PREDICTIVE MODELING

Predictive modeling is about building models for regression and classification business problem.
Linear Regression – This model is used to predict continuous target variable.
Logistic Regression – This model is used to predict Classification target variable.
Linear Discriminant Analysis – This model is use to predict classification target variable as well as dimension reduction.

BY: SAHIL SAXENA

Table of Contents

PROBLEM – 1. Linear Regression 2

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. 3

Q-1.1. Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA, duplicate values). Perform Univariate and Bivariate Analysis 3

Q-1.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Check for the possibility of combining the sub levels of a ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning. 12

Q-1.3 Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning. 14

Q-1.4 Inference: Basis on these predictions, what are the business insights and recommendations. 17

Problem 2: Logistic Regression and LDA. 18

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. 18

Q-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. 18

Q -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). 25

Q-2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized. 29

Q-2.4 Inference: Basis on these predictions, what are the insights and recommendations. 33

TABLE 1. 1 – Top 5 Rows 3

TABLE 1. 2 – Top 5 Rows of after dropping Unnamed:0 4

TABLE 1. 3 – Data Information 4

TABLE 1. 4 – Data Description 5

TABLE 1. 5 – 0 Value in data. 13

TABLE 1. 6 – Data After Encoding...14

TABLE 1. 7 – Model Comparison 15

TABLE 1. 8 – Model Stats Report. 16

Table 2. 1 – Top 5 rows of data 18

Table 2. 2 top 5 rows after dropping 19

Table 2. 3 – Data Information 19

Table 2. 4 – Data Description 20

Table 2. 5 – Top 5 Rows after Encoding 25

Table 2. 6 – Probability Prediction for 0's and 1's for Logistic Regression 27

Table 2. 7 - Probability Prediction for 0's and 1's for LDA 28

Table 2. 8 – Model Comparison...33

Figure 1. 1 – Box Plot of Variables...6

Figure 1. 2 – Distribution plot of variable 7

Figure 1. 3 – Correlation Plot 8

Figure 1. 4 – Pair Plot 9

Figure 1. 5 – Count plot of Variables 10

Figure 1. 6 – Bivariate Plots of Independent Variable. 11

Figure 1. 7 – Box Plot after outlier treatment. 12

Figure 1. 8 – Plots for Business Insights 17

Figure 2. 1 – Distribution plot 20

Figure 2. 2 – Boxplot of independent numeric variables...21

Figure 2. 3 – Count plot of Categorical Independent variables. 22

Figure 2. 4 – Box plot Holiday Package Vs Salary and Holiday package Vs age 22

Figure 2. 5 – Plot of Holiday Package Vs Education and Holiday Vs Foreign. 23

Figure 2. 6 - Plot of Holiday Package Vs Number of young children and Holiday package Vs Number of old children ...23

Figure 2. 7 – Pair Plot of Independent Variables 24

Figure 2. 8 – Correlation plot of Independent Variables. 25

Figure 2. 9 - Holiday Package Vs Salary and Holiday Package Vs Number of young Children 33

Figure 2. 10 - Holiday Package Vs Education and Holiday Package Vs Foriegn 34

PROBLEM – 1. Linear Regression

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

Data Dictionary:

Variable Name and their Description

Carat: Carat weight of the cubic zirconia.

Cut: Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal.

Color: Colour of the cubic zirconia. With D being the worst and J the best.

Clarity: Clarity refers to the absence of the Inclusions and Blemishes. (In order from Worst to Best in terms of avg price) IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, l1

Depth: The Height of cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter.

Table: The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter.

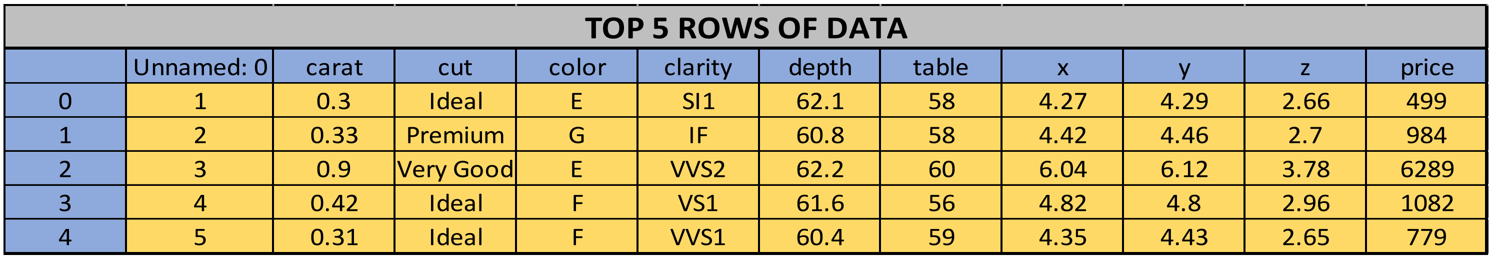
Price: The Price of the cubic zirconia.

X: Length of the cubic zirconia in mm.

Y: Width of the cubic zirconia in mm.

Z: Height of the cubic zirconia in mm.

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



TOP 5 ROWS OF DATA

	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
0	1	0.3	Ideal	E	SI1	62.1	58	4.27	4.29	2.66	499
1	2	0.33	Premium	G	IF	60.8	58	4.42	4.46	2.7	984
2	3	0.9	Very Good	E	VVS2	62.2	60	6.04	6.12	3.78	6289
3	4	0.42	Ideal	F	VS1	61.6	56	4.82	4.8	2.96	1082
4	5	0.31	Ideal	F	VVS1	60.4	59	4.35	4.43	2.65	779

TABLE 1. 1 – Top 5 Rows

THE UNNAMED : 0 COLUMN IS JUST AN INDEX REPRESENTATION OF THE DATA WHICH IS OF NO USE. SO WE ARE DROPPING THE COLUMN FROM FURTHER ANALYSIS.

Top 5 rows are shown below after dropping the Unnamed :0 column from the data frame.

TOP 5 ROWS OF DATA

	carat	cut	color	clarity	depth	table	x	y	z	price
0	0.3	Ideal	E	SI1	62.1	58	4.27	4.29	2.66	499
1	0.33	Premium	G	IF	60.8	58	4.42	4.46	2.7	984
2	0.9	Very Good	E	VVS2	62.2	60	6.04	6.12	3.78	6289
3	0.42	Ideal	F	VS1	61.6	56	4.82	4.8	2.96	1082
4	0.31	Ideal	F	VVS1	60.4	59	4.35	4.43	2.65	779

TABLE 1. 2 – Top 5 Rows of after dropping Unnamed:0

EXPLORATORY DATA ANALYSIS (EDA)

DATA INFORMATION

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26967 entries, 0 to 26966
Data columns (total 11 columns):

#	Column	Count	Non-Null	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(2), object(3)
memory usage: 2.3+ MB

TABLE 1. 3 – Data Information

As the data information

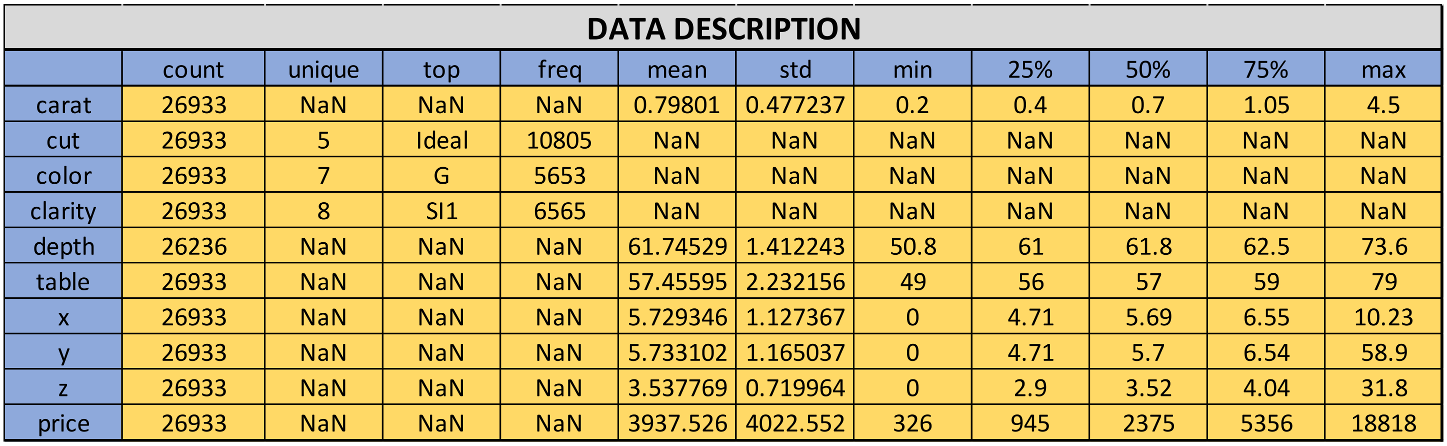
1. data frame has 26967 rows and 11 columns.
2. There is missing value present in the data frame.
3. There are three types of data types present int64(2), float64(6), object (3)
4. All the columns datatype are correct as per their data so there no anomalies present in the data frame.
5. This data frame consumes around 2.3+ MB for storage.

```
MISSING VALUES

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

There is only one column 'depth' which has 697 missing values present. So, we need to check it and impute as per the information

DUPLICATE ROWS

There are 34 duplicate rows present in the data and as these duplicate rows can influence my model performance and make it biased towards duplicate rows. So, for further analysis we will be dropping duplicate rows.



DATA DESCRIPTION

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
carat	26933	NaN	NaN	NaN	0.79801	0.477237	0.2	0.4	0.7	1.05	4.5
cut	26933	5	Ideal	10805	NaN	NaN	NaN	NaN	NaN	NaN	NaN
color	26933	7	G	5653	NaN	NaN	NaN	NaN	NaN	NaN	NaN
clarity	26933	8	SI1	6565	NaN	NaN	NaN	NaN	NaN	NaN	NaN
depth	26236	NaN	NaN	NaN	61.74529	1.412243	50.8	61	61.8	62.5	73.6
table	26933	NaN	NaN	NaN	57.45595	2.232156	49	56	57	59	79
x	26933	NaN	NaN	NaN	5.729346	1.127367	0	4.71	5.69	6.55	10.23
y	26933	NaN	NaN	NaN	5.733102	1.165037	0	4.71	5.7	6.54	58.9
z	26933	NaN	NaN	NaN	3.537769	0.719964	0	2.9	3.52	4.04	31.8
price	26933	NaN	NaN	NaN	3937.526	4022.552	326	945	2375	5356	18818

TABLE 1. 4 – Data Description

DATA DESCRIPTION

1. The minimum carat diamond in the data is 0.2 carat and maximum are 4.5 carat diamond.
2. Maximum diamond data present belongs the best cut category which is ideal.
3. The minimum and the maximum depth of the diamond is 50.8 to 73.6.
4. In x, y, z variable min dimension value present is 0 which is not possible, so we need check this data whether to impute it or drop it.
5. The starting price of diamond is 326 and can goes up to 18818 for a good quality diamond.

UNIVARIATE ANALYSIS.

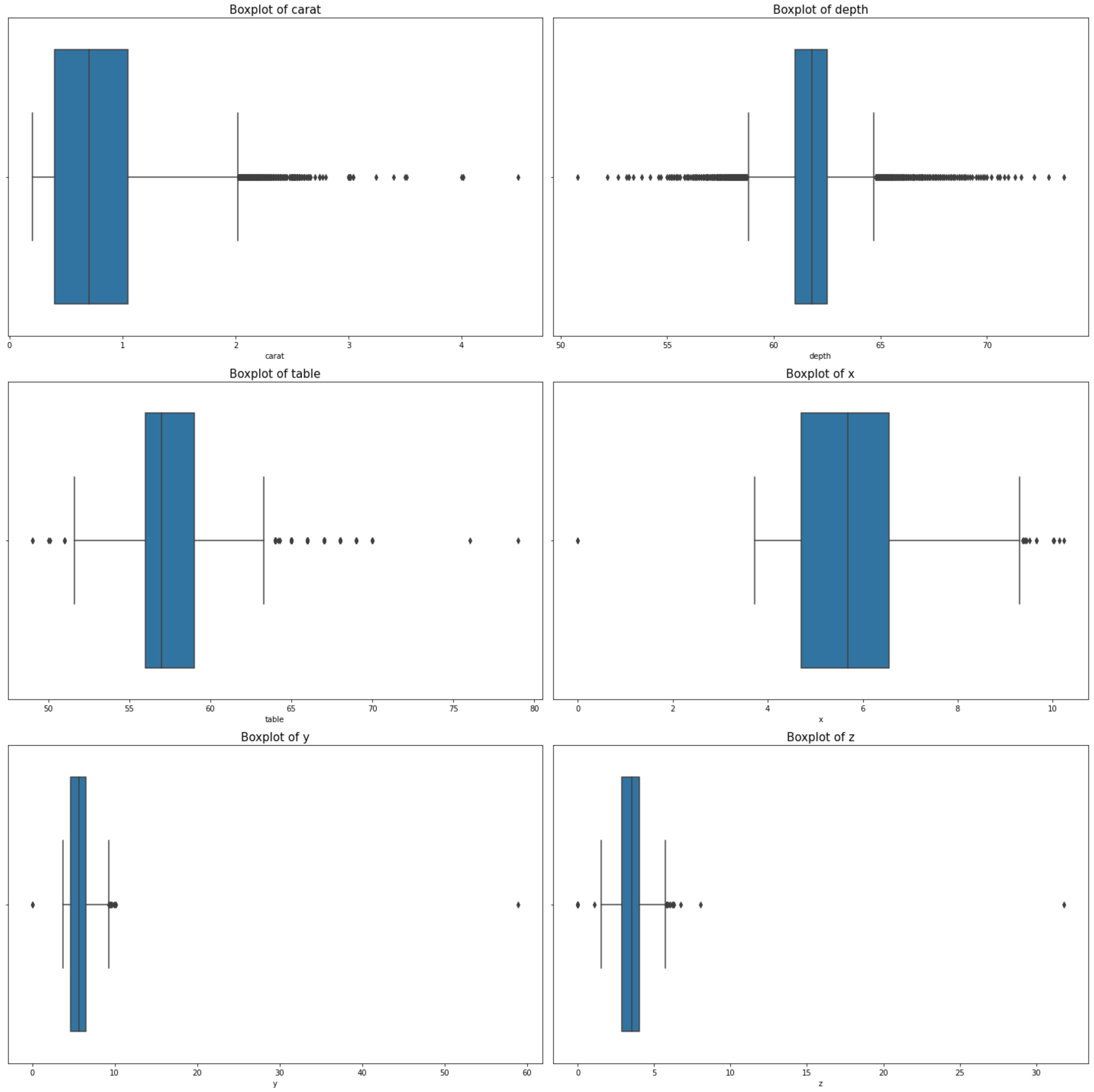


Figure 1. 1 – Box Plot of Variables

As we can check from variables box plot that all the variables have outliers.

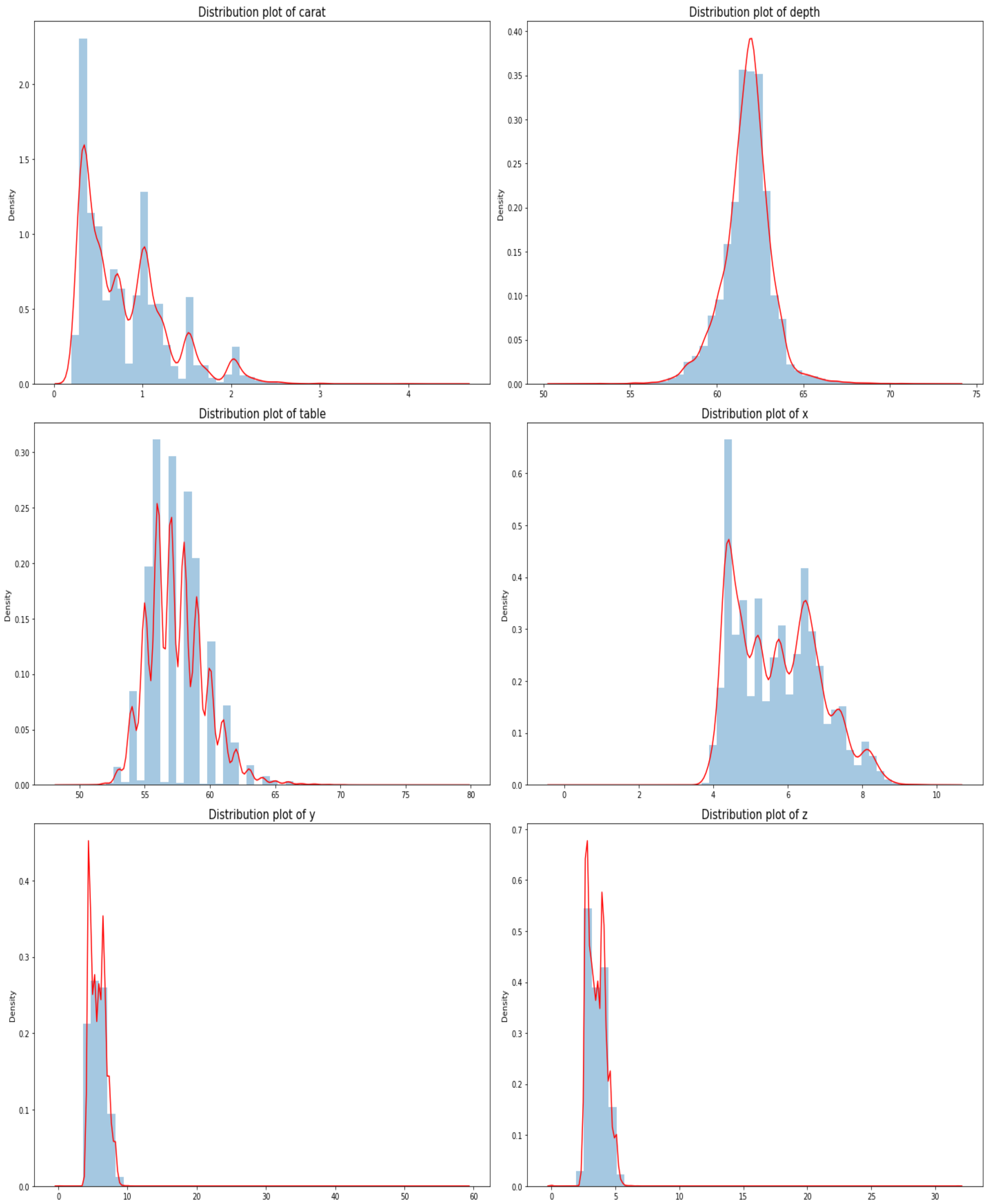


Figure 1. 2 – Distribution plot of variable

```
Skewness of carat is 1.1147891
Skewness of depth is -0.026085
Skewness of table is 0.7658050
Skewness of x is 0.39228983097
Skewness of y is 3.86776408706
Skewness of z is 2.58066466754
```

The above plot and stats represent,

1. The distribution of depth column is almost normal as its skewness value is almost near to zero.
2. The X column is moderately right skewed.
3. Rest all the columns are highly

BIVARIATE ANALYSIS.

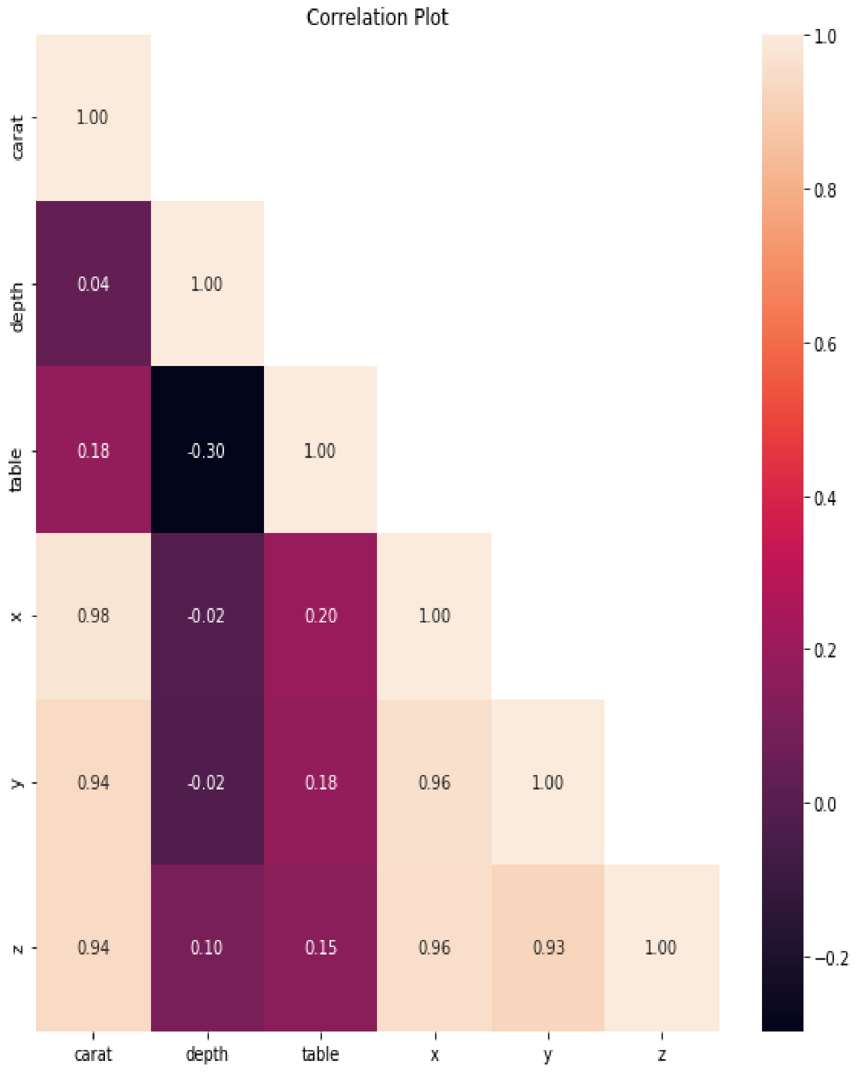


Figure 1. 3 – Correlation Plot

From the Correlation plot we can infer the same result as of pair plot

1. PRICE is highly correlated with X, Y, Z and CARAT, so can be good predictors for price.
2. CARAT is also high positively correlated with X, Y, Z.

PAIR PLOT

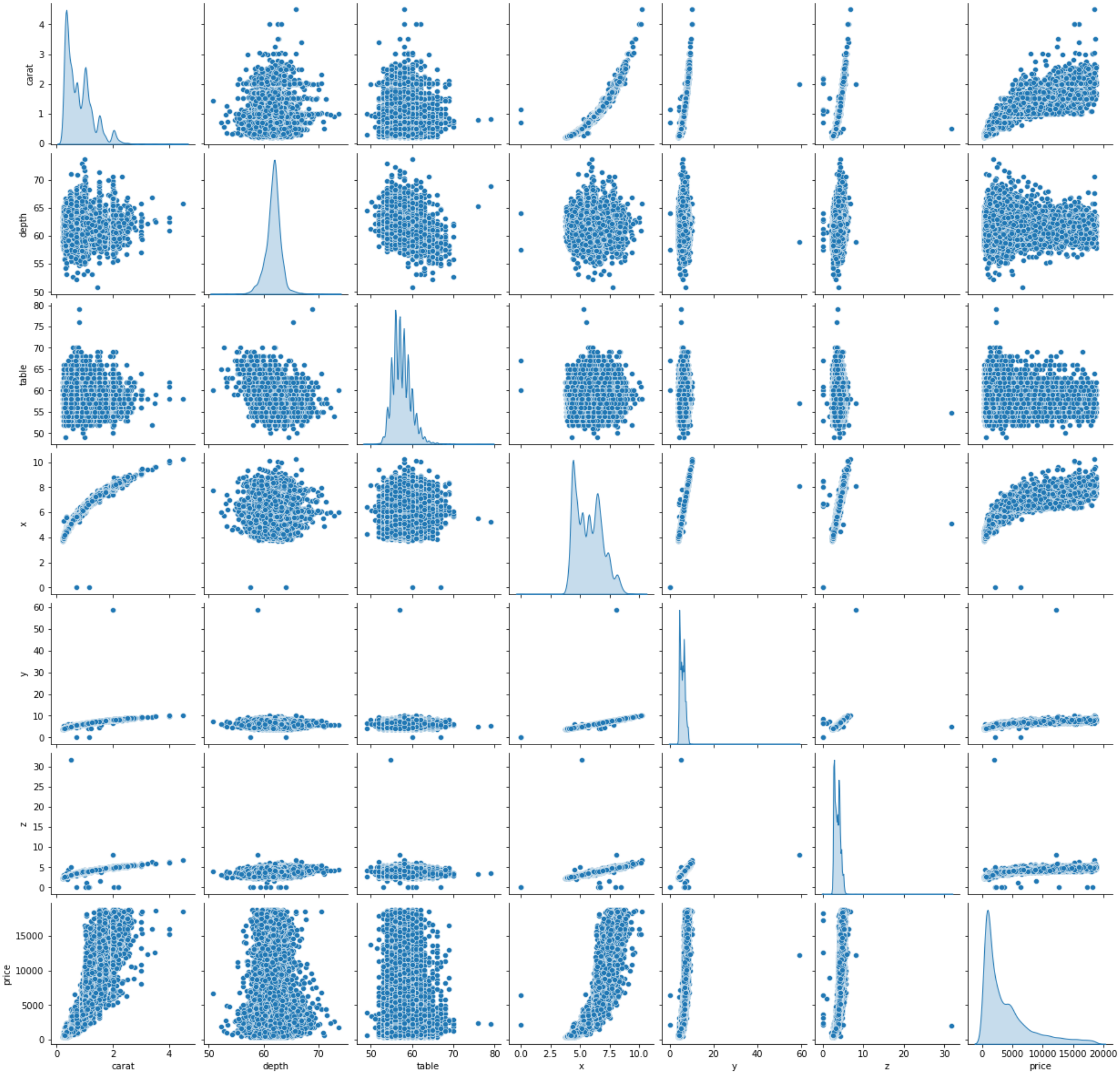


Figure 1. 4 – Pair Plot

From the pair plot we can check that,

1. Price is having linear relation with many independent variables such as X, Y, Z and CARAT.
2. CARAT is also highly related with X, Y, Z.
3. This relation indicates that these variables might play very important role in predicting the prices of the stones.

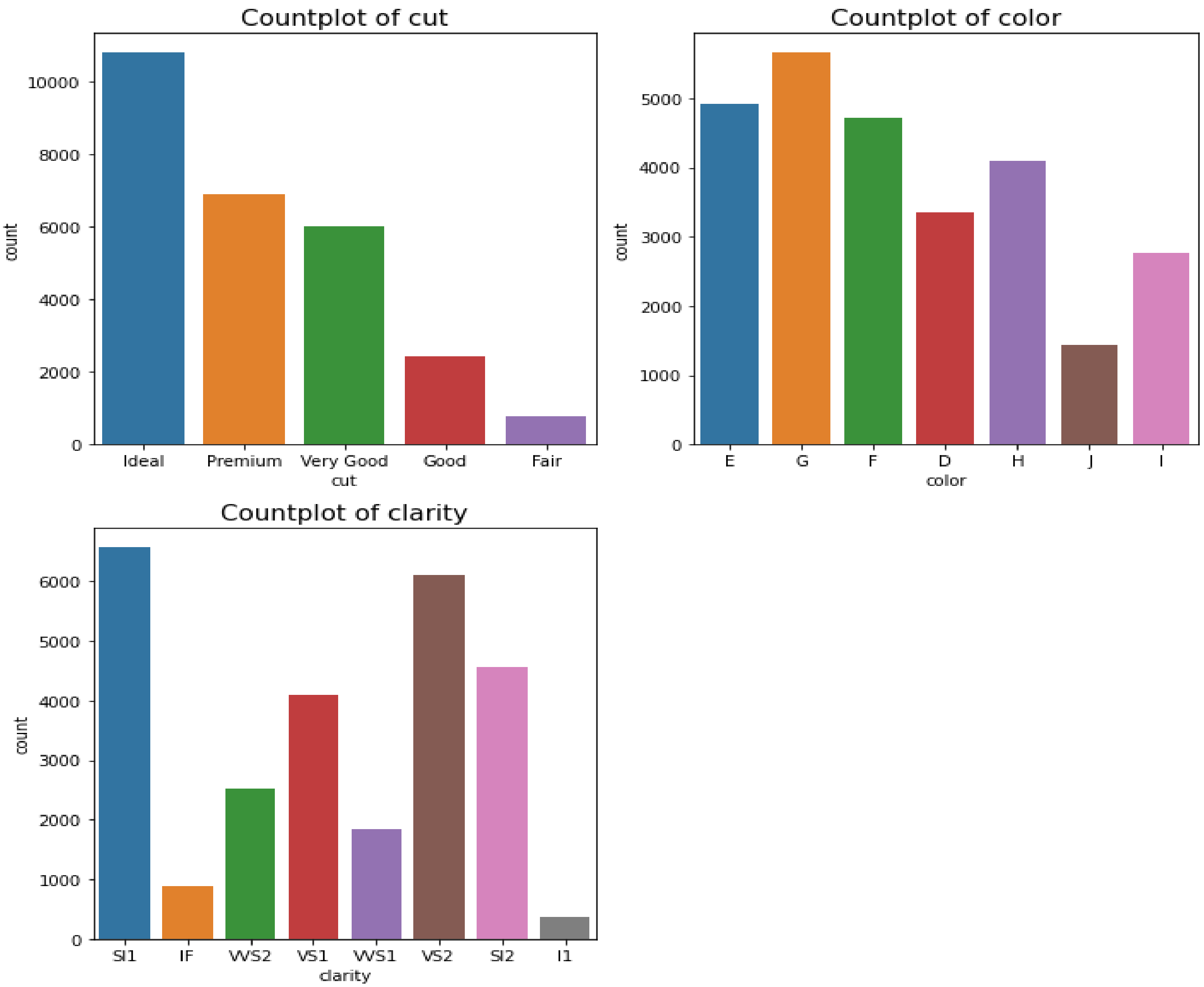


Figure 1. 5 – Count plot of Variables

From the above plot we can check,

1. For cut, Ideal is the most selling cut diamond among other cuts and fair is the least sold diamond cuts.
2. G color diamond are the most selling diamond, it is moderately color in ranking.
3. SI1 is the most selling clarity of diamond among others, I1 is the best clarity but still its least selling.

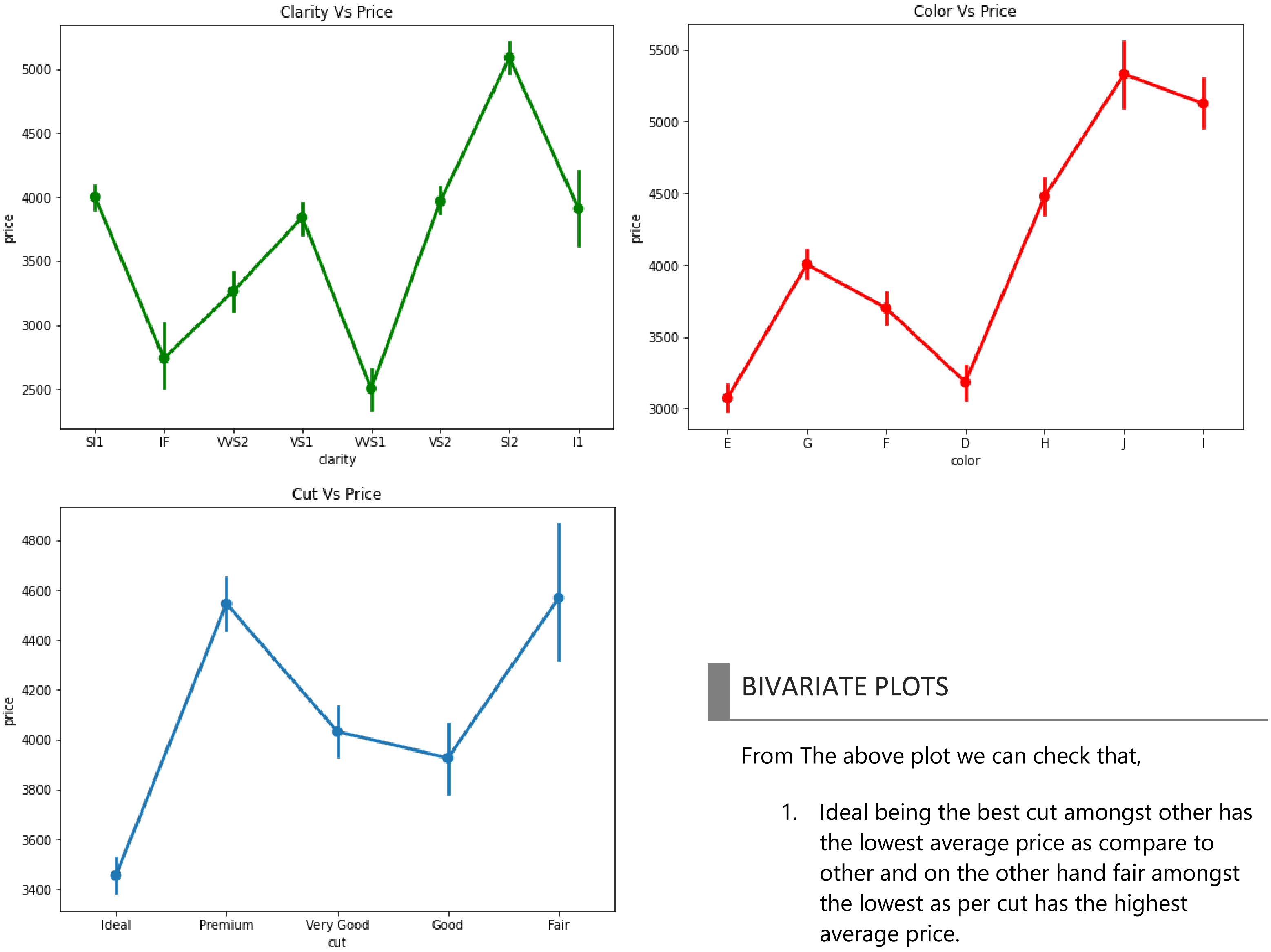


Figure 1. 6 – Bivariate Plots of Independent Variable.

BIVARIATE PLOTS

From The above plot we can check that,

1. Ideal being the best cut amongst other has the lowest average price as compare to other and on the other hand fair amongst the lowest as per cut has the highest average price.
2. As the order of color of stones goes from lower to higher (D --> J) the average price also increase with the quality if color.
3. Same as Color as the quality of clarity of stones increasing so do the price of the stones.

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

Solution:

As linear regression model is sensitive to the outliers and every independent continuous variables have outliers present, First we will treat the outliers and then imputing the missing values.

OUTLIER TREATMENT.

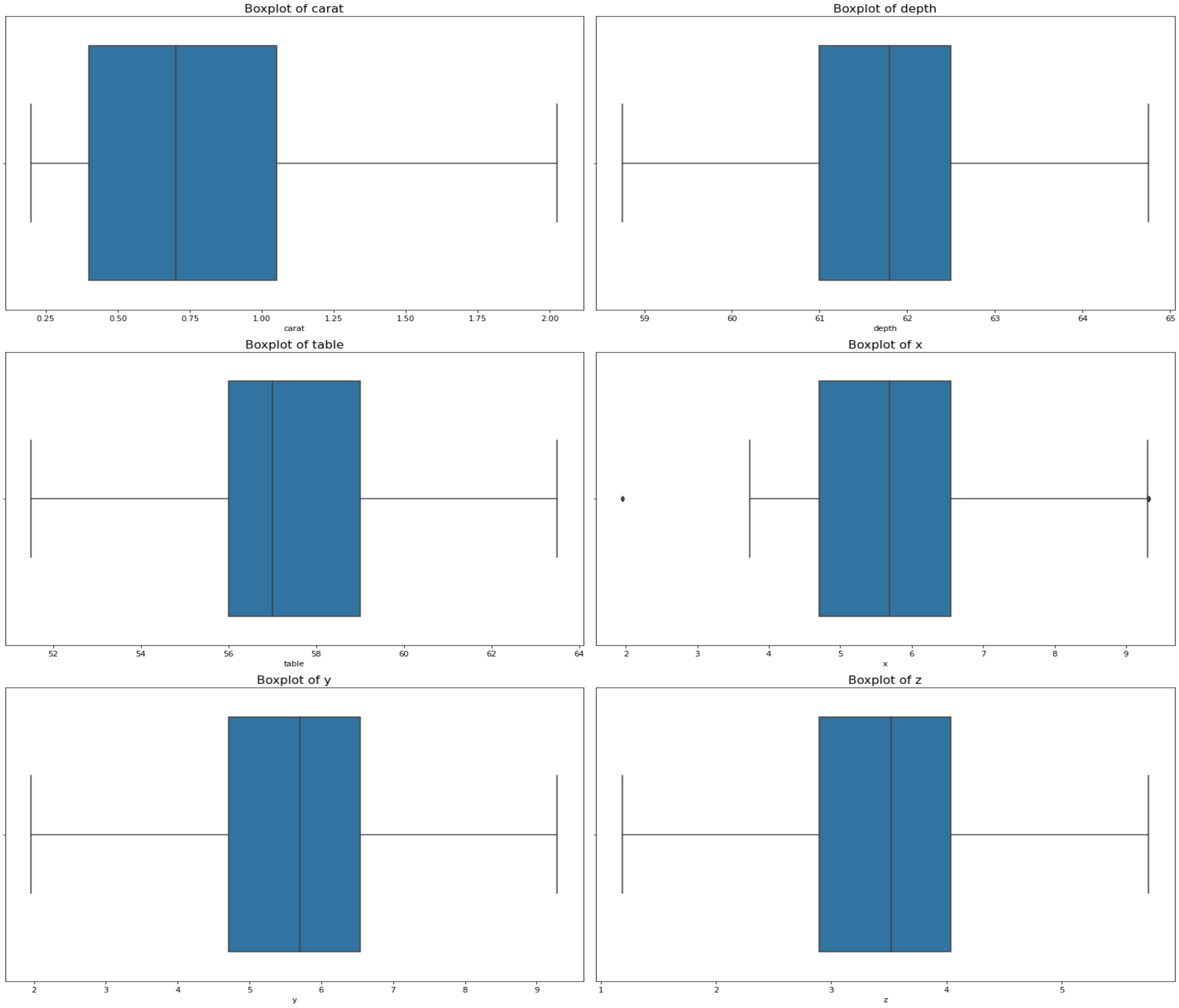


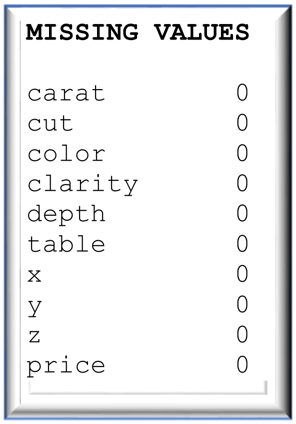
Figure 1. 7 – Box Plot after outlier treatment.

As we can check from the above figure of boxplot that outliers have been treated properly using upper and lower range

MISSING VALUE TREATMENT.

Now as we have treated the outliers in the data, we can use mean of each variable to fill the null values in each variable.

As only depth column has missing value present in them, so let’s fill those null values,



```
MISSING VALUES

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

As we can check there are no more missing value present in the data frame.

*** NOTE***

For filling up the null values we have used fillna function using mean of depth column.

0 VALUES IN DATA.

0 VALUES IN DATA										
	carat	cut	color	clarity	depth	table	x	y	z	price
5821	0.71	Good	F	SI2	64.1	60	0	0	0	2130
6034	2.02	Premium	H	VS2	62.7	53	8.02	7.95	0	18207
10827	2.2	Premium	H	SI1	61.2	59	8.42	8.37	0	17265
12498	2.18	Premium	H	SI2	59.4	61	8.49	8.45	0	12631
12689	1.1	Premium	G	SI2	63	59	6.5	6.47	0	3696
17506	1.14	Fair	G	VS1	57.5	67	0	0	0	6381
18194	1.01	Premium	H	I1	58.1	59	6.66	6.6	0	3167
23758	1.12	Premium	G	I1	60.4	59	6.71	6.67	0	2383

TABLE 1. 5 – 0 Value in data.

As we can see in the table Variable X, Y, Z have some 0 values in them, also showing in the data description table as minimum value.

As we treated outliers in the data these values also got treated and changed these 0 to the lower limit of the distributions. Now no more 0 values present in the data frame.

SUB LEVEL COMBINING.

There are number of ordinal data present in the data which can be combined. There are many possibilities of combining or not combining the sublevel. We chose not to combine the data as these are ordinal data and average price of each category is different.

It will not be fair to combine the data all these categories could be good predictors of the price, so I will be moving forward without combining the different sublevels.

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

Solution:

ENCODING THE DATA

As the data present are in ordinal form, so we will be numbering them as per their label and converting them into continuous variables. 1 is the worst or lowest

For Cut,

Fair = 1
Good = 2
Very Good = 3
Premium = 4
Ideal = 5

For Color

D – 1
E – 2
F – 3
G – 4
H –5
I – 6
J – 7

For Clarity,

IF – 1
VVS1 – 2
VVS2 – 3
VS1 – 4
VS2 – 5
SI2 – 6
SI1 – 7
I1 – 8

AFTER ENCODING

	carat	cut	color	clarity	depth	table	x	y	z
0	0.3	5	2	6	62.1	58	4.27	4.29	2.66
1	0.33	4	4	1	60.8	58	4.42	4.46	2.7
2	0.9	3	2	3	62.2	60	6.04	6.12	3.78
3	0.42	5	3	4	61.6	56	4.82	4.8	2.96
4	0.31	5	3	2	60.4	59	4.35	4.43	2.65

TABLE 1. 6 – Data After Encoding

DATA INFORMATION

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26967 entries, 0 to 26966
Data columns (total 11 columns):

#	Column	Count	Non-Null	Dtype
---	------	--------------	-----	-----
0	carat	26967	non-null	float64
1	cut	26967	non-null	int64
2	color	26967	non-null	int64
3	clarity	26967	non-null	int64
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(2), object(3)
memory usage: 2.3+ MB

We have converted all in object data type into int 64 for further creating a model.

SPLITTING THE DATA INTO TRAIN AND TEST DATA

X_train (18853, 9)
X_test (8080, 9)
y_train (18853, 9)
y_test (8080,)

For Train and Test split we have used train and test split from sklearn.

X_train – Contains independent variables (predictors) for training the model.

X_test – Contains independent variable for testing the model.

y_train – Contains dependent variable (Target) for training the model.

y_test – contains dependent variable for testing the model.

MODEL COMPARISON					
MODEL	RMSE		R-squared		R_Adjusted
	TRAIN	TEST	TRAIN	TEST	
Model - 1	1162.002	1156.824	0.916	0.918	0.916
Model -2	1318.64	1310.85	0.8917	0.8936	0.892
Model - 3	1215.8	1208.66	0.9079	0.9096	0.908
Model - 4	1269.007	1263.814	0.899	0.902	0.9

TABLE 1. 7 – Model Comparison

As we can check from the model comparison,

RMSE

Model 1 RMSE score for train and test is 1162.002 and 1156.824, Which is very less as compare to other models. So, in RMSE model 1 performance is best.

R – SQUARED

For model 1 R-squared value for train and test is 0.916 and 0.918, which the best amongst any other model also. So, we can Model 1 is the best and robust model.

R – ADJUSTED

For R- adjusted also model 1 performance is the best with 0.916 value.

CONCLUSION

"Model 1" is the best in all the evaluation parameters, so for this business model we will be choosing model 1.

MODELS

Model 1

1st model is created after treating outliers and no sublevel category combination is performed.

Model 2

2nd model is created after just sublevel category combining in to each other.

Model 3

3rd model is created with no outlier treatment in data and no combination of sublevel categories in categorical variables.

Model 4

4th model is created with Outlier treatment in data and combining sublevel of categories of categorical variables.

Following is stats report for the chosen model i.e., Model -1

OLS Regression Results

==

Dep.Variable:	price	R-squared:	0.916
Model:	OLS	Adj.R-squared:	0.916
Method:	Least Squares	F-statistic:	2.28E+04
Date:	Sun,19 dec 2021	Prob (F-statistic):	0
Time:	13:46:45	Log-Likelihood:	-1.598E+05
No.Observations:	18853	AIC:	3.20E+05
Df Residuals:	18843	BIC:	3.20E+05
Df Model:	9		
Covariance Type:	nonrobust		

==

coef	std	err	t	P>|t|	[0.025	0.975]
--						
Intercept	1.15E+04	889.686	12.96	0	9786.888	1.33E+04
carat	1.35E+04	103.838	130.315	0	1.33E+04	1.37E+04
cut	135.2292	9.379	14.419	0	116.846	153.612
color	-333.2608	5.254	-63.426	0	-343.56	-322.962
clarity	-485.1131	5.701	-85.089	0	-496.288	-473.938
depth	-50.0758	11.408	-4.39	0	-72.436	-27.716
table	-27.2071	4.994	-5.448	0	-36.996	-17.418
x	-2589.3989	153.183	-16.904	0	-2889.652	-2289.146
y	1301.4601	151.332	8.6	0	1004.835	1598.085
z	-910.8909	123.294	-7.388	0	-1152.559	-669.223

==

Omnibus:	3630.861	Durbin-Watson:	1.973
Prob(Omnibus):	0	Jarque-Bera (JB):	32131.902
Skew:	0.675	Prob(JB):	0
Kurtosis:	9.252	Cond.No.	8.99E+03

==

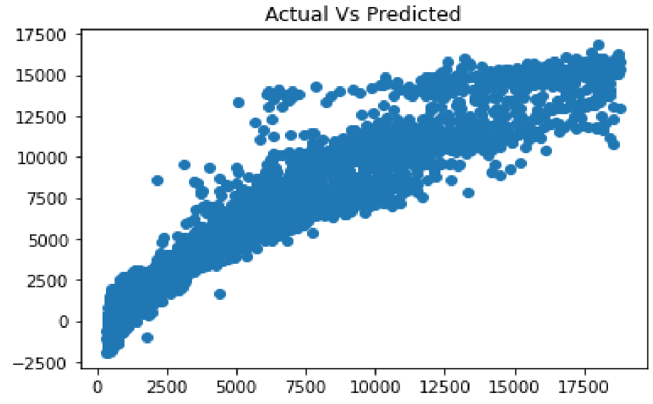
Notes:

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

[2] The condition number is large, 8.99e+03. This might indicate that there are strong multicollinearity or other numerical problems.

TABLE 1. 8 – Model Stats Report.

From the above report we can check the p value for all variable is less than 0.05, So we can say t hat we reject the null hypothesis that at least one variable is influencing the target variable.



This can be interpreted as if there is 1 unit increase in carat there will be 13531.598 unit will increase in price.

Similarly, if there is 1 unit will increase in x there will be 2589.399 unit will decrease in price.

Most influencing factors are

Carat, x, y, z, clarity

y = 115303.754 + (13531.598 * carat) + (135.23 * cut) + (-333.26 * color) + (-485.113 * clarity) + (-50.075 * depth) + (-27.207 * table) + (-2589 .399 * x) + (1301.460 * y) + (-910.89 * z)

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

Solution:

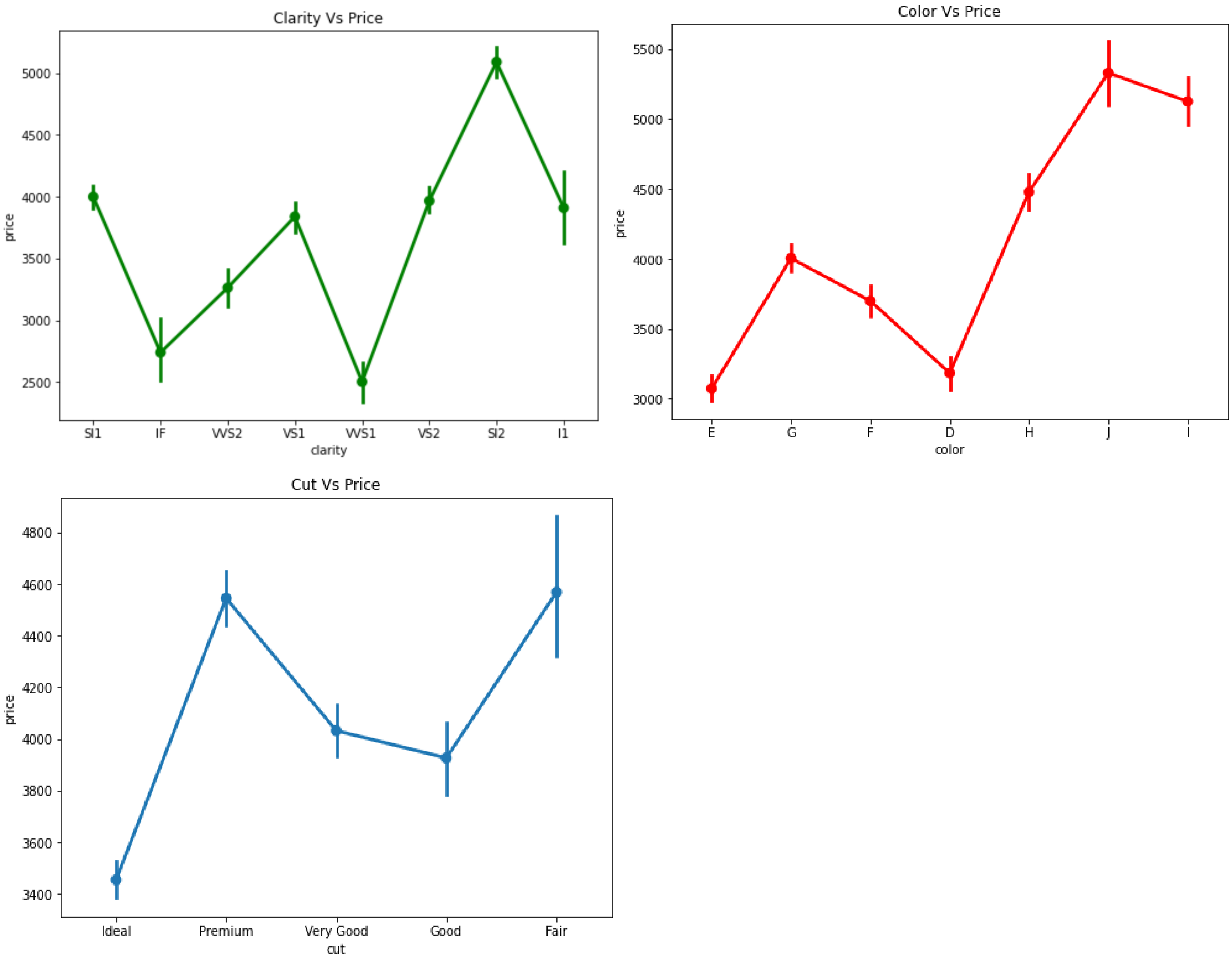


Figure 1. 8 – Plots for Business Insights

Business Insights

1. As we can check ideal is highest quality cut but still its average price is very less, we can try to increase the price of ideal cut diamonds.
2. As we saw from the model carat is most influencing the price, we can check which low quality diamonds have good carat and we can sell them with larger profit margin.
3. Ideal and premium cut diamonds are the highest selling diamonds so we can increase the quality and price for these diamonds
4. There is different combination like ideal cut, moderate color and moderate clarity levels have very high price in the market, we can look for this kind of combination and sell them at good price for higher profit margins.
5. Fair cut diamonds have very high average price in the market so we can look for different combinations of dimensions and quality with that and sell them at good profit.

Problem 2: Logistic Regression and LDA.

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.

Data Dictionary:

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

Q-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.

Solution:

TOP 5 ROWS OF DATA FRAME

	Unnamed: 0	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	1	no	48412	30	8	1	1	no
1	2	yes	37207	45	8	0	1	no
2	3	no	58022	46	9	0	0	no
3	4	no	66503	31	11	2	0	no
4	5	no	66734	44	12	0	2	no

Table 2. 1 – Top 5 rows of data

The Unnamed :0 column is a redundant column and this will not help in any of the data analysis, so we dropping this column.

After dropping the column top 5 rows of the data frame.

TOP 5 ROWS OF DATA FRAME							
	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 2. 2 top 5 rows after dropping

EXPLORATORY DATA ANALYSIS (EDA)

1. There 872 rows and 08 columns present in the data frame.
2. There are no null values present in data frame.
3. There are two types of datatypes present int64 (6), object (2)
4. As all the datatype is correct as per the data so there are no anomalies in the data frame.
5. The memory usage for this data frame is 54.6+ KB.

DATA INFORMATION

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
Data columns (total 8 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(6), object(2)
memory usage: 54.6+ KB
```

Table 2. 3 – Data Information

DUPLICATE ROWS

THERE ARE NO DUPLICATE ROWS PRESENT IN THE DATA FRAME.

DATA DESCRIPTION

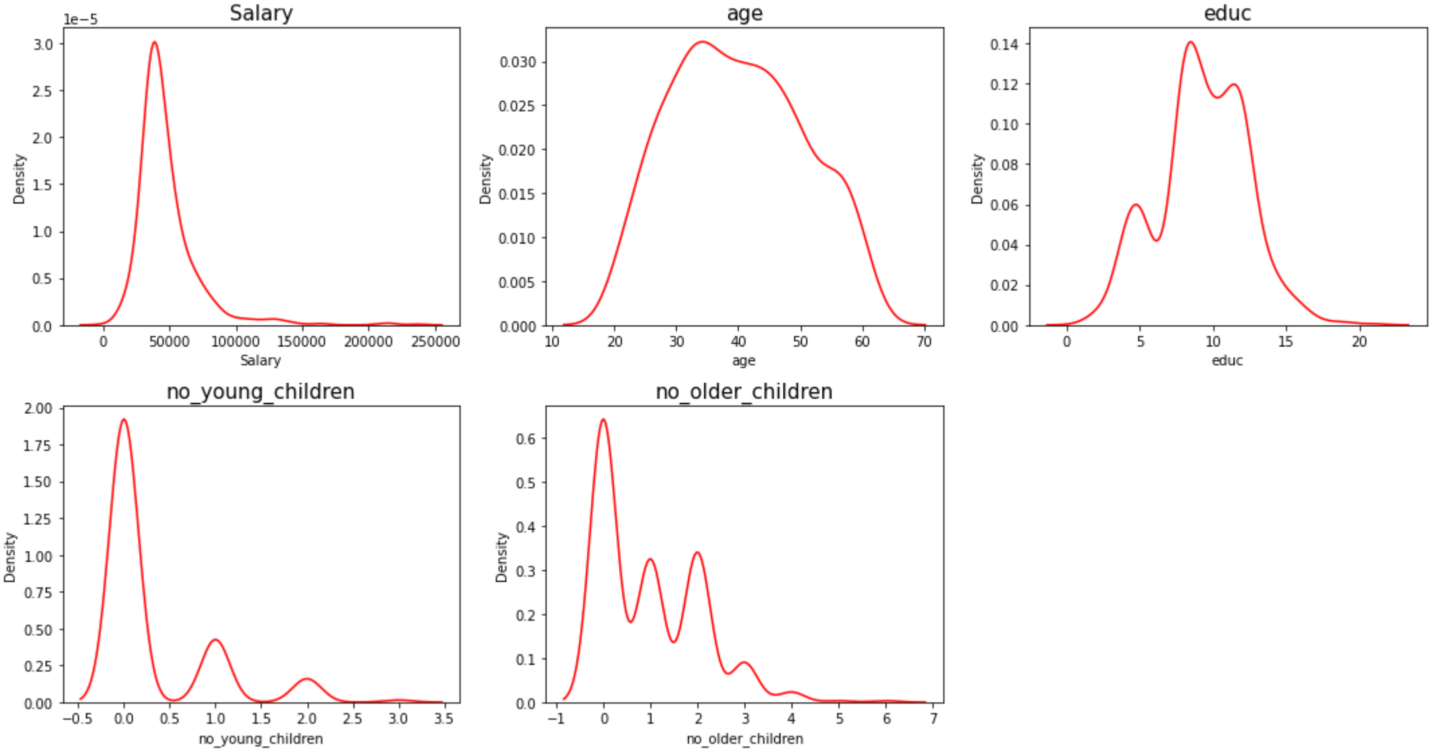
	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Holliday_Package	872	2	no	471	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Salary	872	NaN	NaN	NaN	47729.17202	23418.66853	1322	35324	41903.5	53469.5	236961
age	872	NaN	NaN	NaN	39.955275	10.551675	20	32	39	48	62
educ	872	NaN	NaN	NaN	9.307339	3.036259	1	8	9	12	21
no_young_children	872	NaN	NaN	NaN	0.311927	0.61287	0	0	0	0	3
no_older_children	872	NaN	NaN	NaN	0.982798	1.086786	0	0	1	2	6
foreign	872	2	no	656	NaN	NaN	NaN	NaN	NaN	NaN	NaN

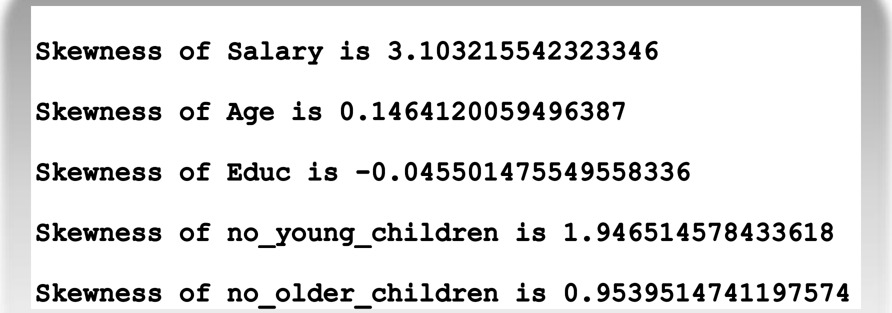
Table 2. 4 – Data Description

1. The higher percentage of employees around 54% has not bought the company's holiday package.
2. There are low as wells as high package salary employees in the company as their salary ranges from 1300 to 236961.
3. The minimum age of an employee is 20 years and maximum 62 years.
4. The are some good educated employees working in the company as education years ranges from 1 to 21 years.
5. Maximum number of young (> 7 years) children that employees have is 3.
6. There are 25 % of employees working in the company are foreigners.

UNIVARIATE ANALYSIS

Figure 2. 1 – Distribution plot





```
Skewness of Salary is 3.103215542323346

Skewness of Age is 0.1464120059496387

Skewness of Educ is -0.045501475549558336

Skewness of no_young_children is 1.946514578433618

Skewness of no_older_children is 0.9539514741197574
```

As we can see from the plot and the stats value:

1. Age and Educ have value near to 0, so they have light skewness present but we can assume and say that they are normally distributed.
2. Salary is very highly right skewed as its high value of skewness.
3. Number of children column are expected to have skewness some customers have children and some don't.

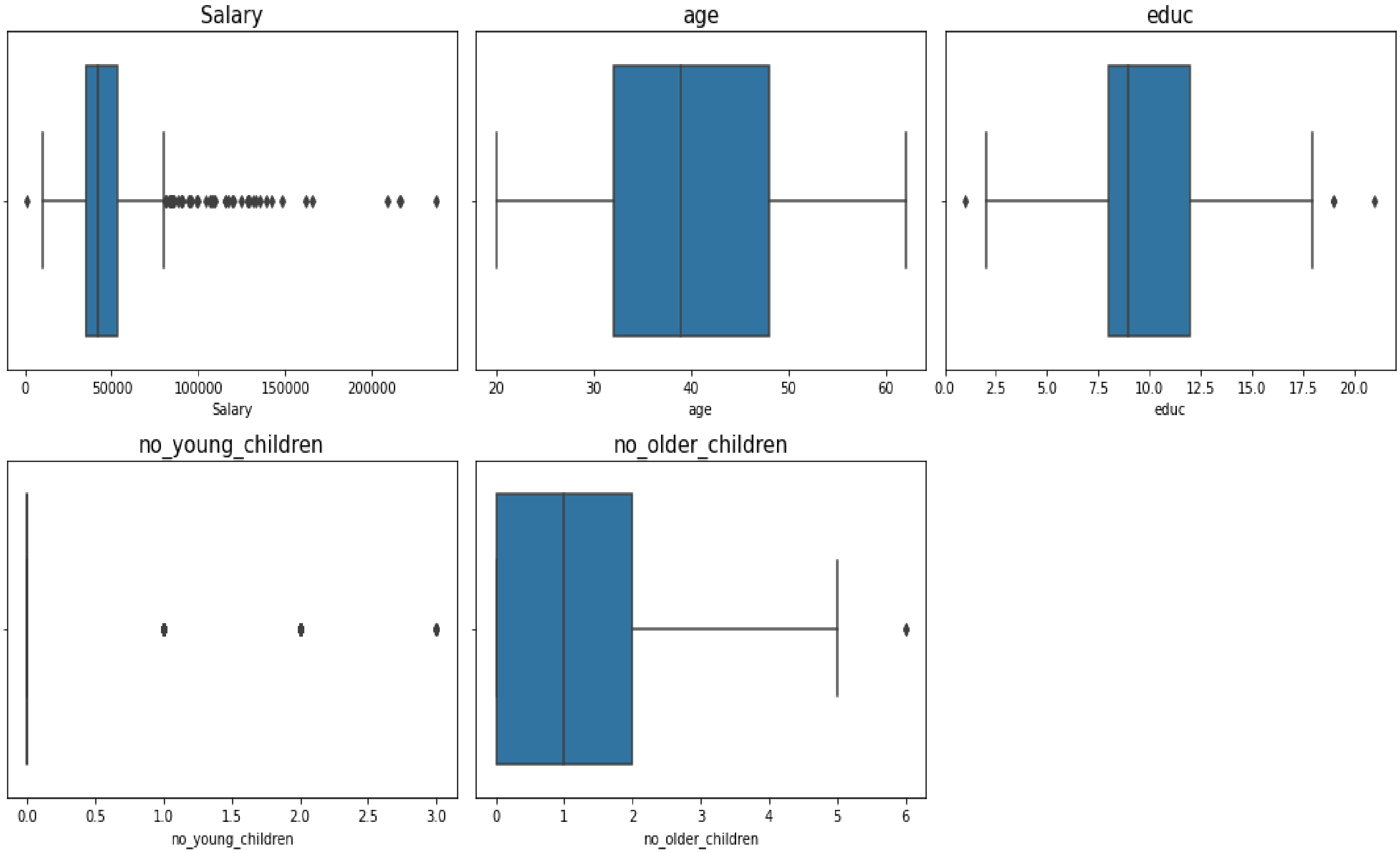


Figure 2. 2 – Boxplot of independent numeric variables.

As we can check boxplots of individual independent variable:

1. Salary has too many outliers.
2. Rest all other numerical variables has a smaller number of outliers.

As we can see from the plots

1. There are higher percentage of employees who has not bought the company's holiday packages.
2. There are more 75 % local and 25 % of foreigner employees work in the company.

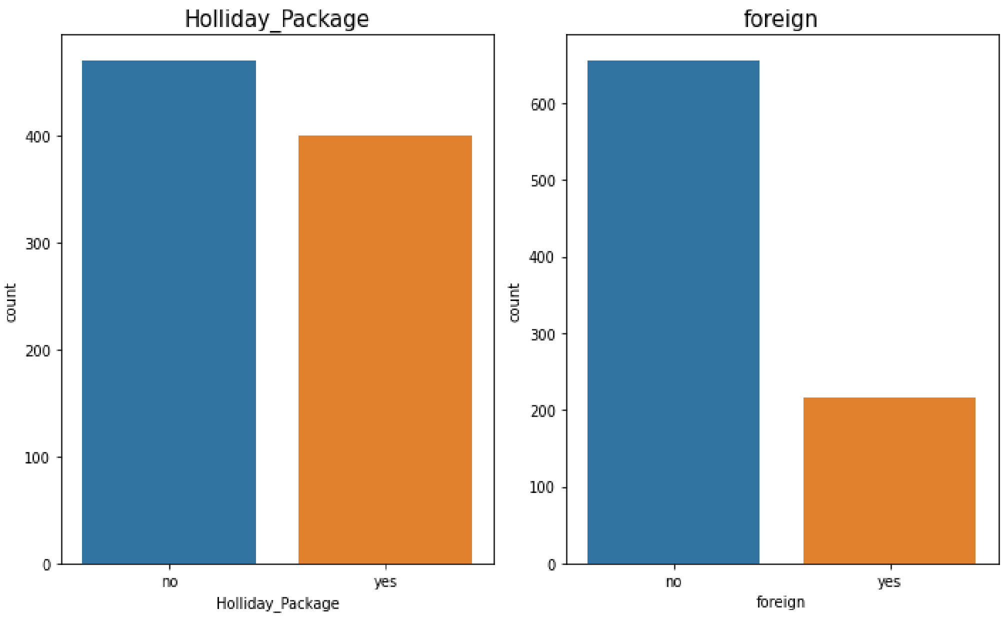


Figure 2. 3 – Count plot of Categorical Independent variables.

BIVARIATE ANALYSIS

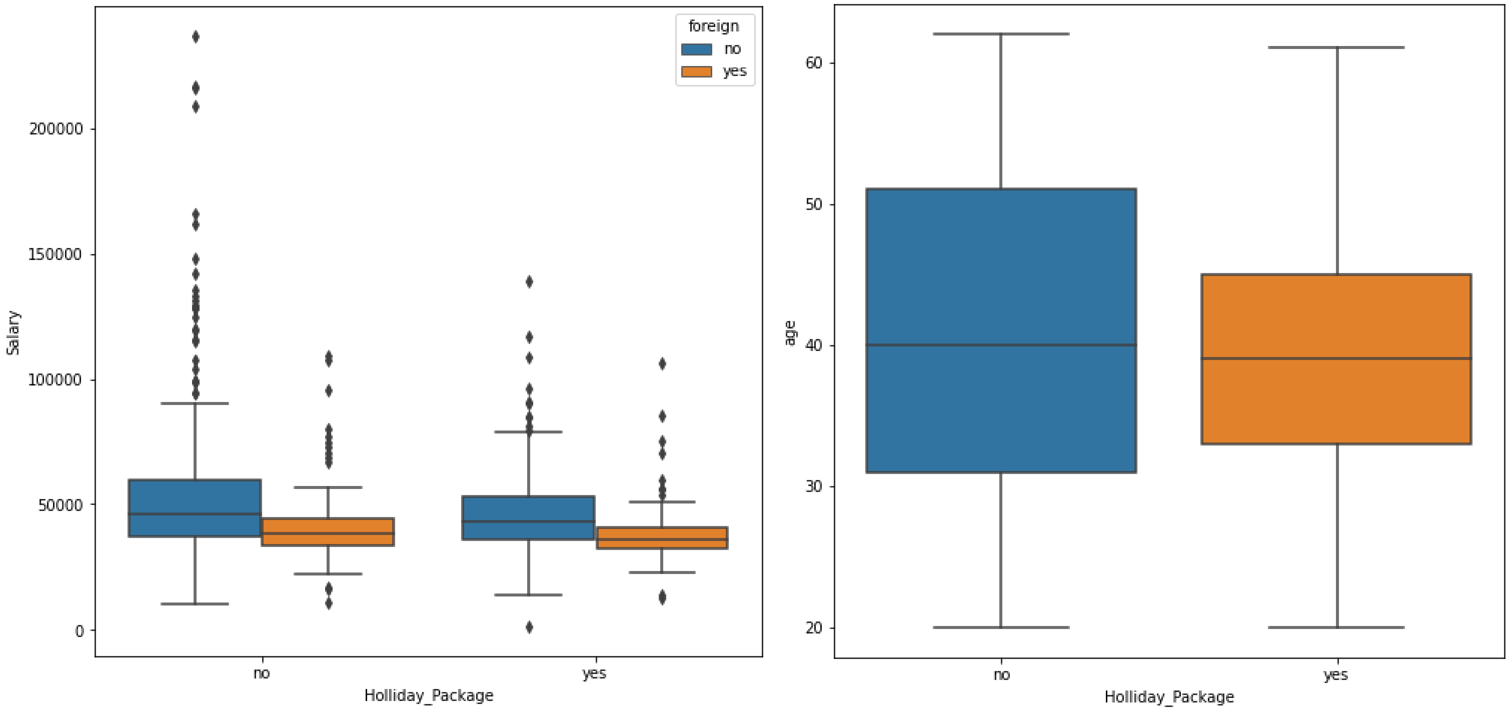


Figure 2. 4 – Box plot Holiday Package Vs Salary and Holiday package Vs age

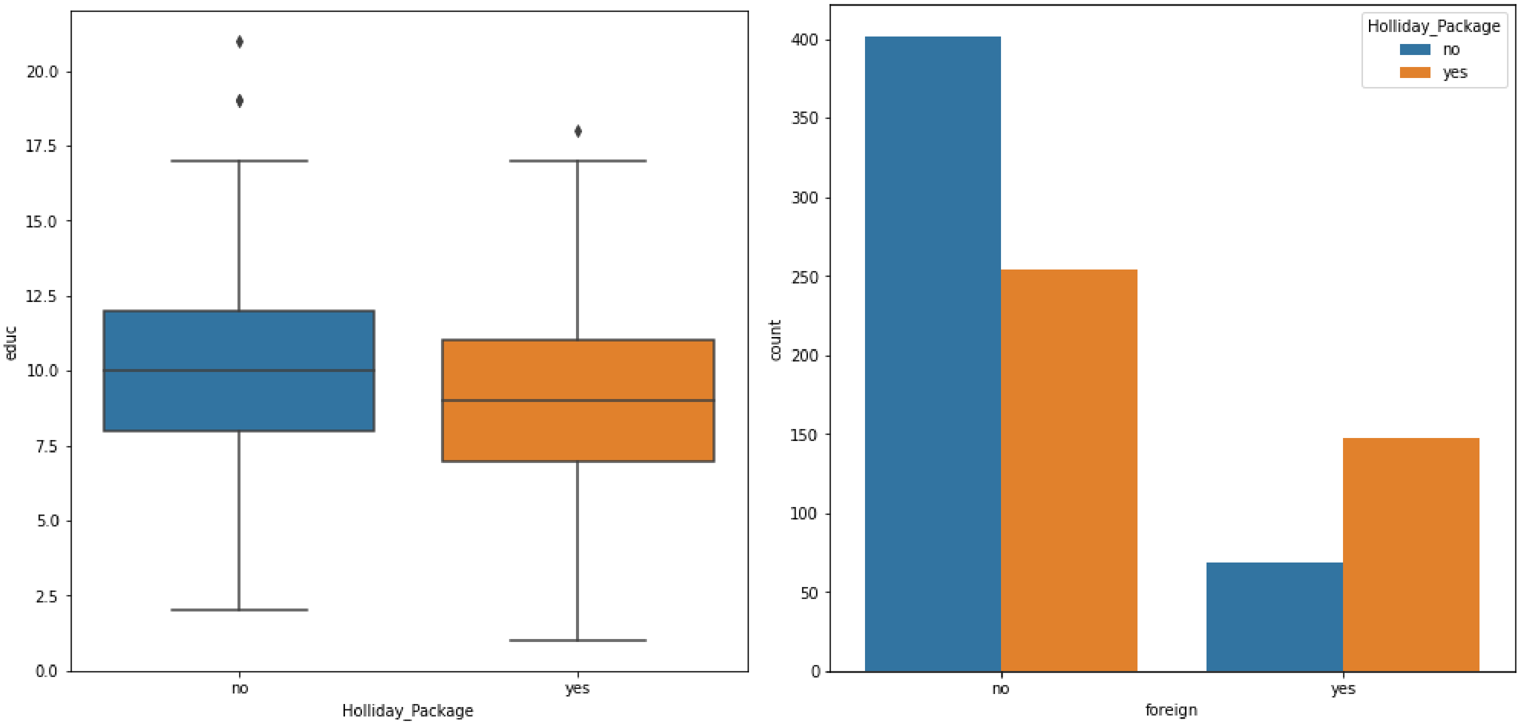


Figure 2. 5 – Plot of Holiday Package Vs Education and Holiday Vs Foreign.

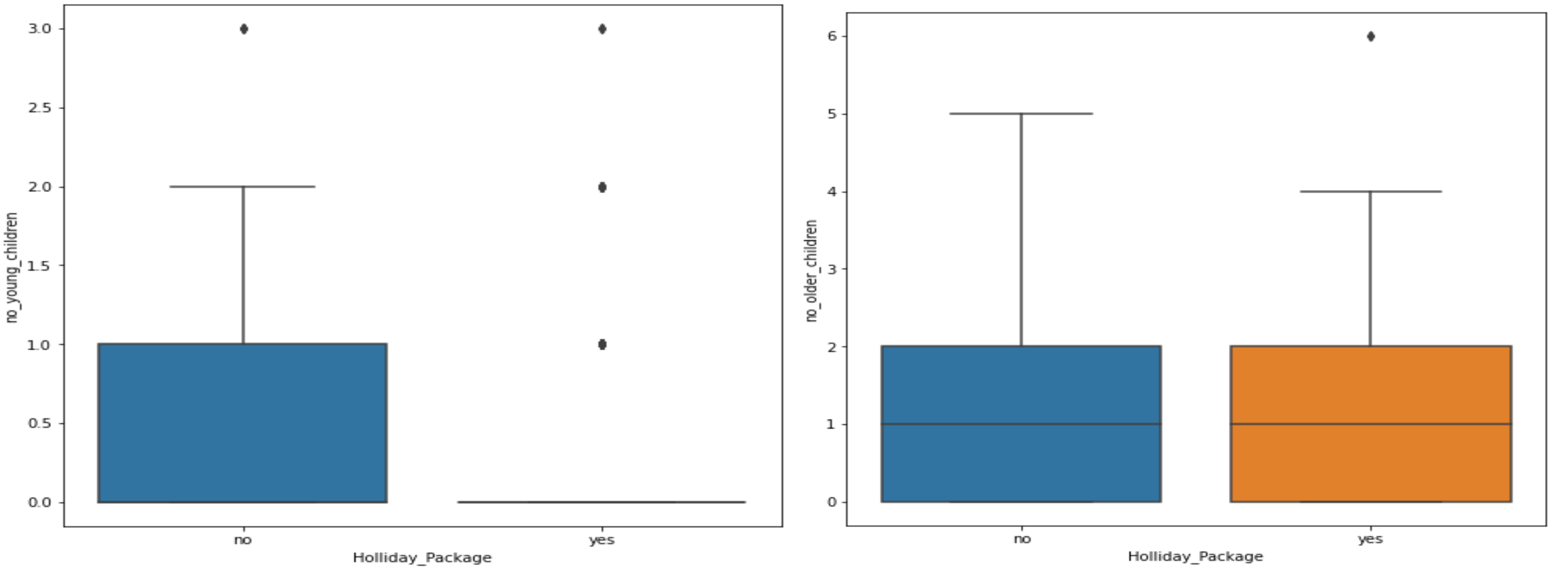


Figure 2. 6 - Plot of Holiday Package Vs Number of young children and Holiday package Vs Number of old children

As we can check from above plots,

1. The median and distribution of employee’s salary who have purchased the package(yes) and who have not purchase(no) are slightly different, so Salary can be an influencer for decision making.
2. The age median values of yes and no employees are same, so age might not be a good factor.
3. employees have less year of education are more likely to buy the holiday package.
4. Employees who have younger children (> 3 years) are less likely to buy holiday package.
5. Employees having older children does not make influence whether to buy a holiday package or not.
6. As there are only 25 % foreign employees in the company but still amongst them there is higher percentage of employees buying holiday package.

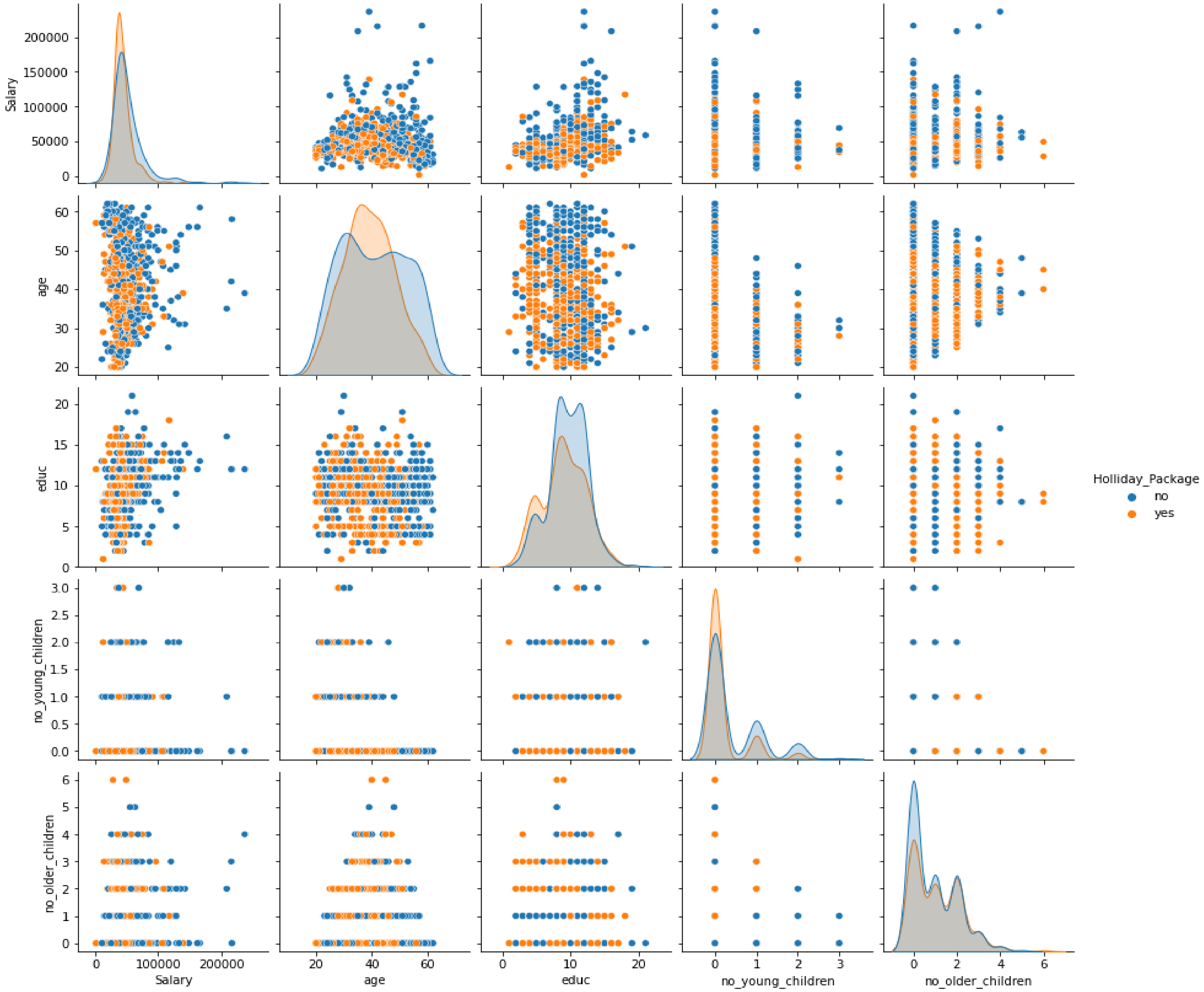
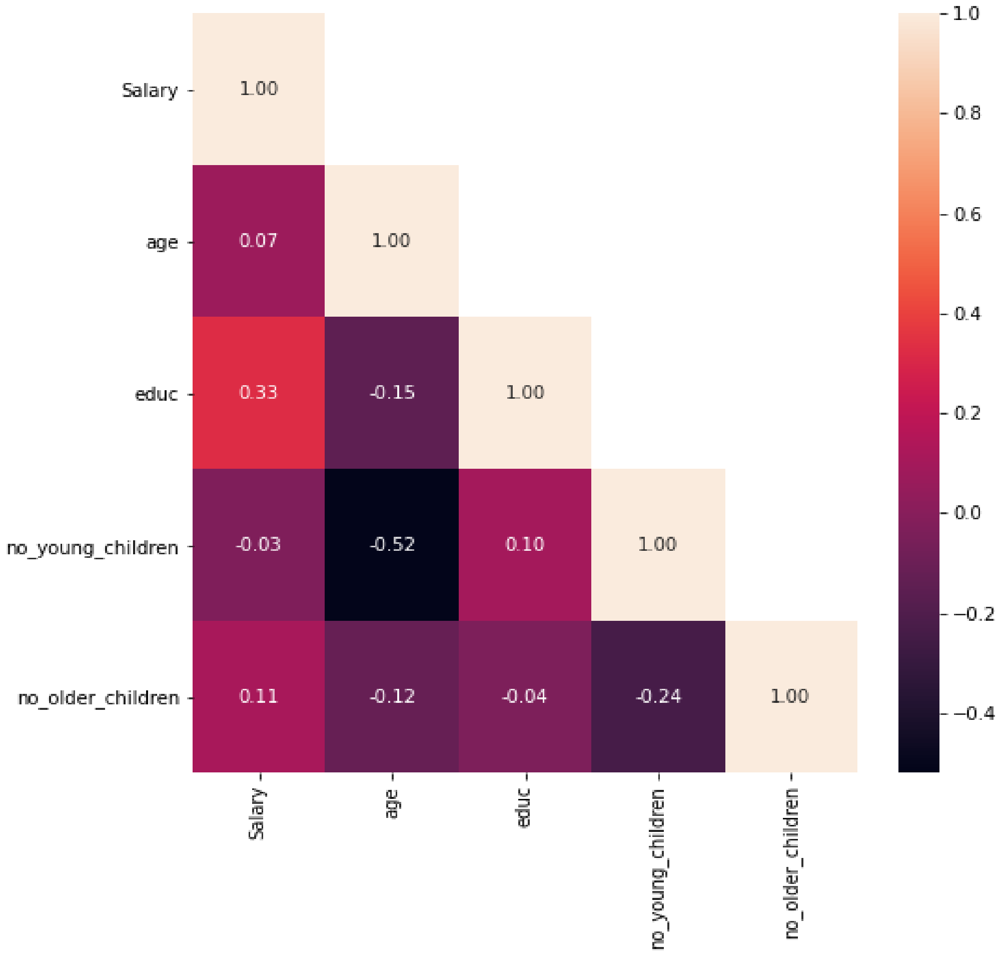


Figure 2. 7 – Pair Plot of Independent Variables

1. There is no correlation present in the data
2. The distribution of holiday package in all the independent variable are overlapping, this might affect the model performance.



From The correlation Plot,

The magnitude of correlation between independent variables,

Only Educ with salary and number of young children are moderately correlated, but this won't affect our model performance.

Figure 2. 8 – Correlation plot of Independent Variables.

Q -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).

Solution:

<u>STRING VALUES ENCODING</u>

TOP 5 ROWS AFTER ENCODING.							
Sr No.	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	0	48412	30	8	1	1	0
1	1	37207	45	8	0	1	0
2	0	58022	46	9	0	0	0
3	0	66503	31	11	2	0	0
4	0	66734	44	12	0	2	0

Table 2. 5 – Top 5 Rows after Encoding

We have assigned 0 and 1's value to categorical variables as well as target variable.

for Holliday_Package

0 - No , 1 - Yes

for foreign

0 - No , 1 - Yes

Note – For Encoding the string values (object data type) we have used **LABEL ENCODING** from sklearn.

CHECKING DATA BALANCING.

```
HOLIDAY PACKAGE

0    0.540138
1    0.459862
```

As we can check that the target variable is well balanced in two categories with 54 % in 0's and 46 % in 1's

SPLITTING THE DATA INTO TRAIN AND TEST (70:30)

X_train (610, 6)

X_test (262, 6)

y_train (610,)

y_test (262,)

For Train and Test split we have used train and test split from sklearn.

X_train – Contains independent variables (predictors) for training the model.

X_test – Contains independent variable for testing the model.

y_train – Contains dependent variable (Target) for training the model.

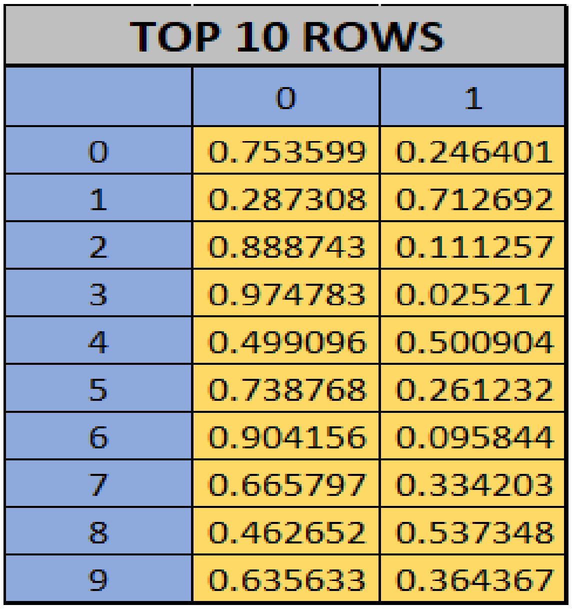
y_test – contains dependent variable for testing the model.

BUILDING LOGISTIC REGRESSION MODEL.

For building the logistic regression model we have used different combinations of hyper parameters which gave us different results but the best model performance was captured with the following parameters,

Hyper parameters:

1. Penalty – l2
2. Solver – newton-cg
3. Tol – 0.0001
4. Max iterations – 10000.



TOP 10 ROWS

	0	1
0	0.753599	0.246401
1	0.287308	0.712692
2	0.888743	0.111257
3	0.974783	0.025217
4	0.499096	0.500904
5	0.738768	0.261232
6	0.904156	0.095844
7	0.665797	0.334203
8	0.462652	0.537348
9	0.635633	0.364367

Table 2. 6 – Probability Prediction for 0's and 1's for Logistic Regression

DESCRIPTION OF HYPER - PARAMETERS

L2 PENALTY

L2 penalty is also called as Ridge Regression. Ridge regression adds ''squared magnitude" of coefficient as penalty term to the loss function.

NEWTON-CG

Newton-cg solver calculates Hessian Explicitly.

TOLERANCE

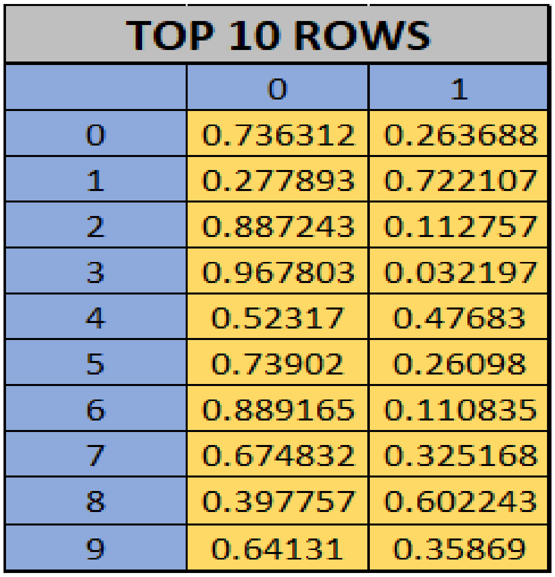
The tolerance set for the loss function is 0.0001.

BUILDING LINEAR DISCRIMINANT ANALYSIS MODEL. (LDA)

For building the linear discriminant analysis (LDA) model we have used different combinations of hyper parameters which gave us different results but the best model performance was achieved with the following parameters,

Hyper Parameters:

1. Solver –Singular value decomposition (SVD)



TOP 10 ROWS		
	0	1
0	0.736312	0.263688
1	0.277893	0.722107
2	0.887243	0.112757
3	0.967803	0.032197
4	0.52317	0.47683
5	0.73902	0.26098
6	0.889165	0.110835
7	0.674832	0.325168
8	0.397757	0.602243
9	0.64131	0.35869

Table 2. 7 - Probability Prediction for 0’s and 1’s for LDA

DESCRIPTION OF HYPER – PARAMETERS

SOLVER:

SVD

The technique is based on fast approximate singular value decomposition (SVD) which has deep connections with low rank approximation of the data matrix.

LSQR

The ‘lsqr’ solver is an efficient algorithm that works only for classification and it supports shrinkage.

EIGEN

The “eigen” solver is based on the optimization of the between class scatter t within class scatter ratio. It can be used for the both classification and transform, and it supports shrinkage.

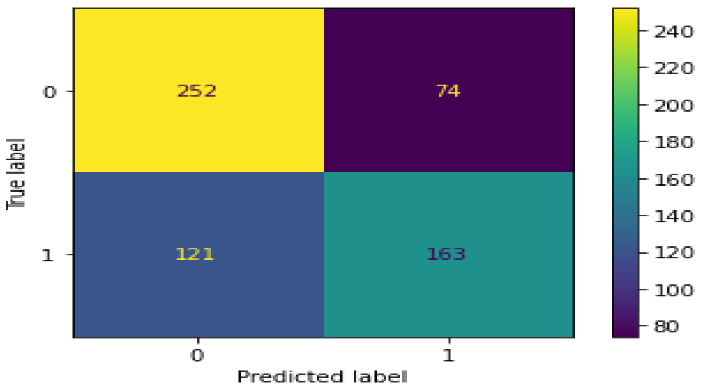
Q-2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.

Solution:

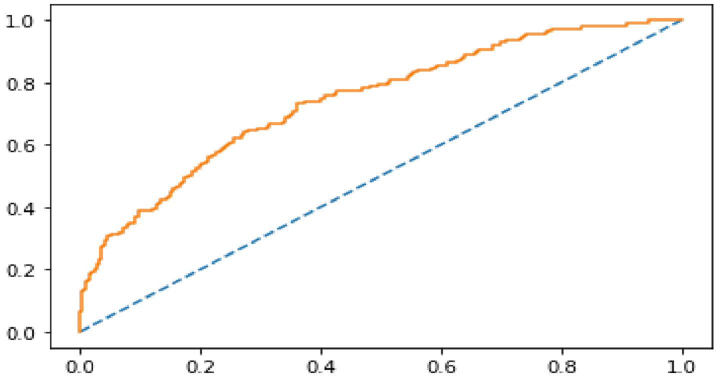
Performance Evaluation of Logistic regression model.

FOR TRAIN DATA

ACCURACY - 0.680327868852459



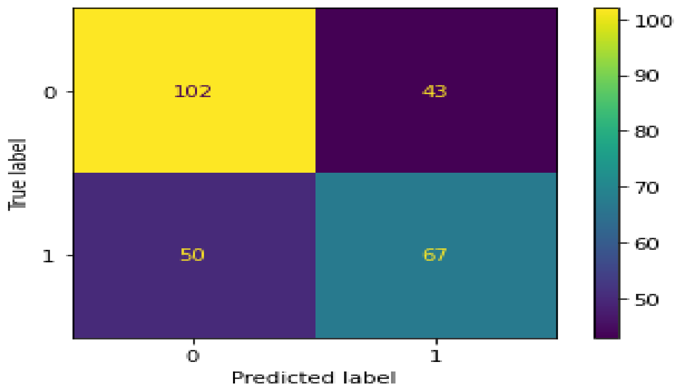
```
        Classification Report for Train Data

    precision     recall    f1-score    support

0    0.68         0.77        0.72        326
1    0.69         0.57        0.63        284

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



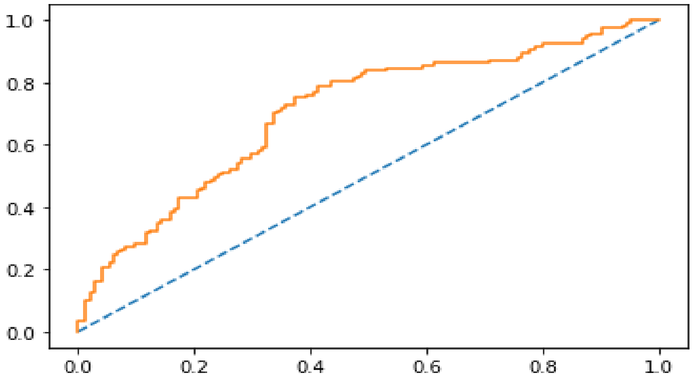
```
AUC Value for Train: 0.743
```

FOR TEST DATA

ACCURACY - 0.6450381679389313



```
        Classification Report for Test Data

    precision     recall    f1-score    support

0     0.67         0.70       0.69        145
1     0.61         0.57       0.59        117

accuracy                      0.65         262
macro avg     0.64     0.64   0.64         262
weighted avg  0.64     0.65   0.64         262
```



```
AUC Value for test: 0.705
```

CONFUSION MATRIX:

For Train Data
True Negative – 252
False Negative – 121
True Positive – 163
False Positive - 74

For Test Data
True Negative – 102
False Negative – 50
True Positive – 67
False Positive - 43

As we can check from the confusion matrix the model performance is good but not very good because for training and testing both prediction false negative rate is high.

CLASSIFICATION REPORT:

As we can see the Precision values for 0's and 1's are almost similar for train data as 0.68 & 0.69 , but for test data there is a bit drop of 0's and 1's as 0.67 & 0.61. train and test data has almost similar values with a bit of drop in 1's.

Similar pattern can be seen for recall values also for train – 0.77 & 0.57 for 0's and 1's. For test 0's, 1's – 0.70 & 0.57.

F1 Score also have similar values in both train and test with values of 0's and 1's , 0.72,0.63 and 0.69,0.59 with a drop of 3 to 4% only we can say model performance for test and train is moderate.

ACCURACY:

Accuracy of the model for both train and test are good for the model performance which is 68% and 65%, As the data is balanced, we can rely on the accuracy for model performance.

ROC AND AUC VALUES:

As the area under the curve of the model for train and test data is also good and there is not much difference as 74.3% and 70.4%

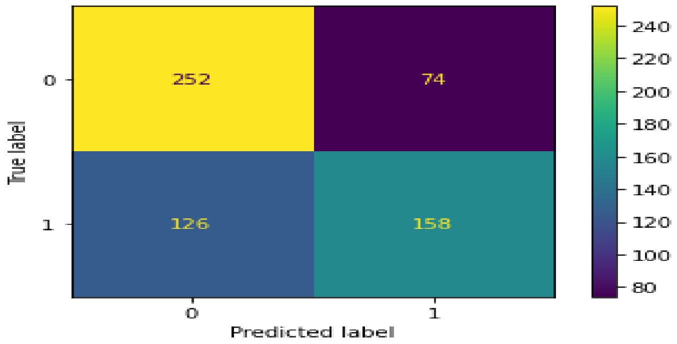
CONCLUSION:

From the above performance evaluation parameters, we can conclude that model is a robust model. There is no overfitting or under fitting present for the data. As it's a balanced data we can measure the model performance from accuracy and from accuracy model is performing good.

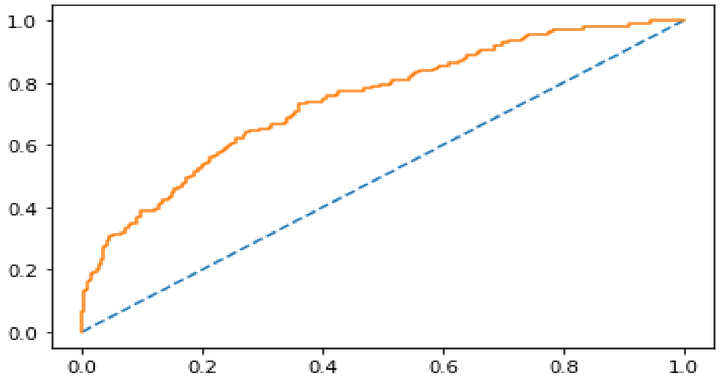
Performance Evaluation for Linear Discriminant Analysis (LDA) model.

FOR TRAIN DATA

ACCURACY - 0.6721311475409836



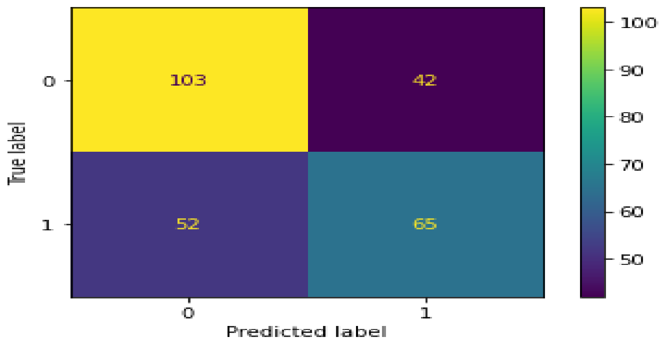
```
     Classification Report for Train Data

   precision     recall    f1-score    support

0     0.67        0.77       0.72        326
1     0.68        0.56       0.61        284

accuracy                        0.67    610
macro avg     0.67    0.66      0.66    610
weighted avg  0.67    0.67      0.67    610
```



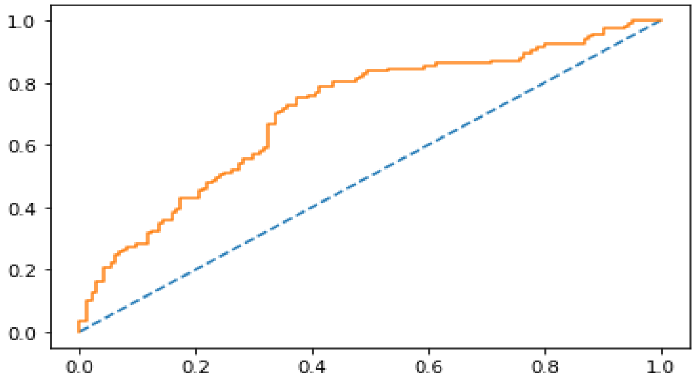
AUC Value for Train : 0.742

FOR TEST DATA

ACCURACY - 0.6412213740458015



```
     Classification Report for Test Data

   precision     recall    f1-score    support

0     0.66         0.71       0.69        145
1     0.61         0.56       0.58        117

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



AUC Value for test : 0.703

CONFUSION MATRIX:

For Train Data
True Negative – 252
False Negative – 126
True Positive – 158
False Positive - 74

For Test Data
True Negative – 103
False Negative – 52
True Positive – 65
False Positive - 42

As we can check from the confusion matrix the model performance is good but not very good because for training and testing both prediction false negative rate is high.

CLASSIFICATION REPORT:

As we can see the Precision values for 0's and 1's are almost similar for train data as 0.67 & 0.68, but for test data there is a bit drop of 0's and 1's as 0.66 & 0.61. train and test data has almost similar values with a bit of drop in 1's.

Similar pattern can be seen for recall values also for train – 0.77 & 0.56 for 0's and 1's. For test 0's, 1's – 0.71 & 0.56.

F1 Score also have similar values in both train and test with values of 0's and 1's, 0.72,0.61 and 0.69,0.58 with a drop of 3 to 4% only we can say model performance for test and train is moderate.

ACCURACY:

Accuracy of the model for both train and test are good for the model performance which is 67% and 64%, As the data is balanced, we can rely on the accuracy for model performance.

ROC AND AUC VALUES:

As the area under the curve of the model for train and test data is also good and there is not much difference as 74.2% and 70.3%

CONCLUSION:

From the above performance evaluation parameters, we can conclude that model is a robust model. There is no overfitting or under fitting present for the data. As it's a balanced data we can measure the model performance from accuracy and from accuracy model is performing good.

MODEL COMPARISON

MODELS PERFORMANCE COMPARISON											
		PRECISION		RECALL		F1 SCORE		ACCURACY		ROC SCORE	
		TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
Logistic Regression	0	0.68	0.67	0.77	0.7	0.72	0.69	0.68	0.65	0.743	0.705
	1	0.69	0.61	0.57	0.57	0.63	0.59				
Linear Discriminative Analysis	0	0.67	0.66	0.77	0.71	0.72	0.69	0.67	0.64	0.743	0.703
	1	0.68	0.61	0.56	0.56	0.61	0.58				

Table 2. 8 – Model Comparison

As we can see from the above table both the model performance values are almost same, but still logistic regression values are higher as compare to linear discriminative analysis. As this is the balanced data accuracy is the best measure for both the models.

As logistic regression accuracy is higher than LDA and so as other evaluation parameters, we chose to work with logistic regression model for this business problem.

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

Solution:

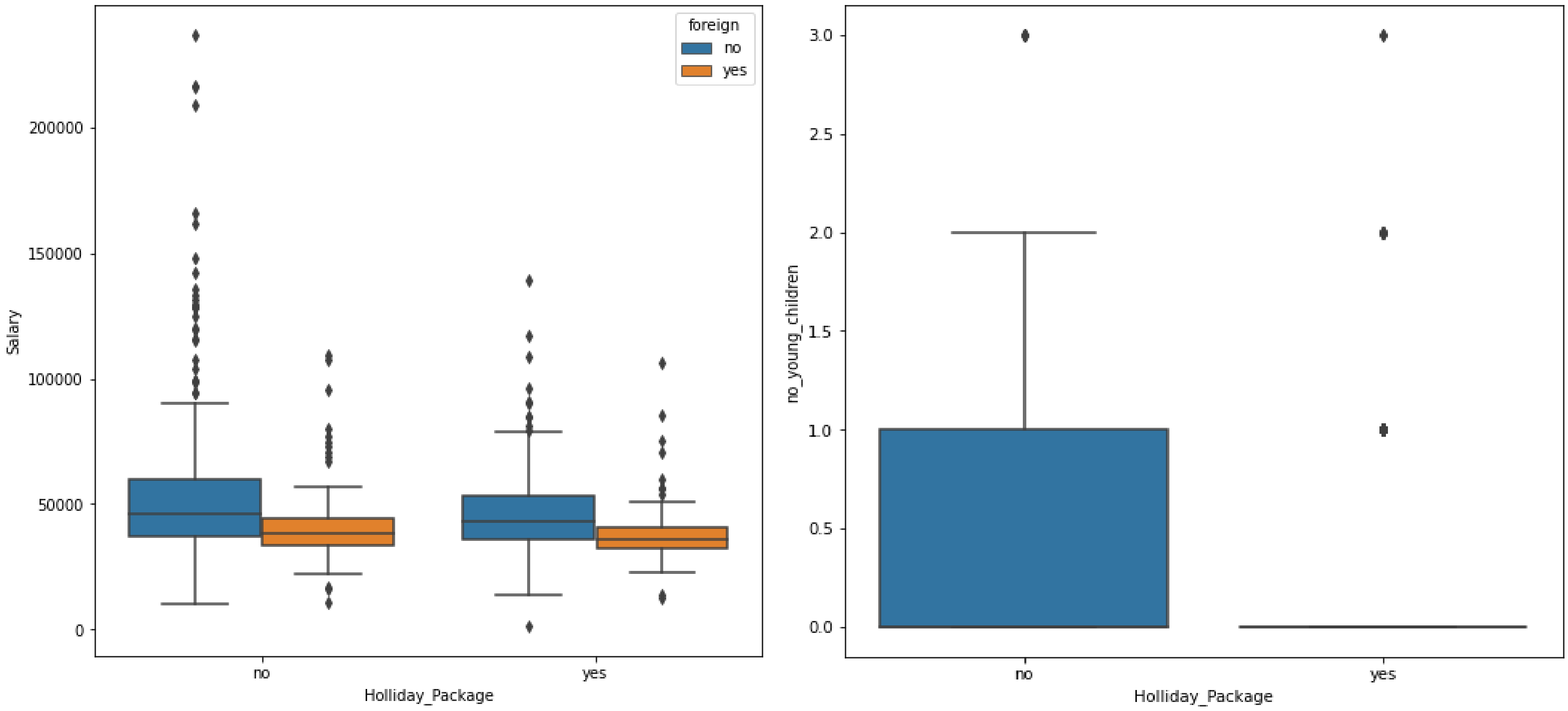


Figure 2. 9 - Holiday Package Vs Salary and Holiday Package Vs Number of young Children

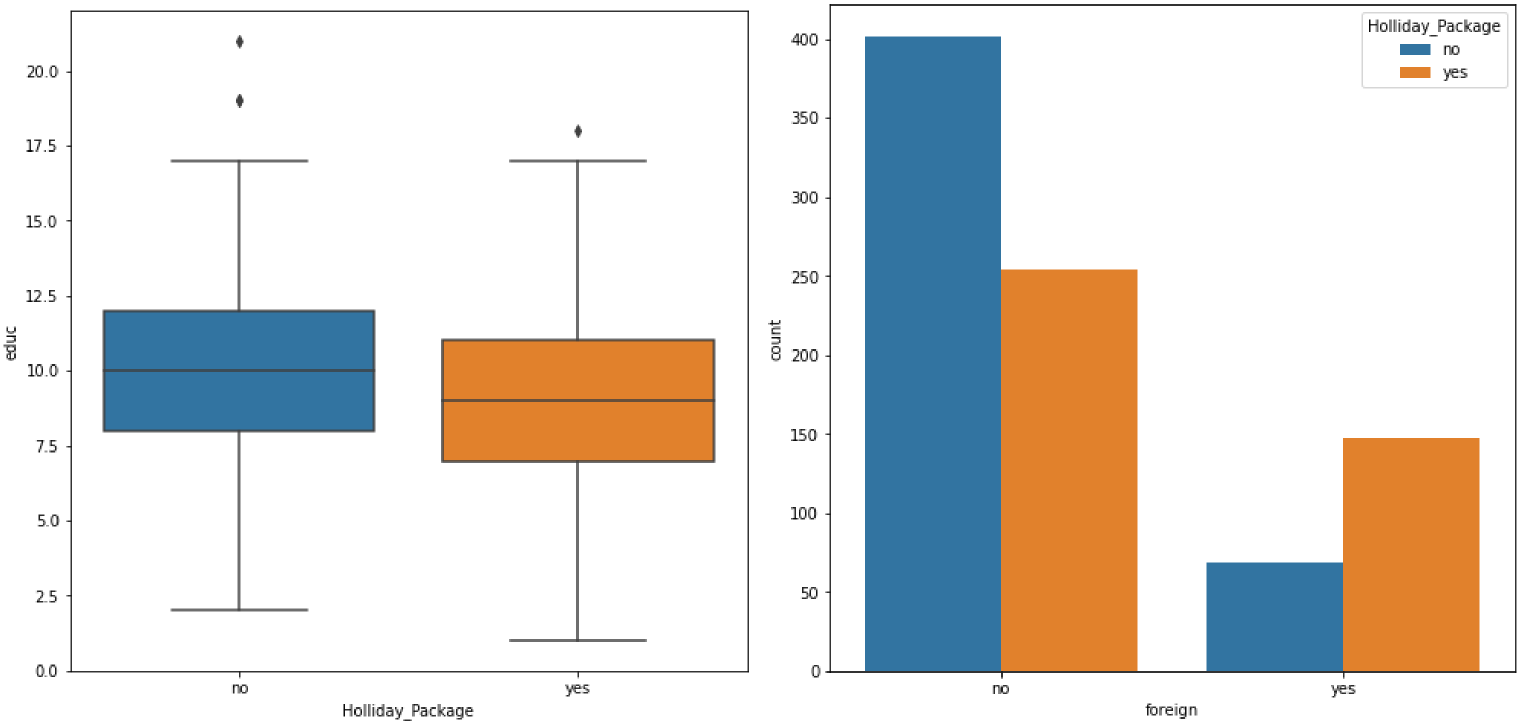


Figure 2. 10 - Holiday Package Vs Education and Holiday Package Vs Foriegn

BUSINESS INSIGHTS:

Important Factors that are influencing the decision whether to buy holiday package are following

1. ***SALARY***

As salary is an important factor for predict whether an employee will buy the holiday package or not.

BUSINESS SOLUTION

In this business we have observed that employees having lesser lower salary are more likely to holiday package. So, the tour and travel company should focus on selling their products with employees having lower salary.

2. ***EDUCATION***

Education is also an important factor in this business problem to predict the decision.

BUSINESS SOLUTION.

In this business problem we have observed that employees with a smaller number of years in education are more interested and likely the holiday package. So, Tour and travel company should focus on employees with less number years in education.

3. ***NUMBER OF CHILDREN.***

   Age of children are also very important for decision making.

   BUSINESS SOLUTION.

   Employees having younger kids (> 7 years) are less likely to buy any holiday package but as the kids of these employees grow older than 7 years than there is an equal chance, whether employee will buy or not. So, Employees with older kids' company can focus on them to sell their products.

4. ***FOREIGN.***

   If an employee is a foreigner or not has a huge impact on decision making.

   BUSINESS SOLUTION.

   In the given data there are a smaller number of foreigner employees working in the company but still there are higher chances of that employees to buy a holiday package. So, company should focus on foreigner employee to sell their products.