Project: Predictive Modelling

Name: Varsha Srinivasan

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

PROBLEM 1: Linear Regression

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.

Introduction

The purpose of this whole exercise is to perform exploratory data analysis and perform Regression using Linear Regression. The dataset consists of 27000 rows with their features like Carat, Cut , Color , Clarity, Depth, Table, X,Y,Z and dependent variable Price. Insights are derived and recommendations are made.

Data Description

1. Carat: Carat weight of the cubic zirconia
2. Cut: Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal.
3. Color: Colour of the cubic zirconia. With D being the worst and J the best.
4. 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.
5. Depth: The Height of cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter.
6. Table: The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter.
7. X : Length of the cubic zirconia in mm.
8. Y : Width of the cubic zirconia in mm.
9. Z : Height of the cubic zirconia in mm.

10.Target Variable - Price: The Price of the cubic zirconia.

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.

Sample of the dataset:

	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 1: Sample of the Dataset1

The data is read from the csv file and the above tables shows the first 5 rows of the dataset. The price is the target variable. It is in continuous form hence Linear regression Is applied.

EXPLORATORY DATA ANALYSIS

Data Type and Missing Values

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26967 entries, 0 to 26966
Data columns (total 10 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   carat    26967 non-null  float64
 1   cut      26967 non-null  object
 2   color    26967 non-null  object
 3   clarity  26967 non-null  object
 4   depth    26270 non-null  float64
 5   table    26967 non-null  float64
 6   x        26967 non-null  float64
 7   y        26967 non-null  float64
 8   z        26967 non-null  float64
 9   price    26967 non-null  int64
dtypes: float64(6), int64(1), object(3)
memory usage: 2.1+ MB
```

Table 2: Dataset Info1

All the variables have continuous values except for cut,color and clarity which has object data type.It has 29697 rows and 10 columns.

There are 34 duplicates in the dataset. Sample duplicates are:

	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

Table 3: Duplicate Sample

The duplicates are removed.

```
carat        0
cut          0
color        0
clarity      0
depth      697
table        0
x            0
y            0
z            0
price        0
dtype: int64
```

There are 697 null values in the depth variable of the dataset

	carat	cut	color	clarity	depth	table	x	y	z	price
count	26933.000000	26933	26933	26933	26236.000000	26933.000000	26933.000000	26933.000000	26933.000000	26933.000000
unique	NaN	5	7	8	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	Ideal	G	SI1	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	10805	5653	6565	NaN	NaN	NaN	NaN	NaN	NaN
mean	0.798010	NaN	NaN	NaN	61.745285	57.455950	5.729346	5.733102	3.537769	3937.526120
std	0.477237	NaN	NaN	NaN	1.412243	2.232156	1.127367	1.165037	0.719964	4022.551862
min	0.200000	NaN	NaN	NaN	50.800000	49.000000	0.000000	0.000000	0.000000	326.000000
25%	0.400000	NaN	NaN	NaN	61.000000	56.000000	4.710000	4.710000	2.900000	945.000000
50%	0.700000	NaN	NaN	NaN	61.800000	57.000000	5.690000	5.700000	3.520000	2375.000000
75%	1.050000	NaN	NaN	NaN	62.500000	59.000000	6.550000	6.540000	4.040000	5356.000000
max	4.500000	NaN	NaN	NaN	73.600000	79.000000	10.230000	58.900000	31.800000	18818.000000

Table 4:Description of Dataset1

The total number of entries is 26933 after removing 34 duplicates in all features except depth which has 26236 entries as additional 697 non-null values in the dataset are present. All the variables have continuous values except for cut, color and clarity which has object data type. Carat has a mean of around 0.80. Depth have a mean of 61.75. Table has a mean of 57.46. x has a mean of 5.73. y has a mean of 5.73. z has a mean of 3.54. Price has a mean of 3937.53. Cut has the Ideal value the most. Color has G value the most. Clarity has the SI1 value the most.

Outliers Proportions

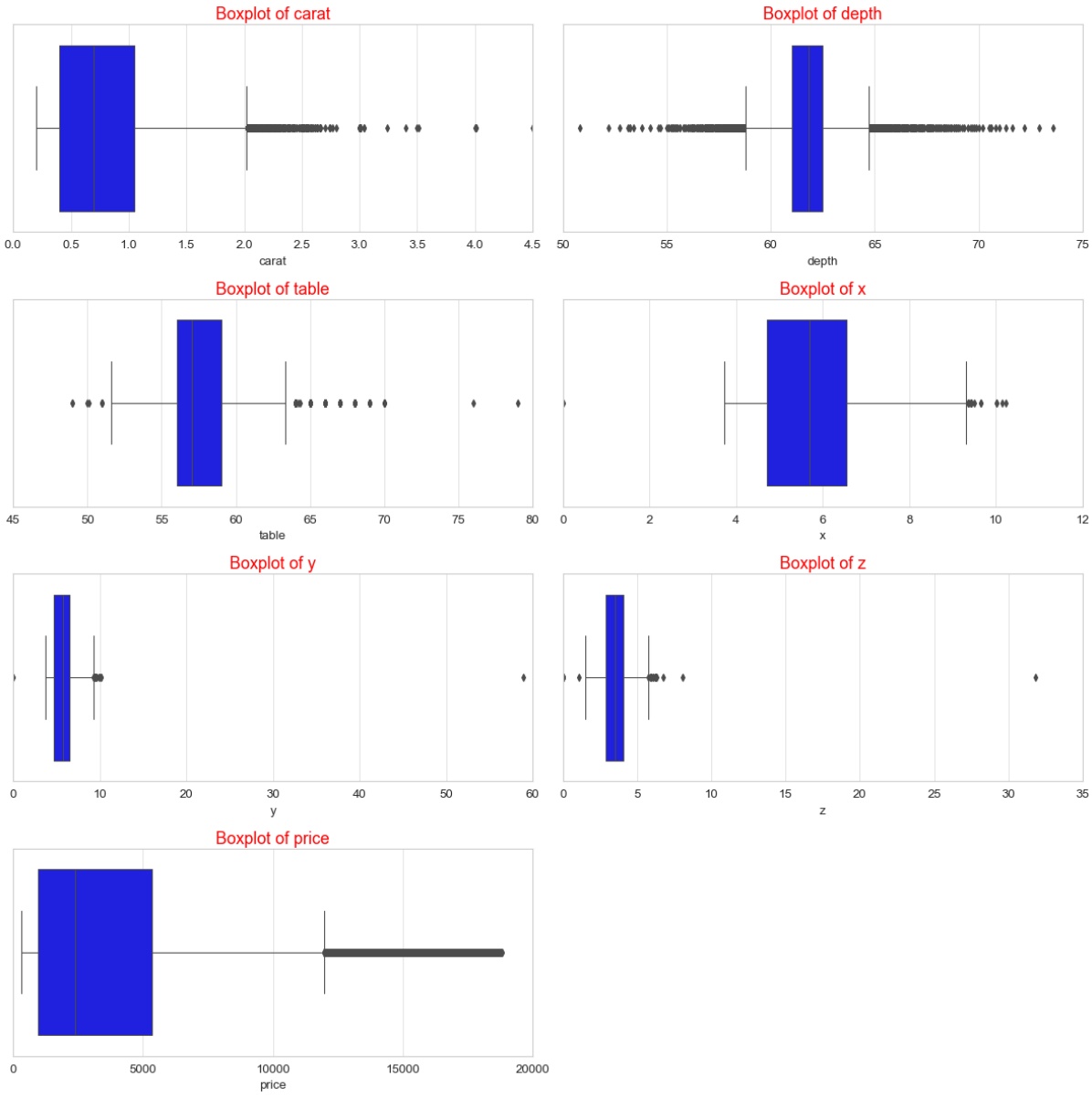

Figure 1: Boxplot with Outliers

There are outliers in all the features. Table has a median of around 57. Depth has a median of around 61. Carat has a median of 0.7.x, y has a median of around of 5.7, z has a median of around 3.52. Price has a median of around 2375.

```
Lower outliers in carat is :  -0.5750000000000001
Upper outliers in carat is :  2.0250000000000004
Number of outliers in carat upper :  657
Number of outliers in carat lower :  0
% of Outlier in carat upper:  2 %
% of Outlier in carat lower:  0 %
-----------------------------------------------------
Lower outliers in depth is :  58.75
Upper outliers in depth is :  64.75
Number of outliers in depth upper :  486
Number of outliers in depth lower :  733
% of Outlier in depth upper:  2 %
% of Outlier in depth lower:  3 %
-----------------------------------------------------
Lower outliers in table is :  51.5
Upper outliers in table is :  63.5
Number of outliers in table upper :  310
Number of outliers in table lower :  8
% of Outlier in table upper:  1 %
% of Outlier in table lower:  0 %
-----------------------------------------------------
Lower outliers in x is :  1.9500000000000002
Upper outliers in x is :  9.309999999999999
Number of outliers in x upper :  12
Number of outliers in x lower :  2
% of Outlier in x upper:  0 %
% of Outlier in x lower:  0 %
-----------------------------------------------------
Lower outliers in y is :  1.9649999999999999
Upper outliers in y is :  9.285
Number of outliers in y upper :  12
Number of outliers in y lower :  2
% of Outlier in y upper:  0 %
% of Outlier in y lower:  0 %
-----------------------------------------------------
Lower outliers in z is :  1.1899999999999997
Upper outliers in z is :  5.75
Number of outliers in z upper :  13
Number of outliers in z lower :  9
% of Outlier in z upper:  0 %
% of Outlier in z lower:  0 %
-----------------------------------------------------
Lower outliers in price is :  -5671.5
Upper outliers in price is :  11972.5
Number of outliers in price upper :  1778
Number of outliers in price lower :  0
% of Outlier in price upper:  7 %
% of Outlier in price lower:  0 %
-----------------------------------------------------
```

Univariate Analysis

Box Plot without Outliers

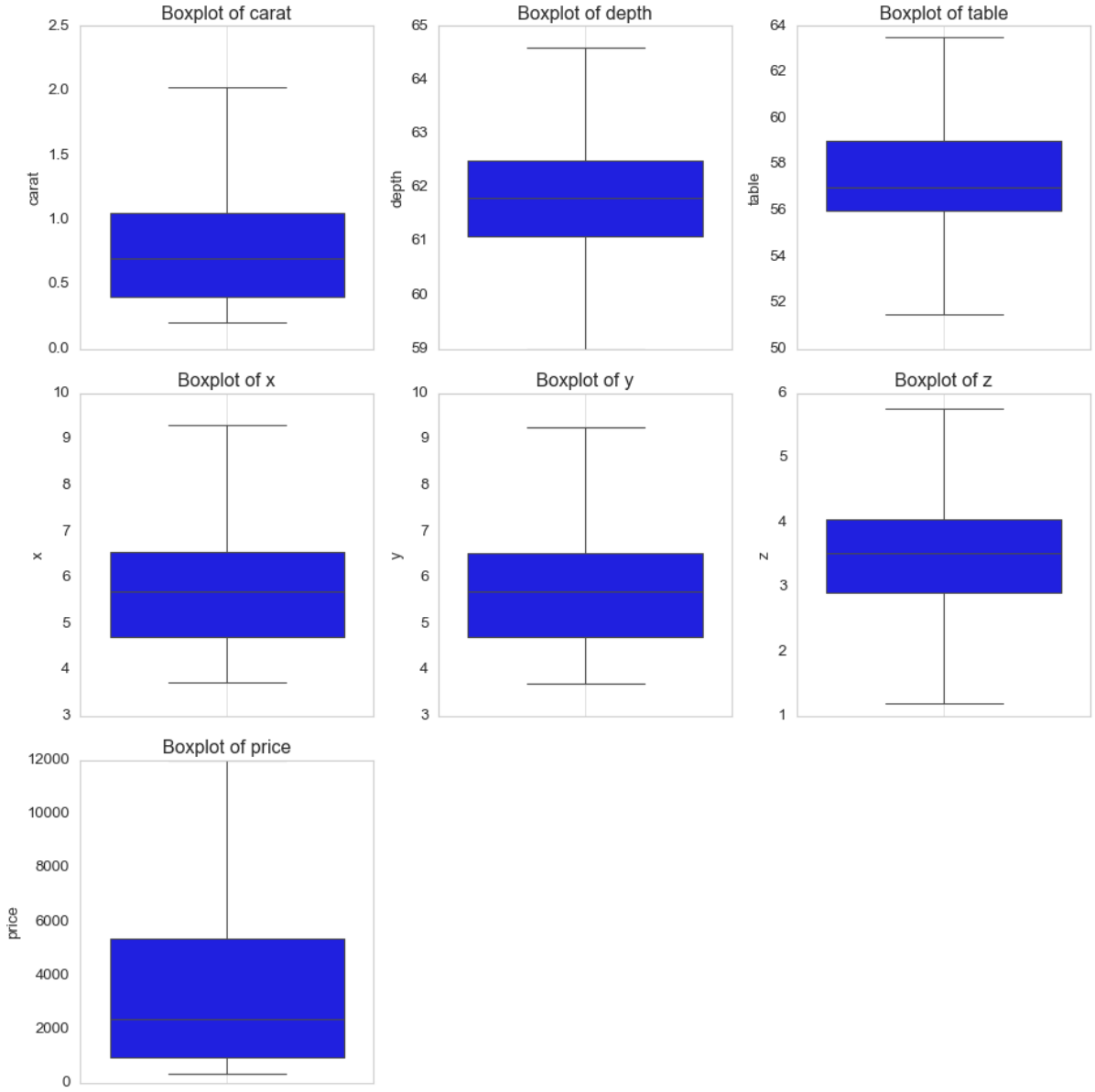

Figure 2:Boxplot without Outliers

Model 5 and 6 makes use of data with no outliers.

Analysis of Carat

```
Description of carat
--------------------------------------------------------------------------------
-
count    26933.000000
mean         0.798010
```

```
std          0.477237
min          0.200000
25%          0.400000
50%          0.700000
75%          1.050000
max          4.500000
Name: carat, dtype: float64

Interquartile range (IQR) of spending is  0.65
Range of values:  4.3

Distribution of carat
------------------------------------------------------------------------------
-
```

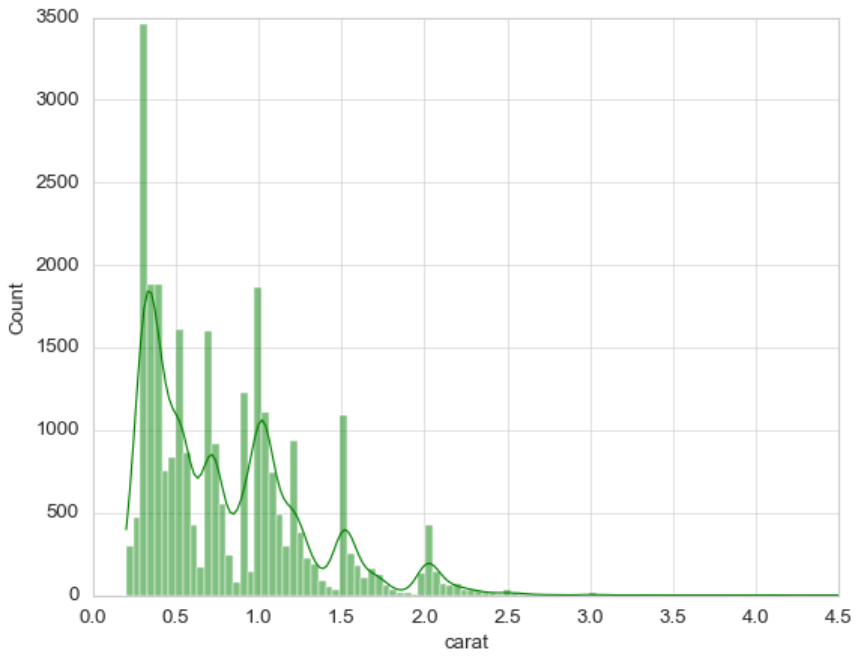

Figure 3: Distribution of carat

carat is extremely right skewed.

Analysis of Depth

```
Description of depth
------------------------------------------------------------------------------
-
count    26236.000000
mean        61.745285
std          1.412243
min         50.800000
25%         61.000000
50%         61.800000
```

```
75%          62.500000
max          73.600000
Name: depth, dtype: float64

Interquartile range (IQR) of spending is  nan
Range of values:  22.8

Distribution of depth
------------------------------------------------------------------------------
-
```

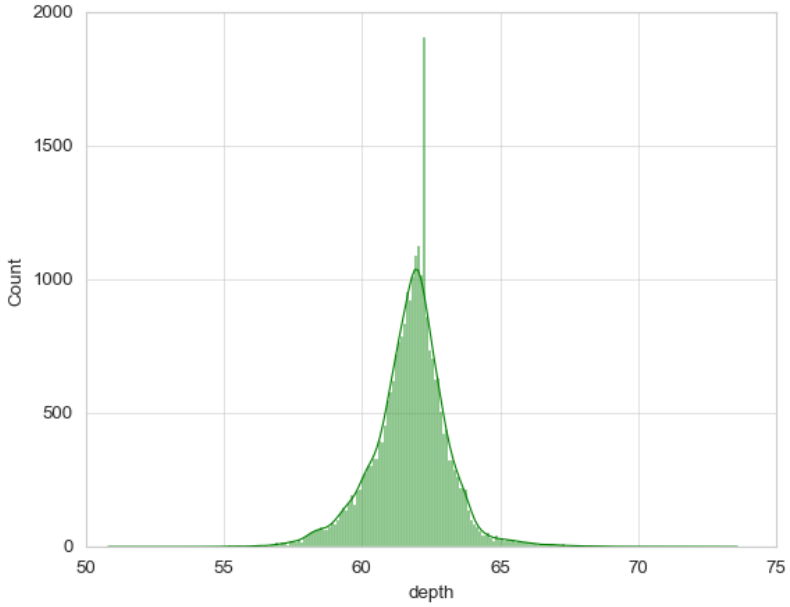

Figure 4: Distribution of depth

Depth is almost normally distributed.

Analysis of Table

```
Description of table
------------------------------------------------------------------------------
-
count    26933.000000
mean        57.455950
std          2.232156
min         49.000000
25%         56.000000
50%         57.000000
75%         59.000000
max         79.000000
Name: table, dtype: float64

Interquartile range (IQR) of spending is  3.0
Range of values:  30.0
```

```
Distribution of table
-------------------------------------------------------------------------------
-
```

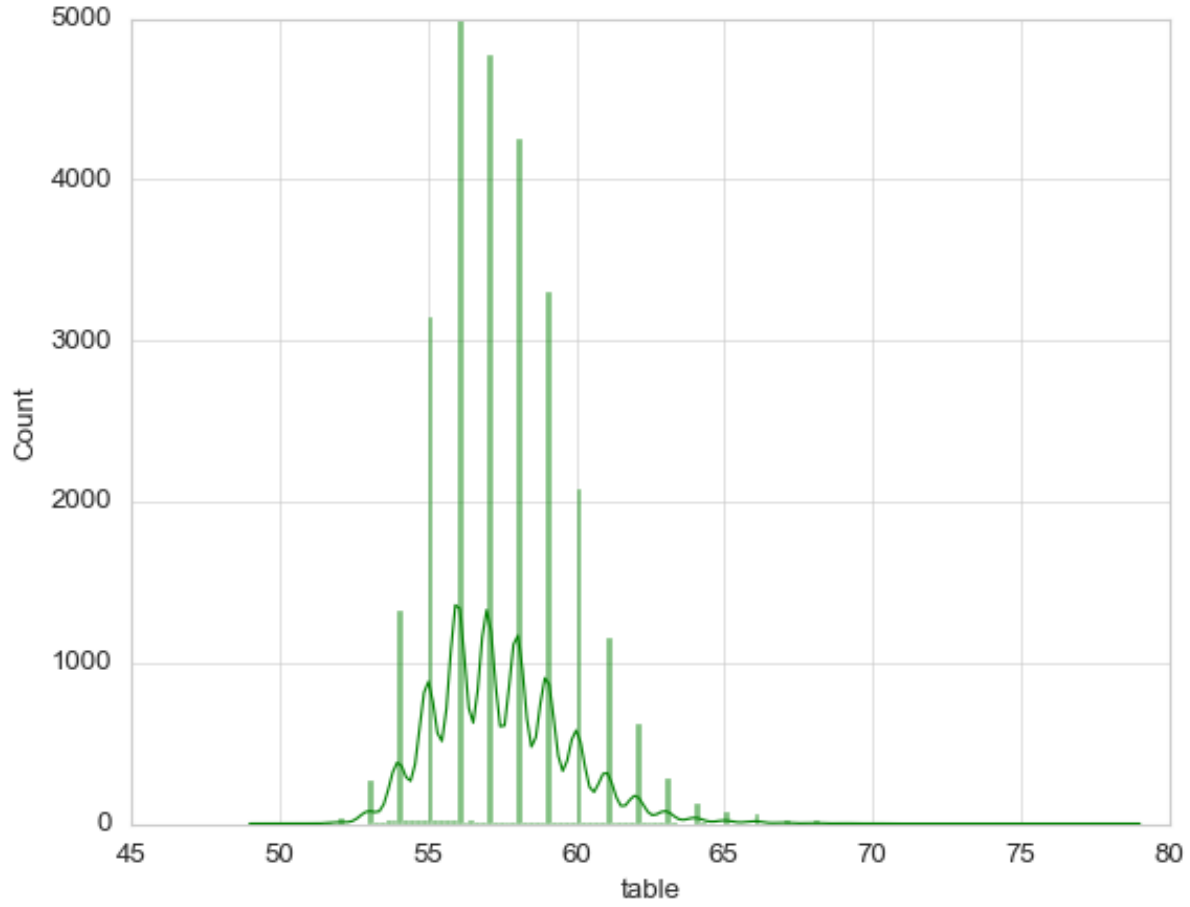

Figure 5: Distribution of table

Table is slightly right skewed.

Analysis of x

```
Description of x
-------------------------------------------------------------------------------
-
count    26933.000000
mean         5.729346
std          1.127367
min          0.000000
25%          4.710000
50%          5.690000
75%          6.550000
max         10.230000
Name: x, dtype: float64
```

```
Interquartile range (IQR) of spending is  1.84
Range of values:  10.23

Distribution of x
--------------------------------------------------------------------------------
-
```

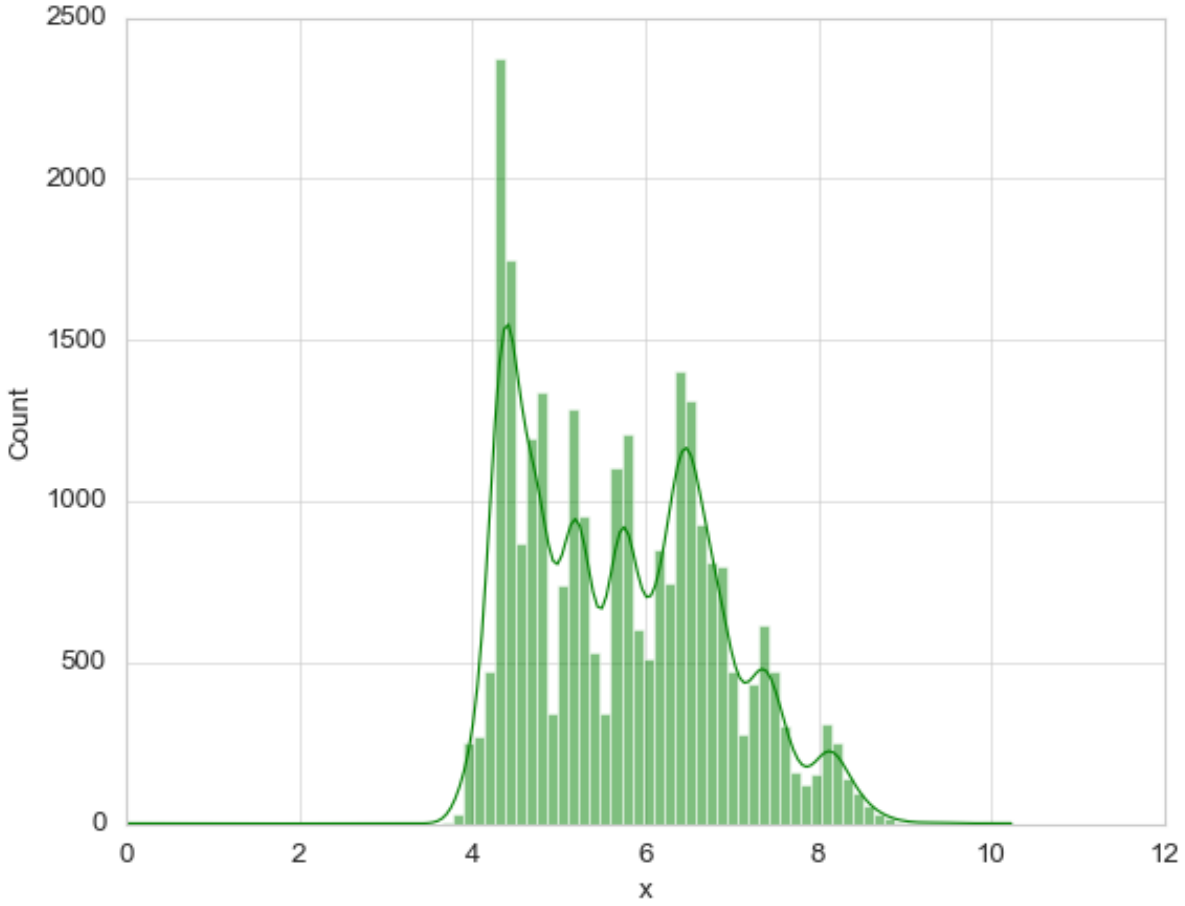

Figure 6: Distribution of x

X is nearly symmetrically distributed.

Analysis of y

```
Description of y
--------------------------------------------------------------------------------
-
count    26933.000000
mean         5.733102
std          1.165037
min          0.000000
25%          4.710000
50%          5.700000
75%          6.540000
max         58.900000
Name: y, dtype: float64
```

```
Interquartile range (IQR) of spending is  1.83
Range of values:  58.9

Distribution of y
--------------------------------------------------------------------------------
-
```

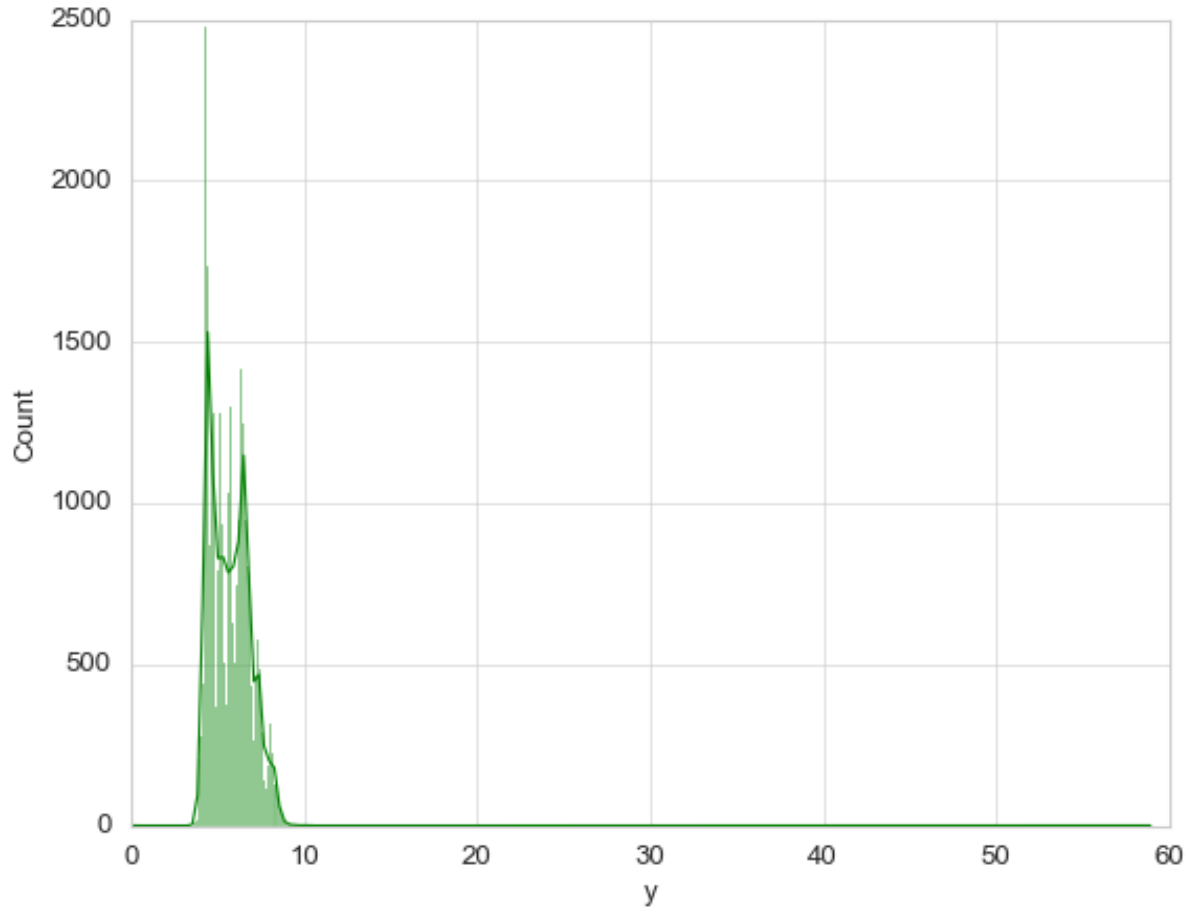

Figure 7:Distribution of y

y is extremely right skewed.

Analysis of z

```
Description of z
--------------------------------------------------------------------------------
-
count    26933.000000
mean         3.537769
std          0.719964
min          0.000000
25%          2.900000
50%          3.520000
75%          4.040000
max         31.800000
Name: z, dtype: float64
```

```
Interquartile range (IQR) of spending is  1.14
Range of values:  31.8

Distribution of z
--------------------------------------------------------------------------------
-
```

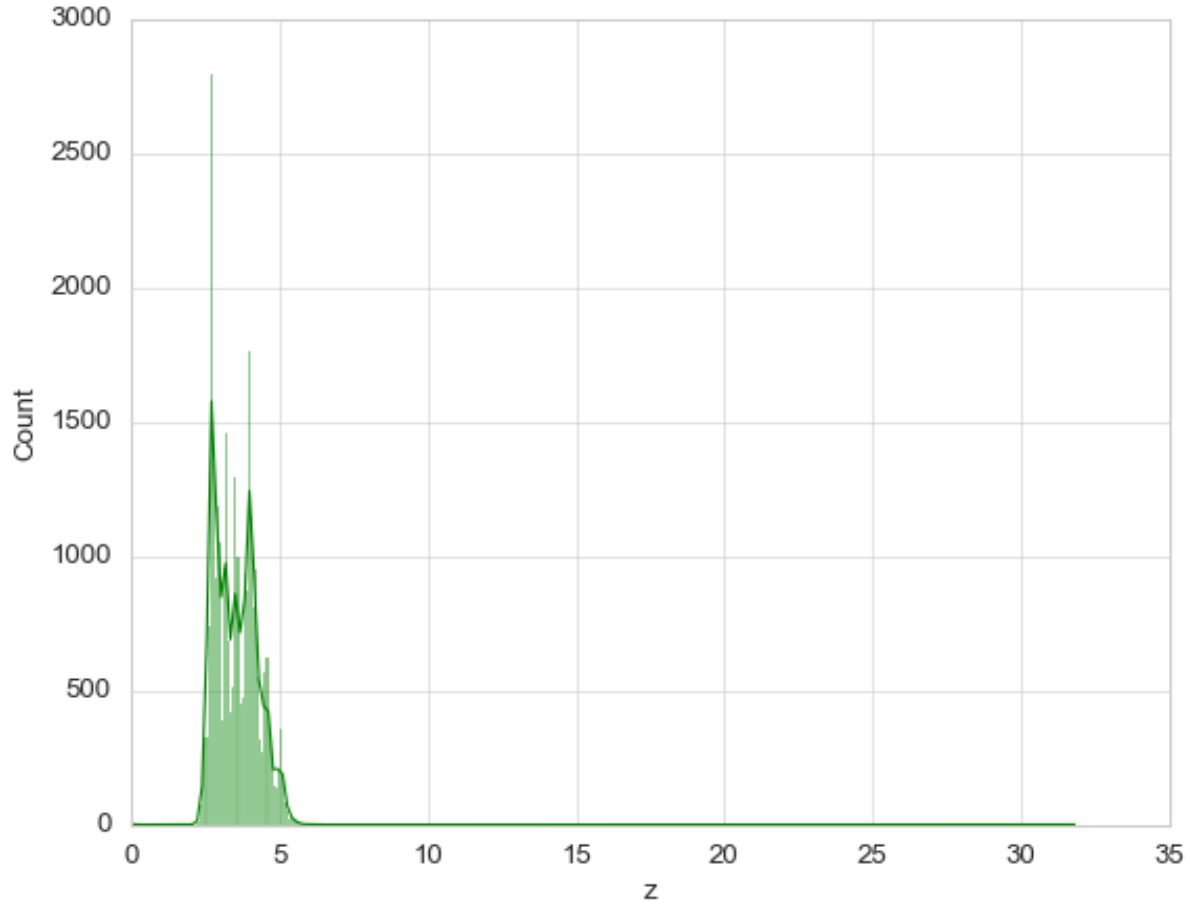

Figure 8:Distribution of z

Z is extremely right skewed.

Analysis of price

```
Description of price
--------------------------------------------------------------------------------
-
count    26933.000000
mean      3937.526120
std       4022.551862
min        326.000000
25%        945.000000
50%       2375.000000
75%       5356.000000
max      18818.000000
```

```
Name: price, dtype: float64

Interquartile range (IQR) of spending is  4411.0
Range of values:  18492

Distribution of price
-------------------------------------------------------------------------------
-
```

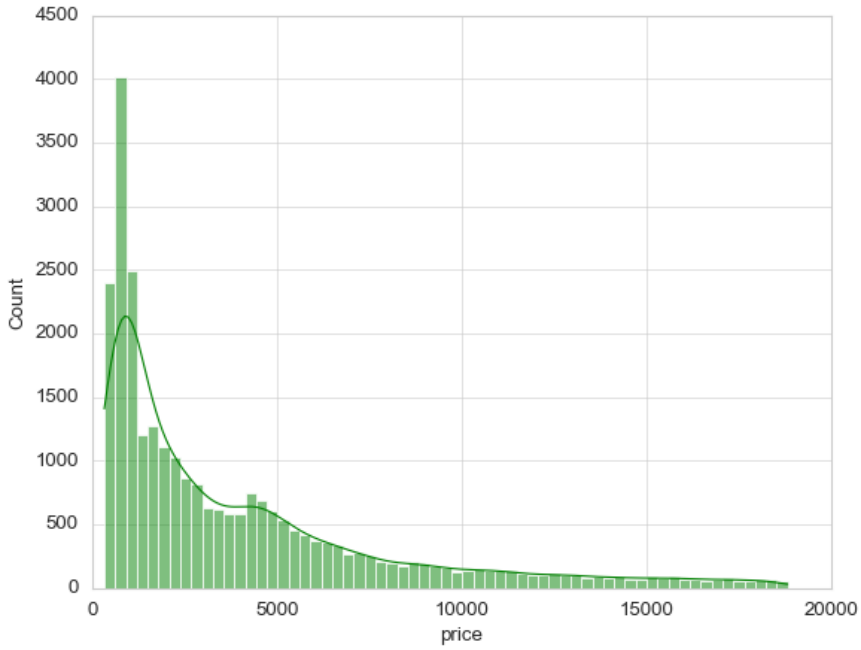

Figure 9:Distribution of price

price is extremely right skewed.

Analysis of cut

```
Value Count of cut
-------------------------------------------------------------------------------
-
Ideal        10805
Premium       6886
Very Good     6027
Good          2435
Fair           780
Name: cut, dtype: int64

Description of cut
-------------------------------------------------------------------------------
-
count     26933
unique        5
top       Ideal
freq      10805
```

```
Name: cut, dtype: object

Countplot of cut
--------------------------------------------------------------------------------
-
```

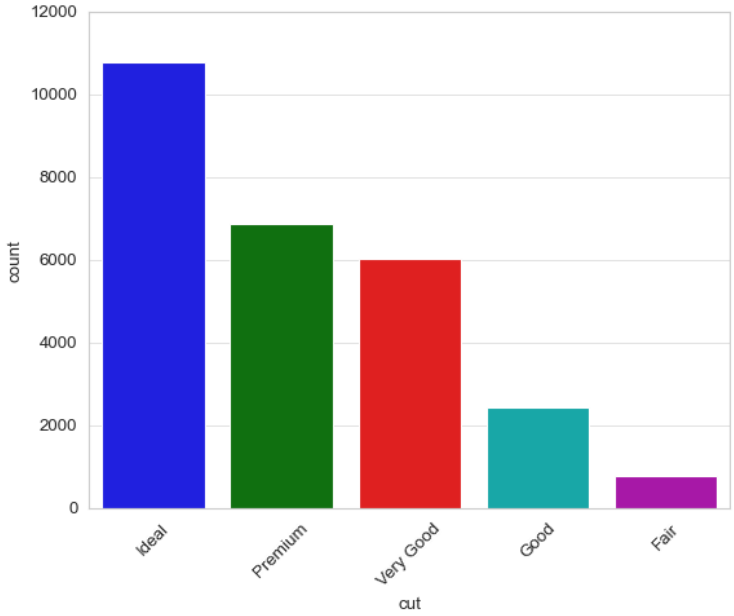

Figure 10: Countplot of Cut

Ideal is the most frequent value.

Analysis of color

```
Value Count of color
--------------------------------------------------------------------------------
-
G    5653
E    4916
F    4723
H    4095
D    3341
I    2765
J    1440
Name: color, dtype: int64

Description of color
--------------------------------------------------------------------------------
-
count     26933
unique        7
top           G
```

```
freq       5653
Name: color, dtype: object

Countplot of color
--------------------------------------------------------------------------------
-
```

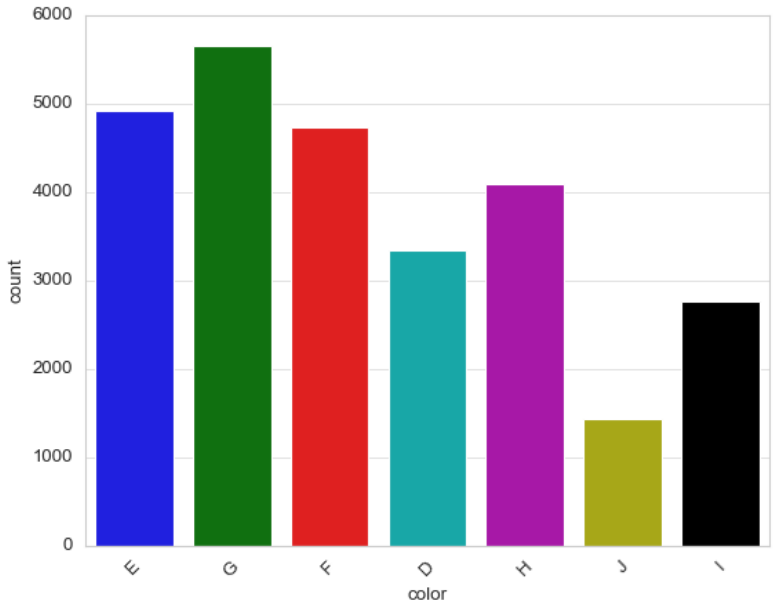

Figure 11:Countplot of color

G is the most frequent value.

Analysis of clarity

```
Value Count of clarity
--------------------------------------------------------------------------------
-
SI1      6565
VS2      6093
SI2      4564
VS1      4087
VVS2     2530
VVS1     1839
IF        891
I1        364
Name: clarity, dtype: int64

Description of clarity
--------------------------------------------------------------------------------
-
count     26933
unique        8
top         SI1
```

```
freq       6565
Name: clarity, dtype: object

Countplot of clarity
--------------------------------------------------------------------------------
-
```

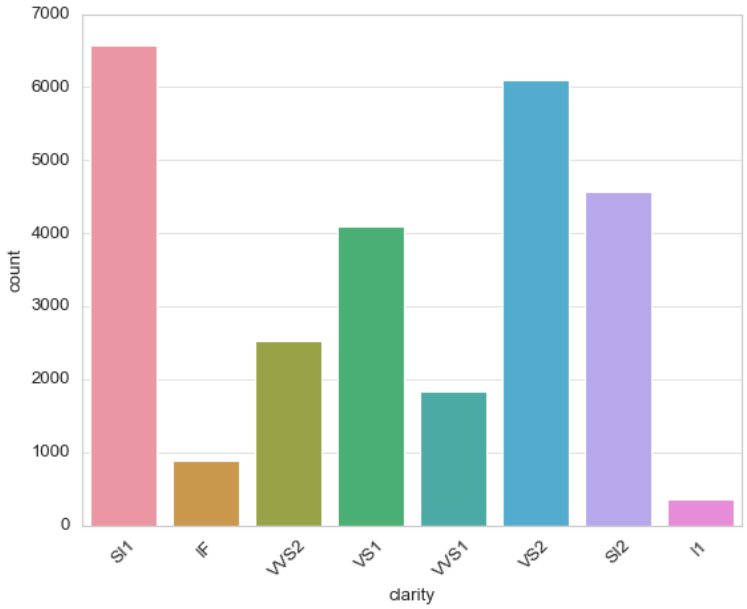

Figure 12:Countplot of clarity

SI1 is the most frequent value in Clarity.

Skewness and Kurtosis

```
Skewness of carat is 1.11
Kurtosis of carat is 1.21
Skewness of depth is -0.03
Kurtosis of depth is 3.68
Skewness of table is 0.77
Kurtosis of table is 1.58
Skewness of x is 0.39
Kurtosis of x is -0.68
Skewness of y is 3.87
Kurtosis of y is 160.04
Skewness of z is 2.58
Kurtosis of z is 87.42
Skewness of price is 1.62
Kurtosis of price is 2.15
```

Skewness essentially measures the symmetry of the distribution. In positively skewed, the mean of the data is greater than the median as a large number of data-pushed on the right-hand side. In negatively skewed, the mean of the data is less than the median as a large number of data-pushed on the left-hand.

If the skewness is between -0.5 & 0.5, the data are nearly symmetrical. If the skewness is between -1 & -0.5 (negative/left skewed) or between 0.5 & 1(positive/right skewed), the data are slightly skewed. If the skewness is lower than -1 (negative/left skewed) or greater than 1 (positive/right skewed), the data are extremely skewed.

maximum spent in single shopping, probability of full payment and current balance are nearly symmetrically distributed.

Kurtosis refers to the degree of presence of outliers in the distribution. If kurtosis > 3, then it is called as Leptokurtic or heavy-tailed distribution as the kurtosis is more than normal distribution. If kurtosis = 3, then it is called as Mesokurtic as the kurtosis is same as the normal distribution. If kurtosis < 3, then it is called as Platykurtic or short-tailed distribution as the kurtosis is less than normal distribution.

BIVARIATE ANALYSIS

PAIR PLOT

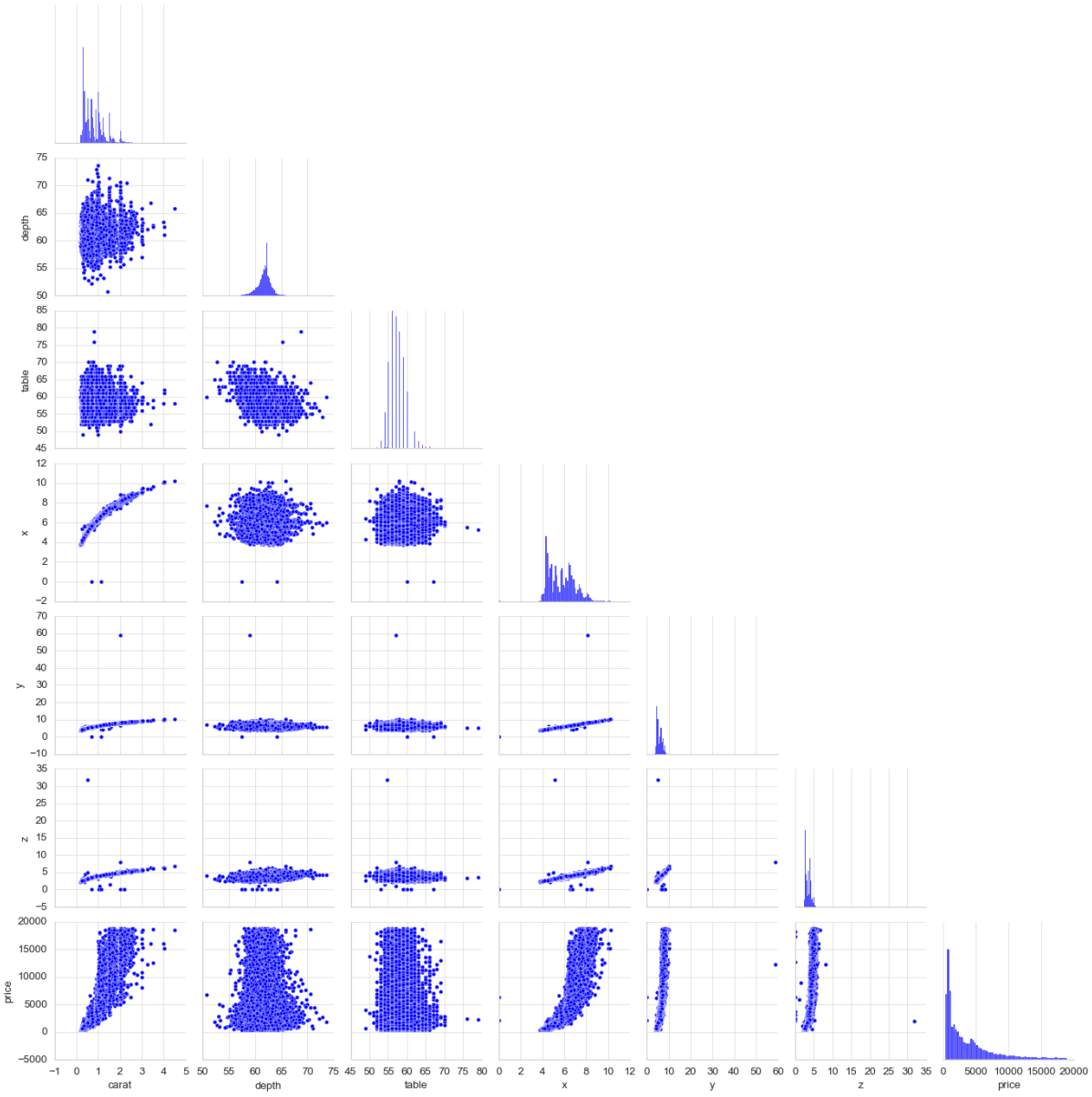

Figure 13: Pairplot

CORRELATION HEATMAP

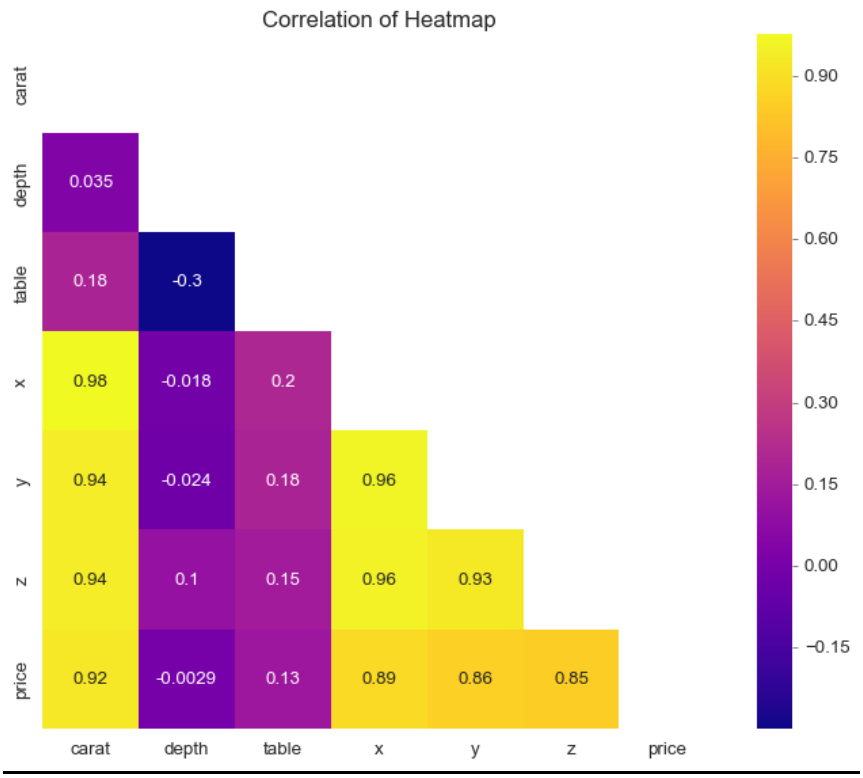

Figure 14: Correlation Heatmap

From pairplot and heatmap we can infer that there is a strong positive correlation between price and other variables such as carat,x,y,z. table is slightly positively correlated with price,x,y,z. depth is slightly negatively correlated with table,x,y,price. Depth and z are slightly positively correlated.

Boxplots

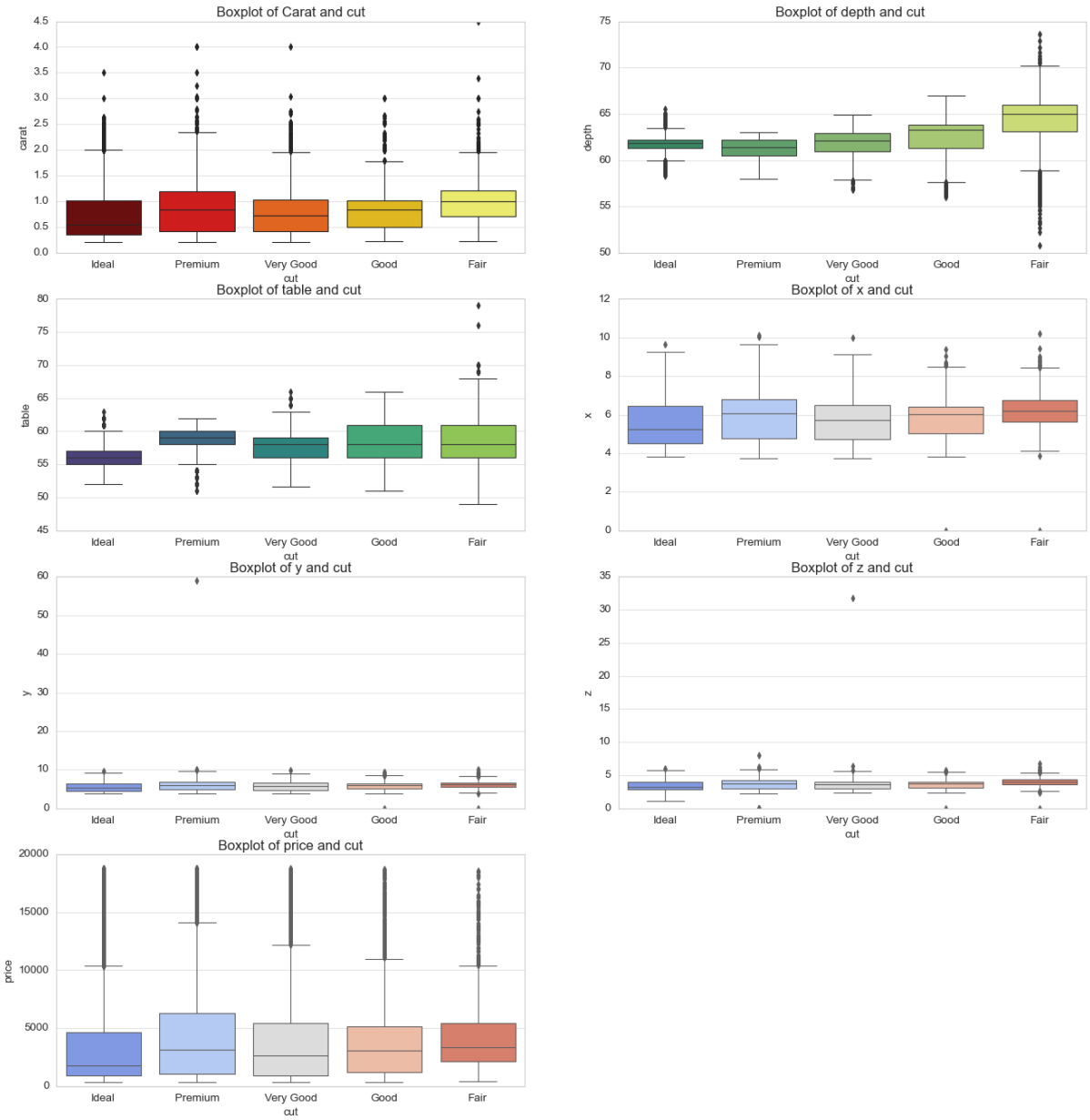

Figure 15: Boxplots of cut and other variables

cubic zirconia with Fair cut has a median depth of around 65. Premium cut cubic zirconia has a higher carat value than others.

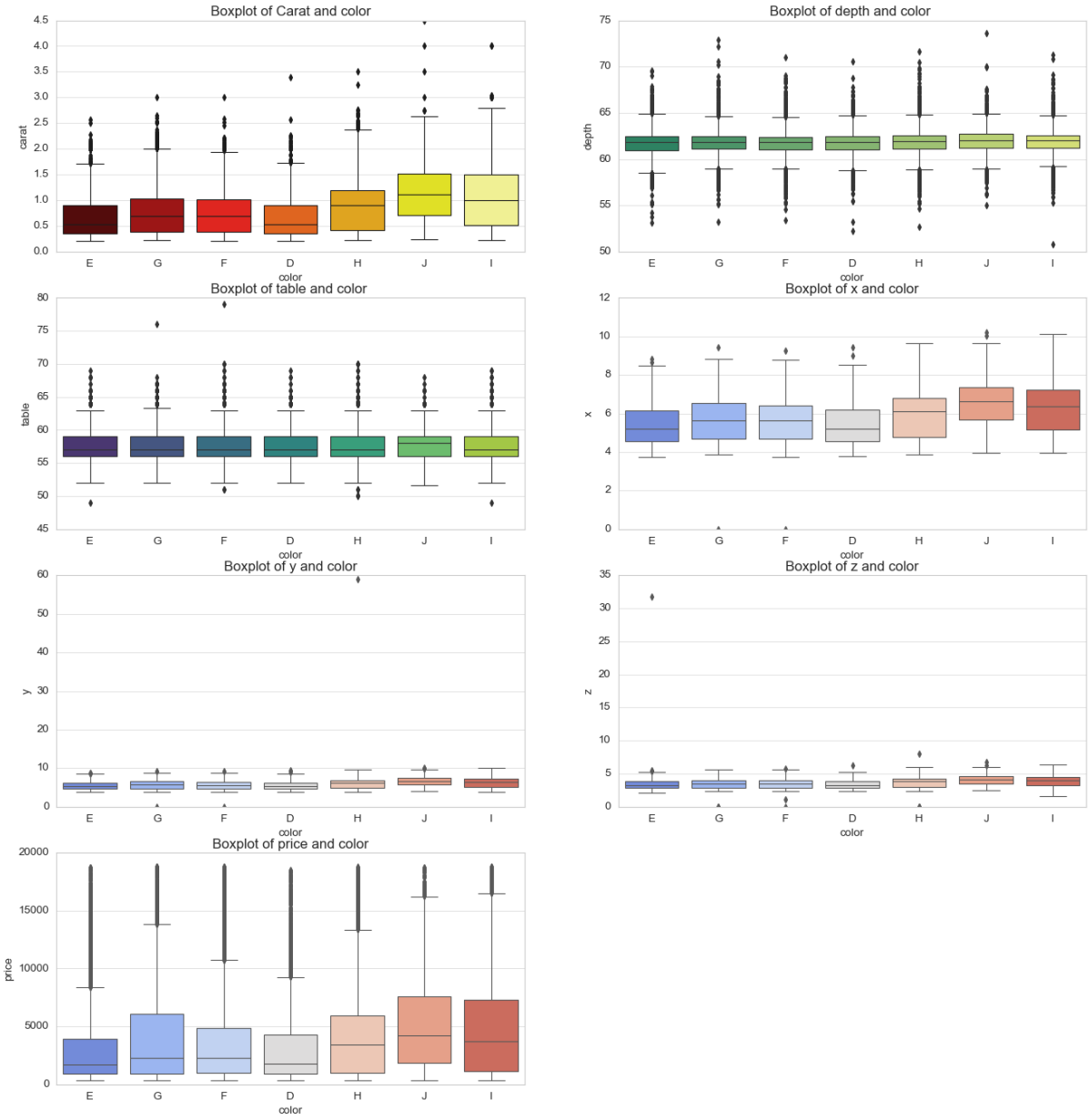

Figure 16:Boxplots of color and other variables

The one with color J has a high median price and high median length value than others.

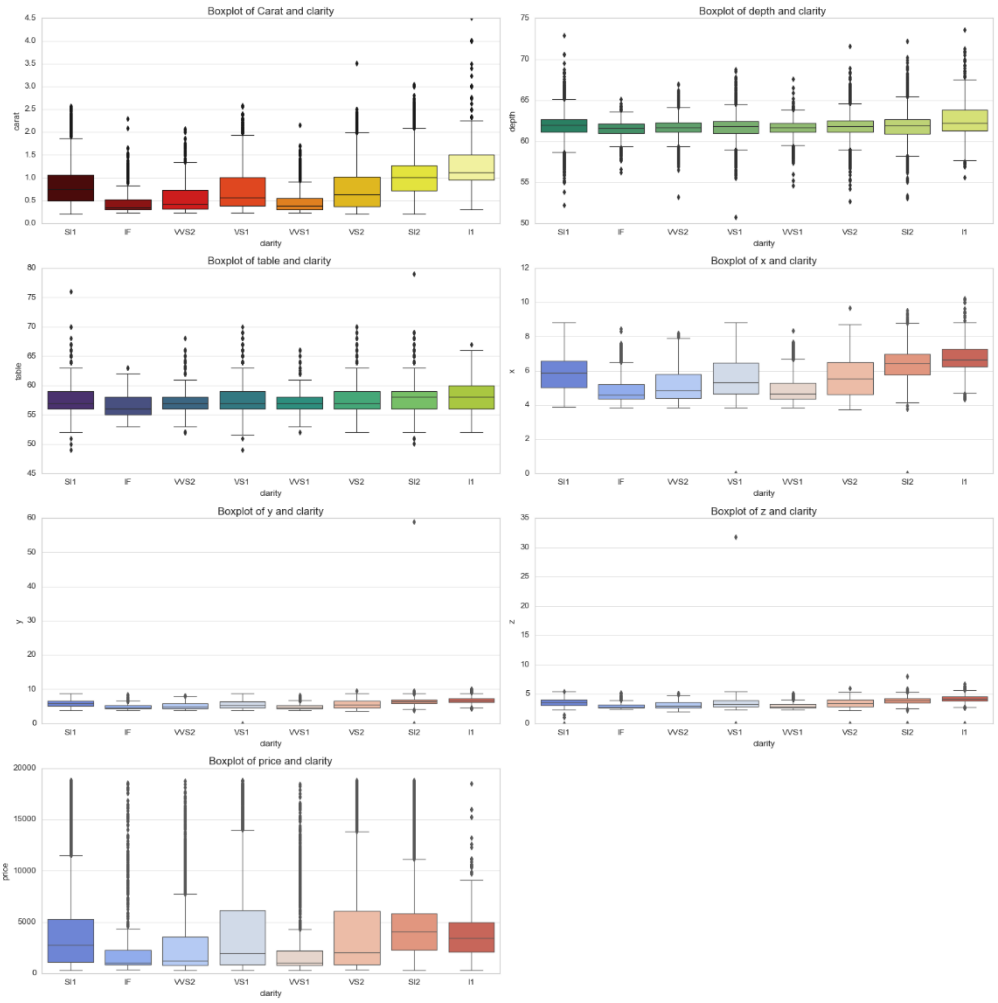

Figure 17:Boxplots of clarity and other variables

Most of the cubic zirconia with I1 cut has a higher carat value than others.

Multivariate Analysis

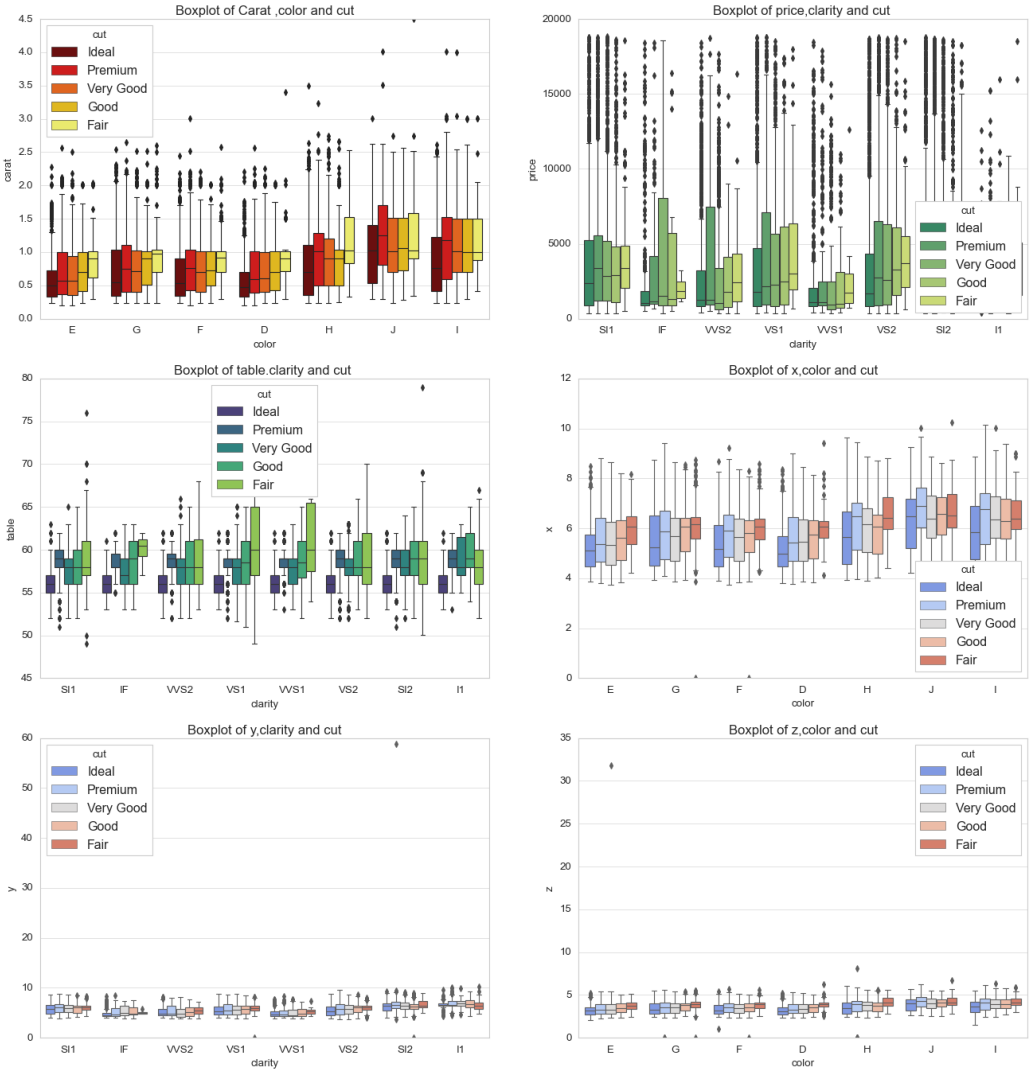

Figure 18: Boxplots for Multivariate Analysis1

Cut Premium with Color J has a higher median carat value. Most of the cubic zirconia with Very Good cut ,clarity VVS2 has a higher price range.

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.

Depth column has null values which is less than 5% of the data which are imputed with the median since we are using the data with outliers.

Sample records with imputed value is shown below:

	carat	cut	color	clarity	depth	table	x	y	z	price
26	0.34	Ideal	D	SI1	61.8	57.0	4.50	4.44	2.74	803
86	0.74	Ideal	E	SI2	61.8	59.0	5.92	5.97	3.52	2501
117	1.00	Premium	F	SI1	61.8	59.0	6.40	6.36	4.00	5292
148	1.11	Premium	E	SI2	61.8	61.0	6.66	6.61	4.09	4177
163	1.00	Very Good	F	VS2	61.8	55.0	6.39	6.44	3.99	6340

Table 5: Sample table with imputed values

X,y,z columns have some records with 0 as a value.

	carat	cut	color	clarity	depth	table	x	y	z	price
5821	0.71	Good	F	SI2	64.1	60.0	0.0	0.0	0.0	2130
17506	1.14	Fair	G	VS1	57.5	67.0	0.0	0.0	0.0	6381

Table 6: Records with x as 0

	carat	cut	color	clarity	depth	table	x	y	z	price
5821	0.71	Good	F	SI2	64.1	60.0	0.0	0.0	0.0	2130
17506	1.14	Fair	G	VS1	57.5	67.0	0.0	0.0	0.0	6381

Table 7:Records with y as 0

	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 8:Records with z as 0

x,y and z represents length,width and height of the cubic zirconia in mm respectively.Hence its not possible to have 0 as its values. We have to change them by imputing with the median value.

The VIF values of numeric variables are:

```
carat ---> 112.60408541595991
depth ---> 505.9482255351656
table ---> 500.8486876073657
x ---> 1077.283065551717
y ---> 347.8760665780185
z ---> 373.86986405430247
price ---> 13.536038629355726
```

All variables have a vif value greater than 5.A value greater than 5 indicates potentially severe correlation between a given explanatory variable and other explanatory variables in the model.

Cut can be re-categorised into Ideal,Premium and Good. Premium and Very Good is combined as Premium. Good and Fair are combined as Good. This is combined based on the x values.

Clarity can be re-categorised into IF,VVS1,VS1,Sl1 and l1. VVS1 and VVS2 is combined as VVS1. VS1 and VS2 are combined as VS1. Sl1 and Sl2 are combined as Sl1. It is done based on Carat value.

Color can be re-categorised into D, F, H and J. E and F is combined as F. G and H are combined as H. I and J are combined as J.It is done based on Carat value.

Model 3 and 4 uses this combined encoding.

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.

The features with object data types color, cut and clarity are ordinal variables. So they are encoded with integers in increasing order as follows and stored in a dictionary object.

Cut:

```
{'Fair': 0, 'Good': 1, 'Very Good': 2, 'Premium': 3, 'Ideal': 4}
```

Color**:**

```
{'D': 0, 'E': 1, 'F': 2, 'G': 3, 'H': 4, 'I': 5, 'J': 6}
```

Clarity**:**

```
{'IF': 0,
 'VVS1': 1,
 'VVS2': 2,
 'VS1': 3,
 'VS2': 4,
 'SI1': 5,
 'SI2': 6,
 'I1': 7}
```

Model 1,2,5 and 6 uses this encoding.

Model 1 – Sklearn LR model

```
LinearRegression()
```

The coefficients are :

```
The coefficient for carat is 11131.971721301814
The coefficient for cut is 107.28987644346527
The coefficient for color is -334.09629862475043
The coefficient for clarity is -504.5027691971778
The coefficient for depth is -82.03515027559568
The coefficient for table is -30.081791696879492
The coefficient for x is -984.3267054058507
The coefficient for y is 10.011040432150162
The coefficient for z is -45.305242609687845
```

The intercept for our model is 10142.448725297465

RMSE for train data is 1211.3777081621438
RMSE for test data is 1229.2135811232442

R-squared for train data is 0.9087174183809241
R-squared for test data is 0.9080038175630045

Adjusted R-squared for train data is 0.9086738189947026
Adjusted R-squared for test data is 0.9079012195900265

Scatterplot

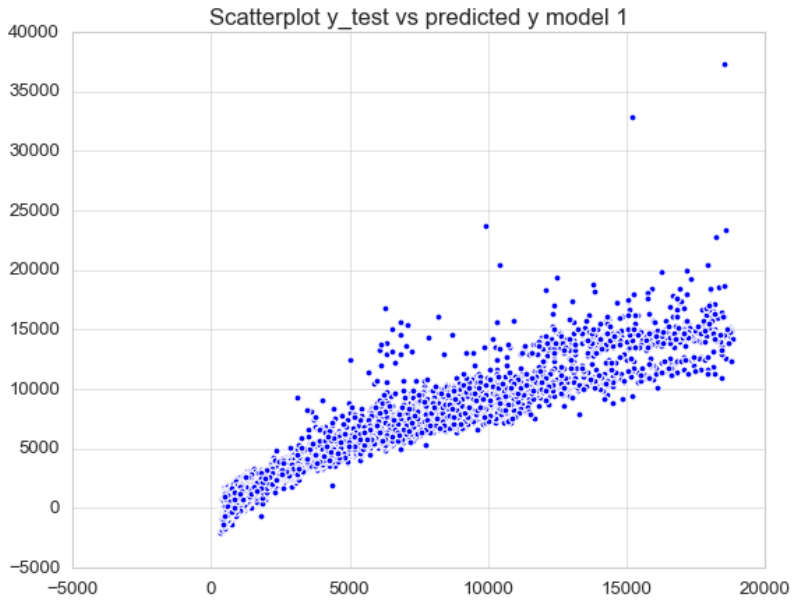

Figure 19: Model 1 - Scatterplot of y test with predicted y

The scatterplot is linear in nature.

Scaling of data isn't required in Linear Regression as the coefficients take care of the difference in scale of the independent variables.

After scaling the score of the model is nearly the same as the unscaled one for both train(0.909) and test(0.908).

Model 2- OLS

OLS is built for unscaled data with the initial encoded data with expression :

```
'price ~ carat + cut + color + clarity + depth + table + x + y + z'
```

The results are as follows:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.909
Model:	OLS	Adj. R-squared:	0.909
Method:	Least Squares	F-statistic:	2.084e+04
Date:	Sun, 10 Apr 2022	Prob (F-statistic):	0.00
Time:	16:26:01	Log-Likelihood:	-1.6060e+05
No. Observations:	18853	AIC:	3.212e+05
Df Residuals:	18843	BIC:	3.213e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.014e+04	710.811	14.269	0.000	8749.196	1.15e+04
carat	1.113e+04	93.346	119.255	0.000	1.09e+04	1.13e+04
cut	107.2899	9.739	11.017	0.000	88.201	126.378
color	-334.0963	5.480	-60.962	0.000	-344.838	-323.354
clarity	-504.5028	5.916	-85.283	0.000	-516.098	-492.908
depth	-82.0352	7.890	-10.397	0.000	-97.500	-66.570
table	-30.0818	4.984	-6.036	0.000	-39.851	-20.313
x	-984.3267	50.690	-19.419	0.000	-1083.684	-884.970
y	10.0110	23.763	0.421	0.674	-36.567	56.589
z	-45.3052	41.512	-1.091	0.275	-126.672	36.062

Omnibus:	4008.482	Durbin-Watson:	1.981
Prob(Omnibus):	0.000	Jarque-Bera (JB):	166200.863
Skew:	0.083	Prob(JB):	0.00
Kurtosis:	17.545	Cond. No.	6.86e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.86e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

Table 9: Model 2 – OLS Results

Coefficients having p-values less than alpha(0.05) are statistically significant.
Carat,cut,color,clarity,depth,table and x are significant variables.

RMSE for train data is 1211.3777081621438
RMSE for test data is 1229.2135811232413

R-squared for train data is 0.909
R-squared for test data is 0.9080038175630047

Adjusted R-squared for train data is 0.909
Adjusted R-squared for test data is 0.9086738189947028

Scatterplot

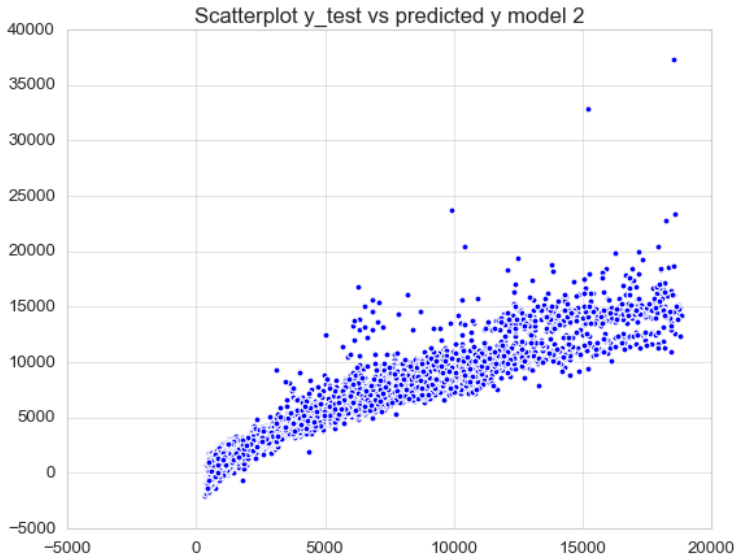

Figure 20:Model 2 - Scatterplot of y test with predicted y

The scatterplot is linear in nature.

Model 3 – Sklearn Combined Encoding with Outliers

Cut can be re-categorised into Ideal,Premium and Good. Premium and Very Good is combined as Premium. Good and Fair are combined as Good.

Clarity can be re-categorised into IF,VVS1,VS1,Sl1 and l1. VVS1 and VVS2 is combined as VVS1. VS1 and VS2 are combined as VS1. Sl1 and Sl2 are combined as Sl1.

Color can be re-categorised into D, F, H and J. E and F is combined as F. G and H are combined as H. I and J are combined as J.

Cut:

```
{'Fair': 0, 'Good': 0, 'Very Good': 1, 'Premium': 1, 'Ideal': 2
```

Color:

```
{'D': 0, 'E': 1, 'F': 1, 'G': 2, 'H': 2, 'I': 3, 'J': 3}
```

Clarity:

```
{'IF': 0,
 'VVS1': 1,
 'VVS2': 1,
 'VS1': 2,
 'VS2': 2,
 'SI1': 3,
 'SI2': 3,
 'I1': 4}
```

The coefficient is :

```
The coefficient for carat is 11012.268557051391
The coefficient for cut is 160.09275590511635
The coefficient for color is -561.6644837255302
The coefficient for clarity is -921.0631812094824
The coefficient for depth is -93.0516970843568
The coefficient for table is -32.66871489742376
The coefficient for x is -992.7095778770644
The coefficient for y is 0.9890563225999723
The coefficient for z is -31.37023889306916
```

The intercept for our model is 11273.041552645002

RMSE for train data is 1251.5258901665827
RMSE for test data is 1263.2575828987135

R-squared for train data is 0.9025664703645613
R-squared for test data is 0.9028374434338761

Adjusted R-squared for train data is 0.9025199330951923
Adjusted R-squared for test data is 0.9027290837053636

Scatterplot

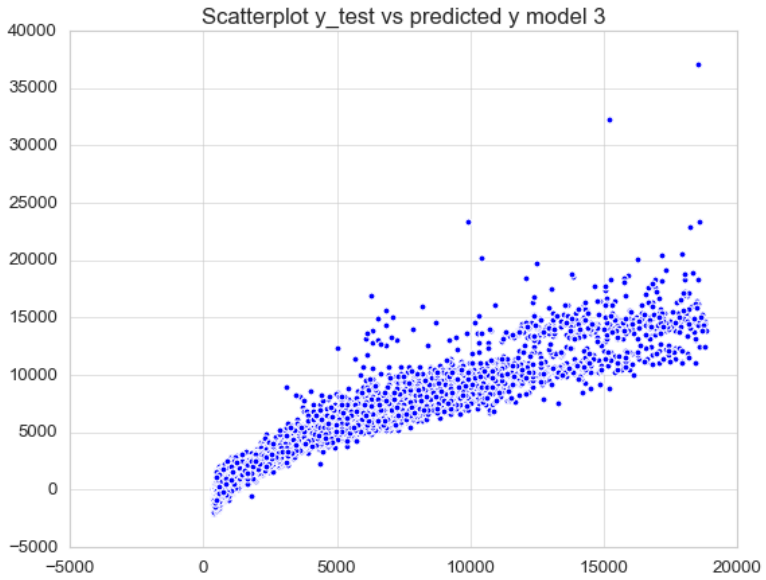

Figure 21:Model 3 - Scatterplot of y test with predicted y

The scatterplot is linear in nature.

Model 4 – OLS Combined Encoding with Outliers

From the correlation table we can see that the depth variable has a very low correlation(-0.0029) with price which indicates it isn't very important to predict the target variable. Hence it is removed in the expression of OLS.

```
'price ~ carat + cut + color + clarity +  table + x + y + z'
```

OLS Regression Results

Dep. Variable:	price	R-squared:	0.902
Model:	OLS	Adj. R-squared:	0.902
Method:	Least Squares	F-statistic:	2.165e+04
Date:	Sun, 10 Apr 2022	Prob (F-statistic):	0.00
Time:	16:26:02	Log-Likelihood:	-1.6128e+05
No. Observations:	18853	AIC:	3.226e+05
Df Residuals:	18844	BIC:	3.226e+05
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3643.3341	319.317	11.410	0.000	3017.445	4269.223
carat	1.076e+04	94.199	114.194	0.000	1.06e+04	1.09e+04
cut	224.7648	16.552	13.579	0.000	192.321	257.209
color	-566.1654	10.689	-52.966	0.000	-587.117	-545.214
clarity	-933.6202	11.801	-79.111	0.000	-956.752	-910.488
table	-7.8544	4.922	-1.596	0.111	-17.501	1.793
x	-812.8654	50.182	-16.198	0.000	-911.226	-714.505
y	22.1858	24.564	0.903	0.366	-25.962	70.333
z	-183.3806	40.955	-4.478	0.000	-263.657	-103.105

Omnibus:	3831.737	Durbin-Watson:	1.980
Prob(Omnibus):	0.000	Jarque-Bera (JB):	111938.200
Skew:	0.269	Prob(JB):	0.00
Kurtosis:	14.925	Cond. No.	2.05e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.05e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Table 10: Model 4 - OLS Results

Coefficients having p-values less than alpha(0.05) are statistically significant. Carat,cut,color,clarity,z and x are significant variables.

RMSE for train data is 1255.9575508945982
RMSE for test data is 1268.2762022065328

R-squared for train data is 0.902
R-squared for test data is 0.9020639028806737

Adjusted R-squared for train data is 0.902
Adjusted R-squared for test data is 0.9019546804675295

Scatterplot

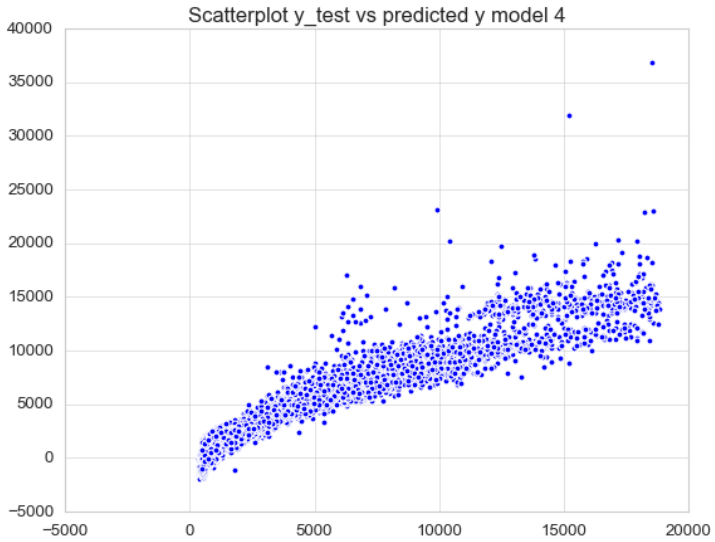

Figure 22:Model 4 - Scatterplot of y test with predicted y

The scatterplot is linear in nature.

Model 5 – Sklearn LR-without Outliers - initial encoding

The outliers are removed and Linear Regression model is built.

```
LinearRegression()
```

The coefficients are:

```
The coefficient for carat is 13713.21426177974
The coefficient for cut is 133.3457039610511
The coefficient for color is -333.8268465840487
The coefficient for clarity is -483.911960800029
The coefficient for depth is 3.9239010421218254
The coefficient for table is -28.035878741661772
The coefficient for x is -2347.4235932445317
The coefficient for y is 1544.5033054305238
The coefficient for z is -1819.3683369303121
```

The intercept for our model is 7854.618473168935

RMSE for train data is 1158.5516819397058
RMSE for test data is 1115.6454874564984

R-squared for train data is 0.9165051703009585
R-squared for test data is 0.896867478747077

Adjusted R-squared for train data is 0.9164652905860886
Adjusted R-squared for test data is 0.8967524610653822

Scatterplot

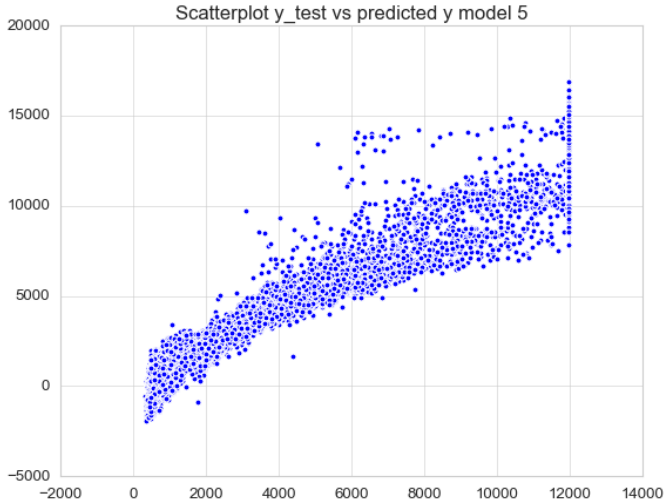

Figure 23:Model 5 - Scatterplot of y test with predicted y

The scatterplot is linear in nature.

Model 6 - OLS without Outliers-initial encoding

OLS model is built without Outliers with initial encoding with increasing order

```
'price ~ carat + cut + color + clarity + depth + table + x + y + z'
```

OLS Regression Results

Dep. Variable:	price	R-squared:	0.917
Model:	OLS	Adj. R-squared:	0.916
Method:	Least Squares	F-statistic:	2.298e+04
Date:	Sun, 10 Apr 2022	Prob (F-statistic):	0.00
Time:	16:26:05	Log-Likelihood:	-1.5976e+05
No. Observations:	18853	AIC:	3.195e+05
Df Residuals:	18843	BIC:	3.196e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	7854.6185	1012.808	7.755	0.000	5869.425	9839.812
carat	1.371e+04	105.138	130.431	0.000	1.35e+04	1.39e+04
cut	133.3457	9.370	14.231	0.000	114.980	151.712
color	-333.8268	5.239	-63.719	0.000	-344.096	-323.558
clarity	-483.9120	5.686	-85.111	0.000	-495.056	-472.768
depth	3.9239	14.033	0.280	0.780	-23.583	31.431
table	-28.0359	4.979	-5.630	0.000	-37.796	-18.276
x	-2347.4236	157.417	-14.912	0.000	-2655.975	-2038.873
y	1544.5033	155.131	9.956	0.000	1240.433	1848.574
z	-1819.3683	174.294	-10.438	0.000	-2161.001	-1477.736

Omnibus:	3678.694	Durbin-Watson:	1.976
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31292.476
Skew:	0.700	Prob(JB):	0.00
Kurtosis:	9.154	Cond. No.	1.03e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.03e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Table 11:Model 6 - OLS Results

Coefficients having p-values less than alpha(0.05) are statistically significant. Carat,cut,color,clarity,z,y,table and x are significant variables.

RMSE for train data is 1158.551681939704
RMSE for test data is 1115.6454874564613

R-squared for train data is 0.917
R-squared for test data is 0.8968674787470838

Adjusted R-squared for train data is 0.916
Adjusted R-squared for test data is 0.8967524610653892

Scatterplot

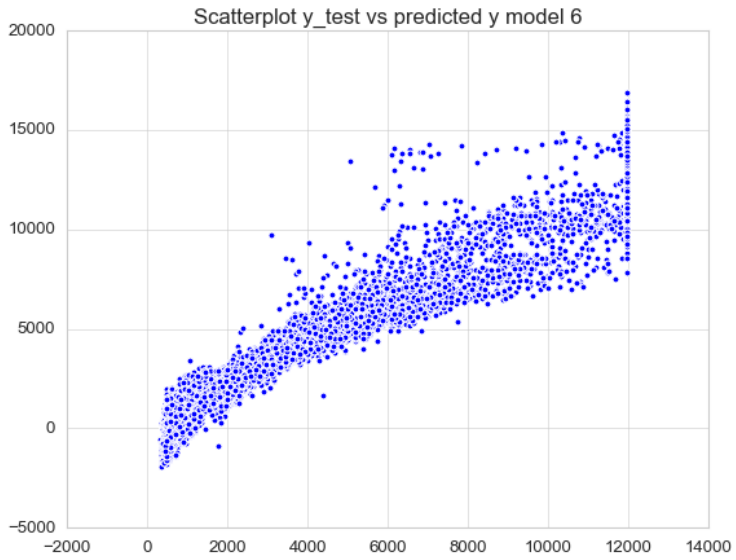

Table 12:Model 6- Scatterplot of y test with predicted y

The scatterplot is linear in nature.

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

The performance metrics of each model is shown in the below table.

	Model 1 Train	Model 1 Test	Model 2 Train	Model 2 Test	Model 3 Train	Model 3 Test	Model 4 Train	Model 4 Test	Model 5 Train	Model 5 Test	Model 6 Train	Model 6 Test
RMSE	1211.378	1229.214	1211.378	1229.214	1251.526	1263.258	1255.958	1268.276	1158.552	1115.645	1158.552	1115.645
R-squared	0.909	0.908	0.909	0.908	0.903	0.903	0.902	0.902	0.917	0.897	0.917	0.897
Adj R-squared	0.909	0.908	0.909	0.909	0.903	0.903	0.902	0.902	0.916	0.897	0.916	0.897

Table 13: Performance Metrics - LR

Both the sklearn and ols models are giving better results after treating the outliers. Hence Model 5 and 6 are the most optimised.R-squared 92.00% shows a good accuracy which means 92% of the price is explained by the model.

Business Insights and Recommendations

Increase in carat weight of the diamond will increase the price of the diamond considerably.

Width of the diamond in mm also plays an important factor. As the width increases the price also increases. Length and height has the effect on price.

Brighter the color of the diamond, the price increases.

Gem Stones Ltd can collect more data which helps in building better models for price prediction.

Most of the cubic zirconia with Very Good cut ,clarity VVS2 has a higher price range. For getting a high profit share the company can concentrate on these factors

The 5 most important attributes are carat,length,width,color and clarity.

PROBLEM 2 : Logistic Regression and LDA

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.

Introduction

The purpose of this whole exercise is to perform exploratory data analysis and predict whether an employee will opt for the package or not on the basis of the information given in the data set.The prediction is done using Logistic Regression and Linear Discriminant Analysis(LDA). Using the predictions recommendations are to be provided to the management. The dataset consists of 872 rows with their features like Salary, Age, Edu, Channel, no_young_children , no_older_children , Foreign and Holiday_Package.

Data Description

1. Target variable- Holiday_Package: Opted for Holiday Package (Yes/No)
2. Salary : Employee salary
3. Age : Age in years
4. Edu : Years of formal education
5. no_young_children : The number of young children (younger than 7 years)
6. no_older_children : Number of older children
7. 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.

Sample of the dataset

	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 14: Sample of the Dataset2

The data is read from the csv file and the above tables shows the first 5 rows of the dataset.

Here Holliday_Package is the Target variable which is categorical. Hence classification using Logistic Regression and Linear Discriminant Analysis is done. The first row isn't important and so it is dropped from the dataset.

EXPLORATORY DATA ANALYSIS

Data Type and Missing Values

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Holliday_Package   872 non-null    object
 1   Salary             872 non-null    int64
 2   age                872 non-null    int64
 3   educ               872 non-null    int64
 4   no_young_children  872 non-null    int64
 5   no_older_children  872 non-null    int64
 6   foreign            872 non-null    object
dtypes: int64(5), object(2)
memory usage: 47.8+ KB
```

Table 15:Dataset Info2

Holiday _Package and foreign are of object data type. Remaining features are of numeric datatype. There are 872 entries with 7 columns of which

Holiday_Package is a target variable. There are 872 entries non-null values for all columns which means there are no missing values

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

There are no missing values and duplicates.

Description of dataset

We can see that 75% of the data in no_young_children are 0s. There are 2 categories of foreign and holiday package with most frequent values as 'no'.

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
count	872	872.000000	872.000000	872.000000	872.000000	872.000000	872
unique	2	NaN	NaN	NaN	NaN	NaN	2
top	no	NaN	NaN	NaN	NaN	NaN	no
freq	471	NaN	NaN	NaN	NaN	NaN	656
mean	NaN	47729.172018	39.955275	9.307339	0.311927	0.982798	NaN
std	NaN	23418.668531	10.551675	3.036259	0.612870	1.086786	NaN
min	NaN	1322.000000	20.000000	1.000000	0.000000	0.000000	NaN
25%	NaN	35324.000000	32.000000	8.000000	0.000000	0.000000	NaN
50%	NaN	41903.500000	39.000000	9.000000	0.000000	1.000000	NaN
75%	NaN	53469.500000	48.000000	12.000000	0.000000	2.000000	NaN
max	NaN	236961.000000	62.000000	21.000000	3.000000	6.000000	NaN

Table 16: Description of Dataset2

Univariate Analysis

Boxplot

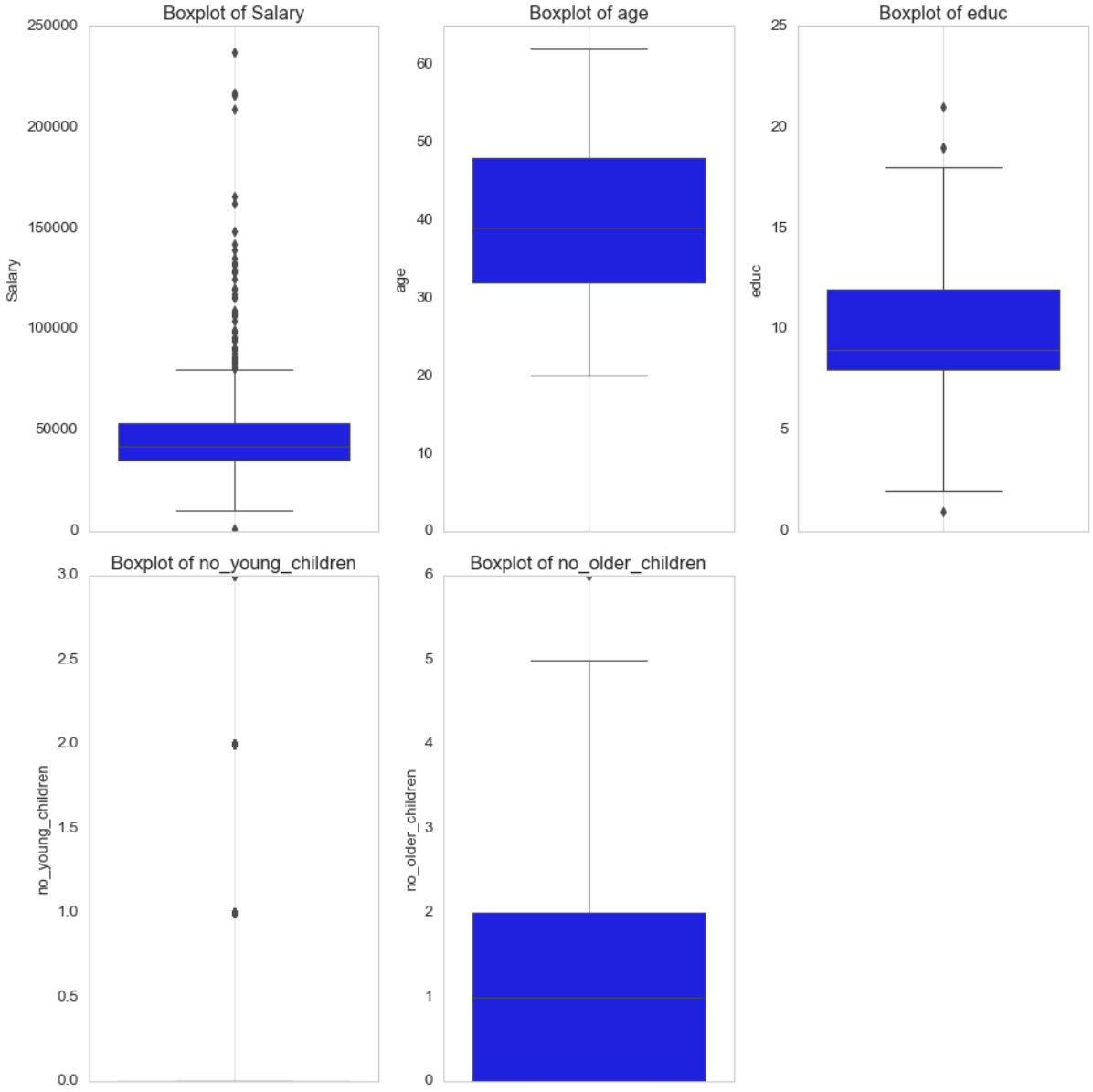

Figure 24:Univariate Analysis - Boxplot

There are outliers in all the features except age which need not be treated as they look like genuine values. Median of age is around 40. no_older_children has a median of 1. educ has a median around 9. Salary has a median around 40000.

Proportion of Outliers

```
Lower outliers in Salary is :  8105.75
Upper outliers in Salary is :  80687.75
Number of outliers in Salary upper :  56
Number of outliers in Salary lower :  1
% of Outlier in Salary upper:  6 %
% of Outlier in Salary lower:  0 %
--------------------------------------------------
Lower outliers in age is :  8.0
Upper outliers in age is :  72.0
Number of outliers in age upper :  0
Number of outliers in age lower :  0
% of Outlier in age upper:  0 %
% of Outlier in age lower:  0 %
--------------------------------------------------
Lower outliers in educ is :  2.0
Upper outliers in educ is :  18.0
Number of outliers in educ upper :  3
Number of outliers in educ lower :  1
% of Outlier in educ upper:  0 %
% of Outlier in educ lower:  0 %
--------------------------------------------------
Lower outliers in no_young_children is :  0.0
Upper outliers in no_young_children is :  0.0
Number of outliers in no_young_children upper :  207
Number of outliers in no_young_children lower :  0
% of Outlier in no_young_children upper:  24 %
% of Outlier in no_young_children lower:  0 %
--------------------------------------------------
Lower outliers in no_older_children is :  -3.0
Upper outliers in no_older_children is :  5.0
Number of outliers in no_older_children upper :  2
Number of outliers in no_older_children lower :  0
% of Outlier in no_older_children upper:  0 %
% of Outlier in no_older_children lower:  0 %
--------------------------------------------------
```

The above image shows the proportion of outliers of each feature in the dataset.

Analysis of no_young_children:

```
Value Count of no_young_children
```

```
--------------------------------------------------------------------------------
0    665
1    147
2     55
3      5
Name: no_young_children, dtype: int64

Description of no_young_children
--------------------------------------------------------------------------------
count    872.000000
mean       0.311927
std        0.612870
min        0.000000
25%        0.000000
50%        0.000000
75%        0.000000
max        3.000000
Name: no_young_children, dtype: float64

Countplot of no_young_children
--------------------------------------------------------------------------------
-
```

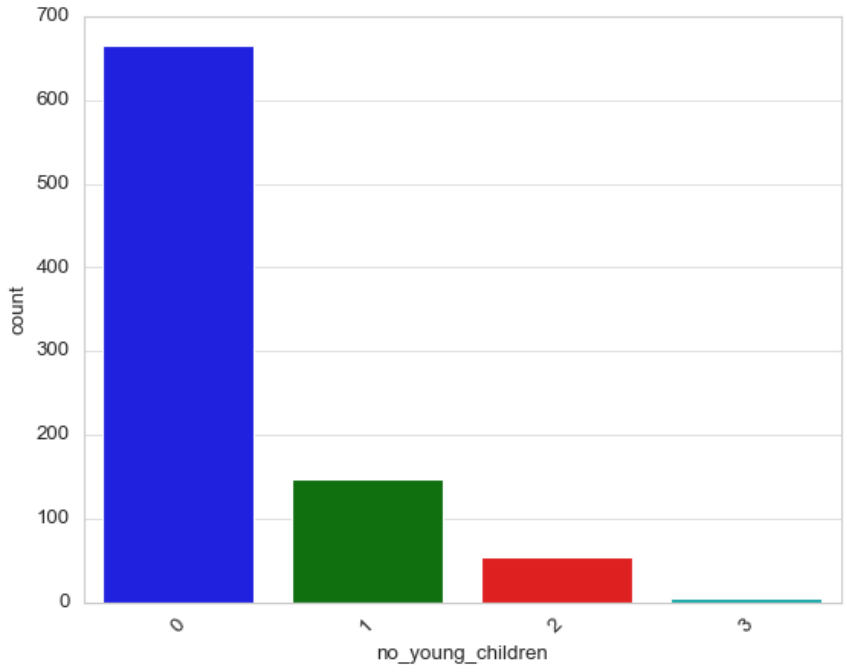

Figure 25:Countplot of no_young_children

```
Interquartile range (IQR) of is  0.0
Range of values:  3

Distribution of no_young_children
--------------------------------------------------------------------------------
```

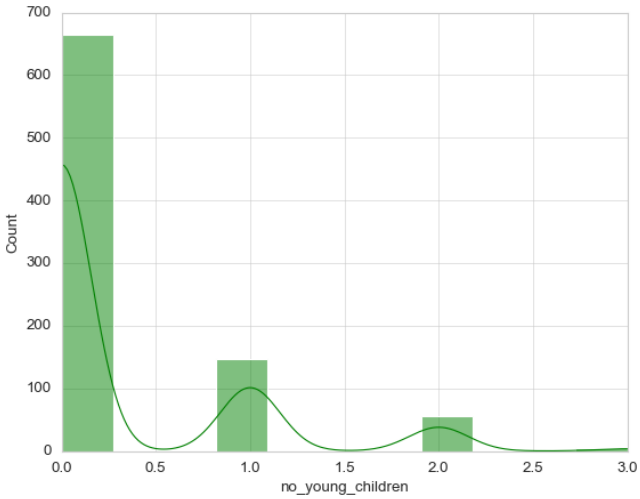

Figure 26: Distribution of no_young_children

Most of the employees have no children younger than 7 years. This feature is extremely right skewed.

Analysis of no_older_children

```
Value Count of no_older_children
--------------------------------------------------------------------------------
-
0    393
2    208
1    198
3     55
4     14
5      2
6      2
Name: no_older_children, dtype: int64

Description of no_older_children
--------------------------------------------------------------------------------
-
count    872.000000
mean       0.982798
std        1.086786
min        0.000000
25%        0.000000
50%        1.000000
75%        2.000000
max        6.000000
Name: no_older_children, dtype: float64

Countplot of no_older_children
```

-

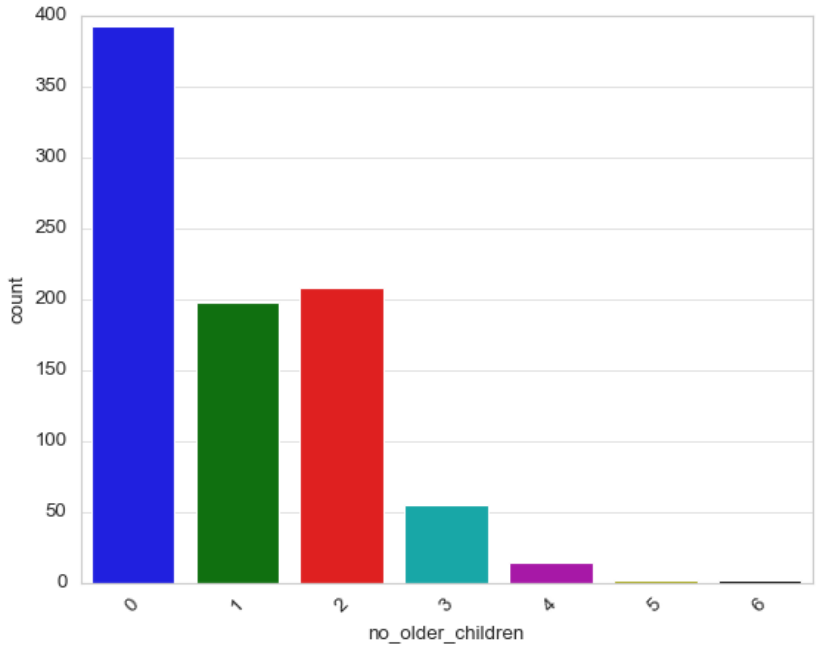

Figure 27: Countplot of no_older_children

```
Interquartile range (IQR) of is  2.0
Range of values:  6

Distribution of no_older_children
```

-

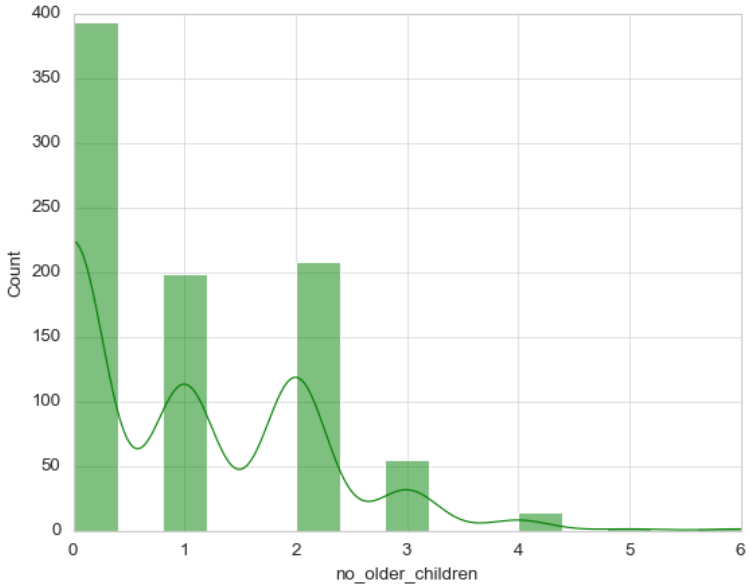

Figure 28: Distribution of no_older_children

Most of the employees have no older children and few of them have 4 older children. This feature is slightly right skewed.

Analysis of foreign

```
Value Count of foreign
--------------------------------------------------------------------------------
-
no      656
yes     216
Name: foreign, dtype: int64

Description of foreign
--------------------------------------------------------------------------------
-
count     872
unique      2
top        no
freq      656
Name: foreign, dtype: object

Countplot of foreign
--------------------------------------------------------------------------------
-
```

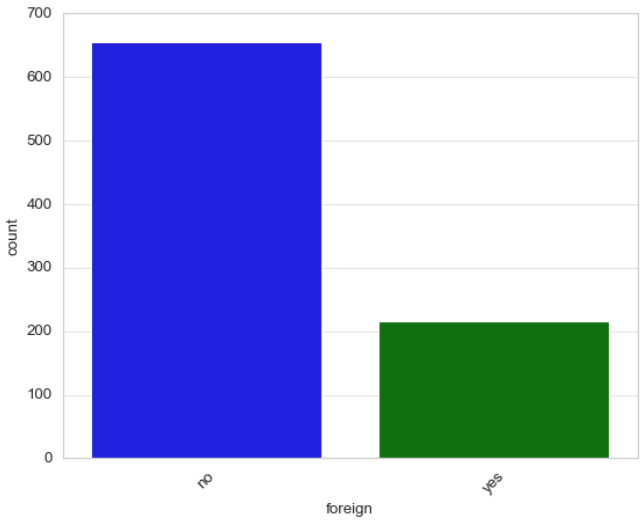

Figure 29: Countplot of foreign

Most of them aren't foreigners.

Analysis of Holliday_Package

```
Value Count of Holliday_Package
```

```
------------------------------------------------------------------------------
no     471
yes    401
Name: Holliday_Package, dtype: int64

Description of Holliday_Package
------------------------------------------------------------------------------
-
count     872
unique      2
top        no
freq      471
Name: Holliday_Package, dtype: object

Countplot of Holliday_Package
------------------------------------------------------------------------------
-
```

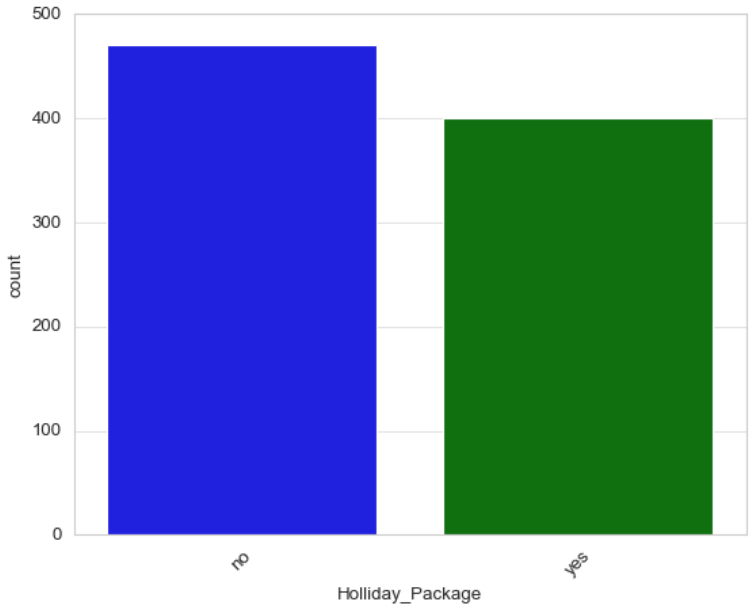

Figure 30: Countplot of Holliday_Package

Mostly people didn't choose the package.

Analysis of Salary

```
Description of Salary
------------------------------------------------------------------------------
count       872.000000
mean      47729.172018
std       23418.668531
min        1322.000000
25%       35324.000000
50%       41903.500000
```

```
75%       53469.500000
max      236961.000000
Name: Salary, dtype: float64

Interquartile range (IQR) of is  18145.5
Range of values:  235639

Distribution of Salary
------------------------------------------------------------------------------
-
```

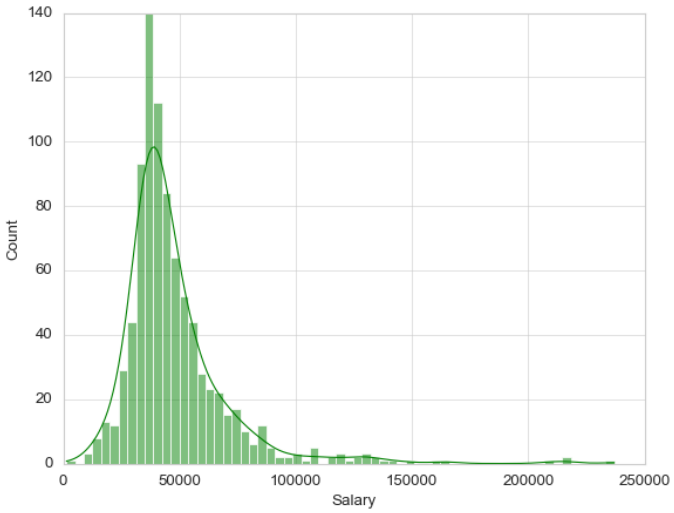

Figure 31:Distribution of Salary

Salary is extremely right skewed.

Analysis of age

```
Description of age
------------------------------------------------------------------------------
-
count    872.000000
mean      39.955275
std       10.551675
min       20.000000
25%       32.000000
50%       39.000000
75%       48.000000
max       62.000000
Name: age, dtype: float64

Interquartile range (IQR) of is  16.0
Range of values:  42

Distribution of age
```

```
-------------------------------------------------------------------------------
-
```

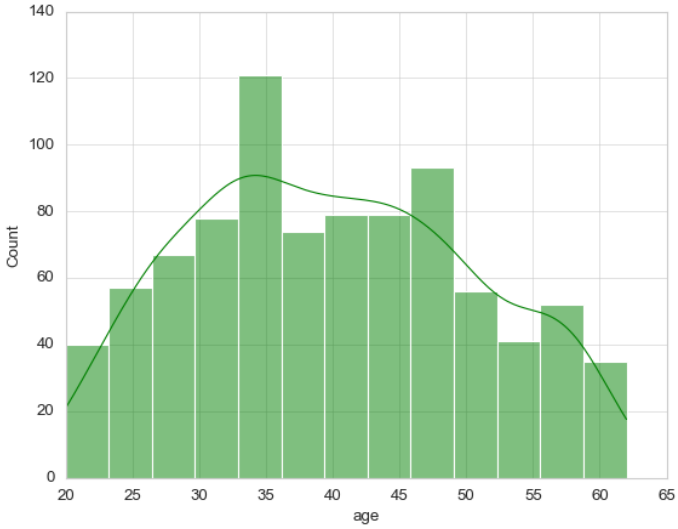

Figure 32:Distribution of age

Age is nearly symmetrical in distribution.

Analysis of educ

```
Description of educ
-------------------------------------------------------------------------------
-
count    872.000000
mean       9.307339
std        3.036259
min        1.000000
25%        8.000000
50%        9.000000
75%       12.000000
max       21.000000
Name: educ, dtype: float64

Interquartile range (IQR) of is  4.0
Range of values:  20

Distribution of educ
-------------------------------------------------------------------------------
```

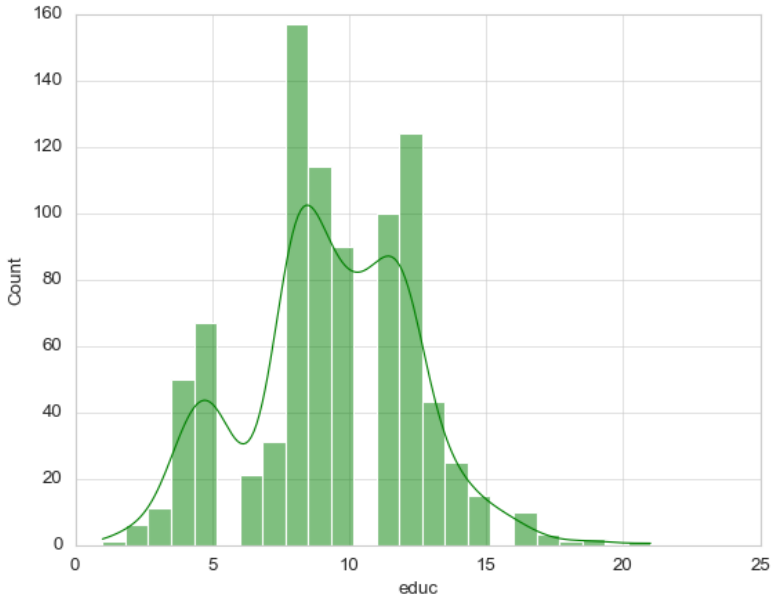

Figure 33:Distribution of educ

Educ is nearly symmetrically distributed.

Skewsness and Kurtosis

```
Skewness of Salary is 3.1
Kurtosis of Salary is 15.85
Skewness of age is 0.15
Kurtosis of age is -0.91
Skewness of educ is -0.05
Kurtosis of educ is 0.01
Skewness of no_young_children is 1.95
Kurtosis of no_young_children is 3.11
Skewness of no_older_children is 0.95
Kurtosis of no_older_children is 0.68
```

Skewness essentially measures the symmetry of the distribution. In positively skewed, the mean of the data is greater than the median as a large number of data-pushed on the right-hand side. In negatively skewed, the mean of the data is less than the median as a large number of data-pushed on the left-hand.

If the skewness is between -0.5 & 0.5, the data are nearly symmetrical. If the skewness is between -1 & -0.5 (negative/left skewed) or between 0.5 & 1(positive/right skewed), the data are slightly skewed. If the skewness is lower than -1 (negative/left skewed) or greater than 1 (positive/right skewed), the data are extremely skewed.

Kurtosis refers to the degree of presence of outliers in the distribution. If kurtosis > 3, then it is called as Leptokurtic or heavy-tailed distribution as the kurtosis is more than normal distribution. If kurtosis = 3, then it is called as Mesokurtic as the kurtosis is same as the normal distribution. If kurtosis < 3, then it is called as Platykurtic or short-tailed distribution as the kurtosis is less than normal distribution.

Bivariate Analysis

Countplot

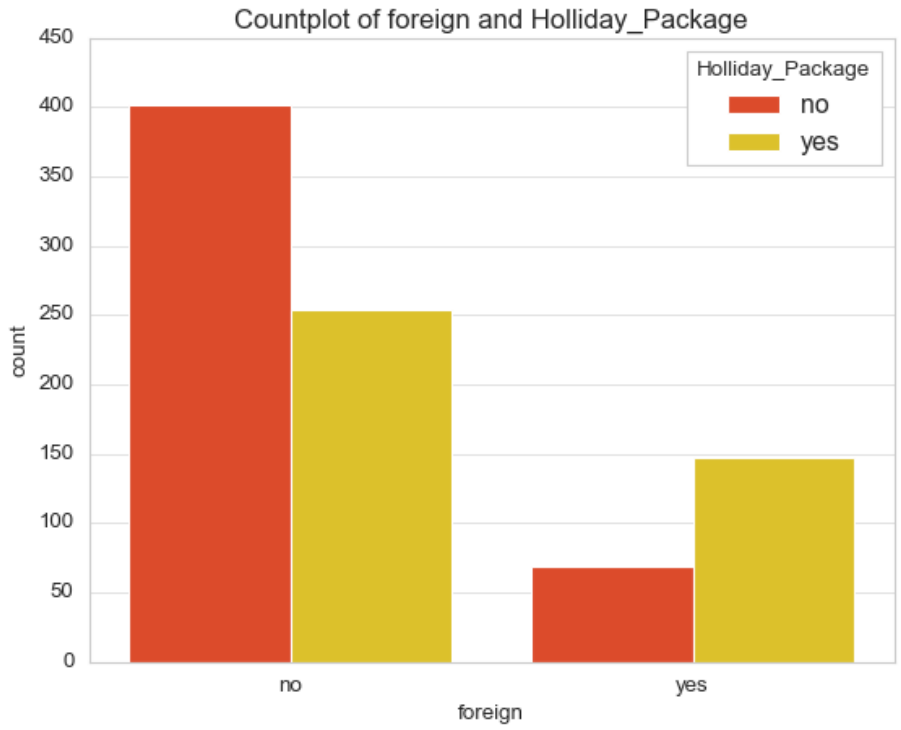

Figure 34:Countplot of Holliday_Package and foreign

Non-foreigners opt the holiday package more than the foreigners

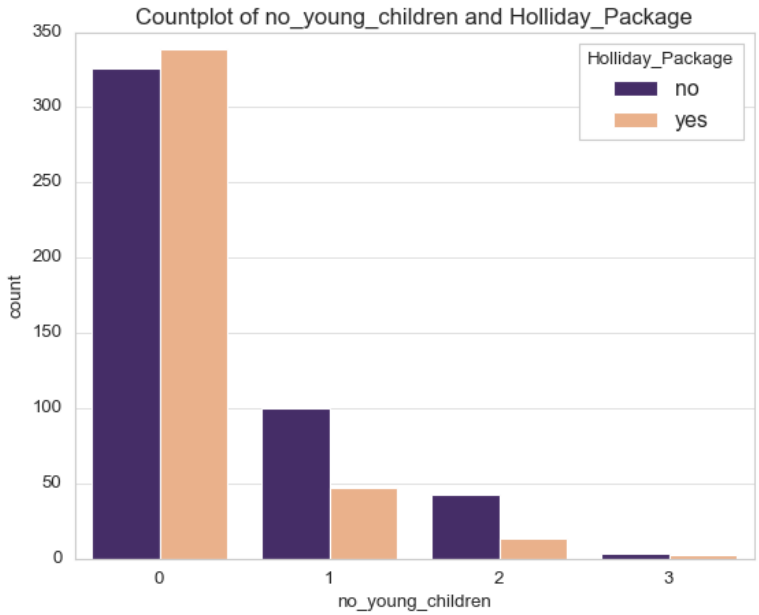

Figure 35:Countplot of Holliday_Package and no_young_children

People with no younger children are willing to opt for the package than the people with younger children

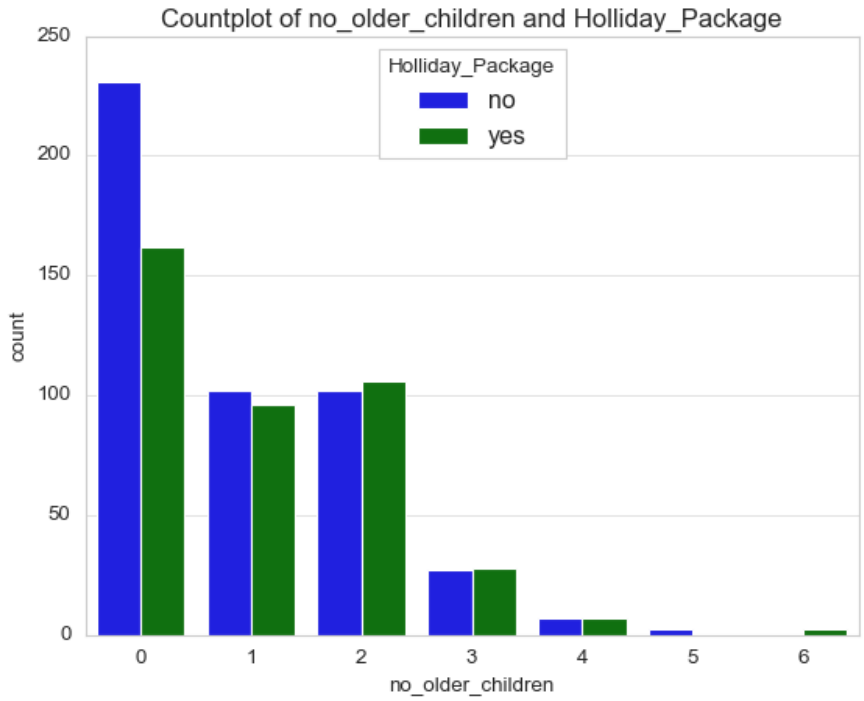

Figure 36:Countplot of Holliday_Package and no_older_children

People with 2,3 or 4 older children are willing to opt package than not to.

Boxplot

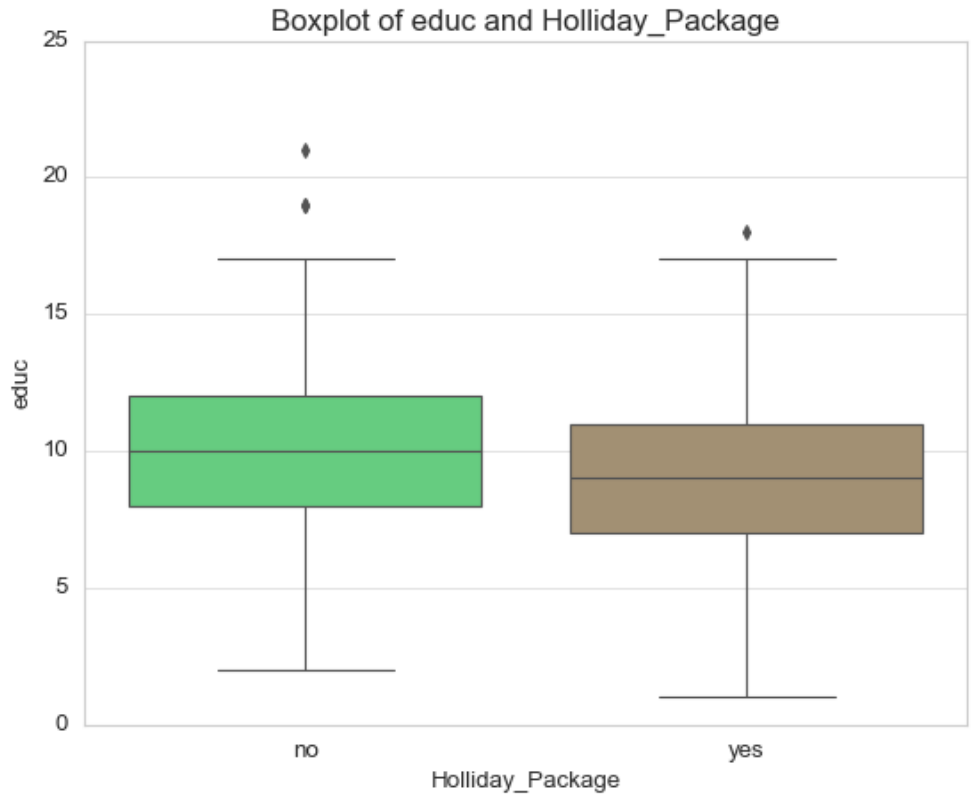

Figure 37:Boxplot of Holliday_Package and educ

Years of formal education doesn't have too much effect on the decision of opting the holiday package as the median value is nearly the same for both category.

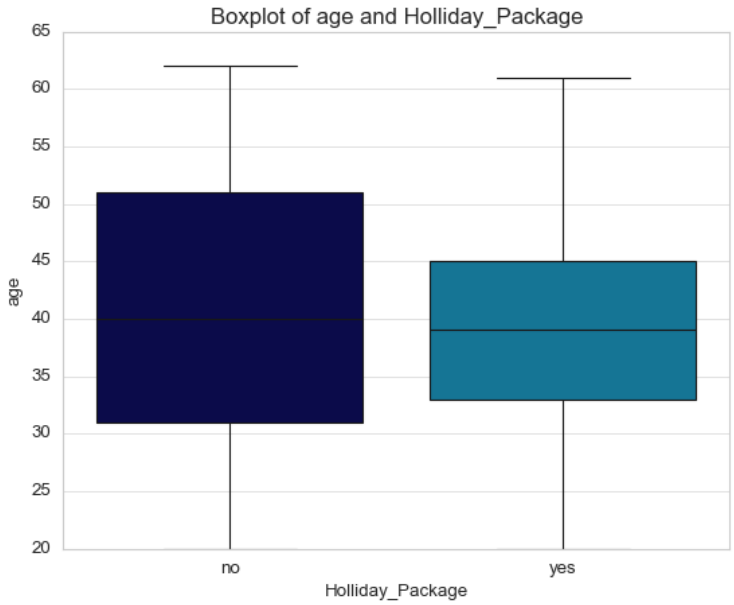

Figure 38: Boxplot of age and Holliday_Package

Most employees above age 45 doesn't opt for the holiday package

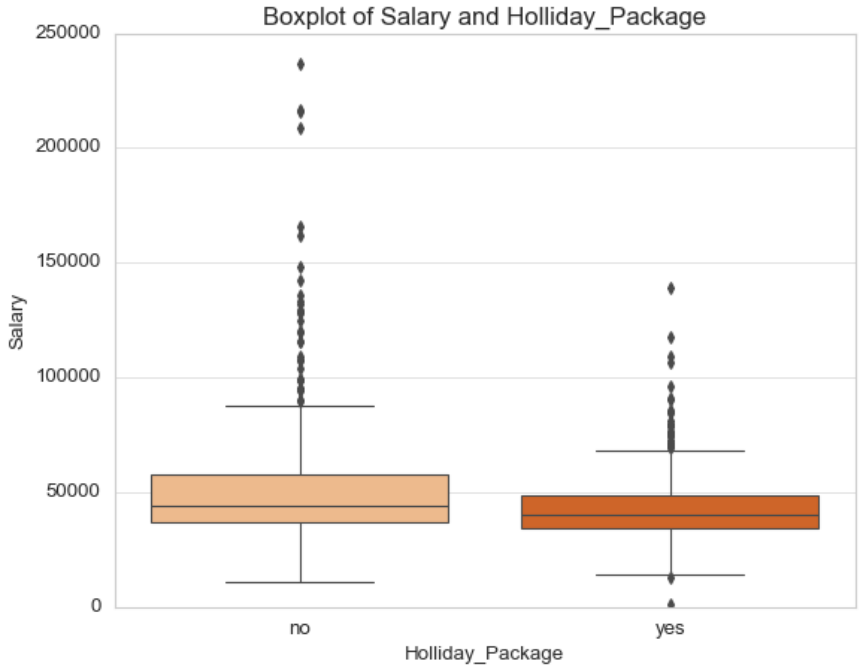

Figure 39:Boxplot of Holliday_Package and Salary

Employees having high salary(above 50000) mostly don't opt for holiday package

PAIRPLOT

Figure 40:Pairplot2

As the age increases the people with younger children opting for the holiday package decreases.

CORRELATION HEATMAP

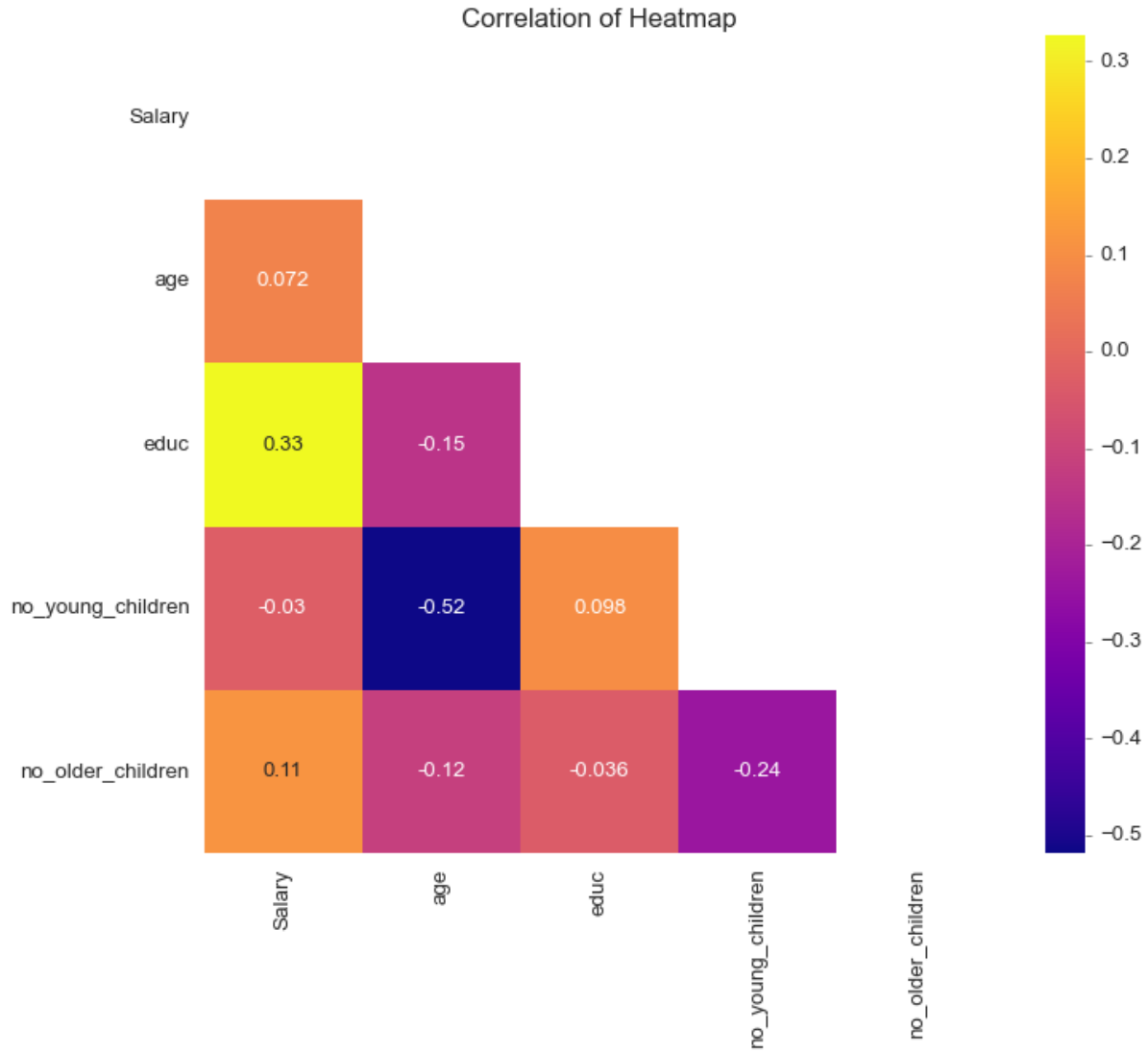

Figure 41:Correlation Heatmap2

From the heatmap and pairplot we can say that age has a significant negative correlation with no_young_children which explains the fact that as a person ages the number of younger children they have decreases.

Multivariate Analysis

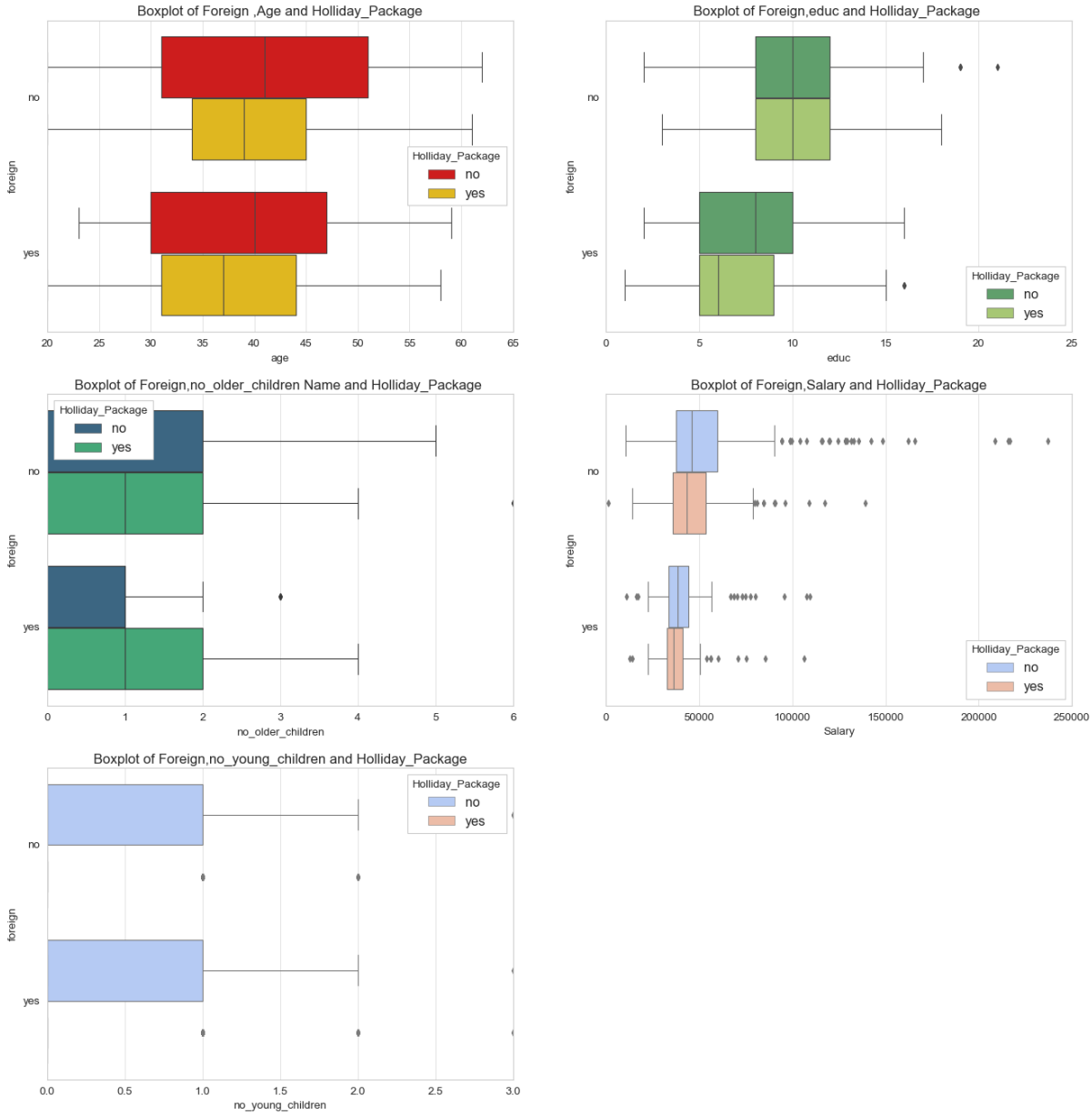

Figure 42:Multivariate Analysis – LDA and LogReg

Foreigners with less older children are not opting the package. Foreigners willing to opt for package has less years in formal education than non foreigners.

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)

Target variable Holliday_Package is encoded using LabelEncoder(yes-1,no-0). The sample data after encoding is:

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

Table 17: LabelEncoding of Holliday_Package

Categorical encoding is done for foreign variable. The sample data is as follows:

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	0	48412	30	8	1	1	0
1	1	37207	45	8	0	1	0
2	0	58022	46	9	0	0	0
3	0	66503	31	11	2	0	0
4	0	66734	44	12	0	2	0

Table 18: Categorical encoding of foreign

Splitting of data into training and test data is done at 70:30 ratio. The proportion of target classes in the whole data,train and test data is the same.

```
y labels value counts percentage:
  0    0.540138
1    0.459862
Name: Holliday_Package, dtype: float64
 train labels value counts:
 0    0.534426
1    0.465574
Name: Holliday_Package, dtype: float64
 test labels value counts:
 0    0.553435
1    0.446565
Name: Holliday_Package, dtype: float64
```

Model 1 : Logistic Regression Default Model

```
LogisticRegression()
```

Solver parameter mentions the algorithm to use in the optimization problem. Default is ‘lbfgs’.The default penalty value is 'l2' which adds a L2 penalty term and it goes well with lbfgs solver.

Output of predicting test data is:

```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
       1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
       1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0,
       1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0])
```

Model 2 : Linear Discriminant Analysis(LDA) Default Model

```
LinearDiscriminantAnalysis()
```

In [56]:

Defaulter solver ‘svd’ is used. svd means Singular value decomposition which does not compute the covariance matrix, therefore this solver is recommended for data with a large number of features.

Output of predicting test data is:

```
array([0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1,
       1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0,
       0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1,
       1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1, 0,
       1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1,
       1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0,
       0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1,
       0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0,
       1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0,
       0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 0,
       0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1,
       0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0])
```

Model 3 : GridSearch - Logistic Regression

GridSearch helps to loop through predefined hyperparameters and fit your model on the training set and we can select the best combination of parameters from the listed hyperparameters.

```
GridSearchCV(cv=3, estimator=LogisticRegression(max_iter=10000, random_state=1),
             n_jobs=-1,
             param_grid={'penalty': ['l2', 'none', 'l1'],
                         'solver': ['sag', 'lbfgs', 'saga', 'newton-cg'],
                         'tol': [0.0001, 1e-05]},
             scoring='recall')
```

tol means Tolerance for stopping criteria. 'none' means no penalty is added. 'l2' means to add a L2 penalty term and it is the default choice. 'l1' means to add a L1 penalty term.

newton-cg is a solver which calculates Hessian explicitly which can be computationally expensive in high dimensions.

sag: Stands for Stochastic Average Gradient Descent. More efficient solver with large datasets.

saga: Saga is a variant of Sag and it can be used with l1 Regularization. It's a quite time-efficient solver and usually the go-to solver with very large datasets.

If n_jobs is set to -1, all processors are used.
Scoring strategy which is recall is used to evaluate the performance of the cross-validated model on the test set.

cv = 3 specifies the number of folds in a (Stratified)KFold.

max_iter = 10000 is the maximum number of iterations taken for the solvers to converge.

The best combination of parameters are:

```
{'penalty': 'l2', 'solver': 'newton-cg', 'tol': 0.0001}

LogisticRegression(max_iter=10000, random_state=1, solver='newton-cg')
```

The probability of test classes for each record in the test dataset.

	0	1
0	0.753599	0.246401
1	0.287308	0.712692
2	0.888743	0.111257
3	0.974783	0.025217
4	0.499096	0.500904

Table 19: LogReg Test data label probabilities

Model 4: GridSearch – LDA

```
GridSearchCV(cv=3, estimator=LinearDiscriminantAnalysis(), n_jobs=-1,
             param_grid={'solver': ['svd', 'lsqr', 'eigen']}, scoring='recall')
```

Solver : 'svd': Singular value decomposition (default). Does not compute the covariance matrix, therefore this solver is recommended for data with a large number of features.

'lsqr': Least squares solution. Can be combined with shrinkage or custom covariance estimator.

'eigen': Eigenvalue decomposition. Can be combined with shrinkage or custom covariance estimator.

If n_jobs is set to -1, all processors are used.
Scoring strategy which is recall is used to evaluate the performance of the cross-validated model on the test set.
cv = 3 specifies the number of folds in a (Stratified)KFold.

The best combination of parameters are:

```
{'solver': 'svd'}

LinearDiscriminantAnalysis()
```

The probability of test classes for each record in the test dataset.

	0	1
0	0.736312	0.263688
1	0.277893	0.722107
2	0.887243	0.112757
3	0.967803	0.032197
4	0.523170	0.476830

Table 20:LDA Test data label probabilities

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 *:*
TN / True Negative: when a case was negative and predicted negative(Top Left of Confusion Matrix)
TP / True Positive: when a case was positive and predicted positive(Bottom Right)
FN / False Negative: when a case was positive but predicted negative(Type 2 error) (Bottom Left)
FP / False Positive: when a case was negative but predicted positive(Type 1 error)(Top Right)

In this problem **False Negative** is an important metric as it denotes claimed cases as unclaimed ones which generates loss for the Insurance firm.

Classification Report
Precision is the ability of a classifier not to label an instance positive that is actually negative.

$$Precision = TP/(TP + FP)$$

Equation 1:Precision

Recall is the fraction of positives that were correctly identified.

$$Recall = TP/(TP+FN)$$

Equation 2:Recall

F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0.

$$F1\ Score = 2*(Recall * Precision) / (Recall + Precision)$$

Equation 3:F1 Score

Recall is the important metric as it is inversely proportional to Type 2(False Negative) error. In this problem decreasing the Type 2 error is important.

Model 1 - Logistic Regression Default Model

Classification Report – Train Data

	precision	recall	f1-score	support
0	0.53	0.90	0.67	326
1	0.42	0.08	0.14	284
accuracy			0.52	610
macro avg	0.47	0.49	0.40	610
weighted avg	0.48	0.52	0.42	610

Table 21:Classification Report-Model1-Train Data

Classification Report – Test Data

	precision	recall	f1-score	support
0	0.55	0.89	0.68	145
1	0.38	0.09	0.14	117
accuracy			0.53	262
macro avg	0.47	0.49	0.41	262
weighted avg	0.47	0.53	0.44	262

Table 22:Classification Report-Model1-Test Data

Confusion Matrix – Train Data

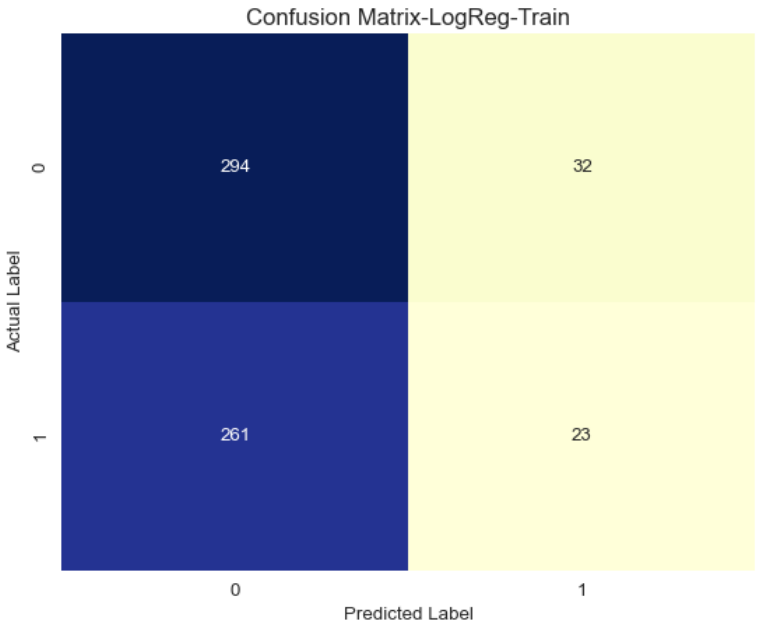

Figure 43:Confusion Matrix-Model1-Train Data

Confusion Matrix – Test Data

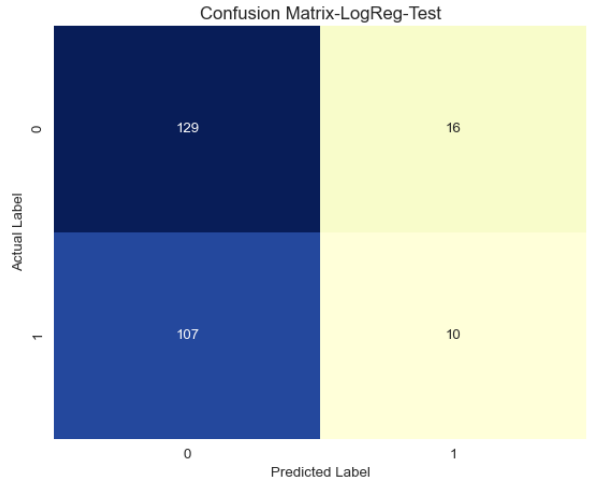

Figure 44:Confusion Matrix-Model1-Test Data

Accuracy for Train Data – 0.52

Accuracy for Test Data – 0.53

The accuracies are nearly the same. Hence there is no underfitting and Overfitting. It has a good score hence a good model.

ROC Curve and AUC Score – Train Data

AUC Score for logReg train data: 0.567

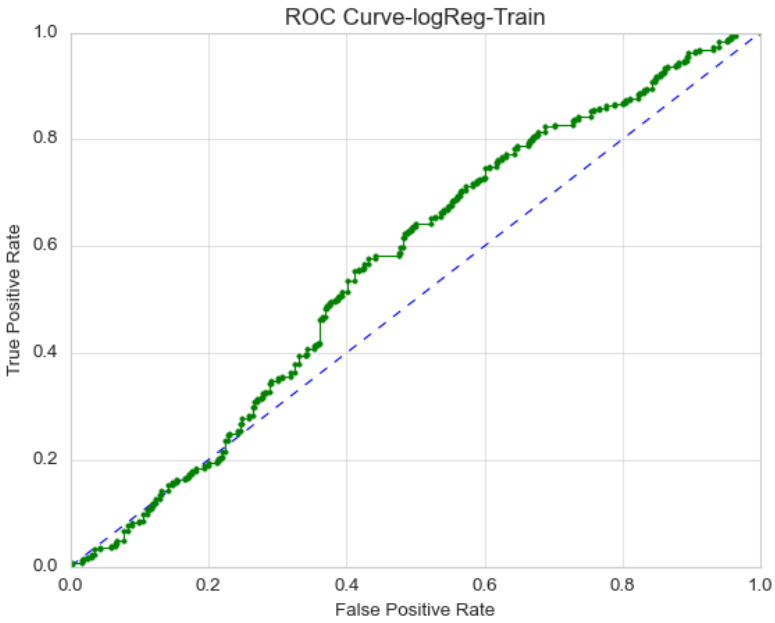

Figure 45:ROC Curve-Model1-Train Data

ROC Curve and AUC Score – Test Data

AUC Score for logReg test data: 0.627

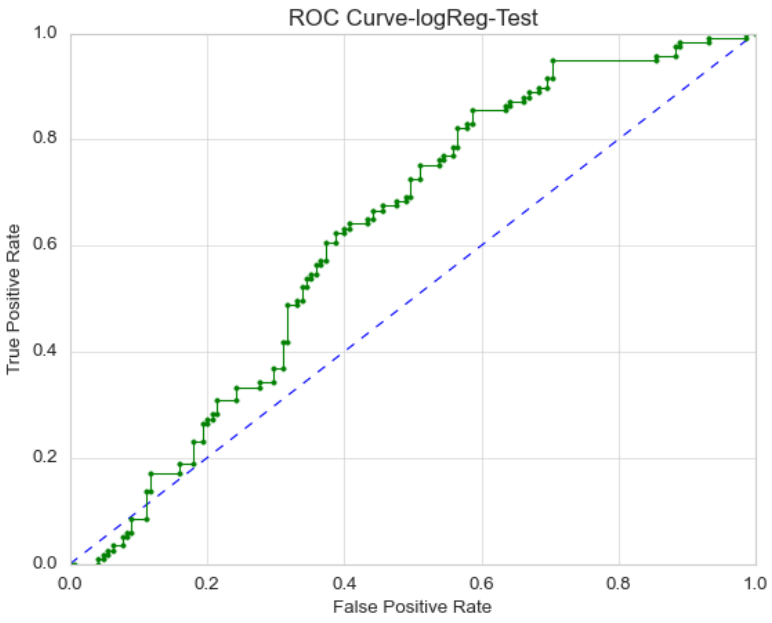

Figure 46:ROC Curve- Model1 -Test Data

Model 2 : Linear Discriminant Analysis(LDA) Default Model

Classification Report – Train Data

	precision	recall	f1-score	support
0	0.67	0.77	0.72	326
1	0.68	0.56	0.61	284
accuracy			0.67	610
macro avg	0.67	0.66	0.66	610
weighted avg	0.67	0.67	0.67	610

Table 23:Classification Report-Model2-Train Data

Classification Report – Test Data

	precision	recall	f1-score	support
0	0.66	0.71	0.69	145
1	0.61	0.56	0.58	117
accuracy			0.64	262
macro avg	0.64	0.63	0.63	262
weighted avg	0.64	0.64	0.64	262

Table 24:Classification Report-Model2-Test Data

Confusion Matrix – Train Data

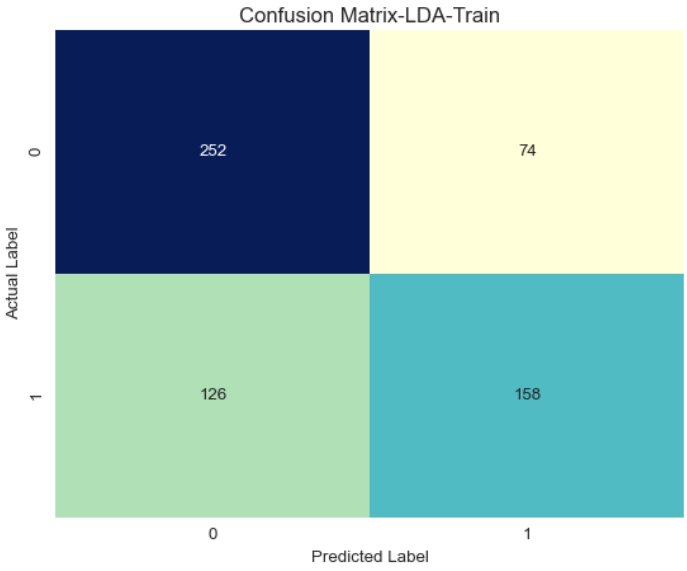

Figure 47: Confusion Matrix-Model2-Train Data

Confusion Matrix – Test Data

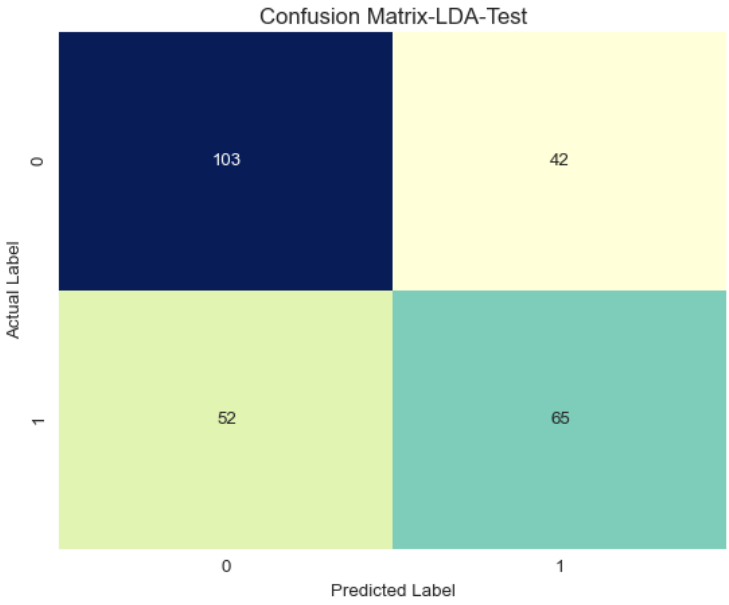

Figure 48:Confusion Matrix-Model2-Test Data

Accuracy for Train Data – 0.67

Accuracy for Test Data – 0.64

The accuracies are nearly the same. Hence there is no underfitting and Overfitting. It has a good score hence a good model.

ROC Curve and AUC Score – Train Data

AUC Score for LDA train data: 0.742

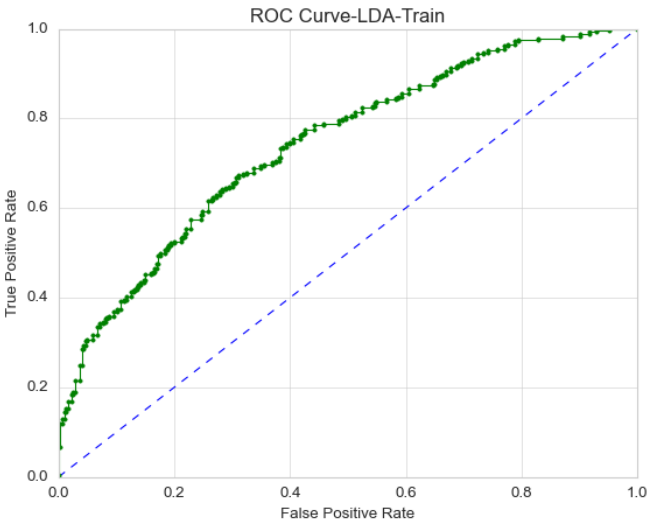

Figure 49:ROC Curve-Model2-Train Data

ROC Curve and AUC Score – Test Data

AUC Score for LDA test data: 0.703

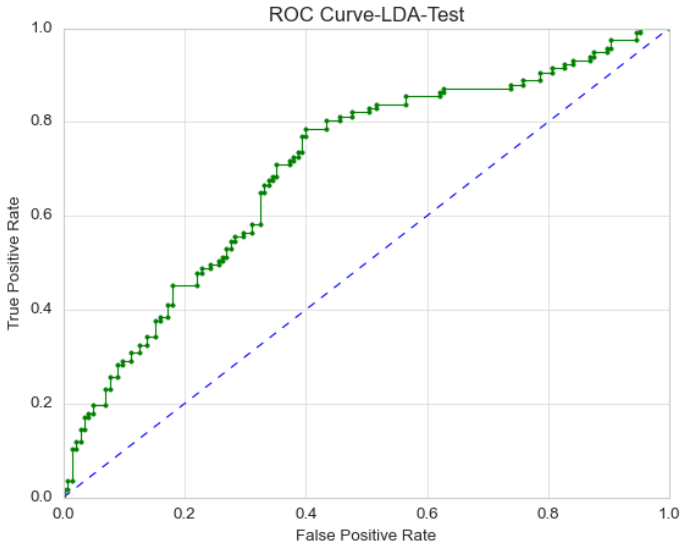

Figure 50:ROC Curve-Model2-Test Data

Model 3 : GridSearch - Logistic Regression

Classification Report – Train Data

	precision	recall	f1-score	support
0	0.68	0.77	0.72	326
1	0.69	0.57	0.63	284
accuracy			0.68	610
macro avg	0.68	0.67	0.67	610
weighted avg	0.68	0.68	0.68	610

Table 25:Classification Report-Model3-Train Data

Classification Report – Test Data

	precision	recall	f1-score	support
0	0.67	0.70	0.69	145
1	0.61	0.57	0.59	117
accuracy			0.65	262
macro avg	0.64	0.64	0.64	262
weighted avg	0.64	0.65	0.64	262

Table 26:Classification Report-Model3-Test Data

Confusion Matrix – Train Data

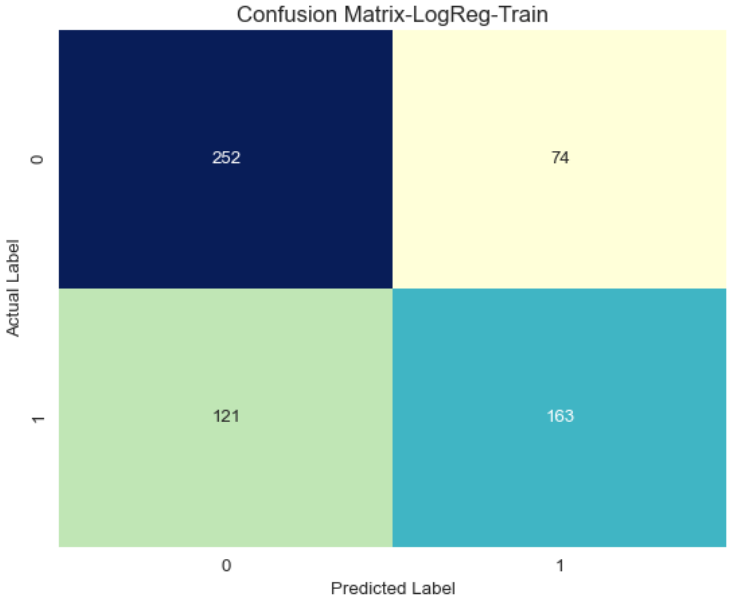

Figure 51: Confusion Matrix-Model3-Train Data

Confusion Matrix – Test Data

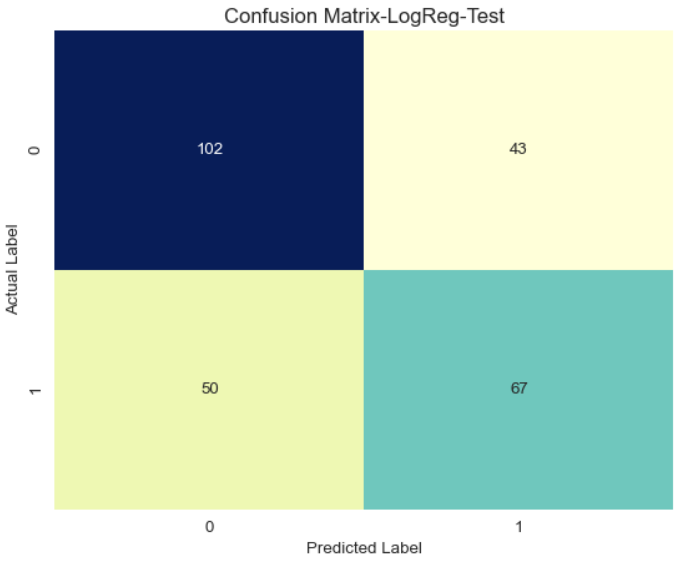

Figure 52:Confusion Matrix- Model3-Test Data

Accuracy for Train Data – 0.68

Accuracy for Test Data – 0.65

The accuracies are nearly the same. Hence there is no underfitting and Overfitting. It has a good score hence a good model.

ROC Curve and AUC Score – Train Data

AUC Score for logReg train data: 0.743

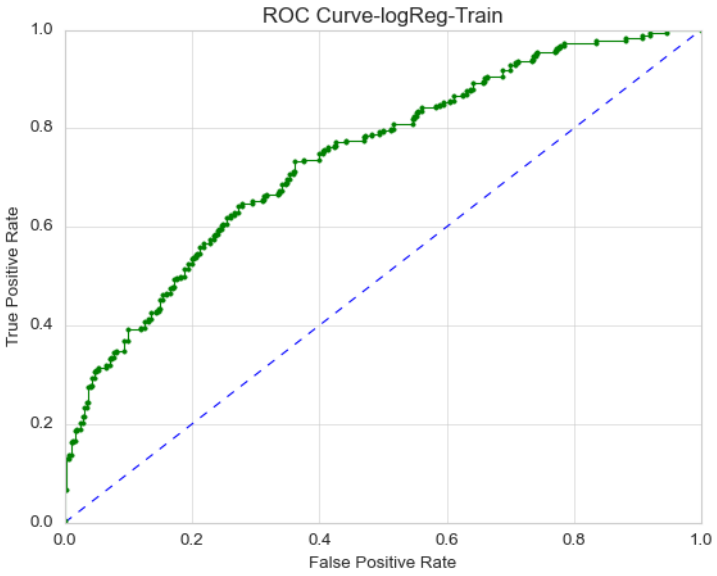

Figure 53:ROC Curve- Model3-Train Data

ROC Curve and AUC Score – Test Data

AUC Score for logReg test data: 0.705

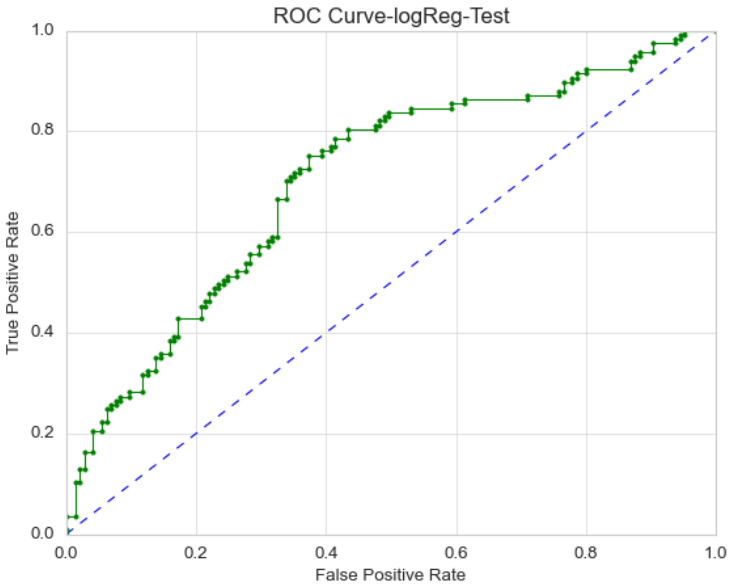

Figure 54: ROC Curve- Model3-Test Data

Model 4: GridSearch – LDA

Since GridSearch gave the model with default parameters the model is further built by trying different current value for probability. 0.4 is chosen as the best values based on a good balance of precision, recall and f1 score.

Classification Report – Train Data

Classification Report of the custom cut-off train data:

	precision	recall	f1-score	support
0	0.73	0.60	0.66	326
1	0.62	0.75	0.68	284
accuracy			0.67	610
macro avg	0.67	0.67	0.67	610
weighted avg	0.68	0.67	0.67	610

Table 27:Classification Report-Model4-Train Data

Classification Report – Test Data

	precision	recall	f1-score	support
0	0.77	0.59	0.67	145
1	0.61	0.79	0.69	117
accuracy			0.68	262
macro avg	0.69	0.69	0.68	262
weighted avg	0.70	0.68	0.68	262

Table 28:Classification Report-Model4-Test Data

Confusion Matrix – Train Data

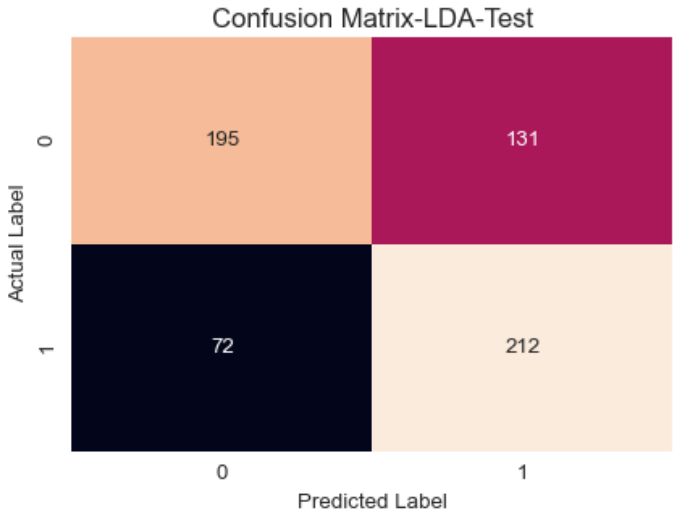

Figure 55: Confusion Matrix-Model4-Train Data

Confusion Matrix – Test Data

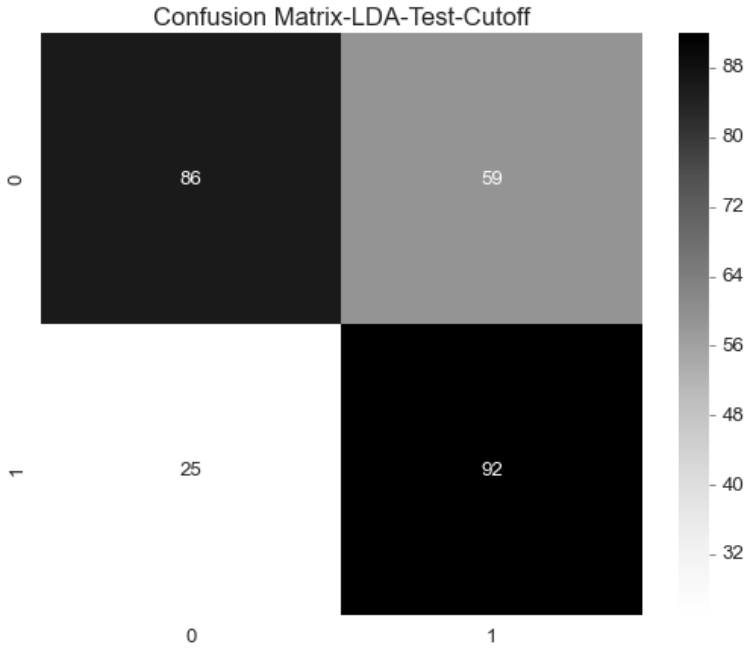

Figure 56:Confusion Matrix-Model4-Test Data

Accuracy for Train Data – 0.67

Accuracy for Test Data – 0.68

The accuracies are nearly the same. Hence there is no underfitting and Overfitting. It has a good score hence a good model.

ROC Curve and AUC Score – Train Data

AUC Score for LDA train data: 0.672

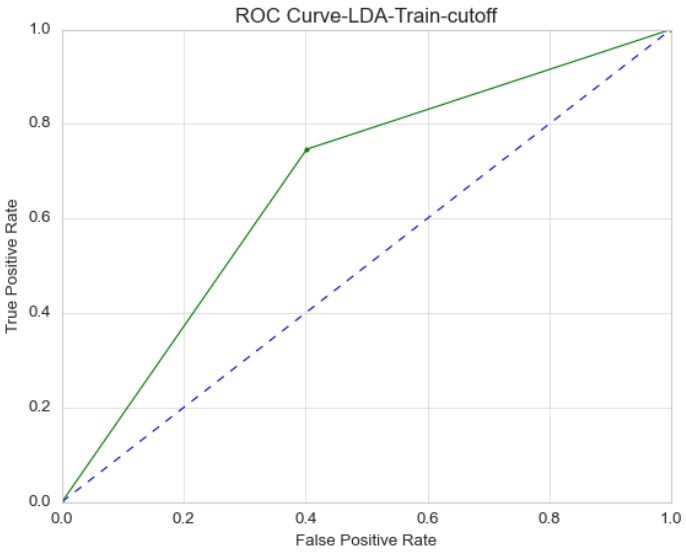

Figure 57:ROC Curve-Model4-Train Data

ROC Curve and AUC Score – Test Data

AUC Score for LDA test data: 0.690

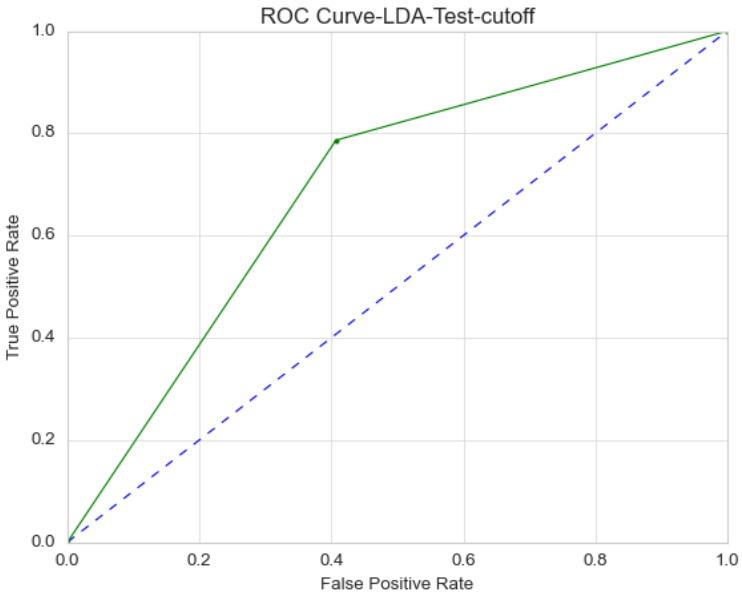

Figure 58:ROC Curve-Model4-Test Data

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

	Model 1 Train	Model 1 Test	Model 2 Train	Model 2 Test	Model 3 Train	Model 3 Test	Model 4 Train	Model 4 Test
Accuracy	0.52	0.53	0.67	0.64	0.68	0.65	0.67	0.68
AUC	0.57	0.63	0.74	0.70	0.74	0.70	0.67	0.69
Recall	0.08	0.09	0.56	0.56	0.57	0.57	0.75	0.79
Precision	0.42	0.38	0.68	0.61	0.69	0.61	0.62	0.61
F1 Score	0.14	0.14	0.61	0.58	0.63	0.59	0.68	0.69

Table 29: Performance Metrics – LDA and LogReg

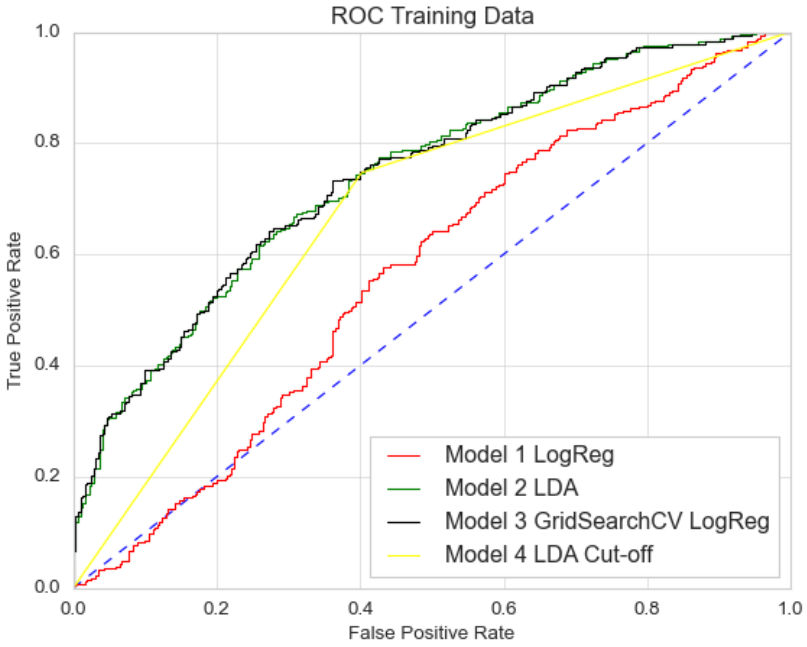

Figure 59:ROC Curve-All 3 Models-Train Data

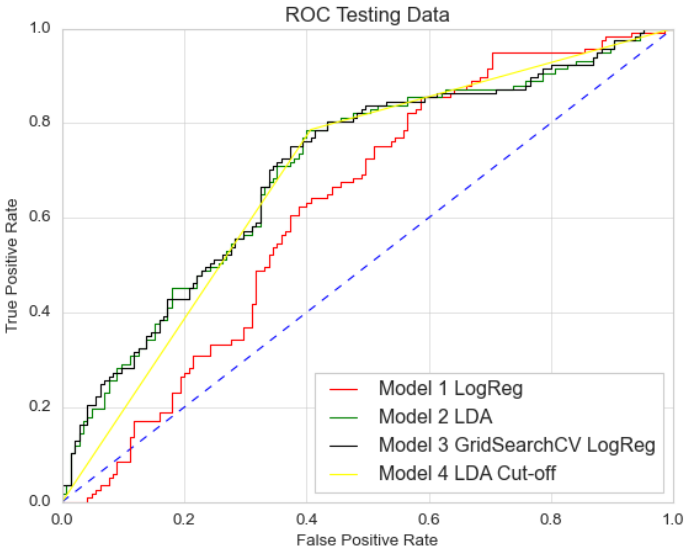

Figure 60:ROC Curve-All 3 Models-Test Data

Out of all the 4 models, Model 4 has slightly better performance than the others which can be inferred from the above performance metrics. It also has a better recall value than others. So model 4 is the best/optimized model.

Recommendations

• Features such as gender, place of interest such as hill station, beaches, dessert and forest safari, etc could be considered as additional features .

• The travel company can provide customised packages based on age groups, salary to attract employees to opt for respective packages

• Foreign nationals can be given offers and specific packages based on their interest which will vary from that of non foreigners.

• The travel company can identify new features which can positively correlate the opting for a holiday package or not so that a better model can be built.

• Offers can be given based on number of children the employee has .

----------X-------X-------X----------