PROJECT

PREDICTIVE MODELING

Table of contents

List of Figures

List of Tables

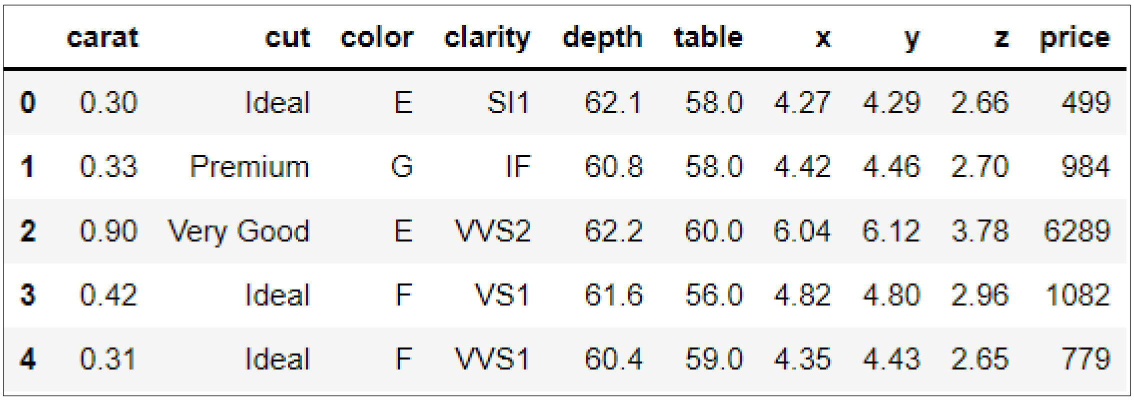
PROBLEM 1 - LINEAR REGRESSION

Problem Statement:

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

Q1.1. Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA, duplicate values). Perform Univariate and Bivariate Analysis.

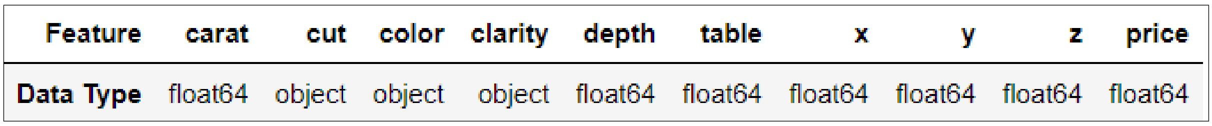
Sample of the Dataset:



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

Table 1. Sample of Cubic Zirconia Dataset.

Data Types of Variables in the Dataset:



Feature	carat	cut	color	clarity	depth	table	x	y	z	price
Data Type	float64	object	object	object	float64	float64	float64	float64	float64	float64

Table 2. Data Types of All Features in the Cubic Zirconia Dataset.

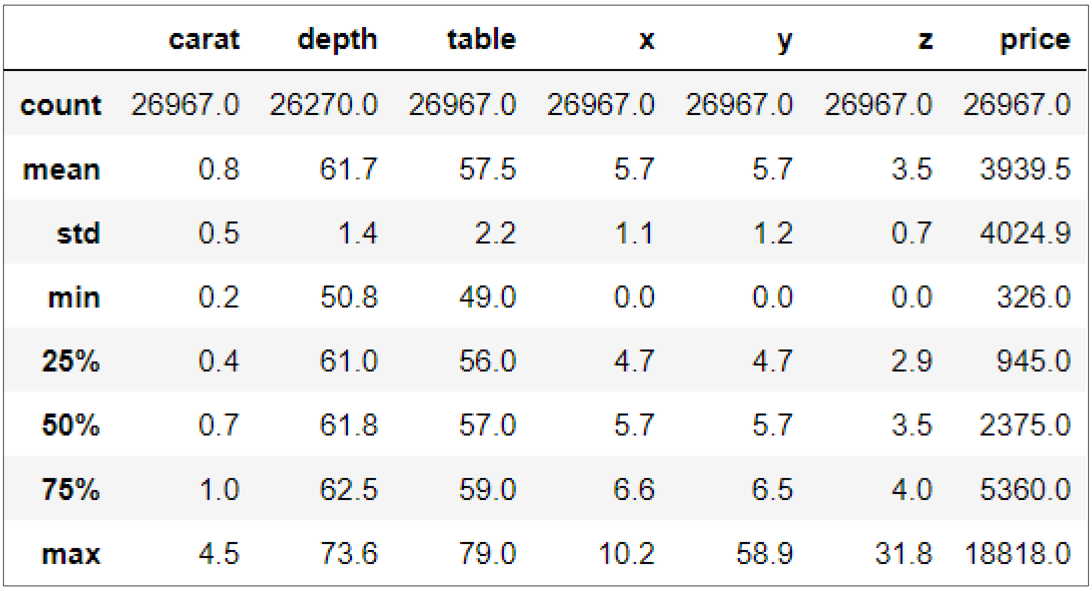
Insights:

1. There are 10 features (columns) with 26967 observations (rows) in the dataset.
2. The dataset has three categorical features i.e., Cut, colour and clarity and their data type are object.

3. The dataset has seven continuous numerical features i.e., carat, depth, table, x, y, z and price and their data type are float64.
4. The target variable in this dataset is price.

Description of the Dataset

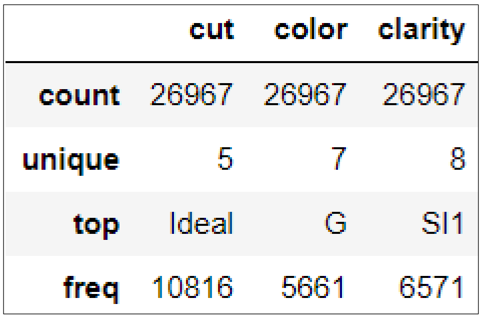
Numerical Features:



	carat	depth	table	x	y	z	price
count	26967.0	26270.0	26967.0	26967.0	26967.0	26967.0	26967.0
mean	0.8	61.7	57.5	5.7	5.7	3.5	3939.5
std	0.5	1.4	2.2	1.1	1.2	0.7	4024.9
min	0.2	50.8	49.0	0.0	0.0	0.0	326.0
25%	0.4	61.0	56.0	4.7	4.7	2.9	945.0
50%	0.7	61.8	57.0	5.7	5.7	3.5	2375.0
75%	1.0	62.5	59.0	6.6	6.5	4.0	5360.0
max	4.5	73.6	79.0	10.2	58.9	31.8	18818.0

Table 3. Description of Numerical Features in Cubic Zirconia Dataset.

Categorical Features:



	cut	color	clarity
count	26967	26967	26967
unique	5	7	8
top	Ideal	G	SI1
freq	10816	5661	6571

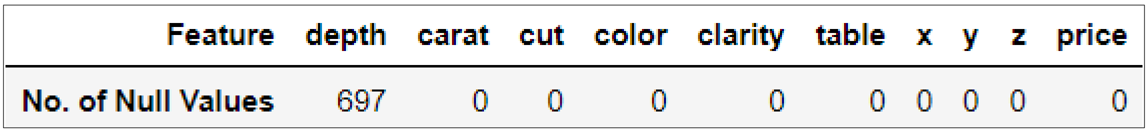
Table 4. Description of Categorical Features in Cubic Zirconia Dataset.

Exploratory Data Analysis

Checking and Dropping Duplicate Observations:

1. There are 34 duplicate observations in the given dataset.
2. The duplicate observations have been dropped from the dataset and the index has been reset.
3. The number of rows after dropping the duplicate observations is 26933.

Checking for Null values in the Dataset:



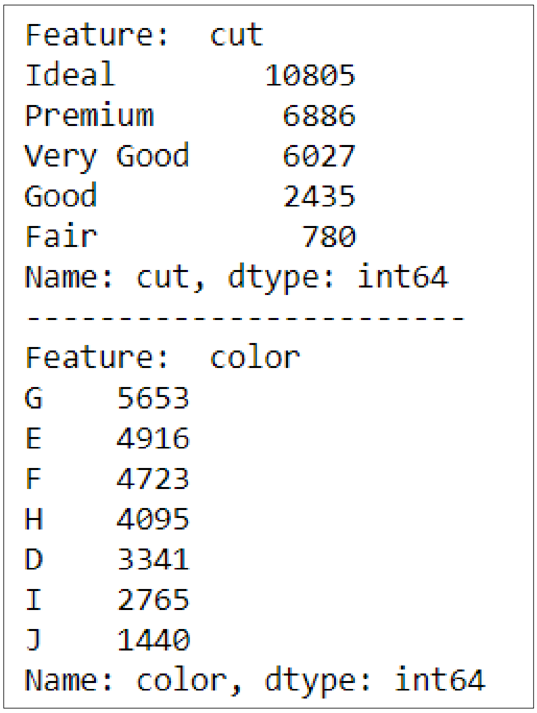
Feature	depth	carat	cut	color	clarity	table	x	y	z	price
No. of Null Values	697	0	0	0	0	0	0	0	0	0

Table 5. Null Values in the Cubic Zirconia Dataset.

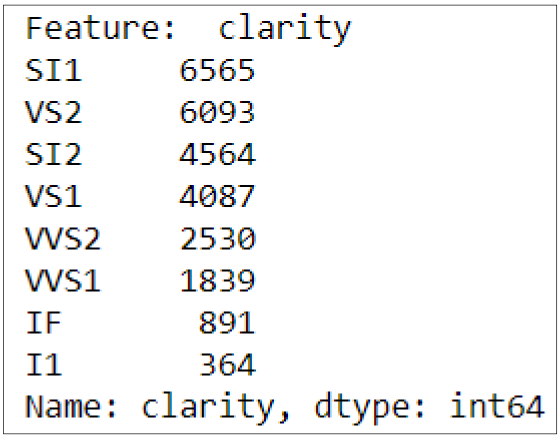
- Depth feature has 697 null values in it.

Checking for Anomalies

Checking Value Counts of Categorical Variables

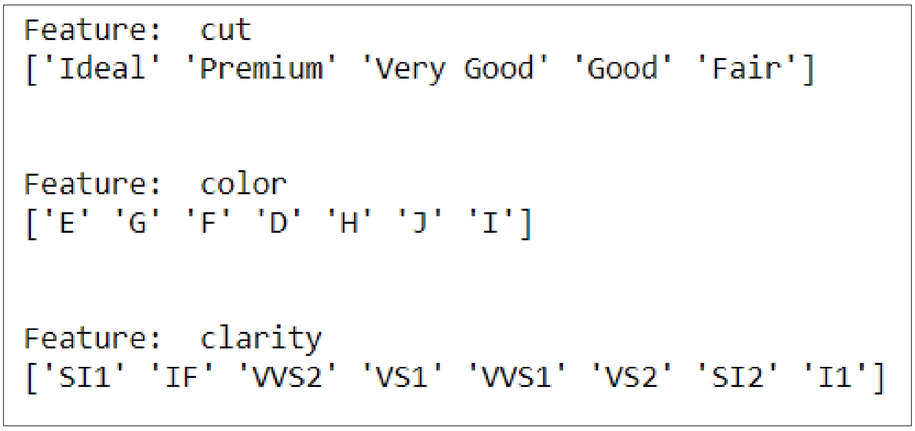


```
Feature:  cut
Ideal        10805
Premium       6886
Very Good     6027
Good          2435
Fair           780
Name: cut, dtype: int64
------------------------
Feature:  color
G    5653
E    4916
F    4723
H    4095
D    3341
I    2765
J    1440
Name: color, dtype: int64
```



```
Feature:  clarity
SI1     6565
VS2     6093
SI2     4564
VS1     4087
VVS2    2530
VVS1    1839
IF       891
I1       364
Name: clarity, dtype: int64
```

Checking Unique Entries of Categorical Variables

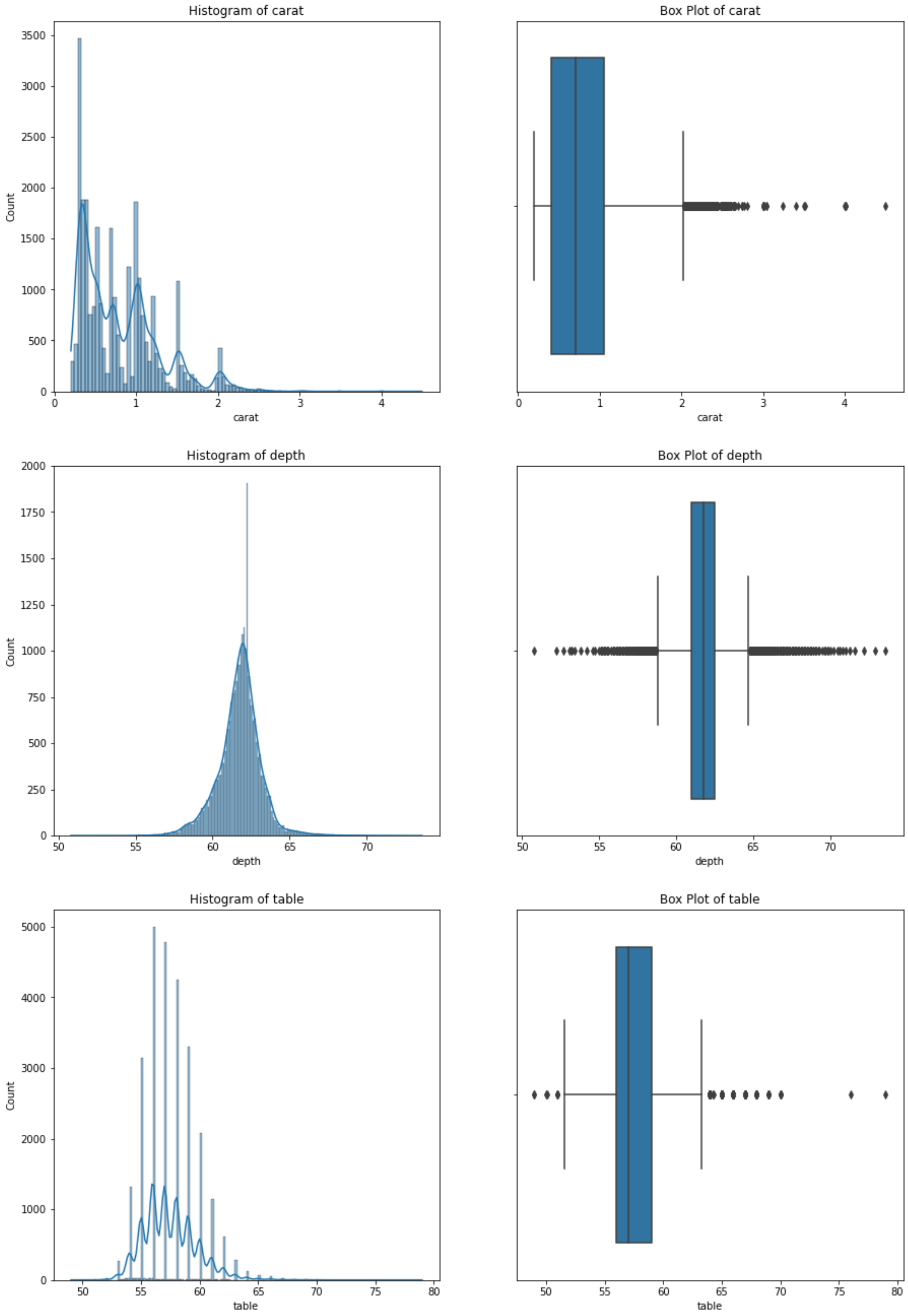


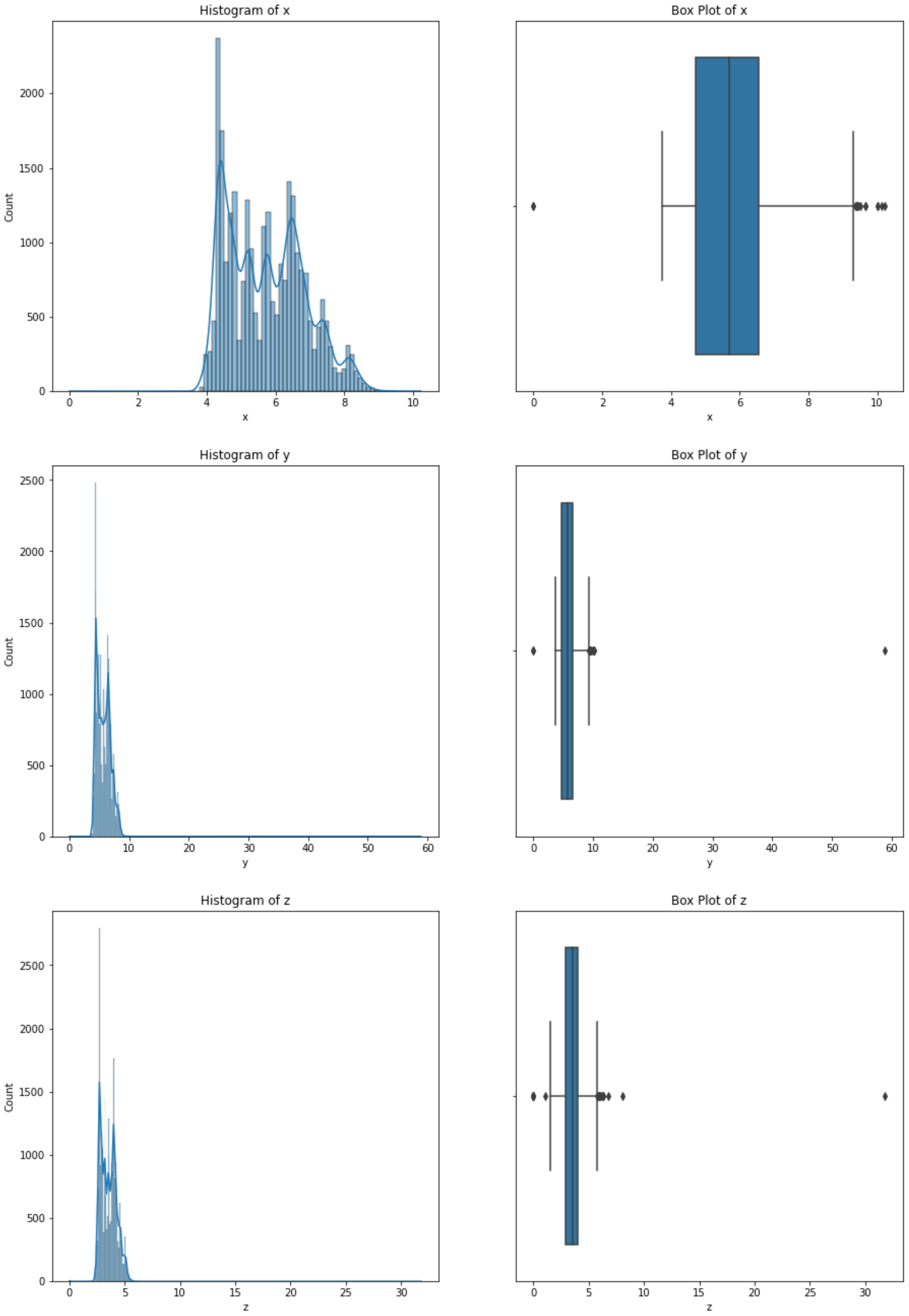
```
Feature:  cut
['Ideal' 'Premium' 'Very Good' 'Good' 'Fair']


Feature:  color
['E' 'G' 'F' 'D' 'H' 'J' 'I']


Feature:  clarity
['SI1' 'IF' 'VVS2' 'VS1' 'VVS1' 'VS2' 'SI2' 'I1']
```

Insights:

1. There are no anomalies in the sublevels of categorical features.
2. There are no sublevels in all categorical features with negligible count.

Univariate Analysis – Distribution Plots:




Histogram of x
Count
2000
1500
1000
500
0
0
2
4
6
8
10
x
Box Plot of x
0
2
4
6
8
10
x
Histogram of y
Count
2500
2000
1500
1000
500
0
0
10
20
30
40
50
60
y
Box Plot of y
0
10
20
30
40
50
60
y
Histogram of z
Count
2500
2000
1500
1000
500
0
0
5
10
15
20
25
30
z
Box Plot of z
0
5
10
15
20
25
30
z

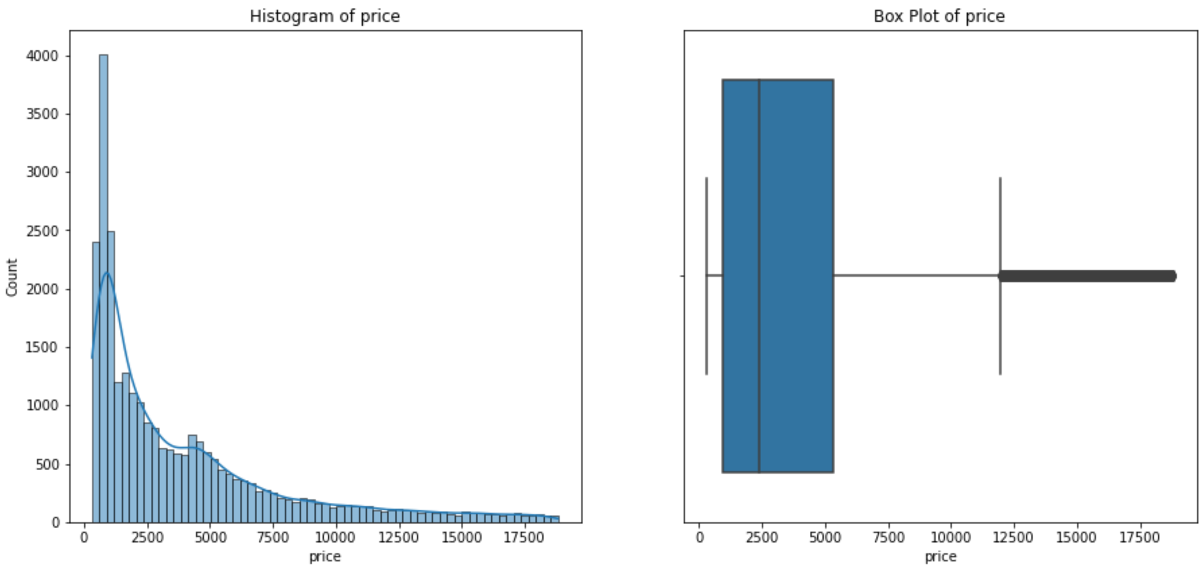
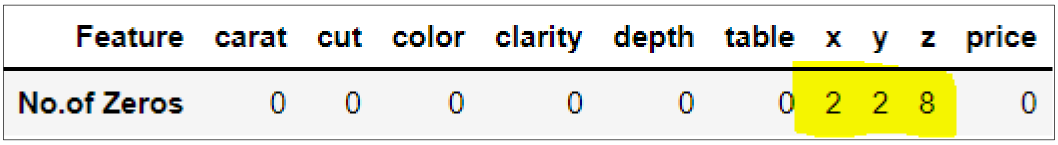


Figure 1. Histograms and Box Plots for Numerical Features in Cubic Zirconia Dataset.

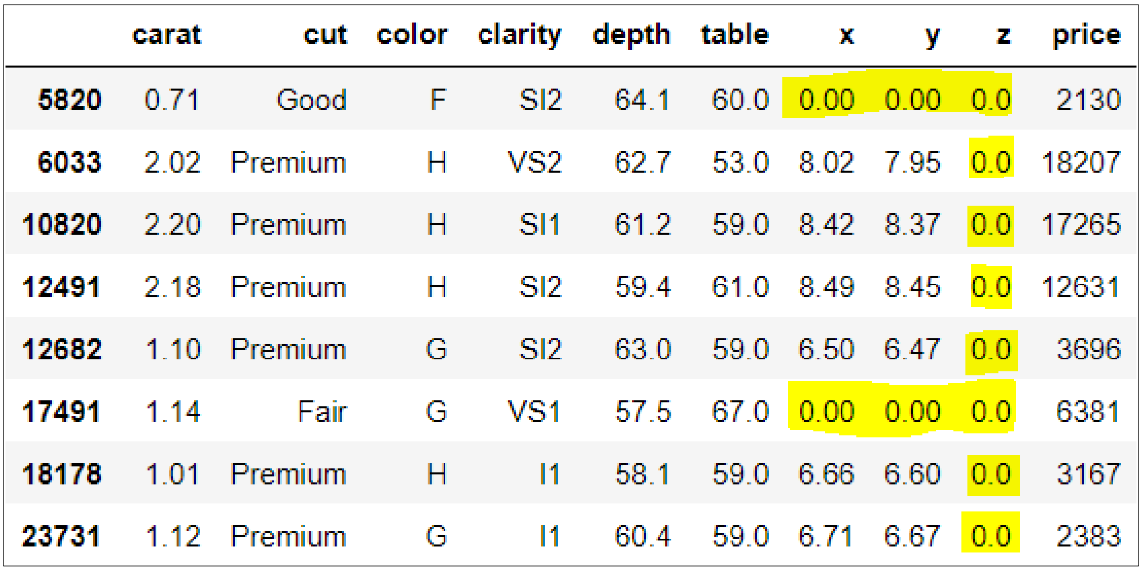
Inferences:

1. There are zero values in **x, y and z features** but these features can not be equal to zero because **they are representing dimensions of the cubic zirconia**.



Feature	carat	cut	color	clarity	depth	table	x	y	z	price
No.of Zeros	0	0	0	0	0	0	2	2	8	0

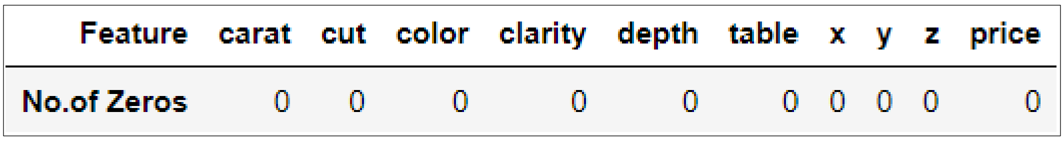
Table 6. Number of Zeros in Each Feature of Cubic Zirconia Dataset.



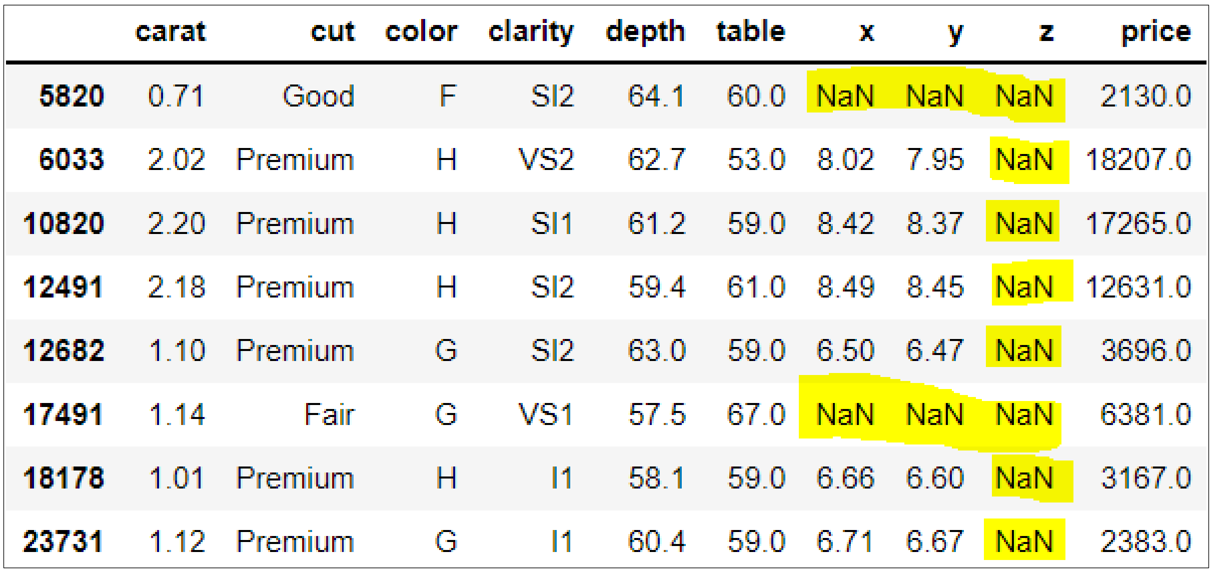
	carat	cut	color	clarity	depth	table	x	y	z	price
5820	0.71	Good	F	SI2	64.1	60.0	0.00	0.00	0.0	2130
6033	2.02	Premium	H	VS2	62.7	53.0	8.02	7.95	0.0	18207
10820	2.20	Premium	H	SI1	61.2	59.0	8.42	8.37	0.0	17265
12491	2.18	Premium	H	SI2	59.4	61.0	8.49	8.45	0.0	12631
12682	1.10	Premium	G	SI2	63.0	59.0	6.50	6.47	0.0	3696
17491	1.14	Fair	G	VS1	57.5	67.0	0.00	0.00	0.0	6381
18178	1.01	Premium	H	I1	58.1	59.0	6.66	6.60	0.0	3167
23731	1.12	Premium	G	I1	60.4	59.0	6.71	6.67	0.0	2383

Table 7. Observations having Zeros in x, y and z Features of Cubic Zirconia Dataset.

2. As the number of zeros are very less, zeros can be dropped from the dataset. But in this dataset, **zeros are imputed with null values** and later these null values in each feature are treated with respective median values.

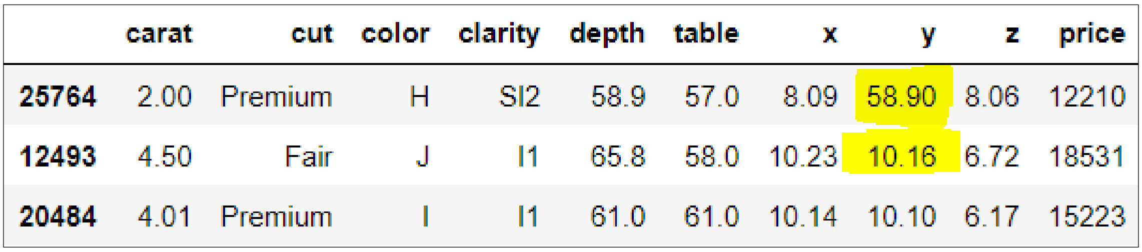


Feature	carat	cut	color	clarity	depth	table	x	y	z	price
No.of Zeros	0	0	0	0	0	0	0	0	0	0



	carat	cut	color	clarity	depth	table	x	y	z	price
5820	0.71	Good	F	SI2	64.1	60.0	NaN	NaN	NaN	2130.0
6033	2.02	Premium	H	VS2	62.7	53.0	8.02	7.95	NaN	18207.0
10820	2.20	Premium	H	SI1	61.2	59.0	8.42	8.37	NaN	17265.0
12491	2.18	Premium	H	SI2	59.4	61.0	8.49	8.45	NaN	12631.0
12682	1.10	Premium	G	SI2	63.0	59.0	6.50	6.47	NaN	3696.0
17491	1.14	Fair	G	VS1	57.5	67.0	NaN	NaN	NaN	6381.0
18178	1.01	Premium	H	I1	58.1	59.0	6.66	6.60	NaN	3167.0
23731	1.12	Premium	G	I1	60.4	59.0	6.71	6.67	NaN	2383.0

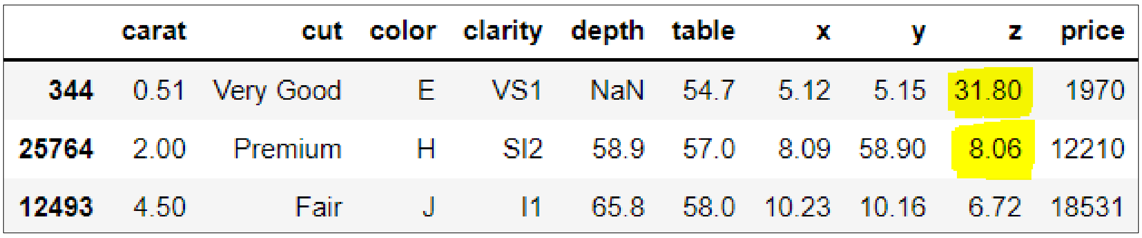
3. **y** feature has an entry equal to 58.9 which is far away from the whisker. It can be noticed in box plot of y feature. It can be a data entry mistake and can be replaced with nearest value i.e., 10.16.



	carat	cut	color	clarity	depth	table	x	y	z	price
25764	2.00	Premium	H	SI2	58.9	57.0	8.09	58.90	8.06	12210
12493	4.50	Fair	J	I1	65.8	58.0	10.23	10.16	6.72	18531
20484	4.01	Premium	I	I1	61.0	61.0	10.14	10.10	6.17	15223

Table 8. Observations having Anomalies in y Feature of Cubic Zirconia Dataset.

4. Similarly, **z** feature has an entry equal to 31.8 which is far away from the whisker. It can be noticed in box plot of z feature. It can be a data entry mistake and can be replaced with nearest value i.e., 8.06.



	carat	cut	color	clarity	depth	table	x	y	z	price
344	0.51	Very Good	E	VS1	NaN	54.7	5.12	5.15	31.80	1970
25764	2.00	Premium	H	SI2	58.9	57.0	8.09	58.90	8.06	12210
12493	4.50	Fair	J	I1	65.8	58.0	10.23	10.16	6.72	18531

Table 9. Observations having Anomalies in z Feature of Cubic Zirconia Dataset.

After treating above anomalies, box plots are plotted again for x, y and z features to confirm that anomalies treated properly and to verify the modified distribution variables.

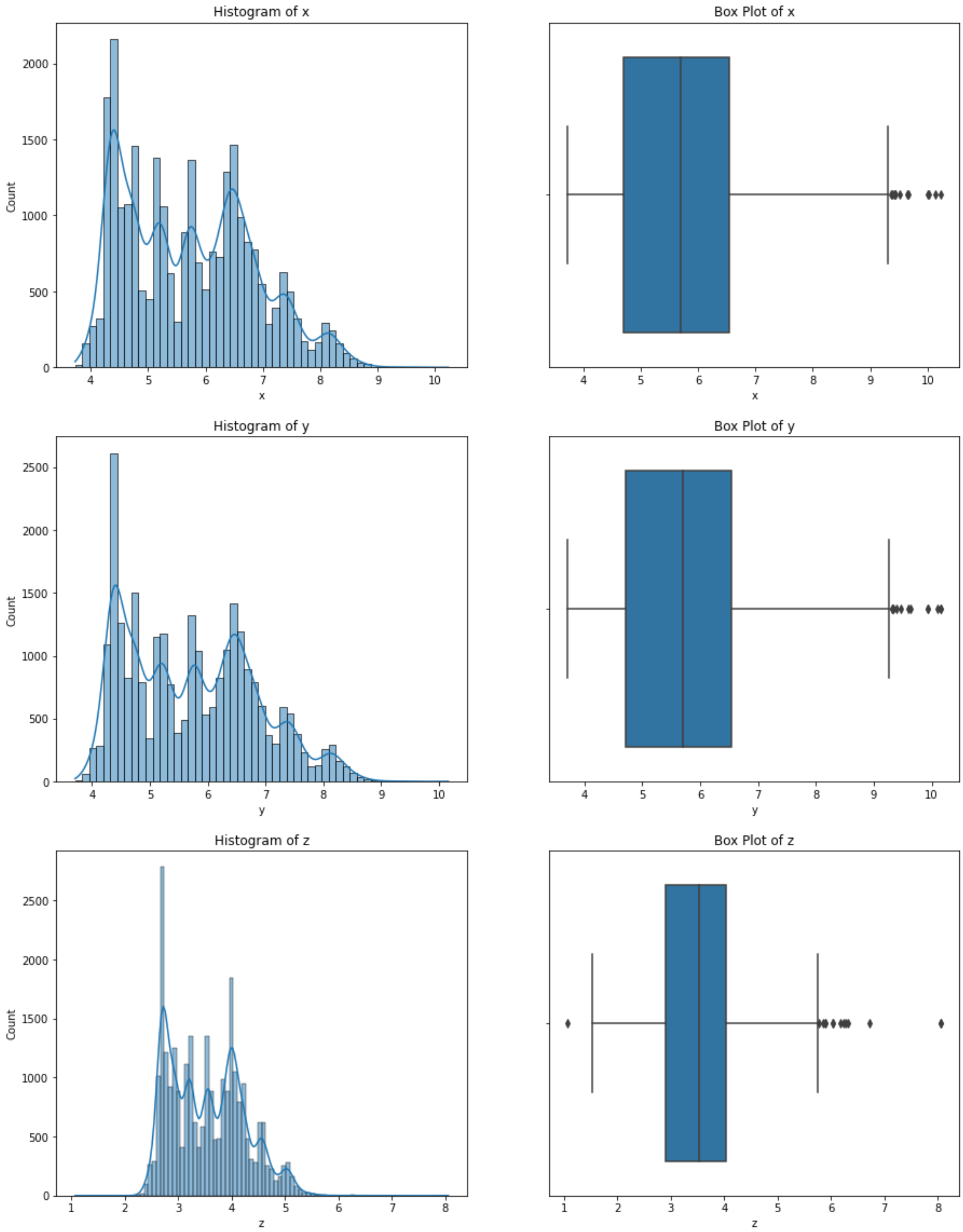
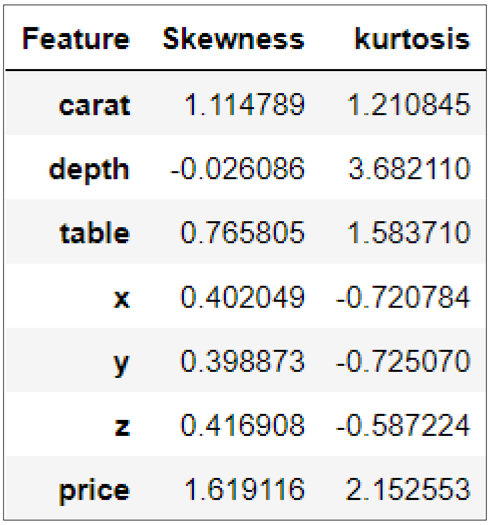


Figure 2. Histograms and Box Plots of x, y and z features after treating the anomalies.

Skewness & Kurtosis:

- Skewness is a measure of lack of symmetry in a distribution.

- Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.



Feature	Skewness	kurtosis
carat	1.114789	1.210845
depth	-0.026086	3.682110
table	0.765805	1.583710
x	0.402049	-0.720784
y	0.398873	-0.725070
z	0.416908	-0.587224
price	1.619116	2.152553

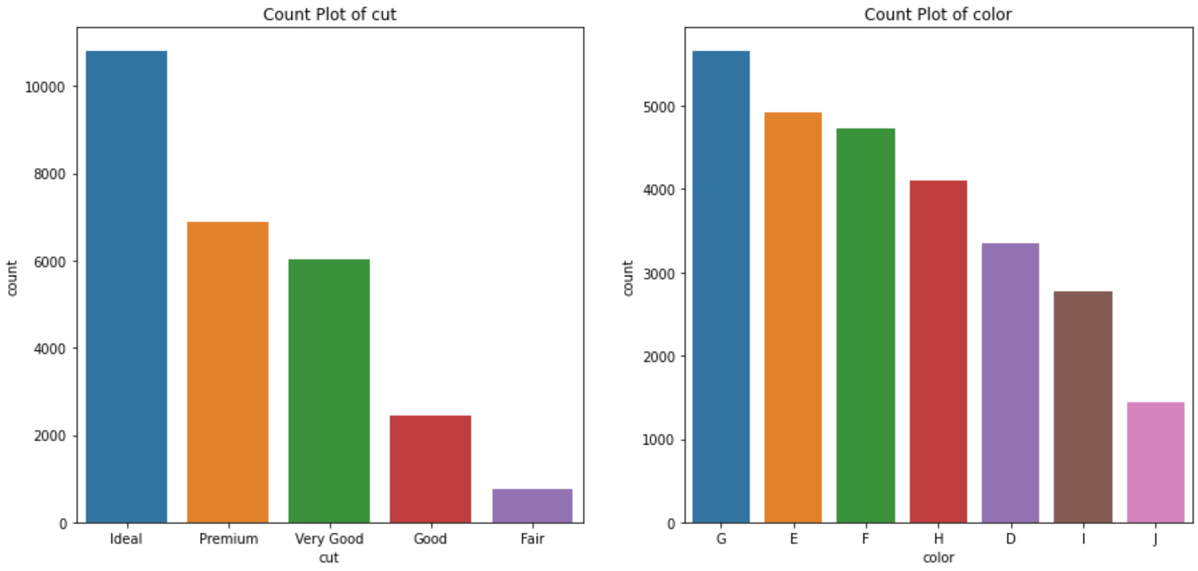
Table 10. Skewness and Kurtosis of Numeric Features in Cubic Zirconia Dataset.

Insights:

From above plots and tables, we can conclude below points,

1. Except depth feature, all other features are right skewed distributions (Positively skewed).
2. Except x, y and z, all remaining features have positive kurtosis.
3. x, y and z features have negative kurtosis.

Count Plot of Categorical Features:



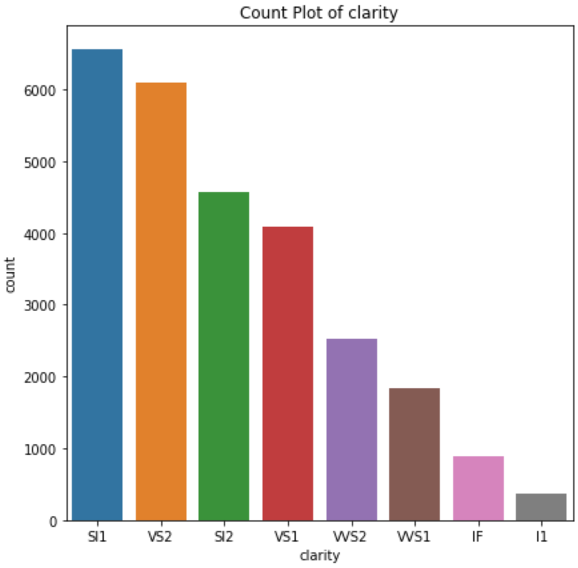


Figure 3. Count Plot of Categorical Features

Insights:

From above count plots, we can conclude below points.

1. Maximum number of cubic zirconia have Ideal quality and minimum number of cubic zirconia have Fair quality. Decreasing order of count as below.
   Ideal > Premium > Very Good > Good > Fair
2. Maximum number of cubic zirconia have colour "G" and minimum number of cubic zirconia have colour "J". Decreasing order of count as below.
   G > E > F > H > D> I > J >
3. Maximum number of cubic zirconia have clarity "SI1" and minimum number of cubic zirconia have clarity "I1". Decreasing order of count as below.
   SI1 > VS2 > SI2 > VS1> VVS2> VVS1 > IF>I1

Bivariate Analysis between Numerical Features

Pair Plot of Numerical Features:

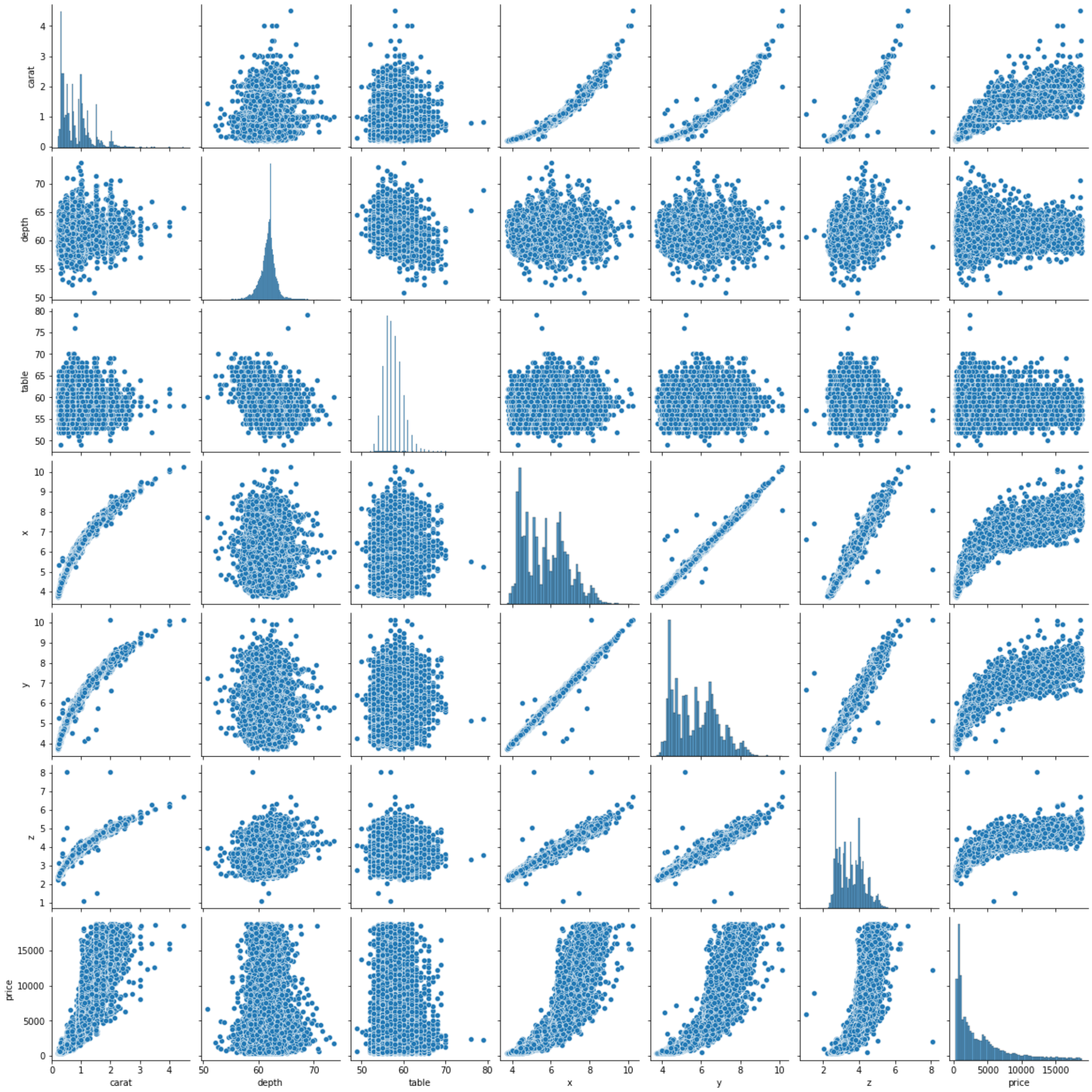


Figure 4. Pair Plot for Numeric Features in Cubic Zirconia Dataset.

Heat Map:

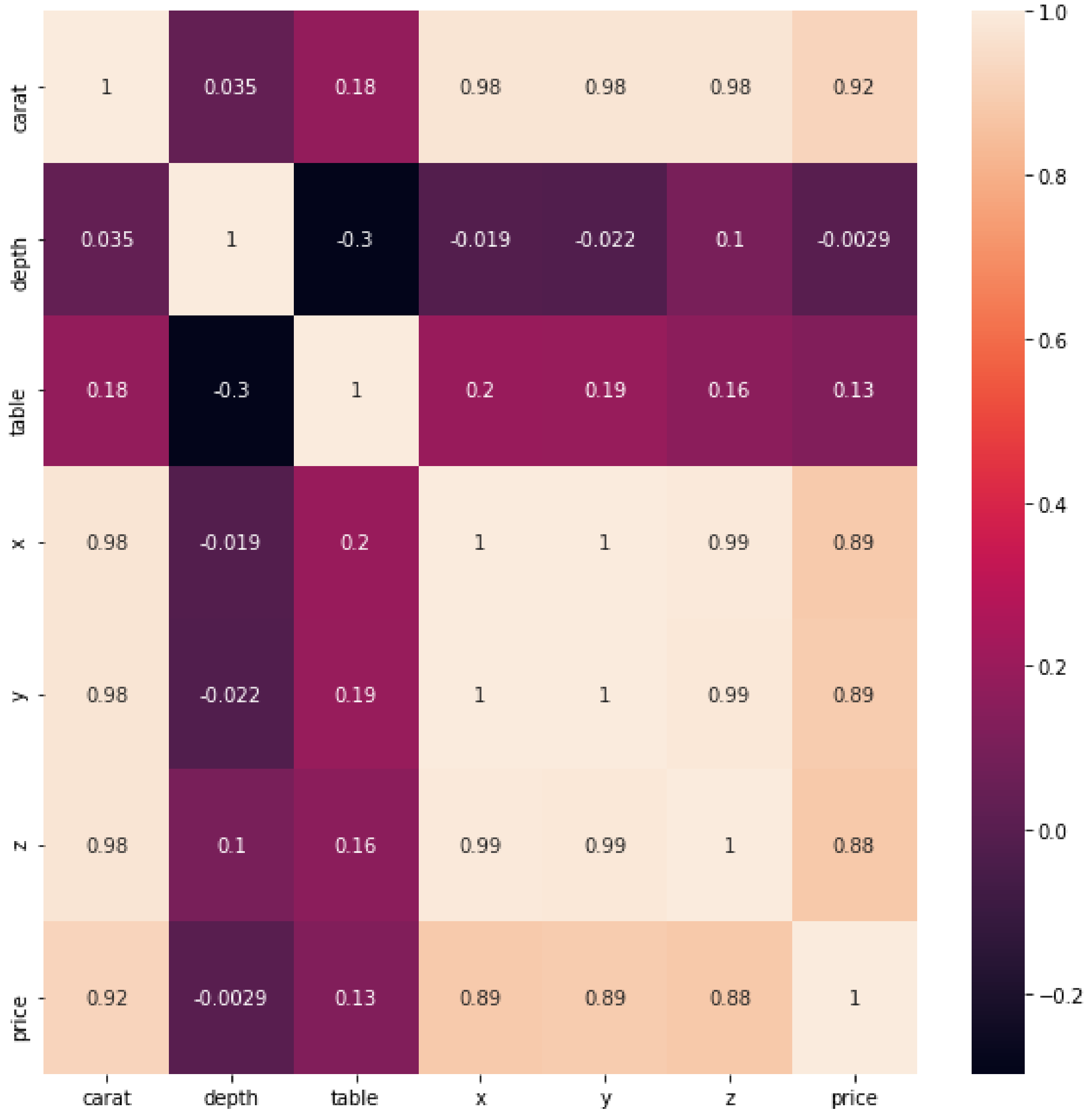


Figure 5. Heatmap for Numeric Features in Cubic Zirconia Dataset.

Insights:

From above Pair-Plot and Heatmap, we can write below inferences,

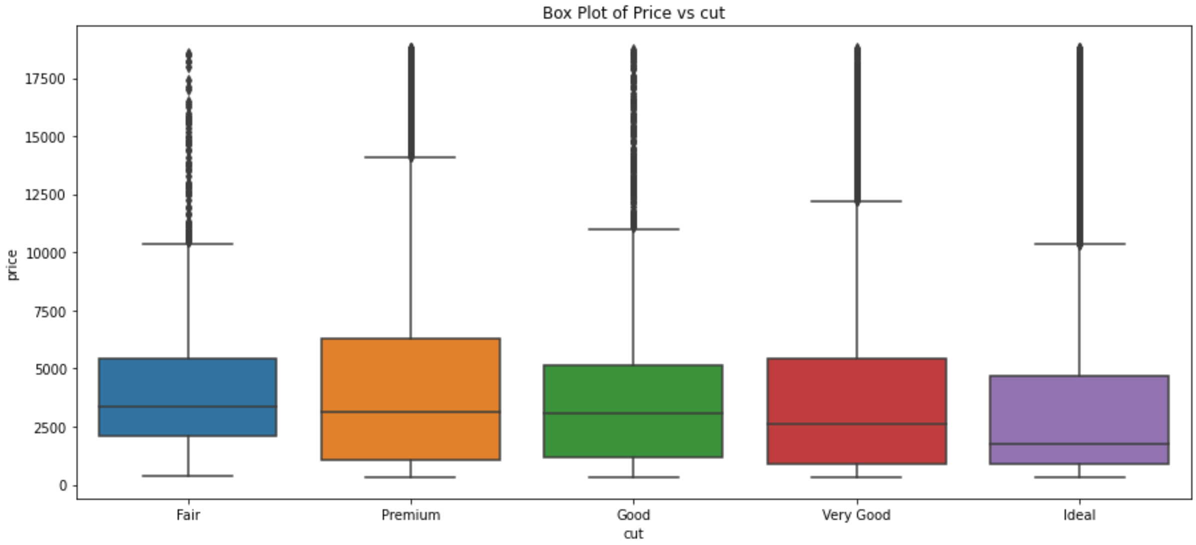
1. Few predictors **have strong correlation between them** like x & y (1), x & z (0.99), y & z (0.99), carat & x (0.98), carat & y (0.98) and carat & z (0.98). It leads to multicollinearity problem. The model performance can be checked by dropping few variables from them to reduce multicollinearity.
2. Few predictors **have weak correlation between them** like depth & carat (0.035), depth & x (-0.019) and depth & y (-0.022).

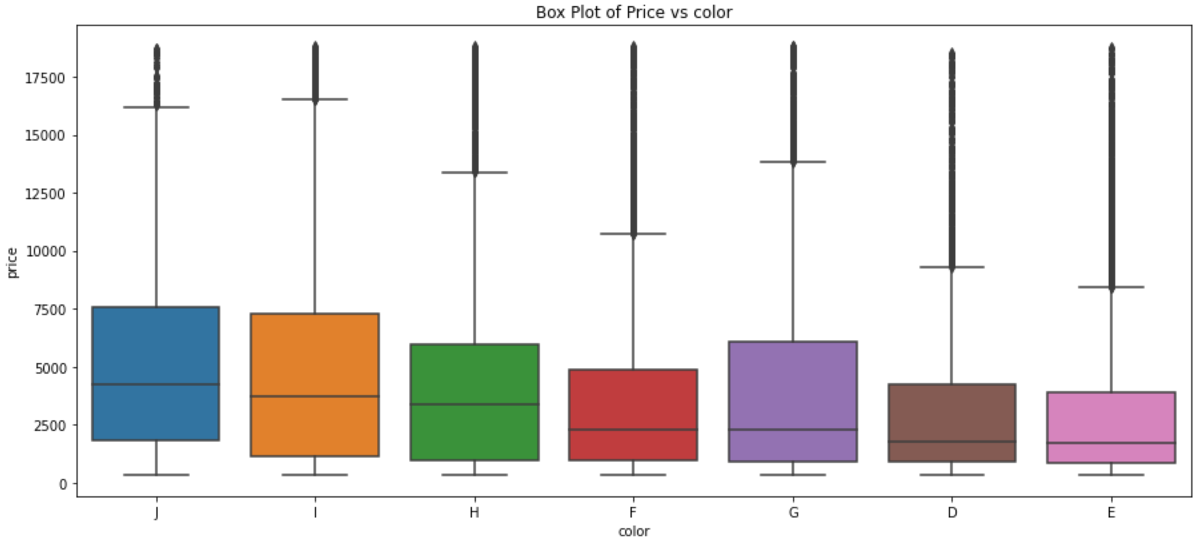
3. Few predictors like carat (0.92), x (0.89), y (0.89) and z (0.88) have **strong correlation with target variable** (price). Hence, these features may be considered as strong predictors after verifying the coefficients (slopes) obtained in the model.
4. Few predictors like depth (-0.0029) and table (0.13) **have weak correlation with target variable (price)**. Hence, these features may be considered as weak predictors after verifying the coefficients (slopes) obtained in the model.

Bivariate Analysis between Target Variable and Categorical Variables

Box Plots of Price vs Categorical Features:

Box plots are plotted below to do bivariate analysis between target variable (price) and categorical variables.





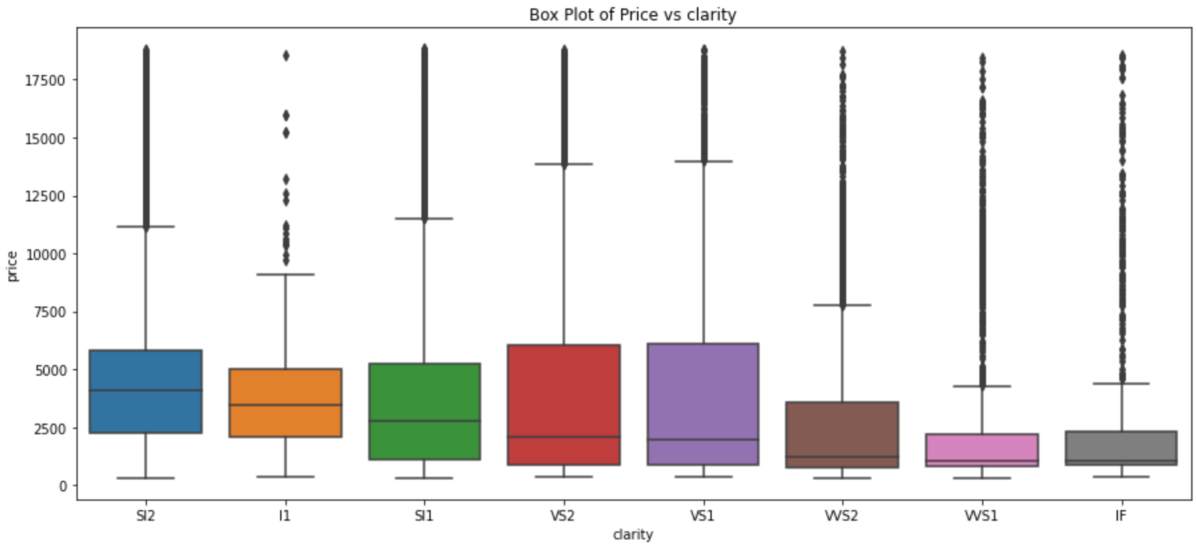


Figure 6. Box Plots of Price vs Categorical Features in Cubic Zirconia Dataset.

Insights:

From above bar plots, we can write below inferences,

1. Median price of Cubic Zirconia with fair quality is maximum and median price of Cubic Zirconia with Ideal quality is minimum. Decreasing order of median price of cubic zirconia with different cut quality is given below.

   Fair > Premium > Good > Very Good > Ideal

   It can be noticed that as cut quality of cubic zirconia increases, median price decreases except for premium cut quality. Usually, it is not expected.

2. Median price of Cubic Zirconia with ‘J’ colour is maximum and median price of Cubic Zirconia with ‘E’ colour is minimum. Decreasing order of median price of cubic zirconia with different colour is given below.

   J > I > H > F > G > D > E

   It can be noticed that as colour quality of cubic zirconia increases, median price also increases.

3. Median price of Cubic Zirconia with SI2 clarity is maximum and median price of Cubic Zirconia with IF clarity is minimum. Decreasing order of median price of cubic zirconia with different clarity is given below.

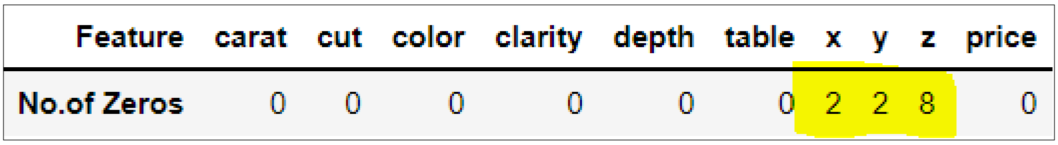
   SI2 > I1 > SI1 > VS2 > VS1 > VVS2 > VVS1 > IF

   It can be noticed that as clarity of cubic zirconia increases, median price also increases except for I1 clarity.

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

Checking for Zeros in the Dataset:

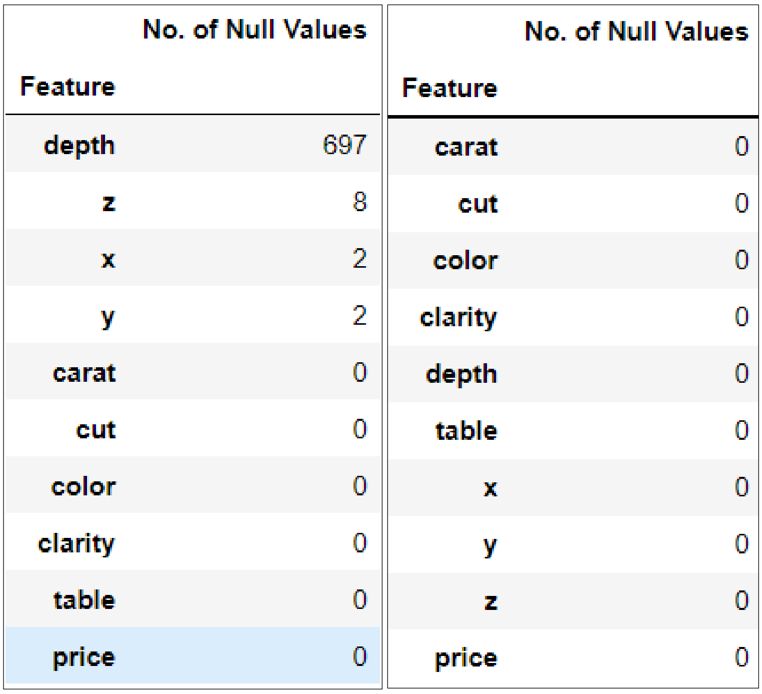
There are zero values in **x, y and z features** (Refer table 7 & 8) but these features should not be equal to zero because **they are representing dimensions of the cubic zirconia**. **These features do not have any meaning if they are equal to zero.** As the number of zeros are very less, zeros **can be dropped** from the dataset. But in this dataset, **zeros have been imputed with null values** in the EDA section and these null values in each feature will be imputed with respective median values in this section.



Feature	carat	cut	color	clarity	depth	table	x	y	z	price
No.of Zeros	0	0	0	0	0	0	2	2	8	0

Imputing Null Values:

There are null values in depth, x, y and z features. They are continuous numerical features and they have outliers in it. Hence, null Values have been imputed with respective median values.



Feature	No. of Null Values
depth	697
z	8
x	2
y	2
carat	0
cut	0
color	0
clarity	0
table	0
price	0

Feature	No. of Null Values
carat	0
cut	0
color	0
clarity	0
depth	0
table	0
x	0
y	0
z	0
price	0

Table 11. Number of Null Values Before and After Treatment.

Checking for Outliers in the Dataset:

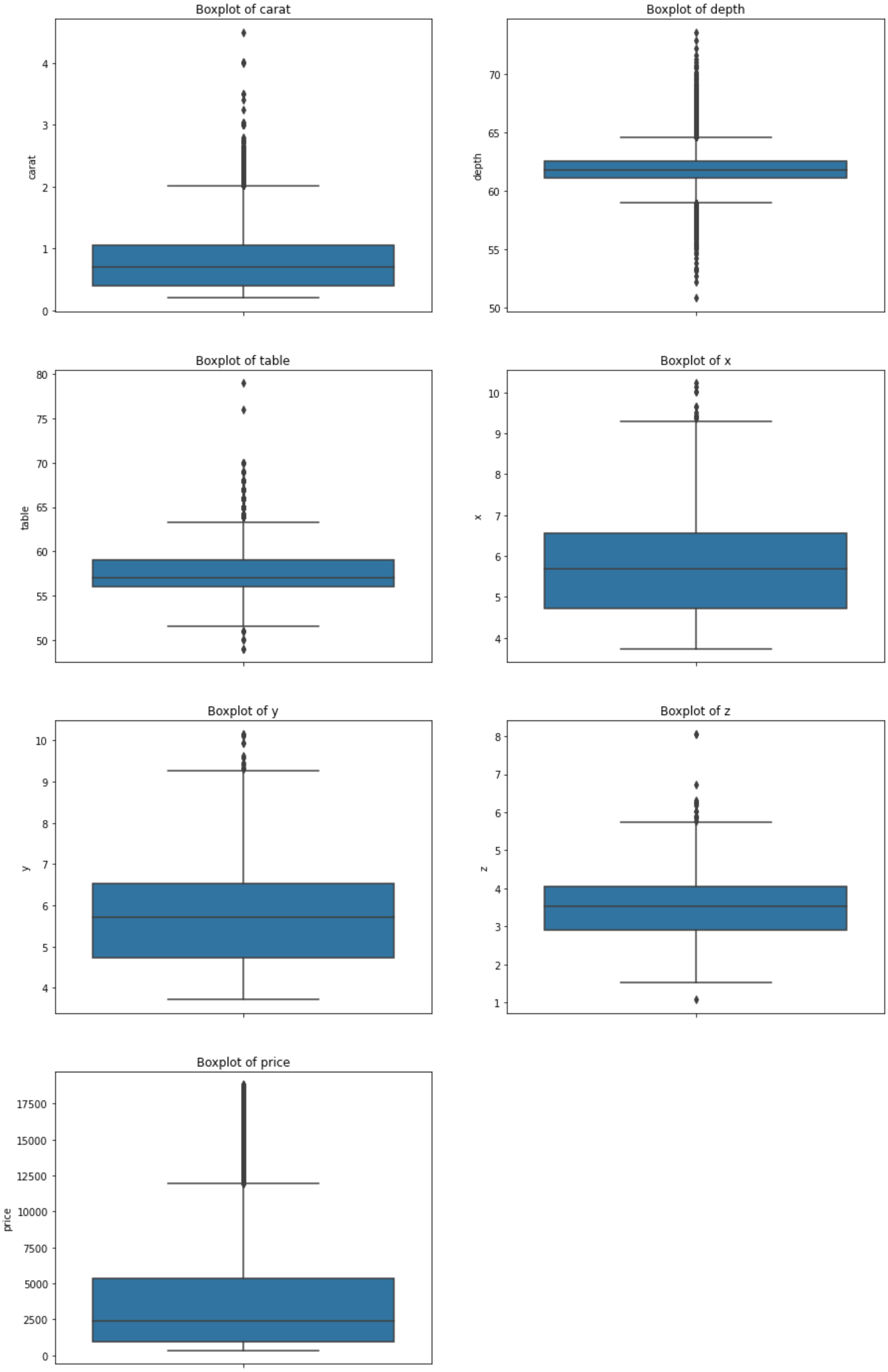
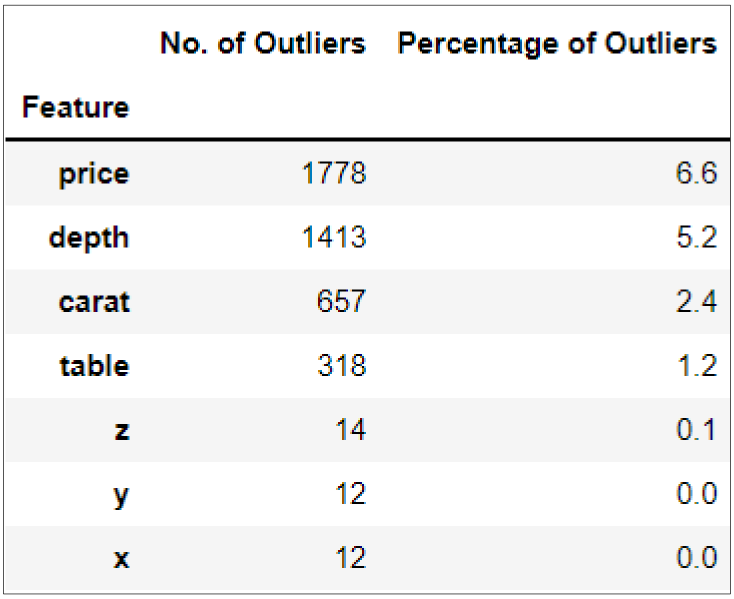


Figure 7. Box Plots of Numerical Features in Cubic Zirconia Dataset.



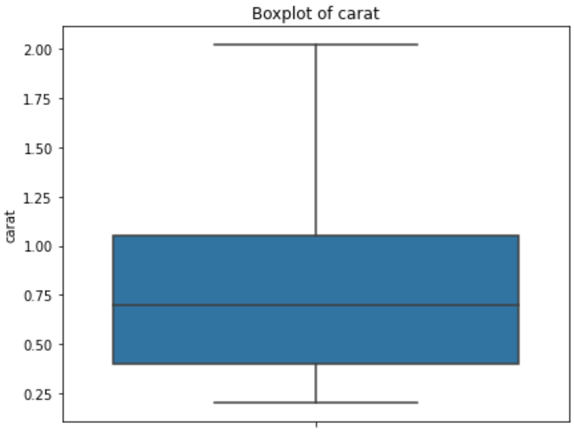
Feature	No. of Outliers	Percentage of Outliers
price	1778	6.6
depth	1413	5.2
carat	657	2.4
table	318	1.2
z	14	0.1
y	12	0.0
x	12	0.0

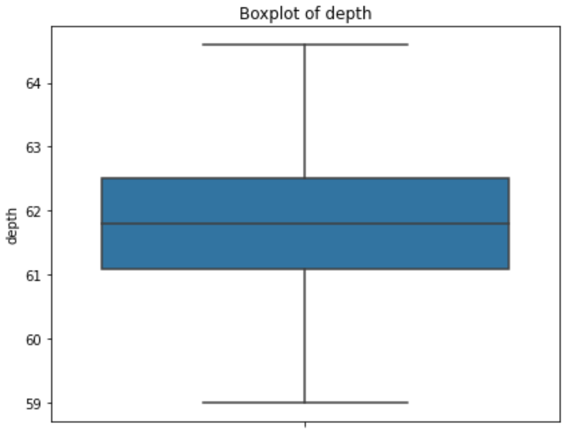
Table 12. Number of Outliers and Percentage of Outliers in Cubic Zirconia Dataset.

Insights:

1. There are outliers in almost all the features in the dataset.
2. Price, depth, carat and table features have 6.6%, 5.2%, 2.4% and 1.2% of outliers respectively.
3. Z feature has 0.1% of outliers in it.

As linear regression model is influenced by presence of outliers, Outliers are treated by capping and flooring method to its nearest whiskers. In this dataset, outliers are treated and box plots are drawn again to verify it.





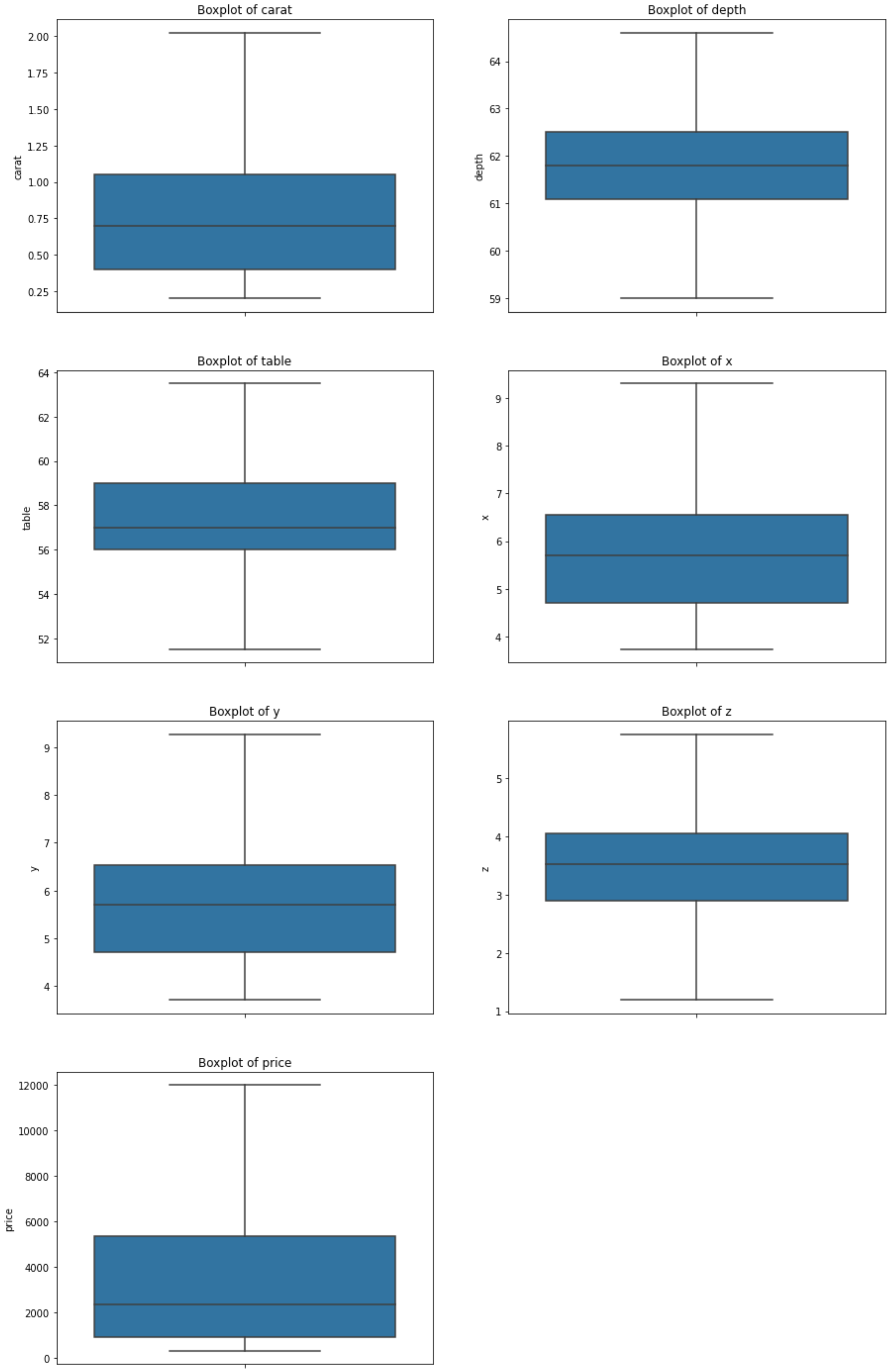


Figure 8. Box Plots of Numerical Features after removing outliers.

By observing above box plots, it can be inferred that there are **no more outliers** in the dataset and outliers have been treated properly.

Checking for the possibility of combining the sub levels of ordinal variables:

To check the possibility of combining the sub levels of ordinal variables, we can compare the mean price of cubic zirconia with different sublevels in each categorical variable by using bar plot of price vs categorical variables.

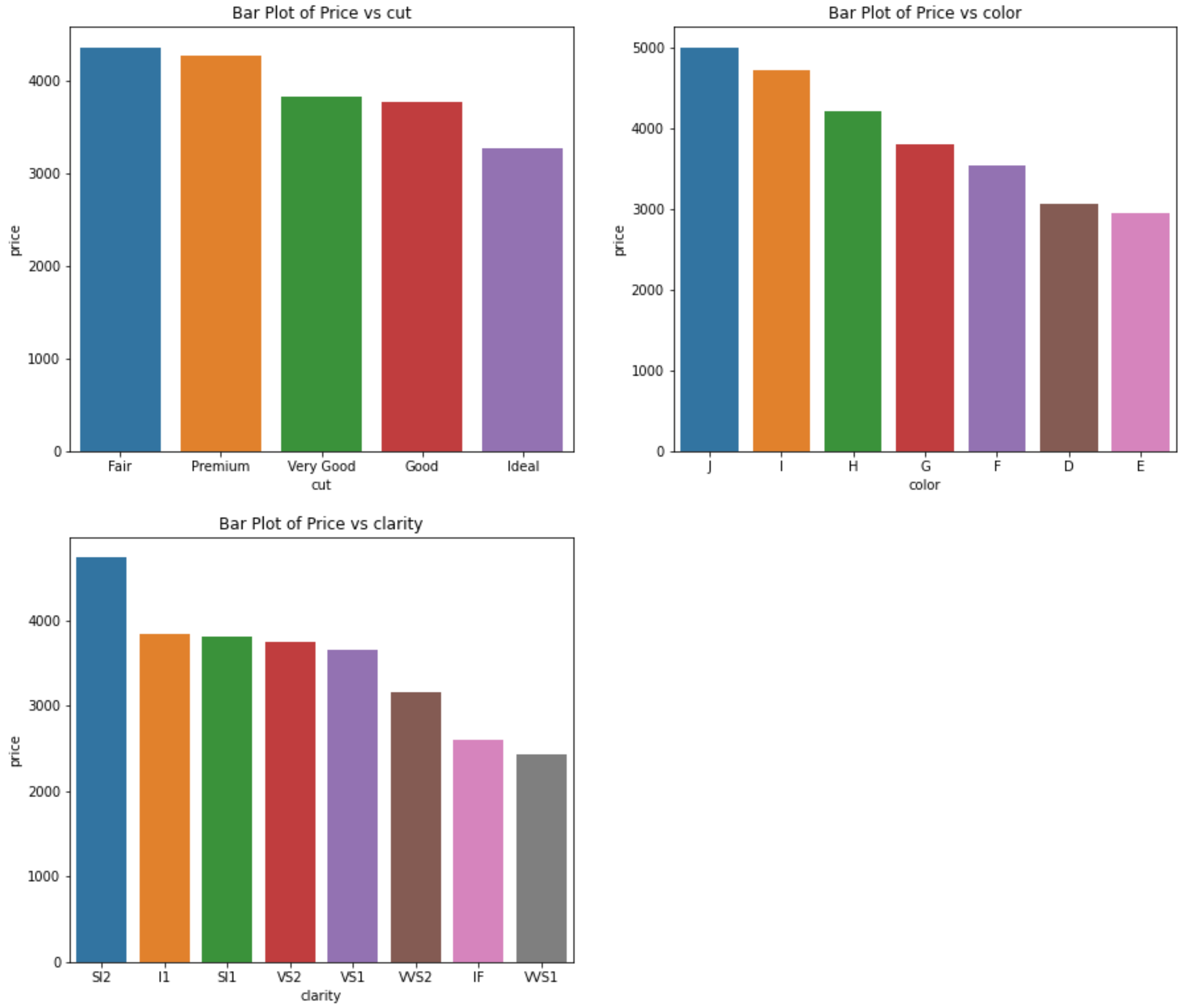


Figure 9. Bar Plots of Price vs Categorical Features.

1. **Cut Feature:** As mean prices of cubic zirconia with **very good cut quality and good cut quality** are almost same, these two sublevels can be combined and **renamed as very good** in cut quality feature.
2. **Colour Feature:**
   a. Mean prices of cubic zirconia with "D" & "E" colours are almost same. "D" & "E" colours can be combined to new colour "DE".

b. Mean prices of cubic zirconia with “F” & “G” colours are almost same. “F” & “G” colours can be combined to new colour “FG”.

c. Mean prices of cubic zirconia with “I” & “J” colours are almost same. “I” & “J” colours can be combined to new colour “IJ”.

3. **Clarity Feature:**

a. Mean prices of cubic zirconia with clarity “IF” & “VVS1” are almost same. “IF” & “VVS1” can be combined to new clarity “IF-VVS1”.

b. Mean prices of cubic zirconia with clarity “VS1”, “VS2” and “SI1” are almost same. These can be combined to new clarity “VS1-VS2-SI1”.

Bar plot of price vs categorical variables are shown below after combining the sublevels as per above discussion.

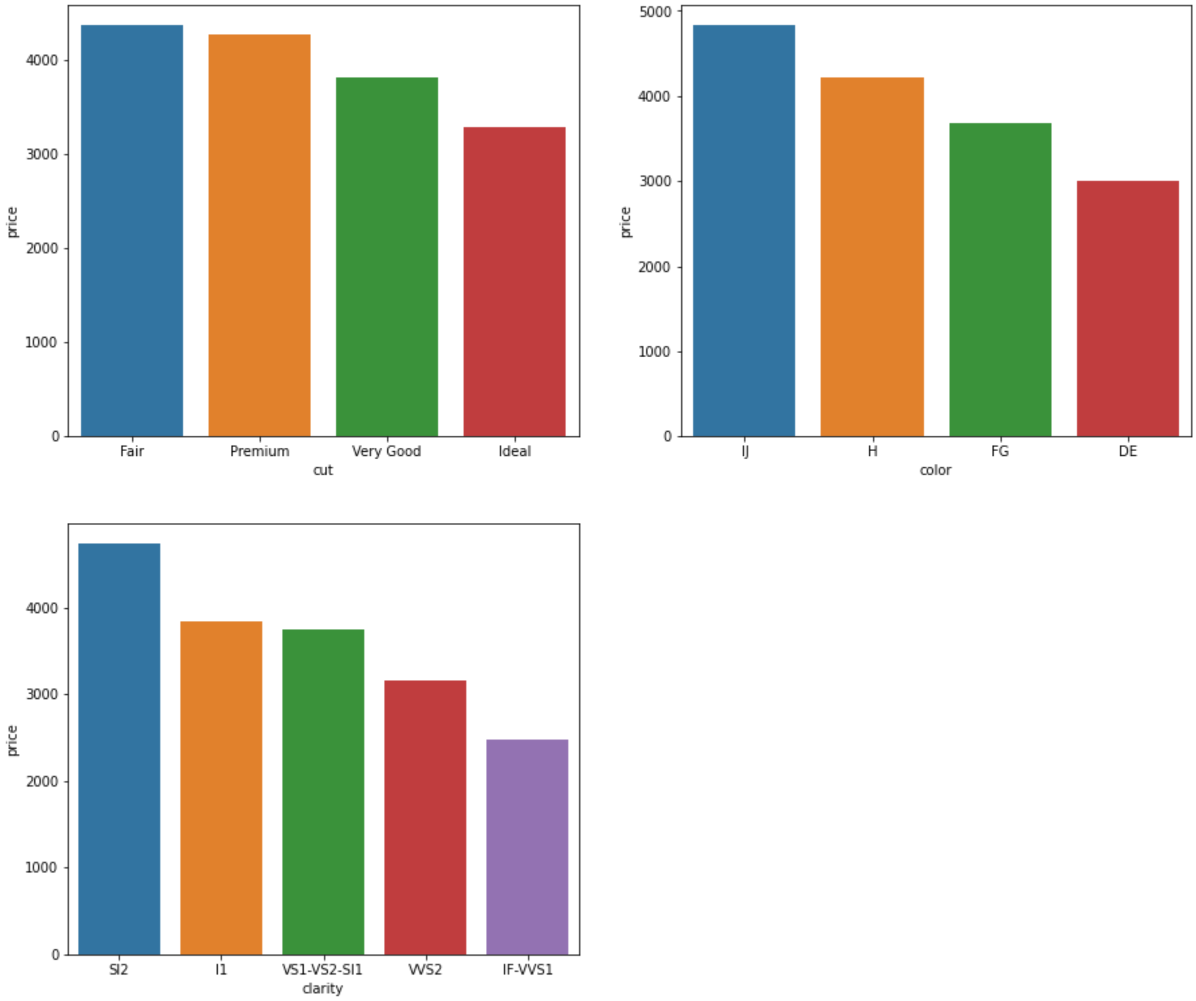


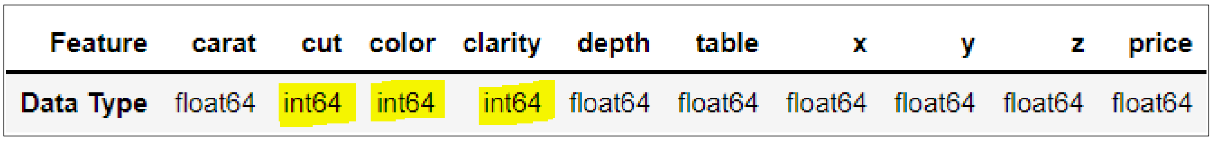
Figure 10. Bar Plots of Price vs Categorical Features after Combining Sublevels

Q1.3. Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from stats model. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adjusted Rsquare. Compare these models and select the best one with appropriate reasoning.

Sublevels in each ordinal categorical variable are encoded with integers **as per priority specified** in the question.

1. In cut feature, 'Ideal', 'Premium', 'Very Good' and 'Fair' sublevels are encoded with 4,3,2 and 1.
2. In colour feature, 'DE','FG','H' and 'IJ' sublevels are encoded with 1, 2, 3 and 4.
3. In clarity feature, 'IF-VVS1', 'VVS2', 'VS1-VS2-SI1','SI2' and 'I1' sublevels are encoded with 1, 2, 3, 4 and 5.

Let us check the data types of all features after encoding the categorical variables into integers.



Feature	carat	cut	color	clarity	depth	table	x	y	z	price
Data Type	float64	int64	int64	int64	float64	float64	float64	float64	float64	float64

Table 13. Data Types of Features after integer encoding.

LINEAR REGRESSION

Let us predict the price of the cubic zirconia by building linear regression models by three different ways.

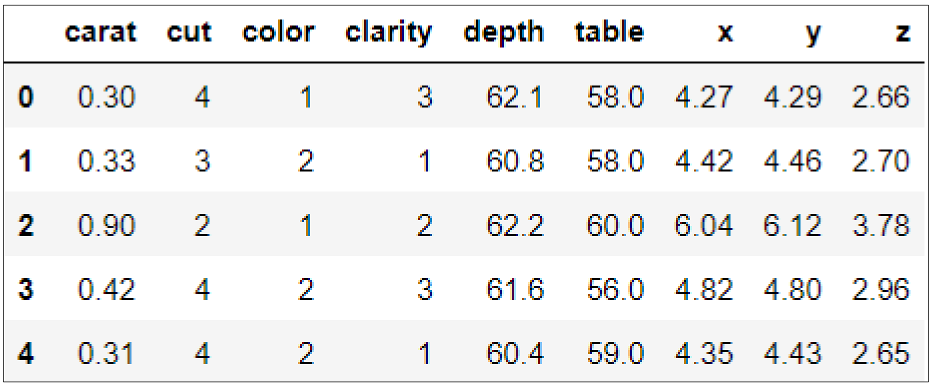
Model 1: Linear regression including all features

Model 2: Linear regression by dropping depth feature

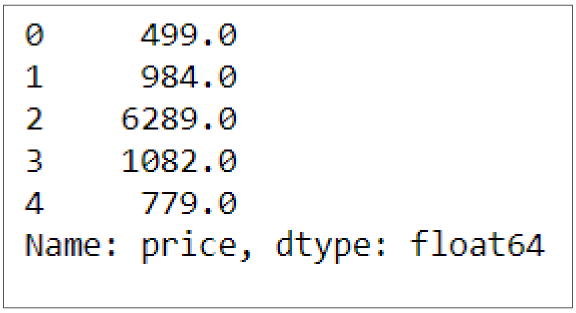
Model 3: Linear regression by dropping x, y, z, depth and table features

Model 1: Building Linear Regression Model Including All Features

Splitting the dataset in predictor and target variable



	carat	cut	color	clarity	depth	table	x	y	z
0	0.30	4	1	3	62.1	58.0	4.27	4.29	2.66
1	0.33	3	2	1	60.8	58.0	4.42	4.46	2.70
2	0.90	2	1	2	62.2	60.0	6.04	6.12	3.78
3	0.42	4	2	3	61.6	56.0	4.82	4.80	2.96
4	0.31	4	2	1	60.4	59.0	4.35	4.43	2.65

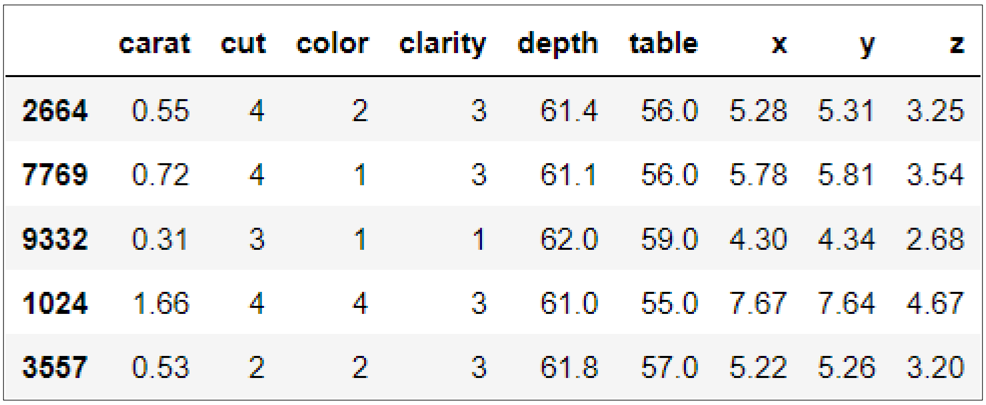


```
0      499.0
1      984.0
2     6289.0
3     1082.0
4      779.0
Name: price, dtype: float64
```

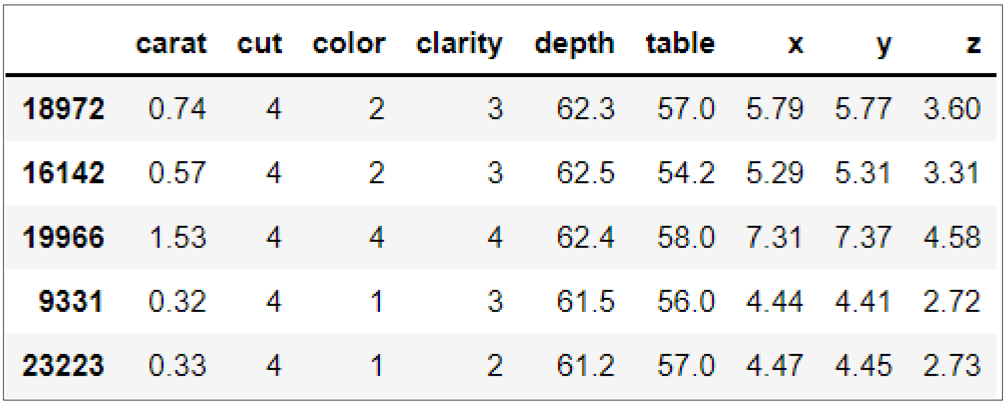
Table 14. Sample of Predictor and Target Variables in Cubic Zirconia Datasets

Splitting the data into Train and Test Sets

Predictor and target dataset have been divided into train and test sets in the ratio of 70:30.

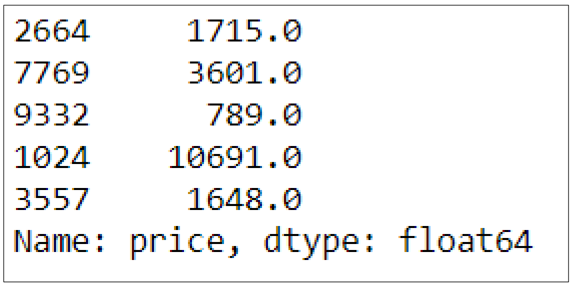


	carat	cut	color	clarity	depth	table	x	y	z
2664	0.55	4	2	3	61.4	56.0	5.28	5.31	3.25
7769	0.72	4	1	3	61.1	56.0	5.78	5.81	3.54
9332	0.31	3	1	1	62.0	59.0	4.30	4.34	2.68
1024	1.66	4	4	3	61.0	55.0	7.67	7.64	4.67
3557	0.53	2	2	3	61.8	57.0	5.22	5.26	3.20

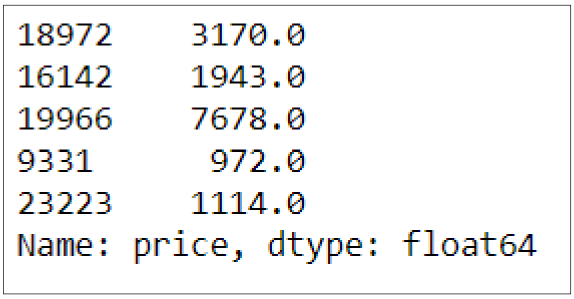


	carat	cut	color	clarity	depth	table	x	y	z
18972	0.74	4	2	3	62.3	57.0	5.79	5.77	3.60
16142	0.57	4	2	3	62.5	54.2	5.29	5.31	3.31
19966	1.53	4	4	4	62.4	58.0	7.31	7.37	4.58
9331	0.32	4	1	3	61.5	56.0	4.44	4.41	2.72
23223	0.33	4	1	2	61.2	57.0	4.47	4.45	2.73

Table 15. Samples of Predictors Train and Test Datasets.

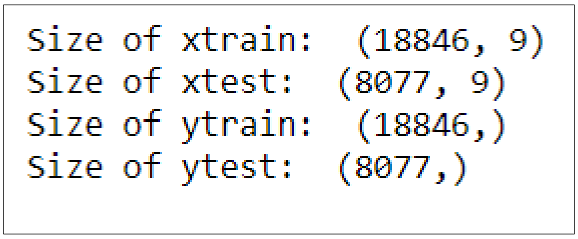


```
2664      1715.0
7769      3601.0
9332       789.0
1024     10691.0
3557      1648.0
Name: price, dtype: float64
```



```
18972     3170.0
16142     1943.0
19966     7678.0
9331       972.0
23223     1114.0
Name: price, dtype: float64
```

Sample of Target Train and Test Datasets.



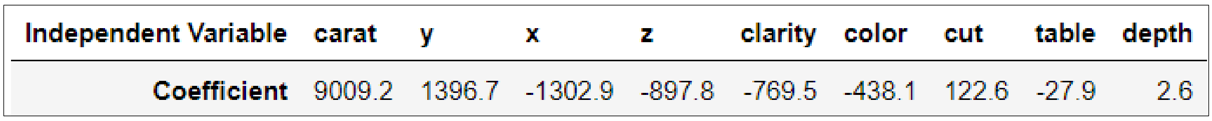
```
Size of xtrain:  (18846, 9)
Size of xtest:  (8077, 9)
Size of ytrain:  (18846,)
Size of ytest:  (8077,)
```

Size of Train and Test Datasets.

Linear Regression by Using Sci-kit learn library and Stats Model

- After splitting the dataset into train and test parts, linear regression model is built by using both Sci-kit learn library and stats model. It can be noticed that **performance metrics obtained for the models built in Sci-kit learn and stats models are same but few additional metrics (like adjusted Rsquare) are obtained in stats model.**
- Train datasets of predictors and target variables are used to fit the model.
- Target variable is predicted for both train and test datasets.
- Model has been evaluated based on performance metrics calculated for both train and test datasets.

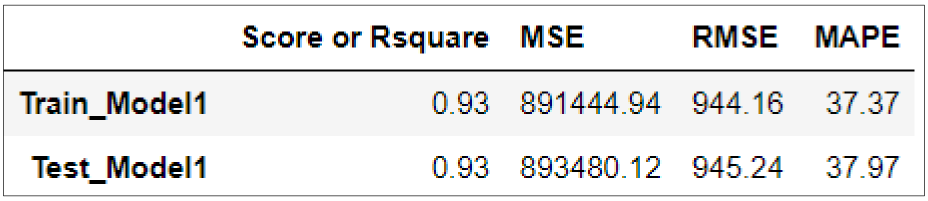
Coefficients for Model 1:



Independent Variable	carat	y	x	z	clarity	color	cut	table	depth
Coefficient	9009.2	1396.7	-1302.9	-897.8	-769.5	-438.1	122.6	-27.9	2.6

Table 16. Coefficients obtained in Model 1 of Linear Regression.

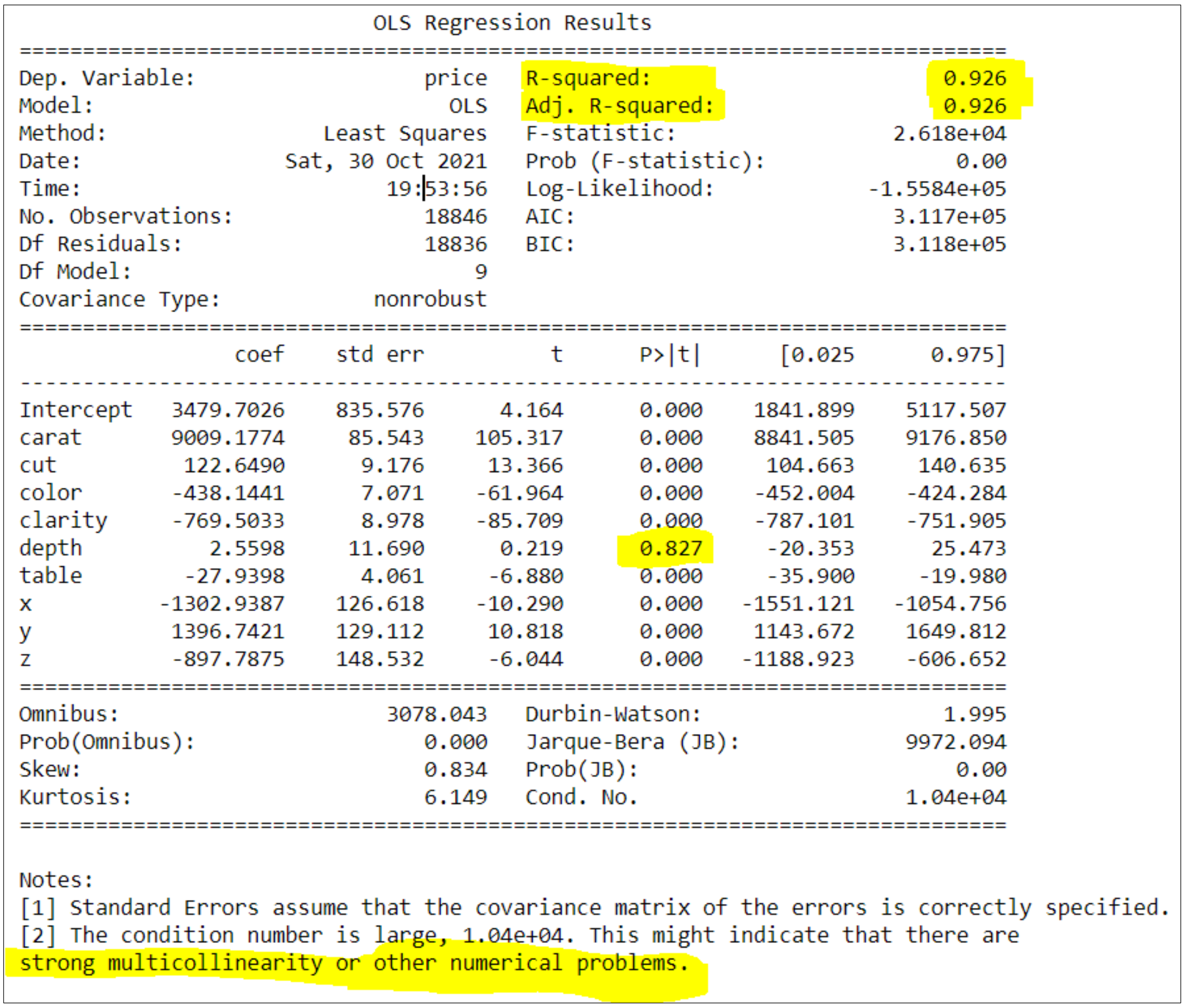
Performance Metrics for Model 1:



	Score or Rsquare	MSE	RMSE	MAPE
Train_Model1	0.93	891444.94	944.16	37.37
Test_Model1	0.93	893480.12	945.24	37.97

Table 17. Performance metrics obtained in Model 1 of Linear Regression.

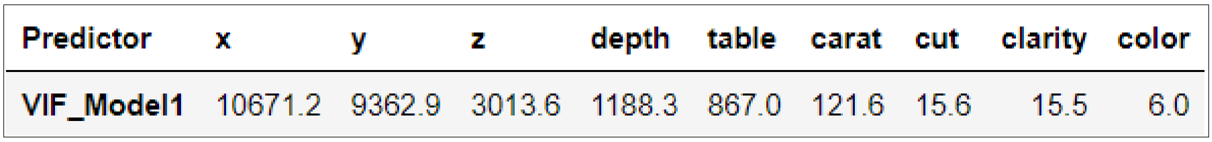
Summary of stats model for Model 1:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.926
Model:                            OLS   Adj. R-squared:                  0.926
Method:                 Least Squares   F-statistic:                 2.618e+04
Date:                Sat, 30 Oct 2021   Prob (F-statistic):               0.00
Time:                        19:53:56   Log-Likelihood:            -1.5584e+05
No. Observations:               18846   AIC:                         3.117e+05
Df Residuals:                   18836   BIC:                         3.118e+05
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   3479.7026    835.576      4.164      0.000    1841.899    5117.507
carat       9009.1774     85.543    105.317      0.000    8841.505    9176.850
cut          122.6490      9.176     13.366      0.000     104.663     140.635
color       -438.1441      7.071    -61.964      0.000    -452.004    -424.284
clarity     -769.5033      8.978    -85.709      0.000    -787.101    -751.905
depth          2.5598     11.690      0.219      0.827     -20.353      25.473
table        -27.9398      4.061     -6.880      0.000     -35.900     -19.980
x          -1302.9387    126.618    -10.290      0.000   -1551.121   -1054.756
y           1396.7421    129.112     10.818      0.000    1143.672    1649.812
z           -897.7875    148.532     -6.044      0.000   -1188.923    -606.652
==============================================================================
Omnibus:                     3078.043   Durbin-Watson:                   1.995
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             9972.094
Skew:                           0.834   Prob(JB):                         0.00
Kurtosis:                       6.149   Cond. No.                     1.04e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.04e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Table 18. Summary of stats model for Model 1.

Variance Inflation Factor for Model 1:



Predictor	x	y	z	depth	table	carat	cut	clarity	color
VIF_Model1	10671.2	9362.9	3013.6	1188.3	867.0	121.6	15.6	15.5	6.0

Variance Inflation Factor for Model 1.

Inferences:

- Rsquare (0.926) and adjusted Rsquare (0.926) are same. **Hence, there are no much weak predictors in dataset.**
- Model score for both train (0.93) and test (0.93) are same. Hence, there is **no over fitting in the model.** As the model score is more than 90%, it can used for predictions.
- RMSE for test dataset is 945.24 which is approximately 24% of mean price (3939.5).
- MAPE (Mean absolute percentage error) for test dataset is 37.97%.

- The intercept obtained in this model is 3479.7
- The list of features in descending order of importance (based on coefficients) is given below.

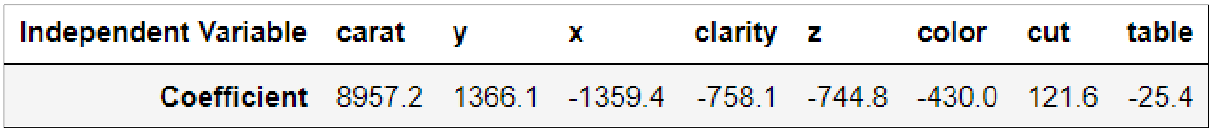
Carat > y > x > z > clarity > colour > cut > table > depth

- Carat is the strong predictor of price of cubic zirconia.
- P-Value for depth feature in the stats model is 0.827 and coefficient is 2.6 which indicate that there is no correlation between depth and price and **depth is a weak predictor of price**. Appropriate action is taken to address this issue and **will be discussed in model 2.**
- Variance Inflation Factor for all the features is more than 10 except colour feature. It implies that there is a **multicollinearity between predictor variables**. Appropriate action is taken to address this issue and **will be discussed in model 3.**

Model 2: Building Linear Regression Model by Dropping Depth Feature

As the P-Value for depth feature in the stats model is 0.827 and coefficient is 2.6 which indicate that there is no correlation between depth and price**. Depth feature is dropped from the dataset** and linear regression model is built by using both Sci-kit learn library and stats model. Model has been evaluated based on performance metrics calculated for both train and test datasets.

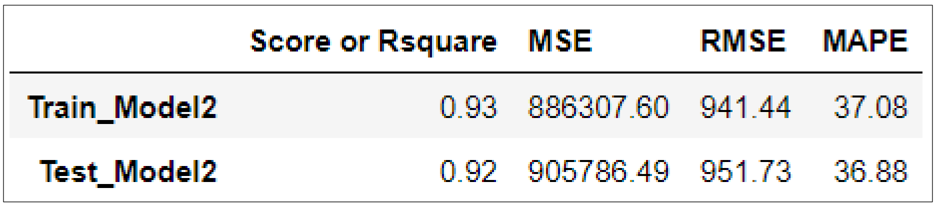
Coefficients for Model 2:



Independent Variable	carat	y	x	clarity	z	color	cut	table
Coefficient	8957.2	1366.1	-1359.4	-758.1	-744.8	-430.0	121.6	-25.4

Table 19. Coefficients obtained in Model 2 of Linear Regression.

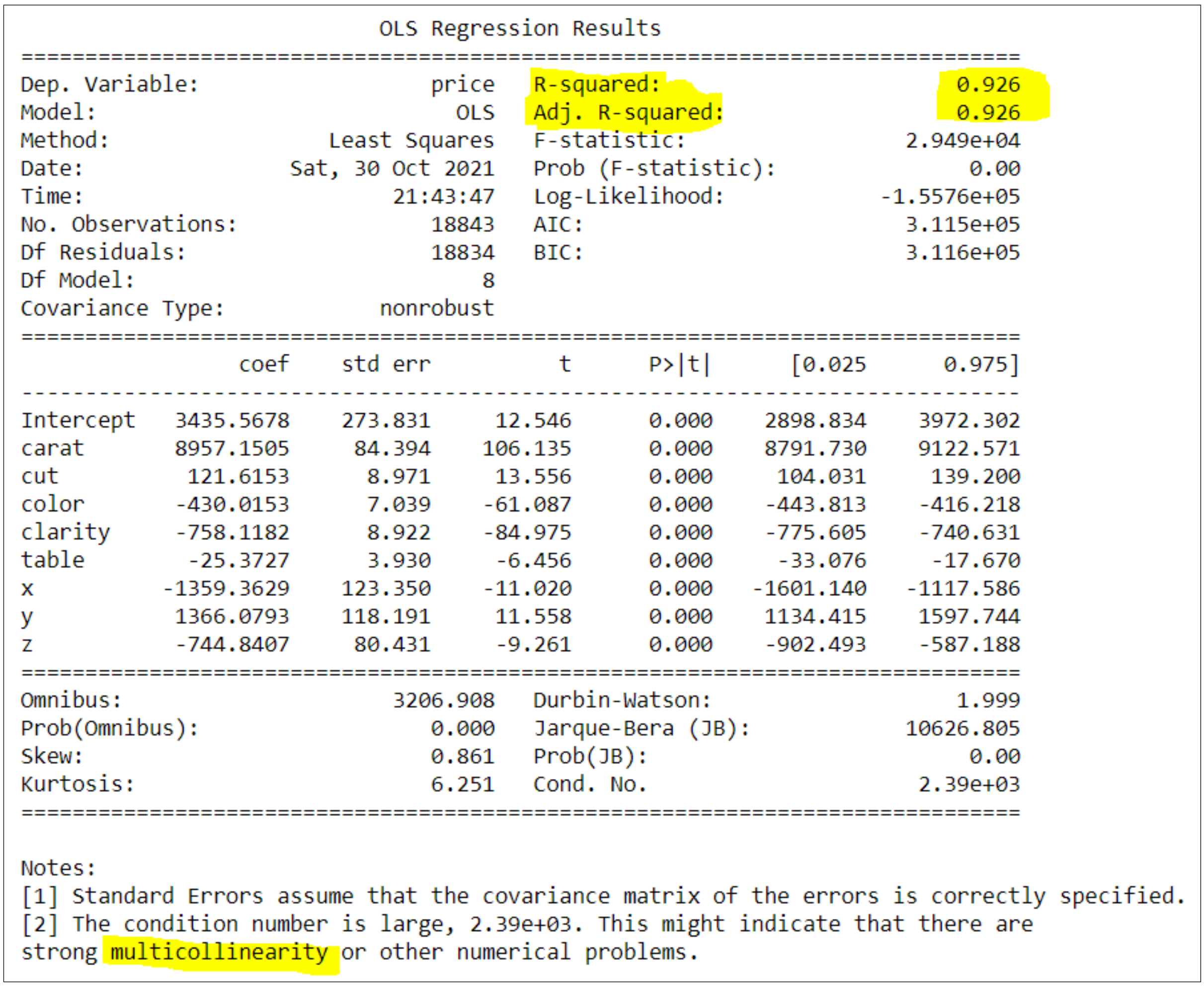
Performance Metrics for Model 2:



	Score or Rsquare	MSE	RMSE	MAPE
Train_Model2	0.93	886307.60	941.44	37.08
Test_Model2	0.92	905786.49	951.73	36.88

Table 20. Performance metrics obtained in Model 2 of Linear Regression.

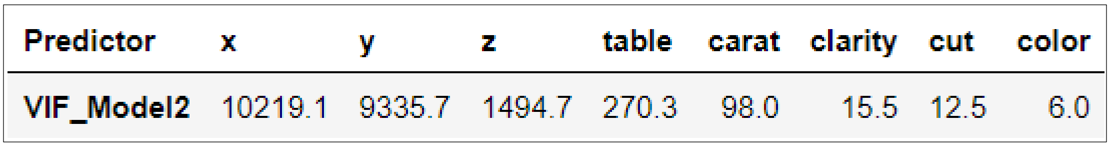
Summary of stats model for Model 2:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.926
Model:                            OLS   Adj. R-squared:                  0.926
Method:                 Least Squares   F-statistic:                 2.949e+04
Date:                Sat, 30 Oct 2021   Prob (F-statistic):               0.00
Time:                        21:43:47   Log-Likelihood:            -1.5576e+05
No. Observations:               18843   AIC:                         3.115e+05
Df Residuals:                   18834   BIC:                         3.116e+05
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   3435.5678    273.831     12.546      0.000    2898.834    3972.302
carat       8957.1505     84.394    106.135      0.000    8791.730    9122.571
cut          121.6153      8.971     13.556      0.000     104.031     139.200
color       -430.0153      7.039    -61.087      0.000    -443.813    -416.218
clarity     -758.1182      8.922    -84.975      0.000    -775.605    -740.631
table        -25.3727      3.930     -6.456      0.000     -33.076     -17.670
x          -1359.3629    123.350    -11.020      0.000   -1601.140   -1117.586
y           1366.0793    118.191     11.558      0.000    1134.415    1597.744
z           -744.8407     80.431     -9.261      0.000    -902.493    -587.188
==============================================================================
Omnibus:                     3206.908   Durbin-Watson:                   1.999
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            10626.805
Skew:                           0.861   Prob(JB):                         0.00
Kurtosis:                       6.251   Cond. No.                     2.39e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.39e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Table 21. Summary of stats model for Model 2.

Variance Inflation Factor for Model 2:



Predictor	x	y	z	table	carat	clarity	cut	color
VIF_Model2	10219.1	9335.7	1494.7	270.3	98.0	15.5	12.5	6.0

Table 22. Variance Inflation Factor for Model 2.

Inferences:

- Rsquare (0.926) and adjusted Rsquare (0.926) are same. **Hence, there are no more weak predictors in dataset.**
- Model score for both train (0.93) and test (0.92) are comparable. Hence, there is **no over fitting in the model.** As the model score is more than 90%, it can used for predictions.
- RMSE for test dataset is 951.73 which is approximately 24% of mean price (3939.5).
- MAPE (Mean absolute percentage error) for test dataset is 36.88%.
- The intercept obtained in this model is 3435.6

- The list of features in descending order of importance (based on coefficients) is given below.

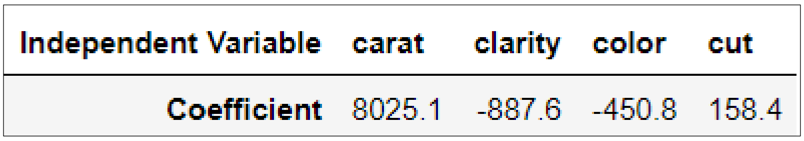
Carat > **y** > **x** > **clarity** > **z** > **colour** > **cut** > **table**

- Carat is the strong predictor of price of cubic zirconia.
- Variance Inflation Factor for all the features is more than 10. It implies that there is a **multicollinearity between predictor variables**. Appropriate action is taken to address this issue and **will be discussed in model 3.**

Model 3: Building Linear Regression Model by Dropping x, y, z, Depth and table Features

- As we discussed in both model 1 and model 2, we have **multicollinearity issue** in both of them. As x, y, z and carat features are **strongly correlated** and x, y, z features have **high VIF than carat** in model 1 and model 2. **x, y and z features are dropped** from the dataset to reduce multicollinearity.
- As **Depth and table are weakly correlated with price** and these features have **low coefficients.** Hence, Depth and table are considered as weak predictors of price and these features are dropped for further analysis.
- linear regression model is built by using both Sci-kit learn library and stats model. Model has been evaluated based on performance metrics calculated for both train and test datasets.

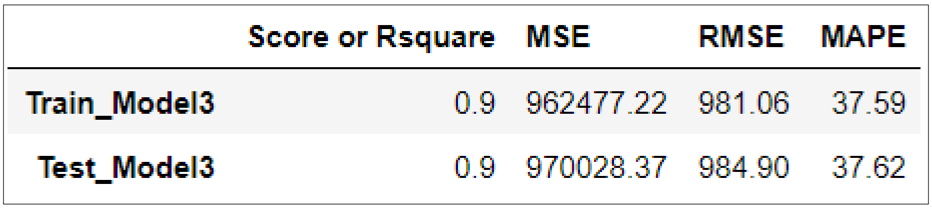
Coefficients for Model 3:



Independent Variable	carat	clarity	color	cut
Coefficient	8025.1	-887.6	-450.8	158.4

Table 23. Coefficients obtained in Model 3 of Linear Regression.

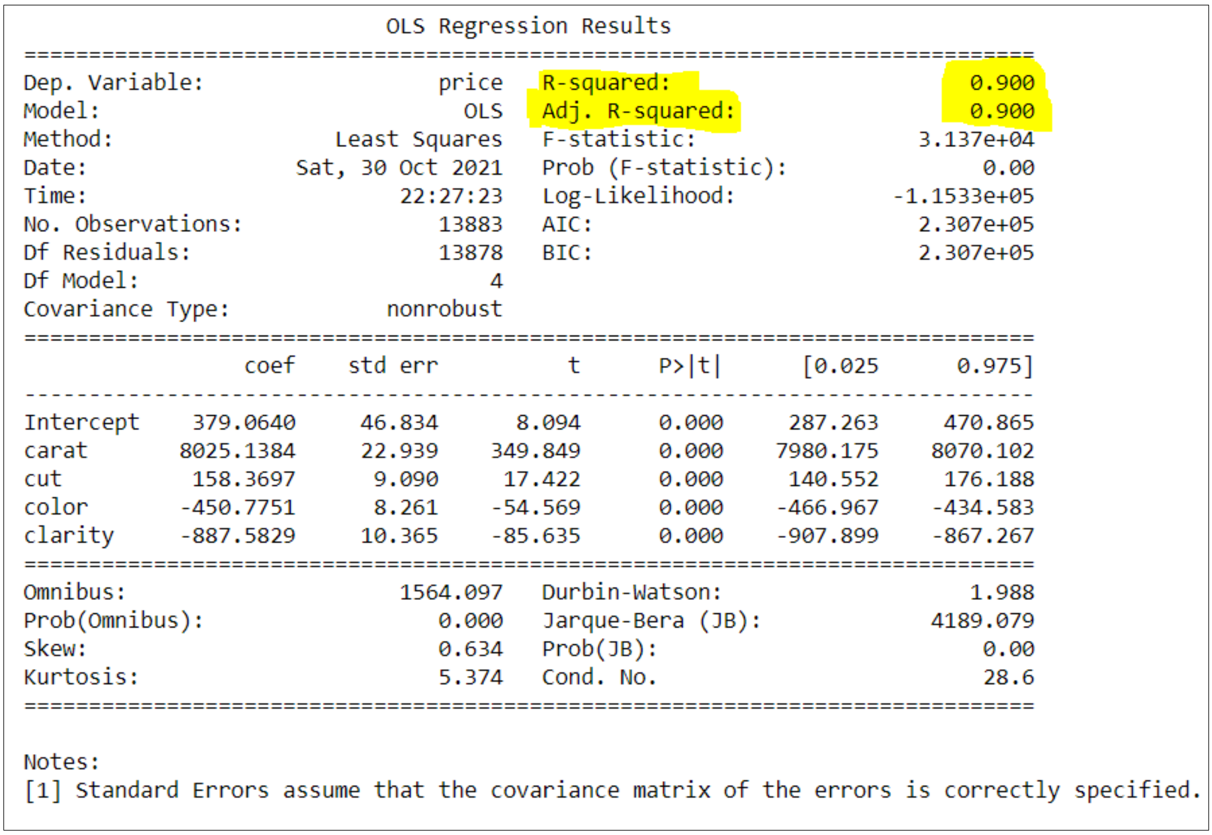
Performance Metrics for Model 3:



	Score or Rsquare	MSE	RMSE	MAPE
Train_Model3	0.9	962477.22	981.06	37.59
Test_Model3	0.9	970028.37	984.90	37.62

Table 24. Performance metrics obtained in Model 3 of Linear Regression.

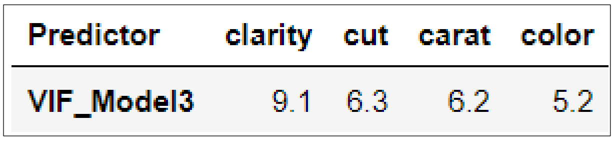
Summary of stats model for Model 3:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.900
Model:                            OLS   Adj. R-squared:                  0.900
Method:                 Least Squares   F-statistic:                 3.137e+04
Date:                Sat, 30 Oct 2021   Prob (F-statistic):               0.00
Time:                        22:27:23   Log-Likelihood:            -1.1533e+05
No. Observations:               13883   AIC:                         2.307e+05
Df Residuals:                   13878   BIC:                         2.307e+05
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    379.0640     46.834      8.094      0.000     287.263     470.865
carat       8025.1384     22.939    349.849      0.000    7980.175    8070.102
cut          158.3697      9.090     17.422      0.000     140.552     176.188
color       -450.7751      8.261    -54.569      0.000    -466.967    -434.583
clarity     -887.5829     10.365    -85.635      0.000    -907.899    -867.267
==============================================================================
Omnibus:                     1564.097   Durbin-Watson:                   1.988
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4189.079
Skew:                           0.634   Prob(JB):                         0.00
Kurtosis:                       5.374   Cond. No.                         28.6
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Table 25. Summary of stats model for Model 3.

Variance Inflation Factor for Model 3:



Predictor	clarity	cut	carat	color
VIF_Model3	9.1	6.3	6.2	5.2

Table 26. Variance Inflation Factor for Model 3.

Inferences:

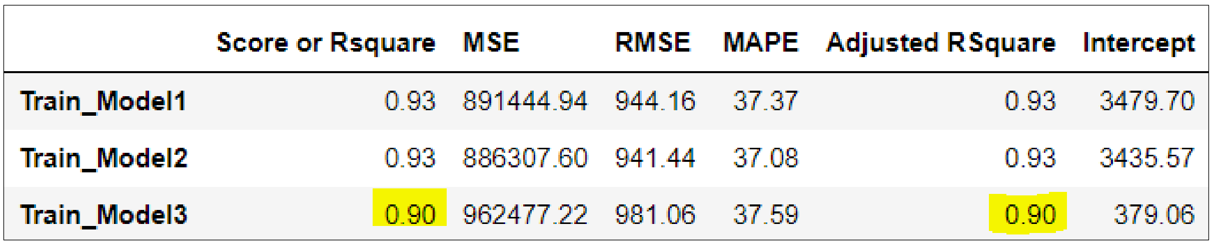
- Rsquare (0.9) and adjusted Rsquare (0.9) are same. **Hence, there are no more weak predictors in dataset.**
- Model score for both train (0.9) and test (0.9) are same. Hence, there is **no over fitting in the model.** As the model score is more than equal to 90%, it can used for predictions.
- RMSE for test dataset is 984.9 which is approximately 25% of mean price (3939.5).
- MAPE (Mean absolute percentage error) for test dataset is 37.62%.
- The intercept obtained in this model is 379.06
- The list of features in descending order of importance (based on coefficients) is given below.

Carat > clarity > colour > cut

- Carat is the strong predictor of price of cubic zirconia.
- Variance Inflation Factor for all the features is less than 10. It implies that there is **No multicollinearity between predictor variables**.

COMPARISON OF MODEL 1, MODEL 2 AND MODEL 3

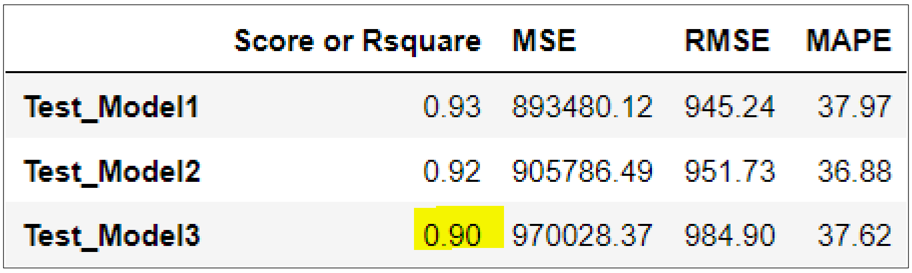
Comparison of Performance Metrics on Train Dataset:



	Score or Rsquare	MSE	RMSE	MAPE	Adjusted RSquare	Intercept
Train_Model1	0.93	891444.94	944.16	37.37	0.93	3479.70
Train_Model2	0.93	886307.60	941.44	37.08	0.93	3435.57
Train_Model3	0.90	962477.22	981.06	37.59	0.90	379.06

Table 27. Comparison of Performance Metrics on Train Dataset

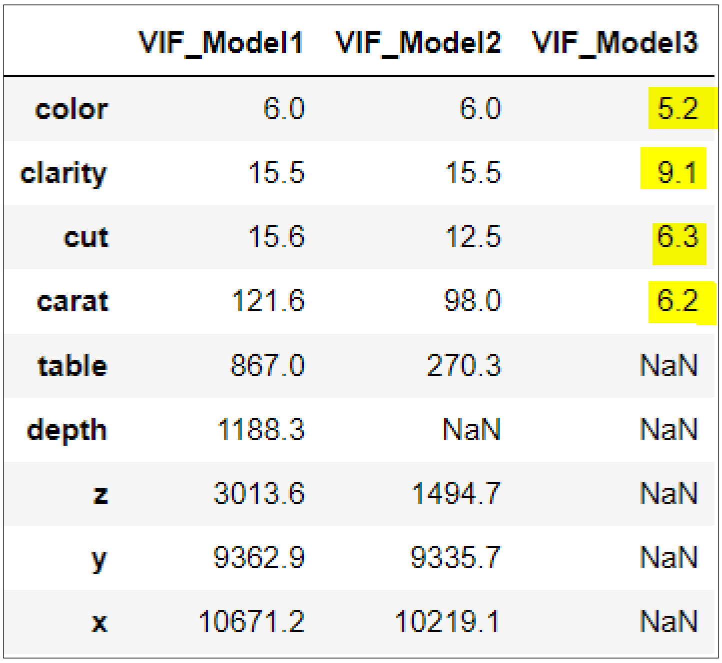
Comparison of Performance Metrics on Test Dataset:



	Score or Rsquare	MSE	RMSE	MAPE
Test_Model1	0.93	893480.12	945.24	37.97
Test_Model2	0.92	905786.49	951.73	36.88
Test_Model3	0.90	970028.37	984.90	37.62

Table 28. Comparison of Performance Metrics on Test Dataset

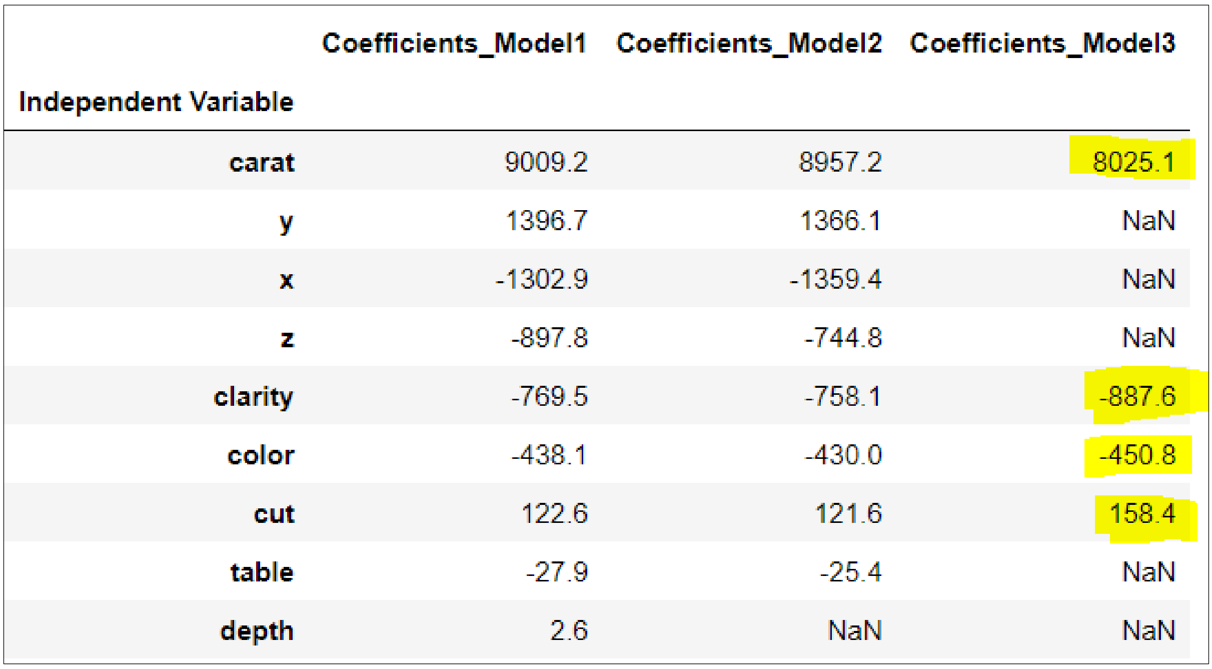
Comparison of Variance Inflation Factor



	VIF_Model1	VIF_Model2	VIF_Model3
color	6.0	6.0	5.2
clarity	15.5	15.5	9.1
cut	15.6	12.5	6.3
carat	121.6	98.0	6.2
table	867.0	270.3	NaN
depth	1188.3	NaN	NaN
z	3013.6	1494.7	NaN
y	9362.9	9335.7	NaN
x	10671.2	10219.1	NaN

Table 29. Comparison of Variance Inflation Factor

Comparison of Coefficients:



	Coefficients_Model1	Coefficients_Model2	Coefficients_Model3
Independent Variable			
carat	9009.2	8957.2	8025.1
y	1396.7	1366.1	NaN
x	-1302.9	-1359.4	NaN
z	-897.8	-744.8	NaN
clarity	-769.5	-758.1	-887.6
color	-438.1	-430.0	-450.8
cut	122.6	121.6	158.4
table	-27.9	-25.4	NaN
depth	2.6	NaN	NaN

Table 30. Comparison of Coefficients

Conclusions:

Notation:

Model 1: Linear regression including all features

Model 2: Linear regression by dropping depth feature

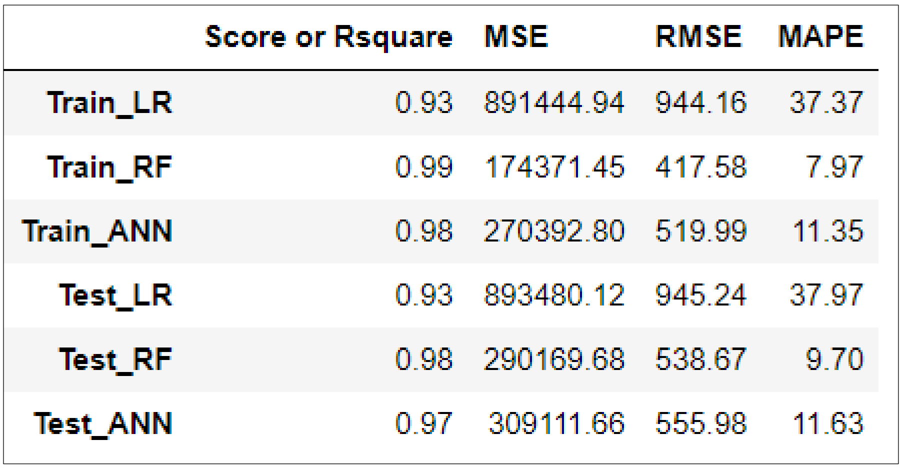
Model 3: Linear regression by dropping x, y, z, depth and table features

1. As **Rsquare of the model1 (0.93) is maximum** and also there is no much difference in RMSE and MAPE for different models, If the exact relation between each predictor and price is not required (**i.e., Accepting multicollinearity**), **then model 1 is the best one.**
2. As **Rsquare of the model3 (0.9) is slightly less than model1 (0.93)** and also there is no much difference in RMSE and MAPE for different models, If the exact relation between each predictor and price is required **(i.e., Not accepting multicollinearity**), **then model 3 is the best one. Additionally, number of independent variables required to predict the price is decreased in this model.**
3. Variance Inflation Factors (VIF) are **decreased to less than 10 in model 3**. Model 3 is the best one to avoid multicollinearity.
4. Neither Rsquare (0.92) is improved nor RMSE is decreased in model 2 compared to model1 and also there is multicollinearity in model2. Hence, model2 is never preferred.
5. The strong predictors in **the decreasing order of their importance** based on their coefficients is as below as per model 3.

Carat (8025.1) > clarity (-887.6) > colour (-450.8) > cut (158.4)

Additionally, let us compare the performance metrics of Linear Regression model with other regression model like Random Forest Regression and Artificial Neural Networks Regression to find the best model.

- After splitting the dataset into train and test parts, Regression model is built by using Random Forest Regressor.
- Additionally, above train and test datasets of predictor variables are scaled by Z-Score method to fit the model in Artificial Neural Networks Regression.
- Best hyper parameters for Random Forest Regression are listed below.
  a. Maximum depth: 12
  b. Maximum features: 7
  c. Minimum samples in Leaf node: 5
  d. Minimum samples split node: 10
  e. Number of estimators: 20
- Best hyper parameters for Artificial Neural Networks Regression are listed below.
  a. Hidden layer sizes: 50
  b. Maximum Iterations: 50
  c. Solver: lbfgs
- Target variable is predicted for both train and test datasets in both the models.
- Models has been evaluated based on performance metrics calculated for both train and test datasets.



	Score or Rsquare	MSE	RMSE	MAPE
Train_LR	0.93	891444.94	944.16	37.37
Train_RF	0.99	174371.45	417.58	7.97
Train_ANN	0.98	270392.80	519.99	11.35
Test_LR	0.93	893480.12	945.24	37.97
Test_RF	0.98	290169.68	538.67	9.70
Test_ANN	0.97	309111.66	555.98	11.63

From above table, it can be noticed than random forest regression and ANN regression are performing better than linear regression.

Q1.4. Inference: Basis on these predictions, what are the business insights and recommendations.

(Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.)

Steps Performed

The following are the steps perfoermed in this project

1. Few duplicate records are found in the gven dataset and they are dropped in EDA section.
2. Anamolies, null values and outlies in the dataset are treated with suitable methods.
3. Few sublevels in different categorcal variabls are combined by checking the possibility
4. Data set has beeen slitted into train and test ration in the ratio of 70:30
5. Linear regression is performed and evaluated by buliding three different linear regression models.
6. Additionally, regression is performed in radom forest and neural networks models also.

Business Interpretations and Actionable Insights

Notation:

Model 1: Linear regression including all features

Model 2: Linear regression by dropping depth feature

Model 3: Linear regression by dropping x, y, z, depth and table features

1. As **Rsquare of the model1 (0.93) is maximum** and also there is no much difference in RMSE and MAPE for different models, If the exact relation between each predictor and price is not required (**i.e., Accepting multicollinearity**), **then model 1 is the best one.**
2. As **Rsquare of the model3 (0.9) is slightly less than model1 (0.93)** and also there is no much difference in RMSE and MAPE for different models, If the exact relation between each predictor and price is required **(i.e., Not accepting multicollinearity**), **then model 3 is the best one. Additionally, number of independent variables required to predict the price is decreased in this model.**
3. Variance Inflation Factors (VIF) are **decreased to less than 10 in model 3**. Model 3 is the best one to avoid multicollinearity.
4. Neither Rsquare (0.92) is improved nor RMSE is decreased in model 2 compared to model1 and also there is multicollinearity in model2. Hence, model2 is never preferred.

5. The strong predictors in **the decreasing order of their importance** based on their coefficients is as below as per model 3.

   Carat (8025.1) > clarity (-887.6) > colour (-450.8) > cut (158.4)

6. Median price of Cubic Zirconia with fair quality is maximum and median price of Cubic Zirconia with Ideal quality is minimum. Decreasing order of median price of cubic zirconia with different cut quality is given below.

   Fair > Premium > Good > Very Good > Ideal

   It can be noticed that as cut quality of cubic zirconia increases, median price decreases except for premium cut quality. Usually, it is not expected. As the price is more, it is better to produce cubic zirconia with fair and premium cut quality to increase the profit.

7. Median price of Cubic Zirconia with ‘J’ colour is maximum and median price of Cubic Zirconia with ‘E’ colour is minimum. Decreasing order of median price of cubic zirconia with different colour is given below.

   J > I > H > F > G > D > E

   It can be noticed that as colour quality of cubic zirconia increases, median price also increases. As the price is more, it is better to produce cubic zirconia with J, I and H colours to increase the profit.

8. Median price of Cubic Zirconia with SI2 clarity is maximum and median price of Cubic Zirconia with IF clarity is minimum. Decreasing order of median price of cubic zirconia with different clarity is given below.

   SI2 > I1 > SI1 > VS2 > VS1 > VVS2 > VVS1 > IF

   It can be noticed that as clarity of cubic zirconia increases, median price also increases except for I1 clarity. As the price is more, it is better to produce cubic zirconia with SI2, I1 and SI1 clarity to increase the profit.

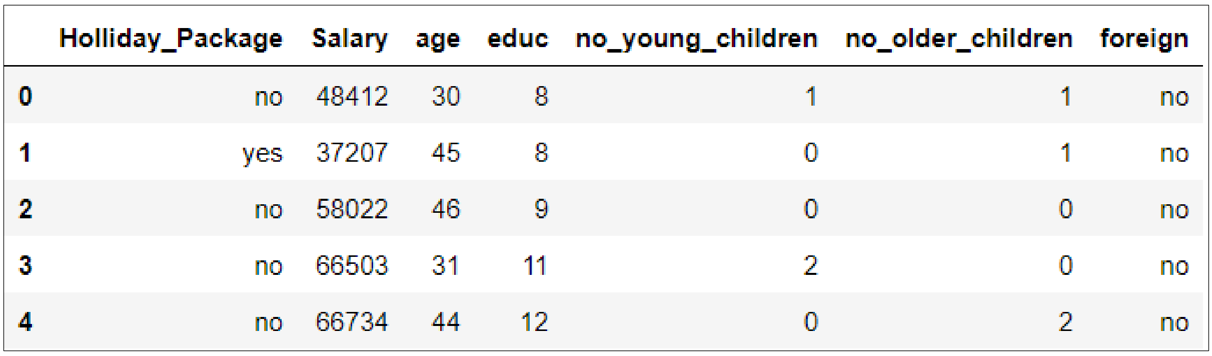
PROBLEM 2: LOGISTIC REGRESSION & LDA

Problem Statement:

You are hired by a tour and travel agency which deals in selling holiday packages. You are provided details of 872 employees of a company. Among these employees, some opted for the package and some didn't. You have to help the company in predicting whether an employee will opt for the package or not on the basis of the information given in the data set. Also, find out the important factors on the basis of which the company will focus on particular employees to sell their packages.

Q2.1. Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it? Perform Univariate and Bivariate Analysis. Do exploratory data analysis.

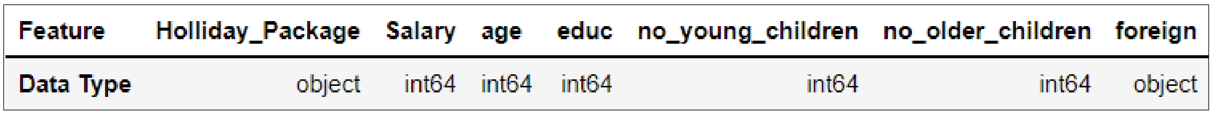
Sample of the Dataset:



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

Table 31. Sample of Holiday Package Dataset.

Data Types of Variables in the Dataset:



Feature	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
Data Type	object	int64	int64	int64	int64	int64	object

Table 32. Data Types of All Features in the Holiday Package Dataset.

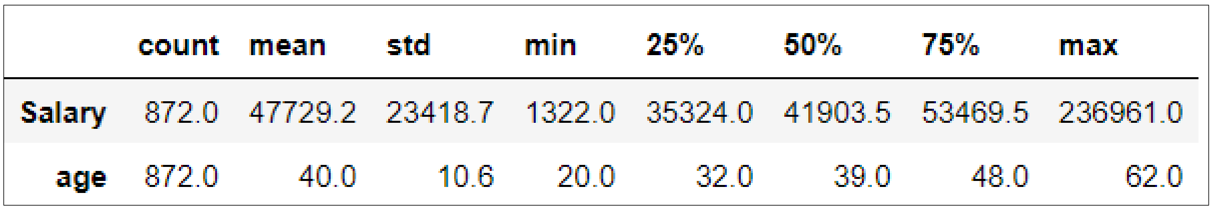
Insights:

1. There are 7 features (columns) with 872 observations (rows) in the dataset.
2. The dataset has two categorical features i.e., holiday package and foreign and their data type are object.
3. The dataset has two continuous numerical features i.e., salary and age and their data type are int64.

4. The dataset has three discrete numerical features i.e., education, no. of young children and no. of old children and their data type are int64.
5. The target variable in this dataset is holiday package.

Description of the Dataset

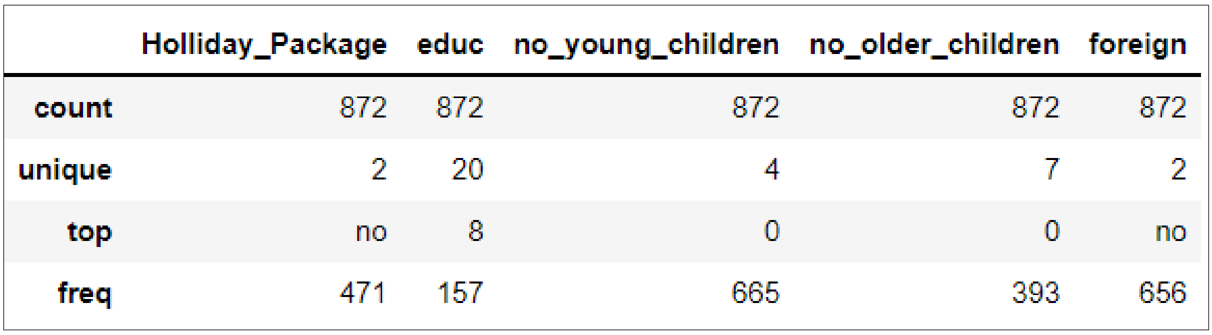
Continuous Numerical Features:



	count	mean	std	min	25%	50%	75%	max
Salary	872.0	47729.2	23418.7	1322.0	35324.0	41903.5	53469.5	236961.0
age	872.0	40.0	10.6	20.0	32.0	39.0	48.0	62.0

Table 33. Description of Numerical Features in Holiday Package Dataset.

Discrete Numerical and Categorical Features:

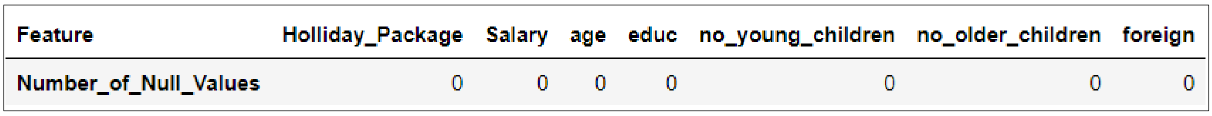


	Holliday_Package	educ	no_young_children	no_older_children	foreign
count	872	872	872	872	872
unique	2	20	4	7	2
top	no	8	0	0	no
freq	471	157	665	393	656

Table 34. Description of Discrete Numerical and Categorical Features in Holiday Package Dataset.

Exploratory Data Analysis

Checking Null values in the Dataset:



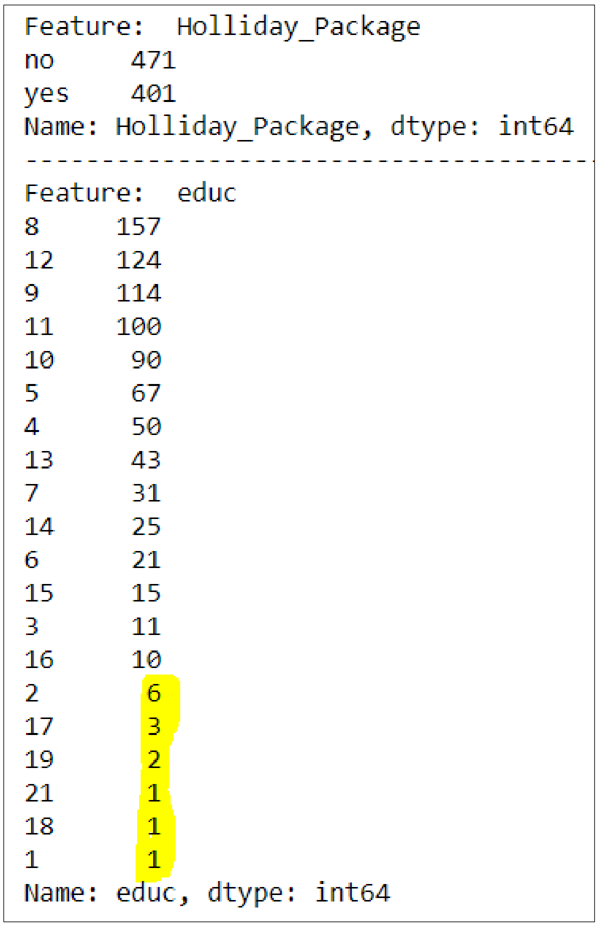
Feature	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
Number_of_Null_Values	0	0	0	0	0	0	0

Table 35. Null Values in the Holiday Package Dataset.

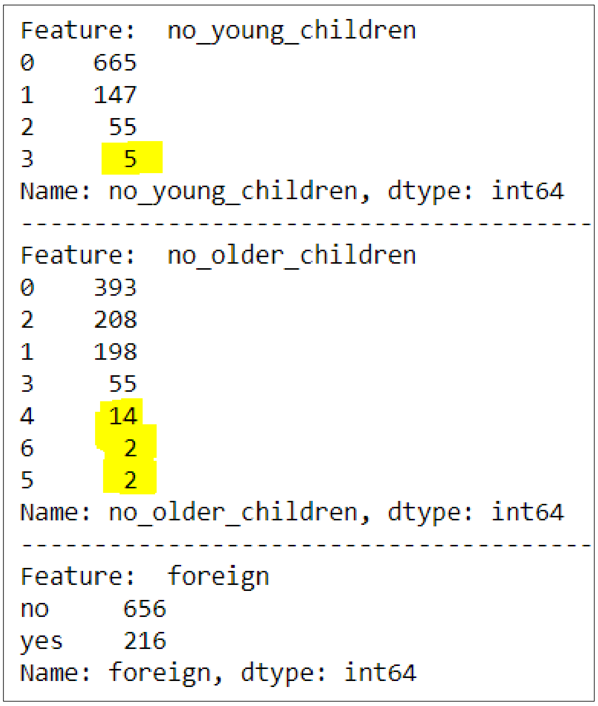
- There are **no duplicate records** in the dataset.
- There are **no null values** in the dataset.

Checking for Anomalies

Checking Value Counts in Discrete Numerical and Categorical Variables

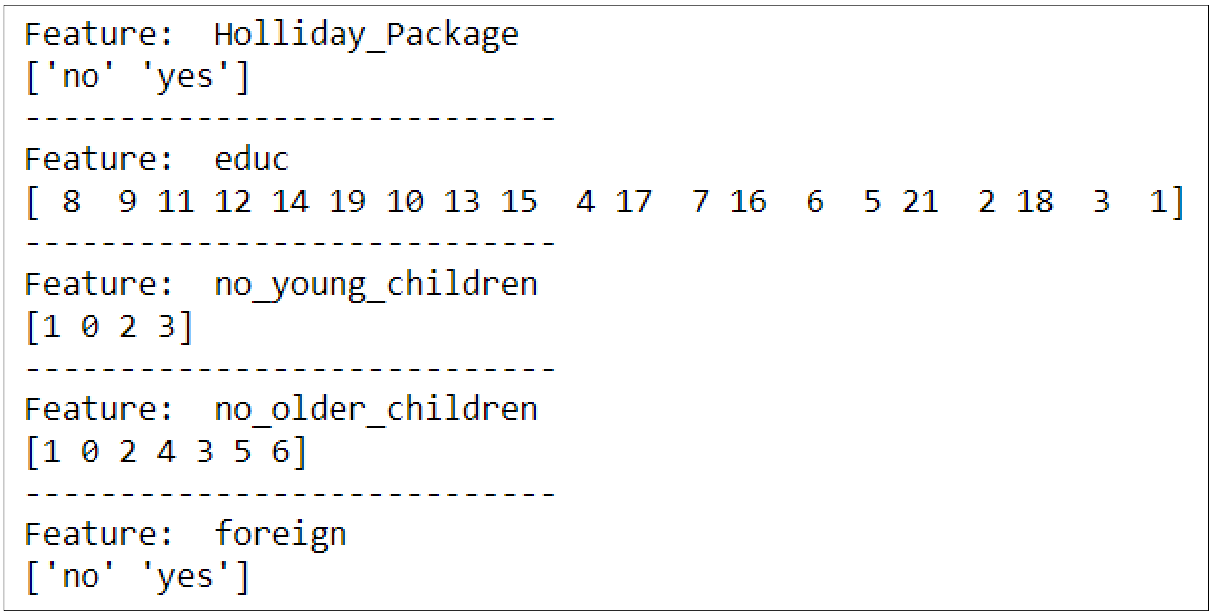


```
Feature:  Holliday_Package
no      471
yes     401
Name: Holliday_Package, dtype: int64
----------------------------------------
Feature:  educ
8     157
12    124
9     114
11    100
10     90
5      67
4      50
13     43
7      31
14     25
6      21
15     15
3      11
16     10
2       6
17      3
19      2
21      1
18      1
1       1
Name: educ, dtype: int64
```



```
Feature:  no_young_children
0    665
1    147
2     55
3      5
Name: no_young_children, dtype: int64
----------------------------------------
Feature:  no_older_children
0    393
2    208
1    198
3     55
4     14
6      2
5      2
Name: no_older_children, dtype: int64
----------------------------------------
Feature:  foreign
no      656
yes     216
Name: foreign, dtype: int64
```

Checking Unique Entries in Discrete Numerical and Categorical Variables



```
Feature:  Holliday_Package
['no' 'yes']
----------------------------
Feature:  educ
[ 8  9 11 12 14 19 10 13 15  4 17  7 16  6  5 21  2 18  3  1]
----------------------------
Feature:  no_young_children
[1 0 2 3]
----------------------------
Feature:  no_older_children
[1 0 2 4 3 5 6]
----------------------------
Feature:  foreign
['no' 'yes']
```

Insights:

1. There are no anomalies in the sublevels of discrete numerical and categorical features.

2. There are few sublevels in education, no. of young children and no. of older children features with negligible count.

Checking Outliers in Continuous Numerical Features:

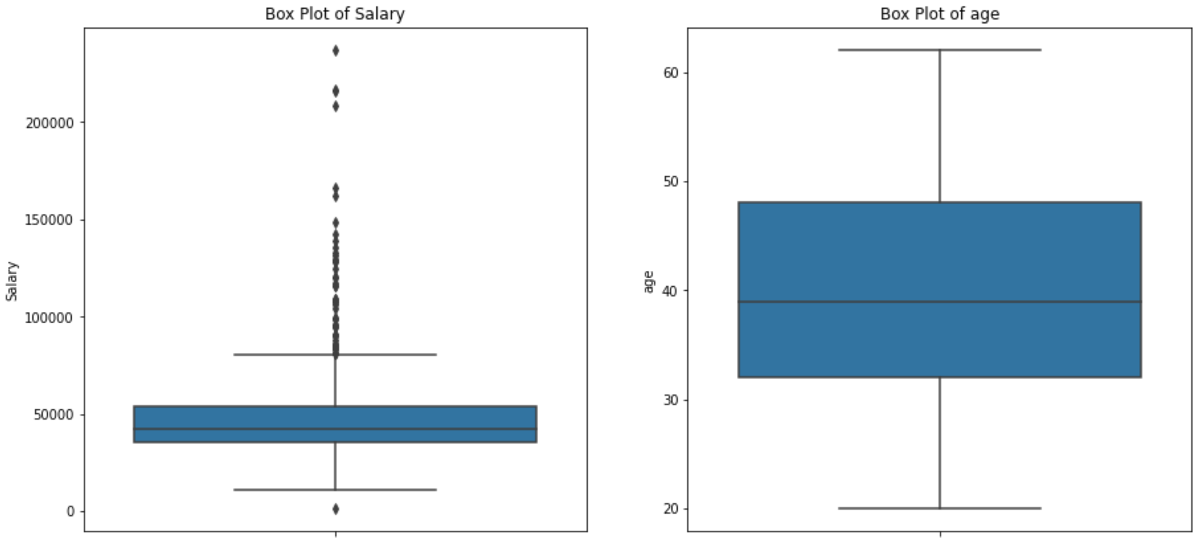
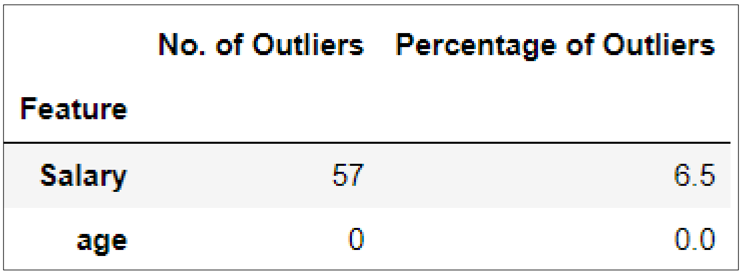


Figure 11. Box Plots of Continuous Numerical Features in Holiday Package Dataset.



	No. of Outliers	Percentage of Outliers
Feature		
Salary	57	6.5
age	0	0.0

Table 36. Number of Outliers and Percentage of Outliers in Holiday Package Dataset.

Insights:

1. There are 6.5% of outliers in salary feature.
2. As outliers' percentage is less, outliers in salary feature are not treated.

Univariate Analysis

Histogram and Boxplot of Continuous Numerical Features:

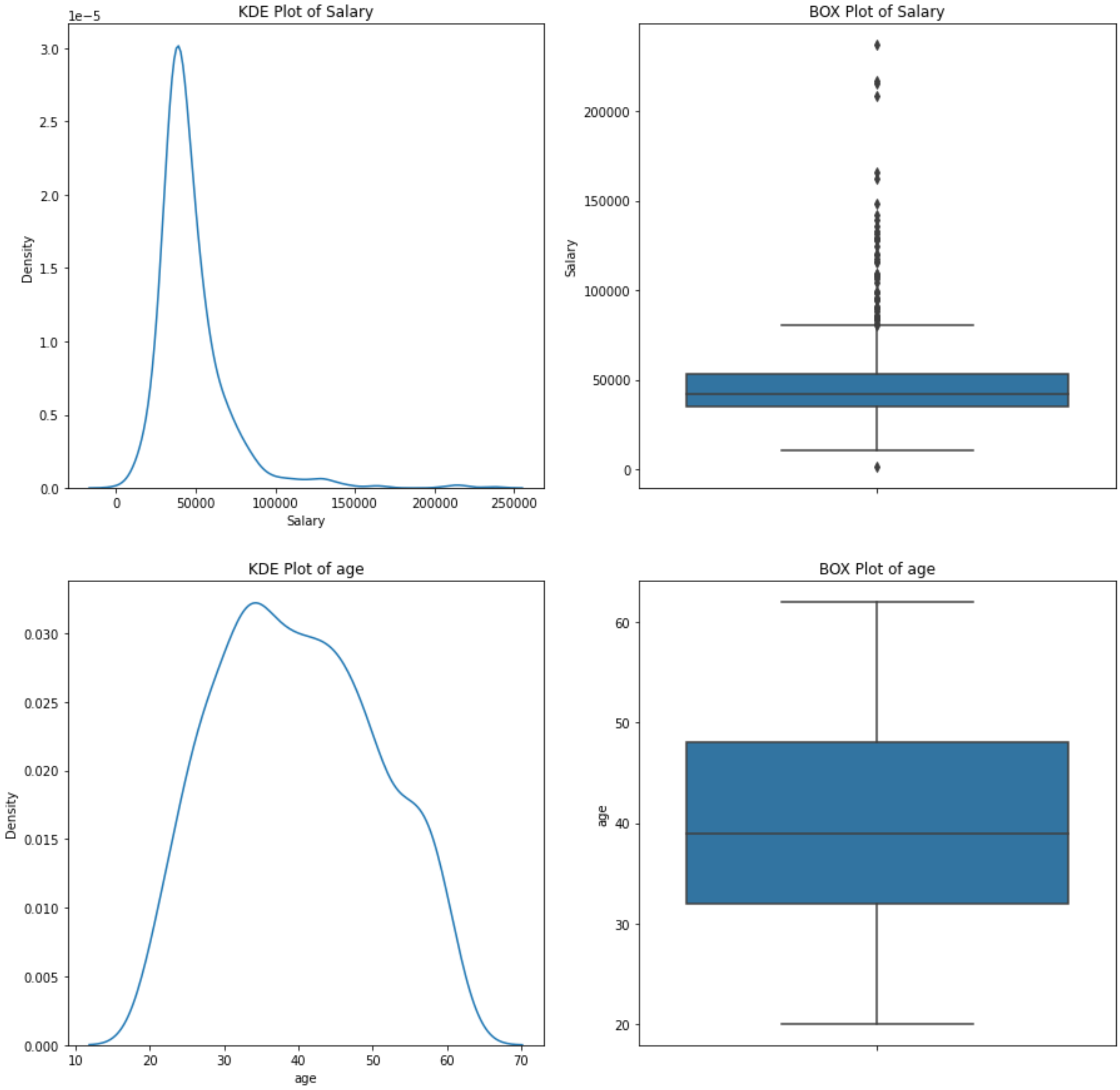
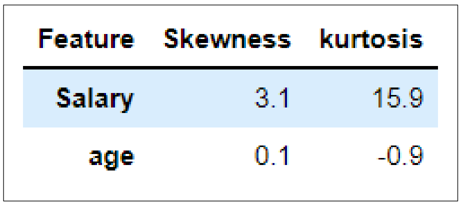


Figure 12. Histograms and Box Plots for Continuous Numerical Features.

Skewness & Kurtosis:

- Skewness is a measure of lack of symmetry in a distribution.
- Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.



Feature	Skewness	kurtosis
Salary	3.1	15.9
age	0.1	-0.9

Table 37. Skewness and Kurtosis of Numeric Features in Holiday Package Dataset.

Insights:

From above plots and tables, we can conclude below points,

1. Both salary and age are right skewed distributions (Positively skewed).
2. Salary feature has positive kurtosis.
3. Age feature has negative kurtosis.

Histogram and Boxplot of Numerical Features with Holliday Package as Hue

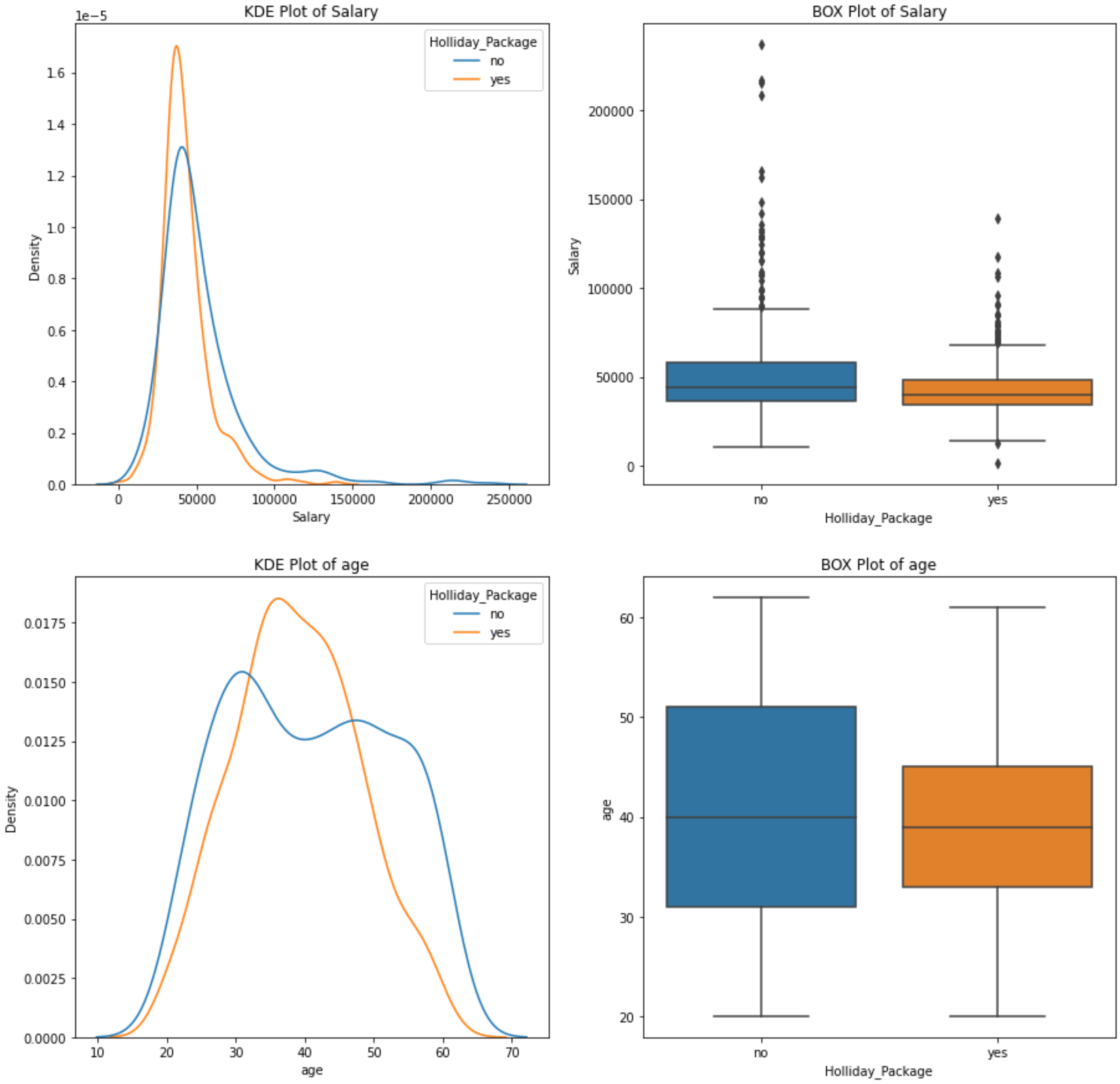
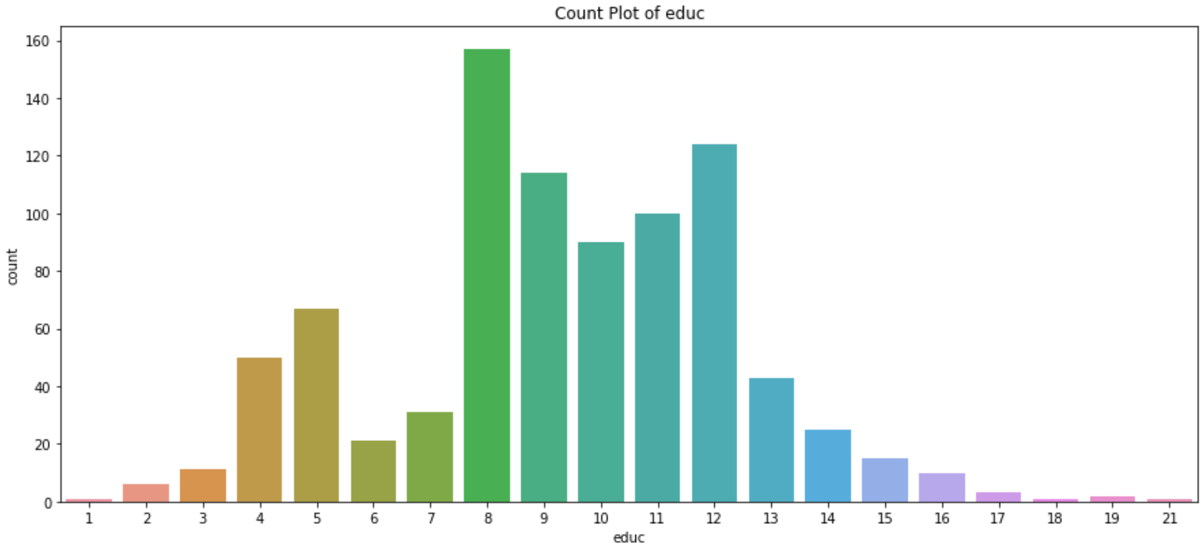


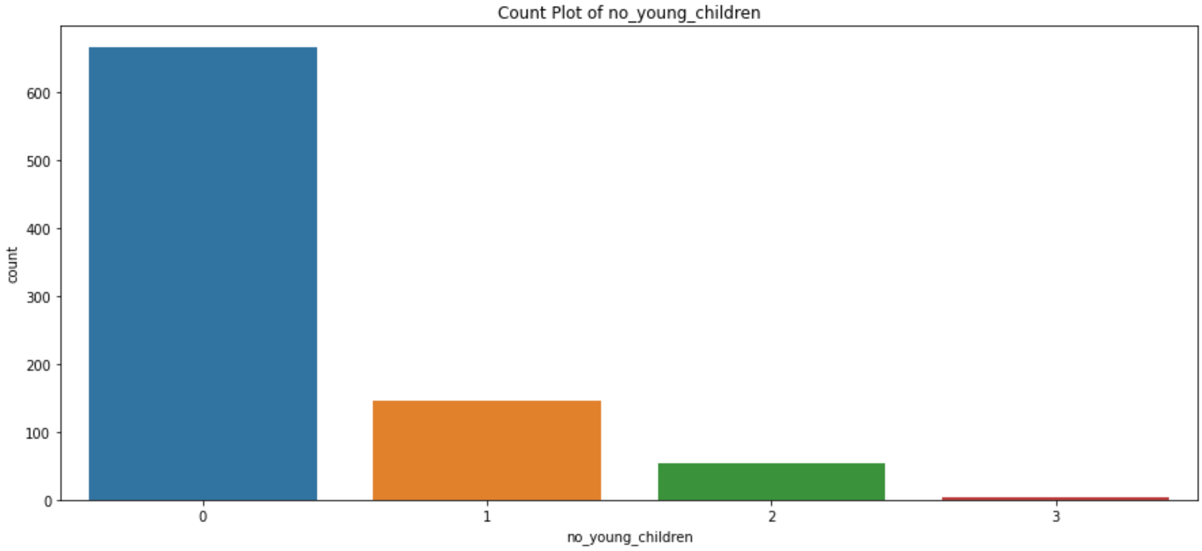
Figure 13. Histograms and Box Plots for Continuous Numerical Features with Holiday Package as Hue.

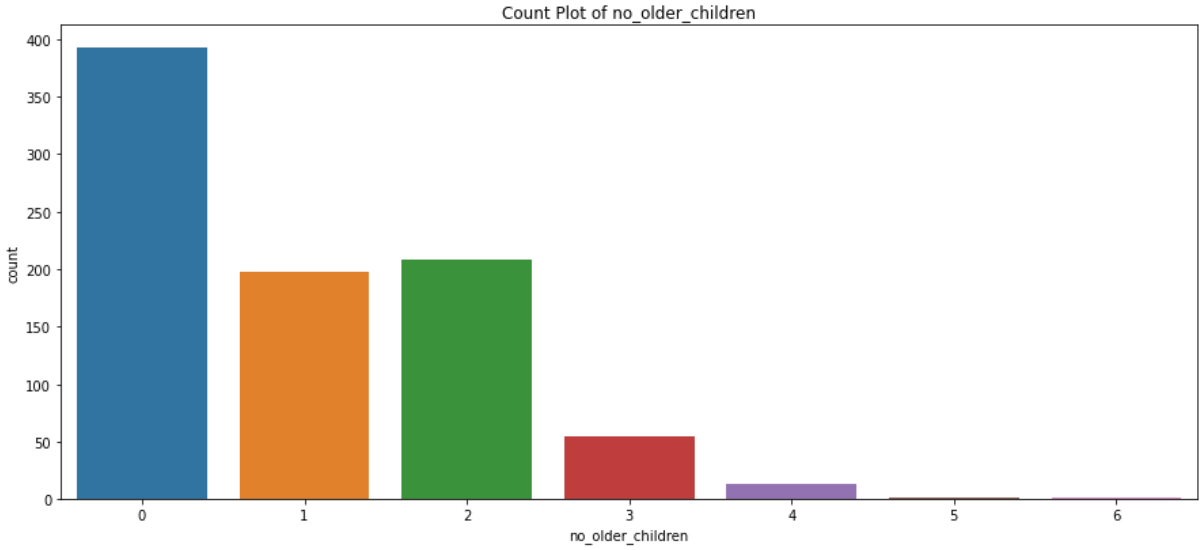
Inferences:

1. Both salary and age have similar distributions in both the classes of target feature.
2. Salary has outliers in both the classes of target feature.

Count Plots of Discrete Numerical and Categorical Features







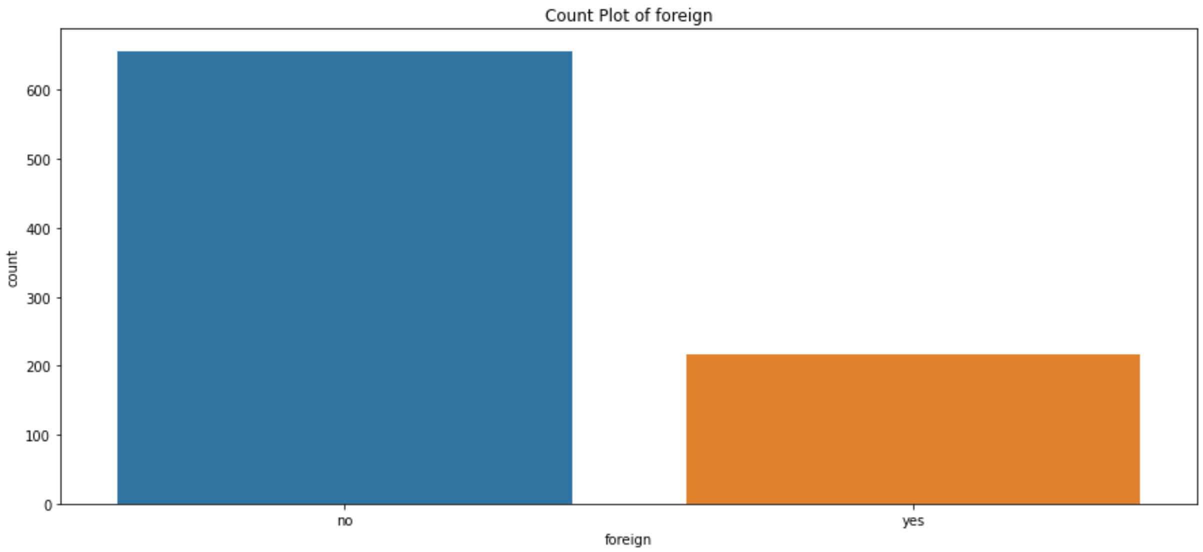


Figure 14. Count Plots of Discrete Numerical and Categorical Features

Insights:

From above count plots, we can conclude below points.

1. There are a greater number of employees with 8-12th education. There are a smaller number of employees with lower-level education and higher-level education.
2. Maximum number of employees with zero young children and minimum number of employees with three young children. Decreasing order of count as below.

   0 > 1 > 2 > 3
3. Maximum number of employees with zero older children and minimum number of employees with six older children. Decreasing order of count as below.

   0 > 2 > 1 > 3 > 4 > 5 > 6
4. There are a smaller number of foreign employees.

Bivariate Analysis between Numerical Features

Pair Plot of Numerical Features with Holiday Package as Hue

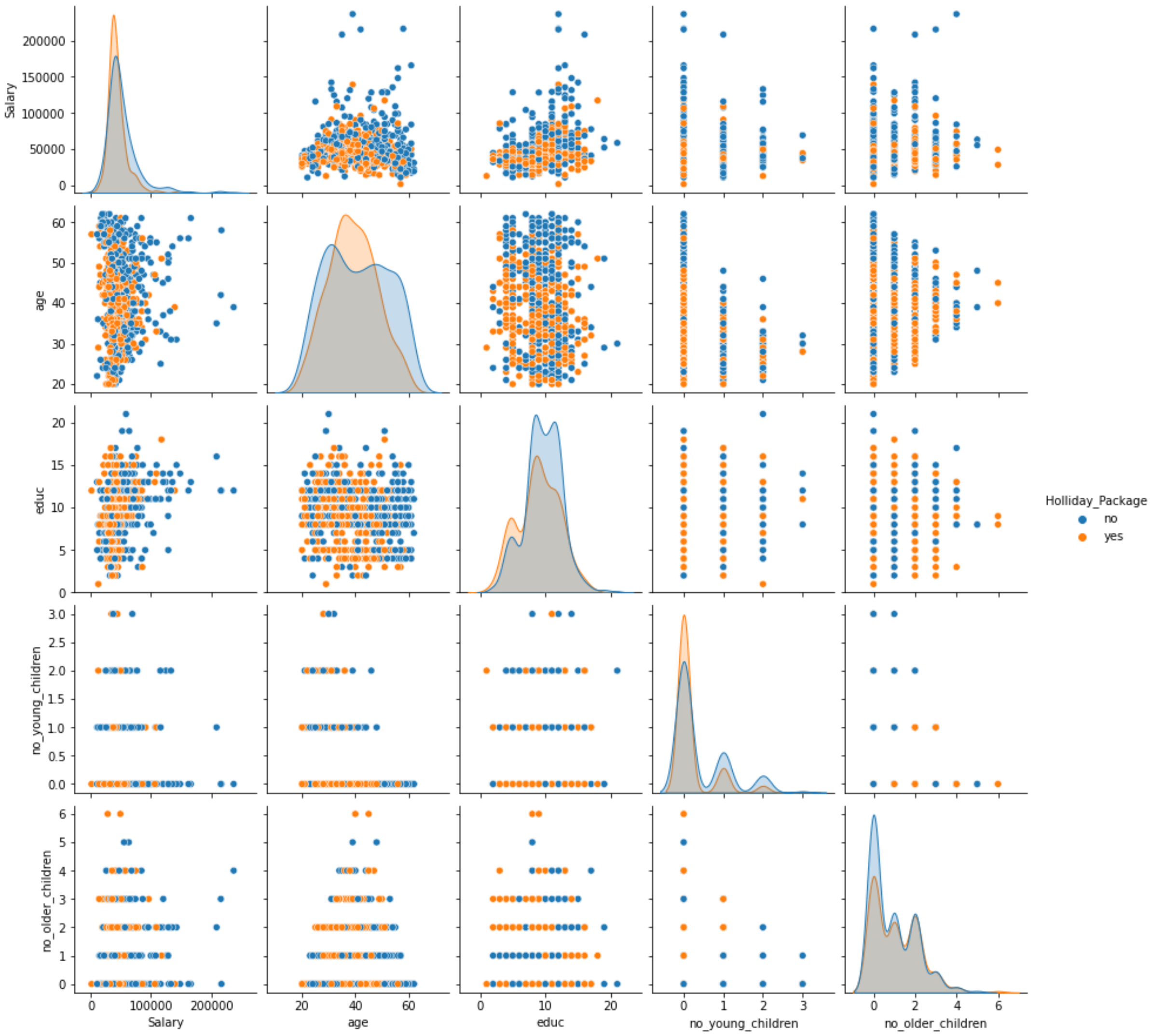


Figure 15. Pair Plot for Numeric Features in Holiday Package Dataset.

Heat Map

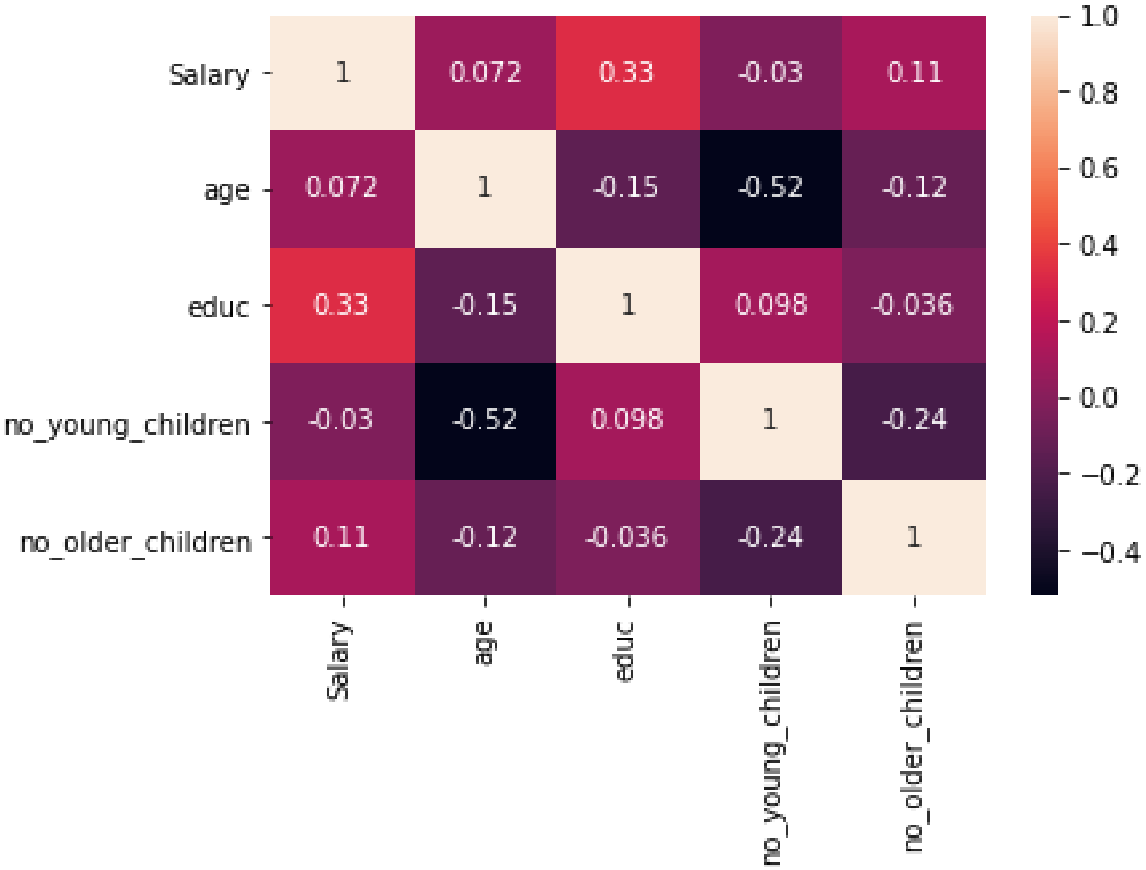


Figure 16. Heatmap for Numeric Features in Holiday Package Dataset.

Note:

From above Pair-Plot and Heatmap, it can be noticed that there is **no significant correlation** between the predictor variables.

Box Plots of Continuous Numerical Features Vs Holliday Package

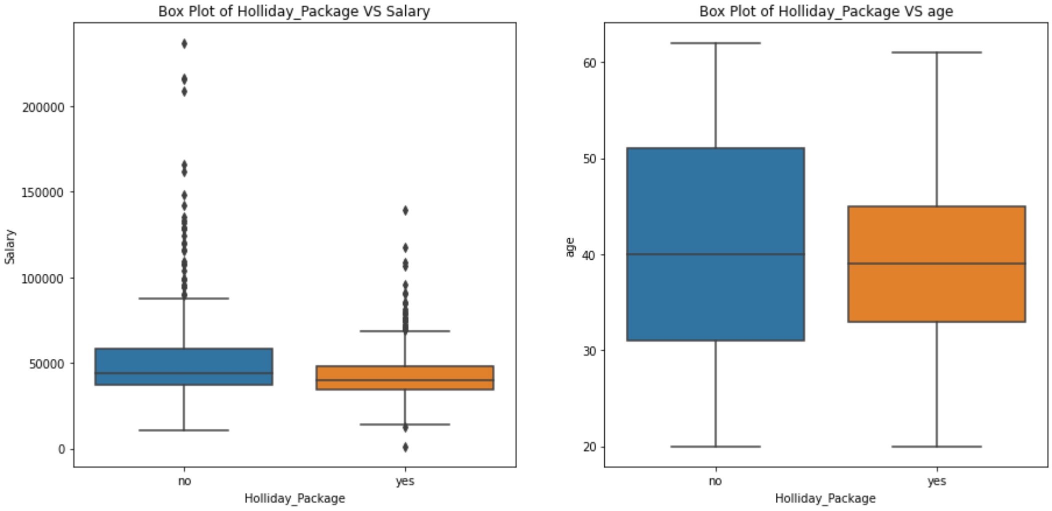


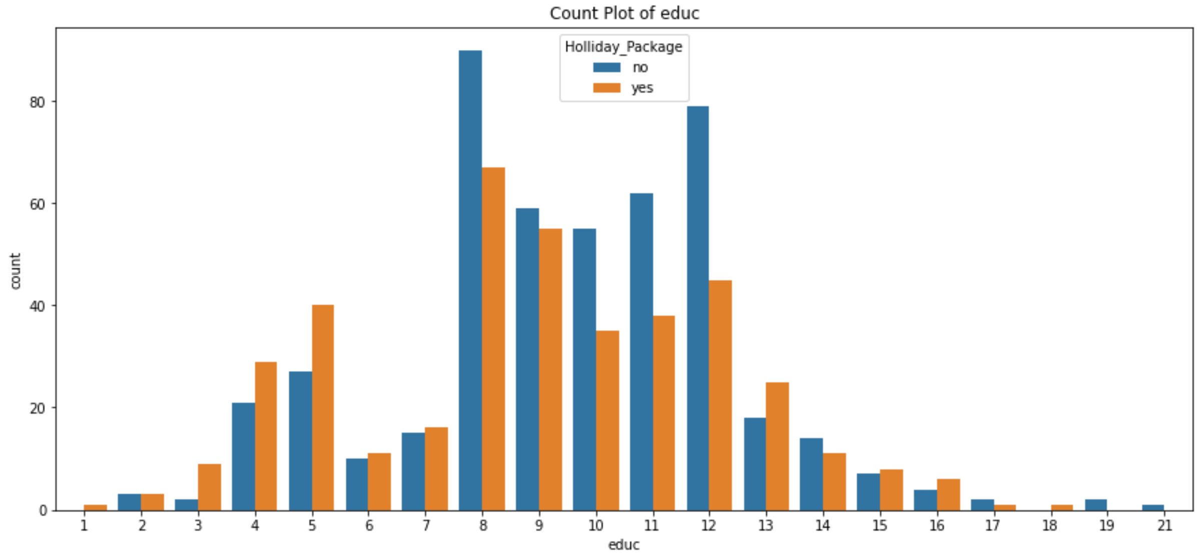
Figure 17. Box Plots of Continuous Numerical Features Vs Holliday Package

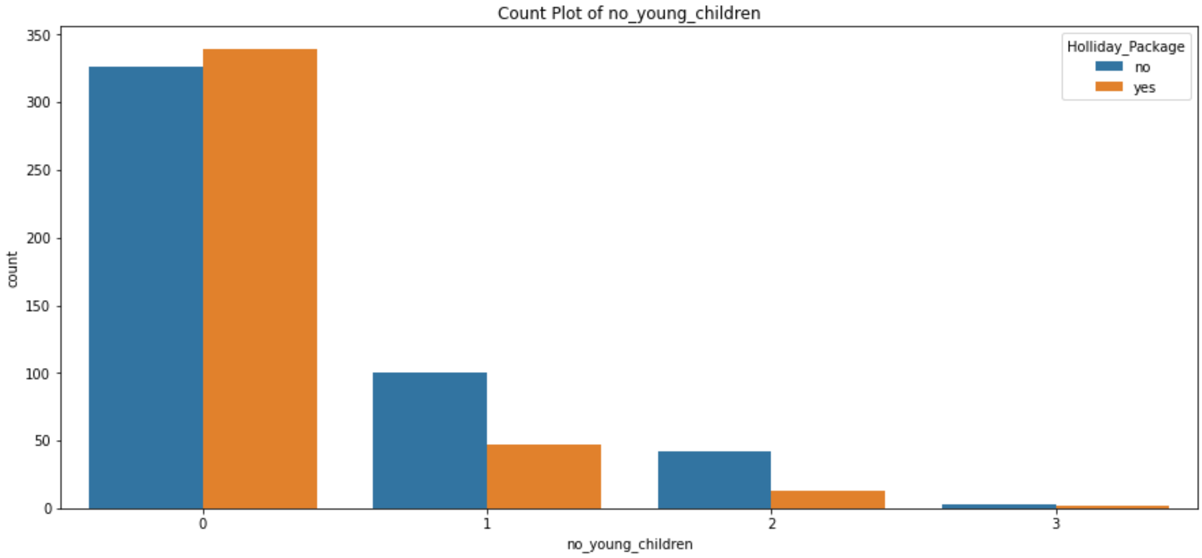
Insights:

From above bar plots, we can write below inferences,

1. Median salary of employees those who have taken holiday package is slightly less than that of employees not taken holiday package.
2. Median age of employees those who have taken holiday package is slightly less than that of employees not taken holiday package.

Count Plots of Discrete Numerical and Categorical Features with Holliday Package as Hue





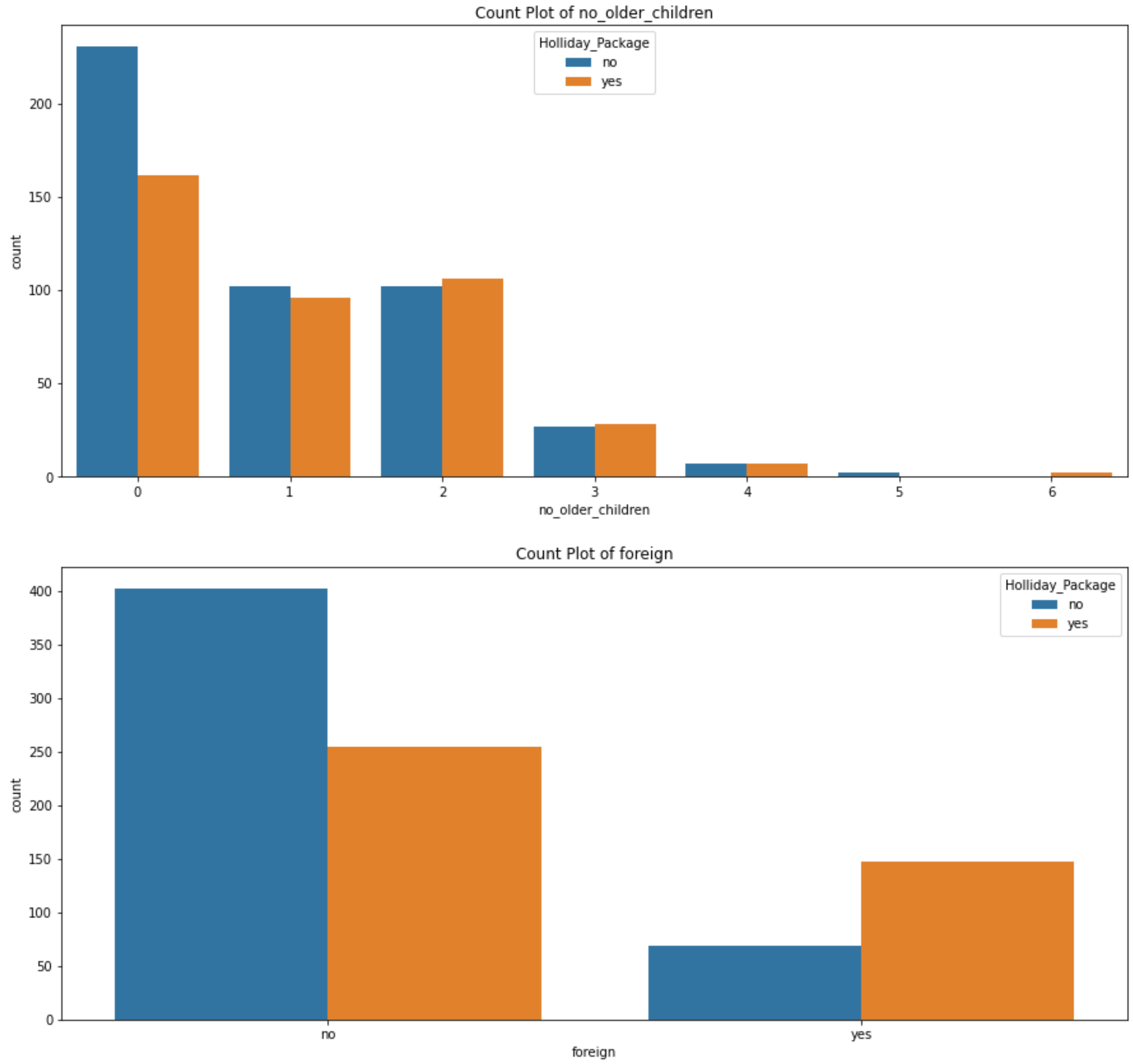


Figure 18. Count Plots of Discrete Numerical and Categorical Features with Holliday Package as Hue

Insights:

From above count plots, we can conclude below points.

1. The employees with both lower-level education and higher-level education are preferring to take holiday package more than employees with 8-12th education.
2. The employees with zero young children are preferring to take holiday package over employees with more young children.
3. The employees with more no. of older children are preferring to take holiday package over employees with less no. of older children.
4. Foreign employees are preferring to take holiday package over non-foreign employees.

Distribution of Classes in Target Feature

The percentage of employees taken holiday package is 46%

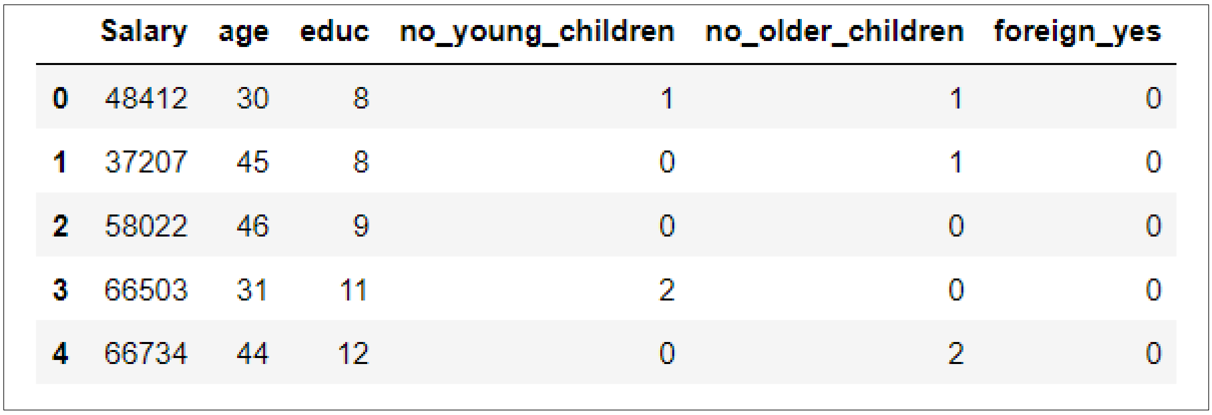
The percentage of employees not taken holiday package is 54%

The data is well balanced with the classes in target feature. We can proceed with model building process.

Q2.2. Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis).

Splitting the Dataset into Predictors and Target Variable

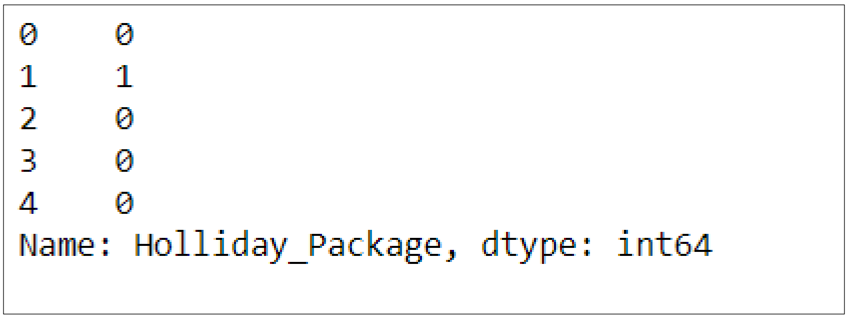
Sample of Predictors Dataset



	Salary	age	educ	no_young_children	no_older_children	foreign_yes
0	48412	30	8	1	1	0
1	37207	45	8	0	1	0
2	58022	46	9	0	0	0
3	66503	31	11	2	0	0
4	66734	44	12	0	2	0

Table 37. Sample of Predictors Dataset.

Sample of Target Feature



```
0    0
1    1
2    0
3    0
4    0
Name: Holliday_Package, dtype: int64
```

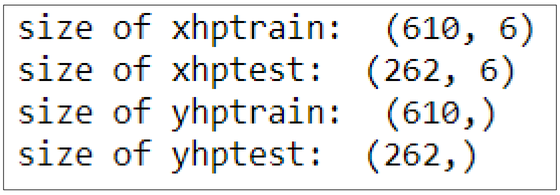
Table 38. Sample of Target Feature.

Splitting the Data into Train and Test Sets

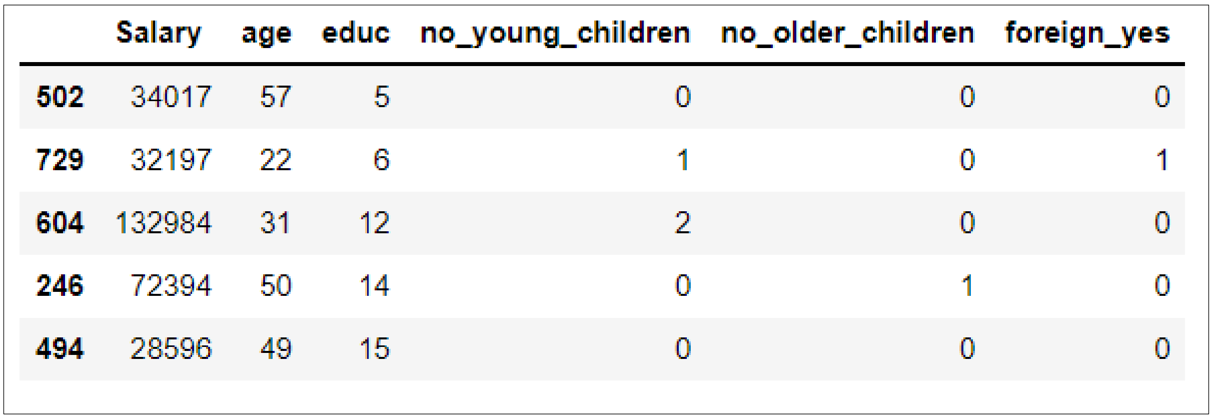
- Both independent and target datasets have been divided into train and test sets.
- No. of observations in test set is selected as 0.3 times of total data points.

- Then no. of observations in train set will be 0.7 times of total data points.

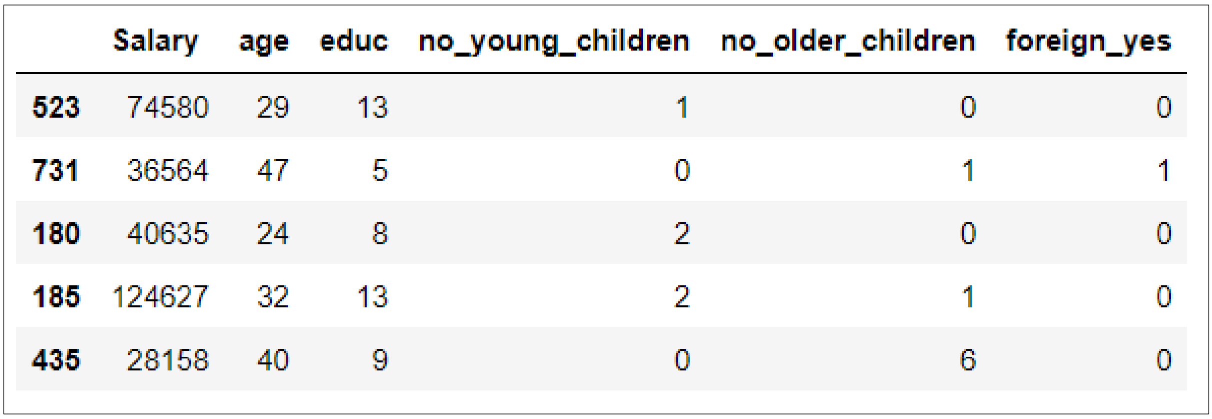
Checking the Training and Test Data:



```
size of xhptrain:  (610, 6)
size of xhptest:  (262, 6)
size of yhptrain:  (610,)
size of yhptest:  (262,)
```

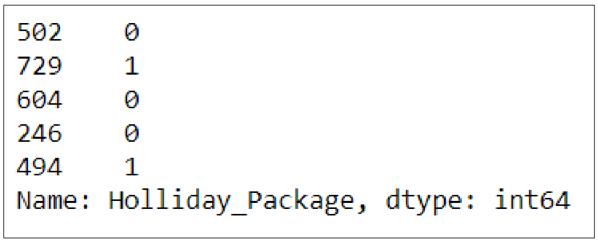


	Salary	age	educ	no_young_children	no_older_children	foreign_yes
502	34017	57	5	0	0	0
729	32197	22	6	1	0	1
604	132984	31	12	2	0	0
246	72394	50	14	0	1	0
494	28596	49	15	0	0	0

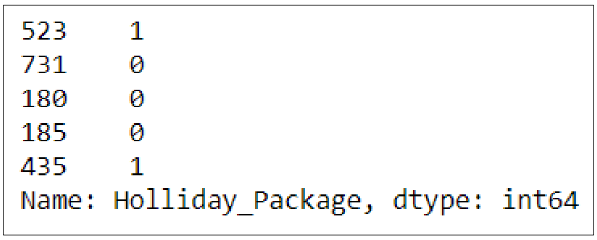


	Salary	age	educ	no_young_children	no_older_children	foreign_yes
523	74580	29	13	1	0	0
731	36564	47	5	0	1	1
180	40635	24	8	2	0	0
185	124627	32	13	2	1	0
435	28158	40	9	0	6	0

Table 39. Samples of Predictors Train and Predictors Test Datasets.



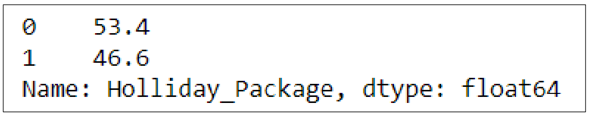
```
502    0
729    1
604    0
246    0
494    1
Name: Holliday_Package, dtype: int64
```



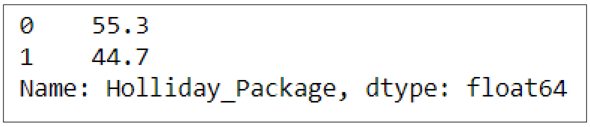
```
523    1
731    0
180    0
185    0
435    1
Name: Holliday_Package, dtype: int64
```

Table 40. Samples of Target Train and Target Test Data.

Distribution of Target Class in Train and Test sets



```
0    53.4
1    46.6
Name: Holliday_Package, dtype: float64
```



```
0    55.3
1    44.7
Name: Holliday_Package, dtype: float64
```

Table 41. Distribution of Target Class in Train and Test sets.

.

From above table, we can notice that target class (0s and 1s) is almost uniformly distributed between train and test datasets.

Building Logistic Regression Model

Below Hyper Parameters have been selected in Logistic Regression model to optimize by using GridSearchCV.

Maximum Iterations: [50,100,200]

Solver: [newton-cg, lbfgs, liglinear, sag, saga]

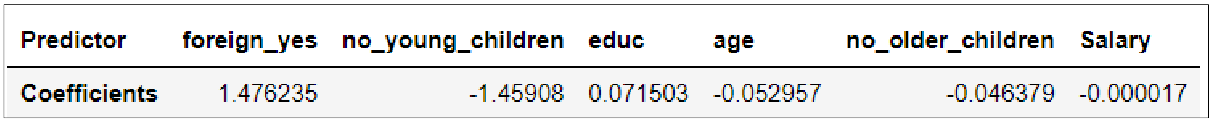
Best Parameters:

Below are the best parameters obtained in Logistic Regression model by using GridSearchCV

Maximum Iterations: 100

Solver: newton-cg

Coefficients:



Predictor	foreign_yes	no_young_children	educ	age	no_older_children	Salary
Coefficients	1.476235	-1.45908	0.071503	-0.052957	-0.046379	-0.000017

Table 42. Coefficients of Features in Logistic Regression Model.

From above table, we can notice that foreign and no. of young children are the most important features in classifying or predicting the class of target variable.

Building Linear Discriminant Analysis (LDA) Model

Below Hyper Parameters have been selected in LDA model to optimize by using GridSearchCV.

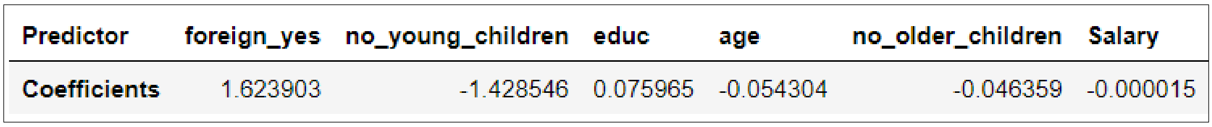
Solver: [svd, lsqr, eigen]

Best Parameters

Below are the best parameters obtained in LDA by using GridSearchCV

Solver: svd

Coefficients



Predictor	foreign_yes	no_young_children	educ	age	no_older_children	Salary
Coefficients	1.623903	-1.428546	0.075965	-0.054304	-0.046359	-0.000015

Coefficients of Features in LDA Model.

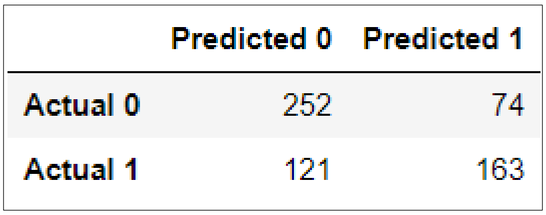
From above table, we can notice that foreign and no. of young children are the most important features in classifying or predicting the class of target variable.

Q2.3. Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.

Logistic Regression Model Evaluation

Model Evaluation Based on Train Set:

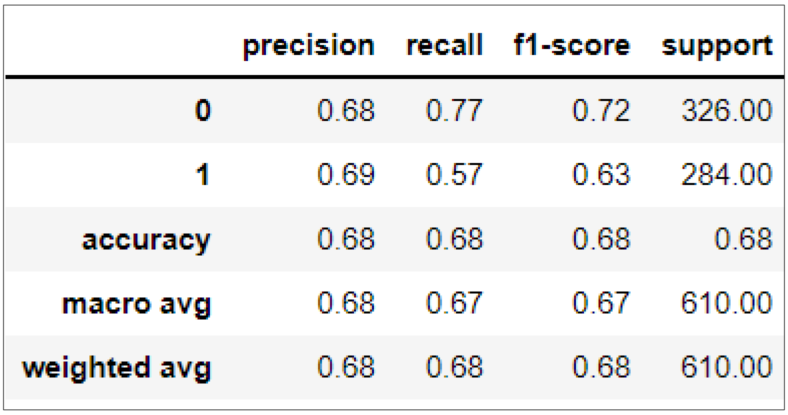
Confusion Matrix:



	Predicted 0	Predicted 1
Actual 0	252	74
Actual 1	121	163

Table 43. Confusion Matrix for Train Dataset in Logistic Regression model.

Classification Report:



	precision	recall	f1-score	support
0	0.68	0.77	0.72	326.00
1	0.69	0.57	0.63	284.00
accuracy	0.68	0.68	0.68	0.68
macro avg	0.68	0.67	0.67	610.00
weighted avg	0.68	0.68	0.68	610.00

Table 44. Classification Report for Train Dataset in Logistic Regression model.

Accuracy:

Accuracy of the model is 0.68

ROC AUC Score:

ROC AUC Score of the model is 0.74

ROC Curve:

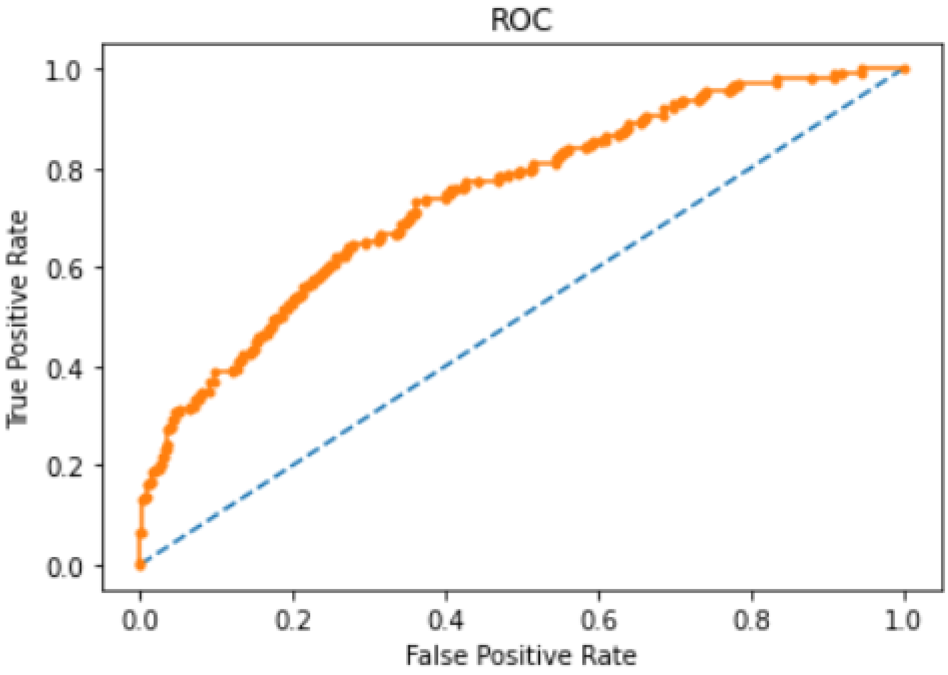
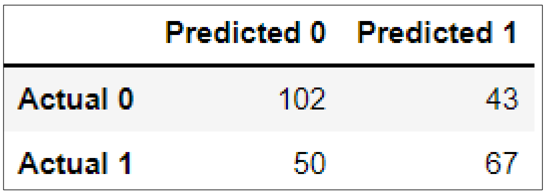


Figure 19. ROC Curve for Train Dataset in Logistic Regression model.

Model Evaluation Based on Test Set

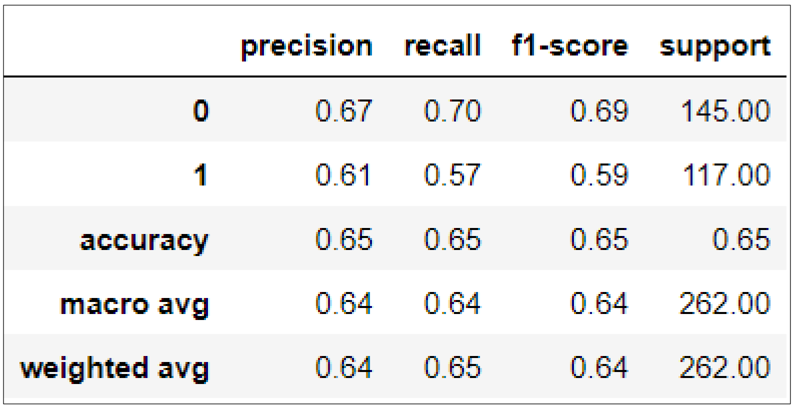
Confusion Matrix:



	Predicted 0	Predicted 1
Actual 0	102	43
Actual 1	50	67

Table 45. Confusion Matrix for Test Dataset in Logistic Regression model.

Classification Report:



	precision	recall	f1-score	support
0	0.67	0.70	0.69	145.00
1	0.61	0.57	0.59	117.00
accuracy	0.65	0.65	0.65	0.65
macro avg	0.64	0.64	0.64	262.00
weighted avg	0.64	0.65	0.64	262.00

Table 46. Classification Report for Test Dataset in Logistic Regression model.

Accuracy:

Accuracy of the model is 0.65

ROC AUC Score:

ROC AUC Score of the model is 0.7

ROC Curve:

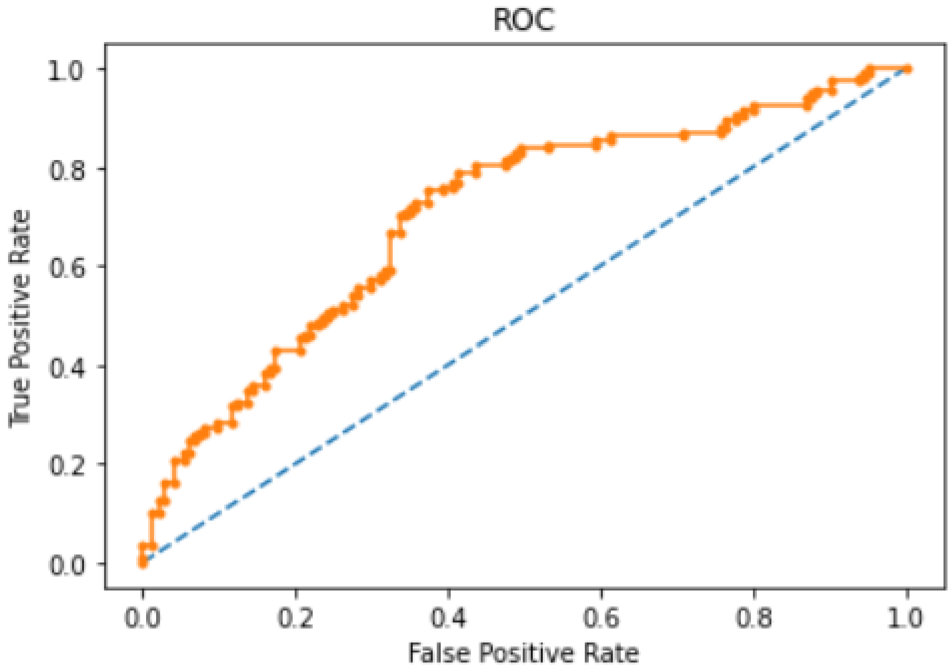


Figure 20. ROC Curve for Test Dataset in Logistic Regression model.

Conclusions:

Data	Random Forest Model				
	Precision for Class 1	**Recall for Class 1**	**F1 Score for Class 1**	**AUC for Class 1**	**Accuracy**
Train Dataset	0.69	0.57	0.63	0.74	0.68
Test Dataset	0.61	0.57	0.59	0.7	0.65

Table 47. Performance Metrics of Logistic Regression model

From above table, we can conclude below points.

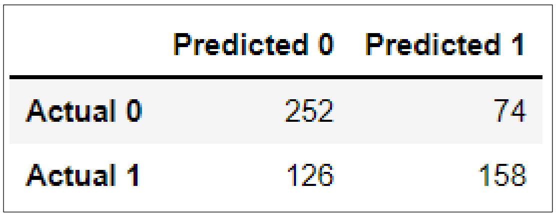
1. The performance metrics like precision, recall, F1 score, AUC Score and Accuracy for test dataset are approaching train dataset. Hence, there is no over fitting in the model.

2. Accuracy of the model on test dataset (0.65) is less than 0.75. Hence, the model is considered can reviewed with the business to improve accuracy.
3. ROC AUC Score of the model on test dataset (0.7) is little low. We should discuss with the business to improve this metric.
4. Precision (0.61), Recall (0.57) and F1 score (0.59) on test dataset are not upto the mark. We should discuss with the business to improve these metrics before considering for predictions.

Linear Discriminant Analysis (LDA) Model Evaluation

Model Evaluation Based on Train Set:

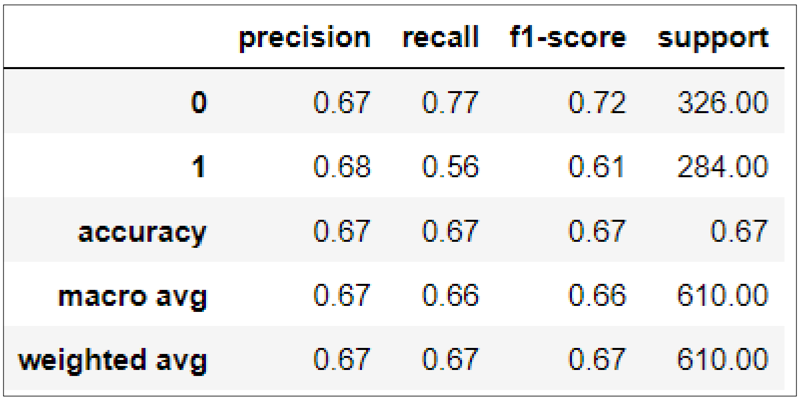
Confusion Matrix:



	Predicted 0	Predicted 1
Actual 0	252	74
Actual 1	126	158

Table 48. Confusion Matrix for Train Dataset in LDA model.

Classification Report:



	precision	recall	f1-score	support
0	0.67	0.77	0.72	326.00
1	0.68	0.56	0.61	284.00
accuracy	0.67	0.67	0.67	0.67
macro avg	0.67	0.66	0.66	610.00
weighted avg	0.67	0.67	0.67	610.00

Table 49. Classification Report for Train Dataset in LDA model.

Accuracy:

Accuracy of the model is 0.67

ROC AUC Score:

ROC AUC Score of the model is 0.74

ROC Curve:

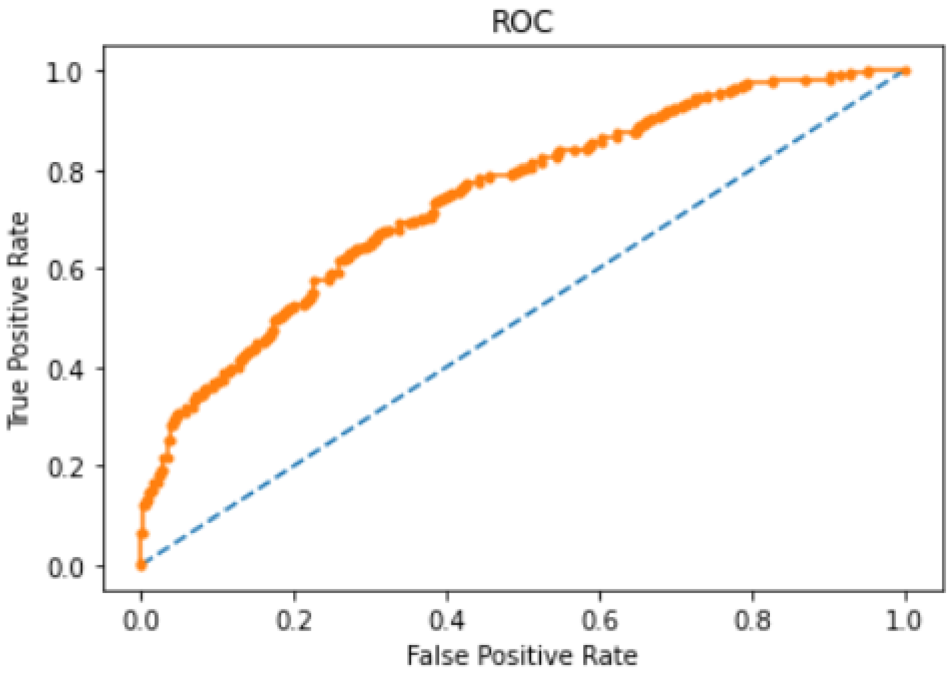
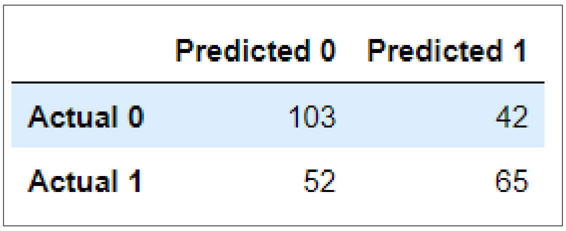


Figure 21. ROC Curve for Train Dataset in LDA model.

Model Evaluation Based on Test Set

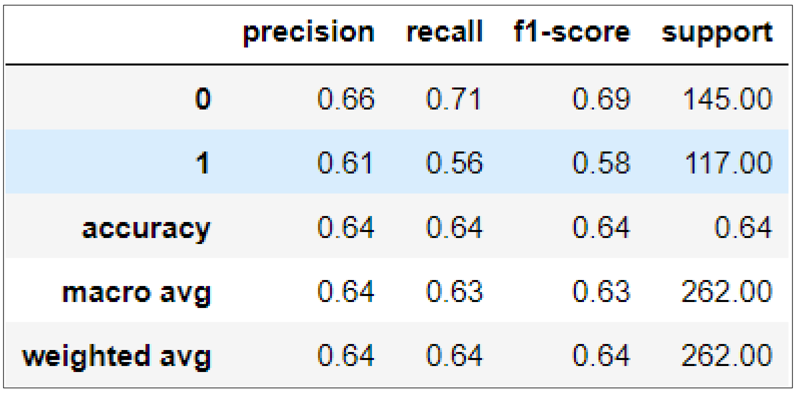
Confusion Matrix:



	Predicted 0	Predicted 1
Actual 0	103	42
Actual 1	52	65

Table 50. Confusion Matrix for Test Dataset in LDA model.

Classification Report:



	precision	recall	f1-score	support
0	0.66	0.71	0.69	145.00
1	0.61	0.56	0.58	117.00
accuracy	0.64	0.64	0.64	0.64
macro avg	0.64	0.63	0.63	262.00
weighted avg	0.64	0.64	0.64	262.00

Table 51. Classification Report for Test Dataset in LDA model.

Accuracy:

Accuracy of the model is 0.64

ROC AUC Score:

ROC AUC Score of the model is 0.7

ROC Curve:

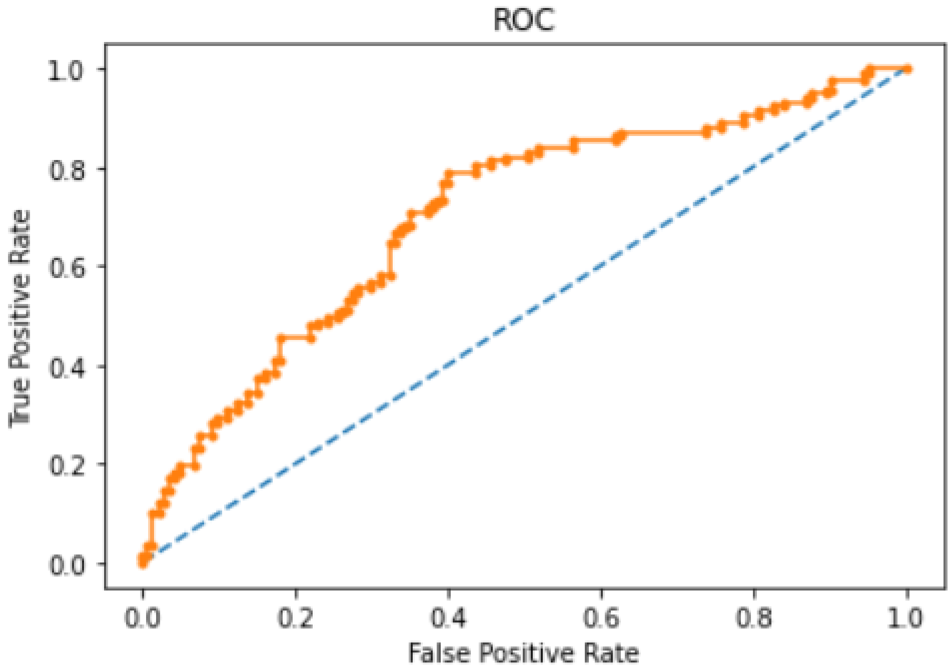


Figure 22. ROC Curve for Test Dataset in LDA model.

Conclusions:

Data	Artificial Neural Network Model				
	Precision for Class 1	Recall for Class 1	F1 Score for Class 1	AUC for Class 1	Accuracy
Train Dataset	0.68	0.56	0.61	0.74	0.67
Test Dataset	0.61	0.56	0.58	0.7	0.64

Table 52. Performance Metrics of LDA model

From above table, we can conclude below points.

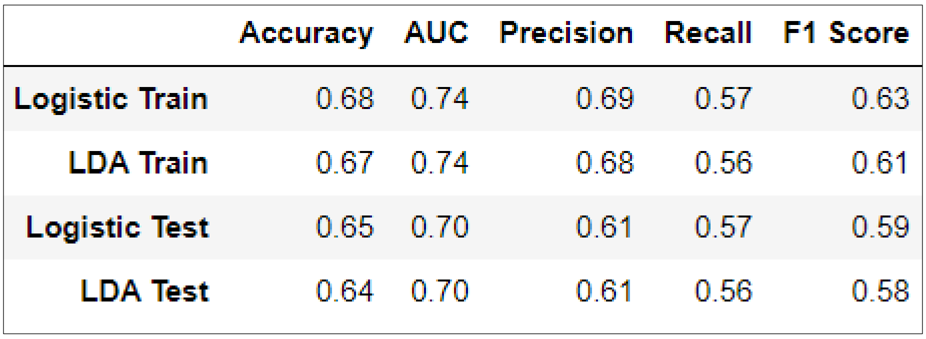
1. The performance metrics like precision, recall, F1 score, AUC Score and Accuracy for test dataset are approaching train dataset. Hence, there is no over fitting in the model.

2. Accuracy of the model on test dataset (0.64) is less than 0.75. Hence, the model is considered can reviewed with the business to improve accuracy.
3. ROC AUC Score of the model on test dataset (0.7) is little low. We should discuss with the business to improve this metric.
4. Precision (0.61), Recall (0.56) and F1 score (0.58) on test dataset are not upto the mark. We should discuss with the business to improve these metrics before considering for predictions.

Importance of Performance Metrics:

✓ F1 score is a harmonic mean of recall and precision and it is more important than recall and precision. Because F1 score takes care of both of them. If either recall or precision decreases, F1 score automatically decreases drastically. Hence, F1 score is most important metric in evaluating model performance and deciding validity of the model.

Comparison of Logistic Regression and LDA Models



	Accuracy	AUC	Precision	Recall	F1 Score
Logistic Train	0.68	0.74	0.69	0.57	0.63
LDA Train	0.67	0.74	0.68	0.56	0.61
Logistic Test	0.65	0.70	0.61	0.57	0.59
LDA Test	0.64	0.70	0.61	0.56	0.58

Table 53. Comparison of Performance Metrics of two models.

Inferences:

From above table and below ROC plots, we can derive below inferences.

1. For all the two models, the performance metrics of train dataset are approaching to the test dataset. Hence, there is no over fitting in any one of these models.
2. The performance metrics of test datasets for all the two models is almost equal but they are not significant enough to use for predictions. We shall discuss with business with the business to improve the performance metrics. Both the models are performing equally good, we can select anyone of them.

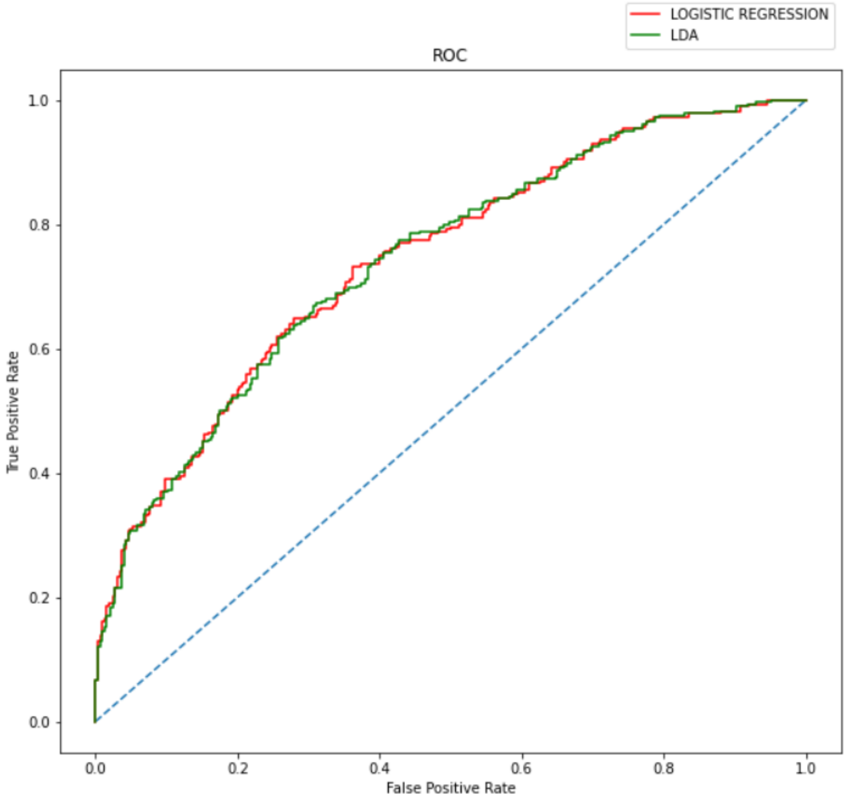


Figure 23. Comparison of ROC Curves for two models on Train Dataset

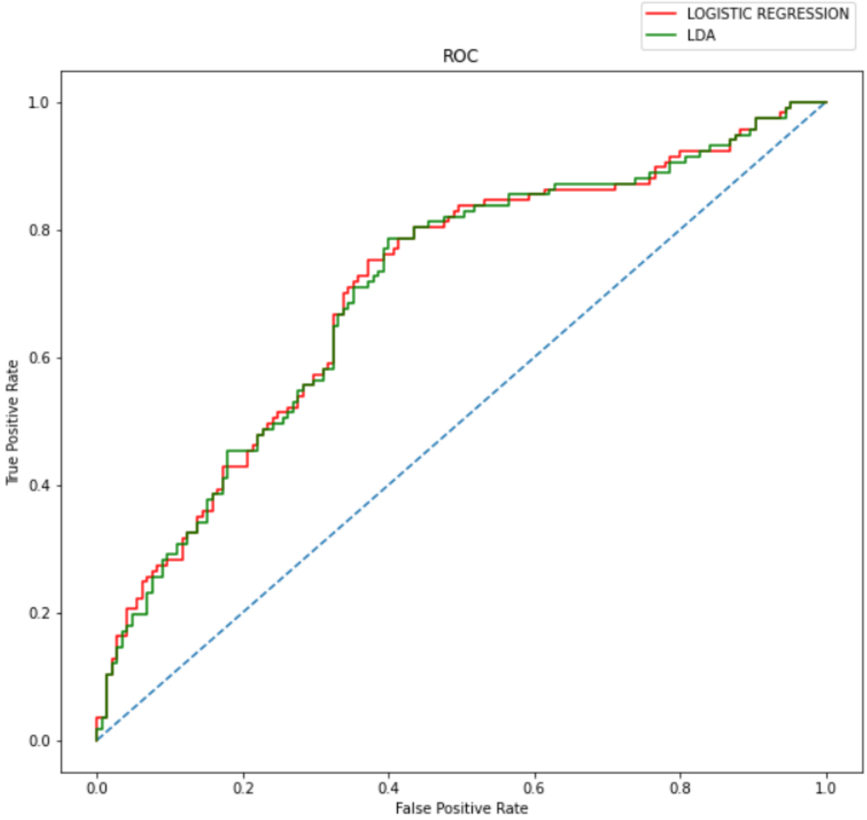


Figure 24. Comparison of ROC Curves for two models on Test Dataset

Q2.4. Inference: Basis on these predictions, what are the insights and recommendations.

(Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.)

Steps Performed

The following are the steps perfoermed in this project

1. Anamolies, null values and outlies in the dataset are checked in EDA section.
2. Data set has beeen slitted into train and test ration in the ratio of 70:30
3. Prediction models are built by using logistic regresion and linear discriminant analysis.
4. Models are evaluated on train and test datasets by using suitable performance metrics.

Business Interpretations and Actionable Insights

1. For all the two models, the performance metrics of train dataset are approaching to the test dataset. Hence, there is no over fitting in any one of these models.
2. The performance metrics of test datasets for all the two models is almost equal but they are not significant enough to use for predictions. We shall discuss with business with the business to improve the performance metrics. Both the models are performing equally good, we can select anyone of them.
3. The employees with both lower-level education and higher-level education are preferring to take holiday package more than employees with 8-12th education. We can focus on that particular group of people to increase sales of holiday packages.
4. The employees with zero young children are preferring to take holiday package over employees with more young children. We can focus on employees with zero young children to increase sales of holiday packages.
5. The employees with more no. of older children are preferring to take holiday package over employees with less no. of older children. We can focus on employees with more no. of older children to increase sales of holiday packages.
6. Foreign employees are preferring to take holiday package over non-foreign employees. We can focus on foreign employees to increase sales of holiday packages.