Hierarchical Clustering

USArrests dataset

```
# husarrests.r
#
df0 <- USArrests
str(df0)
```

```
## 'data.frame':    50 obs. of  4 variables:
##  $ Murder  : num  13.2 10 8.1 8.8 9 7.9 3.3 5.9 15.4 17.4 ...
##  $ Assault : int  236 263 294 190 276 204 110 238 335 211 ...
##  $ UrbanPop: int  58 48 80 50 91 78 77 72 80 60 ...
##  $ Rape    : num  21.2 44.5 31 19.5 40.6 38.7 11.1 15.8 31.9 25.8 ...
```

```
head(df0)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

```
#
# scale dataframe
#
df <- scale(df0)
head(df)
```

```
##                Murder   Assault   UrbanPop         Rape
## Alabama    1.24256408 0.7828393 -0.5209066 -0.003416473
## Alaska     0.50786248 1.1068225 -1.2117642  2.484202941
## Arizona    0.07163341 1.4788032  0.9989801  1.042878388
## Arkansas   0.23234938 0.2308680 -1.0735927 -0.184916602
## California 0.27826823 1.2628144  1.7589234  2.067820292
## Colorado   0.02571456 0.3988593  0.8608085  1.864967207
```

```
#
# find distances
#
distance = dist(df)
head(distance)
```

```
## [1] 2.703754 2.293520 1.289810 3.263110 2.651067 3.215297
```

```
length(distance)
```

```
## [1] 1225
```

```
#
#
# distance in a matrix display
#
distmat = as.matrix(distance)
dim(distmat)
```

```
## [1] 50 50
```

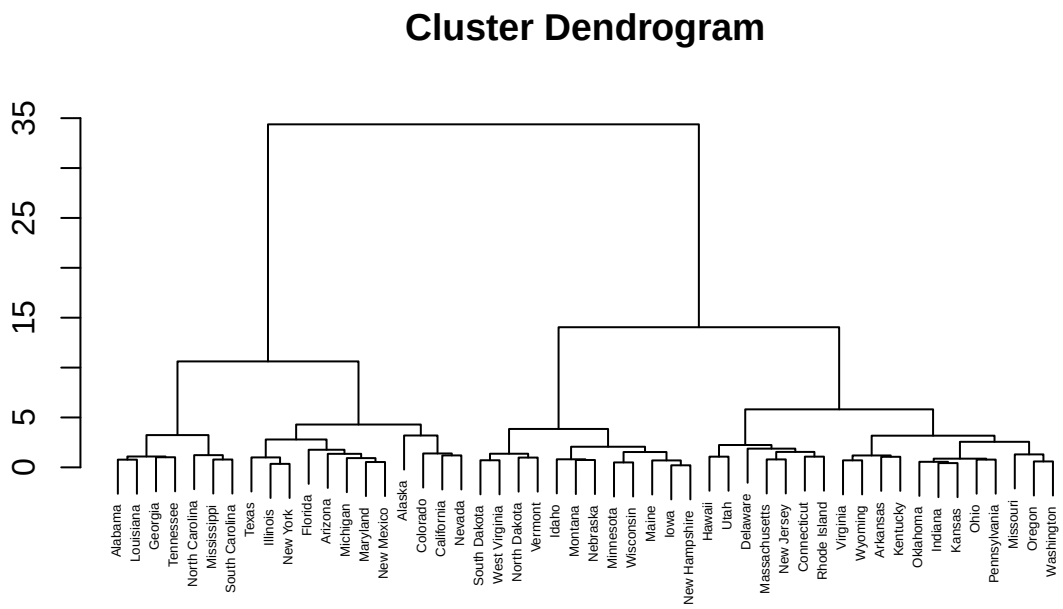
```
distmat[1:7,1:7]
```

```
##              Alabama   Alaska  Arizona Arkansas California Colorado Connecticut
## Alabama     0.000000 2.703754 2.293520 1.289810   3.263110 2.651067    3.215297
## Alaska      2.703754 0.000000 2.700643 2.826039   3.012541 2.326519    4.739912
## Arizona     2.293520 2.700643 0.000000 2.717758   1.310484 1.365031    3.262858
## Arkansas    1.289810 2.826039 2.717758 0.000000   3.763641 2.831051    2.607639
## California  3.263110 3.012541 1.310484 3.763641   0.000000 1.287619    4.066390
## Colorado    2.651067 2.326519 1.365031 2.831051   1.287619 0.000000    3.327992
## Connecticut 3.215297 4.739912 3.262858 2.607639   4.066390 3.327992    0.000000
```

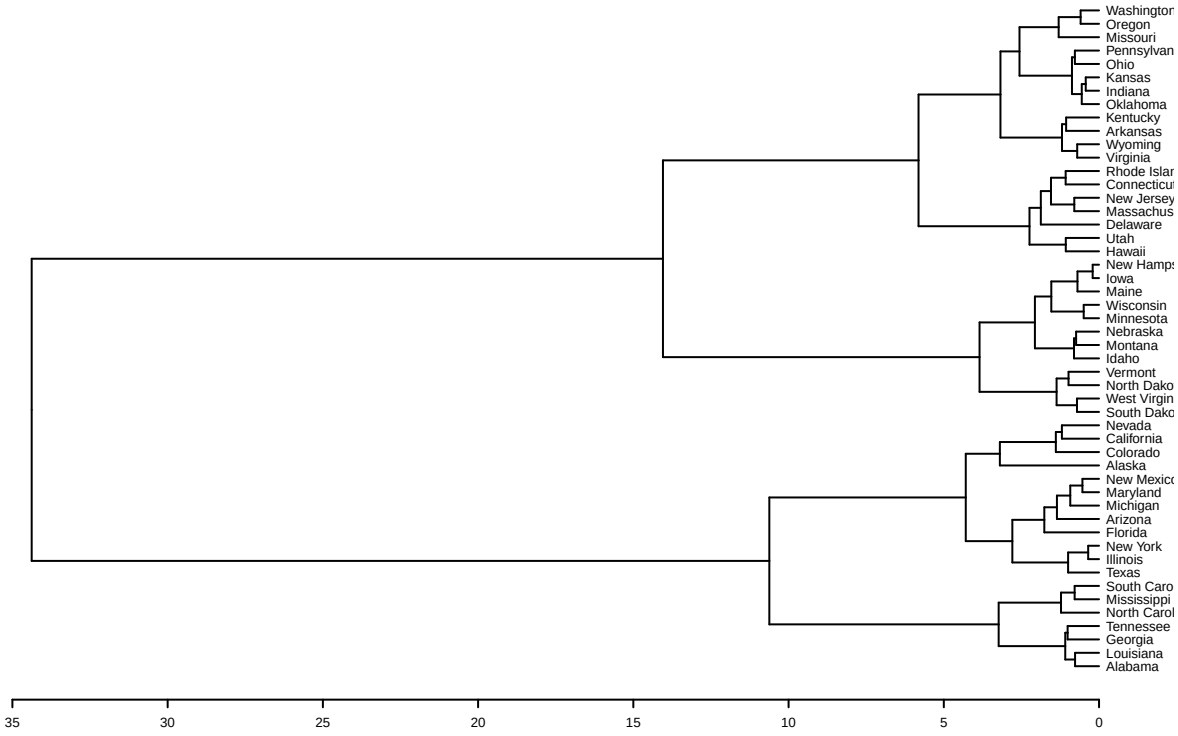
```
#
# Hierarchical clustering - Ward
#
h1 = hclust(distance, method ='ward.D')
str(h1)
```

```
## List of 7
##  $ merge      : int [1:49, 1:2] -15 -13 -14 -23 -20 -36 -37 -19 -46 -41 ...
##  $ height     : num [1:49] 0.206 0.35 0.429 0.494 0.535 ...
##  $ order      : int [1:50] 1 18 10 42 33 24 40 43 13 32 ...
##  $ labels     : chr [1:50] "Alabama" "Alaska" "Arizona" "Arkansas" ...
##  $ method     : chr "ward.D"
##  $ call       : language hclust(d = distance, method = "ward.D")
##  $ dist.method: chr "euclidean"
##  - attr(*, "class")= chr "hclust"
```

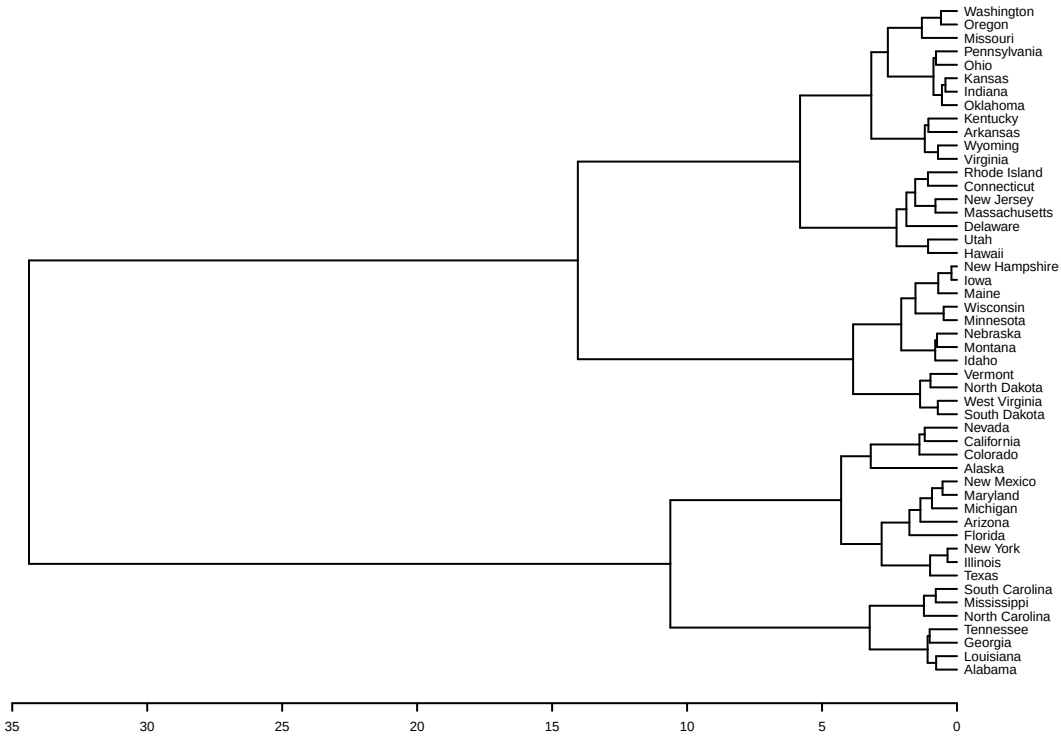
```
#
# Dendrogram - Ward
#
plot(h1,cex=0.4,xlab="",sub="",ylab="")
```




```
#
par(cex=0.4)
plot(as.dendrogram(h1),horiz=T)
```




```
#
par(cex=0.4, mar=c(5, 8, 4, 8))
plot(as.dendrogram(h1),horiz=T)
```




```
#
# CUT the dendrograms to 4 clusters
#
cut1 = cutree(h1,k=4)
head(cut1)
```

```
##    Alabama     Alaska    Arizona   Arkansas California   Colorado
##          1          2          2          3          2          2
```

```
#
# dataframe with cluster numbers
df1 = data.frame(df,cluster = cut1)
head(df1)
```

```
##               Murder   Assault   UrbanPop         Rape cluster
## Alabama    1.24256408 0.7828393 -0.5209066 -0.003416473       1
## Alaska     0.50786248 1.1068225 -1.2117642  2.484202941       2
## Arizona    0.07163341 1.4788032  0.9989801  1.042878388       2
## Arkansas   0.23234938 0.2308680 -1.0735927 -0.184916602       3
## California 0.27826823 1.2628144  1.7589234  2.067820292       2
## Colorado   0.02571456 0.3988593  0.8608085  1.864967207       2
```

```
#
# number of members per cluster
#
table(cut1)
```

```
## cut1
##  1  2  3  4
##  7 12 19 12
```

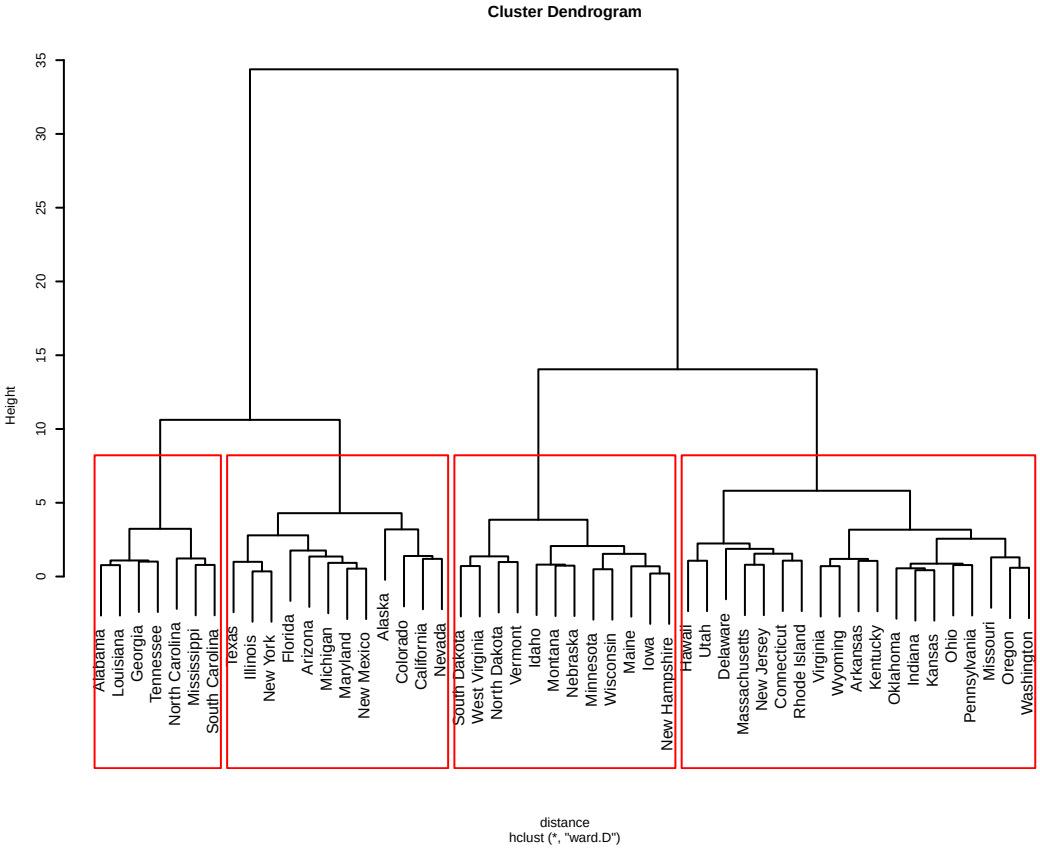
```
#
# members of cluster 1 (use rownames of original dataframe)
#
rownames(df)[1:5]
```

```
## [1] "Alabama"    "Alaska"     "Arizona"    "Arkansas"   "California"
```

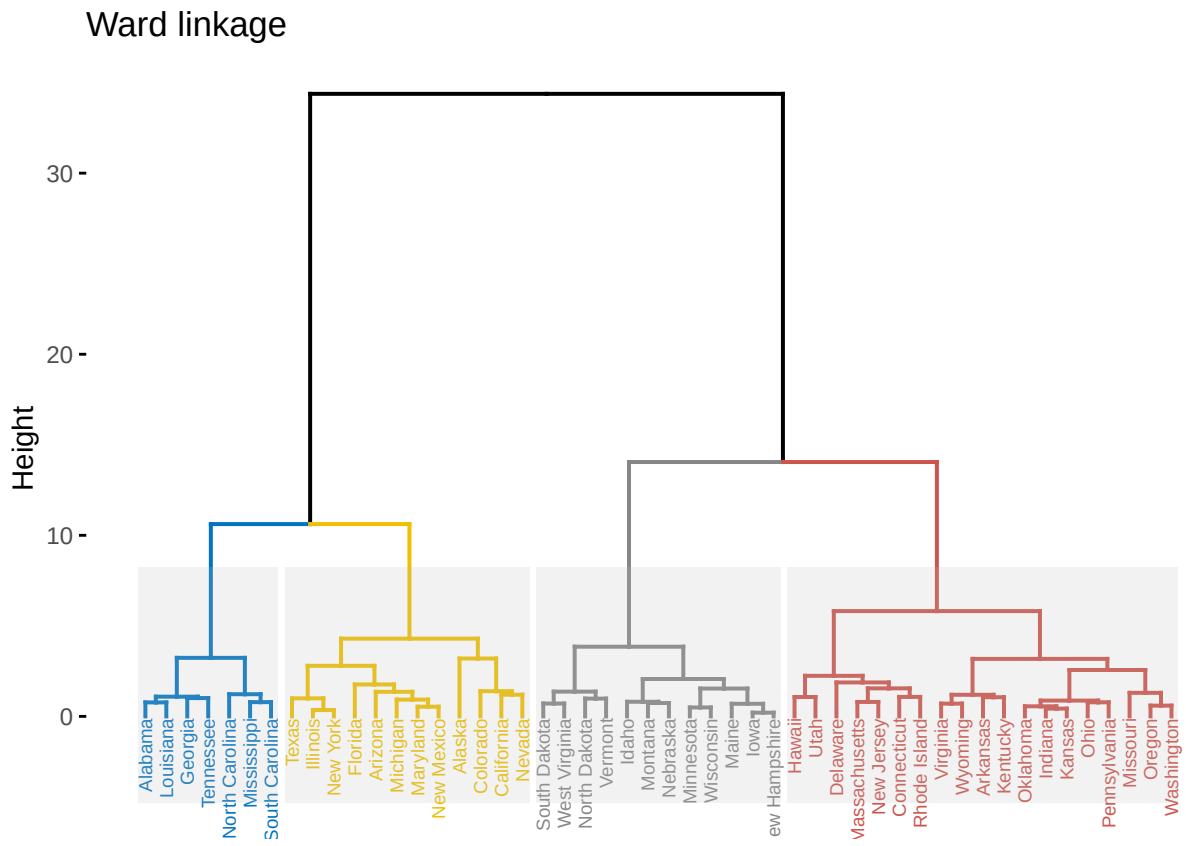
```
#
rownames(df)[cut1 == 1]
```

```
## [1] "Alabama"        "Georgia"        "Louisiana"      "Mississippi"
## [5] "North Carolina" "South Carolina" "Tennessee"
```

```
#
# dendrogram with cut - Ward linkage
#
plot(h1,cex=1.2)
rect.hclust(h1,k=4,border="red")
#
# library factoextra
#
library(factoextra)
```

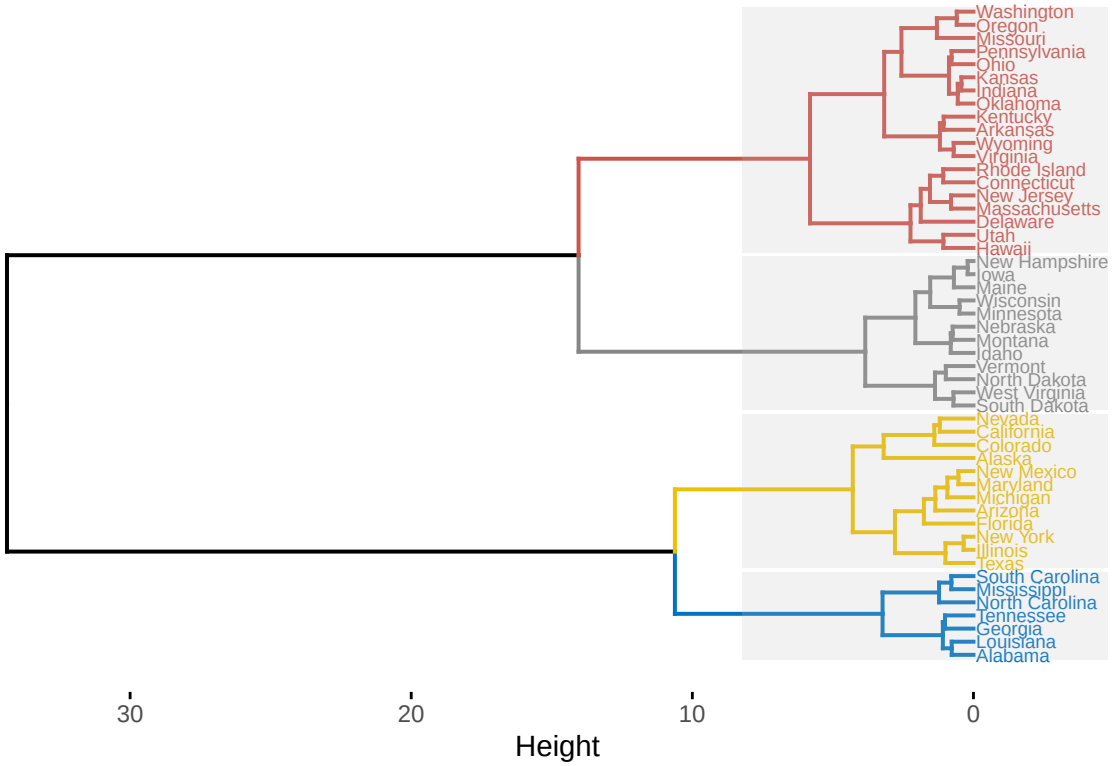


```
fviz_dend(h1,k=4,cex = 0.5, main="Ward linkage",
          k_colors = "jco", rect = T, rect_fill = T)
```

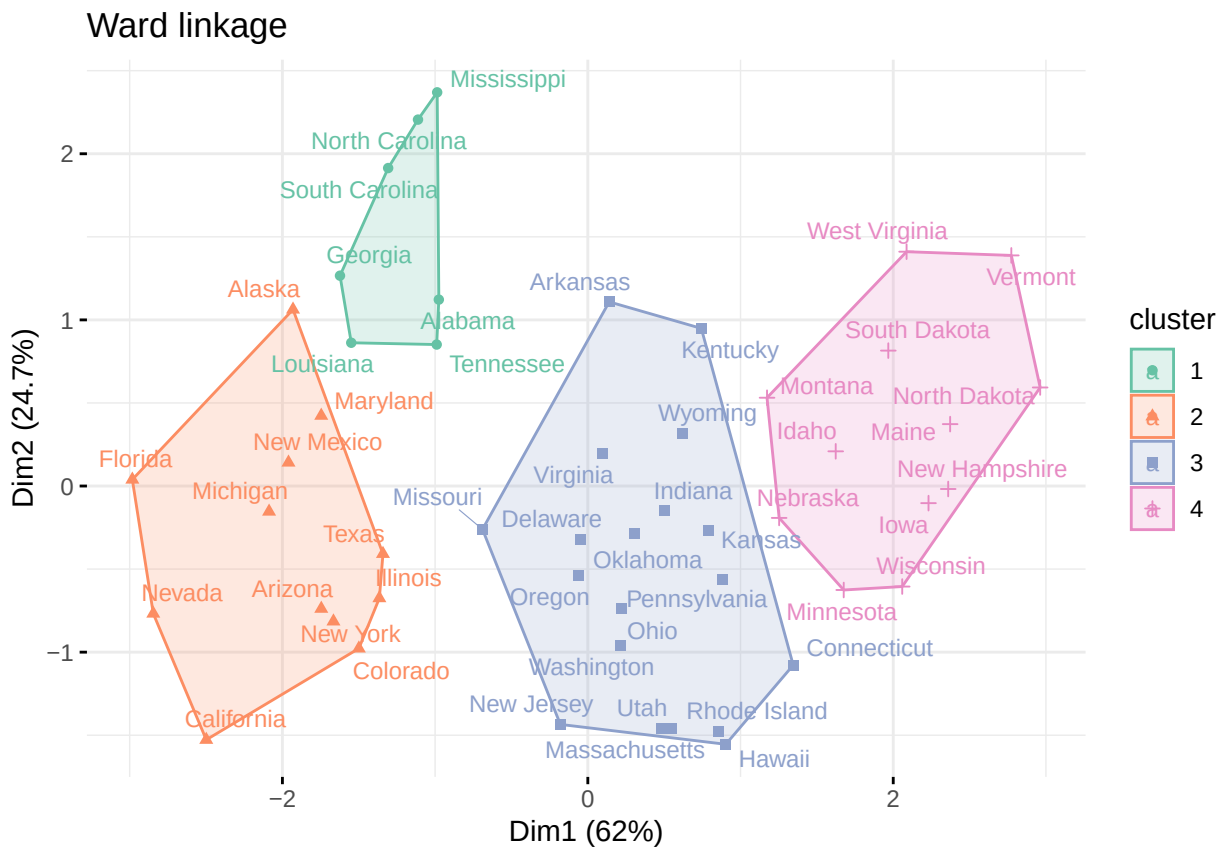



```
#
# dendrogram horizontal - Ward linkage
fviz_dend(h1,k=4,cex = 0.5, main="Ward linkage",horiz=T,
          k_colors = "jco", rect = T, rect_fill = T)
```

Ward linkage



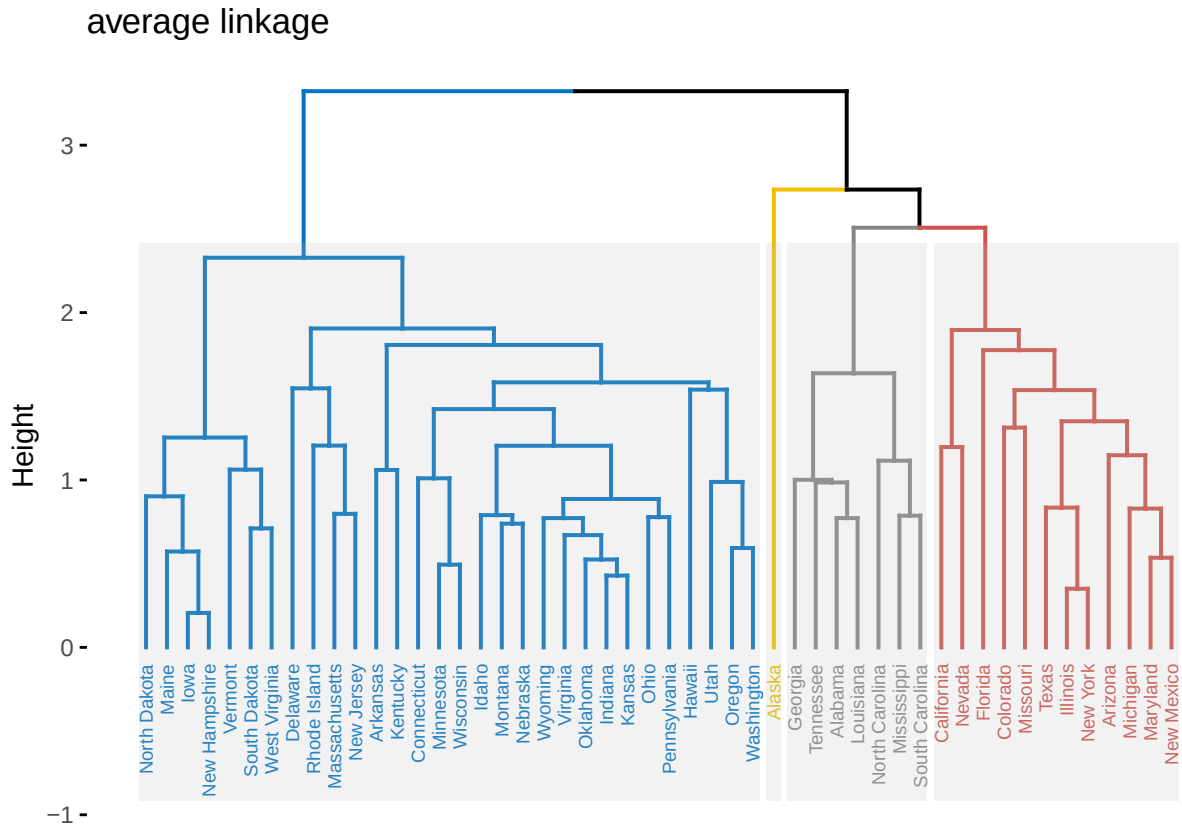

```
#
# scatterplot - Ward linkage
#
fviz_cluster(list(data = df, cluster = cut1),main="Ward linkage",
             palette = "Set2",show.clust.cent = F, labelsize = 10,
             repel = T, # Avoid label overplotting (slow)
             ggtheme = theme_minimal()
             )
```



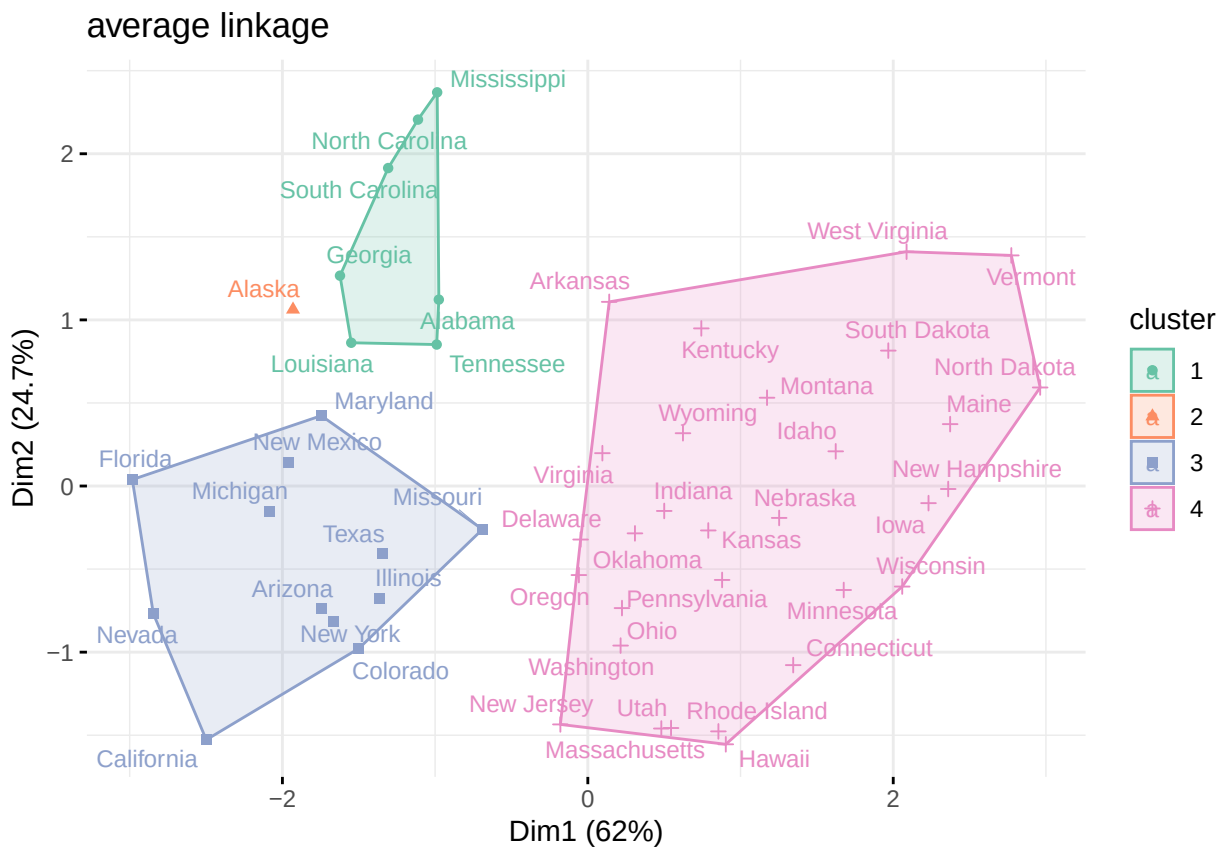

```
#
# cut dendrogram with average linkage
#
h2 = hclust(distance, method ='average')
cut2 = cutree(h2,k=4)
table(cut2)
```

```
## cut2
##  1  2  3  4
##  7  1 12 30
```

```
#
# dendrogram - average linkage
#
fviz_dend(h2,k=4,cex = 0.5,main="average linkage",
          k_colors = "jco", rect = T, rect_fill = T)
```




```
#
# scatterplot - average linkage
#
fviz_cluster(list(data = df, cluster = cut2),main="average linkage",
             palette = "Set2",show.clust.cent = F, labelsize = 10,
             repel = T, # Avoid label overlap (slow)
             ggtheme = theme_minimal()
             )
```




```
# COPHENETIC distances
c1 = cophenetic(h1)
#
# display them as a matrix
#
c1mat = as.matrix(c1)
c1mat[1:6,1:6]
```

```
##             Alabama    Alaska   Arizona Arkansas California  Colorado
## Alabama     0.00000 10.619899 10.619899 34.37957  10.619899 10.619899
## Alaska     10.61990  0.000000  4.292862 34.37957   3.194472  3.194472
## Arizona    10.61990  4.292862  0.000000 34.37957   4.292862  4.292862
## Arkansas   34.37957 34.379570 34.379570  0.00000  34.379570 34.379570
## California 10.61990  3.194472  4.292862 34.37957   0.000000  1.392043
## Colorado   10.61990  3.194472  4.292862 34.37957   1.392043  0.000000
```

```
#
# compare with Euclidean distances
#
distmat[1:6,1:6]
```

```
##             Alabama   Alaska  Arizona Arkansas California Colorado
## Alabama    0.000000 2.703754 2.293520 1.289810   3.263110 2.651067
## Alaska     2.703754 0.000000 2.700643 2.826039   3.012541 2.326519
## Arizona    2.293520 2.700643 0.000000 2.717758   1.310484 1.365031
```

```
## Arkansas   1.289810 2.826039 2.717758 0.000000   3.763641 2.831051
## California 3.263110 3.012541 1.310484 3.763641   0.000000 1.287619
## Colorado   2.651067 2.326519 1.365031 2.831051   1.287619 0.000000
```

```
#
# CPCC = correlation (euclidean distances, cophenetic distances)
#
cor(distance,c1)
```

```
## [1] 0.6844016
```

```
#
# Hierarchical clustering - average
#
c2 = cophenetic(h2)
cor(distance,c2)
```

```
## [1] 0.7180382
```

```
#
# average linkage distances are closer to euclidean distances
# than ward linkage distances
#
```