Hierarchical clustering

rajendra

12/11/2019

```
library(ISLR)
library(readr)
library(tidyverse)
```

```
## -- Attaching packages ------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v ggplot2 3.2.1     v forcats 0.4.0
```

```
## -- Conflicts ---------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(cluster)
library(factoextra)
```

```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at
https://goo.gl/13EFCZ
```

```
library(dendextend)
```

```
##
## ---------------------
## Welcome to dendextend version 1.12.0
## 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
## Or contact: <tal.galili@gmail.com>
##
##  To suppress this message use:
suppressPackageStartupMessages(library(dendextend))
## ---------------------
```

```
##
## Attaching package: 'dendextend'
## The following object is masked from 'package:stats':
##
##     cutree
```

```r
library(fpc)
```

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

```
## Parsed with column specification:
## cols(
##   name = col_character(),
##   mfr = col_character(),
##   type = col_character(),
##   calories = col_double(),
##   protein = col_double(),
##   fat = col_double(),
##   sodium = col_double(),
##   fiber = col_double(),
##   carbo = col_double(),
##   sugars = col_double(),
##   potass = col_double(),
##   vitamins = col_double(),
##   shelf = col_double(),
##   weight = col_double(),
##   cups = col_double(),
##   rating = col_double()
## )
```

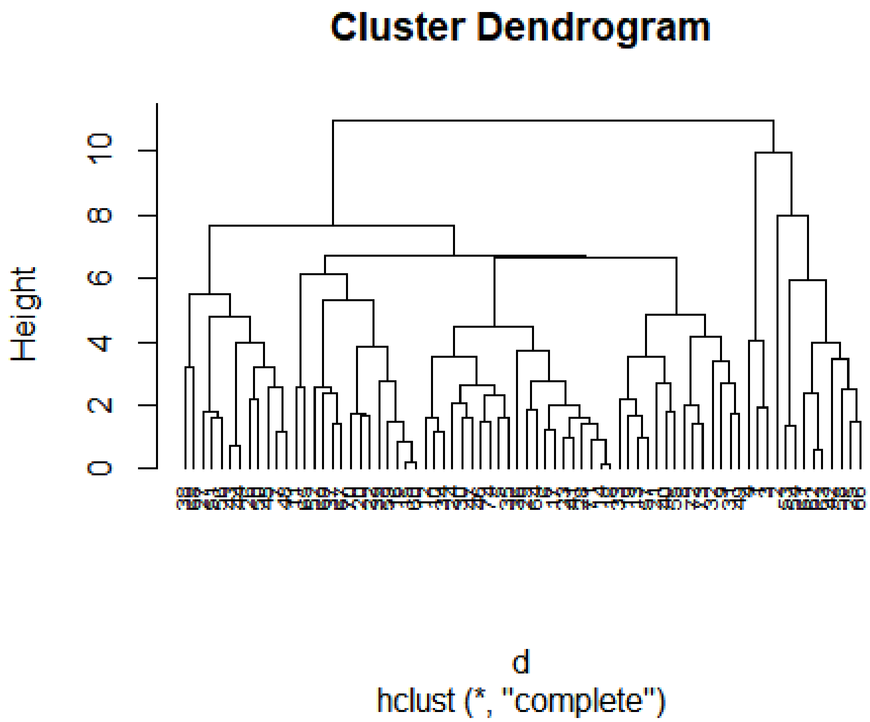
```r
View(cereals)
set.seed(123)
summary(cereals)
```

```
##      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
##
```

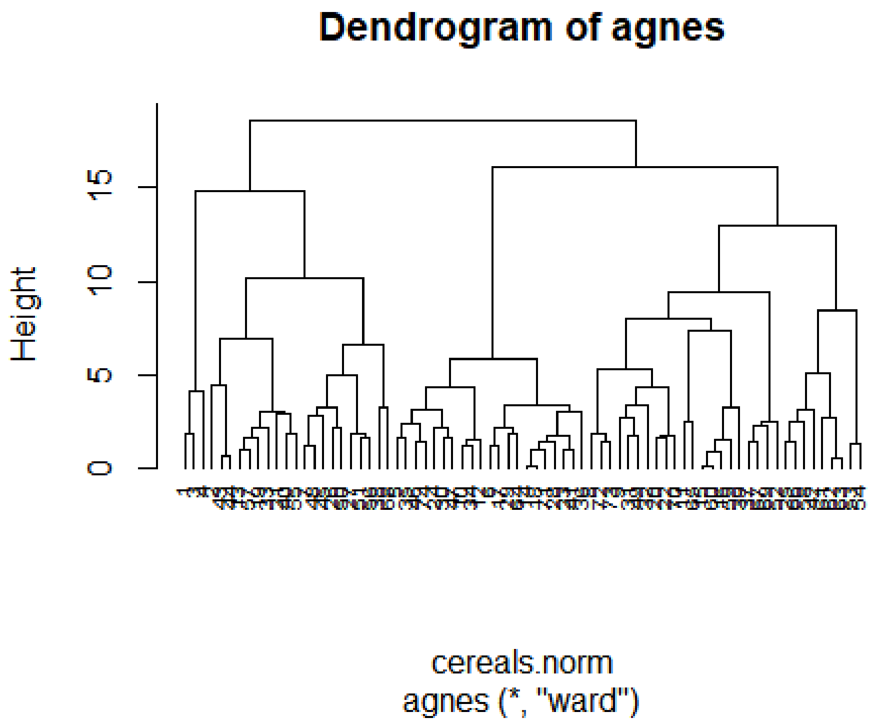
```
cereals.norm <- cereals[,-c(1:3)]#normaliizing the dataset
cereals.norm <- na.omit(cereals.norm)#Ommiitting na values
cereals.norm <- scale(cereals.norm)
str(cereals.norm)
```

```
##  num [1:74, 1:13] -1.866 0.654 -1.866 -2.874 0.15 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : NULL
##   ..$ : chr [1:13] "calories" "protein" "fat" "sodium" ...
##  - attr(*, "scaled:center")= Named num [1:13] 107.03 2.51 1 162.36 2.18
...
##   ..- attr(*, "names")= chr [1:13] "calories" "protein" "fat" "sodium" ...
##  - attr(*, "scaled:scale")= Named num [1:13] 19.84 1.08 1.01 82.77 2.42
...
##   ..- attr(*, "names")= chr [1:13] "calories" "protein" "fat" "sodium" ...
```

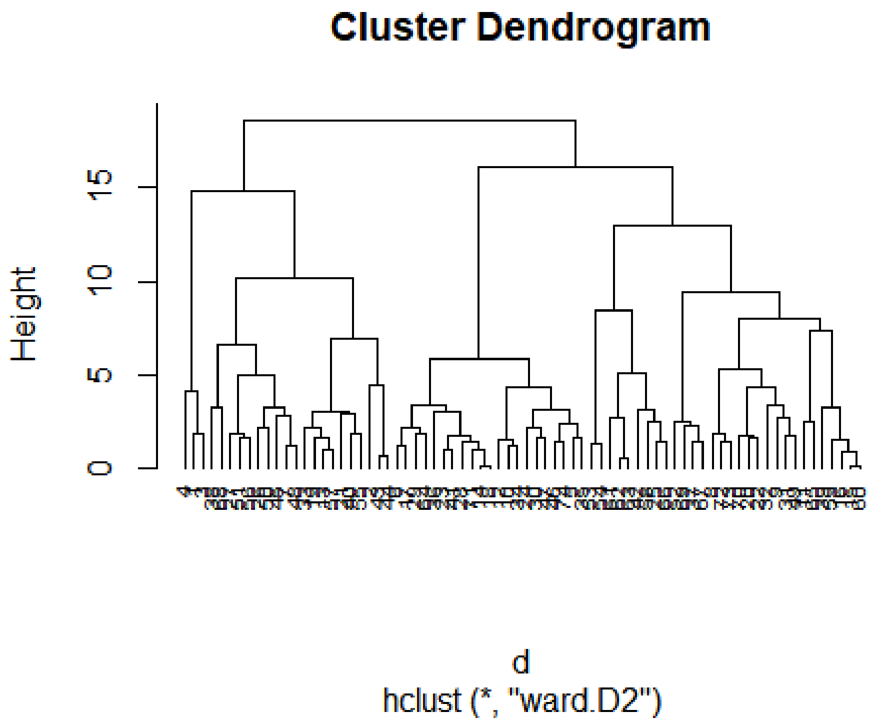
```
# Dissimilarity matrix
d <- dist(cereals.norm, method = "euclidean")

# Hierarchical clustering using Complete Linkage
hc1 <- hclust(d, method = "complete" )

# Plot the obtained dendrogram
plot(hc1, cex = 0.6, hang = -1)
```



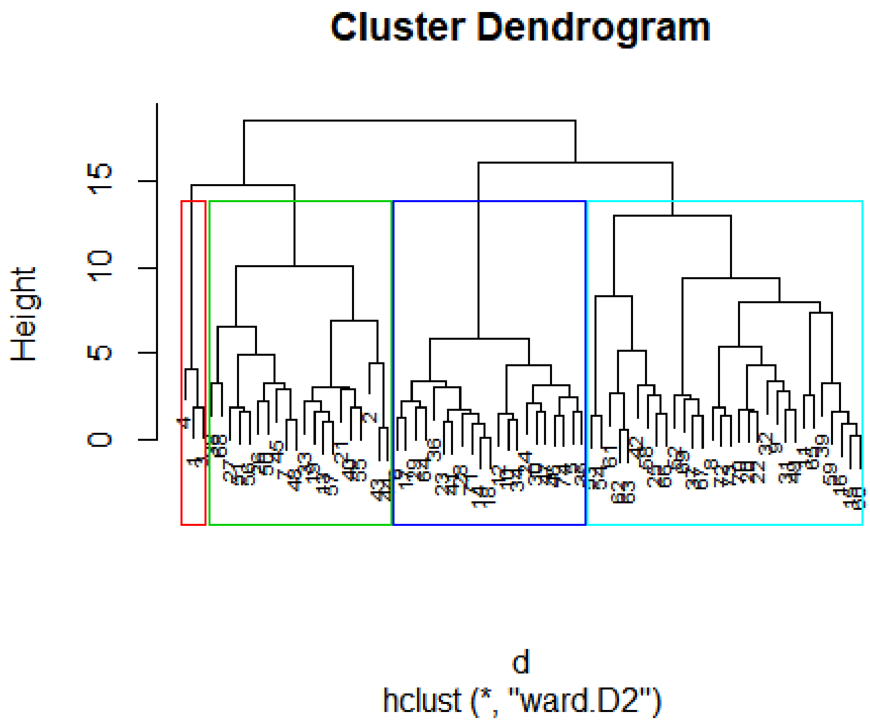
```r
# Dissimilarity matrix
d <- dist(cereals.norm, method = "euclidean")

# Compute with agnes and with different linkage methods
hc_single <- agnes(cereals.norm, method = "single")
hc_complete <- agnes(cereals.norm, method = "complete")
hc_average <- agnes(cereals.norm, method = "average")
hc_ward <- agnes(cereals.norm, method = "ward")

# Compare Agglomerative coefficients
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)

## [1] 0.9046042

hc2 <- agnes(cereals.norm, method = "ward")
pltree(hc2, cex = 0.6, hang = -1, main = "Dendrogram of agnes")
```



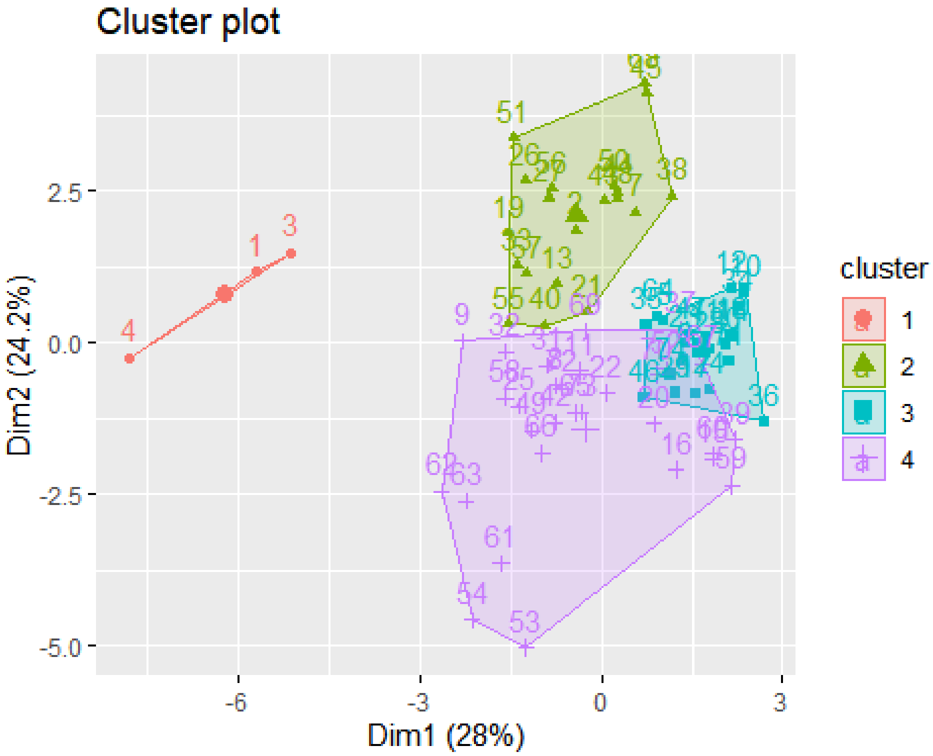
```
d <- dist(cereals.norm, method = "euclidean")

# Hierarchical clustering using Ward Linkage
hc3 <- hclust(d, method = "ward.D2" )

# Plot the obtained dendrogram
plot(hc3, cex = 0.6, hang = -1)
```



```
#From the dendogram,when we cut the longest length we are obtaining the
optimal number of clusters as 6

hcluster <- cutree(hc3, k = 4)
plot(hc3, cex = 0.6)
rect.hclust(hc3, k = 4, border = 2:5)
```



```
fviz_cluster(list(data = cereals.norm, cluster = hcluster))
```



```
#cluster stabilities of all 4 clusters
hclust_stability <- clusterboot(cereals.norm, clustermethod=hclustCBI,
method="ward.D2", k=4, count = FALSE)
hclust_stability
```

```
## * Cluster stability assessment *
## Cluster method:  hclust/cutree
## Full clustering results are given as parameter result
## of the clusterboot object, which also provides further statistics
## of the resampling results.
## Number of resampling runs:  100
##
## Number of clusters found in data:  4
##
##  Clusterwise Jaccard bootstrap (omitting multiple points) mean:
## [1] 0.5651665 0.7875223 0.8663548 0.6777744
## dissolved:
## [1] 49  7  5 27
## recovered:
## [1] 51 61 79 43
```

```
#Analyze the clustering results
clusters <- hclust_stability$result$partition

#Cluster stability values
hclust_stability$bootmean
```

```
## [1] 0.5651665 0.7875223 0.8663548 0.6777744
```

```
library(caret)
```

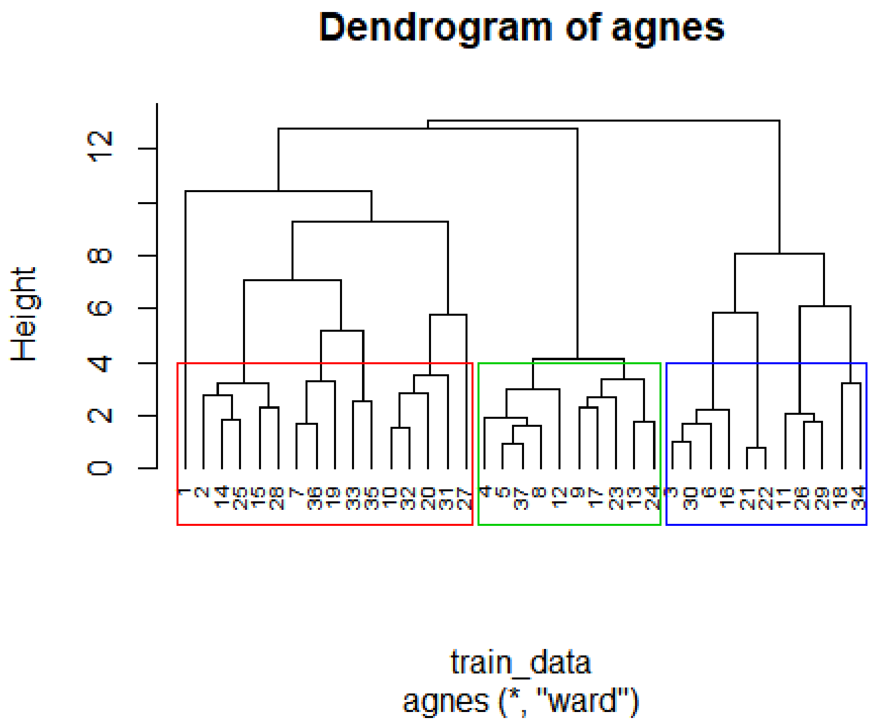
```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

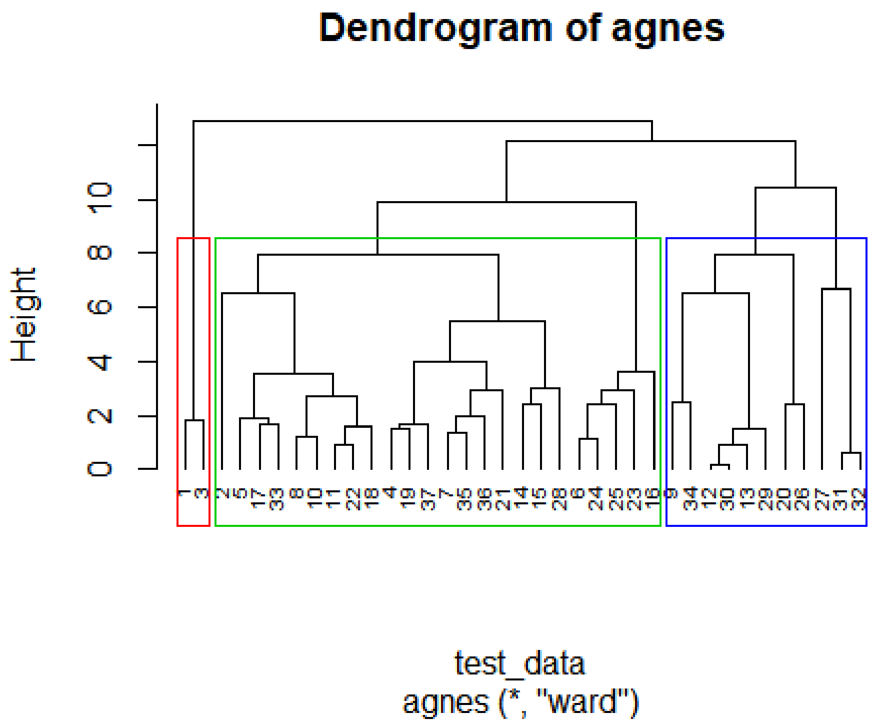
```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
set.seed(123)
C<-cereals
C1<-na.omit(C)

C_index<-createDataPartition(C1$calories,p=0.5,list=FALSE)
train_data<-C1[C_index,]
test_data<-C1[-C_index,]
train_data<-scale(train_data[,-c(1:3)])
test_data<-scale(test_data[,-c(1:3)])

# Compute with agnes and with different linkage methods
```

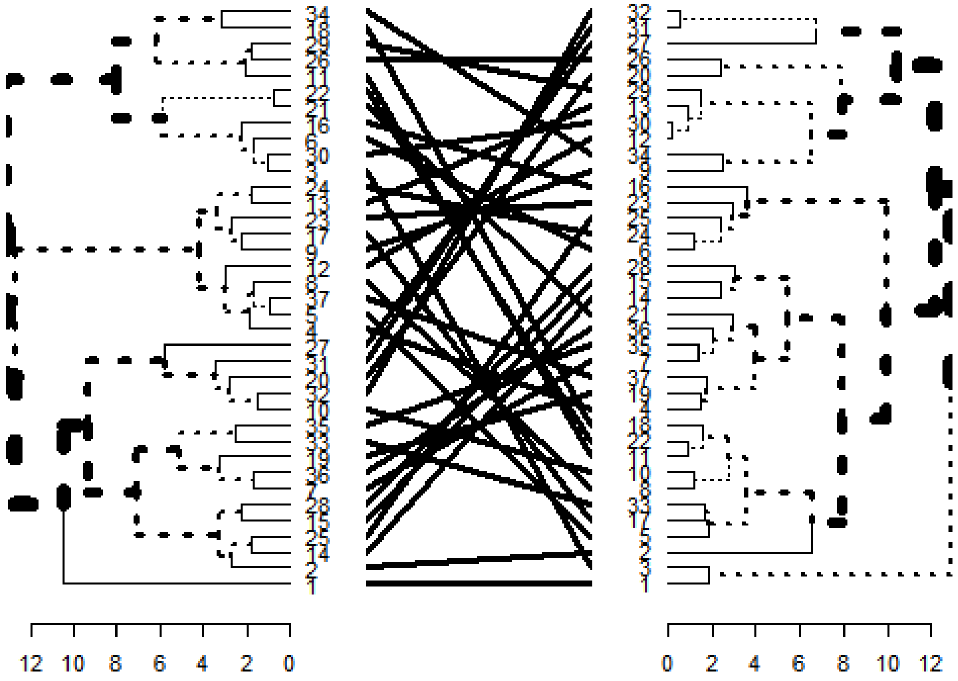
```r
hc_single <- agnes(train_data, method = "single")
hc_complete <- agnes(train_data, method = "complete")
hc_average <- agnes(train_data, method = "average")
hc_ward <- agnes(train_data, method = "ward")

# Compare Agglomerative coefficients
print(hc_single$ac)

## [1] 0.6482111

print(hc_complete$ac)

## [1] 0.7717234

print(hc_average$ac)

## [1] 0.7358853

print(hc_ward$ac)

## [1] 0.8199919

# Compute with agnes and with different linkage methods
hc_single1 <- agnes(test_data, method = "single")
hc_complete1 <- agnes(test_data, method = "complete")
hc_average1 <- agnes(test_data, method = "average")
hc_ward1 <- agnes(test_data, method = "ward")

# Compare Agglomerative coefficients
print(hc_single1$ac)

## [1] 0.6713887

print(hc_complete1$ac)

## [1] 0.8142103

print(hc_average1$ac)

## [1] 0.7402693

print(hc_ward1$ac)

## [1] 0.8457967

pltree(hc_ward,cex=0.6,hang=-1,main="Dendrogram of agnes")
rect.hclust(hc_ward, k = 3, border = 2:5)
```



```
pltree(hc_ward1,cex=0.6,hang=-1,main="Dendrogram of agnes")
rect.hclust(hc_ward1, k = 3, border = 2:5)
```



```
tanglegram(as.dendrogram(hc_ward),as.dendrogram(hc_ward1))
```



```
cor_cophenetic(as.dendrogram(hc_ward),as.dendrogram(hc_ward1))

## [1] 0.09342934

cor_bakers_gamma(as.dendrogram(hc_ward),as.dendrogram(hc_ward1))

## [1] 0.07586184

#Since the stability values are near to zero the above model is not stable

result<-cbind(C1,hcluster)
result[result$hcluster==1,]

##                         name mfr type calories protein fat sodium fiber
## 1                  100%_Bran   N    C       70       4   1    130    10
## 3                   All-Bran   K    C       70       4   1    260     9
## 4 All-Bran_with_Extra_Fiber   K    C       50       4   0    140    14
##   carbo sugars potass vitamins shelf weight cups   rating hcluster
## 1     5      6    280       25     3      1 0.33 68.40297        1
## 3     7      5    320       25     3      1 0.33 59.42551        1
## 4     8      0    330       25     3      1 0.50 93.70491        1

result[result$hcluster==2,]

##                                        name mfr type calories protein fat
## 2                         100%_Natural_Bran   Q    C      120       3   5
## 7                                   Basic_4   G    C      130       3   2
## 13                               Clusters   G    C      110       3   2
## 19                     Cracklin'_Oat_Bran   K    C      110       3   3
## 21                 Crispy_Wheat_&_Raisins   G    C      100       2   1
## 26 Fruit_&_Fibre_Dates,_Walnuts,_and_Oats   P    C      120       3   2
## 27                          Fruitful_Bran   K    C      120       3   0
## 33                     Great_Grains_Pecan   P    C      120       3   3
## 38                 Just_Right_Fruit_&_Nut   K    C      140       3   1
## 40                                   Life   Q    C      100       4   2
## 43       Muesli_Raisins,_Dates,_&_Almonds   R    C      150       4   3
## 44      Muesli_Raisins,_Peaches,_&_Pecans   R    C      150       4   3
## 45                   Mueslix_Crispy_Blend   K    C      160       3   2
## 48              Nutri-Grain_Almond-Raisin   K    C      140       3   2
## 50                   Oatmeal_Raisin_Crisp   G    C      130       3   2
## 51                  Post_Nat._Raisin_Bran   P    C      120       3   1
## 55                     Quaker_Oat_Squares   Q    C      100       4   1
## 56                            Raisin_Bran   K    C      120       3   1
## 57                        Raisin_Nut_Bran   G    C      100       3   2
## 68                      Total_Raisin_Bran   G    C      140       3   1
##    sodium fiber carbo sugars potass vitamins shelf weight cups   rating
## 2      15   2.0   8.0      8    135        0     3   1.00 1.00 33.98368
## 7     210   2.0  18.0      8    100       25     3   1.33 0.75 37.03856
## 13    140   2.0  13.0      7    105       25     3   1.00 0.50 40.40021
## 19    140   4.0  10.0      7    160       25     3   1.00 0.50 40.44877
## 21    140   2.0  11.0     10    120       25     3   1.00 0.75 36.17620
## 26    160   5.0  12.0     10    200       25     3   1.25 0.67 40.91705
## 27    240   5.0  14.0     12    190       25     3   1.33 0.67 41.01549
## 33     75   3.0  13.0      4    100       25     3   1.00 0.33 45.81172
## 38    170   2.0  20.0      9     95      100     3   1.30 0.75 36.47151
## 40    150   2.0  12.0      6     95       25     2   1.00 0.67 45.32807
## 43     95   3.0  16.0     11    170       25     3   1.00 1.00 37.13686
## 44    150   3.0  16.0     11    170       25     3   1.00 1.00 34.13976
## 45    150   3.0  17.0     13    160       25     3   1.50 0.67 30.31335
## 48    220   3.0  21.0      7    130       25     3   1.33 0.67 40.69232
## 50    170   1.5  13.5     10    120       25     3   1.25 0.50 30.45084
## 51    200   6.0  11.0     14    260       25     3   1.33 0.67 37.84059
## 55    135   2.0  14.0      6    110       25     3   1.00 0.50 49.51187
## 56    210   5.0  14.0     12    240       25     2   1.33 0.75 39.25920
## 57    140   2.5  10.5      8    140       25     3   1.00 0.50 39.70340
## 68    190   4.0  15.0     14    230      100     3   1.50 1.00 28.59278
##    hcluster
## 2         2
## 7         2
## 13        2
## 19        2
## 21        2
## 26        2
## 27        2
## 33        2
## 38        2
## 40        2
## 43        2
## 44        2
## 45        2
## 48        2
## 50        2
## 51        2
## 55        2
## 56        2
## 57        2
## 68        2
```

```
result[result$hcluster==3,]
```

```
##                        name mfr type calories protein fat sodium fiber
## 5  Apple_Cinnamon_Cheerios   G    C      110       2   2    180   1.5
## 6              Apple_Jacks   K    C      110       2   0    125   1.0
## 10            Cap'n'Crunch   Q    C      120       1   2    220   0.0
## 12    Cinnamon_Toast_Crunch   G    C      120       1   3    210   0.0
## 14             Cocoa_Puffs   G    C      110       1   1    180   0.0
## 17               Corn_Pops   K    C      110       1   0     90   1.0
## 18           Count_Chocula   G    C      110       1   1    180   0.0
## 23             Froot_Loops   K    C      110       2   1    125   1.0
## 24          Frosted_Flakes   K    C      110       1   0    200   1.0
## 28          Fruity_Pebbles   P    C      110       1   1    135   0.0
## 29            Golden_Crisp   P    C      100       2   0     45   0.0
## 30          Golden_Grahams   G    C      110       1   1    280   0.0
## 34         Honey_Graham_Ohs   Q    C      120       1   2    220   1.0
## 35       Honey_Nut_Cheerios   G    C      110       3   1    250   1.5
## 36              Honey-comb   P    C      110       1   0    180   0.0
## 41            Lucky_Charms   G    C      110       2   1    180   0.0
## 46    Multi-Grain_Cheerios   G    C      100       2   1    220   2.0
## 47        Nut&Honey_Crunch   K    C      120       2   1    190   0.0
## 64                  Smacks   K    C      110       2   1     70   1.0
## 71                    Trix   G    C      110       1   1    140   0.0
## 74      Wheaties_Honey_Gold   G    C      110       2   1    200   1.0
##    carbo sugars potass vitamins shelf weight cups   rating hcluster
## 5   10.5     10     70       25     1      1 0.75 29.50954        3
## 6   11.0     14     30       25     2      1 1.00 33.17409        3
## 10  12.0     12     35       25     2      1 0.75 18.04285        3
## 12  13.0      9     45       25     2      1 0.75 19.82357        3
## 14  12.0     13     55       25     2      1 1.00 22.73645        3
## 17  13.0     12     20       25     2      1 1.00 35.78279        3
## 18  12.0     13     65       25     2      1 1.00 22.39651        3
## 23  11.0     13     30       25     2      1 1.00 32.20758        3
## 24  14.0     11     25       25     1      1 0.75 31.43597        3
## 28  13.0     12     25       25     2      1 0.75 28.02576        3
## 29  11.0     15     40       25     1      1 0.88 35.25244        3
## 30  15.0      9     45       25     2      1 0.75 23.80404        3
## 34  12.0     11     45       25     2      1 1.00 21.87129        3
## 35  11.5     10     90       25     1      1 0.75 31.07222        3
## 36  14.0     11     35       25     1      1 1.33 28.74241        3
## 41  12.0     12     55       25     2      1 1.00 26.73451        3
## 46  15.0      6     90       25     1      1 1.00 40.10596        3
## 47  15.0      9     40       25     2      1 0.67 29.92429        3
## 64   9.0     15     40       25     2      1 0.75 31.23005        3
## 71  13.0     12     25       25     2      1 1.00 27.75330        3
## 74  16.0      8     60       25     1      1 0.75 36.18756        3
```

```
result[result$hcluster==4,]
```

```
##                           name mfr type calories protein fat sodium fiber
## 8                    Bran_Chex   R    C       90       2   1    200     4
## 9                  Bran_Flakes   P    C       90       3   0    210     5
## 11                    Cheerios   G    C      110       6   2    290     2
## 15                   Corn_Chex   R    C      110       2   0    280     0
## 16                 Corn_Flakes   K    C      100       2   0    290     1
## 20                     Crispix   K    C      110       2   0    220     1
## 22                 Double_Chex   R    C      100       2   0    190     1
## 25         Frosted_Mini-Wheats   K    C      100       3   0      0     3
## 31           Grape_Nuts_Flakes   P    C      100       3   1    140     3
## 32                  Grape-Nuts   P    C      110       3   0    170     3
## 37 Just_Right_Crunchy__Nuggets   K    C      110       2   1    170     1
## 39                         Kix   G    C      110       2   1    260     0
## 42                       Maypo   A    H      100       4   1      0     0
## 49           Nutri-grain_Wheat   K    C       90       3   0    170     3
## 52                  Product_19   K    C      100       3   0    320     1
## 53                 Puffed_Rice   Q    C       50       1   0      0     0
## 54                Puffed_Wheat   Q    C       50       2   0      0     1
## 58              Raisin_Squares   K    C       90       2   0      0     2
## 59                   Rice_Chex   R    C      110       1   0    240     0
## 60               Rice_Krispies   K    C      110       2   0    290     0
## 61              Shredded_Wheat   N    C       80       2   0      0     3
## 62      Shredded_Wheat_'n'Bran   N    C       90       3   0      0     4
## 63   Shredded_Wheat_spoon_size   N    C       90       3   0      0     3
## 65                   Special_K   K    C      110       6   0    230     1
## 66     Strawberry_Fruit_Wheats   N    C       90       2   0     15     3
## 67           Total_Corn_Flakes   G    C      110       2   1    200     0
## 69           Total_Whole_Grain   G    C      100       3   1    200     3
## 70                     Triples   G    C      110       2   1    250     0
## 72                  Wheat_Chex   R    C      100       3   1    230     3
## 73                    Wheaties   G    C      100       3   1    200     3
##    carbo sugars potass vitamins shelf weight cups   rating hcluster
## 8     15      6    125       25     1   1.00 0.67 49.12025        4
## 9     13      5    190       25     3   1.00 0.67 53.31381        4
## 11    17      1    105       25     1   1.00 1.25 50.76500        4
## 15    22      3     25       25     1   1.00 1.00 41.44502        4
## 16    21      2     35       25     1   1.00 1.00 45.86332        4
## 20    21      3     30       25     3   1.00 1.00 46.89564        4
## 22    18      5     80       25     3   1.00 0.75 44.33086        4
## 25    14      7    100       25     2   1.00 0.80 58.34514        4
## 31    15       5     85        25      3    1.00 0.88 52.07690         4
## 32    17       3     90        25      3    1.00 0.25 53.37101         4
## 37    17       6     60       100      3    1.00 1.00 36.52368         4
## 39    21       3     40        25      2    1.00 1.50 39.24111         4
## 42    16       3     95        25      2    1.00 1.00 54.85092         4
## 49    18       2     90        25      3    1.00 1.00 59.64284         4
## 52    20       3     45       100      3    1.00 1.00 41.50354         4
## 53    13       0     15         0      3    0.50 1.00 60.75611         4
## 54    10       0     50         0      3    0.50 1.00 63.00565         4
## 58    15       6    110        25      3    1.00 0.50 55.33314         4
## 59    23       2     30        25      1    1.00 1.13 41.99893         4
## 60    22       3     35        25      1    1.00 1.00 40.56016         4
## 61    16       0     95         0      1    0.83 1.00 68.23588         4
## 62    19       0    140         0      1    1.00 0.67 74.47295         4
## 63    20       0    120         0      1    1.00 0.67 72.80179         4
## 65    16       3     55        25      1    1.00 1.00 53.13132         4
## 66    15       5     90        25      2    1.00 1.00 59.36399         4
## 67    21       3     35       100      3    1.00 1.00 38.83975         4
## 69    16       3    110       100      3    1.00 1.00 46.65884         4
## 70    21       3     60        25      3    1.00 0.75 39.10617         4
## 72    17       3    115        25      1    1.00 0.67 49.78744         4
## 73    17       3    110        25      1    1.00 1.00 51.59219         4
```

```
#From the above results we can say that elementary public schools belongs to
cluster 1 because it has highest rating.We need to normalize the data set
because the data set is having diffferent range values.
```