Zirconia Price Prediction

Project – Final Report

Table of Contents

List of Tables

List of Figures

EXECUTIVE SUMMARY

1. INTRODUCTION

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.

2. DATA DESCRIPTION

Field Name	Description	Detail	Data Type
Carat	Weight of the cubic zirconia	Carat	Numeric
Cut	Describe cut quality of the cubic zirconia	Quality in increasing order Fair, Good, Very Good, Premium, Ideal	Categorical (Ordinal)
Colour	Colour of the cubic zirconia	D being the worst and J the best	Categorical (Ordinal)
Clarity	Cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes	(In order from Best to Worst, IF = flawless, I1= level 1 inclusion) IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1	Categorical (Ordinal)
Depth	The Height of a cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter		Numeric
Table	The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter		Numeric
Price	Price of the cubic zirconia		Numeric
X	Length of the cubic zirconia	in mm	Numeric
Y	Width of the cubic zirconia	in mm	Numeric
Z	Height of the cubic zirconia	in mm	Numeric

Table 1: Description of dependent and independent variable

3. EXPLORATORY DATA ANALYSIS

3.1 Data Preparation

Dataset has 26967 rows and 10 features. Cut, Colour and Clarity are object types, price (dependent variable) is integer type and all other are float64 type.

Sample data

carat	cut	color	clarity	depth	table	x	y	z	price
0.3	Fair	D	SI2	62.1	58	4.27	4.29	2.66	499
0.33	Premium	G	VVS1	60.8	58	4.42	4.46	2.7	984
0.9	Very Good	D	VVS2	62.2	60	6.04	6.12	3.78	6289
0.42	Fair	F	VS1	61.6	56	4.82	4.8	2.96	1082
0.31	Fair	F	IF	60.4	59	4.35	4.43	2.65	779

Table 2: Sample data

Data Summary

	carat	cut	colour	clarity	depth	table	X	Y	z	price
count	26967	26967	26967	26967	26270	26967	26967	26967	26967	26967
unique	-	5	7	8	-	-	-	-	-	-
Top	-	Fair	G	SI2	-	-	-	-	-	-
freq	-	10816	5661	6571	-	-	-	-	-	-
mean	0.798	-	-	-	61.75	57.46	5.73	5.73	3.54	3939.52
Std	0.478	-	-	-	1.41	2.23	1.13	1.17	0.72	4024.86
Min	0.2	-	-	-	50.8	49	0	0	0	326
0.25 Percentile	0.4	-	-	-	61	56	4.71	4.71	2.9	945
0.5 Percentile	0.7	-	-	-	61.8	57	5.69	5.71	3.52	2375
0.75 Percentile	1.05	-	-	-	62.5	59	6.55	6.54	4.04	5360
max	4.5	-	-	-	73.6	79	10.23	58.9	31.8	18818

Table 3: Data summary

Key Observations:

- Variables X, Y and Z, which are length, width and height of the stone, are having some zero value that is not possible and thus needs to be treated to avoid inclusion of erroneous data while building model.
- Cut variable has 5 unique values, color has 7 unique values and clarity has 8 unique values.
- Since the mean and median values are very far apart the variables seem to be skewed
- By looking at the dataset, it appears that there are outliers in the variables. The same is visible from the distribution of 5 values (min, 25 percentile, 50 percentile, 75 percentile and maximum)

Action:

- We will be removing all the records where length or width or height is zero

As listed below there are 9 records with zero length/width/height

carat	cut	color	clarity	depth	table	x	y	z	price
0.71	Good	F	I1	64.1	60	0	0	0	2130
2.02	Premium	H	SI1	62.7	53	8.02	7.95	0	18207
0.71	Good	F	I1	64.1	60	0	0	0	2130
2.2	Premium	H	SI2	61.2	59	8.42	8.37	0	17265
2.18	Premium	H	I1	59.4	61	8.49	8.45	0	12631
1.1	Premium	G	I1	63	59	6.5	6.47	0	3696
1.14	Ideal	G	VS1	57.5	67	0	0	0	6381
1.01	Premium	H	VS2	58.1	59	6.66	6.6	0	3167
1.12	Premium	G	VS2	60.4	59	6.71	6.67	0	2383

Table 4: Records with zero value

Duplicate records:
Dataset has 34 duplicate records and all these records are deleted from the data set. Duplicate data in structured data can be kept if you see it is reinforcing the outcome of data distribution. Duplicate inputs result in some distribution across your output and thus you need to retain that distribution. In this problem as the duplicate data are limited, it will not influence the outcome, it is better to keep them aside.

Dataset does have null values in “depth” feature.

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

Action:

We will impute data for null.

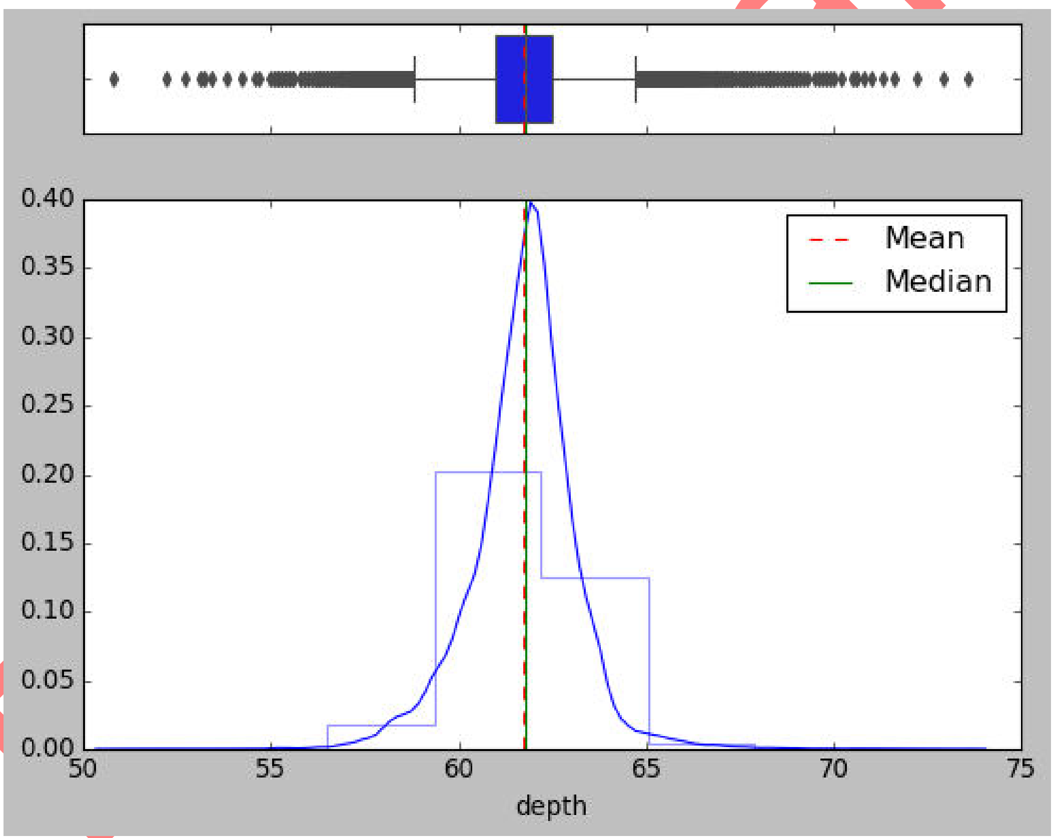


Figure 1: Depth boxplot and distribution plot

Depth Mean = 61.74 Depth Median = 61.8

As percentage of null value in variable depth is on lower side, we will go with the basic methodology of imputing the null value with mean or median. Depth has outliers and in many literatures you will find that in such scenario median is more preferred. In this case as the distribution is symmetric, mean and median are very close and thus it does not makes any difference.

We will impute the null values with median of depth.

3.2 Univariate Analysis

3.2.1 Univariate Analysis - Categorical Variable

CUT : 5	
Ideal	779
Good	2434
Very Good	6027
Premium	6880
Fair	10805

Table 5: Cut – Unique counts

COLOR : 7	
J	1440
I	2765
E	3341
H	4091
F	4722
D	4916
G	5650

Table 6: Colour – Unique counts

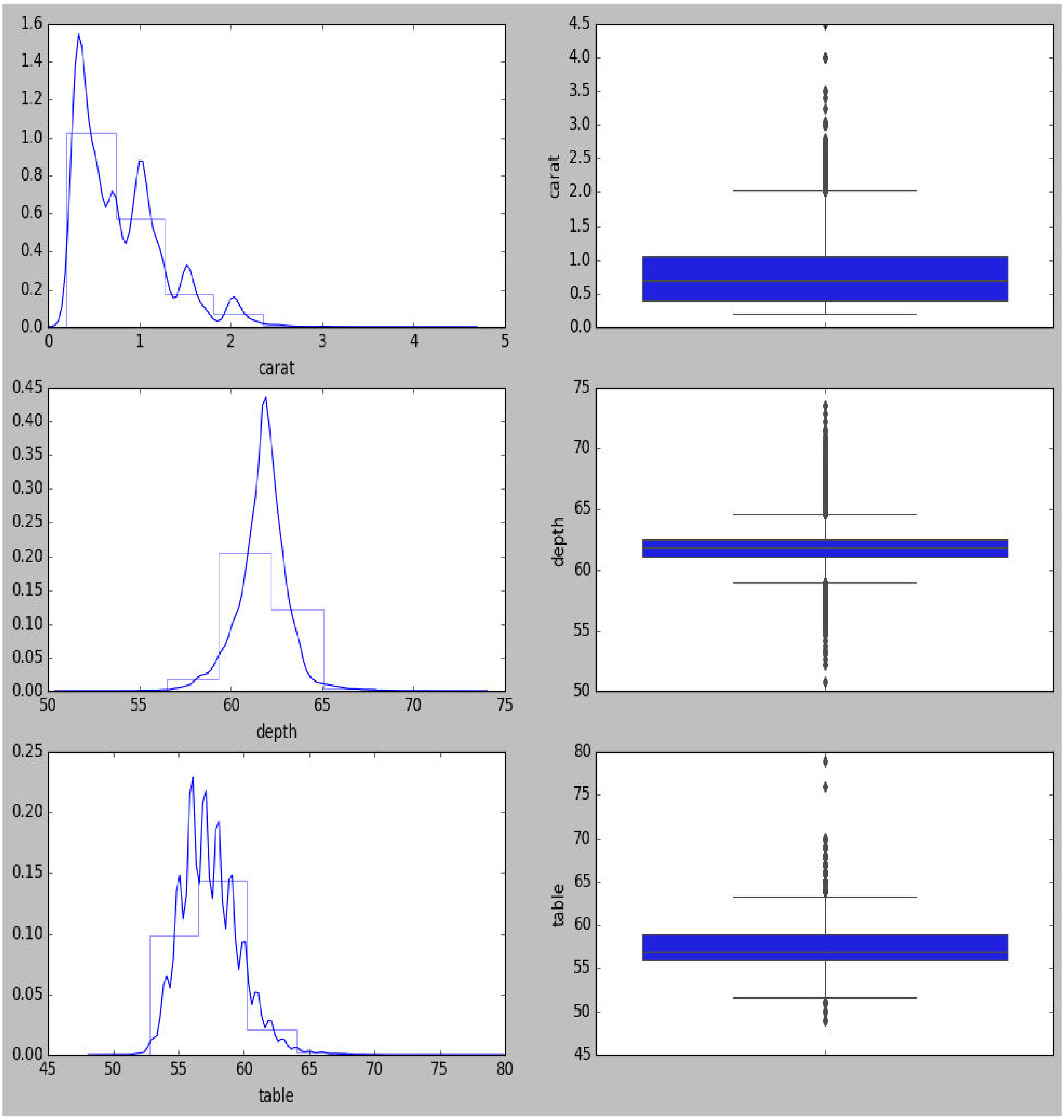
CLARITY : 8	
VS2	362
VVS1	891
IF	1839
VVS2	2530
VS1	4086
I1	4561
SI1	6092
SI2	6564

Table 7: Cut –Unique counts

Key Observations:

- Cut variable has 5 variants and Ideal type is highly present in the dataset
- Cut type: Ideal has the highest count followed by premium, very good, good and fair
- Color variable has 7 variants and with G and J being the most and least no of observations
- Color: G has the highest count followed by E, F, H, D, I and J.
- Clarity variable has 8 variants with S1 being the most frequent in the dataset
- Clarity: SI1 has higher contribution followed by VS2, SI2, VS1, VVS2, VVS1, IF and I1.

3.2.2 Univariate Analysis - Continuous Variable



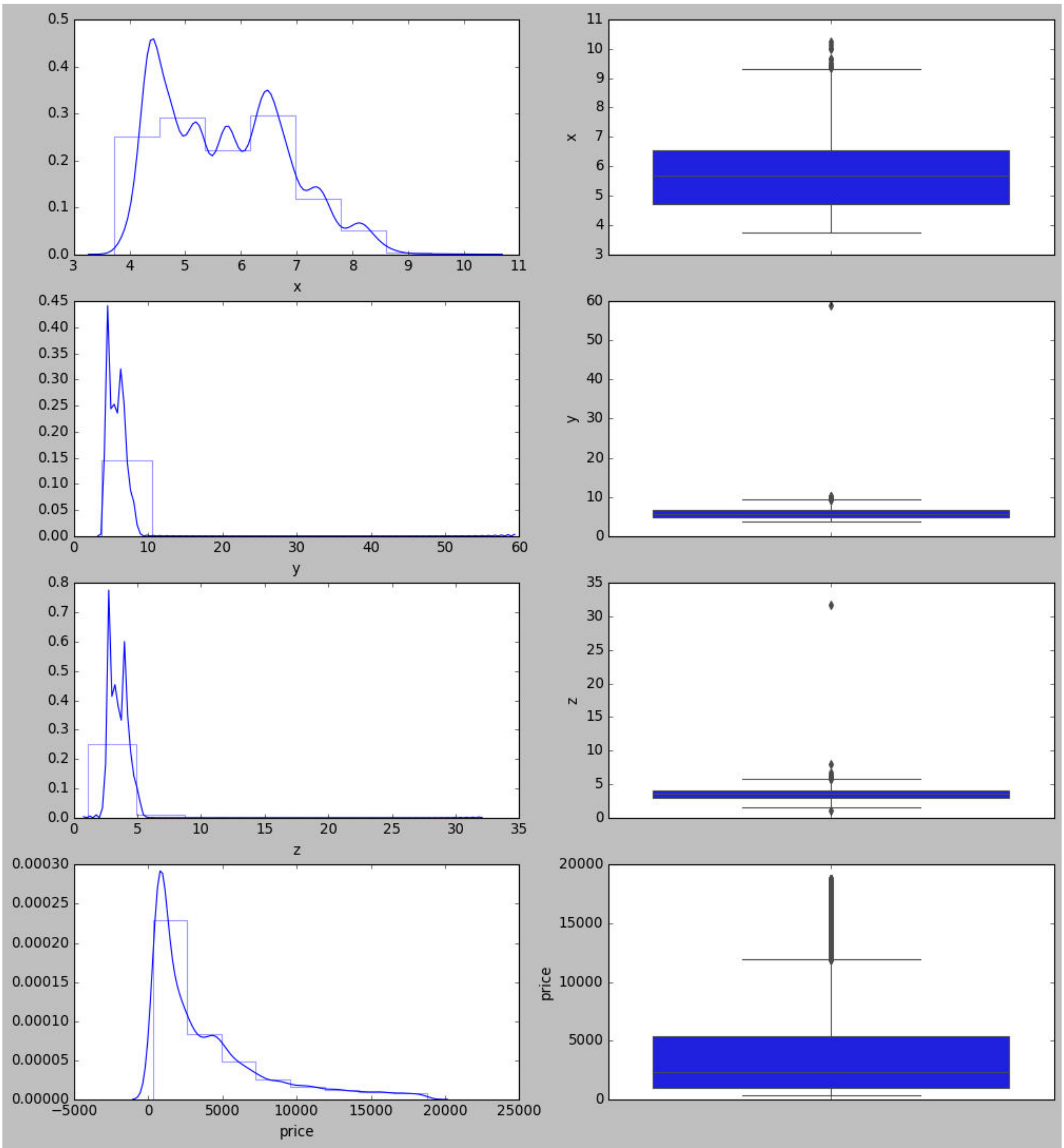


Figure 2: Distribution plot and box plot of continuous variable

Skewness

carat	1.114871
depth	-0.028403
table	0.764890
x	0.402010
y	3.888607
z	2.639529
price	1.619055

Key Observations:

- All the continuous variables have outliers. Which means all variables have values which are out of the range of (Q1 – 1.5* IQR) to (Q3 +1.5 * IQR) as shown below in the figure. However, as there is no value which seems to erroneous, we will not remove these values.

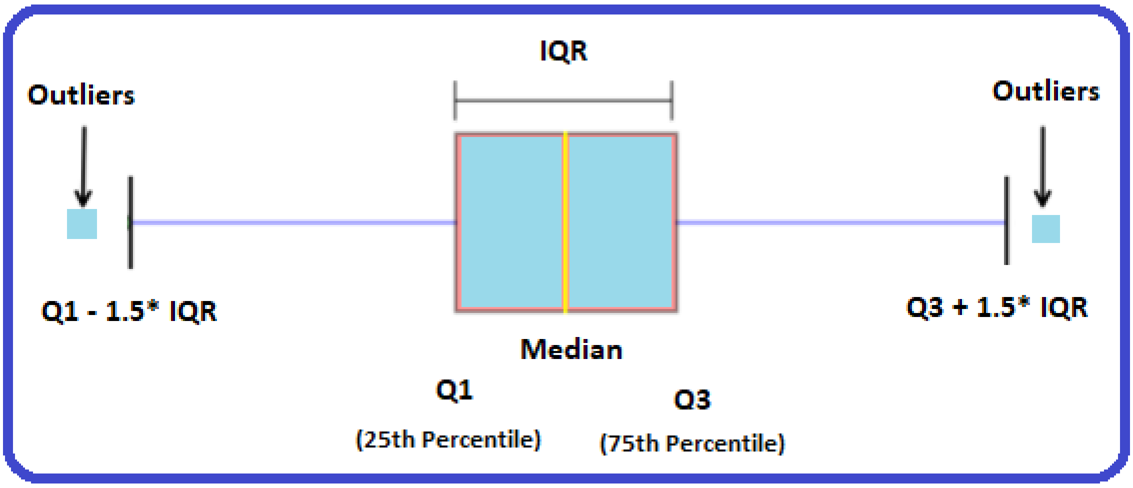


Figure 3: IQR concept

- y, z, price and carat are right skewed with skewness more than 1
- Depth looks to have a symmetric distribution

3.3 Bivariate Analysis

3.3.1 Dependent variable – Independent categorical variable

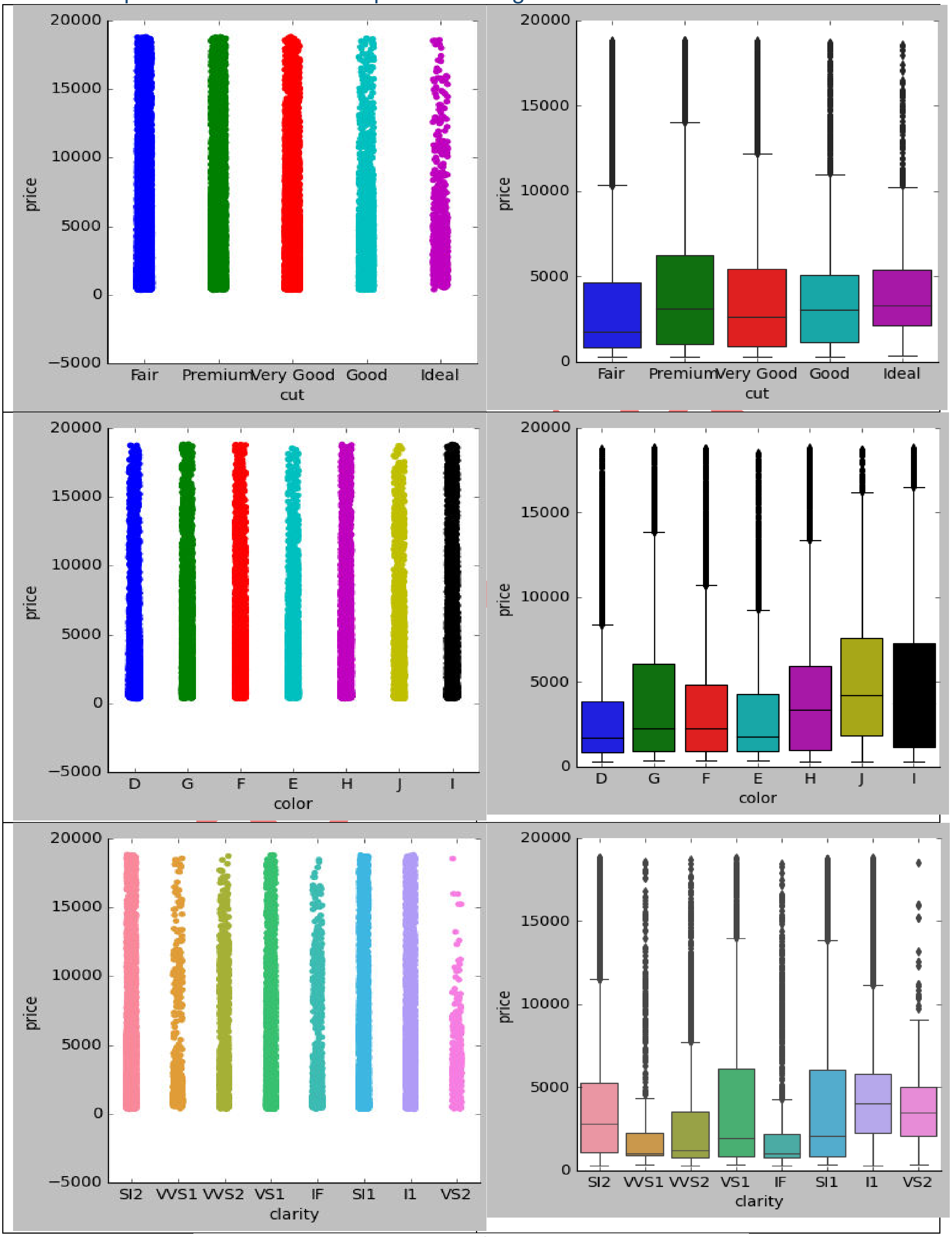
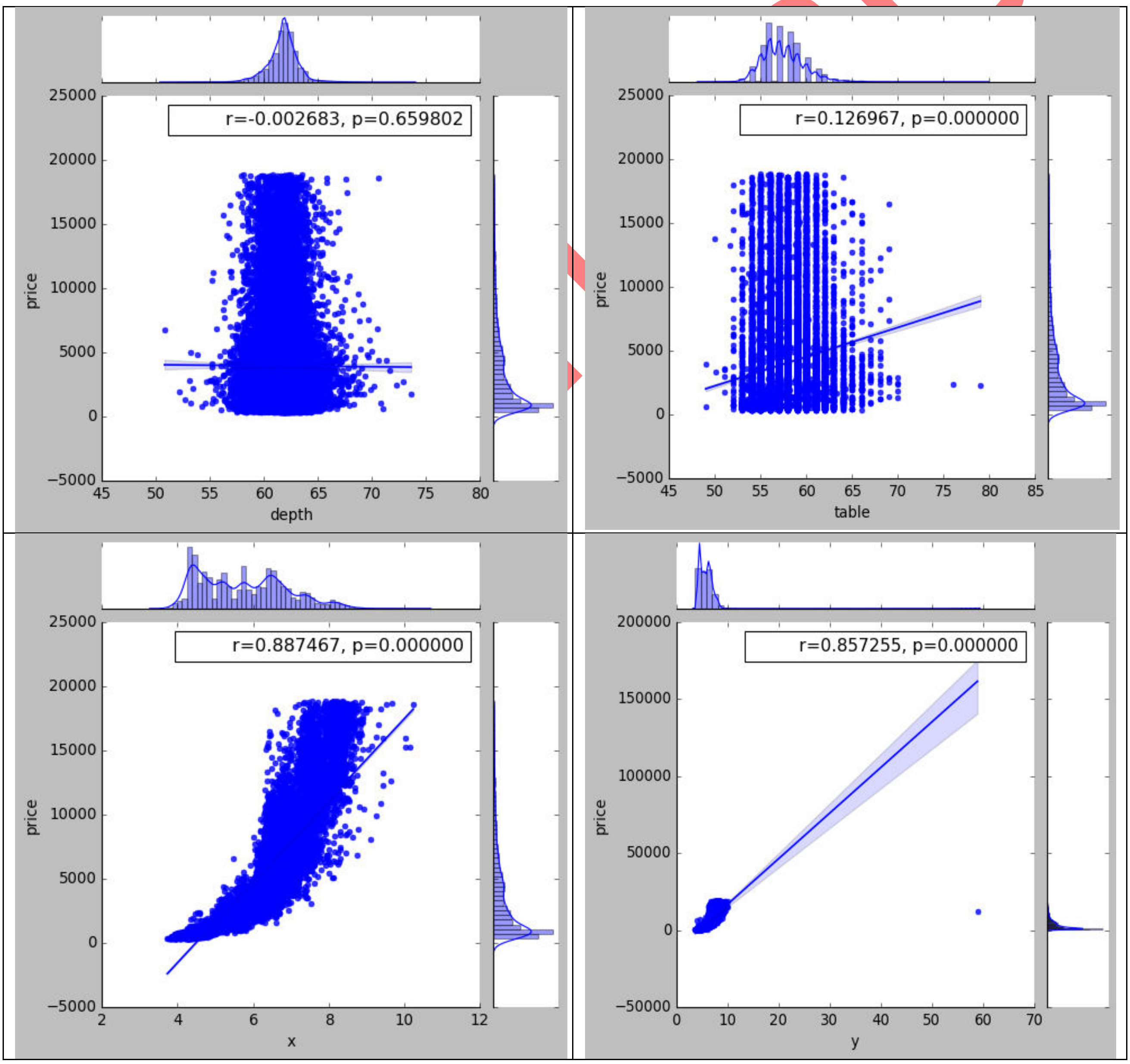


Figure 4: Bivariate analysis (Price - Categorical Independent Variable)

Key Observations:

- If we see box plot of variable cut vs price, the range and median of price of all the categories of stone with respect to stone are somewhat same. Good and Ideal have slight different distribution as compared to other category. They are right skewed and have outliers at higher price, which can be seen both in the box plot bot and is more clear in the density plot.
- Different category of color are showing different range and median (can be seen in box plot (Color vs Price) but the difference is small. In density plot of color the difference is not visible that much.
- Among the categorical variable, clarity is showing a strong predictor as compared to others. Both the plots are suggesting the same.

3.3.2 Dependent variable – Independent continuous variable



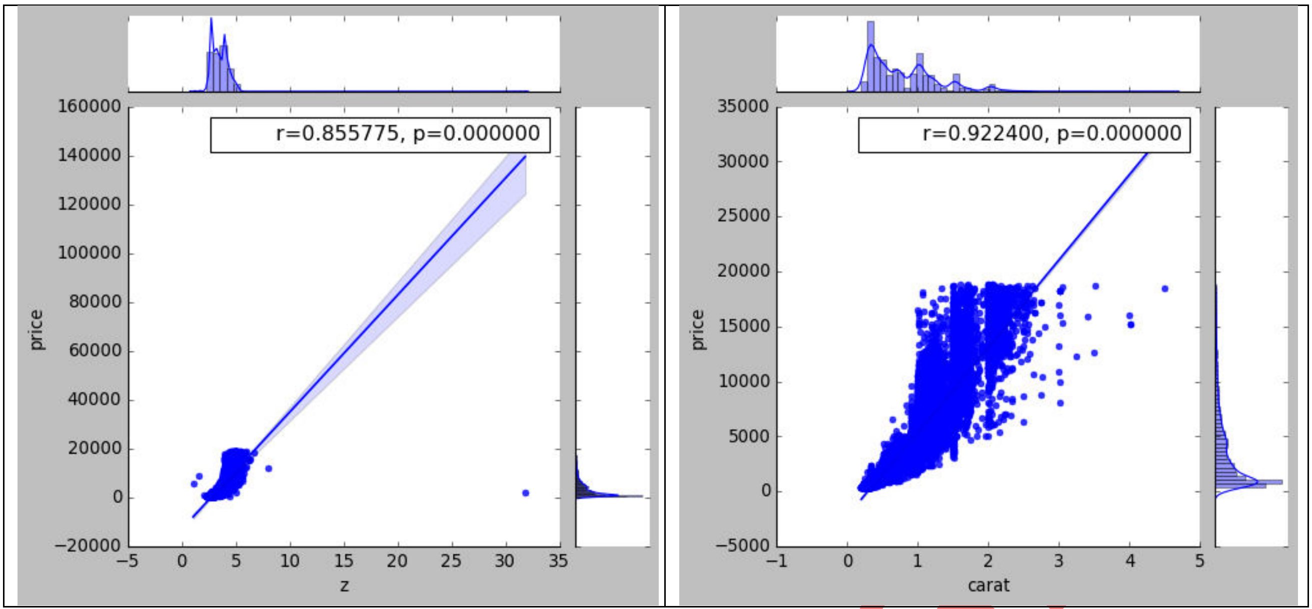


Figure 5: Bivariate analysis (Price – Continuous Independent Variable)

Key Observations:

- Depth vs Price – The Pearson correlation coefficient of 0.002 suggest that there is hardly any correlation between depth and variable and with it high p-value of 0.659 suggest that the probability that the correlation between them in the sample data occurred by chance.
- Table vs Price – Very weak correlation found between Table and Price as Pearson correlation coefficient of 0.12 and the p-value is zero that suggest this correlation is not limited to chance.
- X ,Y,Z vs Price – As can be seen above X, Y and Z are highly correlated to Price.(High Pearson correlation and p-value = 0)
- Carat vs Price – Carat is having the maximum Pearson correlation coefficient (0.92) with price (target variable).
- As we can see from the plot, there are 2 records where y & z value is too high as compared to other values. These could be responsible for pulling the regression line down a bit towards the extreme value. So we will get rid of these records.

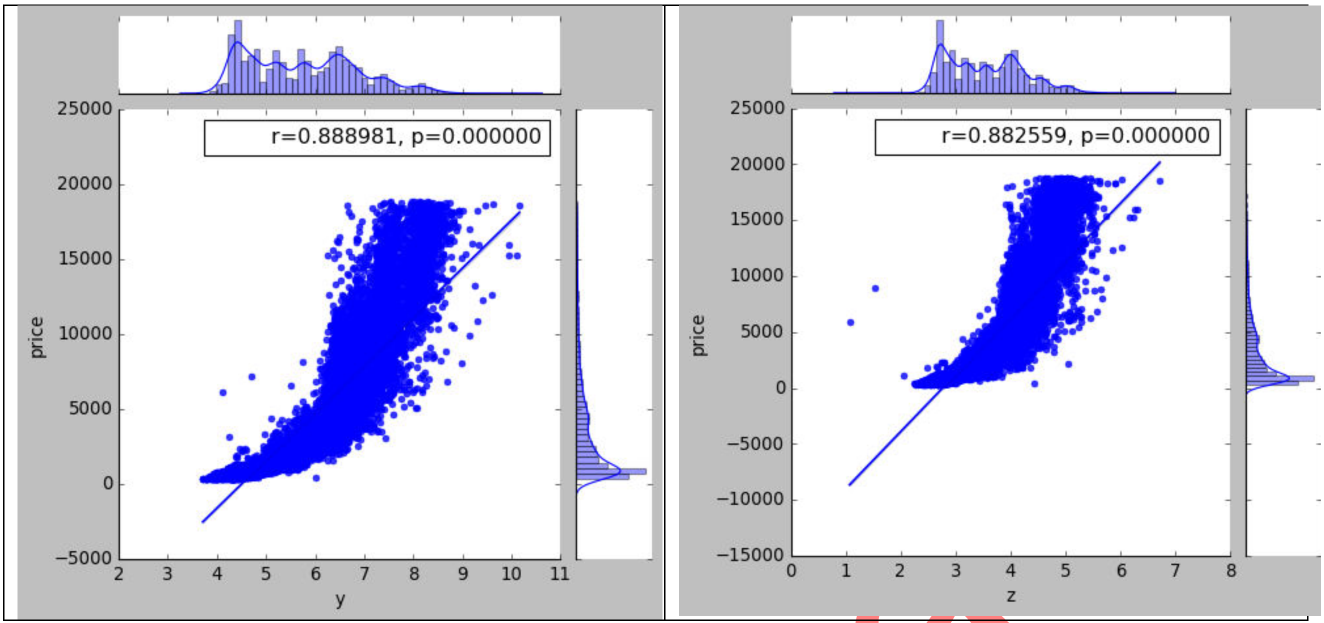


Figure 6: Feature 'y' & 'z' after removing the extreme values

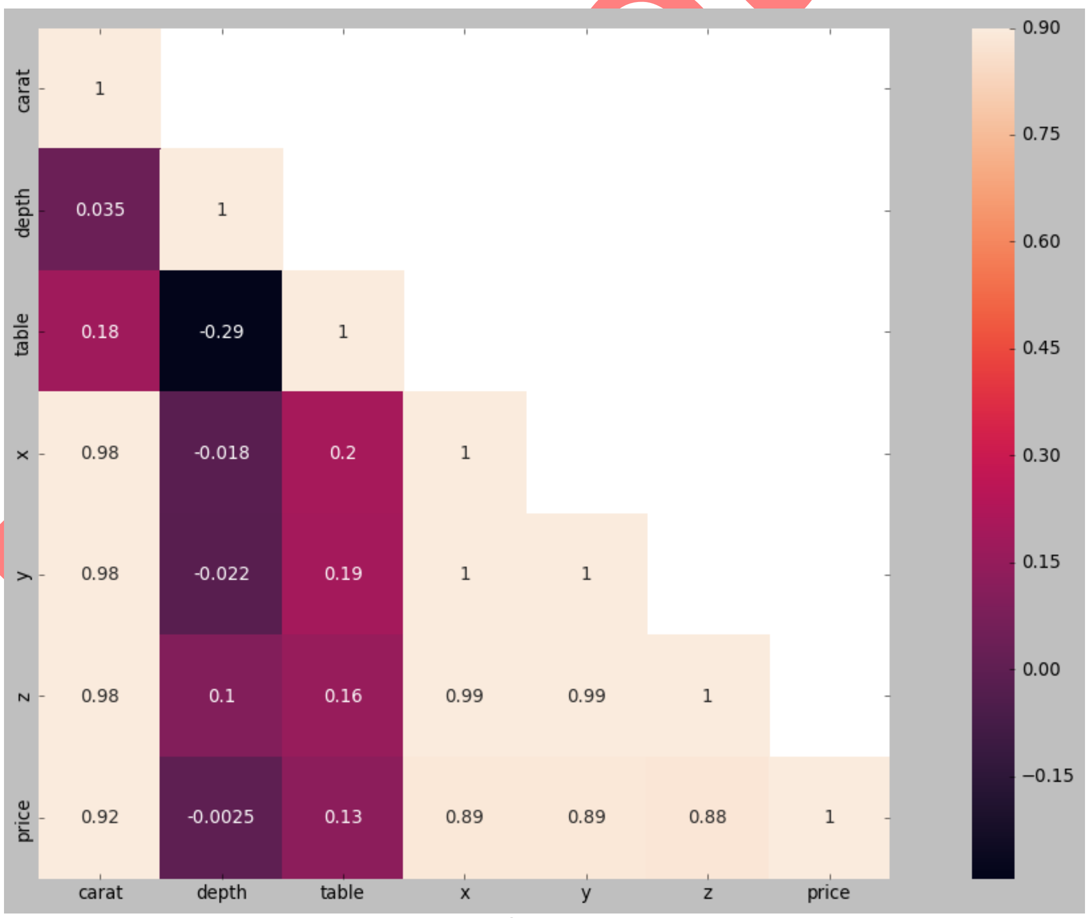


Figure 7: Correlation Matrix

Key Observations:

- X, Y, Z, Carat and Price are highly correlated to each other (Can cause multicollinearity in the model if used together)
- Depth and table don't have strong correlation with any dependent or with target variable

Checking the scope of merging ordinal categories together

- The difference between the mean values for **D & E** as well as between **J & I** look comparative low. Therefore, there is definitely a scope of clubbing these categories into one. Therefore, we will club **D & E to D_E and I & J to I_J.**
- Similarly, the difference between the mean values for **Good** & **Very Good** as well as between **Ideal & Premium** look comparative low. Therefore, there is definitely a scope of clubbing these categories into one. Therefore, we will club **Good & Very Good to Good_Very_Good and Ideal & Premium to Ideal_Premium.**
- Moreover, the difference between the mean values for **SI1 & SI2** as well as between **VS1 & VS2** look comparative low. Therefore, there is definitely a scope of clubbing these categories into one. Therefore, we will club **SI1 & SI2 to SI1_2 and VS1 & VS2 to VS1_2.**

4. MODEL DEVELOPMENT

Encoding Categorical Variable (Cut, Colour and Clarity):
All the categorical variables are ordinal. The order of each variable is linked to their quality and thus to their price. Thus, we have encoded all the categorical variables in the order of their importance/effect on the price.

Model 1: The first model is made using all the variables

The intercept for our model is 4594.23
The coefficient for carat is 10775.43
The coefficient for cut is -51.64
The coefficient for color is -351.786
The coefficient for clarity is -446.18
The coefficient for depth is 46.13
The coefficient for table is -62.78
The coefficient for x is -1651.38
The coefficient for y is 2612.5
The coefficient for z is -3104.54

Model Performance

R^2 Train	0.8886
R^2 Test	0.9003
RMSE Train	1338
RMSE Test	1278

Table 8: Model 1 performance measure

Key Observations:

As we had seen, some of the independent variables are correlated and thus we can see they are causing problem of multicollinearity in the model. The above model coefficients also indicating the problem of multicollinearity.

1. x is positively correlated to price but coefficient is negative
2. z is positively correlated to price but coefficient is negative
3. Table is positively correlated to price but coefficient is negative

OLS output:

```
                    OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.889
Model:                            OLS   Adj. R-squared:                  0.889
Method:                 Least Squares   F-statistic:                 1.670e+04
Date:                Fri, 20 Aug 2021   Prob (F-statistic):               0.00
Time:                        16:25:28   Log-Likelihood:            -1.6241e+05
No. Observations:               18846   AIC:                         3.248e+05
Df Residuals:                   18836   BIC:                         3.249e+05
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
```

```
------------------------------------------------------------------------------
Intercept   4594.2323   1347.732     3.409     0.001    1952.557    7235.907
carat       1.078e+04    102.099   105.539     0.000    1.06e+04     1.1e+04
cut          -51.6383     14.761    -3.498     0.000     -80.572     -22.705
color       -351.7855      7.179   -48.999     0.000    -365.858    -337.713
clarity     -446.1803      7.901   -56.473     0.000    -461.667    -430.694
depth         46.1268     20.585     2.241     0.025       5.778      86.475
table        -62.7799      5.506   -11.401     0.000     -73.573     -51.987
x          -1651.3818    201.462    -8.197     0.000   -2046.265   -1256.498
y           2612.4986    204.970    12.746     0.000    2210.738    3014.259
z          -3104.5440    320.381    -9.690     0.000   -3732.520   -2476.568
==============================================================================
Omnibus:                     4890.805   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           273957.515
Skew:                          -0.384   Prob(JB):                         0.00
Kurtosis:                      21.663   Cond. No.                     1.20e+04
==============================================================================
```

From the above summary of the model, we can see that all of the features are significant as none of the features have pvalues > 0.05.

In addition, the output from the sklearn's Linear Regression & statsmodel's OLS are similar. Therefore, we will continue using sklearn's Linear Regression model for further analysis.

<u>Model 2 (Model using carat and price)</u> – As we had seen carat will be a strong predictor variable for price, therefore it makes sense to make a model keeping only carat as predictor.

The intercept for our model is -2249.62
The coefficient for carat is 7758.38

Model Performance

R^2 Train	0.8477
R^2 Test	0.8577
RMSE Train	1564
RMSE Test	1527

Table 9: Model 2 performance measure

Comment: As compared to the first model where R^2 was 88.86%, by just using carat we have achieved R^2 84.77%. Thus, the second model with only one variable carat explains 85% of response variable variation, which is just 4% less of full model (including all dependent variable).

OLS output:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.848
Model:                            OLS   Adj. R-squared:                  0.848
Method:                 Least Squares   F-statistic:                 1.049e+05
Date:                Wed, 15 Sep 2021   Prob (F-statistic):               0.00
Time:                        15:27:26   Log-Likelihood:            -1.6536e+05
No. Observations:               18846   AIC:                         3.307e+05
Df Residuals:                   18844   BIC:                         3.307e+05
Df Model:                           1
Covariance Type:            nonrobust
```

```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -2249.6171     22.225   -101.221      0.000   -2293.180   -2206.054
carat       7758.3843     23.949    323.955      0.000    7711.442    7805.326
==============================================================================
Omnibus:                     4877.955   Durbin-Watson:                   1.996
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            51969.901
Skew:                           0.938   Prob(JB):                         0.00
Kurtosis:                      10.916   Cond. No.                         3.64
==============================================================================
```

Model 3 (Model using carat, cut, colour, clarity and price): We are excluding depth and table because of the poor relationship with price found in EDA and we are excluding X,Y and Z as they are highly collinear with carat and will cause multicollinearity as we have seen in our first model.

The intercept for our model is 516.99
The coefficient for carat is 8539.17
The coefficient for cut is -167.04
The coefficient for color is -346.59
The coefficient for clarity is -483.37

Model Performance

R^2 Train	0.8833
R^2 Test	0.8932
RMSE Train	1370
RMSE Test	1322

Table 10: Model 3 performance measure

Comment: R^2 and RMSE are very close to full model. Thus, we have removed variables without affecting the model performance.

OLS output:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.883
Model:                            OLS   Adj. R-squared:                  0.883
Method:                 Least Squares   F-statistic:                 3.566e+04
Date:                Wed, 15 Sep 2021   Prob (F-statistic):               0.00
Time:                        15:27:26   Log-Likelihood:            -1.6285e+05
No. Observations:               18846   AIC:                         3.257e+05
Df Residuals:                   18841   BIC:                         3.258e+05
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    516.9915     42.837     12.069      0.000     433.027     600.956
carat       8539.1694     23.429    364.475      0.000    8493.247    8585.092
cut         -167.0450     12.461    -13.406      0.000    -191.469    -142.621
color       -346.5907      7.332    -47.271      0.000    -360.962    -332.219
clarity     -483.3701      7.957    -60.747      0.000    -498.967    -467.774
==============================================================================
Omnibus:                     3609.261   Durbin-Watson:                   1.994
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           109049.162
```

```
Skew:                     0.083   Prob(JB):                  0.00
Kurtosis:                14.783   Cond. No.                  25.4
==============================================================================
```

<u>Model 4 (Model using carat, cut, colour, clarity, depth, table and price):</u> Including all variable excluding x, y and z, which are correlated to carat

The intercept for our model is 10086.60
The coefficient for carat is 8568.91
The coefficient for cut is -91.50
The coefficient for color is -344.90
The coefficient for clarity is -477.12
The coefficient for depth is -102.33
The coefficient for table is -60.05

Model Performance

R^2 Train	0.8846
R^2 Test	0.8947
RMSE Train	1361
RMSE Test	1313

Table 11: Model 4 performance measure

Comment: This model was made just to compare how this combination works. As we knew, the additional variables are not helping the model to predict better.

OLS Output:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.885
Model:                            OLS   Adj. R-squared:                  0.885
Method:                 Least Squares   F-statistic:                 2.409e+04
Date:                Wed, 15 Sep 2021   Prob (F-statistic):               0.00
Time:                        15:27:27   Log-Likelihood:            -1.6274e+05
No. Observations:               18846   AIC:                         3.255e+05
Df Residuals:                   18839   BIC:                         3.255e+05
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   1.009e+04    639.883     15.763      0.000    8832.376    1.13e+04
carat       8568.9077     23.432    365.687      0.000    8522.978    8614.837
cut          -91.4964     14.562     -6.283      0.000    -120.038     -62.955
color       -344.9036      7.297    -47.268      0.000    -359.206    -330.601
clarity     -477.1200      7.922    -60.231      0.000    -492.647    -461.593
depth       -102.3330      7.600    -13.464      0.000    -117.231     -87.435
table        -60.0464      5.600    -10.723      0.000     -71.022     -49.070
==============================================================================
Omnibus:                     3614.483   Durbin-Watson:                   1.996
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           108479.626
Skew:                           0.101   Prob(JB):                         0.00
Kurtosis:                      14.752   Cond. No.                     5.45e+03
```

==

Feature Engineering
As the dimensions - X, Y and Z are correlated to each other (We have also seen it in the correlation matrix), so we can replace these 3 features with one single feature – (X*Y*Z), quite close to volume. So let us replace volume feature to our best model (Model 3) with Carat as carat is highly correlated with all these three variables.

Model 5 (Model using cut, colour, clarity, (X*Y*Z) and price): Replacing carat with X*Y*Z in model 3

The intercept for our model is 374.89
The coefficient for cut is -127.32
The coefficient for color is -346.69
The coefficient for clarity is -480.88
The coefficient for X*Y*Z is 52.91

Model Performance

R^2 Train	0.8864
R^2 Test	0.8946
RMSE Train	1351
RMSE Test	1313

Table 12: Model 5 performance measure

Comment: The model seems quite stable. Can't see big difference in train and test result as compared to the model where we were using carat in place of X*Y*Z

OLS Output:

```
                 OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.886
Model:                            OLS   Adj. R-squared:                  0.886
Method:                 Least Squares   F-statistic:                 3.677e+04
Date:                Wed, 15 Sep 2021   Prob (F-statistic):               0.00
Time:                        15:27:28   Log-Likelihood:            -1.6260e+05
No. Observations:               18846   AIC:                         3.252e+05
Df Residuals:                   18841   BIC:                         3.252e+05
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    374.8942     42.249      8.873      0.000     292.082     457.706
cut         -127.3184     12.279    -10.369      0.000    -151.387    -103.250
color       -346.6937      7.232    -47.939      0.000    -360.869    -332.518
clarity     -480.8779      7.846    -61.287      0.000    -496.257    -465.498
X_Y_Z         52.9063      0.143    370.156      0.000      52.626      53.186
==============================================================================
Omnibus:                     3532.118   Durbin-Watson:                   1.994
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            87602.703
Skew:                           0.222   Prob(JB):                         0.00
Kurtosis:                      13.553   Cond. No.                         658.
```

==

Models after applying log transformation

We have seen there are some skewed variables in the model, which may influence our model performance. As we have analysed, all the variables are right skewed. Thus to improve the model performance we apply log transformation and convert them to closer to normal distribution.

Applying log transformation on price (skewness changes from 1.619055 to 0.128091)

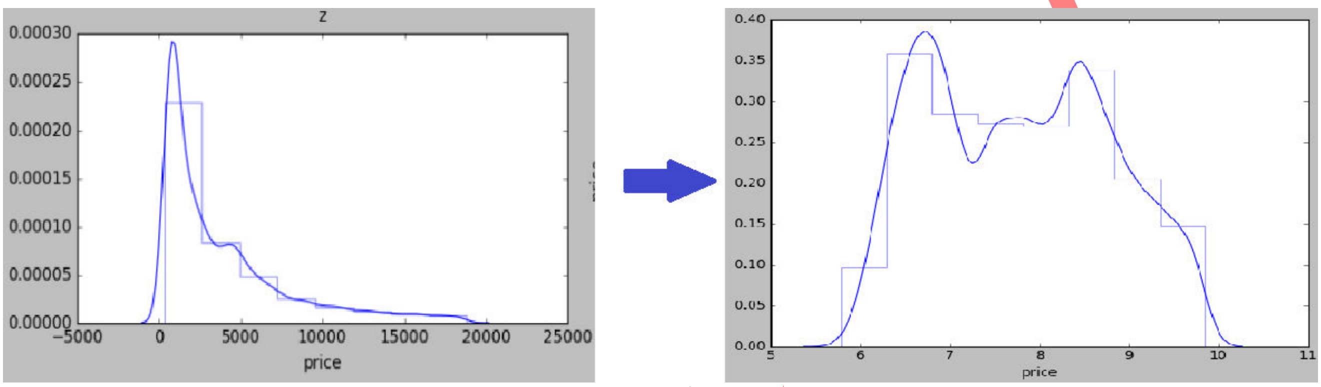


Figure 8: Price log transformation

Applying log transformation on Carat (skewness changes from 1.114871 to 0.104376)

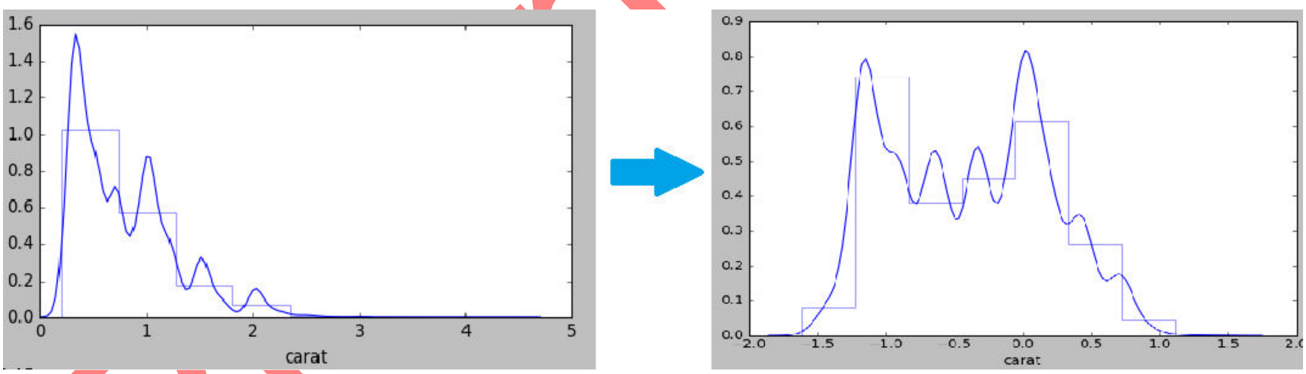


Figure 9: Carat log transformation

Model 6 (Model using log carat, cut, colour, clarity and log price): Applying it on model 3

The intercept for our model is 9.34
The coefficient for carat is 1.84
The coefficient for cut is -0.03
The coefficient for color is -0.09
The coefficient for clarity is -0.12

Model Performance

RMSE Train	1423
RMSE Test	1162
R^2 Train	0.9667

R^2 Test	0.9672

Table 13: Model 6 performance measure

Comment: Prediction capability of model improved significantly.

OLS Output:

```
                         OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.967
Model:                            OLS   Adj. R-squared:                  0.967
Method:                 Least Squares   F-statistic:                 1.369e+05
Date:                Wed, 15 Sep 2021   Prob (F-statistic):               0.00
Time:                        15:27:29   Log-Likelihood:                 4976.9
No. Observations:               18846   AIC:                            -9944.
Df Residuals:                   18841   BIC:                            -9905.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      9.3388      0.007   1339.744      0.000       9.325       9.353
carat          1.8365      0.003    709.641      0.000       1.831       1.842
cut           -0.0300      0.002    -17.755      0.000      -0.033      -0.027
color         -0.0863      0.001    -87.578      0.000      -0.088      -0.084
clarity       -0.1214      0.001   -110.956      0.000      -0.124      -0.119
==============================================================================
Omnibus:                     4076.390   Durbin-Watson:                   1.993
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            20403.879
Skew:                          -0.960   Prob(JB):                         0.00
Kurtosis:                       7.722   Cond. No.                         30.8
==============================================================================
```

<u>Model 7 (Model using cut, colour, clarity, log(x*y*z) and log price):</u> Applying it on model 3

The intercept for our model is -0.12
The coefficient for cut is -0.02
The coefficient for color is -0.09
The coefficient for clarity is -0.12
The coefficient for X*Y*Z is 1.85

Model Performance

RMSE Train	1393
RMSE Test	1158
R^2 Train	0.9667
R^2 Test	0.9665

Table 14: Model 7 performance measure

Comment: The model is quite stable. The model performance for train data is similar as compared to our earlier models but at the same time, the model is performing very well for the test data or the unseen data. Which means the model is reliable and its performance can be predictable when the model is deployed.

OLS Output:

```
                     OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.967
Model:                            OLS   Adj. R-squared:                  0.967
Method:                 Least Squares   F-statistic:                 1.368e+05
Date:                Wed, 15 Sep 2021   Prob (F-statistic):               0.00
Time:                        15:27:29   Log-Likelihood:                 4968.1
No. Observations:               18846   AIC:                            -9926.
Df Residuals:                   18841   BIC:                            -9887.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.1171      0.011    -10.532      0.000      -0.139      -0.095
cut           -0.0186      0.002    -11.016      0.000      -0.022      -0.015
color         -0.0862      0.001    -87.421      0.000      -0.088      -0.084
clarity       -0.1196      0.001   -109.403      0.000      -0.122      -0.118
X_Y_Z          1.8511      0.003    709.295      0.000       1.846       1.856
==============================================================================
Omnibus:                     3702.112   Durbin-Watson:                   1.991
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            27565.356
Skew:                          -0.747   Prob(JB):                         0.00
Kurtosis:                       8.734   Cond. No.                         62.1
==============================================================================
```

<u>Model 8 (Model using low VIF values):</u> Model with only those features that have low VIF values to check for multicollinearity.

VIF values:

variables	VIF
carat	2168.870401
cut	9.475077
color	4.969718
clarity	14.238904
depth	1041.604398
table	938.780952
x	11058.672177
y	11457.113484
z	3385.152611
X*Y*Z	2339.900662

Table 15: VIF values

Model using only low vif value features:
The intercept for our model is 11.82
The coefficient for cut is 413.69
The coefficient for color is 478.83
The coefficient for clarity is 419.13

Model Performance

RMSE Train	3886
RMSE Test	3910
R^2 Train	0.0605
R^2 Test	0.0663

Table 16: Model 8 performance measure

Comment: We can see that the model has become quite unstable post dropping the variables with high multicollinearity. To get rid of the multicollinearity, we can also look at how we can apply PCA for dimensionality reduction

OLS Output:

```
                    OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.061
Model:                            OLS   Adj. R-squared:                  0.060
Method:                 Least Squares   F-statistic:                     404.9
Date:                Wed, 15 Sep 2021   Prob (F-statistic):          6.60e-255
Time:                        15:27:30   Log-Likelihood:            -1.8251e+05
No. Observations:               18846   AIC:                         3.650e+05
Df Residuals:                   18842   BIC:                         3.651e+05
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     11.8185    121.478      0.097      0.922    -226.290     249.927
cut          413.6870     35.065     11.798      0.000     344.957     482.417
color        478.8334     19.786     24.201      0.000     440.051     517.615
clarity      419.1271     21.456     19.535      0.000     377.072     461.182
==============================================================================
Omnibus:                     5300.525   Durbin-Watson:                   1.995
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            12260.300
Skew:                           1.607   Prob(JB):                         0.00
Kurtosis:                       5.298   Cond. No.                         25.1
==============================================================================
```

5. MODEL PERFORMANCE

Model	Predictors	Target	R^2 Train	R^2 Test	RMSE Train	RMSE Test	Adj. R^2
Model 1	carat, cut, colour, clarity, depth, table, x, y, z	Price	0.889	0.900	1338	1278	0.889
Model 2	carat	Price	0.848	0.858	1564	1527	0.848
Model 3	carat, cut, colour, clarity	Price	0.883	0.893	1370	1322	0.883
Model 4	carat, cut, colour, clarity, depth, table	Price	0.885	0.895	1361	1313	0.885
Model 5	cut, colour, clarity, x*y*z	Price	0.886	0.895	1351	1313	0.886
Model 6	Log carat, cut, colour, clarity	Log Price	0.967	0.967	1423	1162	0.967
Model 7	cut, colour, clarity, Log (x*y*z)	Log Price	0.967	0.967	1393	1158	0.967
Model 8	cut, color, clarity	Price	0.061	0.066	3886	3910	0.060

Table 27: Model performance comparison

Model Selection

Model for prediction

If we compared all the models than Model 4 and Model 5 are doing better with respect to prediction. To select best among them it would be better to have more data for training, validating and testing. As of now, Model 5 looks to be more balance. However, Model 4 is not bad as it is giving similar result but will require further analysis with more data.

Model for prescriptive analysis

Model 5 with only 4 independent variable is most suitable. Model performance very close to full model. It is the simplest model with no transformation and with least variable. There is no multicollinearity, as the independent variable in the model are not correlated among each other as observed in EDA section.

Model 5 (Model using cut, colour, clarity, (X*Y*Z) and price): Replacing carat with X*Y*Z in model 3

The intercept for our model is 374.89
The coefficient for cut is -127.32
The coefficient for color is -346.69
The coefficient for clarity is -480.88
The coefficient for X*Y*Z is 52.91

6. CONCLUSION

- The important feature for price prediction of zirconia stone from the data set provided is coming out to be x*y*z, cut, clarity and color. Among them x*y*z is dominating price prediction.
- Would advise to work with more variable and data to get better and stable model.
- Before using this model, full fledge testing is of the model is advised.
- Looking at the heat map, variable cut is not playing any role in price determination. The company needs to look into it in detail. Is this phenomenon specific to company or it a general phenomenon. For that, the company needs to take data from the market and see similar trends are there or not. If they do not find similar trends than they have to find why cut is not adding value to price of stone.
- High dependency of price on carat also needs to be analyzed in detail with help of subject matter expert.
- For prediction we will choose model 5 as analyzed in model performance section
- Also for prescriptive analysis, model 5 is most suitable.
- Coefficient X*Y*Z of model 5 is positive. Therefore, with increase in the volume of stone price increases

Cause of concern

- Cut, color and clarity are quality of stone. Which means that the company is not able to demand any premium from the market for this product on basis of its artisanship, brand in the market and service quality. Which should be cause of concern for the company.

Short-term strategy

- The company need to work on these three features (cut, color and clarity) to increase the revenue and focus less on other parameters, which are not able to influence price.

Long-term strategy

- The company needs to find other features with which they can influence the price better way and increase their profitability. The cut may not look a differentiator in this data but if a company establishes a brand in the market than rather demanding price on cost of production the company can demand premium on the quality of workmanship and service they provide.
- Further market research needs to be done to see how competitors are doing.