PREDICTIVE MODELING

DSBA Nov Batch A

Abstract

Two case studies cubic zirconia price prediction and Holiday package purchase prediction have been solved using Linear regression model, Logistic regression and Linear Discriminant Analysis

Jyoti Ranjan Padhiary

[Email address]

Table of Contents

PREDICTIVE MODELING

Problem – 1 Linear Regression

You are hired by a company Gem Stones co ltd, which is a cubic zirconia manufacturer. You are provided with the dataset containing the prices and other attributes of almost 27,000 cubic zirconia (which is an inexpensive diamond alternative with many of the same qualities as a diamond). The company is earning different profits on different prize slots. You have to help the company in predicting the price for the stone on the bases of the details given in the dataset so it can distinguish between higher profitable stones and lower profitable stones so as to have better profit share. Also, provide them with the best 5 attributes that are most important.

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

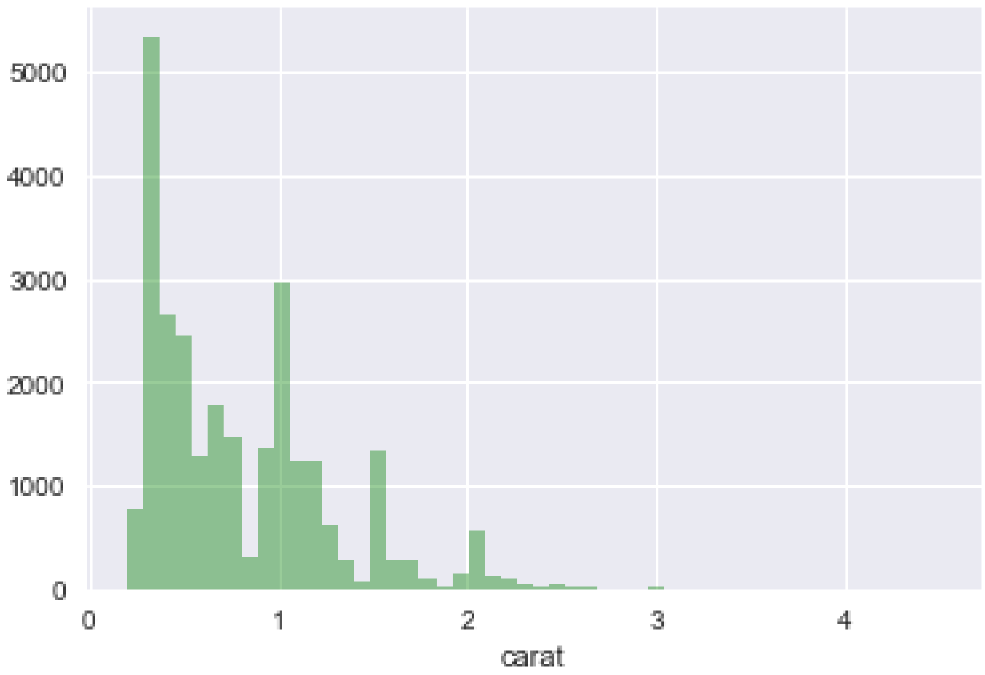
There are 26967 observations and 10 variables. One un identified variable “Unnamed: 0” is found which has been dropped from the data frame.

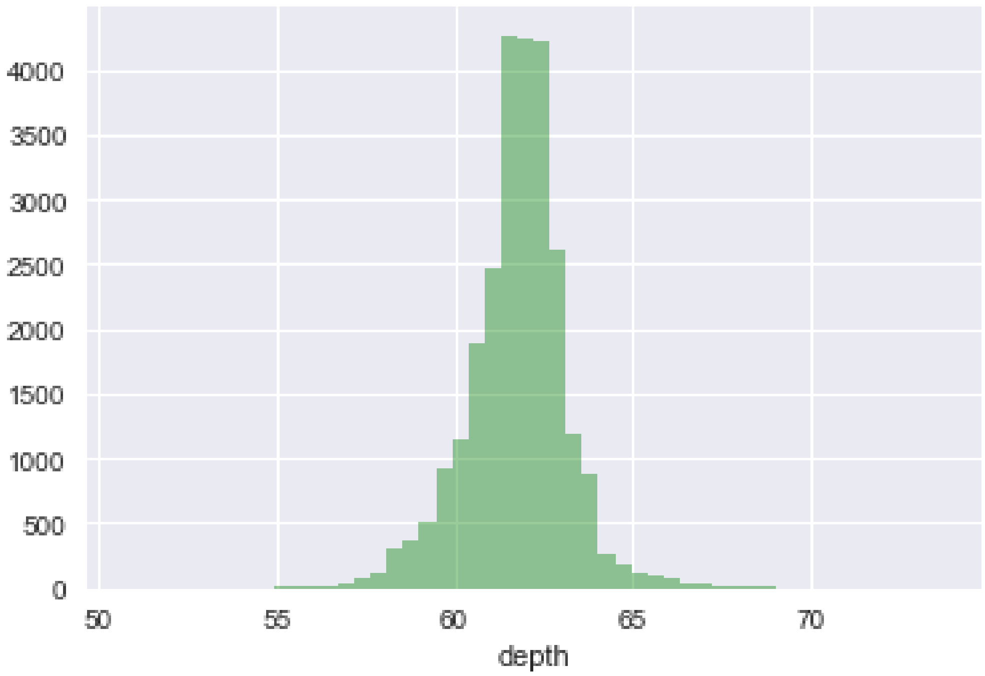
Out of 10 variables 3 are categorical and 7 numerical. The target variable is “Price”. There are some duplicates found which has been dropped.

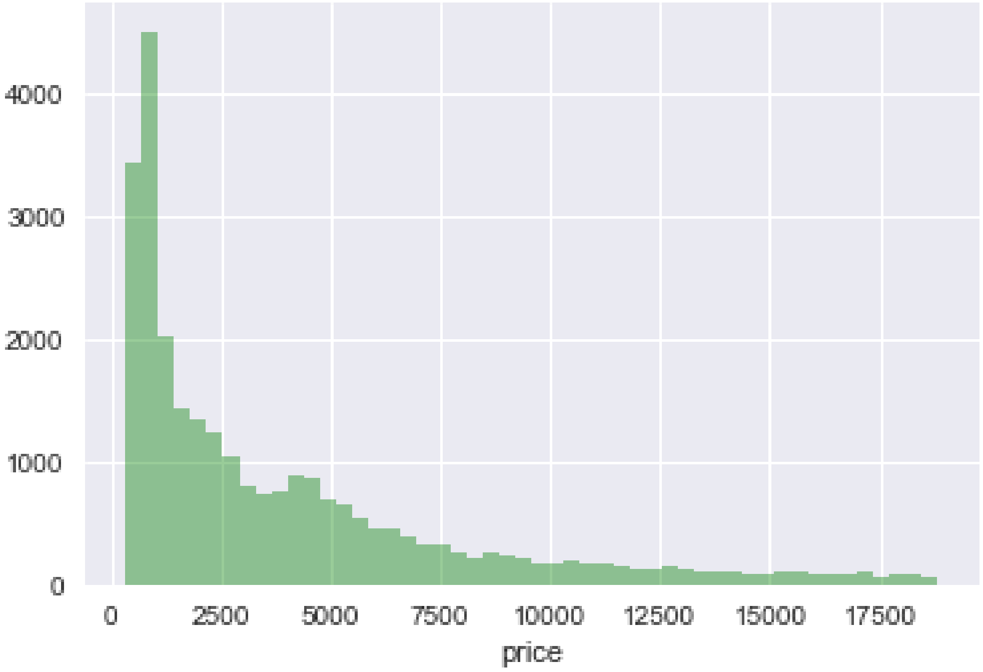
We will start with exploratory data analysis by conducting univariate, bivariate and multivariate analysis.

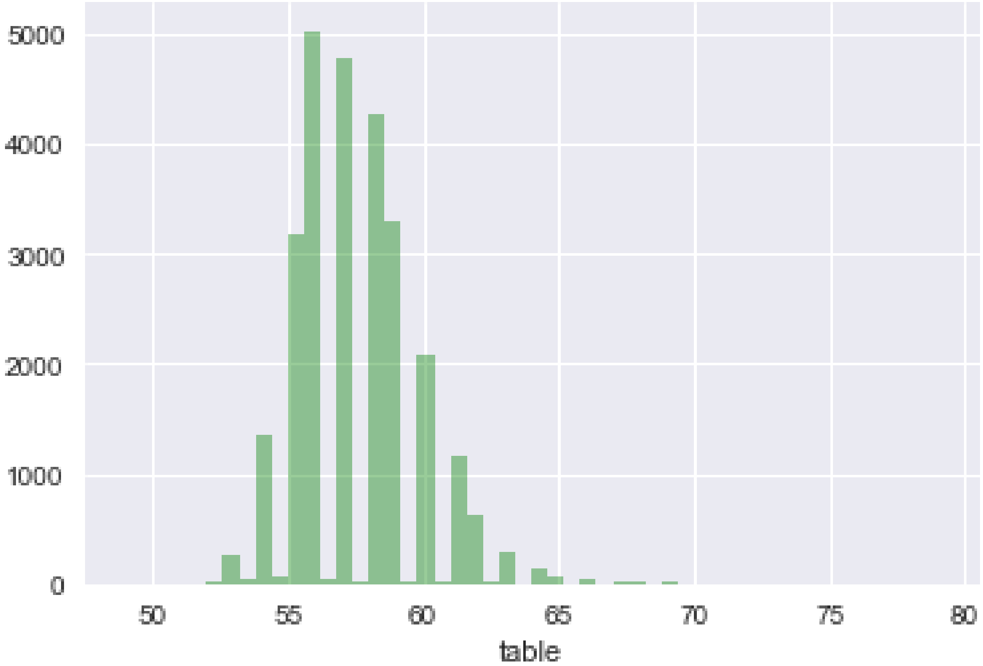
Univariate Analysis

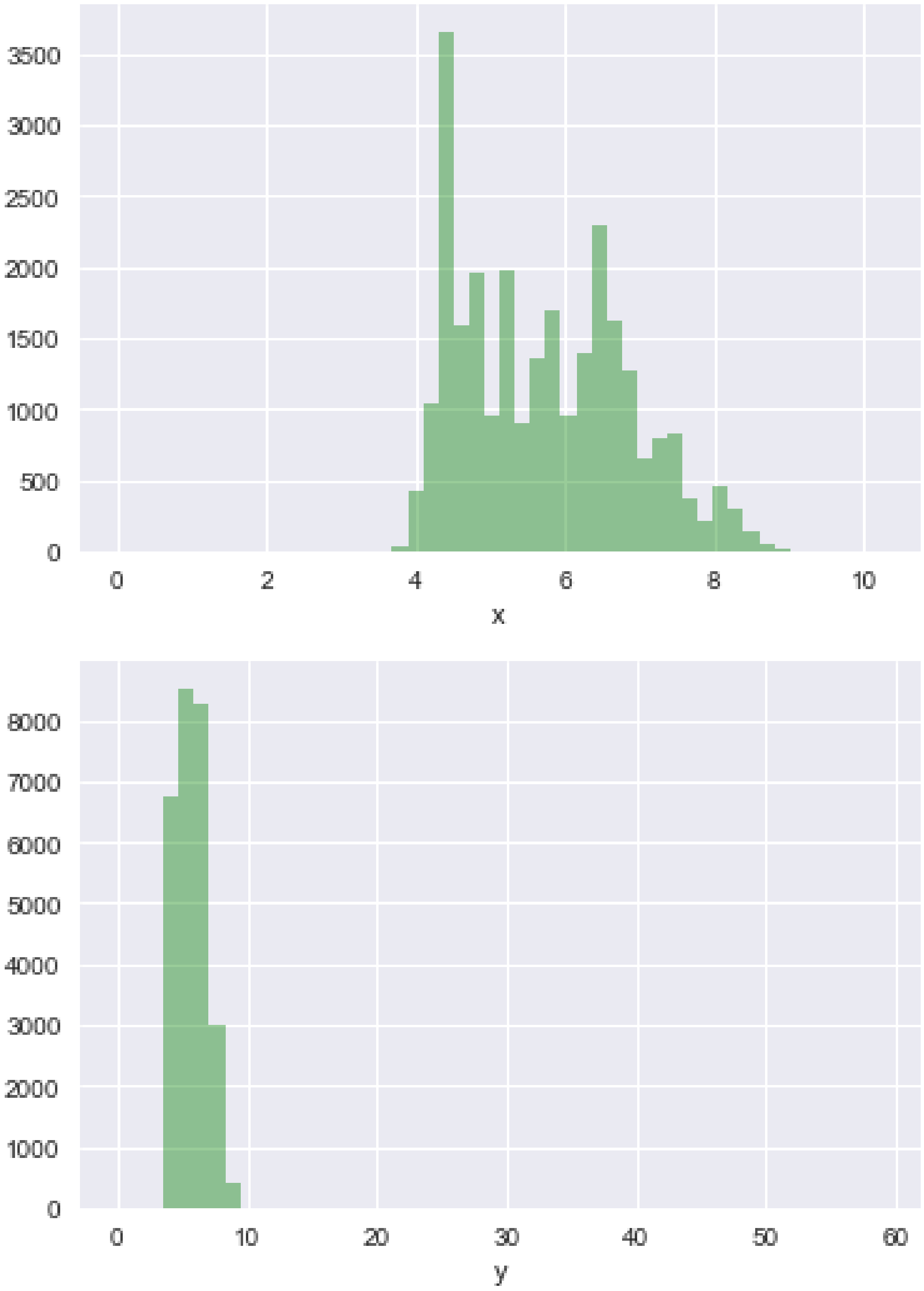
Numerical Variables

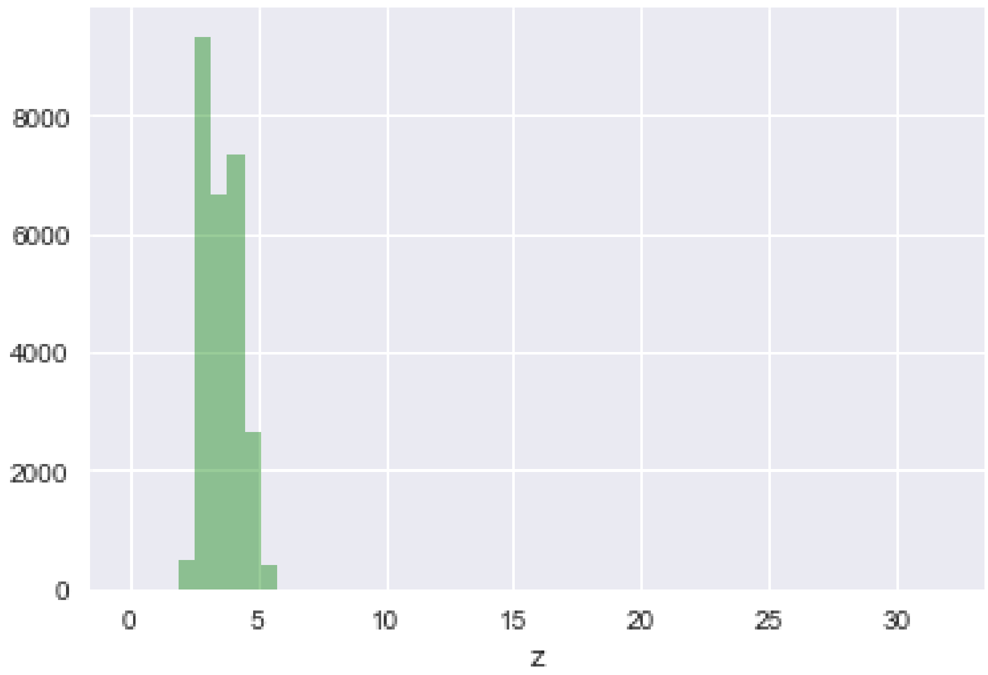

5000
4000
3000
2000
1000
0
0
1
2
3
4
carat


4000
3500
3000
2500
2000
1500
1000
500
0
50
55
60
65
70
depth

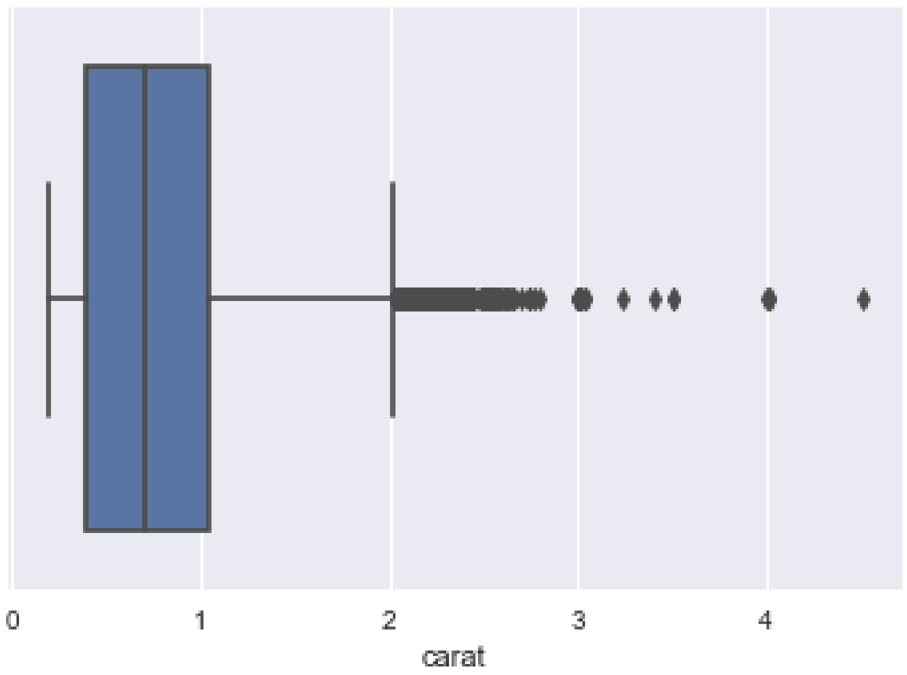

4000
3000
2000
1000
0
0
2500
5000
7500
10000
12500
15000
17500
price

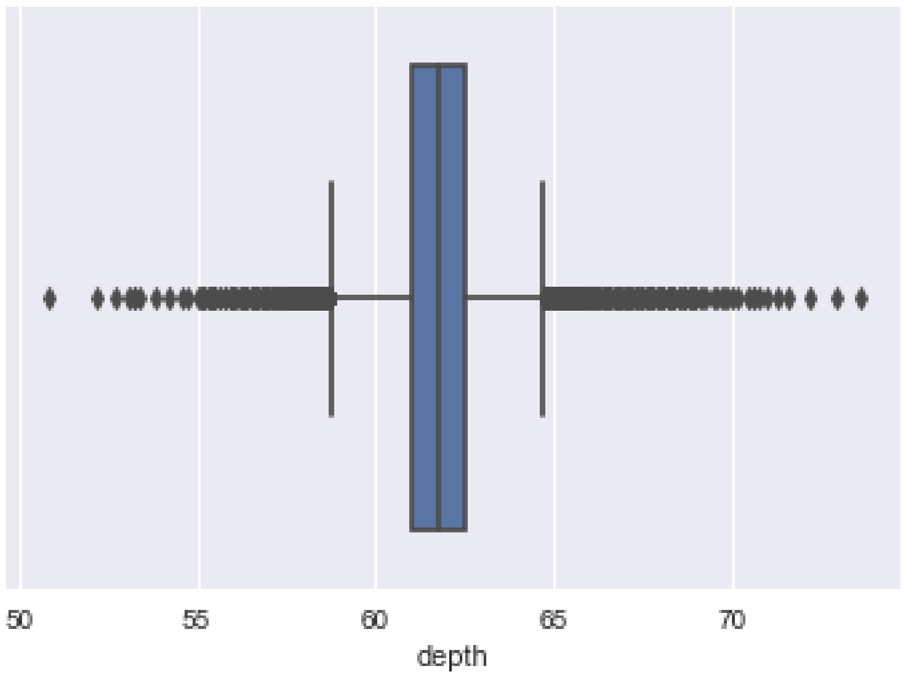

5000
4000
3000
2000
1000
0
50
55
60
65
70
75
80
table

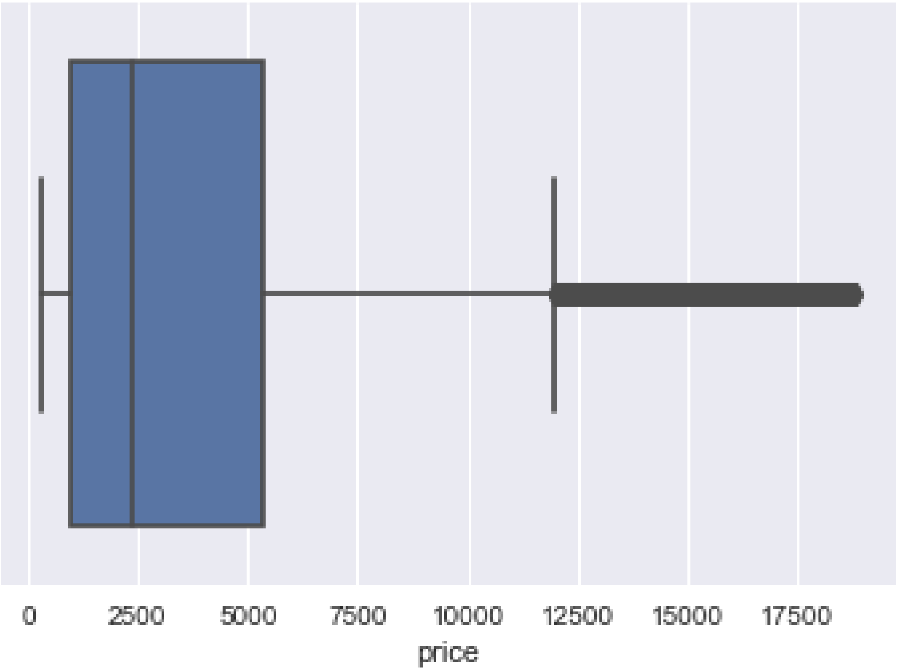

3500
3000
2500
2000
1500
1000
500
0
0
2
4
6
8
10
x
8000
7000
6000
5000
4000
3000
2000
1000
0
0
10
20
30
40
50
60
y

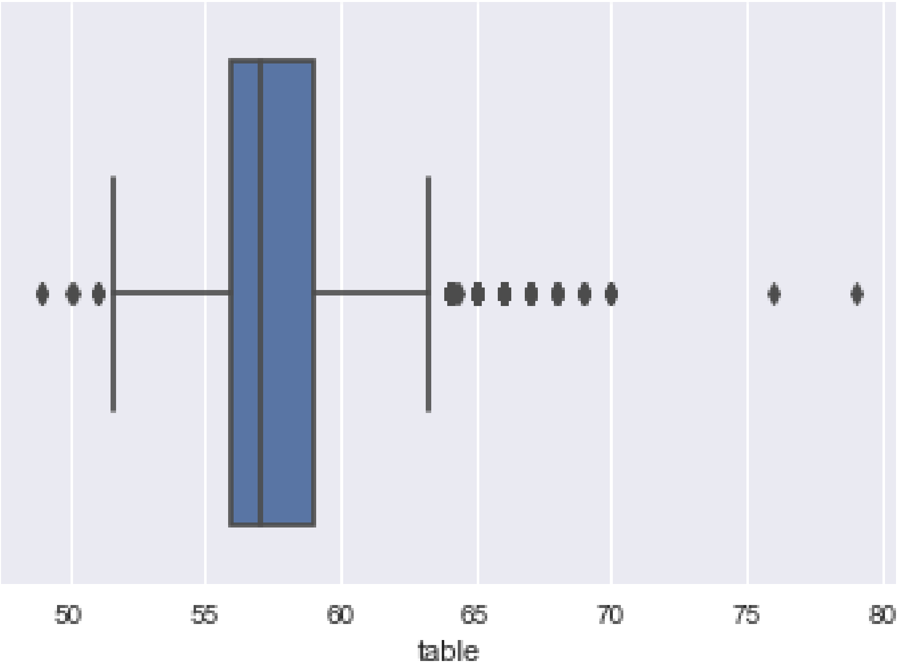


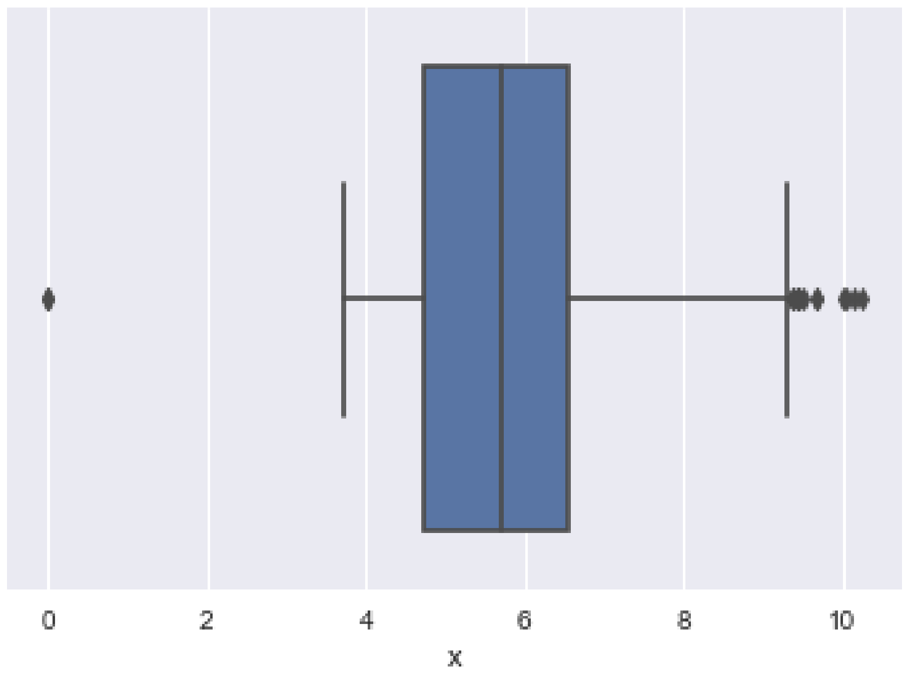
Observations: We have plotted the distribution plots of all the numerical variables. It can be observed that price is right skewed and depth is left skewed. No variables follow a normal distribution.

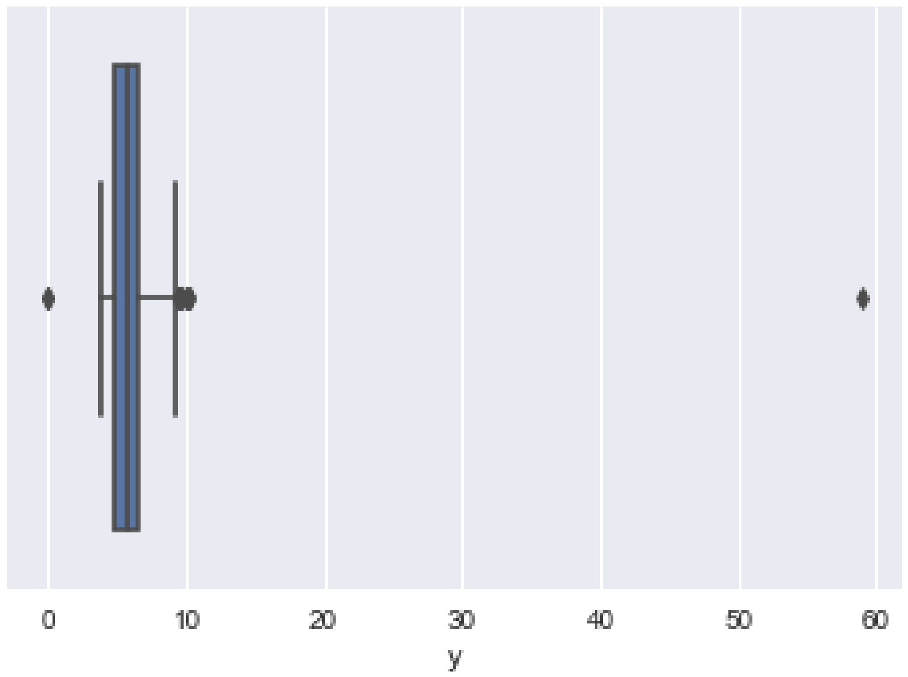

0
1
2
3
4
carat

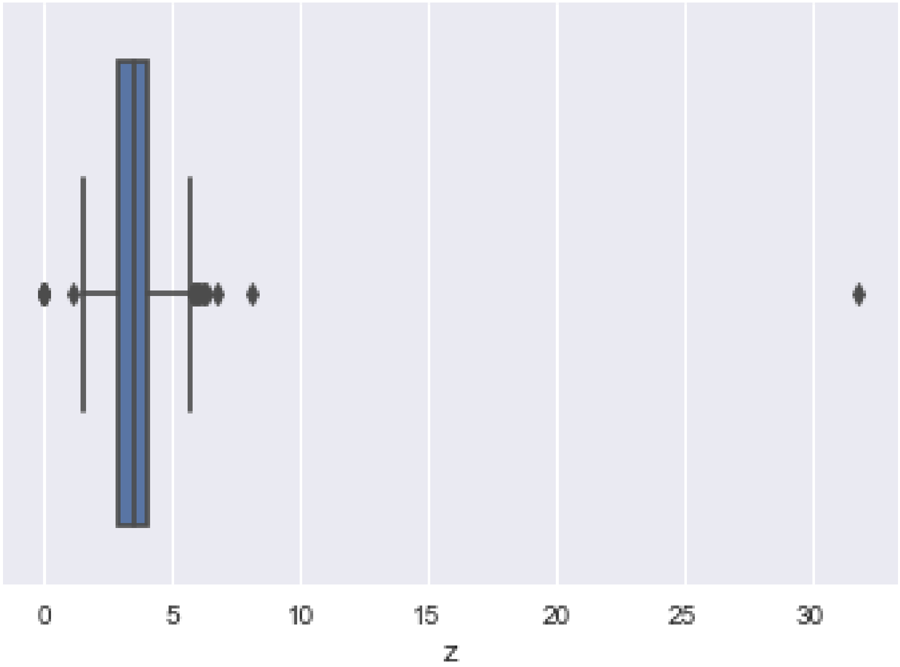

50
55
60
65
70
depth


0
2500
5000
7500
10000
12500
15000
17500
price


50
55
60
65
70
75
80
table

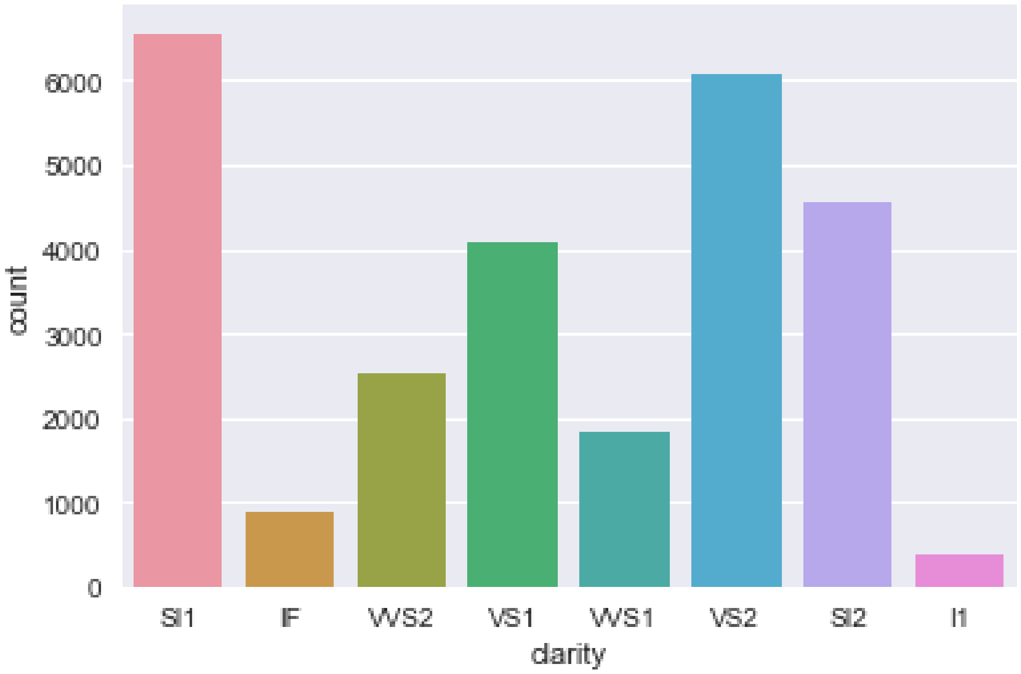

0
2
4
6
8
10
x

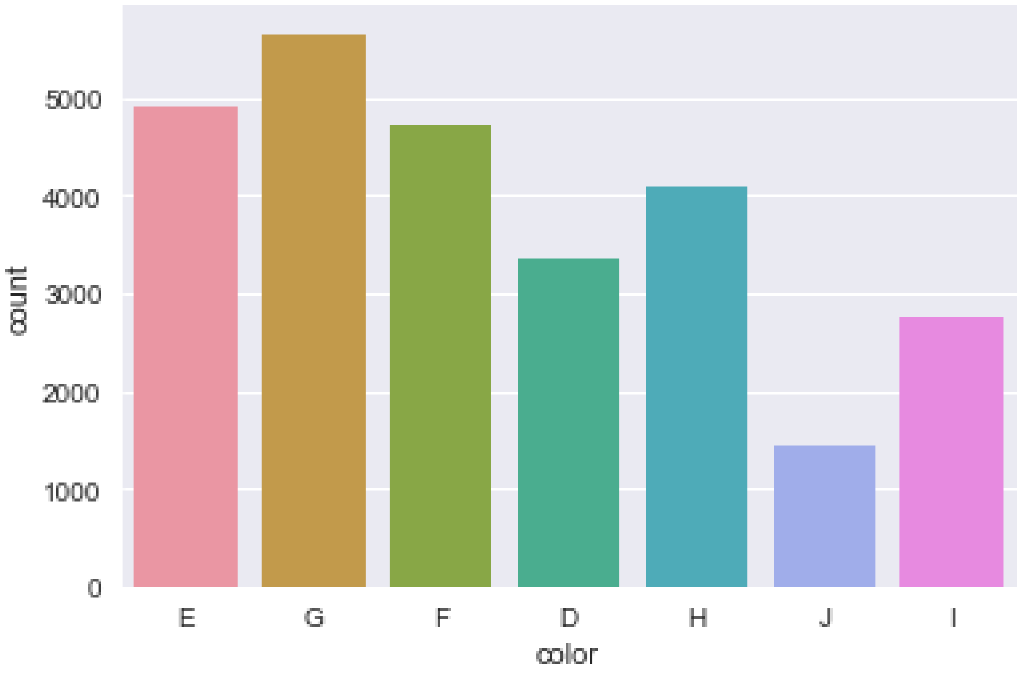

0
10
20
30
40
50
60
y

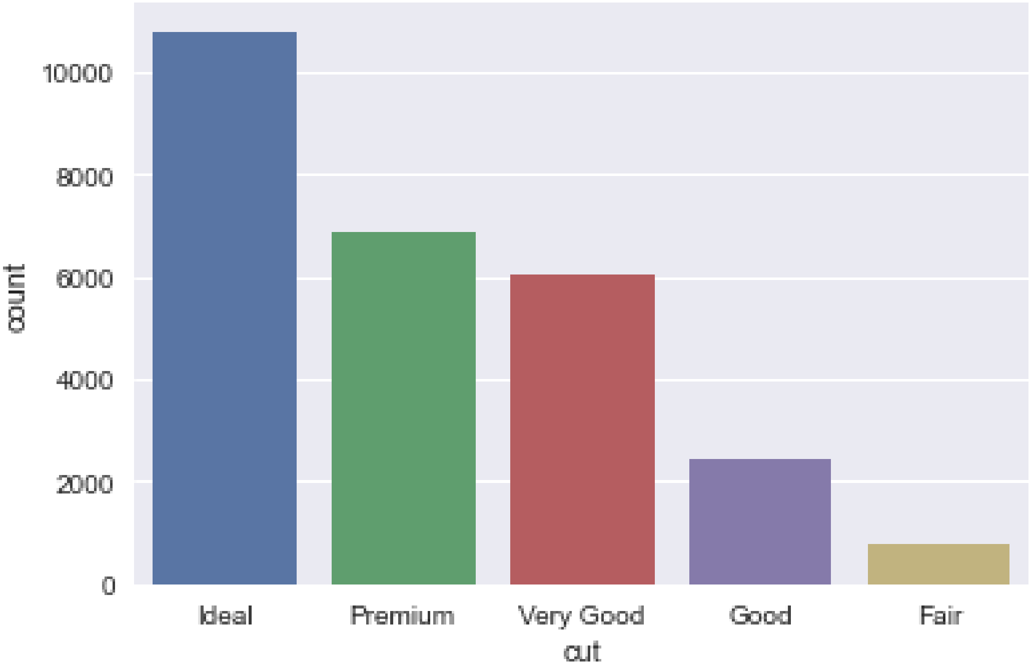


We have plotted the box plots of all the variables above and it is observed that there are outliers in all the variables.

Categorical Variable


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


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



Count plot of categorical variables are plotted above.

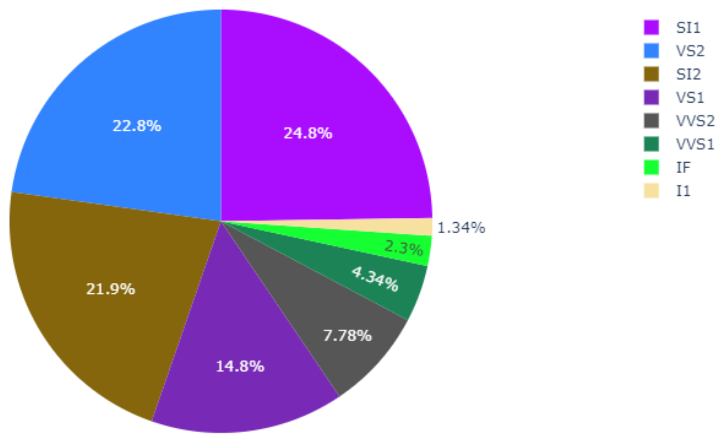
Color : Max no of stones are having G color

Cut : Max no of stones are having ideal cut

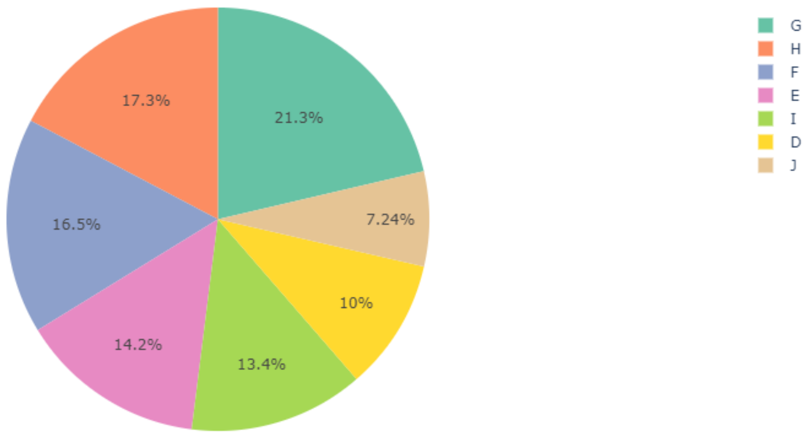
Clarity : max no of stones are having clarity SI1

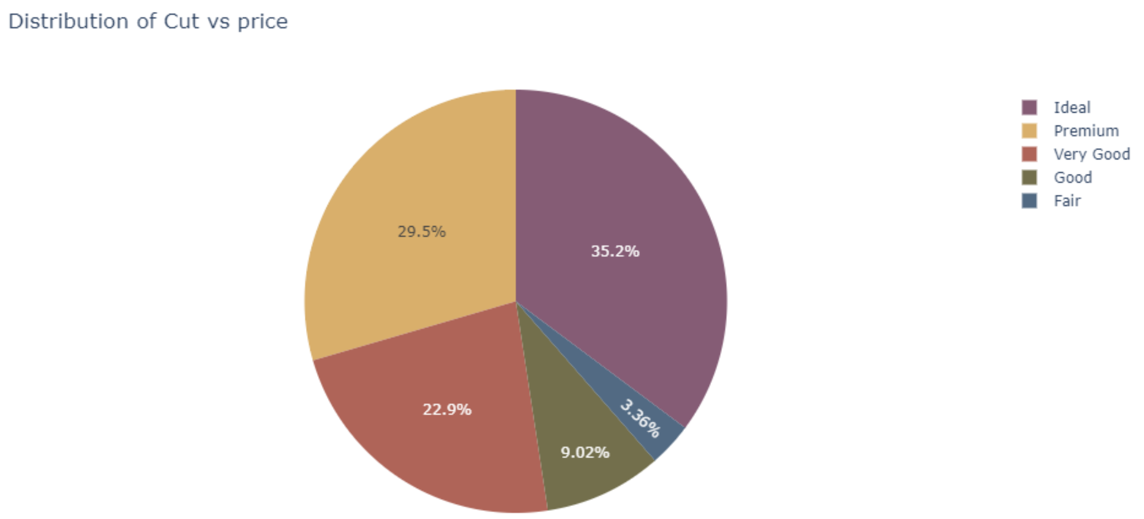
Bivariate Analysis

Distribution of Clarity vs price



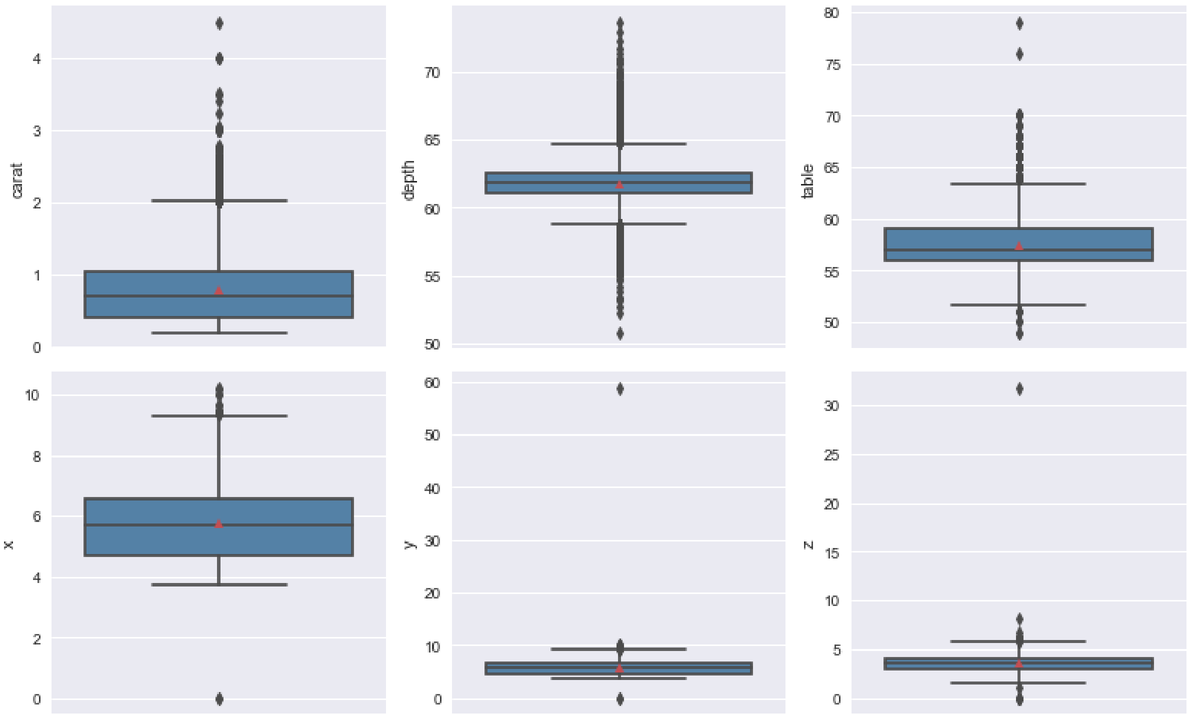
Distribution of Color vs price

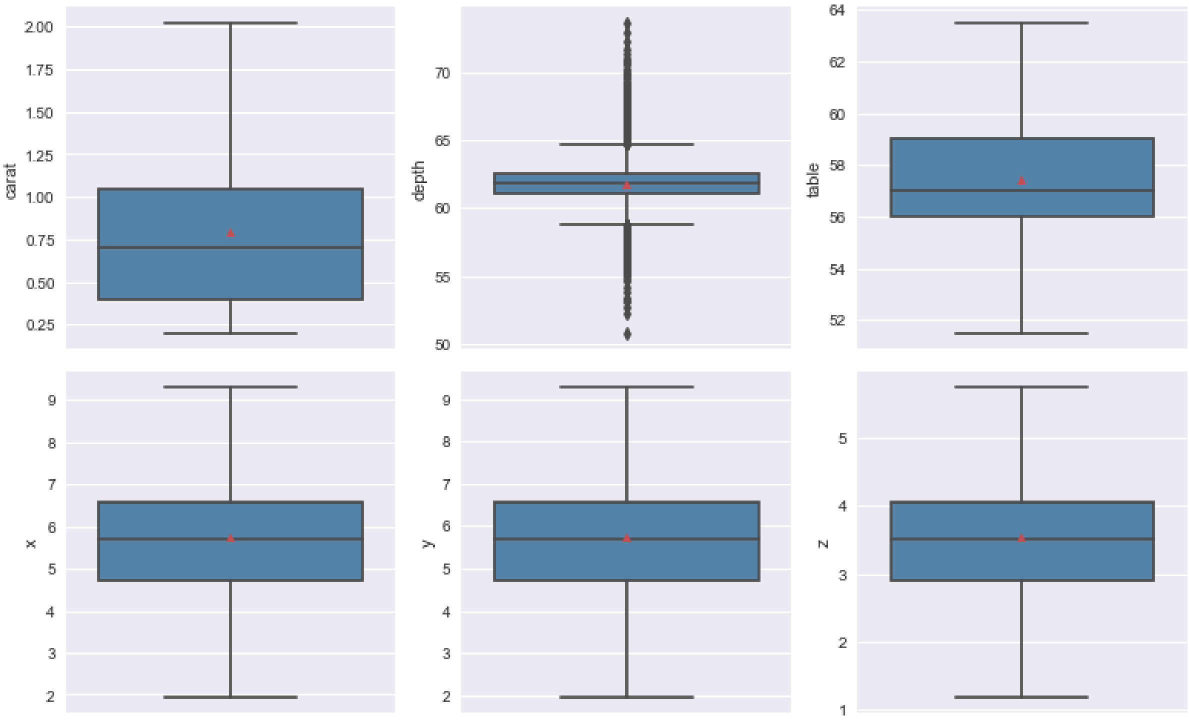


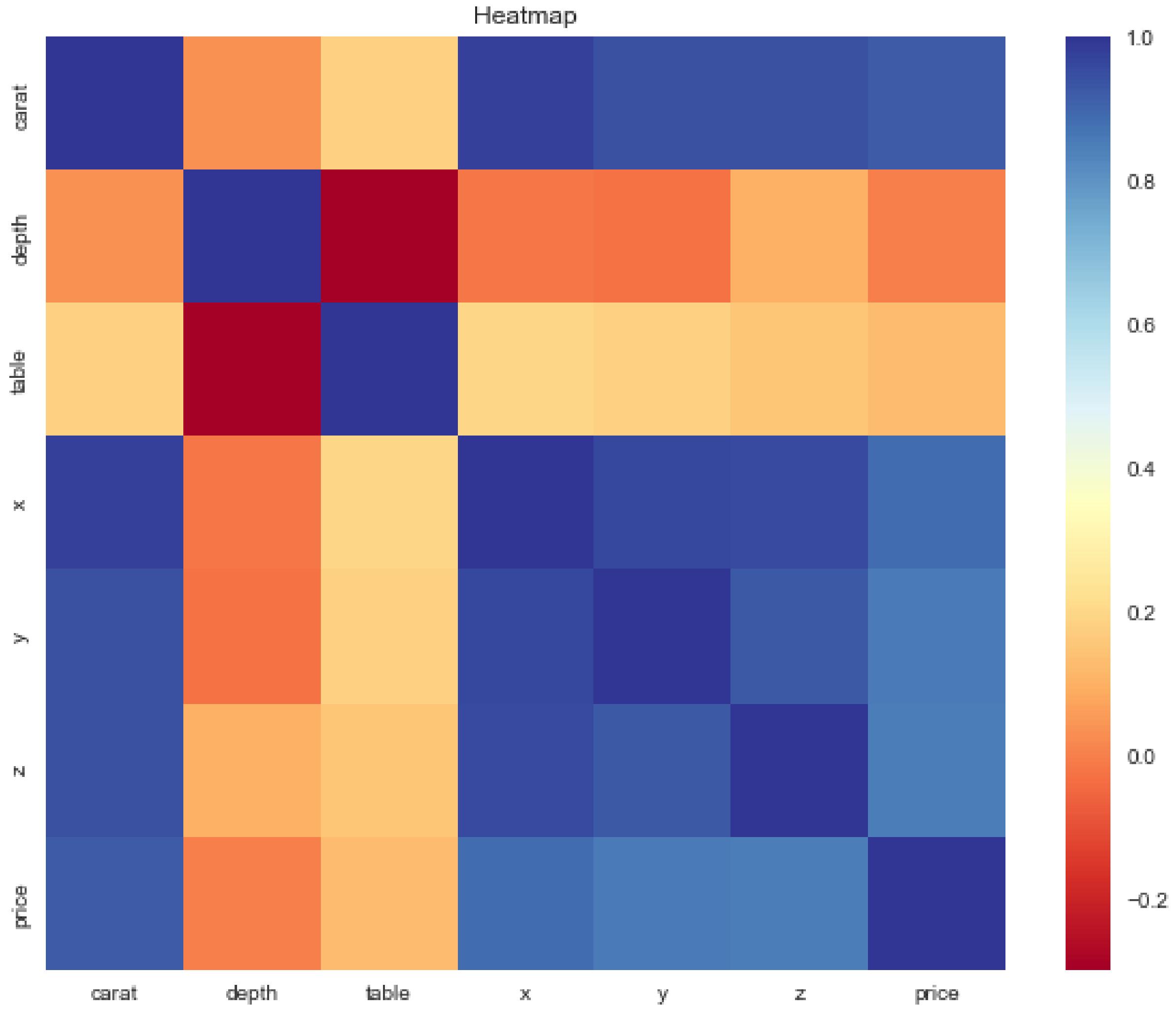


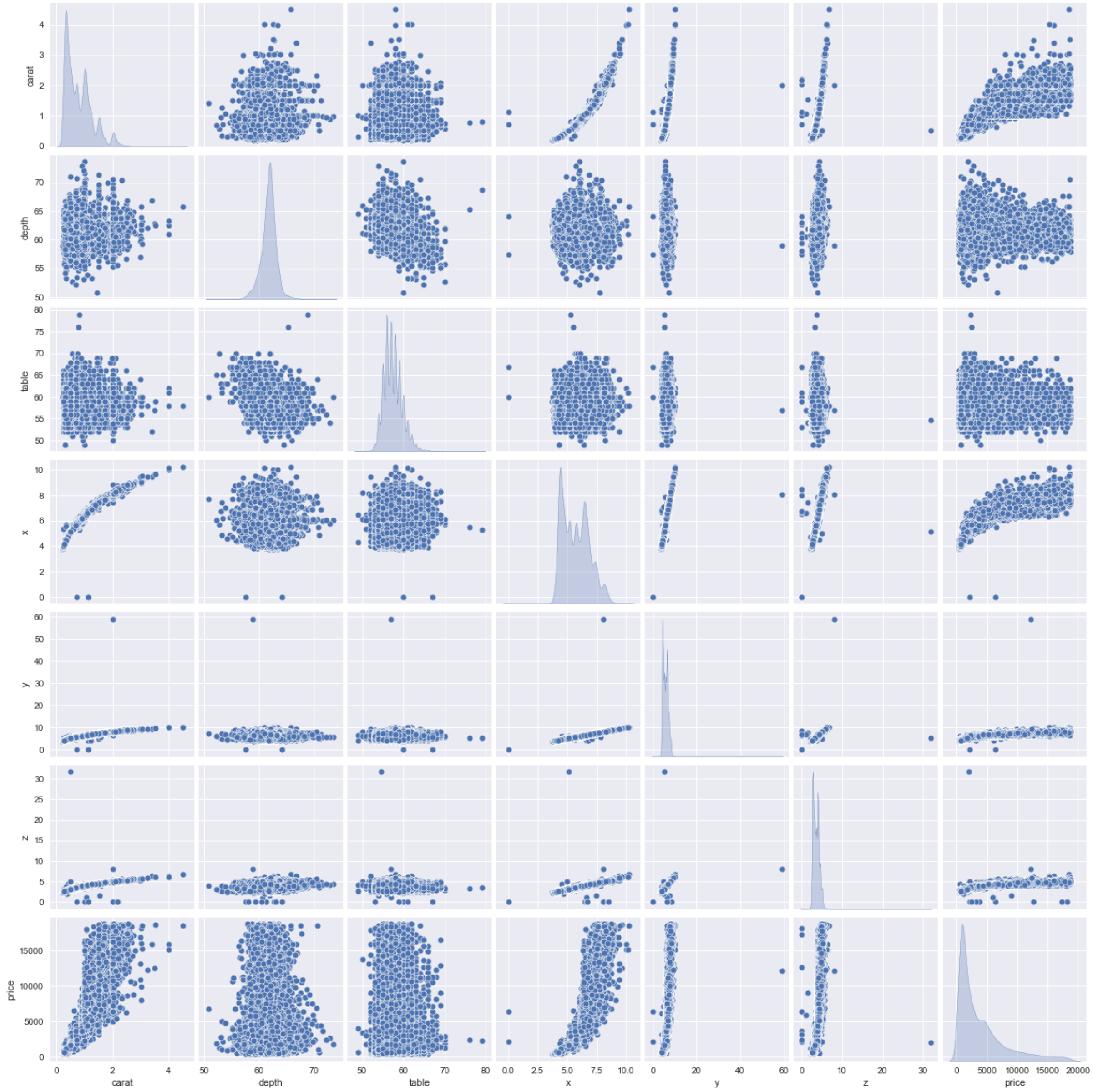
We have plotted the pie plot of three categorical variables against the price.

Multivariate Analysis


carat
0
1
2
3
4
depth
50
55
60
65
70
table
50
55
60
65
70
75
80
x
0
2
4
6
8
10
y
0
10
20
30
40
50
60
z
0
5
10
15
20
25
30


carat
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
depth
50
55
60
65
70
table
52
54
56
58
60
62
64
x
2
3
4
5
6
7
8
9
y
2
3
4
5
6
7
8
9
z
1
2
3
4
5

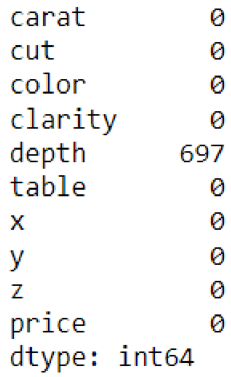

Heatmap
carat
depth
table
x
y
z
price
carat
depth
table
x
y
z
price
1.0
0.8
0.6
0.4
0.2
0.0
−0.2



From the multi variate analysis it can be observed that there is high correlation between carat ,x, y, & z and price.

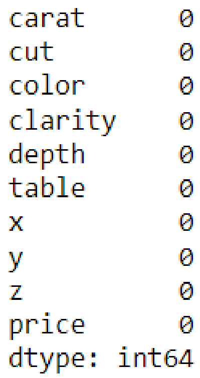
1.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Check for the possibility of combining the sub levels of a ordinal variables and take actions accordingly. Explain why you are combining these sub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

There are null values present in depth.



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

We have imputed the same using median



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

There are some rows where all the values in x, y and z are zero, these are mistake values so it has been dropped.

There are three categorical variables are they have sublevel, from the EDA part it can be observed that the sublevels are quite significantly impacting the target variable so combining them is not a good idea.

Each sublevel is different in its own so it does not have any similarity with other variables so it can not be combined.

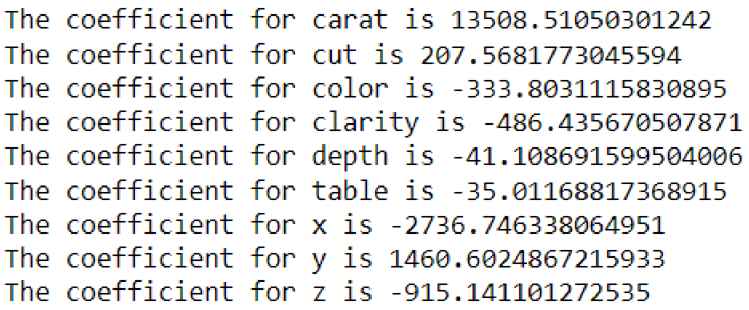
1.3 Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning.

The categorical variables have been encoded as per 1 being the worst and the highest no being the best.

Linear Regression Model Using SK learn

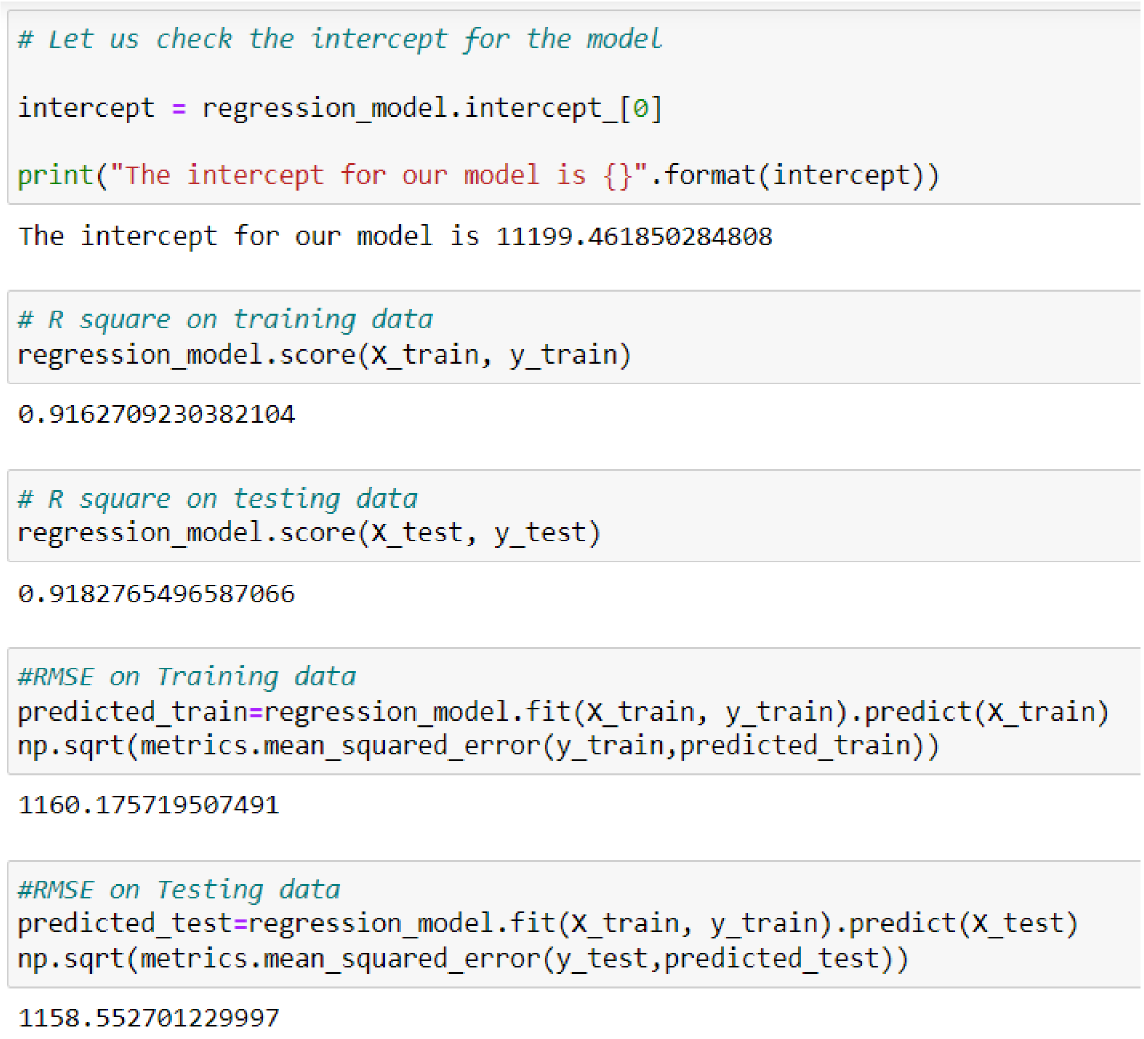
It is split using train test split by 70:30 ration for train and test data.

The linear regression model has been built and the below coefficients are observed.



```
The coefficient for carat is 13508.51050301242
The coefficient for cut is 207.5681773045594
The coefficient for color is -333.8031115830895
The coefficient for clarity is -486.435670507871
The coefficient for depth is -41.108691599504006
The coefficient for table is -35.01168817368915
The coefficient for x is -2736.746338064951
The coefficient for y is 1460.6024867215933
The coefficient for z is -915.141101272535
```

Metrics of Regular LR Model



```
# Let us check the intercept for the model

intercept = regression_model.intercept_[0]

print("The intercept for our model is {}".format(intercept))
```

```
The intercept for our model is 11199.461850284808
```

```
# R square on training data
regression_model.score(X_train, y_train)
```

```
0.9162709230382104
```

```
# R square on testing data
regression_model.score(X_test, y_test)
```

```
0.9182765496587066
```

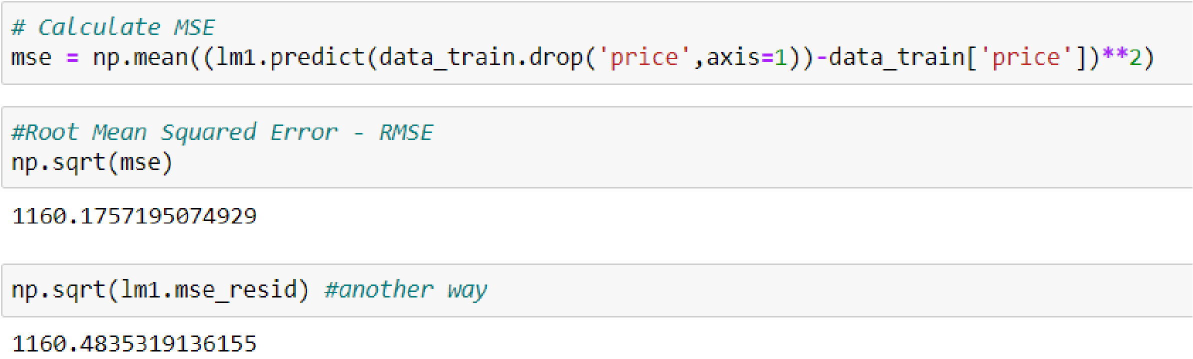
```
#RMSE on Training data
predicted_train=regression_model.fit(X_train, y_train).predict(X_train)
np.sqrt(metrics.mean_squared_error(y_train,predicted_train))
```

```
1160.175719507491
```

```
#RMSE on Testing data
predicted_test=regression_model.fit(X_train, y_train).predict(X_test)
np.sqrt(metrics.mean_squared_error(y_test,predicted_test))
```

```
1158.552701229997
```

Linear Regression Model Using Stats Model



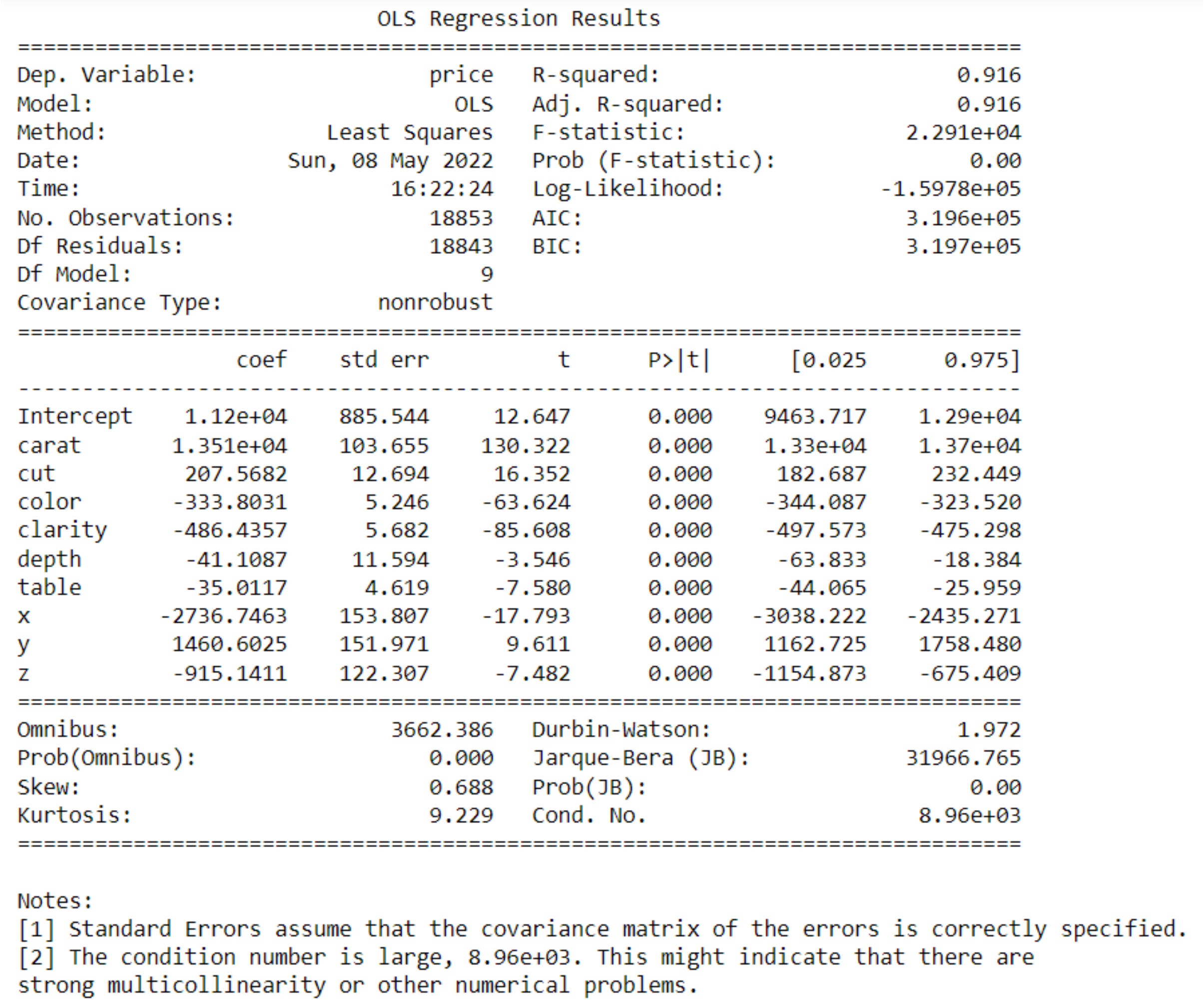
```
# Calculate MSE
mse = np.mean((lm1.predict(data_train.drop('price',axis=1))-data_train['price'])**2)
```

```
#Root Mean Squared Error - RMSE
np.sqrt(mse)
```

```
1160.1757195074929
```

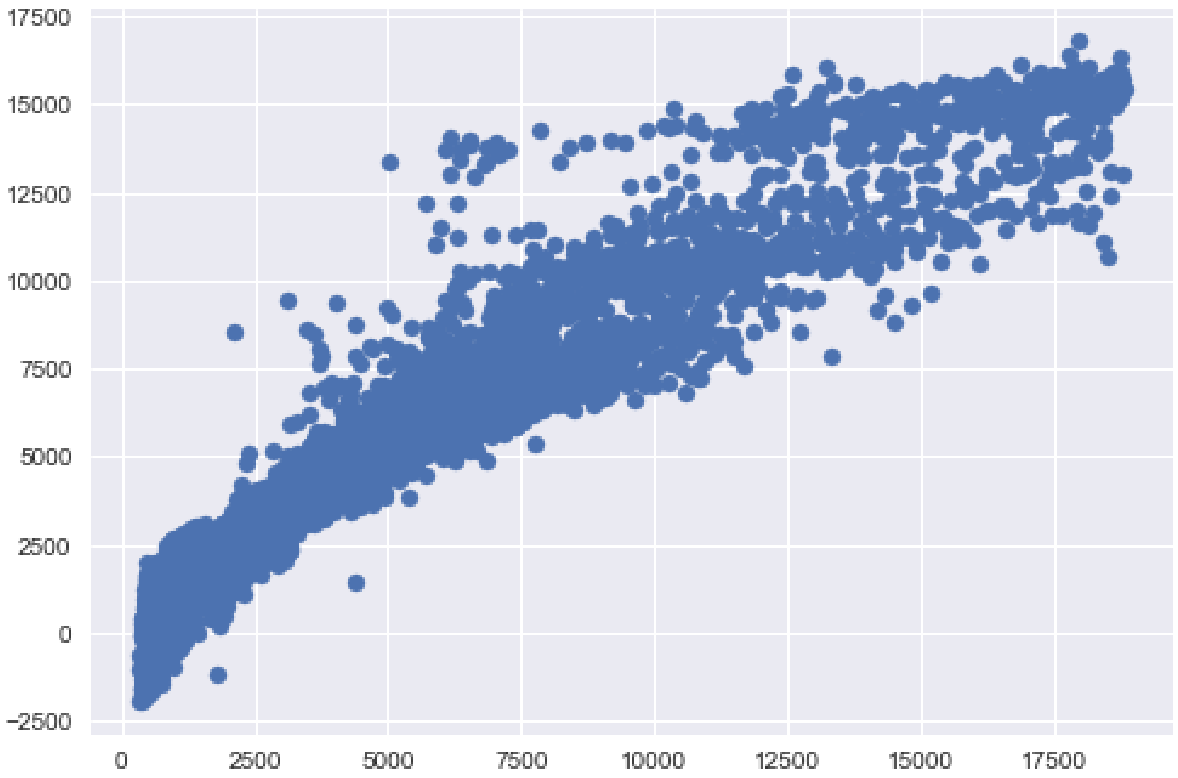
```
np.sqrt(lm1.mse_resid) #another way
```

```
1160.4835319136155
```



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                  price   R-squared:                       0.916
Model:                            OLS   Adj. R-squared:                  0.916
Method:                 Least Squares   F-statistic:                 2.291e+04
Date:                Sun, 08 May 2022   Prob (F-statistic):               0.00
Time:                        16:22:24   Log-Likelihood:            -1.5978e+05
No. Observations:               18853   AIC:                         3.196e+05
Df Residuals:                   18843   BIC:                         3.197e+05
Df Model:                           9                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     1.12e+04    885.544     12.647      0.000    9463.717    1.29e+04
carat        1.351e+04    103.655    130.322      0.000     1.33e+04    1.37e+04
cut           207.5682     12.694     16.352      0.000     182.687     232.449
color        -333.8031      5.246    -63.624      0.000    -344.087    -323.520
clarity      -486.4357      5.682    -85.608      0.000    -497.573    -475.298
depth         -41.1087     11.594     -3.546      0.000     -63.833     -18.384
table         -35.0117      4.619     -7.580      0.000     -44.065     -25.959
x           -2736.7463    153.807    -17.793      0.000   -3038.222   -2435.271
y            1460.6025    151.971      9.611      0.000    1162.725    1758.480
z            -915.1411    122.307     -7.482      0.000   -1154.873    -675.409
==============================================================================
Omnibus:                     3662.386   Durbin-Watson:                   1.972
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            31966.765
Skew:                           0.688   Prob(JB):                         0.00
Kurtosis:                       9.229   Cond. No.                     8.96e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.96e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

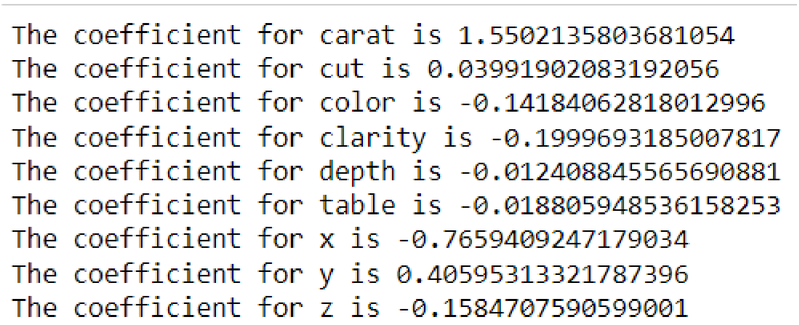
Scatter plot of Predicted Price using stats model LR



Linear Regression Model by Scaling the data

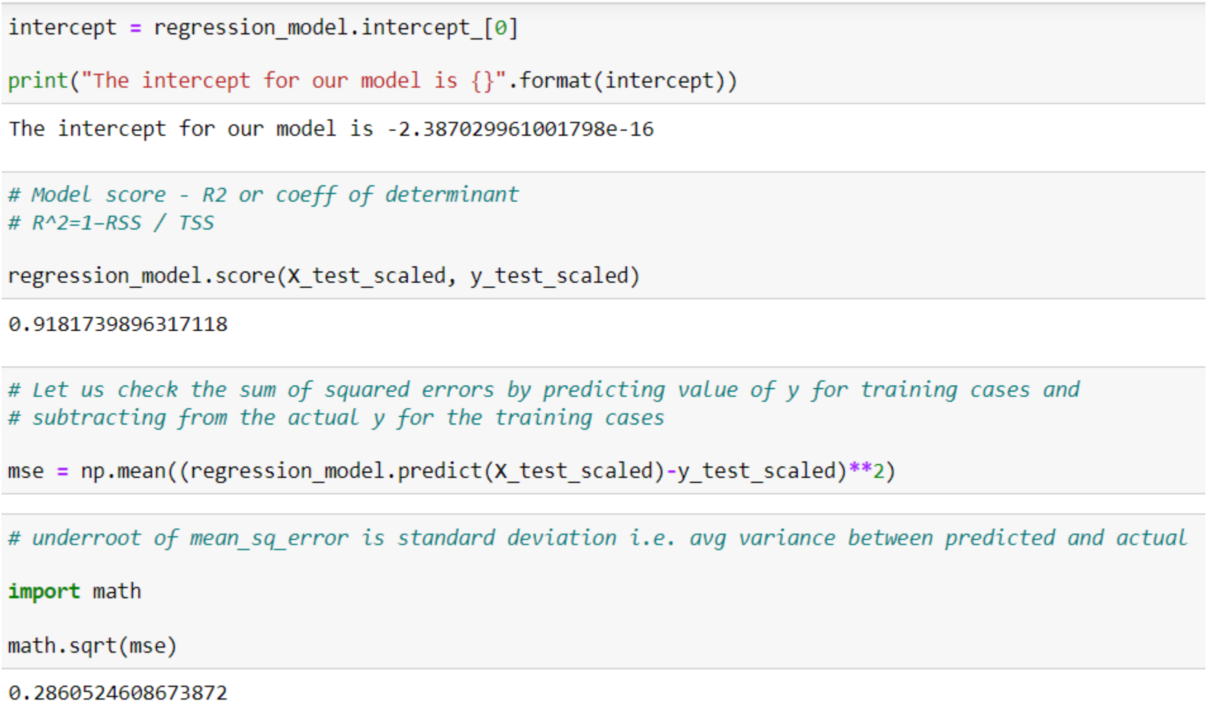
We have scaled the data using z score and then fit transform it with the model.

Below is the coefficients observed



```
The coefficient for carat is 1.5502135803681054
The coefficient for cut is 0.03991902083192056
The coefficient for color is -0.14184062818012996
The coefficient for clarity is -0.1999693185007817
The coefficient for depth is -0.012408845565690881
The coefficient for table is -0.018805948536158253
The coefficient for x is -0.7659409247179034
The coefficient for y is 0.40595313321787396
The coefficient for z is -0.1584707590599001
```

The metrics of MSE and RMSE are as per below



```
intercept = regression_model.intercept_[0]

print("The intercept for our model is {}".format(intercept))
```

```
The intercept for our model is -2.387029961001798e-16
```

```
# Model score - R2 or coeff of determinant
# R^2=1-RSS / TSS

regression_model.score(X_test_scaled, y_test_scaled)
```

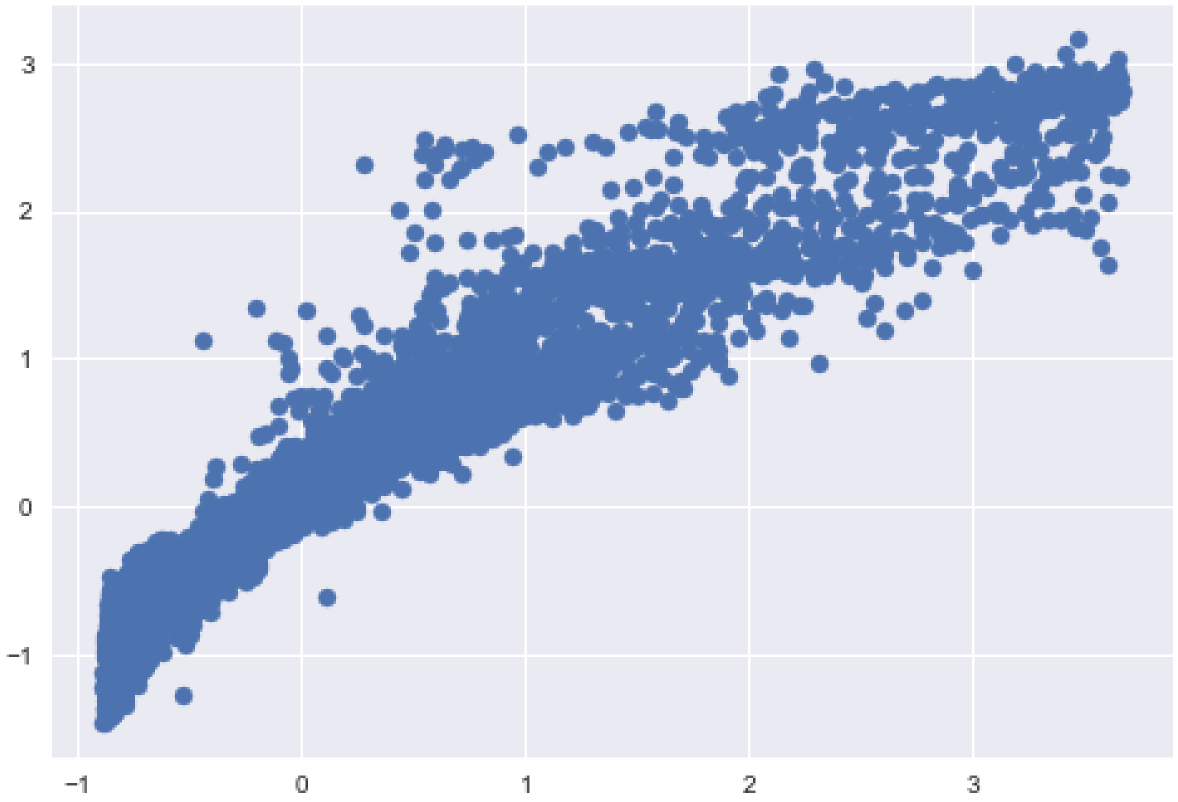
```
0.9181739896317118
```

```
# Let us check the sum of squared errors by predicting value of y for training cases and
# subtracting from the actual y for the training cases

mse = np.mean((regression_model.predict(X_test_scaled)-y_test_scaled)**2)
```

```
# underroot of mean_sq_error is standard deviation i.e. avg variance between predicted and actual

import math

math.sqrt(mse)
```

```
0.2860524608673872
```

Scatter plot of predicted price in scaled LR model.



	LR by using SK learn	LR by using Stats Model	LR by scaling the data
R square		0.916	0.918
R square training	0.916		
R square testing	0.918		
Adjusted R square		0.916	
RMSE		1160.17	0.28
RMSE training	1160.17		
RMSE testing	1158.55		

Considering the above values of R square we can say that LR model after scaling the data is best.

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

From the above linear regression model it can be observed that the best factors affecting price of the cubic zirconia are carat, cut, width, color and depth.

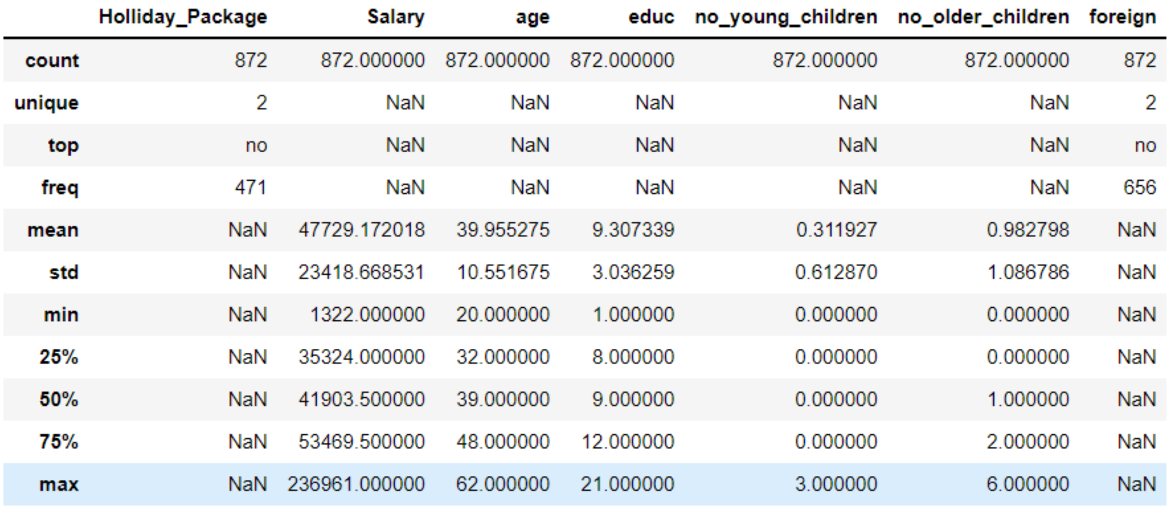
The company should focus on the attributes of length, height, table and clarity as they are not able to make the product profitable.

Length, width and height are strongly correlated to price of the stone.

Problem-2 Linear Discriminant Analysis and Logistic Regression

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.

There are 872 observations 7 variables. No duplicate values no missing values.



	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
count	872	872.000000	872.000000	872.000000	872.000000	872.000000	872
unique	2	NaN	NaN	NaN	NaN	NaN	2
top	no	NaN	NaN	NaN	NaN	NaN	no
freq	471	NaN	NaN	NaN	NaN	NaN	656
mean	NaN	47729.172018	39.955275	9.307339	0.311927	0.982798	NaN
std	NaN	23418.668531	10.551675	3.036259	0.612870	1.086786	NaN
min	NaN	1322.000000	20.000000	1.000000	0.000000	0.000000	NaN
25%	NaN	35324.000000	32.000000	8.000000	0.000000	0.000000	NaN
50%	NaN	41903.500000	39.000000	9.000000	0.000000	1.000000	NaN
75%	NaN	53469.500000	48.000000	12.000000	0.000000	2.000000	NaN
max	NaN	236961.000000	62.000000	21.000000	3.000000	6.000000	NaN

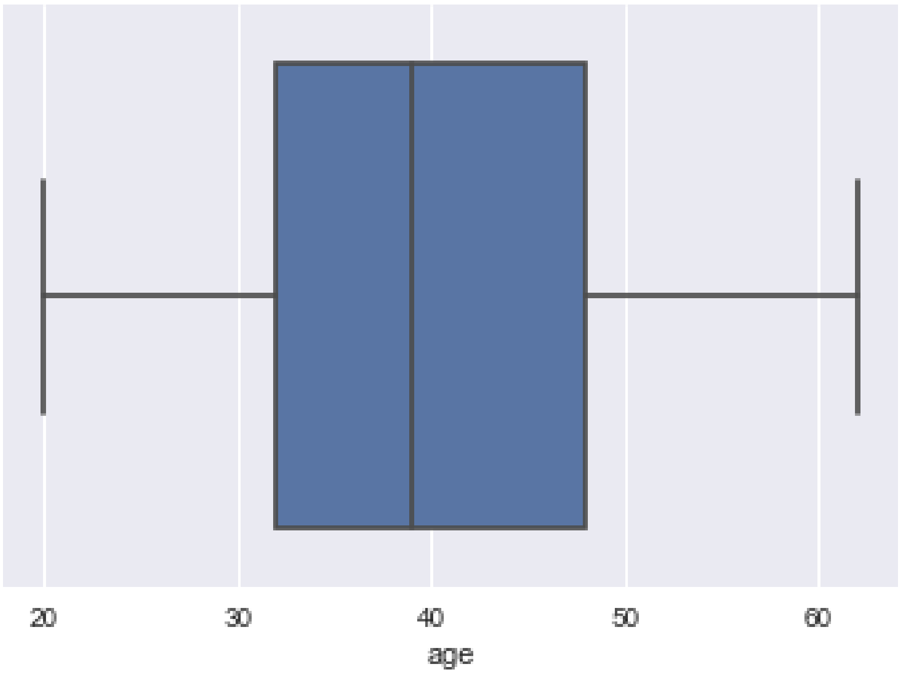
Out of 7 variables two are categorical and 5 numerical.

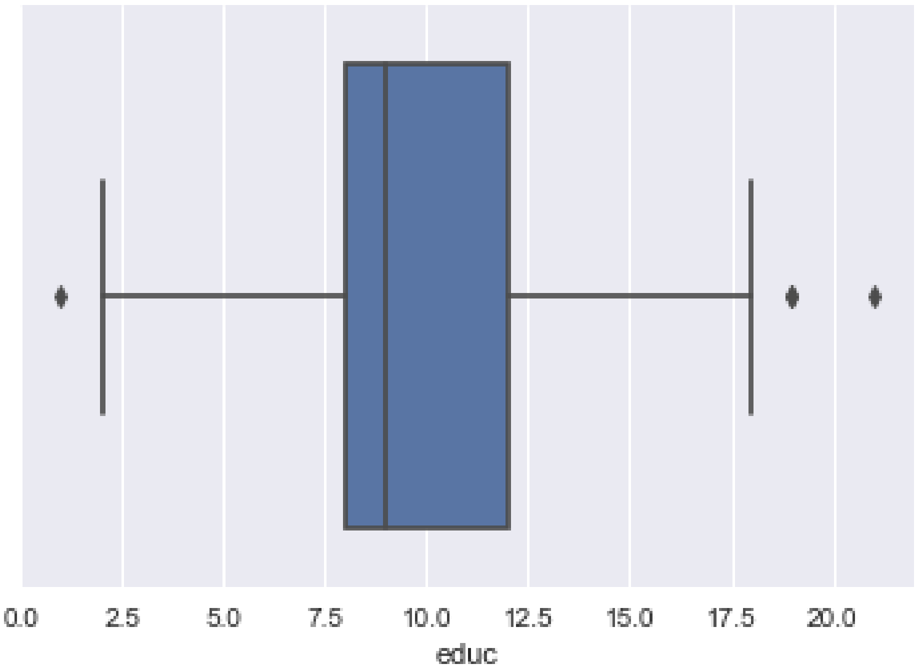
The target variable is Holiday package.

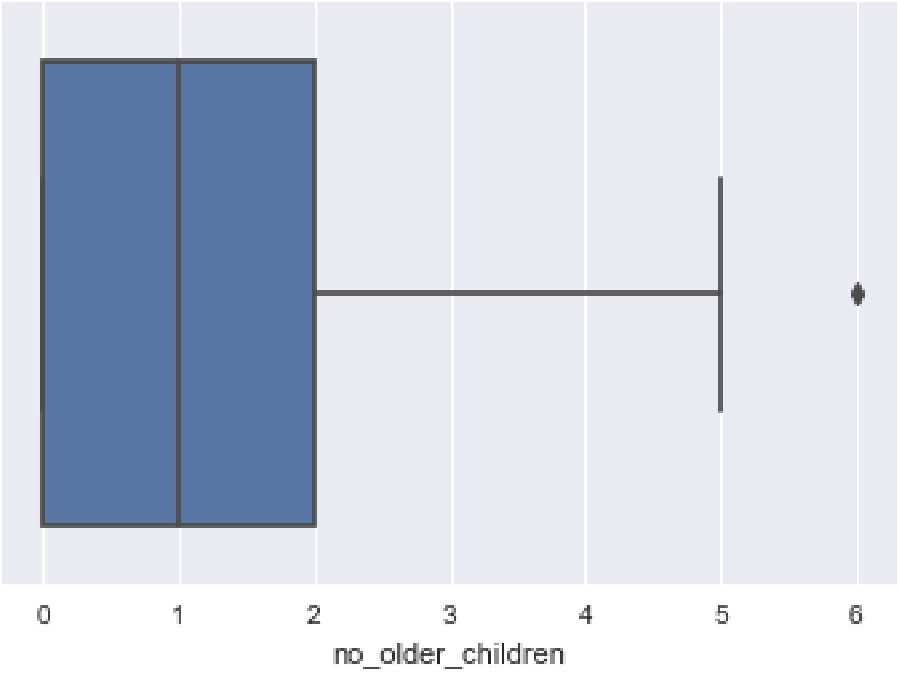
Exploratory data analysis

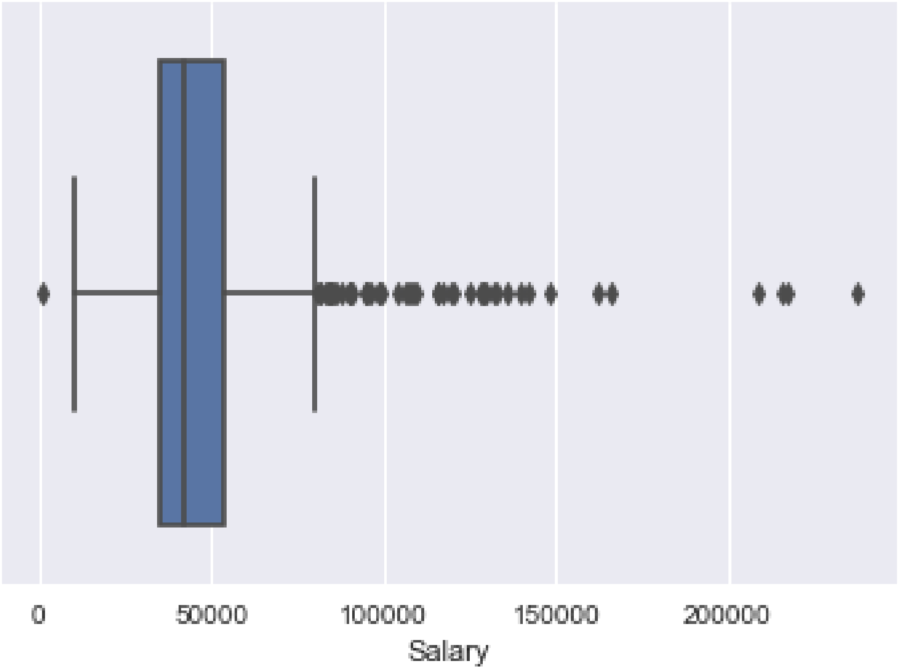
Univariate Analysis

Numerical Variables

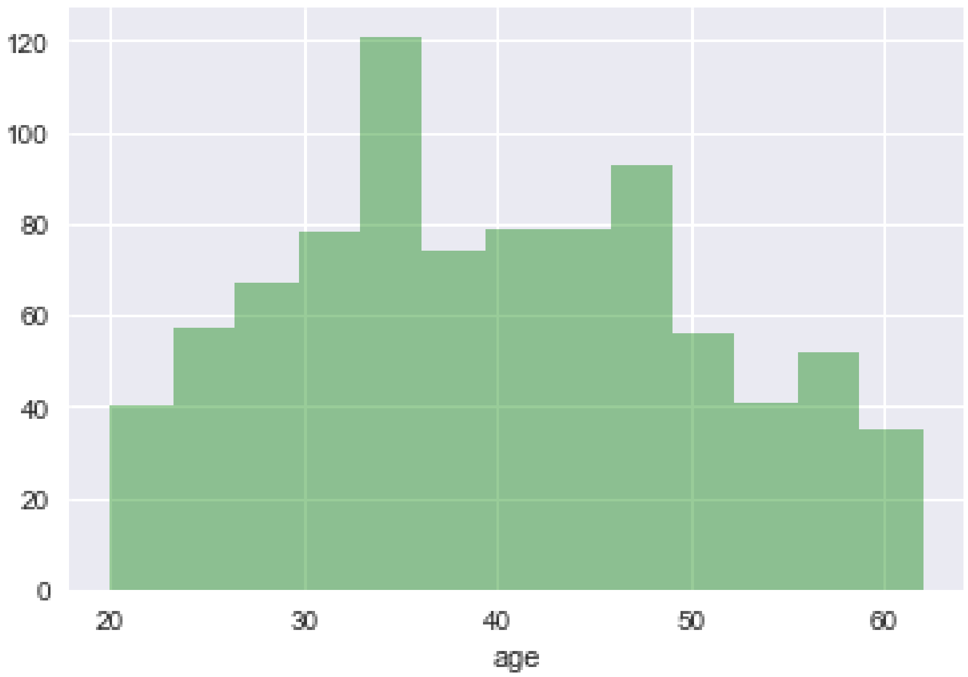

20
30
40
50
60
age

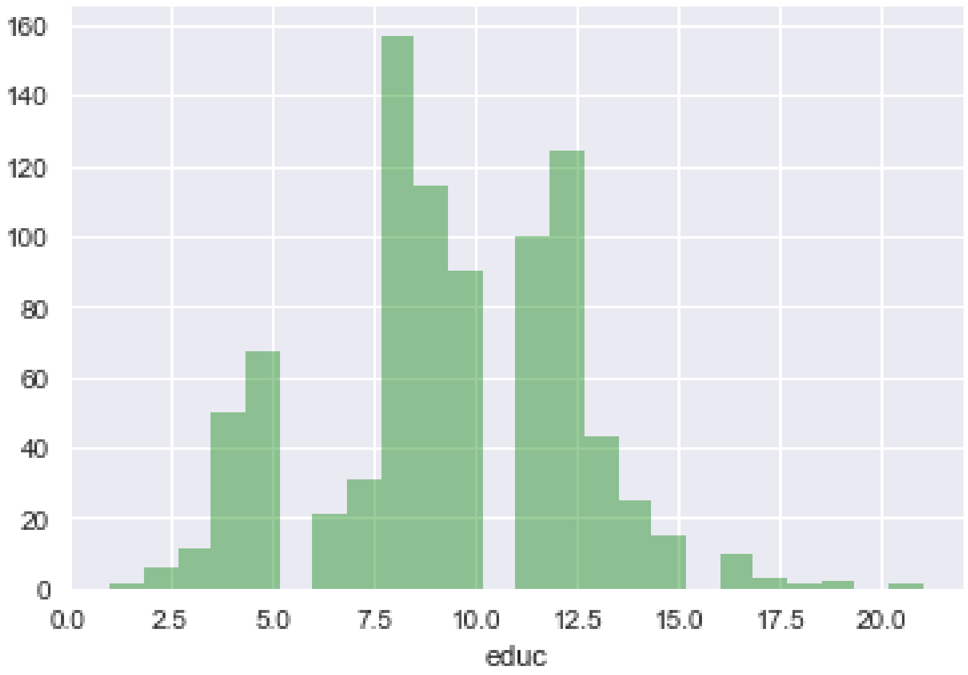

0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
educ

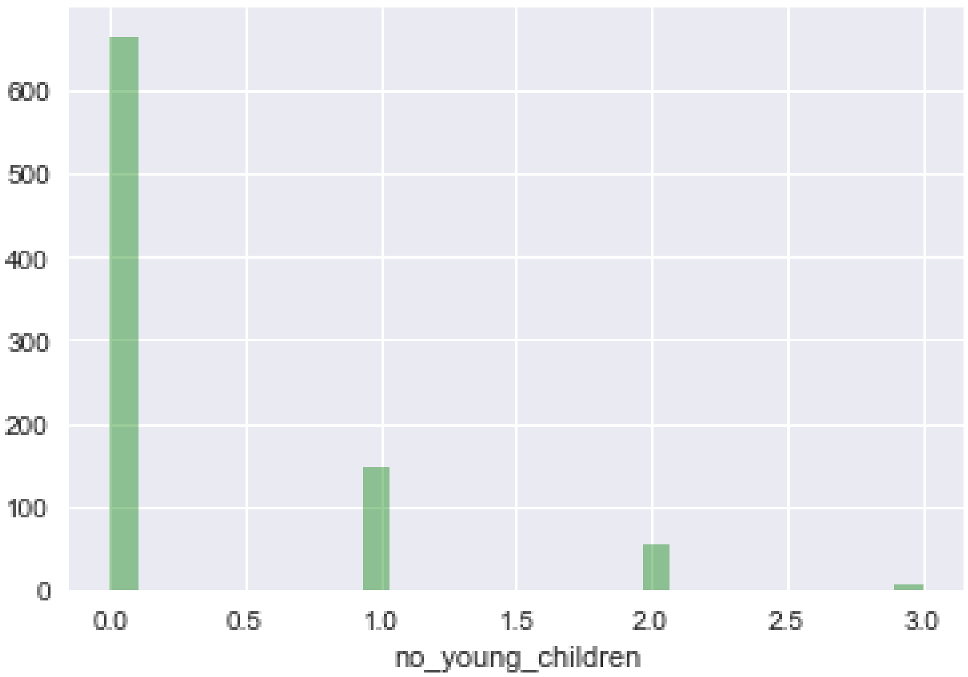


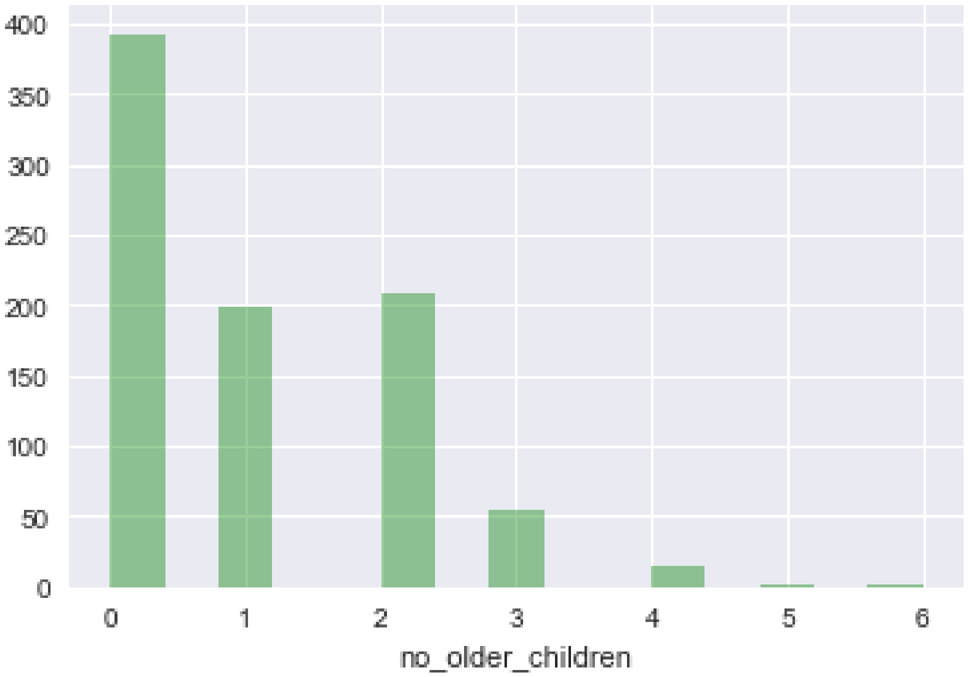


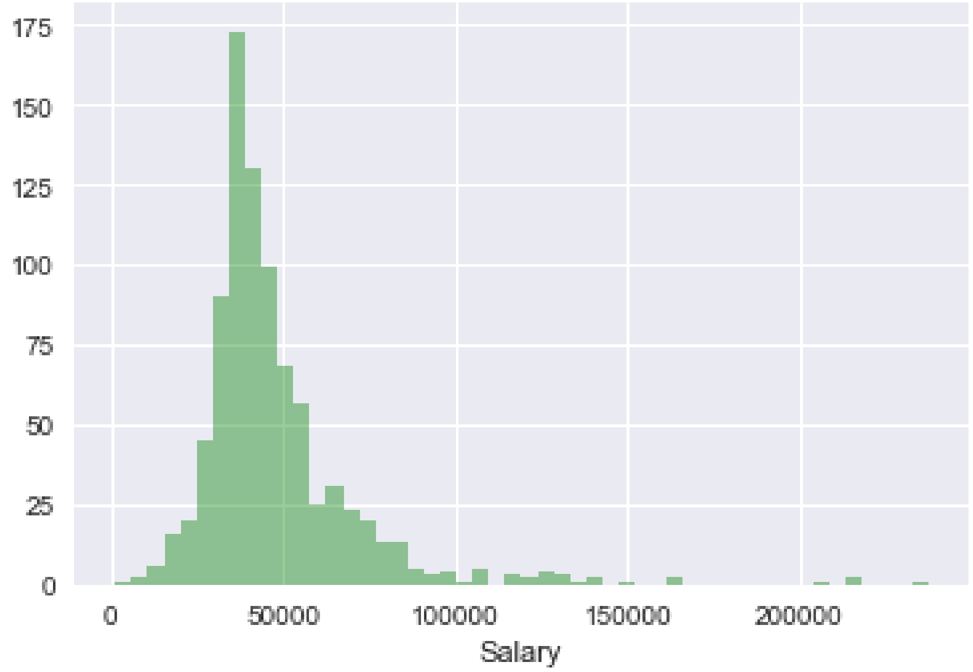
From the above box plots it can be observed that salary, and education years have outliers.


120
100
80
60
40
20
0
20
30
40
50
60
age


160
140
120
100
80
60
40
20
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
educ

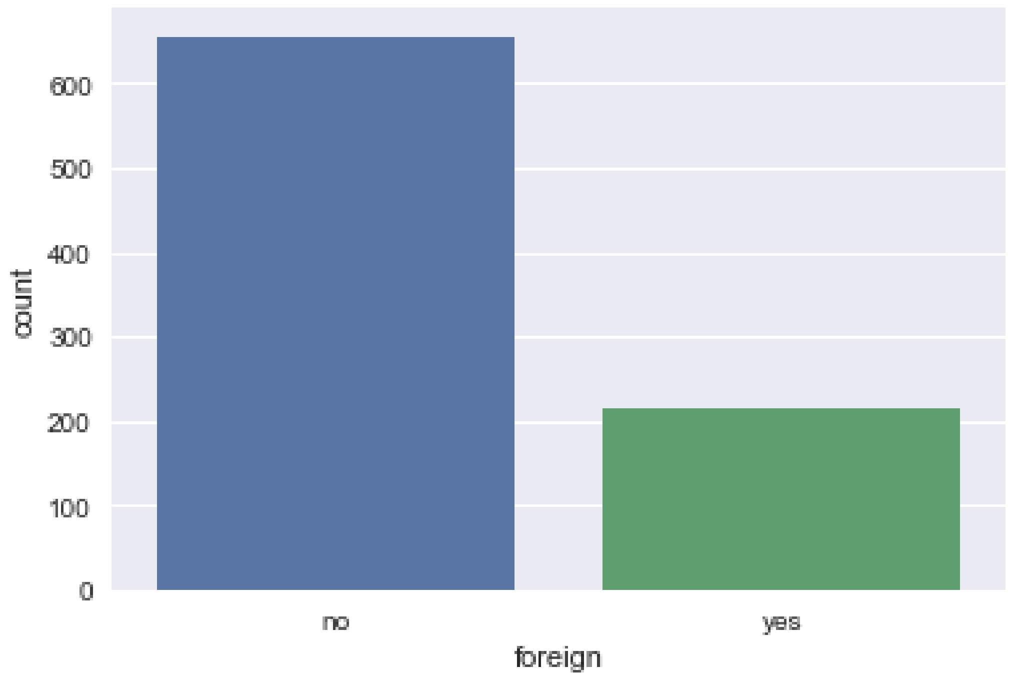

600
500
400
300
200
100
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
no_young_children

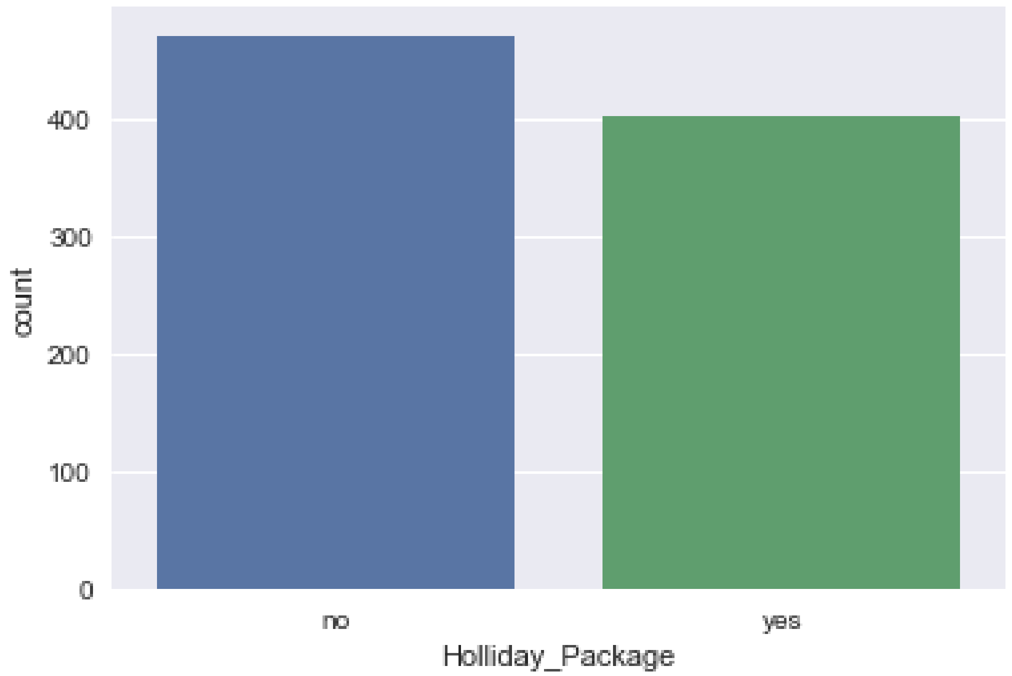

400
350
300
250
200
150
100
50
0
0
1
2
3
4
5
6
no_older_children



From the above distribution plot it can be observed that salary is left skewed and no variable is normally distributed.

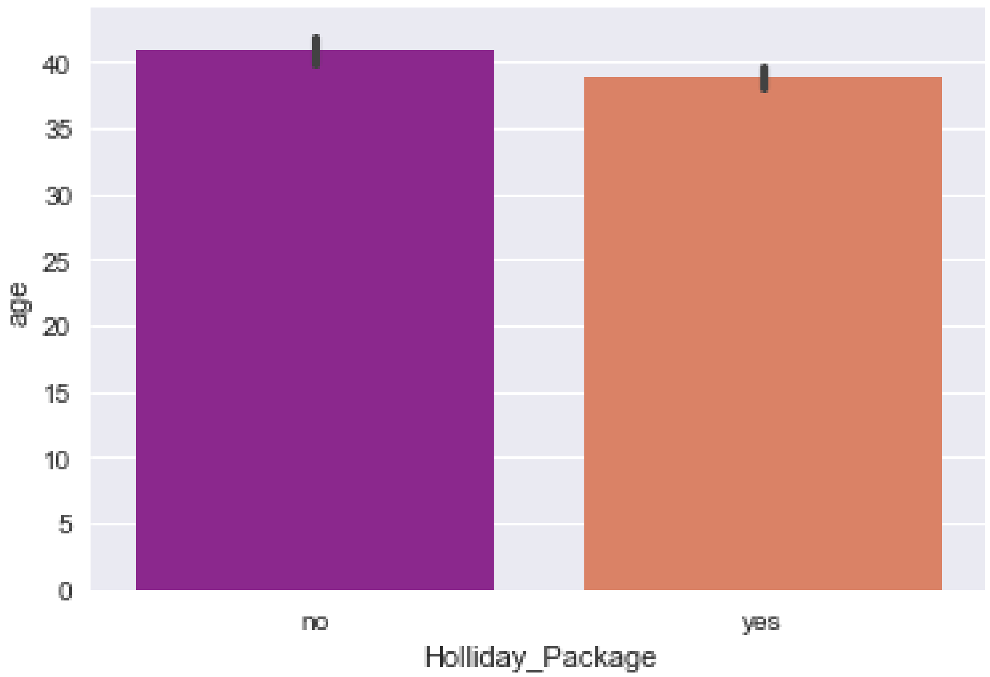
Categorical Variable

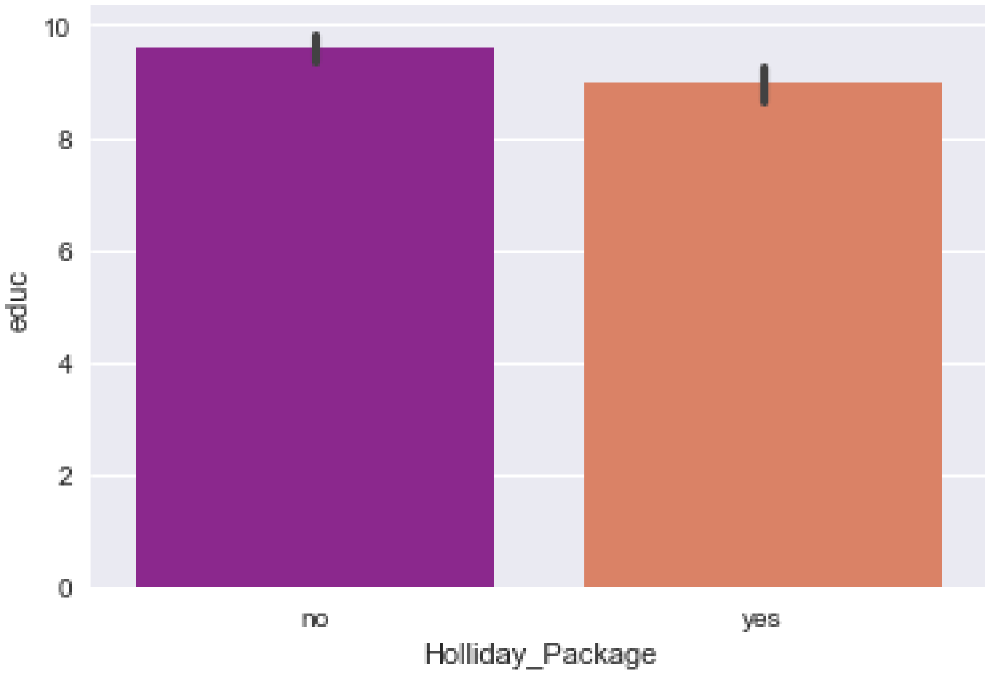


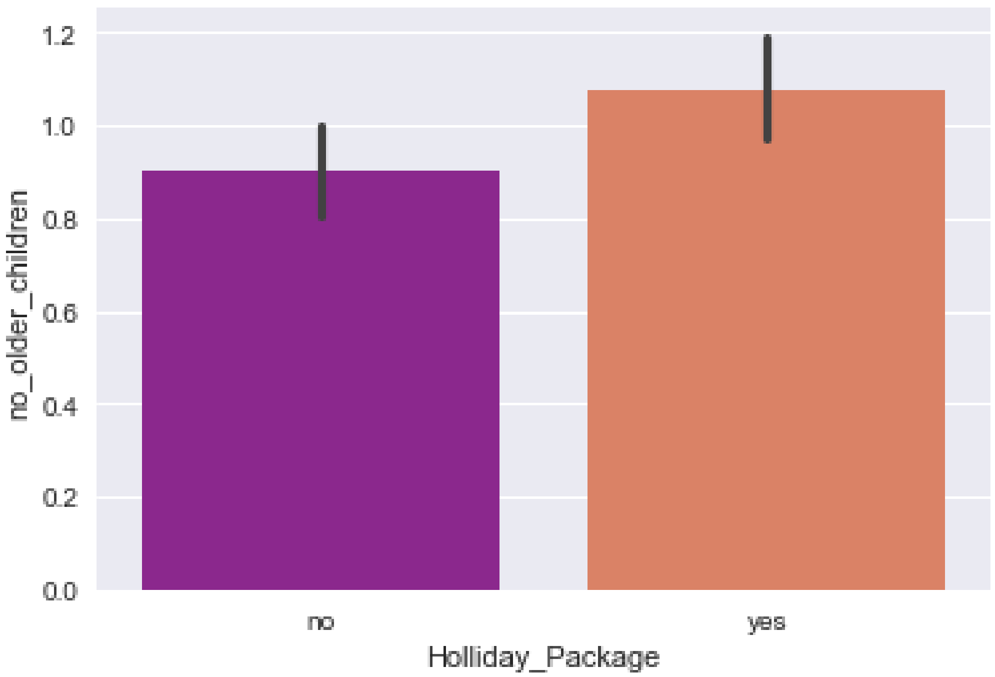


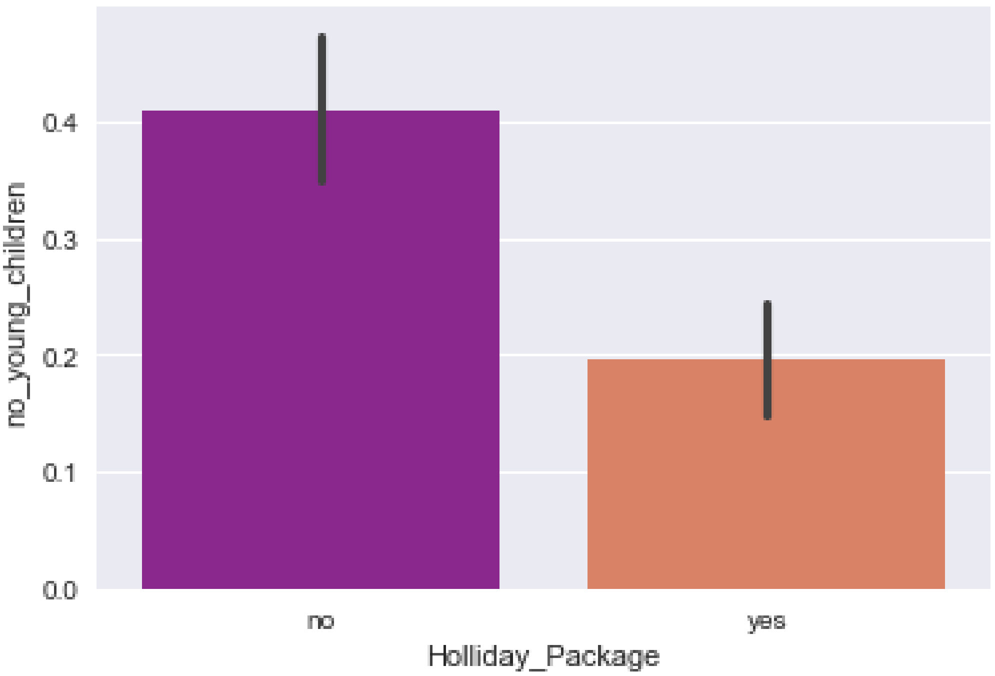
More people have not taken the package and those who have taken are not foreigners.

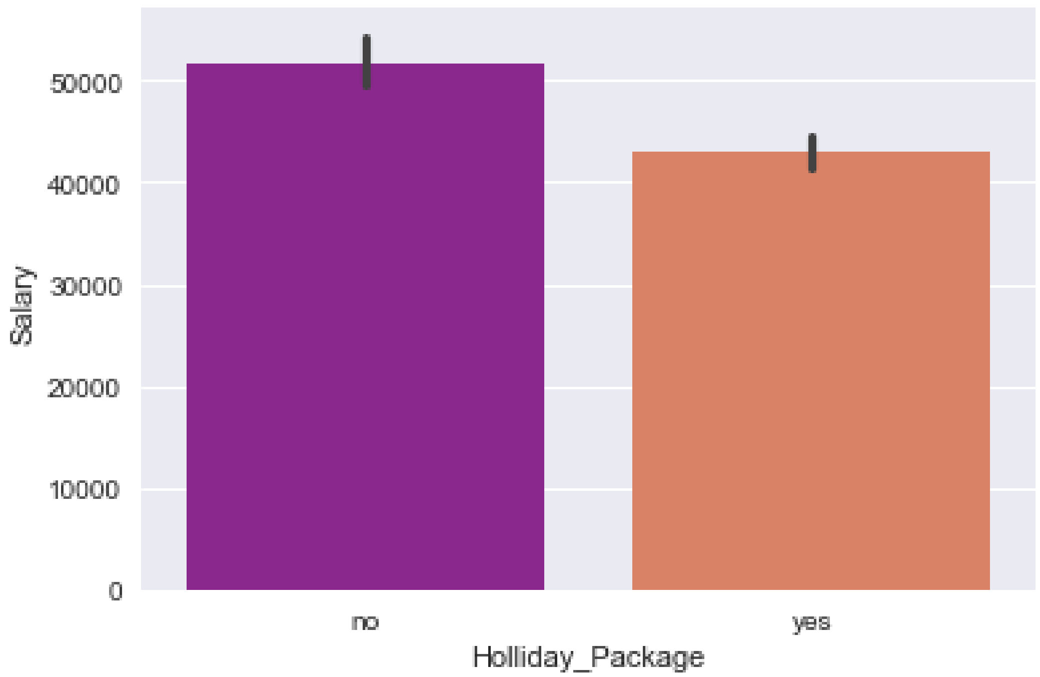
Bivariate Analysis

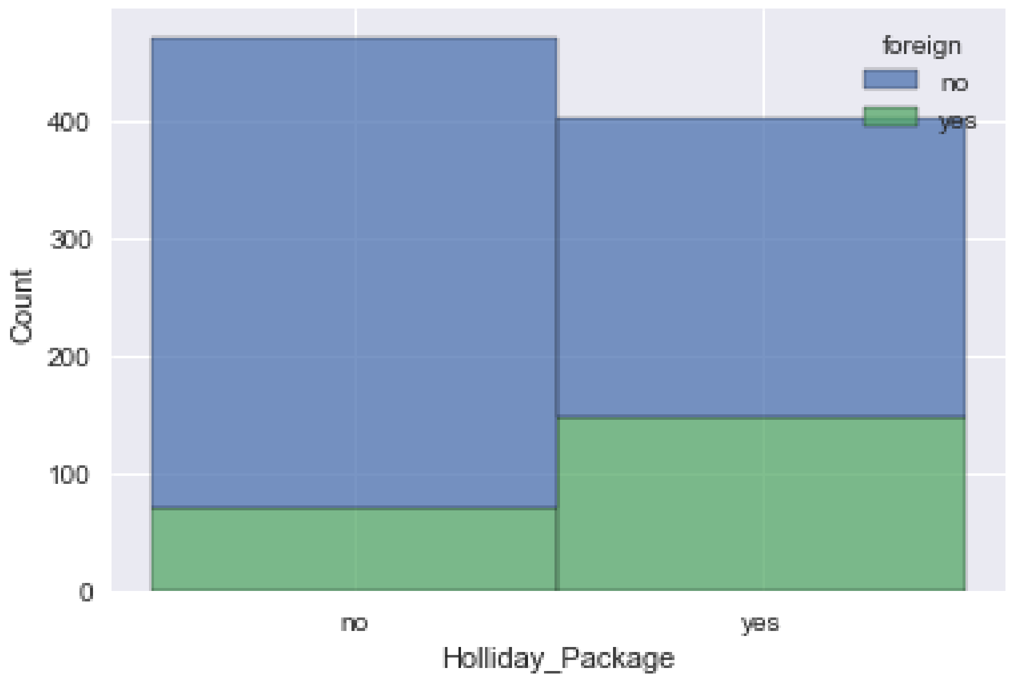

age
0
5
10
15
20
25
30
35
40
no
yes
Holliday_Package


educ
0
2
4
6
8
10
no
yes
Holliday_Package


1.2
1.0
0.8
0.6
0.4
0.2
0.0
no_older_children
no
yes
Holliday_Package


0.4
0.3
0.2
0.1
0.0
no_young_children
no
yes
Holliday_Package





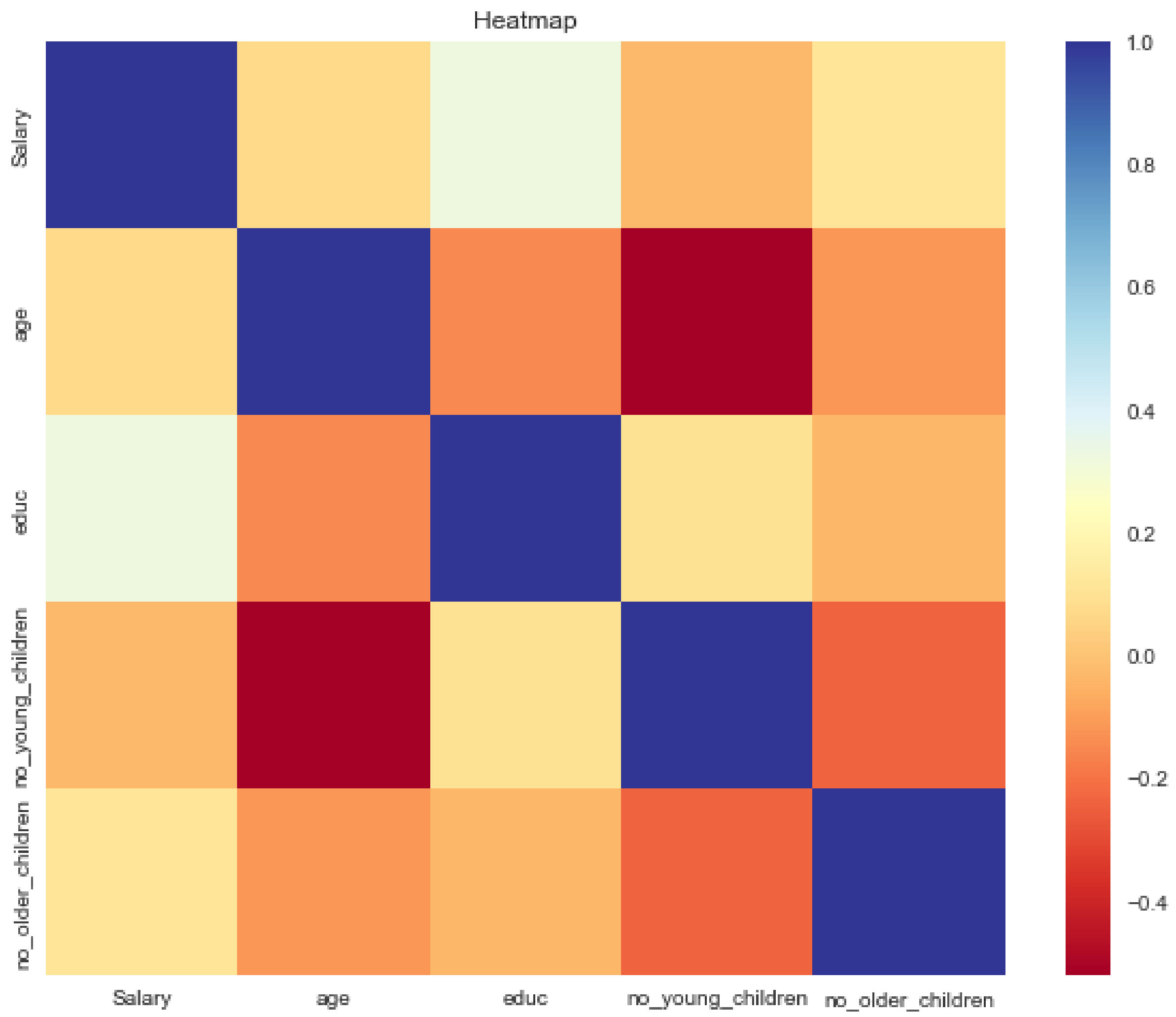
From the above bar plots it can be observed that the average age of people who have taken the package is less than that of those who have not taken.

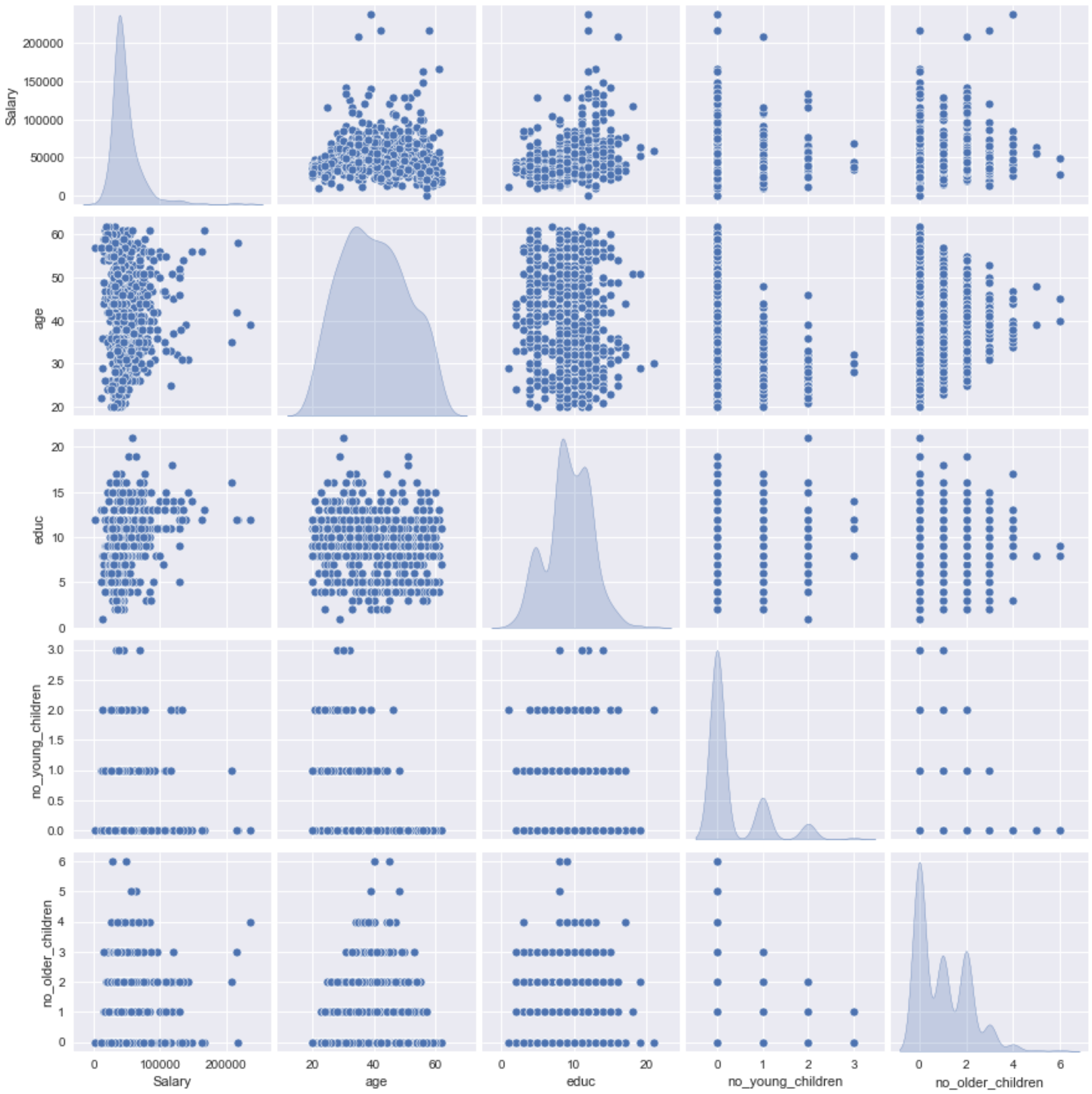
Average years of education is less for the people who have taken the package than those who have not taken.

Average no of older children is more for those who have taken the package and no of younger children are less.

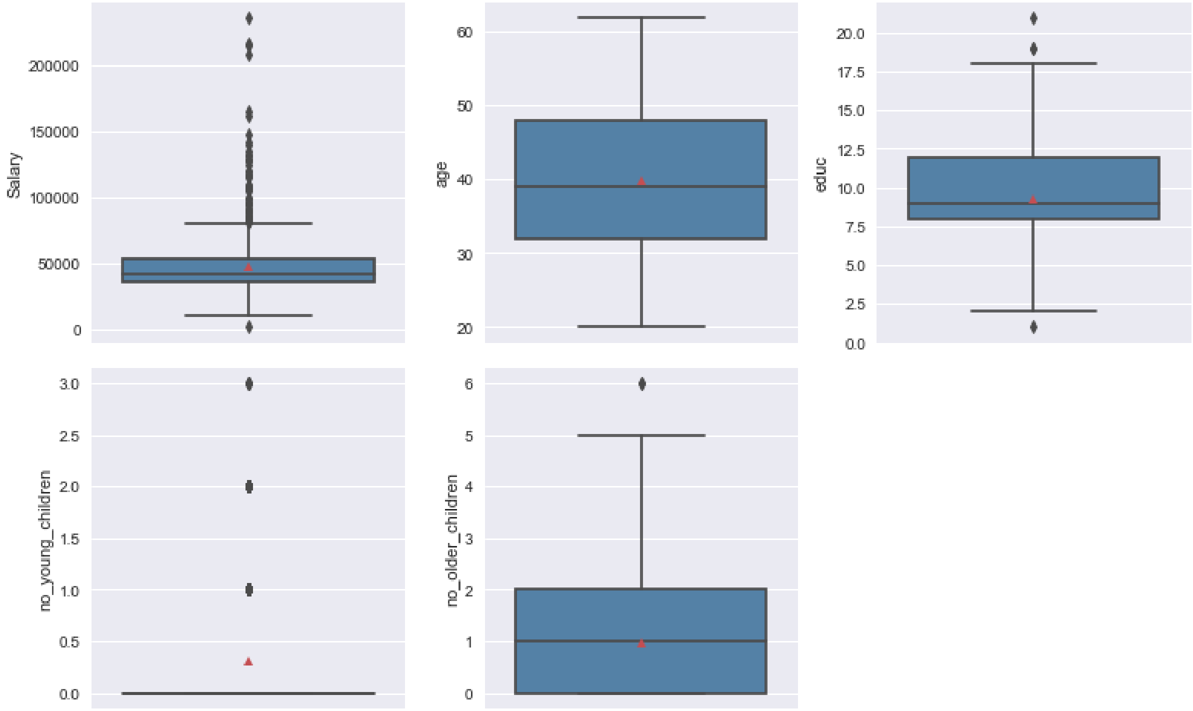
Average salary is less for those who have taken the package

Multivariate Analysis

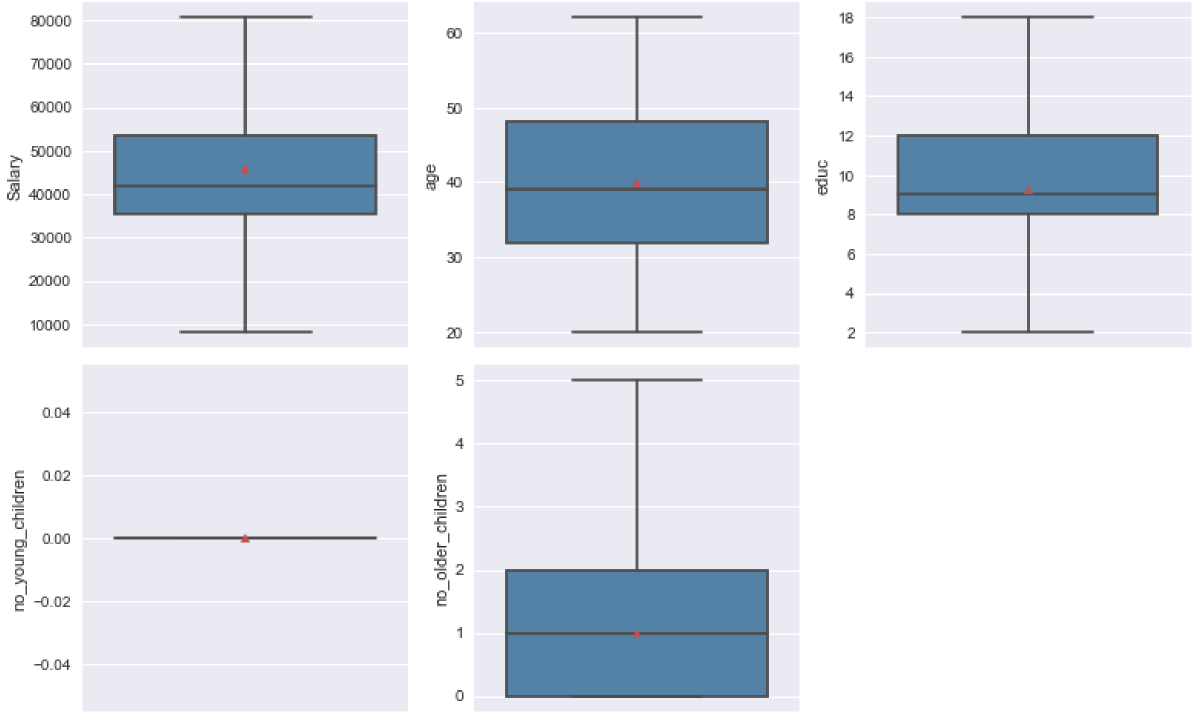

Heatmap
Salary
age
educ
no_young_children
no_older_children
Salary
age
educ
no_young_children
no_older_children
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4



From the above pair plot and correlation plot it can be observed that there is no high correlation between the variables.

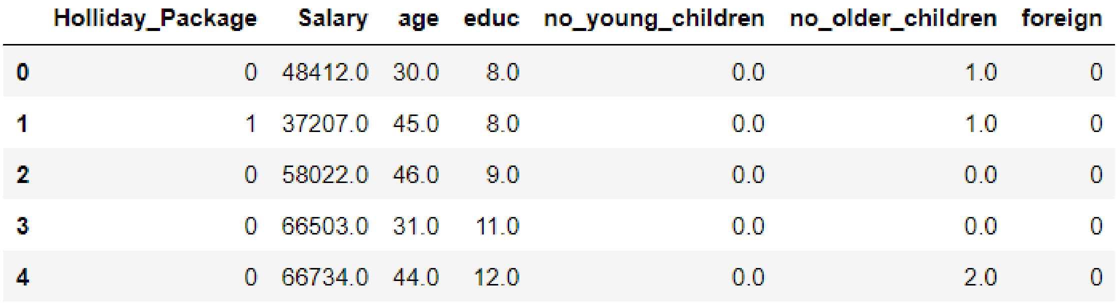


From the above box plot it can be observed that there are outliers which needs to be treated.



We have removed the outliers.

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).



	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	0	48412.0	30.0	8.0	0.0	1.0	0
1	1	37207.0	45.0	8.0	0.0	1.0	0
2	0	58022.0	46.0	9.0	0.0	0.0	0
3	0	66503.0	31.0	11.0	0.0	0.0	0
4	0	66734.0	44.0	12.0	0.0	2.0	0

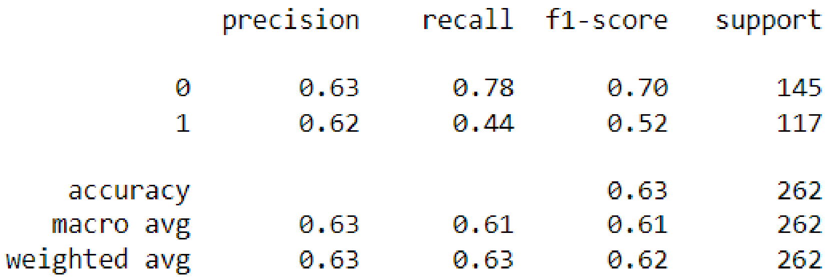
We have encoded the categorical variables. as 1 for yes and 0 for No

The data has been split into 70:30 ratios for train and test set.

Logistic regression and Linear discriminant model built and the data has been fitted into that model.

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.

LDA Classification Report

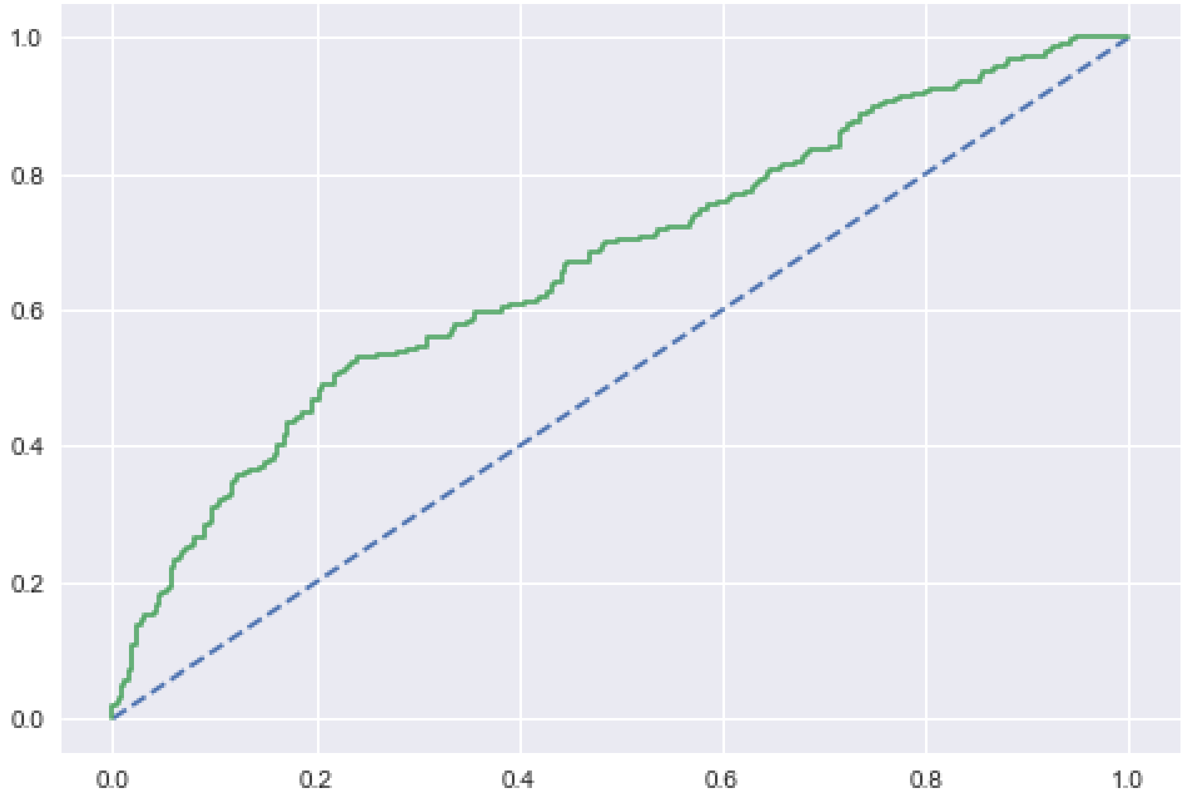


```
              precision    recall  f1-score   support

           0       0.63      0.78      0.70       145
           1       0.62      0.44      0.52       117

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

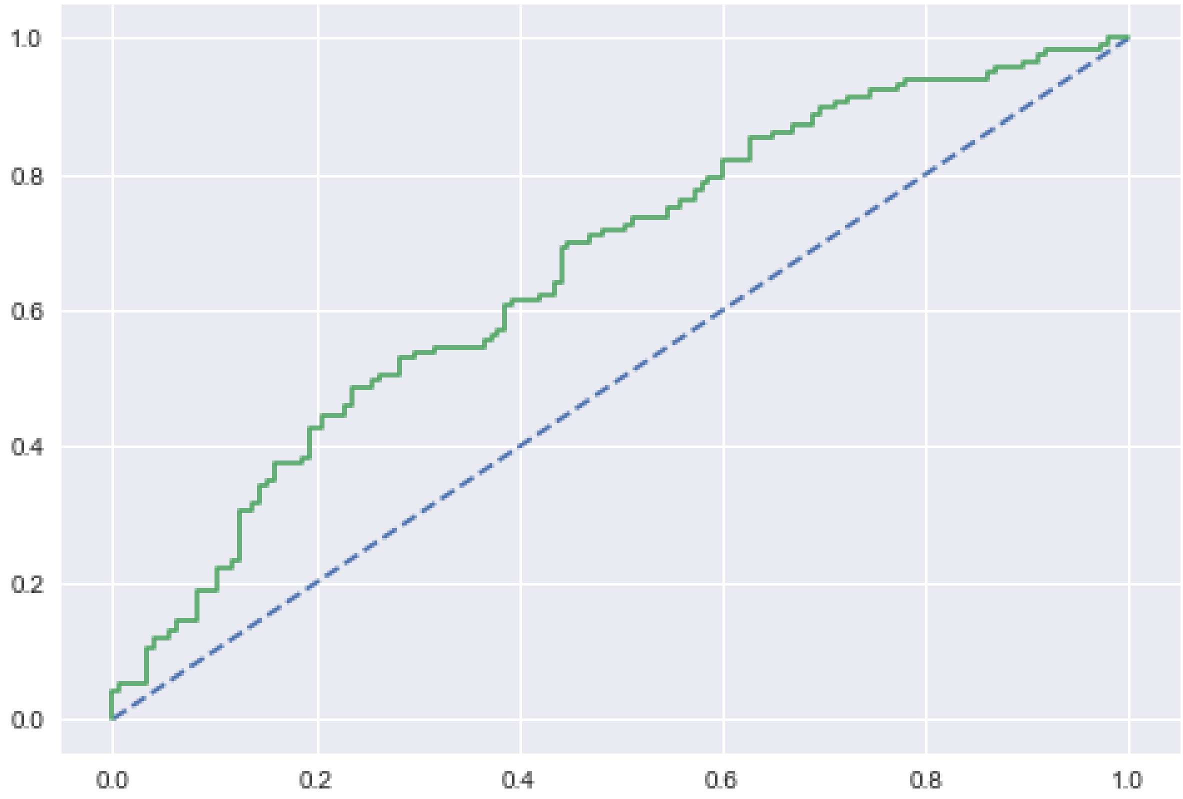
We are getting 0.63 as accuracy in LDA model with 0.44 recall score.

Logistics regression analysis performance metrics are as per below

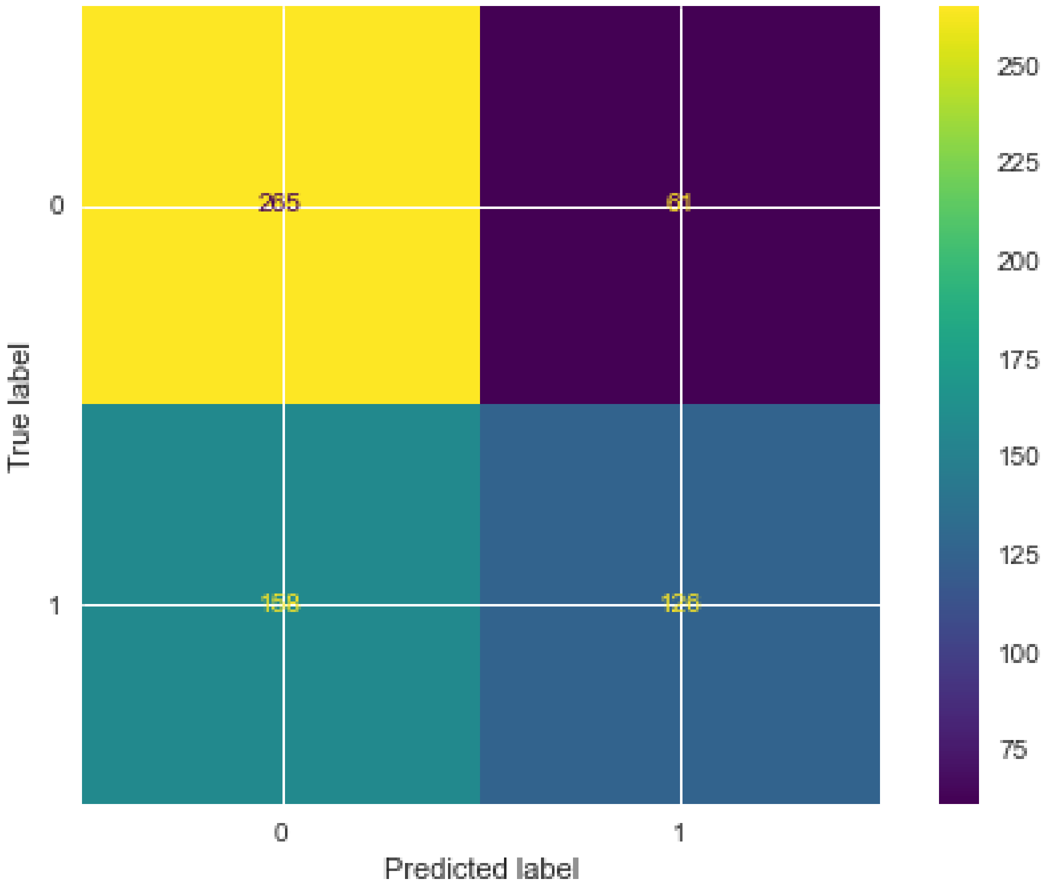
AUC & ROC curve of training data



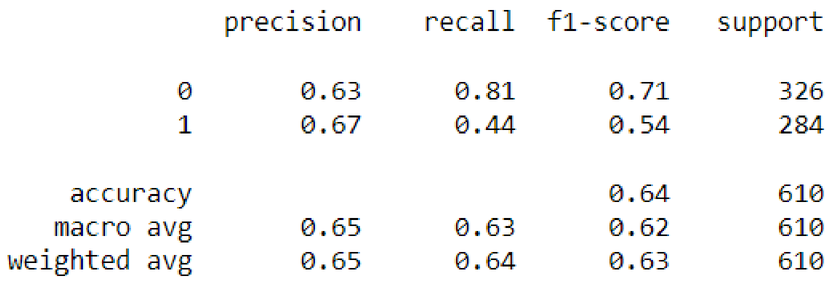
AUC & ROC curve of testing data



Confusion matrix of train data

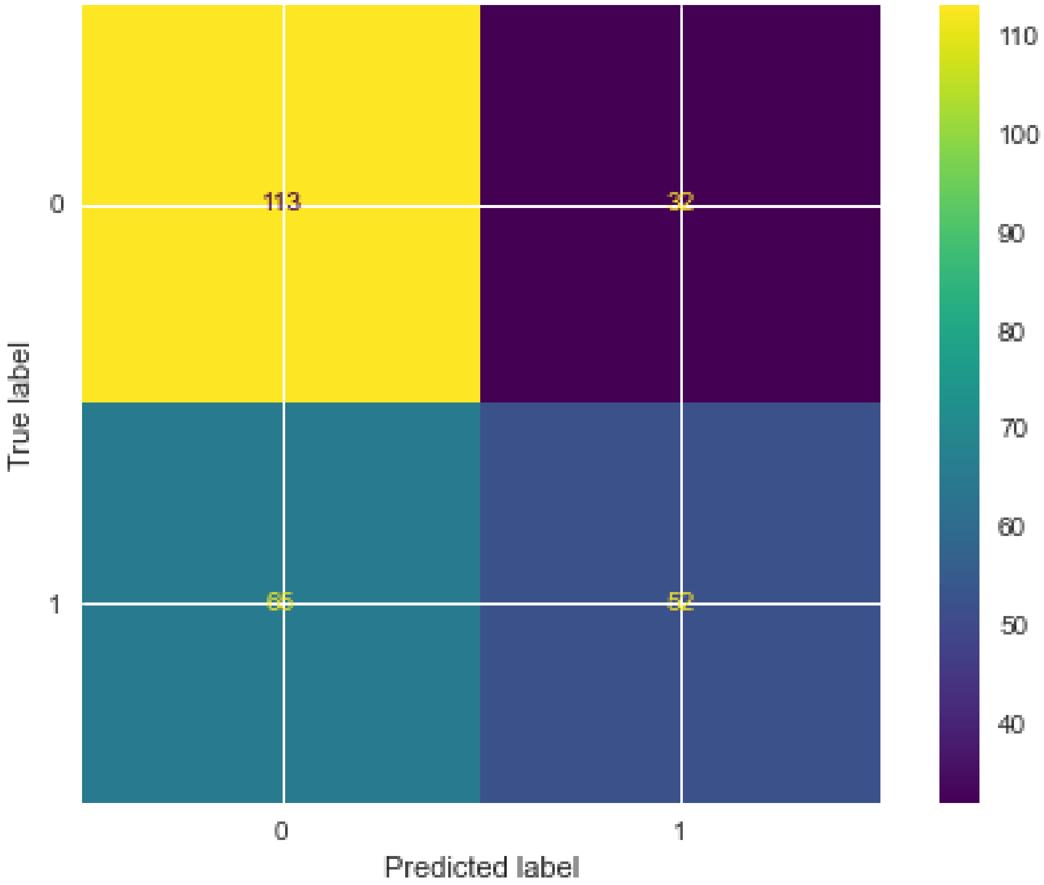


Classification report of train data



```
              precision    recall  f1-score   support

           0       0.63      0.81      0.71       326
           1       0.67      0.44      0.54       284

    accuracy                           0.64       610
   macro avg       0.65      0.63      0.62       610
weighted avg       0.65      0.64      0.63       610
```

Confusion Matrix for testing data

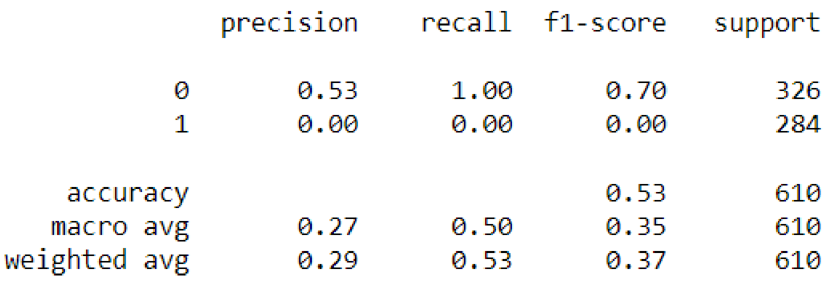


Classification report of testing data

```
              precision    recall  f1-score   support

           0       0.63      0.78      0.70       145
           1       0.62      0.44      0.52       117

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

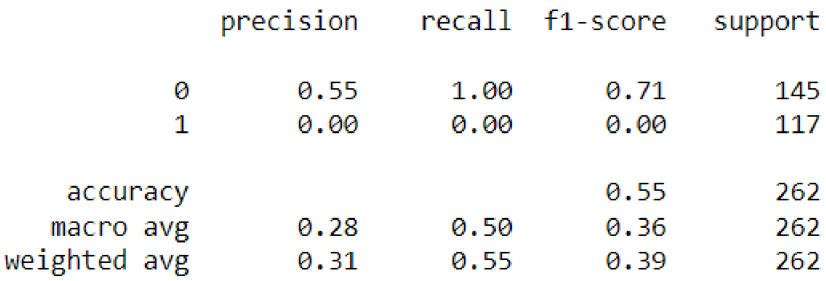
We have performed a Grid search cv also for optimizing the model and improving the accuracy.

Classification report on train data after grid search cv



```
              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
```

Classification report on test data after grid search cv



```
              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
```

But after grid search cv the accuracy is decreasing.

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

	LDA	LR train	LR test	LR after Grid Search CV train	LR after Grid Search CV test
Recall	0.44	0.44	0.44	0.00	0.00
Accuracy	0.63	0.64	0.63	0.53	0.55
Precision	0.62	0.67	0.62	0.00	0.00

From the above table we can observe that Logistics regression is the best model.

Company should focus on employees who are younger , have comparatively less salary , early married and having more children above 7 yrs of age to sell the packages.

These people are in the less salary zone which means they don’t handle much important job profile and hence have time to spend with family.