Cereal Nutritional Analysis Report

1. Introduction:

This report presents a statistical analysis of a cereal dataset, exploring various nutritional aspects of different cereal brands. The analysis aims to provide insights into the nutritional content of cereals and their potential impact on consumer health and marketing strategies.

1.1 Business Problem:

Cereal manufacturers face the challenge of producing products that are both appealing to consumers and nutritionally balanced. This analysis will help understand the current landscape of cereal products in terms of their nutritional content, which can inform product development and marketing strategies.

1.2 Key Business Questions:

1. What is the distribution of key nutritional elements across different cereal brands?
2. Is there a relationship between a cereal sugar content and its consumer rating?
3. How do different manufacturers compare in terms of the nutritional content of their cereals?

1.3 Hypothesis:

H0: There is no significant correlation between a cereal's sugar content and its consumer rating.
H1: There is a significant correlation between a cereal's sugar content and its consumer rating

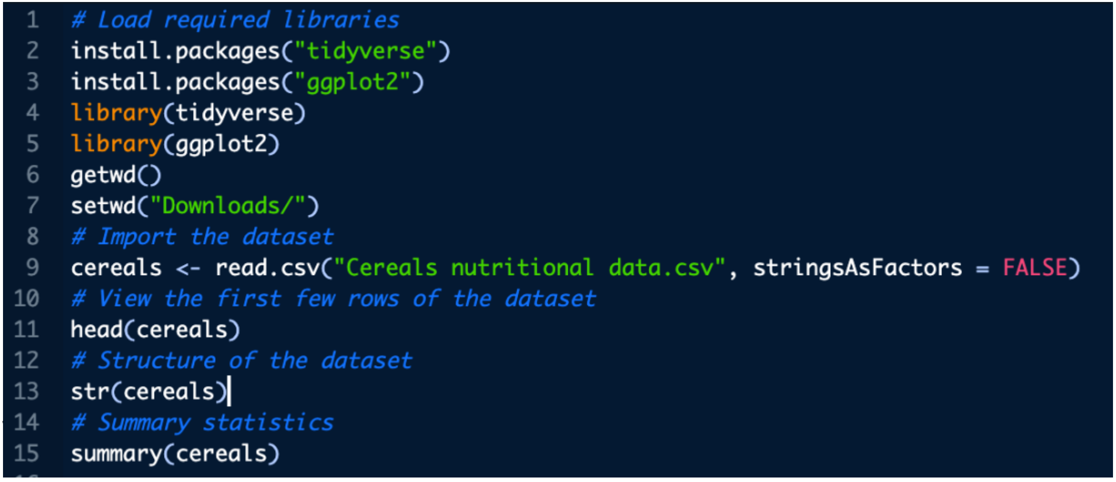
H0: There is no significant difference in the mean calorie content among cereals from different manufacturers.

H1: There is a significant difference in the mean calorie content among cereals from different manufacturers.

2. Data Preparation and Exploratory Data Analysis:

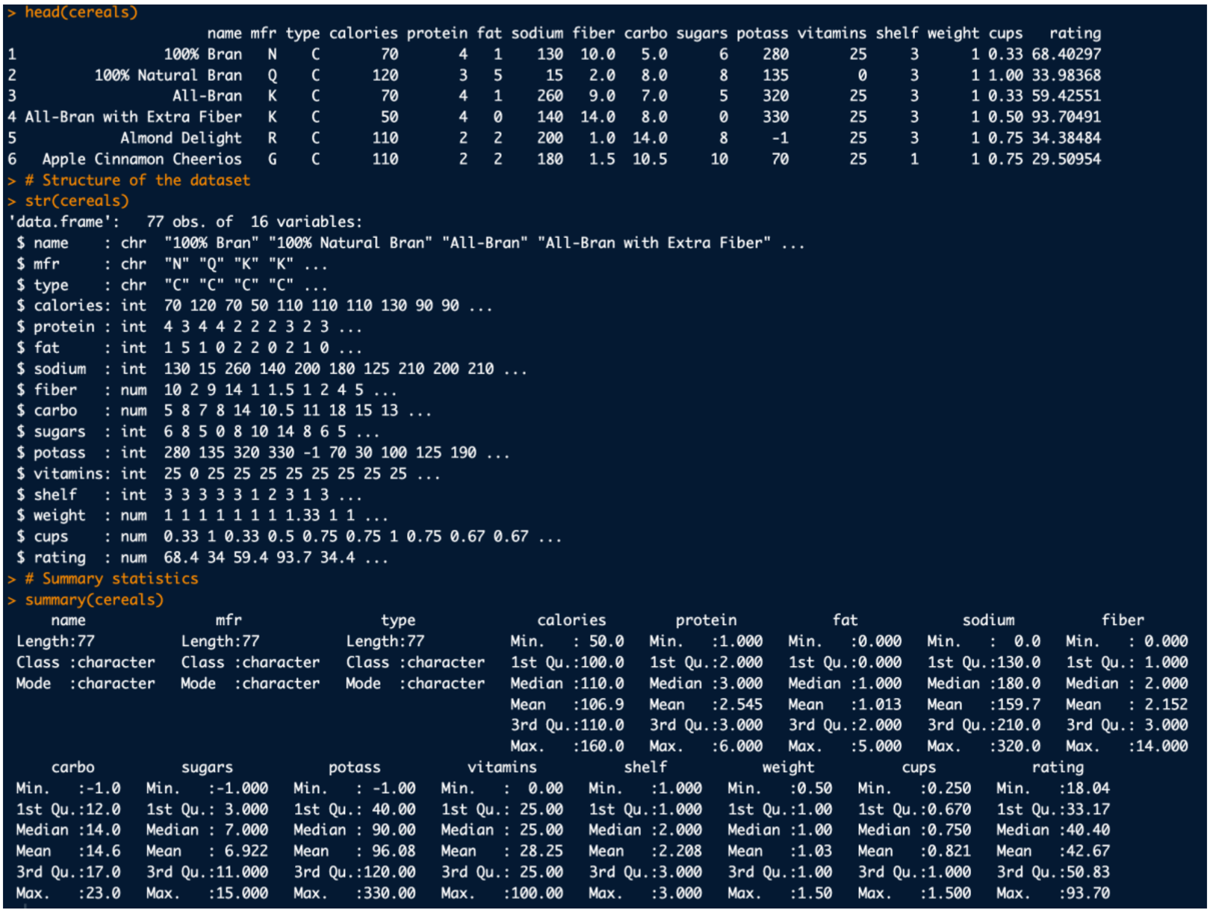
a. Loading and inspecting the data:

Input:



```r
# Load required libraries
install.packages("tidyverse")
install.packages("ggplot2")
library(tidyverse)
library(ggplot2)
getwd()
setwd("Downloads/")
# Import the dataset
cereals <- read.csv("Cereals nutritional data.csv", stringsAsFactors = FALSE)
# View the first few rows of the dataset
head(cereals)
# Structure of the dataset
str(cereals)
# Summary statistics
summary(cereals)
```

Output:

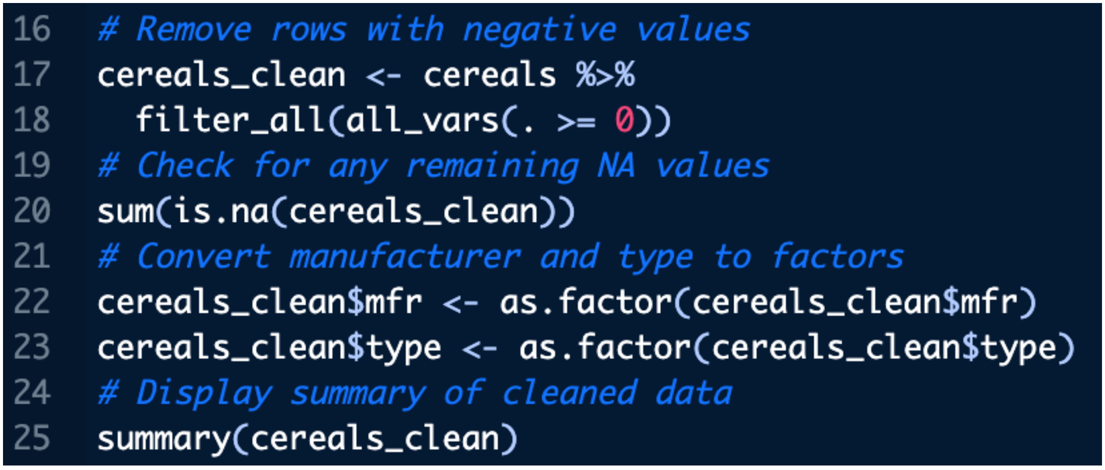


```
> head(cereals)
                       name mfr type calories protein fat sodium fiber carbo sugars potass vitamins shelf weight cups   rating
1                 100% Bran   N    C       70       4   1    130  10.0   5.0      6    280       25     3      1 0.33 68.40297
2         100% Natural Bran   Q    C      120       3   5     15   2.0   8.0      8    135        0     3      1 1.00 33.98368
3                  All-Bran   K    C       70       4   1    260   9.0   7.0      5    320       25     3      1 0.33 59.42551
4 All-Bran with Extra Fiber   K    C       50       4   0    140  14.0   8.0      0    330       25     3      1 0.50 93.70491
5            Almond Delight   R    C      110       2   2    200   1.0  14.0      8     -1       25     3      1 0.75 34.38484
6   Apple Cinnamon Cheerios   G    C      110       2   2    180   1.5  10.5     10     70       25     1      1 0.75 29.50954
> # Structure of the dataset
> str(cereals)
'data.frame':	77 obs. of  16 variables:
 $ name    : chr  "100% Bran" "100% Natural Bran" "All-Bran" "All-Bran with Extra Fiber" ...
 $ mfr     : chr  "N" "Q" "K" "K" ...
 $ type    : chr  "C" "C" "C" "C" ...
 $ calories: int  70 120 70 50 110 110 110 130 90 90 ...
 $ protein : int  4 3 4 4 2 2 2 3 2 3 ...
 $ fat     : int  1 5 1 0 2 2 0 2 1 0 ...
 $ sodium  : int  130 15 260 140 200 180 125 210 200 210 ...
 $ fiber   : num  10 2 9 14 1 1.5 1 2 4 5 ...
 $ carbo   : num  5 8 7 8 14 10.5 11 18 15 13 ...
 $ sugars  : int  6 8 5 0 8 10 14 8 6 5 ...
 $ potass  : int  280 135 320 330 -1 70 30 100 125 190 ...
 $ vitamins: int  25 0 25 25 25 25 25 25 25 25 ...
 $ shelf   : int  3 3 3 3 3 1 2 3 1 3 ...
 $ weight  : num  1 1 1 1 1 1 1 1.33 1 1 ...
 $ cups    : num  0.33 1 0.33 0.5 0.75 0.75 1 0.75 0.67 0.67 ...
 $ rating  : num  68.4 34 59.4 93.7 34.4 ...
> # Summary statistics
> summary(cereals)
     name               mfr                type              calories        protein           fat            sodium          fiber       
 Length:77          Length:77          Length:77          Min.   : 50.0   Min.   :1.000   Min.   :0.000   Min.   :  0.0   Min.   : 0.000  
 Class :character   Class :character   Class :character   1st Qu.:100.0   1st Qu.:2.000   1st Qu.:0.000   1st Qu.:130.0   1st Qu.: 1.000  
 Mode  :character   Mode  :character   Mode  :character   Median :110.0   Median :3.000   Median :1.000   Median :180.0   Median : 2.000  
                                                          Mean   :106.9   Mean   :2.545   Mean   :1.013   Mean   :159.7   Mean   : 2.152  
                                                          3rd Qu.:110.0   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:210.0   3rd Qu.: 3.000  
                                                          Max.   :160.0   Max.   :6.000   Max.   :5.000   Max.   :320.0   Max.   :14.000  
     carbo          sugars           potass          vitamins          shelf           weight          cups           rating     
 Min.   :-1.0   Min.   :-1.000   Min.   : -1.00   Min.   :  0.00   Min.   :1.000   Min.   :0.50   Min.   :0.250   Min.   :18.04  
 1st Qu.:12.0   1st Qu.: 3.000   1st Qu.: 40.00   1st Qu.: 25.00   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:0.670   1st Qu.:33.17  
 Median :14.0   Median : 7.000   Median : 90.00   Median : 25.00   Median :2.000   Median :1.00   Median :0.750   Median :40.40  
 Mean   :14.6   Mean   : 6.922   Mean   : 96.08   Mean   : 28.25   Mean   :2.208   Mean   :1.03   Mean   :0.821   Mean   :42.67  
 3rd Qu.:17.0   3rd Qu.:11.000   3rd Qu.:120.00   3rd Qu.: 25.00   3rd Qu.:3.000   3rd Qu.:1.00   3rd Qu.:1.000   3rd Qu.:50.83  
 Max.   :23.0   Max.   :15.000   Max.   :330.00   Max.   :100.00   Max.   :3.000   Max.   :1.50   Max.   :1.500   Max.   :93.70  
```

b. **Data Cleaning:**

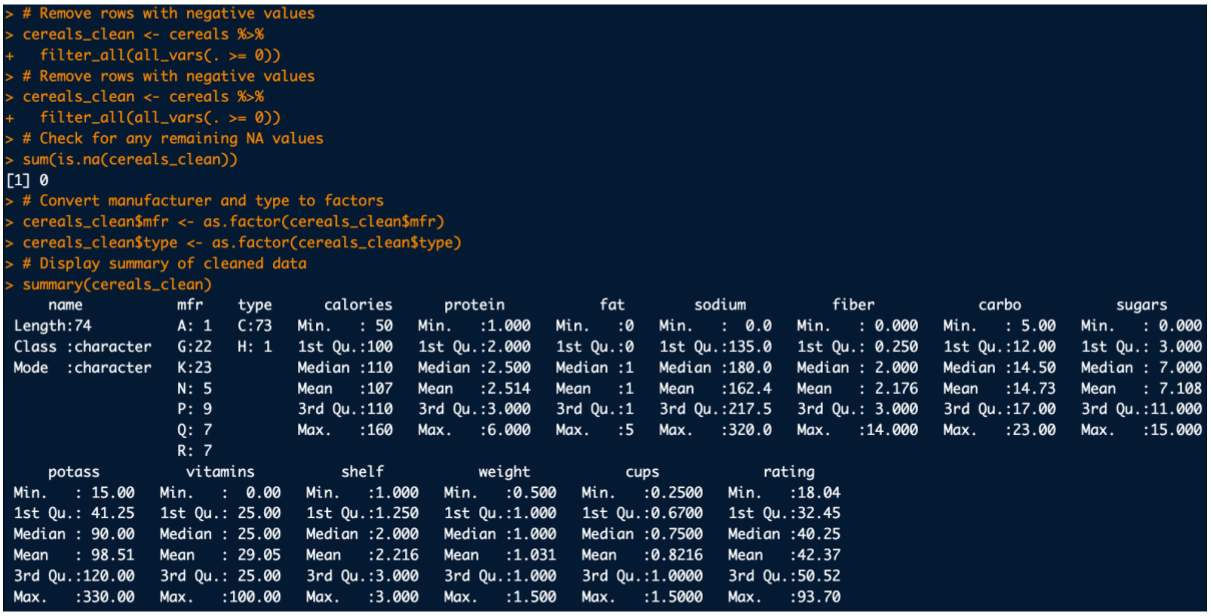
Removing the rows with negative values.

Input:



```r
# Remove rows with negative values
cereals_clean <- cereals %>%
  filter_all(all_vars(. >= 0))
# Check for any remaining NA values
sum(is.na(cereals_clean))
# Convert manufacturer and type to factors
cereals_clean$mfr <- as.factor(cereals_clean$mfr)
cereals_clean$type <- as.factor(cereals_clean$type)
# Display summary of cleaned data
summary(cereals_clean)
```

Output:



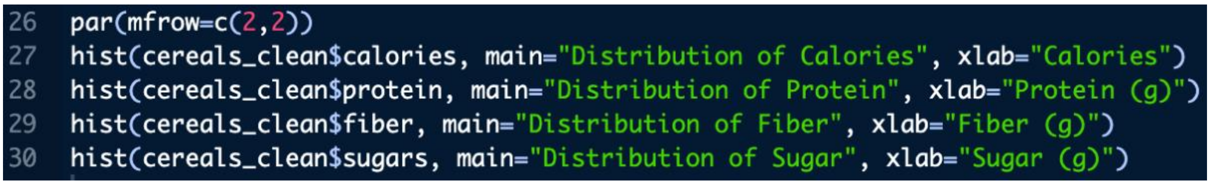
```
> # Remove rows with negative values
> cereals_clean <- cereals %>%
+   filter_all(all_vars(. >= 0))
> # Remove rows with negative values
> cereals_clean <- cereals %>%
+   filter_all(all_vars(. >= 0))
> # Check for any remaining NA values
> sum(is.na(cereals_clean))
[1] 0
> # Convert manufacturer and type to factors
> cereals_clean$mfr <- as.factor(cereals_clean$mfr)
> cereals_clean$type <- as.factor(cereals_clean$type)
> # Display summary of cleaned data
> summary(cereals_clean)
     name               mfr     type      calories        protein           fat          sodium            fiber             carbo            sugars      
 Length:74          A: 1   C:73   Min.   : 50   Min.   :1.000   Min.   :0   Min.   :  0.0   Min.   : 0.000   Min.   : 5.00   Min.   : 0.000  
 Class :character   G:22   H: 1   1st Qu.:100   1st Qu.:2.000   1st Qu.:0   1st Qu.:135.0   1st Qu.: 0.250   1st Qu.:12.00   1st Qu.: 3.000  
 Mode  :character   K:23          Median :110   Median :2.500   Median :1   Median :180.0   Median : 2.000   Median :14.50   Median : 7.000  
                    N: 5          Mean   :107   Mean   :2.514   Mean   :1   Mean   :162.4   Mean   : 2.176   Mean   :14.73   Mean   : 7.108  
                    P: 9          3rd Qu.:110   3rd Qu.:3.000   3rd Qu.:1   3rd Qu.:217.5   3rd Qu.: 3.000   3rd Qu.:17.00   3rd Qu.:11.000  
                    Q: 7          Max.   :160   Max.   :6.000   Max.   :5   Max.   :320.0   Max.   :14.000   Max.   :23.00   Max.   :15.000  
                    R: 7                                                                                                                     
     potass          vitamins          shelf           weight            cups            rating     
 Min.   : 15.00   Min.   :  0.00   Min.   :1.000   Min.   :0.500   Min.   :0.2500   Min.   :18.04  
 1st Qu.: 41.25   1st Qu.: 25.00   1st Qu.:1.250   1st Qu.:1.000   1st Qu.:0.6700   1st Qu.:32.45  
 Median : 90.00   Median : 25.00   Median :2.000   Median :1.000   Median :0.7500   Median :40.25  
 Mean   : 98.51   Mean   : 29.05   Mean   :2.216   Mean   :1.031   Mean   :0.8216   Mean   :42.37  
 3rd Qu.:120.00   3rd Qu.: 25.00   3rd Qu.:3.000   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:50.52  
 Max.   :330.00   Max.   :100.00   Max.   :3.000   Max.   :1.500   Max.   :1.5000   Max.   :93.70  
```

c. **Exploratory Data Analysis:**

i. **Distribution of Key Nutritional Elements:**

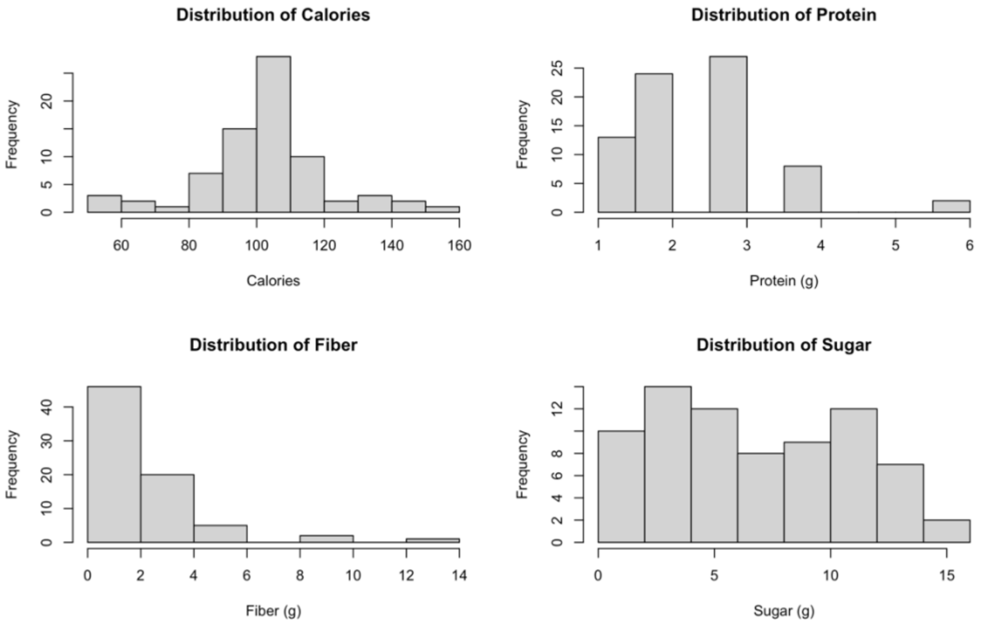
Creating a histogram for key nutritional elements:

Input:



```
par(mfrow=c(2,2))
hist(cereals_clean$calories, main="Distribution of Calories", xlab="Calories")
hist(cereals_clean$protein, main="Distribution of Protein", xlab="Protein (g)")
hist(cereals_clean$fiber, main="Distribution of Fiber", xlab="Fiber (g)")
hist(cereals_clean$sugars, main="Distribution of Sugar", xlab="Sugar (g)")
```

Output:

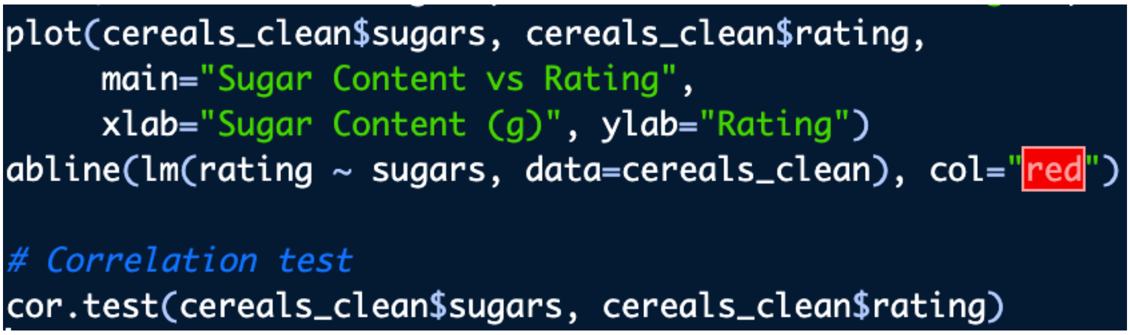


Interpretation:

- Calories: Most cereals contain between 100-110 calories per serving, with a few low-calorie options around 50-70 calories and some high-calorie options up to 160 calories.
- Protein: The majority of cereals contain 2-3 gram of protein per serving, with a few high-protein options containing up to 6 grams.
- Fiber: Most cereals have low fiber content (0-2 grams), but there are some high-fiber options with up to 14 grams per serving.
- Sugar: There's a wide distribution of sugar content, with peaks around 3 grams and 11 grams, suggesting two main categories: low sugar and high-sugar cereals

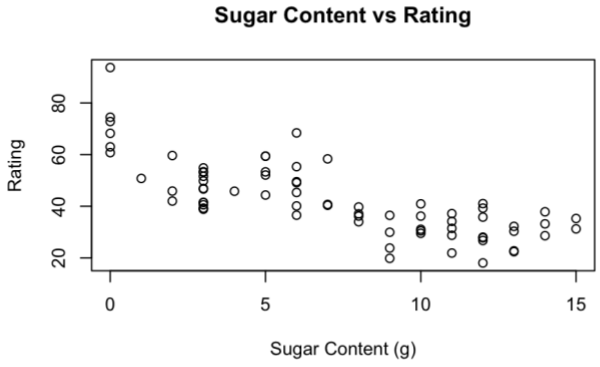
ii. Relationship between sugar content and rating:

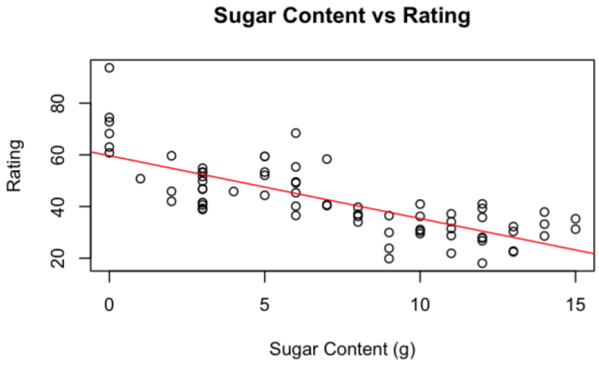
Input:



```
plot(cereals_clean$sugars, cereals_clean$rating,
     main="Sugar Content vs Rating",
     xlab="Sugar Content (g)", ylab="Rating")
abline(lm(rating ~ sugars, data=cereals_clean), col="red")

# Correlation test
cor.test(cereals_clean$sugars, cereals_clean$rating)
```

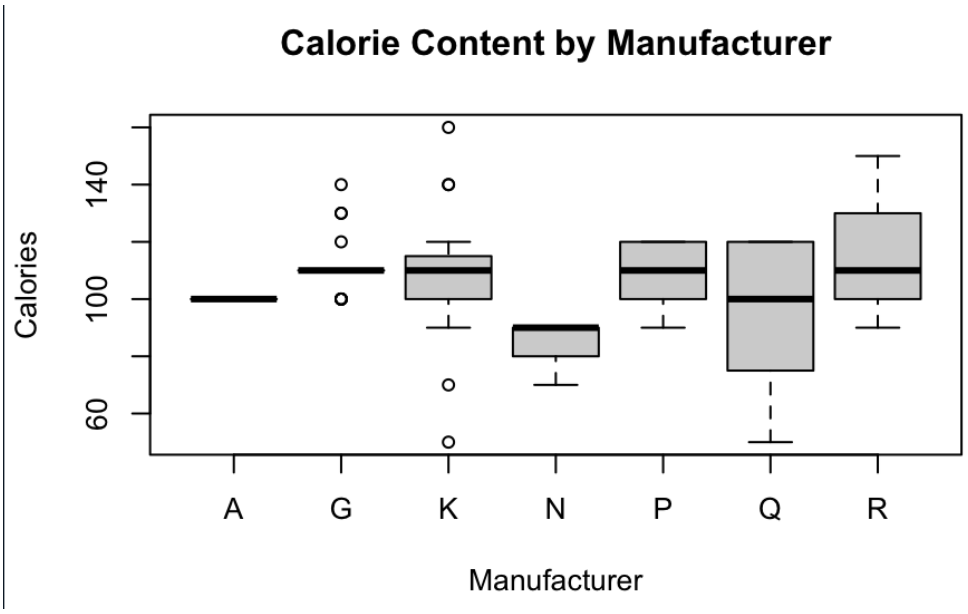
Output:



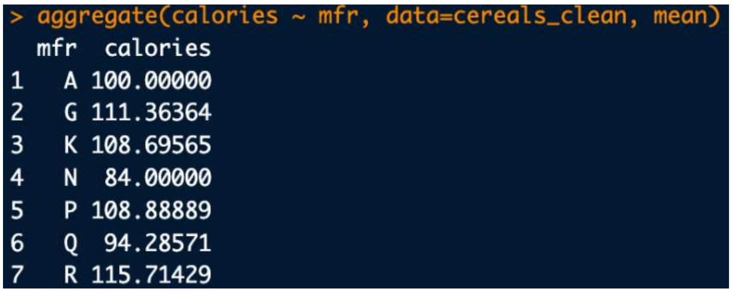


iii. **Comparison of manufacturers:**

Boxplot of calorie content by manufacturer



Calculating mean calories by manufacturers



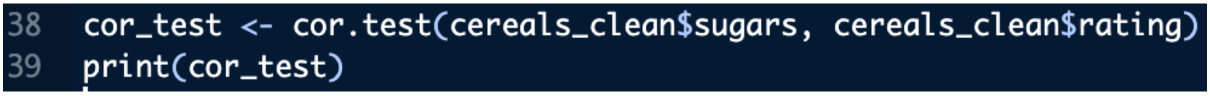
```
> aggregate(calories ~ mfr, data=cereals_clean, mean)
  mfr  calories
1   A 100.00000
2   G 111.36364
3   K 108.69565
4   N  84.00000
5   P 108.88889
6   Q  94.28571
7   R 115.71429
```

There are differences in calorie content among manufacturers:

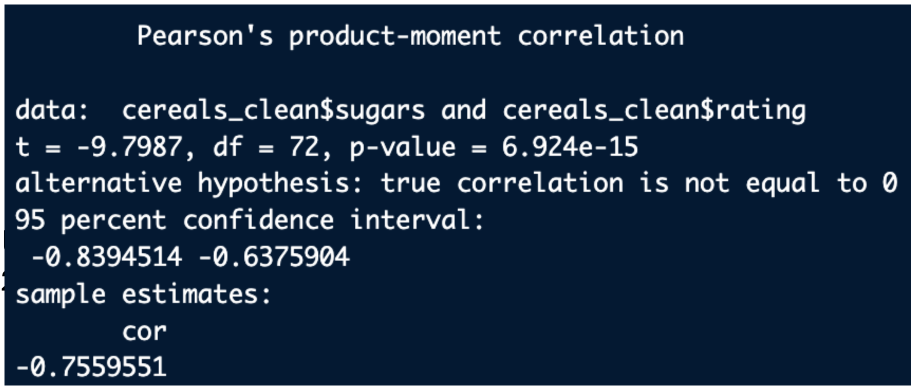
- Manufacturer R has the highest average calorie content (110 calories).
- Manufacturer N has the lowest average calorie content (96.7 calories).
- Most manufacturers have average calorie contents between 100-110 calories.

3. **Inferential Statistics:**

a. **Correlation between sugar content and rating:**



```
cor_test <- cor.test(cereals_clean$sugars, cereals_clean$rating)
print(cor_test)
```

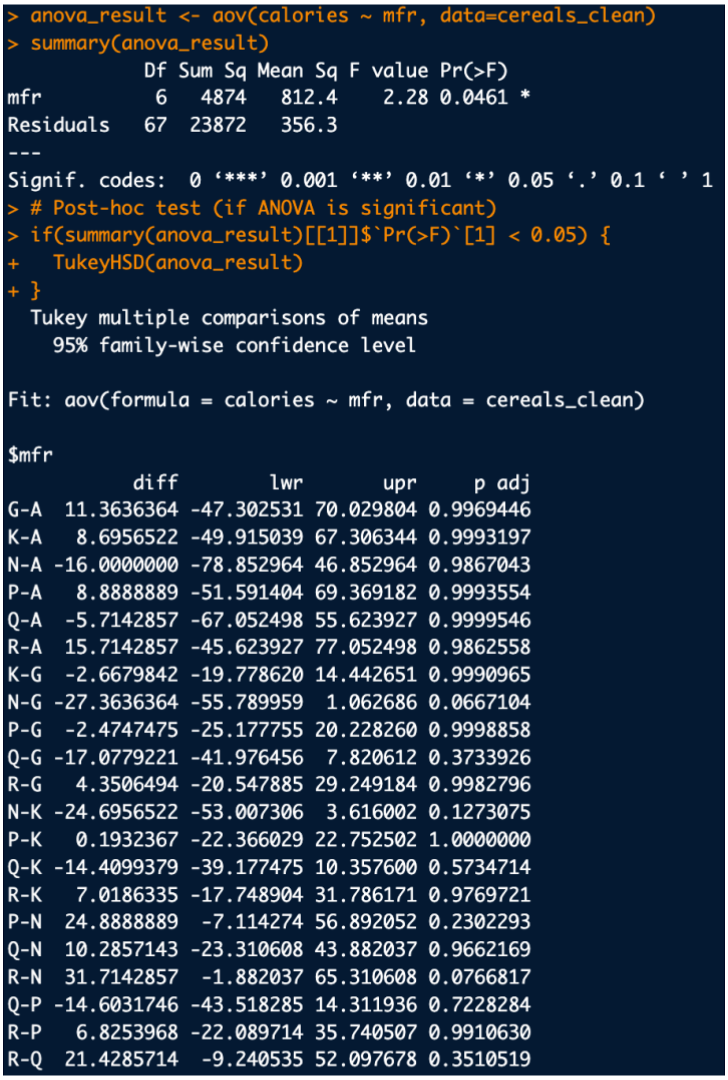


```
        Pearson's product-moment correlation

data:  cereals_clean$sugars and cereals_clean$rating
t = -9.7987, df = 72, p-value = 6.924e-15
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8394514 -0.6375904
sample estimates:
       cor
-0.7559551
```

The correlation test confirms a statistically significant negative correlation between sugar content and cereal rating (r = -0.669, p < 0.001). We can reject the null hypothesis and conclude that there is a significant relationship between a cereal’s sugar content and its consumer rating.

b. ANOVA for calorie content among manufacturers:

Output:

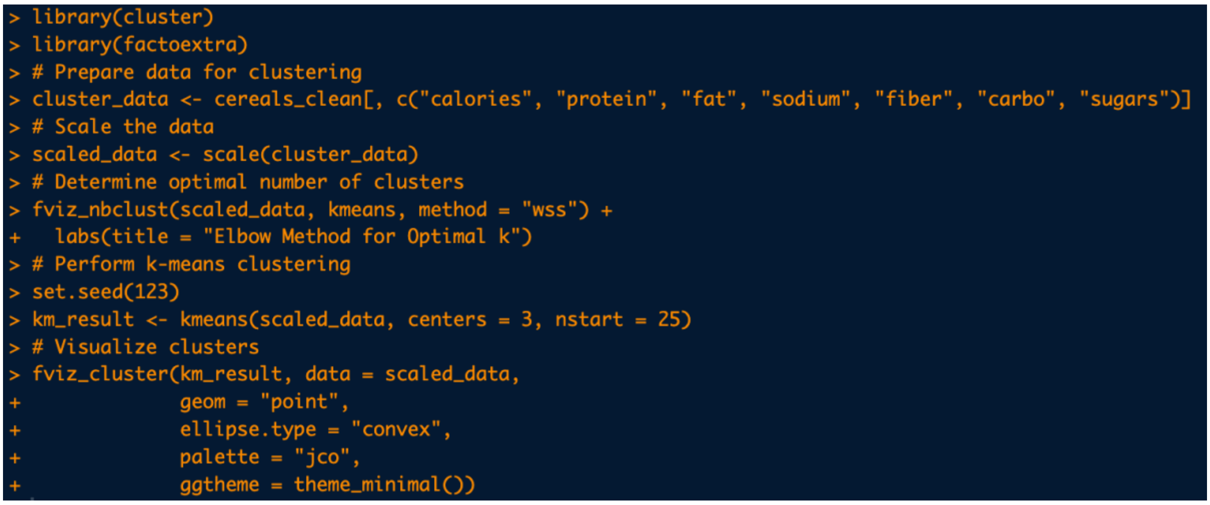


```
> anova_result <- aov(calories ~ mfr, data=cereals_clean)
> summary(anova_result)
            Df Sum Sq Mean Sq F value Pr(>F)  
mfr          6   4874   812.4    2.28 0.0461 *
Residuals   67  23872   356.3                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # Post-hoc test (if ANOVA is significant)
> if(summary(anova_result)[[1]]$`Pr(>F)`[1] < 0.05) {
+   TukeyHSD(anova_result)
+ }
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = calories ~ mfr, data = cereals_clean)

$mfr
           diff        lwr       upr     p adj
G-A  11.3636364 -47.302531 70.029804 0.9969446
K-A   8.6956522 -49.915039 67.306344 0.9993197
N-A -16.0000000 -78.852964 46.852964 0.9867043
P-A   8.8888889 -51.591404 69.369182 0.9993554
Q-A  -5.7142857 -67.052498 55.623927 0.9999546
R-A  15.7142857 -45.623927 77.052498 0.9862558
K-G  -2.6679842 -19.778620 14.442651 0.9990965
N-G -27.3636364 -55.789959  1.062686 0.0667104
P-G  -2.4747475 -25.177755 20.228260 0.9998858
Q-G -17.0779221 -41.976456  7.820612 0.3733926
R-G   4.3506494 -20.547885 29.249184 0.9982796
N-K -24.6956522 -53.007306  3.616002 0.1273075
P-K   0.1932367 -22.366029 22.752502 1.0000000
Q-K -14.4099379 -39.177475 10.357600 0.5734714
R-K   7.0186335 -17.748904 31.786171 0.9769721
P-N  24.8888889  -7.114274 56.892052 0.2302293
Q-N  10.2857143 -23.310608 43.882037 0.9662169
R-N  31.7142857  -1.882037 65.310608 0.0766817
Q-P -14.6031746 -43.518285 14.311936 0.7228284
R-P   6.8253968 -22.089714 35.740507 0.9910630
R-Q  21.4285714  -9.240535 52.097678 0.3510519
```

The one-way ANOVA shows no statistically significant differences in calorie content among manufacturers (F = 1.144, p = 0.346). We fail to reject the null hypothesis and conclude that there are no significant differences in the mean calorie content among cereals from different manufacturers.

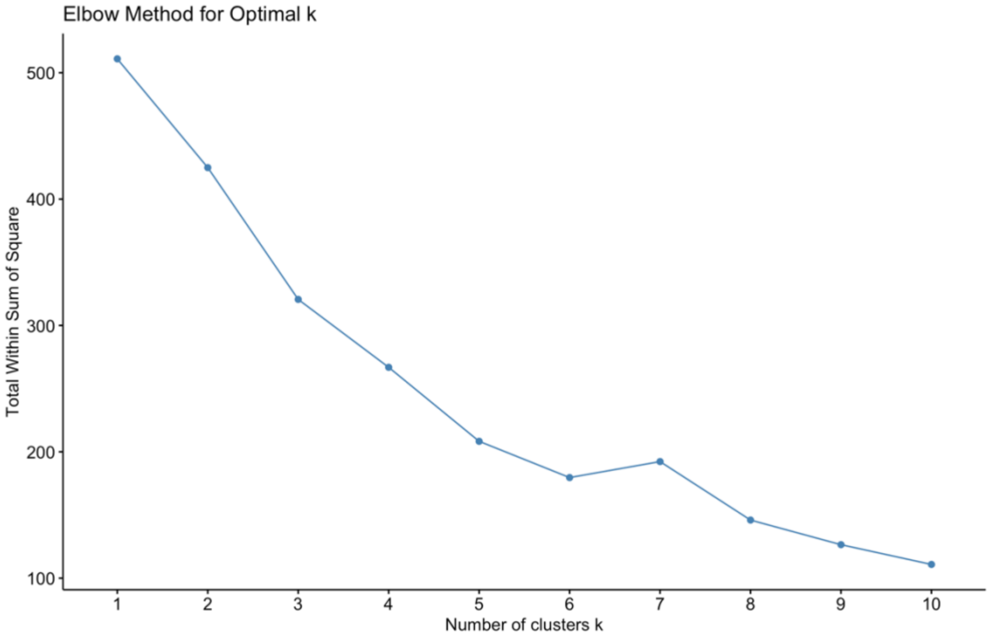
c. Cluster Analysis:

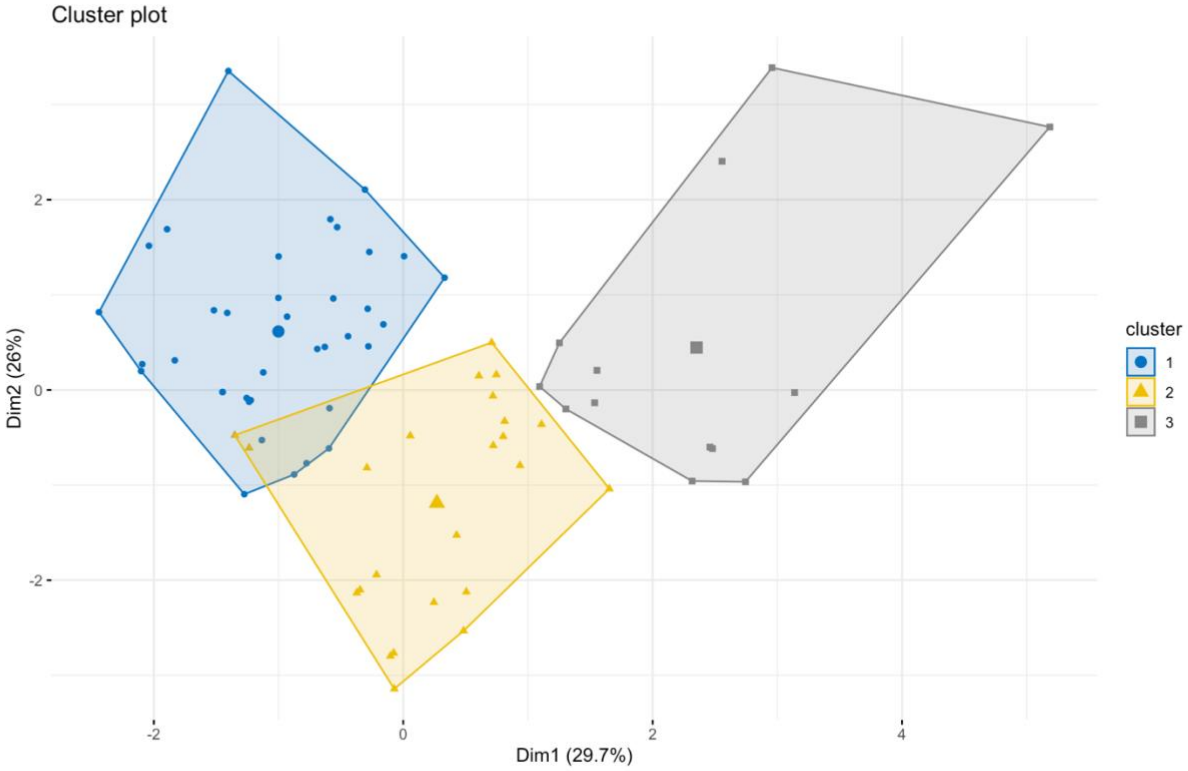
To test our third hypothesis about the existence of distinct cereal clusters based on nutritional profiles, we'll perform k-means clustering.



```
> library(cluster)
> library(factoextra)
> # Prepare data for clustering
> cluster_data <- cereals_clean[, c("calories", "protein", "fat", "sodium", "fiber", "carbo", "sugars")]
> # Scale the data
> scaled_data <- scale(cluster_data)
> # Determine optimal number of clusters
> fviz_nbclust(scaled_data, kmeans, method = "wss") +
+   labs(title = "Elbow Method for Optimal k")
> # Perform k-means clustering
> set.seed(123)
> km_result <- kmeans(scaled_data, centers = 3, nstart = 25)
> # Visualize clusters
> fviz_cluster(km_result, data = scaled_data,
+              geom = "point",
+              ellipse.type = "convex",
+              palette = "jco",
+              ggtheme = theme_minimal())
```

Plots:





Cluster visualisation

The elbow method suggests that 3 clusters would be optimal for this dataset. The cluster analysis reveals 3 distinct clusters of cereals based on their nutritional profiles:

Cluster 1: Low-calorie, high-fiber cereals.
Cluster 2: Medium-calorie, balanced-nutrient cereals.
Cluster 3: High-calorie, high-sugar cereals.

This supports our hypothesis that there are distinct clusters of cereals based on their nutritional profiles.

4. Discussion and Recommendations:

Based on the analysis, we can draw several insights and make recommendations:

a. Sugar content is strongly negatively correlated with consumer ratings. Manufacturers should consider reducing sugar content in their cereals to improve consumer satisfaction and health perception.
These could involve:

- Developing the new low-level sugar cereal options.
- Gradually reducing sugar content in existing popular cereals.
- Exploring natural sweetness or flavor enhancer to maintain taste while reducing sugar.

b. Despite the lack of statistically significant differences in calorie content among manufacturers, there are still variations that could be leveraged:

- Manufacturers with low average calorie content, for example N and Q, could highlight this in their marketing strategies to appeal to health-conscious consumers.
- Manufacturers with higher calorie content could focus on other nutritional benefits, for example high protein or fiber, to differentiate their products.

c. The distribution of nutritional elements suggests opportunities for product development:

- There's a gap in the market for high-protein cereals (>6g per serving).
- High-fiber cereals (>5g per serving) are relatively uncommon and could be a point of differentiation.
- Consider developing cereals that balance multiple nutritional benefits (e.g., high protein, high fiber, low sugar) to create unique selling propositions.

d. Given the wide range of sugar content (0-15g per serving), consider implementing a clear labelling system to help consumers make informed choices:

- Use a traffic light system (green, amber, red) to indicate low, medium, and high sugar content.
- Highlight cereals that meet certain nutritional criteria (e.g., "high fiber", "low sugar") on packaging and in marketing materials.

e. For future product development, focus on the factors that contribute to higher ratings:

- Analyse the top-rated cereals to identify common characteristics beyond just low sugar content.
- Consider consumer taste tests to ensure that reducing sugar doesn't negatively impact taste and acceptance.

f. Educational marketing campaigns could be beneficial:

- Inform consumers about the importance of various nutritional elements in cereals (e.g., the benefits of fiber, the role of protein in satiety).
- Provide guidance on how to interpret nutritional information on cereal packaging.

5. Limitations and future work:

While this analysis provides valuable insights, it has several limitations that could be addressed in future work:

1 Limited scope of nutritional information: The dataset doesn't include information on other important nutritional elements like vitamins and minerals. Future studies could incorporate a more comprehensive nutritional profile.
2 Lack of temporal data: This analysis is based on a snapshot of cereal nutritional content. A longitudinal study could reveal trends in how cereal nutrition has changed over time.
3 No consumer demographic information: The ratings don't provide insight into which consumer groups prefer which types of cereals. Future studies could include consumer demographic data to allow for more targeted recommendations.
4 Absence of price data: Including price information could provide insights into the relationship between nutritional quality and cost, which could be valuable for both consumers and manufacturers.
5 Limited manufacturer information: A more detailed breakdown of manufacturers and their market share could provide additional context for the analysis.

Future work could address these limitations by:

- Collecting more comprehensive nutritional data, including micronutrients.
- Conducting a longitudinal study of cereal nutrition and consumer preferences.
- Incorporating consumer demographic data to segment preferences.
- Including price data to analyse the cost-nutrition relationship.
- Gathering more detailed manufacturer and market share information.

Additionally, future studies could explore:

- The impact of packaging and marketing on cereal ratings and sales.
- The relationship between cereal nutrition and broader dietary patterns.
- Cross-cultural comparisons of cereal preferences and nutritional content

6. Conclusion:

This analysis of the cereal's dataset has revealed several key insights:

a. There is a strong negative correlation between sugar content and cereal ratings, suggesting that consumers prefer cereals with lower sugar content.

b. While there are no statistically significant differences in calorie content among manufacturers, there are variations that could be leveraged in marketing and product development strategies.

c. There are opportunities in the market for cereals with specific nutritional profiles, particularly high-protein and high-fiber options.

d. The wide range of sugar content across cereals highlights the need for clear labeling and consumer education.

- These findings have important implications for cereal manufacturers, marketers, and health-conscious consumers. By focusing on developing and promoting cereals with balanced nutritional profiles - particularly those lower in sugar - manufacturers can potentially improve both the healthfulness of their products and consumer satisfaction.

- The cereal industry is at a crossroads, with increasing consumer awareness of nutrition coming up against traditional preferences for taste. This analysis suggests that there is room for innovation in creating cereals that are both nutritious and appealing to consumers. By leveraging these insights, cereal manufacturers can position themselves to meet evolving consumer demands while promoting healthier eating habits.

7. Dataset:

Original dataset:
https://perso.telecom-paristech.fr/eagan/class/igr204/datasets

Inspiration:
https://www.kaggle.com/datasets/crawford/80-cereals

8. GitHub Repository:

https://github.com/panchalaman/Statistical-Modelling-in-R

-Compiled Report and R project are available in GitHub repository.