BIS581

library(ggplot2)  
mydiamond <- diamonds

#Summary data

summary(mydiamond)

## carat cut color clarity depth   
## Min. :0.2000 Fair : 1610 D: 6775 SI1 :13065 Min. :43.00   
## 1st Qu.:0.4000 Good : 4906 E: 9797 VS2 :12258 1st Qu.:61.00   
## Median :0.7000 Very Good:12082 F: 9542 SI2 : 9194 Median :61.80   
## Mean :0.7979 Premium :13791 G:11292 VS1 : 8171 Mean :61.75   
## 3rd Qu.:1.0400 Ideal :21551 H: 8304 VVS2 : 5066 3rd Qu.:62.50   
## Max. :5.0100 I: 5422 VVS1 : 3655 Max. :79.00   
## J: 2808 (Other): 2531   
## table price x y   
## Min. :43.00 Min. : 326 Min. : 0.000 Min. : 0.000   
## 1st Qu.:56.00 1st Qu.: 950 1st Qu.: 4.710 1st Qu.: 4.720   
## Median :57.00 Median : 2401 Median : 5.700 Median : 5.710   
## Mean :57.46 Mean : 3933 Mean : 5.731 Mean : 5.735   
## 3rd Qu.:59.00 3rd Qu.: 5324 3rd Qu.: 6.540 3rd Qu.: 6.540   
## Max. :95.00 Max. :18823 Max. :10.740 Max. :58.900   
##   
## z   
## Min. : 0.000   
## 1st Qu.: 2.910   
## Median : 3.530   
## Mean : 3.539   
## 3rd Qu.: 4.040   
## Max. :31.800   
##

#Get structures

str(mydiamond)

## tibble [53,940 × 10] (S3: tbl\_df/tbl/data.frame)  
## $ carat : num [1:53940] 0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...  
## $ cut : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 ...  
## $ color : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...  
## $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 5 ...  
## $ depth : num [1:53940] 61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...  
## $ table : num [1:53940] 55 61 65 58 58 57 57 55 61 61 ...  
## $ price : int [1:53940] 326 326 327 334 335 336 336 337 337 338 ...  
## $ x : num [1:53940] 3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...  
## $ y : num [1:53940] 3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...  
## $ z : num [1:53940] 2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...

How many records are in the dataset?

cat("Number of rows:", dim(mydiamond)[1])

## Number of rows: 53940

What is the largest diamond by weight (carat)?

cat("The largest diamond by weight is: ", max(mydiamond$carat))

## The largest diamond by weight is: 5.01

Most and least expensive?

cat("The most expensive diamond is: ", max(mydiamond$price))

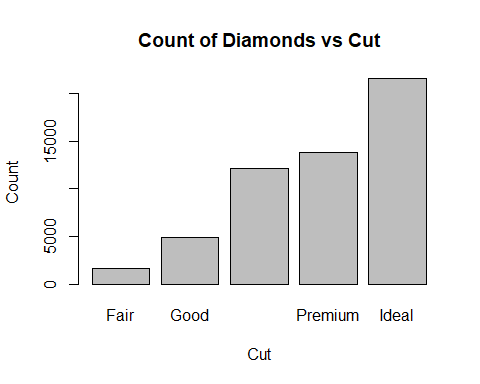
## The most expensive diamond is: 18823

cat("\nThe least expensive diamond is: ", min(mydiamond$price))

##   
## The least expensive diamond is: 326

Plot a bar chart of count of diamonds vs cut. #xlab is label for x axis, ylab is label for y axis, main is the title.

barplot(table(mydiamond$cut), main = "Count of Diamonds vs Cut", xlab = "Cut", ylab = "Count")



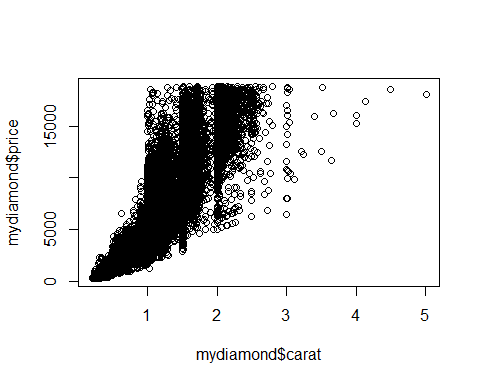
Let’s explore the data a bit. What attributes does the most expensive diamond have? Change max(price) to min(price) to see the least expensive.

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(mydiamond$carat, mydiamond$price)



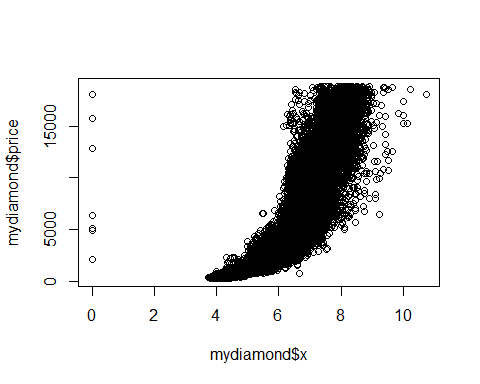
Does it look like carat and price have a linear relationship?

Yes, I would say that carat and price do have a positive linear relationship along with a linear trendline. This plot shows me as the carat size increases the price also increases.

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.

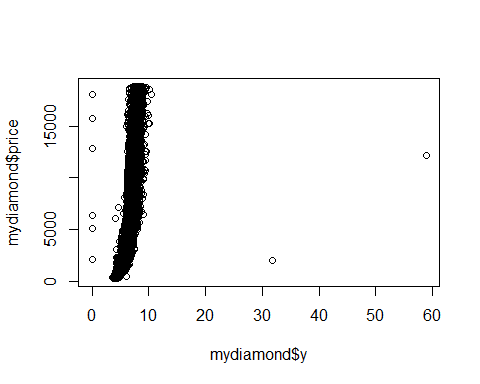
Plot of diamond length vs price

plot(mydiamond$x, mydiamond$price)



Plot of diamond width vs price

plot(mydiamond$y, mydiamond$price)



Plot of diamond width vs price

plot(mydiamond$z, mydiamond$price)

