BDS_KMeans_Protein

Ahmad Mudrik

2024-10-29

Introduction

The data set presents protein consumption data from 25 European countries, detailing the intake (in grams per person per day) of red meat, white meat, eggs, milk, fish, cereals, starch, nuts, and fruits and vegetables. This study aims to demonstrate the clustering procedure using the K-Means method, exploring varying values of 'K' and 'nstart.'

Hypothesis

1. The diets of people living in countries with similar regions or cultures tend to have comparable compositions.

Objective

1. To build a 3-Means clustering model.
2. To build a 5-Means clustering model.
3. To build the optimal cluster models b varying the 'K' and 'nstart' value.

Import, inspect, and standardize data set

Import and inspect data set

```
food <- read.csv("C:\\Users\\mudrik\\Documents\\Rstudio Projects\\BDS_KMeans_Protein\\protein.csv", row
head(food)
```

```
##                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
```

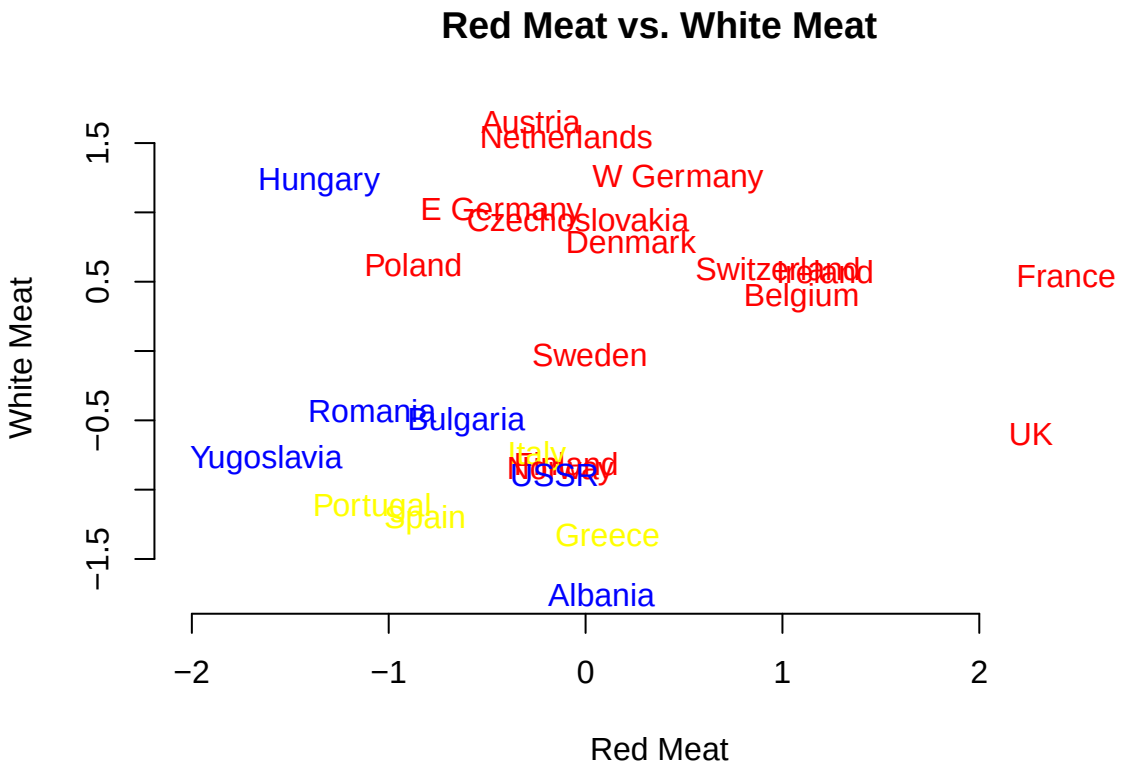
- To apply K-Means clustering in R, the data must be transformed into a numeric matrix, referred to as "x."
- This involves converting any factors, such as food$Country, into dummy variables.
- When importing the data, the row.names=1 function should be utilized to achieve this transformation.

Scale data for standardization

```
xfood <- scale(food)
```

- Units used in the data are converted from "in grams per person per day" to units of standard deviation (sd).
- This minimizes the squared errors across dimensions of x.

Study 1: 3-Means Model



Findings from Study 1

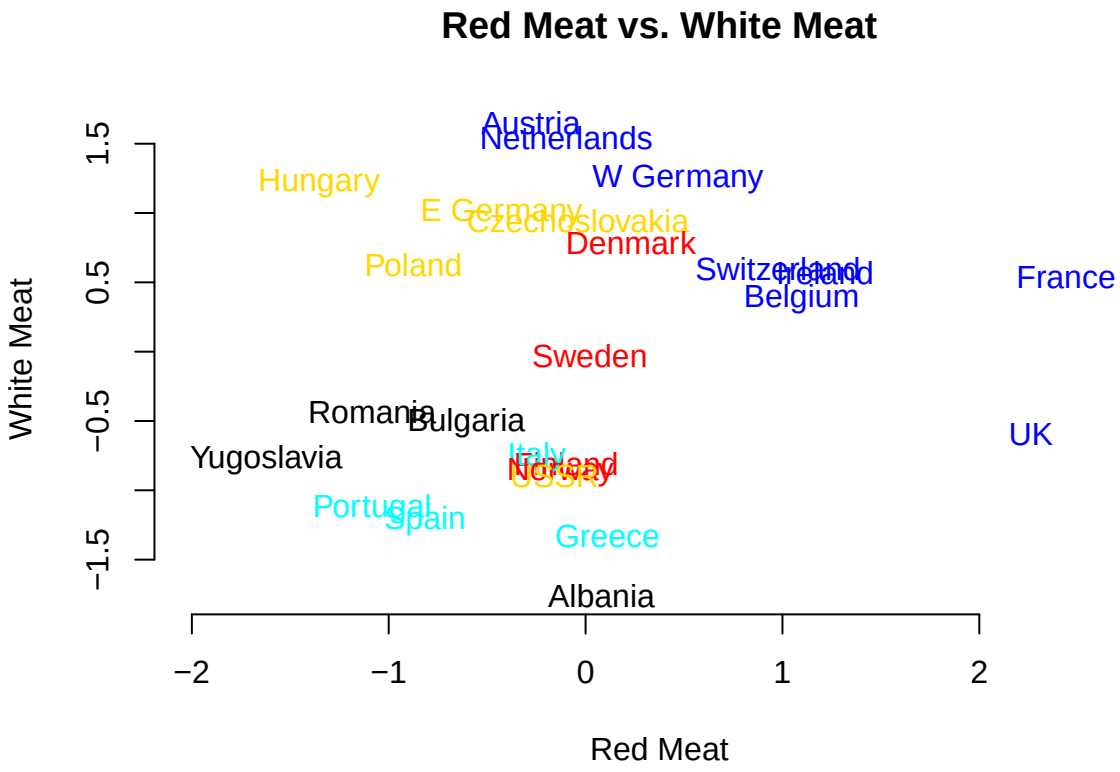
To fit 3-Mean model to the protein data.

- K-Means algorithm takes argument "center" to define K, and "nstart" to determine number of repeats of the algorithm (each corresponding to different random start).
- The minimum deviance found across "nstart" runs is reported to the user.
- The 3-Means model plots Red Meat vs. White Meat data, which showcases *3 clusters*

Study 2: 5-Means Model

```r
grpProtein <- kmeans(xfood, centers=5, nstart=10)

plot(xfood[,"RedMeat"],
     xfood[,"WhiteMeat"],
     xlim=c(-2,2.75),
     type="n",
     main = "Red Meat vs. White Meat",
     xlab="Red Meat",
     ylab="White Meat",
     bty="n")
text(xfood[,"RedMeat"],
     xfood[,"WhiteMeat"],
     labels=rownames(food),
     col=c("red", "gold", "blue", "cyan", "black")[grpProtein$cluster])
```



Findings from Study 2

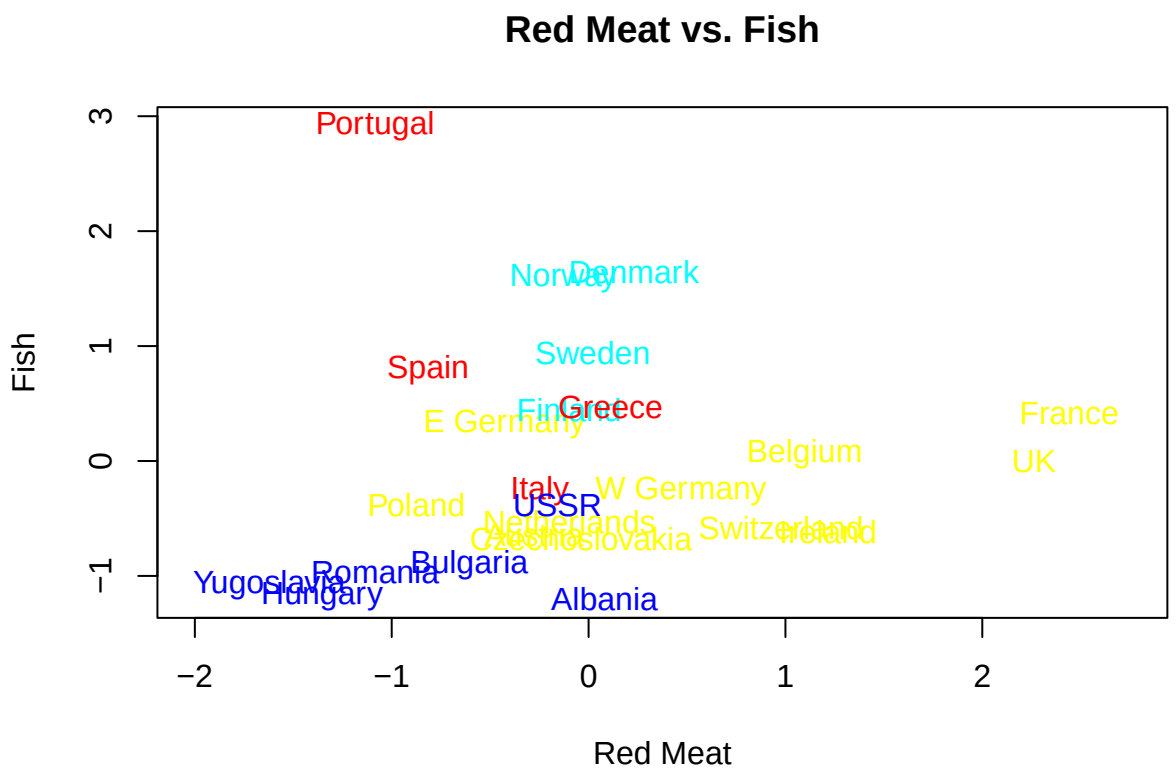
- The clustering analysis was performed on all nine proteins. However, data plotted only compared red and white meat consumption.
- The results highlighted several familiar cultural and/or geographic groupings, which were identified solely through patterns of protein consumption.
- For instance, *West Germany*'s protein consumption clusters with that of *Hungary, Poland, and Czechoslovakia*. Given their geographical proximity, it is reasonable to conclude that their dietary profiles are similar.

- In contrast, *East Germany* is grouped with countries in the Western European region, such as the *Netherlands and Belgium*. This distinction may explain why West and East Germany are classified differently, as each region's diet aligns more closely with that of its respective neighboring countries.

Study 3: 4-Means Model

```
grpProtein <- kmeans(xfood, centers=4, nstart=35)

plot(xfood[,"RedMeat"],
     xfood[,"Fish"],
     xlim=c(-2,2.75),
     type="n",
     main = "Red Meat vs. Fish",
     xlab="Red Meat",
     ylab="Fish")
text(xfood[,"RedMeat"],
     xfood[,"Fish"],
     labels=rownames(food),
     col=c("red", "yellow", "blue", "cyan", "black")[grpProtein$cluster])
```



Findings from Study 3

- The selection of the "k" value is highly subjective; algorithms that automatically determine "k" can be sensitive to assumptions regarding component probability models.

- To achieve relatively homogeneous groups, it is advisable to employ a trial-and-error approach when selecting the "k" value, ultimately choosing a "k" value that produces the most sensible clusters.

END OF DOCUMENT