Hierarchical Clustering

In [1]:

```
# Load the data
data("USArrests")

# Standardize the data
df <- scale(USArrests)

# Show the first 6 rows
head(df, nrow = 6)
```

Out[1]:

A matrix: 6 × 4 of type dbl

	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

Similarity measures

In [2]:

```
# Compute the dissimilarity matrix
# df = the standardized data
res.dist <- dist(df, method = "euclidean")
```

To see easily the distance information between objects, we reformat the results of the function dist() into a matrix using the as.matrix() function.

In [3]:

```
as.matrix(res.dist)[1:6, 1:6]
```

Out[3]:

A matrix: 6 × 6 of type dbl

	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

Linkage

The linkage function takes the distance information, returned by the function dist(), and groups pairs of objects into clusters based on their similarity.

In [4]:

```
res.hc <- hclust(d = res.dist, method = "ward.D2")
```

Dendrogram

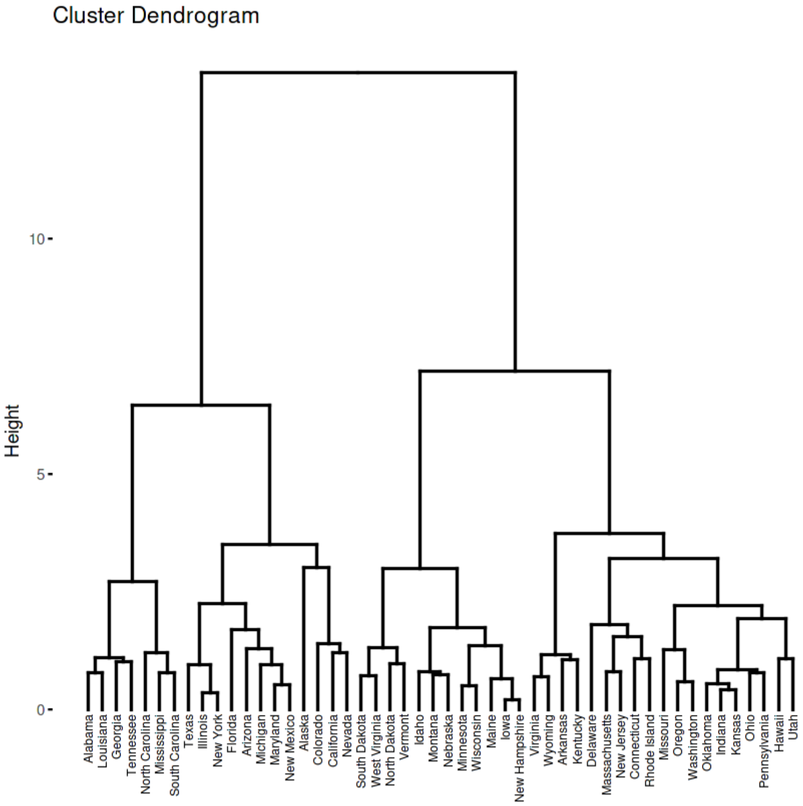
In [5]:

```
# cex: label size
library("factoextra")
fviz_dend(res.hc, cex = 0.5)
```

```
Loading required package: ggplot2

Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/
ve3WBa
```

Out[5]:



In [6]:

```
# Compute cophentic distance
res.coph <- cophenetic(res.hc)

# Correlation between cophenetic distance and
# the original distance
cor(res.dist, res.coph)
```

Out[6]: 0.697526563237039

In [7]:

```
res.hc2 <- hclust(res.dist, method = "average")

cor(res.dist, cophenetic(res.hc2))
```

Out[7]: 0.718038237932047

we see that,the average linkage method gives better correlation.

Cut the dendrogram into different groups

In [8]:

```
# Cut tree into 4 groups
grp <- cutree(res.hc, k = 4)
head(grp, n = 4)
```

Out[8]: **Alabama:** 1 **Alaska:** 2 **Arizona:** 2 **Arkansas:** 3

In [9]:

```
# Number of members in each cluster
table(grp)
```

Out[9]:

```
grp
 1  2  3  4 
 7 12 19 12 
```

In [10]:

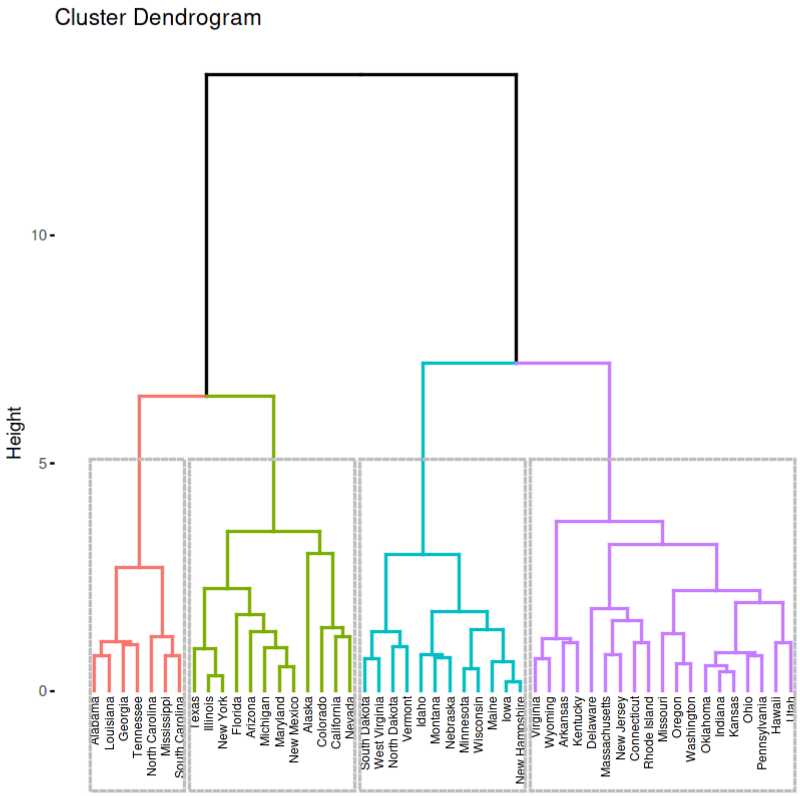
```
# Get the names for the members of cluster 1
rownames(df)[grp == 1]
```

Out[10]: 'Alabama' · 'Georgia' · 'Louisiana' · 'Mississippi' · 'North Carolina' · 'South Carolina' · 'Tennessee'

In [11]:

```
fviz_dend(res.hc, cex = 0.5, k = 4,
 color_labels_by_k = FALSE, rect = TRUE)
```

Out[11]:



Non Hierarchical Clustering

Computing K-means clustering

In [12]:

```
data("USArrests")       # Loading the data set
df <- scale(USArrests) # Scaling the data

# View the firt 3 rows of the data
head(df, n = 3)
```

Out[12]:

A matrix: 3 × 4 of type dbl

	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

In [13]:

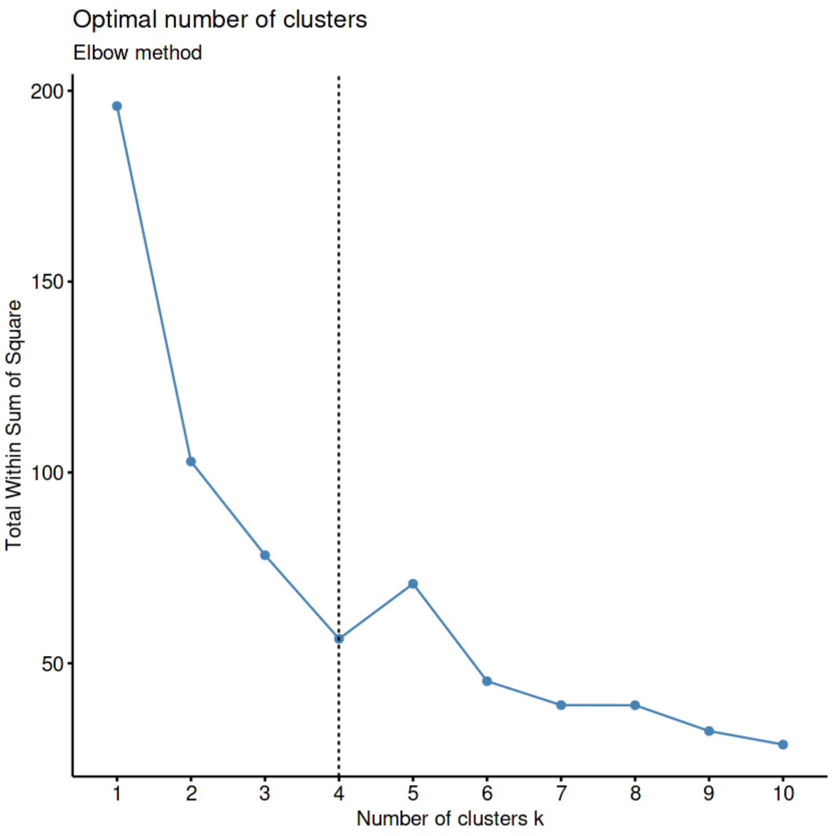
```
library(factoextra)
```

Determine the optimal number of clusters for k-means clustering:

In [14]:

```
# Elbow method
fviz_nbclust(df, kmeans, method = "wss") +
geom_vline(xintercept = 4, linetype = 2)+
labs(subtitle = "Elbow method")
```

Out[14]:



In [15]:

```
# Compute k-means with k = 4
set.seed(123)
km.res <- kmeans(df, 4, nstart = 25)
```

In [16]:

```
# Print the results
print(km.res)
```

```
K-means clustering with 4 clusters of sizes 8, 13, 16, 13

Cluster means:
      Murder    Assault   UrbanPop        Rape
1  1.4118898  0.8743346 -0.8145211  0.01927104
2 -0.9615407 -1.1066010 -0.9301069 -0.96676331
3 -0.4894375 -0.3826001  0.5758298 -0.26165379
4  0.6950701  1.0394414  0.7226370  1.27693964

Clustering vector:
       Alabama         Alaska        Arizona       Arkansas     California 
             1              4              4              1              4 
      Colorado    Connecticut       Delaware        Florida        Georgia 
             4              3              3              4              1 
        Hawaii          Idaho       Illinois        Indiana           Iowa 
             3              2              4              3              2 
        Kansas       Kentucky      Louisiana          Maine       Maryland 
             3              2              1              2              4 
 Massachusetts       Michigan      Minnesota    Mississippi       Missouri 
             3              4              2              1              4 
       Montana       Nebraska         Nevada  New Hampshire     New Jersey 
             2              2              4              2              3 
    New Mexico       New York North Carolina   North Dakota           Ohio 
             4              4              1              2              3 
      Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina 
             3              3              3              3              1 
  South Dakota      Tennessee          Texas           Utah        Vermont 
             2              1              4              3              2 
      Virginia     Washington  West Virginia      Wisconsin        Wyoming 
             3              3              2              2              3 

Within cluster sum of squares by cluster:
[1]  8.316061 11.952463 16.212213 19.922437
 (between_SS / total_SS =  71.2 %)

Available components:

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

In [17]:

```
aggregate(USArrests, by=list(cluster=km.res$cluster), mean)
```

Out[17]:

A data.frame: 4 × 5

cluster	Murder	Assault	UrbanPop	Rape
<int>	<dbl>	<dbl>	<dbl>	<dbl>
1	13.93750	243.62500	53.75000	21.41250
2	3.60000	78.53846	52.07692	12.17692
3	5.65625	138.87500	73.87500	18.78125
4	10.81538	257.38462	76.00000	33.19231

In [18]:

```
dd <- cbind(USArrests, cluster = km.res$cluster)
head(dd)
```

Out[18]:

A data.frame: 6 × 5

	Murder	Assault	UrbanPop	Rape	cluster
	<dbl>	<int>	<int>	<dbl>	<int>
Alabama	13.2	236	58	21.2	1
Alaska	10.0	263	48	44.5	4
Arizona	8.1	294	80	31.0	4
Arkansas	8.8	190	50	19.5	1
California	9.0	276	91	40.6	4
Colorado	7.9	204	78	38.7	4

In [19]:

```
# Cluster number for each of the observations
km.res$cluster
```

Out[19]: **Alabama:** 1 **Alaska:** 4 **Arizona:** 4 **Arkansas:** 1 **California:** 4 **Colorado:** 4 **Connecticut:** 3 **Delaware:** 3 **Florida:** 4 **Georgia:** 1 **Hawaii:** 3 **Idaho:** 2 **Illinois:** 4 **Indiana:** 3 **Iowa:** 2 **Kansas:** 3 **Kentucky:** 2 **Louisiana:** 1 **Maine:** 2 **Maryland:** 4 **Massachusetts:** 3 **Michigan:** 4 **Minnesota:** 2 **Mississippi:** 1 **Missouri:** 4 **Montana:** 2 **Nebraska:** 2 **Nevada:** 4 **New Hampshire:** 2 **New Jersey:** 3 **New Mexico:** 4 **New York:** 4 **North Carolina:** 1 **North Dakota:** 2 **Ohio:** 3 **Oklahoma:** 3 **Oregon:** 3 **Pennsylvania:** 3 **Rhode Island:** 3 **South Carolina:** 1 **South Dakota:** 2 **Tennessee:** 1 **Texas:** 4 **Utah:** 3 **Vermont:** 2 **Virginia:** 3 **Washington:** 3 **West Virginia:** 2 **Wisconsin:** 2 **Wyoming:** 3

In [20]:

```
head(km.res$cluster, 4)
```

Out[20]: **Alabama:** 1 **Alaska:** 4 **Arizona:** 4 **Arkansas:** 1

In [21]:

```
# Cluster size
km.res$size
```

Out[21]: 8 · 13 · 16 · 13

In [22]:

```
# Cluster means
km.res$centers
```

Out[22]:

A matrix: 4 × 4 of type dbl

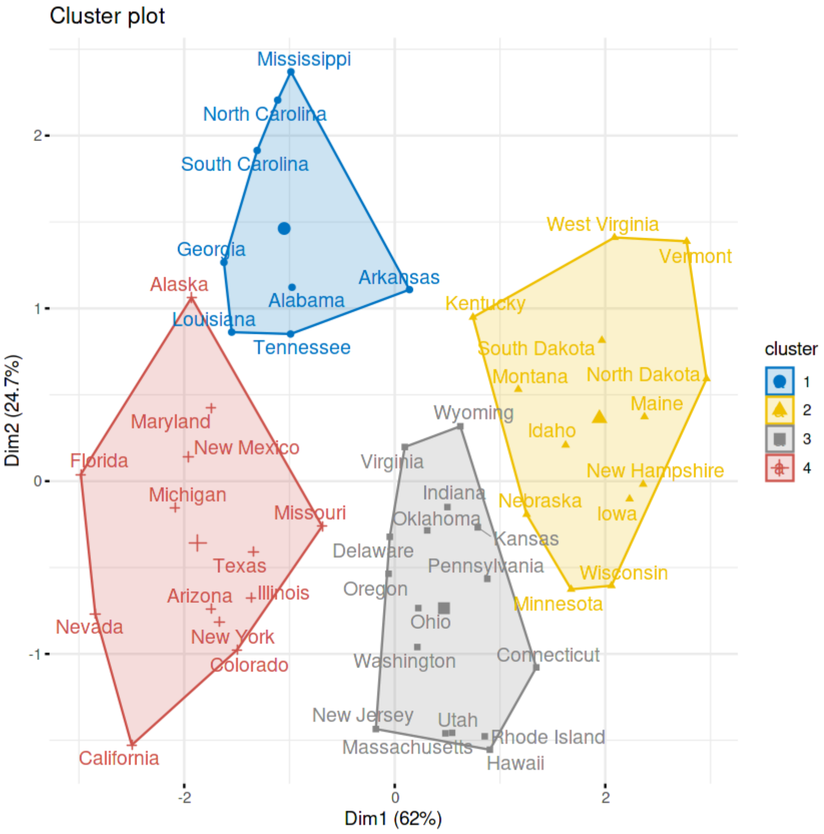
	Murder	Assault	UrbanPop	Rape
1	1.4118898	0.8743346	-0.8145211	0.01927104
2	-0.9615407	-1.1066010	-0.9301069	-0.96676331
3	-0.4894375	-0.3826001	0.5758298	-0.26165379
4	0.6950701	1.0394414	0.7226370	1.27693964

In [24]:

```
set.seed(123)
km.res <- kmeans(df, 4, nstart = 25)

# Visualize
library("factoextra")
fviz_cluster(km.res, data = df,
             ellipse.type = "convex",
             palette = "jco",
             repel = TRUE,
             ggtheme = theme_minimal())
```

Out[24]:



In [0]: