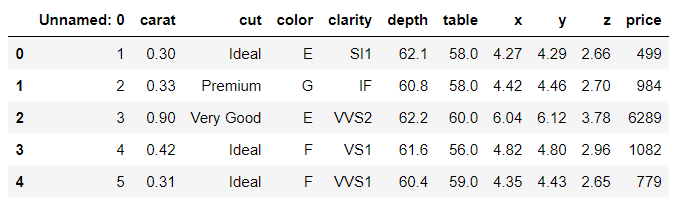
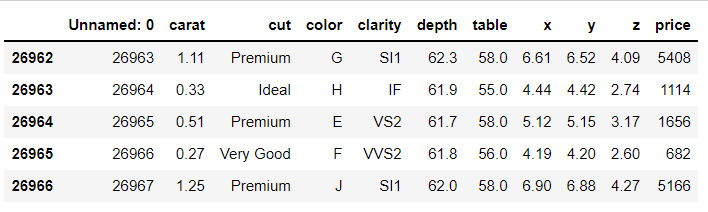
* 1. **Read the data and do exploratory data analysis. Describe the data briefly. Perform Univariate and Bivariate Analysis.**

Check the head and tail of the data to confirm the data frame.

HEAD



TAIL

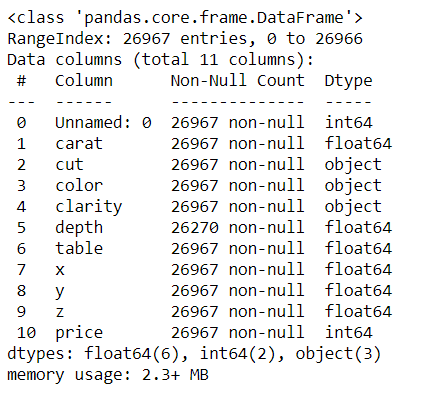


SHAPE

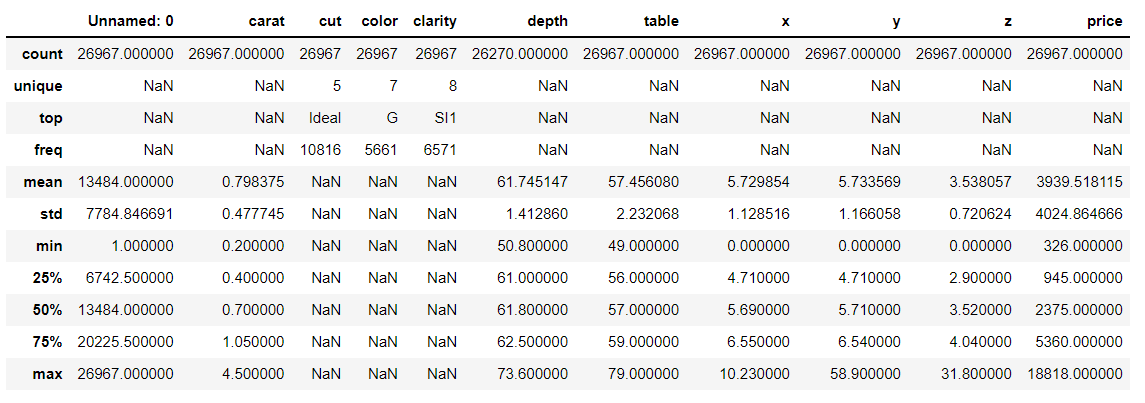


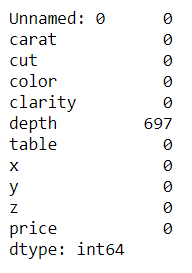
INFO

Unnamed is redundant in our analysis and can be removed later, we have three object and rest integer/float dtypes.



DESCRIBE

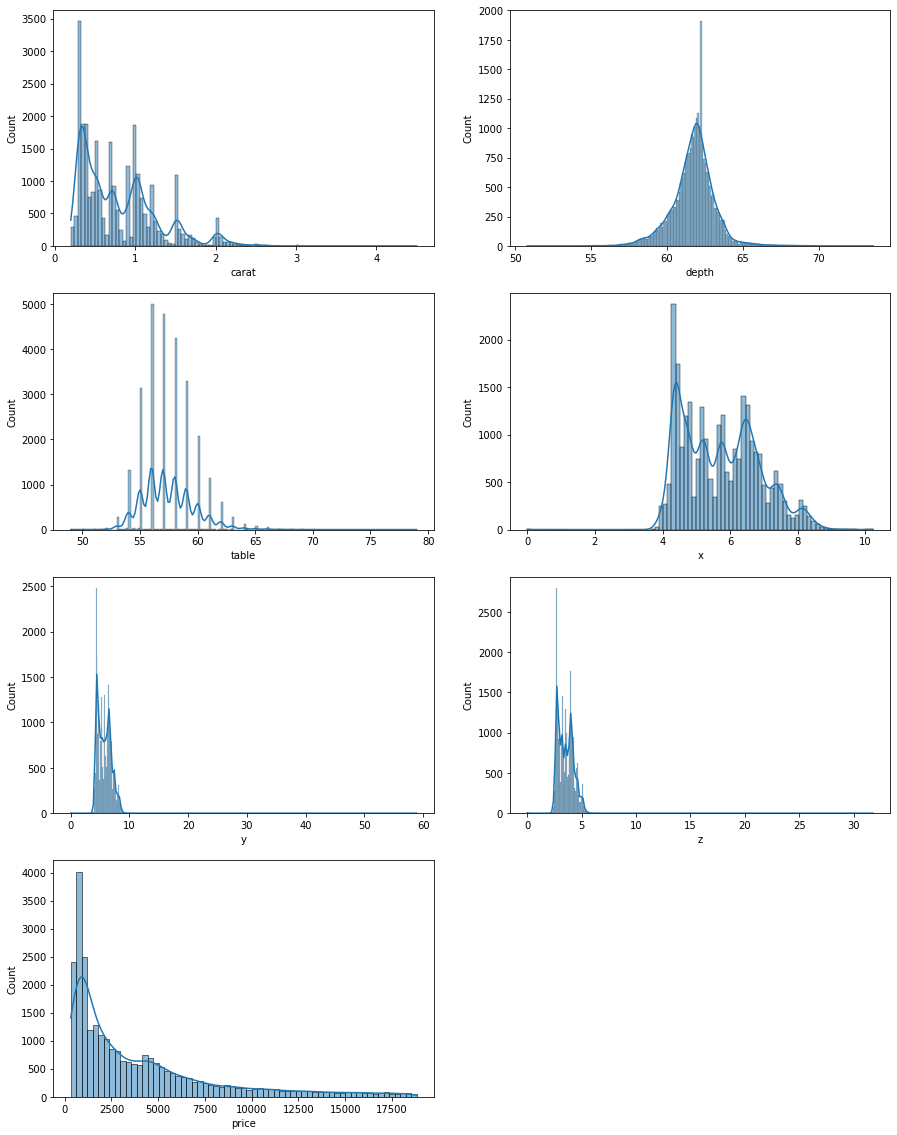


NULL VALUES  


DUPLICATED ROWS



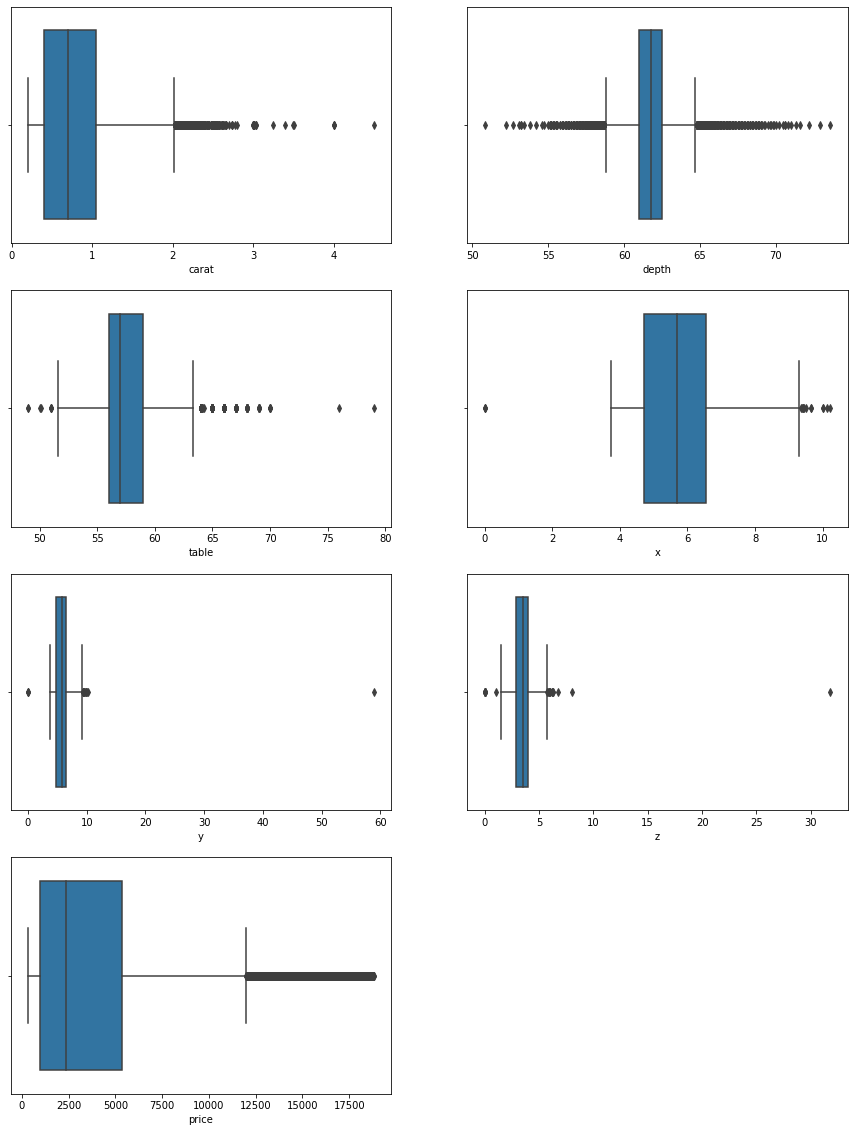
UNIVARIATE AND BIVARIATE ANALYSIS



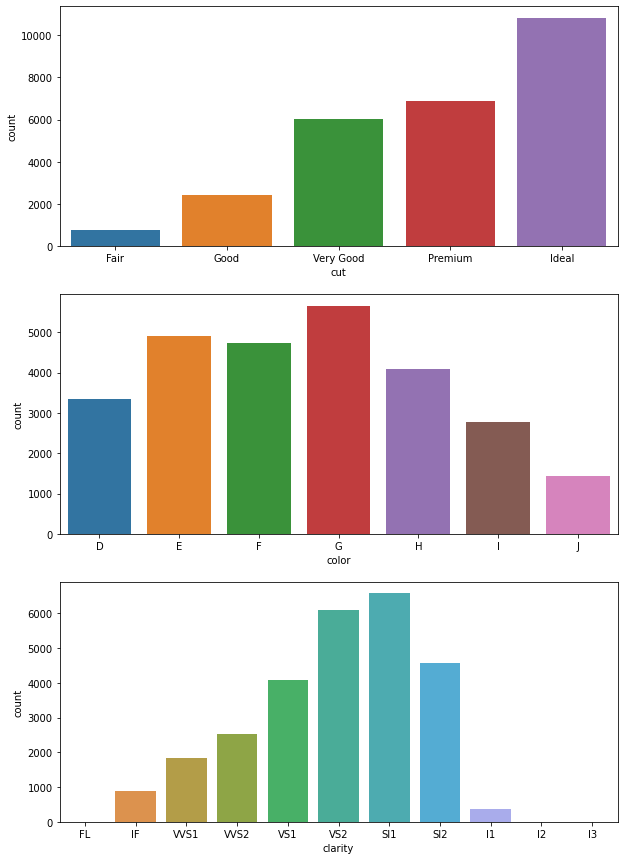
Price is heavily right skewed.

Depth is normally distributed.

Carat, table has multiple peaks, indicating preferred values.

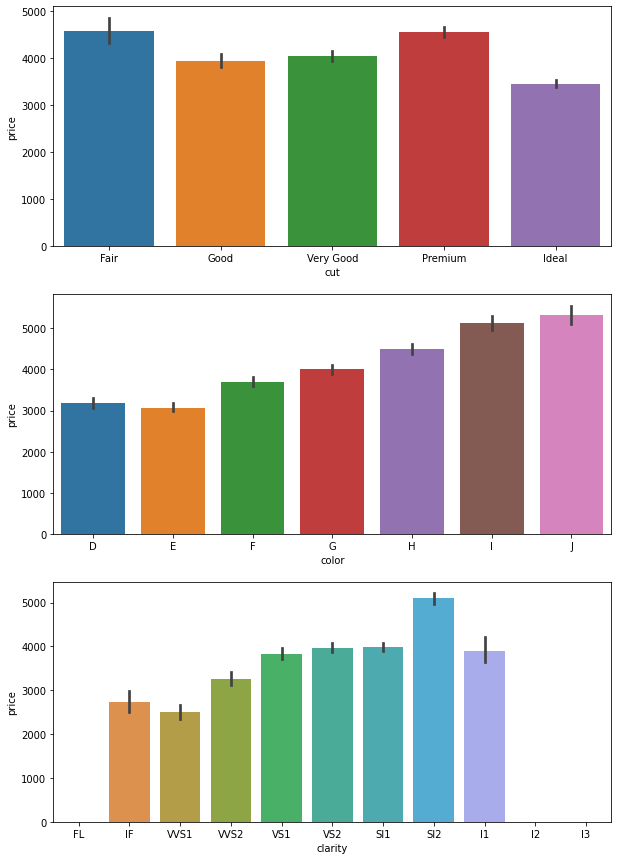


We can notice a lot of outliers in our data.

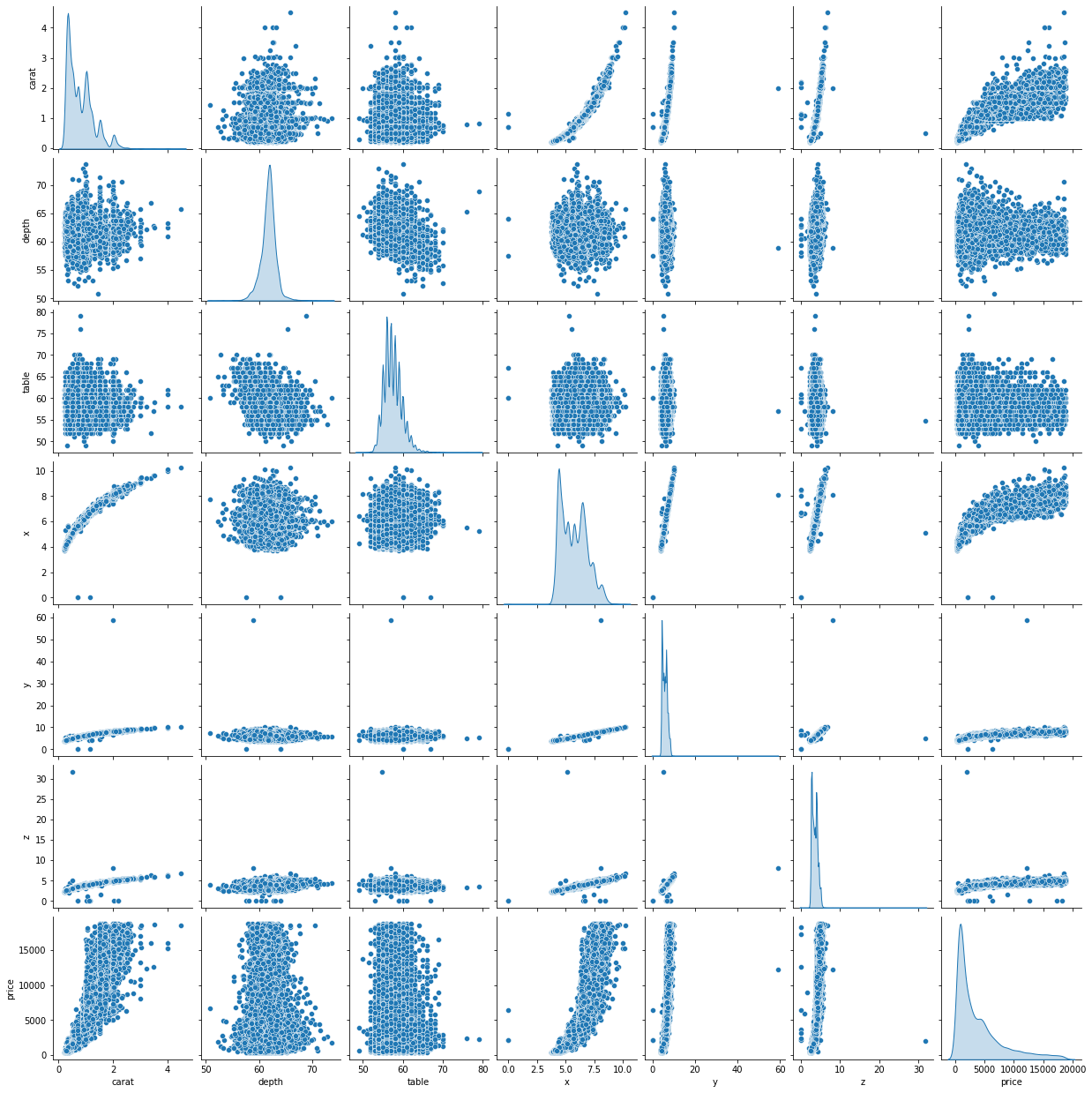
****

**Ideal cut has high count because of lowest price.**

**In clarity- VS1, VS2, SI1, SI2 are having max share. These are generally priced higher and may look more beautiful to the end consumer.**

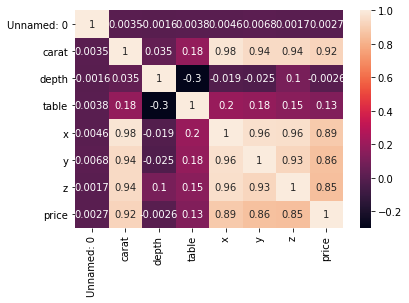
****

**Pair plot:**

****

**Correlation:**

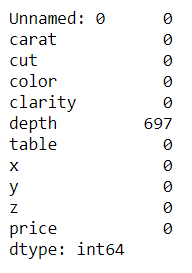
**We can see multi collinearity present in dataset.**

****

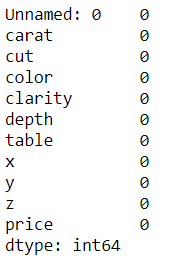
**------------------------------------------------------------------------------------------------------------------------------------------------------------------------------**

* 1. **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?**

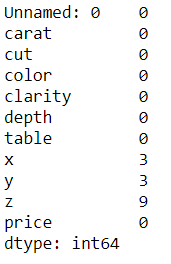
As we saw earlier 697 null values were present in the dataset.



After replacing null values of depth with median.



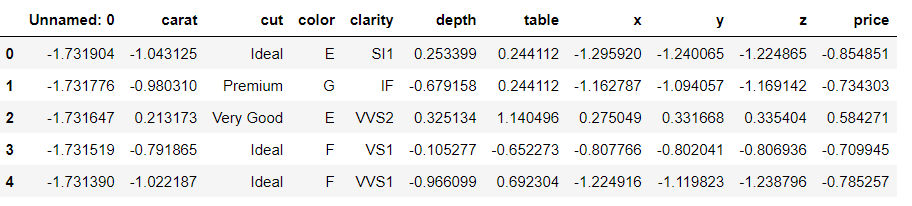
Entries where value is equal to zero.



As these entries are so few, we have removed them from the dataset.

Scaling can be done, as some values are very far apart for example (carat, depth), (carat, x) etc

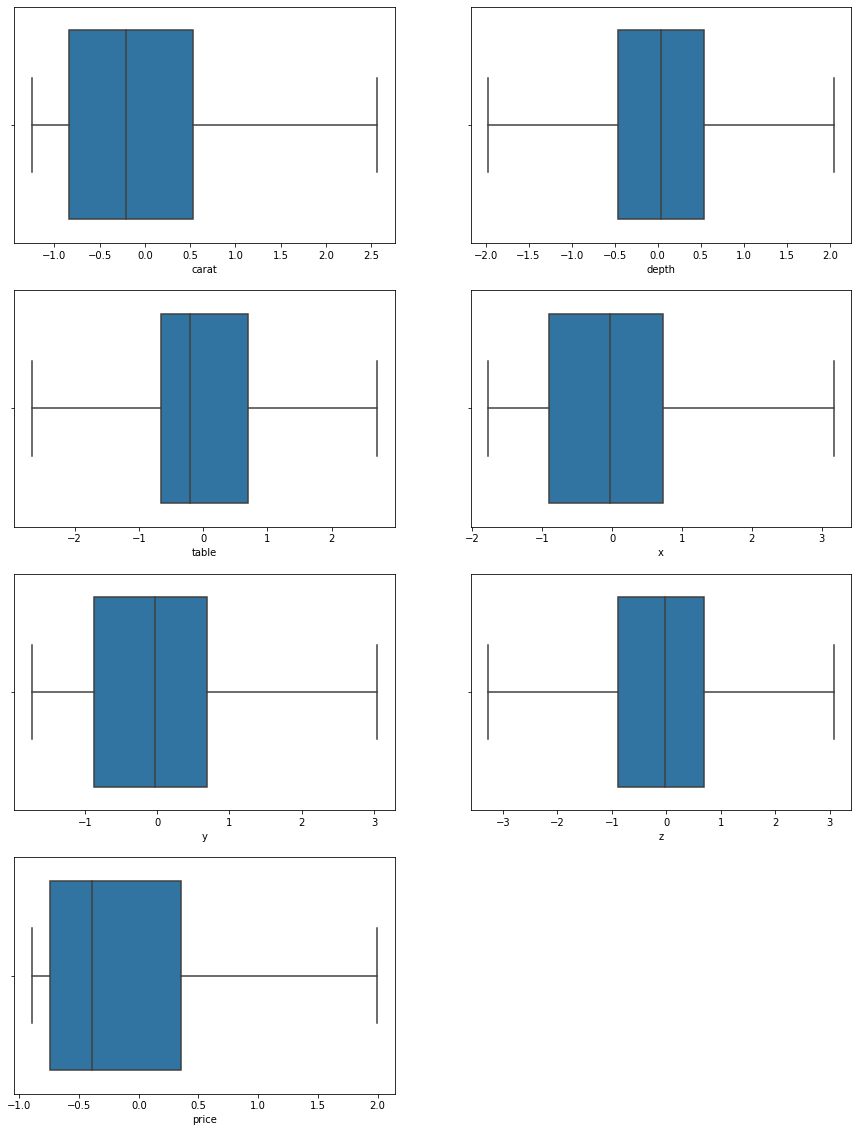
After scaling:



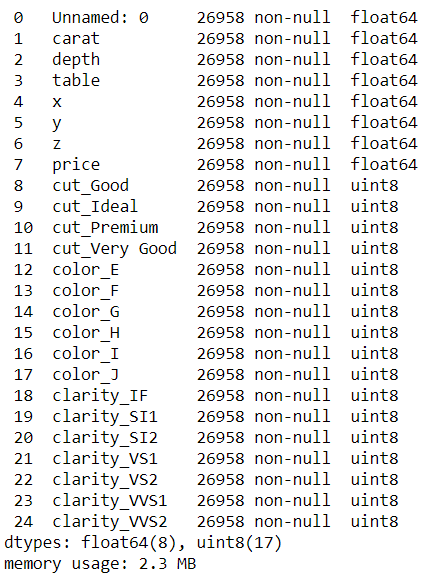
**------------------------------------------------------------------------------------------------------------------------------------------------------------------------------**

* 1. **Encode the data 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.**

After removing outliers:



Converting categorical data to integer using dummy variables:



Remove the unnamed col as it is redundant.

Coefficients after applying linear regression model:

The coefficient for carat is 1.100941784780458

The coefficient for depth is 0.0056051434455705595

The coefficient for table is -0.013319500386803969

The coefficient for x is -0.3050434981963325

The coefficient for y is 0.3039144895792643

The coefficient for z is -0.1391657156798792

The coefficient for cut\_Good is 0.09403402912977914

The coefficient for cut\_Ideal is 0.15231074620567503

The coefficient for cut\_Premium is 0.14852774839849478

The coefficient for cut\_Very Good is 0.12583881878452835

The coefficient for color\_E is -0.04705442233369901

The coefficient for color\_F is -0.062684374391429

The coefficient for color\_G is -0.10072161838356845

The coefficient for color\_H is -0.20767313311661725

The coefficient for color\_I is -0.32395419274627507

The coefficient for color\_J is -0.4685893027501597

The coefficient for clarity\_IF is 0.9997691394634898

The coefficient for clarity\_SI1 is 0.6389785818271321

The coefficient for clarity\_SI2 is 0.4295966234831557

The coefficient for clarity\_VS1 is 0.8380875826737574

The coefficient for clarity\_VS2 is 0.7660244466083612

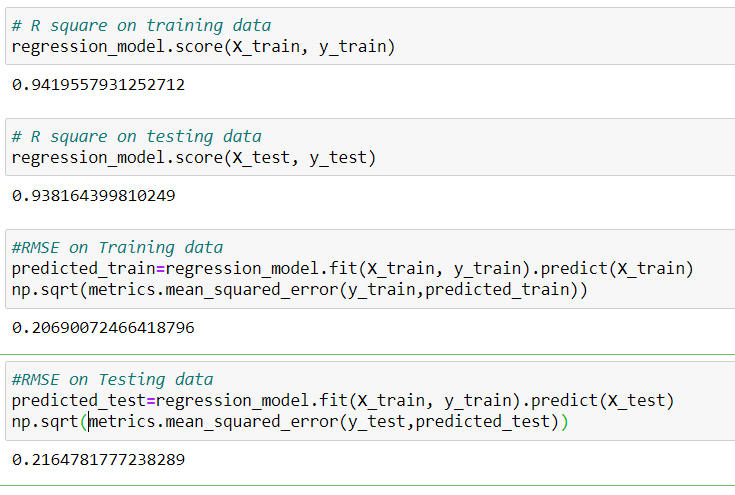
The coefficient for clarity\_VVS1 is 0.9420769630114086

The coefficient for clarity\_VVS2 is 0.931367028841569

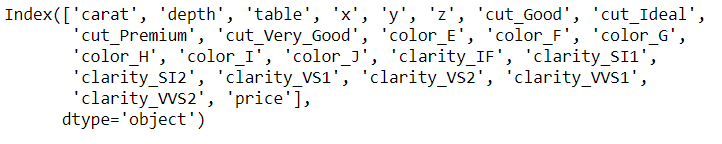
Intercept of the model:

The intercept for our model is -0.7567627863049388

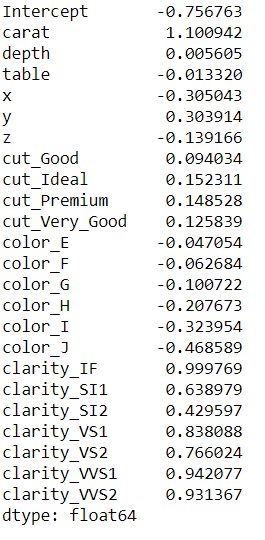
Test errors:

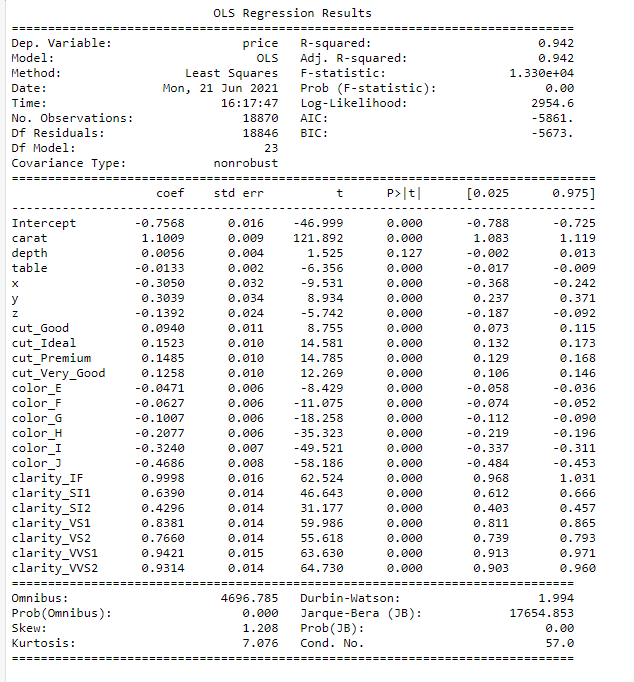


Linear regression using STATSMODEL:

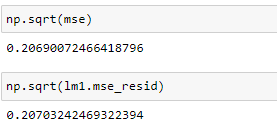


Coeff and intercepts:

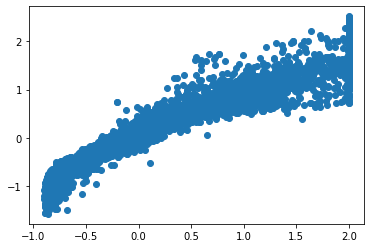


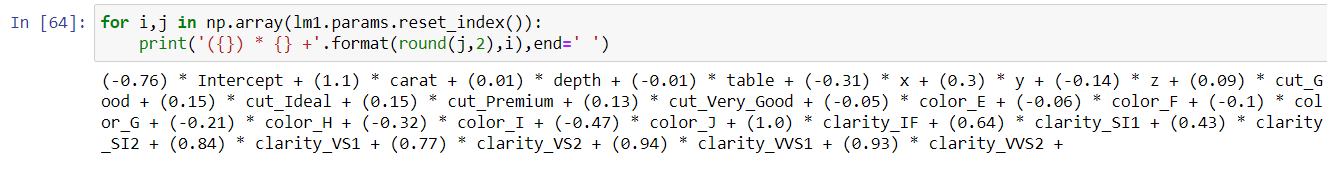


We can use the adjusted R^2 value to say that we are able to explain 94.2 percent of variance in the data. Which is a good result.



Y vs y pred

****



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

Clarity is the best indicator after y parameter.

VS1, VS2, SI1, SI2 these have the majority of sales in clarity. So, the company should focus on them.

The cut is also an aspect to focus on, we saw that Ideal, premium and very good (In decreasing order), so company should focus on these three cuts only.

The business should have their main focus on clarity, as its coefficient is the highest after that focus on y and then cut.

With our model we were able to capture 94.2 % of variation which is very good. But we have discarded the outliers in our observations, this should also be considered while presenting the report.