Homework 4

Luke Bennett

10/9/2021

Problem 1 a)

```
library(UsingR)
```

```
## Loading required package: MASS
## Loading required package: HistData
## Loading required package: Hmisc
## Loading required package: lattice
## Loading required package: survival
## Loading required package: Formula
## Loading required package: ggplot2
##
## Attaching package: 'Hmisc'
## The following objects are masked from 'package:base':
##
##     format.pval, units
##
## Attaching package: 'UsingR'
## The following object is masked from 'package:survival':
##
##     cancer
```

```
data("UScereal")
UScereal$mfr <- as.character(UScereal$mfr)
index_G <- which(UScereal$mfr == "G")
UScereal$mfr[index_G] <- "General Mills"
index_K <- which(UScereal$mfr == "K")
UScereal$mfr[index_K] <- "Kellogs"
index_N <- which(UScereal$mfr == "N")
UScereal$mfr[index_N] <- "Nabisco"
index_P <- which(UScereal$mfr == "P")
UScereal$mfr[index_P] <- "Post"
index_Q <- which(UScereal$mfr == "Q")
UScereal$mfr[index_Q] <- "Quaker Oats"
index_R <- which(UScereal$mfr == "R")
UScereal$mfr[index_R] <- "Ralston Purina"
UScereal$mfr <- as.factor(UScereal$mfr)
```

Problem 1 b)

```
UScereal$shelf <- factor(UScereal$shelf, levels = 1:3, labels = c("low", "medium", "high"))
```

Problem 1 c)

```
UScereal$product <- rownames(UScereal)
rownames(UScereal) <- 1:65
```

Problem 1 str()

```
print(str(UScereal))
```

```
## 'data.frame':    65 obs. of  12 variables:
##  $ mfr      : Factor w/ 6 levels "General Mills",..: 3 2 2 1 2 1 6 4 5 1 ...
##  $ calories : num  212 212 100 147 110 ...
##  $ protein  : num  12.12 12.12 8 2.67 2 ...
##  $ fat      : num  3.03 3.03 0 2.67 0 ...
##  $ sodium   : num  394 788 280 240 125 ...
##  $ fibre    : num  30.3 27.3 28 2 1 ...
##  $ carbo    : num  15.2 21.2 16 14 11 ...
##  $ sugars   : num  18.2 15.2 0 13.3 14 ...
##  $ shelf    : Factor w/ 3 levels "low","medium",..: 3 3 3 1 2 3 1 3 2 1 ...
##  $ potassium: num  848.5 969.7 660 93.3 30 ...
##  $ vitamins : Factor w/ 3 levels "100%","enriched",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ product  : chr  "100% Bran" "All-Bran" "All-Bran with Extra Fiber" "Apple Cinnamon Cheerios" ...
## NULL
```

Problem 2 a)

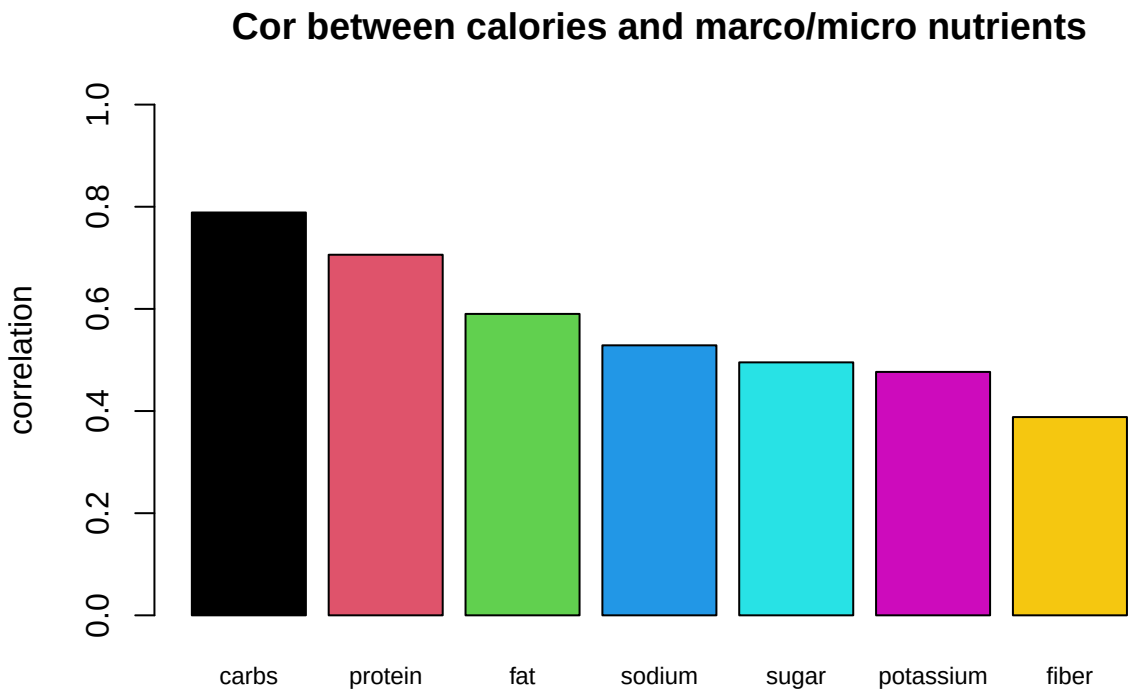
```
calpro <- cor(UScereal$cal, UScereal$protein)
calfat <- cor(UScereal$cal, UScereal$fat)
calsod <- cor(UScereal$cal, UScereal$sodium)
calfib <- cor(UScereal$cal, UScereal$fibre)
calcar <- cor(UScereal$cal, UScereal$carbo)
calsug <- cor(UScereal$cal, UScereal$sugars)
calpot <- cor(UScereal$cal, UScereal$potassium)
```

Correlations listed calories and carbos .789, fat .590, fiber .388, potassium .477, protein .706, sodium .529, sugar .495

Problem 2 b)

```
cors <- c(calpro, calfat, calsod, calfib, calcar, calsug, calpot)
names(cors) <- c("protein", "fat", "sodium", "fiber", "carbs", "sugar", "potassium")
cors <- sort(cors, decreasing = TRUE)
barplot(cors, cex.names = 0.75,
        ylim = c(0, 1),
```

```
        col = 1:7, ylab = "correlation",
        main = "Cor between calories and marco/micro nutrients")
```



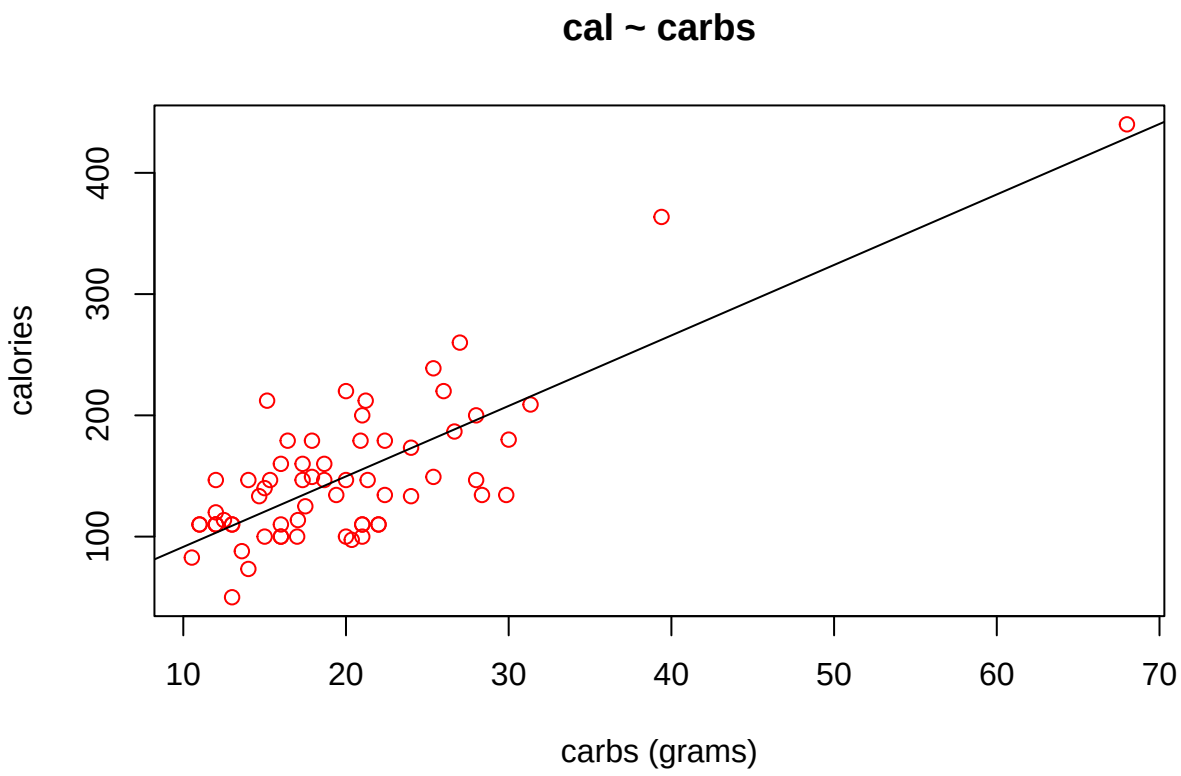
Carbs has the highest correlation.

Problem 2 c)

```
plot(UScereal$calories ~ UScereal$carbo,
     xlab = "carbs (grams)",
     ylab = "calories",
     main = "cal ~ carbs",
     col = "red")
lm(UScereal$calories ~ UScereal$carbo)
```

```
##
## Call:
## lm(formula = UScereal$calories ~ UScereal$carbo)
##
## Coefficients:
##    (Intercept)  UScereal$carbo
##         33.340           5.813
```

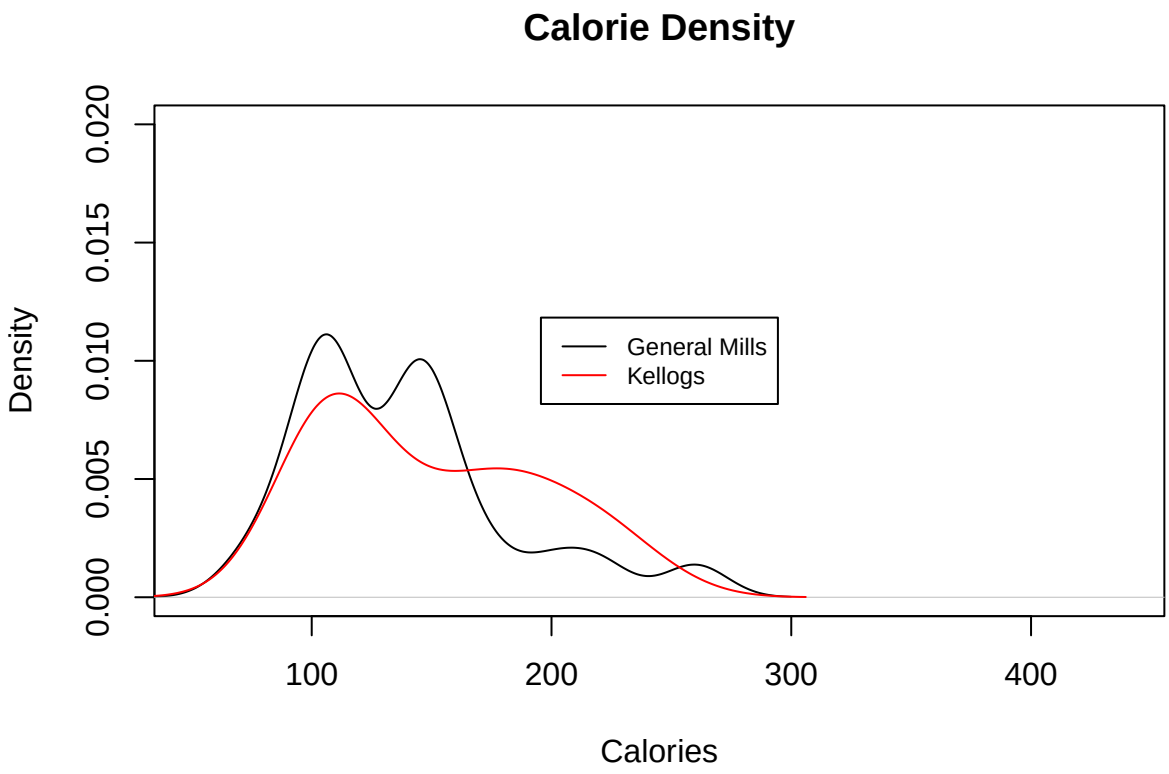
```
abline(a = 33.40, b = 5.813)
```



Y intercept is 33.40 indicated the calories in carbles cereal. slope is how the calories change per each gram of carbs

Problem 2 d)

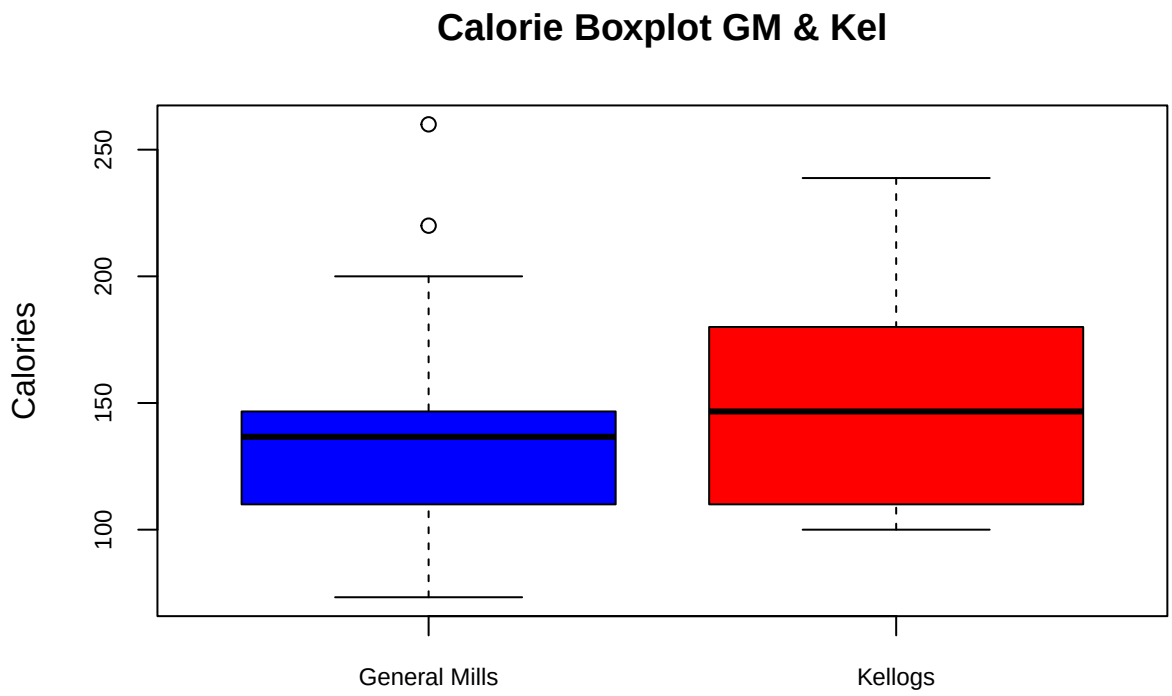
```
caldengm <- density(UScereal$calories[which(UScereal$mfr == "General Mills")])
caldenk <- density(UScereal$calories[which(UScereal$mfr == "Kellogs")])
xmin <- min(UScereal$calories)
xmax <- max(UScereal$calories)
ymin <- 0
plot(caldengm, xlim = c(xmin, xmax), ylim = c(0, 0.02), main = "Calorie Density",
     xlab = "Calories")
lines(caldenk, col = "red")
legend("center", legend = c("General Mills", "Kellogs"), col = c(1, "red"),
       lty = c(1, 1), cex = 0.75)
```



These are probability functions.
They are non-negative for all real numbers and integrate to the number 1.
GM is distributed more heavily with lower calorie, perhaps more sugary cereal.
Kellogs has a less extreme distibution, maybe more hearty cereal.

Problem 2 e)

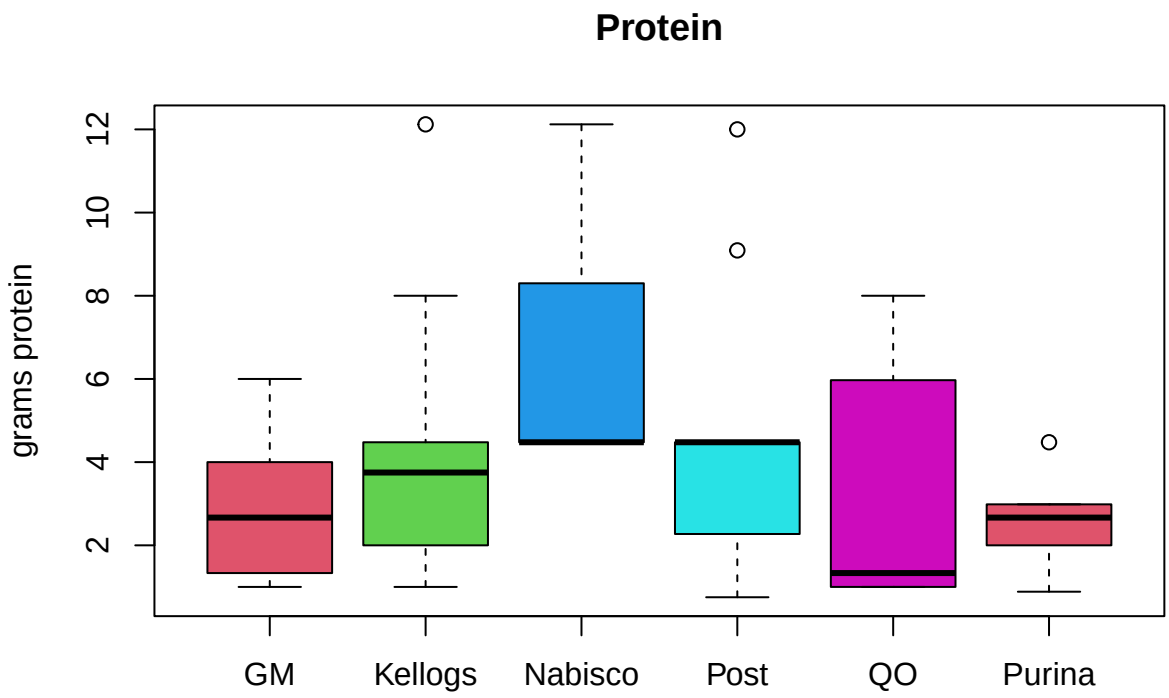
```
gm <- which(UScereal$mfr == "General Mills")
kel <- which(UScereal$mfr == "Kellogs")
boxplot(UScereal$calories[gm], UScereal$calories[kel],
        col = c("blue", "red"),
        main = "Calorie Boxplot GM & Kel",
        ylab = "Calories", cex.axis = 0.75,
        names = c("General Mills", "Kellogs"))
```



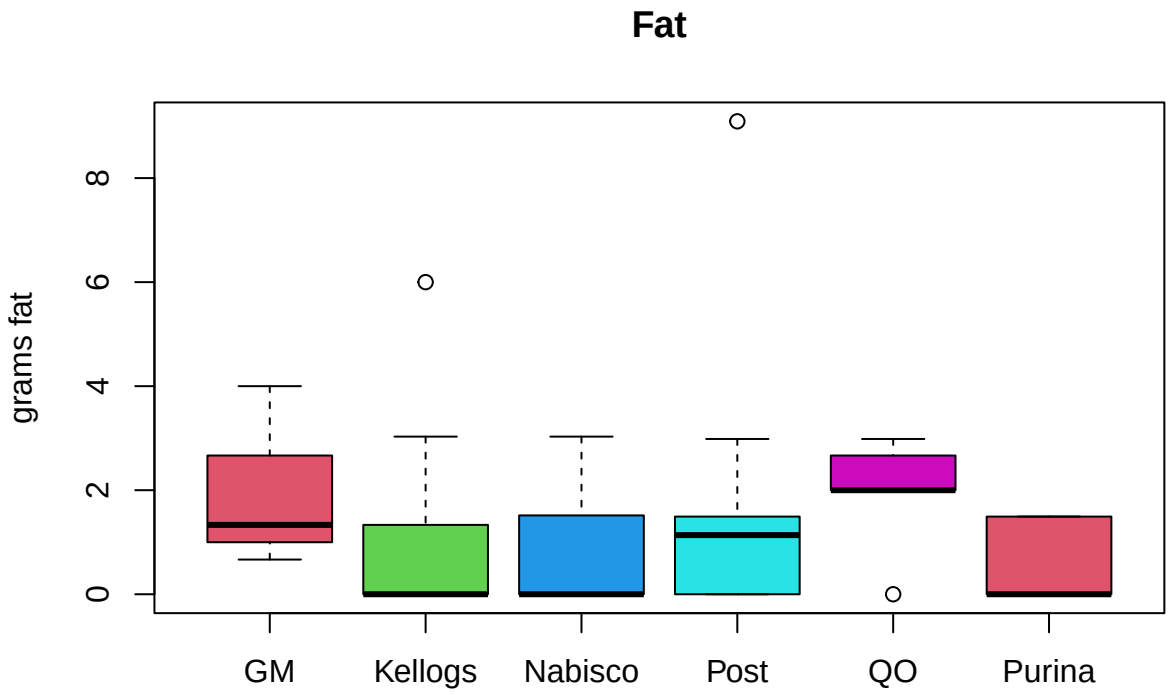
GM has lower caloric yet higher spread cereal.
Kellogs seems to have higher calorie cereal.

Problem 2 f)

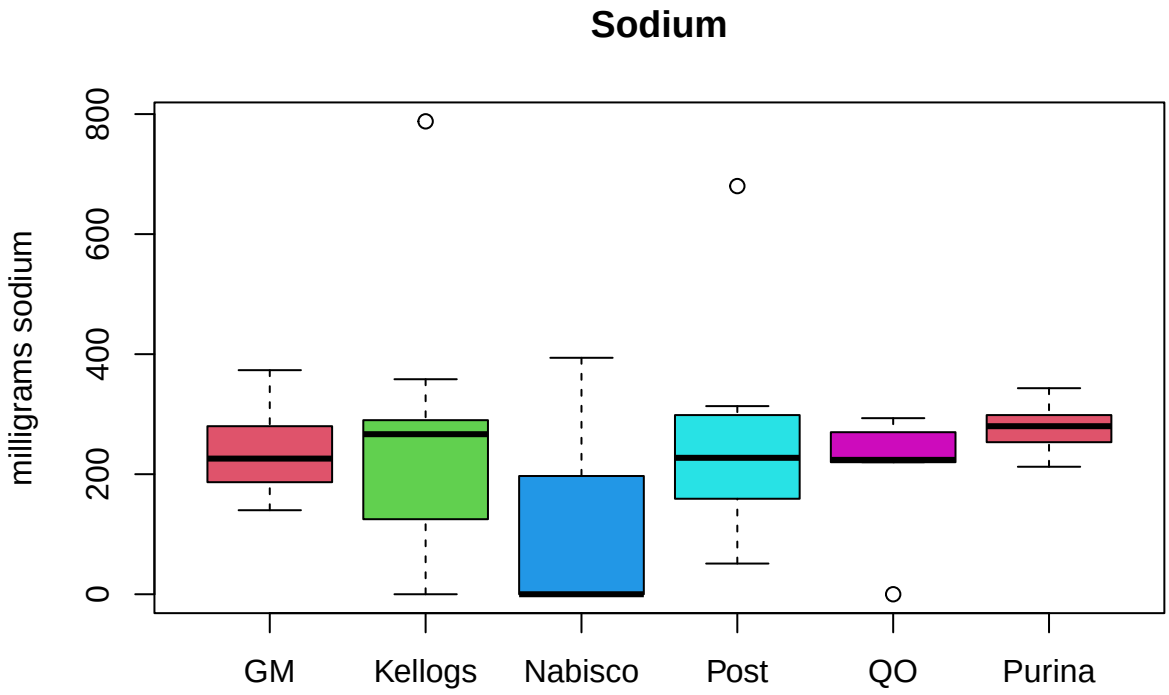
```
gm <- which(UScereal$mfr == "General Mills")
k <- which(UScereal$mfr == "Kellogs")
n <- which(UScereal$mfr == "Nabisco")
p <- which(UScereal$mfr == "Post")
q <- which(UScereal$mfr == "Quaker Oats")
r <- which(UScereal$mfr == "Ralston Purina")
boxplot(UScereal$protein[gm], UScereal$protein[k],
        UScereal$protein[n], UScereal$protein[p],
        UScereal$protein[q], UScereal$protein[r],
        col = 2:6, main= "Protein",
        names  = c("GM", "Kellogs", "Nabisco", "Post", "QO", "Purina"),
        ylab = "grams protein")
```



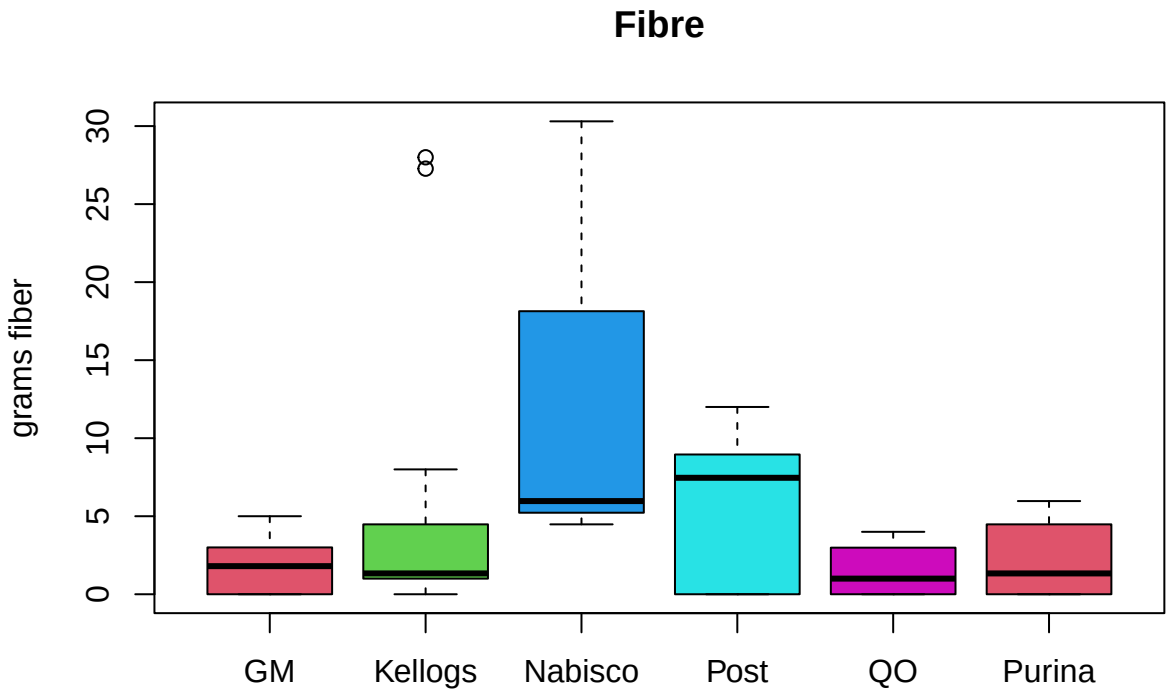

```
boxplot(UScereal$fat[gm], UScereal$fat[k], UScereal$fat[n],
        UScereal$fat[p], UScereal$fat[q], UScereal$fat[r],
        col = 2:6, main= "Fat",
        names  = c("GM", "Kellogs", "Nabisco", "Post", "QO", "Purina"),
        ylab = "grams fat")
```



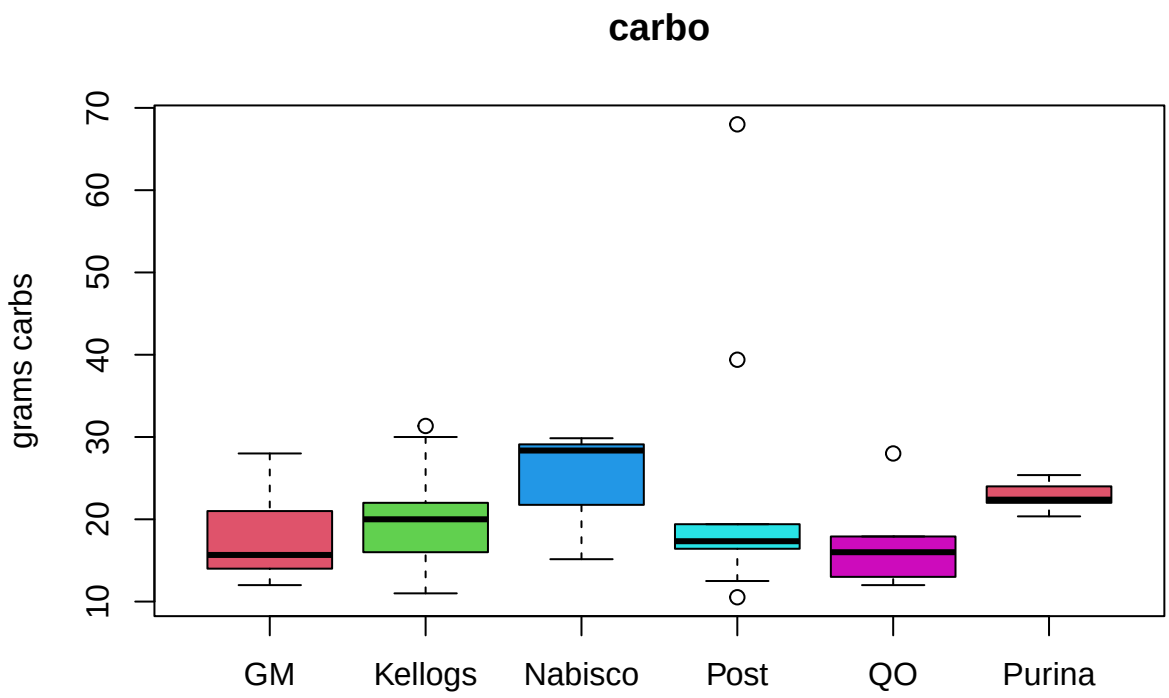

```
boxplot(UScereal$sodium[gm], UScereal$sodium[k],
        UScereal$sodium[n], UScereal$sodium[p],
        UScereal$sodium[q], UScereal$sodium[r],
        col = 2:6, main= "Sodium",
        names  = c("GM", "Kellogs", "Nabisco", "Post", "QO", "Purina"),
        ylab = "milligrams sodium")
```



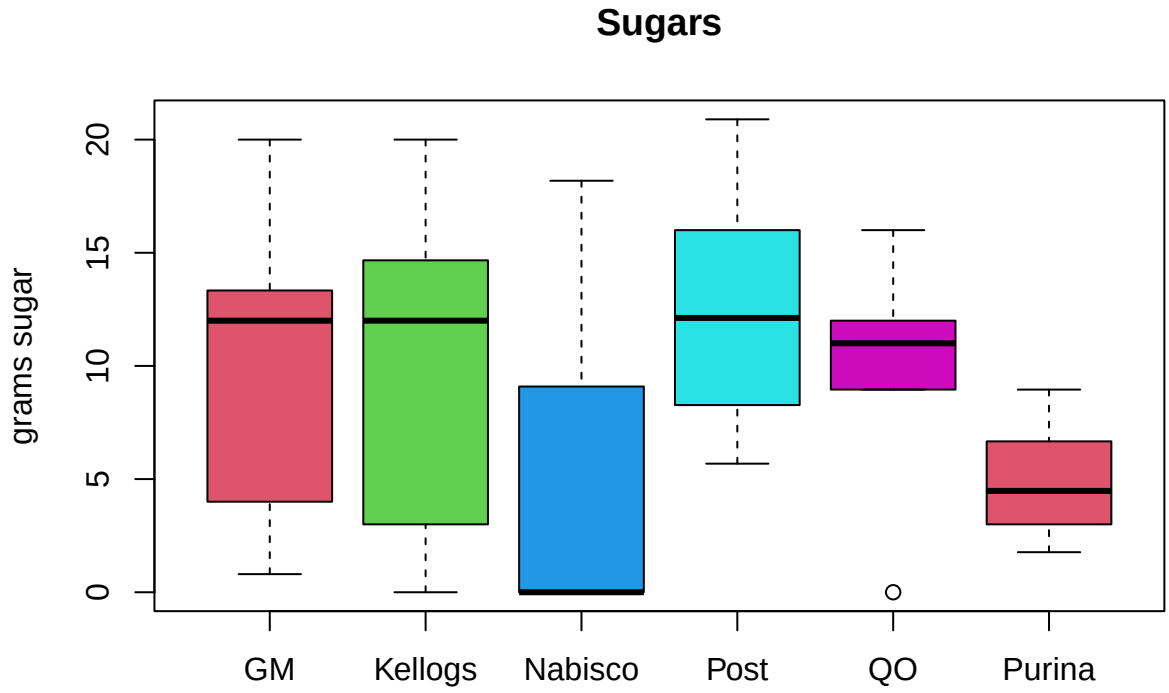

```
boxplot(UScereal$fibre[gm], UScereal$fibre[k], UScereal$fibre[n],
        UScereal$fibre[p],
        UScereal$fibre[q], UScereal$fibre[r],
        col = 2:6, main= "Fibre",
        names  = c("GM", "Kellogs", "Nabisco", "Post", "QO", "Purina"),
        ylab = "grams fiber")
```



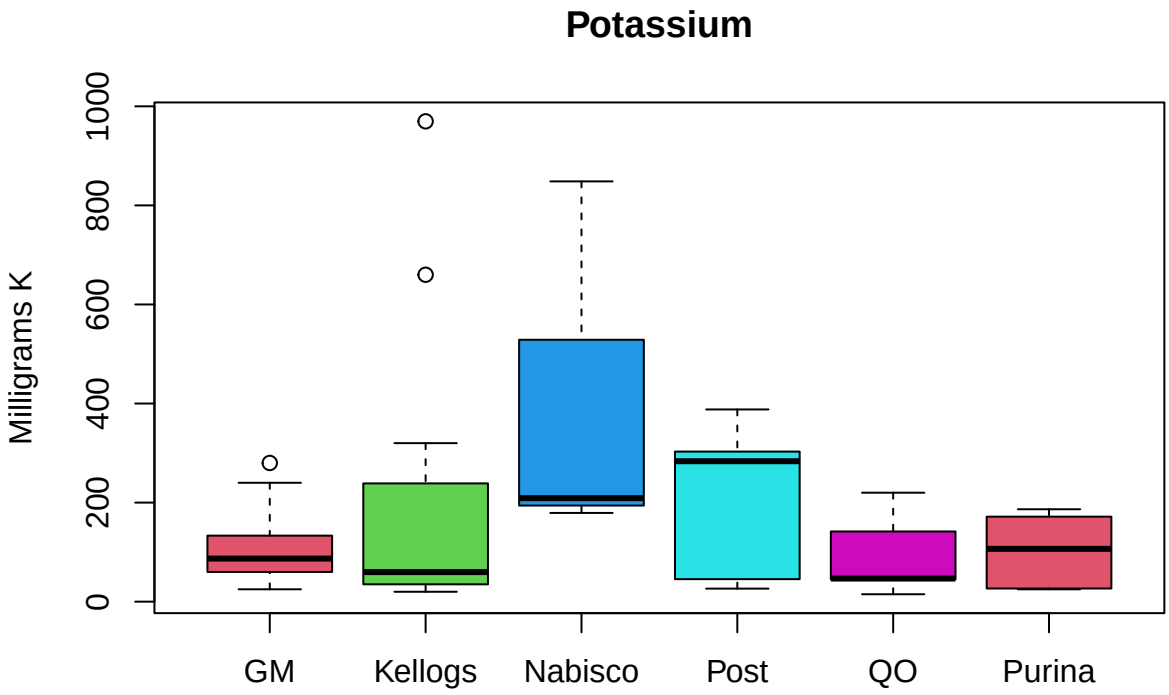

```
boxplot(UScereal$carbo[gm], UScereal$carbo[k],
        UScereal$carbo[n], UScereal$carbo[p],
        UScereal$carbo[q], UScereal$carbo[r],
        col = 2:6, main= "carbo",
        names  = c("GM", "Kellogs", "Nabisco", "Post", "QO", "Purina"),
        ylab = "grams carbs")
```




```
boxplot(UScereal$sugars[gm], UScereal$sugars[k],
        UScereal$sugars[n], UScereal$sugars[p],
        UScereal$sugars[q], UScereal$sugars[r],
        col = 2:6, main= "Sugars",
        names  = c("GM", "Kellogs", "Nabisco", "Post", "QO", "Purina"),
        ylab = "grams sugar")
```



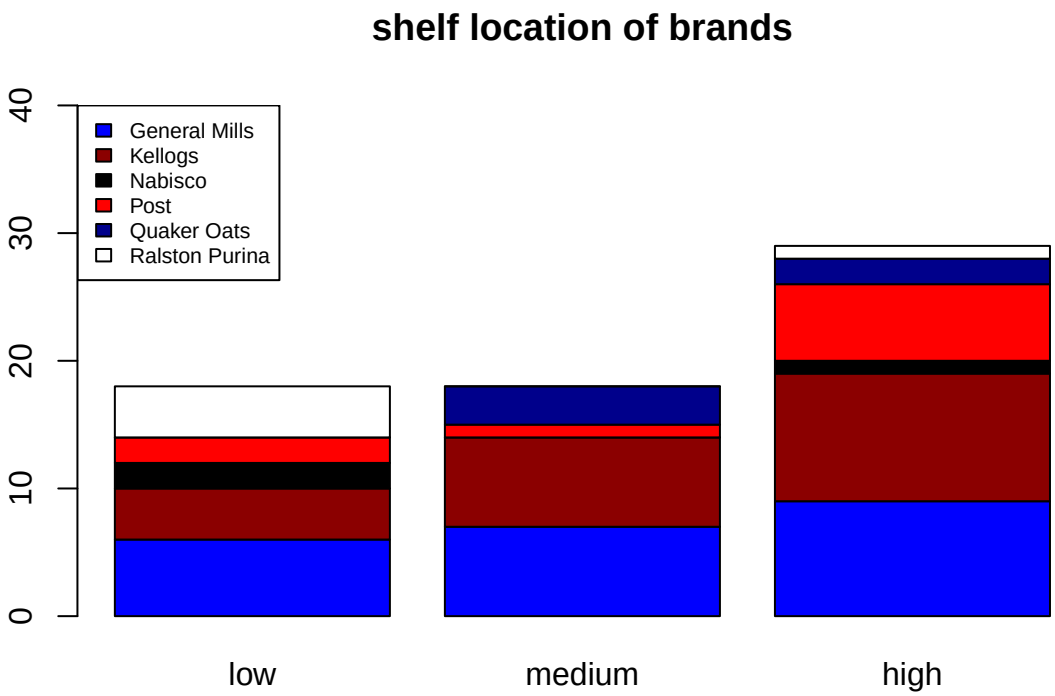

```
boxplot(UScereal$potassium[gm], UScereal$potassium[k],
        UScereal$potassium[n], UScereal$potassium[p],
        UScereal$potassium[q], UScereal$potassium[r],
        col = 2:6, main= "Potassium",
        names  = c("GM", "Kellogs", "Nabisco", "Post", "QO", "Purina"),
        ylab = "Milligrams K")
```



Based on these boxplot, it is clear that Nabisco has the healthiest cereal. High fiber, protein, potassium, but low sugar and sodium.
Average carbs and fat.

Problem 2 g)

```
barplot(table(UScereal$mfr, UScereal$shelf),
        col = c("blue", "darkred", "black", "red", "darkblue", "white"),
        ylim = c(0, 40),
        main = "shelf location of brands"
        )
legend("topleft",
       legend = c("General Mills", "Kellogs", "Nabisco", "Post",
                          "Quaker Oats", "Ralston Purina"),
       fill = c("blue", "darkred", "black", "red", "darkblue", "white"),
       cex = 0.65)
```


shelf location of brands
General Mills
Kellogs
Nabisco
Post
Quaker Oats
Ralston Purina
0
10
20
30
40
low
medium
high