HW5

2019150432 임효진

June 6, 2021

1

(a)

```r
library(stats)
dat1=matrix(c(2, 2, 4, 5, 6, 6, 7, 7, 6, 4), ncol=2, byrow = T)
dist(dat1 , diag = T)
```

```
##           1        2        3        4        5
## 1 0.000000
## 2 3.605551 0.000000
## 3 5.656854 2.236068 0.000000
## 4 7.071068 3.605551 1.414214 0.000000
## 5 4.472136 2.236068 2.000000 3.162278 0.000000
```
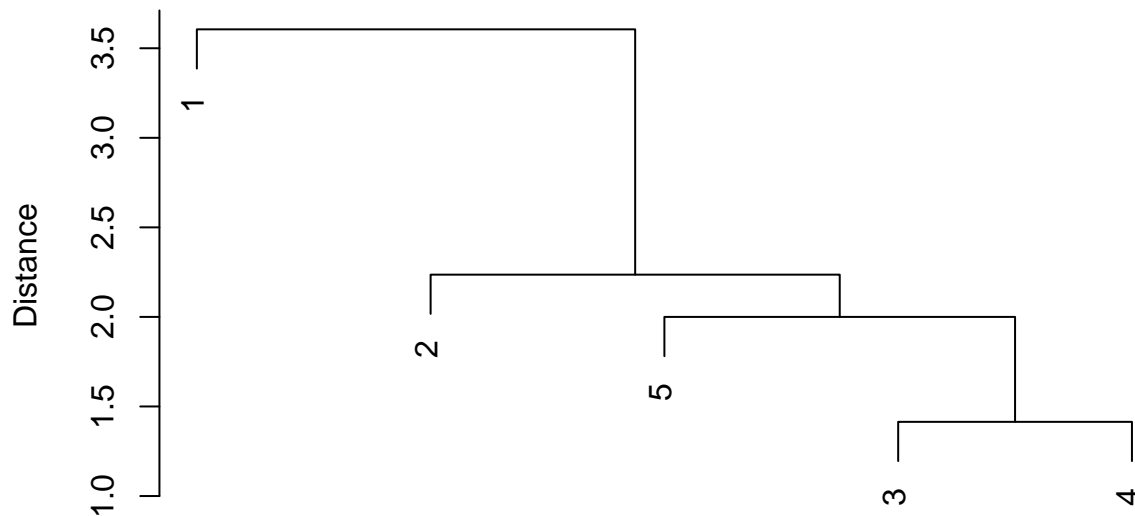
(b)

```r
hclust(dist(dat1),method="single")
```

```
##
## Call:
## hclust(d = dist(dat1), method = "single")
##
## Cluster method   : single
## Distance         : euclidean
## Number of objects: 5
```

```r
plot(hclust(dist(dat1),method="single"),
     xlab=NA,ylab="Distance",
     main="Single linkage dendrogram")
```
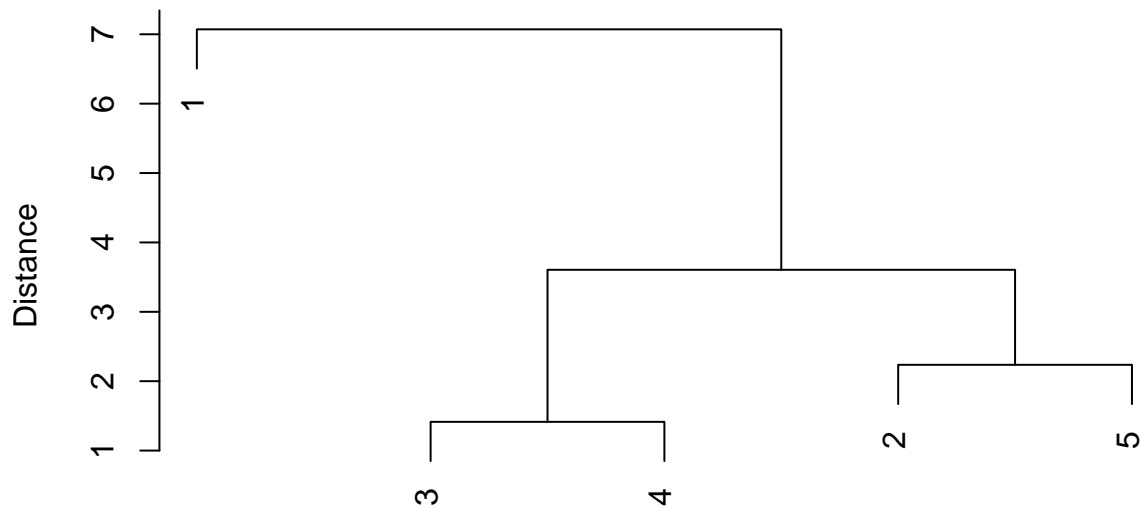
Single linkage dendrogram

hclust (*, "single")

(c)

```r
hclust(dist(dat1),method="complete")
```

```
##
## Call:
## hclust(d = dist(dat1), method = "complete")
##
## Cluster method   : complete
## Distance         : euclidean
## Number of objects: 5
```

```r
plot(hclust(dist(dat1),method="complete"),
     xlab=NA,ylab="Distance",
     main="Complete linkage dendrogram")
```
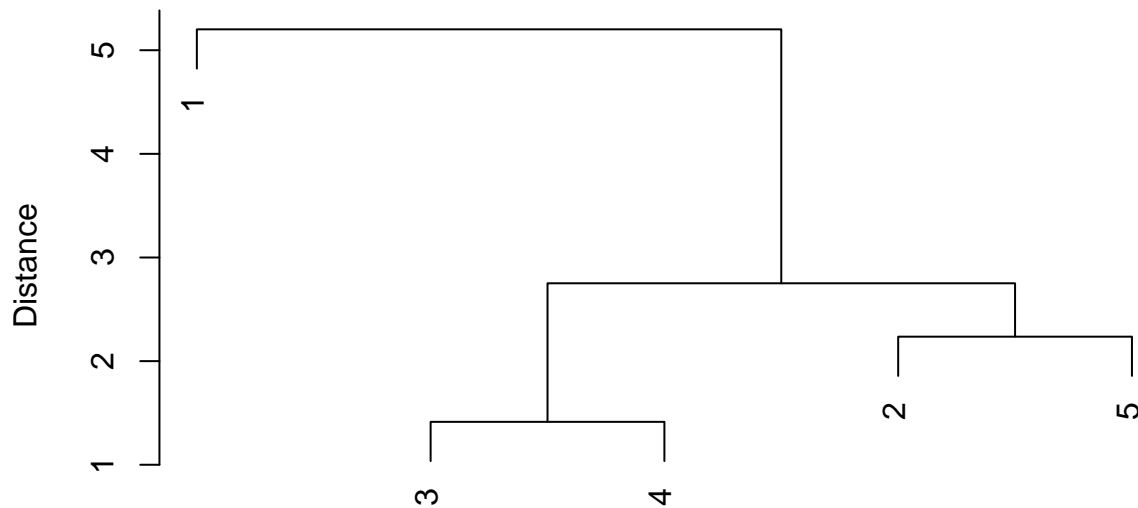
Complete linkage dendrogram

hclust (*, "complete")

(d)

```r
hclust(dist(dat1),method="average")
```

```
##
## Call:
## hclust(d = dist(dat1), method = "average")
##
## Cluster method   : average
## Distance         : euclidean
## Number of objects: 5
```

```r
plot(hclust(dist(dat1),method="average"),
     xlab=NA,ylab="Distance",
     main="Group Average dendrogram")
```

Group Average dendrogram

hclust (*, "average")

(e)

Partitions used with complete linkage and group average are the same but the distances usd to merge groups are different. While complete linkage and group average clustering merged (3, 4) and (2, 5) and (2, 3, 4, 5) sequentially single linkage clustering merged (3, 4) and (3, 4, 5) and (2, 3, 4, 5) in this order.

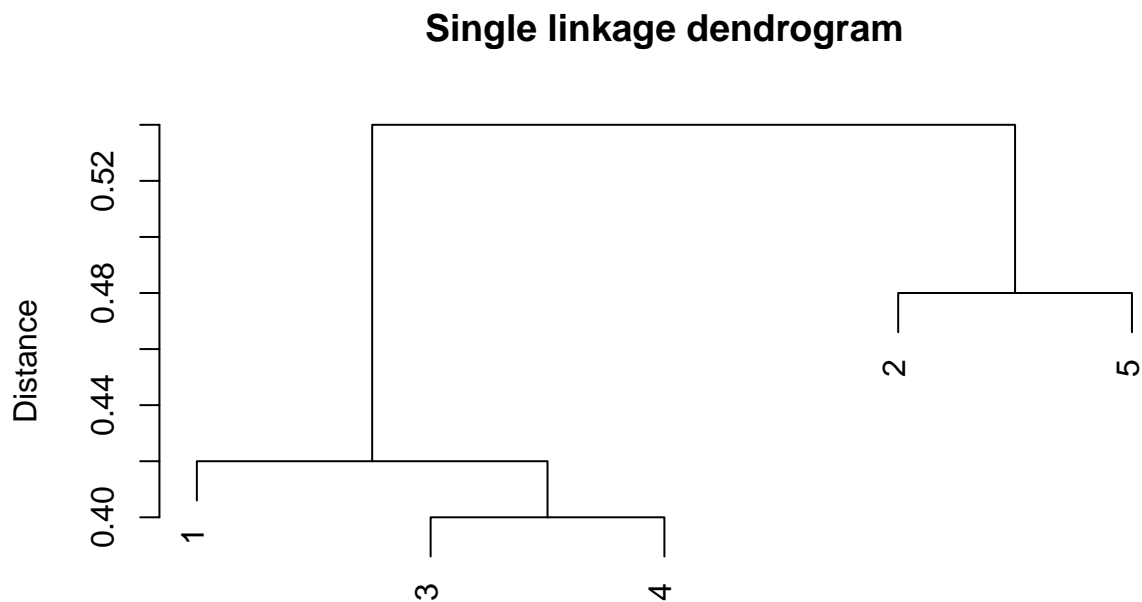## 2

(a)

```
cor=matrix(rep(1, 25), nrow=5)
cor[lower.tri(cor)]=c(0.39, 0.51, 0.58, 0.46, 0.44, 0.38, 0.52,
                      0.60, 0.43, 0.32)
cor[upper.tri(cor, diag = F)]=t(cor)[upper.tri(cor, diag = F)]
dist=as.dist(1-cor)

hclust(dist,method="single")

##
## Call:
## hclust(d = dist, method = "single")
```

4

```
## 
## Cluster method   : single
## Number of objects: 5
```

```
plot(hclust(dist,method="single"),
     xlab=NA,ylab="Distance",
     main="Single linkage dendrogram")
```

Single linkage dendrogram
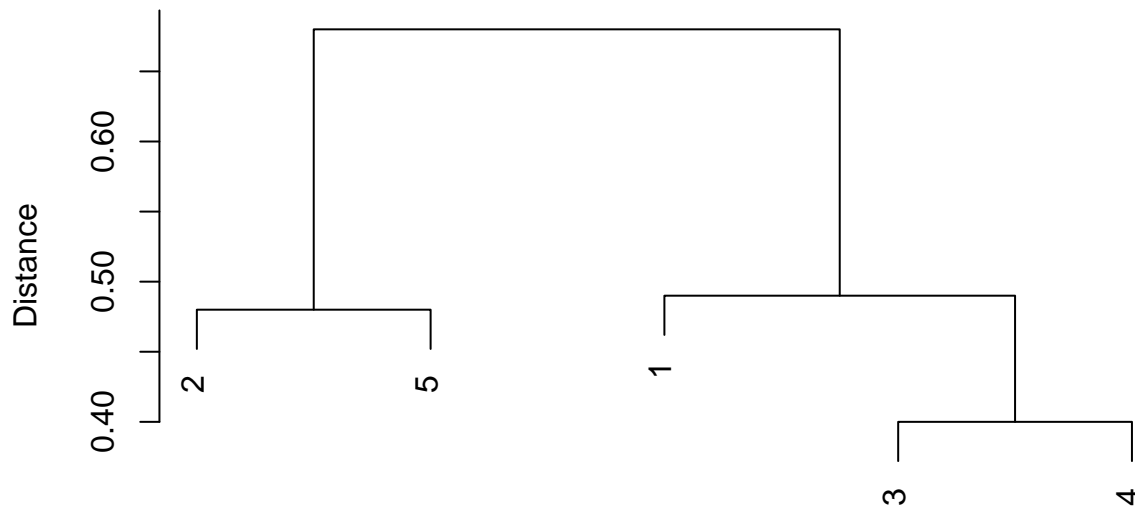

hclust (*, "single")

(b)

```
hclust(dist,method="complete")
```

```
## 
## Call:
## hclust(d = dist, method = "complete")
## 
## Cluster method   : complete
## Number of objects: 5
```

```
plot(hclust(dist,method="complete"),
     xlab=NA,ylab="Distance",
     main="Complete linkage dendrogram")
```
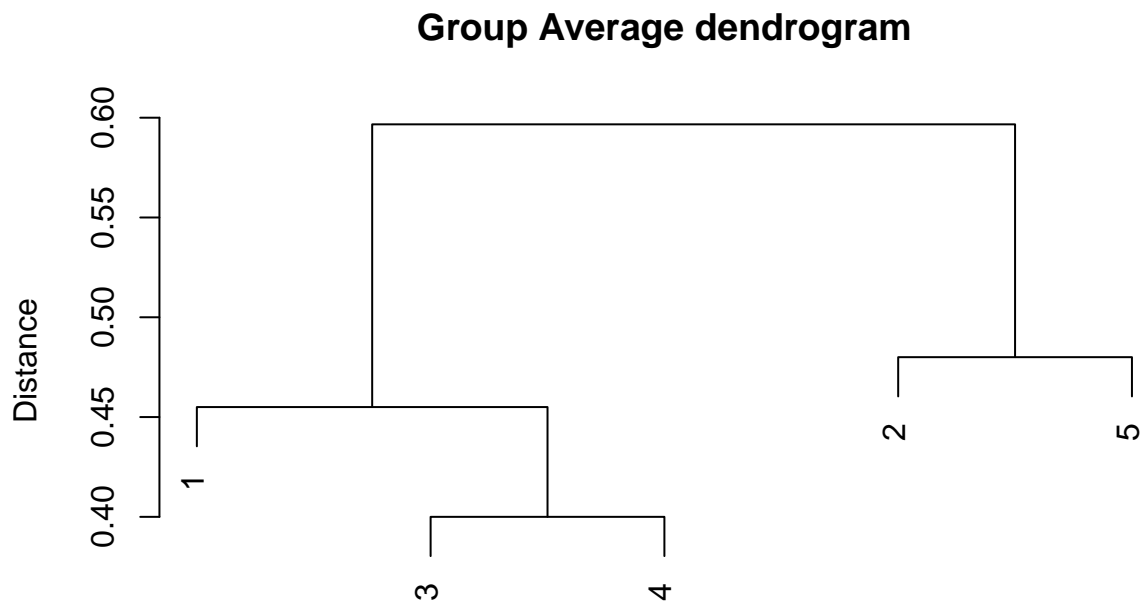
5

Complete linkage dendrogram

hclust (*, "complete")

(c)

```r
hclust(dist,method="average")
```

```
##
## Call:
## hclust(d = dist, method = "average")
##
## Cluster method   : average
## Number of objects: 5
```

```r
plot(hclust(dist,method="average"),
     xlab=NA,ylab="Distance",
     main="Group Average dendrogram")
```

Group Average dendrogram

hclust (*, "average")

(d)

The partitions are the same when using single linkage, complete linkage group average cluster-ing, but the distances used to merge groups are different.

3

(a)

```r
dat3=matrix(c(5, 4, 1, 3, -1, 1, 3, 1), ncol=2, byrow = T)

km1=kmeans(dat3, matrix(c(3, 3.5, 1, 1), ncol=2, byrow = T))
km1
```

```
## K-means clustering with 2 clusters of sizes 2, 2
##
## Cluster means:
##   [,1] [,2]
## 1    4  2.5
## 2    0  2.0
##
## Clustering vector:
```

```
## [1] 1 2 2 1
##
## Within cluster sum of squares by cluster:
## [1] 6.5 4.0
##  (between_SS / total_SS =  60.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

(b)

```r
km2=kmeans(dat3, matrix(c(4, 2.5, 0, 2), ncol=2, byrow = T))
km2
```

```
## K-means clustering with 2 clusters of sizes 2, 2
##
## Cluster means:
##    [,1] [,2]
## 1     4  2.5
## 2     0  2.0
##
## Clustering vector:
## [1] 1 2 2 1
##
## Within cluster sum of squares by cluster:
## [1] 6.5 4.0
##  (between_SS / total_SS =  60.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```r
plot(dat3, col=km2$cluster, cex=2, pch=1, lwd=2)
points(km2$centers, col=1:2, pch=8)
```
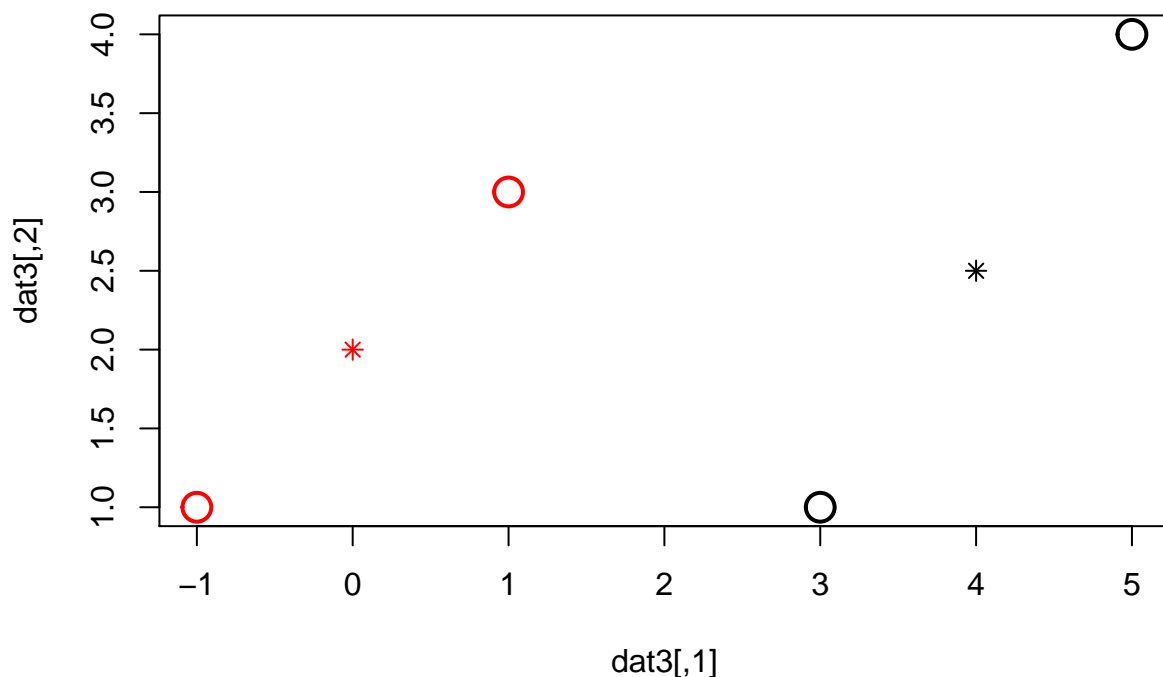
The clustering are the same. Observation 1 and 4 form one cluster and observation 2 and 3 form another. The within cluset SS is 6.5 and 4.0 respectively and the ratio of the between SS and total SS is 60.7 which indicates that the goodness of the classification is not that bad.
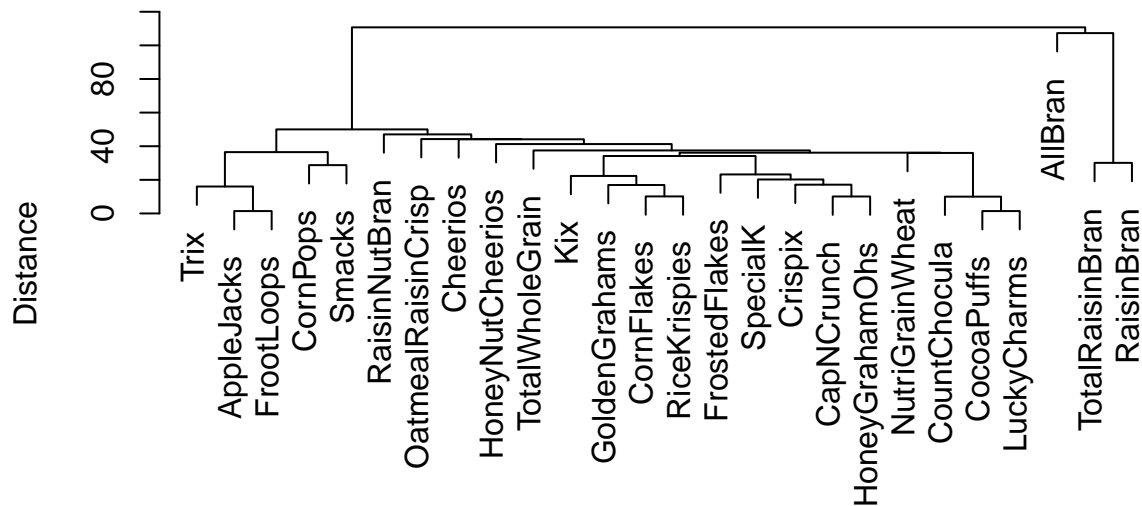
4

(a)

```
dat4=read.table("C:/Users/limj/Downloads/cereal_r.dat",
                header=T)
hclust(dist(dat4[, -c(1, 2)]), method="single")
```

```
##
## Call:
## hclust(d = dist(dat4[, -c(1, 2)]), method = "single")
##
## Cluster method   : single
## Distance         : euclidean
## Number of objects: 26
```

```
plot(hclust(dist(dat4[, -c(1, 2)]), method="single"),
     labels=dat4[,1], xlab=NA, ylab="Distance",
     main="Single linkage dendrogram")
```
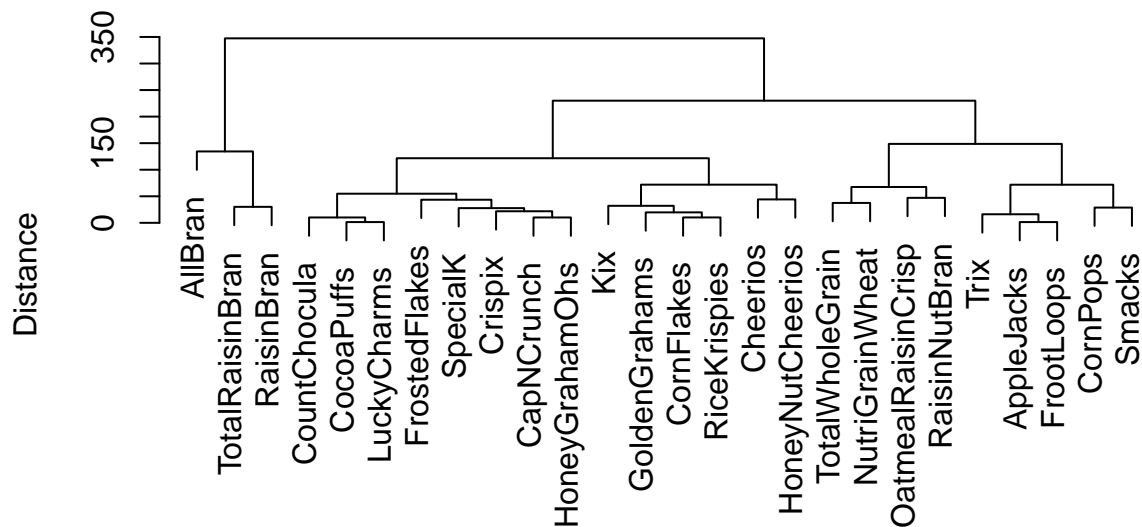
Single linkage dendrogram

hclust (*, "single")

```r
hclust(dist(dat4[, -c(1, 2)]), method="complete")
```

```
##
## Call:
## hclust(d = dist(dat4[, -c(1, 2)]), method = "complete")
##
## Cluster method   : complete
## Distance         : euclidean
## Number of objects: 26
```

```r
plot(hclust(dist(dat4[, -c(1, 2)]), method="complete"),
     labels=dat4[,1], xlab=NA, ylab="Distance",
     main="Complete linkage dendrogram")
```
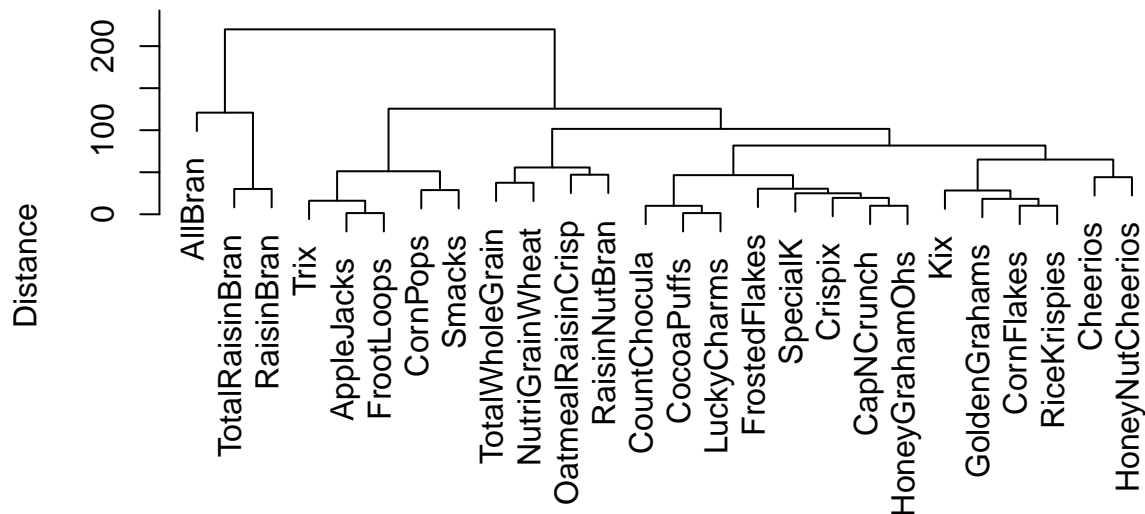
Complete linkage dendrogram

hclust (*, "complete")

```r
hclust(dist(dat4[, -c(1, 2)]), method="average")
```

```
##
## Call:
## hclust(d = dist(dat4[, -c(1, 2)]), method = "average")
##
## Cluster method   : average
## Distance         : euclidean
## Number of objects: 26
```

```r
plot(hclust(dist(dat4[, -c(1, 2)]), method="average"),
     labels=dat4[,1], xlab=NA, ylab="Distance",
     main="Group Average dendrogram")
```

Group Average dendrogram

hclust (*, "average")

The partitions are the same but the distance but the distances used to merge groups are different.

(b)

```
km3=kmeans(dat4[, -c(1, 2)], 3)
km3
```

```
## K-means clustering with 3 clusters of sizes 11, 3, 12
##
## Cluster means:
##    Calories  Protein       Fat Sodium     Fiber Carbohydrates      Sugar
## 1 110.9091 2.454545 0.8181818  250.0 0.7727273      16.59091   6.181818
## 2 110.0000 3.333333 1.0000000  220.0 6.0000000      12.00000  10.333333
## 3 108.3333 2.000000 0.9166667  147.5 1.1666667      12.58333  10.583333
##   Potassium
## 1  49.09091
## 2 263.33333
## 3  65.00000
##
## Clustering vector:
##  [1] 1 3 3 1 1 1 3 3 3 2 3 3 2 3 1 3 1 3 1 3 2 1 3 1 1 1
##
```

```
## Within cluster sum of squares by cluster:
## [1] 18581.91 10164.00 36827.08
##  (between_SS / total_SS =  72.9 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"       "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"        "ifault"
```

3 clusters formed from single linkage, complete linkage, group average, k-means clustering are the same.