PREDICTIVE MODELLING PROJECT-18Th SEPT, 2022

Submitted by,

Deepa .K

Table of Contents

List of Figures:

Problem Statement 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.

Variable Name	**Description**
Carat	Carat weight of the cubic zirconia.
Cut	Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal.
Color	Colour of the cubic zirconia.With D being the 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.

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.

Solution:

The dataset has 10 columns and 26967 entries with some duplicated data and all the entries are filled with float,int and object data types.

Shape: (26967, 10)

Data type: Float,int,object

Length: 9 Columns(dropped unwanted column-Unnamed:0)

Duplication: 34 rows are with duplications.

Missing values: No missing values in the dataset.

Summary dataset: Count, mean, std.deviation, range of values and IQR range.

	carat	depth	table	x	y	z	price
count	26967.000000	26270.000000	26967.000000	26967.000000	26967.000000	26967.000000	26967.000000
mean	0.798375	61.745147	57.456080	5.729854	5.733569	3.538057	3939.518115
std	0.477745	1.412860	2.232068	1.128516	1.166058	0.720624	4024.864666
min	0.200000	50.800000	49.000000	0.000000	0.000000	0.000000	326.000000
25%	0.400000	61.000000	56.000000	4.710000	4.710000	2.900000	945.000000
50%	0.700000	61.800000	57.000000	5.690000	5.710000	3.520000	2375.000000
75%	1.050000	62.500000	59.000000	6.550000	6.540000	4.040000	5360.000000
max	4.500000	73.600000	79.000000	10.230000	58.900000	31.800000	18818.000000

Figure 4: Summary Dataset

Observation:

- The dataset looks good with no missing values in it. The minimum to maximum range value of all the attributes are visible here.

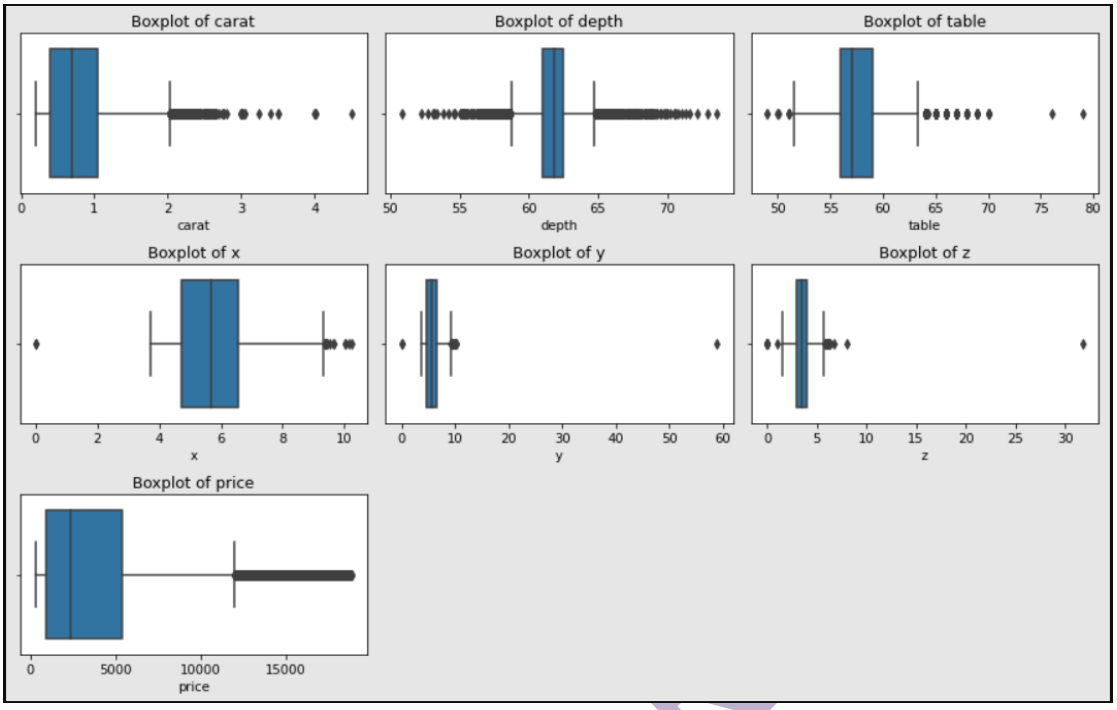

Figure 5:Box plot

Observation:

- For the univariate analysis,boxplot with outliers are represented.Here all the variables show outlier data. This can be treated in the next plot and replaced with median values to retain them within the dataset.

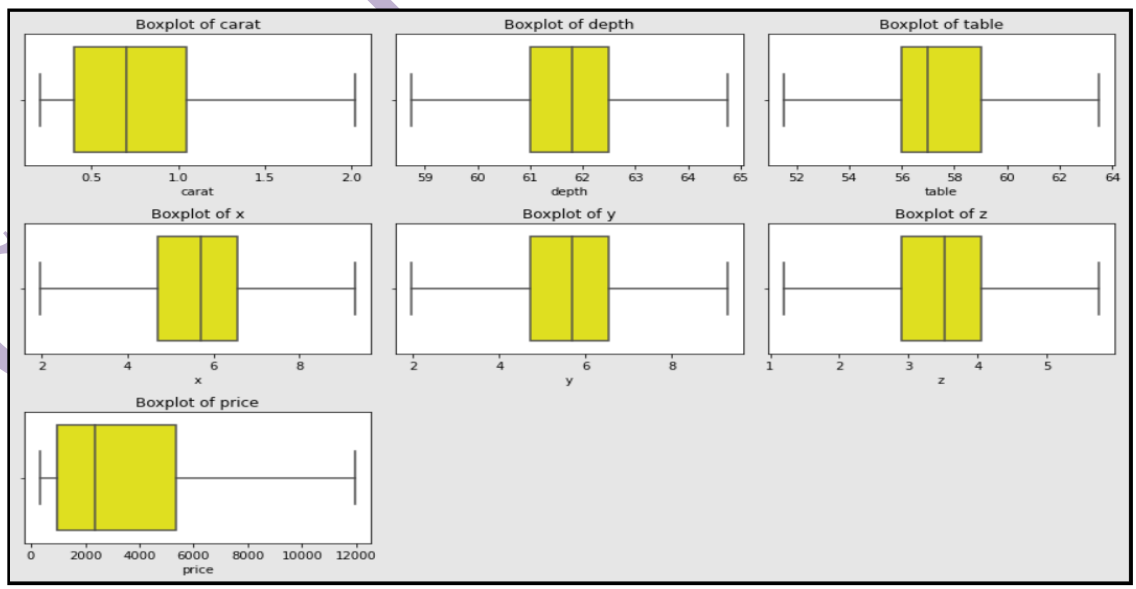

Figure 6: No outlier

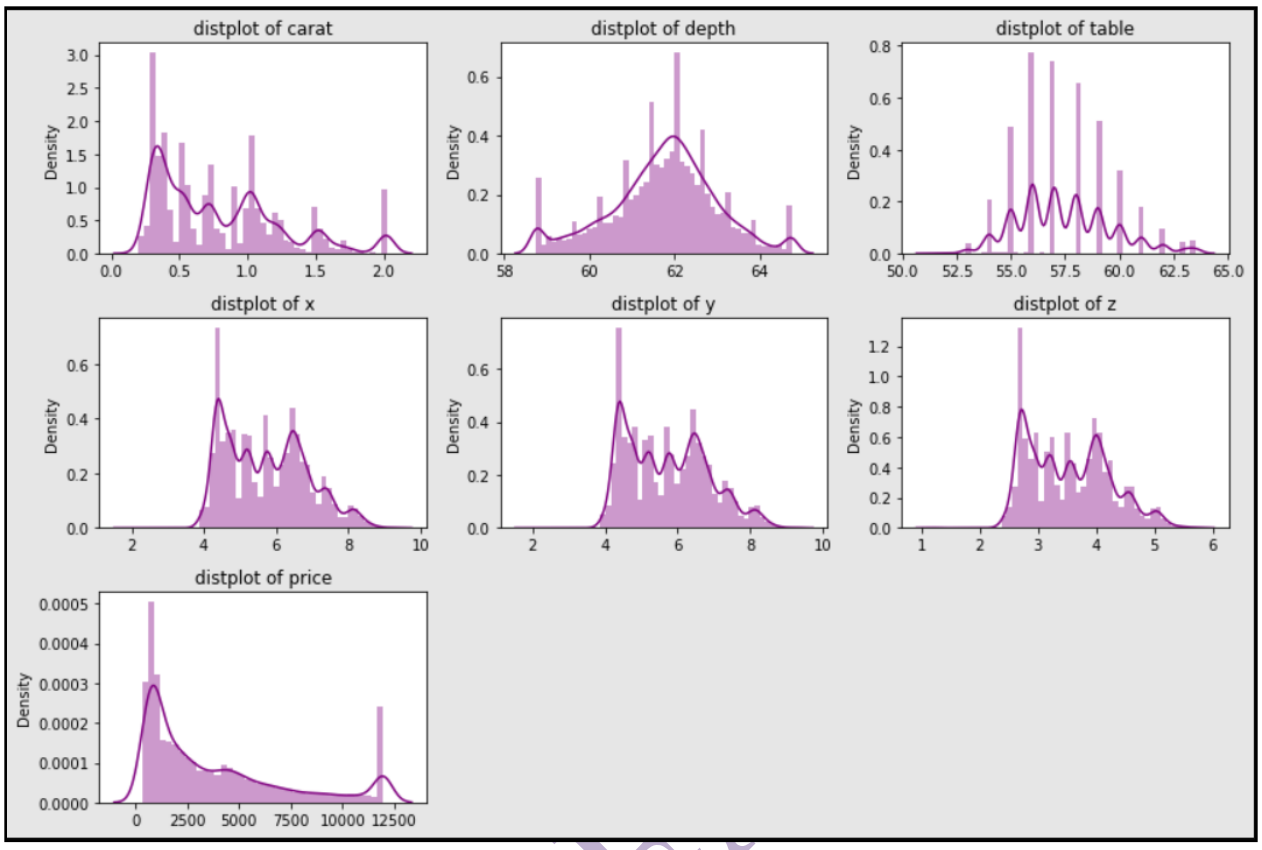

Figure 7:Distplot

Observation:

- The ‘carat’ and ‘Price’ values are right skewed in the above distplot.

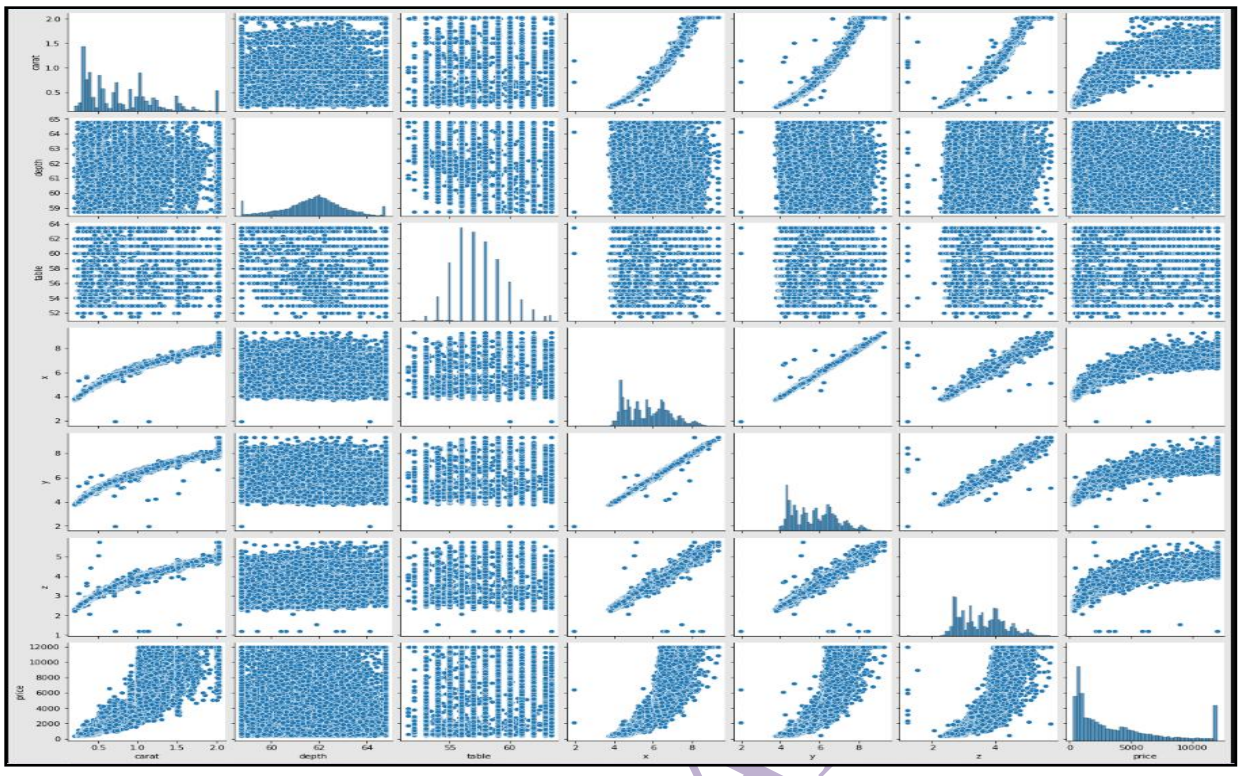

Figure 8:Pair plot

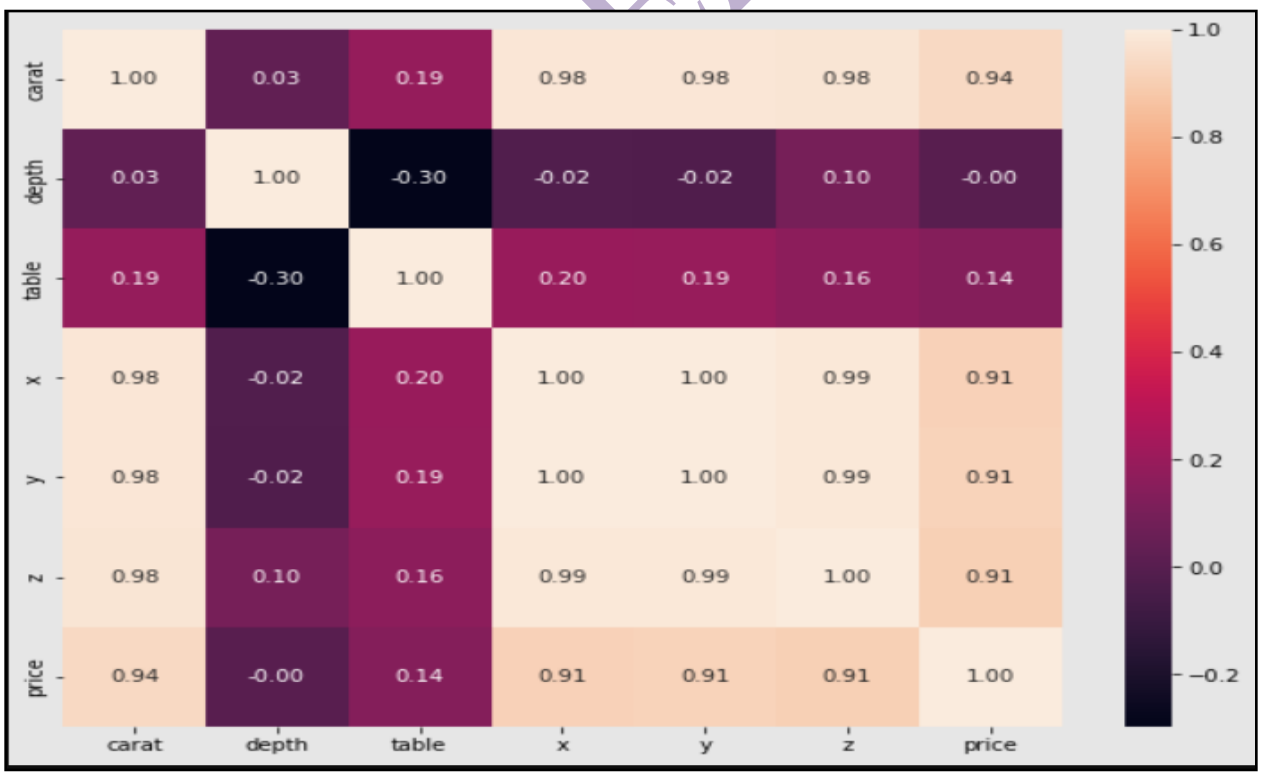

Figure 9:Heat map

Observation:

- There is strong correlation between most variables and price of diamond and exception is seen in depth that has very negligible correlation.

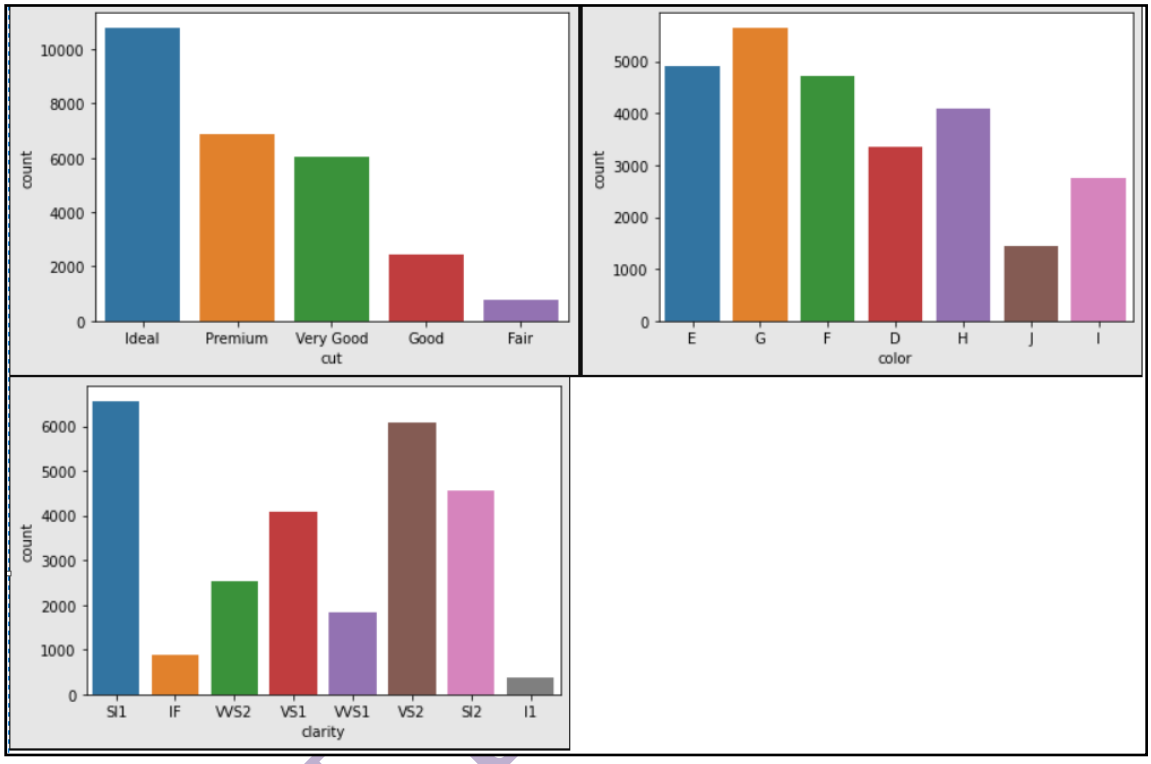

Figure 10: Countplot

Observation:

- Cut: The Premium Cut on Diamonds are the most expensive followed by Very Good Cut. Ideal cut diamonds are available in larger quantity than other cut ones.
- Color: J color code diamonds are the best colored and D refers to the worst color.
- Clarity: The Diamonds clarity with VS1 & VS2 are the best and most expensive.

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:

The values with zero are listed and we can observe there are few 'Zero' values present in the data set on variables X,Y and Z.This indicates that they are faulty values.

As there cannot be dimensionless or 2-dimensional diamonds we have listed these as faulty/wrong values.After imputing with median the shape of dataset gets reduced from (26967,10) to (26933,10).

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 R square, RMSE & Adj R square. Compare these models and select the best one with appropriate reasoning.

Solution:

Linear regression is a way to identify a relationship between two or more variables. We use this relationship to predict the values for one variable for a given set of values to the other variables. The variable which is used in prediction is termed as independent/explanatory/regressor variable where the predicted variable is termed as dependent/target/response/regressand variable.Linear regression assumes that the dependent variable is linearly related to the estimated parameters.

Here we have splitted the dataset into 70:30 ratio to get the best fit model for this problem. The coefficients for each independent variables are listed below.

- The coefficient for carat is 11021.57
- The coefficient for cut is 108.11

- The coefficient for color is 333.86
- The coefficient for clarity is 505.36
- The coefficient for depth is -79.53
- The coefficient for table is -30.35
- The coefficient for x is -939.48
- The coefficient for y is 10.63
- The coefficient for z is -41.27

From the coefficients it is clear that the increase in one unit of each attribute causes change in price of diamond.The one unit increase in carat increases price by 11021.57.The one unit decrease in depth decreases price by -79.53.

R-Square value:

R-Square value(coefficient of determinant) is a statistical measure that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model.

- X-Train,Y-Train: 0.908
- X-Test,Y-Test: 0.907

RMSE (Root mean squared value):

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are and RMSE is measure of how spread out these residuals are.Also how concentrated the data is around the line of best fit.

- X-Train,Y-Train: 1212.41
- X-Test,Y-Test: 1230.27

R-Square values & RMSE of train and test data are almost equal which represents that the model is right fit one.90% of variation in price is explained by predictors in the model for train set.

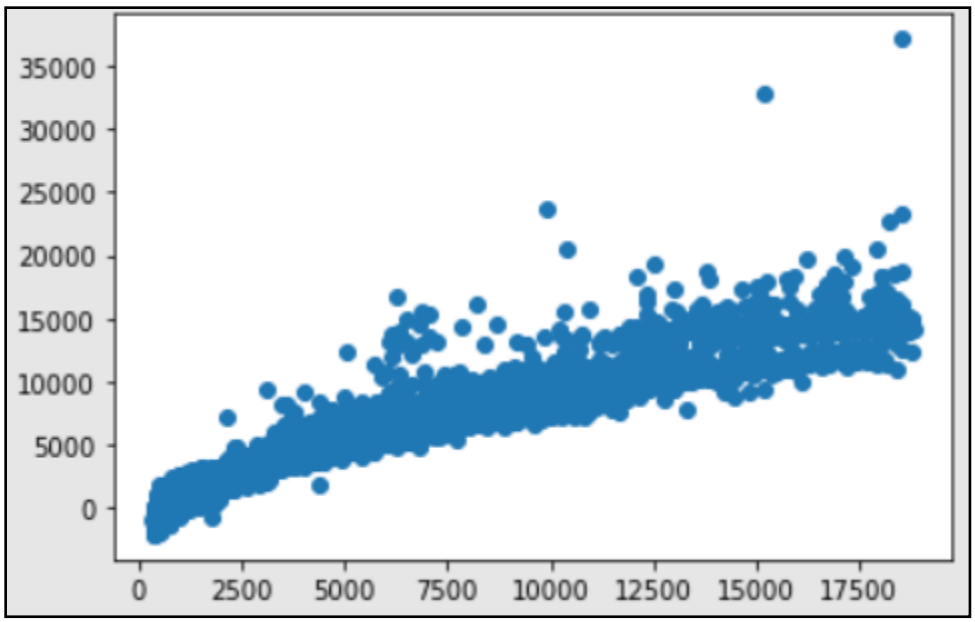

Figure 11:Y(Scatter)

- Strong correlation between the predicted y and actual y is seen in the linear plot.

Linear regression using Statsmodel:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.909
Model:	OLS	Adj. R-squared:	0.909
Method:	Least Squares	F-statistic:	2.080e+04
Date:	Fri, 16 Sep 2022	Prob (F-statistic):	0.00
Time:	19:05:20	Log-Likelihood:	-1.6061e+05
No. Observations:	18853	AIC:	3.212e+05
Df Residuals:	18843	BIC:	3.213e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4276.2134	714.365	5.986	0.000	2875.993	5676.433
carat	1.102e+04	91.410	120.572	0.000	1.08e+04	1.12e+04
cut	108.1127	9.746	11.093	0.000	89.009	127.216
color	333.8688	5.485	60.870	0.000	323.118	344.620
clarity	505.3640	5.919	85.378	0.000	493.762	516.966
depth	-79.5320	7.857	-10.122	0.000	-94.933	-64.131
table	-30.3556	4.988	-6.086	0.000	-40.133	-20.578
x	-939.4875	49.668	-18.915	0.000	-1036.841	-842.134
y	10.6322	23.769	0.447	0.655	-35.957	57.221
z	-41.2730	39.592	-1.042	0.297	-118.876	36.330

Omnibus:	3976.153	Durbin-Watson:	1.980
Prob(Omnibus):	0.000	Jarque-Bera (JB):	158732.003
Skew:	0.103	Prob(JB):	0.00
Kurtosis:	17.214	Cond. No.	6.89e+03

Figure 12:LM Summary

- As we see here the overall P value is less than alpha(0), so rejecting H0 and accepting Ha that atleast 1 regression co-efficient is not '0'. The attribute which are having p value greater than 0.05 are poor predictor for price.

MSE (mean squared value):

The mean squared error (MSE) indicates how close a regression line is to a set of points.

MSE of the model:1212.74

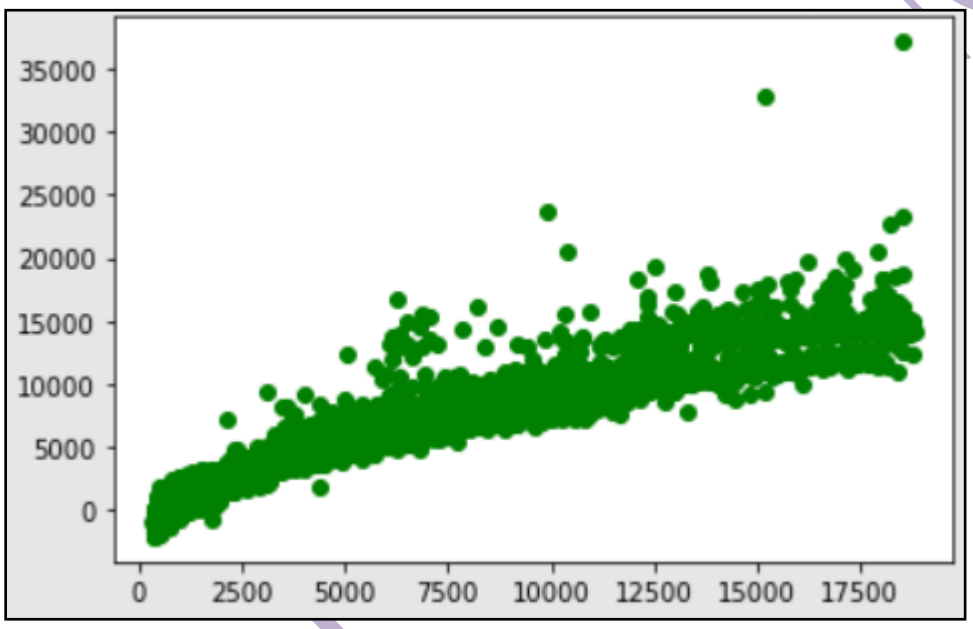

Figure 13:Y test

Linear equation:

- Price=(4276.21) * Intercept + (11021.58) * carat + (108.11) * cut + (333.87) * color + (505.36) * clarity + (-79.53) * depth + (-30.36) * table + (-939.49) * x + (10.63) * y + (-41.27) * z.

Observation:

- The one unit increase in carat increases price by 11021.57.
- The one unit increase in cut increases price by 108.11
- The one unit increase in color increases price by 333.86.
- The one unit increase in clarity increases price by 505.36
- The one unit decrease in depth decreases price by -79.53.
- The one unit decrease in table decreases price by -30.35

- The one unit decrease in x decreases price by -939.48
- The one unit decrease in y increases price by 10.63
- The one unit decrease in z decreases price by -41.27

Here the increase and decrease of price is caused by maintaining all other predictors as constant. The 'y' variable (width) in mm having positive co-efficient indicates that higher the width of the stone is a higher profitable stones.
Finally we can conclude that best 5 attributes that are most important are 'Carat', 'Cut', 'color', 'clarity' and width.

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

Observation:

We can see that the from the linear plot very strong correlation between the predicted y and actual y with high spread data.This indicates some kind noise present in the data set which shows unexplained variances on the output.

Linear regression Performance Metrics:

- Intercept for the model: 4276.21
- R square on training data: 0.908
- R square on testing data: 0.907
- RMSE on Training data: 1212.41
- RMSE on Testing data: 1230.27

As the training data & testing data score are almost inline we can conclude this model is a Right-Fit Model.

The Gem Stones company should consider the features 'Carat', 'Cut', 'color', 'clarity' and width 'y' as most important for predicting the price.
As we can see from the model,higher the width ('y') of the stone higher the price.

The 'Premium Cut' Diamonds are the most Expensive followed by 'Very Good' Cut should consider in higher profitable stones.

The Diamonds clarity with 'VS1' &'VS2' are the most Expensive. So these two category also consider in higher profitable stones.

Problem Statement 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
edu	Years of formal education
no_young_children	The number of young children (younger than 7 years)
no_older_children	Number of older children
foreign	foreigner Yes/No

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:

The dataset has 7 columns and 872 entries with no duplicated data and all the entries are filled with int and object data types.

Shape: (872, 7)

Data type: int,object

Length: **6** Columns(dropped unwanted column-Unnamed:0)

Duplication: No duplications.

Missing values: No missing values in the dataset.

Summary dataset: Count, mean, std.deviation, range of values and IQR range.

	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 14: Summary dataset

Observation:

- The dataset looks good with no missing data in it and there are no duplications.The minimum to maximum range of all the integer type attributes are listed.

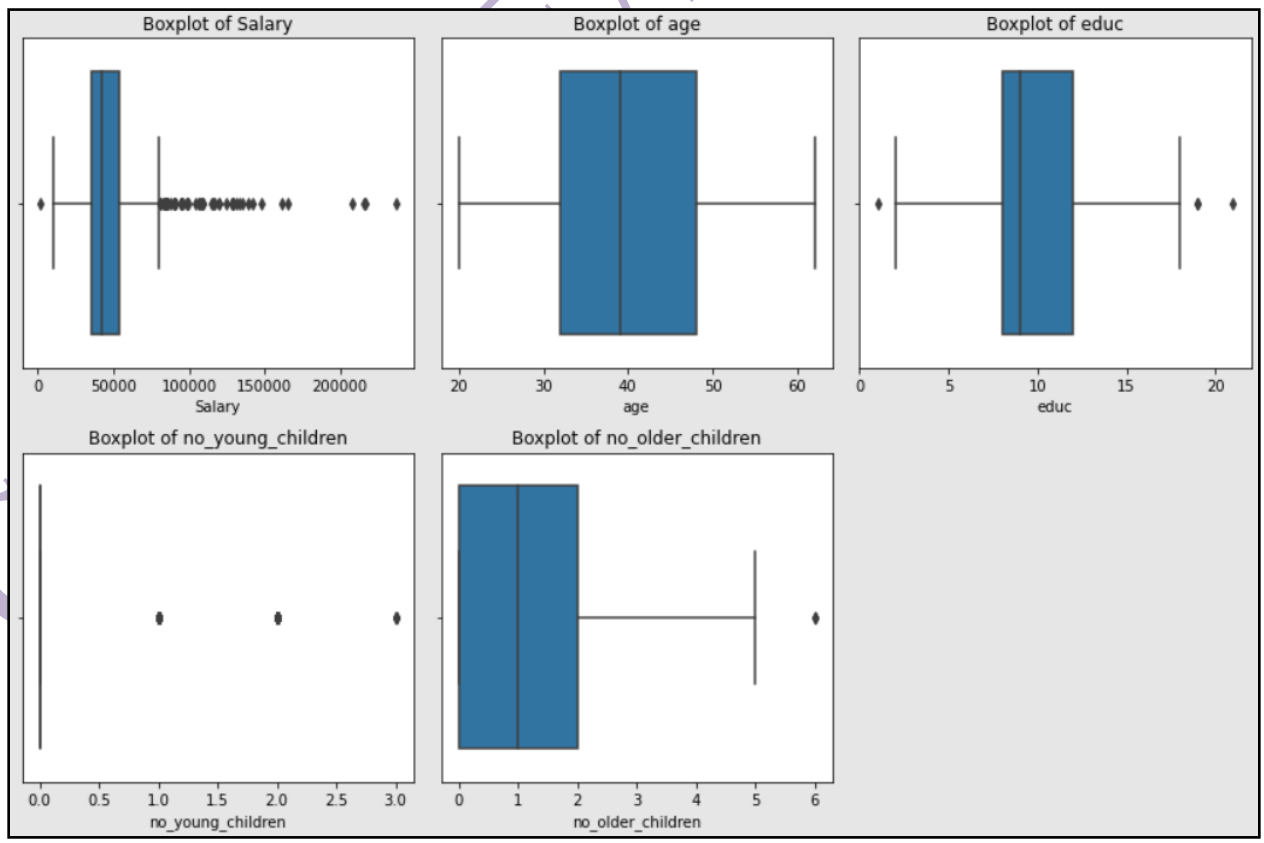

Figure 15:Boxplot

Observation:

- For the univariate analysis, boxplot with outliers are represented.Here all the variables show outlier data. This can be treated in the next plot and replaced with median values to retain them within the dataset.
- Outlier treatment of ‘salary’ attribute is done by imputing it with median value.
- Lower Range of Salary : 8105.75,Upper Range of Salary : 80687.75
- The univariate analysis of categorical valuables are listed using countplot.

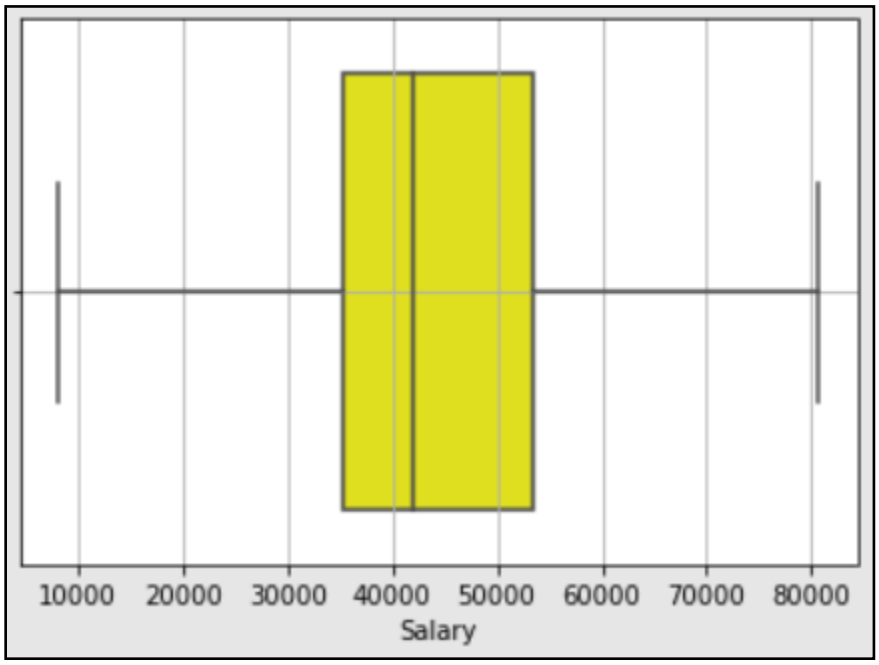

Figure 16:No outlier

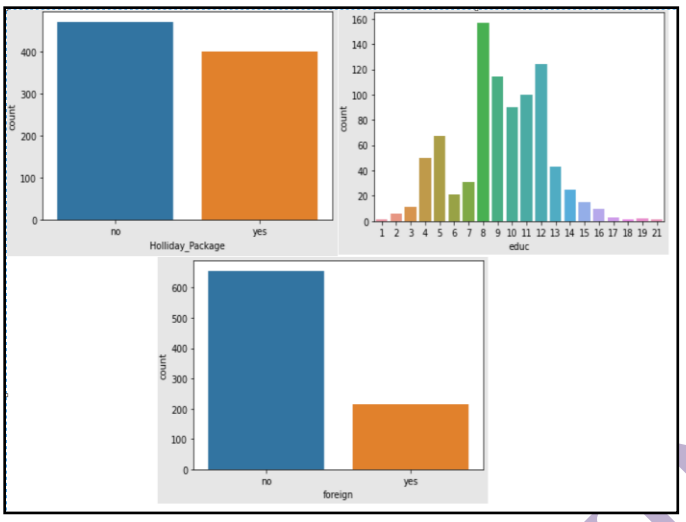

Figure 17:Countplot-Univariate

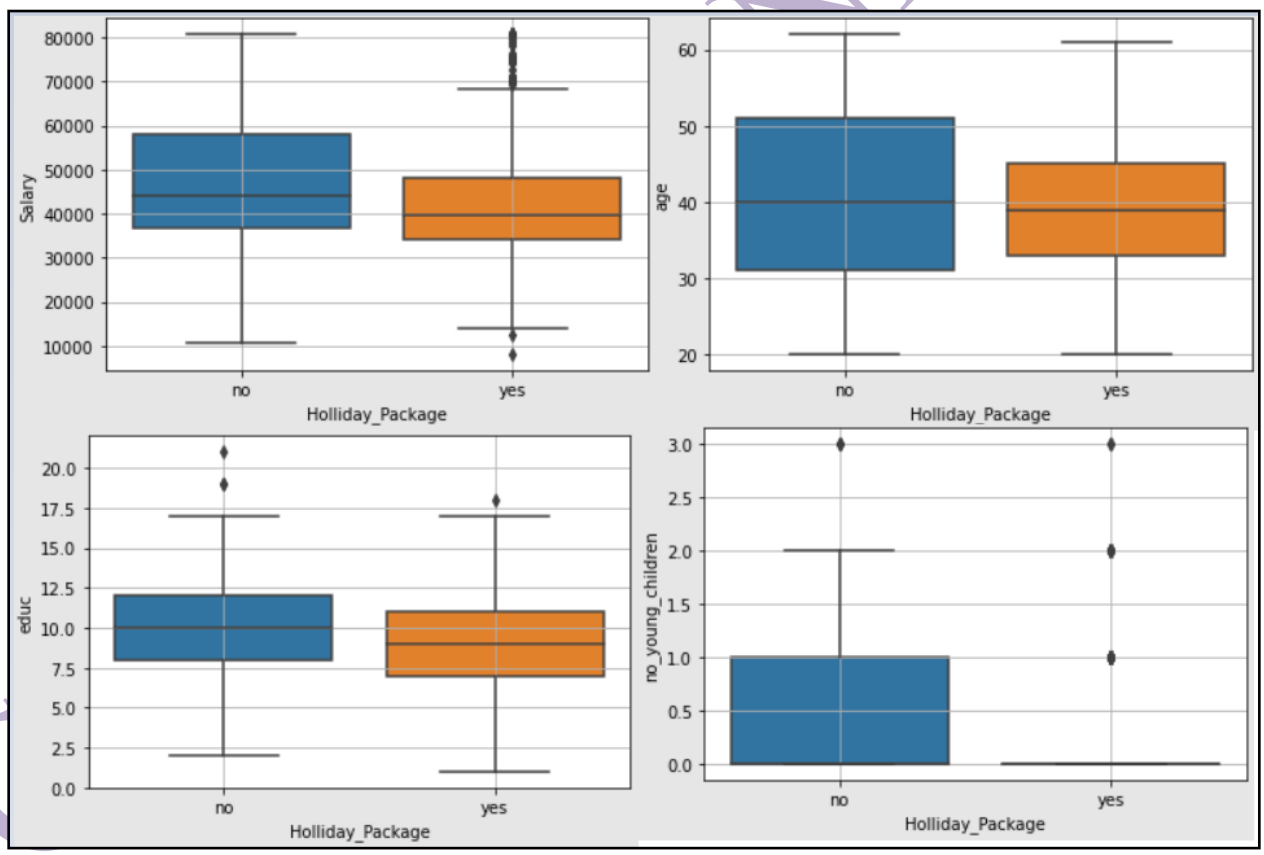

Figure 18:Bivariate-Target

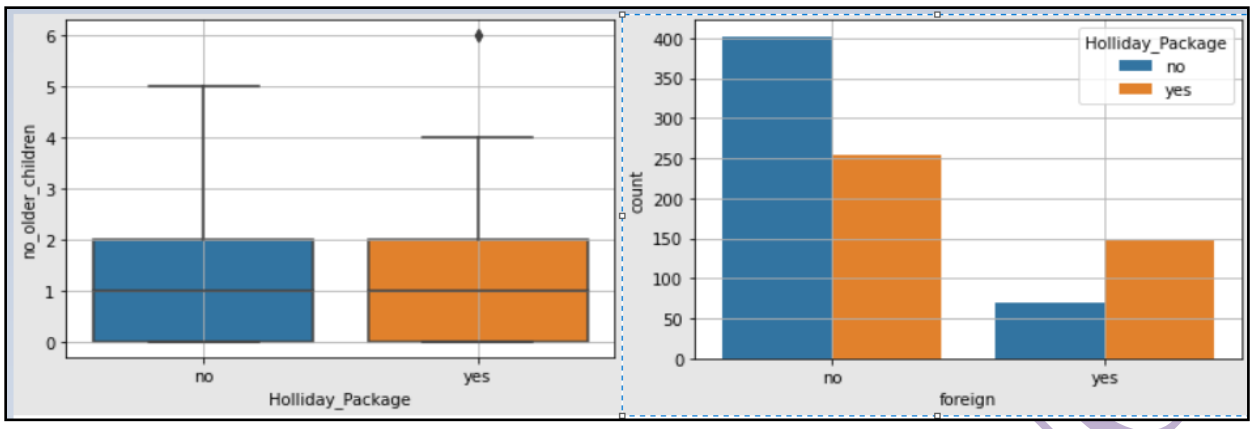

Figure 19:Bivariate-Targets

Observation:

- For the Bivariate analysis, the continuous,categorical and target variables are plotted.The distribution of data is visualized using boxplot.
- People with salary range from 35000 to 49000 opt for holiday package.
- People under the age group 32 -52 doesn't choose holiday package.
- People who had their education of 7-11 years choose the package.
- People with younger children are not opting to take holiday package.
- People with elder children are equally accepting and not accepting the package.
- Foriegners are also interested in taking up the holiday package.

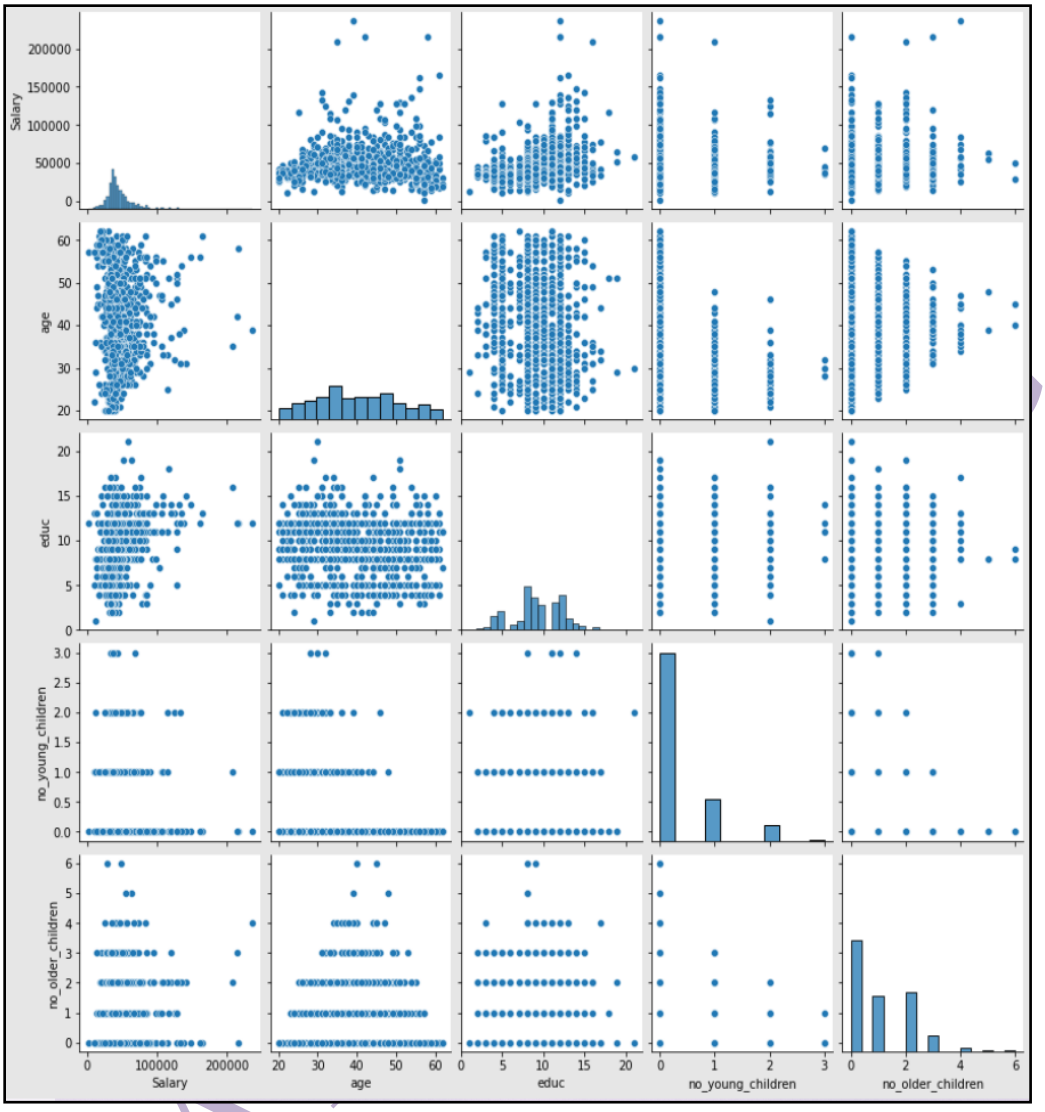

Figure 20:Pairplot

Observation:

- The above pairplot shows correlation between salary,age,education,number of younger children and number of elder children.

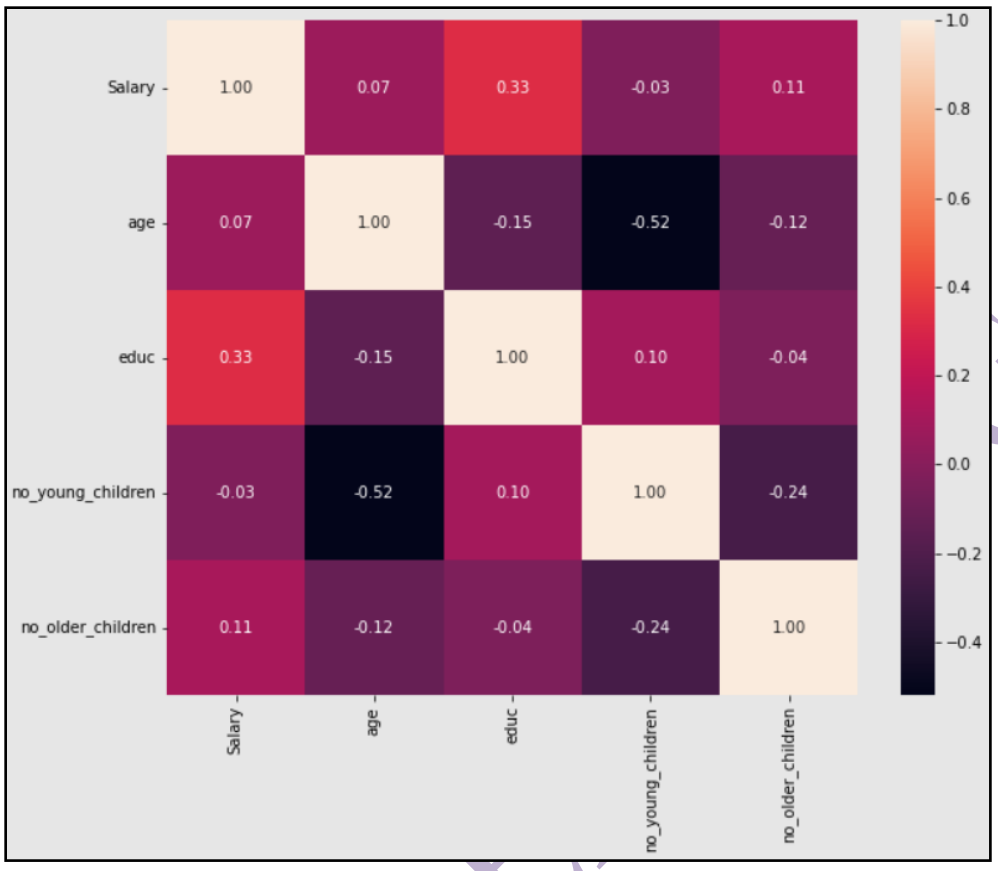

Figure 21:Heatmap

Observation:

- The above heatmap shows hardly any correlation between variables.

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:

Encoding is done to target variable(categorical) to change the datatypes to zero's and one's.

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign_yes
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 22:Encoding

Here the categorical variables 'Holiday_package' and 'Foreign' are encoded with 0 and 1.The variable Foreign is filled with 'yes' or 'no' and the dataset is retained with only the case with'YES'.

Training Split: X_train, Y_train

Test Split: X_test, Y_test.

Logistic Regression is defined as a statistical approach, for calculating the probability outputs for the target labels. In its basic form it is used to classify binary data. Logistic regression is very much similar to linear regression where the explanatory variables(X) are combined with weights to predict a target variable of binary class(y).

Linear Discriminant Analysis (LDA) uses linear combinations of independent variables to predict the class in the response variable. The concept of searching for a linear combination of predictor variables that best separates the classes of the target variable.

The main difference between linear regression and logistic regression is of the type of the target variable.

$$y = 1 /1 + e -(w0+w1x) = e (w0+w1x)/ e (w0+w1x) + 1$$

Dimensions:

No. of rows and columns of the training set for the independent variables: (610, 6)
No. of rows and columns of the training set for the dependent variable: (610,)
No. of rows and columns of the test set for the independent variables: (262, 6)
No. of rows and columns of the test set for the dependent variable: (262,)

Train and Test data Prediction:

The good and bad data in terms of 0's and 1's can be predicted using probability function as shown below.

	0	1
0	0.677845	0.322155
1	0.534493	0.465507
2	0.691845	0.308155
3	0.487745	0.512255
4	0.571939	0.428061

Figure 23:YTest_prob

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:

The logistic probability score function allows us to obtain a predicted probability score of a given event using a logistic regression model.

ROC Curves for Train and test data:

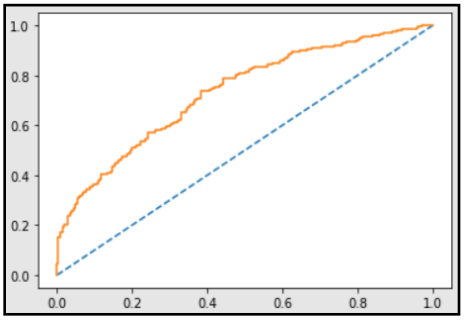

Figure 24:ROC-train

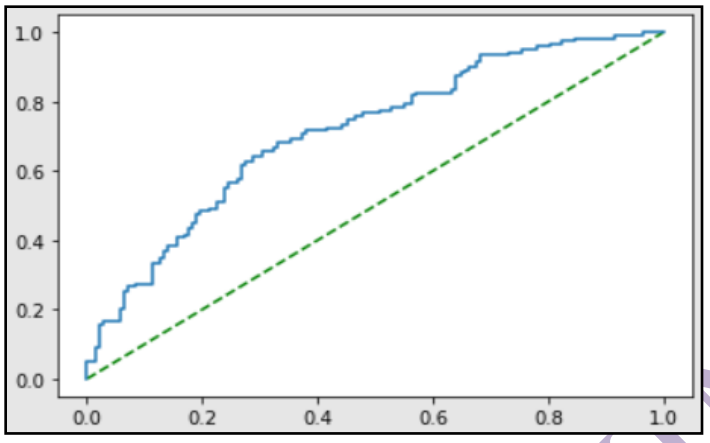

Figure 25:ROC_Test

- AUC score for train and test:(0.731,0.731)
- Confusion matrix(train): array([[244, 85],
  [118, 163]]

	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

Figure 26:Train data Metrics

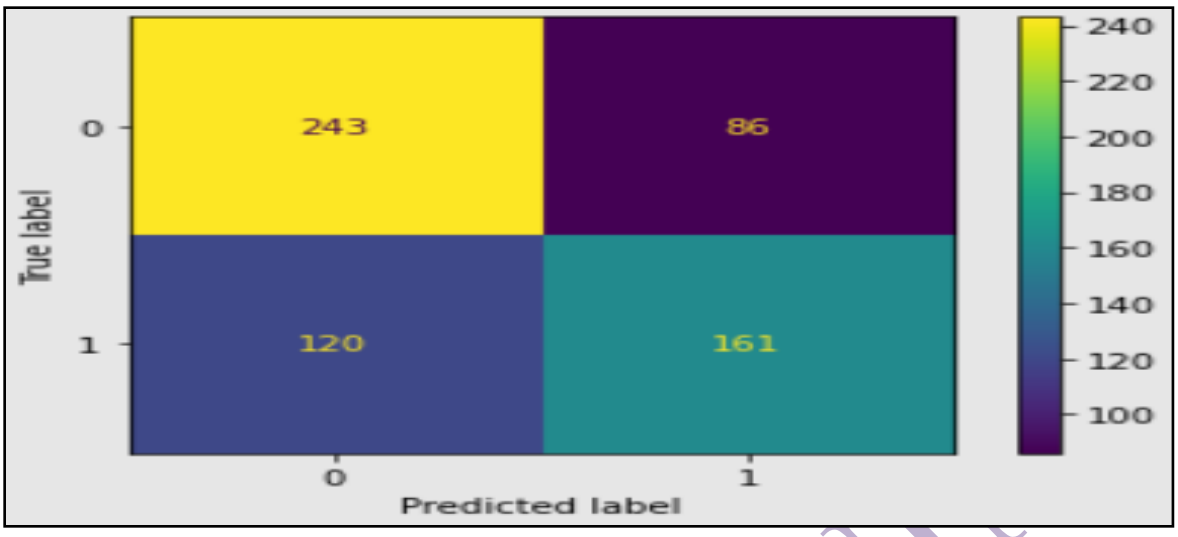

Figure 27:Confusion matrix-train

- Confusion matrix(Test): array([[108, 34],
  [58, 62]]

	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

Figure 28:Test data metrics

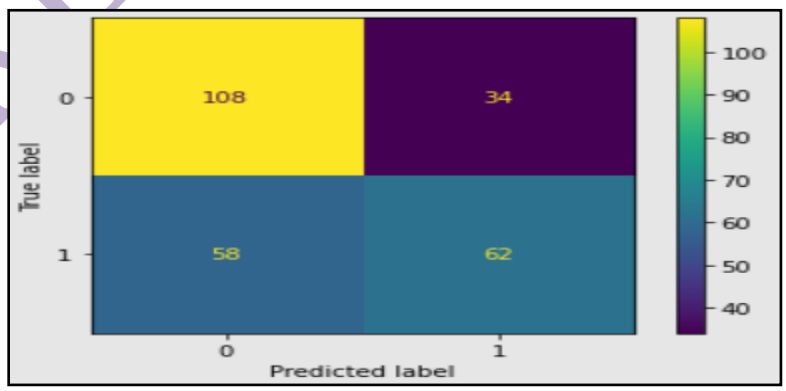

Figure 29:Confusion matrix-Test

The grid search method is used for logistic regression to find the optimal solving and the parameters for solving.

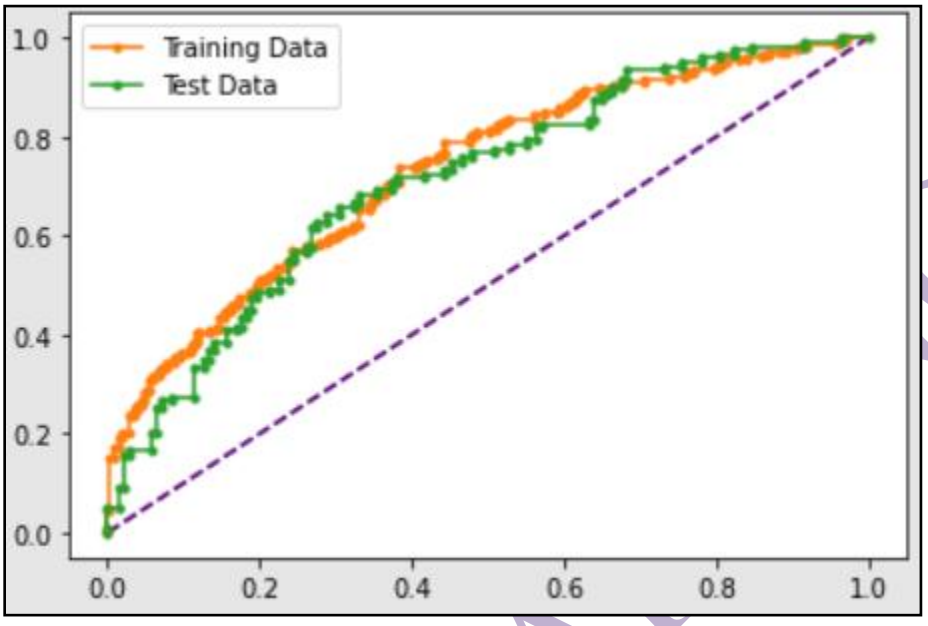

Figure 30:AUC

AUC for the Training Data: 0.731
AUC for the Test Data: 0.714

Observation:

- The comparison of both the models(Logistic and LDA) provides us almost equivalent AUC,ROC scores.
- Both models are working perfect and for the conclusion LDA can be choosen as best.
- Precision tells us how many predictions are actually positive out of all the total positives predicted.
- Recall tells us how many observations of positive class are actually predicted as positive.

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

Solution:

- People above 50 years of age are not interested to take up the holiday packages.
- Age group of 30-50 are interested to choose package.It seems age plays a major role.
- People who have salary less than 50,000 opt for holiday package.
- Education plays an important role in deciding the target.
- Most people who are older wouldn't prefer for holiday package and that can be transferred to any other place which draw their attention.
- Elder children are excited to visit holiday package and it can be enhanced by providing some discounts on it.
- People who earn more than 150000 don't spend much on holidays and their needs can be fulfilled according to their interest.

**********Thank You**********