Predictive Modelling Group Assignment

**Sharmila Sengupta Chowdhry**

#### Problem 1: Linear Regression Introduction

**PART 1**

* 1. Exploratory Data Analysis
  2. Univariate Analysis
  3. Bivariate Analysis
  4. Outliers Detection and Treatment
  5. Converting Categorical Variables into Numeric

#### PART 2

* 1. MODEL 0 (All Variables)
  2. MODEL 1 (Dropped x)
  3. MODEL 2 (Dropped y)
  4. MODEL 3 (Dropped Cut\_Ideal)
  5. MODEL 4 (Dropped z)
  6. MODEL 5 (Dropped clarity\_SI1)
  7. MODEL 6 (Dropped cut\_premium)
  8. MODEL 7 (Dropped cut\_Very\_Good)

#### PART 3

* 1. MODEL 8 (All Variables)
  2. MODEL 9 (Dropped x)
  3. MODEL 10 (Dropped y)
  4. MODEL 11 (Dropped Cut\_Ideal)
  5. MODEL 12 (Dropped z)
  6. MODEL 13 (Dropped clarity\_SI1)
  7. MODEL 14 (Dropped cut\_premium)
  8. MODEL 15 (Dropped cut\_Very\_Good) TRAIN-TEST - Model Summary

**Business Recommendation**

# Problem 1: Linear Regression

*You are hired by a company named Gem Stones Co Ltd, which is a cubic zirconia manufacturer. You are provided with the dataset containing the prices and other attributes of approximately 27,000 pieces of cubic zirconia (which is an inexpensive synthesized diamond alternative with similar qualities of a diamond).*

*Your objective is to accurately predict prices of the zircon pieces. Since the company profits at a different rate at different price levels, for revenue management, it is important that prices are predicted as accurately as possible. At the same time, it is important to understand which of the predictors are more important in determining the price*

# Introduction

Linear Regression techniques are widely used because of its simplicity and ability to explain the relationship of independent variables to the dependent variable. Linear regression is a linear model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that (y) can be calculated from a linear combination of the input variables (x).

To begin our analysis of the data, we would start data cleaning and Exploratory Data Analysis (EDA). We will start EDA by exploring the nature of all the variables, identify the response and the predictors, apply appropriate methods to determine whether there is any duplicate observation or missing data and how to treat them, and whether the variables have a symmetric or skewed distribution.

We will also conduct both univariate and bivariate analyses and pre-processing of data to explore relationships between the predictors as well as the predictor and target variables. For any regression problem (linear or logistic), the dependence of the response on the predictors will be thoroughly

# PART 1

***Objective:*** *The very first step of any data analysis assignment is to do the exploratory data analysis (EDA). Once you have understood the nature of all the variables, identified the response and the predictors, apply appropriate methods to determine whether there is any duplicate observation or missing data and whether the variables have symmetric or skewed distribution. Note that data may contain various types of attributes and numerical and/or visual data summarization techniques need to be appropriately decided. Both univariate and bivariate analyses and pre-processing of data are important. Check for outliers and comment on removing or keeping them while model building. Since this is a regression problem, the dependence of the response on the predictors needs to be thoroughly investigated.*

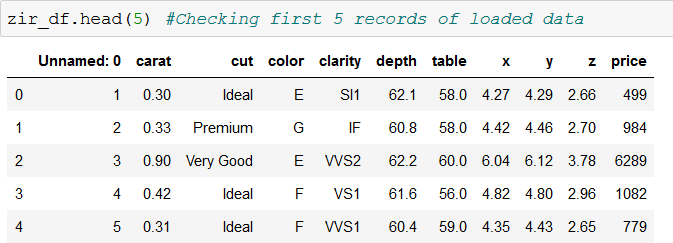
## Exploratory Data Analysis

We have imported all the necessary libraries. Then we read the dataset and conducted basic Exploratory Data Analysis. Dataset given for problem 1 is **“cubic\_zirconia.csv”**.

Price is our target (dependent variable) while rest of the variables are independent.

### Dataset loading into dataframe and checking records

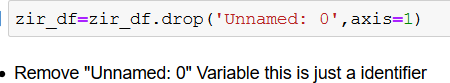
* Checked few records after loading dataset into Dataframe



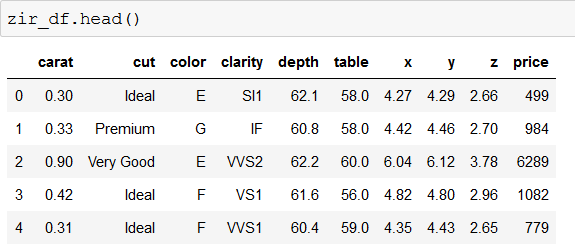
* We observed there is one unnamed column **“Unnamed: 0”**
* Dataframe has 26967 rows and 11 columns

### Dropping unnecessary column

* We are removing column **"Unnamed: 0"** as it is not required in model building.



* Checked few records after removing column **"Unnamed: 0"**



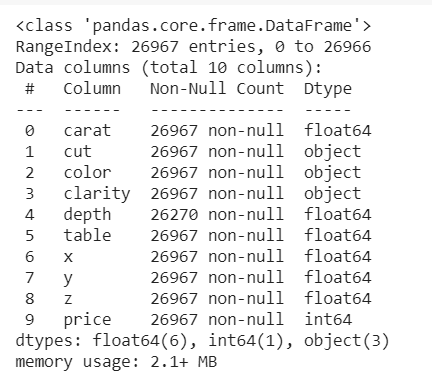
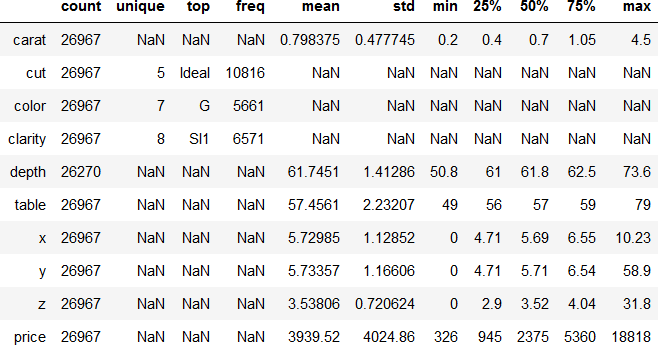
* Now we have **26967** rows and **10** columns.

### Summary of basic statistics

* In total, we have 26967 records across 10 columns in the dataset. The target / dependent / response variable is Price, which is an integer data type. The other columns are feature / independent / predictor variables
* Out of the 10 columns, 3 are object type, 6 are float and 1 is int.
  + Predictors of Object Data type: Cut, Type, Color, Clarity.
  + Predictors of Float Data Type: Carat, Depth, Table, x, y, z.
  + Predictors of Int Type: Price
* We observed **“depth”** variable is having 26270 count while rest of the variables 26967 which means **“depth”**

variable has missing values.

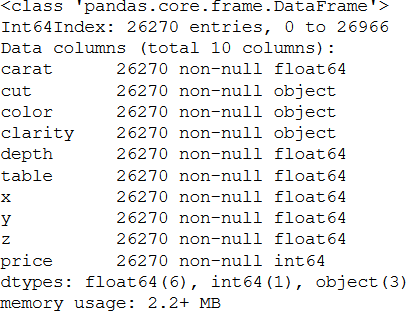
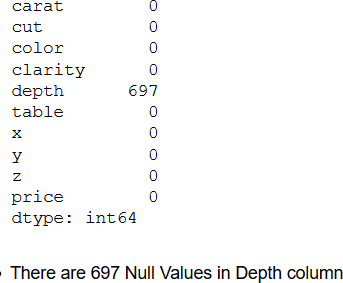
* There is no significant difference in mean and median of independent variables.
* Standard deviation is also minimum.



* Summary of basic statistical details of categorical variables.
* Variable **“cut”** has five unique levels. Out of these five levels **“Ideal”** is top one.
* Variable **“color”** has seven unique levels. Out of these seven levels **“G”** is top one.
* Variable **“clarity”** has eight unique levels. Out of these eight levels **“SI1”** is top one.

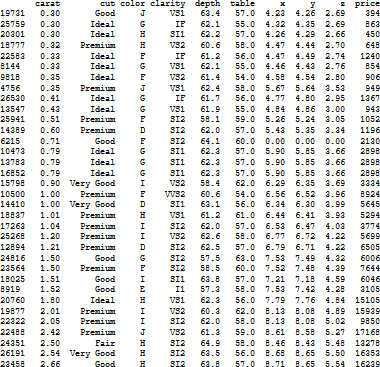
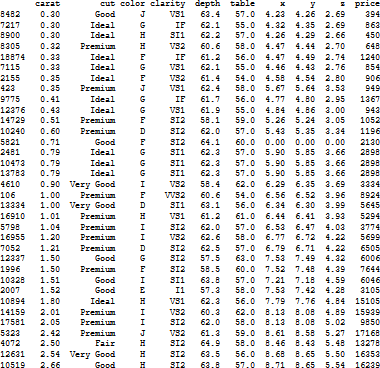
### Missing values

* + - * Missing value produce biased estimates that results in invalid conclusions, thereby compromising the predictability of the study.
      * After checking for null values, we found that there are 697 missing values in depth. We can directly remove those records we have still sufficient observation to learn the Model (Impute them with median or mean, as it will not affect the depth but other variables).
      * Dropped the records with “depth” as null and verified that the records are successfully dropped.

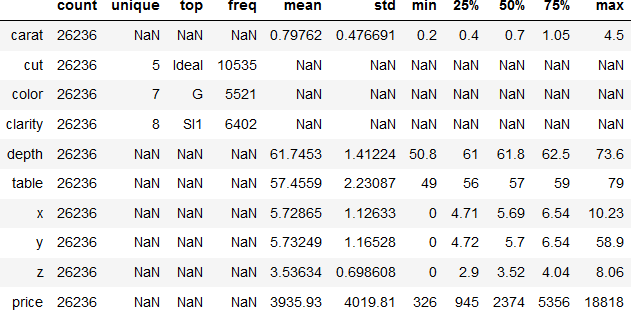
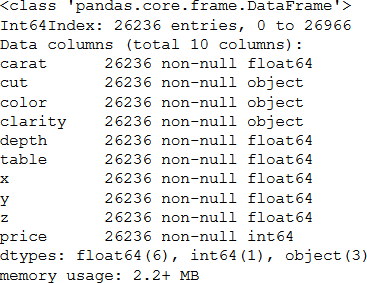


### 1.1.5 Checking duplicate rows in dataframe

* After checking the duplicates in the data we found that there are 34 duplicates in the data.



* From the above two snippets we can see that row number 8482 and 19731 are duplicates. In the same way, row number 7217 and 25759 and 8900 and 20301 are duplicates.
* Out of the total 26270 records, there are 34 duplicates which is approximately 0.0012% of the total data. It is a very small contribution therefore; we decided to drop the duplicates, hence losing 0.0012% of data from the data set.



As per above, we can see that after dropping duplicates from the dataset we now have a total of 26,236records. None of the continuous variables look to be normally distributed since their mean is not equal to median (50%). Although for variable depth, table, x, y & z the mean and median are very to each other.

## Univariate Analysis

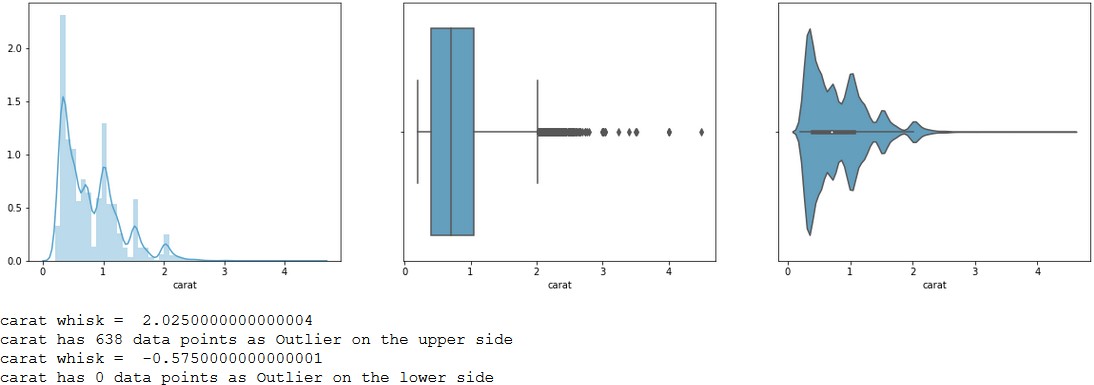
Univariate analysis refers to the analysis of a single variable. The main purpose of univariate analysis is to summarize and find patterns in the data variable

### Univariate Analysis of Continuous Variables

We start univariate analysis with a set of three graphs, a distribution/density plot, a boxplot and a violin plot, all to check the distribution of data in the feature, as well as the presence and extent of outliers. Like missing data, outliers

substantially affect the model, and thus must be dealt with carefully. We elaborate on this in an upcoming section. We begin with analyzing the continuous variables in the listed order: Carat, Depth, Table, Price, x, y and z

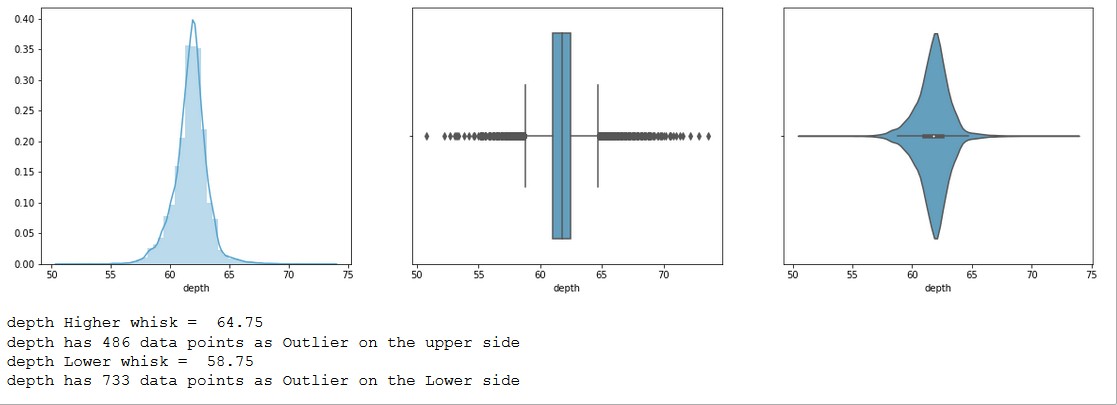
### Univariate Analysis of Carat



The univariate analysis of Carat reveals that most of the observations are in between 0 and around 2. There are more outliers on the upper side of distribution, which explains the right skew in the violin plot. We do not know whether these are valid outliers or invalid ones. The treatment of outliers is subjective. There are multiple fluctuations in the data.

Most of the data points in the carat are between 0 and 2. There are some data points that are between 2 to 5, which is responsible for the skew.

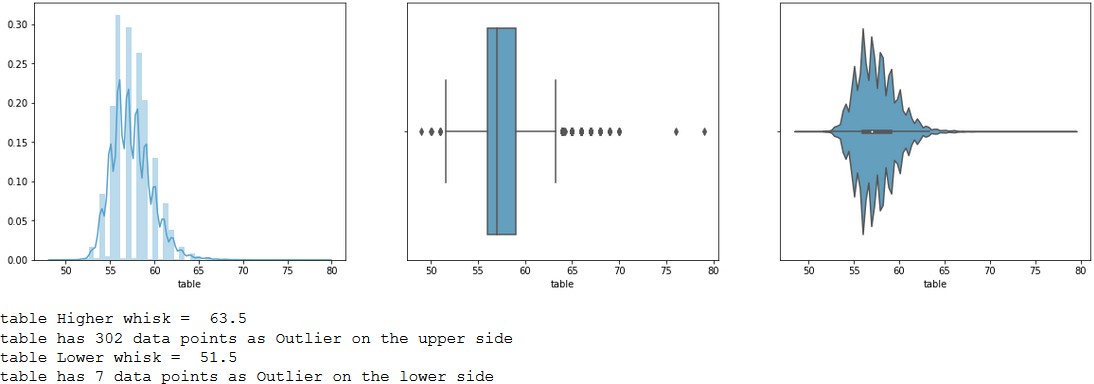
### Univariate Analysis of Depth



The univariate analysis of Depth reveals that most of the observations are in between 60 and around 65. There are more outliers on the both side of distribution. It seems from the above graph that the data distribution is almost normally distributed.

Most of the data points in the depth are between 58 and 65.

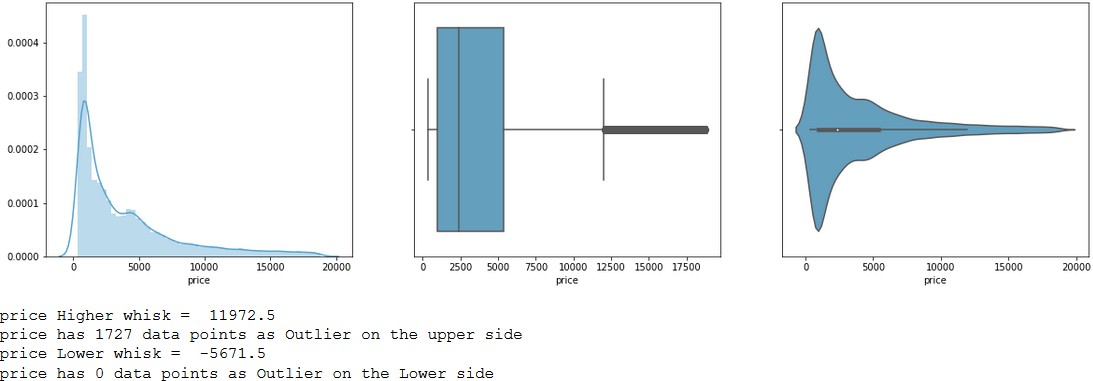
### Univariate Analysis of Table



The univariate analysis of Table reveals that most of the observations are in between 53 and around 63 and the distribution is right skewed. It seems Outliers are on the both sides of the whisker but more outliers on the higher side.

Most of the data points in the table are between 51 and 63. There are some data points that are between 48 to 50 & 63 to 79, which is responsible for the skew.

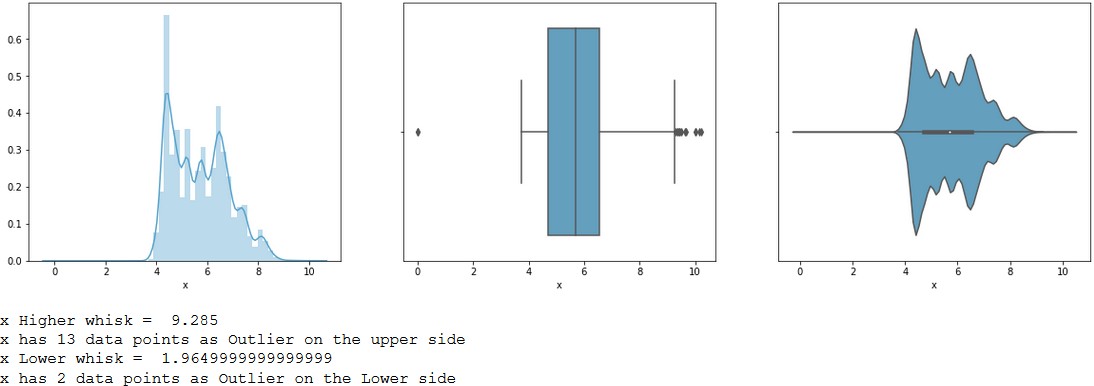
### Univariate Analysis of Price



From the violin plot it can be seen that, most of the observations are in between 0 and 10k and the distribution is right skewed .There are outliers on the upper side of distribution, which explains the right skew in the violin plot.

From the boxplot, most of the data points in Price are between 0 and 12k. There are some data points, which are between 12k & 18k, which are responsible for the skew.

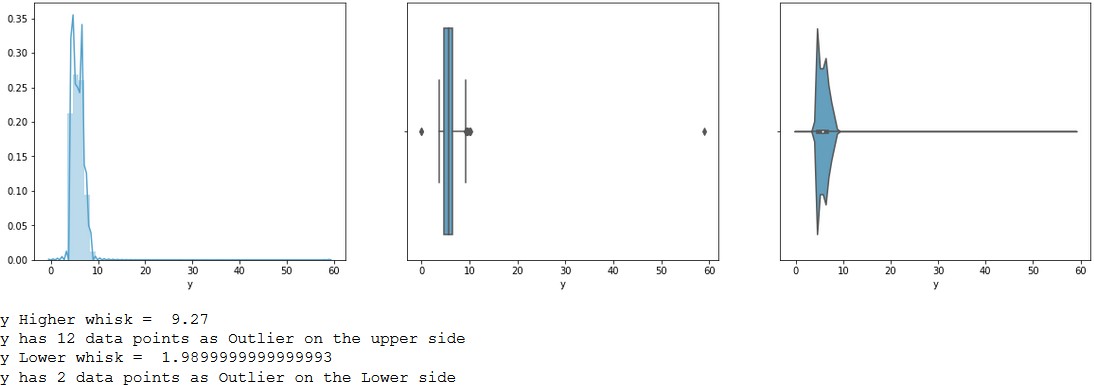
### Univariate Analysis of X



From the violin plot it can be seen that, most of the observations are in between 4 and 8 and the distribution is right skewed .There are outliers on the upper side of distribution. Lot of fluctuation in the data can be seen.

One extreme outlier is on the left side of the whisker. Seems like an invalid outlier, as x cannot be zero.

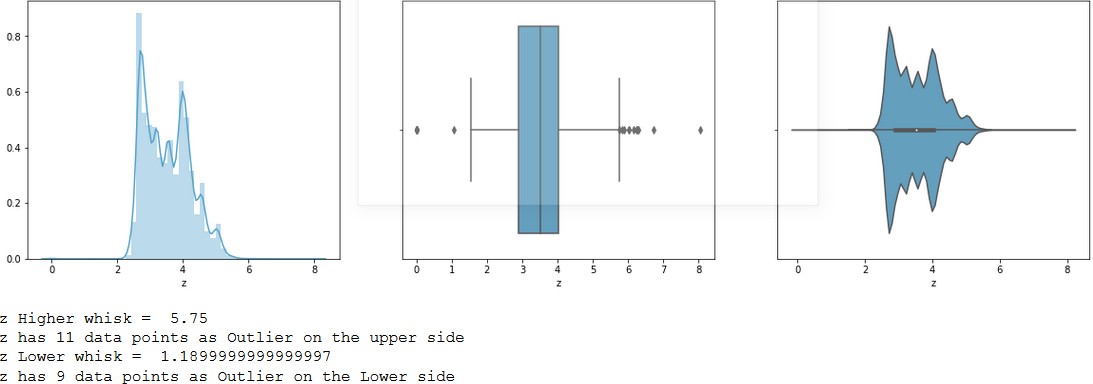
### Univariate Analysis of Y



 From the above plot, it can be seen that the range for this variable is very low and the distribution is right skewed .There are outliers on the upper side of distribution. It seems Outliers are on the both sides of the whisker but more outliers are on the higher side. One extreme outlier is on the right side of the whisker.

From the boxplot, it seems one extreme outlier is on the left side of the whisker. Seems like an invalid outlier, as y cannot be zero..

### Univariate Analysis of Z



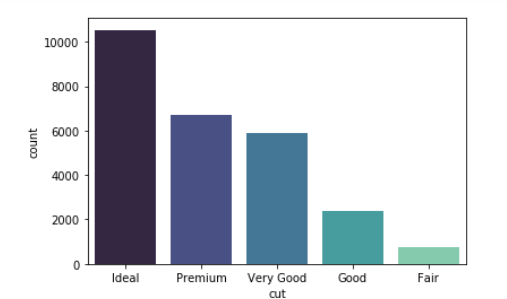
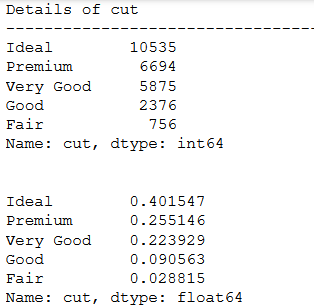
 From the above plot, it can be seen that the range for this variable is very low and the distribution is right skewed .There are outliers on the upper side of distribution. It seems Outliers are on the both sides of the whisker but more outliers are on the higher side.

One extreme outlier is on the right side of the whisker. From the boxplot, it seems one extreme outlier is on the left side of the whisker. Seems like an invalid outlier, as y cannot be zero.

### Univariate Analysis of Categorical Variables

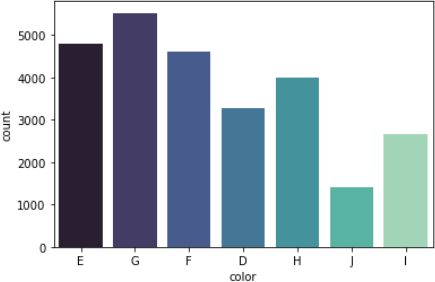
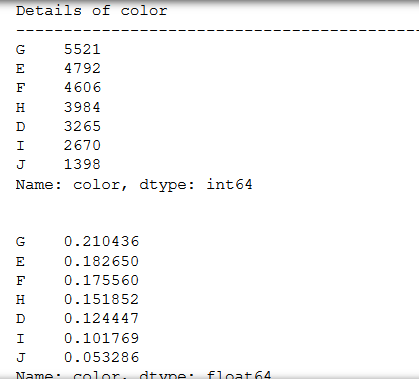
In Python, categorical variables are analysed through a function called countplot, pulled from the seaborn library We analysed the frequency of class distribution in each of the categorical features, Cut, Color Clarity

### Univariate Analysis of Cut



Cut constitutes of 40% (10535) Ideal, 26% (6694) Premium, 22% (5875) for Very Good, 9% (2376) for Good and 2% (756) High Injury. 87% of the data is contributed by Idea, Premium and Very good diamond cut quality

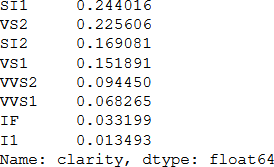
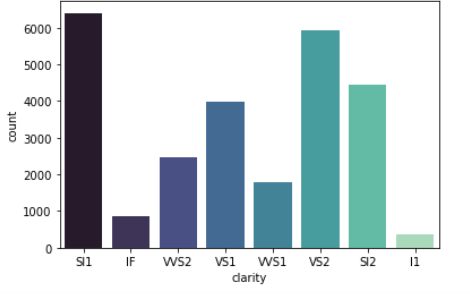
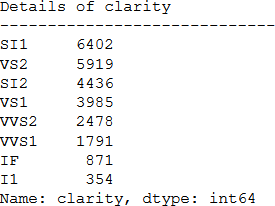
### Univariate Analysis of Color



Color constitutes of 21% (5521) color G, 18% (4792) color E, 18% (4606) color F, 15% (3984) color H, 12% (3265) color D, 10% (2670) color F and 5% (1398) for Color J. It seems color J has the lowest share amongst other colors. It seems that it belongs to the higher price of the section.

### Univariate Analysis of Clarity

Clarity constitutes of 24% (6402) SI1, 22.5% (5919) VS2, 17% (4436) SI2, 15% (3985) VS1, 9% (2478) VVS2,

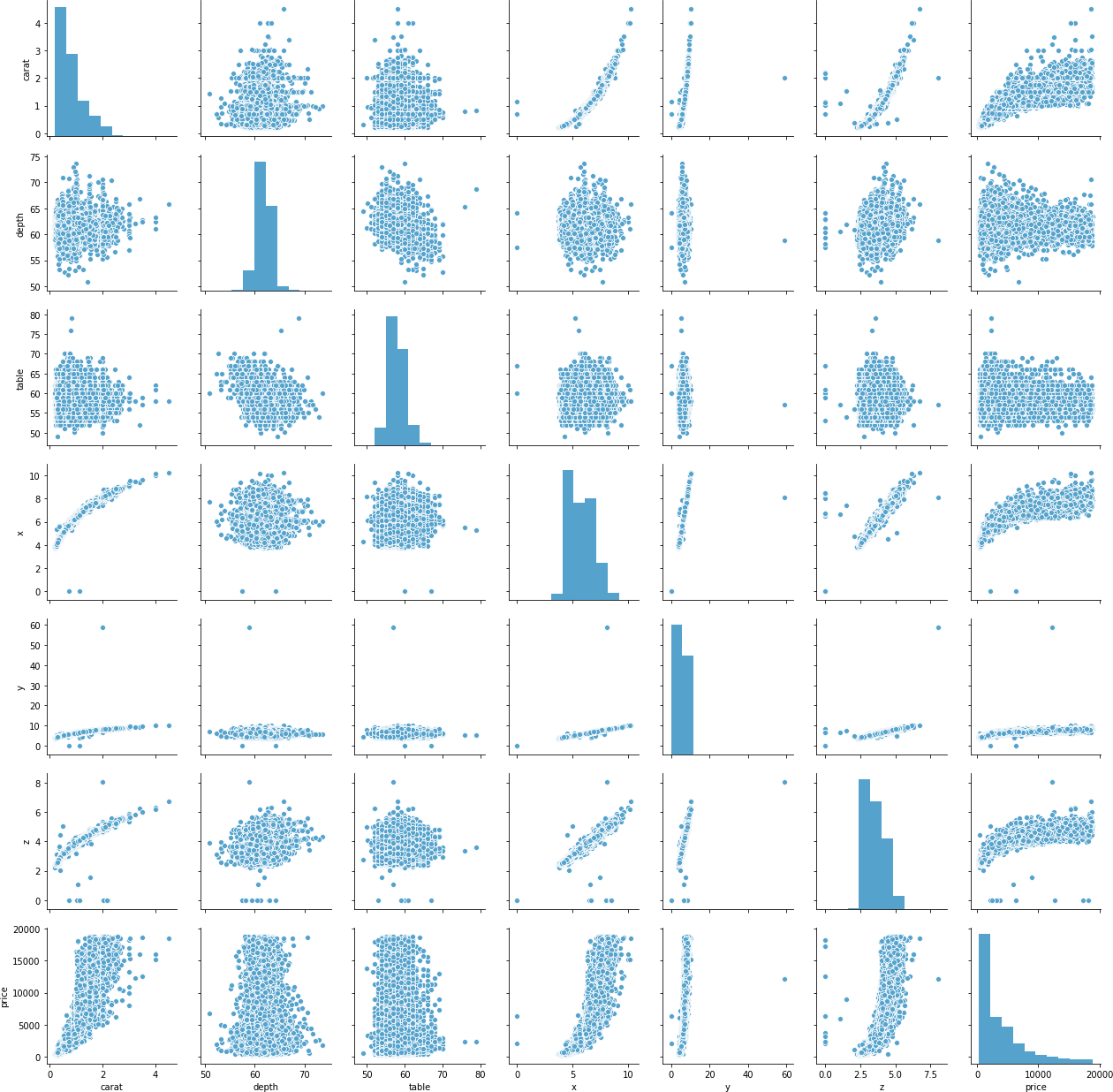


6% (1791) VVS1, 3% (871) IF and 1% (354) L1. There are no manufacturing of Flawless and Worst clarity diamond present in the dataset. The manufacturer is manufacturing more no of diamonds with SL1 and VS2 clarity

## Bivariate Analysis

### Bivariate Analysis of Continuous Variables

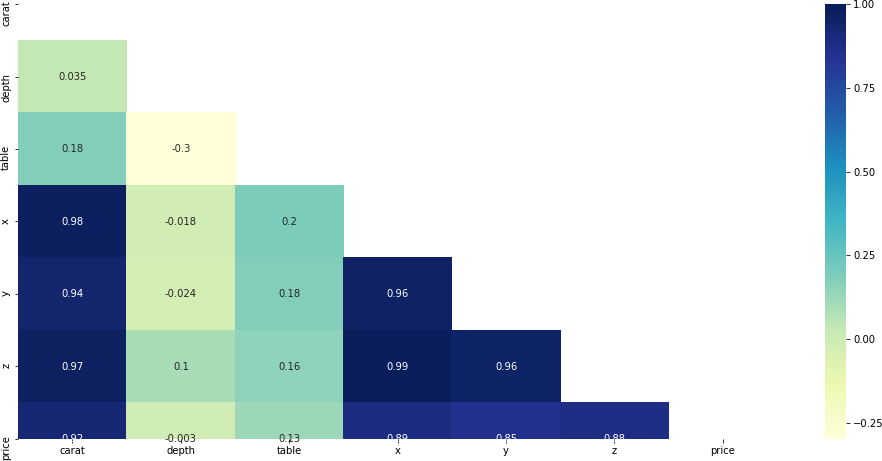
* + - 1. **Scatter Plot**



Observation from the pairplot:

* + - * + There a very high correlation between Price and x, y and z. There seems an exponential relationship. Slight increase in x,y or z will have huge impact on the price:
        + There seems a positive correlation between Price and Carat.
        + x, y and z all seems to have a positive correlation with carat.

### Heatmap

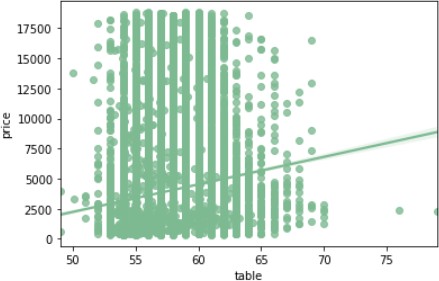
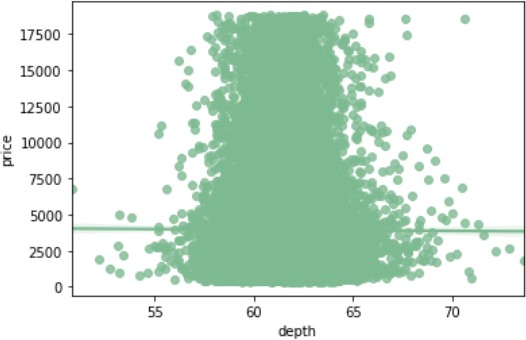
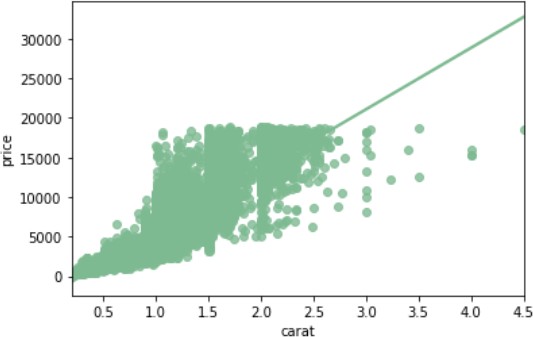


Observation from the heatmap:

* + - * + From the heatmap we can see that Price has a high positive correlation with Carat, x, y and z.
        + x, y and z has high positive correlation among themselves

### Regplot

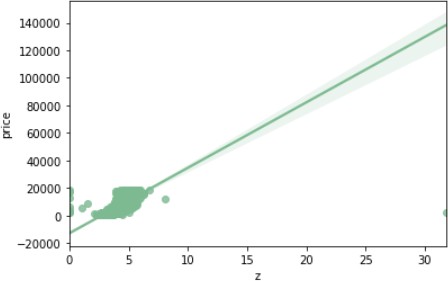
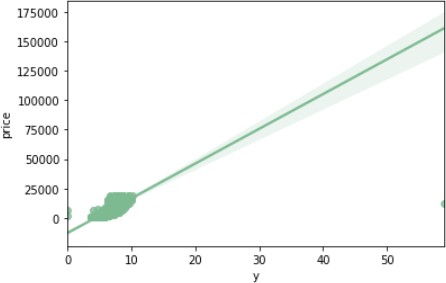
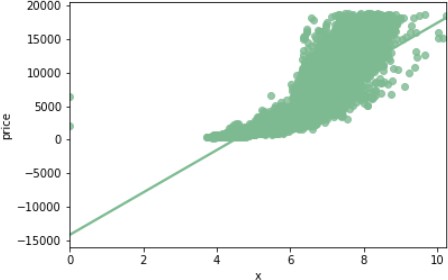
**Price & Carat Price & Depth Price & Table**



Observation from the Regplot:

* + - * + Price has positive correlation with carat
        + Few data points of higher carat rating are deviating from the regression line
        + At any value Depth has no or very low correlation with price. Therefore, the regression line is a straight.
        + From the Heatmap shown above a negative corr of -0.003 can be seen which is close to zero and that explains an almost straight line.
        + Most of the data points of Depth are between price level 55 to 70.

### Price & X Price & Y Price & Z

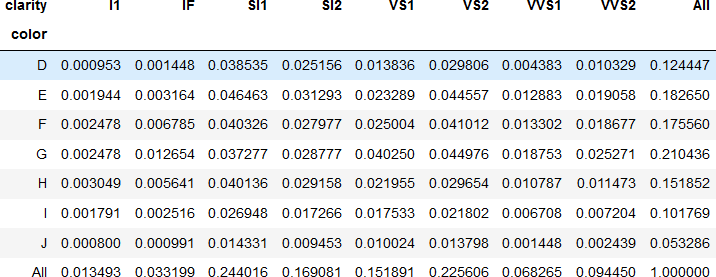
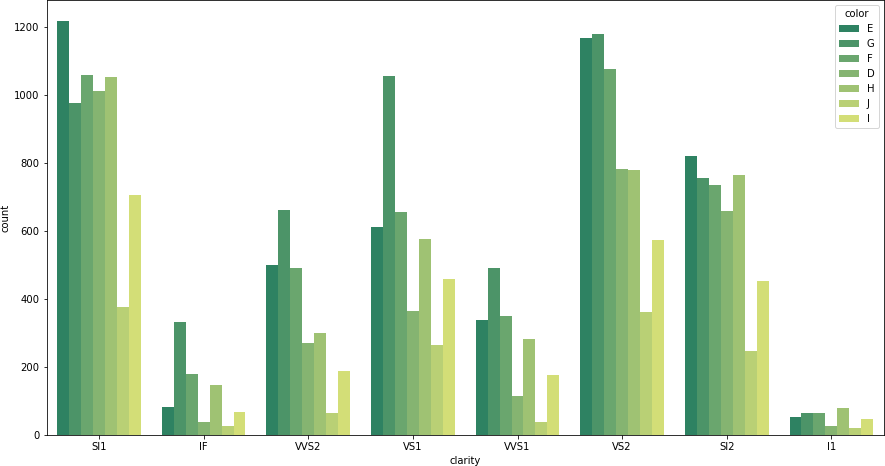


Observation from the Regplot:

* + - * + As per the reg plot there is a strong positive correlation between Price and x,y and z. We can confirm this from the heatmap as well
        + x(Length) and z has more spread in the data as compared to y.
        + Due to presence of extreme outlier in y( width) the spread of y is more than what appears.

### Bivariate Analysis of Categorical Variables

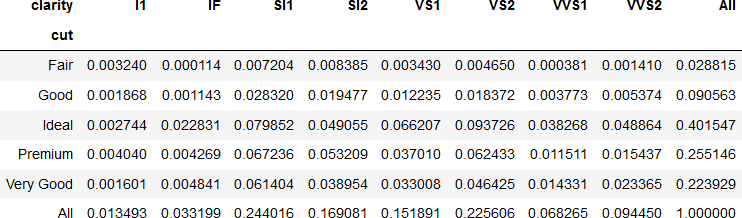
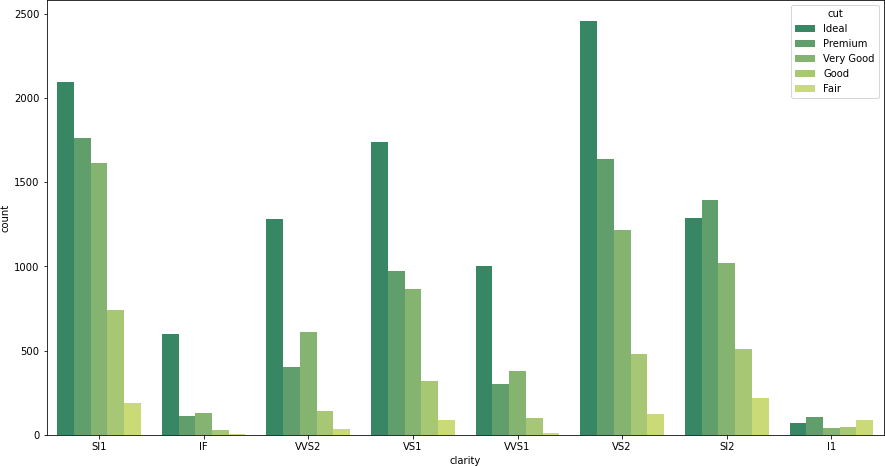
* + - 1. **Clarity & Color**



Observation:

* + - * + SL1 and VS2 are in majority of numbers
        + Color G, E and F seems to have more contribution than the others do.
        + Color D, J and I have the least contribution

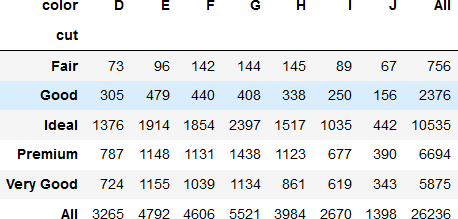
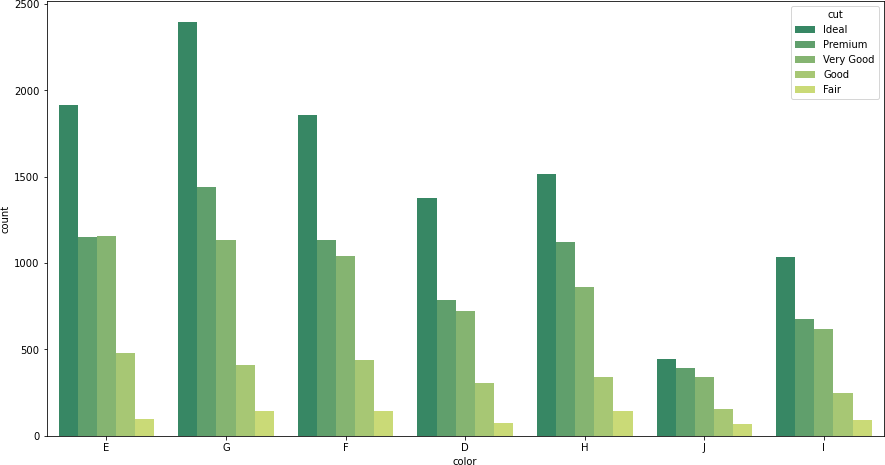
### Clarity & Cut



Observation:

* + - * + From the bar graph, we can see that VVS1 and VVS2 Clarity has more Ideal cut diamonds. As per the data, these are the best three qualities of diamond.
        + I1 being the worst clarity, shows that the contribution of all the cuts are almost similar. Although Premium still has the highest contribution in I1.
        + Most of the dataset in clarity belongs to Ideal cut category (40%), followed by Premium and Very good cut quality
        + SI1 and VS2 seems to have a major contribution of 24% and 22%, respectively.
        + From the bar plot it can be seen that in each Clarity Ideal cut is in majority. Except for I1 & SI2

### Color & Cut

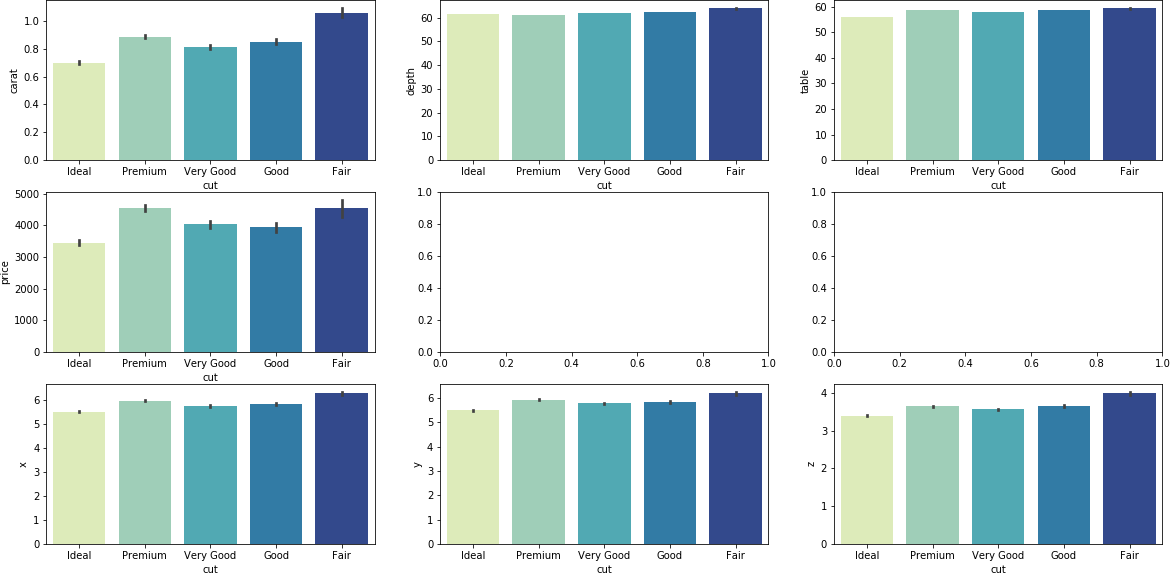


Observation:

* + - * + From the barplot, it can be seen that in each Clarity Ideal cut is in majority.
        + Color G being the highest contributor of ideal cut data points followed by E and F.

### Bivariate Analysis of Categorical & Continuous Variables

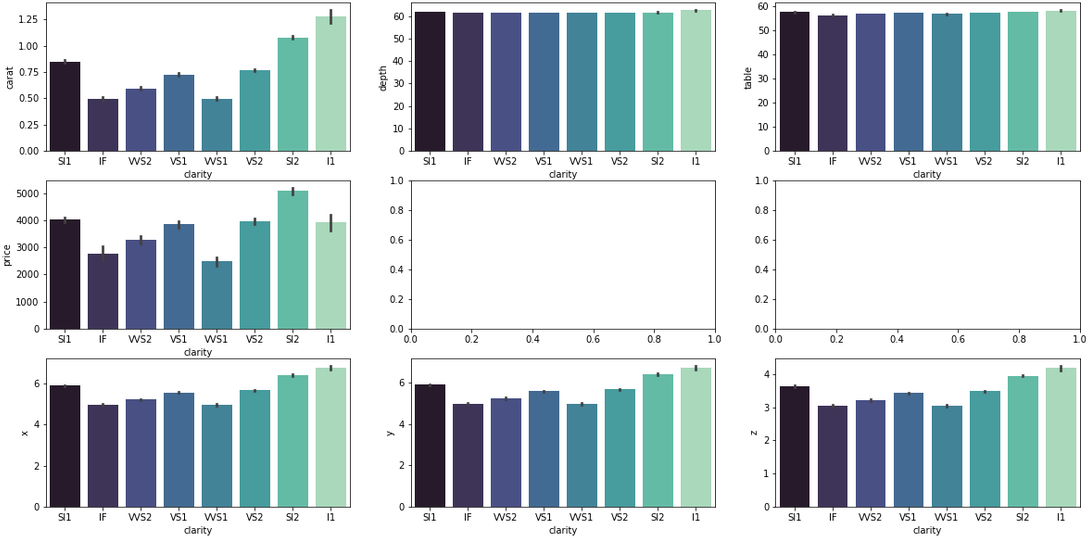
* + - 1. **Only Cut**



Observation:

* + - * + Fair being the lowest quality diamond shown highest average carat rating as more compared to others. Also, Ideal cut being the best quality cut diamond has less carat average
        + Depth and Table showing no significant difference between the cut categories
        + Premium cut and Fair cut seems to have higher price than the others. Ideal cut has the lowest price as compared to the others.
        + x, y and z has the similar behavior among cut

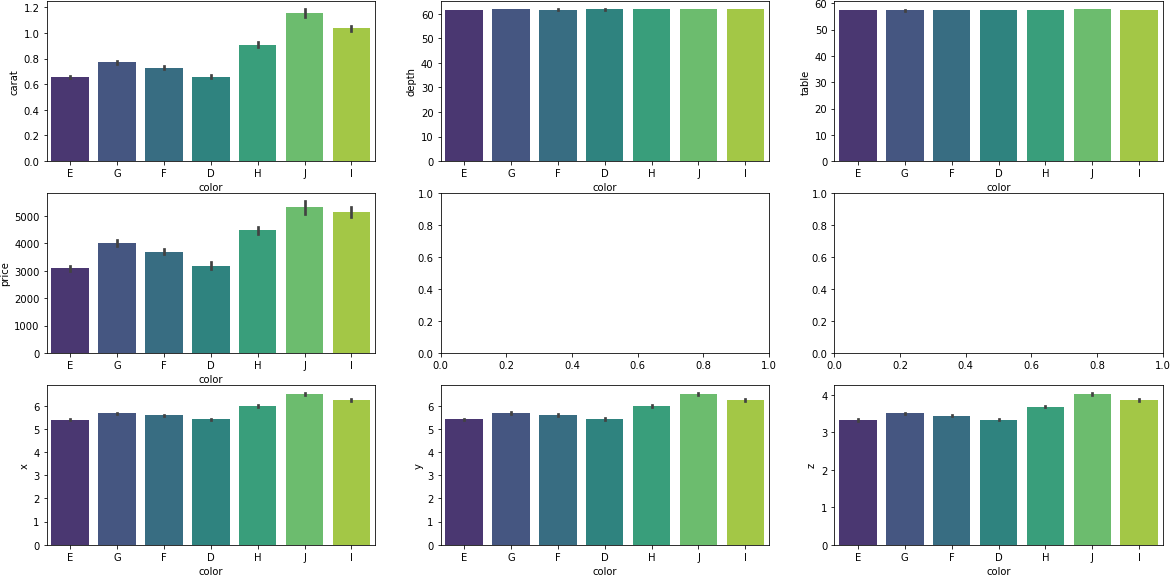
### Only Clarity



Observation:

* + - * + I1 being the worst Clarity has max carat weights meaning it contains more number of blemishes.
        + SI2 Clarity has the highest price. Whereas IF, VVS1 and VVS2 being the top3 clarity has lower price than the other clarity diamonds.
        + I1 clarity diamonds seems to be bigger in size compared to others as there x, y and z values are more than others.

### Only Color



Observation:

* + - * + From the bar graph it seems that -Color J has high carat weight followed by color I. while color E and D has the lower caret weights compared to others.
        + Color J and Color I has the highest price diamonds.
        + Diamonds belonging to Color E and Color D has the lowest Price.
        + Diamonds belonging to Color J and I seems to be bigger in size than other colors. Rest are almost similar in size

## Outliers Detection and Treatment

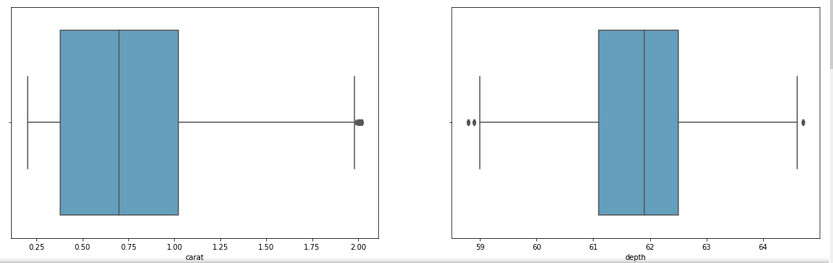
We’ll fetch the number of outliers from every Variable from upper Range (Q3 + 1.5IQR) and Lower Range (Q1 - 1.5IQR)

#### Observations on Outliers Detected:

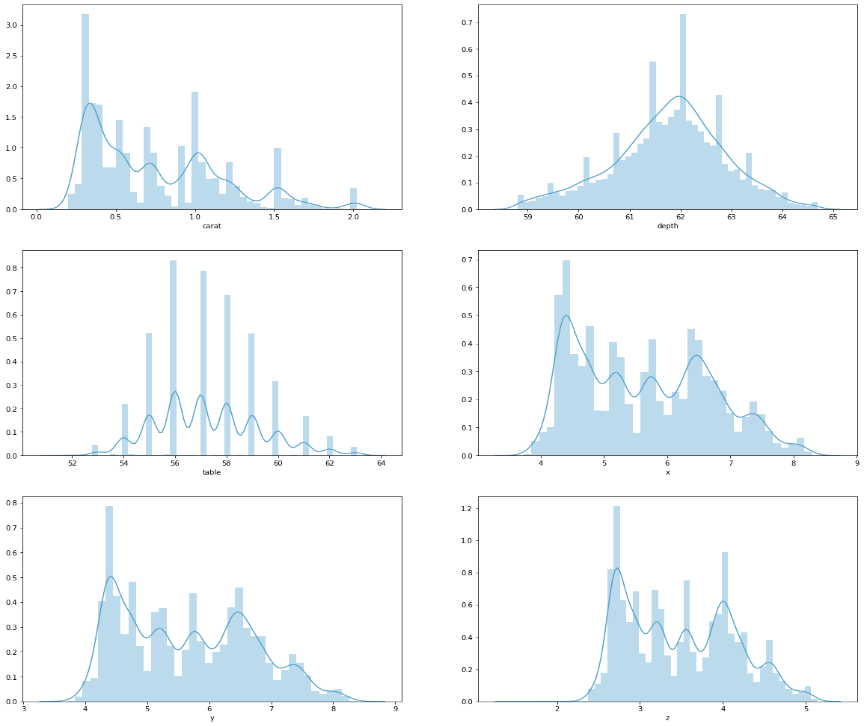
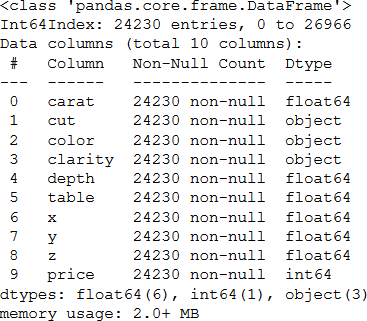
* For **Carat** higher whisker is at 2.025 and Lower -0.575. It has 638 values on the higher side of the whisker and 0 values on the lower side of the whisker.
* For **Depth** higher whisker is at 64.75 and Lower 58.75. It has 486 values on the higher side of the whisker and 733 values on the lower side of the whisker.
* For **Table** higher whisker is at 63.5 and Lower 51.5. It has 302 values on the higher side of the whisker and 7 values on the lower side of the whisker.
* For **x** higher whisker is at 9.285 and Lower 1.96. It has 13 values on the higher side of the whisker and 2 values on the lower side of the whisker.
* For **y** higher whisker is at 9.27 and Lower -1.98. It has 12 values on the higher side of the whisker and 2 values on the lower side of the whisker.
* For **z** higher whisker is at 5.75 and Lower 1.19. It has 11 values on the higher side of the whisker and 9 values on the lower side of the whisker.

#### Outlier Treatment:

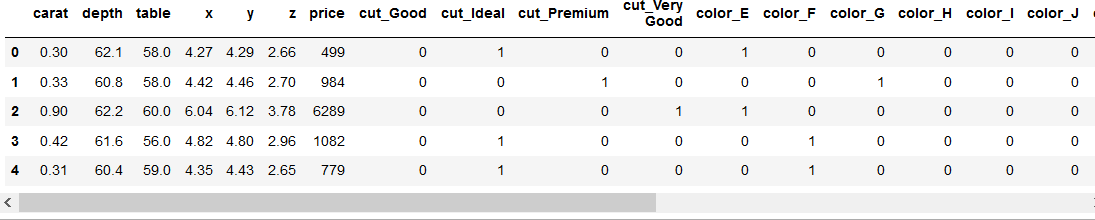
* As we have many outliers. Let us remove the records, which has value higher than upper bound and less than lower bound, and check the number of observations.
* If only few hundred records are dropped then we would have still sufficient records for the model to learn.
* After removing outliers, we have 24,230 records, so overall we have removed 2K records during Missing value, Duplicate and Outlier Treatment
* However, despite the outlier treatment, we still have outliers in the data



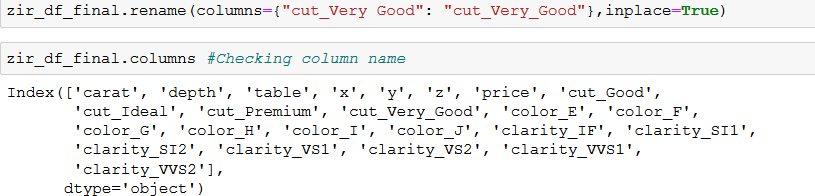
**Data Info and Histogram after Outlier Treatment**:



## Converting Categorical Variables into Numeric



Renaming columns 'cut\_Very Good' to ''cut\_Very\_Good' as there should not be any space in variable name



# PART 2

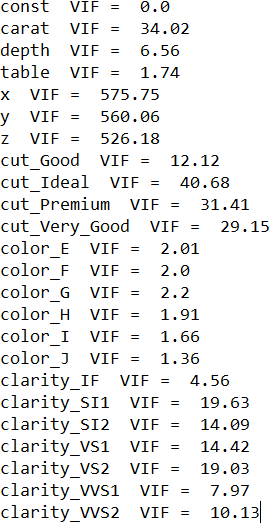
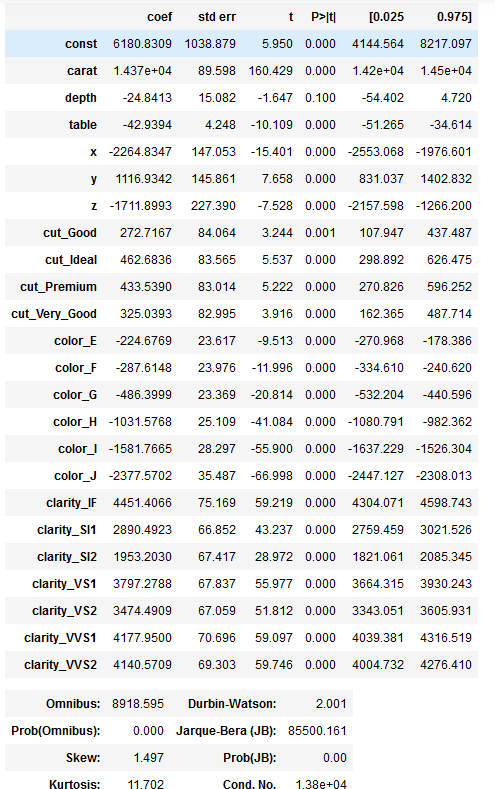
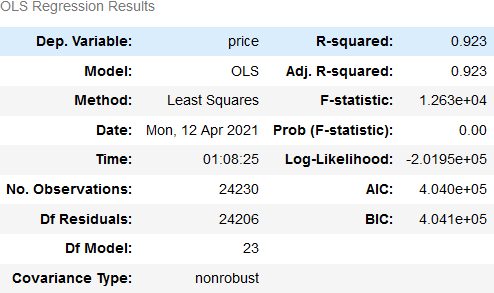
**Objective:** *Use Full Data to develop a model to identify significant predictors. Check whether the proposed model is free of multicollinearity. Apply variable selection method as required. Show all intermediate models leading to the final model. Justify your choice of the final model. Which are the significant predictors?*

Linear regression is the appropriate regression (prediction) analysis to conduct when the dependent variable is Continuous. As our dependent variable i.e., Price, is a continuous variable so we need not need to perform encoding to it. However, for other independent variables in the dataset we have used Encoding technique to convert all the categorical variables into numeric values.

Points on modelling process:

* We are using full data in Model building.
* Missing values are already treated and by following dropping the null values. And there was no significant change in the statistics
* We have treated the outliers and dropped the records having outliers. We have sufficient number of records for the model building

## MODEL 0 (All Variables)



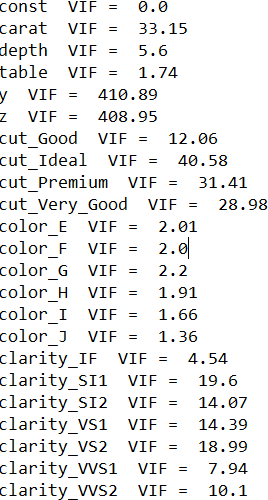
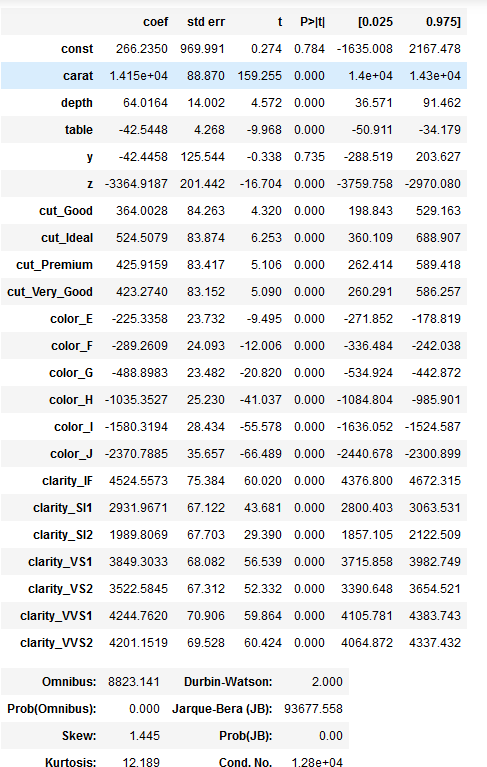
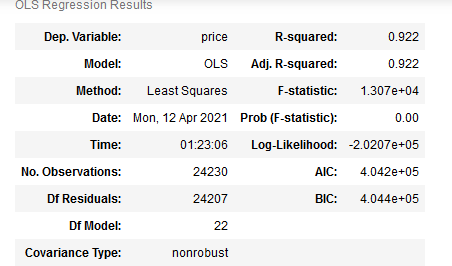
Observation:

* + - Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. In general, VIF greater than 5 have high/severe multicollinearity between variables.

Since, multiple variable are having VIF value greater than 5. Therefore, first we will need to remove them, as it will cause our model to have multicollinearity.

* + - P-Values help in determining if the variable is significant or not. We will be looking at VIF and P-value to get rid of the Independent variables, which are causing multicollinearity.
    - Looking that VIF values we can interpret that we will have to drop multiple variables in order to get to the final model.
    - Here x has highest VIF and this is because of high correlation between x, y, z and Carat. Let’s drop x and re-run the model
    - From the Model output, Durbin Watson (D) is coming out as 2, which means the there is no auto correlation among the residuals.

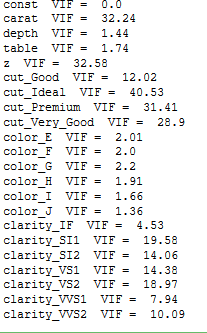
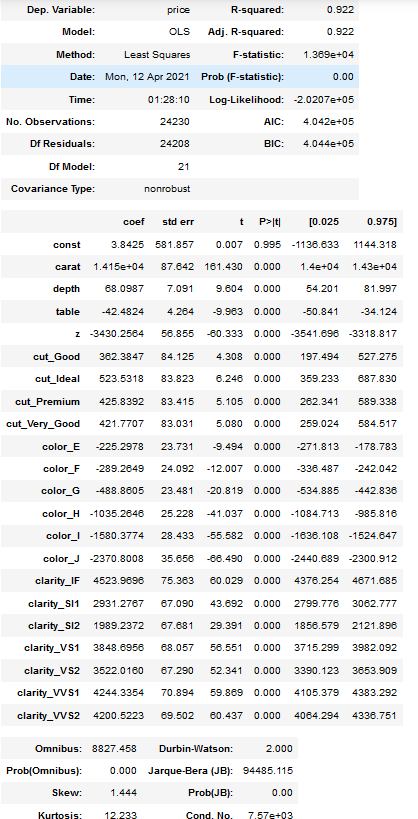
## MODEL 1 (Dropped x)



Observation:

* + - After dropping x variable, still we can see that there is very high VIF values for variable y and z.
    - In addition, other variables have high VIF values.
    - Let us build model by dropping variable y, having the highest VIF value of 410.89.

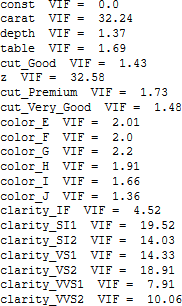
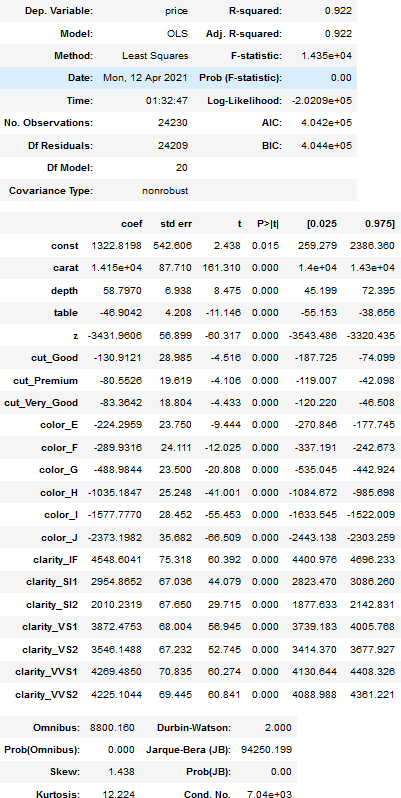
## MODEL 2 (Dropped y)



Observation:

* + - From above VIF values we can interpret that after removing y variable, the VIF value of z has got down from 408.95 to 32.58
    - In this iteration we found that Cut\_Ideal is having the highest VIF value of 40.53. We will be building model by dropping this variable and see what effect it has on the model

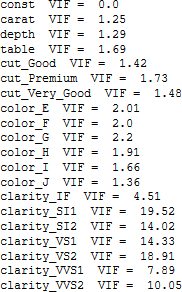
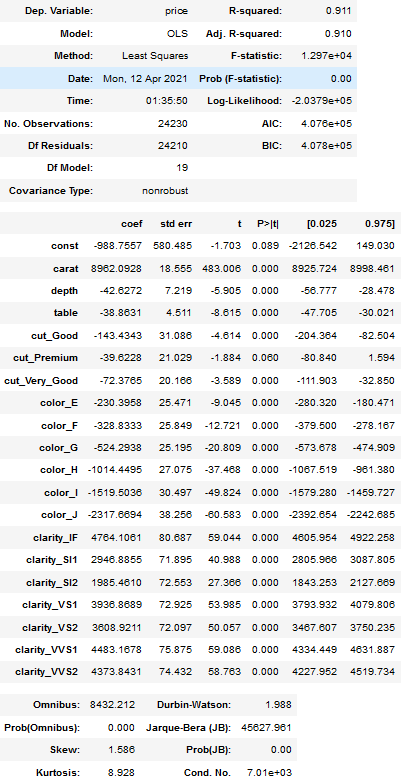
## MODEL 3 (Dropped Cut\_Ideal)



Observation:

* + - From above VIF values we can interpret that after removing cut\_Ideal variable, the VIF value of cut\_Premium, cut\_Very\_Good and cut\_Good has got down to the acceptance range
    - In this iteration we found that variable z is having the highest VIF value of 32.58. We will be building model by dropping this variable and see what effect it has on the model

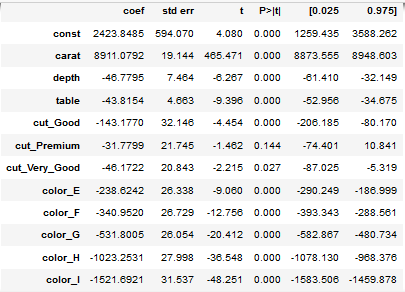
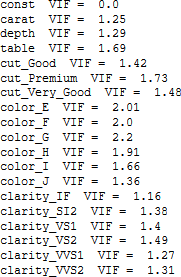
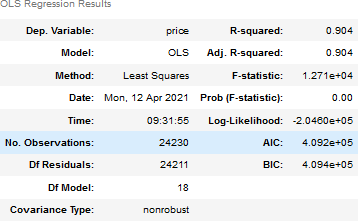
## MODEL 4 (Dropped z)



Observation:

* + - From above VIF values we can interpret that after removing variable z, the VIF value of Carat has got down to the acceptance range. Although we see high VIF value in Clarity columns.
    - In this iteration we found that variable clarity\_SI1 is having the highest VIF value of 19.52. We will be building model by dropping this variable and see what effect it has on the model.

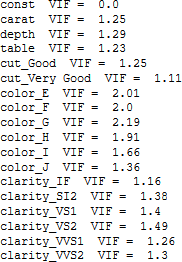
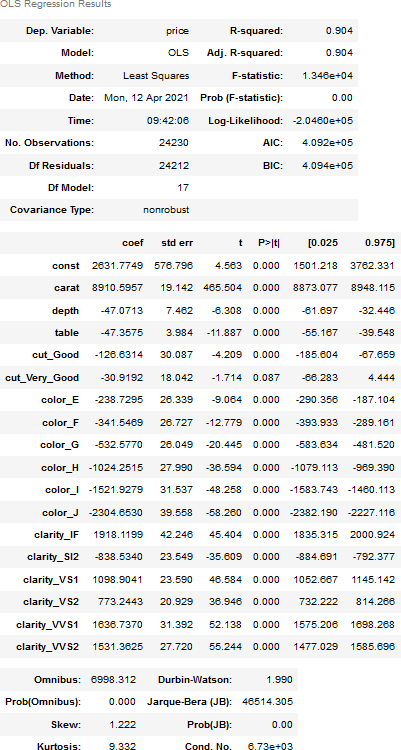
## MODEL 5 (Dropped clarity\_SI1)



Observation:

* + - From above VIF values we can interpret that after removing variable Clarity\_SI1, the VIF value of all the remaining variables are in the acceptance range.
    - Although the VIF values has got down to the range of 0 and 2.2, we can see from the model that P-value for cut\_premium is 0.144. Having a p-value greater than 0.05 implies that the variable is not significantly affecting the Dependent variable.
    - In this iteration we will be building model by dropping variable cut\_premium and see what effect it has on the model.

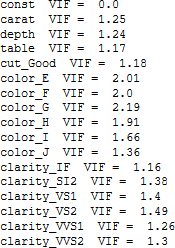
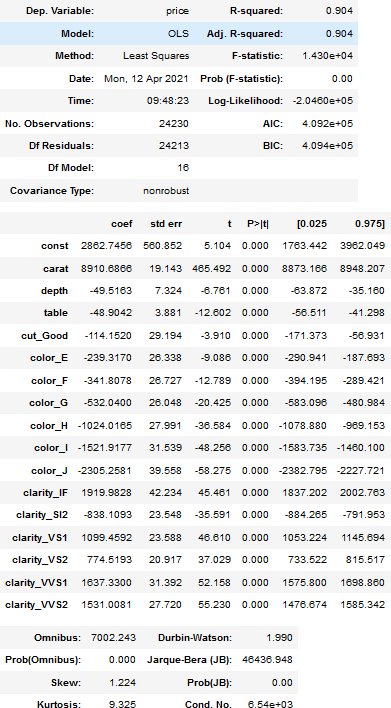
## MODEL 6 (Dropped cut\_premium)



Observation:

* + - From above VIF values we can interpret that the VIF value of all the remaining variables are in the acceptance range.
    - Although the VIF values has got down to the range of 0 and 2.2, we can see from the model that P-value for cut\_Very\_Good is 0.087. Having a p-value greater than 0.05 implies that the variable is not significantly affecting the Dependent variable.
    - In this iteration we will be building model by dropping variable cut\_Very\_Good and see what effect it has on the model.

## MODEL 7 (Dropped cut\_Very\_Good)



Observation:

* + - This model seems to be perfect as compare to other 6 models above.
    - Equation for the model: Price = 2862.7456 + 8910.6866\*carat -49.5163\*depth -48.9042\*table - 114.1520\*cut\_Good – 239.3170\*color\_E -341.8078\*color\_F – 532.04\*color\_G – 1024.0165\*color\_H - 1521.9177\*color\_I -2305.2581\*color\_J + 1919.9828\*clarity\_IF – 838.1093\*clarity\_SI2 + 1099.4592\*clarity\_VS1 + 774.5193\*clarity\_VS2 + 1637.3300\*clarity\_VVS1 + 1531.0081\*clarity\_VVS2
    - Variables which are positively impacting the price are:
      * carat
      * clarity\_if
      * clarity\_VS1
      * clarity\_VS2
      * clarity\_VVS1
      * clarity\_VVS2
    - From above VIF values we can interpret that the VIF value of all the remaining variables are in the acceptance range.
    - Also, the p-values for all the variables are less than 0.05, which implies all the remaining variables are significantly affecting the dependent variable.
    - For the Durbin-Watson (DB) test, we’re looking for a value between 1.5–2.5. A few things to know regarding the DB test:
      * 2: No Autocorrelation
      * 0–1.9: Positive Autocorrelation
      * 2.1–4: Negative Autocorrelation
      * Looking at the OLS Summary, we can see the Durbin-Watson score is 1.935, this score tells us there is no correlation between the model residuals
      * As we do not have any high VIF value and p-values are also under considerable range, we’ll be calling this out final model and will be doing further metric calculations on this model only.

# PART 3

***Objective:*** *If prediction accuracy of the price is the only objective, then you may want to divide the data into a training and a test set, chosen randomly, and use the training set to develop a model and test set to validate your model. Use the models developed in Part (2) to compare accuracy in training and test sets. Compare the final model of Part (2) and the proposed one in Part (3). Which model provides the most accurate prediction? If the model found in Part (2) is different from the proposed model in Part (3), give an explanation.*

The train-test split is a technique for evaluating the performance of a machine learning algorithm. It can be used for classification or regression problems and can be used for any supervised learning algorithm.

### Splitting Data into training (70%) and test (30%)



Observation:

* + - Train and Test dataset are uniformly distributed

Using the same variable that we used to build our final model in previous Models from 0 to 7. We are going to use the same variables to build the Linear Regression Model using Sklearn and then we will compare both of them to see which of them is better

* 1. **MODEL 8 (All Variables)**

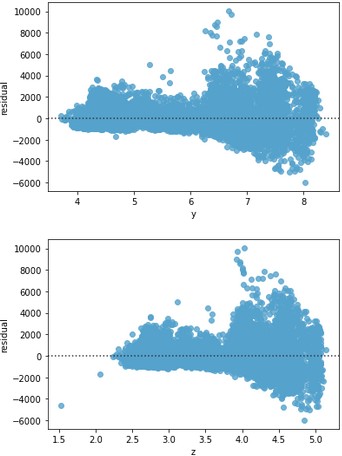
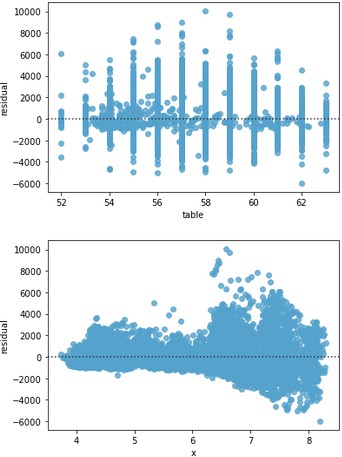
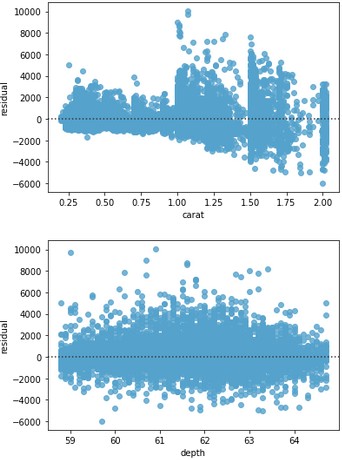


Observation:

* + - R-value of Model 8 is .922
    - RMSE for Train Data Model 8 is 1,024,746
    - RMSE of Test Data Model 8 is 999,852

### Checking Assumptions for Model 8

* + - 1. **Checking for Homoscadicity**



Observation:

* + - * + We can see here that Variable x, y and z are showing the funnel shape, which indicates the Hetroscadecity in Data. So, we will Remove all the variable with Highest Correlation and rebuild the Mode

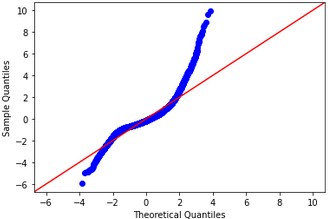
### Checking for Auto Correlation



Observation:

* + - * + Negative Correlation is observed

### Checking for Normality



Observation:

* + - * + The points clearly follow another shape than the straight line hence not normal

## MODEL 9 (Dropped x)



Observation:

* + - R-value of Model 9 is .921 is dropped from .922
    - RMSE for Train Data Model 9 is 1,036,237
    - RMSE of Test Data Model 9 is 1,008,770

### Checking Assumptions for Model 9

* + - 1. **Checking for Homoscadicity**

|  |  |  |
| --- | --- | --- |
|  |  |  |

Observation:

* + - * + Still y and z showing funnel shape. Let’s drop these variable along with all the other insignificant variables and rebuild the mode

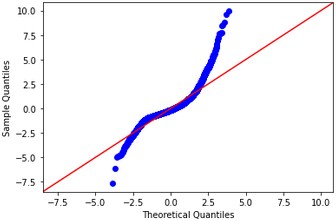
### Checking for Auto Correlation



Observation:

* + - * + Negative Correlation is observed

### Checking for Normality



Observation:

* + - * + The points clearly follow another shape than the straight line hence not normal

## MODEL 10 (Dropped y)



Observation:

* + - R-value of Model 10 is .921 has been reduced
    - RMSE for Train Data Model 10 is 1,037,427 has increased
    - RMSE of Test Data Model 10 is 1,003,152 has reduced

Dropping variables having high VIF values as followed from model 3-7. Intermediate models and respective RMSE Values are as below

## MODEL 11 (Dropped Cut\_Ideal)



* 1. **MODEL 12 (Dropped z)**



## MODEL 13 (Dropped clarity\_SI1)



* 1. **MODEL 14 (Dropped cut\_premium)**



## MODEL 15 (Dropped cut\_Very\_Good)



### Checking Assumptions for Model 15

* + - 1. **Checking for Homoscadicity**

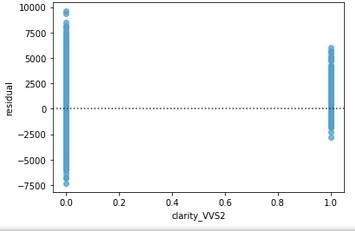
|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

Observation:



* + - * + Errors are heteroscedastic (i.e. OLS assumption is violated), then it will be difficult to trust the standard errors of the model

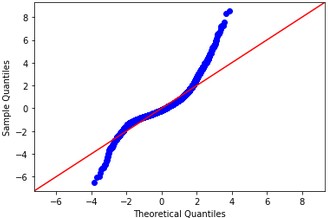
### Checking for Auto Correlation



Observation:

* + - * + Negative Correlation is observed

### Checking for Normality



Observation:

* + - * + The points clearly follow another shape than the straight line hence not normal

## TRAIN-TEST - Model Summary

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model Score** | **Model8** | **Model9** | **Model10** | **Model11** | **Model12** | **Model13** | **Model14** | **Model15** |
| **Adjusted R Square** | 0.9220534 | 0.9211793 | 0.9210887 | 0.9209552 | 0.9100328 | 0.9037828 | 0.9037811 | 0.9037762 |
| **RMSE**  **For Train Data** | 1024746 | 1036237 | 1037428 | 1039183 | 1182778 | 1264945 | 1264968 | 1265032 |
| **RMSE**  **For Test Data** | 999853 | 1008771 | 1003152 | 1004568 | 1183612 | 1265162 | 1265369 | 1265666 |

Now that we have run again 8 Linear Regression Models **(with split data),** we can summarize as below;

We have selected Model 15 because it performed better compared to other 7 models (Model 8 - 14). Although RMSE is high for Model 15 but its Adjusted Rsq. is lower than other models. Hence we recommend using Model 15.

# Business Recommendation

The requirement here is to accurately predict the prices of zircon and hence we have use Linear Regression model to understand the relationship between the independent and dependent variables. Also, we have been able to capture the variables which are significant or sensitive towards target variable i.e. price and also taken out the variables which have least impact on the price.

Below are the key recommendation to the business:

1. Carat is the most significant variable for price prediction. With the unit increase in Carat there is positive uplift in the price. While variable Clarity has also direct and positive impact on the price of zirconia. Hence, company should have a clear focus on Carat and Clarity of Zirconia to get high profit margins.
2. Colour variable has a negative impact on price. Thus company needs to track the variation in colour of zirconia in order to eliminate any losses. This also indicates that a colourless zirconia might have a higher positive impact on prices.
3. Ideal Cut is having a greater positive impact on the price as compare to Premium
4. There are many similar variables is data, hence dimensionality reduction technique like PCA can be applied before creating the model.
5. There is high collinearity in the data, hence there needs to be a thorough investigation required on the related variables and eliminate the multi-collinearity
6. RMSE is very high even after trying number of iterations on the initial model we can further try other models like Random Forest, Artificial Neural Network to achieve higher accuracy for price prediction.