PREDICTIVE MODELLING

Project

[02/10/2022]

[SUDHEENDRA K]
[PGP DSBA]

PROJECT PM

1.1 Contents[i]

PROJECT PM

PROJECT PM

Problem 1: Linear Regression

PROJECT PM

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 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.

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.**

PROJECT PM

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.**

Table 1. 1 Independent Variables

- *Imported the required libraries.*

	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
26957	26958	2.09	Premium	H	SI2	60.6	59.0	8.27	8.22	5.00	17805
26958	26959	1.37	Premium	E	SI2	61.0	57.0	7.25	7.19	4.40	6751
26959	26960	1.05	Very Good	E	SI2	63.2	59.0	6.43	6.36	4.04	4281
26960	26961	1.10	Very Good	D	SI2	NaN	63.0	6.76	6.69	3.94	4361
26961	26962	0.25	Premium	F	VVS2	62.0	59.0	4.04	3.99	2.49	740
26962	26963	1.11	Premium	G	SI1	62.3	58.0	6.61	6.52	4.09	5408
26963	26964	0.33	Ideal	H	IF	61.9	55.0	4.44	4.42	2.74	1114
26964	26965	0.51	Premium	E	VS2	61.7	58.0	5.12	5.15	3.17	1656
26965	26966	0.27	Very Good	F	VVS2	61.8	56.0	4.19	4.20	2.60	682
26966	26967	1.25	Premium	J	SI1	62.0	58.0	6.90	6.88	4.27	5166

Table 1. 2 Tail of the dataset

Starting on with loading the data, we could see the tail of the data set from the above figure and can get a glimpse of the dataset and let's do a detailed EDA

PROJECT PM

Exploratory Data Analysis (EDA)

- *EDA is understanding the data sets by summarizing their main characteristics often plotting them visually.*

- *This step is very important especially when we arrive at modelling the data. Plotting in EDA consists of Histograms, box plot, pair plot and many more.*

- *It often takes much time to explore the data. Through the process of EDA, we can define the problem statement or definition on our data set which is very important.*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26967 entries, 0 to 26966
Data columns (total 11 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Unnamed: 0  26967 non-null  int64
 1   carat       26967 non-null  float64
 2   cut         26967 non-null  object
 3   color       26967 non-null  object
 4   clarity     26967 non-null  object
 5   depth       26270 non-null  float64
 6   table       26967 non-null  float64
 7   x           26967 non-null  float64
 8   y           26967 non-null  float64
 9   z           26967 non-null  float64
 10  price       26967 non-null  int64
dtypes: float64(6), int64(2), object(3)
memory usage: 2.3+ MB
```

Table 1. 3 Dataset Information

- *Data set has 26,967 rows with 11 variables.*

PROJECT PM

- *Column indicating row number (Unnamed:0) cannot be used for analysis and needs to be deleted.*
- *Excluding row number data set has 3 categorical variables and 7 numerical variables*
- *Price is dependent variable and other 9 independent (predictive variables)*
- *There are 697 'Null Values' in variable 'depth'*
- *Checking for Duplicates: - There are 43 duplicate rows in the dataset*
- *We will drop the first column 'Unnamed: 0' column as this is not important for our study.*

	count	mean	std	min	25%	50%	75%	max
carat	26967.0	0.793593	0.462431	0.200	0.40	0.70	1.05	2.025
cut	26967.0	1.627953	0.539964	0.000	1.00	2.00	2.00	2.000
color	26967.0	1.931546	1.068930	0.000	1.00	2.00	3.00	4.000
clarity	26967.0	2.212037	0.846867	0.000	2.00	2.00	3.00	4.000
depth	26967.0	61.750502	1.218929	59.000	61.10	61.80	62.50	64.600
table	26967.0	57.435699	2.157125	51.500	56.00	57.00	59.00	63.500
x	26967.0	5.729903	1.127023	1.950	4.71	5.69	6.55	9.310
y	26967.0	5.731798	1.118970	1.965	4.71	5.71	6.54	9.285
z	26967.0	3.537261	0.697278	1.190	2.90	3.52	4.04	5.750
price	26967.0	3737.914136	3470.888236	326.000	945.00	2375.00	5360.00	11982.500

Table 1. 4 Data Description

From the above table the overall statistics can be observed.

PROJECT PM

Univariate Analysis

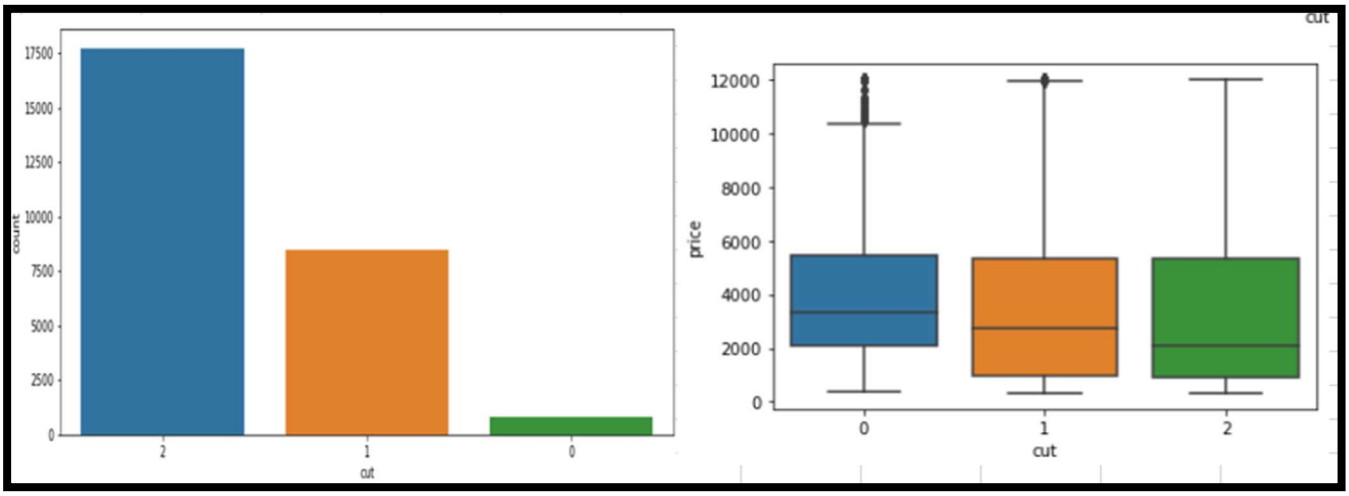


Figure 1. 1 Cut Variable

- *Looking at the above unique values for variable "Cut" we see the ranking given for each unique value like " Fair, Good, Ideal, Premium, Very Good "*
- *But I shall cut down the ordinal level to only three levels that is to Fair(0), Good&VeryGood(1) and Ideal&Premium(2)*
- *For the cut variable we see the most sold is Ideal&Premium cut type gems and least sold is Fair cut gems*
- *All cut type gems have outliers with respect to price*
- *Slightly less priced seems to be Ideal type and premium cut type to be slightly more expensive*

PROJECT PM

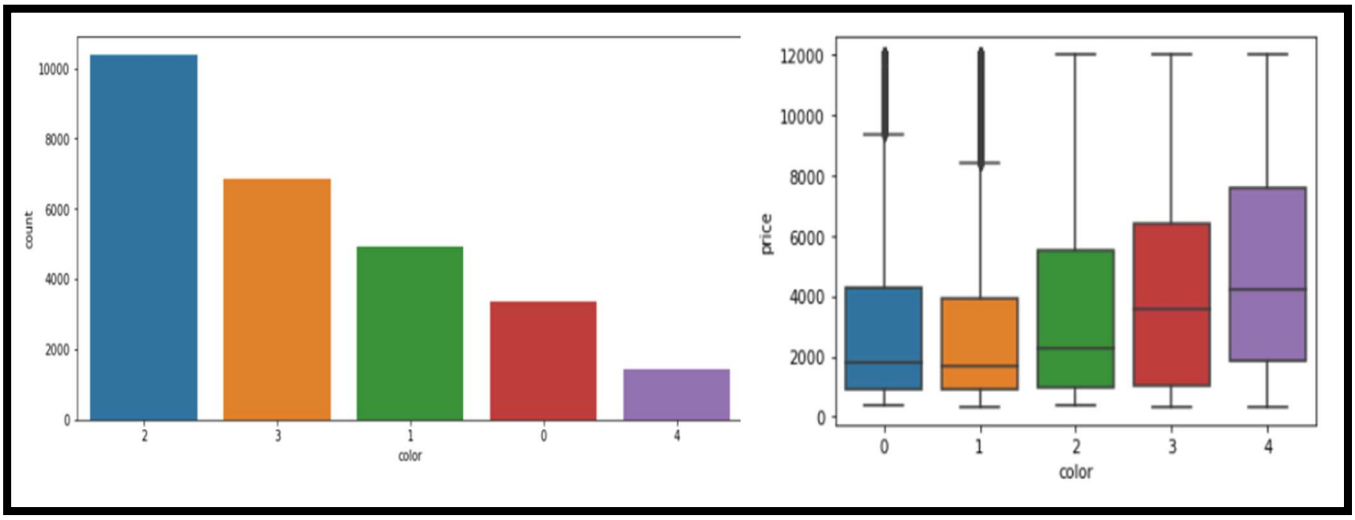


Figure 1. 2 Colour Variable

- *For the color variable we see the most sold is G colored gems and least is J colored gems*
- *All color type gems have outliers with respect to price*
- *However, the least priced seems to be E type; J and I colored gems seems to be more expensive*

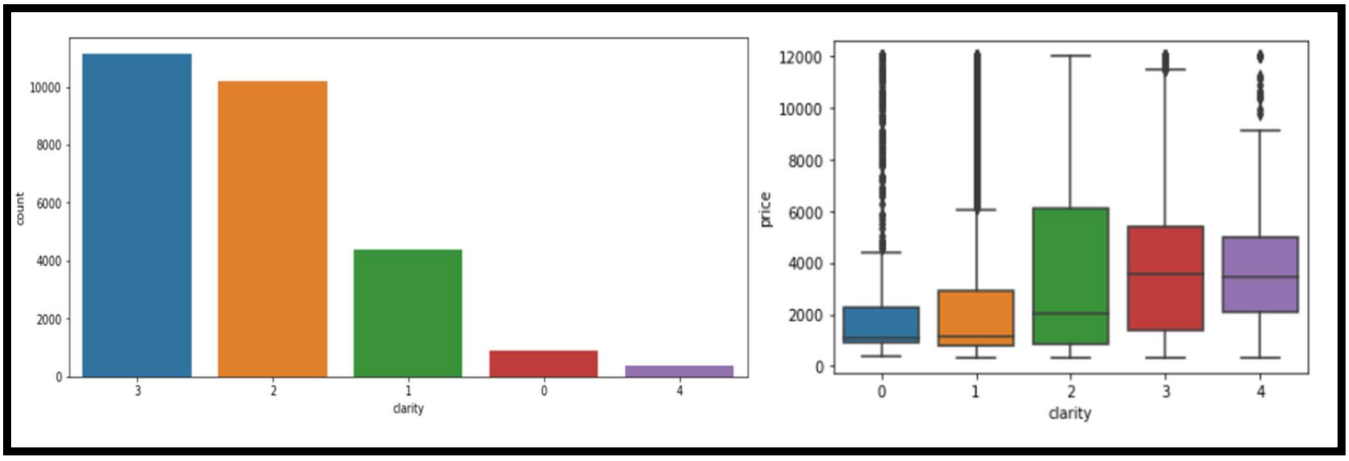


Figure 1. 3 Clarity Variable

PROJECT PM

- *For the clarity variable we see the most sold is SI1 clarity gems and least is I1 clarity gems*

- *All clarity type gems have outliers with respect to price*

- *Slightly less priced seems to be SI1 type; VS2 and SI2 clarity stones seems to be more expensive*

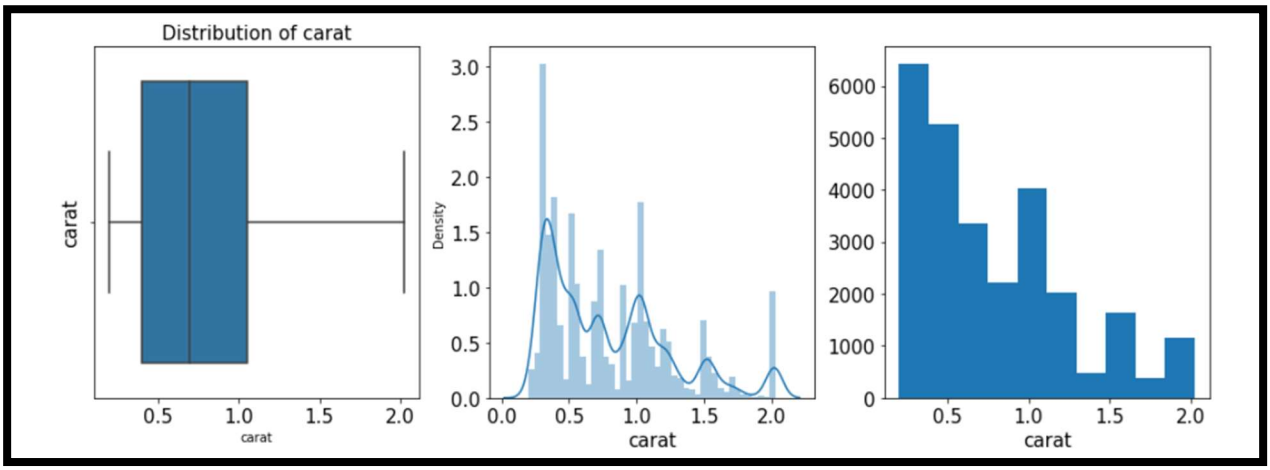


Figure 1. 4 Carat Variable

- *Carat an independent variable, and it ranges from 0.2 to 2.025. mean value is around 0.79 and 75% of the stones are of 2.00 carat value.*

- *Standard deviation is around 0.4624 which shows that the data is skewed and has a right tailed curve. Which means that majority of the stones are of lower carat. There are very few stones above 1.05 carat.*

PROJECT PM

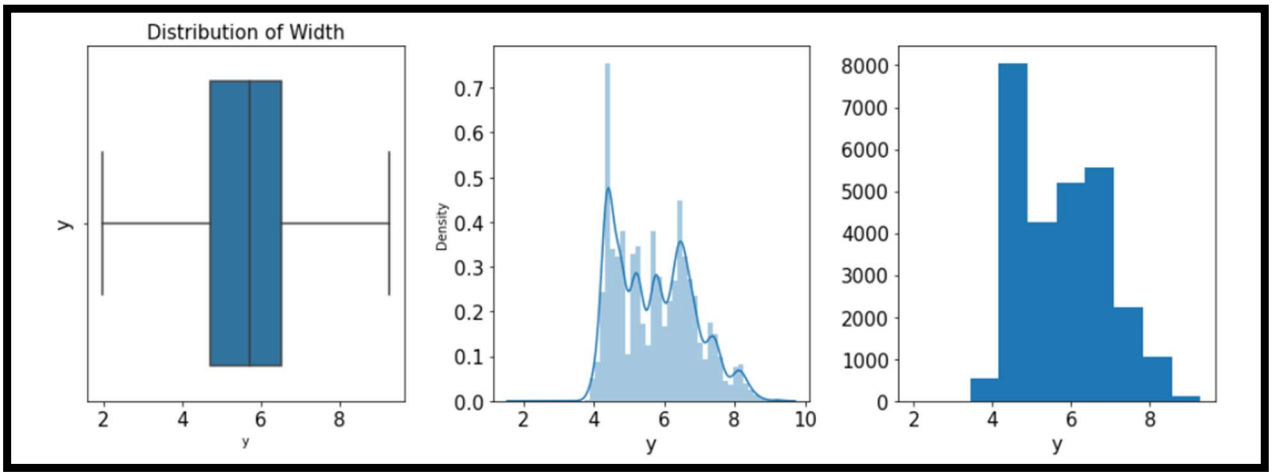


Figure 1. 5 Width Variable

- *Width is an independent variable, and it ranges from 1.9650 to 9.210. mean value is around 5.73 and 75% of the stones are of 6.54 width.*

- *Standard deviation is around 1.1189 which shows that the data is skewed and has a right tailed curve. Which means that majority of the stones are of lower width.*

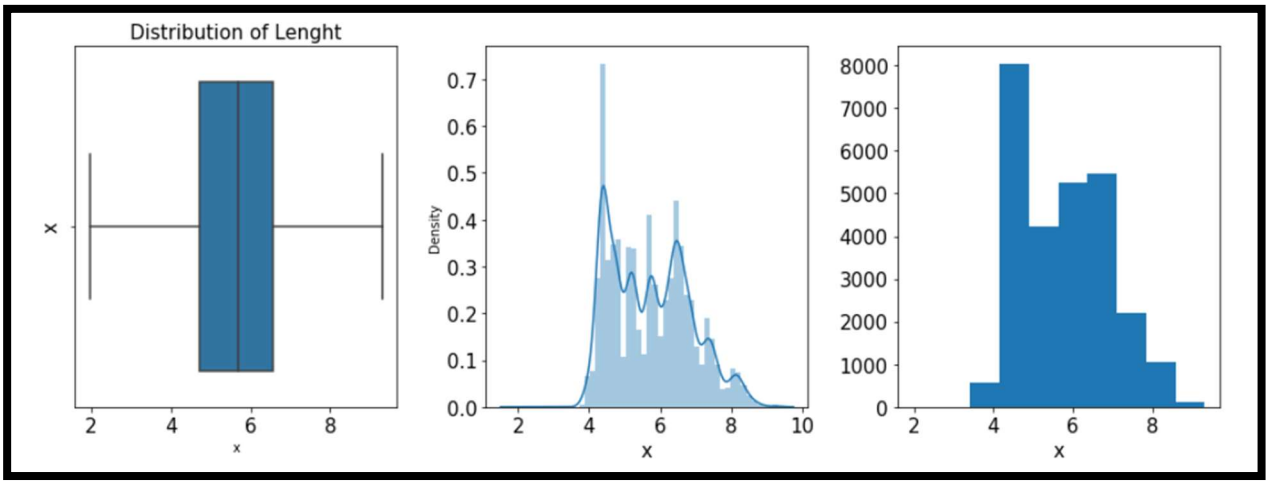


Figure 1. 6 Length Variable

- *Lenght is an independent variable, and it ranges from 1.950 to 9.310. mean value is around 5.72 and 75% of the stones are of 6.55 lenght.*

PROJECT PM

- *Standard deviation is around 1.127 which shows that the data is skewed and has a right tailed curve. Which means that majority of the stones are of lower length.*

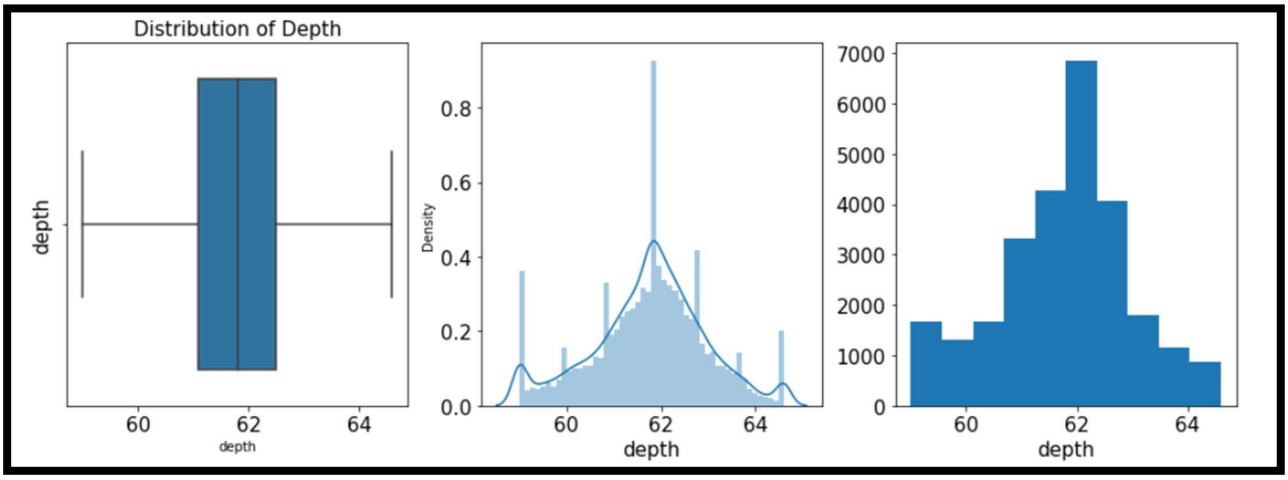


Figure 1. 7 Depth Variable

- *Depth, the percentage height of cubic zirconia stones is in the range of 50.80 to 73.60.*
- *Average height of the stones is 61.80 25% of the stones are 61 and 75% of the stones are 62.5.*

- *Standard deviation of the height of the stones is 1.4. Standard deviation is indicating a normal distribution*

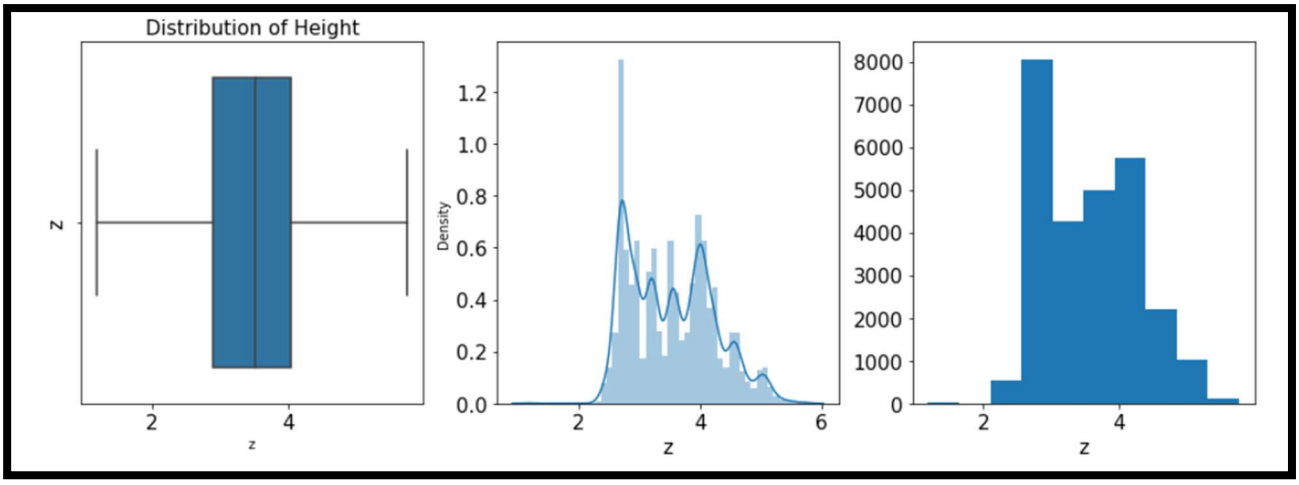


Figure 1. 8 Height Variable

- *Height is an independent variable, and it ranges from 1.190 to 5.750. mean value is around 3.5372 and 75% of the stones are of 4.04 height.*

- *Standard deviation is around 0.697 which shows that the data is skewed and has a right tailed curve. Which means that majority of the stones are of lower height.*

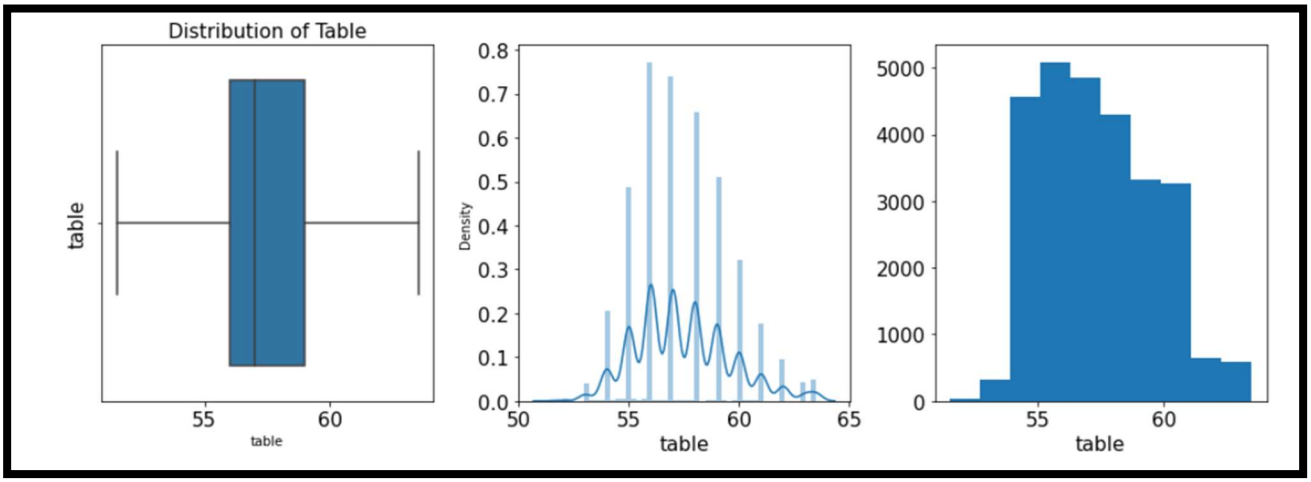


Figure 1. 9 Table Variable

- *Table, the percentage width of cubic Zirconia is in the range of 51 to 63.*

- *Average is around 57. 43% of stones are below 56 and 75% of the stones have a width of less than 59.*

- *Standard deviation is 2.15. Thus the data does not show normal distribution and is similar to carat with most of the stones having less width*

- *This shows outliers are present in the variable.*

PROJECT PM

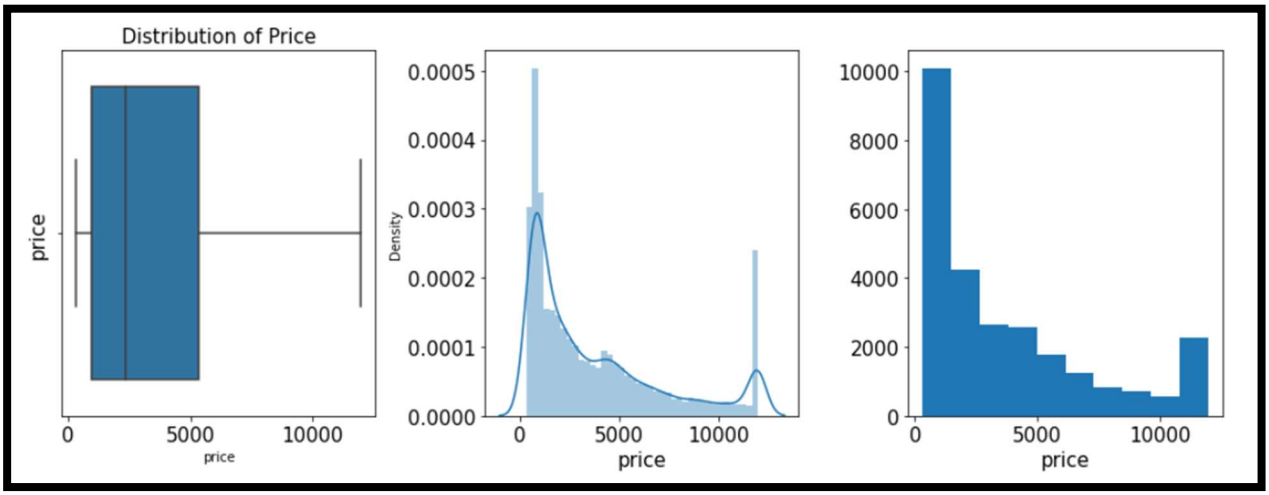


Figure 1. 10 Price Variable

- *Price is the Predicted variable.*
- *Prices are in the range of 3470.88 to 11982.*
- *Median price of stones is 2375, while 25% of the stones are priced below 945.*
- *75% of the stones are in the price range of 5360.*
- *Standard deviation of the price is 3470. Indicating prices of majority of the stones are in lower range as the distribution is right skewed.*
- *Variables x, y, and z seems to follow a normal distribution with a few outliers.*

PROJECT PM

Bivariate Analysis

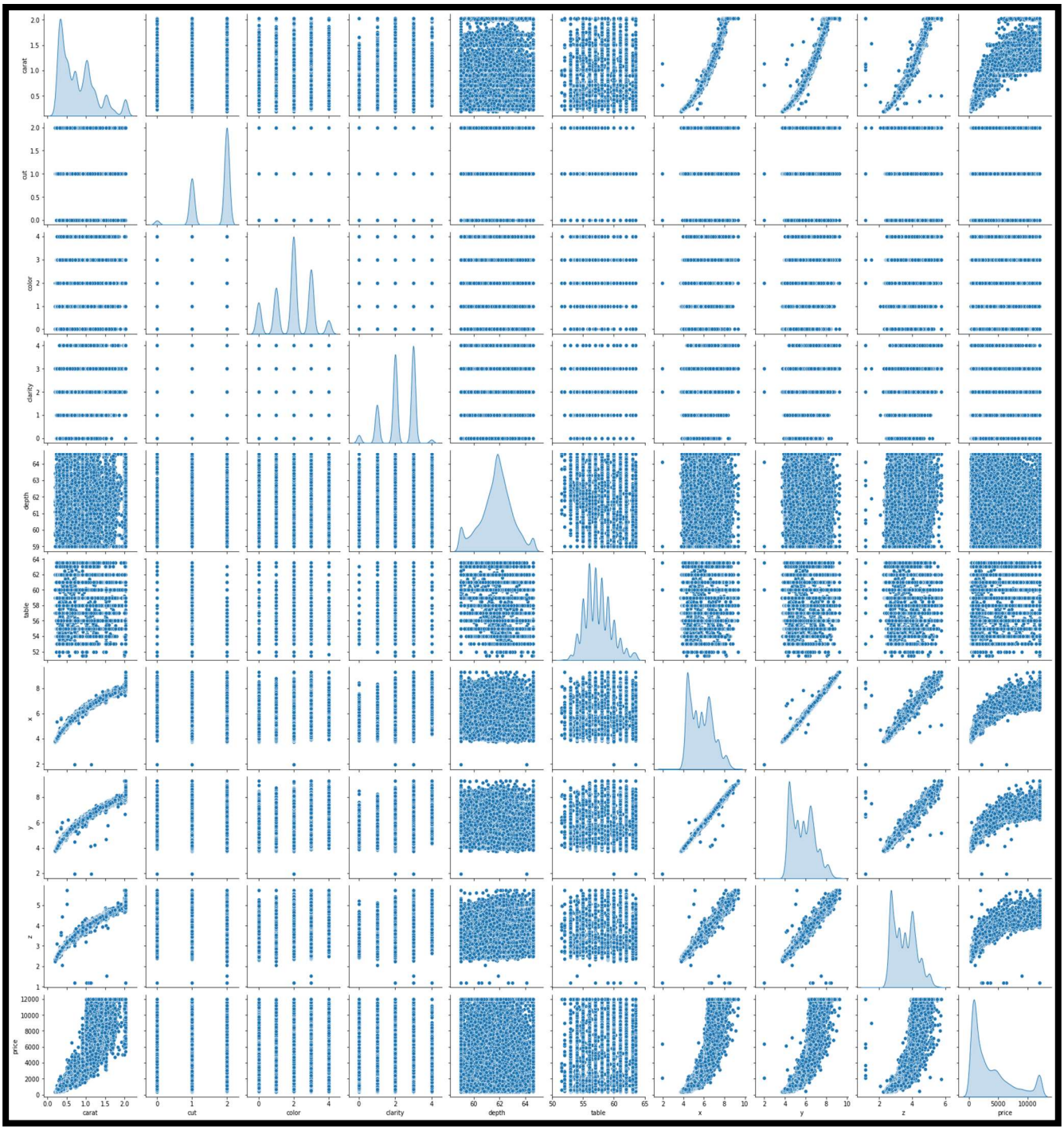


Figure 1. 11 Pair Plot

PROJECT PM

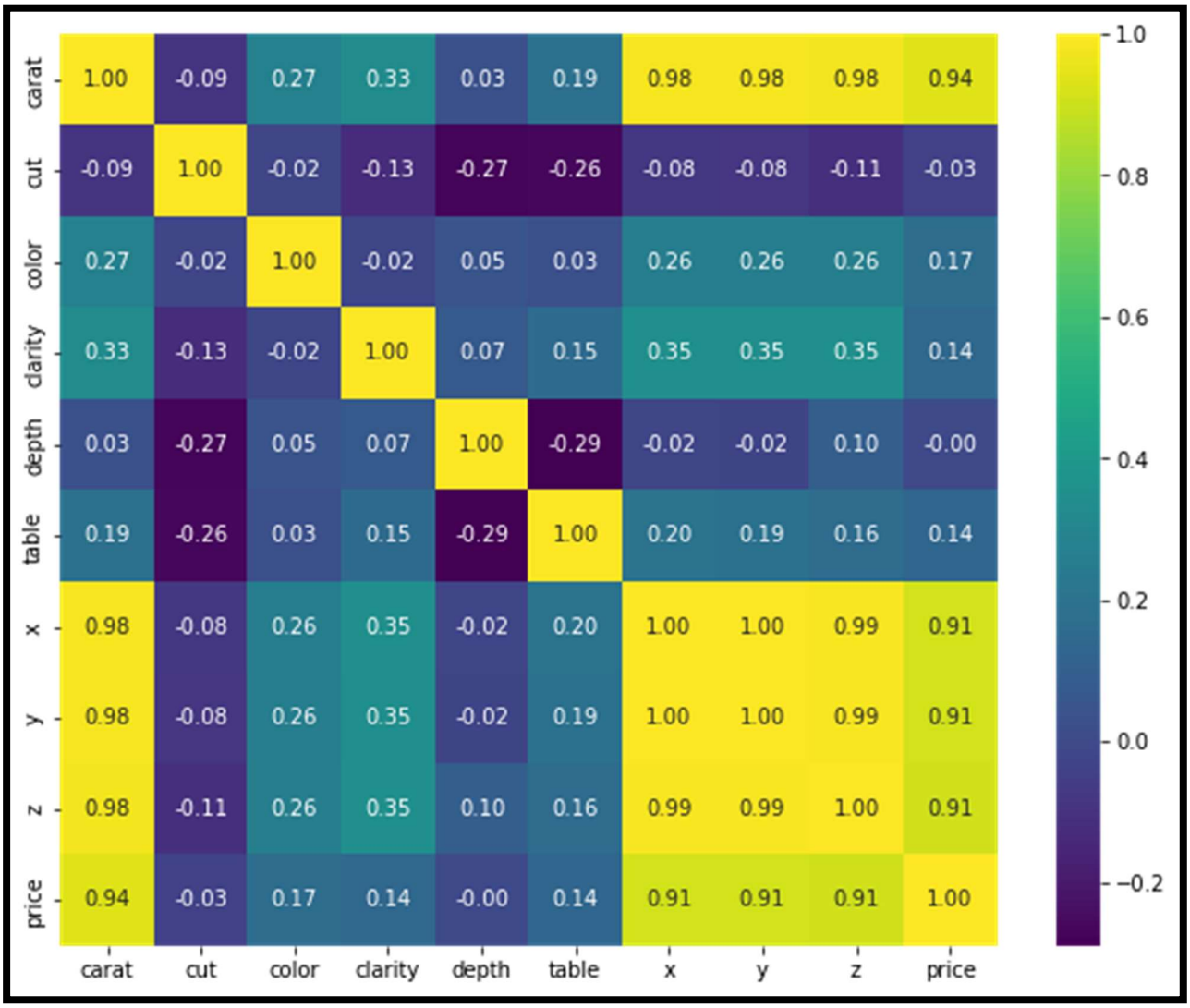


Figure 1. 12 Heat Map

- *Pair plot allows us to see both distribution of single variable and relationships between two variables.*

- *We can also observe the correlation value from the heat map.*

Conclusion of EDA:

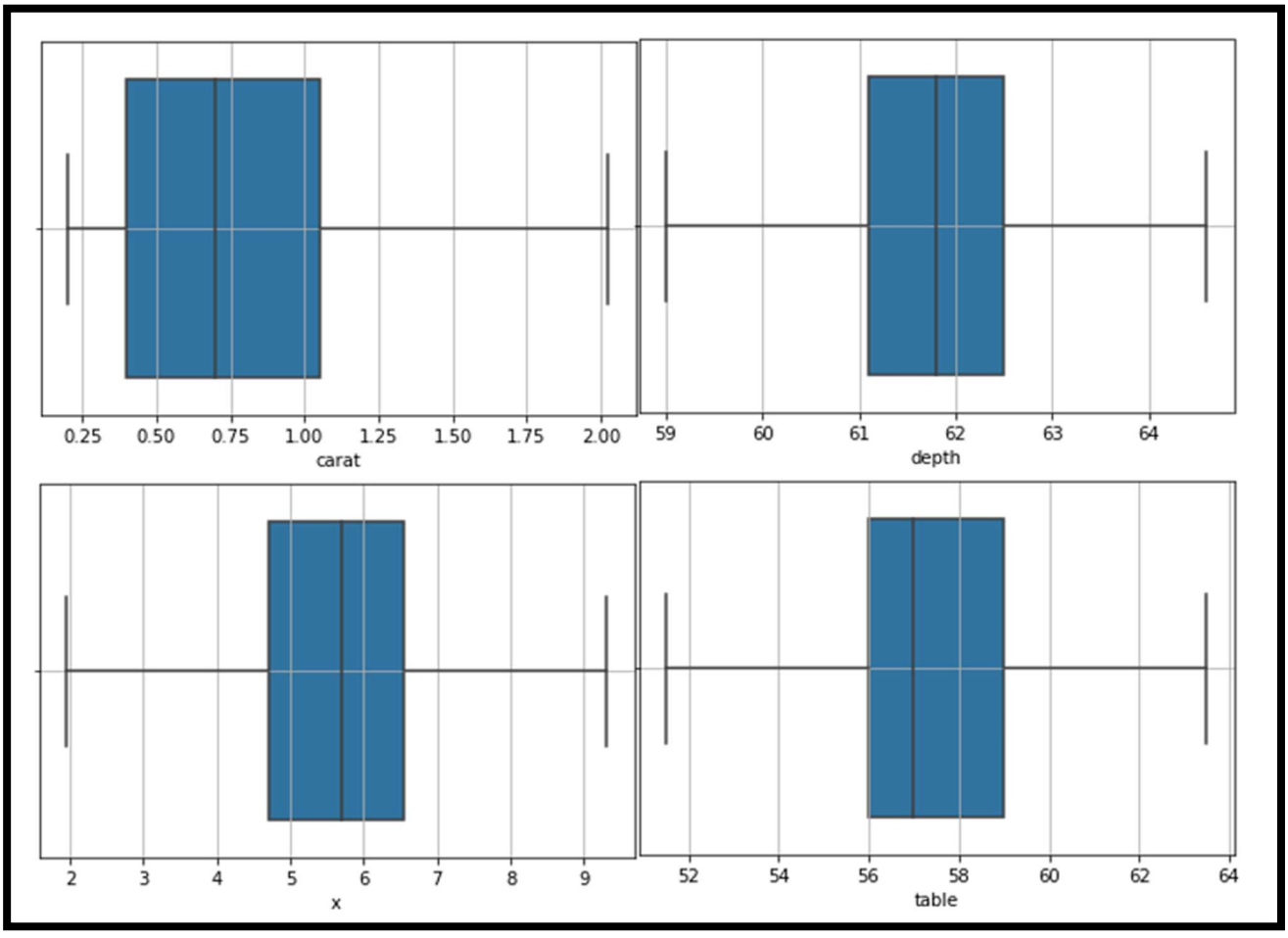
PROJECT PM

- *Price – This variable gives the continuous output with the price of the cubic zirconia stones. This will be our Target Variable*
  .
- *Carat, depth, table, x, y, z variables are numerical or continuous variables.*

- *Cut, Clarity and colour are categorical variables.*

- *We will drop the first column 'Unnamed: 0' column as this is not important for our study which leaves the shape of the dataset with 26967 rows & 10 Columns.*

- *Shall be treating outliers from IQR method and have glimpse of boxplot after treating outliers from the below figure*



PROJECT PM

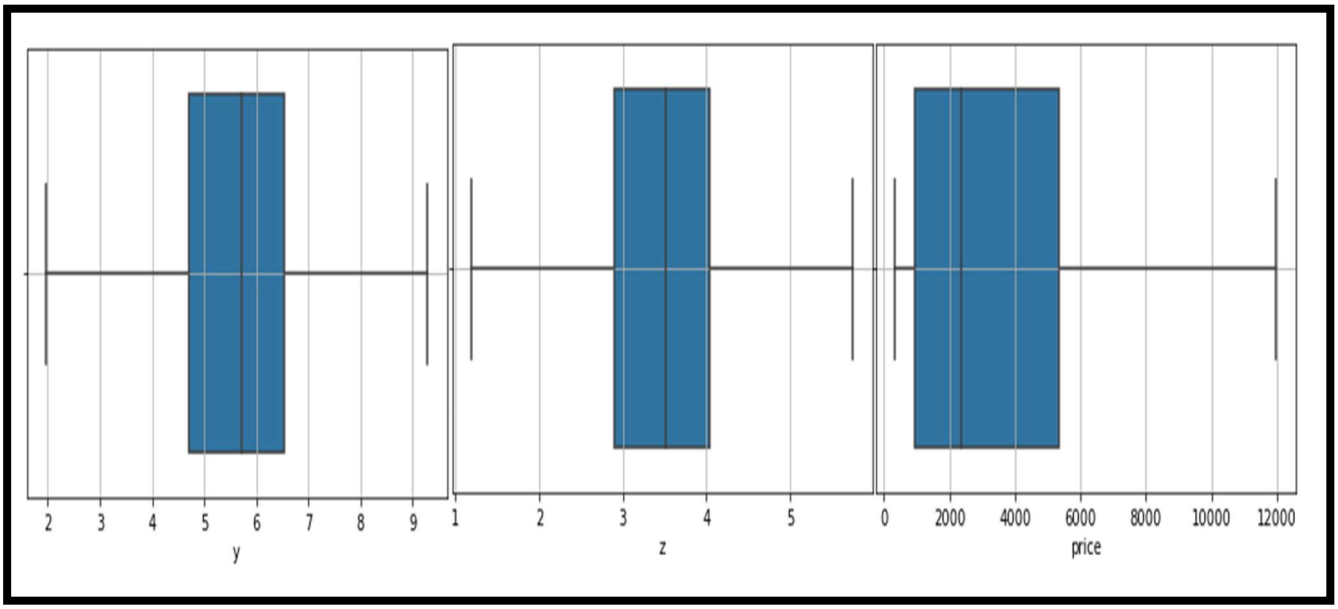


Figure 1. 13 Boxplot after outliers' treatment

1.3 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.

- *We will drop the first column 'Unnamed: 0' column as this is not important for our study*

PROJECT PM

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

```
3    11146
2    10192
1     4370
0      894
4      365
Name: clarity, dtype: int64
```

```
2    17715
1     8471
0      781
Name: cut, dtype: int64
```

```
2    10390
3     6873
1     4917
0     3344
4     1443
Name: color, dtype: int64
```

Variables counts

Snippet 1. 1 Null & Categorical Variable counts

- *Only in 'depth 697missing values are present which we will impute by its median values and we could see from the above figure that every null value is imputed and found 0 in all the variables*

```
<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  int64
 2   color    26967 non-null  int64
 3   clarity  26967 non-null  int64
 4   depth    26967 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  float64
dtypes: float64(7), int64(3)
memory usage: 2.1 MB
```

Table 1. 5 Information after Converting all into numerical datatypes

- *We now have a dataset with 26,967 rows and 10 variable columns with all numerical variables*

PROJECT PM

- *Please note, cantering/scaling does not affect our statistical inference in regression models - the estimates are adjusted appropriately, and the p-values will be the same therefore, ill not doing scaling process*

1.4 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.

- *It shall be hard to pick an order and the variable seems to be having a direct impact on the price variable.*

- *In one-hot encoding, the integer encoded variable is removed and a new binary variable is added for each unique integer value.*

- *A one hot encoding allows the representation of categorical data to be more expressive. Many machine learning algorithms cannot work with categorical data directly and hence, the categories must be converted into numbers.*

- *This is required for both input and output variables that are categorical. The only disadvantage being, that for high cardinality, the feature space can really blow up quickly and we will then need to struggle with dimensionality.*

- *For this case study, I have used one-hot encoding for color and clarity independent variables using the function referred as dummy encoding, which takes into consideration the (Kn-1) encoding. Dummy encoding converts it into n-1 variables.*

PROJECT PM

	carat	cut	color	depth	table	x	y	z	price	clarity_0	clarity_1	clarity_2	clarity_3	clarity_4
0	0.30	2	1	62.1	58.0	4.27	4.29	2.66	499.0	0	0	0	1	0
1	0.33	2	2	60.8	58.0	4.42	4.46	2.70	984.0	1	0	0	0	0
2	0.90	1	1	62.2	60.0	6.04	6.12	3.78	6289.0	0	1	0	0	0
3	0.42	2	2	61.6	56.0	4.82	4.80	2.96	1082.0	0	0	1	0	0
4	0.31	2	2	60.4	59.0	4.35	4.43	2.65	779.0	0	1	0	0	0

Table 1. 6 Head of the data after Encoding

- *One-hot encoding ends up with kn variables, while dummy encoding usually ends up with kn-1 variables. This will also help us to deal with the issue of multidimensionality, if needed.*

- *Head of the data can be seen from the above table after encoding.*

	carat	cut	color	depth	table	x	y	z	clarity_0	clarity_1	clarity_2	clarity_3	clarity_4
0	0.30	2	1	62.1	58.0	4.27	4.29	2.66	0	0	0	1	0
1	0.33	2	2	60.8	58.0	4.42	4.46	2.70	1	0	0	0	0
2	0.90	1	1	62.2	60.0	6.04	6.12	3.78	0	1	0	0	0
3	0.42	2	2	61.6	56.0	4.82	4.80	2.96	0	0	1	0	0
4	0.31	2	2	60.4	59.0	4.35	4.43	2.65	0	1	0	0	0

X-HEAD

	price
0	499.0
1	984.0
2	6289.0
3	1082.0
4	779.0

Y-HEAD

Table 1. 7 Splitting the data into dependent and independent variable

- *Splitting the data into dependent(Y) and independent(X) variable for further process to modelling*

- *Copy all the predictor variables into X data frame and copy target into the y data frame.*

- *Using the dependent variable, we split the X and Y data frames into training set and test set.*

PROJECT PM

```
The coefficient for carat is 8782.24770740607
The coefficient for cut is 201.13620218118095
The coefficient for color is -387.6295437727388
The coefficient for depth is -16.324546980177992
The coefficient for table is -27.297628476020584
The coefficient for x is -1655.5189911224459
The coefficient for y is 1381.306603412922
The coefficient for z is -180.003696480504
The coefficient for clarity_0 is 1379.0564210900925
The coefficient for clarity_1 is 1145.250841182612
The coefficient for clarity_2 is 543.6153883359477
The coefficient for clarity_3 is -413.99345520321043
The coefficient for clarity_4 is -2653.9291954054584

The intercept for our model is 1750.3835494016575
```

Snippet 1. 2 Coefficients and Intercept of the model

- *For this we use the Sklearn package and then split X and Y in 70:30 ration and then invoke the linear regression function and find the best fit model on training data.*

- *The intercept for our model is 1750.3835494016575*

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

```
0.9304131253151822
```

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

```
0.9287566116845567
```

Snippet 1. 3 R-Square score on training and testing data

PROJECT PM

- *R square on training data : 0.9304*
- *R square on testing data : 0.9287*

- *R square is the percentage of the response variable variation that is explained by a linear model and computed by the formula as:*

- *R-square = Explained Variation / Total Variation.*

- *It is always between 0 and 100%, in which 0% indicates that the model explains none of the variability of the response data around its mean and 100% indicates that the model explains all the variability of the response data around its mean.*

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

```
917.5667264501519
```

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

```
921.701330479256
```

Snippet 1. 4 RMSE on Training and testing data

- *In the regression model we can see the R-square value on training and test data respectively as 0.930413 and 0.9287566*

- *The RMSE on training and test data respectively is 917.56672 and 921.70133047.*

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

PROJECT PM

OLS Regression Results

Dep. Variable:	price	R-squared:	0.930
Model:	OLS	Adj. R-squared:	0.930
Method:	Least Squares	F-statistic:	2.102e+04
Date:	Sat, 01 Oct 2022	Prob (F-statistic):	0.00
Time:	23:42:15	Log-Likelihood:	-1.5555e+05
No. Observations:	18876	AIC:	3.111e+05
Df Residuals:	18863	BIC:	3.112e+05
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1458.6530	559.614	2.607	0.009	361.758	2555.547
carat	8782.2477	80.834	108.645	0.000	8623.805	8940.690
cut	201.1362	14.186	14.178	0.000	173.330	228.942
color	-387.6295	6.564	-59.050	0.000	-400.496	-374.763
depth	-16.3245	8.701	-1.876	0.061	-33.378	0.729
table	-27.2976	3.644	-7.491	0.000	-34.440	-20.155
x	-1655.5190	133.443	-12.406	0.000	-1917.080	-1393.958
y	1381.3066	133.889	10.317	0.000	1118.871	1643.742
z	-180.0037	88.365	-2.037	0.042	-353.206	-6.801
clarity_0	1670.7870	114.535	14.588	0.000	1446.288	1895.286
clarity_1	1436.9814	112.576	12.765	0.000	1216.322	1657.641
clarity_2	835.3460	113.227	7.378	0.000	613.411	1057.281
clarity_3	-122.2629	114.104	-1.072	0.284	-345.917	101.391
clarity_4	-2362.1986	124.084	-19.037	0.000	-2605.415	-2118.982

Omnibus:	3909.153	Durbin-Watson:	1.979
Prob(Omnibus):	0.000	Jarque-Bera (JB):	11631.697
Skew:	1.078	Prob(JB):	0.00
Kurtosis:	6.184	Cond. No.	5.73e+17

Table 1. 8 OLR Regression Summary

- *From the OLR summary, we can see that the p value is 0.061 for depth variable, which is much greater than 0.05 and clarity_3 is of .284 which means this attribute is of no use.*

- *The sample linear equation after running the model:*

(1458.65) * Intercept + (8782.25) * carat + (201.14) * cut + (-387.63) * color + (-16.32) * depth + (-27.3) * table + (-1655.52) * x + (1381.31) * y + (-180.0) * z + (1670.79) * clarity_0 + (1436.98) * clarity_1 + (835.35) * clarity_2 + (-122.26) * clarity_3 + (-2362.2) * clarity_4 +

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

```
917.5667264501532
```

```
np.sqrt(lm1.mse_resid)
```

```
917.8828562533132
```

Snippet 1. 5 Root Mean Squared Error

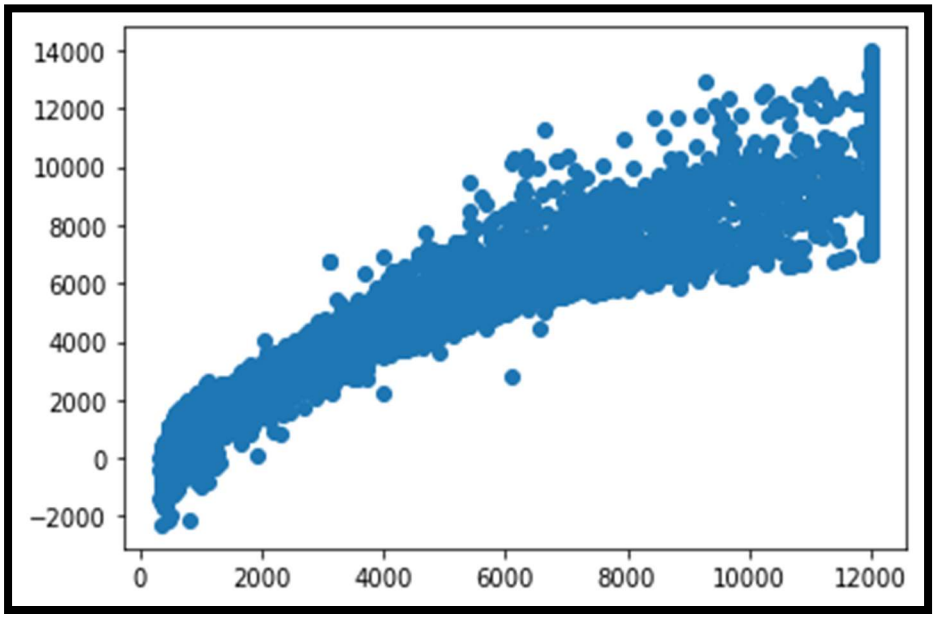


Figure 1. 14 Prediction on the test dataset

- *Scatterplot on test data between dependent variable – price - and independent variables*

PROJECT PM

```
carat ---> 31.391970954394754
cut ---> 1.3045691587767188
color ---> 1.101699261016668
depth ---> 2.7699573340022448
table ---> 1.3676517936505284
x ---> 398.1974120960518
y ---> 381.7907160728694
z ---> 105.09587970643595
clarity_0 ---> 350.97252651675234
clarity_1 ---> 1717.3495640187066
clarity_2 ---> 4018.391384027688
clarity_3 ---> 4406.376992928564
clarity_4 ---> 145.30622863460954
```

Snippet 1. 6 Checking Multicollinearity using VIF

Check Multi-collinearity using VIF

We can observe very strong multi collinearity present in the data set when ideally it should be within 1 to 5.

1.5 Inference: Basis on these predictions, what are the business insights and recommendations. Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

- *We can see that the from the linear plot, very strong corelation between the predicted y and actual y. But there are lots of spread. That indicates some kind noise present on the data set that is Unexplained variances on the output.*

Linear regression Performance Metrics:

- *Intercept for the model: 1750.3835*
- *R square on training data: 0.93041*

PROJECT PM

- *R square on testing data: 0.92875*
- *RMSE on Training data: 917.56672*
- *RMSE on Testing data: 921.70133*

- *From the above matrix, as the training data & testing data score are almost inline, we can conclude this model is a Right-Fit Model.*

- *Finally, 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.*

- *When 'carat' increases by 1 unit, diamond price increases by 8782.2477 units, keeping all other predictors constant.*

- *When 'cut' increases by 1 unit, diamond price increases by 201.1362 units, keeping all other predictors constant.*

- *When 'colour' increases by 1 unit, diamond price decreases by -387.6295 units, keeping all other predictors constant.*

- *When 'y' increases by 1 unit, diamond price increases by 1381.3066 units, keeping all other predictors constant.*

- *There are also some negative co-efficient values, we can see the 'X' i.e., Length of the cubic zirconia in mm. having negative co-efficient -1655. And the p value is less than 0.05, so can conclude that as higher the length of the stone is a lower profitable stone.*

- *Similarly, for the 'z' variable having negative co-efficient i.e., -180.0037 and the p value is less than 0.05, so we can conclude that as higher the 'z' of the stone is a lower profitable stone.*

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.

- *As we can see from the model Higher the width('y') of the stone is higher the price.*

- *So, the stones having higher width('y') should consider in higher profitable stones.*

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

- *The Diamonds clarity with 'VS1' &'VS2' are the most expensive. So, these two categories also consider in higher profitable stones.*

- *As we see for 'X' i.e., Length. of the stone, higher the length of the stone is lower the price.*

- *So higher the Length('x') of the stone are lower is the profitable higher the 'z' i.e. Height of the stone is, lower the price. This is because if a Diamond's Height is too large Diamond will become 'Dark' in appearance because it will no longer return an Attractive amount of light.*

As expected, Carat is a strong predictor of the overall price of the stone. Clarity refers to the absence of the Inclusions and Blemishes and has emerged as a strong predictor of price as well. Clarity of stone types IF, VVS_1, VVS_2 and vs1 are helping the firm put an expensive price cap on the stones.

Color of the stones such H, I and J won't be helping the firm put an expensive price cap on such stones. The company should instead focus on stones of color D, E and F to command relative higher price points and support sales. This also can indicate that company should be looking to come up with new color stones like clear stones or a different color/unique color that helps impact the price positively.

The company should focus on the stone's carat and clarity so as to increase their prices. Ideal customers will also contribute to more profits. The marketing efforts

can make use of educating customers about the importance of a better carat score and importance of clarity index. Post this, the company can make segments, and target the customer based on their income/paying capacity etc, which can be further studied.

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
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

Table 2. 1 Variables in the dataset

PROJECT PM

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.

EDA

- *Data Summary and Exploratory Data Analysis:*
- *Checking if the data is being imported properly*

	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. 2 Head of the dataset

- *Head: The top 5 rows of the dataset are viewed using head () function*

```
<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
```

Table 2. 3 Data Information

PROJECT PM

- *Dimension of the Dataset: The Dimension or shape of the dataset can be shown using shape function. It shows that the dataset given to us has 872 rows and 8 columns or variables.*

- *Structure of the Dataset: Structure of the dataset can be computed using .info() function.*

- *Data Types are of Integer and Object*

- *We shall drop the first column 'Unnamed: 0' column as this is not important for our study. The shape would be – 872 rows and 7 columns and lets get into the description of the data*

	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	872.0	NaN	NaN	NaN	47729.172018	23418.668531	1322.0	35324.0	41903.5	53469.5	236961.0
age	872.0	NaN	NaN	NaN	39.955275	10.551675	20.0	32.0	39.0	48.0	62.0
educ	872.0	NaN	NaN	NaN	9.307339	3.036259	1.0	8.0	9.0	12.0	21.0
no_young_children	872.0	NaN	NaN	NaN	0.311927	0.61287	0.0	0.0	0.0	0.0	3.0
no_older_children	872.0	NaN	NaN	NaN	0.982798	1.086786	0.0	0.0	1.0	2.0	6.0
foreign	872	2	no	656	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 2. 4 Description of the data

- From description of the data set we would get to know the overall summary of the data set.

- Holiday Package – This variable is a categorical Variable. output with the This will be our Target Variable.

- Salary, age, educ, no_young_children, no_older_children, variables are numerical or continuous variables.

```
Holliday_Package      0
Salary                0
age                   0
educ                  0
no_young_children     0
no_older_children     0
foreign               0
dtype: int64

Number of duplicate rows = 0
```

Snippet 2. 1 Null & Duplicate Values

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

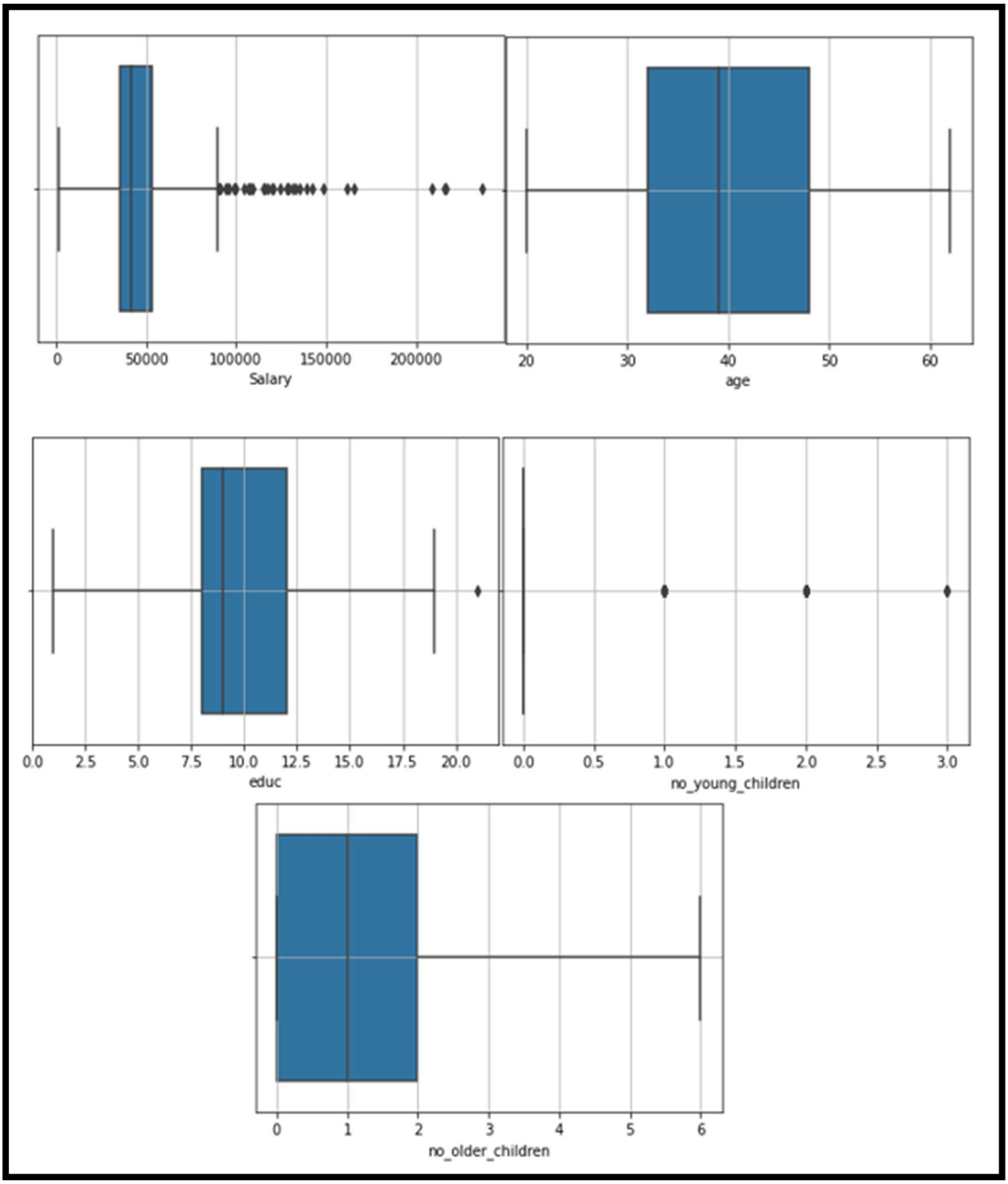


Figure 2. 1 Outliers through Boxplot

PROJECT PM

- *We can observer that there are significant outliers present in variable Salary, however there are minimal outliers in other variables like educ, no. of young children & no. of older children.*
- *There are no outliers in variable age.*
- *Treatment of outliers by IQR method.*

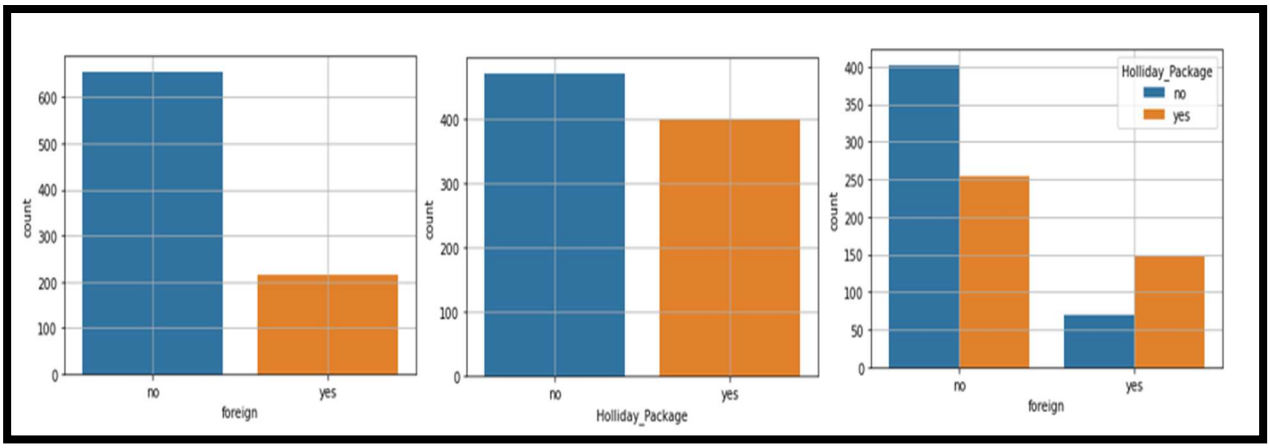


Figure 2. 2 Categorical Variable

- *Foreign is a categorical variable*
- *We can observe that most of the employees are not opting for the holiday package and comparatively less are interested in the package.*
- *This implies we have a dataset which is fairly balanced.*
- *We can observe that higher number of the employees are not Foreigners and pretty less are foreigners.*
- *Holliday_Package: The distribution seems to be fine, with 54% for no and 46% for yes.*
- *Foreign: The data is imbalanced with more skewed towards no and relatively a smaller shared for yes.*

- *Both the variables can be encoded into numerical values for model creation analytical purposes.*

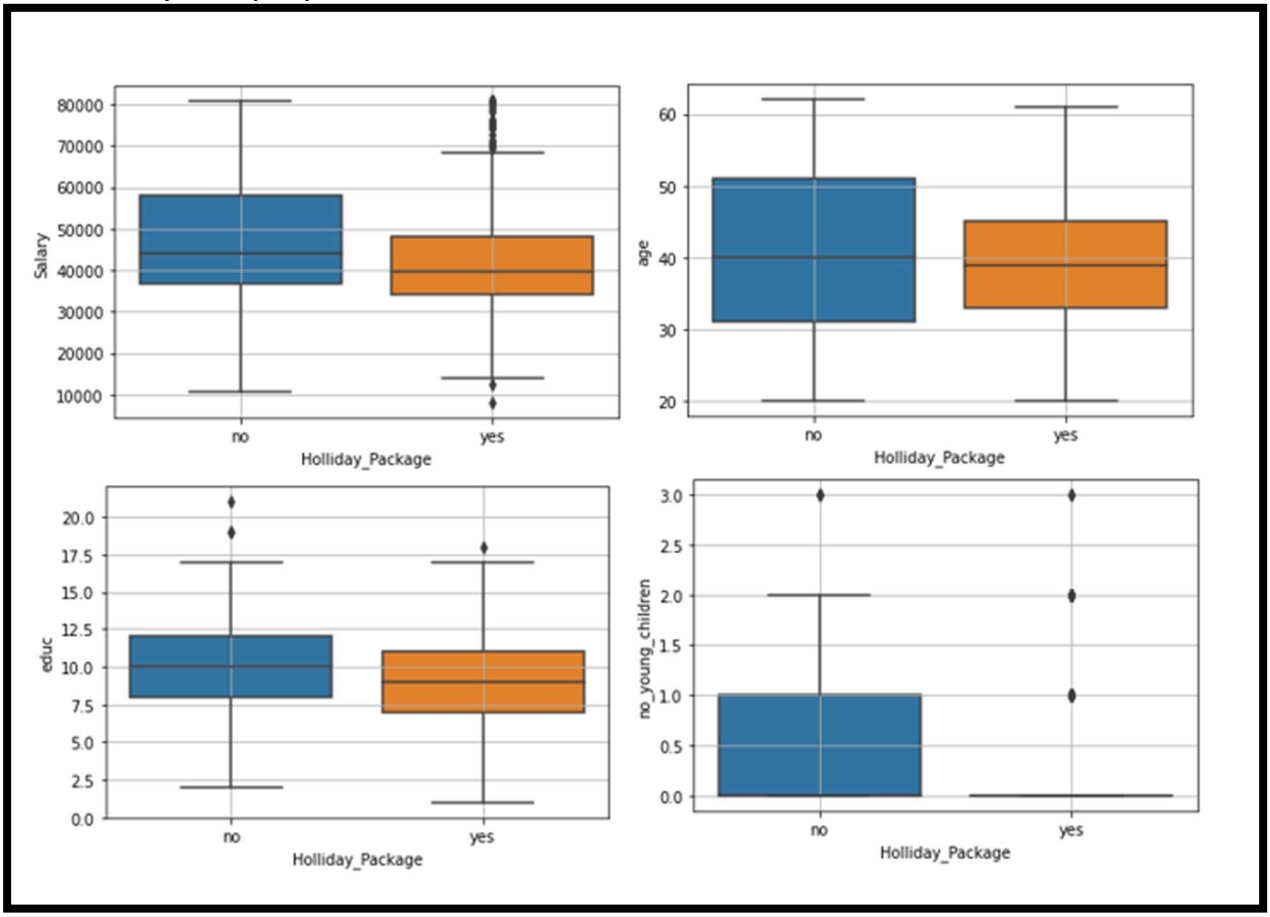


Figure 2. 3 Bivariate analysis

- *Salary ranges from 1322 to 236961. Average salary of employees is around 47729 with a standard deviation of 23418.*

- *Standard deviation indicates that the data is not normally distributed.*

- *Skew of 0.71 indicates that the data is right skewed and there are few employees earning more than an average of 47729. 75% of the employees are earning below 53469 while 255 of the employees are earning 35324.*

- *Age of the employee ranges from 20 to 62.*

PROJECT PM

- *Median is around 39. 25% of the employees are below 32 and 25% of the employees are above 48. Standard deviation is around 10.*

- *Standard deviation indicates almost normal distribution.*

- *Years of formal education ranges from 1 to 21 years. 25% of the population has formal education for 8 years, while the median is around 9 years. 75% of the employees have formal education of 12 years. Standard deviation of the education is around 3. This variable is also indicating skewness in the data.*

- *While performing the bivariate analysis we observe that Salary for employees opting for holiday package and for not opting for holiday package is similar in nature. However, the distribution is fairly spread out for people not opting for holiday packages.*

- *Holiday Package & educ This variable is also showing a similar pattern. This means education is likely not to be a variable for influencing holiday packages for employees.*

- *We observe that employees with less years of formal education(1 to 7 years) and higher education are not opting for the Holiday package as compared to employees with formal education of 8 year to 12 years.*

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign_yes
Holliday_Package	1.000000	-0.180214	-0.092311	-0.102552	-0.173115	0.080286	0.254096
Salary	-0.180214	1.000000	0.047029	0.352726	-0.034360	0.121993	-0.239387
age	-0.092311	0.047029	1.000000	-0.149294	-0.519093	-0.116205	-0.107148
educ	-0.102552	0.352726	-0.149294	1.000000	0.098350	-0.036321	-0.419678
no_young_children	-0.173115	-0.034360	-0.519093	0.098350	1.000000	-0.238428	0.085111
no_older_children	0.080286	0.121993	-0.116205	-0.036321	-0.238428	1.000000	0.021317
foreign_yes	0.254096	-0.239387	-0.107148	-0.419678	0.085111	0.021317	1.000000

Table 2. 5 Correlation between the variables

PROJECT PM

- *Checked for data Correlation.*

- *We will see correlation between independent variables to see which factors might influence choice of holiday package.*

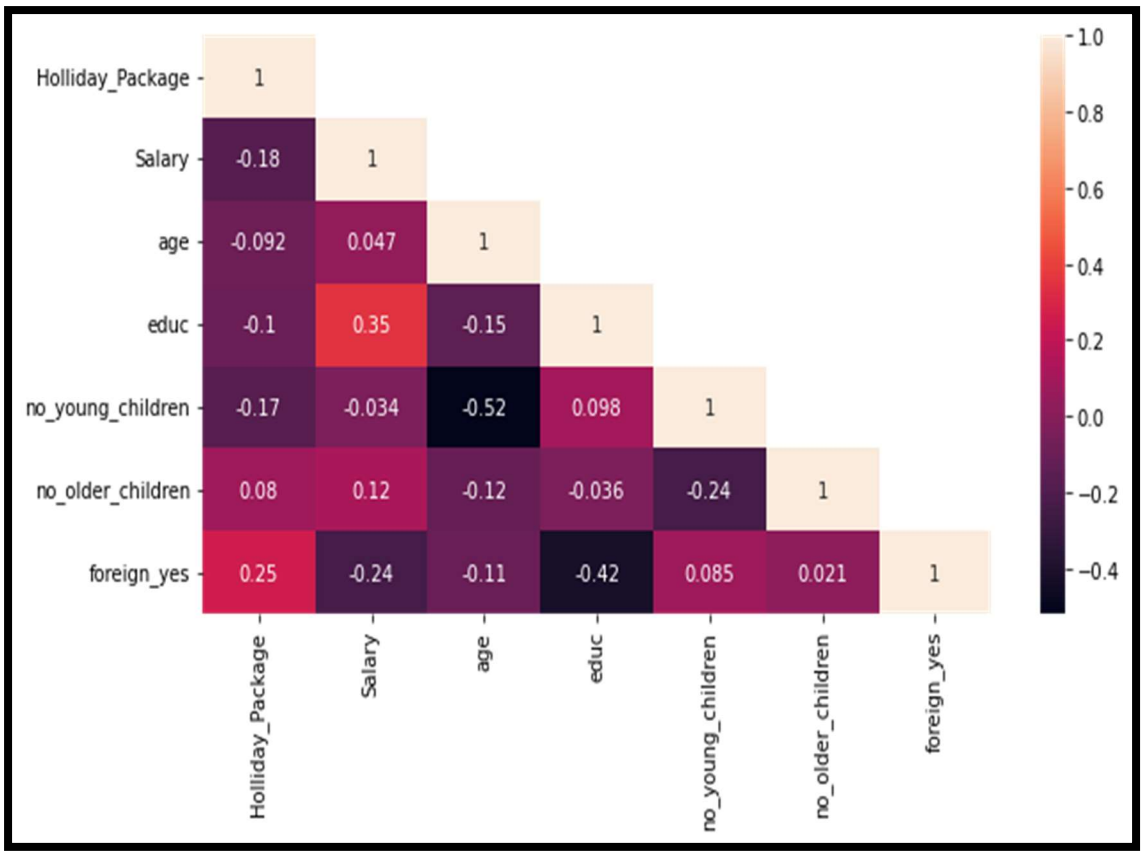


Figure 2. 4 Heat Map

- *We can relate there isn't any strong correlation between any variables. Salary and education display moderate corelation and no_older_children is somewhat correlated with salary variable. However, there are no strong correlation in the data set.*

- *Before we proceed with the model creation, let us read the other part of the data to see how the numerical data also impacts the model. I will first look at the data correlation to quickly identify the variable importance using the heatmap.*

PROJECT PM

- *We can relate there isn't any strong correlation between any variables. Salary and education display moderate corelation and no_older_children is somewhat correlated with salary variable. However, there are no strong correlation in the data set.*

PROJECT PM

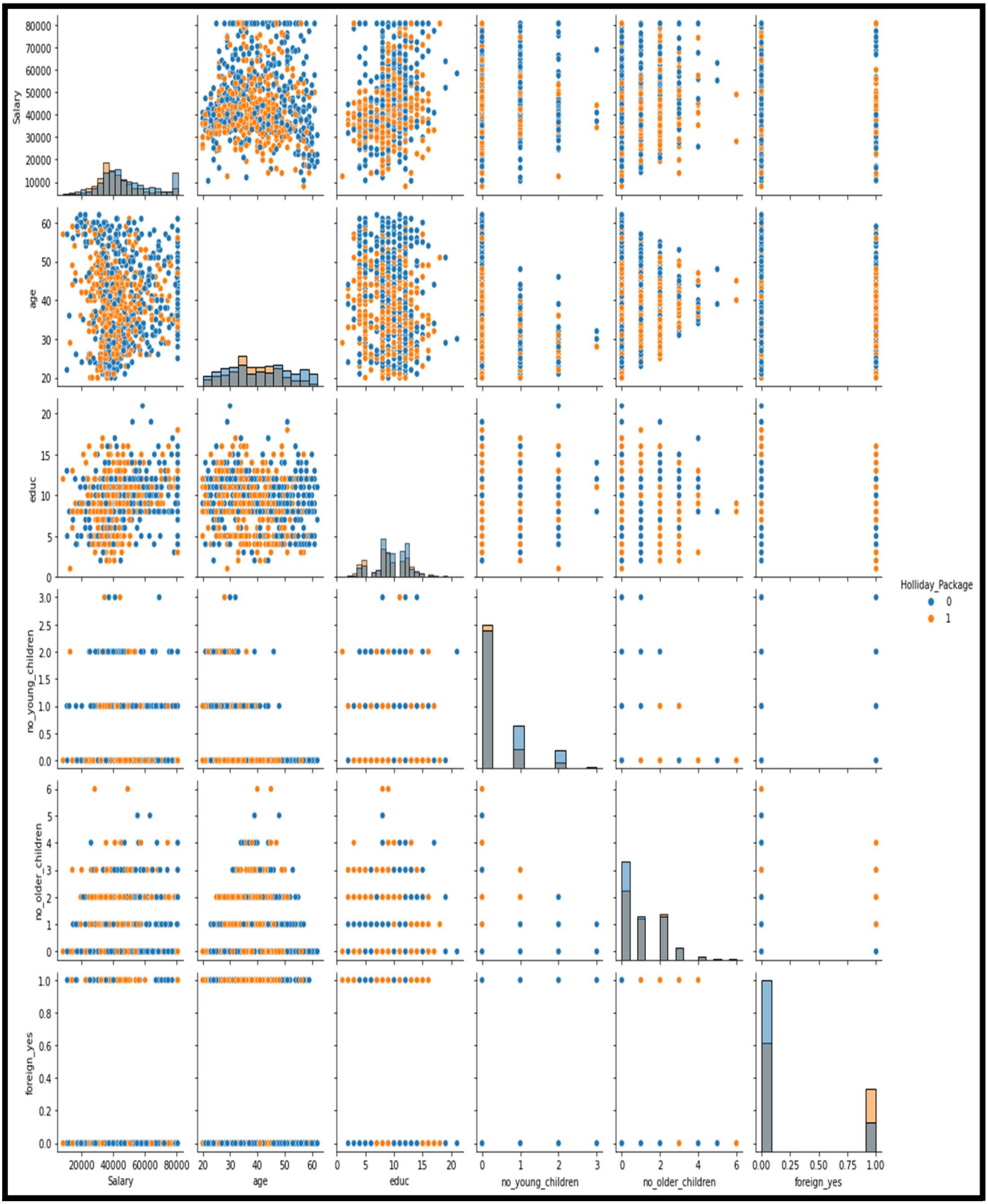

Salary
age
educ
no_young_children
no_older_children
foreign_yes
Holliday_Package
0
1

Figure 2. 5 Pair plot

- *Checking pairwise distribution of the continuous variables: [Salary, age,educ, no. of young children, 'no of older children]*

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).

- *While Linear Regression helps us predicting continuous target variable, Logistic Regression helps us for predicting a discrete a target variable. Logistic Regression is one of the "white-box" algorithms which helps us in determining the probability values and the corresponding cut-offs. Logistic regression is used to solve such problem which gives us the corresponding probability outputs and then we can decide the appropriate cut-off points to get the target class outputs.*

- *Precisely 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).*

```
df1['Holliday_Package'].value_counts()

0    471
1    401
Name: Holliday_Package, dtype: int64
```

- ***Snippet 2. 2 Value Counts of Holiday Variable***

PROJECT PM

- *Evaluation of Logistic regression model- Performance measurement of classification algorithms is judge by confusion matrix which comprise the classification count values of actual and predicted labels.*

  Pros and cons of Logistics Regression:

- *Pros- Logistic regression classification model is simple and easily scalable for multiple classes.*
- *Cons- Classifier constructs linear boundaries and the interpretation of coefficients value is difficult.*

- *Split X and y into training and test set in 70:30 ratio. This implies 70% of the total data will be used for training purposes and remaining 30% will be used for test purposes*

- *Split X and y into training and test set in 70:30 ratio. This implies 70% of the total data will be used for training purposes and remaining 30% will be used for test purposes.*

```
<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    int8
 1   Salary             872 non-null    float64
 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: float64(1), int64(4), int8(1), object(1)
memory usage: 41.9+ KB
```

Table 2. 6 Data Information after converting all variables into numerics

	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

Table 2. 7 Head of the data after encoding

- *Here we have done ONE HOT ENCODING to create dummy variables and we can see all values for foreign_yes are 0.*

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

X-HEAD

```
0    0
1    1
2    0
3    0
4    0
Name: Holliday_Package, dtype: int8
```

Y-HEAD

Table 2. 8 Splitting the data

- *The data proportion seems to be reasonable and we can continue with our model building as next steps.*

- *As next steps, we will initiate the LogisticRegression function and will then fit the Logistic Regression model. There after we will predict on the training and test data set.*

PROJECT PM

	Salary	age	educ	no_young_children	no_older_children	foreign_yes
821	38974.0	47	12	0	2	1
805	40270.0	33	8	2	0	1
322	32573.0	30	11	1	0	0
701	43839.0	43	11	0	1	1
773	33060.0	40	5	1	1	1
...	...	...	...	...	...	...
594	42369.0	47	9	0	1	0
297	44207.0	45	12	0	2	0
76	50291.0	34	10	0	2	0
831	33434.0	44	7	0	1	1
187	36832.0	40	8	0	2	0

610 rows × 6 columns

Table 2. 9 Prediction on Independent variable train data

	Salary	age	educ	no_young_children	no_older_children	foreign_yes
264	25118.0	58	8	0	0	0
189	40913.0	20	9	1	0	0
643	28446.0	58	8	0	0	0
65	36072.0	35	4	0	2	0
241	52736.0	40	10	0	3	0
...	...	...	...	...	...	...
165	34878.0	29	14	1	1	0
100	61159.0	38	10	0	3	0
503	41167.0	44	9	0	2	0
431	41769.0	43	9	0	0	0
119	46856.0	44	9	0	3	0

262 rows × 6 columns

Table 2. 10 Prediction on Independent variable test data

PROJECT PM

```
array([1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0,
       1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1,
       1, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1,
       0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0,
       0, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0,
       1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0,
       1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0,
       1, 0, 0, 1, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0,
       1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1,
       1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0,
       0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1,
       1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0,
       0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0,
       0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1,
       0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1,
       0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 0,
       1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0,
       1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1,
       0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1,
       0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1,
       0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1,
       0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0,
       0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0,
       0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1,
       0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1,
       1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 0,
       0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 0], dtype=int8)
```

Snippet 2. 3 Dependent Variable Prediction on training set

```
array([0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0,
       0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0,
       0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1,
       0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1,
       0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1,
       0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0,
       0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0,
       0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
       1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0,
       1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0,
       0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0,
       1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
      dtype=int8)
```

Snippet 2. 4 Dependent Variable Prediction on testing set

	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

Snippet 2. 5 Head of the dependent variable probability predication

```
Number of rows and columns of the training set for the independent variables: (610, 6)
Number of rows and columns of the training set for the dependent variable: (610,)
Number of rows and columns of the test set for the independent variables: (262, 6)
Number of rows and columns of the test set for the dependent variable: (262,)
```

Snippet 2. 6 Dimension of the split

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.

Performance Metrics For Logistic Regression

```
# Accuracy - Training Data
model.score(X_train, Y_train)
```

```
0.6622950819672131
```

Snippet 2. 7 Accuracy on training Data

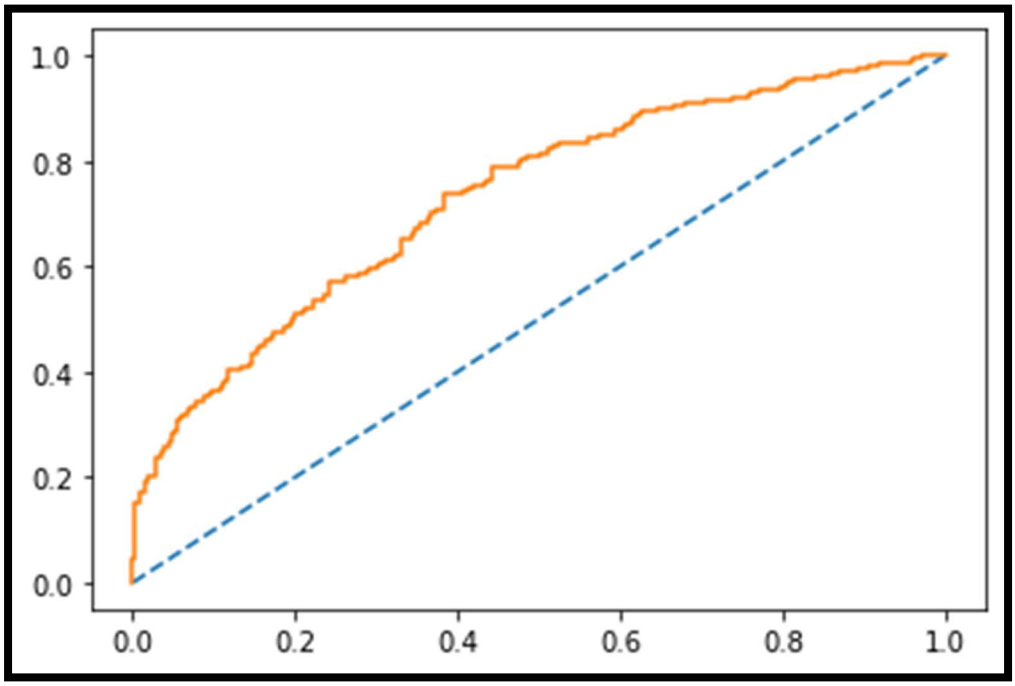


Figure 2. 6 AUC and ROC for the training data

- *ROC Curve- Receiver Operating Characteristic(ROC) measures the performance of models by evaluating the trade-offs between sensitivity (true positive rate) and false (1- specificity) or false positive rate.*

- *AUC - The area under curve (AUC) is another measure for classification models is based on ROC. It is the measure of accuracy judged by the area under the curve for ROC.*

- *Performance Matrix of Logistics Regression model:*

```
# Accuracy - Test Data
model.score(X_test, Y_test)
```

```
0.648854961832061
```

Snippet 2. 8 Accuracy on test data

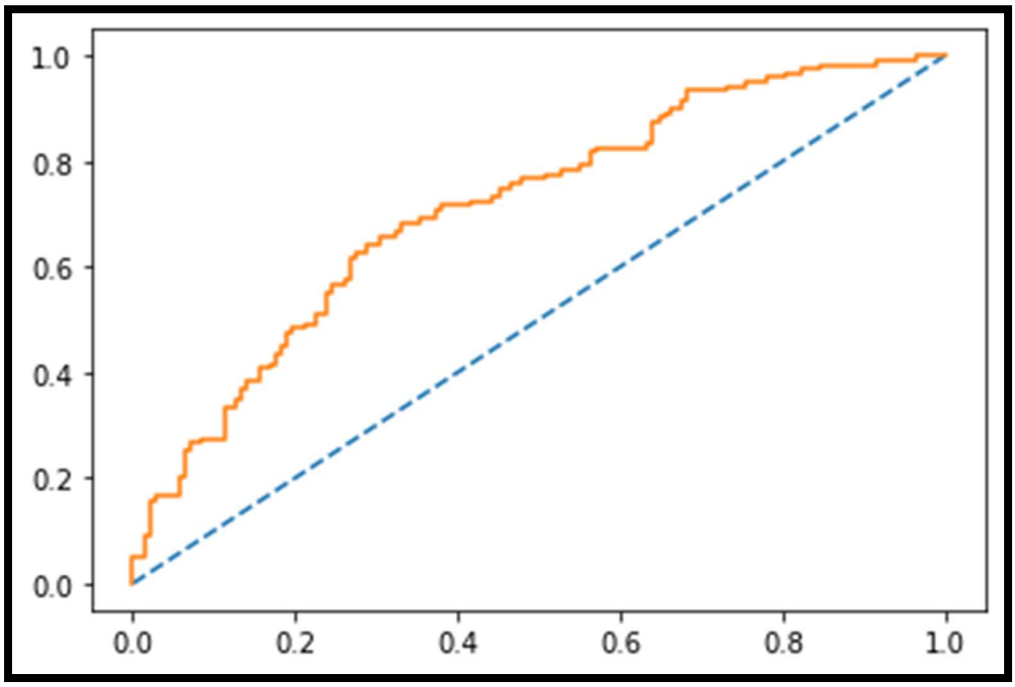
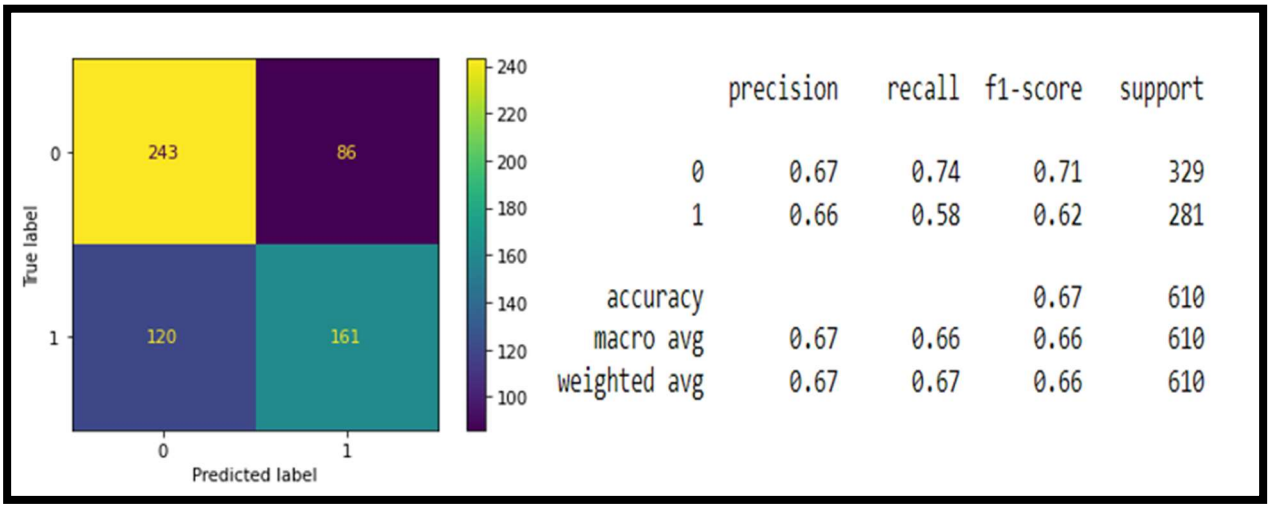


Figure 2. 7 AUC and ROC for the test data

- *Accuracy on – Test Data AUC: 0.731*

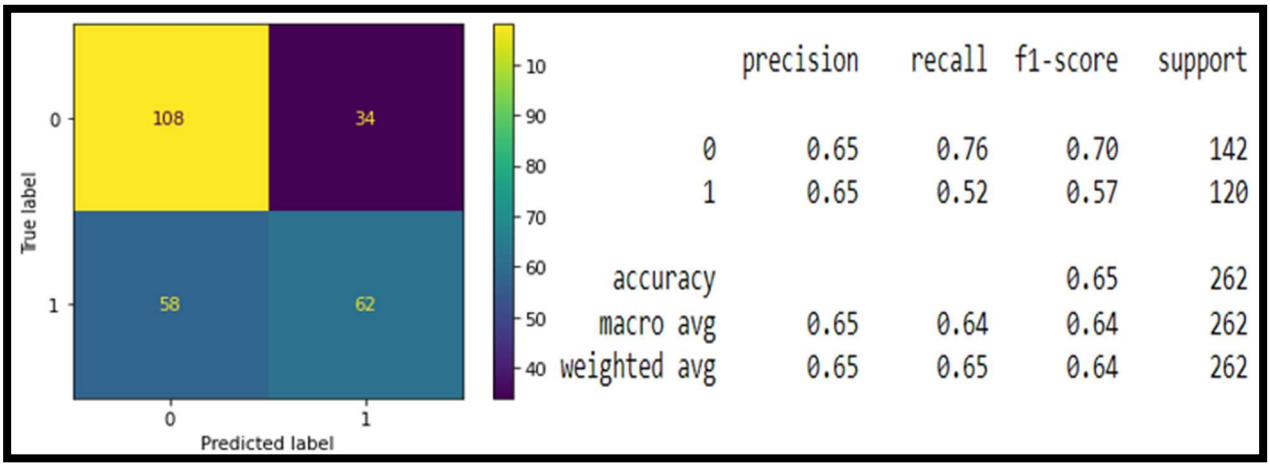


	precision	recall	f1-score	support
0	0.67	0.74	0.71	329
1	0.66	0.58	0.62	281
accuracy			0.67	610
macro avg	0.67	0.66	0.66	610
weighted avg	0.67	0.67	0.66	610

Figure 2. 8 Confusion Matrix and Classification Report on training data

- *Confusion matrix on the training data*
- *Here we see that precision for 1 is 0.66, recall is 0.58 accuracy is 0.67 and f1 score is 0.62.*

Confusion matrix on the test data



	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 2. 9 Confusion Matrix and Classification Report on testing data

- *Confusion matrix on the training data*
- *Here we see that precision for 1 is 0.65, recall is 0.52 accuracy is 0.65 and f1 score is 0.57.*

PROJECT PM

- *While looking the metrices for both training and the test data, it seems the accuracy scores are same on both models at approximately 66%. Our model is close enough to be treated as a right fit model. The current model is not struggling with being an over fit model or an under fit model. The AUC scores for both the training and test data is also same at 73%.*

- *The model performance is good on F1 score as well with training data performing better at 62% while the test data gave a F1 score of 57%.*

- *To summarize, our model scores are not very good but seems to be a right fit model and seems to be avoiding the underfit and overfit scenarios. The training data seems to be performing a bit better when compared with test data though the difference is not much. The point to note is that F1 score seems to perform better on Training data and dips slightly in the test data.*

Performance Metrics For LDA

PROJECT PM

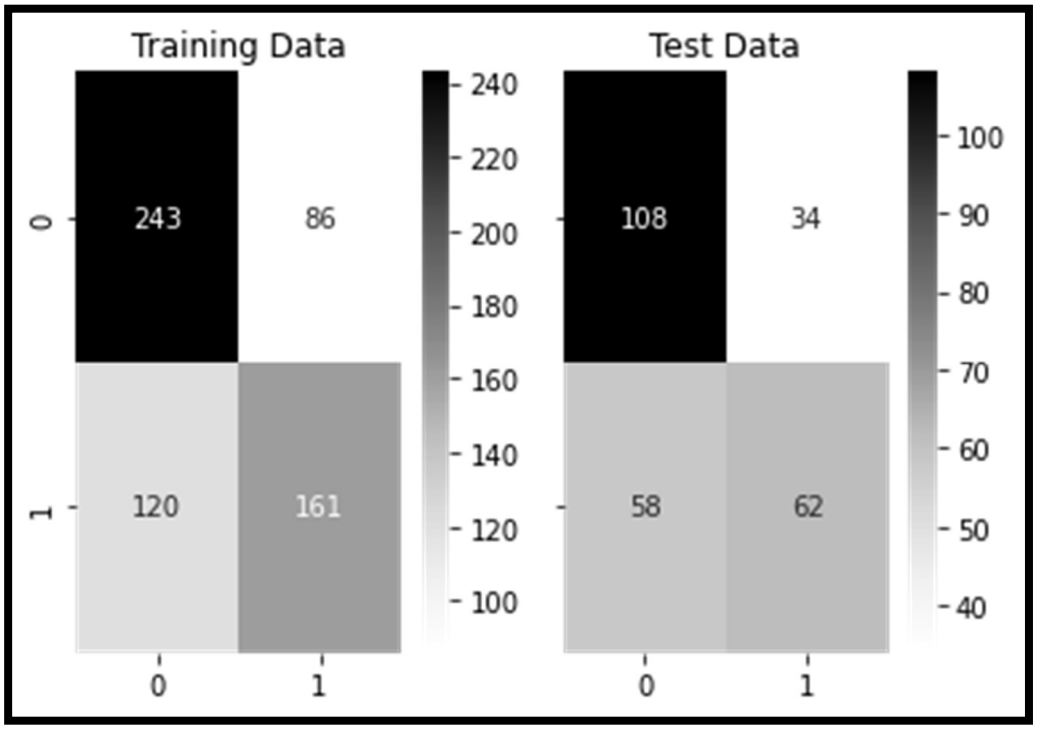


Figure 2. 10 Model Evaluation

- *The accuracy score of the training data and test data is at 66% and 65% respectively.*

- *This is almost similar to the Logistic Regression model result so far.*

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.67	0.74	0.70	329
1	0.65	0.57	0.61	281
accuracy			0.66	610
macro avg	0.66	0.66	0.66	610
weighted avg	0.66	0.66	0.66	610

Classification Report of the test data:

	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

Table 2. 11 Clasification Report on Training and Testing Data

- *The AUC scores are marginally lower for the test data, else they are also almost similar to the Logistic Regression model.*

- *F1 scores are 61% and 57% for train and test data, respectively, which again is almost close to the logistic regression model.*

- *From the below image:*
- *AUC for the Training Data: 0.740 or 74.0%*
- *AUC for the Test Data: 0.725 or 72.5%*

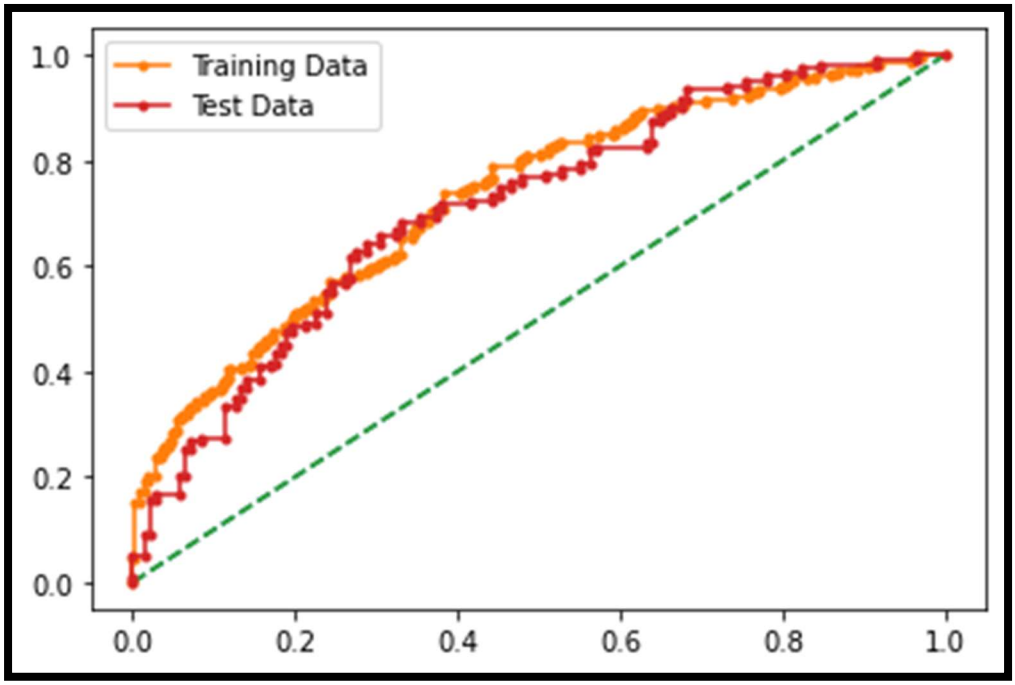


Figure 2. 11 AUC on Training and Testing Data

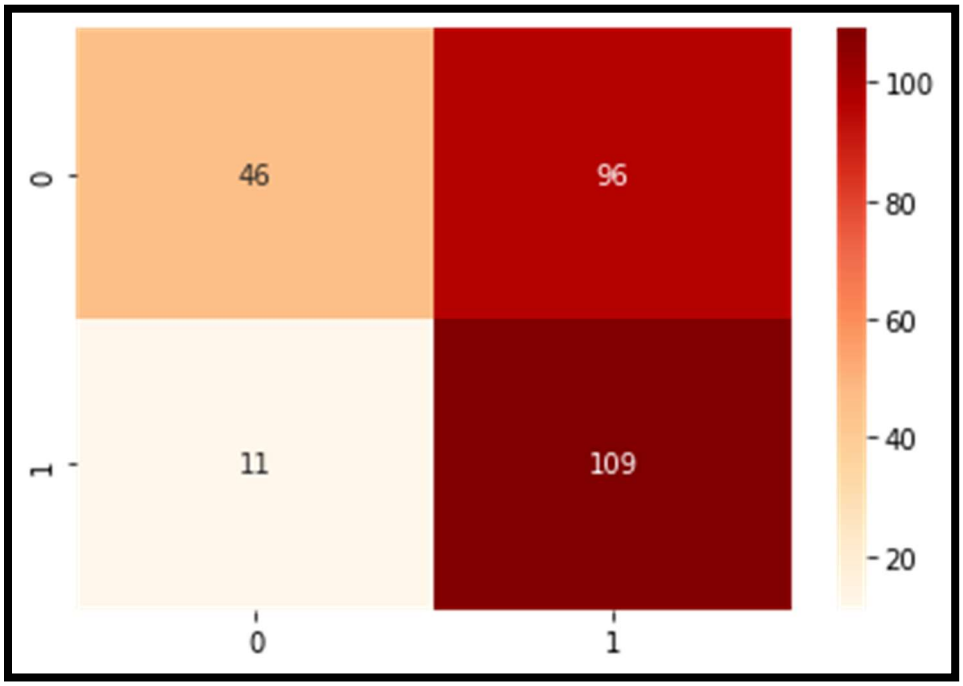


Figure 2. 12 Dependent variable prediction

- *Custom cut off for the LDA model:*

- *Comparison of the Classification report:*

- *Looking at the custom classification report for the test data, we can see that we have managed to retain our accuracy scores and have been able to improve our F1 score from 57% to now a 67%.*

Classification Report of the default cut-off test data:

	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

Classification Report of the Holidaypackage cut-off test data:

	precision	recall	f1-score	support
0	0.81	0.32	0.46	142
1	0.53	0.91	0.67	120
accuracy			0.59	262
macro avg	0.67	0.62	0.57	262
weighted avg	0.68	0.59	0.56	262

Table 2. 12 Cut – off Report

- *As stated above, both the models – Logistics and LDA offers almost similar results. While LDA offers flexibility to control or change the important*

metrices such as precision, recall and F1 score by changing the custom cut-off.

- *Like in this case study, the moment we changed the cut off to 40%, we were able to improve our precision, recall and F1 scores considerably. Further, this is up to the business if they would allow the play with the custom cut off values or no.*

- *Though for this case study, I have chosen to proceed with Logistics Regression as its is easier to implement, interpret, and very efficient to train. Also, our dependent variable is following a binary classification of classes, and hence it is ideal for us to rely on the logistic regression model to study the test case at hand.*

- *Logistic regression is a classification algorithm used to find the probability of event success and event failure. It is used when the dependent variable is binary(0/1, True/False, Yes/No) in nature. It learns a linear relationship from the given dataset and then introduces a non-linearity in the form of the Sigmoid function.*

2.4 Inference: Basis on these predictions, what are the insights and recommendations. Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

- *So, we had been given a problem where we had to find out whether the employees will opt for a holiday package or not. We looked in the data using logistic regression and LDA.*

- *We found out that the results using both the methods is same. Predictions were done using both the models.*

PROJECT PM

As interpretation

- *While doing EDA we found out that most of the employees who are above 50 don't opt for holiday packages. It seems like they are not interested in holiday packages at all.*

- *Employees who are in the age gap of 30 to 50 opt for holiday packages. It seems like young people believe in spending on holiday packages so age here plays a very important role in deciding whether they will opt for package or not*

- *Also, people who have salary less than 50000 opt for holiday packages. So, salary is also a deciding factor for the holiday package.*

- *Education also plays an important role in deciding the holiday packages.*

- *To improve our customer base, we need to look into those factors.*

- *There is no plausible effect of salary, age, and education on the prediction for Holliday_packages. These variables don't seem to impact the decision to opt for holiday packages as we couldn't establish a strong relation of these variables with the target variable.*

- *Foreign has emerged as a strong predictor with a positive coefficient value. The log likelihood or likelihood of a foreigner opting for a holiday package is high.*

- *no_young_children variable is negating the probability for opting for holiday packages, especially for couple with number of young children at The company can try to bin salary ranges to see if they can derive some more meaningful interpretations out of that variable. May be club the salary or age in different buckets and see if there is some plausible impact on the predictor variable. OR else, the business can use some different model techniques to do a deep dive.*

PROJECT PM

Recommendation:

- *The company should really focus on foreigners to drive the sales of their holiday packages as that's where the majority of conversions are going to come in.*

- *The company can try to direct their marketing efforts or offers toward foreigners for a better conversion opting for holiday packages.*

- *The company should also stay away from targeting parents with younger children. The chances of selling to parents with 2 younger children is probably the lowest. This also gels with the fact that parents try and avoid visiting with younger children.*

- *If the firm wants to target parents with older children, that still might end up giving favourable return for their marketing efforts then spent on couples with younger children.*

- *As we already have the customer base who are of the age of 30 to 50 so we need to look for the options and target the older people and the people who are earning more than 150000. As we know most of the people who are older prefer to visit religious places so it would be better if we target those places and provide them with packages where they can visit religious places.*

- *We can also look into the family dynamics of the people of the older people if the older people have elder children e.g 30 to 40 they can use the holiday packages so the deal should include the family package.*

- *People who earn more than 150000 don't spend much on the holiday packages, they tend to go for lavish holidays and we can provide them with customized packages according to their wish, such as fancy hotels, longer vacations, personal cars during the holiday to attract such employees.*

- *Plus, such people who earn more than 150000 we can provide them extra facilities according to their own wishes at the moment.*

PROJECT PM

- *In this project we started with EDA, descriptive statistics and did null value condition check, we performed Univariate and Bivariate Analysis. did exploratory data analysis, we treated outliers then we moved on to Logistic regression. We encoded the data (having string values) for Modelling. We split data into train and test (70:30) and finally we applied Logistic Regression and LDA (linear discriminant analysis).*