Assignment 5 Nikesh Sapkota

2022-04-17

Hierarchical Clustering using Cereals dataset

```
#Importing dataset

library(readr)
library(tidyverse)
```

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

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

## v ggplot2 3.3.5     v dplyr   1.0.8
## v tibble  3.1.6     v stringr 1.4.0
## v tidyr   1.2.0     v forcats 0.5.1
## v purrr   0.3.4

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

## 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()
```

```
library(cluster)
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

```
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
```

```
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(RColorBrewer)

Cereals <- read_csv("Cereals.csv")
```

```
## Rows: 77 Columns: 16
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): name, mfr, type
## dbl (13): calories, protein, fat, sodium, fiber, carbo, sugars, potass, vita...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
View(Cereals)
```

```
summary(Cereals)
```

```
##      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  
##                                                                 
```

The dataset Cereals.csv includes nutritional information, store display, and consumer ratings for 77 breakfast cereals.

Data Preprocessing. Remove all cereals with missing values.

```
colSums(is.na(Cereals))
```

```
##     name      mfr     type calories  protein      fat   sodium    fiber 
##        0        0        0        0        0        0        0        0 
##    carbo   sugars   potass vitamins    shelf   weight     cups   rating 
##        1        1        2        0        0        0        0        0
```

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

```
Cereals_num <- Cereals_data %>% select_if(is.numeric)

head(Cereals_num)
```

```
## # A tibble: 6 x 13
##   calories protein   fat sodium fiber carbo sugars potass vitamins shelf weight
##      <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>  <dbl>  <dbl>    <dbl> <dbl>  <dbl>
## 1       70       4     1    130    10     5        6    280       25     3      1
## 2      120       3     5     15     2     8        8    135        0     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
## 5      110       2     2    180   1.5  10.5     10     70       25     1      1
## 6      110       2     0    125   1    11       14     30       25     2      1
## # ... with 2 more variables: cups <dbl>, rating <dbl>
```

```
scaled_cereals <- as.data.frame(scale(Cereals_num))
```

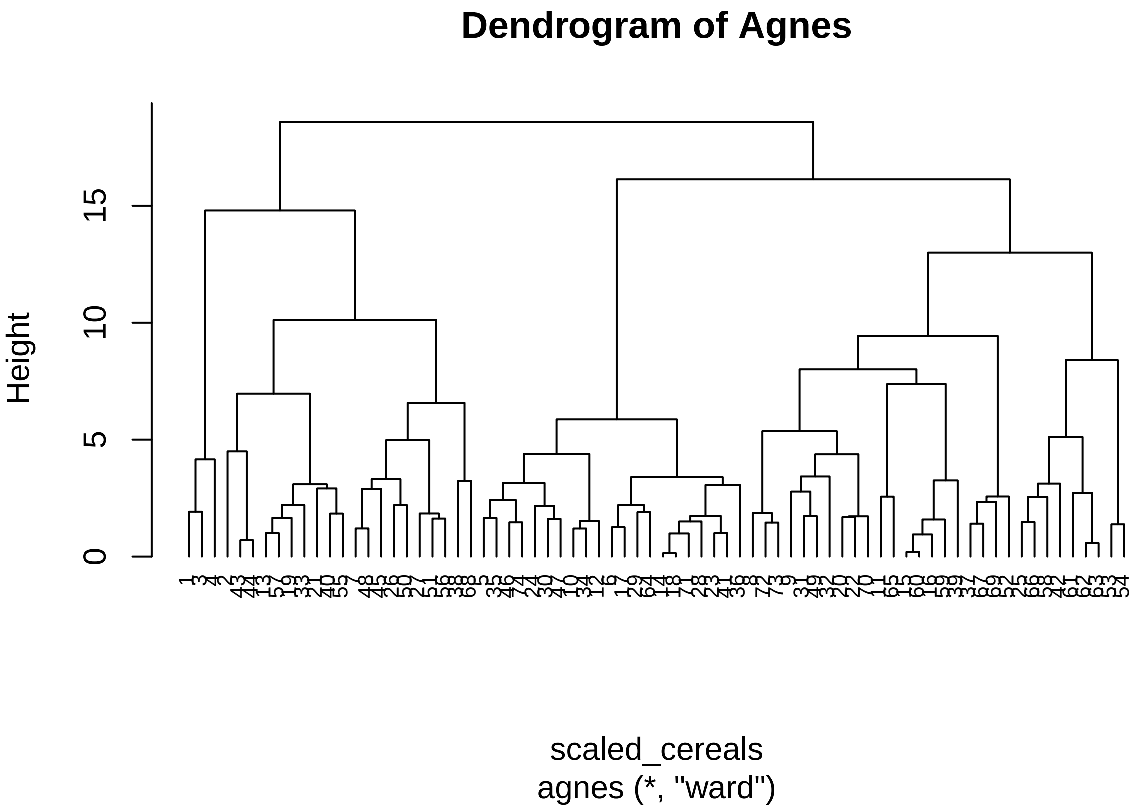
1. Apply hierarchical clustering to the data using Euclidean distance to the normalized measurements. Use Agnes to compare the clustering from single linkage, complete linkage, average linkage, and Ward. Choose the best method.

```
methods <- c( "average", "single", "complete", "ward")

names(methods) <- c( "average", "single", "complete", "ward")

linkage<- function(x) { agnes(scaled_cereals, metric = "euclidean",
                                     method = x)$ac}

map_dbl(methods , linkage)
```

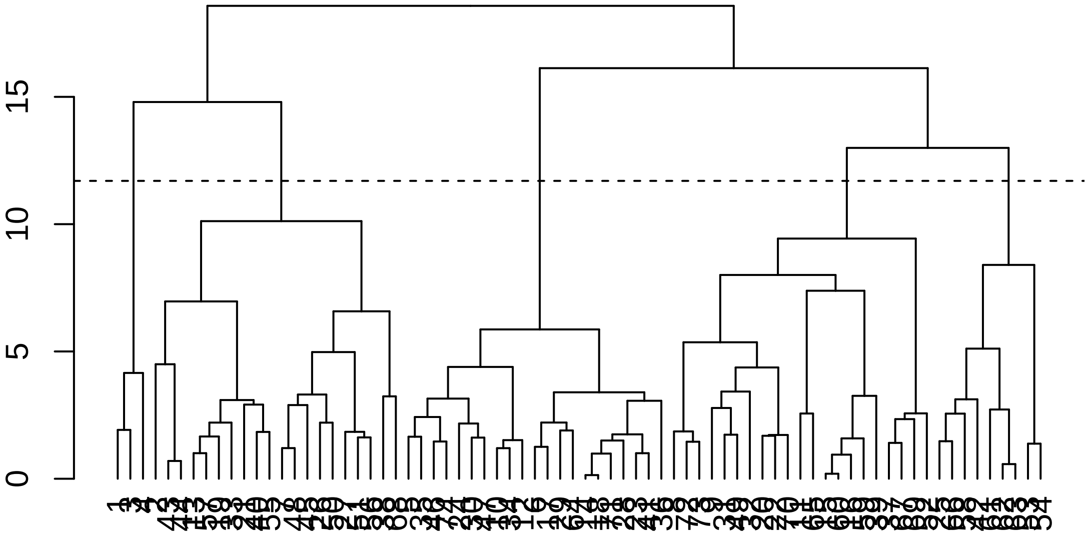
```
##   average    single  complete      ward
## 0.7766075 0.6067859 0.8353712 0.9046042
```

2. How many clusters would you choose?

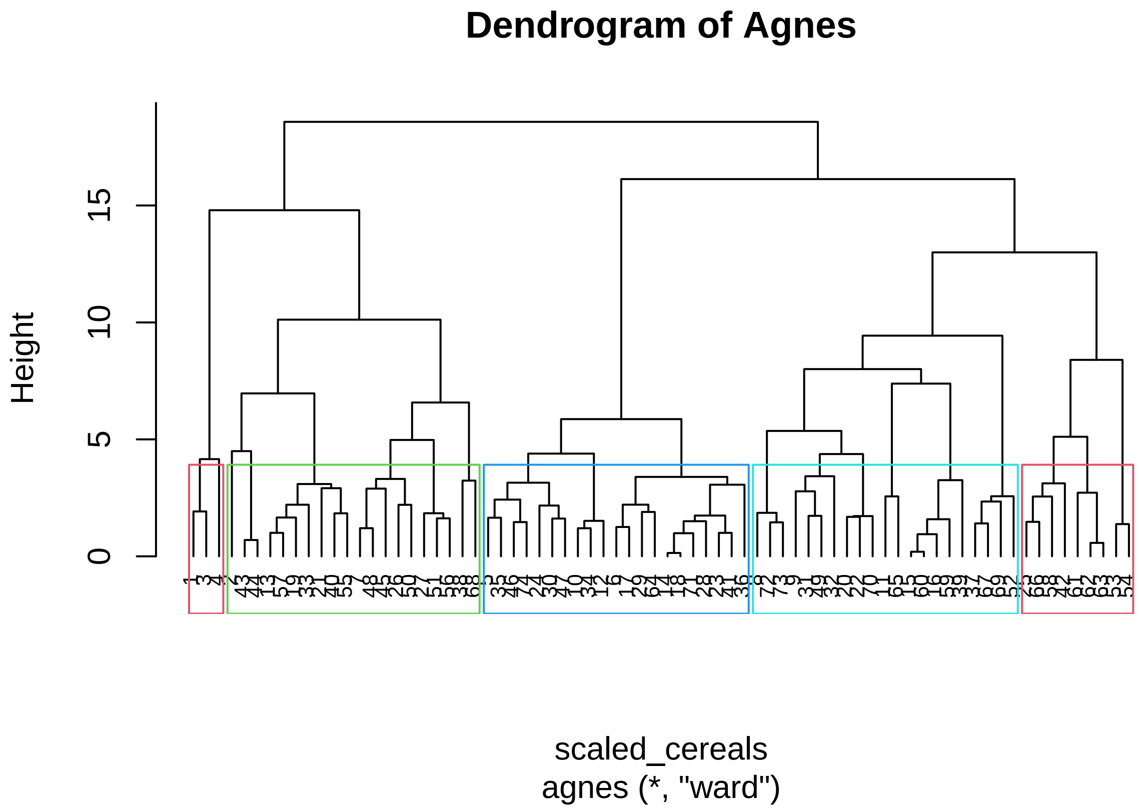
```
Hierarchical_cereals <- agnes(scaled_cereals, method = "ward")

pltree(Hierarchical_cereals, cex = 0.7, hang = -1, main = "Dendrogram of Agnes")
```



```
plot(as.dendrogram(Hierarchical_cereals))
abline(h = 11.7, lty = 2)
```



```
pltree(Hierarchical_cereals, cex = 0.7, hang = -1, main = "Dendrogram of Agnes")

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



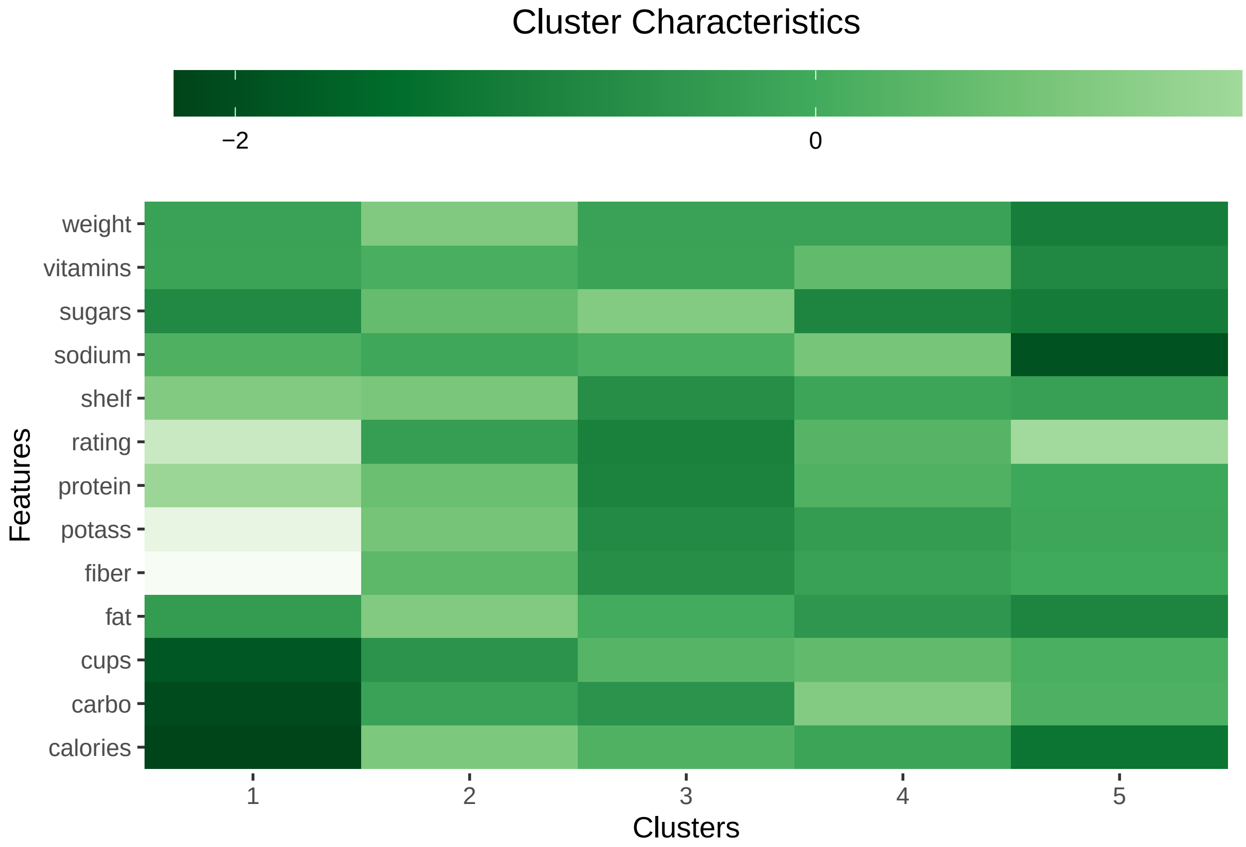
The optimal number of clusters is determined using hierarchical clustering. The optimal number of clusters can be determined by examining the greatest height difference. As a result of the above analysis, the optimal number of clusters is “k = 5”.

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.”

```
cluster_assignment <- cutree(Hierarchical_cereals, k=5)
cereals_clustered <- mutate(scaled_cereals, cluster = cluster_assignment)

split_data <- split(cereals_clustered, cereals_clustered$cluster)
split_means <- lapply(split_data, colMeans)
(centroids <- do.call(rbind, split_means))
```

```
##     calories     protein        fat      sodium       fiber      carbo
## 1 -2.2018711  1.38174776 -0.3310734  0.17279012  3.64131237 -2.0718749
## 2  0.8553248  0.59163927  0.9435592 -0.08898011  0.38141771 -0.2003584
## 3  0.1978117 -0.91996886  0.0000000  0.12101140 -0.66198437 -0.5423583
## 4 -0.1621407  0.18662567 -0.4729620  0.77112209 -0.21003997  0.9626860
## 5 -1.2499969 -0.06420242 -0.8828625 -1.94150793 -0.02664224  0.1551013
##       sugars     potass    vitamins      shelf     weight       cups     rating
## 1 -0.7894824  2.9837813 -0.18184220  0.9419715 -0.2008324 -1.8452553  2.2426479
## 2  0.5143002  0.7475659  0.09849786  0.8217889  0.9235649 -0.5477863 -0.2928786
## 3  0.9583619 -0.7415648 -0.18184220 -0.6604628 -0.2008324  0.2779676 -0.9636465
## 4 -0.8659505 -0.3485391  0.45893508 -0.1453946 -0.2008324  0.4577648  0.2916795
## 5 -1.0953551 -0.1122758 -0.80482011 -0.2598542 -1.0482044  0.1156788  1.4712151
##   cluster
## 1       1
## 2       2
## 3       3
## 4       4
## 5       5
```

```
Hierarchical_palette <-
  colorRampPalette(rev(brewer.pal(9, 'Greens')), space = 'Lab')
data.frame(centroids) %>% gather("features", "values",-cluster) %>%
  ggplot(aes(
    x = factor(cluster),
    y = features,
    fill = values
  )) +
  geom_tile() + theme_classic() +
  theme(
    axis.line = element_blank(),
    legend.position = "top",
    legend.justification = "left",
    plot.title = element_text(hjust = 0.5),
    legend.title = element_blank(),
    legend.key.width = unit(4.5, "cm")
  ) +
  scale_x_discrete(expand = c(0, 0)) +
  scale_fill_gradientn(colours = Hierarchical_palette(100)) +
  labs(title = "Cluster Characteristics",
       x = "Clusters",
       y = "Features",
       fill = "Centroids")
```



From the graph above, we can see that each cluster pattern is distinct. The analysis for each of the five clusters may be seen below.

- Cluster 1 (Bran Cereals): Cereals in Cluster 1 are “rich in vitamins, protein, potassium, fibers, and moderate vitamins,” have “few carbohydrates, sugar, and calories,” and have a “high rating and good shelf life,” among other things.
- Cluster2 (Hot Cereals): Cereals in Cluster 2 include “excellent vitamins, protein, potassium, fibre, and calories,” but “high sugar, fat, and weight.”
- Cluster3 (Sugary Cereals): Cereals in Cluster 3 are “heavy in sugar, sodium, carbohydrate, and fat,” as well as “poor in vitamins, protein, potassium, and fiber” when compared to other clusters.
- Cluster4 (Organic Cereals): Cereals are “rich in all components,” but also “heavy in sodium and carbs” when compared to other clusters.
- Cluster5 (Whole Grain Cereals): Cereals in Cluster5 are “low in sodium and sugars” in comparison to other clusters.

Conclus

There are certain grains that are superior to others. Only a few cereals are promoted exclusively for children, and some of them contain up to 50% sugar. The packaging of these products can also be deceitful because it emphasizes only the positive aspects of the product, such as added fiber or essential vitamins. Healthy cereals, on the other hand, are free of sugar and come in a variety of colors and shapes. According to studies, less sugar, salt, and fiber are all beneficial to both children and adults. We may deduce that cluster1 is favorable to children based on the prior cluster analysis and data. As a result, this can be suggested for daily lunches in elementary schools. The data must also be standardized such that each variable has the same scale. The model may be skewed toward the variables with greater magnitudes if the variables’ scales aren’t the same.