BIS581

The following uses the diamonds dataset from the GGPLOT2 library AUTHOR: GREESHMA CHOWDARY PEDDINENI DATE: 01/19/2025

library(ggplot2)  
mydiamonds <- diamonds

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.

barplot(table(diamonds$cut), main = "Count of Diamonds by Cut", xlab = "Cut", ylab = "Count", col = "Green")

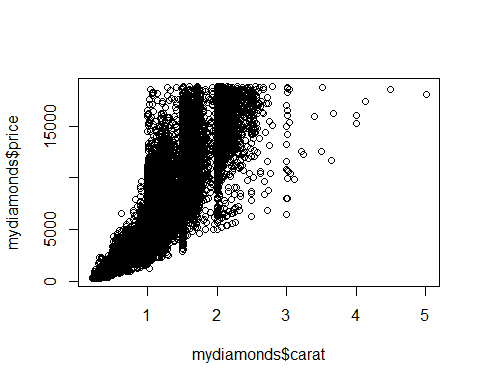
Let’s explore the data a bit. What attributes does the most expensive diamond have? Change max(price) to min(price) to see the least expensive.

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

Create a plot of carat vs price.

plot(mydiamonds$carat, mydiamonds$price)

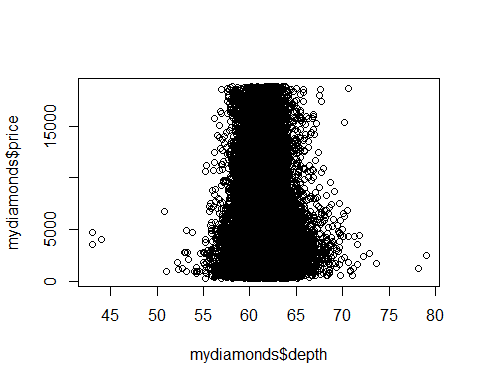


Does it look like carat and price have a linear relationship?

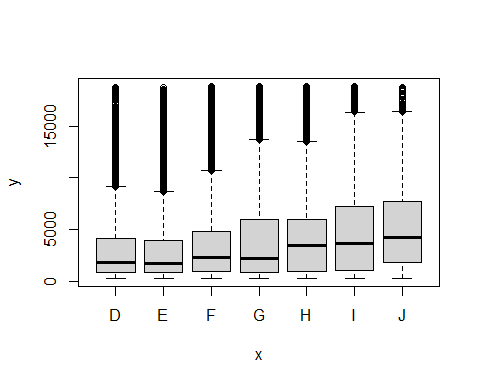
Answer: The carat and price plot doesn’t look like a linear relationship. Because we cannot split the data into half. But it looks like exponential, we can make it into half.

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.

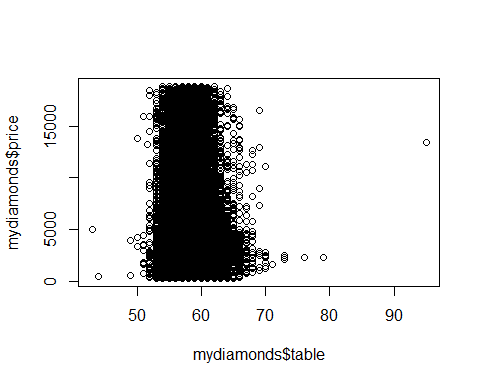
plot(mydiamonds$depth, mydiamonds$price)



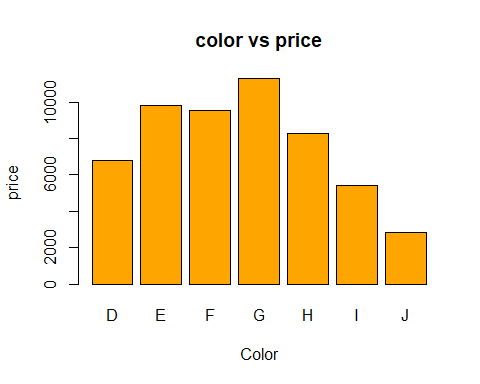
plot(mydiamonds$color, mydiamonds$price)



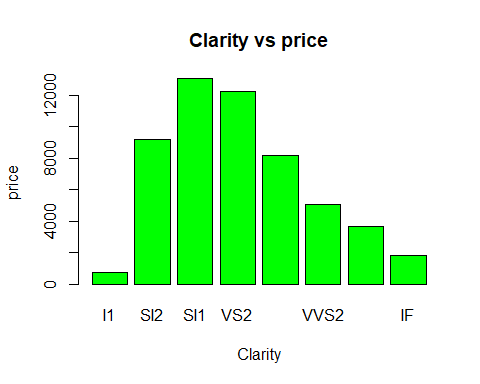
plot(mydiamonds$table, mydiamonds$price)



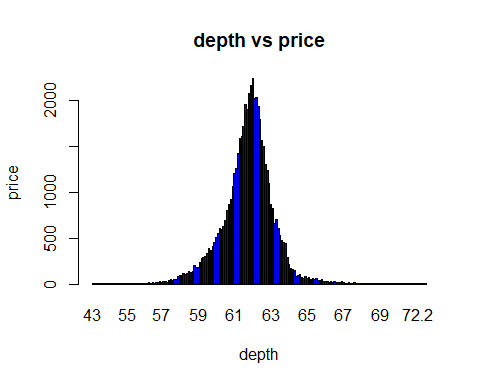
barplot(table(diamonds$color), main = "color vs price", xlab = "Color", ylab = "price", col = "Orange")



barplot(table(diamonds$clarity), main = "Clarity vs price", xlab = "Clarity", ylab = "price", col = "Green")



barplot(table(diamonds$depth), main = "depth vs price", xlab = "depth", ylab = "price", col = "Blue")



By the bar charts we can conclude that which one is highest and lowest. But from plots we cannot able to find anything because the variables are categorical. From my analysis we can say that bar charts shows exact information that what we want.I kept both plots here to compare that which one’s are better.