Aseem Malhotra

101557400

STA 20010

Statistical Computing

Assignment 1:

Modelling Diamond prices

**Table of Content**

|  |  |
| --- | --- |
| **Chapters** | Page Number |
| **1. INTRODUCTION** | Page 1 |
| * 1. **Purpose** | Page 1 |
| * 1. **Background** | Page 1 |
| **1.3 Investigation** | Page 1 |
| 1. **Exploration** | Page 2 |
| **2.1 Carat** | Page 2 |
| **2.2 Cut** | Page 3 |
| * 1. **Color** | Page 5 |
| **2.4 Clarity** | Page 7 |
| **3.0 Regression** | Page 9 |
| **3.1 Regression Model 1** | Page 9 |
| **3.2 Regression Model 2** | Page 10 |
| **3.3 Regression Model 3** | Page 11 |
| **3.4 Regression Model 4** | Page 11 |
| **4. Conclusion** | Page 12 |

**1. INTRODUCTION**

* 1. **Purpose**

The purpose of this report is to find the strength of the relationship between the price of a diamond and the four major categories in which a diamond can be classified – Carat, Cut, Clarity and Color.

We need to find which relationship is strongest and find a regression model that can be used to predict price of a diamond based on the four C’s.

* 1. **Background**

We have a dataset of almost 54,000 diamonds prices. For each diamond, we have:

* Carat of the diamond which is the unit of measurement of weight of diamond. The bigger the diamond the higher it will cost.
* We have the Cut of the diamond which is divided into 5 categories Ideal, Premium, Very Good, Good and Fair with Ideal being the most expensive and Fair being the least expensive.
* The Clarity of the diamonds has eight factors I1, SI1, SI2, VS2, VS1, VVS2, VVS1, IF ranging from least to most clarity in that order.
* The last classification is the Color of the diamond divided into 7 sub-groups from D to J, with D (colorless) being considered the highest quality and J the worst.
  1. **Investigation**

For the first part, we would consider each individual variable and its effect on the price of diamonds. We will use graphs, charts and statistical calculation for find the strength of the relationship between each C-variable and price of diamonds.

Than we will try and come up with regression equation/model that will show how much each variable contribute to the price of the diamond. We will use this regression equation to predict price of diamond based on their variables and check the accuracy of our models.

In the end bases on the individual variable analysis and regression model we will rank the variables in order of their effect on diamond price and consider any limitation of our analysis.

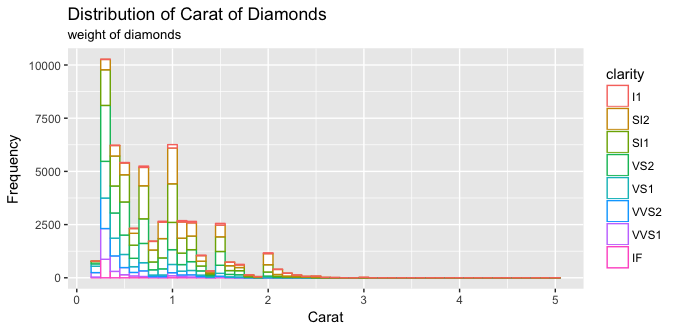
1. **Exploration**

**2.1 Carat**

Diamonds are weighed in units of carat. 1 carat is equal to 0.2 grams. A summary of carats is given in *Table1* below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Table1*: summary of carat | | | | | |
| Min. | 1st Qu | Median | Mean | 3rd Qu. | Max. |
| 0.2 | 0.4 | 0.7 | 0.79 | 1.04 | 5.01 |

Half of the diamonds weigh below 0.7 carats and the other half between 0.7 to 5.01. The middle half of carat is between 0.4 and 1.04 carats. We can see from *figure1* that distribution of carats of diamonds in positively skewed with a long tail between 2 to 5 carats.

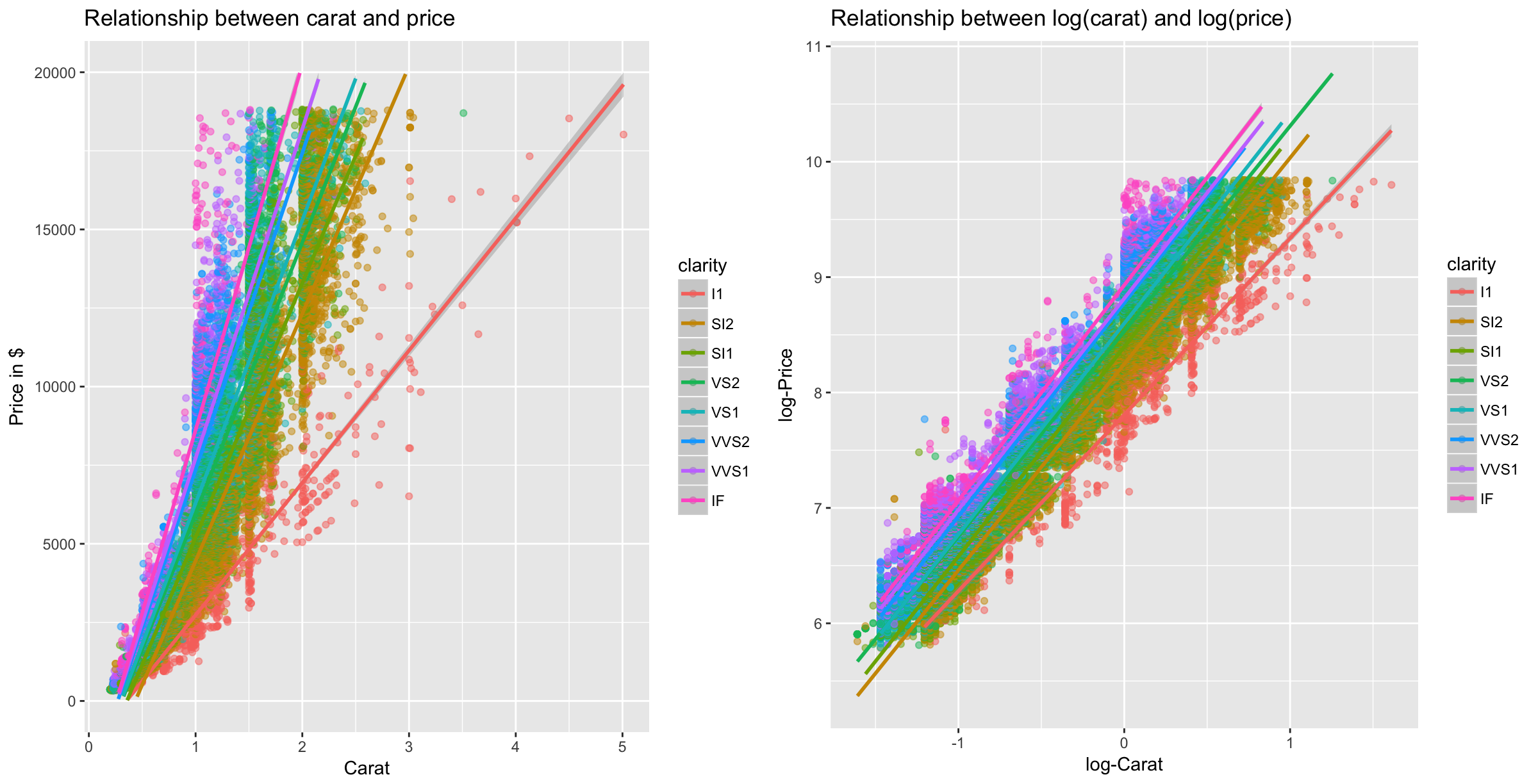


*Figure1*: Distribution of Carat of diamonds.

**Relationship between Carat and Price**

*Figure 2* shows the relationship between Carats and their respective Price of diamonds. Carat has a strong, positive, and linear relationship with price with r = 0.92 (figure 2 (a)). as the distribution is skewed talking a log value of both price and carat will help reduce the outliers. The log-Price and log-Carat have a strong, positive and linear relationship with r = 0.97.

We can be confident that carats have a Significant impact on price of diamonds.



1. (b)

*Figure 2*: Relationship between Carats and Price (by Clarity).

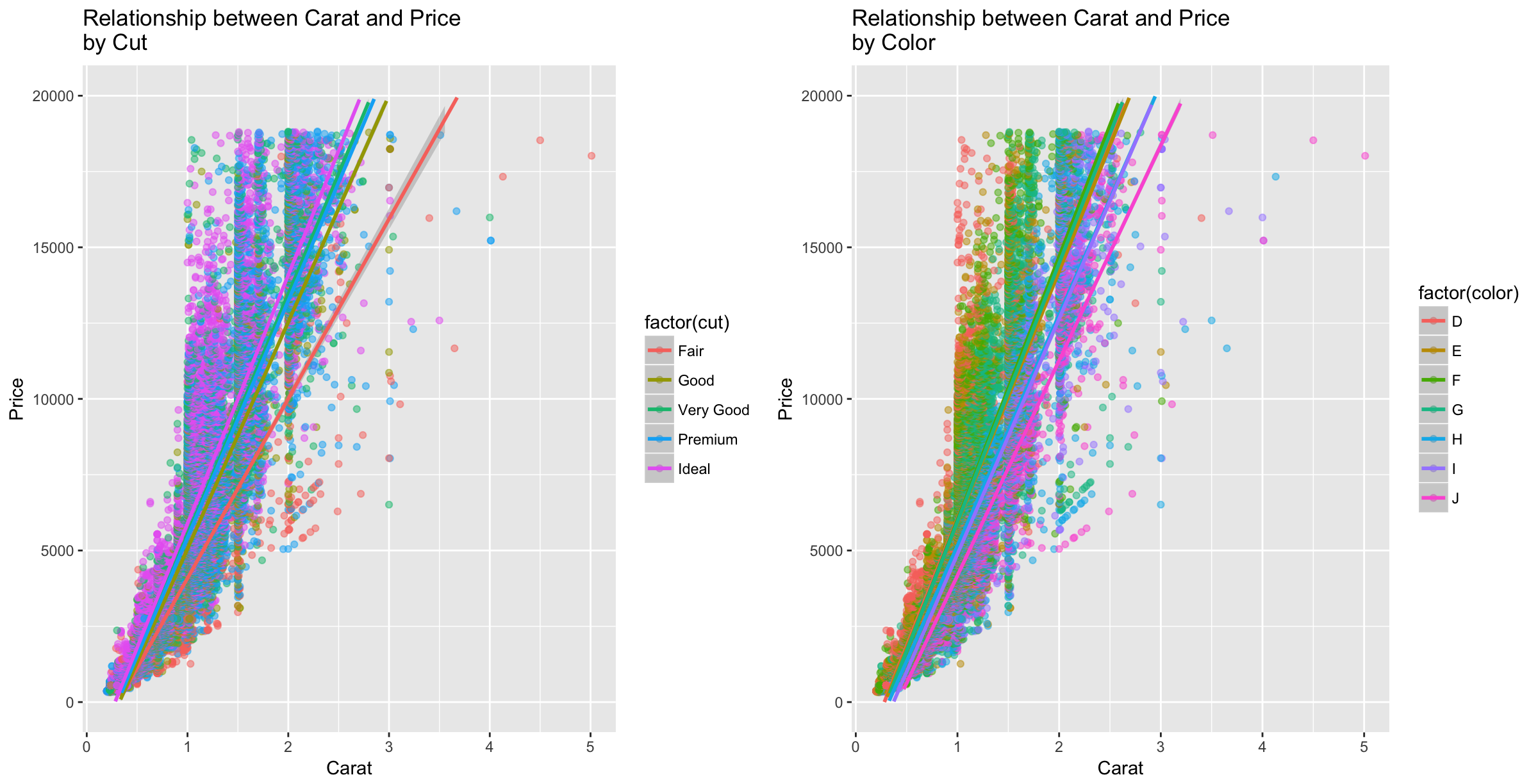
**Relationship between Carat and Cut, Color and Clarity**

We will now look at the how the distribution of other three factors, Clarity, Cut and Color are related to the Carat. *Figure 2* (a) the diamonds are distributed by their clarity and the highest and most expensive diamonds are also the smallest (IF). As clarity decreases the carats increases and majority of the biggest diamonds are of lowest grade clarity (I1).

A similar trend can be seen for cut and color of diamonds but not to the same extent. *Figure 3(a)* shows the distribution by cuts of diamond, the Fair and Good graded cut (lower grade) weigh more than higher graded Ideal and premium cuts.

We can see the same characteristics for distribution by color of diamonds (*Figure 3 (b))*. The

Higher graded colors D to G weigh less than lower graded colors H to J. But this variance is not as large as for cut and clarity.



1. (b)

*Figure 3*: Relationship between Carats and Price (by Cut and Color)

As we have already seen the strong relationship with carat and price the relationship of price and other factors can be distorted by its size. So, for each other factor we will compare ratios of price and carat (price/carat) with the factors and distribution of each variable for each quartile of carat.

Quartile 1: Diamonds less than equal to 0.4 carats.

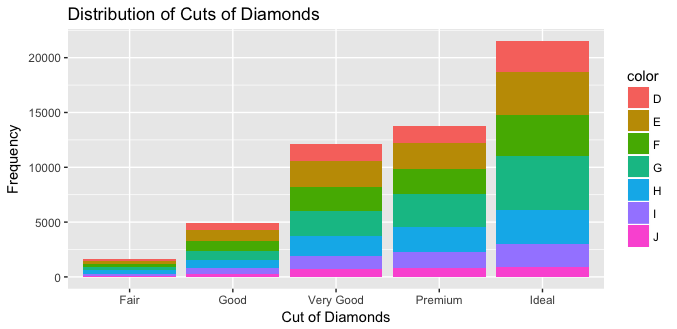
Quartile 2: Diamonds more than 0.4 carats but less than equal to 0.7 carats.

Quartile 3: Diamonds more than 0.7 carats but less than equal to 1.04 carats.

Quartile 4: Diamonds greater than 1.04 carats.

**2.2 Cut**

the cut of the diamonds can vary in 5 different way from fair (lowest grade) to Ideal (highest grade). Majority of the diamonds are higher graded with higher grade from Very good to ideal (*Figure 4)* and a small number are Fair and Good graded.

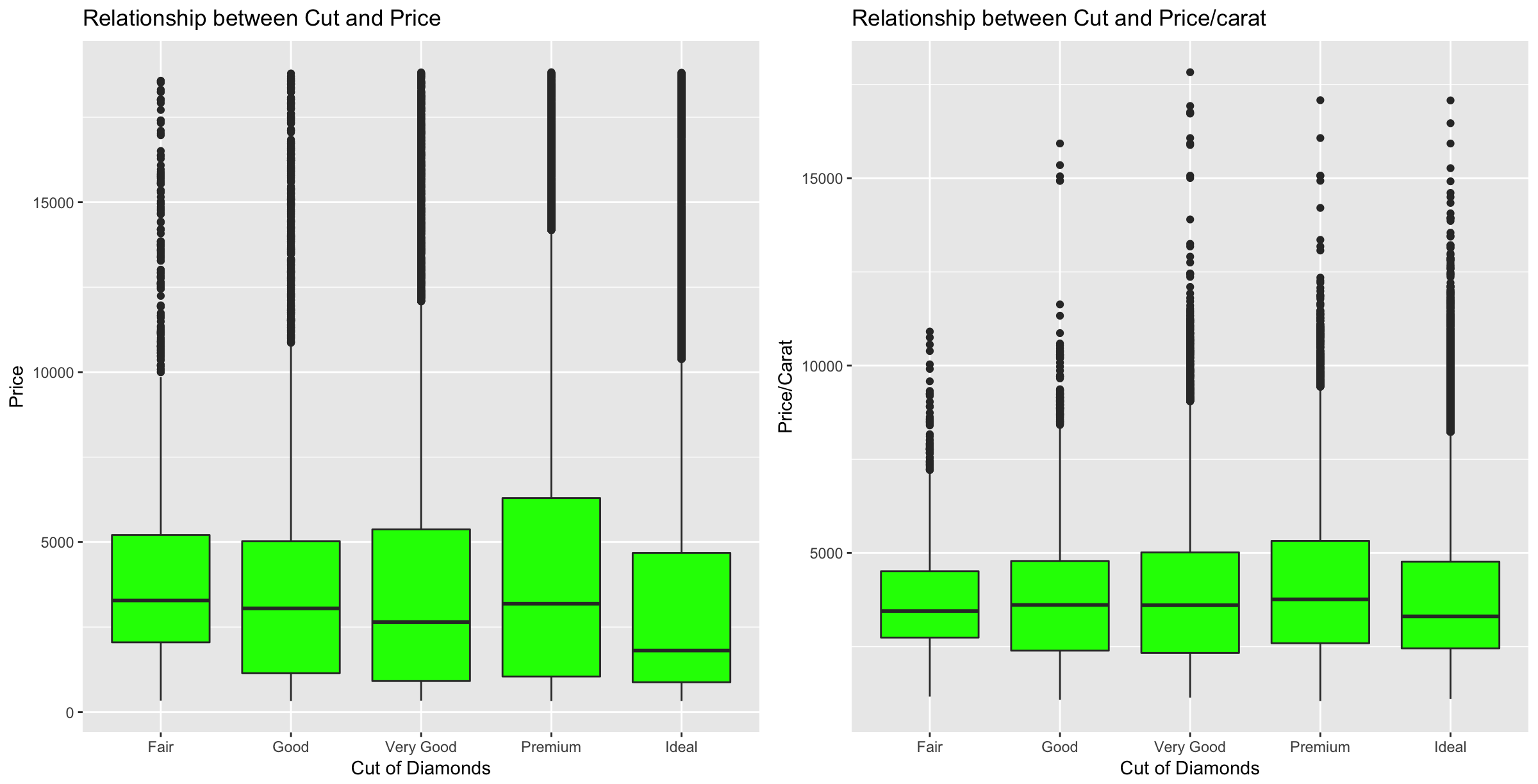


*Figure4:* Distribution of Cuts.

We expected the grade to reflect in the prices but it does the reverse, see *Table 2* below. Premium graded cuts have the highest mean Price of $4584 followed by Fair cut $4359. And Ideal cut has the lowest mean price of $ 3458. But as predicted this can be due to difference in mean weight. Mean weight for Fair cut is 1.05 carat and it decreases with grade to 0.703 carat for Ideal grade.

|  |  |  |  |
| --- | --- | --- | --- |
| *Table 2*: Mean Price, Carat and ratio(price/Carat) by Cut | | | |
| Cut | Mean Price ($) | Mean carat | Mean Price/Carat |
| Fair | 4359 | 1.05 | 3767 |
| Good | 3929 | 0.849 | 3860 |
| Very Good | 3982 | 0.806 | 4014 |
| Premium | 4584 | 0.892 | 4223 |
| Ideal | 3458 | 0.703 | 3920 |

When we consider the ratio of price/carat it gives us a clearer picture of effect of cut on prices. The ratio is least for fair cut 3767 and increases with grade. This difference can also be seen in *figure 5* below where *figure 5(a)* shows a decrease in price as with respect to increase in grade of cut but *figure 5(b)* shows a more constant price to cut relationship.



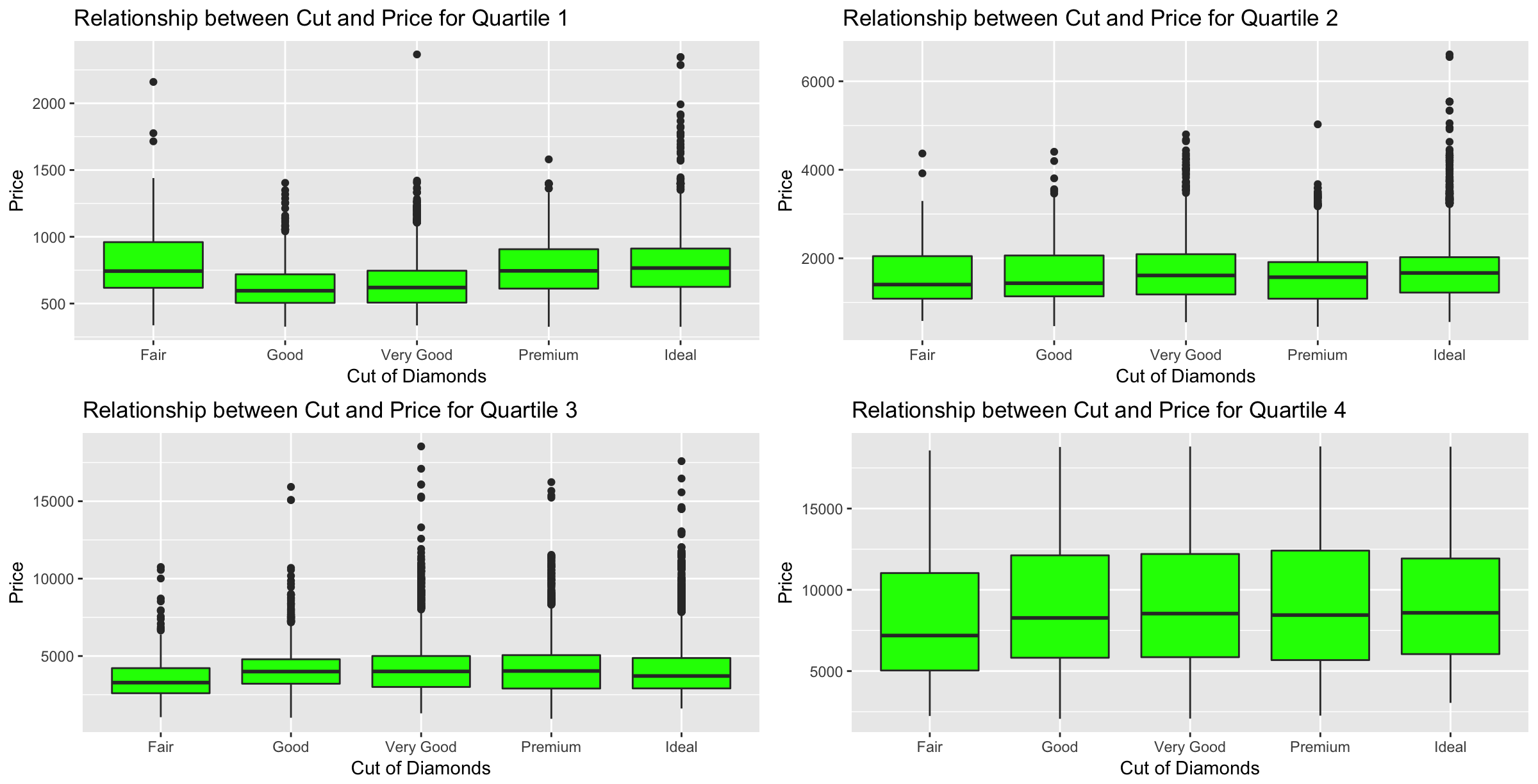
1. (b)

*Figure5:* Comparing price and cuts and Ratio of price/carat to cuts.

*Figure 6* shows the relationship between price and cut for each quartile of carat. For Q1 the mean price of Fair cut is higher than then Good and Very Good cut and about the same as mean price for Premium and Ideal, but this could just be because

* It’s difficult to differentiate between cuts in smaller diamonds.
* The sample of fair graded diamonds is very small compared to other grades so it might not be accurate.

As size increases on Q2, Q3, and Q4 we can see mean price of higher cuts are more than lower cuts.

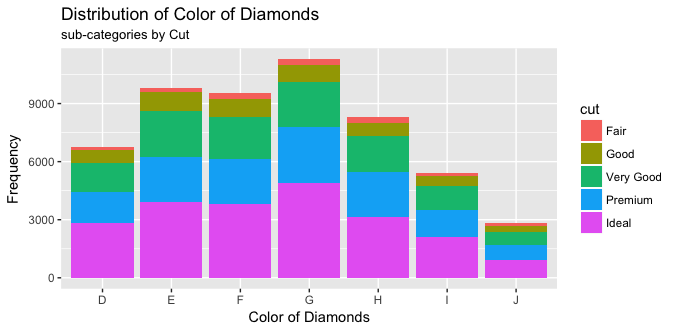


*Figure 6:* relationship between price and cut for each Quartile.

So, we can conclude there is a positive but weak relationship between price and cut of diamonds.

* 1. **Color**

The colors of diamonds are graded between D and J, with D, E and F being totally colorless are quite rare and expensive and G, H, I and J treading around near colorless and more common and less expensive.

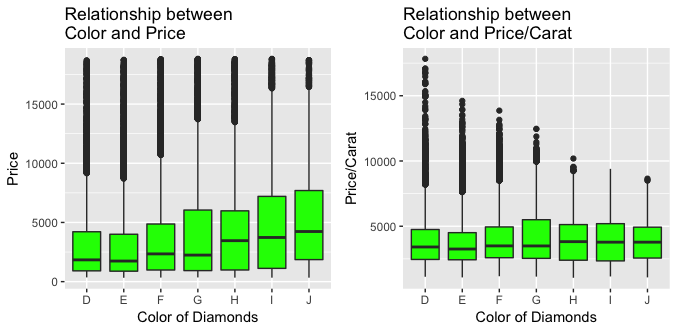


*Figure7:* Distribution of Color.

When we compare the mean prices per color we see (*Table 3)* that contrary to expectation price for higher graded color (D to F) is considerably less then lower graded color (G to J). But higher graded color also weight (0.658 carat) about half of lowest graded color (1.162 carat).

|  |  |  |  |
| --- | --- | --- | --- |
| *Table 3*: Mean Price, Carat and ratio(price/Carat) by Color | | | |
| Color | Mean Price ($) | Mean Carat | Mean Price/Carat |
| D | 3170 | 0.658 | 3953 |
| E | 3077 | 0.658 | 3805 |
| F | 3725 | 0.737 | 4135 |
| G | 3999 | 0.771 | 4163 |
| H | 4487 | 0.912 | 4008 |
| I | 5092 | 1.027 | 3996 |
| J | 5324 | 1.162 | 3826 |

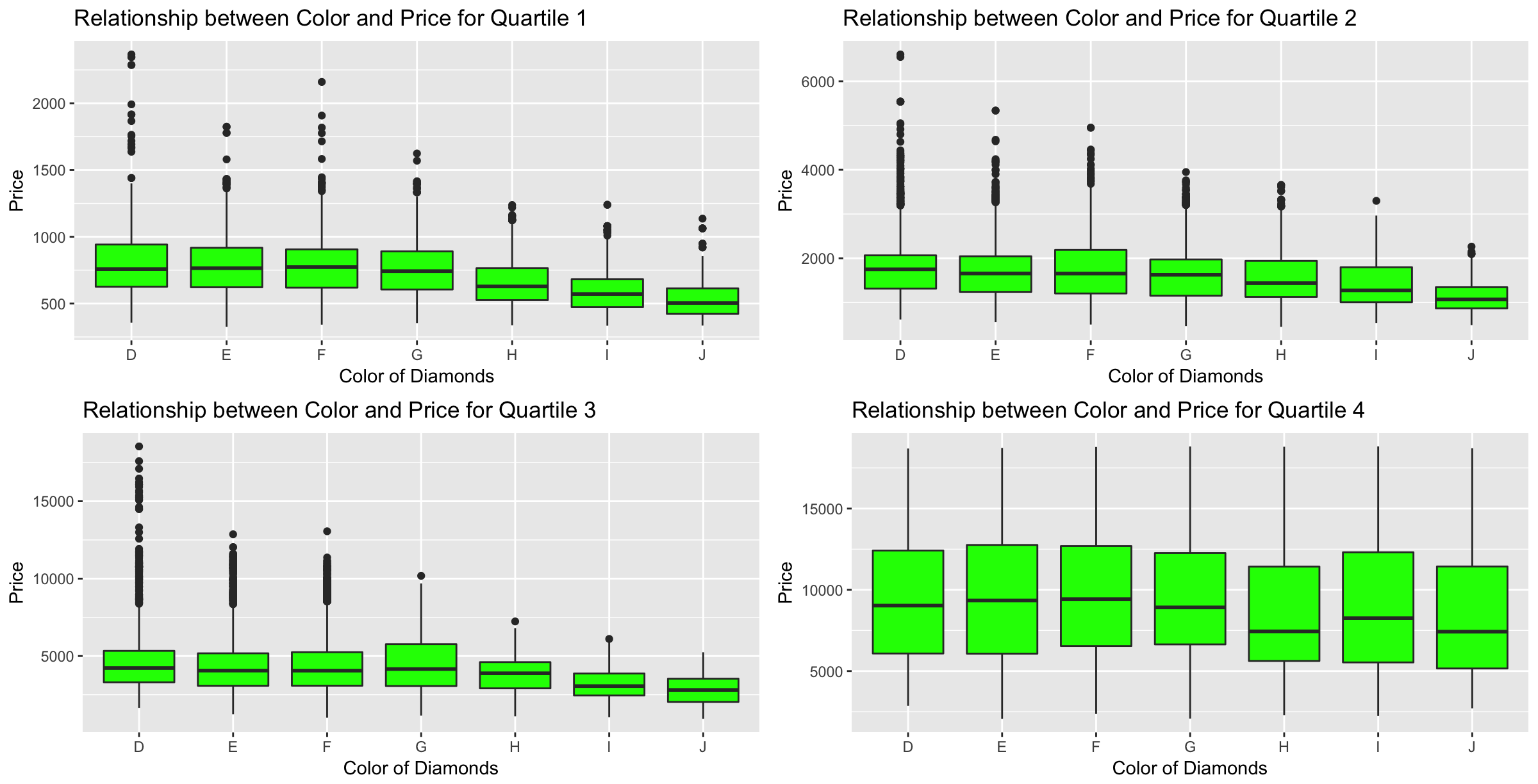
Comparing the ratios of price/carat the difference in that we saw in mean prices evens out and all ratios are between 3805 to 4163. This can also be seen in *Figure 8.* The boxplot (a) comparing the prices rises as we go lower in graded but boxplot (b) comparing price/carat rations looks even for all colors.



1. (b)

*Figure8:* Comparing price and color and Ratio of price/carat to color.

*Figure 9* clearly shows that for quartile 1-3 the mean price colorless diamonds are slightly higher then diamonds with color and for heavy weight diamonds the prices variation is very small.

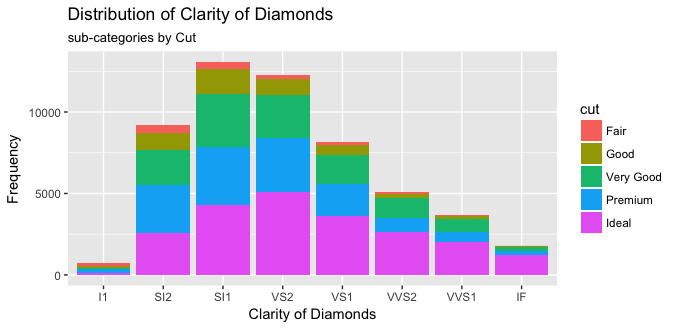


*Figure 9:* relationship between price and color for each Quartile.

We can conclude that color does effect price but this relationship does not seem big.

**2.4 Clarity**

The clarity of diamond is graded from I1 (lowest grade) to IF (highest graded). The sample for I1 graded diamonds is very small compared to others so its estimation cannot be as accurate as others.

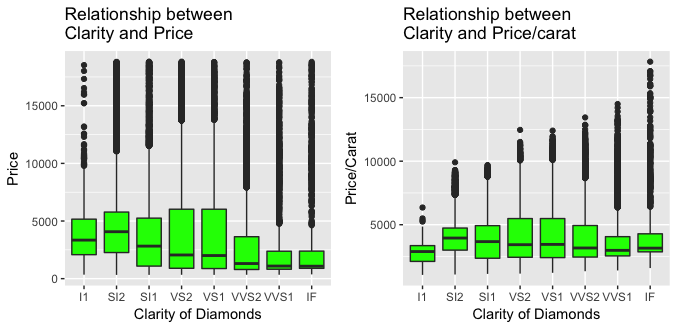


*Figure10:* Distribution of Color.

The mean price of lowest clarity graded diamond (*see Table 4 below)* I1 ($3924) and SI2 ($5063) is considerably more than highest clarity graded diamonds, VVS1 ($2523) and IF ($2865). But so is the weight with I1 and SI2 weighing 1.284 carats respectively and 1.078 carats and IF and VVS1 weighing 0.505 carats and 0.503 carats respectively.

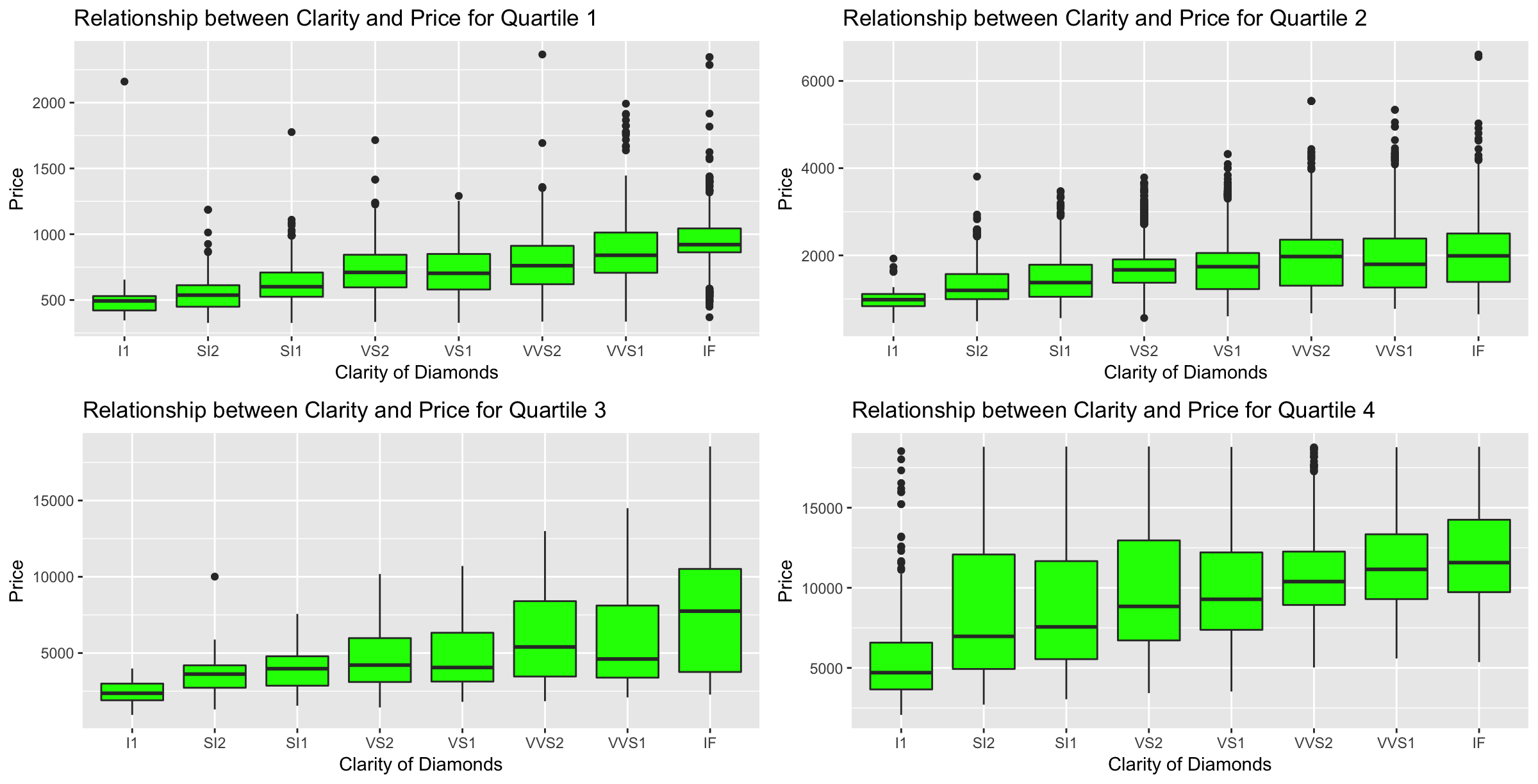
|  |  |  |  |
| --- | --- | --- | --- |
| *Table 4*: Mean Price, Carat and ratio(price/Carat) by Clarity | | | |
| Clarity | Mean Price $ | Mean Carat | Mean Price/Carat |
| I1 | 3924 | 1.284 | 2796 |
| SI2 | 5063 | 1.078 | 4011 |
| SI1 | 3996 | 0.851 | 3849 |
| VS2 | 3925 | 0.764 | 4081 |
| VS1 | 3839 | 0.727 | 4156 |
| VVS2 | 3284 | 0.596 | 4204 |
| VVS1 | 2523 | 0.503 | 3851 |
| IF | 2865 | 0.505 | 4260 |

When We compare the rations of price/carat for each clarity of diamonds we can see that trend reverses and a small increase in price with increase in clarity. *Figure 11* will show this more clearly where boxplot (a) (comparing price with clarity) is decreasing with grade whereas boxplot (b) (comparing ratio of price/carat) shows the true picture.



*Figure11:* Comparing price and clarity and Ratio of price/carat to clarity.

When we compare price and clarity for each quartile (*Figure 12 below)* we can see a strong relationship between grade and price. Price increases as we move up in grade for each weight class.



*Figure12:* relationship between price and clarity for each Quartile.

We can conclude that Clarity has a stronger relationship with price than cut and color but not as strong as carat.

**3.0 Regression**

We can use the four C variables can create a regression model to predict price of diamonds. Since carat had such a strong relationship with price we will produce 4 regression models one for each quartile. We will choose the model depending on the weight of diamonds.

The regression model will be in the form:

**= b0 + b1\*(Carat) + b2(Cut) + b3(Clarity) + b4(Color)**

Where the price will be estimated by adding the intercept coefficient of the model (b0) to product of weight of the diamond (carat) by its coefficient (b1) and coefficients of respective cut, color and clarity factor.

e.g. the coefficient for carat for regression model 1 (quartile 1 – 0.2carats to 0.4 carats) is $3076.99. it means that for each additional 0.1 carat increase in weight the price of diamond will increase by $307.70 when all other variables are statistically controlled.

As cut, clarity and color are ordinal-categorical variables with more than 2 factors, every factor of each variable will have a coefficient and will be added (or subtracted if negative) from the regression equation. The coefficient of the base grade i.e. Fair for cut, I1 for clarity and D for color will be 0 as its already computed in the intercept coefficient.

e.g. In regression model 1 (Q1) the cut grade in is Premium price will reduce by $20.19. If the clarity is VS1, $289.73 will be added in estimated price. If the color is D no amount will be added or subtracted from the estimated price.

**3.1 Regression Model 1** (Quartile 1 <= 0.4 carat)

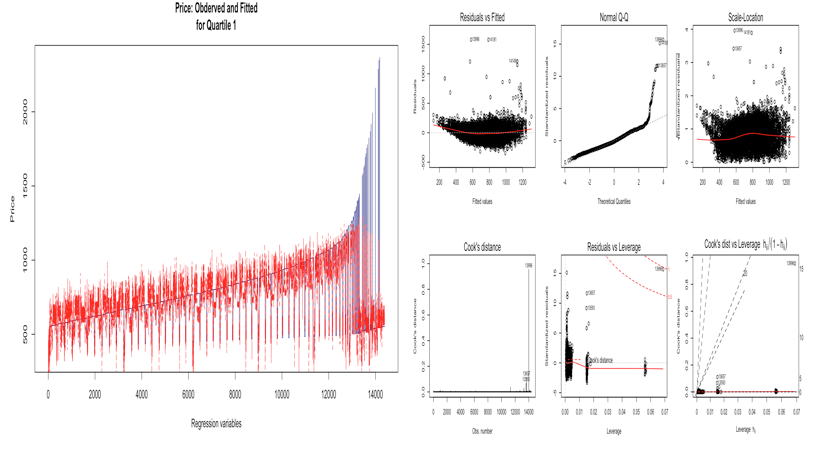
**1 = -**381.6 + Carat\*3076.99 +

Cut [0 (Fair), -120.60 (Good), -121.28 (Very Good), -20.19 (Premium), -55.22 (Ideal)] + Clarity [ 0(I1), 26.64 (SI2), 141.47 (SI1), 246.61 (VS2), 289.73 (VS1), 367.07 (VVS2), 444.49 (VVS1), 526.05(IF)] +

Color [0(D), -49.51(E), -84.06(F), -129.15(G), -196.86(H), -272.71(I), -349.24(J)]

Residual standard error: 104.4 on 14372 degrees of freedom. Multiple R-squared: 0.7504, Adjusted R-squared: 0.7501. F-statistic: 2400 on 18 and 14372 DF, p-value: < 2.2e-16

75.04% of variation in price can be explained by this regression model. The coefficient for Cut is decreasing as we move up on grade (it should be increasing), that because in *Figure 6* quartile 1 Fair cut had the highest mean price. We can see from *Figure 13* that towards the end of our scale there is considerable difference in actual price and price predicted by our model. This could be due to a long tail in the sample.



*Figure 13:* Observed vs Fitted value

**3.2 Regression Model 2** (Quartile 2 > .4 & <= 0.7 carat)

**2 = -**2431.57 + Carat\*5966.37 +

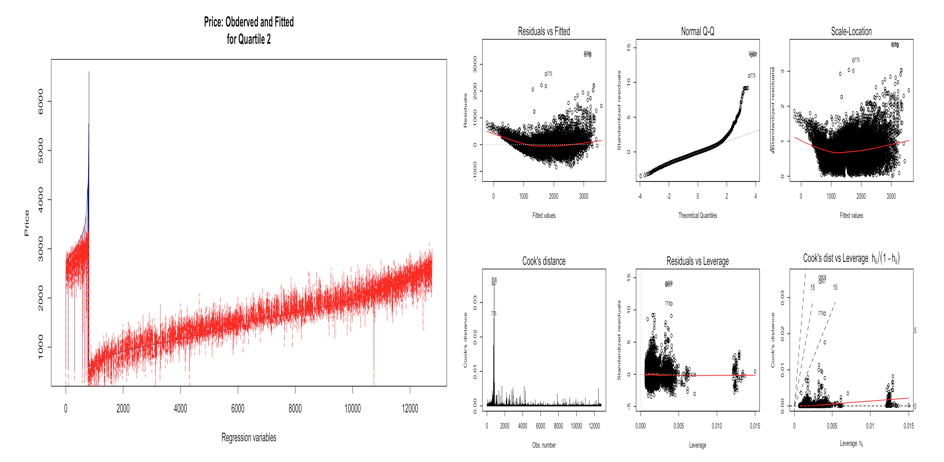
Cut [0 (Fair), 110.43 (Good), 157.50 (Very Good), 191.92 (Premium), 237.93 (Ideal)] +

Clarity [ 0(I1), 496.75 (SI2), 722.39 (SI1), 932.43 (VS2), 1091.10 (VS1), 1349.21 (VVS2), 1460.76 (VVS1), 1645.18 (IF)] +

Color [0(D), -134.70 (E), -223.61 (F), -348.04 (G), -481.06 (H), -690.63 (I), -784.33 (J)]

Residual standard error: 240.7 on 12752 degrees of freedom. Multiple R-squared: 0.8616, Adjusted R-squared: 0.8614. F-statistic: 4411 on 18 and 12752 DF, p-value: < 2.2e-16.

86.16% of variation in price can be explained by this regression model. There is lot of inconsistency in prediction price at the beginning of our model for Q2, like the one at the end of regression model for Q1.



*Figure 14:* Observed vs Fitted value

**3.3 Regression Model 3** (Quartile 3 > .7 & <= 1.04 carat)

**3 = -**7267.48 + Carat\*10,622.16 +

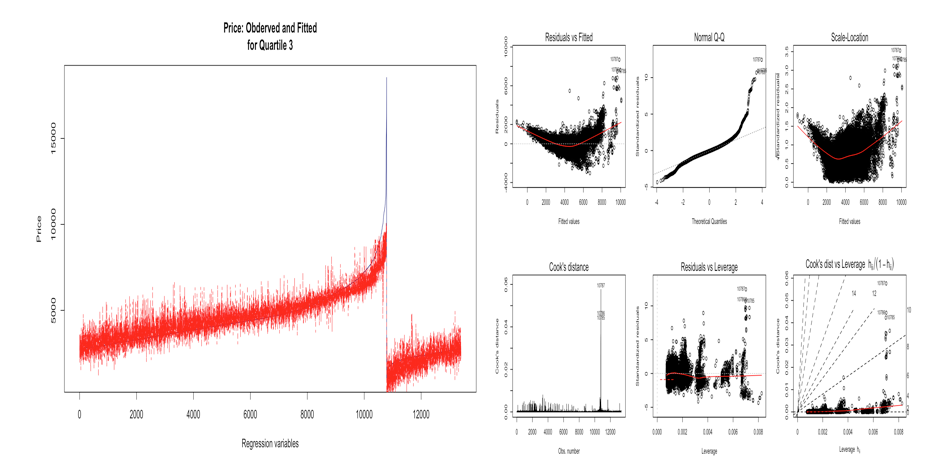
Cut [0 (Fair), 367.01 (Good), 638.26 (Very Good), 562.74 (Premium), 851.31 (Ideal)] +

Clarity [ 0(I1), 1074.43 (SI2), 1709.63 (SI1), 2400.20 (VS2), 2789.65 (VS1), 3793.36 (VVS2), 4030.56 (VVS1), 5429.10 (IF)] +

Color [0(D), -271.09 (E), -447.50 (F), -750.90 (G), -1181.41 (H), -1610.96 (I), -1866.17 (J)]

Residual standard error: 700.9 on 13380 degrees of freedom. Multiple R-squared: 0.8344, Adjusted R-squared: 0.8341 F-statistic: 3744 on 18 and 13380 DF, p-value: < 2.2e-16.

83.44% of variation in price can be explained by this regression model. The variance at the end of observed vs fitted model could be due to outliers.



*Figure 15:* Observed vs Fitted value

**3.4 Regression Model 4** (Quartile 4 > 1.04 carat)

**4 = -**11005.01 + Carat\*10406.86 +

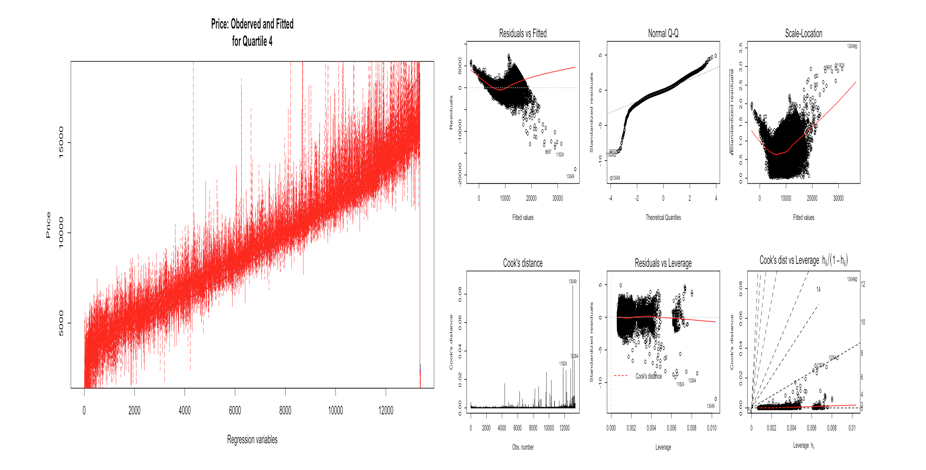
Cut [0 (Fair), 979.29 (Good), 1408.85 (Very Good), 1251.68 (Premium), 1724.89 (Ideal)] +

Clarity [ 0(I1), 3897.31 (SI2), 5438.30 (SI1), 6598.94 (VS2), 7218.21 (VS1), 8332.22 (VVS2), 9121.04 (VVS1), 9866.47 (IF)] +

Color [0(D), -191.96 (E), -467.71 (F), -1264.91 (G), -2184.84 (H), -3029.20 (I), -4446.42 (J)]

Residual standard error: 1502 on 13360 degrees of freedom. Multiple R-squared: 0.8617, Adjusted R-squared: 0.8615. F-statistic: 4623 on 18 and 13360 DF, p-value: < 2.2e-16

86.17% of variation in price can be explained by this regression model. We had a long tail for diamonds weighing from 1.04 carats to 5.1 carts and that could be the cause of Normal Q-Q plots inconsistency.



*Figure 16:* Observed vs Fitted value

**4. Conclusion**

we can derive the following conclusions from our research in diamond database:

* Carat has the strongest relationship with price. With a co-relation or = 0.92 and a positive, linear relationship of all the variables it has the most effect on price.
* Clarity has the second strongest relationship with price. When we compare the constancy of increase in mean price/carat ratio (*Table4*) as we go from lowest grade clarity to highest and compare this increase over the 4 quartiles in *figure 13* it only comes after carat.
* Color has the third strongest relationship with price of diamonds. *Figure 9* shows that prices for colors H, I and J are slightly lower than D, E, F at least in Q1, Q2 and Q3.
* Cut has the least strongest relationship with price. The mean of price/carat ration does not vary significantly across different cuts and mean price is consistent along all quartiles except Q1 where fair cut is the highest valued (also the one with highest coefficient in regression model 1) *Figure 6.*

Limitations to our analysis

* Small sample for certain factors within variables effect the accuracy of their relationship with Price. E.g. Fair cut has a small sample and shows inconstancy in regression moles 1. We have a small sample of diamonds with clarity I1 so its mean price/carat ration might not reflect the population.
* There’s considerable difference between our predicted price and actual price. And the normal Q-Q line is often skewed.
* We have not considered the relationship between the three categorical variables. If we consider relationship between Cut, Color and Clarity we might discover more pattern that will help predict the diamond prices more accurately.