```r
# Assignment 5 setup
library(cluster)
```

```
## Warning: package 'cluster' was built under R version 4.0.3
```

```r
library(stats)
library(caret)
```

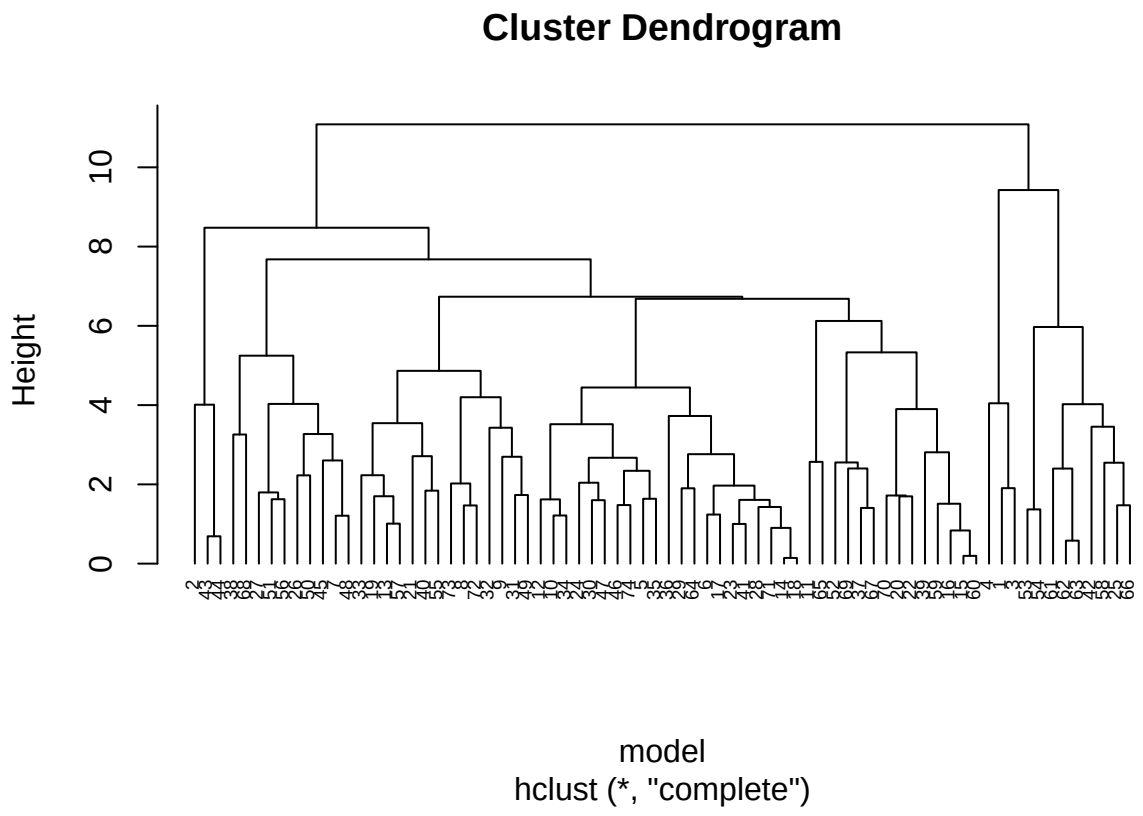
```
## Loading required package: lattice

## Loading required package: ggplot2
```

```r
library(ISLR)
```

```r
# Data Clean-up: Pre-processing
Cereals <- read.csv("Cereals.csv")
Cereals_Numeric <- Cereals[,(4:16)] # removes categorical variables
scaledcereals <- scale(Cereals_Numeric) # scales data
Cereals_Cleaned <- scaledcereals[complete.cases(scaledcereals),] # removes NAs after scaling data
```

```r
# Part 1.
model <- dist(Cereals_Cleaned, method = "euclidean")
model2 <-hclust(model, method = "complete")
plot(model2, cex = 0.6, hang = -1)
```



```r
# set models for each linkage method
md_single <- agnes(Cereals_Cleaned, method = "single")
md_complete <- agnes(Cereals_Cleaned, method = "complete")
md_average <- agnes(Cereals_Cleaned, method = "average")
md_ward <- agnes(Cereals_Cleaned, method = "ward")
# show results to determine best method
md_single$ac
```

```
## [1] 0.6094447
```

```r
md_complete$ac
```

```
## [1] 0.8413498
```

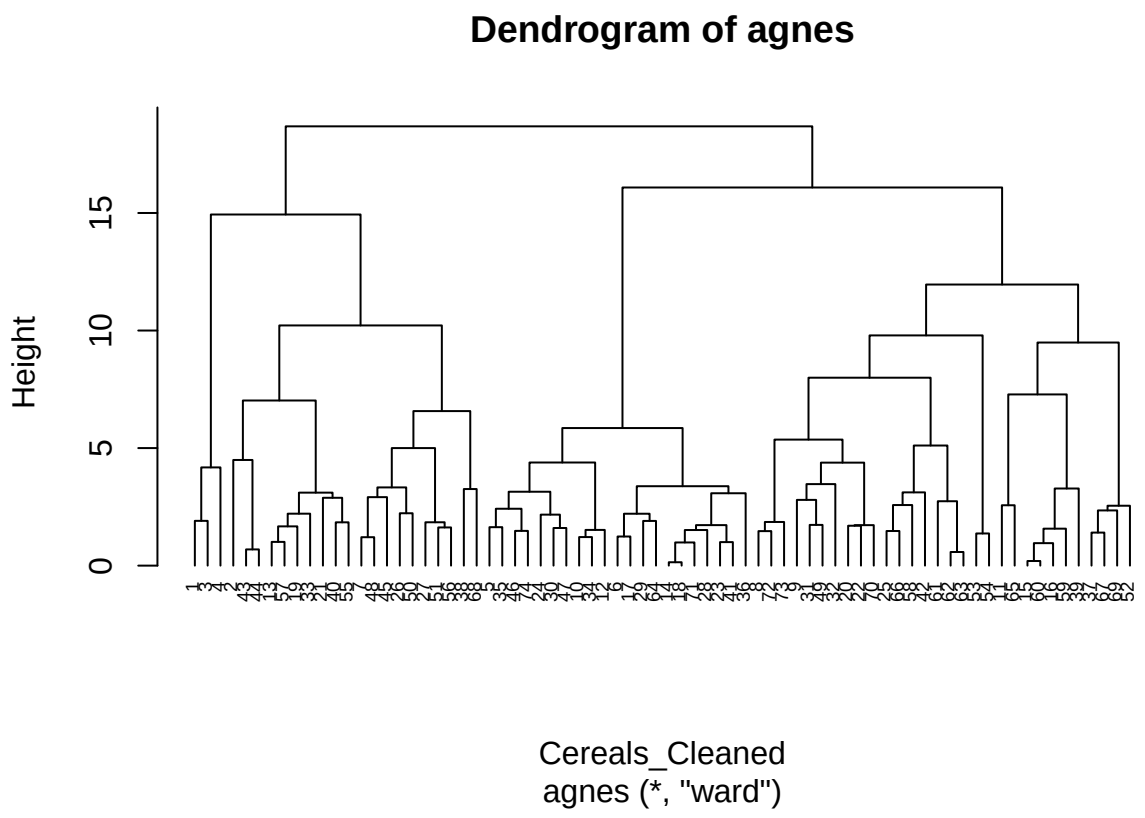
```r
md_average$ac
```

```
## [1] 0.7814484
```

```r
md_ward$ac
```

```
## [1] 0.9049881
```

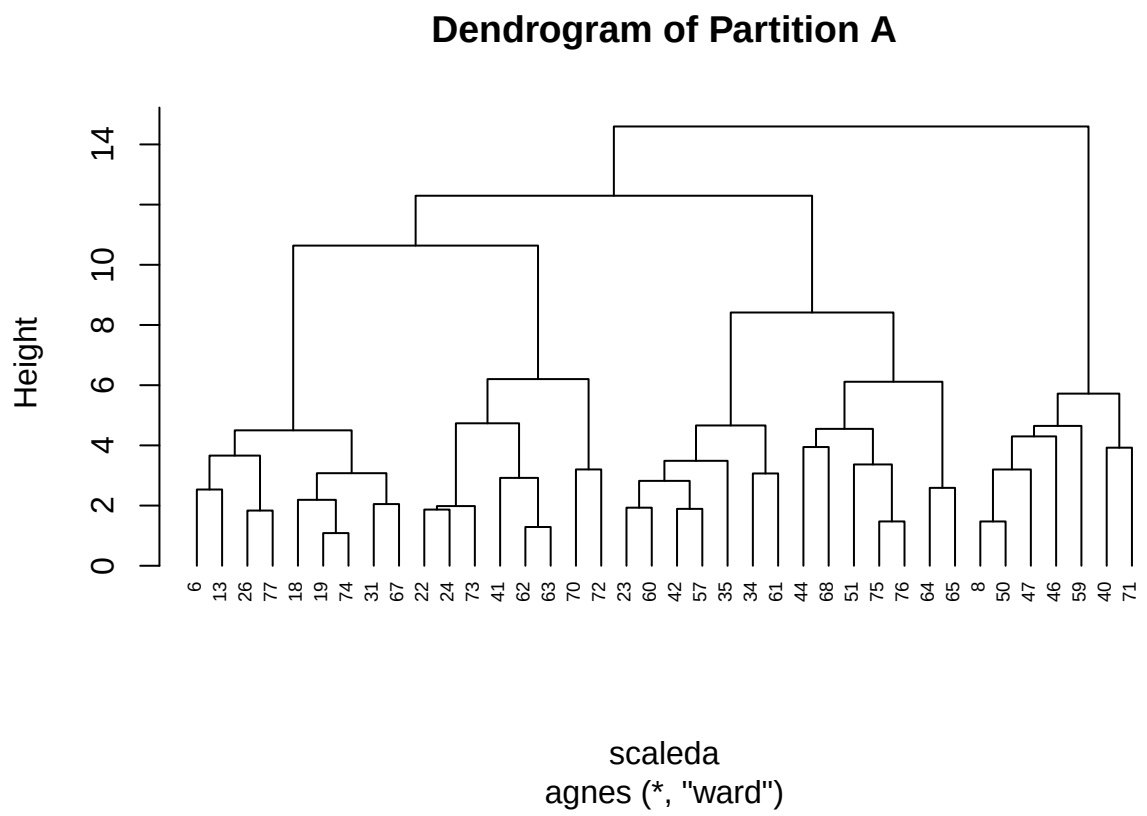
```r
# "Ward" is the best method here
pltree(md_ward, cex = 0.6, hang = -1, main = "Dendrogram of agnes") # Dendrogram using ward
```



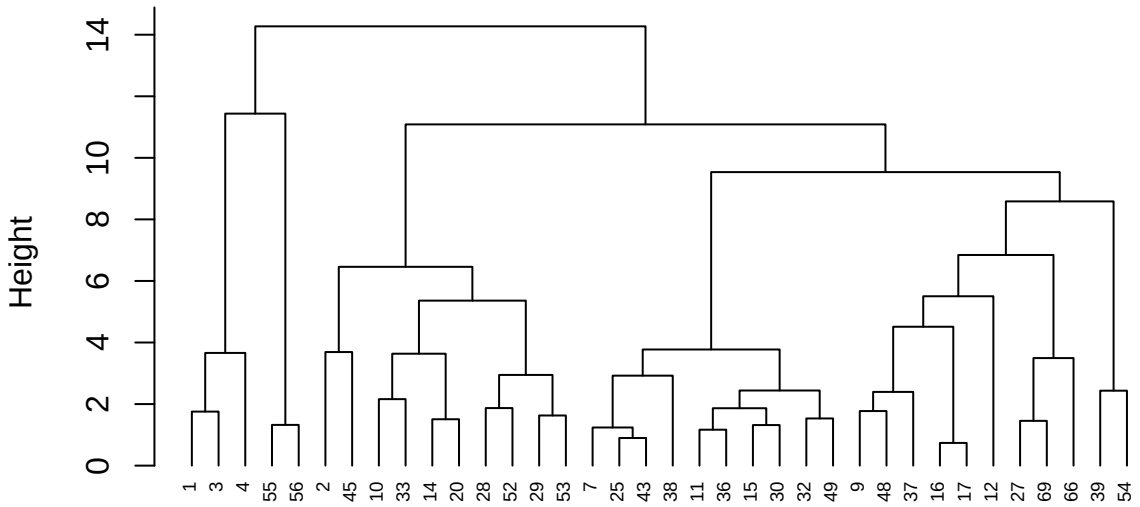
```r
# Part 2.
# While both are unsupervised, there are several major differences between them.
# K-means requires knowledge beforehand of the number of clusters, while
# hierarchical clustering does not require this knowledge. K-means does not use
# hierarchies, while HC is build around it. K-means uses the distance between different
# data-points, while HC builds clusters based on the closest points to other points.

# Part 3.
# I would use 3 clusters.
```

```r
# Part 4.
Cer_index = createDataPartition(Cereals_Numeric$calories, p=0.5, list = FALSE)
Cer_a = Cereals_Numeric[Cer_index,]
cer_b = Cereals_Numeric[-Cer_index,]
scaleda <- scale(Cer_a)
scaledb <- scale(cer_b)
scaleda <- scaleda[complete.cases(scaleda),]
scaledb <- scaledb[complete.cases(scaledb),]
a_ward <- agnes(scaleda, method = "ward")
b_ward <- agnes(scaledb, method = "ward")
pltree(a_ward, cex = 0.6, hang = -1, main = "Dendrogram of Partition A")
```




```r
pltree(b_ward, cex = 0.6, hang = -1, main = "Dendrogram of Partition B")
```




```
# On the surface, the clusters do not seem to be stable. The shapes of the dendrograms
# are significantly different from each other, and the best number of clusters seems
# to change. However, there is much more stability when it comes to individual results.
# Most cereals are clustered nearby to the same ones in each partition as they are in
# the general model. For example, numbers 1-4, which have some of the closest
# values of any cereal, and clustered nearby in partitions A and B too.

# Part 5.
# The data should be normalized. This prevents variables with a larger scale from
# biasing the results. To build a cluster built specifically around health, only
# the relevant variables should be used. Fat and sugar may be most important,
# while fiber may be considered less important, and weight and shelf being excluded
# entirely. Based on the clusters built above, the 1st cluster is the healthiest,
# containing cereals like all the bran cereals.

# Part 6.
# The biggest advantage of HC vs K-means is that it is easier to visualize and understand.
# The hierarchical nature allows for dendrograms, which make is easier to determine the
# number of clusters. The hierarchy also provides more information about the relationship
# between different variables.
```