Siddhi Padekar

29.08.2021

GEM sTORES cO. lTD

Data Analysis

Table of Contents

|  |  |
| --- | --- |
| Summary | 2 |
| Introduction | 2 |
| Data Dictionary | 2 |
| Sample of the dataset | 3 |
| Exploratory Data Analysis | 3 |
| Univariate Analysis , Bivariate analysis, Multivariate Analysis | 7 |
| 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? | 12 |
| 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. | 13 |
| Inference: Basis on these predictions, what are the business insights and recommendations. | 16 |

Executive Summary

Gem Stones co ltd, which is a cubic zirconia manufacturer. Cubic zirconia is an inexpensive diamond alternative with many of the same qualities as a diamond. The dataset contains the prices and other attributes of almost 27,000 cubic zirconia. In this problem statement, we will study how to predict the price for the stone and how to distinguish between higher profitable stones and lower profitable stones to have better profit share

Introduction

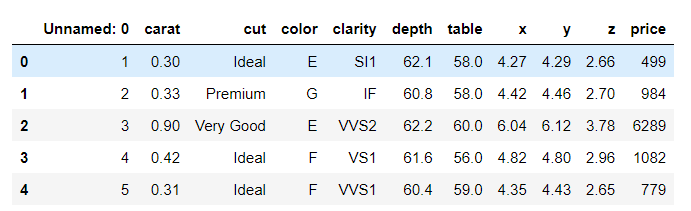
The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset, predict the price for the stone on the bases of the details given in the dataset using linear regression and Check the performance of Predictions on Train and Test sets using Rsquare, RMSE.

Data Dictionary:

Below is the brief description of data set

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Carat | Carat weight of the cubic zirconia. |
| Cut | Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal. |
| Color | Colour of the cubic zirconia.With D being the worst and J the best. |
| Clarity | cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst, IF = flawless, l1= level 1 inclusion) IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, l1 |
| Depth | The Height of cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter. |
| Table | The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter. |
| Price | the Price of the cubic zirconia. |
| X | Length of the cubic zirconia in mm. |
| Y | Width of the cubic zirconia in mm. |
| Z | Height of the cubic zirconia in mm. |

Sample of the dataset:



Above is sample details of data set showing first 5 rows.

Price is the target variable as we need to predict price of Cubic zirconia and all other are predictor variable

Exploratory Data Analysis

Data columns (total 11 columns):

Unnamed: 0 26967 non-null int64

Carat 26967 non-null float64

Cut 26967 non-null object

Color 26967 non-null object

Clarity 26967 non-null object

Depth 26270 non-null float64

Table 26967 non-null float64

X 26967 non-null float64

Y 26967 non-null float64

Z 26967 non-null float64

Price 26967 non-null int64

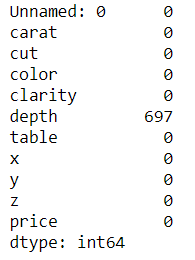
dtypes: float64(6), int64(2), object(3)

memory usage: 2.3+ MB

Shape of Data (26967, 11)

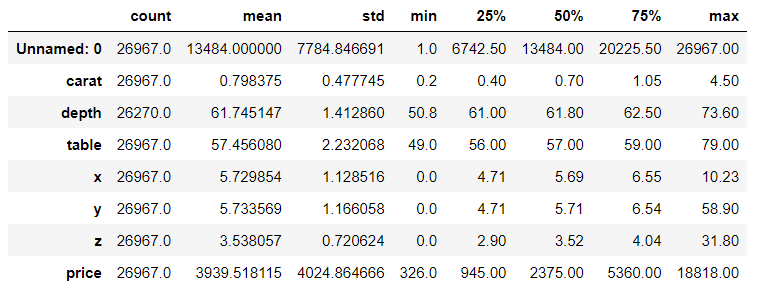
1. The data set contains 26967 row, 11 columns .
2. In the given data set there are 2 Integer type features, 6 Float type features. 3 Object type features.

Check for missing values in the dataset:

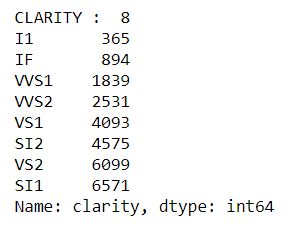
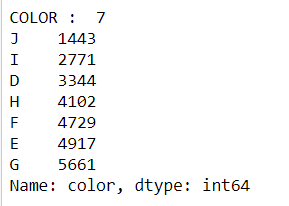
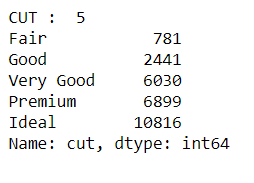


* There are Nul values present only in depth column of data set

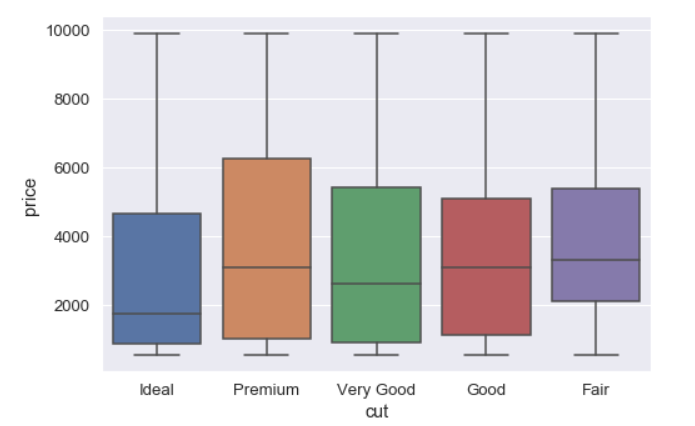
Mean, Standard deviation , Min, Max

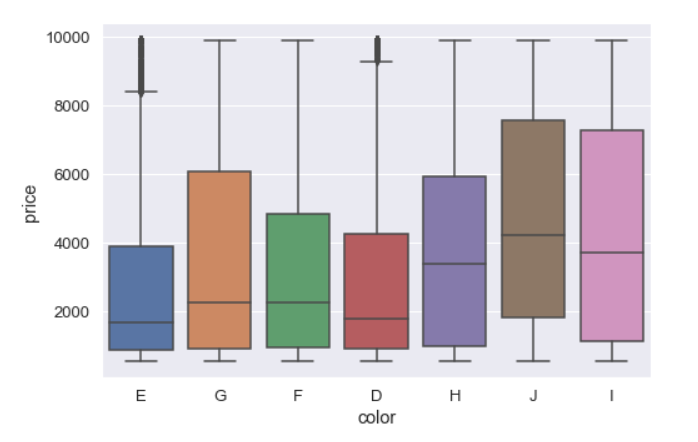


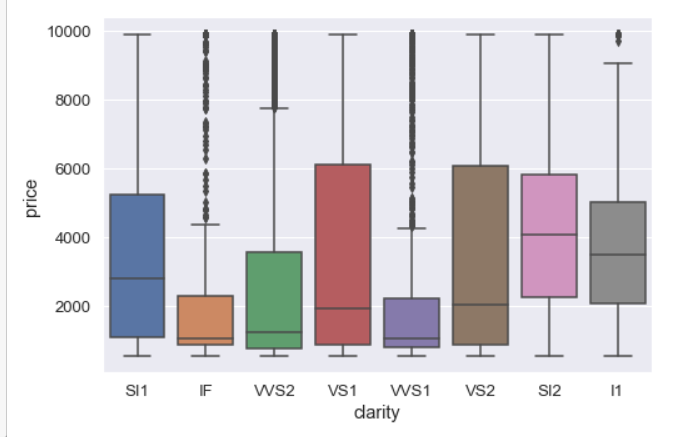
Categorising the object variable (Cut , Color, Clarity)



* Cut is categorised into 5 heads – Fair ,Good , Very Good, Premium , Ideal ( on the Top is Ideal )
* Color is categorised into 7 heads – J,I,D,H,F,E,G ( on the Top is G)
* Clarity is categorised into 8 heads – I1,IF,VVS1,VVS2,VS1, SI2,VS2,SI1 ( on the Top is SI1)







Check for Duplicate values and Removing Duplicate values in the dataset:

Number of duplicate rows = 33

Before removing duplicate values - (26958, 10) – 26958 Rows and 10 Columns

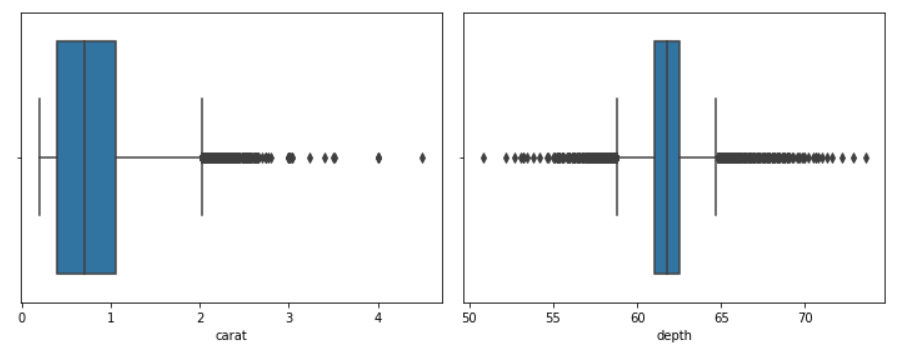
After removing duplicate values- (26925, 10) - 26925 Rows and 10 Columns

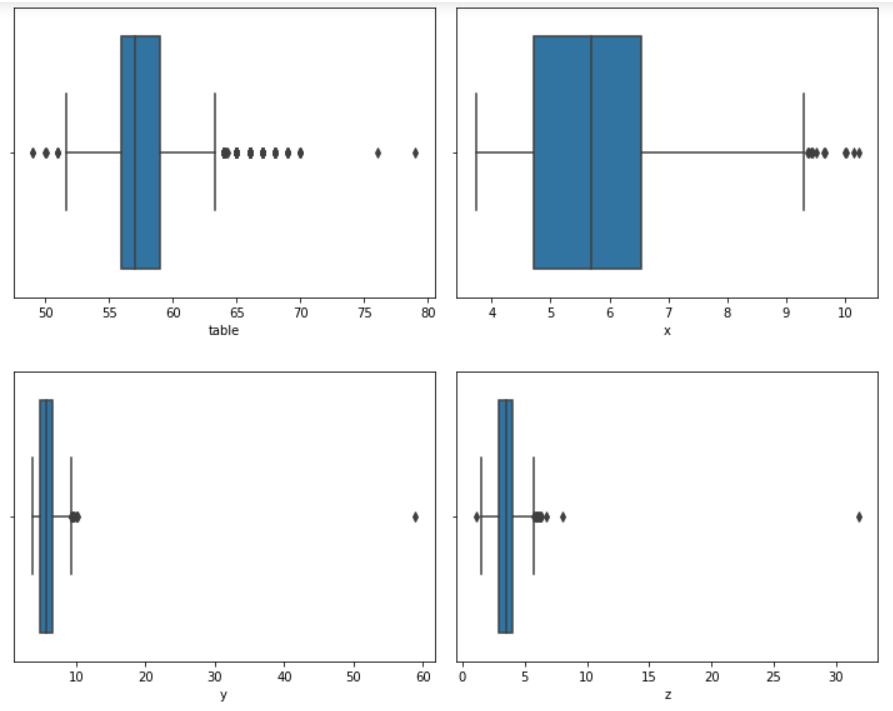
We have dropped duplicate values as they are very small in numbers.

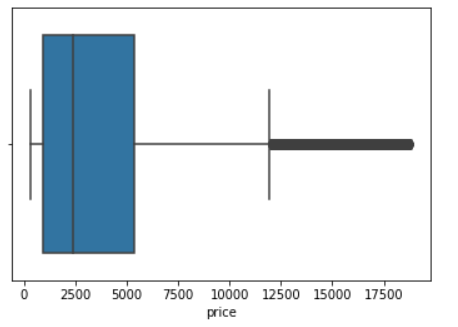
Note- we have dropped first column from the dataset ("Unnamed: 0") as this is only serial numbers

Outliers

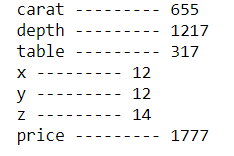
An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. A value that "lies outside" (is much smaller or larger than) most of the other values in a set of data.



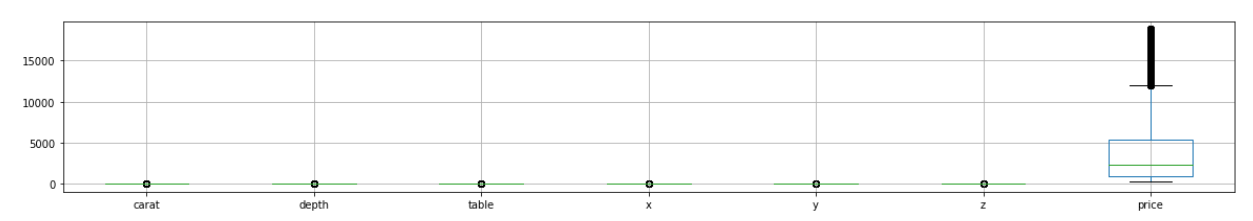




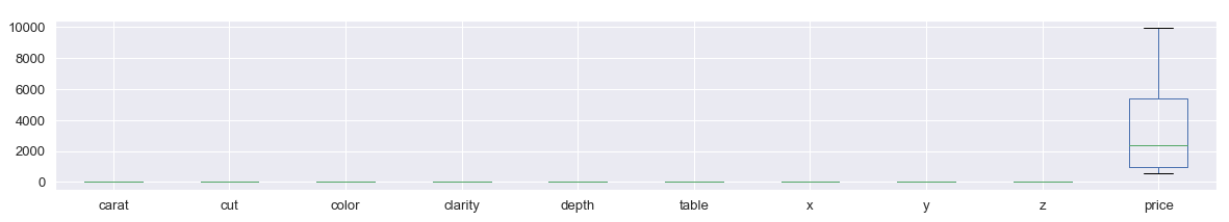
There are significant number of outliers in all the attributes



Before removing Outliers

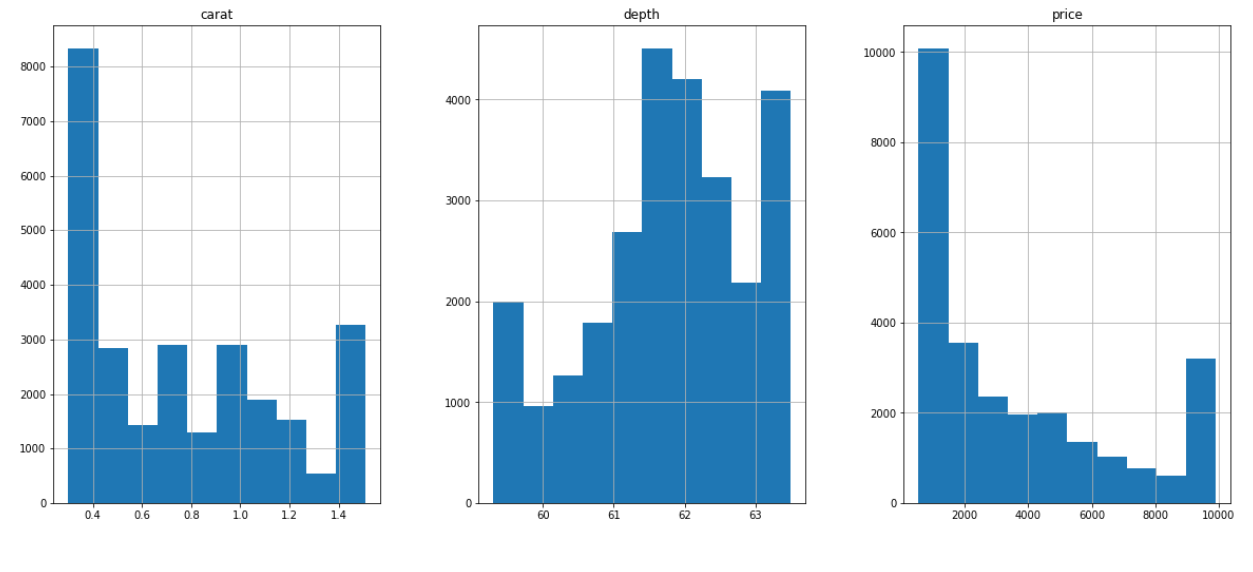


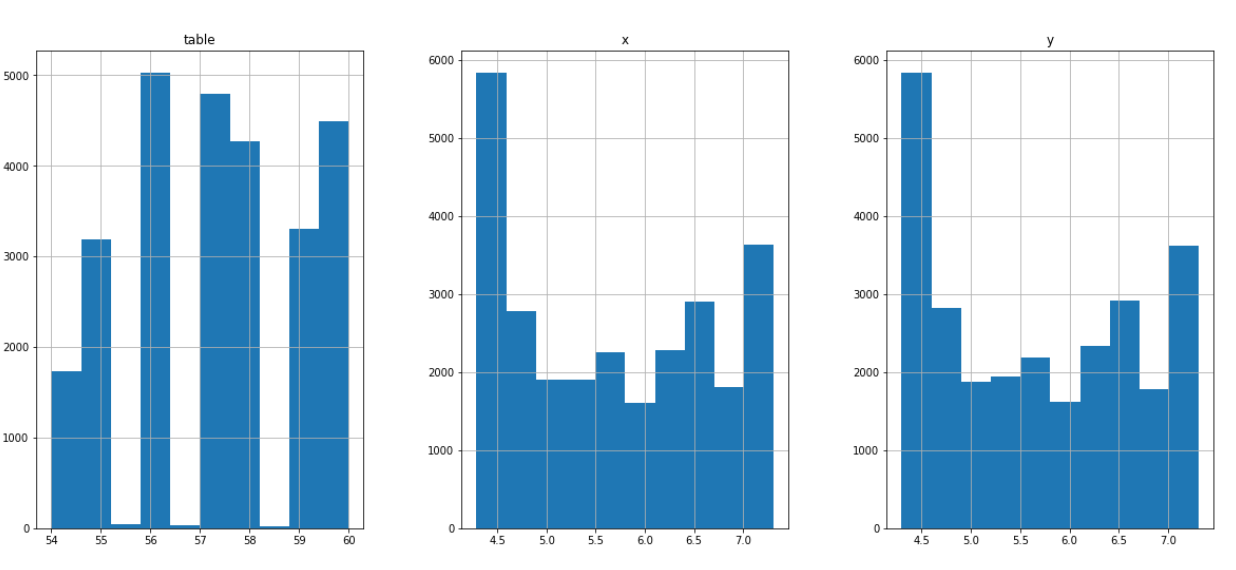
After removing outliers

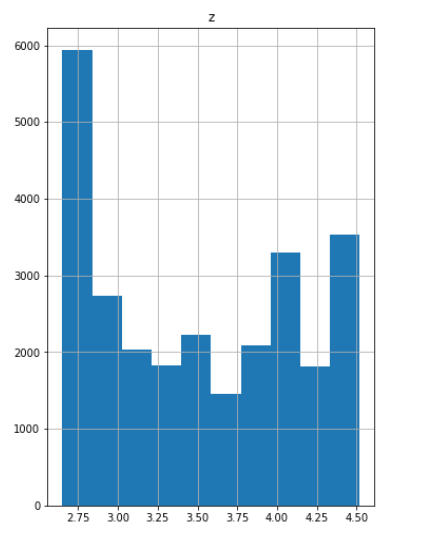


Univariate Analysis

Histograms



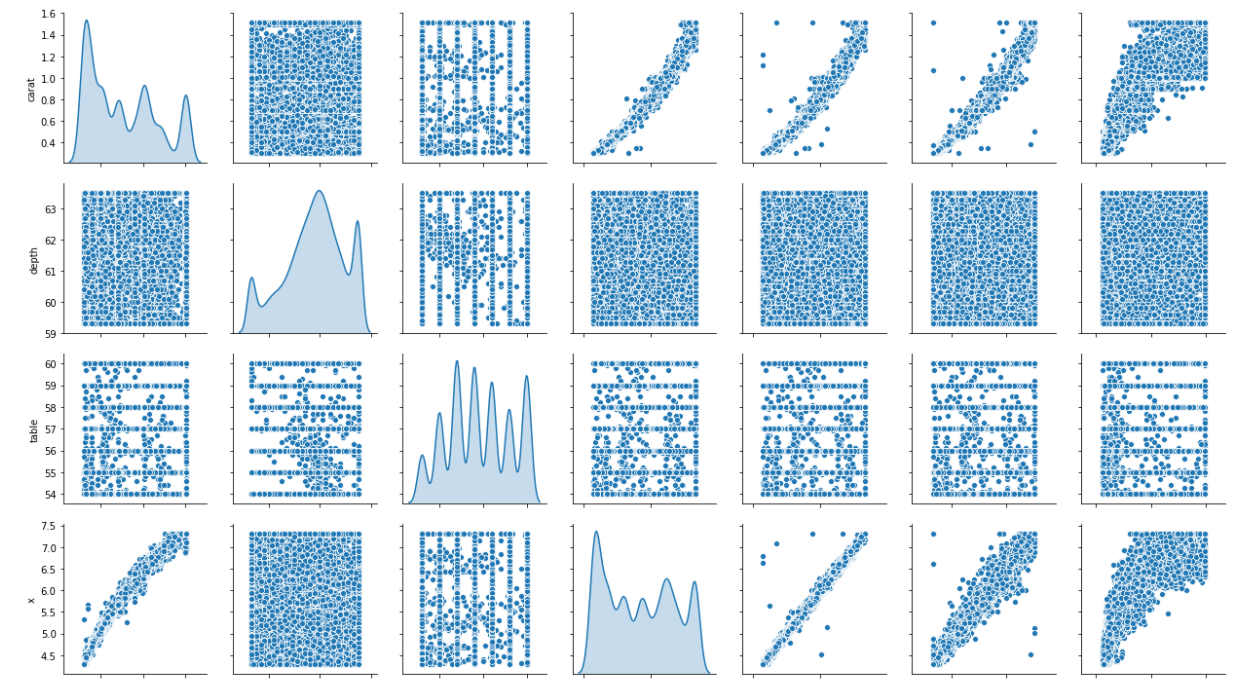


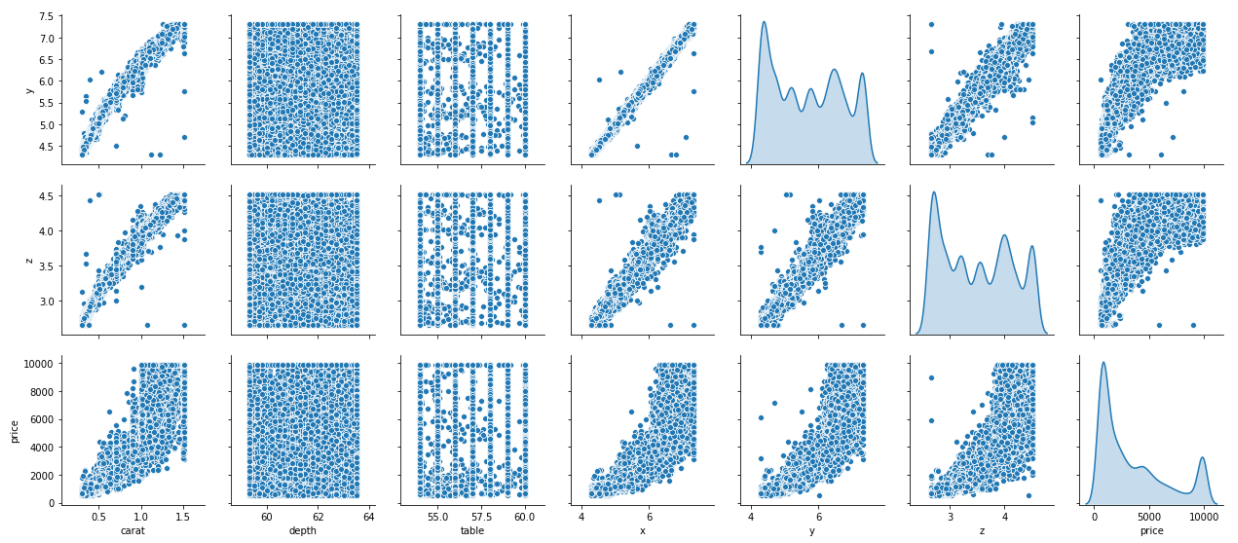


Bivariate Analysis

Pair Plot

Pair plot visualizes given data to find the relationship between them where the variables can be continuous or categorical.





Multivariate analysis

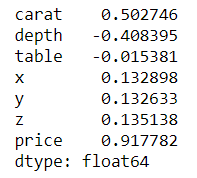
Correlation Plot



### The variable 'carat', 'x','y', 'z' correlates with target variable 'price' .

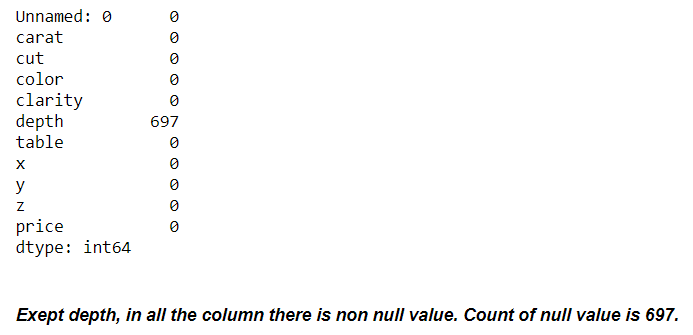
From the correlation plot, we can see that various variable 'carat', 'x','y', 'z' correlates with target variable 'price' . Correlation values near to 1 are highly positively correlated. Correlation values near to 0 are not correlated to each other.

Skewness of every attribute



### We can see that the target feature "price" are heavily "right-skewed".

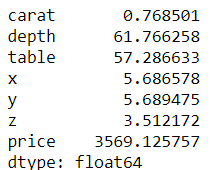
Imputing Null values



In this case, the mean or median values can be used for imputing the missing or null values of the continuous numerical variables as the mean and median values does not have much difference.

The mean (average) of a data set is found by adding all numbers in the data set and then dividing by the number of values in the set. The median is the middle value when a data set is ordered from least to greatest.

Hence, we have imputed null values with **mean** in the data set.



Are there any values Equal to 0?

Yes there are values equal to 0, however they are faulty data as length , width ,height of any Cubic zirconia gem can not be 0 . Hence we have removed them from the data.

Number of rows with x == 0: 3

Number of rows with y == 0: 3

Number of rows with z == 0: 9

Number of rows with depth == 0: 0

Do you think scaling is necessary in this case?

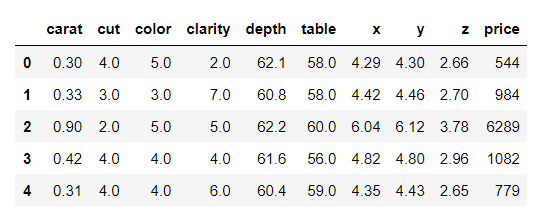
Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. Standardization is another scaling technique where the values are centred on the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation. It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.

In the given data set data set, we can see the all the variable are scaled on different parameters. Like price varies in 1000s unit (Rupees) and depth varies in 100s unit (cms) and table are in 100s unit (percentages) , and carat is in 10s (weights) . Therefore, it is necessary to scale or standardise the data to allow each variable to be compared on a common scale.

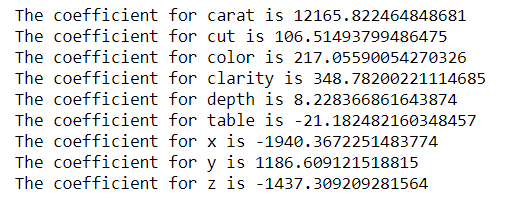
However, in this case it is not necessary to scale the data. We will get a solution whether we apply some kind of linear scaling or not. When number of features becomes large, it helps is running model quickly else the starting point would be very far from minima, if the scaling is not done in pre-processing.

**Hence, we have now processed the model without scaling also, we will check the output with scaled data of regression model output.**

Sample data set after converting object variable into float.



**Linear regression** after splitting data into train and test



We can obverse that there is a direct relation between following independent variables and price.

A unit increases in carat the price also increases with 12165.82 .

A unit increase in cut the price increases by 106.51

A unit increase in color the price increases by 217.05

A unit increase in clarity the price increases by 348.78

A unit increase in depth the price increases by 8.22

A unit increase in y the price increases by 1186.609

However,

A unit decrease in table the price also decrease with -21.18.

A unit decrease in x the price also decrease with -1940.36

A unit decrease in z the price also decrease with –1437.30

The intercept (often labeled as constant) is the point where the function crosses the y-axis. In some analysis, the regression model only becomes significant when we remove the intercept, and the regression line reduces to Y = bX + error. If X equals 0, the intercept is simply the expected mean value of Y at that value. If X never equals 0, then the intercept has no intrinsic meaning. In scientific research, the purpose of a regression model is to understand the relationship between predictors and the response.  If so, and if X never = 0, there is no interest in the intercept. It doesn’t tell you anything about the relationship between X and Y.

The intercept for our model is 2142.1746382196725 . This concludes that our other variables are 0 our price is 2142.17 no interest in the intercept. It doesn’t tell you anything about the relationship between X and Y. hence we need to make it 0 by doing zscore technique.

**Performance Metrics:**

R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.The R-squared value R2 is always between 0 and 1 inclusive.

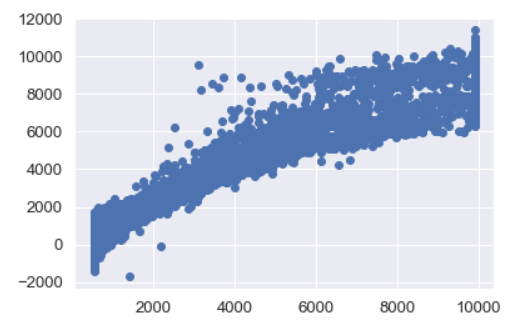
R square on training data is 0.9381845880910002

R square on test data is 0.9390427010465159

0% indicates that the model explains none of the variability of the response data around its mean.100% indicates that the model explains all the variability of the response data around its mean. In this regression model we can see the R-square value on Training and Test data respectively 0.938184588091000 and 0.9390427010465159.

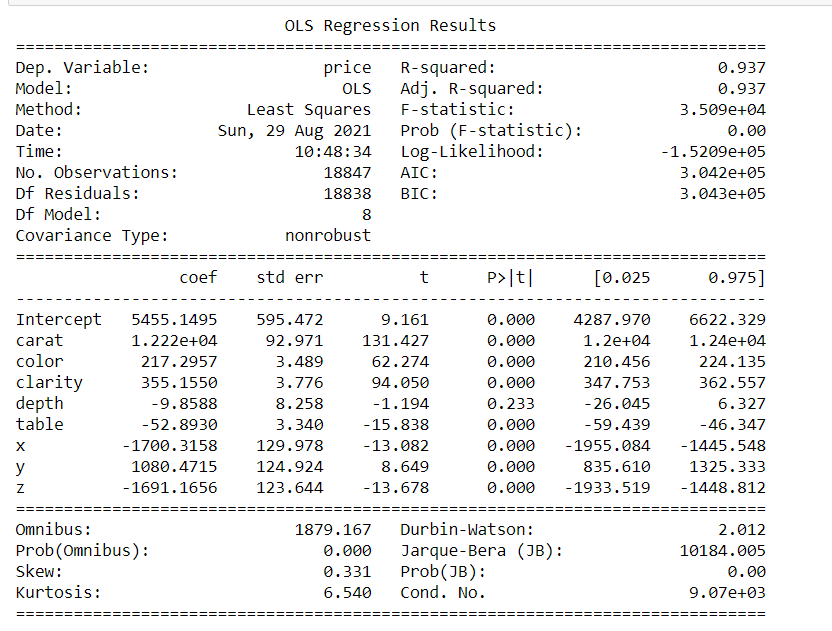
RMSE on Training data is 766.7476323246534

RMSE on Testing data is 767.3860552453051



As the training data & testing data score are almost inline, we can conclude this model is a Right-Fit Model.

Regression model summary -



If we assume null hypothesis is true, i.e there is no relationship between this variable with price. Now we can ask what is the probability of finding this co-efficient in this drawn sample if in the real world the co-efficient is zero. As we see here 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'.

For an example: we can see the p value is showing 0.233 for 'depth' variable, which is much higher than 0.05. That means this dimension is useless. So we can say that the attribute which are having p value greater than 0.05 are poor predictor for price.

Final Regression equation

(5455.15) \* Intercept + (12218.91) \* carat + (217.3) \* color + (355.16) \* clarity + (-9.86) \* depth + (-52.89) \* table + (-1700.32) \* x + (1080.47) \* y + (-1691.17) \* z +

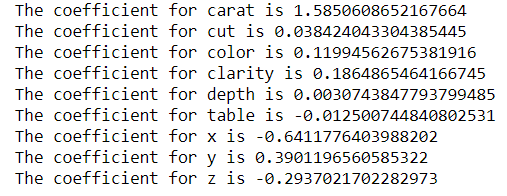
1. When carat increases by 1 unit, diamond price increases by 12218.91 units, keeping all other predictors constant.
2. When color increases by 1 unit, diamond price increases by 217.3 units, keeping all other predictors constant.
3. When clarity increases by 1 unit, diamond price increases by 355.16 units, keeping all other predictors constant.
4. When y increases by 1 unit, diamond price increases by 1080.47 units, keeping all other predictors constant.

As per model these five attributes that are most important attributes 'Carat', 'Cut', 'color','clarity' and width i.e 'y' for predicting the price.

There are also some negative co-efficient values, for instance, corresponding co-efficient (-1700.32) for 'x', (-9.86) for depth , (- -1691.17) for z and (-52.89) for table This implies, these are inversely proportional with diamond price.

After applying Zscore

Linear Regression



The intercept for our model is -2.2577061658186334e-16

### Inference:

we can see that the from the linear plot, very strong corelation between the predicted y and actual y. But there are lots of spread. That indicates some kind noise present on the data set i.e Unexplained variances on the output.

### Impact of scaling:

Now we can observe by applying z score the intercept became -2.2577061658186334e-16. Earlier it was -2142.1746382196725 . the co-efficient has changed and became nearly zero however the overall accuracy still same.

### Multi collinearity:

We can observe there are very strong multi collinearity present in the data set.

### Recommendations:

#### The Gem Stones Company should consider the features 'Carat', 'Cut', 'color', 'clarity' and width i.e 'y' as most important for predicting the price.

To distinguish between higher profitable stones and lower profitable stones so as to have better profit share.

As we can see from the model Higher the width('y') of the stone is higher the price.

#### So the stones having higher width('y') should consider in higher profitable stones.

The 'Premium Cut' on Diamonds are the most Expensive, followed by 'Very Good' Cut, these should consider in higher profitable stones.

The Diamonds clarity with 'VS1' &'VS2' are the most Expensive. So these two category also consider in higher profitable stones.

As we see for 'X' i.e Length of the stone, higher the length of the stone is lower the price.

#### So higher the Length('x') of the stone are lower is the profitability. higher the 'z' i.e Height of the stone is, lower the price.

#### This is because if a Diamond's Height is too large Diamond will become 'Dark' in appearance because it will no longer return an Attractive amount of light. That is why

#### Stones with higher 'z' is also are lower in profitability.