Hierarchical Clustering

Arbuda Sivani

4/17/2022

```
Cereals <- read.csv("~/ML/Assignment/Assignment_5/Cereals.csv")
View(Cereals)
```

```
#Loading libraries
library(cluster)
```

```
## Warning: package 'cluster' was built under R version 4.1.3
```

```
library(caret)
```

```
## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.1.3

## Loading required package: lattice
```

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.1.3

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

```
library(dendextend)
```

```
## Warning: package 'dendextend' was built under R version 4.1.3

##
## ---------------------
## Welcome to dendextend version 1.15.2
## Type citation('dendextend') for how to cite the package.
##
## Type browseVignettes(package = 'dendextend') for the package vignette.
## The github page is: https://github.com/talgalili/dendextend/
##
## Suggestions and bug-reports can be submitted at: https://github.com/talgalili/dendextend/issues
## You may ask questions at stackoverflow, use the r and dendextend tags:
##   https://stackoverflow.com/questions/tagged/dendextend
##
##  To suppress this message use:  suppressPackageStartupMessages(library(dendextend))
## ---------------------
```

```
##
## Attaching package: 'dendextend'
```

```
## The following object is masked from 'package:stats':
##
##     cutree
```

```
#Removing missing values
Cereals_DF <- data.frame(Cereals[,4:16])
head(Cereals_DF)
```

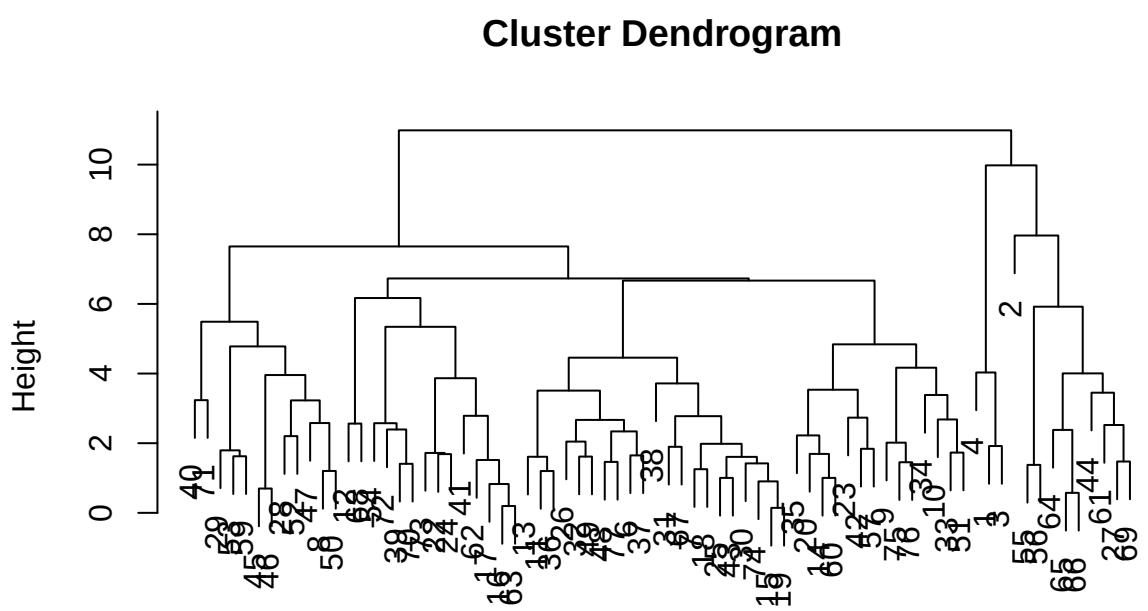
```
##   calories protein fat sodium fiber carbo sugars potass vitamins shelf weight
## 1       70       4   1    130  10.0   5.0      6    280       25     3      1
## 2      120       3   5     15   2.0   8.0      8    135        0     3      1
## 3       70       4   1    260   9.0   7.0      5    320       25     3      1
## 4       50       4   0    140  14.0   8.0      0    330       25     3      1
## 5      110       2   2    200   1.0  14.0      8     NA       25     3      1
## 6      110       2   2    180   1.5  10.5     10     70       25     1      1
##   cups   rating
## 1 0.33 68.40297
## 2 1.00 33.98368
## 3 0.33 59.42551
## 4 0.50 93.70491
## 5 0.75 34.38484
## 6 0.75 29.50954
```

```
Cereals_NA <- na.omit(Cereals)
```

```
#Normalising and Scaling the data
Cereals_norm <- scale(Cereals_NA[,4:16])
```

#1.Apply hierarchical clustering to the data using Euclidean distance to the normalizedmeasurements. Use Agnes to compare the clustering from single linkage, complete linkage, average linkage, and Ward. Choose the best method.

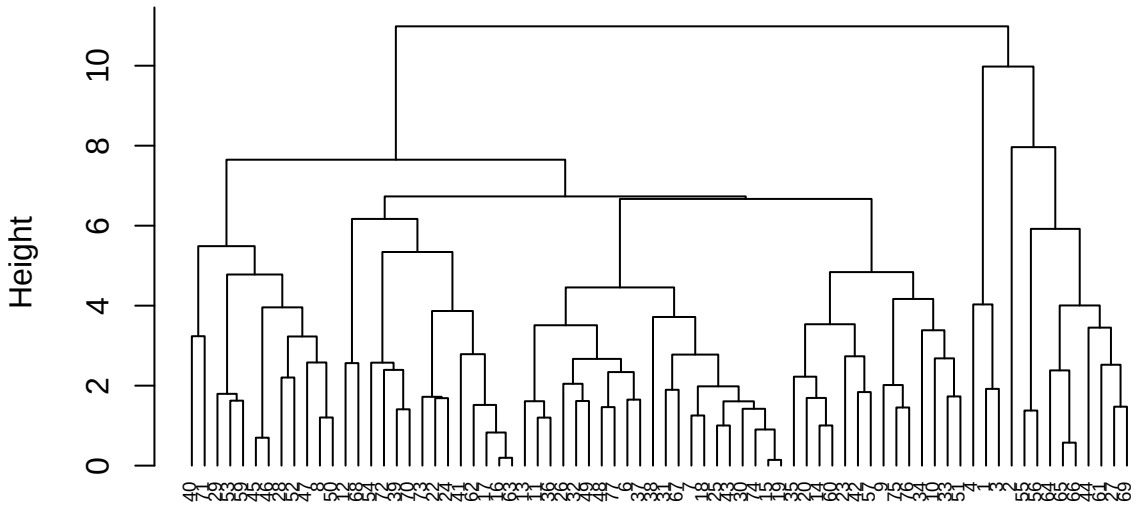
```
d <- dist(Cereals_norm, method = "euclidean")
HC <- hclust(d, method = "complete")
plot(HC)
```




```
round(HC$height, 3)
```

```
##  [1]  0.143  0.196  0.575  0.698  0.828  0.904  1.003  1.004  1.201  1.203
## [11]  1.254  1.378  1.408  1.421  1.454  1.463  1.474  1.517  1.608  1.611
## [21]  1.616  1.625  1.650  1.687  1.692  1.720  1.730  1.795  1.839  1.897
## [31]  1.919  1.982  2.015  2.046  2.203  2.224  2.339  2.381  2.394  2.522
## [41]  2.563  2.574  2.579  2.668  2.682  2.734  2.776  2.787  3.229  3.236
## [51]  3.385  3.451  3.510  3.535  3.717  3.866  3.957  4.005  4.031  4.168
## [61]  4.456  4.779  4.839  5.342  5.488  5.920  6.169  6.669  6.731  7.650
## [71]  7.964  9.979 10.984
```

```
plot(HC, cex=0.6, hang = -1)
```



d
hclust (*, "complete")

```
#Using Agnes to compare the clustering from single linkage, complete linkage, average linkage and Ward.
hc_single <- agnes(Cereals_norm, method = "single")
hc_complete <- agnes(Cereals_norm, method = "complete")
hc_average <- agnes(Cereals_norm, method = "average")
hc_ward <- agnes(Cereals_norm, method = "ward")

#Comparing the agnes coefficients for Single, complete, average and ward method
print(hc_single$ac)
```

```
## [1] 0.6067859
```

```
print(hc_complete$ac)
```

```
## [1] 0.8353712
```

```
print(hc_average$ac)
```

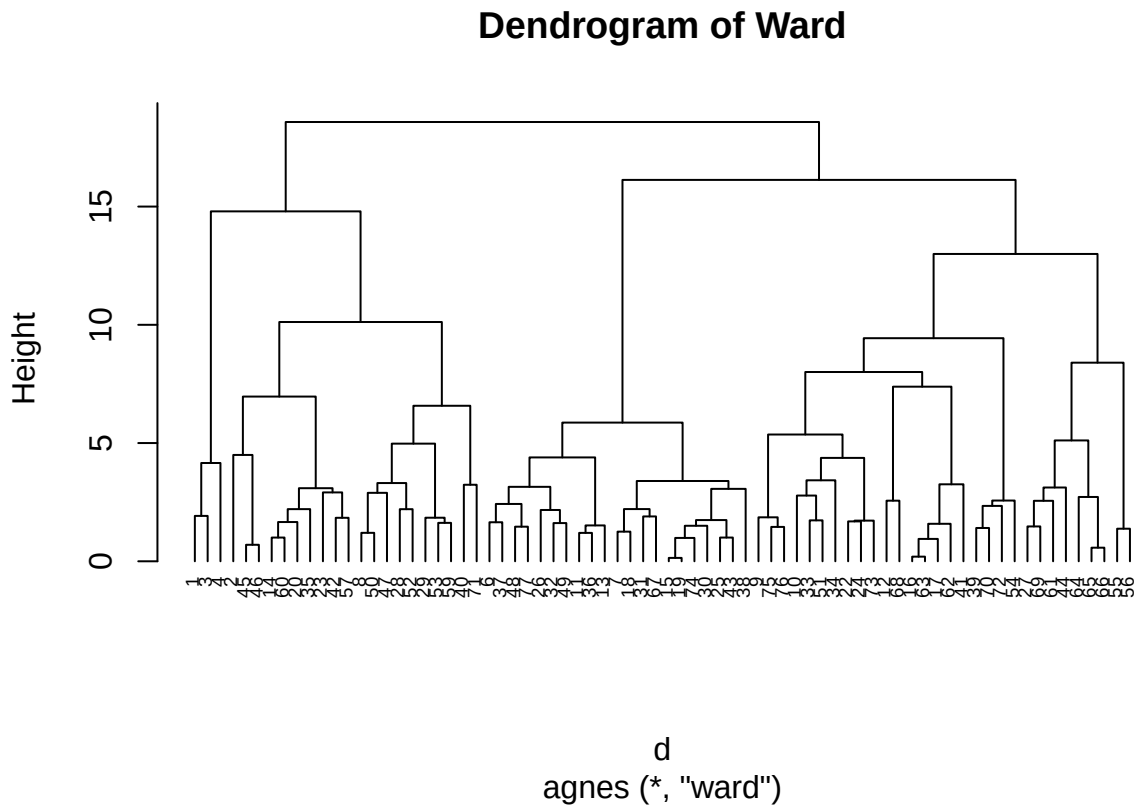
```
## [1] 0.7766075
```

```
print(hc_ward$ac)
```

```
## [1] 0.9046042
```

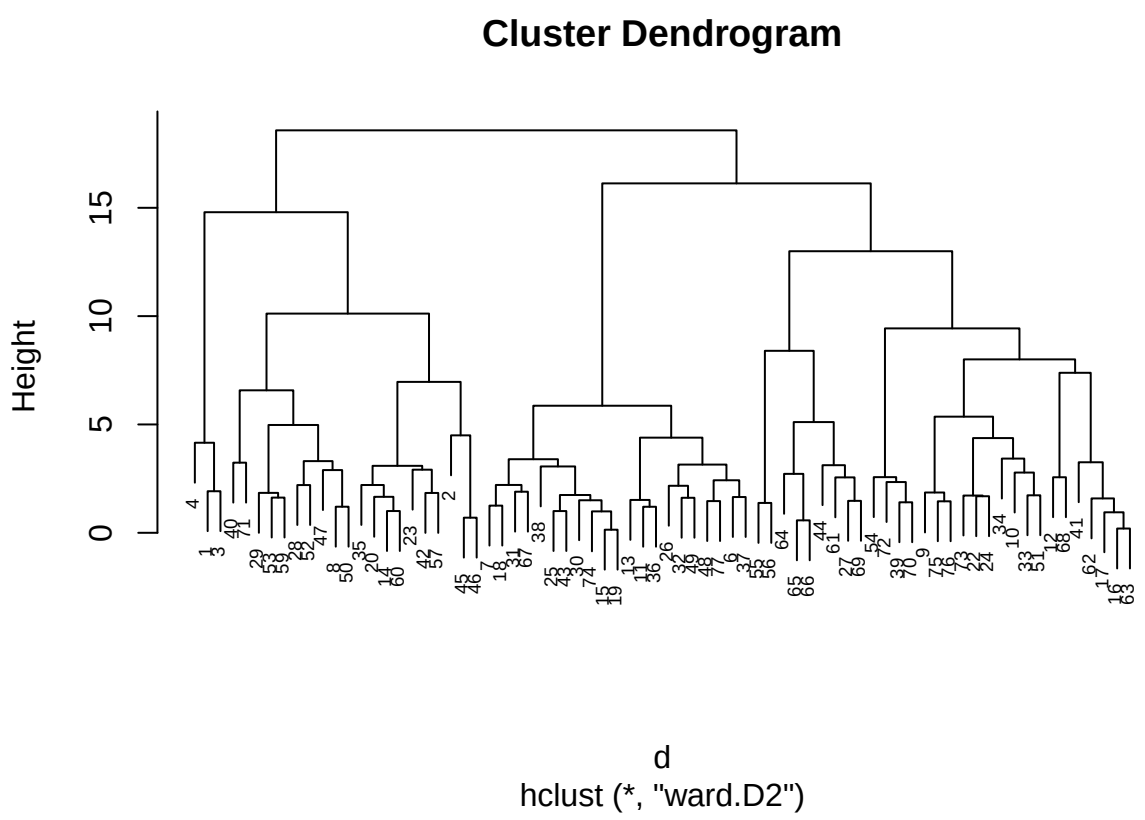
#From the above result, we can see that the ward method has the highest agnes coefficient of 0.904. Hence, ward method is taken as the best.

```
hc_ward <- agnes(d, method = "ward")
pltree(hc_ward, cex=0.6, hang=-1,main = "Dendrogram of Ward")
```

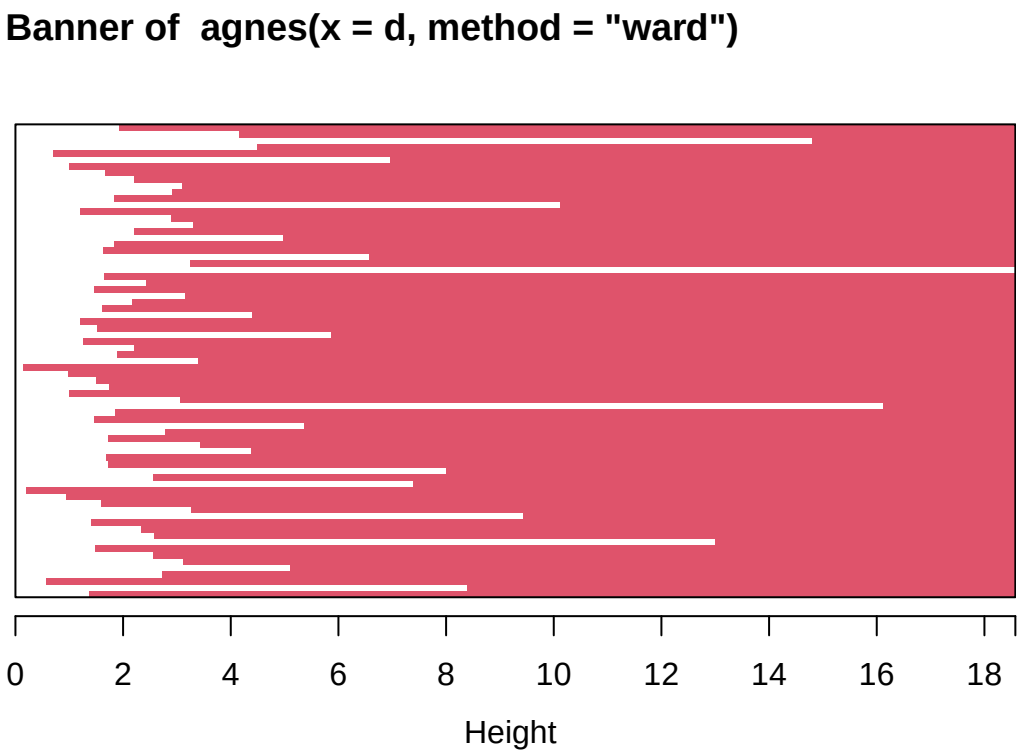


#2.How many clusters would you choose?

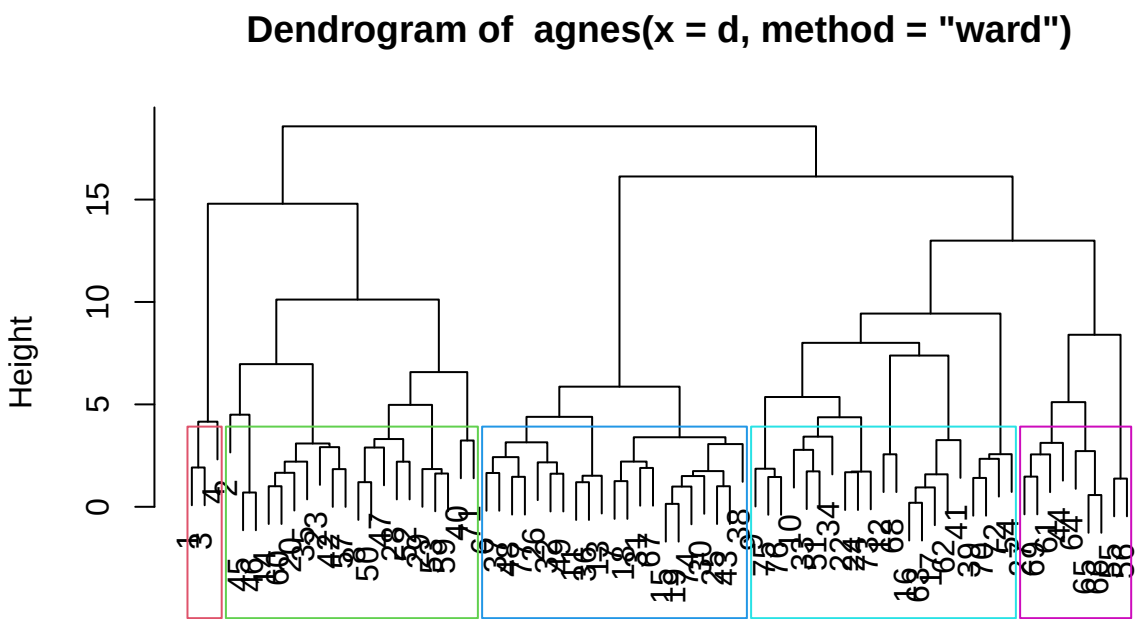
```
HC_1 <- hclust(d, method = "ward.D2")
plot(HC_1,cex=0.6)
```




```
plot(hc_ward)
```




```
rect.hclust(hc_ward, k=5, border = 2:10)
```




```
Group <- cutree(HC_1, k=5)
table(Group)
```

```
## Group
##  1  2  3  4  5
##  3 20 21 21  9
```

```
DF <- as.data.frame(cbind(Cereals_norm,Group))

#Visualize the clusters on a scatterplot
fviz_cluster(list(data=Cereals_norm, cluster = Group))
```

#To determine the value of k, the larget difference in height can be used to calculate the value of k. Hence, k= is the optimal number of clusters.

#3.The elementary public schools would like to choose a set of cereals to include in their daily cafeterias. Every day a different cereal is offered, but all cereals should support a healthy diet. For this goal, you are requested to find a cluster of "healthy cereals."

```
New_Cereals <- Cereals
New_Cereals_NA <- na.omit(New_Cereals)
Clust <- cbind(New_Cereals_NA, Group)

#Finding out the cluster of healthy cereals
mean(Clust[Clust$Group==1,"rating"])
```

```
## [1] 73.84446
```

```
mean(Clust[Clust$Group==2,"rating"])
```

```
## [1] 38.26161
```

```
mean(Clust[Clust$Group==3,"rating"])
```

```
## [1] 28.84825
```

```
mean(Clust[Clust$Group==4,"rating"])
```

```
## [1] 46.46513
```

```
mean(Clust[Clust$Group==5,"rating"])
```

```
## [1] 63.0184
```

#After looking at the above results it is clear that Cluster1 has the highest mean(73.84) which implicates that cluster1 is a healthy cluster