Cereals

EB04 - Group 2

Zhanica Arrindell (596427) Vera Gak Anagrova (772713)
Aleksandra Tatko (648925) Ly Le (644801)

2025-09-28

Contents

1 Introduction **1**

2 Cereal Ratings by Calorie Level **2**

3 Rating vs Sugar & Rating vs Calories **3**

4 Comparison two main manufactures **4**

5 Apendix **5**

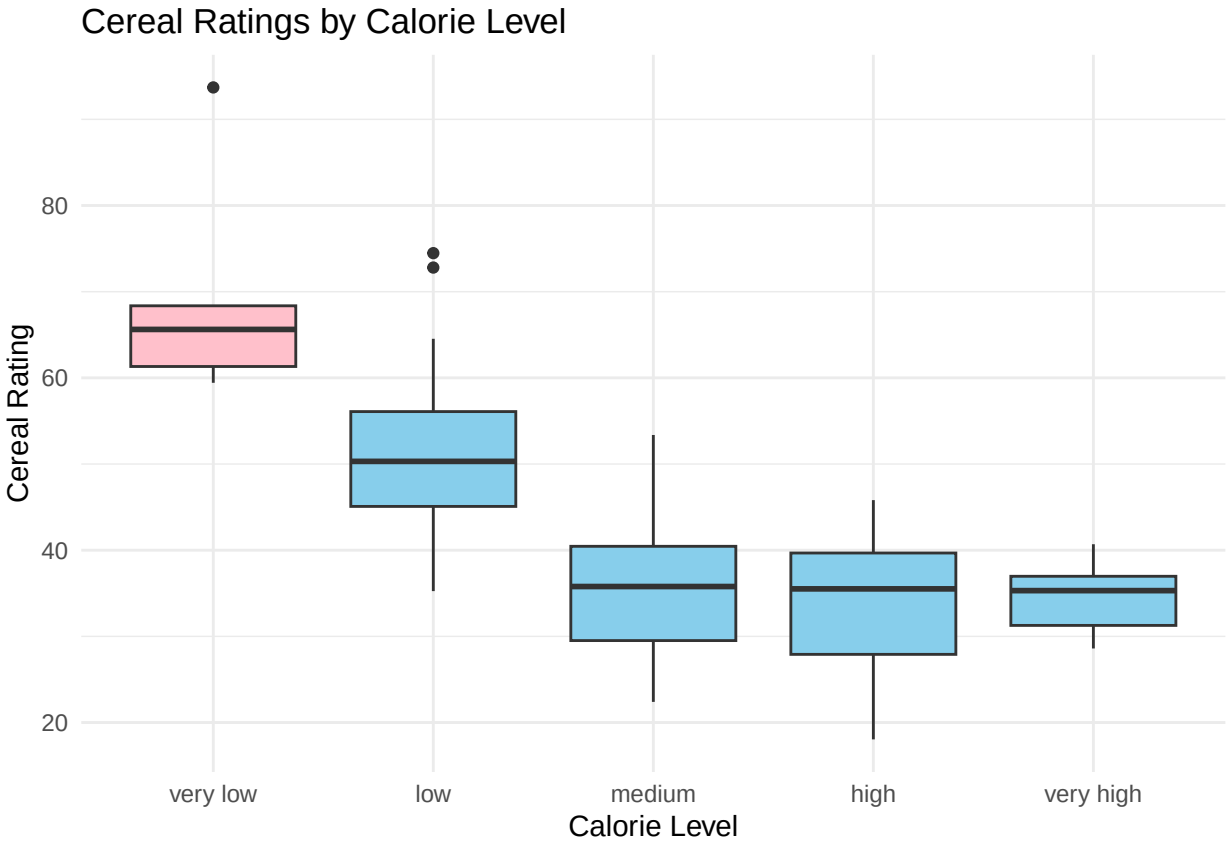
1 Introduction

This report presents the analysis of the Cereal dataset, following tasks A–C from the assignment instructions. All plots are included in the main body, while the full R code can be found in the appendix.

2 Cereal Ratings by Calorie Level

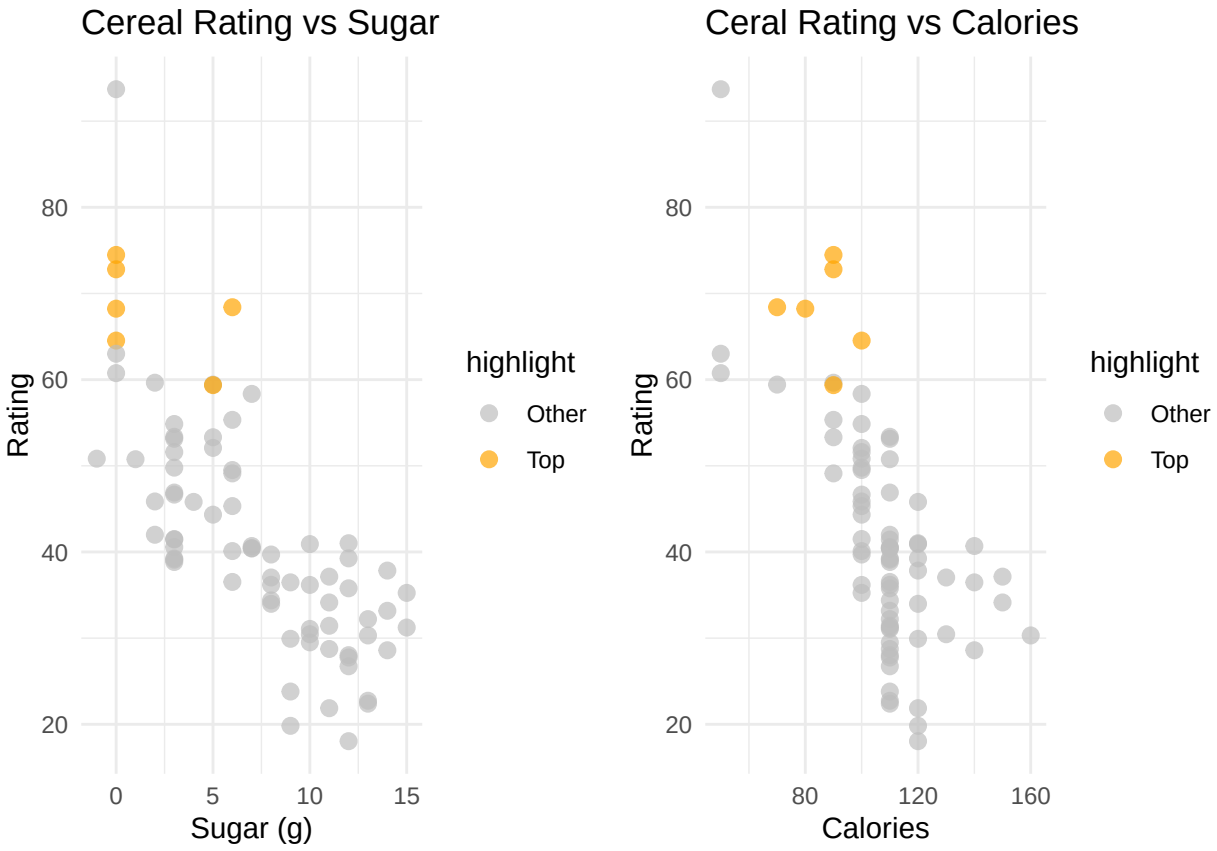
We created an ordered factor named cal_level with five levels that groups the cereals based on calories per serving, being: - very low: 80 or less calories per serving - low: 81-100 calories per serving - medium: 101-110 calories per serving - high: 111-130 calories per serving - very high: more than 130 calories per serving

We created a boxplot with the ratings for each of the five calorie levels with the levels ordered from "very low" to "very high". We customized the plot having the highest rated box displayed in a color different from the other boxes.



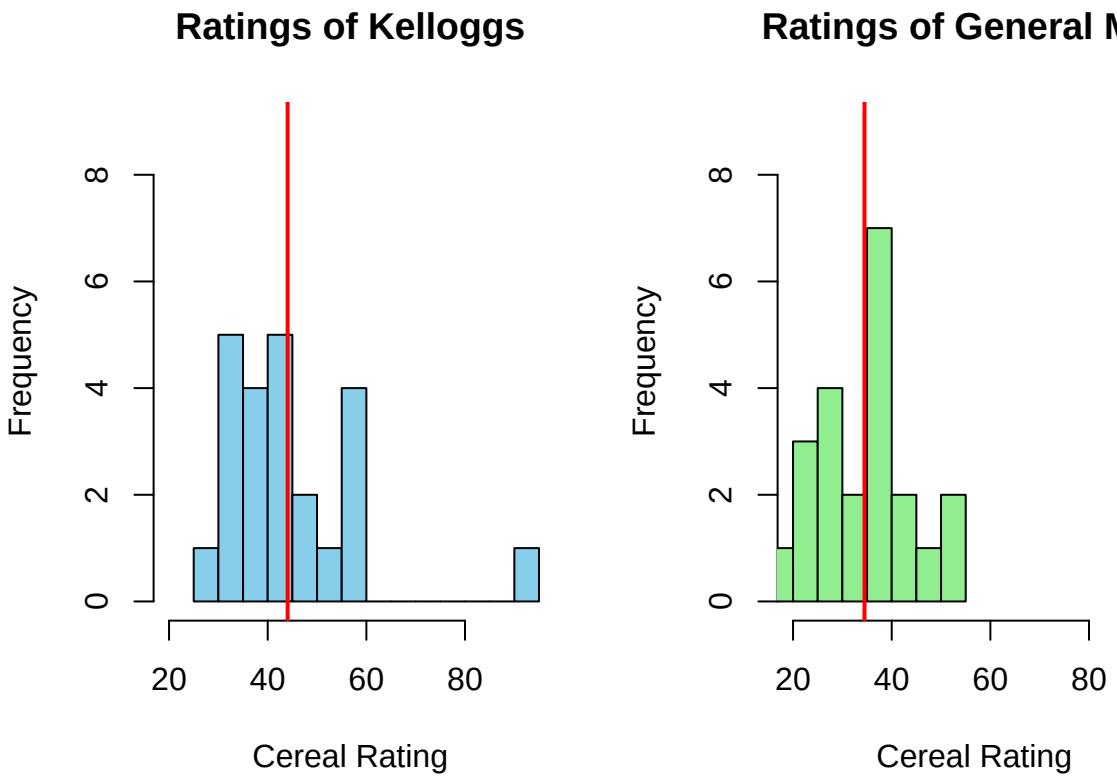
3 Rating vs Sugar & Rating vs Calories

We created a figure with two scatter plots next to each other to investigate the relation between rating and nutritional value: - The first plot: shows the cereals ratings versus sugar quantity per serving - The second plot: one show the ratings versus number of calories per serving. The scatter plots includes all 77 cereals. Cereals from the highest rated manufacturer are highlighted by a different color.



4 Comparison two main manufactures

We compares the cereals of two main manufacturers. We found two manufacturers with the most cereals. We created one figure with two plots next to each other: - The first shows a histogram of ratings of the first manufacturer's cereals - The second shows a histogram of ratings for the second manufacturer's cereals. The mean rating in both plots by a vertical, red line.



5 Apendix

```r
#get data
library(readr)
library(ggplot2)
library(gridExtra)
Cereals <- read_csv("Cereals.csv")
str(Cereals)       # structure of the dataset



# Question 1
# Assign text categories based on calories
cal_text <- ifelse(Cereals$calories <= 80, "very low",
                   ifelse(Cereals$calories >= 81 & Cereals$calories <= 100, "low",
                          ifelse(Cereals$calories >= 101 & Cereals$calories <= 110, "medium",
                                 ifelse(Cereals$calories >= 111 & Cereals$calories <= 130, "high",
                                        "very high"))))

# Convert to an ordered factor and add column to data frame
Cereals$cal_level <- factor(cal_text,
                            levels = c("very low", "low", "medium", "high", "very high"),
                            ordered = TRUE)

# Average rating per calorie level
avg_ratings <- aggregate(rating ~ cal_level, data = Cereals, mean)

# Identify the level with the highest average rating
top_level <- avg_ratings$cal_level[which.max(avg_ratings$rating)]

# Define colors
box_colors <- rep("skyblue", length(levels(Cereals$cal_level)))
box_colors[which(levels(Cereals$cal_level) == top_level)] <- "pink"

# Plot
ggplot(Cereals, aes(x = cal_level, y = rating, fill = cal_level)) +
  geom_boxplot() +
  scale_fill_manual(values = box_colors) +
  labs(title = "Cereal Ratings by Calorie Level",
       x = "Calorie Level",
       y = "Cereal Rating") +
  theme_minimal() +
  theme(legend.position = "none")


# Question 2

# Calculate mean rating per manufacturer
mean_ratings <- tapply(Cereals$rating, Cereals$mfr, mean, na.rm = TRUE)
# Identify top manufacturer
top_mfr <- with(Cereals, names(sort(tapply(rating, mfr, mean), decreasing = TRUE))[1])
Cereals$highlight <- ifelse(Cereals$mfr == top_mfr, "Top", "Other")
```

```r
# Define color palette
colors <- c("Top" = "orange", "Other" = "gray")

# Create plots
p1 <- ggplot(Cereals, aes(sugars, rating, color = highlight)) +
  geom_point(size = 2.5, alpha = 0.7) +
  scale_color_manual(values = colors) +
  labs(title = "Cereal Rating vs Sugar", x = "Sugar (g)", y = "Rating") +
  theme_minimal()

p2 <- ggplot(Cereals, aes(calories, rating, color = highlight)) +
  geom_point(size = 2.5, alpha = 0.7) +
  scale_color_manual(values = colors) +
  labs(title = "Ceral Rating vs Calories", x = "Calories", y = "Rating") +
  theme_minimal()

# Display side by side
grid.arrange(p1, p2, ncol = 2)


# Question 3
# Count cereals per manufacturer
counts <- table(Cereals$mfr)
# Sort descending and get the top 2
top2 <- names(sort(counts, decreasing = TRUE))[1:2]
top2
# ratings of this manufacturers
ratings1 <- Cereals$rating[Cereals$mfr == top2[1]]
ratings2 <- Cereals$rating[Cereals$mfr == top2[2]]
# Common x-axis: min and max ratings
xrange <- range(c(ratings1, ratings2))
# Common y-axis: maximum histogram count
yrange <- range(0, max(hist(ratings1, plot = FALSE)$counts,
                       hist(ratings2, plot = FALSE)$counts))

par(mfrow = c(1, 2), mar = c(5, 4, 4, 2))  # 1 row, 2 columns, adjust margins

hist(ratings1,
     breaks = 10,                  # number of bins
     xlim = xrange,
     ylim = yrange,
     col = "skyblue",
     main = "Ratings of Kelloggs",
     xlab = "Cereal Rating",
     ylab = "Frequency")
abline(v = mean(ratings1), col = "red", lwd = 2)  # vertical line at mean

# Histogram for second manufacturer
hist(ratings2,
     breaks = 10,
     xlim = xrange,
     ylim = yrange,
     col = "lightgreen",
```

```
     main = "Ratings of General Mills",
     xlab = "Cereal Rating",
     ylab = "Frequency")
abline(v = mean(ratings2), col = "red", lwd = 2)  # vertical line at mean
```