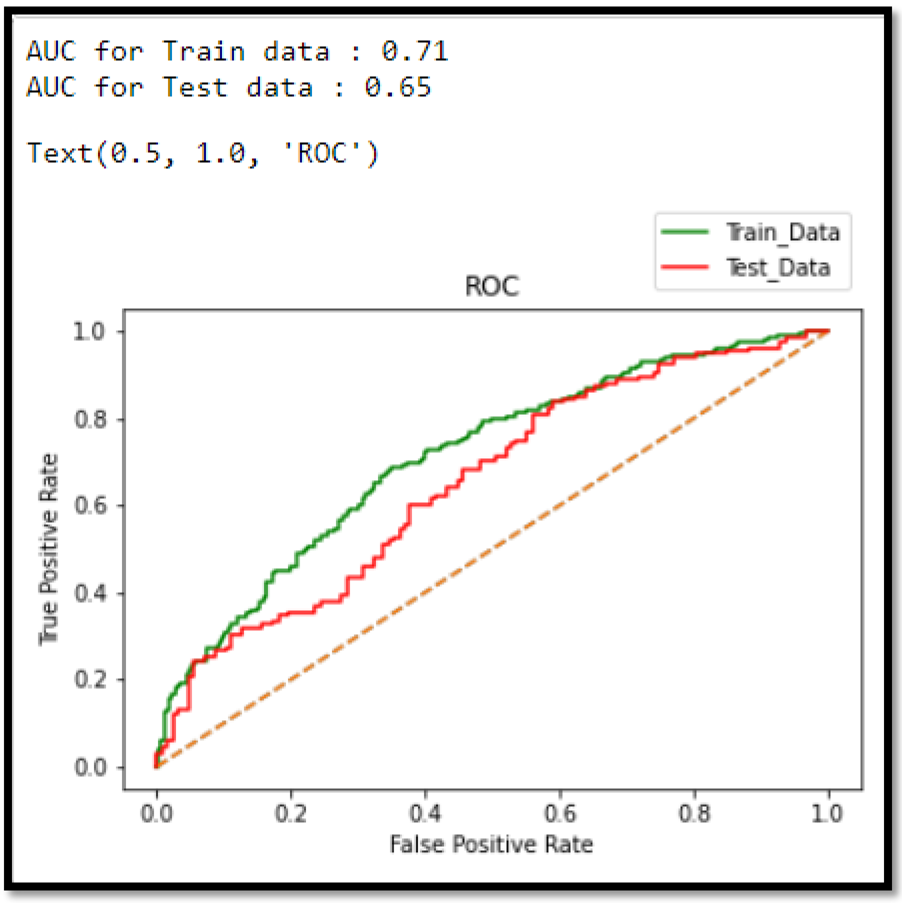
```
              precision    recall  f1-score   support

           0       0.67      0.75      0.71       344
           1       0.62      0.53      0.57       266

    accuracy                           0.66       610
   macro avg       0.65      0.64      0.64       610
weighted avg       0.65      0.66      0.65       610
```

Figure 47. Confusion Matrix and Classification Report - Train data



```
              precision    recall  f1-score   support

           0       0.53      0.72      0.61       127
           1       0.60      0.40      0.48       135

    accuracy                           0.55       262
   macro avg       0.56      0.56      0.54       262
weighted avg       0.57      0.55      0.54       262
```

Figure 48. Confusion Matrix and Classification Report - Test data



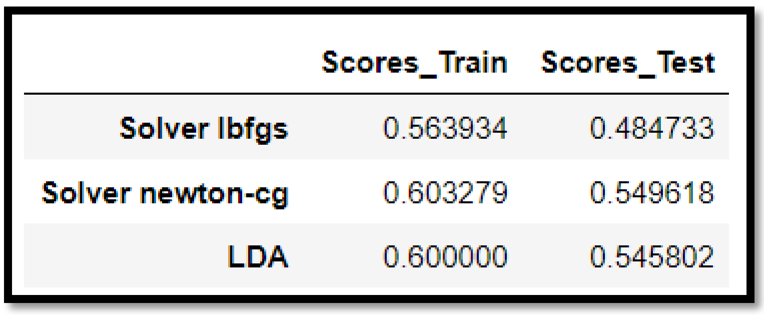
```
AUC for Train data : 0.71
AUC for Test data : 0.65

Text(0.5, 1.0, 'ROC')
```

Figure 49. AUC and ROC

The model has not shown any improvement after removing foreign column.

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.

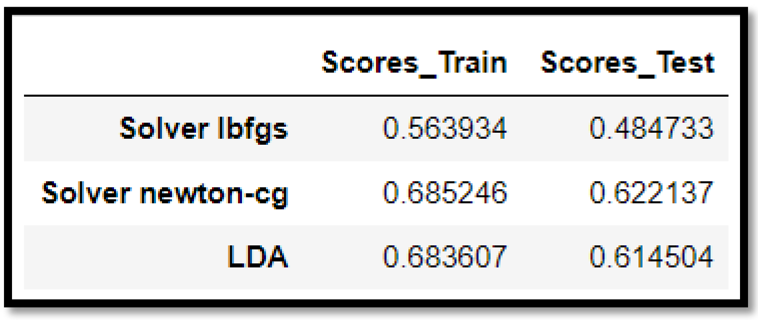
Model Performance before Outlier Treatment



	Scores_Train	Scores_Test
Solver lbfgs	0.563934	0.484733
Solver newton-cg	0.603279	0.549618
LDA	0.600000	0.545802

Table 20. Model Scores – Before Outlier Treatment

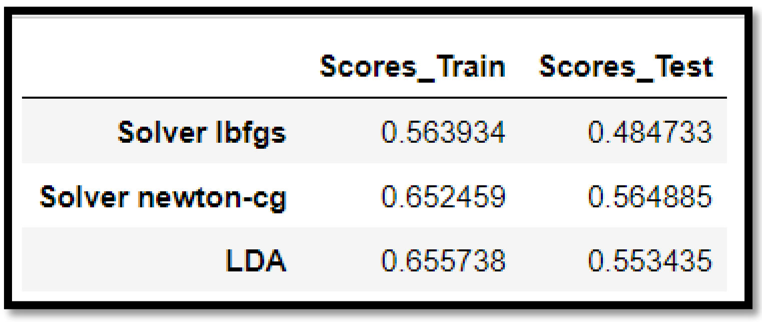
Model Performance post Outlier Treatment



	Scores_Train	Scores_Test
Solver lbfgs	0.563934	0.484733
Solver newton-cg	0.685246	0.622137
LDA	0.683607	0.614504

Table 21. Model Scores – Post Outlier Treatment

Model Performance when "foreign" column removed



	Scores_Train	Scores_Test
Solver lbfgs	0.563934	0.484733
Solver newton-cg	0.652459	0.564885
LDA	0.655738	0.553435

Table 22. Model Scores – foreign column removed

Comparison of models and Inferences

Below are the observations and inferences based on model performance scores:

1. Model has performed the best post outlier treatment.
2. **62.21%** accuracy scores are obtained for model which uses **solver = "newton-cg"** post outlier treatment. This is the best possible accuracy so far compared to other models.
3. **61.45%** accuracy scores are obtained for **LDA** model post outlier treatment which is the second best model.
4. The model predicts whether an employee will opt for a holiday package or not **62.21%** accurately most of the times.
5. However the data shows that there is a **class balance** between yes and no classes.
6. Model still needs to improve as **38.8%** the model might predict incorrectly.
7. More data needs to be collected in order to build better models.
8. Looks like the Tours and Travel Company is providing holiday packages with **exuberant prices** without any discounts or offers which is why we see that most of the employees are not opting for the holiday package. They need to come up with different offers and schemes which are **budget friendly** so that more employees might opt for a holiday package.