K-Means Clustering Analysis of Protein Consumption

Amandeep Randhawa

2025-05-25

**Introduction**

This project applies **K-means clustering** to analyze protein consumption patterns across 25 European countries. The dataset includes intake percentages of various protein sources such as red meat, white meat, eggs, milk, fish, cereals, starch, nuts, and fruits & vegetables. Our objective is to uncover distinct dietary patterns and group countries into meaningful clusters based on their protein preferences.

We begin by exploring a simplified 2-variable model (RedMeat and WhiteMeat), followed by a full-dimensional clustering using all available variables. Visualizations accompany each step to support interpretation and reveal patterns.

# Install packages only if not already installed  
if (!require(factoextra)) install.packages("factoextra", dependencies = TRUE)

## Loading required package: factoextra

## Loading required package: ggplot2

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

if (!require(ggplot2)) install.packages("ggplot2")  
if (!require(gridExtra)) install.packages("gridExtra")

## Loading required package: gridExtra

if (!require(cluster)) install.packages("cluster")

## Loading required package: cluster

# Load libraries  
library(ggplot2)  
library(gridExtra)  
library(cluster)  
library(factoextra)

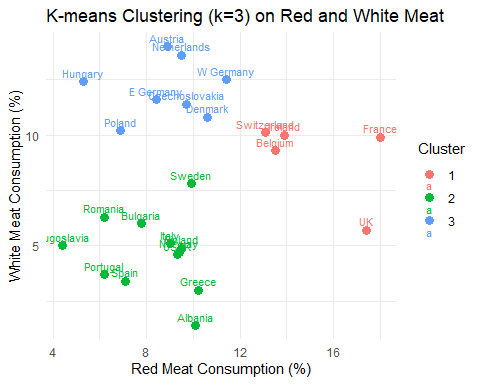
# Define file path (use forward slashes for portability)  
file\_path <- "C:/Users/amand/OneDrive/Documents/course 2 Data Mining U of G/Data mining Course 2/Assignment 2/protein.csv"  
  
# Check if file exists  
if (!file.exists(file\_path)) {  
 stop("Dataset file not found. Please check the file path.")  
}  
  
# Load dataset (assuming comma-separated; adjust sep = "\t" if tab-separated)  
protein\_data <- read.csv(file\_path, header = TRUE, row.names = 1)  
  
# Verify column names  
expected\_columns <- c("RedMeat", "WhiteMeat", "Eggs", "Milk", "Fish", "Cereals", "Starch", "Nuts", "Fr.Veg")  
if (!all(expected\_columns %in% colnames(protein\_data))) {  
 stop("Dataset does not contain all expected columns.")  
}  
  
# Check for missing values  
if (any(is.na(protein\_data))) {  
 stop("Dataset contains missing values. Please handle them before proceeding.")  
}  
  
# View first few rows  
head(protein\_data)

## RedMeat WhiteMeat Eggs Milk Fish Cereals Starch Nuts Fr.Veg  
## Albania 10.1 1.4 0.5 8.9 0.2 42.3 0.6 5.5 1.7  
## Austria 8.9 14.0 4.3 19.9 2.1 28.0 3.6 1.3 4.3  
## Belgium 13.5 9.3 4.1 17.5 4.5 26.6 5.7 2.1 4.0  
## Bulgaria 7.8 6.0 1.6 8.3 1.2 56.7 1.1 3.7 4.2  
## Czechoslovakia 9.7 11.4 2.8 12.5 2.0 34.3 5.0 1.1 4.0  
## Denmark 10.6 10.8 3.7 25.0 9.9 21.9 4.8 0.7 2.4

# Subset RedMeat and WhiteMeat  
meat\_data <- protein\_data[, c("RedMeat", "WhiteMeat")]  
  
# Apply K-means clustering with k = 3  
set.seed(123)  
k3 <- kmeans(meat\_data, centers = 3, nstart = 25)  
  
# Add cluster assignments to data  
meat\_data$Cluster <- as.factor(k3$cluster)  
  
# Print cluster assignments  
print(k3$cluster)

## Albania Austria Belgium Bulgaria Czechoslovakia   
## 2 3 1 2 3   
## Denmark E Germany Finland France Greece   
## 3 3 2 1 2   
## Hungary Ireland Italy Netherlands Norway   
## 3 1 2 3 2   
## Poland Portugal Romania Spain Sweden   
## 3 2 2 2 2   
## Switzerland UK USSR W Germany Yugoslavia   
## 1 1 2 3 2

# Visualize clusters  
ggplot(meat\_data, aes(x = RedMeat, y = WhiteMeat, color = Cluster, label = rownames(meat\_data))) +  
 geom\_point(size = 3) +  
 geom\_text(vjust = -0.5, size = 3) +  
 labs(title = "K-means Clustering (k=3) on Red and White Meat",  
 x = "Red Meat Consumption (%)",  
 y = "White Meat Consumption (%)") +  
 theme\_minimal()



Interpretation of Clustering Results

The K-means clustering results are used to understand the dietary patterns of European countries based on their protein consumption.

K-means Clustering with k=3 (RedMeat and WhiteMeat)

The k=3 clustering on RedMeat and WhiteMeat divides the 25 countries into three groups, as shown in the scatter plot. The cluster assignments and plot suggest:

Cluster 1 (red): Countries like France, UK, Belgium, and Switzerland stand out with higher RedMeat (10–16%) and moderate WhiteMeat (6–10%). These Western European spots seem to love their red meat, maybe thanks to wealth or tradition.

Cluster 2 (green): This group, including Sweden, Italy, Greece, Albania, Bulgaria, Romania, Portugal, Yugoslavia, and Spain, keeps RedMeat lower (4–10%) and WhiteMeat modest (4–8%). It’s a mix of Southern and Eastern countries, possibly leaning on plants or other proteins instead.

Cluster 3 (blue): Austria, W Germany, E Germany, Czechoslovakia, Denmark, Poland, Hungary, and Netherlands fall here, with moderate RedMeat (8–12%) and higher WhiteMeat (8–12%). This Central and Northern crew might prefer poultry or balanced diets.

# Normalize the data  
protein\_scaled <- scale(protein\_data)  
  
# Apply K-means clustering with k = 7  
set.seed(123)  
k7 <- kmeans(protein\_scaled, centers = 7, nstart = 25)  
  
# Add cluster assignments to original data  
protein\_data$Cluster <- as.factor(k7$cluster)  
  
# Summarize cluster sizes  
cat("Cluster sizes:\n")

## Cluster sizes:

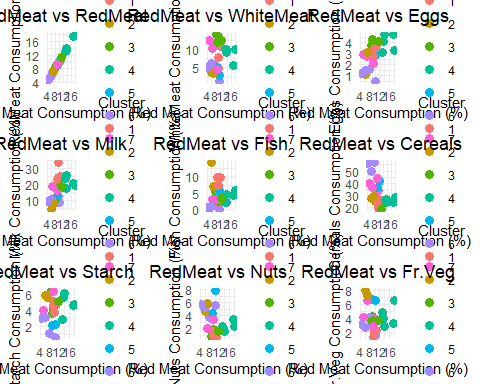
table(k7$cluster)

##   
## 1 2 3 4 5 6 7   
## 4 2 4 5 2 4 4

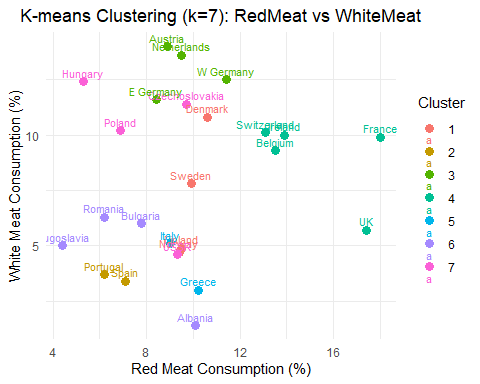
# List of variables to plot against RedMeat  
variables <- setdiff(colnames(protein\_data), c("Cluster", "Country"))  
  
# Generate scatter plots  
plots <- list()  
for (i in seq\_along(variables)) {  
 var <- variables[i]  
 p <- ggplot(protein\_data, aes\_string(x = "RedMeat", y = var, color = "Cluster")) +  
 geom\_point(size = 3) +  
 labs(title = paste("RedMeat vs", var),  
 x = "Red Meat Consumption (%)",  
 y = paste(var, "Consumption (%)")) +  
 theme\_minimal() +  
 theme(plot.title = element\_text(hjust = 0.5))  
 plots[[i]] <- p  
}

## Warning: `aes\_string()` was deprecated in ggplot2 3.0.0.  
## ℹ Please use tidy evaluation idioms with `aes()`.  
## ℹ See also `vignette("ggplot2-in-packages")` for more information.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.

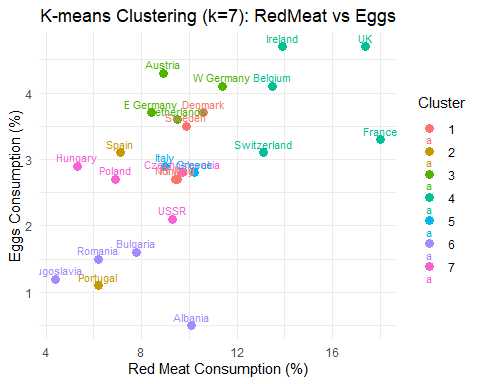
# Display plots in a 3x3 grid  
grid.arrange(grobs = plots, ncol = 3)



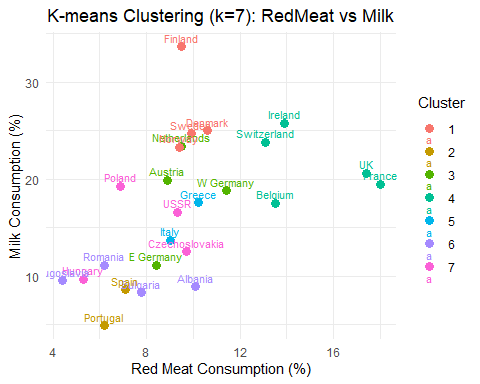
ggplot(protein\_data, aes(x = RedMeat, y = WhiteMeat, color = Cluster)) +  
 geom\_point(size = 3) +  
 geom\_text(aes(label = rownames(protein\_data)), vjust = -0.5, size = 3) +  
 labs(title = "K-means Clustering (k=7): RedMeat vs WhiteMeat",  
 x = "Red Meat Consumption (%)",  
 y = "White Meat Consumption (%)") +  
 theme\_minimal() +  
 theme(plot.title = element\_text(hjust = 0.5))



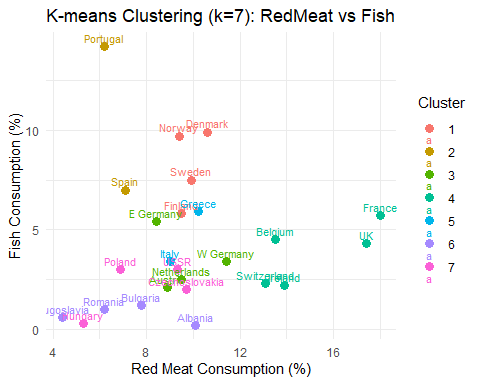
ggplot(protein\_data, aes(x = RedMeat, y = Eggs, color = Cluster)) +  
 geom\_point(size = 3) +  
 geom\_text(aes(label = rownames(protein\_data)), vjust = -0.5, size = 3) +  
 labs(title = "K-means Clustering (k=7): RedMeat vs Eggs",  
 x = "Red Meat Consumption (%)",  
 y = "Eggs Consumption (%)") +  
 theme\_minimal() +  
 theme(plot.title = element\_text(hjust = 0.5))



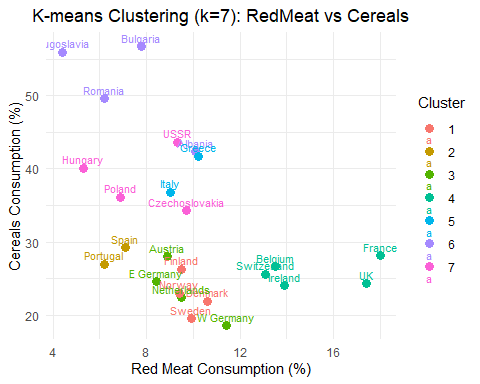
ggplot(protein\_data, aes(x = RedMeat, y = Milk, color = Cluster)) +  
 geom\_point(size = 3) +  
 geom\_text(aes(label = rownames(protein\_data)), vjust = -0.5, size = 3) +  
 labs(title = "K-means Clustering (k=7): RedMeat vs Milk",  
 x = "Red Meat Consumption (%)",  
 y = "Milk Consumption (%)") +  
 theme\_minimal() +  
 theme(plot.title = element\_text(hjust = 0.5))



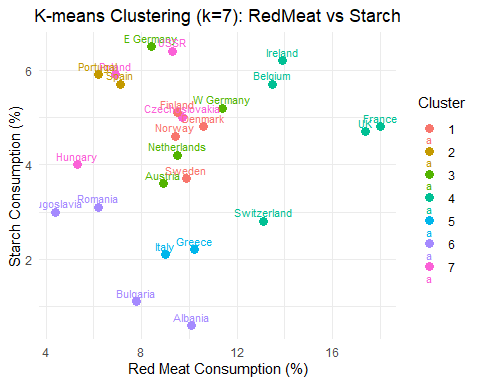
ggplot(protein\_data, aes(x = RedMeat, y = Fish, color = Cluster)) +  
 geom\_point(size = 3) +  
 geom\_text(aes(label = rownames(protein\_data)), vjust = -0.5, size = 3) +  
 labs(title = "K-means Clustering (k=7): RedMeat vs Fish",  
 x = "Red Meat Consumption (%)",  
 y = "Fish Consumption (%)") +  
 theme\_minimal() +  
 theme(plot.title = element\_text(hjust = 0.5))



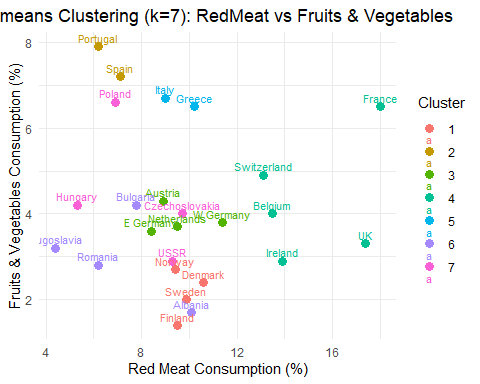
ggplot(protein\_data, aes(x = RedMeat, y = Cereals, color = Cluster)) +  
 geom\_point(size = 3) +  
 geom\_text(aes(label = rownames(protein\_data)), vjust = -0.5, size = 3) +  
 labs(title = "K-means Clustering (k=7): RedMeat vs Cereals",  
 x = "Red Meat Consumption (%)",  
 y = "Cereals Consumption (%)") +  
 theme\_minimal() +  
 theme(plot.title = element\_text(hjust = 0.5))



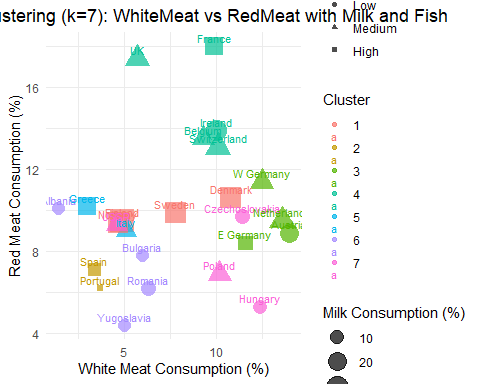
ggplot(protein\_data, aes(x = RedMeat, y = Starch, color = Cluster)) +  
 geom\_point(size = 3) +  
 geom\_text(aes(label = rownames(protein\_data)), vjust = -0.5, size = 3) +  
 labs(title = "K-means Clustering (k=7): RedMeat vs Starch",  
 x = "Red Meat Consumption (%)",  
 y = "Starch Consumption (%)") +  
 theme\_minimal() +  
 theme(plot.title = element\_text(hjust = 0.5))



ggplot(protein\_data, aes(x = RedMeat, y = Fr.Veg, color = Cluster)) +  
 geom\_point(size = 3) +  
 geom\_text(aes(label = rownames(protein\_data)), vjust = -0.5, size = 3) +  
 labs(title = "K-means Clustering (k=7): RedMeat vs Fruits & Vegetables",  
 x = "Red Meat Consumption (%)",  
 y = "Fruits & Vegetables Consumption (%)") +  
 theme\_minimal() +  
 theme(plot.title = element\_text(hjust = 0.5))



# Categorize Fish consumption into 3 levels (low, medium, high) for point shapes  
protein\_data$Fish\_Category <- cut(protein\_data$Fish,   
 breaks = quantile(protein\_data$Fish, probs = c(0, 0.33, 0.66, 1)),  
 labels = c("Low", "Medium", "High"),  
 include.lowest = TRUE)  
  
# Create complex scatter plot  
ggplot(protein\_data, aes(x = WhiteMeat, y = RedMeat, color = Cluster, size = Milk, shape = Fish\_Category)) +  
 geom\_point(alpha = 0.7) +  
 geom\_text(aes(label = rownames(protein\_data)), vjust = -0.5, size = 3) +  
 scale\_size\_continuous(range = c(2, 8), name = "Milk Consumption (%)") +  
 scale\_shape\_manual(values = c(16, 17, 15), name = "Fish Consumption") +  
 labs(title = "K-means Clustering (k=7): WhiteMeat vs RedMeat with Milk and Fish",  
 x = "White Meat Consumption (%)",  
 y = "Red Meat Consumption (%)") +  
 theme\_minimal() +  
 theme(plot.title = element\_text(hjust = 0.5),  
 legend.position = "right")



# Add country names to the dataset  
protein\_data$Country <- rownames(protein\_data)  
  
# Display sorted data by cluster  
View(protein\_data[order(protein\_data$Cluster), c("Country", "Cluster", expected\_columns)])  
  
# Summarize cluster centroids  
cat("Cluster centroids (scaled):\n")

## Cluster centroids (scaled):

print(k7$centers)

## RedMeat WhiteMeat Eggs Milk Fish Cereals  
## 1 0.006572897 -0.2290150 0.19147892 1.3458748 1.1582546 -0.8722721  
## 2 -0.949484801 -1.1764767 -0.74802044 -1.4583242 1.8562639 -0.3779572  
## 3 -0.083057512 1.3613671 0.88491892 0.1671964 -0.2745013 -0.8062116  
## 4 1.599006499 0.2988565 0.93413079 0.6091128 -0.1422470 -0.5948180  
## 5 -0.068119111 -1.0411250 -0.07694947 -0.2057585 0.1075669 0.6380079  
## 6 -0.807569986 -0.8719354 -1.55330561 -1.0783324 -1.0386379 1.7200335  
## 7 -0.605901566 0.4748136 -0.27827076 -0.3640885 -0.6492221 0.5719474  
## Starch Nuts Fr.Veg  
## 1 0.1676780 -0.95533923 -1.1148048  
## 2 0.9326321 1.12203258 1.8925628  
## 3 0.3665660 -0.86720831 -0.1585451  
## 4 0.3451473 -0.34849486 0.1020010  
## 5 -1.3010340 1.49973655 1.3659270  
## 6 -1.4234267 0.99613126 -0.6436044  
## 7 0.6419495 -0.04884971 0.1602082

**Conclusion**

Through K-means clustering, we identified meaningful dietary clusters among European countries.

* The **3-cluster solution** based on *RedMeat* and *WhiteMeat* highlighted three distinct consumption behaviors, with Western Europe leaning toward red meat, Southern and Eastern countries showing moderate intake, and Central/Northern Europe displaying a balance.
* The **7-cluster solution** using all variables revealed more nuanced groupings, capturing differences in fish, milk, cereal, and fruit/vegetable consumption.
* Advanced scatter plots provided layered insights, such as the relationship between milk and fish intake relative to meat consumption.

This analysis demonstrates how unsupervised learning techniques like K-means can segment populations based on real-world behaviors, offering valuable insights for public health, food policy, and market segmentation.