**CUBIC ZIRCONIA PRICE PREDICTION CASE STUDY**

**SUBMITTED BY: RAGHAV GUPTA**

Table of Contents

[Problem 1 : Cubic Zirconia Price Prediction Case Study 4](#_Toc72760222)

[Problem Statement: 4](#_Toc72760223)

[Data Description: 4](#_Toc72760224)

[Domain: 4](#_Toc72760225)

[Context: 4](#_Toc72760226)

[Attribute Information: 4](#_Toc72760227)

[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. 5](#_Toc72760228)

[Basic EDA summary:- 5](#_Toc72760229)

[Univariate Analysis 5](#_Toc72760230)

[Summary 5](#_Toc72760231)

[Histogram 6](#_Toc72760232)

[Box Plots 7](#_Toc72760233)

[Log transformation of price 9](#_Toc72760234)

[Bivariate and Multivariate analysis 10](#_Toc72760235)

[Correlation matrix 12](#_Toc72760236)

[Heat Map 13](#_Toc72760237)

[Pairplots 14](#_Toc72760238)

[VIF Checking for Multicollinearity 15](#_Toc72760239)

[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? 17](#_Toc72760240)

[Is Scaling necessary 18](#_Toc72760241)

[1.3 Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Linear regression. Performance Metrics: Check the performance of Predictions on Train and Test sets using Rsquare, RMSE. 19](#_Toc72760242)

[Splitting data into training and test set 19](#_Toc72760243)

[Linear Regression using Sklearn 19](#_Toc72760244)

[Linear Regression using statsmodels 20](#_Toc72760245)

[PCA technique 21](#_Toc72760246)

[Linear Regression Model after PCA 23](#_Toc72760247)

[Running other regression models 24](#_Toc72760248)

[Analysing top 5 features visually 28](#_Toc72760249)

[K-Mean clustering for differentiating high and low price cubic zirconia 29](#_Toc72760250)

[Conclusion 31](#_Toc72760251)

[1.4 Inference: Basis on these predictions, what are the business insights and recommendations. 32](#_Toc72760252)

# Problem 1 : Cubic Zirconia Price Prediction Case Study

## ****Problem Statement:****

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

## Data Description:

The dataset contains 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 details can be found in the following csv file: cubic\_zirconia.csv

## Domain:

Manfucturing (cubic Zirconia)

## Context:

The company is earning different profits on different prize slots. We 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, we need to provide them with the best 5 attributes that are most important.

## Attribute Information:

* Carat : Carat weight of the cubic zirconia.
* Cut : Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal.
* Color : Colour of the cubic zirconia.With D being the best and J the worst.
* Clarity : cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst, FL = flawless, I3= level 3 inclusions) FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, I2, I3
* Depth : The Height of a cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter.
* Table : The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter.
* Price : the Price of the cubic zirconia.
* X : Length of the cubic zirconia in mm.
* Y : Width of the cubic zirconia in mm.
* Z : Height of the cubic zirconia in mm.

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

### Basic EDA summary:-

* Data contains 26967 observations and 10 features.
* All columns are numeric variables except cut, color and clarity which are categorical in nature
* There were total of 34 duplicate rows present in the data set (0.12%). These appeared to be genuinely duplicate rows and its since it percentage was very less these duplicates rows were removed from the dataset
* There were 697 rows for which depth attribute has missing value.
* There were 8 rows for which physical dimension of cubic zirconia i.e. length/width/height was 0. Since these dimensions cannot be 0, these rows have missing values.
* From box plot and data summary (shown below in Univariate analysis), it can be seen that there are outliers in all the numeric columns.

### Univariate Analysis

### Summary

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **carat** | **depth** | **table** | **x** | **y** | **z** | **price** |
| **count** | 26933 | 26236 | 26933 | 26933 | 26933 | 26933 | 26933 |
| **mean** | 0.798 | 61.7453 | 57.4559 | 5.7293 | 5.7331 | 3.5378 | 3937.526 |
| **std** | 0.4772 | 1.4122 | 2.2322 | 1.1274 | 1.165 | 0.72 | 4022.552 |
| **min** | 0.2 | 50.8 | 49 | 0 | 0 | 0 | 326 |
| **25%** | 0.4 | 61 | 56 | 4.71 | 4.71 | 2.9 | 945 |
| **50%** | 0.7 | 61.8 | 57 | 5.69 | 5.7 | 3.52 | 2375 |
| **75%** | 1.05 | 62.5 | 59 | 6.55 | 6.54 | 4.04 | 5356 |
| **max** | 4.5 | 73.6 | 79 | 10.23 | 58.9 | 31.8 | 18818 |
| **CV** | 0.6 | NaN | 0.04 | 0.2 | 0.2 | 0.2 | 1.02 |
| **Skew** | 1.11 | -0.03 | 0.77 | 0.39 | 3.87 | 2.58 | 1.62 |
| **IQR** | 0.65 | 1.5 | 3 | 1.84 | 1.83 | 1.14 | 4411 |
| **UR** | 2.03 | 64.75 | 63.5 | 9.31 | 9.29 | 5.75 | 11972.5 |
| **LR** | -0.58 | 58.75 | 51.5 | 1.95 | 1.96 | 1.19 | -5671.5 |

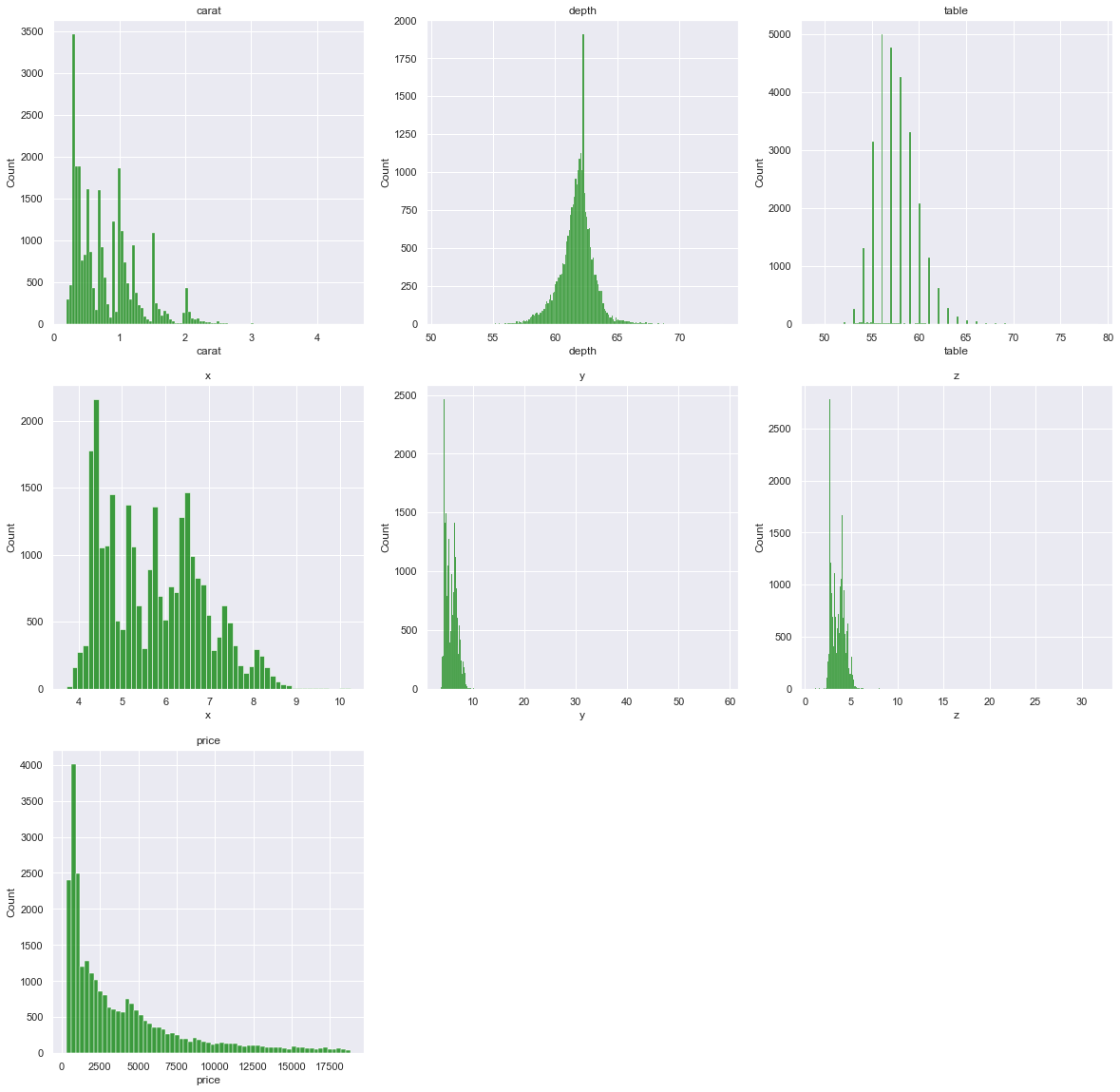
From summary, we can see that :-

* x, y and z have min values 0. These correspond to Length, width and height which cannot be 0 for zirconia diamond.
* Except price, mean and median values appear to be similar for other attributes.
* All the columns are right skewed except depth
* Average price is higher than the middle most value in price. So there appears to be few zirconium which lies in higher price zone as compared to the others.
* Looking at the LR and UR, it appears that these columns do contains outliers.

### Histogram

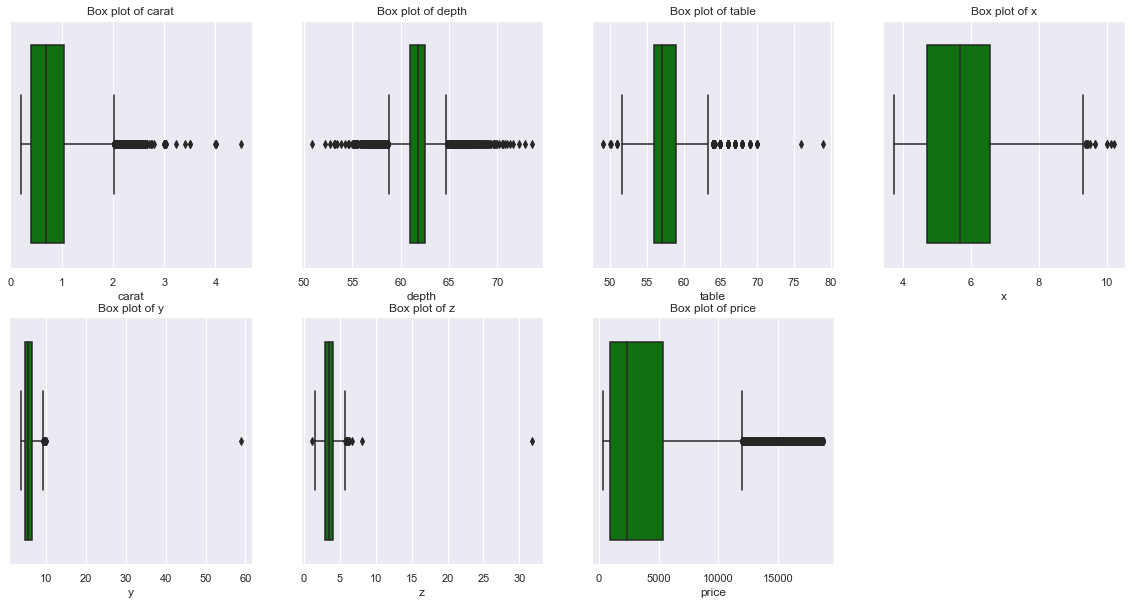
As seen in the histograms (shown below in the figure),

* For majority of the zirconium, carot lies between 0.2 to 1. There is few occurrenc of Above 2 carats diamonds.
* Length(x) of the cubic zirconia (in mm) is varying from 4 to 8
* Width(y) of the cubic zirconia (in mm) is varying from 5 to 8
* Height(z) of the cubic zirconia (in mm) is varying from 2.5 to 5
* Majority of Depth ranges from 60 to 64
* Majority of table value ranges from 55 to 63.



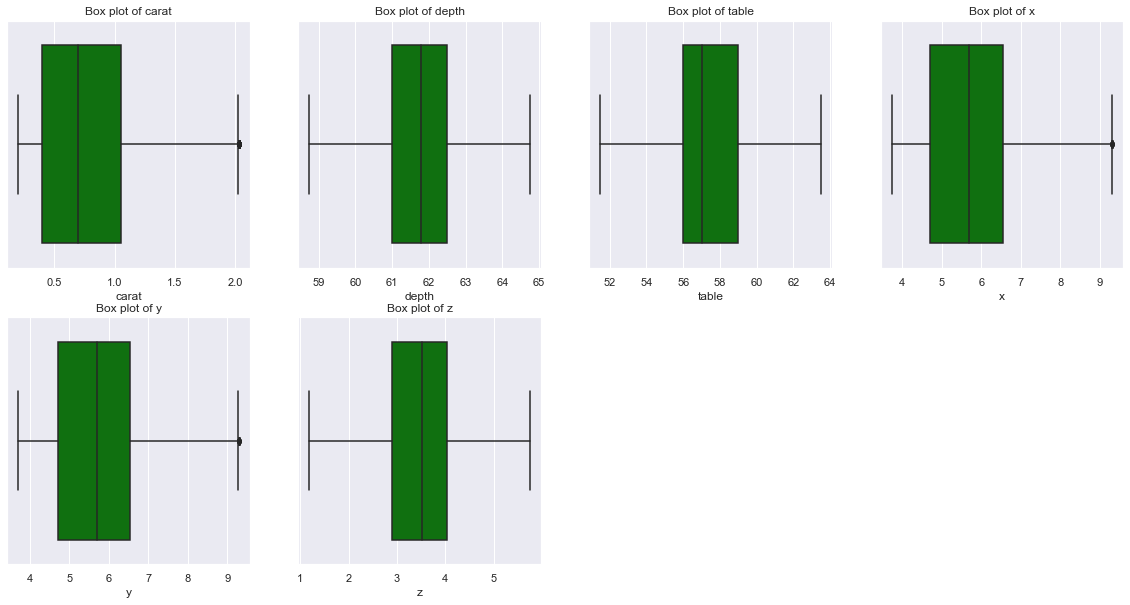
### Box Plots

As evident from the box plots (shown below) , there are outliers in all the numeric columns.

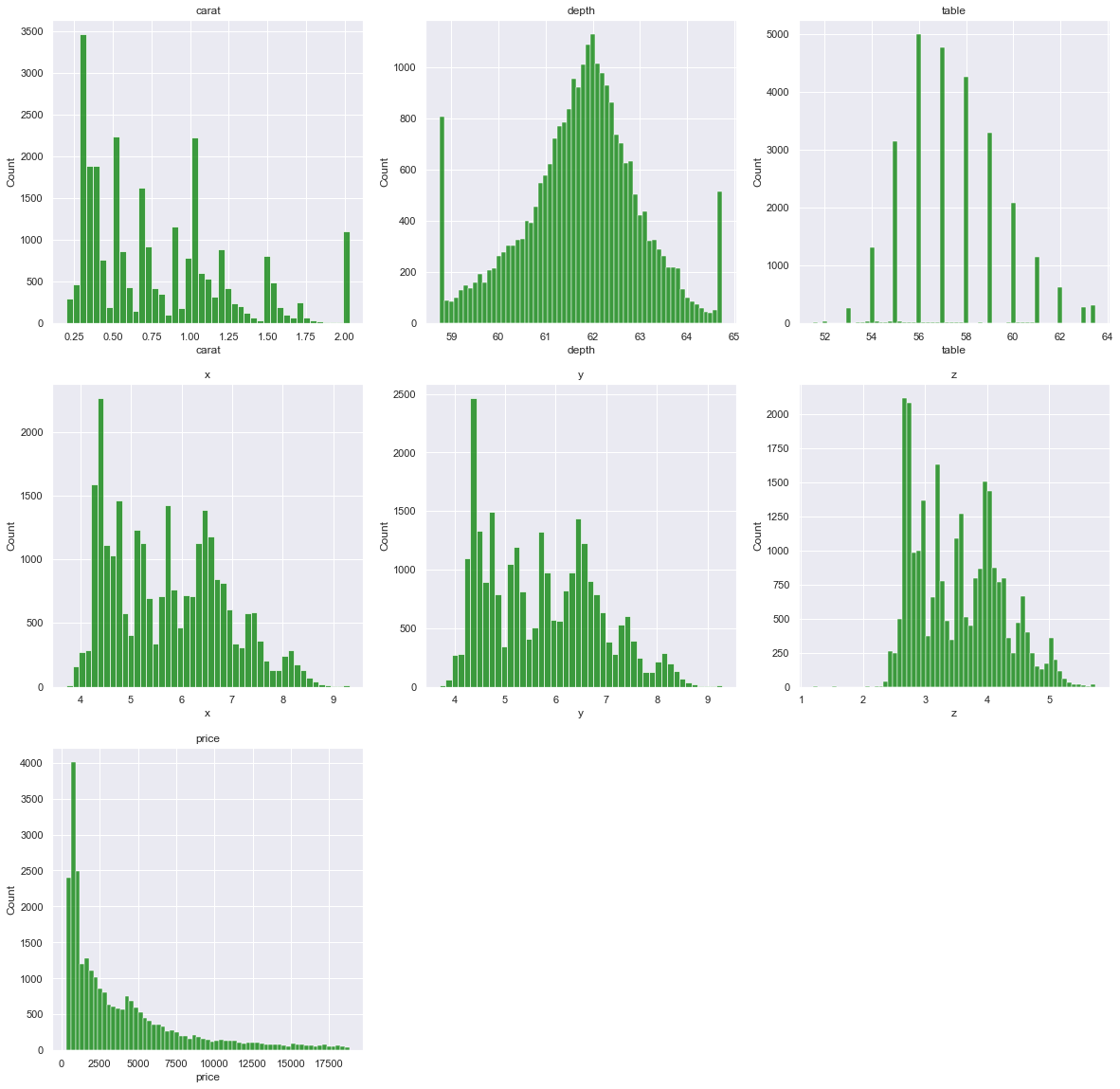
****

We have treated the outliers by capping and flooring technique.

After outlier treatment, box plots and histogram plots are as shown below :-

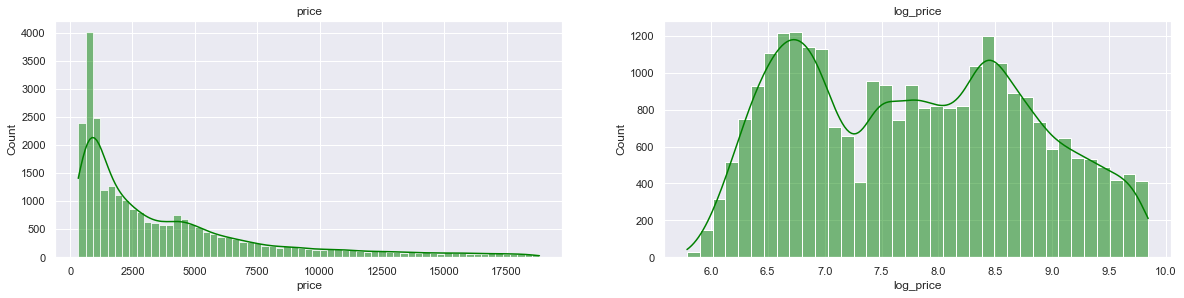


We can see that now there is no outlier present in the numeric columns



### Log transformation of price

Since price columns has outliers, have extreme values on the higher side i.e right skewed. It’s better to proceed with the log transformation of pricecolumn, it would reduce the skewness of the dependent variable.



We can see that although price and log\_price both are not normal but steep value(huge spike) have been reduced in log\_price graph. Data is more evenly distributed in log\_price as compared to price graph.

**Checking Skewness**

Skew of log price :- 0.13

Skew of price :- 1.62

We can see that the Skewness has been reduced after performing log transformation and its very much near to 0 now.

### Bivariate and Multivariate analysis

#### Checking the unique values for categorical variables

Attribute: cut with 5 different values

Fair 780

Good 2435

Very Good 6027

Premium 6886

Ideal 10805

Name: cut, dtype: int64

Attribute color with 7 different values

J 1440

I 2765

D 3341

H 4095

F 4723

E 4916

G 5653

Name: color, dtype: int64

Attribute clarity with 8 values

I1 364

IF 891

VVS1 1839

VVS2 2530

VS1 4087

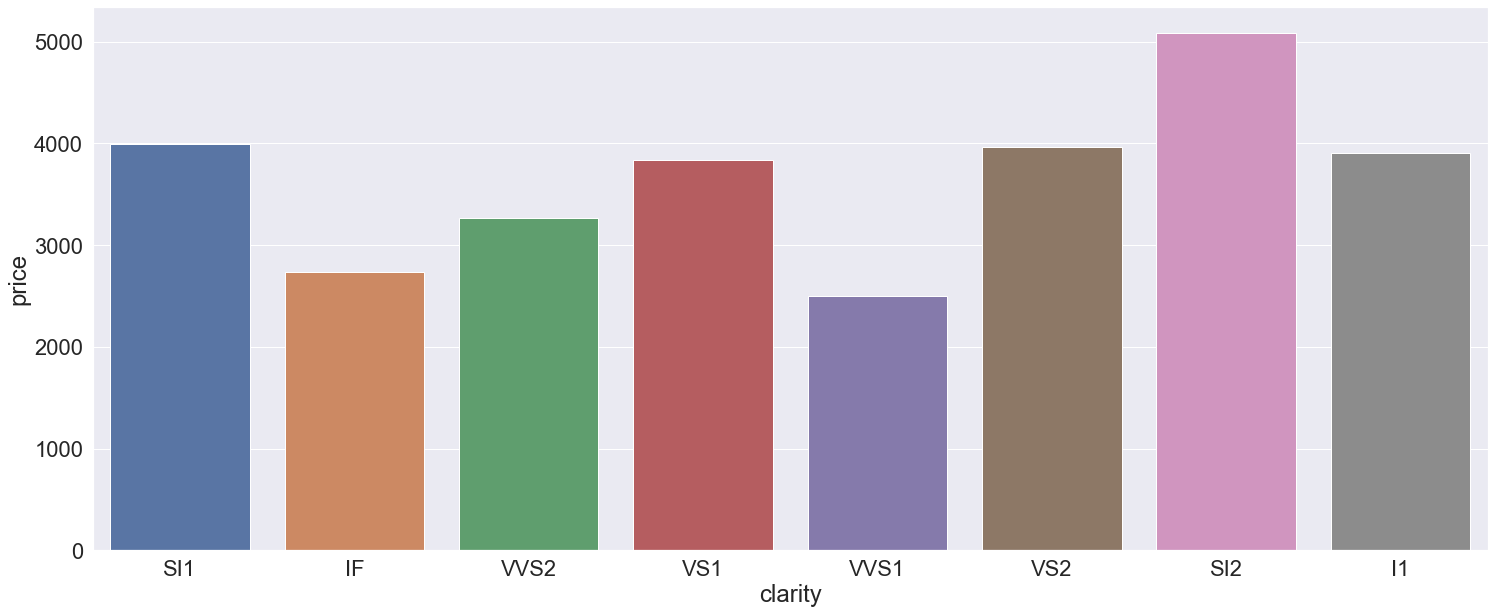
SI2 4564

VS2 6093

SI1 6565

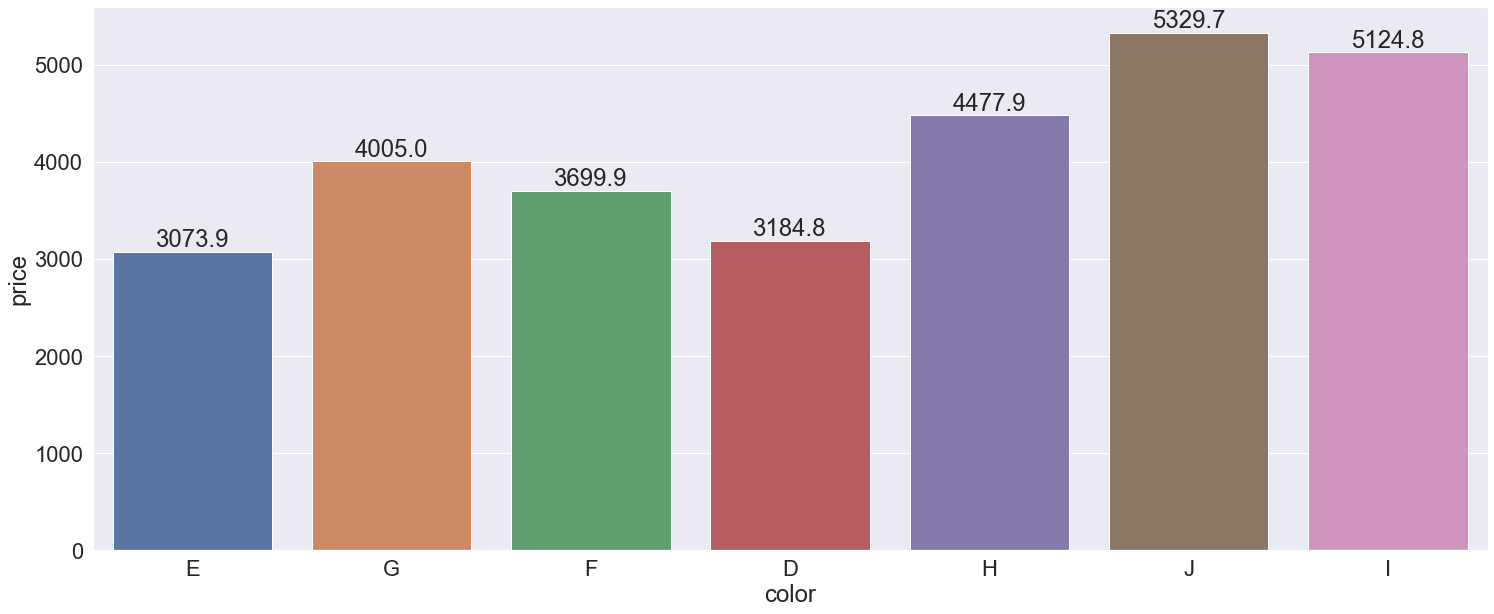
Name: clarity, dtype: int64

#### Bar Plots



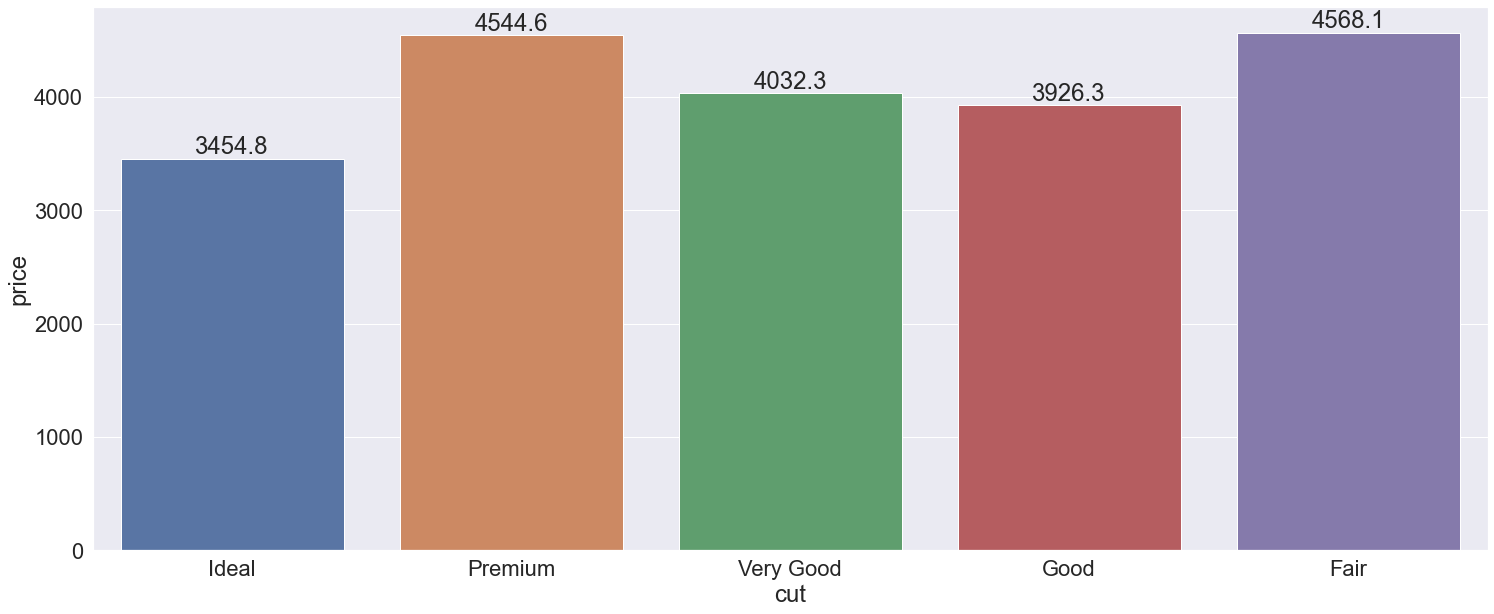
We can see that the mean price level is highest for SI2 clarity and its lowest for VVS1

Important thing to notice here is that IF is second best clarity quality but still its mean price is among the lower side. Best clarity quality 'FL' data is not present.



Price is lowest for color E and its highest for color l

Here too, best color 'D' is also among the lower price range.



Here too , best quality cut 'Ideal' has the lowest mean price

### Correlation matrix

carat depth table x y z price log\_price

carat 1.00 0.03 0.19 0.98 0.98 0.98 0.92 0.93

depth 0.03 1.00 -0.29 -0.02 -0.02 0.09 -0.00 0.00

table 0.19 -0.29 1.00 0.20 0.19 0.16 0.13 0.16

x 0.98 -0.02 0.20 1.00 1.00 0.99 0.89 0.96

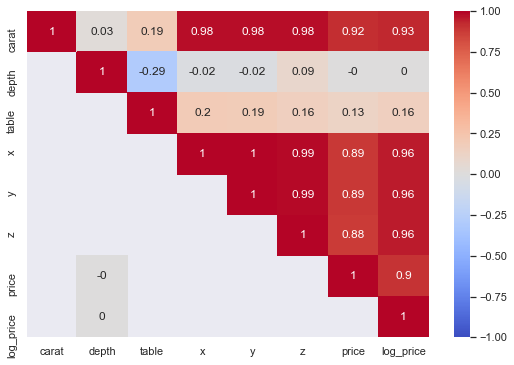
y 0.98 -0.02 0.19 1.00 1.00 0.99 0.89 0.96

z 0.98 0.09 0.16 0.99 0.99 1.00 0.88 0.96

price 0.92 -0.00 0.13 0.89 0.89 0.88 1.00 0.90

log\_price 0.93 0.00 0.16 0.96 0.96 0.96 0.90 1.00

### Heat Map

****

We can see in heatmap & correlation matrix that

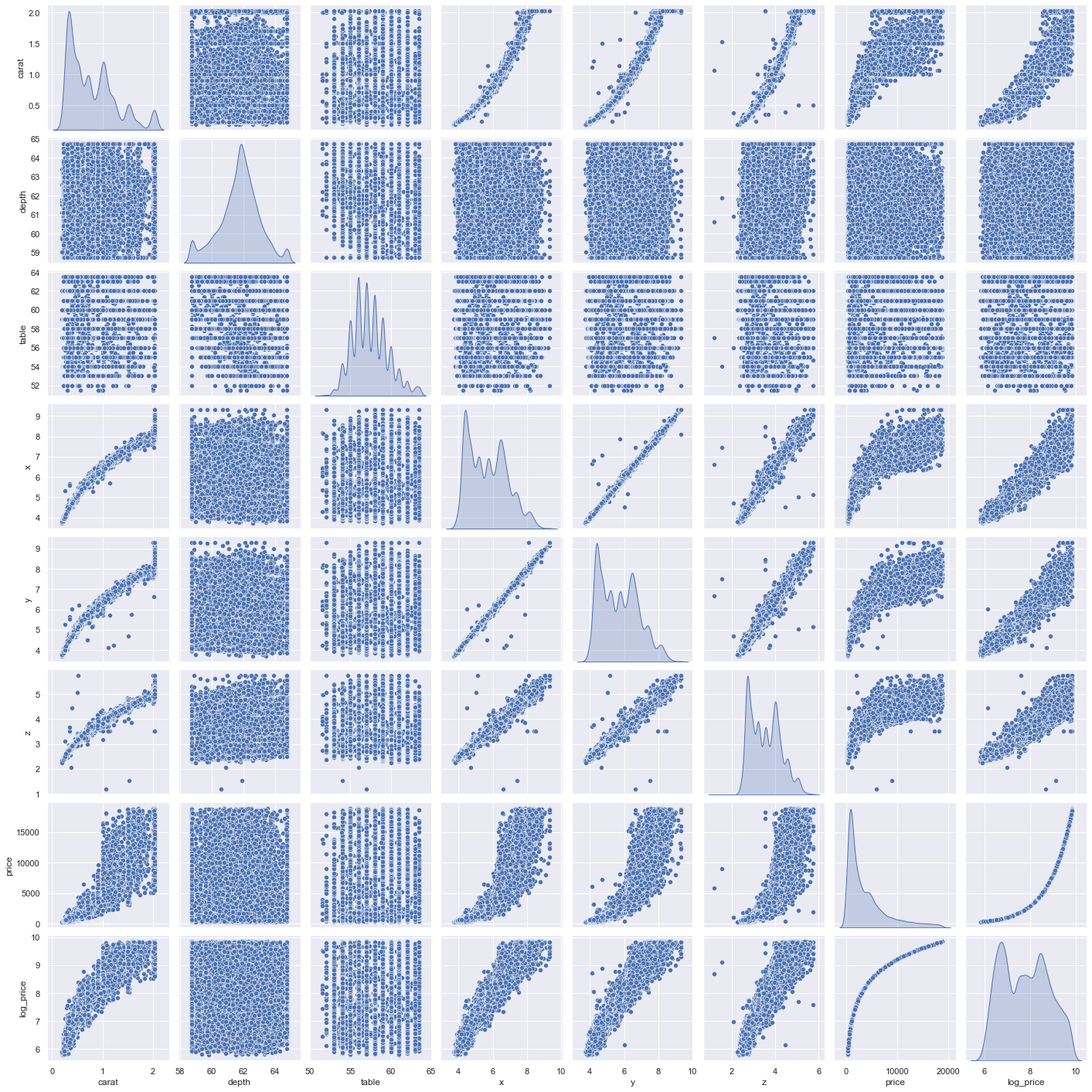
* Price has strong correlation with length , width height , carat.
* There is no correlation of depth on price. It has slighly negative correlation with table attribute
* Carat, length, width, height has strong correlation with each other . This means independent attributes have correlation with each other

Thus this dataset has multicolinearity problem

### Pairplots

To check the correlation of 2 columns in more detail we can draw **pairplots /Scatter**

**Diagram**



As depicted in heat map of correlation matrix, we can see that price increases with increase in carat, x,y,z. Carat weight and length, width and height appears to have some structural correlation. As size of zirconium increases, its carat weight increases and if carat weight increases then its length/width/height should also increase

In distribution of carat, x, y & z there are multiple peaks. So it appears different class of data is present in these columns

Another important thing to notice here is that for table and depth there appears to be no correlation with price. We need to check whether these attributes have evidence to show that there is any relationship of price with these attributes

### VIF Checking for Multicollinearity

Output for VIF is as shown below :-

Variables VIF

0 carat 117.290747

1 table 696.217447

2 x 10214.077576

3 y 9247.015287

4 z 2601.306261

5 depth 931.535862

We can see that VIF is greater than 5 for all the columns. Hence all these independent columns

are highly correlated with each other and this data has a multicollinearity problem

Removing x and then checking the VIF

Variables VIF

0 carat 115.453366

1 table 655.889207

2 y 2021.917644

3 z 2433.025133

4 depth 911.203786

Removing z and then checking the VIF

Variables VIF

0 carat 92.408862

1 table 517.696218

2 y 647.570069

3 depth 526.038654

Removing y and then checking the VIF

Variables VIF

0 carat 4.048380

1 depth 446.207901

2 table 454.949118

Removing table and then checking the VIF

Variables VIF

0 carat 3.946744

1 depth 3.946744

Now we can see that there is no mutlicollinearity in the dataset but all the attributes have been reduced to 2.

So let’s proceed with Linear Regerssion model on this dataset (ignoring the multicollinearity assumption which is most often violated)

## 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?

#### Checking for 0 value in x, y and z

#### carat cut color clarity depth table x y z price

#### 5822 0.71 Good F SI2 64.1 60.0 0.00 0.00 0.0 2130

#### 6035 2.02 Premium H VS2 62.7 53.0 8.02 7.95 0.0 18207

#### 10828 2.20 Premium H SI1 61.2 59.0 8.42 8.37 0.0 17265

#### 12499 2.18 Premium H SI2 59.4 61.0 8.49 8.45 0.0 12631

#### 12690 1.10 Premium G SI2 63.0 59.0 6.50 6.47 0.0 3696

#### 17507 1.14 Fair G VS1 57.5 67.0 0.00 0.00 0.0 6381

#### 18195 1.01 Premium H I1 58.1 59.0 6.66 6.60 0.0 3167

#### 23759 1.12 Premium G I1 60.4 59.0 6.71 6.67 0.0

There are 8 rows for which length/width/height is 0. Since these values cannot be 0 , these are missing values too. Count of these rows is very less. We can either drop them or impute with the median values

#### Imputing missing values

As only Depth , x, y and z column contains the Null values and it contains outliers. We are imputing it with the Median.

We have used sklearn **Simple Imputer to impute the median values to the null values.**

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 carat 26933 non-null float64

1 cut 26933 non-null object

2 color 26933 non-null object

3 clarity 26933 non-null object

4 depth 26933 non-null float64

5 table 26933 non-null float64

6 x 26933 non-null float64

7 y 26933 non-null float64

8 z 26933 non-null float64

9 price 26933 non-null float64

After imputing, we can see that there are no null values

### Is Scaling necessary

Scaling is not required for the linear regression model. Scaling centers the data but it would not affect the accuracy or residual errors. Coefficients/intercepts value may change due to centring of data but error would not be affected by the scaling. Hence its not required here.

## 1.3 Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Linear regression. Performance Metrics: Check the performance of Predictions on Train and Test sets using Rsquare, RMSE.

### Splitting data into training and test set

We have used two methods for applying the linear regression model

1. Sklearn method :- After extracting the Target column, data set is split into 70:30 ratio i.e. 70% of input observations into train data set for building the model and 30% observations into test data set for testing and validating the model.
2. Smstats method:- Without extracting the Target column, data set is split into 70:30 ratio i.e. 70% of input observations into train data set for building the model and 30% observations into test data set for testing and validating the model.

### Linear Regression using Sklearn

After applying the linear regression model using sklearn performance metrics of predictions on train and test data set is as shown below :-

#### R square on training data

0.9763153115042822

#### R square on testing data

0.9759279412402047

#### RMSE on Train data

1763.685852397315

#### RMSE on Testing data

1597.3490579838347

#### Coefficients for each of the independent attributes

The coefficient for carat is -1.0202873340043548

The coefficient for cut is 0.026650203007287158

The coefficient for color is 0.08029557576207384

The coefficient for clarity is 0.12128135115597725

The coefficient for depth is 0.04797794002096829

The coefficient for table is 0.00996335613264138

The coefficient for x is 0.7372589515317397

The coefficient for y is 0.45437413118636044

The coefficient for z is 0.3193892009319381

#### Checking the intercept for the model

The intercept for our model is -3.864234287361503

On checking the output, we have confirmed that there was no negative price prediction for train/test dataset. If we don’t do log transformation of price, then we were getting the negative price predictions for few rows.

### Linear Regression using statsmodels

After applying the linear regression model using sklearn performance metrics of predictions on train and test data set is as shown below :-

OLS Regression Results

==============================================================================

Dep. Variable: log\_price **R-squared: 0.976**

Model: OLS **Adj. R-squared: 0.976**

Method: Least Squares F-statistic: 8.630e+04

Date: Sat, 24 Apr 2021 Prob (F-statistic): 0.00

Time: 01:08:09 Log-Likelihood: 8202.1

No. Observations: 18853 AIC: -1.638e+04

Df Residuals: 18843 BIC: -1.631e+04

Df Model: 9

Covariance Type: nonrobust

==============================================================================

coef std err t P>|t| [0.025 0.975]

------------------------------------------------------------------------------

Intercept -3.8642 0.137 -28.226 0.000 -4.133 -3.596

carat -1.0203 0.014 -71.992 0.000 -1.048 -0.993

cut 0.0267 0.001 21.081 0.000 0.024 0.029

color 0.0803 0.001 113.377 0.000 0.079 0.082

clarity 0.1213 0.001 157.813 0.000 0.120 0.123

depth 0.0480 0.002 25.463 0.000 0.044 0.052

table 0.0100 0.001 14.799 0.000 0.009 0.011

x 0.7373 0.021 34.602 0.000 0.695 0.779

y 0.4544 0.021 21.632 0.000 0.413 0.496

z 0.3194 0.024 13.374 0.000 0.273 0.366

==============================================================================

Omnibus: 4444.444 Durbin-Watson: 1.983

Prob(Omnibus): 0.000 Jarque-Bera (JB): 44217.729

Skew: -0.848 Prob(JB): 0.00

Kurtosis: 10.308 Cond. No. 1.03e+04

==============================================================================

Warnings:

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

[2] The condition number is large, 1.03e+04. This might indicate that there are

strong multicollinearity or other numerical problems.

The overall P value is less than alpha, so rejecting H0 and accepting Ha that atleast 1 regression co-efficient is not 0. Here all regression co-efficient are not 0

Further p value for each coefficient is less than the alpha (0.05), hence all the coefficients are important in deciding the target variables.

#### R square on training data

0.9763153115042822

#### RMSE on Train data

1763.6858523972169

#### RMSE on Testing data

1597.3490579837476

#### Linear Equation

Log price = (-3.86) + (-1.02) \* carat + (0.03) \* cut + (0.08) \* color + (0.12) \* clarity + (0.05) \* depth + (0.01) \* table + (0.74) \* x + (0.45) \* y + (0.32) \* z

When carat increases by 1 unit, log\_price decreases by 0.98 units, keeping all other predictors constant. This is absurd as we have seen in the analysis that price increases with the carat. Coefficient of x ,y, and z is positive which means log price increases with the increase in x,y,z value. These are correlated with each other and carat. Due to high multicollinearity in the data set, as the value of carat will increase, value of x,y and z will increase. Overall effect results in increasing the price. Due to high multicollinearity these coefficients are not stable and there could be multiple coefficient values which will give the same predicted value but the coefficients value would be different. Hence we should not rely on these coefficients.

So I have gone ahead to perform PCA to reduce multicollinearity in the dataset but it would be bit difficult to comprehend the coefficients of the new components created out of PCA.

### PCA technique

#### Scaling

For PCA technique scaling of data is required. Hence we have used Sklearn standardscalar to perform scaling on the data set.

#### Bartlett Test of Sphericity

Output

chi\_square\_value, p\_value : (292921.1425906154, 0.0)

Output of p-value is low here. Hence we shall go with alternative hypothesis. It means that dimension reduction is possible in the dataset given

#### Kaiser-Meyer-Olkin (KMO) Test

kmo\_model : 0.7721386736349162

Overall Measure of Sample Adequacy (MSA) is greater than 0.5. Hence we have adequate sample to proceed with PCA

#### PCA from sklearn's decomposition class and find Principal Components

PCs Proportion Of Variance Standard Deviation Cumulative Proportion

PC1 47.92 2.08 47.92

PC2 15.53 1.18 63.46

PC3 13.53 1.10 76.98

PC4 10.97 0.99 87.96

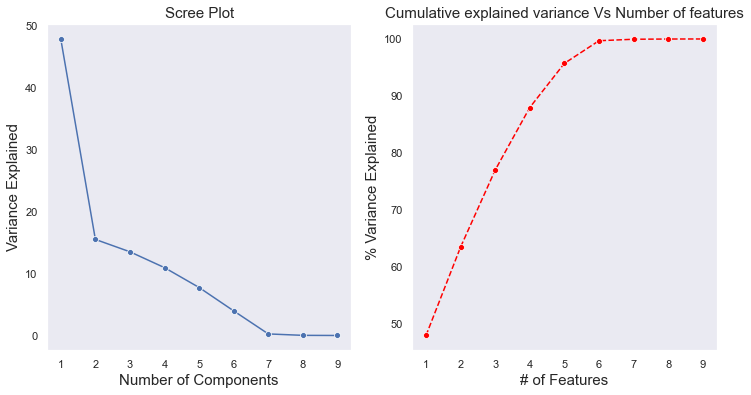
PC5 7.76 0.84 95.71

PC6 3.98 0.60 99.69

PC7 0.26 0.15 99.95

PC8 0.03 0.06 99.99

PC9 0.01 0.04 100.00



From these graphs we can see that if we chose 6 PC's then 99% variance is getting explained and our dimension is getting reduced from 9 to 6.

**Hence we have chosen 6 PC’s to capture 99% variance and removing the multicollinearity in the dataset**

#### Correlation between components and features



We can see that PC1 is a combination of Carat and its length, width , height PC2 is a combination of table and cut

#### Checking correlation of dataset after PCA dimension reduction

PC1 PC2 PC3 PC4 PC5 PC6

PC1 1.0 -0.00 -0.0 -0.00 -0.00 0.00

PC2 -0.0 1.00 0.0 -0.00 0.01 0.00

PC3 -0.0 0.00 1.0 0.00 0.00 -0.00

PC4 -0.0 -0.00 0.0 1.00 -0.00 0.01

PC5 -0.0 0.01 0.0 -0.00 1.00 -0.00

PC6 0.0 0.00 -0.0 0.01 -0.00 1.00

#### Checking VIF

Variables VIF

0 PC1 1.000043

1 PC2 1.000056

2 PC3 1.000020

3 PC4 1.000079

4 PC5 1.000066

5 PC6 1.000086

We can see that VIF is 1 for all the PC's. Hence multicolinearity has been removed.

### Linear Regression Model after PCA

#### R square on training data

0.9598691080691446

#### R square on testing data

0.9579513261268167

#### RMSE on Train data

2127.5550472650098

#### RMSE on Testing data

2157.4699108376485

#### Coefficients for each of the independent attributes

The coefficient for PC1 is 0.4502114007573328

The coefficient for PC2 is -0.11685823373711127

The coefficient for PC3 is -0.08450759301734595

The coefficient for PC4 is -0.059037742690766115

The coefficient for PC5 is 0.3555184644737734

The coefficient for PC6 is -0.02962329575342412

#### Checking the intercept for the model

The intercept for our model is 7.783340582033334

### Running other regression models

We have run the decision tree regressor, Random Forest regressor, Artificial Nuero network MLP regressor for predicting.

Summary of performance metrics of all the models is as shown below :-

Train RMSE Test RMSE Training Score Test Score

Linear Reg bef PCA 1763.685852 1597.349058 0.976315 0.975928

Linear Reg after PCA 2127.555047 2157.469911 0.959869 0.957951

Decision Tree Reg 3.534912 768.168353 0.999995 0.983475

Random Forest Reg 216.739108 568.248918 0.998776 0.991144

ANN Reg 599.872909 638.599335 0.989860 0.988100

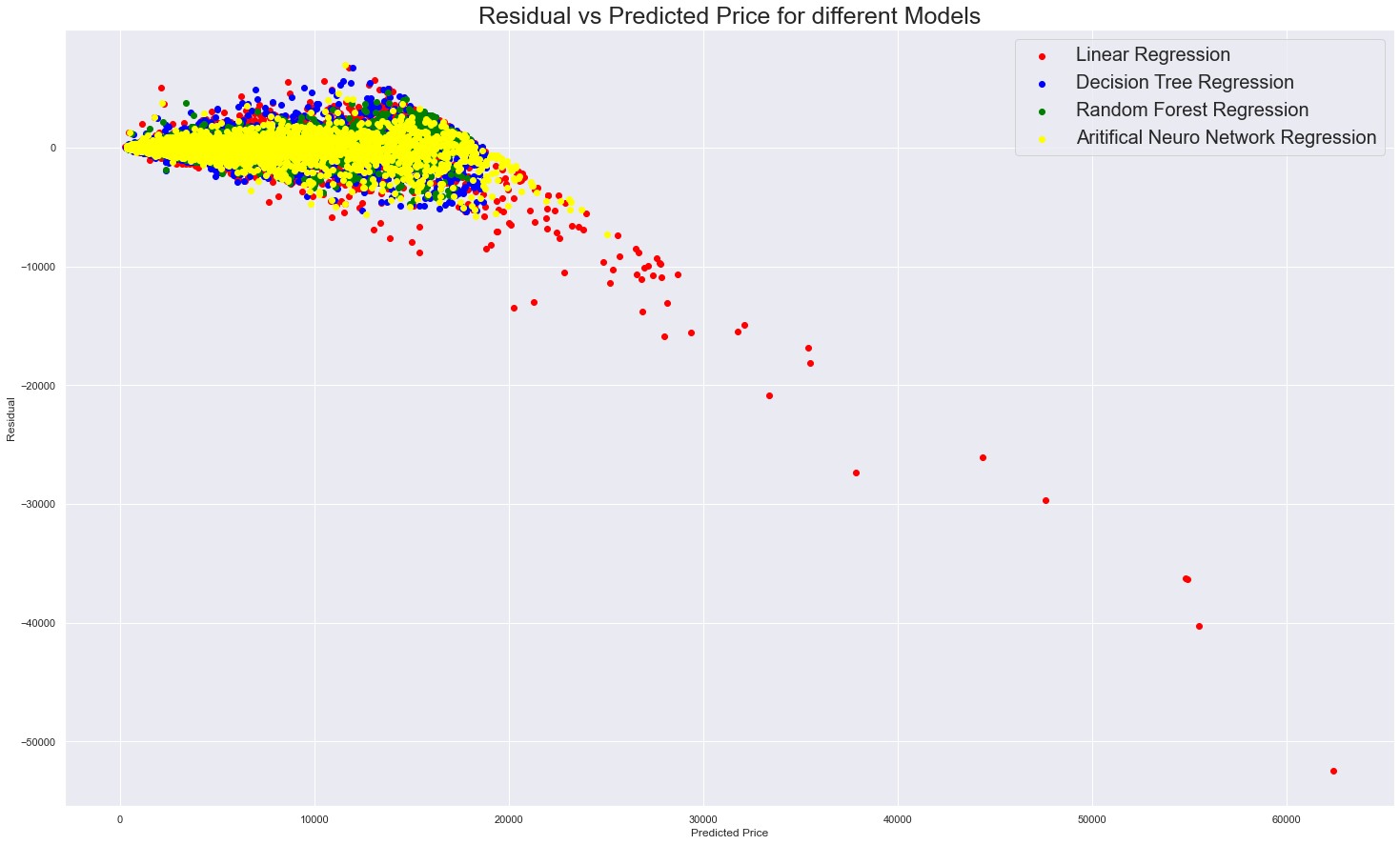
**On comparing all the models, it looks like that no model is over-fitting/under fitting.**

**All models test and train score are comparable and within 1-2% range. Hence no pruning is required here.**

**We can see that Decision Tree, Random Forest and Artificial Neural Network are giving good results and have r2 results close to 99%.**

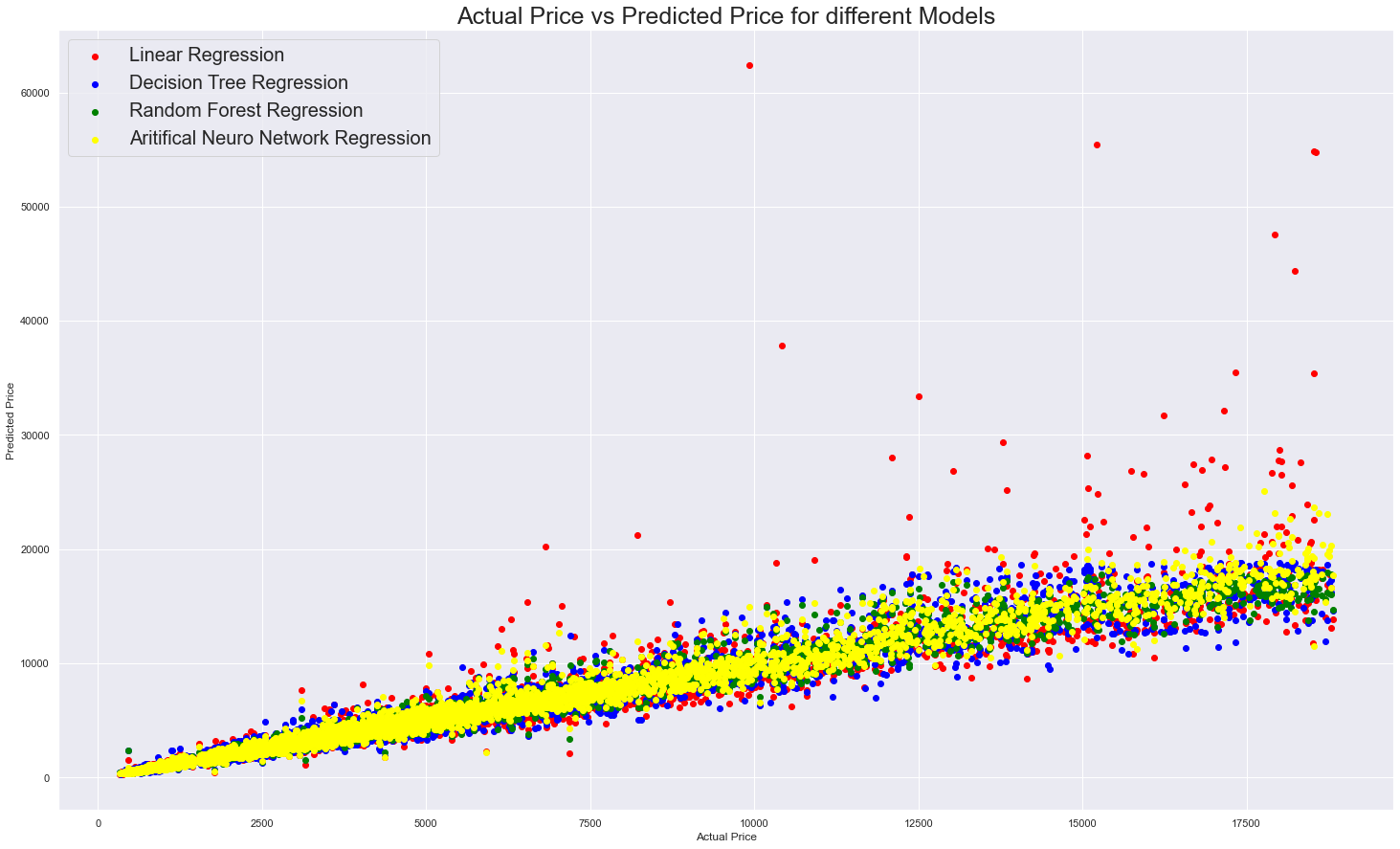
**Among all these models we will go for Random forest which has around 99% R2 score and Root mean square error is also these least (568) for test dataset**

#### Residual Vs Predicted Price plot for different Models



We can see that residual is quite high for Linear regression model as compared to other models. For Random Forest (green in colour) we can see that residue is near 0 for different predicted price values.

#### Actual price Vs predicted price comparison for all models

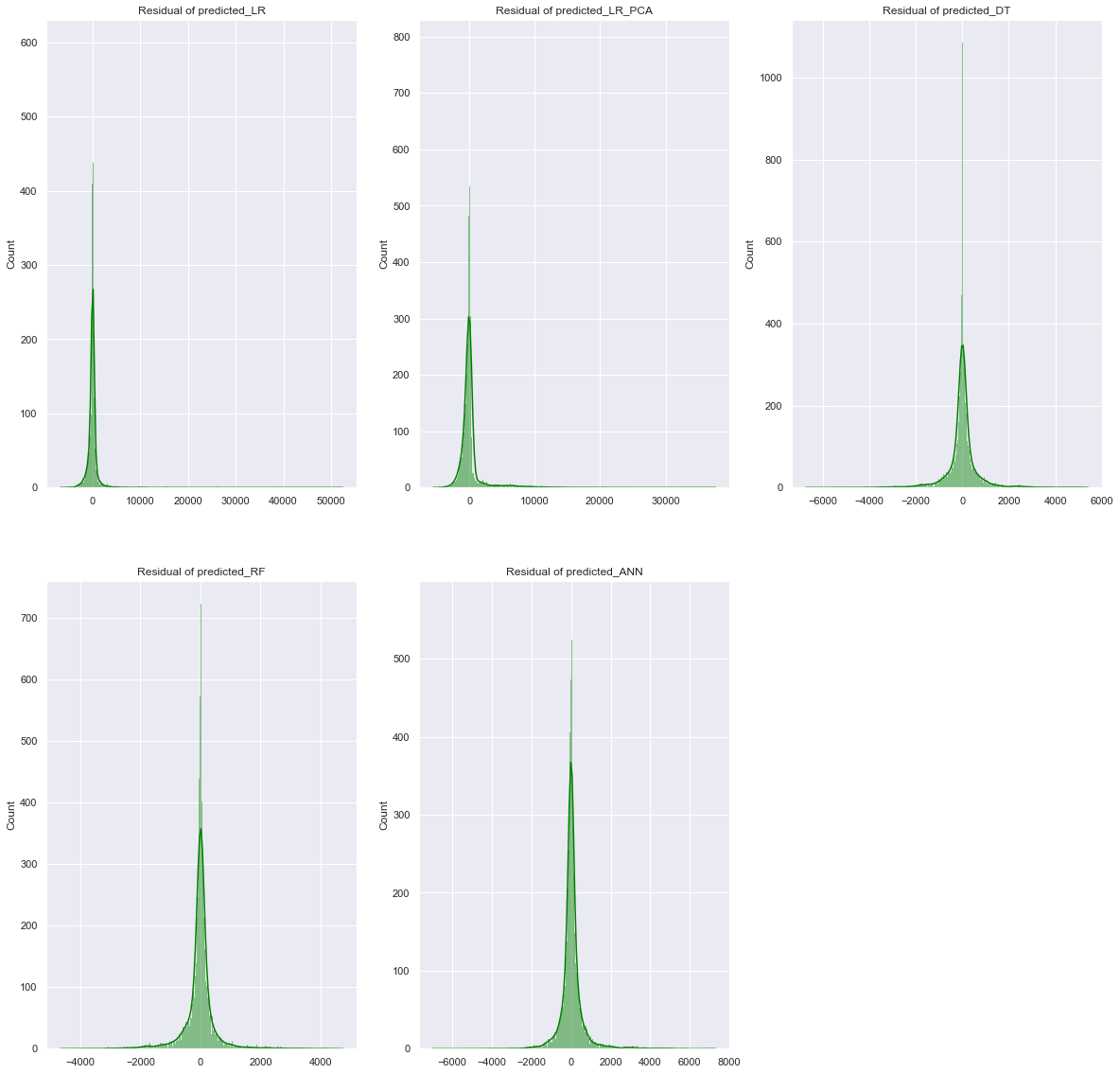


We can see that Liner Regression model has some of the predicted price very far from other models.

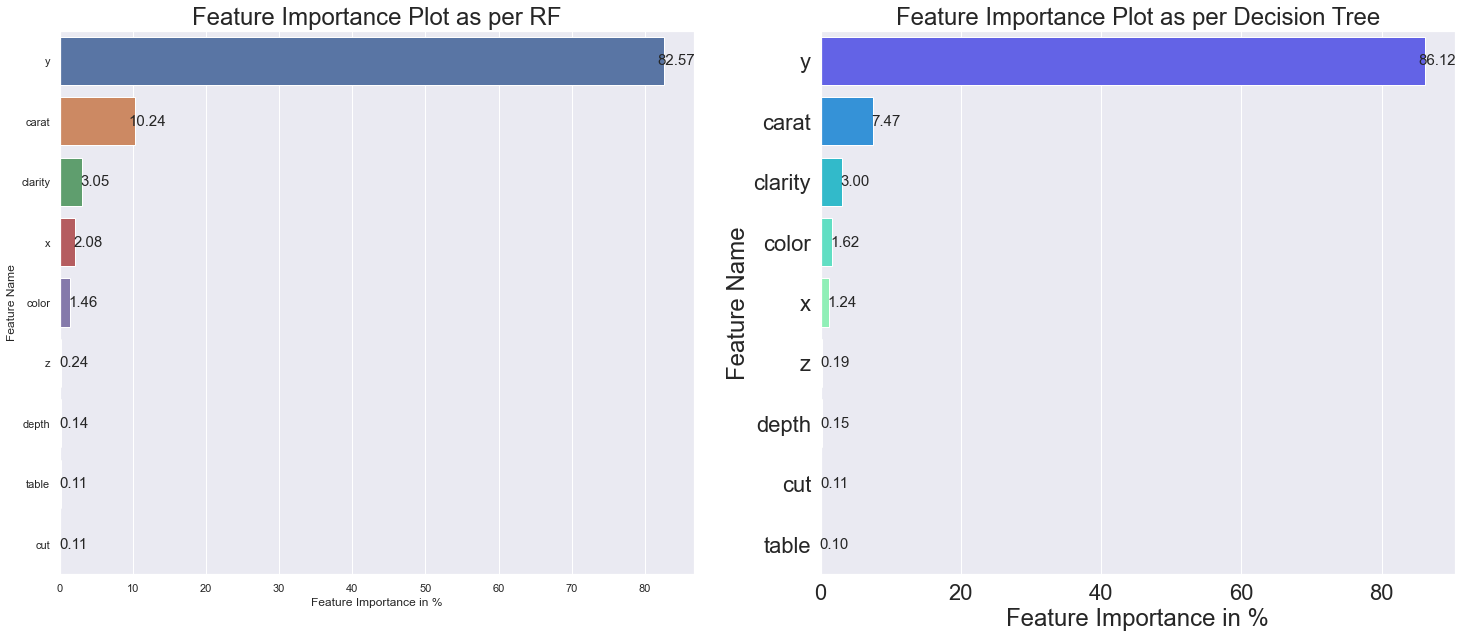
Random Forest and ANN model points are quite close to each other and error seems to be less in these models

#### Error (Residue) distribution plot for all the models

As seen in the histograms below, Linear regression model error distribution plot has some positive skewness. All other models seems to be centrally distributed.



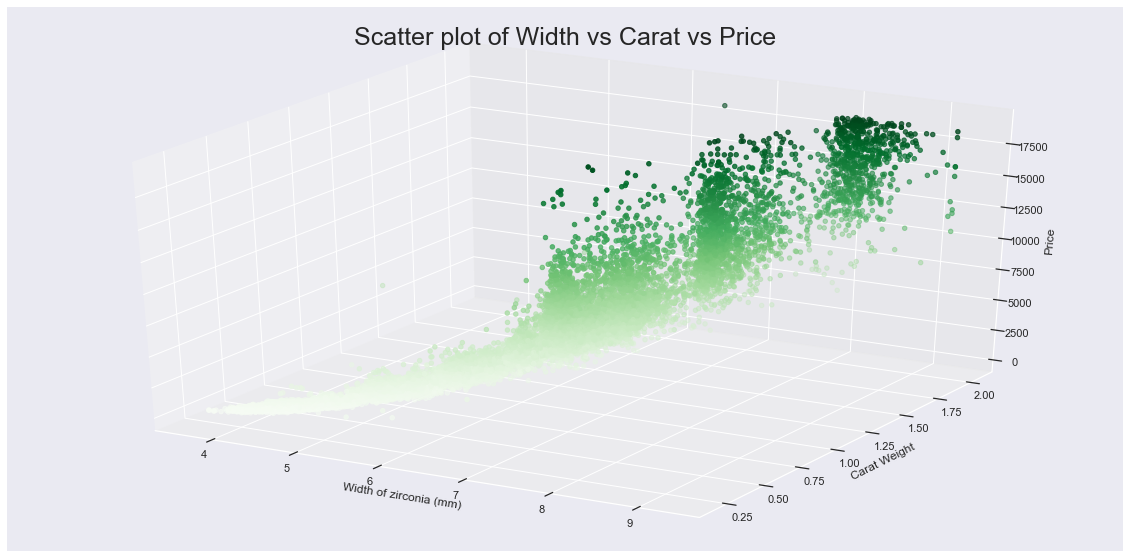
#### Variable Importance



We can see as per Random Forest which is giving best results, top feature which are helping in predicting the price are :-

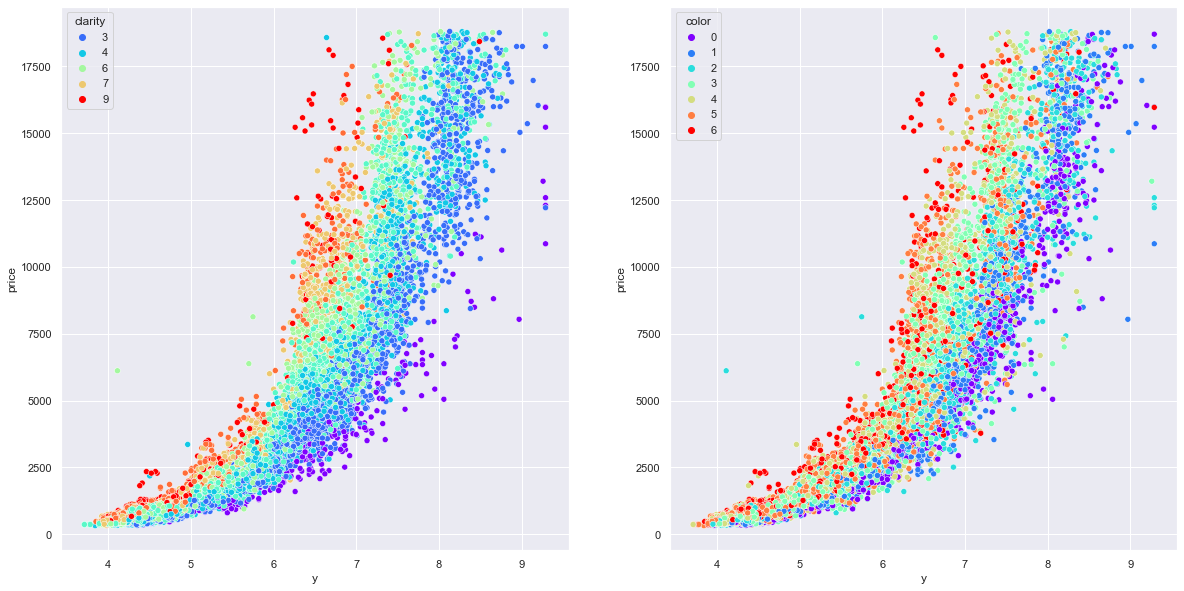
* y (width)
* carat (weight)
* clarity
* x (length)
* color

### Analysing top 5 features visually



From above graph , we can see that as width and carat weight increases, price increases.

Top price levels of cubic zirconia have high width and high carat weight

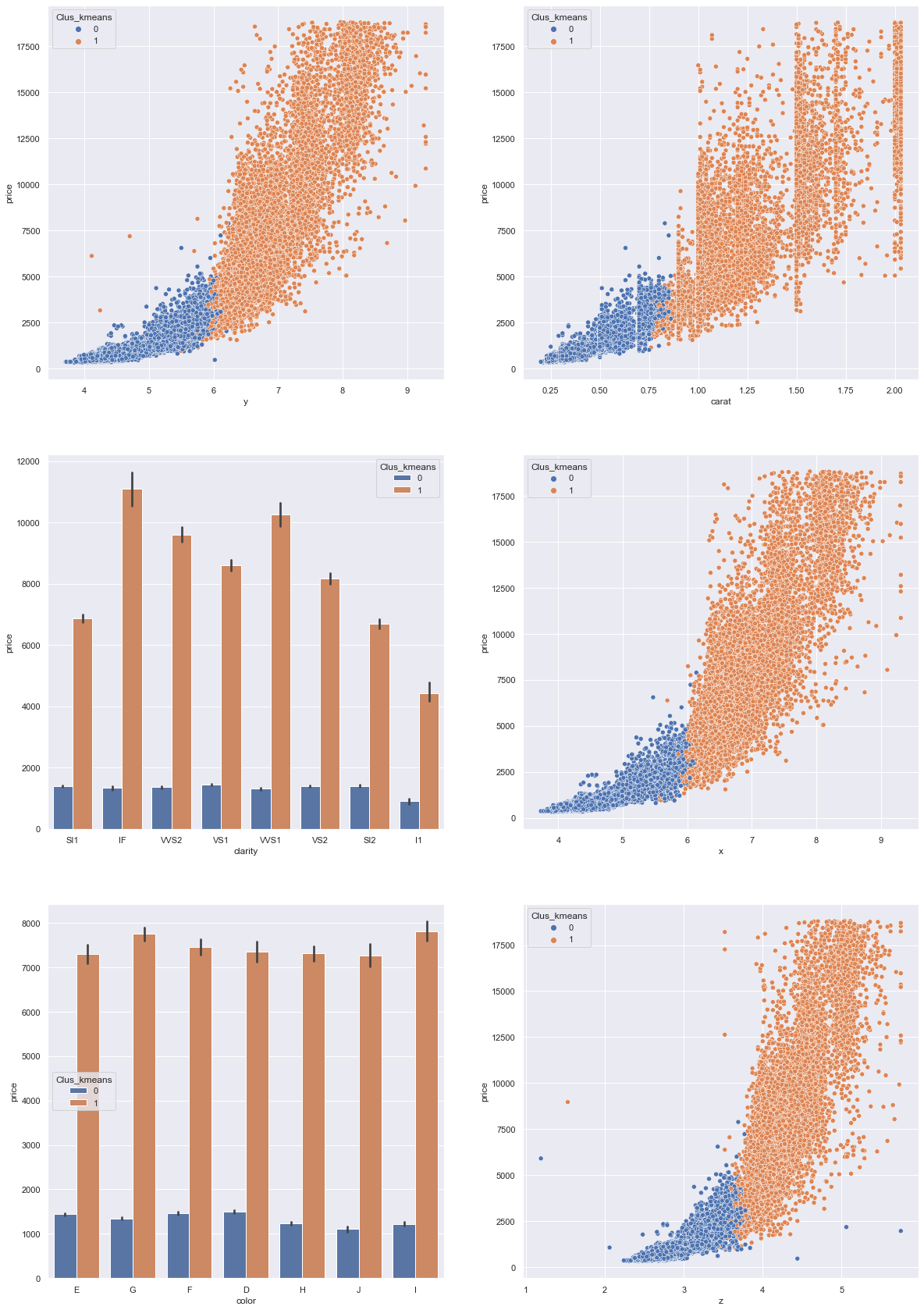


From above graphs, we can see that if we keep width as a constant, then price increases for better quality of clarity of cubic zirconia and better quality of color i.e. it means that individually these categorical variables have not much of importance but in combination with width, price increases as their quality range increases.

### K-Mean clustering for differentiating high and low price cubic zirconia

We have performed K-Mean clustering algorithm to clearly distinguish high price and low price cubic zirconia i.e. we have created 2 clusters.

Visually segmentation created is as shown below :-



From these graphs , we can conclude that price is in higher range for following cubic zirconia specifications :-

* Length > 6 mm
* Width > 6 mm
* Height > 4 mm
* Carat weight> 0.8
* Clarity better than l1

On similar grounds, cubic zirconia specification for lower price range is

* Length < 6 mm
* Width < 6 mm
* Height < 4 mm
* Carat weight < 0.8
* Clarity l1 or lower

### Conclusion

The final Linear Regression equation is

**log price = b0 + b1 \* carat + b2 \* cut + b3 \* color + b4 \* clarity + b5 \* depth + b6 \* table + b7 \* x + b8 \* y + b9 \* z**

**log price = (-3.99) + (-0.98) \* carat + (0.02) \* cut + (0.08) \* color + (0.12) \* clarity + (0.05) \* depth + (0.01) \* table + (0.75) \* x + (0.47) \* y + (0.19) \* z**

When carat increases by 1 unit, log\_price decreases by 0.98 units, keeping all other predictors constant. This is absurd as we have seen in the analysis that price increases with the carat.

Coefficient of x,y, and z is positive which means log price increases with the increase in x,y,z value. These are correlated with each other and carat. Due to high multicollinearity in the data set , as the value of carat will increase, value of x,y and z will increase. Overall effect results in increasing the price. Due to high multicollinearity these coefficients are not stable and there could be multiple coefffcient values which will give the same predicted value but the coefficients value would be different. Hence we should not rely on these coefficients.

After performing PCA to reduce multicollinearity in the dataset, new equation created is as shown below :-

**log\_price = (7.78) + (0.45) \* PC1 + (-0.12) \* PC2 + (-0.08) \* PC3 + (-0.06) \* PC4 + (0.36) \* PC5 + (-0.03) \* PC6**We can see that PC1 has the most important coefficient (0.45) for PC1, where PC equation is defined as PC1 : 0.43 \* carat + 0.43 *x + 0.43 \*y + 0.43*z - 0.09 \* cut - 0.14 \*color - 0.19 \*clarity + 0.01 \* depth + 0.11 \* table

We can see that PC1 is primarily the factor of carat , length, width and height of the zirocnia. So as these will increase, price of zirconia will significantly increase.

If we check the root mean square error of the predicted price, its coming as very high for linear regression model (1600) as compared to the other models. On comparing the results with other regressor models, we can see that Decision Tree, Random Forest and Aritificial Nueral Network are giving good results and have R2 results close to 99%.

Among all these models we should opt for Randorm forest regressor which has around 99% R2 score and Root mean square error is also these least (568) for test dataset. We have plotted various graphs showing residual error vs predicted, actual vs predicted, residual error distribution for all the models. Visually we can see that Random Forest regressor is giving the least error.

As per random forest, top feature which are helping in predicting the price are :-

* y (width)
* carat (weight)
* clarity
* x (length)
* color

As the width increases from 4mm to 9mm, price of cubic zirconia increases from 400 to 19000 approx. which is very steep rise. Thus it’s a very important factor

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

As per the analysis, our recommendation to business is :-

1) We have predicted the price of cubiz zirconia using multiple regressor models. After comparing several models, we have shortlisted that the Random Forest Regressor is giving the best results.

R2 score of the the model as 99% and Root Mean square error for the test data set is around 568.

2) Top 5 features which are contributing to the high price of cubic zirconia are :-

y (width)

carat (weight)

clarity

x (length)

color

3) Width, length, height and carat weight have a very strong correlation with each other. These are physical dimensions of the cubic zirconia, if we are increasing the one property, then the other has to be increased proportionately. Due to this relationship, it might be difficult to provide the linear equation with stable coefficients for the price prediction. Still we have obtained the linear equation as shown below :-

**log price = (-3.99) + (-0.98) \* carat + (0.02) \* cut + (0.08) \* color + (0.12) \* clarity + (0.05) \* depth + (0.01) \* table + (0.75) \* x + (0.47) \* y + (0.19) \* z**

4) We have found that Cut, Depth and Table attributes importance in predicting the price of cubic zirconia is the least as per the given data. So Business can check and analyse these features using some more dataset to understand the impact.

5) Cubic zirconia manufacturer should manufacture zirconia with the following specifications for higher price range and higher profits :-

Length > 6 mm

Width > 6 mm

Height > 4 mm

Carat weight> 0.8

Clarity better than l1

6) Cubic zirconia manufacturer should check the quantities sold for the following specifications :-

Length < 6 mm

Width < 6 mm

Height < 4 mm

Carat weight < 0.8

Clarity l1 or lower

For these specifications, price is low and if selling quantities are not high then these cubic zirconia manufacturing should be discontinued.