DeSimone_MS64060_Assignment 5

Heather DeSimone

4/9/2022

##First I have loaded in my data frame and removed the cereals that are missing information.

```
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(caret)

## Warning: package 'caret' was built under R version 4.1.2

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.1.2

## Loading required package: lattice

library(class)
library(ISLR)

## Warning: package 'ISLR' was built under R version 4.1.1

library(tidyverse)

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

## -- Attaching packages --------------------------------------- tidyverse
1.3.1 --

## v tibble  3.1.2     v purrr   0.3.4
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## Warning: package 'tidyr' was built under R version 4.1.2

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

```
## Warning: package 'stringr' was built under R version 4.1.2

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

## -- Conflicts ---------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()

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(stats)
DF=read.csv("C:/Users/hdesi/Desktop/MBA/Machine Learning/Cereals.csv")
DF <- na.omit(DF) ##Remove cereals missing data
DF$mfr<-NULL ##Not needed
DF$type<-NULL ##Not needed
rownames(DF) <- DF$name ##Change row name to cereal name rather than numeric
value
DF$name<-NULL
head(DF)

##                          calories protein fat sodium fiber carbo sugars
potass
## 100%_Bran                      70       4   1    130  10.0   5.0      6
280
## 100%_Natural_Bran             120       3   5     15   2.0   8.0      8
135
## All-Bran                       70       4   1    260   9.0   7.0      5
320
## All-Bran_with_Extra_Fiber      50       4   0    140  14.0   8.0      0
330
## Apple_Cinnamon_Cheerios       110       2   2    180   1.5  10.5     10
70
## Apple_Jacks                   110       2   0    125   1.0  11.0     14
30
##                          vitamins shelf weight cups   rating
## 100%_Bran                      25     3      1 0.33 68.40297
## 100%_Natural_Bran               0     3      1 1.00 33.98368
## All-Bran                       25     3      1 0.33 59.42551
## All-Bran_with_Extra_Fiber      25     3      1 0.50 93.70491
## Apple_Cinnamon_Cheerios        25     1      1 0.75 29.50954
## Apple_Jacks                    25     2      1 1.00 33.17409

sapply(DF, class) ##Making sure variables are numerical
```

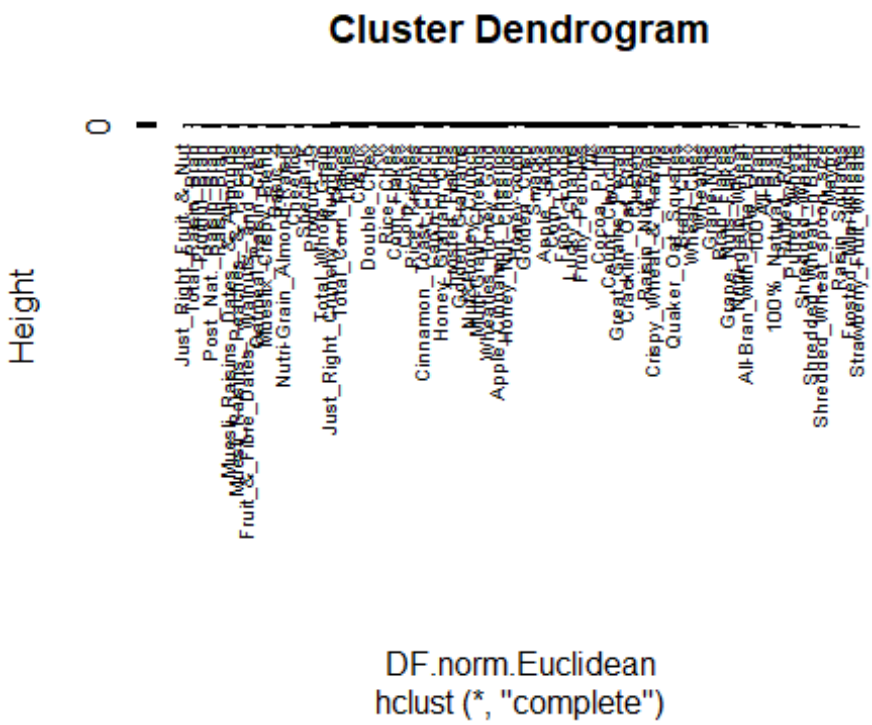
```
##  calories   protein       fat    sodium     fiber     carbo    sugars
potass
## "integer" "integer" "integer" "integer" "numeric" "numeric" "integer"
"integer"
##  vitamins     shelf    weight      cups    rating
## "integer" "integer" "numeric" "numeric" "numeric"
```

##Creating data frame for normalization

```
DF.norm <- data.frame(DF)

head(DF.norm)

##                           calories protein fat sodium fiber carbo sugars
potass
## 100%_Bran                       70       4   1    130  10.0   5.0      6
280
## 100%_Natural_Bran              120       3   5     15   2.0   8.0      8
135
## All-Bran                        70       4   1    260   9.0   7.0      5
320
## All-Bran_with_Extra_Fiber       50       4   0    140  14.0   8.0      0
330
## Apple_Cinnamon_Cheerios        110       2   2    180   1.5  10.5     10
70
## Apple_Jacks                    110       2   0    125   1.0  11.0     14
30
##                           vitamins shelf weight cups   rating
## 100%_Bran                       25     3      1 0.33 68.40297
## 100%_Natural_Bran                0     3      1 1.00 33.98368
## All-Bran                        25     3      1 0.33 59.42551
## All-Bran_with_Extra_Fiber       25     3      1 0.50 93.70491
## Apple_Cinnamon_Cheerios         25     1      1 0.75 29.50954
## Apple_Jacks                     25     2      1 1.00 33.17409
```

##I will perform hierarchical clustering using Euclidean Distance

```
DF.norm <- scale(DF) ##Data normalization
DF.norm.Euclidean <- dist(DF.norm, method = "euclidean")
hc1 <- hclust(DF.norm.Euclidean, method = "complete")

plot(hc1, cex = .6, hang = -1) ##Plotting the cluster Dendrogram using all
variables still in dataset
```

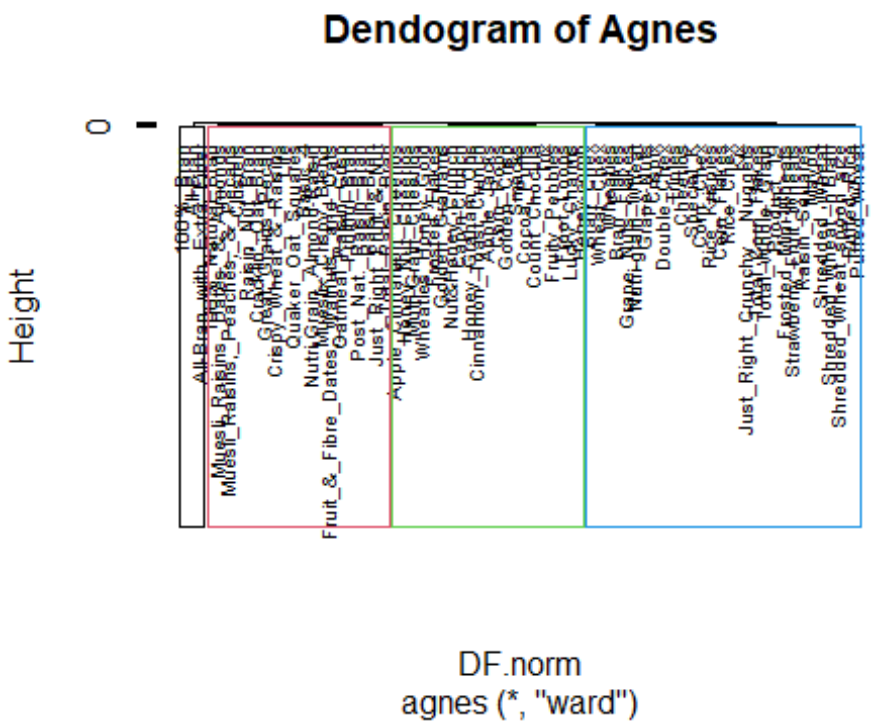


##I will now use Agnes to compare clustering methods to find the best one

```
library(cluster)
hc_single <- agnes(DF.norm, method = "single")
hc_complete <- agnes(DF.norm, method = "complete")
hc_average <- agnes(DF.norm, method = "average")
hc_ward <- agnes(DF.norm, method = "ward") ##Ward is the best 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) ##closest to 1

## [1] 0.9046042
```

##I will now create my Agnes Dendrogram

```
pltree(hc_ward, cex = 0.6, hang = -1, main = "Dendogram of Agnes")
rect.hclust(hc_ward, k = 4, border = 1:4) ##4 clusters
```



##Now I want to cluster my data by unhealthy variables. For our purposes, we will assume that cereals high in calories, fat, sugar, and sodium are unhealthy.

```
DF.Unhealthy <- DF[c(1,3,4,7)] ##Calories, fat, sodium, sugar
head(DF.Unhealthy)
```

```
##                              calories fat sodium sugars
## 100%_Bran                          70   1    130      6
## 100%_Natural_Bran                 120   5     15      8
## All-Bran                           70   1    260      5
## All-Bran_with_Extra_Fiber          50   0    140      0
## Apple_Cinnamon_Cheerios           110   2    180     10
## Apple_Jacks                       110   0    125     14
```

##Finding best Agnes method

```
unhealthy_single <- agnes(DF.Unhealthy, method = "single")
unhealthy_complete <- agnes(DF.Unhealthy, method = "complete")
unhealthy_average <- agnes(DF.Unhealthy, method = "average")
unhealthy_ward <- agnes(DF.Unhealthy, method = "ward") ##Best method

print(unhealthy_single$ac)

## [1] 0.7794119

print(unhealthy_complete$ac)

## [1] 0.967792
```

```
print(unhealthy_average$ac)
```

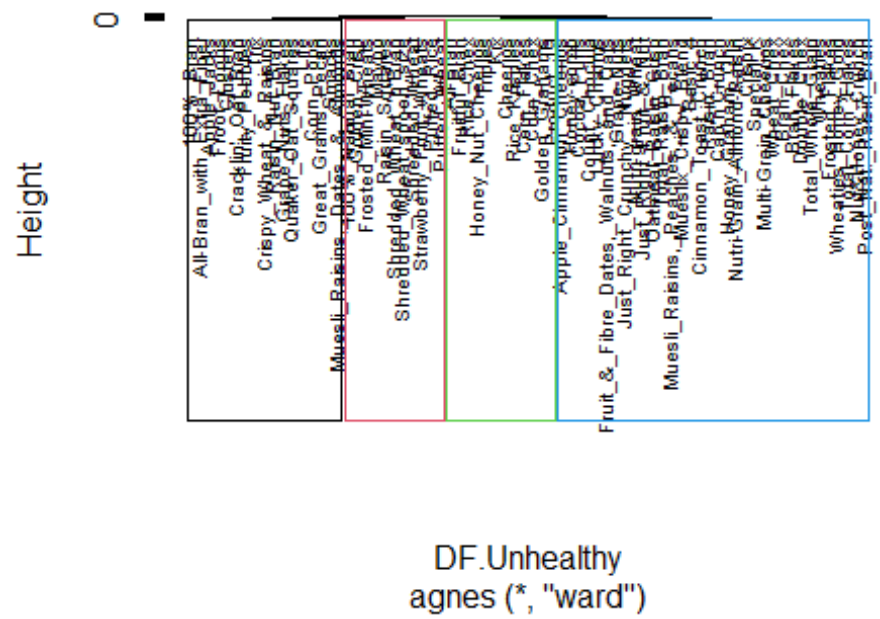
```
## [1] 0.9423144
```

```
print(unhealthy_ward$ac)
```

```
## [1] 0.9868955
```

##Ward was the best method for clustering. Now we will create our dendograph to look at our clusters for unhealthy variables.

```
pltree(unhealthy_ward, cex = 0.6, hang = -1, main = "Dendograph Using
Unhealthy Variables:Fat, Calories, Sugar & Sodium")
rect.hclust(unhealthy_ward, k = 4, border = 1:4)
```



##So far, it looks like the healthiest cluster is cluster 1(black) and the least healthy is cluster 4 (blue)

##Now we will cluster based on healthy variables. Those high in protein, fiber, and vitamins are most healthy.

```
DF.Healthy <- DF[c(2,5,9)] ##Protein, fiber, vitamins
head(DF.Healthy)
```

```
##                            protein fiber vitamins
## 100%_Bran                        4  10.0       25
## 100%_Natural_Bran                3   2.0        0
## All-Bran                         4   9.0       25
```

```
## All-Bran_with_Extra_Fiber       4  14.0       25
## Apple_Cinnamon_Cheerios         2   1.5       25
## Apple_Jacks                     2   1.0       25
```

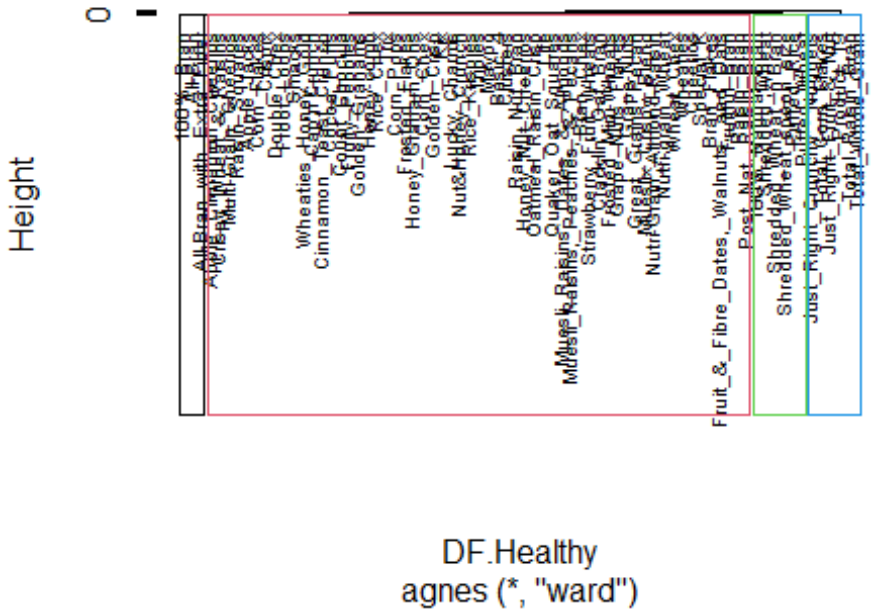
##Finding best Agnes method

```
healthy_single <- agnes(DF.Healthy, method = "single")
healthy_complete <- agnes(DF.Healthy, method = "complete")
healthy_average <- agnes(DF.Healthy, method = "average")
healthy_ward <- agnes(DF.Healthy, method = "ward")##Best method

print(healthy_single$ac)

## [1] 0.9950214

print(healthy_complete$ac)

## [1] 0.9957495

print(healthy_average$ac)

## [1] 0.9948298

print(healthy_ward$ac)

## [1] 0.9983455
```

##Ward was the best method for clustering. Now we will create our dendograph to look at our clusters for healthy variables.

```
pltree(healthy_ward, cex = 0.6, hang = -1, main = "Dendograph Using Healthy
Variables:Protein, Fiber & Vitamins")
rect.hclust(healthy_ward, k = 4, border = 1:4)
```

ndograph Using Healthy Variables:Protein, Fiber & V



##Clustering is a bit uneven, but it looks like cluster 1 (Black) is the healthiest and there are repeat cereals in this healthy cluster that were also in the healthy cluster is our last dendograph

##Now we will look at all of our health related variables together. Our cluster will consist of protein, fiber, vitamins, calories, fat, sugar, and sodium

```
DF.TotalHealth <- DF[c(1,2,3,4,5,7,9)]
head(DF.TotalHealth)
```

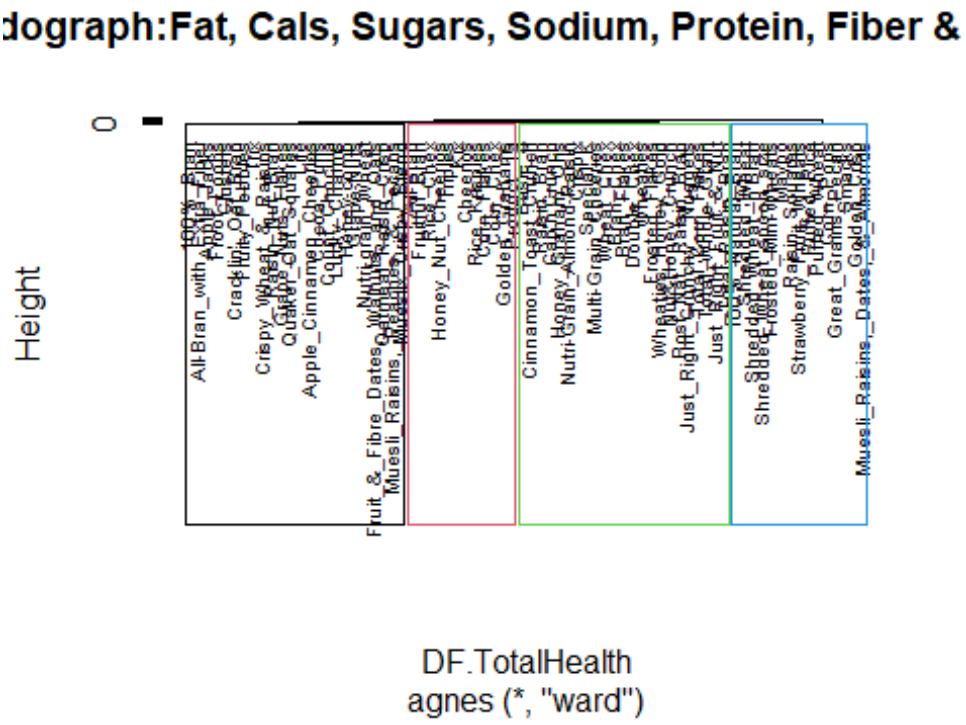
```
##                            calories protein fat sodium fiber sugars
vitamins
## 100%_Bran                        70       4   1    130  10.0       6
25
## 100%_Natural_Bran               120       3   5     15   2.0       8
0
## All-Bran                         70       4   1    260   9.0       5
25
## All-Bran_with_Extra_Fiber        50       4   0    140  14.0       0
25
## Apple_Cinnamon_Cheerios         110       2   2    180   1.5      10
25
## Apple_Jacks                     110       2   0    125   1.0      14
25
```

##Finding best Agnes

```
TotalHealth_single <- agnes(DF.TotalHealth, method = "single")
TotalHealth_complete <- agnes(DF.TotalHealth, method = "complete")
TotalHealth_average <- agnes(DF.TotalHealth, method = "average")
TotalHealth_ward <- agnes(DF.TotalHealth, method = "ward") ##Best Method

print(TotalHealth_single$ac)

## [1] 0.8615331

print(TotalHealth_complete$ac)

## [1] 0.9604174

print(TotalHealth_average$ac)

## [1] 0.9306789

print(TotalHealth_ward$ac)

## [1] 0.9837845
```

##Ward is again the best method. Now we will look at our dendograph in which we clustered based on all health variables (good and bad)

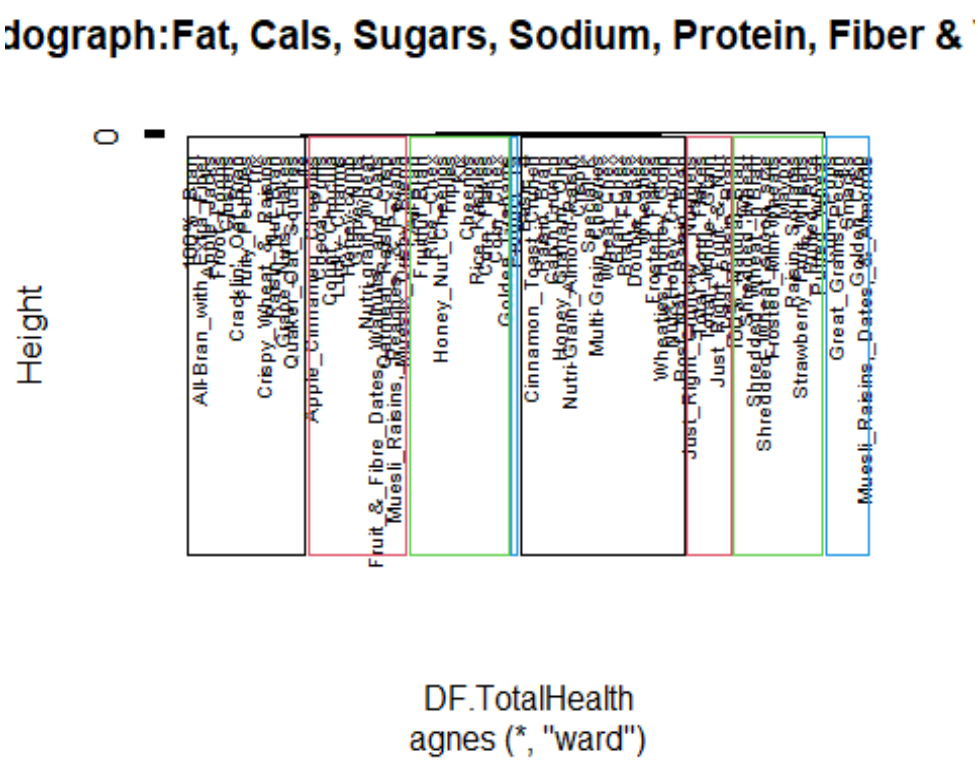
```
pltree(TotalHealth_ward, cex = 0.6, hang = -1, main = "Dendograph:Fat, Cals,
Sugars, Sodium, Protein, Fiber & Vitamins")
rect.hclust(TotalHealth_ward, k = 4, border = 1:4)
```



##This dendograph looks very similar to the 1st one we did (With unhealthy variables)
##Cluster 1 (black) looks to be the overall healthiest cereals

##I want to create more than 4 cluster using this same overall health model

```
pltree(TotalHealth_ward, cex = 0.6, hang = -1, main = "Dendograph:Fat, Cals,
Sugars, Sodium, Protein, Fiber & Vitamins")
rect.hclust(TotalHealth_ward, k = 8, border = 1:4) ##8 clusters
```



##This 8 cluster model gives us a much more condensed list. With 100% bran in all of our dendograph healthy clusters, we will assume that the cluster this cereal falls into is the healthiest group - based on sugar, fat, calories, sodium, protein, fiber, and vitamins. The school could use any of these dendographs to base their decision on, depending on what they are looking for and what they consider healthy. Some people feel that a diet low in calories, fat, sugar, and sodium is healthy even if those foods are low in nutrients. Some people feel that a diet high in vitamins, fiber, and protein are healthy even if they have higher calories, fat, sugar. and sodium. I believe that the last dendograph should be used (8 clusters) as it takes all of these variables into consideration.