PREDICTIVE MODELING PROJECT REPORT

By Karthik Sreeram R

DECEMBER 5, 2020

UNIVERSITY OF TEXAS AT AUSTIN AND GREAT LAKES

Purpose
This document is the business report for my final project in the subject “Predictive Modeling “ This document gives us a detailed explanation of various approaches used, their insight and inferences. Tools used analysis: Python and Jupiter notebook. Packages used: NumPy, pandas, seaborn, os, matplotlib, SciPy, stats model, sklearn and sweetviz

Contents

Problem 1: Linear Regression

Business scenario

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.

Data Dictionary:

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, I3= level 3 inclusions) FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, I2, I3
Depth	The Height of a 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.

1.1) Read the data and do exploratory data analysis. Describe the data briefly. Perform Univariate and Bivariate Analysis.

a.) Dataset Head

]:

	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

Inference :

Dataset has 11 columns .
The first column (Unnamed column :0) is of no use for analysis and can be removed .

b.) Type of the variables in dataset

```
 #   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
```

Inference:

Depth has null value. Other columns has no null value .
Cut, color and clarity are object data type ie contains strings value. Remining columns are of numerical datatype (either integer or float).
Dataset has 26967 observations (Rows of data)

c.) Summary of the dataset:

	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

Inference:

All variables (columns) have different scale. For most variables there is a huge difference between the 75thpercentile and maximum value. So, there is chance for outliers.

Remove unwanted column:

Removed the unwanted unnamed column. It has no value.

d) Data has any duplicates?

```
Number of Duplicates  0
```

Inference:

Dataset has no duplicated data.

e) Numerical Data has any observation with value zero?

```
carat      False
cut        False
color      False
clarity    False
depth      False
table      False
x          False
y          False
z          False
price      False
dtype: bool
```

. Inference:

Dataset has no numerical observation with zero value data.

f) Count of null values in the data set.

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

Inference:

Only depth column has 697 null values. Remining columns has no null values

g) Value count for non-numerical columns

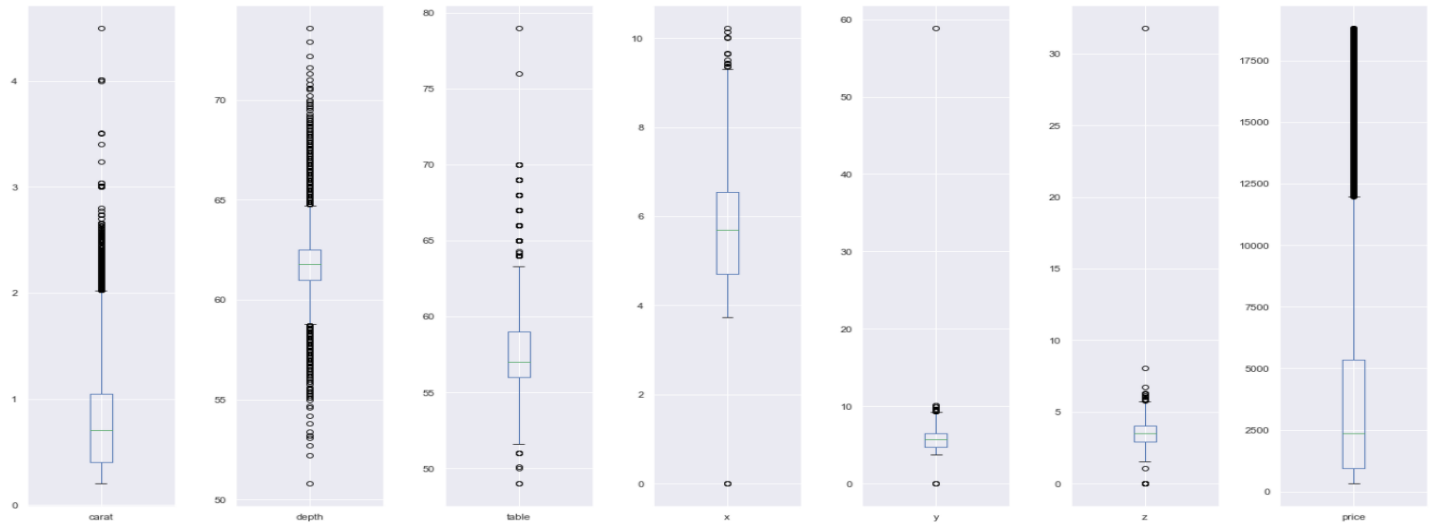
```
CUT :  5
Fair            781
Good           2441
Very Good      6030
Premium        6899
Ideal         10816
Name: cut, dtype: int64


COLOR :  7
J     1443
I     2771
D     3344
H     4102
F     4729
E     4917
G     5661
Name: color, dtype: int64


CLARITY :  8
I1         365
IF         894
VVS1      1839
VVS2      2531
VS1       4093
SI2       4575
VS2       6099
SI1       6571
Name: clarity, dtype: int64
```

Inference:

There are three non-numerical columns. All the three columns can be ordered.

h) Outlier Analysis:

Boxplot (Before removing outlier):




```
the number of outliers are  892
```

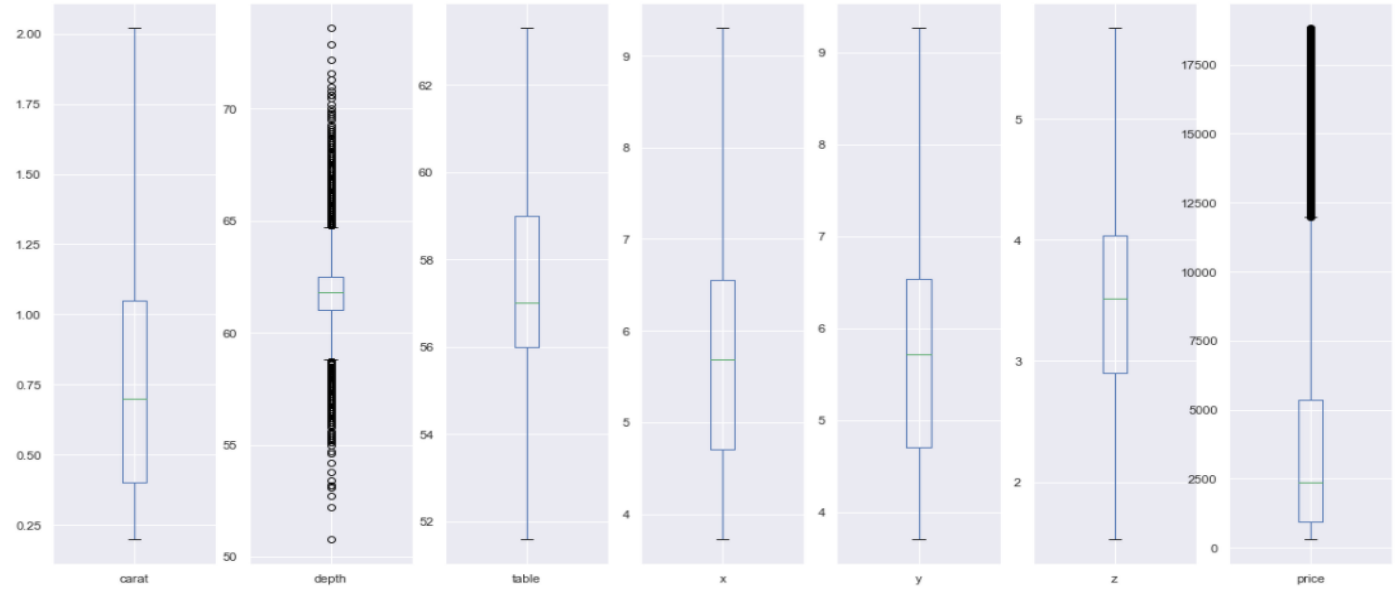
(Except target price column)

Inference:

Dataset has some outliers in all columns. Will be treating outliers for all column except target Price column.

After Handling Outliers:

Box plot



Summary of data set after handling outliers:

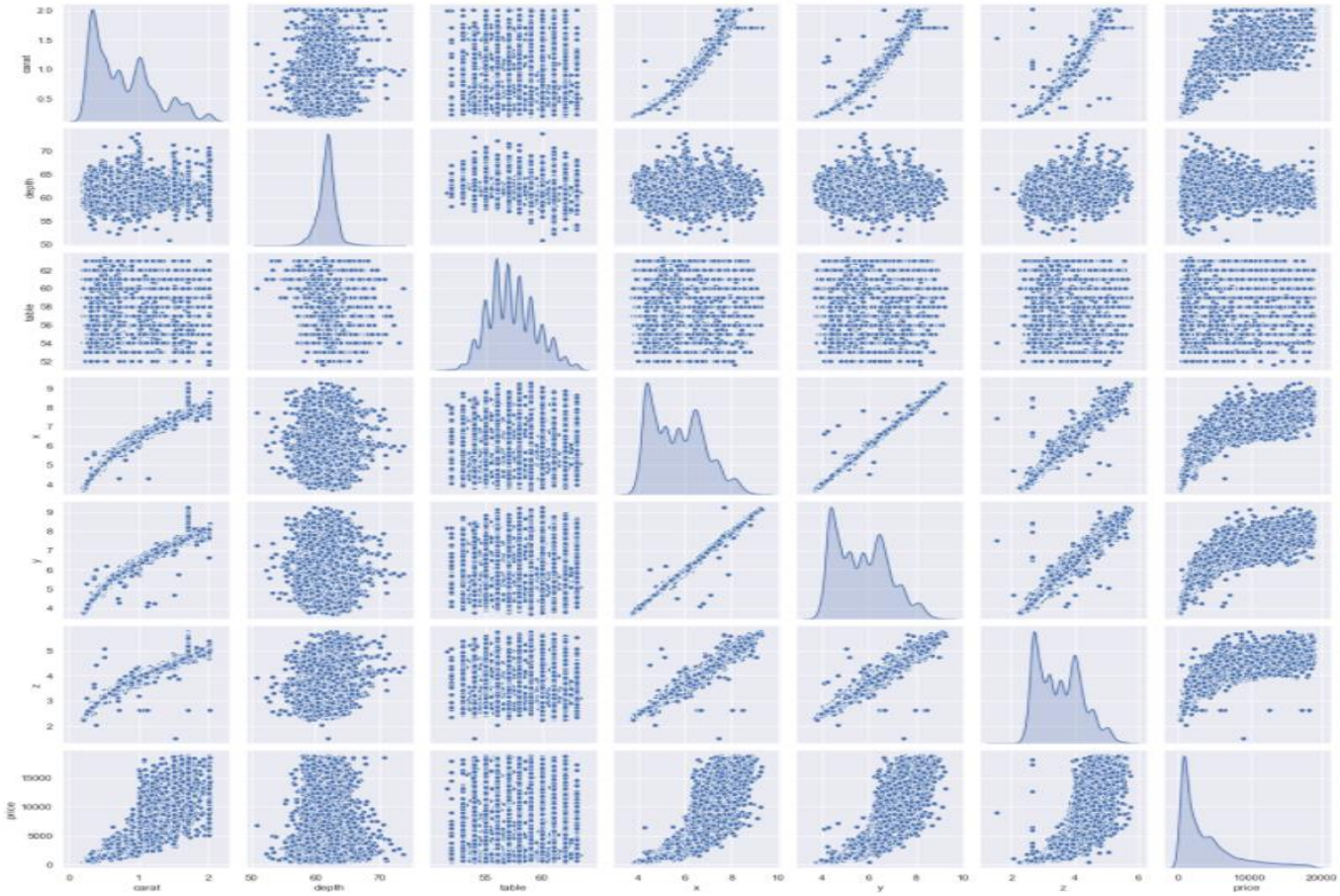
	carat	depth	table	x	y	z	price
count	26967.000000	26270.000000	26967.000000	26967.000000	26967.000000	26967.000000	26967.000000
mean	0.785860	61.745147	57.407702	5.729438	5.731334	3.537316	3939.518115
std	0.444042	1.412860	2.090151	1.124638	1.116593	0.694826	4024.864666
min	0.200000	50.800000	51.600000	3.730000	3.710000	1.530000	326.000000
25%	0.400000	61.000000	56.000000	4.710000	4.710000	2.900000	945.000000
50%	0.700000	61.800000	57.000000	5.690000	5.710000	3.520000	2375.000000
75%	1.050000	62.500000	59.000000	6.550000	6.540000	4.040000	5360.000000
max	2.020000	73.600000	63.300000	9.300000	9.260000	5.750000	18818.000000

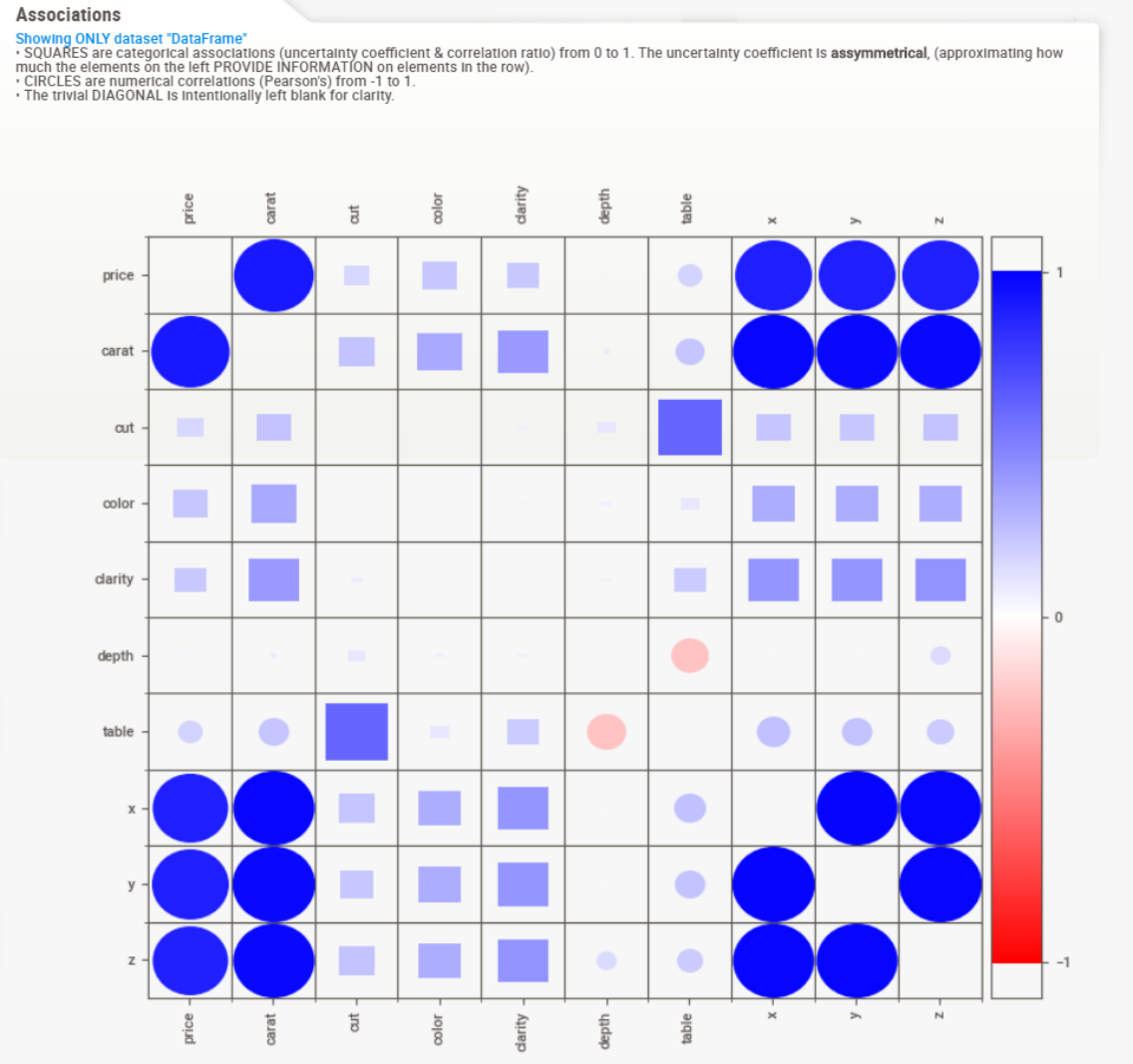
Inference:

The outliers in independent columns are handled and replaced. No changes are made for target price column.

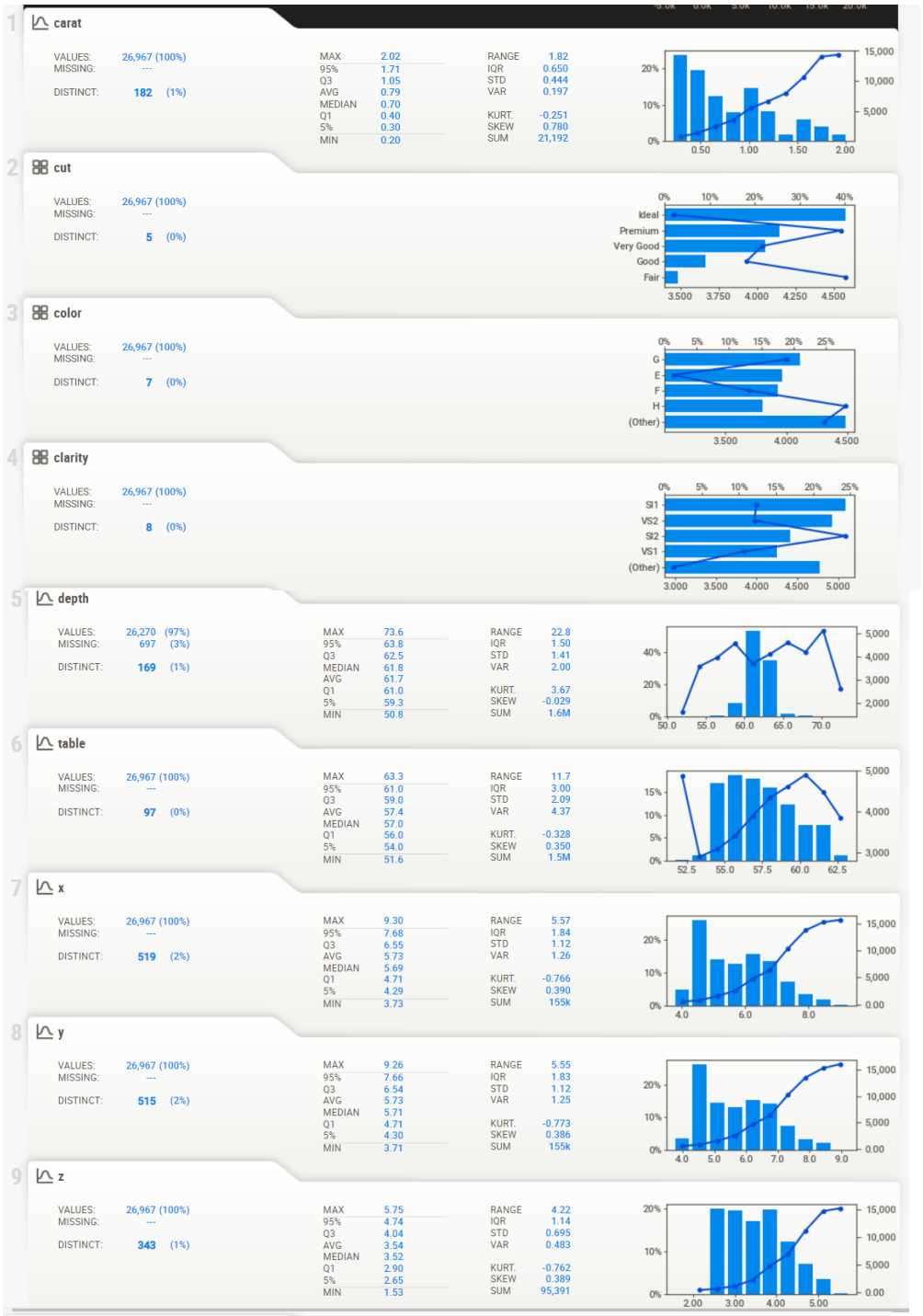
i)Graphical Analysis:

A.) Multivariate Analysis


carat
depth
table
x
y
z
price



B.) Uni-variate and Bi-variate Analysis


1 carat
VALUES: 26,967 (100%)
MISSING: ---
DISTINCT: 182 (1%)
MAX 2.02
95% 1.71
Q3 1.05
AVG 0.79
MEDIAN 0.70
Q1 0.40
5% 0.30
MIN 0.20
RANGE 1.82
IQR 0.650
STD 0.444
VAR 0.197
KURT. -0.251
SKEW 0.780
SUM 21,192
2 cut
VALUES: 26,967 (100%)
MISSING: ---
DISTINCT: 5 (0%)
Ideal
Premium
Very Good
Good
Fair
3 color
VALUES: 26,967 (100%)
MISSING: ---
DISTINCT: 7 (0%)
G
E
F
H
(Other)
4 clarity
VALUES: 26,967 (100%)
MISSING: ---
DISTINCT: 8 (0%)
SI1
VS2
SI2
VS1
(Other)
5 depth
VALUES: 26,270 (97%)
MISSING: 697 (3%)
DISTINCT: 169 (1%)
MAX 73.6
95% 63.8
Q3 62.5
MEDIAN 61.8
AVG 61.7
Q1 61.0
5% 59.3
MIN 50.8
RANGE 22.8
IQR 1.50
STD 1.41
VAR 2.00
KURT. 3.67
SKEW -0.029
SUM 1.6M
6 table
VALUES: 26,967 (100%)
MISSING: ---
DISTINCT: 97 (0%)
MAX 63.3
95% 61.0
Q3 59.0
AVG 57.4
MEDIAN 57.0
Q1 56.0
5% 54.0
MIN 51.6
RANGE 11.7
IQR 3.00
STD 2.09
VAR 4.37
KURT. -0.328
SKEW 0.350
SUM 1.5M
7 x
VALUES: 26,967 (100%)
MISSING: ---
DISTINCT: 519 (2%)
MAX 9.30
95% 7.68
Q3 6.55
AVG 5.73
MEDIAN 5.69
Q1 4.71
5% 4.29
MIN 3.73
RANGE 5.57
IQR 1.84
STD 1.12
VAR 1.26
KURT. -0.766
SKEW 0.390
SUM 155k
8 y
VALUES: 26,967 (100%)
MISSING: ---
DISTINCT: 515 (2%)
MAX 9.26
95% 7.66
Q3 6.54
AVG 5.73
MEDIAN 5.71
Q1 4.71
5% 4.30
MIN 3.71
RANGE 5.55
IQR 1.83
STD 1.12
VAR 1.25
KURT. -0.773
SKEW 0.386
SUM 155k
9 z
VALUES: 26,967 (100%)
MISSING: ---
DISTINCT: 343 (1%)
MAX 5.75
95% 4.74
Q3 4.04
AVG 3.54
MEDIAN 3.52
Q1 2.90
5% 2.65
MIN 1.53
RANGE 4.22
IQR 1.14
STD 0.695
VAR 0.483
KURT. -0.762
SKEW 0.389
SUM 95,391
price
MISSING: ---
Auto
5
15
30
NUMERICAL ASSOCIATIONS
(PEARSON, -1 to 1)
carat 0.91
y 0.89
x 0.89
z 0.88
table 0.13
depth -0.00
price 0.00
CATEGORICAL ASSOCIATIONS
(CORRELATION RATIO, 0 to 1)
color 0.18
clarity 0.17
cut 0.11

Insights:

There is a high numerical correlation between target column price and carat, x, y, z.
There is a high categorical correlation between price and color / clarity
According to the visualization, there is no much correlation between price and depth
There is a small correlation between price vs table/cut
Only the price (target variable) and depth have uniform distribution. Other variable has random distribution with multiple peaks. This may be due to multiple groups in the other variables.
Price (Target variable) Is right Skewed.
Business Insights:
These information suggest that as price increases the sales (quantity) reduces.
The carat and dimensions (x(length), y (width) and z (height)) plays the huge impact for the pricing than other variables.

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?

a.)Does the data seta as any null values?

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

Inference:

Depth column has null values.

b.) Does the data seta has zero as values?

```
carat      False
cut        False
color      False
clarity    False
depth      False
table      False
x          False
y          False
z          False
price      False
dtype: bool
```

Inference:

No. There is no zero value in the dataset.

c.) Do they null values in the dataset has any meaning or do we need to change them or drop them?

As the column depth is least corelated (in fact no correlation) with the target variable, imputing the null values of the depth column will not change the results of prediction much. But if I drop them, then it may affect the other column which may affect the prediction. even though the count of na value is around 2 percentage of the data set I will impute it with most frequent value.

Any null value after imputing the data?

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

Inference:

Null value data is imputed with frequent value.

d) Do the data require scaling ?

	carat	depth	table	x	y	z	price
count	26967.000000	26270.000000	26967.000000	26967.000000	26967.000000	26967.000000	26967.000000
mean	0.785860	61.745147	57.407702	5.729438	5.731334	3.537316	3939.518115
std	0.444042	1.412860	2.090151	1.124638	1.116593	0.694826	4024.864666
min	0.200000	50.800000	51.600000	3.730000	3.710000	1.530000	326.000000
25%	0.400000	61.000000	56.000000	4.710000	4.710000	2.900000	945.000000
50%	0.700000	61.800000	57.000000	5.690000	5.710000	3.520000	2375.000000
75%	1.050000	62.500000	59.000000	6.550000	6.540000	4.040000	5360.000000
max	2.020000	73.600000	63.300000	9.300000	9.260000	5.750000	18818.000000

Inference:

Yes, scaling is recommended. the data columns have different scales .so we need to scale the data set. It will also help us to center the variables and make predators have the value mean 0 (at least near zero). So it will help us to interpret the intercept term as the expected value of price (target value) when the predictor values are set to their means.

Insights:

Data had null values and it is replaced by frequent value.
Data has no zero value.
Scaling is required.

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.

a) Encoding the string values

As the category columns are ordered in nature, I have the encoded using ordinal encoder.

For the column cut:

Fair:1, Good:2, Very Good:3, Premium:4, Ideal:5

For the column Color (D being the best and J being worst) :

D:7, E:6, F:5, G:4, H:3, I:2, J :1

For the column Clarity (In order from Best to Worst, FL = flawless, I3= level 3 inclusions) :

FL:11, IF:10, VVS1:9, VVS2:8, VS1:7, VS2:6, SI1:5, SI2:4, I1:3, I2:2, I3:1

Head of the data after encoding.

]:

	carat	cut	color	clarity	depth	table	x	y	z	price
0	0.3	5	6	5	62.1	58	4.27	4.29	2.66	499
1	0.33	4	4	10	60.8	58	4.42	4.46	2.7	984
2	0.9	3	6	8	62.2	60	6.04	6.12	3.78	6289
3	0.42	5	5	7	61.6	56	4.82	4.8	2.96	1082
4	0.31	5	5	9	60.4	59	4.35	4.43	2.65	779

Data information:

Have changed the encoded data to numerical data type .

```
<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  int32  
 2   color    26967 non-null  int32  
 3   clarity  26967 non-null  int32  
 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), int32(3)
memory usage: 1.7 MB
```

Data summary after encoding:

[81]:

	count	mean	std	min	25%	50%	75%	max
carat	26967.0	0.785860	0.444042	0.20	0.40	0.70	1.05	2.02
cut	26967.0	3.909556	1.113229	1.00	3.00	4.00	5.00	5.00
color	26967.0	4.393889	1.705992	1.00	3.00	4.00	6.00	7.00
clarity	26967.0	6.053102	1.647042	3.00	5.00	6.00	7.00	10.00
depth	26967.0	61.751734	1.395068	50.80	61.10	61.90	62.50	73.60
table	26967.0	57.407702	2.090151	51.60	56.00	57.00	59.00	63.30
x	26967.0	5.729438	1.124638	3.73	4.71	5.69	6.55	9.30
y	26967.0	5.731334	1.116593	3.71	4.71	5.71	6.54	9.26
z	26967.0	3.537316	0.694826	1.53	2.90	3.52	4.04	5.75
price	26967.0	3939.518115	4024.864666	326.00	945.00	2375.00	5360.00	18818.00

Inference :

String columns are encoded to numerical format.(ordinal)

b.) Data Split:

A.) Split target variable and independent variable.

1.) Independent variables (X). head

	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

2.)The Target variable – price (Y). head

	price
0	499.0
1	984.0
2	6289.0
3	1082.0
4	779.0

B.) Splitting the data into train and test variables

(70:30 , used random state =1)

1.) X_train.head

	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

2.) Y_train.head

	price
11687	1061.0
9728	6320.0
1936	536.0
26220	2214.0
18445	2575.0

3.) 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	62.0	56.0	4.39	4.44	2.66
19837	1.20	3	3	7	62.0	57.0	6.77	6.81	4.21

4.) Y_test.head

	price
18031	10671.0
26051	11607.0
16279	1133.0
16466	626.0
19837	6177.0

Inference:

Have split the data into train and split.

After splitting the data, have scaled the data using zscore .

c.) Linear regression model

Successfully built the model for linear regression.

```
LinearRegression()
```

Coefficients:

```
The coefficient for carat is 1.281008834659617
The coefficient for cut is 0.04408688255896204
The coefficient for color is 0.12342511078790361
The coefficient for clarity is 0.19166493838032628
The coefficient for depth is -0.0038375354580184253
The coefficient for table is -0.015396399190677689
The coefficient for x is -0.5358366115152883
The coefficient for y is 0.44065827297784943
The coefficient for z is -0.16422719315423384
```

d.)Performance Metrics: Check the performance of Predictions on Train and Test sets using Rsquare, RMSE

A.)Rsquare :

1.) Train variable:

```
0.888699333687784
```

2.) Test variable:

```
0.8836528787741355
```

B.)RMSE:

1.) Train variable :

```
0.3336175449706086
```

2.) Test Variable :

```
0.341096938165479
```

Inference :

Data is split into train and test after scaling it . Linear Regression is built for the given data set . The rsquare and RMSE are calculated.

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

The coefficient of each variable is :

```
The coefficient for carat is 1.281008834659617
The coefficient for cut is 0.04408688255896204
The coefficient for color is 0.12342511078790361
The coefficient for clarity is 0.19166493838032628
The coefficient for depth is -0.0038375354580184253
The coefficient for table is -0.015396399190677689
The coefficient for x is -0.5358366115152883
The coefficient for y is 0.44065827297784943
The coefficient for z is -0.16422719315423384
```

Business Insights and recommendation:

zirconia **Price = (0.000000000000000336) * intercept +(1.28) *Carat +(0.440) *cut + (0.1234) *color+(0.191) *clarity +(-0.0038) * depth +(-0.0153) *table+(-0.535) *x+(0.440)*y+(-0.164)*z**

The most important factors which determines the price of zirconia are Carat, x (length (negative coefficient)), y (width), clarity and z (height (negative coefficient))

When Carat increases by 1 unit, price of zirconia increases by 1.28 units, keeping all other predictors constant.

when X(Length.) increases by 1 unit, prices decreases by 0.535 units, keeping all other predictors constant.

when y (width) increases by 1 unit, prices increases by 0.440 units, keeping all other predictors constant.

when clarity increases by 1 unit, prices increases by 0.191 units, keeping all other predictors constant.

when Z(Height.) increases by 1 unit, prices decreases by 0.164 units, keeping all other predictors constant.

When Color increases by 1 unit, price of zirconia increases by 0.12 units, keeping all other predictors constant.

Problem 2: Logistic Regression and LDA

Business scenario

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.

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

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.

a.) Dataset Head

]:

	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

Inference:

Dataset has 8 columns.
The first column (Unnamed column :0) is of no use for analysis and can be removed.

b.) Summary of the dataset:

	Unnamed: 0	Salary	age	educ	no_young_children	no_older_children
count	872.000000	872.000000	872.000000	872.000000	872.000000	872.000000
mean	436.500000	47729.172018	39.955275	9.307339	0.311927	0.982798
std	251.869014	23418.668531	10.551675	3.036259	0.612870	1.086786
min	1.000000	1322.000000	20.000000	1.000000	0.000000	0.000000
25%	218.750000	35324.000000	32.000000	8.000000	0.000000	0.000000
50%	436.500000	41903.500000	39.000000	9.000000	0.000000	1.000000
75%	654.250000	53469.500000	48.000000	12.000000	0.000000	2.000000
max	872.000000	236961.000000	62.000000	21.000000	3.000000	6.000000

Inference:

All variables (columns) have different scale. For most variables there is a huge difference between the 75thpercentile and maximum value compared to the 50 percentile and 75 percentiles So, there is chance for outliers.

c.) Type of the variables in dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
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
```

Inference:

Dataset has no null values.
Holiday_package and foreign columns are of object data type i.e. contains strings value. Remining columns are of numerical datatype (either integer).
Dataset has 872 observations (Rows of data)

d.) Dataset has any null values.

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

Inference:

Dataset has no null values.

e.) EDA Data has any duplicities?

```
Number of duplicate rows = 0
```

Inference:

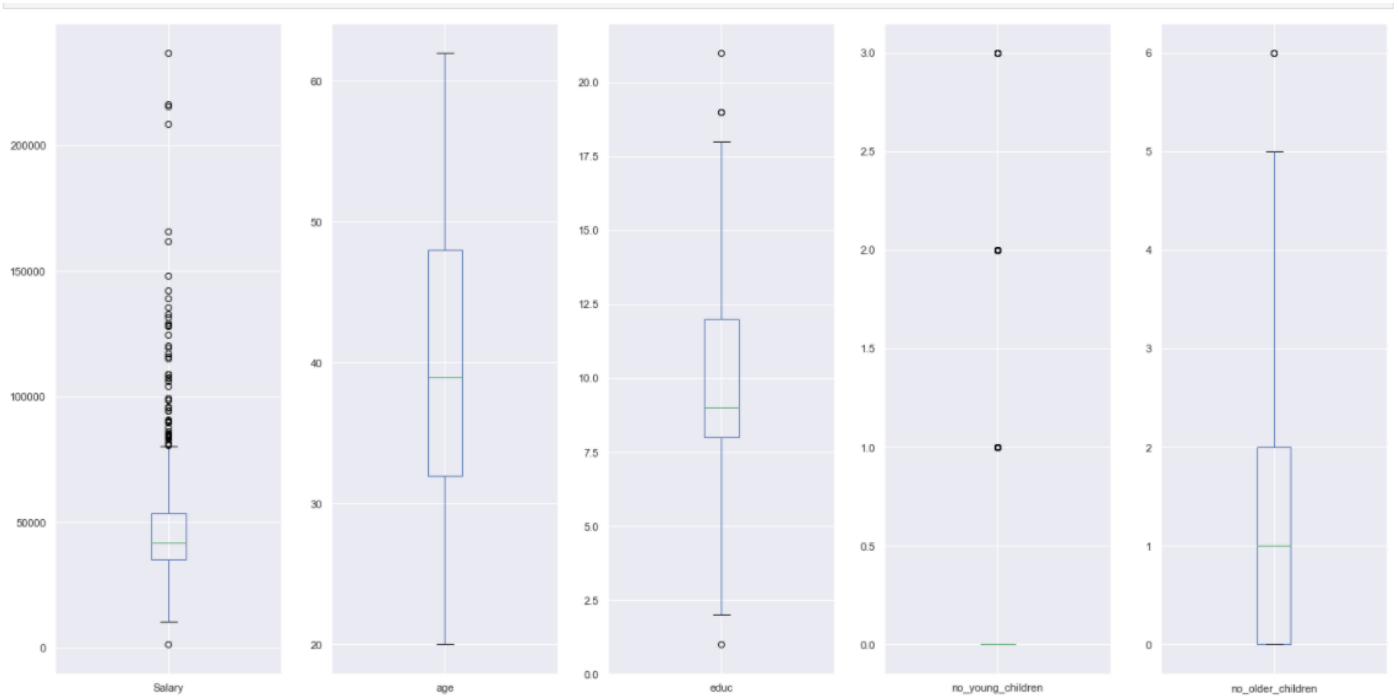
Dataset has no duplicated rows.

f.) Remove the Unnamed column.

Head of dataset after removing unwanted column:

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

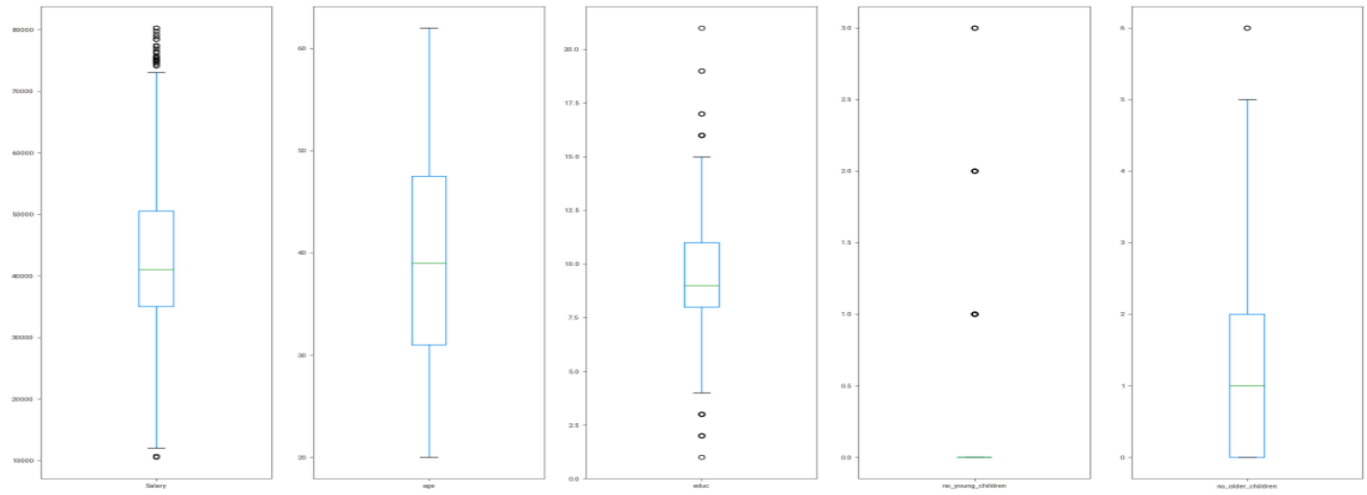
g) Check for outliers



Inference:

Only Salary and age is continuous variable. others are categorical. Only salary variables have outlier in the continuous variable

h) Boxplot of the dataset after handling Outliers



Inference:

Removed the outliers present in the salary column

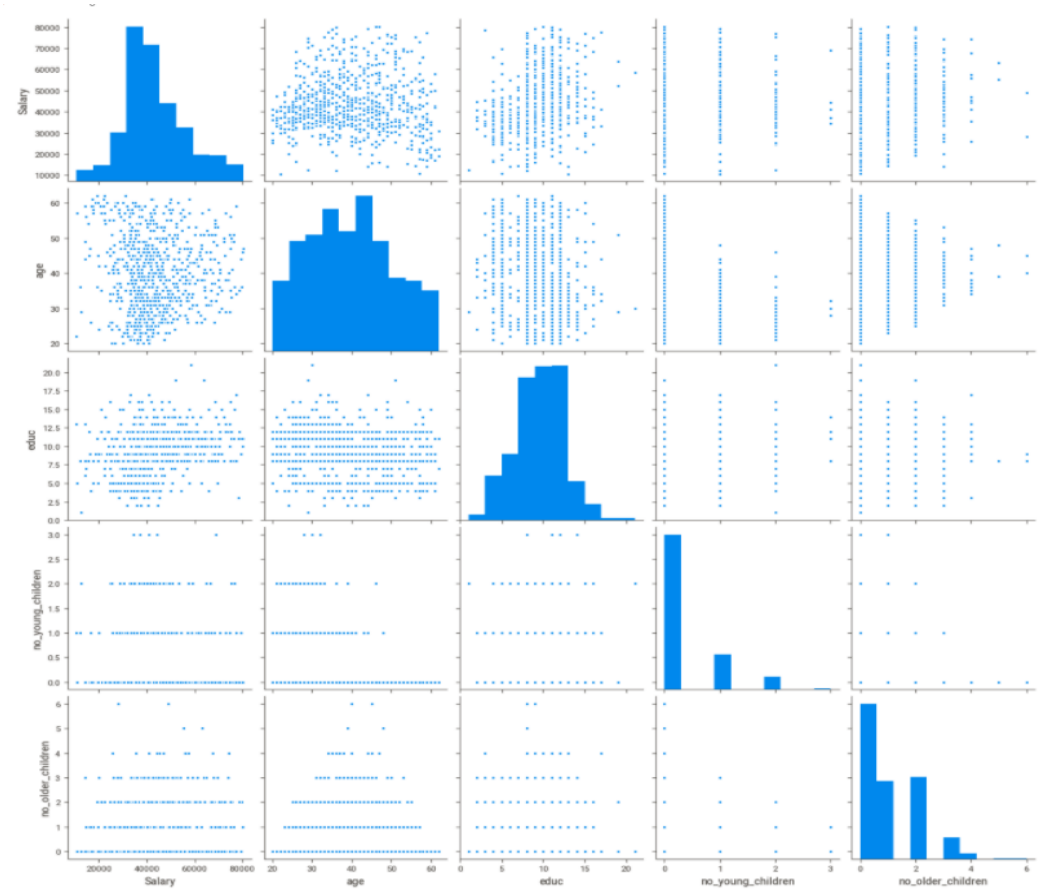
i)Value counts of categorical variables:

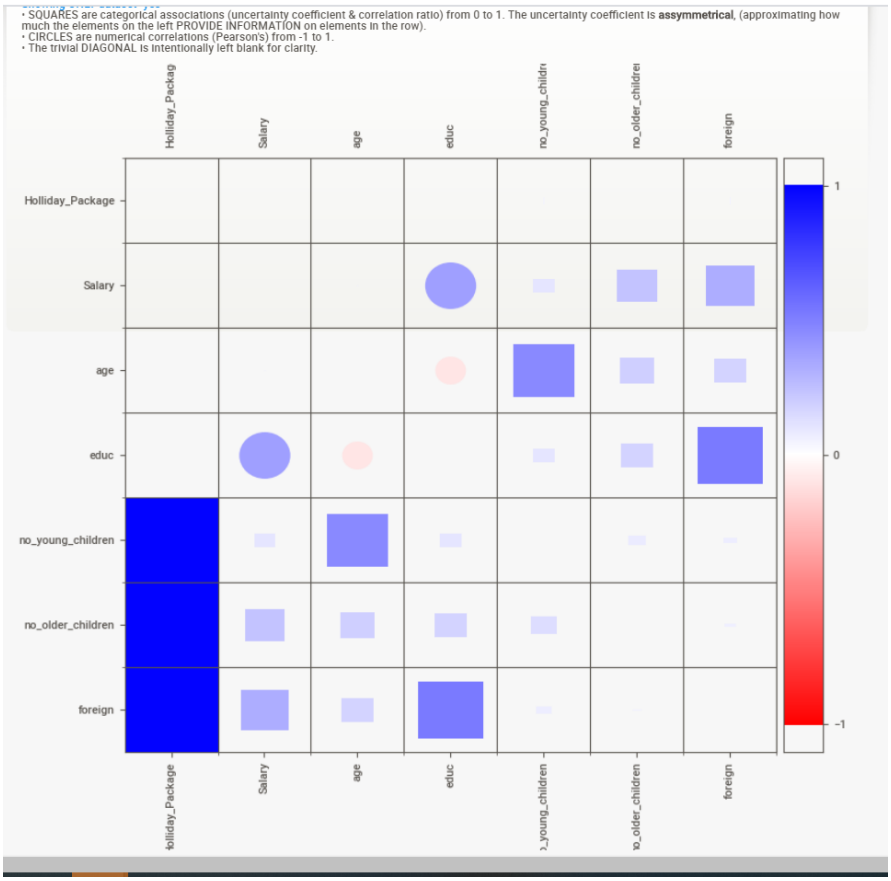
```
HOLLIDAY_PACKAGE :  2
yes    389
no     426
Name: Holliday_Package, dtype: int64


FOREIGN :  2
yes    211
no     604
Name: foreign, dtype: int64
```

```
no_young_children
0    617
1    141
2     52
3      5
Name: no_young_children, dtype: int64
```

i)Graphical Analysis:

A.) Multivariate Analysis

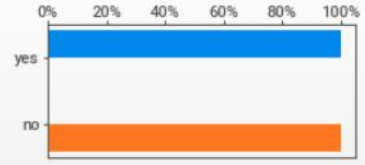

Salary
age
educ
no_young_children
no_older_children


• SQUARES are categorical associations (uncertainty coefficient & correlation ratio) from 0 to 1. The uncertainty coefficient is assymmetrical, (approximating how much the elements on the left PROVIDE INFORMATION on elements in the row).
• CIRCLES are numerical correlations (Pearson's) from -1 to 1.
• The trivial DIAGONAL is intentionally left blank for clarity.
Holliday_Package
Salary
age
educ
no_young_children
no_older_children
foreign
1
0
-1

B.) Univariate and Bivariate Analysis

1 **Holliday_Package**

VALUES:	389 (100%)	426 (100%)
MISSING:	---	---
DISTINCT:	1 (0%)	1 (0%)

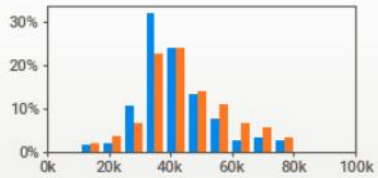


2 **Salary**

VALUES:	389 (100%)	426 (100%)
MISSING:	---	---
DISTINCT:	388 (99%)	423 (99%)

MAX	79,732	80,297
95%	67,718	69,770
Q3	47,361	52,940
AVG	41,554	44,787
MEDIAN	39,218	42,352
Q1	34,293	36,277
5%	25,325	22,754
MIN	12,578	10,551

RANGE	67,154	69,746
IQR	13,068	16,663
STD	12,156	13,750
VAR	147.8M	189.1M
KURT.	0.964	-0.068
SKEW	0.800	0.357
SUM	16.2M	19.1M

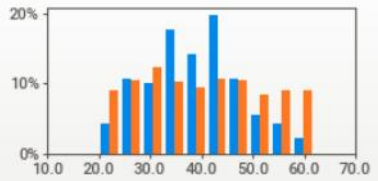


3 **age**

VALUES:	389 (100%)	426 (100%)
MISSING:	---	---
DISTINCT:	42 (11%)	43 (10%)

MAX	61.0	62.0
95%	55.0	59.0
Q3	45.0	50.0
AVG	38.8	40.4
MEDIAN	38.0	40.0
Q1	32.0	30.0
5%	25.0	23.0
MIN	20.0	20.0

RANGE	41.0	42.0
IQR	13.0	20.0
STD	9.05	11.7
VAR	81.9	137
KURT.	-0.521	-1.20
SKEW	0.120	0.112
SUM	15,096	17,217

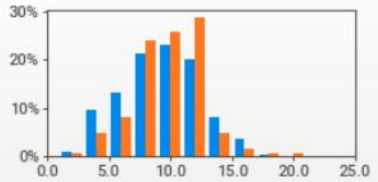


4 **educ**

VALUES:	389 (100%)	426 (100%)
MISSING:	---	---
DISTINCT:	17 (4%)	18 (4%)

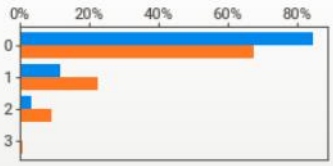
MAX	17.0	21.0
95%	14.0	13.8
Q3	11.0	11.0
MEDIAN	9.0	9.0
AVG	8.9	9.4
Q1	7.0	8.0
5%	4.0	4.0
MIN	1.0	2.0

RANGE	16.0	19.0
IQR	4.00	3.00
STD	3.16	2.79
VAR	9.98	7.81
KURT.	-0.519	0.935
SKEW	-0.010	0.088
SUM	3,451	4,011



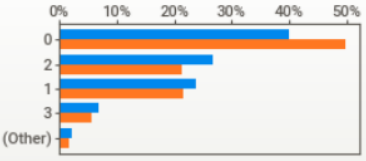
5 **no_young_children**

VALUES:	389 (100%)	426 (100%)
MISSING:	---	---
DISTINCT:	4 (1%)	4 (1%)



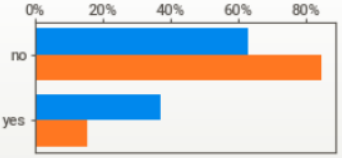
6 **no_older_children**

VALUES:	389 (100%)	426 (100%)
MISSING:	---	---
DISTINCT:	6 (2%)	6 (1%)



7 **foreign**

VALUES:	389 (100%)	426 (100%)
MISSING:	---	---
DISTINCT:	2 (1%)	2 (0%)



Insights:

There is high chance for them to take the package if the employee salary is between 30k to 40 k.
This suggest that package pricing is average pricing.

There Is a higher chance of taking up the package if the employee age is between age of 25 of to 50 years.
After 50 years there is a higher chance of telling no.

If the employee has no young children then there is huge chance to tell yes.

If employee is a foreigner there is a huge chance to tell yes.

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).

a) Encode data

Head of dataset after encoding

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

Inference:

Foreign column is encoded to zero (if no) and one (if yes)

b) Data Split:

```
numpy.matrix
```

Inference:

Data is successfully split into train and test (70:30) and random state is 1

c) Data Apply Logistic Regression and Linear discriminant analysis

A.) Logistic Regression

```
LogisticRegression(solver='liblinear')
```

B.) LDA

```
LinearDiscriminantAnalysis()
```

Inference:

Data is encoded , split into training an testing and model for logistic regression and LDA is built

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.

a) Performance Metrics for Logistic Regression

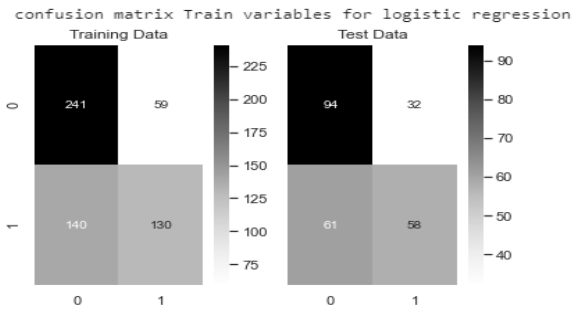
Accuracy score for Training data

```
Accuracy score  for Logistic regression  train variables
0.6508771929824562
```

Accuracy score for Testing data

```
Accuracy score for Logistic regression test variables
0.6204081632653061
```

Confusion Matrix for Logistic Regression



Classification Report for Logistic Regression:

```
Logistic regression Classfication report
Classification Report of the training data:
              precision    recall  f1-score   support

          no       0.63      0.80      0.71       300
         yes       0.69      0.48      0.57       270

    accuracy                           0.65       570
   macro avg       0.66      0.64      0.64       570
weighted avg       0.66      0.65      0.64       570


Classification Report of the test data:
              precision    recall  f1-score   support

          no       0.61      0.75      0.67       126
         yes       0.64      0.49      0.56       119

    accuracy                           0.62       245
   macro avg       0.63      0.62      0.61       245
weighted avg       0.62      0.62      0.61       245
```

ROC curve and ROC_AUC score for Logistic Regression

```
AUC and ROC FOR Logistic regression
AUC for the Training Data: 0.737
AUC for the Test Data: 0.665
```

b) Performance Metrics for LDA

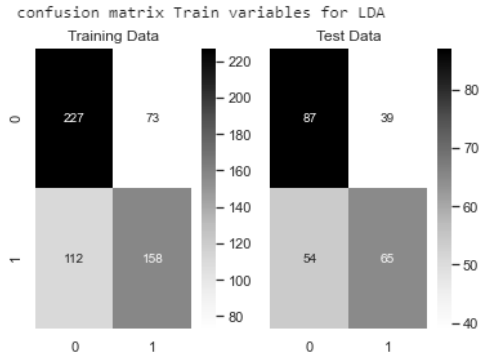
Accuracy score for Training data

```
Accuracy score  for LDA  train variables
0.6754385964912281
```

Accuracy score for Testing data

```
Accuracy score for LDA test variables
0.6204081632653061
```

Confusion Matrix For LDA



Classification Report For LDA

```
LDA Classfication report
Classification Report of the training data:

              precision    recall  f1-score   support

          no       0.67      0.76      0.71       300
         yes       0.68      0.59      0.63       270

    accuracy                           0.68       570
   macro avg       0.68      0.67      0.67       570
weighted avg       0.68      0.68      0.67       570


Classification Report of the test data:

              precision    recall  f1-score   support

          no       0.62      0.69      0.65       126
         yes       0.62      0.55      0.58       119

    accuracy                           0.62       245
   macro avg       0.62      0.62      0.62       245
weighted avg       0.62      0.62      0.62       245
```

ROC curve and ROC_AUC score for LDA

```
AUC and ROC FOR LDA
AUC for the Training Data: 0.743
AUC for the Test Data: 0.670
```

b) Performance Metrics for Final Model: Compare Both the models and write inference which model is best/optimized.

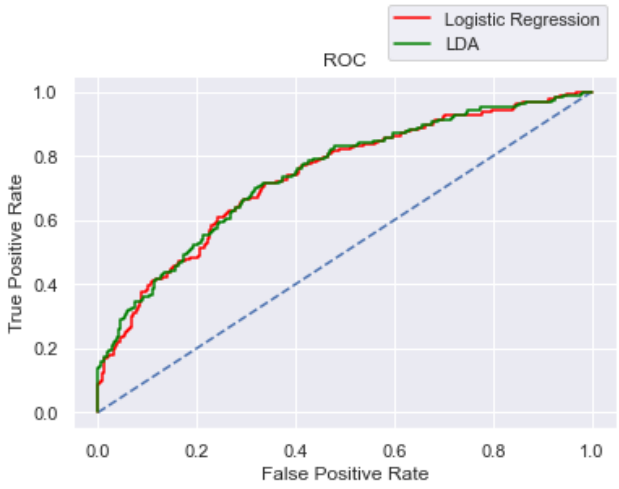
Comparison in Table form:

	Logistic reg Train	Logistic reg Test	LDA Train	LDA Test
Accuracy	0.65	0.62	0.68	0.62
AUC	0.74	0.67	0.74	0.67
Recall	0.48	0.49	0.59	0.55
Precision	0.69	0.64	0.68	0.62
F1 Score	0.57	0.56	0.63	0.58

ROC curve and ROC_AUC score comparison

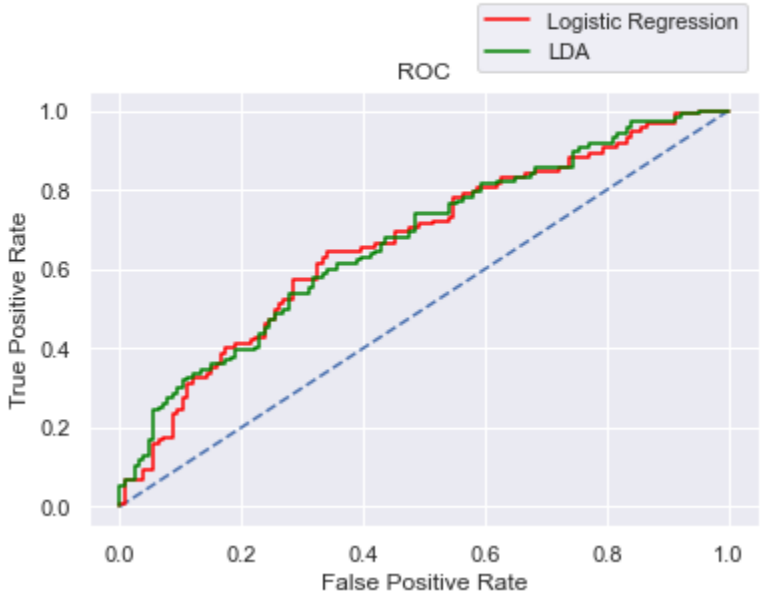
A.) Train data

```
ROC curve for Train data
<matplotlib.legend.Legend at 0x21596cd4580>
```



B.) Test Data:

```
ROC curve for Test data
[3]: <matplotlib.legend.Legend at 0x2159cd140d0>
```



Inference:

Based on comparing the performance metrics, Linear discriminant analysis (LDA) performs better than the Logistic regression because it has the best recall rate .Even accuracy is more for LDA.So it is the best model .

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

The Linear discriminant analysis model will be able to predicting whether an employee will opt for the package or not with around 70 percent accuracy.

Business Insights:

The important factors which determine whether an employee will opt in for package are

Salary, Age, no of young children and foreign.

The company must focus on the people who earns between 30 to 40 k and between age 25 to 50 years and if they have no children, there is a huge chance for them to opt in for a package.

Recommendation:

The greatest number of people who are opting in for the package has a salary of range between 30 to 40 k. It suggests that the package is of average price with medium level facilities
So, if they add some additional luxury packages with facilities like booking in star hotels, luxury cars etc. it may help to increase the sales of packages to a higher income group.

The analysis shows that a greater number of foreigners opt in for packages than the non-foreigners. This along with the previous analysis which shows that most of the people are from salary group of 30 to 50k(so it is not expensive package) suggest that packages provided are either of local sightseeing place or of less interest to the non-foreigners. So, suggest the company to add some more activities or places in their packages.

The analysis shows that data if the employee as no young children, there is more Chace of them taking up the package. As count of children increases, the willingness to opt in for a holiday package decreases. So, I suggest the company to provide additional discounts or children attractiveness for the employee who has young children to boost up the chance of them opting in for the package .