PREDICTIVE MODELING

Table of content

Problem 1

Problem 2

List of tables

List of figures

BUSINESS REPORT

Problem 1: Linear regression

1.1 Read 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
...	...	...	...	...	...	...	...	...	...	...	...
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

26967 rows × 11 columns

Shape of the data

(26967, 11)

Number of rows are 26967
Number of columns are 11

Checking the null values and data types

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

Out of 11 columns
Unnamed: 0, Price are integer data types
Carat, depth, table, x, y, z are float data types
Color and clarity are object data types
There are 697 missing values in the depth column

Description of the data

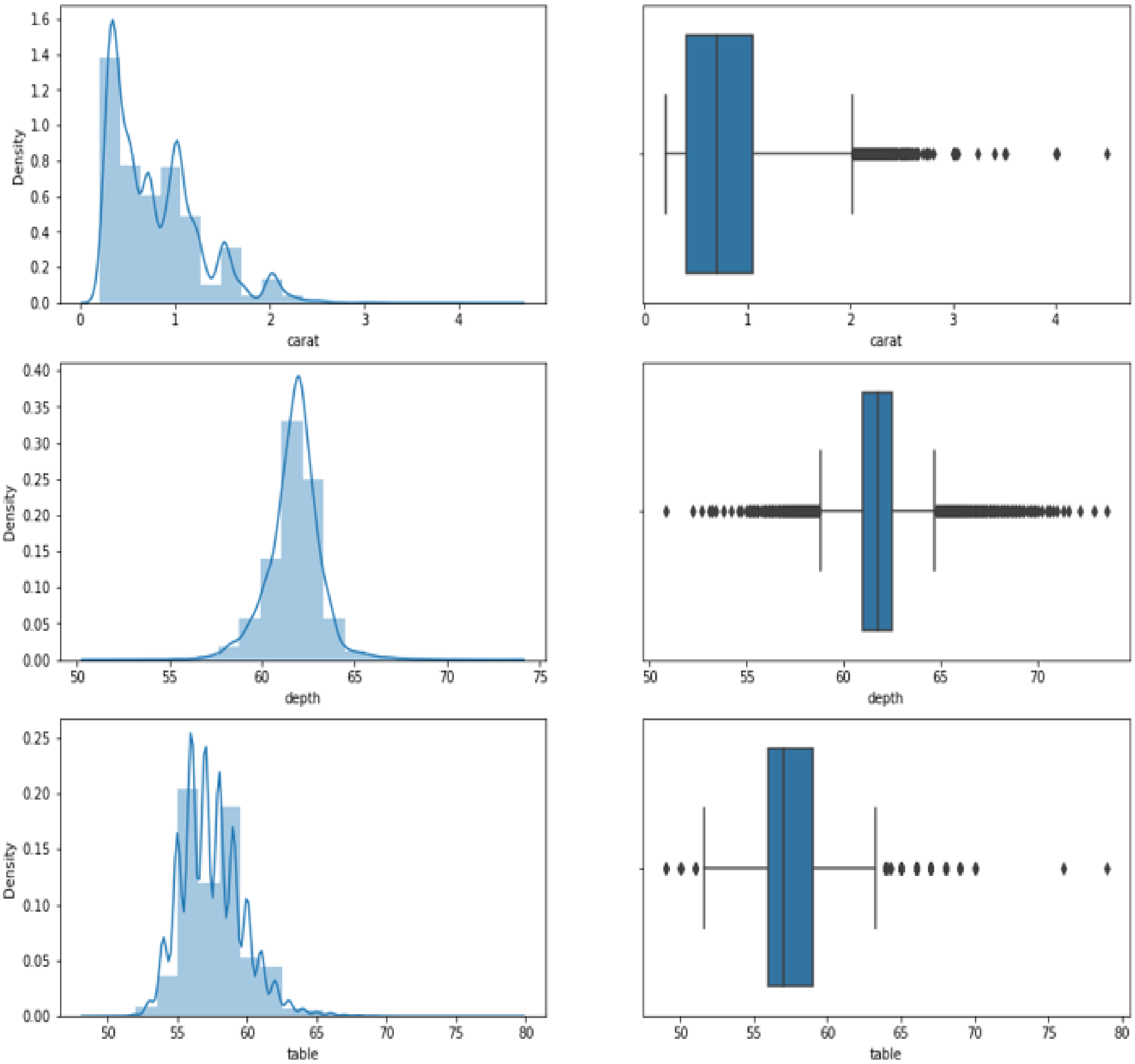
	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
count	26967.000000	26967.000000	26967	26967	26967	26270.000000	26967.000000	26967.000000	26967.000000	26967.000000	26967.000000
unique	NaN	NaN	5	7	8	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	NaN	Ideal	G	SI1	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	NaN	10816	5661	6571	NaN	NaN	NaN	NaN	NaN	NaN
mean	13484.000000	0.798375	NaN	NaN	NaN	61.745147	57.456080	5.729854	5.733569	3.538057	3939.518115
std	7784.846691	0.477745	NaN	NaN	NaN	1.412860	2.232068	1.128516	1.166058	0.720624	4024.864666
min	1.000000	0.200000	NaN	NaN	NaN	50.800000	49.000000	0.000000	0.000000	0.000000	326.000000
25%	6742.500000	0.400000	NaN	NaN	NaN	61.000000	56.000000	4.710000	4.710000	2.900000	945.000000
50%	13484.000000	0.700000	NaN	NaN	NaN	61.800000	57.000000	5.690000	5.710000	3.520000	2375.000000
75%	20225.500000	1.050000	NaN	NaN	NaN	62.500000	59.000000	6.550000	6.540000	4.040000	5360.000000
max	26967.000000	4.500000	NaN	NaN	NaN	73.600000	79.000000	10.230000	58.900000	31.800000	18818.000000

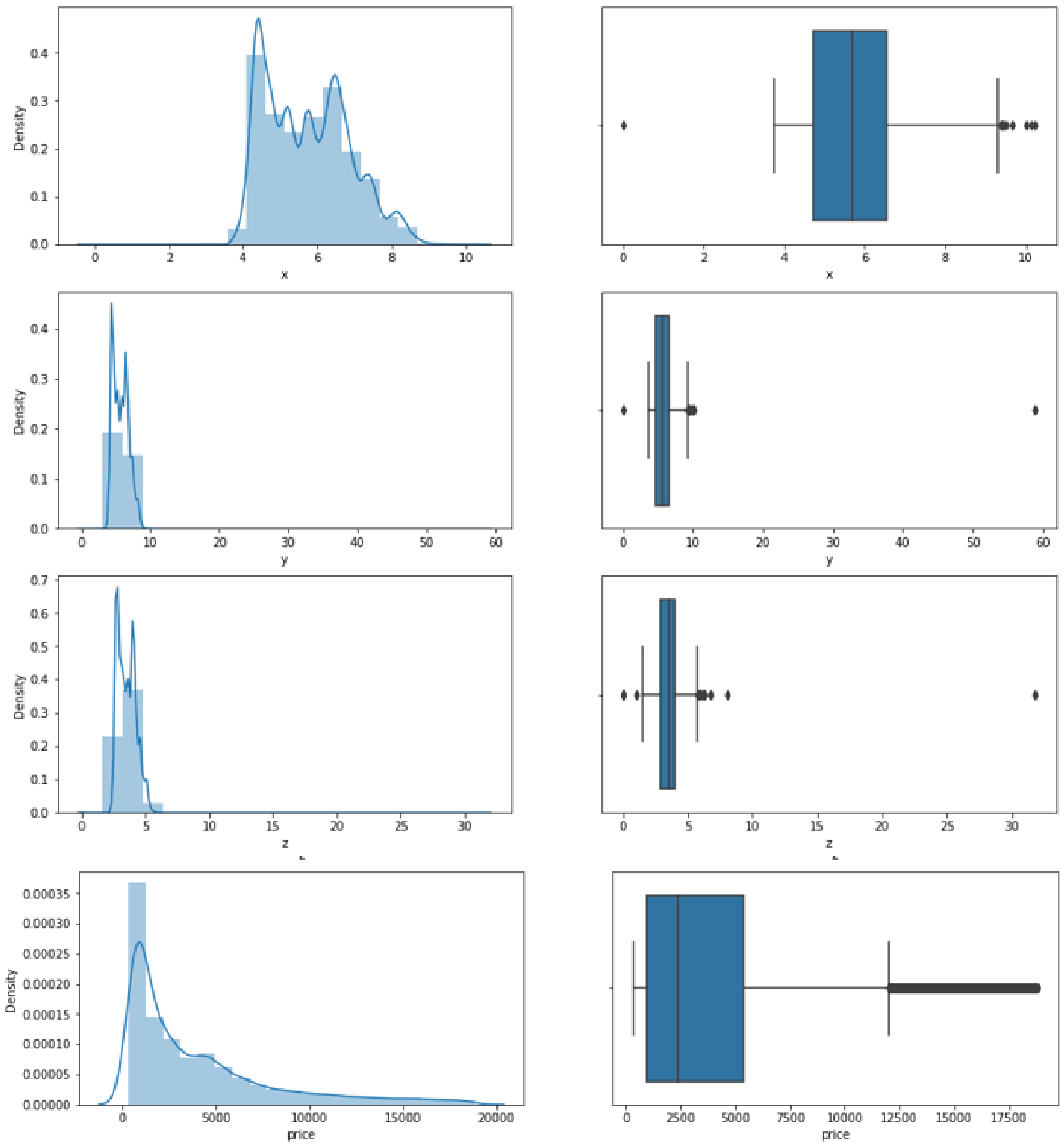
Checking the duplicate

There are no duplicated rows in the data set

UNIVARIATE ANALYSIS

FIGURE 1: UNIVARIATE ANALYSIS





Above figure shows univariate analysis of the variables
Carat variable: It is slightly right skewed as the outliers are present
Depth variable: It is close to normal as the outliers are present but does not impact the mean
Table variable: It is close to normal as the outliers are present but does not impact the mean
X variable: It follows a normal distribution
Y variable: It follows a normal distribution
Z variable: It follows a normal distribution
Price variable: It is right skewed as the outliers are present impact the mean

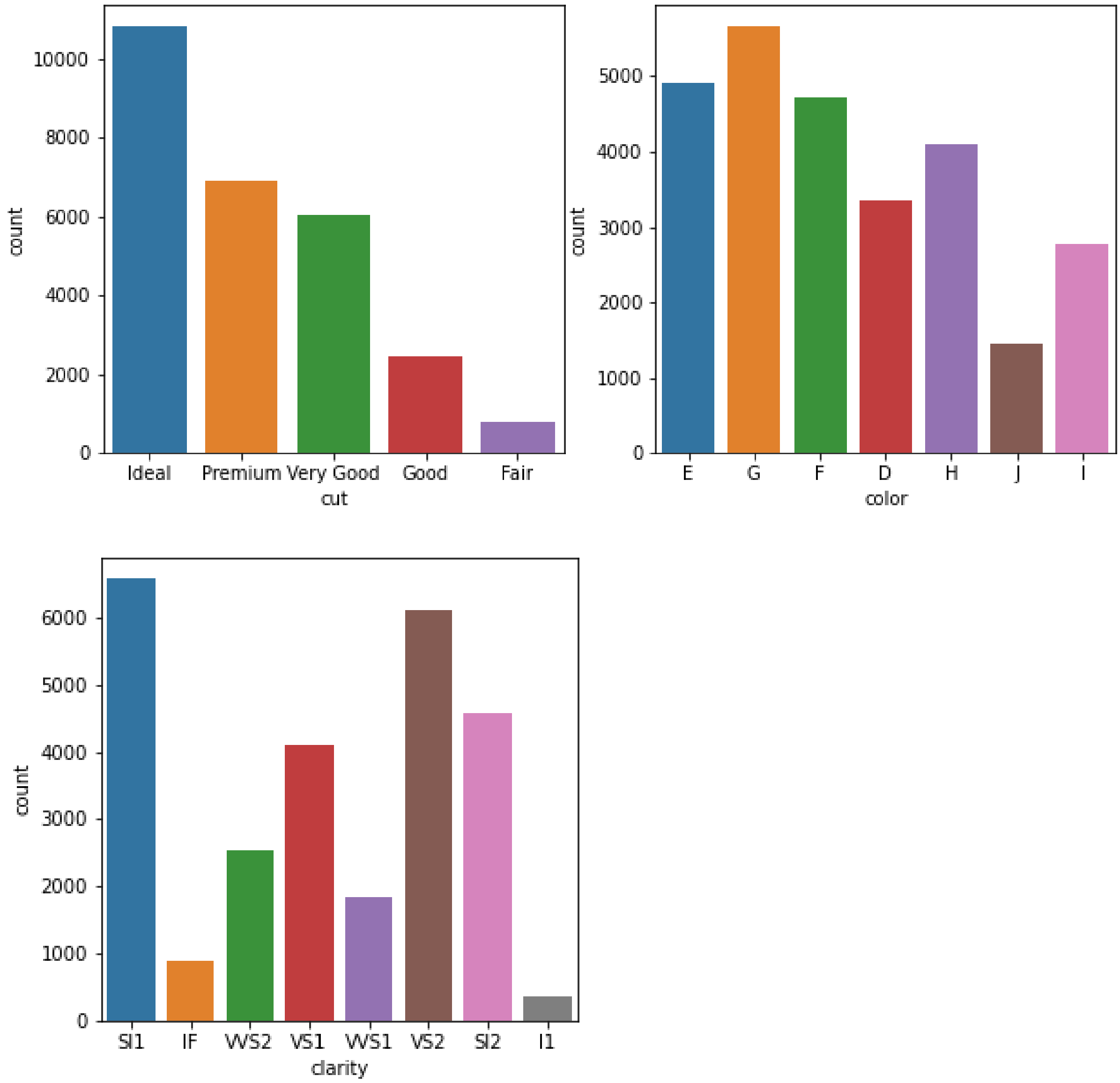
Outliers proportion
Carat variable: 2.45%
Depth variable: 5.26%
Table variable: 1.17%
X variable: 0.05%

Y variable: 0.05%
Z variable: 0.08%
Price variable: 6.59%

UNIVARIATE ANALYSIS FOR CATEGORICAL VARIABLE

COUNT PLOT

figure 2: count plot

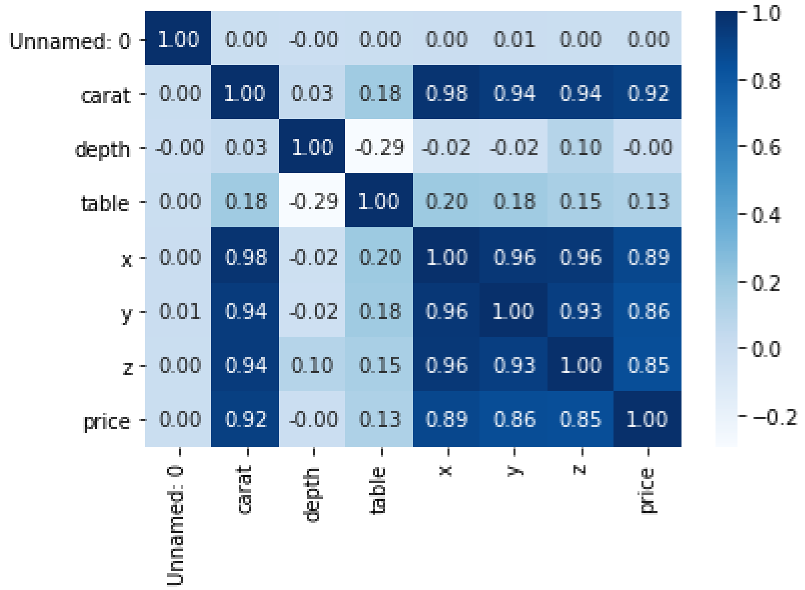


Cut: Fair < good < very good < Premium < Ideal
Color: J > I > D > H > F > E > G
Clarity: I1 < IF < WS1 < WS2 < VS1 < SI2 < VS2 < SI1

BIVARIATE ANALYSIS

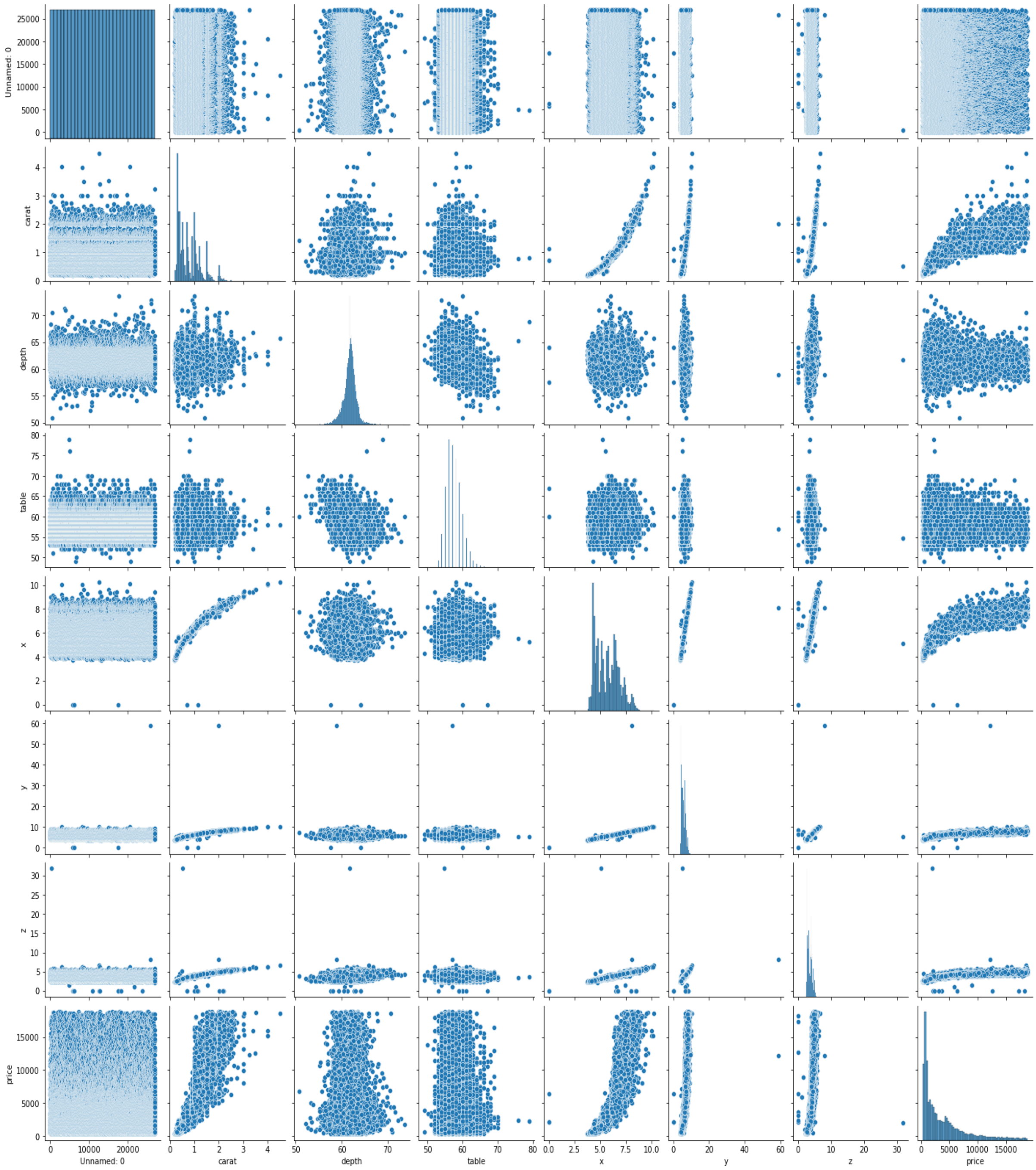
HEAT MAP

FIGURE 3: HEAT MAP



From the above figure Variable Carat is high correlated to variables x, y, z and price with correlation of 0.98, 0.94, 0.94, 0.92.
Variable x is highly correlated to variables y, z and price with a correlation of 0.96, 0.96 and 0.89.
Variable y is highly correlated to variables z and price with a correlation of 0.93 and 0.86
Variable z is correlated to variable price with a correlation of 0.85.

FIGURE 4: PAIRPLOT



From the above pair plot as the variable carat increases variables x, y and z also increases As the variable price increases variables x, y and z increases and reaches to its maximum peak value. There are outliers present in variables x, y and z

Variable carat and price is also related as carat increases the price also increases

1.2 imputing the null values

	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
26	27	0.34	Ideal	D	SI1	NaN	57.0	4.50	4.44	2.74	803
86	87	0.74	Ideal	E	SI2	NaN	59.0	5.92	5.97	3.52	2501
117	118	1.00	Premium	F	SI1	NaN	59.0	6.40	6.36	4.00	5292
148	149	1.11	Premium	E	SI2	NaN	61.0	6.66	6.61	4.09	4177
163	164	1.00	Very Good	F	VS2	NaN	55.0	6.39	6.44	3.99	6340
...	...	...	...	...	...	...	...	...	...	...	...
26848	26849	1.22	Very Good	H	VS1	NaN	59.0	6.91	6.85	4.29	7673
26854	26855	1.29	Premium	I	VS2	NaN	58.0	7.12	7.03	4.27	6321
26879	26880	0.51	Very Good	E	SI1	NaN	58.0	5.10	5.13	3.12	1343
26923	26924	0.51	Ideal	D	VS2	NaN	57.0	5.12	5.09	3.18	1882
26960	26961	1.10	Very Good	D	SI2	NaN	63.0	6.76	6.69	3.94	4361

697 rows × 11 columns

There are 697 missing values in the depth variable. As the variable depth is close to normal distribution we replace all the missing value by the median value which is 61.7. after imputing the null values data are as follows

	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
26	27	0.34	Ideal	D	SI1	61.7	57.0	4.50	4.44	2.74	803
86	87	0.74	Ideal	E	SI2	61.7	59.0	5.92	5.97	3.52	2501
117	118	1.00	Premium	F	SI1	61.7	59.0	6.40	6.36	4.00	5292
148	149	1.11	Premium	E	SI2	61.7	61.0	6.66	6.61	4.09	4177
163	164	1.00	Very Good	F	VS2	61.7	55.0	6.39	6.44	3.99	6340

Here are the first five rows of the missing values of the data sets. Therefore, we replace all the 697 missing values by the median values.

	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
5821	5822	0.71	Good	F	SI2	64.1	60.0	0.0	0.0	0.0	2130
6215	6216	0.71	Good	F	SI2	64.1	60.0	0.0	0.0	0.0	2130
17506	17507	1.14	Fair	G	VS1	57.5	67.0	0.0	0.0	0.0	6381

There are 3 rows in the data set which has variable x, y and z as zero value. As the dependent variable price has value which is to be predicted by the model. And all other variables including carat, cut, color, clarity, depth and table has values so the variable x, y and z cannot be 0.
We replace all the zero values by the lower limit from the box plot

After imputing,

	carat	cut	color	clarity	depth	table	x	y	z	price
5821	0.71	Good	F	SI2	64.1	60.0	1.95	1.965	1.19	2130.0
6215	0.71	Good	F	SI2	64.1	60.0	1.95	1.965	1.19	2130.0
17506	1.14	Fair	G	VS1	59.0	63.5	1.95	1.965	1.19	6381.0

1.3 Encoding the data

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

Scaling the data

	carat	cut	color	clarity	depth	table	x	y	z	price
0	-1.067407	4	1	5	0.288851	0.261603	-1.295386	-1.288528	-1.258146	-0.933183
1	-1.002532	3	3	0	-0.777680	0.261603	-1.162290	-1.136600	-1.200779	-0.793447
2	0.230108	2	1	2	0.370892	1.188780	0.275152	0.346935	0.348130	0.735009
3	-0.807904	4	2	3	-0.121353	-0.665574	-0.807366	-0.832743	-0.827893	-0.765211
4	-1.045782	4	2	1	-1.105843	0.725192	-1.224402	-1.163411	-1.272487	-0.852511

Creating dummy variables for model building

	carat	depth	table	x	y	z	price	cut_1	cut_2	cut_3	...	color_4	color_5	color_6	clarity_1	clarity_2	clarity_3	clarity_4	clarity_5	clarity_6	clarity_7
0	-1.067407	0.288851	0.261603	-1.295386	-1.288528	-1.258146	-0.933183	0	0	0	...	0	0	0	0	0	0	0	1	0	0
1	-1.002532	-0.777680	0.261603	-1.162290	-1.136600	-1.200779	-0.793447	0	0	1	...	0	0	0	0	0	0	0	0	0	0
2	0.230108	0.370892	1.188780	0.275152	0.346935	0.348130	0.735009	0	1	0	...	0	0	0	0	1	0	0	0	0	0
3	-0.807904	-0.121353	-0.665574	-0.807366	-0.832743	-0.827893	-0.765211	0	0	0	...	0	0	0	0	0	1	0	0	0	0
4	-1.045782	-1.105843	0.725192	-1.224402	-1.163411	-1.272487	-0.852511	0	0	0	...	0	0	0	1	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
26962	0.684239	0.452933	0.261603	0.780919	0.704413	0.792724	0.481179	0	0	1	...	0	0	0	0	0	0	0	1	0	0
26963	-1.002532	0.124770	-1.129162	-1.144544	-1.172348	-1.143412	-0.755992	0	0	0	...	1	0	0	0	0	0	0	0	0	0
26964	-0.613277	-0.039312	0.261603	-0.541173	-0.519950	-0.526717	-0.599833	0	0	1	...	0	0	0	0	0	0	1	0	0	0
26965	-1.132283	0.042729	-0.665574	-1.366371	-1.368961	-1.344196	-0.880458	0	1	0	...	0	0	0	0	1	0	0	0	0	0
26966	0.986992	0.206810	0.261603	1.038239	1.026143	1.050876	0.411454	0	0	1	...	0	0	1	0	0	0	0	1	0	0

26967 rows × 24 columns

Splitting the data into train set and testing set at 70:30
Training data

Xtrain1: (18876, 23)
Train_labels1: (18876, 1)
Testing data
Xtest1: (8091, 23)
Test_labels1: (8091,1)

Applying linear regression model
Linear regression model using sklearn
Linear regression model is used to predict the continuous form of the dependent variable. In this model we need the predict the price of cubic zirconia.
First we train the model using training data and then use for testing data for predictions. The output of the model is in the form y = mx + c
Where x is the variable
M is the coefficient of the variable
C is the intercept
Linear model uses gradient descent approach to find the best fit line which gives the minimum error with coefficients

We get the following coefficient of the variables

```
the coefficient of carat is 1.219840104421633
the coefficient of depth is -0.005959591827335875
the coefficient of table is -0.014071598528856052
the coefficient of x is -0.4270213702679114
the coefficient of y is 0.2935180900578898
the coefficient of z is -0.02179800525689941
the coefficient of cut_1 is 0.10903833456096067
the coefficient of cut_2 is 0.14532724674818837
the coefficient of cut_3 is 0.17462008019535863
the coefficient of cut_4 is 0.18114077380663807
the coefficient of color_1 is -0.05615037825674192
the coefficient of color_2 is -0.0779449796025534
the coefficient of color_3 is -0.12156499215469985
the coefficient of color_4 is -0.24319837169212216
the coefficient of color_5 is -0.3816598850345283
the coefficient of color_6 is -0.5519346720844236
the coefficient of clarity_1 is -0.05708402647817809
the coefficient of clarity_2 is -0.06658298151946244
the coefficient of clarity_3 is -0.18273945122610658
the coefficient of clarity_4 is -0.2609799299990168
the coefficient of clarity_5 is -0.4166005965950975
the coefficient of clarity_6 is -0.6513256933409982
the coefficient of clarity_7 is -1.1688597490306822
```

Above are the coefficient of all the variables which is achieved by finding the best fit line with the price variable by gradient descent approach.
Variable carat has the highest positive coefficient 1.21. Every one unit increase in the carat the price goes up by 1.21 unit keeping all the variables constant
Variable clarity has the negative coefficient -1.16. Means every one-unit increase in the clarity price goes down by -1.16 unit keeping all other variables constant

The intercept of our model is 0.32

The determinant of coefficient (R^2) is 0.9408 for the training set
The determinant of coefficient (R^2) is 0.9403 for the testing set
Where R^2 determines how good the model is
For training data,
Root mean square error is 0.2437
For testing data,
Root mean square error is 0.2498

Perform check of the variables using stats model
This model uses ordinary least square method to calculate the minimum error and the best fit line

MODEL 1

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.941
Model:                            OLS   Adj. R-squared:                  0.941
Method:                 Least Squares   F-statistic:                 1.304e+04
Date:                Sat, 30 Oct 2021   Prob (F-statistic):               0.00
Time:                        19:49:36   Log-Likelihood:                -136.98
No. Observations:               18876   AIC:                             322.0
Df Residuals:                   18852   BIC:                             510.3
Df Model:                          23
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.3217      0.016     20.293      0.000       0.291       0.353
carat          1.2198      0.010    122.036      0.000       1.200       1.239
depth         -0.0060      0.003     -2.084      0.037      -0.012      -0.000
table         -0.0141      0.002     -5.872      0.000      -0.019      -0.009
x             -0.4270      0.043     -9.927      0.000      -0.511      -0.343
y              0.2935      0.043      6.780      0.000       0.209       0.378
z             -0.0218      0.016     -1.325      0.185      -0.054       0.010
cut_1          0.1090      0.013      8.722      0.000       0.085       0.134
cut_2          0.1453      0.012     12.142      0.000       0.122       0.169
cut_3          0.1746      0.012     14.962      0.000       0.152       0.197
cut_4          0.1811      0.012     14.907      0.000       0.157       0.205
color_1       -0.0562      0.007     -8.572      0.000      -0.069      -0.043
color_2       -0.0779      0.007    -11.803      0.000      -0.091      -0.065
color_3       -0.1216      0.006    -18.839      0.000      -0.134      -0.109
color_4       -0.2432      0.007    -35.303      0.000      -0.257      -0.230
color_5       -0.3817      0.008    -49.848      0.000      -0.397      -0.367
color_6       -0.5519      0.010    -58.068      0.000      -0.571      -0.533
clarity_1     -0.0571      0.012     -4.806      0.000      -0.080      -0.034
clarity_2     -0.0666      0.011     -5.856      0.000      -0.089      -0.044
clarity_3     -0.1827      0.011    -16.941      0.000      -0.204      -0.162
clarity_4     -0.2610      0.011    -24.710      0.000      -0.282      -0.240
clarity_5     -0.4166      0.011    -39.169      0.000      -0.437      -0.396
clarity_6     -0.6513      0.011    -58.746      0.000      -0.673      -0.630
clarity_7     -1.1689      0.019    -61.701      0.000      -1.206      -1.132
==============================================================================
Omnibus:                     4697.007   Durbin-Watson:                   1.982
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            17423.603
Skew:                           1.212   Prob(JB):                         0.00
Kurtosis:                       7.034   Cond. No.                         67.7
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Below is the linear equation of the stats model 1

(0.32)*Intercept + (1.22)*carat + (-0.01)*depth + (-0.01)*table + (-0.43)*x + (0.29)*y + (-0.02)*z + (0.11)*cut_1 + (0.15)*cut_2 + (0.17)*cut_3 + (0.18)*cut_4 + (-0.06)*color_1 + (-0.08)*color_2 + (-0.12)*color_3 + (-0.24)*color_4 + (-0.38)*color_5 + (-0.55)*color_6 + (-0.06)*clarity_1 + (-0.07)*clarity_2 + (-0.18)*clarity_3 + (-0.26)*clarity_4 + (-0.42)*clarity_5 + (-0.65)*clarity_6 + (-1.17)*clarity_7 +

MODEL 2

This model 2 contains all the variables except z variable. From the model 1, the pvalue of z is 0.185 which is greater than value of alpha which 0.05. therefore, we fail to reject the null hypothesis.
Null hypothesis – there is no correlation between the independent and the dependent variable
Alternative hypothesis – there is correlation between the independent and the dependent variable.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.941
Model:                            OLS   Adj. R-squared:                  0.941
Method:                 Least Squares   F-statistic:                 1.363e+04
Date:                Sat, 30 Oct 2021   Prob (F-statistic):               0.00
Time:                        20:12:33   Log-Likelihood:                -137.86
No. Observations:               18876   AIC:                             321.7
Df Residuals:                   18853   BIC:                             502.2
Df Model:                          22
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.3206      0.016     20.251      0.000       0.290       0.352
carat          1.2194      0.010    122.048      0.000       1.200       1.239
depth         -0.0085      0.002     -3.937      0.000      -0.013      -0.004
table         -0.0140      0.002     -5.858      0.000      -0.019      -0.009
x             -0.4350      0.043    -10.212      0.000      -0.518      -0.351
y              0.2803      0.042      6.653      0.000       0.198       0.363
cut_1          0.1106      0.012      8.886      0.000       0.086       0.135
cut_2          0.1465      0.012     12.273      0.000       0.123       0.170
cut_3          0.1757      0.012     15.085      0.000       0.153       0.198
cut_4          0.1823      0.012     15.039      0.000       0.159       0.206
color_1       -0.0562      0.007     -8.581      0.000      -0.069      -0.043
color_2       -0.0779      0.007    -11.802      0.000      -0.091      -0.065
color_3       -0.1216      0.006    -18.841      0.000      -0.134      -0.109
color_4       -0.2431      0.007    -35.290      0.000      -0.257      -0.230
color_5       -0.3816      0.008    -49.842      0.000      -0.397      -0.367
color_6       -0.5520      0.010    -58.077      0.000      -0.571      -0.533
clarity_1     -0.0571      0.012     -4.810      0.000      -0.080      -0.034
clarity_2     -0.0666      0.011     -5.854      0.000      -0.089      -0.044
clarity_3     -0.1828      0.011    -16.946      0.000      -0.204      -0.162
clarity_4     -0.2610      0.011    -24.709      0.000      -0.282      -0.240
clarity_5     -0.4166      0.011    -39.167      0.000      -0.437      -0.396
clarity_6     -0.6514      0.011    -58.750      0.000      -0.673      -0.630
clarity_7     -1.1692      0.019    -61.727      0.000      -1.206      -1.132
==============================================================================
Omnibus:                     4695.499   Durbin-Watson:                   1.982
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            17411.712
Skew:                           1.212   Prob(JB):                         0.00
Kurtosis:                       7.033   Cond. No.                         59.0
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Below is the linear equation for model 2

```
(0.32)*Intercept + (1.22)*carat + (-0.01)*depth + (-0.01)*table + (-0.43)*x + (0.28)*y + (0.11)*cut_1 + (0.15)*cut_2 + (0.18)*c
ut_3 + (0.18)*cut_4 + (-0.06)*color_1 + (-0.08)*color_2 + (-0.12)*color_3 + (-0.24)*color_4 + (-0.38)*color_5 + (-0.55)*color_6
+ (-0.06)*clarity_1 + (-0.07)*clarity_2 + (-0.18)*clarity_3 + (-0.26)*clarity_4 + (-0.42)*clarity_5 + (-0.65)*clarity_6 + (-1.1
7)*clarity_7 +
```

Comparison of both the models

Table 1 : comparison of models

	Model 1		Model 2	
	Training set	**Testing set**	**Training set**	**Testing set**
R Square	0.941	0.941	0.941	0.941
RMSE	0.2437	0.2498	0.2437	0.2430
Adj Rsquare	0.941	0.941	0.941	0.941

From the above table both model 1 and model 2 have the same determinant of coefficient and adj Rsquare which is 0.941.
Root mean squared error for model 1
Training set: 0.2437
Testing set: 0.2498
Root mean squared error for model 2
Training set: 0.2437
Testing set: 0.2430
As model 1 contains z variable which has no correlation with the output variable price and model 2 contains all the variables which has a correlation with the output variable price. Therefore, we select model 2

1.4 Inferences and recommendations

Variable carat has the highest positive coefficient 1.21. Every one-unit increase in the carat, the price goes up by 1.21 unit keeping all the variables constant
Variable clarity_7 has the negative coefficient -1.16. Means every one-unit increase in the clarity_7 price goes down by -1.16 unit keeping all other variables constant

There are many variables having positive and negative coefficients which increases and decreases the price of the cubic
With the carat variable we can predict the higher price of cubic which can be grouped in higher profitable stones and lower profitable stones
With the clarity_7 variable we group the lower and higher profitable stones.
Five variables which are good predictors of price variable
Carat: 1.21
Y: 0.28
Cut_1: 0.11
Clarity_6: -0.65
Clarity_7: -1.16

- Selling of the stones based on the higher carat value will be more profitable.
- Avoid selling stones based on clarity_7 which result in low profit.
- Give more discounts on more profitable stones to attract more customers and gain more profit.
- Introduce more design as demanded by the customers based on the carat in order to obtain more profits.

2.1 Read 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
...	...	...	...	...	...	...	...	...
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

872 rows × 8 columns

Shape of the data

(872, 7)

Checking the null value

```
<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
```

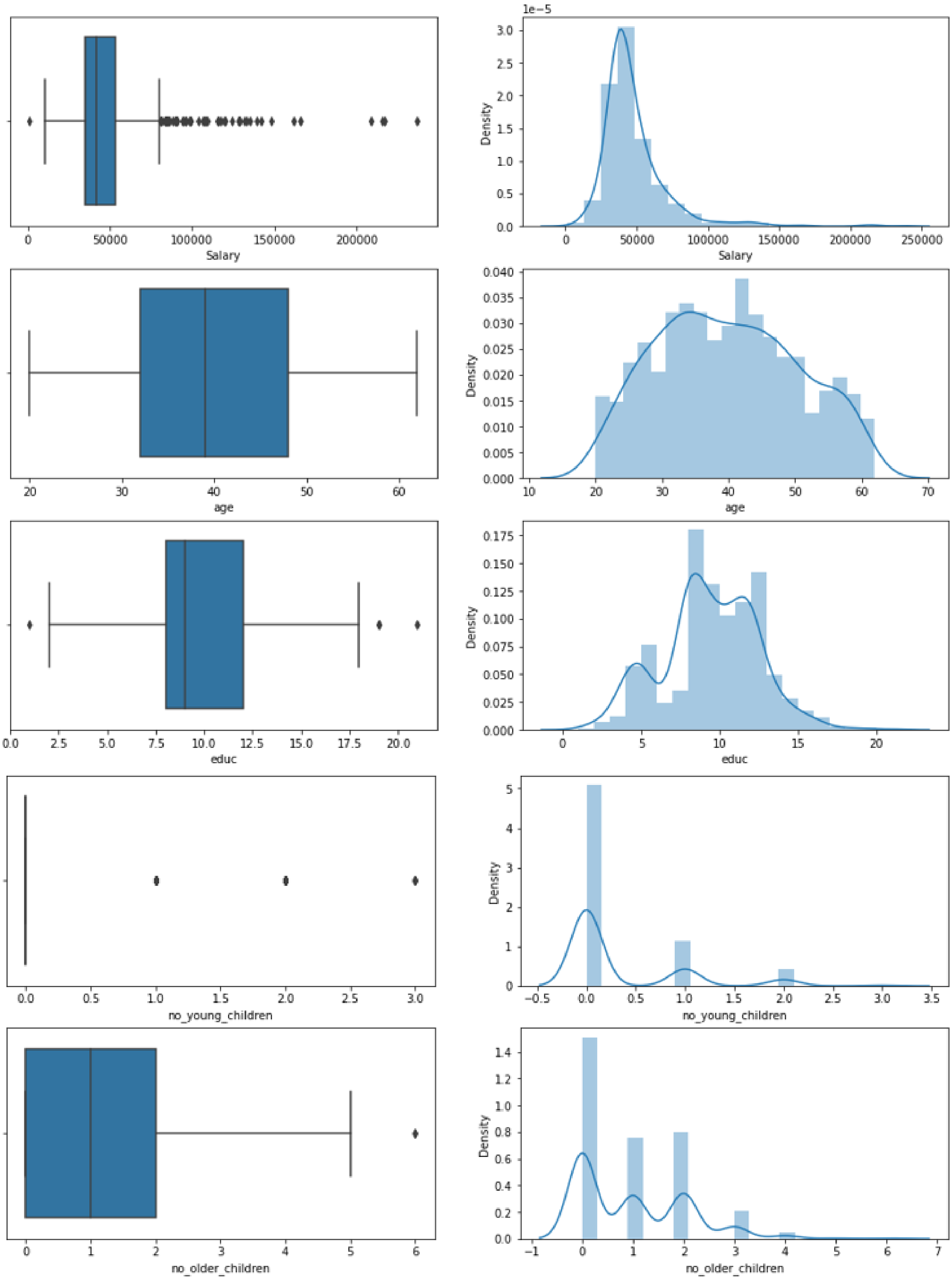
There are 7 columns and 872 entries in each columns. There are no missing values in the columns. There are two columns of object data type and 5 columns of integer data types.

Descriptive statistics of the data

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

Univariate analysis

figure 5: univariate analysis



The above figure shows univariate analysis of the variables using box plot and distribution plot.
Variable salary is right skewed as more number of outliers are present and mean is higher than median

Variable age is close to normal and distribution is normal
Variable edu is close to normal as it does not have extreme outliers
Variable **no**_young_children is right skewed as the outliers are present
Variable no_older_children is a normally distributed

Percentage of Outliers

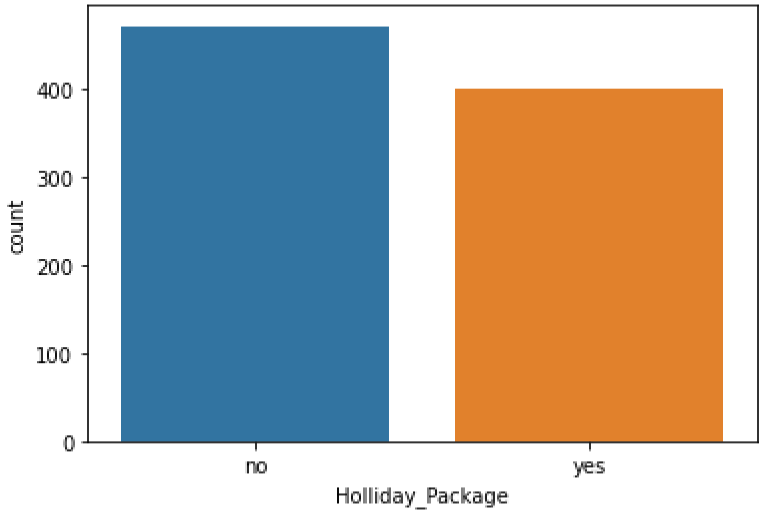
	outliers%
Holliday_Package	0.00
Salary	6.54
age	0.00
educ	0.46
foreign	0.00
no_older_children	0.23
no_young_children	23.74

No_young_children has the highest number of outliers 23.74%
Salary has the 6.54% of outliers
Educ has 0.46% of outliers
No_older_children has 0.23% of outliers which is the lowest of all the variables
Holliday_Package and foreign variable has no outliers

Count plot for categorical variable

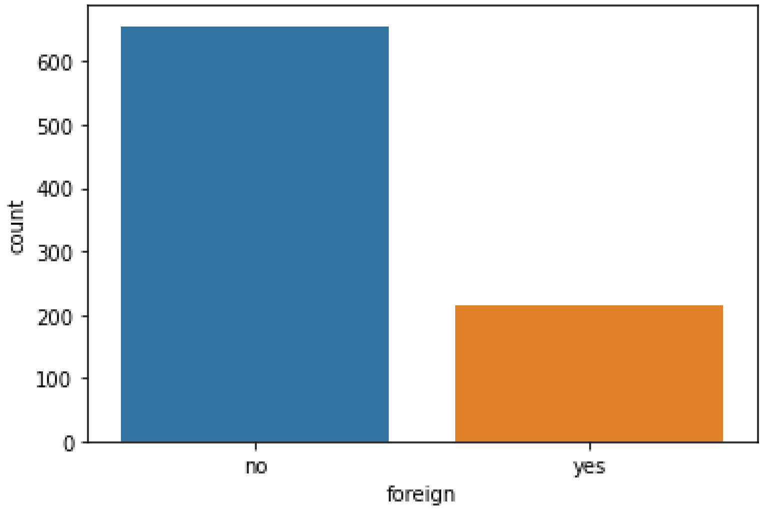
Variable holliday_package

figure 6: count plot



There are 471 employees have not opted for package
There are 401 employees have opted for the package

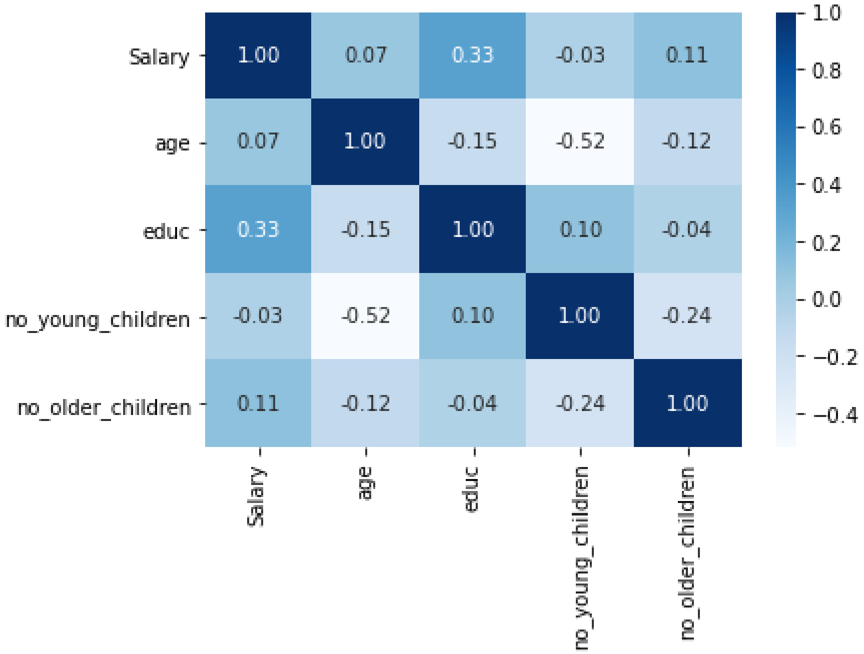
figure 7: count plot



There are 656 not a foreign employee
There are 216 foreign employee

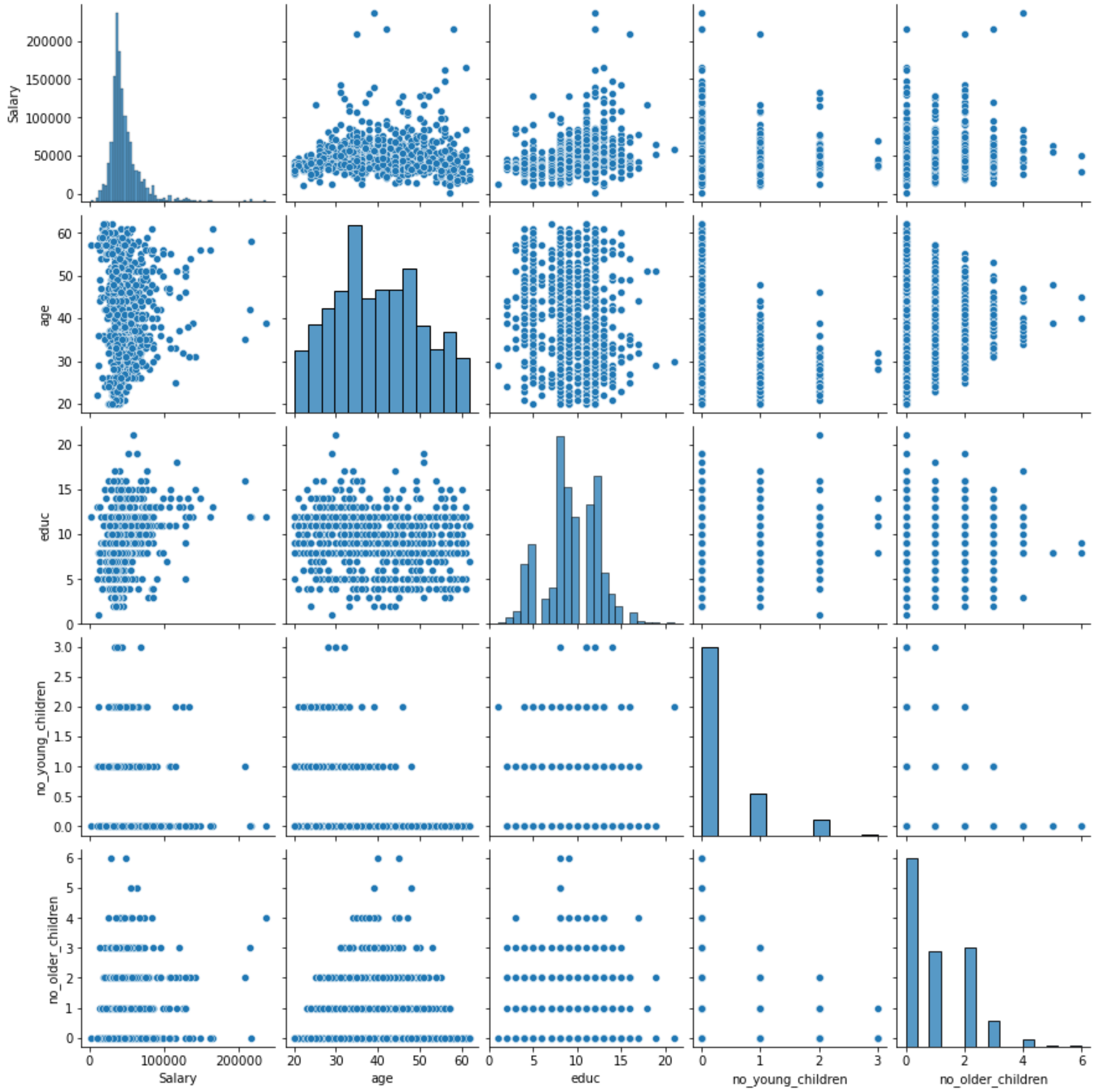
Bivariate analysis

figure 8: heat map



Variable salary and educ has a correlation of 0.33 which is not a strong correlation
And other variables has not a strong correlation

figure 9: pair plot



Variables has no linear relationship

2.2 Encoding the data

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign_yes
0	0	48412.0	30.0	8.0	0.0	1.0	0
1	1	37207.0	45.0	8.0	0.0	1.0	0
2	0	58022.0	46.0	9.0	0.0	0.0	0
3	0	66503.0	31.0	11.0	0.0	0.0	0
4	0	66734.0	44.0	12.0	0.0	2.0	0
...	...	...	...	...	...	...	...
867	0	40030.0	24.0	4.0	0.0	1.0	1
868	1	32137.0	48.0	8.0	0.0	0.0	1
869	0	25178.0	24.0	6.0	0.0	0.0	1
870	1	55958.0	41.0	10.0	0.0	1.0	1
871	0	74659.0	51.0	10.0	0.0	0.0	1

872 rows × 7 columns

Splitting the data into 70: 30

X_train head

	Salary	age	educ	no_young_children	no_older_children	foreign_yes
821	38974.0	47.0	12.0	0.0	2.0	1
805	40270.0	33.0	8.0	0.0	0.0	1
322	32573.0	30.0	11.0	0.0	0.0	0
701	43839.0	43.0	11.0	0.0	1.0	1
773	33060.0	40.0	5.0	0.0	1.0	1

X_train shape (610,6)

Train labels head

	Holliday_Package
821	0
805	0
322	0
701	1
773	1

Train_labels shape (610,1)

X_test head

	Salary	age	educ	no_young_children	no_older_children	foreign_yes
264	25118.0	58.0	8.0	0.0	0.0	0
189	40913.0	20.0	9.0	0.0	0.0	0
643	28446.0	58.0	8.0	0.0	0.0	0
65	36072.0	35.0	4.0	0.0	2.0	0
241	52736.0	40.0	10.0	0.0	3.0	0

X_test shape (262, 6)

Test_labels head

	Holliday_Package
264	1
189	0
643	0
65	1
241	0

Test_labels shape (262,1)

Applying logistic regression model

Logistic regression model internally uses linear equation to find the intercept and coefficient and then it is converted to the classes using activation function. It uses sigmoid curve to calculate the probability depending on the defined threshold. Any value greater than threshold will be considered as 1 and the value less than threshold will be considered as 0. Threshold value is usually 0.5 and it can be adjusted accordingly. It uses log of odds to convert into the probability. Log of odds is the linear equation having intercept and coefficient.
Building logistic regression with the following parameters are as follows

Logistic Regression (max_iter=10000, n_jobs=2, penalty='none', solver='newton-cg',
verbose=True)
max_iter – number of steps taken by the model to minimize error to find the best fit sigmoid curve.
N_jobs – number of processor used to train the model
Solver – optimization technique used to solve

By applying grid search we get following parameters

{'max_iter': 1000, 'n_jobs': 2, 'penalty': 'l2', 'solver': 'newton-cg'}

Applying linear discriminant analysis model

Linear discriminant analysis creates a linear line between the classes to separate 0 and 1. It defines which observation belongs to the class. It also separates multiclass dependent variable. Linear discriminant analysis uses Bayes theorem to calculate the posterior probability from the prior probability. $P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$

P(A|B) = posterior probability
P(B|A) = prior probability
P(B) = condition at any given condition

Comparison of both the model based on performance metrics

- accuracy

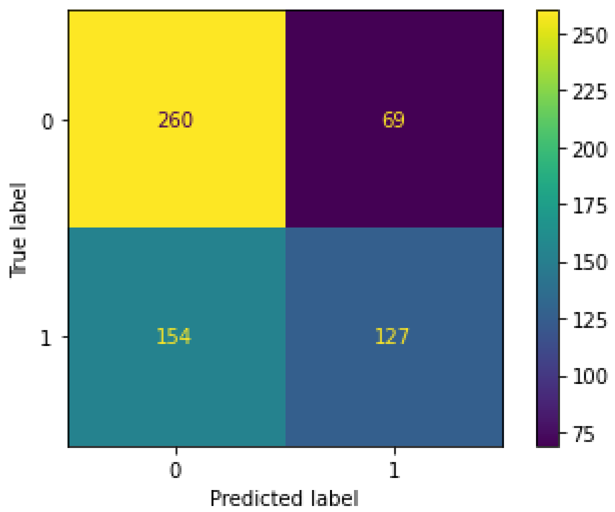
table 2: comparison of accuracy

	Logistic regression		**Linear discriminant analysis**	
	Train set	**Test set**	**Train set**	**Test set**
Accuracy	0.63	0.66	0.63	0.66

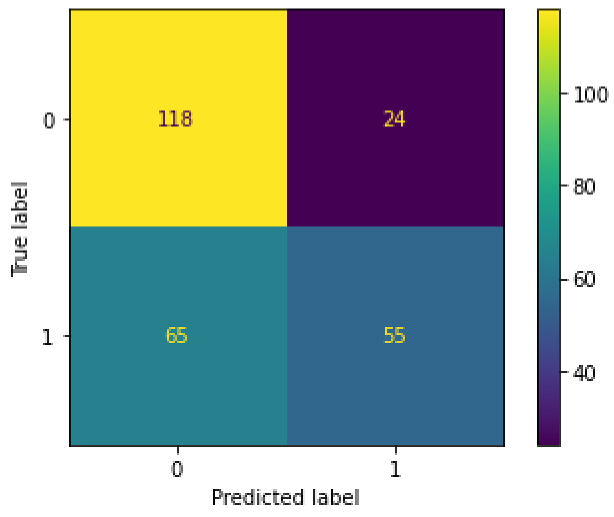
Accuracy for the both models are same for the training and testing data

- Confusion metrics

Logistic regression
Training data

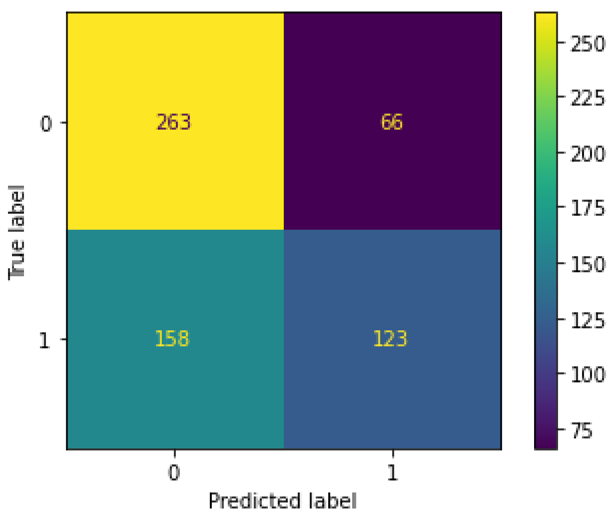


Testing data

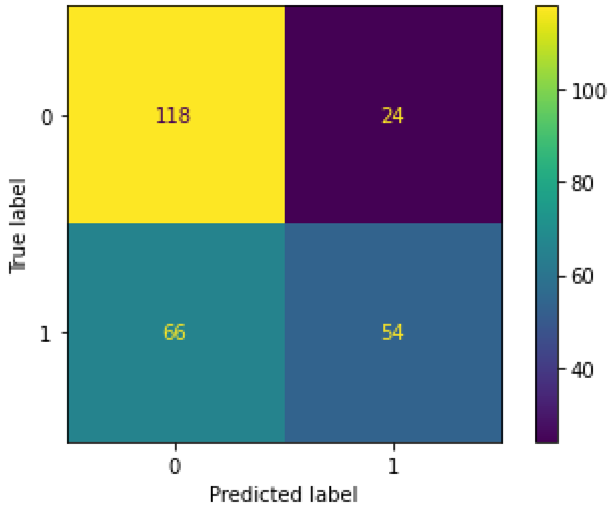


Linear discriminant analysis

- Training data



Testing data



From the confusion metrics true positive cases and false negative cases of both the model are almost same for the training and the testing set
Logistic regression
Train data
TP: 127
FN: 154
Test data
TP: 55
FN: 65
Linear discriminant model
TP: 123
FN: 158
Test data
TP: 54
FN: 66

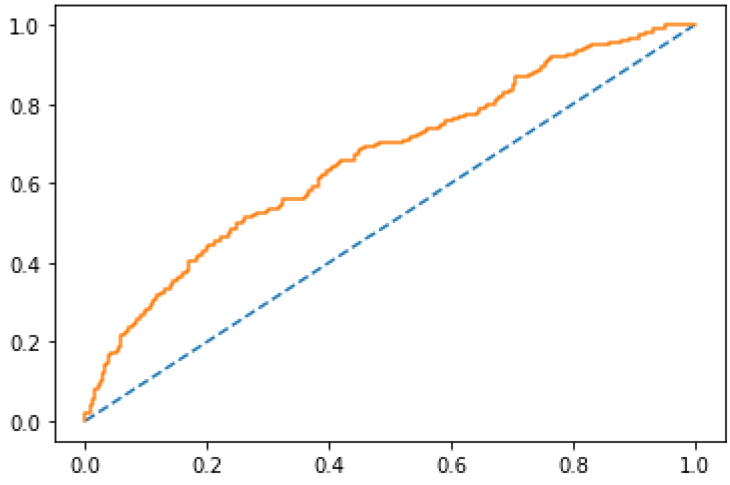
- ROC curve and ROC_AUC score

Logistic regression
Training set

Figure 10: auc score and roc_auc curve

```
auc score is 0.661

[<matplotlib.lines.Line2D at 0x227652b8670>]
```

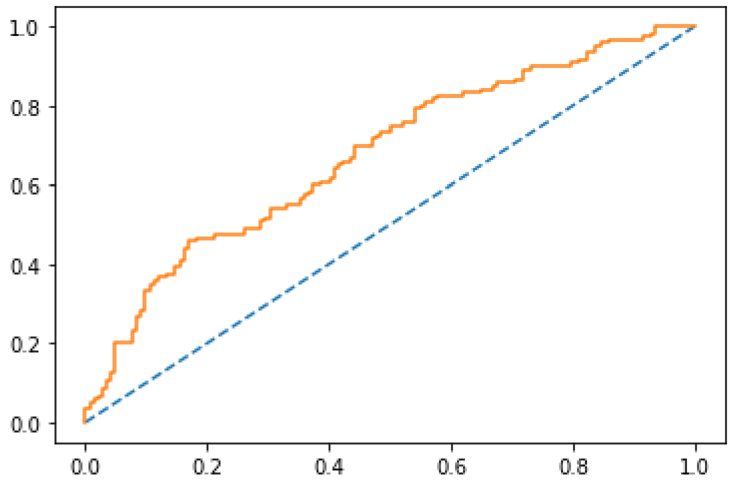


Testing data

figure 11: auc score and roc_auc curve

```
auc score is 0.675

[<matplotlib.lines.Line2D at 0x1de551ac160>]
```

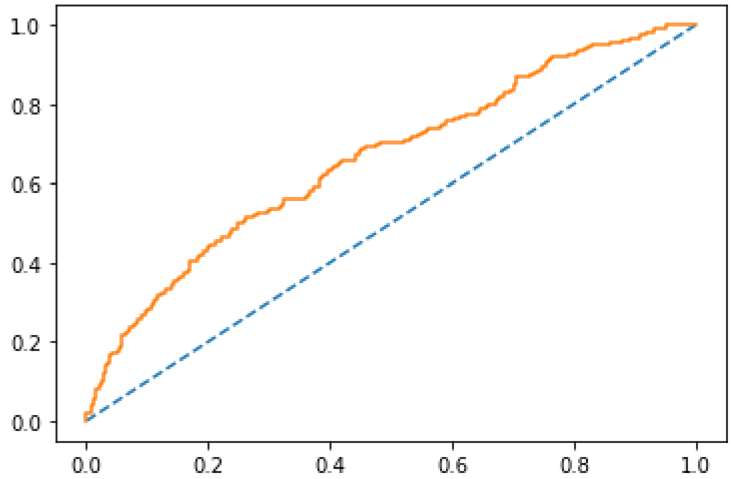


Linear discriminant analysis
Training data

figure 12: auc score and roc_auc curve

```
auc score is 0.661

[<matplotlib.lines.Line2D at 0x227652b8670>]
```

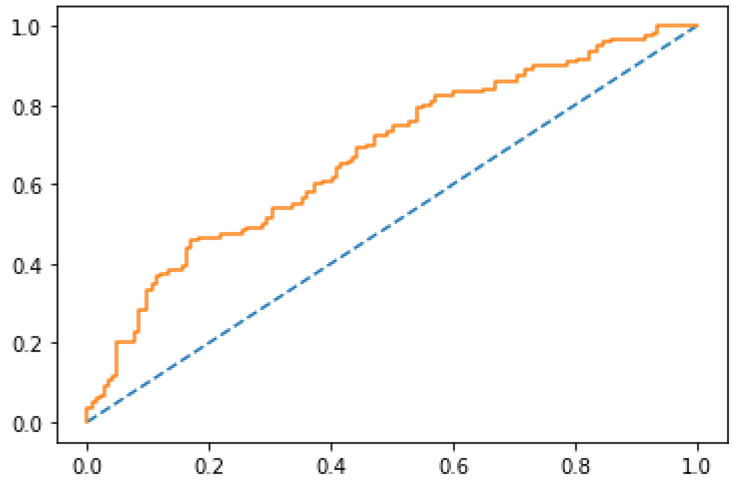


Testing data

figure 13: auc score and roc_auc curve

```
auc score is 0.675

[<matplotlib.lines.Line2D at 0x2276533e0a0>]
```



On comparison of both the models, performance of both the models are almost same.
As the precision of logistics model is 0.46 and for linear discriminant analysis is 0.45.
We can select logistics regression but cannot put into production as the accuracy is not high.
As data is balanced we can select the model based on the accuracy. The accuracy of the model can be increased by tuning the parameters using Grid search.

2.4 Inferences and recommendations

Both the models logistics regression and linear discriminant analysis performance is almost same as the accuracy and precision of the model is low. It cannot put into production, we can tune the parameters of the model and then check the performance.

- Employees having higher salary should be targeted by giving discount on the package.
- Employees having more number of younger children and number of older children should be provided with a good and reasonable family package to increase the sale of the package
- Employees of younger age should be targeted as they usually opt for holiday package with friends.
- Non foreigners should be targeted as numbers are more who have not opted for package. As only 40% of the non-foreigners employees have opted for the package