Problem 1: Linear Regression

Gem Stones co ltd company which is a cubic zirconia manufacturer wants to predict 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.

Data Dictionary:

Variable Name	**Description**
Carat	Carat weight of the cubic zirconia.
Cut	Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal.
Color	Colour of the cubic zirconia.With D being the best and J the worst.
Clarity	cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst, FL = flawless, I3= level 3 inclusions) FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, I2, I3
Depth	The Height of a cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter.
Table	The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter.
Price	the Price of the cubic zirconia.
X	Length of the cubic zirconia in mm.
Y	Width of the cubic zirconia in mm.
Z	Height of the cubic zirconia in mm.

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

Top 5 rows of the data set-

	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

A summary of the dataset

```
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)
memory usage: 2.3+ MB
```

Inferences-

i) The dataset has total of 10 variables-
- 6 variables(carat, depth, table, x, y, z) are float and continuous.
- 3 variables(cut, color, clarity) are objects and categorical.
- The price variable is integer and continuous.

ii) Shape of the dataset -
- Number of columns- 11

- Number of rows- 26967

iii) There are 697 null values in depth variable.
iv) There are no duplicates.

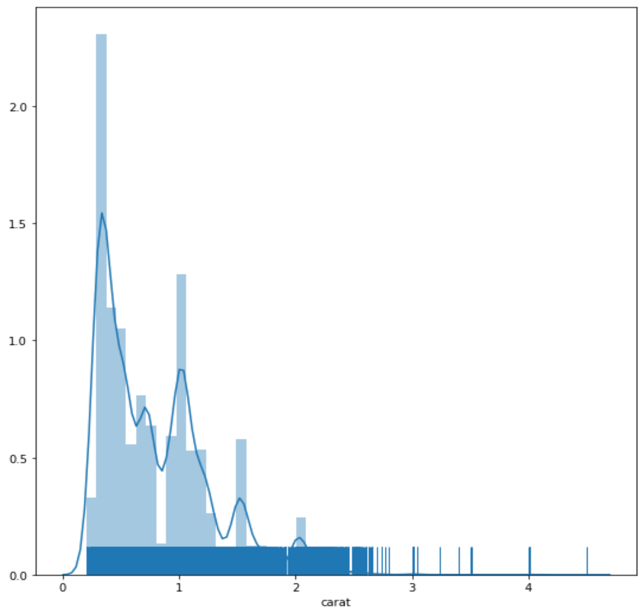
Univariate Analysis-

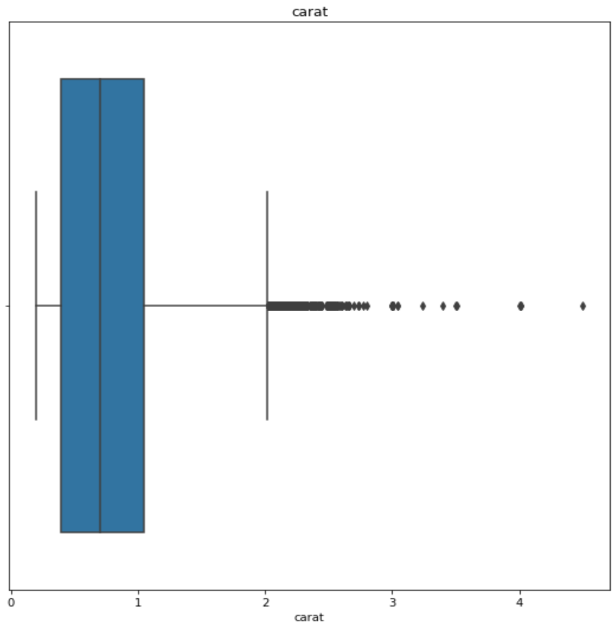
i) For continuous variables- using distribution plot and box plot

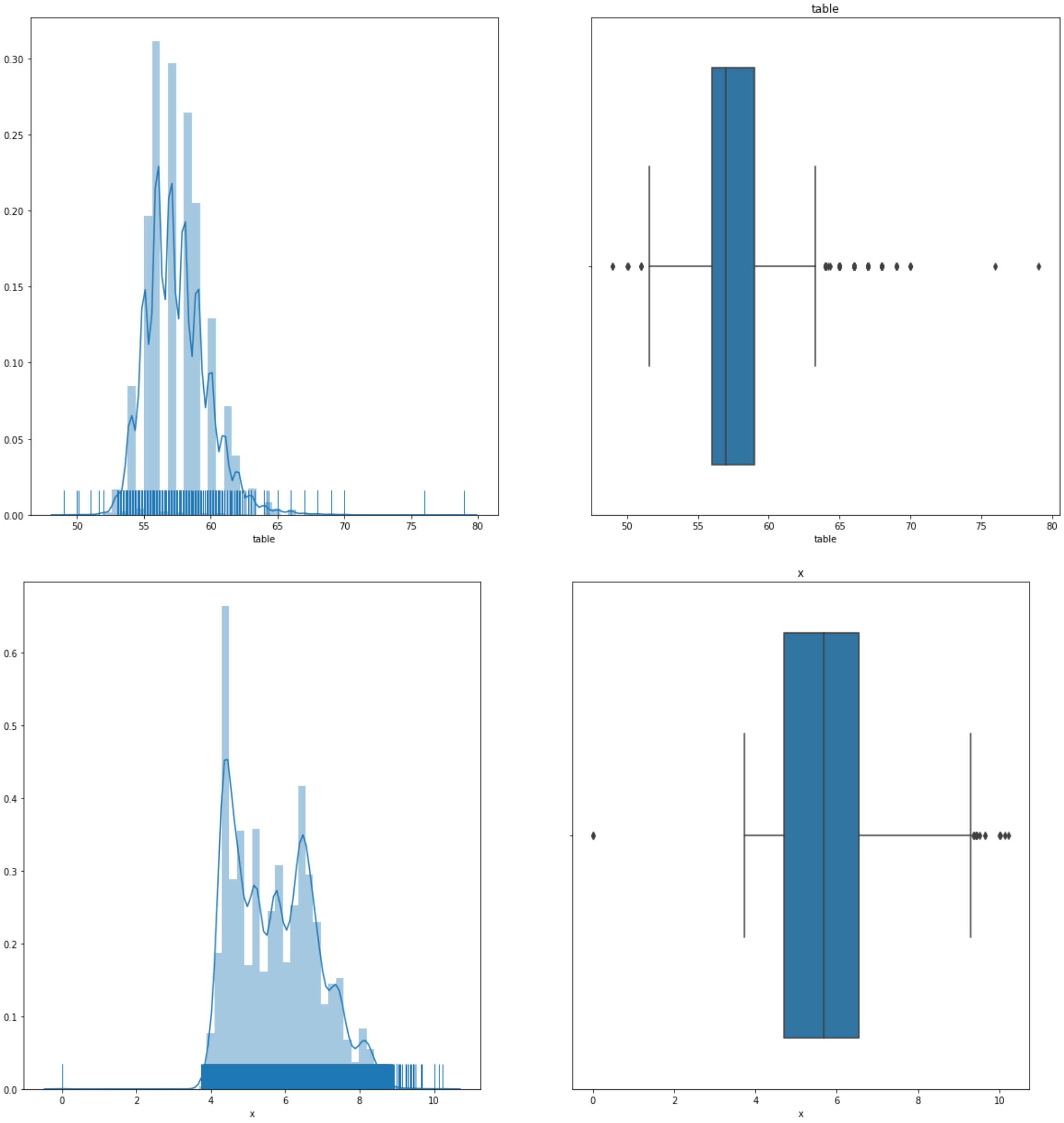
Summary statistics of continuous variables

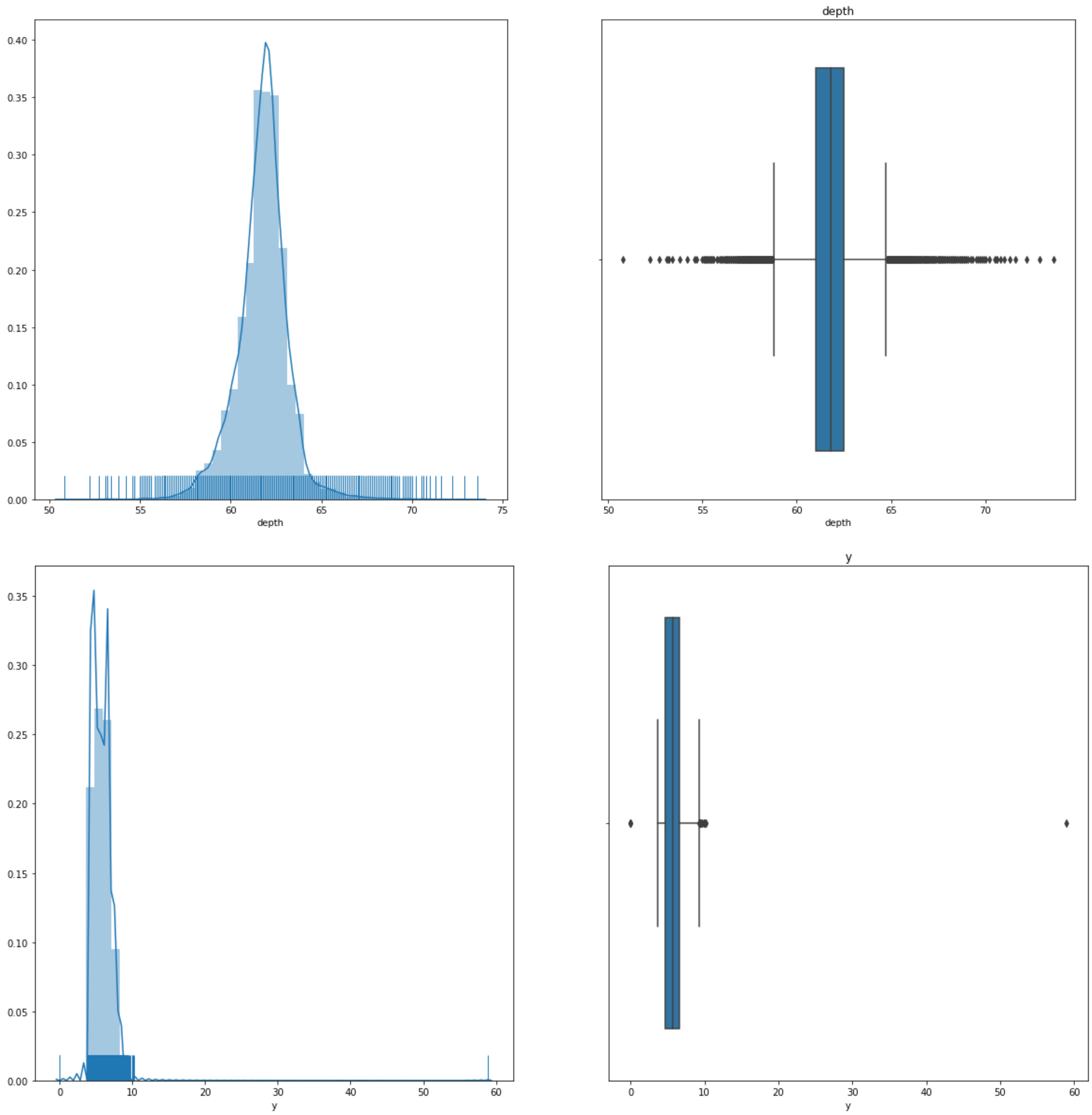
	count	mean	std	min	25%	50%	75%	max
carat	26967.0	0.798375	0.477745	0.2	0.40	0.70	1.05	4.50
depth	26270.0	61.745147	1.412860	50.8	61.00	61.80	62.50	73.60
table	26967.0	57.456080	2.232068	49.0	56.00	57.00	59.00	79.00
x	26967.0	5.729854	1.128516	0.0	4.71	5.69	6.55	10.23
y	26967.0	5.733569	1.166058	0.0	4.71	5.71	6.54	58.90
z	26967.0	3.538057	0.720624	0.0	2.90	3.52	4.04	31.80
price	26967.0	3939.518115	4024.864666	326.0	945.00	2375.00	5360.00	18818.00

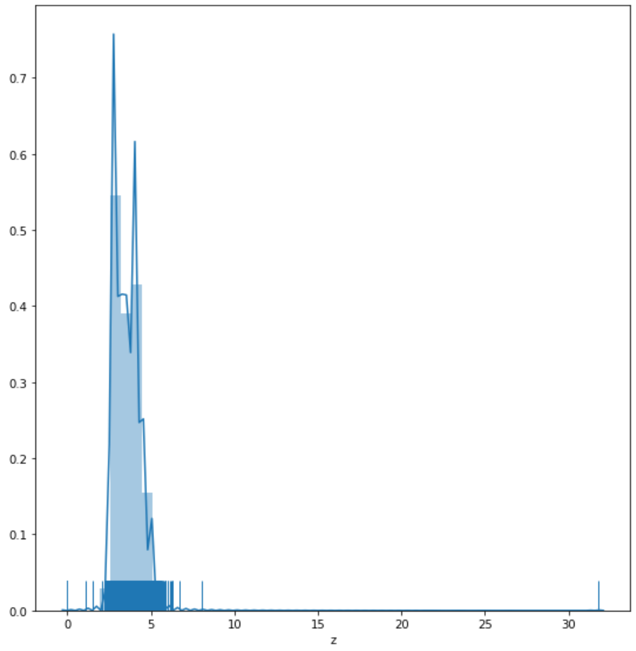
Distribution plots and Box plots

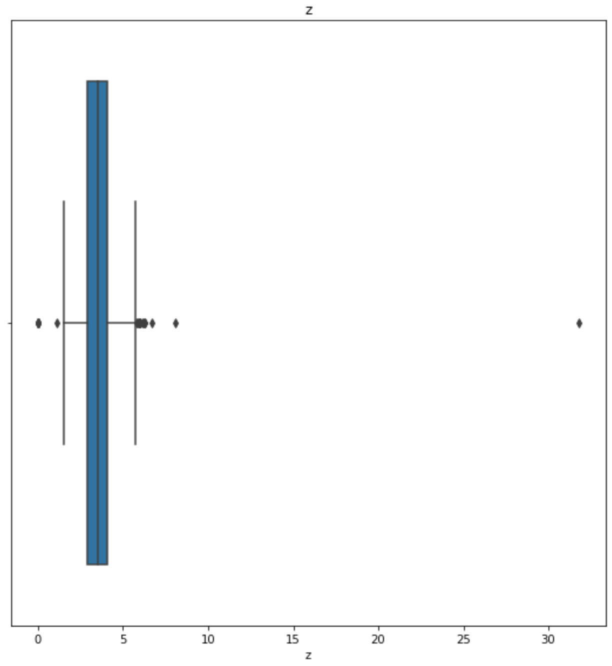


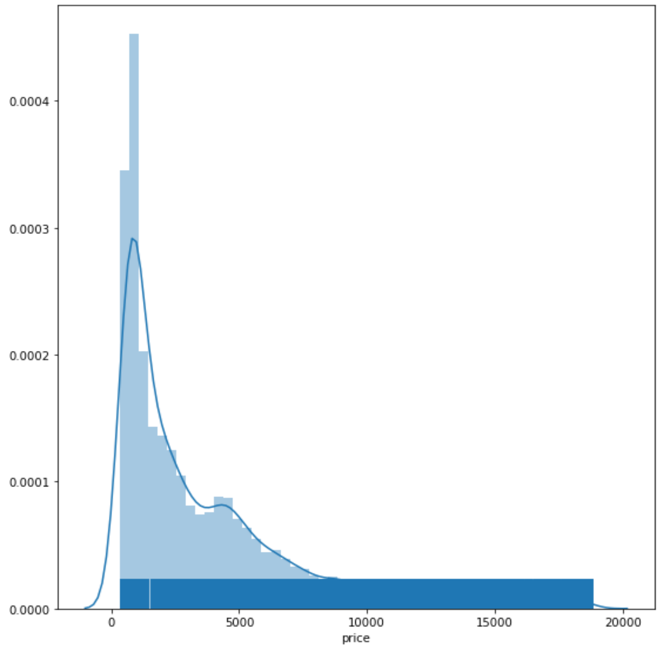


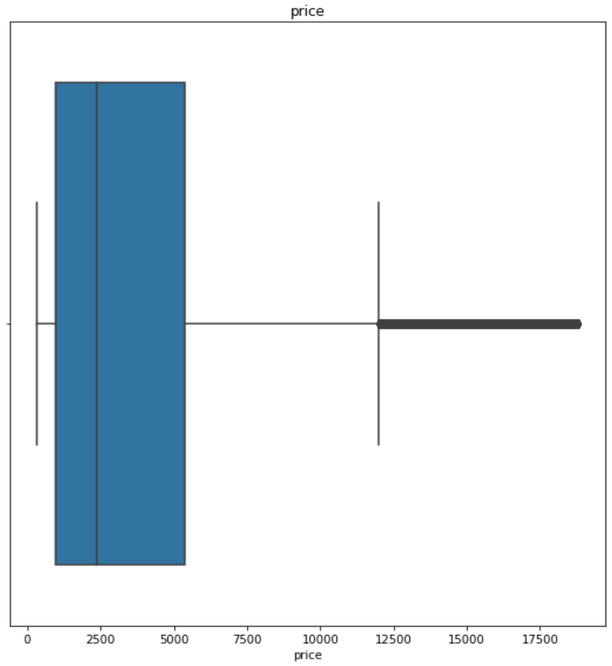

0.30
0.25
0.20
0.15
0.10
0.05
0.00
50
55
60
65
70
75
80
table
table
50
55
60
65
70
75
80
table
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
2
4
6
8
10
x
x
0
2
4
6
8
10
x


0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
50
55
60
65
70
75
depth
depth
50
55
60
65
70
depth
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
10
20
30
40
50
60
y
y
0
10
20
30
40
50
60
y









Kurtosis and Skewness in the dataset

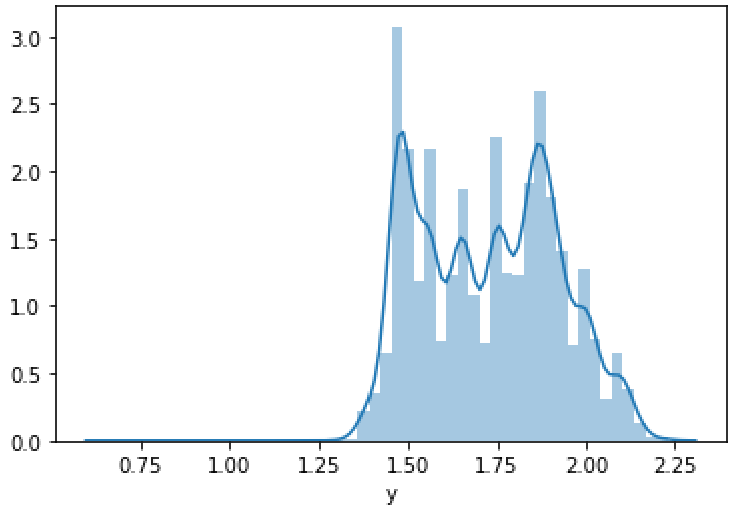
	Kurtosis	skewness
Carat	1.215364	1.116481
Depth	3.674431	-0.028618
Table	1.582166	0.765758
X	-0.657825	0.387986
Y	159.291616	3.850189
Z	87.006350	2.568257
Price	2.148617	1.618550

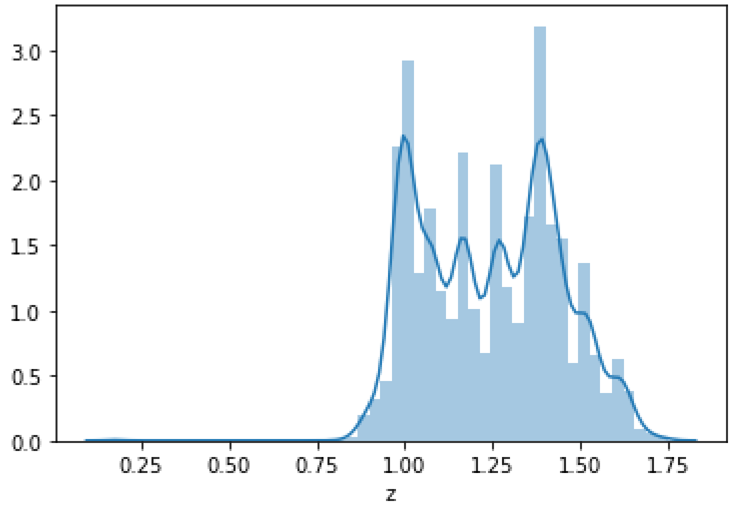
Inference-

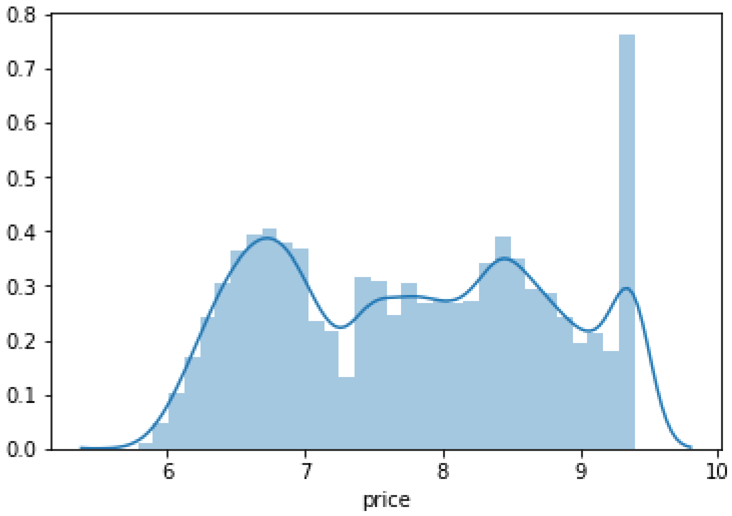
i) Carat,depth,table and x variables are almost normally distributed.
ii) y, z , and price variables are positively skewed.
iii) All variables are having outliers.

As one of the assumptions of linear regression is that data is normally distributed, we are going to correct skewness and kurtosis by using log function. As outliers can affect the outcome of the model we are going to treat outliers by capping the outliers values between -1.5 and +1.5 IQR(interquartile range).

Distribution plot after correction of skewness and kurtosis-



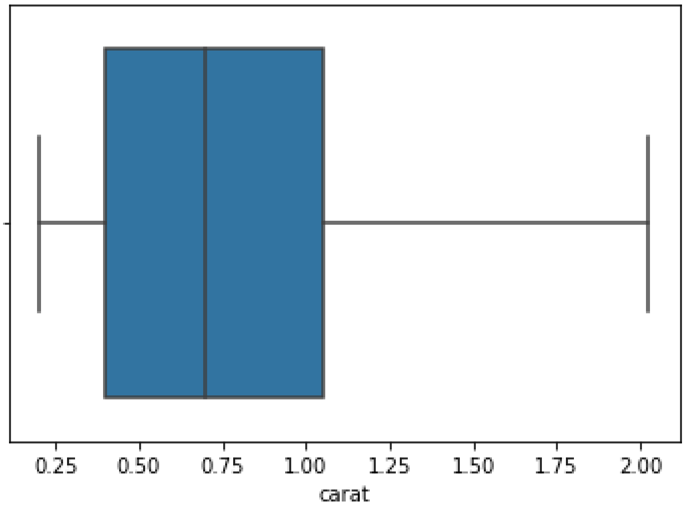


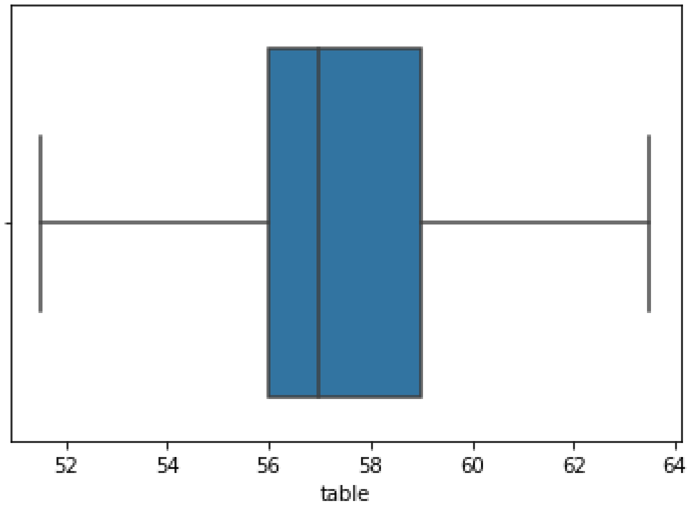


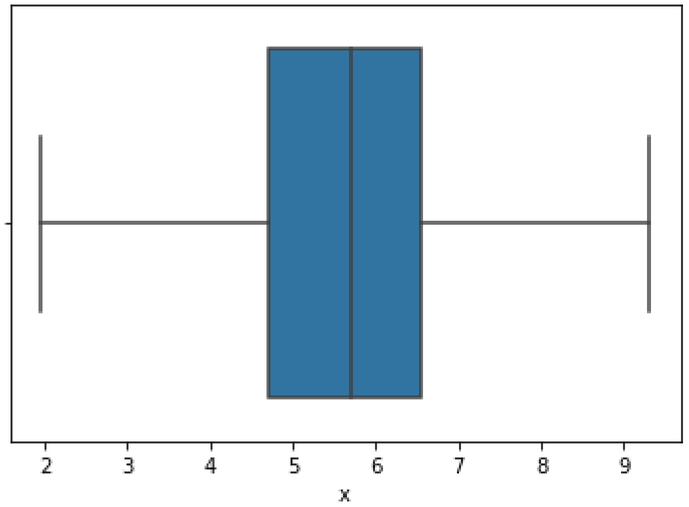
Skewness and kurtosis after correction

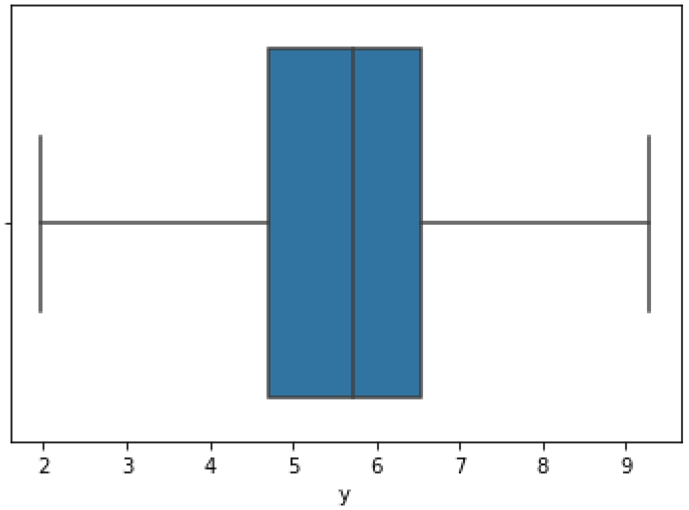
	Kurtosis	skewness
Carat	1.215364	1.116481
Depth	3.674431	-0.028618
Table	1.582166	0.765758
X	-0.657825	0.387986
Y	-0.737128	0.390750
Z	-0.70219	0.38419
Price	0.2294	1.158

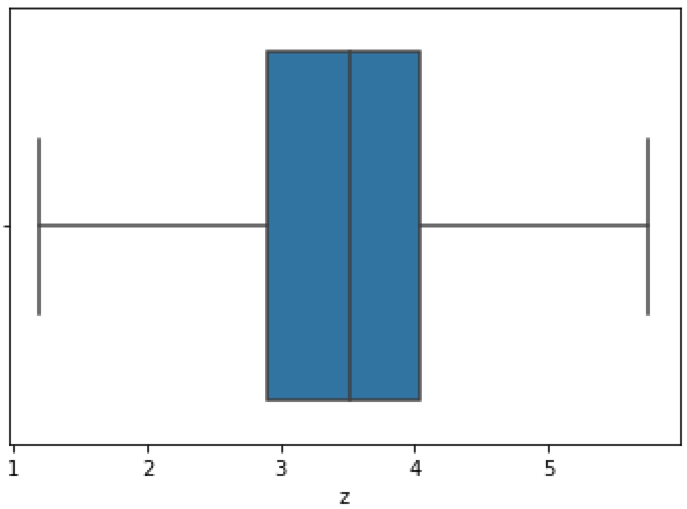
Box plots after treatment of outliers

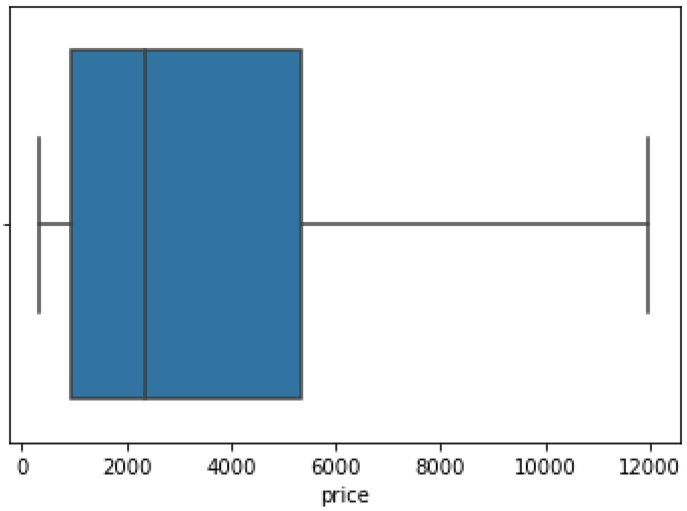


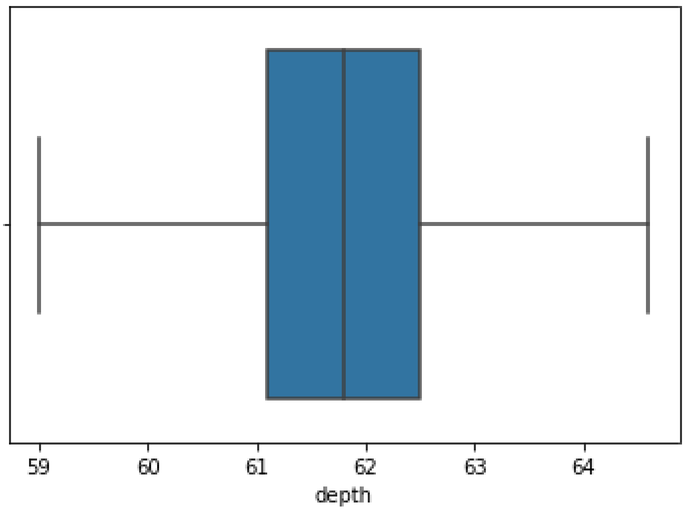

52
54
56
58
60
62
64
table


2
3
4
5
6
7
8
9
x


2
3
4
5
6
7
8
9
y

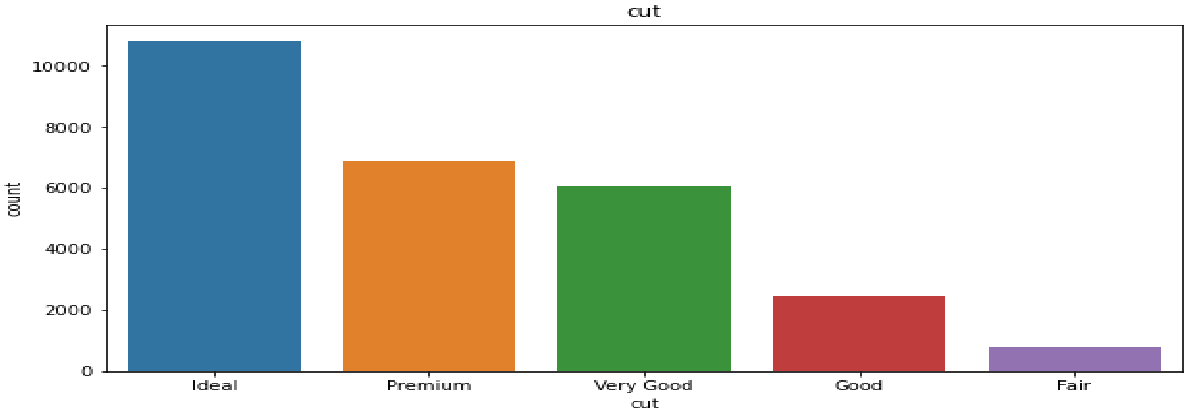


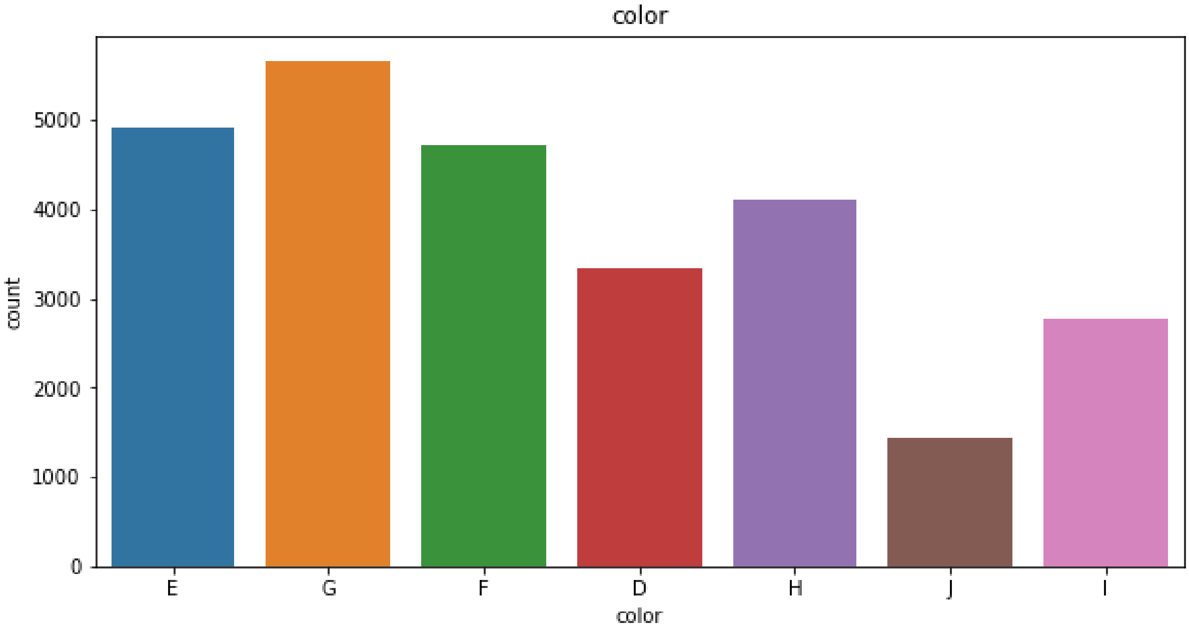


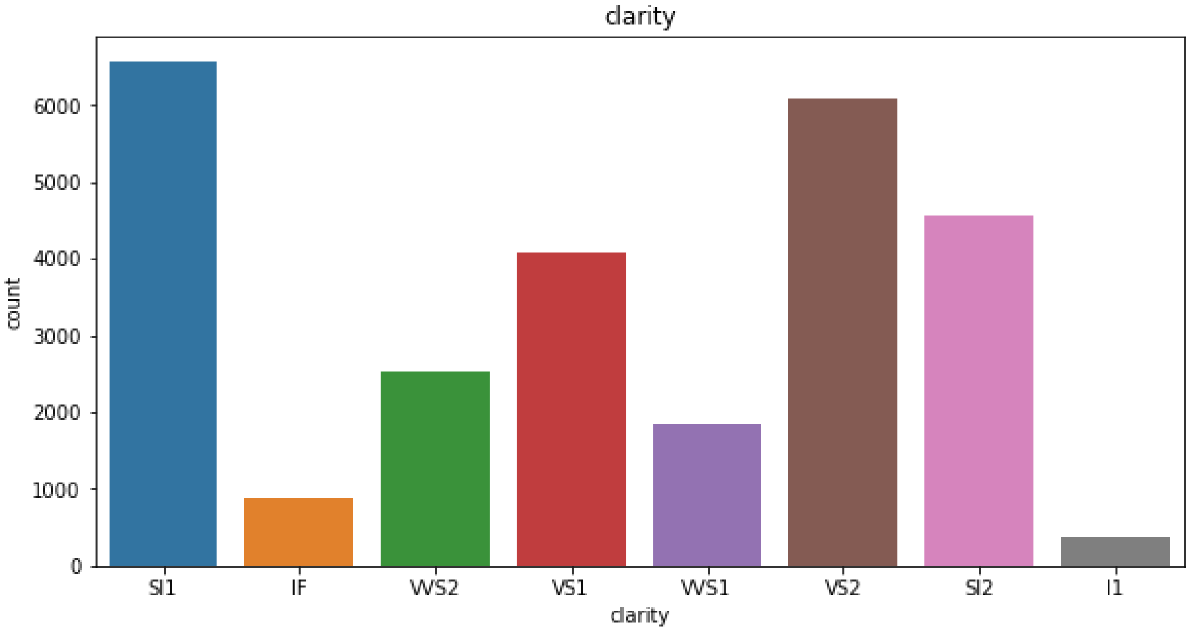


ii) Categorical variables- For categorical variables using count plots

Count plot


cut
count
10000
8000
6000
4000
2000
0
Ideal
Premium
Very Good
Good
Fair
cut


color
count
5000
4000
3000
2000
1000
0
E
G
F
D
H
J
I
color


clarity
count
6000
5000
4000
3000
2000
1000
0
SI1
IF
VVS2
VS1
VVS1
VS2
SI2
I1
clarity

Inference- i) counts of cut in decreasing order ideal>premium>very good>good>fair.

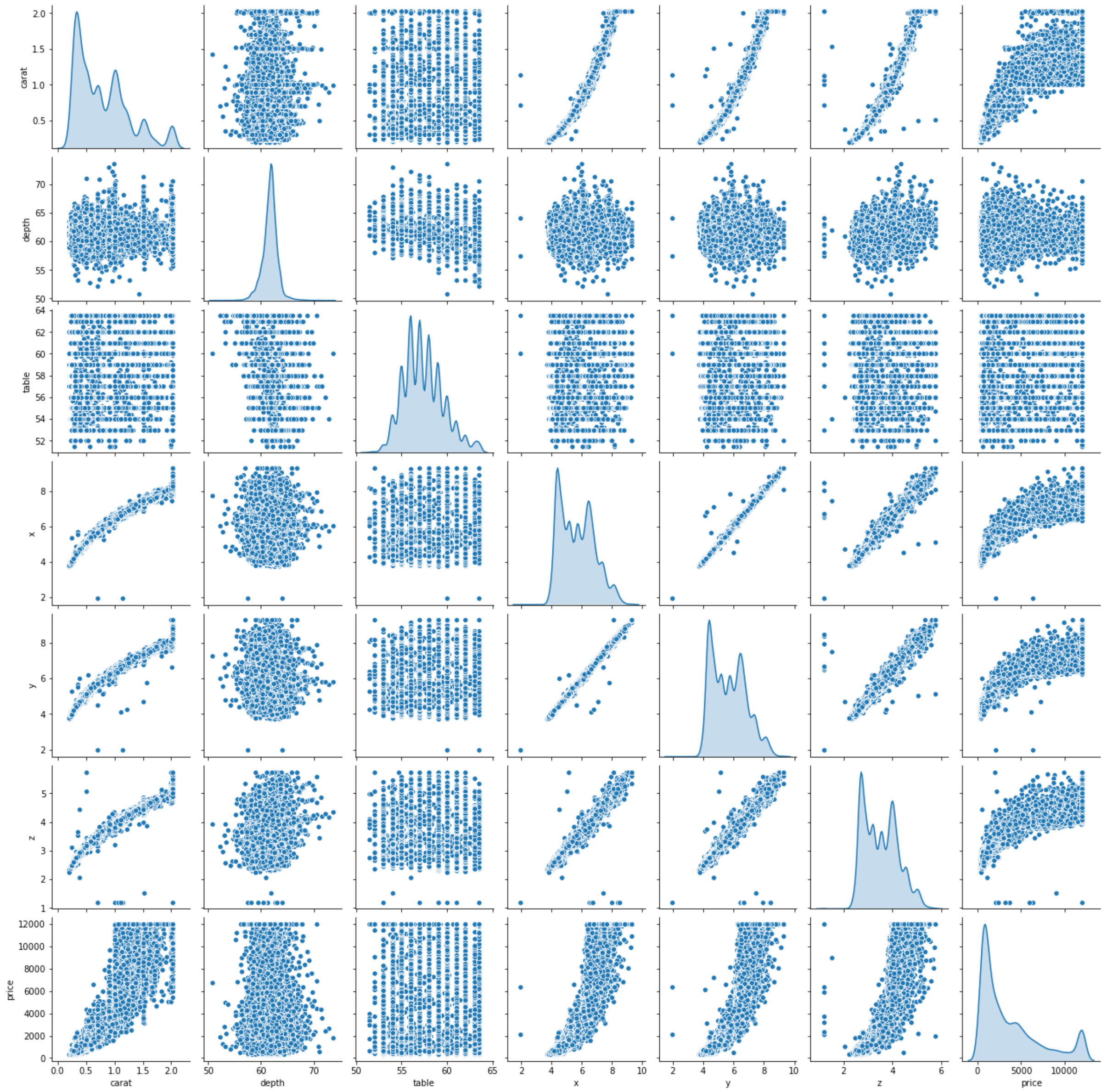
ii) Counts of clarity in decreasing order G>E>F>H>D>I>J.

iii) Counts of clarity in decreasing order SI1>VS2>SI2>VS1>WS2>WS1>IF>I1.

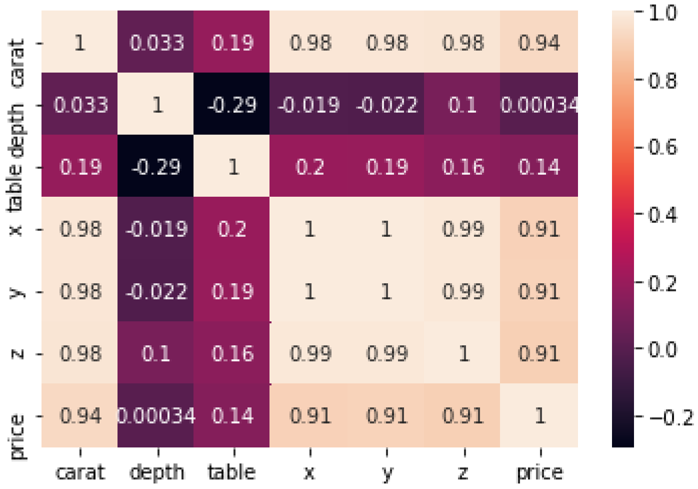
Bivariate Analysis

i) continuous variable- To check correlation between different continuous variables using pair plot and heat map

Pair plot



Heat Map



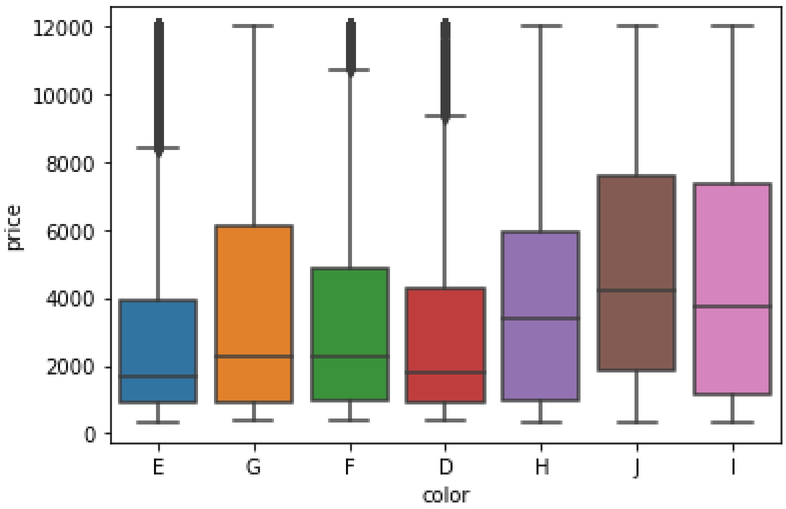
Inference-

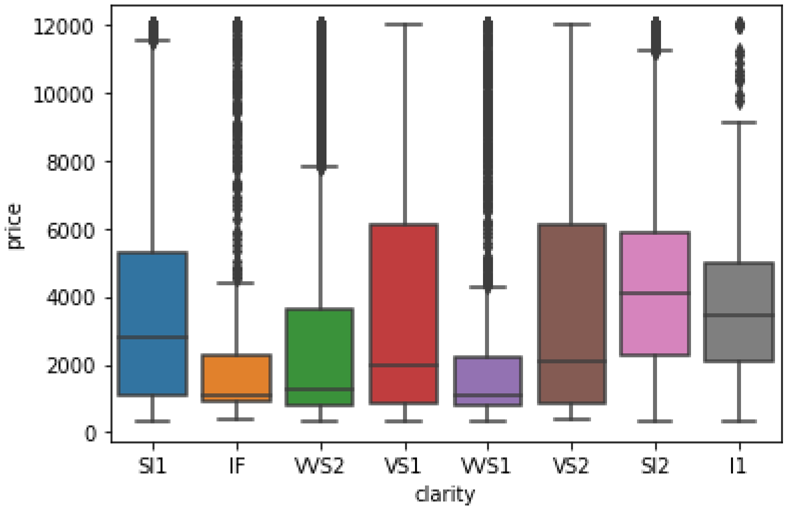
i) There is a positive linear correlation of price which is the target variable with carat, x, y and z variables.

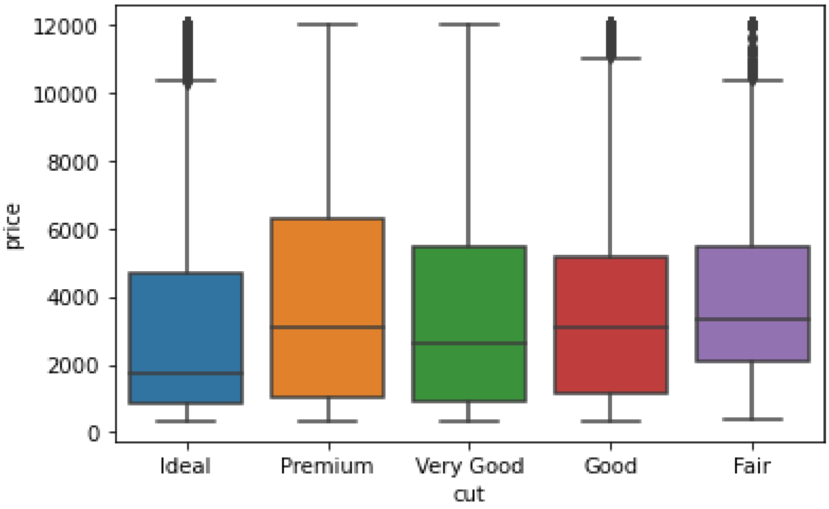
ii) There is a positive linear correlation between carat, x ,y and z variables.

ii) Categorical variables- Using boxplot by keeping target variable(price) on y axis and categorical variables on x axis.

Box Plot







Inference-

i) Median price of stones based on colour in decreasing order J>I>H>G=F>D=E.
ii) MEdian price of stones based on clarity in decreasing order SI2>LI1>SI1>VS1=VS2>WS2>WS1>IF
III) Median price of stones based on cut in decreasing order premium=fair>good>very good>ideal.

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

i) There are 697 null values in the depth column. As null values are 3% of the data we are going to replace it with some values. As this column has outliers and mean value can affect the outliers ,replacement of null has been done by using median value.

ii) X=Length of the cubic zirconia in mm.
   Y= Width of the cubic zirconia in mm.
   Z= Height of the cubic zirconia in mm.

Above all variables have 0 values in some rows. As the measurement of stone cannot be 0 and they have a positive correlation with target variable(price) we are going to replace these 0 values. After correction of outliers these 0 values have been capped into the lower margin of 1.5 IQR.

Summary statistics before replacement of 0 values-

	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	26967.0	13484.000000	7784.846691	1.0	6742.50	13484.00	20225.50	26967.00
carat	26967.0	0.798375	0.477745	0.2	0.40	0.70	1.05	4.50
depth	26270.0	61.745147	1.412860	50.8	61.00	61.80	62.50	73.60
table	26967.0	57.456080	2.232068	49.0	56.00	57.00	59.00	79.00
x	26967.0	5.729854	1.128516	0.0	4.71	5.69	6.55	10.23
y	26967.0	5.733569	1.166058	0.0	4.71	5.71	6.54	58.90
z	26967.0	3.538057	0.720624	0.0	2.90	3.52	4.04	31.80
price	26967.0	3939.518115	4024.864666	326.0	945.00	2375.00	5360.00	18818.00

Summary statistics after replacement of 0 values-

	count	mean	std	min	25%	50%	75%	max
carat	26967.0	0.793593	0.462431	0.200	0.40	0.70	1.05	2.025
depth	26967.0	61.750502	1.218929	59.000	61.10	61.80	62.50	64.600
table	26967.0	57.435699	2.157125	51.500	56.00	57.00	59.00	63.500
x	26967.0	5.729903	1.127023	1.950	4.71	5.69	6.55	9.310

y	26967.0	5.731798	1.118970	1.965	4.71	5.71	6.54	9.285
z	26967.0	3.537261	0.697278	1.190	2.90	3.52	4.04	5.750
price	26967.0	3737.914136	3470.888236	326.000	945.00	2375.00	5360.00	11982.500

iii) scaling- Linear regression uses linear combination which can be expressed as

y=mx+c

y=Dependent/Target variable
m=coefficients/weights
x=independent variable
c=intercept/bias/constant

Weights are calculated by independent variables and it gets inclined towards variables with higher values and thus affects y. Hence scaling is important in linear regression.
Here the z score from scipy.stats is used for scaling of the data.

Top 5 rows of the data frame after scaling-

	carat	cut	color	clarity	depth	table	x	y	z	price
0	-1.07	-0.54	-0.94	-1.06	0.29	0.26	-1.30	-1.29	-1.26	-0.93
1	-1.00	0.43	0.23	-1.64	-0.78	0.26	-1.16	-1.14	-1.20	-0.79
2	0.23	1.41	-0.94	1.84	0.37	1.19	0.28	0.35	0.35	0.73
3	-0.81	-0.54	-0.36	0.10	-0.12	-0.67	-0.81	-0.83	-0.83	-0.77
4	-1.05	-0.54	-0.36	1.26	-1.11	0.73	-1.22	-1.16	-1.27	-0.85
5	0.49	-0.54	-1.53	0.68	-0.21	-0.67	0.65	0.68	0.65	1.66

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

-As linear regression models can not take string values,here data having string values are converted into numerical values by using label encoding.

-After performing encoding the data set has been scaled by using z score method.
- Data is split into then split into train and test set in 70:30 ratio by using train_test_split method from sklearn.
-Here we have made an OLS regression model and passed a train and test set for prediction.

Ordinary Least Square(OLS)- This model uses linear least square method and estimates the coefficients and bias by minimizing the sum of squared errors. It defines the linear regression model as a line which, while passing through the distribution of datasets, minimizes the sum of squared errors between observed and predicted value.

OLS regression result of testing data set-

```
==============================================================================
=====
Dep. Variable:              price          R-squared:                  0.910
Model:                        OLS          Adj. R-squared:             0.910
Method:             Least Squares          F-statistic:            2.129e+04
Date:            Sun, 27 Jun 2021          Prob (F-statistic):          0.00
Time:                    14:51:52          Log-Likelihood:           -4060.3
No. Observations:           18876          AIC:                        8141.
Df Residuals:               18866          BIC:                        8219.
Df Model:                       9
Covariance Type:        nonrobust
==============================================================================
=====
                 coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------
Intercept      0.0034      0.002      1.555      0.120      -0.001       0.008
carat          1.2228      0.012    100.228      0.000       1.199       1.247
cut            0.0123      0.002      5.481      0.000       0.008       0.017
color         -0.1120      0.002    -48.805      0.000      -0.116      -0.107
clarity        0.1239      0.002     54.759      0.000       0.119       0.128
depth         -0.0324      0.003     -9.714      0.000      -0.039      -0.026
table         -0.0468      0.002    -19.613      0.000      -0.052      -0.042
x             -0.6837      0.048    -14.230      0.000      -0.778      -0.590
y              0.5159      0.048     10.764      0.000       0.422       0.610
z             -0.0516      0.020     -2.561      0.010      -0.091      -0.012
==============================================================================
=====
Omnibus:                 4847.936      Durbin-Watson:                1.984
Prob(Omnibus):              0.000      Jarque-Bera (JB):         24178.418
Skew:                       1.156      Prob(JB):                      0.00
Kurtosis:                   8.040      Cond. No.                      61.5
```

OLS regression result of testing data set

```
                    OLS Regression Results
==============================================================================
========
Dep. Variable:              price          R-squared:                 0.908
Model:                       OLS          Adj. R-squared:             0.908
Method:                Least Squares         F-statistic:             8862.
Date:             Sun, 27 Jun 2021        Prob (F-statistic):          0.00
Time:                    14:52:19         Log-Likelihood:           -1786.6
No. Observations:            8091              AIC:                     3593.
Df Residuals:              8081               BIC:                      3663.
Df Model:                    9
Covariance Type:         nonrobust
==============================================================================
========
                 coef    std err        t         P>|t|     [0.025     0.975]
------------------------------------------------------------------------------
Intercept   -0.0079     0.003     -2.361     0.018     -0.015     -0.001
carat        1.1986     0.019     63.543     0.000      1.162      1.236
cut          0.0182     0.004      5.176     0.000       0.011      0.025
color       -0.1130     0.004    -32.129     0.000     -0.120     -0.106
clarity      0.1303     0.003     37.724     0.000       0.124      0.137
depth       -0.0159     0.007     -2.276     0.023      -0.030     -0.002
table       -0.0421     0.004    -11.639     0.000     -0.049     -0.035
x           -0.5114     0.059     -8.700     0.000      -0.627     -0.396
y            0.4903     0.055      8.957     0.000       0.383      0.598
z            -0.1811     0.052     -3.514     0.000     -0.282     -0.080
==============================================================================
========
Omnibus:                  1993.190   Durbin-Watson:               2.010
Prob(Omnibus):               0.000   Jarque-Bera (JB):         10972.411
Skew:                     1.070       Prob(JB):                  0.00
```

Inference- Both train and test data have correctly fit into the model.

	Train data	Test data
R squared	0.91	0.90
Adjusted r squared	0.91	0.90

90% of the time model will make the correct prediction.

- Coefficients and intercepts which will make best fit line are as following

	Train data	Test data
Intercept	0.0034	-0.079
Carat	1.2228	1.1986
Cut	0.0123	0.0182
Color	-0.1120	-0.1130
Clarity	0.1239	0.1303
Depth	-0.0324	-0.0159
Table	-0.0468	-0.0421
X	-0.6837	0.5114
Y	0.5159	0.4903
Z	-0.0516	-0.1811

- As the overall p value of the model is 0.00 which is <0.05 ,this model is reliable to use.
- Price of stones is showing a positive correlation with carat, cut , clarity, x
and y and p values of these variables are 0.00(>0.05) these correlations are real and just because of coincidence.

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

Inference- with prediction of 90% this model will be correct in pricing of the time and the model is reliable to use.
Recommendation- Price of stones is highly related to following attributes
i) Carat
ii)Cut
iii)Clarity
iv)X
V) Y
So if values of the above attributes are more price of stones will increase and they can make more profits from them.

Problem 2-Logistic Regression and Linear Discriminant Analysis

A tour and travel company which deals in holiday packages wants to predict whether an employee will opt for the holiday package or not on the basis of the information given in the data set of the 872 employees of the company.

Data Dictionary:

Variable Name	**Description**
Holiday_Package	Opted for Holiday Package yes/no?
Salary	Employee salary
age	Age in years
edu	Years of formal education
no_young_children	The number of young children (younger than 7 years)
no_older_children	Number of older children
foreign	foreigner Yes/No

Question 2.1- Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis.

Top 5 rows of data set-

	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

Summary of data set

```
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)
memory usage: 54.6+ KB
```

Inference-

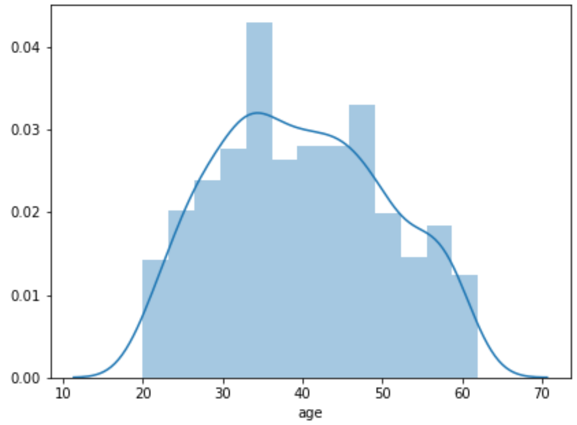
i) Number of rows - 872
 - Number of columns- 8
ii) Salary, age and educ variables are numeric and continuous.
 - No_young_children and no_older_children variables are numeric and categorical.
 - Holliday_package and foreign variables are object and categorical.
iii) There are no null values.
iv)There are no duplicates.

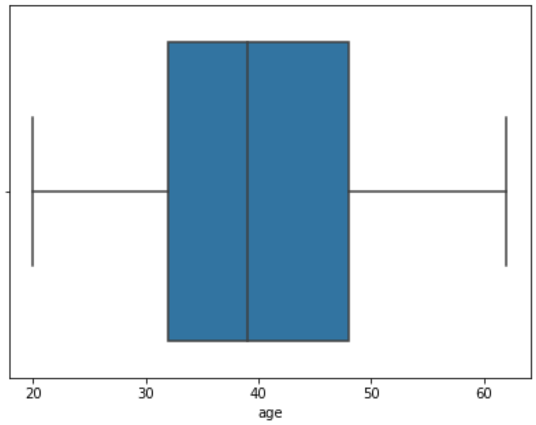
Univariate Analysis-

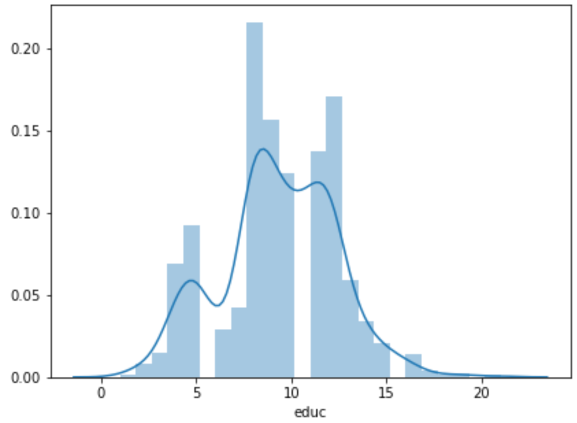
Summary statistic of data set-

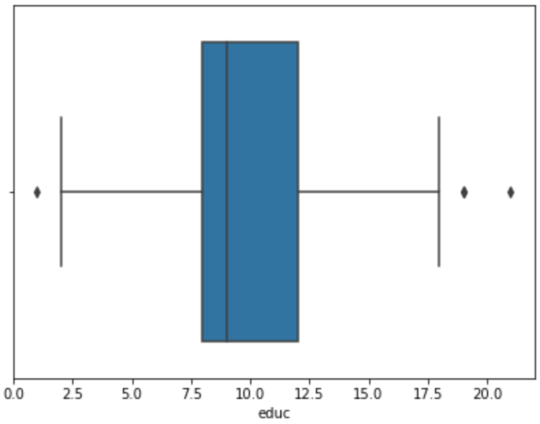
	Unnamed: 0	Salary	age	educ	no_young_children	no_older_children
count	872.000000	872.000000	872.000000	872.000000	872.000000	872.000000
mean	436.500000	47729.172018	39.955275	9.307339	0.311927	0.982798
std	251.869014	23418.668531	10.551675	3.036259	0.612870	1.086786
min	1.000000	1322.000000	20.000000	1.000000	0.000000	0.000000
25%	218.750000	35324.000000	32.000000	8.000000	0.000000	0.000000
50%	436.500000	41903.500000	39.000000	9.000000	0.000000	1.000000
75%	654.250000	53469.500000	48.000000	12.000000	0.000000	2.000000
max	872.000000	236961.00000	62.000000	21.000000	3.000000	6.000000

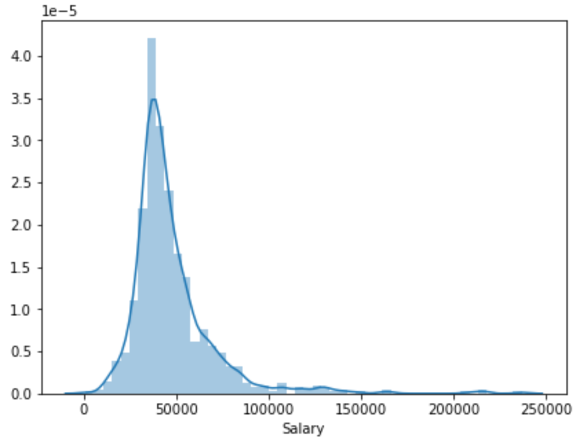
i) continuous variables- For continuous variables using distribution plot and box plot for outliers.

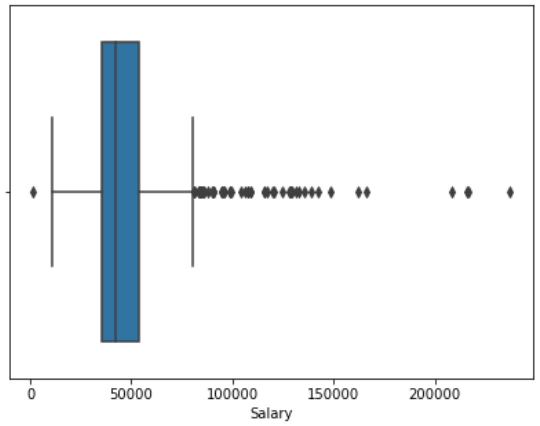












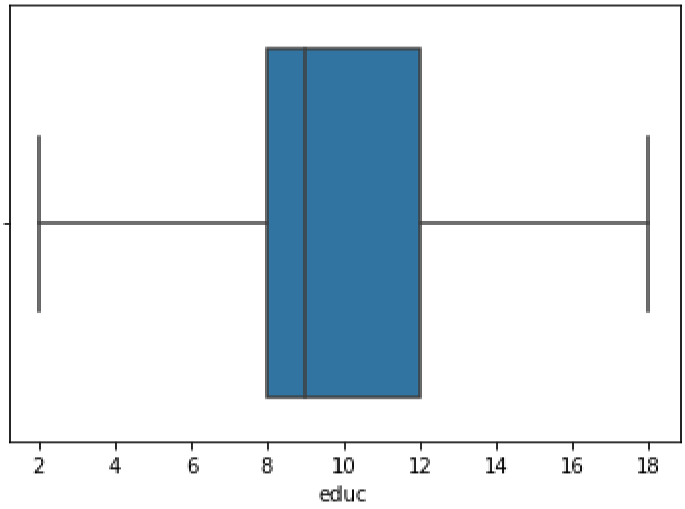
Skewness and kurtosis

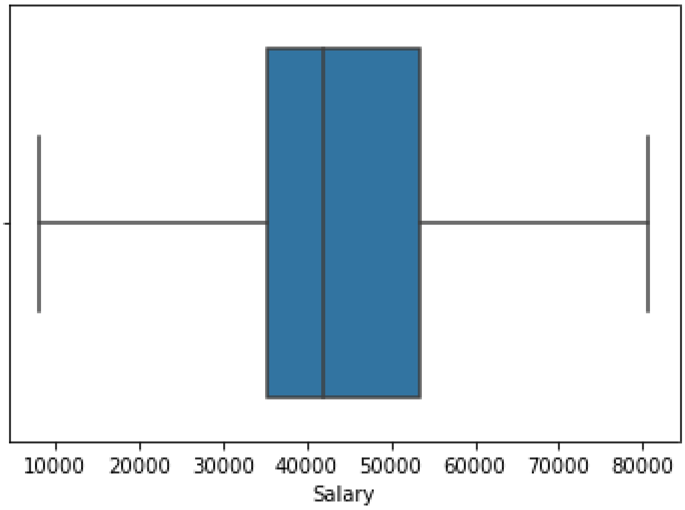
	Skewness	Kurtosis
Salary	3.103	15.852
Age	0.146	-0.909
Educ	-0.045	0.005

Inference- i) All continuous variables salary are normally distributed. Salary variable is positively skewed.
ii) All variables except age are having outliers in both upper and lower range.

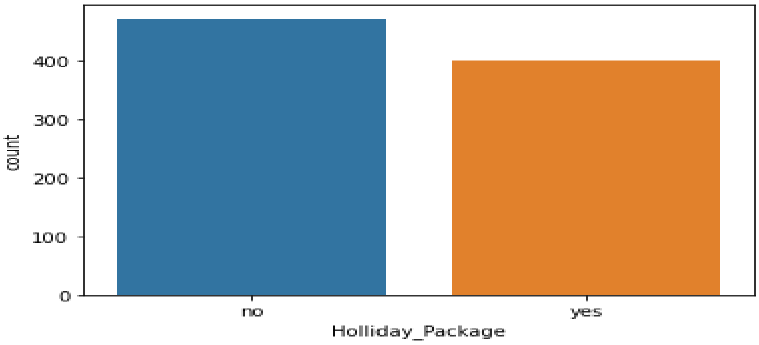
As presence of outliers can affect performance of the model, we are going to treat outliers by capping outliers between -1.5 and +1.5 times of IQR.

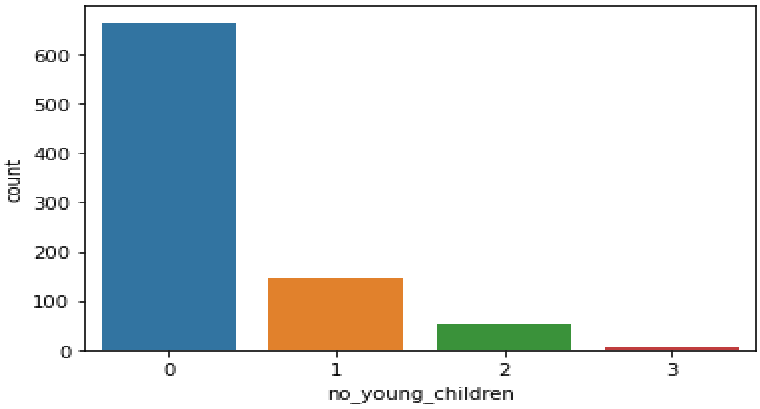
Box plots after treatment of outliers-

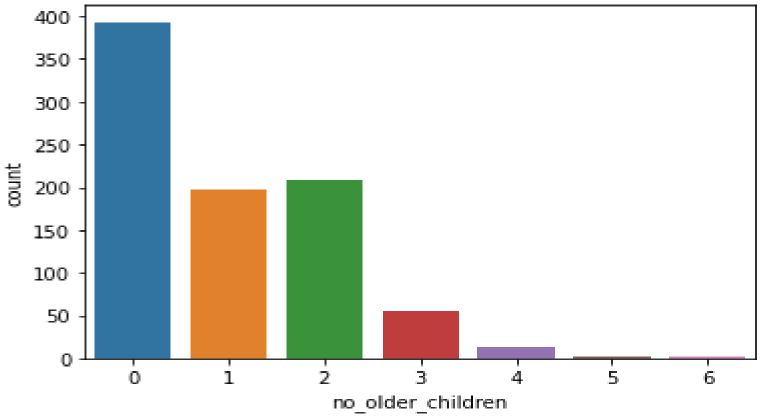


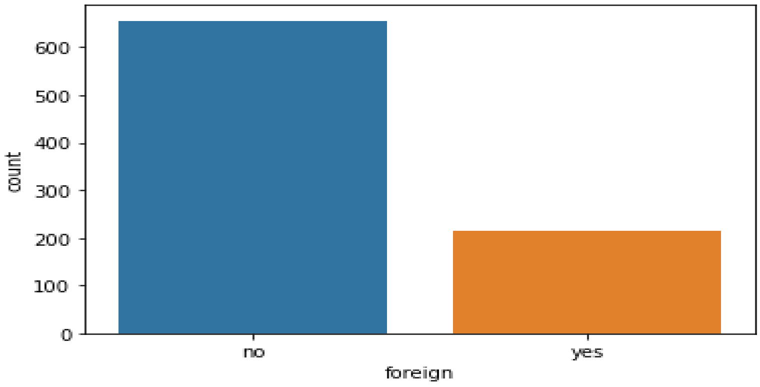


ii) Categorical variables- for categorical variables using count plots





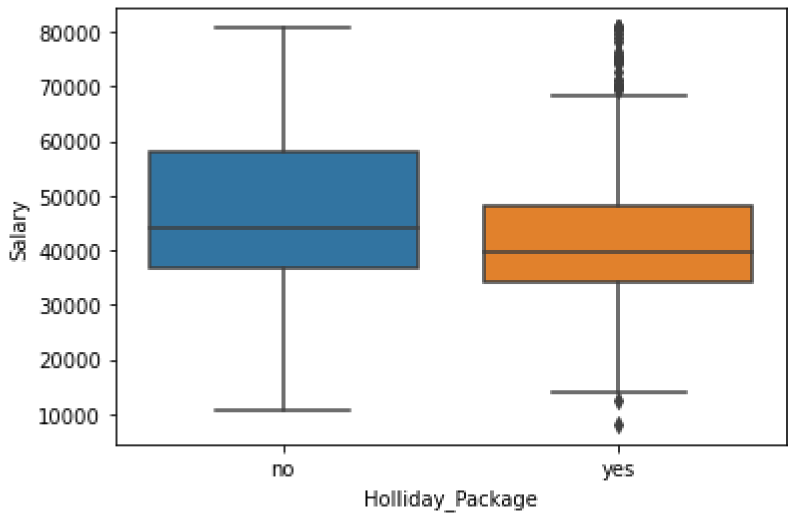


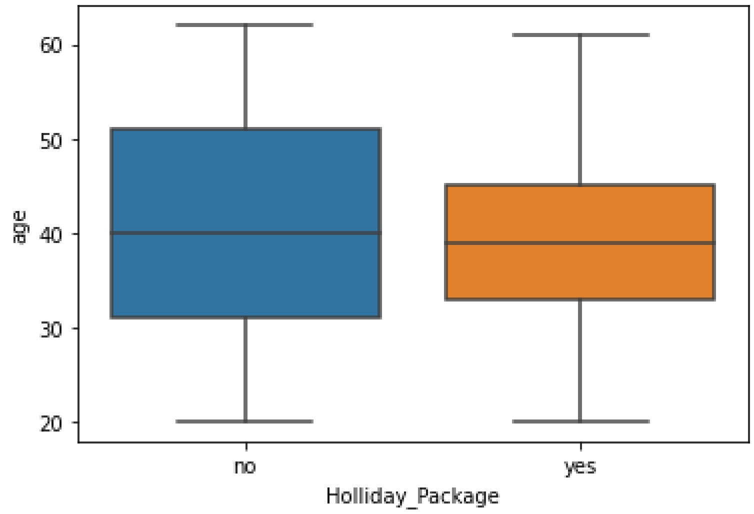


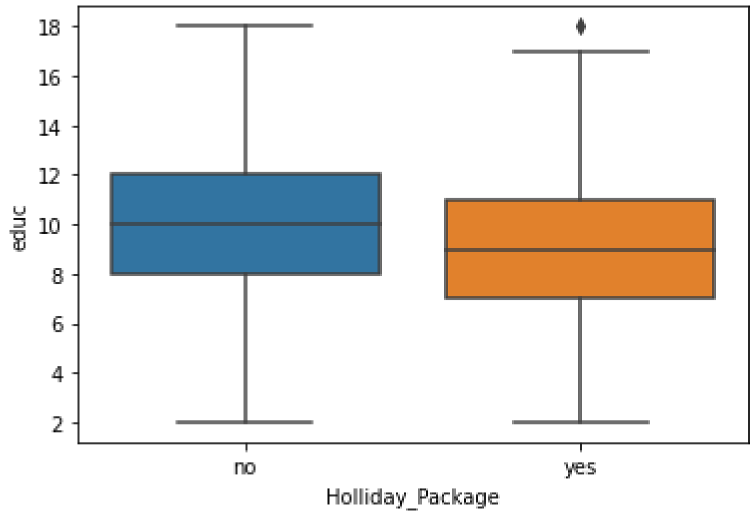
Inference- i)Number of employees who has taken package is lesser than who has not taken packages.
ii) Counts of employees decreases as no_young_children increases.
iii) Counts of employees decreases as no_older_children increases.
iv) Counts of foreigner employees are less than non foreign employees.

Bivariate Analysis-

i) Target and continuous variables- Using box plot with target variable on x- axis and continuous variable on y-axis.

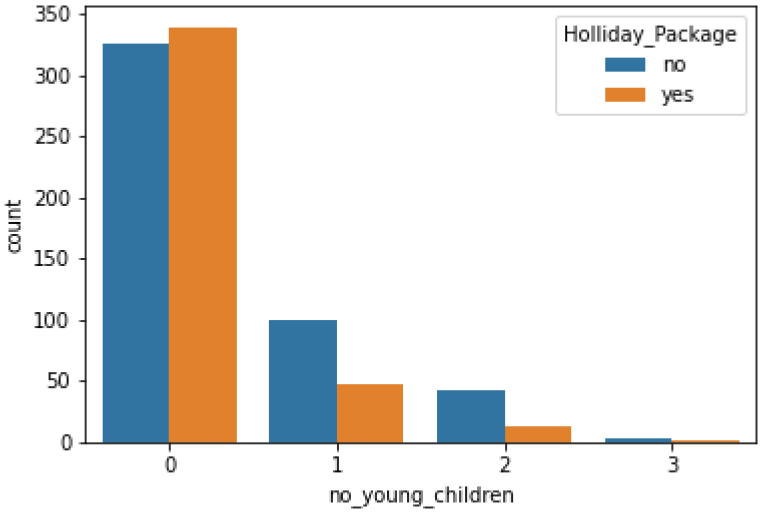


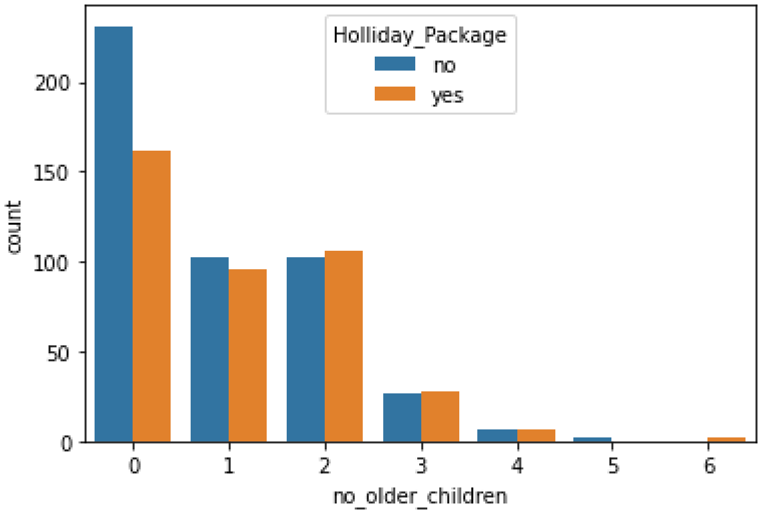


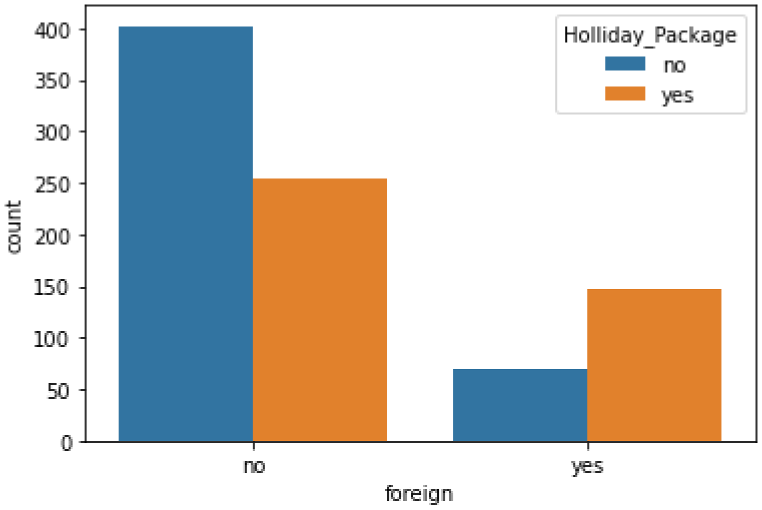


Inference- i) Median salary of employees taking holiday_package is lesser than those who are not taking packages.
ii) Median age of employees who are taking packages is less than 40 and not taking packages is 40.
iii) Median education level of employees taking packages is less than those who are not taking packages.

ii) For Target and categorical variables- using count plot with categorical variable on x-axis and target variable as hue.







Inference--no_young_children-

	Taken package	Not taken package
0 young children	around 340	around 325
1 young child	around 40	around 100
2 young children	around 20	around 20
3 young children	around 5	around 10

Majority of employees with young children have not taken the package

- no_older_children-

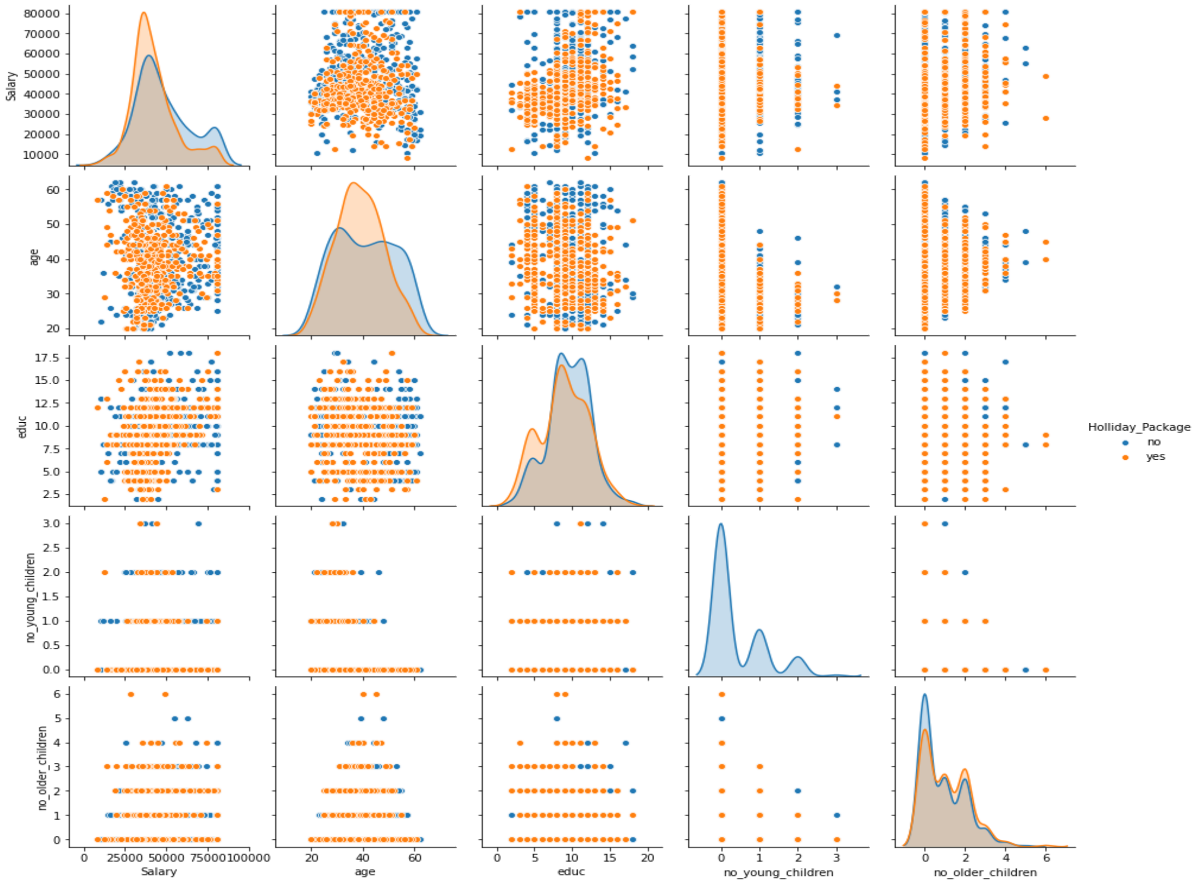
	Taken package	Not taken package
0 old children	around 160	>200
1 old child	around 90	around 100
2 old children	around 100	around 95
3 old children	around 25	around 25
4 old children	around 10	around 10
5 old children	0	around 5
6 old children	around 5	0

Majority of employees with older children have not taken the package

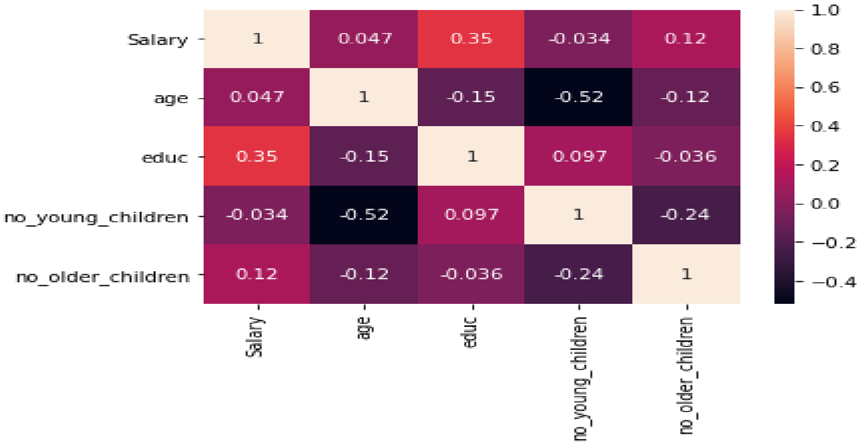
- majority of the employees who have not taken packages are not foreign employees .who has taken packages majority are foreign employees

iii) For continuous variables- for correlation between continuous variables using pair plot and heatmap and using target variable as hue.

Pair plot



Heat Map



Inference- By looking at the scatter plots and heat map we can see that there is a very weak correlation between different variables.

- By looking at diagonal distribution graphs in pair plot we can find out which variable can be a good predictor for classification of target variable based on how well the graphs are separated.

- No_older _children is a poor predictor because graphs of two classes of target variable are overlapping each other.
- No_younger_children can be a very good predictor because it is separating classes vey well.
- Salary, age and educ are weak predictors for separating classes.

Question 2.2-Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis).

-Data having string values are converted into numerical values by using label encoding.
-Then data is separated into independent variables by assigning x and dependent variable by assigning y.
-Data is split into train and test in 70:30 ratio by using train_test_split method from sklearn library.
-Logistic Regression model- Logistic regression model works by assigning probabilities to different classes to which a query point is likely to belong.To do so , it learns from the training set a vector of weight and bias.Each weight is assigned to one input variable. To classify a query point,the classifier takes the weight sum of features and bias to represent the evidence of the query point belonging to the class of interest.

$$Z=WX+B$$

W = Weight
X = input variable
B= Bias

To transform the z value into probability, Z is passed through sigmoid function

- The algorithm uses a cross entropy loss function to find optimal weight and bias across entire data set put together.
- Here we have made a logistic regression model by using OVR(one vs rest) scheme and regularized it by using L2 method with ‘ibfgs’ solver.

- Linear Discriminant Analysis- LDA model uses Bayes’ Theorem to estimate probabilities. They make predictions upon the probability that a new input dataset belongs to each class. The class which has the highest probability is considered as the output class and then the LDA makes a prediction.

Question 2.3- Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.

i) Logistic Regression-

- Train data-

 Accuracy- 0.53

Confusion matrix- True negative- 326
False negative- 284
True positive- 0
False negative- 0

Classification report-

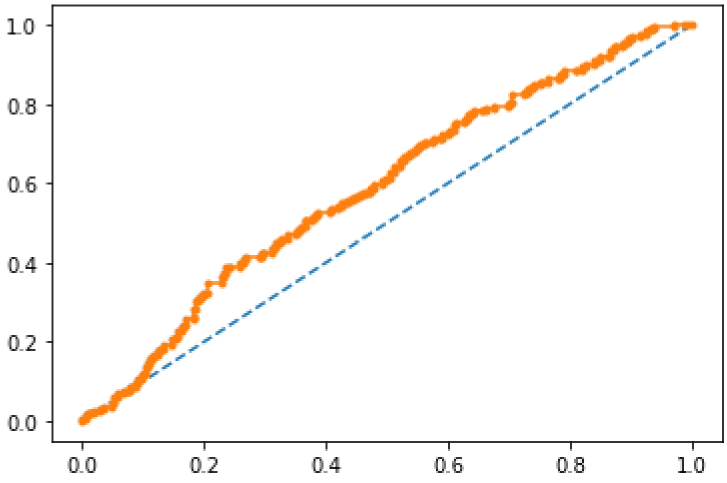
```
           precision    recall  f1-score   support

       0      0.53      1.00      0.70       326
       1      0.00      0.00      0.00       284

   accuracy                          0.53       610
  macro avg       0.27      0.50      0.35       610
weighted avg       0.29      0.53      0.37       610
```

Roc_auc_score- 0.5909

Roc curve-



- Test data-

Accuracy- 0.55
Confusion matrix- True negative- 145
False positive- 117
True positive- 0
False negative- 0

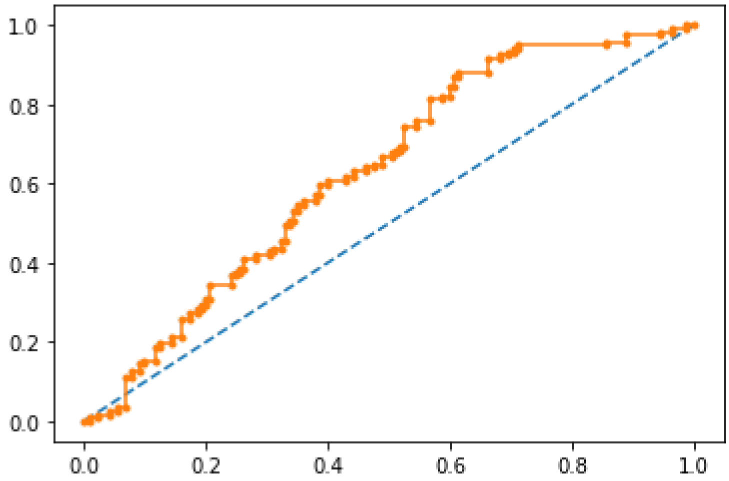
Classification report-
```
              precision    recall  f1-score   support

           0       0.55      1.00      0.71       145
           1       0.00      0.00      0.00       117

    accuracy                           0.55       262
   macro avg       0.28      0.50      0.36       262
weighted avg       0.31      0.55      0.39       262
```

Roc_auc_score- 0.632

Roc curve-



Inference- As scaling has not been done, it has affected the outcome of the model. Underfitting of data has happened as prediction of both train and test is poor with just 55% of accuracy.

ii) Linear Discriminant Analysis-

Train data-

Accuracy- 0.68
Confusion matrix- True negative- 254
False negative- 126
True positive- 158
False positive- 72

Classification report-

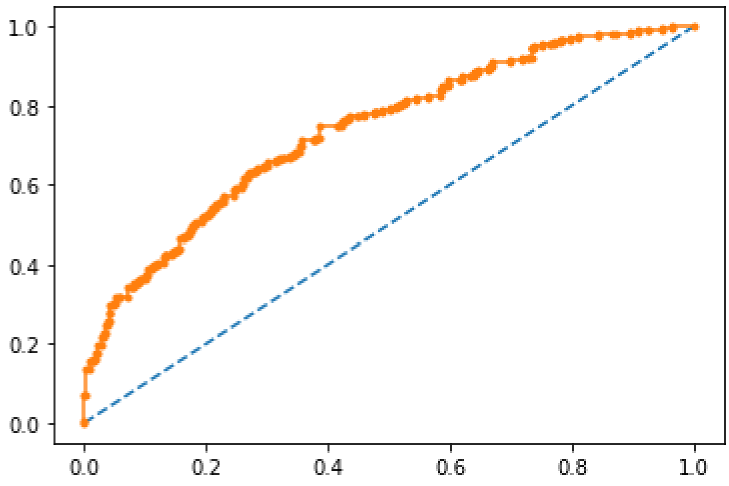
```
              precision    recall  f1-score   support

           0       0.67      0.78      0.72       326
           1       0.69      0.56      0.61       284

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

Roc_auc_score- 0.739

Roc_curve-



<u>Test data</u>-

Accuracy-0.64

Confusion matrix- True negative- 102
False negative- 52
True positive- 65
False positive- 43

Classification report-

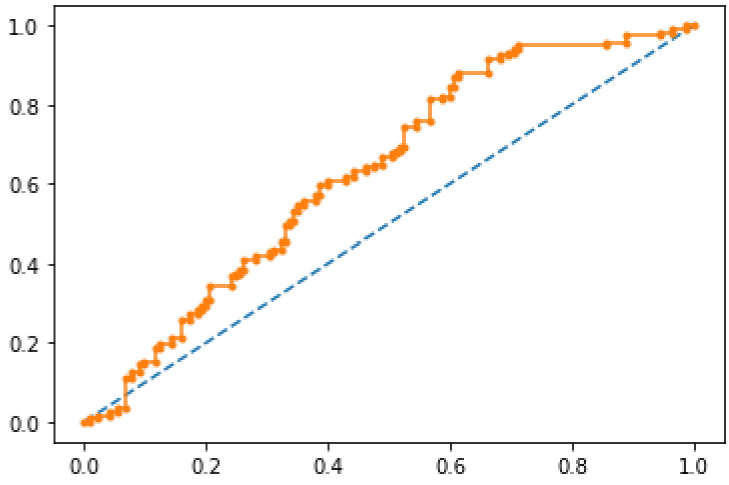
```
              precision    recall  f1-score   support

           0       0.66      0.70      0.68       145
           1       0.60      0.56      0.58       117

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

Roc_auc_score- 0.632

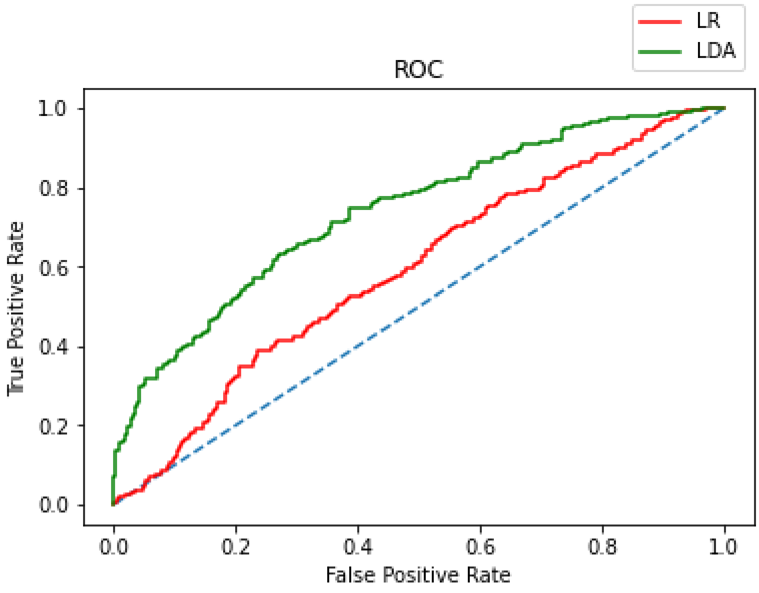
Roc_curve-



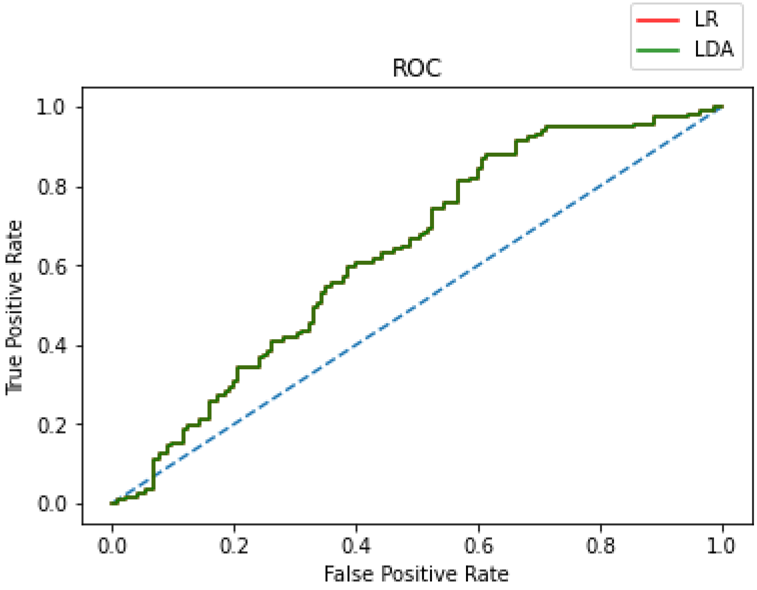
Inference- the data is well fit because it is giving almost the same result on both train and test. There is poor performance of the model because scaling of data has not been done and it has affected the outcome of the model. This model has accuracy of 0.64, precision of 0.60 and recall of 0.56.

Comparison of roc_curve

i) Train data-



ii) Test data-



Inference- Linear discriminant analysis is better than logistic regression in solving this problem.

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

Inference - LDA model is better than logistic regression model in solving this problem.
Recommendations- Salary, age, education and no_young _ children are good predictors for making the decision.