Predictive Modeling - Business Report

Rohan R. Khade

Table of Contents

Chapter 1. Problem 1: Linear Regression - 6 -

1.1 Problem Statement - 6 -

1.2 Introduction - 6 -

1.2.1 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. - 6 -

1.2.1.1 Univariate Analysis - 8 -

1.2.1.2 Bivariate & Multivariate Analysis - 11 -

1.2.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Check for the possibility of combining the sub levels of an ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning. - 16 -

1.2.3 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 chck 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. - 19 -

1.2.4 Inference: Basis on these predictions, what are the business insights and recommendations. - 24 -

Chapter 2. Problem 2: Logistic Regression and LDA - 25 -

2.1 Problem Statement - 25 -

2.2 Introduction - 25 -

2.2.1 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. - 25 -

2.2.1.1 Univariate Analysis - 27 -

2.2.1.2 Bivariate and Multivariate Analysis - 30 -

2.2.2 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). - 35 -

2.2.3 Performance Metrics: Comment and Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score, classification reports for each model. - 36 -

2.2.3.1 Logistic Regression - 36 -

2.2.3.1.1 Classification Report - 38 -

2.2.3.1.2 Confusion Matrix - 38 -

2.2.3.1.3 Classification Report (Using Gridsearch with best parameters) .... - 41 -
2.2.3.1.4 Confusion Matrix (Using Gridsearch with best parameters)........ - 41 -
2.2.3.2 Linear Discriminant Analysis (LDA) .... - 42 -
2.2.3.2.1 Classification Report .... - 43 -
2.2.4 Inference: Basis on these predictions, what are the insights and recommendations. .... - 45 -

List of Tables

Table 1 Dataframe: df (with head function) .... - 6 -
Table 2 Dataframe: df (with describe with inclue all function) .... - 7 -
Table 3 Dataframe: df (with head function) .... - 16 -
Table 4 X_train dataset (with head function) .... - 19 -
Table 5 X_test dataset (with head function) .... - 20 -
Table 6 Dataframe: df1 (with head function) .... - 25 -
Table 7 Dataframe: df1 (with describe function) .... - 26 -
Table 8 Comparative Analysis for Three Models .... - 44 -

List of Figures

Figure 1. Dataset information .... - 7 -
Figure 2. Carat data series: Description & graphical representation .... - 8 -
Figure 3. Depth data series: Description & graphical representation .... - 8 -
Figure 4. Table data series: Description & graphical representation .... - 9 -
Figure 5. x data series: Description & graphical representation .... - 9 -
Figure 6. y data series: Description & graphical representation .... - 9 -
Figure 7. z data series: Description & graphical representation .... - 10 -
Figure 8. Price data series: Description & graphical representation .... - 10 -
Figure 9. Pie chart analysis of the three variables with object data type .... - 11 -
Figure 10. Correlation matrix of the seven variables with integer data type .... - 11 -
Figure 11. Pairplot........ - 12 -
Figure 12. Scatterplot for Carat & Price variable with Cut as hue .... - 13 -

Figure 13. Bivariate analysis for categorical variable: Cut - 14 -
Figure 14. Bivariate analysis for categorical variable: Clarity - 14 -
Figure 15. Bivariate analysis for categorical variable: Color - 14 -
Figure 16. Counts for Categorical Variables - 15 -
Figure 17. Box plot with outliers - 17 -
Figure 18. Box plot post outlier treatment - 18 -
Figure 19. Coefficient Analysis: Train dataset - 20 -
Figure 20. Intercept Calculation - 20 -
Figure 21. Model train dataset score - 20 -
Figure 22. Model test dataset score - 20 -
Figure 23. OLS params - 21 -
Figure 24. OLS Regression Results - 21 -
Figure 25. Scatterplot for Test Data - 22 -
Figure 26. Coefficient Analysis: Scaled train dataset - 22 -
Figure 27. Intercept Calculation on Scaled Data - 22 -
Figure 28. Mean Squared Error: Scaled data - 22 -
Figure 29. Scatterplot for Scaled Test Data - 23 -
Figure 30. VIF Calculations - 23 -
Figure 31. Dataset information - 26 -
Figure 32. Holliday Package - 27 -
Figure 33. Foreign - 27 -
Figure 34. Swarmplot with salary and age as hue for holliday package - 28 -
Figure 35. Count plot for educ with holliday package as hue - 28 -
Figure 36. Count plot for no_young_children with holliday package as hue - 29 -
Figure 37. Count plot for no_older_children with holliday package as hue - 29 -
Figure 38. Pairplot (Claimed variable as hue) - 30 -
Figure 39. VIF Calculation - 31 -
Figure 40. Box Plot: Salary (without outlier treatment) - 31 -
Figure 41. Box Plot: Age (without outlier treatment) - 31 -
Figure 42. Box Plot: Educ (without outlier treatment) - 32 -
Figure 43. Box Plot: no_younger_children (without outlier treatment) - 32 -
Figure 44. Box Plot: no_older_children (without outlier treatment) - 33 -

Back to Top

Figure 45. Box Plot: Salary (with outlier treatment) - 33 -
Figure 46. Correlation Matrix - 34 -
Figure 47. Correlation Matrix Heatmap - 34 -
Figure 48. Label Encoding - 35 -
Figure 49. Dummy Encoding - 35 -
Figure 50. X, y Split dataset (Train and Test) - 35 -
Figure 51. Logistic Regression Model - 35 -
Figure 52. Linear Discriminant Analysis (LDA) - 36 -
Figure 53. X train and y train model score - 36 -
Figure 54. X test and y test model score - 36 -
Figure 55. ytest_predict_prob - 36 -
Figure 56. Train label ROC curve - 37 -
Figure 57. Test label ROC curve - 37 -
Figure 58. Train label - 38 -
Figure 59. Test label - 39 -
Figure 60. X train and y train model score (Using Gridsearch with best parameters) - 39 -
Figure 61. X test and y test model score (Using Gridsearch with best parameters) - 39 -
Figure 62. ytest_predict_prob (Using Gridsearch with best parameters) - 39 -
Figure 63. Train label ROC curve (Using Gridsearch with best parameters) - 40 -
Figure 64. Test label ROC curve (Using Gridsearch with best parameters) - 40 -
Figure 65. Train label - 41 -
Figure 66. Test label - 42 -
Figure 67. Train label ROC curve - 42 -
Figure 68. Test label ROC curve - 43 -

Chapter 1. Problem 1: Linear Regression

1.1 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.

1.2 Introduction

The dataset has 26,967 rows and 11 columns (However, as we have dropped the unnamed column, we have only 10 columns left). The columns of the dataset include spending, advance payments, probability of full payment, current balance, credit limit, minimum (min) payment amount (amt), and maximum (max) spent in single shopping. The dataset provides a list of customers surveyed to understand the best promotional offer that can be offered by the bank to them. We have 34 duplicte values as part of the data set and we have dropped them as they are repeat values included as part of the data set.

1.2.1 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.

Table 1 Dataframe: df (with head function)

	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

Back to Top

Table 2 Dataframe: df (with describe with inclue all function)

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
carat	26967.0	NaN	NaN	NaN	0.798375	0.477745	0.2	0.4	0.7	1.05	4.5
cut	26967	5	Ideal	10816	NaN	NaN	NaN	NaN	NaN	NaN	NaN
color	26967	7	G	5661	NaN	NaN	NaN	NaN	NaN	NaN	NaN
clarity	26967	8	SI1	6571	NaN	NaN	NaN	NaN	NaN	NaN	NaN
depth	26270.0	NaN	NaN	NaN	61.745147	1.41286	50.8	61.0	61.8	62.5	73.6
table	26967.0	NaN	NaN	NaN	57.45608	2.232068	49.0	56.0	57.0	59.0	79.0
x	26967.0	NaN	NaN	NaN	5.729854	1.128516	0.0	4.71	5.69	6.55	10.23
y	26967.0	NaN	NaN	NaN	5.733569	1.166058	0.0	4.71	5.71	6.54	58.9
z	26967.0	NaN	NaN	NaN	3.538057	0.720624	0.0	2.9	3.52	4.04	31.8
price	26967.0	NaN	NaN	NaN	3939.518115	4024.864666	326.0	945.0	2375.0	5360.0	18818.0

Figure 1. Dataset information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26967 entries, 0 to 26966
Data columns (total 10 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   carat    26967 non-null  float64
 1   cut      26967 non-null  object
 2   color    26967 non-null  object
 3   clarity  26967 non-null  object
 4   depth    26270 non-null  float64
 5   table    26967 non-null  float64
 6   x        26967 non-null  float64
 7   y        26967 non-null  float64
 8   z        26967 non-null  float64
 9   price    26967 non-null  int64
dtypes: float64(6), int64(1), object(3)
memory usage: 2.1+ MB
```

The dataset has a depth variable which has null values count of 697. However, the other variables have no null values as we can see they have 26,967 entries for each column. If we look at the data types we have float, object, and integer types.

Back to Top

1.2.1.1 Univariate Analysis

Figure 2. Carat data series: Description & graphical representation

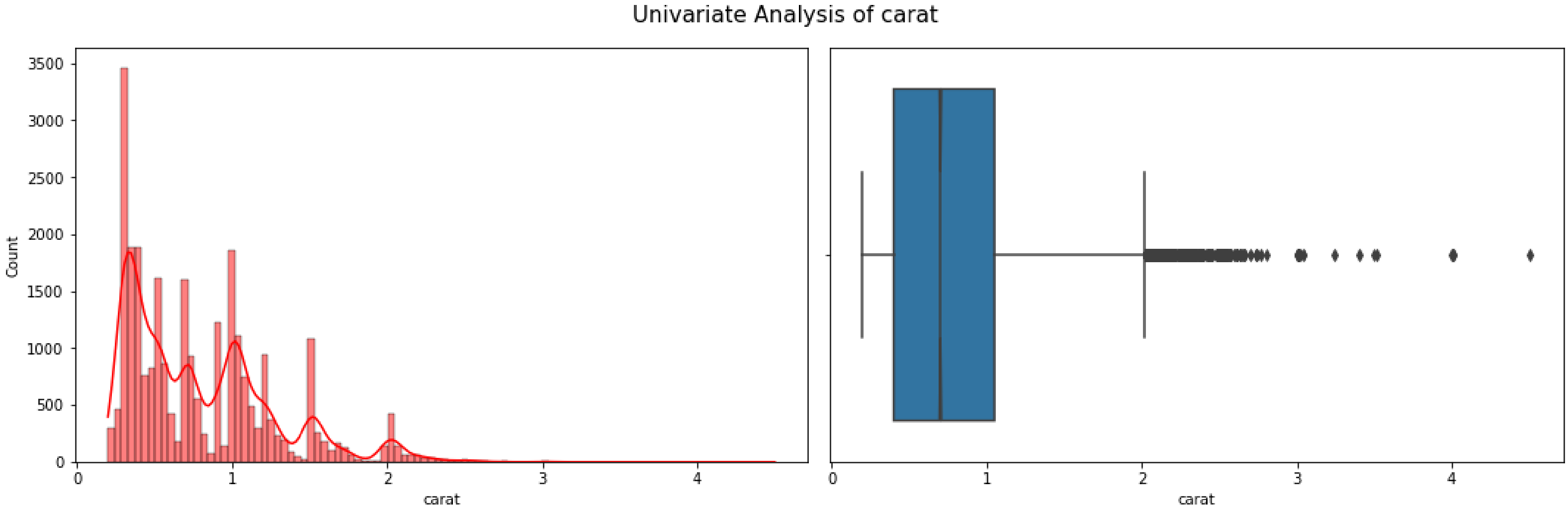


Figure 3. Depth data series: Description & graphical representation

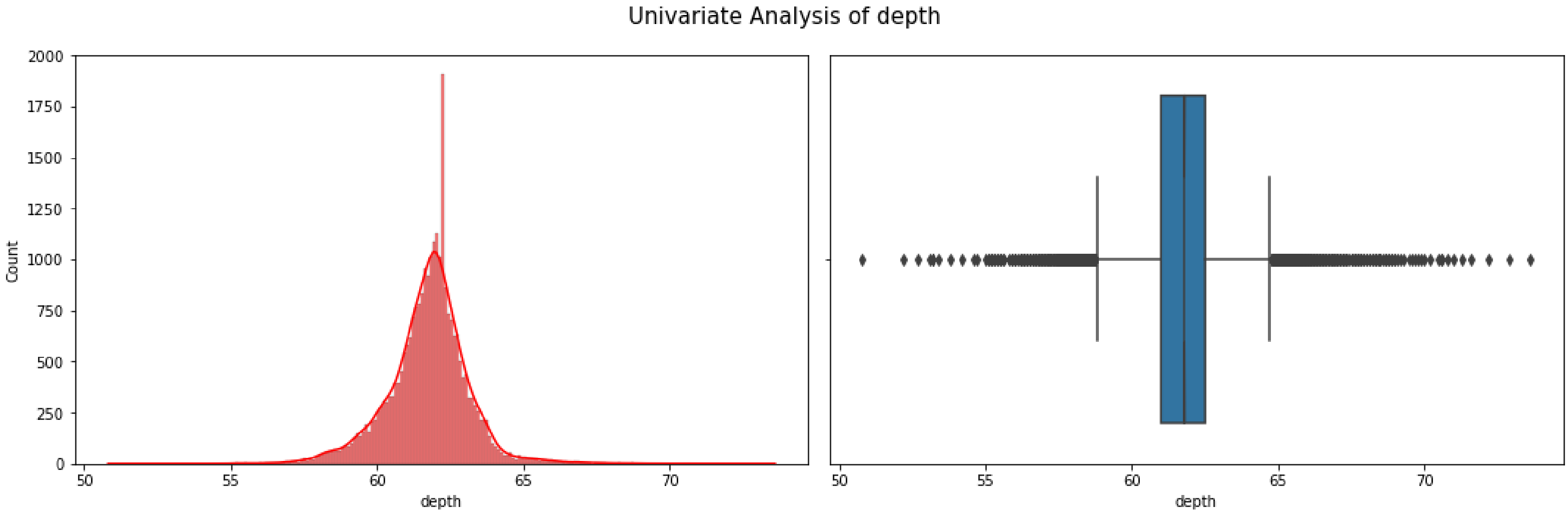


Figure 4. Table data series: Description & graphical representation

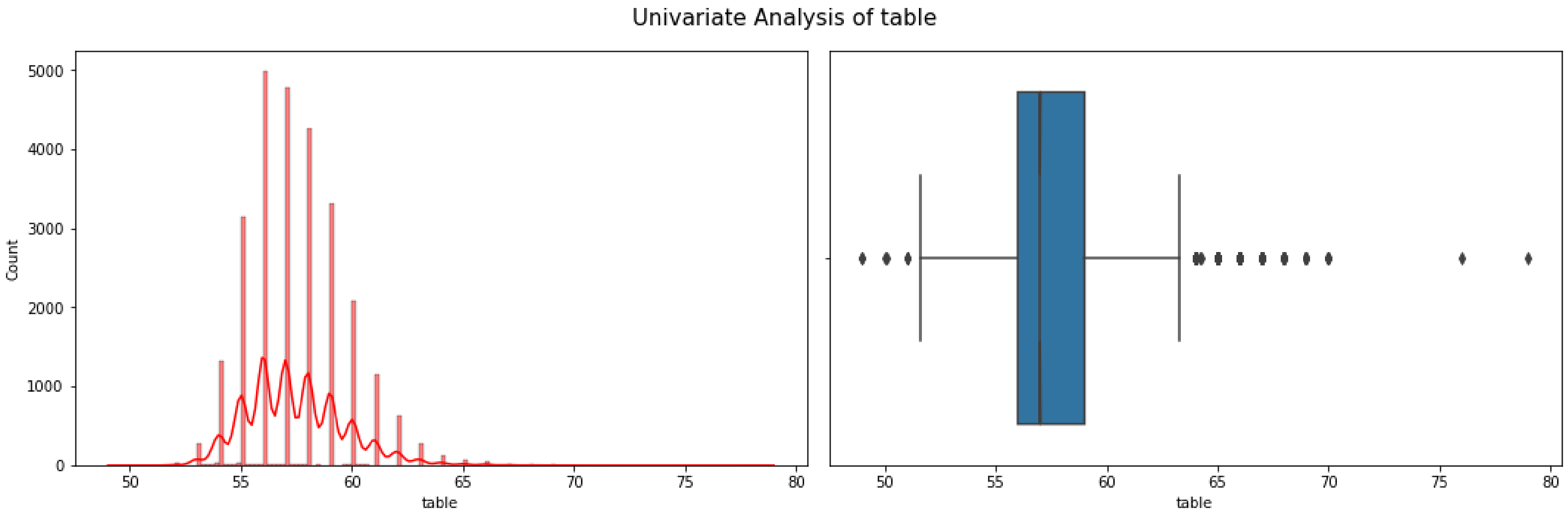


Figure 5. x data series: Description & graphical representation

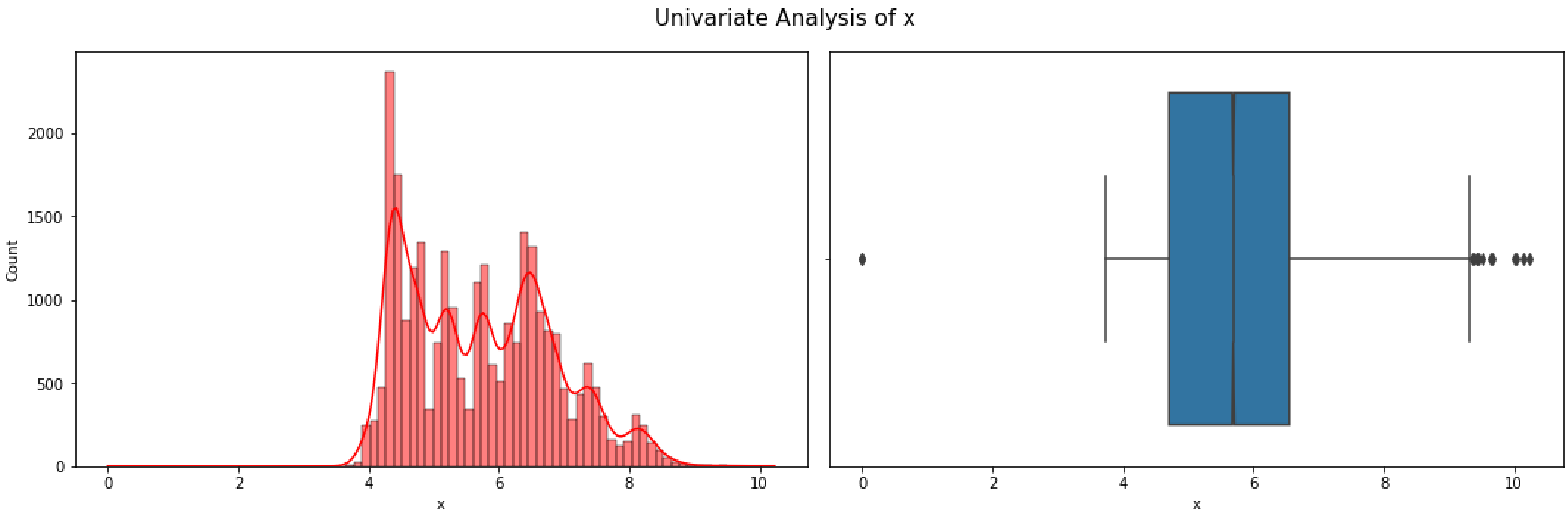
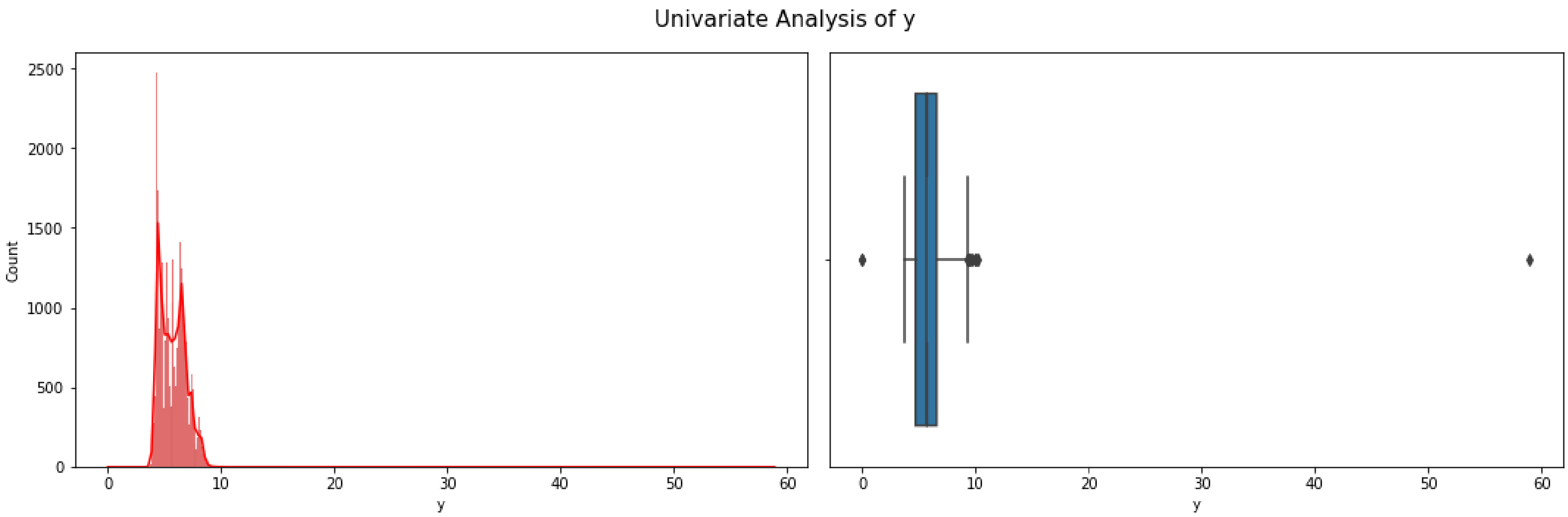


Figure 6. y data series: Description & graphical representation



Back to Top

Figure 7. z data series: Description & graphical representation

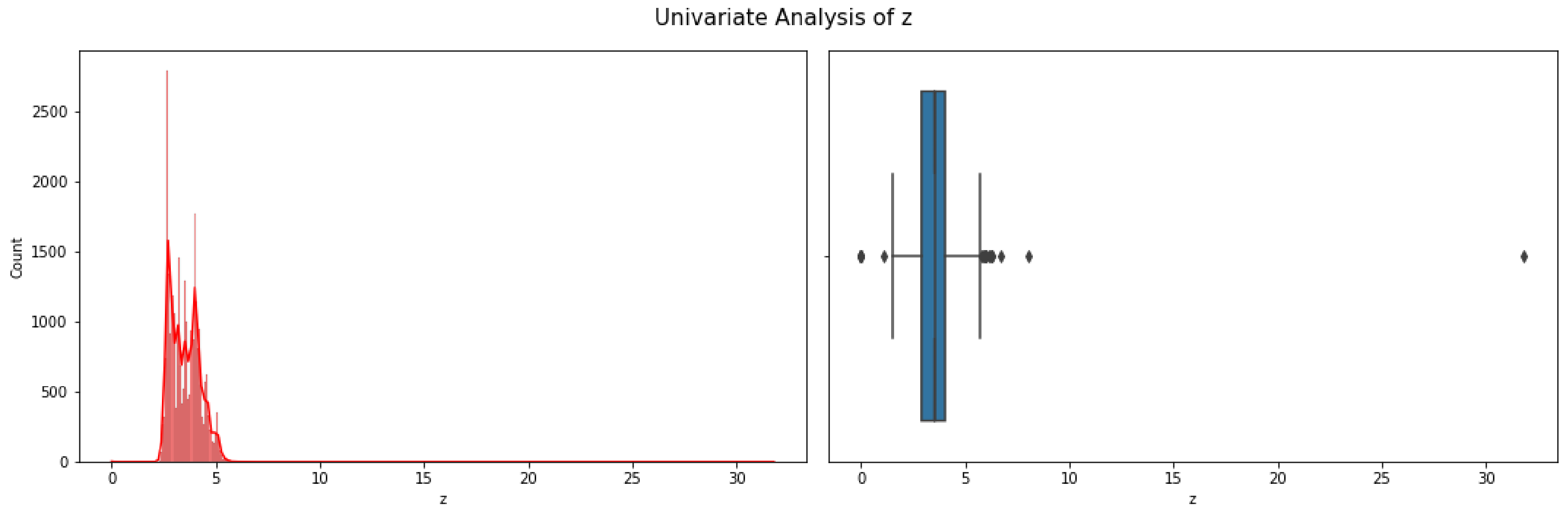


Figure 8. Price data series: Description & graphical representation

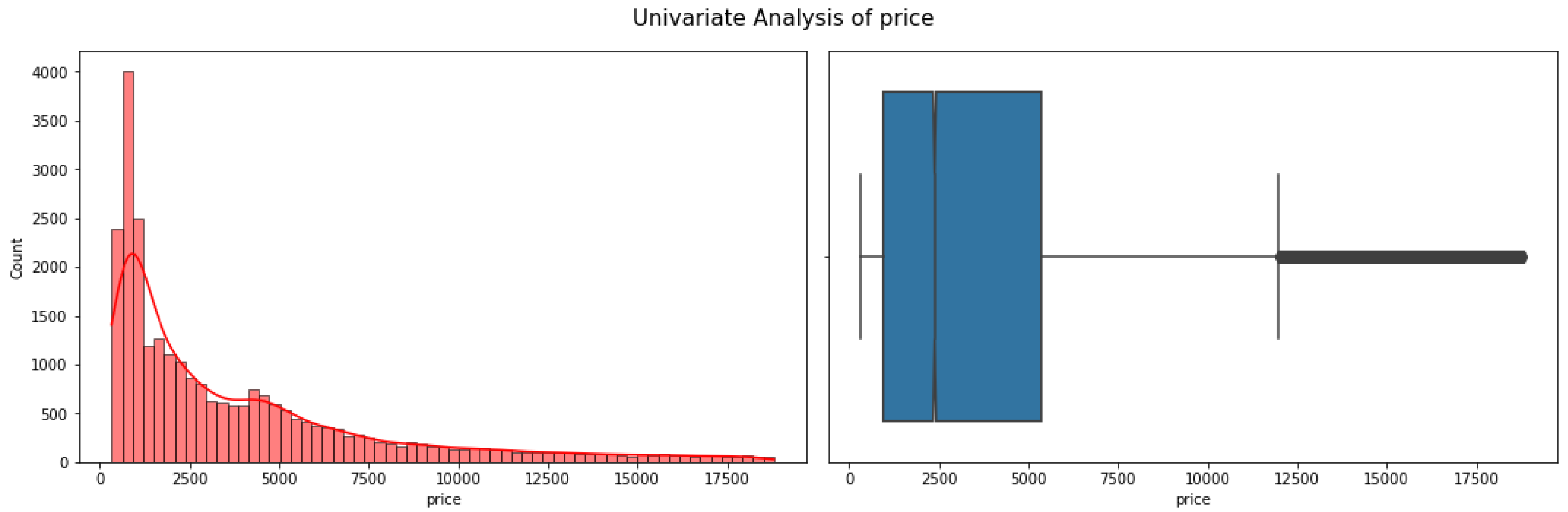
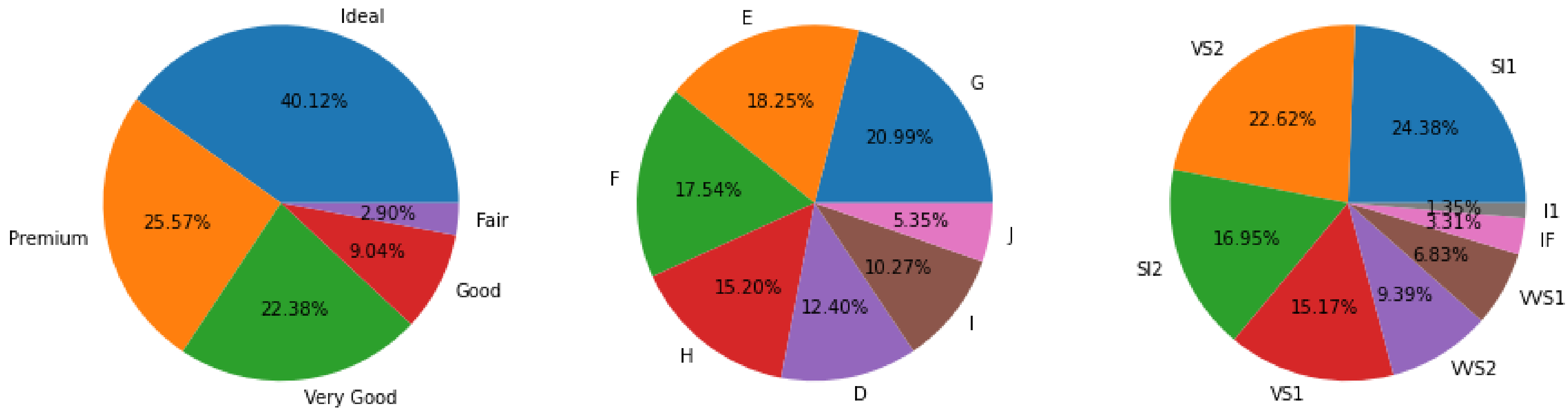
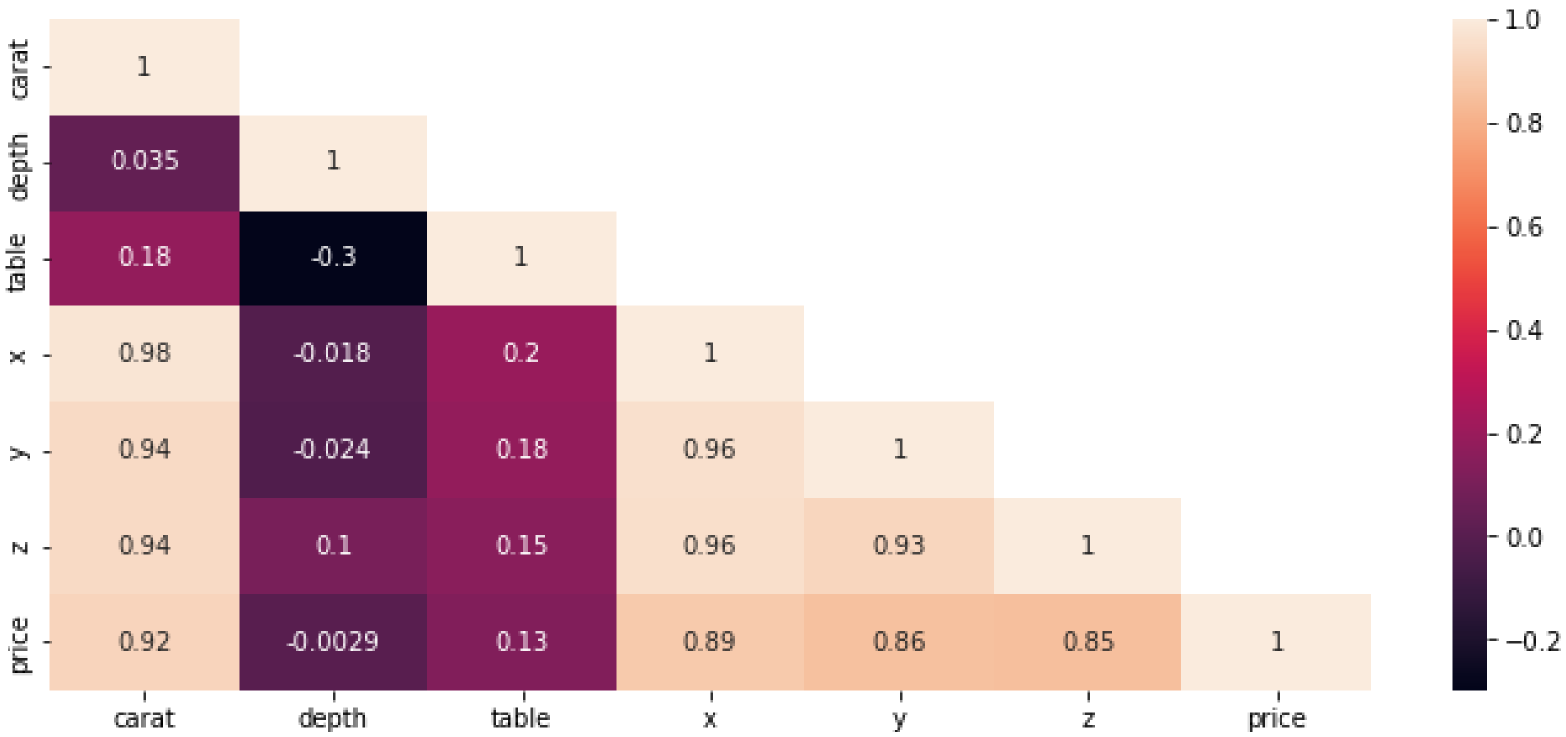


Figure 9. **Pie chart analysis of the three variables with object data type**



1.2.1.2 Bivariate & Multivariate Analysis

Figure 10. **Correlation matrix of the seven variables with integer data type**



Back to Top

Figure 11. Pairplot

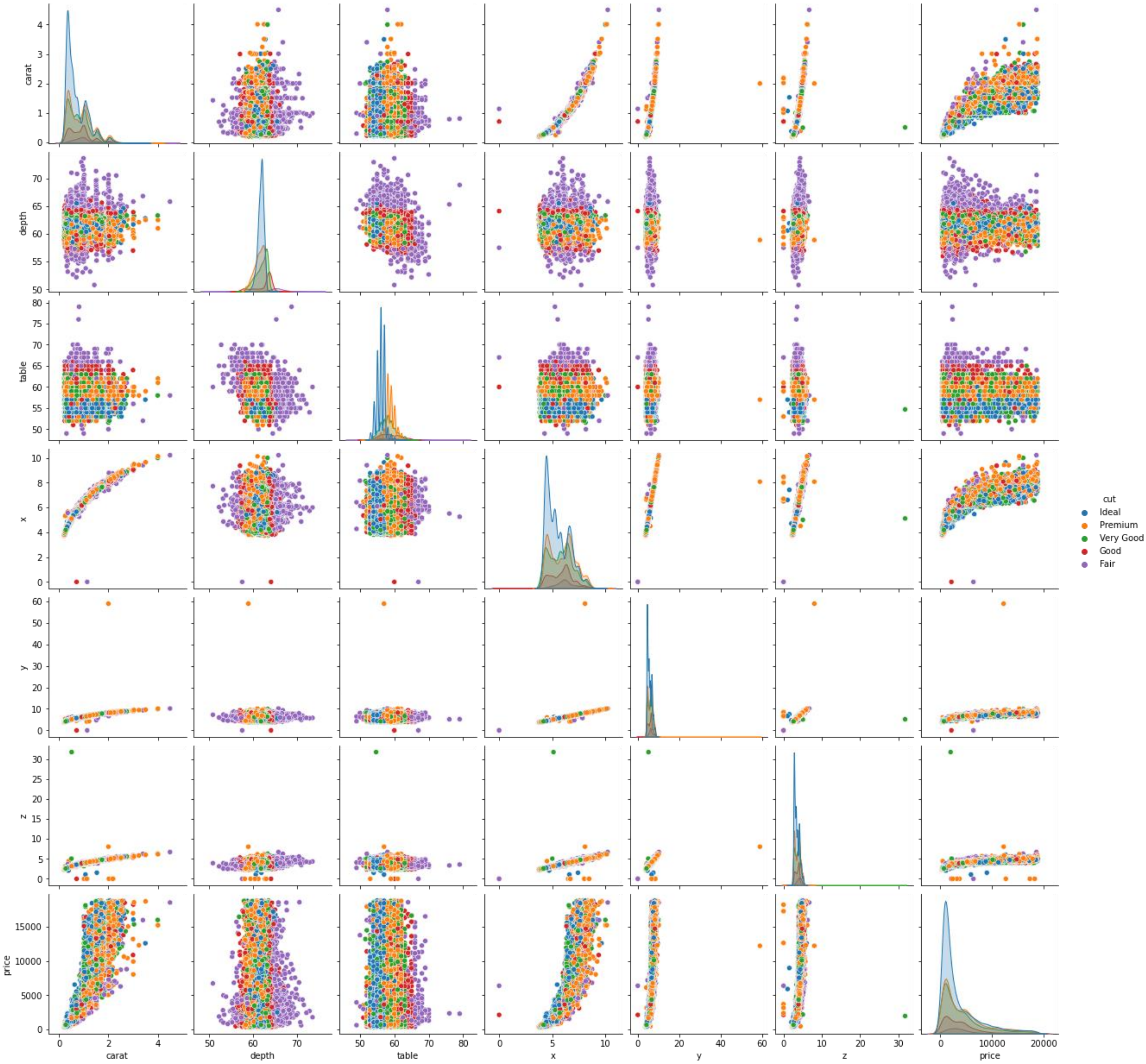


Figure 12. Scatterplot for Carat & Price variable with Cut as hue

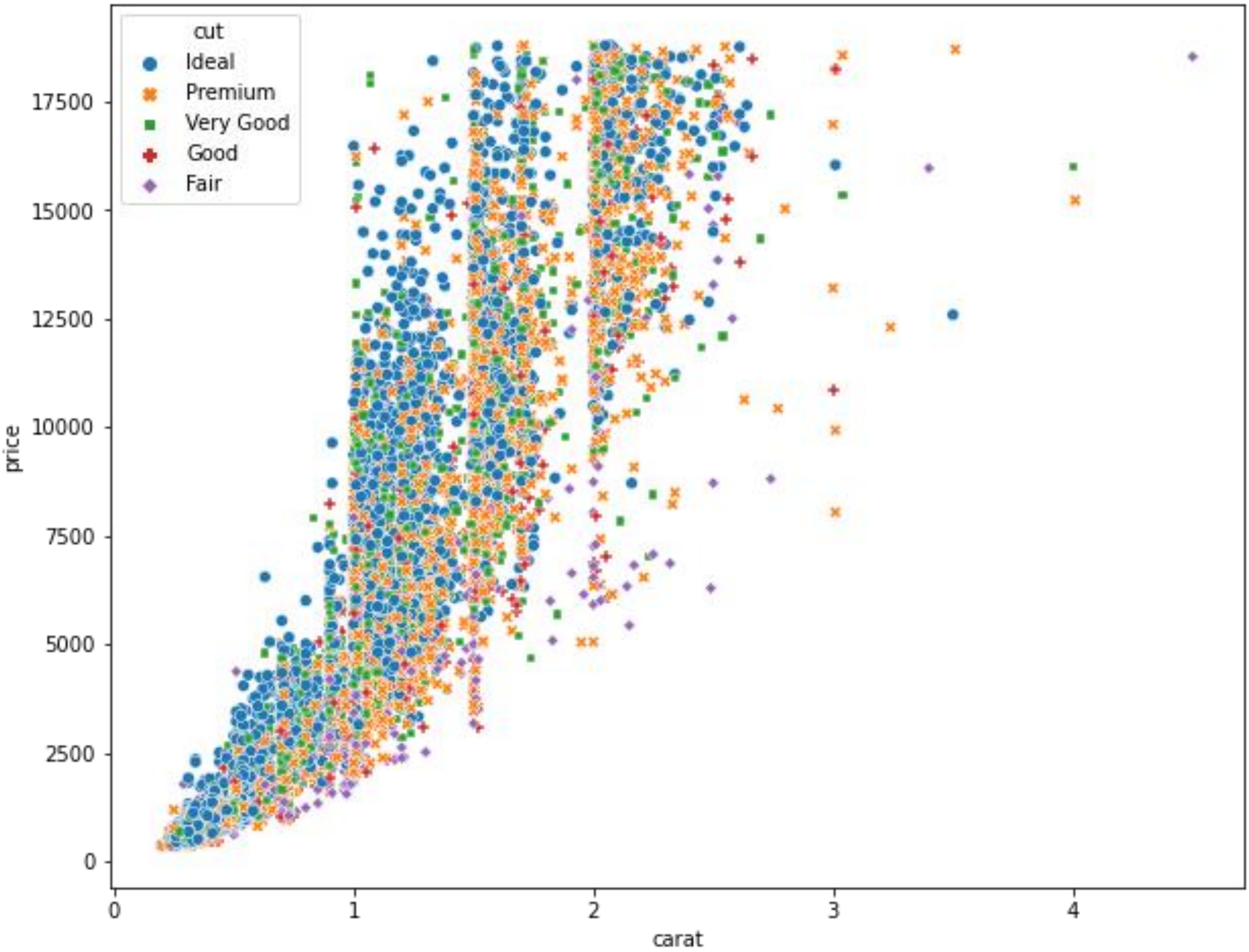


Figure 13. Bivariate analysis for categorical variable: Cut

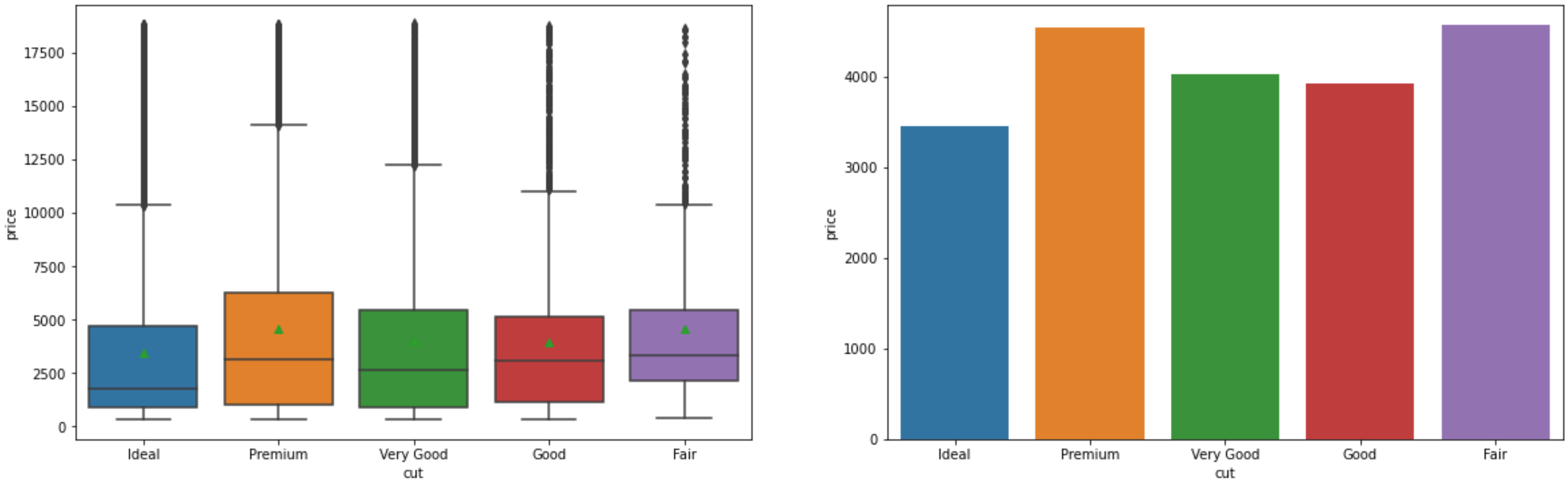


Figure 14. Bivariate analysis for categorical variable: Clarity

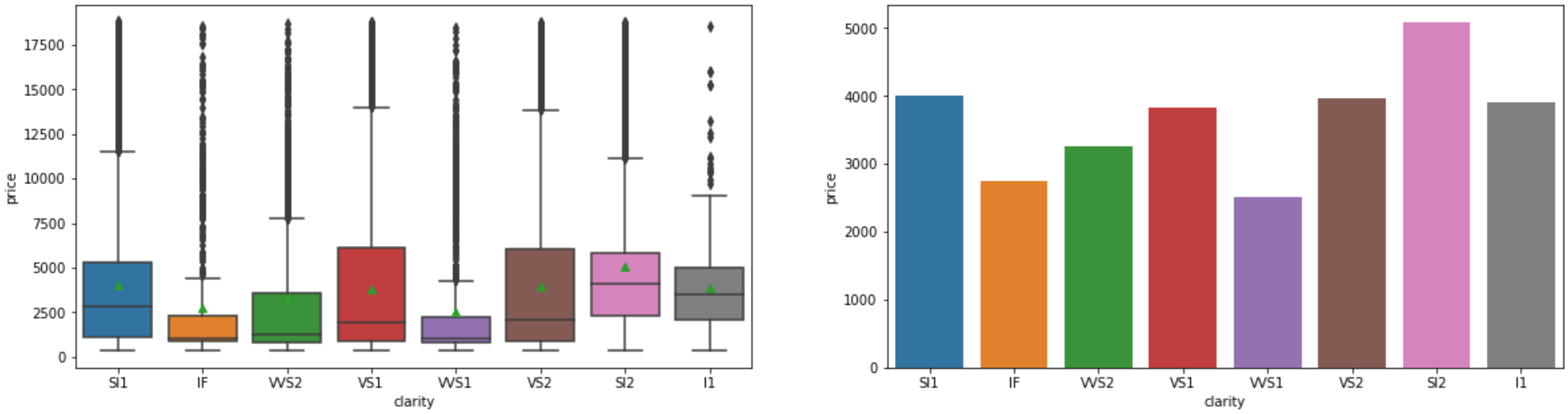
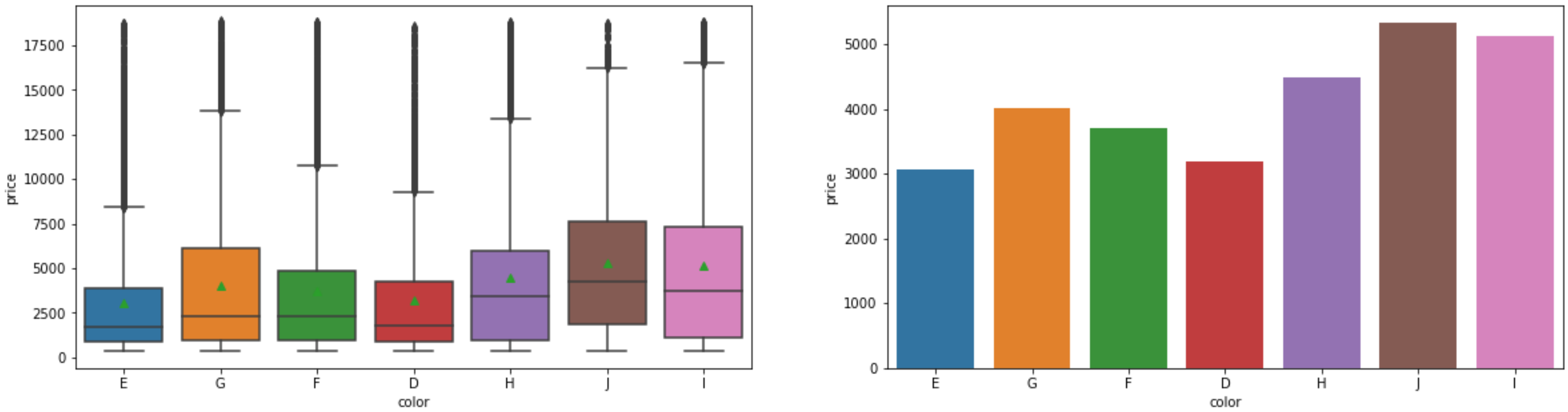


Figure 15. Bivariate analysis for categorical variable: Color



[Back to Top]

Figure 16. Counts for Categorical Variables

```
cut :

 Ideal        10805
Premium       6886
Very Good     6027
Good          2435
Fair           780
Name: cut, dtype: int64


color :

 G    5653
E    4916
F    4723
H    4095
D    3341
I    2765
J    1440
Name: color, dtype: int64


clarity :

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


Back to Top

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

As depth has null values, we have imputed them with mean using fillna function. We have also identified variables x, y, z, which are stone dimesions are zero. As height, width, and length of the stone cannot be zero, we have imputed them with values. Now, we have a dataframe with 26,925 rows and 10 columns with no null values.

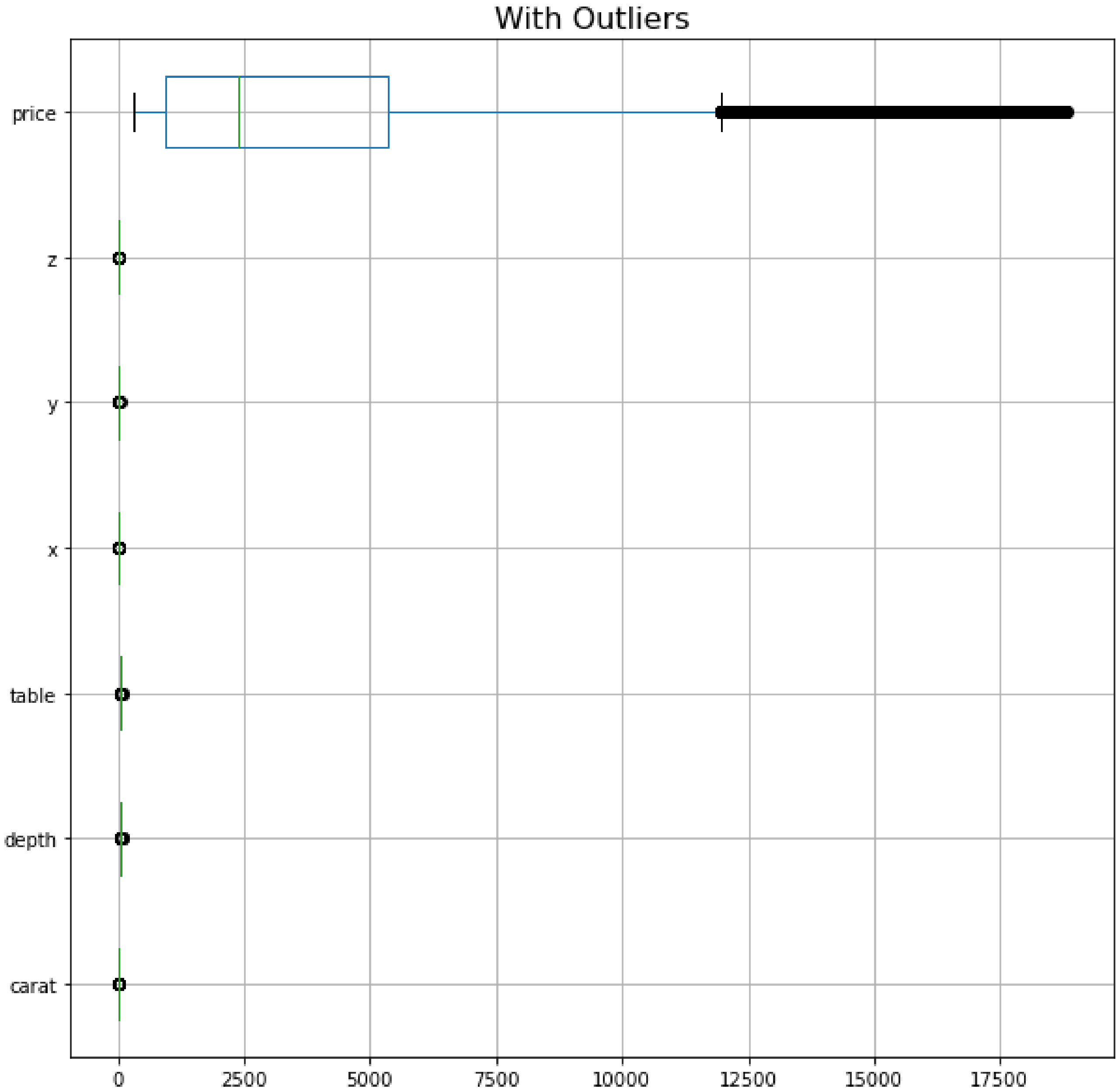
There are clarity types including IF, I1, S1, S2, VS1, VS2, VVS1, and VVS2, wherein we have combined the sublevels in clarity S1 & S2 into S1, VS1 & VS2 into VS, and VVS1 & VVS2 into VVS. For further analysis, we have to convert the caregorical variables in binary vectors which can be done using three methods one-hot enconding, dummy encoding, or manually provide numbers to the categories. We have done the same, below:

Table 3 Dataframe: df (with head function)

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

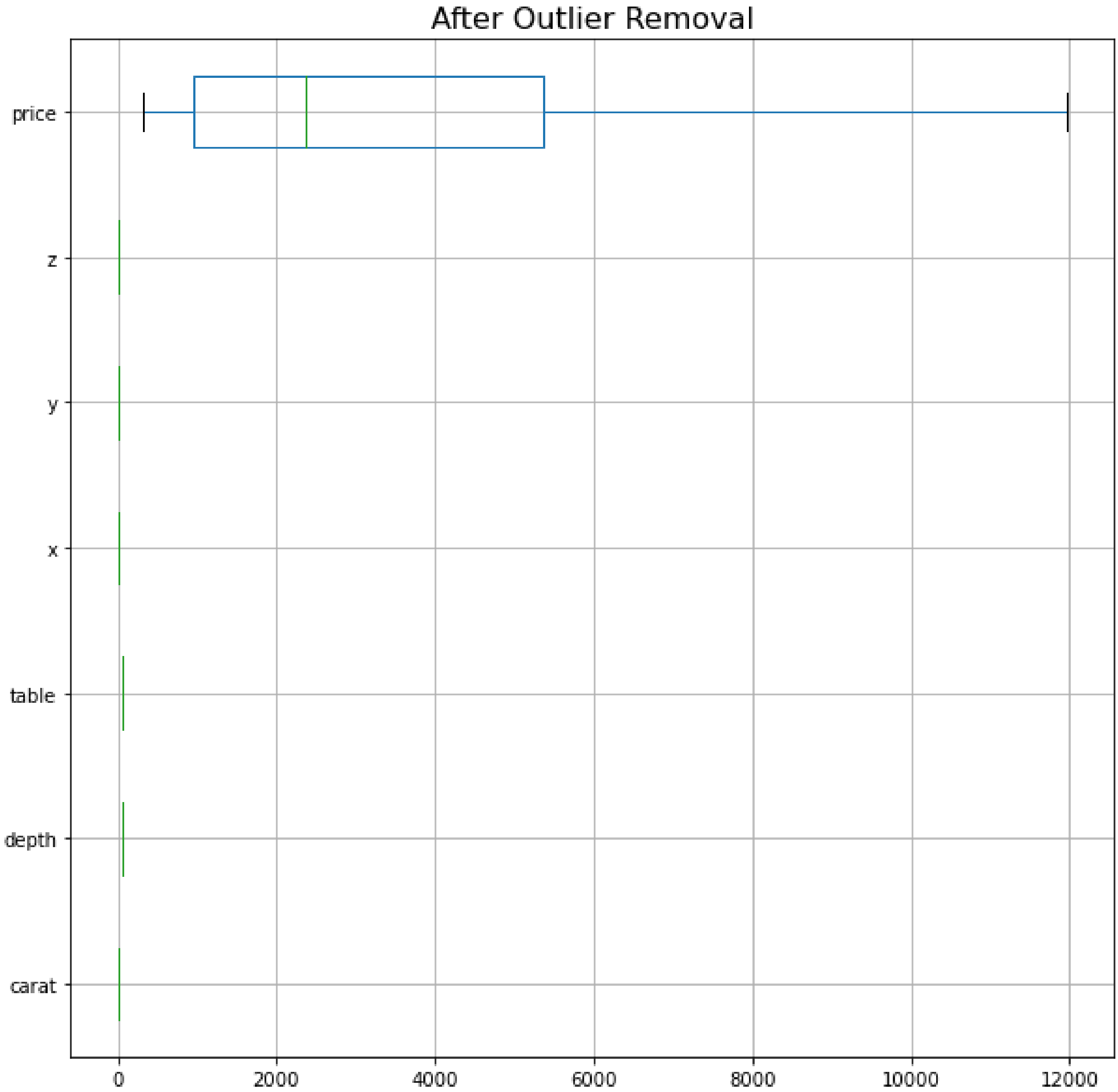
In addition, to treat the outliers for better analysis, we have used to impute lower range and upper range with the Inter Quartile Range (IQR). We have taken 5, 25, 75 percentiles of the column to treat the outliers. We have calculated IQR range and minimum threshold while calculating the lower bound and upper bound values to treat the outliers on the left of the lower whisker and right of the upper whisker respectively.

Figure 17. Box plot with outliers



Back to Top

Figure 18. Box plot post outlier treatment



The dataset shows that there are good amount of outliers present for several variables and skewness is measured for every attributes and post conducting the univariate analysis, we can see that dependent variable “price” and independent variable “carat” are rightly-skewed.

1.2.3 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.

We have dropped the dependent variable from the dataframe and formed a new dataframe “X” and we have created a new dataframe “y” with only the dependent variable that is price. Post doing that we have split the data into train and test with 70% of the overall dataset for training and 30% as test data using X & y variable.

We have called the linear regression function to conduct further analysis and find the coefficients on the X_train data. We have also determined the intercept for our linear regression model, identified the score with train and test data for X and y.

Table 4 X_train dataset (with head function)

	carat	cut	color	clarity	depth	table	x	y	z
5030	1.10	1	1	2	63.3	56.0	6.53	6.58	4.15
12108	1.01	2	0	2	64.0	56.0	6.30	6.38	4.06
20181	0.67	1	5	3	60.7	61.4	5.60	5.64	3.41
4712	0.76	1	3	2	59.0	63.0	6.05	5.97	3.47
2548	1.01	3	3	3	62.8	59.0	6.37	6.34	3.99
...	...	...	...	...	...	...	...	...	...
10965	0.53	2	3	2	61.0	55.0	5.21	5.32	3.21
17309	1.35	4	3	3	62.7	57.0	7.02	7.07	4.42
5193	1.22	3	5	3	60.6	61.0	6.94	6.88	4.19
12182	0.56	4	1	2	62.8	58.0	5.31	5.26	3.32
235	1.21	3	4	0	62.2	58.0	6.83	6.80	4.24

18847 rows × 9 columns

Back to Top

Table 5 X_test dataset (with head function)

	carat	cut	color	clarity	depth	table	x	y	z
11971	1.510	2	5	2	63.0	59.0	7.26	7.31	4.59
3294	1.020	3	3	2	60.8	58.0	6.50	6.46	3.94
25427	2.025	3	0	2	60.0	58.0	8.31	8.23	4.96
709	1.710	2	2	3	61.9	61.0	7.61	7.67	4.73
8010	1.500	1	4	3	63.9	59.0	7.25	7.18	4.61

Figure 19. Coefficient Analysis: Train dataset

```
The coefficient for carat is 9409.661275873284
The coefficient for cut is 156.50625964369178
The coefficient for color is -227.1973537195468
The coefficient for clarity is 425.4417343185445
The coefficient for depth is -6.745788707654608
The coefficient for table is -33.98837227650518
The coefficient for x is -2555.8564233651978
The coefficient for y is 2419.800348768334
The coefficient for z is -1025.1345513440062
```

Figure 20. Intercept Calculation

```
The intercept for our model is 2061.9762330956155
```

Figure 21. Model train dataset score

```
0.907681262834382
```

Figure 22. Model test dataset score

```
0.910243112128656
```

We have called the ordinary least squares (OLS) module from statsmodel into smf. The class estimates a multi-variate regression model and provides a variety of fit-statistics.

Figure 23. OLS params

```
Intercept    2061.976233
carat        9409.661276
depth          -6.745789
table         -33.988372
x           -2555.856423
y            2419.800349
z           -1025.134551
color        -227.197354
clarity       425.441734
cut           156.506260
dtype: float64
```

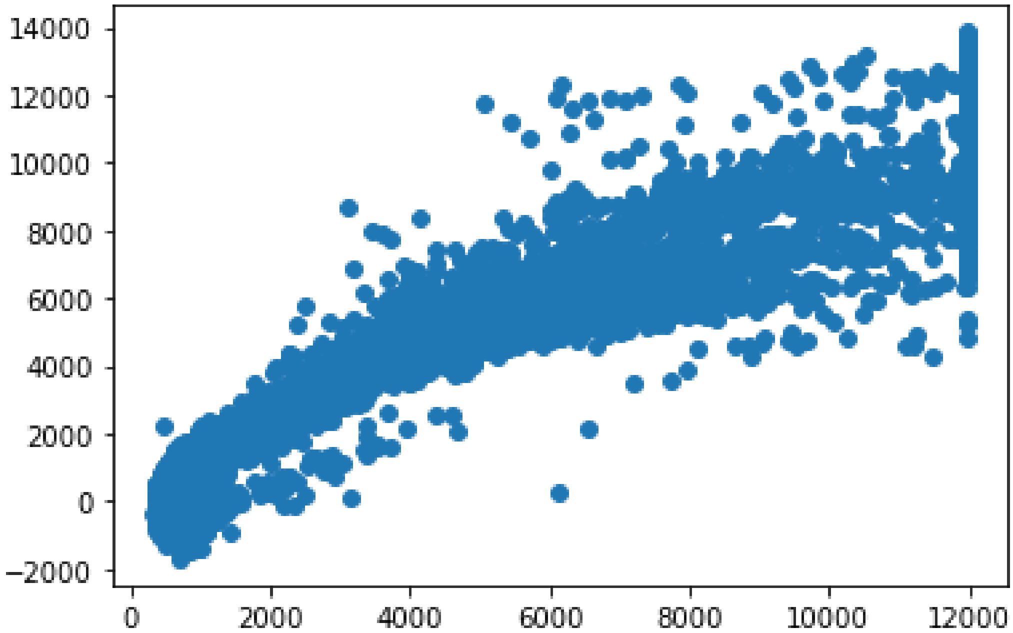
We have also printed the summary of OLS regression results to analyze the model

Figure 24. OLS Regression Results

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.908
Model:                            OLS   Adj. R-squared:                  0.908
Method:                 Least Squares   F-statistic:                 2.058e+04
Date:                Tue, 09 Aug 2022   Prob (F-statistic):               0.00
Time:                        19:08:30   Log-Likelihood:            -1.5787e+05
No. Observations:               18847   AIC:                         3.158e+05
Df Residuals:                   18837   BIC:                         3.158e+05
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   2061.9762    923.092      2.234      0.026     252.633    3871.320
carat       9409.6613     95.364     98.671      0.000    9222.739    9596.583
depth         -6.7458     12.840     -0.525      0.599     -31.914      18.422
table        -33.9884      4.529     -7.505      0.000     -42.865     -25.112
x          -2555.8564    155.904    -16.394      0.000   -2861.443   -2250.270
y           2419.8003    153.665     15.747      0.000    2118.602    2720.998
z          -1025.1346    161.138     -6.362      0.000   -1340.980    -709.289
color       -227.1974      4.716    -48.181      0.000    -236.440    -217.955
clarity      425.4417      8.853     48.055      0.000     408.089     442.795
cut          156.5063      8.450     18.522      0.000     139.944     173.068
==============================================================================
Omnibus:                     5917.366   Durbin-Watson:                   2.006
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            35162.848
Skew:                           1.380   Prob(JB):                         0.00
Kurtosis:                       9.096   Cond. No.                     1.03e+04
==============================================================================
```

We have calculated the mean squared error on the test data X and y and plotted the scatter plot graph of the test data:

Figure 25. Scatterplot for Test Data



We have scaled the split dataset to check the improvement in the model and identify if it can be done using zscore. We again using the linear regression model and determine the coefficient on the scaled train data. We have also calculated the intercept and mean squared error for the scaled data

Figure 26. Coefficient Analysis: Scaled train dataset

```
The coefficient for carat is 1.2512899513477598
The coefficient for cut is 0.050346889645917485
The coefficient for color is -0.1119604098217393
The coefficient for clarity is 0.10926011236806346
The coefficient for depth is -0.002378022918964528
The coefficient for table is -0.021193979860987224
The coefficient for x is -0.8286154429657883
The coefficient for y is 0.7791564312279321
The coefficient for z is -0.2053906720562874
```

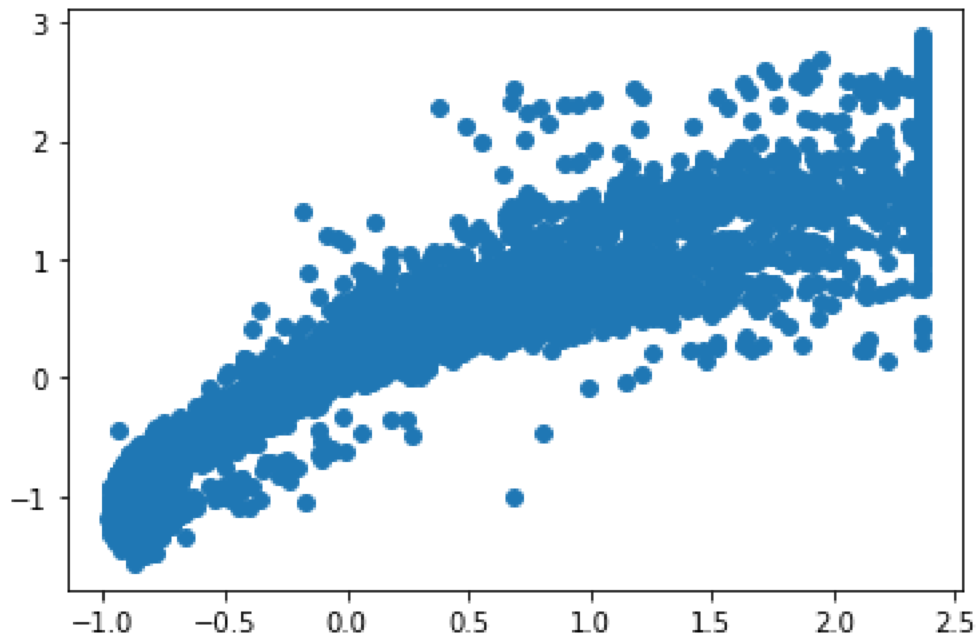
Figure 27. Intercept Calculation on Scaled Data

```
The intercept for our model is -8.700879107820098e-16
```

Figure 28. Mean Squared Error: Scaled data

```
0.2995917690605196
```

Figure 29. Scatterplot for Scaled Test Data



We have also calculated variance inflation factors below on X dataset

Figure 30. VIF Calculations

```
carat ---> 122.12104511477665
cut ---> 10.181138884451808
color ---> 3.6693970158880136
clarity ---> 10.27320983628489
depth ---> 1208.281297081905
table ---> 874.0702428119733
x ---> 10607.077935100035
y ---> 9322.733506762192
z ---> 3289.392854874187
```

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

- **Observations:**
    - We can observe there are very strong multi collinearity present in the data set. We can see R-squared:0.931 and Adj. R-squared: 0.931 are same. The overall P value is less than alpha
    - We can conclude that Best 5 attributes that are most important are 'Carat', 'Cut', 'colour', clarity' and width i.e., 'y' for predicting the price
    - We can see that the p value is 0.599 for depth variable, which is much greater than 0.05. That means this attribute is of no use
    - We can also observe more the width of the stone, it will have higher price. In addition, as we see for 'x' i.e., Length. of the stone, higher the length of the stone is lower the price.
- **Recommendations:**
    - The Gem Stones company should consider the features 'Carat', 'Cut', 'colour', 'clarity' and width i.e., 'y' as most important for predicting the price. To distinguish between higher profitable stones and lower profitable stones so as to have better profit share.
    - The 'premium’ on gemstones are the most expensive, followed by 'very good', these should consider in higher profitable stones
    - Higher the length('x') of the stone is lower is the profitability, higher the 'z' i.e height of the stone then lower the price. This is because if a gemstone’s height is too large it will become 'dark' in appearance because it will no longer return an attractive amount of light.

Chapter 2. Problem 2: Logistic Regression and LDA

2.1 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.

2.2 Introduction

The dataset has 872 rows and 7 columns after dropping unnamed column. The columns of the dataset include age, salary, holiday package, educ, no_young_children, no_older_children, and foreign.

2.2.1 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.

Table 6 Dataframe: df1 (with head function)

	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 7 Dataframe: df1 (with describe function)

	count	mean	std	min	25%	50%	75%	max
Salary	872.0	47729.172018	23418.668531	1322.0	35324.0	41903.5	53469.5	236961.0
age	872.0	39.955275	10.551675	20.0	32.0	39.0	48.0	62.0
educ	872.0	9.307339	3.036259	1.0	8.0	9.0	12.0	21.0
no_young_children	872.0	0.311927	0.612870	0.0	0.0	0.0	0.0	3.0
no_older_children	872.0	0.982798	1.086786	0.0	0.0	1.0	2.0	6.0

Figure 31. Dataset information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   Holliday_Package   872 non-null    object
 1   Salary             872 non-null    int64 
 2   age                872 non-null    int64 
 3   educ               872 non-null    int64 
 4   no_young_children  872 non-null    int64 
 5   no_older_children  872 non-null    int64 
 6   foreign            872 non-null    object
dtypes: int64(5), object(2)
memory usage: 47.8+ KB
```

The dataset has no null values as the total number of rows are 872 and the data types are in integer and object form. We have also tried to identify the duplicates and **we have no duplicate entries**. In addition, we have also checked if there are any null values, and there are no null values in the dataset.

2.2.1.1 Univariate Analysis

To analyze each of the relevant columns for object data types, we have given a value counts function with outputs below:

Figure 32. Holliday Package

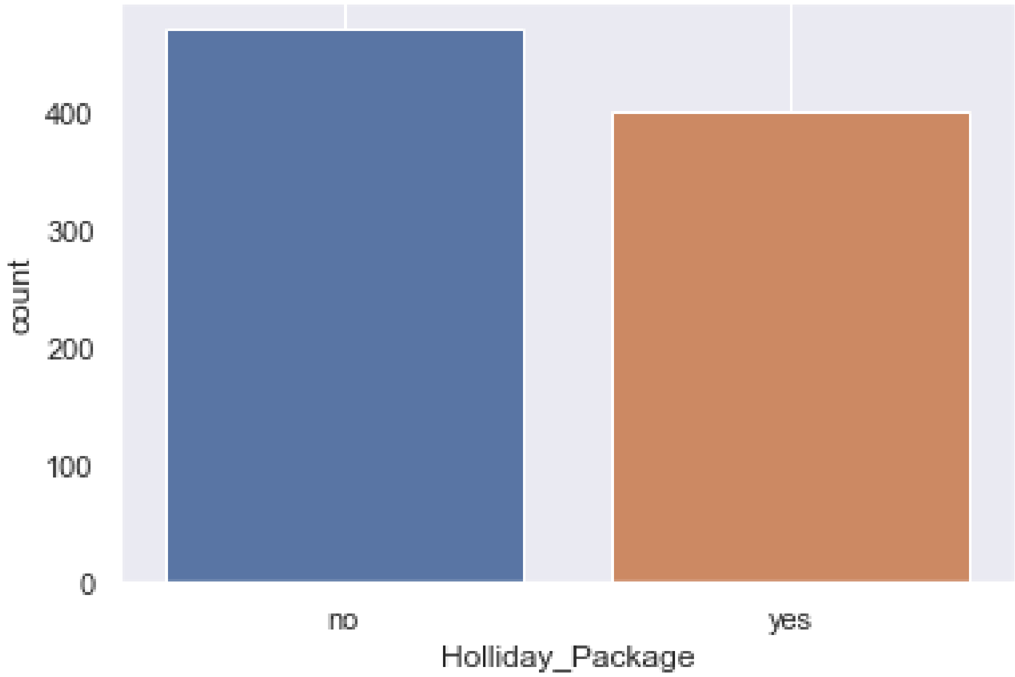


Figure 33. Foreign

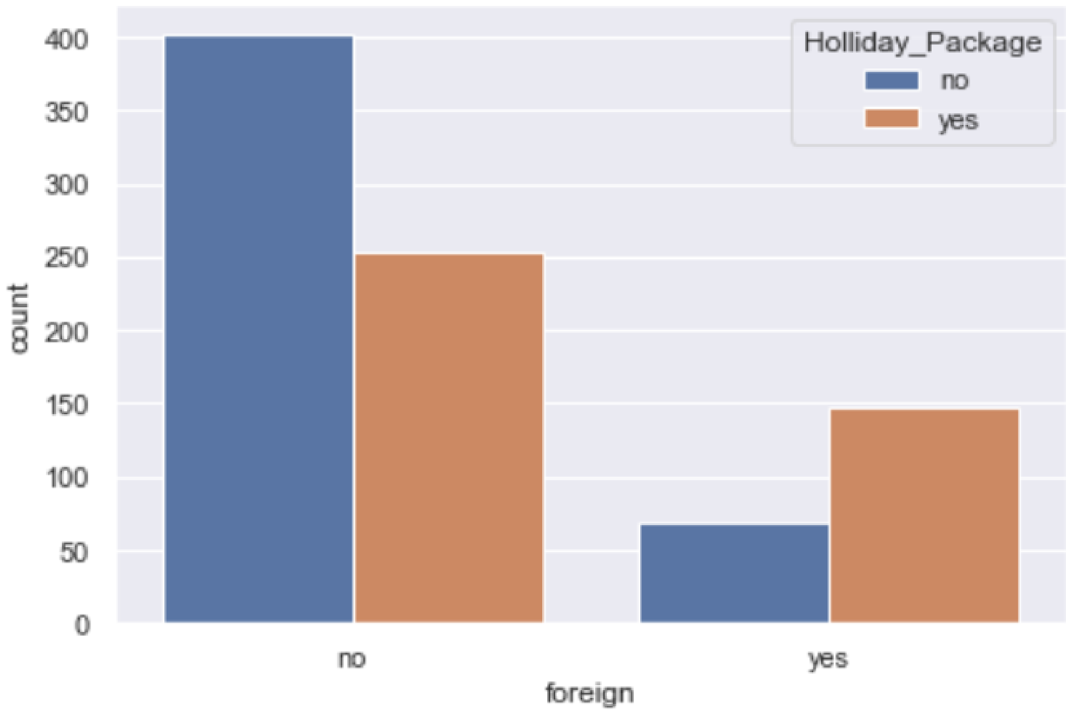


Figure 34. Swarmplot with salary and age as hue for holliday package

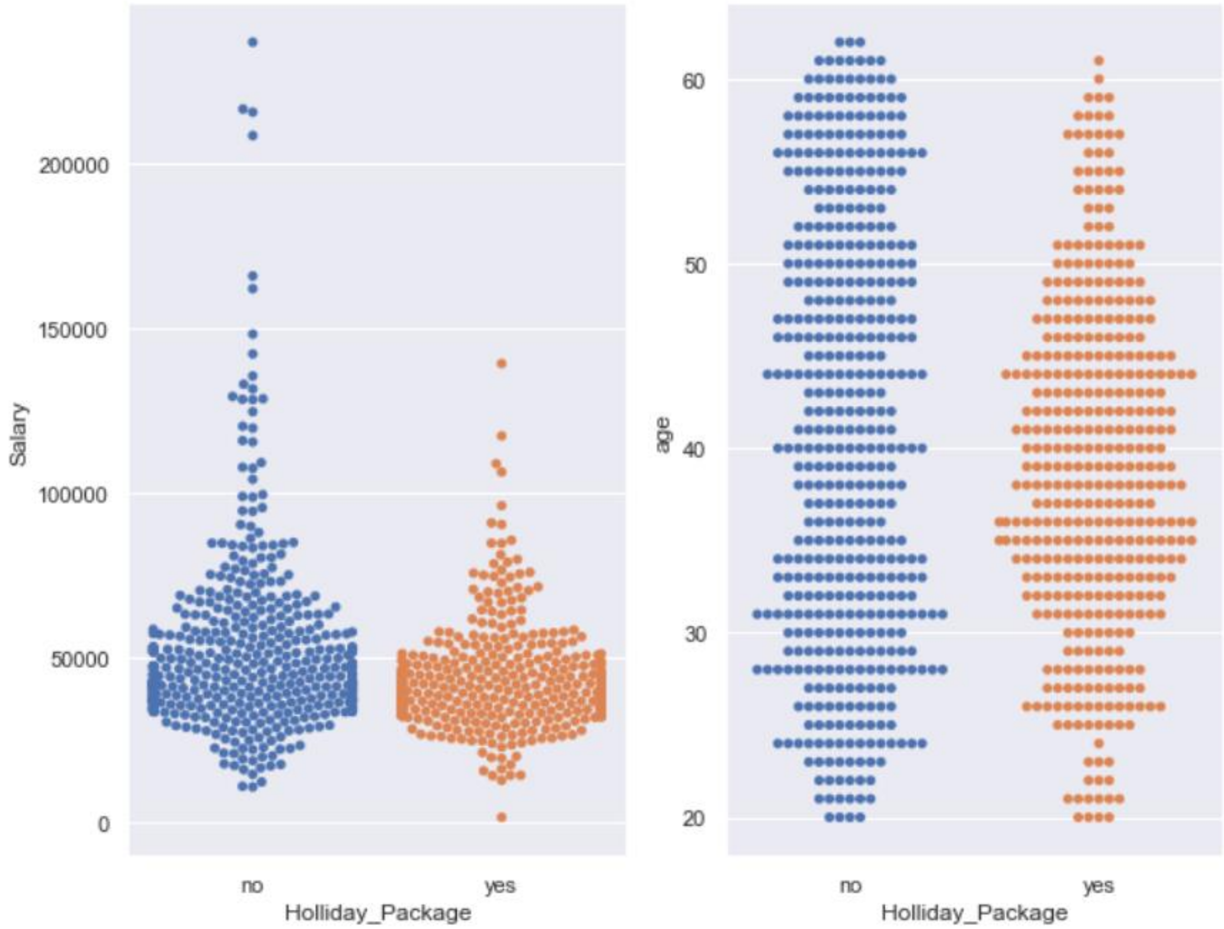
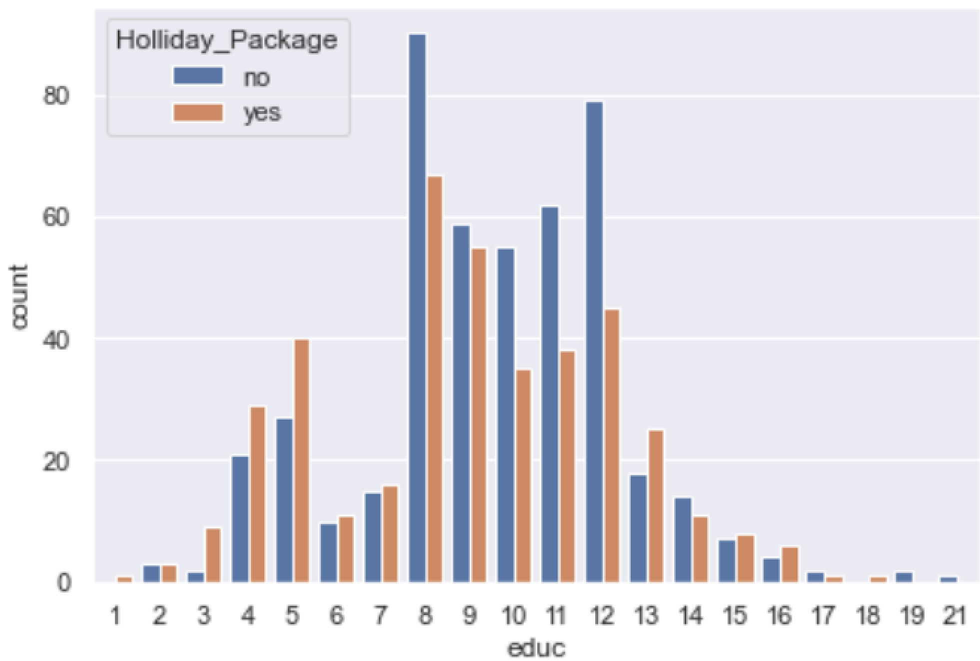


Figure 35. Count plot for educ with holliday package as hue



Back to Top

Figure 36. **Count plot for no_young_children with holliday package as hue**

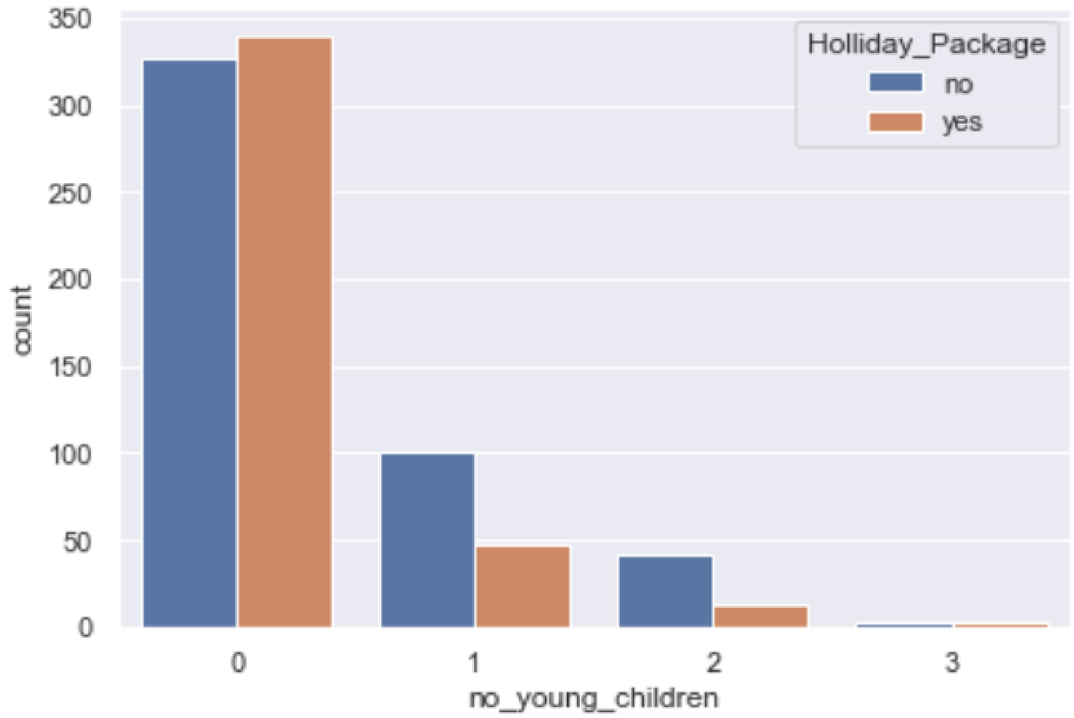
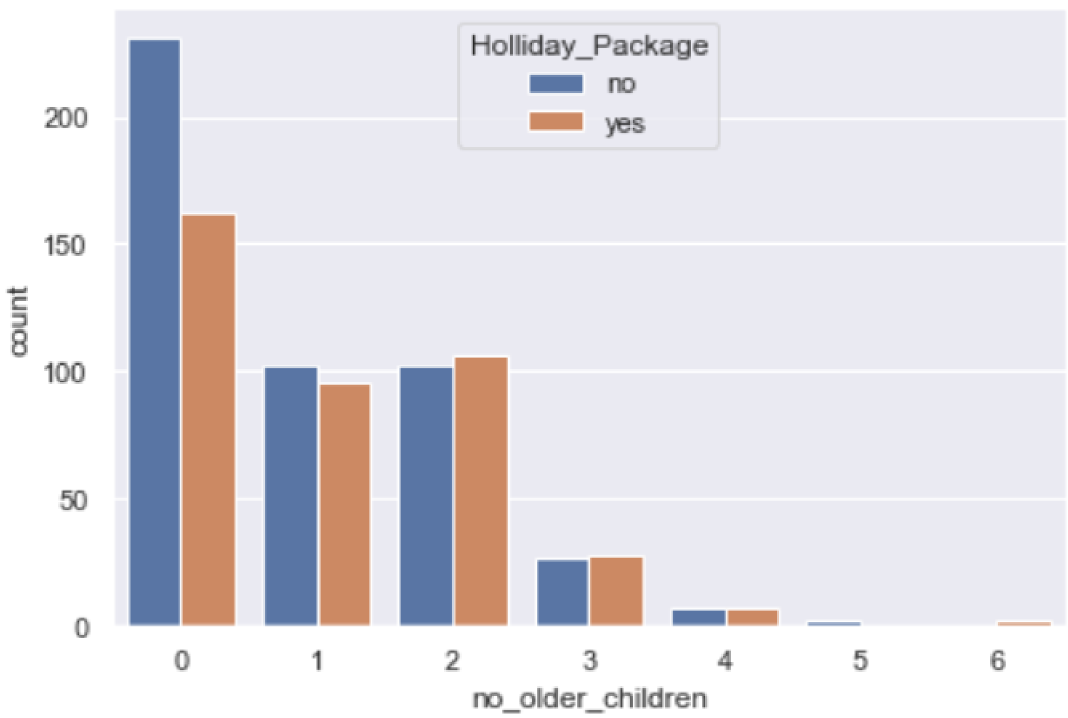


Figure 37. **Count plot for no_older_children with holliday package as hue**



2.2.1.2 Bivariate and Multivariate Analysis

Figure 38. Pairplot (Claimed variable as hue)

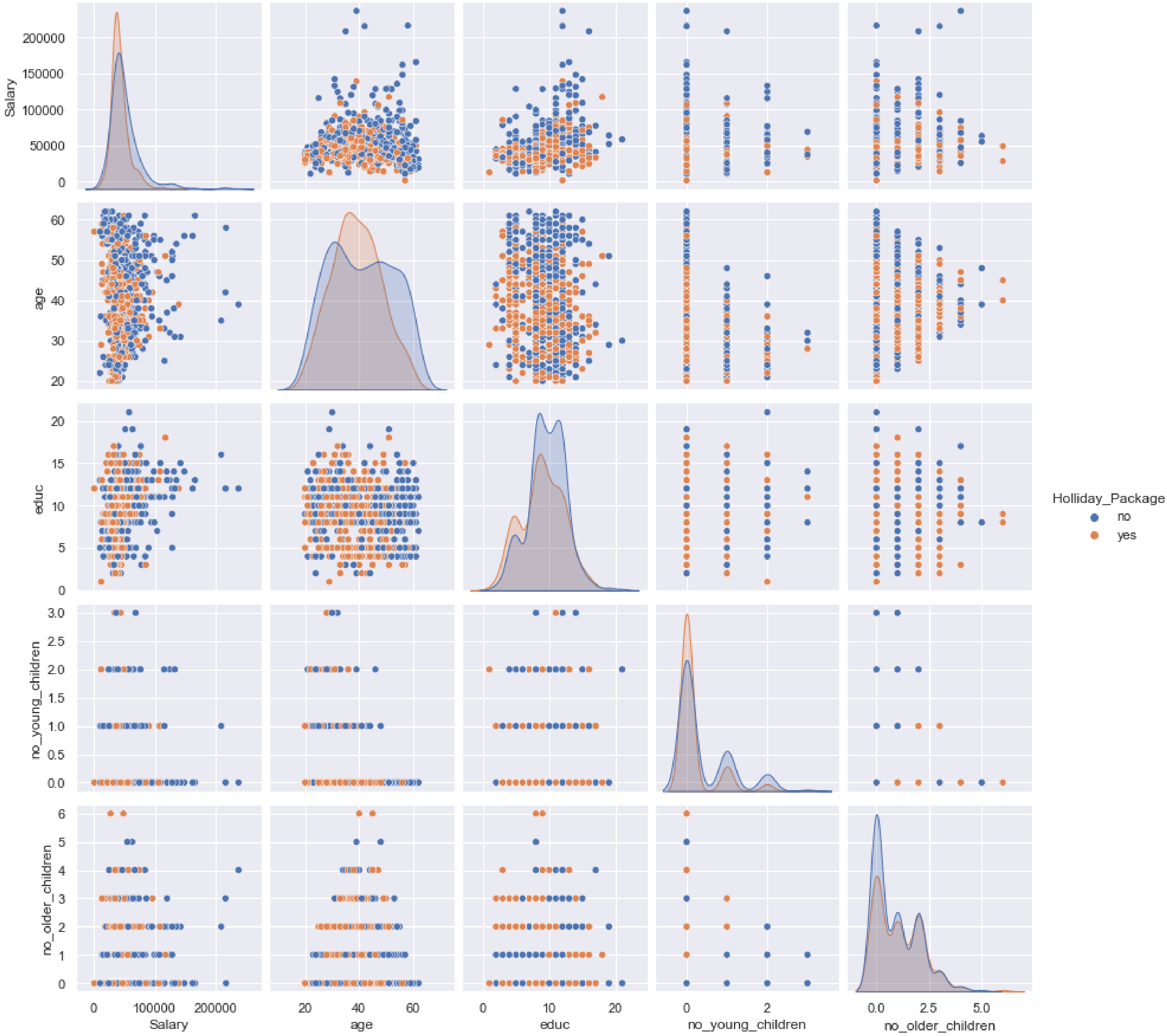


Figure 39. VIF Calculation

	Variables	VIF
0	Salary	6.027872
1	age	6.832751
2	educ	8.890845
3	no_young_children	1.403995
4	no_older_children	1.817912

Figure 40. Box Plot: Salary (without outlier treatment)

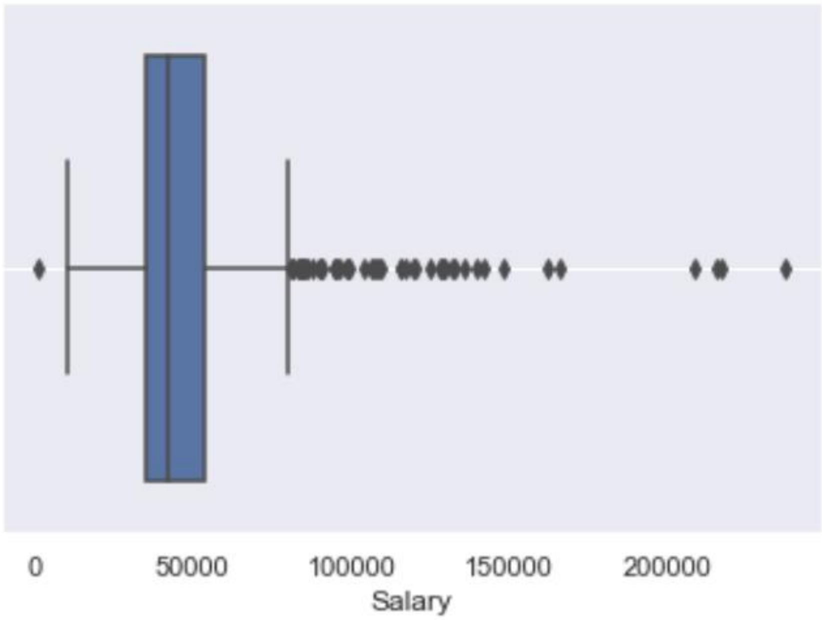


Figure 41. Box Plot: Age (without outlier treatment)

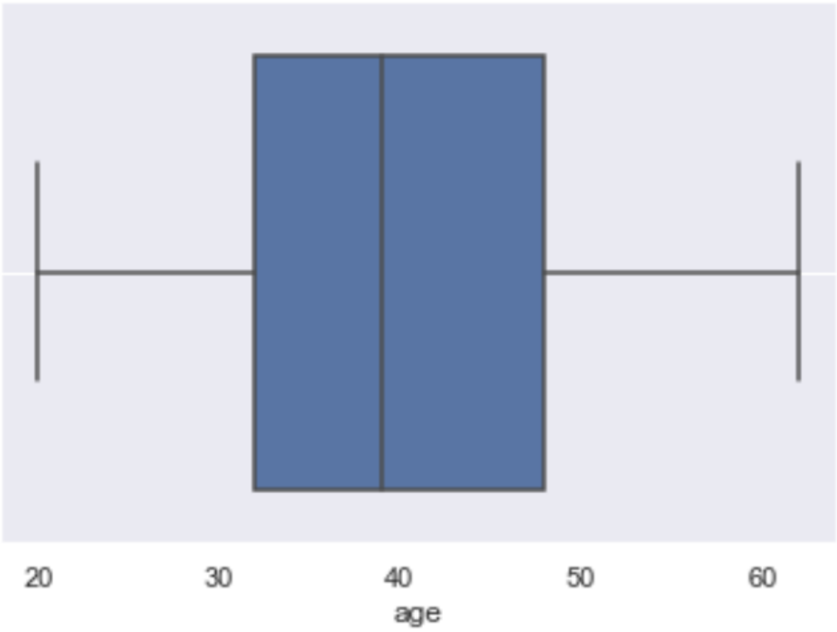


Figure 42. Box Plot: Educ (without outlier treatment)

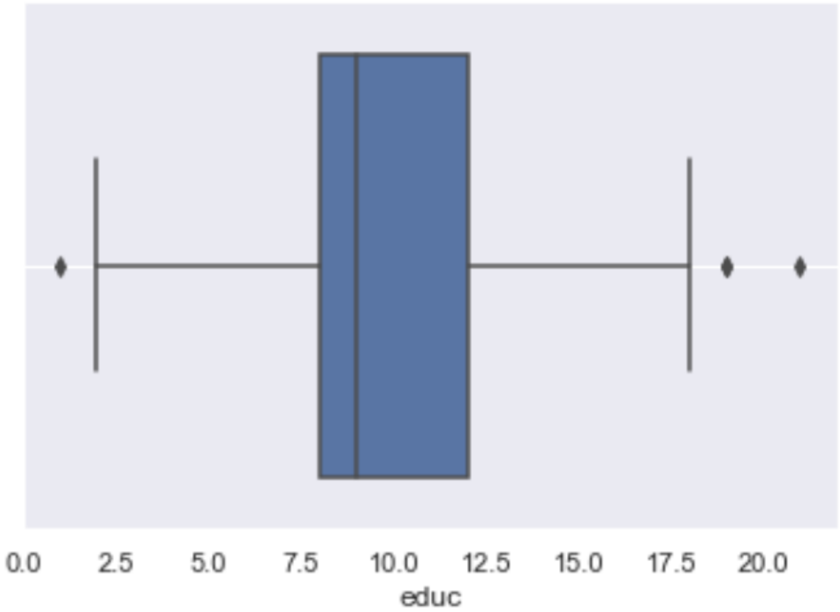


Figure 43. Box Plot: no_younger_children (without outlier treatment)

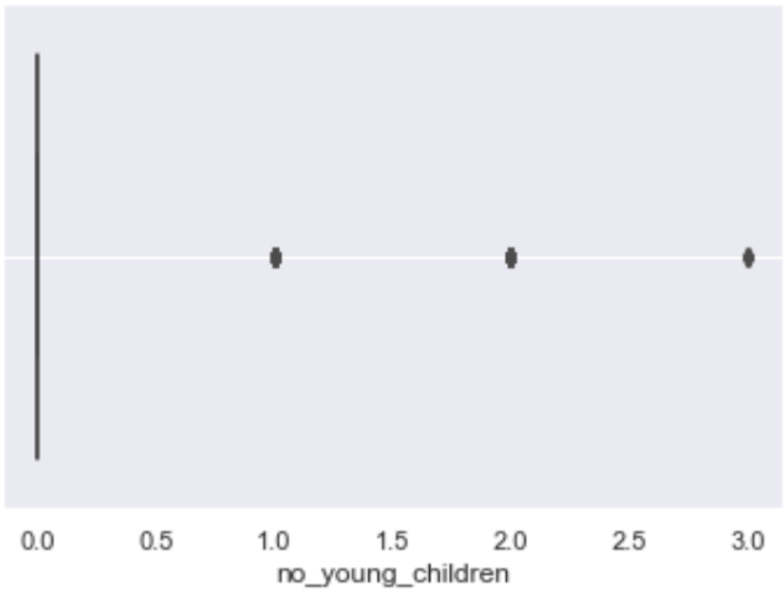


Figure 44. Box Plot: no_older_children (without outlier treatment)

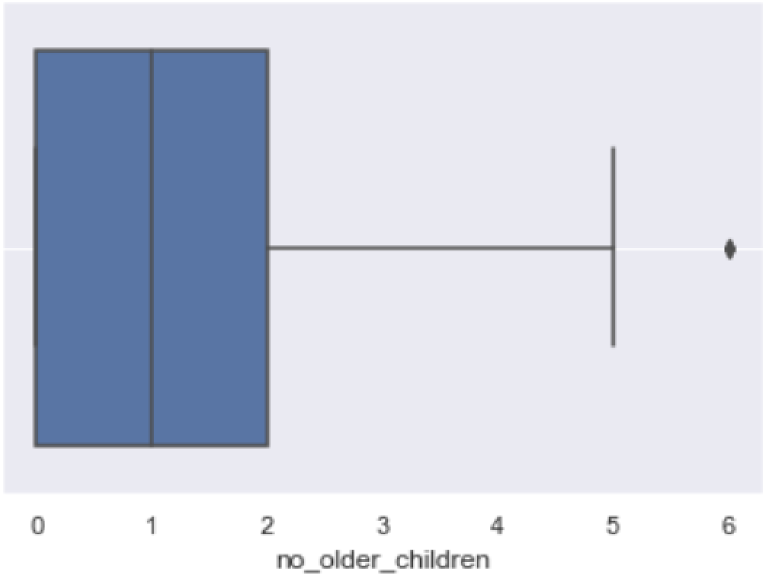


Figure 45. Box Plot: Salary (with outlier treatment)

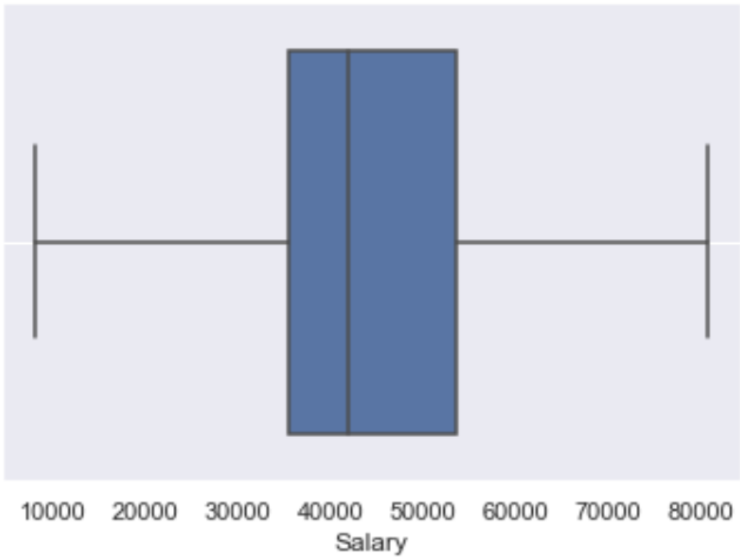
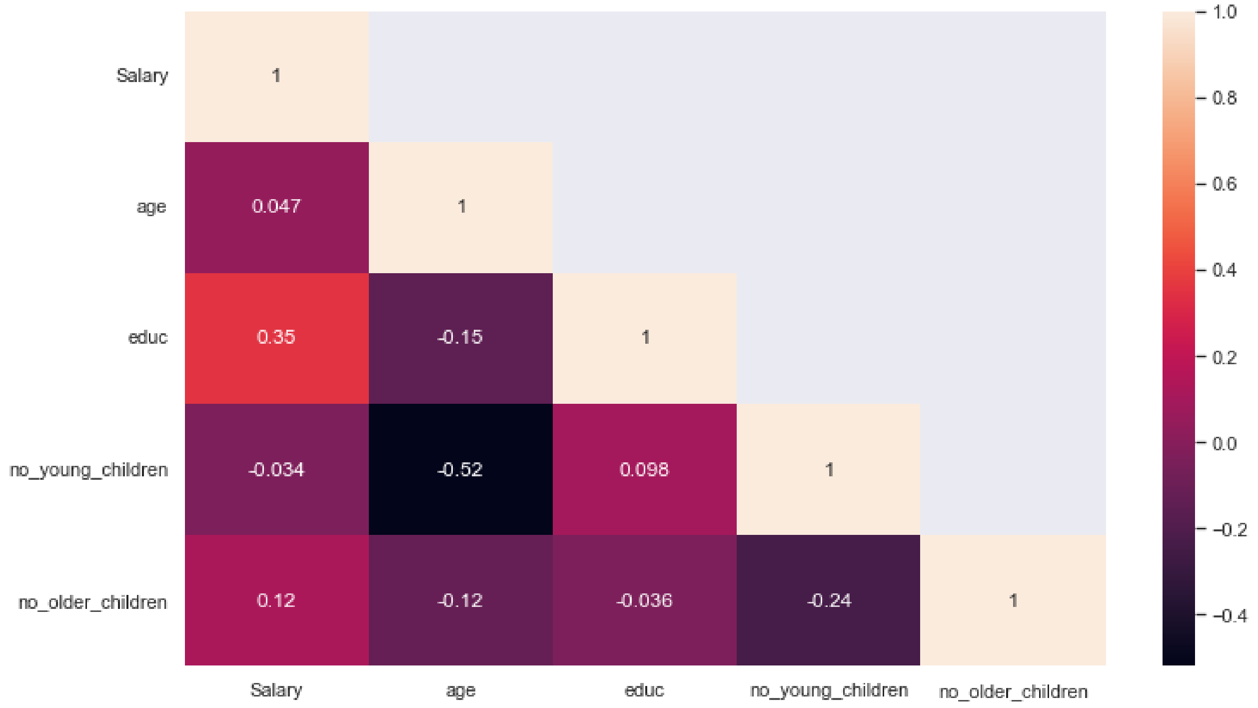


Figure 46. Correlation Matrix

	Salary	age	educ	no_young_children	no_older_children
Salary	1.000000	0.047029	0.352726	-0.034360	0.121993
age	0.047029	1.000000	-0.149294	-0.519093	-0.116205
educ	0.352726	-0.149294	1.000000	0.098350	-0.036321
no_young_children	-0.034360	-0.519093	0.098350	1.000000	-0.238428
no_older_children	0.121993	-0.116205	-0.036321	-0.238428	1.000000

Figure 47. Correlation Matrix Heatmap



Back to Top

2.2.2 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).

Figure 48. Label Encoding

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	0	48412.0	30	8	1	1	0
1	1	37207.0	45	8	0	1	0
2	0	58022.0	46	9	0	0	0
3	0	66503.0	31	11	2	0	0
4	0	66734.0	44	12	0	2	0

Figure 49. Dummy Encoding

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	0	48412.0	30	8	1	1	0
1	1	37207.0	45	8	0	1	0
2	0	58022.0	46	9	0	0	0
3	0	66503.0	31	11	2	0	0
4	0	66734.0	44	12	0	2	0

Figure 50. X, y Split dataset (Train and Test)

```
Shape of X_train (610, 6)
Shape of X_test (262, 6)
Shape of y_train (610,)
Shape of y_test (262,)
Shape of df1 dataframe (872, 7)
```

Figure 51. Logistic Regression Model

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

Figure 52. Linear Discriminant Analysis (LDA)

```
▾ LinearDiscriminantAnalysis
LinearDiscriminantAnalysis()
```

2.2.3 Performance Metrics: Comment and Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score, classification reports for each model.

2.2.3.1 Logistic Regression

Figure 53. X train and y train model score

```
0.6672131147540984
```

Figure 54. X test and y test model score

```
0.648854961832061
```

Figure 55. ytest_predict_prob

	0	1
0	0.677850	0.322150
1	0.534541	0.465459
2	0.691849	0.308151
3	0.487796	0.512204
4	0.571939	0.428061

Train label AUC score: 0.733

Test label AUC score: 0.733

Figure 56. **Train label ROC curve**

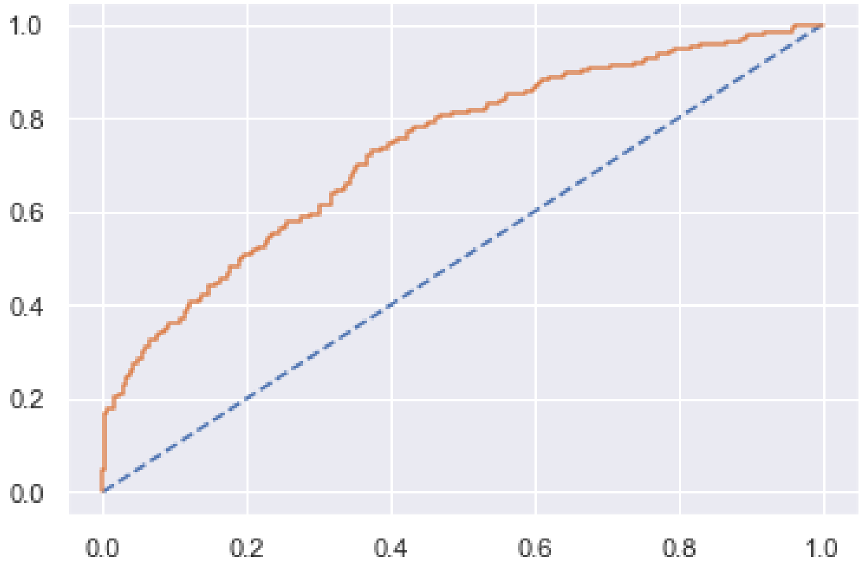
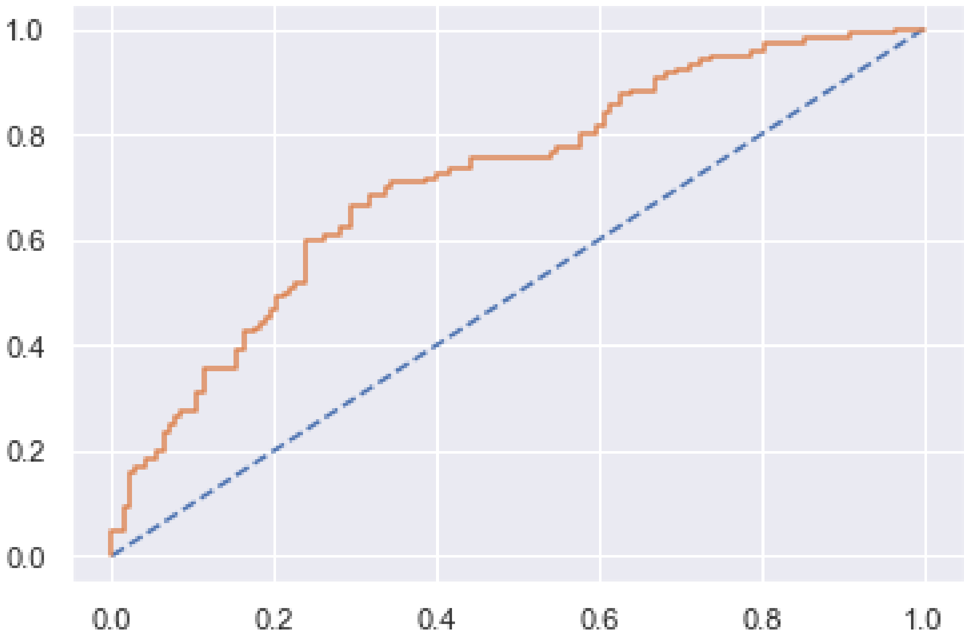


Figure 57. **Test label ROC curve**



2.2.3.1.1 Classification Report

Train Label:

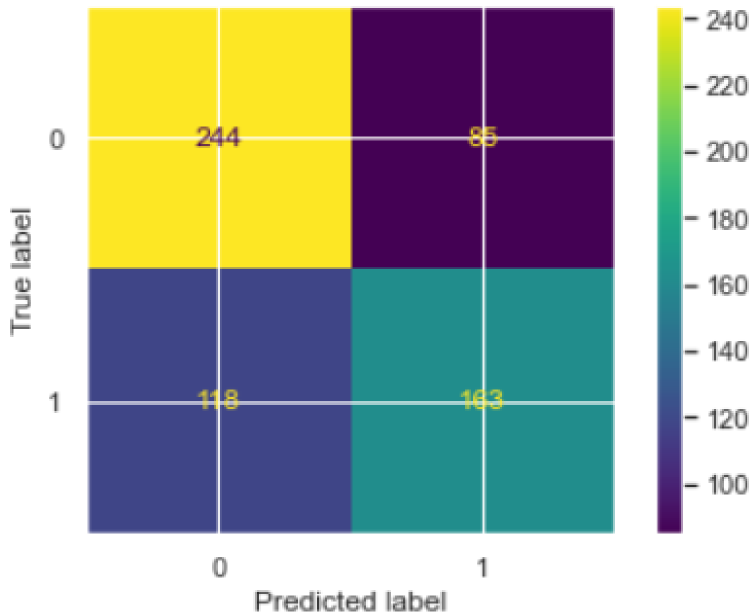
	precision	recall	f1-score	support
0	0.67	0.74	0.71	329
1	0.66	0.58	0.62	281
accuracy			0.67	610
macro avg	0.67	0.66	0.66	610
weighted avg	0.67	0.67	0.66	610

Test Label:

	precision	recall	f1-score	support
0	0.65	0.76	0.70	142
1	0.65	0.52	0.57	120
accuracy			0.65	262
macro avg	0.65	0.64	0.64	262
weighted avg	0.65	0.65	0.64	262

2.2.3.1.2 Confusion Matrix

Figure 58. Train label



Back to Top

Figure 59. Test label

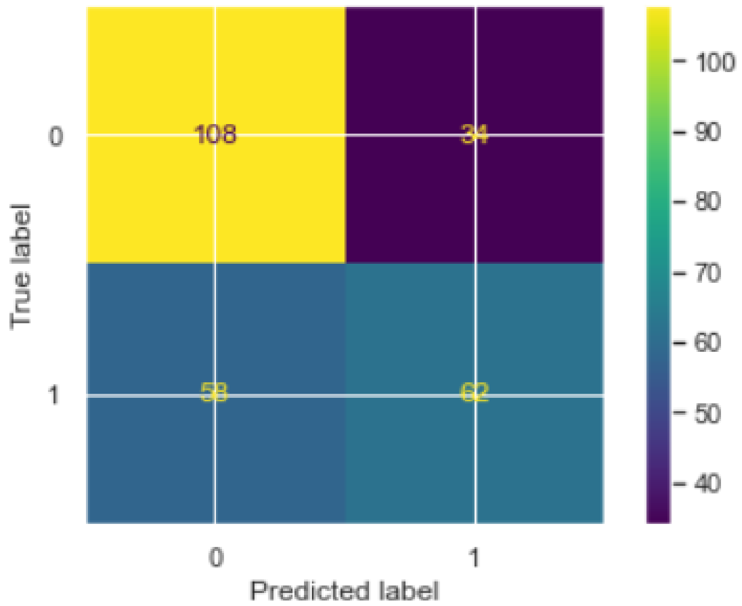


Figure 60. X train and y train model score (Using Gridsearch with best parameters)

0.659016393442623

Figure 61. X test and y test model score (Using Gridsearch with best parameters)

0.6641221374045801

Figure 62. ytest_predict_prob (Using Gridsearch with best parameters)

	0	1
0	0.668853	0.331147
1	0.627745	0.372255
2	0.681034	0.318966
3	0.586411	0.413589
4	0.557742	0.442258

Train label AUC score: 0.729

Test label AUC score: 0.729

Figure 63. Train label ROC curve (Using Gridsearch with best parameters)

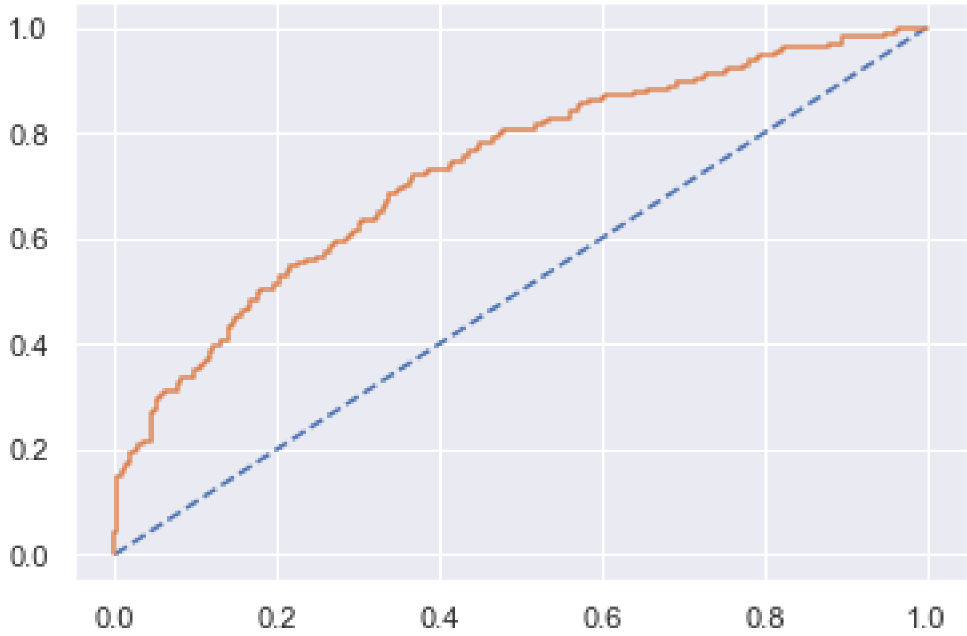
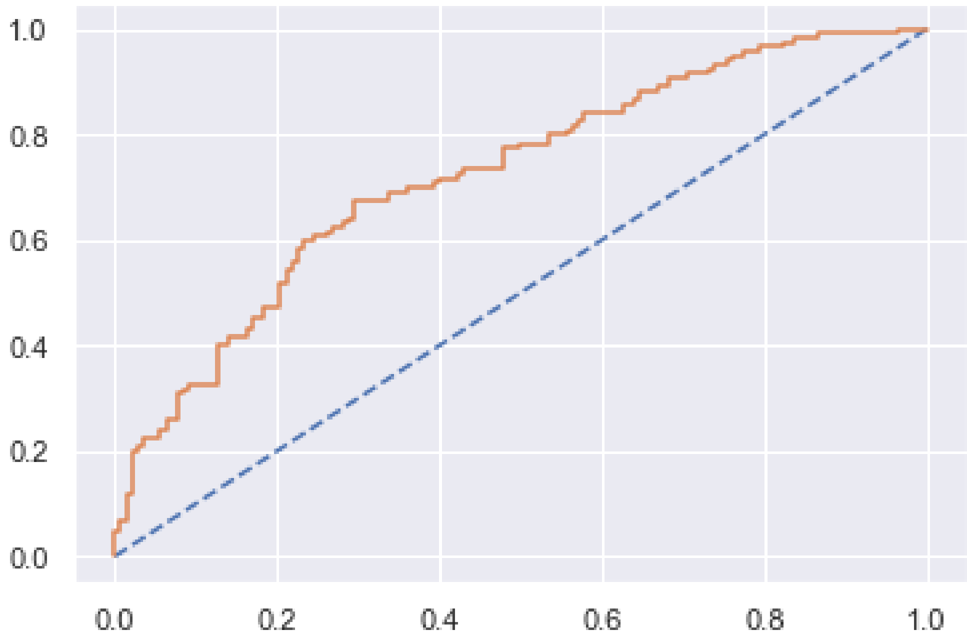


Figure 64. Test label ROC curve (Using Gridsearch with best parameters)



2.2.3.1.3 Classification Report (Using Gridsearch with best parameters)

Train Label:

	precision	recall	f1-score	support
0	0.66	0.76	0.71	329
1	0.66	0.54	0.59	281
accuracy			0.66	610
macro avg	0.66	0.65	0.65	610
weighted avg	0.66	0.66	0.65	610

Test Label:

	precision	recall	f1-score	support
0	0.66	0.79	0.72	142
1	0.67	0.52	0.58	120
accuracy			0.66	262
macro avg	0.67	0.65	0.65	262
weighted avg	0.67	0.66	0.66	262

2.2.3.1.4 Confusion Matrix (Using Gridsearch with best parameters)

Figure 65. Train label

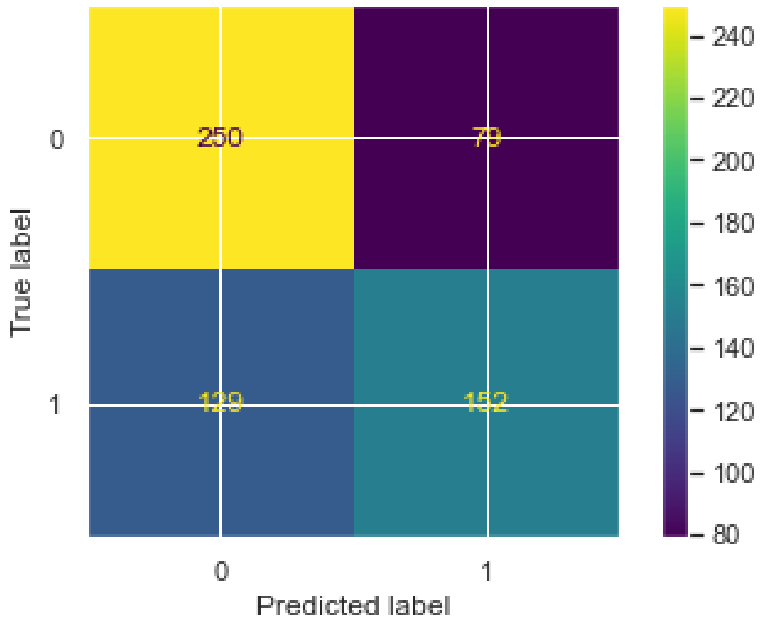
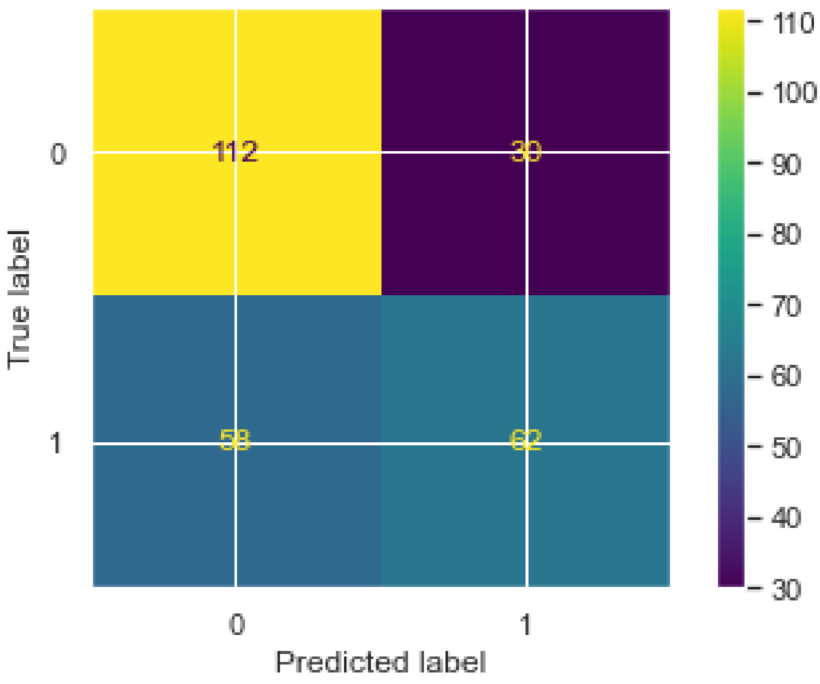


Figure 66. Test label



2.2.3.2 Linear Discriminant Analysis (LDA)

Train label AUC score: 0.730

Test label AUC score: 0.730

Figure 67. Train label ROC curve

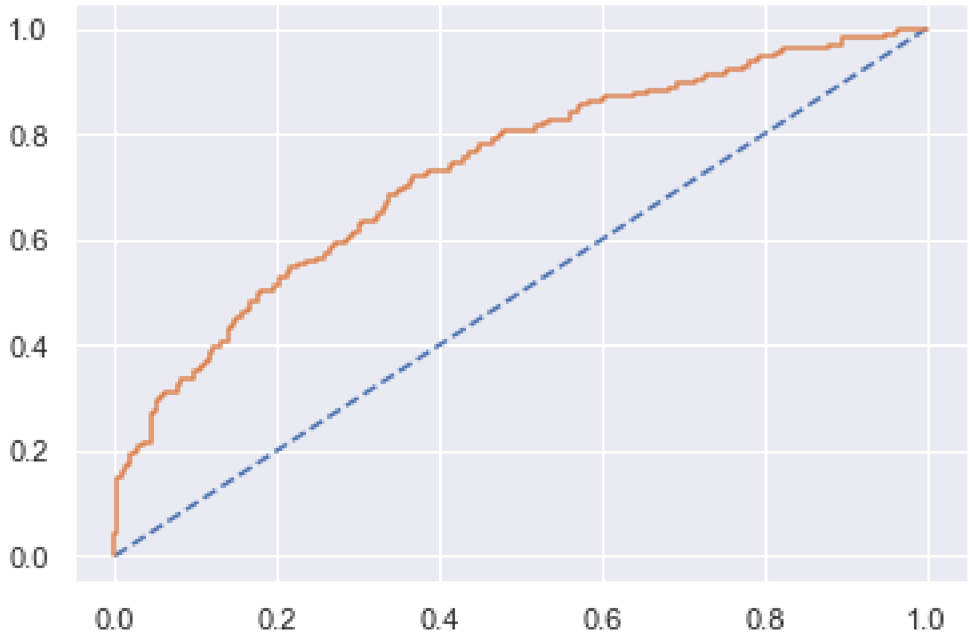
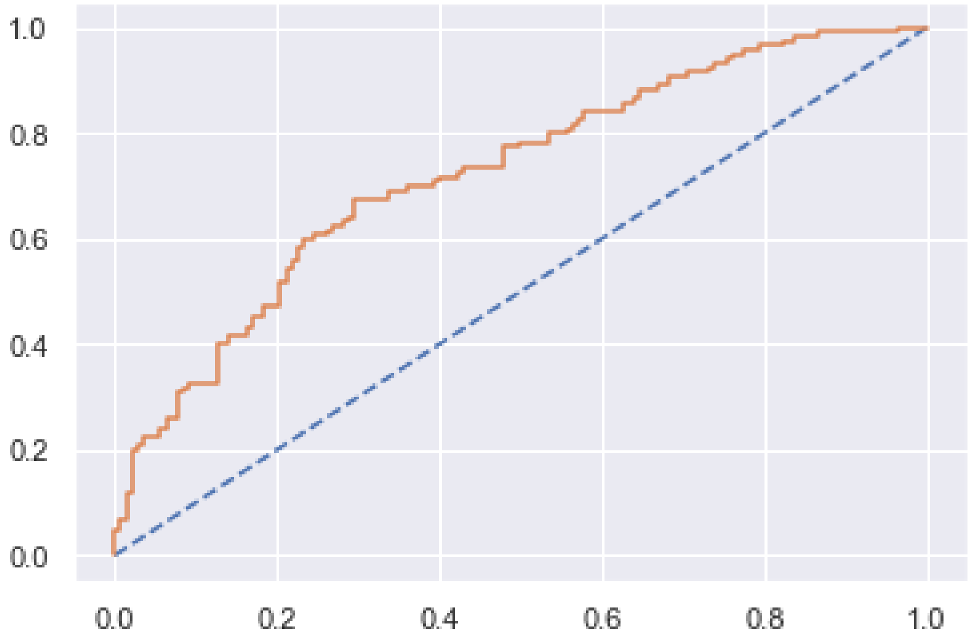


Figure 68. Test label ROC curve



2.2.3.2.1 Classification Report

Train Label:

	precision	recall	f1-score	support
0	0.67	0.77	0.71	329
1	0.67	0.56	0.61	281
accuracy			0.67	610
macro avg	0.67	0.66	0.66	610
weighted avg	0.67	0.67	0.66	610

Test Label:

	precision	recall	f1-score	support
0	0.67	0.79	0.72	142
1	0.68	0.53	0.60	120
accuracy			0.67	262
macro avg	0.67	0.66	0.66	262
weighted avg	0.67	0.67	0.67	262

Final Model: Compare all the models and write an inference which model is best/optimized:

Table 8 Comparative Analysis for Three Models

Scores	Dataset	Logistic Regression		Logistic Regression (Best Parameters)		LDA	
		0	1	0	1	0	1
Accuracy	**Train**	0.67		0.66		0.67	
	Test	0.65		0.66		0.67	
Recall	**Train**	0.74	0.58	0.76	0.54	0.77	0.56
	Test	0.76	0.52	0.79	0.52	0.79	0.53
Precision	**Train**	0.67	0.66	0.66	0.66	0.67	0.67
	Test	0.65	0.65	0.66	0.67	0.67	0.68
F1 Score	**Train**	0.71	0.62	0.71	0.59	0.71	0.61
	Test	0.70	0.57	0.72	0.58	0.72	0.60
AUC Score	**Train**	0.733		0.729		0.730	
	Test	0.733		0.729		0.730	

Note: 0 = No
1 = Yes

Linear Discriminant Analysis (LDA) will be the best model for the given dataset.

Looking at the aforementioned table, we can see that the LDA model has better accuracy; and better recall, precision, and F1 score for test data. As recall is a ratio between true positives and false negatives, so recall value closer to 1 means depicts better model performance. In addition, as F1 score helps in classification of positives and negatives, higher F1 score means better model performance. As a result, we should consider LDA model for the given dataset.

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

- Most employees over the age of 50 do not choose holiday packages. They don't seem interested in holiday packages at all
- Employees who are in the age gap of 30 to 50 years choose holiday packages. Young people seem to believe that I spend on package holidays, so age plays a very important role here in deciding whether to choose a package or not.
- People who have salary less than 50000 choose holiday packages. So salary is also a deciding factor for the holiday package. Education also plays an important role in deciding on holiday packages.
- As we already have a customer base between the ages of 30 and 50, we need to look for opportunities to target older people and people who earn more than 150,000
- As we know, most of the elderly people prefer to visit religious places, so it would be better if we target these places and provide them with packages where they can visit religious places
- We can also look at the family dynamics of elderly people, if elderly people have older children, e.g. 30 to 40 years old, they can take advantage of holiday packages, so the deal should include a family package
- People who earn more than 150,000 don't spend much on vacation packages, tend to go on lavish vacations and we can provide them with customized packages according to their wishes like luxury hotels, longer vacations, private cars during vacations to attract such employees

The End