Balaram Tripathy

Project - Predictive Modeling

Problem 1: Linear Regression

Qn 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.

Ans 1.1. EDA of the Data Cubic Zirconia as follows.

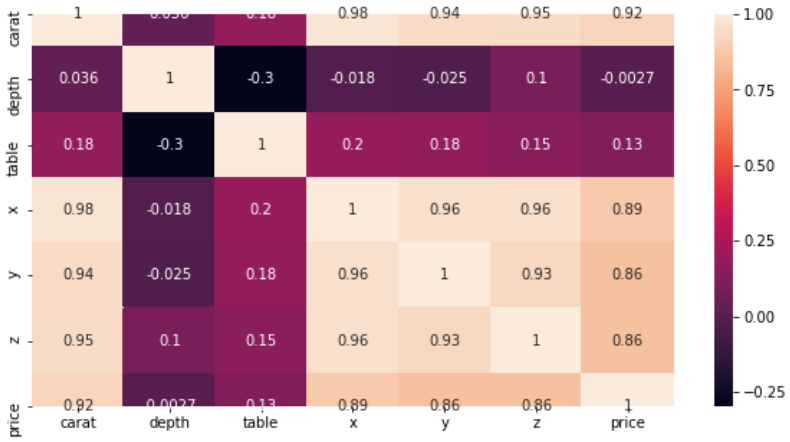
Shape of the data: It has 11 variables (including unnamed-index Variable)
and 26967 observations (including Header).
Delete Variable: 1st unnamed variable is no use for the analysis. So, we will delete.
Duplicate Values: After we remove the Unnamed column, we got 33 duplicate values.
So, we can delete them.
Data Types: Categorical Variables: cut, color, clarity.
Continuous Variables: carat, depth, table, x, y, z, price.
Missing Values (Null Values): Variable "depth" has 697 missing values.
Target Variable: price

Insights of Univariate Analysis:
Price range 344 to 1944 has maximum sales. As the price increase the no. of sales is decrease.
I.e. low-cost cubic zirconia has a good demand.

Cut - Ideal cut products are more in demand. And Fair cut products less in demand.
Color – Color G has more in demand and J has less demand.
Clarity - Clarity SI1 has more in demand and I1 has less demand.

Outliers: We have observed Outlier are present in the dataset. So, we need to do Outlier treatment required for this data.

Bi Variate Analysis:
Observed that carat, x, y and z variables have good correlation with price (dependent variable).
We can drop "depth" and "table" also. Less correlation with price and in gemstone we can define from carat, x (length), y(width) and z(height).



Qn 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?

Ans 1.2.

0 values: Observed that the minimum value of x (length), y(width) and z(height) is zero and It doesn't make any sense to have length\width\height of a diamond to be zero.
The maximum zeros we can see in variable z(height) has 9 records, which is 0.033% of total records.
- So, we can delete those observations.

Missing Values (Null Values): Variable "depth" has 697 missing values.
Which is 2.59% of total observations.

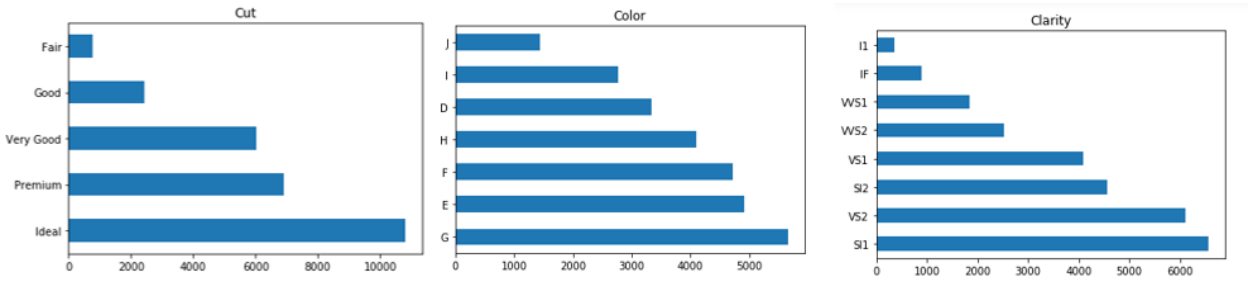
Missing Value treatment: Mean and Median of the variable "depth" is nearly same (i.e. 61.8). So, we will fill the missing values by the mean of "depth".

Scaling:
Scaling not required, after the data conversion we do on point 1.3 (below).
Even also we do scaling, result not change.

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

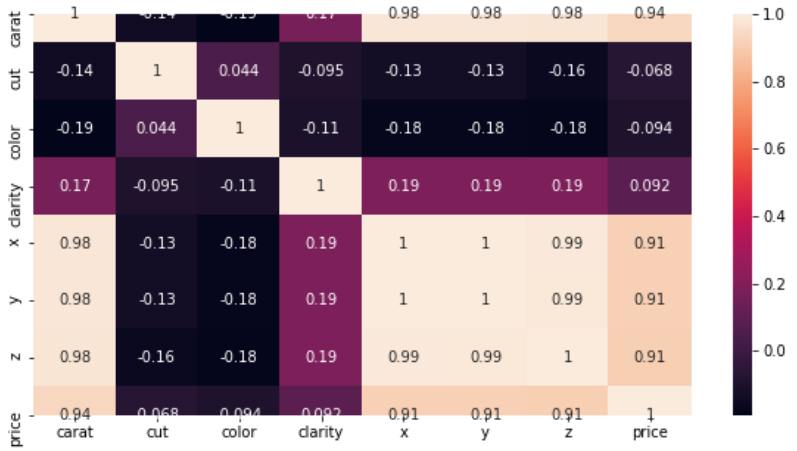
Ans 1.3.

Encoding / Convert Categorical data to Continuous data type:
We have 3 categorical variables ("cut", "color" and "clarity").



As we can observe the ascending trend in values in all three categorical variables.
So, we can manually convert them to continuous variables based on each categorical value (w.r.t to their ascending count).

Correlation after the above conversion.



"cut", "color" and "clarity" has less correlation with price, but they are important features for gemstone. As carat makes value count from "cut", "color" and "clarity" of a gemstone, we can have calculated Features (like cut/carat) and drop these variables.

Alternatively, we can do Label Encoder or One-hot Encoding or create dummies.

Data Split:
Out of 26925 observations. We split the train and test set with 70:30 ratio.
We got 18847 observation in Train set and 8078 in test set.

R-Square Value (After Liner regression):
R-Square of Train data is 0.89
And R-Square of Test data is 0.89

Both provide a good result of 89%. I.e. the model is normal (neither overfit or nor underfit).

RMSE value (After Liner regression):
RMSE value of Train data is 1158.09
RMSE value of Test data is 1163.01

Not much difference (considering RMSE value for this model is high, in 4 digit).

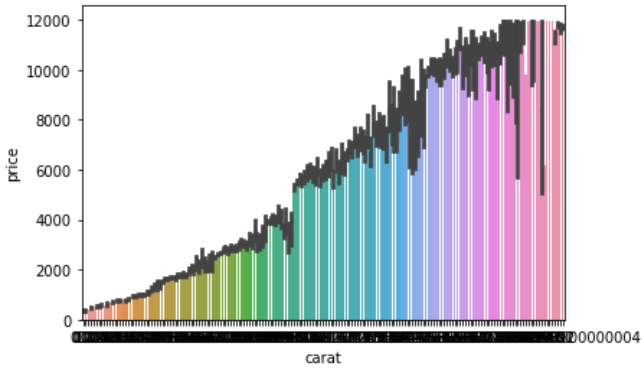
Following is the model formula (with Intercept, and coefficient of independent variables).

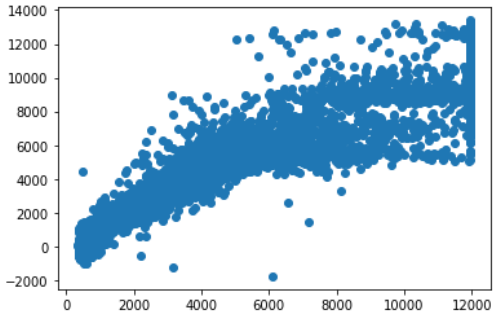
(-3988.35) * Intercept + (7432.69) * carat + (-2576.31) * x + (3156.13) * y + (-619.01) * z + (73.5) * cut_carat + (50.46) * color_carat + (-25.51) * clarity_carat

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

Ans 1.4.

A. We have observed there is moderate positive correlation between carat and cubic zirconia price. When carat increases, Price also increase.



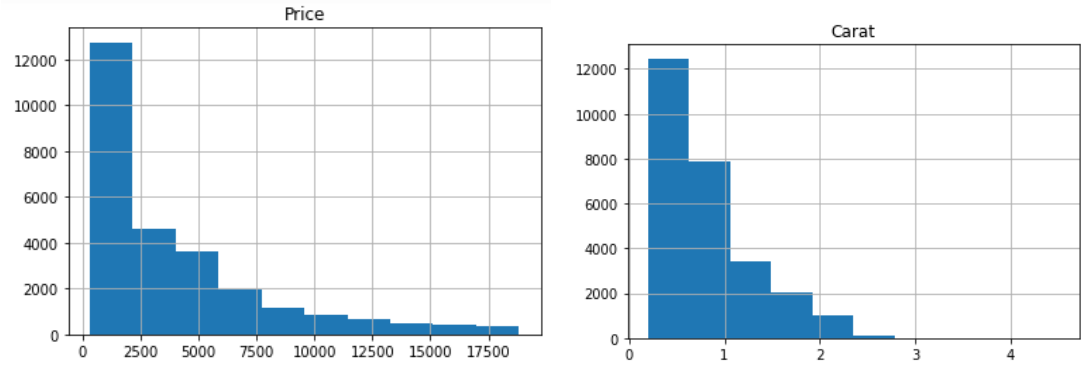


Model shows increase in carat Price also increases. If the cubic zirconia is increase by 1 carat, price will increase by 7432.69.

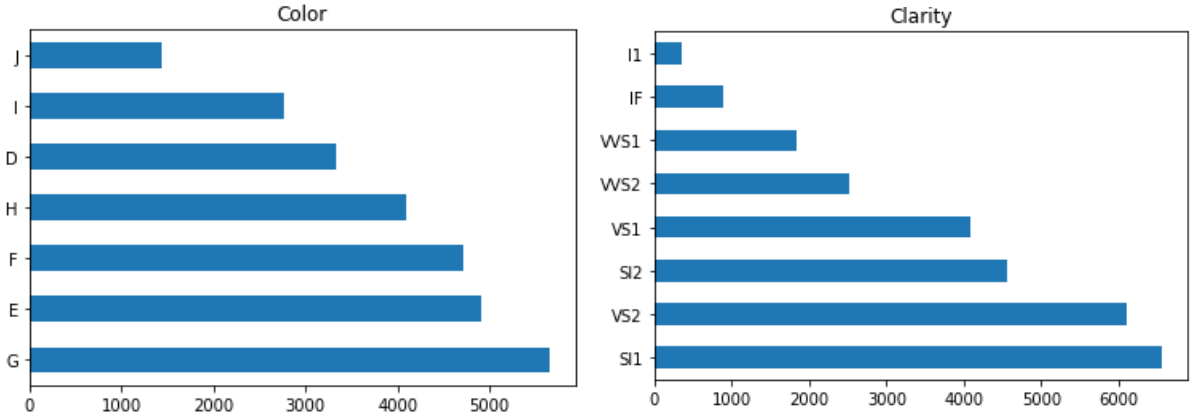
B. We have observed Cubiz Zironica which has less than 1 carat, average price/carat are 4 times higher than higher carat products (including actual and predicted price).
Also, we have observed Less carat and less price Cubic zironica are more in demand.

So, business should carry on less carat products as most sales product.

And same time Business should do some activity on how to increase the sales on high carat products.



C. We have observed Color G and Clairity SI1 has good demand.



D. 5 best important attributes are – ‘Carat”, ‘Width”(y), ‘Clarity’, ‘Color” and “Cut”.