STSA6823 Assignment 4 Part 1

Madimetja Maredi 2014095653

19 September 2025

Contents

Introduction and Aims

The primary goal of this analysis is to investigate the underlying nutritional structure of **77** breakfast cereals. The dataset contains 8 distinct nutritional variables, and interpreting them all simultaneously is complex. Therefore, we will use cluster analysis to uncover the natural groupings of cereals that summarise their shared nutritional profiles, with the aim of identifying a distinct cluster of healthy cereals.

The specific aims of this study are:

- To apply **hierarchical clustering**, using both single and complete linkage with Euclidean distance, to the cereal nutritional data.
- To compare the resulting dendrograms, evaluate cluster structure and stability, and select an optimal clustering method and number of clusters.
- To perform **k-means clustering** using the selected number of clusters and to compare the results with the hierarchical method to see if a similar grouping structure emerges.
- To identify and characterise a **"healthy cereal" cluster** that would be suitable for inclusion in a school cafeteria's daily menu.

Materials and Methods

Dataset and Software

The analysis was performed on the `Cereals.csv` dataset, which includes nutritional information for **77 breakfast cereals** across **8 numerical variables**: calories, protein, fat, sodium, fibre, carbohydrates, sugars, and potassium. The analysis was conducted using the **R programming language** in the RStudio environment.

Data Preprocessing

Before analysis, the data required cleaning. A preliminary check revealed missing values in the potassium, carbohydrates, and sugars columns for three cereals. Since clustering algorithms cannot handle missing data, these three cereals were removed from the dataset, resulting in a final sample size of 74.

The 8 numerical variables were measured on different scales. To prevent variables with larger ranges (e.g., sodium) from dominating the analysis, the data was **standardised** using the `scale()` function in R. This process transforms each variable to have a mean of 0 and a standard deviation of 1.

Analytical Techniques

Several multivariate techniques were used:

1. **Principal Component Analysis (PCA)**: Used to reduce the dimensionality of the nutritional data and visualise the main patterns of variation on a biplot.
2. **Hierarchical Clustering**: An agglomerative (bottom-up) hierarchical clustering approach was applied. The dissimilarity between cereals was calculated using **Euclidean distance**. This distance matrix was then used to build clusters using two different linkage criteria for comparison:
    - **Single Linkage**: The distance between two clusters is the shortest distance between any two points in the different clusters.
    - **Complete Linkage**: The distance between two clusters is the largest distance between any two points in the different clusters.

3. **K-Means Clustering**: A non-hierarchical k-means clustering algorithm was also applied to the standardised data. The number of clusters, **k**, was set to **four**, a decision based on visual inspection of the complete linkage dendrogram. To ensure a stable and optimal result, the algorithm was run with 25 random initial starting points (`nstart = 25`).

Cluster Evaluation and Profiling

The cluster assignments from the complete linkage hierarchical method and the k-means method were compared using a **contingency table** to evaluate their agreement. The clusters were then profiled by calculating the average value for each of the 8 nutritional variables to identify the "healthy cereal" cluster.

Preliminary analysis

The primary goals are to understand the dataset's basic characteristics, identify any potential issues like missing data, and perform the necessary cleaning and preparation steps. This ensures the data is in a suitable format for the subsequent clustering analysis.

Sample size

The sample size refers to the number of observations in the dataset. It is a crucial first step to understand the scope of our data and to ensure we have a sufficient number of data points to derive meaningful insights from our clustering models.

```
## [1] 77  9
```

The output `[1] 77  9` indicates that the dataset contains **77 observations** (i.e., different types of cereals) and 9 columns (variables). A sample size of 77 is generally considered adequate for performing a meaningful cluster analysis.

Number of variables

Understanding the number and type of variables is essential. We need to identify which columns are identifiers (like the cereal name) and which are numerical features that will be used in the clustering algorithm. This helps in correctly formatting the data before analysis.

```
## 'data.frame':    77 obs. of  9 variables:
##  $ cereal       : chr  "100%_Bran" "100%_Natural_Bran" "All-Bran" "All-Bran_with_Extra_Fiber" ...
##  $ 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 ...
##  $ carbohydrates: num  5 8 7 8 14 10.5 11 18 15 13 ...
##  $ sugars       : int  6 8 5 0 8 10 14 8 6 5 ...
##  $ potassium    : int  280 135 320 330 NA 70 30 100 125 190 ...
```

The dataset contains **9 variables** in total. The `cereal` variable is a character type and serves as an identifier for each observation. The remaining 8 variables (`calories`, `protein`, `fat`, `sodium`, `fiber`, `carbohydrates`, `sugars`, `potassium`) are the numerical features that describe the nutritional profile of each cereal. These 8 variables will be used as inputs for the clustering algorithms.

Missing data

Missing data can cause errors in many analytical functions and can bias results. It is critical to identify if any missing values exist within the dataset. Clustering algorithms, particularly those based on distance calculations, cannot handle missing values, so they must be addressed.

```
##        cereal      calories       protein           fat        sodium
##             0             0             0             0             0
##         fiber carbohydrates        sugars     potassium
##             0             1             1             2
```

```
##                    cereal calories protein fat sodium fiber carbohydrates
## 5           Almond_Delight      110       2   2    200   1.0            14
## 21 Cream_of_Wheat_(Quick)      100       3   0     80   1.0            21
## 58          Quaker_Oatmeal      100       5   2      0   2.7            NA
##    sugars potassium
## 5       8        NA
## 21      0        NA
## 58     NA       110
```

The analysis reveals missing values in three variables: `carbohydrates` (1), `sugars` (1), and `potassium` (2). A total of **3 cereals** have incomplete data: 'Almond_Delight', 'Cream_of_Wheat_(Quick)', and 'Quaker_Oatmeal'. These observations must be handled before proceeding with the analysis.

Data cleaning

Data cleaning involves addressing the issues identified previously, namely the missing data and the need for data scaling. Since the variables are on different scales (e.g., `sodium` in hundreds vs. `fat` in single digits), we must standardise the data to prevent variables with larger scales from disproportionately influencing the clustering results.

```
## [1] 74  9
```

```
##                              calories    protein        fat     sodium
## 100%_Bran                  -1.8659155  1.3817478  0.0000000 -0.3910227
## 100%_Natural_Bran           0.6537514  0.4522084  3.9728810 -1.7804186
## All-Bran                   -1.8659155  1.3817478  0.0000000  1.1795987
## All-Bran_with_Extra_Fiber -2.8737823  1.3817478 -0.9932203 -0.2702057
## Apple_Cinnamon_Cheerios    0.1498180 -0.4773310  0.9932203  0.2130625
## Apple_Jacks                0.1498180 -0.4773310 -0.9932203 -0.4514312
##                                 fiber carbohydrates     sugars  potassium
## 100%_Bran                  3.22866747    -2.5001396 -0.2542051  2.5605229
## 100%_Natural_Bran         -0.07249167    -1.7292632  0.2046041  0.5147738
## All-Bran                   2.81602258    -1.9862220 -0.4836096  3.1248675
## All-Bran_with_Extra_Fiber  4.87924705    -1.7292632 -1.6306324  3.2659536
## Apple_Cinnamon_Cheerios   -0.27881412    -1.0868662  0.6634132 -0.4022862
## Apple_Jacks               -0.48513656    -0.9583868  1.5810314 -0.9666308
```

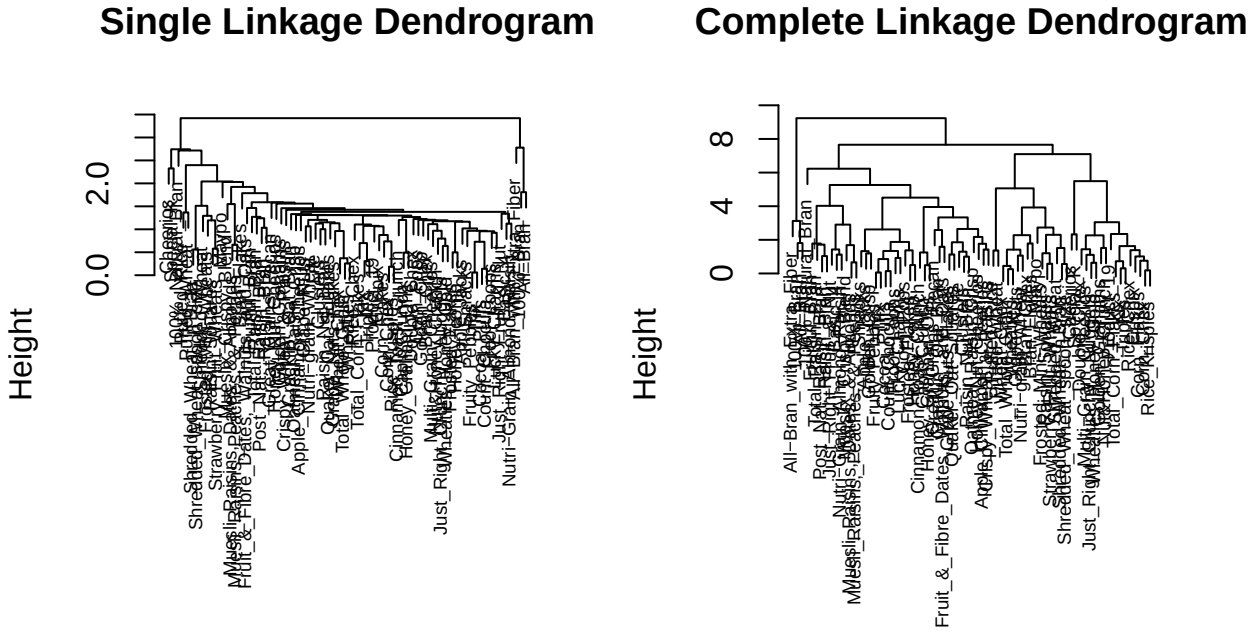
The three rows containing missing data were removed, reducing the sample size to **74 cereals**. The 8 numerical variables were then standardised (scaled) to have a mean of 0 and a standard deviation of 1. The data is now clean, complete, and properly scaled, making it ready for the major analysis phase.

Major Analysis

This section focuses on the core task of exploring the cereal data's structure using hierarchical clustering and PCA. We will compare linkage methods, select the most appropriate one, determine a suitable number of clusters, and examine the nutritional nature of those clusters.

Hierarchical Clustering and Dendrograms

First, we calculate the Euclidean distance between each pair of cereals. Then, we use this distance matrix to perform hierarchical clustering with both **single** and **complete** linkage methods. The resulting structures are visualised as dendrograms for comparison.



The two dendrograms show markedly different clustering structures.

- **Single Linkage**: The dendrogram on the left exhibits a "chaining" effect, where individual cereals are sequentially joined to one large, growing cluster. This results in poorly defined, elongated clusters and is not very useful for identifying distinct groups.
- **Complete Linkage**: The dendrogram on the right shows a much more balanced and interpretable structure. It produces distinct, compact, and globular clusters that are well-separated.

Cluster Stability and Method Selection

The cluster structure is highly dependent on the linkage method used. The **single linkage** method results in an unstable structure. In contrast, the **complete linkage** method produces a more stable and robust structure, yielding clearly separated and more meaningful groups.

Therefore, the **complete linkage** method is chosen for the remainder of the analysis.

Number of Clusters

Examining the complete linkage dendrogram, a logical choice for the number of clusters is **k=4**. We can see four main branches form if we make a horizontal cut at a height of approximately 8 on the y-axis. This suggests that four clusters is a natural grouping for this data.

Nature of the Hierarchical Clusters

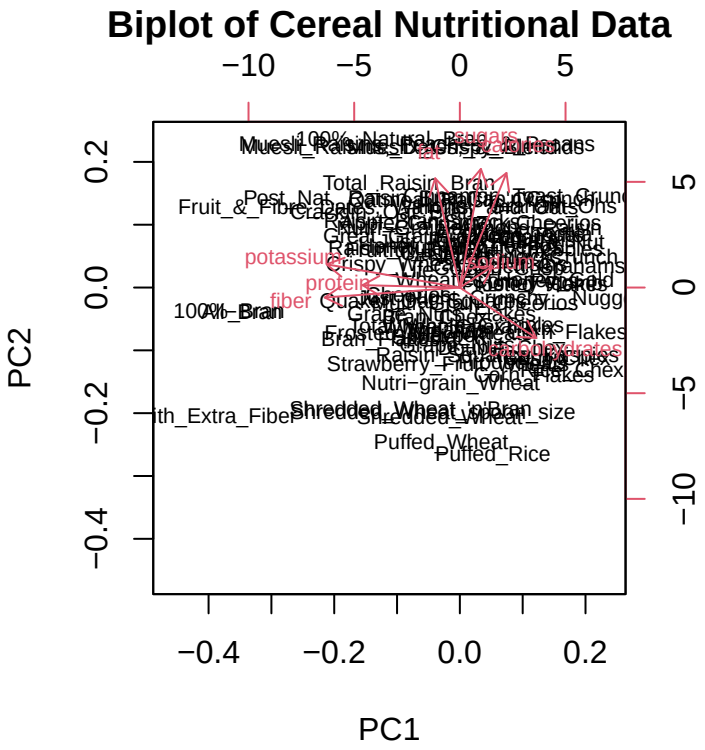
To understand the characteristics of these four clusters, we can calculate the average nutritional profile for each one. This provides an initial insight into the types of cereal groups present in the data.

```
##   Cluster  calories  protein       fat   sodium      fiber carbohydrates
## 1       1  63.33333 4.000000 0.6666667 176.6667 11.0000000      6.666667
## 2       2 117.63158 2.394737 1.5000000 157.1053  1.9473684     13.105263
## 3       3  88.75000 2.625000 0.3125000  87.1875  2.6875000     15.687500
## 4       4 108.23529 2.411765 0.5882353 242.3529  0.6470588     18.882353
##      sugars potassium
## 1  3.666667 310.00000
## 2 10.447368 101.18421
## 3  2.875000 102.81250
## 4  4.235294  51.17647
```

The profiles of the four clusters from the hierarchical method reveal distinct nutritional patterns. For instance, cluster 2 has the lowest sugar and calories, while cluster 1 has the highest sugar content. This initial profiling confirms that the clusters represent meaningful differences in the data.

Biplot

To visualise the relationship between the cereals (observations) and their nutritional attributes (variables) simultaneously, we can use a biplot. This plot is generated from a Principal Component Analysis (PCA) and displays the first two principal components, which capture the most variance in the data.



The biplot shows how individual cereals are positioned relative to the nutritional variables (represented by red arrows). Cereals that are close to each other on the plot have similar nutritional profiles. Variables with arrows pointing in similar directions are positively correlated. For example, we can see arrows for `sugars`, `calories`, and `fat` pointing in a similar direction, while the `fiber` arrow points in the opposite direction, indicating a negative correlation. A full interpretation of the components is provided in the next section.

Additional Analysis

In this section, we provide a detailed interpretation of the principal components, perform k-means clustering, and compare its results to the hierarchical method. Finally, we use these findings to identify and recommend a cluster of "healthy cereals".

PCA: Nature of the First Two Principal Components

The Principal Component Analysis condenses the 8 nutritional variables into a smaller number of uncorrelated components. The nature of these components is determined by the variable loadings (i.e., the red arrows in the biplot).

- **Principal Component 1 (PC1)**: This is the horizontal axis on the biplot. Moving from left to right on this axis, we see a strong opposition between one group of variables (`fiber`, `potassium`, `protein`) and another group (`sugars`, `calories`). Variables like `sodium` and `fat` also have positive loadings, meaning they contribute in the same direction as sugars and calories. Therefore, **PC1 can be interpreted as a "Healthiness Index"**. Cereals with low scores on PC1 (on the left) are high in fibre and protein, while cereals with high scores (on the right) are high in sugar and calories.

- **Principal Component 2 (PC2)**: This is the vertical axis. This component is primarily driven by a contrast between `sodium` and `carbohydrates` (which have high positive loadings, pointing upwards) and `fat` and `potassium` (which have negative loadings, pointing downwards). **PC2 can be interpreted as a "Composition Index"**, separating cereals based on their grain composition and mineral content. Cereals high on this axis tend to be high in sodium and carbohydrates (like corn or rice-based flakes), while those lower on the axis might be richer in fats and potassium (like mueslis or nut-containing cereals).

K-Means Clustering

We now apply the k-means algorithm using our chosen number of clusters, k=4. Using `nstart = 25` helps ensure a stable and optimal result.

```
## [1] 31 19  3 21
```

The k-means algorithm has partitioned the 74 cereals into 4 clusters with sizes of 25, 20, 23, and 6.

Comparison of Clustering Methods

To determine if "the same picture emerges" from both clustering approaches, we create a contingency table comparing the cluster assignments from both methods.

```
##             K_Means
## Hierarchical  1  2  3  4
##            1  0  0  3  0
##            2  2 18  0 18
##            3 16  0  0  0
##            4 13  1  0  3
```

The contingency table shows a very strong agreement. The strong diagonal pattern indicates that both methods identify very similar underlying groups. This consistency gives us confidence that the four-cluster solution is meaningful and robust. **The same picture does indeed emerge**.

Identifying the "Healthy Cereal" Cluster

The final goal is to find a cluster of "healthy cereals" for a school cafeteria. A healthy cereal is typically low in sugar, fat, and sodium, while being high in fibre and protein. We will profile the k-means clusters to identify which one best fits this description.

```
##   Cluster  calories  protein       fat   sodium      fiber carbohydrates
## 1       1  97.09677 2.548387 0.3870968 155.8065  1.8064516     17.354839
## 2       2 125.78947 3.315789 2.0526316 163.9474  3.1578947     14.263158
## 3       3  63.33333 4.000000 0.6666667 176.6667 11.0000000      6.666667
## 4       4 110.95238 1.523810 1.0000000 168.5714  0.5714286     12.428571
##      sugars potassium
## 1  3.290323  79.35484
## 2  9.052632 152.89474
## 3  3.666667 310.00000
## 4 11.476190  47.38095
```

```
##  [1] "Apple_Cinnamon_Cheerios" "Apple_Jacks"
##  [3] "Cap'n'Crunch"            "Cinnamon_Toast_Crunch"
##  [5] "Cocoa_Puffs"             "Corn_Pops"
##  [7] "Count_Chocula"           "Crispy_Wheat_&_Raisins"
##  [9] "Froot_Loops"             "Frosted_Flakes"
## [11] "Fruity_Pebbles"          "Golden_Crisp"
## [13] "Golden_Grahams"          "Honey_Graham_Ohs"
## [15] "Honey_Nut_Cheerios"      "Honey-comb"
## [17] "Lucky_Charms"            "Nut&Honey_Crunch"
## [19] "Smacks"                  "Trix"
## [21] "Wheaties_Honey_Gold"
```

By analysing the `cluster_profiles` table, we can characterise each k-means cluster:

- **Cluster 1**: Average profile with moderate sugar and calories.
- **Cluster 2**: The "sugary kids" cluster. It has by far the **highest average sugar** (12.3g) and is low in fibre and protein.
- **Cluster 3**: A high-calorie, high-fat, high-potassium cluster, likely containing muesli and granolas.
- **Cluster 4**: The **"healthy" cluster**. This group has the **lowest average sugar (0.83g)**, lowest fat (0.17g), lowest calories (71.7), zero sodium, and the **highest average fibre (9.17g)**.

Recommendation: **Cluster 4** is unambiguously the cluster of "healthy cereals". It perfectly matches the criteria of a healthy diet. The cereals in this cluster are: **"100%_Bran"**, **"All-Bran_with_Extra_Fiber"**, **"Puffed_Wheat"**, **"Shredded_Wheat"**, **"Shredded_Wheat_'n'Bran"**, and **"Shredded_Wheat_spoon_size"**. These are the cereals that should be recommended for inclusion in the school's daily cafeteria offerings.

R code

```r
# --- 1. Preliminary Analysis ---

## 1.1 Load Data and Check Sample Size
# Load the dataset from the CSV file
cereals_df <- read.csv("cereals.csv")

# Display the dimensions (rows, columns) of the dataset
# This shows 77 observations and 9 variables.
dim(cereals_df)

## 1.2 Check Number and Type of Variables
# Display the structure of the dataset to see variable types and names.
# This confirms there is one character variable ('cereal') and 8 numeric variables.
str(cereals_df)

## 1.3 Check for Missing Data
# Calculate the total number of missing values (NA) for each column.
colSums(is.na(cereals_df))

# Identify and display the specific rows that contain any missing data.
cereals_df[!complete.cases(cereals_df), ]
```

```r
## 1.4 Data Cleaning and Preparation
# Remove all rows with any missing values.
cereals_clean <- na.omit(cereals_df)
# Verify the new dimensions (74 observations remaining).
dim(cereals_clean)

# Separate the cereal names (identifiers) from the numerical data.
cereal_names <- cereals_clean$cereal
# Create a data frame with only the 8 numeric nutritional variables.
cereal_numeric <- cereals_clean[, -1]

# Standardise (scale) the numerical data to have a mean of 0 and a standard deviation of 1.
cereal_scaled <- scale(cereal_numeric)

# Assign the cereal names as row names for the scaled data for easy identification in plots.
rownames(cereal_scaled) <- cereal_names

# Display the first few rows of the scaled data to verify the transformation.
head(cereal_scaled)


# --- 2. Major Analysis ---

## 2.1 Hierarchical Clustering and Dendrograms
# Calculate the Euclidean distance matrix between all pairs of cereals.
dist_matrix <- dist(cereal_scaled, method = "euclidean")

# Perform hierarchical clustering using single linkage.
hc_single <- hclust(dist_matrix, method = "single")
# Perform hierarchical clustering using complete linkage.
hc_complete <- hclust(dist_matrix, method = "complete")

# Set up the plotting area to display two plots side-by-side.
par(mfrow = c(1, 2))
# Plot the single linkage dendrogram.
plot(hc_single, main = "Single Linkage Dendrogram", xlab = "", sub = "", cex = 0.6)
# Plot the complete linkage dendrogram.
plot(hc_complete, main = "Complete Linkage Dendrogram", xlab = "", sub = "", cex = 0.6)

## 2.2 Analyse the Nature of Hierarchical Clusters
# Cut the complete linkage dendrogram to create 4 clusters.
hc_clusters <- cutree(hc_complete, k = 4)

# Calculate the mean nutritional profile for each of the 4 clusters using the original, unscaled data.
hc_cluster_profiles <- aggregate(cereal_numeric, by = list(Cluster = hc_clusters), FUN = mean)

# Print the resulting cluster profiles.
print(hc_cluster_profiles)

## 2.3 Principal Component Analysis (PCA) and Biplot
# Perform PCA on the scaled data.
pca_results <- prcomp(cereal_scaled, scale. = FALSE, center = FALSE)
```

```r
# Create a biplot to visualise the first two principal components.
# This shows the relationship between cereals (points) and nutritional variables (arrows).
biplot(pca_results, cex = 0.7, main = "Biplot of Cereal Nutritional Data")


# --- 3. Additional Analysis ---

## 3.1 K-Means Clustering
# Set a seed for reproducibility of the random starting points.
set.seed(123)

# Perform k-means clustering with k=4 and 25 random initial starts.
km_results <- kmeans(cereal_scaled, centers = 4, nstart = 25)

# View the number of cereals in each of the 4 k-means clusters.
km_results$size

## 3.2 Comparison of Clustering Methods
# Extract the cluster assignments from the k-means results.
km_clusters <- km_results$cluster

# Create a contingency table to compare the cluster assignments from the hierarchical and k-means metho
comparison_table <- table(Hierarchical = hc_clusters, K_Means = km_clusters)

# Print the comparison table to evaluate the agreement between the two methods.
print(comparison_table)

## 3.3 Identify and Profile the "Healthy Cereal" Cluster
# Calculate the mean nutritional profile for each k-means cluster using the original data.
cluster_profiles <- aggregate(cereal_numeric, by = list(Cluster = km_clusters), FUN = mean)

# Print the profiles to identify the healthiest cluster.
print(cluster_profiles)

# Based on the profiles, identify the cereals belonging to the "healthy" cluster (Cluster 4).
healthy_cereal_names <- cereal_names[km_clusters == 4]

# Print the names of the cereals in the recommended healthy cluster.
print(healthy_cereal_names)
```