BUSINESS REPORT

Problem 1: Linear Regression

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.

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.

- Importing data into df –

Head of 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

We can say -

1. There is no use of column unnamed:0

2. Our dependent variable will be price for Linear Regression model.

Shape of data –

(26967, 11)

BUSINESS REPORT

After dropping Unnamed:0 column from dataset, checking the info-

```
<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
```

We can say -

1. Dataset has 4 object type/categorical variables namely cut, color, clarity
2. Dataset has 6 float type variables namely carat, depth, table, x, y, z, price
3. Dataset has 1 int type variable which is price.

Checking 5-point summary -

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
carat	26967	NaN	NaN	NaN	0.798375	0.477745	0.2	0.4	0.7	1.05	4.5
cut	26967	5	Ideal	10816	NaN	NaN	NaN	NaN	NaN	NaN	NaN
color	26967	7	G	5661	NaN	NaN	NaN	NaN	NaN	NaN	NaN
clarity	26967	8	SI1	6571	NaN	NaN	NaN	NaN	NaN	NaN	NaN
depth	26270	NaN	NaN	NaN	61.7451	1.41286	50.8	61	61.8	62.5	73.6
table	26967	NaN	NaN	NaN	57.4561	2.23207	49	56	57	59	79
x	26967	NaN	NaN	NaN	5.72985	1.12852	0	4.71	5.69	6.55	10.23
y	26967	NaN	NaN	NaN	5.73357	1.16606	0	4.71	5.71	6.54	58.9
z	26967	NaN	NaN	NaN	3.53806	0.720624	0	2.9	3.52	4.04	31.8
price	26967	NaN	NaN	NaN	3939.52	4024.86	326	945	2375	5360	18818

BUSINESS REPORT

We can say –

1. There are total 26967 records

2. Depth has 26270 records. It states that it has some missing values.

3. Cut, colour, clarity has some unique values

4. By looking at 5-point summary it seems that some variables have outliers.

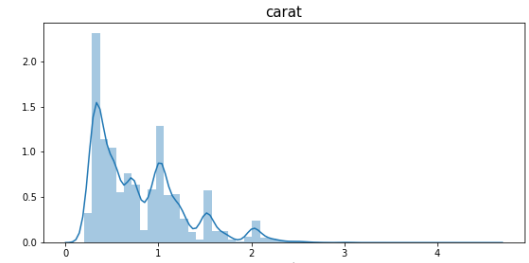
Checking Null values –

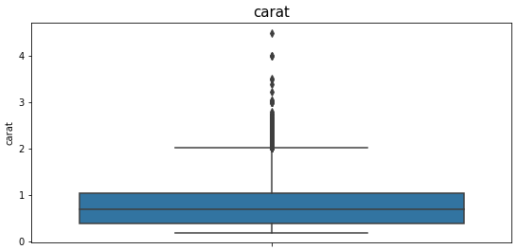
```
carat        0
cut          0
color        0
clarity      0
depth      697
table        0
x            0
y            0
z            0
price        0
dtype: int64
```

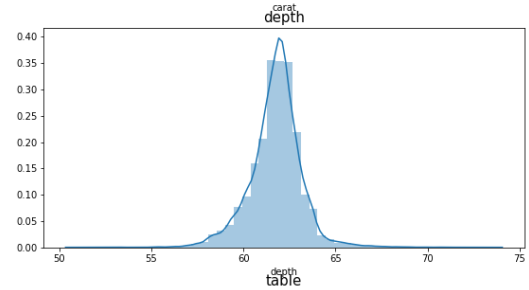
We can see that column depth has null/missing values.

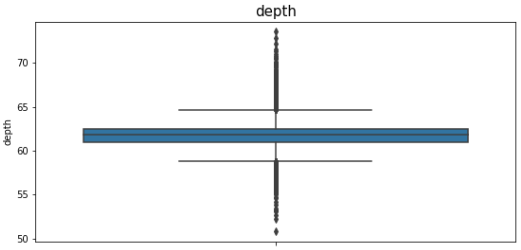
Univariate Analysis-

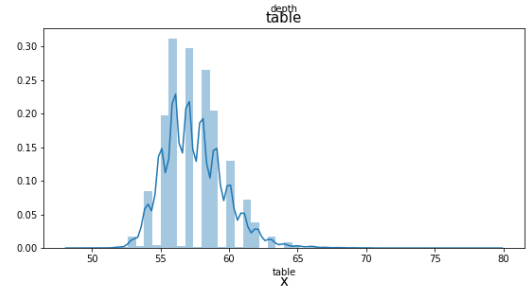
BUSINESS REPORT

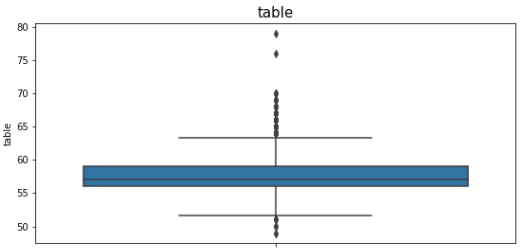


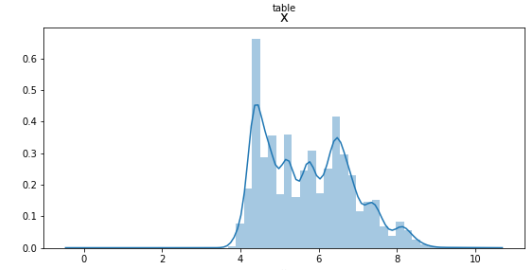


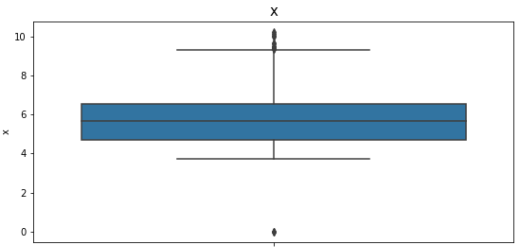


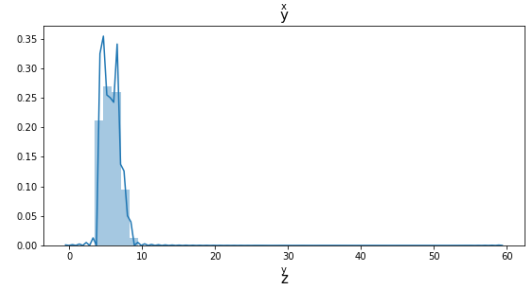


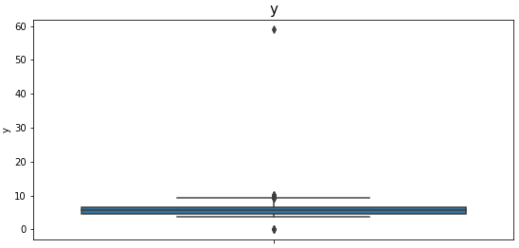


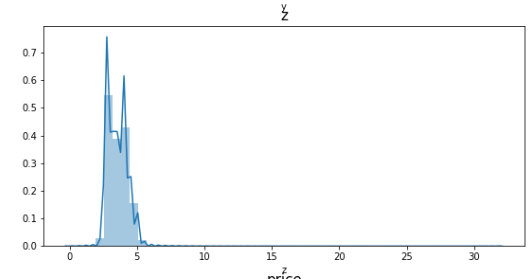


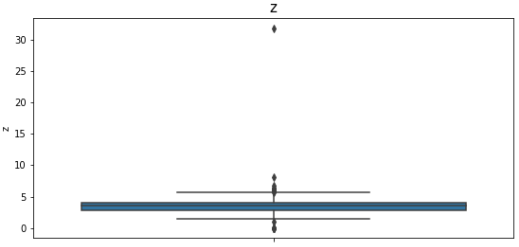


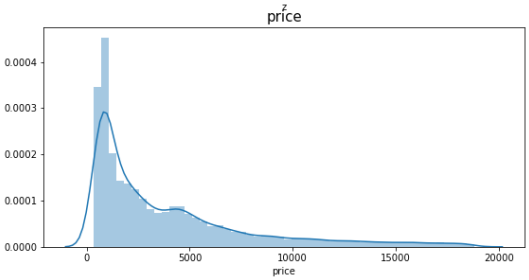


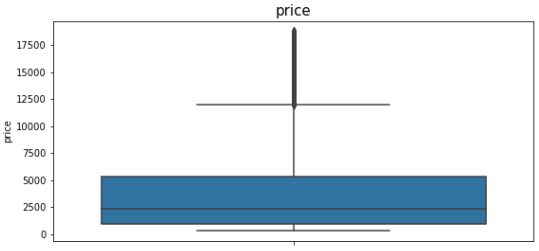












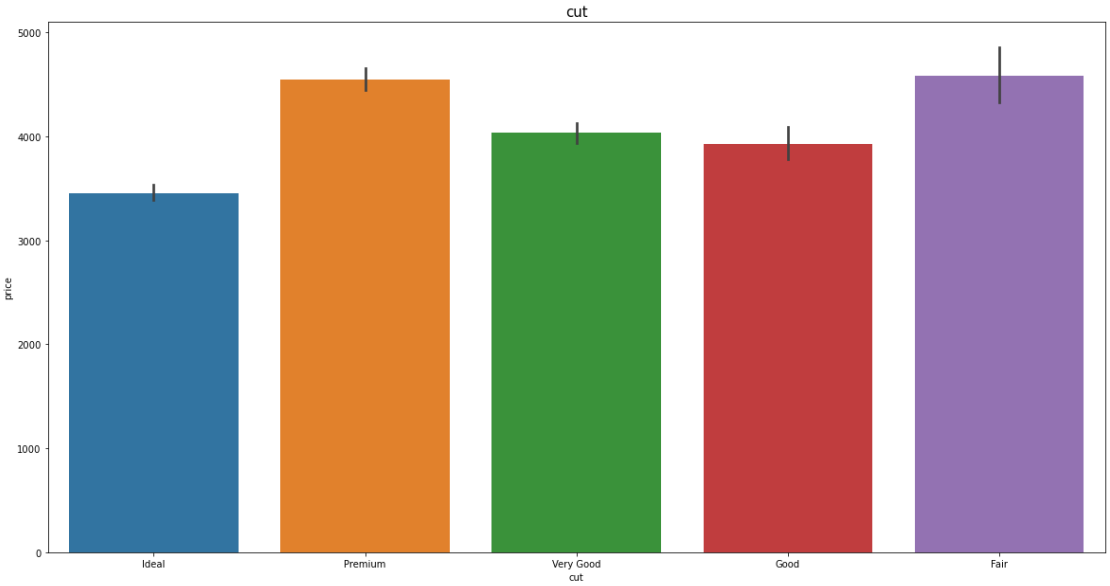
BUSINESS REPORT

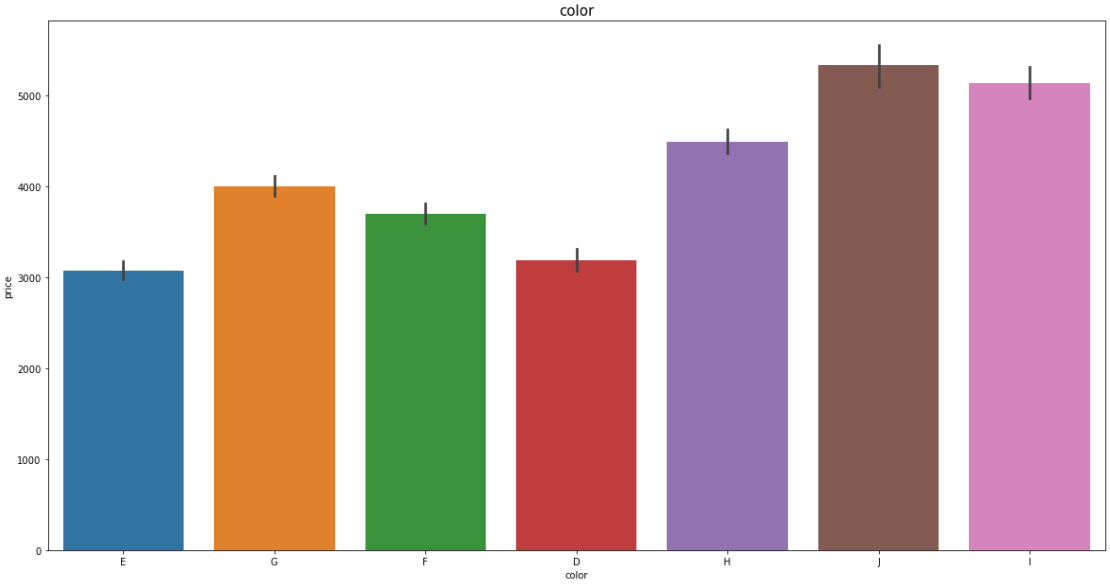
We can say –

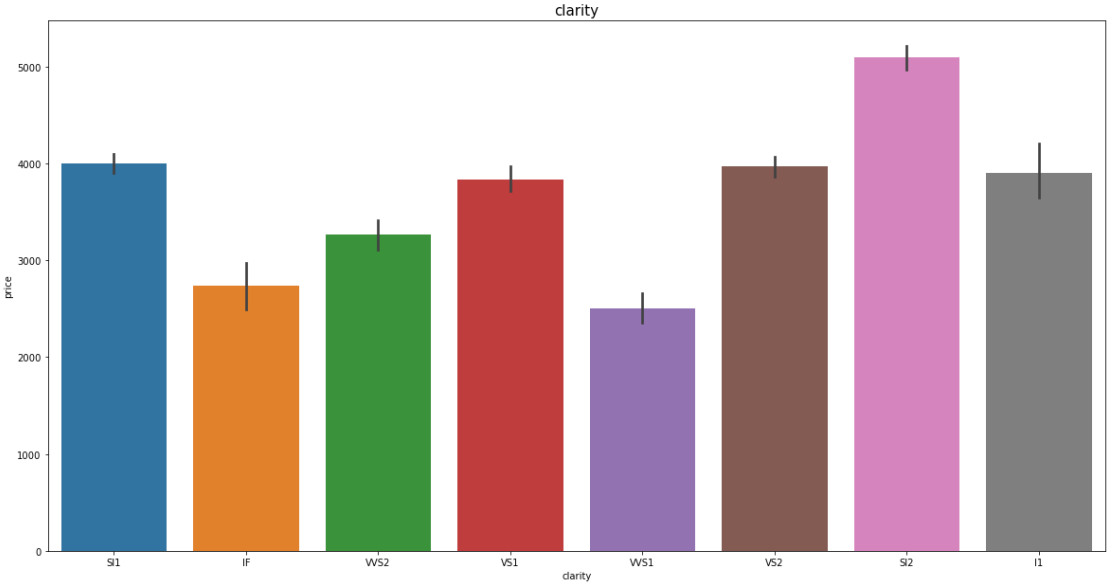
1. Distribution of column carat is right skewed and has outliers.
2. Distribution of column depth looks normally distributed and also has outliers.
3. Distribution of column table is right skewed and also has outliers.
4. Distribution of column x is right skewed and also has outliers.
5. Distribution of column y is right skewed and also has outliers.
6. Distribution of column z is right skewed and also has outliers.
7. Distribution of column price is right skewed and also has outliers.

Bi- Variate Analysis-

BUSINESS REPORT

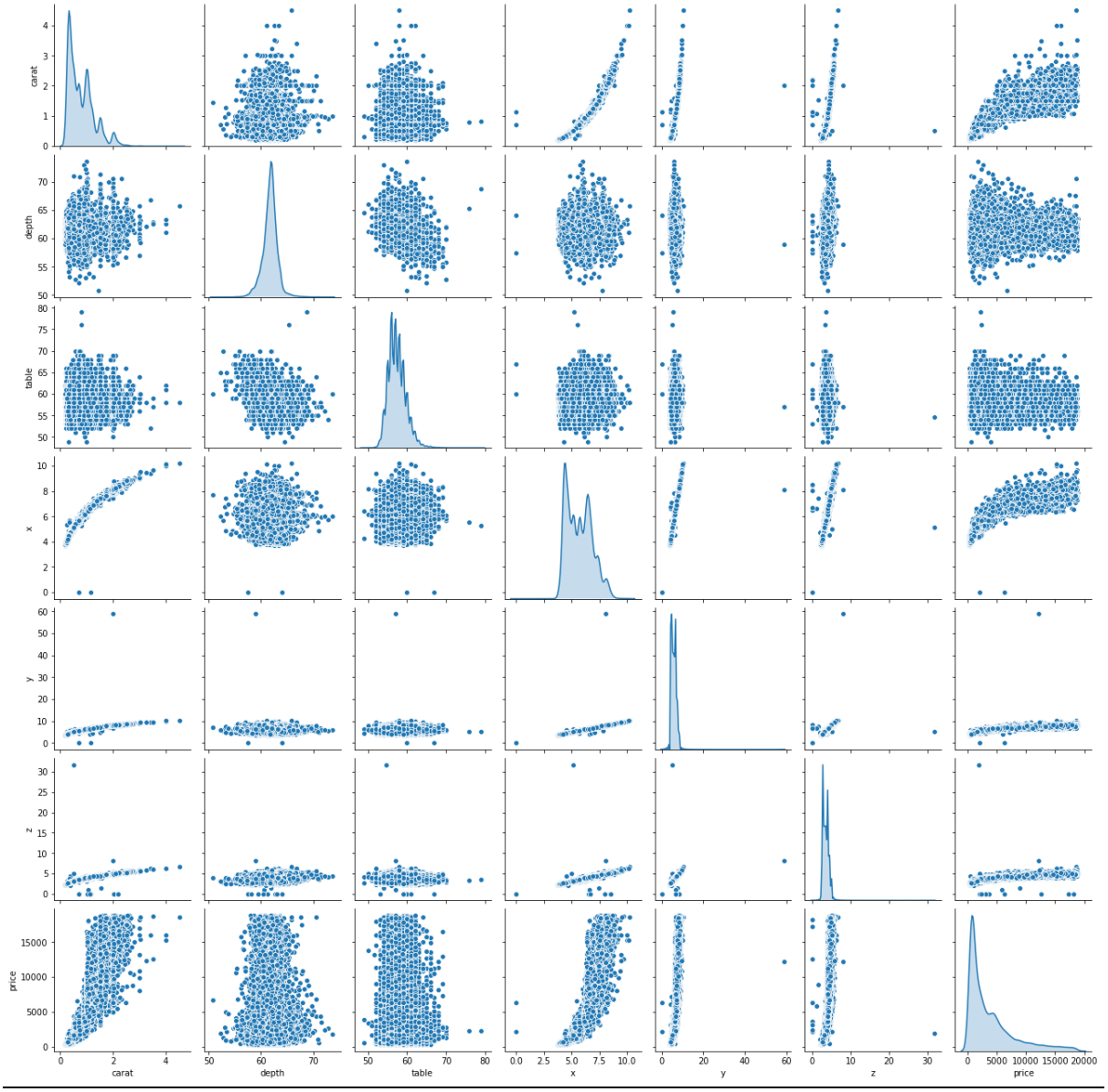






BUSINESS REPORT

Multivariate Analysis –



BUSINESS REPORT

Heatmap –

We can say carat, x, y, Z variables have correlation with each other.

Checking Outliers

```
carat        662
clarity        0
color          0
cut            0
depth       1225
price       1779
table        318
x             15
y             15
z             23
```

There are outliers present in the carat , depth , price ,table ,x, y, z columns and will be removed.

BUSINESS REPORT

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? Do you think scaling is necessary in this case?

Checking Null Values –

```
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 depth column.

As there are outliers in depth variable, we should impute the null values with median.

After Imputing –

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

Regarding Scaling –

Yes, we need scaling in this as dimensions of cubic zirconia diamond that is length, width and height are in mm and to scaling helps in normalizing the data.

It is always good idea to do scaling as the model performance significantly increases.

BUSINESS REPORT

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

Encoding of data –

['Ideal' 'Premium' 'Very Good' 'Good' 'Fair']

['E' 'G' 'F' 'D' 'H' 'J' 'I']

['SI1' 'IF' 'VVS2' 'VS1' 'VVS1' 'VS2' 'SI2' 'I1']

After Encoding checking the data head –

	carat	cut	color	clarity	depth	table	x	y	z	price
0	0.30	5	6	5	62.1	58.0	4.27	4.29	2.66	499.0
1	0.33	4	4	10	60.8	58.0	4.42	4.46	2.70	984.0
2	0.90	3	6	8	62.2	60.0	6.04	6.12	3.78	6289.0
3	0.42	5	5	7	61.6	56.0	4.82	4.80	2.96	1082.0
4	0.31	5	5	9	60.4	59.0	4.35	4.43	2.65	779.0

Checking the info –

```
<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  int64
 2   color    26967 non-null  int64
 3   clarity  26967 non-null  int64
 4   depth    26967 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  float64
dtypes: float64(7), int64(3)
memory usage: 2.1 MB
```

BUSINESS REPORT

Separating independent and dependent variables –

	carat	cut	color	clarity	depth	table	x	y	z
0	0.30	5	6	5	62.1	58.0	4.27	4.29	2.66
1	0.33	4	4	10	60.8	58.0	4.42	4.46	2.70
2	0.90	3	6	8	62.2	60.0	6.04	6.12	3.78
3	0.42	5	5	7	61.6	56.0	4.82	4.80	2.96
4	0.31	5	5	9	60.4	59.0	4.35	4.43	2.65

	price
0	499.0
1	984.0
2	6289.0
3	1082.0
4	779.0

Splitting data into Train and Test data –

X_Train head –

BUSINESS REPORT

	carat	cut	color	clarity	depth	table	x	y	z
11687	0.41	5	2	8	62.3	56.0	4.77	4.73	2.96
9728	1.71	5	1	5	62.8	57.0	7.58	7.55	4.75
1936	0.33	2	5	5	61.8	62.0	4.40	4.45	2.74
26220	0.70	3	3	5	62.8	57.0	5.61	5.66	3.54
18445	0.70	5	7	4	62.1	56.0	5.67	5.71	3.53
...	...	...	...	...	...	...	...	...	...
10955	0.21	4	7	6	61.6	59.0	3.82	3.78	2.34
17289	0.31	4	3	6	61.9	58.0	4.39	4.37	2.71
5192	2.01	5	6	4	62.2	57.0	8.04	8.09	5.02
12172	1.54	5	2	4	60.3	59.0	7.43	7.46	4.49
235	1.21	4	3	10	62.2	58.0	6.83	6.80	4.24

18876 rows × 9 columns

X Test head –

	carat	cut	color	clarity	depth	table	x	y	z
18031	2.01	1	2	4	66.5	61.0	7.81	7.75	5.17
26051	1.51	4	5	5	62.2	59.0	7.34	7.30	4.55
16279	0.50	3	3	5	60.9	61.0	5.06	5.15	3.11
16466	0.31	5	6	7	61.8	56.0	4.39	4.44	2.66
19837	1.20	3	3	7	62.0	57.0	6.77	6.81	4.21
...	...	...	...	...	...	...	...	...	...
9716	0.32	5	2	10	62.2	57.0	4.37	4.35	2.71
9944	0.70	4	3	8	62.1	58.0	5.71	5.66	3.53
9858	0.71	5	4	4	61.1	56.0	5.75	5.77	3.52
4075	1.51	3	1	4	62.8	59.0	7.23	7.26	4.55
17732	0.33	3	7	8	60.8	58.0	4.47	4.48	2.72

8091 rows × 9 columns

BUSINESS REPORT

After Scaling the data –

X – Train head -

	carat	cut	color	clarity	depth	table	x	y	z
11687	-0.828710	0.987756	-1.401996	1.187046	0.395675	-0.669095	-0.851279	-0.895171	-0.827069
9728	1.983254	0.987756	-1.988276	-0.642867	0.754075	-0.204361	1.643548	1.626182	1.740161
1936	-1.001754	-1.707630	0.356843	-0.642867	0.037275	2.119307	-1.179779	-1.145518	-1.142594
26220	-0.201426	-0.809168	-0.815717	-0.642867	0.754075	-0.204361	-0.105494	-0.063661	0.004771
18445	-0.201426	0.987756	1.529403	-1.252839	0.252315	-0.669095	-0.052224	-0.018956	-0.009571
...	...	...	...	...	...	...	...	...	...
10955	-1.261320	0.089294	1.529403	-0.032896	-0.106084	0.725106	-1.694725	-1.744562	-1.716277
17289	-1.045015	0.089294	-0.815717	-0.032896	0.108955	0.260372	-1.188657	-1.217045	-1.185620
5192	2.632168	0.987756	0.943123	-1.252839	0.323995	-0.204361	2.051954	2.108994	2.127396
12172	1.615535	0.987756	-1.401996	-1.252839	-1.037923	0.725106	1.510372	1.545714	1.367267
235	0.901729	0.089294	-0.815717	2.406988	0.323995	0.260372	0.977669	0.955610	1.008715

18876 rows × 9 columns

X-Test head –

	carat	cut	color	clarity	depth	table	x	y	z
18031	2.626661	-2.631987	-1.406184	-1.232242	3.414634	1.647329	1.840904	1.800533	2.339552
26051	1.546003	0.062497	0.351673	-0.631491	0.327897	0.725417	1.424445	1.398790	1.450408
16279	-0.636926	-0.835664	-0.820232	-0.631491	-0.605303	1.647329	-0.595826	-0.520650	-0.614703
16466	-1.047576	0.960659	0.937625	0.570012	0.040759	-0.657450	-1.189502	-1.154511	-1.260050
19837	0.875995	-0.835664	-0.820232	0.570012	0.184328	-0.196494	0.919377	0.961336	0.962812
...	...	...	...	...	...	...	...	...	...
9716	-1.025963	0.960659	-1.406184	2.372266	0.327897	-0.196494	-1.207223	-1.234860	-1.188345
9944	-0.204663	0.062497	-0.820232	1.170763	0.256112	0.264461	-0.019871	-0.065341	-0.012379
9858	-0.183049	0.960659	-0.234280	-1.232242	-0.461733	-0.657450	0.015572	0.032863	-0.026720
4075	1.546003	-0.835664	-1.992136	-1.232242	0.758604	0.725417	1.326975	1.363079	1.450408
17732	-1.004349	-0.835664	1.523577	1.170763	-0.677087	0.264461	-1.118615	-1.118800	-1.174003

8091 rows × 9 columns

BUSINESS REPORT

Coefficient for each of the independent variables –

```
The coefficient for carat is 1.178540910600686
The coefficient for cut is 0.034729711277533014
The coefficient for color is 0.13416117479808776
The coefficient for clarity is 0.20358085654236643
The coefficient for depth is -0.010286781466753308
The coefficient for table is -0.01110628969808213
The coefficient for x is -0.4880138318099923
The coefficient for y is 0.38324720303191334
The coefficient for z is -0.020798557388008263
```

Intercept of Model –

```
The intercept for our model is -5.746644888976901e-16
```

Coefficient of Determinant for train data - 0.9317074064221308

Coefficient of Determinant for test data - 0.9298441146329623

BUSINESS REPORT

Summary of the model -

OLS Regression Results

Dep. Variable:	price	R-squared:	0.932
Model:	OLS	Adj. R-squared:	0.932
Method:	Least Squares	F-statistic:	2.860e+04
Date:	Sun, 30 Aug 2020	Prob (F-statistic):	0.00
Time:	00:20:40	Log-Likelihood:	-1452.7
No. Observations:	18876	AIC:	2925.
Df Residuals:	18866	BIC:	3004.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.214e-17	0.002	6.38e-15	1.000	-0.004	0.004
carat	1.1785	0.011	110.437	0.000	1.158	1.199
cut	0.0347	0.002	14.961	0.000	0.030	0.039
color	0.1342	0.002	66.574	0.000	0.130	0.138
clarity	0.2036	0.002	96.315	0.000	0.199	0.208
depth	-0.0103	0.003	-3.245	0.001	-0.017	-0.004
table	-0.0111	0.002	-4.586	0.000	-0.016	-0.006
x	-0.4880	0.042	-11.623	0.000	-0.570	-0.406
y	0.3832	0.042	9.094	0.000	0.301	0.466
z	-0.0208	0.018	-1.138	0.255	-0.057	0.015

Omnibus:	2652.334	Durbin-Watson:	1.984
Prob(Omnibus):	0.000	Jarque-Bera (JB):	10110.567
Skew:	0.672	Prob(JB):	0.00
Kurtosis:	6.324	Cond. No.	62.8

From the summary we can see that the R square value is 0.932

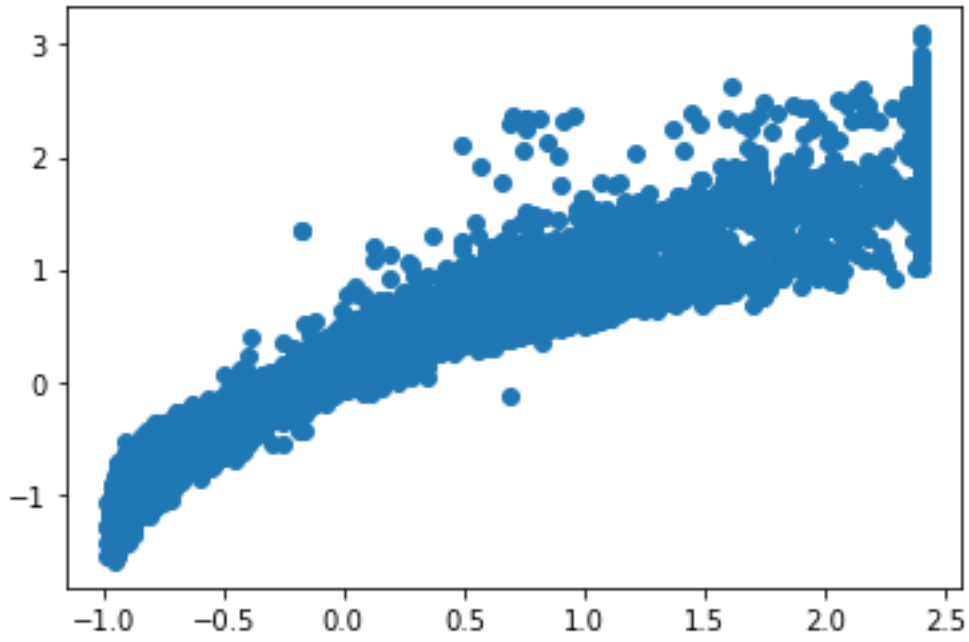
After calculating MSE for test data - 0.2648695629305823

After calculating MSE for train data - 0.26132851657993456

RMSE score for train and test data are same.

Scatter plot of the dependent variable (price)

BUSINESS REPORT



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

```
(0.0) * Intercept + (1.18) * carat + (0.03) * cut + (0.13) * color + (0
.2) * clarity + (-0.01) * depth + (-0.01) * table + (-0.49) * x + (0.38
) * y + (-0.02) * z +
```

From above analysis we can say –

There are also some negative co-efficient values, for instance, table has its corresponding co-efficient as -0.49. This implies, the Average Diameter makes the price decreases by 49 units.

The five important parameters which will help the company to decide the profitable price for cubic zirconia diamond

They are -

1.Average Diameter plays an important role in increasing and decreasing the price

2.Y which is the width of the cubic zirconia diamond

3.Carat which is the weight of the cubic zirconia diamond (has Highest Weightage)

4.X which is the length of the cubic zirconia diamond

5.Clarity which is the absence of the Inclusions and Blemishes.

6.Color of the cubic zirconia diamond (has Lowest Weightage).

BUSINESS REPORT

Problem 2: Logistic Regression and LDA

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.

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.

- Importing dataset Holiday_Package.csv into df2 –

Head of 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

From the table above we can state that column Unnamed:0 can be dropped.

Shape of data –

(872, 8)

Dataset consists of 872 rows and 8 columns.

BUSINESS REPORT

Checking info of the dataset-

```
<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    int8
 1   Salary             872 non-null    float64
 2   age                872 non-null    int64
 3   educ               872 non-null    float64
 4   no_young_children  872 non-null    float64
 5   no_older_children  872 non-null    float64
 6   foreign            872 non-null    int8
dtypes: float64(4), int64(1), int8(2)
memory usage: 35.9 KB
```

We can say:

1. Data has 2 object type/categorical variables which are Holliday_Package, foreign

2. There are 5 int type variables namely Salary, age, educ, no_young_children, no_older_children.

Checking 5-point summary of data using describe –

	count	mean	std	min	25%	50%	75%	max
Holliday_Package	872.0	0.459862	0.498672	0.00	0.0	0.0	1.0	1.00
Salary	872.0	45608.336869	15699.745151	8105.75	35324.0	41903.5	53469.5	80687.75
age	872.0	39.955275	10.551675	20.00	32.0	39.0	48.0	62.00
educ	872.0	9.302752	3.014712	2.00	8.0	9.0	12.0	18.00
no_young_children	872.0	0.000000	0.000000	0.00	0.0	0.0	0.0	0.00
no_older_children	872.0	0.980505	1.077197	0.00	0.0	1.0	2.0	5.00
foreign	872.0	0.247706	0.431928	0.00	0.0	0.0	0.0	1.00

We can say -

1. There are total 872 rows in the dataset.

2. Salary variable having the count value 872 and the value of

3.mean, standard deviation, minimum, 25%, 50%, 75% and max are 47729.2, 23418.7, 1322, 35324, 41903.5, 53469.5, 236961 respectively.

4. age variable having the count value 872 and the value of mean, standard deviation, minimum, 25%, 50%, 75% and max are 39.9553, 10.5517, 20, 32, 39, 48, 62 respectively.

5. educ variable having the count value 872 and the value of mean, standard deviation, minimum, 25%, 50%, 75% and max are 9.30734, 3.03626, 1, 8, 9, 12, 21 respectively.

BUSINESS REPORT

6. no_young_children variable having the count value 872 and the value of

7.mean, standard deviation, minimum, 25%, 50%, 75% and max are 0.311927, 0.61287, 0, 0, 0, 0, 3 respectively.

8. no_older_children variable having the count value 872 and the value of

9.mean, standard deviation, minimum, 25%, 50%, 75% and max are 0.982798, 1.08679, 0, 0, 1, 2, 6 respectively.

10. foreign being the categorical value have 2 unique values and the top one is no with a frequency of 656.

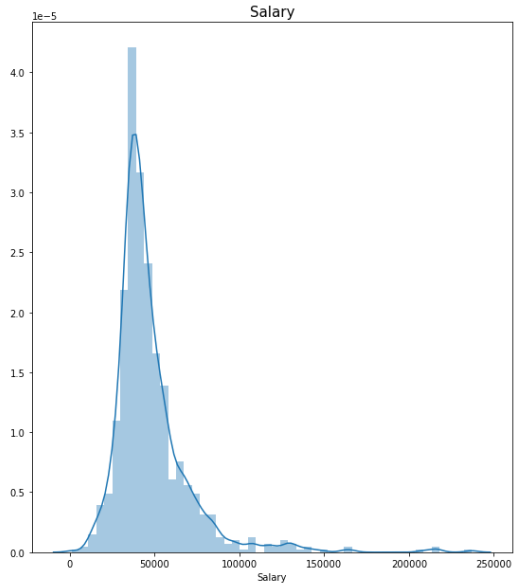
Checking for null values –

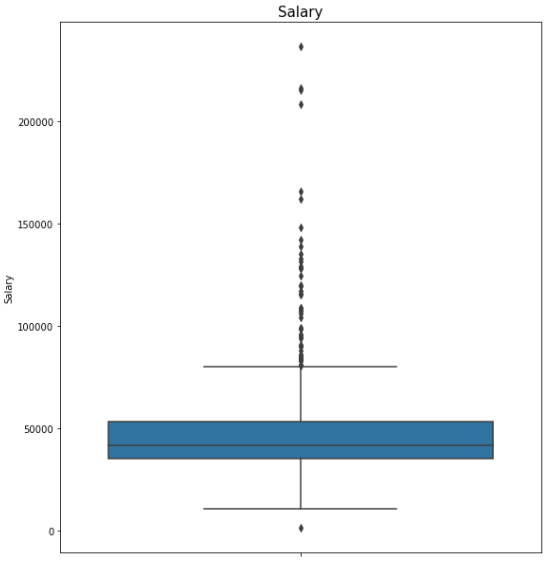
```
Holliday_Package     0
Salary               0
age                  0
educ                 0
no_young_children    0
no_older_children    0
foreign              0
dtype: int64
```

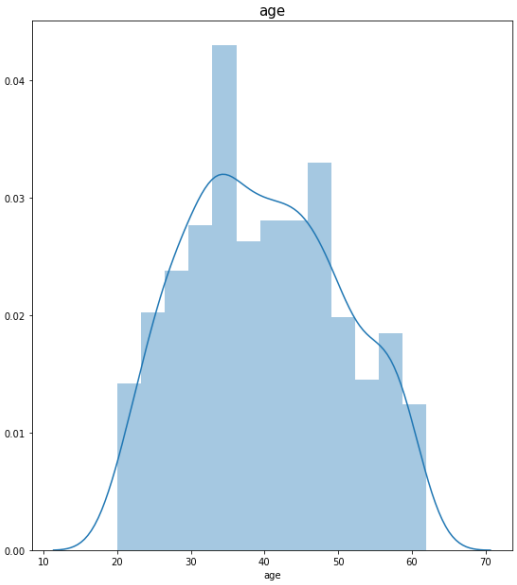
We can say there are no null values in the dataset.

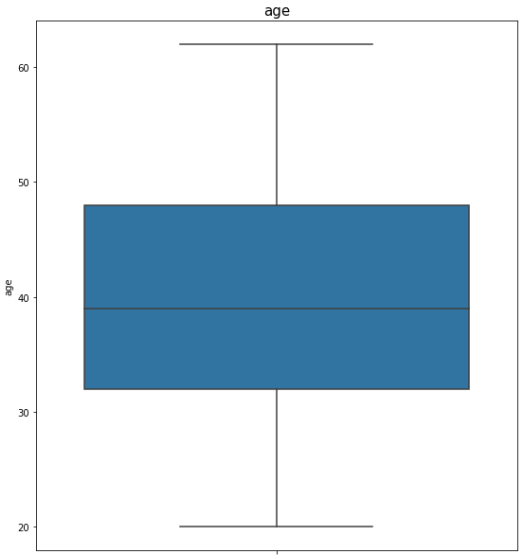
Univariate Analysis

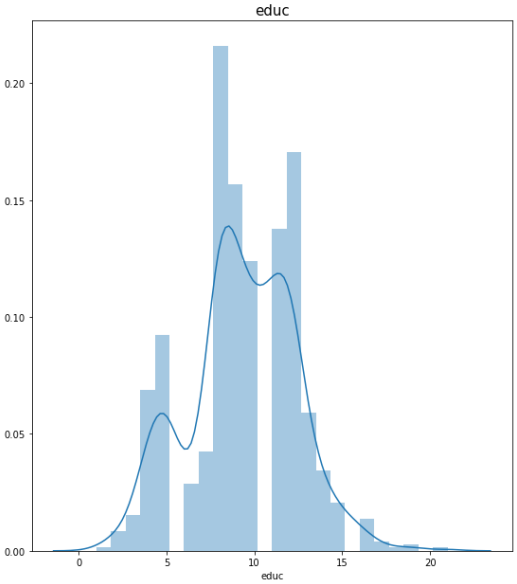
BUSINESS REPORT

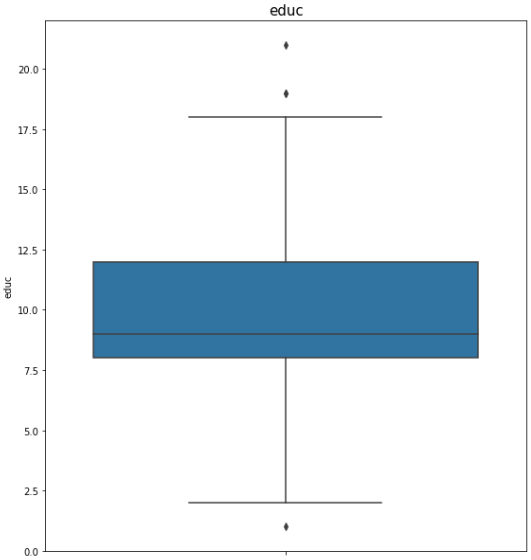












BUSINESS REPORT

We can say -

1. Distribution of column variable salary is right skewed and has outliers

2. Distribution of column variable age looks normally distributed but boxplot shows there are no outliers.

3. Distribution of column variable educ is right skewed and also has outliers.

Checking outliers column wise –

```
Holliday_Package       0
Salary                57
Unnamed: 0             0
age                    0
educ                   4
foreign                0
no_older_children      2
no_young_children    207
dtype: int64
```

We can say -

1. Salary has 57 outliers.

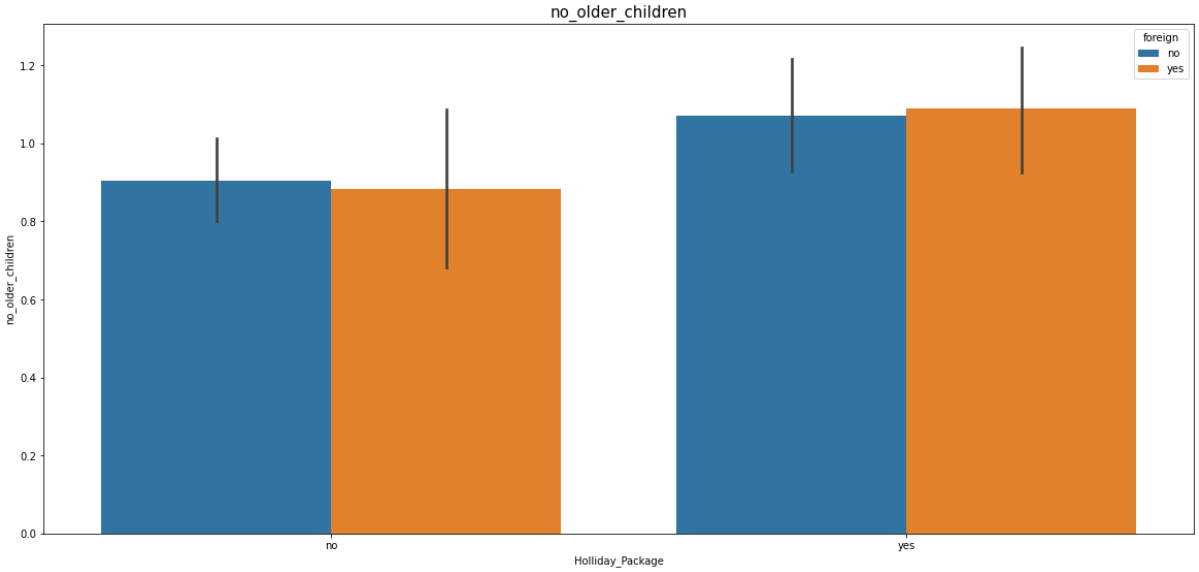
2. educ has 4 outliers.

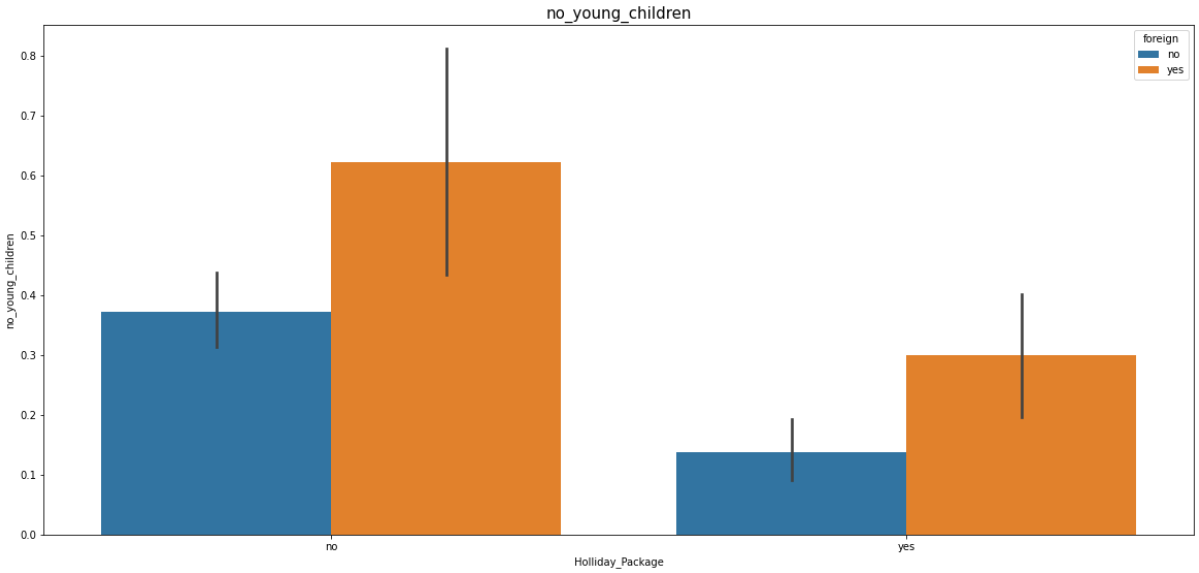
3. no_older_children have 2 outliers.

4. no_young_children have 207 outliers.

BUSINESS REPORT

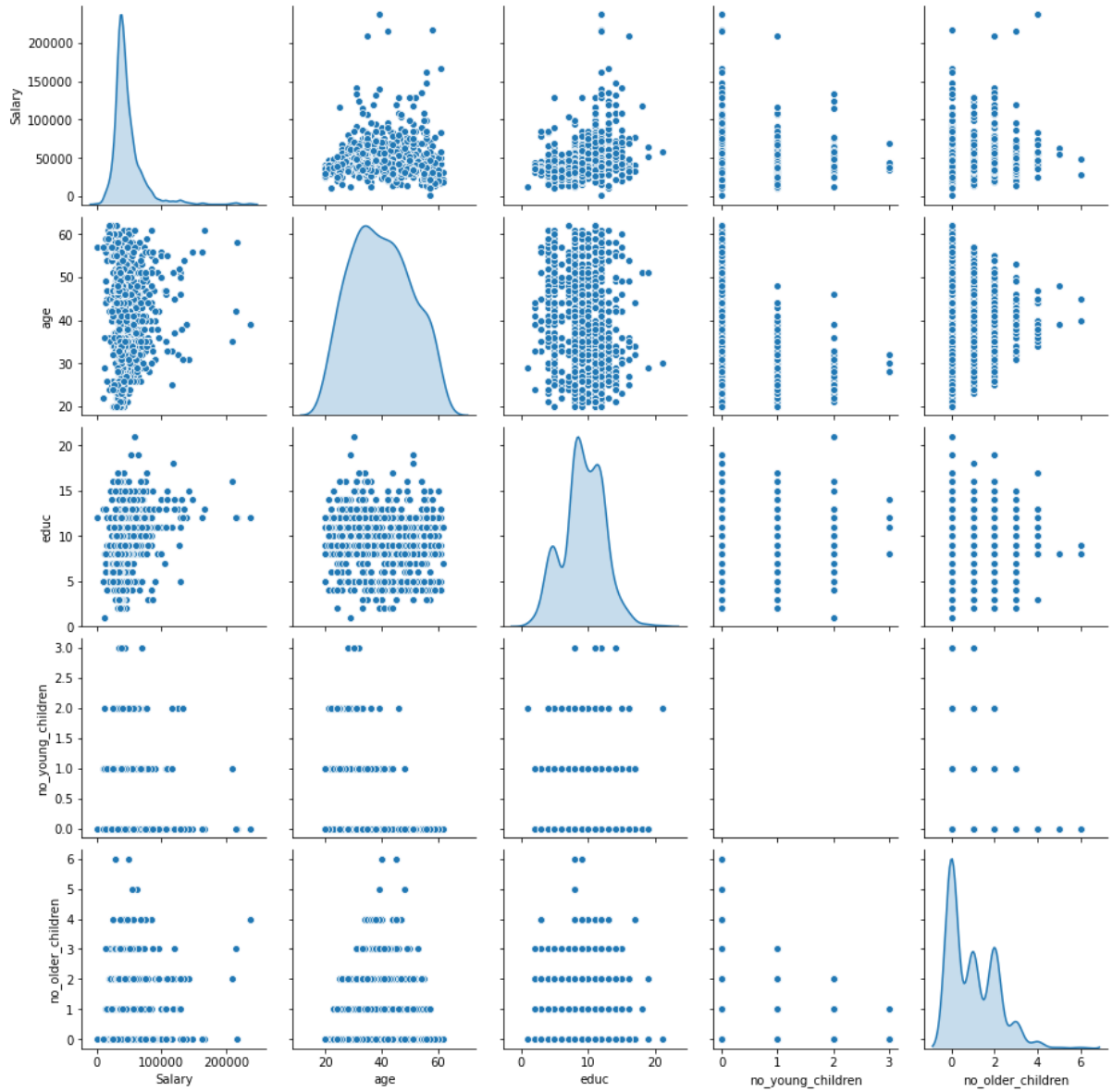
Bivariate Analysis





BUSINESS REPORT

Multivariate Analysis-



BUSINESS REPORT

Heatmap –

BUSINESS REPORT

Outliers Treatment

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

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).

Converting data types to integer datatypes

```
<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    int8
 1   Salary             872 non-null    float64
 2   age                872 non-null    int64
 3   educ               872 non-null    float64
 4   no_young_children  872 non-null    float64
 5   no_older_children  872 non-null    float64
 6   foreign            872 non-null    int8
dtypes: float64(4), int64(1), int8(2)
memory usage: 35.9 KB
```

Separating independent and target variables –

X.head()

	Salary	age	educ	no_young_children	no_older_children	foreign
0	48412.0	30	8.0	0.0	1.0	0
1	37207.0	45	8.0	0.0	1.0	0
2	58022.0	46	9.0	0.0	0.0	0
3	66503.0	31	11.0	0.0	0.0	0
4	66734.0	44	12.0	0.0	2.0	0

BUSINESS REPORT

y.head()

```
0    0
1    1
2    0
3    0
4    0
Name: Holliday_Package, dtype: int8
```

Splitting data into train and test and building Logistic Regression and LDA.

df3.corr()

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign	Prediction
Holliday_Package	1.000000	-0.180214	-0.092311	-0.101116	NaN	0.078691	0.254096	0.259676
Salary	-0.180214	1.000000	0.047029	0.352913	NaN	0.124030	-0.239387	-0.364258
age	-0.092311	0.047029	1.000000	-0.149682	NaN	-0.117754	-0.107148	-0.164919
educ	-0.101116	0.352913	-0.149682	1.000000	NaN	-0.035656	-0.420922	-0.351553
no_young_children	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
no_older_children	0.078691	0.124030	-0.117754	-0.035656	NaN	1.000000	0.022729	0.174012
foreign	0.254096	-0.239387	-0.107148	-0.420922	NaN	0.022729	1.000000	0.858111
Prediction	0.259676	-0.364258	-0.164919	-0.351553	NaN	0.174012	0.858111	1.000000

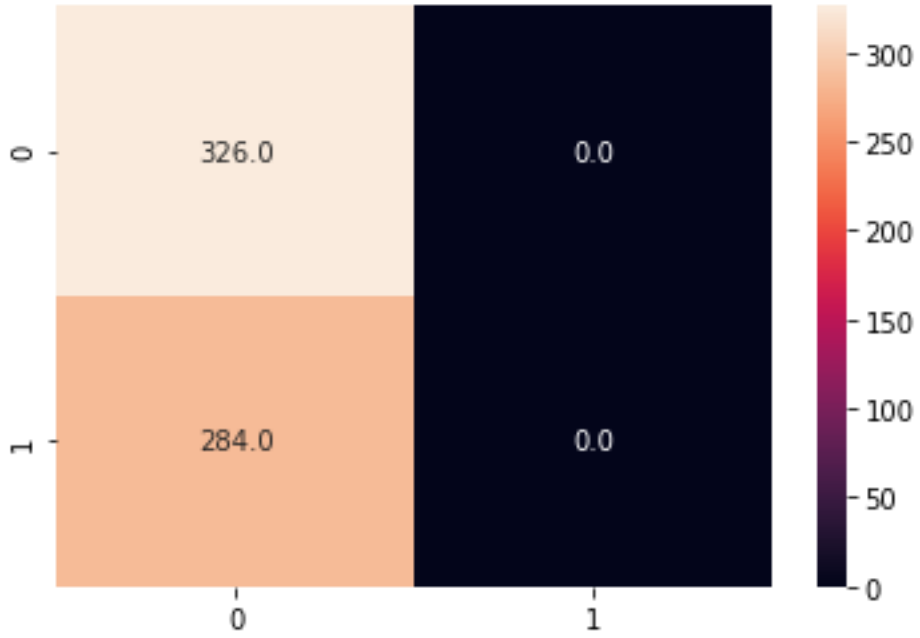
2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.

Performance Metrics for Logistic Regression Model

For Train Data –

BUSINESS REPORT

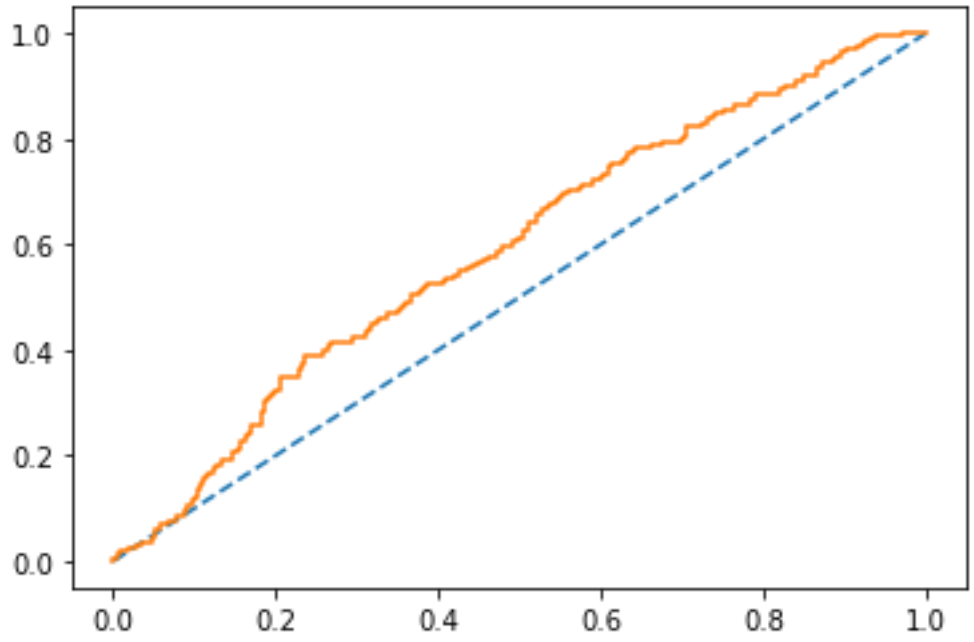
Confusion matrix –



Classification Report-

	precision	recall	f1-score	support
0	0.53	1.00	0.70	326
1	0.00	0.00	0.00	284
accuracy			0.53	610
macro avg	0.27	0.50	0.35	610
weighted avg	0.29	0.53	0.37	610

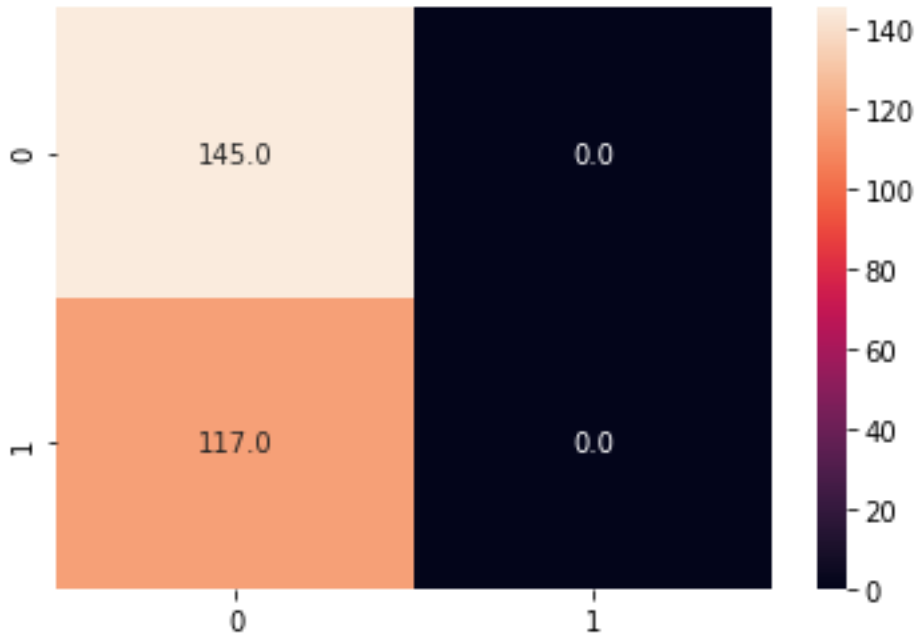
BUSINESS REPORT



AUC = 0.591

For Test Data

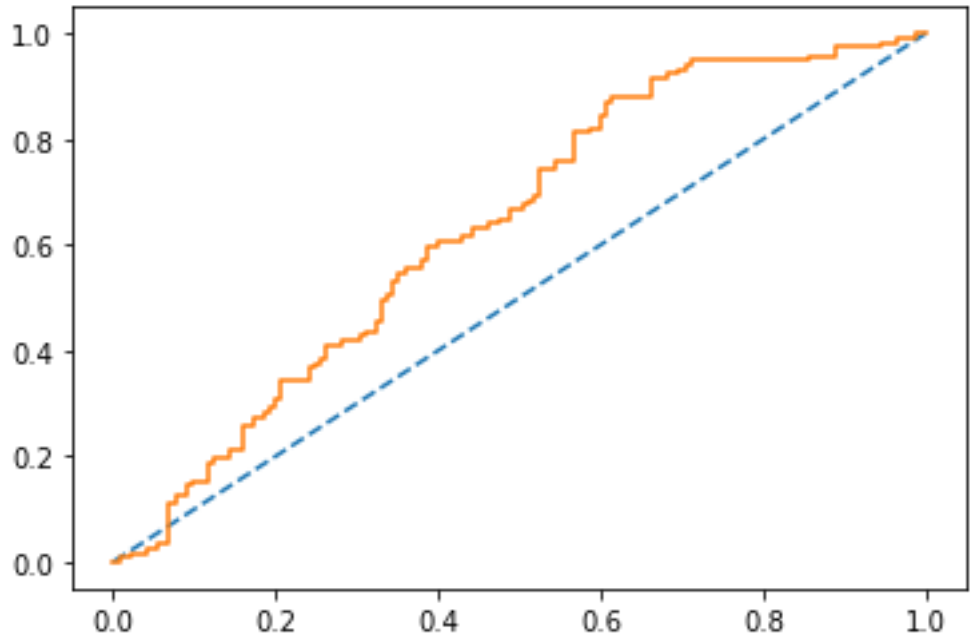
Confusion matrix –



BUSINESS REPORT

Classification Report-

	precision	recall	f1-score	support
0	0.55	1.00	0.71	145
1	0.00	0.00	0.00	117
accuracy			0.55	262
macro avg	0.28	0.50	0.36	262
weighted avg	0.31	0.55	0.39	262



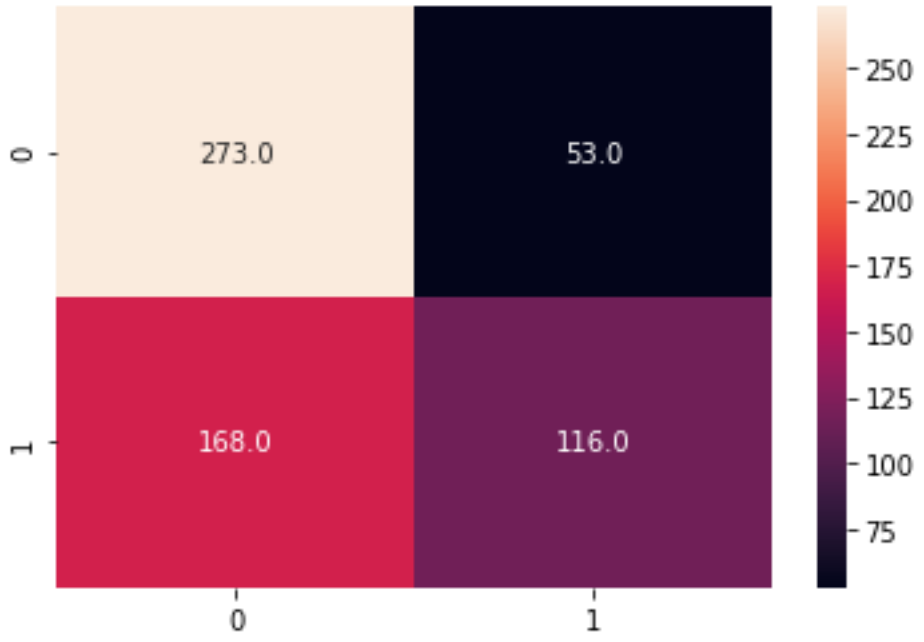
AUC - 0.633

Performance Metrics for LDA Model -

For Train Data

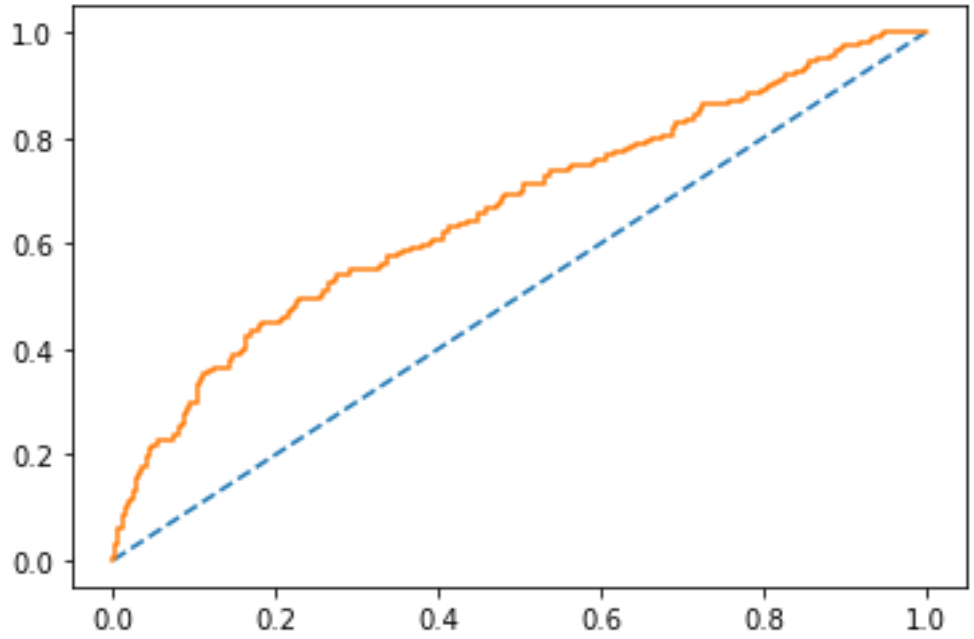
Confusion Matrix –

BUSINESS REPORT



Classification Report -

	precision	recall	f1-score	support
0	0.62	0.84	0.71	326
1	0.69	0.41	0.51	284
accuracy			0.64	610
macro avg	0.65	0.62	0.61	610
weighted avg	0.65	0.64	0.62	610

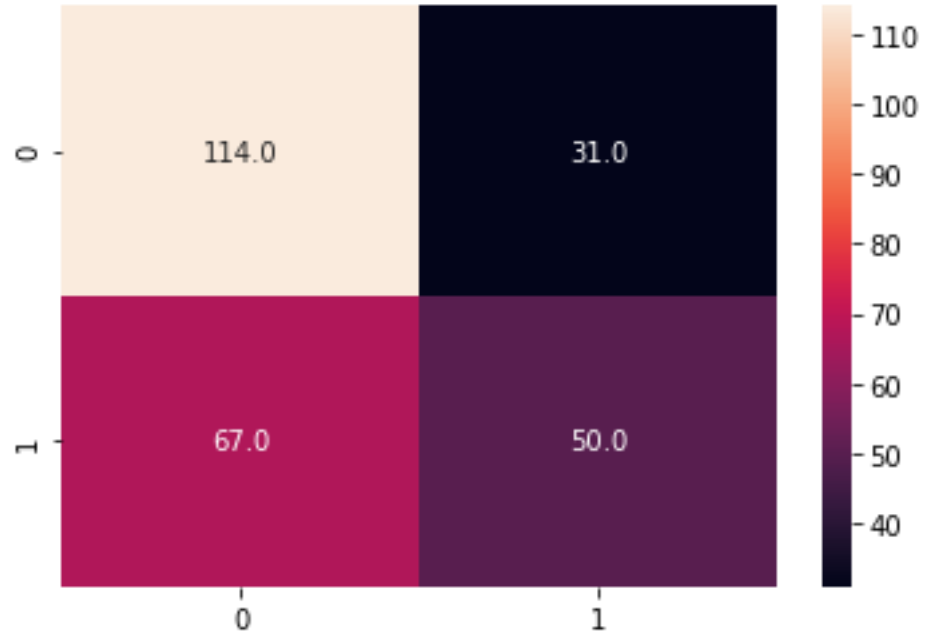


AUC = 0.662

BUSINESS REPORT

For Test Data-

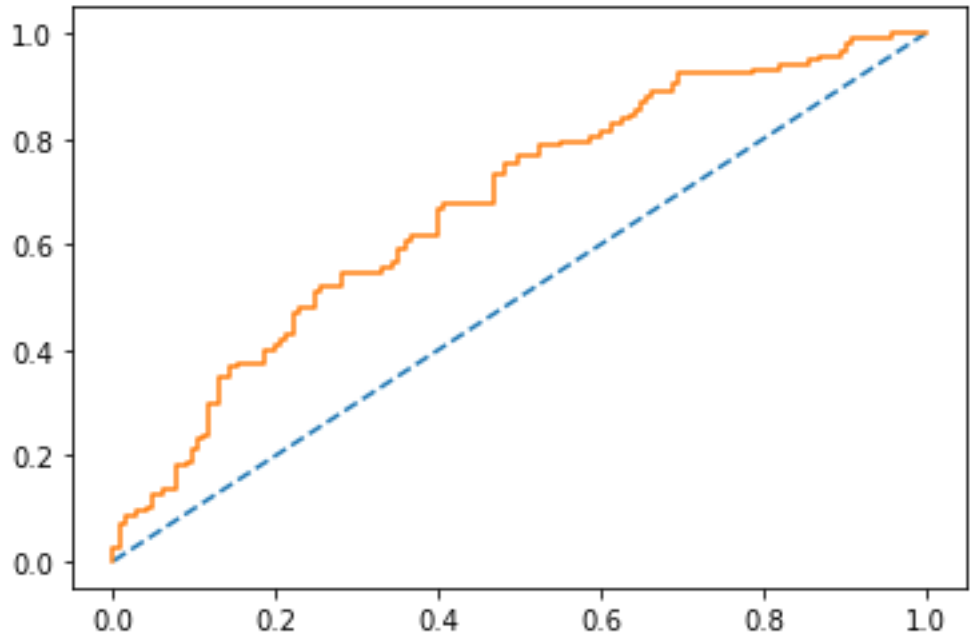
Confusion Matrix-



Classification Report -

	precision	recall	f1-score	support
0	0.63	0.79	0.70	145
1	0.62	0.43	0.51	117
accuracy			0.63	262
macro avg	0.62	0.61	0.60	262
weighted avg	0.62	0.63	0.61	262

BUSINESS REPORT



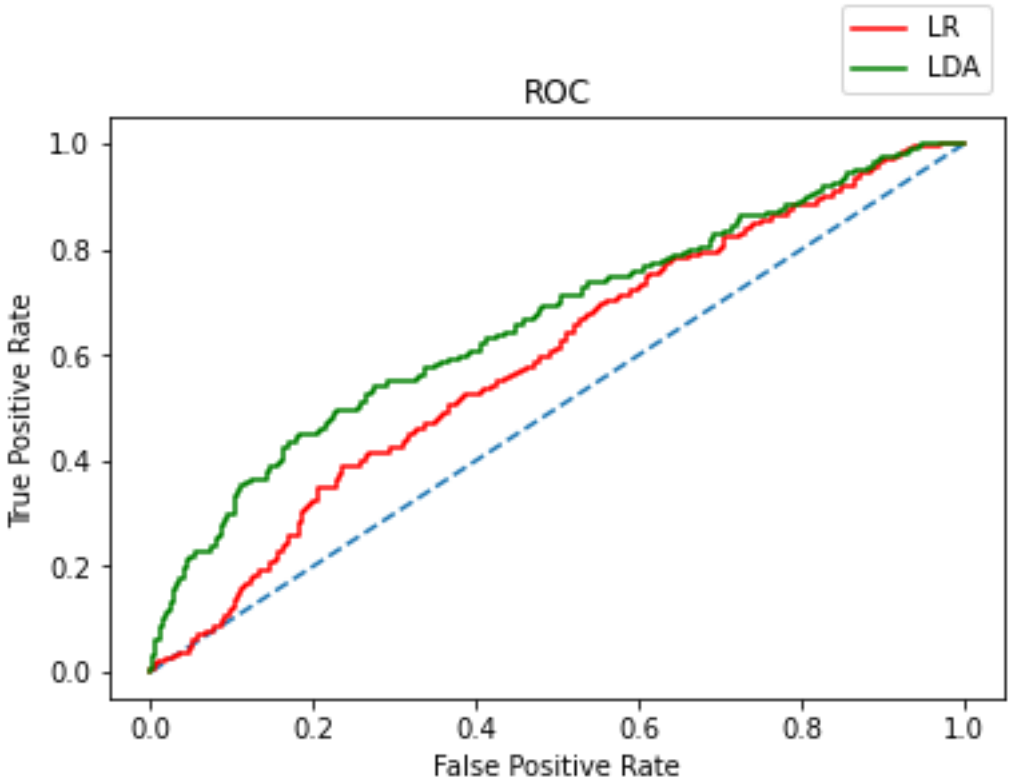
AUC = 0.674

Comapring both LR Model and LDA Model

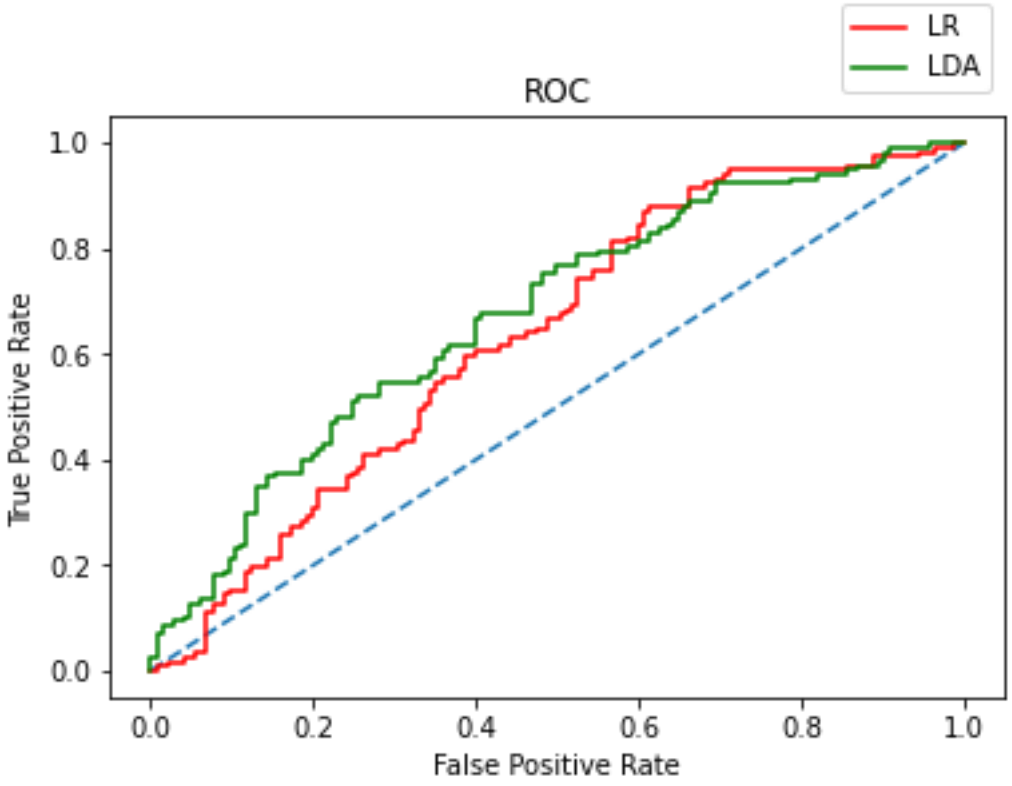
	LR Train	LR Test	LDA Train	LDA Test
Accuracy	0.53	0.55	0.64	0.63
AUC	0.59	0.63	0.66	0.67
Recall	0.00	0.00	0.51	0.51
Precision	0.00	0.00	0.41	0.43
F1 Score	0.00	0.00	0.69	0.62

BUSINESS REPORT

For Train Data



For Test Data



BUSINESS REPORT

We can say -

1. Accuracy for train data for Logistic Regression Model is 53%, for Linear Discriminant Analysis Model is 64%.

2. Accuracy for test data for Logistic Regression Model is 55%, for Linear Discriminant Analysis Model is 63%.

3. AUC Score for train data for Logistic Regression Model is 59%, for Linear Discriminant Analysis Model is 66%.

4. AUC Score for test data for Logistic Regression Model is 63%, for Linear Discriminant Analysis Model is 67%.

5. Recall for train data for Logistic Regression Model is 0%, for Linear Discriminant Analysis Model is 51%.

6. Recall for test data for Logistic Regression Model is 0%, for Linear Discriminant Analysis Model is 51%.

7. Precision for train data for Logistic Regression Model is 0%, for Linear Discriminant Analysis Model is 41%.

8. Precision for test data for Logistic Regression Model is 0%, for Linear Discriminant Analysis Model is 43%.

9. F1 Score for train data for Logistic Regression Model is 0%, for Linear Discriminant Analysis Model is 69%.

10. F1 Score for test data for Logistic Regression Model is 0%, for Linear Discriminant Analysis Model is 62%.

As we can see Performance Metrics of Linear Discriminant Analysis Model is far better than Logistic Regression Model and can be used for making any future predictions.

BUSINESS REPORT

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

```
The coefficient for Salary is -4.135560275930995e-06
The coefficient for age is -0.0020325651700645177
The coefficient for educ is 0.014704572939515846
The coefficient for no_young_children is 2.220446049250313e-16
The coefficient for no_older_children is 0.042320617662464505
The coefficient for foreign is 0.31074104185168466
```

Insights and recommendations –

1. As we can see the important parameters on the basis of which the company will focus on particular employees to sell their packages are foreign which has highest weightage and no_older_children which has weightage of 0.042320617662464505

2. We can state from analysis:
   The employees having a smaller number of young children are likely to opt for holiday package

3. The employees having less who are foreigner are likely to opt for holiday package.

4. From the business point of view , we can target the employees who have less young children and are foreigner.