Problem 1 – Linear Regression

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 Dictionary for Market Segmentation:

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

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.

Reading the data

The data is read, and the **top five records** of the dataset are observed to get an overview.

	carat	cut	color	clarity	depth	table	x	y	z	price
1	0.30	Ideal	E	SI1	62.1	58.0	4.27	4.29	2.66	499
2	0.33	Premium	G	IF	60.8	58.0	4.42	4.46	2.70	984
3	0.90	Very Good	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289
4	0.42	Ideal	F	VS1	61.6	56.0	4.82	4.80	2.96	1082
5	0.31	Ideal	F	VVS1	60.4	59.0	4.35	4.43	2.65	779

Exploratory Data Analysis

The column properties are observed

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 26967 entries, 1 to 26967
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: 2.3+ MB
```

The descriptive statistics is also taken

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
carat	26967	NaN	NaN	NaN	0.798375	0.477745	0.2	0.4	0.7	1.05	4.5
cut	26967	5	Ideal	10816	NaN	NaN	NaN	NaN	NaN	NaN	NaN
color	26967	7	G	5661	NaN	NaN	NaN	NaN	NaN	NaN	NaN
clarity	26967	8	SI1	6571	NaN	NaN	NaN	NaN	NaN	NaN	NaN
depth	26270	NaN	NaN	NaN	61.7451	1.41286	50.8	61	61.8	62.5	73.6
table	26967	NaN	NaN	NaN	57.4561	2.23207	49	56	57	59	79
x	26967	NaN	NaN	NaN	5.72985	1.12852	0	4.71	5.69	6.55	10.23
y	26967	NaN	NaN	NaN	5.73357	1.16606	0	4.71	5.71	6.54	58.9
z	26967	NaN	NaN	NaN	3.53806	0.720624	0	2.9	3.52	4.04	31.8
price	26967	NaN	NaN	NaN	3939.52	4024.86	326	945	2375	5360	18818

Below are the inferences from the descriptive statistics:

- There are totally 26967 observations and 10 features
- There are 3 object type features – cut, color and clarity and the remaining are numeric - carat, depth, table, x, y , z,price.
- The variable cut has 5 unique values , color has 7 and clarity has 8.
- The frequency of the values for the object data types are as follows:
  - The most frequent value for cut is 'Ideal' occuring 10816 times.
  - The most frequent value for color is 'G' occuring 5661 times.
  - The most frequent value for clarity is 'SI1' occuring 6571 times.
- The value ranges for the independent variables are all different since some represent weight and others, the meaurements of the diamond.
- About 75% of the observations have around 1.05 carats (weight) and 4.5 carats is the maximum weight.

- There are 697 null values in the dataset for depth variable.
- There are 34 duplicate observations.

Actions done after basic EDA

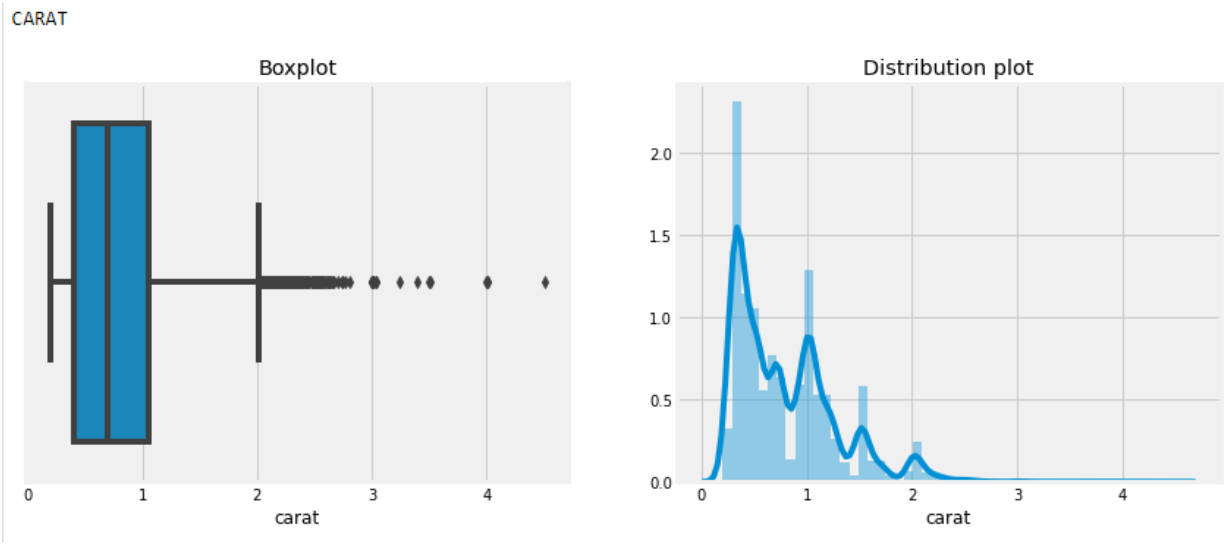
The duplicate records are removed, and now, there are 26933 unique records.

Univariate Analysis for Numeric variables:

The univariate analysis (using boxplot and distribution plot to check skewness and outliers) is done each variable.

CARAT:

- The boxplot shows that there are outliers.
- The distribution plot shows that the data distribution is right-skewed.



DEPTH:

- The boxplot shows that there are outliers.
- The distribution plot shows that the data distribution is a bit left-skewed.

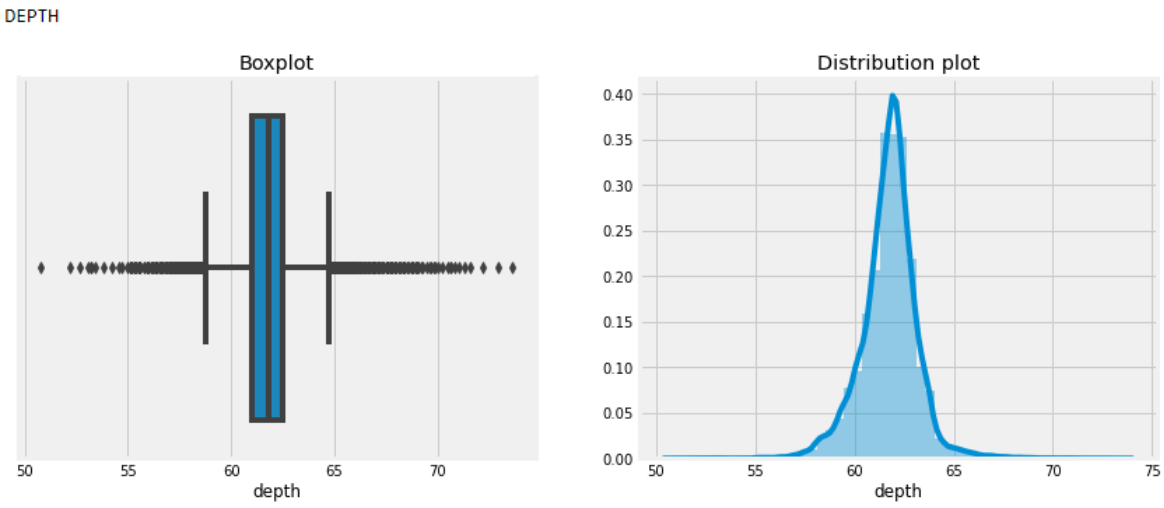
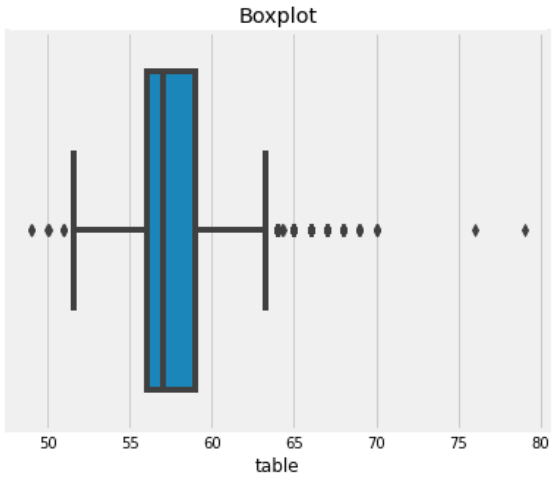
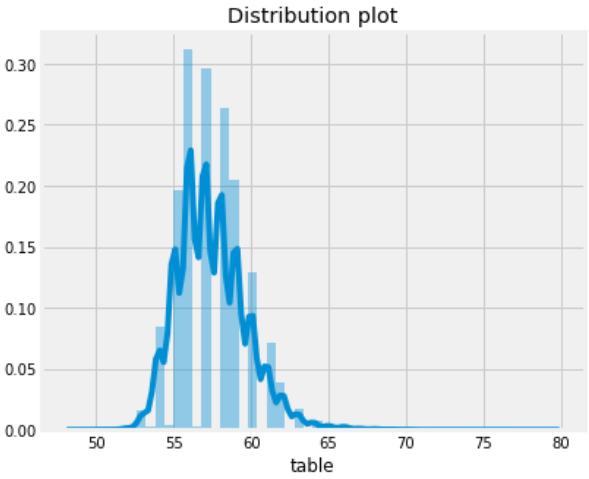


TABLE:

- The boxplot shows that there are outliers.
- The distribution plot shows that the data distribution is a bit right skewed..

TABLE

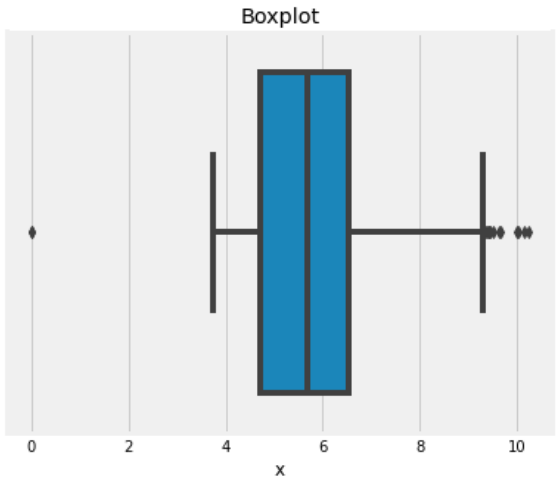


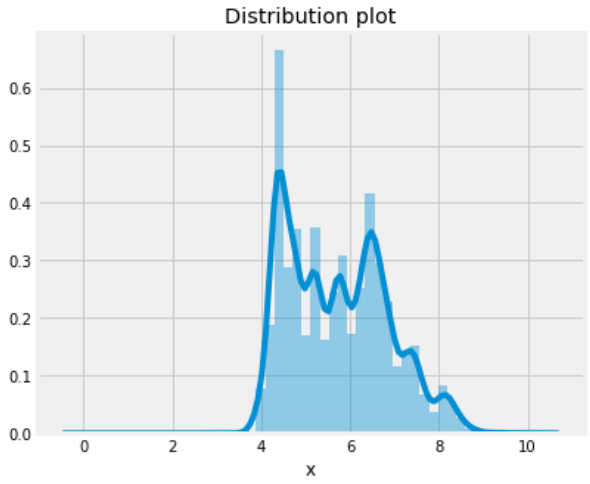


X:

- The boxplot shows that there are outliers.
- The distribution plot shows that the data distribution is not exactly normal.

X

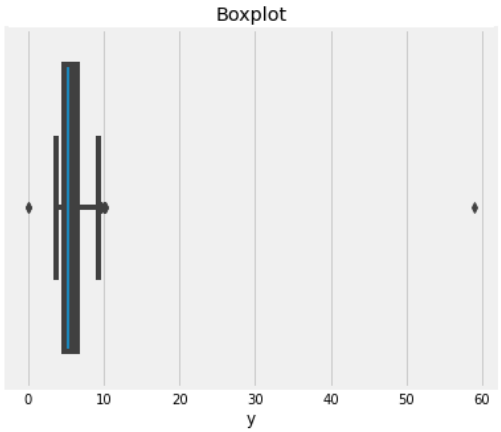


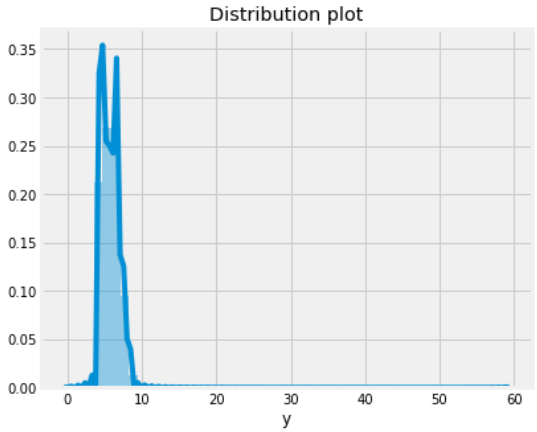


Y:

- The boxplot shows that there are a few outliers.
- The distribution plot shows that the data distribution is highly right-skewed.

Y

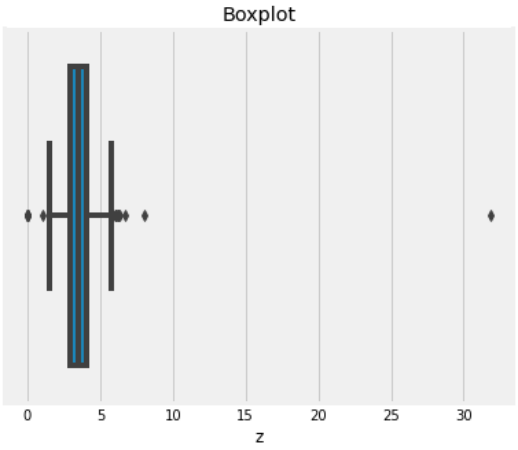


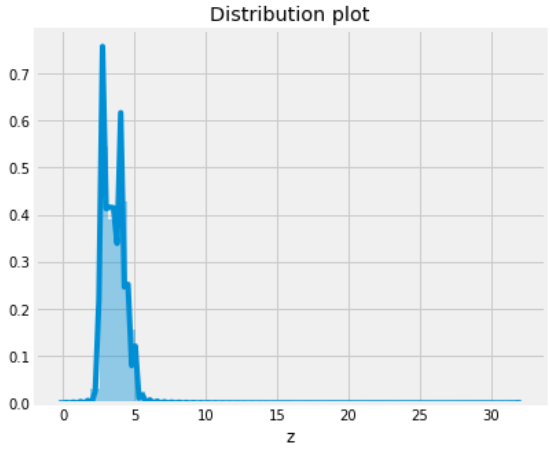


Z:

- The boxplot shows that there are a few outliers.
- The distribution plot shows that the data distribution is highly right-skewed.

Z

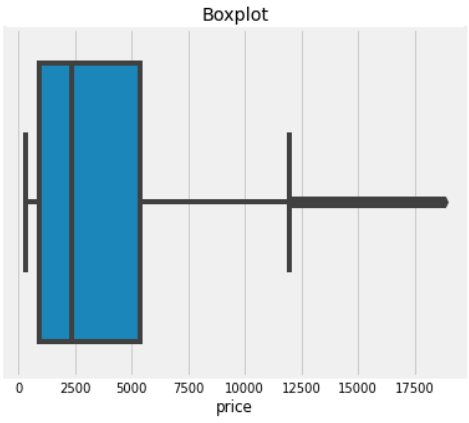


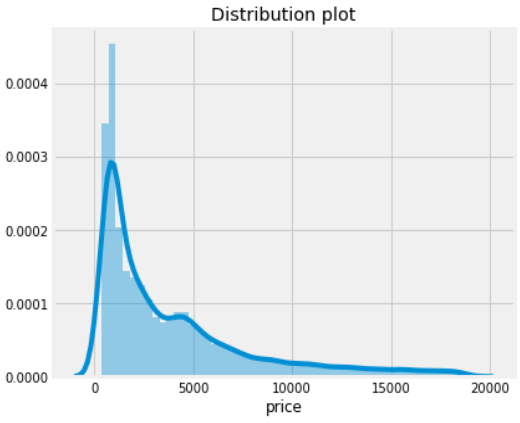


PRICE:

- The boxplot shows that there are a few outliers.
- The distribution plot shows that the data distribution is highly right-skewed.

PRICE





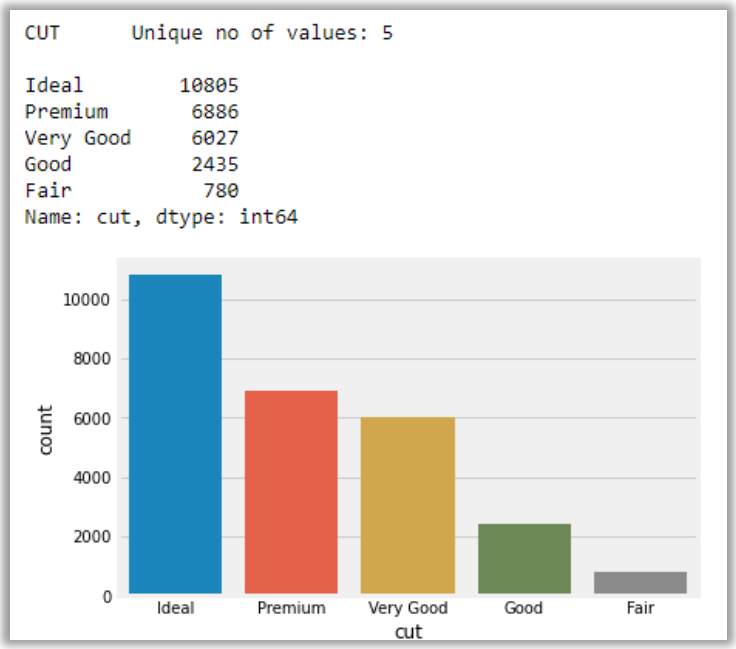
Univariate Analysis for Categorical variables:

CUT:

- There are totally 5 unique values – Ideal, Premium, Very Good, Good and Fair.
- Most of the observations(10805) in the dataset have cut type as Ideal.

```
CUT      Unique no of values: 5

Ideal        10805
Premium       6886
Very Good     6027
Good          2435
Fair           780
Name: cut, dtype: int64
```

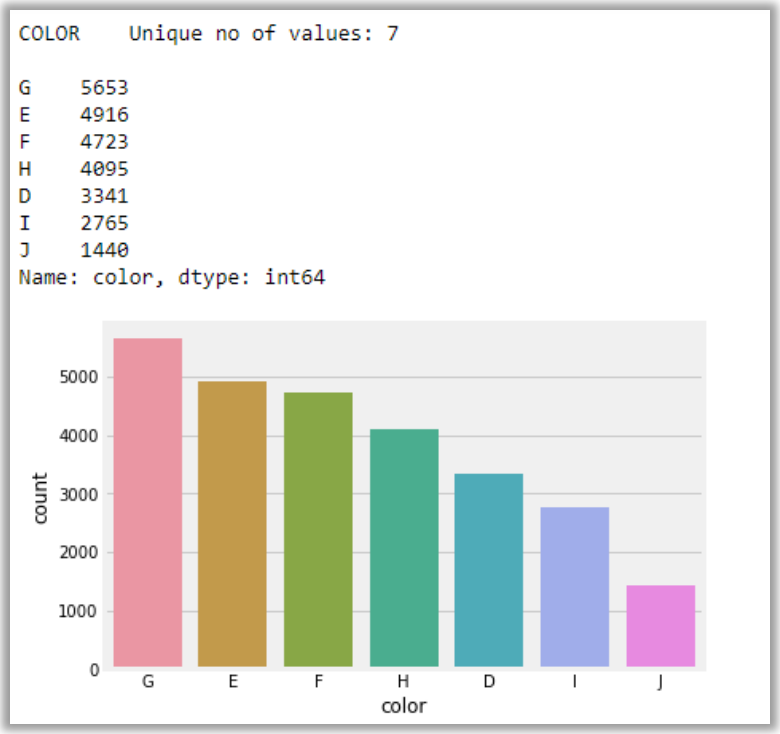


CUT:

- There are totally 7 unique values – D,E,F,G,H,I and J.
- Most of the observations(5653) in the dataset have color 'G'.

```
COLOR    Unique no of values: 7

G    5653
E    4916
F    4723
H    4095
D    3341
I    2765
J    1440
Name: color, dtype: int64
```

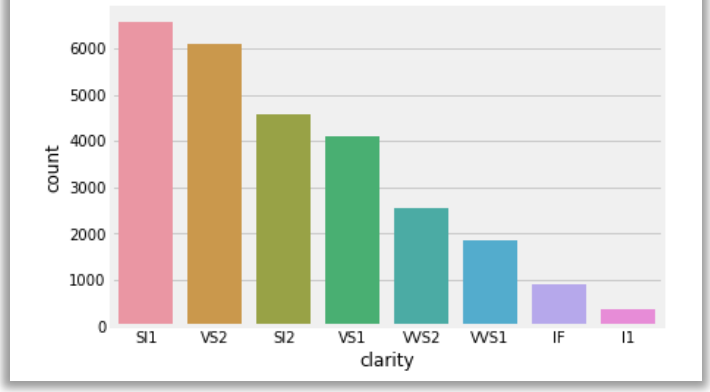


CUT:

- There are totally 8 unique values – 'SI1', 'IF', 'VVS2', 'VS1', 'VVS1', 'VS2', 'SI2' and 'I1'.
- Most of the observations(6565) in the dataset have clarity 'SI1'.

```
CLARITY          Unique no of values: 8

SI1      6565
VS2      6093
SI2      4564
VS1      4087
VVS2     2530
VVS1     1839
IF        891
I1        364
Name: clarity, dtype: int64
```

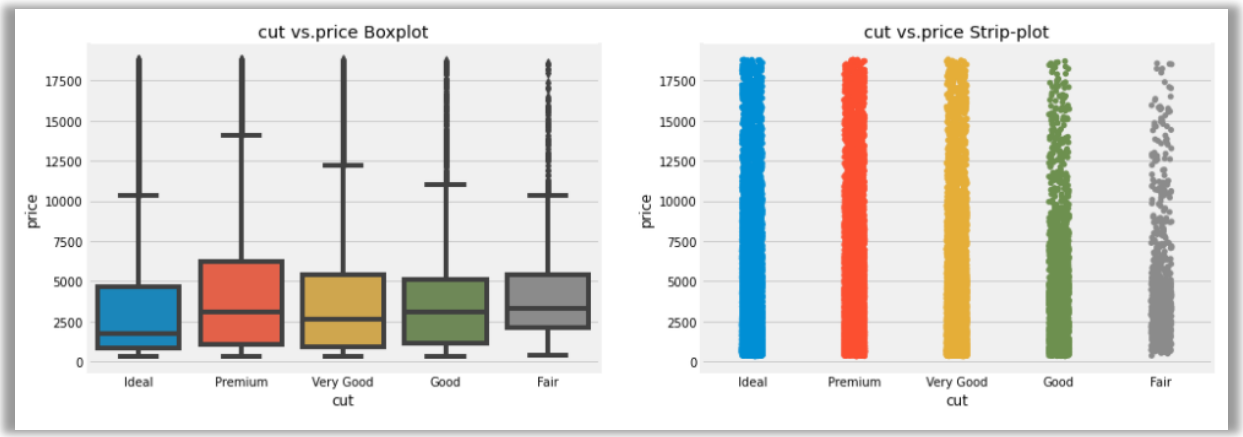


Bivariate Analysis:

A bivariate analysis is done between the price and the different categorical variables.

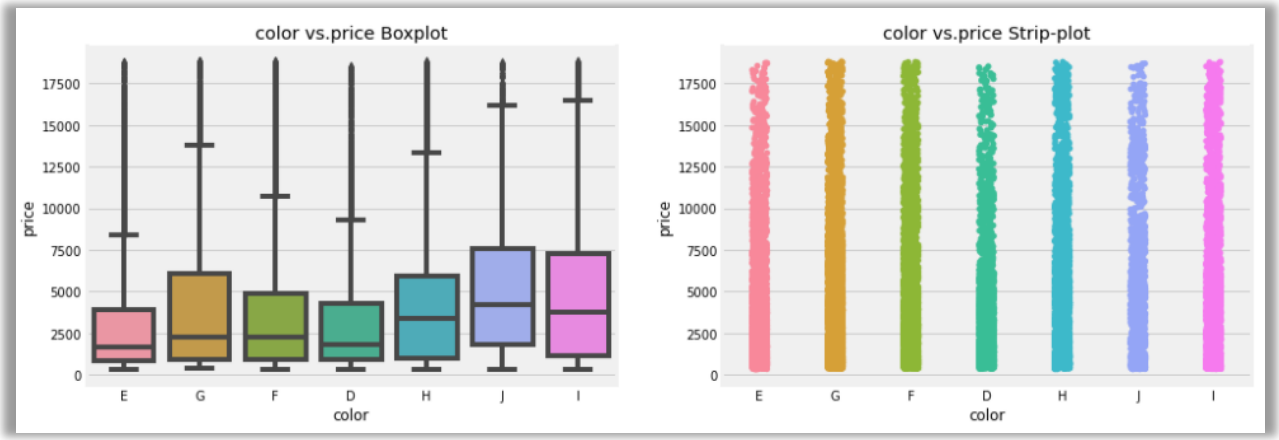
CUT Vs. PRICE:

- The "cut vs. price Boxplot" shows that Premium cut zirconia have higher prices.
- The "cut vs. price Strip-plot" shows that zirconia having 'Ideal', ''Permium' and 'Very Good' cuts have relatively the same price range.
    - Zirconia having 'Good' cuts mostly have prices around 10000, however, there are some data points where the prices are above 10000.
    - Data points having 'Fair' cuts mostly have prices around 6000, however, are a few surprising cases where the prices are above this value.



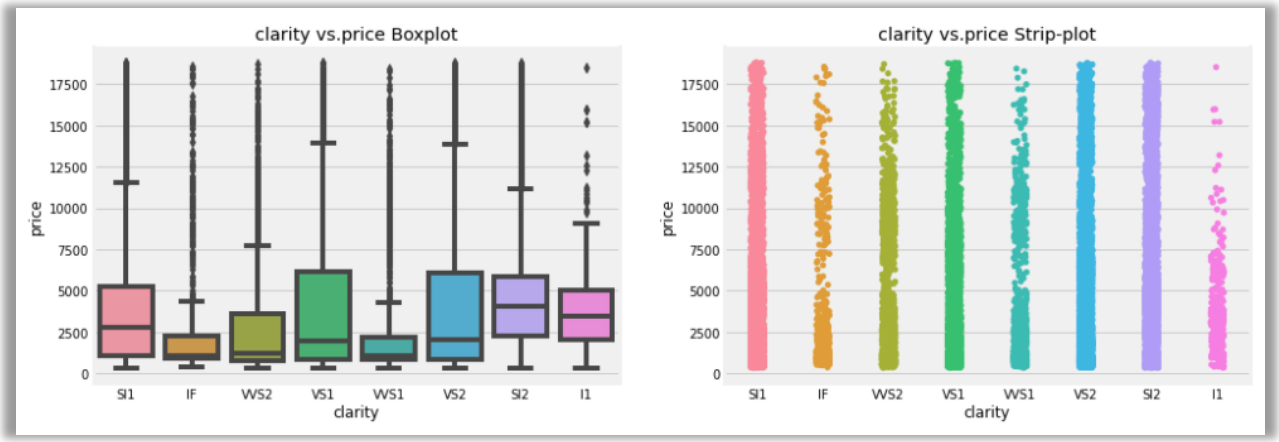
COLOR Vs. PRICE:

- The "color vs. price Boxplot" shows that there are a large amount of outliers (big price ranges) for all the color categories, with 'D' and 'E' having the higest amount of outliers.
- The "color vs. price Strip-plot" shows that the proce ranges for all the color categories appear to be relatively the same.
    - 'D' has prices mostly around 12500 and there is also a cluster of data points having prices between 15000 to 17500.
    - 'E' has prices mostly around 15000 and there are a few data points having prices above this value.

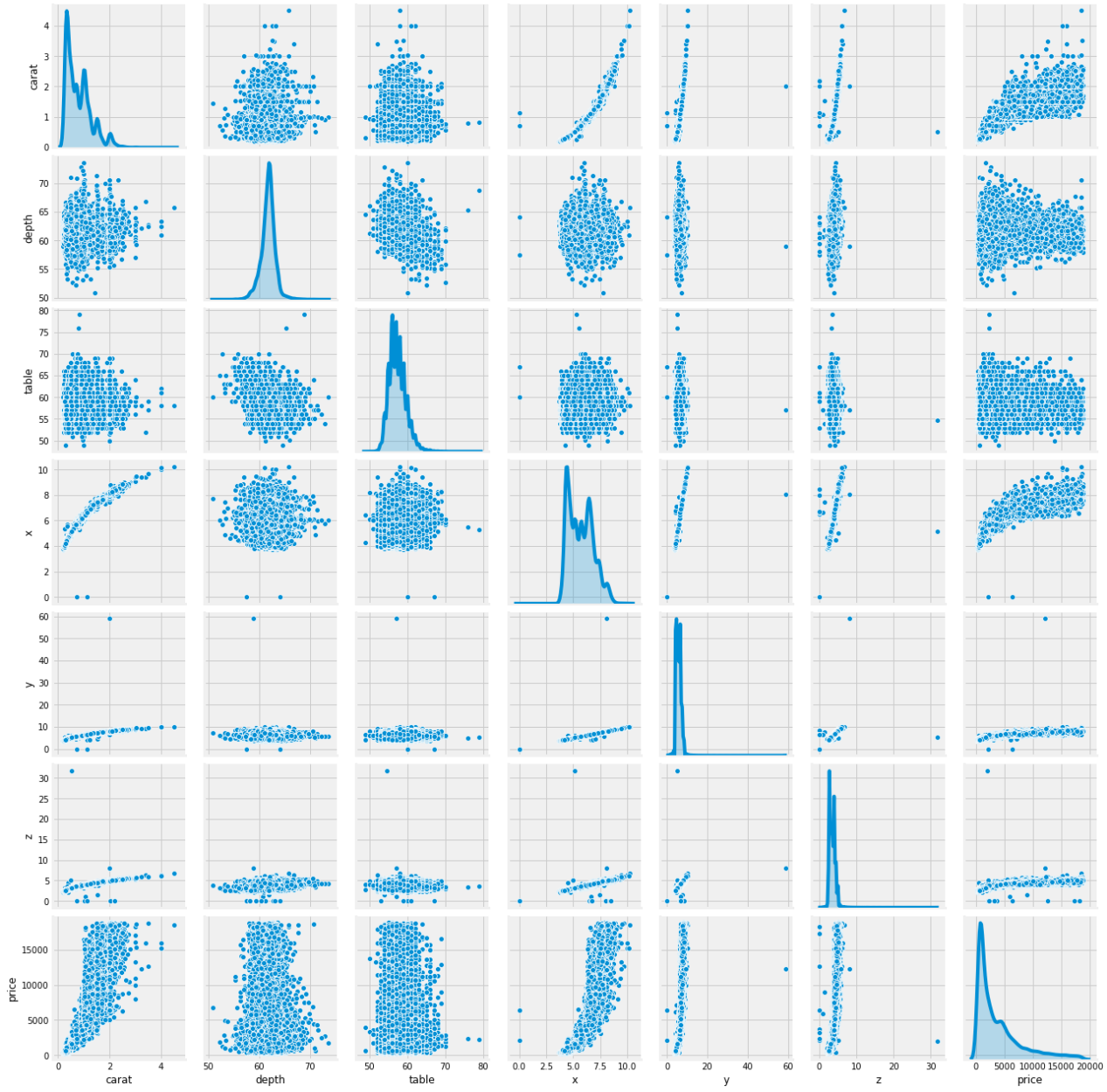


CLARITY Vs. PRICE:

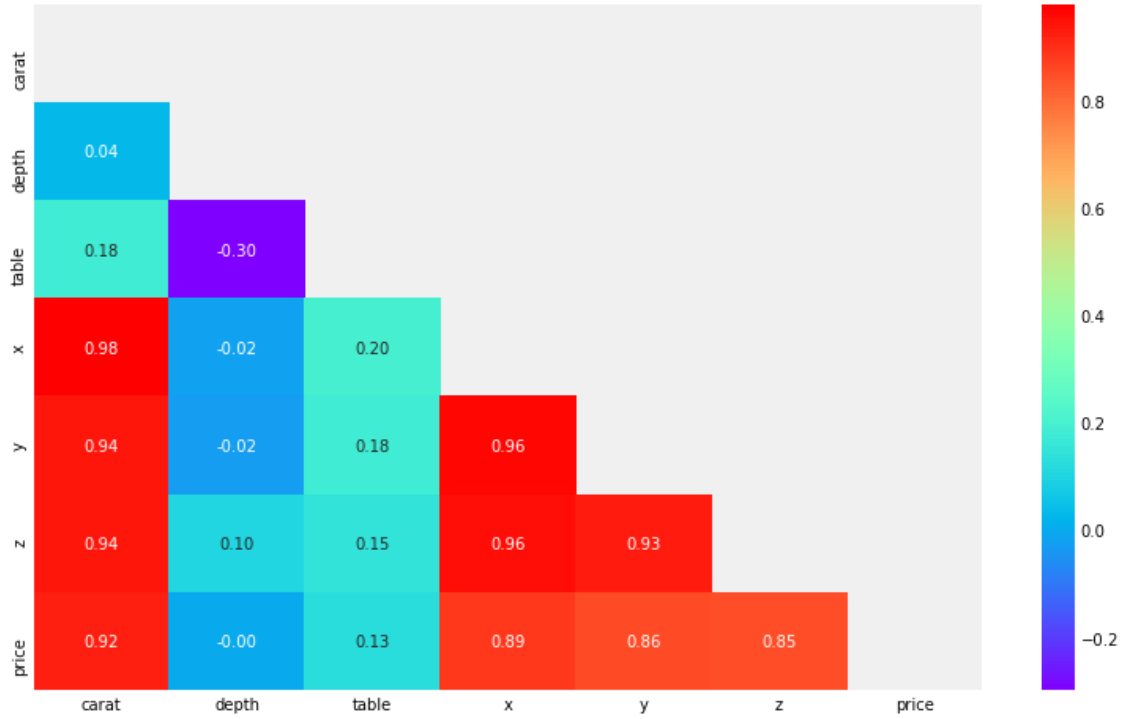
- The "clarity vs. price Boxplot" displays a large amount of outliers for clarity 'VVS1' i.e. they have prices even above the 75th percentile.
- The "clarity vs. price Strip-plot" shows that the prices for zirconia having 'IF' and 'I1' clarity are very less – below 5000 and 7500 respectively, however, there are some points above those ranges.



The below pairplot shows that features x, y, z and price seem to have a strong relationship with the target variable ‘price’.



The same can be confirmed in the below **Heatmap** which shows the correlations based on the Pearson method.



Inferences from Univariate/Bivariate analysis:

- The Univariate analysis shows that all the variables do not have a normal distribution.
  - It will be checked later if the independent variables can be normalized.

- The Bivariate analysis shows that features carat, x, y and z have very strong linear correlations with price, so the Linear Regression model can be used.
  - The features 'depth' and 'table' can be dropped.
  - The categorical variables cut, color and clarity will also be considered for the linear regression model.

1.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Do you think scaling is necessary in this case?

There are no null values in in the variables, however there are a few observations having x, y and z as 0 (as observed in the descriptive statistics).

	carat	cut	color	clarity	x	y	z	price
5822	0.71	Good	F	SI2	0.00	0.00	0.0	2130
6035	2.02	Premium	H	VS2	8.02	7.95	0.0	18207
10828	2.20	Premium	H	SI1	8.42	8.37	0.0	17265
12499	2.18	Premium	H	SI2	8.49	8.45	0.0	12631
12690	1.10	Premium	G	SI2	6.50	6.47	0.0	3696
17507	1.14	Fair	G	VS1	0.00	0.00	0.0	6381
18195	1.01	Premium	H	I1	6.66	6.60	0.0	3167
23759	1.12	Premium	G	I1	6.71	6.67	0.0	2383

The variables x, y and z correspond to the length, width and height of the zirconia. It makes no sense for an object to have its dimensions as 0, so the zeros will be replaced with 'NaN' and then replaced with the median of the respective column.

Scaling is necessary for this case study, as Linear Regression is a supervised machine learning model which is also a distance-based model i.e. it utilizes the Stochastic Gradient algorithm where different sets of weights (coefficients) and bias are computed at different points of the slope in order to find the best set of weights and bias. If the value ranges are different for all the features, then, there is a change where some variables may dominate the others in the contribution to the model. Hence, standard scaling will be done after splitting the data into the train and test sets in order to make all the feature variables contribute equally to the regression model.

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.

Encoding Categorical features:

The categorical columns (cut, color and clarity) are encoded based on their rank i.e. values are mapped to each category. For example, the cut 'Fair' is the lowest in terms of rank, hence 0 is assigned to this category.

The below screenshot shows the result of encoding the categorical variables.

	carat	cut	color	clarity	x	y	z	price
1	0.30	4	5	2	4.27	4.29	2.66	499
2	0.33	3	3	7	4.42	4.46	2.70	984
3	0.90	2	5	5	6.04	6.12	3.78	6289
4	0.42	4	4	4	4.82	4.80	2.96	1082
5	0.31	4	4	6	4.35	4.43	2.65	779

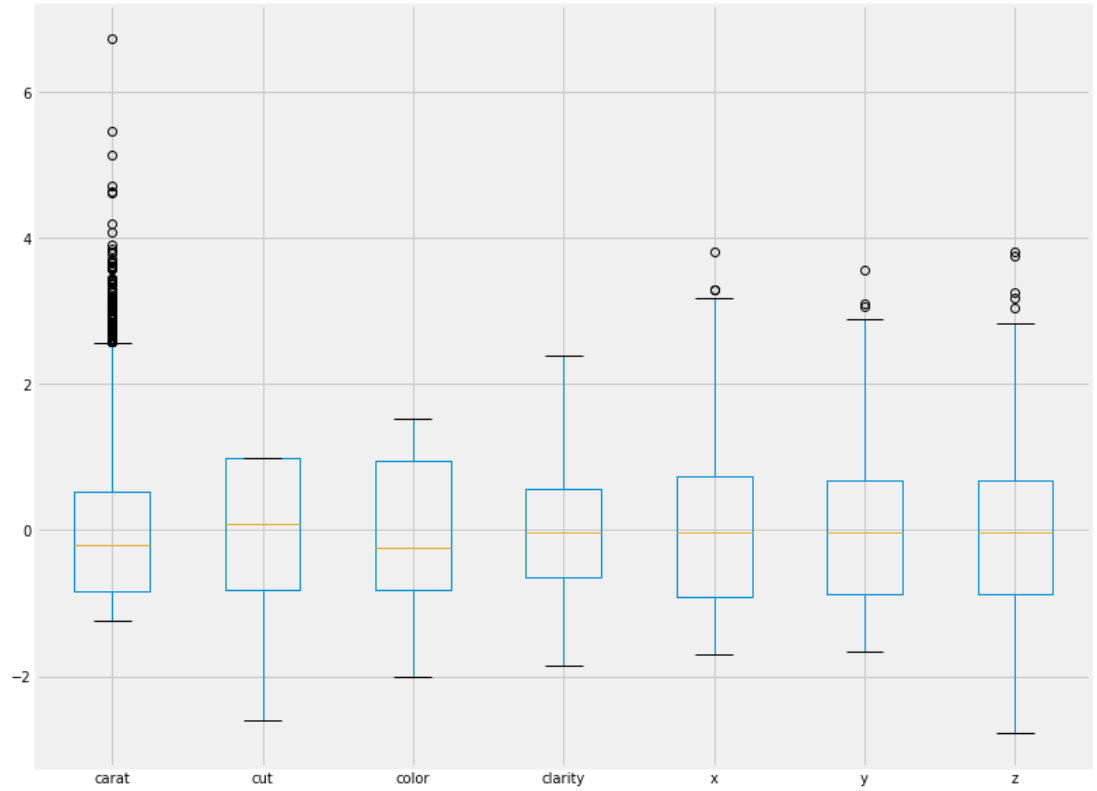
Data split:

The independent variables (all the features other than 'price') are taken as 'X', and the dependent variable 'price' is taken as 'y'. The 'X' and 'y' data are split into train and test data sets. The split is such that 70% of the data is for training and 30% is for testing.

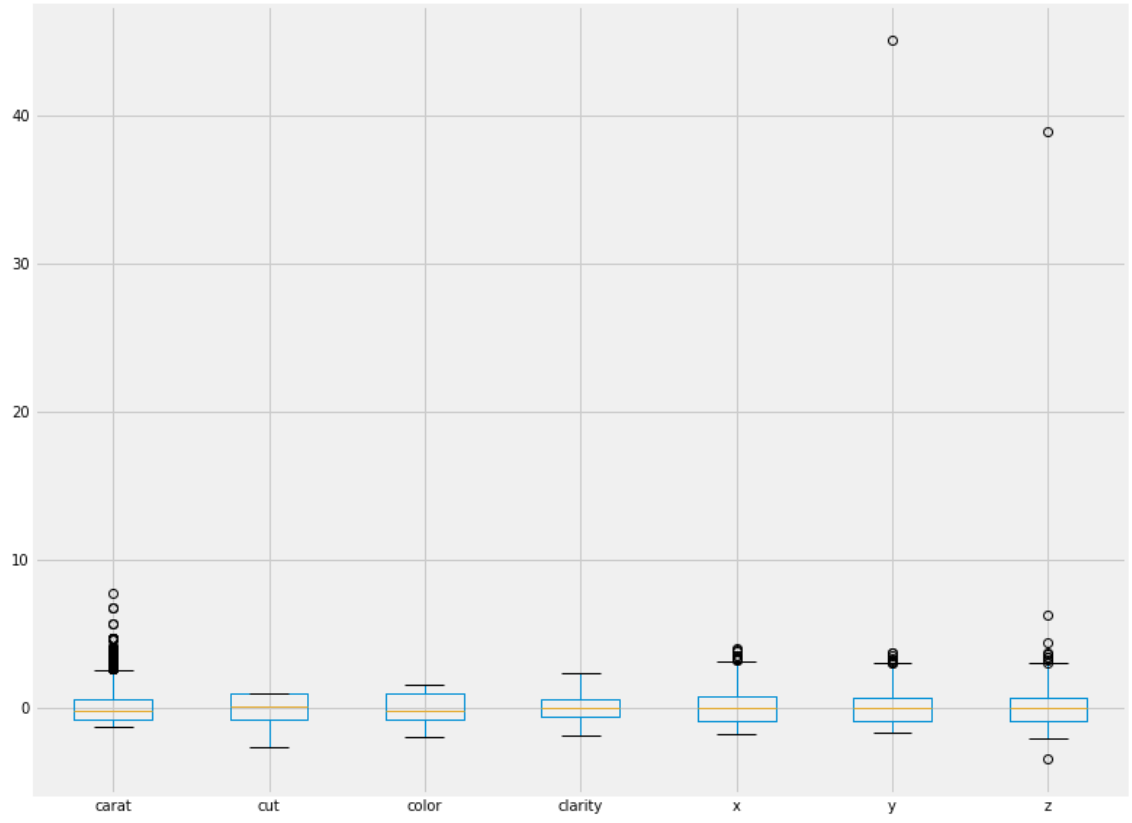
After the data has been split, standard scaling is applied to the dependent data train and test sets so that they have a data distribution of mean 0 and standard deviation 1.

From the below screenshots, we can see that the features have been scaled for X_train and X_test

X_train



X_test



Linear Regression:

The linear regression model is applied to the train sets for the dependent and independent variables i.e. X_train and y_train.

Performance Metrics:

The predictions are made for the train and the test data sets related to the independent variables.

And the performance metrics are summarized in the below table

Metrics	Train	Test
R-Square	90.7	91.01
Adj. R-Square	90.69	91.01
RMSE	1225.35	1208.7

Inferences are made from the above table:

- The R-Square values for the train and test data are almost equal.
- The Adjusted R-Square values for train and test data are also almost equal.
- The Root Mean Square Error (RMSE) values for the train and test sets are similar. The RMSE for the test set is slightly less than that of the train set, meaning that the model works well on the test set.
- The Adjusted R-Square and the R-Square values for the train and test sets are almost similar, meaning that no variables having poor impact on the dependent variable (price) were input into the model.

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

Insights:

- Around 75% of the observations have the zirconia weight (in carats) as 1.05
- The prices for zirconia having cuts 'Ideal', 'Premium' and 'Very Good' are high.
- The features – carat, x, y and z have strong positive-linear correlations with price, with carat having the highest correlation.
- Below the linear equation displaying the relationship between the independent variables and the target variable price.
- **(5103.21 * carat) + (181.01 * cut) + (569.97 * color) + (836.48 * clarity) + (-840.07 * x) + (40.2 * y) + (-137.2 * z)**
- The coefficients before the variable name are good indicators of the impact of the feature on price.
- For example: The coefficient for carat is 5103.21, which means that for every unit increase in carat (while keeping other variables constant), there is 5103.21 units increase in price. Hence, carat is a significant variable for price prediction.
- ***The top 5 attributes for predicting price (based on the coefficients of the above equation) are:***

  carat, x, clarity, color, cut

Recommendations:

- It appears that most customers are interested in buying 1.05 carats zirconia, so zirconia that around this weight can be sold more.
- Zirconia having cuts 'Ideal', 'Premium' and 'Very Good' can be produced more.
- Zirconia with clarity 'SI1', 'SI2' and 'VVS2' are recommended since there are a lot of data points with high price ranges for these particular clarity values.