PGP - Data Science and Business Analytics

PREDICTIVE MODELLING PROJECT REPORT

Akshaya Nallathambi

th August, 2021

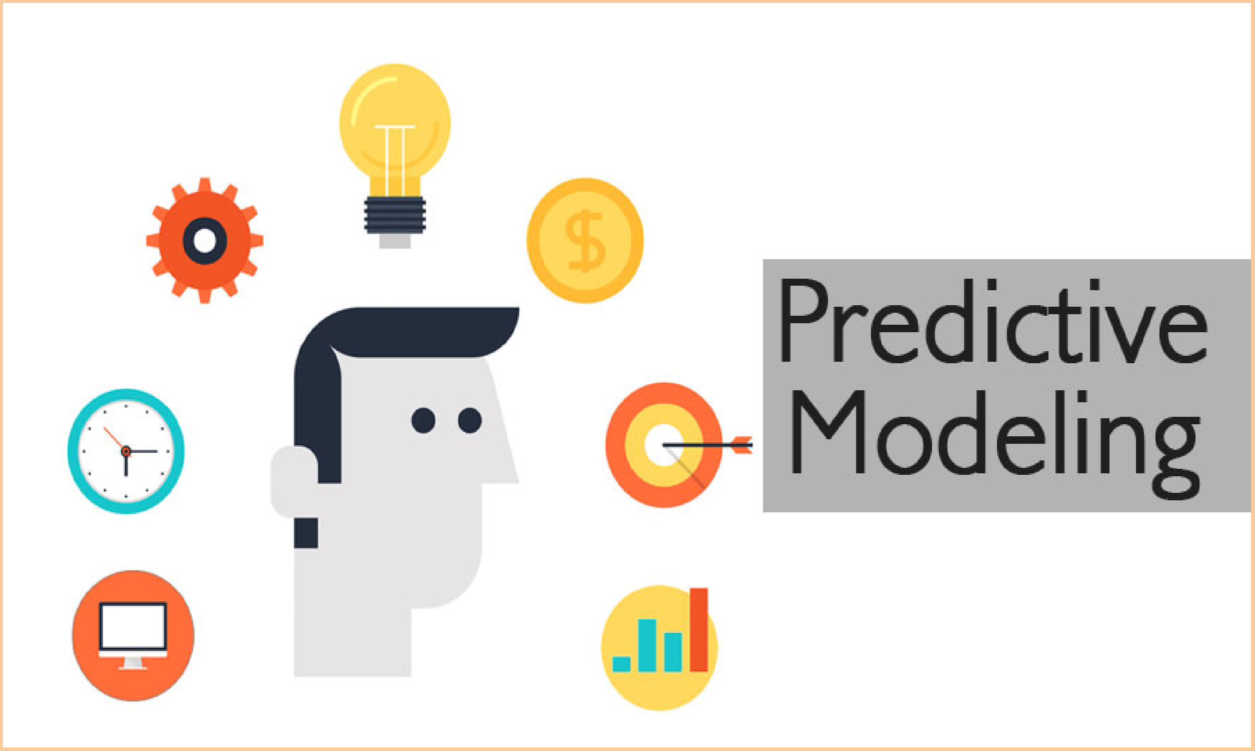


Table Of Contents

Problem 1

Problem 2

List of Figures

List of Tables

List of Graphs

Problem 1

Problem statement-

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 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: Cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst, IF = flawless, I1= level 1 inclusion) IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, I1

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.

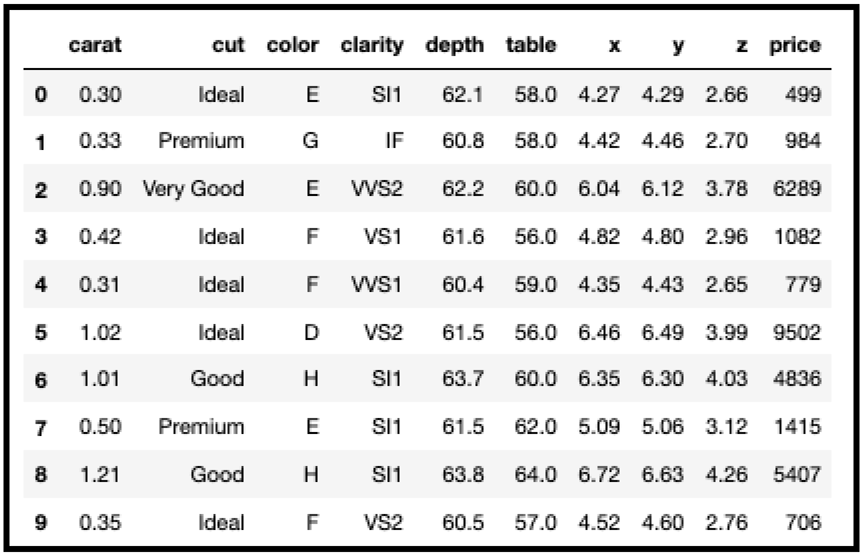
X: Length of the cubic zirconia in mm.

Y: Width of the cubic zirconia in mm.

Z: Height of the cubic zirconia in mm.

Sample of the dataset-

FIGURE 1



	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
5	1.02	Ideal	D	VS2	61.5	56.0	6.46	6.49	3.99	9502
6	1.01	Good	H	SI1	63.7	60.0	6.35	6.30	4.03	4836
7	0.50	Premium	E	SI1	61.5	62.0	5.09	5.06	3.12	1415
8	1.21	Good	H	SI1	63.8	64.0	6.72	6.63	4.26	5407
9	0.35	Ideal	F	VS2	60.5	57.0	4.52	4.60	2.76	706

There are 10 variables out of which 6 are float values, 1 is int value and 3 are object values . The dataset contains 697 null values in depth column.

Types of variables in the data frame-

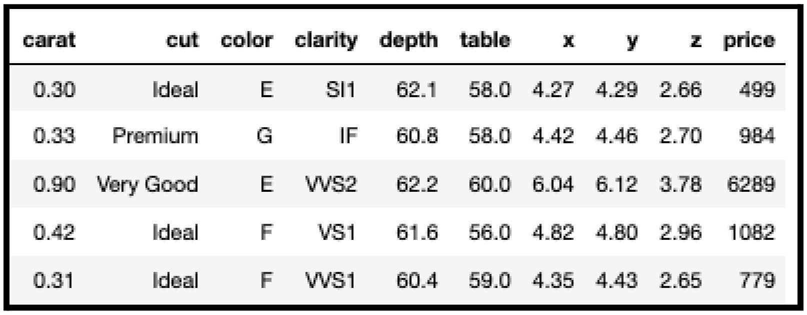
TABLE 1

carat	float64	Continuous
cut	object	Categorical
color	object	Categorical
clarity	object	Categorical
depth	float64	Continuous
table	float64	Continuous
X	float64	Continuous
Y	float64	Continuous
Z	float64	Continuous
price	int64	Continuous

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

The below table gives the first 5 rows of sample data.

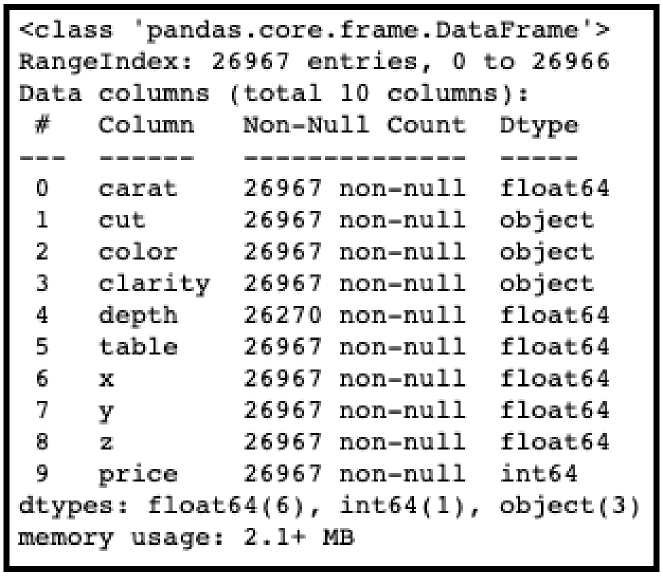
FIGURE 2



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

The image below gives the basic information of the data set. It is clear that the variables have int, float and object data types with 10 columns and 26967 rows. The dataset contains 697 null values in depth column which have been treated by imputing them with the median of depth column. The memory usage is 2.1+ MB.
The dataset also does not contain any duplicates.

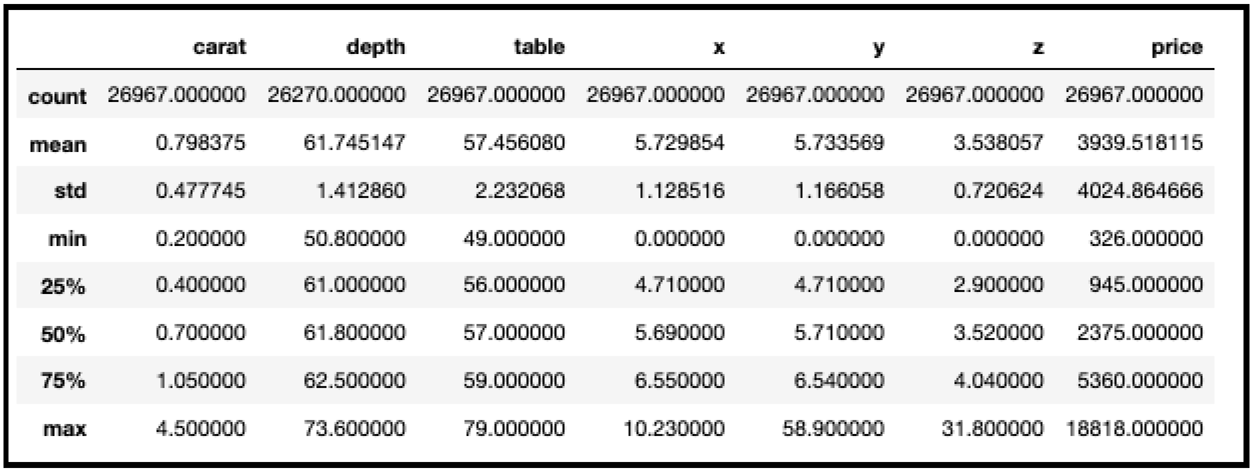
FIGURE 3



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26967 entries, 0 to 26966
Data columns (total 10 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   carat     26967 non-null  float64
 1   cut       26967 non-null  object
 2   color     26967 non-null  object
 3   clarity   26967 non-null  object
 4   depth     26270 non-null  float64
 5   table     26967 non-null  float64
 6   x         26967 non-null  float64
 7   y         26967 non-null  float64
 8   z         26967 non-null  float64
 9   price     26967 non-null  int64
dtypes: float64(6), int64(1), object(3)
memory usage: 2.1+ MB
```

The below image gives the five point summary of the continuous variables in the data set. It is clear that the data needs scaling as the numbers are of different magnitude.

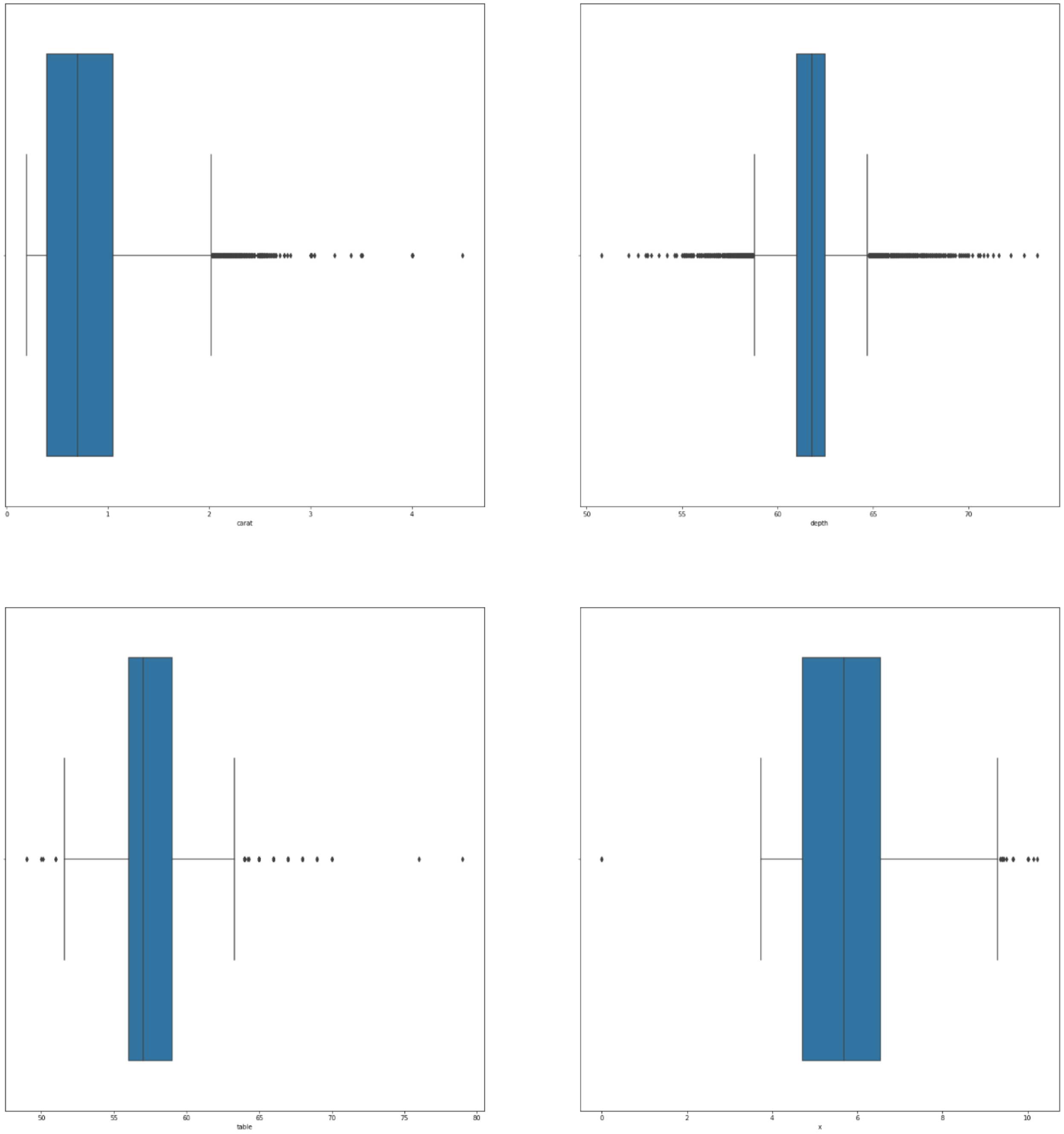
FIGURE 4

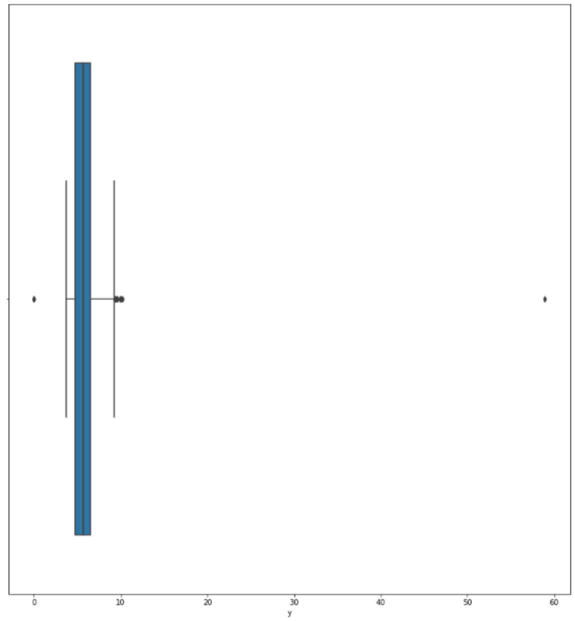


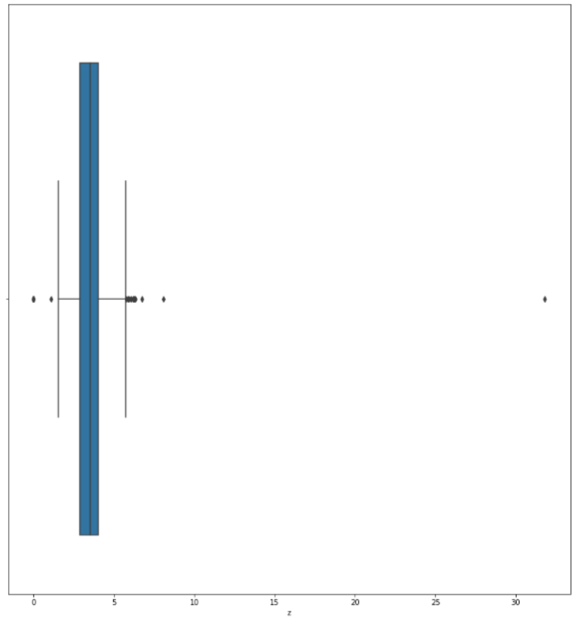
	carat	depth	table	x	y	z	price
count	26967.000000	26270.000000	26967.000000	26967.000000	26967.000000	26967.000000	26967.000000
mean	0.798375	61.745147	57.456080	5.729854	5.733569	3.538057	3939.518115
std	0.477745	1.412860	2.232068	1.128516	1.166058	0.720624	4024.864666
min	0.200000	50.800000	49.000000	0.000000	0.000000	0.000000	326.000000
25%	0.400000	61.000000	56.000000	4.710000	4.710000	2.900000	945.000000
50%	0.700000	61.800000	57.000000	5.690000	5.710000	3.520000	2375.000000
75%	1.050000	62.500000	59.000000	6.550000	6.540000	4.040000	5360.000000
max	4.500000	73.600000	79.000000	10.230000	58.900000	31.800000	18818.000000

There are outliers in all the continuous variables. This is evident from the box plots below,

GRAPH 1







There are both categorical variables and continuous variables in the given data set. So univariate and bivariate analysis can be done in both.

Univariate analysis:

Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships (unlike regression) and it’s major purpose is to describe; It takes data, summarizes that data and finds patterns in the data. [1]

The histograms are used for numerical variables and count plots are used for categorical variables to perform univariate analysis.

It is clear from the graph (GRAPH 2) that all the numerical variables are skewed.

Skewness:

Skewness is a number that indicates to what extent a variable is asymmetrically distributed. The below table represents the level of skewness and the respective skewness value. [2]

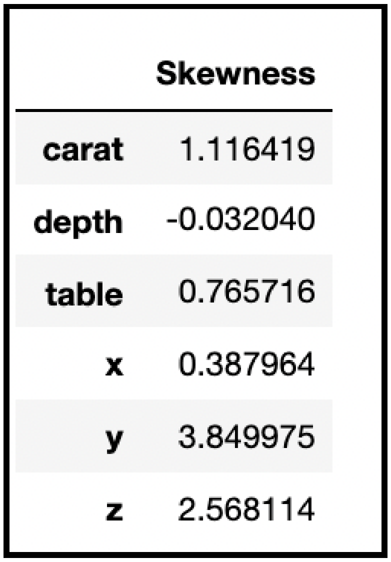
TABLE 2

Skewness level	Value
Symmetrical or Not Skewed	0
Less Skewed Data	± 0.5 to 1
Highly Skewed Data	Greater than ±1

When the skewness value is positive it is considered as right skewed data and when the skewness value is negative it is considered as left skewed data.

The table below shows the skewness value corresponding to each variable in the given data set.

FIGURE 5

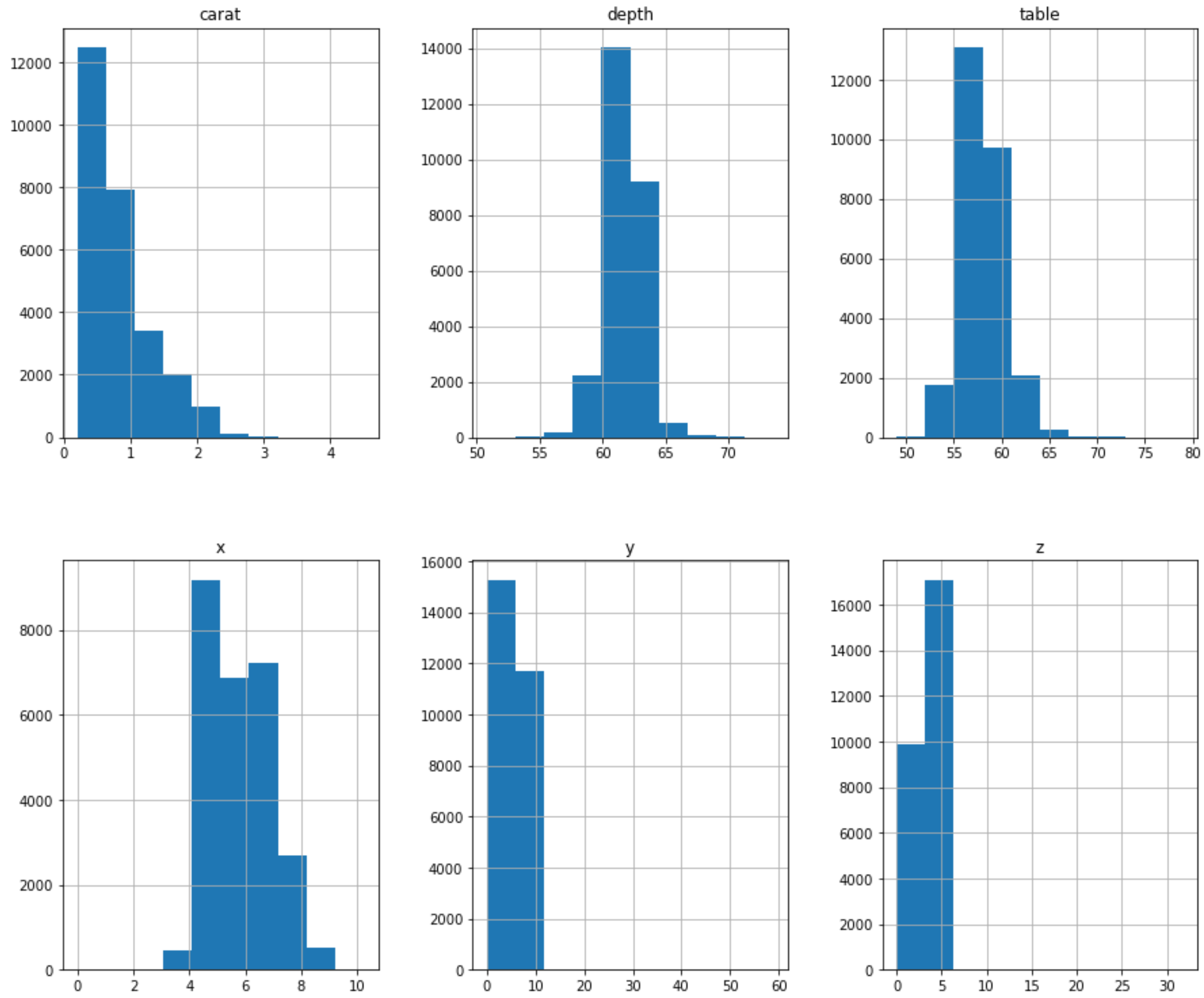


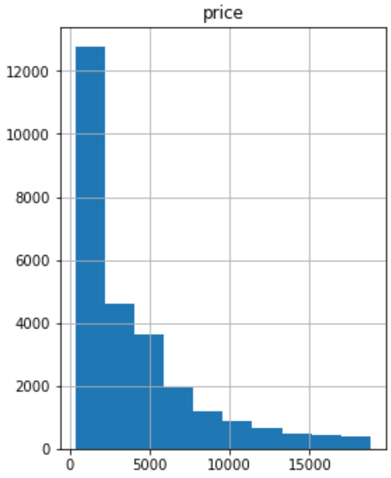
	Skewness
carat	1.116419
depth	-0.032040
table	0.765716
x	0.387964
y	3.849975
z	2.568114

All the variables have a positive skew value except for “depth”. Therefore, all the other variables above are right skewed variables and “depth” is left skewed variable.

Univariate Analysis for Continuous variables-

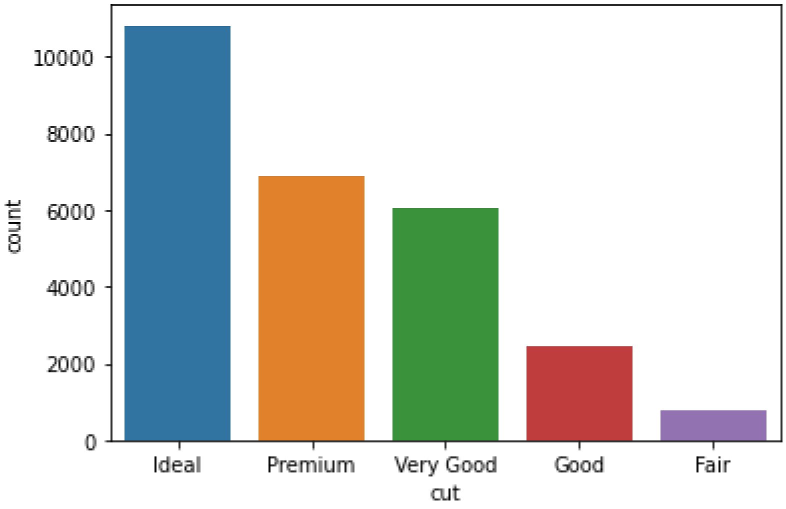
GRAPH 2



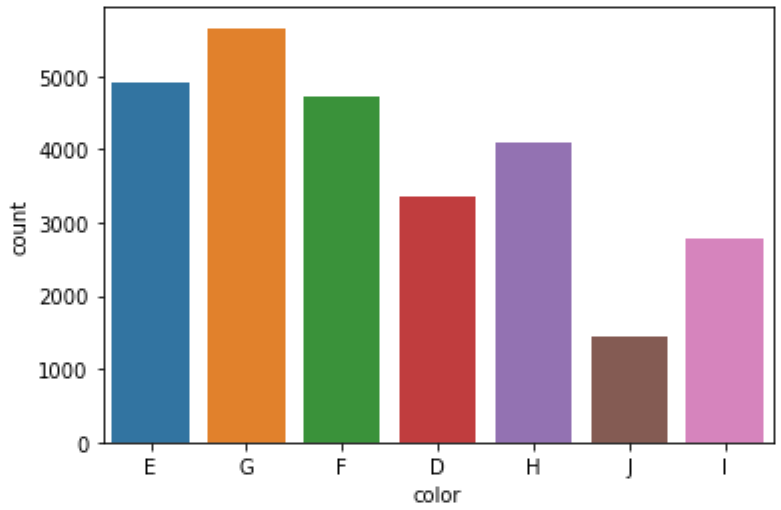


Univariate Analysis for Categorical variables-

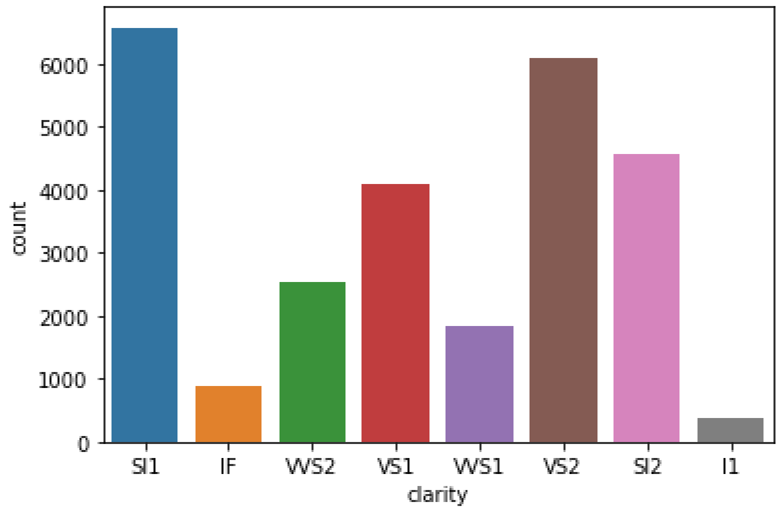
GRAPH 3



GRAPH 4



GRAPH 5



Bivariate analysis:

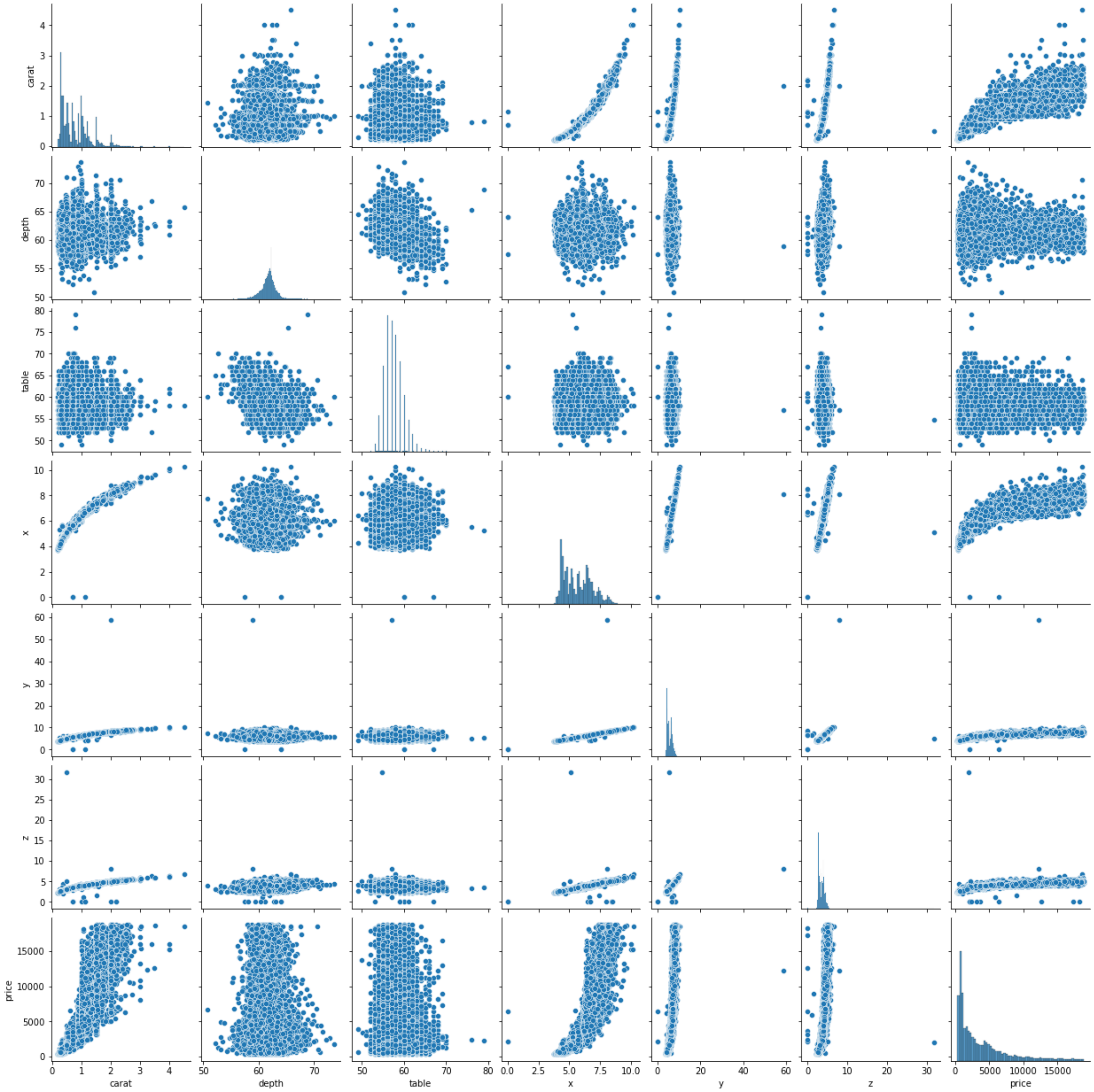
Bivariate analysis is one of the simplest forms of quantitative (statistical) analysis. It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical relationship between them. Bivariate analysis can be helpful in testing simple hypotheses of association. [3]

The pairplot is generally used for numerical variables and box plots are used for categorical with numerical variables to perform bivariate analysis.

A pairplot plots a pairwise relationship in a dataset. The pairplot function creates a grid of Axes such that each variable in data will be shared in the y-axis across a single row and in the x-axis across a single column. [4]

A boxplot is a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”). It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed. [5]

GRAPH 6



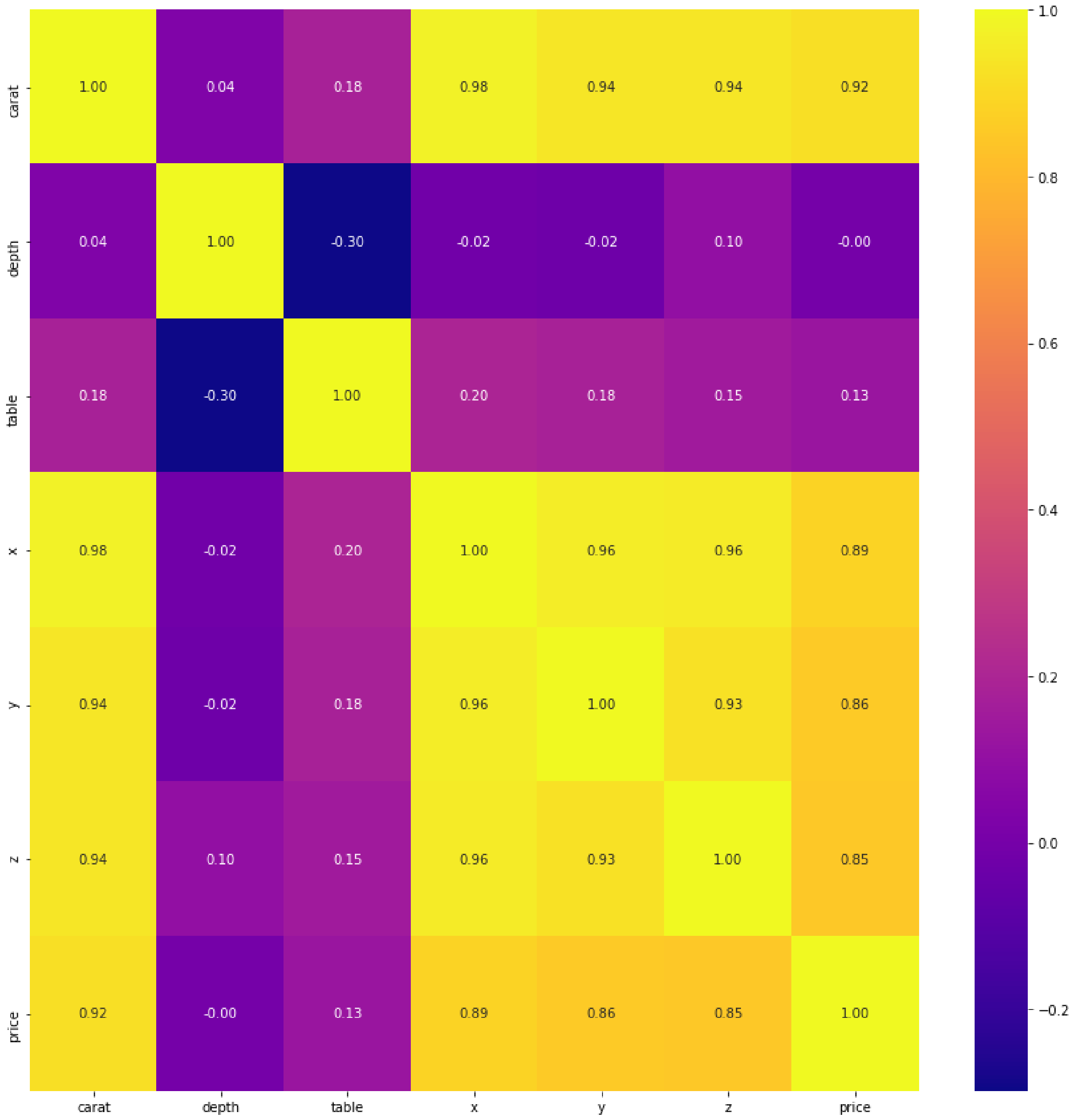
The following are the findings from the pairplot generated -

- The variable carat is highly correlated with the variables x, y, z and price.
- The variable x is highly correlated with the variables y, z and price.

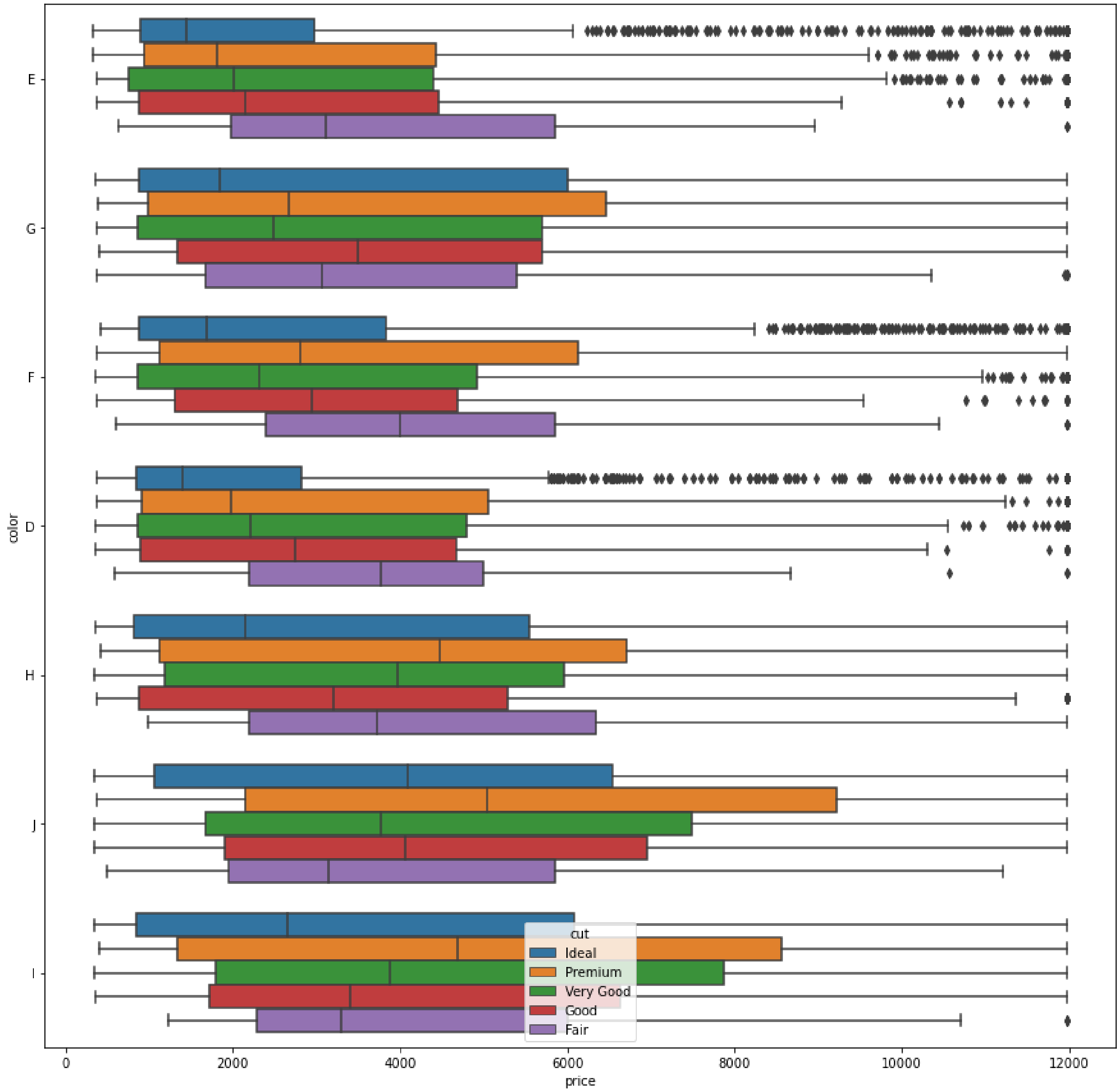
- The variable y is highly correlated with the variables z and price.
- The variables z and price are highly correlated to each other.

The heat map can also be used to check the association between two variables. All the boxes with a value higher than 0.8 are highly correlated. The heat map for all the numerical variable is below,

GRAPH 7



GRAPH 8



The below are the findings from the boxplot generated -

- The premium cut zirconium has the highest price in all colors.
- The color D, F and E has had the extreme price that has the ideal cut.
- The color E with the fair cut has the maximum value in the box plot without considering the outliers.

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

The dataframe contains 26967 records of data evenly spread. The dataset contains 697 null values in the “depth” column which has been treated by imputing them with the median of the “depth” column.

The columns “x”,”y” and “z” have zero values in it. Also, they are highly correlated with each other and with the column “carat”. Hence these columns can be dropped.

Scaling:

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. [6]

Scaling converts variables with different scales of measurements into a single scale. This is done only for the numerical variables.

The data is scaled using the formula $\frac{X-\mu}{\sigma}$.

μ: Mean

σ: Standard deviation

The process of scaling is necessary in the given data set as the variables of the data set are of different scales i.e. one variable has one digit number and other has five digit numbers. For e.g. in our data set “price” has values in five digits and “carat” has only one digit number. Since the data in these variables are of different scales, it is tough to compare these variables. Therefore scaling is done in the given data set.

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

We encode the object data types into categorical variables by encoding them as numerical variables to perform the modeling. There are three object type variables in the given data set that need encoding.

The below table gives the encoded value for each unique data in the respective variable.

VARIABLE: cut

TABLE 3

Fair	0
Good	1
Very Good	2
Premium	3
Ideal	4

VARIABLE: color

TABLE 4

J	0
I	1
H	2
G	3
F	4

E	5
D	6

VARIABLE: clarity

TABLE 5

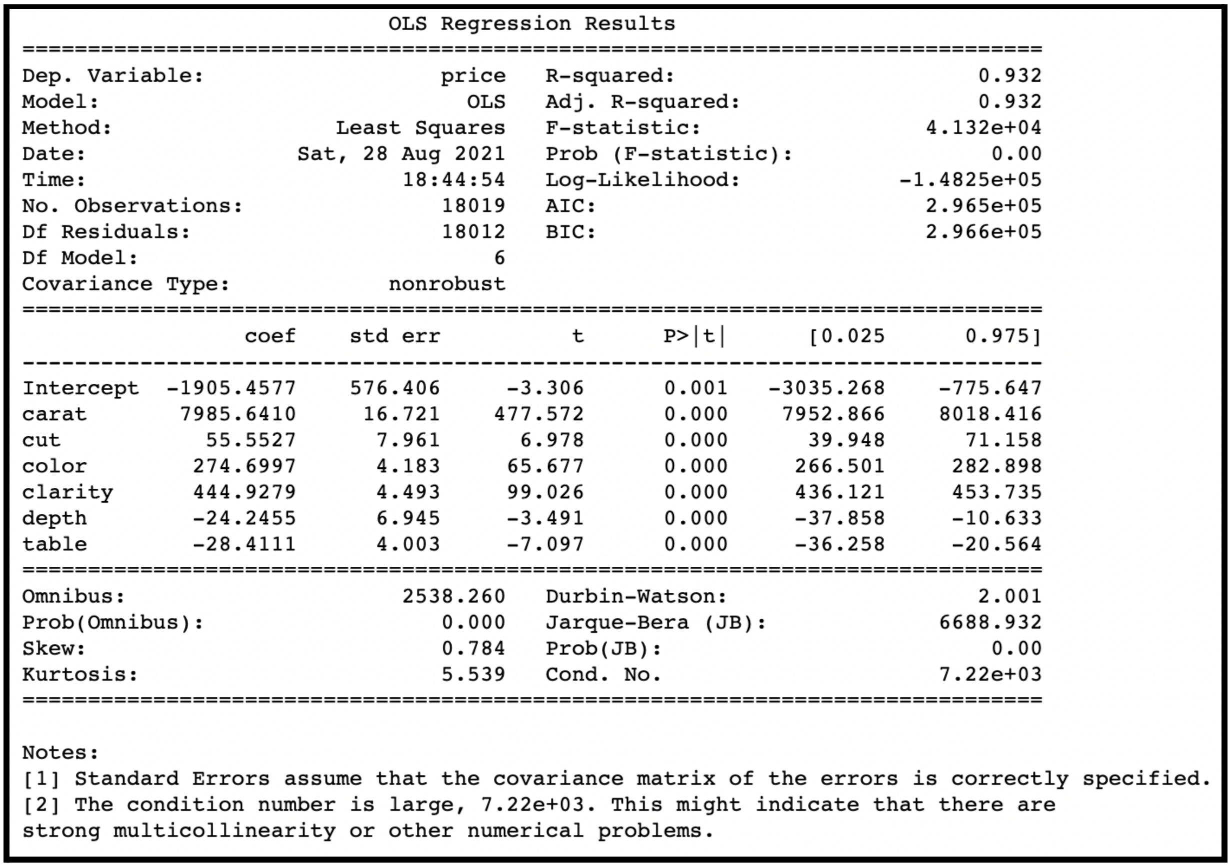
I1	0
SI2	1
SI1	2
VS2	3
VS1	4
VVS2	5
VVS1	6
IF	7

In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable. [7]

From the given data, 30 percent is taken for the test size and 70 percent is taken for the training. The target variable here is “price”. The data is fitted into the Linear Regression Model.

Linear model summary –

FIGURE 6

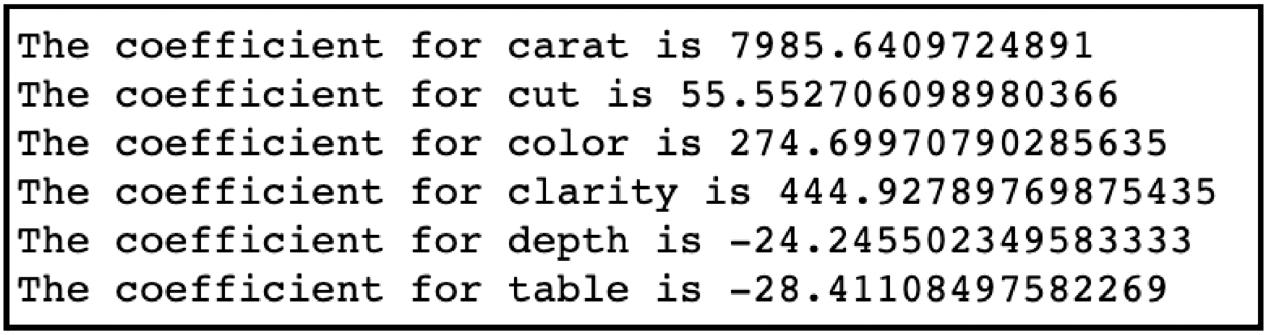


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.932
Model:                            OLS   Adj. R-squared:                  0.932
Method:                 Least Squares   F-statistic:                 4.132e+04
Date:                Sat, 28 Aug 2021   Prob (F-statistic):               0.00
Time:                        18:44:54   Log-Likelihood:            -1.4825e+05
No. Observations:               18019   AIC:                         2.965e+05
Df Residuals:                   18012   BIC:                         2.966e+05
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -1905.4577    576.406     -3.306      0.001   -3035.268    -775.647
carat       7985.6410     16.721    477.572      0.000    7952.866    8018.416
cut           55.5527      7.961      6.978      0.000      39.948      71.158
color        274.6997      4.183     65.677      0.000     266.501     282.898
clarity      444.9279      4.493     99.026      0.000     436.121     453.735
depth        -24.2455      6.945     -3.491      0.000     -37.858     -10.633
table        -28.4111      4.003     -7.097      0.000     -36.258     -20.564
==============================================================================
Omnibus:                     2538.260   Durbin-Watson:                   2.001
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6688.932
Skew:                           0.784   Prob(JB):                         0.00
Kurtosis:                       5.539   Cond. No.                     7.22e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.22e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

BEFORE SCALING:

The coefficient for each independent attribute is given below-

FIGURE 7



```
The coefficient for carat is 7985.6409724891
The coefficient for cut is 55.552706098980366
The coefficient for color is 274.69970790285635
The coefficient for clarity is 444.92789769875435
The coefficient for depth is -24.245502349583333
The coefficient for table is -28.41108497582269
```

The intercept for our model is -1905.45765795932.

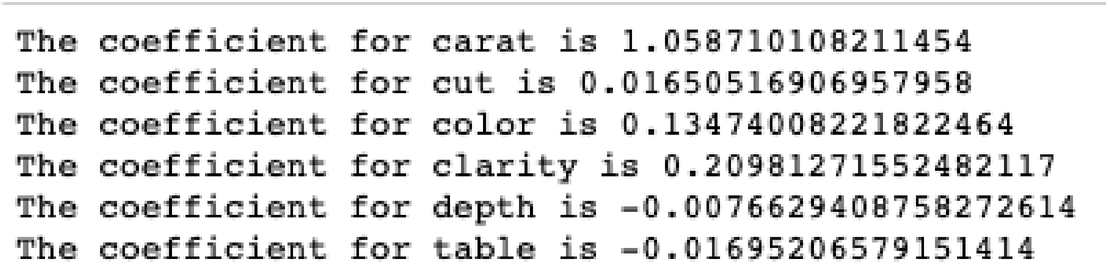
The R square for training data is 0.9322660677114132 and for testing data is 0.931957072360339.

The RMSE score for training data is 905.4332671439 and for testing data is 906.4573329601515.

AFTER SCALING:

The coefficient for each independent attribute is given below-

FIGURE 8



```
The coefficient for carat is 1.058710108211454
The coefficient for cut is 0.016505169069579 58
The coefficient for color is 0.13474008221822464
The coefficient for clarity is 0.20981271552482117
The coefficient for depth is -0.007662940875827261 4
The coefficient for table is -0.016952065791514 14
```

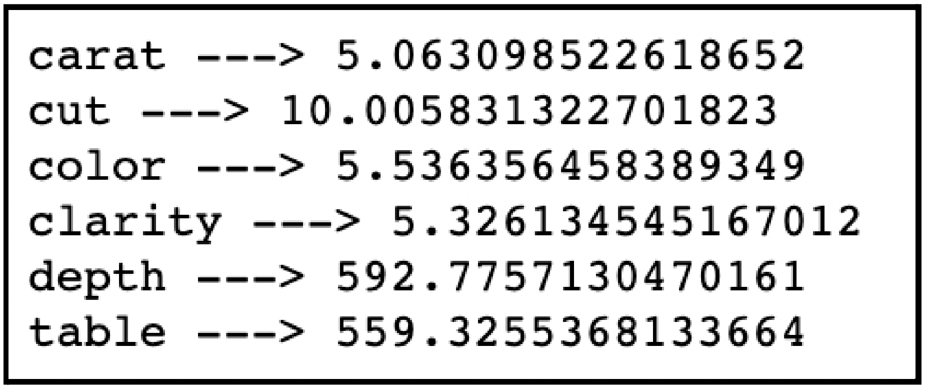
The intercept for our model is -1.1052488343132788e-16.

The R square for training data is 0.9322660677114132 and for testing data is 0.9319541071286874.

The RMSE score for training data is 0.2602574346461341 and for testing data is 0.26085607693000473.

The variance inflation factor on each variable using the model is given below-

FIGURE 9



```
carat ---> 5.063098522618652
cut ---> 10.005831322701823
color ---> 5.536356458389349
clarity ---> 5.326134545167012
depth ---> 592.7757130470161
table ---> 559.3255368133664
```

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

Some observations made from the model, and insights that can be provided are :

- We are able to understand that the diamond is being produced artificially (x,y,z dimensions) to compete viably with naturally available diamond resources
- Price seems to be dependent on carat, weight. For every unit increase in carat, there seems to be a rise in the price of the diamond
- Similarly clarity also seems to influence price – for every unit increase in clarity, the diamond price is reduced by almost 1/4th
- Colour also seems to influence the price – indicating the brightness of the crystal
- A minimum Cut is evident – which shows that the process requires basic level of machine to increase the cut and depth of the diamond

Problem 2

Problem statement-

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 Description-

Holiday_Package: Opted for Holiday Package yes/no?

Salary: Employee salary

Age: Age in years

Edu: Years of formal education

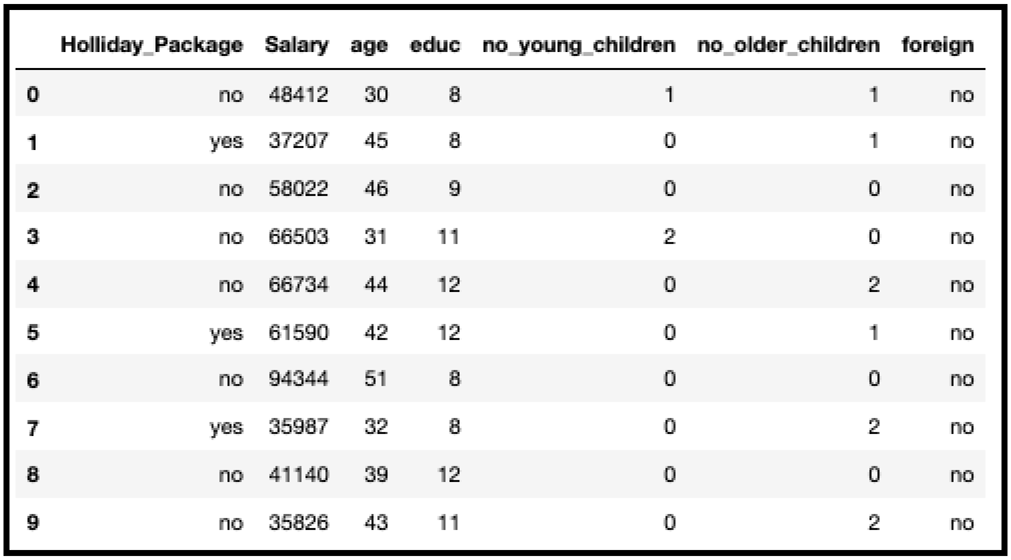
No_young_children: The number of young children (younger than 7 years)

No_older_children: Number of older children

Foreign: foreigner Yes/No

Sample of the dataset-

FIGURE 10



	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	no	48412	30	8	1	1	no
1	yes	37207	45	8	0	1	no
2	no	58022	46	9	0	0	no
3	no	66503	31	11	2	0	no
4	no	66734	44	12	0	2	no
5	yes	61590	42	12	0	1	no
6	no	94344	51	8	0	0	no
7	yes	35987	32	8	0	2	no
8	no	41140	39	12	0	0	no
9	no	35826	43	11	0	2	no

There are 7 variables out of which 2 are object type and 5 are int values. The data given is for 872 individuals. There are no null values.

Types of variables in the data frame-

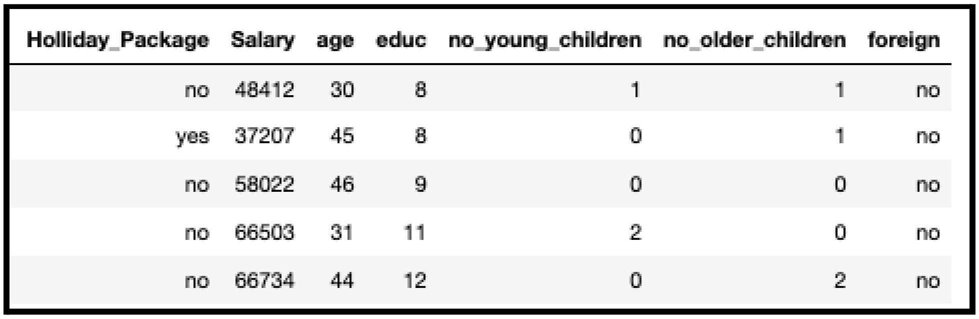
TABLE 6

Holliday_Package	object	Categorical
Salary	int64	Continuous
age	int64	Continuous
educ	int64	Continuous
no_young_children	int64	Continuous
no_older_children	int64	Continuous
foreign	object	Categorical

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.

The below table gives the first 5 rows of sample data.

FIGURE 11

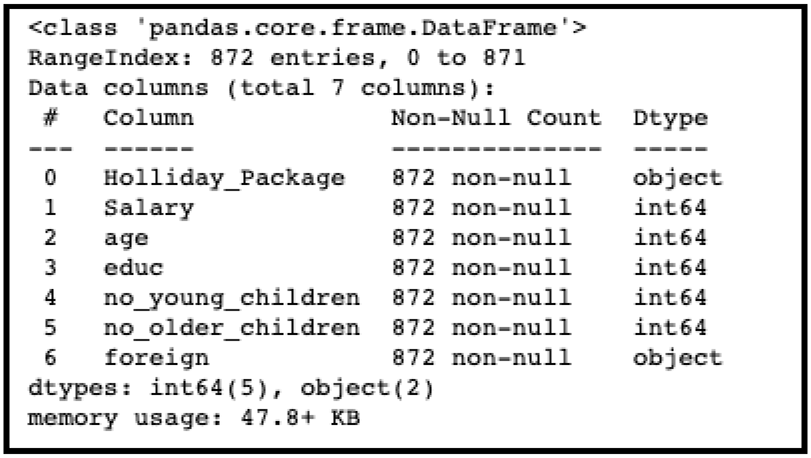


Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
no	48412	30	8	1	1	no
yes	37207	45	8	0	1	no
no	58022	46	9	0	0	no
no	66503	31	11	2	0	no
no	66734	44	12	0	2	no

The image below gives the basic information of the data set. It is clear that the variables have int and object data types with 7 columns and 872 rows. The dataset has no null values. The memory usage is 47.8+ KB.

The dataset also does not contain any duplicates.

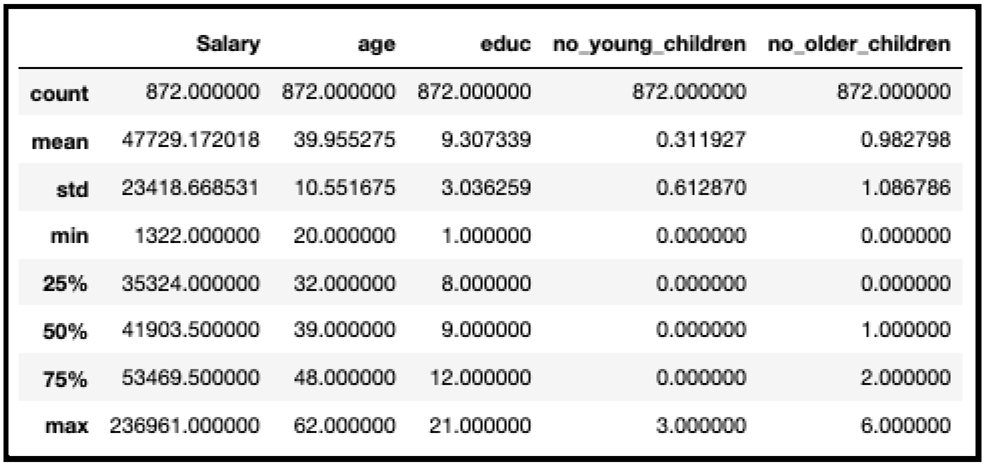
FIGURE 12



```
<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)
memory usage: 47.8+ KB
```

The below image gives the five point summary of the continuous variables in the data set. It is clear that the data needs scaling as the numbers are of different magnitude.

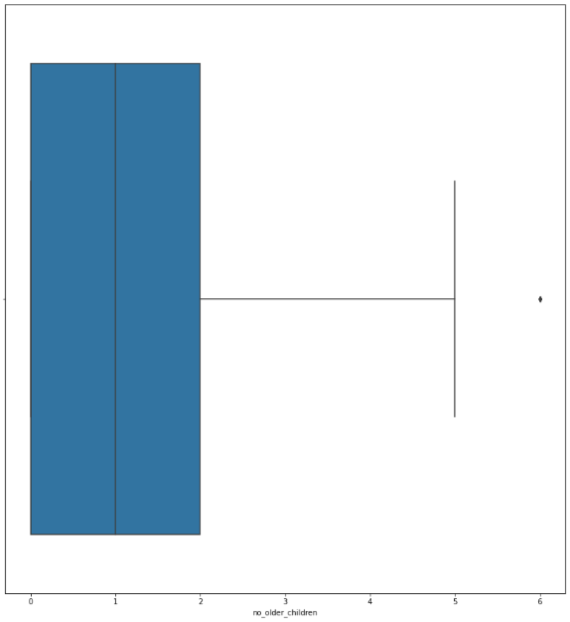
FIGURE 13

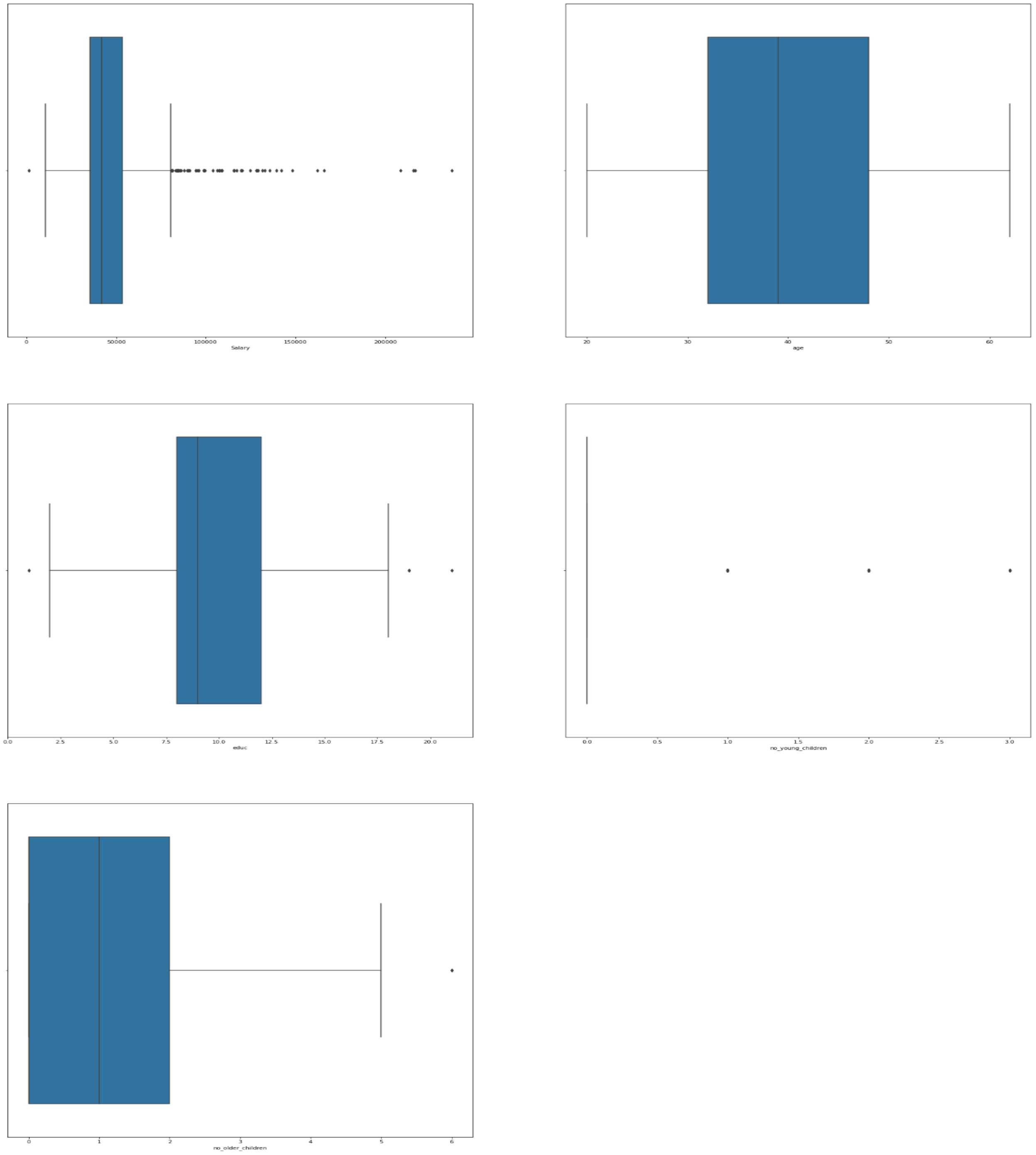


	Salary	age	educ	no_young_children	no_older_children
count	872.000000	872.000000	872.000000	872.000000	872.000000
mean	47729.172018	39.955275	9.307339	0.311927	0.982798
std	23418.668531	10.551675	3.036259	0.612870	1.086786
min	1322.000000	20.000000	1.000000	0.000000	0.000000
25%	35324.000000	32.000000	8.000000	0.000000	0.000000
50%	41903.500000	39.000000	9.000000	0.000000	1.000000
75%	53469.500000	48.000000	12.000000	0.000000	2.000000
max	236961.000000	62.000000	21.000000	3.000000	6.000000

There are outliers in all the continuous variables except “age”. This is evident from the box plots below,

GRAPH 9




0
50000
100000
150000
200000
Salary
20
30
40
50
60
age
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
educ
0.0
0.5
1.0
1.5
2.0
2.5
3.0
no_young_children
0
1
2
3
4
5
6
no_older_children

There are both categorical variables and continuous variables in the given data set. So univariate and bivariate analysis can be done in both.

Univariate analysis:

Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships (unlike regression) and it’s major purpose is to describe; It takes data, summarizes that data and finds patterns in the data. [1]

The histograms are used for numerical variables and count plots are used for categorical variables to perform univariate analysis.

It is clear from the graph (*GRAPH 10*) that all the numerical variables are skewed.

Skewness:

Skewness is a number that indicates to what extent a variable is asymmetrically distributed. The below table represents the level of skewness and the respective skewness value. [2]

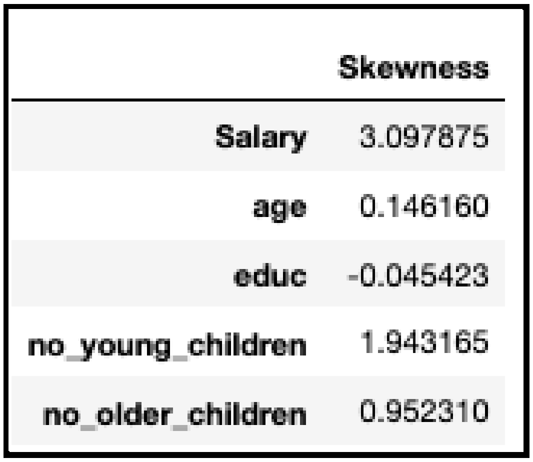
TABLE 7

Skewness level	**Value**
Symmetrical or Not Skewed	0
Less Skewed Data	± 0.5 to 1
Highly Skewed Data	Greater than ±1

When the skewness value is positive it is considered as right skewed data and when the skewness value is negative it is considered as left skewed data.

The table below shows the skewness value corresponding to each variable in the given data set.

FIGURE 14

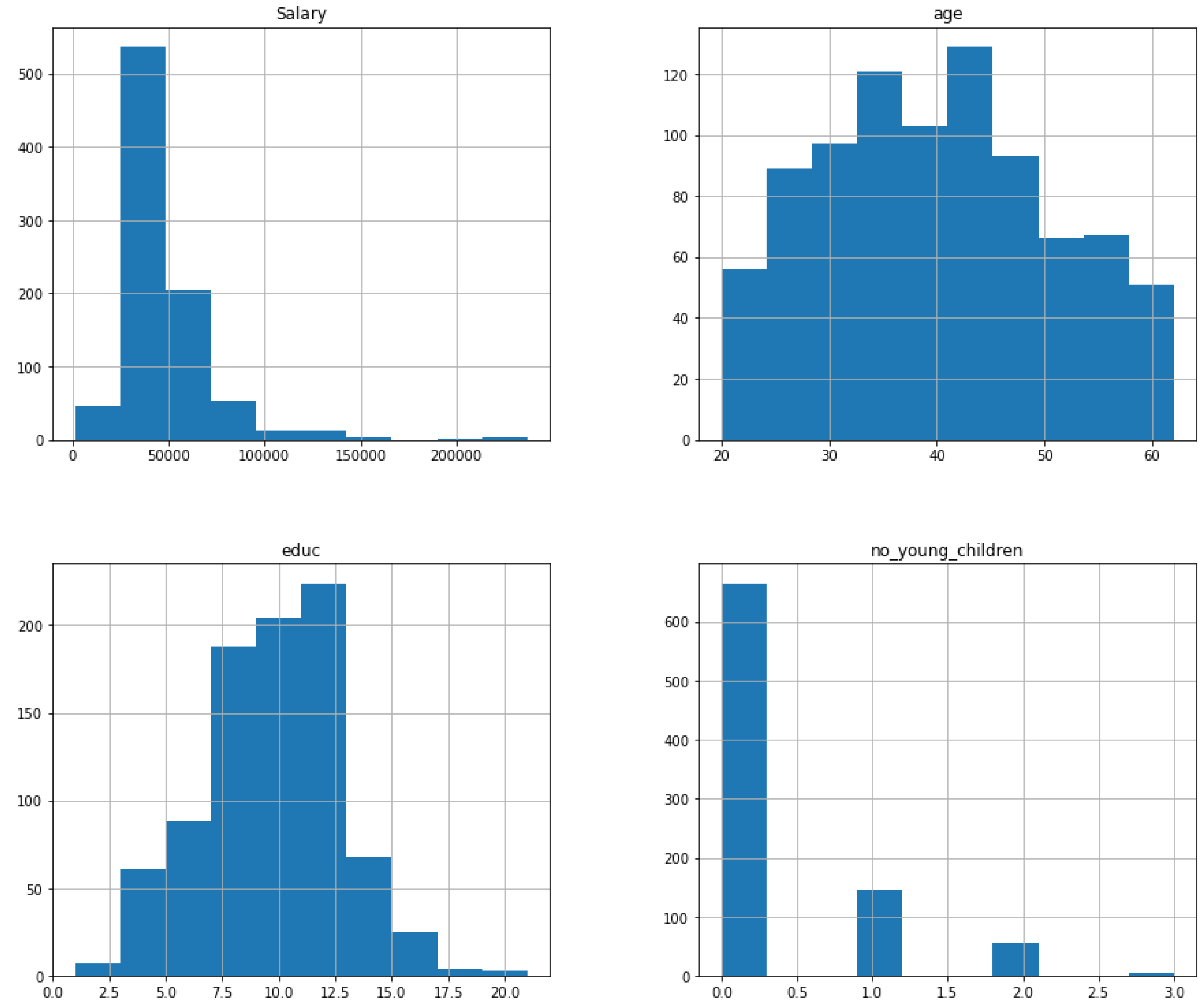


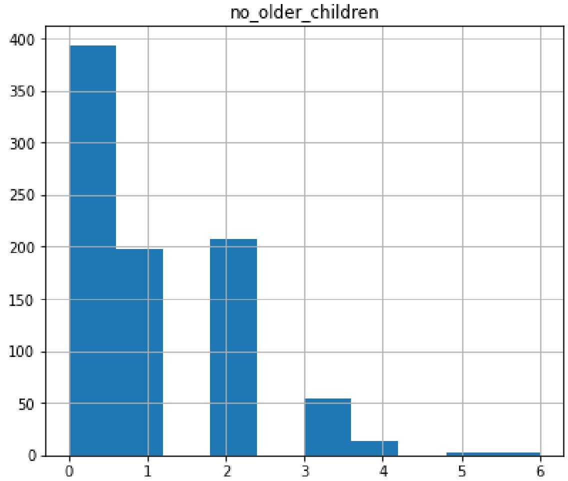
	Skewness
Salary	3.097875
age	0.146160
educ	-0.045423
no_young_children	1.943165
no_older_children	0.952310

All the variables have a positive skew value except for “educ”. Therefore, all the other variables above are right skewed variables and “educ” is left skewed variable.

Univariate Analysis for Continuous variables-

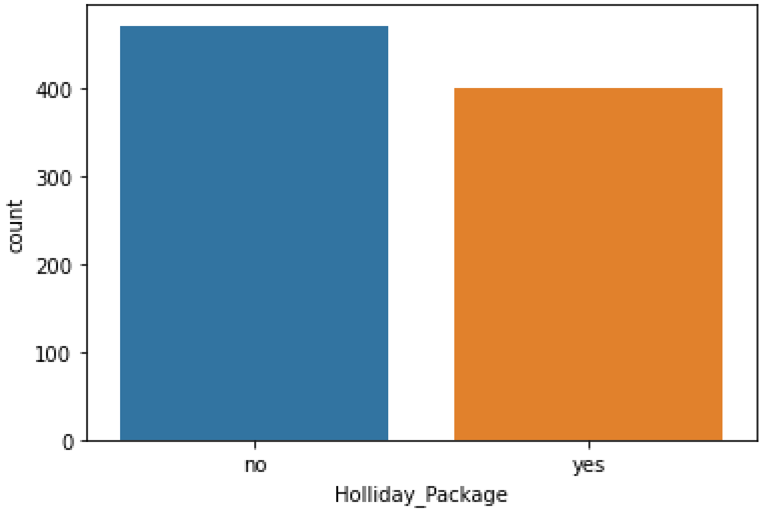
GRAPH 10



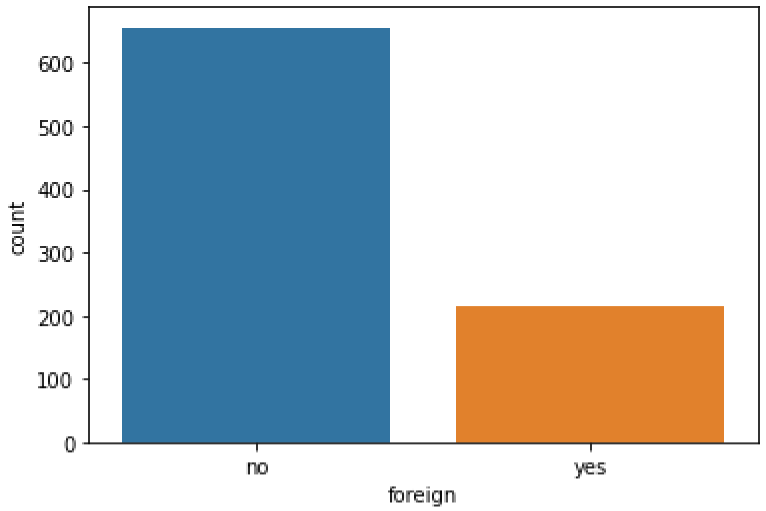


Univariate Analysis for Categorical variables-

GRAPH 11



GRAPH 12



Bivariate analysis:

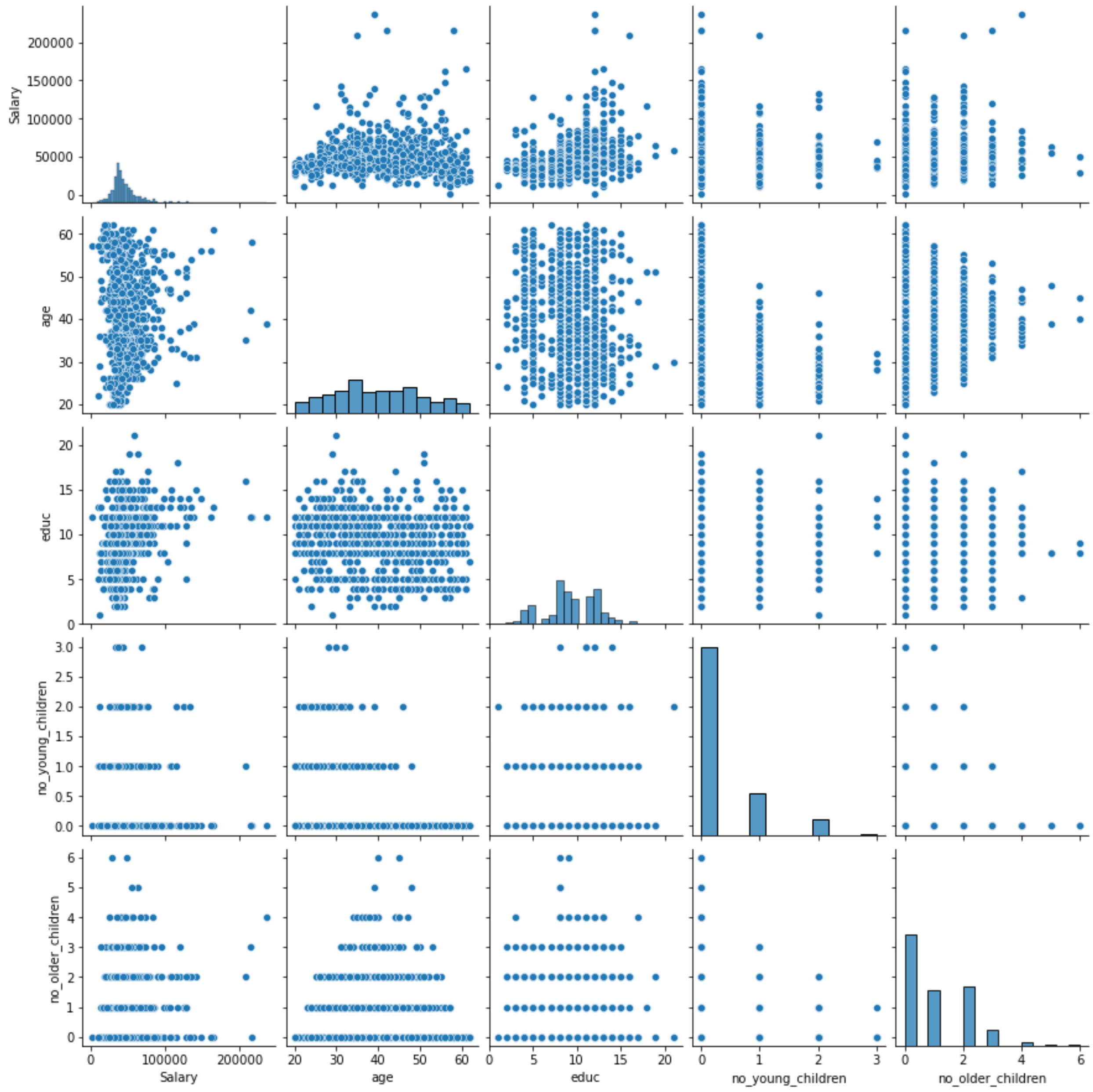
Bivariate analysis is one of the simplest forms of quantitative (statistical) analysis. It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical relationship between them. Bivariate analysis can be helpful in testing simple hypotheses of association. [3]

The pairplot is generally used for numerical variables and box plots are used for categorical with numerical variables to perform bivariate analysis.

A pairplot plots a pairwise relationship in a dataset. The pairplot function creates a grid of Axes such that each variable in data will be shared in the y-axis across a single row and in the x-axis across a single column. [4]

A boxplot is a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”). It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed. [5]

GRAPH 13

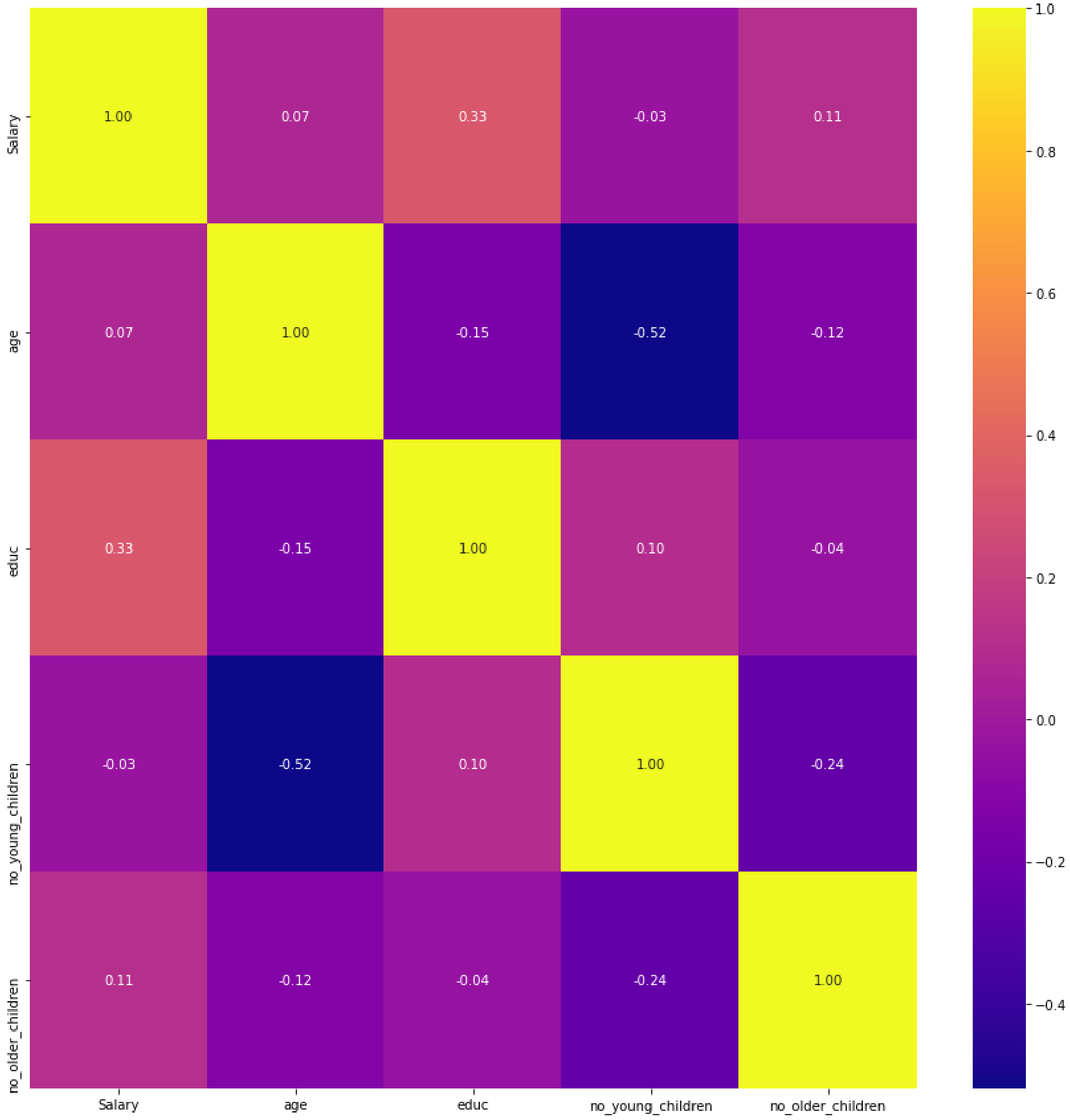


The below are the findings from the pairplot generated -

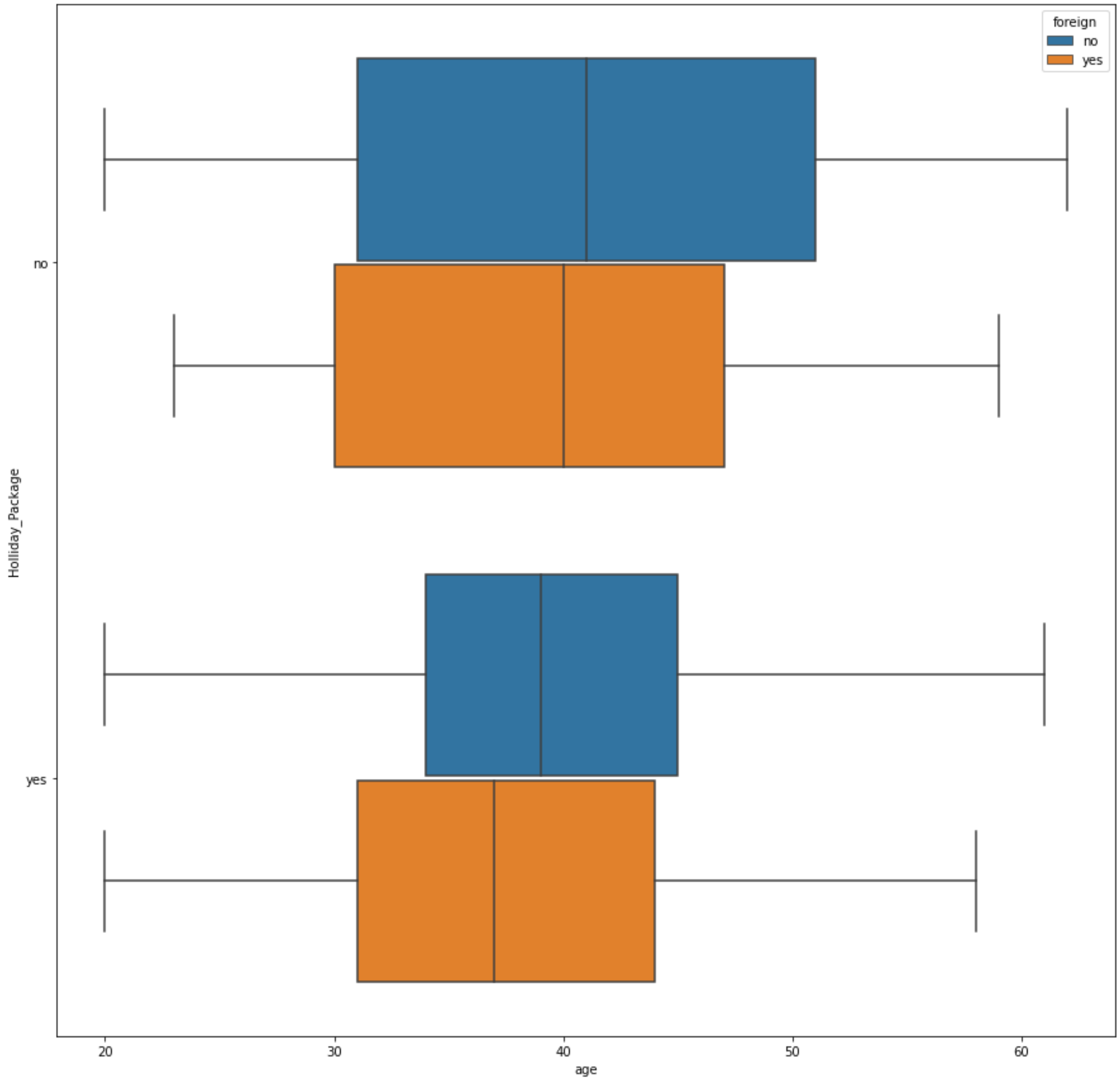
- None of the variables are correlated with each other

The heat map can also be used to check the association between two variables. All the boxes with a value higher than 0.8 are highly correlated. It is clear that none of the variables have a value more than 0.8. The heat map for all the numerical variable is below,

GRAPH 14



GRAPH 15



The below are the findings from the boxplot generated -

- The individuals of age between 35 and 45 seems to take the holiday package and they are foreigners.

- Majority of the individuals between the age 30 to 50 neither take the holiday package nor are foreigners.

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).

The dataframe contains 872 records of data evenly spread. The dataset contains no null values.

We convert all object types into numerical variables to make it convenient to work on modeling. There are two object type variables in the given data set that need encoding.

The below table gives the encoded value for each unique data in the respective variable.

VARIABLE: Holliday_Package

TABLE 8

yes	0
no	1

VARIABLE: foreign

TABLE 9

no	0
yes	1

LOGISTIC REGRESSION:

In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1, with a sum of one. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". [8]

LINEAR DISCRIMINANT ANALYSIS:

Linear Discriminant Analysis (LDA) is a dimensionality reduction technique. As the name implies dimensionality reduction techniques reduce the number of dimensions (i.e. variables) in a dataset while retaining as much information as possible. For instance, suppose that we plotted the relationship between two variables where each color represents a different class. [9]

From the given data, 30 percent is taken for the test size and 70 percent is taken for the training. The target variable here is “Holliday_Package”. The data is fitted into the Logistic Regression Model and Linear Discriminant Analysis(LDA).

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 – How accurately / cleanly does the model classify the data points. Lesser the false predictions, more the accuracy.

Sensitivity / Recall – How many of the actual True data points are identified as True data points by the model . Remember, False Negatives are those data points which should have been identified as True.

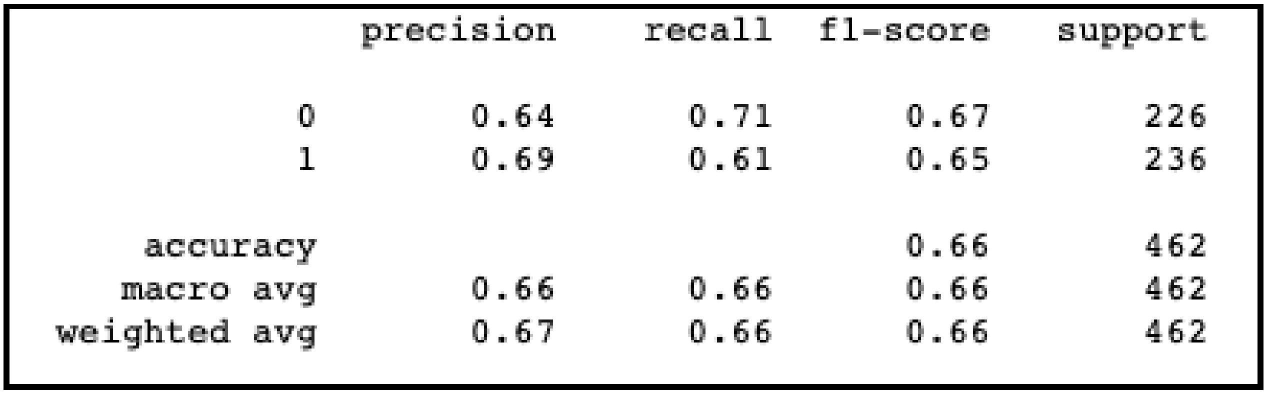
Specificity – How many of the actual Negative data points are identified as negative by the model

Precision – Among the points identified as Positive by the model, how many are really Positive

LOGISTIC REGRESSION:

The below is the classification report of the training data-

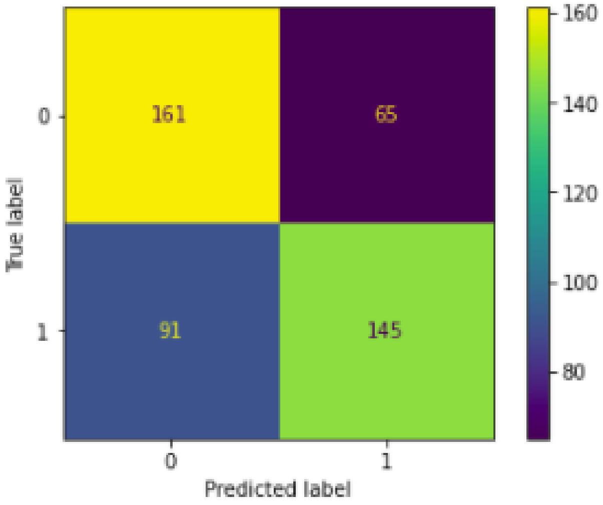
FIGURE 15



```
              precision    recall  f1-score   support

           0       0.64      0.71      0.67       226
           1       0.69      0.61      0.65       236

    accuracy                           0.66       462
   macro avg       0.66      0.66      0.66       462
weighted avg       0.67      0.66      0.66       462
```

The below is the confusion matrix for training data-

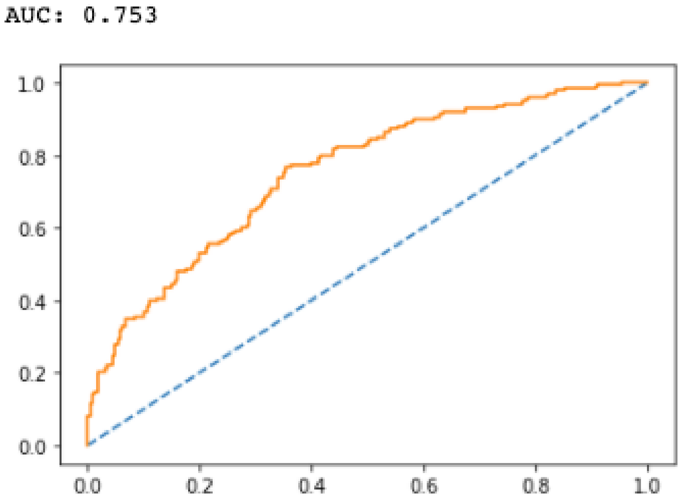
FIGURE 16



The training data accuracy is 66.23%.

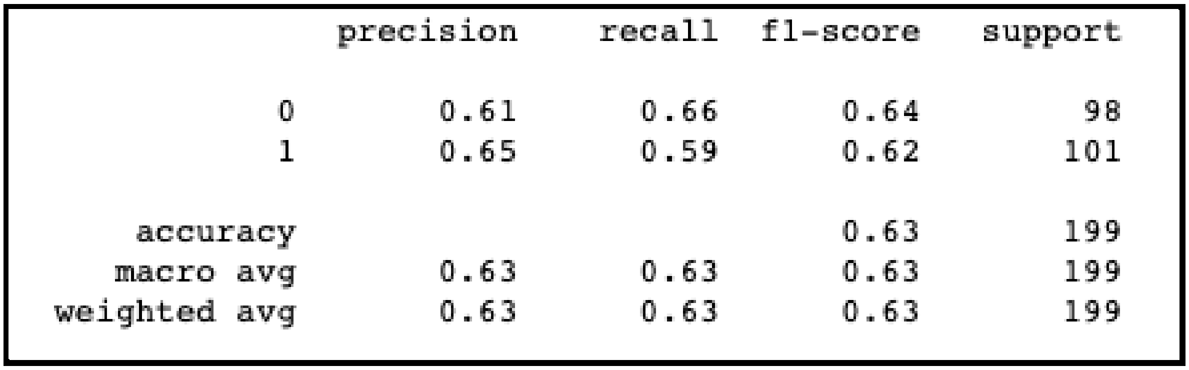
ROC curve and AUC score for training data-

GRAPH 16



The below is the classification report of the testing data-

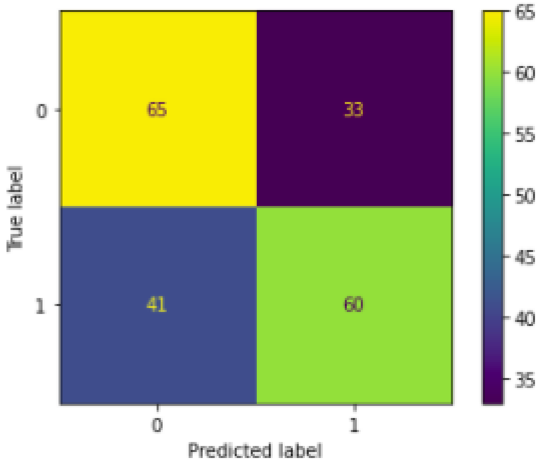
FIGURE 17



```
              precision    recall  f1-score   support

           0       0.61      0.66      0.64        98
           1       0.65      0.59      0.62       101

    accuracy                           0.63       199
   macro avg       0.63      0.63      0.63       199
weighted avg       0.63      0.63      0.63       199
```

The below is the confusion matrix for testing data-

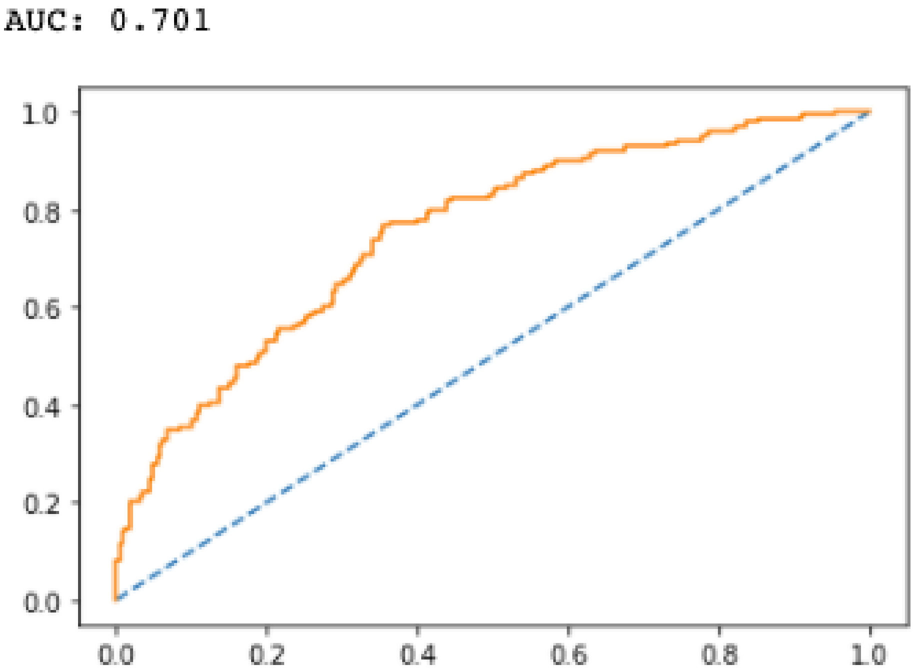
FIGURE 18



The training data accuracy is 62.81%.

ROC curve and AUC score for testing data-

GRAPH 17

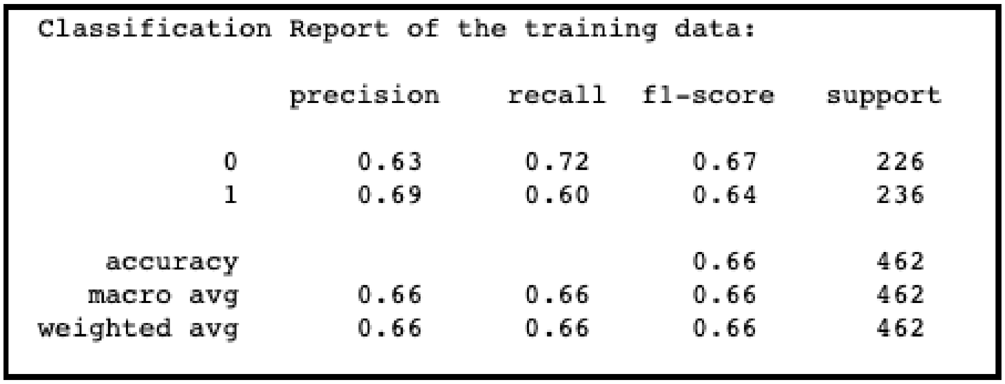


```
AUC: 0.701
```

LINEAR DISCRIMINANT ANALYSIS:

The below is the classification report of the training data-

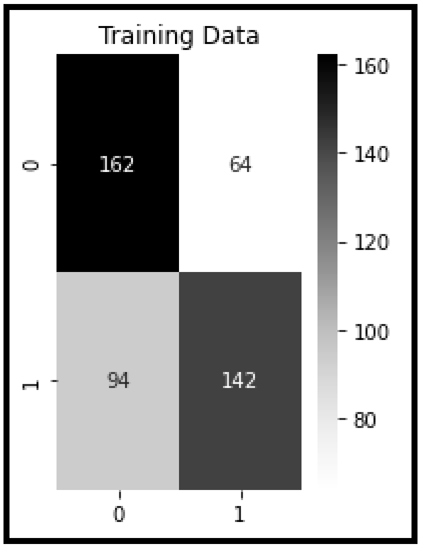
FIGURE 19



```
Classification Report of the training data:

              precision    recall  f1-score   support

           0       0.63      0.72      0.67       226
           1       0.69      0.60      0.64       236

    accuracy                           0.66       462
   macro avg       0.66      0.66      0.66       462
weighted avg       0.66      0.66      0.66       462
```

The below is the confusion matrix for training data-

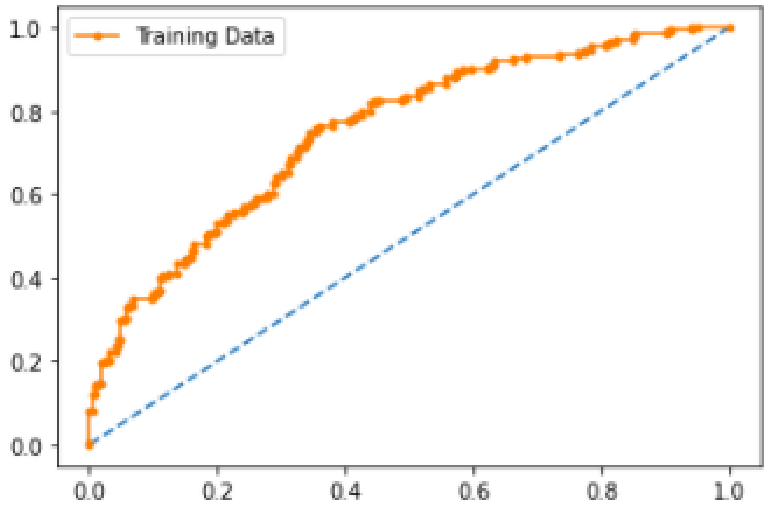
FIGURE 20



ROC curve and AUC score for training data-

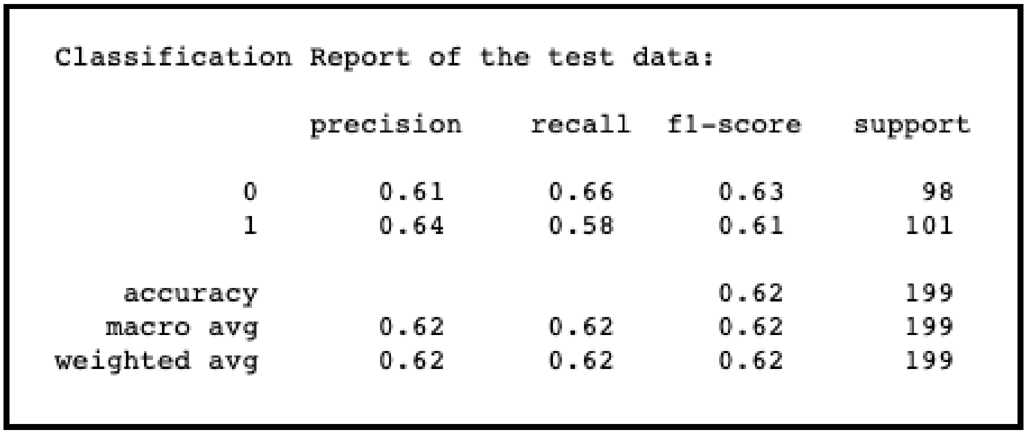
GRAPH 18

```
AUC for the Training Data: 0.753
```



The below is the classification report of the testing data-

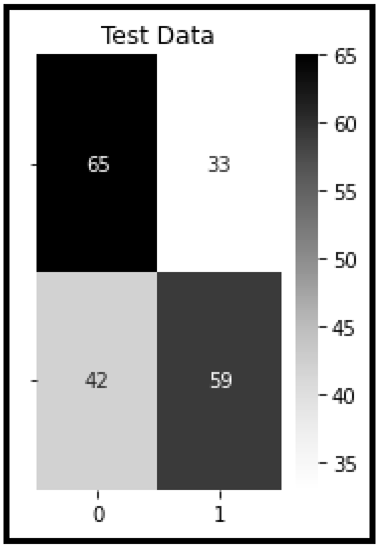
FIGURE 21



```
Classification Report of the test data:

              precision    recall  f1-score   support

           0       0.61      0.66      0.63        98
           1       0.64      0.58      0.61       101

    accuracy                           0.62       199
   macro avg       0.62      0.62      0.62       199
weighted avg       0.62      0.62      0.62       199
```

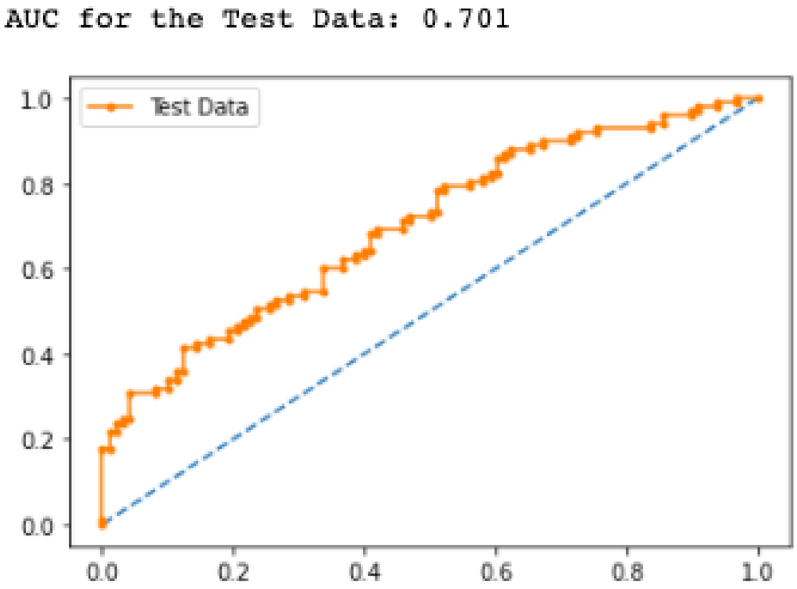
The below is the confusion matrix for testing data-

FIGURE 22



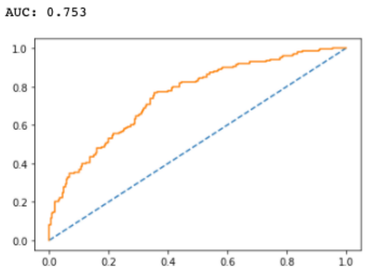
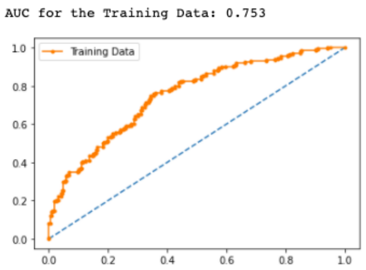
ROC curve and AUC score for testing data-

GRAPH 19

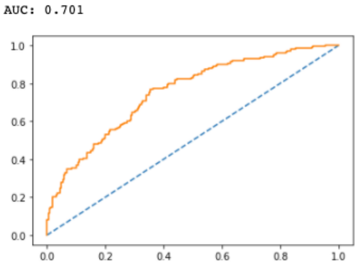
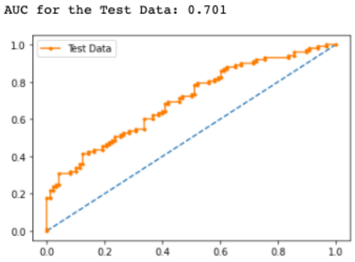


```
AUC for the Test Data: 0.701
```

Training data-

	Logistic Regression	LDA
Precision	0.69	0.69
Recall	0.61	0.60
F1 score	0.65	0.64
Accuracy	66.23%	65.8%
AUC score	0.753	0.753
ROC curve		

Testing data-

	Logistic Regression	LDA
Precision	0.65	0.64
Recall	0.59	0.58
F1 score	0.62	0.61
Accuracy	62.81%	62.31%
AUC score	0.701	0.701
ROC curve	 AUC: 0.701 	 AUC for the Test Data: 0.701

Inference:

Based on the table of comparison displayed above, it is clear that the Logistic Regression model has the best accuracy in both training and testing data. The F1 score is high which means both type 1 and type 2 errors are reduced. Also, the recall score is high in both training and testing which means type 2 error is low.

The Precision value is almost the same for both the models which means type 1 error is low in this.

The Area under the curve (AUC) is the same in both the models.

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

Some observations made from the model, and insights that can be provided are:

- Employees with 2 or lesser children tend to travel when their children are young. Not many travel with grown-up children. This might be to make use of maximum benefits of travelling with “Children” and not with “Adults. This might also be a strategic move to reduce travel expenses at an early age
- Employees without any children tend to travel around the age of 45 and they almost do not travel after crossing 50 years of age
- For employees willing to travel more, it can be observed that the average age of employees are around 35 years
- Education seems to influence salary of employees – however it does not seem to influence the willingness to opt for vacations
- Around 58% of foreigners tend to travel frequently – and this might even be to utilize the holiday season or travel to their home countries once in a while
- We can summarise and safely say that employees with young children tend to travel more than ones with older children. There can be various reasons that we can think of – such as conflicting calendars between parents and children as they grow up / different holiday seasons and so on. Also for employees without children, they tend to travel when they are around 45, and almost do not travel once they cross 50 years of age

References

Websites-

[1] https://www.statisticshowto.com/univariate/

[2] https://www.spss-tutorials.com/skewness/

[3] https://en.wikipedia.org/wiki/Bivariate_analysis

[4] https://pythonbasics.org/seaborn-pairplot/

[5] https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd5

[6] https://en.wikipedia.org/wiki/Feature_scaling

[7] https://en.wikipedia.org/wiki/Linear_regression

[8] https://en.wikipedia.org/wiki/Logistic_regression

[9] https://towardsdatascience.com/linear-discriminant-analysis-in-python-76b8b17817c2

End of Project