Assignment_5

Kevin Gardner

Due 4/17/2022

———————————————————

Following is the link to my GitHub account:

https://github.com/Kgardner22/64060_-kgardner

———————————————————

IMPORT THE DATA:

```
cereals.df <- read.csv('C:/R/MyData/Cereals.csv', header = T, sep = ',')

summary(cereals.df)

##      name               mfr                type              calories
##  Length:77          Length:77          Length:77          Min.   : 50.0
##  Class :character   Class :character   Class :character   1st Qu.:100.0
##  Mode  :character   Mode  :character   Mode  :character   Median :110.0
##                                                           Mean   :106.9
##                                                           3rd Qu.:110.0
##                                                           Max.   :160.0
##
##     protein           fat            sodium          fiber
##  Min.   :1.000   Min.   :0.000   Min.   :  0.0   Min.   : 0.000
##  1st Qu.:2.000   1st Qu.:0.000   1st Qu.:130.0   1st Qu.: 1.000
##  Median :3.000   Median :1.000   Median :180.0   Median : 2.000
##  Mean   :2.545   Mean   :1.013   Mean   :159.7   Mean   : 2.152
##  3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:210.0   3rd Qu.: 3.000
##  Max.   :6.000   Max.   :5.000   Max.   :320.0   Max.   :14.000
##
##      carbo          sugars           potass          vitamins
##  Min.   : 5.0   Min.   : 0.000   Min.   : 15.00   Min.   :  0.00
##  1st Qu.:12.0   1st Qu.: 3.000   1st Qu.: 42.50   1st Qu.: 25.00
##  Median :14.5   Median : 7.000   Median : 90.00   Median : 25.00
##  Mean   :14.8   Mean   : 7.026   Mean   : 98.67   Mean   : 28.25
##  3rd Qu.:17.0   3rd Qu.:11.000   3rd Qu.:120.00   3rd Qu.: 25.00
##  Max.   :23.0   Max.   :15.000   Max.   :330.00   Max.   :100.00
##  NA's   :1      NA's   :1        NA's   :2
##      shelf           weight          cups           rating
##  Min.   :1.000   Min.   :0.50   Min.   :0.250   Min.   :18.04
##  1st Qu.:1.000   1st Qu.:1.00   1st Qu.:0.670   1st Qu.:33.17
##  Median :2.000   Median :1.00   Median :0.750   Median :40.40
##  Mean   :2.208   Mean   :1.03   Mean   :0.821   Mean   :42.67
##  3rd Qu.:3.000   3rd Qu.:1.00   3rd Qu.:1.000   3rd Qu.:50.83
##  Max.   :3.000   Max.   :1.50   Max.   :1.500   Max.   :93.70
##
```

REMOVE ALL CEREALS WITH MISSING VALUES:

```
cereals.df <- na.omit(cereals.df) #Remove NA (missing) values
```

SET ROW NAMES IN THE DATAFRAME:

```
# set row names to the name column
row.names(cereals.df) <- cereals.df[,1]

# remove the name column as a variable
cereals.df <- cereals.df[,-1]
```

NORMALIZE THE DATA:

```
# normalize all numeric variables (Columns 3 - 15)
cereals.df.norm <- sapply(cereals.df[,c(3:15)], scale)

# add row names: cereals
row.names(cereals.df.norm) <- row.names(cereals.df)
```

APPLY HIERARCHICAL CLUSTERING USING AGNES AND FOUR LINKAGE MEASURES:

Apply hierarchical clustering to the data using Euclidean distance to the normalized measurements.

```
# load the cluster package so we can use agnes
library(cluster)

# compute normalized distance based on all numeric variables
d.norm <- dist(cereals.df.norm, method = "euclidean")

hc_single <- agnes(d.norm, method = "single") # uses single linkage
hc_complete <- agnes(d.norm, method = "complete") #uses complete linkage
hc_average <- agnes(d.norm, method = "average") #uses average linkage
hc_ward <- agnes(d.norm, method = "ward") #uses Ward's method
```

COMPARE AGGLOMERATIVE COEFFICIENTS

```
print(hc_single$ac)

## [1] 0.6067859
```

```r
print(hc_complete$ac)

## [1] 0.8353712

print(hc_average$ac)

## [1] 0.7766075

print(hc_ward$ac)

## [1] 0.9046042
```

COMPARE THE RESULTS:

Compare the results of the Agglomerative coefficients (AC) of the four methods.

Single Linkage (hc_single): AC = 0.6067859

Complete Linkage (hc_complete): AC = 0.8353712

Average Linkage (hc_average): AC = 0.7766075

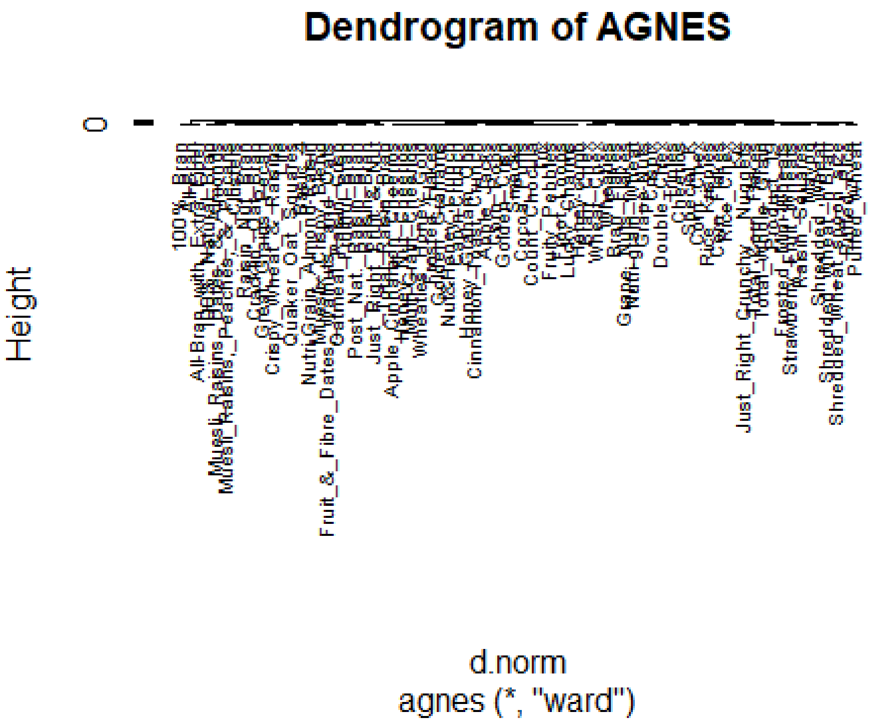
Ward's Method (hc_ward): AC = 0.9046042

In comparing the Agglomerative coefficients (AC), we see Ward's Method is the best linkage method as it results in the highest Agglomeritve coefficient value.

PLOT THE DENDROGRAM:

To help answer the question of "how many clusters would you choose", let's first review the dendrogram using Ward's Method for hierarchical clustering since this proved to be the best linkage method.

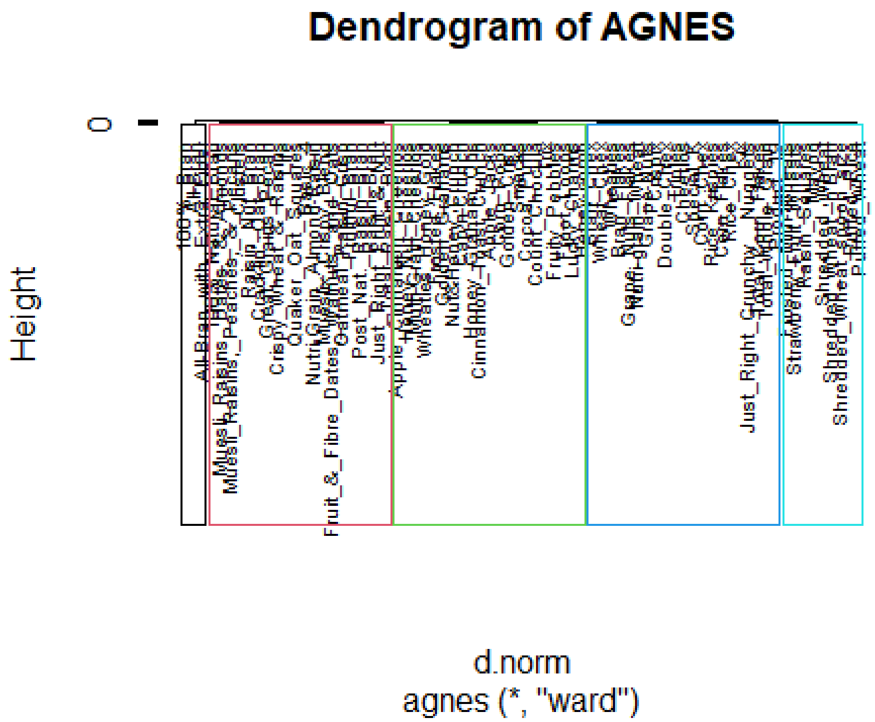
```r
pltree(hc_ward, cex = 0.6, hang = -1, main = "Dendrogram of AGNES")
```



In reviewing the dendrogram, the cereals appear to fall logically into five (5) clusters by setting a cutoff of 12.

Let’s now visualize the five (5) clusters:

```
pltree(hc_ward, cex = 0.6, hang = -1, main = "Dendrogram of AGNES")
rect.hclust(hc_ward, k=5, border = 1:5)
```



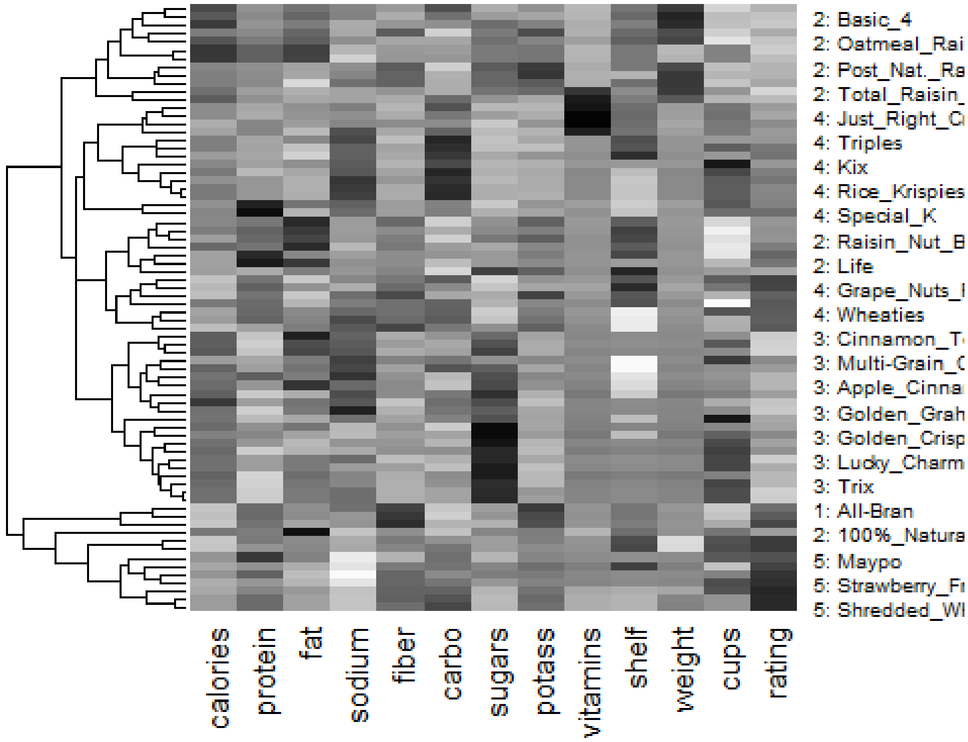
COMPUTE CLUSTER MEMBERSHIP BY "CUTTING" THE DENDROGRAM

```
memb <- cutree(hc_ward, k = 5)
memb

##  [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 5 2 2 3 3 3 4 4 2 3 3
3 4 2
## [39] 4 2 3 5 2 2 2 3 3 2 4 2 2 4 5 5 2 2 2 5 4 4 5 5 5 3 4 5 4 2 4 4 3 4 4
3
```

CREATING A HEATMAP

The following will create a heatmap for the 74 cereals (in rows). The rows are sorted by the five clusters from the Ward linkage clustering. Darker cells denote higher values within a column

```
# set labels as cluster membership and cereal name
row.names(cereals.df.norm) <- paste(memb, ": ", row.names(cereals.df), sep =
"")

# plot heatmap
# rev() reverses the color mapping to large = dark
heatmap(as.matrix(cereals.df.norm), Colv = NA,
        col=rev(paste("gray",1:99,sep="")))
```



INTERPRETING THE DATA:

In reviewing the 13 variables for all 74 cereals by cluster, we can identify the following cluster commonalities:

Cluster 1: High in Fiber and Potassium with good customer satisfaction rating

Cluster 2: High in Calories and Fat (and lowest in Vitamins)

Cluster 3: High in Sugars and Calories

Cluster 4: High in Carbohydrates ("Carbs") and slightly higher in Sodium

Cluster 5: Higher in Protein, Fiber, Carbs, and Potassium and highest in Customer Satisfaction Rating

The elementary public schools would like to identify a group/cluster of "healthy cereals" to include in their daily cafeterias. To make this recommendation, we first need to understand which factors are desirable to consider a cereal "healthy".

A google search reveals "healthy cereals" should include:

```
 * Whole grains

 * High in Fiber, protein and nutrients/vitamins

 * Carbs are a main source of energy and help fuel our brains and vital
organs
```

Based on our above interpretation of the five (5) clusters, the healthiest group of cereals is:

```
 Cluster 5:
 This cluster has higher levels of protein, fiber, carbs, and potassium while
having the highest customer satisfaction ratings. There are nine cereals in
this grouping providing a good selection of "healthy cereals"for elementary
children.
```