Homework 4

Kevin Jin

3/10/2022

Problem 1

Part (a)

```
# Load data
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")
# Rename factor levels
levels(UScereal$mfr) <- c("General Mills",
                          "Kelloggs",
                          "Nabisco",
                          "Post",
                          "Quaker Oats",
                          "Ralston Purina")
```

Part (b)

```
# Factorize shelf variable
UScereal$shelf <- factor(UScereal$shelf,
                         labels = c("Lower", "Middle", "Upper"))
```

Part (c)

```
# Create product variable
library(dplyr)
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:Hmisc':
##
##     src, summarize
## The following object is masked from 'package:MASS':
##
##     select
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
UScereal <- mutate(UScereal, product = rownames(UScereal))
```

Problem 2

Part (a)

```
# Calculate correlation between calories and seven nutrition metrics
# Store Pearson coefficients in data frame for easier plotting
correlation <- data.frame(c("Protein", "Fat", "Sodium",
                            "Fibre", "Carbohydrates", "Sugars", "Potassium"),
           c(cor(x = UScereal$calories, y = UScereal$protein),
                 cor(x = UScereal$calories, y = UScereal$fat),
                 cor(x = UScereal$calories, y = UScereal$sodium),
                 cor(x = UScereal$calories, y = UScereal$fibre),
                 cor(x = UScereal$calories, y = UScereal$carbo),
                 cor(x = UScereal$calories, y = UScereal$sugars),
                 cor(x = UScereal$calories, y = UScereal$potassium)))
names(correlation) <- c("Variables", "Pearson")
correlation$Variables <- factor(correlation$Variables)
correlation
```

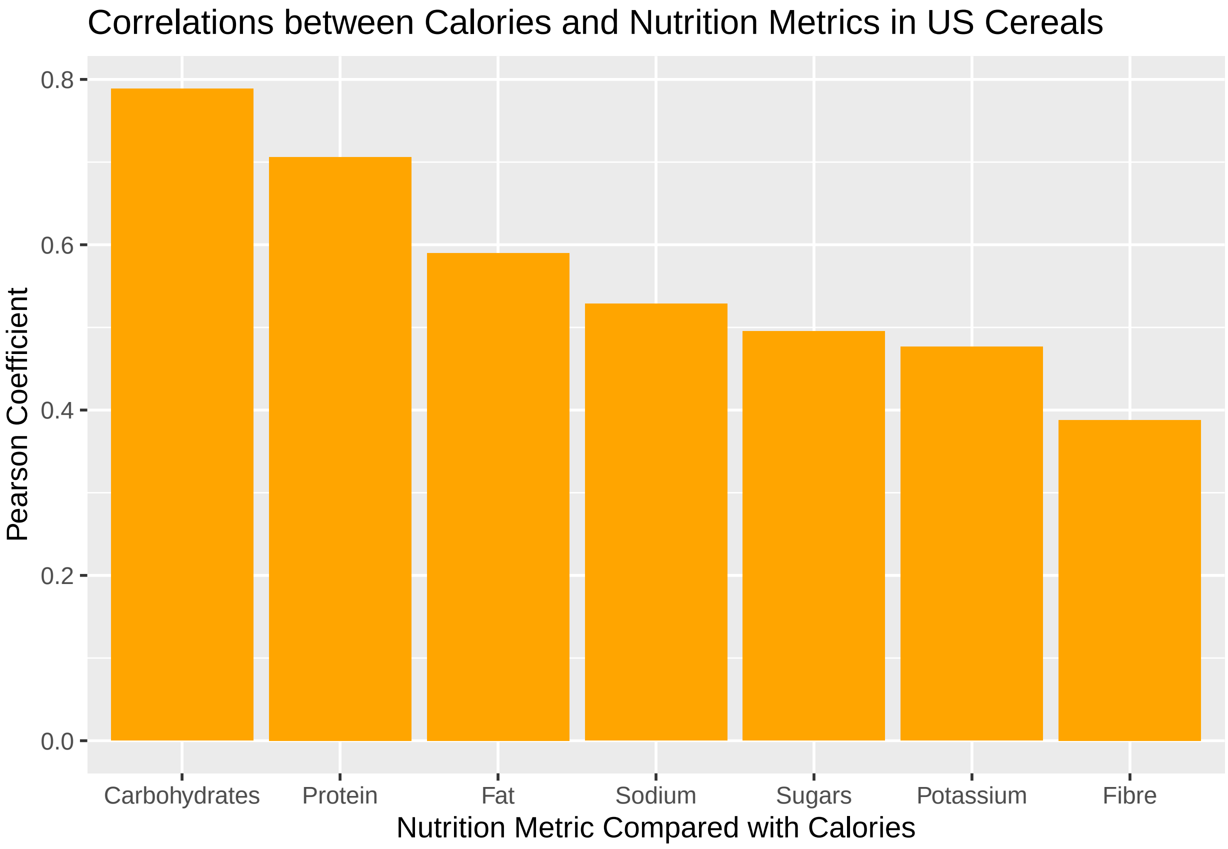
```
##       Variables   Pearson
## 1       Protein 0.7060105
## 2           Fat 0.5901757
## 3        Sodium 0.5286552
## 4         Fibre 0.3882179
## 5 Carbohydrates 0.7887227
## 6        Sugars 0.4952942
## 7     Potassium 0.4765955
```

Variables	Pearson Coefficient (2 d.p.)
Calories ~ Protein	0.71
Calories ~ Fat	0.59

Variables	Pearson Coefficient (2 d.p.)
Calories ~ Sodium	0.53
Calories ~ Fibre	0.39
Calories ~ Carbohydrates	0.79
Calories ~ Sugars	0.50
Calories ~ Potassium	0.48

Part (b)

```r
# Draw bar plot of correlations between calories and nutrition facts
ggplot(data = correlation, mapping = aes(x = reorder(Variables, -Pearson),
                                         y = Pearson)) +
  geom_bar(stat = "identity",
           fill = "orange") +
  ggtitle("Correlations between Calories and Nutrition Metrics in US Cereals") +
  xlab("Nutrition Metric Compared with Calories") +
  ylab("Pearson Coefficient")
```

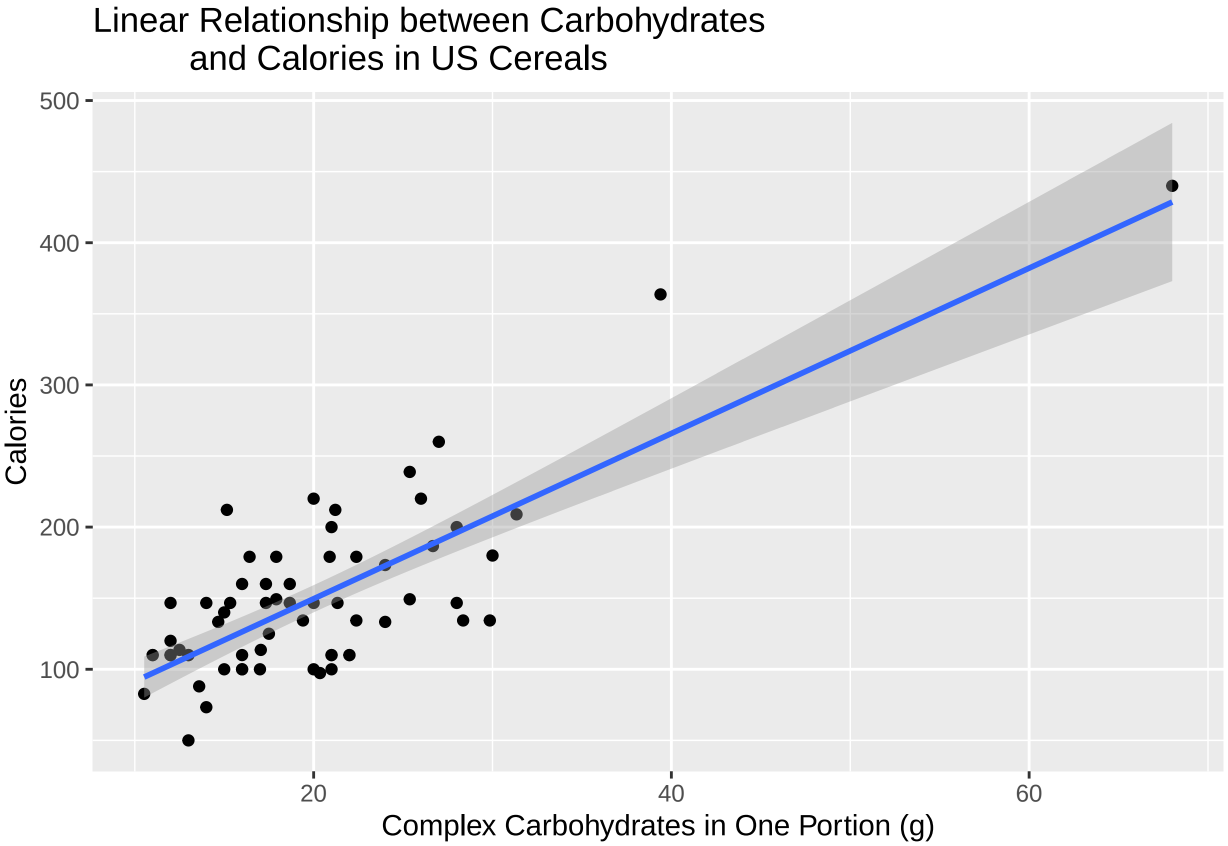


Conclusion: Carbohydrates have the highest positive correlation with calories.

Part (c)

```r
# Plot linear relationship between carbohydrates and calories
ggplot(data = UScereal, mapping = aes(x = carbo, y = calories)) +
  geom_point() +
  geom_smooth(method = "lm", fullrange = TRUE) +
  ggtitle("Linear Relationship between Carbohydrates
          and Calories in US Cereals") +
  xlab("Complex Carbohydrates in One Portion (g)") +
  ylab("Calories")
```

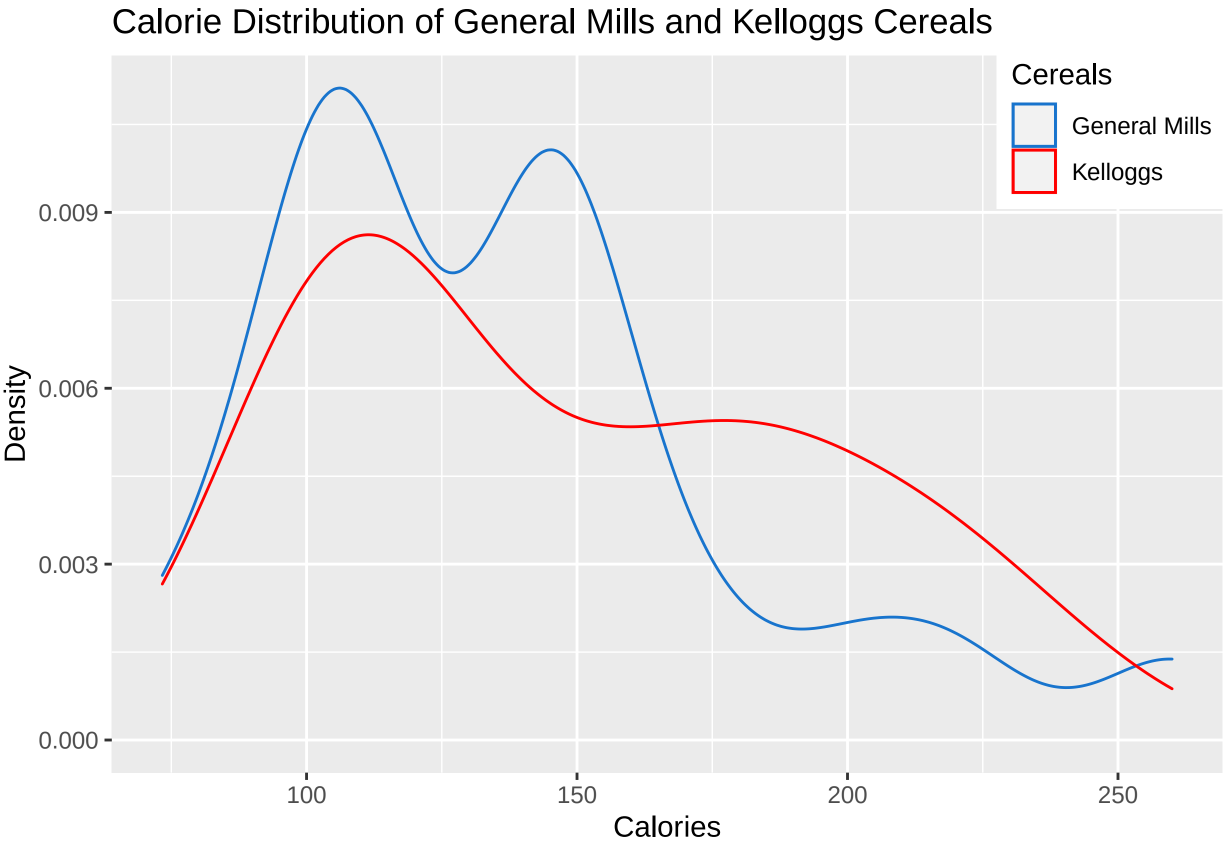
```
## `geom_smooth()` using formula 'y ~ x'
```



Conclusion: In this plot, the slope represents number of calories per gram of complex carbohydrates in one portion of cereal, and the intercept represents the average number of calories in one portion with zero grams of carbohydrates.

Part (d)

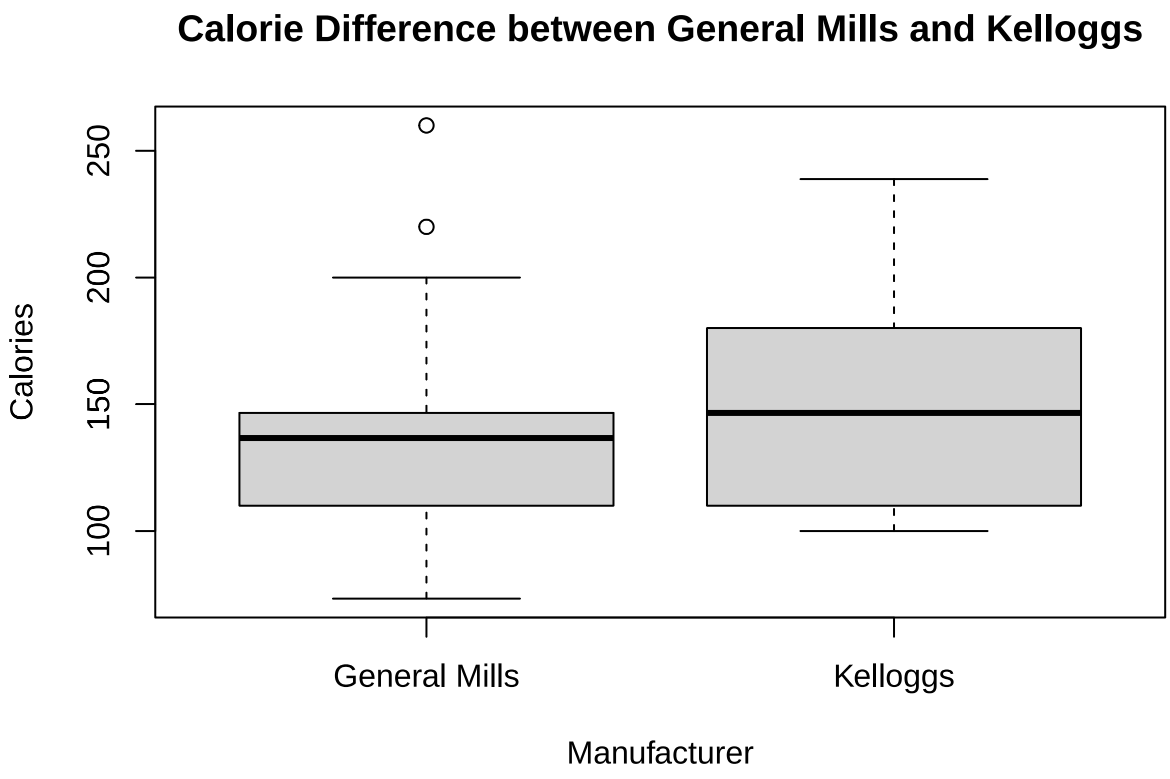
```
# Extract General Mills and Kelloggs rows (cereals)
gm <- filter(UScereal, mfr == "General Mills")
kel <- filter(UScereal, mfr == "Kelloggs")

# Draw density plot
ggplot() +
  geom_density(data = gm,
               mapping = aes(x = calories, color = "General Mills")) +
  geom_density(data = kel,
               mapping = aes(x = calories, color = "Kelloggs")) +
  ggtitle("Calorie Distribution of General Mills and Kelloggs Cereals") +
  xlab("Calories") +
  ylab("Density") +
  theme(legend.position = c(0.9, 0.9)) +
  scale_color_manual(values = c("dodgerblue3", "red")) +
  guides(color = guide_legend("Cereals"))
```



Conclusion: The General Mills density curve is left-skewed with a bimodal distribution, suggesting a lower number of calories in general, while the Kelloggs curve is also left-skewed but less steep, also suggesting a lower number of calories but with less skew. Both curves are asymmetric.

Part (e)

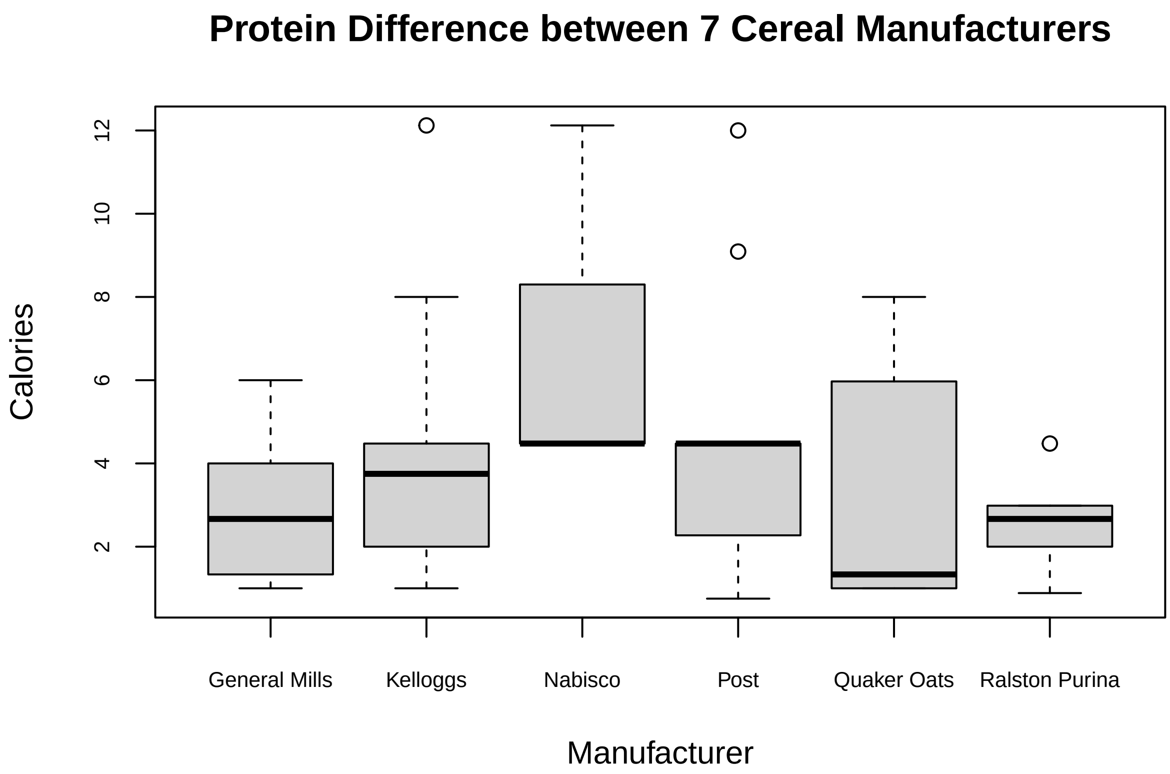
```
gm_cal <- gm$calories
kel_cal <- kel$calories
length(kel_cal) <- length(gm_cal) # Pad shorter dataframe to combine

# Boxplot to show significant difference in calories between GM and Kelloggs
boxplot(gm_cal, kel_cal,
        names = c("General Mills", "Kelloggs"),
        main = "Calorie Difference between General Mills and Kelloggs",
        xlab = "Manufacturer",
        ylab = "Calories")
```



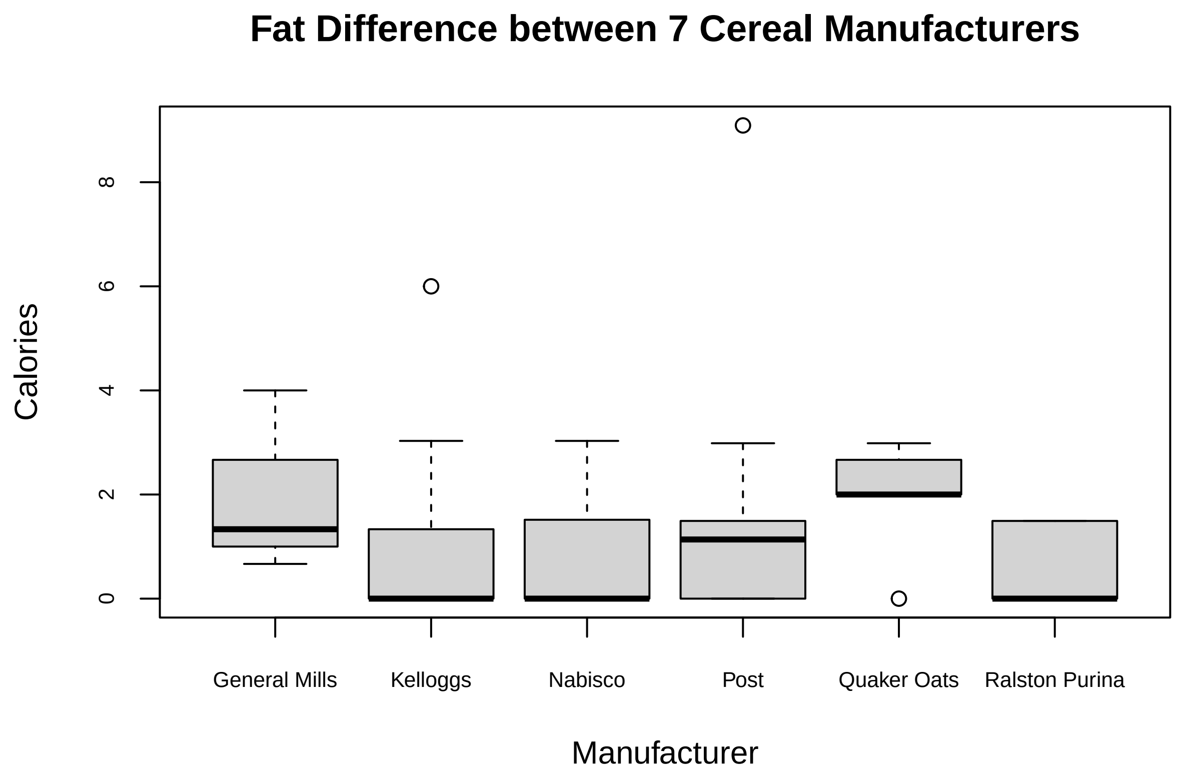
Conclusion: There does appear to be a slight difference in median calorie content between General Mills and Kelloggs and a larger IQR for Kelloggs, but not a significant difference in calorie distribution overall considering the overlap in distributions.

Part (f)

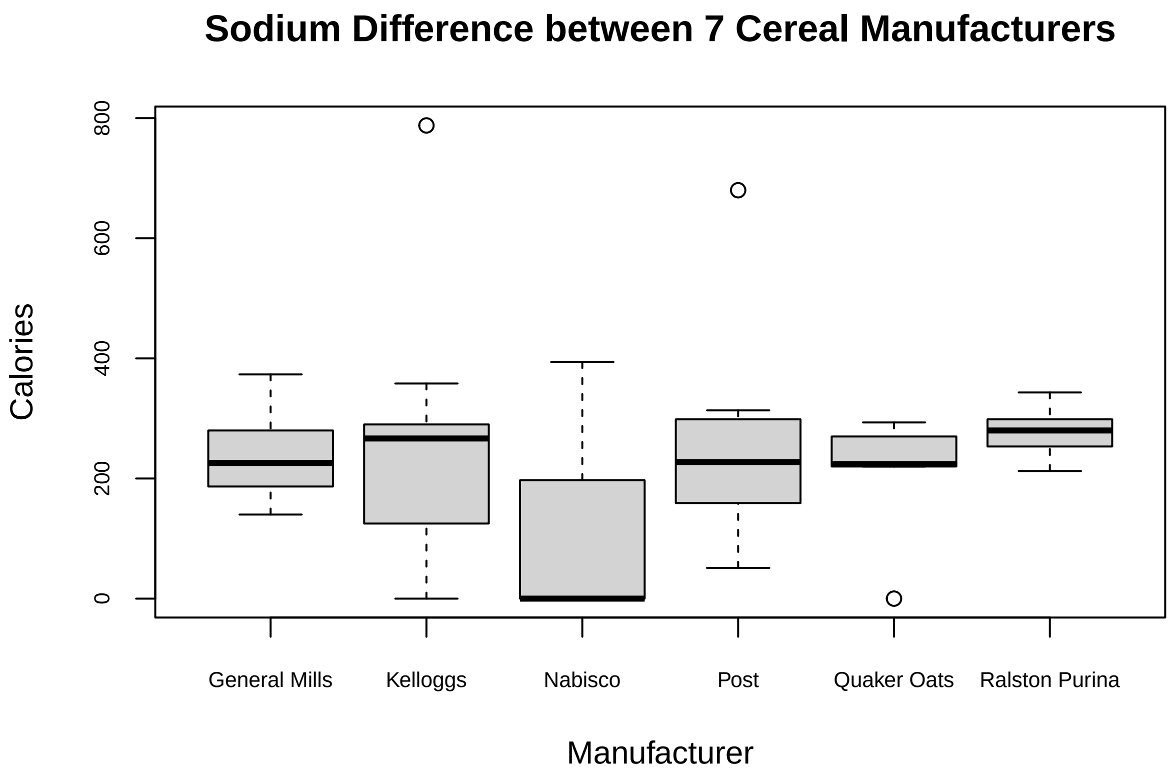
```
# Side by side box plots for each nutritional metric
# Protein
boxplot(UScereal$protein ~ UScereal$mfr,
        main = "Protein Difference between 7 Cereal Manufacturers",
        xlab = "Manufacturer",
        ylab = "Calories",
        cex.axis = 0.65)
```



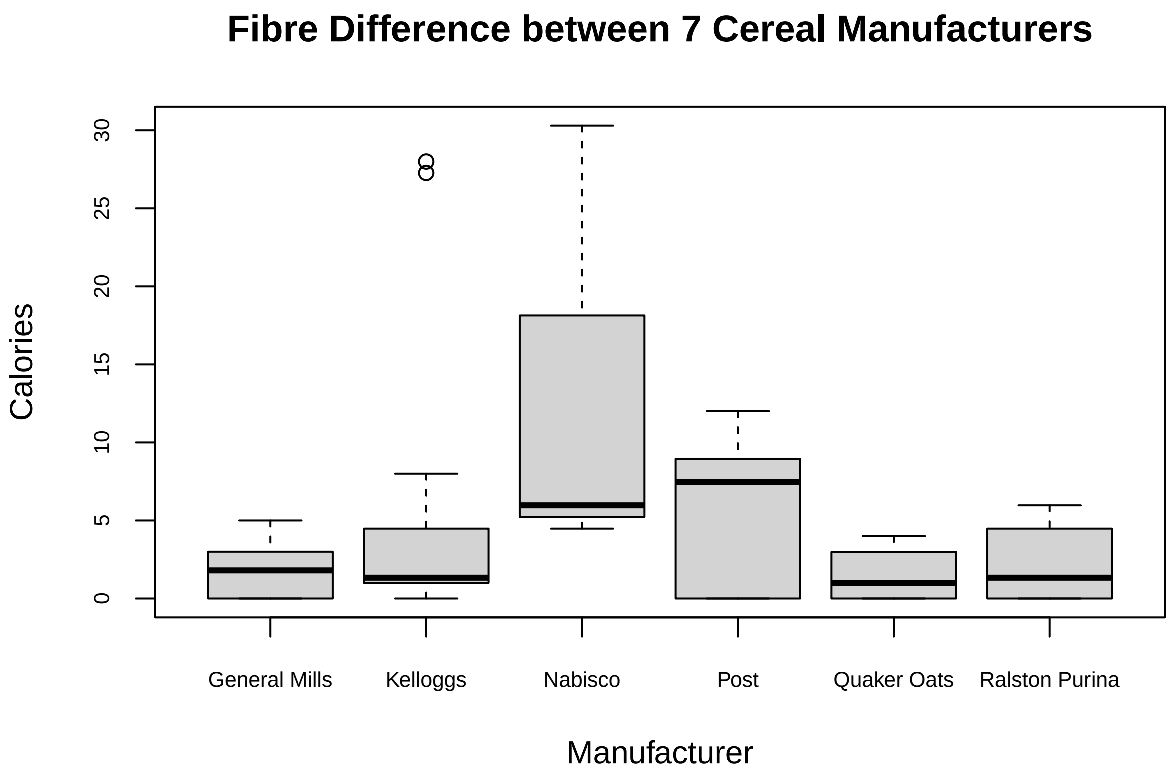
```
# Fat
boxplot(UScereal$fat ~ UScereal$mfr,
        main = "Fat Difference between 7 Cereal Manufacturers",
        xlab = "Manufacturer",
        ylab = "Calories",
        cex.axis = 0.65)
```



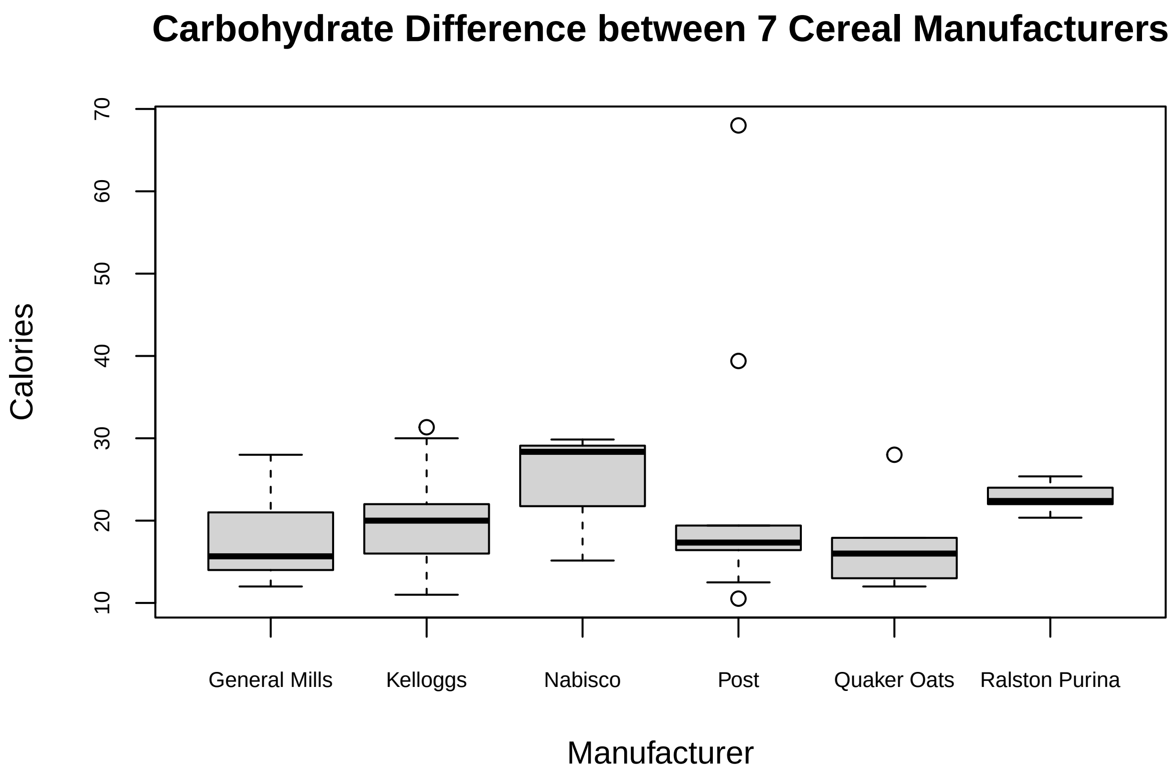
```r
# Sodium
boxplot(UScereal$sodium ~ UScereal$mfr,
        main = "Sodium Difference between 7 Cereal Manufacturers",
        xlab = "Manufacturer",
        ylab = "Calories",
        cex.axis = 0.65)
```



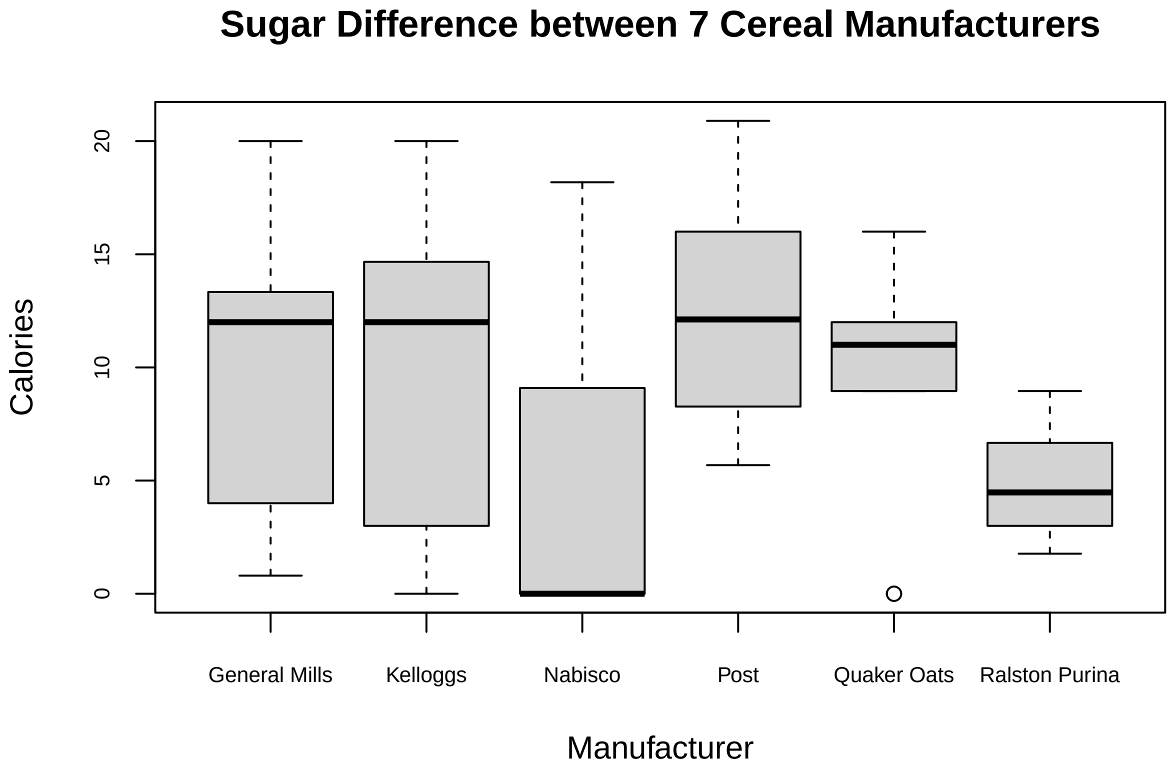
```
# Fibre
boxplot(UScereal$fibre ~ UScereal$mfr,
        main = "Fibre Difference between 7 Cereal Manufacturers",
        xlab = "Manufacturer",
        ylab = "Calories",
        cex.axis = 0.65)
```



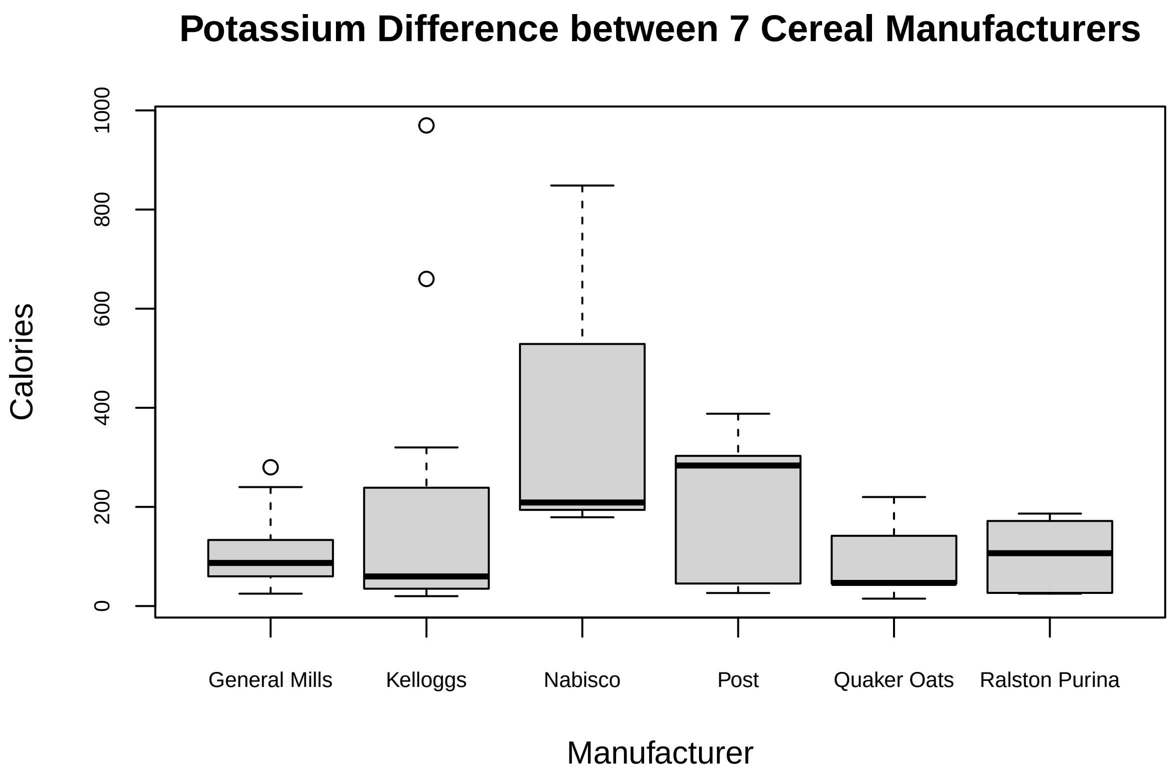
```r
# Carbo
boxplot(UScereal$carbo ~ UScereal$mfr,
        main = "Carbohydrate Difference between 7 Cereal Manufacturers",
        xlab = "Manufacturer",
        ylab = "Calories",
        cex.axis = 0.65)
```



```
# Sugars
boxplot(UScereal$sugars ~ UScereal$mfr,
        main = "Sugar Difference between 7 Cereal Manufacturers",
        xlab = "Manufacturer",
        ylab = "Calories",
        cex.axis = 0.65)
```



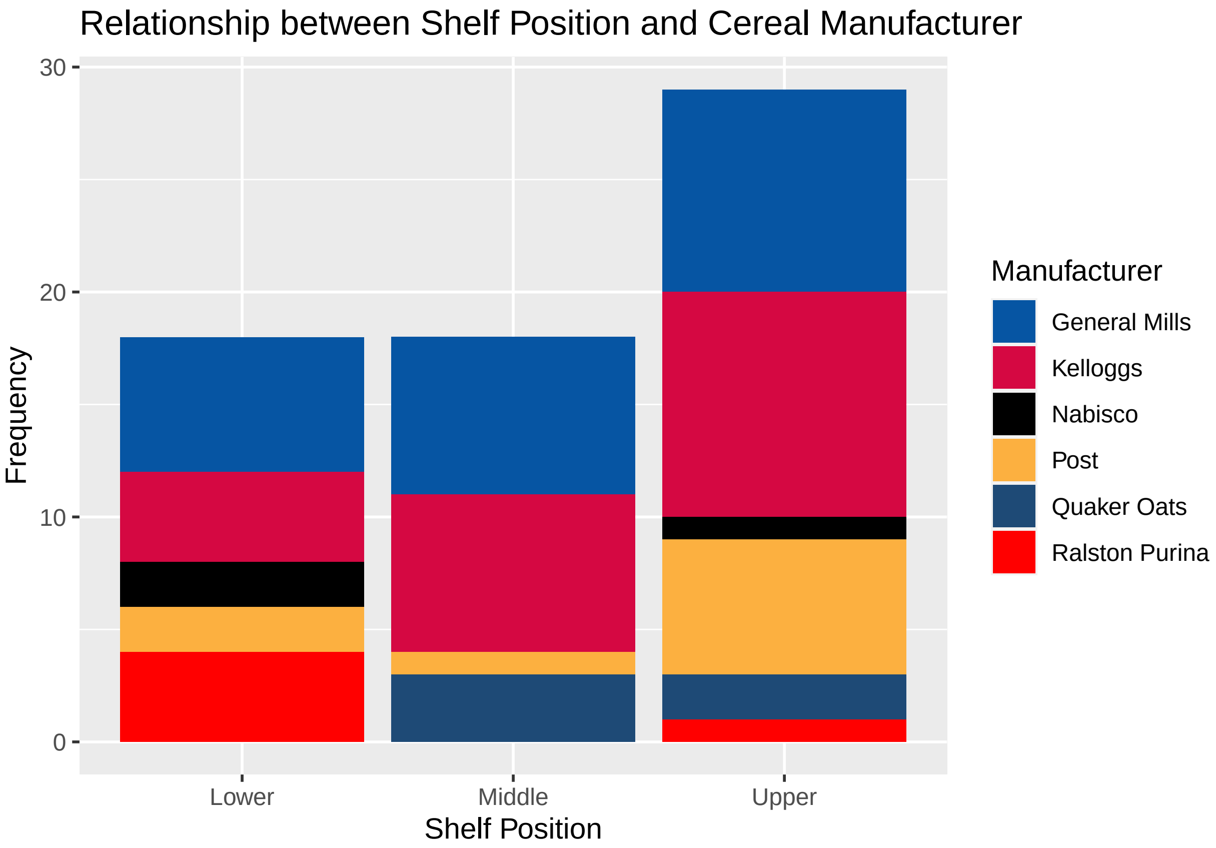
```
# Potassium
boxplot(UScereal$potassium ~ UScereal$mfr,
        main = "Potassium Difference between 7 Cereal Manufacturers",
        xlab = "Manufacturer",
        ylab = "Calories",
        cex.axis = 0.65)
```



Conclusion: Based on its high protein content, low fat, low sodium, high fibre, high carbohydrates, low sugar, and high potassium, out of all cereal manufacturers, Nabisco is aiming to make the healthiest cereals.

Part (g)

```
# Draw stacked bar plot showing relationship between shelf position and mfr
ggplot(UScereal) +
  geom_bar(aes(x = shelf, fill = mfr)) +
  ggtitle("Relationship between Shelf Position and Cereal Manufacturer") +
  xlab("Shelf Position") +
  ylab("Frequency") +
  guides(fill=guide_legend(title="Manufacturer")) +
  scale_fill_manual(values = c("#0655A3", "#D50842", "black", "#FCB040",
                               "#1E4A76", "red"))
```



Conclusion: General Mills and Kelloggs occupy most of the shelf space among all three shelf positions. The middle shelf has the narrowest distribution of brands.