jstadden_5

Jared Stadden

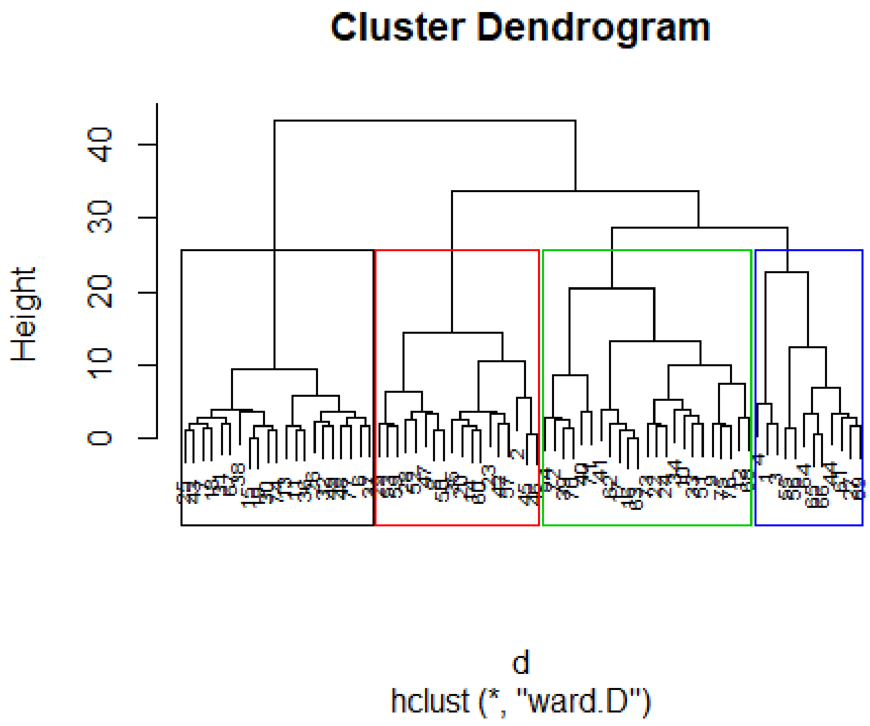
4/19/2020

R Markdown

```
#Read data into R
original = read.csv("C:\\Users\\jared\\Desktop\\Cereals.csv")

cereal <- na.omit(original)

#install.packages("stats")
#install.packages("cluster")
#install.packages("factoextra")

library(stats)
library(cluster)

## Warning: package 'cluster' was built under R version 3.6.3

library(ISLR)

#Normalize data
cereal[,4:16] <- scale(cereal[,4:16])
df <- cereal

#Applying various methods
hc_single <- agnes(df, method = "single")
hc_complete <- agnes(df, method = "complete")
hc_average <- agnes(df, method = "average")
hc_ward <- agnes(df, method = "ward")

#Choose method with output closest to 1
print(hc_single$ac)

## [1] 0.5192072

print(hc_complete$ac)

## [1] 0.9480521

print(hc_average$ac)

## [1] 0.9009936

print(hc_ward$ac)

## [1] 0.9812112
```

1. Ward method is best since its value is closest to 1.

```
#Applying Ward Heirarchical-Clustering
d <- dist(df, method = "euclidean")

## Warning in dist(df, method = "euclidean"): NAs introduced by coercion

hc_ward <- hclust(d, method = "ward")

## The "ward" method has been renamed to "ward.D"; note new "ward.D2"

#Plot Dendrogram
plot(hc_ward, cex = 0.6)

#Dendrogram with rectangles for each cluster
rect.hclust(hc_ward, k = 4, border = 1:4)
```



2. Based on the dendrogram I am going to use 4 clusters

```
#Assigning clusters
memb <- cutree(hc_ward, k=4)
df$cluster <- memb

#Splitting the data
set.seed(123)
Index_Train<- sample(seq_len(nrow(cereal)),size=floor(0.6*nrow(cereal)))
Train<-cereal[Index_Train,]
Validation<-cereal[-Index_Train,]


#Clustering
d2 <- dist(Train, method = "euclidean")
```

```
## Warning in dist(Train, method = "euclidean"): NAs introduced by coercion
```

```
hc_ward2 <- hclust(d2, method = "ward")
```

```
## The "ward" method has been renamed to "ward.D"; note new "ward.D2"
```

```
memb <- cutree(hc_ward2, k=4)
Train$cluster <- memb
head(Train)
```

```
##                      name mfr type   calories    protein       fat
sodium
## 33     Grape_Nuts_Flakes   P    C -0.3541153  0.4522084 0.0000000 -
0.27020566
## 53 Post_Nat._Raisin_Bran   P    C  0.6537514  0.4522084 0.0000000
0.45469653
## 15           Cocoa_Puffs   G    C  0.1498180 -1.4068705 0.0000000
0.21306247
## 70     Total_Corn_Flakes   G    C  0.1498180 -0.4773310 0.0000000
0.45469653
## 44                 Maypo   A    H -0.3541153  1.3817478 0.0000000 -
1.96164410
## 52  Oatmeal_Raisin_Crisp   G    C  1.1576848  0.4522084 0.9932203
0.09224544
##         fiber       carbo     sugars      potass   vitamins      shelf
## 33  0.3401532  0.06944832 -0.4836096 -0.19065695 -0.1818422  0.9419715
## 53  1.5780879 -0.95838683  1.5810314  2.27835060 -0.1818422  0.9419715
## 15 -0.8977815 -0.70142805  1.3516269 -0.61391539 -0.1818422 -0.2598542
## 70 -0.8977815  1.61120105 -0.9424187 -0.89608768  3.1822385  0.9419715
## 44 -0.8977815  0.32640711 -0.9424187 -0.04957081 -0.1818422 -0.2598542
## 52 -0.2788141 -0.31598986  0.6634132  0.30314456 -0.1818422  0.9419715
##        weight       cups     rating cluster
## 33 -0.2008324  0.2476647  0.6915569       1
## 53  1.9501886 -0.6432404 -0.3228791       2
## 15 -0.2008324  0.7567534 -1.3991551       3
## 70 -0.2008324  0.7567534 -0.2516826       4
## 44 -0.2008324  0.7567534  0.8892251       1
## 52  1.4287290 -1.3644493 -0.8494505       2
```

3. Couldn’t get centroids to work, so unable to complete part 3

4. The data should be normalized. When data is not normalized the distance calculations of the clustering method place extra importance on variables with larger values. When assessing the healthiness of cereal we wouldn’t want potassium to be overvalued compared to more important indicators such as vitamins, sugars, or fat. Normalization would ensure that clusters are created more fairly across measures of health with different units and scales but equal importance.