BUSINESS REPORT

ON

PREDICTIVE MODELLING

By Kshitij Nishant

Table of Contents

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.

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.

Answer:

After reading the data,

	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
0	1	0.30	Ideal	E	SI1	62.1	58.0	4.27	4.29	2.66	499
1	2	0.33	Premium	G	IF	60.8	58.0	4.42	4.46	2.70	984
2	3	0.90	Very Good	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289
3	4	0.42	Ideal	F	VS1	61.6	56.0	4.82	4.80	2.96	1082
4	5	0.31	Ideal	F	VVS1	60.4	59.0	4.35	4.43	2.65	779
5	6	1.02	Ideal	D	VS2	61.5	56.0	6.46	6.49	3.99	9502
6	7	1.01	Good	H	SI1	63.7	60.0	6.35	6.30	4.03	4836
7	8	0.50	Premium	E	SI1	61.5	62.0	5.09	5.06	3.12	1415
8	9	1.21	Good	H	SI1	63.8	64.0	6.72	6.63	4.26	5407
9	10	0.35	Ideal	F	VS2	60.5	57.0	4.52	4.60	2.76	706

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.

Description of data:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
carat	26967.0	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	26270.0	NaN	NaN	NaN	61.745147	1.41286	50.8	61.0	61.8	62.5	73.6
table	26967.0	NaN	NaN	NaN	57.45608	2.232068	49.0	56.0	57.0	59.0	79.0
x	26967.0	NaN	NaN	NaN	5.729854	1.128516	0.0	4.71	5.69	6.55	10.23
y	26967.0	NaN	NaN	NaN	5.733569	1.166058	0.0	4.71	5.71	6.54	58.9
z	26967.0	NaN	NaN	NaN	3.538057	0.720624	0.0	2.9	3.52	4.04	31.8
price	26967.0	NaN	NaN	NaN	3939.518115	4024.864666	326.0	945.0	2375.0	5360.0	18818.0

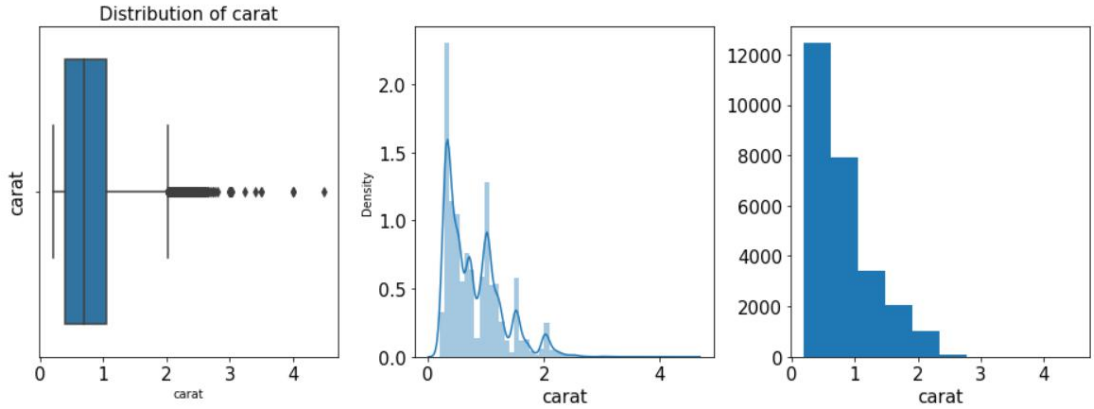
Exploratory Data Analysis

A. Univariate analysis

Continuous Variables

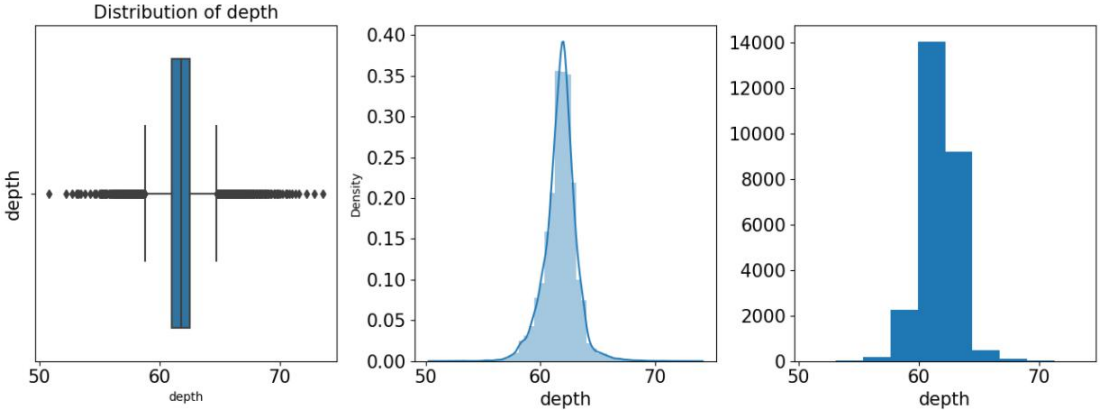
1) Carat

Range of values: 4.3
Minimum carat: -1.2525215828139256
Maximum carat: 3.0
Mean value: -0.004145716416463627
Median value: -0.2059197863853835
Standard deviation: 0.9844981513847829
Null values: False



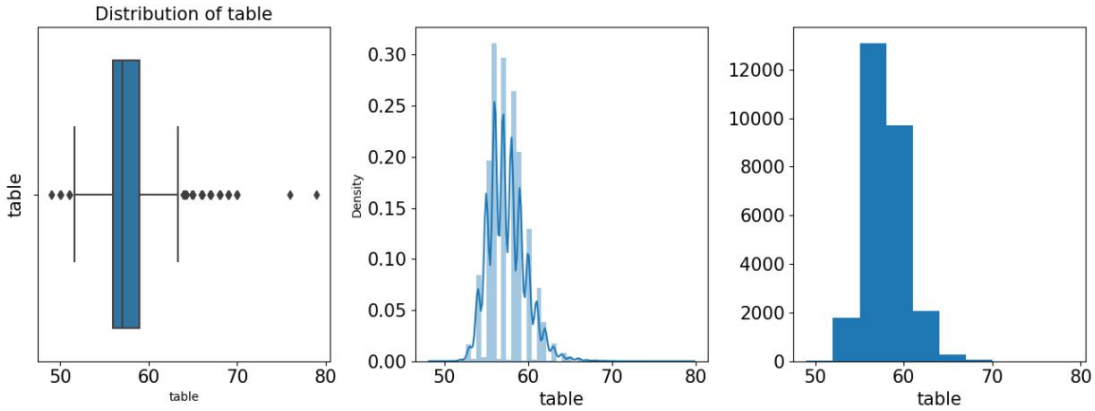
2) Depth

Range of values: 22.799999999999997
Minimum depth: 50.8
Maximum depth: 73.6
Mean value: 61.745146555006194
Median value: 61.8
Standard deviation: 1.4128602381425932
Null values: True



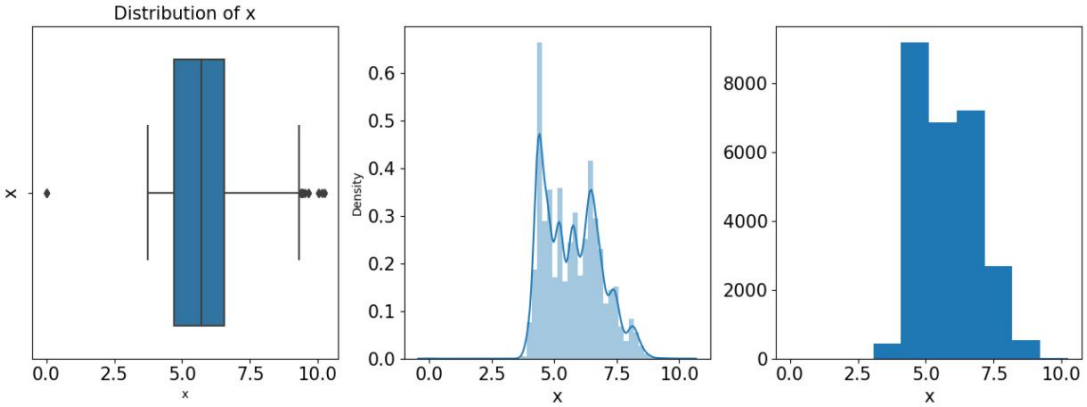
3) Table

Range of values: 30.0
Minimum table: 49.0
Maximum table: 79.0
Mean value: 57.45607965290908
Median value: 57.0
Standard deviation: 2.2320679090295075
Null values: False



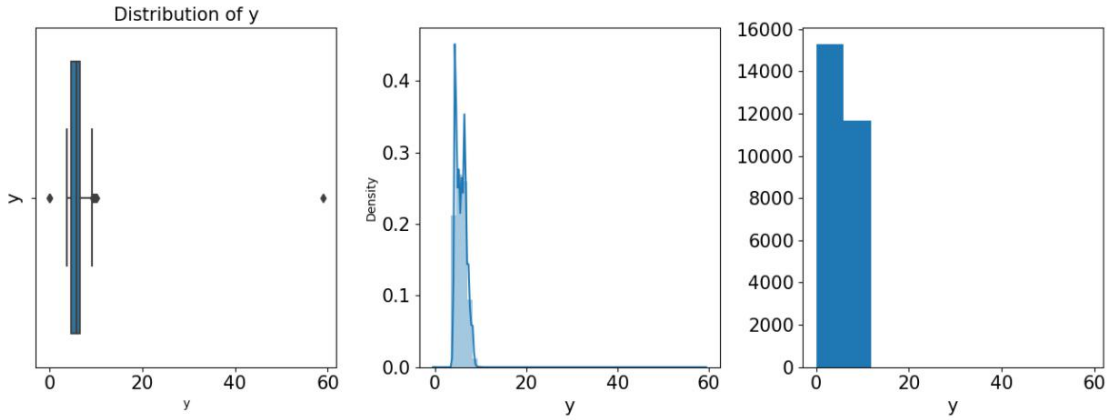
4) X

Range of values: 10.23
Minimum x: 0.0
Maximum x: 10.23
Mean value: 5.729853524678309
Median value: 5.69
Standard deviation: 1.1285163776477767
Null values: False



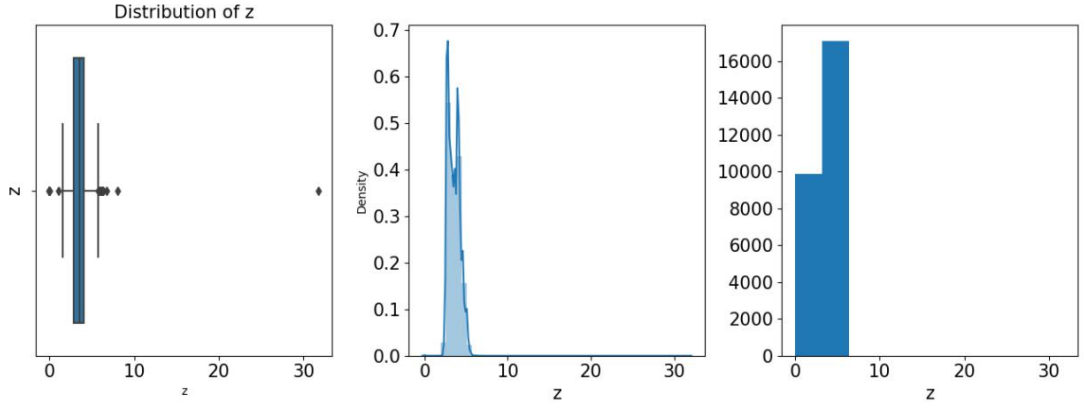
5) Y

Range of values: 58.9
Minimum y: 0.0
Maximum y: 58.9
Mean value: 5.733568806318799
Median value: 5.71
Standard deviation: 1.1660575299260496
Null values: False



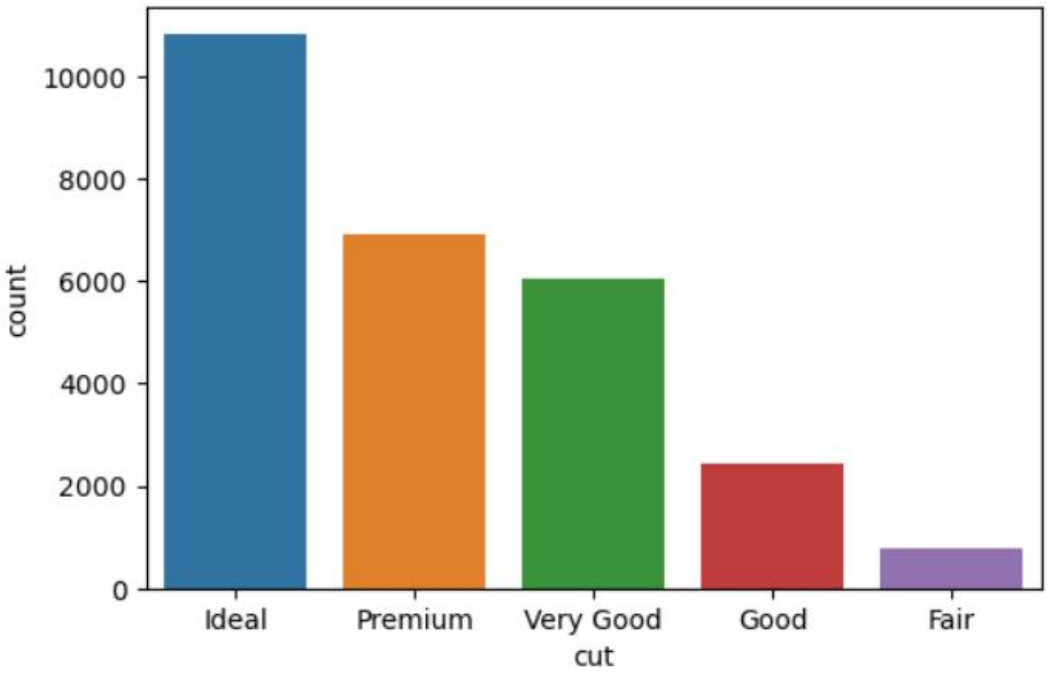
6) Z

Range of values: 31.8
Minimum z: 0.0
Maximum z: 31.8
Mean value: 3.5380572551637184
Median value: 3.52
Standard deviation: 0.7206236256427411
Null values: False



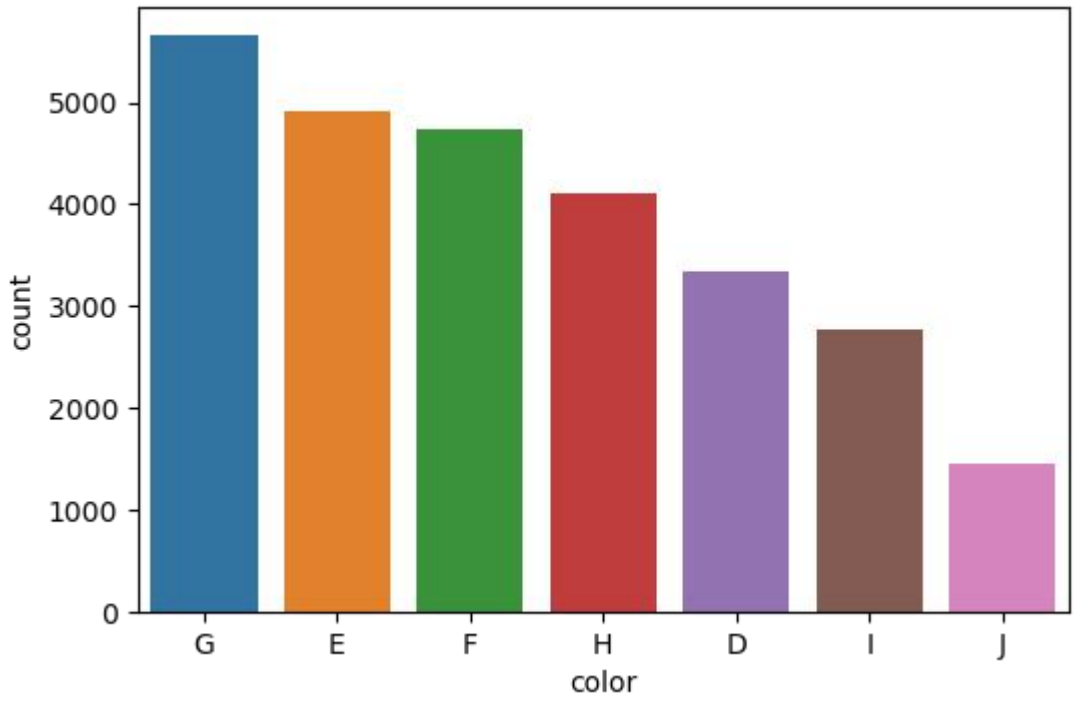
Categorical Variables

1) CUT



No. of Ideal zirconia has most numbers

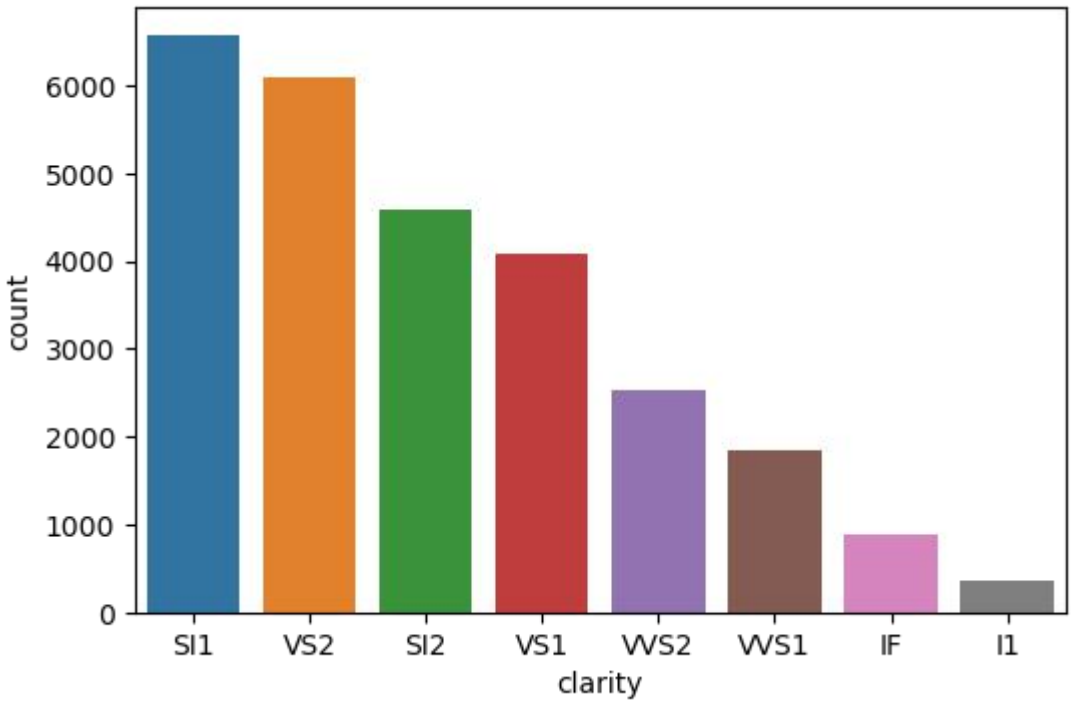
2) Color



G color zirconia is most in abundance. Followed by E & F respectively.

With D being the best and J the worst.

3) Clarity

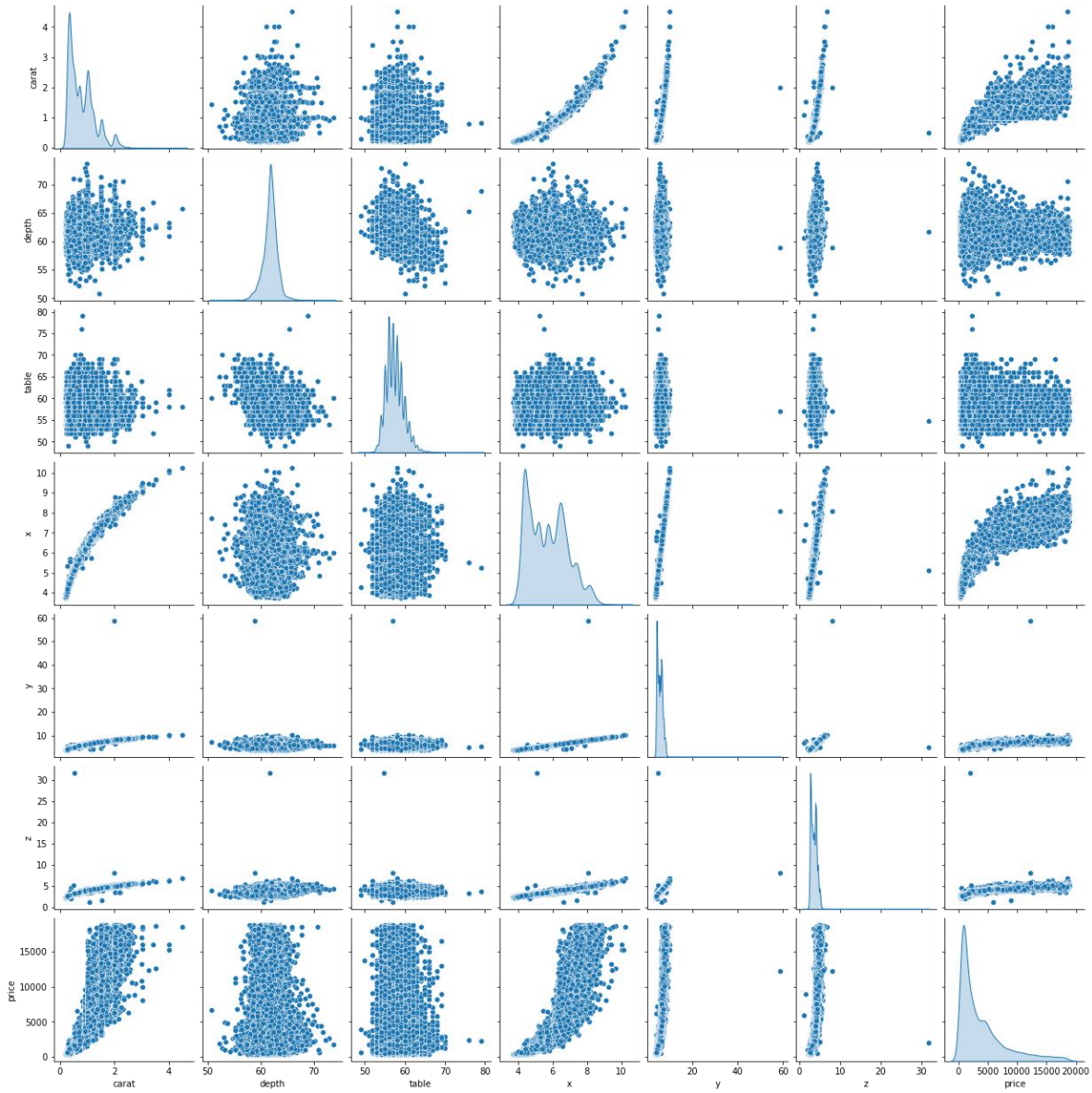


Zirconia with Clarity corresponding to SI1 is most in number. Followed by VS2 and SI2.

Clarity cubic zirconia Clarity refers to the absence of the Inclusions.

In order from Best to Worst, FL = flawless, I3= level 3 inclusions: FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, I2, I3

B. MULTIVARIATE ANALYSIS

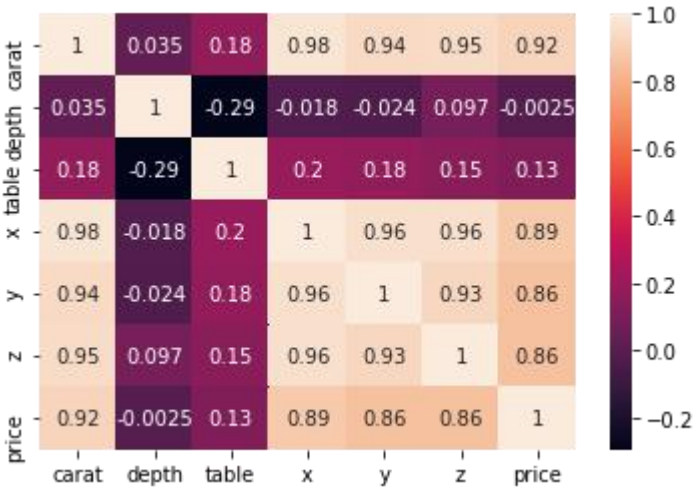


Point to be noted:

1.There's a high correlation between independent variables such as x,y,z . This may create multicollinearity problem.

2.There's a high linear correlation between dependent variables and independent variables. Except for depth and table.

3. Variables such as depth , table , x are normally distributed. Other variables have skewness in it.



1)By looking at heatmap, we can further confirm the inferences that we made above. Correlations between DV price and IV, carat,x,y,z are high.

2) Few of the above mentioned IV have high correlations between them.This may create a multicollinearity issue while creating Model.

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 a ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning.

Answer:

Number of zero values of each variable,

Number of zero values for the carat is 0
Number of zero values for the cut is 0
Number of zero values for the color is 0
Number of zero values for the clarity is 0
Number of zero values for the depth is 0
Number of zero values for the table is 0
Number of zero values for the x is 3
Number of zero values for the y is 3
Number of zero values for the z is 9
Number of zero values for the price is 0

There are records which have zeroes in it. It's an error value and has to be treated as well. As x,y,z cannot be zero for any zirconia as it defines length,breadth,height respectively.

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

There are records for column z which has zeroes in it. This values can be accepted as these records have x,y values which shows that these zirconia might be a thin sheet, hence height field is mentioned

as zero. But for the sake of calculation purpose we will going to impute all the x,y,z columns with their respective mean values,

	carat	cut	color	clarity	depth	table	x	y	z	price
5821	0.71	Good	F	SI2	64.1	60.0	0.00	0.00	0.0	2130
6034	2.02	Premium	H	VS2	62.7	53.0	8.02	7.95	0.0	18207
6215	0.71	Good	F	SI2	64.1	60.0	0.00	0.00	0.0	2130
10827	2.20	Premium	H	SI1	61.2	59.0	8.42	8.37	0.0	17265
12498	2.18	Premium	H	SI2	59.4	61.0	8.49	8.45	0.0	12631
12689	1.10	Premium	G	SI2	63.0	59.0	6.50	6.47	0.0	3696
17506	1.14	Fair	G	VS1	57.5	67.0	0.00	0.00	0.0	6381
18194	1.01	Premium	H	I1	58.1	59.0	6.66	6.60	0.0	3167
23758	1.12	Premium	G	I1	60.4	59.0	6.71	6.67	0.0	2383

Changing datatype from Object to numeric and assigining a codes for each category:

Column Name: cut
['Ideal', 'Premium', 'Very Good', 'Good', 'Fair']
Categories (5, object): ['Fair', 'Good', 'Ideal', 'Premium', 'Very Good']
[2 3 4 1 0]

Column Name: color
['E', 'G', 'F', 'D', 'H', 'J', 'I']
Categories (7, object): ['D', 'E', 'F', 'G', 'H', 'I', 'J']
[1 3 2 0 4 6 5]

Column Name: clarity
['SI1', 'IF', 'VVS2', 'VS1', 'VVS1', 'VS2', 'SI2', 'I1']
Categories (8, object): ['I1', 'IF', 'SI1', 'SI2', 'VS1', 'VS2', 'VVS1', 'VVS2']
[2 1 7 4 6 5 3 0]

```
<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  int8   
 2   color    26967 non-null  int8   
 3   clarity  26967 non-null  int8   
 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), int8(3)
memory usage: 1.5 MB
```

1.3 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.

Answer:

The coefficient for carat is 1.4836192853139354
The coefficient for cut is 0.009047385078917823
The coefficient for color is -0.06850690194834504
The coefficient for clarity is 0.06967514363446227
The coefficient for depth is -0.03456802492987976
The coefficient for table is -0.048497123303384376
The coefficient for x is -0.7301350656346718
The coefficient for y is 0.4466943355168739
The coefficient for z is -0.19667738759196302

Intercept : -0.11158015121069696

Accuracy score of Training Data: 90.23%

Accuracy score of Test Data: 90.17%

Overall accuracy score of Training and Testing Data is similar.

After applying Stats Model,

OLS Regression Results

Dep. Variable:	price	R-squared:	0.902
Model:	OLS	Adj. R-squared:	0.902
Method:	Least Squares	F-statistic:	1.937e+04
Date:	Mon, 24 Jan 2022	Prob (F-statistic):	0.00
Time:	00:05:01	Log-Likelihood:	-4444.6
No. Observations:	18876	AIC:	8909.
Df Residuals:	18866	BIC:	8988.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.1116	0.009	-12.867	0.000	-0.129	-0.095
carat	1.4836	0.012	121.397	0.000	1.460	1.508
cut	0.0090	0.002	4.043	0.000	0.005	0.013
color	-0.0685	0.001	-49.987	0.000	-0.071	-0.066
clarity	0.0697	0.001	52.155	0.000	0.067	0.072
depth	-0.0346	0.005	-7.503	0.000	-0.044	-0.026
table	-0.0485	0.002	-19.506	0.000	-0.053	-0.044
x	-0.7301	0.043	-17.166	0.000	-0.814	-0.647
y	0.4467	0.044	10.234	0.000	0.361	0.532
z	-0.1967	0.032	-6.220	0.000	-0.259	-0.135

Omnibus:	5232.485	Durbin-Watson:	1.968
Prob(Omnibus):	0.000	Jarque-Bera (JB):	51074.776
Skew:	1.051	Prob(JB):	0.00
Kurtosis:	10.780	Cond. No.	142.

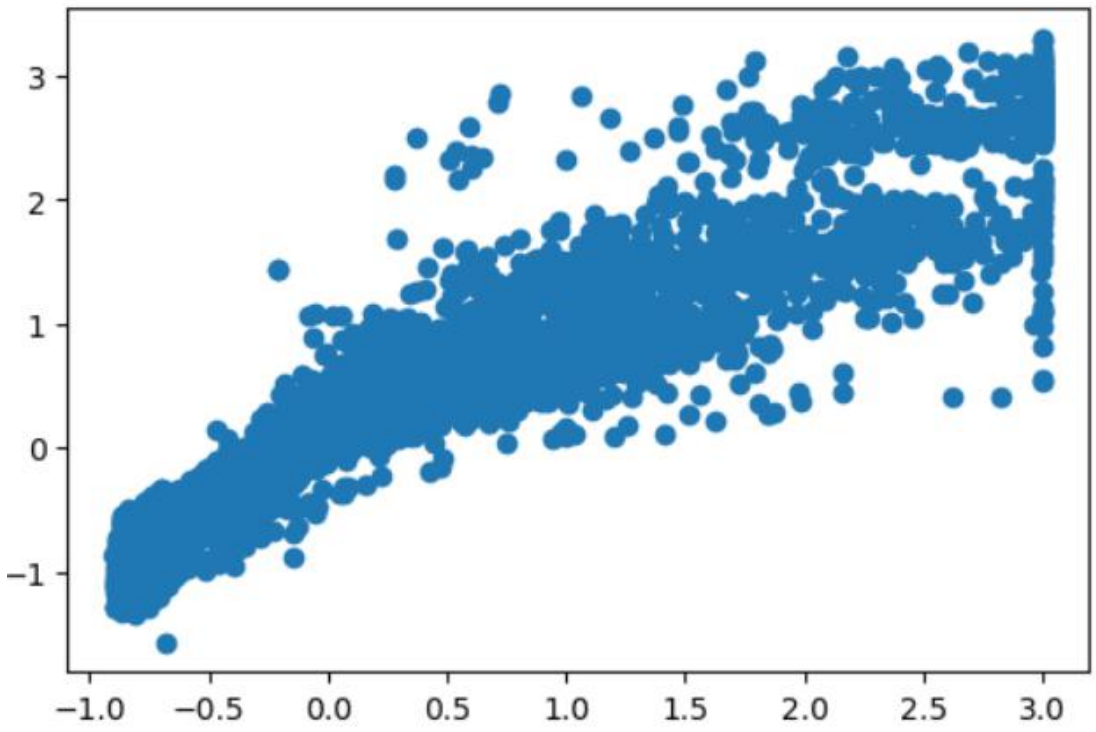
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Here are the following observations :-
1) R-square and Adjusted R-square is giving us the good results.
2) P values for all the variables are within significance level.

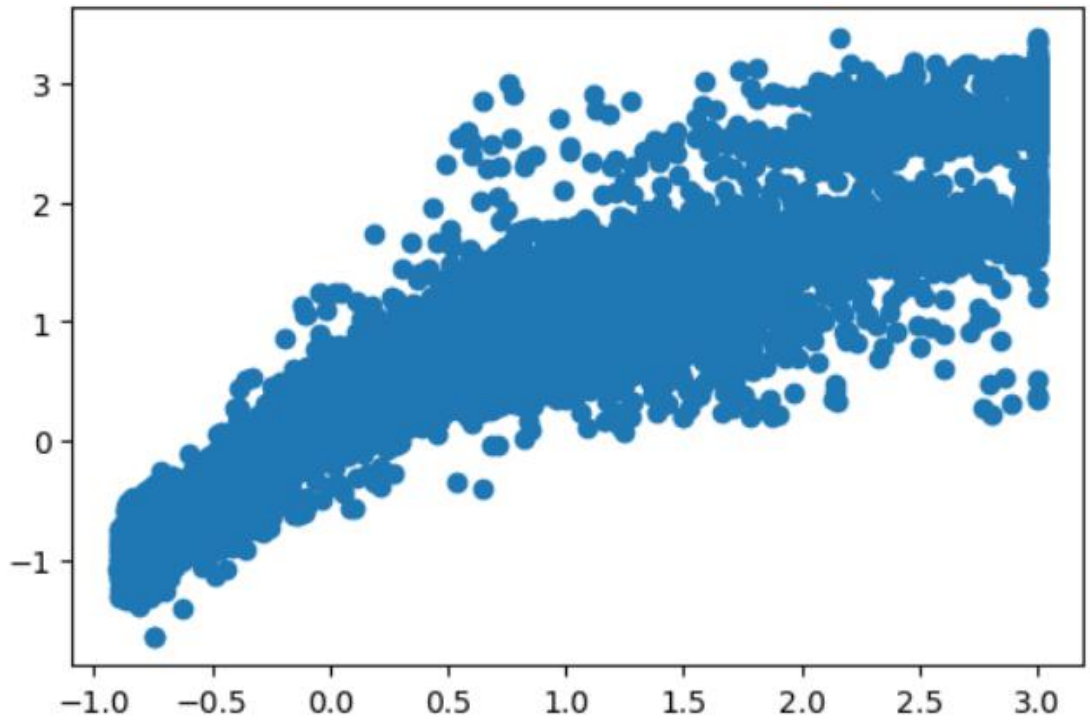
MSE for training is 1.4176336778420113e-33 and testing is 0.09170851041562096.

Root mean squared error for training is 3.765147643641629e-17 and Testing is 0.3028341302026919.

Scatter Plot for Test Data,



Scatter Plot for Train data,



As we can observe there is multicollinearity present between the data. VIF should be less than 5 for each variables.

But its very high x,y,z & carat variable. That was evident as well from heatmap.

We will try to create the model once again using only carat variable and ignoring x,y,z variables.

carat : 28.79898918828712
cut : 4.3818240632570875
color : 3.11001334505159
clarity : 4.295107057323167
depth : 4.55893169017134
table : 1.1822579286587427
x : 409.0838121520042
y : 402.46192440827264
z : 241.36801201458135

After recreating the formula by ignoring x,y,z variables. And calculating different parameters,

OLS Regression Results

Dep. Variable:	price	R-squared:	0.893
Model:	OLS	Adj. R-squared:	0.893
Method:	Least Squares	F-statistic:	2.624e+04
Date:	Mon, 24 Jan 2022	Prob (F-statistic):	0.00
Time:	00:05:02	Log-Likelihood:	-5310.2
No. Observations:	18876	AIC:	1.063e+04
Df Residuals:	18869	BIC:	1.069e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.1594	0.009	-17.749	0.000	-0.177	-0.142
carat	0.9951	0.003	386.151	0.000	0.990	1.000
cut	0.0146	0.002	6.274	0.000	0.010	0.019
color	-0.0656	0.001	-45.771	0.000	-0.068	-0.063
clarity	0.0758	0.001	54.518	0.000	0.073	0.079
depth	-0.0340	0.003	-13.018	0.000	-0.039	-0.029
table	-0.0525	0.003	-20.370	0.000	-0.058	-0.047

Omnibus:	4528.234	Durbin-Watson:	1.972
Prob(Omnibus):	0.000	Jarque-Bera (JB):	24227.124
Skew:	1.053	Prob(JB):	0.00
Kurtosis:	8.135	Cond. No.	22.0

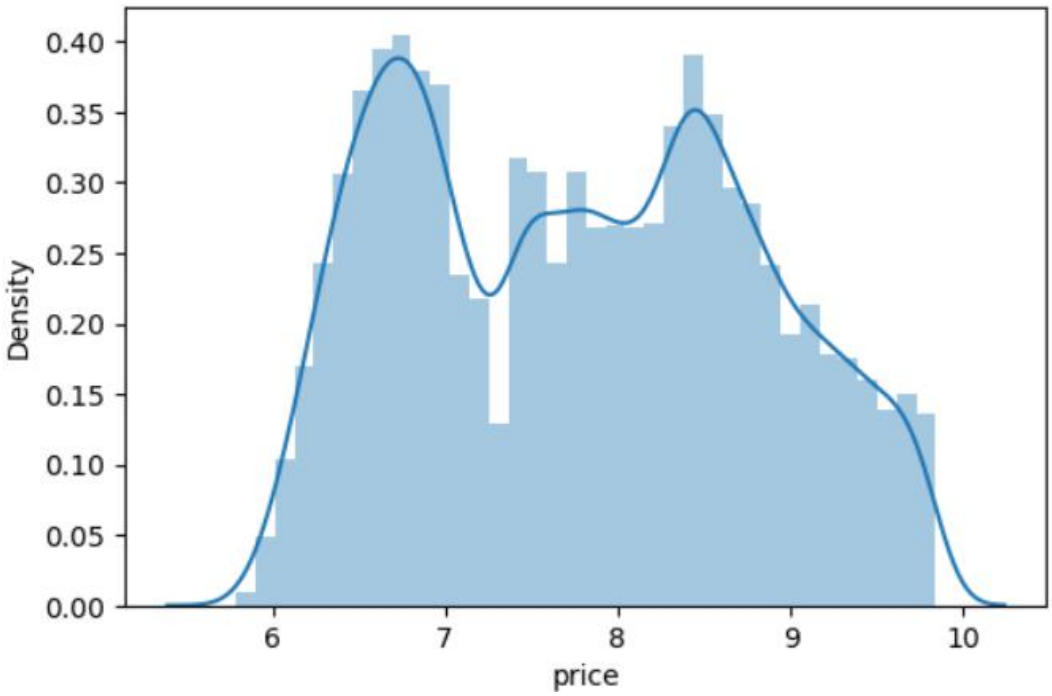
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

We can see there's minor changes in R squared value. Which means we can make model by using these variables as well.

We can we see that all the VIF values are within 5 which shows there's no collinearity present between dependent variables,

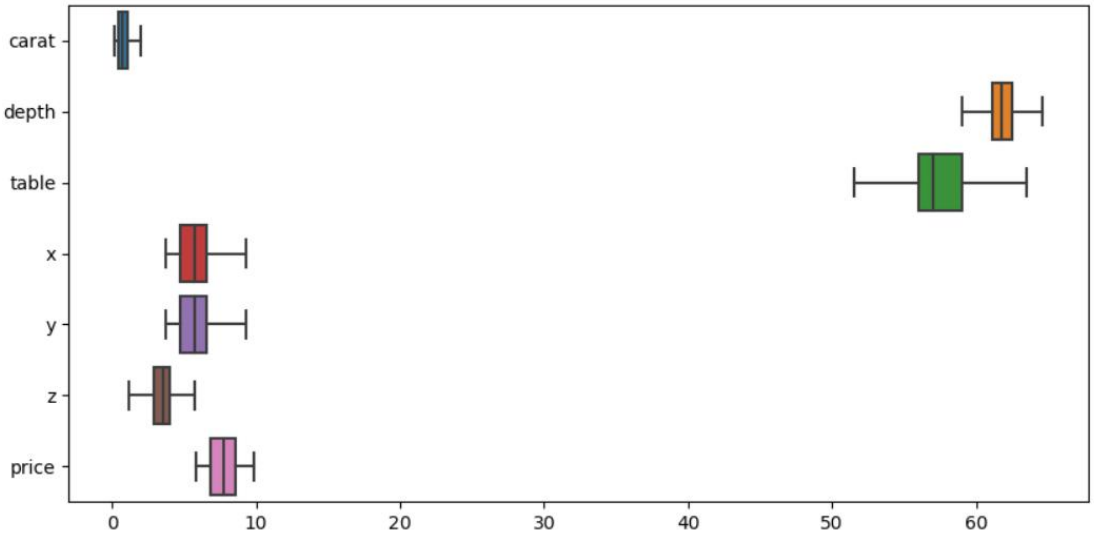
carat : 1.1868161549194873
cut : 4.3541864426985475
color : 3.1021684502722158
clarity : 4.2823030371685915
depth : 1.119765909273898
table : 1.1614370830194474

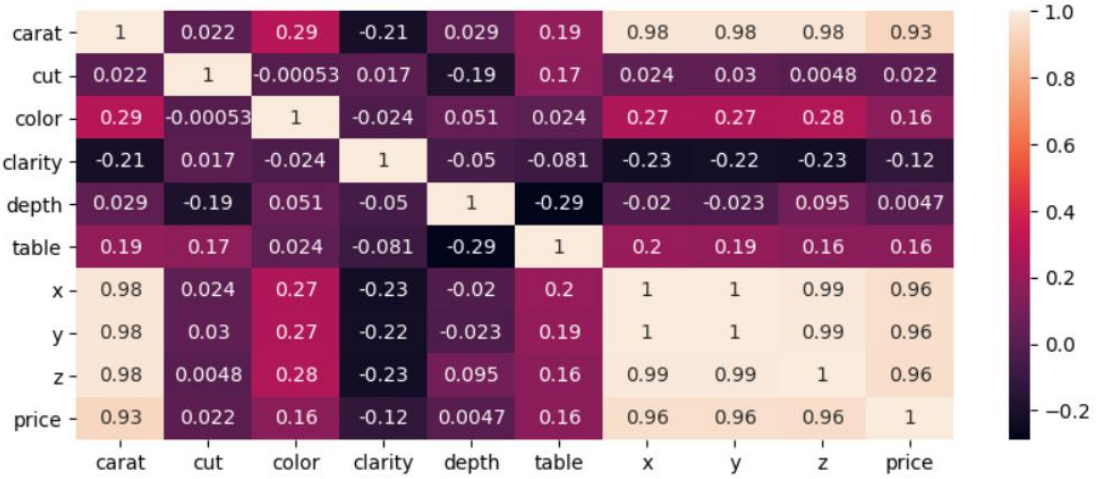
I made a few more models and to see if we can achieve higher accuracy or not. By changing outlier treatment technique and doing some feature engineering,



It’s somewhat normally distributed.

This time I treated the outliers before making the model.





Multicollinearity is exists.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.947
Model:	OLS	Adj. R-squared:	0.947
Method:	Least Squares	F-statistic:	1.127e+05
Date:	Mon, 24 Jan 2022	Prob (F-statistic):	0.00
Time:	00:05:05	Log-Likelihood:	635.06
No. Observations:	18853	AIC:	-1262.
Df Residuals:	18849	BIC:	-1231.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.4524	0.011	230.799	0.000	2.432	2.473
color	-0.0706	0.001	-67.667	0.000	-0.073	-0.069
clarity	0.0620	0.001	61.020	0.000	0.060	0.064
x	0.9206	0.002	569.216	0.000	0.917	0.924

Omnibus:	777.729	Durbin-Watson:	1.995
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2550.538
Skew:	0.024	Prob(JB):	0.00
Kurtosis:	4.801	Cond. No.	47.7

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

With various permutation and combination I got this results with gives us the best accuracy along with the least VIF scores.
Making model mentioning all the features gave us the result of 95.1 percent but gave very high VIF value as well.
Hence, chose above features to get the most optimum results.

VIF is also not very high,

color : 3.654174232574307
clarity : 4.340332635216736
x : 6.693204279345197

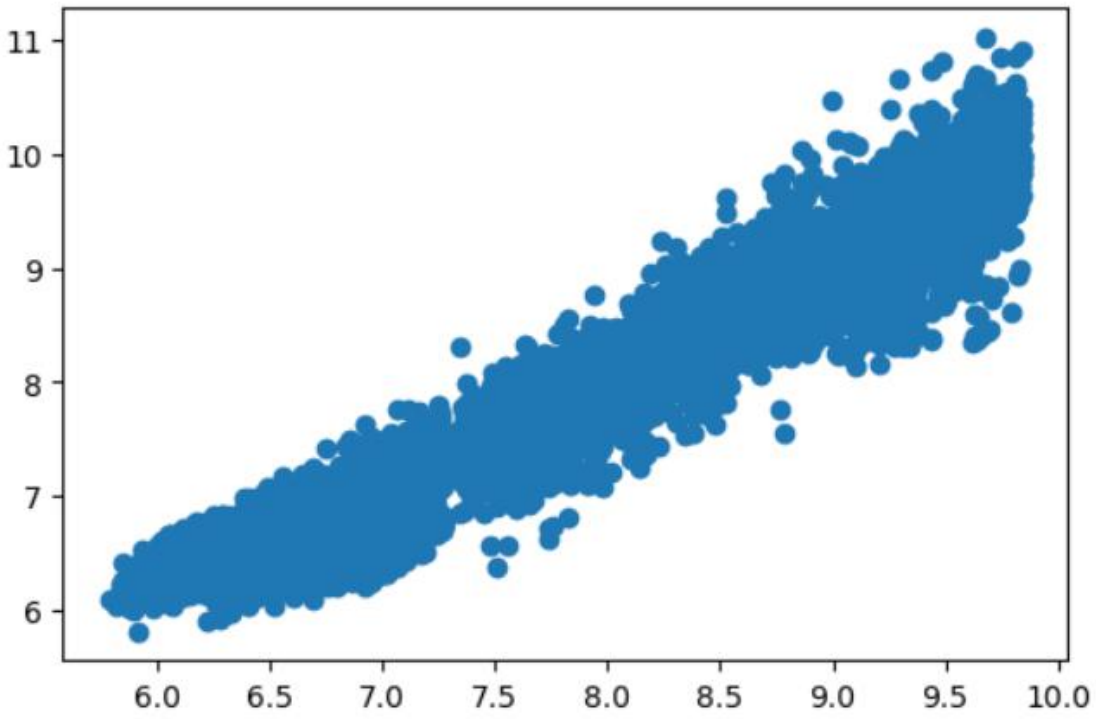
Training set accuracy: 94.72%

Testing set accuracy: 94.68%

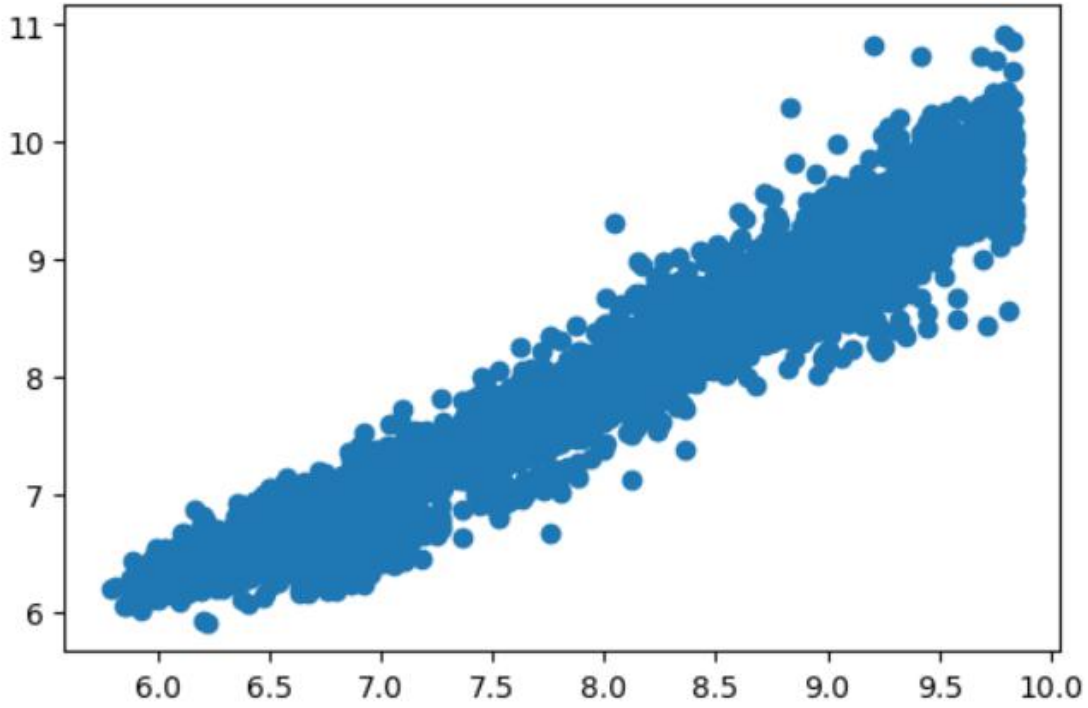
Root mean squared error for Test set is 0.23573060618431754

Root mean squared error for Training set is 2.5157124920164795e-16

Scatter plot for training set:



Scatter plot for testing set:



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

Answer:

The linear equation that is formed using coefficients and intercept is :-
The coefficient for color is -0.07063375156153978

The coefficient for clarity is 0.0619693025627653

The coefficient for x is 0.9206095882387271

The intercept is 2.45243

log (Price) = -0.0706 color + 0.0619clarity + 0.9206*x + 2.4524
math.exp(Price) can be used to get the actual price

Inference
1.The linear regression equation that we form is log (Price) = -0.0706 color + 0.0619clarity + 0.9206x + 2.4524 or Price =e^ (-0.0706 color + 0.0619clarity + 0.9206x + 2.4524).

2.Looking into the equation, we can say that the x variable which is the length of the zirconia is the most important factor for deciding the price.

3.As there's a good positive correlation between carat, x, y, z we can say that all the features are of utmost important to determine price. As if any of the features get deteriorated, it will going to affect the other correlated features, resulting in effecting the price.

4.Clarity has also a major role in deciding the price of zirconia.

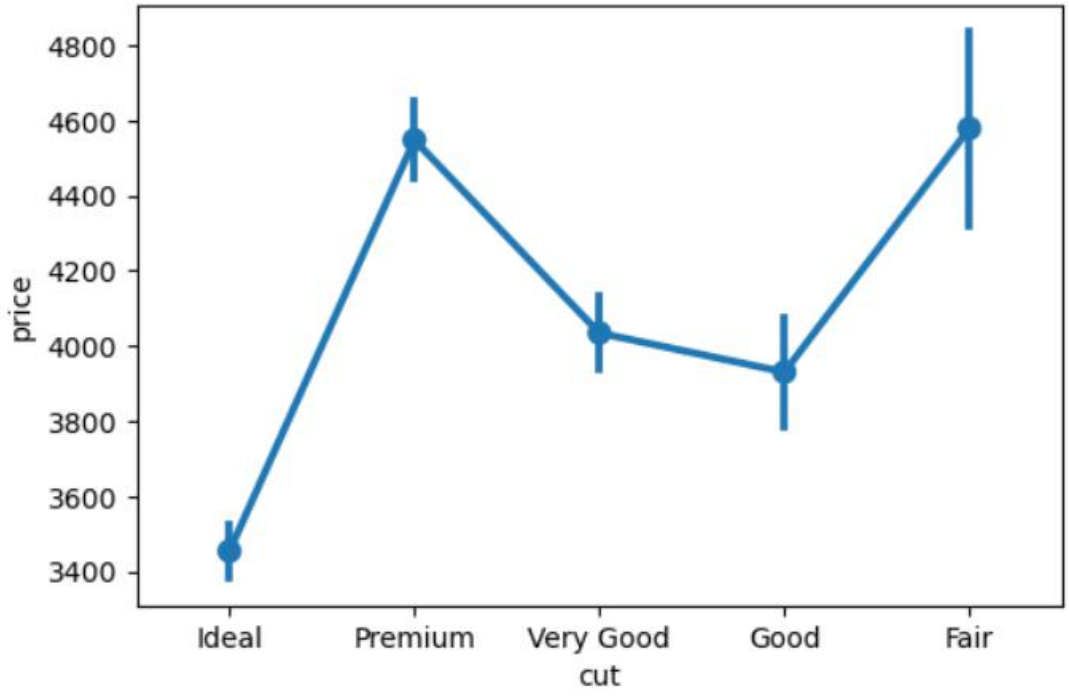
SUGGESTIONS

1.From looking at the countplot of clarity , we can see that the production of high clarity zirconia is less. There's no zirconia with FL remarks. Moreover zirconia with IF, VVS1, VVS2 labels are less.

2.Company should try to provide clearer zirconia as it directly impacts the price.

3.As the price of zirconia is hugely dependent on x & x and carat is highly correlated. The company should try to manufacture high carat zirconia to get better price.

4. The price of Premium and Fair cut is maximum. Hence we should try to maximize the manufacturing of zirconia with such cuts.



End of Business Report 1

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.

Answer:

After reading 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
5	6	yes	61590	42	12	0	1	no
6	7	no	94344	51	8	0	0	no
7	8	yes	35987	32	8	0	2	no
8	9	no	41140	39	12	0	0	no
9	10	no	35826	43	11	0	2	no

Holiday_Package :Opted for Holiday Package yes/no?
Salary: Employee salary
Age: Age in years
Educ: 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

Description of data:

	count	mean	std	min	25%	50%	75%	max
Salary	872.0	47729.172018	23418.668531	1322.0	35324.0	41903.5	53469.5	236961.0
age	872.0	39.955275	10.551675	20.0	32.0	39.0	48.0	62.0
educ	872.0	9.307339	3.036259	1.0	8.0	9.0	12.0	21.0
no_young_children	872.0	0.311927	0.612870	0.0	0.0	0.0	0.0	3.0
no_older_children	872.0	0.982798	1.086786	0.0	0.0	1.0	2.0	6.0

No Null values after check,

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

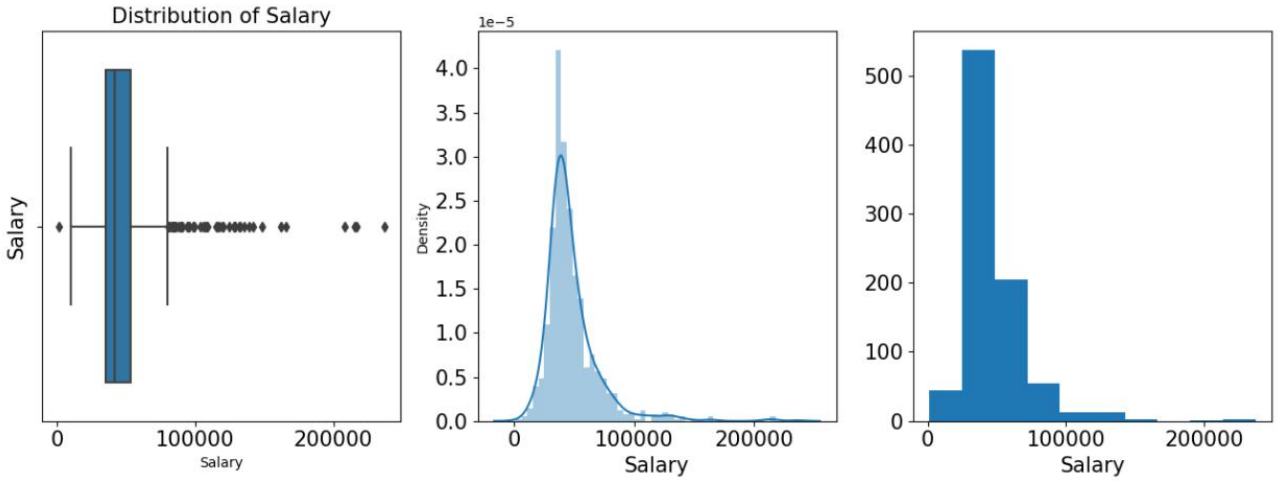
EDA

A. Univariate:

Continuous variables

1) Salary

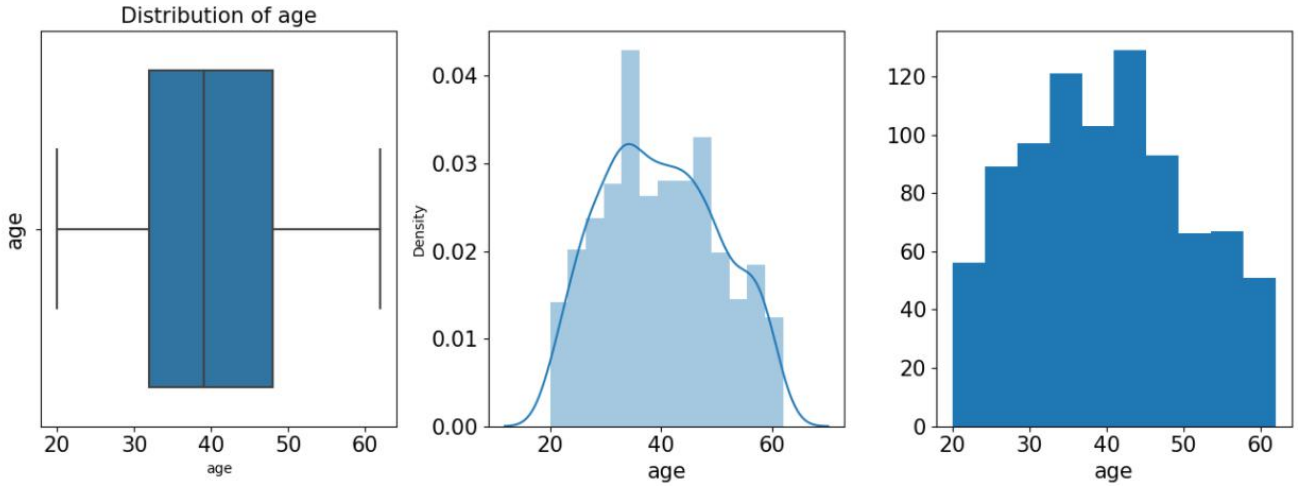
Range of values: 235639
Minimum Salary: 1322
Maximum Salary: 236961
Mean value: 47729.172018348625
Median value: 41903.5
Standard deviation: 23418.66853107387
Null values: False



2) Age

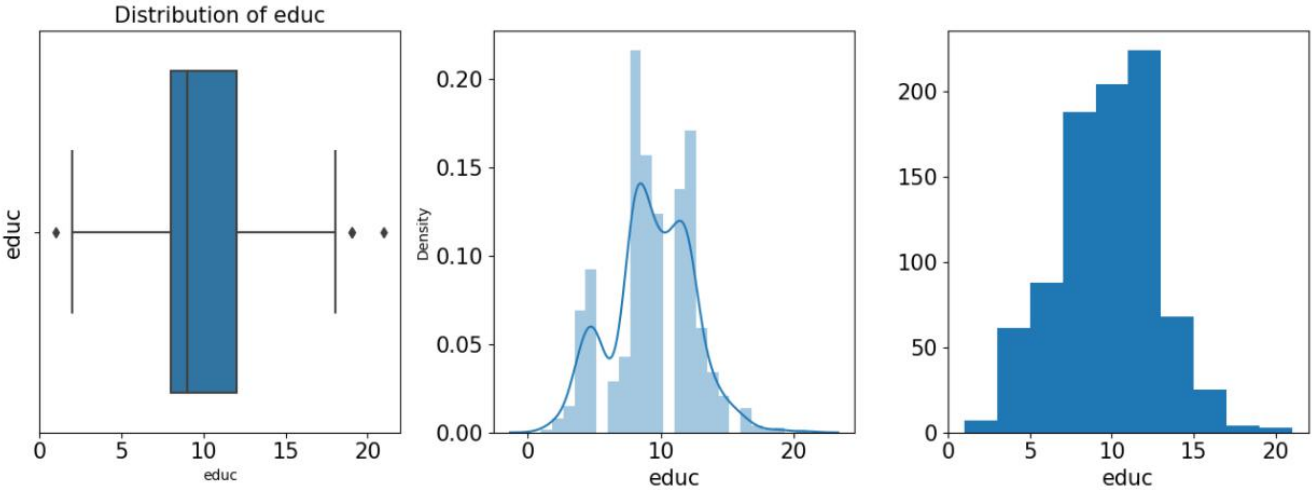
Range of values: 42

Minimum age: 20
Maximum age: 62
Mean value: 39.955275229357795
Median value: 39.0
Standard deviation: 10.551674590487607
Null values: False



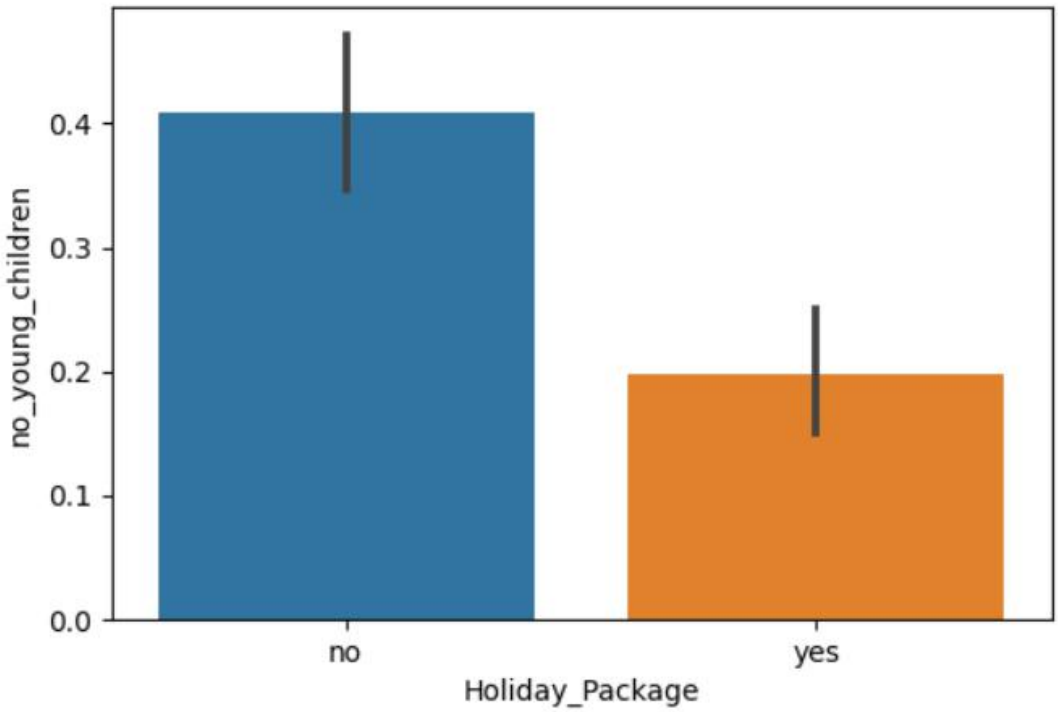
3) Education

Range of values: 20
Minimum educ: 1
Maximum educ: 21
Mean value: 9.307339449541285
Median value: 9.0
Standard deviation: 3.0362586930870448
Null values: False

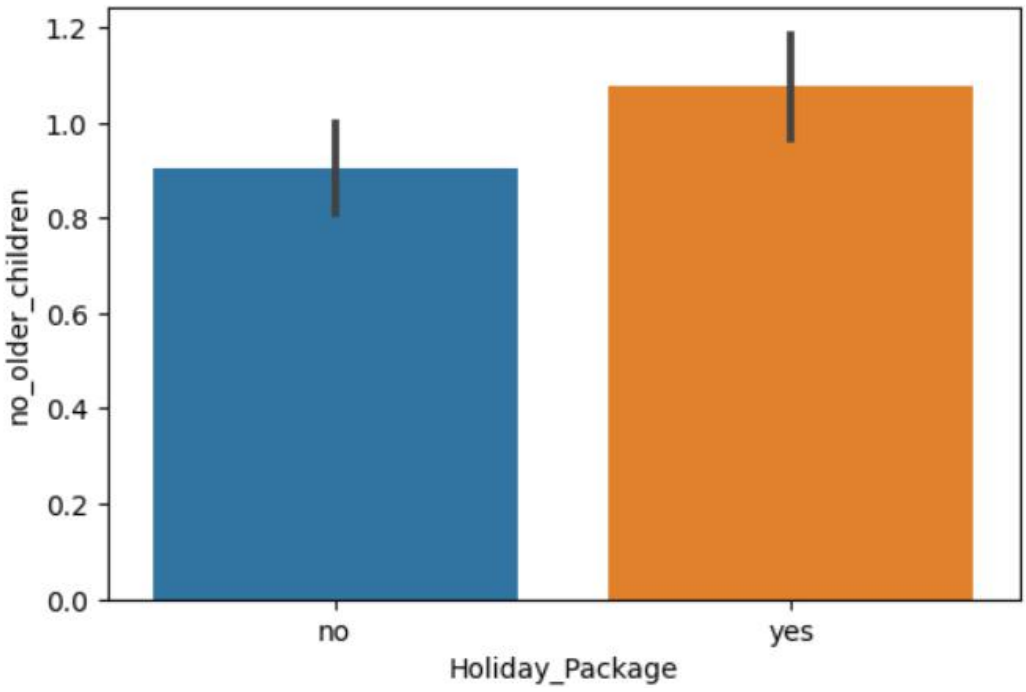


Categorical Variables

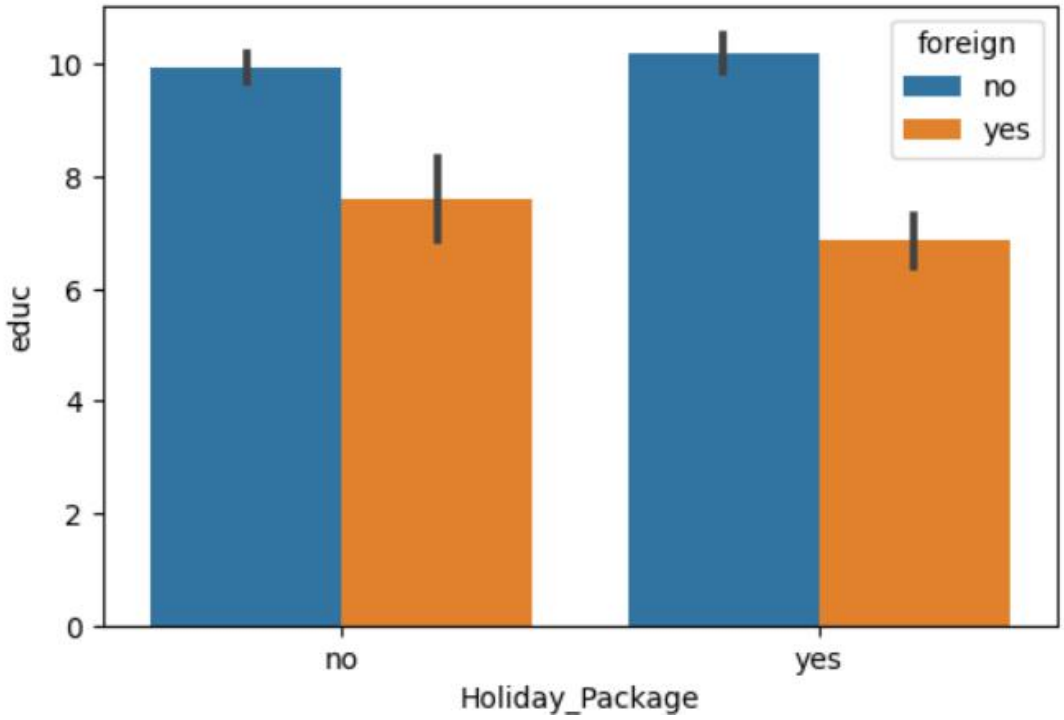
1. Bar Plots between Holiday_Package and various other variables in dataset:



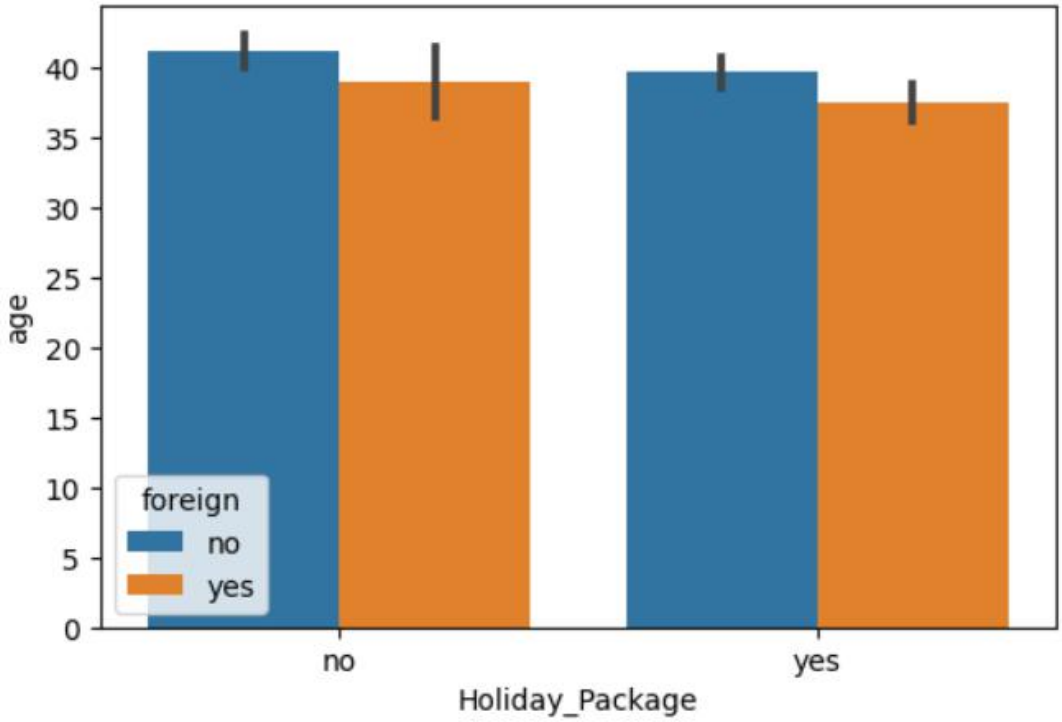
2.



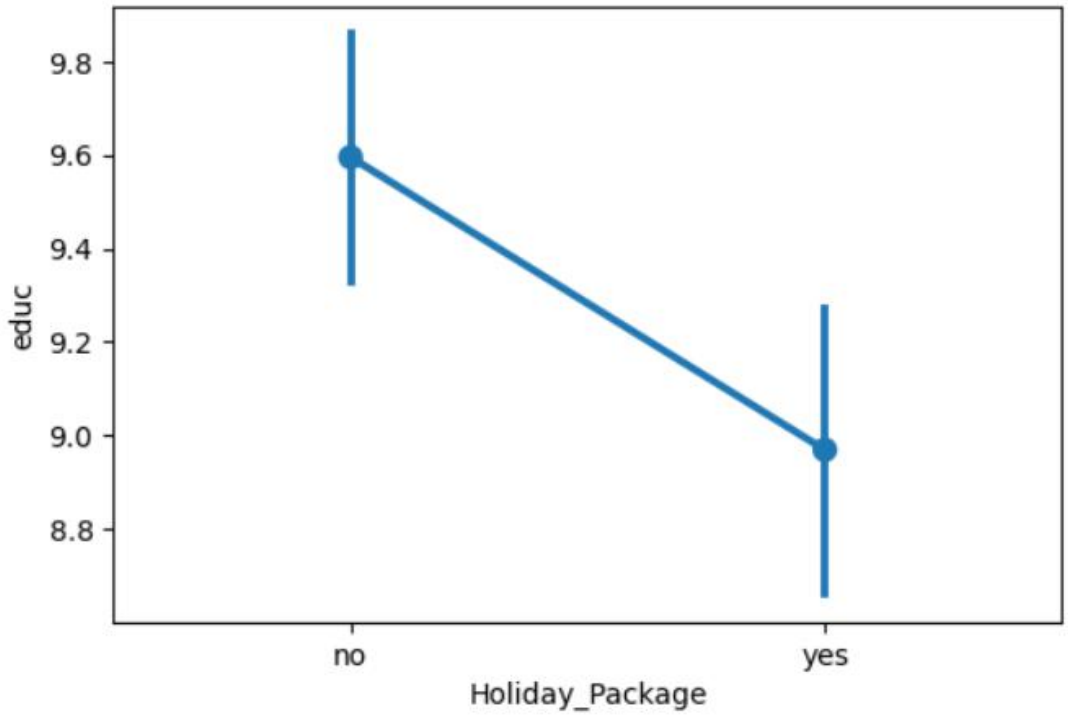
3.



4.

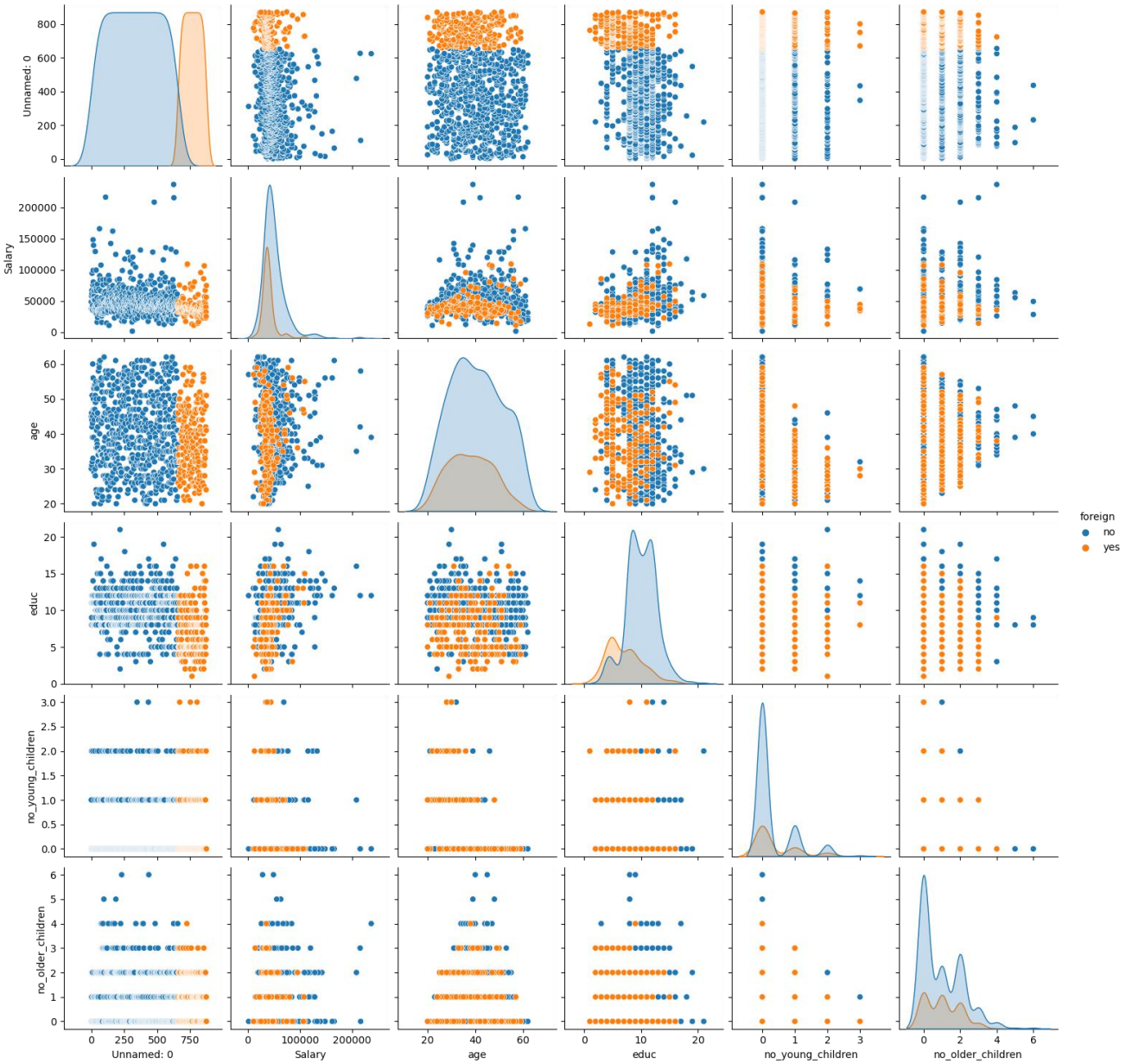


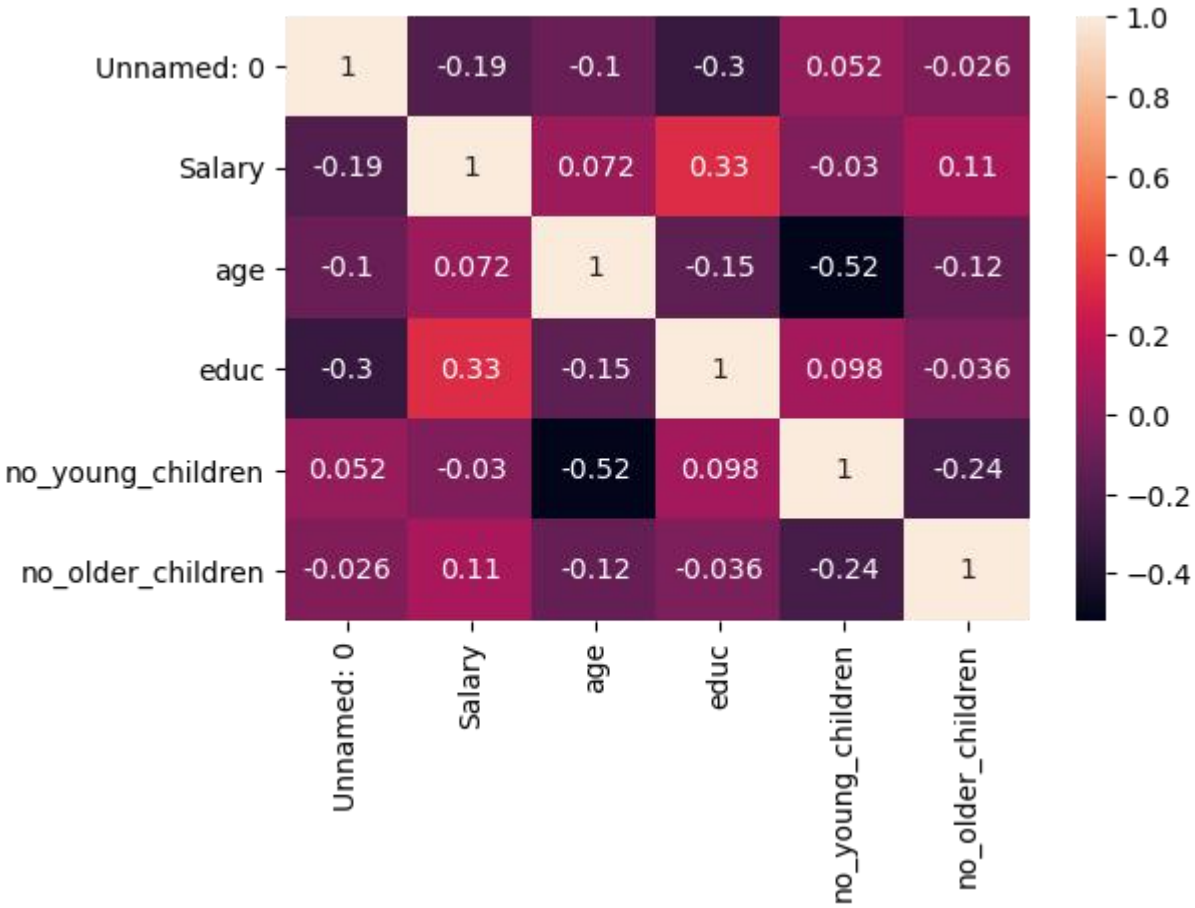
5. Pointplot between Education and Holiday_Package



B. Multivariate Analysis

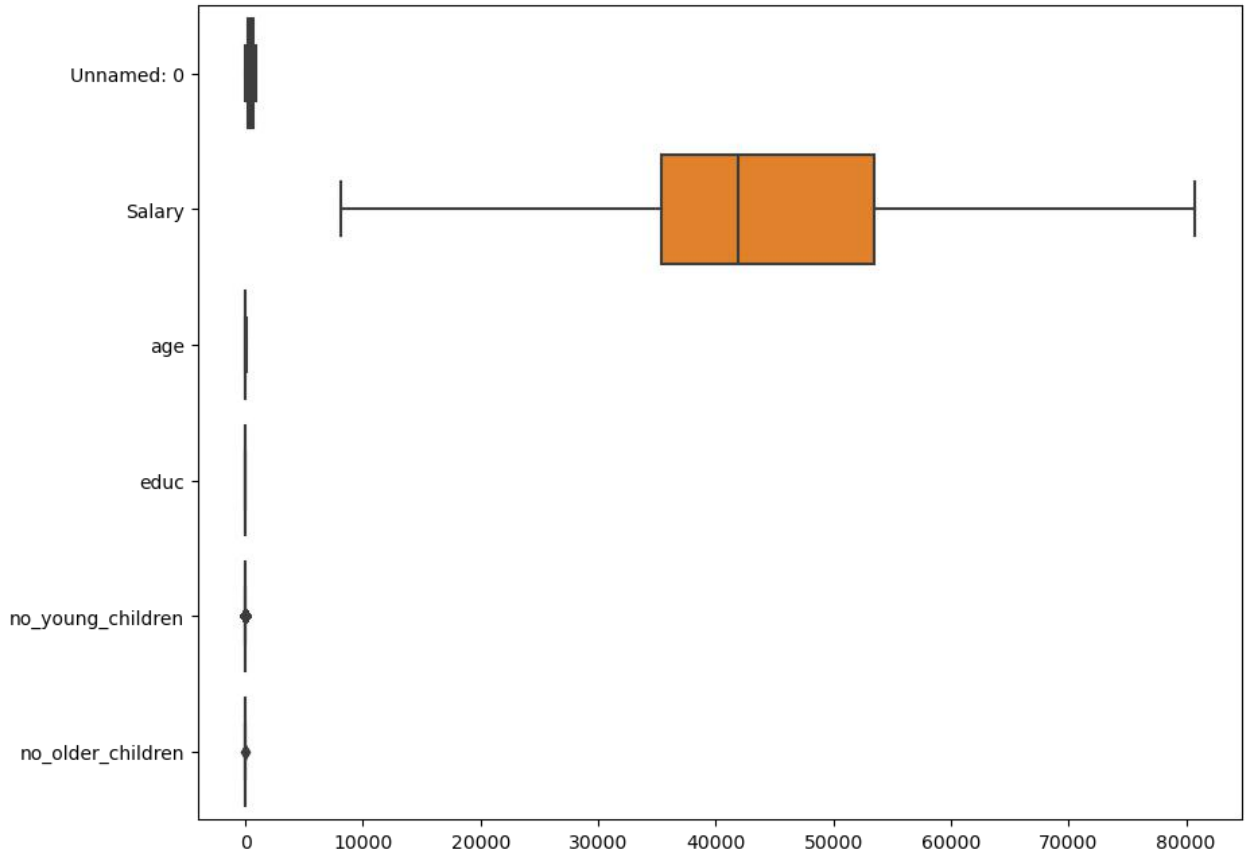
Eyeballing on the data,we see there's no such features which completely separate two required classes. There's an overlap between every features. But still features such as Salary and educ can be slightly useful for separating classes.


Unnamed: 0
Salary
age
educ
no_young_children
no_older_children
foreign
no
yes



No good correlations to be seen in a heatmap.

Then I treated the outliers and shown the boxplot:



Then I changed the object datatype to numeric datatype,

```
<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   Holiday_Package    872 non-null    int8
 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    int8
dtypes: int64(6), int8(2)
memory usage: 42.7 KB
```

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).

Answer

In [171]:
```
X=df2.copy()
X.drop('Holiday_Package',axis=1,inplace=True)
Y=df2.pop('Holiday_Package')
#Creating new dataset containing IV and DV in it
```

In [172]:
```
X.head()
```

Out[172]:

	Unnamed: 0	Salary	age	educ	no_young_children	no_older_children	foreign
0	1	48412	30	8	1	1	0
1	2	37207	45	8	0	1	0
2	3	58022	46	9	0	0	0
3	4	66503	31	11	2	0	0
4	5	66734	44	12	0	2	0

In [173]:
```
X_train,X_test,train_labels,test_labels,=train_test_split(X,Y,test_size=0.30,random_state=1)
#Splting the data into training and testing dataset
```

In [174]:
```
clf=LogisticRegression()
```

In [175]:
```
clf.fit(X_train,train_labels)
#Fitting the Trained Model into training dataset
```

Out[175]:
```
LogisticRegression()
```

```
#Making of LDA model

lda=LinearDiscriminantAnalysis(tol=0.00001,solver='svd')
model=lda.fit(X_train,train_labels)
model
#Fitting lda model on training data.
```

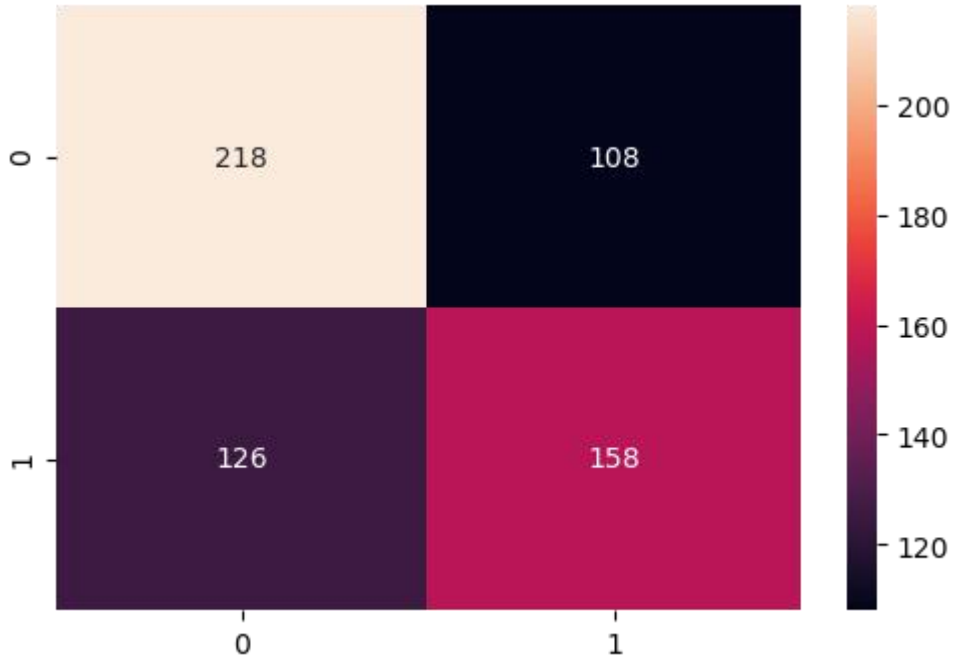
```
LinearDiscriminantAnalysis(tol=1e-05)
```

I built the model as asked in the question without any scaling.

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.

Answer

Confusion Matrix of Train set:

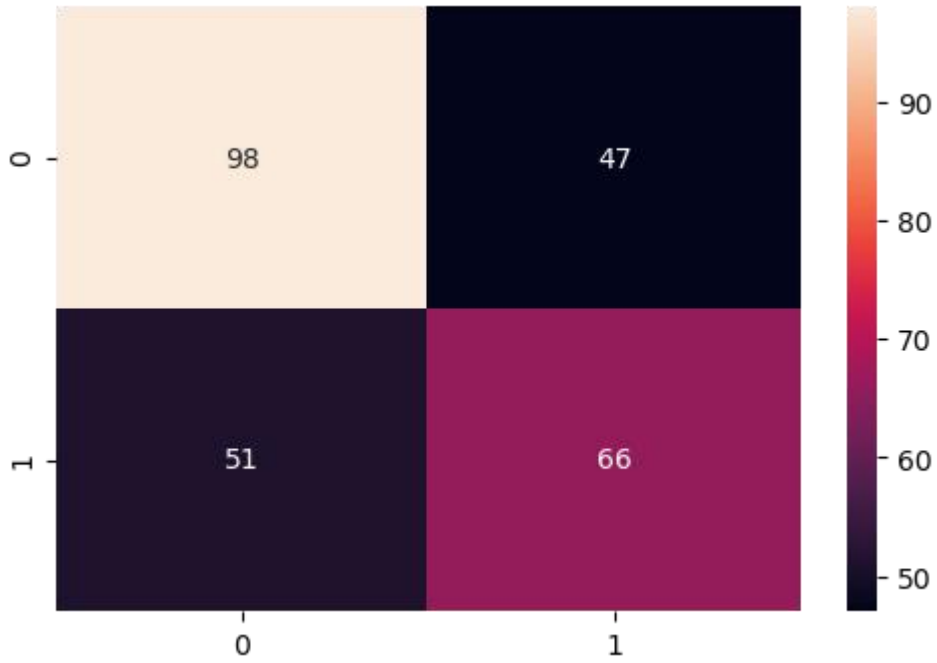


Classification Report on Train Data:

	precision	recall	f1-score	support
0	0.63	0.67	0.65	326
1	0.59	0.56	0.57	284
accuracy			0.62	610
macro avg	0.61	0.61	0.61	610
weighted avg	0.62	0.62	0.62	610

Accuracy is poor.
Recall is very poor.

Confusion Matrix on Test set:



Number of False negative is very high . Which proves this model is not at all suitable as it incorrectly determines our potential customers.

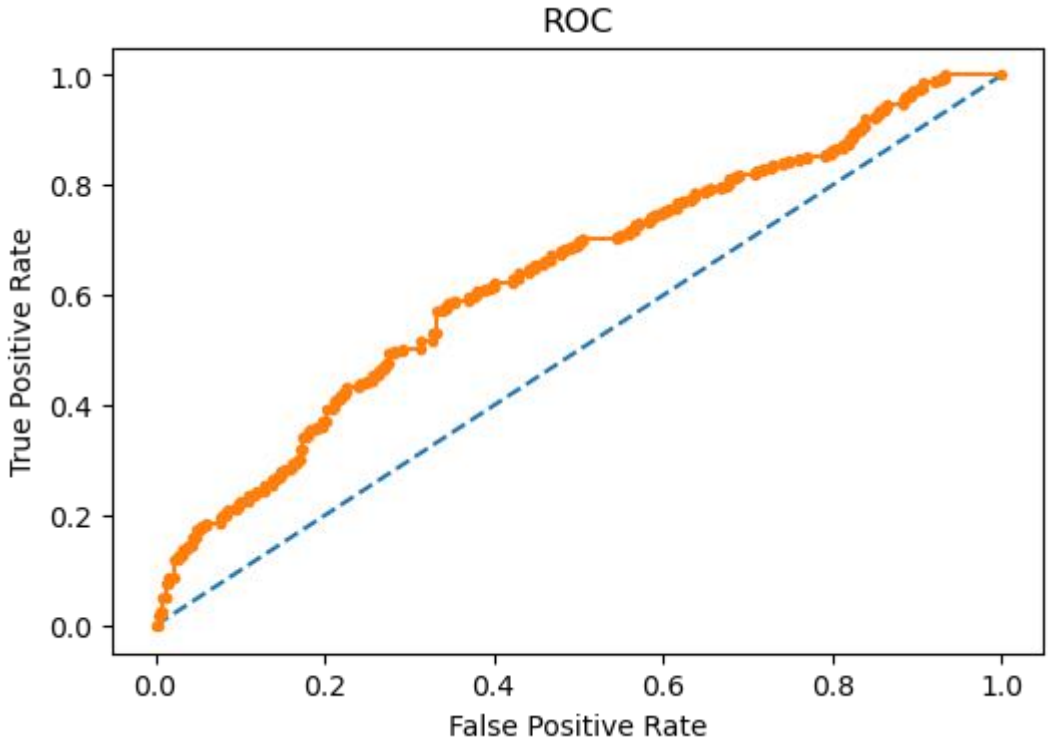
Classification Report on Test Set:

	precision	recall	f1-score	support
0	0.66	0.68	0.67	145
1	0.58	0.56	0.57	117
accuracy			0.63	262
macro avg	0.62	0.62	0.62	262
weighted avg	0.62	0.63	0.63	262

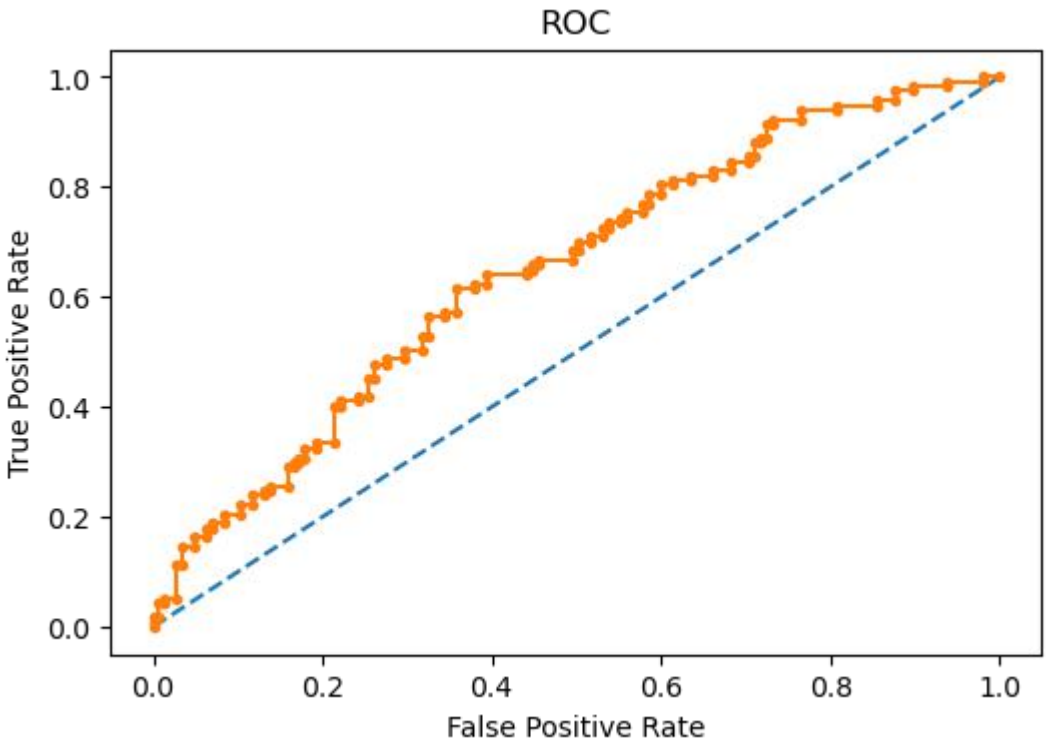
Poor results as that of testing dataset.

AUC score of Train set: 63.66%
AUC score of Test set: 64.91%

ROC curve of Train Set:

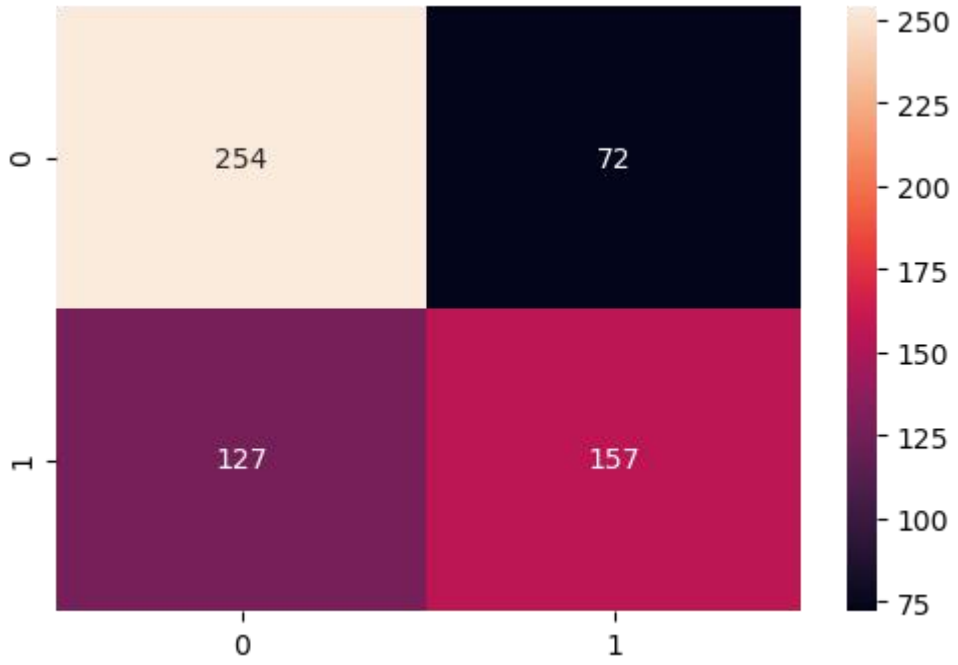


ROC curve of Test Set:



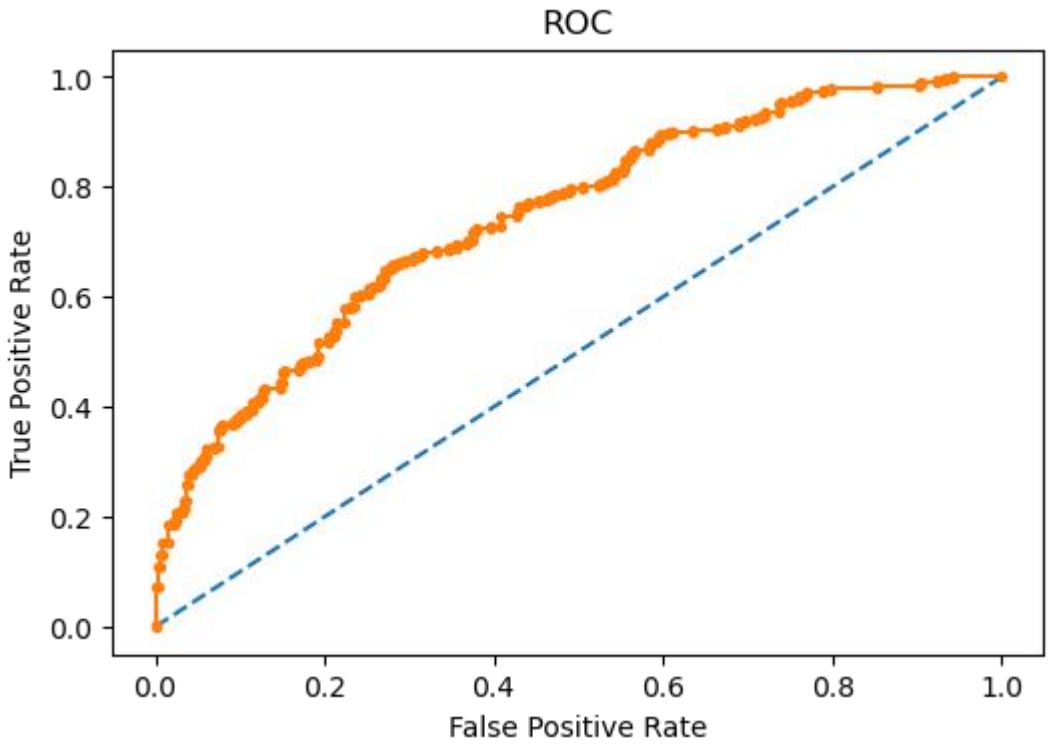
Now I made an LDA model to fit the train data and test data.
Then made a Confusion Matrices and Classification reports as well:

Train Data:

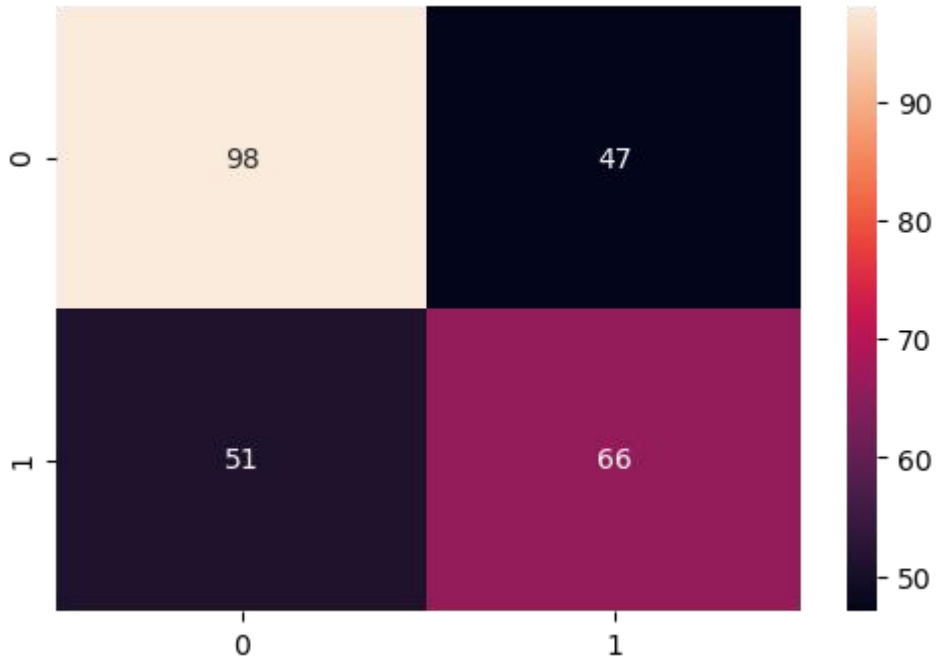


	precision	recall	f1-score	support
0	0.67	0.78	0.72	326
1	0.69	0.55	0.61	284
accuracy			0.67	610
macro avg	0.68	0.67	0.67	610
weighted avg	0.68	0.67	0.67	610

AUC score for training dataset: 74.60%

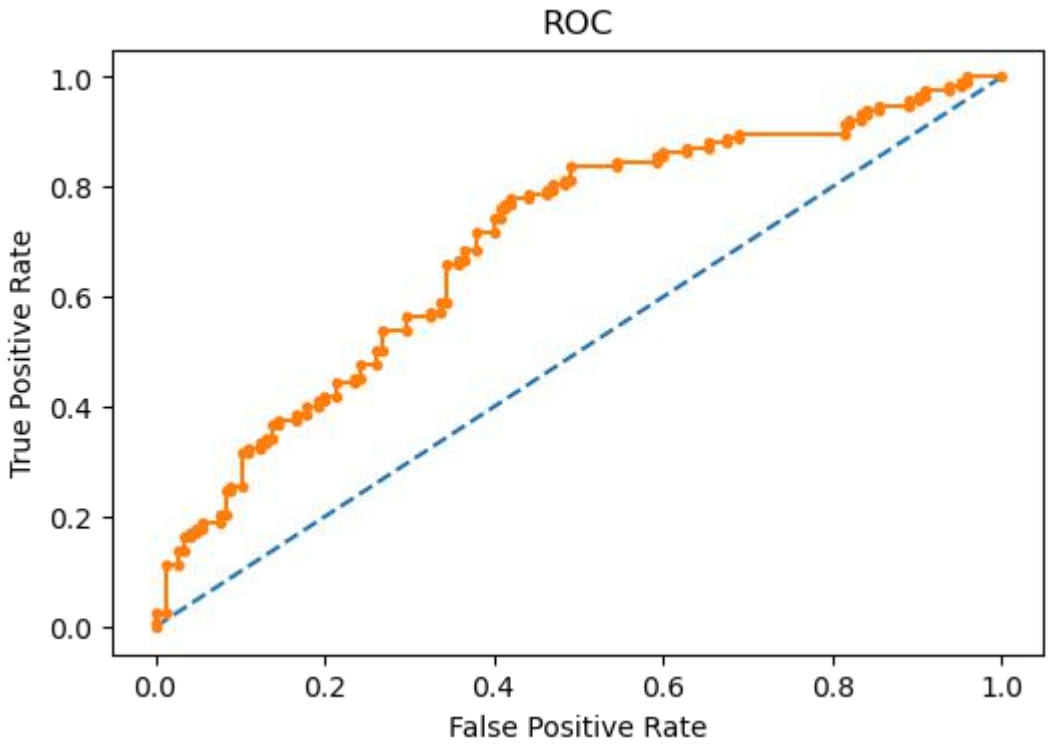


Test Data:



	precision	recall	f1-score	support
0	0.66	0.68	0.67	145
1	0.58	0.56	0.57	117
accuracy			0.63	262
macro avg	0.62	0.62	0.62	262
weighted avg	0.62	0.63	0.63	262

AUC score for test dataset: 69.66%



Accuracy and Recall are higher as compared to Logistic Model.

Looking into all the important parameters such as Accuracy, Recall, AUC score and ROC curve , LDA is performing better as compared to Logistic Regression model.

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

Answer:

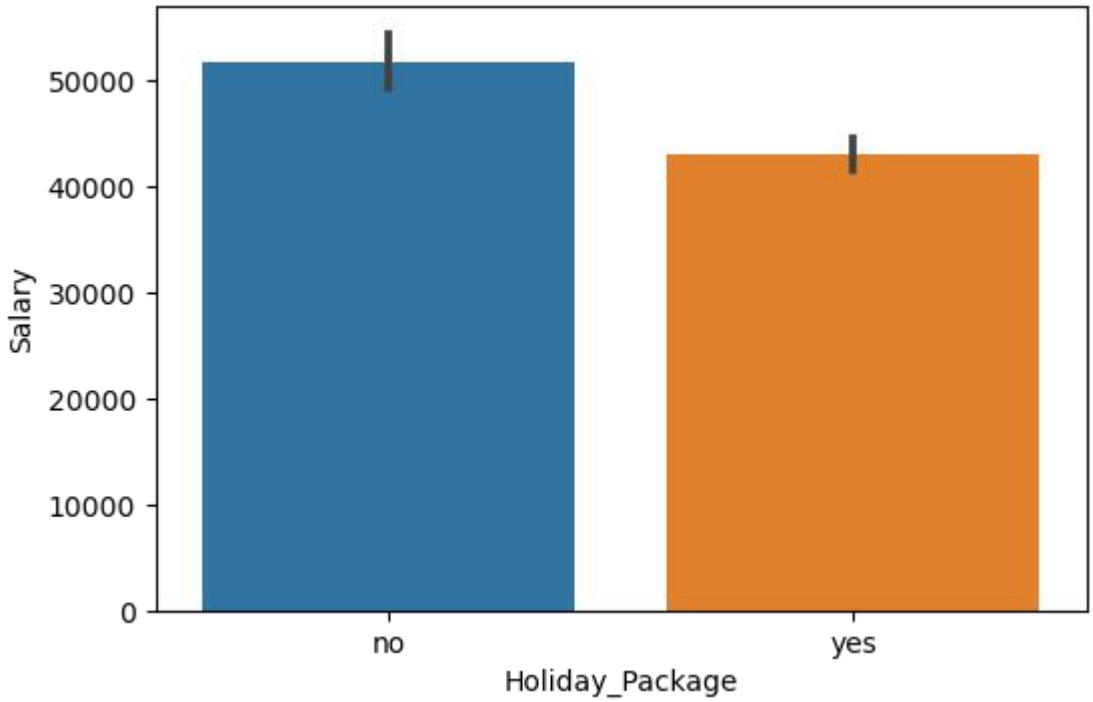
INSIGHTS

1. Looking into all the important parameters such as Accuracy, Recall, AUC score and ROC curve, LDA is performing better as compared to Logistic Regression mode.

2. But the accuracy that we are getting is still not good to make any predictions. Hence we should try some more models such as neural networks, random forests to choose the optimum model for our predictions.

3. We should try to gather some more data to make the model better and more robust.

4. We can change outlier treatment techniques such as scaling the data and treating those values which are above and below +3 & -3 SD respectively. It was not done here as we were not asked to scale the data.



End Of Business Report 2