Assignment 5

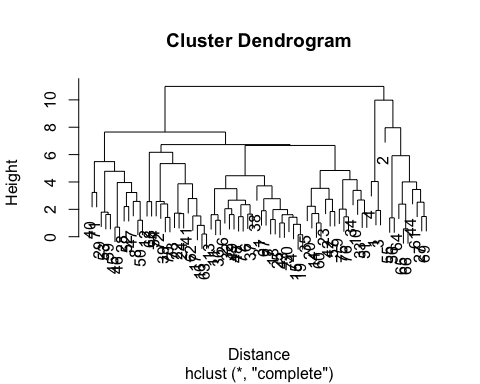
2022-11-20

library(knitr)  
library(stats)  
library(cluster)

cereals <- read.csv("~/Downloads/Cereals.csv")

numericaldata <- data.frame(cereals[,4:16])  
OmitMissing <- na.omit(numericaldata)  
Normalise <- scale(OmitMissing)

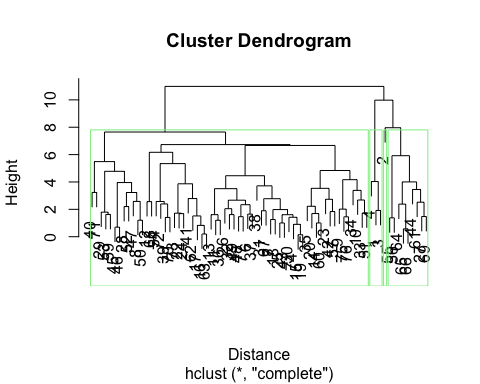
Distance <- dist(Normalise, method = "euclidean")  
  
Clustering\_heirarchial <- hclust(Distance, method = "complete")  
plot(Clustering\_heirarchial)



round(Clustering\_heirarchial$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(Clustering\_heirarchial)  
rect.hclust(Clustering\_heirarchial, k = 4, border = "lightgreen")



singleCH <- agnes(Normalise, method = "single")  
completeCH <- agnes(Normalise, method = "complete")  
averageCH <- agnes(Normalise, method = "average")  
wardCH <- agnes(Normalise, method = "ward")  
  
print(singleCH$ac)

## [1] 0.6067859

print(completeCH$ac)

## [1] 0.8353712

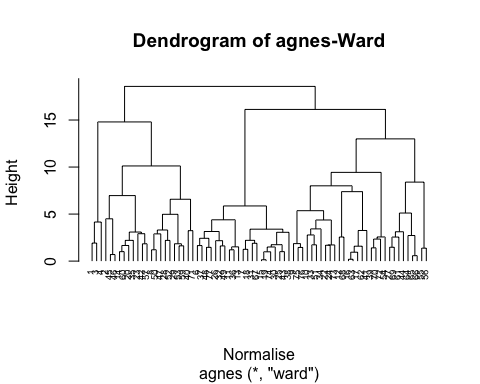
print(averageCH$ac)

## [1] 0.7766075

print(wardCH$ac)

## [1] 0.9046042

pltree(wardCH, cex = 0.6, hang = -1, main = 'Dendrogram of agnes-Ward')



cutree(wardCH, k = 4)

## [1] 1 2 1 1 3 3 2 4 4 3 4 3 2 3 4 4 3 3 2 4 2 4 3 3 4 2 2 3 3 3 4 4 2 3 3 3 4 2  
## [39] 4 2 3 4 2 2 2 3 3 2 4 2 2 4 4 4 2 2 2 4 4 4 4 4 4 3 4 4 4 2 4 4 3 4 4 3

As we can see from this ward is the best method to use.

HCuster1 <- hclust(Distance, method = "ward.D2")  
subgrp <- cutree(HCuster1, k = 4)  
table(subgrp)

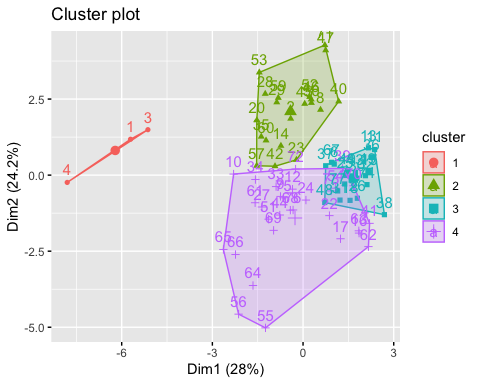
## subgrp  
## 1 2 3 4   
## 3 20 21 30

library("factoextra")

## Loading required package: ggplot2

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

cereals <- as.data.frame(cbind(Normalise, subgrp))  
  
fviz\_cluster(list(data = Normalise, cluster = subgrp))

 The optimal number of clusters is four.

Newcereals <- numericaldata  
Newcereals\_omit <- na.omit(Newcereals)  
Clust <- cbind(Newcereals\_omit, subgrp)  
Clust[Clust$subgrp==1,]

## calories protein fat sodium fiber carbo sugars potass vitamins shelf weight  
## 1 70 4 1 130 10 5 6 280 25 3 1  
## 3 70 4 1 260 9 7 5 320 25 3 1  
## 4 50 4 0 140 14 8 0 330 25 3 1  
## cups rating subgrp  
## 1 0.33 68.40297 1  
## 3 0.33 59.42551 1  
## 4 0.50 93.70491 1

mean(Clust[Clust$subgrp==1, "rating"])

## [1] 73.84446

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

## [1] 38.26161

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

## [1] 28.84825

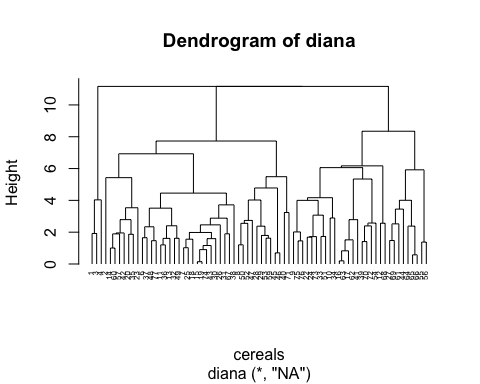
mean(Clust[Clust$subgrp==4, "rating"])

## [1] 51.43111

hc\_diana <- diana(cereals)  
  
hc\_diana$dc

## [1] 0.8386493

pltree(hc\_diana, cex = 0.5, hang = -1, main = 'Dendrogram of diana')



Newcereals\_omit$rating > 65

## [1] TRUE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [61] TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [73] FALSE FALSE

As we can see there are 5 Cereals that score and overall rating of above 65. These five cereals should be offered in schools. These cereals are 1, 4, 64, 65, and 66.