Clustering

Issues

The dataset called "USArrests" provides information on five characteristics for each state in the United States. These characteristics are the state name, murder rate, assault rate, urban population percentage (which measures the proportion of the population living in urban areas), and rape rate.

Using the USArrests data addressed the following:

- A principal component analysis, including a discussion of the interpretation of the principal components.
- A clustering of the data, using k-means clustering for suitable k.
- A hierarchical clustering of the data, with interpretations of the clusters in the hierarchy.

Findings

The results of Principal Component Analysis (PCA) indicate that there is a substantial association between serious crimes and the degree of urbanization in the first loading vector, but a weaker correlation in the second loading vector. Accordingly, murder, assault, and rape are crimes that frequently occur together in states, but there is minimal connection between these crimes and the urban population.

Using k-means clustering, we found that raising the value of "k" causes a decrease in the "tot.withinss," demonstrating that the algorithm has successfully found clusters that are homogeneous and well-separated from one another.

Furthermore, the application of hierarchical clustering led to the development of dendrograms, which are tree-like structures. When complete and average linkage methods were used, the dendrograms were more evenly distributed.

Discussions

In principal component analysis, We have seen that the first principal component accounts for 62.0% of the data variance, the second principle component for 24.7%, and so on.

Code from hierarchical clustering produced balanced dendrograms from complete and average linkage, but from single linkage it produced a bad dendrogram.

Appendix A: Method

Using the PCA technique, a dataset's dimensionality can be reduced by spotting patterns in the data. We first compute the mean and variance of the dataset's variables before using PCA. We do PCA in R using the prcomp() function if the variables' mean and variance values differ. The center, scale, rotation, sdev, and x values are provided by doing this. The variables' mean and standard deviation are shown by the scale's center and length. Using the biplot() function, the first two principal components can then be plotted.

Additionally, we can figure out the variance and the percentage of variance that each primary component explains. Use the plot() method to visualize the share of variation that each component contributes to.

We begin with a matrix-formatted dataset that has been divided into two equal halves to perform k-means clustering. The data is clustered using R's tool with function kmeans(). After that, plot the data, giving each observation a color according on which cluster it belongs to. This procedure is repeated with k=3 and the tot.withinss value—a metric for the sum of all within-cluster squares—is noted. We make inferences about the effectiveness of the clustering algorithm based on our observations.

Appendix B: Results

```
> pr.out <- prcomp (data , scale = TRUE)
> names (pr.out)
[1] "sdev"     "rotation" "center"   "scale"    "x"
>
> pr.out$center
  Murder  Assault UrbanPop     Rape
   7.788  170.760   65.540   21.232
> pr.out$scale
   Murder   Assault  UrbanPop      Rape
 4.355510 83.337661 14.474763  9.366385
> pr.out$rotation
                PC1        PC2        PC3         PC4
Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
> dim (pr.out$x)
[1] 50  4
```

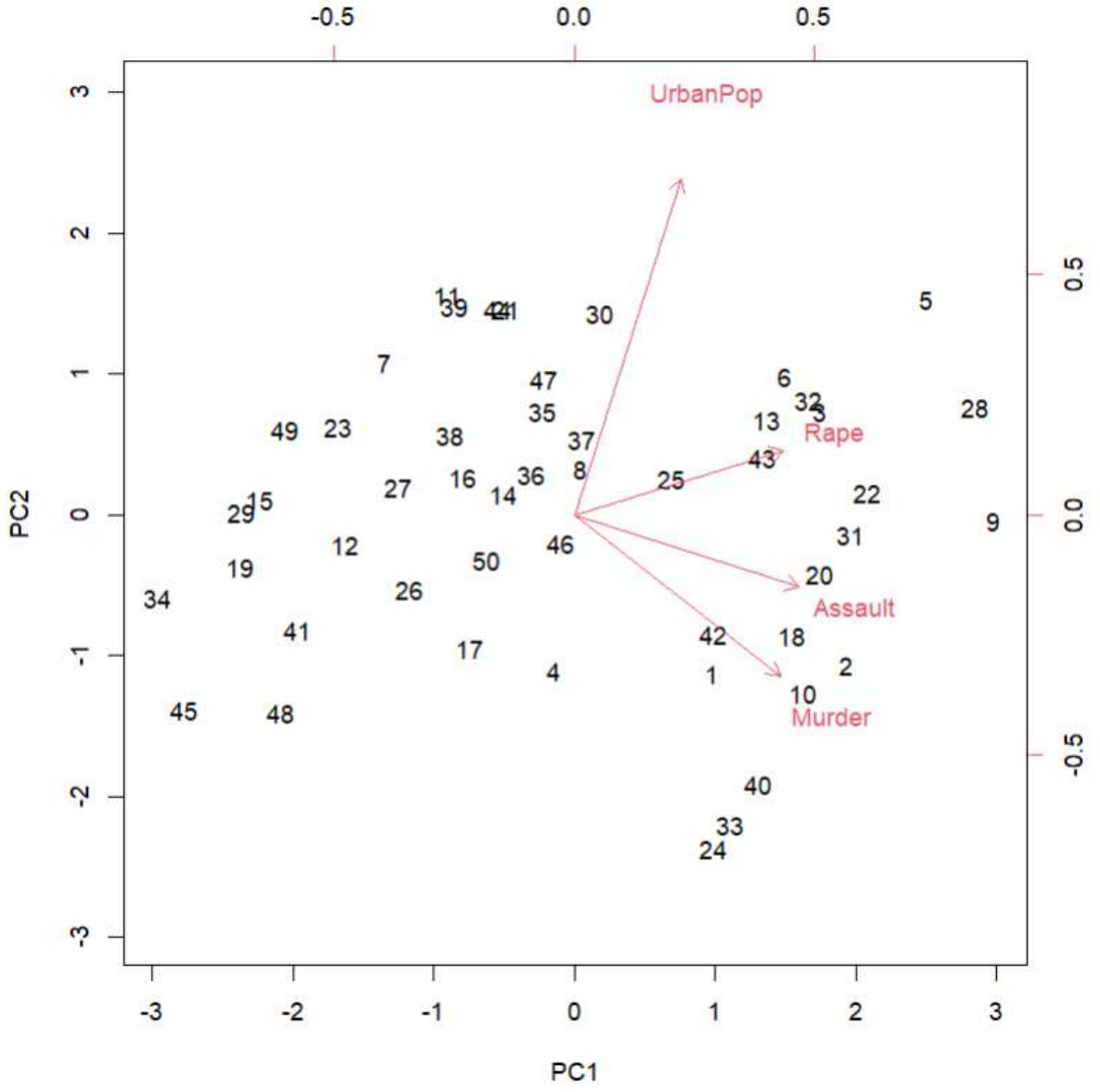

Figure 1. Biplot for first two principal components

```
> pr.var <- pr.out$sdev^2
> pr.var
[1] 2.4802416 0.9897652 0.3565632 0.1734301
>
> pve <- pr.var / sum (pr.var)
> pve
[1] 0.62006039 0.24744129 0.08914080 0.04335752
```

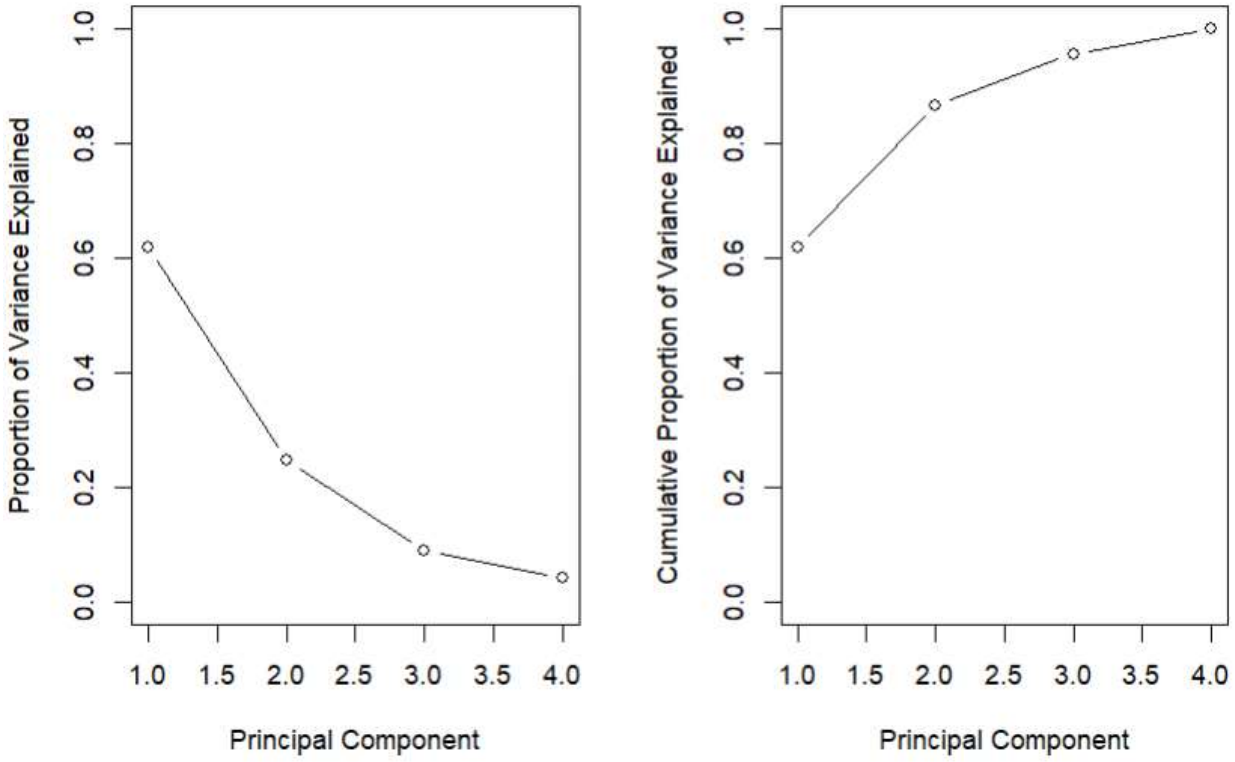

Figure 2. Variance plots

k-means clustering for k=2, k=3.

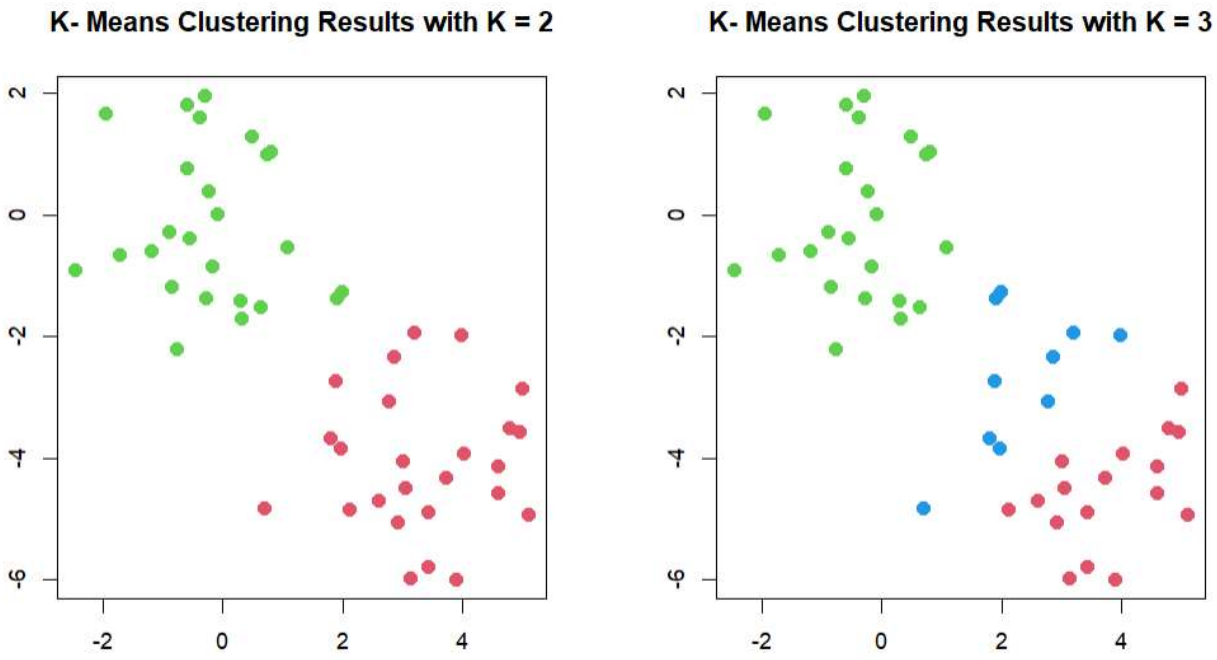

Figure 3. k-means clustering for k=2, k=3.

```
> km.out <- kmeans (x, 3, nstart = 20)
> km.out
K-means clustering with 3 clusters of sizes 17, 10, 23

Cluster means:
        [,1]        [,2]
1  3.7789567 -4.56200798
2  2.3001545 -2.69622023
3 -0.3820397 -0.08740753

Clustering vector:
 [1] 1 2 1 2 1 1 1 2 1 2 1 2 1 2 1 2 1 1 1 1 1 2 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 2 3 2 3 3 3 3

Within cluster sum of squares by cluster:
[1] 25.74089 19.56137 52.67700
 (between_SS / total_SS =  79.3 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "b
etweenss"
[7] "size"         "iter"         "ifault"
> km.out$tot.withinss
[1] 97.97927
```

```
> km.out$withinss
[1] 25.74089 19.56137 52.67700
> set.seed (4)
> km.out <- kmeans (x, 3, nstart = 20)
> km.out
K-means clustering with 3 clusters of sizes 17, 23, 10

Cluster means:
         [,1]        [,2]
1  3.7789567 -4.56200798
2 -0.3820397 -0.08740753
3  2.3001545 -2.69622023

Clustering vector:
 [1] 1 3 1 3 1 1 1 3 1 3 1 3 1 3 1 3 1 1 1 1 1 3 1 1 1 2 2 2 2 2 2 2 2 2 2 2
[37] 2 2 2 2 2 2 2 3 2 3 2 2 2 2

Within cluster sum of squares by cluster:
[1] 25.74089 52.67700 19.56137
 (between_SS / total_SS =  79.3 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
```

Hierarchical clustering

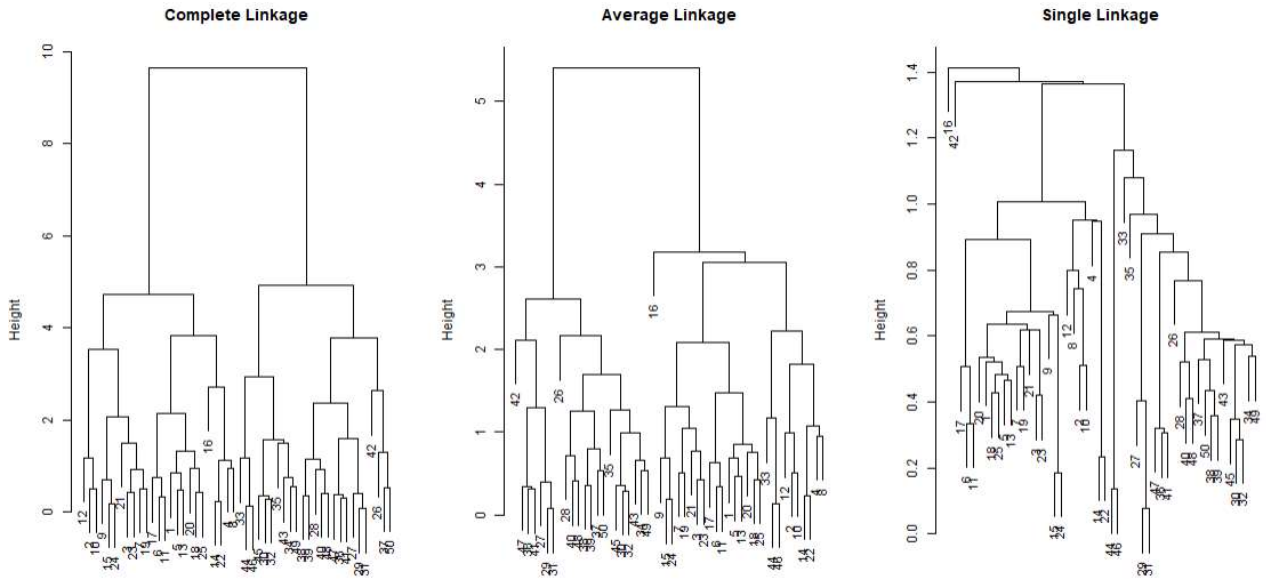

Figure 4. Hierarchical clustering performing on complete, average and single linkage

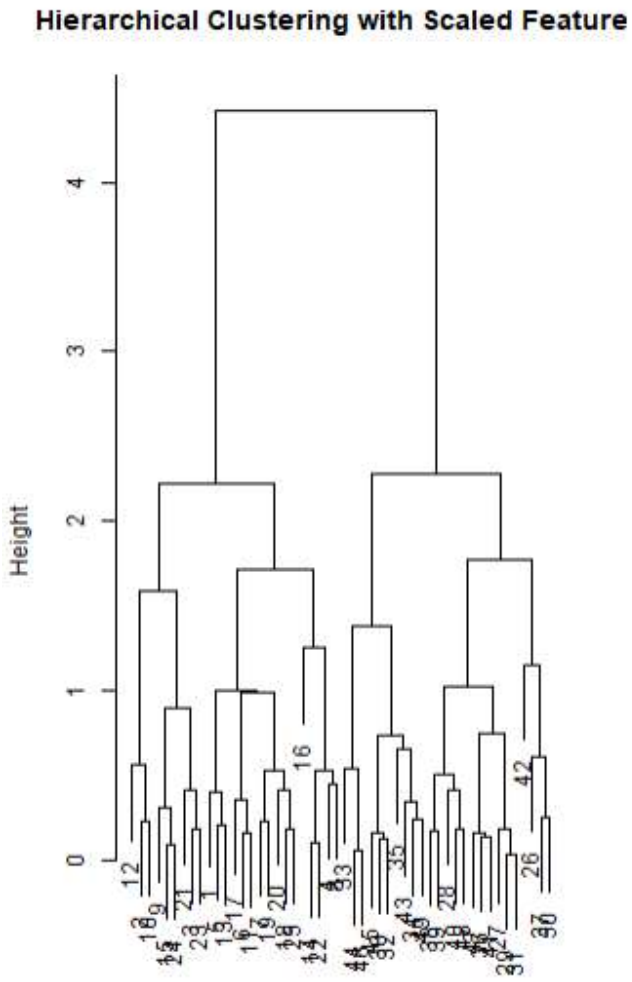

Figure 5. Hierarchical clustering with scaled features

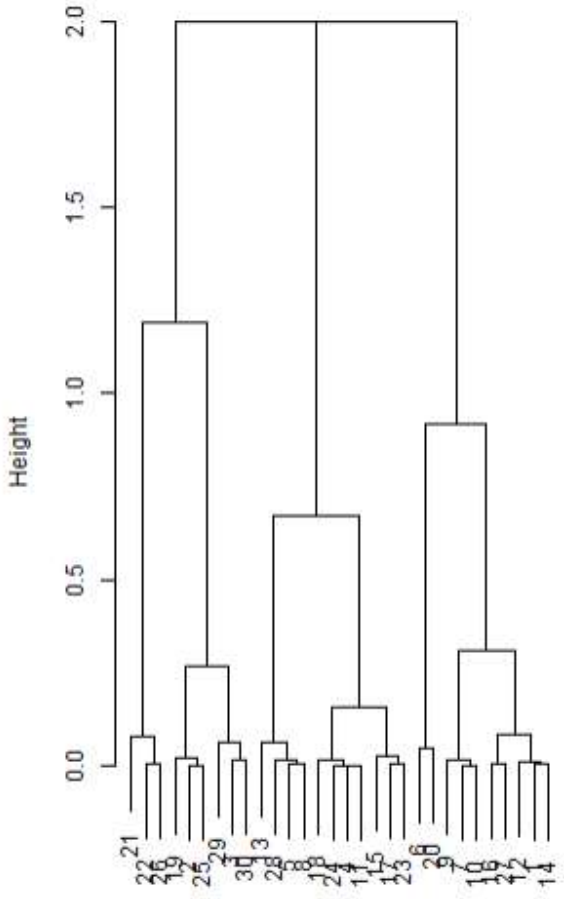

Figure 6. Complete linkage with correlation-based distance.

Appendix C: Code

PCA code

```
install.packages('readxl')

library(readxl)

file <-"E:\\Assignments\\MTH 522\\Project 4\\Project 4.xlsx"

data <- read_excel(file, sheet = 1)

data
```

```
str(data)
summary(data)
colnames(data)
sapply(data, class)
data <- data.frame(data)
sapply(data, class)
data[,1] <- as.numeric(data[,1])
sapply(data, class)
states<-row.names(data)
states
names(data)
rownames(data)
ncol(data)
nrow(data)

apply (data , 2, mean)
apply (data , 2, var)
data <- data[, -1]
pr.out <- prcomp (data , scale = TRUE)
names (pr.out)

pr.out$center
```

```
pr.out$scale
pr.out$rotation
dim (pr.out$x)
biplot (pr.out , scale = 0)

pr.out$rotation = -pr.out$rotation
pr.out$x = -pr.out$x
biplot (pr.out , scale = 0)
pr.out$sdev



pr.var <- pr.out$sdev^2
pr.var

pve <- pr.var / sum (pr.var)
pve



par (mfrow = c(1, 2))
plot (pve , xlab = " Principal Component ",
     ylab = " Proportion of Variance Explained ", ylim = c(0, 1),
     type = "b")
```

```
plot ( cumsum (pve), xlab = " Principal Component ",
      ylab = " Cumulative Proportion of Variance Explained ",
      ylim = c(0, 1), type = "b")

a <- c(1, 2, 8, -3)
cumsum (a)
```

Code for K-meams and hierarchical clustering

```
install.packages('readxl')
library(readxl)
file <-"E:\\Assignments\\MTH 522\\Project 4\\Project 4.xlsx"
data <- read_excel(file, sheet = 1)
data

set.seed (2)
x <- matrix ( rnorm (50 * 2), ncol = 2)
x[1:25, 1] <- x[1:25, 1] + 3
x[1:25, 2] <- x[1:25, 2] - 4

km.out <- kmeans(x, 2, nstart = 20)
km.out
```

```
km.out$cluster



par (mfrow = c(1, 2))
plot (x, col = (km.out$cluster + 1),
      main = "K- Means Clustering Results with K = 2",
      xlab = "", ylab = "", pch = 20, cex = 2)

set.seed (4)
km.out <- kmeans (x, 3, nstart = 20)
km.out

plot (x, col = (km.out$cluster + 1),
     main = "K- Means Clustering Results with K = 3",
     xlab = "", ylab = "", pch = 20, cex = 2)
km.out$tot.withinss
km.out$withinss
```

```
hc.complete <- hclust ( dist (x), method = "complete")
hc.average <- hclust ( dist (x), method = "average")
hc.single <- hclust ( dist (x), method = "single")

par (mfrow = c(1, 3))
plot (hc.complete, main = " Complete Linkage ",
     xlab = "", sub = "", cex = .9)

plot (hc.average , main = " Average Linkage ",
     xlab = "", sub = "", cex = .9)

plot (hc.single, main = " Single Linkage ",
     xlab = "", sub = "", cex = .9)

cutree (hc.complete, 2)
cutree (hc.average , 2)
cutree (hc.single, 2)
cutree (hc.single, 4)

xsc <- scale (x)
plot ( hclust ( dist (xsc), method = "complete") ,
       main = " Hierarchical Clustering with Scaled Features ")
```

```
x <- matrix ( rnorm (30 * 3), ncol = 3)
dd <- as.dist (1 - cor (t(x)))
plot ( hclust (dd, method = "complete") ,
      main = " Complete Linkage with Correlation - Based Distance ",
      xlab = "", sub = "")
```