Aseem Malhotra

101557400

STA 20010

Statistical Computing

Modelling Diamond prices

Table of Content

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

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 the 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 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 that will show how much each variable contribute to the price of the diamond.

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.

1. **Exploration**
   1. Carat

Diamonds are weighed in units of carat. I 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.

|  |
| --- |
| p <- ggplot(diamonds, aes(carat, color = clarity))    p + geom\_histogram(binwidth = .1, fill = "white") + ggtitle("Distribution of Carat of Diamonds", subtitle = "weight of diamonds") + ylab("Frequency") + xlab("Carat") |

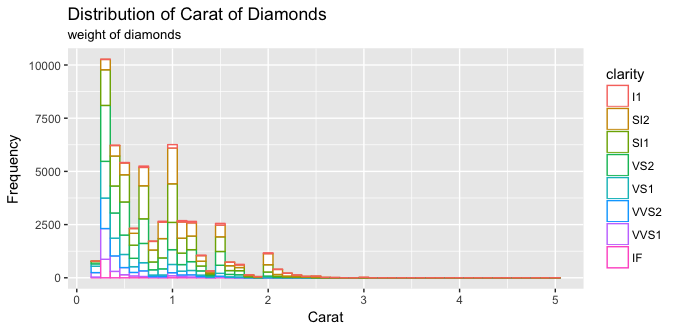
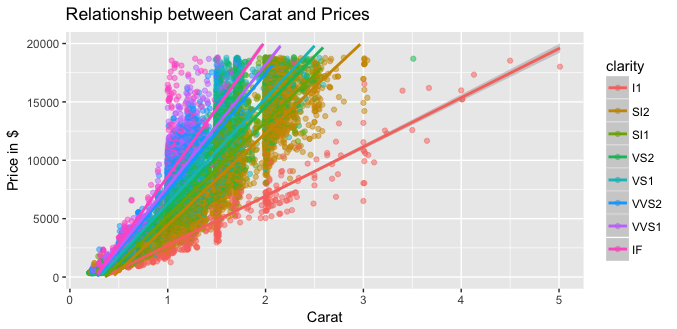


figure1: Distribution of Carat of diamonds.

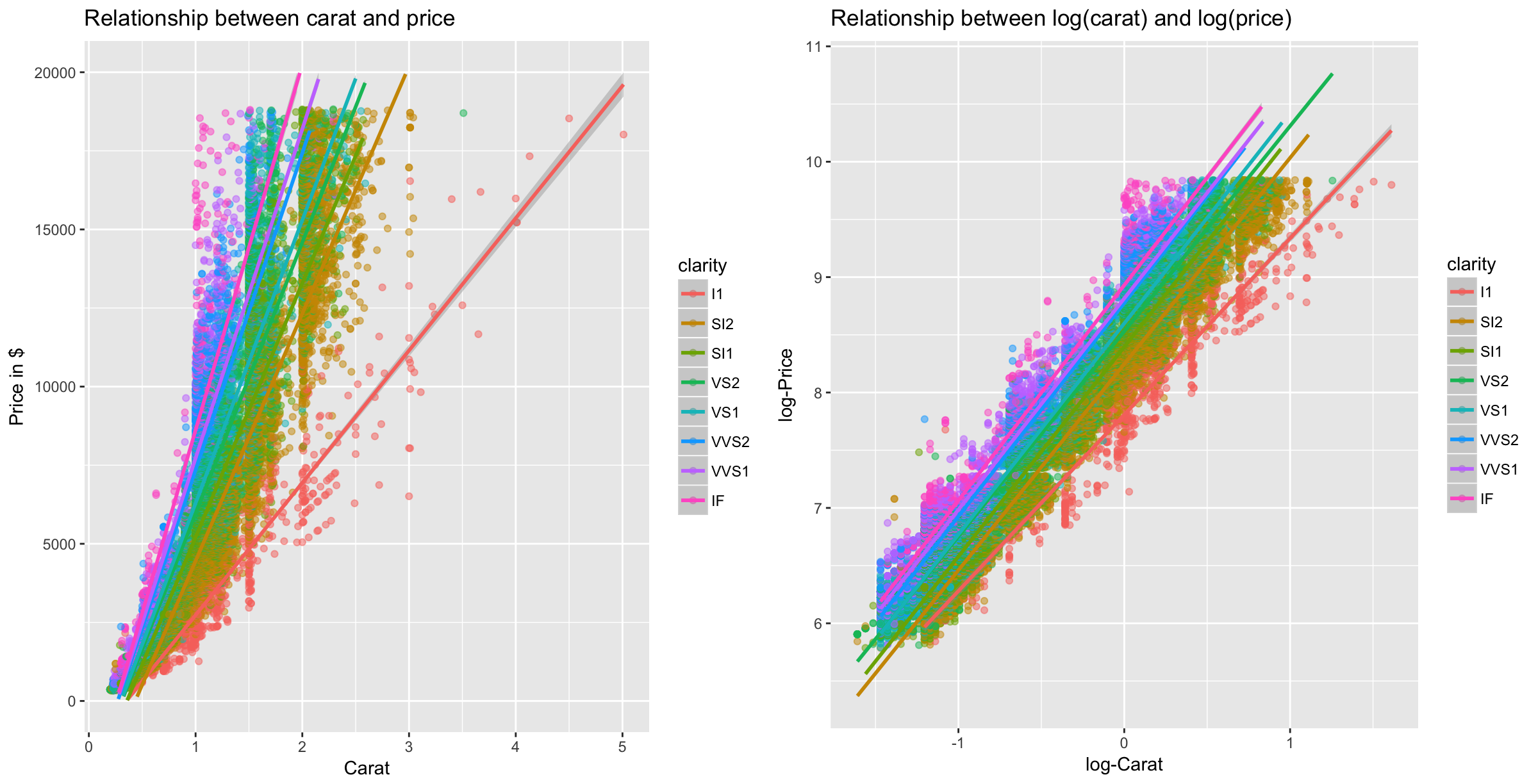
Relationship to price

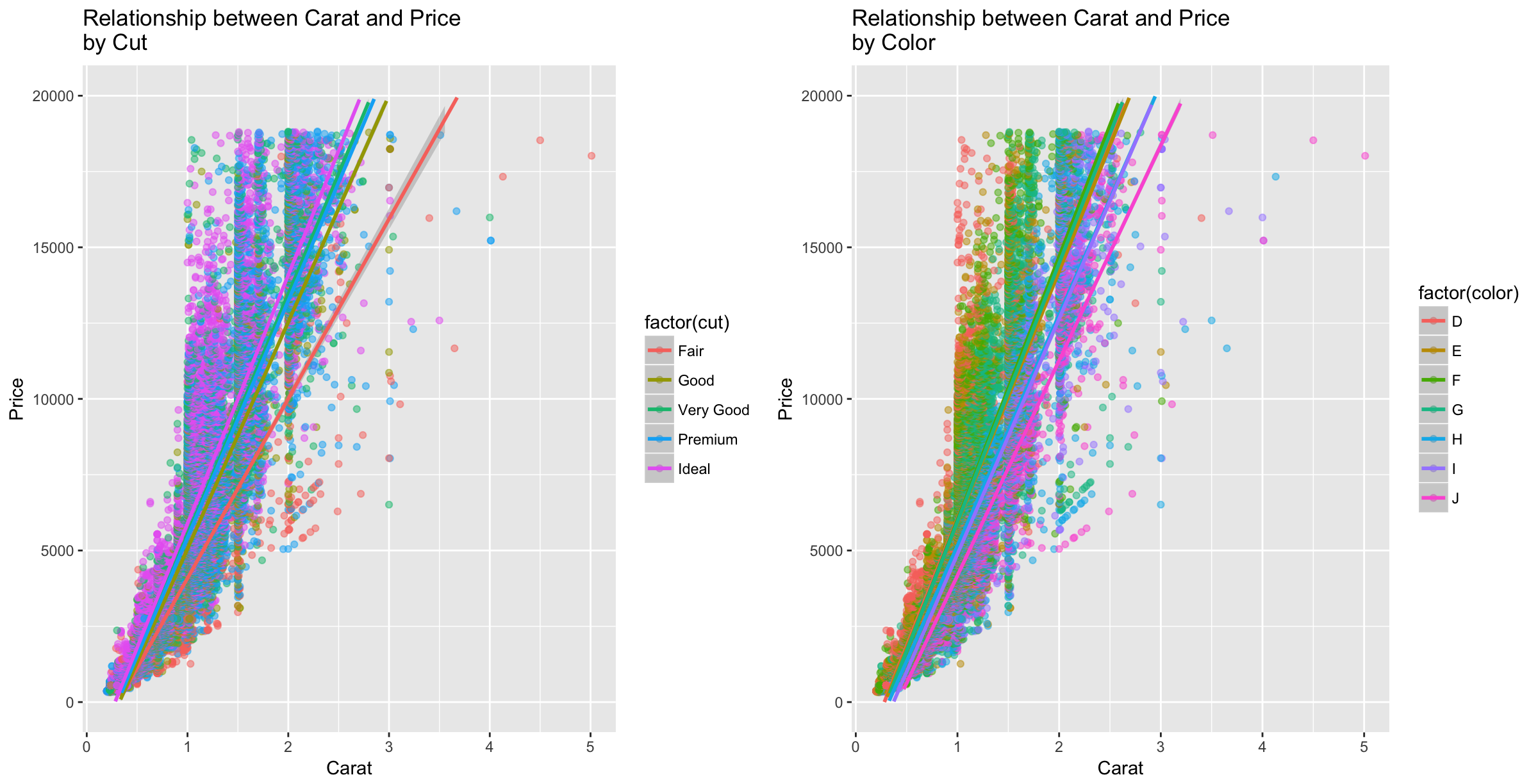
|  |
| --- |
| p <- ggplot(diamonds, aes(carat, price, color = clarity))  p + geom\_point(alpha = .5) + geom\_smooth(method = "lm") + ggtitle("Relationship between carat and price") +  xlab("Carat") + ylab("Price in $") + ylim(0, 20000) |

Carat has a strong, positive, and linear relationship with price with r = 0.92, and between log of both values is 0.97. We can be confident that carat has a huge impact on price.



|  |
| --- |
| p <- ggplot(diamonds, aes(log(carat), log(price), color = clarity))  z2 <- p + geom\_point(alpha = .5) + geom\_smooth(method = "lm") + ggtitle("Relationship between log(carat) and log(price)") +  + xlab("log-Carat") + ylab("log-Price")  grid.arrange(z1, z2, ncol=2) |

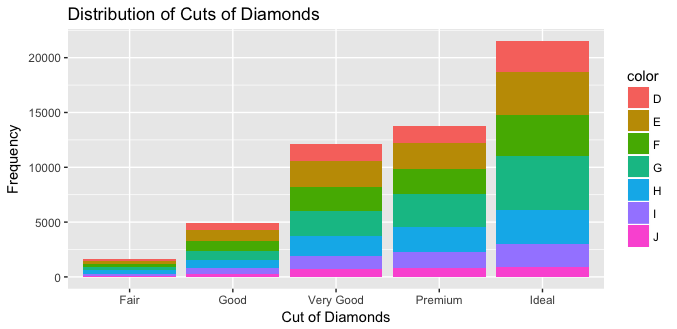




2.2 Cut

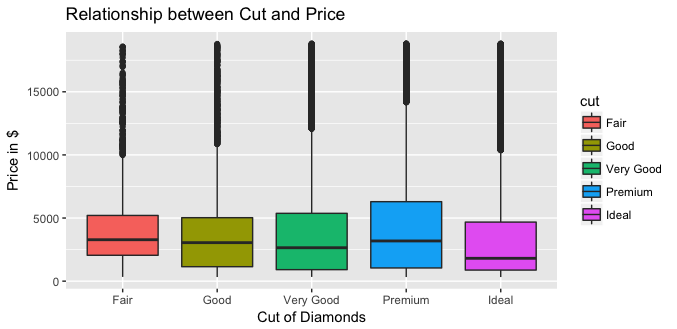
the cut of the diamonds can vary in 5 different way from fair (lowest grade) to Ideal (highest grade).

|  |
| --- |
| p <- ggplot(diamonds, aes(factor(cut), fill = color))  p + geom\_bar() + ggtitle("Distribution of Cuts of Diamonds") + xlab("Cut of Diamonds") + ylab("Frequency") |



We expected the grade to reflect in the prices but it does not reflect in the figure below. Premium graded cuts has a the highest mean Price of $4584 followed by Fair cut $4359. And Ideal cut has the lowest mean price of $ 3458.

|  |
| --- |
| p <- ggplot(diamonds, aes(factor(cut), price, fill= cut ))  p + geom\_boxplot() + ggtitle("Relationship between Cut and Price") + xlab("Cut of Diamonds") + ylab("Price in $") |

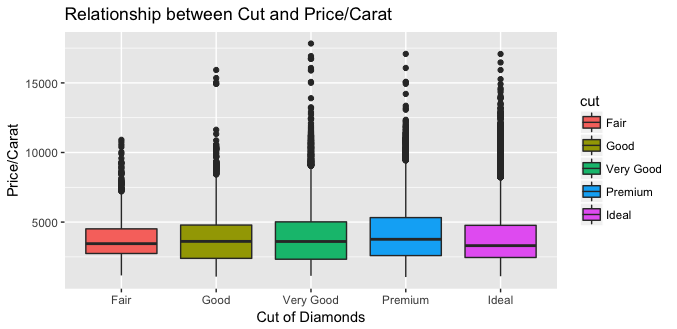


|  |
| --- |
| by(price, cut, summary)  by(carat, carat, summary)  by(price/carat, cut, summary) |

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

We will calculate the relationship between price with (Price/Carat) in figure below and we can see a more even boxplot.

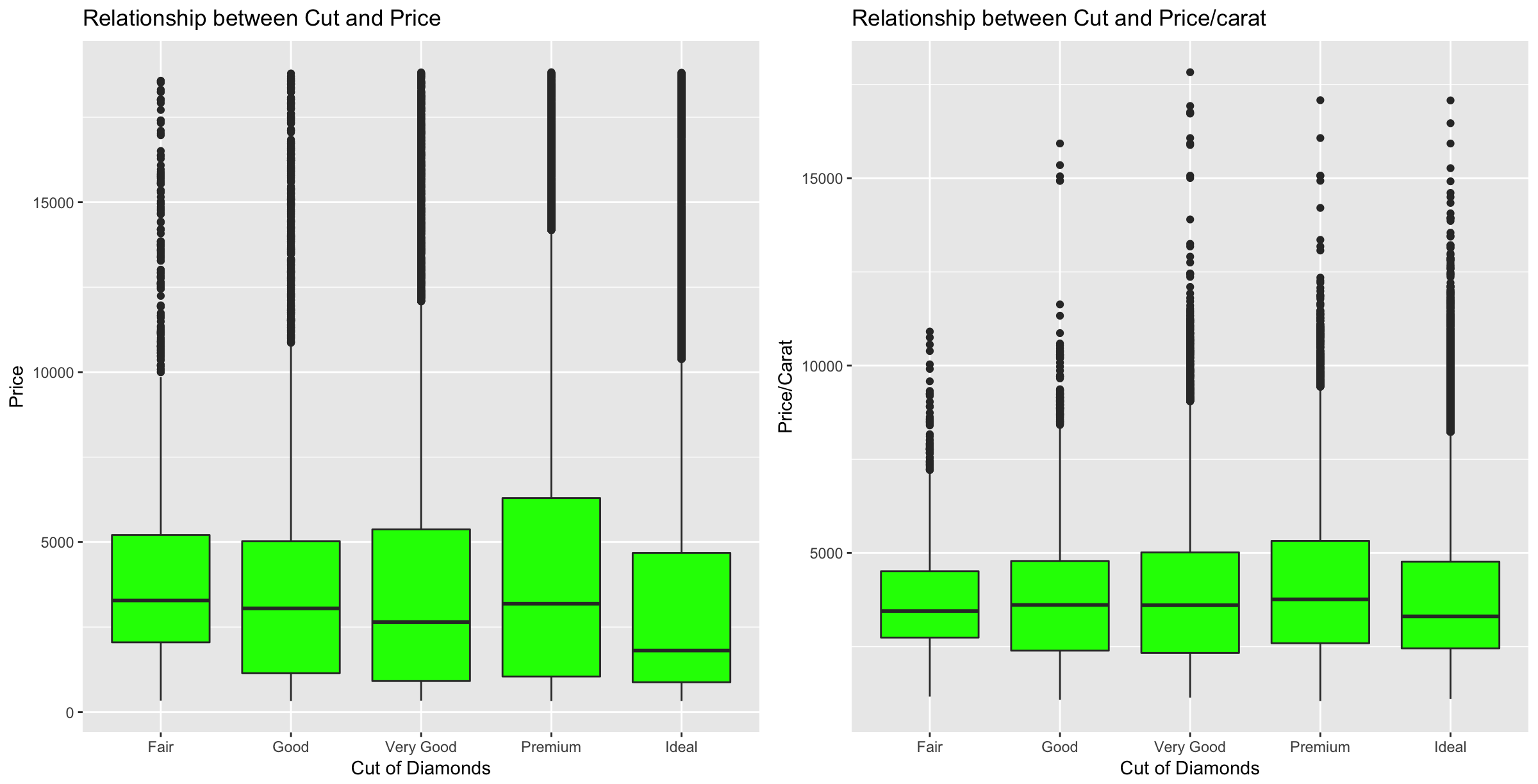
|  |
| --- |
| p <- ggplot(diamonds, aes(factor(cut), (price/carat), fill= cut ))  p + geom\_boxplot() + ggtitle("Relationship between Cut and Price/Carat") + xlab("Cut of Diamonds") + ylab("Price/Carat") |



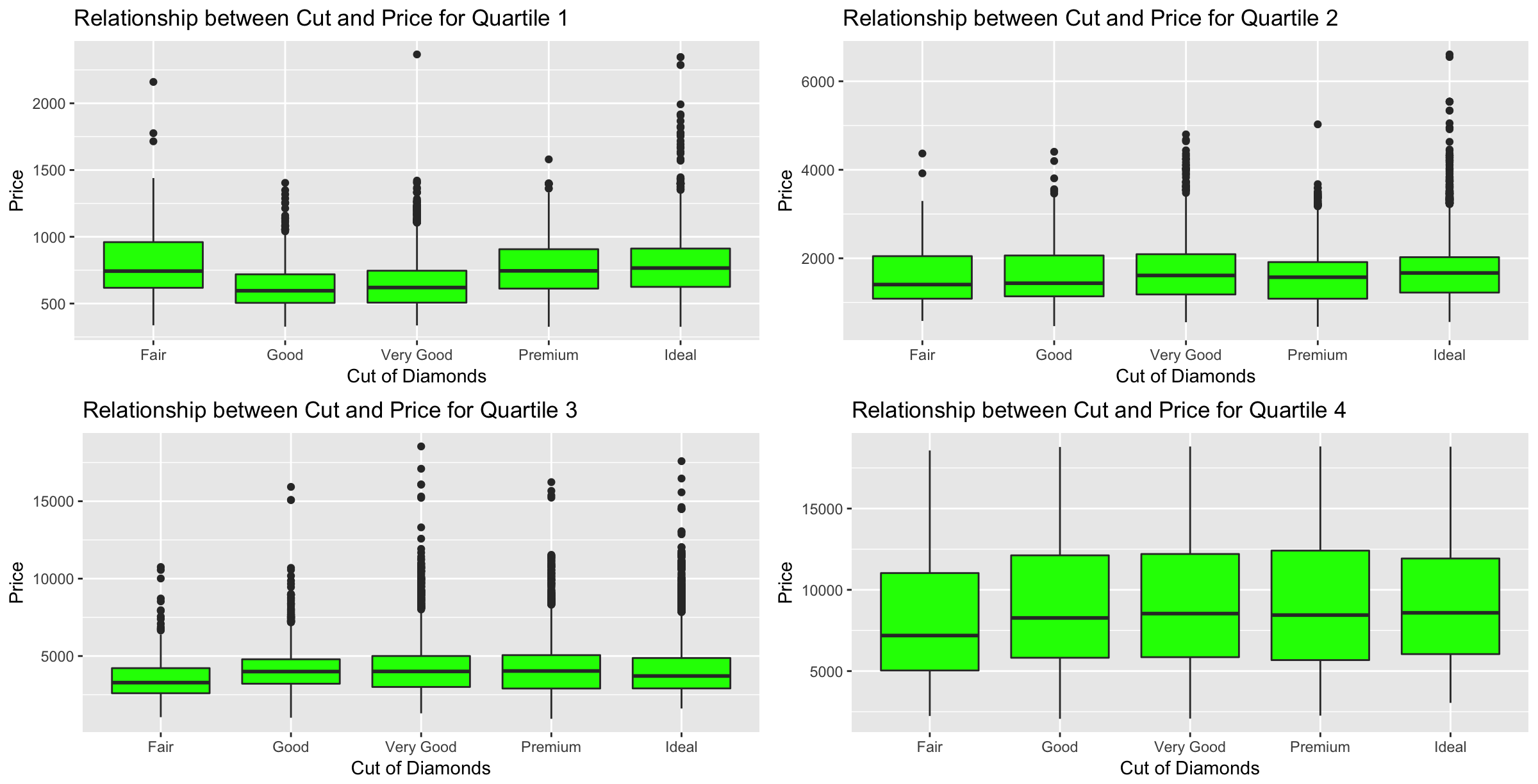
the effect of cut on price if far less than anticipated.

grid.arrange(p1, p2, ncol = 2)

|  |
| --- |
| p1 <- ggplot(diamonds, aes(factor(cut), price))  p1 <- p1 + geom\_boxplot(fill = "green") + ggtitle("Relationship between Cut and Price") + xlab("Cut of Diamonds") + ylab("Price")  p2 <- ggplot(diamonds, aes(factor(cut), price/carat))  p2 <- p2 + geom\_boxplot(fill = "green") + ggtitle("Relationship between Cut and Price/carat") + xlab("Cut of Diamonds") + ylab("Price/Carat")  grid.arrange(p1, p2, ncol = 2) |



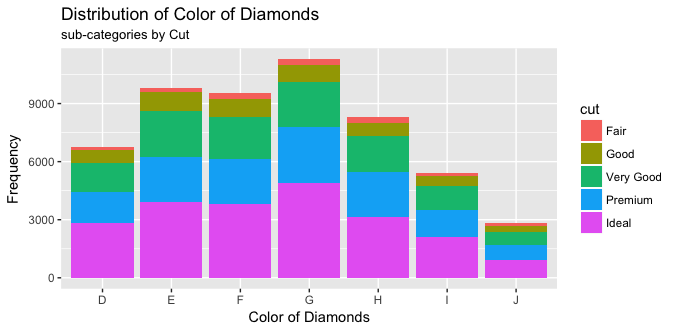
|  |
| --- |
| p1 <- ggplot(q1, aes(factor(cut), (price)))  a1 <- p1 + geom\_boxplot(fill = "green") + ggtitle("Relationship between Cut and Price for Quartile 1") + xlab("Cut of Diamonds") + ylab("Price")  grid.arrange(a1, a2, a3, a4, ncol=2) |



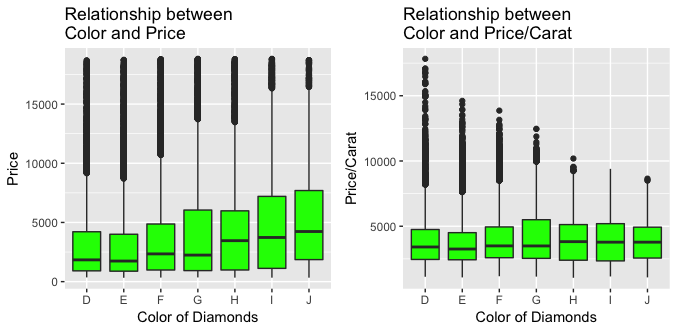
* 1. Color

The color of a 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.

|  |
| --- |
| > p <- ggplot(diamonds, aes(factor(color), fill = cut))  > p + geom\_bar() + ggtitle("Distribution of Color of Diamonds", subtitle = "sub-categories by Cut") + xlab("Color of Diamonds") + ylab("Frequency") |



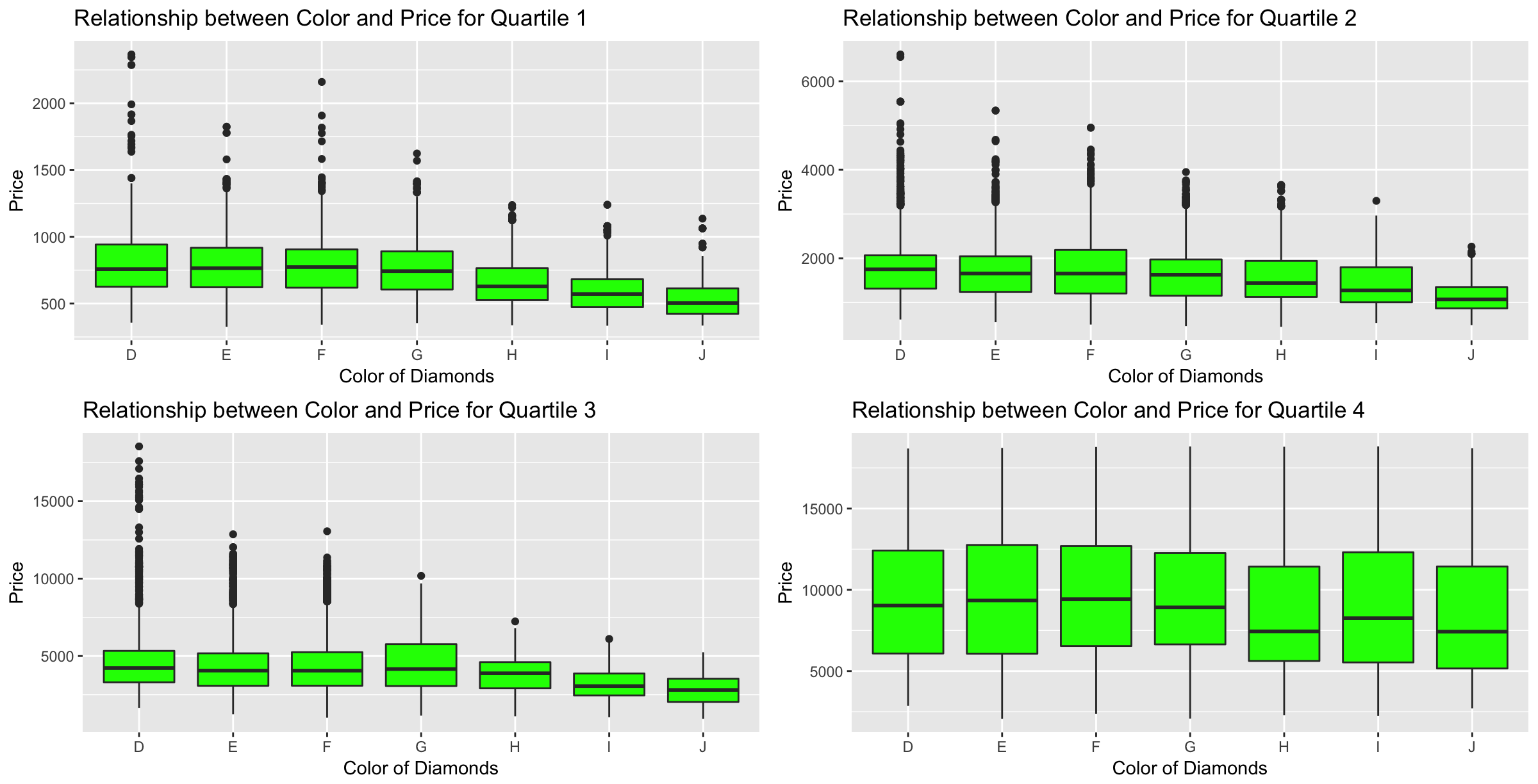
|  |
| --- |
| p2 <- ggplot(diamonds, aes(factor(color), price/carat))  p2 <- p2 + geom\_boxplot(fill = "green") + ggtitle("Relationship between\nColor and Price/Carat") + xlab("Color of Diamonds") + ylab("Price/Carat")  p1 <- ggplot(diamonds, aes(factor(color), price))  p1 <- p1 + geom\_boxplot(fill = "green") + ggtitle("Relationship between\nColor and Price") + xlab("Color of Diamonds") + ylab("Price")  grid.arrange(p1, p2, ncol = 2) |



|  |
| --- |
| by(price, color, summary)  by(carat, color, summary)  by(price/carat, color, summary) |

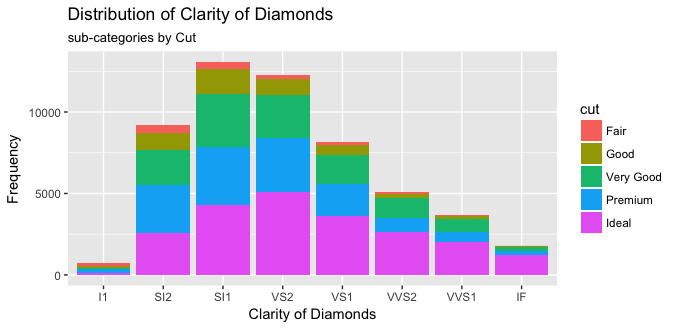
|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| Color | Mean Price ($) | Mean Carat | Mean Price/Carat |
| D | 3170 | 0.658 | 3953 |
| E | 3077 | 0.658 | 3805 |
| F | 3725 | 0.737 | 4135 |
| G | 3999 | 0.771 | 4163 |
| H | 4487 | 0.912 | 4008 |
| I | 5092 | 1.027 | 3996 |
| J | 5324 | 1.162 |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

|  |
| --- |
| a4 <- p1 + geom\_boxplot(fill = "green") + ggtitle("Relationship between Color and Price for Quartile 4") + xlab("Color of Diamonds") + ylab("Price")  > grid.arrange(a1, a2, a3, a4, ncol=2) |

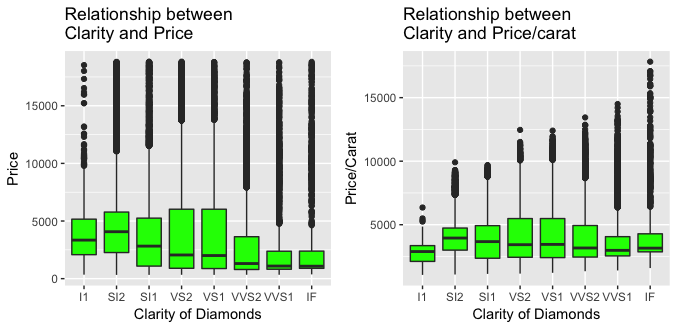


* 1. Clarity

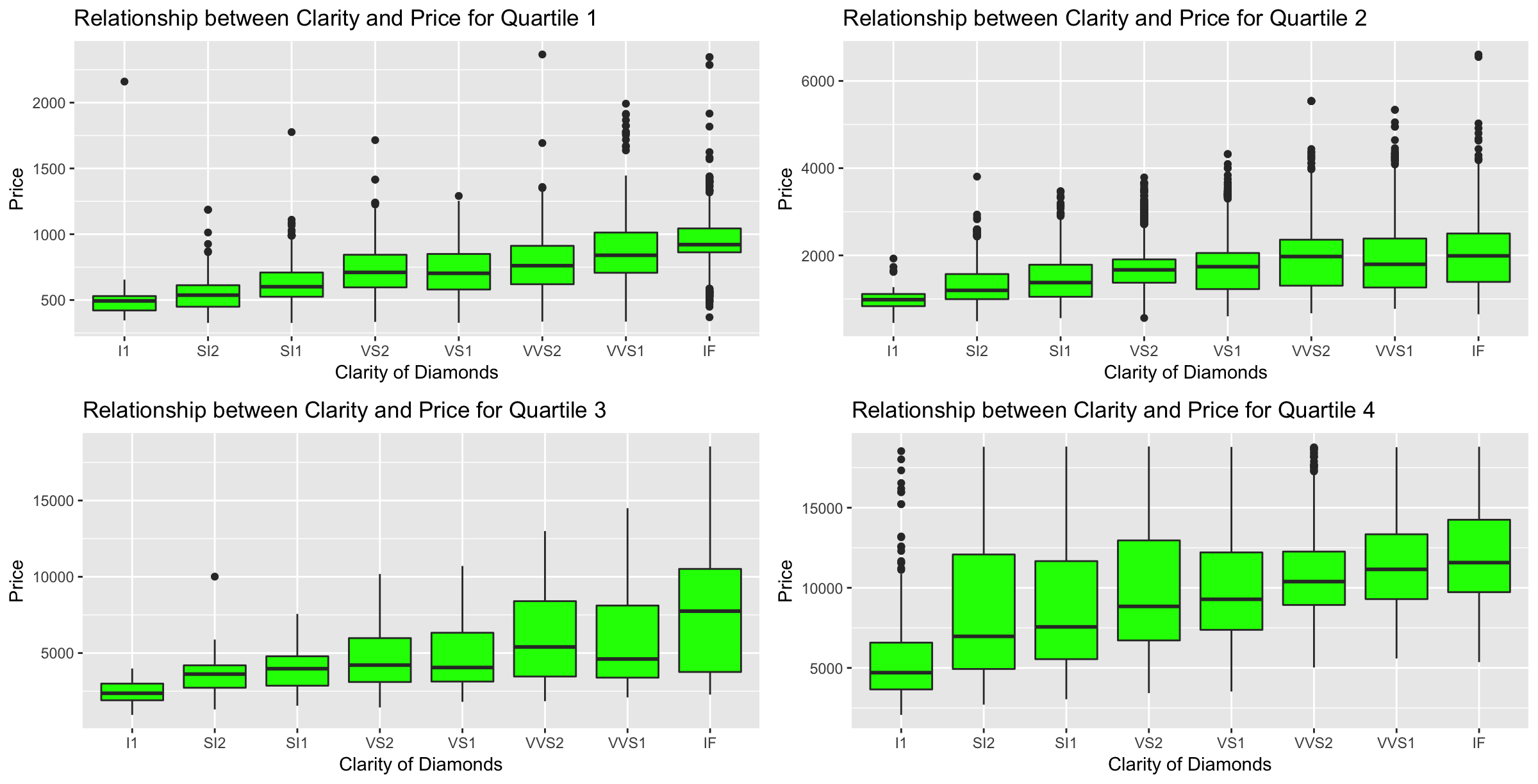
|  |
| --- |
| > p <- ggplot(diamonds, aes(factor(clarity), fill = cut))  > p + geom\_bar() + ggtitle("Distribution of Clarity of Diamonds", subtitle = "sub-categories by Cut") + xlab("Clarity of Diamonds") + ylab("Frequency") |



|  |
| --- |
| p1 <- ggplot(diamonds, aes(factor(clarity), price))  p1 <- p1 + geom\_boxplot(fill = "green") + ggtitle("Relationship between\nClarity and Price") + xlab("Clarity of Diamonds") + ylab("Price")  p2 <- ggplot(diamonds, aes(factor(clarity), price/carat))  p2 <- p2 + geom\_boxplot(fill = "green") + ggtitle("Relationship between\nClarity and Price/carat") + xlab("Clarity of Diamonds") + ylab("Price/Carat")  grid.arrange(p1, p2, ncol = 2) |



|  |
| --- |
| > p1 <- ggplot(q4, aes(factor(clarity), (price)))  > a4 <- p1 + geom\_boxplot(fill = "green") + ggtitle("Relationship between Clarity and Price for Quartile 4") + xlab("Clarity of Diamonds") + ylab("Price")  > grid.arrange(a1, a2, a3, a4, ncol=2) |



|  |
| --- |
| by(price, diamonds$clarity, summary)  by(carat, diamonds$clarity, summary)  by(price/carat, diamonds$clarity, summary) |

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