Clustering on European Protein Consumption

Athira

2025-05-25

Contents

1 Importing Libraries and load input files or European protein consumption dataset 1

2 PROGRAM FOR CLUSTERING REDMEAT and WHITEMEAT USING Kmean ALGORITHM AND PLOT USING DIFFERENT GRAPHS 2

2.1 Plot RedMeat and Whitemeat Using ggplot with datapoints as 25 Countries with 3 clusters . 3

2.2 Ploting same graph after normalizing using scale() option 4

2.3 Ploting the same graph using Fviz_cluster option 5

2.4 Results inferred from above graph and table 6

3 PROGRAM FOR CLUSTERING 9 PROTEIN INTO 7 CLUSTERS USING KMeans 6

3.1 Generate a table with 25 countries and kmean clusters of 9 proteins 7

3.2 Plot the graph with Eggs in X-axis and Milk in Y-axis along with kmeans values generated from 9 protein data. 7

3.3 Inference from above Plot 8

3.4 ploting the graph with Fish on X-axis and Cereals on Y-axis and also influenced by all 9 protein consumption 8

3.5 Inference from above Plot 9

3.6 Ploting the graph with Starch on X-axis and Nuts on Y-axis and also influenced by all 9 protein consumption 10

3.7 Inference From Plot 10

3.8 Graph showing 9 protein consumption based on dimention 11

1 Importing Libraries and load input files or European protein consumption dataset

```
library(cluster)
```

```
## Warning: package 'cluster' was built under R version 4.4.3
```

```r
  library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.4.3

## Warning: package 'ggplot2' was built under R version 4.4.3

## Warning: package 'tibble' was built under R version 4.4.3

## Warning: package 'tidyr' was built under R version 4.4.3

## Warning: package 'readr' was built under R version 4.4.3

## Warning: package 'purrr' was built under R version 4.4.3

## Warning: package 'dplyr' was built under R version 4.4.3

## Warning: package 'stringr' was built under R version 4.4.3

## Warning: package 'forcats' was built under R version 4.4.3

## Warning: package 'lubridate' was built under R version 4.4.3

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- 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
```

```r
  library(dplyr)
  library(ggplot2)
  library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.4.3

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

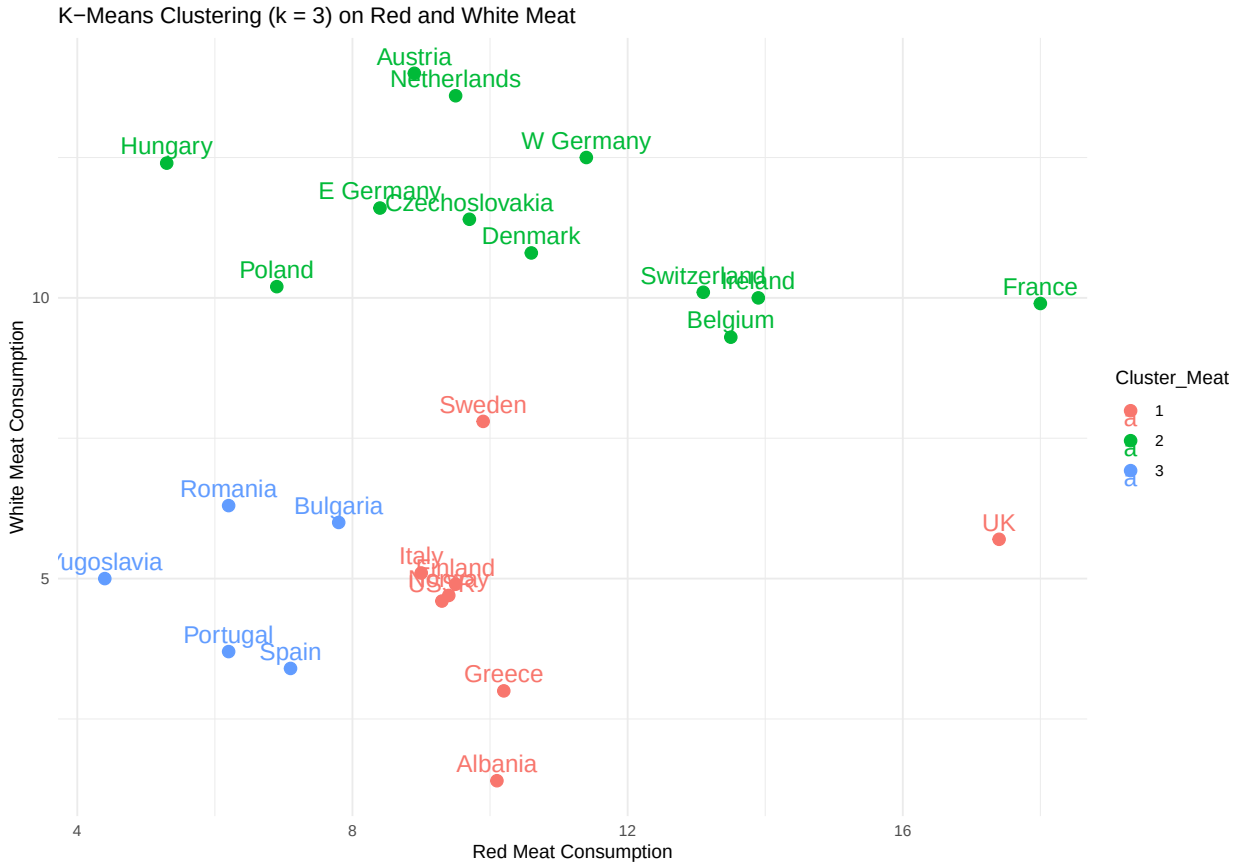
```r
#Load CSV data set and making a copy
  protein_1<-read.csv("protein.csv",fileEncoding = "latin1")
  protein_2<-protein_1
```

2 PROGRAM FOR CLUSTERING REDMEAT and WHITEMEAT USING Kmean ALGORITHM AND PLOT USING DIFFERENT GRAPHS

```r
#Select an object with the required fields for Kmeans
  red_white_meat<-protein_2 %>% select(RedMeat,WhiteMeat)
#For labeling get the country details in other variable
  Country<-protein_2$Country
#Calculate Kmean for the red_white_meat object(Redmeat and whitemeat) with 3 clusters and 25 iterations
  kmean.result <- kmeans(red_white_meat,3,25)
#Adding labels to cluster after making into factor datatype
  protein_2$Cluster_Meat <- as.factor(kmean.result$cluster)
```

2.1 Plot RedMeat and Whitemeat Using ggplot with datapoints as 25 Countries with 3 clusters

```r
ggplot(protein_2, aes(x = RedMeat, y = WhiteMeat,
                      color = Cluster_Meat, label = Country)) +
                      geom_point(size = 3) +geom_text(vjust = -0.5, size = 5) +
                      labs(title = "K-Means Clustering (k = 3) on Red and White Meat",
                     x = "Red Meat Consumption", y = "White Meat Consumption") +
                     theme_minimal()
```

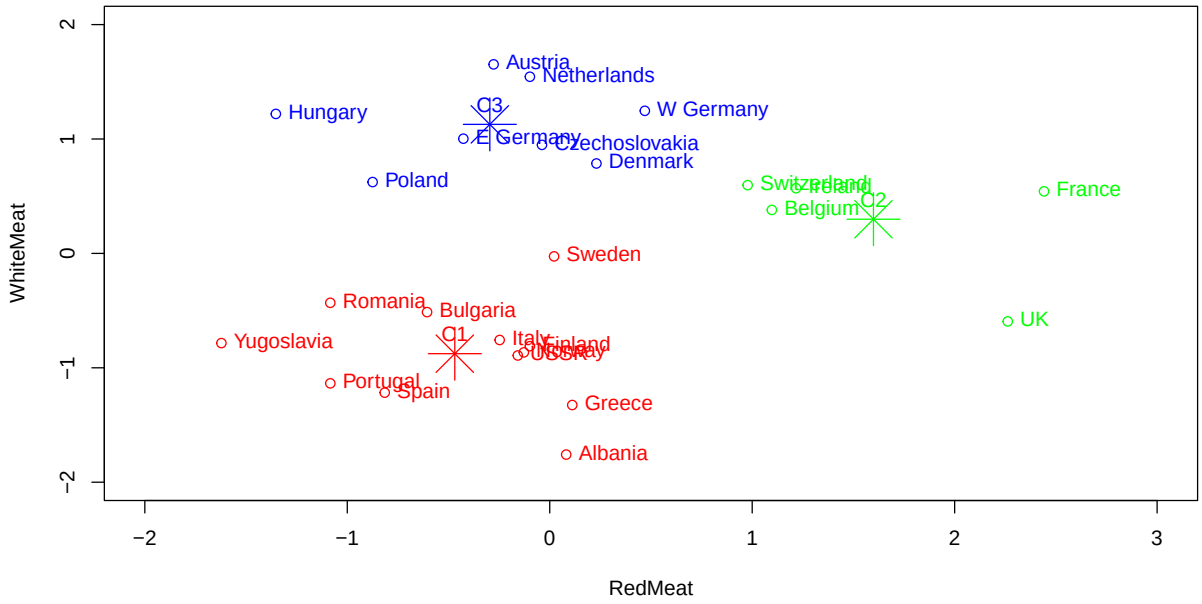


2.2 Ploting same graph after normalizing using scale() option

```
#scaling the whitemeat and red meat protein values for getting normalized values
  cluster_colors<-c("Red","green","blue")
  red_white_meat_scaled<-red_white_meat %>%scale()
#Generating Kmeans with normalized values
  kmean.result <- kmeans(red_white_meat_scaled,3,25)
#plot a table to find similarity between countries and their redmeat and whitemeat consuption
  table(protein_2$Country,kmean.result$cluster)
```

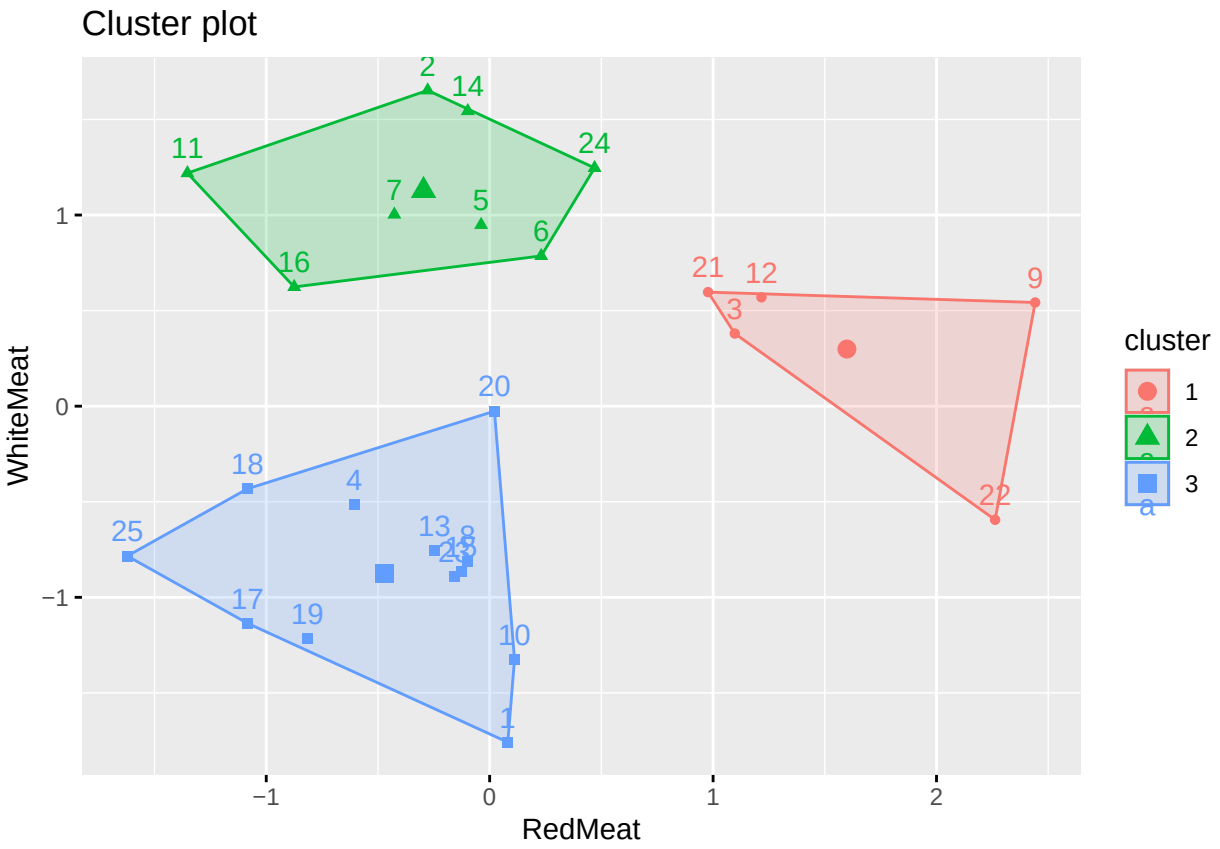
```
##
##                 1 2 3
##   Albania       1 0 0
##   Austria       0 0 1
##   Belgium       0 1 0
##   Bulgaria      1 0 0
##   Czechoslovakia 0 0 1
##   Denmark       0 0 1
##   E Germany     0 0 1
##   Finland       1 0 0
##   France        0 1 0
##   Greece        1 0 0
##   Hungary       0 0 1
##   Ireland       0 1 0
##   Italy         1 0 0
##   Netherlands   0 0 1
##   Norway        1 0 0
##   Poland        0 0 1
##   Portugal      1 0 0
##   Romania       1 0 0
##   Spain         1 0 0
##   Sweden        1 0 0
##   Switzerland   0 1 0
##   UK            0 1 0
##   USSR          1 0 0
##   W Germany     0 0 1
##   Yugoslavia    1 0 0
```

```
#Plot the graph based on clusters, centroids and counties with RedMeat in x-axis and WhiteMeat in y-axi
  plot(red_white_meat_scaled[,"RedMeat"],red_white_meat_scaled[,"WhiteMeat"],
       col=cluster_colors[kmean.result$cluster],xlim=c(-2,3),ylim = c(-2,2),
       xlab ="RedMeat",ylab = "WhiteMeat" )
  points(kmean.result$centers[,c("RedMeat", "WhiteMeat")],
         pch = 8, cex=4,col=cluster_colors[1:3])
  text(kmean.result$centers, labels = paste("C", 1:3, sep=""),
       pch=8, pos = 3, cex = 1,col = cluster_colors[1:3])
  text(red_white_meat_scaled[,"RedMeat"], red_white_meat_scaled[,"WhiteMeat"],
       labels = Country,col = cluster_colors[kmean.result$cluster],pos = 4,cex = 1)
```



2.3 Ploting the same graph using Fviz_cluster option

```
#Generate the same plot with fviz_cluster, which shows boundaries for each cluster and centriods
  fviz_cluster(kmeans(red_white_meat,centers = 3,100),data=red_white_meat)
```



2.4 Results inferred from above graph and table

- RedMeat is plotted in X-axis and WhiteMeat in Y-axis
- 25 countries are plotted into 3 clusters
- C1 cluster[Sweden,Romania,Bulgaria,Yugoslavia,Italy,USSR,Norway,Finland, Greece,Albania, Portugal,Spain], having low WhiteMeat and RedMeat consumption.
- C2 cluster[Switzerland, France, Belgarium,UK, Finland] having high redMeat consumption and medium WhiteMeat consuption.
- c3 cluster[Austria,Netherlands,W Germany,Hungary,EGermany,Czechoslovakia,Denmark,Poland] low RedMeat and High WhiteMeat Consumption.

3 PROGRAM FOR CLUSTERING 9 PROTEIN INTO 7 CLUSTERS USING KMeans

Normalize dataset contains 9 protein and clustured using Kmeans algorithm with 7 clusters

```
#Removing the Country names from data for analystical purpose
  protein_1$Country<-NULL
#Scaling or Normalizing the data
  protein_1_scaled<-protein_1 %>% scale()
#For the pupose of ploting generate an object with different colors
  cluster_colors<-c("Red","blue","green","brown","violet","black","purple")
```

```
#Apply kmeans algorithm on above scaled data
  kmean.result1<-kmeans(protein_1_scaled,7,100)
```

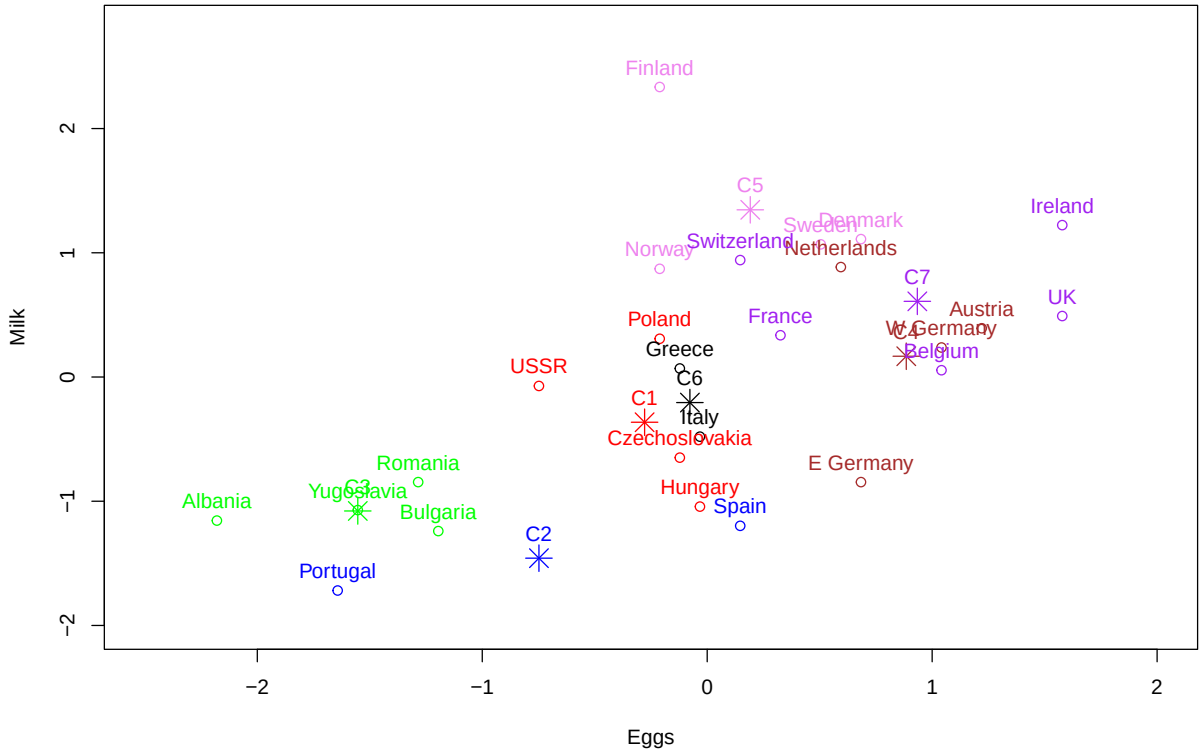
3.1 Generate a table with 25 countries and kmean clusters of 9 proteins

```
#Plot a table to find relationship between European Countries and protein consumption
  table(protein_2$Country,kmean.result1$cluster)
```

```
##
##                 1 2 3 4 5 6 7
##   Albania       0 0 1 0 0 0 0
##   Austria       0 0 0 1 0 0 0
##   Belgium       0 0 0 0 0 0 1
##   Bulgaria      0 0 1 0 0 0 0
##   Czechoslovakia 1 0 0 0 0 0 0
##   Denmark       0 0 0 0 1 0 0
##   E Germany     0 0 0 1 0 0 0
##   Finland       0 0 0 0 1 0 0
##   France        0 0 0 0 0 0 1
##   Greece        0 0 0 0 0 1 0
##   Hungary       1 0 0 0 0 0 0
##   Ireland       0 0 0 0 0 0 1
##   Italy         0 0 0 0 0 1 0
##   Netherlands   0 0 0 1 0 0 0
##   Norway        0 0 0 0 1 0 0
##   Poland        1 0 0 0 0 0 0
##   Portugal      0 1 0 0 0 0 0
##   Romania       0 0 1 0 0 0 0
##   Spain         0 1 0 0 0 0 0
##   Sweden        0 0 0 0 1 0 0
##   Switzerland   0 0 0 0 0 0 1
##   UK            0 0 0 0 0 0 1
##   USSR          1 0 0 0 0 0 0
##   W Germany     0 0 0 1 0 0 0
##   Yugoslavia    0 0 1 0 0 0 0
```

3.2 Plot the graph with Eggs in X-axis and Milk in Y-axis along with kmeans values generated from 9 protein data.

```
#piloting the graph with Eggs on X-axis and Milk on Y-axis and also influenced by all 9 protein consump
#Created 7 clusters with countries as datapoints and each cluster having  centroids
  plot(protein_1_scaled[,"Eggs"],protein_1_scaled[,"Milk"],col=cluster_colors[kmean.result1$cluster],xl
  points(kmean.result1$centers[,c("Eggs", "Milk")], col = cluster_colors[1:7], pch = 8, cex=2,pos=3)
  text(kmean.result1$centers[,"Eggs"],kmean.result1$centers[,"Milk"]+.2,labels = paste("C", 1:7, sep=""
  text(protein_1_scaled[,"Eggs"], protein_1_scaled[,"Milk"],
     labels = Country,
     pos = 3, cex = 1, col = cluster_colors[kmean.result1$cluster])
```



3.3 Inference from above Plot

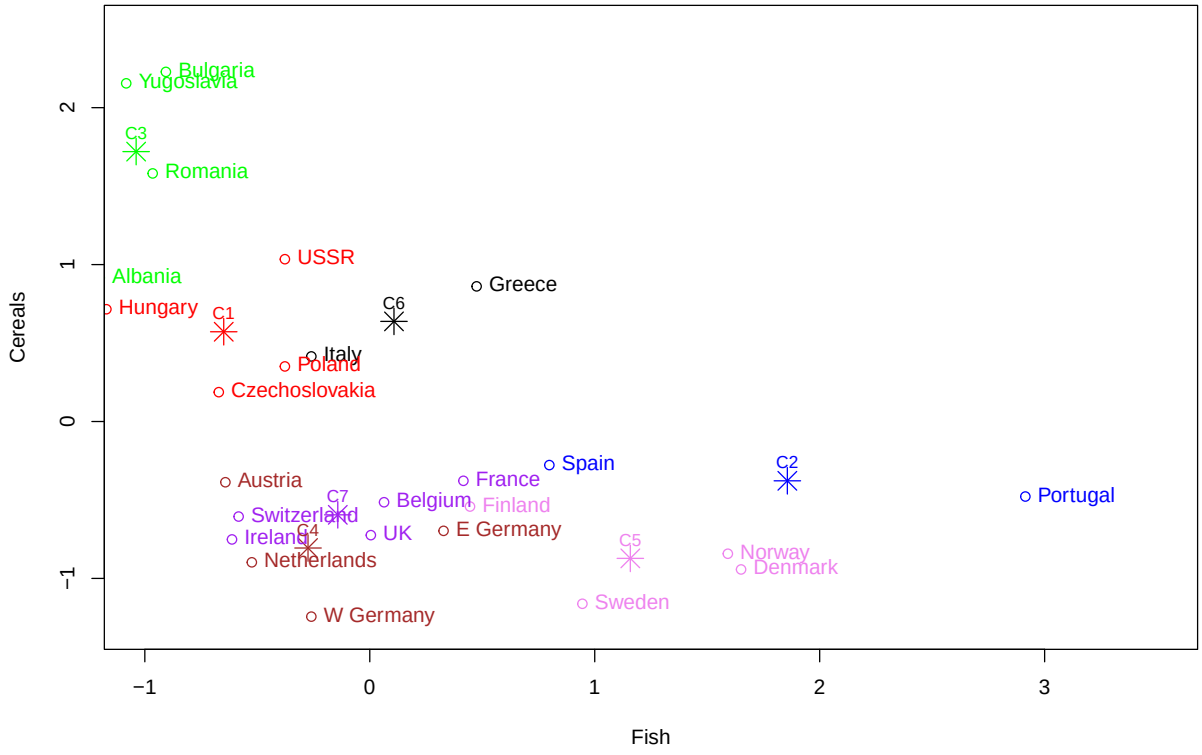
- plot Eggs and Milk in X and y-axis.Also have the effect of 9 protein
- for cluster C1[Portugal, Spain], having similar egg and mik consumption.
- Cluster C2[Albania,Yugoslavia,Romania, Bulgaria,Hungary,USSR] having very low egg and milk consumption
- Cluster c3[Czechoslovakia,Poland,E Germany,W Germany,Austria,Netherland,Poland] milk and Egg medium consumption
- Cluster C4[Finland,Denmark, Sweden,Norway] having high milk consumption and medium egg consumption
- Cluster C5[Ireland,Belgium, UK] having high milk and egg consumption
- Cluster C6[Greece, Italy] having low egg and milk consumption
- Cluster C7[France, Switzerland] have more than medium egg and milk consumption

3.4 ploting the graph with Fish on X-axis and Cereals on Y-axis and also influenced by all 9 protein consumption

- Created 7 clusters with Countries as data points and each cluster having centroids.

```
plot(protein_1_scaled[,"Fish"],protein_1_scaled[,"Cereals"],
     col=cluster_colors[kmean.result1$cluster],xlim=c(-1,3.5),
     ylim = c(-1.3,2.5),xlab = "Fish",ylab = "Cereals")
  points(kmean.result1$centers[,c("Fish", "Cereals")], col = cluster_colors[1:7],
```

```
        pch = 8, cex=2)
text(protein_1_scaled[,"Fish"], protein_1_scaled[,"Cereals"],
   labels = Country,
   pos = 4, cex = 1, col = cluster_colors[kmean.result1$cluster])
text(kmean.result1$centers[,"Fish"],kmean.result1$centers[,"Cereals"],
     labels = paste("C", 1:7, sep=""),col = cluster_colors[1:7],
     pos = 3, cex = 0.8)
```



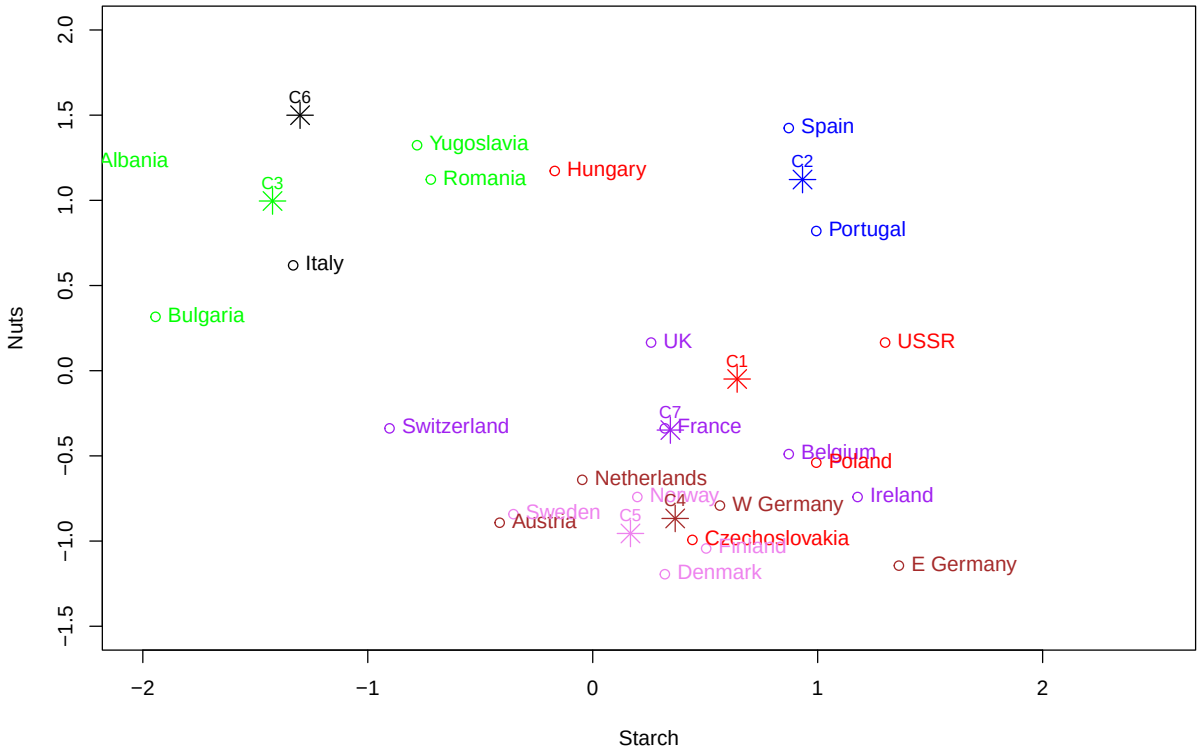
3.5 Inference from above Plot

- Cluster C1[Spain and Portugal] having similar Fish and Cereal consumpton
- Cluster C2[Bulgaria,Yugoslavia, Romania, Albania, Hungary,USSR] having low fish and high Cereals consumption
- Cluster C3[Poland,Czechoslovakia, Austria, E Germany,W Germany,Netherlands] having low Fish and Cereals consumption
- Cluster C4[Norway,Denmark, Sweden] having high Fish and low Cereals consumption.
- Cluster C5[Belgium,Ireland, UK] having similarity and have low fish and Cereals Consumption.
- Cluster C6[Greece,Italy] having medium fish and Cereal consumption.
- Cluster C7[Switzerland,France] having comparitively low Fish and Cereals Consumption

3.6 Ploting the graph with Starch on X-axis and Nuts on Y-axis and also influenced by all 9 protein consumption

- Created 7 clusters with countries as data points and each cluster having centroids

```
plot(protein_1_scaled[,"Starch"],protein_1_scaled[,"Nuts"],col=cluster_colors[kmean.result1$cluster], xl
  points(kmean.result1$centers[,c("Starch", "Nuts")], col = cluster_colors[1:7], pch = 8, cex=2)
  text(protein_1_scaled[,"Starch"],protein_1_scaled[,"Nuts"],
     labels = protein_2$Country,
     pos = 4, cex = 1, col = cluster_colors[kmean.result1$cluster])
  text(kmean.result1$centers[,"Starch"],kmean.result1$centers[,"Nuts"], labels = paste("C", 1:7, sep="")
```



3.7 Inference From Plot

- Cluster C1[Spain, Portugal] having more Starch and Nuts Consumption.
- Cluster C2[Albania,Bulgaria,Yugoslaviya,Romania, Hungary,USSR] having low Nuts consumption but starch consumption is scattered around
- Cluster C3[Poland,E Germany, W Germany,Netherland, Austria] having high starch and low Nuts consumption
- Cluster C4[Norway,Sweden, Finland, Denmark] having medium starch and low Nuts consumption
- Cluster C5[UK,Belgium, Ireland] having high starch and low Nuts consumption. Similar to C3 but other proteins are the effective factor
- Cluster C6[Italy] having medium high Nuts and low Starch consumption.
- Cluster C7[Switzerland, France] having medium low starch and Nuts consumption.

3.8 Graph showing 9 protein consumption based on dimention

```
#Plot a graph with 9 protein consumption, x axis and y axis plot is based on Dimensions
  fviz_cluster(kmean.result1,data = protein_1_scaled,geom="point")
```