Untitled

Jacob Fabian

2022-11-25

```
library(readr)
library(cluster)
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
```

```
setwd("~/Downloads")
```

```
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)
```

```
head(Cereals)
```

```
## # A tibble: 6 x 16
##   name        mfr   type  calor~1 protein   fat sodium fiber carbo sugars potass
##   <chr>       <chr> <chr>   <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>  <dbl>  <dbl>
## 1 100%_Bran   N     C          70       4     1    130  10       5      6    280
## 2 100%_Natur~ Q     C         120       3     5     15   2       8      8    135
## 3 All-Bran    K     C          70       4     1    260   9       7      5    320
## 4 All-Bran_w~ K     C          50       4     0    140  14       8      0    330
```

```
## 5 Almond_Del~ R     C         110       2     2    200   1    14         8     NA
## 6 Apple_Cinn~ G     C         110       2     2    180   1.5  10.5      10     70
## # ... with 5 more variables: vitamins <dbl>, shelf <dbl>, weight <dbl>,
## #   cups <dbl>, rating <dbl>, and abbreviated variable name 1: calories
```

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

```
## # A tibble: 6 x 16
##   name        mfr   type  calor~1 protein   fat sodium fiber carbo sugars potass
##   <chr>       <chr> <chr>   <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>  <dbl>  <dbl>
## 1 100%_Bran   N     C          70       4     1    130  10     5        6    280
## 2 100%_Natur~ Q     C         120       3     5     15   2     8        8    135
## 3 All-Bran    K     C          70       4     1    260   9     7        5    320
## 4 All-Bran_w~ K     C          50       4     0    140  14     8        0    330
## 5 Apple_Cinn~ G     C         110       2     2    180   1.5  10.5     10     70
## 6 Apple_Jacks K     C         110       2     0    125   1    11       14     30
## # ... with 5 more variables: vitamins <dbl>, shelf <dbl>, weight <dbl>,
## #   cups <dbl>, rating <dbl>, and abbreviated variable name 1: calories
```

```
Cereals.norm <- Cereals %>%
    as_tibble() %>%
    mutate(across(where(is.numeric), scale))
```

```
distance <- dist(Cereals.norm[4:16], method = "euclidean")
hc_single <- agnes(Cereals.norm[4:16], method = "single")
hc_complete <- agnes(Cereals.norm[4:16], method = "complete")
hc_average <- agnes(Cereals.norm[4:16], method = "average")
hc_ward <- agnes(Cereals.norm[4:16], method = "ward")
```

```
print(hc_single)
```

```
## Call:     agnes(x = Cereals.norm[4:16], method = "single")
## Agglomerative coefficient:  0.6067859
## Order of objects:
##  [1]  1  3  4  2  5 35  6 14 18 71 41 23 28 17 10 34 12 64 46 74 47  8 72 73 30
## [26] 24 29 36  7 48 50 26 27 51 56 13 57 19 55 33 40 21 31 49 20 22 70 32 15 60
## [51] 16 59  9 25 66 58 42 61 62 63 39 45 11 65 43 44 37 67 69 52 38 68 53 54
## Height (summary):
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1431  1.3777  1.7695  1.8668  2.2787  4.0361
##
## Available components:
## [1] "order"  "height" "ac"     "merge"  "diss"   "call"   "method" "data"
```

```
print(hc_complete)
```

```
## Call:     agnes(x = Cereals.norm[4:16], method = "complete")
## Agglomerative coefficient:  0.8353712
## Order of objects:
##  [1]  1  3  4  2 25 66 58 42 61 62 63 53 54  5 35 46 74 24 30 47 10 34 12  6 17
```

```
## [26] 14 18 71 28 23 41 29 64 36  8 72 73  9 31 49 32 13 57 19 33 21 40 55 11 65
## [51] 15 60 16 59 39 20 22 70 37 67 69 52  7 48 45 26 50 43 44 27 51 56 38 68
## Height (summary):
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1431  1.6076  2.3389  2.9321  3.7169 10.9839
##
## Available components:
## [1] "order"  "height" "ac"     "merge"  "diss"   "call"   "method" "data"
```

```
print(hc_average)
```

```
## Call:     agnes(x = Cereals.norm[4:16], method = "average")
## Agglomerative coefficient:  0.7766075
## Order of objects:
##  [1]  1  3  4  2  5 35 46 74 24 30 47  6 17 14 18 71 23 41 28 29 64 10 34 12 36
## [26]  8 72 73  9 32 20 22 70 31 49 13 57 19 33 40 55 21 15 60 16 59 39 25 66 58
## [51] 42 61 62 63  7 48 50 45 26 27 51 56 43 44 37 67 69 52 38 68 11 65 53 54
## Height (summary):
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1431  1.4633  2.0666  2.4461  2.9445  7.7243
##
## Available components:
## [1] "order"  "height" "ac"     "merge"  "diss"   "call"   "method" "data"
```

```
print(hc_ward)
```

```
## Call:     agnes(x = Cereals.norm[4:16], method = "ward")
## Agglomerative coefficient:  0.9046042
## Order of objects:
##  [1]  1  3  4  2 43 44 13 57 19 33 21 40 55  7 48 45 26 50 27 51 56 38 68  5 35
## [26] 46 74 24 30 47 10 34 12  6 17 29 64 14 18 71 28 23 41 36  8 72 73  9 31 49
## [51] 32 20 22 70 11 65 15 60 16 59 39 37 67 69 52 25 66 58 42 61 62 63 53 54
## Height (summary):
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1431  1.5858  2.3422  3.6092  4.1559 18.5749
##
## Available components:
## [1] "order"  "height" "ac"     "merge"  "diss"   "call"   "method" "data"
```

###The highest coefficient will be the best method

Single - 0.607

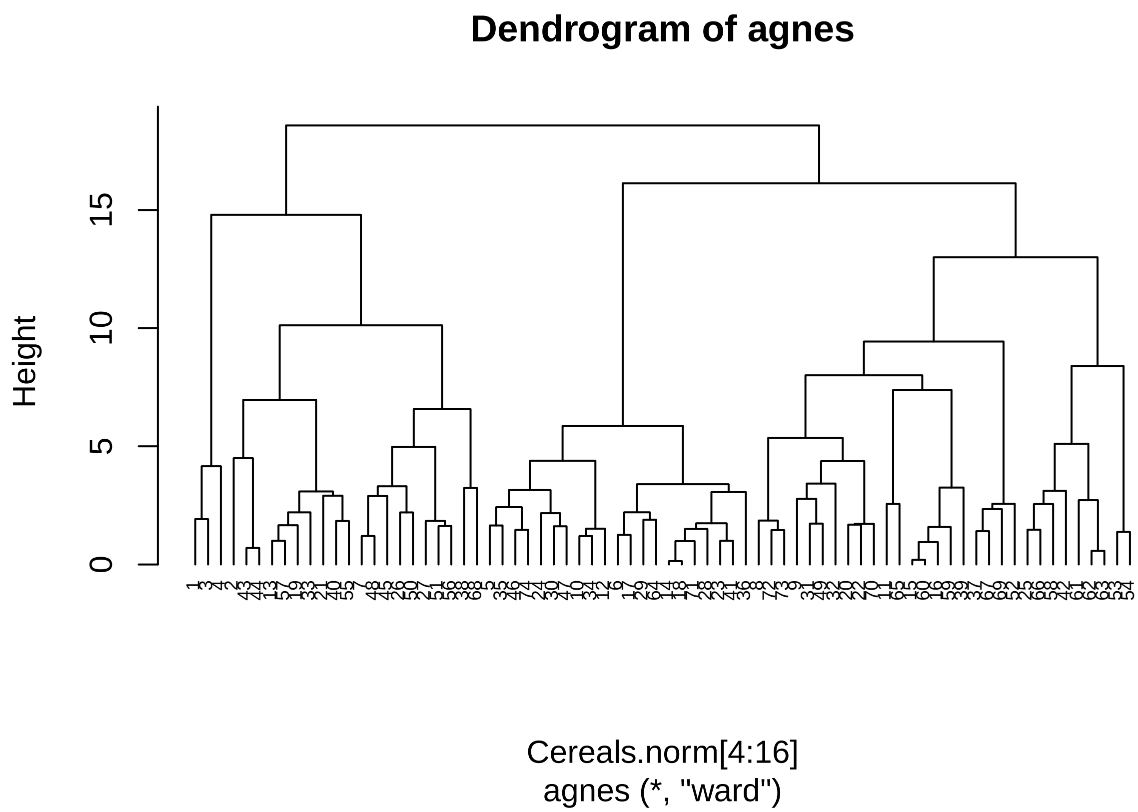
Compete - 0.835

Average - 0.777

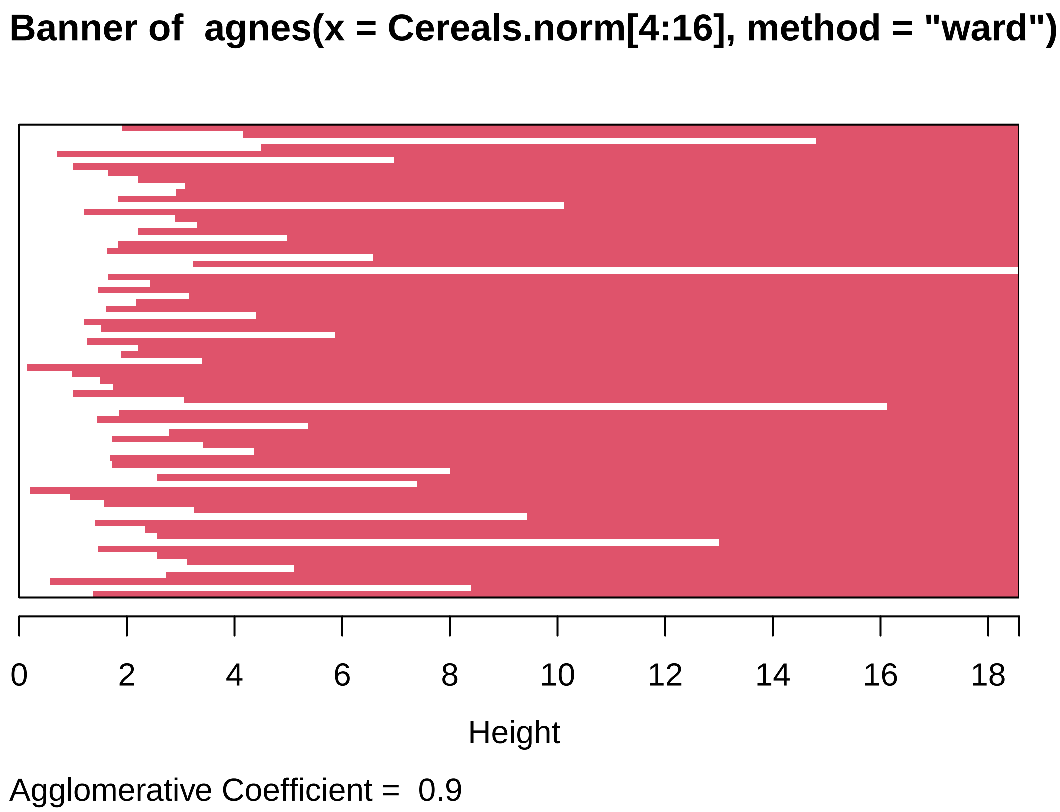
Ward - 0.904

Since Ward has the highest coefficient we will look at that

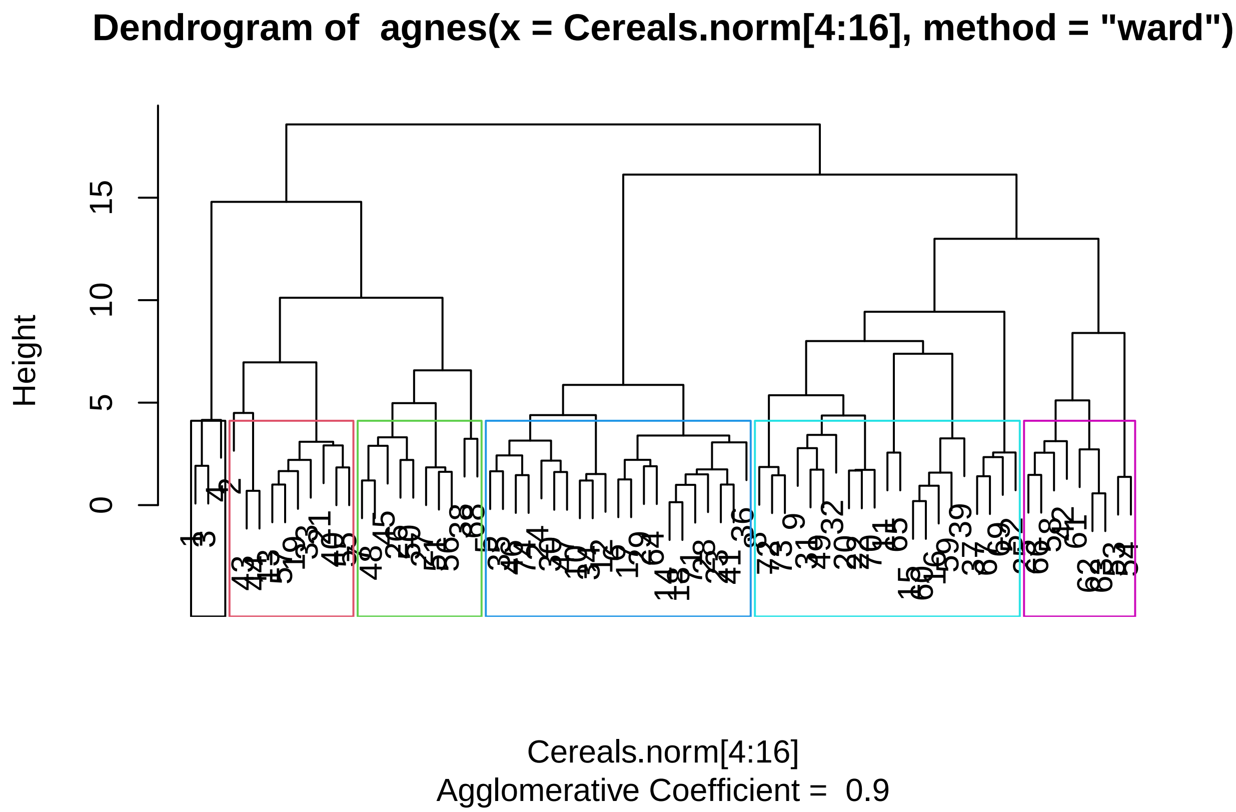
```r
pltree(hc_ward, cex = 0.6, hang = -1, main = "Dendrogram of agnes")
```




```r
plot(hc_ward)
```




```
rect.hclust(hc_ward, k = 6, border = 1:6)
```



6 Clusters would be the best

```
model <- kmeans(Cereals.norm[4:16], centers = 6, nstart = 25)
100 * model$betweenss / model$totss
```

```
## [1] 58.62927
```

58.63% stay in their cluster.

```
cl <- kmeans(Cereals[4:12], centers = 6, nstart = 25)
Cereals <- data.frame(Cereals, cl$cluster)
cl$centers
```

```
##   calories  protein       fat   sodium    fiber    carbo    sugars    potass
## 1  95.0000 3.500000 0.8333333 188.3333 8.000000 10.00000  8.500000 276.66667
## 2 119.2857 3.071429 1.7142857 163.5714 3.214286 14.00000  8.785714 149.28571
## 3 110.4000 2.240000 1.0000000 194.8000 1.260000 15.42000  7.280000  67.40000
## 4 108.0000 2.400000 0.6000000 275.0000 0.550000 19.35000  3.900000  51.00000
## 5  86.0000 2.500000 0.6000000   3.0000 2.100000 14.60000  2.900000  95.00000
## 6 108.8889 1.888889 0.8888889 105.0000 1.111111 12.11111 11.333333  43.88889
##   vitamins
```

```
## 1      37.5
## 2      25.0
## 3      37.0
## 4      32.5
## 5      10.0
## 6      25.0
```

The data should be standardized since when it comes to what we eat, we should value what we put into our body. Cluster 1 is probably the most healthy, since it is high in protien, fiber, potassium and higher in vitamins. And less carboydrates and low calories.