PPROJECT REPORT- PREDICTIVE MODELING

JUNE 19, 2022

SHARJIL SHAH
Great Learning, G11 Jan_22A

TABLE OF CONTENTS

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 worst and J the best.
Clarity	Clarity refers to the absence of the Inclusions and Blemishes. (In order from Worst to Best in terms of avg price) IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, l1
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.

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

Ans: Checking if data has flown in properly:

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

Tail of data:

	carat	cut	color	clarity	depth	table	x	y	z	price
26962	1.11	Premium	G	SI1	62.3	58.0	6.61	6.52	4.09	5408
26963	0.33	Ideal	H	IF	61.9	55.0	4.44	4.42	2.74	1114
26964	0.51	Premium	E	VS2	61.7	58.0	5.12	5.15	3.17	1656
26965	0.27	Very Good	F	VVS2	61.8	56.0	4.19	4.20	2.60	682
26966	1.25	Premium	J	SI1	62.0	58.0	6.90	6.88	4.27	5166

Shape:

```
(26967, 10)
```

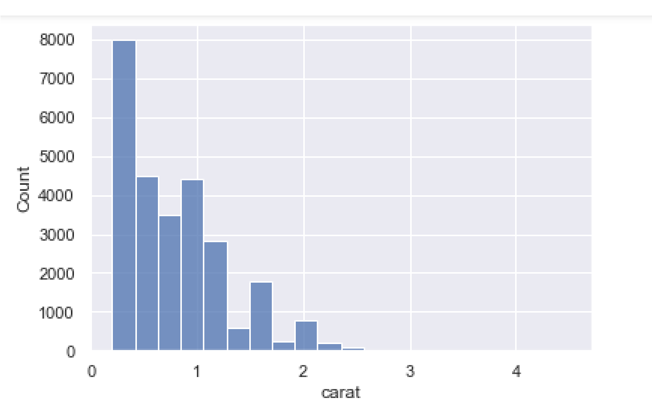
Description of data:

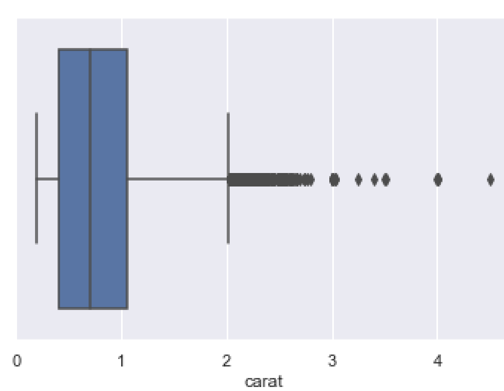
	carat	depth	table	x	y	z	price
count	26967.000000	26270.000000	26967.000000	26967.000000	26967.000000	26967.000000	26967.000000
mean	0.785860	61.745147	57.407702	5.729438	5.731334	3.537316	3939.518115
std	0.444042	1.412860	2.090151	1.124638	1.116593	0.694826	4024.864666
min	0.200000	50.800000	51.600000	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

Data Info: Dataset has int, float and object data types

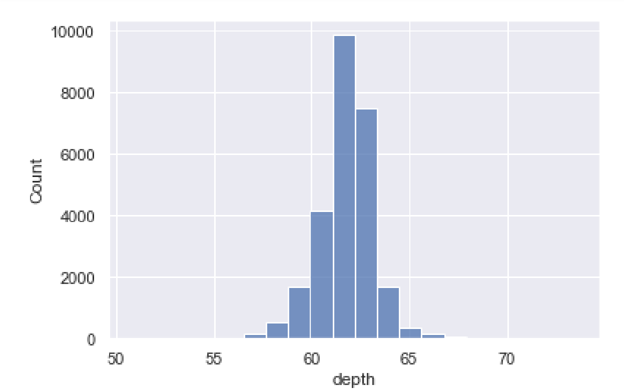
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26967 entries, 0 to 26966
Data columns (total 11 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Unnamed: 0  26967 non-null  int64
 1   carat       26967 non-null  float64
 2   cut         26967 non-null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

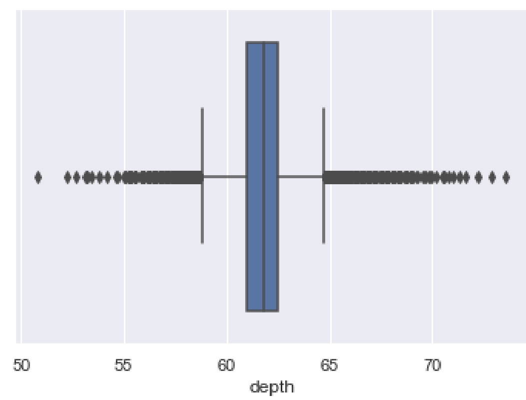
Univariate and Bivariate Analysis





The distribution of data in carat seems to positively skewed, as there are multiple peaks points in the distribution there could multimode and the box plot of carat seems to have large number of outliers. In the range of 0 to 1 where majority of data lies.

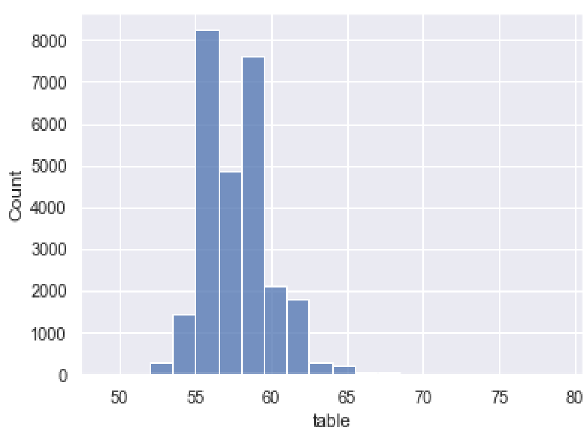


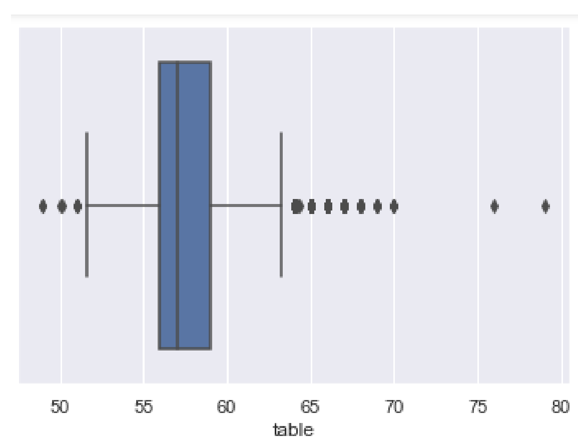


The distribution of depth seems to be normal distribution,

The depth ranges from 55 to 65

The box plot of the depth distribution holds many outliers.

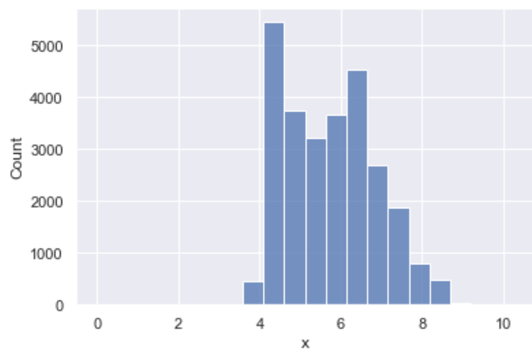


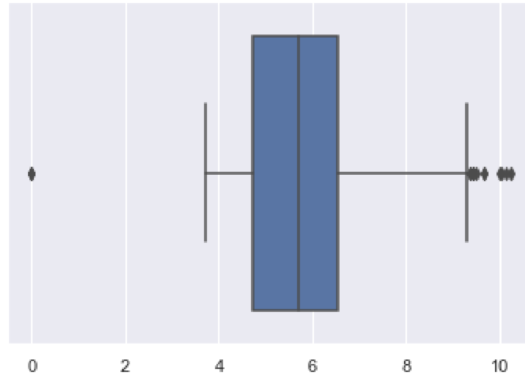


The distribution of table also seems to be positively skewed

The box plot of table has outliers

The data distribution where there is maximum distribution is between 55 to 65

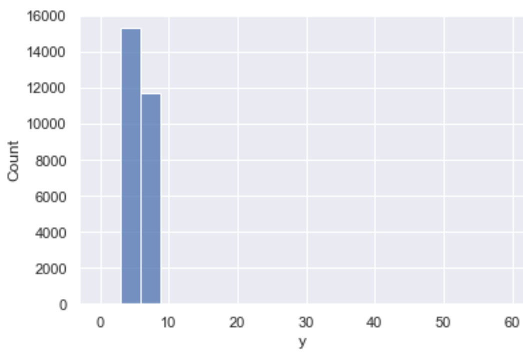


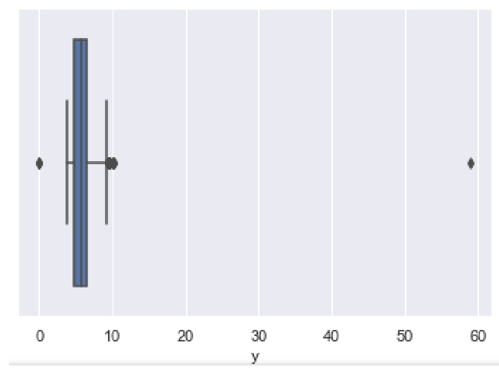


The distribution of x (Length of the cubic zirconia in mm.) is positively skewed

The box plot of the data consists of many outliers

The distribution rages from 4 to 8

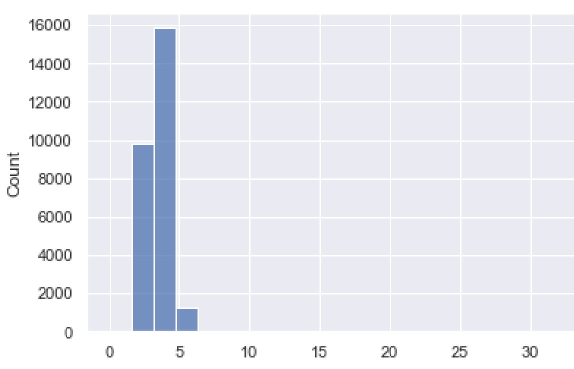


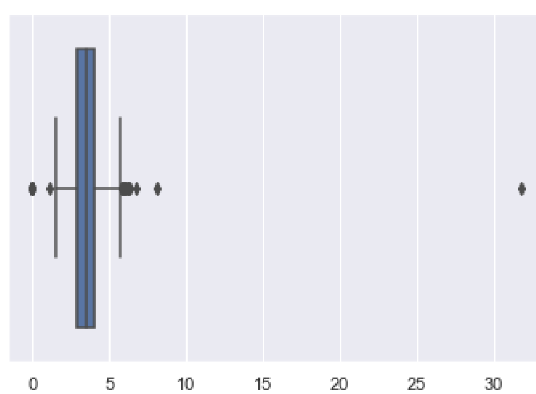


The distribution of Y (Width of the cubic zirconia in mm.) is positively skewed

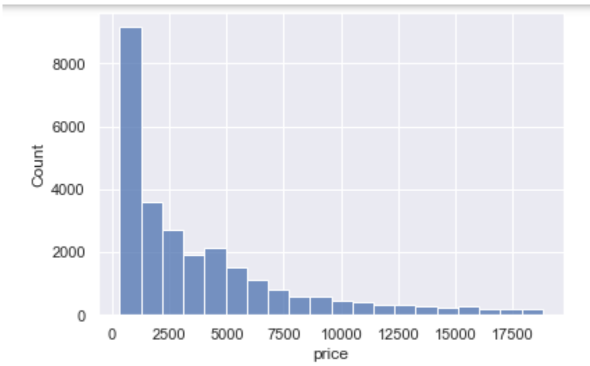
The box plot also consists of outliers

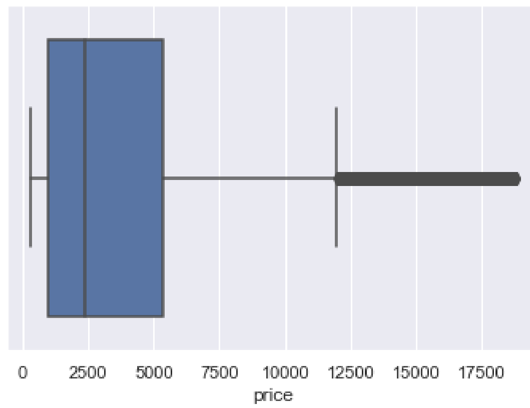
The distribution too much positively skewed. The skewness may be due to the diamonds are always made in specific shape. There might not be too much sizes in the market





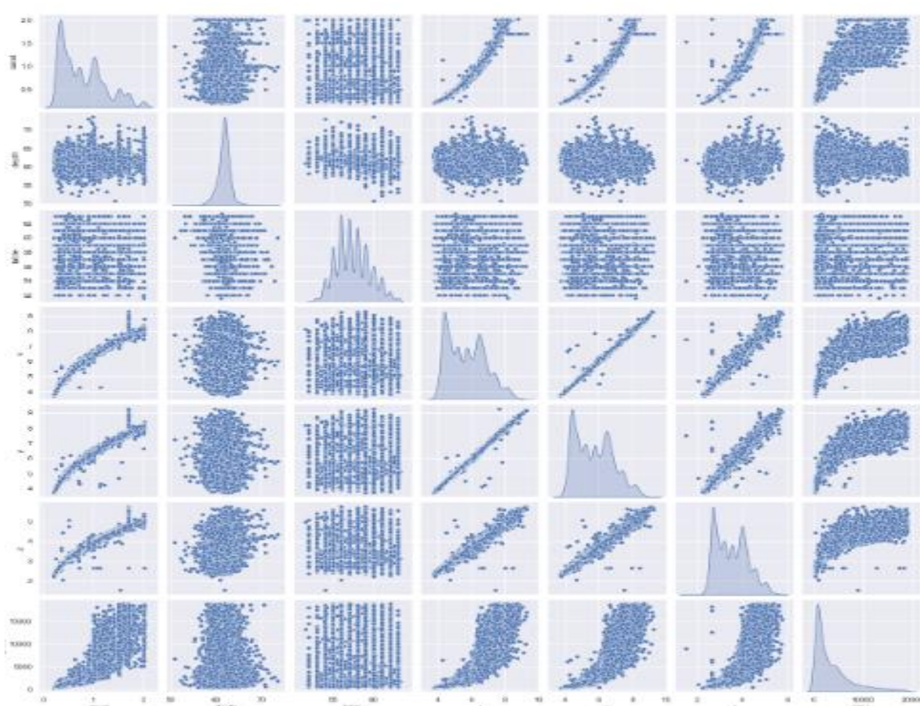
The distribution of z (Height of the cubic zirconia in mm.) is positively skewed The box plot also consists of outliers The distribution too much positively skewed. The skewness may be due to the diamonds are always made in specific shape. There might not be too much sizes in the market

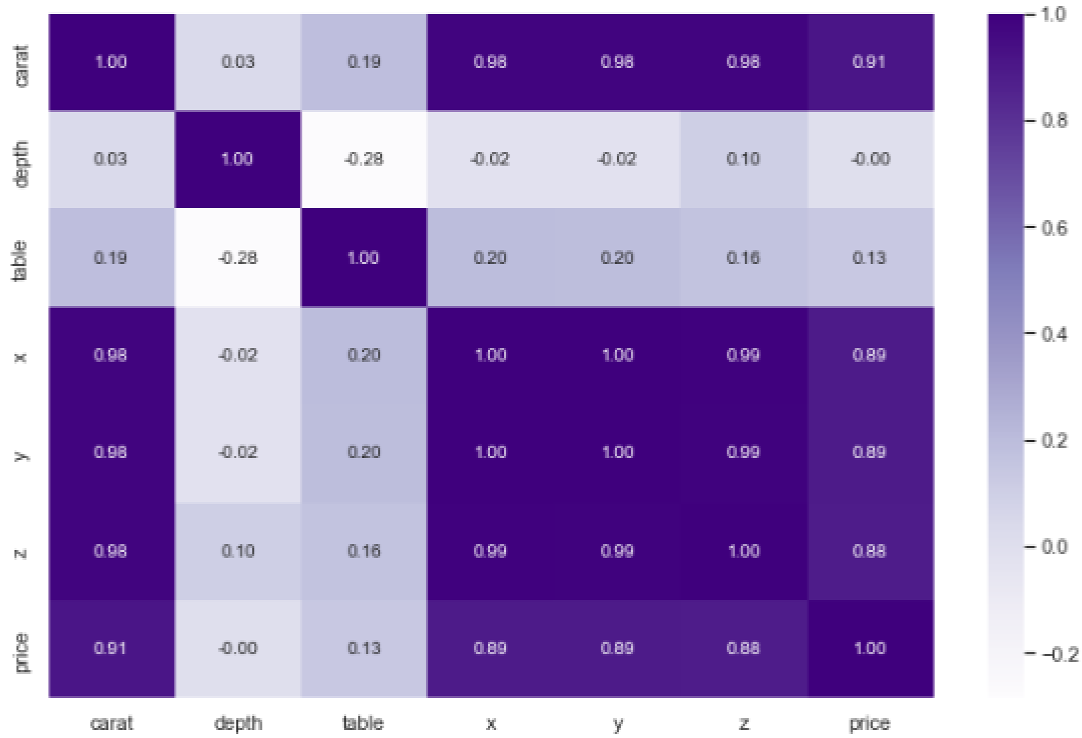




The price has seems to be positively skewed. The skew is positive The price has outliers in the data The price distribution is from rs 100 to 8000.

Multivariate Analysis





This matrix clearly shows the presence of multi collinearity in the dataset.

1.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Check for the possibility of combining the sub levels of an ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning.

Ans: Based on the below, all columns except for depth has no null values.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26967 entries, 0 to 26966
Data columns (total 10 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   carat    26967 non-null  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
```

Yes we have Null values in depth, since depth being continuous variable mean or median imputation can be done. The percentage of Null values is less than 5%, we can also drop these if we want. After median imputation, we don't have any null values in the dataset.

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

Checking if there is value that is "0"

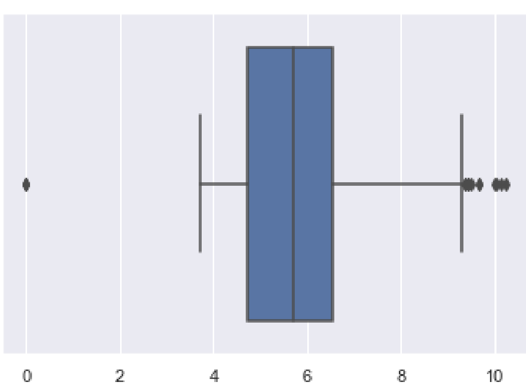
We have certain rows having values zero, the x, y, z are the dimensions of a diamond so this can't take into model. As there are very less rows.

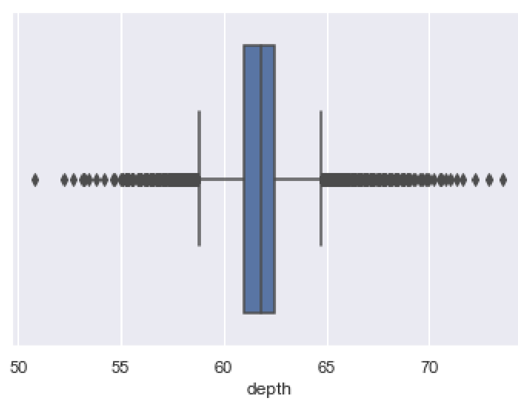
We can drop these rows as don't have any meaning in model building

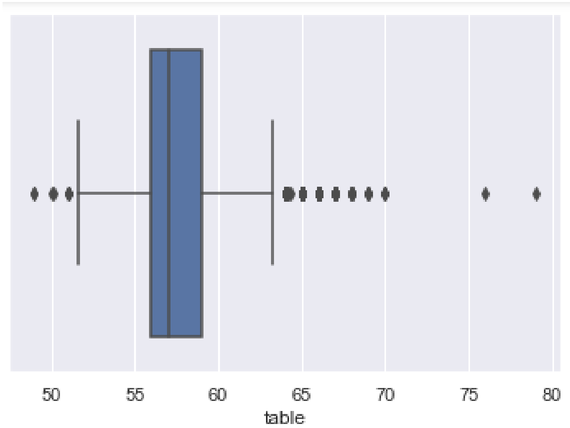
SCALING

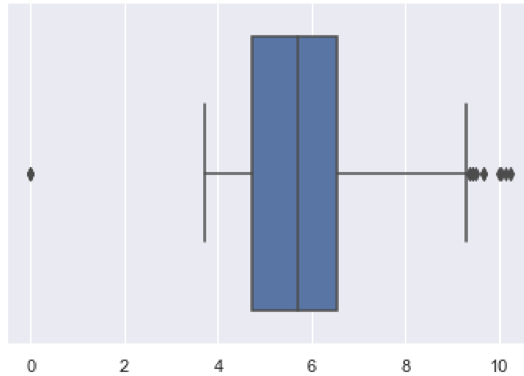
Scaling can be useful to reduce or check the multi collinearity in the data, so if scaling is not applied I find the VIF – variance inflation factor values very high. Which indicates presence of multi collinearity. These values are calculated after building the model of linear regression. To understand the multi collinearity in the model. The scaling had no impact in model score or coefficients of attributes nor the intercept.

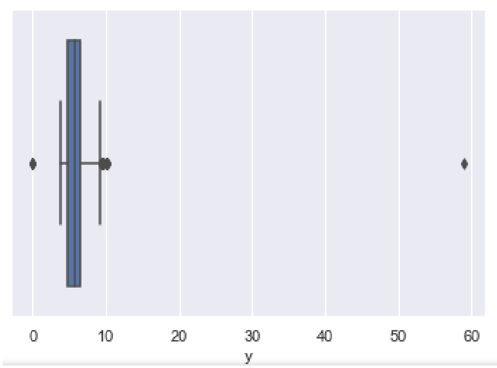
CHECKING THE OUTLIERS IN THE DATA

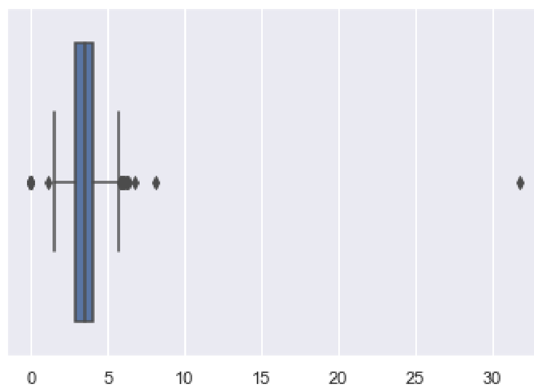


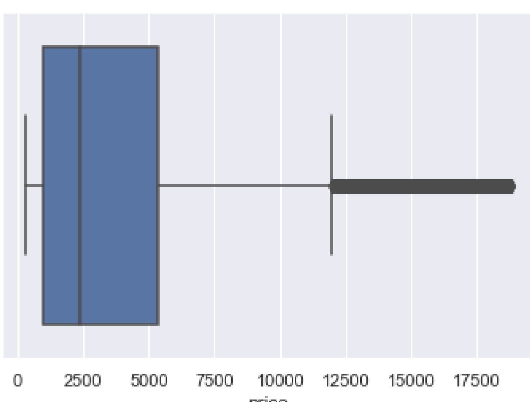












After imputation is done, we see that there are no null values present

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

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

	carat	cut	color	clarity	depth	table	x	y	z	price
0	0.3	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	2.7	984
2	0.9	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	4.8	2.96	1082
4	0.31	Ideal	F	VVS1	60.4	59.0	4.35	4.43	2.65	779

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

```
<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  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
```

	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

1.2 Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning.

	carat	cut	color	clarity	depth	table	x	y	z	price
0	0.3	5	E	SI1	62.1	58.0	4.27	4.29	2.66	499
1	0.33	4	G	IF	60.8	58.0	4.42	4.46	2.7	984
2	0.9	3	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289
3	0.42	5	F	VS1	61.6	56.0	4.82	4.8	2.96	1082
4	0.31	5	F	VVS1	60.4	59.0	4.35	4.43	2.65	779

	carat	cut	color	clarity	depth	table	x	y	z	price
0	0.3	5	6	SI1	62.1	58.0	4.27	4.29	2.66	499
1	0.33	4	4	IF	60.8	58.0	4.42	4.46	2.7	984
2	0.9	3	6	VVS2	62.2	60.0	6.04	6.12	3.78	6289
3	0.42	5	5	VS1	61.6	56.0	4.82	4.8	2.96	1082
4	0.31	5	5	VVS1	60.4	59.0	4.35	4.43	2.65	779

	carat	cut	color	clarity	depth	table	x	y	z	price
0	0.3	5	6	5	62.1	58.0	4.27	4.29	2.66	499
1	0.33	4	4	10	60.8	58.0	4.42	4.46	2.7	984
2	0.9	3	6	8	62.2	60.0	6.04	6.12	3.78	6289
3	0.42	5	5	7	61.6	56.0	4.82	4.8	2.96	1082
4	0.31	5	5	9	60.4	59.0	4.35	4.43	2.65	779

	carat	cut	color	clarity	depth	table	x	y	z	price
0	-1.094197	0.979550	0.94147	-0.639402	0.249646	0.283381	-1.297720	-1.290856	-1.262665	-0.854832
1	-1.026634	0.081246	-0.23089	2.396400	-0.682226	0.283381	-1.164341	-1.138605	-1.205096	-0.734329
2	0.257052	-0.817058	0.94147	1.182079	0.321328	1.240267	0.276149	0.348088	0.349280	0.583753
3	-0.823947	0.979550	0.35529	0.574919	-0.108766	-0.673506	-0.808665	-0.834101	-0.830894	-0.709979
4	-1.071676	0.979550	0.35529	1.789239	-0.968956	0.761824	-1.226585	-1.165473	-1.277057	-0.785263

```
<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
```

	count	mean	std	min	25%	50%	75%	max
carat	26967.0	-1.186263e-16	1.000019	-1.319405	-0.868988	-0.193364	0.594864	2.779383
cut	26967.0	4.798569e-16	1.000019	-2.613667	-0.817058	0.081246	0.979550	0.979550
color	26967.0	1.487790e-16	1.000019	-1.989430	-0.817070	-0.230890	0.941470	1.527650
clarity	26967.0	6.182452e-17	1.000019	-1.853722	-0.639402	-0.032241	0.574919	2.396400
depth	26967.0	-4.311599e-16	1.000019	-7.850470	-0.467179	0.106281	0.536376	8.493127
table	26967.0	-7.827470e-16	1.000019	-2.778656	-0.673506	-0.195062	0.761824	2.819130
x	26967.0	-2.734161e-16	1.000019	-1.777884	-0.906476	-0.035068	0.729637	3.174915
y	26967.0	-2.663514e-16	1.000019	-1.810303	-0.914705	-0.019107	0.724240	3.160267
z	26967.0	-7.779919e-16	1.000019	-2.889003	-0.917248	-0.024921	0.723482	3.184577
price	26967.0	-2.910285e-17	1.000019	-0.897815	-0.744018	-0.388720	0.352933	3.696710

```
      carat       cut     color   clarity     depth     table         x  \
0 -1.094197  0.979550  0.94147 -0.639402  0.249646  0.283381 -1.297720
1 -1.026634  0.081246 -0.23089  2.396400 -0.682226  0.283381 -1.164341
2  0.257052 -0.817058  0.94147  1.182079  0.321328  1.240267  0.276149
3 -0.823947  0.979550  0.35529  0.574919 -0.108766 -0.673506 -0.808665
4 -1.071676  0.979550  0.35529  1.789239 -0.968956  0.761824 -1.226585

          y         z
0 -1.290856 -1.262665
1 -1.138605 -1.205096
2  0.348088  0.349280
3 -0.834101 -0.830894
4 -1.165473 -1.277057
```

```
      price
0 -0.854832
1 -0.734329
2  0.583753
3 -0.709979
4 -0.785263
```

```
          carat       cut     color  clarity     depth     table         x  \
11687 -0.846468  0.979550 -1.40325  1.182079  0.393011 -0.673506 -0.853124
9728   2.081238  0.979550 -1.98943 -0.639402  0.751423 -0.195062  1.645505
1936  -1.026634 -1.715362  0.35529 -0.639402  0.034599  2.197154 -1.182125
26220 -0.193364 -0.817058 -0.81707 -0.639402  0.751423 -0.195062 -0.106203
18445 -0.193364  0.979550  1.52765 -1.246562  0.249646 -0.673506 -0.052852

              y         z
11687 -0.896793 -0.830894
9728   1.628794  1.745340
1936  -1.147561 -1.147526
26220 -0.063887  0.003863
18445 -0.019107 -0.010529
          carat       cut     color  clarity     depth     table         x  \
18031  2.756863 -2.613667 -1.40325 -1.246562  3.403673  1.718711  1.850019
26051  1.630822  0.081246  0.35529 -0.639402  0.321328  0.761824  1.432099
16279 -0.643780 -0.817058 -0.81707 -0.639402 -0.610543  1.718711 -0.595259
16466 -1.071676  0.979550  0.94147  0.574919  0.177963 -0.673506 -1.191017
19837  0.932677 -0.817058 -0.81707  0.574919  0.177963 -0.195062  0.925259

              y         z
18031  1.807913  2.349820
26051  1.404894  1.457493
16279 -0.520642 -0.615008
16466 -1.156517 -1.262665
19837  0.966051  0.968152
          price
11687 -0.715197
9728   0.591455
1936  -0.845639
26220 -0.428723
18445 -0.339028
          price
18031  1.672505
26051  1.905064
16279 -0.697308
16466 -0.823277
19837  0.555925
```

```
The coefficient for carat is 1.2801213328224776
The coefficient for cut is 0.044061306493l8646
The coefficient for color is 0.1233528534141011
The coefficient for clarity is 0.1924067541374261
The coefficient for depth is -0.003832957799618479 6
The coefficient for table is -0.015416741736580713
The coefficient for x is -0.5361037488818727
The coefficient for y is 0.44081340476733066
The coefficient for z is -0.16420841159037242
```

```
The intercept for our model is 0.0015672526389941363
```

R square on train data:

```
Out[46]: 0.8886993336877839
```

R square on test data:

```
Out[47]: 0.883659588050507
```

RMSE on Training data:

```
Out[48]: 0.33336543663305496
```

RMSE on Test data:

```
Out[49]: 0.34168755937542916
```

We still find we have multi collinearity in the dataset, to drop these values to lower level we can drop columns after doing stats model.

From stats model we can understand the features that do not contribute to the Model

We can remove those features after that the Vif Values will be reduced

Ideal value of VIF is less than 5%.

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.

1.4 Inference: Basis on these predictions, what are the business insights and recommendations. Please explain and summarize the various steps performed in this project. There should be proper business interpretation and actionable insights present.

We had a business problem to predict the price of the stone and provide insights for the company on the profits on different prize slots. From the EDA analysis we could understand the cut, ideal cut had number 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 l1, l2, l3 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.

For better accuracy dropping depth column in iteration for better results.

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

PROBLEM 2: LINEAR REGRESSION

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.

Data Dictionary

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.

Loading all the necessary library for the model building.

Now, reading the head and tail of the dataset to check whether data has been properly fed.

Head of the data

	Unnamed: 0	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	1	no	48412	30	8	1	1	no
1	2	yes	37207	45	8	0	1	no
2	3	no	58022	46	9	0	0	no
3	4	no	66503	31	11	2	0	no
4	5	no	66734	44	12	0	2	no

Info

```
<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
```

Describe

	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

We have integer and continuous data, Holiday package is our target variable Salary, age, educ and number young children, number older children of employee have the went to foreign, these are the attributes we have to cross examine and help the company predict weather the person will opt for holiday package or not.

1 No null values in the dataset,
2 We have integer and object data

Null Value Check

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

There are no NULL values found in the Data

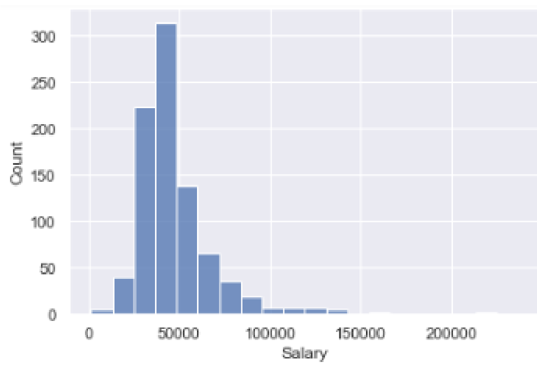
Check for Duplicate values

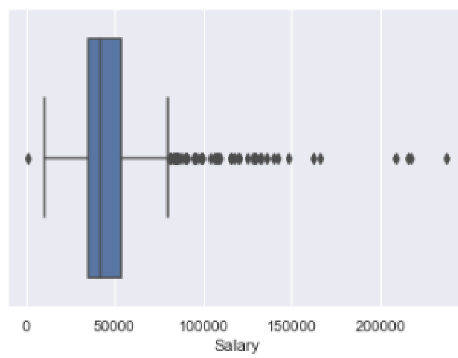
```
Number of duplicate rows = 0
```

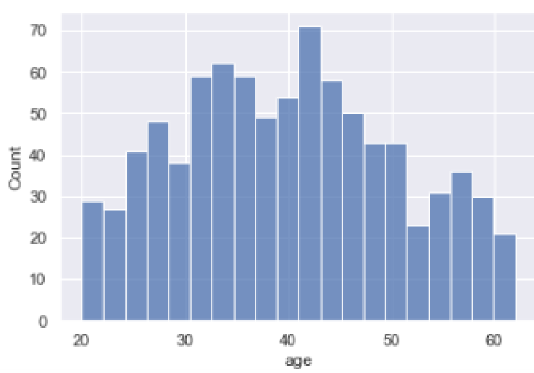
Drop unnamed column

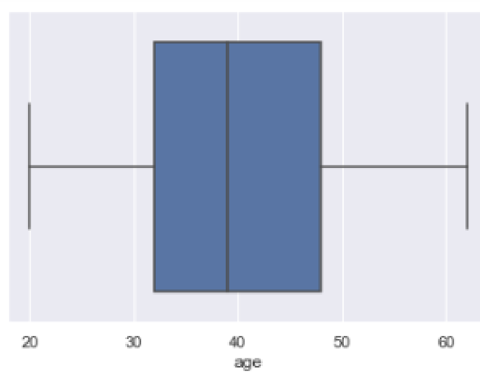
	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	no	48412	30	8	1	1	no
1	yes	37207	45	8	0	1	no
2	no	58022	46	9	0	0	no
3	no	66503	31	11	2	0	no
4	no	66734	44	12	0	2	no

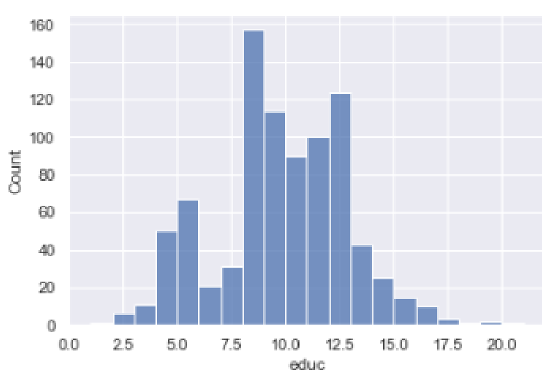
Univariate/Bivariate Analysis

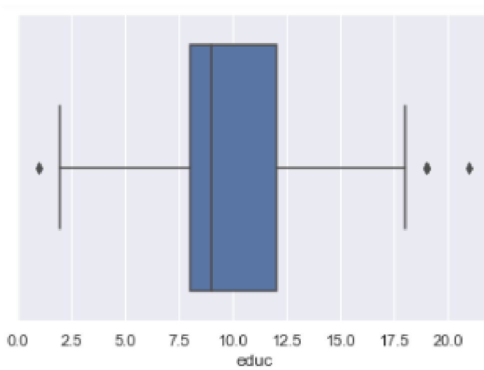


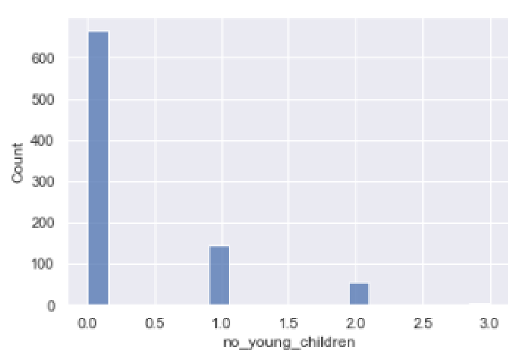


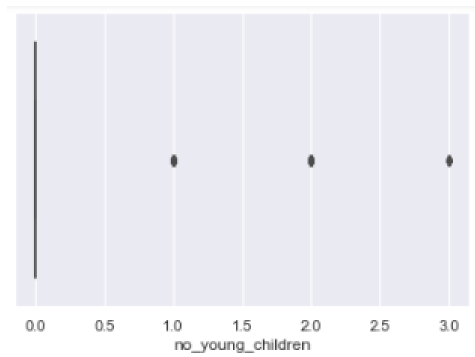


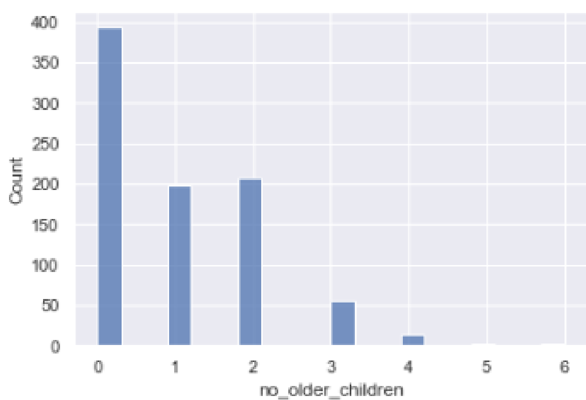


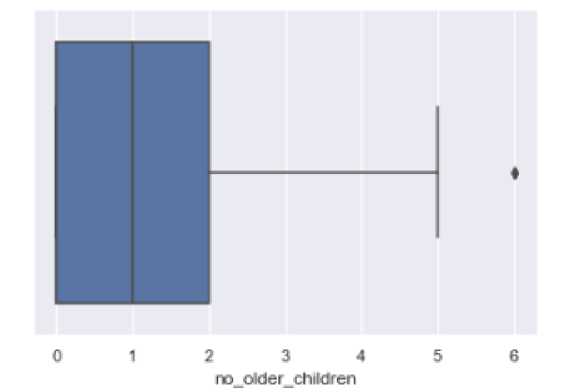


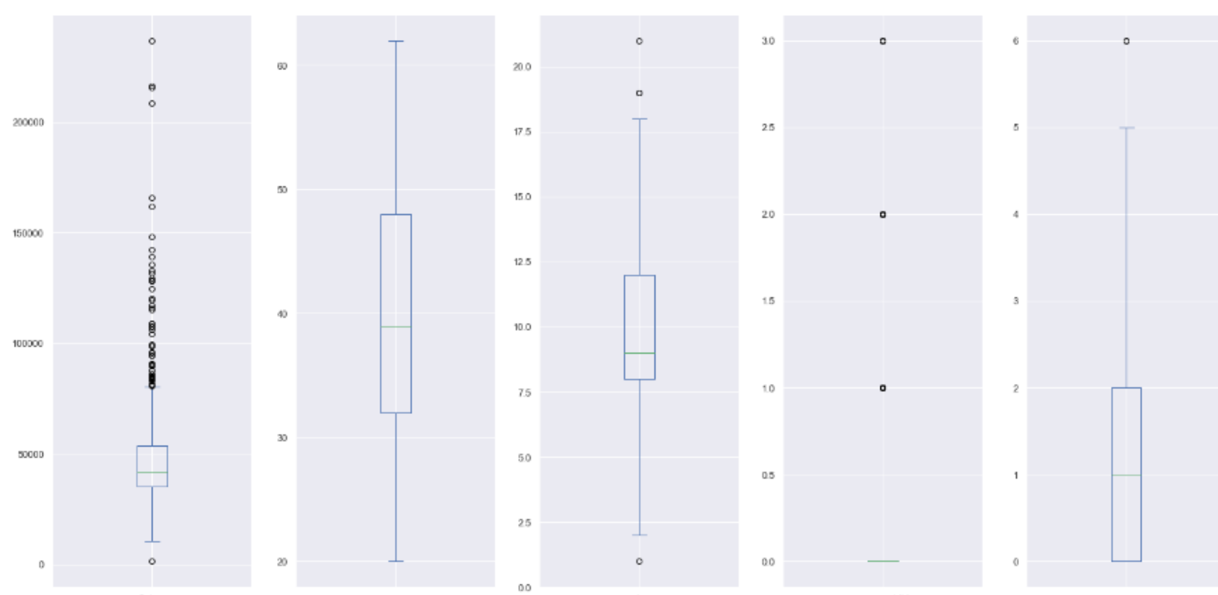








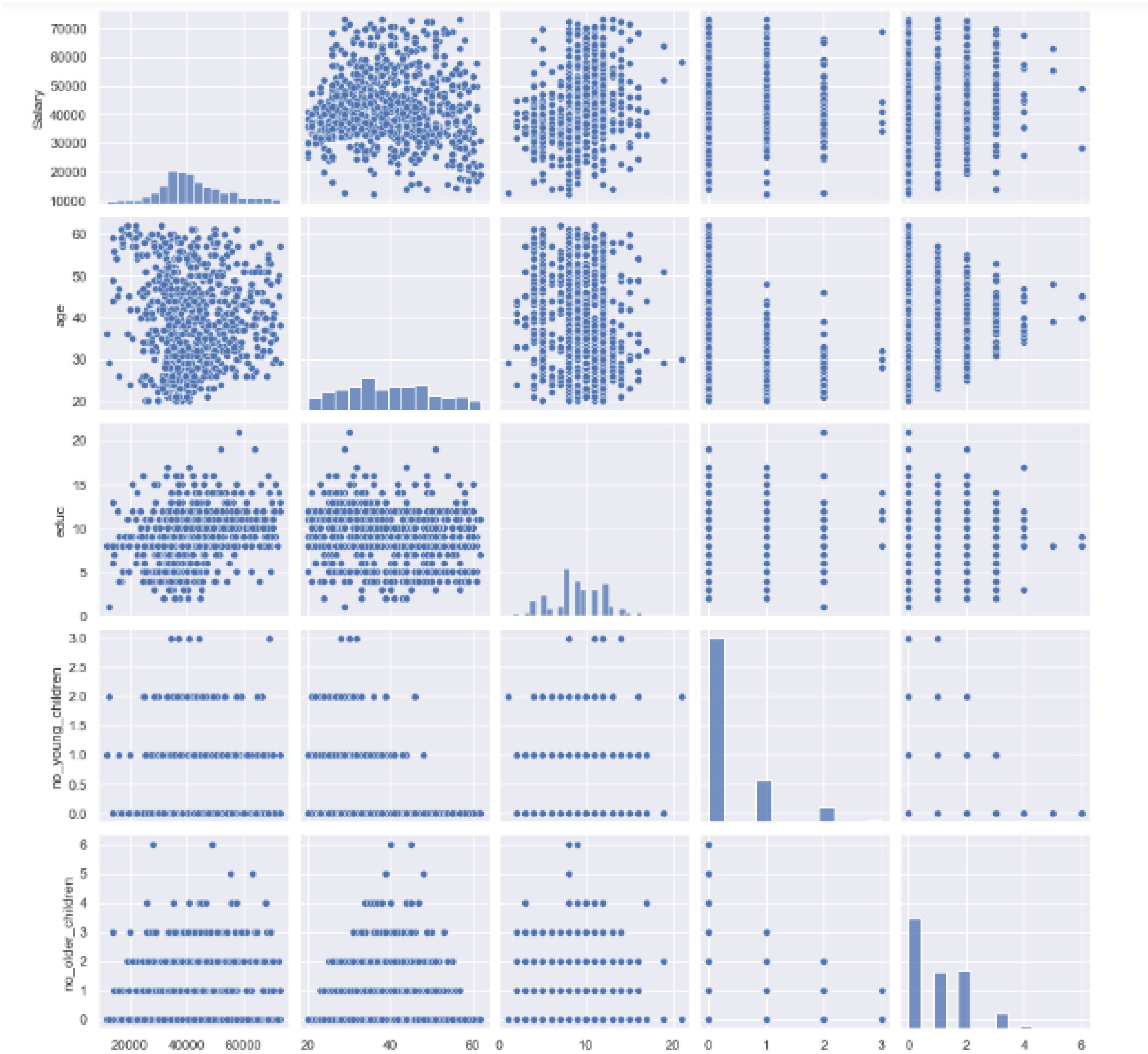




- From Holiday v/s Salary, We can see employee below salary 150000 have always opted for holiday package.

- From Age v/s Salary, 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.
- Based on the analysis, it looks like only 45% people are interested in holiday package

Multivariate Analysis



There is no correlation between the data, the data seems to be normal. There is no huge difference in the data distribution among the holiday package, I don't see any clear two different distribution in the data.

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).

	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

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 789 entries, 0 to 870
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Holliday_Package   789 non-null    object
 1   Salary             789 non-null    int64
 2   age                789 non-null    int64
 3   educ               789 non-null    int64
 4   no_young_children  789 non-null    int64
 5   no_older_children  789 non-null    int64
 6   foreign            789 non-null    object
dtypes: int64(5), object(2)
memory usage: 81.6+ KB
```

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

```
Accuracy score for Logistic regression train variables
0.644927536231884
```

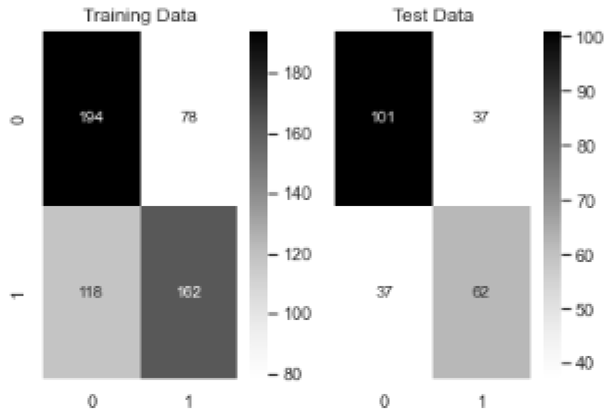

```
Accuracy score for Logistic regression train variables
0.644927536231884
```


```
Accuracy score for Logistic regression test variables

0.6877637130801688
```

Confusion Matrix Train Variables for Logistic regression

```
confusion matrix Train variables for logistic regression
```



Logistic regression Classification report

```
Logistic regression Classfication report
Classification Report of the training data:

              precision    recall  f1-score   support

          no       0.62      0.71      0.66       272
         yes       0.68      0.58      0.62       280

    accuracy                           0.64       552
   macro avg       0.65      0.65      0.64       552
weighted avg       0.65      0.64      0.64       552


Classification Report of the test data:

              precision    recall  f1-score   support

          no       0.73      0.73      0.73       138
         yes       0.63      0.63      0.63        99

    accuracy                           0.69       237
   macro avg       0.68      0.68      0.68       237
weighted avg       0.69      0.69      0.69       237
```

```
AUC and ROC FOR Logistic regression
AUC for the Training Data: 0.701
AUC for the Test Data: 0.763
```

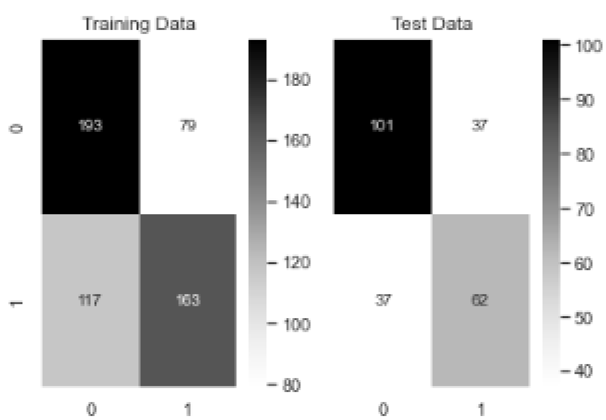
ROC curve for the model


```
Accuracy score for LDA test variables

0.6877637130801688
```

Confusion matrix train variables for LDA

```
confusion matrix Train variables for LDA
```



LDA classification report

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

              precision    recall  f1-score   support

          no       0.62      0.71      0.66       272
         yes       0.67      0.58      0.62       280

    accuracy                           0.64       552
   macro avg       0.65      0.65      0.64       552
weighted avg       0.65      0.64      0.64       552

Classification Report of the test data:

              precision    recall  f1-score   support

          no       0.73      0.73      0.73       138
         yes       0.63      0.63      0.63        99

    accuracy                           0.69       237
   macro avg       0.68      0.68      0.68       237
weighted avg       0.69      0.69      0.69       237
```

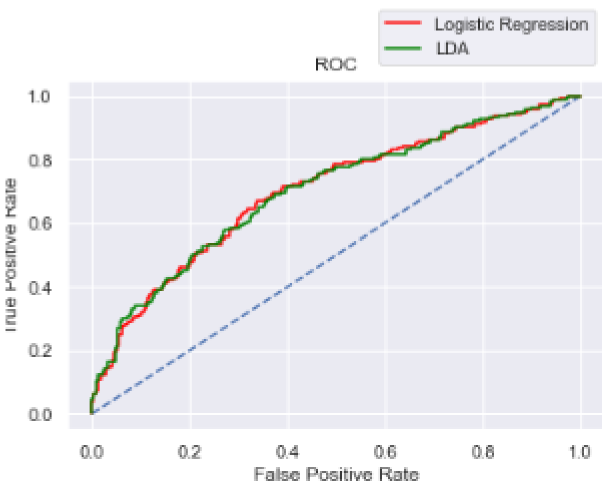
```
AUC and ROC FOR LDA
AUC for the Training Data: 0.700
AUC for the Test Data: 0.767
```

Roc curve for the model

	precision	recall	f1-score	support
no	0.621795	0.713235	0.664384	272.000000
yes	0.675000	0.578571	0.623077	280.000000
accuracy	0.644928	0.644928	0.644928	0.644928
macro avg	0.648397	0.645903	0.643730	552.000000
weighted avg	0.648783	0.644928	0.643431	552.000000

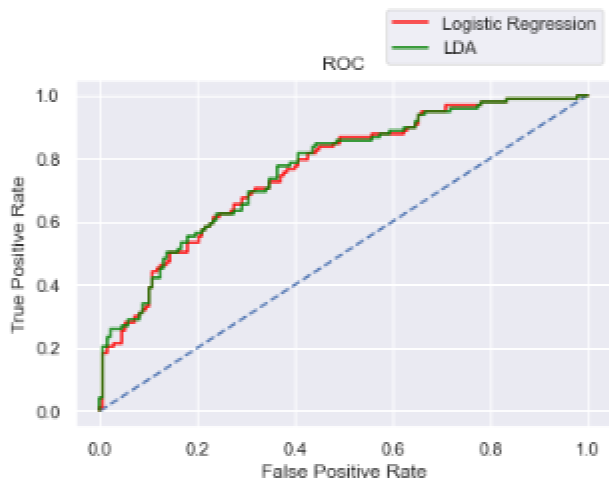
	precision	recall	f1-score	support
no	0.731884	0.731884	0.731884	138.000000
yes	0.626263	0.626263	0.626263	99.000000
accuracy	0.687764	0.687764	0.687764	0.687764
macro avg	0.679073	0.679073	0.679073	237.000000
weighted avg	0.687764	0.687764	0.687764	237.000000

	Logistic reg Train	Logistic reg Test	LDA Train	LDA Test
Accuracy	0.64	0.69	0.64	0.69
AUC	0.70	0.76	0.70	0.77
Recall	0.58	0.63	0.58	0.63
Precision	0.68	0.63	0.67	0.63
F1 Score	0.62	0.63	0.62	0.63



```
ROC curve for Test data
```

```
<matplotlib.legend.Legend at 0x276b7389dc0>
```



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

Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

We had a business problem where we need to predict whether an employee would opt for a holiday package or not.

For this problem we had done predictions both using logistic regression and linear discriminant analysis. Since both the techniques are giving the same results.

The EDA analysis clearly indicates certain criteria that where we could find people aged above 50 are less interested in holiday packages.

This is one of the observations we found that the aged people are less interested in holiday packages.

People ranging from the age 30 to 50 generally opt for holiday packages.

Employee age between 50 and 60 usually less interested to opt a holiday package, whereas employee aged 30 to 50 and salary less than 50000 people are comparatively opt more holiday packages.

The important factors deciding the predictions are salary, age and education

Recommendations :

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

2. For people earning more than 150000 can be provided vacation holiday packages.

3. For employee having more number of older children can be provided with packages in holiday vacation places.