**ADTA 5130**

**FINAL PROJECT REPORT**

Price Behavioral Patterns of Diamonds

SAI VARUN TEJA MUDUMBA

# Introduction

Diamonds are one of the best-known and most sought-after gemstones [1]. They are well known from the ancient times and are used in jewelry. India was the first country to mine diamonds and it was the main hub of the gemstones until Brazil discovered them and started mining for diamonds. Diamonds are thought to have been first recognized and mined in India, where significant alluvial deposits of the stone could be found many centuries ago along the rivers Penner, Krishna, and Godavari [2]. Buddhist works dating from the 4th century BC describe the diamond as a well-known and precious stone but do not mention the details of diamond cutting [1]. The different characteristics of the diamonds which makes them unique from others are its strength, excellent refractive properties, brilliance.

Diamonds are a component of carbon which has its atoms arranged in a crystal format and they are formed deep inside the bottom layers of the earth because of the intense pressure and temperatures. The carbon atoms of the diamonds are packed so rigidly inside such that the color of the diamond varies drastically even with the presence of small percentages of impurities. Since they are formed only in the deeper layers of the earth and that too at higher temperatures and pressures, it makes the extraction process a difficult task. And since the temperatures and the pressure requirements need to be met, there are not formed so common and are rarely found in only some parts of the world. Because of all these reasons, they stand to be one of the costliest gems available in the world. The actual price of the diamonds is not just limited by this, but it is mainly dependent on the grading developed in the 20th century which computes the characteristics of the diamonds commonly based on the four C’s namely the Carat, Clarity, Color, and Cut. Most of the consumers buy the diamonds based on their preferences of different values of these four C’s. As per the 2014 results from the Harvard Atlas of Economic Complexity, India and Belgium are the topmost countries in the world in diamond exports and they contribute 20% of the total diamond exports each.

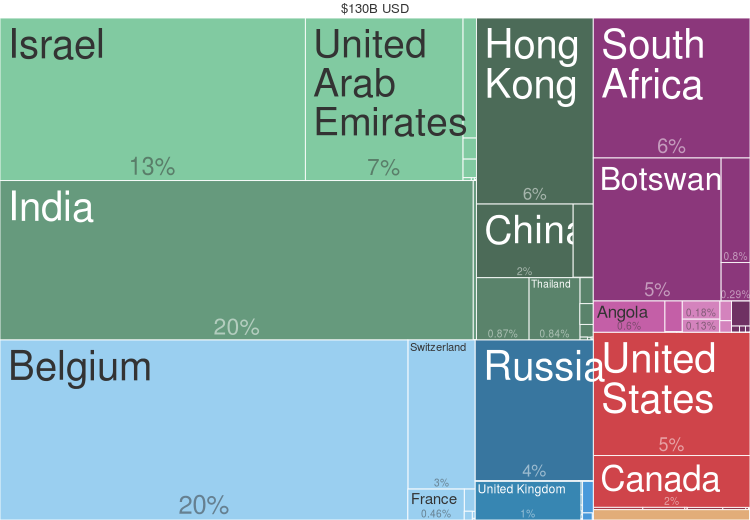


Fig.1. Diamond exports by country (2014) from [Harvard Atlas of Economic Complexity](http://atlas.cid.harvard.edu/explore/tree_map/export/show/all/7102/2014/) [2].

**The four C’s:**

The four C’s namely the Carat, Clarity, Color, Cut are used as the characteristics by the gemologists and they developed a grading procedure in the early twentieth century to classify the diamonds and define their value.

**Carat:**

The carat of the diamond gives us the details about the mass of the diamonds and one carat equals 200 units in milligrams. Of all the grading characteristics available, there might be certain differences in the metrics they use in the classification such as color or clarity, but the carat is defined the same throughout the world, so it determines a strong relationship with the price. The price of the diamond is expected to increase as the carat weight increases since larger carat weights imply larger weight diamonds and hence the price of larger diamonds are obviously higher because they are very rare to be found.



Fig.2. Diamond Carat Weight Measures of a Diamond [3]

**Clarity:**

Clarity defines the perfection of the diamond. It gives us the information about the structural details of the diamond such as cracks or the presence of other materials. These inclusions in the diamonds are graded into categories such as SI1, SI2, VS1, VS2, VVS1, VVS2, I1, IF. These categories are determined by the Gemological Institute of America (GIA).

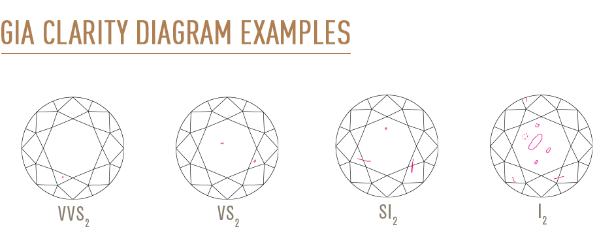


Fig.3. Diamond Clarity Weight Measures of a Diamond [3]

**Cut:**

Cut refers not to a diamond's shape (e.g. round, oval, pear, etc.) but to a diamond's proportions, symmetry, and polish [5]. The cut of a diamond describes the quality of workmanship and the angles to which a diamond is cut [1]. The price of the diamond is dependent on the type of the cut. A good cut can increase the price per carat of the diamond and a poor cut can decrease the price of the diamond accordingly.



Fig.4. The different Diamond Cuts [1]

**Color:**

Color explains about the saturation and the depth of the color of the diamond. The color of the diamonds varies from Colorless to the light yellow. The colorless diamonds price higher as they are very rarely found, and the color grading is divided into 22 grades starting from letter D to Z.

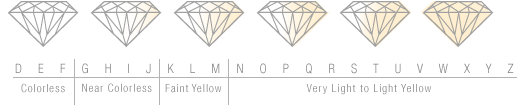
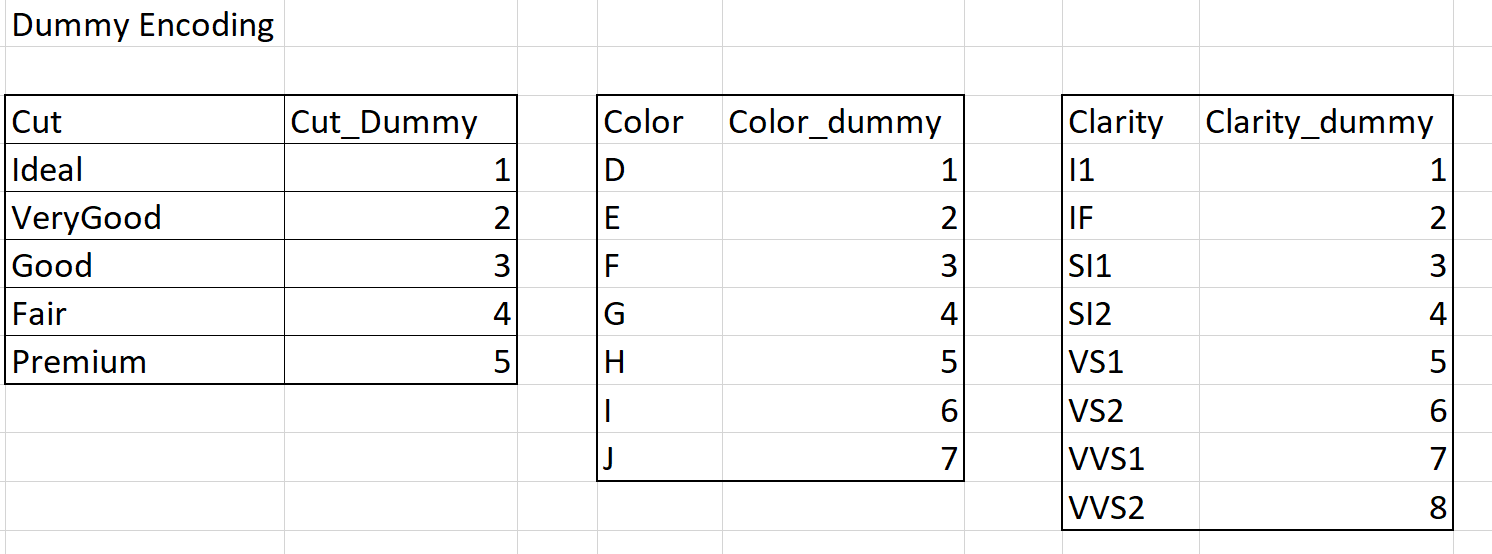


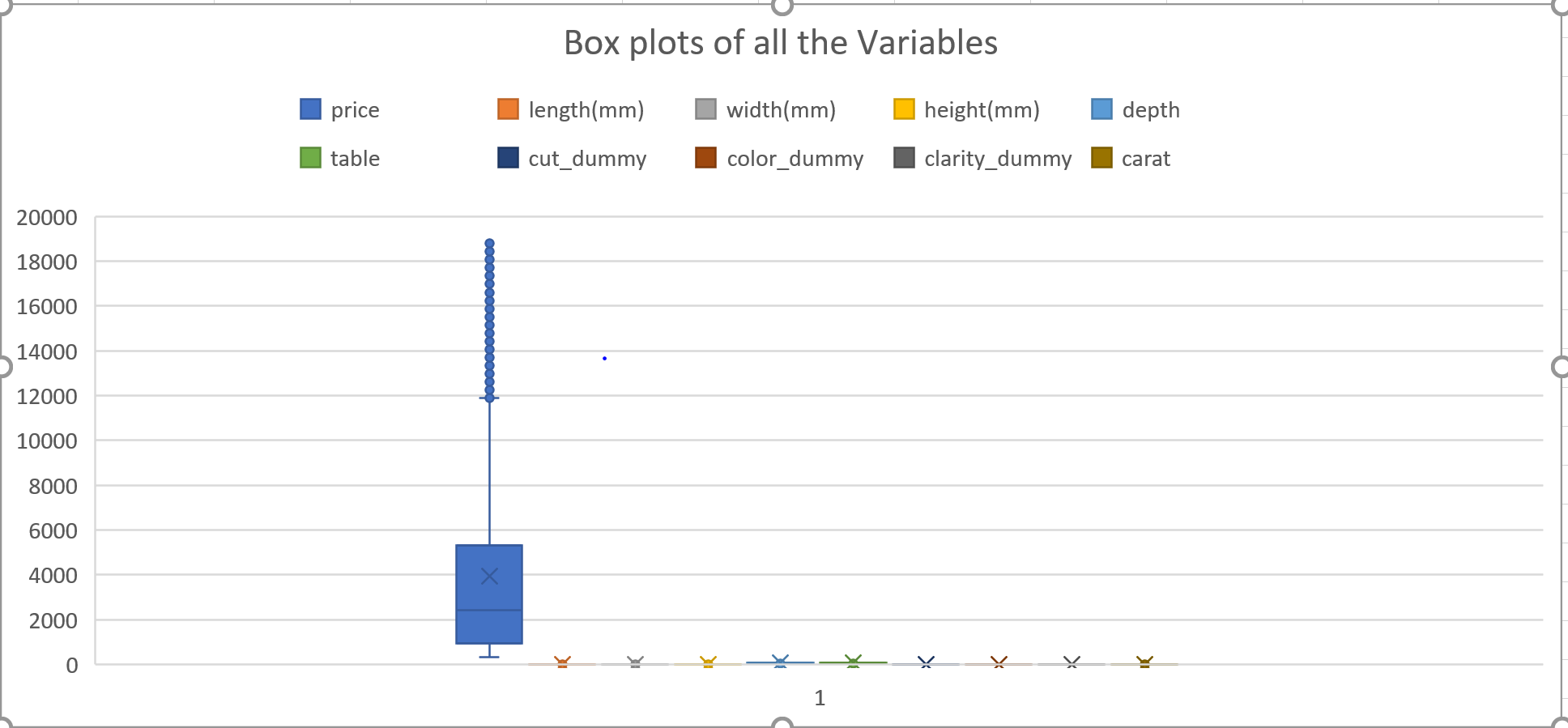
Fig.5. The different Color Gradings of Diamonds [7]

# Exploratory Data Analysis

Before proceeding with building the regression analysis, it is an important step to look at the data distribution of the variables in the dataset. As we look at the data set it has both Qualitative and Quantitative variables and it is important for us to treat the Qualitative (Categorical) variables by creating Dummy variables before we run the regression. The dummy codes for the above variables are assigned as shown in the below picture:

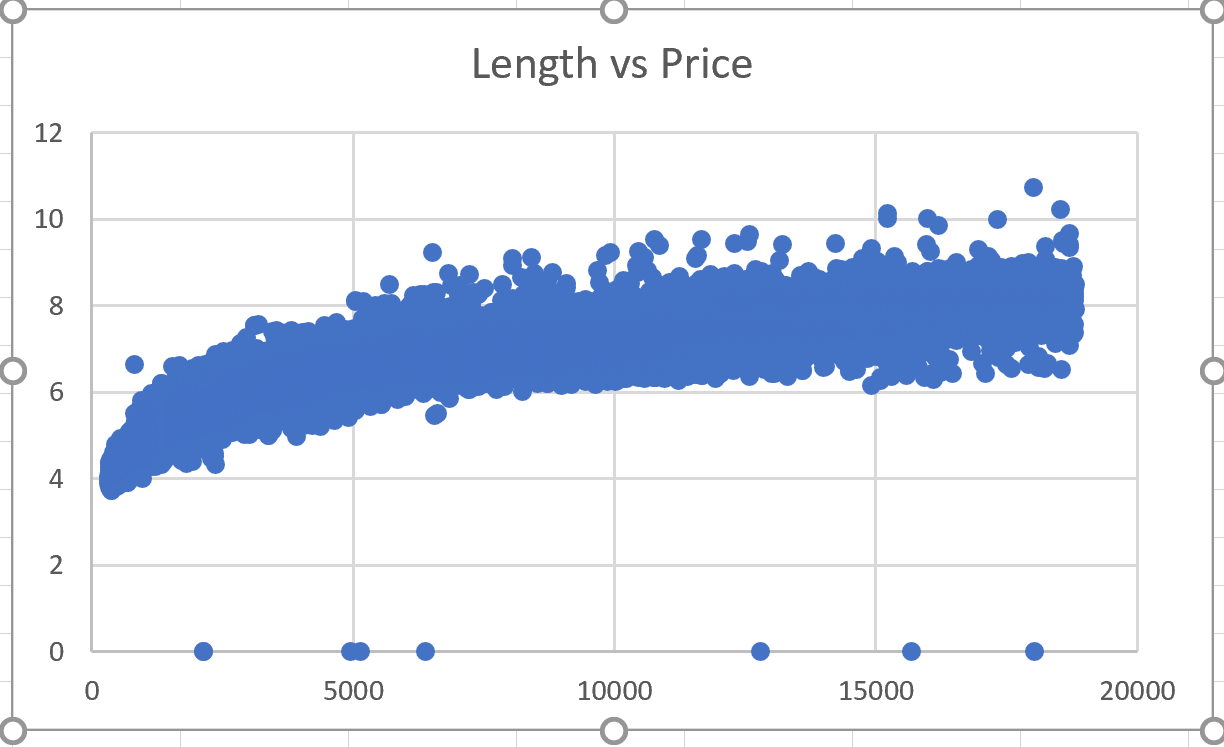
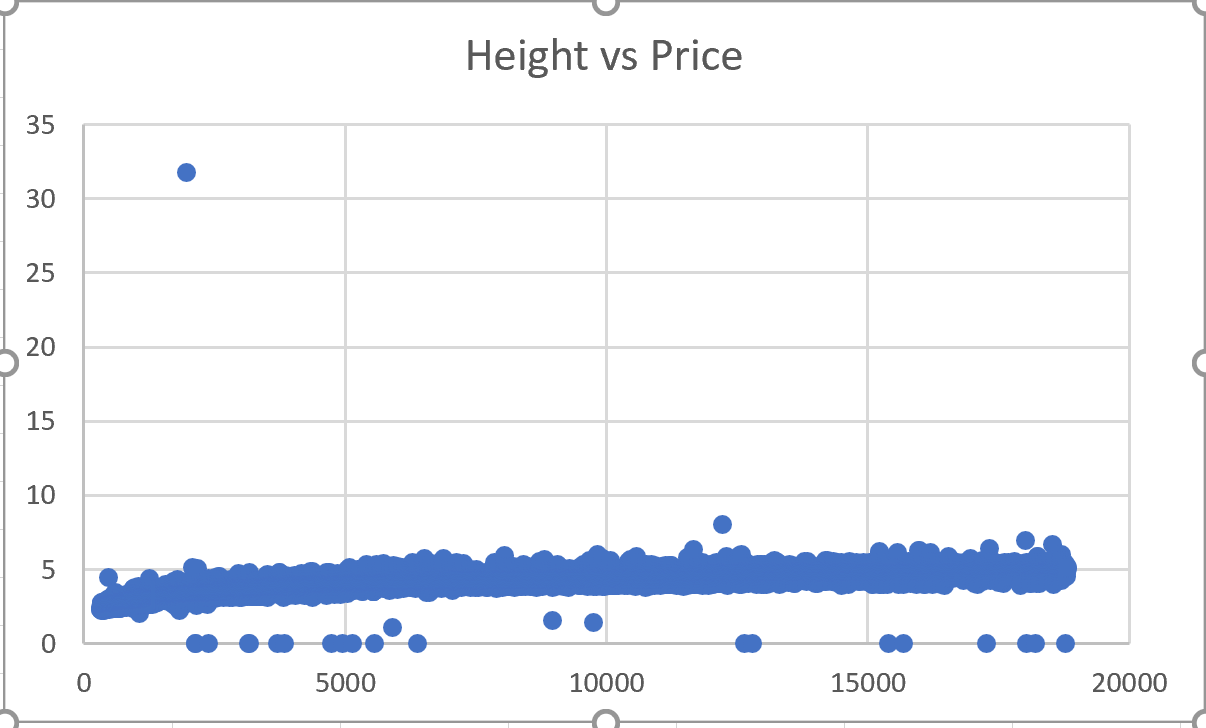


After treating the Cut, Clarity and Color variables with the dummy coding, we can have a box plot to look at the distribution of the data. The box plot is as shown below

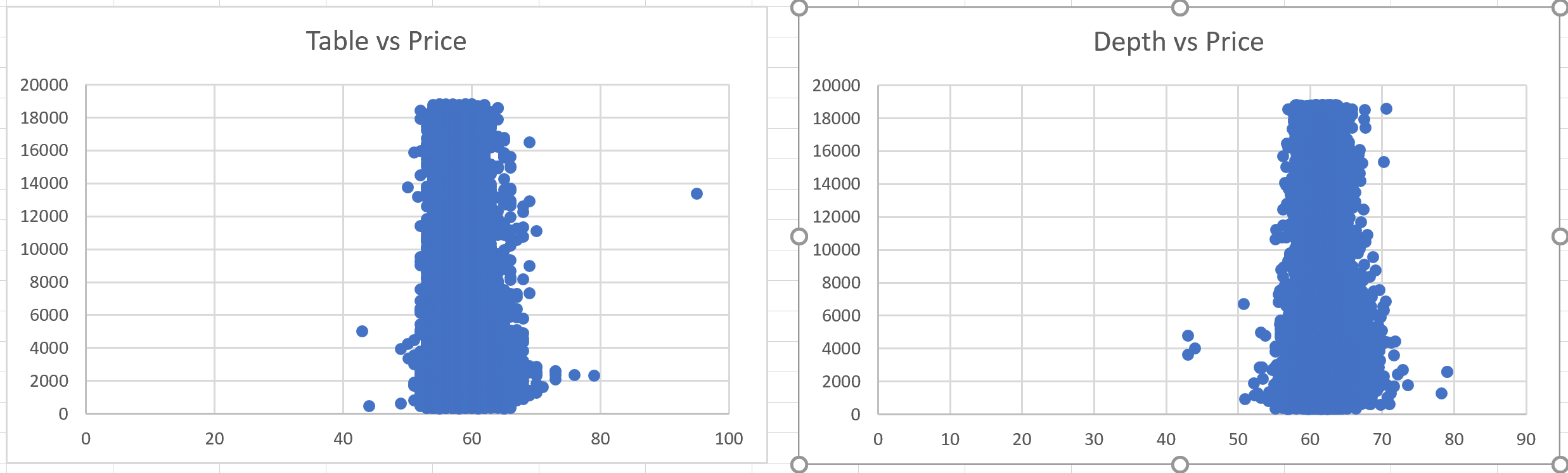


We can see that all the variables except Price are in the same scale and the Price variable is on the higher scale than the others with a mean value of 4000 and most of its prices are between 1000 and 5000, and it also has a small portion of records which had a price range of 12000 – 19000.

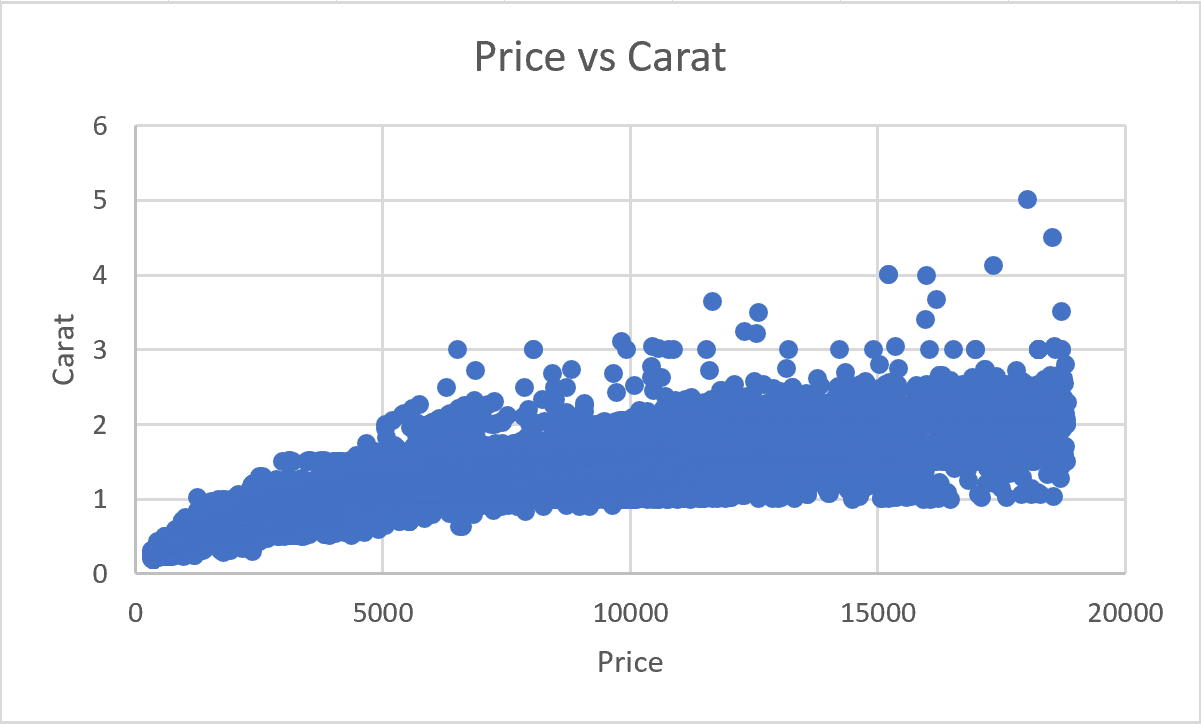
The other variables in the data set other than Carat, Cut, Clarity, and Color are its dimensions such as Length, Height, Width, Depth and Table. The Length, Width, Height has the values of the lengths, heights and widths of the diamonds in mm. If we look at the relationship between these dimensions with the price, we see that they have a linear relation with the price.

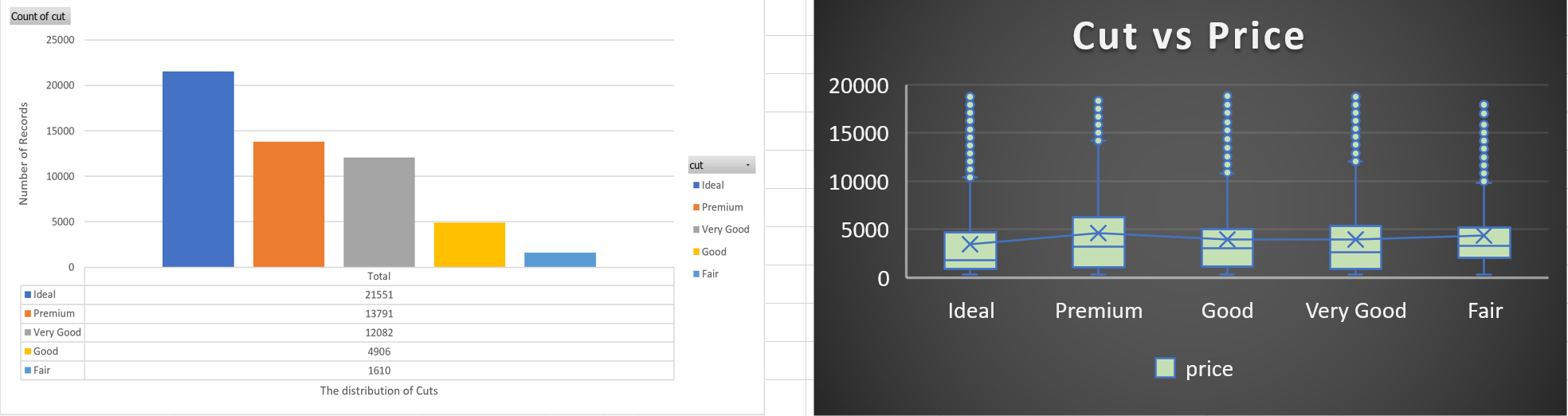
The Table and the Depth variables in the dataset give us the widths of the diamonds compared to its widest points and the heights of the diamonds from the table divided by its average griddle diameter. These values alter with the refraction properties of the light and low values of this make the diamond unattractive or less attractive. We can observe that the price can change drastically for the same or slight change in the value of the Table and Depth.

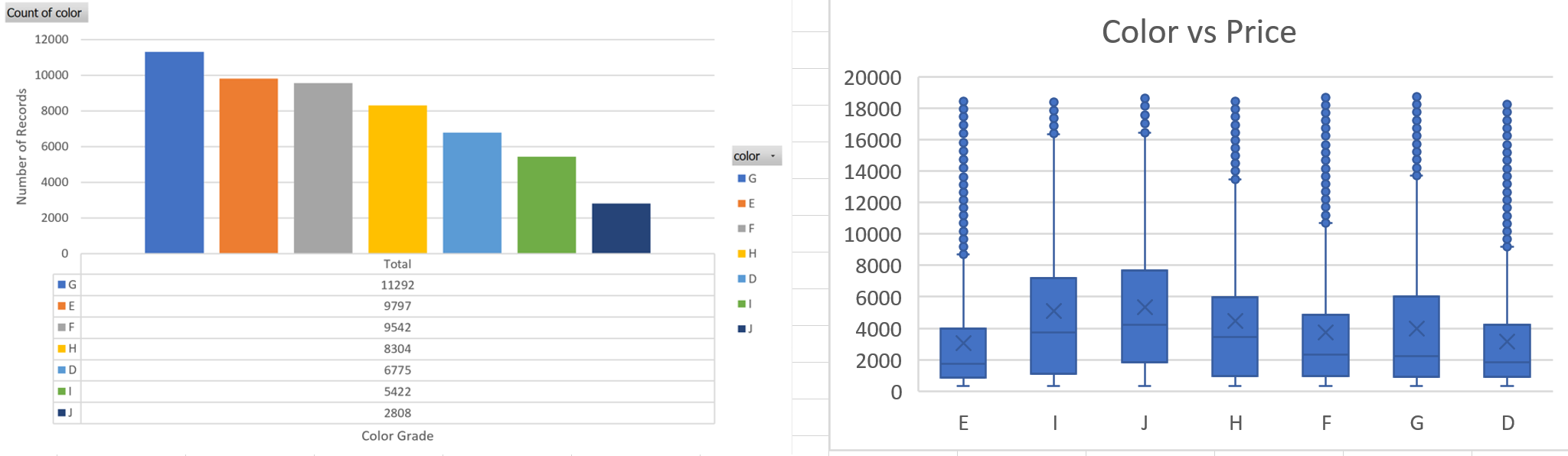


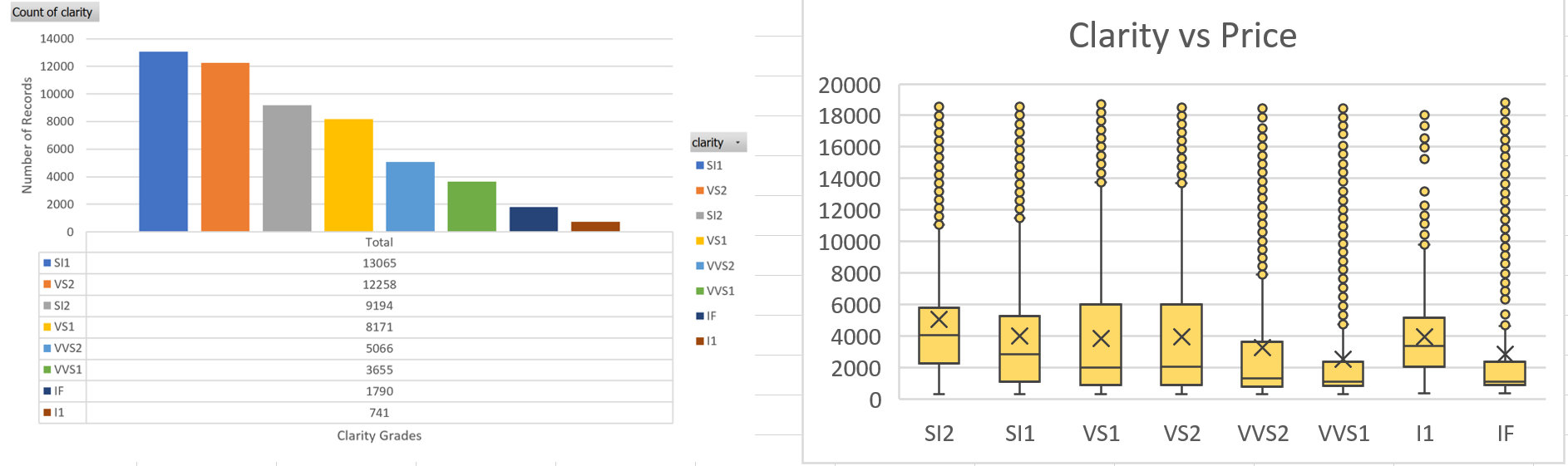
As we already know from the diamonds grading procedure that the four C’s affect the price to a greater percentage. Carat among them seems to alter the price and there exists a strong relationship between the Price and Carat weight.



Let us look at the distribution of the other variables such as Cut, Clarity and Color and their relationship with the price. The graphs show that the dataset has more records with the cut type of Ideal, color grade of G, and a clarity grade of SI1. The price ranges of the Premium cut type has a greater price compared to other cuts, the J color grade has a greater price compared to other colors and similarly, the clarity grades of VS1 and VS2 have greater price compared to other clarity grades. Their distribution is as shown below:

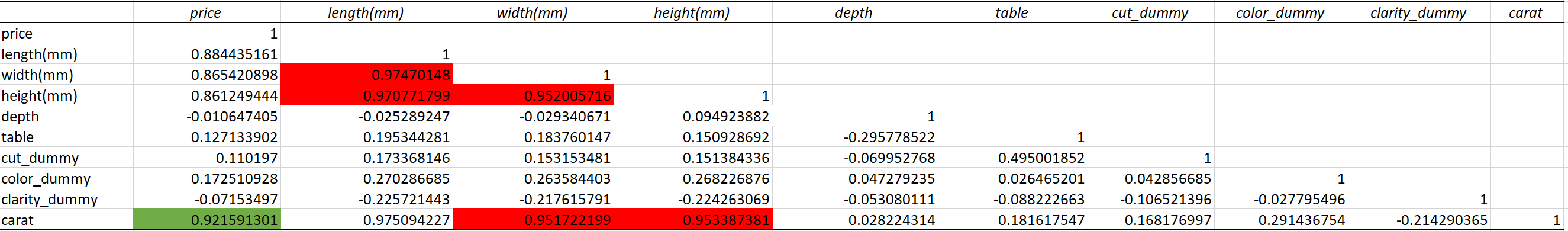






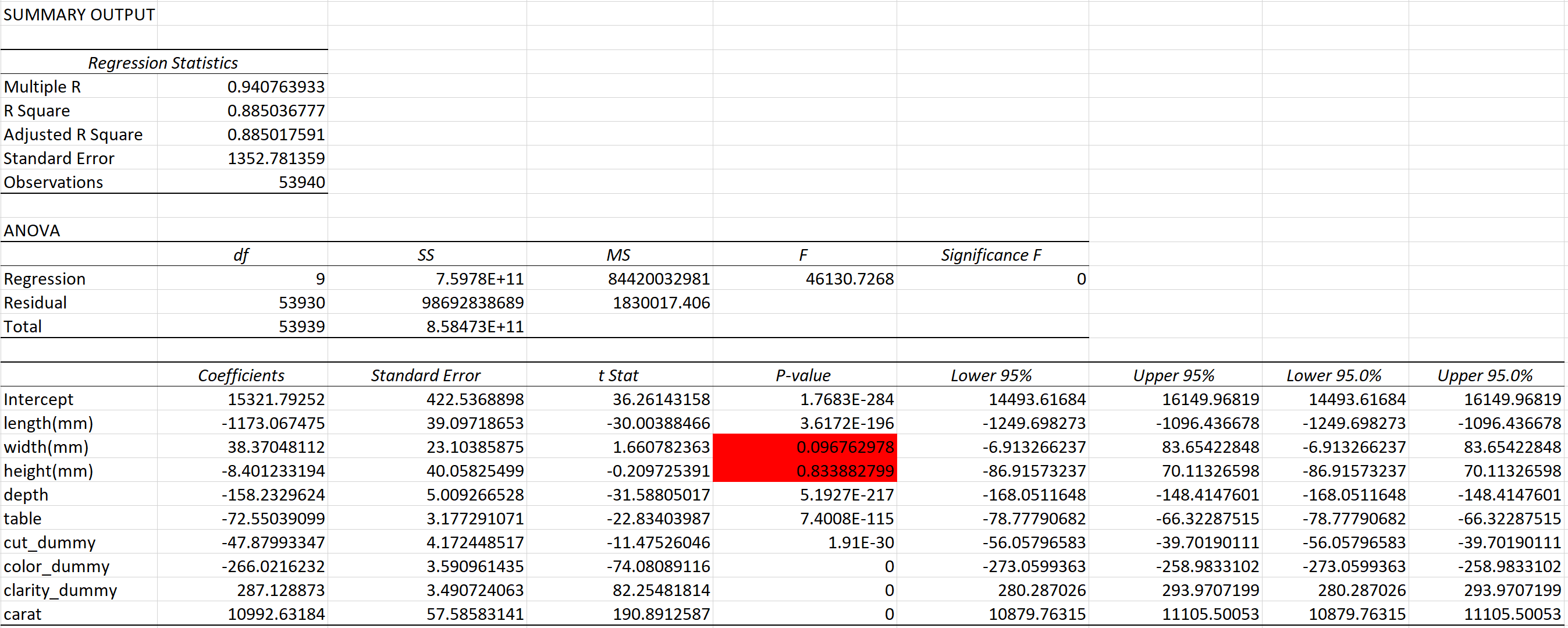
# Data Analysis and Insights

From the above explorations on the data, it is evident that our target variable Price is related to our independent variables such as the four C’s, its dimensions etc., Building a Correlation matrix aids in analyzing the correlation between the variables and check for Multicollinearity. The matrix also helps in answering our research question of finding out the most correlated variable and the least correlated variable with the price.



We can conclude that the price is highly correlated with the carat weight followed by its dimensions. We can also infer from the table that there exists a negative correlation between the price and the depth, price, and clarity and they are the least correlated attributes with the price. It is also evident that there exists a strong correlation between some of the independent variables and are highlighted in red in the above picture. Since carat is highly correlated with the price and the dimensions, we cannot have dimensions in our regression model since it results in an unreliable output. So, we need to eliminate them to get some significant price predictions.

If we run the Regression Model by taking all the inputs, we get the following Regression output.

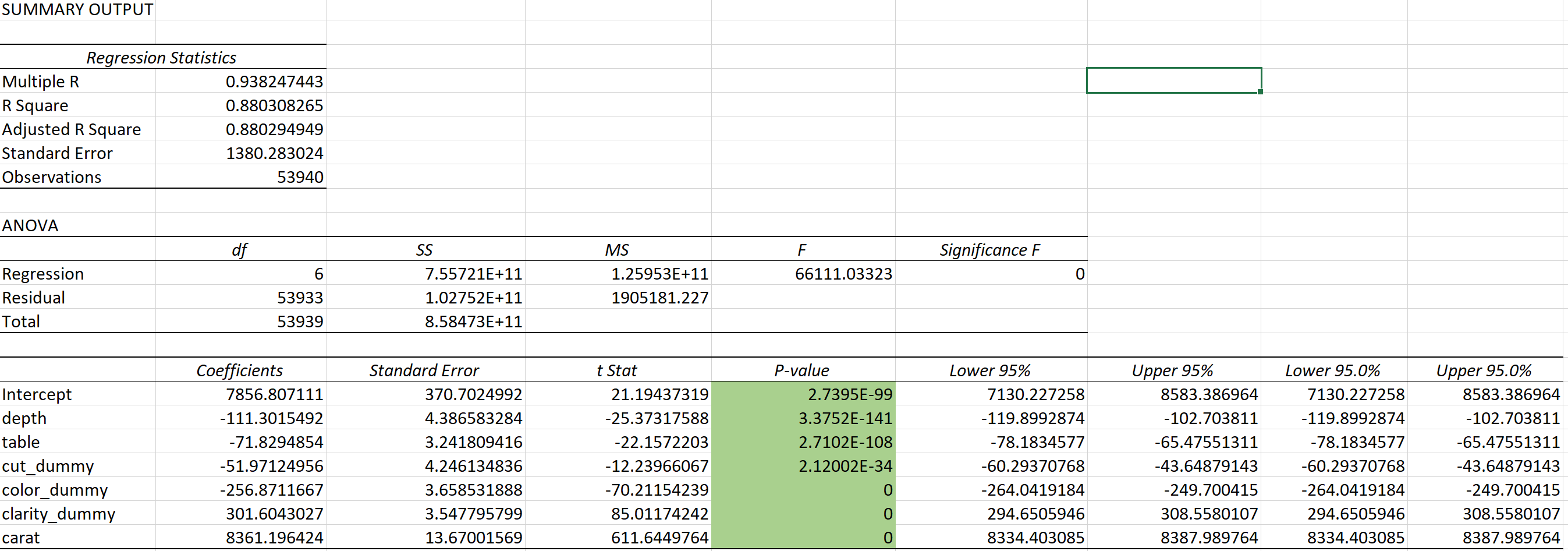


With all the variables as inputs, they’re able to explain 88.5% of the variances in the price and the Price equation is given as:

Price=15321.8+10992.63(Carat)+287.13(Clarity\_dum)-266(Color\_dum)-47.9(Cut\_dum)-72.5(table)-158.2(depth)-8.4(height)+38.3(width)-1173(length)

Since there exists Multicollinearity between Carat, length, width and height we can have only one of them in the model. And from the above equation it is clear that width and height are insignificant (p-value>0.05) so we cannot have them in the model. So in between Carat and length, Carat is highly correlated with the price, so we proceed with Carat removing length.

The regression model without length, width and height resulted in the R2 value of 88.5%, which tells us that only 0.5% of the variance in price is explained by them. And the Regression table is as shown below:



The equation is modified as

Price=7856.8+8361.2(Carat)+301.6(Clarity\_dum)-256.9(Color\_dum)-51.97(Cut\_dum)-71.82(table)-111.3(depth)

This explains that as the carat increases by one unit, it alters the price by $8361 and in the similar fashion based on the Clarity grading adds a multiple of $301.6 based on the type of Clarity grading, Color subtracts a multiple of $256.9 based on the type of color and even the cut subtracts a multiple of $51.97 based on the type of cut. Table and depth have a negative impact on the price and it decreases the price by $71.82 and $111.3 respectively.

* For example, if we need to predict the price of the diamond with the following values:

Carat:2, Clarity: I1(clarity\_dummy=1), Color: D (clarity\_dummy=1), Cut: Ideal(cut\_dum=1), Table: 62, Depth: 58.1

We get the price calculated as:

Price=7856.8+8361.2(2) +301.6(1)-256.9(1)-51.97(1)-71.82(62)-111.3(58.1) => $ **13651.91374**

So, by using this model of price prediction of diamonds, the customers can make use of this model and estimate the price of the diamonds that they wish to buy by putting in the carat weights, color, clarity, cut, the table and depth values. And with this estimated price details, they can compare it to the retailer said price and this aids them in saving their money instead of spending a huge amount of money than its worth. This data presented does not take into consideration of the supply and demand factors. If the supply is less than the demand, then there is an increase in the price of the diamonds. So, this data lacks the power to estimate the prices in such scenarios. And, the taxes on goods like diamonds depends on the countries and this data lacks the taxes information based on the country as well which makes it incompatible to predict the accurate diamond prices. So, assuming data covers all the other limitations then this can guide the buyers helps them to track the diamond prices and save them from greedy retailers.

### References

Diamond (gemstone). (2019, March 06). Retrieved from <https://en.wikipedia.org/wiki/Diamond_(gemstone)>

Diamond (gemstone). (2019, March 06). Retrieved from <https://en.wikipedia.org/wiki/Diamond_(gemstone)>

What is Diamond Carat | The 4Cs of Diamond Quality by GIA. (n.d.). Retrieved from <https://4cs.gia.edu/en-us/diamond-carat-weight/>

How Carat Weight Affects Diamond Price. (2015, November 20) by Joe Leone. Retrieved from https://www.diamondlighthouse.com/blog/2014/10/23/how-carat-weight-affects-diamond-price/

Diamond Cut. (n.d.). Retrieved from: <https://www.lumeradiamonds.com/diamond-education/diamond-cut>

10 Different Cuts of Diamonds by Henry Jones. Retrieved from: <https://www.quora.com/What-are-the-different-cuts-of-diamonds>

Diamonds, E. (n.d.). The 4 C's. Retrieved from https://enchanteddiamonds.com/education/understanding-diamond-color

Significance of Diamond Depth And Table Percentages. (n.d.). Retrieved from https://beyond4cs.com/grading/depth-and-table-values/