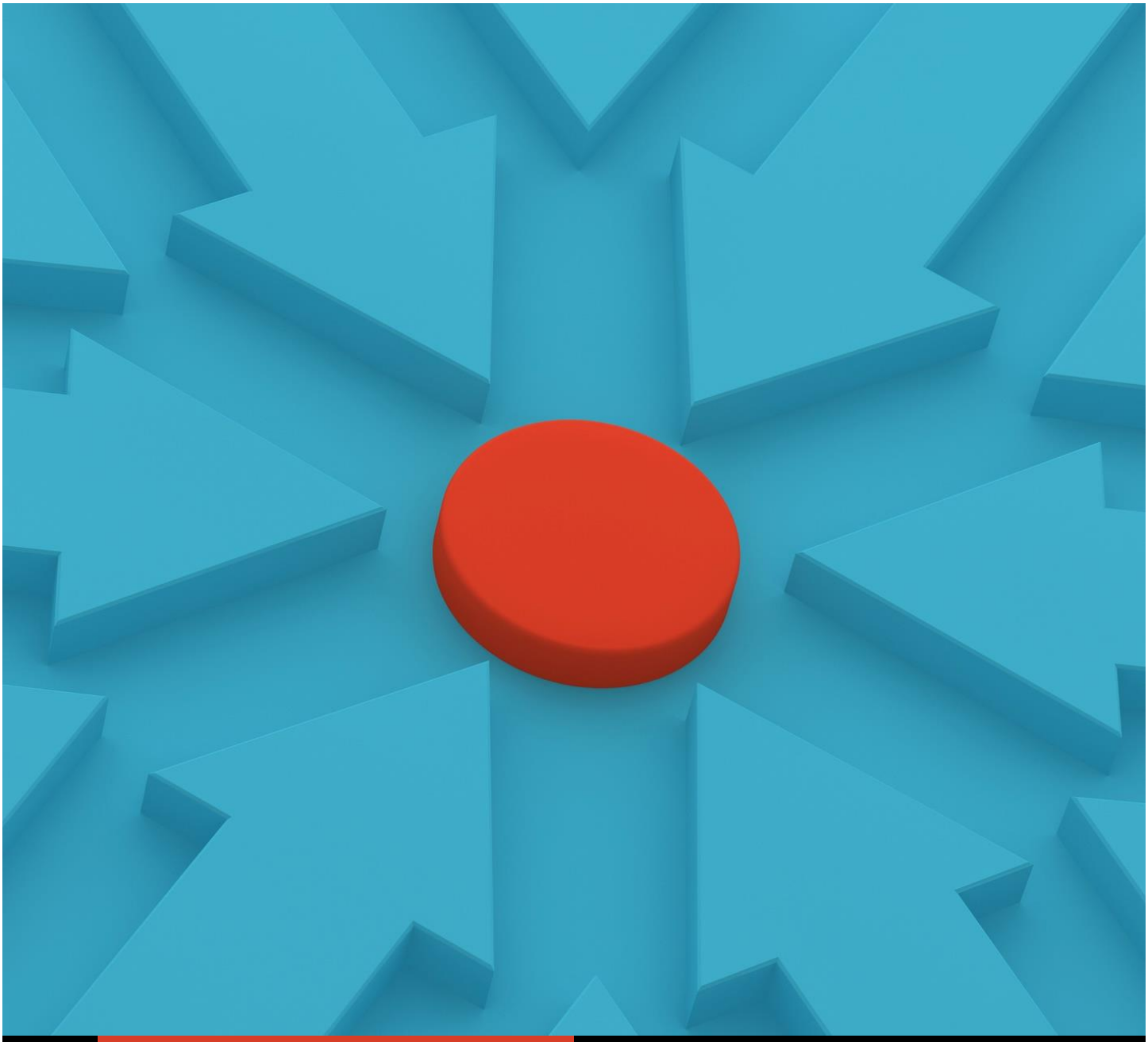


PREDICTIVE MODELLING

PROJECT REPORT

SUNDAR RAM S
PGPDSBA
ONLINE DEC_C 2021
05-JUN-2022

Contents

TABLE OF FIGURES

TABLE OF TABLES

TABLE OF EQUATIONS

PROBLEM 1 – GEM STONES COMPANY

Problem Statement

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

Sample of the Dataset

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

Table 1 Sample of the Dataset

The data consists of 10 columns having the price and other characteristics of cubic zirconia for 26967 samples.

Dataset 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.

EXPLORATORY DATA ANALYSIS

Variable type

Carat	: float64
Cut	: object
Color	: object
Clarity	: object
Depth	: float64
Table	: float64
Price	: int64
X	: float64
Y	: float64
Z	: float64

Variable type is the type of data each column is holding. Here we are able to notice that Price column is of type int64, color, clarity & cut are of type object and rest all are of the type float64.

Check for duplicated records in the dataset

On exploring the dataset, it is found that there are 34 duplicate records. We need to ensure the dataset free from duplicated records. Hence, the duplicated records are removed and updated dataset contains 26933 records.

	carat	cut	color	clarity	depth	table	x	y	z	price
4756	0.35	Premium	J	VS1	62.4	58.0	5.67	5.64	3.53	949
6215	0.71	Good	F	SI2	64.1	60.0	0.00	0.00	0.00	2130
8144	0.33	Ideal	G	VS1	62.1	55.0	4.46	4.43	2.76	854
8919	1.52	Good	E	I1	57.3	58.0	7.53	7.42	4.28	3105
9818	0.35	Ideal	F	VS2	61.4	54.0	4.58	4.54	2.80	906
10473	0.79	Ideal	G	SI1	62.3	57.0	5.90	5.85	3.66	2898
10500	1.00	Premium	F	VVS2	60.6	54.0	6.56	6.52	3.96	8924
12894	1.21	Premium	D	SI2	62.5	57.0	6.79	6.71	4.22	6505
13547	0.43	Ideal	G	VS1	61.9	55.0	4.84	4.86	3.00	943
13783	0.79	Ideal	G	SI1	62.3	57.0	5.90	5.85	3.66	2898
14389	0.60	Premium	D	SI2	62.0	57.0	5.43	5.35	3.34	1196
14410	1.00	Very Good	D	SI1	63.1	56.0	6.34	6.30	3.99	5645
15798	0.90	Very Good	I	VS2	58.4	62.0	6.29	6.35	3.69	3334
16852	0.79	Ideal	G	SI1	62.3	57.0	5.90	5.85	3.66	2898
17263	1.04	Premium	I	SI2	62.0	57.0	6.53	6.47	4.03	3774

Table 2 Sample Duplicated records in the Dataset

Check for missing values in the dataset

Carat : 26933 non - null

Cut : 26933 non - null

Color : 26933 non - null

Clarity : 26933 non - null

Depth : 26236 non - null

Table : 26933 non - null

Price : 26933 non - null

X : 26933 non - null

Y : 26933 non - null

Z : 26933 non - null

It can be seen that there are 697 missing values in the depth variable.

Summary of the dataset

Summarizing briefly, the dataset has a total of 26933 records. Mean carat is 0.8, mean depth is 61.75, mean table is 57.46, mean x value is 5.73, mean y value is 5.73, mean z value is 3.54 & the mean price is 3937.53.

From the below table, we can see that the minimum value is 0 for x, y & z values, which has no meaning to it. This might probably be some missing values, which are entered as 0. These are bad data and need to be treated before moving on with the predictive modelling. We can also see that the Mean and median are almost having the same value, which may be a sign for normal distribution.

	carat	depth	table	x	y	z	price
count	26933.00	26236.00	26933.00	26933.00	26933.00	26933.00	26933.00
mean	0.80	61.75	57.46	5.73	5.73	3.54	3937.53
std	0.48	1.41	2.23	1.13	1.17	0.72	4022.55
min	0.20	50.80	49.00	0.00	0.00	0.00	326.00
25%	0.40	61.00	56.00	4.71	4.71	2.90	945.00
50%	0.70	61.80	57.00	5.69	5.70	3.52	2375.00
75%	1.05	62.50	59.00	6.55	6.54	4.04	5356.00
max	4.50	73.60	79.00	10.23	58.90	31.80	18818.00

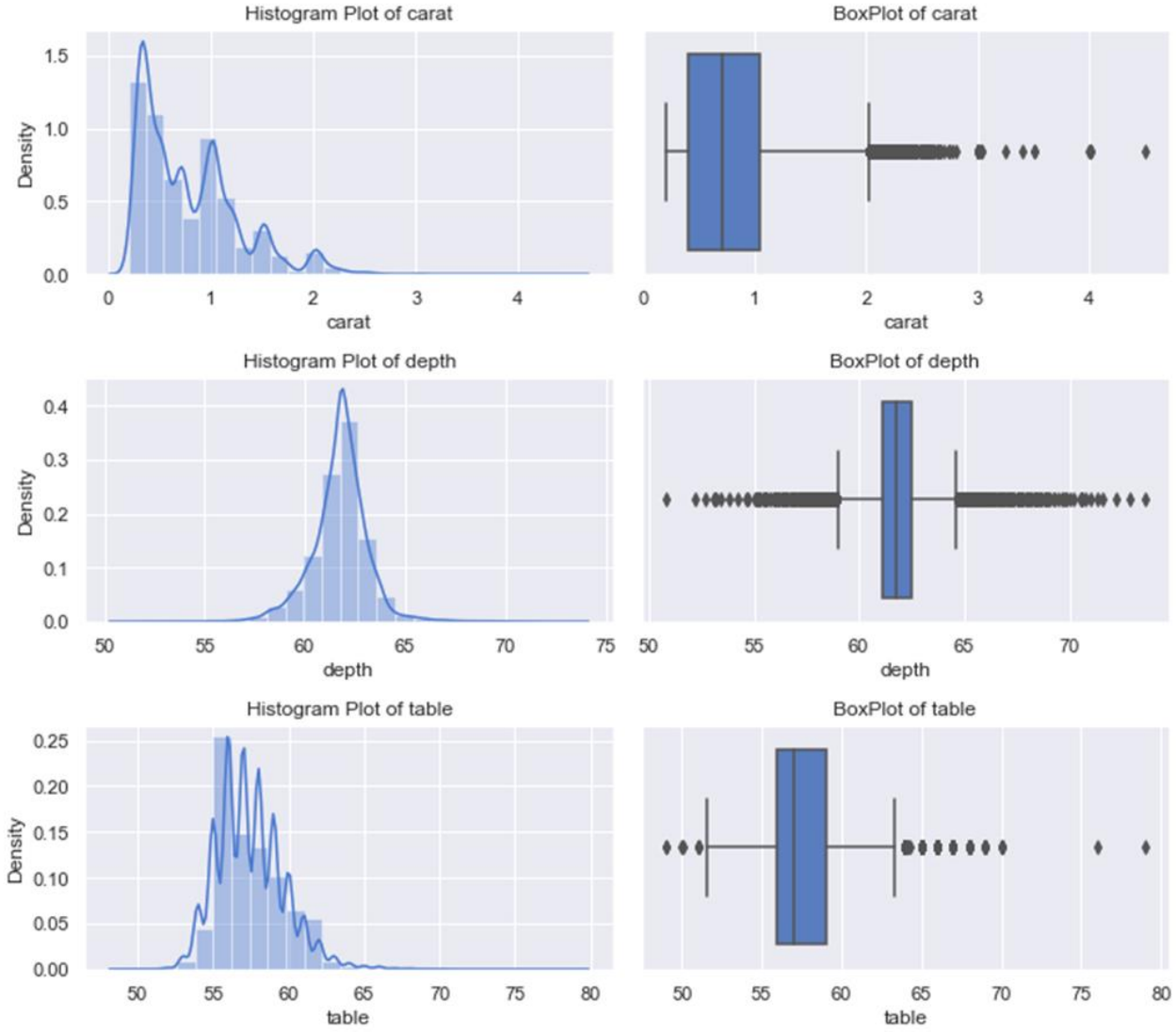
Table 3 Statistical summary of the dataset

DATA VISUALIZATION

Univariate Analysis – Histograms, Box Plots & Count Plots

From the below histogram having the Kernel Density function, we are able to understand the distribution of data in the variables. For all the continuous variables, except depth & price, there are multiple nodes and hence multimodal distribution These indicate presence of certain patterns in the dataset. However, we will not be exploring the patterns. For Depth, it is unimodal, while for the price, it is bi modal. Also, we can see that the distributions are too peaked which indicates the kurtosis value for the distributions might be higher than normal. Except depth all the distributions are skewed towards the left, while the depth is not skewed.

From the box plot, it evident that there are outliers in all the variables. It is important to treat these outliers before proceeding with modelling, as the outliers affect prediction.



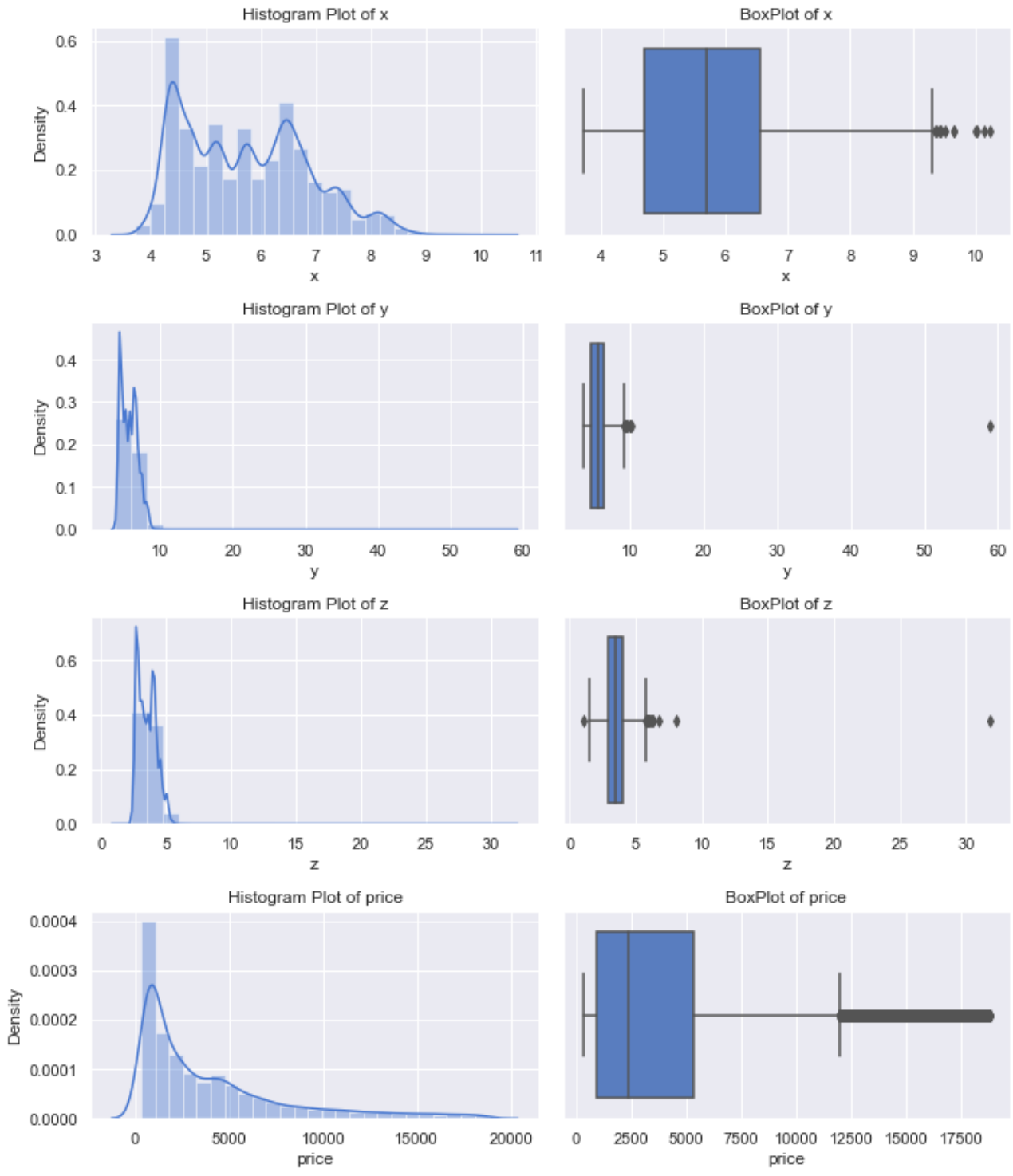
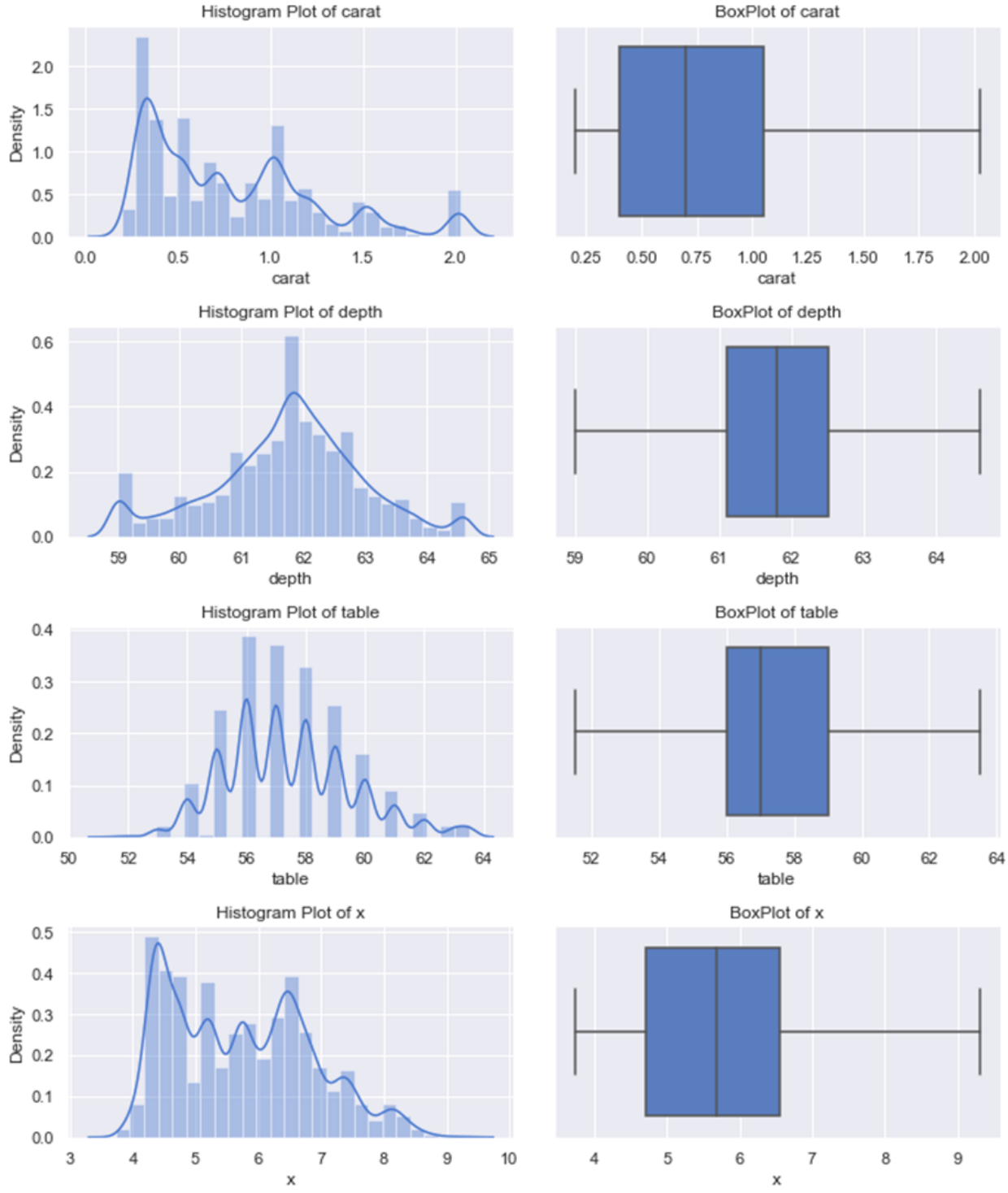


Figure 1 Histogram & Box Plots of continuous variables

Outlier treatment is done by replacing the outlier values with the values of upper & lower whiskers, depending up on the value of the outlier. It is replaced with lower whisker, if the value is less than the lower whisker & is replaced with the upper whisker if the value is higher than the upper whisker. Lower whisker value is [Q1 - 1.5(Inter Quartile Range)] & Upper whisker value is [Q3 + 1.5(Inter Quartile Range)].

Post removal of outliers, we can see that the distribution of depth & price has become tri-modal. This is because the values of whiskers have become more. Also from the box plot we are able to understand that all the outliers are removed.



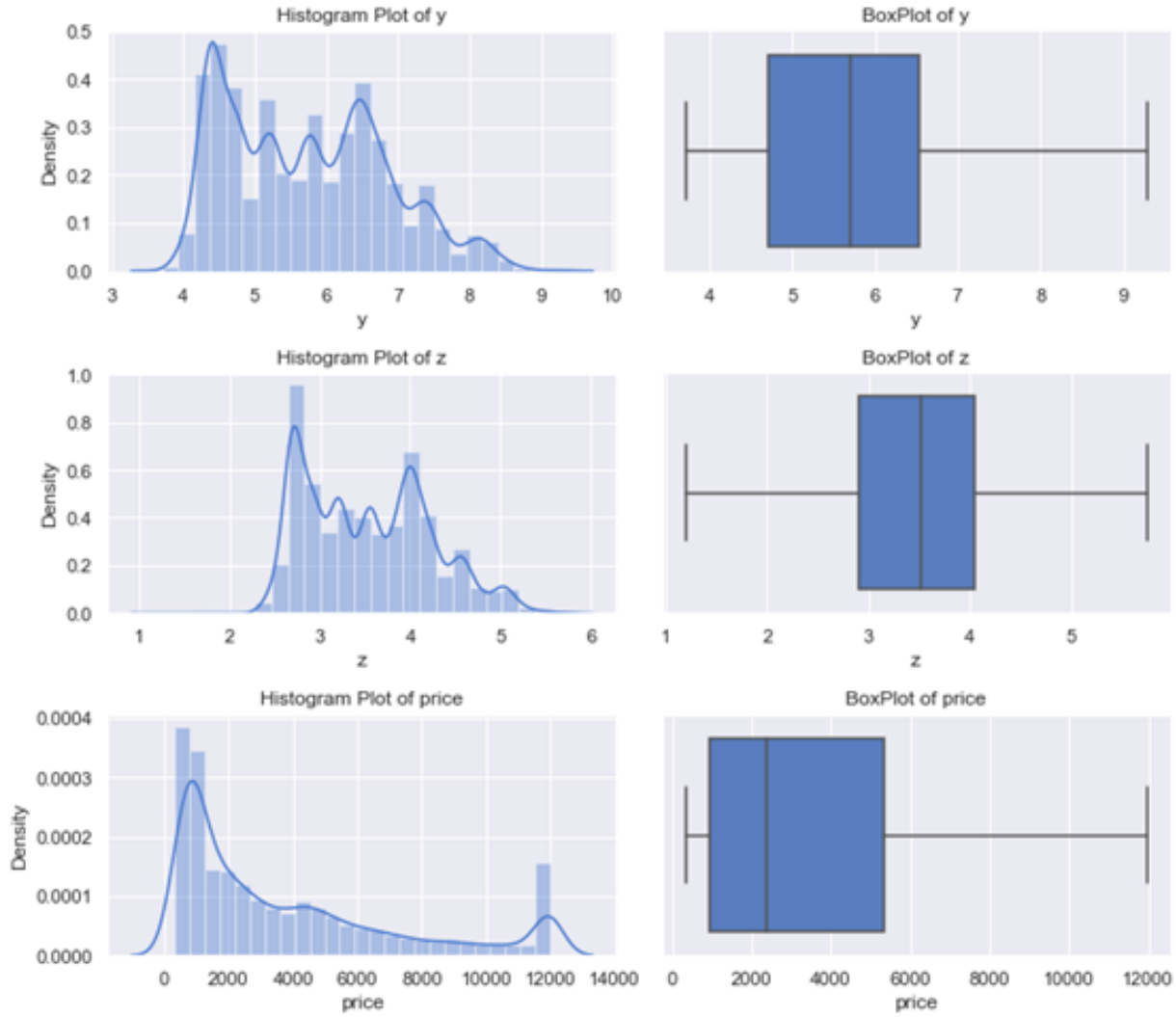


Figure 2 Histogram & Box plots post outlier treatment

Univariate visualization of Categorical variables is carried out using the count plots.

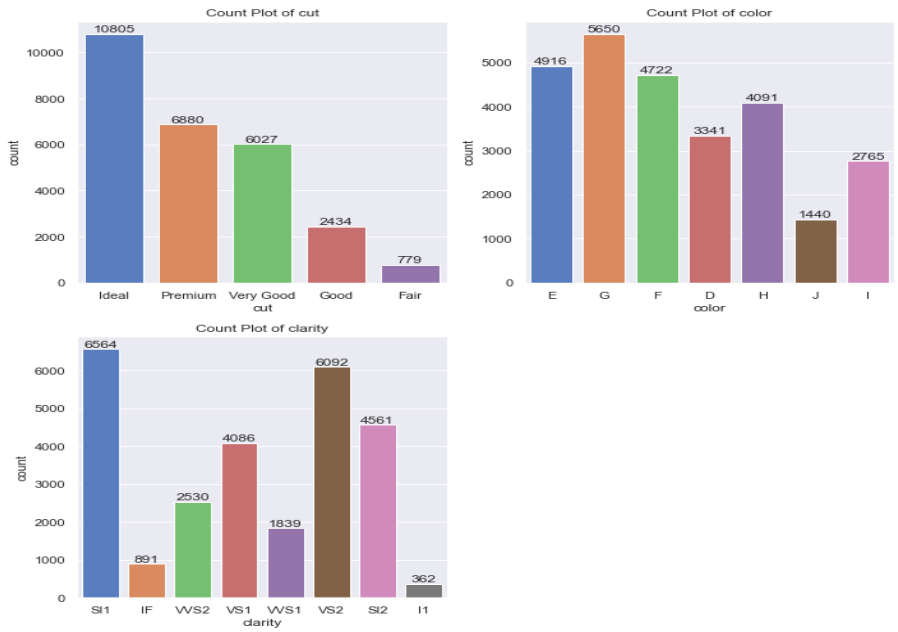


Figure 3 Count plot of categorical variables

Bivariate Analysis - Correlation Plot & Pair Plot

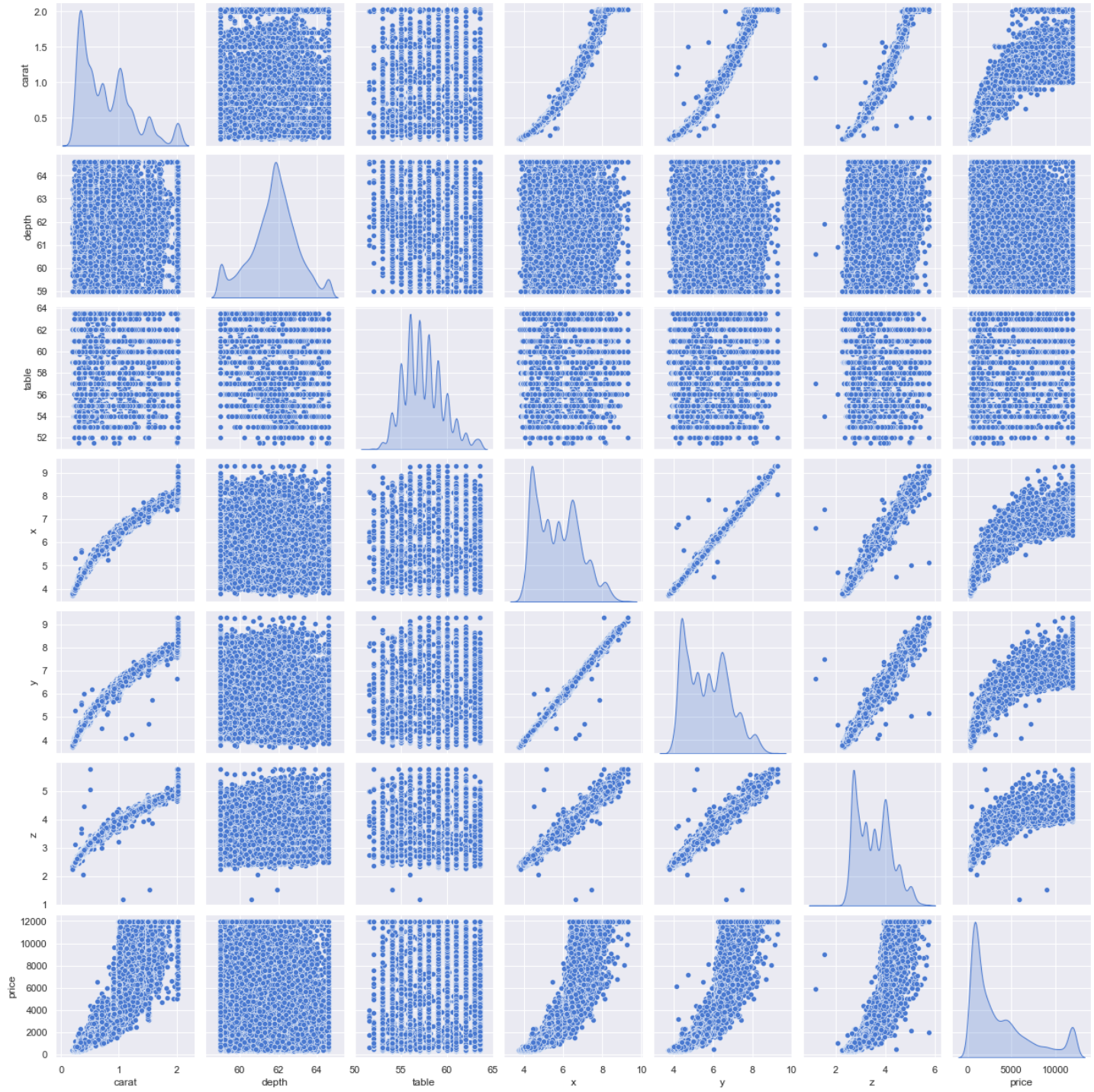


Figure 4 Pair Plot of continuous variables

From the above pair plot it can be seen that carat, x, y & z have a linear relationship with the price variable. As the variables carat, x, y, z increases, prices also increase.

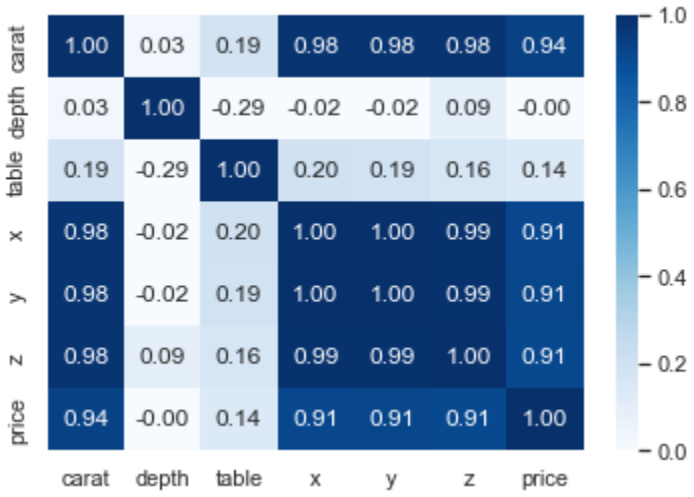


Figure 5 Correlation plot of continuous variables

The figure 5 shows the correlation plot (heat map) of the continuous variables present in the dataset. The figure indicates there is perfect correlation between x & y variables. Also, good correlation amongst carat, x, y, z & price. This indicates that carat and size (dimensions) of the cubic zirconia play a major role in deciding the price of the stone.

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.

IMPUTING NULL VALUES & CLEANING BAD DATA

We saw that there are 697 null values in the depth variables. Null values affect prediction. Hence, it would be the best to impute null values with the median value. From the summary of the dataset, we know that the median value is 61.80. Hence, we can impute the null values with 61.80.

We also saw that there are some meaning less data (0) in the x, y & z variables. These need to be cleaned. On exploring, it is found that the records indexed 5821,6034,10827,12498,12689,17506,18194,23758 have 0 in either x, y or z. As there are only 8 records with 0, we can drop them.

	carat	cut	color	clarity	depth	table	x	y	z	price
5821	0.71	Good	F	SI2	64.1	60.0	0.00	0.00	0.0	2130
6034	2.02	Premium	H	VS2	62.7	53.0	8.02	7.95	0.0	18207
10827	2.20	Premium	H	SI1	61.2	59.0	8.42	8.37	0.0	17265
12498	2.18	Premium	H	SI2	59.4	61.0	8.49	8.45	0.0	12631
12689	1.10	Premium	G	SI2	63.0	59.0	6.50	6.47	0.0	3696
17506	1.14	Fair	G	VS1	57.5	67.0	0.00	0.00	0.0	6381
18194	1.01	Premium	H	I1	58.1	59.0	6.66	6.60	0.0	3167
23758	1.12	Premium	G	I1	60.4	59.0	6.71	6.67	0.0	2383

Table 3 Records with 0s in x, y, z

The summary of the cleaned data set is as in the below table. We can see no major change/ difference after cleaning the data. Hence, the dataset remains unaffected.

	carat	depth	table	x	y	z	price
count	26925.00	26925.00	26925.00	26925.00	26925.00	26925.00	26925.00
mean	0.80	61.75	57.46	5.73	5.73	3.54	3936.25
std	0.48	1.39	2.23	1.13	1.16	0.72	4020.98
min	0.20	50.80	49.00	3.73	3.71	1.07	326.00
25%	0.40	61.10	56.00	4.71	4.71	2.90	945.00
50%	0.70	61.80	57.00	5.69	5.70	3.52	2373.00
75%	1.05	62.50	59.00	6.55	6.54	4.04	5353.00
max	4.50	73.60	79.00	10.23	58.90	31.80	18818.00

Table 4 Summary of the dataset post cleaning/ imputing

COMBINING SUBLEVELS OF ORDINAL VARIABLES

Here, there are 3 object type variables, that are ordinal in nature. However, having too many ordinals might affect prediction as the model has to learn huge data variations and the number of records to get the model trained become less. Hence, we can combine the ordinals to select, meaningful divisions, so that the model can be better built to predict.

Here, we are combining only the color and clarity as they have 7 & 8 ordinals respectively. However, there are only 5 ordinals in cut, hence we can leave it as it is.

Clarity & Color are each clustered as Worst, Better and Best. The claritiy level, from its terminology seems to have a pattern. Hence, I1 is categorized as Best, SI2, SI1, VS2, VS1 are clustered as better & VVS2, VVS1, IF are clustered as Worst. In the color variable, I & J have very less frequency, hence they are categorized as best.

Going by ordinal and spread of data amongst the other ordinals D, E & F are clustered as Worst & G&H are clustered as better.

The spread of data amongst the color and clarity can be visualized in the count plot as below.

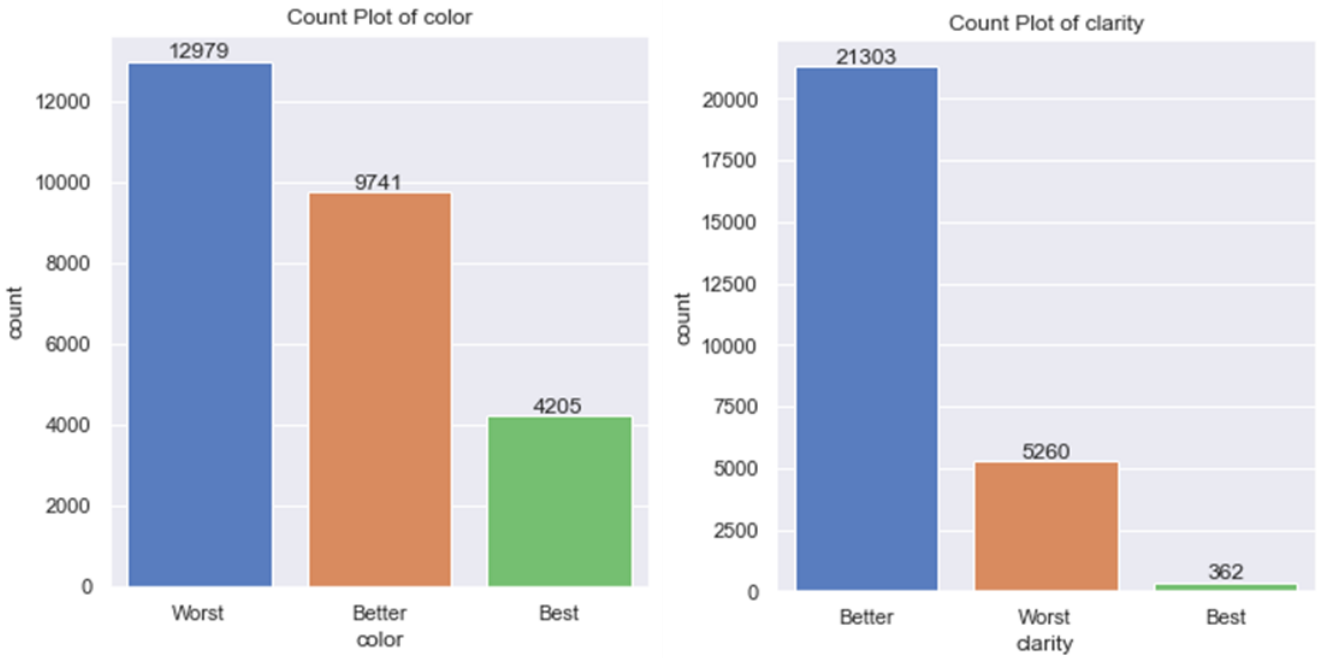


Table 5 Count plot of Color & Clarity post clustering

Clarity		Color	
Data	**Cluster**	**Data**	**Cluster**
I1	Best	D	Worst
SI2	Better	E	Worst
SI1	Better	F	Worst
VS2	Better	G	Better
VS1	Better	H	Better
VVS2	Worst	I	Best
VVS1	Worst	J	Best
IF	Worst		

Table 6 Cluster Reference of Clarity & Color

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.

ENCODING

Linear Regression or any regression in that case deal only with numbers. Hence the object type variables in the dataset need to be converted to numerical type. Here, the cut, clarity & color are object type variables and need to be converted to numerical type.

We can go ahead with either label encoding or one hot encoding or a mix of both. However, we will go with a mix of both where in the cut uses label encoding and for clarity & color we will go with One hot encoding. The reason is cut variable has 5 groups while clarity and color have only 3 groups. One hot encoding increases the number of variables, which is however not going to affect the modelling.

Label encoding details as in the below table & a sample data after performing one hot encoding is also shown in the table.

Cut	Label
Fair	1
Good	2
Very Good	3
Premium	4
Ideal	5

Table 7 Label Encoding of Cut

	carat	cut	depth	table	x	y	z	price	color_Better	color_Worst	clarity_Better	clarity_Worst
0	0.30	5	62.1	58.0	4.27	4.29	2.66	499.0	0	1	1	0
1	0.33	4	60.8	58.0	4.42	4.46	2.70	984.0	1	0	0	1
2	0.90	3	62.2	60.0	6.04	6.12	3.78	6289.0	0	1	0	1
3	0.42	5	61.6	56.0	4.82	4.80	2.96	1082.0	0	1	1	0
4	0.31	5	60.4	59.0	4.35	4.43	2.65	779.0	0	1	0	1

Table 8 Encoded Sample dataset

LINEAR REGRESSION MODEL

- The term "regression" refers to predicting a real number.
- The term "linear" in the name “linear regression” refers to the fact that the method models data with linear combination of the explanatory variables.
- A linear combination is an expression where one or more variables are scaled by a constant factor and added together.
- In the case of simplest linear regression with a single explanatory variable, the linear combination used in linear regression can be expressed as:

Dependent variable value = (weight *independent variable) + constant

Equation 1 Linear Regression model equation format

- It is the straight line in the scatter plot of the variables
- For a linear model to be built, there must be correlation amongst the dependent variables, which we saw in the Correlation & pair plot.
- Based on the correlation a scatter plot can be obtained and there can be infinite straight lines that can fit in the scatter plot as a linear model.
- The best fit line can be found out using the Gradient Descent method.
- Gradient descent methods use partial derivatives on the parameters (slope and intercept) to minimize sum of squared errors
- The line that represents the model may not touch all the points in the scatter plot
- The vertical distance between a point and the line is the error (shown in yeelow in Figure 6) in prediction of the model
- The line which gives least sum of squared errors across all the data points put together is considered as the best fit line.

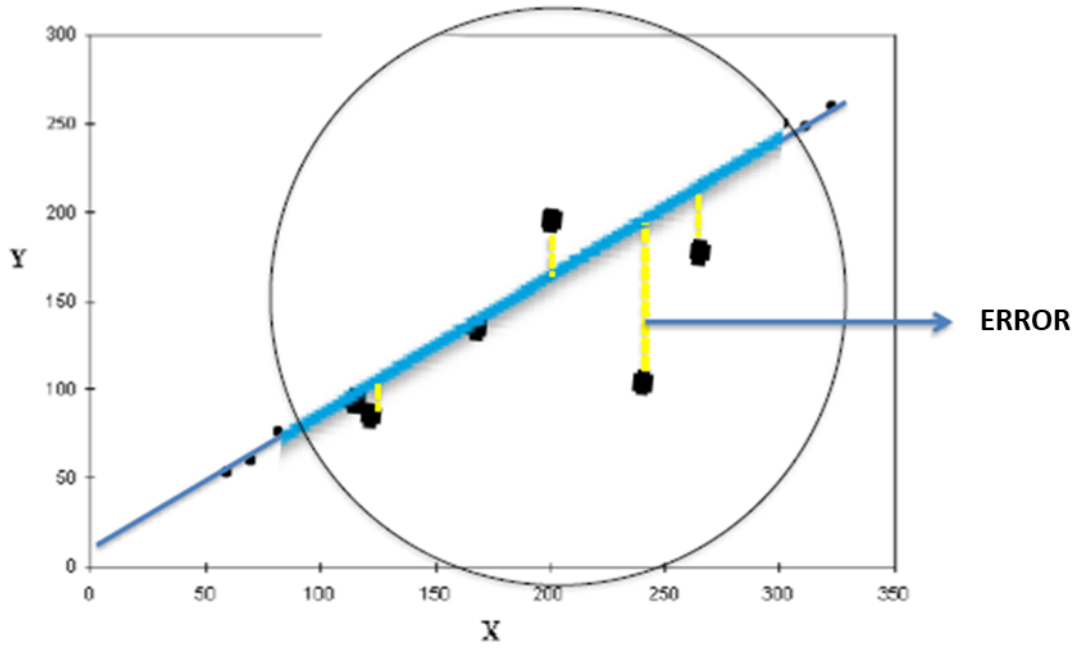


Figure 6 Best fit line & Error Rep

- The best fit line will always go through that point in the features space where the Xbar(blue vertical line) and ybar(blue horizontal line) meet
- Coefficient of determinant –determines the fitness of a linear model. The closer the points get to the line, the R^2 (coefficient of determinant) tends to 1, the better the model is

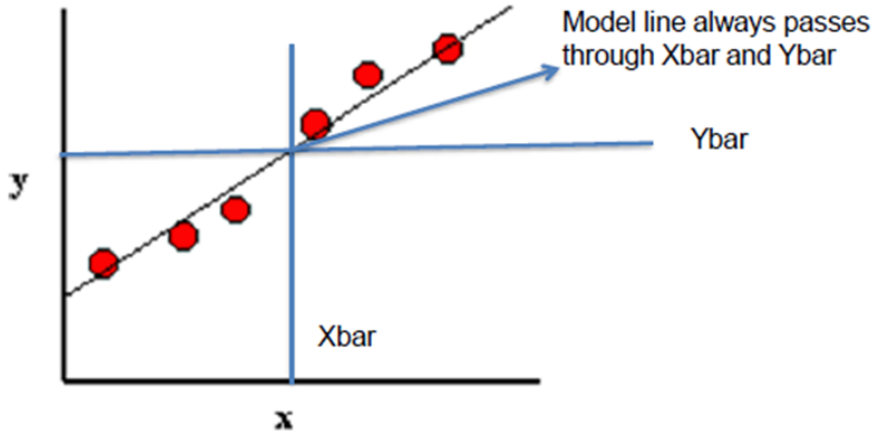


Figure 7 Best fit line passing through Xbar and ybar

ERRORS IN LINEAR REGRESSION MODEL

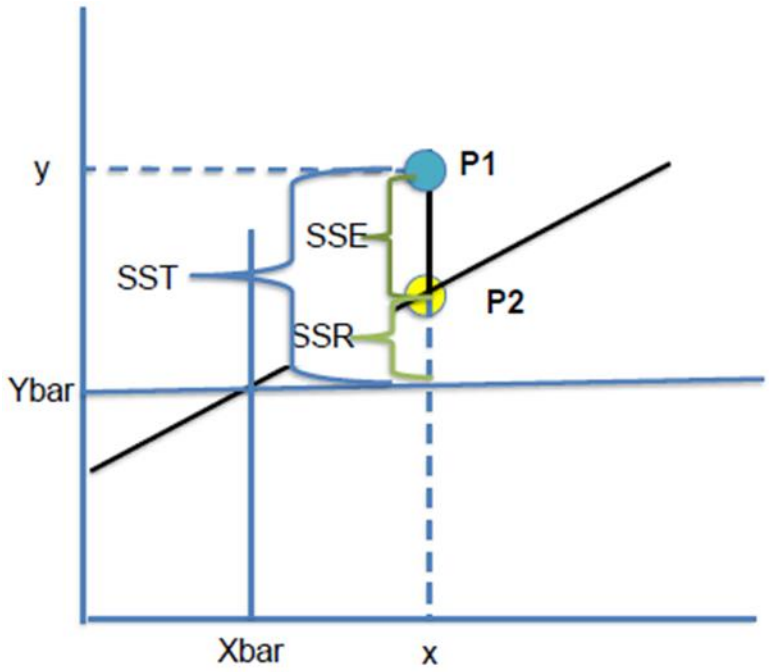


Figure 8 Representation of Errors in Linear Regression model

P1	Original y data point for given x
P2	Estimated y value for given x
Ybar	Average of all Y values in data set
SST	Sum of Square error Total (SST), Variance of P1 from Ybar (Y Ybar)^2
SSR	Regression error (p2 ybar)^2 (portion SST captured by regression model)
SST	captured by regression model)
SSE	Residual error (p1 p2)^2

Table 9 Description of Error Representaiton

- That model is the most fit where every data point lies on the line. i.e. SSE = 0 for all data points
- Hence SSR should be equal to SST i.e. SSR/SST should be 1.
- Poor fit will mean large SSE. SSR/SST will be close to 0
- SSR / SST is called as r^2 (r square) or coefficient of determination
- r^2 is always between 0 and 1 and is a measure of utility of the regression model.

STRUCTURE OF A LINEAR REGRESSION MODEL

$$Y = m_1X_1 + m_2X_2 + \ldots\ldots + m_nX_n + C + e$$

Equation 2 Multi variate Linear Regression Structure

- Y = Dependent / target / predicted variable
- X_i = Independent/ predictor variable
- m_i = coefficients for the independent / predictor variable
- C = constant / intercept / bias
- e = residual error / unexplained variance / difference between actual and prediction

Linear Regression models can be built using two ways

- SKLEARN
- SCIPYSTATS

LINEAR REGRESSION MODEL USING SKLEARN

Here, in the cubic zirconia data set, the data is first split in to predictor variables and the target variable as X & y respectively. Then using the train_test_split library in sklearn, we shall be able to split the data in to 70% training & 30% testing.

Then by using the LinearRegression model library from sklearn.linear_model, we will be able to fit the train data in to our model. By fitting the data, the model gets trained using the training set. The independent attributes have different units and scales of measurements. Hence, It is always the best to scale all the dimensions using z scores or some other method to address the problem of different scales.

We know that the coefficients of the predictor variables (weights), determine how much influence it has in predicting the target variable. The coefficients of different variables are as below.

Feature	Iteration 1 - W/o scaling	Iteration 2 - Scaled Data
Carat Coefficient	9196.61207	1.22295
Cut Coefficient	136.29811	0.04384
Depth Coefficient	-2.16499	-0.00076
Table Coefficient	-22.72956	-0.01417
X Coefficient	-1927.73679	-0.62497
Y Coefficient	1749.71021	0.56339
Z Coefficient	-735.90071	-0.14744
Color_Better Coefficient	851.66890	0.11835
Color_Worst Coefficient	1189.72103	0.17190
Clarity_Better Coefficient	2569.21428	0.30170
Clarity_Worst Coefficient	3586.70699	0.41082
Intercept	-2651.07505	-6.78329
R^2 - Train Data	0.91723	0.91723
R^2 - Test Data	0.91860	0.91858
RMSE - Train Data	994.89659	0.28768
RMSE - Test Data	994.30275	0.28534

Table 10 SKLearn – Coefficients, Intercept, R^2 & RMSE

It can be seen that, there are no major difference due to scaling of dataset. However, the intercept, which is of no meaning in the linear model is found to be reduced. The model is also properly fit that is evident from the R^2 values in train & test data.

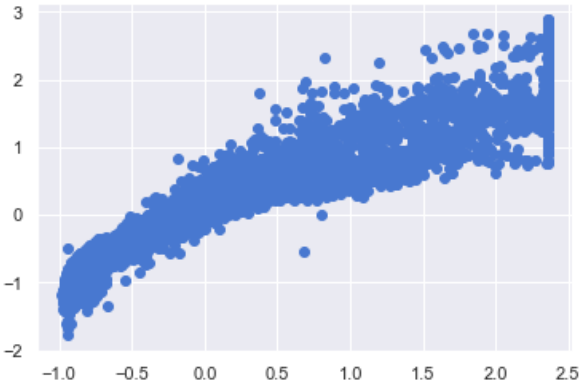


Figure 9 Linear Regression Plot using SK Learn

LINEAR REGRESSION MODEL USING STATS MODEL

Hypothesis Testing

- To establish the reliability of the coefficients, we need hypothesis testing
- The Null hypothesis (H0) claims there is no relation between predictor & predicted variables. That means the coefficient is 0.
- At 95% confidence level, if the p value is < .05, we reject the H0 i.e., probability of finding these coefficients in sample is very low
- If p value is >=.05, we do not have sufficient evidence in the data to reject the H0 and hence we do not reject H0.
- We believe H0 is likely to be true in the universe

Model Building:

R^2 is not a reliable metric as it always increases with addition of more attributes even if the attributes have no influence on the predicted variable. Instead, we use adjusted R^2 which removes the statistical chance that improves R^2. SKlearn does not provide a facility for adjusted R^2... so we use statsmodel, a library that gives results similar. This library expects the X and Y to be given in one single data frame.

The predictor and predicted attributes are again merged as a single data frame & split in to train (70%) and test (30%) data. Then the training Xs & Ys and Testing Xs and Ys are merged separately to meet the statsmodel requirement.

The scipy stats uses the Ordinary Least Squares method to build the model. The formula is built and fit in the linear model. The linear model summary is as below. It can be seen that the intercepts and the coefficients values remain the same.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.917
Model:	OLS	Adj. R-squared:	0.917
Method:	Least Squares	F-statistic:	1.898e+04
Date:	Sun, 05 Jun 2022	Prob (F-statistic):	0.00
Time:	15:19:57	Log-Likelihood:	-1.5684e+05
No. Observations:	18847	AIC:	3.137e+05
Df Residuals:	18835	BIC:	3.138e+05
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2651.0751	879.802	-3.013	0.003	-4375.567	-926.583
carat	9196.6121	90.779	101.308	0.000	9018.677	9374.547
cut	136.2981	8.019	16.998	0.000	120.581	152.015
depth	-2.1650	12.173	-0.178	0.859	-26.025	21.695
table	-22.7296	4.291	-5.296	0.000	-31.141	-14.318
x	-1927.7368	148.488	-12.982	0.000	-2218.786	-1636.688
y	1749.7102	146.531	11.941	0.000	1462.497	2036.924
z	-735.9007	152.845	-4.815	0.000	-1035.491	-436.311
color_Better	851.6689	22.353	38.100	0.000	807.854	895.484
color_Worst	1189.7210	22.092	53.852	0.000	1146.418	1233.024
clarity_Better	2569.2143	63.973	40.161	0.000	2443.821	2694.608
clarity_Worst	3586.7070	66.613	53.844	0.000	3456.140	3717.274

Omnibus:	3937.359	Durbin-Watson:	2.019
Prob(Omnibus):	0.000	Jarque-Bera (JB):	11440.752
Skew:	1.095	Prob(JB):	0.00
Kurtosis:	6.126	Cond. No.	1.04e+04

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

Here, we can see that the p value is < 0.05 for all the attributes except depth. Hence here, the depth is not so useful in predicting the price. In this particular, dataset, we are not having any difference in the R^2 and Adj R^2 values. Hence, this prediction model comes out to be a good model. This model is dependent on t-distribution which has a mean of 0. The RMSE value is found to be 994.89659.

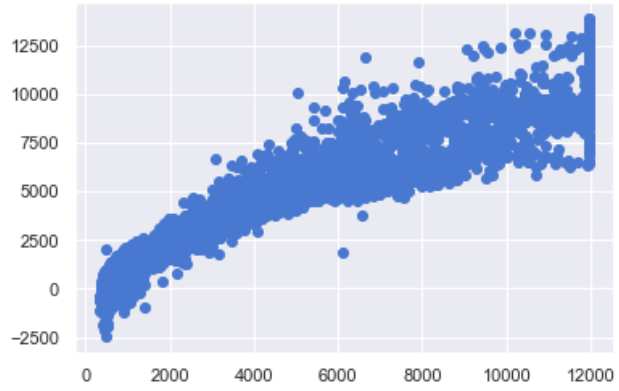


Figure 10 Linear regression built using statsmodel

However, out of the 3 models built, the model built using stats model can be considered the best, as it takes in to account the adjusted R^2 which does not change based on the number of attributes.

The final Linear model Regression equation obtained is as below

Price = (-2651.08) * Intercept + (9196.61) * carat + (136.3) * cut + (-2.16) * depth + (-22.73) * table + (-1927.74) * x + (1749.71) * y + (-735.9) * z + (851.67) * color_Better + (1189.72) * color_Worst + (2569.21) * clarity_Better + (3586.71) * clarity_Worst

Equation 3 Linear Regression Equation for Cubic Zirconia dataset

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

Given the Adj R^2 value of 0.917, this model is definitely going to be a very good model in predicting the price of a product. Also, it can be seen that carat, y, z color, clarity and x play the major role in deciding the price of the product. This is significant from their coefficient values in the Equation 3. Also, on analyzing the RMSE value, it is 994. Given this RMSE, it would have an impact on the sales of zirconia. Hence, the business should have an additional caution before pricing the model. Pricing greater by 1000 may have an impact on sales, while pricing less by 1000 would impact profitability. The company has been making profits majorly in the price range 2500 to 4000 that can be seen from the mean and median values of price. This price range has a mean carat of 0.8, mean y value of 5.9 and mean z value of 3.6. The company must also focus on improving the color and clarity for the products of this carat and dimensions at this average price band to attract consumer segment.

PROBLEM 2 – LOGISTIC REGRESSION & LINEAR DISCRIMINANT ANALYSIS

Problem Statement

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

Sample of the Dataset

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

Table 11 Sample of Holiday Dataset

The data consists of 7 columns having the employee details for 872 employees.

Dataset 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.

EXPLORATORY DATA ANALYSIS

Variable type

Holiday_Package	: Object
Salary	: int64
age	: int64
edu	: int64
no_young_children	: int64
no_older_children	: int64
foreign	: Object

Variable type is the type of data each column is holding. Here we are able to notice that Holiday_Package & foreign columns is of type object & remaining are of type int64.

Check for duplicated records in the dataset

On exploring the dataset, it is found that there are no duplicate records.

Check for missing values in the dataset

Holiday_Package	: 872 non-null
Salary	: 872 non-null
age	: 872 non-null
edu	: 872 non-null
no_young_children	: 872 non-null
no_older_children	: 872 non-null
foreign	: 872 non-null

There are no null values in the entire dataset.

Summary of the dataset

Summarizing briefly, the dataset has a total of 872 records. Mean Salary is 47.7 k with a standard deviation of 23.4 k, mean age is 39.9, mean education years is 9.3. Number of young children and number of older children are categorical and hence can be ignored while summarizing the means.

From the below table, we can see that the Mean and median are almost having the same value, for age and education years, which may be a sign for normal distribution. However, salary seems to be skewed.

	Salary	age	educ	no_young_children	no_older_children
count	872.000000	872.000000	872.000000	872.000000	872.000000
mean	47729.172018	39.955275	9.307339	0.311927	0.982798
std	23418.668531	10.551675	3.036259	0.612870	1.086786
min	1322.000000	20.000000	1.000000	0.000000	0.000000
25%	35324.000000	32.000000	8.000000	0.000000	0.000000
50%	41903.500000	39.000000	9.000000	0.000000	1.000000
75%	53469.500000	48.000000	12.000000	0.000000	2.000000
max	236961.000000	62.000000	21.000000	3.000000	6.000000

Table 12 Summary of the holiday dataset

DATA VISUALIZATION

Univariate Analysis – Box Plots, Histograms & Count Plots

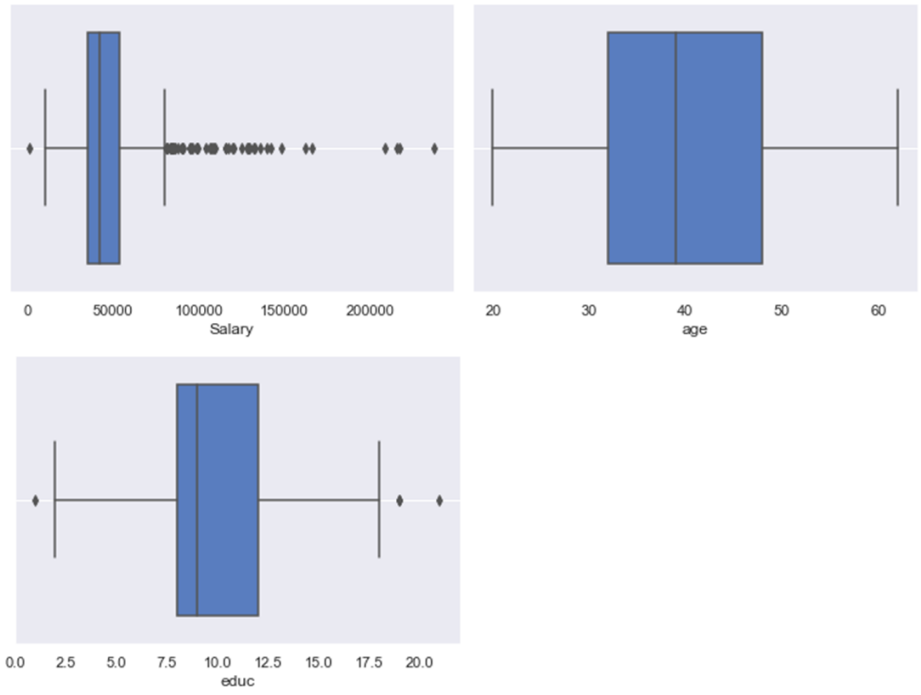


Figure 11 Box Plot - Holiday Dataset

From figure 11, we are able to understand that there are many outliers in the Salary variable, no outliers in the age and 3 outliers in the education years.

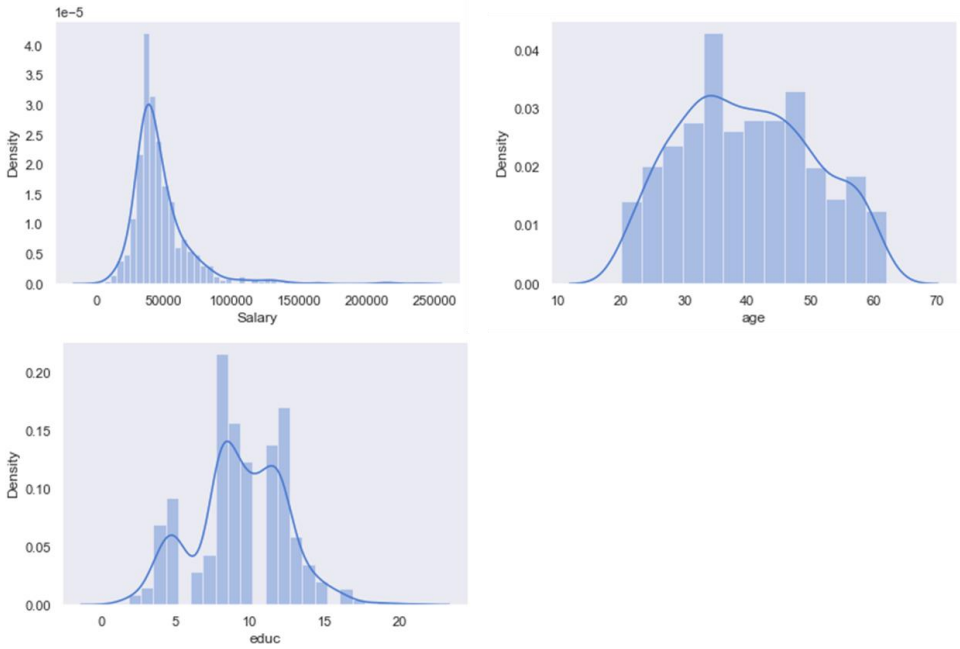


Figure 12 Histogram - Holiday Dataset

From figure 12, we can see that, salary variable is right skewed, with higher kurtosis that is evident by its peak. The age is more or less flat with multiple nodes. Age is more or less normally distributed with slight skew towards right. The Education years are multimodal.

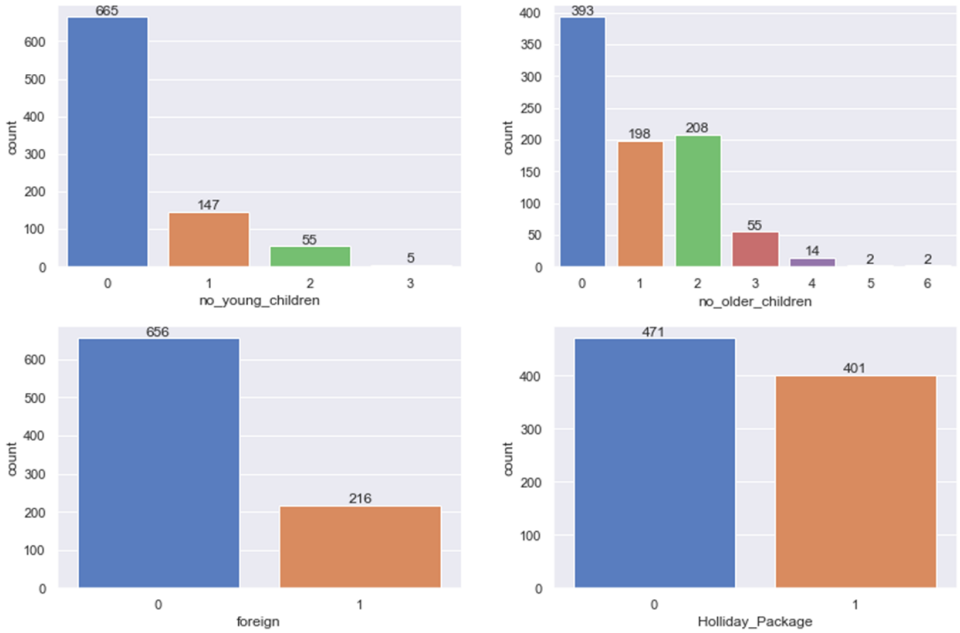


Figure 13 Count Plot - Holiday Dataset

From figure 13, we can see that, most of the employees have zero young & old children. Very few have 3 young children. Similarly, very few have 4, 5 & 6 old children. There are less foreigners in the employees. The Yes & No category of the Holiday Package are almost balanced. However, number of employees accepting the holiday package are less in comparison with those rejecting the package.

Bivariate Analysis – Box Plots, Count Plots, Heatmap & Pair plot

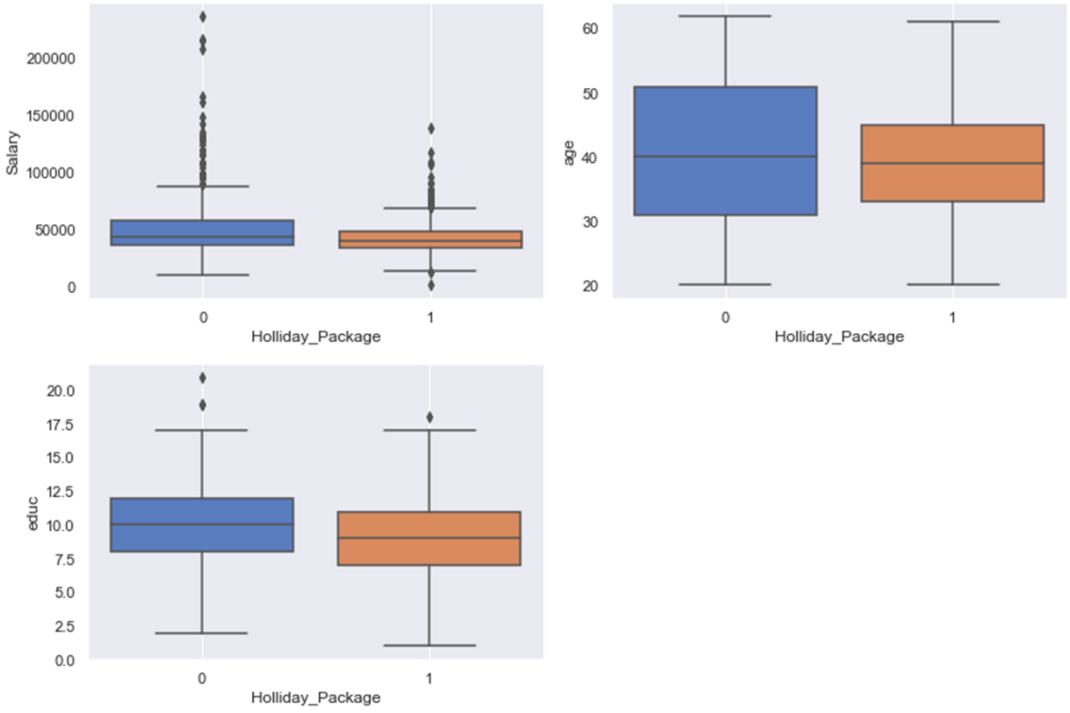


Figure 14 Box Plot - Multi variate

From Figure 14, we can see that, there are more outliers in Salary variable are in category rejecting the holiday package. Similarly, the age band of people rejecting the package is higher than people accepting. There are 2 outliers in the education years for the category of employees rejecting the package while there is only 1 outlier in the other.

From figure 15, we can see that, employees having no young children are found to be accepting the holiday packages. In case of employees with older children, those with 2 & 3 older children are found to accept the package in majority. Coming to nationality, foreigners are found to be more inclined towards accepting the holiday package.

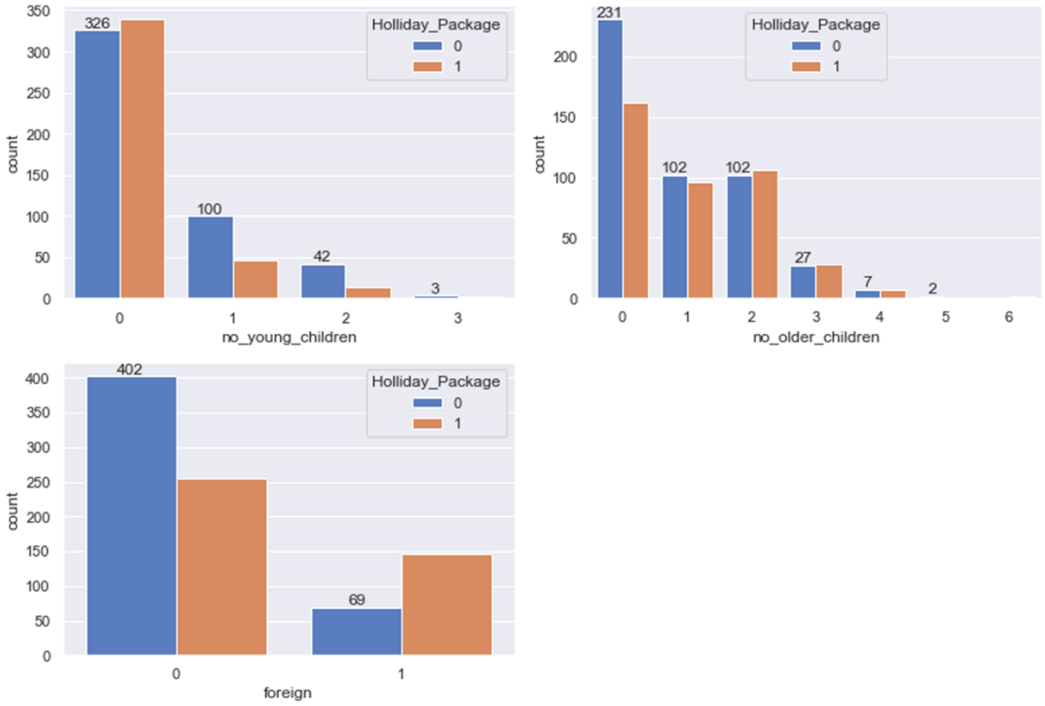


Figure 15 Count Plot - Multi variate

From figure 16, we can see that, we can see that there is very negligible correlation amongst the variables.

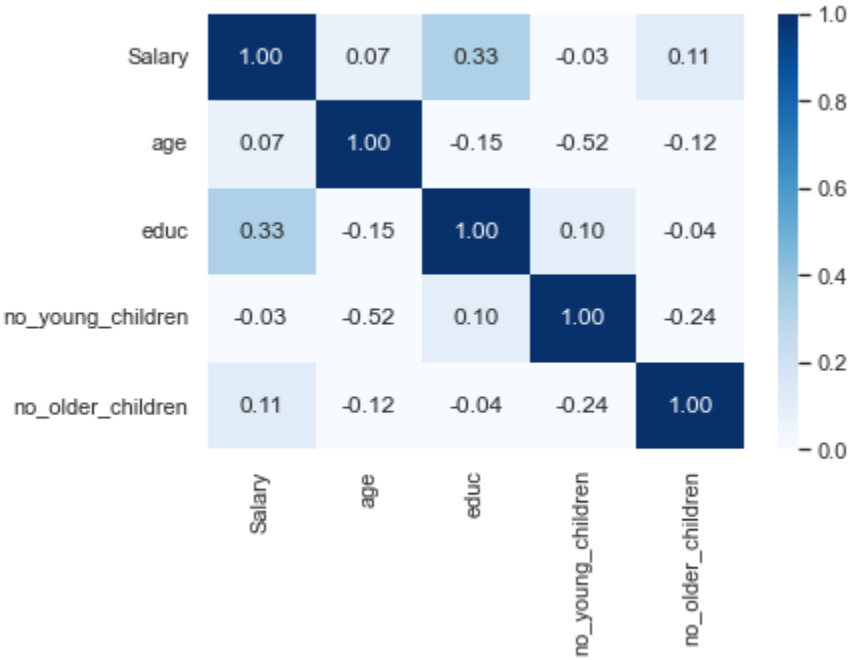


Figure 16 Heat Map - Holiday Dataset

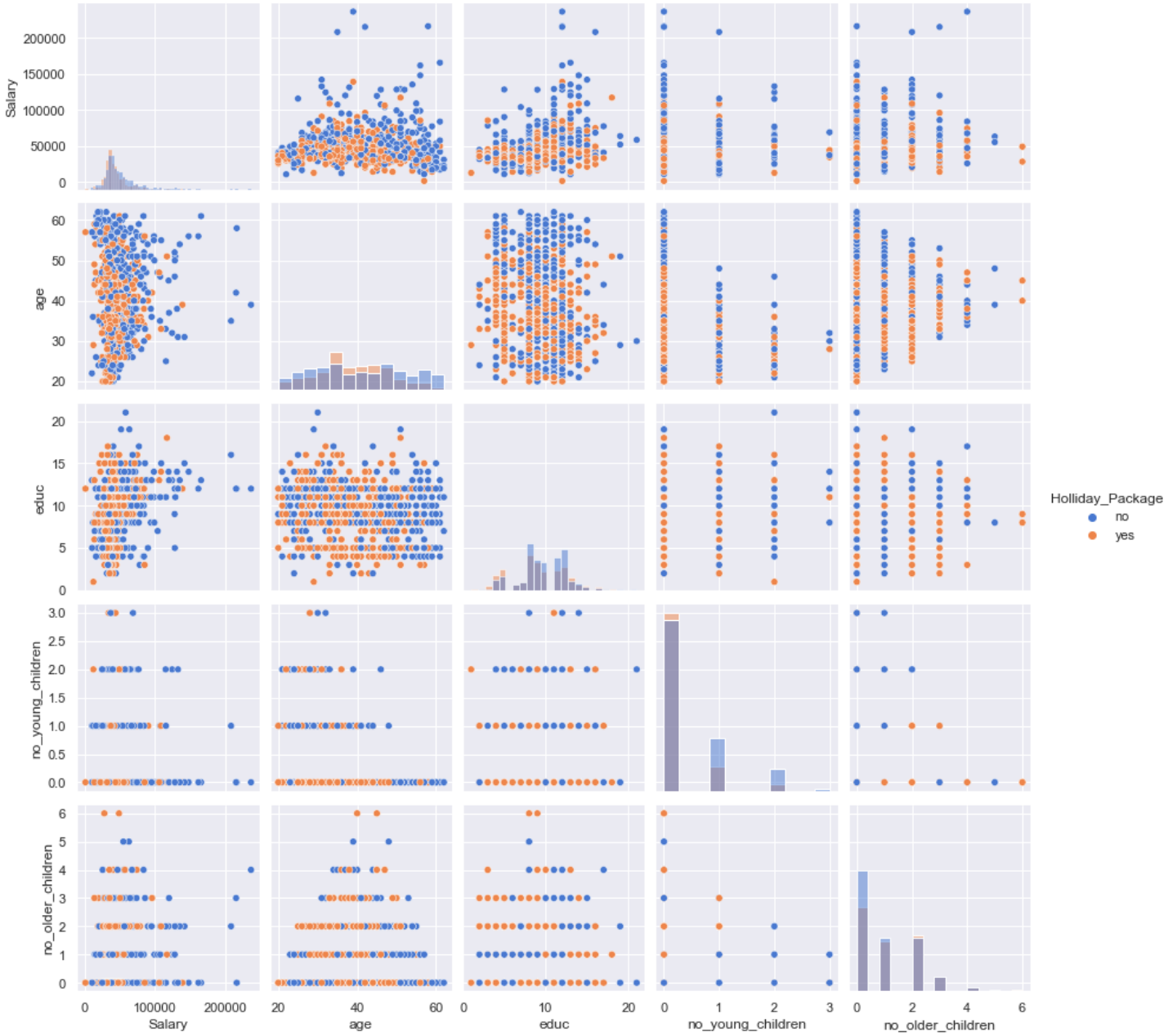


Figure 17 Pair Plot with hue - Holiday Dataset

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).

ENCODING

Logistic Regression or Linear Discriminant Analysis deal only with numbers. Hence the object type variables in the dataset need to be converted to numerical type. Here, the holiday_package & foreign are object type variables and need to be converted to numerical type.

We can go ahead with label encoding as both are Yes & No type, assigning 0 for No and 1 for Yes.

SPLITTING DATA

Holiday Package being the predicted variable is stored as a series separately in a variable y, while the other predictor variables are stored as data frame in x. Then using the train_test_split library in sklearn, we shall be able to split the data in to 70% training & 30% testing.

LOGISTIC REGRESSION

- It is a supervised learning method for classification that establishes relation between dependent class variable and independent variables using regression
- The dependent variable is categorical i.e. it can take only integral values representing different classes
- The probabilities describing the possible outcomes of a query point are modelled using a logistic function
- Belongs to family of discriminative classifiers. They rely on attributes which discriminate the classes well
- Logistic Regression assigns probabilities to different classes.
- To do so, it learns from the training set a vector of weights and bias.
- Each weight (wi) is assigned to one input feature Xi
- The weight assigned to each feature represents how important that feature is for classification decision
- The weights can be positive i.e., direct correlation of the feature with the class of interest, while a negative weight indicates inverse relation with the class of interest
- A simple Logistic regression is represented as $Z = w.x + b$

Equation 4 Logistic regression Model

- Since the weights are running numbers and so is the bias term, Z can take values from –infinity to + infinity
- To transform the value of Z into probability (range between 0 and 1) , Z is passed through Sigmoid function (mathematical transformation)
- The algorithm uses cross-entropy loss function (negative log likelihood loss) to find the most optimal weights and bias across entire data set put together (N records)

- Most optimal weights and bias would be those that minimize overall all training error i.e., misclassification in the training data

$$logLoss = \frac{-1}{N}\sum_{i=1}^{N}(y_i(logp_i) + (1 - y_i)log(1 - p_i))$$

Equation 5 Cross Entropy Loss function

- Incorrect classification will add large magnitude to the loss function while correct classification will contribute very minimal to the loss function

LOGISTIC REGRESSION MODEL USING SKLEARN

By using the LogisticRegression model library from sklearn.linear_model, we will be able to fit the train data in to our model. By fitting the data, the model gets trained using the training set. By using the appropriate python codes, we will be able to predict the classes and also the probability of that record being predicted to that class.

LINEAR DISCRIMINANT ANALYSIS

- Discriminant Analysis is used for classifying observations to a class or category based on predictor (independent) variables of the data.
- Discriminant Analysis creates a model to predict future observations where the classes are known.
- LDA uses linear combinations of independent variables to predict the class in the response variable of a given observation. LDA assumes that the independent variables(p) are normally distributed and there is equal variance/ covariance for the classes. LDA is popular because it can be used for both classification and dimensionality reduction.
- When these assumptions are satisfied, LDA creates a linear decision boundary. Note that based on many research study, it is observed that LDA performs well when these assumptions are violated.
- LDA is based upon the concept of searching for a linear combination of predictor variables that best separates the classes of the target variable.
- The LDA model gives linear combinations of the predictor variables as follows:

$$DS = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \ldots + \beta kXk$$

Equation 6 LDA Structure

Where:

- DS = Discriminant Score
- β's = Discriminant weight (coefficients)
- X's = Explanatory (Predictor or independent) variables

- The weights (or coefficients) are estimated so that the groups are separated as clearly as possible on the values of the discriminant functions.
- LDA constructs an equation which minimizes the possibility of misclassifying cases into their respective classes.
- LDA comes to our rescue in situations where logistic regression is unstable
  - Classes are well separated – Logistic Regression lacks stability when the classes are well - separated. That's when LDA comes in.
  - The data is small.
- We have more than two classes – LDA is a better choice whenever multi- class classification is required.
- LDA Compute the d-dimensional mean vectors for the different classes of the dataset
- Calculate between - class variance, the separability between the mean of different classes
- Calculate within – class variance, the separability between the mean and sample of each classes
- Compute the eigen vectors (e1, e2, ..., en) and the corresponding eigen values (λ1, λ2, ...,λn) for the scatter matrices. An eigen vector, corresponding to a real non-zero eigen value, points in a direction that is stretched by the transformation and the eigen value is the factor by which it is stretched. Negative eigen value indicates the direction is reversed.

- Eigen vector, v of a matrix A is the vector for which the following equation is satisfied: $Av = \lambda v$, where $\lambda$is a scalar value called the eigen value. This implies that the linear transformation A on vector v is completely defined by λ.
- Sort the eigen vectors by decreasing eigen values and choose k eigen vectors with the largest eigen values to form a n X k dimensional matrix W
- Construct a lower dimensional space projection using Fisher's criterion, which maximizes the between class variance and minimizes the within – class variance.
- LDA model uses Bayes' Theorem to estimate probabilities. They make predictions upon the probability that a new input dataset belongs to each class. The class which has the highest probability is considered as the output class and then the LDA makes a prediction.
- The prediction is made simply by the use of Bayes' theorem which estimates the probability of the output class given the input. They also make use of the probability of each class and also the data belonging to that class.

LDA MODEL USING SKLEARN

By using the LinearDiscriminantAnalysis model library from sklearn.discriminant_analysis, we will be able to fit the train data in to our model. By fitting the data, the model gets trained using the training set. By using the appropriate python codes, we will be able to predict the classes and also the probability of that record being predicted to that class.

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.

ACCURACY, F1 SCORE, PRECISION & RECALL

Here, the subject of interest for us is the employees who are going to accept the package. Hence considering 1, the accuracy, F1 score, precision & recall are mapped in the table below.

		Logistic Regression		LDA	
		Train Data	Test Data	Train Data	Test Data
For model	Accuracy	0.67	0.65	0.66	0.64
For 1	Precision	0.66	0.65	0.65	0.64
	Recall	0.58	0.52	0.58	0.49
	F1 Score	0.62	0.58	0.61	0.56

Table 13 Classification Report - LDA & Logistic Regression

It can be seen that all the parameters are better performed in Logistic Regression model. However, we need to focus on precision here considering the subject of interest.

CONFUSION MATRIX

Logistic Regression Test Data			
Confusion Matrix		Predicted	
		0	1
Actual	0	109	33
	1	58	62

LDA Test Data			
Confusion Matrix		Predicted	
		0	1
Actual	0	109	33
	1	61	59

Table 14 Confusion Matrix - Test Data - LDA & Logistic Regression

From Table 14, it is evident that the True Positives are high in the Logistic Regression test data compared to the LDA Test data.

ROC CURVE & ROC_AUC SCORE

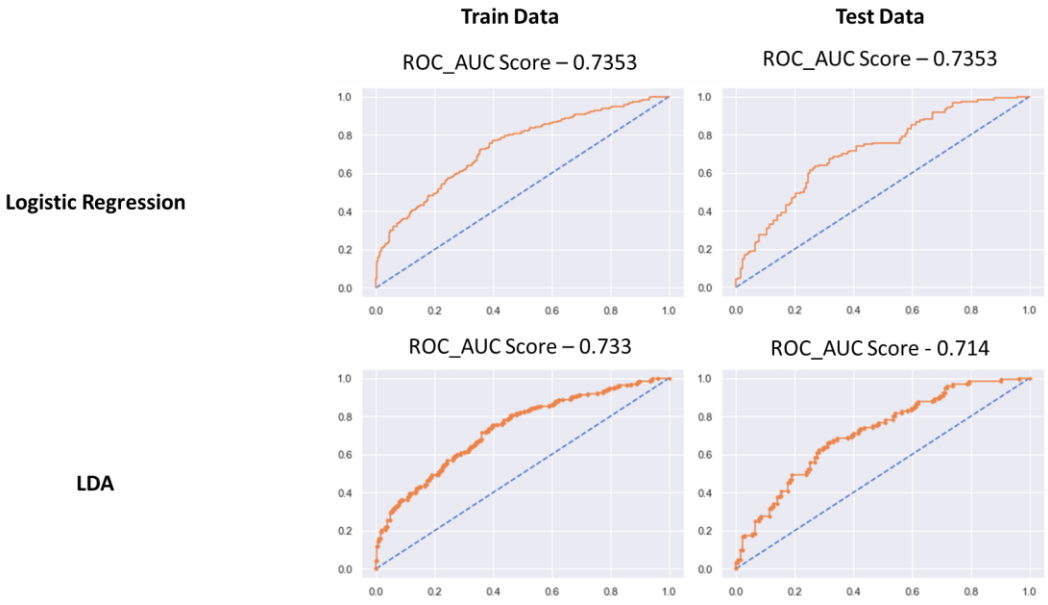


Figure 18 ROC Curve - LDA & Logistic Regression

From Figure 18, it can be seen that the ROC_AUC score is higher for the Logistic Regression Model.

Concluding from the above inferences, it can be seen that Logistic regression is better performing in this case of Holiday Package.

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

- Thus, through Logistic Regression & LDA, we were able to predict the Class of people who were likely to buy the holiday package.
- Employees having no young children are found to be accepting the holiday packages.
- In case of employees with older children, those with 2 & 3 older children are found to accept the package in majority.
- Coming to nationality, foreigners are found to be more inclined towards accepting the holiday package.
- Hence, the company must focus on employees with the above traits.

END