K-Means Clustering

Amran A

#Introduction #The objective of this project in R is to perform clustering analysis on protein consumption data for different European countries. The visualization and clustering analysis in this project provide valuable insights into European protein consumption patterns across different geographic regions and protein sources. By clustering the data based on similarities, distinct clusters are represented by different colors. One plot focuses on red and white meat consumption, revealing similarities in consumption patterns among European countries. Overall, these findings contribute to our understanding of European protein consumption by highlighting similarities and differences among regions and protein sources.

#Importing datasets

```
## Set CRAN mirror
options(repos = "https://cran.rstudio.com")

install.packages(c("cluster", "factoextra"))
```

```
## Warning in readRDS(dest): lzma decoder corrupt data
```

```
##
## The downloaded binary packages are in
##  /var/folders/4n/tlpw44pd0pd33dmw7845d09m0000gn/T//RtmpYW8XUy/downloaded_packages
```

```
#loading packages
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

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

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

```r
# Importing the data sets
protein <- read.csv("protein.csv", header=TRUE, row.names=1)
head(protein, 10)
```

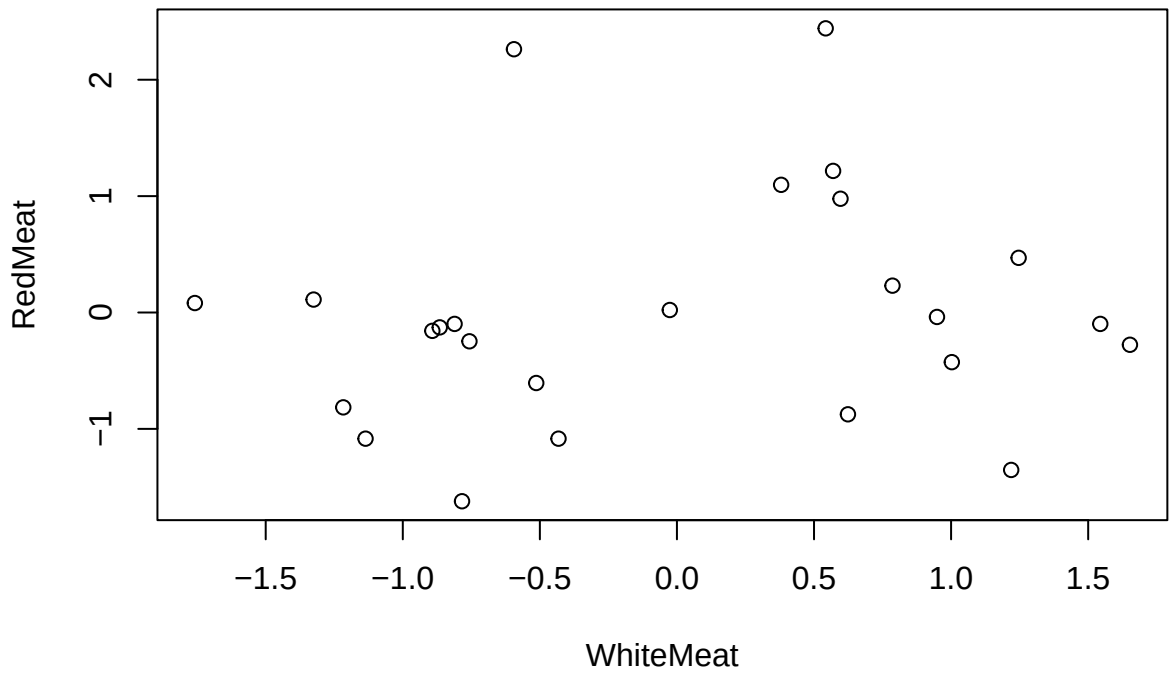
```
##                RedMeat WhiteMeat Eggs Milk Fish Cereals Starch Nuts Fr.Veg
## Albania           10.1       1.4  0.5  8.9  0.2    42.3    0.6  5.5    1.7
## Austria            8.9      14.0  4.3 19.9  2.1    28.0    3.6  1.3    4.3
## Belgium           13.5       9.3  4.1 17.5  4.5    26.6    5.7  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
## E Germany          8.4      11.6  3.7 11.1  5.4    24.6    6.5  0.8    3.6
## Finland            9.5       4.9  2.7 33.7  5.8    26.3    5.1  1.0    1.4
## France            18.0       9.9  3.3 19.5  5.7    28.1    4.8  2.4    6.5
## Greece            10.2       3.0  2.8 17.6  5.9    41.7    2.2  7.8    6.5
```

```r
# using the scale function to center the variables and make them comparable
protein_scaled <- scale(protein, center=TRUE, scale=TRUE)
```

```r
#Clustering red and white meat: The variables "RedMeat" and "WhiteMeat" are extracted from the scaled d
red_white = protein_scaled[,c("WhiteMeat","RedMeat")]
head(red_white)
```

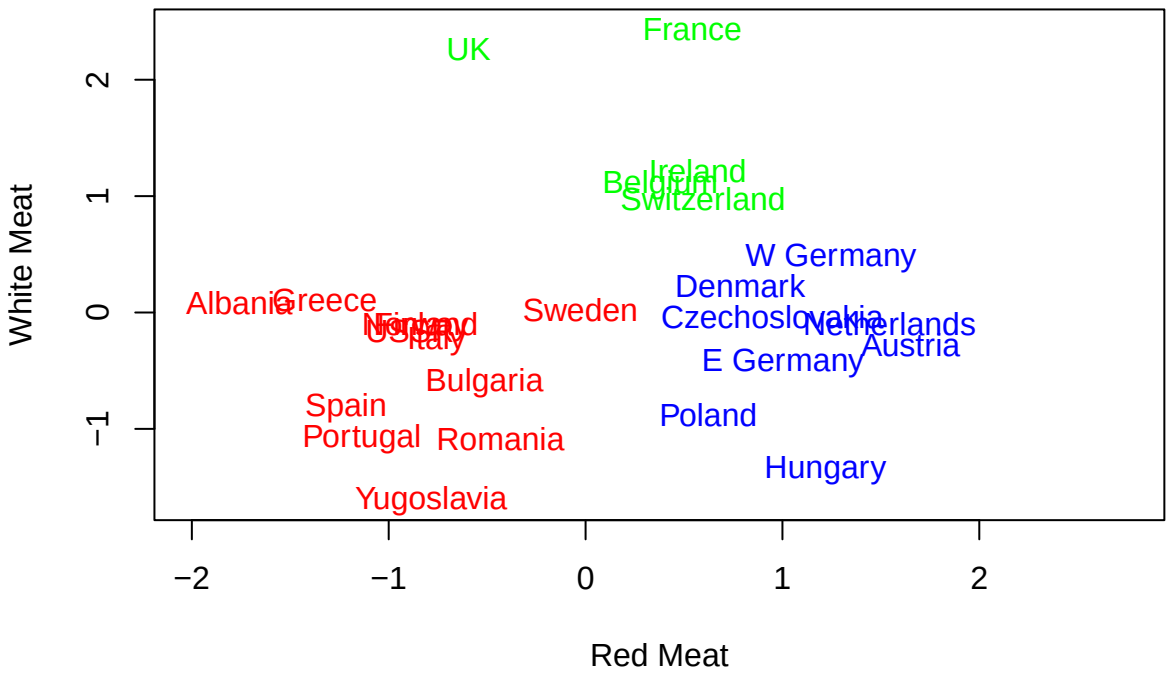
```
##                 WhiteMeat     RedMeat
## Albania        -1.7584889  0.08126490
## Austria         1.6523731 -0.27725673
## Belgium         0.3800675  1.09707621
## Bulgaria       -0.5132535 -0.60590157
## Czechoslovakia  0.9485445 -0.03824231
## Denmark         0.7861225  0.23064892
```

```r
plot(red_white)
```




```
# Used k-means to make 3 clusters
cluster_redwhite <- kmeans(red_white, centers=3)

# Plotting with the labels
plot(red_white, xlim=c(-2,2.75),
     type="n", xlab="Red Meat", ylab="White Meat")
text(red_white, labels=rownames(red_white),
     col=rainbow(3)[cluster_redwhite$cluster])
```




```
#Clustering all nine protein groups: The k-means clustering algorithm is applied to all nine protein gr
cluster_all <- kmeans(protein_scaled, centers=7, nstart=30)
names(cluster_all)
```

```
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
# Extract some of the information from the fitted model
cluster_all$centers
```

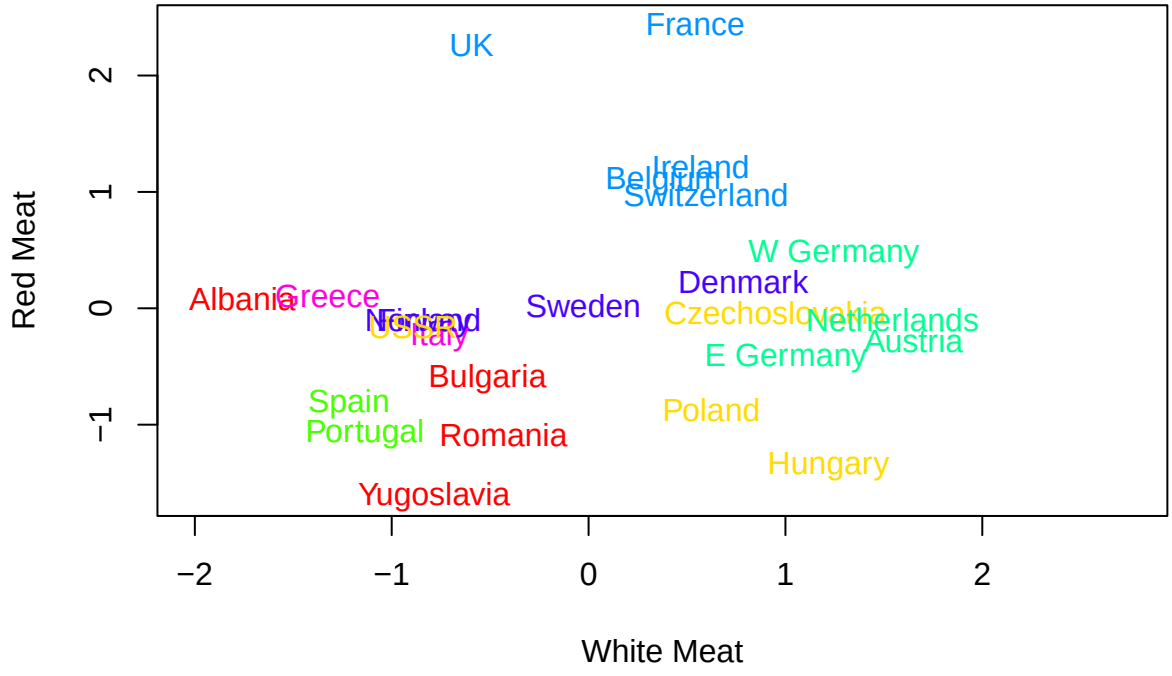
```
##        RedMeat  WhiteMeat        Eggs       Milk       Fish    Cereals
## 1 -0.807569986 -0.8719354 -1.55330561 -1.0783324 -1.0386379  1.7200335
## 2 -0.605901566  0.4748136 -0.27827076 -0.3640885 -0.6492221  0.5719474
## 3 -0.949484801 -1.1764767 -0.74802044 -1.4583242  1.8562639 -0.3779572
##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

```
## 7 -1.3010340  1.49973655  1.3659270
```

```
cluster_all$cluster
```

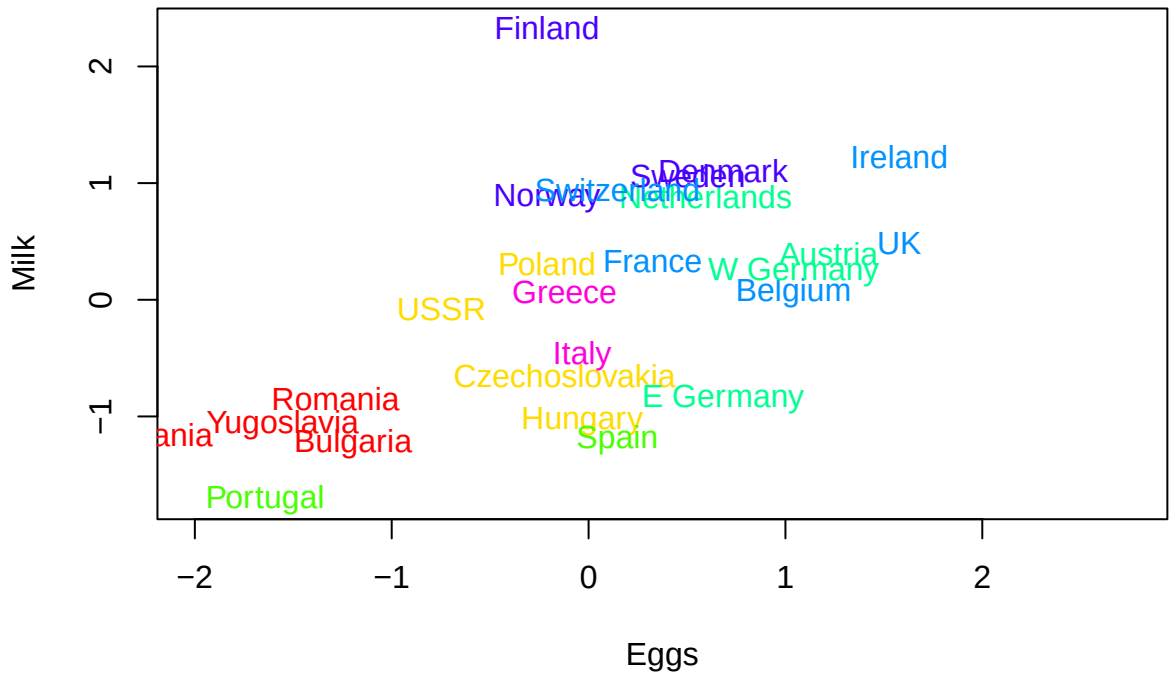
```
##        Albania        Austria        Belgium       Bulgaria Czechoslovakia
##              1              4              5              1              2
##        Denmark      E Germany        Finland         France         Greece
##              6              4              6              5              7
##        Hungary        Ireland          Italy    Netherlands         Norway
##              2              5              7              4              6
##         Poland       Portugal        Romania          Spain         Sweden
##              2              3              1              3              6
##    Switzerland             UK           USSR      W Germany     Yugoslavia
##              5              5              2              4              1
```

```
# Resulting clusters (coloured) are on a scatter plot of white meat against red meat.
  # Plotting the clustering on the red-white meat axes
plot(protein_scaled[,"WhiteMeat"], protein_scaled[,"RedMeat"], xlim=c(-2,2.75),
     type="n", ylab="Red Meat", xlab="White Meat")
text(protein_scaled[,"WhiteMeat"], protein_scaled[,"RedMeat"], labels=rownames(protein),
     col=rainbow(7)[cluster_all$cluster])
```

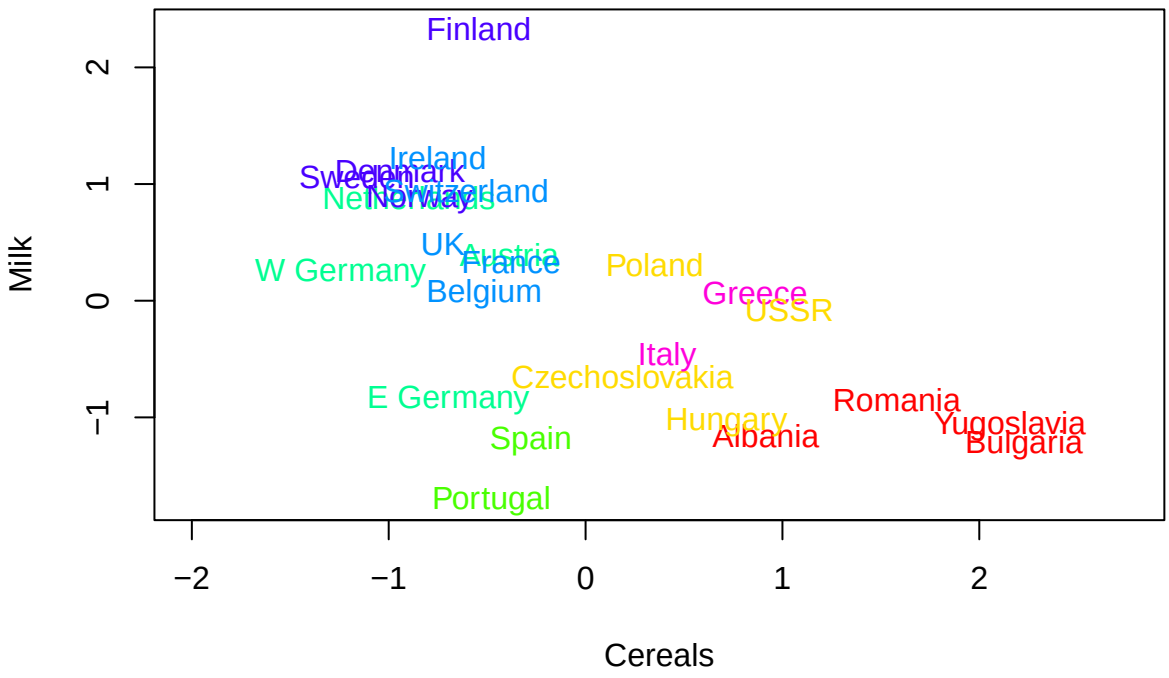



```
# Plotting milk vs Eggs variables
plot(protein_scaled[,"Eggs"], protein_scaled[,"Milk"], xlim=c(-2,2.75),
```

```
     type="n", xlab="Eggs", ylab="Milk")
text(protein_scaled[,"Eggs"], protein_scaled[,"Milk"], labels=rownames(protein),
         col=rainbow(7)[cluster_all$cluster])
```



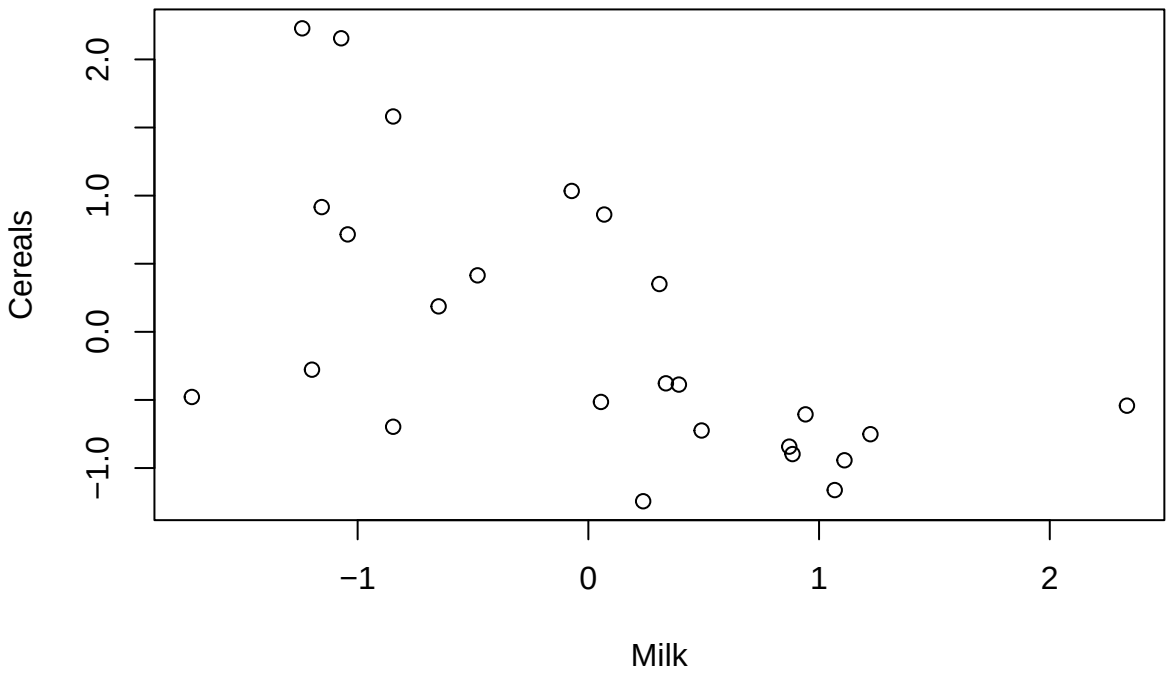

```
# Plotting Milk vs Cereals variables
plot(protein_scaled[,"Cereals"], protein_scaled[,"Milk"], xlim=c(-2,2.75),
     type="n", xlab="Cereals", ylab="Milk")
text(protein_scaled[,"Cereals"], protein_scaled[,"Milk"], labels=rownames(protein),
     col=rainbow(7)[cluster_all$cluster])
```




```
#clustering  Milk and Cereal  (p=2) and using 3 clusters (k=3)
  # Milk and Cereal variables
milk_cereals = protein_scaled[,c("Milk","Cereals")]
head(milk_cereals)
```

```
##                        Milk    Cereals
## Albania          -1.15573814  0.9159176
## Austria           0.39237676 -0.3870690
## Belgium           0.05460623 -0.5146342
## Bulgaria         -1.24018077  2.2280161
## Czechoslovakia -0.64908235  0.1869740
## Denmark           1.11013912 -0.9428885
```

```
plot(milk_cereals)
```




```
# Used k-means to make 3 clusters
cluster_milkcereals <- kmeans(milk_cereals, centers=3)
```

Results analysis

#The visualization results for all of these show European protein consumption for different sources of proteins in different geographic regions. The data is clustered based on similarities, and the resulting clusters are represented by different colors. In my results, one of the plots demonstrates the consumption of red and white meat in different European countries. The countries are clustered together based on their protein (red and white meat) consumption values, revealing similarities in consumption patterns among them, for example, countries like Germany, Austria, and Ireland have a high consumption of Red Meat, and countries such as the United Kingdom and France have a high consumption of White Meat. Another one of the plot's examines the consumption of milk and eggs in European countries, which is clustered based on similarities in the resulting values for consumption, for example, countries such as portugal and Romania have a low consumption of both Milk and Eggs, whereas countries such as the United Kingdom and Ireland have a high consumption of both milk and Eggs. Additionally, I included a plot that visualizes milk vs cereal protein consumption in Europe, clustering the regions based on their similarities in consumption amounts, for this plot we have countries such as Finland with a high consumption of milk but a low consumption of cereal, whereas we also have countries such as Bulgaria with a high consumption of cereal but a low consumption of milk.

#In summary, the visualization and clustering analysis provide insights into European protein consumption patterns for various protein sources. The clusters allow us to identify groups of countries with similar consumption behaviors, contributing to our understanding of protein consumption across different regions.

#PDF knitting with package install.packages("tinytex") tinytex::install_tinytex()

“‘