BIS581

The following uses the diamonds dataset from the GGPLOT2 library

library(ggplot2)  
mydiamonds <- diamonds

You can paste in a new “code chunk” by pressing CTRL- ALT - I so for each question listed below, put in a new code chunk to answer that question. Knit your notebook to Word and submit the word document AND the .RMD file on blackboard.

How many records are in the dataset?

nrow(mydiamonds)

## [1] 53940

What is the largest diamond by weight (carat)?

max(mydiamonds$carat)

## [1] 5.01

Most and least expensive?

max(mydiamonds$price)

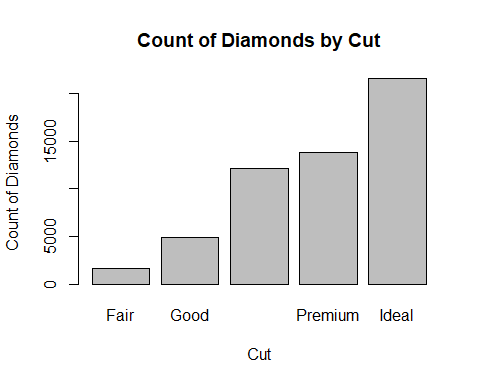
## [1] 18823

min(mydiamonds$price)

## [1] 326

Plot a bar chart of count of diamonds vs cut.

cut\_counts <- table(mydiamonds$cut)  
barplot(cut\_counts,main = "Count of Diamonds by Cut",xlab = "Cut", ylab = "Count of Diamonds")



Let’s explore the data a bit. What attributes does the most expensive diamond have? Change max(price) to min(price) to see the least expensive.

subset(diamonds, price == max(price))

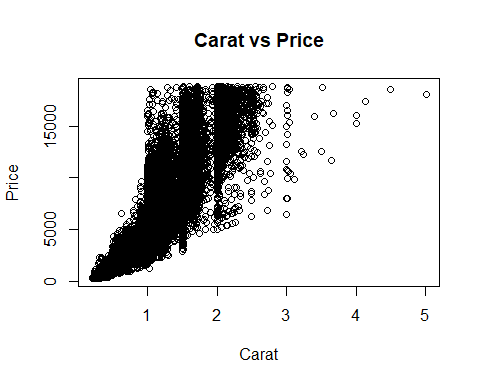
## # A tibble: 1 × 10  
## carat cut color clarity depth table price x y z  
## <dbl> <ord> <ord> <ord> <dbl> <dbl> <int> <dbl> <dbl> <dbl>  
## 1 2.29 Premium I VS2 60.8 60 18823 8.5 8.47 5.16

subset(mydiamonds, price == min(price))

## # A tibble: 2 × 10  
## carat cut color clarity depth table price x y z  
## <dbl> <ord> <ord> <ord> <dbl> <dbl> <int> <dbl> <dbl> <dbl>  
## 1 0.23 Ideal E SI2 61.5 55 326 3.95 3.98 2.43  
## 2 0.21 Premium E SI1 59.8 61 326 3.89 3.84 2.31

Create a plot of carat vs price.

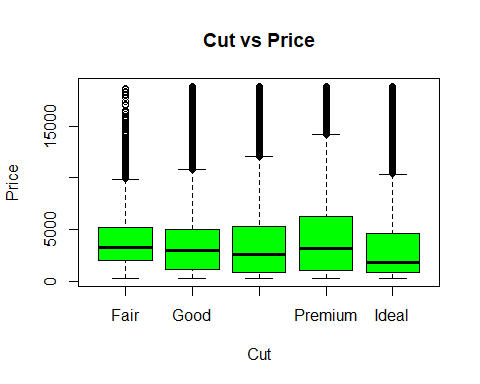
plot(mydiamonds$carat, mydiamonds$price, main = "Carat vs Price",xlab = "Carat",ylab = "Price")



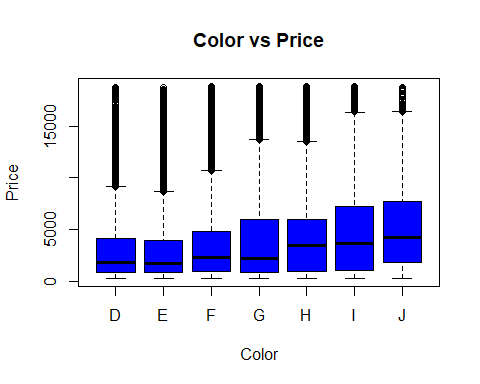
Does it look like carat and price have a linear relationship? In the diamond dataset, there is a significant relationship between carat and price, with larger diamonds often costing more. A scatter plot of carat compared to price shows a rising trend, indicating that as the carat value grows, so does the price. However, the connection may not be completely linear because additional factors like as cut, clarity, and color influence the price. To further understand this link, we may employ techniques such as analysis of variance or linear regression.

Create three other plots of other variables vs price. The point of exploratory analysis (know your data) is to do just that, explore. You might have to plot more than three to find variables that plot correctly. Please realize too that scatter plots (or line) are for continuous variables and not for categorical variables. See the ggplot2 intro for references. Please try to pick three variables that you think have a strong influence in the price of the diamond. The main point for this is to make a model later on.

plot(mydiamonds$price ~ mydiamonds$cut,main = "Cut vs Price",xlab = "Cut",ylab = "Price",col = "green")



plot(mydiamonds$price ~ mydiamonds$color,main = "Color vs Price",xlab = "Color",ylab = "Price",col = "blue")



plot(mydiamonds$price ~ mydiamonds$clarity,main = "Clarity vs Price",xlab = "Clarity",ylab = "Price",col = "red")

