Problem 1 Linear Regression

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

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

Data set :

carat	cut	color	clarity	depth	table	x	y	z	price
0.3	Ideal	E	SI1	62.1	58	4.27	4.29	2.66	499
0.33	Premium	G	IF	60.8	58	4.42	4.46	2.7	984
0.9	Very Good	E	VVS2	62.2	60	6.04	6.12	3.78	6289
0.42	Ideal	F	VS1	61.6	56	4.82	4.8	2.96	1082
0.31	Ideal	F	VVS1	60.4	59	4.35	4.43	2.65	779
1.02	Ideal	D	VS2	61.5	56	6.46	6.49	3.99	9502
1.01	Good	H	SI1	63.7	60	6.35	6.3	4.03	4836
0.5	Premium	E	SI1	61.5	62	5.09	5.06	3.12	1415
1.21	Good	H	SI1	63.8	64	6.72	6.63	4.26	5407
0.35	Ideal	F	VS2	60.5	57	4.52	4.6	2.76	706
0.32	Ideal	E	VS2	61.6	56	4.4	4.43	2.72	637
1.1	Premium	D	SI1	60.7	55	6.74	6.71	4.08	6468
0.5	Good	E	VS1	61.1	58.2	5.08	5.12	3.11	1932
0.71	Ideal	D	SI2	61.6	55	5.74	5.76	3.54	2767
1.5	Fair	G	VS2	66.2	53	7.12	7.08	4.7	10644
0.31	Ideal	G	VS2	61.6	55	4.37	4.39	2.7	544
0.34	Ideal	G	SI1	61.2	57	4.56	4.53	2.78	650
1.01	Ideal	D	VS2	59.8	56	6.52	6.49	3.89	7127
0.9	Good	D	SI1	61.9	64	6	6.09	3.74	3567
0.54	Premium	G	VS2	60	59	5.42	5.22	3.19	1637
1.04	Premium	D	VVS2	61.1	60	6.54	6.51	3.99	10984
0.4	Ideal	F	VS2	62.9	57	4.72	4.69	2.96	1080
1.52	Ideal	D	SI2	62.7	56	7.35	7.28	4.59	8631
1.19	Ideal	J	SI2	61.7	56	6.8	6.85	4.21	4508
0.66	Ideal	H	SI1	62.4	58	5.53	5.56	3.46	1609
1.5	Premium	H	SI2	61.4	62	7.4	7.25	4.49	7187

We are provided with the above data set of 26967 rows and 10 columns. Of the above columns, six columns are float data type , three columns are object data type and remaining one column is of integer data type.

Column **'depth'** has 697 null values and remaining columns have 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  float64
 1   cut      26967 non-null  object
 2   color    26967 non-null  object
 3   clarity  26967 non-null  object
 4   depth    26270 non-null  float64
 5   table    26967 non-null  float64
 6   x        26967 non-null  float64
 7   y        26967 non-null  float64
 8   z        26967 non-null  float64
 9   price    26967 non-null  int64
dtypes: float64(6), int64(1), object(3)
memory usage: 2.1+ MB
```

Descriptive statistics for the dataset:

	carat	cut	color	clarity	depth	table	x	y	z	price
count	26967.000000	26967	26967	26967	26270.000000	26967.000000	26967.000000	26967.000000	26967.000000	26967.000000
unique	NaN	5	7	8	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	Ideal	G	SI1	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	10816	5661	6571	NaN	NaN	NaN	NaN	NaN	NaN
mean	0.798375	NaN	NaN	NaN	61.745147	57.456080	5.729854	5.733569	3.538057	3939.518115
std	0.477745	NaN	NaN	NaN	1.412860	2.232068	1.128516	1.166058	0.720624	4024.864666
min	0.200000	NaN	NaN	NaN	50.800000	49.000000	0.000000	0.000000	0.000000	326.000000
25%	0.400000	NaN	NaN	NaN	61.000000	56.000000	4.710000	4.710000	2.900000	945.000000
50%	0.700000	NaN	NaN	NaN	61.800000	57.000000	5.690000	5.710000	3.520000	2375.000000
75%	1.050000	NaN	NaN	NaN	62.500000	59.000000	6.550000	6.540000	4.040000	5360.000000
max	4.500000	NaN	NaN	NaN	73.600000	79.000000	10.230000	58.900000	31.800000	18818.000000

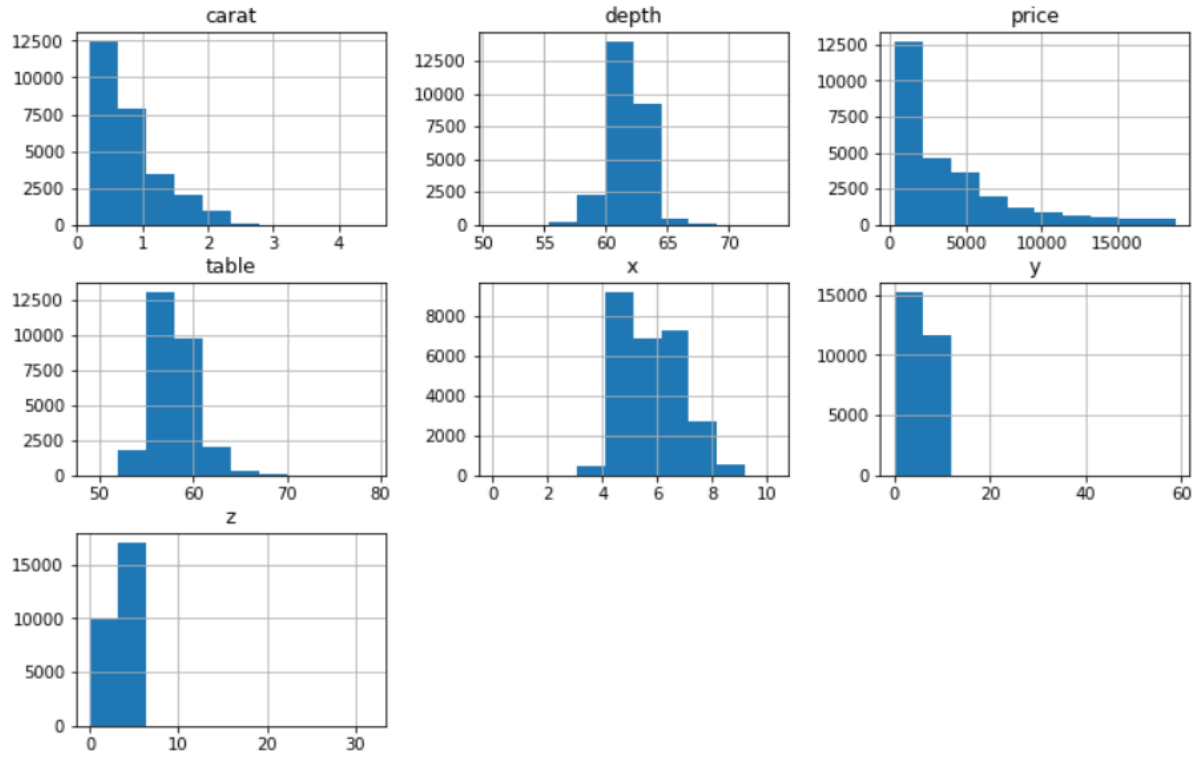
We have Columns **'cut','color','clarity'** as **categorical type** data and column **'price'** as **integer type** and remaining all columns are **float type** data.

As per the details resulted from the descriptive statistics of the dataset, we can find that:

All the variables except variable 'depth' have a **value count** of **26967** with **no null values** and variable 'depth' has a value count of 26270 with 697 null values.

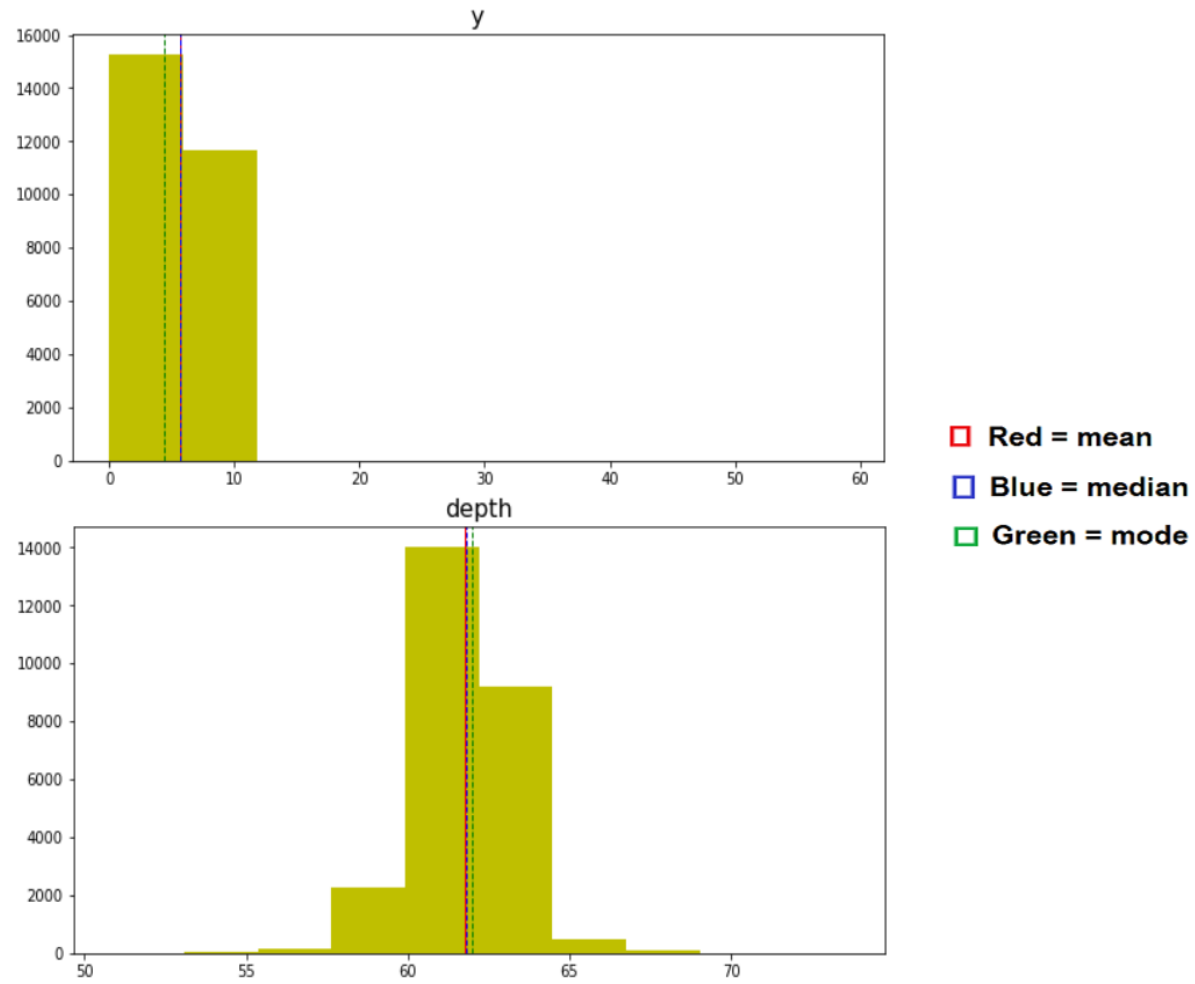
Of the entire dataset, column **'price'** has **highest max** value of 18818 and columns **'x','y' and 'z'** have **least min** value of 0.00

Columns **'price'** and **'carat'** have highest mean value – 3939.518 and least mean value – 0.79837 respectively.

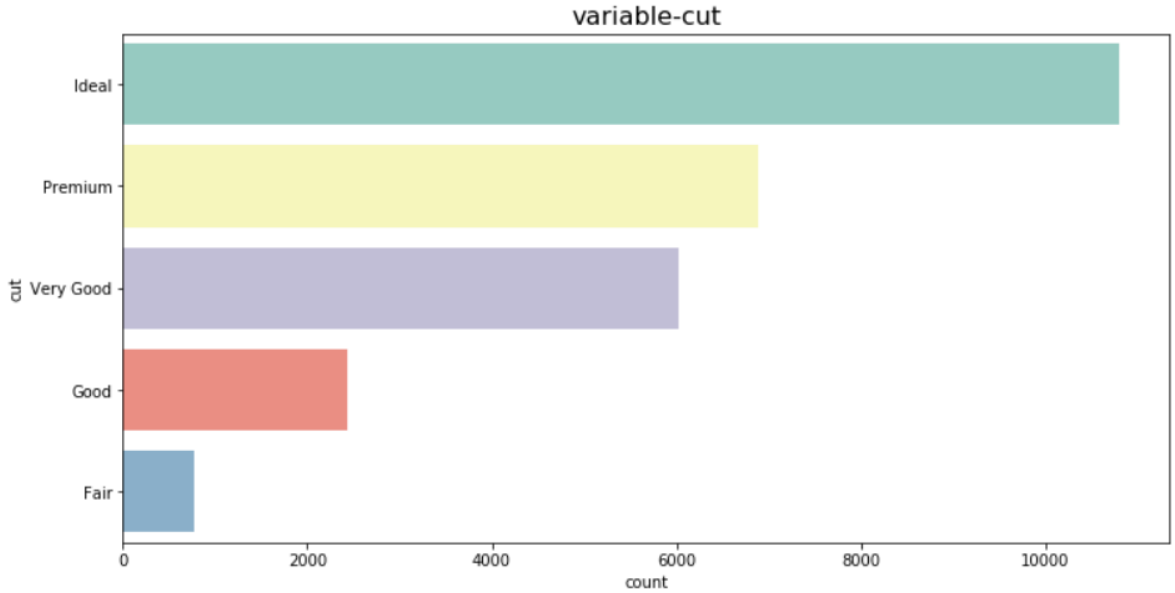


From the above histograms of the variables , we can see that majority of the variables are not symmetrical.

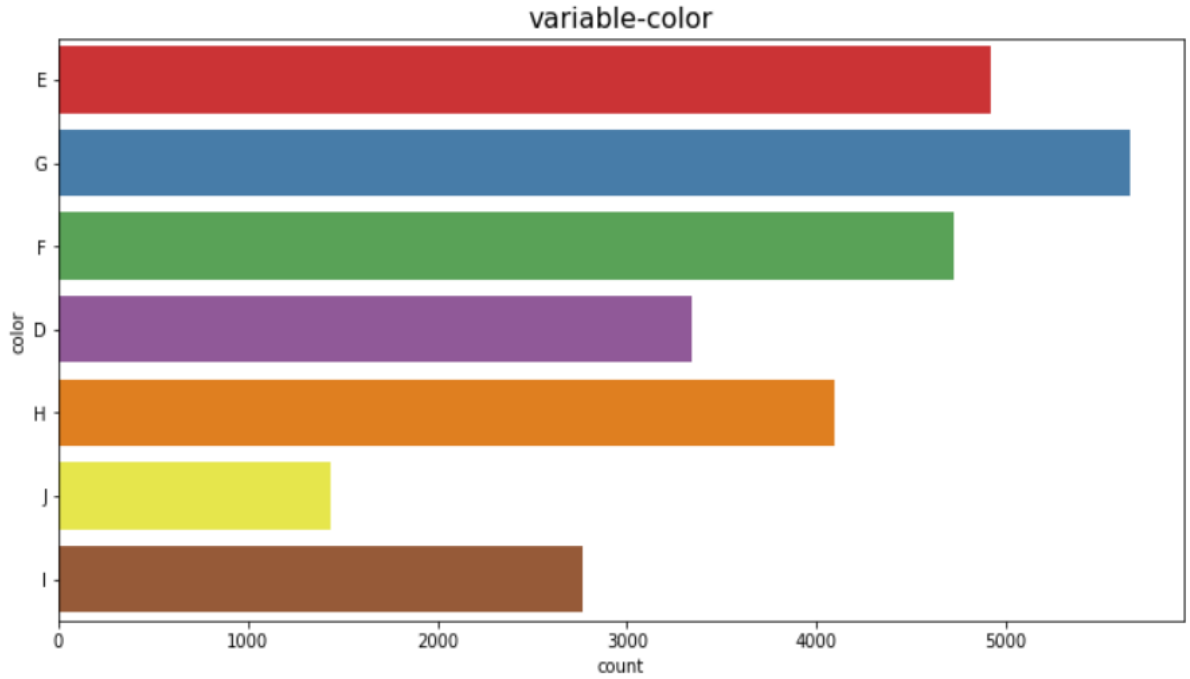
Among all the variables , Variable **'y'** is highly **right skewed** (skew = 3.867764) and Varible **'depth'** is highly **left skewed** (skew = -0.026086).



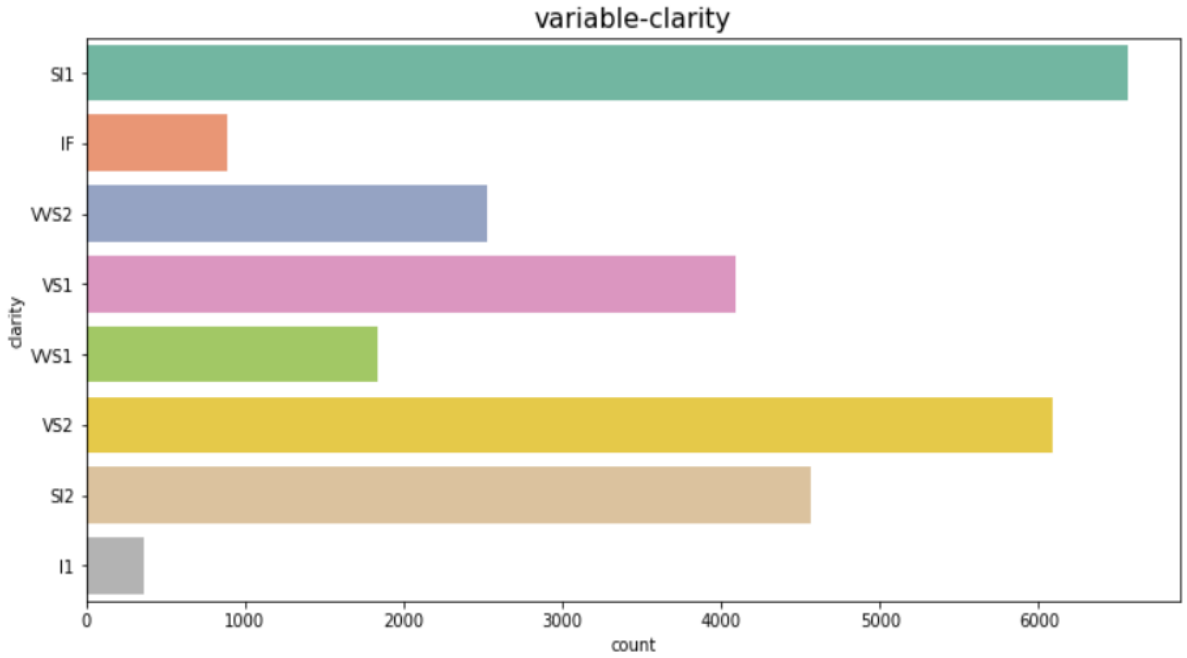
Categorical Data type:



Variable **'cut'** has highest presence of **'Ideal cut'** quality of the cubic zirconia and least presence of **'Fair cut'** quality



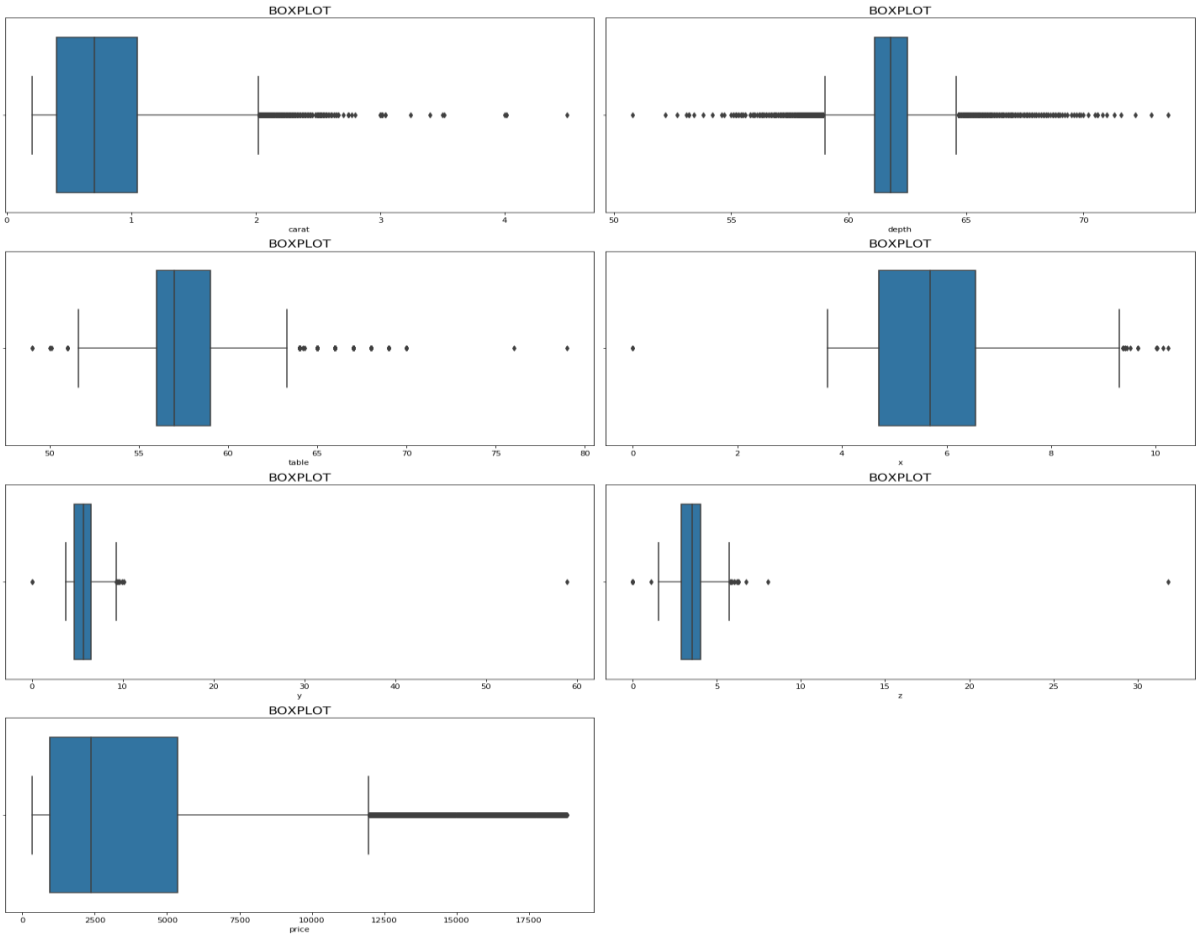
Variable **'color'** has highest presence of **'G'** color of the cubic zirconia and least presence of **'J'** color



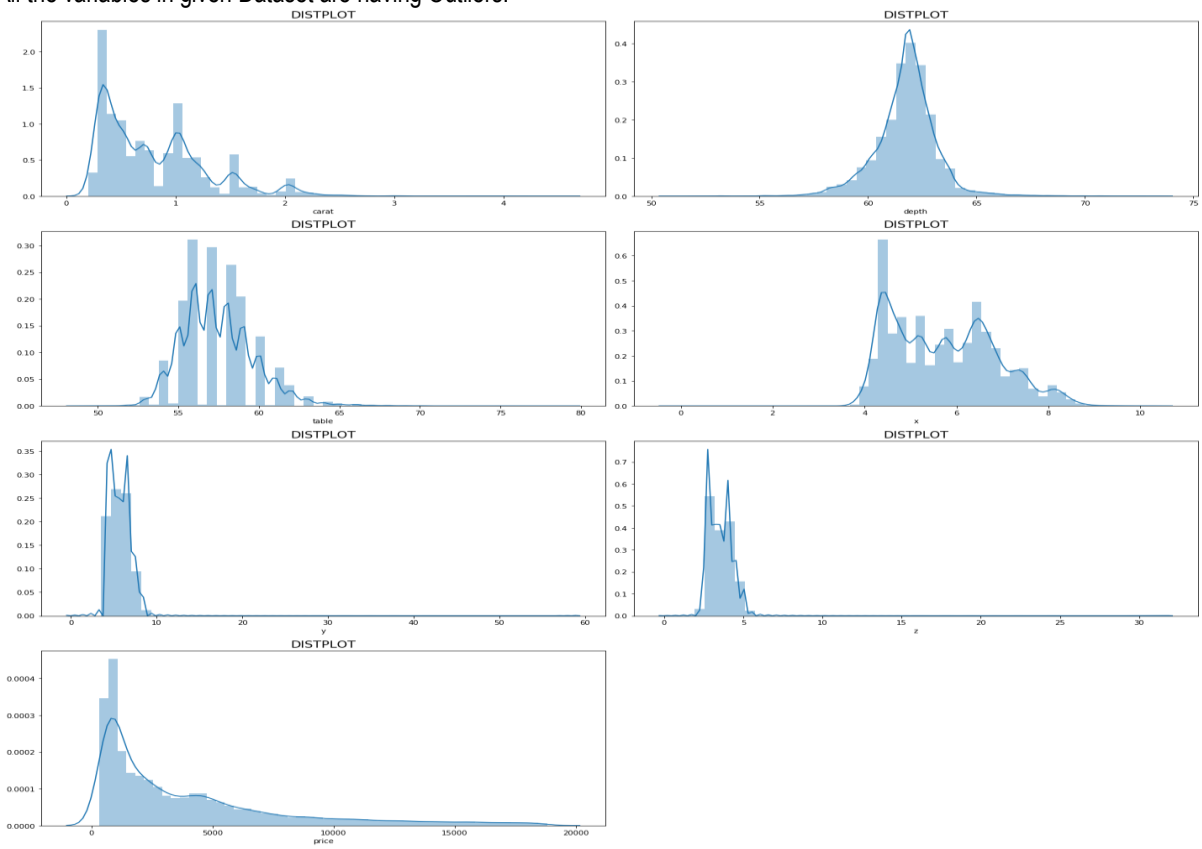
Variable **'clarity'** has highest presence of **'SI1'** clarity of the cubic zirconia and least presence of **'I1'** clarity

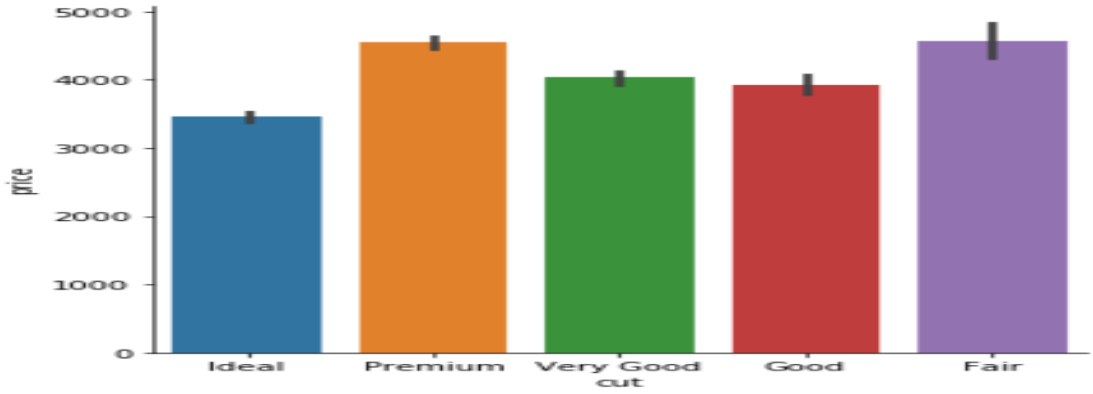
Skewness:

```
carat    1.114789
depth   -0.026086
table    0.765805
x        0.392290
y        3.867764
z        2.580665
price    1.619116
dtype: float64
```

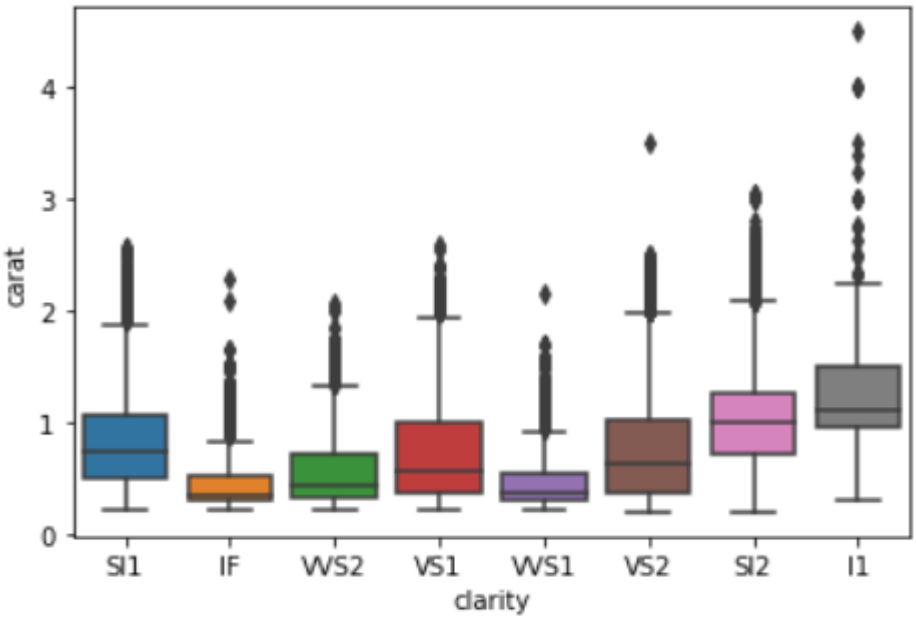


All the variables in given Dataset are having Outliers.

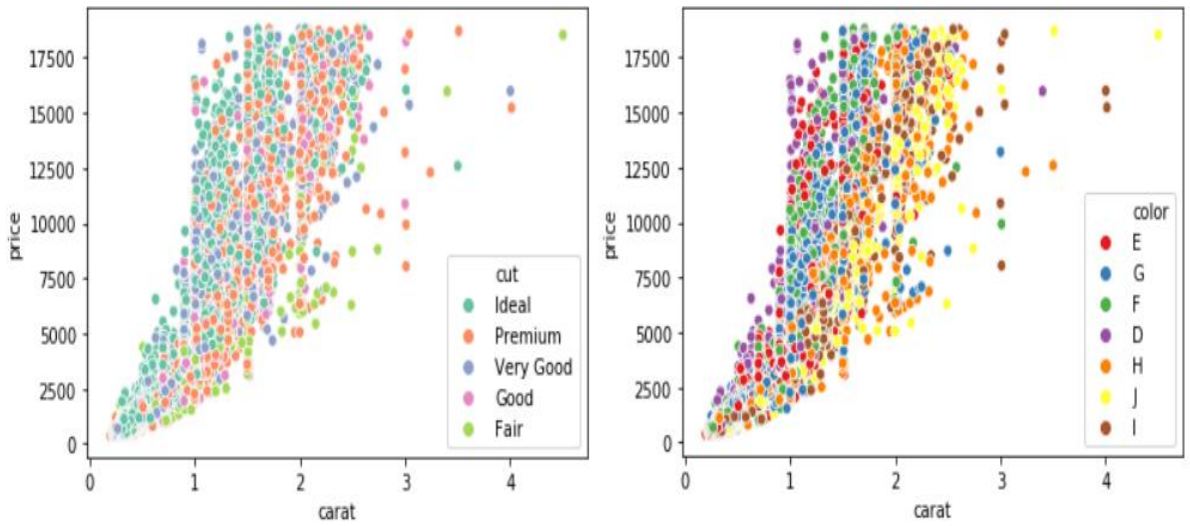




'Premium' and 'Fair' have highest price among all the varieties.



Clarity 'I1' has highest median value and clarity 'WS1' has the least median value.



As per above plots, we can see that 'Ideal quality' is on lower side and 'Fair quality' is on higher side.

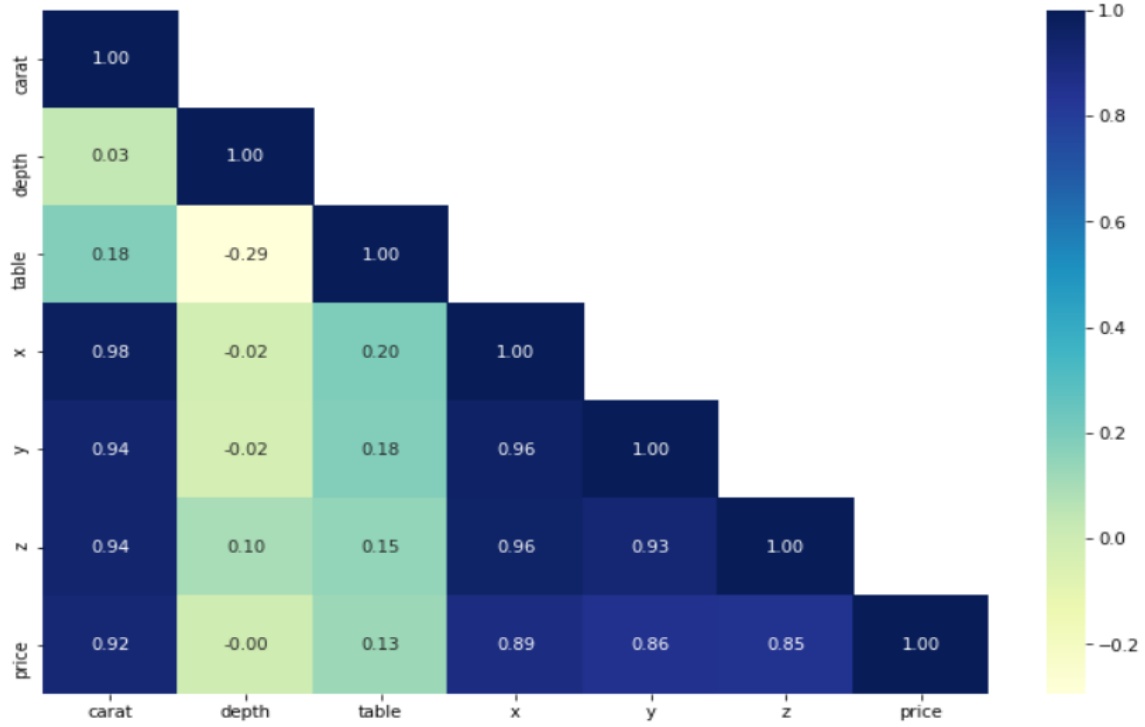
Also 'color D' is on lower side and 'color J' is on higher side.

Multivariate Analysis:

We have the following correlation among the different variables given in the dataset.

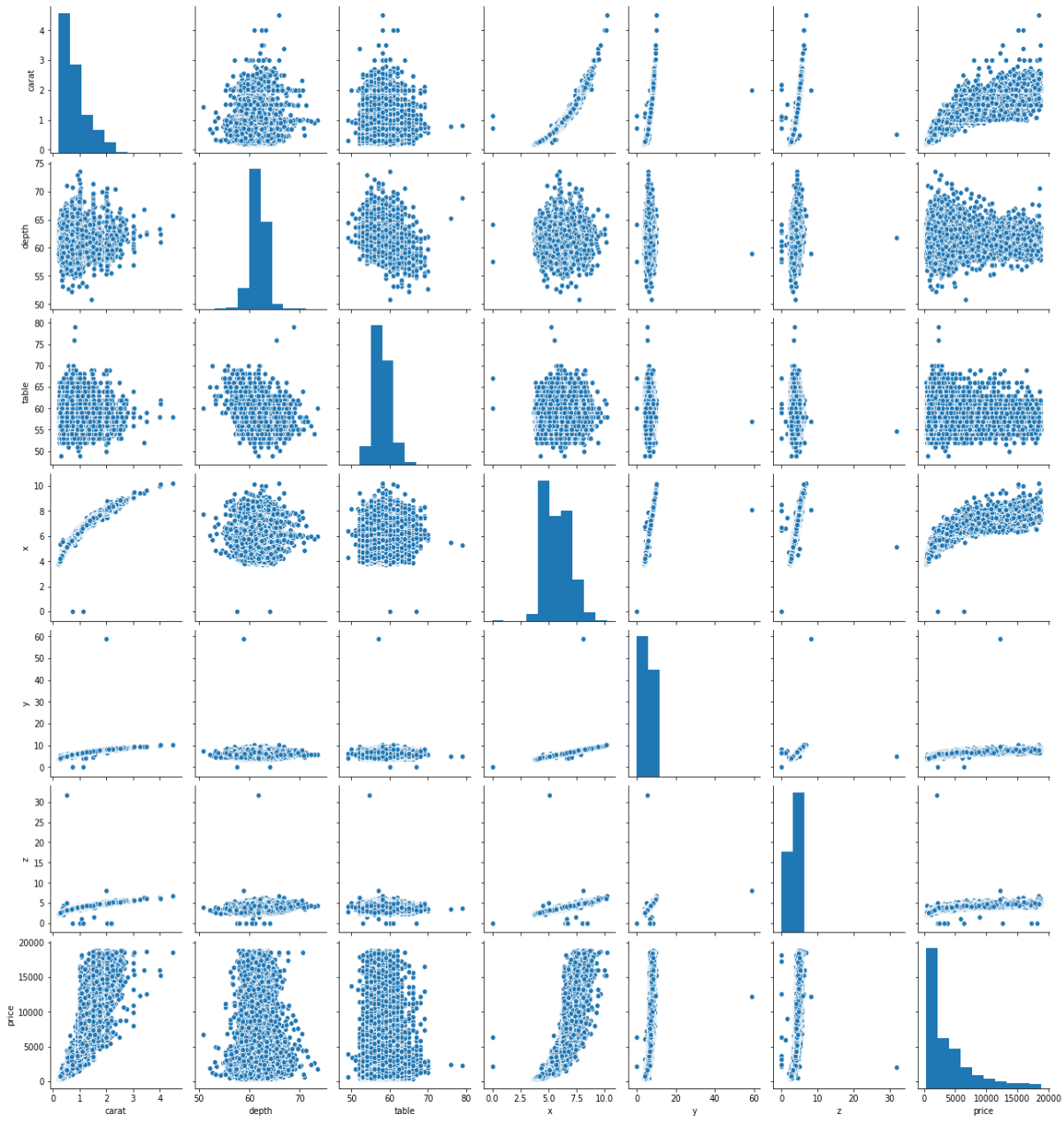
	carat	depth	table	x	y	z	price
carat	1.000000	0.034772	0.181539	0.976858	0.941442	0.940982	0.922409
depth	0.034772	1.000000	-0.293713	-0.018120	-0.024119	0.097733	-0.002840
table	0.181539	-0.293713	1.000000	0.196254	0.182352	0.148994	0.126844
x	0.976858	-0.018120	0.196254	1.000000	0.962601	0.956490	0.886554
y	0.941442	-0.024119	0.182352	0.962601	1.000000	0.928725	0.856441
z	0.940982	0.097733	0.148994	0.956490	0.928725	1.000000	0.850682
price	0.922409	-0.002840	0.126844	0.886554	0.856441	0.850682	1.000000

HeatMap:



From the above map , we can see that many columns are co-related to each other and there is **highest positive correlation** (0.96) between variables **'x'** and **'y'** . Also there is **highest negative correlation**(- 0.29) between variables **'table'** and **'depth'**.

Pairplot:



In the above plot scatter diagrams are plotted for all the numerical columns in the dataset. From the visual representation , we can understand the degree of correlation between any two columns of the given dataset.

Variables **'x','y' and 'z'** show positive linear correlation with Variable **'carat'**.

Variables **'y' , 'z'** show positive linear correlation with Variable '**depth'**.

Variables **'y' , 'z'** show positive linear correlation with Variable '**x'.**

Variable **'z'** shows positive linear correlation with Variable '**y'**

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

Among all the variables of the given dataset, Column **'depth'** has 697 null values and remaining columns have no Null values.

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

We impute the null values of the variable 'depth' with its median value.

	carat	cut	color	clarity	depth	table	x	y	z	price
min	0.200000	NaN	NaN	NaN	50.800000	49.000000	0.000000	0.000000	0.000000	326.000000

There is presence of **zero values** in variables **'x','y' and 'z'** of the given dataset. Practically zero values for length, breadth and height of cubic zirconia is not possible. So, we impute these values with lower range values (Q1-(1.5 * IQR)) of the respective variables.

As most of the variables are highly skewed towards either right or left and also there is much difference in the magnitude of value in the dataset between values among different variables. We prefer to use z score method to minimize the effect of both skewness and variability due to difference between high and low magnitude value of the data. Also scaling of variables does not affect the accuracy of the model.

Scaled Variables :

	carat	cut	color	clarity	depth	table	x	y	z
0	-1.067471	-0.541748	-0.940999	-1.063351	0.286857	0.261676	-1.295847	-1.288982	-1.258616
1	-1.002552	0.434559	0.231435	-1.643046	-0.780022	0.261676	-1.162650	-1.136943	-1.201206
2	0.230898	1.410866	-0.940999	1.835122	0.368925	1.188856	0.275874	0.347673	0.348871
3	-0.807797	-0.541748	-0.354782	0.096038	-0.123481	-0.665503	-0.807459	-0.832865	-0.828039
4	-1.045831	-0.541748	-0.354782	1.255428	-1.108292	0.725266	-1.224808	-1.163774	-1.272969

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

The given dataset contains three categorical variables 'cut','color' and 'quality'.

CUT : 5	
Fair	780
Good	2435
Very Good	6027
Premium	6886
Ideal	10805

COLOR : 7	
J	1440
I	2765
D	3341
H	4095
F	4723
E	4916
G	5653

CLARITY : 8	
I1	364
IF	891
VVS1	1839
VVS2	2530
VS1	4087
SI2	4564
VS2	6093
SI1	6565

We have to encode the data in these variables to use them in the models,

	carat	cut	color	clarity	depth	table	x	y	z
0	0.30	2	1	2	62.1	58.0	4.27	4.29	2.66
1	0.33	3	3	1	60.8	58.0	4.42	4.46	2.70
2	0.90	4	1	7	62.2	60.0	6.04	6.12	3.78
3	0.42	2	2	4	61.6	56.0	4.82	4.80	2.96
4	0.31	2	2	6	60.4	59.0	4.35	4.43	2.65
...	...	...	...	...	...	...	...	...	...
26962	1.11	3	3	2	62.3	58.0	6.61	6.52	4.09
26963	0.33	2	4	1	61.9	55.0	4.44	4.42	2.74
26964	0.51	3	1	5	61.7	58.0	5.12	5.15	3.17
26965	0.27	4	2	7	61.8	56.0	4.19	4.20	2.60
26966	1.25	3	6	2	62.0	58.0	6.90	6.88	4.27

Splitting data into train and test (70:30)

```
X_train.head()
```

	carat	cut	color	clarity	depth	table	x	y	z
22114	-0.980913	1.410866	0.817653	0.675733	0.533060	1.188856	-1.171530	-1.154830	-1.115091
2275	-1.067471	-0.541748	-0.940999	0.675733	-0.451751	-1.129093	-1.224808	-1.271095	-1.272969
19183	-0.634681	-0.541748	0.231435	-1.063351	0.615128	-0.201914	-0.567705	-0.609279	-0.526636
5030	0.663688	-1.518055	-0.940999	-0.483656	1.271669	-0.665503	0.710983	0.759072	0.879916
25414	0.490572	0.434559	-0.354782	-0.483656	-0.533819	2.116035	0.719863	0.678581	0.635922

```
X_test.head()
```

	carat	cut	color	clarity	depth	table	x	y	z
16997	1.009920	0.434559	1.403870	-1.063351	-1.026225	2.116035	1.101694	1.063150	0.951678
24457	0.230898	1.410866	0.231435	0.675733	-2.257239	0.725266	0.551147	0.562316	0.262755
16612	1.247954	-0.541748	-0.354782	0.096038	-1.764833	-0.201914	1.376967	1.331454	1.123909
308	0.101062	-1.518055	0.231435	0.096038	1.517872	-0.201914	0.222596	0.177747	0.363223
26652	2.611242	1.410866	1.990088	0.675733	-0.780022	2.116035	2.096229	2.136366	1.999415

```
y_train.head()
22114     537
2275      844
19183    1240
5030     4065
25414    4057
```

```
y_test.head()
16997     5292
24457     4484
16612    11649
308       3316
26652    13043
```

Performance Metrics:

The coefficient of determination R^2 of the prediction on:

Train set is 0.8950903008189984

Test set is 0.9006355201086821

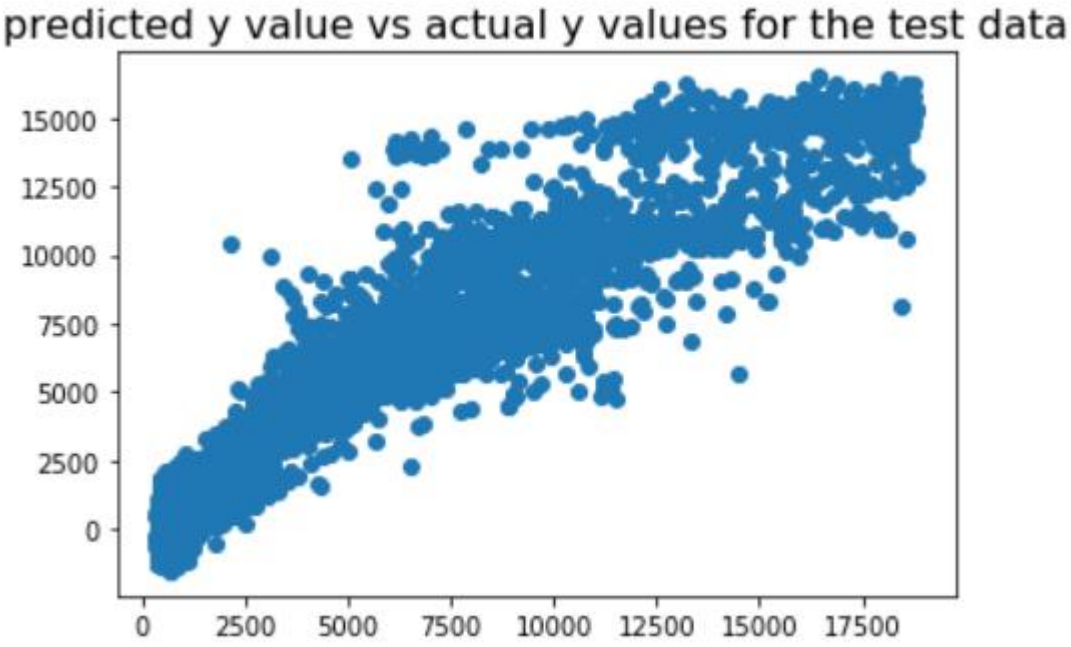
The Root Mean Square Error (RMSE) of the model of the:

Training set is 1298.6539514206925

Testing set is 1277.4915285667225

From the above values we can see that the model has high accuracy as the higher coefficient is an indicator of a better goodness of fit for the observations.

The coefficient of determination R^2 and The Root Mean Square Error (RMSE) for both the training and testing set are quite similar. So, we can conclude that the model is good.



From the above plot, we can see that the predicted 'y' values and actual 'y' values of the test set for the model are almost close. We can conclude that the model's prediction is good.

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

This linear regression model helped in estimating the values of the coefficients used in the representation with the data available to us. With this method, we are able to find out the cause and effect relationship between the independent and dependent variables of the given dataset.

The above model will be helpful in predicting the dependent variables through the independent variables and it assigned coefficients to the every predictor variable to give the best predictive/dependent variable.

As per predictions of the model, we have got the following coefficients for the independent variables of the given dataset.

coef	const	carat	cut	color	clarity	depth	table	x	y	z
	3940	6443	52.88	-484	477.2	-176	-210	-3668	1824	-643

Finally, we have following representation between the independent and dependent variables of the dataset.

Price = (6443*carat)+(52.88*cut)-(484*color)+(477.2*clarity)-(176*depth)-(210*table)-(3668*x)+(1824*y)-(643*z)+3940

The best 5 attributes that are most important from the given dataset are:

1.'**carat'** (Carat weight of the cubic zirconia.)
2.'**x'** (Length of the cubic zirconia in mm.)
3.'**y'** (Width of the cubic zirconia in mm.)
4.'**z'** (Height of the cubic zirconia in mm.)
5.'**color'** (Colour of the cubic zirconia)

Therefore, the company will be able to determine the price of the stone based on the above equation and it can distinguish between higher profitable stones and lower profitable stones by varying the values for the above five important variables to get the best price for the stones.