Assignment 5: Cereals

Nicolas (780037), Maggie (755273), Avel (781169)

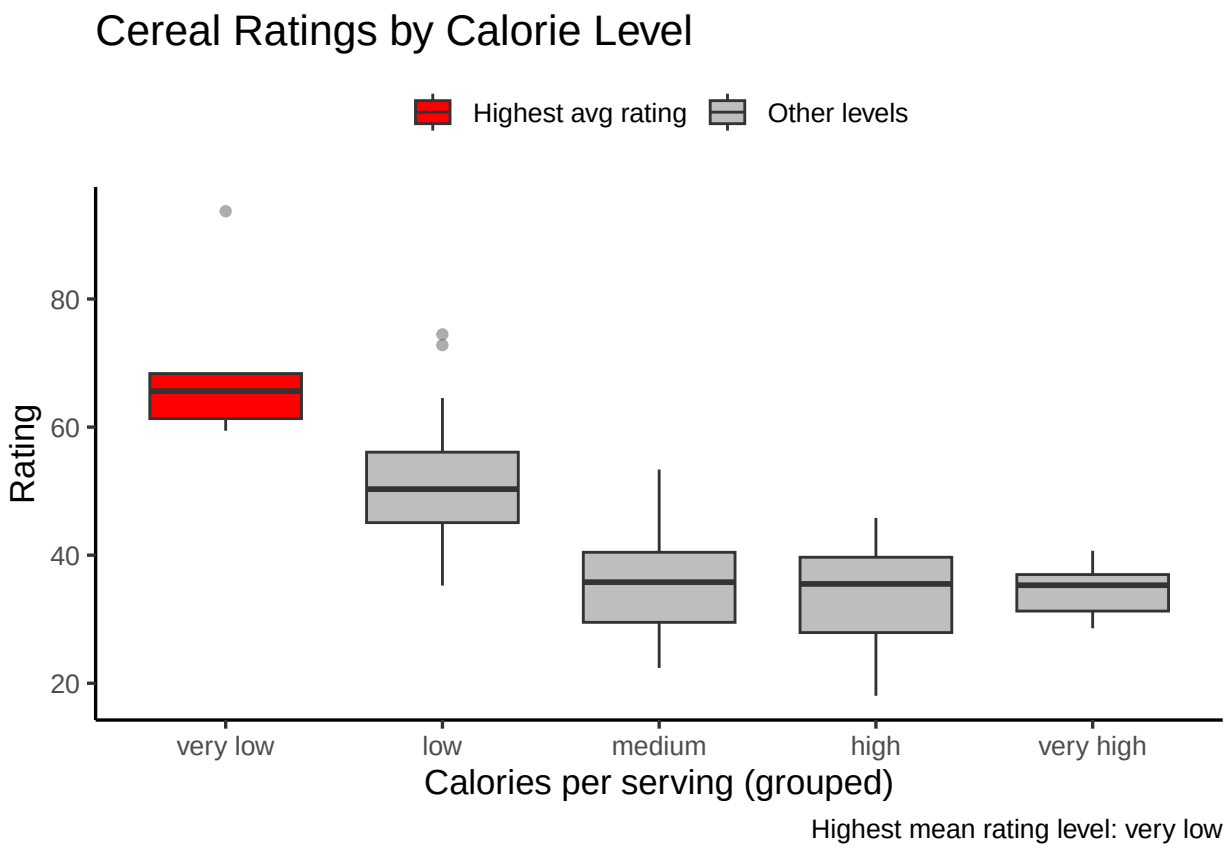
Sep 31, 2025

```
## [1] "/Users/nico/Documents/EUR/Programming/Assignments/Group Assignment/Programming-GroupAssignments"
```

Task A

```
Cereals <- data%>%
  mutate(
    cal_level = case_when(
      calories <= 80 ~ "very low",
      calories <= 100 ~ "low",
      calories <= 110 ~ "medium",
      calories <= 130 ~ "high",
      calories > 130 ~ "very high"
    ),
    cal_level = factor(
      cal_level,
      levels = c("very low", "low", "medium", "high", "very high"),
      ordered = TRUE
    )
  )

Cereals <- Cereals %>%
  mutate(manufacturer = recode(mfr,
                           "A" = "American Home Food Products",
                           "G" = "General Mills",
                           "K" = "Kellogg's",
                           "N" = "Nabisco",
                           "P" = "Post",
                           "Q" = "Quaker Oats",
                           "R" = "Ralston Purina"))

top_level <- Cereals %>%
  group_by(cal_level) %>%
  summarize(mean_rating = mean(rating, na.rm = TRUE), .groups = "drop") %>%
  slice_max(mean_rating, n = 1, with_ties = FALSE) %>%
  pull(cal_level)

Cereals <- Cereals %>%
  mutate(
    highest_group = if_else(cal_level == top_level,
                            "Highest avg rating", "Other levels")
  )
```

```r
# Boxplot
ggplot(Cereals, aes(x = cal_level, y = rating, fill = highest_group)) +
  geom_boxplot(width = 0.7, outlier.alpha = 0.4) +
  scale_fill_manual(
    values = c("Other levels" = "gray",
               "Highest avg rating" = "red"),
    guide = guide_legend(title = NULL)
  ) +
  labs(
    title = "Cereal Ratings by Calorie Level",
    x = "Calories per serving (grouped)",
    y = "Rating",
    caption = paste("Highest mean rating level:", top_level)
  ) +
  theme_classic(base_size = 13) +
  theme(
    legend.position = "top",
    panel.grid.minor = element_blank()
  )
```



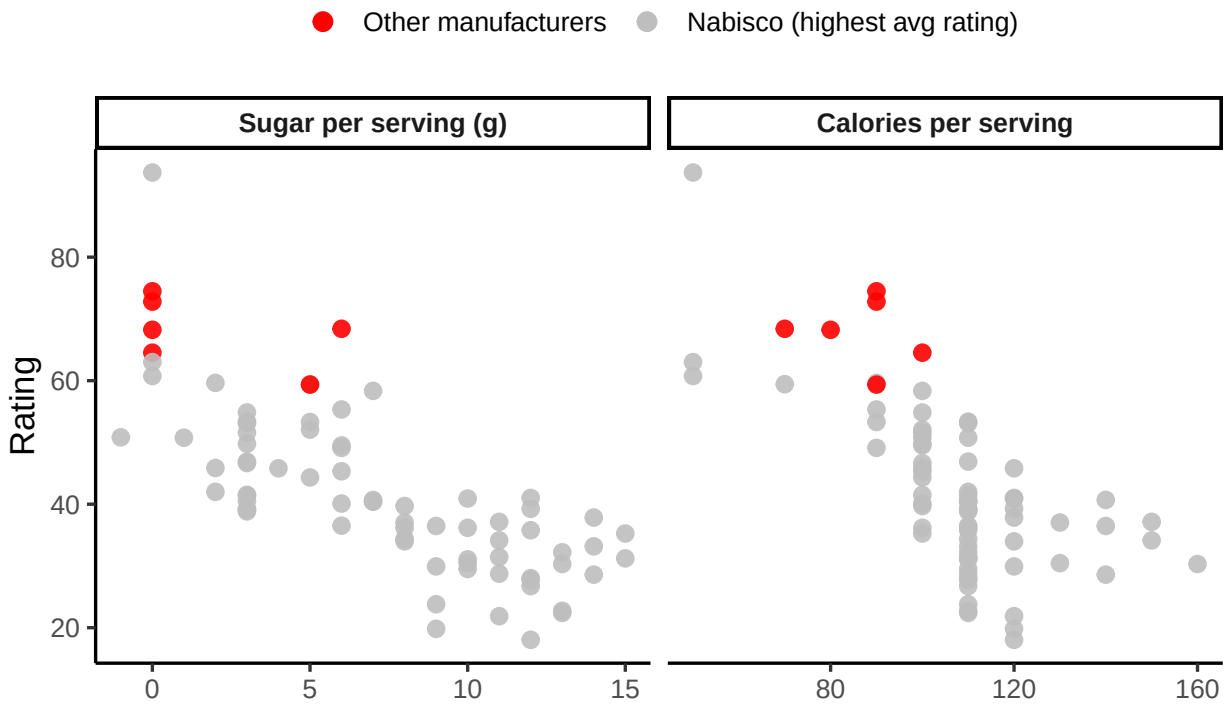

```r
colnames(Cereals)
```

```
##  [1] "name"          "mfr"           "type"          "calories"
##  [5] "protein"       "fat"           "sodium"        "fiber"
##  [9] "carbo"         "sugars"        "potass"        "vitamins"
```

```
## [13] "shelf"         "weight"        "cups"          "rating"
## [17] "cal_level"     "manufacturer"  "highest_group"
```

Task B

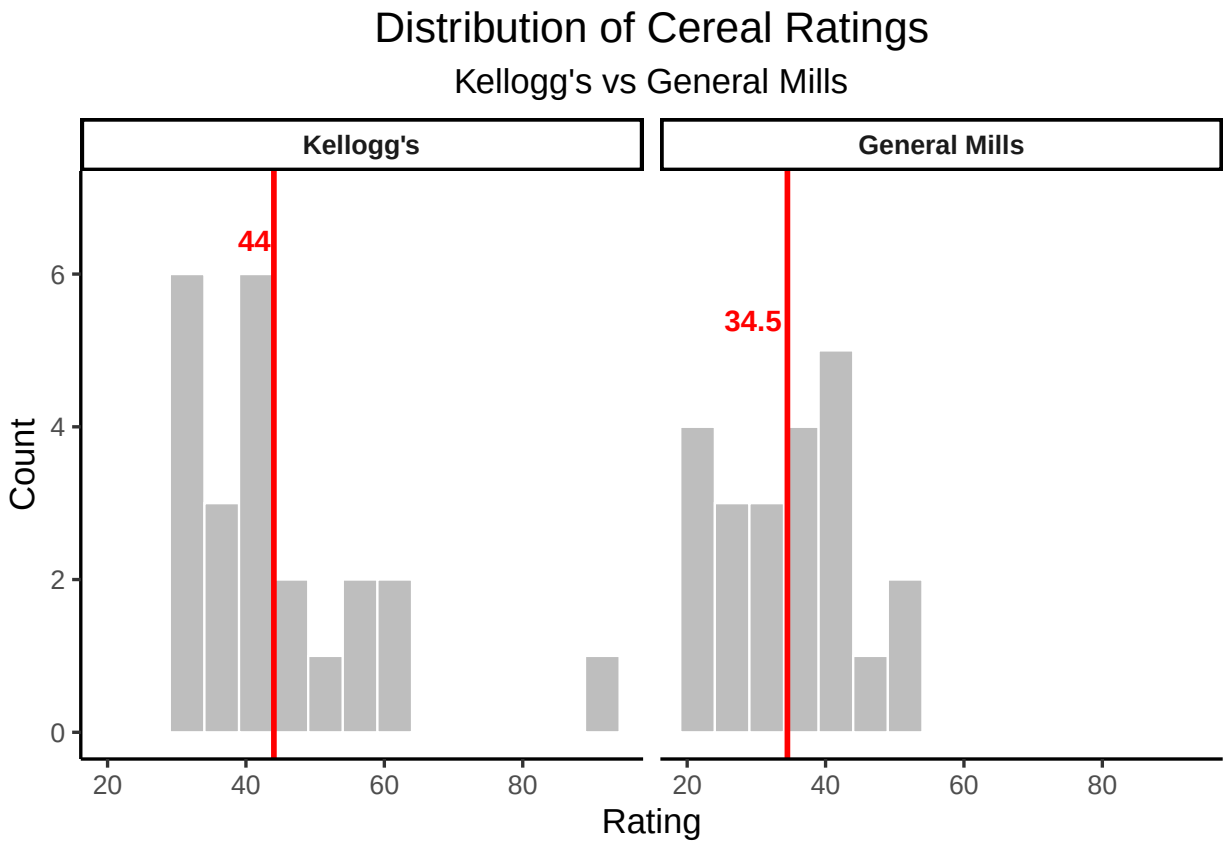
```r
top_mfr <- Cereals %>%
  group_by(manufacturer) %>%
  summarize(mean_rating = mean(rating, na.rm = TRUE), .groups = "drop") %>%
  slice_max(mean_rating, n = 1, with_ties = FALSE) %>%
  pull(manufacturer)

# Create a variable to distinguish.
cereals_flagged <- Cereals %>%
  mutate(
    highlight = if_else(manufacturer == top_mfr, "Highlighted", "Other manufacturers")
  )

# Formatting.
plot_df <- cereals_flagged %>%
  select(rating, sugars, calories, highlight) %>%
  pivot_longer(
    cols = c(sugars, calories),
    names_to = "metric",
    values_to = "value"
  ) %>%
  mutate(
    metric = factor(metric,
                    levels = c("sugars", "calories"),
                    labels = c("Sugar per serving (g)", "Calories per serving"))
  )

# Actual Plot (Only put this chunk in R Markdown.)
ggplot(plot_df, aes(x = value, y = rating, color = highlight)) +
  geom_point(size = 2.6, alpha = 0.9) +
  facet_wrap(~ metric, nrow = 1, scales = "free_x") +
  scale_color_manual(
    values = c("Other manufacturers" = "gray", "Highlighted" = "red"),
    labels = c("Other manufacturers", paste0(top_mfr, " (highest avg rating)")),
    guide = guide_legend(title = NULL, override.aes = list(alpha = 1, size = 3))
  ) +
  labs(
    title = "Ratings vs Sugar and Calories",
    subtitle = paste0(top_mfr, " is the highest-rated manufacturer."),
    x = NULL,
    y = "Rating"
  ) +
  theme_classic(base_size = 13) +
  theme(
    legend.position = "top",
    panel.grid.minor = element_blank(),
    strip.text = element_text(face = "bold"),
    plot.title = element_text(hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5)
  )
```



Task C

```
top2 <- Cereals %>%
  count(manufacturer, sort = TRUE) %>%
  slice_head(n = 2)

top2_mfr <- top2$manufacturer

df2 <- Cereals %>%
  filter(manufacturer %in% top2_mfr) %>%
  mutate(manufacturer = factor(manufacturer, levels = top2_mfr))

binwidth <- 5

means_df <- df2 %>%
  group_by(manufacturer) %>%
  summarize(mean_rating = mean(rating, na.rm = TRUE), .groups = "drop") %>%
  mutate(label = as.character(round(mean_rating, 1)))

breaks <- seq(floor(min(df2$rating, na.rm = TRUE)),
              ceiling(max(df2$rating, na.rm = TRUE)),
              by = binwidth)

max_counts <- df2 %>%
  group_by(manufacturer) %>%
  summarize(
```

```
    max_count = {
      h <- hist(rating, breaks = breaks, plot = FALSE)
      if (length(h$counts)) max(h$counts) else 0
    },
    .groups = "drop"
  )


means_df2 <- means_df %>%
  left_join(max_counts, by = "manufacturer")
y_upper <- ceiling(max(means_df2$max_count) * 1.15)

# Actual Plot

ggplot(df2, aes(x = rating)) +
  geom_histogram(
    binwidth = binwidth,
    breaks = breaks,
    fill = "gray",
    color = "white",
    linewidth = 0.4
  ) +
  geom_vline(
    data = means_df2,
    aes(xintercept = mean_rating),
    color = "red",
    linewidth = 1
  ) +
  geom_text(
    data = means_df2,
    aes(x = mean_rating, y = max_count * 1.05, label = label),
    vjust = 0,
    hjust = 1.1,
    color = "red",
    fontface = "bold",
    inherit.aes = FALSE
  ) +
  facet_wrap(~ manufacturer, nrow = 1, scales = "fixed") +
  coord_cartesian(xlim = range(df2$rating, na.rm = TRUE), ylim = c(0, y_upper)) +
  labs(
    title = "Distribution of Cereal Ratings",
    x = "Rating",
    y = "Count",
    subtitle = paste0(top2_mfr[1], " vs ", top2_mfr[2])
  ) +
  theme_classic(base_size = 13) +
  theme(
    legend.position = "none",
    panel.grid.minor = element_blank(),
    strip.text = element_text(face = "bold"),
    plot.title = element_text(hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5)
  )
```


Distribution of Cereal Ratings
Kellogg's vs General Mills
Kellogg's
General Mills
44
34.5
Count
0
2
4
6
20
40
60
80
Rating