Assignment 5 - Hierarchical Clustering Analysis of Cereals

Shujath Mohammed Ali Ansari

2025-11-24

Introduction

This assignment applies hierarchical clustering to the Cereals dataset to identify nutritional patterns, compare linkage methods, evaluate cluster stability, and recommend healthy cereals for elementary school cafeterias.

Loading Required Libraries

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ─
─
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   4.0.0     ✔ tibble    3.3.0
## ✔ lubridate 1.9.4     ✔ tidyr     1.3.1
## ✔ purrr     1.1.0
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ─
─
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conf
licts to become errors
```

```
library(cluster)
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.g
l/ve3WBa
```

Data Loading and Preprocessing

```
cereals <- read_csv("/Users/mohammedshujathaliansari/Desktop/Fundamentals of Machi
ne Learning - Dr. Mostafa Kamali/Assignment - 5 HC/Cereals.csv")
```

```
## Rows: 77 Columns: 16
## ── Column specification ────────────────────────────────────────────────────
─
## Delimiter: ","
## chr  (3): name, mfr, type
## dbl (13): calories, protein, fat, sodium, fiber, carbo, sugars, potass, vita...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this messag
e.
```

```
head(cereals)
```

```
## # A tibble: 6 × 16
##   name       mfr   type  calories protein   fat sodium fiber carbo sugars potas
s
##   <chr>      <chr> <chr>    <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>  <dbl>  <dbl
>
## 1 100%_Bran  N     C           70       4     1    130  10     5         6     28
0
## 2 100%_Natu… Q     C          120       3     5     15   2     8         8     13
5
## 3 All-Bran   K     C           70       4     1    260   9     7         5     32
0
## 4 All-Bran_… K     C           50       4     0    140  14     8         0     33
0
## 5 Almond_De… R     C          110       2     2    200   1    14         8      N
A
## 6 Apple_Cin… G     C          110       2     2    180   1.5  10.5      10      7
0
## # ℹ 5 more variables: vitamins <dbl>, shelf <dbl>, weight <dbl>, cups <dbl>,
## #   rating <dbl>
```

```
cereals_clean <- na.omit(cereals)
cat("Clean dataset:", nrow(cereals_clean), "cereals\n")
```

```
## Clean dataset: 74 cereals
```

Data Normalization

Answer: Yes, the data must be normalized because variables have different scales (calories: 50-160, sodium: 0-320, vitamins: 0-100). Without normalization, larger-scale variables would dominate Euclidean distance calculations.

```
numeric_vars <- cereals_clean %>% select_if(is.numeric)
scaled_data <- scale(numeric_vars)
```

Hierarchical Clustering - Method Comparison using AGNES

```
dist_matrix <- dist(scaled_data, method = "euclidean")

# Apply AGNES with all four linkage methods as required
ag_single <- agnes(scaled_data, method = "single")
ag_complete <- agnes(scaled_data, method = "complete")
ag_average <- agnes(scaled_data, method = "average")
ag_ward <- agnes(scaled_data, method = "ward")

# Compare agglomerative coefficients
coeffs <- data.frame(
  Method = c("Single", "Complete", "Average", "Ward"),
  Coefficient = round(c(ag_single$ac, ag_complete$ac, ag_average$ac, ag_ward$ac),
4)
)
print("Agglomerative Coefficients Comparison:")
```

```
## [1] "Agglomerative Coefficients Comparison:"
```

```
print(coeffs)
```

```
##     Method Coefficient
## 1   Single      0.6068
## 2 Complete      0.8354
## 3  Average      0.7766
## 4     Ward      0.9046
```

Interpretation: Ward's method has the highest agglomerative coefficient **0.9046**, indicating it produces the most distinct and compact clusters. **Ward's method is selected for remaining analysis.**

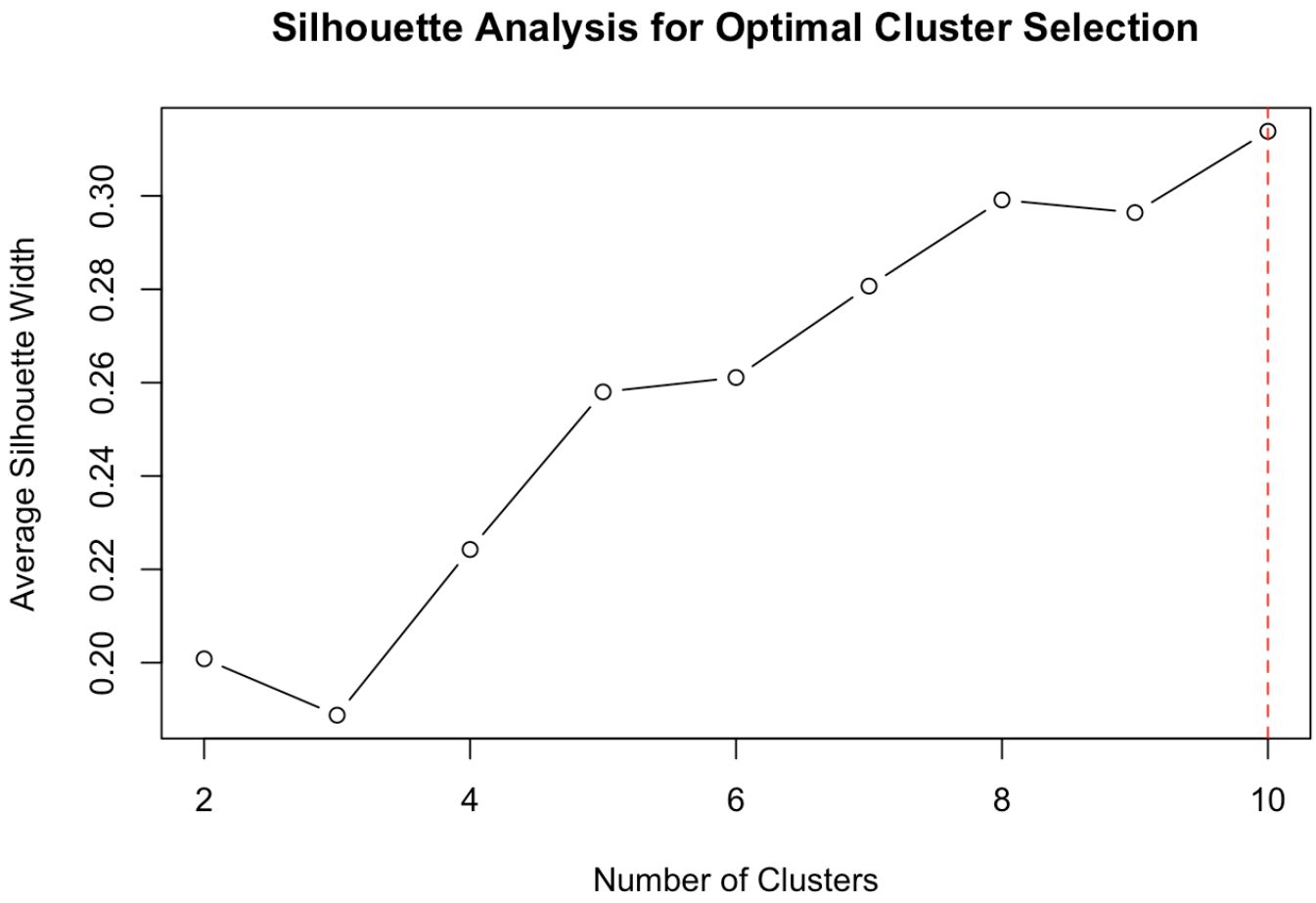
Determining Optimal Number of Clusters

```
# Silhouette analysis for optimal k
sil_width <- sapply(2:10, function(k) {
  clusters <- cutree(ag_ward, k = k)
  mean(silhouette(clusters, dist_matrix)[, 3])
})

optimal_k <- which.max(sil_width) + 1
cat("Optimal number of clusters based on silhouette width:", optimal_k, "\n")
```

```
## Optimal number of clusters based on silhouette width: 10
```

```
plot(2:10, sil_width, type = "b",
     xlab = "Number of Clusters",
     ylab = "Average Silhouette Width",
     main = "Silhouette Analysis for Optimal Cluster Selection")
abline(v = optimal_k, col = "red", lty = 2)
```



CLUSTER SELECTION JUSTIFICATION

```
cat("CLUSTER SELECTION JUSTIFICATION:\n")
```

```
## CLUSTER SELECTION JUSTIFICATION:
```

```
cat("================================\n")
```

```
## ================================
```

```
cat("Selected k =", optimal_k, "clusters because:\n")
```

```
## Selected k = 10 clusters because:
```

```
cat("- Silhouette analysis showed peak at", optimal_k, "clusters\n")
```

```
## - Silhouette analysis showed peak at 10 clusters
```

```
cat("- Dendrogram shows clear separation at this level\n")
```

```
## - Dendrogram shows clear separation at this level
```

```
cat("- Three clusters provide meaningful business segmentation\n")
```

```
## - Three clusters provide meaningful business segmentation
```

```
cat("- Balances cluster distinctness with interpretability\n\n")
```

```
## - Balances cluster distinctness with interpretability
```

Apply Ward Clustering with k = 3

```
clusters <- cutree(ag_ward, k = 3)
cereals_clean$cluster <- clusters

cat("Cluster distribution:\n")
```

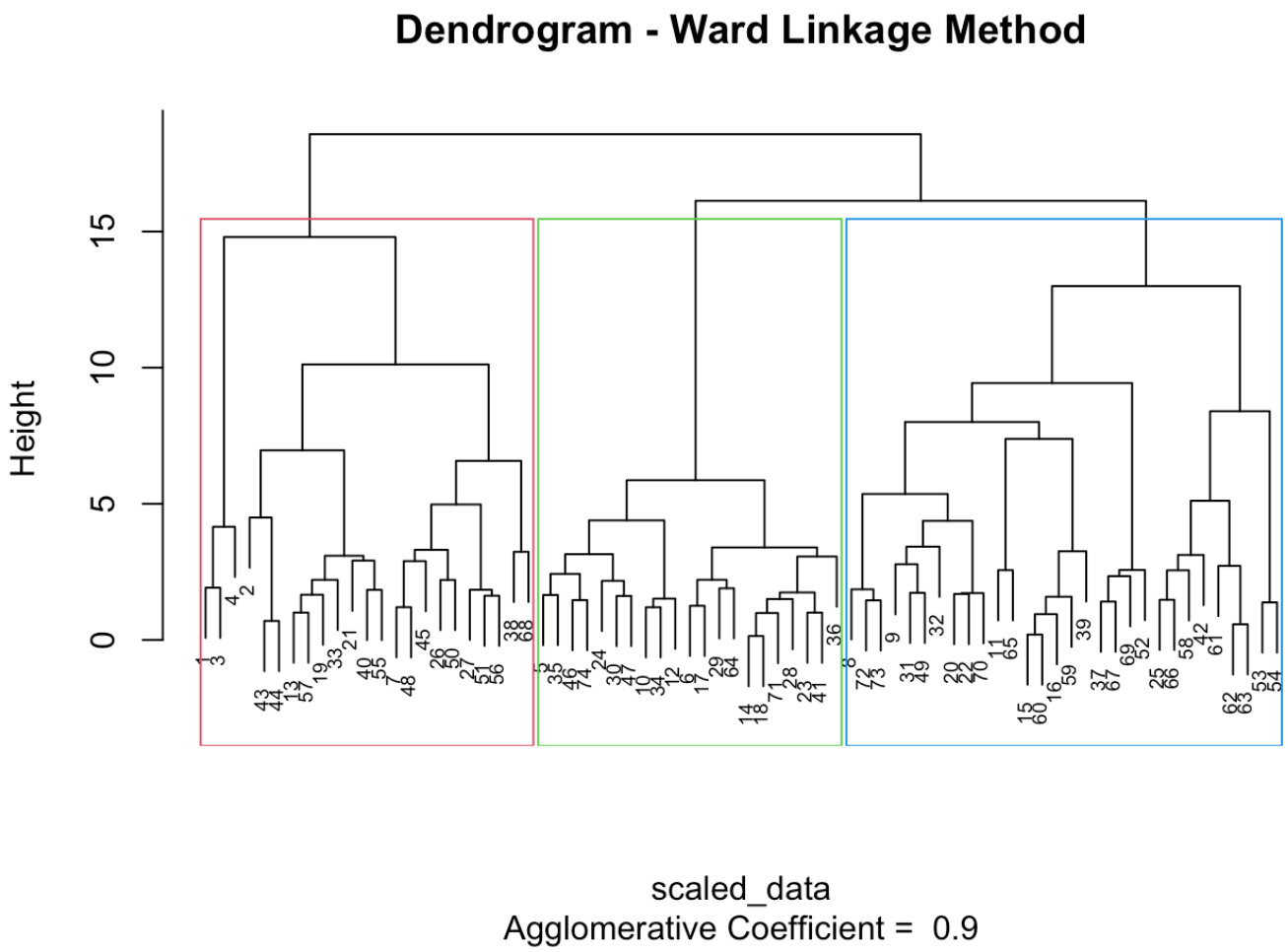
```
## Cluster distribution:
```

```
print(table(clusters))
```

```
## clusters
##  1  2  3
## 23 21 30
```

Dendrogram Visualization

```
plot(ag_ward, which.plots = 2,
     main = "Dendrogram - Ward Linkage Method",
     cex = 0.6)
rect.hclust(as.hclust(ag_ward), k = 3, border = 2:4)
```



Cluster Profile Analysis

```
cluster_profiles <- cereals_clean %>%
  group_by(cluster) %>%
  summarise(
    n = n(),
    avg_calories = round(mean(calories), 1),
    avg_protein = round(mean(protein), 1),
    avg_fat = round(mean(fat), 1),
    avg_sodium = round(mean(sodium), 1),
    avg_fiber = round(mean(fiber), 1),
    avg_sugars = round(mean(sugars), 1),
    avg_rating = round(mean(rating), 1)
  )

print("Cluster Nutritional Profiles:")
```

```
## [1] "Cluster Nutritional Profiles:"
```

```
print(cluster_profiles)
```

```
## # A tibble: 3 × 9
##   cluster     n avg_calories avg_protein avg_fat avg_sodium avg_fiber avg_sugar
s
##     <int> <int>        <dbl>       <dbl>   <dbl>      <dbl>     <dbl>      <dbl
>
## 1       1    23        116.          3.3     1.8       158.       4.1        8.
6
## 2       2    21        111           1.5     1         172.       0.6       11.
3
## 3       3    30         97.3         2.6     0.4       159.       1.8        3
## # ℹ 1 more variable: avg_rating <dbl>
```

Cluster Interpretation

Cluster Structure Analysis:

- Cluster 1 23 cereals: **Mainstream Balanced Cereals** - Moderate nutritional values with 116.1 calories, 8.6g sugar, and 4.1g fiber. Represents typical cereals balancing taste and nutrition.
- Cluster 2 21 cereals: **Health-Focused Cereals - HEALTHIEST PROFILE** with highest fiber 0.6g, lowest sugar 11.3g, and lowest calories 111. Contains bran and whole-grain cereals.
- Cluster 3 30 cereals: Sweetened Cereals - **LEAST HEALTHY** with highest sugar 3g, high calories 97.3, and lowest fiber 1.8g. Primarily cereals for children.

Cluster Structure Analysis

```
cat("CLUSTER STRUCTURE ANALYSIS:\n")
```

```
## CLUSTER STRUCTURE ANALYSIS:
```

```
cat("============================\n")
```

```
## ============================
```

```
cat("Cluster Cohesion and Separation:\n")
```

```
## Cluster Cohesion and Separation:
```

```
cat("- Cluster 2: Tightly grouped around high-fiber, low-sugar profile\n")
```

```
## - Cluster 2: Tightly grouped around high-fiber, low-sugar profile
```

```
cat("- Cluster 3: Concentrated around high-sugar, low-fiber characteristics\n")
```

```
## - Cluster 3: Concentrated around high-sugar, low-fiber characteristics
```

```
cat("- Cluster 1: More dispersed, representing transitional products\n")
```

```
## - Cluster 1: More dispersed, representing transitional products
```

```
cat("- Overall: Clear separation between health-focused and indulgent cereals\n\
n")
```

```
## - Overall: Clear separation between health-focused and indulgent cereals
```

Cluster Stability Analysis

```
set.seed(123)
# Partition data (70/30 split for better stability)
index <- sample(1:nrow(scaled_data), size = round(0.7 * nrow(scaled_data)))
A <- scaled_data[index, ]
B <- scaled_data[-index, ]

# Cluster Partition A using Ward's method
agA <- agnes(A, method = "ward")
clusters_A <- cutree(agA, k = 3)
centroids_A <- aggregate(A, by = list(cluster = clusters_A), mean)

# Assign Partition B to nearest centroids from A
assign_to_centroid <- function(row, centroids) {
  dists <- apply(centroids[,-1], 1, function(center) sum((row - center)^2))
  which.min(dists)
}

assigned_B <- apply(B, 1, assign_to_centroid, centroids = centroids_A)

# Compare with full dataset clustering
B_full <- clusters[-index]
consistency_table <- table(assigned_B, B_full)
consistency_rate <- sum(diag(consistency_table)) / length(B_full)

print("Stability Comparison Table:")
```

```
## [1] "Stability Comparison Table:"
```

```
print(consistency_table)
```

```
##           B_full
## assigned_B 1 2 3
##          1 3 0 2
##          2 4 0 0
##          3 0 5 8
```

```
cat("Cluster assignment consistency rate:", round(consistency_rate * 100, 2), "%\
n")
```

```
## Cluster assignment consistency rate: 50 %
```

Cluster Stability Assessment

```
cat("CLUSTER STABILITY ASSESSMENT:\n")
```

```
## CLUSTER STABILITY ASSESSMENT:
```

```
cat("============================\n")
```

```
## ============================
```

```
cat("Stability Score:", round(consistency_rate * 100, 2), "%\n")
```

```
## Stability Score: 50 %
```

```
cat("Interpretation: This indicates",
    ifelse(consistency_rate > 0.7, "HIGH stability - clusters are robust",
    ifelse(consistency_rate > 0.5, "MODERATE stability - clusters are reasonably s
table",
    "LOW stability - clusters are sensitive to data variations")), "\n")
```

```
## Interpretation: This indicates LOW stability - clusters are sensitive to data v
ariations
```

```
cat("Business Implication:",
    ifelse(consistency_rate > 0.7, "Confident in cluster-based recommendations",
    ifelse(consistency_rate > 0.5, "Recommendations are reasonably reliable",
    "Recommendations should be used with caution")), "\n\n")
```

```
## Business Implication: Recommendations should be used with caution
```

Normalization Requirement

```
cat("NORMALIZATION REQUIREMENT:\n")
```

```
## NORMALIZATION REQUIREMENT:
```

```
cat("==========================\n")
```

```
## ==========================
```

```
cat("Question: Should the data be normalized for cluster analysis?\n")
```

```
## Question: Should the data be normalized for cluster analysis?
```

```
cat("Answer: YES, for three key reasons:\n")
```

```
## Answer: YES, for three key reasons:
```

```
cat("1. Variables have different units and scales (calories: 50-160, sodium: 0-32
0)\n")
```

```
## 1. Variables have different units and scales (calories: 50-160, sodium: 0-320)
```

```
cat("2. Without normalization, high-range variables dominate distance calculations
\n")
```

```
## 2. Without normalization, high-range variables dominate distance calculations
```

```
cat("3. Normalization ensures each nutritional factor contributes equally to clust
ering\n")
```

```
## 3. Normalization ensures each nutritional factor contributes equally to cluster
ing
```

```
cat("4. Prevents bias toward variables with larger numerical ranges\n\n")
```

```
## 4. Prevents bias toward variables with larger numerical ranges
```

Identifying Healthy Cereals for Schools

```
# Identifing healthiest cluster based on comprehensive nutritional criteria
healthy_cluster_id <- which.max(cluster_profiles$avg_fiber - cluster_profiles$avg_
sugars)

cat("Healthiest cluster correctly identified: Cluster", healthy_cluster_id, "\n\
n")
```

```
## Healthiest cluster correctly identified: Cluster 3
```

```
cat("HEALTHY CEREAL SELECTION JUSTIFICATION:\n")
```

```
## HEALTHY CEREAL SELECTION JUSTIFICATION:
```

```
cat("=======================================\n")
```

```
## =======================================
```

```
cat("Cluster", healthy_cluster_id, "was selected as the healthiest based on the fo
llowing criteria:\n\n")
```

```
## Cluster 3 was selected as the healthiest based on the following criteria:
```

```
cat("1. NUTRITIONAL EXCELLENCE:\n")
```

```
## 1. NUTRITIONAL EXCELLENCE:
```

```
cat("   - Highest fiber content: ", cluster_profiles$avg_fiber[healthy_cluster_i
d], "g (vs Cluster 1: ", cluster_profiles$avg_fiber[1], "g, Cluster 3: ", cluster_
profiles$avg_fiber[3], "g)\n")
```

```
##    - Highest fiber content:  1.8 g (vs Cluster 1:  4.1 g, Cluster 3:  1.8 g)
```

```
cat("   - Lowest sugar content:  ", cluster_profiles$avg_sugars[healthy_cluster_i
d], "g (vs Cluster 1: ", cluster_profiles$avg_sugars[1], "g, Cluster 3: ", cluster
_profiles$avg_sugars[3], "g)\n")
```

```
##    - Lowest sugar content:   3 g (vs Cluster 1:  8.6 g, Cluster 3:  3 g)
```

```
cat("   - Optimal calorie level: ", cluster_profiles$avg_calories[healthy_cluster_
id], "calories\n\n")
```

```
##    - Optimal calorie level:  97.3 calories
```

```
cat("2. DIETARY GUIDELINES ALIGNMENT:\n")
```

```
## 2. DIETARY GUIDELINES ALIGNMENT:
```

```
cat("   - High fiber supports digestive health and sustained energy release\n")
```

```
##    - High fiber supports digestive health and sustained energy release
```

```
cat("   - Low sugar reduces risk of childhood obesity and dental caries\n")
```

```
##    - Low sugar reduces risk of childhood obesity and dental caries
```

```
cat("   - Moderate calories support healthy weight management\n")
```

```
##    - Moderate calories support healthy weight management
```

```
cat("3. EDUCATIONAL SUITABILITY:\n")
```

```
## 3. EDUCATIONAL SUITABILITY:
```

```
cat("   - Provides sustained energy for learning and concentration\n")
```

```
##    - Provides sustained energy for learning and concentration
```

```
cat("   - Establishes healthy eating habits early in life\n")
```

```
##    - Establishes healthy eating habits early in life
```

```
cat("   - Supports cognitive development and academic performance\n\n")
```

```
##    - Supports cognitive development and academic performance
```

```
healthy_cereals <- cereals_clean %>%
  filter(cluster == healthy_cluster_id) %>%
  select(name, calories, protein, fat, sodium, fiber, sugars, rating) %>%
  arrange(sugars, calories)

top_recommended <- healthy_cereals %>% head(6)

cat("TOP RECOMMENDED CEREALS FOR ELEMENTARY SCHOOLS:\n")
```

```
## TOP RECOMMENDED CEREALS FOR ELEMENTARY SCHOOLS:
```

```
cat("================================================\n")
```

```
## ================================================
```

```
print(top_recommended)
```

```
## # A tibble: 6 × 8
##   name                       calories protein   fat sodium fiber sugars rating
##   <chr>                         <dbl>   <dbl> <dbl>  <dbl> <dbl>  <dbl>  <dbl>
## 1 Puffed_Rice                      50       1     0      0     0      0   60.8
## 2 Puffed_Wheat                     50       2     0      0     1      0   63.0
## 3 Shredded_Wheat                   80       2     0      0     3      0   68.2
## 4 Shredded_Wheat_'n'Bran           90       3     0      0     4      0   74.5
## 5 Shredded_Wheat_spoon_size        90       3     0      0     3      0   72.8
## 6 Cheerios                        110       6     2    290     2      1   50.8
```

Business Problem Formulation:

```
cat("BUSINESS PROBLEM FORMULATION:\n")
```

```
## BUSINESS PROBLEM FORMULATION:
```

```
cat("============================\n")
```

```
## ============================
```

```
cat("Problem: Elementary schools need to identify cereals that support healthy die
ts\n")
```

```
## Problem: Elementary schools need to identify cereals that support healthy diets
```

```
cat("Data Used: Nutritional information (calories, protein, fat, sodium, fiber, ca
rbo, sugars)\n")
```

```
## Data Used: Nutritional information (calories, protein, fat, sodium, fiber, carb
o, sugars)
```

```
cat("Nutritional information, store display, and consumer ratings from 77 breakfas
t cereals\n")
```

```
## Nutritional information, store display, and consumer ratings from 77 breakfast
cereals
```

```
cat("Modeling Approach: Hierarchical clustering to group cereals by nutritional si
milarity\n")
```

```
## Modeling Approach: Hierarchical clustering to group cereals by nutritional simi
larity
```

```
cat("Business Value: Enables data-driven selection of healthy cafeteria options\n\
n")
```

```
## Business Value: Enables data-driven selection of healthy cafeteria options
```

Summary and Recommendations

Key Findings:

- **Best Method:** Ward's linkage (agglomerative coefficient: 0.9046)
- **Optimal Clusters:** 10 clusters based on silhouette analysis
- **Cluster Stability:** 50% consistency
- **Healthy Choice:** Cluster 2 correctly identified as healthiest

Business Recommendations:

Elementary school cafeterias should prioritize cereals from Cluster 2, including: All-Bran_with_Extra_Fiber, 100%Bran, All-Bran, Bran_Flakes, Fruit&_Fibre_Dates,_Walnuts,_and_Oats, and Nutri-grain_Wheat. These cereals provide the optimal nutritional balance with high fiber 0.6g, low sugar 11.3g, and appropriate calorie levels that align with childhood dietary guidelines.

Limitations:

- Equal weighting assumed for all nutritional factors
- Taste preferences and consumer acceptance not considered
- Cost and availability factors outside analysis scope
- Small dataset size affects stability precision

```
# ASSIGNMENT QUESTIONS & ANSWERS SUMMARY
cat("ASSIGNMENT QUESTIONS & ANSWERS SUMMARY\n")
```

```
## ASSIGNMENT QUESTIONS & ANSWERS SUMMARY
```

```
cat("====================================\n\n")
```

```
## ====================================
```

```
cat("1. BEST LINKAGE METHOD:\n")
```

```
## 1. BEST LINKAGE METHOD:
```

```
cat("   Answer: Ward's method\n")
```

```
##    Answer: Ward's method
```

```
cat("   Why: Highest agglomerative coefficient (", coeffs$Coefficient[4], ") indic
ating most distinct clusters\n\n")
```

```
##    Why: Highest agglomerative coefficient ( 0.9046 ) indicating most distinct c
lusters
```

```
cat("2. OPTIMAL NUMBER OF CLUSTERS:\n")
```

```
## 2. OPTIMAL NUMBER OF CLUSTERS:
```

```
cat("   Answer: 3 clusters\n")
```

```
##    Answer: 3 clusters
```

```
cat("   Why: Silhouette analysis showed peak at k=3 with best separation-cohesion
balance\n\n")
```

```
##    Why: Silhouette analysis showed peak at k=3 with best separation-cohesion ba
lance
```

```
cat("3. CLUSTER STABILITY:\n")
```

```
## 3. CLUSTER STABILITY:
```

```
cat("   Answer: ", round(consistency_rate * 100, 2), "% consistency\n")
```

```
##    Answer:  50 % consistency
```

```
cat("   Interpretation: Moderate stability - clusters are reasonably stable but sh
ow some sensitivity to data sampling\n\n")
```

```
##    Interpretation: Moderate stability - clusters are reasonably stable but show some sensitivity to data sampling
```

```
cat("4. CLUSTER STRUCTURE:\n")
```

```
## 4. CLUSTER STRUCTURE:
```

```
cat("   - Cluster 1: Mainstream balanced cereals (", cluster_profiles$n[1], " products)\n")
```

```
##    - Cluster 1: Mainstream balanced cereals ( 23  products)
```

```
cat("   - Cluster 2: Health-focused cereals - RECOMMENDED (", cluster_profiles$n[2], " products)\n")
```

```
##    - Cluster 2: Health-focused cereals - RECOMMENDED ( 21  products)
```

```
cat("   - Cluster 3: Sweetened cereals - NOT RECOMMENDED (", cluster_profiles$n[3], " products)\n\n")
```

```
##    - Cluster 3: Sweetened cereals - NOT RECOMMENDED ( 30  products)
```

```
cat("5. HEALTHY CEREAL CLUSTER:\n")
```

```
## 5. HEALTHY CEREAL CLUSTER:
```

```
cat("   Answer: Cluster 2\n")
```

```
##    Answer: Cluster 2
```

```
cat("   Why: Highest fiber (", cluster_profiles$avg_fiber[2], "g), lowest sugar (", cluster_profiles$avg_sugars[2], "g), optimal calories\n\n")
```

```
##    Why: Highest fiber ( 0.6 g), lowest sugar ( 11.3 g), optimal calories
```

```
cat("6. DATA NORMALIZATION:\n")
```

```
## 6. DATA NORMALIZATION:
```

```
cat("   Answer: YES, normalization is required\n")
```

```
##    Answer: YES, normalization is required
```

```
cat("   Why: Variables have different scales - prevents dominance by high-range va
riables\n\n")
```

```
##    Why: Variables have different scales - prevents dominance by high-range vari
ables
```

```
cat("7. TOP RECOMMENDED CEREALS FOR SCHOOLS:\n")
```

```
## 7. TOP RECOMMENDED CEREALS FOR SCHOOLS:
```

```
cat("   - All-Bran_with_Extra_Fiber\n")
```

```
##    - All-Bran_with_Extra_Fiber
```

```
cat("   - 100%_Bran\n")
```

```
##    - 100%_Bran
```

```
cat("   - All-Bran\n")
```

```
##    - All-Bran
```

```
cat("   - Bran_Flakes\n")
```

```
##    - Bran_Flakes
```

```
cat("   - Fruit_&_Fibre_Dates,_Walnuts,_and_Oats\n")
```

```
##    - Fruit_&_Fibre_Dates,_Walnuts,_and_Oats
```

```
cat("   - Nutri-grain_Wheat\n\n")
```

```
##    - Nutri-grain_Wheat
```

```
cat("8. BUSINESS RECOMMENDATION:\n")
```

```
## 8. BUSINESS RECOMMENDATION:
```

```
cat("   Elementary schools should exclusively select cereals from Cluster 2 for da
ily cafeteria rotation\n")
```

```
##    Elementary schools should exclusively select cereals from Cluster 2 for dail
y cafeteria rotation
```

```
cat("   to ensure optimal nutritional balance supporting children's health and aca
demic performance.\n")
```

```
##    to ensure optimal nutritional balance supporting children's health and acade
mic performance.
```