

Hierarchical Clustering on USArrests Dataset

EXP 4 : Hierarchical Clustering

AIM

To perform hierarchical clustering on the USArrests dataset, identify optimal clusters using gap statistic, and analyze the characteristics of each cluster.

EXPERIMENTAL SETUP

The dataset used is **USArrests**, which contains statistics, in arrests per 100,000 residents, for assault, murder, and rape in each of the 50 US states in 1973.

LIBRARIES REQUIRED

- factoextra
- cluster

STEP 1: Load Libraries and Dataset

```
# Load required libraries
library(factoextra)

## Loading required package: ggplot2

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

library(cluster)

# Load data
df <- USArrests

# Print number of attributes and number of instances
cat("Number of attributes:", ncol(df), "\n")

## Number of attributes: 4

cat("Number of instances:", nrow(df), "\n")

## Number of instances: 50
```

```

# Remove rows with missing values
df <- na.omit(df)

# Scale each variable to have mean 0 and SD 1
df <- scale(df)

# View first six rows
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

# Define linkage methods
m <- c("average", "single", "complete", "ward")
names(m) <- c("average", "single", "complete", "ward")

# Function to compute agglomerative coefficient
ac <- function(x) {
  agnes(df, method = x)$ac
}

# Calculate agglomerative coefficient for each method
sapply(m, ac)

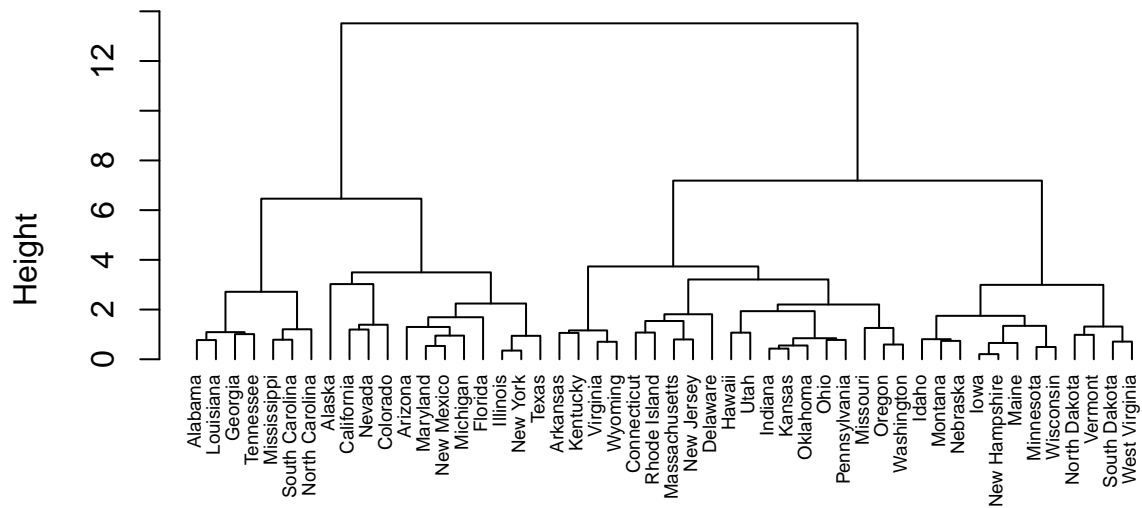
## average single complete ward
## 0.7379371 0.6276128 0.8531583 0.9346210

# Perform clustering using Ward's method
clust <- agnes(df, method = "ward")

# Produce dendrogram
pltree(clust, cex = 0.6, hang = -1, main = "Dendrogram")

```

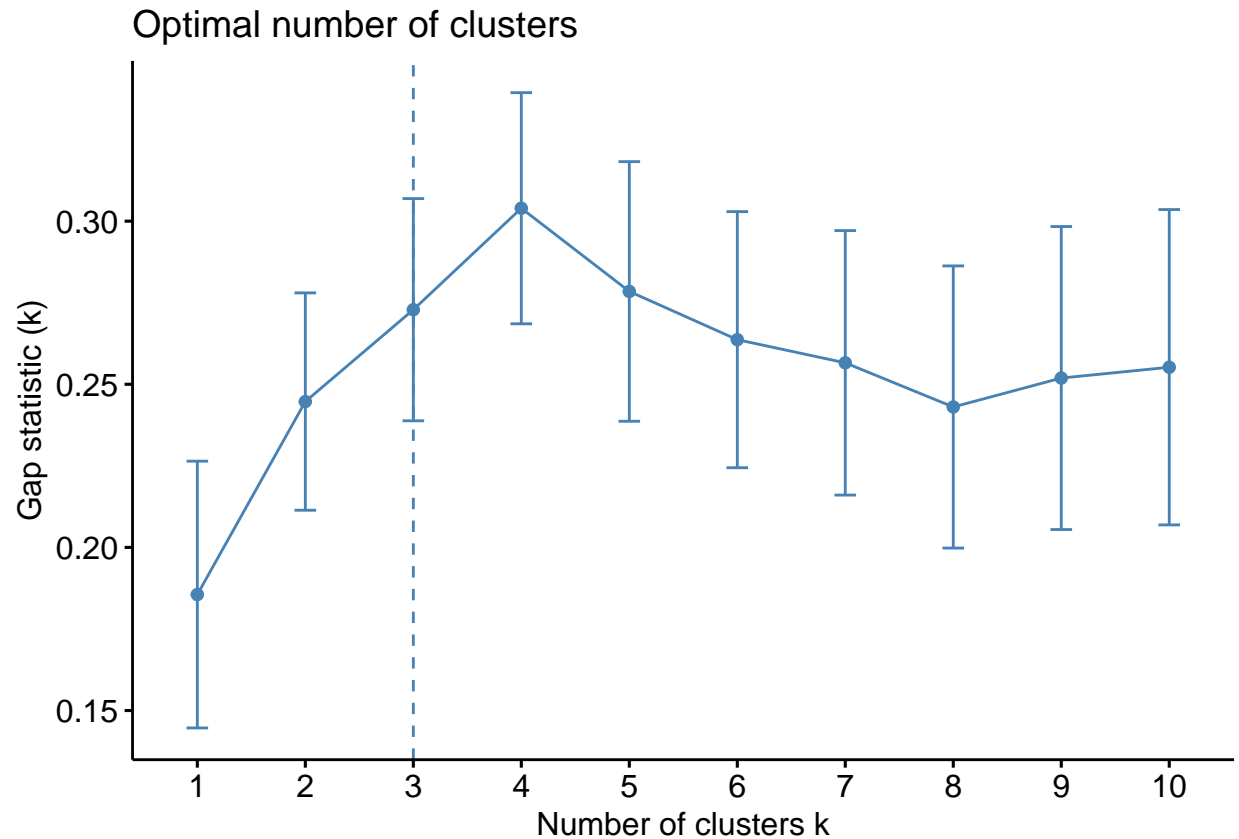
Dendrogram



df
agnes (*, "ward")

```
# Calculate gap statistic
gap_stat <- clusGap(df, FUN = hcut, nstart = 25, K.max = 10, B = 50)

# Plot gap statistic
fviz_gap_stat(gap_stat)
```



```
# Compute distance matrix
d <- dist(df, method = "euclidean")

# Perform hierarchical clustering using Ward.D2 method
final_clust <- hclust(d, method = "ward.D2")

# Cut dendrogram into 4 clusters
groups <- cutree(final_clust, k = 4)

# Number of observations per cluster
table(groups)
```

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

```
# Append cluster labels to original data
final_data <- cbind(USArrests, cluster = groups)

# Display first six rows of clustered data
head(final_data)
```

```
##           Murder Assault UrbanPop Rape cluster
## Alabama    13.2    236      58 21.2        1
## Alaska     10.0    263      48 44.5        2
```

```
## Arizona      8.1      294      80 31.0      2
## Arkansas     8.8      190      50 19.5      3
## California   9.0      276      91 40.6      2
## Colorado     7.9      204      78 38.7      2
```

```
# Mean values per cluster
```

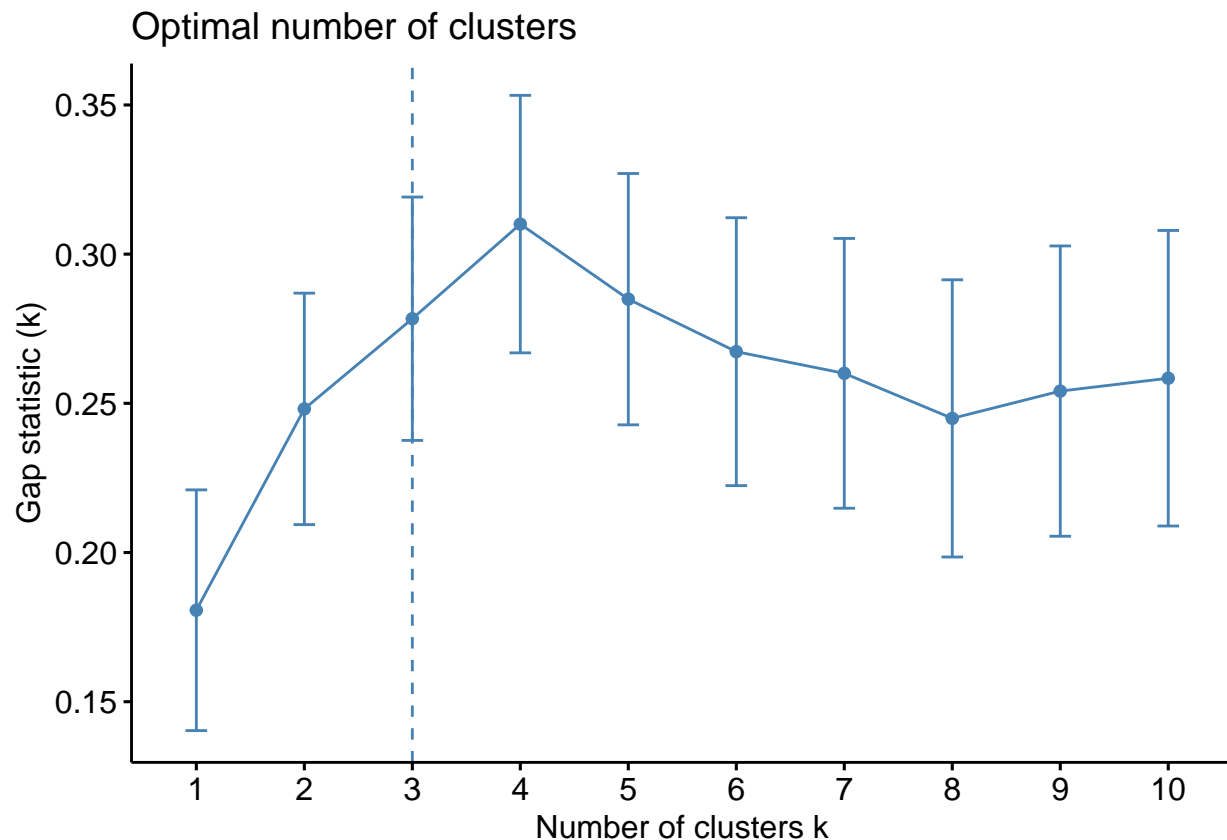
```
aggregate(final_data, by = list(cluster = final_data$cluster), mean)
```

```
##   cluster  Murder  Assault UrbanPop  Rape cluster
## 1      1 14.671429 251.2857 54.28571 21.68571      1
## 2      2 10.966667 264.0000 76.50000 33.60833      2
## 3      3  6.210526 142.0526 71.26316 19.18421      3
## 4      4  3.091667  76.0000 52.08333 11.83333      4
```

```
# Calculate gap statistic again (if needed)
```

```
gap_stat <- clusGap(df, FUN = hcut, nstart = 25, K.max = 10, B = 50)
```

```
fviz_gap_stat(gap_stat)
```



```
# Find optimal number of clusters
```

```
optimal_clusters <- which.max(gap_stat$Tab[, "gap"])
```

```
# Re-perform clustering with optimal number of clusters
```

```
final_clust <- hclust(d, method = "ward.D2")
```

```
groups <- cutree(final_clust, k = optimal_clusters)
```

```

# Append cluster labels
final_data <- cbind(USArrests, cluster = groups)

# Mean values per new cluster
aggregate(final_data, by = list(cluster = final_data$cluster), mean)

```

```

##  cluster      Murder  Assault UrbanPop      Rape cluster
## 1         1 14.671429 251.2857 54.28571 21.68571         1
## 2         2 10.966667 264.0000 76.50000 33.60833         2
## 3         3  6.210526 142.0526 71.26316 19.18421         3
## 4         4  3.091667  76.0000 52.08333 11.83333         4

```

CONCLUSION

The hierarchical clustering analysis grouped states into distinct clusters based on crime statistics. Ward's method was found effective, and the optimal number of clusters was determined using the gap statistic. This method provides insights into regional crime patterns and can support targeted policymaking.