hw 7

Jacqueline Nguyen

2024-06-12

1

```
#a)
data <- read.csv("births.csv", stringsAsFactors = TRUE)
head(data)
```

```
##   Gender Premie weight Apgar1 Fage Mage Feduc Meduc TotPreg Visits   Marital
## 1   Male     No    124      8   31   25    13    14       1     13   Married
## 2 Female     No    177      8   36   26     9    12       2     11 Unmarried
## 3   Male     No    107      3   30   16    12     8       2     10 Unmarried
## 4 Female     No    144      6   33   37    12    14       2     12 Unmarried
## 5   Male     No    117      9   36   33    10    16       2     19   Married
## 6 Female     No     98      4   31   29    14    16       3     20   Married
##   Racemom Racedad Hispmom Hispdad Gained     Habit MomPriorCond BirthDef
## 1   White   White NotHisp NotHisp     40 NonSmoker         None     None
## 2   White   White Mexican Mexican     20 NonSmoker         None     None
## 3   White Unknown Mexican Unknown     70 NonSmoker At Least One     None
## 4   White   White NotHisp NotHisp     50 NonSmoker         None     None
## 5   White   Black NotHisp NotHisp     40 NonSmoker At Least One     None
## 6   White   White NotHisp NotHisp     21 NonSmoker         None     None
##      DelivComp BirthComp
## 1 At Least One      None
## 2 At Least One      None
## 3 At Least One      None
## 4 At Least One      None
## 5         None      None
## 6         None      None
```

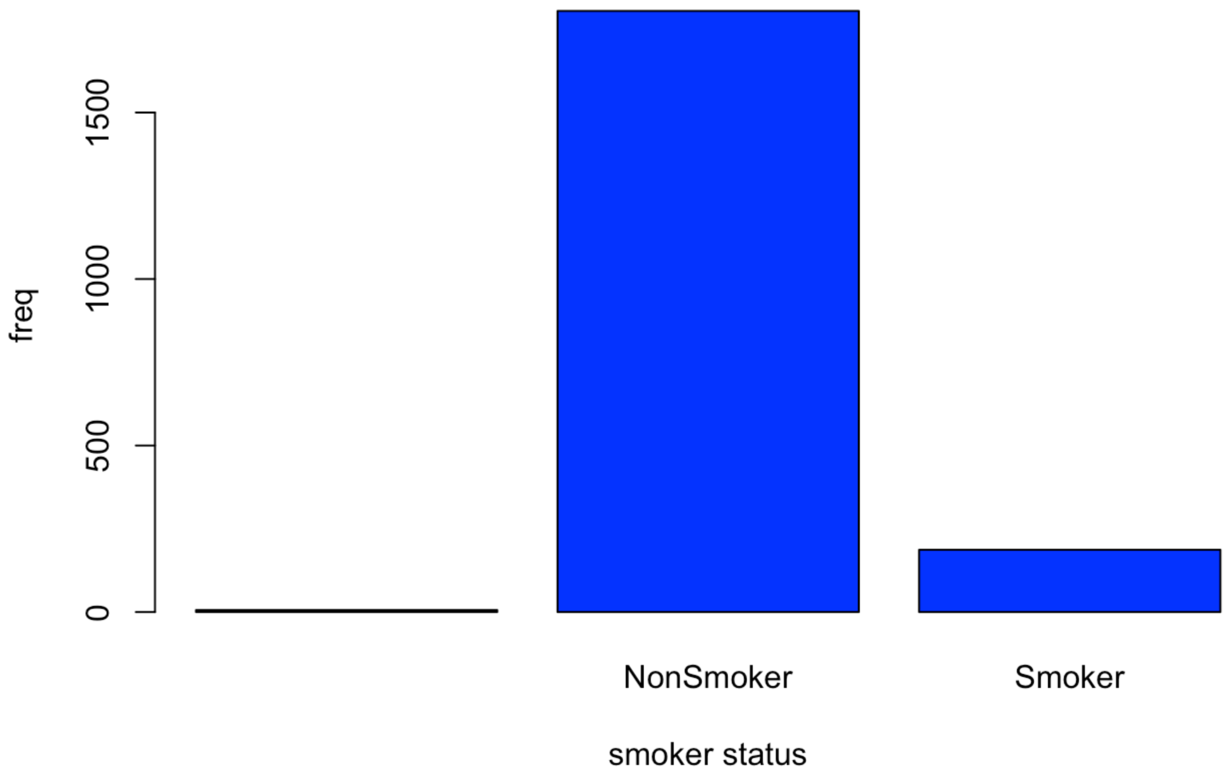
```
#b)
sum(data$Habit == "")
```

```
## [1] 6
```

```
levels(data$Habit)
```

```
## [1] ""          "NonSmoker" "Smoker"
```

```
# "" and 6 observations

#c)
barplot(table(data$Habit), col = "blue", xlab = "smoker status", ylab = "freq")
```



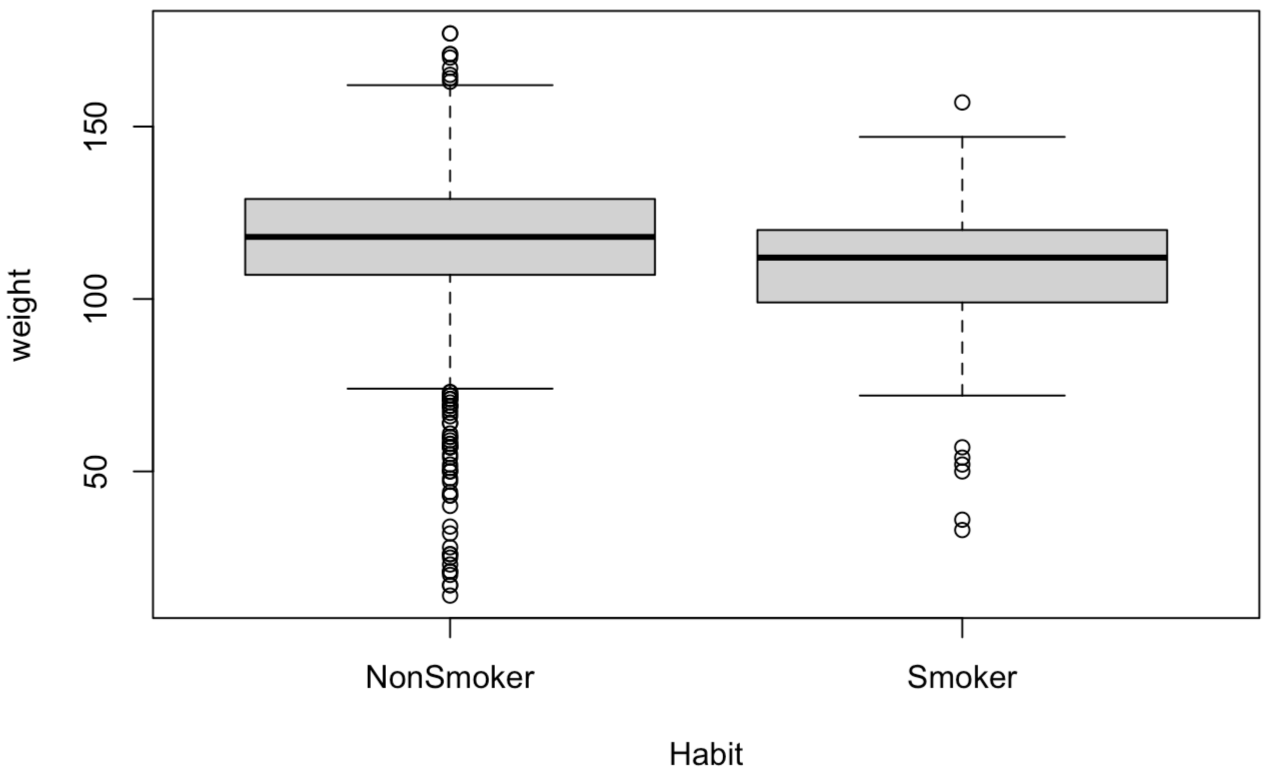
```
#d)
 data_habit_known <- droplevels(subset(data, Habit != ""))
 sum(data_habit_known$Habit == "")
```

```
## [1] 0
```

```
 levels(data_habit_known$Habit)
```

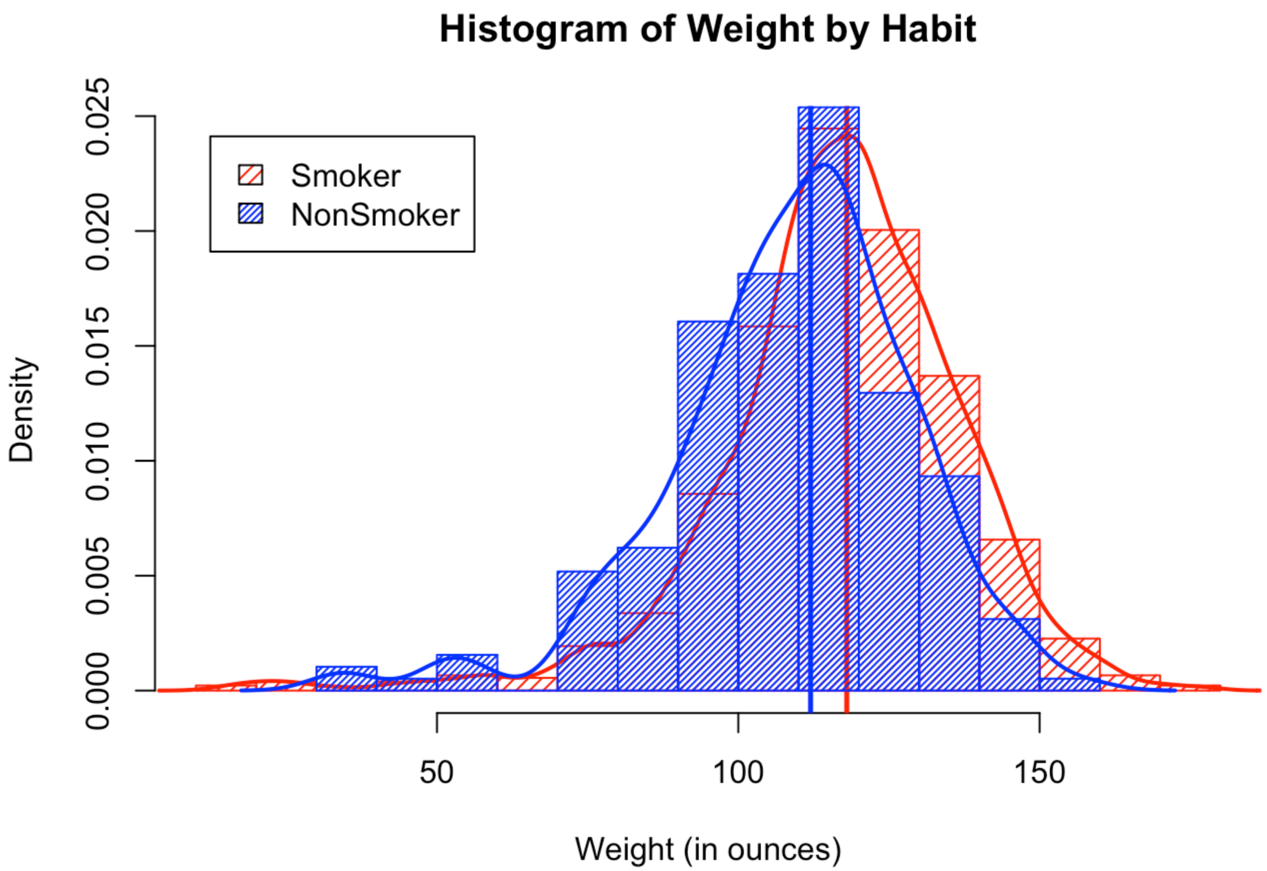
```
## [1] "NonSmoker" "Smoker"
```

```
 #e)
 boxplot(weight ~ Habit, data = data_habit_known)
```



2

```
library(ggplot2)
smoker_data <- droplevels(subset(data, Habit != "Smoker"))
nonsmoker_data <- droplevels(subset(data, Habit != "NonSmoker"))

with(smoker_data, hist(weight,
prob = TRUE, density = 20, col = "red",
xlab = "Weight (in ounces)", main = "Histogram of Weight by Habit",
))
lines(density(smoker_data$weight), lwd = 2, col = "red")
abline(v = median(smoker_data$weight),lwd = 2.5, col = "red")
with(nonsmoker_data, hist(weight,
prob = TRUE, density = 40, col = "blue", add = TRUE
))
lines(density(nonsmoker_data$weight), lwd = 2, col = "blue")
abline(v = median(nonsmoker_data$weight), lwd = 2.5, col = "blue")
legend("topleft", c("Smoker", "NonSmoker"),
density = c(20, 40),
fill = c("red", "blue"),
inset = 0.05
)
```



```
# based  on the plot, do you think there is a significant difference between the typical
weight of a baby born to a mother who smokes and the typical weight of a baby born to a
mother who does not smoke?

# I believe that there is a significant but a difference. I believe that the difference
is significant enough to pose a threat and thereby should be further examined
```

3

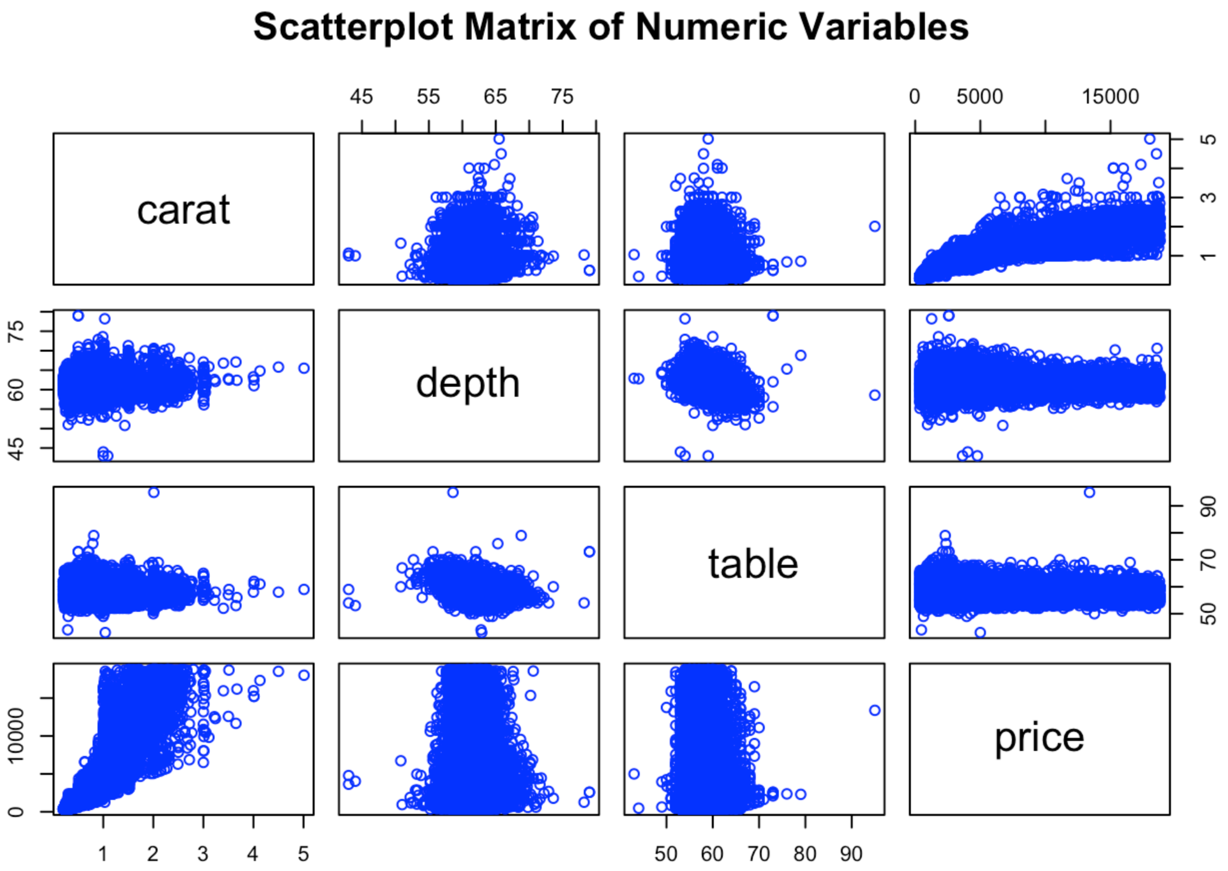
```
library(ggplot2)

diamonds_data <- diamonds

head(diamonds_data)
```

```
## # A tibble: 6 × 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
## 3  0.23 Good      E     VS1      56.9    65   327  4.05  4.07  2.31
## 4  0.29 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
## 5  0.31 Good      J     SI2      63.3    58   335  4.34  4.35  2.75
## 6  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
```

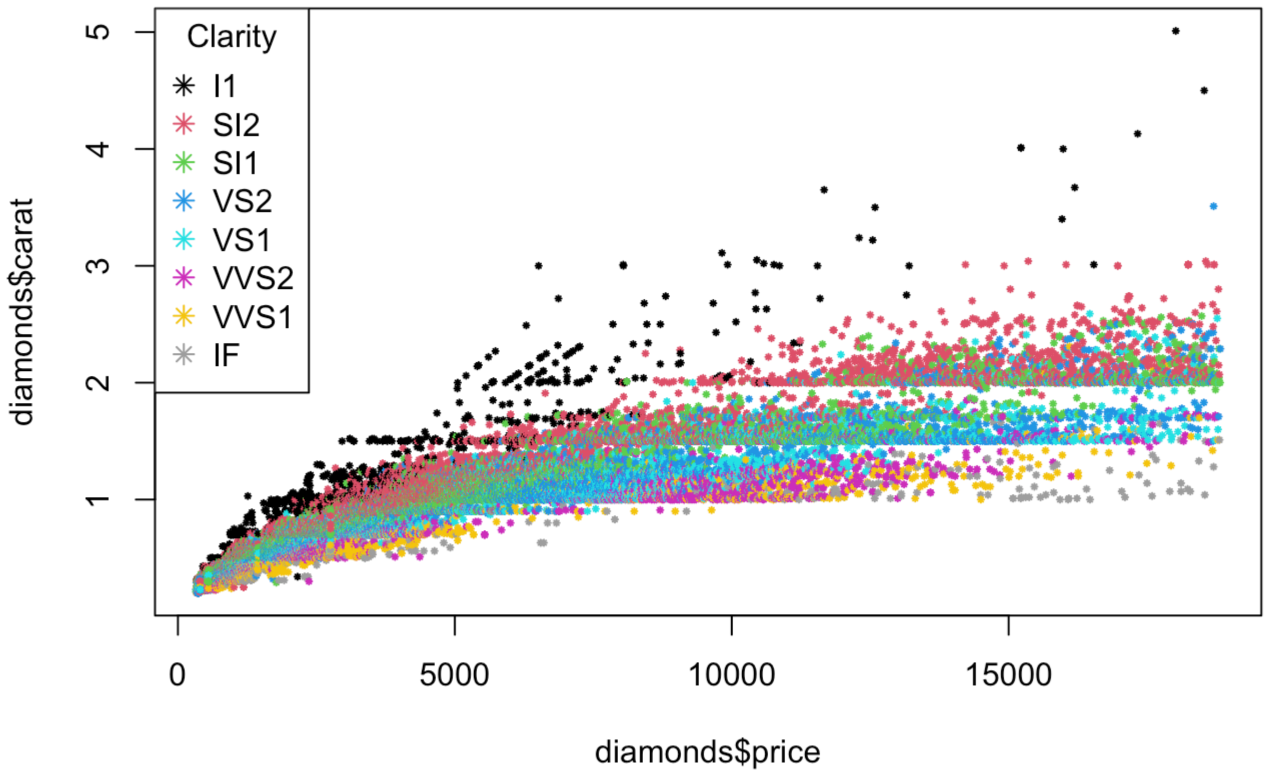
```
#a)

lmat <- lm(carat~price, data = diamonds_data)
lmat
```

```
##
## Call:
## lm(formula = carat ~ price, data = diamonds_data)
##
## Coefficients:
## (Intercept)        price
##   0.3672972    0.0001095
```

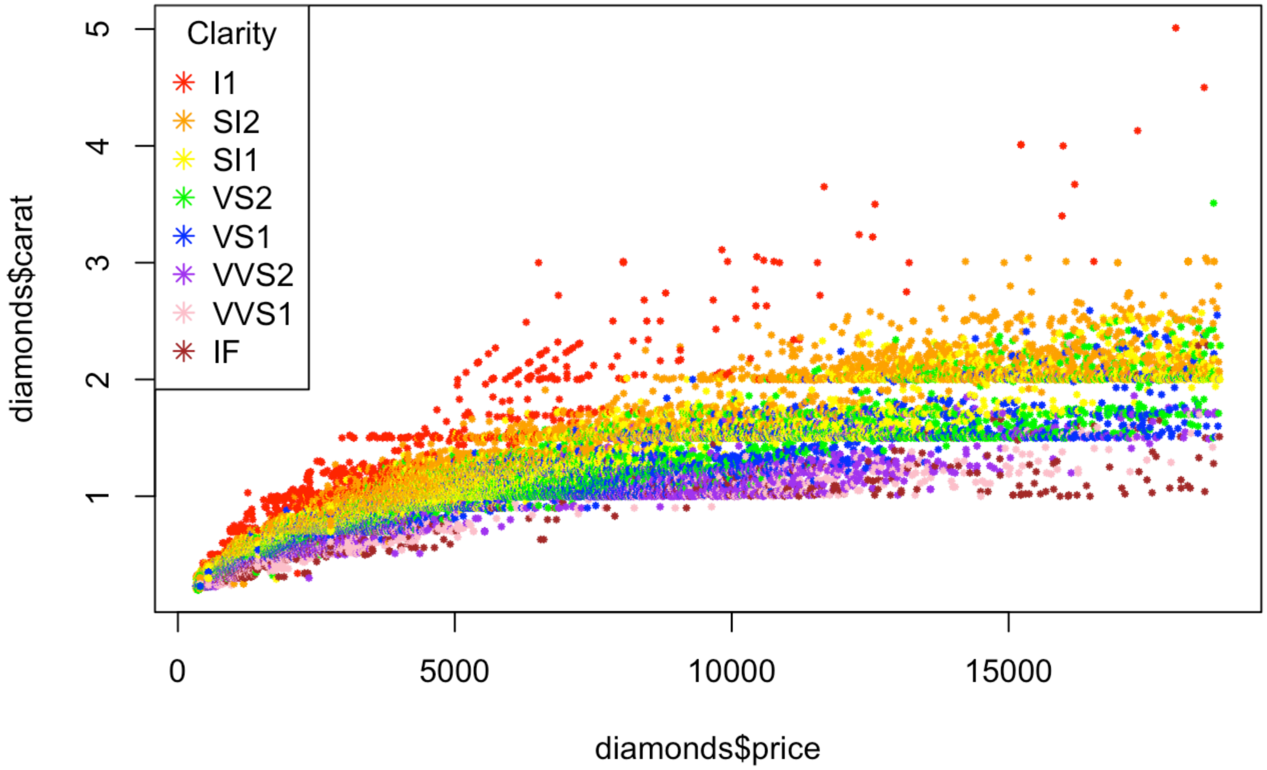
```
pairs(diamonds[, c("carat", "depth", "table", "price")],main = "Scatterplot Matrix of Nu
meric Variables", pch = 1,  col = "blue",)
```



```
#the carat and the price have the strongest relationship though the correlation does not seem to be linear

#b)

plot(diamonds$price, diamonds$carat, pch = 8, cex = 0.3, col = diamonds$clarity)

legend("topleft", legend = levels(diamonds$clarity), col = 1:8, pch = 8, title = "Clarity")
```



```
#The default colors 1 to 8 are chosen because the clarity column in the diamonds dataset is a factor with levels from "I1" to "IF". Each level is assigned a number internally which is 1-8 corresponding to the level in the order it first appears in the data.

#c)
colors <- c("I1" = "red", "SI2" = "orange", "SI1" = "yellow", "VS2" = "green", "VS1" = "blue", "VVS2" = "purple", "VVS1" = "pink", "IF" = "brown")
plot(diamonds$price, diamonds$carat, pch = 8, cex = 0.3, col = colors[diamonds$clarity])
legend("topleft", legend = levels(diamonds$clarity), col = colors, pch = 8, title = "Clarity")
```



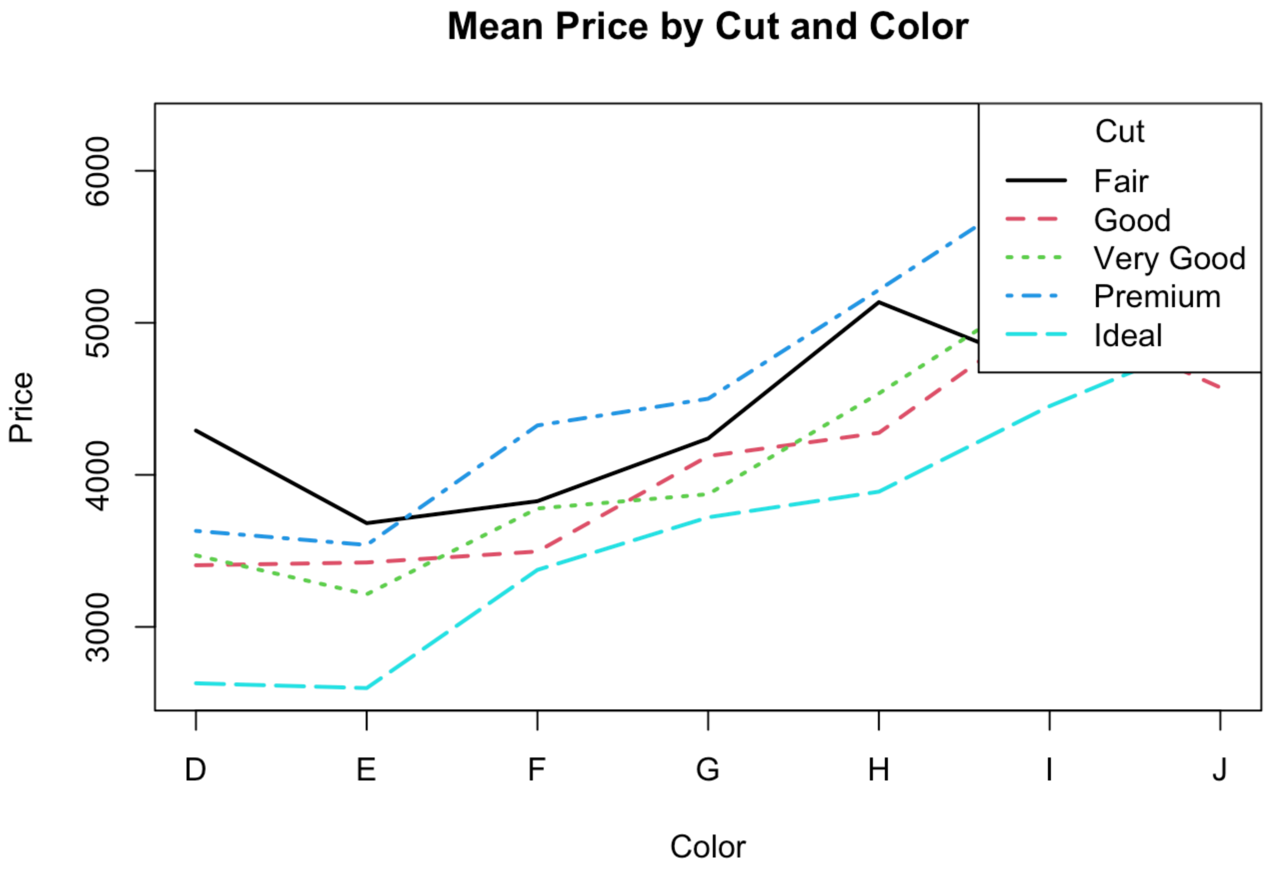
```
#d) The first scatterplot shows that as the carat size increases, the price tends to increase as well. However, there is still a lot of variability in the price for each carat size. The three-way relationship observed in the scatterplot shows that the clarity of a diamond can vary for different combinations of carat and price demonstrating that the clarity of a diamond is determined by both its carat size and price.
```

4

```
mean_price <- aggregate(price ~ color + cut, data = diamonds, FUN = mean)

color_levels <- levels(diamonds$color)
cut_levels <- levels(diamonds$cut)

mean_price_mat <- matrix(NA, nrow = length(color_levels), ncol = length(cut_levels), dimnames = list(color_levels, cut_levels))
for (i in 1:length(color_levels)) {
  for (j in 1:length(cut_levels)) {
    mean_price <- mean(diamonds$price[diamonds$color == color_levels[i] & diamonds$cut == cut_levels[j]])
    mean_price_mat[i, j] <- mean_price
  }
}

mean_price_mat
```

```
##       Fair     Good Very Good  Premium    Ideal
## D 4291.061 3405.382  3470.467 3631.293 2629.095
## E 3682.312 3423.644  3214.652 3538.914 2597.550
## F 3827.003 3495.750  3778.820 4324.890 3374.939
## G 4239.255 4123.482  3872.754 4500.742 3720.706
## H 5135.683 4276.255  4535.390 5216.707 3889.335
## I 4685.446 5078.533  5255.880 5946.181 4451.970
## J 4975.655 4574.173  5103.513 6294.592 4918.186
```

```
#b)

matplot(mean_price_mat, type = "l", lty = 1:5, lwd = 2, col = 1:5,
xaxt = "n", xlab = "Color", ylab = "Price", main = "Mean Price by Cut and Color")

axis(1, at = 1:nrow(mean_price_mat), labels = rownames(mean_price_mat))
legend("topright", legend = colnames(mean_price_mat), lty = 1:5, lwd = 2, col = 1:5, tit
le = "Cut")
```



```
#c)

# The mean price of diamonds does differ for different levels of color and cut. Diamonds
with higher cuts tend to have lower mean prices and diamonds with lower cuts have higher
mean prices. Diamonds with lower colors have higher mean prices, while diamonds with hig
her color have lower mean prices.Though the differences in prices between cut levels are
not as pronounced as the differences between color levels.
```