STAT 3355 - HW 3

2024-03-02

```r
library(ggplot2)
library(dplyr)
```

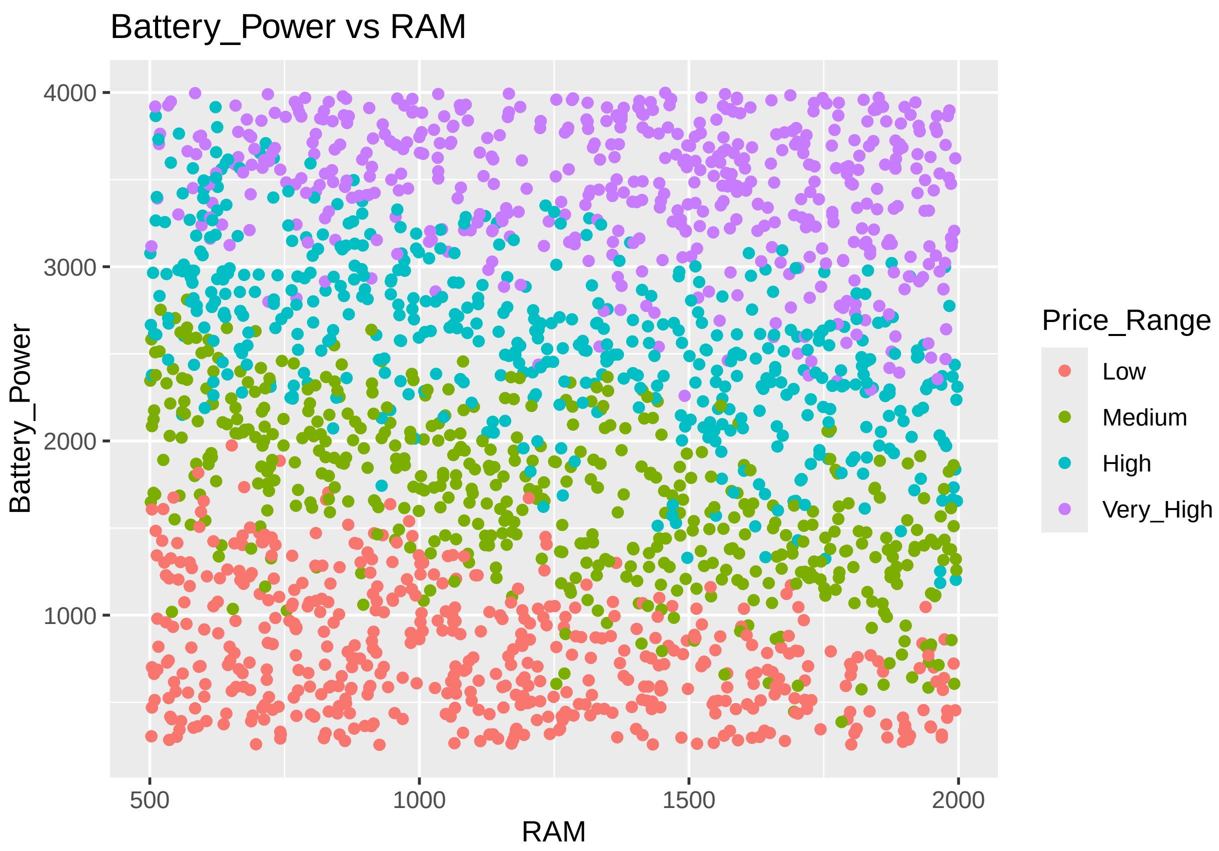
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Problem 1 - (a)

```r
data <- read.csv("/Users/springkim/downloads/train.csv")
data$price_range <- factor(data$price_range, levels = c(0, 1, 2, 3), labels = c("Low", "Medium", "High"
```

Problem 1 - (b)

```r
ggplot(data, aes(x = battery_power, y = ram, col = price_range)) +
  geom_point() +
  labs(x = "RAM", y = "Battery_Power", title = "Battery_Power vs RAM") +
  guides(col = guide_legend(title = "Price_Range"))
```



Problem 1 - (c)

```r
pearsonCorrelation <- cor(data$ram, data$battery_power, method = "pearson")
cat("Pearson Correlation Coefficient:", pearsonCorrelation, "\n")
```

```
## Pearson Correlation Coefficient: -0.0006529264
```

Problem 1 - (d)

```r
priceLow <- data[data$price_range == "Low",]
priceMedium <- data[data$price_range == "Medium",]
priceHigh <- data[data$price_range == "High",]
priceVeryHigh <- data[data$price_range == "Very_High",]
```

Problem 1 - (e)

```r
pearsonCorrelation_priceLow <- cor(priceLow$ram, priceLow$battery_power, method = "pearson")
pearsonCorrelation_priceMedium <- cor(priceMedium$ram, priceMedium$battery_power, method = "pearson")
pearsonCorrelation_priceHigh <- cor(priceHigh$ram, priceHigh$battery_power, method = "pearson")
```

```
pearsonCorrelation_priceVeryHigh <- cor(priceVeryHigh$ram, priceVeryHigh$battery_power, method = "pears
cat("Pearson Correlations by Price\n",
"Low Price:", pearsonCorrelation_priceLow, "\n",
"Medium Price:", pearsonCorrelation_priceMedium, "\n", "High Price:", pearsonCorrelation_priceHigh, "\n
"Very High Price:", pearsonCorrelation_priceVeryHigh, "\n")
```

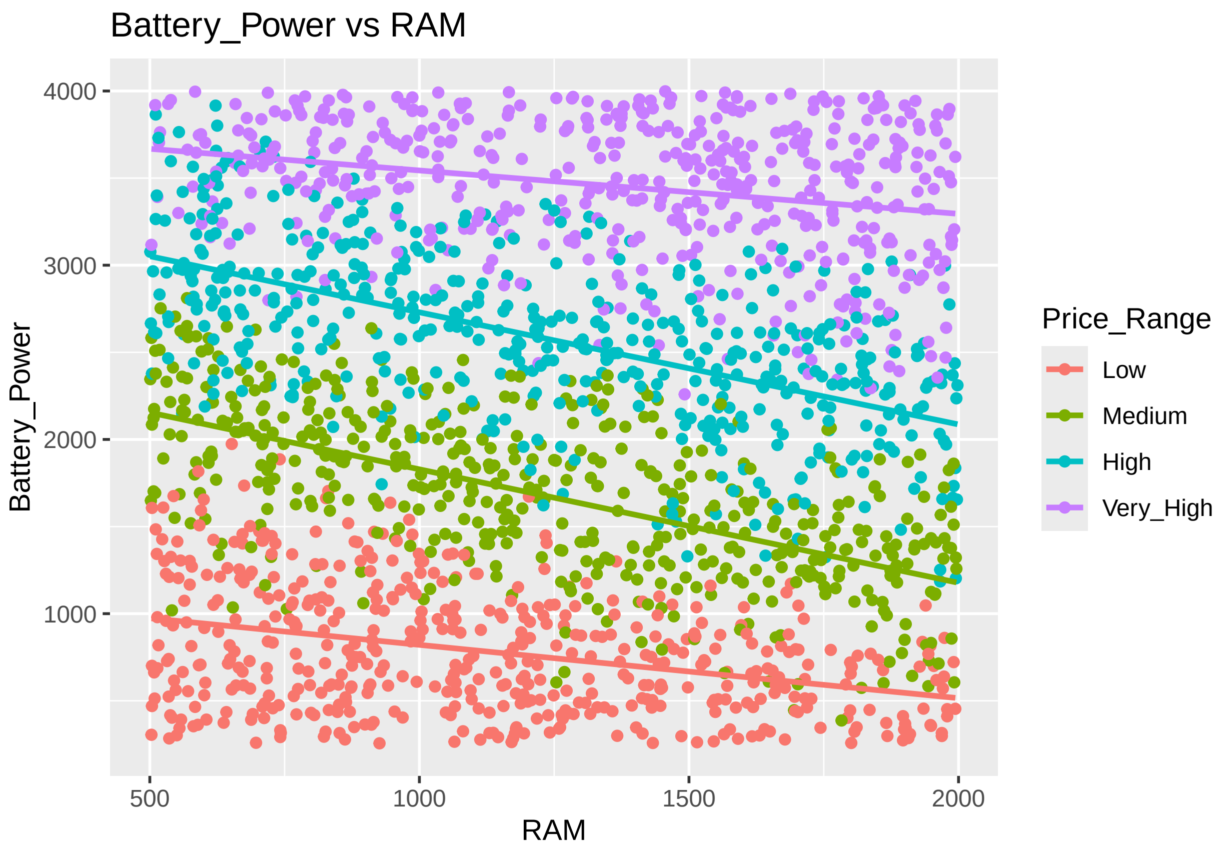
```
## Pearson Correlations by Price
##  Low Price: -0.3465878
##  Medium Price: -0.6133971
##  High Price: -0.5874086
##  Very High Price: -0.2627589
```

```
# I think it's because part (c) calculuated just only one pearson.
```

Problem 1 - (f)

```
ggplot(data, aes(x = battery_power, y = ram, col = price_range)) +
  geom_point() +
  labs(title = "Battery_Power vs RAM", x = "RAM", y = "Battery_Power") +
  guides(col = guide_legend(title = "Price_Range")) +
  geom_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

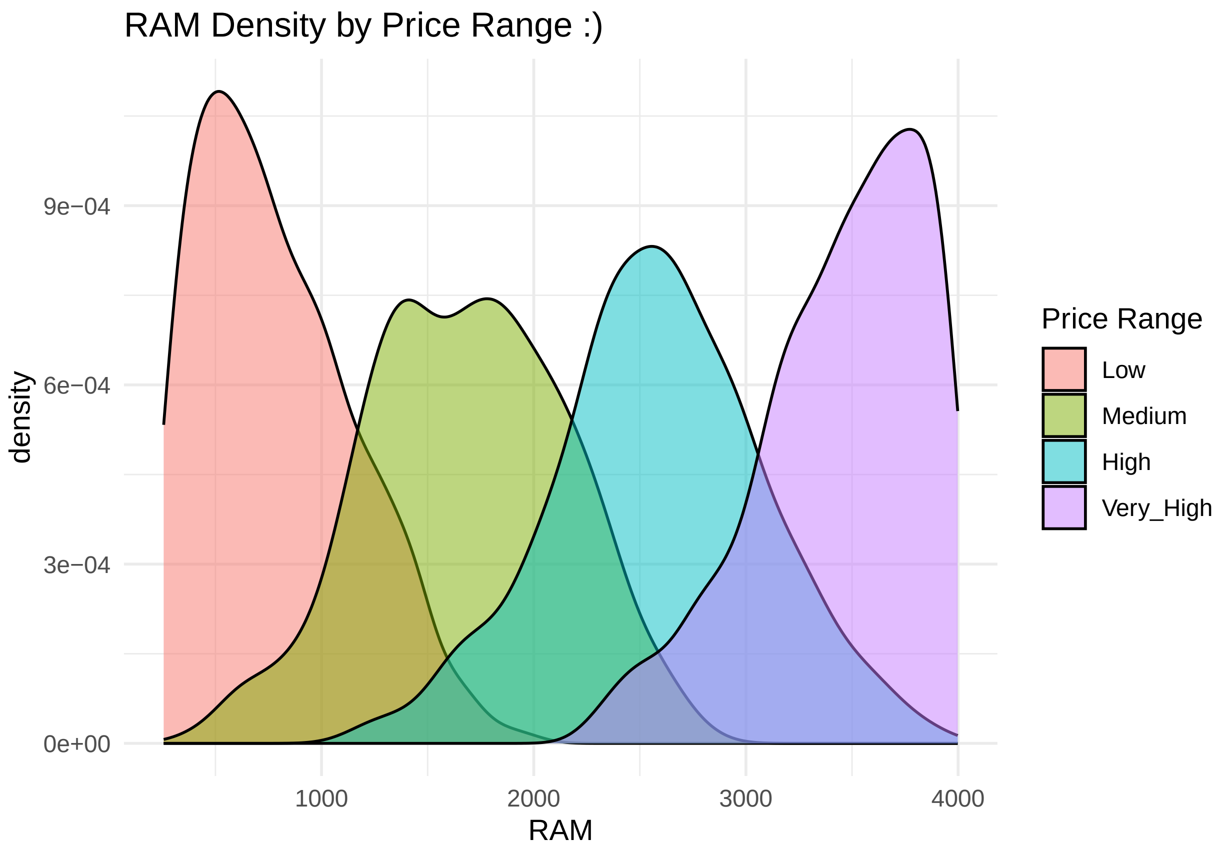


Problem 1 - (g)

```
four_cores <- data[data$n_cores == 4, c("clock_speed", "n_cores")]
six_cores <- data[data$n_cores == 6, c("clock_speed", "n_cores")]
eight_cores <- data[data$n_cores == 8, c("clock_speed", "n_cores")]
four_avg <- round(mean(four_cores$clock_speed), digits = 2)
six_avg <- round(mean(six_cores$clock_speed), digits = 2)
eight_avg <- round(mean(eight_cores$clock_speed), digits = 2)
four_median <- round(median(four_cores$clock_speed), digits = 2)
six_median <- round(median(six_cores$clock_speed), digits = 2)
eight_median <- round(median(eight_cores$clock_speed), digits = 2)
# Because there are consistent factor to average and median clock speed.
```

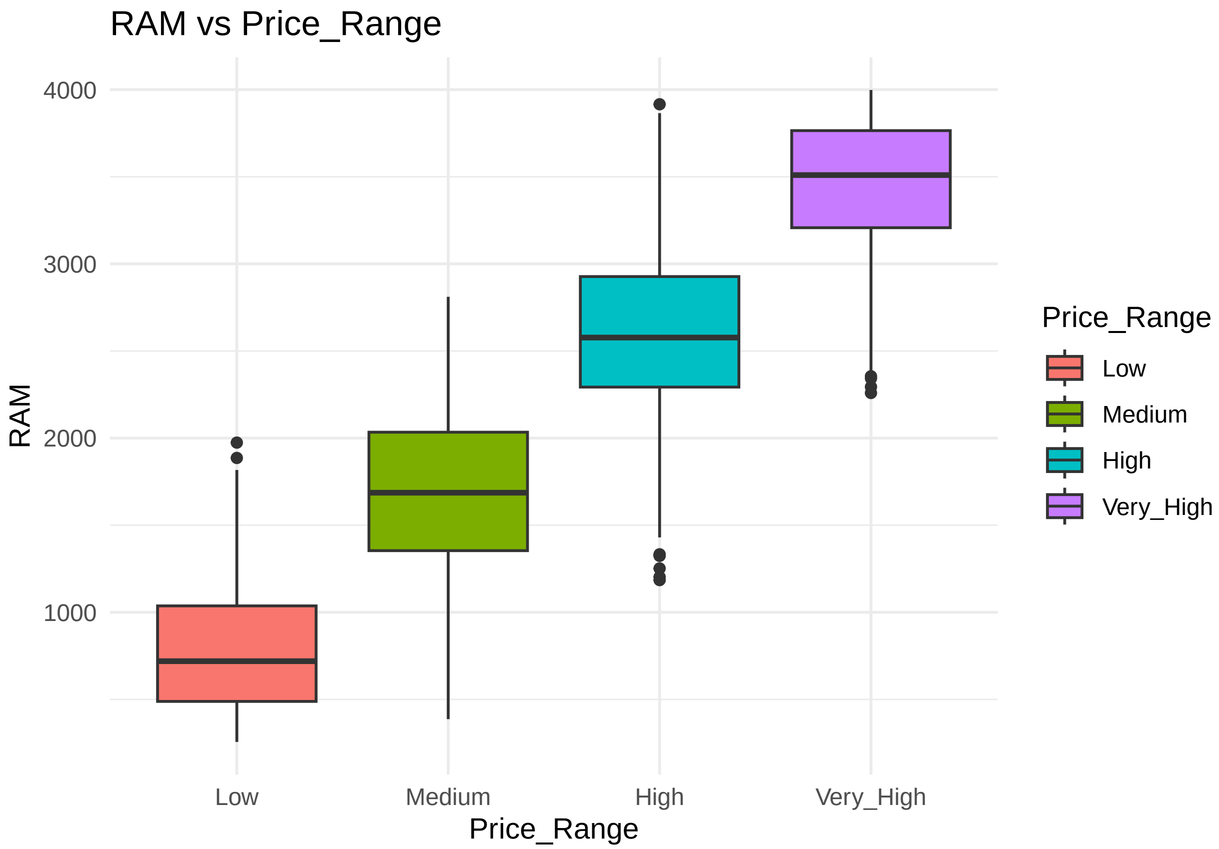
Problem 1 - (h)

```
ggplot(data, aes(x = ram, fill = price_range)) + geom_density(alpha = 0.5) +
  labs(title = "RAM Density by Price Range :)", x = "RAM",vy = "Density",
fill = "Price Range", col = 1:length(levels(data$price_range)), pch = 1) + theme_minimal()
```



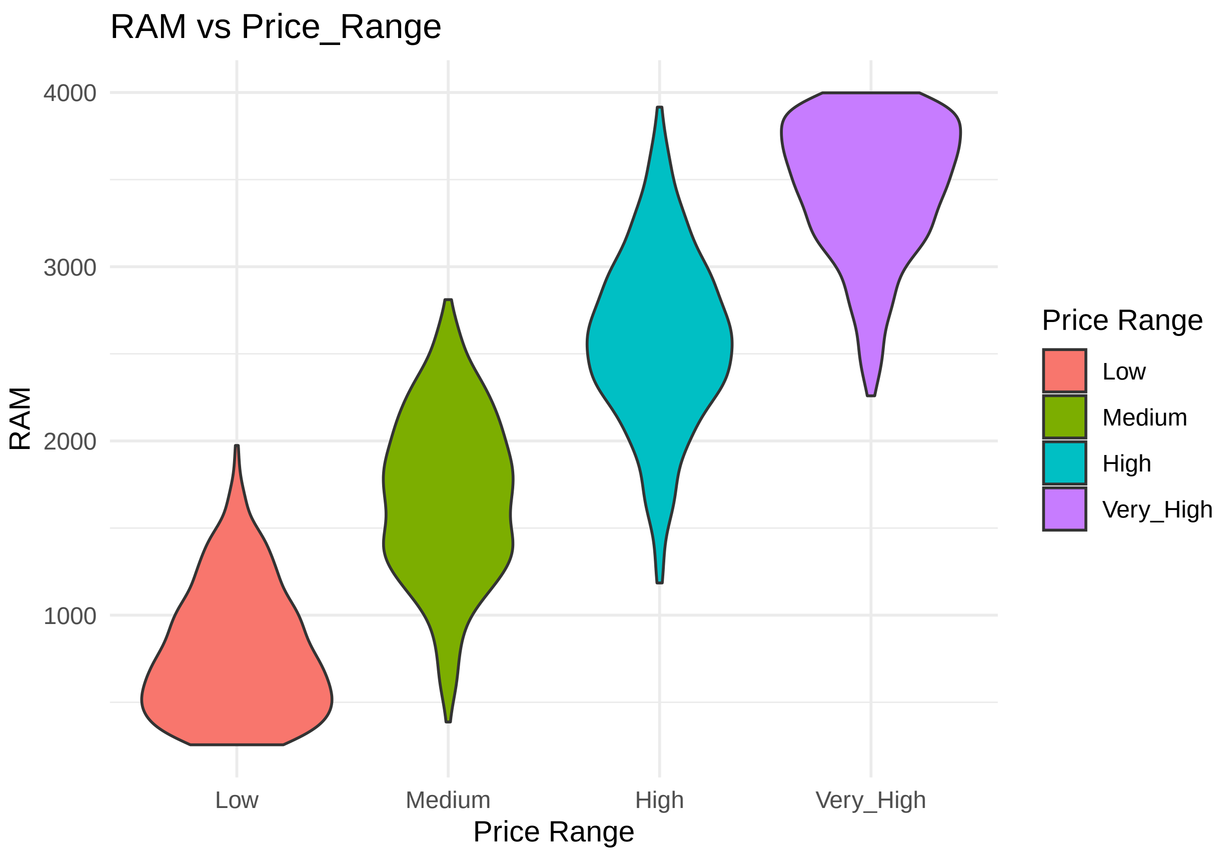
Problem 1 - (i)

```
ggplot(data, aes(x = price_range, y = ram, fill = price_range)) +
  geom_boxplot() +
  labs(title = "RAM vs Price_Range",x = "Price_Range",y = "RAM",
       fill = "Price_Range") + theme_minimal()
```



Problem 1 - (j)

```
ggplot(data, aes(x = price_range, y = ram, fill = price_range)) + geom_violin() +
  labs(title = "RAM vs Price_Range",x = "Price Range", y = "RAM",
       fill = "Price Range") + theme_minimal()
```

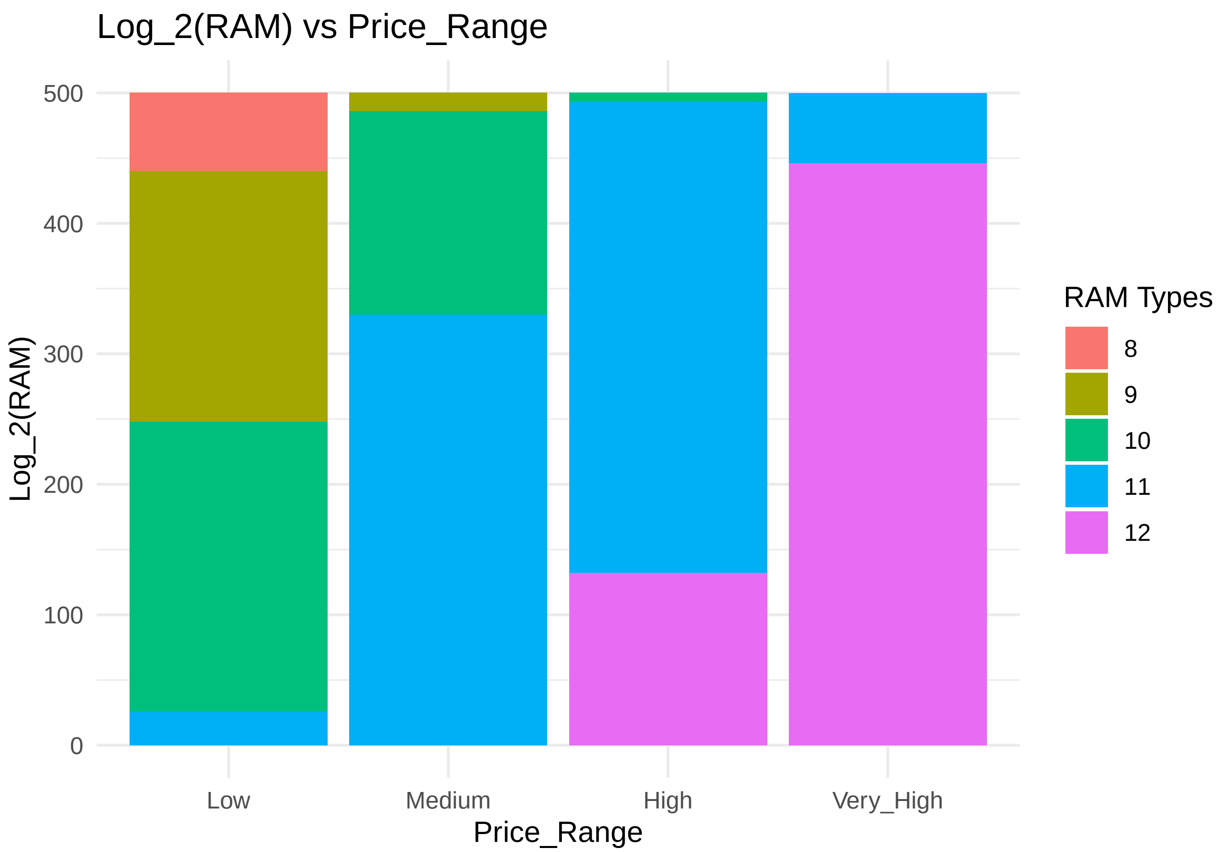


Problem 1 - (k)

```r
data$ram_types <- round(log2(data$ram))
data$ram_types <- factor(data$ram_types)
```

Problem 1 - (l)

```r
ggplot(data, aes(x = price_range, fill = ram_types )) + geom_bar() +
  labs(title = "Log_2(RAM) vs Price_Range",x = "Price_Range", y = "Log_2(RAM)",
       fill = "RAM Types") + theme_minimal()
```



## Problem 2

```
library(ggplot2)
library(dplyr)
```

Problem 2 - (a)

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

Problem 2 - (b)

```
mpg_data$trans <- factor(substr(mpg$trans, 1, 4), levels = c("auto", "manu"))
```

Problem 2 - (c)

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

Problem 2 - (d)

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

Problem 2 - (e)

```
mpg_data$class <- factor(mpg_data$class, levels = c("2seater", "subcompact", "compact", "midsize", "suv
```

Problem 2 - (f)

```
mpg_data$country <- NA
mpg_data <- mpg_data %>%
  mutate(country = case_when(
    manufacturer %in% c("Chevrolet", "Dodge", "Ford", "Jeep", "Lincoln",
                        "Mercury", "Pontiac") ~"US",
    manufacturer %in% c("Honda", "Nissan", "Subaru", "Toyota") ~"Japan",
    manufacturer %in% c("Audi", "Volkswagen") ~"Germany",
    manufacturer %in% c("Hyundai") ~"Korea",
    manufacturer %in% c("land rover") ~"GB"))
```

Problem 2 - (g)

```
#sorted_table <- table(mpg_data$country)[order(-table(mpg_data$country))] barplot(sorted_table, main =
#The most: United States, the least: Great Britian
```

Problem 2 - (h)

```
us_mpg_data <- subset(mpg_data, mpg_data$country == "United States")
mode_cyl <- names(which.max(table(us_mpg_data$cyl)))
mode_trans <- names(which.max(table(us_mpg_data$trans)))
mode_fl <- names(which.max(table(us_mpg_data$fl)))
mode_class <- names(which.max(table(us_mpg_data$class)))
mode_drv <- names(which.max(table(us_mpg_data$drv)))
mode_displ <- names(which.max(table(us_mpg_data$displ)))
```

Problem 2 - (i)

```
japan_mpg_data <- subset(mpg_data, country == "Japan")
us_combined_mpg <- ((us_mpg_data$cty + us_mpg_data$hwy) / 2)
#japan_combined_mpg <- ((japan_mpg_data$cty + japan_mpg_data$hwy) / 2) boxplot(us_combined_mpg, japan_c
```

Problem 2 - (j)

```r
par(mfrow=c(1, 2))
#hist(us_mpg_data$displ,
     #main = "ED: U.S. Cars",
     #xlab = "Engine Displacement",
     #ylab = "Frequency",
     #col = "blue")
#hist(japan_mpg_data$displ,
     #main = "ED: Japanese Cars",
     #xlab = "Engine Displacement",
     #ylab = "Frequency",col = "red")
```