PRREDICTIVE MODELING PROJECT

ANSWER REPORT

NAME – ELTON REBELLO

BATCH – JAN22A

DATE – 19 – 06 –2022

Table Of Contents

List of Tables

List of Figures

Problem 1: Linear Regression

Executive Summary -

Gem Stones co ltd, is a cubic zirconia manufacturer which is an inexpensive diamond alternative with many of the same qualities as a diamond. The company wants us to help them in predicting the price for the stone so that they can distinguish between higher profitable stones and lower profitable stones so as to have better profit share.

Introduction -

The purpose of the problem is to explore the dataset. Do the exploratory analysis. The data consists details of 26967 cubic zirconia stones and their carat, cut, color, clarity, depth, table, Length(X), Width(Y), Height(Z). We will be using Linear Regression to solve the problem.

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 – Color 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, SI1, SI2, I1.
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. X – Length of the cubic zirconia in mm.
8. Y – width of the cubic zirconia in mm.
9. Z – Height of the cubic zirconia in mm.

Sample of the Dataset -

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

Table No 1 – Sample of the Dataset

Dataset has 11 variables with details about the cubic zirconia.

Problem 1.1

Read the data and exploratory data analysis. Describe the data briefly. Perform Univariate and Bivariate Analysis.

Exploratory Data Analysis (EDA) -

Types of Variables in the dataset -

carat	float64
cut	object
color	object
clarity	object
depth	float64
table	float64
x	float64
y	float64
z	float64
price	int64

There is total 26967 rows and 10 columns in the dataset. 6 columns are of float datatype, 3 columns are of object datatype and 1 is of integer datatype.

Missing values in the dataset -

```
RangeIndex: 26967 entries, 0 to 26966
Data columns (total 10 columns):
    carat    26967 non-null  float64
    cut      26967 non-null  object
    color    26967 non-null  object
    clarity  26967 non-null  object
    depth    26270 non-null  float64
    table    26967 non-null  float64
    x        26967 non-null  float64
    y        26967 non-null  float64
    z        26967 non-null  float64
    price    26967 non-null  int64
```

From the above result we can see that there are missing values in depth column.

Univariate Analysis -

Describing the data -

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
carat	26967	NaN	NaN	NaN	0.80	0.48	0.2	0.4	0.7	1.05	4.5
cut	26967	5	Ideal	10816	NaN	NaN	NaN	NaN	NaN	NaN	NaN
color	26967	7	G	5661	NaN	NaN	NaN	NaN	NaN	NaN	NaN
clarity	26967	8	SI1	6571	NaN	NaN	NaN	NaN	NaN	NaN	NaN
depth	26970	NaN	NaN	NaN	61.7	1.41	50.8	61	61.8	62.5	73.6
table	26967	NaN	NaN	NaN	57.5	2.23	49.0	56	57	59	79
x	26967	NaN	NaN	NaN	5.73	1.12	0.0	4.71	5.69	6.55	10.23
y	26967	NaN	NaN	NaN	5.73	1.17	0.0	4.71	5.71	6.54	58.9
z	26967	NaN	NaN	NaN	3.54	0.72	0.0	2.9	3.52	4.04	31.8
price	26967	NaN	NaN	NaN	3939.54	4024.86	326	945	2375	5360	18818

Table No 2 –Description of the data

From the descriptive we can see the mean/median of the variables.

Unique Values of categorical Variables -

CUT : 5
Fair 781
Good 2441
Very Good 6030
Premium 6899
Ideal 10816

COLOR: 7
J 1443
I 2771
D 3344
H 4102
F 4729
E 4917
G 5661

CLARITY : 8
I1 365
IF 894
VVS1 1839
VVS2 2531
VS1 4093
SI2 4575
VS2 6099
SI1 6571

Histogram of carat -

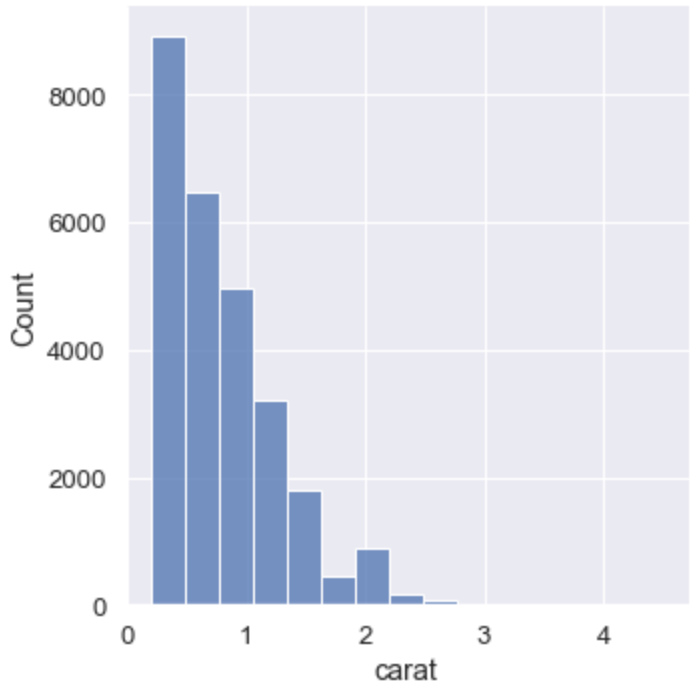

Fig No 1 – histogram of carat

Histogram of depth -

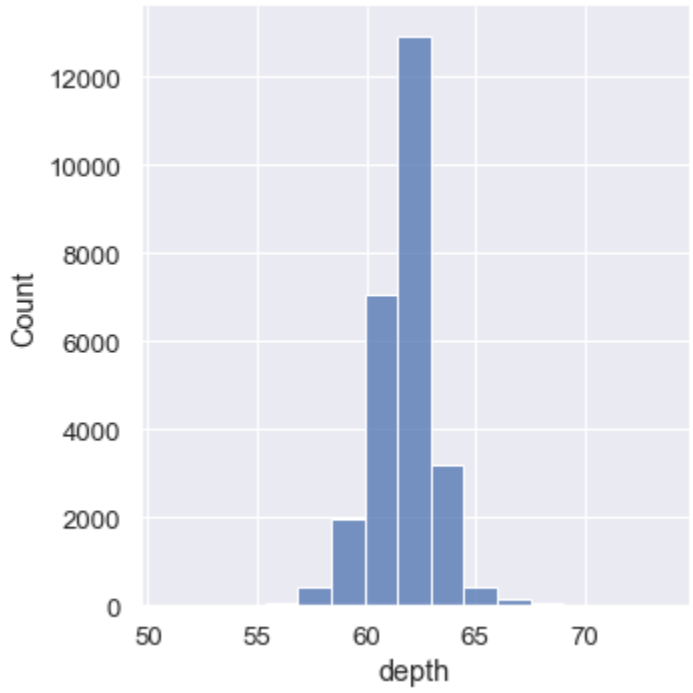

Fig No 2 – histogram of depth

Histogram of table -

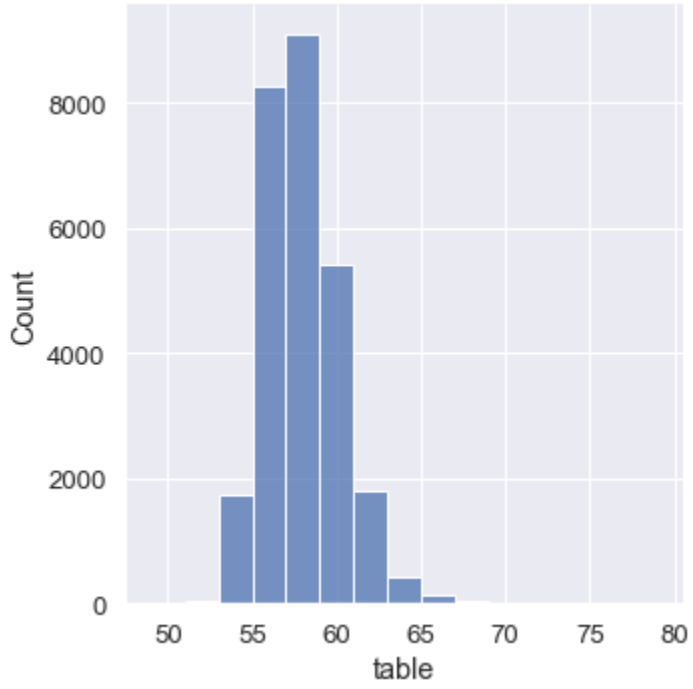

Fig No 3 – histogram of table

Histogram of x -

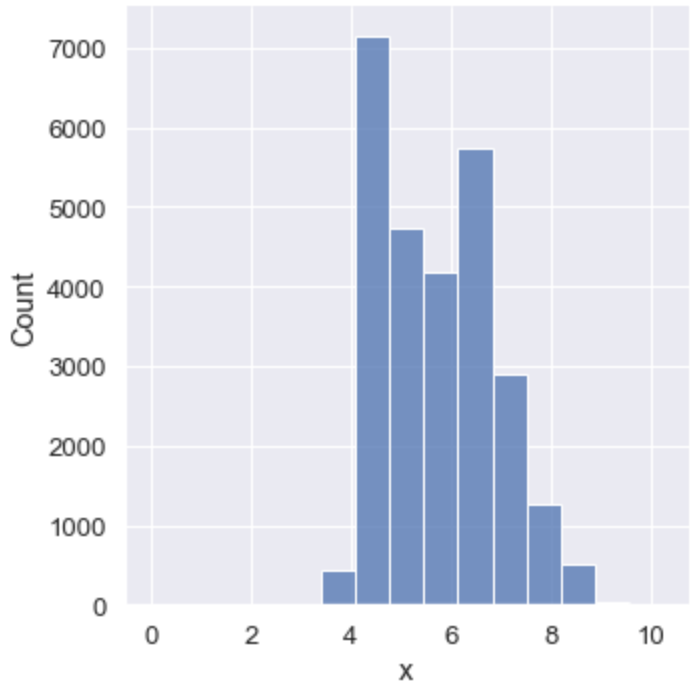

Fig No 4 – histogram of x

Histogram of y -

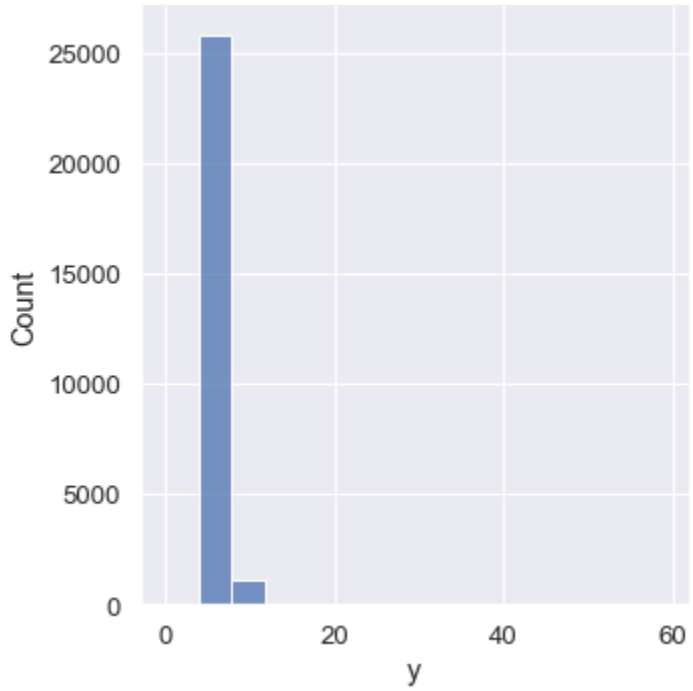

Fig No 5 – histogram of y

Histogram of z -

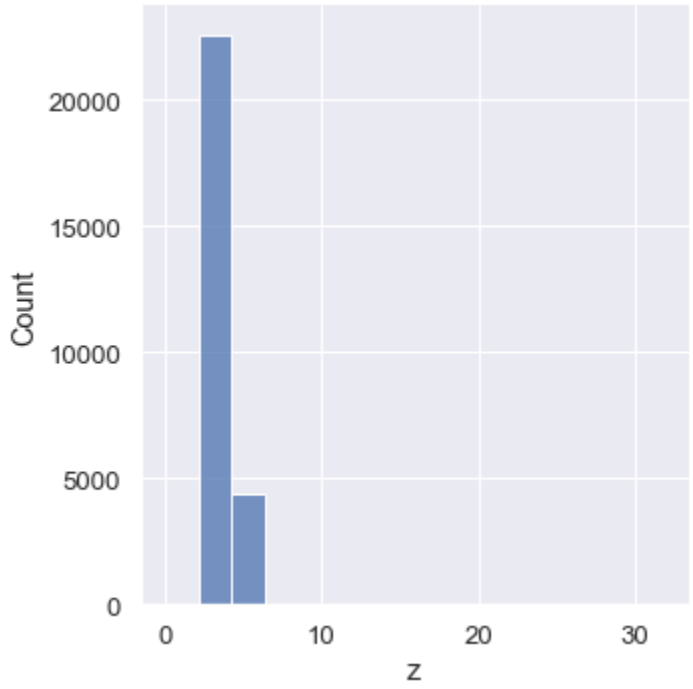

Fig No 6 – histogram of z

Histogram of price -

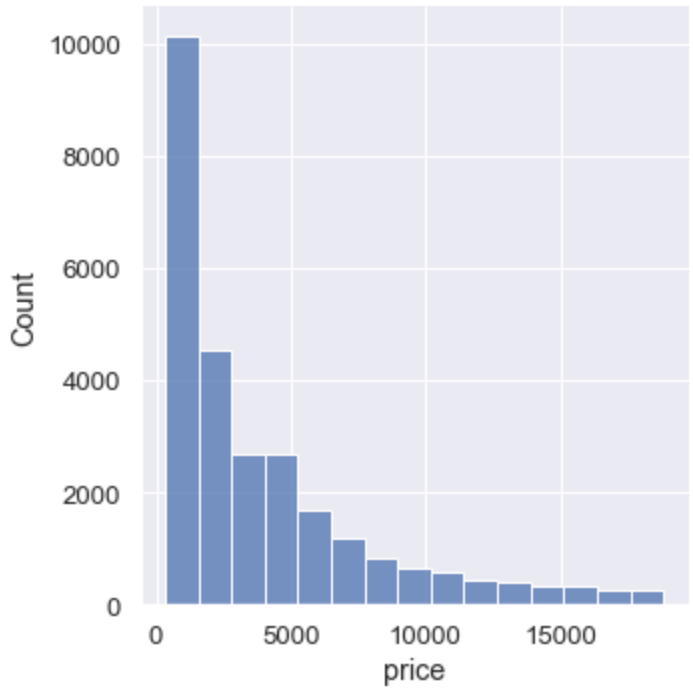

Fig No 7 – histogram of price

Bivariate Analysis -

Correlation Plot -

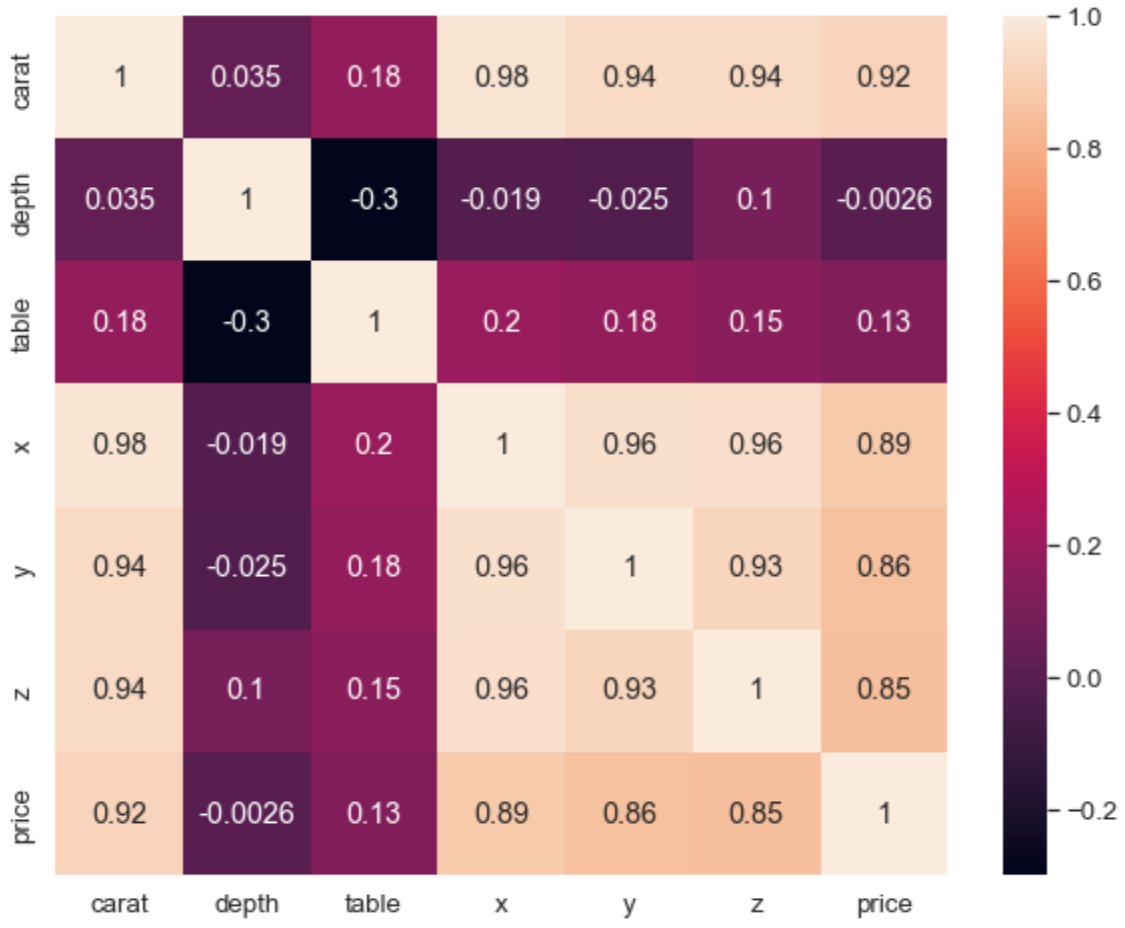

Fig No 8 –correlation plot

From the correlation plot, we can see that various attributes are highly correlated to each other. Correlation values near to 1 or –1 are highly positively correlated and highly negatively correlated respectively. Correlation values near to 0 are not correlated to each other.

Pair plot -

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

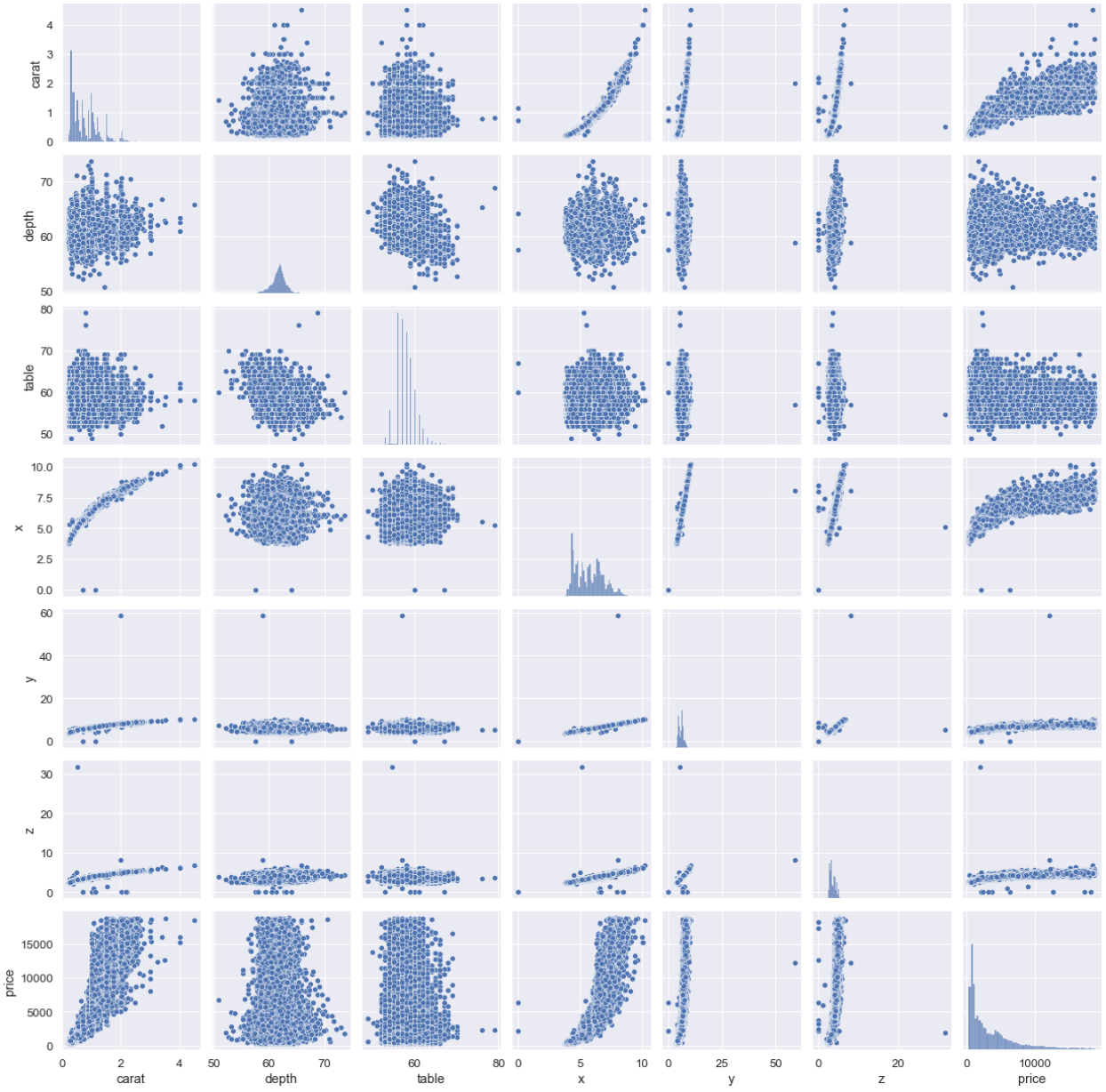

Fig No 9 – Pairplot

Boxplot -

A boxplot is a graph that gives a good indication of how the values in the data are spread out.

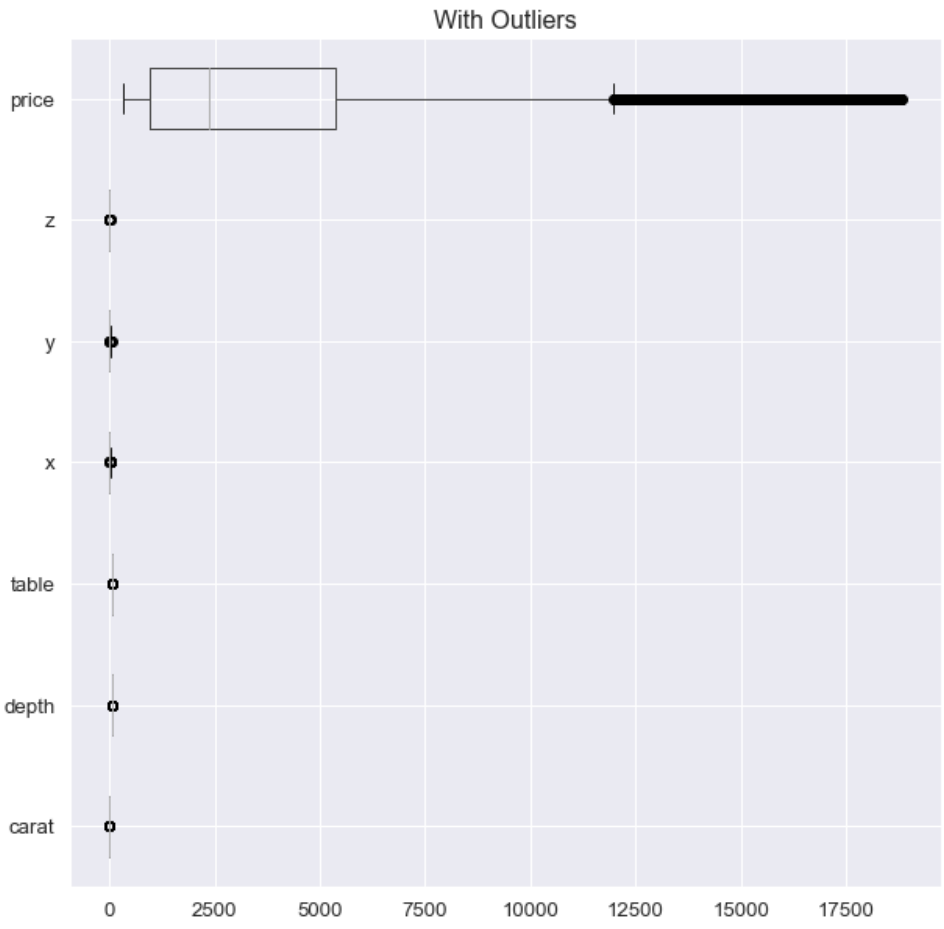

Fig No 10 – Boxplot

There are outliers present in all the variables as per the above diagram.

Boxplot of carat -

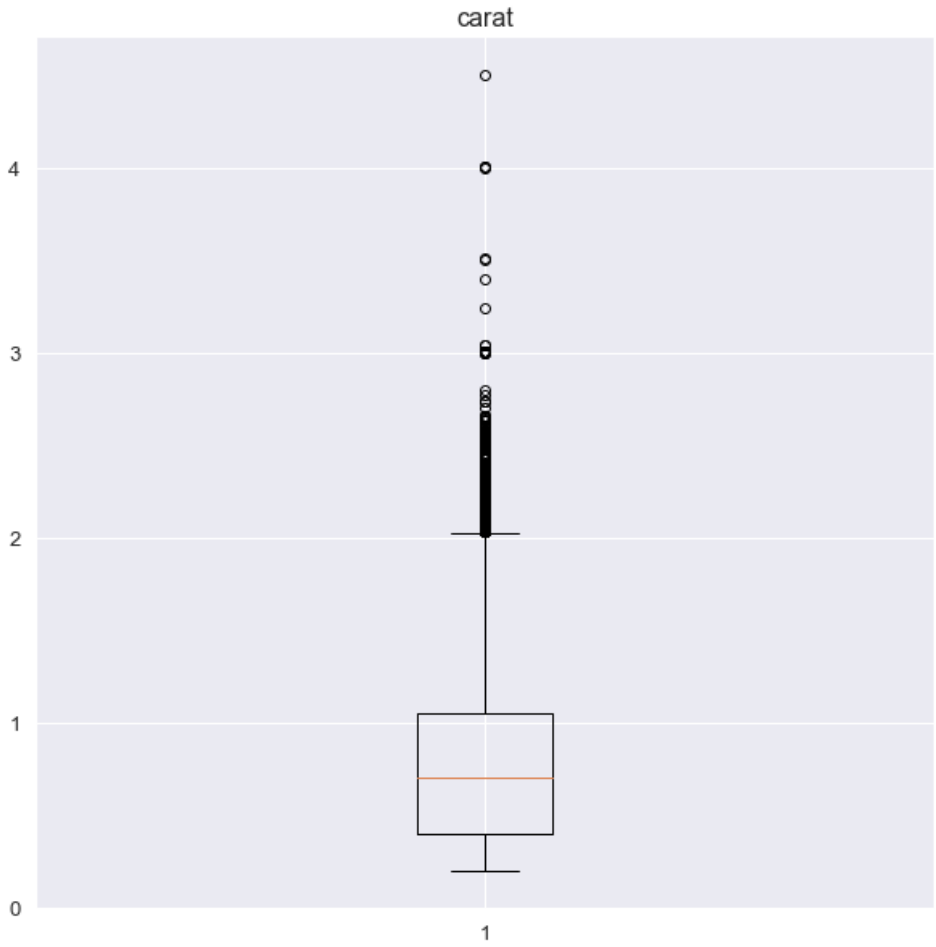

Fig No 11 – Boxplot of carat

Boxplot of depth -

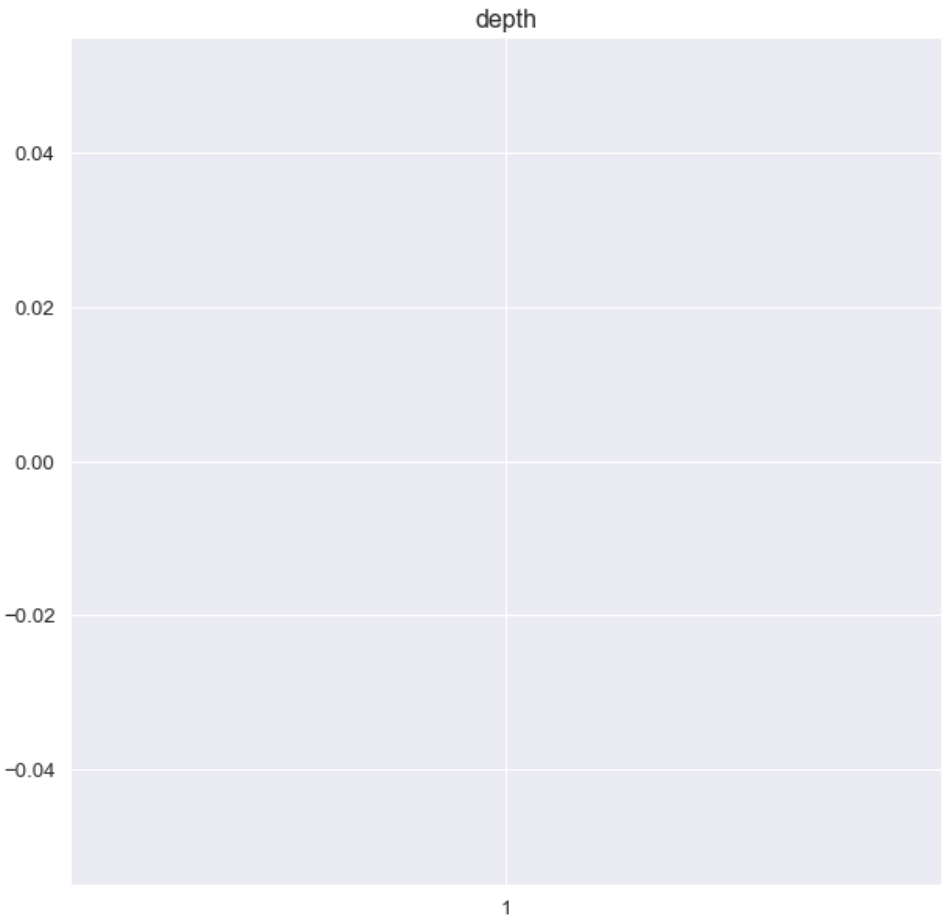

Fig No 12 – Boxplot of depth

Boxplot of table -

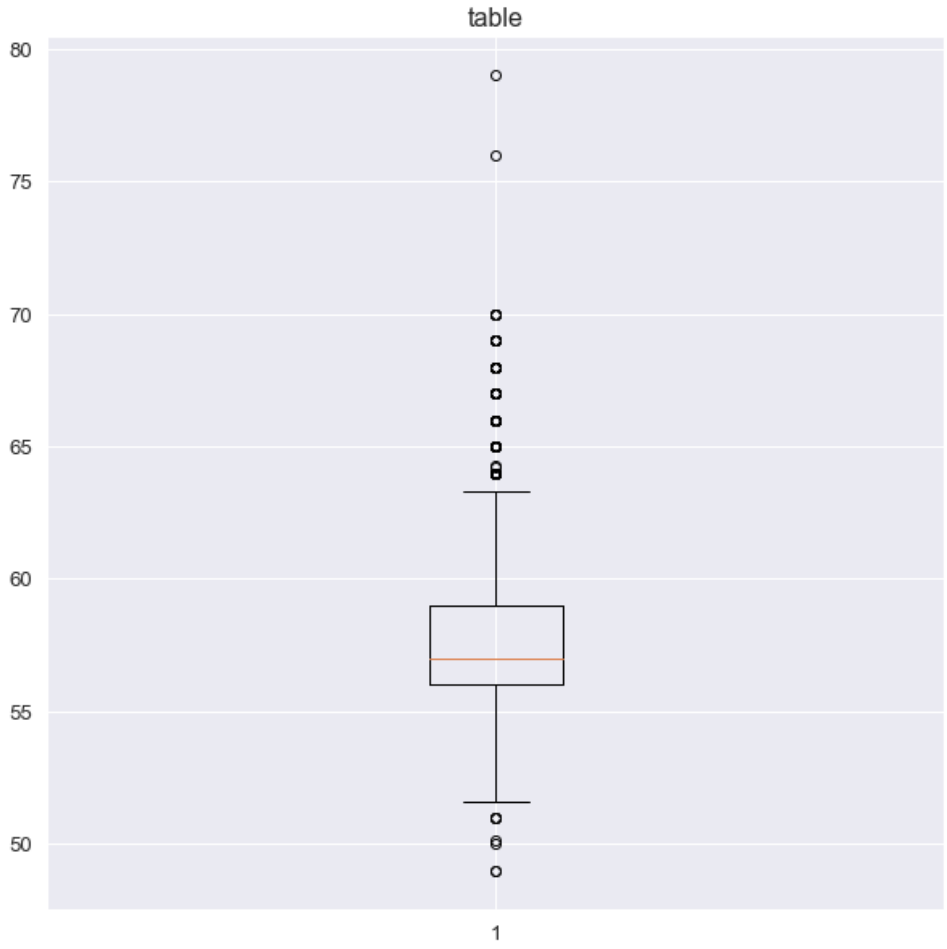

Fig No 13 – Boxplot of table

Boxplot of x -

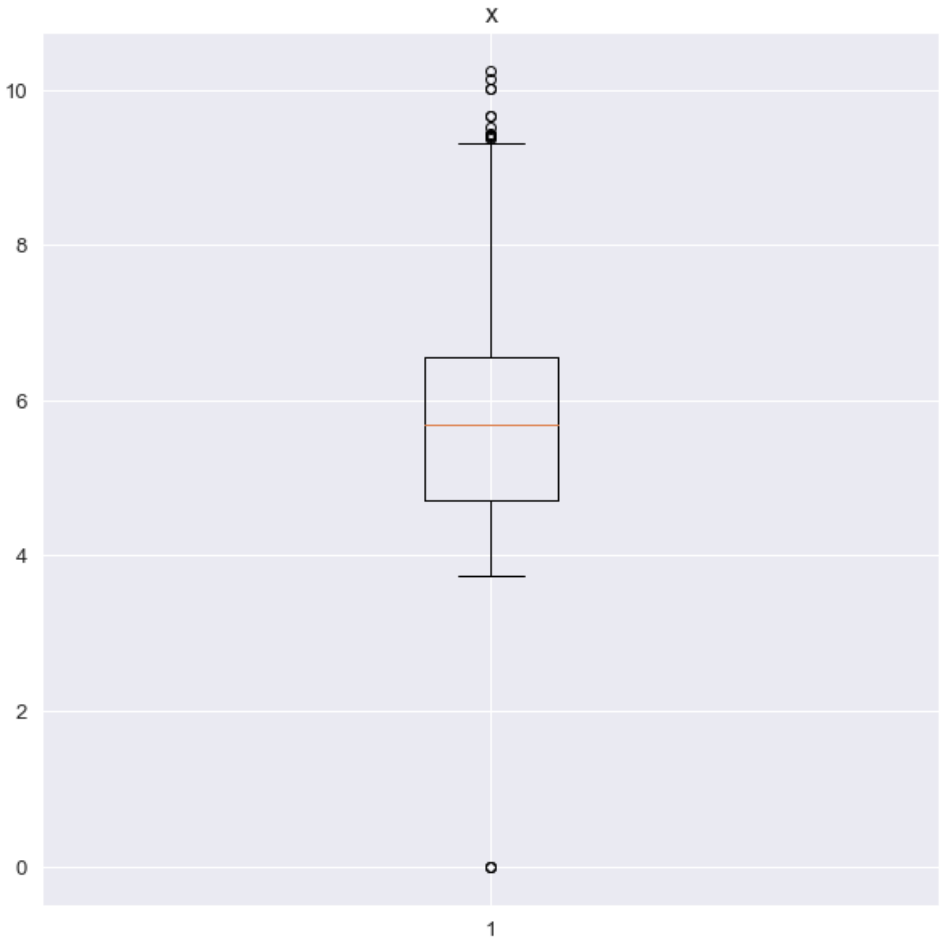

Fig No 14 – Boxplot of x

Boxplot of y -

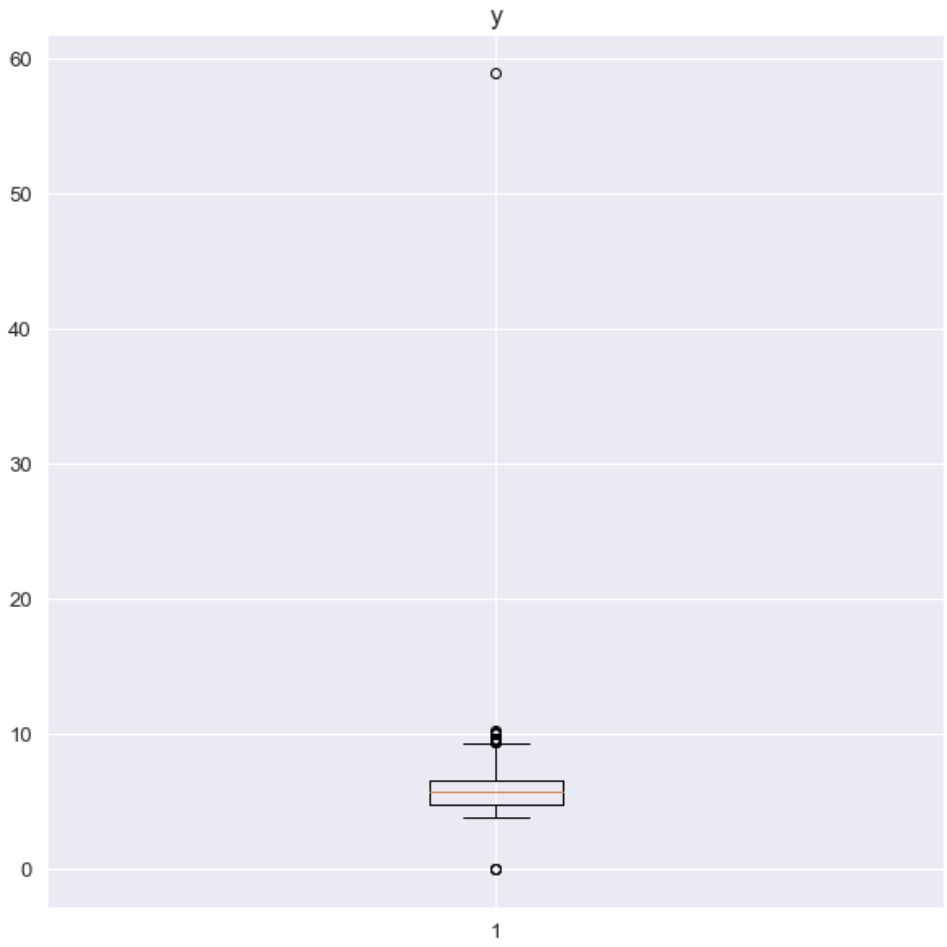

Fig No 15 – Boxplot of y

Boxplot of z -

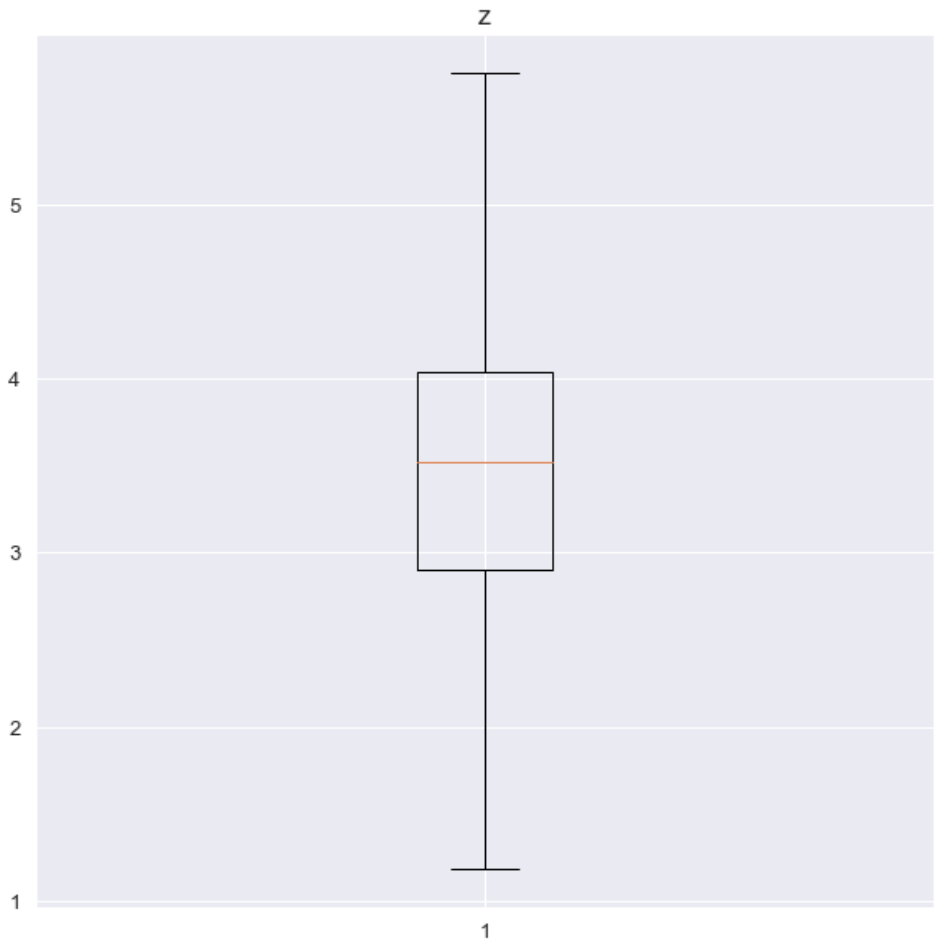

Fig No 16 – Boxplot of z

Boxplot of price -

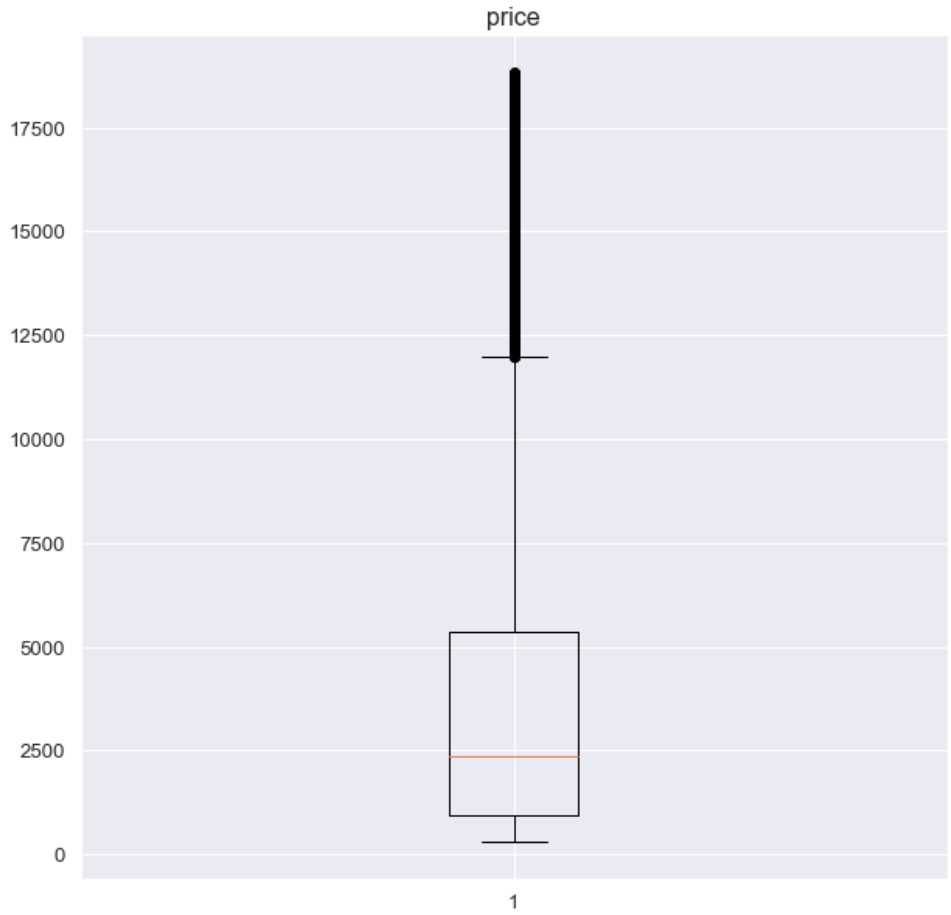

Fig No 17 – Boxplot of price

Duplicate values in the data set -

There are 34 duplicate rows in the dataset which we will drop.

Shape of the dataset Before dropping duplicates (26933, 24)
Shape of the dataset after dropping duplicates (26910, 24)

Problem 1.2

Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to

change them or drop them? Check for the possibility of combining the sub levels of a ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning.

Null values -

There are 697 null values present in depth column. Which we will drop.

Description of the data after imputing null values -

	carat	depth	table	x	y	z	price	...
count	26933	26933	26933	26933	26933	26933	26933	...
mean	0.79	61.75	57.4	5.7	5.7	3.5	3735.8	...
std	0.46	1.39	2.2	1.1	1.1	0.7	3468.2	...
min	0.2	50.8	51.5	1.1	1.1	1.2	326	...
25%	0.4	61.1	56	4.7	4.7	2.9	945	...
50%	0.7	61.8	57	5.7	5.7	3.5	2375	...
75%	1.05	62.5	59	6.6	6.5	4	5356	...
max	2.03	73.6	63.5	9.3	9.2	5.8	11972.5	...

Table No 3 – Description of data after imputing null values

Currently there are no values equal to zero in the dataset, at first there were 8 rows with zero values in them. values equal

to zero were dropped during the outlier treatment done to the dataset.

There were only 8 rows with null values which would not impact much on our dataset, as it is a large one, still the rows were treated.

In this case combining of sublevels is not necessary as each sub level holds its value. In the case of cut there are 3 sub categories fair, good, very good, premium and ideal these categories have their own meaning and combining them is not necessary.

In the case of color there are J, I, D, H, F, E and G, each color is unique therefore they cannot be combined in the same categories.

In the case of clarity there are various levels which effect the price accordingly. Each clarity of cubic zirconia is priced at a different level.

Problem 1.3

Encode the data for modeling. Split the data into train and test (70:30). Apply Linear regression model using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of predictions on Train and Test sets using

Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning.

We have encoded the data having string values for modeling.

The cut, color and clarity column have been encoded using one-hot encoding.

Our target variable will be price for this dataset.

We will split the data into train (70) and test (30).

Once we split the data we will use linear regression function to find the best fit model.

The Intercept for our model is 10315.14

Coefficients for each independent attribute -

The coefficient for carat is 9056.599600104772
The coefficient for depth is -107.20282009866389
The coefficient for table is -49.09783245656273
The coefficient for x is -2898.7463915588514
The coefficient for y is 2155.300639537146
The coefficient for z is -65.34696618719954
The coefficient for cut_Good is -2.0463630789890885e-12
The coefficient for cut_Ideal is 299.64740818624415
The coefficient for cut_Premium is 184.86371888602648
The coefficient for cut_Very Good is 0.0
The coefficient for color_E is 0.0
The coefficient for color_F is 0.0

The coefficient for color_G is 0.0
The coefficient for color_H is 0.0
The coefficient for color_I is 0.0
The coefficient for color_J is 0.0
The coefficient for clarity_IF is 0.0
The coefficient for clarity_SI1 is 0.0
The coefficient for clarity_SI2 is 0.0
The coefficient for clarity_VS1 is 0.0
The coefficient for clarity_VS2 is 0.0
The coefficient for clarity_VVS1 is 0.0
The coefficient for clarity_VVS2 is 0.0

R-squared is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

R Square on the training data is 88.38%

R Square on the test data is 88.87%

Root Mean Square Error (RMSE) -

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.

RMSE on training data is 1181.60

RMSE on test data is 1158.10

Linear Regression using Statsmodel -

Concatenate x and y into a single dataframe

```
Intercept       1.031514e+04
carat           9.056600e+03
depth          -1.072028e+02
table          -4.909783e+01
x              -2.898746e+03
y               2.155301e+03
z              -6.534697e+01
cut_Good        2.757996e-13
cut_Ideal       2.996474e+02
cut_Premium     1.848637e+02
cut_verygood    0.000000e+00
color_E         0.000000e+00
color_F         0.000000e+00
color_G         0.000000e+00
color_H         0.000000e+00
color_I         0.000000e+00
color_J         0.000000e+00
clarity_SI1     0.000000e+00
clarity_SI2     0.000000e+00
clarity_VS1     0.000000e+00
clarity_VS2     0.000000e+00
clarity_VVS1    0.000000e+00
clarity_VVS2    0.000000e+00
```

Summary of the model -

```
OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.884
```

Model:	OLS	Adj. R-squared:	0.884
Method:	Least Squares	F-statistic:	1.791e+04
Date:	Sun, 19 Jun 2022	Prob (F-statistic):	0.00
Time:	10:47:00	Log-Likelihood:	-1.6013e+05
No. Observations:	18853	AIC:	3.203e+05
Df Residuals:	18844	BIC:	3.203e+05
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.032e+04	830.170	12.425	0.000	8687.928	1.19e+04
carat	9056.5996	104.865	86.364	0.000	8851.054	9262.145
depth	-107.2028	10.637	-10.078	0.000	-128.052	-86.354
table	-49.0978	5.244	-9.363	0.000	-59.376	-38.819
x	-2898.7464	163.191	-17.763	0.000	-3218.616	-2578.877
y	2155.3006	163.523	13.180	0.000	1834.780	2475.821
z	-65.3470	130.511	-0.501	0.617	-321.160	190.466
cut_Good	2.758e-13	1.26e-14	21.906	0.000	2.51e-13	3e-13
cut_Ideal	299.6474	24.052	12.458	0.000	252.503	346.791
cut_Premium	184.8637	24.663	7.496	0.000	136.523	233.205
cut_verygood	0	0	nan	nan	0	0
color_E	0	0	nan	nan	0	0
color_F	0	0	nan	nan	0	0
color_G	0	0	nan	nan	0	0
color_H	0	0	nan	nan	0	0
color_I	0	0	nan	nan	0	0
color_J	0	0	nan	nan	0	0
clarity_SI1	0	0	nan	nan	0	0
clarity_SI2	0	0	nan	nan	0	0
clarity_VS1	0	0	nan	nan	0	0
clarity_VS2	0	0	nan	nan	0	0
clarity_VVS1	0	0	nan	nan	0	0
clarity_VVS2	0	0	nan	nan	0	0

Omnibus:	5502.522	Durbin-Watson:	2.015
Prob(Omnibus):	0.000	Jarque-Bera (JB):	29165.021
Skew:	1.308	Prob(JB):	0.00
Kurtosis:	8.503	Cond. No.	inf

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 0. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.

Prediction on the test data -

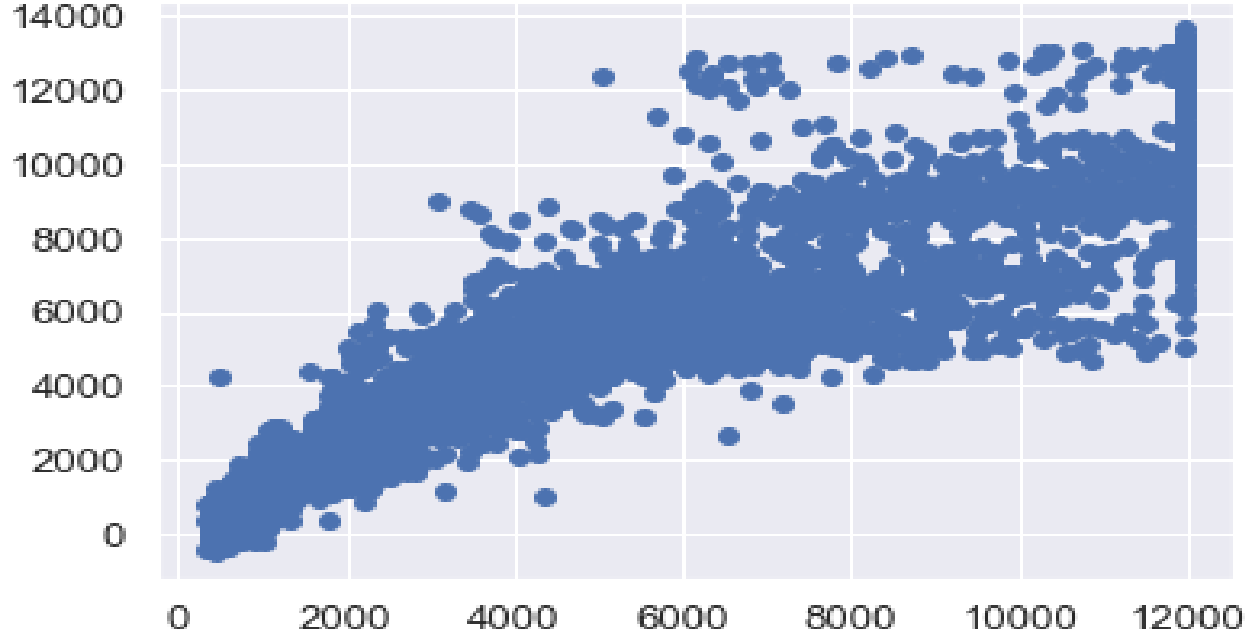

Fig No 18 – Prediction of the data

Problem 1.4

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

From the analysis we can see that the most impact on price was from the carat variable around 9.06

From the analysis I would recommend the business to focus more of the high carat stones as they are more profitable.

The company must also focus on the clarity of the stones, high clarity stones will also impact the price factor.

The color of the stones doesn't affect much of the price focusing on the carat and clarity will increase the business of the company.

Problem 2: Logistic regression and LDA

Executive Summary -

A tour and travel agency which deals in selling holiday packages wants us to help them in predicting whether an employee will opt for the package or not on the basis of information provided. And also find the important factors on the basis of which the company will focus on particular employees to sell their packages.

Introduction -

The purpose of the problem is to explore the dataset. Do the exploratory analysis. The data consists details of 872 employees and their details such as holiday package, salary, age, education children and foreign. We will be using Logistic Regression and LDA to solve the problem.

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 -

	Unnamed:0	Holiday_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 No 4 – Sample of the dataset

Dataset has 8 variables with information about the employees.

Problem 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 analysis.

Exploratory data analysis (EDA) -

Types of Variables in the dataset -

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

There are 872 rows and 7 columns in the dataset. 2 columns are of object data type and 5 columns are of integer data type.

Missing values in the dataset -

```
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
```

From the above result we can see that there are no missing values in the data set.

Univariate Analysis -

Describing the data -

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Holliday_Package	872	2	no	471	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Salary	872	NaN	NaN	NaN	47729.17	23418.67	1322	35324	41903.5	53469.5	236961
age	872	NaN	NaN	NaN	39.95	10.55	20	32	39	48	62
educ	872	NaN	NaN	NaN	9.30	3.03	1	8	9	12	21
No_young_children	872	NaN	NaN	NaN	0.31	0.61	0	0	0	0	3
No_older_children	872	NaN	NaN	NaN	0.98	1.09	0.0	0.0	1.0	2.0	6.0
foreign	872	2	no	656	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table No 5 – Describing the dataset

From the descriptive we can see the mean/median of the variables.

Unique Values of categorical Variables -

HOLLIDAY_PACKAGE: 2
yes 401
no 471
Name: Holliday_Package, dtype: int64

FOREIGN: 2
yes 216
no 656

Histogram of Salary -

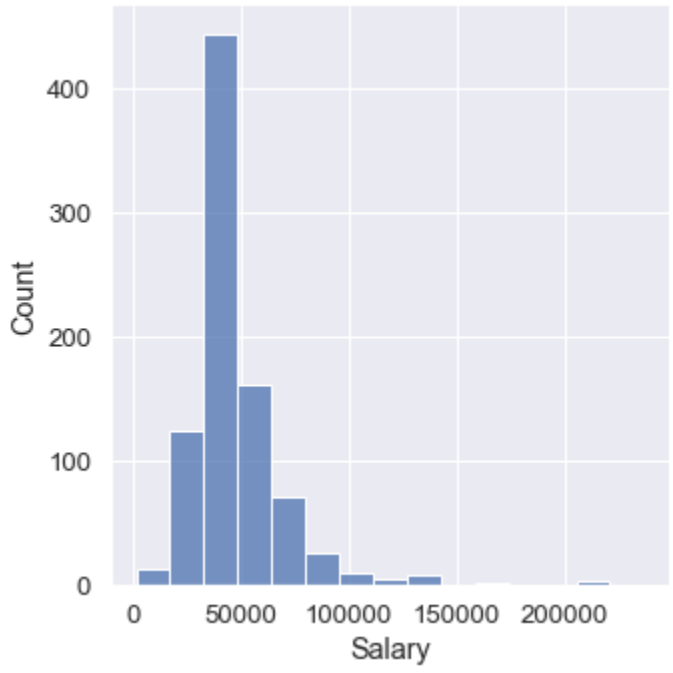

Fig No 19 – Histogram of salary

Histogram of age -

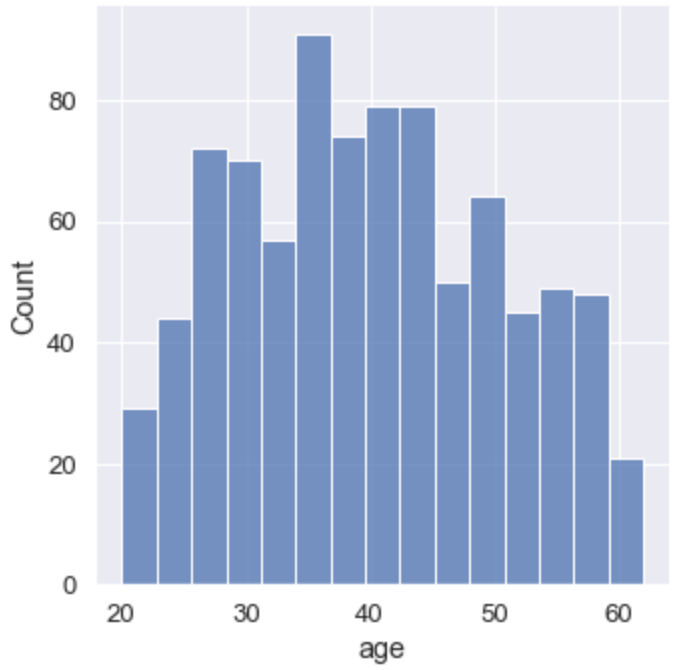

Fig No 20 – Histogram of age

Histogram of educ -

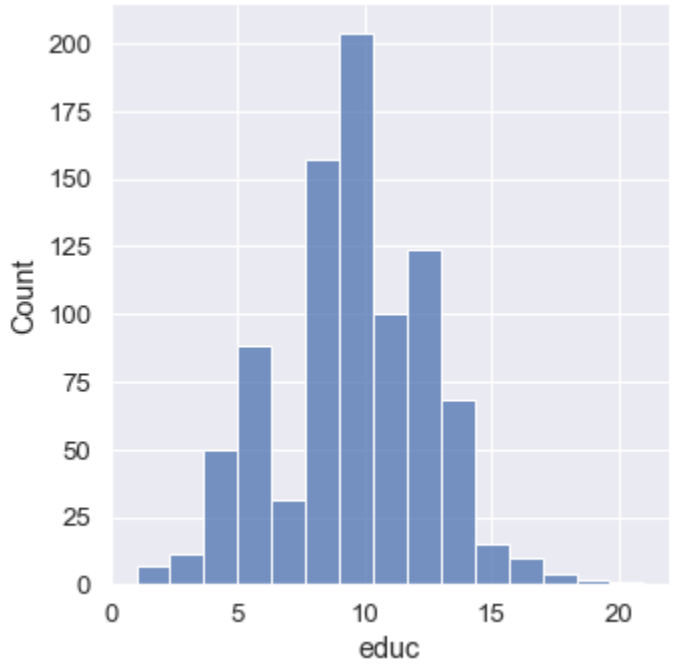

Fig No 21 – Histogram of edu

Histogram of no_young_children -

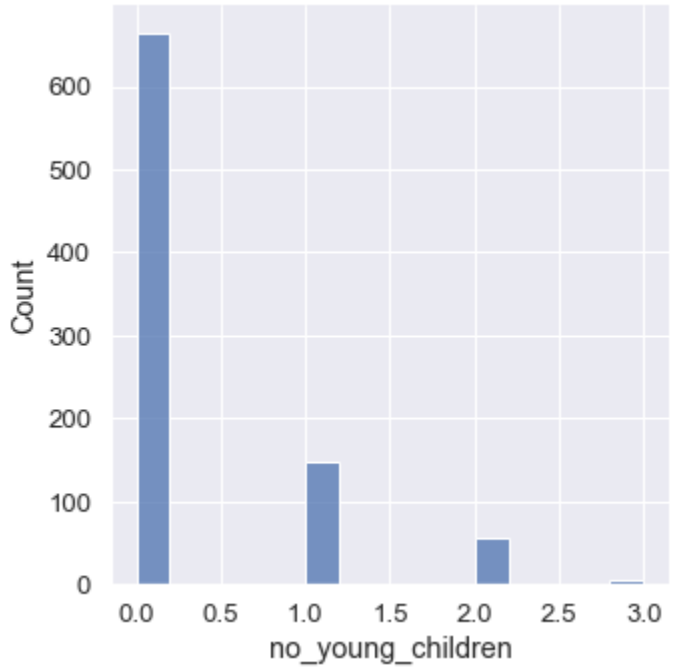

Fig No 22 – Histogram of no_young_children

Histogram of no_older_children -

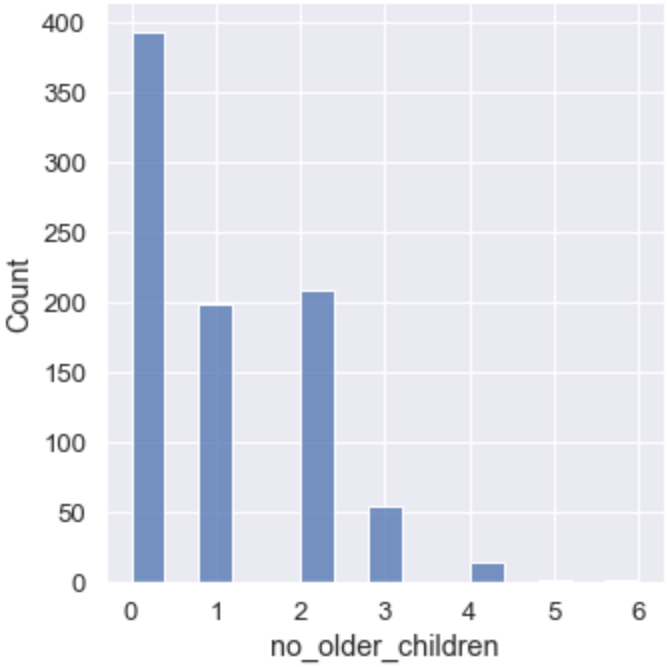

Fig No 23 – Histogram of no_older_children

Bivariate Analysis -

Correlation Plot -

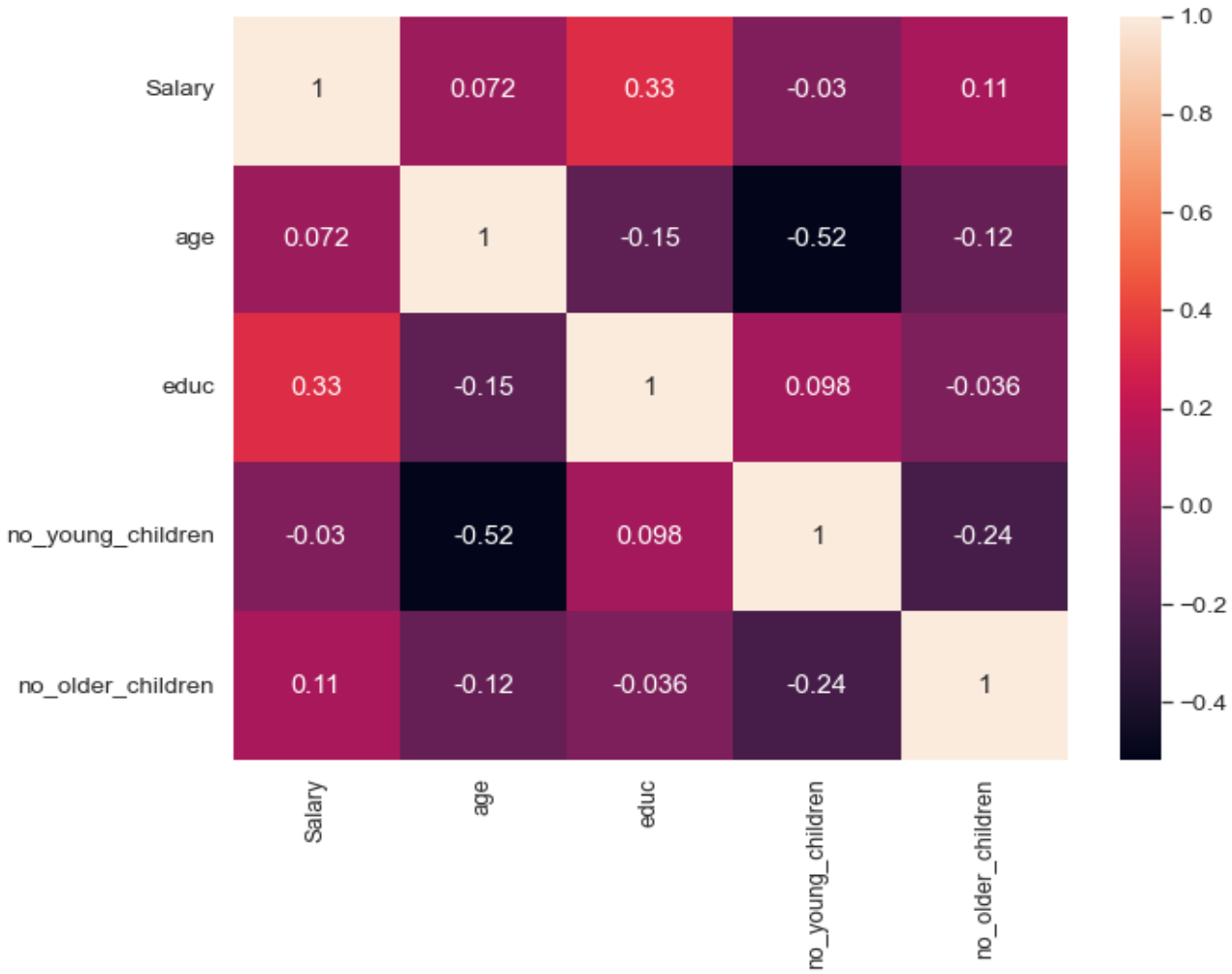

Fig No 24 – Correlation Plot

From the correlation plot, we can see that various attributes are highly correlated to each other. Correlation values near to 1 or –1 are highly positively correlated and highly negatively correlated respectively. Correlation values near to 0 are not correlated to each other.

Pair plot -

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

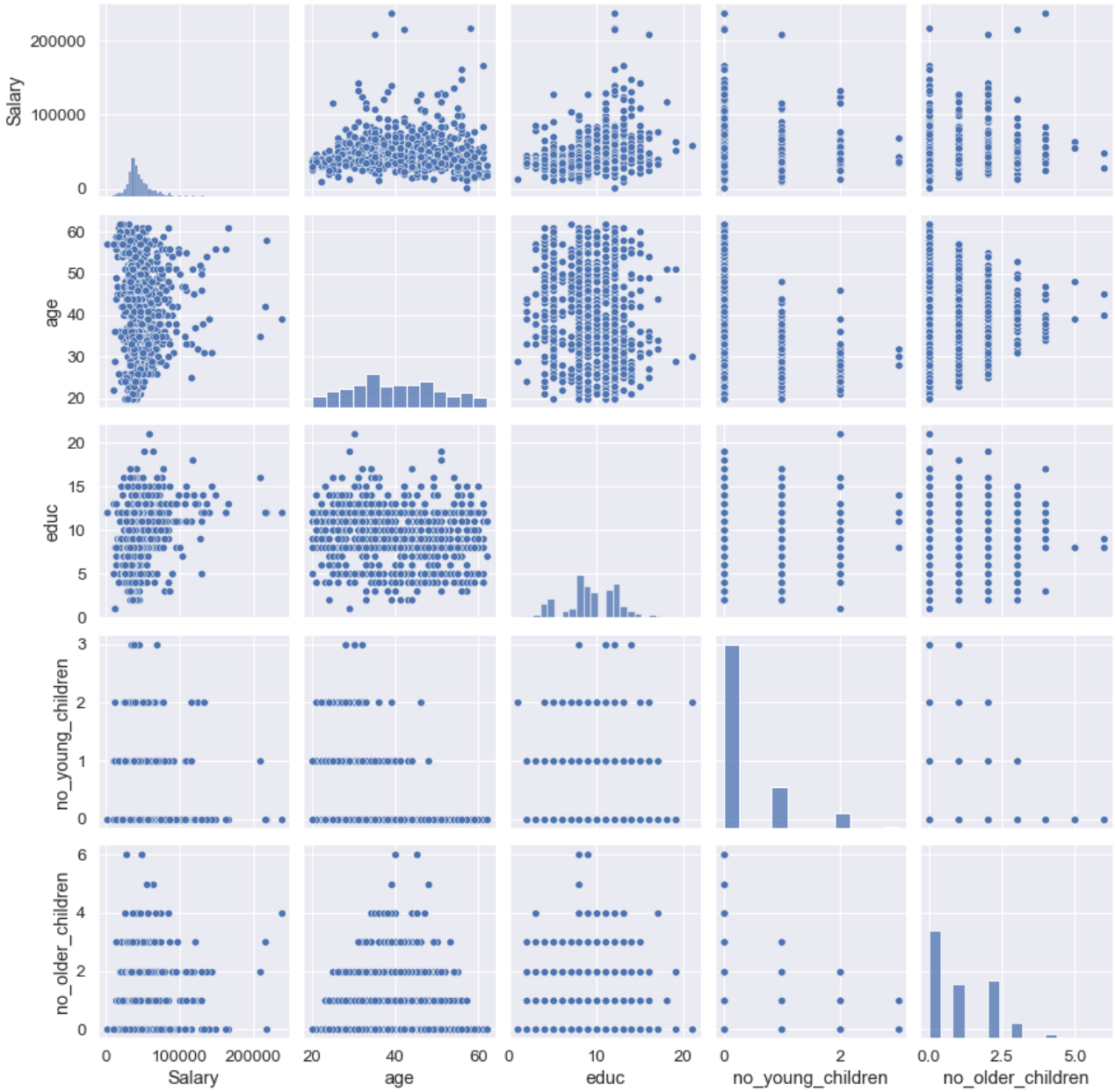

Fig No 25 – Pairplot

Boxplot -

A boxplot is a graph that gives a good indication of how the values in the data are spread out.

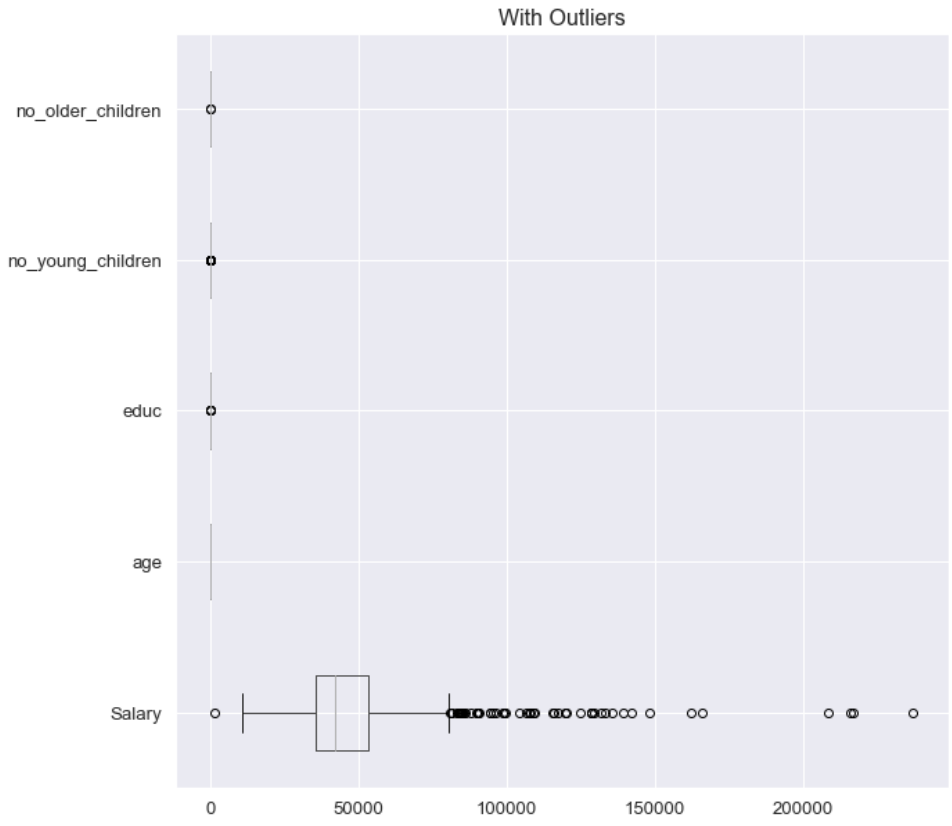

Fig No 26 - Boxplot

Boxplot of Salary -

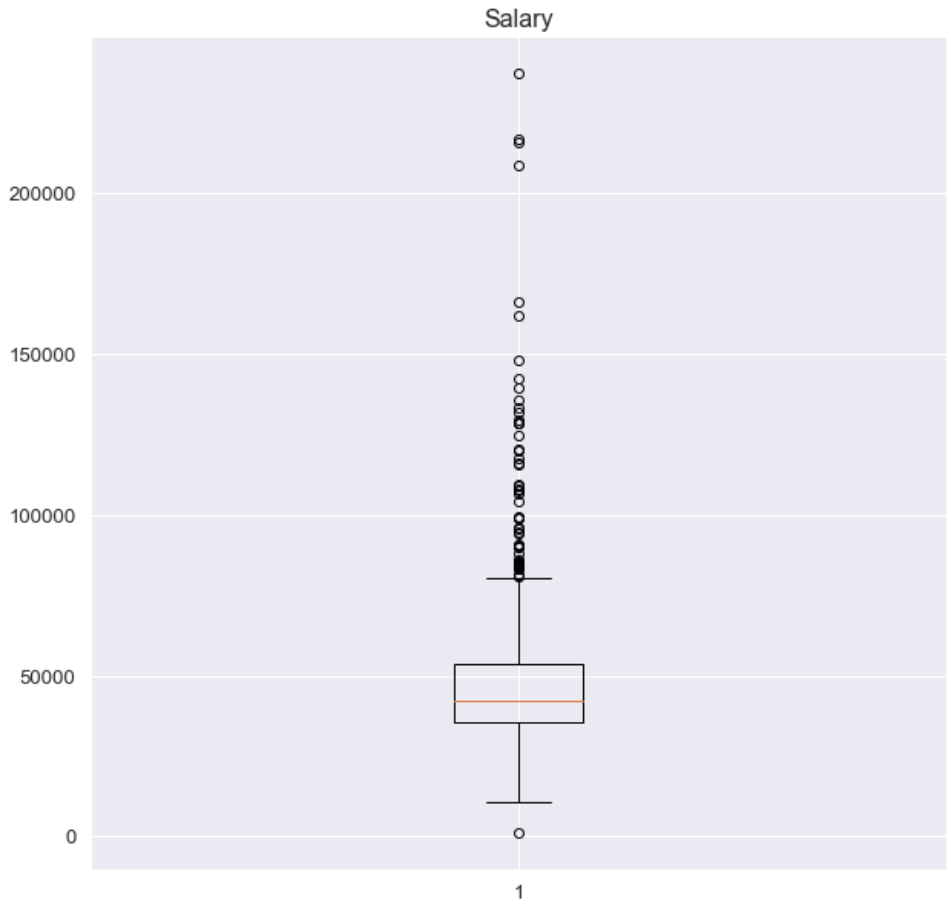

Fig No 27 – Boxplot of Salary

Boxplot of age -

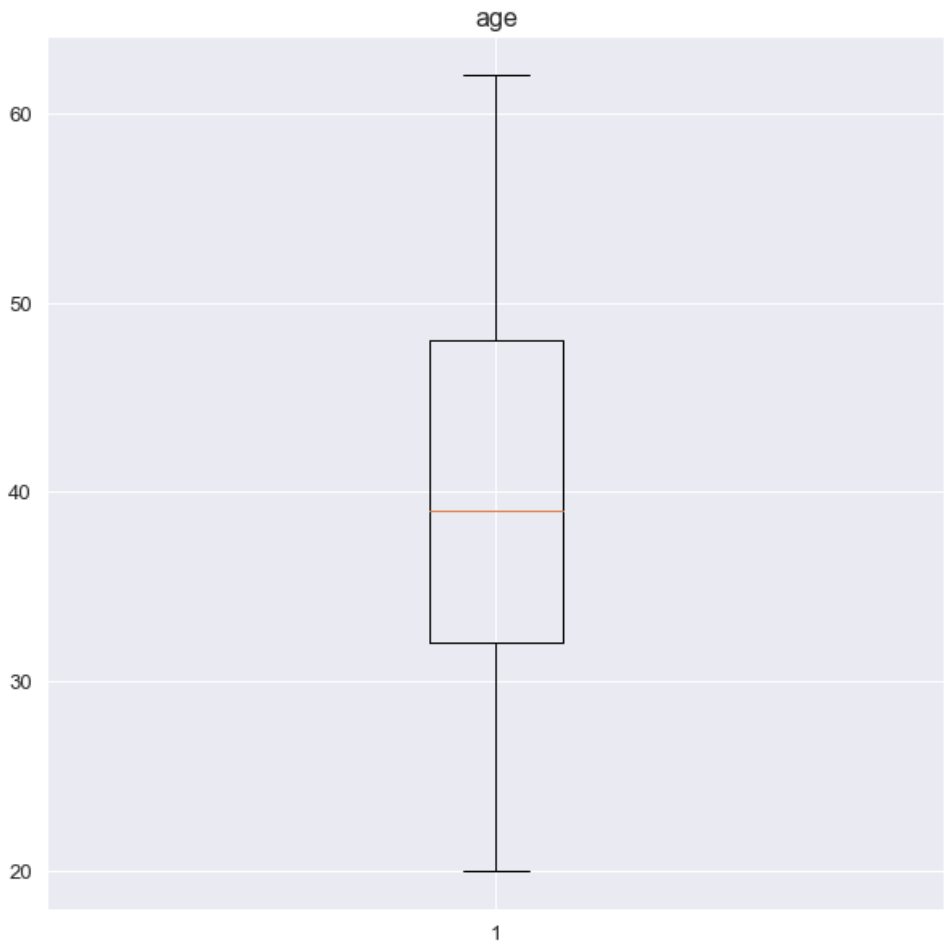

Fig No 28 – Boxplot of age

Boxplot of educ -

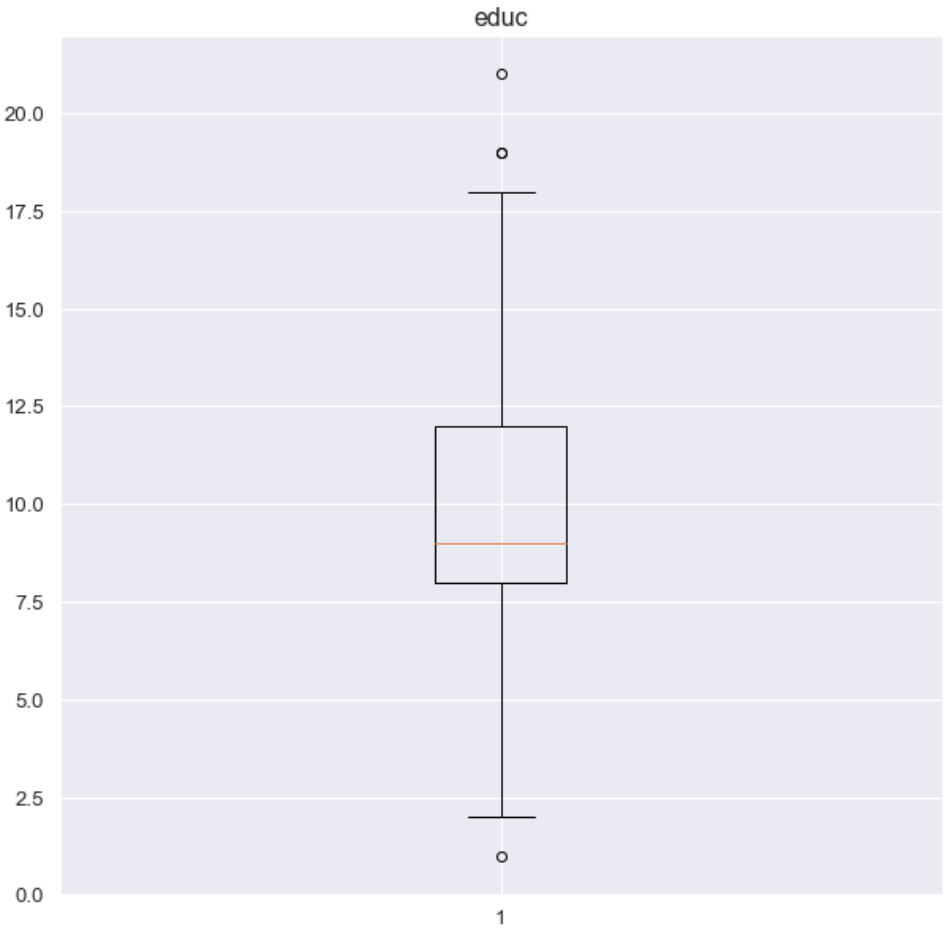

Fig No 29 – Boxplot of educ

Boxplot of no_young_children -

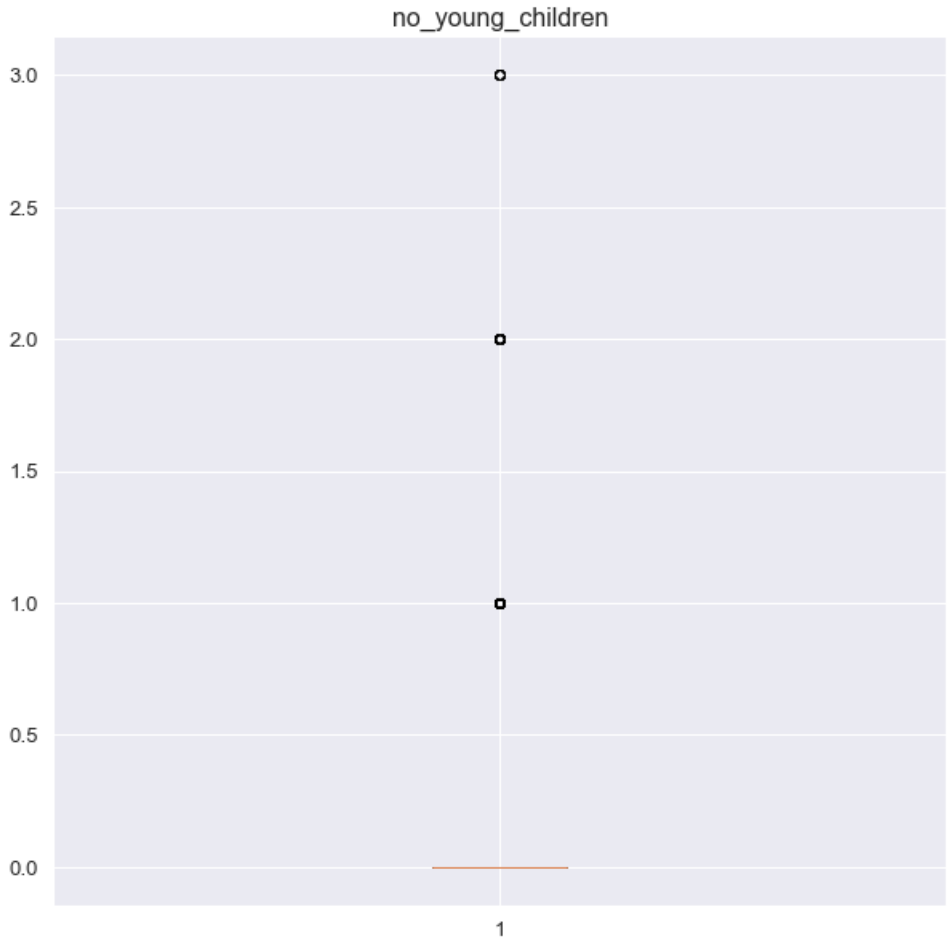

Fig No 30 – Boxplot of no_young_children

Boxplot of no_older_children -

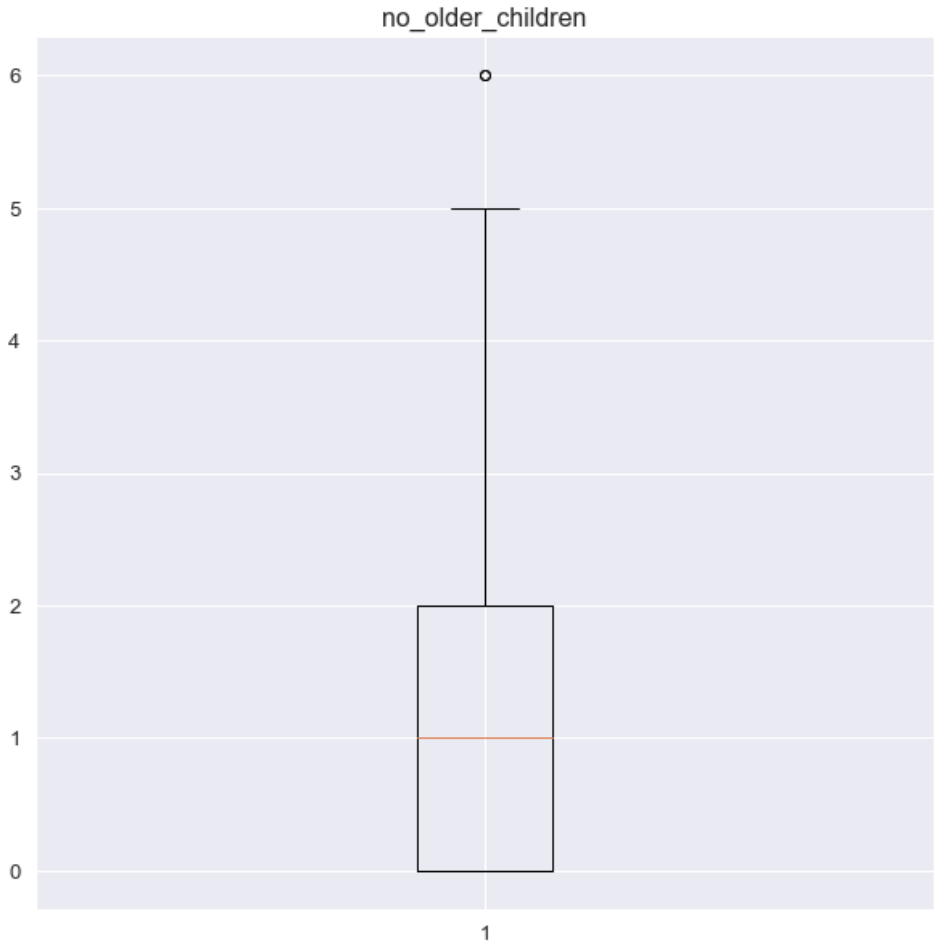

Fig No 31 – Boxplot of no_older_children

Duplicate values in the data set -

There are no Duplicate values in the data set.

Problem 2.2

Do not scale the data. Encode the data for modelling. Data split: split the data into train and test (70:30). Apply Logistic regression and LDA.

We will Encode the Holliday_Package and foreign columns.

After encoding this is how data looks -

	Holliday_ Package	Salary	age	edu	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 No 6 – Data after encoding

We have split the data into 70:30

Once we split the data we apply Logistic Regression and LDA.

Problem 2.3

Performance Metrics: Check the performance on train and test sets using accuracy, confusion matrics, 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.

Logistic regression Model -

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable.

Accuracy of training Model – 66.7%

Accuracy of Test model – 65.3%

The accuracy scores are almost similar.

AUC ROC for training data -

AUC: 0.735

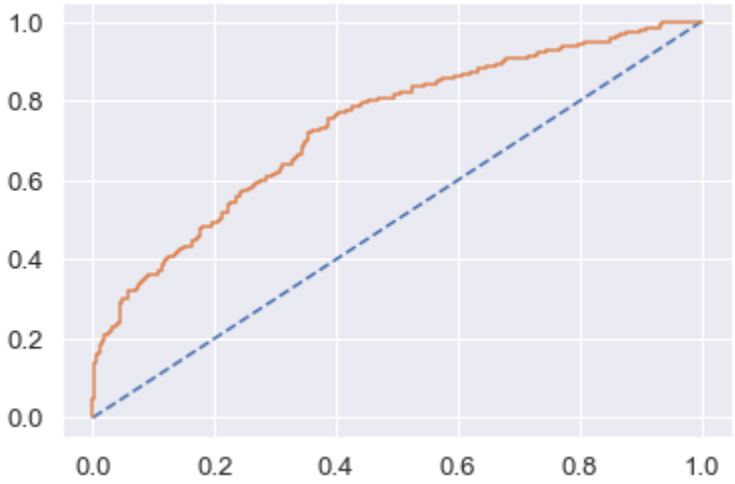

Fig No 32 – AUC ROC for training data

AUC ROC for test data -

AUC: 0.735

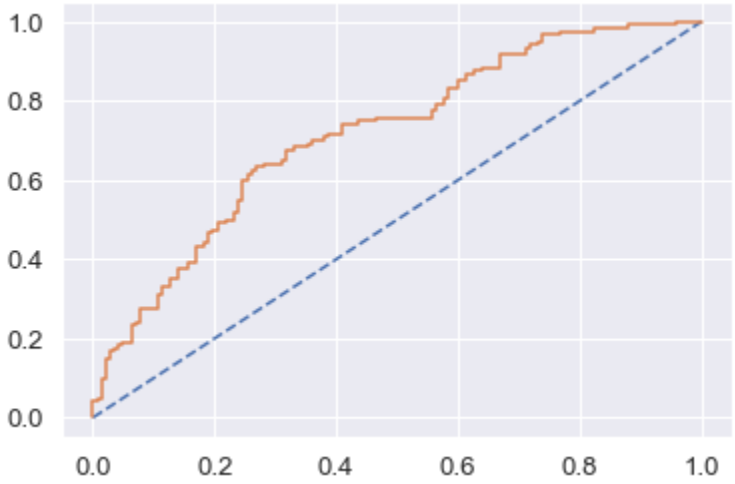

Fig No 33 – AUC ROC for test data

Confusion Matrics for training data -

array ([[244, 85],
        [118, 163]]

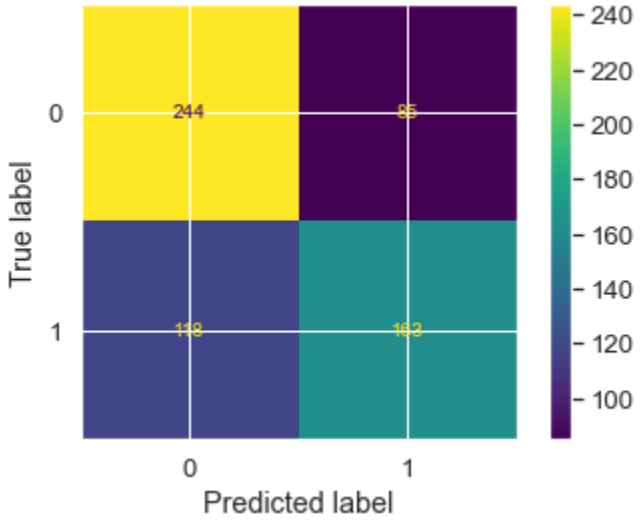

Fig No 34 – Confusion Matrics for training data

Classification report for 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

Confusion Matrics for test data -

array ([[109, 33],
        [58, 62]]

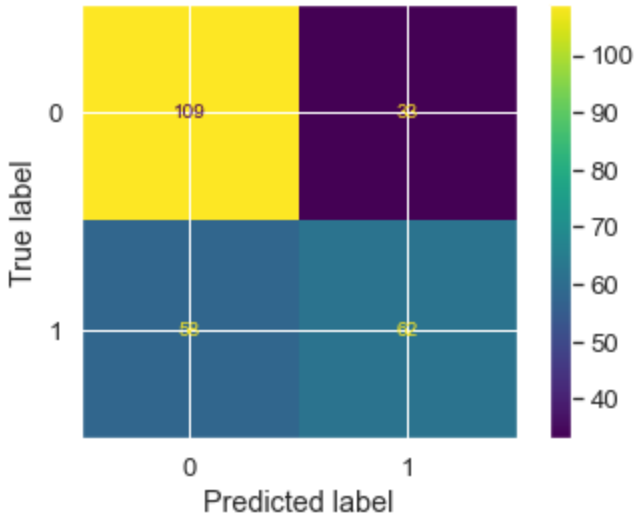

Fig No 35 – Confusion Matrics for test data

Classification report for test 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

Linear Discriminant Analysis (LDA) -

Linear discriminant analysis is primarily used here to reduce the number of features to a more manageable number before classification.

Accuracy of training Model –66 %

Accuracy of Test model – 64%

Confusion Matrics -

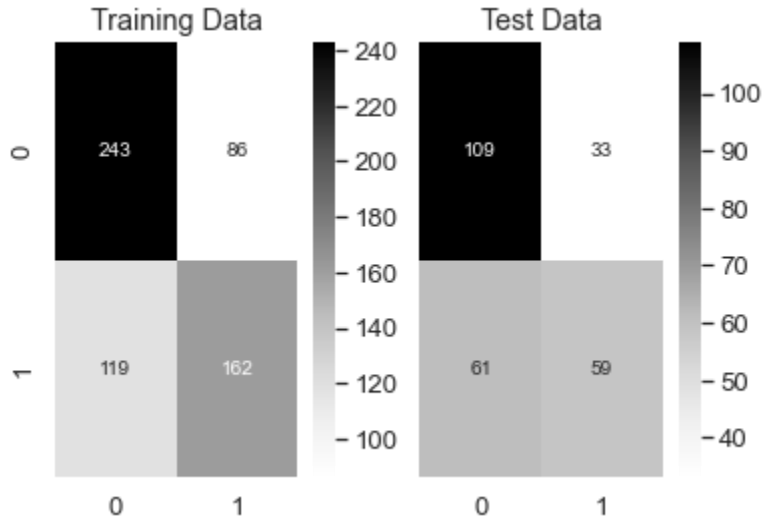

Fig No 36 – Confusion Matrics

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

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

AUC and ROC for the training data -

AUC: 0.733

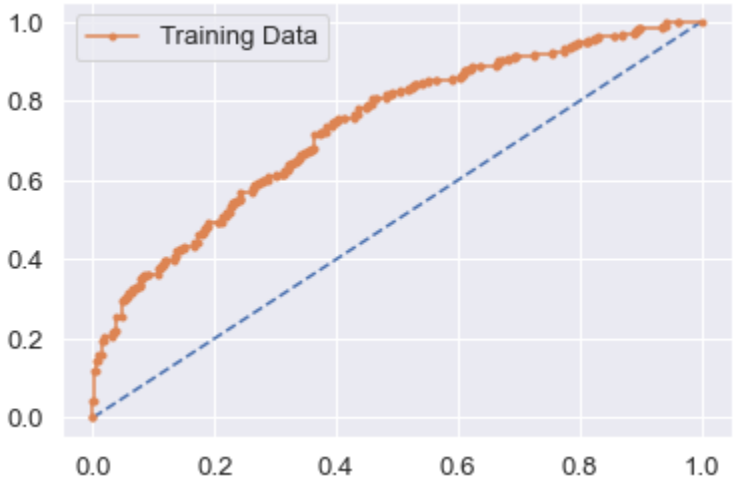

Fig No 37 – AUC ROC for training data

AUC and ROC for test data -

AUC: 0.714

Fig No 38 – AUC ROC for test data

Comparing the classification report of both the models, the accuracy of Logistic regression model appears slightly better than the LDA model. Also, the F1 score is slightly higher in Logistic regression model, therefore I think the Logistic Regression model is the best compare to LD. Logistic regression model is also easy to interpret and efficient to train.

Problem 2.4

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

Looking at the data the total number of employees accepting the holiday package are 401 and saying no are 471.

Most of the employee having young children are saying no to the holiday package.

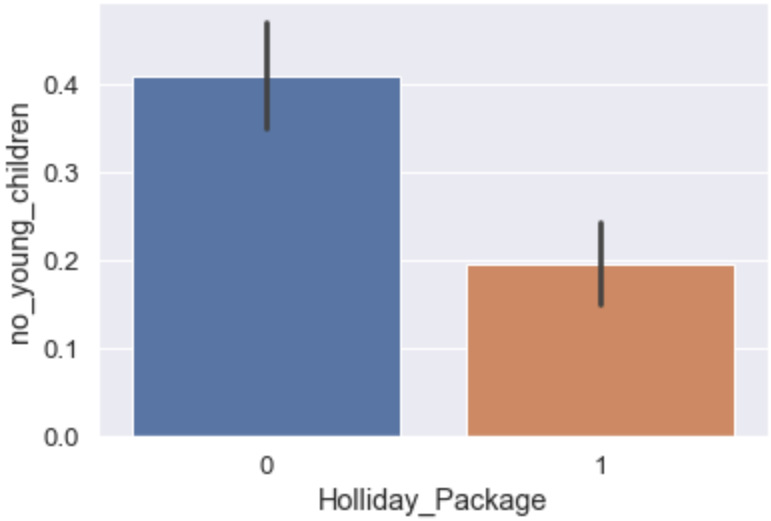

Fig No 39 – Boxplot

More employees opt for the holiday package with older children.

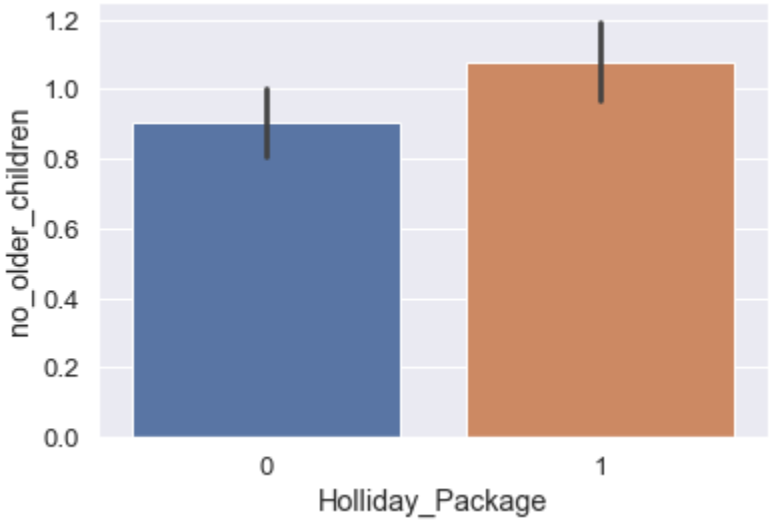

Fig No 40 – Boxplot

Therefore, the company should not focus on the employees with younger children as it is likely they will say no.

The company must target employees with older children as there is a higher chance of them saying yes.

More number of foreigners opt for the holiday package.

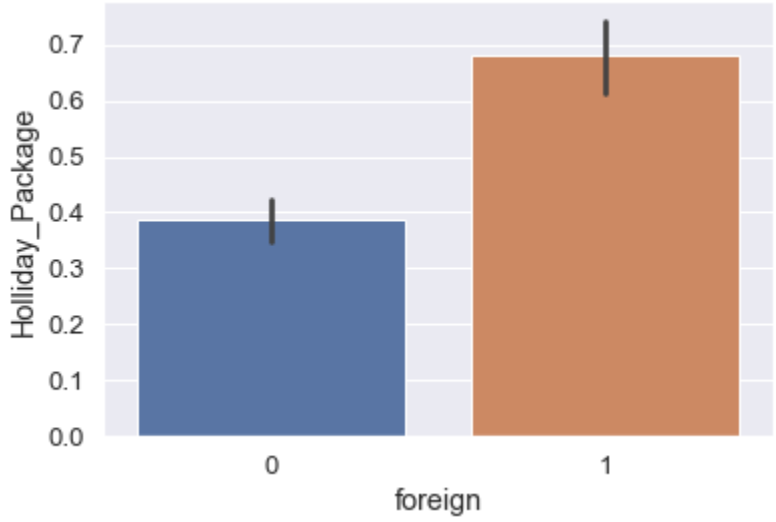

Fig No 41 – Boxplot

The company must focus on targeting the foreigners as they are most likely to say yes.

People with Higher salary tend to say no to the holiday packages whereas people with average salary say yes.

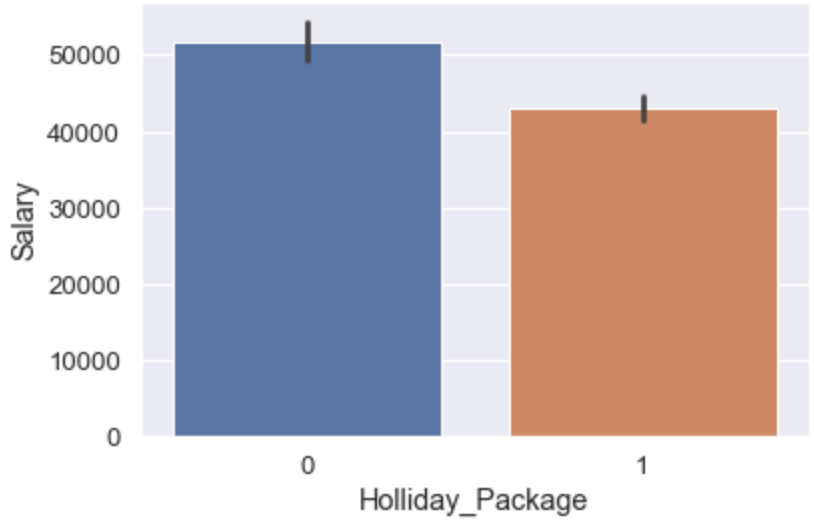

Fig No 42 – Boxplot

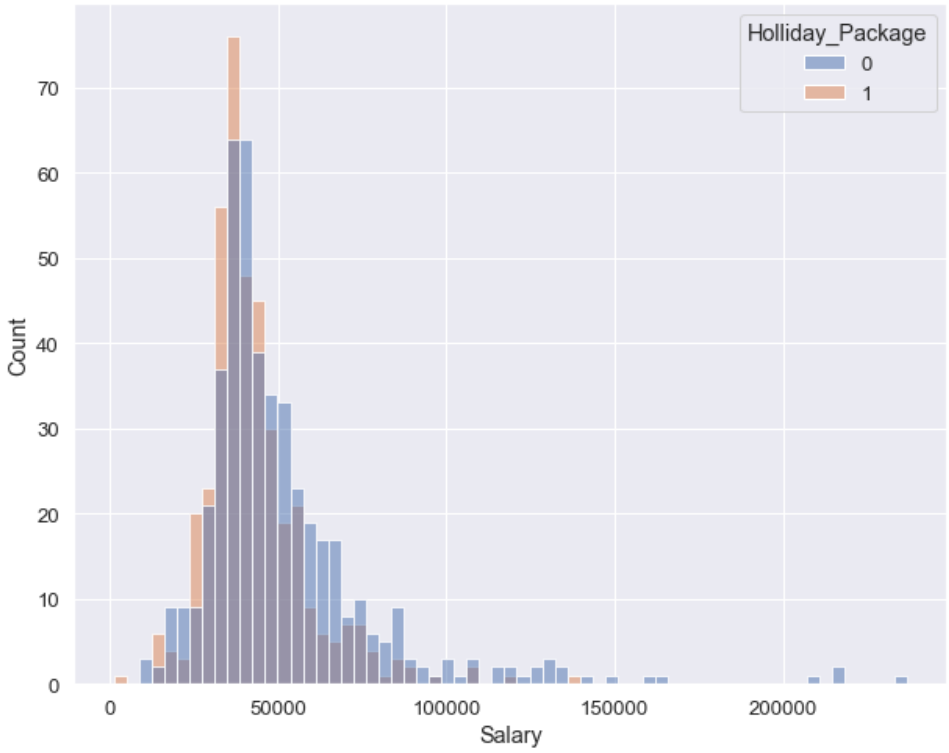

Fig No 43 – Histogram for Salary

The company must focus on the average salary earning employees around 40 to 50K.

* * * * * * *