




---

# Predictive Modeling Business Report

---

 **Name:** Hemant Patidar  
 **Batch:** PGPDSBA Online Sep\_A 2021  
 **Date:** 20/02/2022

## Table of Contents

<b>Linear Regression – GemStone Company</b>	<b>6</b>
Executive Summary	6
Introduction	6
Data Description	6
<b>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.</b>	<b>6</b>
Head & Tail of the Data:	7
Data Information:	7
Data Clean-up:	8
Univariate Data Analysis	8
Data skewness, Kurtosis and Outlier proportion	15
Multivariate Data Analysis	15
<b>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 an ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning.</b>	<b>18</b>
Zeros Count in Features	19
Ordinal values	19
<b>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 &amp; Adj Rsquare. Compare these models and select the best one with appropriate reasoning.</b>	<b>21</b>
Encoding summary	21
Splitting Train vs. Test Data	21
Scikit Learn's Linear Regression	21
R <sup>2</sup> – Coefficient of determinant (Model Score)	22
Outlier treatment & Scaling	23
VIF (Variance Inflation Factor)	25
StatsModel Linear Regression	26
Model Selection –	28
<b>1.4 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.</b>	<b>29</b>
Inference –	29
Insights –	29
Recommendation –	29

<b>Logistic Regression and LDA – Travel Agency</b>	<b>31</b>
Executive Summary	31
Introduction	31
Data Description	31
<b>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.</b>	<b>31</b>
Head & Tail of the Data:	31
Data Information:	32
Outlier Proportion:	33
Skewness & Kurtosis:	33
Univariate Analysis:	33
Multivariate Analysis:	38
<b>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).</b>	<b>40</b>
Data Encoding	40
Target Variable Proportion	41
Splitting Train vs. Test Data	41
Logistic Regression Model -	41
Linear Discriminant Analysis –	44
<b>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.</b>	<b>45</b>
Confusion Matrix –	45
ROC Curve & ROC-AUC Score –	46
Model Comparison –	47
<b>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.</b>	<b>49</b>
Insights derived from EDA and building logistic regression & LDA –	49
Recommendation	49

## List of Figures

Figure I - Univariate Analysis (Carat).....	9
Figure II - Univariate Analysis (Depth) .....	9
Figure III - Univariate Analysis (Table) .....	10
Figure IV - Univariate Analysis (Length) .....	10
Figure V - Univariate Analysis (Width) .....	11
Figure VI - Univariate Analysis (Height) .....	11
Figure VII - Univariate Analysis (Price) .....	12
Figure VIII - Univariate Analysis (Cut).....	12
Figure IX - Price vs. Cut.....	12
Figure X - Univariate Analysis (Color).....	13
Figure XI - Price vs. Color.....	13
Figure XII - Univariate Analysis (Clarity) .....	14
Figure XIII - Price vs. Clarity.....	14
Figure XIV - Pair Plot.....	16
Figure XV - Correlation Matrix (Heat Map).....	17
Figure XVI - Outliers Analysis in Data Frame.....	18
Figure XVII - Summary after Missing/Zero Value Impute .....	19
Figure XVIII - Clarity Chart.....	20
Figure XIX - Actual vs. Predicted values .....	23
Figure XX - Outlier Analysis (Post treatment) .....	24
Figure XXI - OLS Regression Summary .....	27
Figure XXII - Predicted price vs Actual (StatsModel Regression Model).....	28
Figure XXIII – Univariate Analysis (Salary).....	34
Figure XXIV - Univariate Analysis (Age).....	34
Figure XXV - Univariate Analysis (Education).....	35
Figure XXVI - Univariate Analysis (No. Young Children).....	35
Figure XXVII - Count Plot (# of young Children) .....	36
Figure XXVIII - Univariate Analysis (No. Older Children).....	36
Figure XXIX - Count Plot (# Of Older Children).....	37
Figure XXX - Univariate Analysis (Holiday Pakage & Foreign).....	37
Figure XXXI - Pair Plot (Hue - Holiday Package) .....	38
Figure XXXII - Pair Plot (Hue - Foreign).....	39
Figure XXXIII - Holiday Package & Foreign Employee .....	40
Figure XXXIV - Correlation Matrix (Heat Map).....	40
Figure XXXV - Before Outlier Treatment.....	42
Figure XXXVI - After Outlier Treatment.....	42
Figure XXXVII - Confusion Matrix (Logistic Regression) .....	45
Figure XXXVIII - Confusion Matrix (LDA) .....	46
Figure XXXIX - ROC Curve (Logistic Regression) .....	46
Figure XL - ROC Curve (LDA).....	47
Figure XLI - Train Data ROC Curve.....	47
Figure XLII - Test Data ROC Curve .....	48

## List of Tables

Table 1 - Data Dictionary.....	6
Table 2 - Head data (5 Rows) .....	7
Table 3 - Tail Data (5 Rows).....	7
Table 4 - Dataframe Info .....	7
Table 5 - Data Summary.....	8
Table 6 - Cut Data Values .....	13
Table 7 - Color Data Values .....	14
Table 8 - Clarity Data Values .....	15
Table 9 - Outlier Proportion .....	15
Table 10 - Skewness & Kurtosis Value .....	15
Table 11 - Zeros Count .....	19
Table 12 - New Clarity Level Proportion .....	20
Table 13 - Feature Coefficients (Scikit Learn LR Model) .....	21
Table 14 - Regularized Coeff. & Intercept.....	22
Table 15 - Models' R2.....	22
Table 16 - After Scaling Data Summary .....	24
Table 17 - Coefficients on Scaled Data LR Model .....	25
Table 18 - Variance Inflation Factor.....	25
Table 19 - Intercept & Coeff. from StatsModel LR.....	26
Table 20 - Feature Importance .....	28
Table 21 - Data Dictionary (Travel Agency).....	31
Table 22 - Head Data (5 Rows).....	32
Table 23 - Tail Data (5 Rows).....	32
Table 24 - Data Information (Agency Data) .....	32
Table 25 - Data Summary.....	33
Table 26 - Outlier Proportion .....	33
Table 27 - Skew & Kurtosis Values .....	33
Table 28 - Frequency (# Of Young Children) .....	36
Table 29 - Frequency (# Of Older Children) .....	37
Table 30 - Frequency (Holiday Package & Foreign) .....	38
Table 31 - Classification Report with 0.5 prob threshold (Logistic Regression).....	43
Table 32 - Classification Report with 0.45 prob threshold (Logistic Regression).....	43
Table 33 - Feature Coefficient (Logistic Regression).....	44
Table 34 - Feature Coefficients (LDA) .....	44
Table 35 - Classification Report (Train Data) LDA .....	45
Table 36 - Classification Report (Test Data) LDA.....	45
Table 37 - Metrics Comparison .....	48

## Linear Regression – GemStone Company

### Executive Summary

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.

### Introduction

Purpose of our exercise would be to predict price for the stone based on individual predictors and identify higher profitable and lower profitable stones.

### Data Description

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, SI1, SI2, I1
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.

*Table 1 - Data Dictionary*

**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.**

## Head & Tail of the Data:

	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
0	1	0.30	Ideal	E	SI1	62.1	58.0	4.27	4.29	2.66	499
1	2	0.33	Premium	G	IF	60.8	58.0	4.42	4.46	2.70	984
2	3	0.90	Very Good	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289
3	4	0.42	Ideal	F	VS1	61.6	56.0	4.82	4.80	2.96	1082
4	5	0.31	Ideal	F	VVS1	60.4	59.0	4.35	4.43	2.65	779

Table 2 - Head data (5 Rows)

	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
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 3 - Tail Data (5 Rows)

## Data Information:

Column	Non-Null Items	Dtype
unnamed: 0	26967	int64
carat	26967	float64
cut	26967	object
color	26967	object
clarity	26967	object
depth	26270	float64
table	26967	float64
x	26967	float64
y	26967	float64
z	26967	float64
price	26967	int64

Table 4 - Dataframe Info

- We have 26,967 rows and 11 columns in data frame
- **Unnamed: 0** is unnecessary field in the data and represents the index of row
- We have **697** null values in depth feature
- Since x, y and z can be represent as length, width & height in data frame
- Data frame has 3 features as object and rest are numeric
- Data frame has **34** duplicate records

## Data Clean-up:

- Dropped **Unnamed: 0** field
- Removed **34** duplicated records
- Renamed x, y and z as length, width, and height

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
carat	26933.0	NaN	NaN	NaN	0.79801	0.477237	0.2	0.4	0.7	1.05	4.5
cut	26933	5	Ideal	10805	NaN	NaN	NaN	NaN	NaN	NaN	NaN
color	26933	7	G	5653	NaN	NaN	NaN	NaN	NaN	NaN	NaN
clarity	26933	8	SI1	6565	NaN	NaN	NaN	NaN	NaN	NaN	NaN
depth	26236.0	NaN	NaN	NaN	61.745285	1.412243	50.8	61.0	61.8	62.5	73.6
table	26933.0	NaN	NaN	NaN	57.45595	2.232156	49.0	56.0	57.0	59.0	79.0
length	26933.0	NaN	NaN	NaN	5.729346	1.127367	0.0	4.71	5.69	6.55	10.23
width	26933.0	NaN	NaN	NaN	5.733102	1.165037	0.0	4.71	5.7	6.54	58.9
height	26933.0	NaN	NaN	NaN	3.537769	0.719964	0.0	2.9	3.52	4.04	31.8
price	26933.0	NaN	NaN	NaN	3937.52612	4022.551862	326.0	945.0	2375.0	5356.0	18818.0

Table 5 - Data Summary

With above data summary, we can infer below –

- Average carat weight of cubic zirconia is 0.798, minimum weight is 0.2 and maximum is 4.5
- Most of the zirconia are in ideal (best) cut, having ~40% weightage in all data
- Zirconia's color are mostly average (G being a value between D: worst – J: best), almost ~20% of zirconia details in data having average color
- Most of the stones are towards best in clarity (SI1 being value between IF: worst – I1: best), having ~24% weightage in the data
- Average depth calculated in stones are 61.745 (697 missing values), the min-max values are 50.8 – 73.6
- Zirconia's table average % width against it diameter is 57.45, having minimum % of 49 and maximum 79
- Length, width & height of cubic is 1.12, 1.16 and 0.72 respectively, and having zeros in dataset as well
- Price of zirconia varies from 326 to 18,818, having an average value of 3937.52

## Univariate Data Analysis



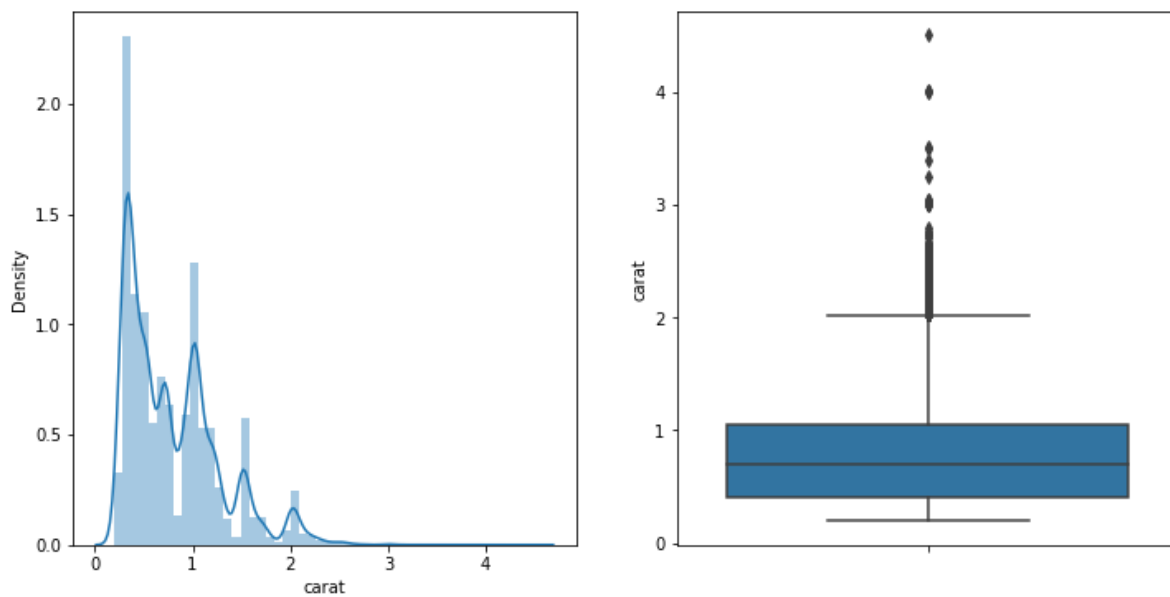


Figure I - Univariate Analysis (Carat)

**Carat:** Data seem positively skewed and lies between a range of 0 – 3, it has several outliers that can be seen in boxplot.

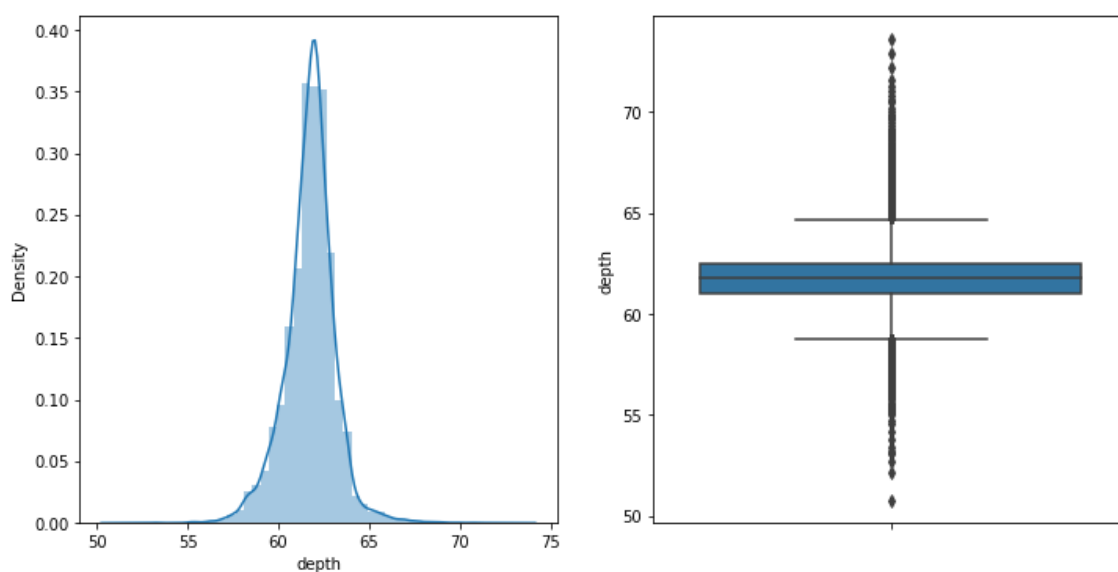


Figure II - Univariate Analysis (Depth)

**Depth:** Data seem normally distributed with bell curve share, but we already know that they are missing values for this feature and by boxplot, we can see that we have outliers at both ends.

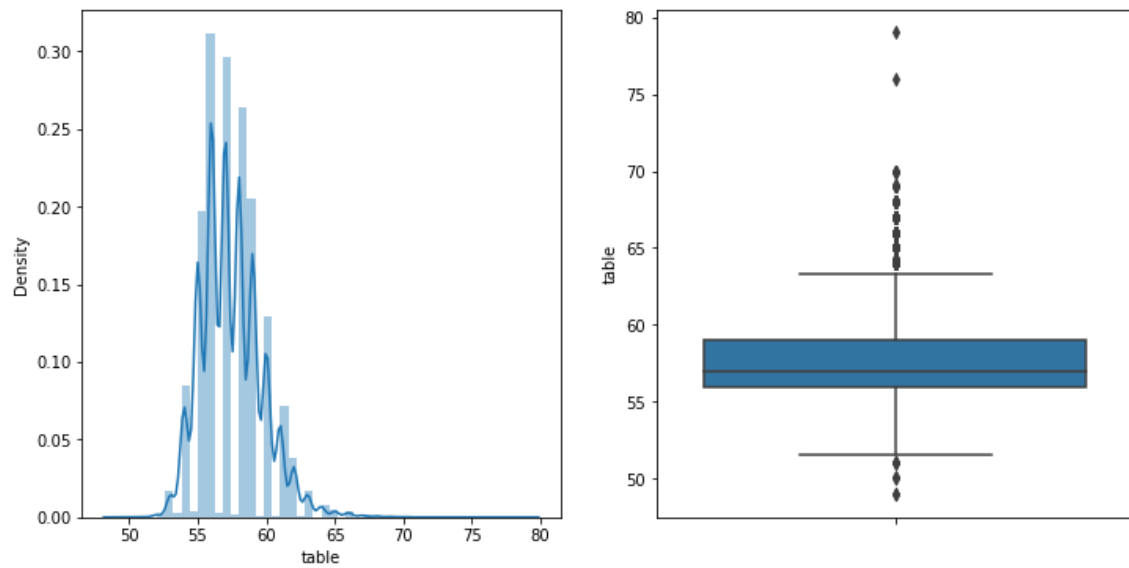


Figure III - Univariate Analysis (Table)

**Table:** With data bins missing at specific intervals, the distribution can be considered normally distributed as bin size increases. The table data has outliers at both ends and median lies somewhere between 55 – 60.

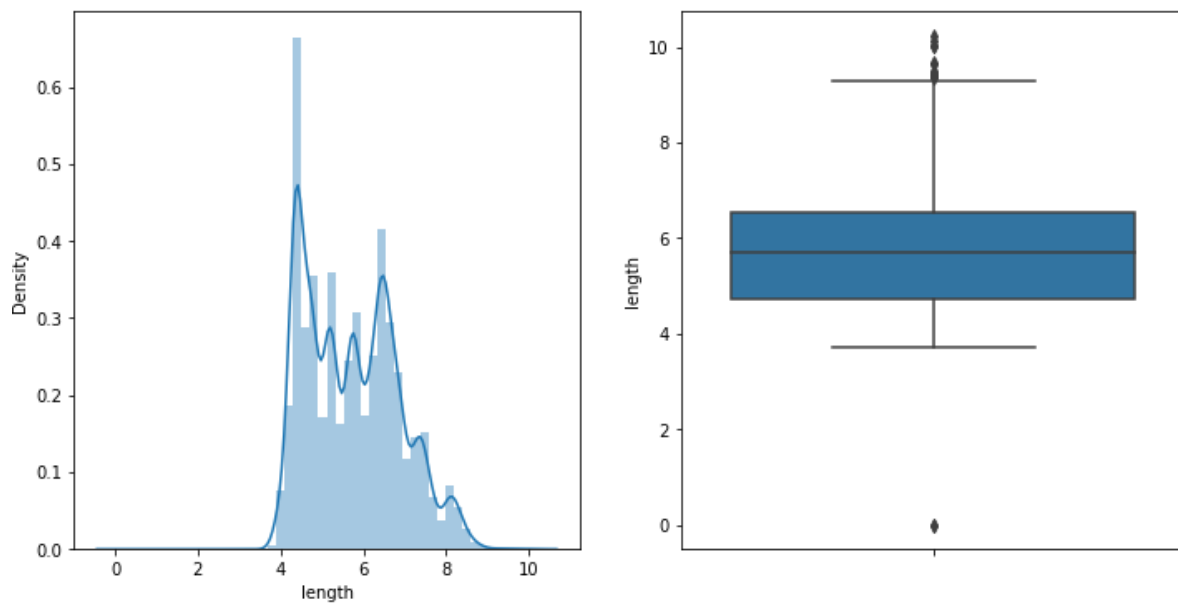


Figure IV - Univariate Analysis (Length)

**Length:** Data has outliers and distribution seem normally distributed. An unusual extreme low value can be seen in boxplot at ~0.

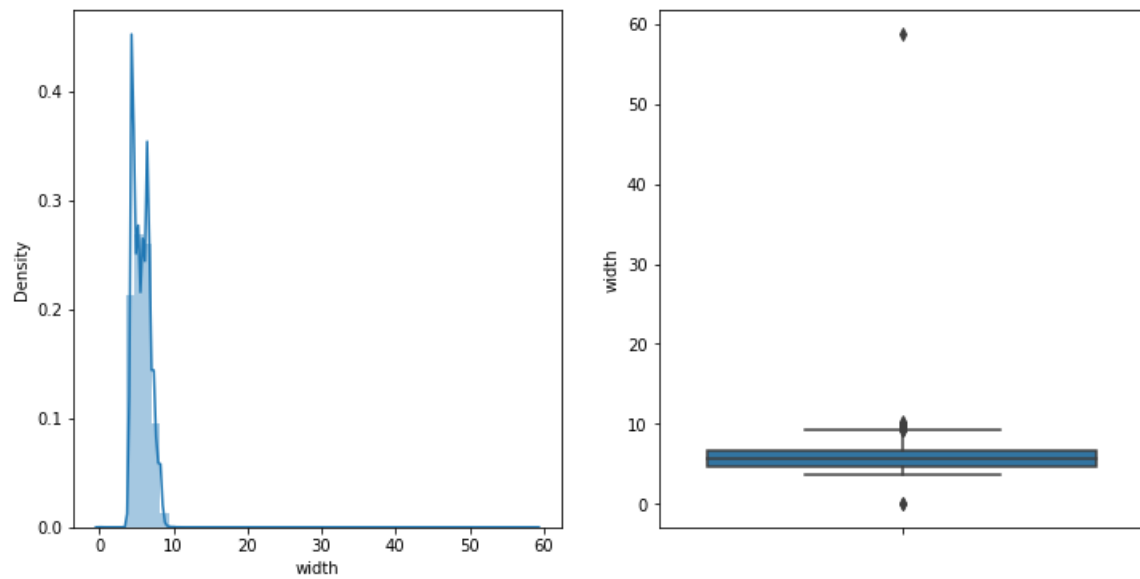


Figure V - Univariate Analysis (Width)

**Width:** Data seem positively skewed and it has outliers. An unusual extreme high value can be seen in boxplot at ~60.

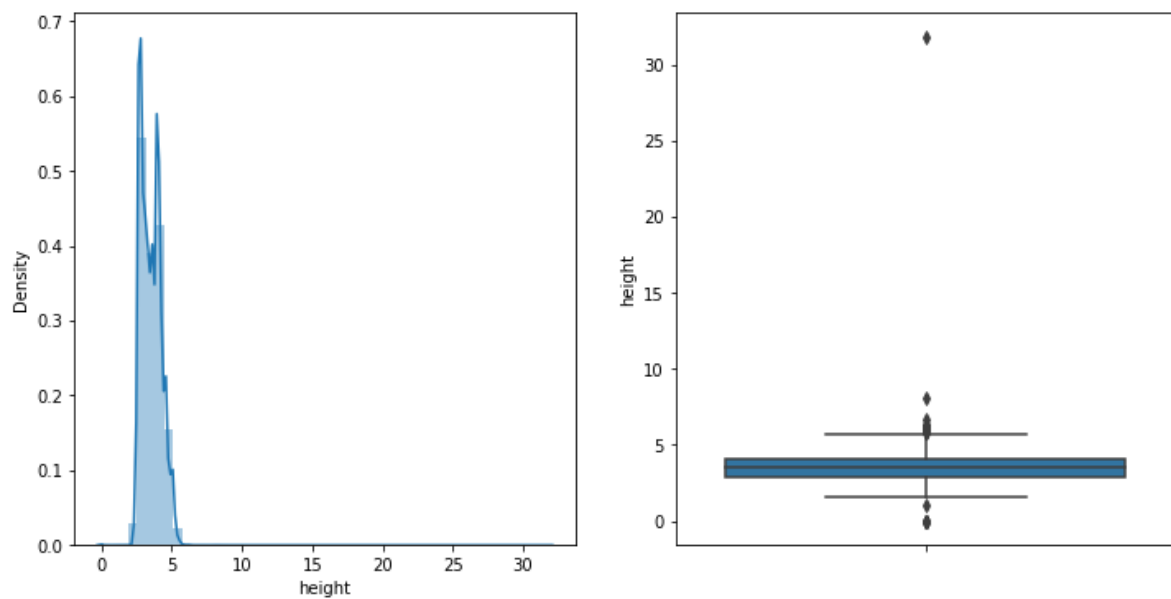


Figure VI - Univariate Analysis (Height)

**Height:** Data have outliers at both ends, and data distribution seem positively skewed within a range of 2-5. An unusual extreme high value can be seen in boxplot at ~30.

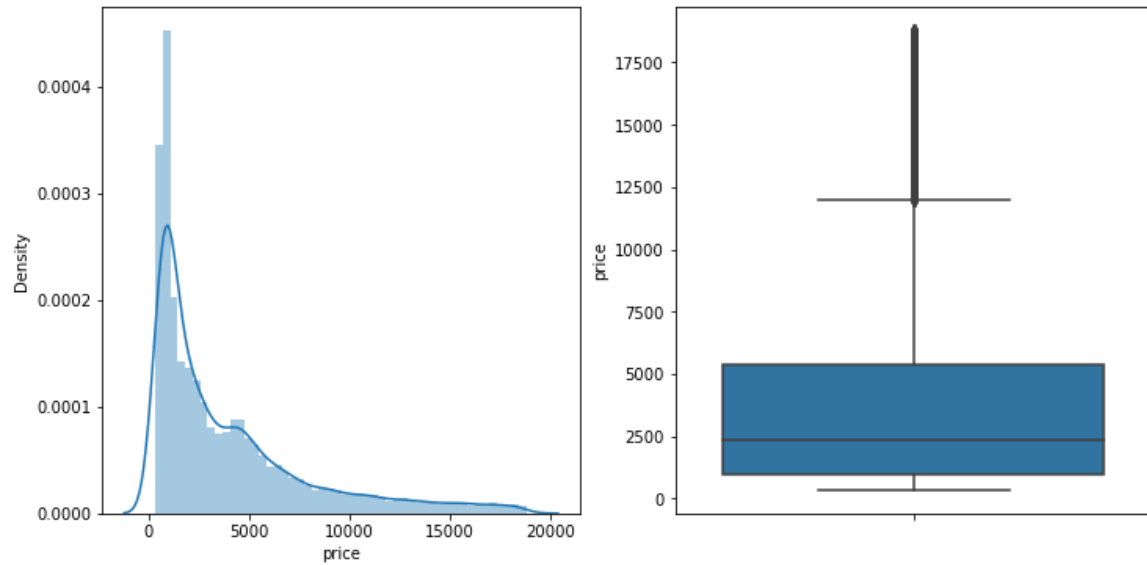


Figure VII - Univariate Analysis (Price)

**Price:** Data seem normally distributed and having outliers at upper threshold.

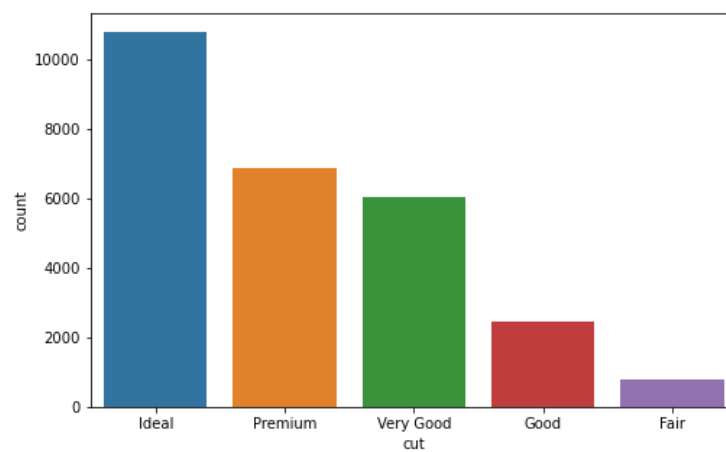


Figure VIII - Univariate Analysis (Cut)

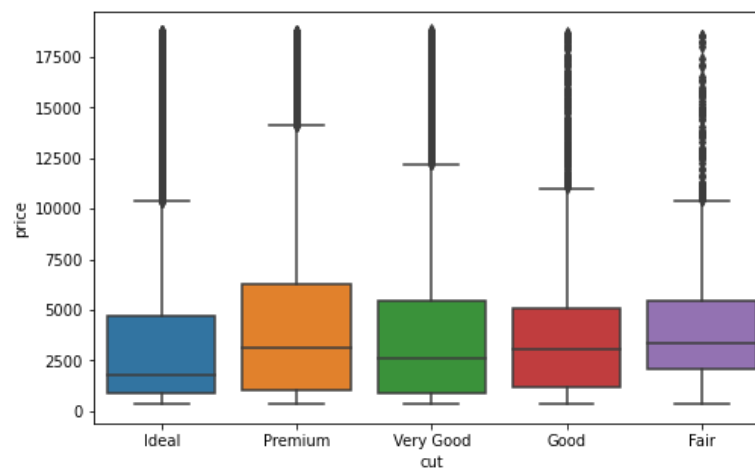


Figure IX - Price vs. Cut

**Cut:** The plot above being sorted on best to bad cut, and we can see that most of the stones are in best cut possible.

Category	Value counts	% in Data
Ideal	10805	40.12%
Premium	6886	25.57%
Very Good	6027	22.38%
Good	2435	9.04%
Fair	780	2.90%

Table 6 - Cut Data Values

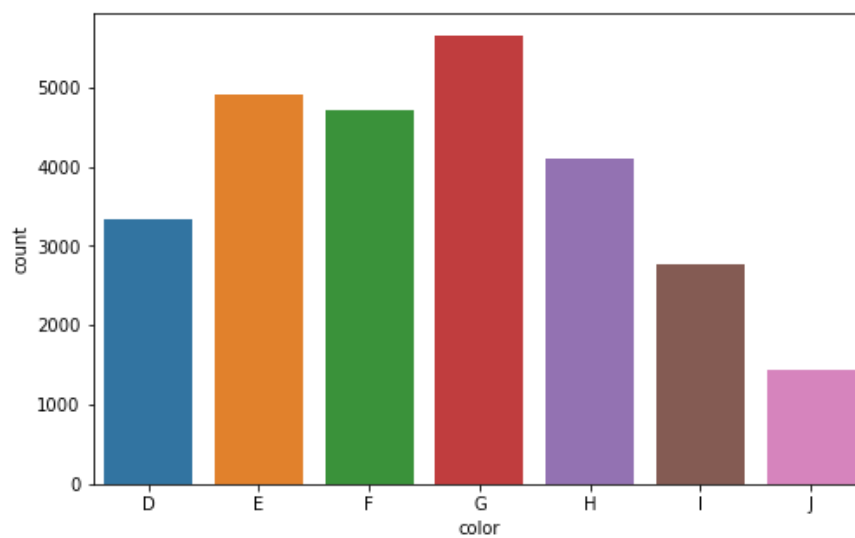


Figure X - Univariate Analysis (Color)

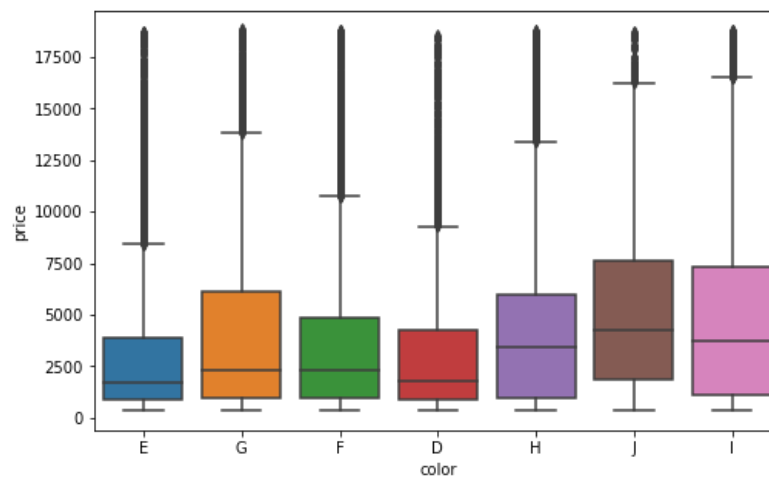


Figure XI - Price vs. Color

**Color:** Sorted in order of bad – good quality, and we can see most of the stones have average color quality as G.

Category	Value counts	% in Data
G	5653	20.99%
E	4916	18.25%
F	4723	17.54%
H	4095	15.20%
D	3341	12.40%
I	2765	10.27%
J	1440	5.35%

Table 7 - Color Data Values

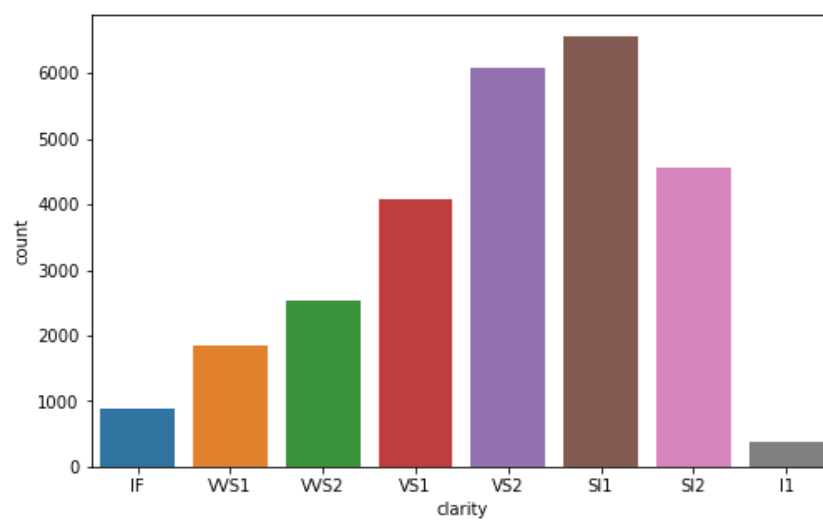


Figure XII - Univariate Analysis (Clarity)

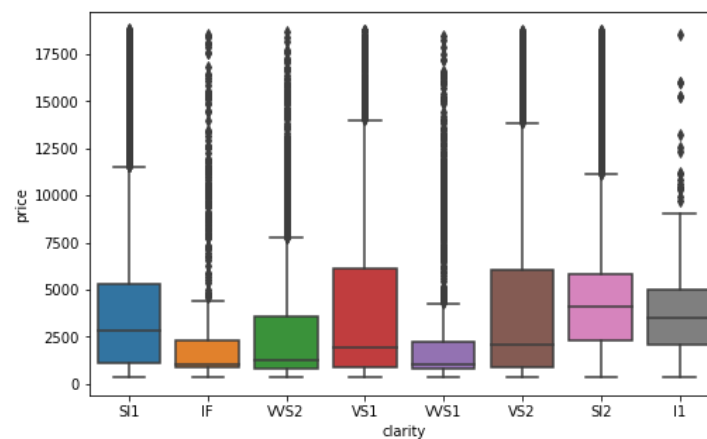


Figure XIII - Price vs. Clarity

**Clarity:** Plot data was sorted in bad to good clarity, and most of the stone are aligned at good clarity and very few of them at best clarity parameter.

Category	Value counts	% in Data
----------	--------------	-----------

SI1	6565	24.38%
VS2	6093	22.62%
SI2	4564	16.95%
VS1	4087	15.17%
VVS2	2530	9.39%
VVS1	1839	6.83%
IF	891	3.31%
I1	364	1.35%

Table 8 - Clarity Data Values

## Data skewness, Kurtosis and Outlier proportion

Outliers are the values beyond  $Q3 + 1.5 \text{ IQR}$  or  $Q1 - 1.5 \text{ IQR}$ , they may affect models or decisions.

Feature	Outliers	% Outlier
carat	657	2.44%
depth	1219	4.65%
table	318	1.18%
length	14	0.05%
width	14	0.05%
height	22	0.08%
price	1778	6.60%

Table 9 - Outlier Proportion

Feature	Skew	Kurtosis
carat	1.115	1.210
depth	-0.026	3.681
table	0.766	1.583
length	0.392	-0.679
width	3.868	160.013
height	2.581	87.407
price	1.619	2.152

Table 10 - Skewness & Kurtosis Value

## Multivariate Data Analysis

- Updated categorical variables into codes
  - **Cut:** Fair:1, Good:2, Very Good:3, Premium:4, Ideal:5

- **Color:** D:1, E:2, F:3, G:4, H:5, I:6, J:7
- **Clarity:** IF:1, VVS1:2, VVS2:3, VS1:4, VS2:5, SI1: 6, SI2:7, I1:8

### Pair plot:

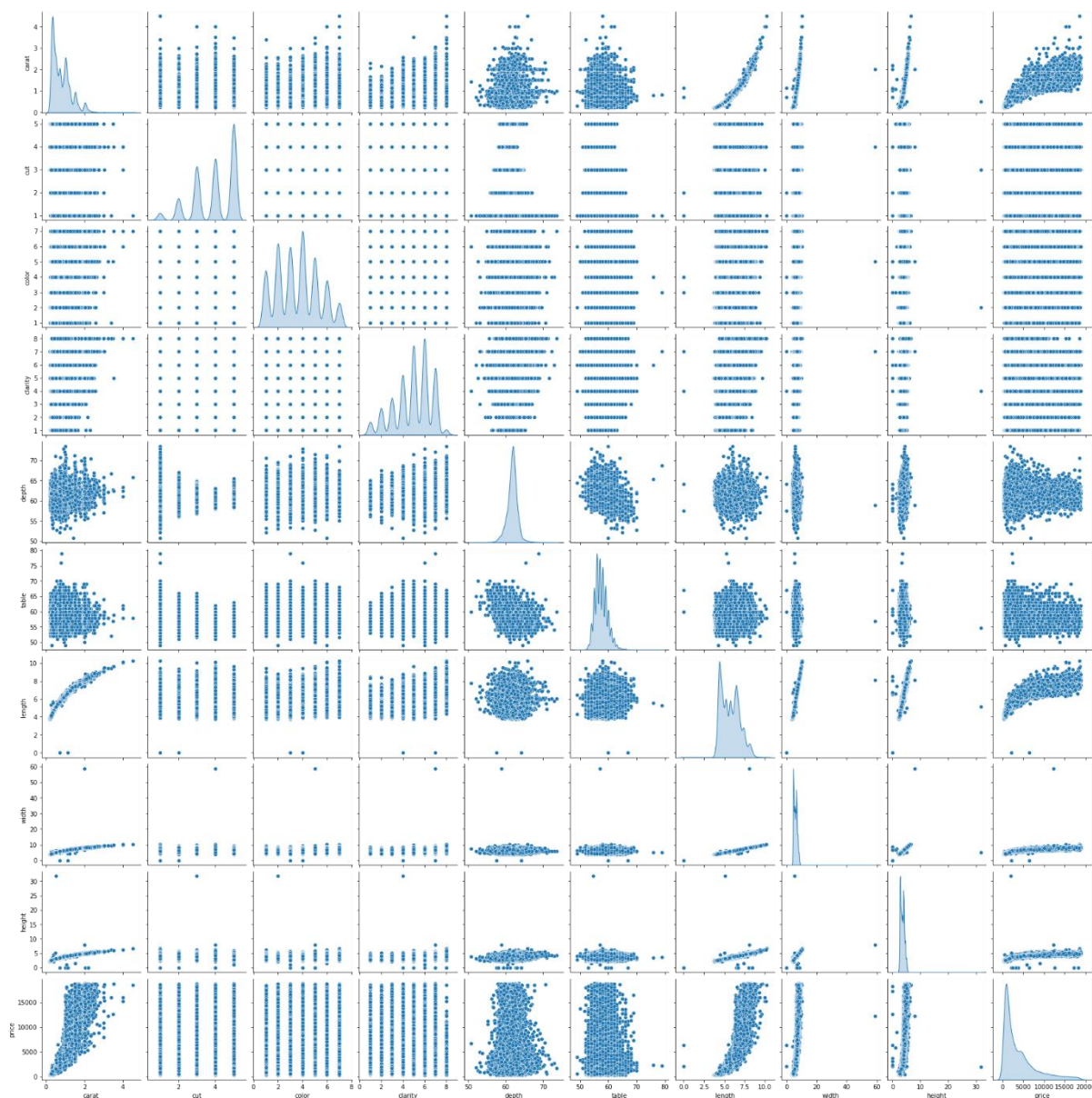


Figure XIV - Pair Plot

- The scatter between price vs. length, length, height, and carat seems following a pattern and they may have some relationship.
- Scatter between Price vs. cut, color, clarity, depth and table are very well spread.
- Depth vs. length has scatter not forming any pattern and data points are cluttered.

### Correlation Matrix (Heat Map):



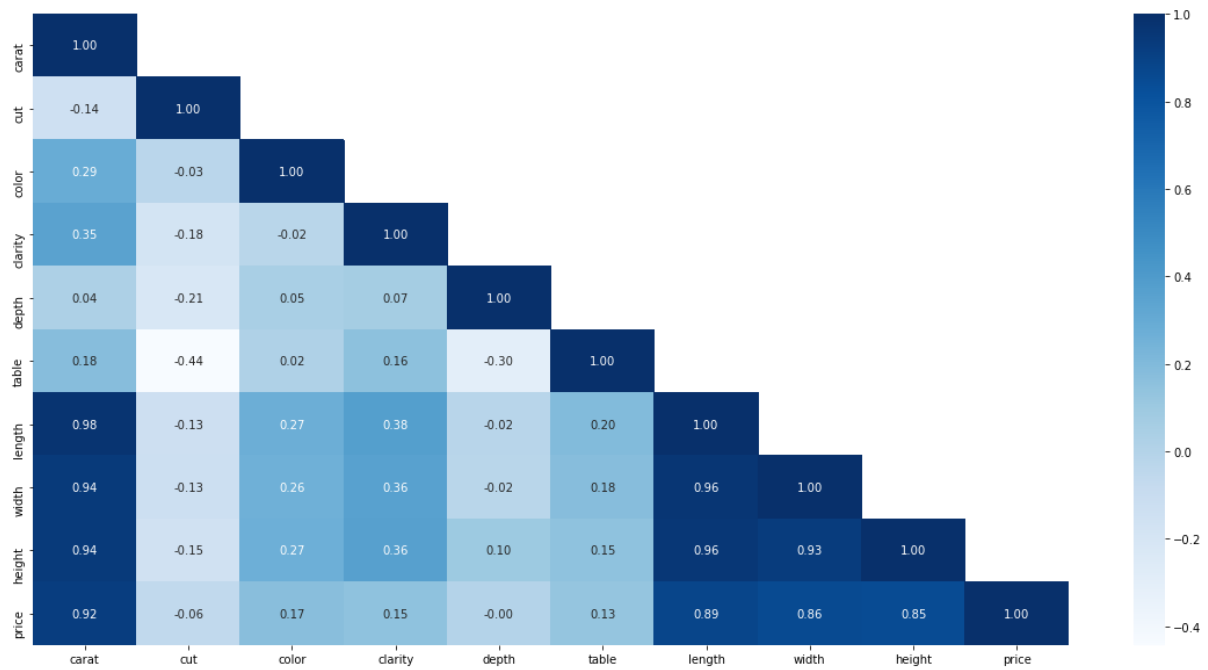


Figure XV - Correlation Matrix (Heat Map)

By above, we can infer below –

- High correlation
  - Length vs. Depth, Carat
  - Height vs. Length, Width, Carat
  - Price vs. Length, Width, Height, Carat
- Low correlation
  - Price vs. Cut, Color, Clarity, table

**Outlier Analysis:**

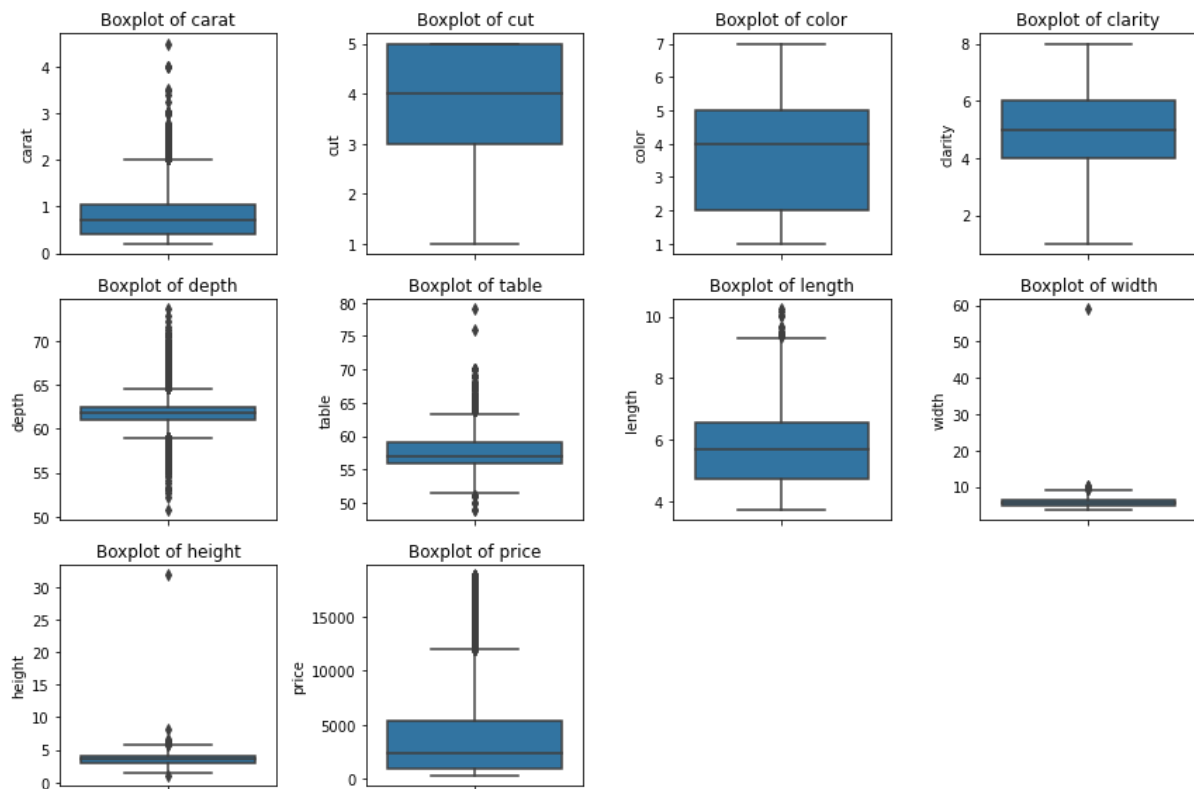


Figure XVI - Outliers Analysis in Data Frame

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 an ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning.

We have 697 missing values in depth data, and since that is continuous/numeric in nature we can impute these missing values with **median**.

There should be a valid value in depth since every stone/cubic zirconia would have some depth... so we cannot drop the missing values.

Also, there are only 2.5% of data missing in there, we are good to impute values if missing proportion is less.

## Zeros Count in Features

Feature	Zero Counts
carat	0
cut	0
color	0
clarity	0
depth	0
table	0
length	2
width	2
height	8
price	0

Table 11 - Zeros Count

We have zeros in length, width and height fields, and these must have a value as cubic zirconia would have >0 value to represent these fields.

We will consider these zeros as they are missing from data frame and impute the values (**with median**) to prevent false prediction from our model.

## Post update data frame summary –

	count	mean	std	min	25%	50%	75%	max
carat	26933.0	0.798010	0.477237	0.20	0.40	0.70	1.05	4.50
cut	26933.0	3.909702	1.113165	1.00	3.00	4.00	5.00	5.00
color	26933.0	3.605206	1.705883	1.00	2.00	4.00	5.00	7.00
clarity	26933.0	4.946423	1.646749	1.00	4.00	5.00	6.00	8.00
depth	26933.0	61.745285	1.393848	50.80	61.10	61.80	62.50	73.60
table	26933.0	57.455950	2.232156	49.00	56.00	57.00	59.00	79.00
length	26933.0	5.729769	1.126285	3.73	4.71	5.69	6.55	10.23
width	26933.0	5.733525	1.163989	3.71	4.72	5.70	6.54	58.90
height	26933.0	3.538815	0.717377	1.07	2.90	3.52	4.04	31.80
price	26933.0	3937.526120	4022.551862	326.00	945.00	2375.00	5356.00	18818.00

Figure XVII - Summary after Missing/Zero Value Impute

## Ordinal values

- **Cut:** Fair, Good, Very Good, Premium, Ideal
  - o Cuts in cubic zirconia having 5 sub-levels and they seem good breakdown, we don't need to combine them.

- **Color:** D, E, F, G, H, I, J
  - Since we don't have enough information about how color levels are coded, we won't combine these as well.
- **Clarity:** IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1
  - Cubic zirconia's clarity chart goes like below –  
And we can see that sub-level VVS1-VVS2, VS1-VS2 and SI1-SI2 nearly represent the same property of zirconia, hence they can be clubbed together.

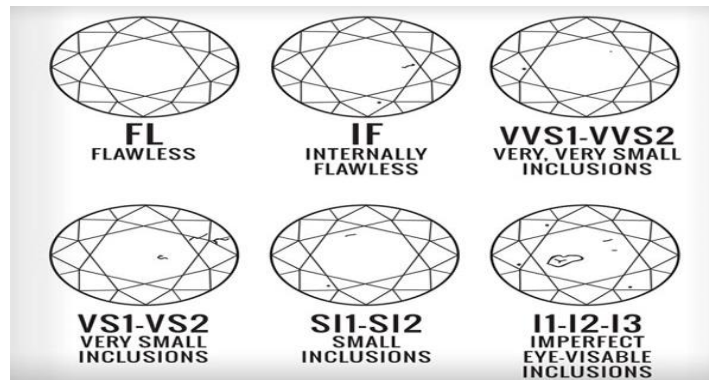


Figure XVIII - Clarity Chart

We will combine these sub-levels in new ordinal value –

IF: 1

VVS1-VVS2: 2

VS1-VS2: 3

SI1-SI2: 4

I1:5

**New proportion –**

Clarity Rank (L-H)	Value counts	% In Data
1	891	3.31%
2	4,369	16.22%
3	10,180	37.80%
4	11,129	41.32%
5	364	1.35%

Table 12 - New Clarity Level Proportion

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.

We have already encoded the data that was categorical in nature and no further transformation needed.

#### Encoding summary

- **Cut:** Fair:1, Good:2, Very Good:3, Premium:4, Ideal:5
- **Color:** D:1, E:2, F:3, G:4, H:5, I:6, J:7
- **Clarity:** IF:1, VVS:2, VS:3, SI:4, I1:5

#### Splitting Train vs. Test Data

We are splitting data in 70 – 30%, our train data will be 70% and test will be 30%.

\*\* Random State will be taken as 123 at all places

#### Scikit Learn's Linear Regression

With built model on training data, we have coefficient and intercept of LR model as below –

##### Intercept –

Intercept of our model is: **11577.556**

##### Coefficients –

Feature	Coefficients
carat	10917.37325
cut	108.136872
color	-314.192338
clarity	-899.728238
depth	-85.555598
table	-35.181439
length	-939.946983
width	7.213324
height	-29.033789

Table 13 - Feature Coefficients (Scikit Learn LR Model)

The intercept & coefficients are large due to price points prediction with feature ranges from 0-10 and 0-100.

### $R^2$ – Coefficient of determinant (Model Score)

The coefficient of determination,  $R^2$  is defined for multiple linear regression similarly, as it is defined for SLR.

$$R^2 = 1 - SSE/SST$$

**On Train Data** – 0.9046

**On Test Data** – 0.9061

The model is not under or overfitted, as it gives us good results on both training and testing data sets.

### Regularizing (Ridge & Lasso) with alpha 0.2

Metrics	Features	Ridge	Lasso
Intercept		11,540.14	11,489.98
Coefficient	carat	10,904.28	10,885.38
	cut	108.16	108.08
	color	-314.11	-313.95
	clarity	-899.86	-899.89
	depth	-85.30	-85.01
	table	-35.16	-35.13
	length	-934.51	-925.23
	width	7.24	4.46
	height	-29.09	-26.80

Table 14 - Regularized Coeff. & Intercept

### $R^2$ Scores –

$R^2$	LR	Ridge	Lasso
Train Data	0.90458	0.90458	0.90458
Test Data	0.90609	0.90609	0.90607

Table 15 - Models'  $R^2$

There is no significant improvement in models' prediction or  $R^2$  from regularization.

### Root mean squared error –

RMSE is a standard way to measure the error of a model in predicting quantitative data, and it is calculated as follows –

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (Predicted_i - Actual_i)^2}{N}}$$

RMSE on Train data – 1245.039

RMSE on Test data – 1226.785

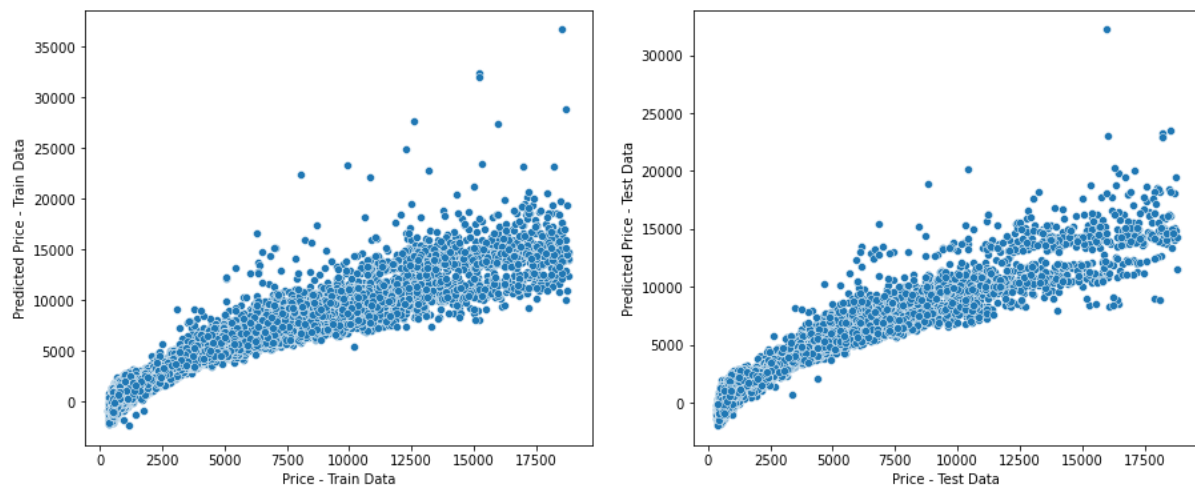


Figure XIX - Actual vs. Predicted values

The scatter plot shows that some predictions are very far from actuals. It must be caused by presence of outliers and data points of different scale.

### Outlier treatment & Scaling

#### Outlier analysis (After treatment) –

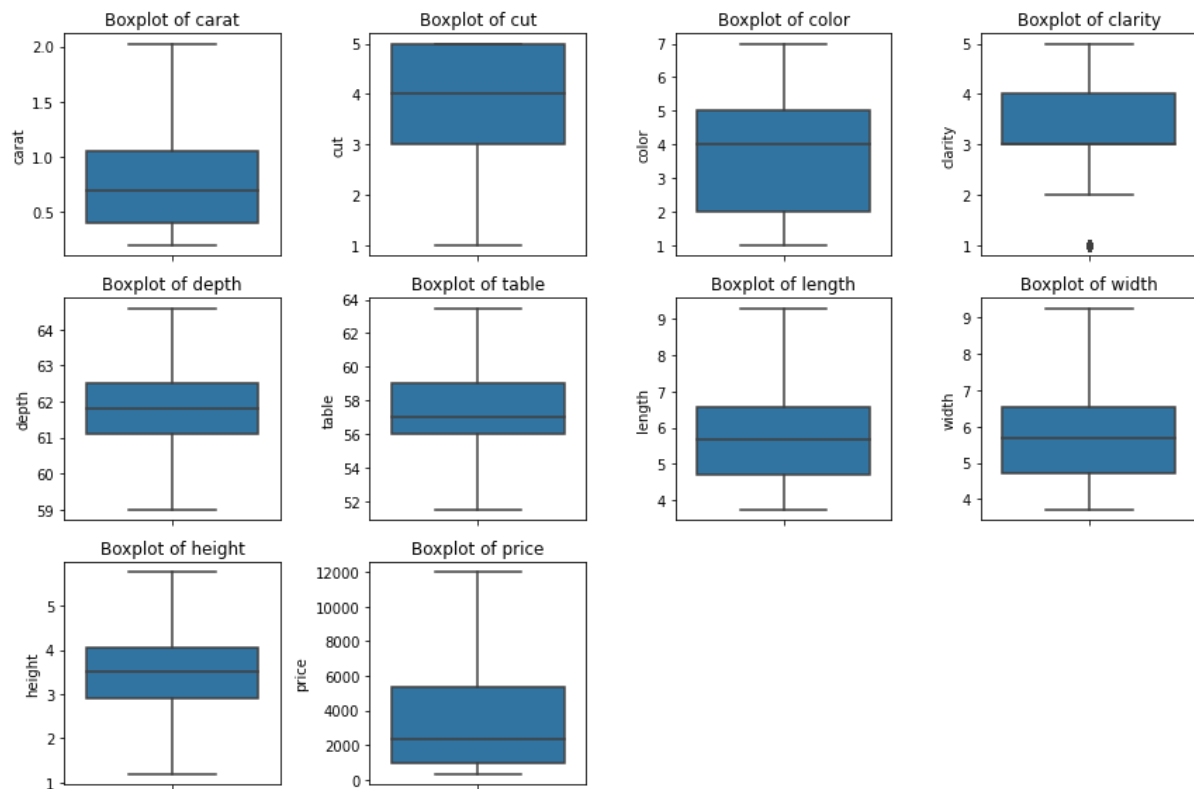


Figure XX - Outlier Analysis (Post treatment)

We can ignore cut, color and clarity plots, because they are categorical variables. We see no outlier in any other plot.

#### After scaling Summary –

	count	mean	std	min	25%	50%	75%	max
carat	26933.0	-2.106922e-16	1.000019	-1.283866	-0.851076	-0.201891	0.555491	2.665341
cut	26933.0	1.225747e-16	1.000019	-2.613949	-0.817237	0.081120	0.979476	0.979476
color	26933.0	2.957654e-16	1.000019	-1.527217	-0.940999	0.231435	0.817653	1.990088
clarity	26933.0	-5.163961e-16	1.000019	-2.612272	-0.250212	-0.250212	0.930818	2.111848
depth	26933.0	1.420076e-16	1.000019	-2.256126	-0.532668	0.041817	0.616303	2.339761
table	26933.0	-9.184186e-16	1.000019	-2.751657	-0.665503	-0.201914	0.725266	2.811420
length	26933.0	-2.094143e-16	1.000019	-1.776344	-0.905762	-0.035179	0.728801	3.180645
width	26933.0	-4.864403e-16	1.000019	-1.808742	-0.905054	-0.028208	0.723375	3.166017
height	26933.0	8.447143e-17	1.000019	-3.375130	-0.916695	-0.025332	0.722262	3.180698
price	26933.0	-1.777890e-17	1.000019	-0.983187	-0.804705	-0.392381	0.467157	2.374950

Table 16 - After Scaling Data Summary

Post outlier treatment and scaling, we built linear model again on that data –

Intercept – 0.00024

Coefficients -



Features	Coefficients
carat	1.18
cut	0.03
color	-0.13
clarity	-0.19
depth	0.01
table	-0.01
length	-0.47
width	0.51
height	-0.18

Table 17 - Coefficients on Scaled Data LR Model

$R^2$  – Coefficient of determinant (Model Score) –

Training data – - 0.95

Testing data – - 0.96

The  $R^2$  has been negative, which comes only where trend was not being followed, hence we can say our model was better without outlier treatment and scaling.

### VIF (Variance Inflation Factor)

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. ...

This ratio is calculated for each independent variable. A high VIF indicates that the associated independent variable is highly collinear with the other variables in the model.

Features	VIF
carat	121.48
cut	17.19
color	6.09
clarity	18.63
depth	1,173.51
table	849.89
length	10,628.17
width	9,367.49
height	3,061.07

Table 18 - Variance Inflation Factor

We have VIF very high on length, width, height and depth, and by pair plot we inferred that these are very much correlated. For prediction from our model, we can exclude these features.

## StatsModel Linear Regression

$R^2$  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  $R^2$

Scikit does not provide a facility for adjusted  $R^2$ , so we use statsmodel, a library that gives results similar to what you obtain in R language.

We StatsModel Linear regression, we get below summary –

Metrics	Features	Value
Intercept		11,577.56
Coefficient	carat	10,917.37
	cut	108.14
	color	-314.19
	clarity	-899.73
	depth	-85.56
	table	-35.18
	length	-939.95
	width	7.21
	height	-29.03

Table 19 - Intercept & Coeff. from StatsModel LR

OLS Regression Results						
=====						
Dep. Variable:	price	R-squared:	0.905			
Model:	OLS	Adj. R-squared:	0.905			
Method:	Least Squares	F-statistic:	1.985e+04			
Date:	Sat, 19 Feb 2022	Prob (F-statistic):	0.00			
Time:	17:46:18	Log-Likelihood:	-1.6112e+05			
No. Observations:	18853	AIC:	3.223e+05			
Df Residuals:	18843	BIC:	3.223e+05			
Df Model:	9					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]
-----						
Intercept	1.158e+04	733.791	15.778	0.000	1.01e+04	1.3e+04
carat	1.092e+04	94.897	115.044	0.000	1.07e+04	1.11e+04
cut	108.1369	10.015	10.797	0.000	88.506	127.767
color	-314.1923	5.607	-56.036	0.000	-325.183	-303.202
clarity	-899.7282	11.761	-76.502	0.000	-922.781	-876.676
depth	-85.5556	8.135	-10.517	0.000	-101.501	-69.610
table	-35.1814	5.141	-6.843	0.000	-45.259	-25.104
length	-939.9470	51.889	-18.114	0.000	-1041.655	-838.239
width	7.2133	24.447	0.295	0.768	-40.704	55.131
height	-29.0338	42.762	-0.679	0.497	-112.852	54.784
=====						
Omnibus:	4199.405	Durbin-Watson:	1.987			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	196213.573			
Skew:	-0.133	Prob(JB):	0.00			
Kurtosis:	18.802	Cond. No.	6.89e+03			
=====						
Notes:						
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.						
[2] The condition number is large, 6.89e+03. This might indicate that there are strong multicollinearity or other numerical problems.						

Figure XXI - OLS Regression Summary

The  $R^2$  has been same as Adjusted  $R^2$ , thus we select this liner model for our prediction. The score 90.5% is good.

From summary, we can see that coefficient against width and height are not statically significant at level ( $\alpha$ ) = 0.05, hence those can be eliminated from our model.

RMSE on Train data – 1245.039

RMSE on Test data – 1223.645

RMSE has been reduced on test data with very low margin from scikit learn LR model. We are good to consider any model as suitable.

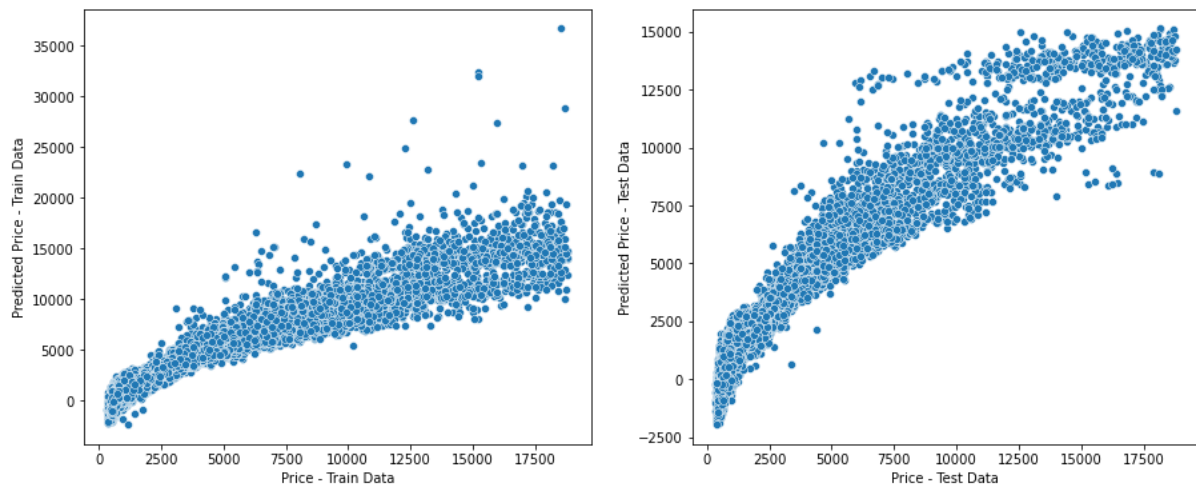


Figure XXII - Predicted price vs Actual (StatsModel Regression Model)

### Model Selection –

Accuracy on both Scikit Learn and StatsModel's Linear regression are ~90%, but plot against Predicted vs. Actual data seem more reliable on scikit learn's linear model as it does not give us negative pricing on test data.

However, RMSE on test data is slightly better on StatsModel's LR model, and we are selecting this having adjusted r-square 90.5%.

We have already concluded that scaling and treatment of outliers are not helping on model's accuracy hence we can identify 5 important independent predictors that are driving the prices.

carat	10,917.37
cut	108.14
color	-314.19
clarity	-899.73
depth	-85.56
table	-35.18
length	-939.95
width	7.21
height	-29.03

Table 20 - Feature Importance

Price of cubic zirconia was driven by Carat, Length, Clarity, Color and Cut significantly.

### FINAL Linear Equation –

$$Y (\text{Price}) = 10917.37 \cdot \text{carat} + 108.13 \cdot \text{cut} - 314.19 \cdot \text{color} - 899.72 \cdot \text{clarity} - 85.55 \cdot \text{depth} - 35.18 \cdot \text{table} - 939.94 \cdot \text{length} + 7.21 \cdot \text{width} - 29.03 \cdot \text{height} + 11577.56$$

1.4 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.

#### Inference –

There is spread between actual vs. predicted profit, which could be noise due to presence of outliers.

We have multicollinearity in data, which can be seen by higher VIFs calculated in Table 18 - Variance Inflation Factor.

With StatsModel OLS method, we have seen that  $R^2$  and Adjusted  $R^2$  are same, and P-value is way lesser than significance level (0.05)

#### Insights –

- In EDA, we have seen 34 duplicate entries and removed them to eliminate biased data
- There were 697 missing records in depth, we have imputed them with median.
- We had 10 dimensionless values in length, width and height, we could have removed or impute them... These were also imputed with median.
- Price of cubic zirconia is heavily influenced by carat, cut, color, clarity and length.
- We had outliers in the data but treating them don't help us getting better accuracy/prediction from model.
- Calculated  $R^2$  and adjusted  $R^2$ , but they don't significantly differ, which concludes that our model from scikit learn can be used.
- With correlation plot (Heat map) we have inferred that price, length, width, height and carat are highly correlated with each other, on the other hand cut, color, clarity has low correlation with others.

#### Recommendation –

- Gemstone company should focus on carat, cut, color, and clarity as they are the driving factors for price
- VVS group of zirconia (clarity VVS1 and VVS2) has high price, company can offer discount on them to increase sale and be profitable.
- The most important feature that drive price is carat, Company can build pricy stones by building stones with high carat.
- Most cubic zirconia is having premium cuts, company can target cutting all stones in premium style to have a higher price.
- The price can be further analysed and fixed to provide us a better differentiator.

- With important features identification company can distinguish higher profitable stones and lower profitable stones so as to have better profit share.

## Logistic Regression and LDA – Travel Agency

### Executive Summary

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.

### Introduction

Purpose of our exercise would be to identify important factors for predicting if employee would opt for package.

### Data Description

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 21 - Data Dictionary (Travel Agency)*

**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.

**Head & Tail of the Data:**

	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 22 - Head Data (5 Rows)

	Unnamed: 0	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
867	868	no	40030	24	4	2	1	yes
868	869	yes	32137	48	8	0	0	yes
869	870	no	25178	24	6	2	0	yes
870	871	yes	55958	41	10	0	1	yes
871	872	no	74659	51	10	0	0	yes

Table 23 - Tail Data (5 Rows)

### Data Information:

Column	Non-Null Items	Dtype
unnamed: 0	872	int64
Holliday_Package	872	object
Salary	872	int64
age	872	int64
educ	872	int64
no_young_children	872	int64
no_older_children	872	int64
foreign	872	object

Table 24 - Data Information (Agency Data)

- We have data with 872 records and 8 features
- But “Unnamed: 0” can be dropped from dataframe
- We don’t have any null/blank values
- We have 2 categorical variables rest are numeric
- We don’t have any duplicated 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 25 - Data Summary

- Average salary of employees is 47,729.17 with deviation of 23,418.66. The minimum salary of an employee is 1322 and maximum is 236,961
- Minimum age of employee is 20 and maximum is 62, and average age of employee is ~40
- Average years of formal education is 9 years, an employee received 21 years of formal education
- One of 872 employee has 3 children younger than 7 years and one employee has 6 children older than 7 years.
- 471 employees were not opted for tour packages
- We have 656 employee that are not foreigner

### Outlier Proportion:

Feature	Outliers	Outlier Proportion
Salary	57	6.54%
age	0	0.00%
educ	4	0.46%
no_young_children	207	23.74%
no_older_children	2	0.23%

Table 26 - Outlier Proportion

### Skewness & Kurtosis:

Feature	Skew	Kurtosis
Salary	3.098	15.755
age	0.146	-0.912
educ	-0.045	-0.001
no_young_children	1.943	3.085
no_older_children	0.952	0.665

Table 27 - Skew &amp; Kurtosis Values

### Univariate Analysis:

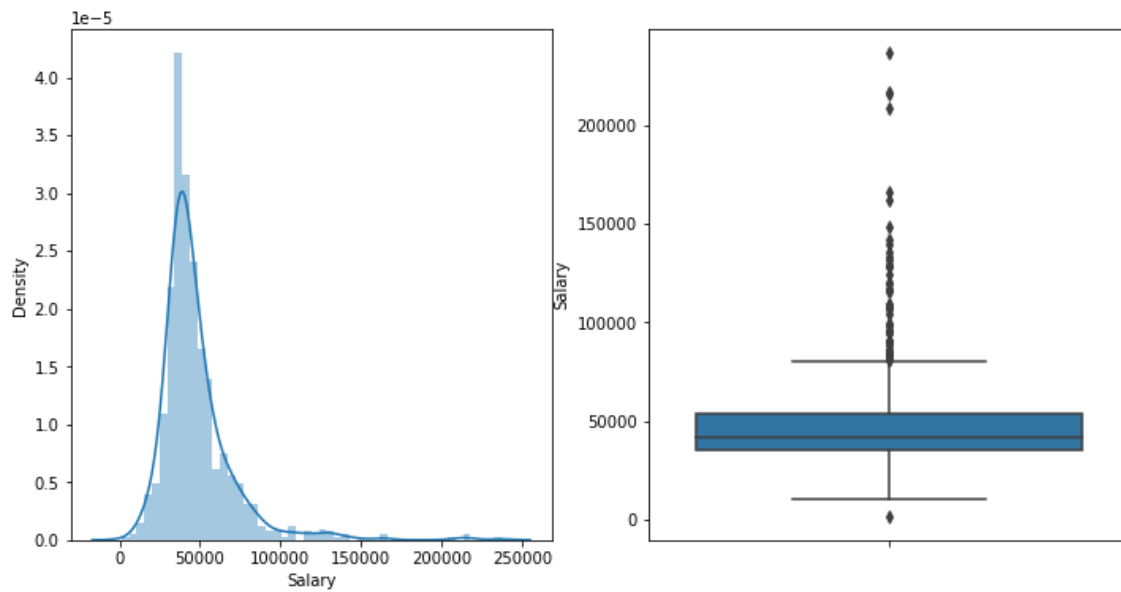


Figure XXIII – Univariate Analysis (Salary)

**Salary:** Data has 6.54% outliers, distribution seem normally distributed and positively skewed, having range from 1.3k – 236k+.

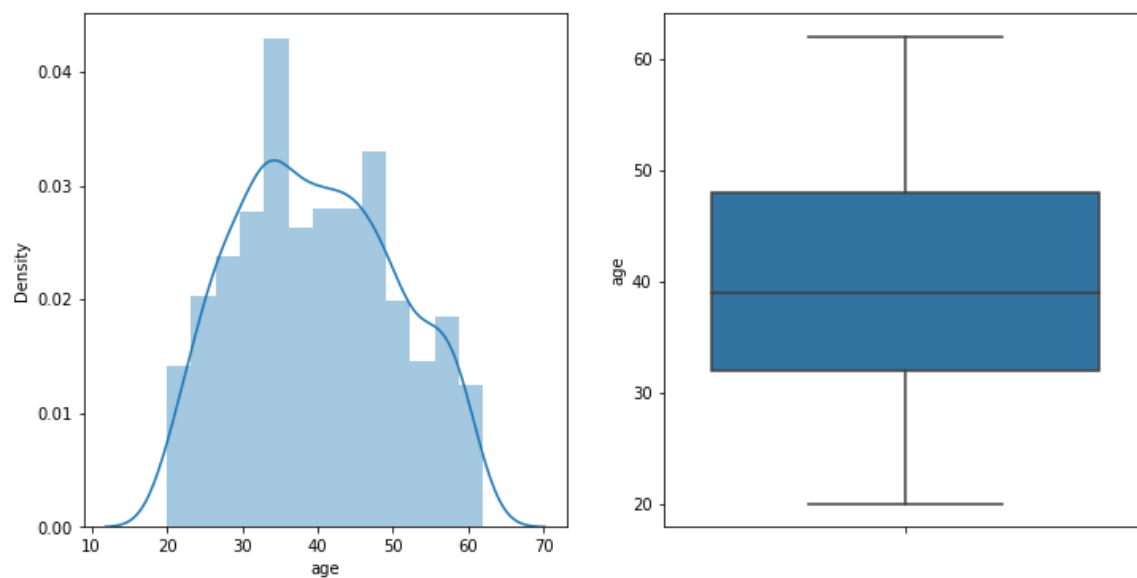


Figure XXIV - Univariate Analysis (Age)

**Age:** Data ranges from 20 to 62, and don't have any outlier. All the values are valid values.

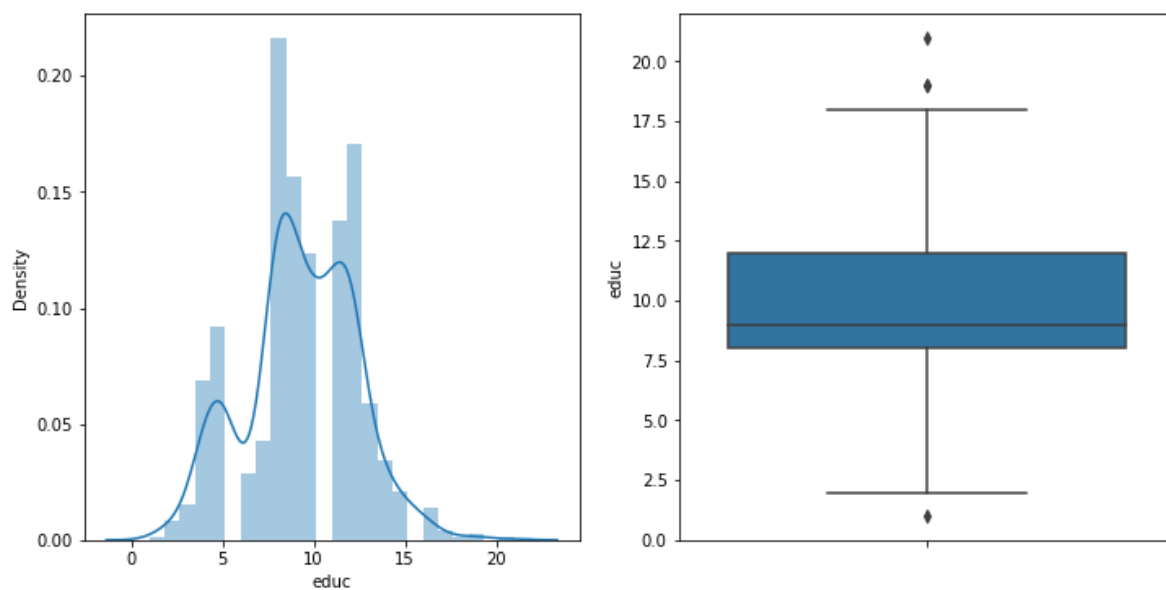


Figure XXV - Univariate Analysis (Education)

**Education:** Formal education years for most of the employees are between 5-10. There are 4 outliers (one having minimum formal education rest 3 breach the threshold)

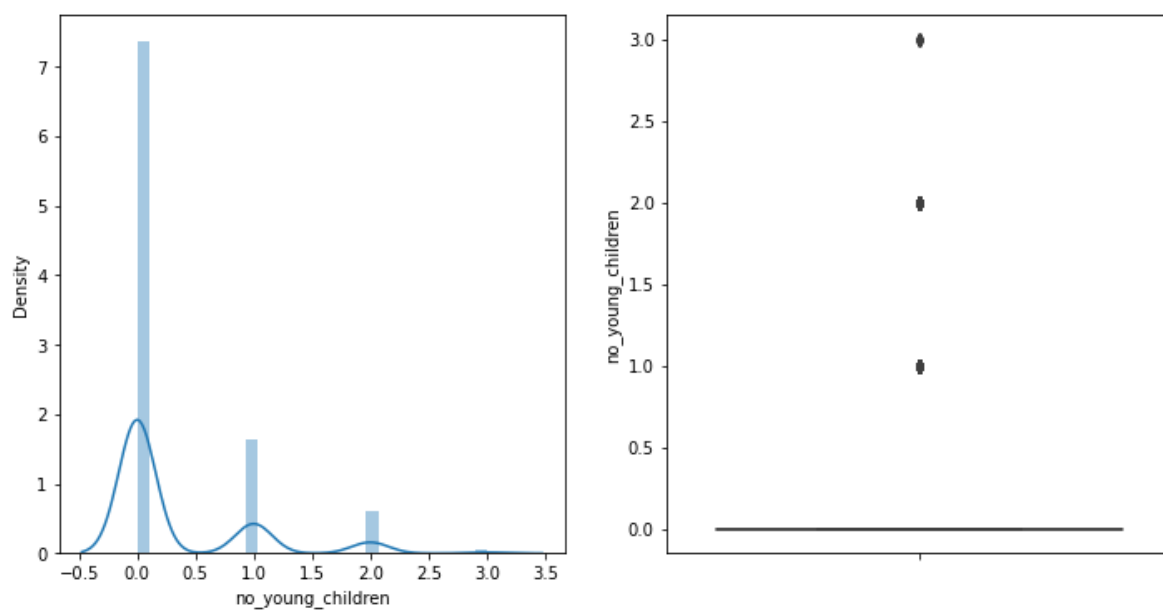


Figure XXVI - Univariate Analysis (No. Young Children)

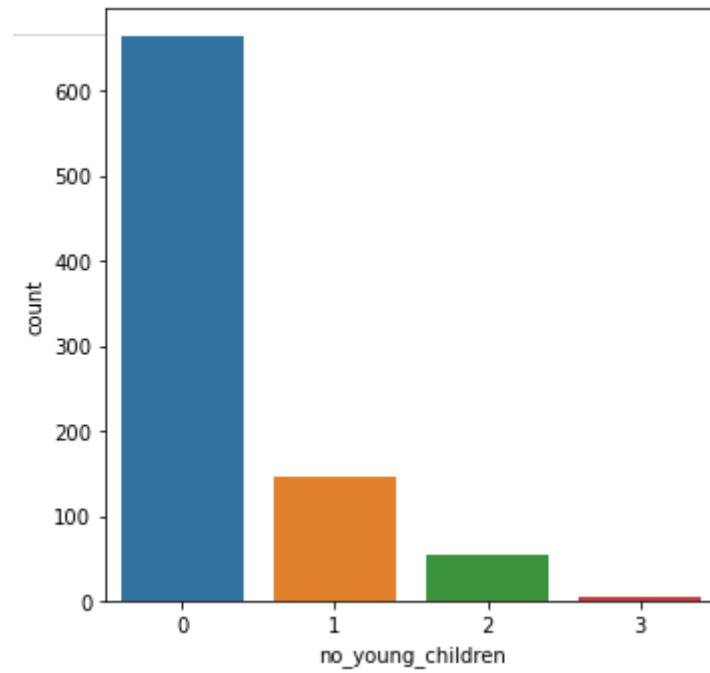


Figure XXVII - Count Plot (# of young Children)

**Number Young Children:** The data seem discrete having values from 0 – 3. Most of the employees don't have younger (< 7 years) child.

Value	Frequency
0	665
1	147
2	55
3	5

Table 28 - Frequency (# Of Young Children)

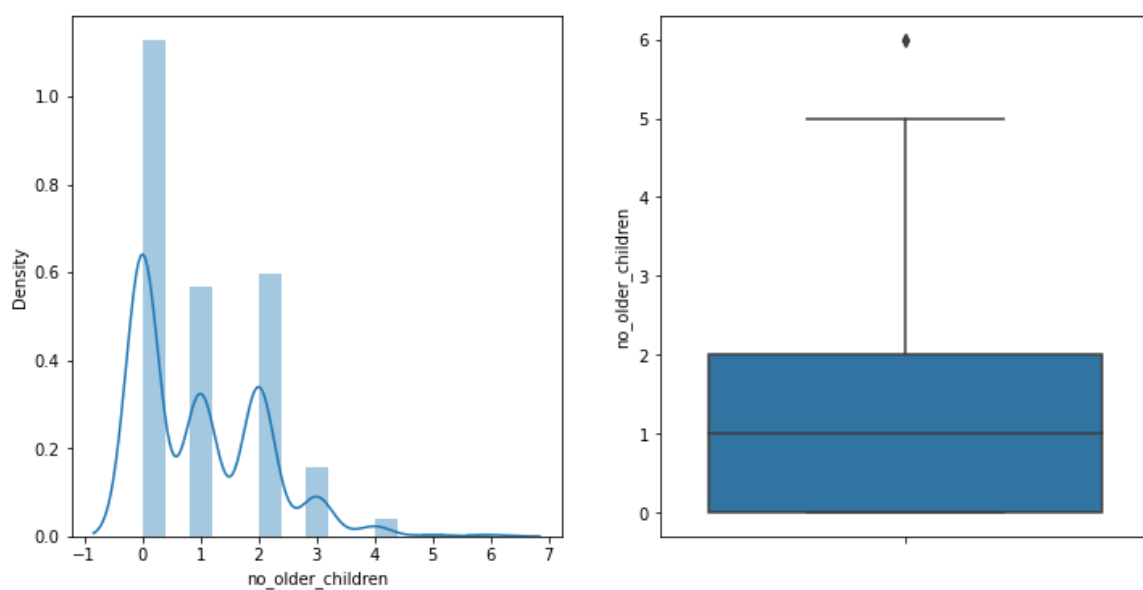


Figure XXVIII - Univariate Analysis (No. Older Children)

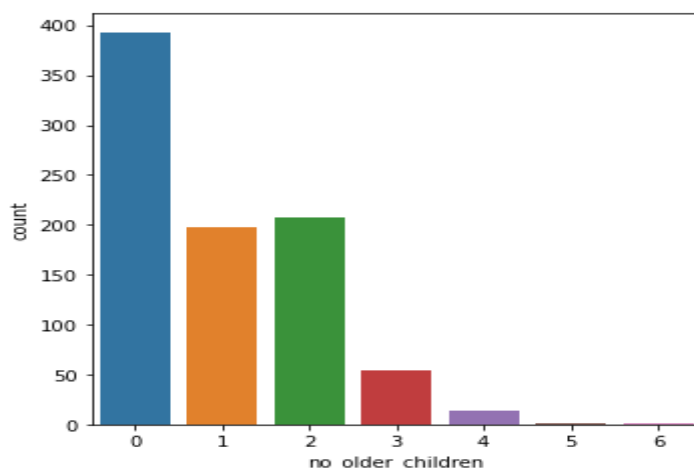


Figure XXIX - Count Plot (# Of Older Children)

Value	Frequency
0	393
2	208
1	198
3	55
4	14
5	2
6	2

Table 29 - Frequency (# Of Older Children)

**Number Older Children:** Most of the employee don't have older child (>7 years) and one of the employees has 6 children older than 7 years.

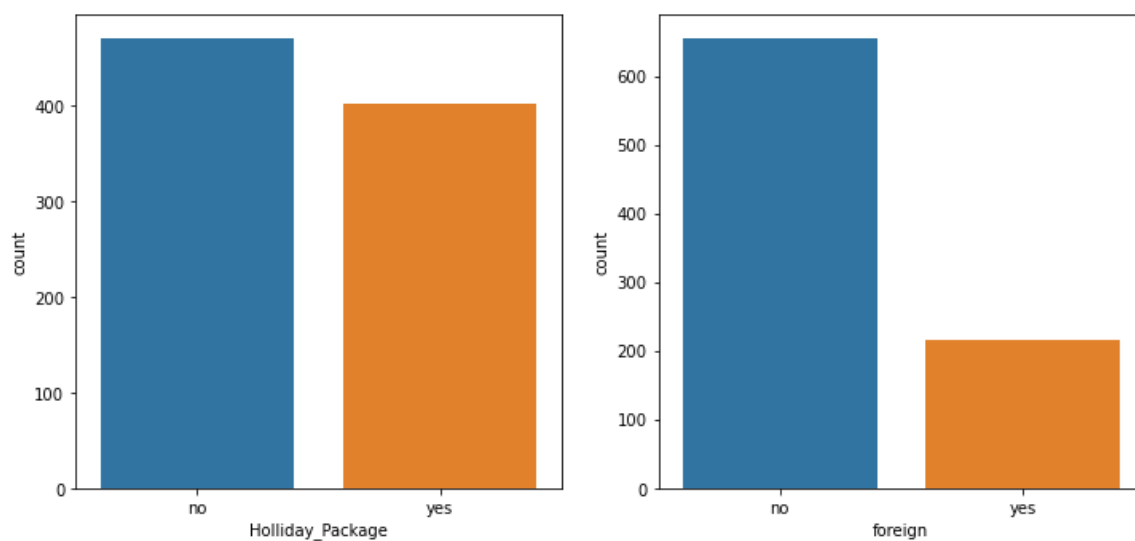


Figure XXX - Univariate Analysis (Holiday Package & Foreign)

Counts	Yes	No
Holiday Package	401	471
Foreign	216	656

Table 30 - Frequency (Holiday Package & Foreign)

**Holiday Package:** Number of employees opted and not opted are 50-50 in nature, the dataset given is not bias to any specific category.

**Foreign:** We have 216 foreigners in the data which is almost 24% of total population.

### Multivariate Analysis:

#### Effect of categorical variables over continuous predictors –

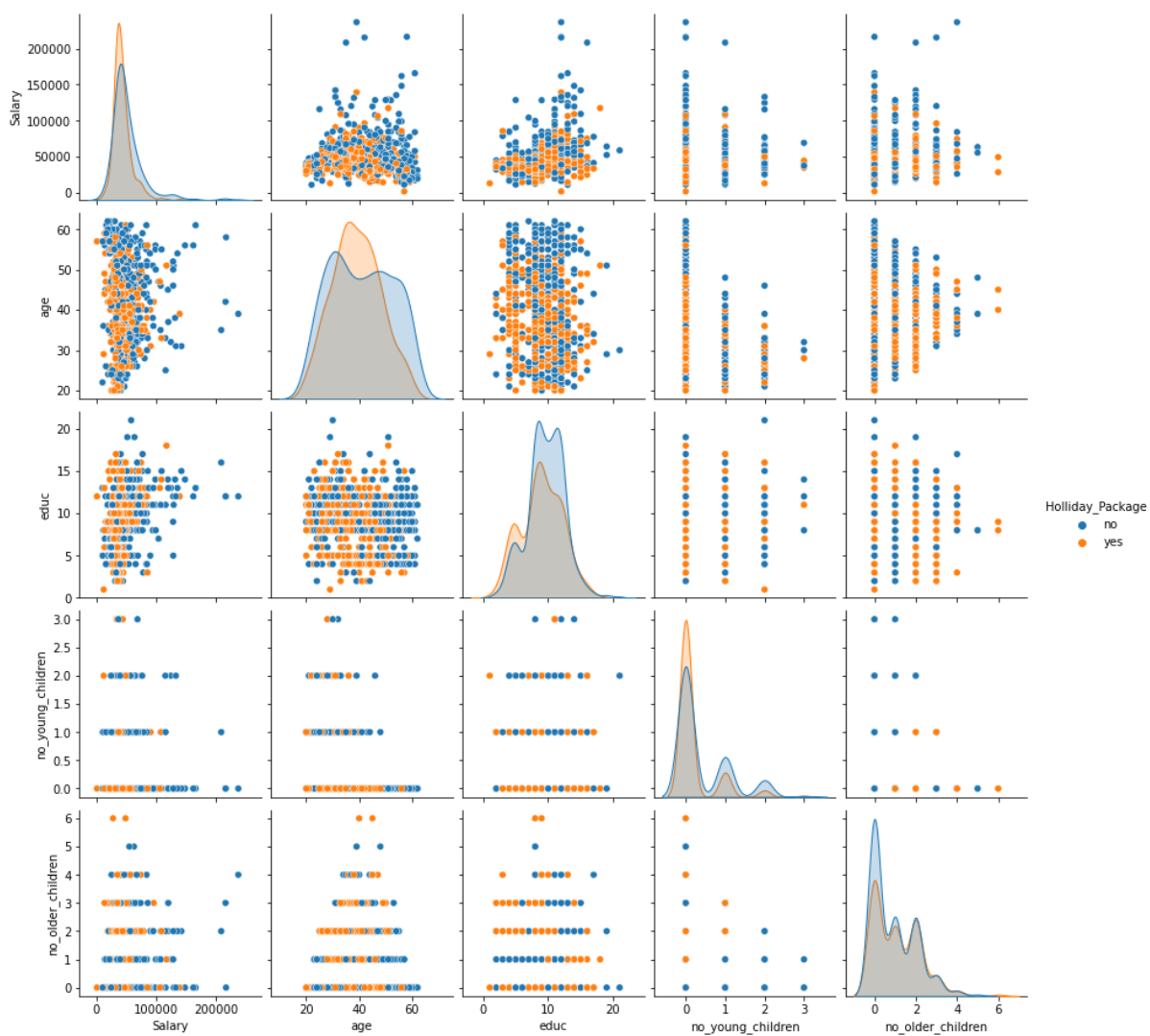


Figure XXXI - Pair Plot (Hue - Holiday Package)

We don't see any pattern of employee opting and not opting the package based on above plot except for young children plot, all distributions seem to follow the same pattern either package was opted or not.

But we do see a pattern that employee who doesn't have any younger children are opting for holiday package more.

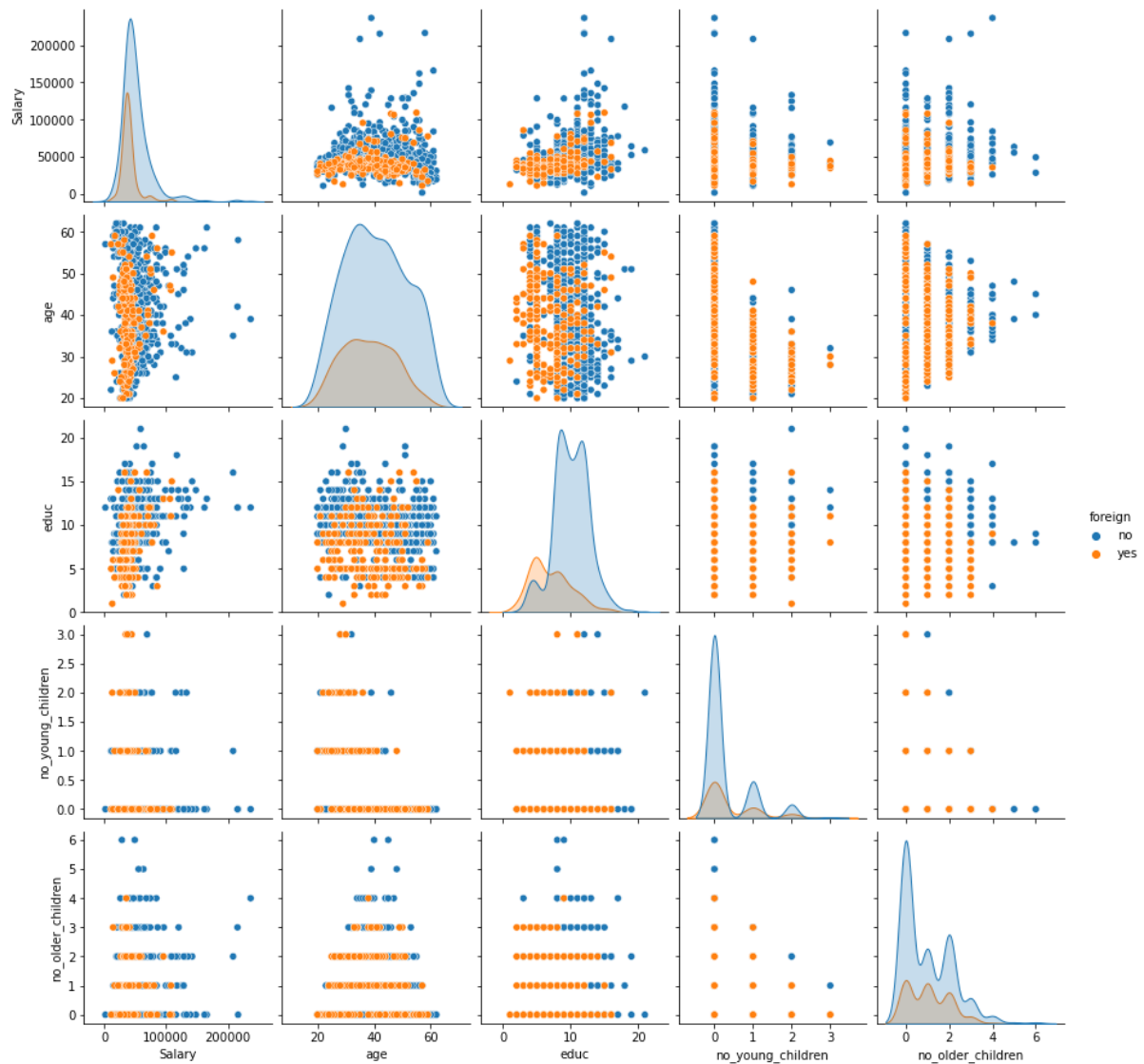


Figure XXXII - Pair Plot (Hue - Foreign)

There are couple of observations can be derived from above –

- Foreigners comparatively have less salary than non-foreigners
- They do not spend more years in formal education

Holiday Package vs Foreigner –

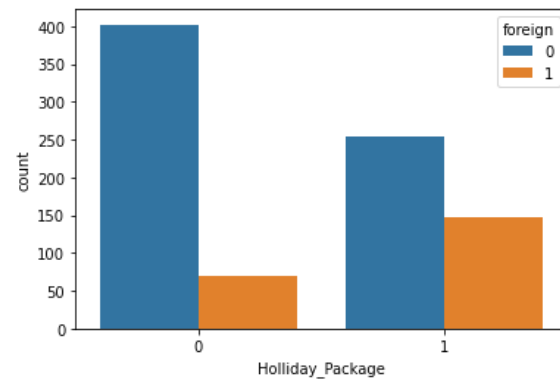


Figure XXXIII - Holiday Package & Foreign Employee

- Foreigner employees mostly opts for holiday packages.

### Correlation Matrix (Heat Map):

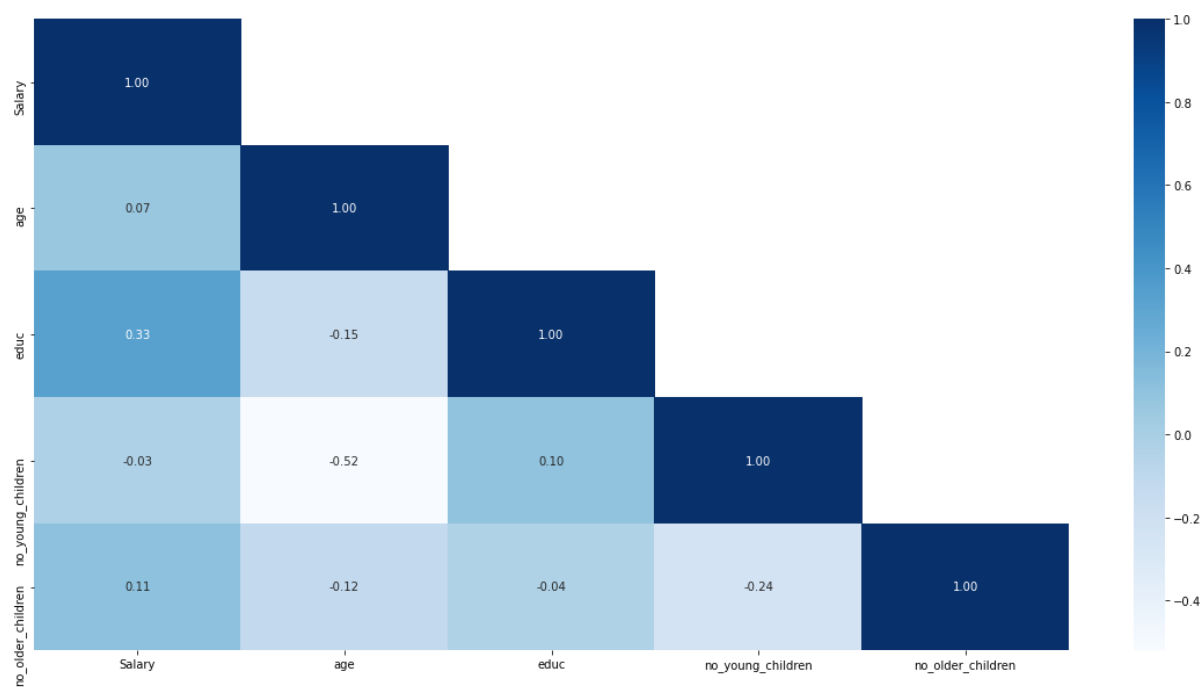


Figure XXXIV - Correlation Matrix (Heat Map)

None of the two predictors are correlated. We don't see any significant correlation ( $>0.70$  or  $<-0.70$ )

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).

### Data Encoding



We have 2 categorical predictors Holiday package and foreign, we have converted them both in 0 – 1 value. If their value field was 'yes', we have converted it to 1 else 0.

### Target Variable Proportion

Opted for Holiday package – 401 (45.98%)

Not Opted for Holiday package – 471 (54.01%)

### Splitting Train vs. Test Data

We are splitting data in 70 – 30%, our train data will be 70% and test will be 30%.

\*\* Random State will be taken as 123 at all places

### Logistic Regression Model -

#### With Outlier Accuracy –

Training data – 54.75%

Testing data – 50%

The model score is very poor as we would have 50% accuracy without any model for 2 predictions.

#### Outlier Treatment –

The number of children (younger or older) are not continuous and have certain discrete values, we can ignore these data points and treat outliers seen on Salary or Education.

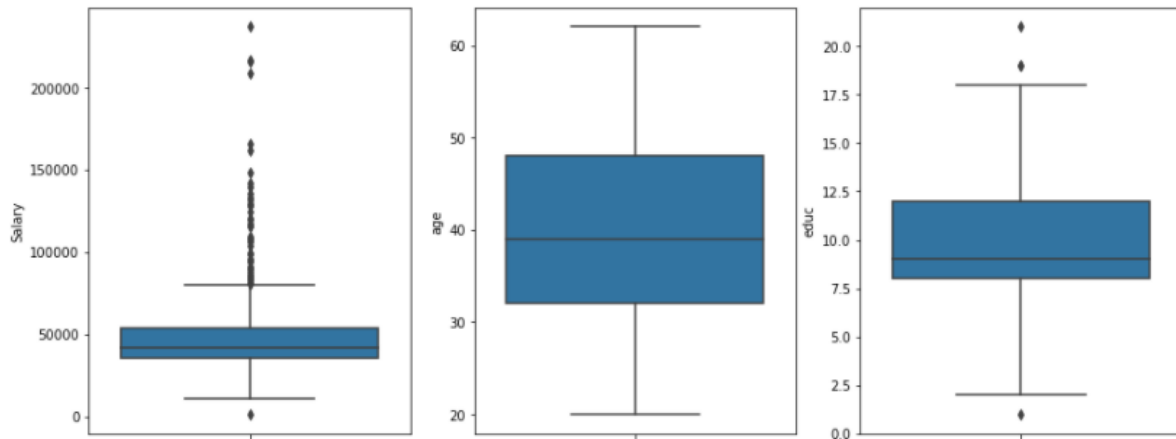


Figure XXXV - Before Outlier Treatment

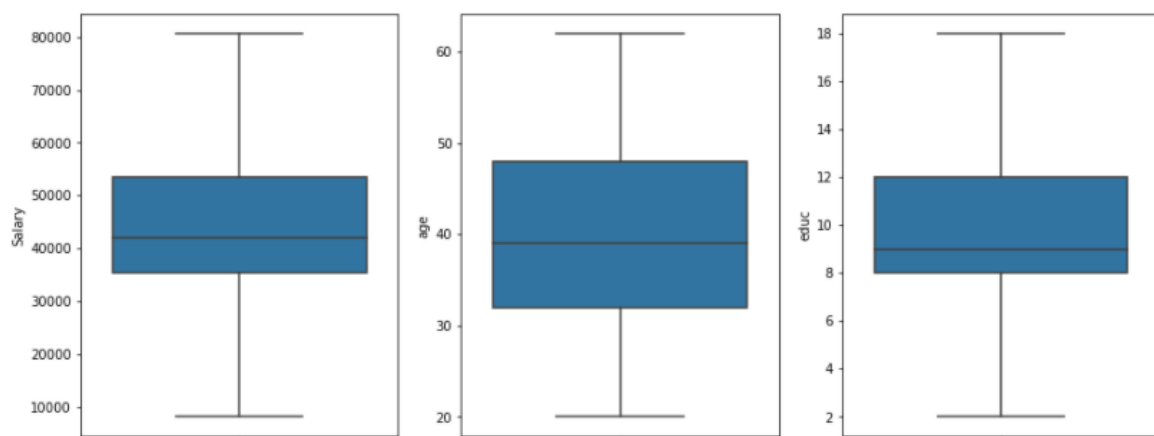


Figure XXXVI - After Outlier Treatment

#### Without Outlier Accuracy –

Training data – 54.26%

Testing data – 53.46%

The accuracy has not been improved by significant number; hence we can try hyper-parameter tuning to see if any parameter provides us better accuracy.

#### Hyper-parameter Passed to iterate –

Penalty - l2 or none

Solver - sag, lbfgs or newton-cg

Tol - 0.0001 or 0.00001

### Best Parameters –

Penalty - l2

Solver – sag

Tol - 0.0001

With best parameters as well, our accuracy stays the same...

	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.54	1.00	0.70	331	0	0.54	1.00	0.70	140
1	0.00	0.00	0.00	279	1	1.00	0.01	0.02	122
accuracy			0.54	610	accuracy			0.54	262
macro avg	0.27	0.50	0.35	610	macro avg	0.77	0.50	0.36	262
weighted avg	0.29	0.54	0.38	610	weighted avg	0.75	0.54	0.38	262
Train Data - Classification Report					Test Data - Classification Report				

Table 31 - Classification Report with 0.5 prob threshold (Logistic Regression)

As seen in above, the model has very poor precision, recall & F1-Score on both train and test data.

To furthermore making our model accurate, we can adjust probability threshold. Which gives us accuracy around –

0.45 was found as best possible threshold...

Training data – 59%

Testing data – 57%

	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.59	0.76	0.67	331	0	0.58	0.70	0.63	140
1	0.57	0.38	0.46	279	1	0.55	0.42	0.47	122
accuracy			0.59	610	accuracy			0.57	262
macro avg	0.58	0.57	0.56	610	macro avg	0.56	0.56	0.55	262
weighted avg	0.58	0.59	0.57	610	weighted avg	0.57	0.57	0.56	262
Train Data - Classification Report					Test Data - Classification Report				

Table 32 - Classification Report with 0.45 prob threshold (Logistic Regression)

### Intercept & Coefficients –

The model has intercept of  $7.28 \times 10^{-5}$  and feature coefficients as below –

Features	Coefficients
Salary	-0.000007

age	0.001410
educ	0.000388
no_young_children	-0.000193
no_older_children	0.000260
foreign	0.000166

Table 33 - Feature Coefficient (Logistic Regression)

Based on above, it seems we have Age as very important factor deciding if employee would opt of holiday package or not, followed by Education & No Older children.

### Linear Discriminant Analysis –

LDA gave us better accuracy here –

Train data – 66.56%

Test data – 68.32%

### Intercept & Coefficients –

Model Intercept: **2.43**

Coefficients –

Features	Coefficients
Salary	-0.000021
age	-0.049215
educ	0.038756
no_young_children	-1.236254
no_older_children	-0.016823
foreign	1.350785

Table 34 - Feature Coefficients (LDA)

Since LDA gives us better accuracy and F1-Score, we can derive insights from LDA model's coefficients that –

- Employee opting for holiday package is highly dependent on younger children he/she is having.
- Other driving factors are nationality (foreigner or not), Age and Education.

	precision	recall	f1-score	support
0	0.67	0.76	0.71	331
1	0.66	0.56	0.60	279
accuracy			0.67	610
macro avg	0.66	0.66	0.66	610
weighted avg	0.66	0.67	0.66	610

Table 35 - Classification Report (Train Data) LDA

	precision	recall	f1-score	support
0	0.67	0.81	0.73	140
1	0.71	0.54	0.61	122
accuracy			0.68	262
macro avg	0.69	0.67	0.67	262
weighted avg	0.69	0.68	0.68	262

Table 36 - Classification Report (Test Data) LDA

**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.

Confusion Matrix –

**Logistic Regression:**

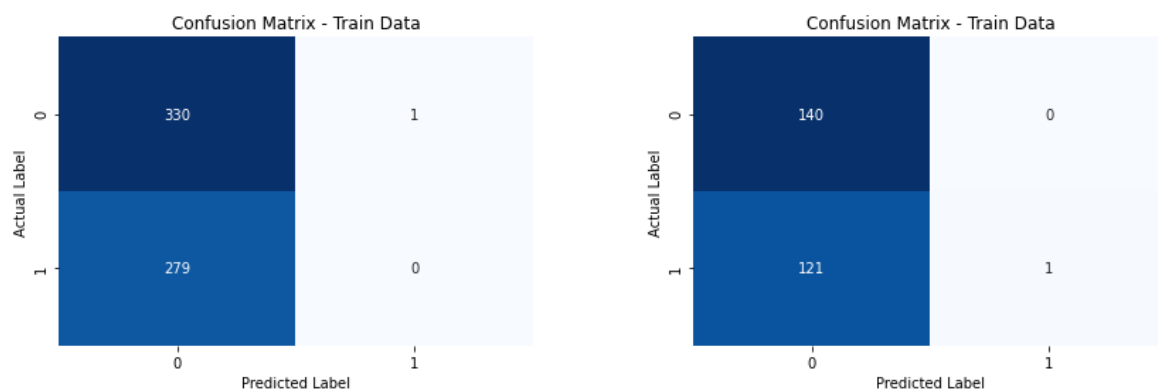


Figure XXXVII - Confusion Matrix (Logistic Regression)

- The confusion matrix shows us that model is not predictive positive labels.

### Linear Discriminant Analysis:

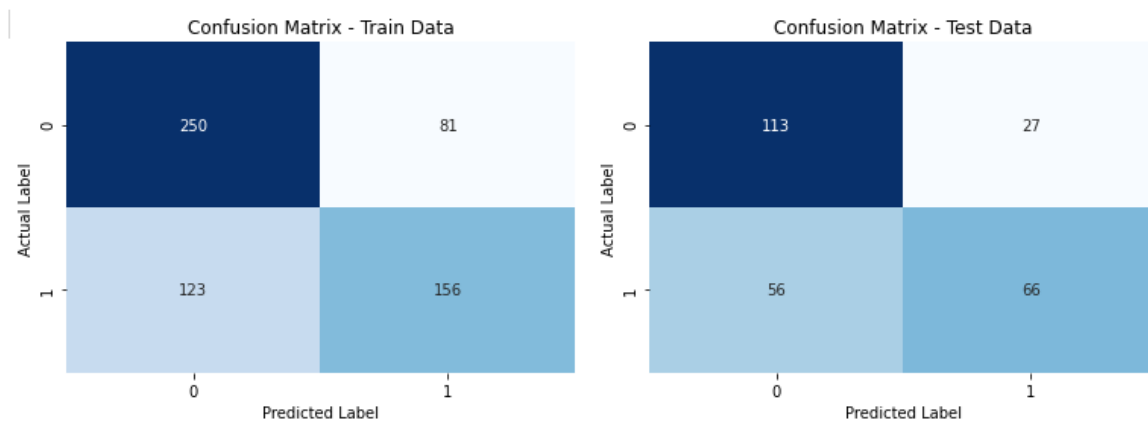


Figure XXXVIII - Confusion Matrix (LDA)

- LDA model predict both positive and negative labels at some extent with better accuracy.

### ROC Curve & ROC-AUC Score –

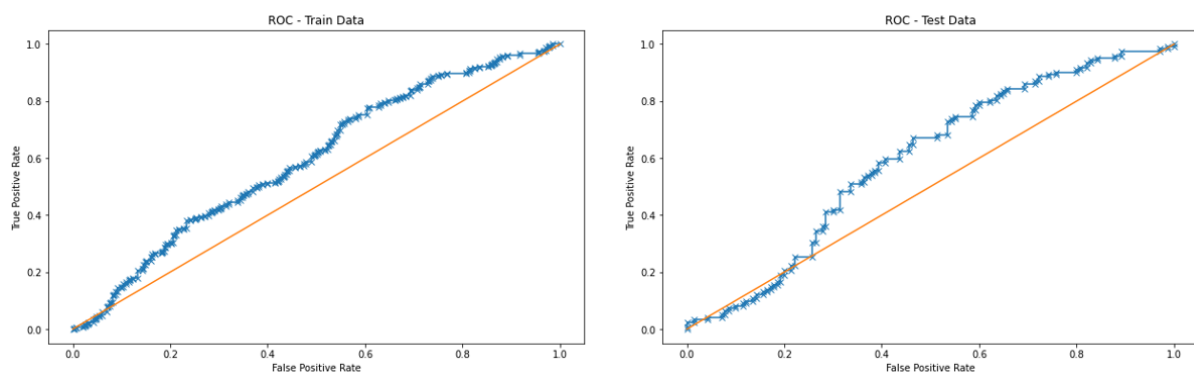


Figure XXXIX - ROC Curve (Logistic Regression)

### ROC-AUC Score:

Training data – 0.5960

Test data – 0.5957

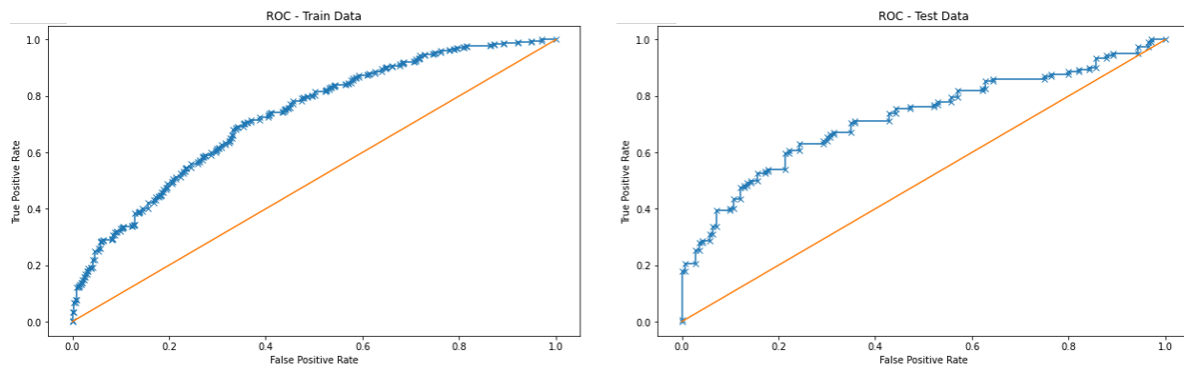


Figure XL - ROC Curve (LDA)

### ROC-AUC Score:

Training data – 0.7285

Test data – 0.7223

### Model Comparison –

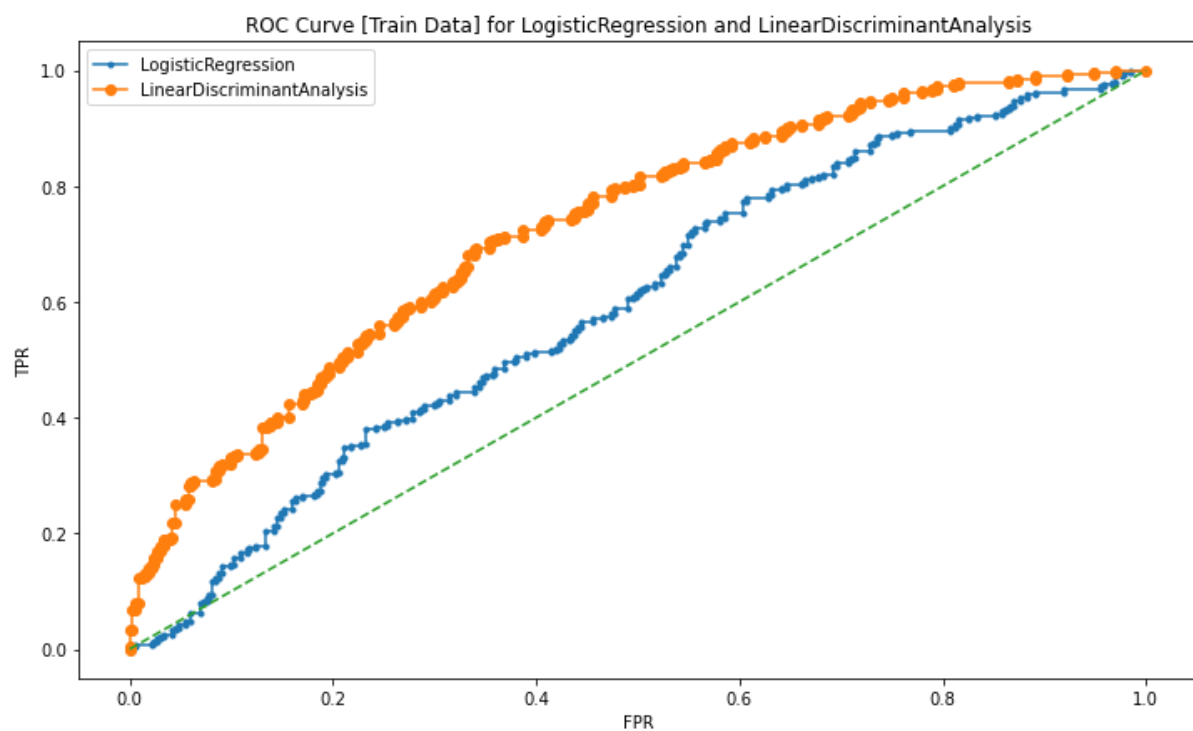


Figure XLI - Train Data ROC Curve

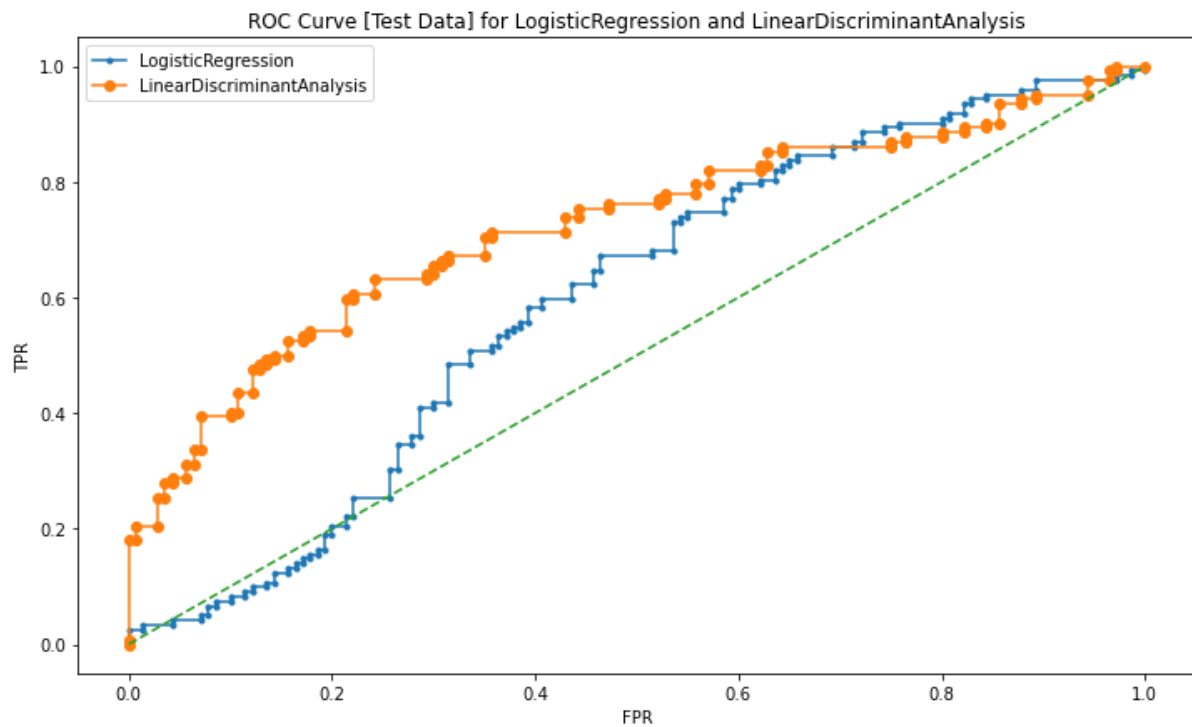


Figure XLII - Test Data ROC Curve

It is evident by above that LDA model is working accurately compared to Logistic Regression model and covers more area under the curve.

Metrics	Logistic Regression		Linear Discriminant Analysis	
	Train Data	Test Data	Train Data	Test Data
Accuracy	0.59	0.57	0.67	0.68
AUC Score	0.59	0.59	0.72	0.72
Recall	0.38	0.42	0.56	0.54
Precision	0.57	0.55	0.66	0.71
F1 Score	0.46	0.47	0.6	0.61

Table 37 - Metrics Comparison

With comparison above, we can infer below –

- LDA model has better accuracy 67-68 % than logistic regression model 57-59%
- LDA model has better AUC score 0.72 compared to regression model 0.59
- Recall, precision and F1-Score, all metrics are improved in LDA model.

Hence, we can conclude that **LDA model is performing better**, and we would select that for our predictions.



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.

**Insights** derived from EDA and building logistic regression & LDA –

- Almost 46% employee opted for Holiday Package
- We have 24% foreigner employees in the company
- Employees were aged from 20 to 62 and having salary as high as 236k+ and as low as 1.3k+
- Most of the employees don't have younger or older children
- Foreigner employee did not receive long formal education whereas non-foreigner employees received formal education for long time
- Employees who have young children usually do not opt for holiday package
- Foreign employees have lesser salary compared to non-foreigners
- The decision/prediction was driven mostly by nationality (foreigner or not), Age and Education
- Foreigner employees are likely to opt for holiday package.
- As age increases, employee loose interest in the holidays
- Employee spent more years in formal education, are likely to opt the holiday package.

### Recommendation

- Agency can focus on below important factor to increase their package sale/profile –
  - Foreign employees – They mostly opt for holiday packages; hence agency can slightly increase their pricing there to book profits.
  - Employees having young children – Company can offer day care in their packages to attract those employees and increase their sales.
  - Employee's Age – Experienced or aged employees doesn't seem to opt for package, agency can offer them discount as per their age to invite them into buying.
  - Education – Employees who have spent more years in formal education, they are buying the package, hence agency can offer discount for employees who have less formal education years spent.

\*\*\*\*\* END \*\*\*\*\*