DATA MINING PROJECT

Submitted by,

JIYA JACOB

PGP-DSBA ONLINE

JULY-B 2021

DATE:19/12/2021

CUBIC ZIRCONIA

CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

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.

INTRODUCTION

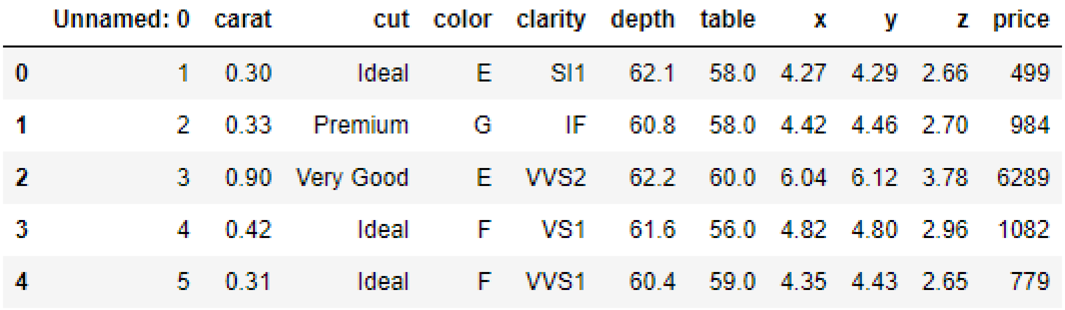
The purpose of this whole exercise is to perform multivariate linear regression on the given data. In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables. Different models of linear regression are tried out by using different encoding techniques, scaling and without scaling and dropping insignificant variables. The variables are found to have multicollinearity and this fact is taken into consideration while creating the linear regression model.

DATA DESCRIPTION

1. Carat: Carat weight of the cubic zirconia.
2. Cut: Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal.
3. Color: Colour of the cubic zirconia. With D being the worst and J the best.
4. 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
5. Depth: The Height of cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter.
6. Table: The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter.
7. Price: The Price of the cubic zirconia.
8. X: Length of the cubic zirconia in mm.
9. Y: Width of the cubic zirconia in mm.
10. Z: Height of the cubic zirconia in mm.

Q1.1: **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.**

a. Sample Dataset.

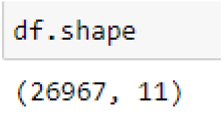


	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
0	1	0.30	Ideal	E	SI1	62.1	58.0	4.27	4.29	2.66	499
1	2	0.33	Premium	G	IF	60.8	58.0	4.42	4.46	2.70	984
2	3	0.90	Very Good	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289
3	4	0.42	Ideal	F	VS1	61.6	56.0	4.82	4.80	2.96	1082
4	5	0.31	Ideal	F	VVS1	60.4	59.0	4.35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

Table 1: Dataset Sample

The above table shows the head of the given dataset, i.e. the first five entries to ensure that the dataset has been loaded without any issues. The data consists of details of 26967 cubic zirconias along with other attributes arranged in 11 columns such as cut, color, clarity, depth, price etc. The first column is an index number ("Unnamed: 0"). As this only a serial no, we can remove it.

b. Shape of the dataset

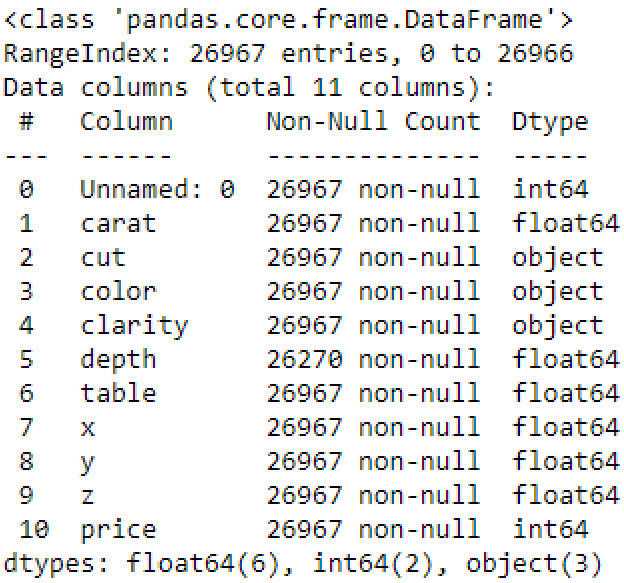


```
df.shape
```

```
(26967, 11)
```

There are total 26967 rows and 11 columns in the given dataset.

c. Checking for the missing values in the dataset.

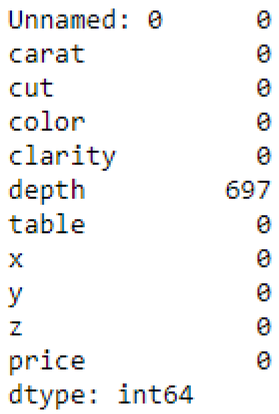


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26967 entries, 0 to 26966
Data columns (total 11 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   Unnamed: 0  26967 non-null  int64  
 1   carat       26967 non-null  float64
 2   cut         26967 non-null  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 2. Checking for the missing values in the dataset

From the above results we can see that there are some missing values present in the depth variable of the given dataset.

d. Checking for the null values in the given dataset

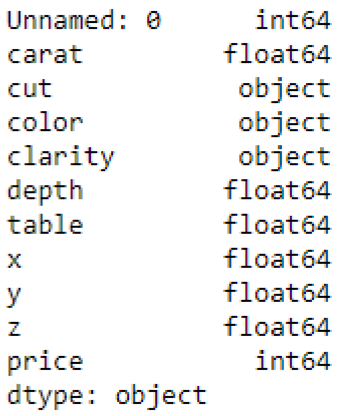


```
Unnamed: 0      0
carat           0
cut             0
color           0
clarity         0
depth         697
table           0
x               0
y               0
z               0
price           0
dtype: int64
```

Table 3. Checking for the null values in the dataset

From the above table we can see that depth variable contains 697 null values, which needs to be imputed.

e. Checking for the datatype of variables present in the dataset.

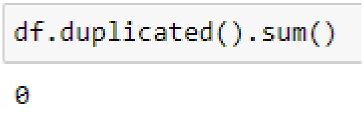


```
Unnamed: 0      int64
carat         float64
cut            object
color          object
clarity        object
depth         float64
table         float64
x             float64
y             float64
z             float64
price           int64
dtype: object
```

Table 4. Checking the data type of variables in the data set

From the above table we can see that out of 11 variables, the data type of the 6 variables in the given data set is float in nature while 3 variables are of object data type and 2 variables are of int data type.

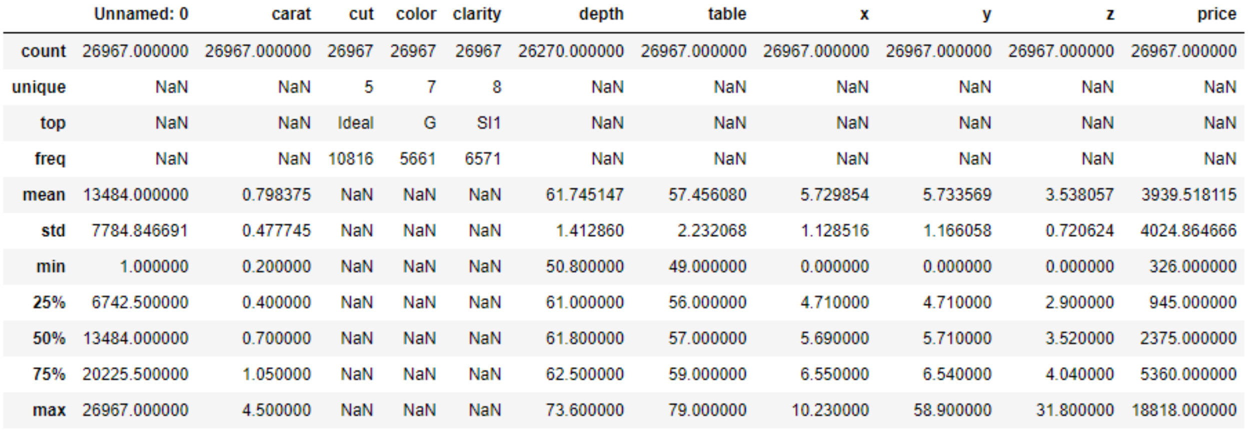
f. Checking for the number of duplicated values



```
df.duplicated().sum()

0
```

From the above output we can see that there are no duplicated values in the given dataset.

g. Checking for the summary of the given dataset.



	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
count	26967.000000	26967.000000	26967	26967	26967	26270.000000	26967.000000	26967.000000	26967.000000	26967.000000	26967.000000
unique	NaN	NaN	5	7	8	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	NaN	Ideal	G	SI1	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	NaN	10816	5661	6571	NaN	NaN	NaN	NaN	NaN	NaN
mean	13484.000000	0.798375	NaN	NaN	NaN	61.745147	57.456080	5.729854	5.733569	3.538057	3939.518115
std	7784.846691	0.477745	NaN	NaN	NaN	1.412860	2.232068	1.128516	1.166058	0.720624	4024.864666
min	1.000000	0.200000	NaN	NaN	NaN	50.800000	49.000000	0.000000	0.000000	0.000000	326.000000
25%	6742.500000	0.400000	NaN	NaN	NaN	61.000000	56.000000	4.710000	4.710000	2.900000	945.000000
50%	13484.000000	0.700000	NaN	NaN	NaN	61.800000	57.000000	5.690000	5.710000	3.520000	2375.000000
75%	20225.500000	1.050000	NaN	NaN	NaN	62.500000	59.000000	6.550000	6.540000	4.040000	5360.000000
max	26967.000000	4.500000	NaN	NaN	NaN	73.600000	79.000000	10.230000	58.900000	31.800000	18818.000000

TABLE 5: SUMMARY OF THE DATA

From the above table, we can see that we have both categorical and continuous data. For categorical data we have cut, color and clarity and for continuous data we have carat, depth, table, x. y, z and price. Price will be target variable. From the summary of the dataset, we can see that mean of the price is the highest while the mean for carat is least among all. The variables x and y share almost the similar mean whereas depth has slightly higher mean than table. Similarly, price has the highest standard deviation whereas carat has the lowest standard deviation, followed by z that has slightly higher standard deviation. Among all variables, carat has the least value for all the statistical measurements.

h. Graphical Representation of univariate and bivariate analysis for continuous columns

1.Carat

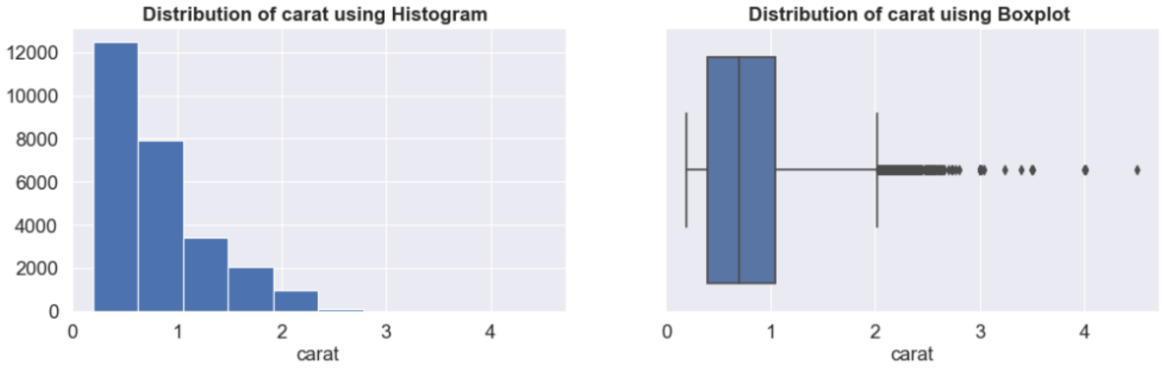


Fig 1: The box plot and the histogram showing the distribution of data of 'Carat' variable

From the univariate analysis using histogram, we see most cubic zirconia has the weight between 0.2 and 0.5 carat, approximately 12000, followed by those cubics that weigh between 0.5 to 1 carat, approximately 8000 and the least number of cubics in the range of 4 to 4.5 carat, which is approximately 2. From the bivariate analysis using boxplot, there are presence of outliers. The second quartile(Q2) or median for the carat variable is about 0.7. The lower or first quartile(Q1) is about 0.4 and the upper or the third quartile(Q3) is about 1.05. The inter quartile range (IQR) for the above boxplot is 0.65

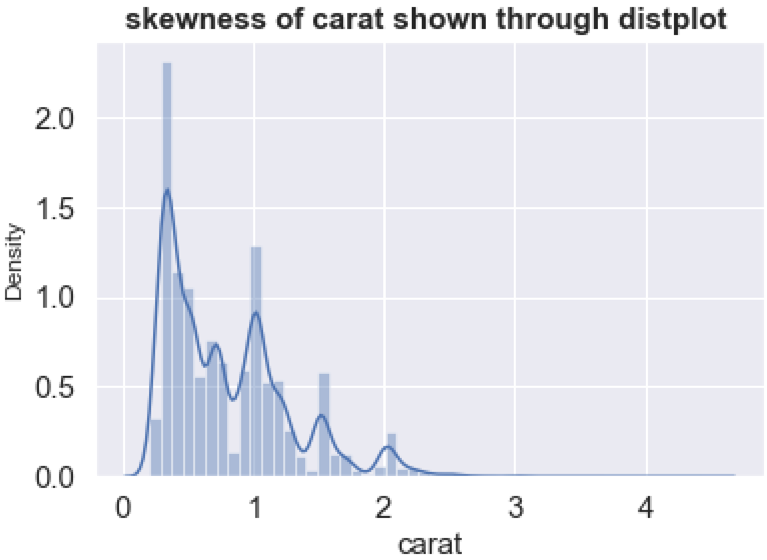


Fig 2: The distplot showing the skewness of data of 'Carat' variable

From the above distplot , we can see that carat is positively skewed or right skewed, with median=0.7,mode=0 and 0.3,mean= 0.798. Carat is a bi modal variable with two modes.

2.Depth

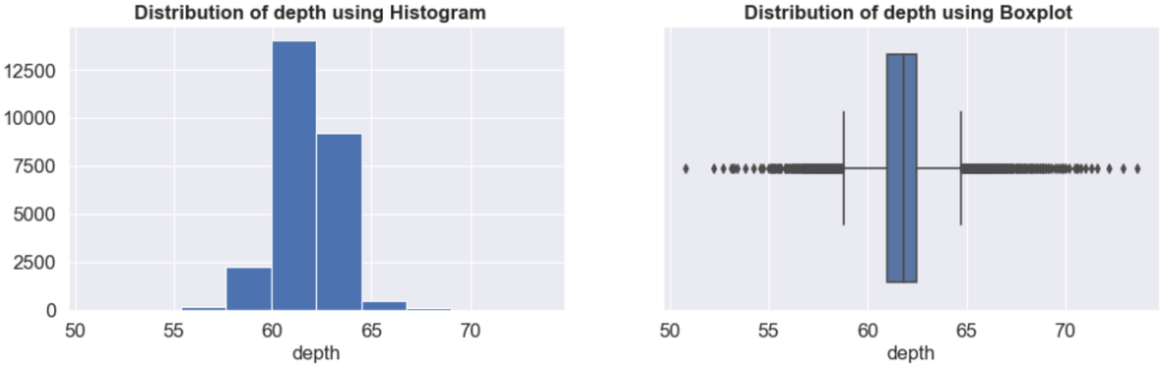


Fig 3: The box plot and the histogram showing the distribution of data of ''Depth'' variable

From the univariate analysis using histogram, we see that the depth of cubics are mostly in the range of 60 to 62 mm, approximately 14,500, followed by those cubics that are having depth in the range of 62 to 65mm, approximately 9500 and the least depth range of cubics are from 67 to 69.5 which is less than 10. From the bivariate analysis using boxplot, there are presence of outliers. The second quartile(Q2) or median for the depth variable is about 61.8. The lower or first quartile(Q1) is about 61.0 and the upper or the third quartile(Q3) is about 62.5. The inter quartile range (IQR) for the above boxplot is 1.5

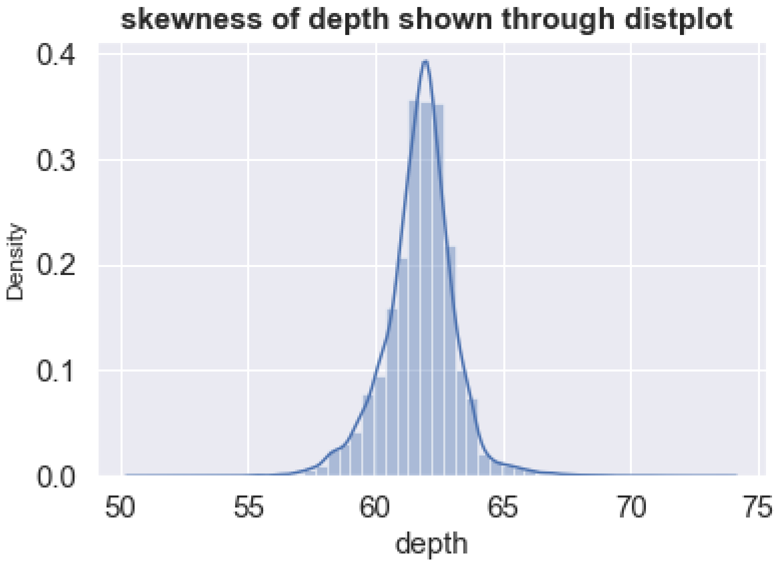


Fig 4: The distplot showing the skewness of data of 'depth' variable

From the above distplot , we can see that depth is negatively skewed or left skewed, with median=61.8,mode=0 and 62.0,mean=61.745. Depth is a bi modal variable with two modes.

3.Table

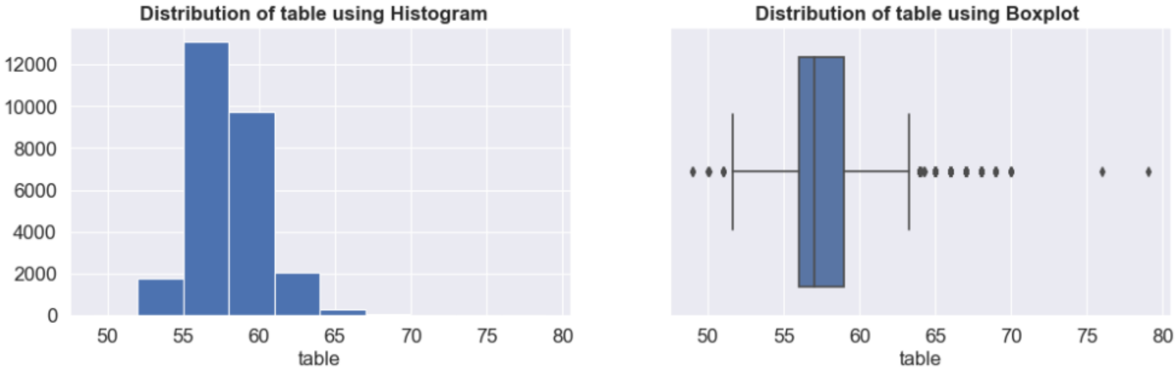


Fig 5: The box plot and the histogram showing the distribution of data of "Table" variable

From the univariate analysis using histogram, we see that the table of the cubic zirconia is plotted along x-axis. Cubics having the table in the range 55 to 58.5 are the most (approximately 13800) The second highest table of the cubics are in the range of 58.5 to 61 mm (10000). The cubics having the table in the range of 64 to 66.5 mm is the least group (approximately 200). From the bivariate analysis using boxplot, there are presence of outliers. The second quartile(Q2) or median for the table variable is about 57. The lower or first quartile(Q1) is about 56 and the upper or the third quartile(Q3) is about 59. The inter quartile range (IQR) for the above boxplot is 3

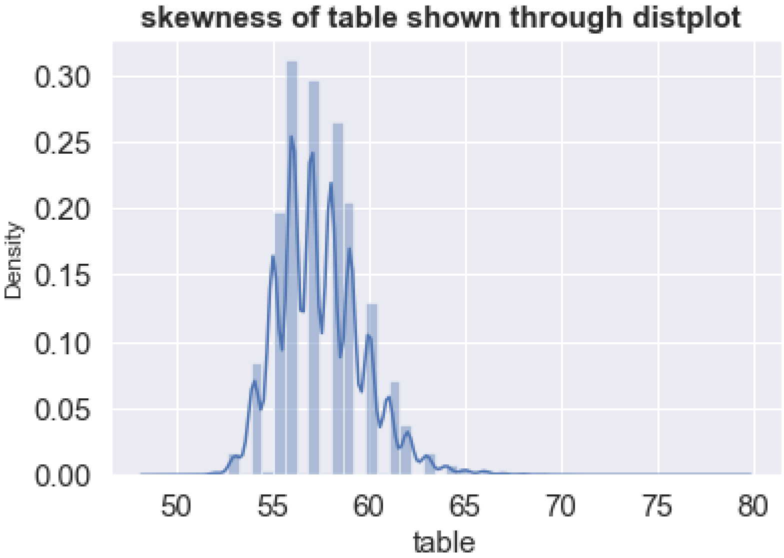


Fig 6: The distplot showing the skewness of data of 'Table' variable

From the above distplot , we can see 'table' is positively skewed or right skewed, with median=57,mode= 0 and 56, mean=57.45

4.x

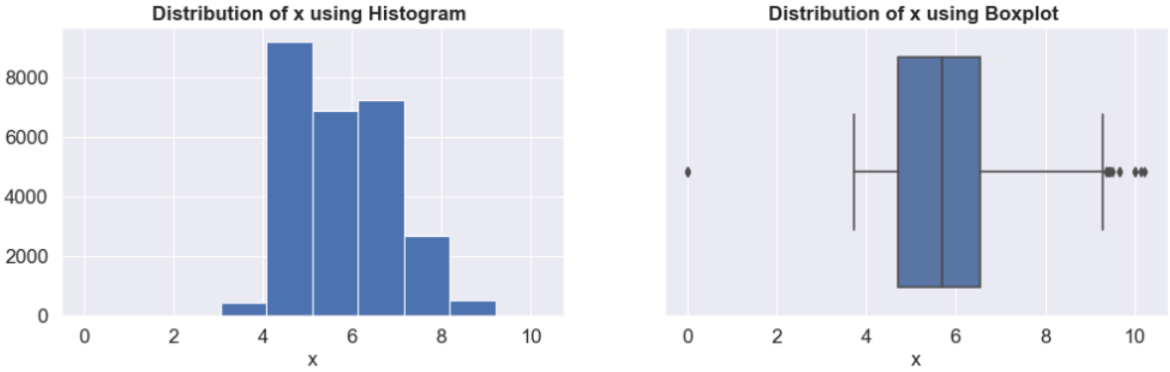


Fig 7: The box plot and the histogram showing the distribution of data of 'x' variable

From the univariate analysis using histogram, we see that the length of the cubic zirconia is plotted along x-axis and labelled as 'x'. Cubics having the length in the range 4 to 5 mm are the most (approximately 9500) The second highest length of the cubics are in the range of 6 to 7mm (7000). The cubics having the length in the range of 3 to 4 mm is the least group (less than 500). From the bivariate analysis using boxplot, there are presence of outliers. The second quartile(Q2) or median for the x variable is about 5.69. The lower or first quartile(Q1) is about 4.71 and the upper or the third quartile(Q3) is about 6.55. The inter quartile range (IQR) for the above boxplot is 1.839

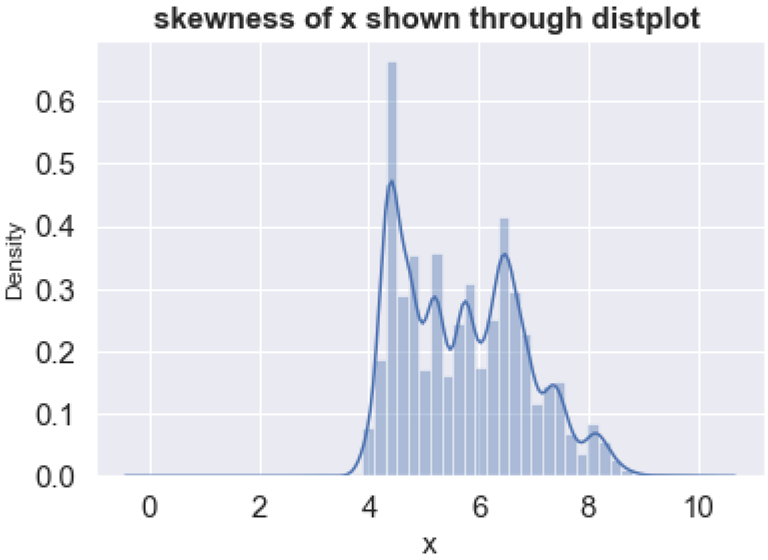


Fig 8: The distplot showing the skewness of data of ‘x’ variable

From the above distplot , we can see that x is positively skewed or right skewed, with median=5.69,mode=0 and 4.38, mean=5.729. x is a bi modal variable with two modes .

5.y

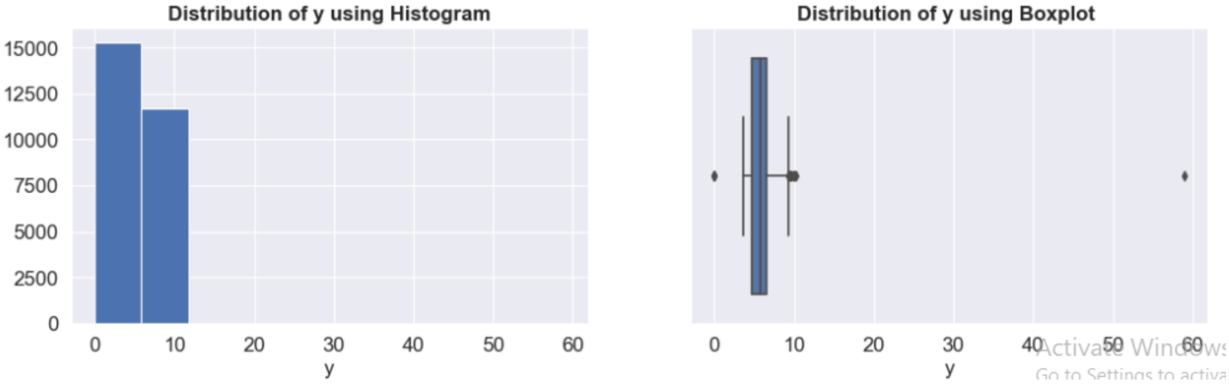


Fig 9: The box plot and the histogram showing the distribution of data of ‘y’ variable

From the univariate analysis using histogram, we see that the width of the cubic zirconia is plotted along x-axis and labelled as ‘y’. It is noticed that only two groups are formed while plotting the ‘y’ variable using histogram. Cubics having the width in the range 0 to 5 mm are the most (approximately 15000). The cubics having the length in the range of 5 to 10 mm is the least group (approximately 12500). From the bivariate analysis using boxplot, there are presence of outliers. The second quartile(Q2) or median for the y variable is about 5.71. The lower or first quartile(Q1) is about 4.71 and the upper or the third quartile(Q3) is about 6.54. The inter quartile range (IQR) for the above boxplot is 1.83

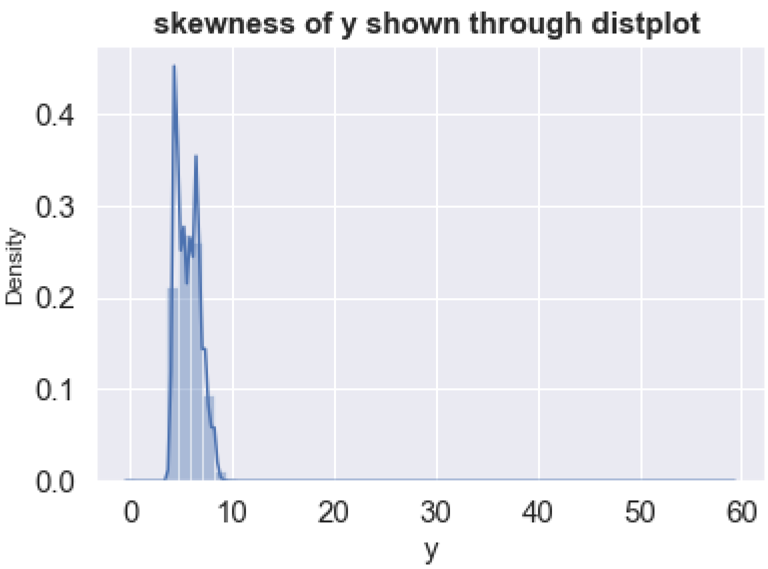


Fig 10: The distplot showing the skewness of data of ‘y’ variable

From the above distplot, we can see that variable y is positively skewed or right skewed, with median=5.71, mode= 0 and 4.35, mean=5.733. y is a bi modal variable with two modes.

6.z

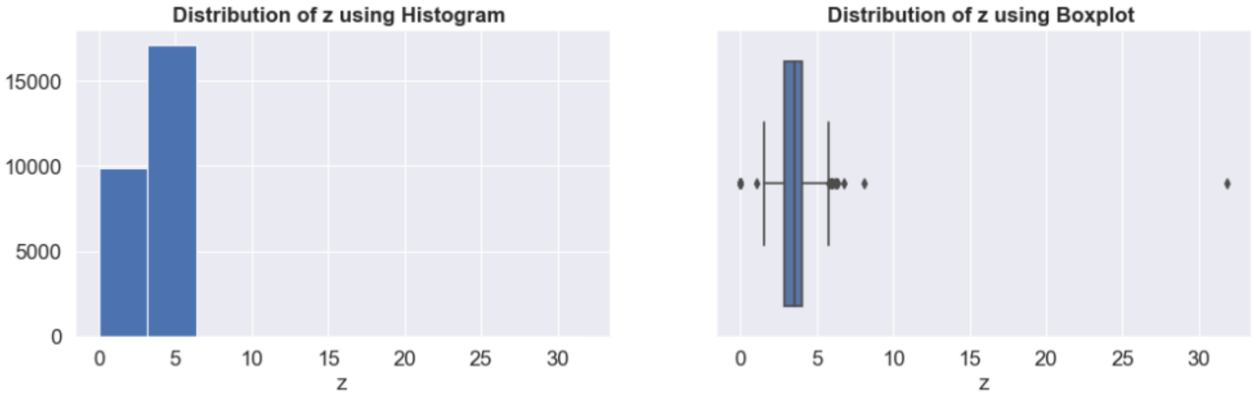


Fig 11: The box plot and the histogram showing the distribution of data of ‘z’ variable

From the univariate analysis using histogram, we see that the height of the cubic zirconia is plotted along x-axis and labelled as ‘z’. It is noticed that only two groups are formed while plotting the ‘z’ variable using histogram. Cubics having the height in the range 4 to 6 mm are the most (approximately 19500). The cubics having the length in the range of 0 to 4 mm is the least group (approximately 10000). From the bivariate analysis using boxplot, there are presence of outliers. The second quartile(Q2) or median for the x variable is about 3.52. The lower or first quartile(Q1) is about 2.9 and the upper or the third quartile(Q3) is about 4.04. The inter quartile range (IQR) for the above boxplot is 1.14

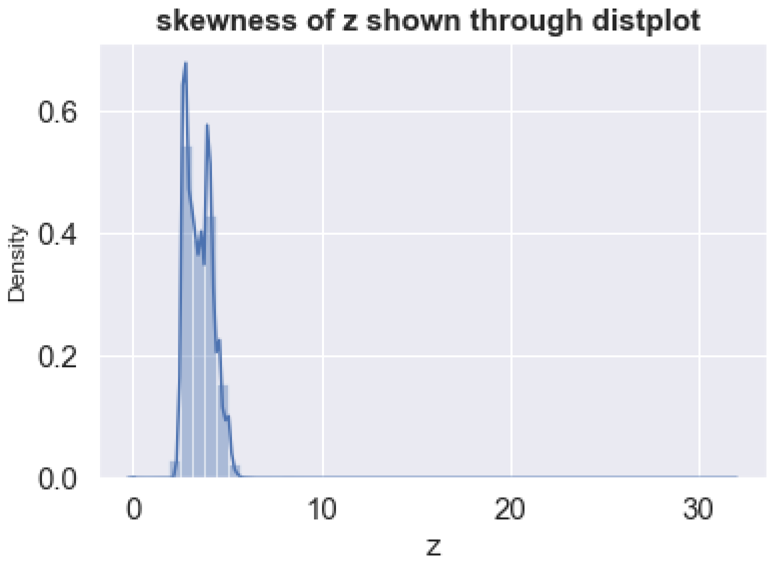


Fig 12: The distplot showing the skewness of data of ‘z’ variable

From the above distplot, we can see that the variable is positively or right skewed, median=3.52, mode=0 and 2.69, mean=3.538. z is a bi modal variable with two modes.

7.Price

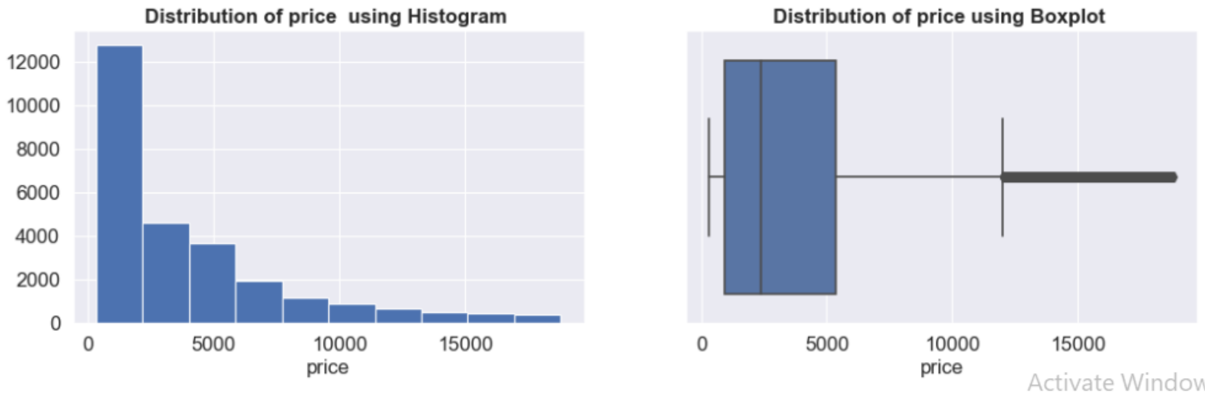


Fig 13: The box plot and the histogram showing the distribution of data of ‘‘Price” variable

From the univariate analysis using histogram, we see most cubic zirconia has the price in the range of 326-2500, approximately 13000, followed by those cubics that has price between 2500 to 4500, approximately 4500 and the least number of cubics in the range of price between 17500 to 19500, which is approximately less than 200. From the bivariate analysis using boxplot, there are presence of outliers. The second quartile(Q2) or median for the price variable is about 2375. The lower or first quartile(Q1) is about 945 and the upper or the third quartile(Q3) is about 5360. The inter quartile range (IQR) for the above boxplot is 4415.

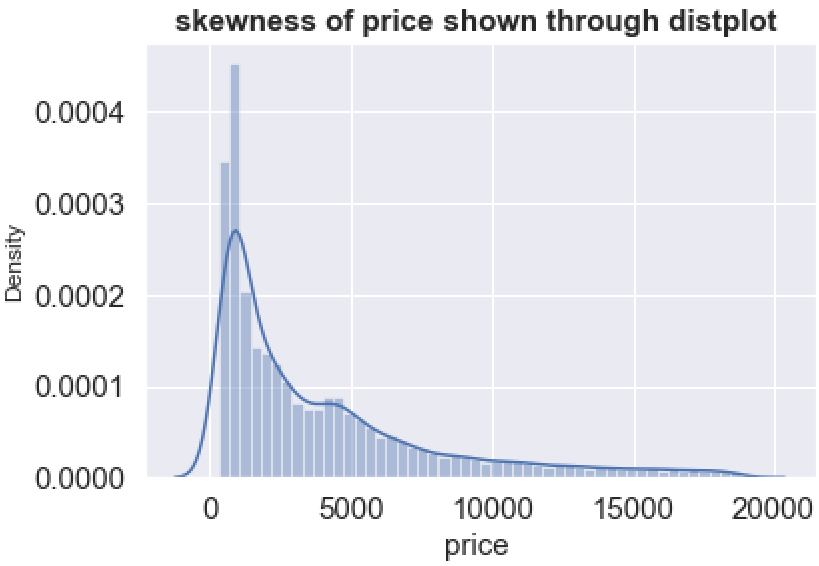


Fig 14: The distplot showing the skewness of data of 'Price 'variable

From the above distplot , we can see that **price variable** is positively skewed or right skewed, with median=2375.0,mode=0 and 544,mean=3939.5. Price is a bi modal variable with two modes.

i. Graphical Representation of univariate and bivariate analysis for categorical columns

1.Cut

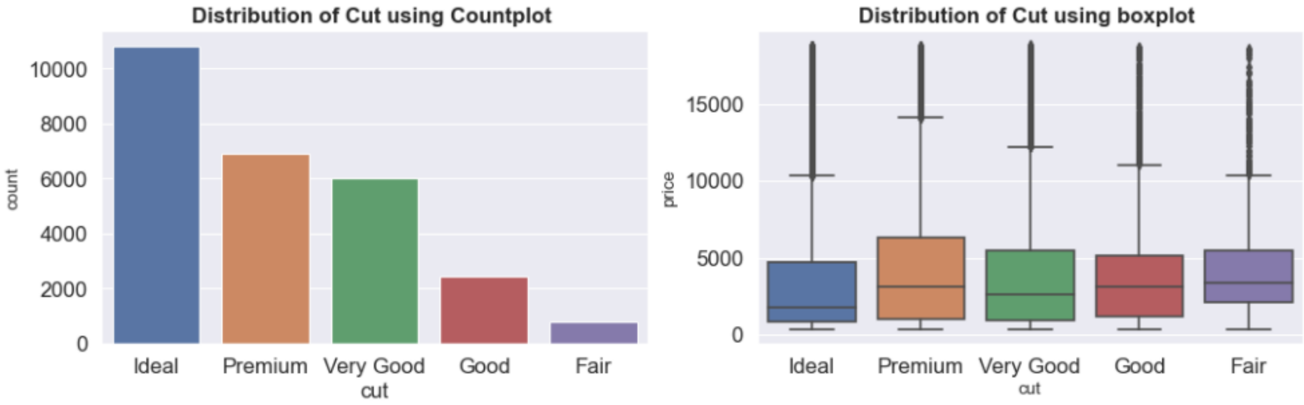


Fig 15: The box plot and the count plot showing the distribution of data of 'Cut' variable

From the univariate analysis using Count plot, we see that among all the levels of cut quality, the greatest number of cubic zirconias is having ideal cut (10816) after which comes the premium cut (6899) and the least being the fair cut (781). From the bivariate analysis using boxplot, we can see that there is presence of outliers in all levels of quality. Across the five different cut quality levels, the median of premium cut is the highest and the ideal cut has the lowest median. The lowest quartile for all the five cut quality levels are the same while the premium cut has the highest third quartile. Ideal cut has the highest number of outliers among all the cut quality levels.

2.Color

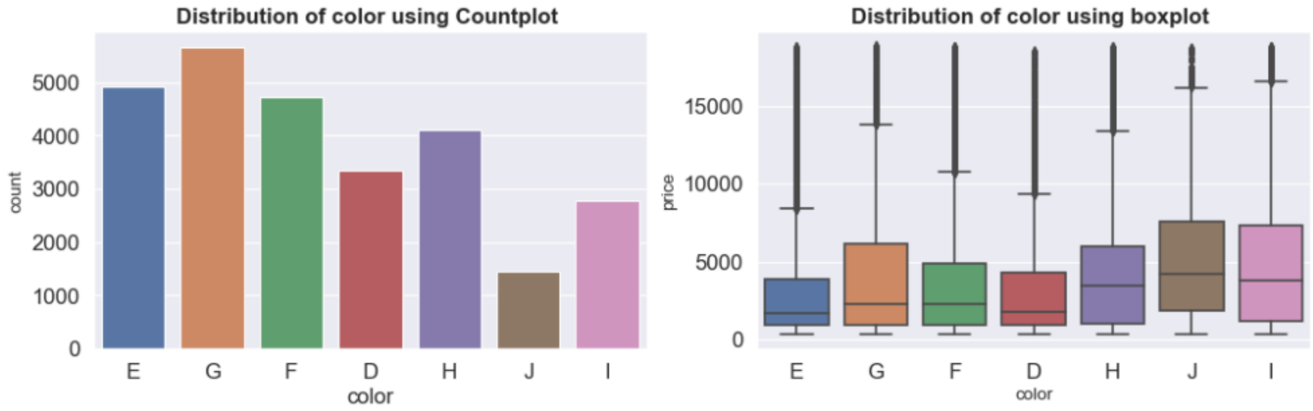


Fig 16: The box plot and the count plot showing the distribution of data of ‘color’ variable

From the univariate analysis using Count plot, we see that among all the levels of color quality here ‘D’ is considered the worst and ‘J’ is the best. The greatest number of cubic zirconias is having color ‘G’ (5661) after which comes the ‘E’ (4917) and the least being the ‘J’ (1443). From the bivariate analysis using boxplot, we can see that there is presence of outliers in all color categories. Across the five different color categorical levels, the median of ‘J’ is the highest and the “E’ has the lowest median. The lowest quartile for all the five cut quality levels are the same while the “J’ has the highest third quartile. “E”has the highest number of outliers among all the color levels.

3.Clarity

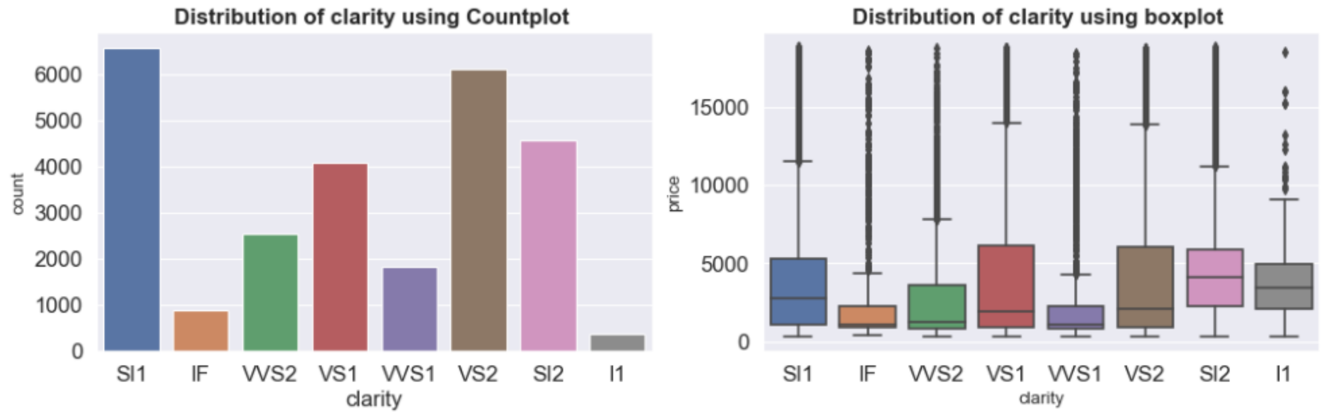


Fig 17: The box plot and the count plot showing the distribution of data of ‘Clarity’ variable

From the univariate analysis using Count plot, we see that among all the levels of clarity quality here ‘I3’ is considered the worst and ‘FL’ is the best. The greatest number of cubic zirconias is having clarity ‘SI1’ (6571) after which comes the ‘VS2’ (6099) and the least being the ‘I1’ (365), followed by “IF”, which is the second least (894). From the bivariate analysis using boxplot, we can see that there is presence of outliers in all clarity categories. Across the five different color categorical levels, the median of ‘SI2’ is the highest and the “IF’ has the lowest median. The lowest quartile for all the five cut quality levels are the same while the “VS2’ has the highest third quartile. “VVS1” has the highest number of outliers among all the color levels.

i. Multivariate analysis

i.1.Heat Map

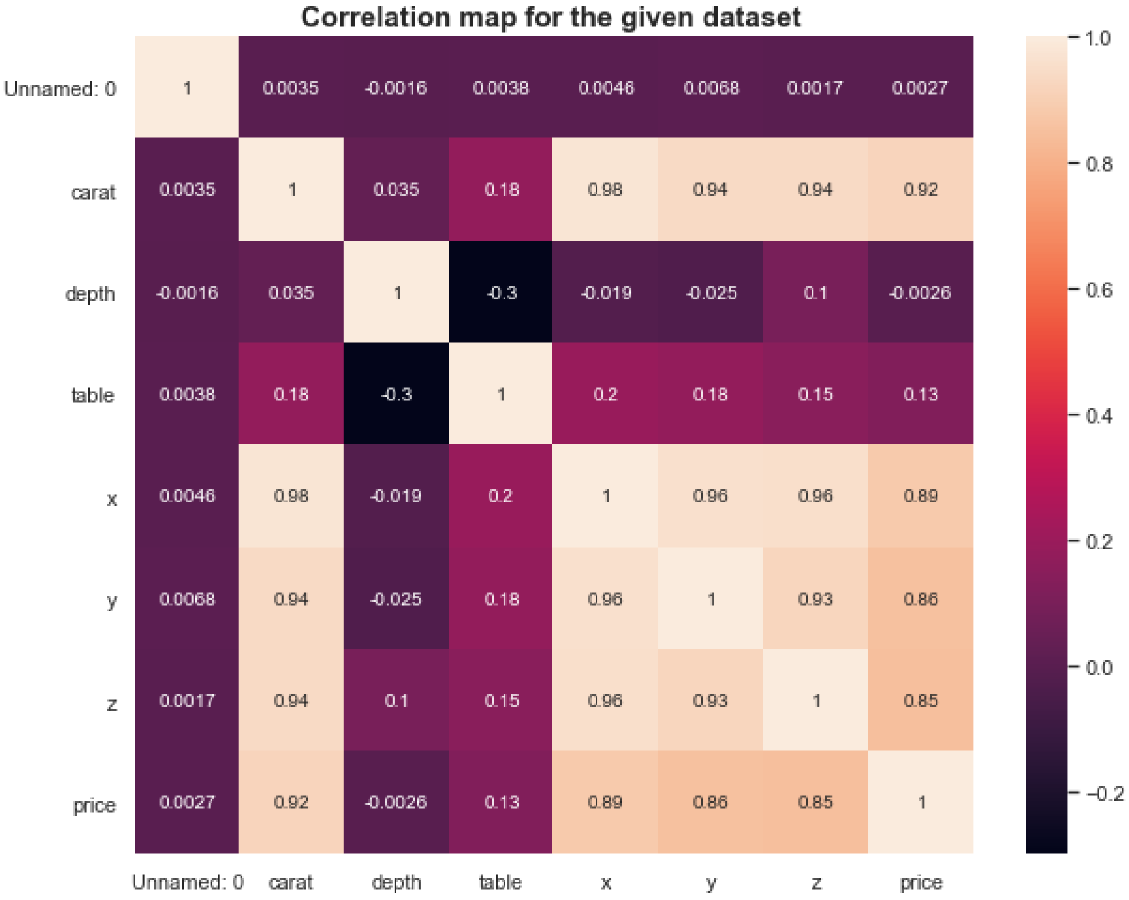
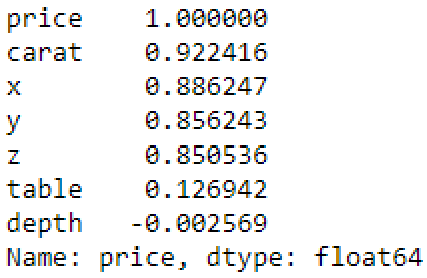


Fig 18: Heat map is portrayed for the multi variate analysis of the data.

The above heat map clearly shows the presence of multi collinearity in the dataset. It is seen that there exists high positive correlation between many components, among which ‘x’ and ‘carat’ has the highest positive correlation followed by ‘y’ and ‘z’ with ‘x’. There are many components with negative correlations among which ‘price’ and ‘depth’ has the maximum negative correlation.



```
price     1.000000
carat     0.922416
x         0.886247
y         0.856243
z         0.850536
table     0.126942
depth    -0.002569
Name: price, dtype: float64
```

TABLE 6: CORRELATION TABLE

The above table shows how each feature affects the price (target variable) of cubic zircoina. It can be inferred that most features correlate with the price of cubic zirconia. The notable exception is "depth" which has a negligible correlation (<1%).

j.2. Pair Plot

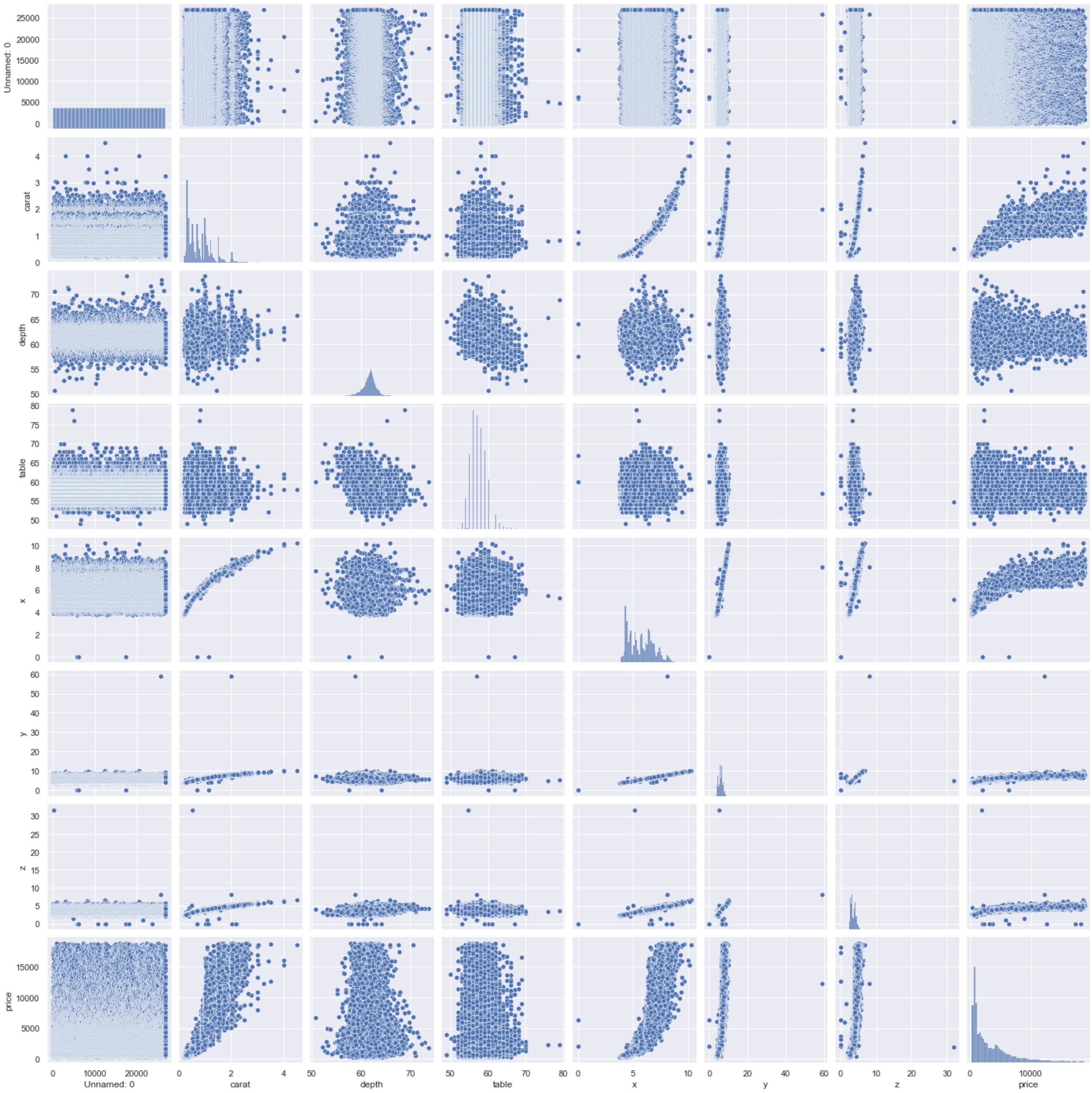
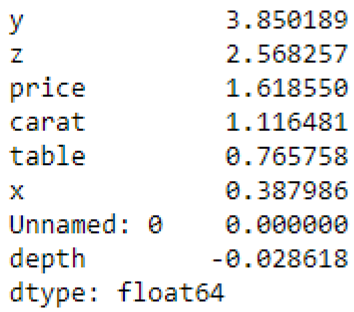


Fig 19: The pair plot showing the multivariate analysis

k. Skewness of the variables



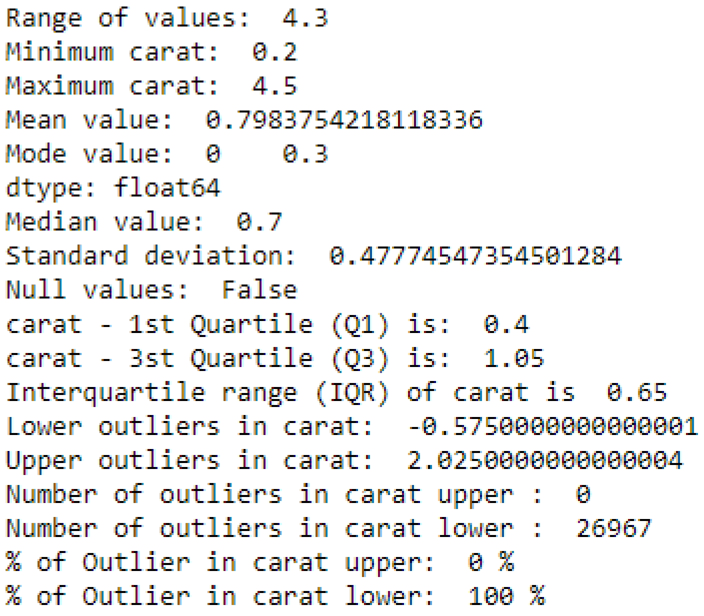
```
y             3.850189
z             2.568257
price         1.618550
carat         1.116481
table         0.765758
x             0.387986
Unnamed: 0    0.000000
depth        -0.028618
dtype: float64
```

Table 7: Skewness of all the variables

From the above given table, we can conclude that the variable ‘depth’ is negatively skewed while the rest of the variables are positively skewed. Out of the positively skewed variables, ‘y’ has the highest skewness followed by ‘z’, then by ‘price’. Among the positively skewed variables it is ‘x’ that has the least positive skewness.

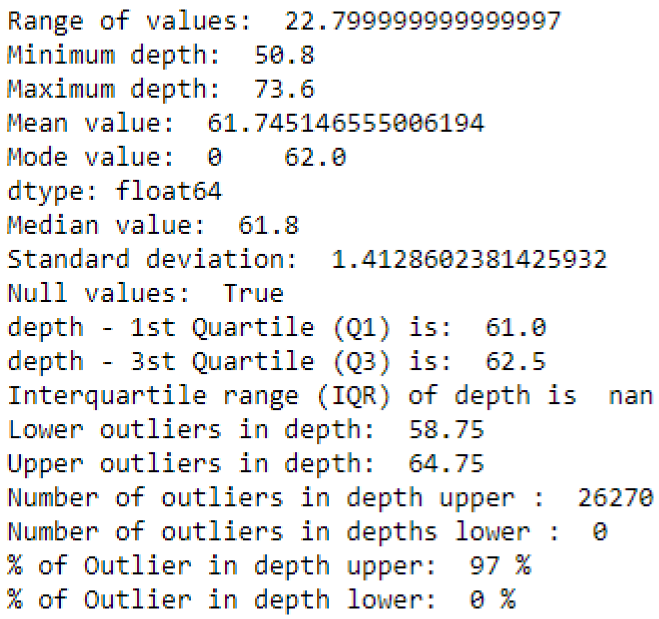
I. Outlier proportion

1.Carat



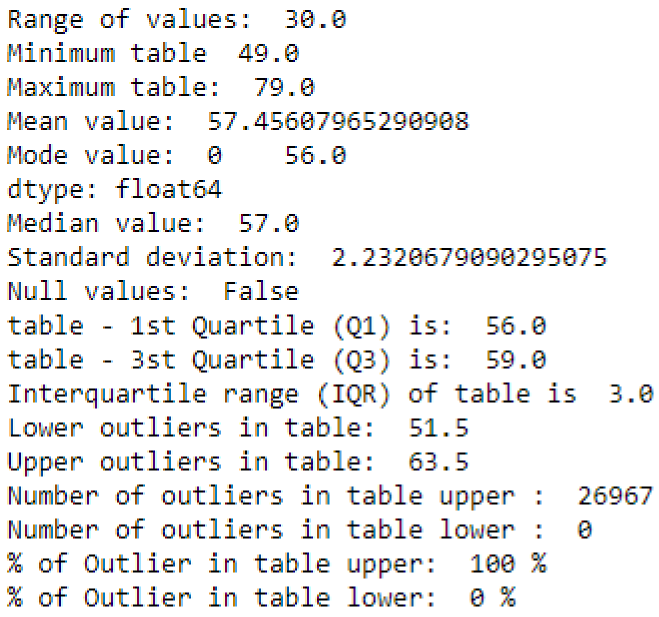
```
Range of values:  4.3
Minimum carat:  0.2
Maximum carat:  4.5
Mean value:  0.7983754218118336
Mode value:  0    0.3
dtype: float64
Median value:  0.7
Standard deviation:  0.47774547354501284
Null values:  False
carat - 1st Quartile (Q1) is:  0.4
carat - 3st Quartile (Q3) is:  1.05
Interquartile range (IQR) of carat is  0.65
Lower outliers in carat:  -0.5750000000000001
Upper outliers in carat:  2.0250000000000004
Number of outliers in carat upper :  0
Number of outliers in carat lower :  26967
% of Outlier in carat upper:  0 %
% of Outlier in carat lower:  100 %
```

2. Depth



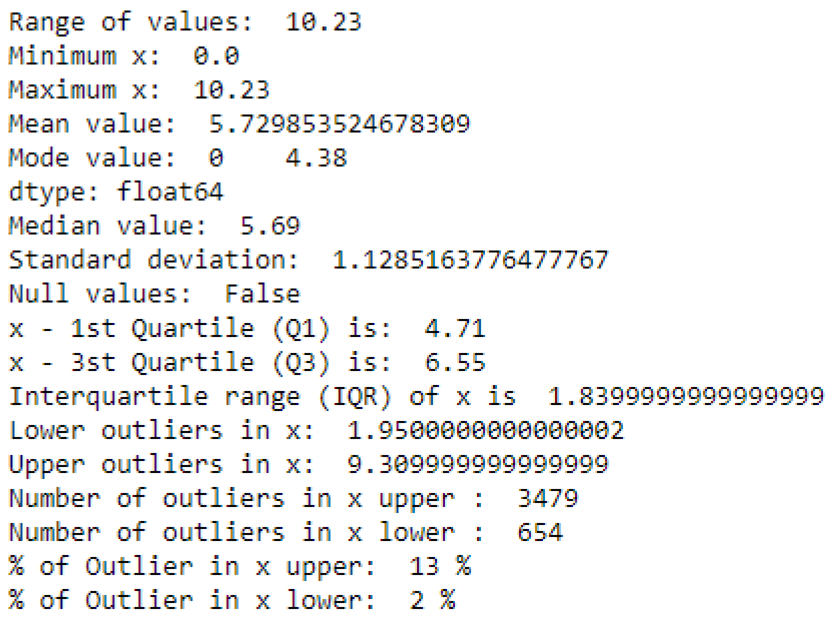
```
Range of values:  22.799999999999997
Minimum depth:  50.8
Maximum depth:  73.6
Mean value:  61.745146555006194
Mode value:  0    62.0
dtype: float64
Median value:  61.8
Standard deviation:  1.4128602381425932
Null values:  True
depth - 1st Quartile (Q1) is:  61.0
depth - 3st Quartile (Q3) is:  62.5
Interquartile range (IQR) of depth is  nan
Lower outliers in depth:  58.75
Upper outliers in depth:  64.75
Number of outliers in depth upper :  26270
Number of outliers in depths lower :  0
% of Outlier in depth upper:  97 %
% of Outlier in depth lower:  0 %
```

3. Table



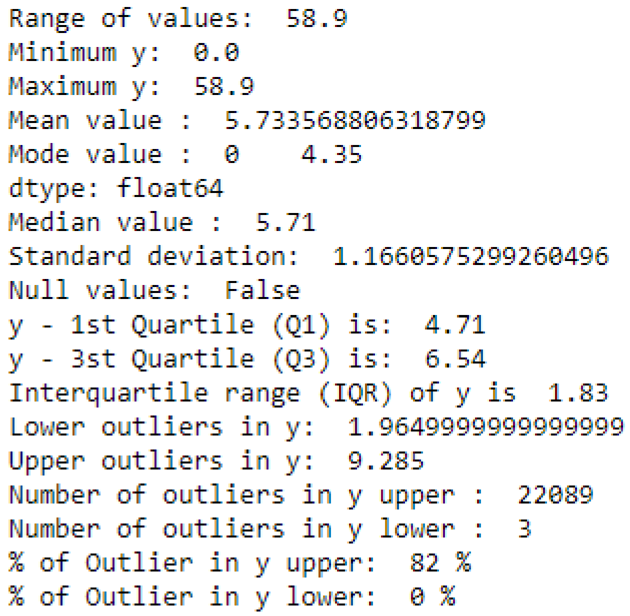
```
Range of values:  30.0
Minimum table  49.0
Maximum table:  79.0
Mean value:  57.45607965290908
Mode value:  0    56.0
dtype: float64
Median value:  57.0
Standard deviation:  2.2320679090295075
Null values:  False
table - 1st Quartile (Q1) is:  56.0
table - 3st Quartile (Q3) is:  59.0
Interquartile range (IQR) of table is  3.0
Lower outliers in table:  51.5
Upper outliers in table:  63.5
Number of outliers in table upper :  26967
Number of outliers in table lower :  0
% of Outlier in table upper:  100 %
% of Outlier in table lower:  0 %
```

4. x



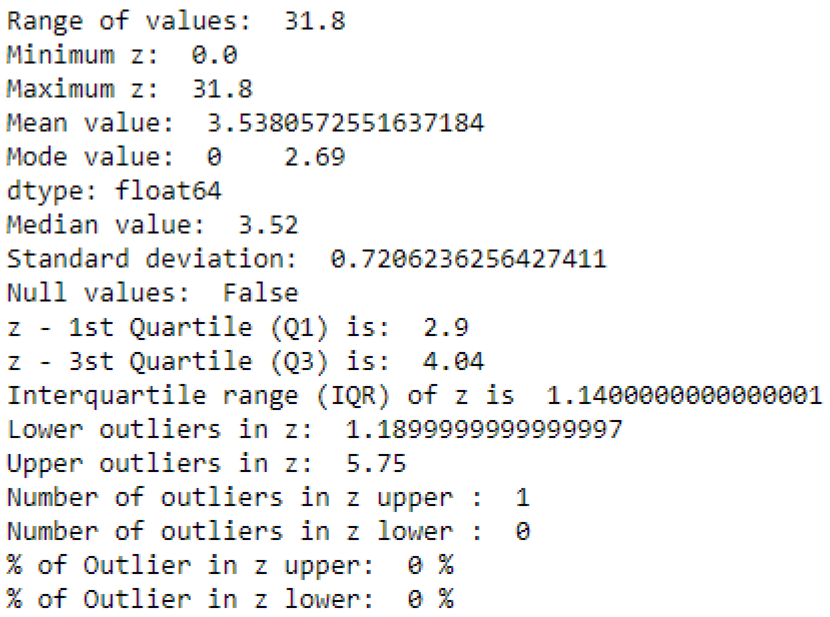
```
Range of values:  10.23
Minimum x:  0.0
Maximum x:  10.23
Mean value:  5.729853524678309
Mode value:  0    4.38
dtype: float64
Median value:  5.69
Standard deviation:  1.1285163776477767
Null values:  False
x - 1st Quartile (Q1) is:  4.71
x - 3st Quartile (Q3) is:  6.55
Interquartile range (IQR) of x is  1.8399999999999999
Lower outliers in x:  1.9500000000000002
Upper outliers in x:  9.309999999999999
Number of outliers in x upper :  3479
Number of outliers in x lower :  654
% of Outlier in x upper:  13 %
% of Outlier in x lower:  2 %
```

5.y



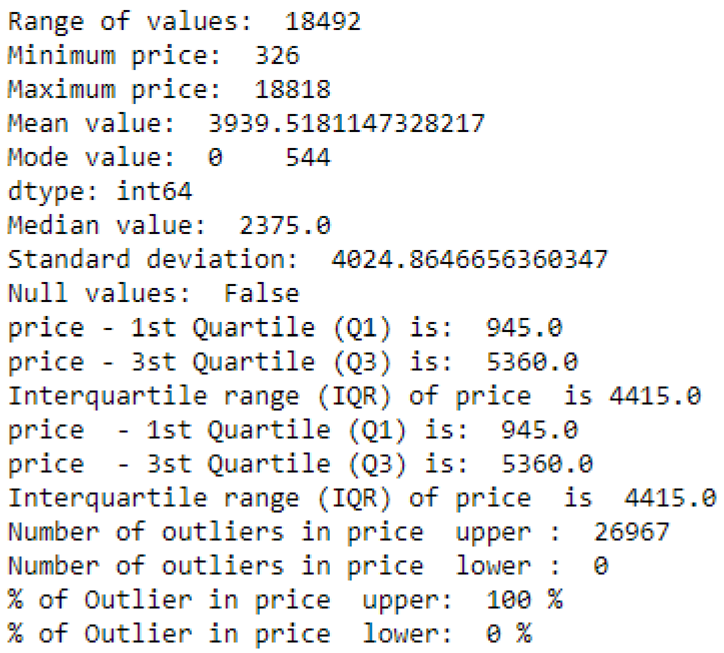
```
Range of values:  58.9
Minimum y:  0.0
Maximum y:  58.9
Mean value :  5.733568806318799
Mode value :  0    4.35
dtype: float64
Median value :  5.71
Standard deviation:  1.1660575299260496
Null values:  False
y - 1st Quartile (Q1) is:  4.71
y - 3st Quartile (Q3) is:  6.54
Interquartile range (IQR) of y is  1.83
Lower outliers in y:  1.9649999999999999
Upper outliers in y:  9.285
Number of outliers in y upper :  22089
Number of outliers in y lower :  3
% of Outlier in y upper:  82 %
% of Outlier in y lower:  0 %
```

6.z



```
Range of values:  31.8
Minimum z:  0.0
Maximum z:  31.8
Mean value:  3.5380572551637184
Mode value:  0    2.69
dtype: float64
Median value:  3.52
Standard deviation:  0.7206236256427411
Null values:  False
z - 1st Quartile (Q1) is:  2.9
z - 3st Quartile (Q3) is:  4.04
Interquartile range (IQR) of z is  1.1400000000000001
Lower outliers in z:  1.1899999999999997
Upper outliers in z:  5.75
Number of outliers in z upper :  1
Number of outliers in z lower :  0
% of Outlier in z upper:  0 %
% of Outlier in z lower:  0 %
```

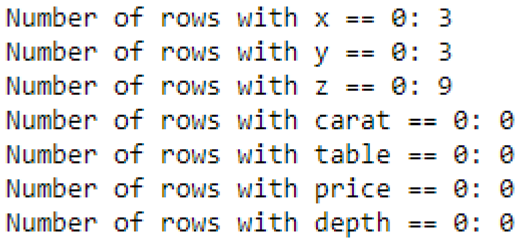
7.Price



```
Range of values:  18492
Minimum price:  326
Maximum price:  18818
Mean value:  3939.5181147328217
Mode value:  0    544
dtype: int64
Median value:  2375.0
Standard deviation:  4024.8646656360347
Null values:  False
price - 1st Quartile (Q1) is:  945.0
price - 3st Quartile (Q3) is:  5360.0
Interquartile range (IQR) of price  is 4415.0
price  - 1st Quartile (Q1) is:  945.0
price  - 3st Quartile (Q3) is:  5360.0
Interquartile range (IQR) of price  is  4415.0
Number of outliers in price  upper :  26967
Number of outliers in price  lower :  0
% of Outlier in price  upper:  100 %
% of Outlier in price  lower:  0 %
```

Q1.2: **Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Check for the possibility of combining the sub levels of an ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning.**

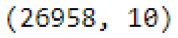
a. Checking for values equal to zero



```
Number of rows with x == 0: 3
Number of rows with y == 0: 3
Number of rows with z == 0: 9
Number of rows with carat == 0: 0
Number of rows with table == 0: 0
Number of rows with price == 0: 0
Number of rows with depth == 0: 0
```

Table 8: Checking for values equal to 0

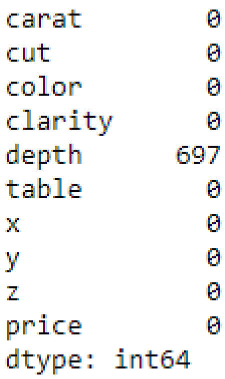
Table above shows the rows which are having the x, y and z values zero. x, y and z values in the given dataset corresponds to length, width and height of the cubic zirconia. Taking into consideration cubic zirconia as a crystal, a crystal is not possible without any of the three parameters i.e, height, length or width. So, the above rows are considered as fault inputs and they have the dropped from the dataset as these rows do not convey any meaningful information in the context of our dataset.



```
(26958, 10)
```

After dropping the rows with x, y, z values equal to zero, the dimension of the given dataset changes and new dimension of the data set is having 26958 rows and 10 columns.

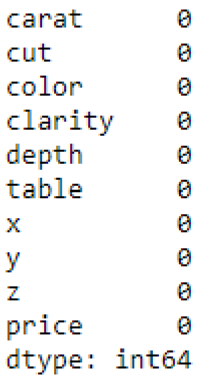
b. Imputing the null values



```
carat         0
cut           0
color         0
clarity       0
depth       697
table         0
x             0
y             0
z             0
price         0
dtype: int64
```

Table: Checking for the null values in the dataset

As we can see from the above table, the 'depth' variable is having 697 null values and these values needs to be imputed before we proceed for the linear regression model. So, we have imputed the null values with the mean of the 'depth' column.

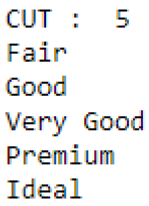


```
carat       0
cut         0
color       0
clarity     0
depth       0
table       0
x           0
y           0
z           0
price       0
dtype: int64
```

Table 9: Checking for the null values in the dataset after imputing the null values.

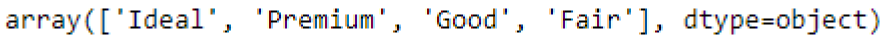
As we see in the above table, after imputing the null values present in the 'depth' variable, now the data set does not have any null values.

c. Combining the sub levels of ordinal variables.



```
CUT :  5
Fair
Good
Very Good
Premium
Ideal
```

The above table shows the sub levels of the ordinal variable 'cut'. This ordinal variable describes the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal.When we look carefully into it, we can see that bot the sub levels, 'Good' and 'Very Good' convey the same meaning. So, both these sub levels are combined into one single level.



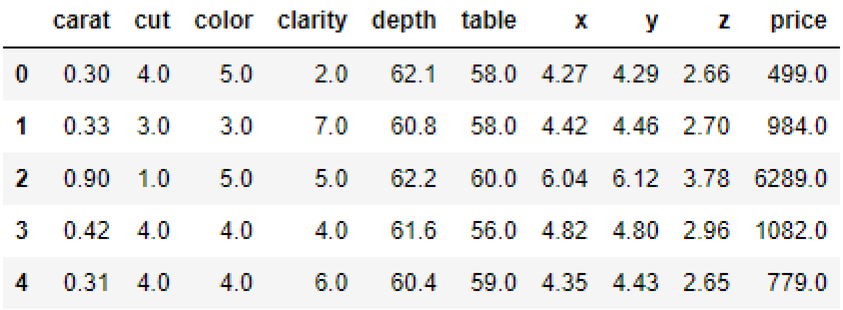
```
array(['Ideal', 'Premium', 'Good', 'Fair'], dtype=object)
```

The above array shows the list of sub levels present in the ordinal variable 'cut' after combining the sub levels- 'Good' and "Very Good'.

Q1.3: 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 stats model. Create multiple models and check the performance of Predictions on Train and Test sets using R square,

RMSE & Adj R square. Compare these models and select the best one with appropriate reasoning.

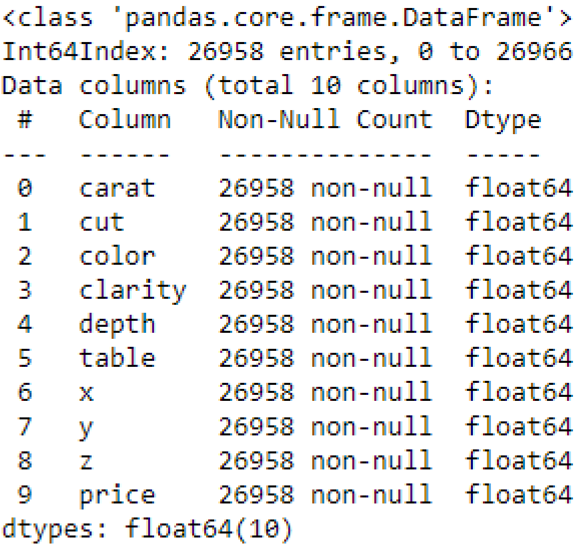
Since the variables in the data set is having outliers, it has been treated. The columns in the data having the string values have been encoded using the label encoding technique.



	carat	cut	color	clarity	depth	table	x	y	z	price
0	0.30	4.0	5.0	2.0	62.1	58.0	4.27	4.29	2.66	499.0
1	0.33	3.0	3.0	7.0	60.8	58.0	4.42	4.46	2.70	984.0
2	0.90	1.0	5.0	5.0	62.2	60.0	6.04	6.12	3.78	6289.0
3	0.42	4.0	4.0	4.0	61.6	56.0	4.82	4.80	2.96	1082.0
4	0.31	4.0	4.0	6.0	60.4	59.0	4.35	4.43	2.65	779.0

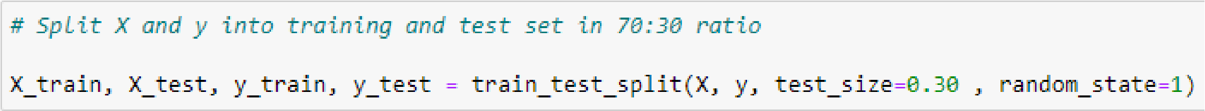
Table 10: Head of the dataset after the data encoding.

After data encoding, all the variables of object data types have been converted into int data types, before modelling. Linear regression model does not take categorical values so that we have encoded categorical values to integer for better results.



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 26958 entries, 0 to 26966
Data columns (total 10 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   carat     26958 non-null  float64
 1   cut       26958 non-null  float64
 2   color     26958 non-null  float64
 3   clarity   26958 non-null  float64
 4   depth     26958 non-null  float64
 5   table     26958 non-null  float64
 6   x         26958 non-null  float64
 7   y         26958 non-null  float64
 8   z         26958 non-null  float64
 9   price     26958 non-null  float64
dtypes: float64(10)
```

Table 11: Information about given data set after data type conversion.



```
# Split X and y into training and test set in 70:30 ratio

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30 , random_state=1)
```

The data has been splitted into train and test data in the ratio 70:30 after dropping the predictor variable ‘price’.

- ROOT MEAN SQUARE ERROR(RMSE)

Root Mean Square Error (RMSE) is a standard way to measure the error of a model in predicting quantitative data. RMSE is most useful when large errors are particularly undesirable. RMSE values between 0.2 and 0.5 shows that the model can relatively predict the data accurately. If the RMSE for the test set is much higher than that of the training set, it is likely that the data has been badly overfit i.e. the model tests well in sample, but has little predictive value when tested out of sample.

- VARIANCE INFLATION FACTOR(VIF)

Variance Inflation Factor is the measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable. VIF is used to detect the severity of multicollinearity in the ordinary least square (OLS) regression analysis. Values of VIF that exceed 10 are often regarded as indicating multicollinearity. A high VIF indicates that the associated independent variable is highly collinear with other variables in the model. The variance inflation factor (VIF) identifies correlation between independent variables and the strength of that correlation

4 Linear regression models were created using different techniques such as standard scaling, label encoding and outlier treatment. Below a table is provided with the observations that is found in each of the model.

	TECHNIQUES	R^2	RMSE	Adj R^2
MODEL 1.	Outlier treatment done	0.932	941.08	0.932
MODEL 2.	Without outlier treatment	0.910	1278.76	0.910
MODEL 3.	With scaling and with outlier treatment	0.932	0.271	0.932
MODEL 4.	Dropping the ‘depth’ variable	0.932	941.08	0.932

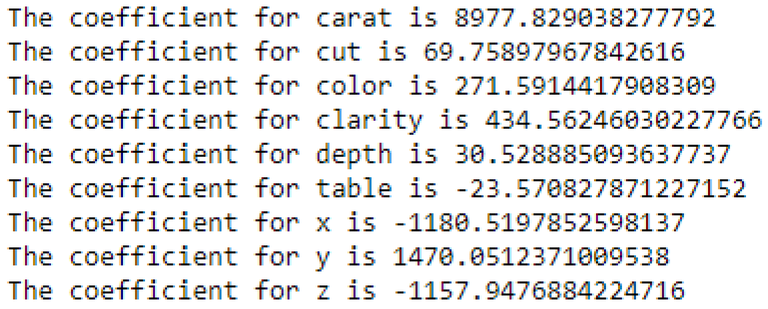
Table 12: Table with different linear regression models and their observations.

Out of the above 4 linear regression models built, I will choose my 1st model, where the outliers are treated as my best model.

- LINEAR REGRESSION USING SCIKIT LEARN

- **OBSERVATIONS**

1. COEFFICIENT TABLE OF INDEPENDENT ATTRIBUTES



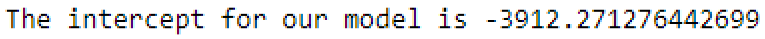
```
The coefficient for carat is 8977.829038277792
The coefficient for cut is 69.75897967842616
The coefficient for color is 271.5914417908309
The coefficient for clarity is 434.56246030227766
The coefficient for depth is 30.528885093637737
The coefficient for table is -23.570827871227152
The coefficient for x is -1180.5197852598137
The coefficient for y is 1470.0512371009538
The coefficient for z is -1157.9476884224716
```

Table 13: **COEFFICIENT TABLE OF INDEPENDENT ATTRIBUTES**

From the above table, we can conclude
The one unit increase in carat increases price by 8977.829038277792.
The one unit increase in cut increases price by 69.75897967842616.
The one unit increase in color increases price by 271.5914417908309.
The one unit increase in clarity increases price by 434.56246030227766.
The one unit increase in y increases price by 1470.0512371009538.
The one unit increase in depth increases price by 30.52888509363773.

But The one unit increase in table decreases price by -23.570827871227152.
The one unit increase in x decreases price by -1180.5197852598137.
The one unit increase in z decreases price by -1157.9476884224716.

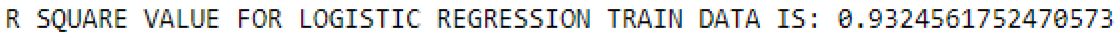
2. INTERCEPT OF THE MODEL



```
The intercept for our model is -3912.271276442699
```

The intercept (often labelled the constant) is the expected mean value of Y when all X=0. If X never equals 0, then the intercept has no intrinsic meaning. The intercept for our model is -3912.271276442699. In present case when the other predictor variable are zero i.e like carat, cut, color, clarity all are zero then the C=-3912. (Y = m1X1 + m2X2+ ….. + mnXn + C + e) that means price is -3912. which is meaningless. We can do Z score or scaling the data and make it nearly zero.

3. R SQUARE VALUE



```
R SQUARE VALUE FOR LOGISTIC REGRESSION TRAIN DATA IS: 0.9324561752470573
```

93% of the variation in the price is explained by the predictors in the model for train set

```
R SQUARE VALUE FOR LOGISTIC REGRESSION TEST DATA IS: 0.9278045054662835
```

- R-square is the percentage of the response variable variation that is explained by a linear model. R-squared is always between 0 and 100%:
- 0% indicates that the model explains none of the variability of the response data around its mean.
- 100% indicates that the model explains all the variability of the response data around its mean.
- Model score - R2 or coeff of determinant.
- R^2=1–RSS / TSS = RegErr /TSS or R-square = Explained variation / Total variation.
- In this regression model we can see the R-square value on Training and Test data respectively 0.9324561752470573 and 0. 0.9278045054662835.

4. RMSE

```
RMSE FOR LOGISTIC REGRESSION TRAIN DATA IS: 897.9583767860294
```

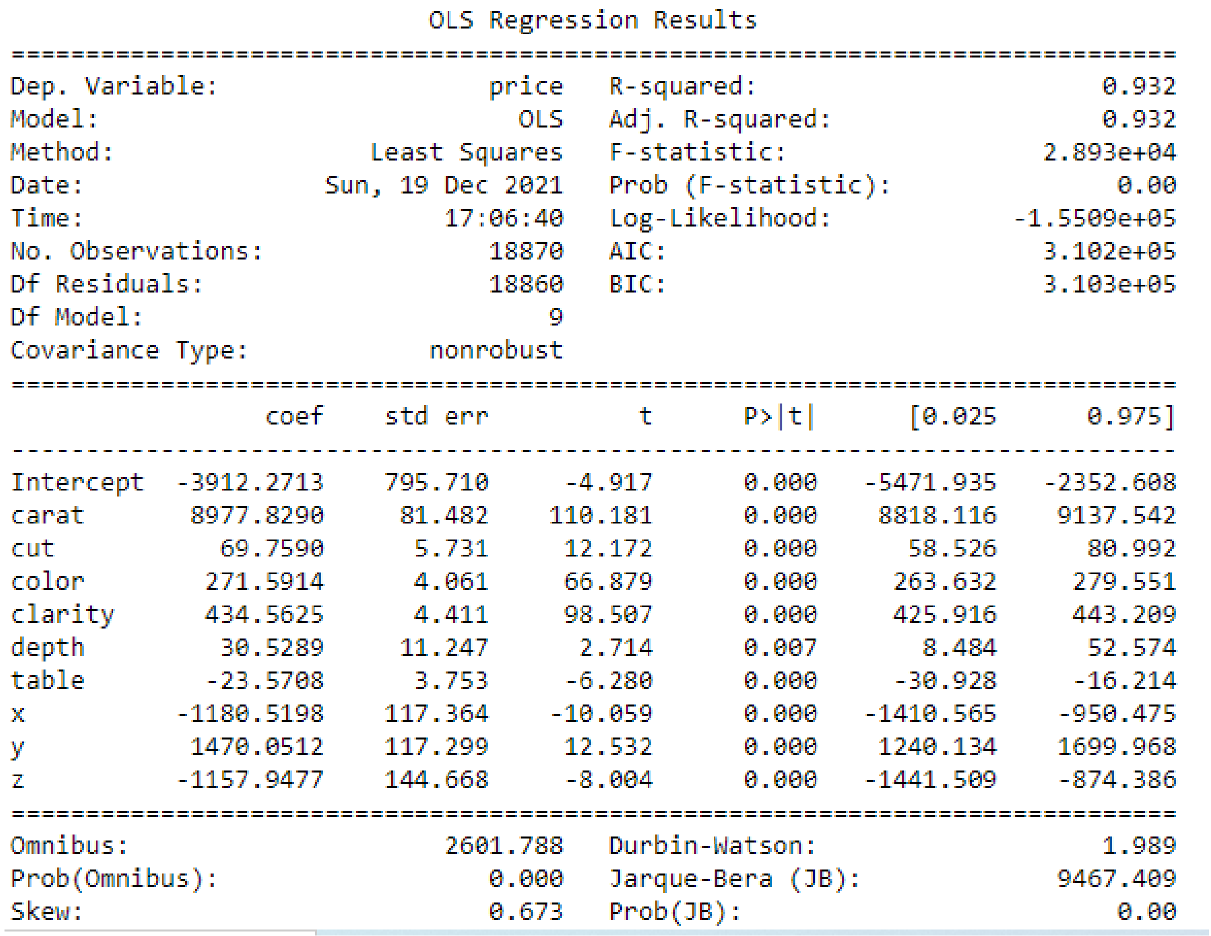
```
RMSE FOR LOGISTIC REGRESSION TEST DATA IS: 941.0889874646965
```

➢ LINEAR REGRESSION USING STATS MODEL

- **OBSERVATIONS**

R^2 is not a reliable metric as it always increases with addition of more attributes even if the attributes have no influence on the predicted variable. Instead we use adjusted R^2 which removes the statistical chance that improves R^2.Scikit does not provide a facility for adjusted R^2, so we use stats model, a library that gives results similar to what we obtain in R language. This library expects the X and Y to be given in one single data frame. Adjusted R-squared metric accounts for the spurious correlations. It decreases when we include attributes into the model that are weak or poor predictors of Y

1. OLS TABLE

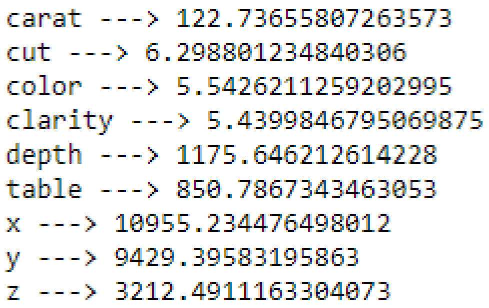


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.932
Model:                            OLS   Adj. R-squared:                  0.932
Method:                 Least Squares   F-statistic:                 2.893e+04
Date:                Sun, 19 Dec 2021   Prob (F-statistic):               0.00
Time:                        17:06:40   Log-Likelihood:            -1.5509e+05
No. Observations:               18870   AIC:                         3.102e+05
Df Residuals:                   18860   BIC:                         3.103e+05
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -3912.2713    795.710     -4.917      0.000   -5471.935   -2352.608
carat       8977.8290     81.482    110.181      0.000    8818.116    9137.542
cut           69.7590      5.731     12.172      0.000      58.526      80.992
color        271.5914      4.061     66.879      0.000     263.632     279.551
clarity      434.5625      4.411     98.507      0.000     425.916     443.209
depth         30.5289     11.247      2.714      0.007       8.484      52.574
table        -23.5708      3.753     -6.280      0.000     -30.928     -16.214
x          -1180.5198    117.364    -10.059      0.000   -1410.565    -950.475
y           1470.0512    117.299     12.532      0.000    1240.134    1699.968
z          -1157.9477    144.668     -8.004      0.000   -1441.509    -874.386
==============================================================================
Omnibus:                     2601.788   Durbin-Watson:                   1.989
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             9467.409
Skew:                           0.673   Prob(JB):                         0.00
```

Table 14: OLS REPORT ON MODEL DATA

From the above ols report, we can see that both the r-squared and Adj R-squared is having the same value-0.932 which implies that the data has been fitted properly into the model we created. P is the conditional probability given H0 is true. Attribute ‘depth’ have P value > 0.05 and hence statistically their coefficients are not reliable. Overall, model P value is lower than 0.05 which means model is reliable after eliminating the useless attributes such as depth.

2. VIF



```
carat ---> 122.73655807263573
cut ---> 6.298801234840306
color ---> 5.5426211259202995
clarity ---> 5.4399846795069875
depth ---> 1175.646212614228
table ---> 850.7867343463053
x ---> 10955.234476498012
y ---> 9429.39583195863
z ---> 3212.4911163304073
```

We can observe there are very strong multi collinearity present in the data set. VIFs start at 1 and have no upper limit a. A value of 1 indicates that there is no correlation between this independent variable and any others b. VIFs between 1 and 5 suggest that there is a moderate correlation, but it is not severe enough to warrant corrective measures c. VIFs greater than 5 represent critical levels of multicollinearity where the coefficients are poorly estimated, and the p-values are questionable.

LINEAR REGRESSION CONCLUSION:

1. intercept for the model: -3912.271276442699
2. R square on training data: 0.9324561752470573
3. R square on testing data: 0.9278045054662835
4. RMSE on Training data: 897.9583767860294
5. RMSE on Testing data: 941.0889874646965

As the training data & testing data score are almost inline, we can conclude this model is a Right-Fit Model.

Q1.4: Inference: Basis on these predictions, what are the business insights and recommendations. Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

We had a business problem to predict the price of the stone and provide insights for the company on the profits on different prize slots. From the EDA analysis we could understand the carat and cut had number profits to the company. The predictions were able to capture 93% variations in the price and it is explained by the predictors in the training set. Using stats model if we could run the model again we can have P values and coefficients which will give us better understanding of the relationship, so that values more 0.05 we can drop those variables and re run the model again for better results.

The equation,

```
(-3912.27) * Intercept + (8977.83) * carat + (69.76) * cut + (271.59) *
color + (434.56) * clarity + (30.53) * depth + (-23.57) * table + (-118
0.52) * x + (1470.05) * y + (-1157.95) * z +
```

Recommendations

1. The carat is the most important attribute and they are the ones which are bringing profits so that we could use marketing based on carat to bring in more profits.

2. The cut of the diamond is the next important attributes. The more finesse the cut of the diamond is, the stone becomes more attractive and its features sharpens thereby attracting the customers and increases the profit.

The best attributes are

1. Carat
2. Cut
3. Color
4. clarity

HOLIDAY PACKAGE

CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

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

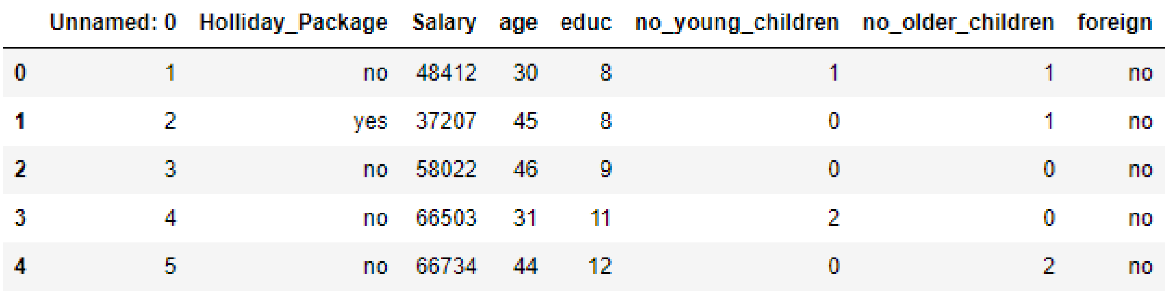
The purpose of this whole exercise is to perform logistic regression and linear discriminant analysis on the given data and evaluate which technique forms the best model for the given dataset. Logistic Regression is a technique usually used for Binary Classification problems. It is used when the data is linearly separable and the outcome is binary or dichotomous in nature. Binary classification refers to predicting the output variable that is discrete in two classes. Linear Discriminant Analysis (LDA) on the other hand is a linear model for classification and dimensionality reduction. LDA is primarily used to reduce the number of features to a more manageable number, thereby reduce the dimensionality before classification.

DATA DESCRIPTION

1. Holiday_Package: Opted for Holiday Package yes/no?
2. Salary: Employee salary
3. Type: Type of tour insurance firms
4. Age: Age in years
5. edu: Years of formal education
6. no_young_children: The number of young children (younger than 7 years)
7. no_older_children: Number of older children
8. foreign: foreigner Yes/No

Q2.1. 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.

a. Sample Dataset.

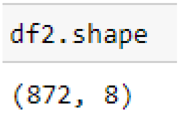


	Unnamed: 0	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	1	no	48412	30	8	1	1	no
1	2	yes	37207	45	8	0	1	no
2	3	no	58022	46	9	0	0	no
3	4	no	66503	31	11	2	0	no
4	5	no	66734	44	12	0	2	no

TABLE 1: Dataset Sample

The above table shows the head of the given dataset, i.e. the first five entries to ensure that the dataset has been loaded without any issues. The data consists of details of 872 employees along with another attributes arranged in 8 columns such as salary, age, edu, foreign, no_young_children , no_older_children and opted for a holliday package or not.

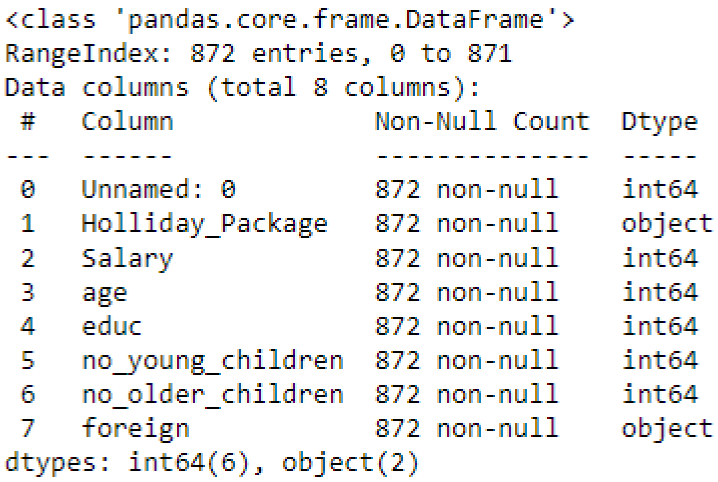
a. Shape of the dataset



```
df2.shape

(872, 8)
```

There are total 872 rows and 8 columns in the dataset.

b. Checking for the missing values in the dataset.

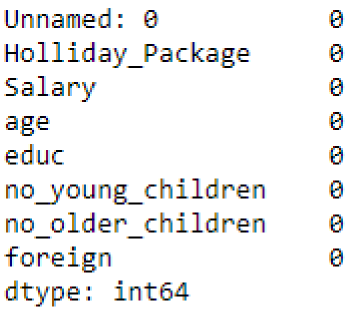


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   Unnamed: 0         872 non-null    int64 
 1   Holliday_Package   872 non-null    object
 2   Salary             872 non-null    int64 
 3   age                872 non-null    int64 
 4   educ               872 non-null    int64 
 5   no_young_children  872 non-null    int64 
 6   no_older_children  872 non-null    int64 
 7   foreign            872 non-null    object
dtypes: int64(6), object(2)
```

Table 2. Checking for the missing values in the dataset

From the above results we can see that there is no missing value present in the dataset.

c. Checking for the null values in the given dataset.

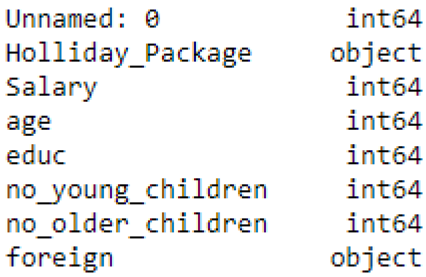


```
Unnamed: 0           0
Holliday_Package     0
Salary               0
age                  0
educ                 0
no_young_children    0
no_older_children    0
foreign              0
dtype: int64
```

Table 3. Checking for the null values in the dataset

From the above table, we can see that there are no null values in the given data set.

d. Checking for the datatypes of variables present in the dataset.

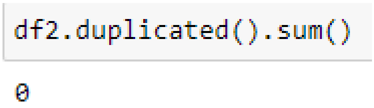


```
Unnamed: 0             int64
Holliday_Package      object
Salary                 int64
age                    int64
educ                   int64
no_young_children      int64
no_older_children      int64
foreign               object
```

Table 4. Checking the data type of variables in the data set

Out of 8, there are 6 variables of int data type and rest 2 variables of object data type.

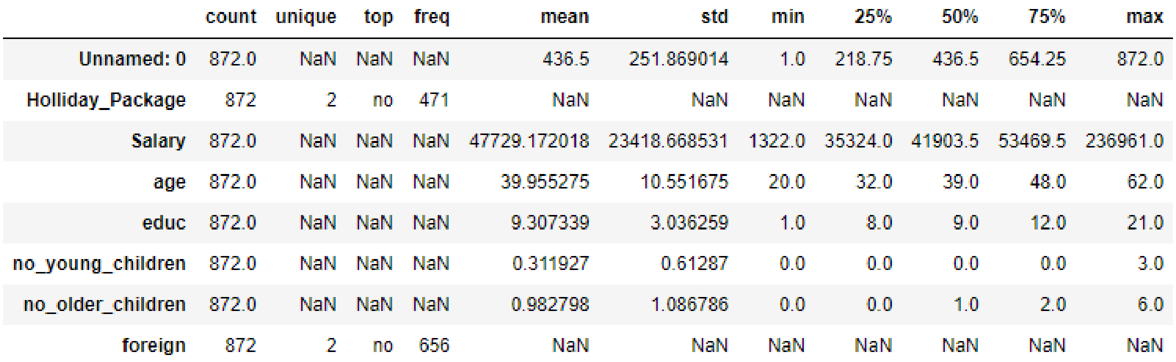
Checking for the number of duplicated values



```
df2.duplicated().sum()

0
```

Here, we can see that there are no duplicated values in the given dataset.

Checking for the summary of the given dataset.



	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Unnamed: 0	872.0	NaN	NaN	NaN	436.5	251.869014	1.0	218.75	436.5	654.25	872.0
Holliday_Package	872	2	no	471	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	236961.0
age	872.0	NaN	NaN	NaN	39.955275	10.551675	20.0	32.0	39.0	48.0	62.0
educ	872.0	NaN	NaN	NaN	9.307339	3.036259	1.0	8.0	9.0	12.0	21.0
no_young_children	872.0	NaN	NaN	NaN	0.311927	0.61287	0.0	0.0	0.0	0.0	3.0
no_older_children	872.0	NaN	NaN	NaN	0.982798	1.086786	0.0	0.0	1.0	2.0	6.0
foreign	872	2	no	656	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 5: The above table shows the summary statistics of the given dataset.

From the above table, we can see that we have both categorical and continuous data. For categorical data we have Holliday_Package and foreign and for continuous data we have no_young_children, no_older_children, salary, age and educ. Holliday_package will be target variable. From the summary of the dataset, we can see that mean of the salary is the highest while the mean for no_young_children is least among all. The variable age has the second highest mean followed by the variable educ. Similarly, salary has the highest standard deviation whereas no_young_children has the lowest standard deviation. Among all variables, no_young_children have the least value for all the statistical measurements.

h. Graphical Representation of univariate and bivariate analysis for continuous columns

h.1. Salary

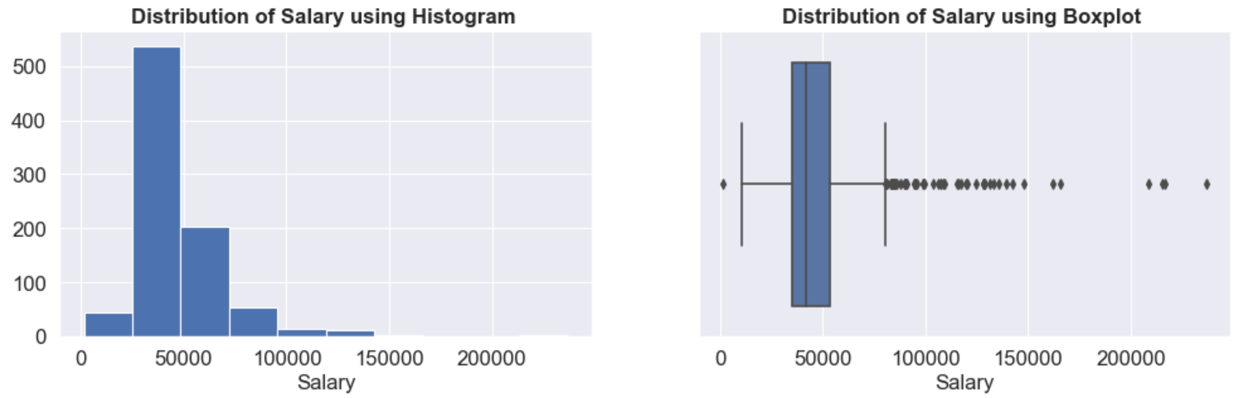


Fig 1: The box plot and the histogram showing the distribution of data of 'Salary' variable

From the univariate analysis using histogram, we see most of the employee salary is between 25000 and 50000, approximately 580 employees, followed by those employees that are having salary between 50000 to 75000, 200 employees and the least number of employees have their salary in the range of 125000 to 145000, which is less than 2 employees. From the bivariate analysis using boxplot, there are presence of outliers. The second quartile(Q2) or median for the salary variable is about 41903.5. The lower or first quartile(Q1) is about 35324 and the upper or the third quartile(Q3) is about 53469.5. The inter quartile range (IQR) for the above boxplot is 18145.5

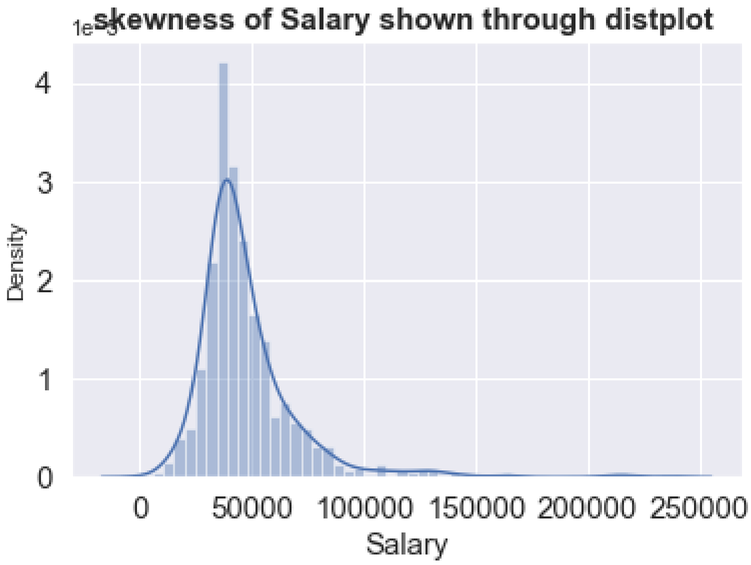


Fig 2: The distplot showing the skewness of data of 'Salary' variable

From the above distplot, we can see that Salary is highly positively skewed or right skewed, with median=41903.5; mean= 47729.17. From this it is clear that Age variable is bimodal with two mode values.

h.2. Age

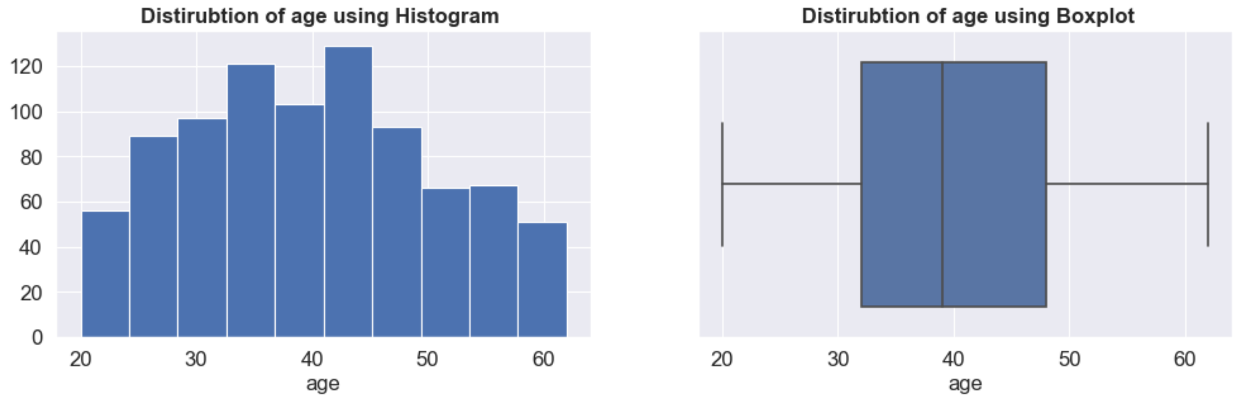


Fig 3: The box plot and the histogram showing the distribution of data of 'Age' variable

From the univariate analysis using histogram, we see that age of most of the employees are between 42 and 45, approximately 136 employees, followed by those employees who are in the age group between 33 to 37, 120 employees and the least number of employees are in the age group between 58 to 62, which is approximately 56 employees. From the bivariate analysis using boxplot, there is no presence of outliers. The second quartile(Q2) or median for the salary variable is about 39. The lower or first quartile(Q1) is about 32 and the upper or the third quartile(Q3) is about 48. The inter quartile range (IQR) for the above boxplot is 16.

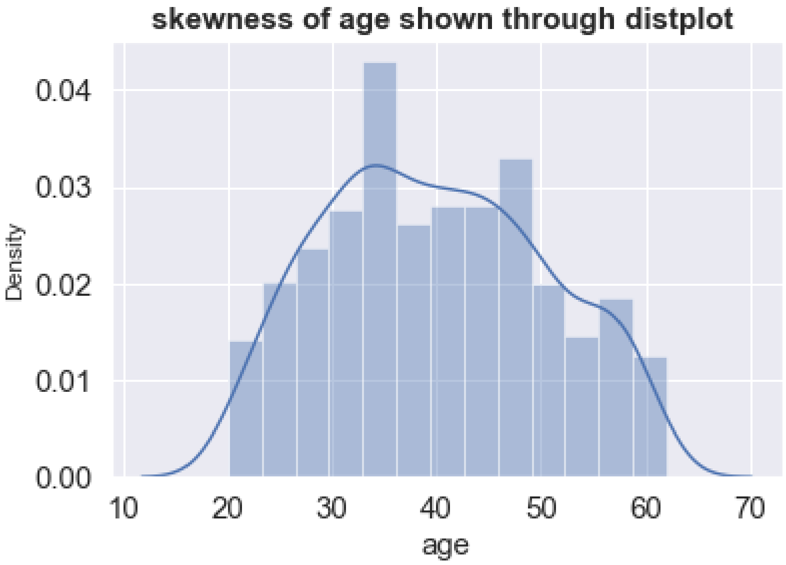


Fig 4: The distplot showing the skewness of data of 'Age' variable

From the above distplot, we can see that age is positively skewed or right skewed, with median=39, mode=0 and 44; mean=39.95. Age is a bimodal variable with two mode values.

h.3. Education(years)

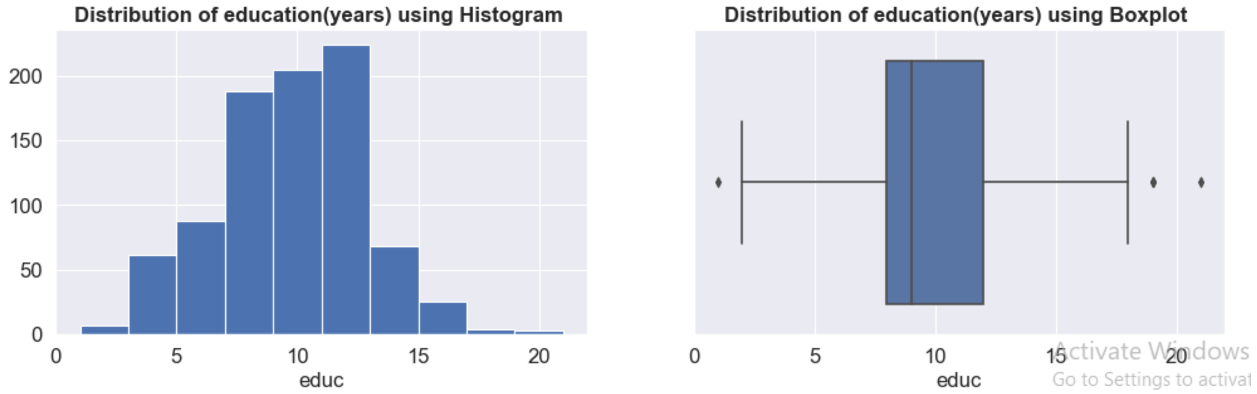


Fig 5: The box plot and the histogram showing the distribution of data of 'Education(years)' variable

From the univariate analysis using histogram, we see that total years of formal education is plotted along the x-axis. Most of the employees have total educational year between 11 and 13, approximately 190 employees, followed by those employees who are having total educational year between 9 and 11, 200 employees and the least number of employees have their total educational year between 19 to 22, which is approximately 5 employees. From the bivariate analysis using boxplot, there is presence of outliers. The second quartile(Q2) or median for the salary variable is about 9. The lower or first quartile(Q1) is about 8 and the upper or the third quartile(Q3) is about 12. The inter quartile range (IQR) for the above boxplot is 4.

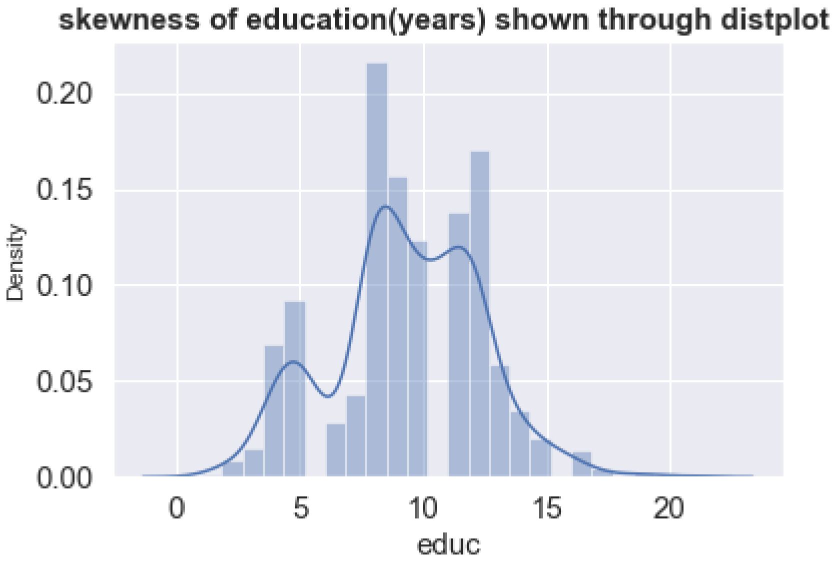


Fig 6: The distplot showing the skewness of data of 'Education(years)' variable

From the above distplot , we can see that education(years)is negatively skewed or left skewed, with median=9.0, mode=0 and 8; mean=9.307. Here also we can see that education(years) is bimodal with two modes.

h.4. no_young_children

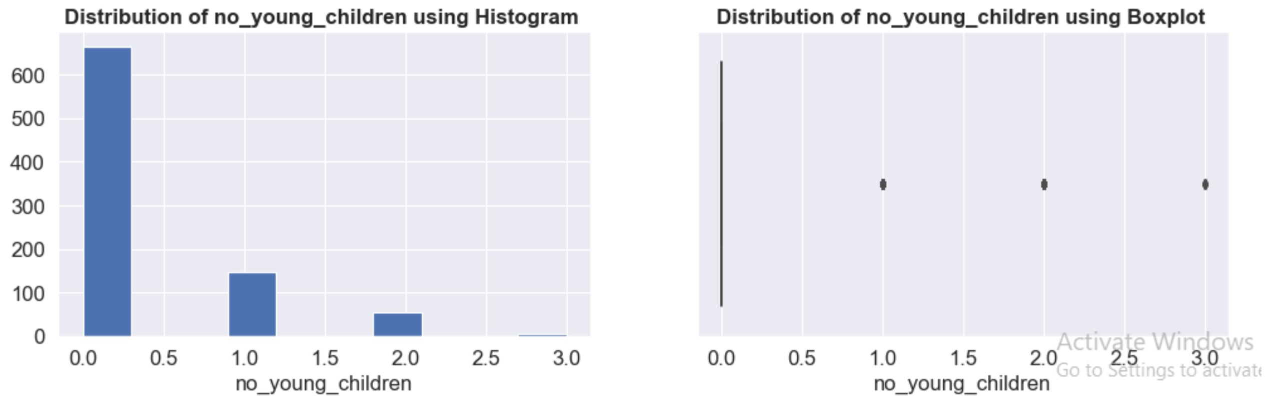


Fig 7: The box plot and the histogram showing the distribution of data of 'no_young_children' variable

From the univariate analysis using histogram, we see that the number of young children (younger than 7 years) is plotted along the x-axis. The minimum number of children is 0 and maximum number of children is 3. Out of the total 872 employees,665 employees do not have children younger than 7 years. 147 employees have one young child while 55 employees have 2 young children and only 3 employees have 3 young children.

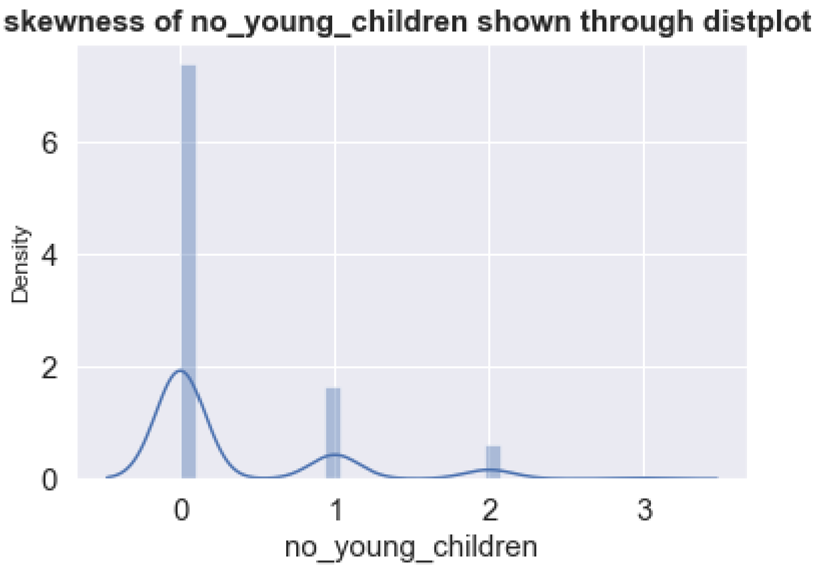


Fig 8: The distplot showing the skewness of data of 'no_young_children' variable

From the above distplot, we can see that no_young_children is positively skewed or right skewed, with mean=0.319

h.5. no_older_children

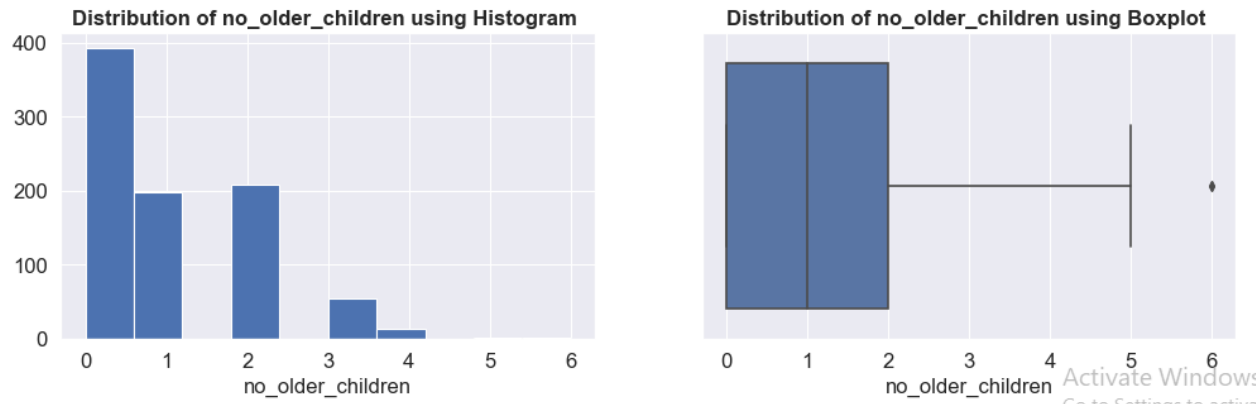


Fig 9: The box plot and the histogram showing the distribution of data of 'no_older_children' variable

From the univariate analysis using histogram, we see that the number of older children is plotted along the x-axis. The minimum number of children is 0 and maximum number of children is 6. Out of the total 872 employees,393 employees do not have older children. 198 employees have one elder child while 208 employees have 2 elder children, 55 employees have 3 elder children, 14 employees have 4 elder children and 2 employees have 5 elder children and another 2 employees have 6 elder children.

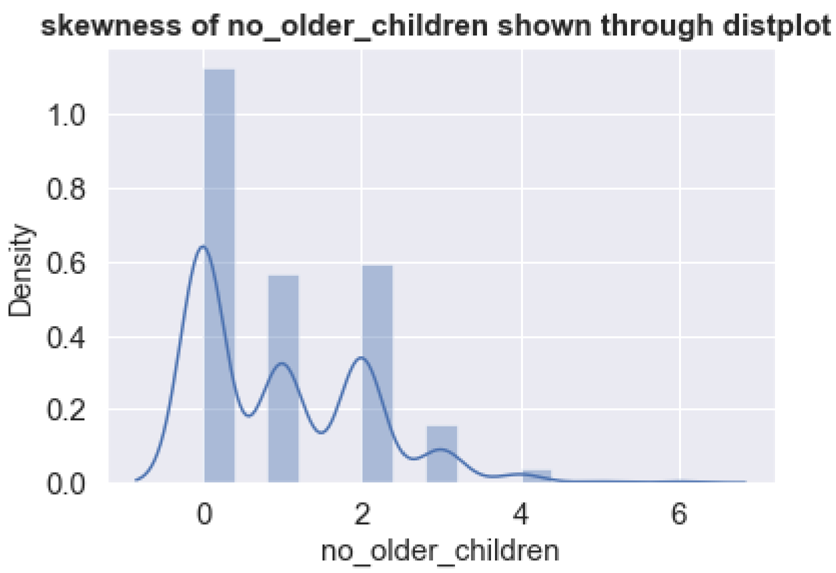


Fig 10: The distplot showing the skewness of data of 'no_older_children' variable

From the above distplot, we can see that no_young_children is positively skewed or right skewed, with mean=0.982

i. Graphical Representation of univariate and bivariate analysis for catagorical columns

i.1. Holliday_Package

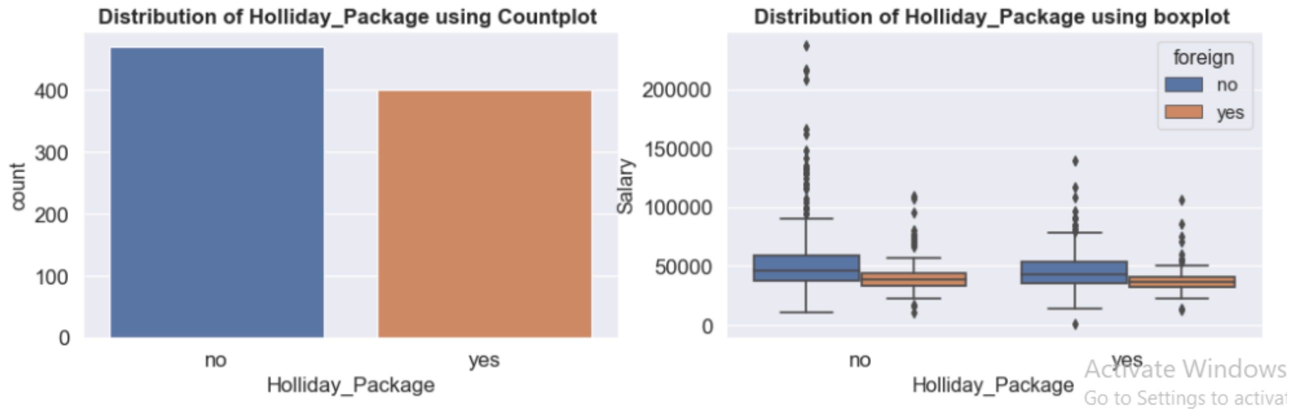


Fig 11: The box plot and the count plot showing the distribution of data of 'Holliday_Package' variable

From the univariate analysis using Count plot, we see that 401 employees have opted for the holiday package offered by the tour and travel agency while 471 employees did not opt for the holiday package. The bivariate analysis using boxplot was constructed keeping the label whether the employee is a foreigner or not. We can see that there is presence of outliers. Across the four different boxes, the median of the native employee rejecting the holiday package is highest while the median of the foreign employee accepting the holiday package is lowest. The employees rejecting the holiday package has more outliers.

i.2. foreign

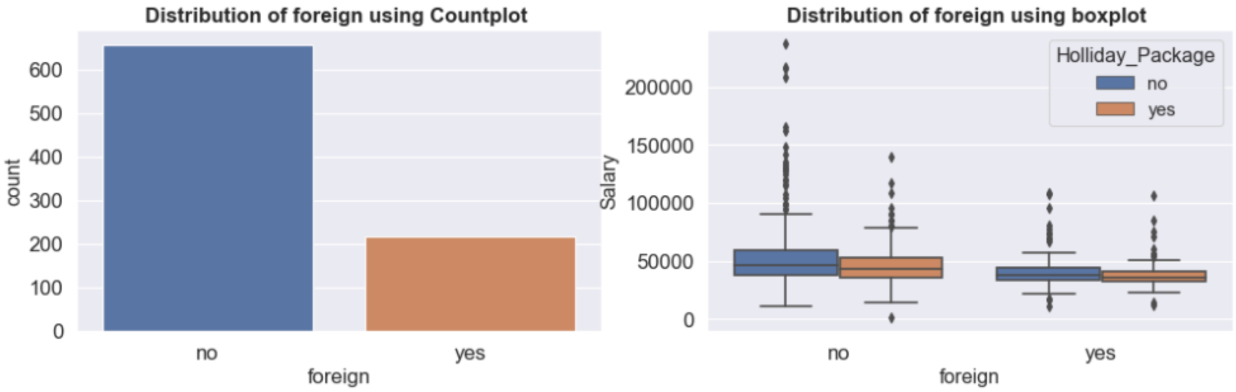


Fig 12: The box plot and the count plot showing the distribution of data of 'foreign' variable

From the univariate analysis using Count plot, we see that 656 employees are native employees while 216 employees are foreigners. The bivariate analysis using boxplot was constructed keeping the label whether the employee has opted for a holiday package or not. We can see that there is presence of outliers. Across the four different boxes, the median of the native employee rejecting the holiday package is highest while the median of the foreign employee accepting the holiday package is lowest. The employees rejecting the holiday package has more outliers.

j. Multivariate analysis

j.1.Heat Map

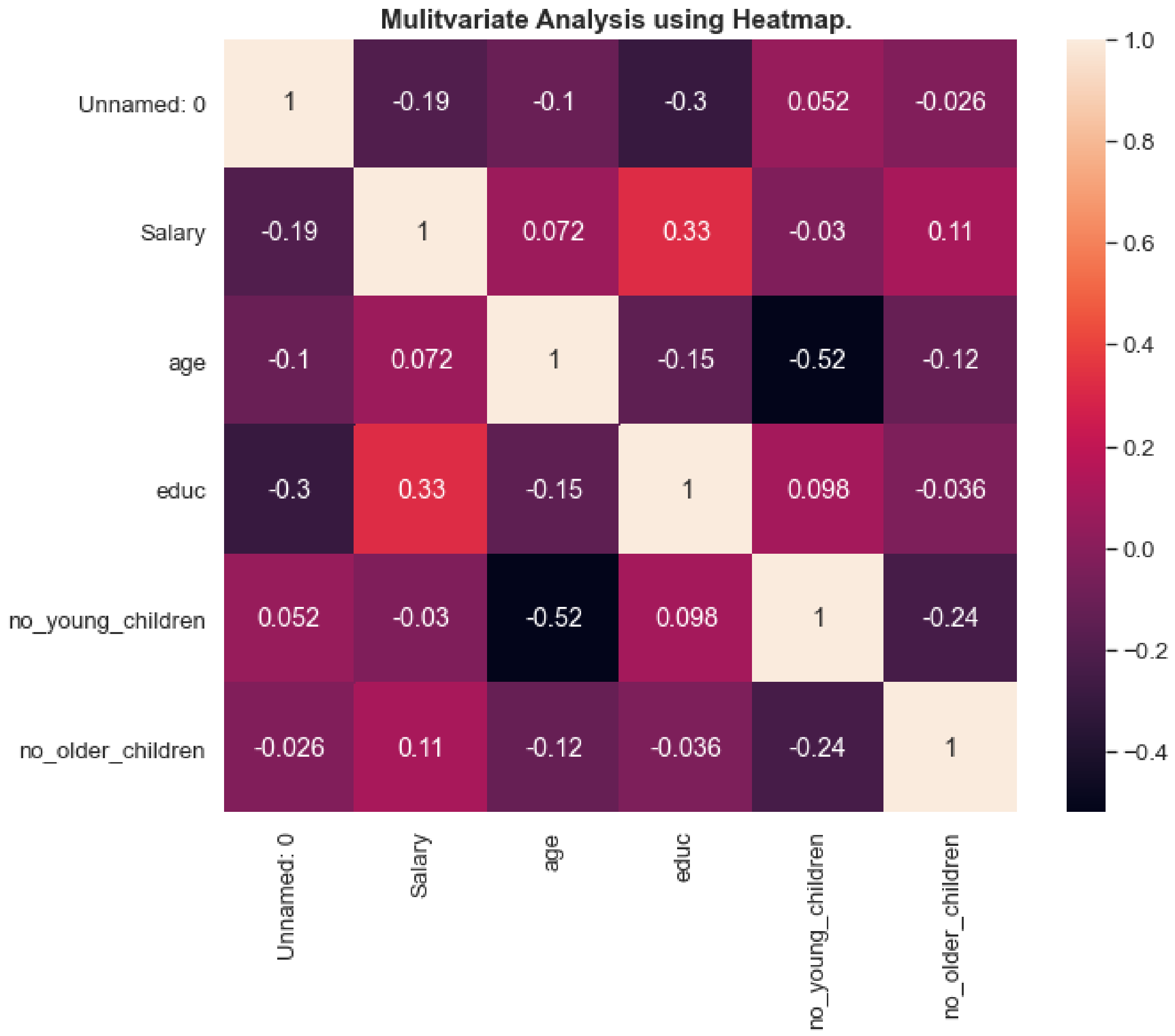


Fig 13: Heat map is portrayed for the multi variate analysis of the data.

The above heat map clearly shows the presence of multi collinearity in the dataset. It is seen that there exists high positive correlation between many components, among which ‘educ’ and ‘salary’ has the highest positive correlation followed by ‘Salary’ and ‘no_older_children’.There are many components with negative correlations among which ‘no_young_children’ and ‘no_older_children’ has the maximum negative correlation.

j.2. Pair Plot

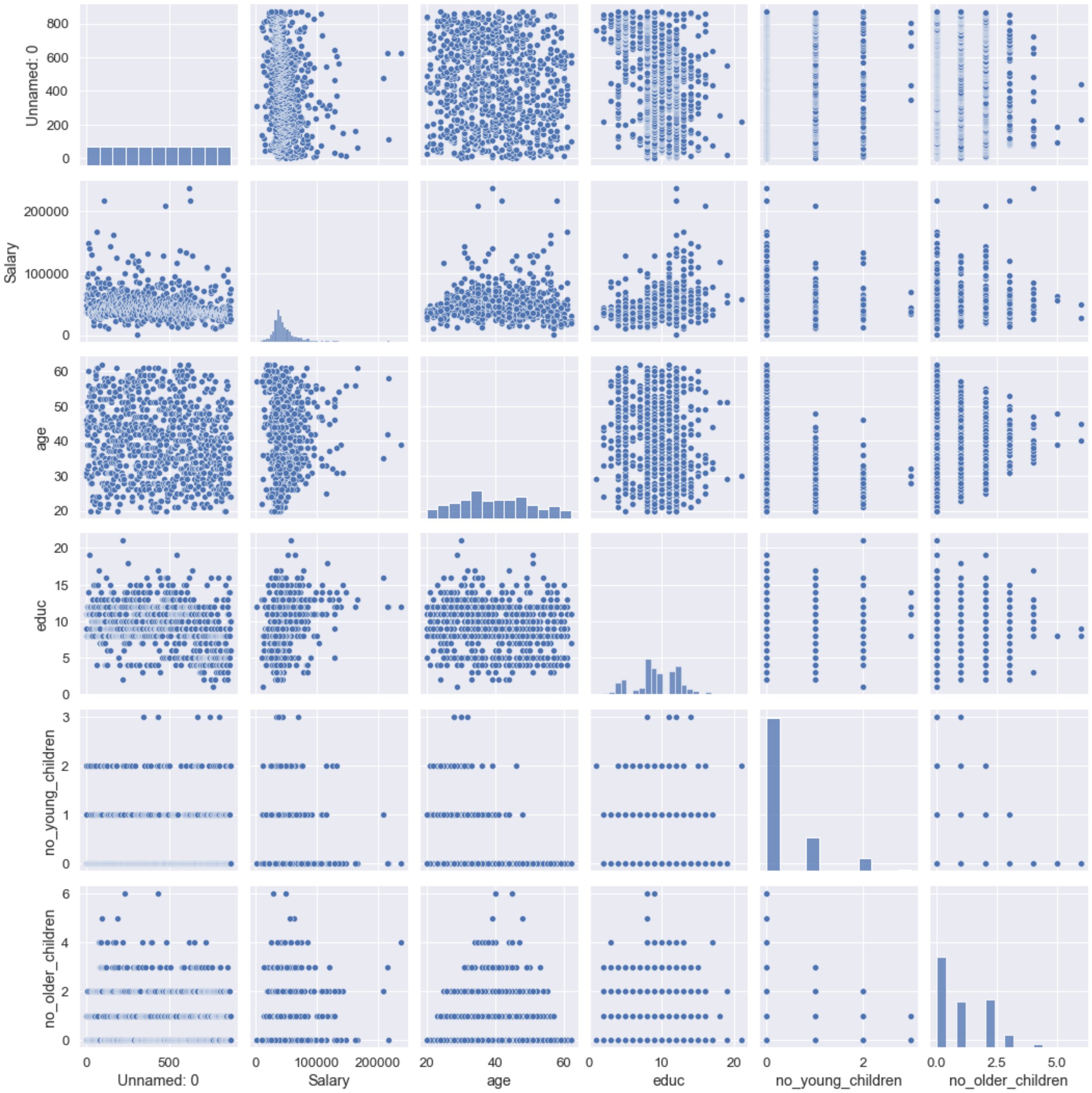


Fig 14: The pair plot showing the multivariate analysis

k. Skewness of the variables

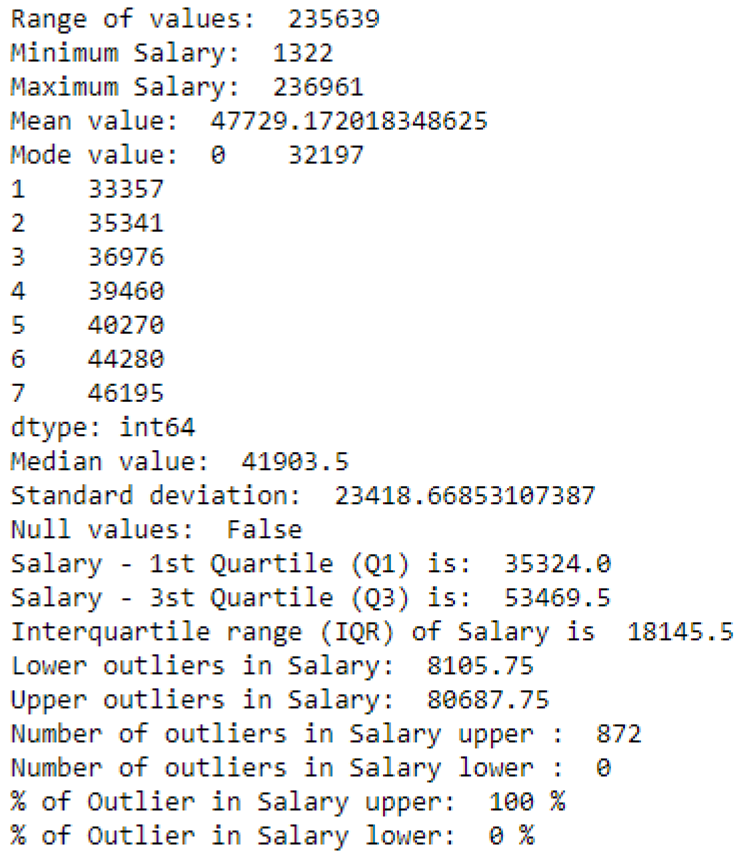
```
Salary               3.103216
age                  0.146412
educ                -0.045501
no_young_children    1.946515
no_older_children    0.953951
```

Table 6: The above table shows the skewness of various variables in the given dataset.

From the above given table, we can conclude that the variable ‘educ’ is negatively skewed while the rest of the variables are positively skewed. Out of the positively skewed variables, Salary has the highest skewness followed by no_young_children, then by no_older_children. Among the positively skewed variables it is age that has the least positive skewness.

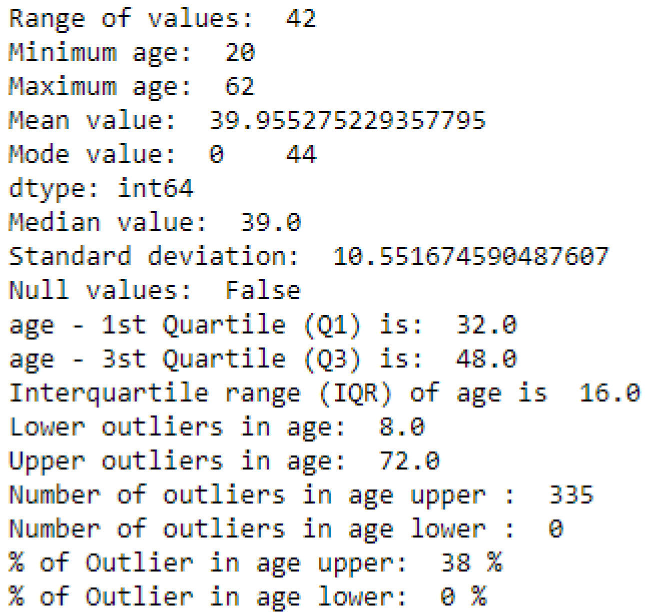
I. Outlier proportion

1. Salary



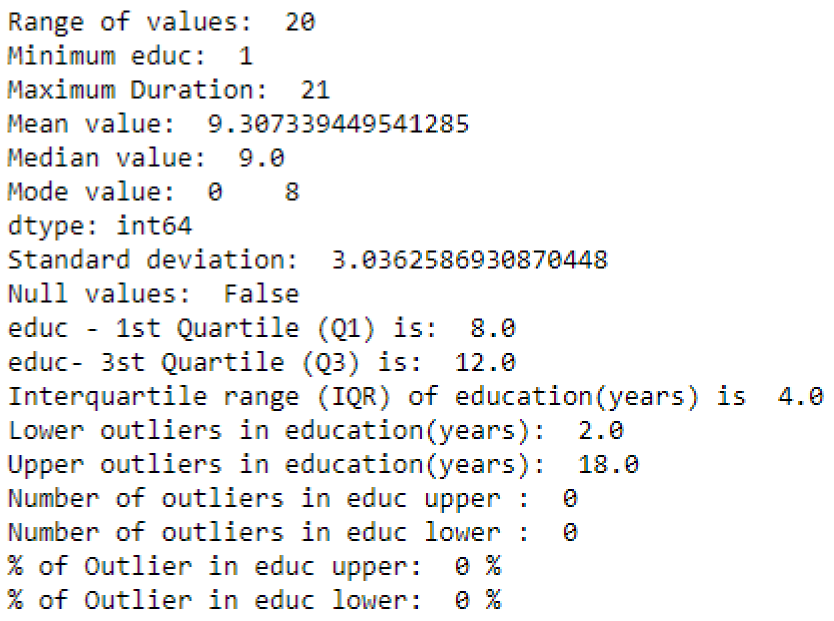
```
Range of values:  235639
Minimum Salary:  1322
Maximum Salary:  236961
Mean value:  47729.172018348625
Mode value:  0    32197
1    33357
2    35341
3    36976
4    39460
5    40270
6    44280
7    46195
dtype: int64
Median value:  41903.5
Standard deviation:  23418.66853107387
Null values:  False
Salary - 1st Quartile (Q1) is:  35324.0
Salary - 3st Quartile (Q3) is:  53469.5
Interquartile range (IQR) of Salary is  18145.5
Lower outliers in Salary:  8105.75
Upper outliers in Salary:  80687.75
Number of outliers in Salary upper :  872
Number of outliers in Salary lower :  0
% of Outlier in Salary upper:  100 %
% of Outlier in Salary lower:  0 %
```

2.Age



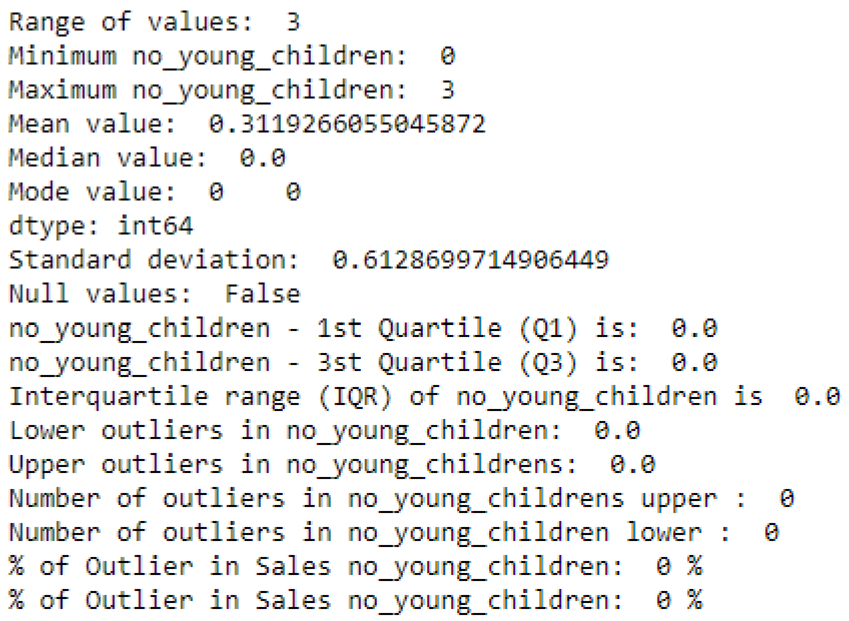
```
Range of values:  42
Minimum age:  20
Maximum age:  62
Mean value:  39.955275229357795
Mode value:  0    44
dtype: int64
Median value:  39.0
Standard deviation:  10.551674590487607
Null values:  False
age - 1st Quartile (Q1) is:  32.0
age - 3st Quartile (Q3) is:  48.0
Interquartile range (IQR) of age is  16.0
Lower outliers in age:  8.0
Upper outliers in age:  72.0
Number of outliers in age upper :  335
Number of outliers in age lower :  0
% of Outlier in age upper:  38 %
% of Outlier in age lower:  0 %
```

3.Educ



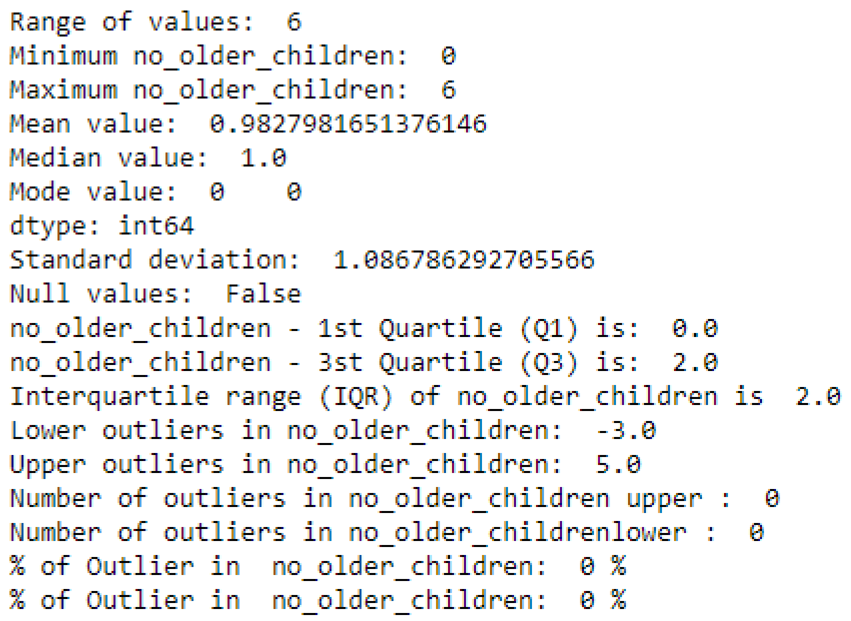
```
Range of values:  20
Minimum educ:  1
Maximum Duration:  21
Mean value:  9.307339449541285
Median value:  9.0
Mode value:  0    8
dtype: int64
Standard deviation:  3.0362586930870448
Null values:  False
educ - 1st Quartile (Q1) is:  8.0
educ- 3st Quartile (Q3) is:  12.0
Interquartile range (IQR) of education(years) is  4.0
Lower outliers in education(years):  2.0
Upper outliers in education(years):  18.0
Number of outliers in educ upper :  0
Number of outliers in educ lower :  0
% of Outlier in educ upper:  0 %
% of Outlier in educ lower:  0 %
```

4.no_young_children



```
Range of values:  3
Minimum no_young_children:  0
Maximum no_young_children:  3
Mean value:  0.3119266055045872
Median value:  0.0
Mode value:  0    0
dtype: int64
Standard deviation:  0.6128699714906449
Null values:  False
no_young_children - 1st Quartile (Q1) is:  0.0
no_young_children - 3st Quartile (Q3) is:  0.0
Interquartile range (IQR) of no_young_children is  0.0
Lower outliers in no_young_children:  0.0
Upper outliers in no_young_childrens:  0.0
Number of outliers in no_young_childrens upper :  0
Number of outliers in no_young_children lower :  0
% of Outlier in Sales no_young_children:  0 %
% of Outlier in Sales no_young_children:  0 %
```

5.no_older_children

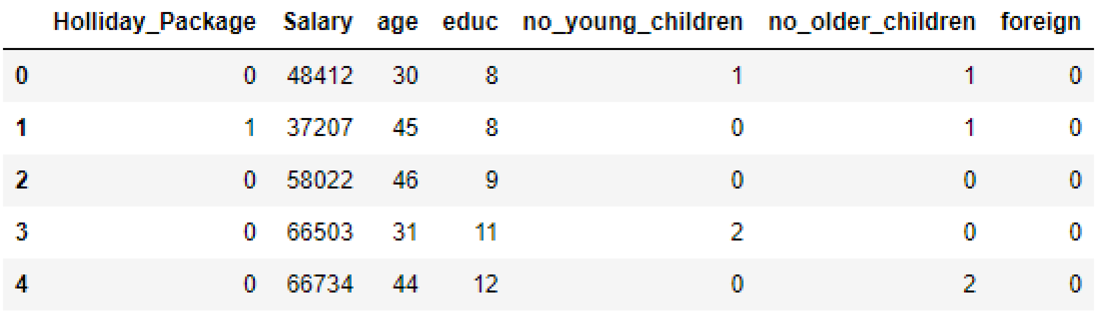


```
Range of values:  6
Minimum no_older_children:  0
Maximum no_older_children:  6
Mean value:  0.9827981651376146
Median value:  1.0
Mode value:  0    0
dtype: int64
Standard deviation:  1.086786292705566
Null values:  False
no_older_children - 1st Quartile (Q1) is:  0.0
no_older_children - 3st Quartile (Q3) is:  2.0
Interquartile range (IQR) of no_older_children is  2.0
Lower outliers in no_older_children:  -3.0
Upper outliers in no_older_children:  5.0
Number of outliers in no_older_children upper :  0
Number of outliers in no_older_childrenlower :  0
% of Outlier in  no_older_children:  0 %
% of Outlier in  no_older_children:  0 %
```

Q2.2. 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).

1.LOGISTIC REGRESSION

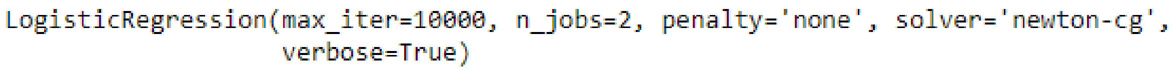
The data has not been scaled. The categorical columns Holliday_Package and foreign have been encoded using the label encoding technique as shown in the above code. Encoding helps the logistic regression model to predict better results.



	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

Table 7: The above table shows the head of the dataset after data encoding has been done.

The given data has been split into the training and testing data with a ratio of 70:30 respectively after dropping the dependant or the predictor variable- Holliday_Package_1.

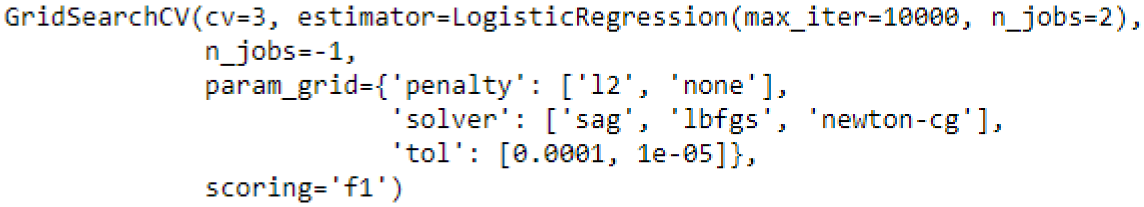


```
LogisticRegression(max_iter=10000, n_jobs=2, penalty='none', solver='newton-cg',
                   verbose=True)
```

The above parameters were used to model the logistic regression.

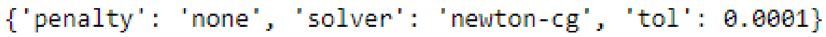
- **GRID SEARCH METHOD**

The grid search method is used for logistic regression to find the optimal parameters for building the model.

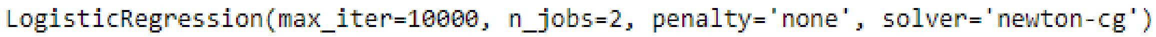


```
GridSearchCV(cv=3, estimator=LogisticRegression(max_iter=10000, n_jobs=2),
             n_jobs=-1,
             param_grid={'penalty': ['l2', 'none'],
                         'solver': ['sag', 'lbfgs', 'newton-cg'],
                         'tol': [0.0001, 1e-05]},
             scoring='f1')
```

The above is the list of parameters that has been loaded into the grid search CV



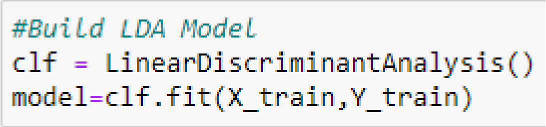
```
{'penalty': 'none', 'solver': 'newton-cg', 'tol': 0.0001}
```



```
LogisticRegression(max_iter=10000, n_jobs=2, penalty='none', solver='newton-cg')
```

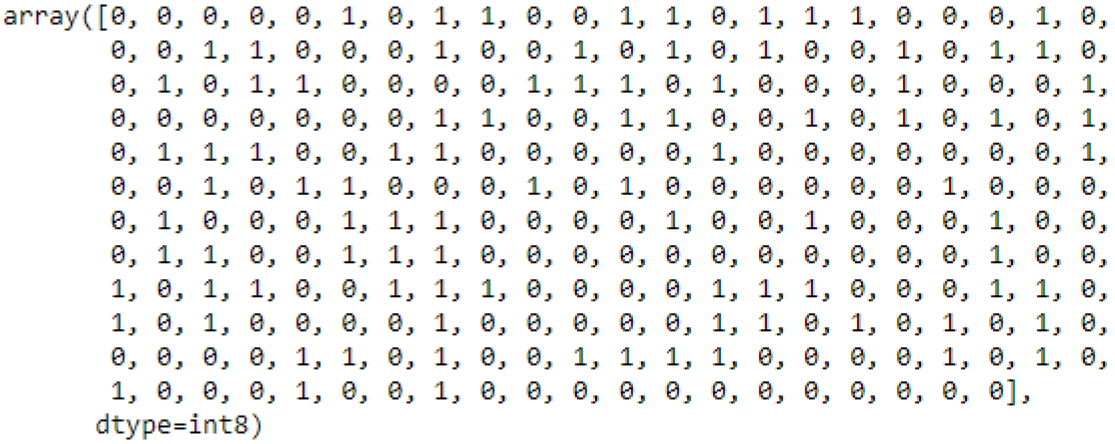
The above set of parameters were considered the best for this modelling.

2.LINEAR DISCRIMINANT ANALYSIS



```
#Build LDA Model
clf = LinearDiscriminantAnalysis()
model=clf.fit(X_train,Y_train)
```

LDA model was created on the given data set after the categorical variables were encoded using one hot encoding technique.



```
array([0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0,
       0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0,
       0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1,
       0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1,
       0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1,
       0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
       0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0,
       0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
       1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0,
       1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0,
       0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0,
       1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
      dtype=int8)
```

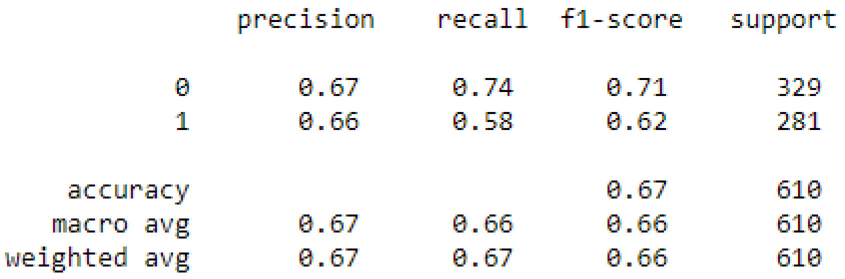
The above array shows the result of test data class prediction with a cut-off value of 0.5.

Q2.3. 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.

1.LOGISTIC REGRESSION

a. Classification Report

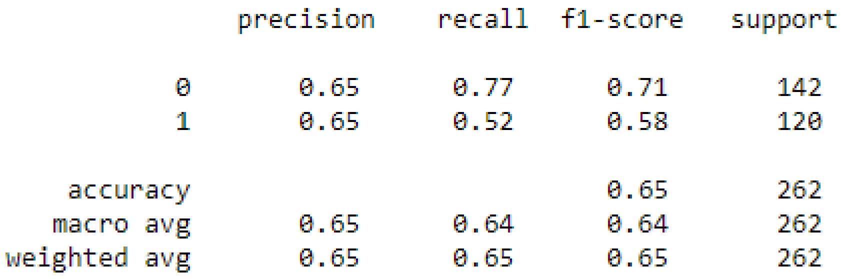
- Training Data



```
              precision    recall  f1-score   support

           0       0.67      0.74      0.71       329
           1       0.66      0.58      0.62       281

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

From the above classification report, we can see that using the logistic regression model, the precision for training data is 0.66, recall is 0.58, f1 score is 0.62 and accuracy is 0.67.

- Testing Data

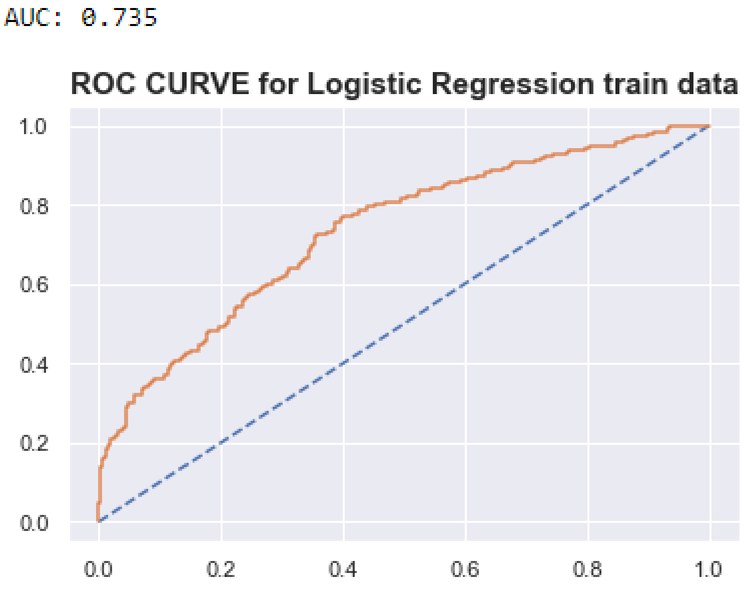


```
              precision    recall  f1-score   support

           0       0.65      0.77      0.71       142
           1       0.65      0.52      0.58       120

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

From the above classification report, we can see that using the Logistic regression model, the precision for testing data is 0.65, recall is 0.52, f1 score is 0.58 and accuracy is 0.65.

b. ROC curve and ROC_AUC score

- Training Data

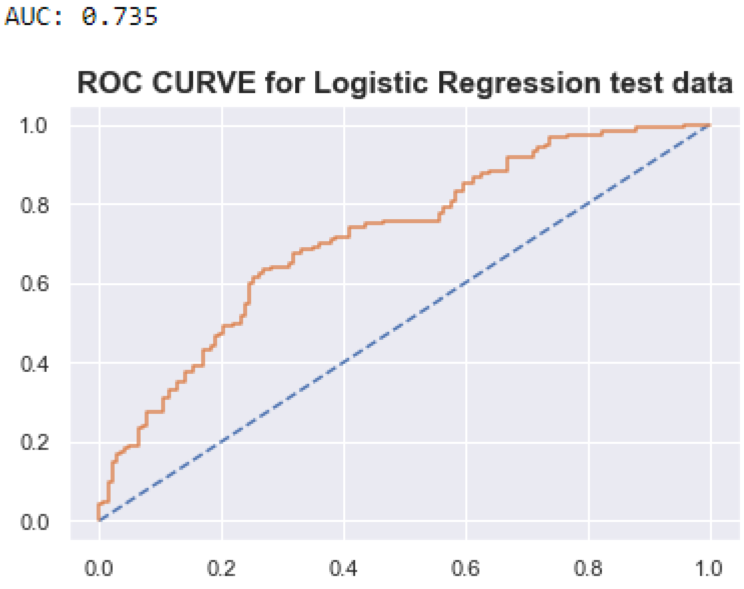


```
AUC: 0.735
```

Fig 15: The ROC curve for the training data using the Logistic regression model

Here, the AUC score is 0.735

- Testing Data



```
AUC: 0.735
```

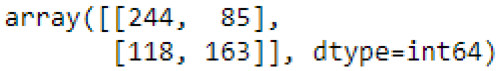
Fig 16: The ROC curve for the testing data using the Logistic regression model

Here, the AUC score is 0.735

From the above ROC_curve for both testing and training data, we can see that the curve is steady for both train and test data. Even the AUC score also remains the same for both the train and test data which means this model is good.

c. Confusion Matrix

Training Data

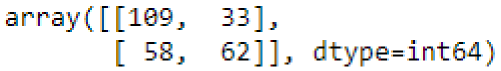


```
array([[244,  85],
       [118, 163]], dtype=int64)
```

Given above is the confusion matrix for Training data using the Logistic regression model.

- Here, True Positive =244, which is actually True (value=positive or 1) and has been predicted True too
- False Negative=85, which is actually True (value=positive or 1), but predicted False (value=negative or 0)
- False Positive =118, which is actually False (value=negative or 0), but predicted True (1)
- True Negative =163, which is actually False and has been predicted False too

Testing Data



```
array([[109,  33],
       [ 58,  62]], dtype=int64)
```

Given above is the confusion matrix for Testing data using the CART model.

- Here, True Positive =109, which is actually True (value=positive or 1) and has been predicted True too
- False Negative=33, which is actually True (value=positive or 1), but predicted False (value=negative or 0)
- False Positive =58, which is actually False (value=negative or 0), but predicted True (1)
- True Negative =62, which is actually False and has been predicted False too

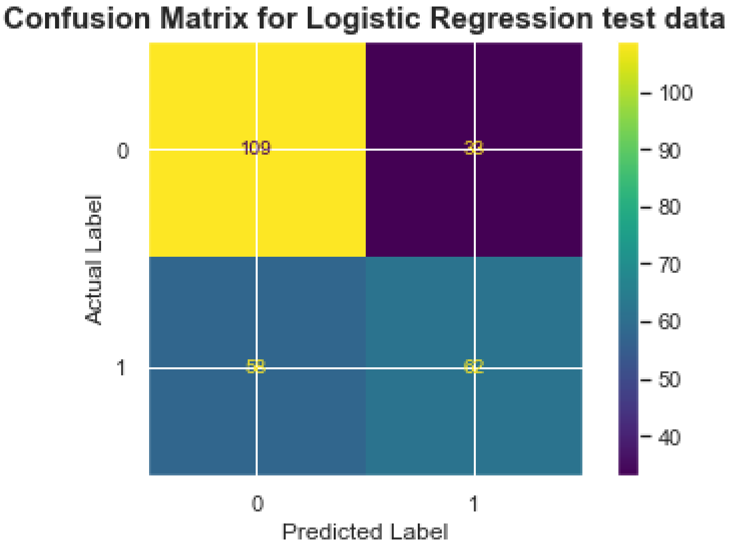
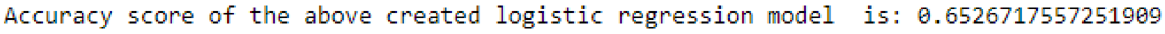


Fig 17: The figure shows the confusion matrix for the Logistic regression test data.

d. Accuracy



```
Accuracy score of the above created logistic regression model  is: 0.6526717557251909
```

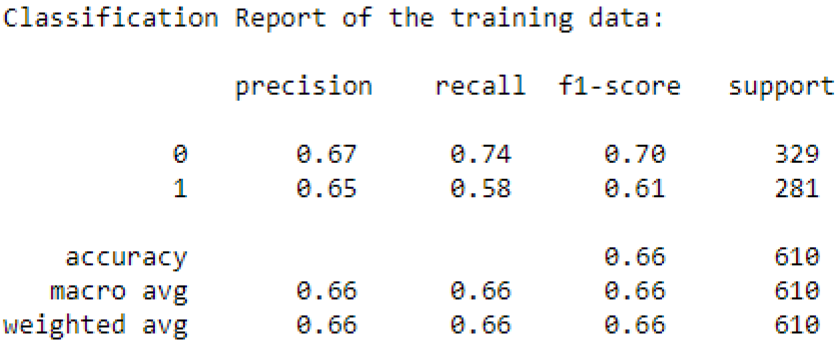
INFERENCES:

The accuracy of the test data model is equal to the train data model. The model is good.

2.LINEAR DISCRIMINANT ANALYSIS

a. Classification Report

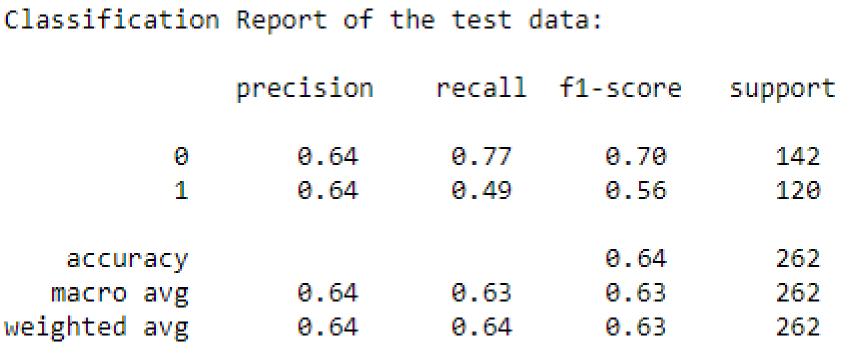
- Training Data



```
Classification Report of the training data:

              precision    recall  f1-score   support

           0       0.67      0.74      0.70       329
           1       0.65      0.58      0.61       281

    accuracy                           0.66       610
   macro avg       0.66      0.66      0.66       610
weighted avg       0.66      0.66      0.66       610
```

From the above classification report, we can see that using the Linear Discriminant Analysis, the precision for training data is 0.65, recall is 0.58, f1 score is 0.61 and accuracy is 0.66

➢ Testing Data

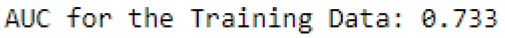


```
Classification Report of the test data:

              precision    recall  f1-score   support

           0       0.64      0.77      0.70       142
           1       0.64      0.49      0.56       120

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

From the above classification report, we can see that using the Linear Discriminant Analysis, the precision for testing data is 0.64, recall is 0.49, f1 score is 0.56 and accuracy is 0.64

b. ROC curve and ROC_AUC score

➢ Training Data



```
AUC for the Training Data: 0.733
```

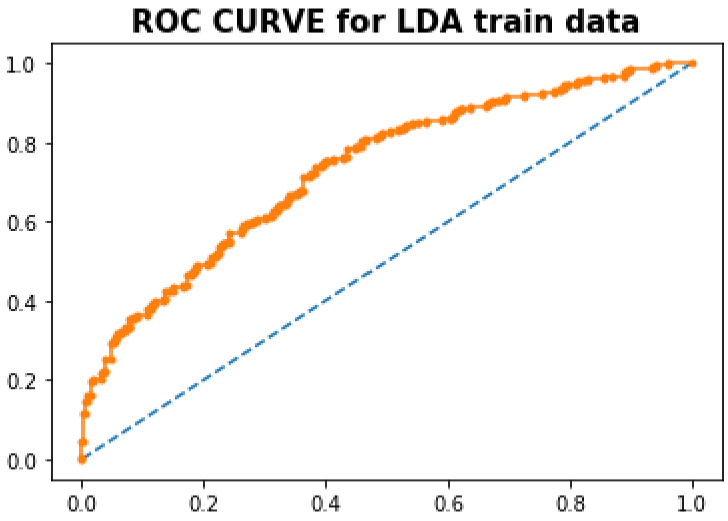
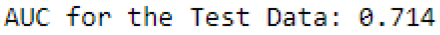


Fig 18: The ROC curve for the training data using the Linear Discriminant Analysis

➢ Testing Data



```
AUC for the Test Data: 0.714
```

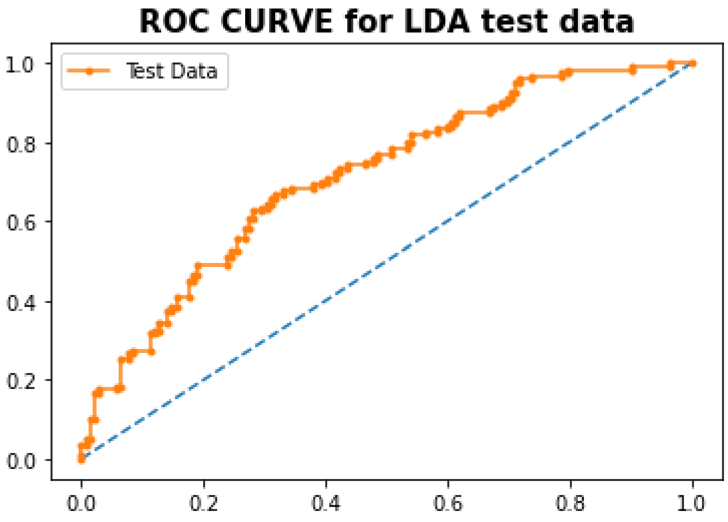
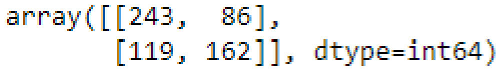


Fig 19: The ROC curve for the testing data using the Linear Discriminant Analysis

From the above ROC_curve for both testing and training data, we can see that the curve becomes flatter for the testing data in comparison to the testing data. Even the AUC score drops from 0.733 to 0.714 which shows that this model is not good enough; this drop is negligible and can be ignored. Thus, we can conclude this model is fair enough.

c. Confusion Matrix

➢ Training Data

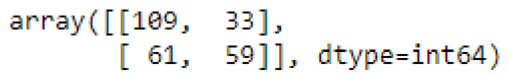


```
array([[243,  86],
       [119, 162]], dtype=int64)
```

Given above is the confusion matrix for Training data using the Random Forest model.

- Here, True Positive =243, which is actually True (value=positive or 1) and has been predicted True too
- False Negative=86, which is actually True (value=positive or 1), but predicted False (value=negative or 0)
- False Positive =119, which is actually False (value=negative or 0), but predicted True (1)
- True Negative =162, which is actually False and has been predicted False too

➢ Testing Data



```
array([[109,  33],
       [ 61,  59]], dtype=int64)
```

Given above is the confusion matrix for Training data using the Random Forest model.

- Here, True Positive =109, which is actually True (value=positive or 1) and has been predicted True too
- False Negative=33, which is actually True (value=positive or 1), but predicted False (value=negative or 0)
- False Positive =61, which is actually False (value=negative or 0), but predicted True (1)
- True Negative =59, which is actually False and has been predicted False too

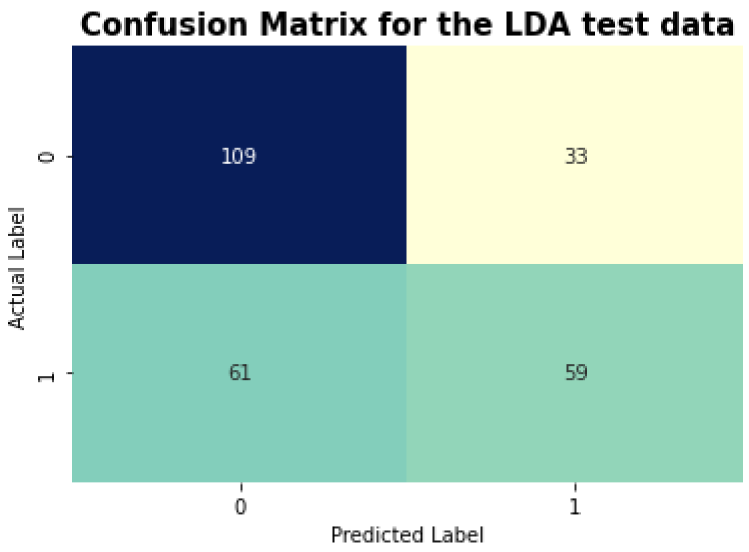
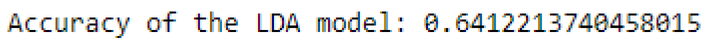


Fig 20: The figure shows the confusion matrix for the LDA test data.

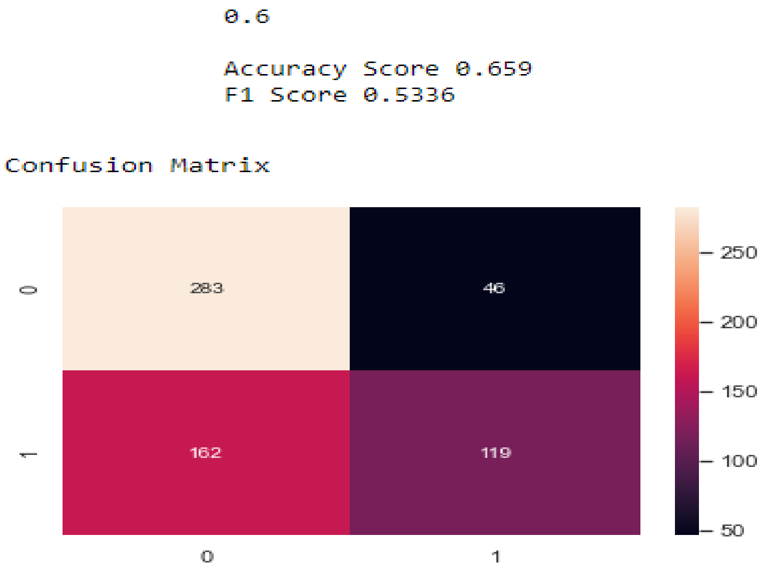
d. Accuracy



```
Accuracy of the LDA model: 0.6412213740458015
```

- <u>CHANGING THE CUT OFF VALUES</u>

We can change the cut off values to check the optimal value that gives better accuracy and F1 score. We will do this exercise only on the training data. The cut off values was changed from 0.1 to 0.9 and the respective accuracies, f1 score and confusion matrix was found. Out of all cut off values we see that 0.5 and 0.6 gives better accuracy than the rest of the custom cut-off values. We have already created the model using the cut off value=0.5 Besides that 0.6 cut-off gives us the best 'f1-score'. Here, we will take the cut-off as 0.6 to get the optimum 'f1' score.



```
0.6

Accuracy Score 0.659
F1 Score 0.5336

Confusion Matrix
```

MODEL COMPARISON

PERFORMANCE METRICS	LOGISTIC REGRESSION TRAIN	LOGISTIC REGRESSION TEST	LDA TRAIN	LDA TEST
ACCURACY	0.667	0.652	0.66	0.64
AUC	0.735	0.735	0.733	0.714
RECALL	0.58	0.52	0.58	0.49
PRECISION	0.66	0.65	0.65	0.64
F1 SCORE	0.62	0.58	0.61	0.56

TABLE 8: Model Comparison

Comparing both these models, we find both results are almost similar, but LDA works better when there is category target variable. LDA is a dimension reduction technique. LDA comes to our rescue in situations where classes are well separated and when the data is small. Here, the given dataset is of small size and the classes are also well separated.

Q2.4. Inference: Basis on these predictions, what are the insights and recommendations. Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

We had a business problem where we need predict whether an employee would opt for a holiday package or not, for this problem we had done predictions both logistic regression and linear discriminant analysis. We found that both are results are same.

The EDA analysis clearly indicates certain criteria where we could find people aged above 50 are not interested much in holiday packages. So, this clearly shows that age is one of the factors in determining whether an employee will opt for holiday packages or not. People ranging from the age 30 to 50 generally opt for holiday packages. Employee age over 50 to 60 have seems to be not taking the holiday package, whereas in the age 30 to 50 and salary less than 50000 people have opted more for holiday package. The salary is also another factor which plays an important role in this context.

The important factors deciding the predictions are salary, age and educ.

Recommendations

1. To attract employees above age 50 to the holiday packages, we can include religious destinations in the packages.

2. For people earning more than 150000 we can provide vacation holiday packages.

3. For employees with a greater number of older children, we can give a discount in the ticket, so that the employees can bring in their family along with them.

4.Employees having lower salary can be given discount on holiday packages which will attract them to choose the package.

THE END