Homework 3

2024-03-01

Question 1

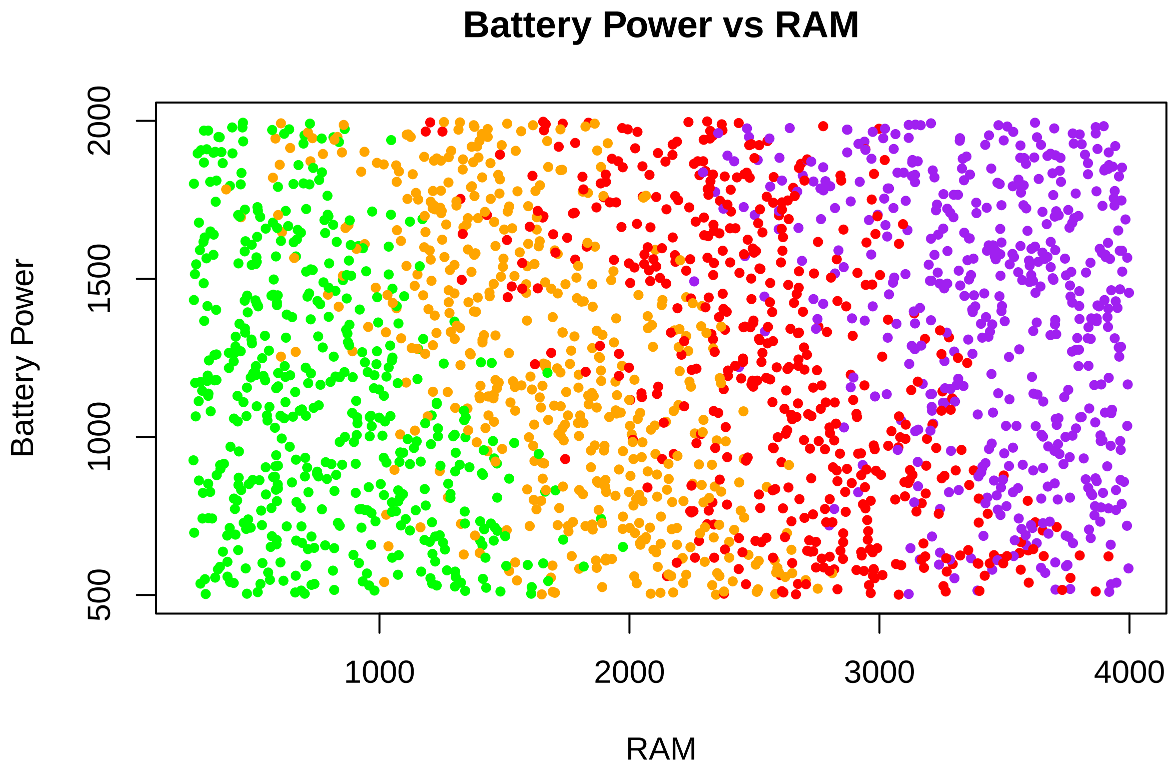
```
setwd("/Users/adi/Desktop/archive")
mobile_data <- read.csv('train.csv',sep=',')
```

a)

```
mobile_data$price_range <- factor(
                                  mobile_data$price_range, levels = c(
                                    "0","1","2","3"), labels = c(
                                      "low", "medium", "high", "very_high"))
```

b)

```
colors <- c("low" = "green", "medium" = "orange", "high" = "red", "very_high" = "purple")
color_code <- colors[mobile_data$price_range]
plot(battery_power ~ ram, data = mobile_data, col = color_code,
     xlab = "RAM", ylab = "Battery Power", pch = 16, cex = 0.7,
     main = "Battery Power vs RAM")
```



c)

```r
cor(mobile_data$ram, mobile_data$battery_power, method = "pearson")
```

```
## [1] -0.0006529264
```

d)

```r
priceLow <- mobile_data[which(mobile_data$price_range == "low"), ]
priceMedium <- mobile_data[which(mobile_data$price_range == "medium"), ]
priceHigh <- mobile_data[which(mobile_data$price_range == "high"), ]
priceVeryhigh <- mobile_data[which(mobile_data$price_range == "very_high"), ]
```

e)

```r
cor(priceLow$ram, priceLow$battery_power, method = "pearson")
```

```
## [1] -0.3465878
```

```r
cor(priceMedium$ram, priceMedium$battery_power, method = "pearson")
```

```
## [1] -0.6133971
```

```r
cor(priceHigh$ram, priceHigh$battery_power, method = "pearson")
```

```
## [1] -0.5874086
```

```r
cor(priceVeryhigh$ram, priceVeryhigh$battery_power, method = "pearson")
```

```
## [1] -0.2627589
```

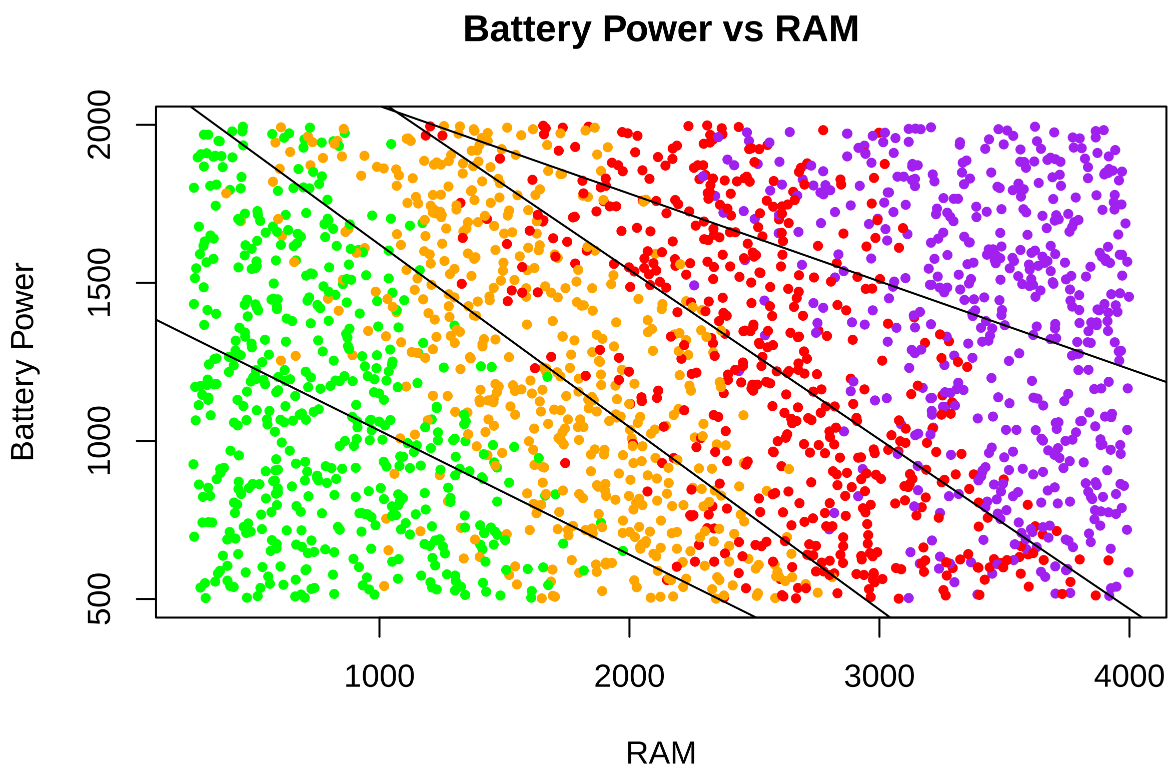
The negative values of the Pearson co-efficients show that the battery power and the RAM of the phones are inversely related. So, the phones with more RAM have less battery power.

The phones at a medium price range have the highest Pearson co-efficient, and the phones at a very high price range have the lowest Pearson co-efficient. This means that at a higher price range, the phones have more battery power even with more RAM.

These results are very different from the result in part (c) because here we are checking for every price range separately. Whereas in part (c), the Pearson co-efficient is very very small, because we are considering all the phones overall, rather than at their different price ranges. So overall, there is a very negligible relation between the RAM and battery power.

f)

```r
plot(battery_power ~ ram, data = mobile_data, col = color_code,
     pch = 16, cex = 0.7, xlab = "RAM", ylab = "Battery Power",
      main = "Battery Power vs RAM")
m_l <- lm(battery_power ~ ram, data = priceLow)
abline(a = coef(m_l)[1], b = coef(m_l)[2])
m_m <- lm(battery_power ~ ram, data = priceMedium)
abline(a = coef(m_m)[1], b = coef(m_m)[2])
m_h <- lm(battery_power ~ ram, data = priceHigh)
abline(a = coef(m_h)[1], b = coef(m_h)[2])
m_vh <- lm(battery_power ~ ram, data = priceVeryhigh)
abline(a = coef(m_vh)[1], b = coef(m_vh)[2])
```



g)

```
round(mean(mobile_data$clock_speed[which(mobile_data$n_cores == 4)]),2)
```

```
## [1] 1.55
```

```
round(median(mean(mobile_data$clock_speed[which(mobile_data$n_cores == 4)])),2)
```

```
## [1] 1.55
```

```
round(mean(mobile_data$clock_speed[which(mobile_data$n_cores == 6)]),2)
```

```
## [1] 1.53
```

```
round(median(mean(mobile_data$clock_speed[which(mobile_data$n_cores == 6)])),2)
```

```
## [1] 1.53
```

```
round(mean(mobile_data$clock_speed[which(mobile_data$n_cores == 8)]),2)
```

```
## [1] 1.51
```

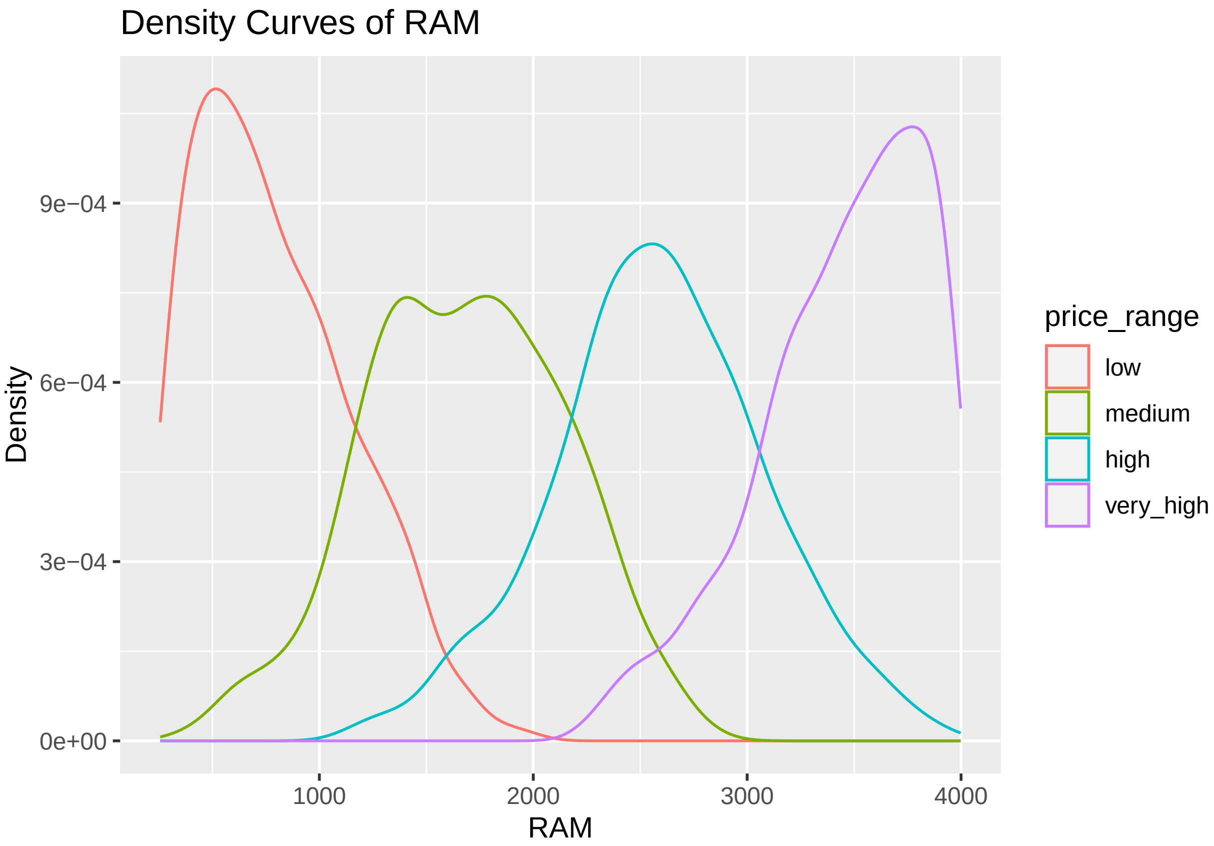
```
round(median(mean(mobile_data$clock_speed[which(mobile_data$n_cores == 8)])),2)
```

```
## [1] 1.51
```

The average and median clock speed doesn’t change because the speeds are symmetrically distributed.

h)

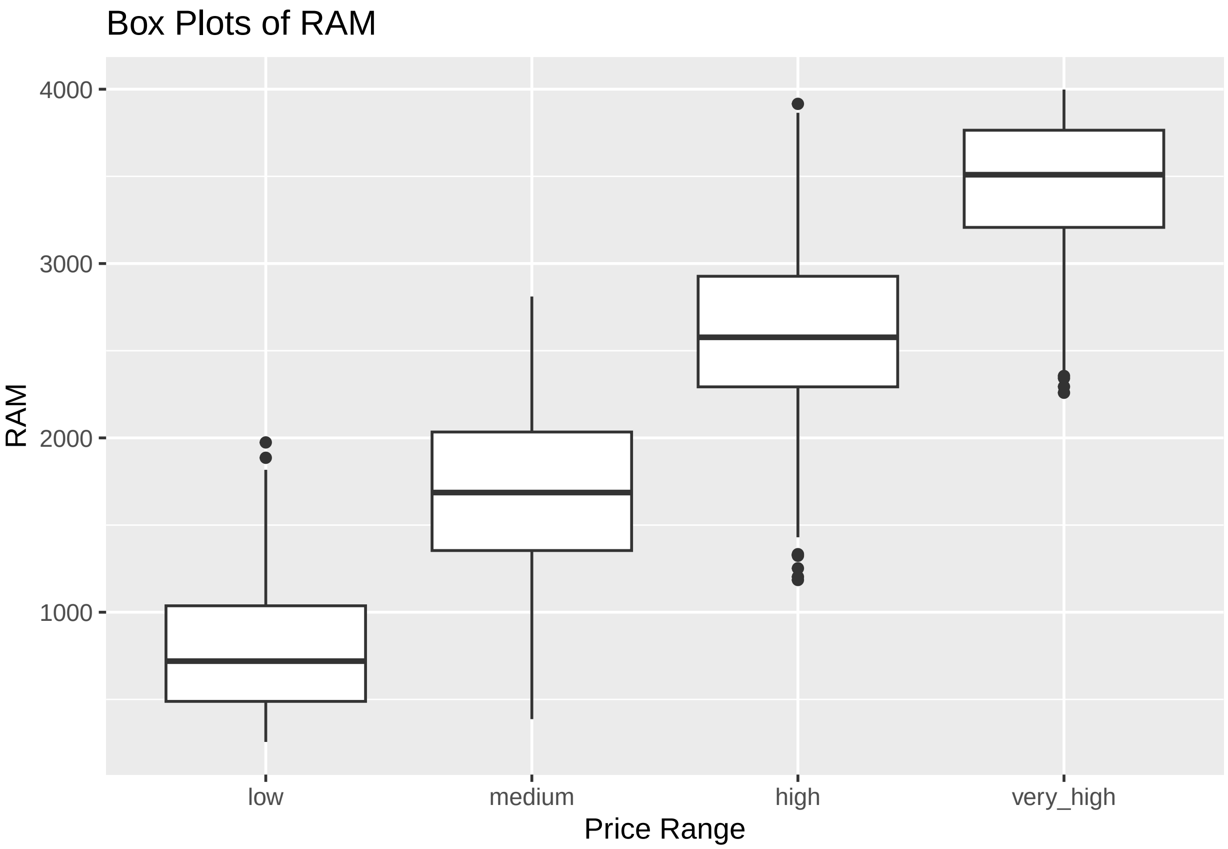
```
library(ggplot2)
ggplot(data = mobile_data) +
geom_density(mapping = aes(
  x = ram, color = price_range)) + labs(
    x = "RAM", y = "Density", title = "Density Curves of RAM")
```



The density curves show that the data for the low price range is positively skewed and for the very high price range is negatively skewed, and the medium and high price ranges have normal distributions.

i)

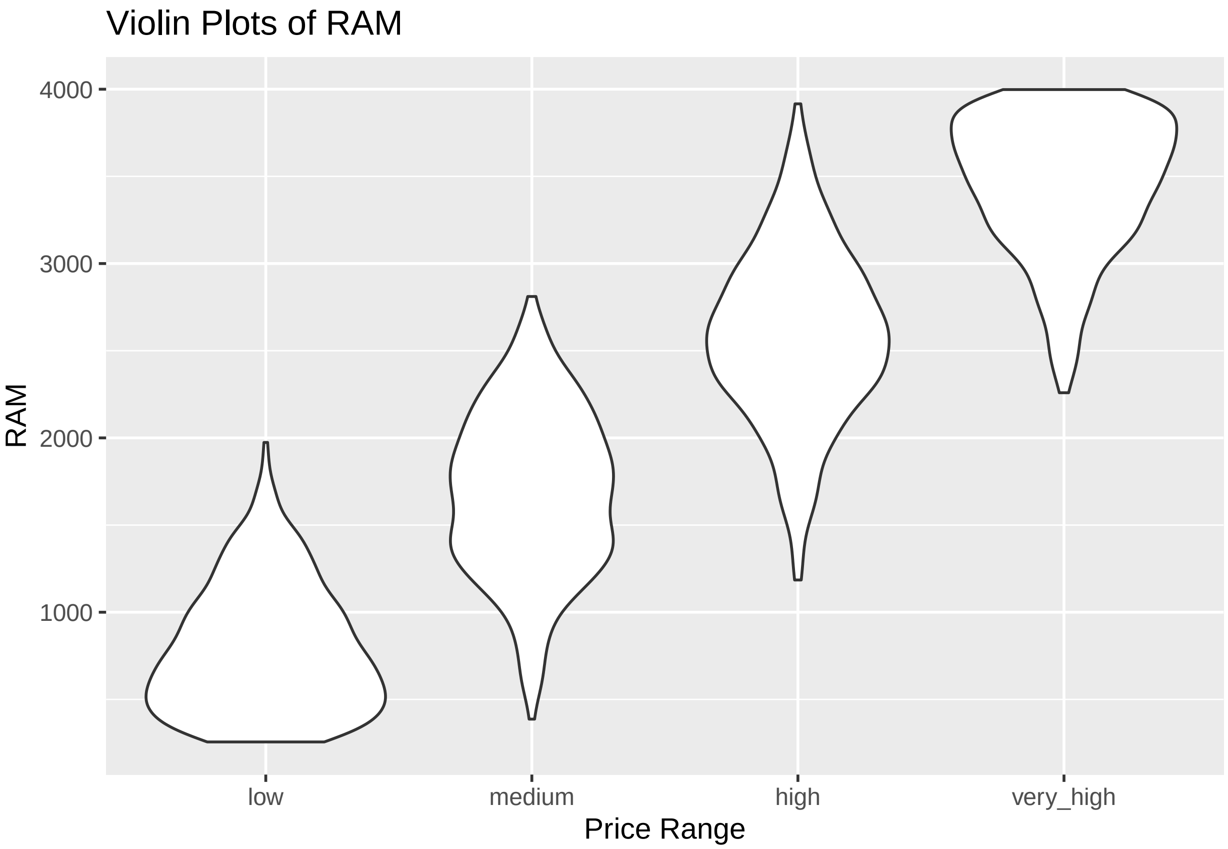
```
library(ggplot2)
ggplot(data = mobile_data) +
geom_boxplot(mapping = aes(
  x = price_range, y = ram)) + labs(
    x = "Price Range", y = "RAM", title = "Box Plots of RAM")
```



The boxplots show that the data for the low price range is positively skewed and for the very high price range is negatively skewed, and the medium and high price ranges have normal distributions.

j)

```
library(ggplot2)
ggplot(data = mobile_data) +
geom_violin(mapping = aes(
  x = price_range, y = ram)) + labs(
    x = "Price Range", y = "RAM", title = "Violin Plots of RAM")
```



The boxplots show that the data for the low price range is positively skewed and for the very high price range is negatively skewed, and the medium and high price ranges have normal distributions.

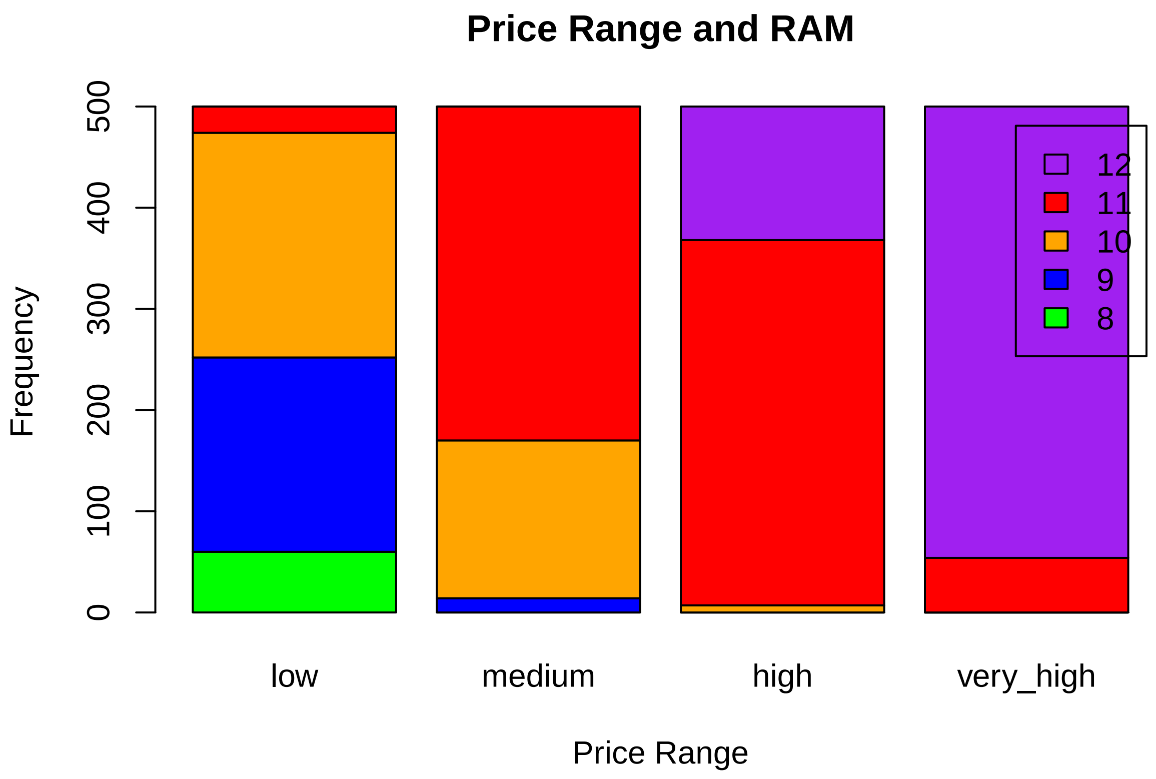
k)

```r
for (i in 1:length(mobile_data$ram)) {
  mobile_data$ram[i] <- round(log2(mobile_data$ram[i]))
}
mobile_data$ram <- factor(mobile_data$ram)
```

This approach makes sense because the RAM values are widely distributed. By taking log2 of the RAM, we have an easier time categorizing the values, which. makes plotting graphs and analyzing of the data easier.

l)

```r
F <- table(mobile_data$ram, mobile_data$price_range)
barplot(F, ylab = "Frequency", xlab = "Price Range", legend.text = TRUE,
        main = "Price Range and RAM", col = c("green", "blue", "orange", "red", "purple"))
```


Price Range and RAM
Frequency
0
100
200
300
400
500
low
medium
high
very_high
Price Range
12
11
10
9
8

Question 2

```r
library(ggplot2)
data(mpg)
mpg_data <- mpg
```

a)

```r
mpg_data$cyl <- factor(mpg_data$cyl, ordered = TRUE,
                       levels = c("4", "5", "6", "8"))
```

b)

```r
trans_sub <- substr(mpg_data$trans, 1, 4)
mpg_data$trans <- factor(trans_sub)
```

c)

```r
mpg_data$drv <- factor(mpg_data$drv, ordered = TRUE,
                       levels = c("f", "r", "4"))
```

d)

```r
mpg_data$fl[which(mpg_data$fl %in% c("e","c"))] <- "other"
mpg_data$fl[which(mpg_data$fl %in% c("r","p"))] <- "gasoline"
mpg_data$fl[which(mpg_data$fl == "d")] <- "diesel"
mpg_data$fl <- factor(mpg_data$fl)
```

e)

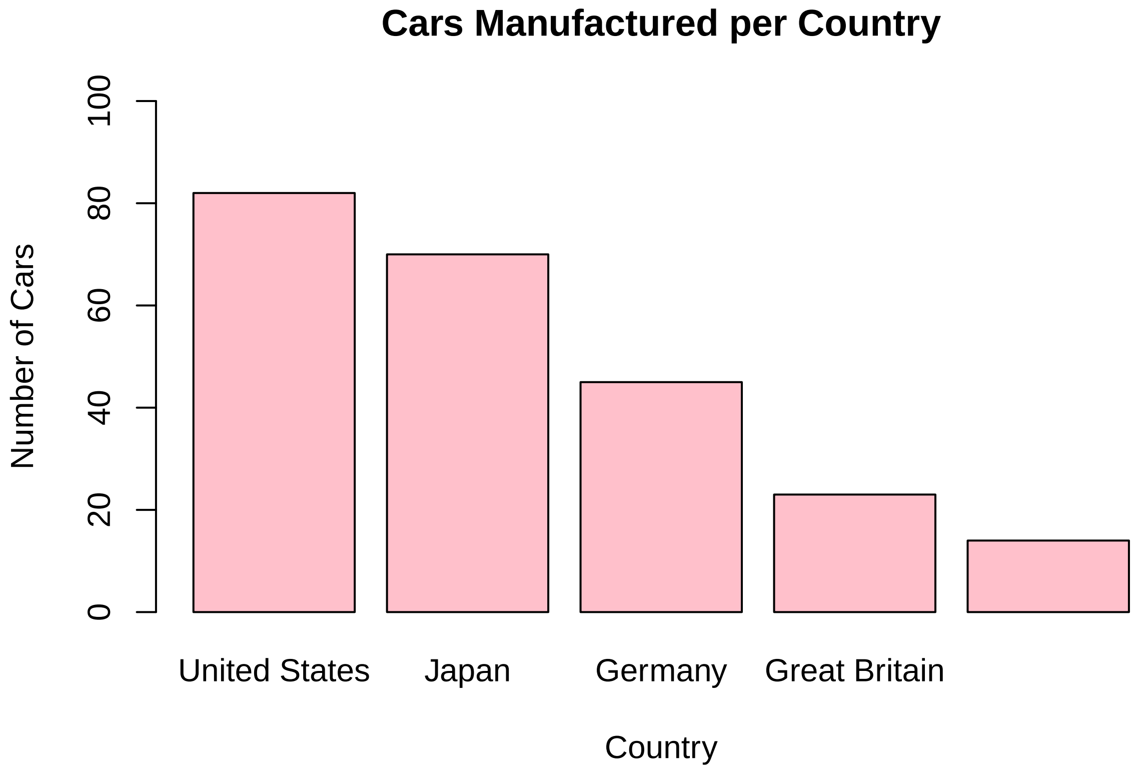
```r
mpg_data$class <- factor(mpg_data$class, ordered = TRUE,
                       levels = c("2seater", "subcompact", "compact",
                                  "midsize", "suv", "minivan", "pickup"))
```

f)

```r
country <- c()
for (i in 1:length(mpg_data$manufacturer)) {
if(mpg_data$manufacturer[i] %in% c(
  "chevorlet", "dodge", "ford", "lincoln", "jeep", "mercury", "pontiac")) {
    country[i] <- "United States"
  }
else if(mpg_data$manufacturer[i] %in% c(
  "honda", "nissan", "subaru", "toyota")) {
   country[i] <- "Japan"
}
else if(mpg_data$manufacturer[i] %in% c(
  "audi", "volkswagen")) {
    country[i] <- "Germany"
  }
else if(mpg_data$manufacturer[i] == "hyundai") {
    country[i] <- "South Korea"
}
else {
    country[i] <- "Great Britain"
  }
}
mpg_data <- cbind(mpg_data,country)
```

g)

```
country_cars <- table(mpg_data$country)
country_cars <- sort(country_cars, decreasing=TRUE)
barplot(country_cars, main = "Cars Manufactured per Country", xlab = "Country",
ylab = "Number of Cars", ylim = c(0,100), col = "pink")
```



United States has the most samples, and South Korea has the least.

h)

```
us_car <- mpg_data[which(mpg_data$country == "United States"), ]
summary(us_car[,c("displ","cyl","trans","drv","fl","class")])
```

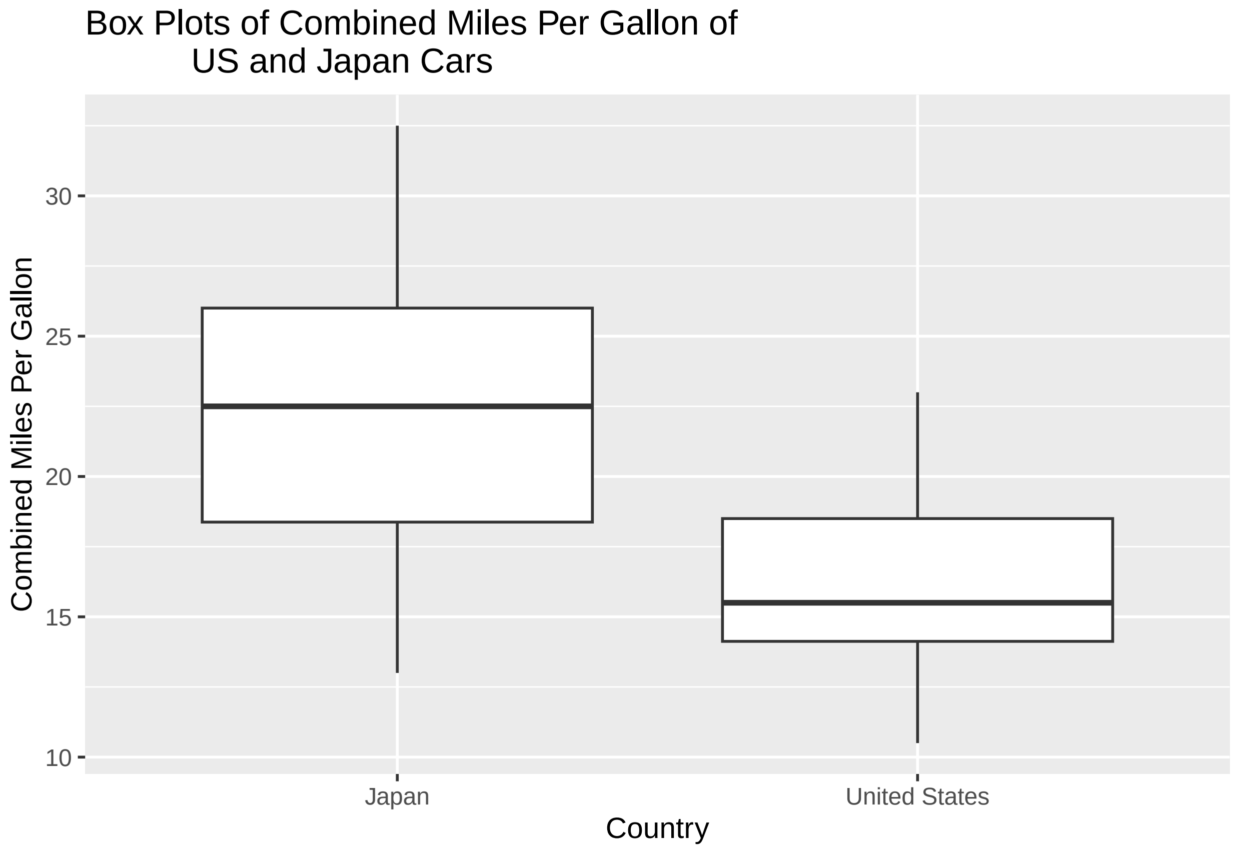
```
##      displ        cyl     trans      drv          fl               class   
##  Min.   :2.400   4: 1   auto:67   f:16   diesel  : 1   2seater   : 0  
##  1st Qu.:3.900   5: 0   manu:15   r:15   gasoline:75   subcompact: 9  
##  Median :4.600   6:34             4:51   other   : 6   compact   : 0  
##  Mean   :4.459   8:47                                  midsize   : 5  
##  3rd Qu.:5.000                                         suv       :31  
##  Max.   :6.100                                          minivan   :11
##                                                         pickup    :26
```

i)

```
library(ggplot2)
cmpg <- (mpg_data$hwy[which(
  mpg_data$country %in% c(
    "United States", "Japan"))] +
     mpg_data$cty[which(
     mpg_data$country %in% c(
    "United States", "Japan"))])/2

us_and_japan <- mpg_data[which(mpg_data$country %in% c("United States","Japan")), ]


ggplot(data = us_and_japan) +
  geom_boxplot(mapping = aes(
    x = country, y = cmpg)) + labs(
      x = "Country", y = "Combined Miles Per Gallon",
           title = "Box Plots of Combined Miles Per Gallon of
           US and Japan Cars")
```



```
us_and_japan <- cbind(cmpg,us_and_japan)

mean(us_and_japan$cmpg[us_and_japan$country == "United States"])
```

```
## [1] 16.21951
```

```r
median(us_and_japan$cmpg[us_and_japan$country=="United States"])
```

```
## [1] 15.5
```

```r
sd(us_and_japan$cmpg[us_and_japan$country=="United States"])
```

```
## [1] 3.008319
```

```r
IQR(us_and_japan$cmpg[us_and_japan$country=="United States"])
```

```
## [1] 4.375
```

```r
mean(us_and_japan$cmpg[us_and_japan$country=="Japan"])
```

```
## [1] 22.66429
```

```r
median(us_and_japan$cmpg[us_and_japan$country=="Japan"])
```

```
## [1] 22.5
```

```r
sd(us_and_japan$cmpg[us_and_japan$country=="Japan"])
```

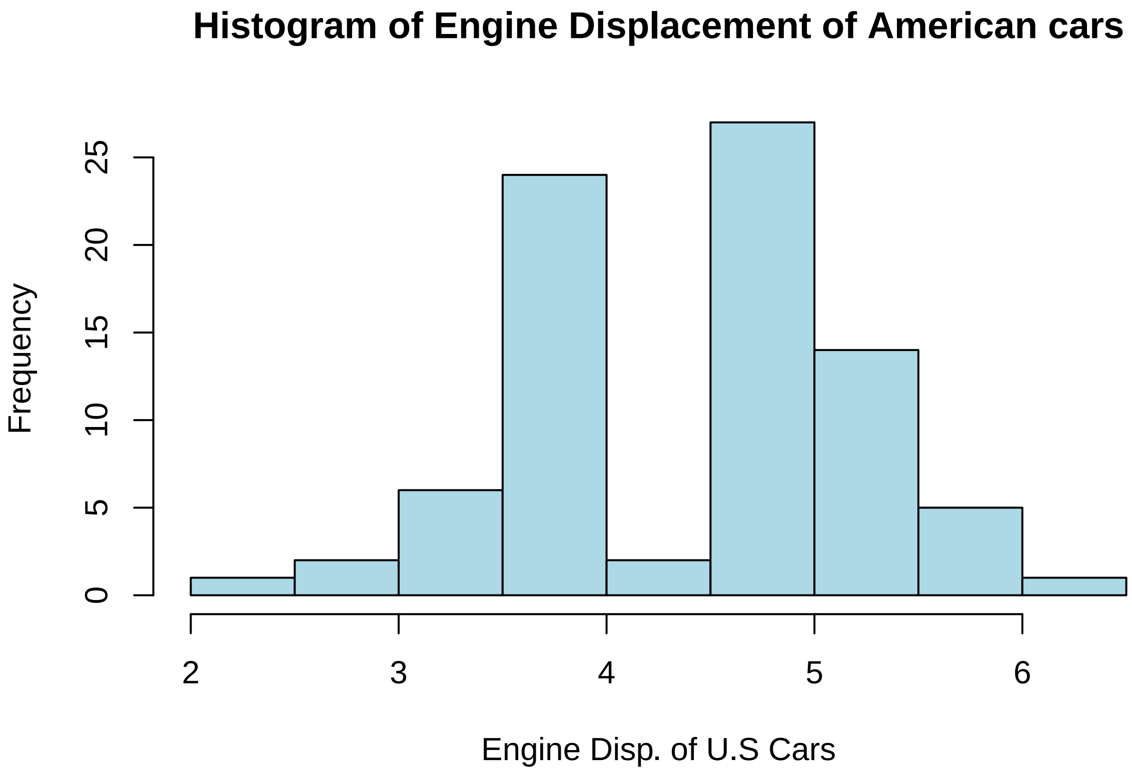
```
## [1] 4.60208
```

```r
IQR(us_and_japan$cmpg[us_and_japan$country=="Japan"])
```

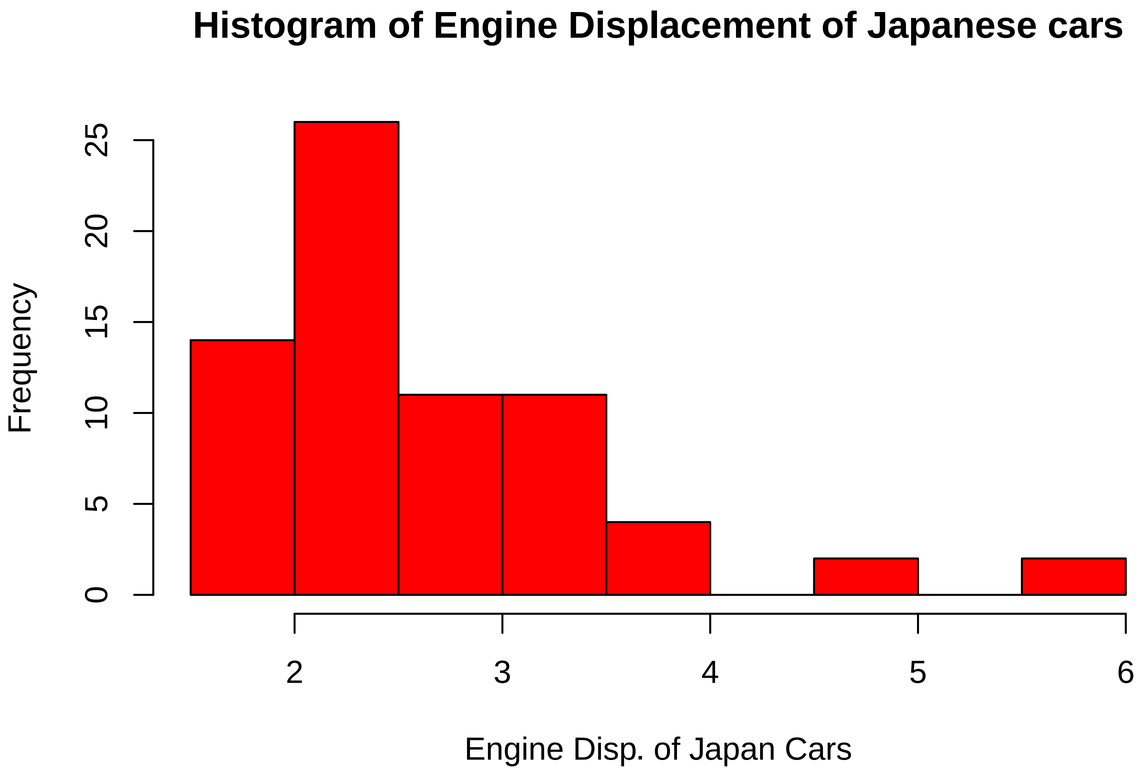
```
## [1] 7.625
```

j)

```r
hist(mpg_data$displ[which(mpg_data$country == "United States")],
     xlab = "Engine Disp. of U.S Cars",
     main = "Histogram of Engine Displacement of American cars", col = "lightblue")
```



```r
hist(mpg_data$displ[which(mpg_data$country == "Japan")],
     xlab = "Engine Disp. of Japan Cars",
     main = "Histogram of Engine Displacement of Japanese cars", col = "red")
```



The histogram of the engine displacement of U.S cars is bi-modal and the histogram of the engine displacement of Japan cars is right skewed.