Hierarchical Clustering

Example

```
# Question: Implement the hierarchical clustering algorithm using the Us arests dataset
# ---
#
# Loading the data set
# ---
#
data("USArrests")
```

Removing the null values

```
# Remove any missing value (i.e, NA values for not available)
# That might be present in the data
# ---
#
df <- na.omit(USArrests)
```

Previewing the dataset

```
# Previewing our dataset
# ---
#
head(df)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

Dataset descriptive statistics

```
# Before hierarchical clustering, we can compute some descriptive statistics
# ---
#
desc_stats <- data.frame(
 Min = apply(df, 2, min), # minimum
 Med = apply(df, 2, median), # median
 Mean = apply(df, 2, mean), # mean
 SD = apply(df, 2, sd), # Standard deviation
 Max = apply(df, 2, max) # Maximum
)
desc_stats <- round(desc_stats, 1)
head(desc_stats)
```

```
##           Min   Med  Mean   SD   Max
## Murder    0.8   7.2   7.8  4.4  17.4
```

```
## Assault  45.0 159.0 170.8 83.3 337.0
## UrbanPop 32.0  66.0  65.5 14.5  91.0
## Rape      7.3  20.1  21.2  9.4  46.0
```

```
# We note that the variables have a large different means and variances.
# This is explained by the fact that the variables are measured in different
# units; Murder, Rape, and Assault are measured as the number of occurrences per 100 000 people,
# and UrbanPop is the percentage of the state's population that lives in an urban area.
# They must be standardized (i.e., scaled) to make them comparable. Recall that,
# standardization consists of transforming the variables such that
# they have mean zero and standard deviation one.
```

Scaling the dataset

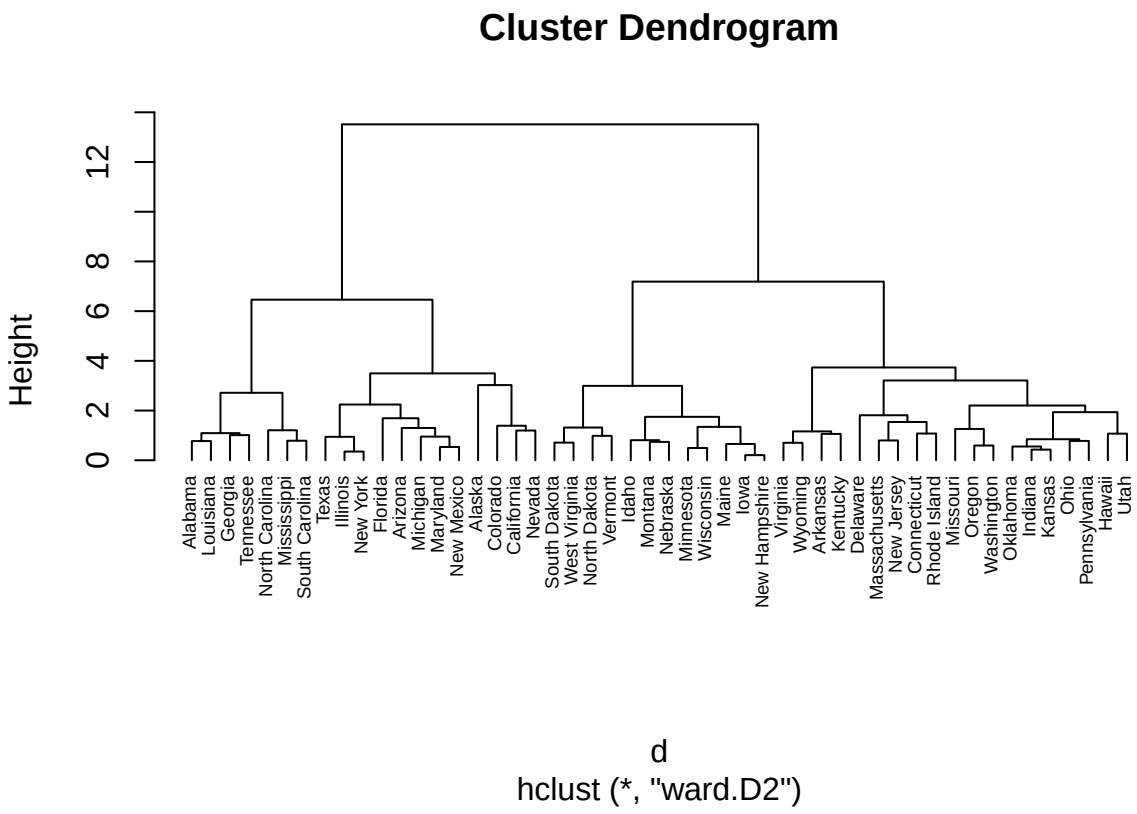
```
# As we don't want the hierarchical clustering result to depend to an arbitrary variable unit,
# we start by scaling the data using the R function scale() as follows
# ---
#
df <- scale(df)
head(df)
```

```
##               Murder   Assault   UrbanPop         Rape
## Alabama    1.24256408 0.7828393 -0.5209066 -0.003416473
## Alaska     0.50786248 1.1068225 -1.2117642  2.484202941
## Arizona    0.07163341 1.4788032  0.9989801  1.042878388
## Arkansas   0.23234938 0.2308680 -1.0735927 -0.184916602
## California 0.27826823 1.2628144  1.7589234  2.067820292
## Colorado   0.02571456 0.3988593  0.8608085  1.864967207
```

```
# We now use the R function hclust() for hierarchical clustering
# ---
#
# First we use the dist() function to compute the Euclidean distance between observations,
# d will be the first argument in the hclust() function dissimilarity matrix
# ---
#
d <- dist(df, method = "euclidean")
# We then hierarchical clustering using the Ward's method
# ---
#
res.hc <- hclust(d, method = "ward.D2" )
```

Plot the dendrogram

```
# Lastly, we plot the obtained dendrogram
# ---
#
plot(res.hc, cex = 0.6, hang = -1)
```



Challenge 1

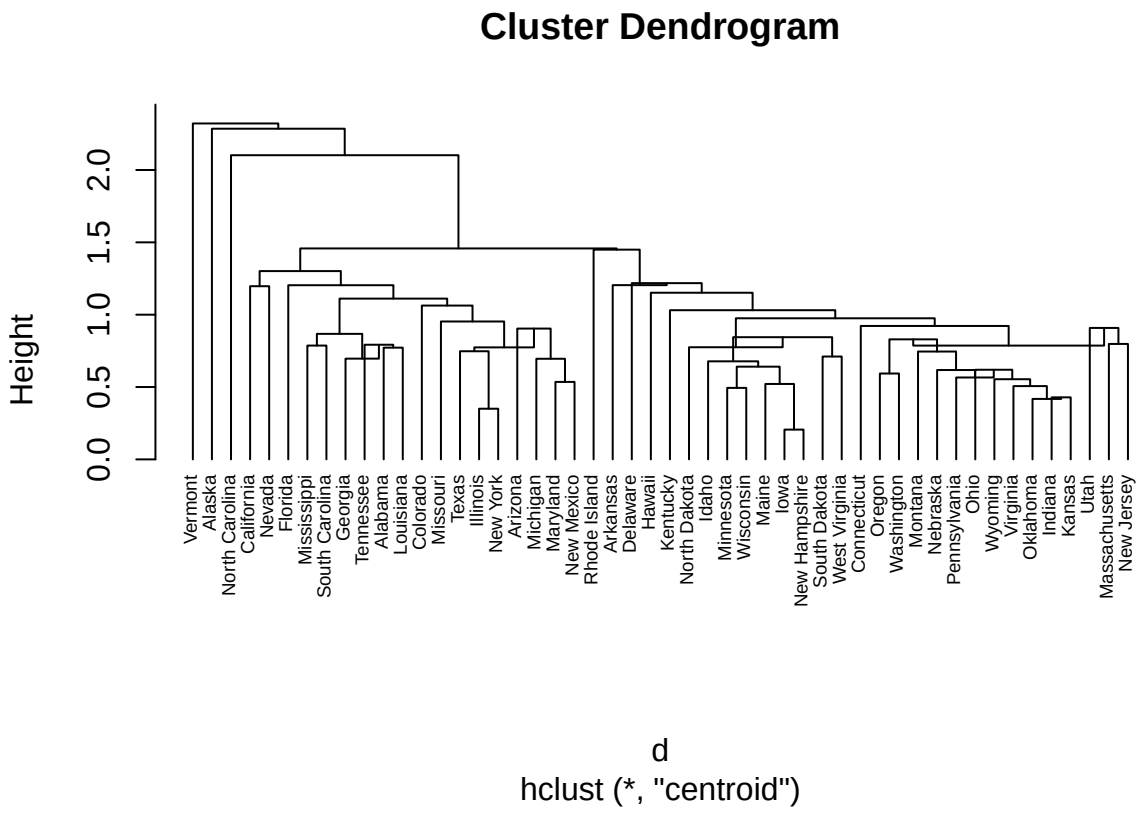
Using different centroid hierachical clustering

```
# Question: Using the USArrests datasets in the above example,
# compute hierarchical clustering with other linkage methods,
# such as single, median, average, centroid, Ward's and McQuitty's.
# ---
# Hint: You can refer to the R documentation in the suggested resources
# We now use the R function hclust() for hierarchical clustering
# ---
#
# First we use the dist() function to compute the Euclidean distance between observations,
# d will be the first argument in the hclust() function dissimilarity matrix
# ---
#
d <- dist(df, method = "euclidean")
# We then hierarchical clustering using the Ward's method
# ---
#
res.hc <- hclust(d, method = "centroid" )
```

Plot the dendrogram

```
# Lastly, we plot the obtained dendrogram
# ---
```

```
#
plot(res.hc, cex = 0.6, hang = -1)
```

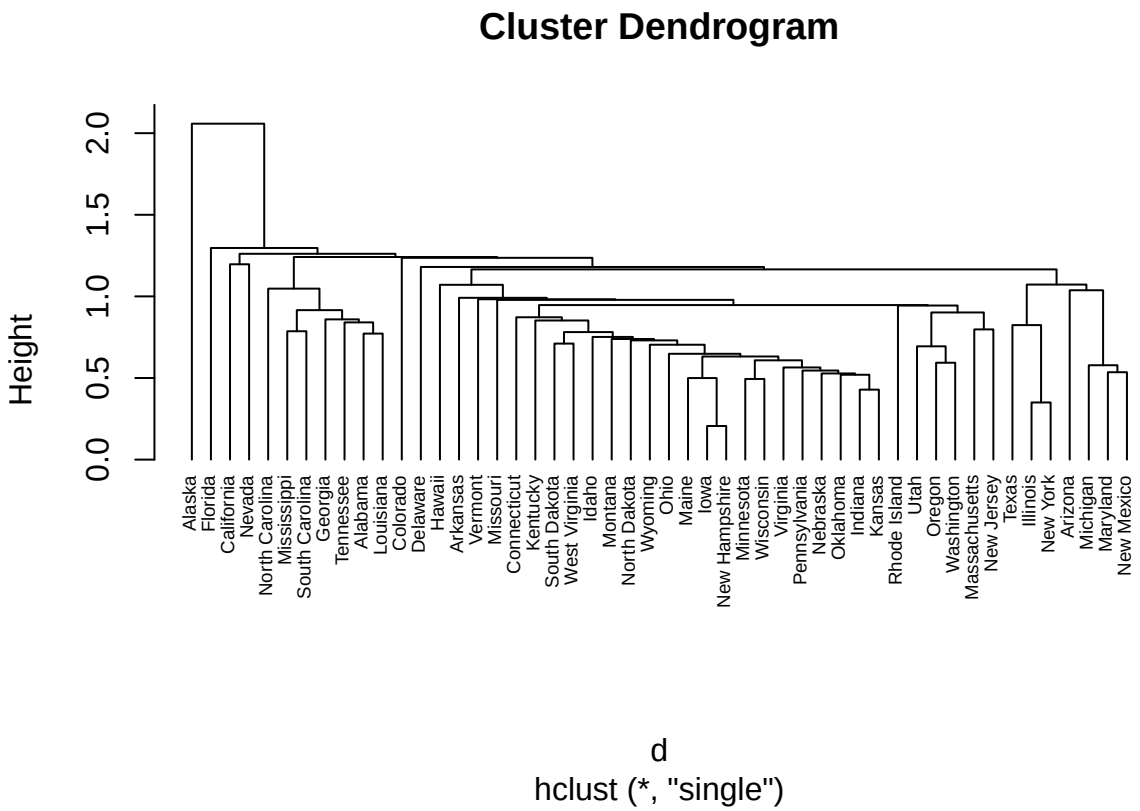


Using different single hierachical clustering

```
# Question: Using the USArrests datasets in the above example,
# compute hierarchical clustering with other linkage methods,
# such as single, median, average, centroid, Ward's and McQuitty's.
# ---
# Hint: You can refer to the R documentation in the suggested resources
# We now use the R function hclust() for hierarchical clustering
# ---
#
# First we use the dist() function to compute the Euclidean distance between observations,
# d will be the first argument in the hclust() function dissimilarity matrix
# ---
#
d <- dist(df, method = "euclidean")
# We then hierarchical clustering using the Ward's method
# ---
#
res.hc <- hclust(d, method = "single" )
```

Plot the dendrogram

```r
# Lastly, we plot the obtained dendrogram
# ---
#
plot(res.hc, cex = 0.6, hang = -1)
```

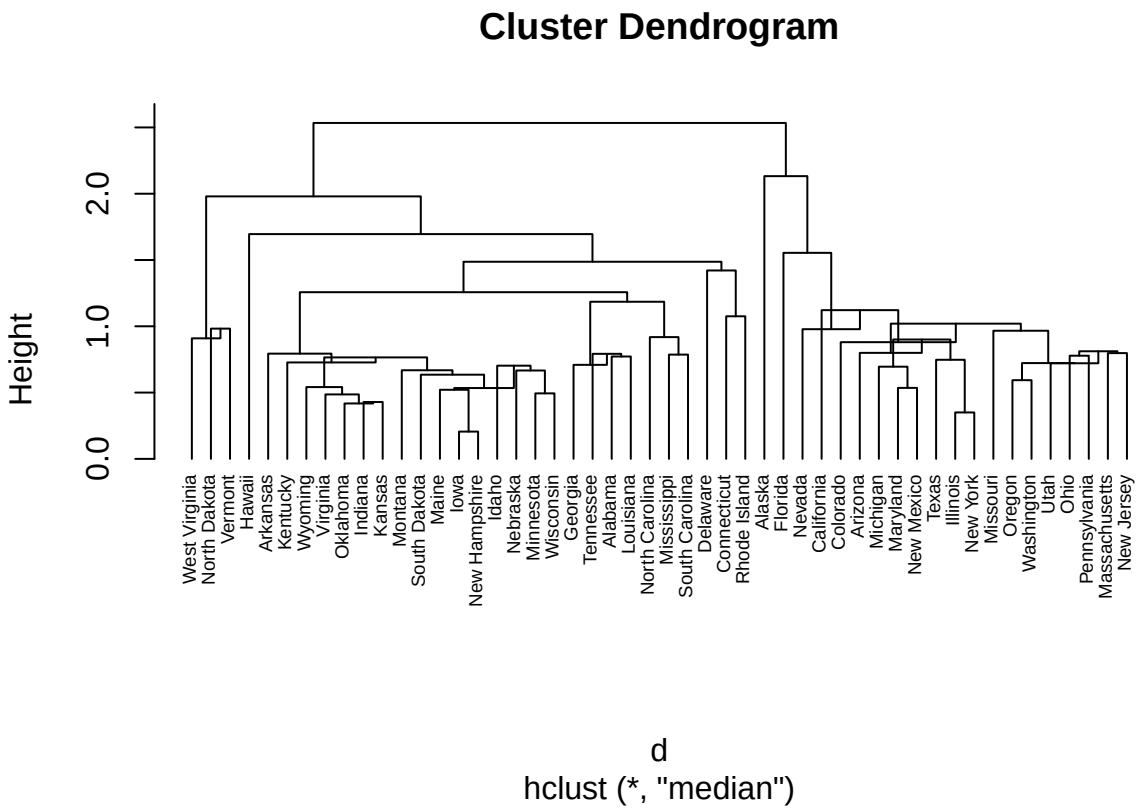


Using different median hierachical clustering

```r
# Question: Using the USArrests datasets in the above example,
# compute hierarchical clustering with other linkage methods,
# such as single, median, average, centroid, Ward's and McQuitty's.
# ---
# Hint: You can refer to the R documentation in the suggested resources
# We now use the R function hclust() for hierarchical clustering
# ---
#
# First we use the dist() function to compute the Euclidean distance between observations,
# d will be the first argument in the hclust() function dissimilarity matrix
# ---
#
d <- dist(df, method = "euclidean")
# We then hierarchical clustering using the Ward's method
# ---
#
res.hc <- hclust(d, method = "median" )
```

Plot the dendrogram

```r
# Lastly, we plot the obtained dendrogram
# ---
#
plot(res.hc, cex = 0.6, hang = -1)
```

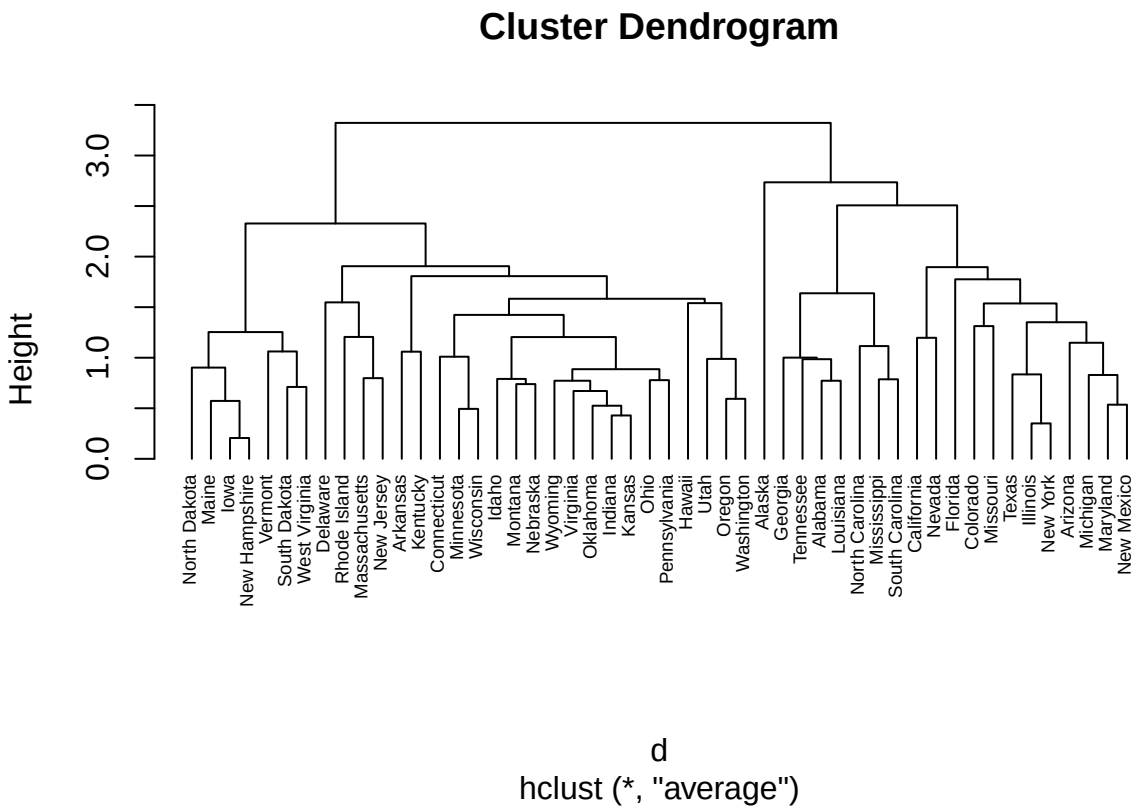


Using different centroid hierachical clustering

```r
# Question: Using the USArrests datasets in the above example,
# compute hierarchical clustering with other linkage methods,
# such as single, median, average, centroid, Ward's and McQuitty's.
# ---
# Hint: You can refer to the R documentation in the suggested resources
# We now use the R function hclust() for hierarchical clustering
# ---
#
# First we use the dist() function to compute the Euclidean distance between observations,
# d will be the first argument in the hclust() function dissimilarity matrix
# ---
#
d <- dist(df, method = "euclidean")
# We then hierarchical clustering using the Ward's method
# ---
#
res.hc <- hclust(d, method = "average" )
```

Plot the dendrogram

```
# Lastly, we plot the obtained dendrogram
# ---
#
plot(res.hc, cex = 0.6, hang = -1)
```



Challenge 2

```
# Question: Perform hierarchical clustering using the mtcars dataset
# ---
#
# Loading our dataset below
# ---
#
df <- mtcars
# Previewing our dataset
# ---
#
head(df)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Dataset descriptive statistics

```
# Before hierarchical clustering, we can compute some descriptive statistics
# ---
#
summary(df)
```

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt             qsec             vs
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

Scaling the dataset

```
# As we don't want the hierarchical clustering result to depend to an arbitrary variable unit,
# we start by scaling the data using the R function scale() as follows
# ---
#
df <- scale(df)
head(df)
```

```
##                           mpg        cyl        disp         hp       drat
## Mazda RX4           0.1508848 -0.1049878 -0.57061982 -0.5350928  0.5675137
## Mazda RX4 Wag       0.1508848 -0.1049878 -0.57061982 -0.5350928  0.5675137
## Datsun 710          0.4495434 -1.2248578 -0.99018209 -0.7830405  0.4739996
## Hornet 4 Drive      0.2172534 -0.1049878  0.22009369 -0.5350928 -0.9661175
## Hornet Sportabout  -0.2307345  1.0148821  1.04308123  0.4129422 -0.8351978
## Valiant            -0.3302874 -0.1049878 -0.04616698 -0.6080186 -1.5646078
##                           wt       qsec         vs         am       gear
## Mazda RX4         -0.610399567 -0.7771651 -0.8680278  1.1899014  0.4235542
## Mazda RX4 Wag     -0.349785269 -0.4637808 -0.8680278  1.1899014  0.4235542
## Datsun 710        -0.917004624  0.4260068  1.1160357  1.1899014  0.4235542
## Hornet 4 Drive    -0.002299538  0.8904872  1.1160357 -0.8141431 -0.9318192
## Hornet Sportabout  0.227654255 -0.4637808 -0.8680278 -0.8141431 -0.9318192
## Valiant            0.248094592  1.3269868  1.1160357 -0.8141431 -0.9318192
##                           carb
```

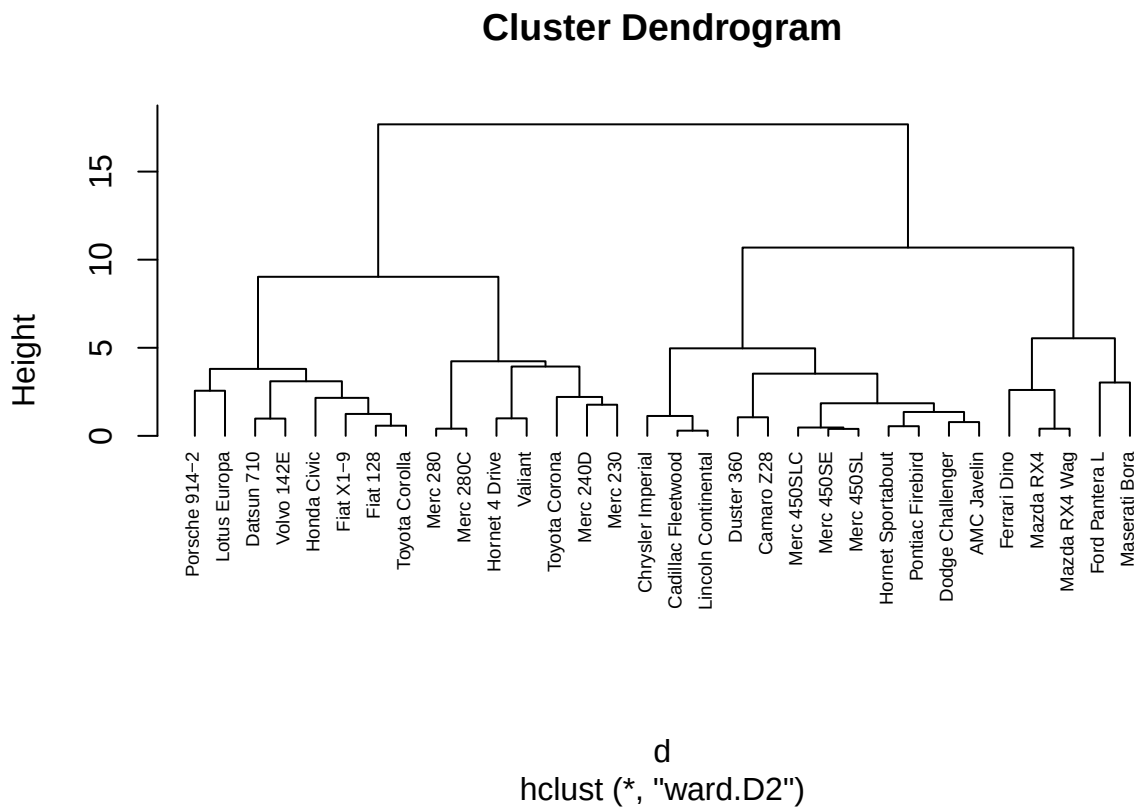
```
## Mazda RX4           0.7352031
## Mazda RX4 Wag       0.7352031
## Datsun 710         -1.1221521
## Hornet 4 Drive     -1.1221521
## Hornet Sportabout -0.5030337
## Valiant            -1.1221521
```

Performing Hierachical clastering

```
# We now use the R function hclust() for hierarchical clustering
# ---
#
# First we use the dist() function to compute the Euclidean distance between observations,
# d will be the first argument in the hclust() function dissimilarity matrix
# ---
#
d <- dist(df, method = "euclidean")
# We then hierarchical clustering using the Ward's method
# ---
#
res.hc <- hclust(d, method = "ward.D2" )
```

Plot the dendrogram

```
# Lastly, we plot the obtained dendrogram
# ---
#
plot(res.hc, cex = 0.6, hang = -1)
```



Challenge 3

```
# Perform hierarchical clustering using the iris dataset
# ---
#
# Loading our dataset below
# ---
#
df <- iris
# Previewing our dataset
# ---
#
head(df)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

Hot Encording our dataset

```
#library(caret)

#dmy <- dummyVars(" ~ .", data = df, fullRank = T)
#df_transformed <- data.frame(predict(dmy, newdata = df))

#head(df_transformed)
#df$Species <- ifelse(df$Species == "setosa",1,0)

#table(df$Species) - BINARY

# CONTINUOUS DATA
df$Species<-as.integer(as.factor(df$Species))
df$Species
```

```
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [75] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3
## [112] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## [149] 3 3
```

```
str(df)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : int  1 1 1 1 1 1 1 1 1 1 ...
```

Dataset descriptive statistics

```
# Before hierarchical clustering, we can compute some descriptive statistics
# ---
#
summary(df)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width       Species
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   Min.   :1
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   1st Qu.:1
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300   Median :2
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199   Mean   :2
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800   3rd Qu.:3
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500   Max.   :3
```

Scaling the dataset

```
# As we don't want the hierarchical clustering result to depend to an arbitrary variable unit,
# we start by scaling the data using the R function scale() as follows
# ---
#
df <- scale(df)
head(df)
```

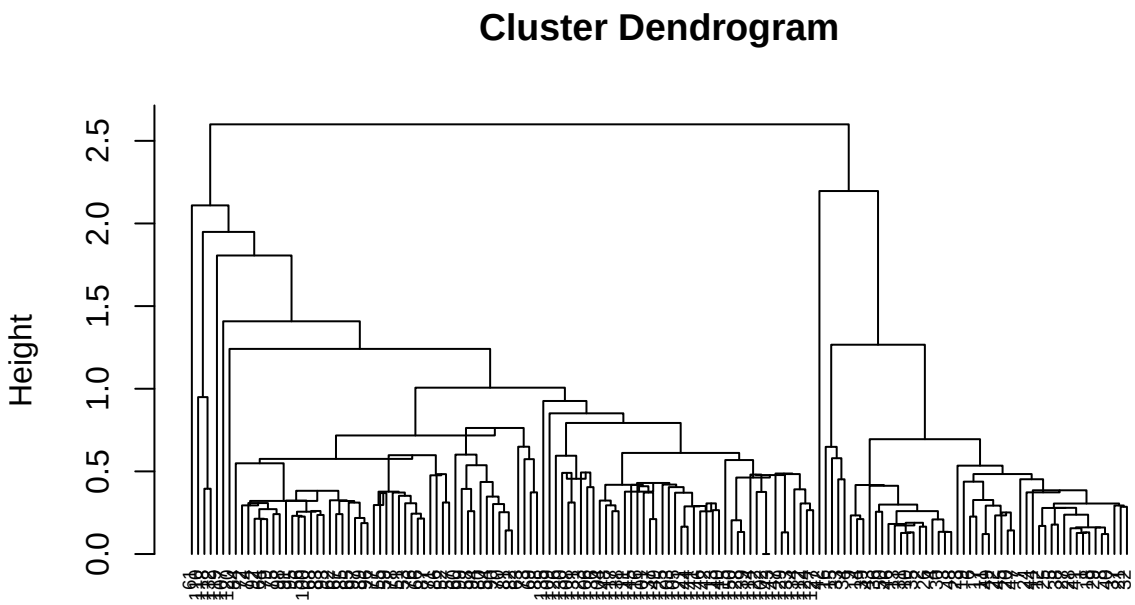
```
##      Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
## [1,]   -0.8976739  1.01560199    -1.335752   -1.311052 -1.220656
## [2,]   -1.1392005 -0.13153881    -1.335752   -1.311052 -1.220656
## [3,]   -1.3807271  0.32731751    -1.392399   -1.311052 -1.220656
## [4,]   -1.5014904  0.09788935    -1.279104   -1.311052 -1.220656
## [5,]   -1.0184372  1.24503015    -1.335752   -1.311052 -1.220656
## [6,]   -0.5353840  1.93331463    -1.165809   -1.048667 -1.220656
```

Performing Hierachical clastering

```
# We now use the R function hclust() for hierarchical clustering
# ---
#
# First we use the dist() function to compute the Euclidean distance between observations,
# d will be the first argument in the hclust() function dissimilarity matrix
# ---
#
d <- dist(df, method = "euclidean")
# We then hierarchical clustering using the Ward's method
# ---
#
res.hc <- hclust(d, method = "centroid" )
```

Plot the dendrogram

```
# Lastly, we plot the obtained dendrogram
# ---
#
plot(res.hc, cex = 0.6, hang = -1)
```



Challenge 4

```
# Perform hierarchical cluster analysis on the given protein consumption.
# ---
# Dataset url = http://bit.ly/HierarchicalClusteringDataset
# ---
#
protein_df <- read.csv("http://bit.ly/HierarchicalClusteringDataset")
head(protein_df)
```

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

```
#
```

Label encoding our dataset, finding summery and checking data type

```
# CONTINUOUS DATA
protein_df$Country<-as.integer(as.factor(protein_df$Country))
head(protein_df$Country)
```

```
## [1] 1 2 3 4 5 6
```

```
summary(protein_df)
```

```
##     Country      RedMeat         WhiteMeat           Eggs            Milk
##  Min.   : 1   Min.   : 4.400   Min.   : 1.400   Min.   :0.500   Min.   : 4.90
##  1st Qu.: 7   1st Qu.: 7.800   1st Qu.: 4.900   1st Qu.:2.700   1st Qu.:11.10
##  Median :13   Median : 9.500   Median : 7.800   Median :2.900   Median :17.60
##  Mean   :13   Mean   : 9.828   Mean   : 7.896   Mean   :2.936   Mean   :17.11
##  3rd Qu.:19   3rd Qu.:10.600   3rd Qu.:10.800   3rd Qu.:3.700   3rd Qu.:23.30
##  Max.   :25   Max.   :18.000   Max.   :14.000   Max.   :4.700   Max.   :33.70
##       Fish           Cereals          Starch           Nuts
##  Min.   : 0.200   Min.   :18.60   Min.   :0.600   Min.   :0.700
##  1st Qu.: 2.100   1st Qu.:24.30   1st Qu.:3.100   1st Qu.:1.500
##  Median : 3.400   Median :28.00   Median :4.700   Median :2.400
##  Mean   : 4.284   Mean   :32.25   Mean   :4.276   Mean   :3.072
##  3rd Qu.: 5.800   3rd Qu.:40.10   3rd Qu.:5.700   3rd Qu.:4.700
##  Max.   :14.200   Max.   :56.70   Max.   :6.500   Max.   :7.800
##      Fr.Veg
##  Min.   :1.400
##  1st Qu.:2.900
##  Median :3.800
##  Mean   :4.136
##  3rd Qu.:4.900
##  Max.   :7.900
```

```
str(protein_df)
```

```
## 'data.frame':    25 obs. of  10 variables:
##  $ Country  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ RedMeat  : num  10.1 8.9 13.5 7.8 9.7 10.6 8.4 9.5 18 10.2 ...
##  $ WhiteMeat: num  1.4 14 9.3 6 11.4 10.8 11.6 4.9 9.9 3 ...
##  $ Eggs     : num  0.5 4.3 4.1 1.6 2.8 3.7 3.7 2.7 3.3 2.8 ...
##  $ Milk     : num  8.9 19.9 17.5 8.3 12.5 25 11.1 33.7 19.5 17.6 ...
##  $ Fish     : num  0.2 2.1 4.5 1.2 2 9.9 5.4 5.8 5.7 5.9 ...
##  $ Cereals  : num  42.3 28 26.6 56.7 34.3 21.9 24.6 26.3 28.1 41.7 ...
##  $ Starch   : num  0.6 3.6 5.7 1.1 5 4.8 6.5 5.1 4.8 2.2 ...
##  $ Nuts     : num  5.5 1.3 2.1 3.7 1.1 0.7 0.8 1 2.4 7.8 ...
##  $ Fr.Veg   : num  1.7 4.3 4 4.2 4 2.4 3.6 1.4 6.5 6.5 ...
```

Scaling the dataset

```
# As we don't want the hierarchical clustering result to depend to an arbitrary variable unit,
# we start by scaling the data using the R function scale() as follows
# ---
#
protein_df <- scale(protein_df)
head(protein_df)
```

```
##         Country     RedMeat  WhiteMeat       Eggs        Milk        Fish
## [1,] -1.6304789  0.08126490 -1.7584889 -2.1796385 -1.15573814 -1.20028213
## [2,] -1.4946057 -0.27725673  1.6523731  1.2204544  0.39237676 -0.64187467
```

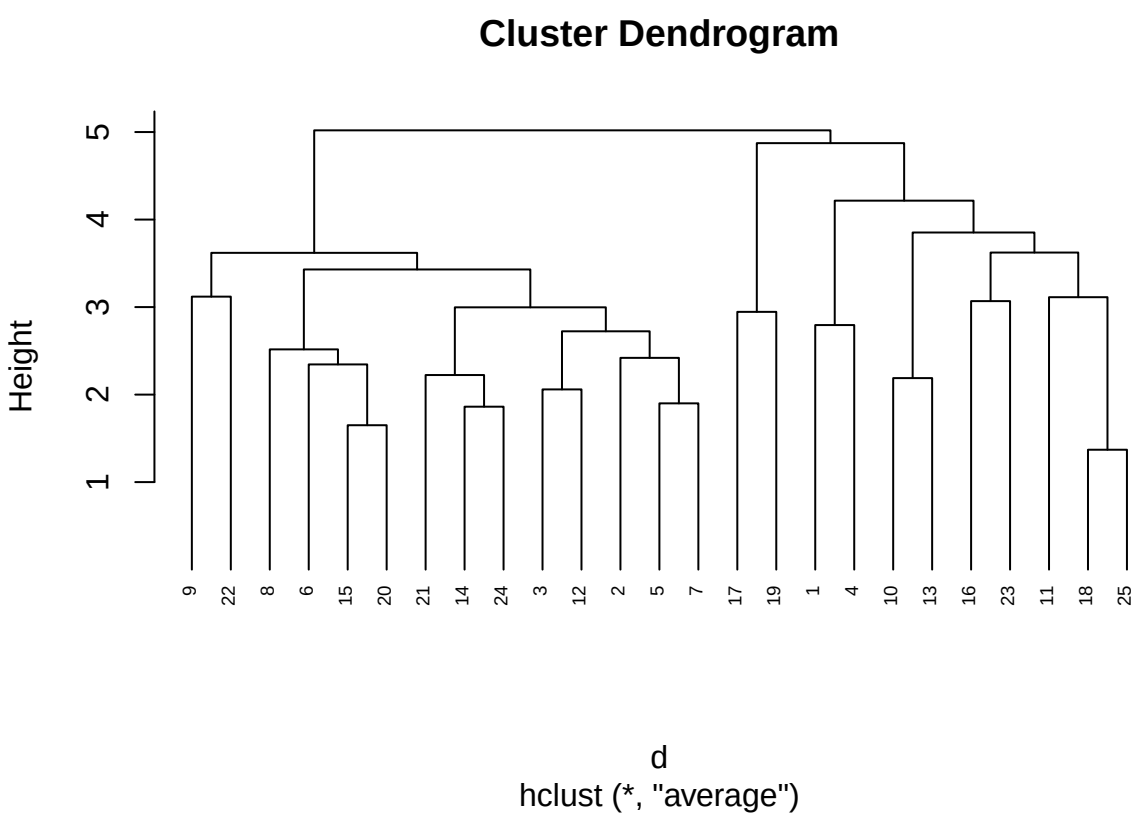
```
## [3,] -1.3587324  1.09707621  0.3800675  1.0415022  0.05460623  0.06348211
## [4,] -1.2228592 -0.60590157 -0.5132535 -1.1954011 -1.24018077 -0.90638347
## [5,] -1.0869860 -0.03824231  0.9485445 -0.1216875 -0.64908235 -0.67126454
## [6,] -0.9511127  0.23064892  0.7861225  0.6835976  1.11013912  1.65053488
##         Cereals     Starch       Nuts      Fr.Veg
## [1,]  0.9159176 -2.2495772  1.2227536 -1.35040507
## [2,] -0.3870690 -0.4136872 -0.8923886  0.09091397
## [3,] -0.5146342  0.8714358 -0.4895043 -0.07539207
## [4,]  2.2280161 -1.9435955  0.3162641  0.03547862
## [5,]  0.1869740  0.4430614 -0.9931096 -0.07539207
## [6,] -0.9428885  0.3206688 -1.1945517 -0.96235764
```

Performing Hierachical clastering Average

```
# We now use the R function hclust() for hierarchical clustering
# ---
#
# First we use the dist() function to compute the Euclidean distance between observations,
# d will be the first argument in the hclust() function dissimilarity matrix
# ---
#
d <- dist(protein_df, method = "euclidean")
# We then hierarchical clustering using the Ward's method
# ---
#
res.hc <- hclust(d, method = "average" )
```

Plot the dendrogram

```
# Lastly, we plot the obtained dendrogram
# ---
#
plot(res.hc, cex = 0.6, hang = -1)
```



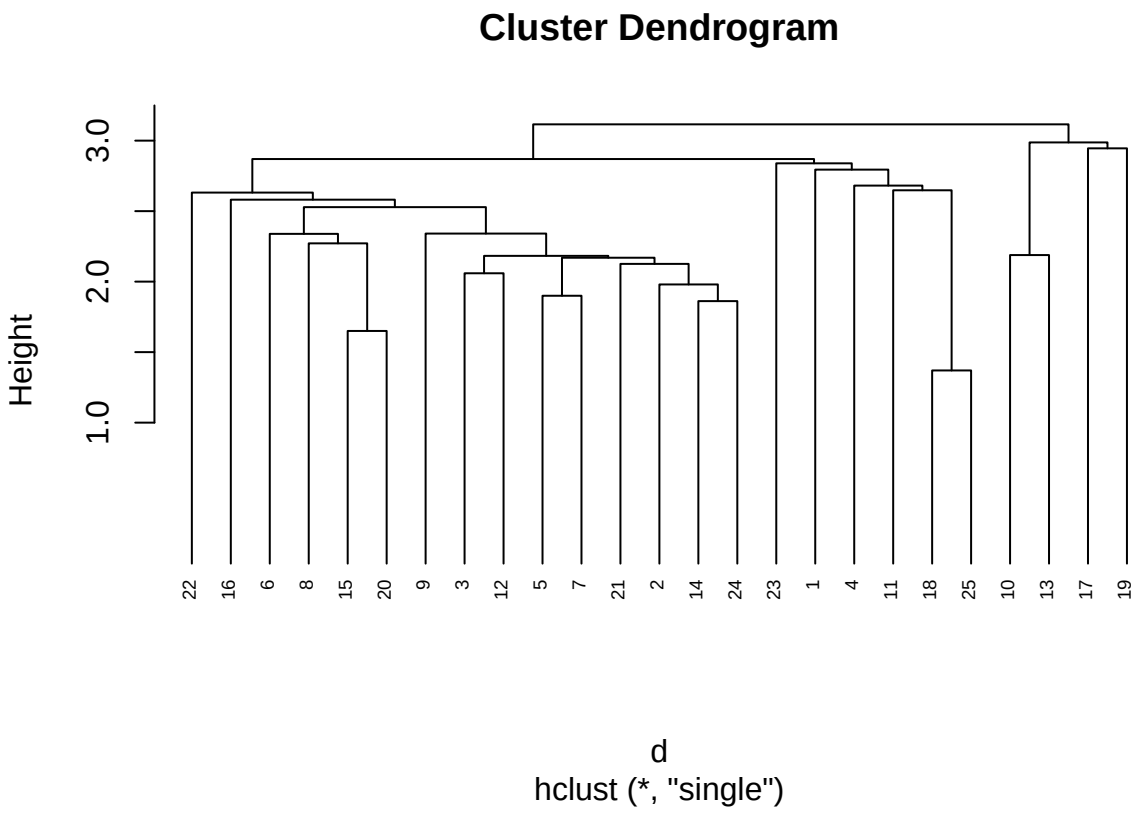
Challenge 4

Performing Hierachical clastering single

```
# We now use the R function hclust() for hierarchical clustering
# ---
#
# First we use the dist() function to compute the Euclidean distance between observations,
# d will be the first argument in the hclust() function dissimilarity matrix
# ---
#
d <- dist(protein_df, method = "euclidean")
# We then hierarchical clustering using the Ward's method
# ---
#
res.hc <- hclust(d, method = "single" )
```

Plot the dendrogram

```
# Lastly, we plot the obtained dendrogram
# ---
#
plot(res.hc, cex = 0.6, hang = -1)
```



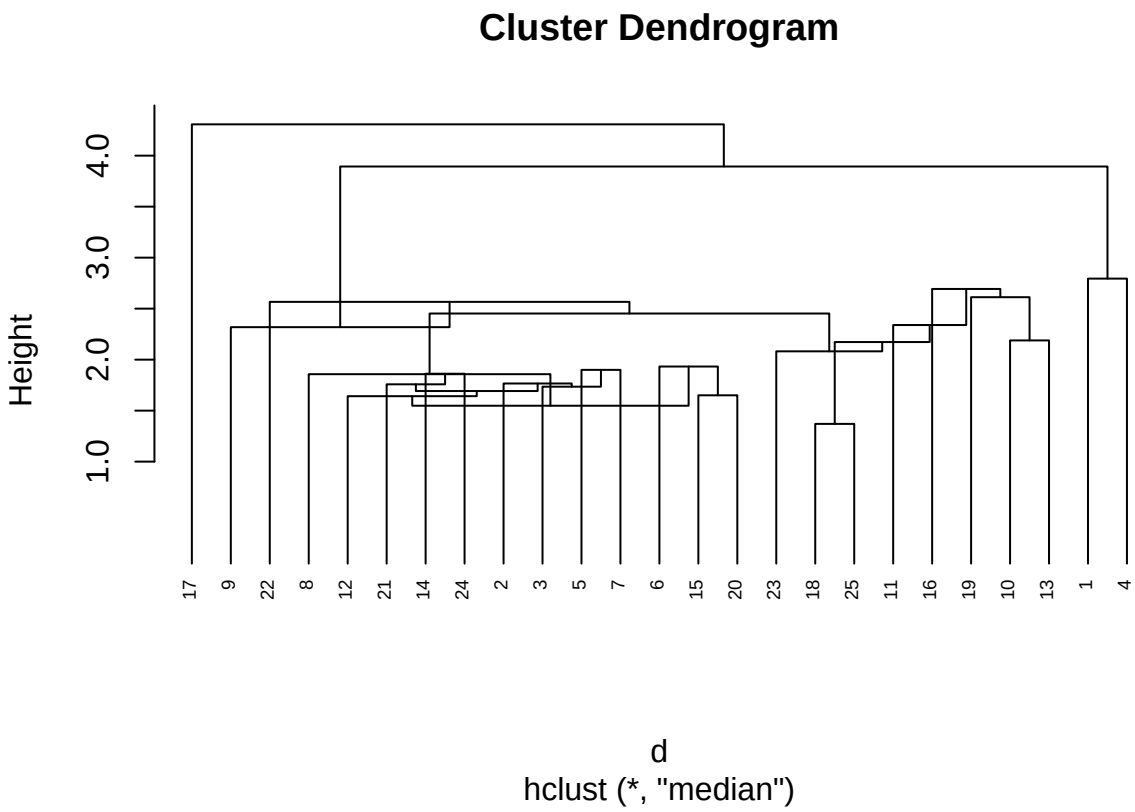
Challenge 4

Performing Hierachical clastering median

```
# We now use the R function hclust() for hierarchical clustering
# ---
#
# First we use the dist() function to compute the Euclidean distance between observations,
# d will be the first argument in the hclust() function dissimilarity matrix
# ---
#
d <- dist(protein_df, method = "euclidean")
# We then hierarchical clustering using the Ward's method
# ---
#
res.hc <- hclust(d, method = "median" )
```

Plot the dendrogram

```
# Lastly, we plot the obtained dendrogram
# ---
#
plot(res.hc, cex = 0.6, hang = -1)
```


Cluster Dendrogram
Height
1.0
2.0
3.0
4.0
17
9
22
8
12
21
14
24
2
3
5
7
6
15
20
23
18
25
11
16
19
10
13
1
4
d
hclust (*, "median")