Assignment 5

Jacob Aylward

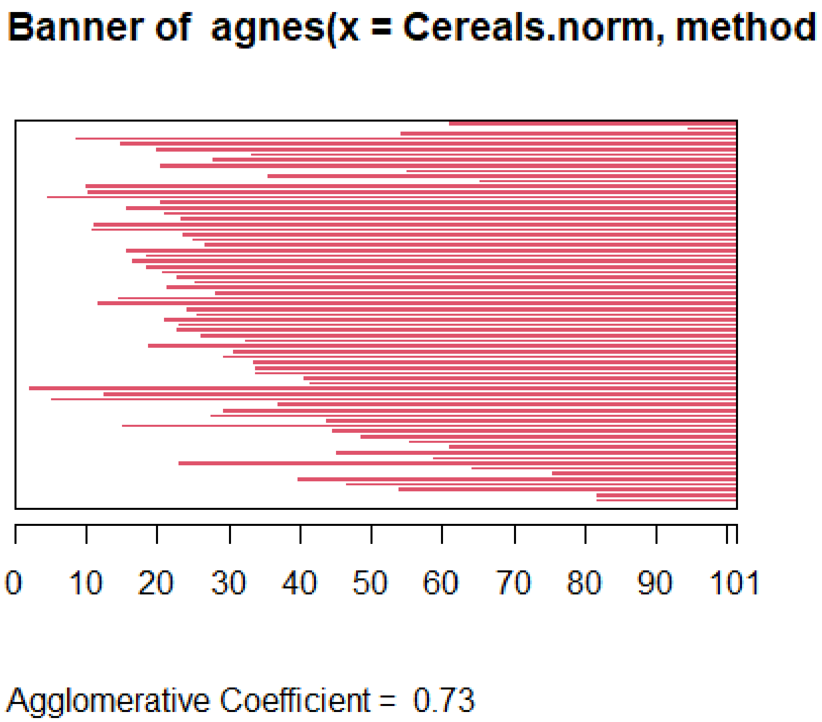
2025-06-22

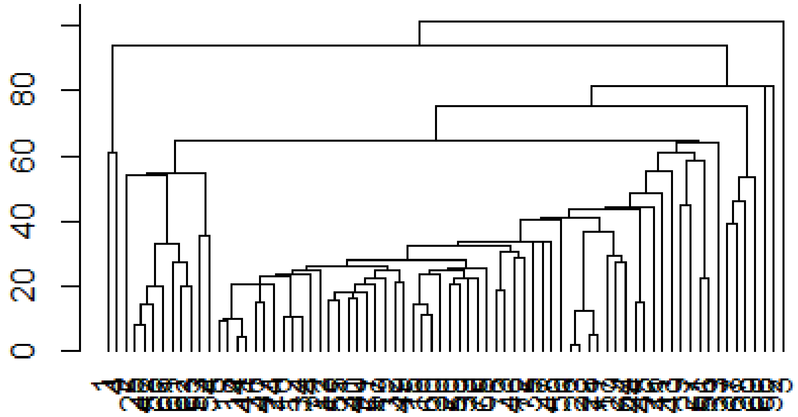
```
library(dplyr)

##
## 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(readr)
Cereals <- read_csv("C:/Users/jacob/Downloads/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)

Cereals <- na.omit(Cereals)

head(Cereals)

## # A tibble: 6 × 16
##   name        mfr   type  calories 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%_Natu… 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_… K     C            50       4     0    140  14     8         0
330
## 5 Apple_Cin… G     C           110       2     2    180   1.5  10.5      10
70
## 6 Apple_Jac… K     C           110       2     0    125   1    11        14
30
## # ℹ 5 more variables: vitamins <dbl>, shelf <dbl>, weight <dbl>, cups
<dbl>,
## #   rating <dbl>
```

```
Cereals.nm <- Cereals %>% select(where(is.numeric))

Cereals.norm <- dist(Cereals.nm, method = "euclidean")

Cereals.scaled <- scale(Cereals.norm)

library(cluster)

sin. <- agnes(Cereals.norm, method = "single")
comp. <- agnes(Cereals.norm, method = "complete")
avg. <- agnes(Cereals.norm, method = "average")
ward. <- agnes(Cereals.norm, method = "ward")

plot(sin., hang = -1, ann = FALSE)
```

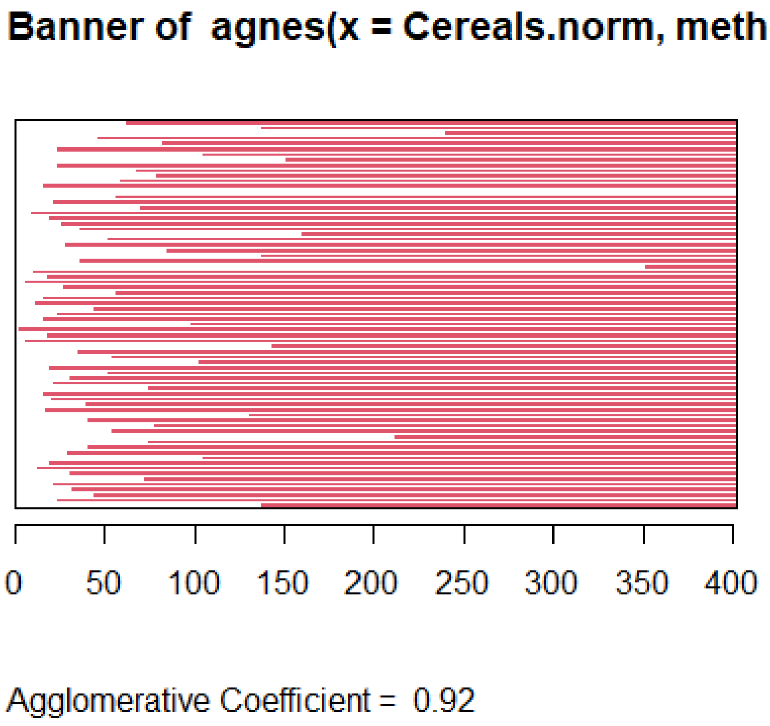
```
## Warning in plot.window(xlim, ylim, log = log, ...): "hang" is not a
graphical
## parameter

## Warning in axis(1, at = at.vals, labels = lab.vals, ...): "hang" is not a
## graphical parameter
```

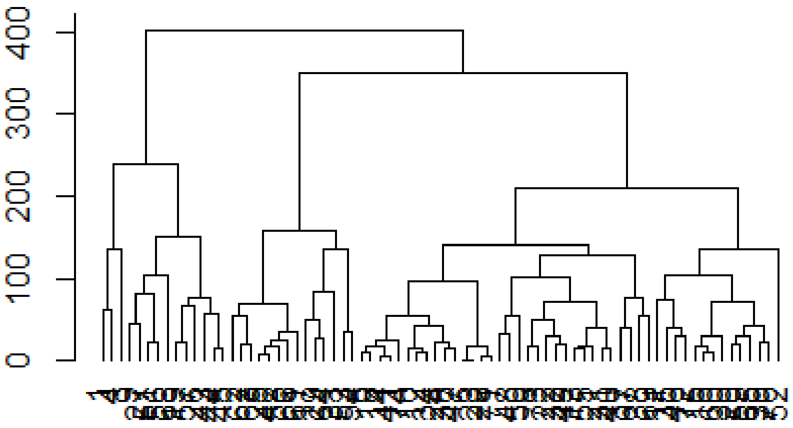




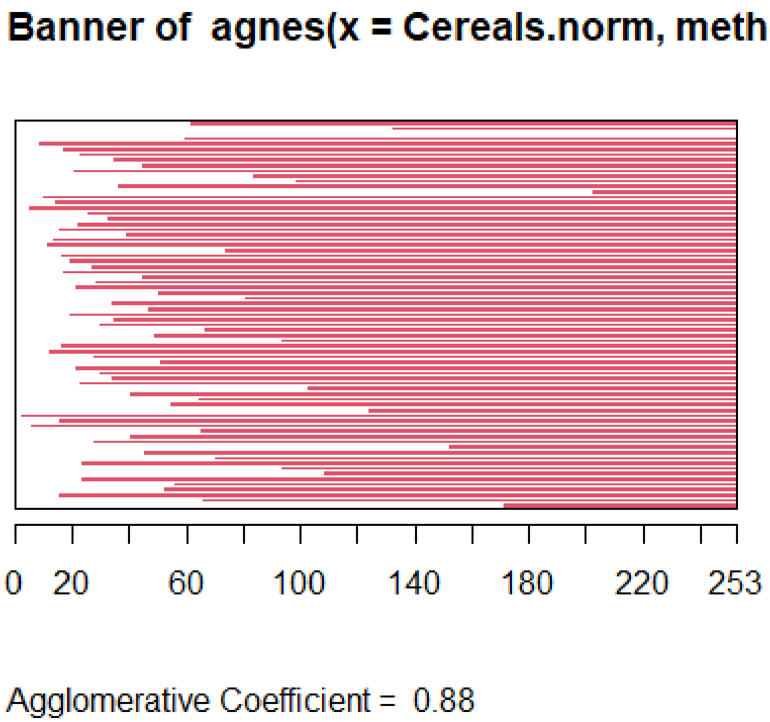
```
plot(comp., hang = -1, ann = FALSE)

## Warning in plot.window(xlim, ylim, log = log, ...): "hang" is not a
graphical
```

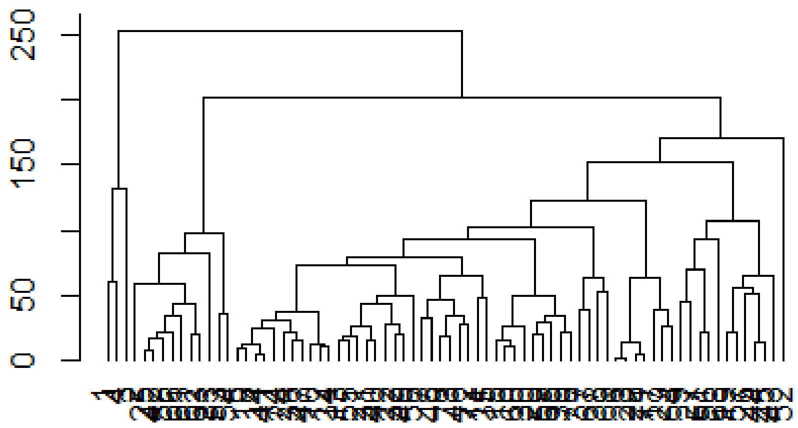
```
## parameter
## Warning in plot.window(xlim, ylim, log = log, ...): "hang" is not a
graphical
## parameter
```





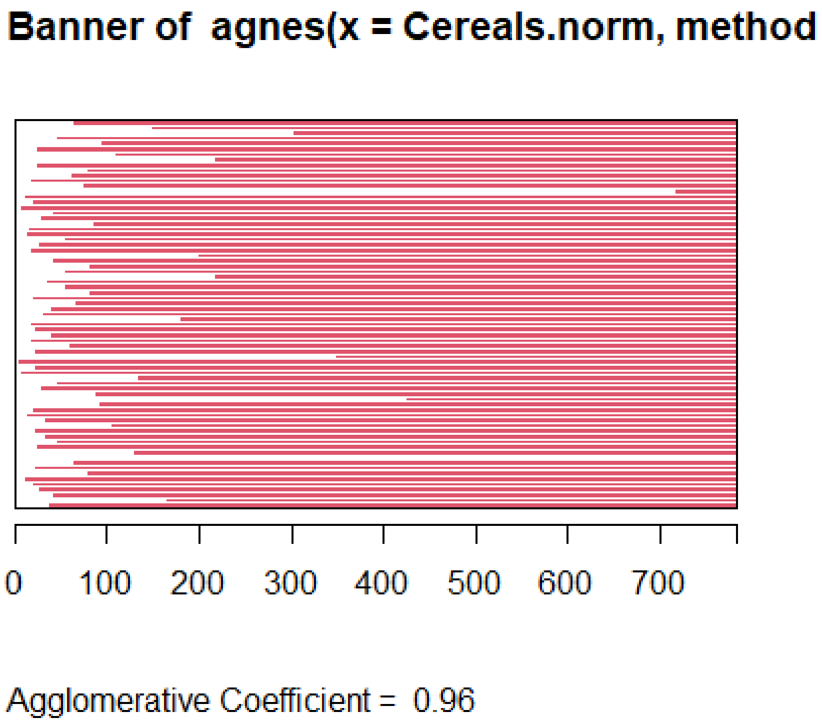
```
plot(avg., hang = -1, ann = FALSE)

## Warning in plot.window(xlim, ylim, log = log, ...): "hang" is not a
graphical
## parameter
## Warning in plot.window(xlim, ylim, log = log, ...): "hang" is not a
graphical
## parameter
```

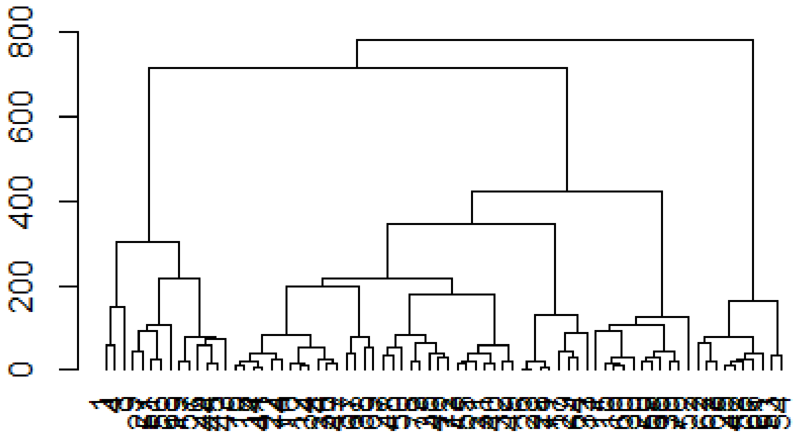




```
plot(ward., hang = -1, ann = FALSE)

## Warning in plot.window(xlim, ylim, log = log, ...): "hang" is not a
graphical
```

```
## parameter
## Warning in plot.window(xlim, ylim, log = log, ...): "hang" is not a
graphical
## parameter
```





```
print(sin.$ac)

## [1] 0.7311616

print(comp.$ac)

## [1] 0.922957

print(avg.$ac)

## [1] 0.8792621

print(ward.$ac)

## [1] 0.9597071
```

Best method is Ward based on the values given above Based on the cluster dendrogram for Ward, I would select 3 clusters (k = 3)

Based on the clusters for the Ward dendrogram the cluster that offers potentially the most “balanced” cereals would be cluster 3 (far right section). With cluster 2 due to the high values for some these are likely to be the least healthy cereals and should be avoided. With the stability of these clusters determining the best cereals or cluster that offers the best will require focus to be on determining which clusters have cereals with the lowest amounts of harmful substances such as too much sugar or sodium

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.” Should the data be normalized? If not, how should they be used in the cluster analysis?

Yes, the data should be normalized due to the high ranges of values in columns such as calories, fiber, and sugar. The larger values of the “unhealthy” variables will ruin the clusters causing cereals to be chosen that are unhealthy. When the data clusters are normalized the healthy cereal options can be chosen from the clusters based on required features such as having low sugar.