Problem 1- Linear Regression

Table of Contents

Contents

Table & Figures

Executive Summary:-

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

Introduction:-

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset and analysis of using method Linear Regression and find higher profitable stones and lower profitable stones find other expect rest detail as per data shown in description.

Data Description:-

Variable Name	Description
Carat	Carat weight of the cubic zirconia.
Cut	Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal.
Color	Colour of the cubic zirconia. With D being the best and J the worst.
Clarity	Cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst, FL = flawless, I1= level 1 inclusion) IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1
Depth	The Height of cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter.
Table	The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter.
Price	The Price of the cubic zirconia.
X	Length of the cubic zirconia in mm.
Y	Width of the cubic zirconia in mm.
Z	Height of the cubic zirconia in mm.

Tab-1.a

Sample of the Dataset:-

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

Fig-1.a

Conclusion | insight: This is Sample of given data all information is mention as per provide by company Gem Stones co ltd. shown in this Fig-1.a.

Data Describe:-

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

Fig-1.0

Conclusion | insight: Data min value of data 0.2 and max value is 18818 and other thing zero is available in data. Required scaling due to data range is show higher.

Q 1.1. Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA). Perform Univariate and Bivariate Analysis.

Solution:-

A) Exploratory data analysis:-

```
Data columns (total 11 columns):
 #   Column       Non-Null Count
---  ------       --------------
 0   Unnamed: 0   26967 non-null
 1   carat        26967 non-null
 2   cut          26967 non-null
 3   color        26967 non-null
 4   clarity      26967 non-null
 5   depth        26270 non-null
 6   table        26967 non-null
 7   x            26967 non-null
 8   y            26967 non-null
 9   z            26967 non-null
 10  price        26967 non-null
dtypes: float64(6), int64(2), obj
memory usage: 2.3+ MB
```

Fig-1.b

Conclusion | insight: Ten types of information given in data as a column. sr.no, carat, cut, color, clarity, depth, table, x, y, z. data is 2.3+ MB and 0 to 26967 row and 10 columns.

B) Data-type:-

```
Unnamed: 0      int64
carat         float64
cut            object
color          object
clarity        object
depth         float64
table         float64
x             float64
y             float64
z             float64
price           int64
dtype: object
```

Fig-1.c

Conclusion | insight: Three type of Data integer, Object, and float. Three column are Object and two column are integer and six are float.

C) Data shape & size:-

```
zirco.shape
(26967, 11)
```

```
zirco.size
296637
```

Fig-1.d

Conclusion | insight: Total 26967 ROW and 11 Column.

D) Types of Object data:-

clarity	
SI1	6571
VS2	6099
SI2	4575
VS1	4093
VVS2	2531
VVS1	1839
IF	894
I1	365

color	
G	5661
E	4917
F	4729
H	4102
D	3344
I	2771
J	1443

cut	
Ideal	10816
Premium	6899
Very Good	6030
Good	2441
Fair	781

Fig-1.e

Conclusion | insight: Three columns the have all types of data shown in fig-1.e.

---------- EDA----------

E) Find Duplicated value:-

```
dups=df1.duplicated()
dups.sum()
```

```
34
```

Fig-1.f.a

Sample of duplicate-

	carat	cut	color	clarity	depth	table	x	y	z	price
4756	0.35	Premium	J	VS1	62.4	58.0	5.67	5.64	3.53	949
6215	0.71	Good	F	SI2	64.1	60.0	0.00	0.00	0.00	2130
8144	0.33	Ideal	G	VS1	62.1	55.0	4.46	4.43	2.76	854
8919	1.52	Good	E	I1	57.3	58.0	7.53	7.42	4.28	3105
9818	0.35	Ideal	F	VS2	61.4	54.0	4.58	4.54	2.80	906
10473	0.79	Ideal	G	SI1	62.3	57.0	5.90	5.85	3.66	2898
10500	1.00	Premium	F	VVS2	60.6	54.0	6.56	6.52	3.96	8924
12894	1.21	Premium	D	SI2	62.5	57.0	6.79	6.71	4.22	6505
13547	0.43	Ideal	G	VS1	61.9	55.0	4.84	4.86	3.00	943
13783	0.79	Ideal	G	SI1	62.3	57.0	5.90	5.85	3.66	2898

Fig-1.f.b

Conclusion | insight: Duplicate 34 value is find in this data.

F) Missing value:-

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

Fig-1.g

Conclusion | insight: In this data seat depth column 697 data is missing find.

G) Zero value available:-

Column - X

	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
6215	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

Column - Y

	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
6215	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

Column – Z

	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
6215	0.71	Good	F	SI2	64.1	60.0	0.00	0.00	0.0	2130
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

Fig-1.h

Conclusion | insight: Find three column zero available show in fig-1.h. In z column higher zero value.

H) Univariate Analysis:-

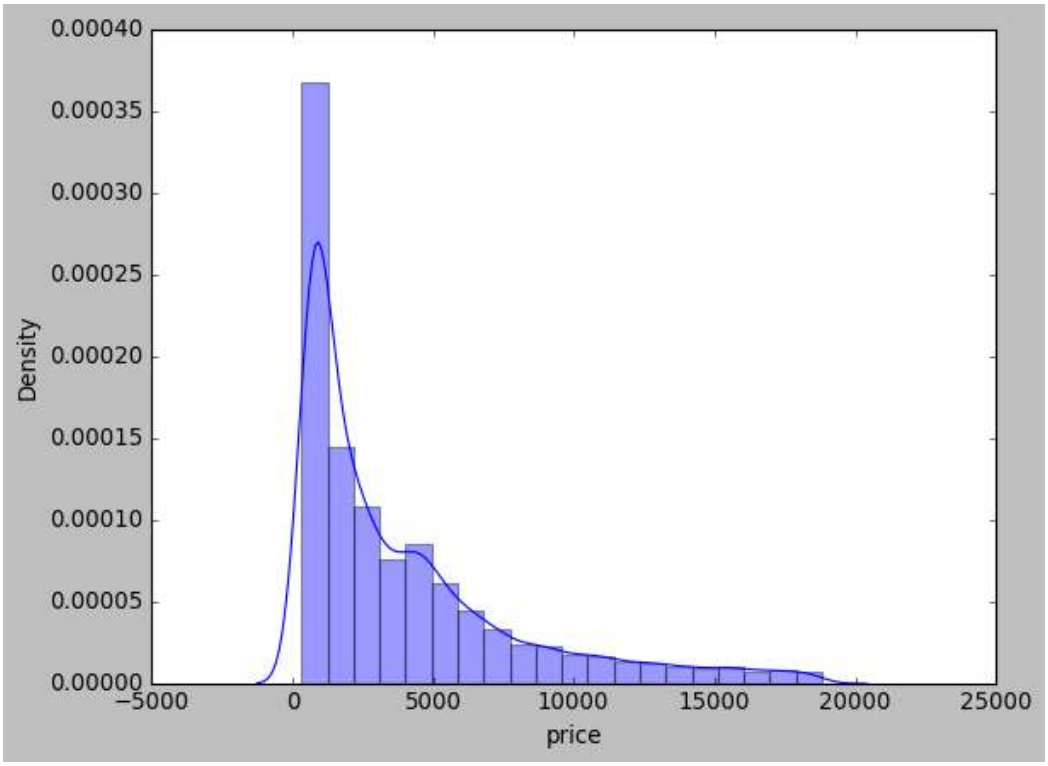


Fig-1.i.1

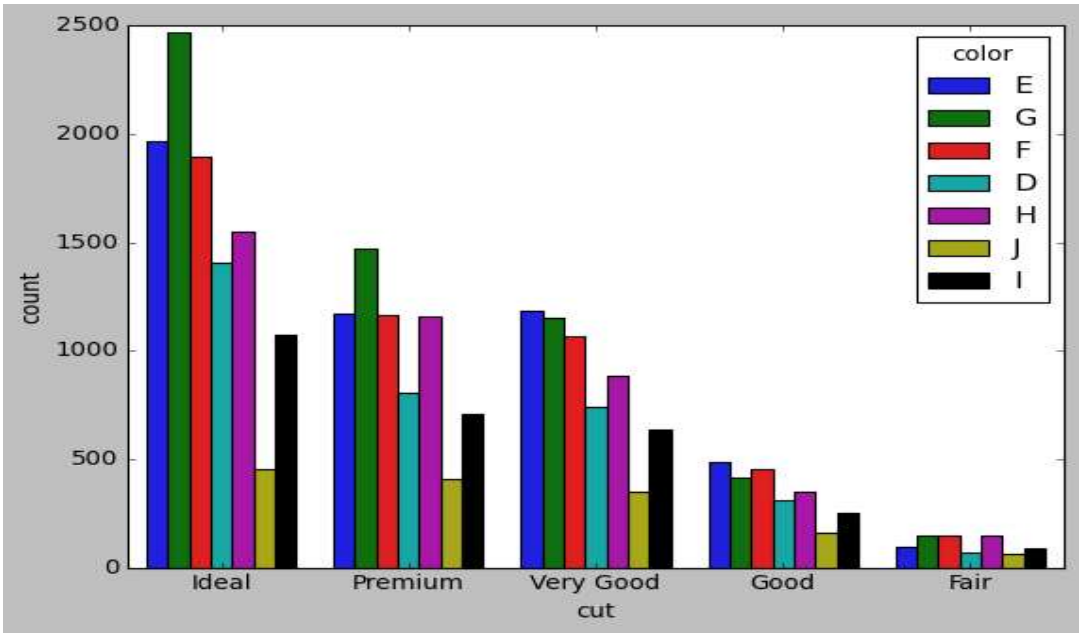


Fig-1.i.2

Conclusion | insight: We are find single| single variable value finding data behaviour show in fig-1.i.1, fig-1.i.2.

I) Bivariate Analysis:-

In single pair of all data

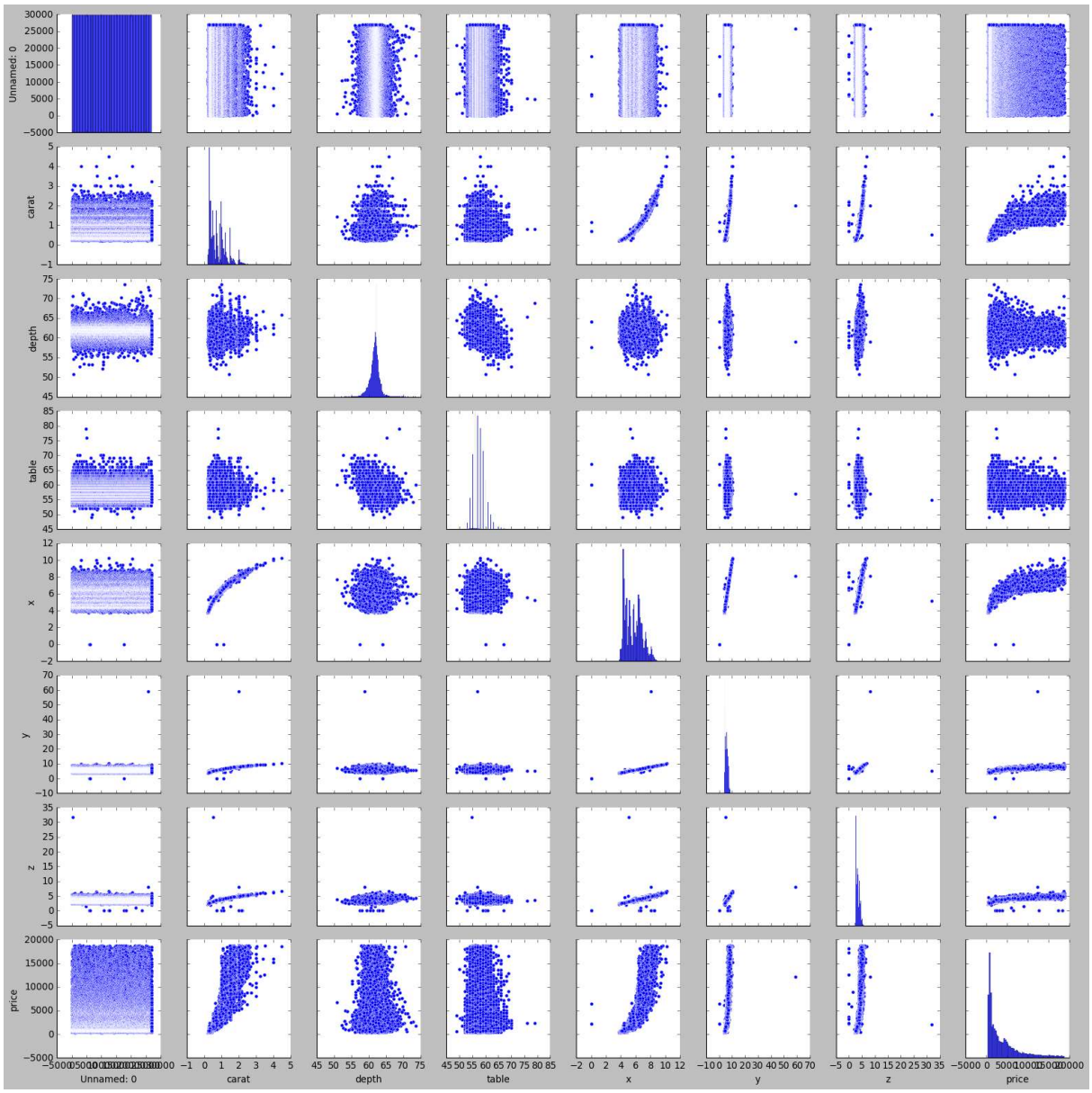


Fig-1.j.1

Outlier check in all data:-

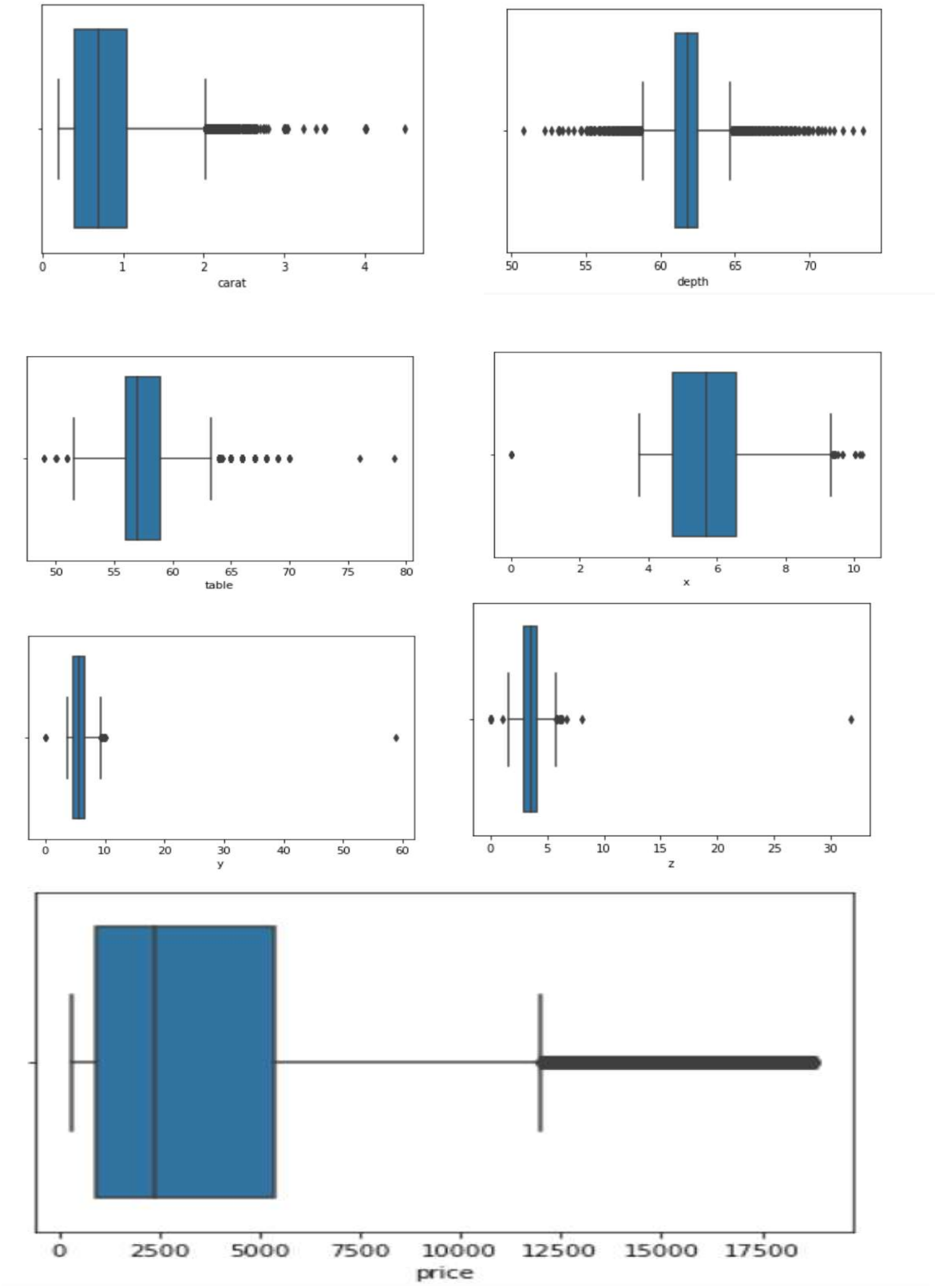


Fig-1.j.2

Correlation of all data:-

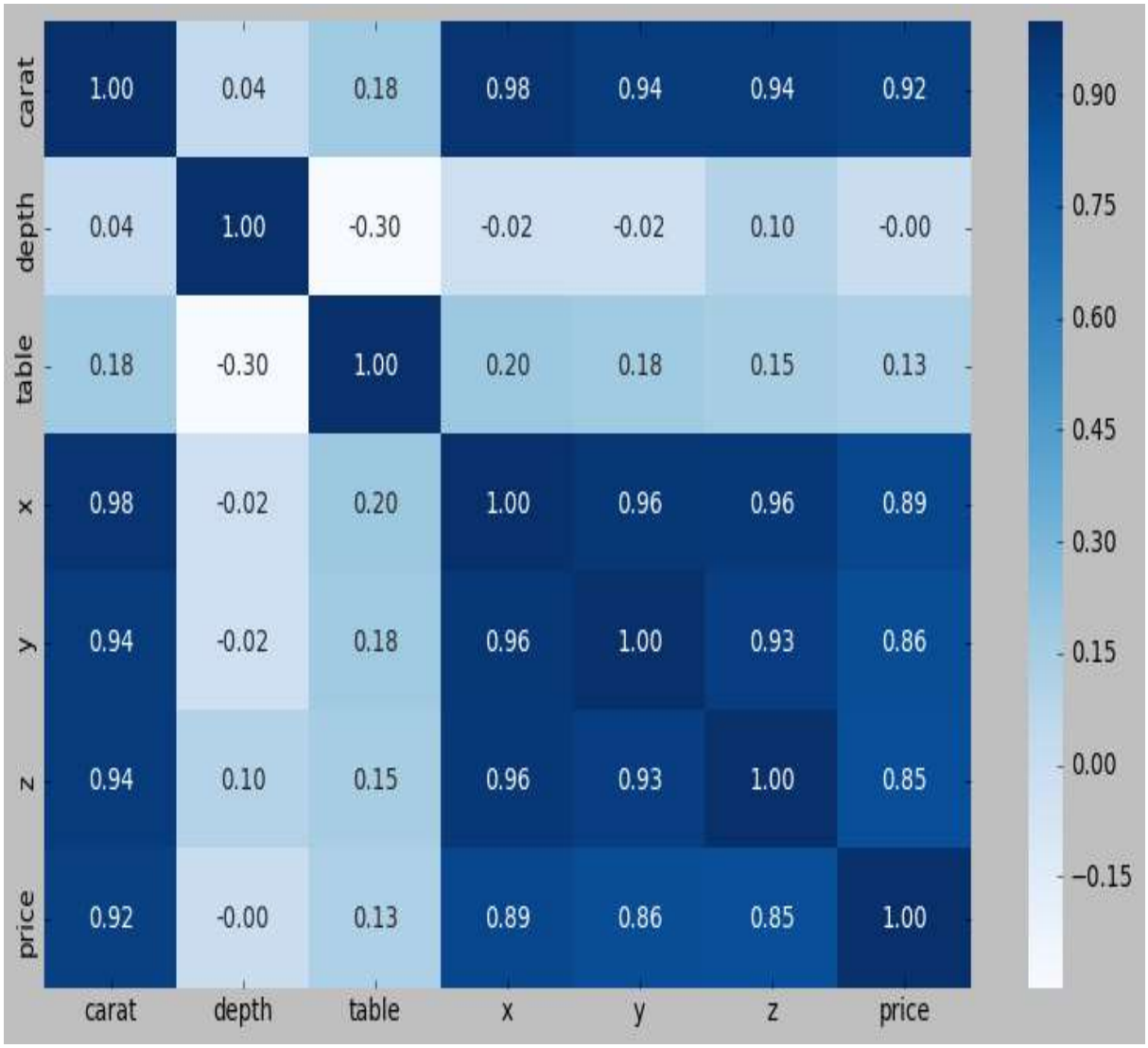


Fig-1.j.2

Conclusion | insight: In all data, we are comparing and showing. What behaving the data .pair plot showing all data and boxplot for outliner check and correlation check each column for every column.

Q1.2 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? Do you think scaling is necessary in this case?

Solution:-

A) Drop column- (Unnamed: 0)

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

Fig-1.2.a

B) Replace missing value:-

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

Fig-1.2.b

C) Treat zero all place:-

carat	cut	color	clarity	depth	table	x	y	z	price

```
df1[df1["x"]==0]
```

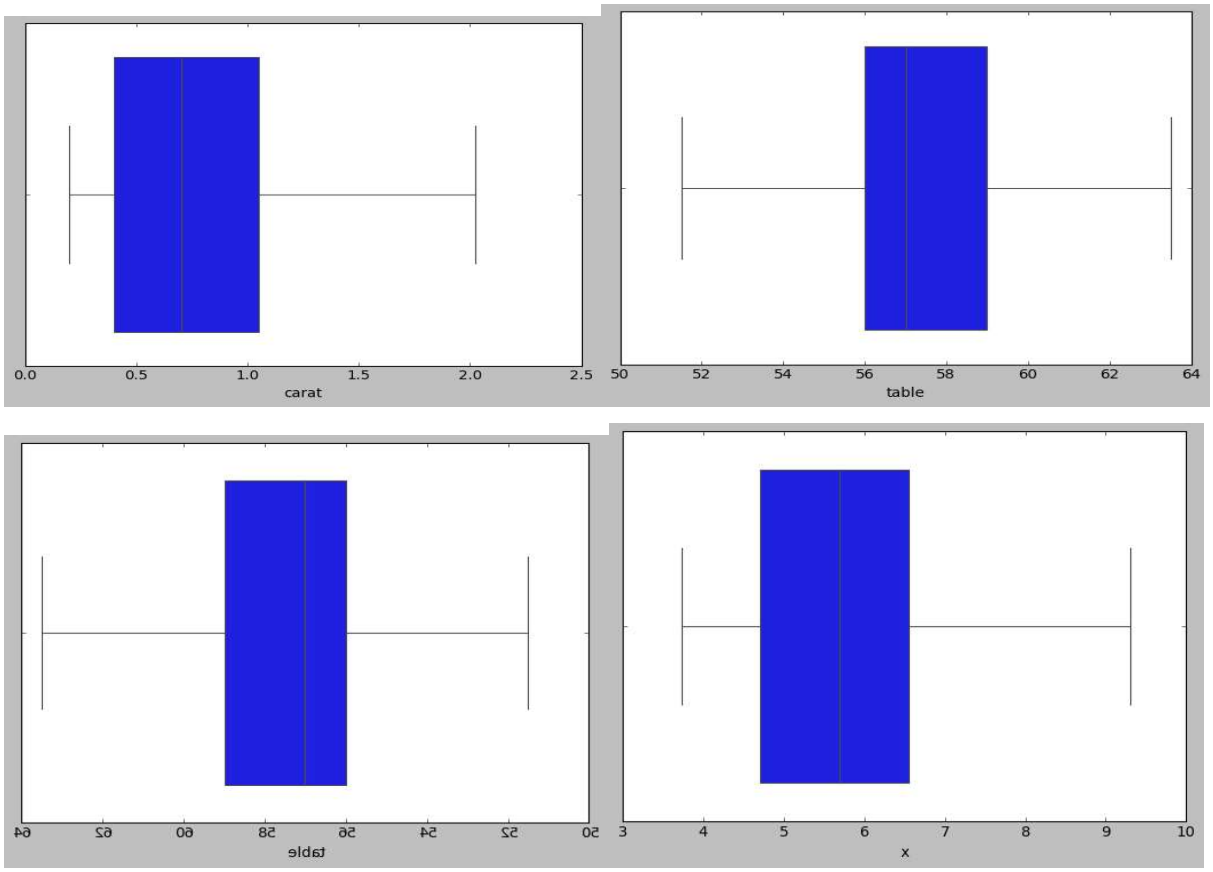
carat	cut	color	clarity	depth	table	x	y	z	price

```
df1[df1["z"]==0]
```

carat	cut	color	clarity	depth	table	x	y	z	price

Fig-1.2.c

D) Treating outliers:-



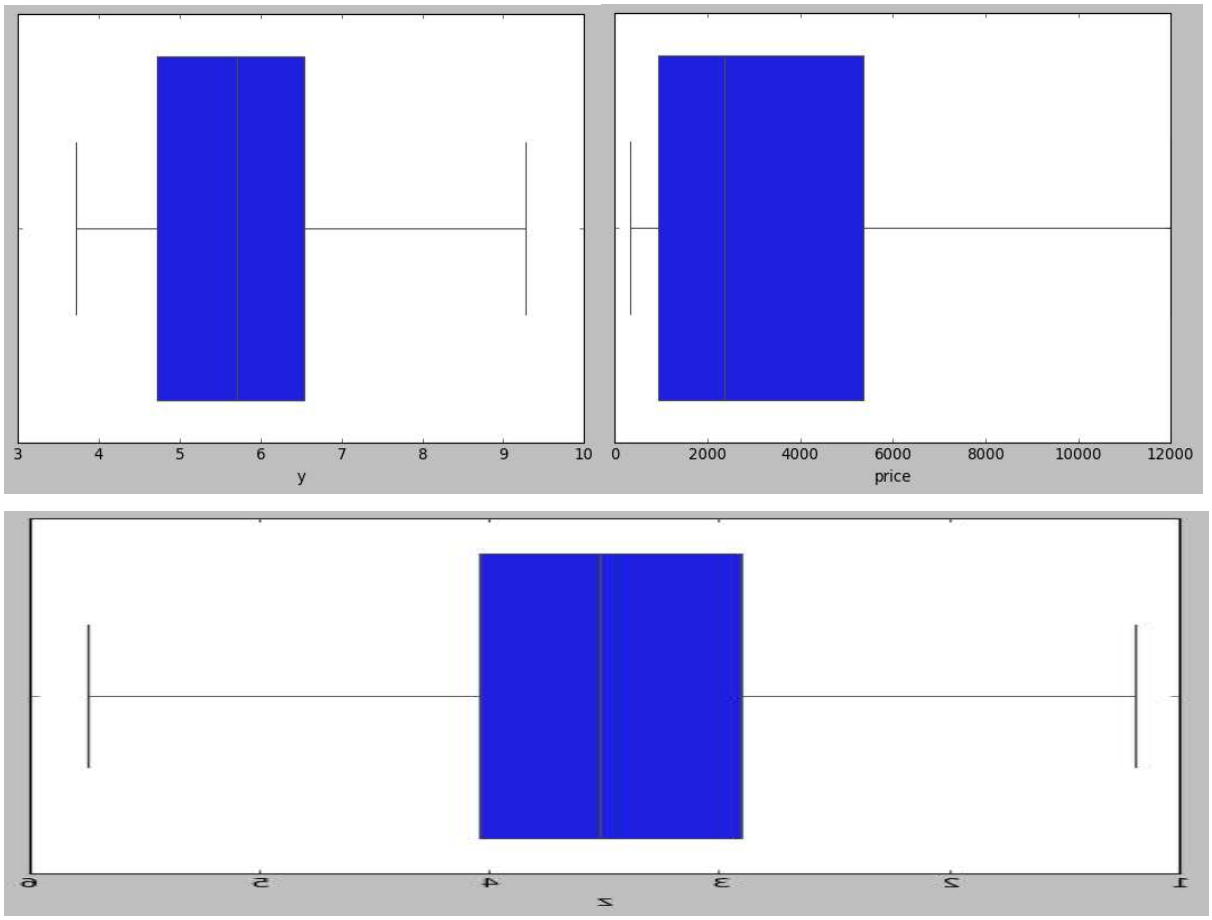


Fig-1.2.d

E) Replace all duplicate value:-

```
df1.duplicated().sum()
```

```
0
```

Fig-1.2.e

F) Convert to Categorical:-

	carat	depth	table	x	y	z	price	cut_Good	cut_Ideal	cut_Premium	...	color_H	color_I	color_J	clarity_IF	clarity_SI1	clarity_SI2	clarity_VS
0	0.30	62.1	58.0	4.27	4.29	2.66	499.0	0	1	0	...	0	0	0	0	1	0	
1	0.33	60.8	58.0	4.42	4.46	2.70	984.0	0	0	1	...	0	0	0	1	0	0	
2	0.90	62.2	60.0	6.04	6.12	3.78	6289.0	0	0	0	...	0	0	0	0	0	0	
3	0.42	61.6	56.0	4.82	4.80	2.96	1082.0	0	1	0	...	0	0	0	0	0	0	
4	0.31	60.4	59.0	4.35	4.43	2.65	779.0	0	1	0	...	0	0	0	0	0	0	

5 rows × 24 columns

Fig-1.2.f

G) Scaling data:-

	carat	depth	table	x	y	z	price	\
0	-1.067534	0.286877	0.261752	-1.296628	-1.289772	-1.261730	-0.	
933027								
1	-1.002580	-0.779916	0.261752	-1.163332	-1.137619	-1.204204	-0.	
793173								
2	0.231554	0.368938	1.188932	0.276264	0.348118	0.349009	0.	
736566								
3	-0.807717	-0.123428	-0.665428	-0.807876	-0.833311	-0.830282	-0.	
764914								
4	-1.045883	-1.108160	0.725342	-1.225536	-1.164469	-1.276112	-0.	
852286								
...	...	...	...	...	...	...	...	
...								
26962	0.686235	0.450999	0.261752	0.782788	0.706127	0.794839	0.	
482523								
26963	-1.002580	0.122755	-1.129018	-1.145559	-1.173420	-1.146677	-0.	
755687								
26964	-0.612854	-0.041367	0.261752	-0.541284	-0.520053	-0.528269	-0.	
599397								
26965	-1.132489	0.040694	-0.665428	-1.367719	-1.370324	-1.348020	-0.	
880257								
26966	0.989355	0.204816	0.261752	1.040493	1.028335	1.053707	0.	
412740								

	cut_Good	cut_Ideal	cut_Premium	...	color_H	color_I	co	
lor_J	\							
0	-0.31531	1.221793	-0.585999	...	-0.423380	-0.338147	-0.2	
37609								
1	-0.31531	-0.818469	1.706488	...	-0.423380	-0.338147	-0.2	
37609								
2	-0.31531	-0.818469	-0.585999	...	-0.423380	-0.338147	-0.2	
37609								
3	-0.31531	1.221793	-0.585999	...	-0.423380	-0.338147	-0.2	
37609								
4	-0.31531	1.221793	-0.585999	...	-0.423380	-0.338147	-0.2	
37609								
...	...	...	...	...	...	...	...	
...								
26962	-0.31531	-0.818469	1.706488	...	-0.423380	-0.338147	-0.2	
37609								
26963	-0.31531	1.221793	-0.585999	...	2.361947	-0.338147	-0.2	
37609								
26964	-0.31531	-0.818469	1.706488	...	-0.423380	-0.338147	-0.2	
37609								
26965	-0.31531	-0.818469	-0.585999	...	-0.423380	-0.338147	-0.2	
37609								
26966	-0.31531	-0.818469	1.706488	...	-0.423380	-0.338147	4.2	
08598								

	clarity_IF	clarity_SI1	clarity_SI2	clarity_VS1	clarity_VS2	\
0	-0.184991	1.761313	-0.451641	-0.422953	-0.540733	
1	5.405654	-0.567758	-0.451641	-0.422953	-0.540733	
2	-0.184991	-0.567758	-0.451641	-0.422953	-0.540733	

3	-0.184991	-0.567758	-0.451641	2.364332	-0.540733
4	-0.184991	-0.567758	-0.451641	-0.422953	-0.540733
...	...	...	...	...	...
26962	-0.184991	1.761313	-0.451641	-0.422953	-0.540733
26963	5.405654	-0.567758	-0.451641	-0.422953	-0.540733
26964	-0.184991	-0.567758	-0.451641	-0.422953	1.849340
26965	-0.184991	-0.567758	-0.451641	-0.422953	-0.540733
26966	-0.184991	1.761313	-0.451641	-0.422953	-0.540733

	clarity_VVS1	clarity_VVS2
0	-0.270743	-0.321957
1	-0.270743	-0.321957
2	-0.270743	3.106009
3	-0.270743	-0.321957
4	3.693534	-0.321957
...	...	...
26962	-0.270743	-0.321957
26963	-0.270743	-0.321957
26964	-0.270743	-0.321957
26965	-0.270743	3.106009
26966	-0.270743	-0.321957

[26927 rows x 24 columns]

Table-1.2.a

Conclusion | insight: all data modifications using zero value treat , outlier treat , missing value treat , duplicate value treat ,drop some value after that our data is ready for perform and finding LR . After looking data scaling is necessary for clustering for k-Mean and Hierarchical clustering also, as per describe data max/Min value Spending 1/18818, and are magnitude all data are not similar, so scaling is required. So we can go for scaling to avoid any performance issue due to the unscaled dataset.

Q1.3 Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply linear regression. Performance Metrics: Check the performance of Predictions on Train and Test sets using Rsquare, RMSE.

Solution:-

Linear regression: - Split data in two part Train & Test

A) Coefficient_train & test:-

```
The coefficient for carat is 1.2305882966567543
The coefficient for depth is 0.0014459929690719868
The coefficient for table is -0.013639708231939687
The coefficient for x is -0.3214591521321091
The coefficient for y is 0.25973210078172104
The coefficient for z is -0.10533359381691018
The coefficient for cut_Good is 0.035198594220606574
The coefficient for cut_Ideal is 0.09291368906273828
The coefficient for cut_Premium is 0.0793527046587005
The coefficient for cut_Very Good is 0.06686385068878037
The coefficient for color_E is -0.02149958578438857
The coefficient for color_F is -0.027147045405626277
The coefficient for color_G is -0.04849184754312909
The coefficient for color_H is -0.08684095874110685
The coefficient for color_I is -0.11509875802594903
The coefficient for color_J is -0.12134930401467273
The coefficient for clarity_IF is 0.21010186299964867
The coefficient for clarity_SI1 is 0.32111616194547404
The coefficient for clarity_SI2 is 0.18955567488824587
The coefficient for clarity_VS1 is 0.35037491255023284
The coefficient for clarity_VS2 is 0.37605773633312284
The coefficient for clarity_VVS1 is 0.2776576039101753
The coefficient for clarity_VVS2 is 0.3184150588953828
```

```
The coefficient for carat is 1.2305882966567543
The coefficient for depth is 0.0014459929690719868
The coefficient for table is -0.013639708231939687
The coefficient for x is -0.3214591521321091
The coefficient for y is 0.25973210078172104
The coefficient for z is -0.10533359381691018
The coefficient for cut_Good is 0.035198594220606574
The coefficient for cut_Ideal is 0.09291368906273828
The coefficient for cut_Premium is 0.0793527046587005
The coefficient for cut_Very Good is 0.06686385068878037
The coefficient for color_E is -0.02149958578438857
The coefficient for color_F is -0.027147045405626277
The coefficient for color_G is -0.04849184754312909
The coefficient for color_H is -0.08684095874110685
The coefficient for color_I is -0.11509875802594903
The coefficient for color_J is -0.12134930401467273
The coefficient for clarity_IF is 0.21010186299964867
The coefficient for clarity_SI1 is 0.32111616194547404
The coefficient for clarity_SI2 is 0.18955567488824587
The coefficient for clarity_VS1 is 0.35037491255023284
The coefficient for clarity_VS2 is 0.37605773633312284
The coefficient for clarity_VVS1 is 0.2776576039101753
The coefficient for clarity_VVS2 is 0.3184150588953828
```

Train	Test

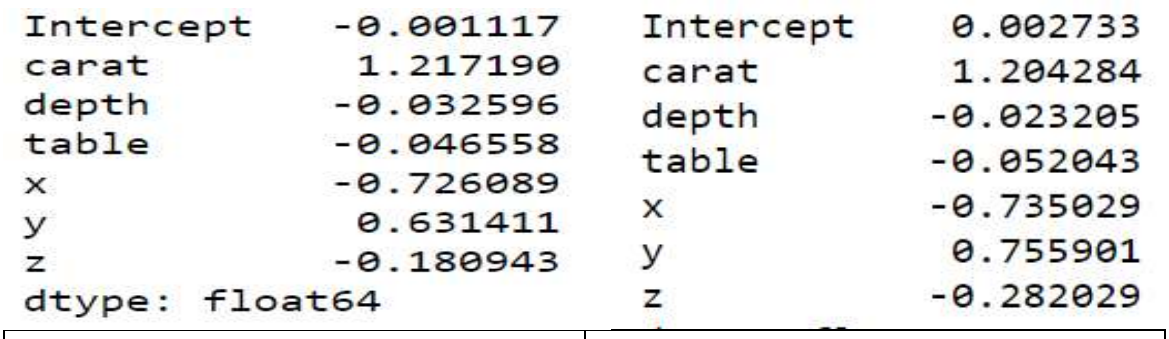
Fig-1.3a

B) Rsquare- Train & test:-

Train data coeff	Test data coeff
0.9413	0.9396

Tab-1.3.a

C) Lm _Train & test:-



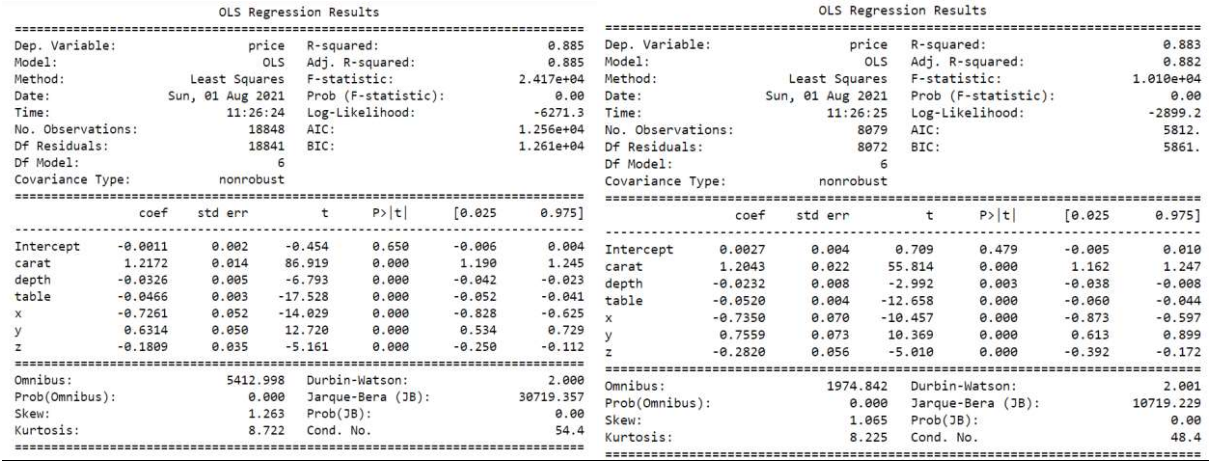
```
Intercept    -0.001117
carat         1.217190
depth        -0.032596
table        -0.046558
x            -0.726089
y             0.631411
z            -0.180943
dtype: float64
```

```
Intercept     0.002733
carat         1.204284
depth        -0.023205
table        -0.052043
x            -0.735029
y             0.755901
z            -0.282029
```

Train	Test

Fig-1.3.b

D) OLS regression result:-



OLS Regression Results

Dep. Variable:	price	R-squared:	0.885
Model:	OLS	Adj. R-squared:	0.885
Method:	Least Squares	F-statistic:	2.417e+04
Date:	Sun, 01 Aug 2021	Prob (F-statistic):	0.00
Time:	11:26:24	Log-Likelihood:	-6271.3
No. Observations:	18848	AIC:	1.256e+04
Df Residuals:	18841	BIC:	1.261e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0011	0.002	-0.454	0.650	-0.006	0.004
carat	1.2172	0.014	86.919	0.000	1.190	1.245
depth	-0.0326	0.005	-6.793	0.000	-0.042	-0.023
table	-0.0466	0.003	-17.528	0.000	-0.052	-0.041
x	-0.7261	0.052	-14.029	0.000	-0.828	-0.625
y	0.6314	0.050	12.720	0.000	0.534	0.729
z	-0.1809	0.035	-5.161	0.000	-0.250	-0.112

Omnibus:	5412.998	Durbin-Watson:	2.000
Prob(Omnibus):	0.000	Jarque-Bera (JB):	30719.357
Skew:	1.263	Prob(JB):	0.00
Kurtosis:	8.722	Cond. No.	54.4

OLS Regression Results

Dep. Variable:	price	R-squared:	0.883
Model:	OLS	Adj. R-squared:	0.882
Method:	Least Squares	F-statistic:	1.010e+04
Date:	Sun, 01 Aug 2021	Prob (F-statistic):	0.00
Time:	11:26:25	Log-Likelihood:	-2899.2
No. Observations:	8079	AIC:	5812.
Df Residuals:	8072	BIC:	5861.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0027	0.004	0.709	0.479	-0.005	0.010
carat	1.2043	0.022	55.814	0.000	1.162	1.247
depth	-0.0232	0.008	-2.992	0.003	-0.038	-0.008
table	-0.0520	0.004	-12.658	0.000	-0.060	-0.044
x	-0.7350	0.070	-10.457	0.000	-0.873	-0.597
y	0.7559	0.073	10.369	0.000	0.613	0.899
z	-0.2820	0.056	-5.010	0.000	-0.392	-0.172

Omnibus:	1974.842	Durbin-Watson:	2.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	10719.229
Skew:	1.065	Prob(JB):	0.00
Kurtosis:	8.225	Cond. No.	48.4

Train	Test

Fig-1.3.c

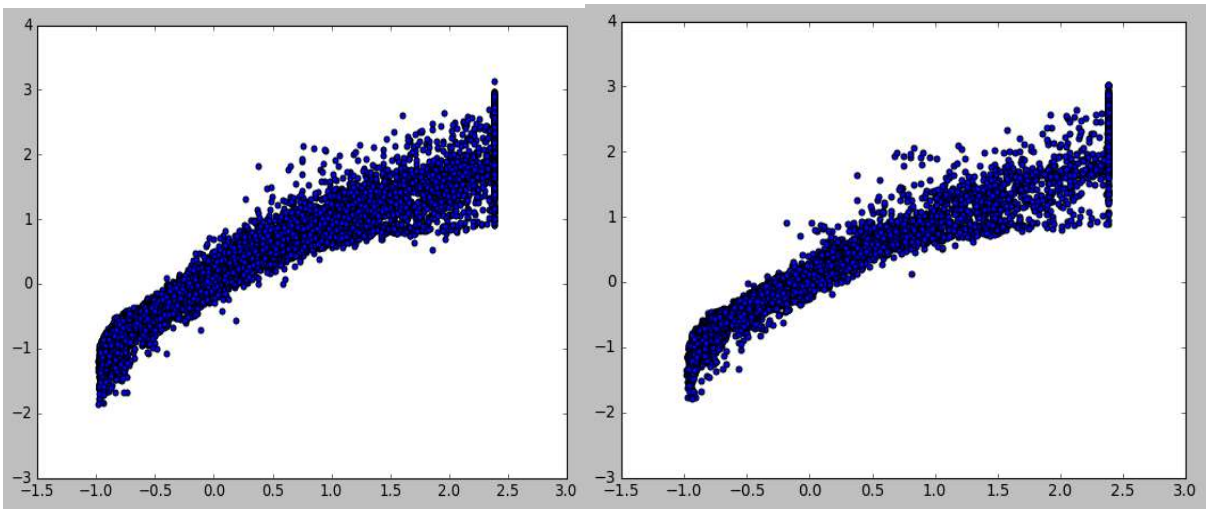
To ideally bring down the values to lower levels we can drop one of the variable that is highly correlated. Dropping variables would bring down the multi collinearity level down.

E) RMSE-report_train & test:-

Train	Test
0.2411	0.2483

Tab-1.3.b

F) LR- linear regression_best fit line:-



Train	Test

Fig-1.3.d

G) Calculation - Train & Test:-

The final Linear Regression equation is\n\nprice = b0 + b1 * carat + b2 * depth + b3 * table + b4 * x + b5 * y + b6 * z \n\nprice.

Train	(-0.0) * Intercept + (1.22) * carat + (-0.03) * depth + (-0.05) * table + (-0.73) * x + (0.63) * y + (-0.18) * z +
Test	(0.0) * Intercept + (1.2) * carat + (-0.02) * depth + (-0 .05) * table + (-0.74) * x + (0.76) * y + (-0.28) * z +

Tab-1.3.c

Conclusion | insight: As per question finding all parameter split data in two form train and test and find Coefficient , Rsquare- Train & test, Lm _Train & test, OLS regression result _train & test RMSE-report_train & test , Linear regression_best fit line and Calculation - Train & Test. After that we absorbed the resolute train and test data is quit-similar. All parameter is required to **final Linear Regression** not only train data 95%, rest test data also.

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

Conclusion | insight:-

We had a business to problem to predict the price of stone and provide insights for the company on the best profits on different prise slot .from EDA analysis we could understand the cut, ideal cut had number of profits to the company. The colours H, I, J have bought profits for the company. In clarity if we could see there were no flawless stones and they were no profits coming from I1, I2, I3 stones. The ideal, premium and very good types of cut were bringing profits where as fair and good are not bringing profits.

The predictions were able to capture 95% variations in the price and it is explained by the predictors in the training set.

Using stats model if we could run the model again we can have P values and coefficients which will give us better understanding of the relationship, so that values more 0.05 we can drop those variables and re run the model again for better results .

```
The coefficient for carat is 1.2305882966567543
The coefficient for depth is 0.0014459929690719868
The coefficient for table is -0.013639708231939687
The coefficient for x is -0.3214591521321091
The coefficient for y is 0.25973210078172104
The coefficient for z is -0.10533359381691018
The coefficient for cut_Good is 0.035198594220606574
The coefficient for cut_Ideal is 0.09291368906273828
The coefficient for cut_Premium is 0.0793527046587005
The coefficient for cut_Very Good is 0.06686385068878037
The coefficient for color_E is -0.02149958578438857
The coefficient for color_F is -0.027147045405626277
The coefficient for color_G is -0.04849184754312909
The coefficient for color_H is -0.08684095874110685
The coefficient for color_I is -0.11509875802594903
The coefficient for color_J is -0.12134930401467273
The coefficient for clarity_IF is 0.21010186299964867
The coefficient for clarity_SI1 is 0.32111616194547404
The coefficient for clarity_SI2 is 0.18955567488824587
The coefficient for clarity_VS1 is 0.35037491255023284
The coefficient for clarity_VS2 is 0.37605773633312284
The coefficient for clarity_VVS1 is 0.2776576039101753
The coefficient for clarity_VVS2 is 0.3184150588953828
```

The result of equation for best profits are shown below mention.

Train	(-0.0) * Intercept + (1.22) * carat + (-0.03) * depth + (-0.05) * table + (-0.73) * x + (0.63) * y + (-0.18) * z +
Test	(0.0) * Intercept + (1.2) * carat + (-0.02) * depth + (-0 .05) * table + (-0.74) * x + (0.76) * y + (-0.28) * z +

Recommendations

1. The ideal, premium, very good cut types are the one which are bringing profits so

That we could use marketing for these to bring in more profits.

2. The clarity of the diamond is the next important attributes the more the clear is the

Stone the profits are more.

3. Best of result is mention intercept (1.22) and carat is (-0.03), depth (-0.05), table (-0.73) coefficients was best result in this data.

4. Carat is (-0.03) than clarity is IF depth 62.1 price 1130 best product best profit in this segment.

Problem 2- Logistic Regression and LDA

Table of Contents

Contents

Table & Figures

Executive Summary:-

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

Introduction:-

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset and analysis of using method Logistic Regression and LDA. Find out the important factors on the basis of which the company will focus on particular employees to sell their packages. 872 employees of a company LDA and LR finding ROC and AUC train and test spit for this problem.

Data Description:-

Variable Name	Description
Holiday_Package	Opted for Holiday Package yes/no?
Salary	Employee salary
age	Age in years
edu	Years of formal education
no_young_children	The number of young children (younger than 7 years)
no_older_children	Number of older children
foreign	foreigner Yes/No

Tab-2.a

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

Fig-2.a

Conclusion | insight: This is Sample of given data all information is mention as per provide by travel agency which deals in selling holiday packages.

Data Describe:-

	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

Fig-2.b

Conclusion | insight: Data min value of data 0 and max value is 62 and other thing zero is available in data.

Q 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.

Solution:-

A) Exploratory data analysis:-

```
Data columns (total 8 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Unnamed: 0          872 non-null    int64
 1   Holliday_Package    872 non-null    object
 2   Salary              872 non-null    int64
 3   age                 872 non-null    int64
 4   educ                872 non-null    int64
 5   no_young_children   872 non-null    int64
 6   no_older_children   872 non-null    int64
 7   foreign             872 non-null    object
dtypes: int64(6), object(2)
memory usage: 54.6+ KB
```

Fig-2.1.b

Conclusion | insight: Ten types of information given in data as a column. sr.no, Holiday_package, salary, age, educ, no_young_childern, no_older_children, foreign, data is 54.6 kb + and 0 to 872 row and 8 columns. Two type of Data integer, Object. Two column are Object and six column are integer.

C) Data shape & size:-

```
df2.size
```

```
6104
```

```
df2.shape
```

```
(872, 7)
```

Fig-2.1.c

Conclusion | insight: Total 872 ROW and 7 Column.

E) Find Duplicated value:-

```
: df2.duplicated().sum()
```

```
: 0
```

Fig-2.1.d

Conclusion | insight: No any Duplicate value is find in this data.

F) Missing value:-

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

Fig-2.1.e

Conclusion | insight: In this data seat depth no any missing value find.

H) Univariate Analysis:-

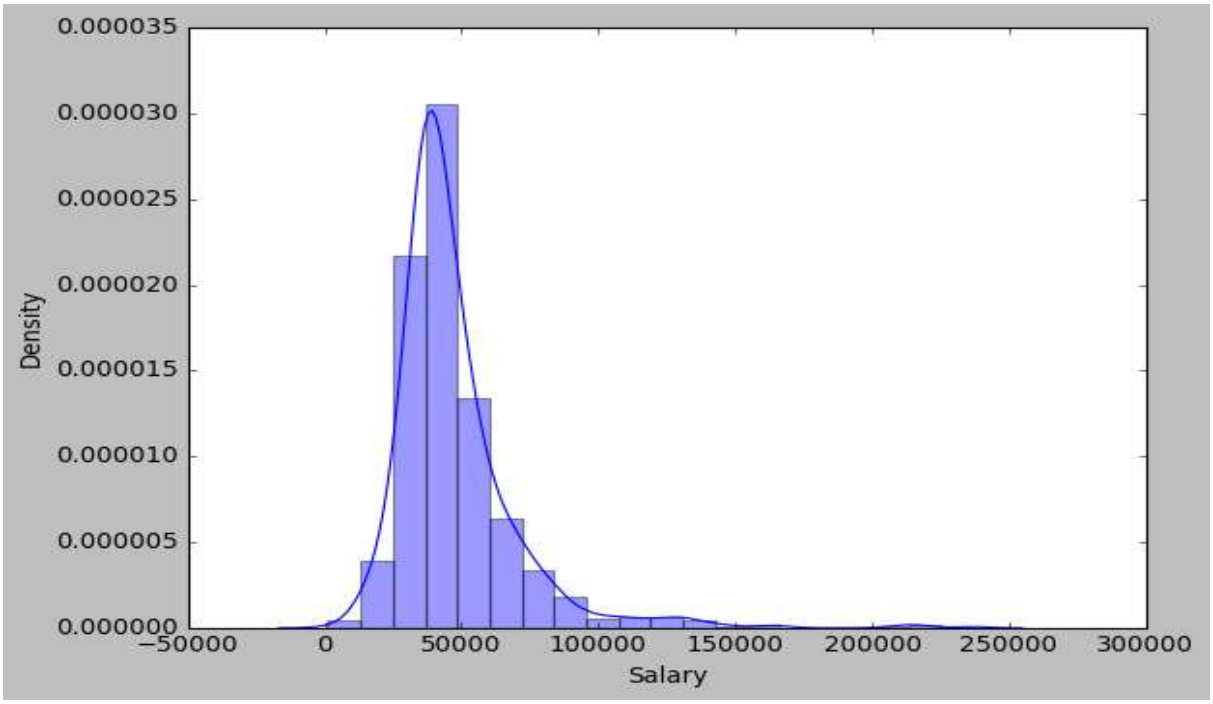


Fig-2.1.f

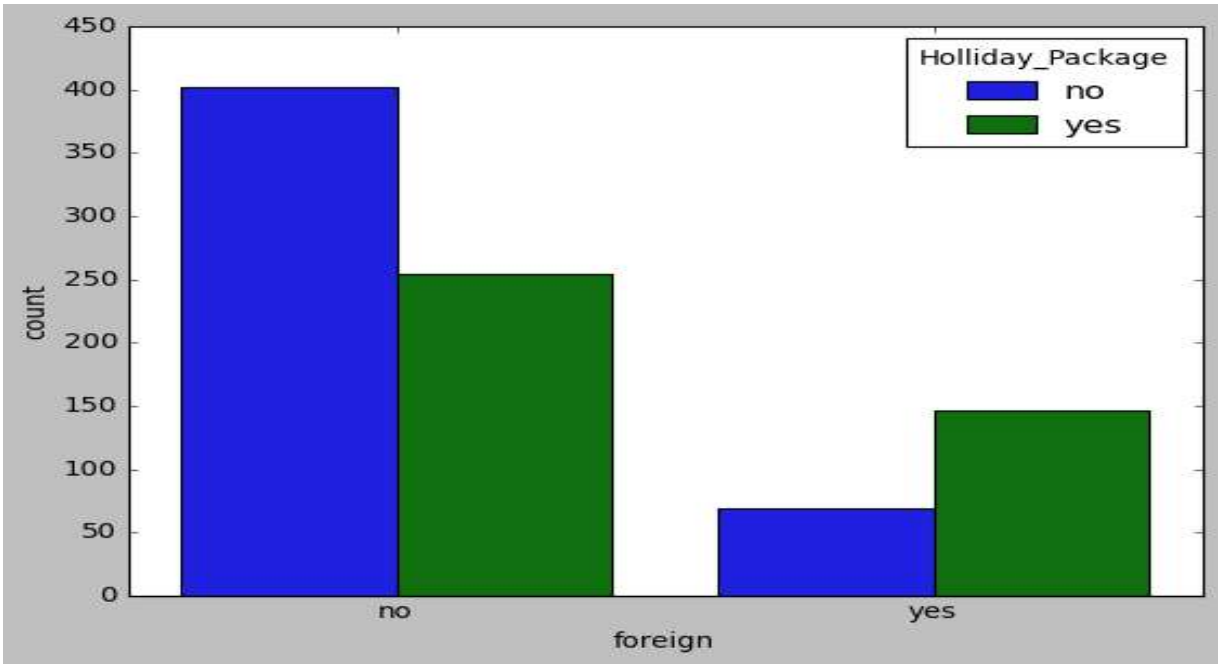


Fig-2.1.g

Conclusion | insight: We are find single| single variable value finding data behaviour show in fig-2.f, fig-2.g foreign | holliday_package or normal.

I) Bivariate Analysis:-

In single pair of all data

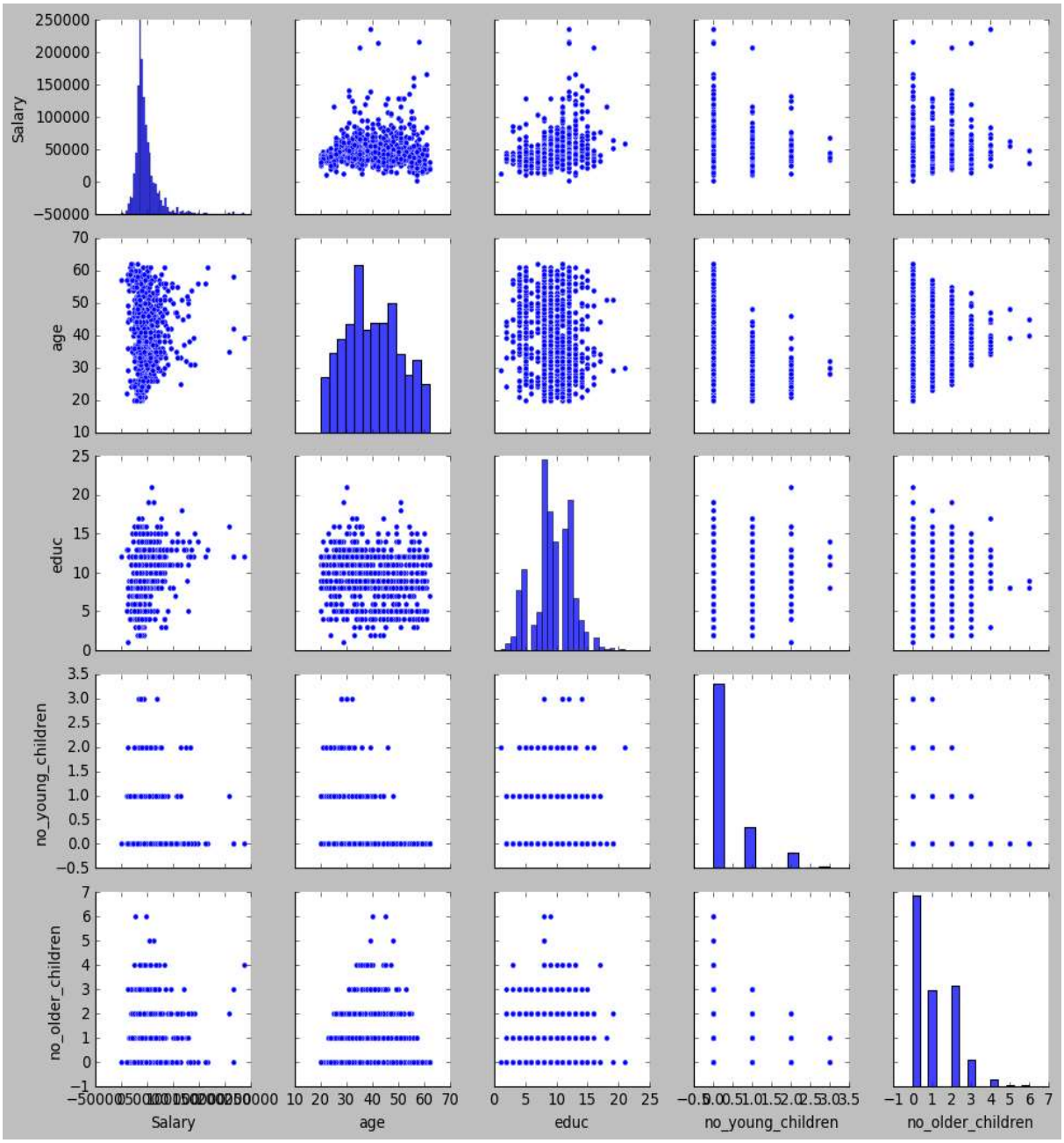


Fig-2.1.h

Outlier check in all data:-

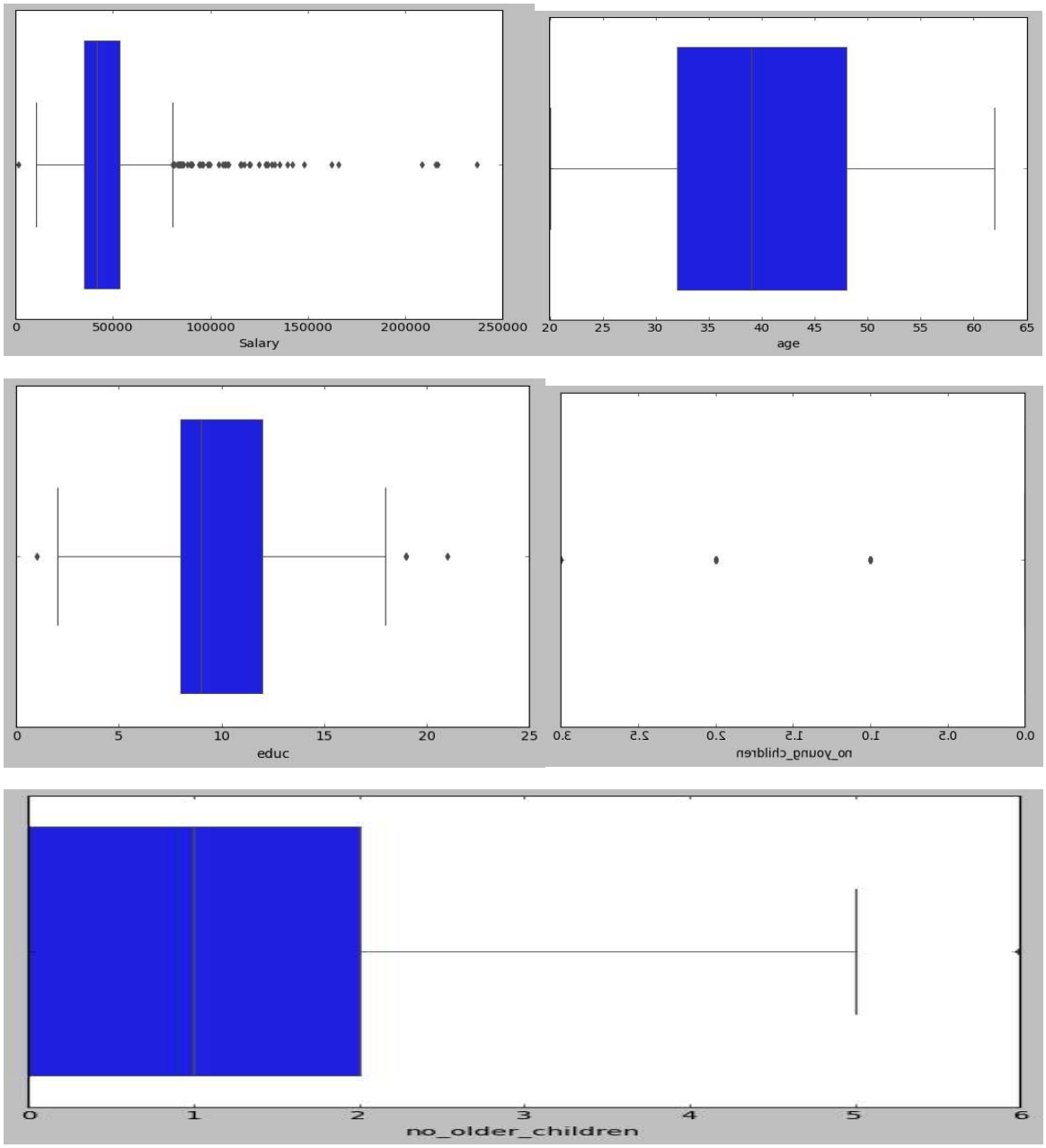


Fig-2.1.i

Conclusion | insight: Required outlier treating fig-2.i show almost outliner hair difference so we are treating all data outliers.

After Treating outliers check all data:-

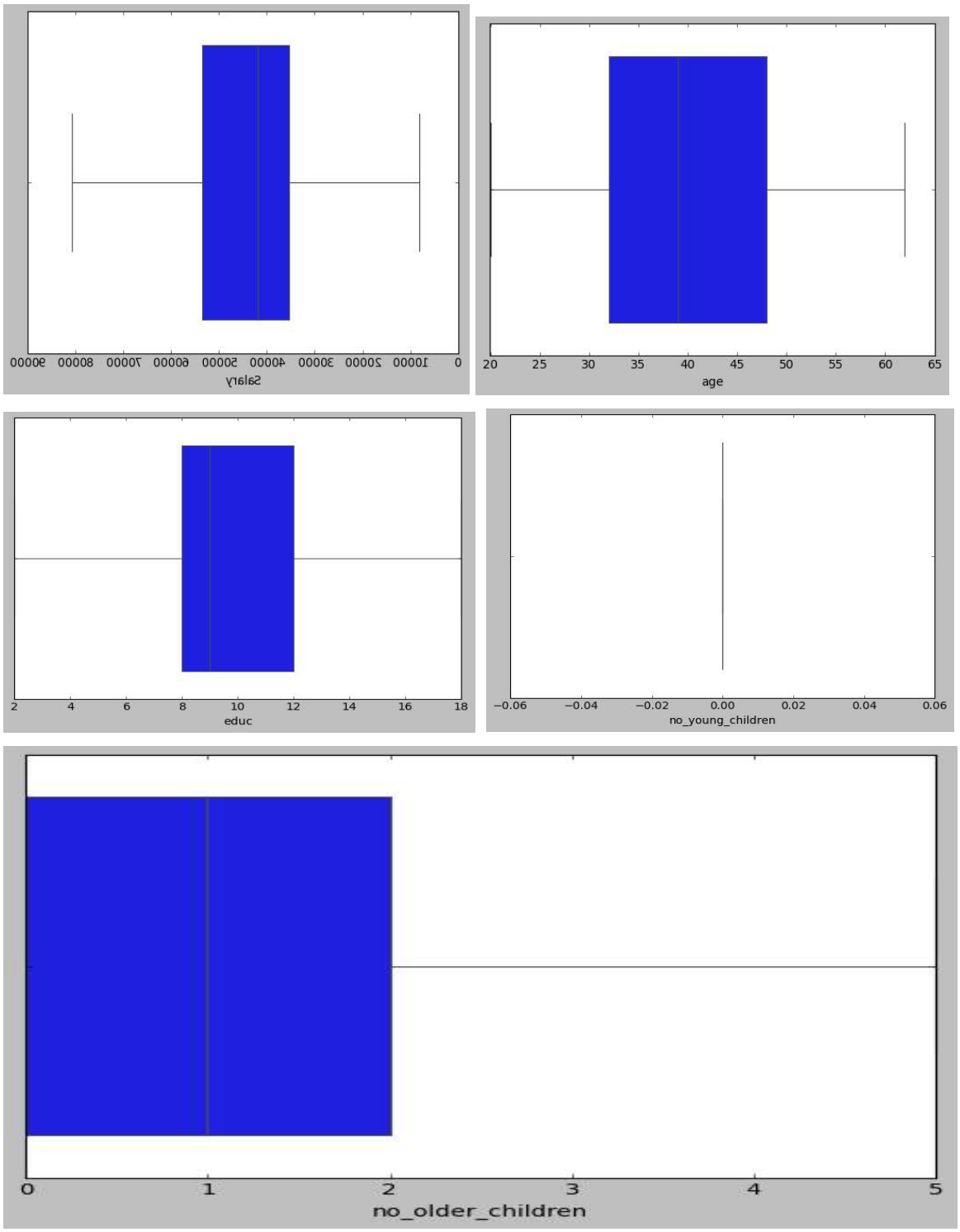


Fig-2.1.j

Correlation of all data:-

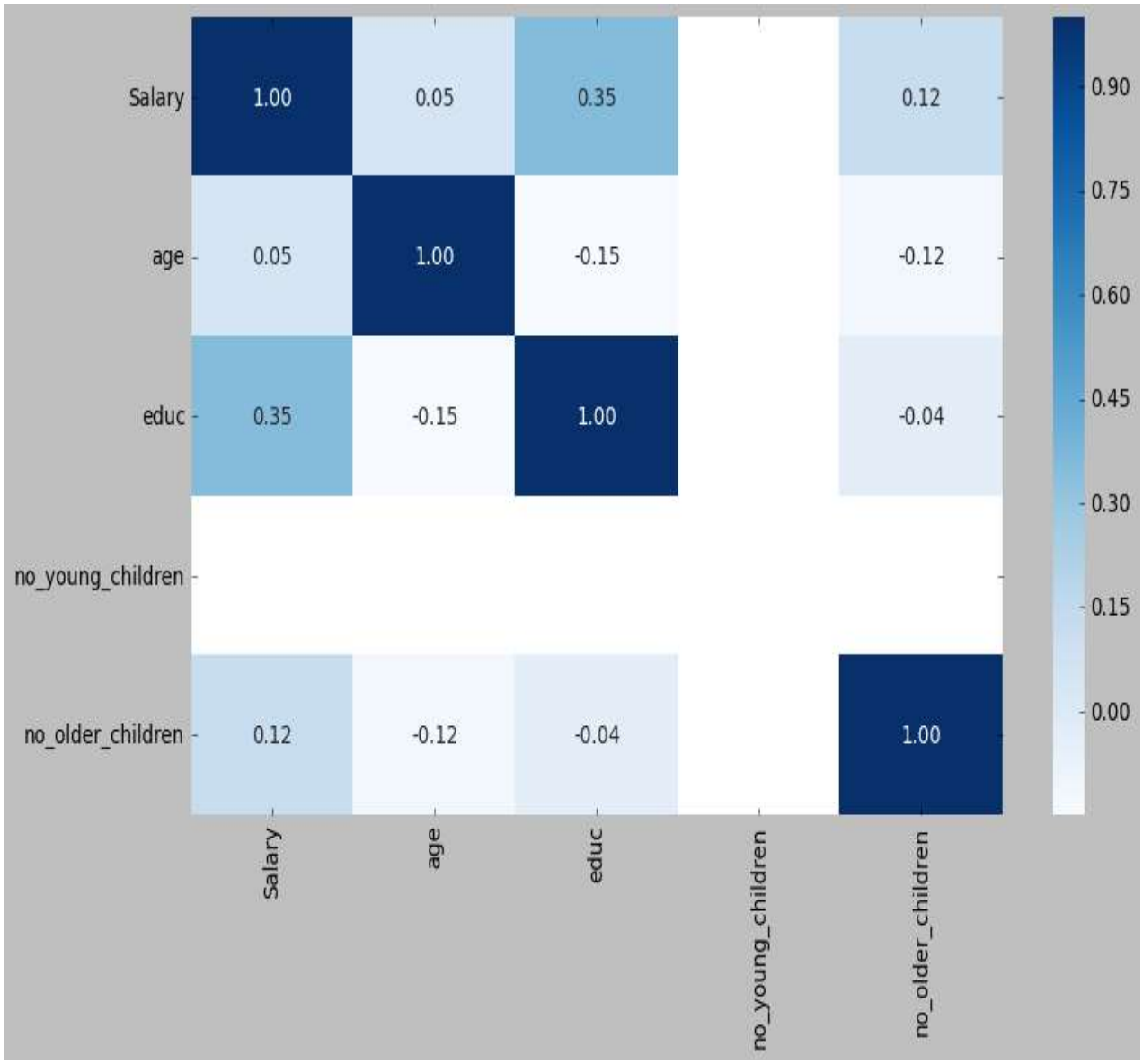


Fig-2.k

Conclusion | insight: In all data, we are comparing and showing. What behaving the data. Pair plot showing all data and boxplot for outlier check and treated outlier and correlation check each column for every column for best result.

Q 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).

Solution:-

Converting categorical to dummy variable in data

	Salary	age	educ	no_young_children	no_older_children	Holliday_Package_yes	foreign_yes
0	48412.00	30.0	8.0	0.0	1.0	0	0
1	37207.00	45.0	8.0	0.0	1.0	1	0
2	58022.00	46.0	9.0	0.0	0.0	0	0
3	66503.00	31.0	11.0	0.0	0.0	0	0
4	66734.00	44.0	12.0	0.0	2.0	0	0
5	61590.00	42.0	12.0	0.0	1.0	1	0
6	80687.75	51.0	8.0	0.0	0.0	0	0
7	35987.00	32.0	8.0	0.0	2.0	1	0
8	41140.00	39.0	12.0	0.0	0.0	0	0
9	35826.00	43.0	11.0	0.0	2.0	0	0

Fig-2.2.a

LR-Logistic Regression predict Train & Test (30:70)

```
model1 = LogisticRegression(solver='newton-cg', max_iter=10000, penalty='none', verbose=True, n_jobs=2)
model1.fit(X_train, y_train)
y_predict_test = model1.predict(X_test)
y_predict_train= model1.predict(X_train)
```

	0	1
0	0.696807	0.303193
1	0.332213	0.667787
2	0.620128	0.379872
3	0.686886	0.313114
4	0.354964	0.645036

X_test

	0	1
0	0.696807	0.303193
1	0.332213	0.667787
2	0.620128	0.379872
3	0.686886	0.313114
4	0.354964	0.645036

X_train

Fig-2.2.b

LDA (linear discriminant analysis) predict Train & Test

```
#Build LDA Model
model2 = LinearDiscriminantAnalysis()
model2.fit(X_train,y_train)
```

```
LinearDiscriminantAnalysis()
```

	0
0	0
1	1
2	0
3	0
4	0

	0
0	0
1	1
2	0
3	0
4	1

X_test	X_train

Tab-2.2.c

Conclusion | insight: The method gives, bilinear solver which is suitable for small datasets. Tolerance and penalty has been found using grid search method predicting the training & testing data.

Q 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.

Solution:-

A).1 LR-Logistic Regression- Train & Test (30:70)

X_Test Data		X_Train Data	
Accuracy_test	0.629	Accuracy train	0.640
Logistic_test_precision	0.62	Logistic_test_precision	0.67
Logistic_test_recall	0.44	Logistic_test_recall	0.44
Logistic_test_f1	0.52	Logistic_test_f1	0.54

Tab-2.3.1a

B) Classification report Train & Test for LR

	precision	recall	f1-score	support
0	0.63	0.78	0.70	145
1	0.62	0.44	0.52	117
accuracy			0.63	262
macro avg	0.63	0.61	0.61	262
weighted avg	0.63	0.63	0.62	262

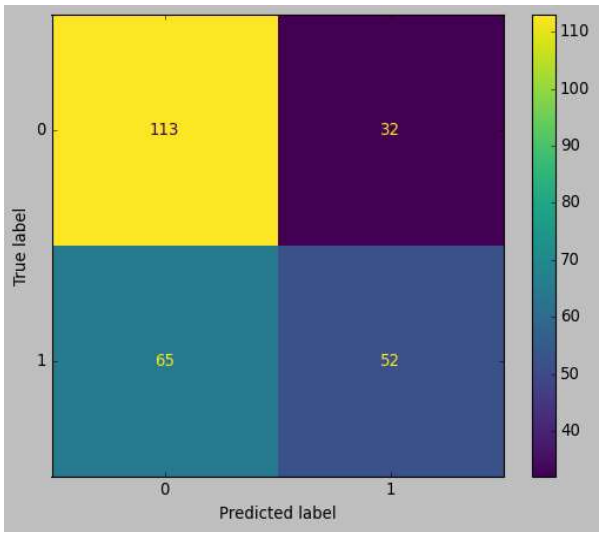
X_Test

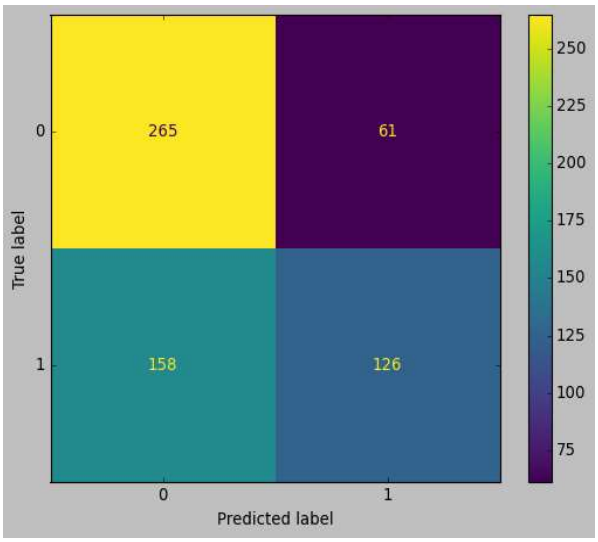
	precision	recall	f1-score	support
0	0.63	0.83	0.71	326
1	0.68	0.43	0.53	284
accuracy			0.64	610
macro avg	0.65	0.63	0.62	610
weighted avg	0.65	0.64	0.63	610

X_Train

Fig-2.3.1.a

C) Confusion matrix Train & Test for LR





X_Test

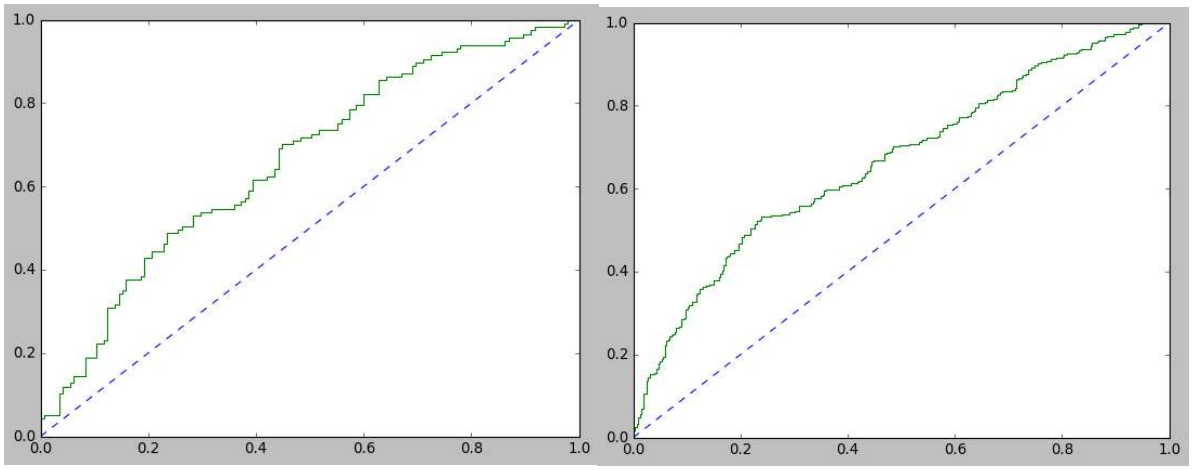
```
[[113  32]
 [ 65  52]]
```

X_Train

```
[[265  61]
 [158 126]]
```

Fig-2.3.1.b

D) AUC & ROC curve Train & Test for LR



Test_AUC- 0.661	Train_AUC- 0.667

Fig – 2.3.1.c

Conclusion | insight: LR:-Confusion matrix fig-2.3.1.b, coefficient of LR data AUC & ROC Train and Test data all are quiet similar also shown in fig-2.3.1.c .the are best observation in all test and train data .

A).2 LDA-(Linear discriminant analysis)-Train & Test

X_Train Data		X_Test Data	
Accuracy_test	0.642	Accuracy train	0.629
Logistic_test_precision	0.68	Logistic_test_precision	0.62
Logistic_test_recall	0.43	Logistic_test_recall	0.44
Logistic_test_f1	0.53	Logistic_test_f1	0.52

Tab – 2.3.2.a

B) Classification report Train & Test for LDA

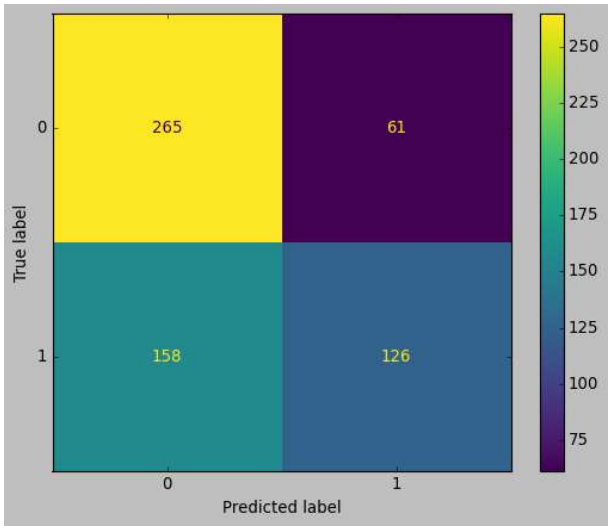
	precision	recall	f1-score	support
0	0.63	0.81	0.71	326
1	0.67	0.44	0.54	284
accuracy			0.64	610
macro avg	0.65	0.63	0.62	610
weighted avg	0.65	0.64	0.63	610

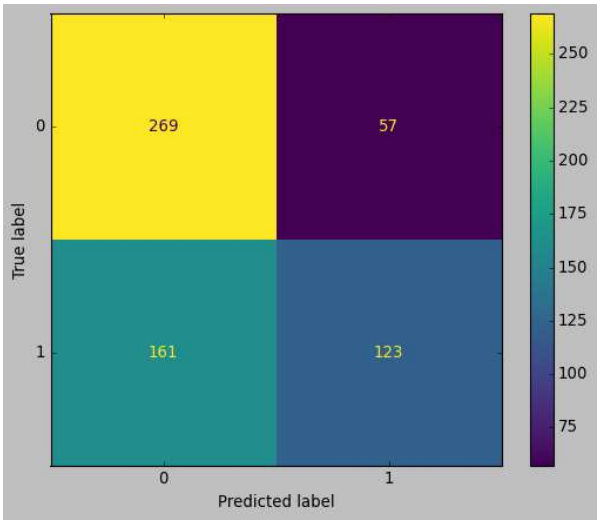
	precision	recall	f1-score	support
0	0.63	0.83	0.71	326
1	0.68	0.43	0.53	284
accuracy			0.64	610
macro avg	0.65	0.63	0.62	610
weighted avg	0.65	0.64	0.63	610

X_Train	X_Test

Fig-2.3.2.b

C) Confusion matrix Train & Test for LDA





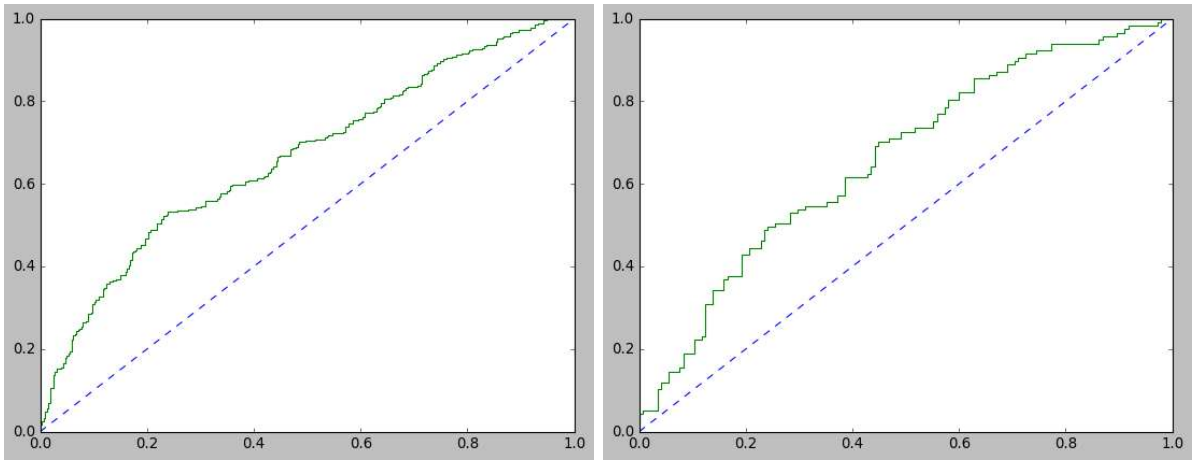
X_Train	X_Test

```
[[265  61]
 [158 126]]
```

```
[[269  57]
 [161 123]]
```

Fig-2.3.2.c

D) AUC & ROC curve Train & Test for LDA



X_Train AUC=0.667	X_Test AUC=0.667

Fig-2.3.2.c

Conclusion | insight: LDA Confusion matrix fig-2.3.2.b, coefficient of LR data AUC & ROC Train and Test data all are quiet similar also shown in fig-2.3.2.c .the are best observation in all test and train data .

A).3 Final Model - Compare LR & LDA

	LR Train	LR Test	LDA Train	LDA Test
Accuracy	0.641	0.630	0.643	0.630
AUC	0.667	0.661	0.667	0.662
Recall	0.440	0.440	0.430	0.440
Precision	0.670	0.620	0.680	0.620
F1 Score	0.540	0.520	0.530	0.520

Fig-2.3.3.a

B) Compare- Train LR & LDA

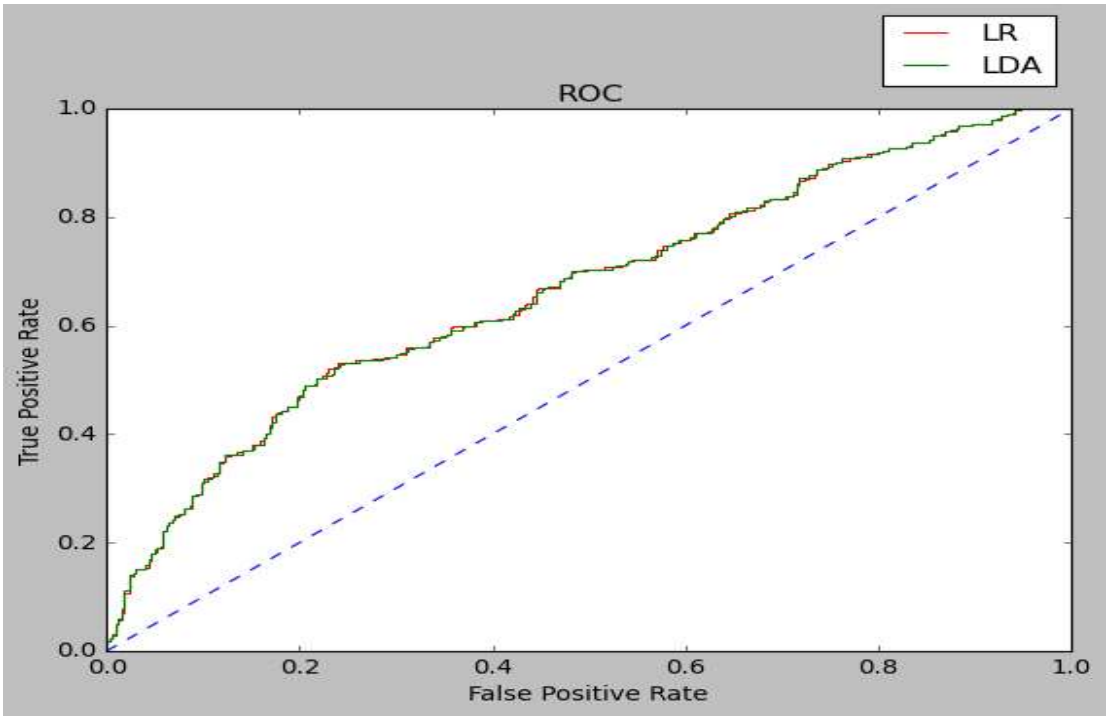


Fig-2.3.3.b

C) Compare- Test LR & LDA

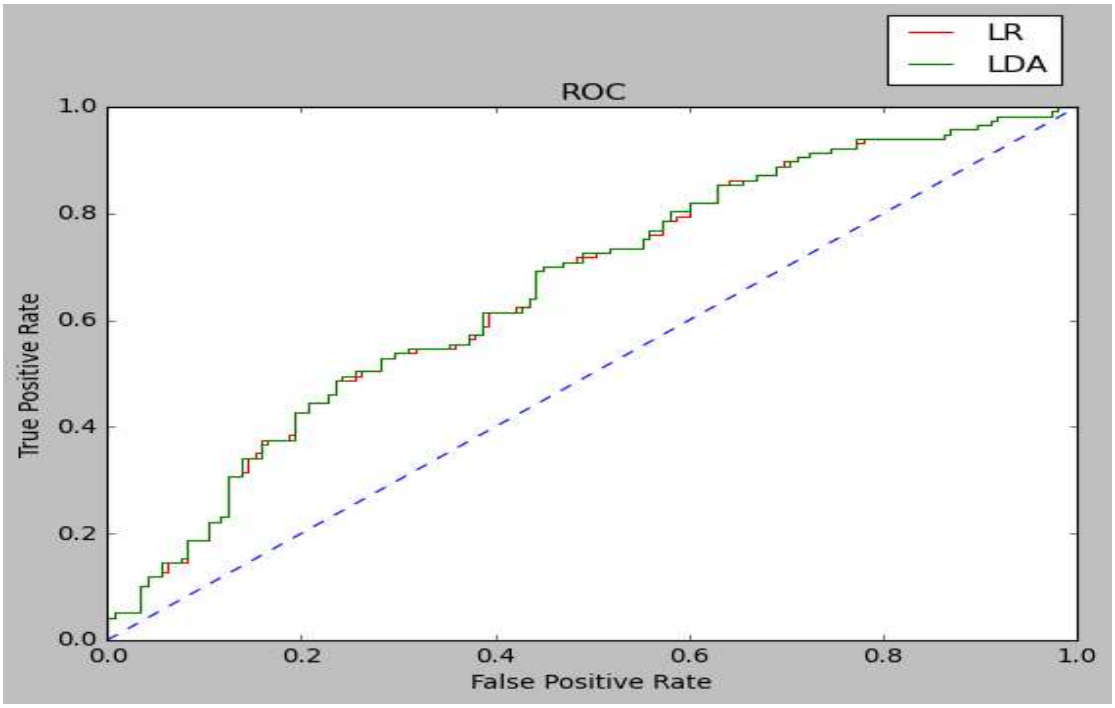


Fig-2.3.3.c

Conclusion | insight: - Comparing both these models, we find both results are same, but LDA works better when there is category target variable.

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

Solution:-

Conclusion | insight: -

We had a business problem where we need predict whether an employee would opt for a holiday package or not, for this problem we had done predictions both logistic regression and linear discriminant analysis. Since both are results are The EDA analysis clearly indicates certain criteria where we could find people aged above 50 are not interested much in holiday packages. So this is one of the we find aged people not opting for holiday packages People ranging from the age 30 to 50 generally opt for holiday packages. Employee age over 50 to 60 have seems to be not taking the holiday package,

Whereas in the age 30 to 50 and salary less than 50000 people have opted more for holiday package. The important factors deciding the predictions are salary, age and educ.

	LR Train	LR Test	LDA Train	LDA Test
Accuracy	0.641	0.630	0.643	0.630
AUC	0.667	0.661	0.667	0.662
Recall	0.440	0.440	0.430	0.440
Precision	0.670	0.620	0.680	0.620
F1 Score	0.540	0.520	0.530	0.520

Recommendations

1. To improve holiday packages over the age above 50 we can provide religious destination places.

2. For people earning more than 150000 we can provide vacation holiday packages.

3. for employee having more than number of older children we can provide packages in holiday vacation places.

4. Most of under holiday packages person age 20 to above and out of country plan and without holiday packages person also.

---- END ---