**Problem 1**: Linear Regression

### Problem Statement:

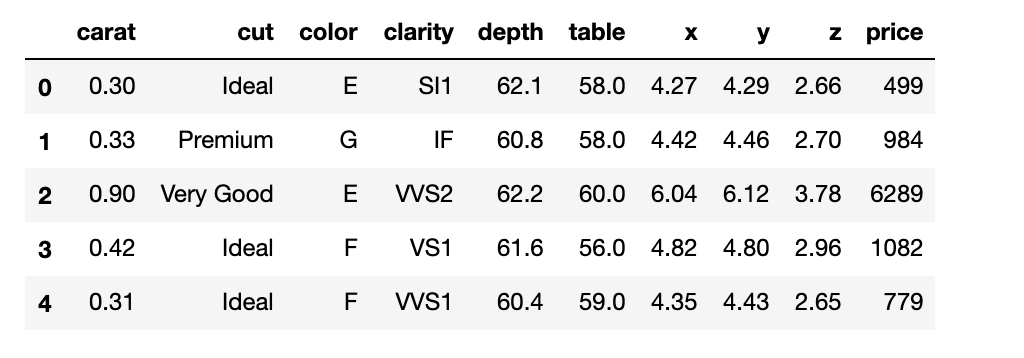
You are hired by a company “Gem Stones co ltd”, which is a cubic zirconia manufacturer. You are provided with the dataset containing the prices and other attributes of almost 27,000 cubic zirconia (which is an inexpensive diamond alternative with many of the same qualities as a diamond). The company is earning different profits on different prize slots. You have to help the company in predicting the price for the stone on the basis 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.

#### 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.

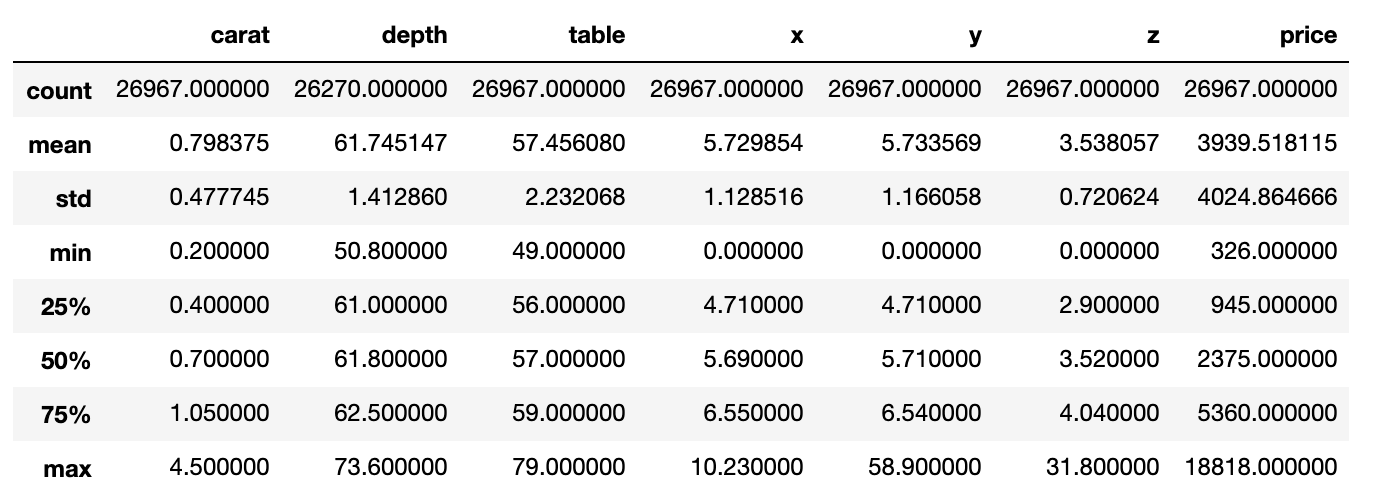
Dataset has 26967 observations and 10 attributes.

* 6 attributes named carat, depth, table, x, y, z are of float64 type
* 3 attributes named cut, color, clarity are of object type
* 1 attribute named price is of integer type

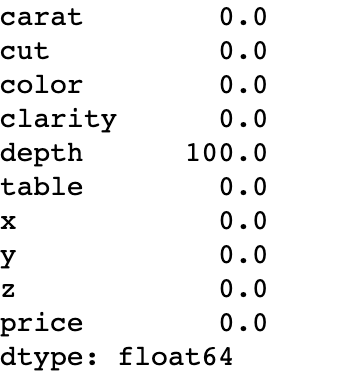
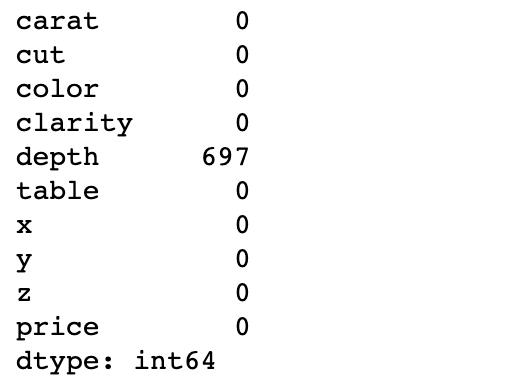
Let’s start the data exploration step with the head function to look at the first 5 rows.



After checking the summary of the data, there seems to be some outliers in the dataset. Let’s check them by doing further exploration.



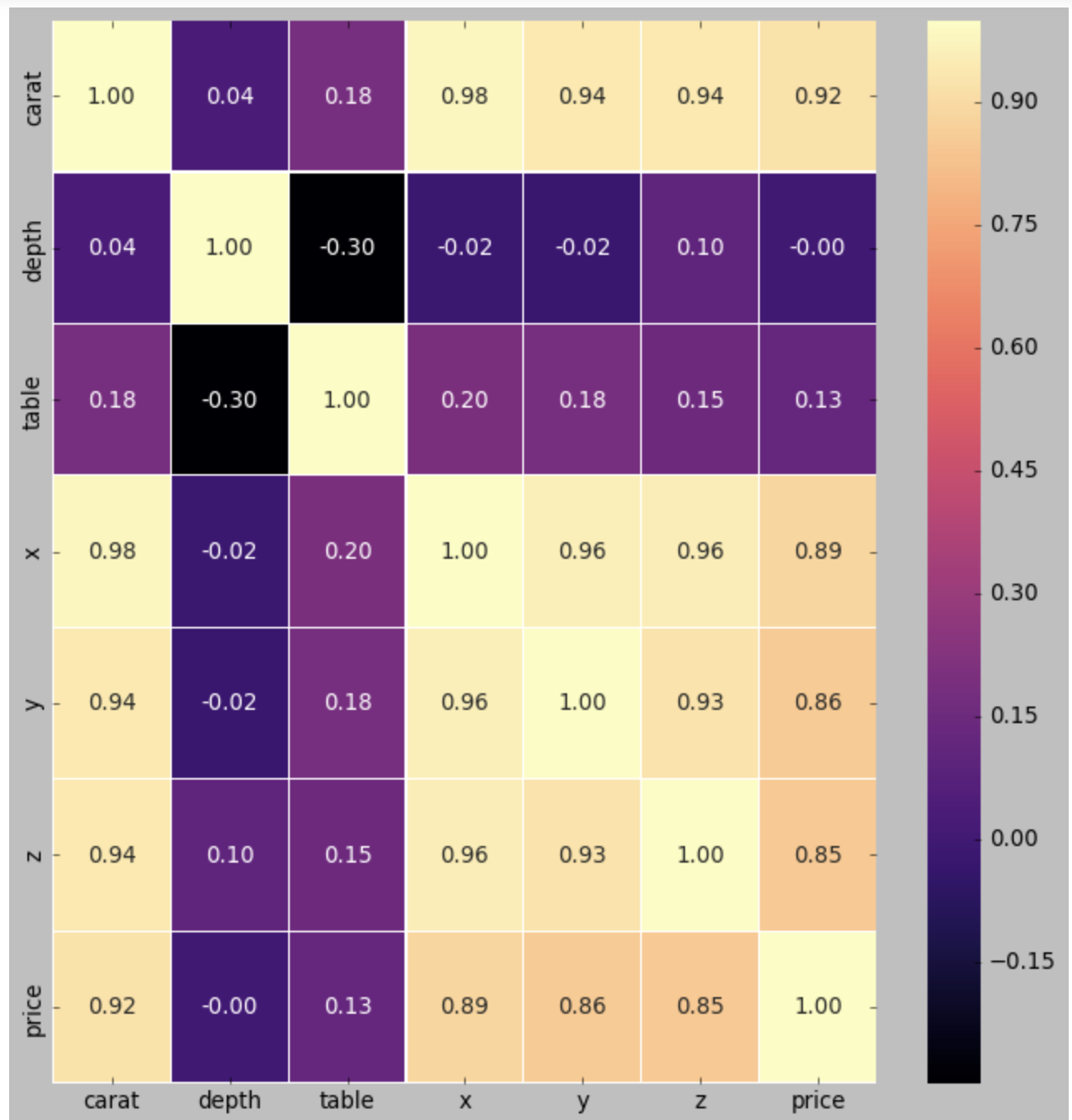
Dataset has null values in the depth column. By checking the percentage of null values, it seems to be that there are null values only in one column that is depth.



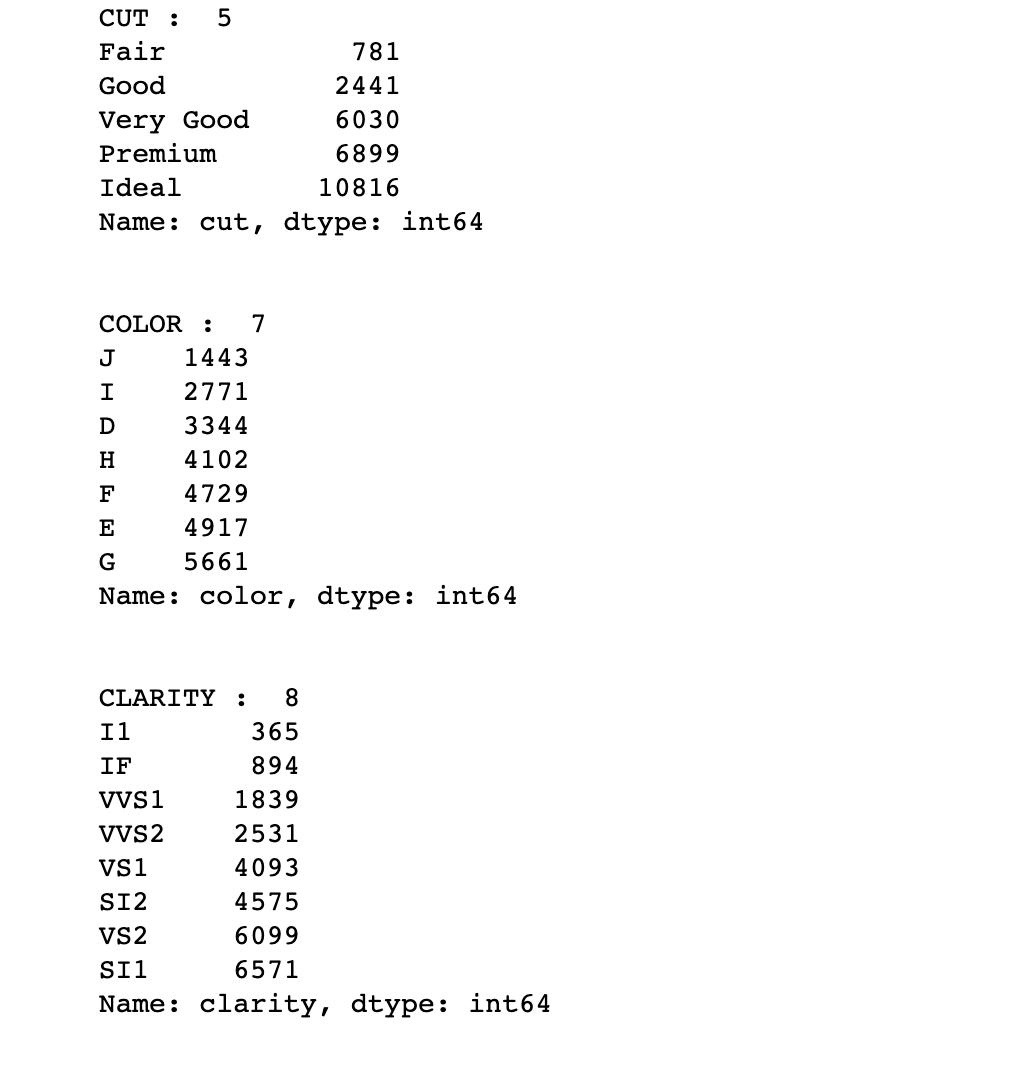
Heat map shows that the below parameters of the cubic zirconia are **highly correlated** with each other

* Carat weight
* Length in mm.
* Width in mm.
* Height in mm.
* Price

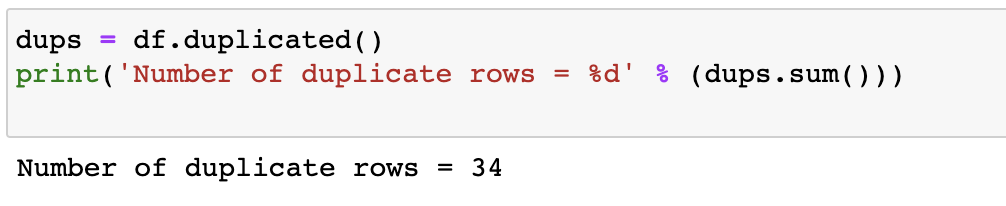
The Height of a cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter is **least correlated** with all the attributes.



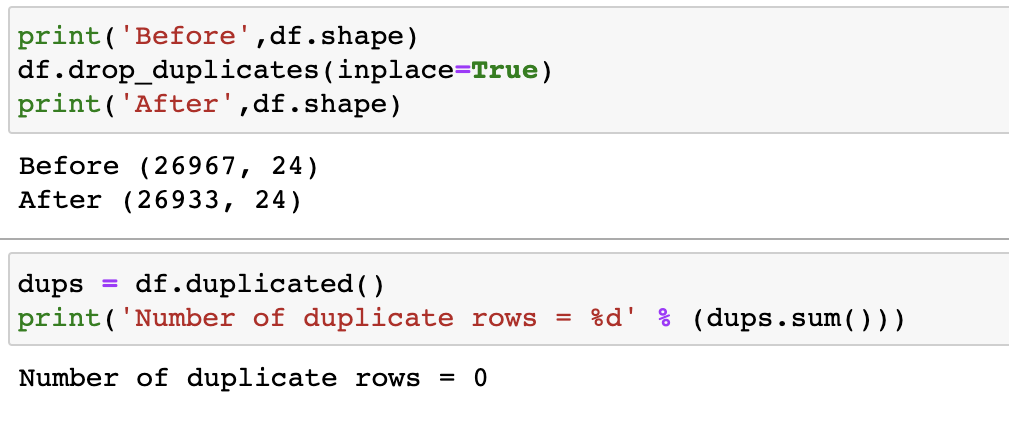
Let’s check the unique values for the object data types - CUT, COLOR and CLARITY



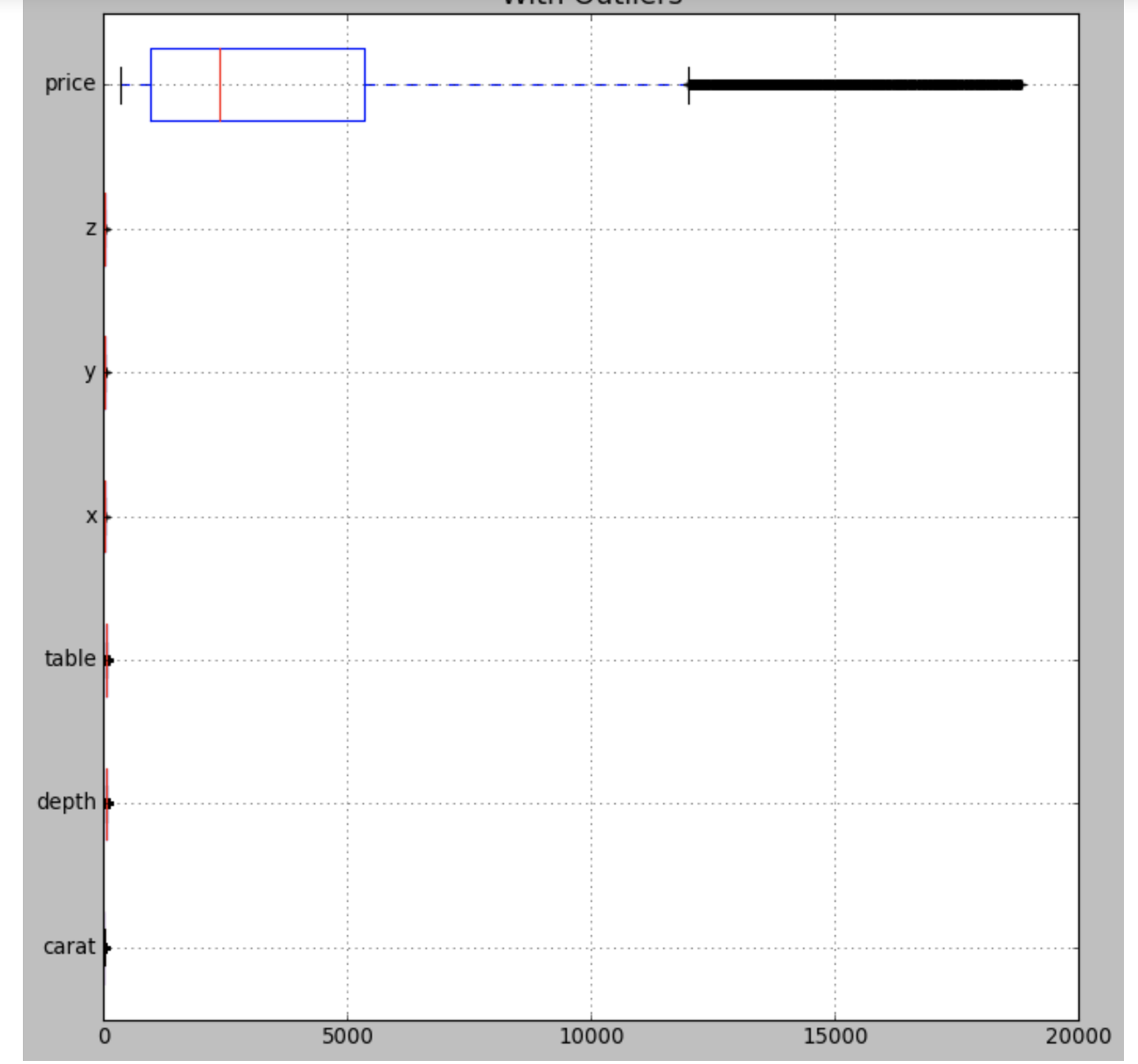
There seems to be 34 duplicates in the dataset.



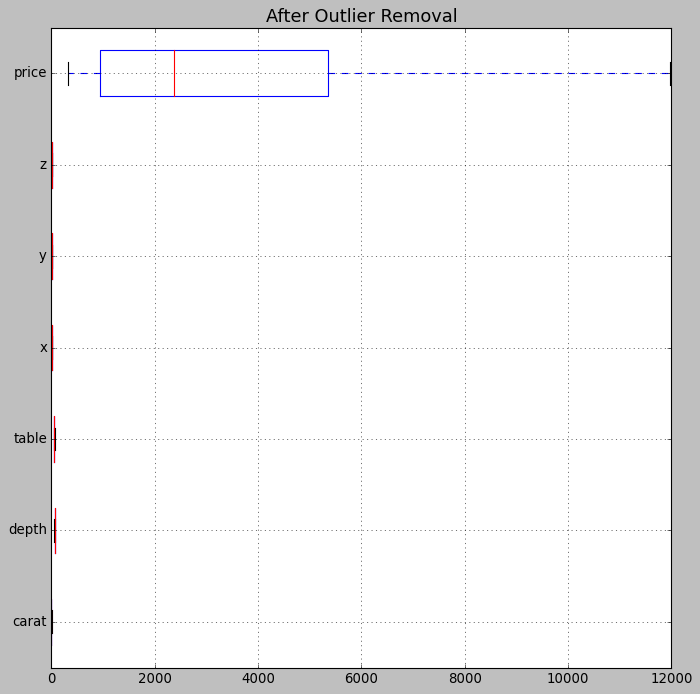
Let's drop them for now. The dataset before and after removing duplicate rows is as shown below.



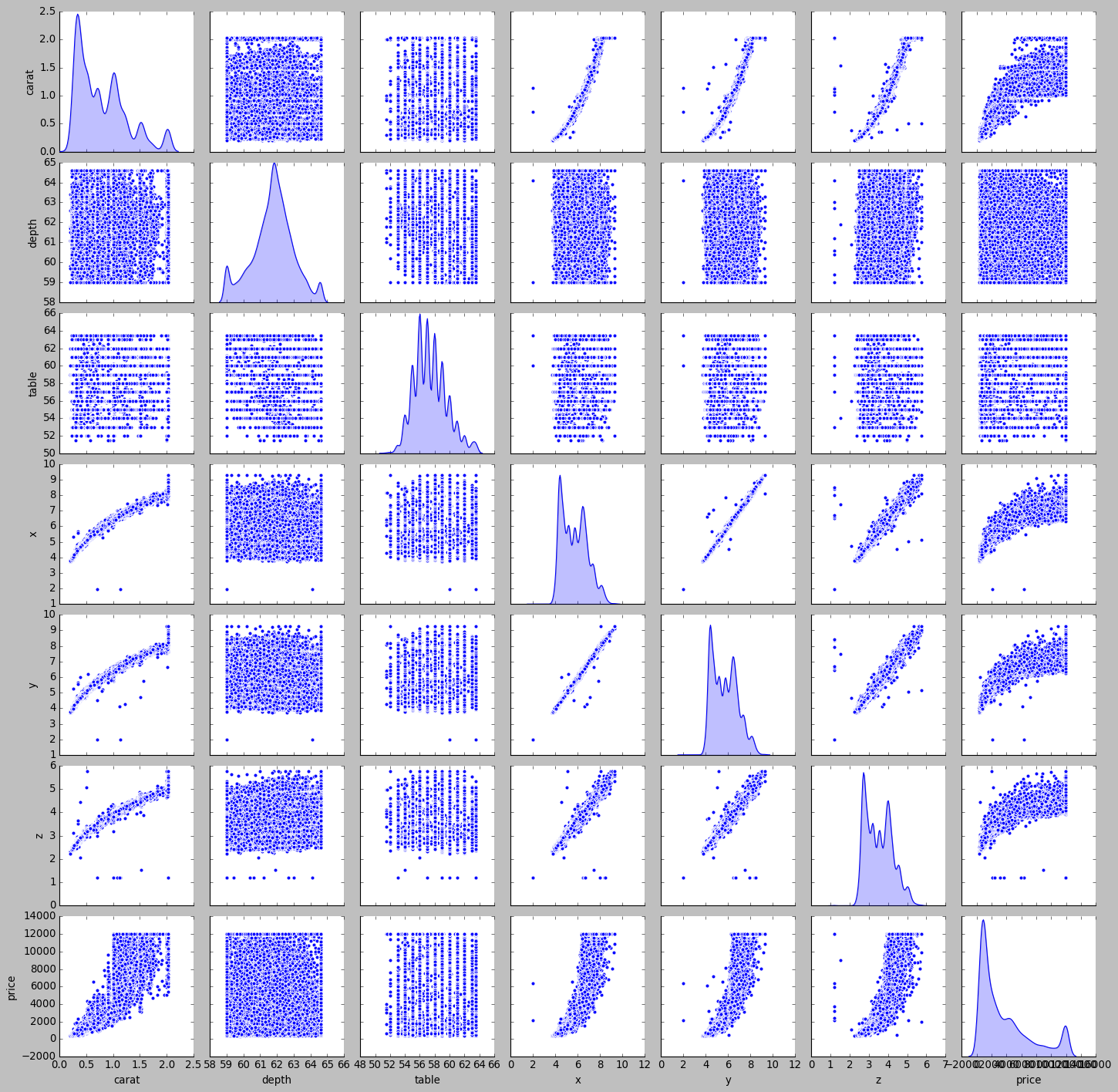
We observed that there are some outliers in the dataset. Upon checking them using box plots, we found that there are outliers in almost every attribute. The box plot is given below:



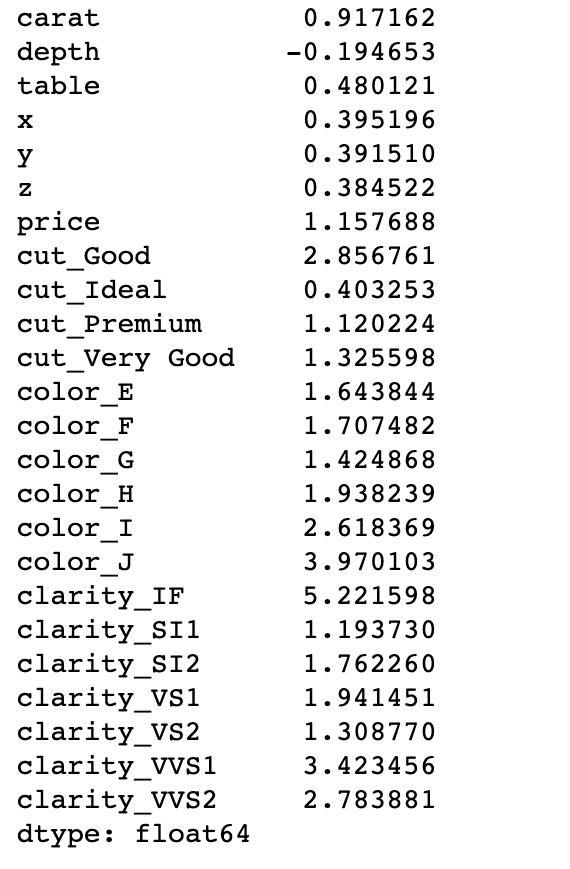
The box plot after removing outliers is as given below:



By checking the pairplot that’s given below, there seems to be some clusters present. We observed that there is some negative relationship between depth and height, width and length of cubic zirconia. carat, length, width, height and price are **positively skewed**.



After checking the skewness by using the skew() function, we can see that the depth attribute is negatively skewed. The screenshot of the result is shown below:



#### 

#### 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 697 null values present in the depth column. Let's replace them with median value.



From the pair plot and heatmap that’s shown in problem 1.1 above, we can see that the attribute depth is least correlated with all other attributes, we can remove it or we can replace the null values with median values. The null values are replaced with its median value.

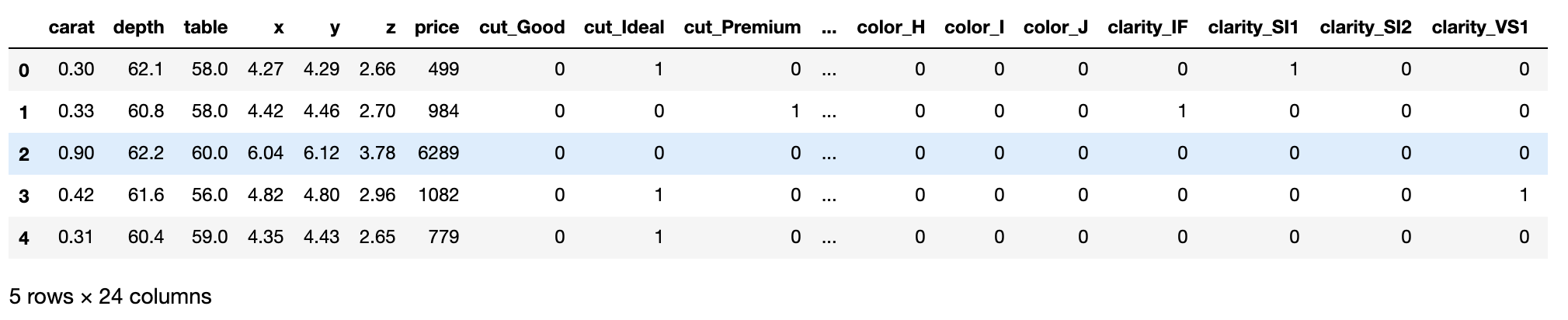
Scaling is necessary because by checking the summary, we can say that min, max and mean are not equal for all the columns. Data is scaled using preprocessing from sklearn library.

#### 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 R-Square, RMSE.

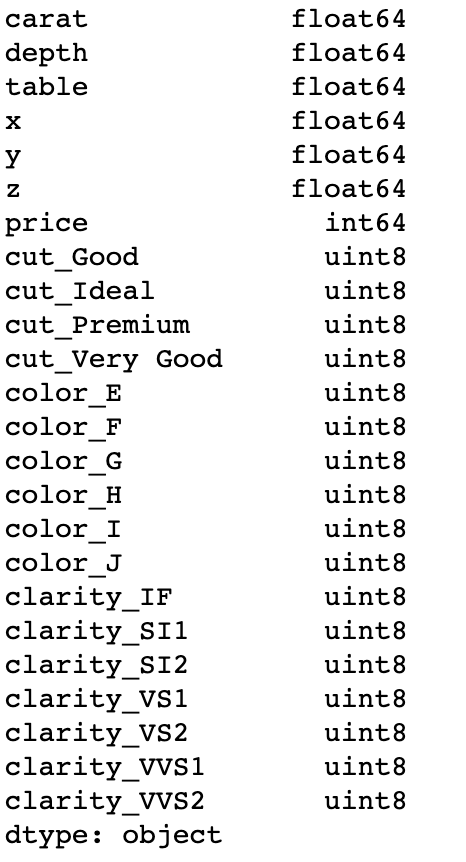
Let's change the object data types by encoding.

One-Hot-Encoding is used to create dummy variables to replace the categories in a categorical variable into features of each category and represent it using 1 or 0 based on the presence or absence of the categorical value in the record.

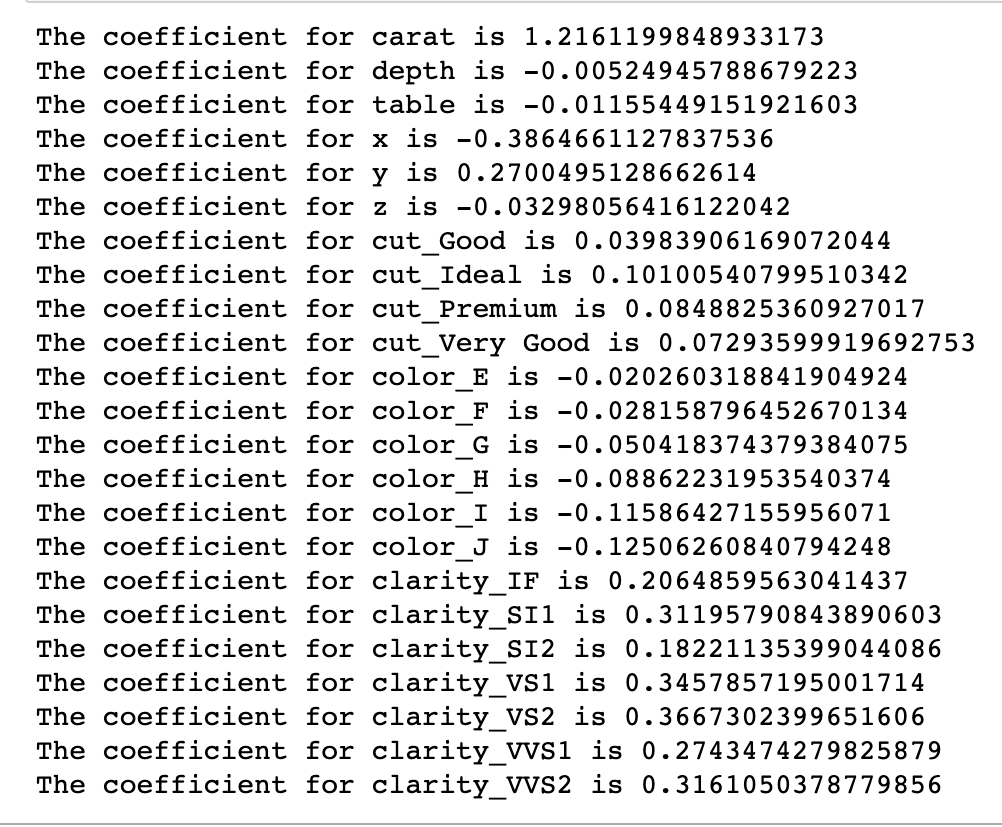
The new dataset is as shown below.



The datatypes of the new dataset are as shown below:



After splitting data, the coefficients for each of the independent attributes are

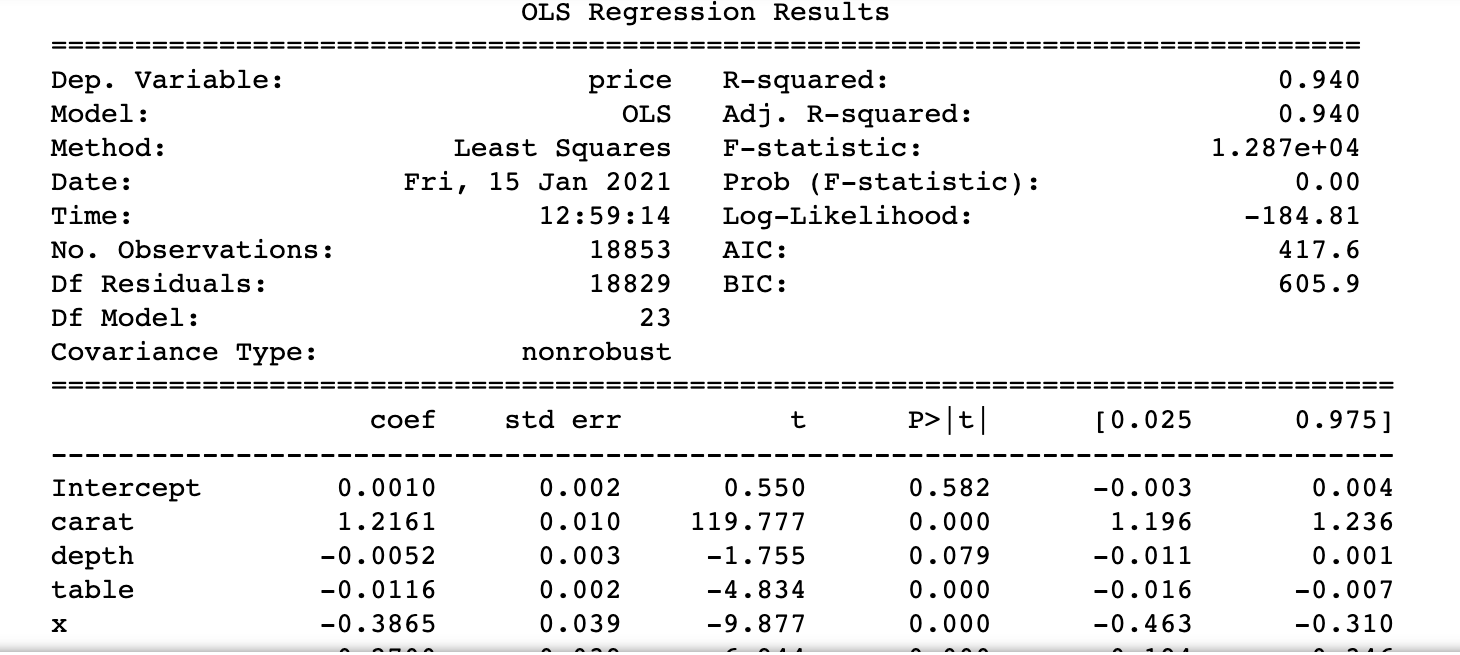


It looks like depth, table, length and width are negatively correlated. But, it shows positive relations. It seems to be that multicollinearity exists.

Let us check the intercept for the model.

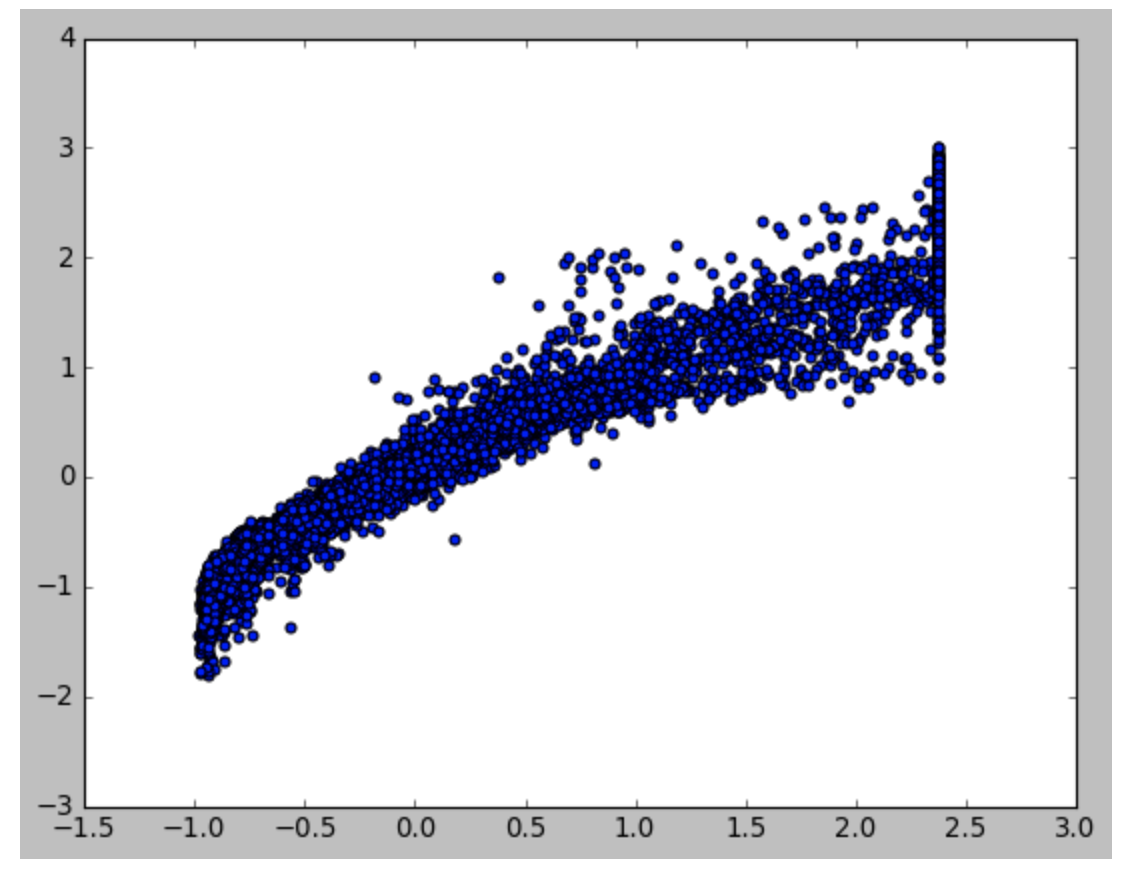
* The intercept for our model = 0.0009803459694918194
* R-Square on training data = 0.9402044588687953
* R-Square on testing data = 0.9419074345242372
* RMSE on training data = 0.24435440092961688
* RMSE on testing data = 0.24143024829860243

We got the same results by using stats model



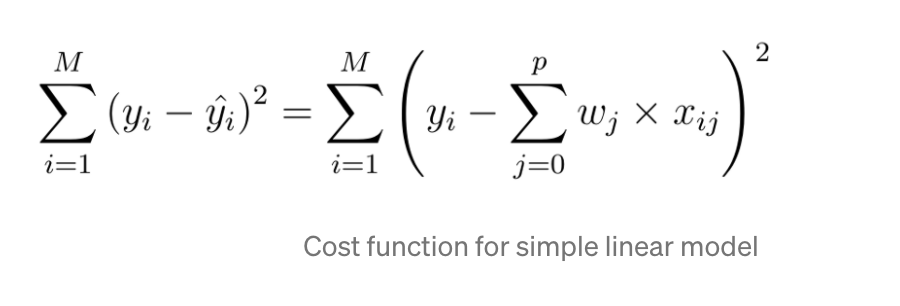
* R-squared = 0.940
* Adj. R-squared = 0.940
* Root Mean Squared Error - RMSE = 0.24451008200785526

Scatter plot for the predicted test data is as given below:

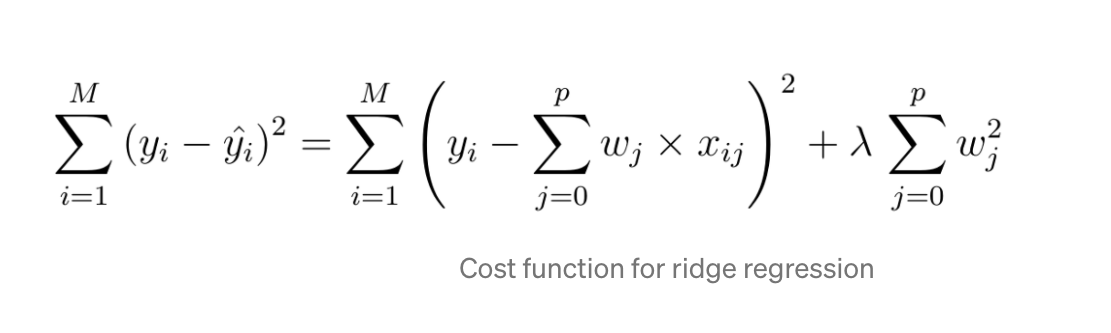


After scaling and applying linear regression, we got the same results as before.

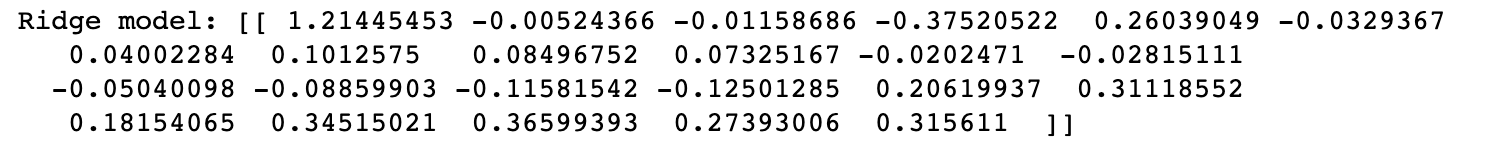
The model seems to be overfit. Let's apply ridge and lasso regression techniques which are some of the simple techniques to reduce model complexity and prevent overfitting that may result from simple linear regression.



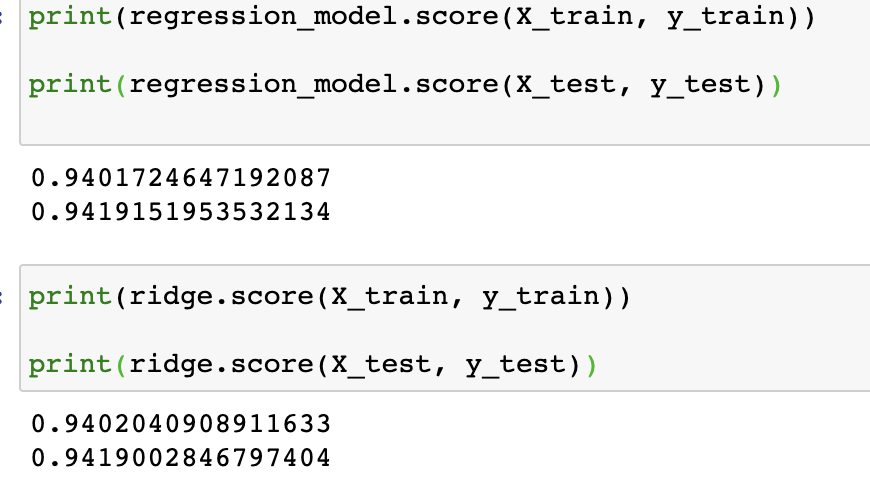
Ridge regression minimizes the cost function. It puts constraint on the coefficients*.* The penalty term regularizes the coefficients such that if the coefficients take large values the optimization function is penalized. So, ridge regression shrinks the coefficients and it helps to reduce the model complexity and multi-collinearity.

****

Higher the alpha value, more restriction on the coefficients; low alpha > more generalization

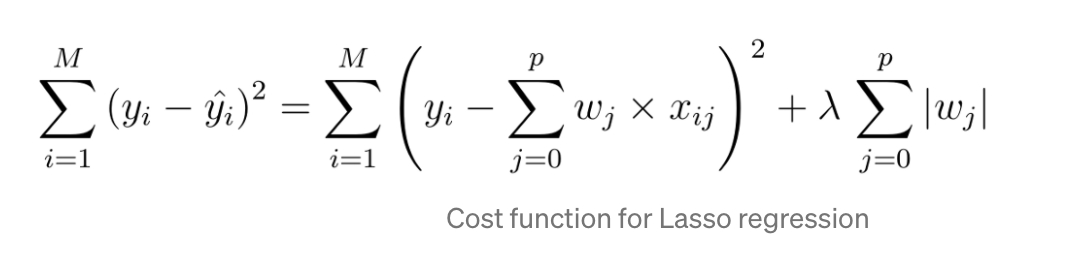
The coefficients for ridge model are ****

The train and test scores for Ridge and linear regression models are almost same

****

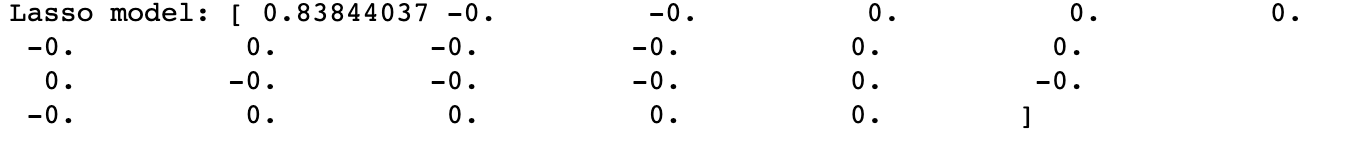
Lets apply lasso regression technique.

The cost function for Lasso (least absolute shrinkage and selection operator) regression can be written as

******

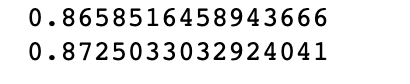
The only difference between ridge and lasso regression is instead of taking the square of the coefficients, magnitudes are taken into account. This type of regularization(L1) can lead to zero coefficients i.e. some of the features are completely neglected for the evaluation of output. Lasso regression not only helps in reducing overfitting but it can help us in feature selection.

The coefficients of lasso model are

****

We Observed that, many of the coefficients have become 0 indicating drop of those dimensions from the model

The train and test score for lasso model is

****

Further, the number of dimensions is much less in LASSO model than ridge or un-regularized model

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

* Linear regression models are overfit for this dataset.
* After regularising using lasso regression technique, we can say that carat, weight, length, width and height of the cubic zirconia are the most important attributes in deciding the good stones that give more profits.

Depth attribute is least important and also lowest profitable if decided based on depth.