

MANOJ KUMAR - 2048015

R-Laboratory 10

12/04/2021

1. Load the necessary packages for clustering.

```
library(tidyverse) # data manipulation
```

```
## — Attaching packages —————  
————— tidyverse 1.3.0 —
```

```
## ✓ ggplot2 3.3.3      ✓ purrr    0.3.4  
## ✓ tibble  3.0.3      ✓ dplyr    1.0.4  
## ✓ tidyr   1.1.1      ✓ stringr  1.4.0  
## ✓ readr   1.4.0      ✓ forcats  0.5.1
```

```
## — Conflicts —————  
————— tidyverse_conflicts() —  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag()     masks stats::lag()
```

```
library(cluster) # clustering algorithms  
library(factoextra) # clustering visualization
```

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

```
library(dendextend) # for comparing two dendrograms
```

```
##  
## -----  
## Welcome to dendextend version 1.14.0  
## Type citation('dendextend') for how to cite the package.  
##  
## Type browseVignettes(package = 'dendextend') for the package vignette.  
## The github page is: https://github.com/talgalili/dendextend/  
##  
## Suggestions and bug-reports can be submitted at: https://github.com/talgalili/dendextend/issues  
## Or contact: <tal.galili@gmail.com>  
##  
## To suppress this message use: suppressPackageStartupMessages(library(dendextend))  
## -----
```

```
##  
## Attaching package: 'dendextend'
```

```
## The following object is masked from 'package:stats':  
##  
##      cutree
```

```
library(corrplot)  # corrplot
```

```
## corrplot 0.84 loaded
```

2.Remove the unnecessary data.

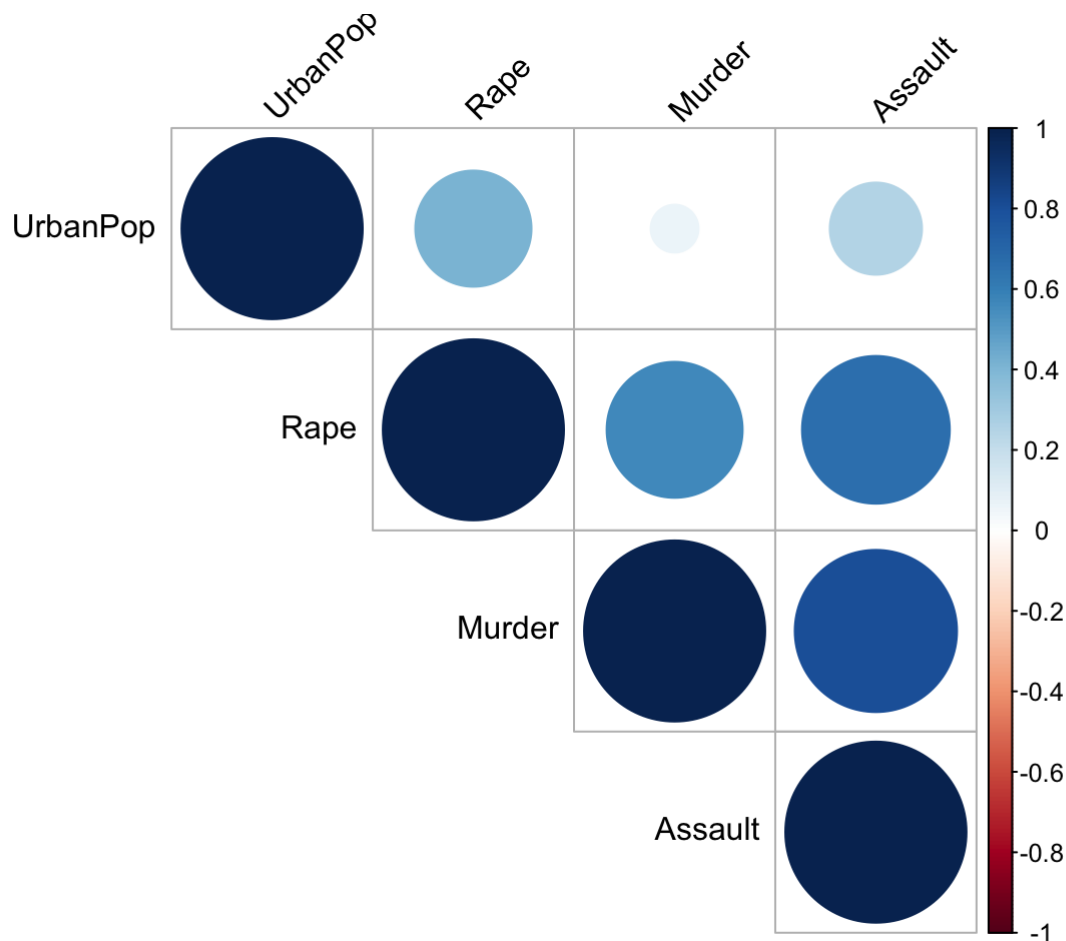
```
df <- USArrests  
df
```

	Murder <dbl>	Assault <int>	UrbanPop <int>	Rape <dbl>				
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				
Connecticut	3.3	110	77	11.1				
Delaware	5.9	238	72	15.8				
Florida	15.4	335	80	31.9				
Georgia	17.4	211	60	25.8				
1-10 of 50 rows		Previous	1	2	3	4	5	Next

```
df <- na.omit(df)
```

```
# Compute correlation matrix
```

```
res <- cor(df)  
corrplot(res, type = "upper", order = "hclust", tl.col = "black", tl.srt = 45)
```



3. Scale/Standardise the data.

```
df <- scale(df)
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
```

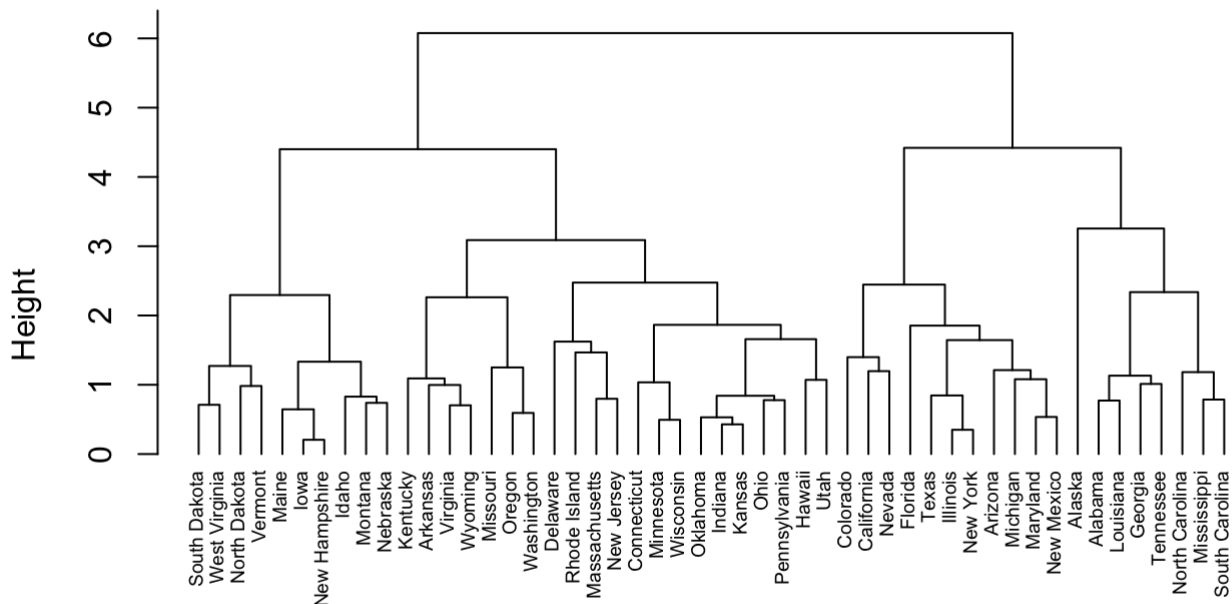
4. Perform Agglomerative Hierarchical Clustering by computing dissimilarity values and perform any hierarchical clustering method like complete linkage and then plot the dendrogram.

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

# Hierarchical clustering using Complete Linkage
hc1 <- hclust(d, method = "complete" )

# Plot the obtained dendrogram
plot(hc1, cex = 0.6, hang = -1)
```

Cluster Dendrogram



d
hclust (*, "complete")

```
# Compute with agnes
hc2 <- agnes(df, method = "complete")

# Agglomerative coefficient
hc2$ac
```

```
## [1] 0.8531583
```

```
# methods to assess
m <- c( "average", "single", "complete", "ward")
names(m) <- c( "average", "single", "complete", "ward")

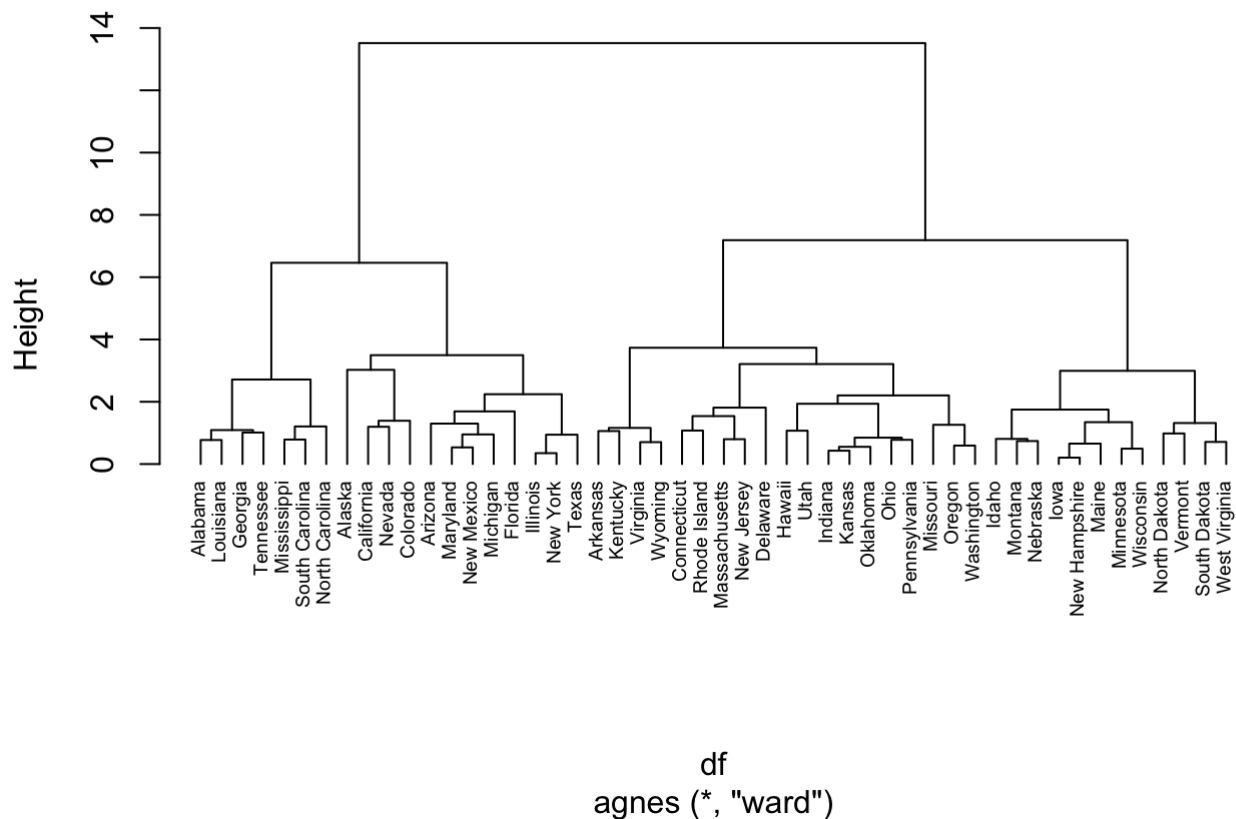
# function to compute coefficient
ac <- function(x) {
  agnes(df, method = x)$ac
}

map_dbl(m, ac)
```

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

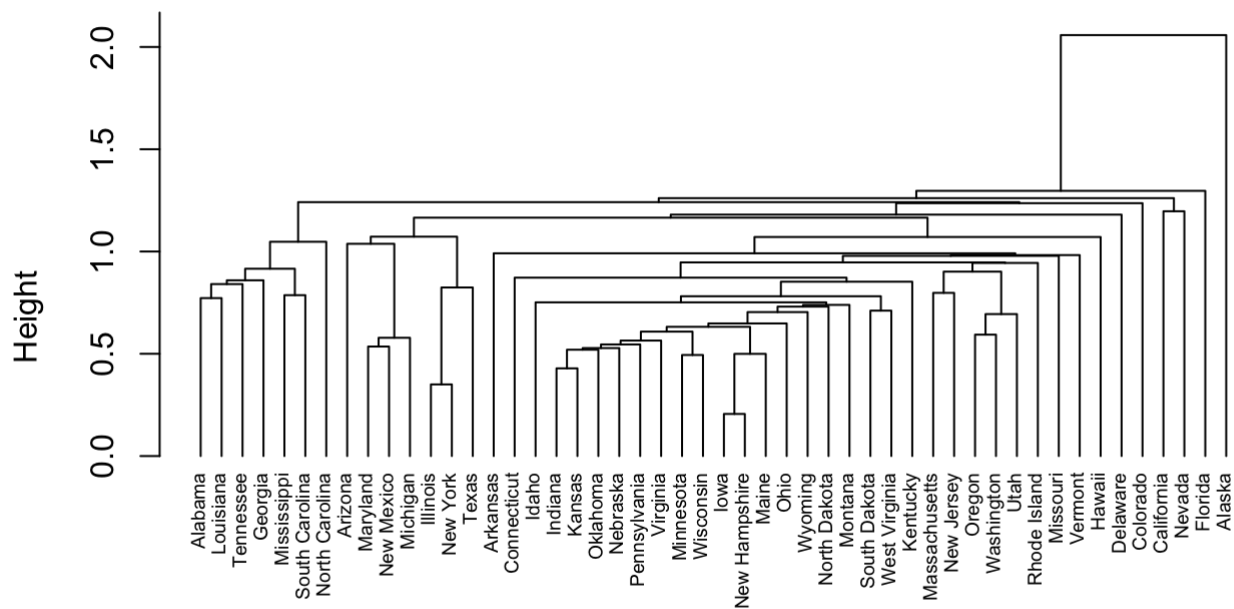
```
hc3 <- agnes(df, method = "ward")
pltree(hc3, cex = 0.6, hang = -1, main = "Dendrogram of agnes - ward")
```

Dendrogram of agnes - ward



```
hc3 <- agnes(df, method = "single")
pltree(hc3, cex = 0.6, hang = -1, main = "Dendrogram of agnes - single")
```

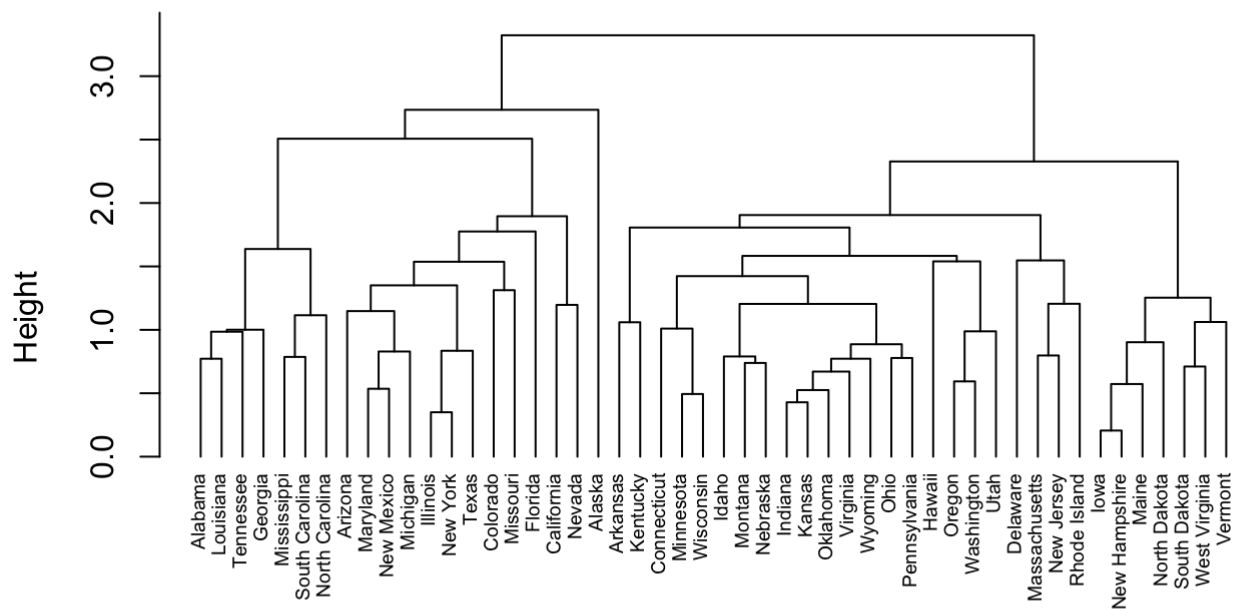
Dendrogram of agnes - single



df
agnes (*, "single")

```
hc3 <- agnes(df, method = "average")
pltree(hc3, cex = 0.6, hang = -1, main = "Dendrogram of agnes - average")
```

Dendrogram of agnes - average



df
agnes (*, "average")

Sub-Group

```
# Ward's method
hc5 <- hclust(d, method = "ward.D2" )

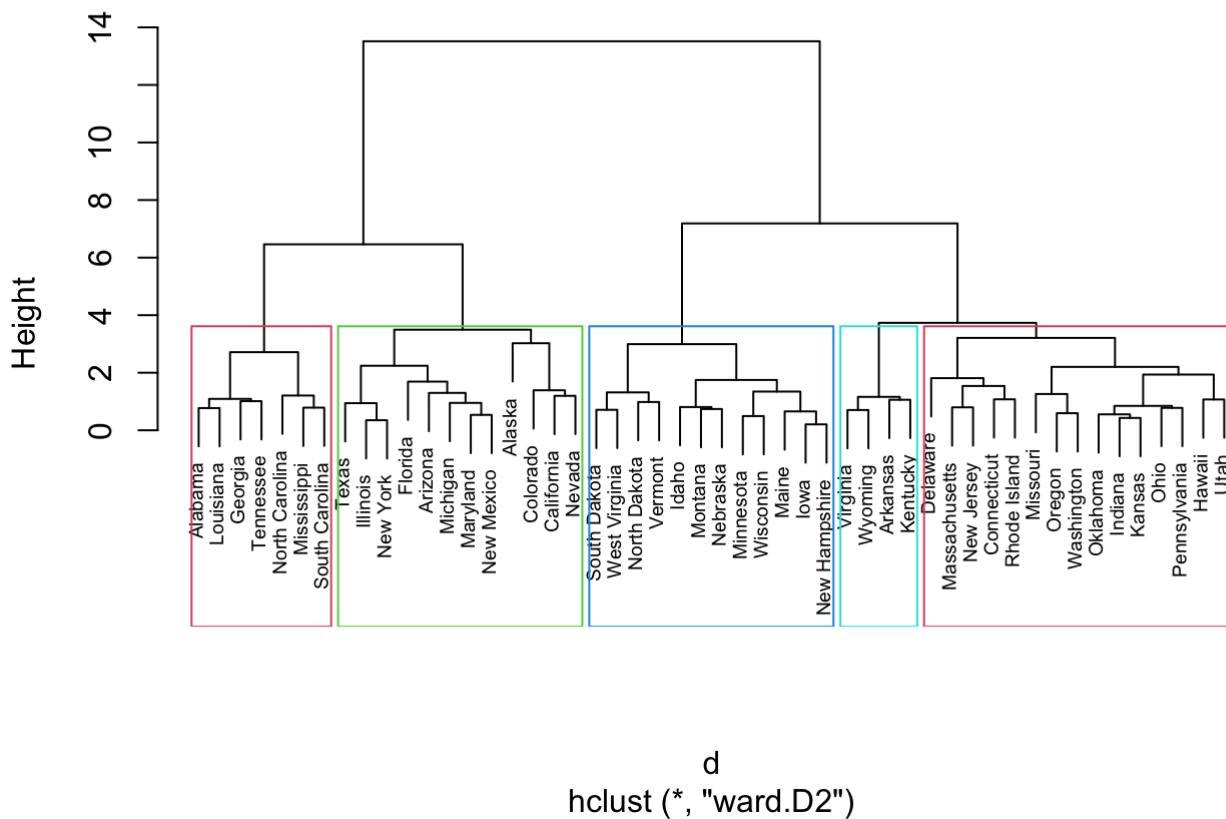
# Cut tree into 4 groups
sub_grp <- cutree(hc5, k = 5)

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

```
## sub_grp
##  1  2  3  4  5
##  7 12  4 15 12
```

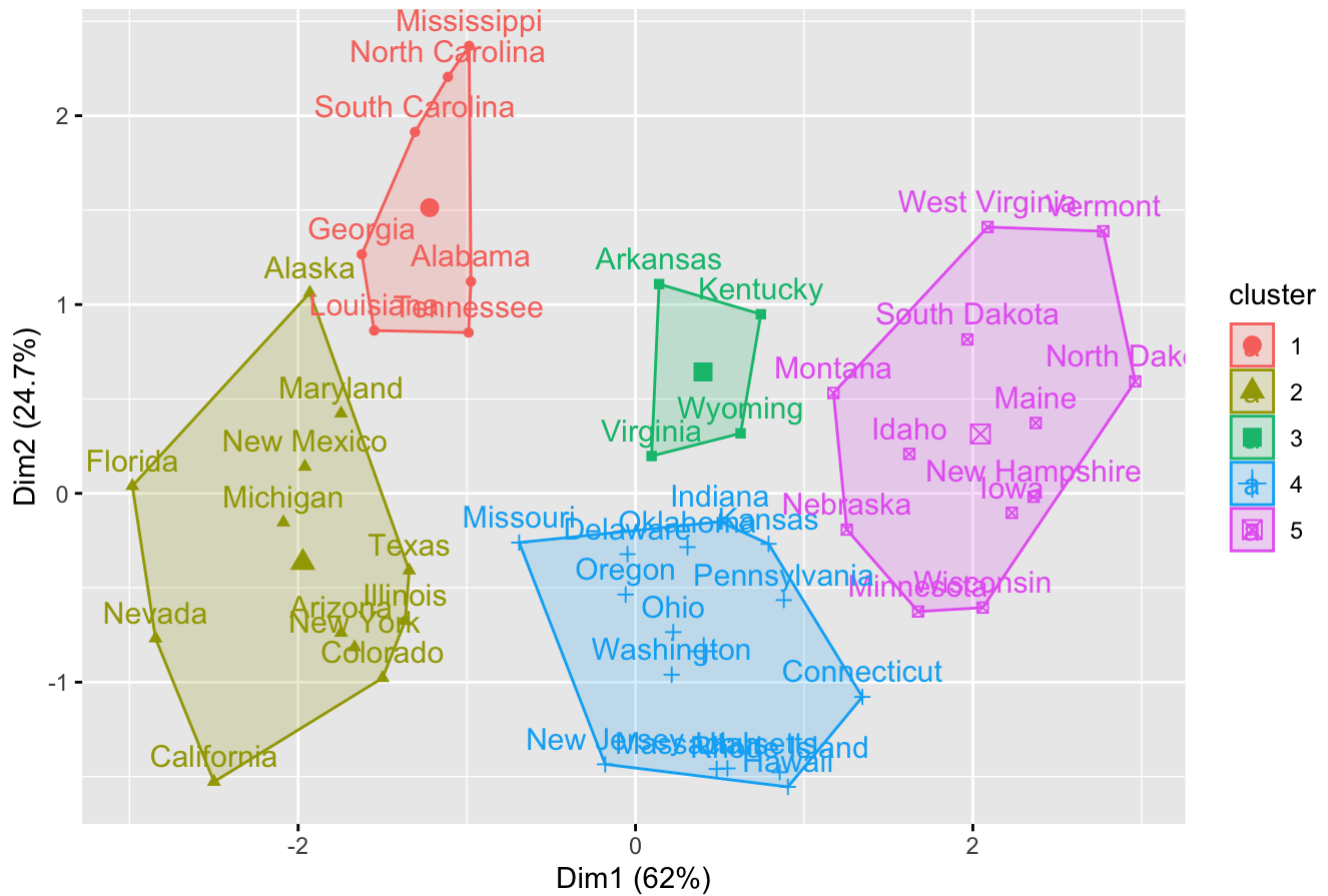
```
plot(hc5, cex = 0.6)
rect.hclust(hc5, k = 5, border = 2:5)
```

Cluster Dendrogram



```
fviz_cluster(list(data = df, cluster = sub_grp))
```

Cluster plot

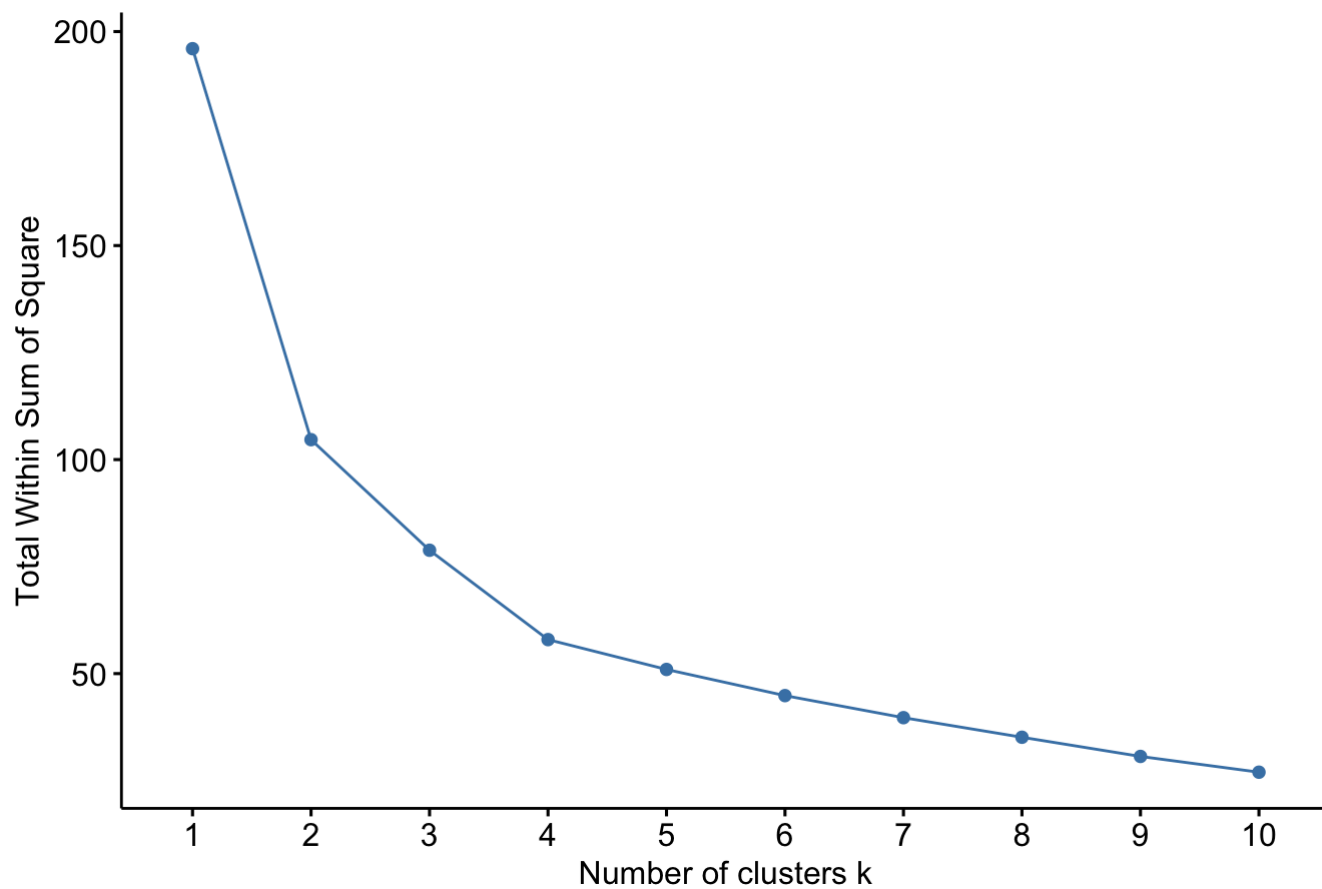


5. Determine optimal number of clusters

Elbow Method

```
fviz_nbclust(df, FUN = hcut, method = "wss")
```

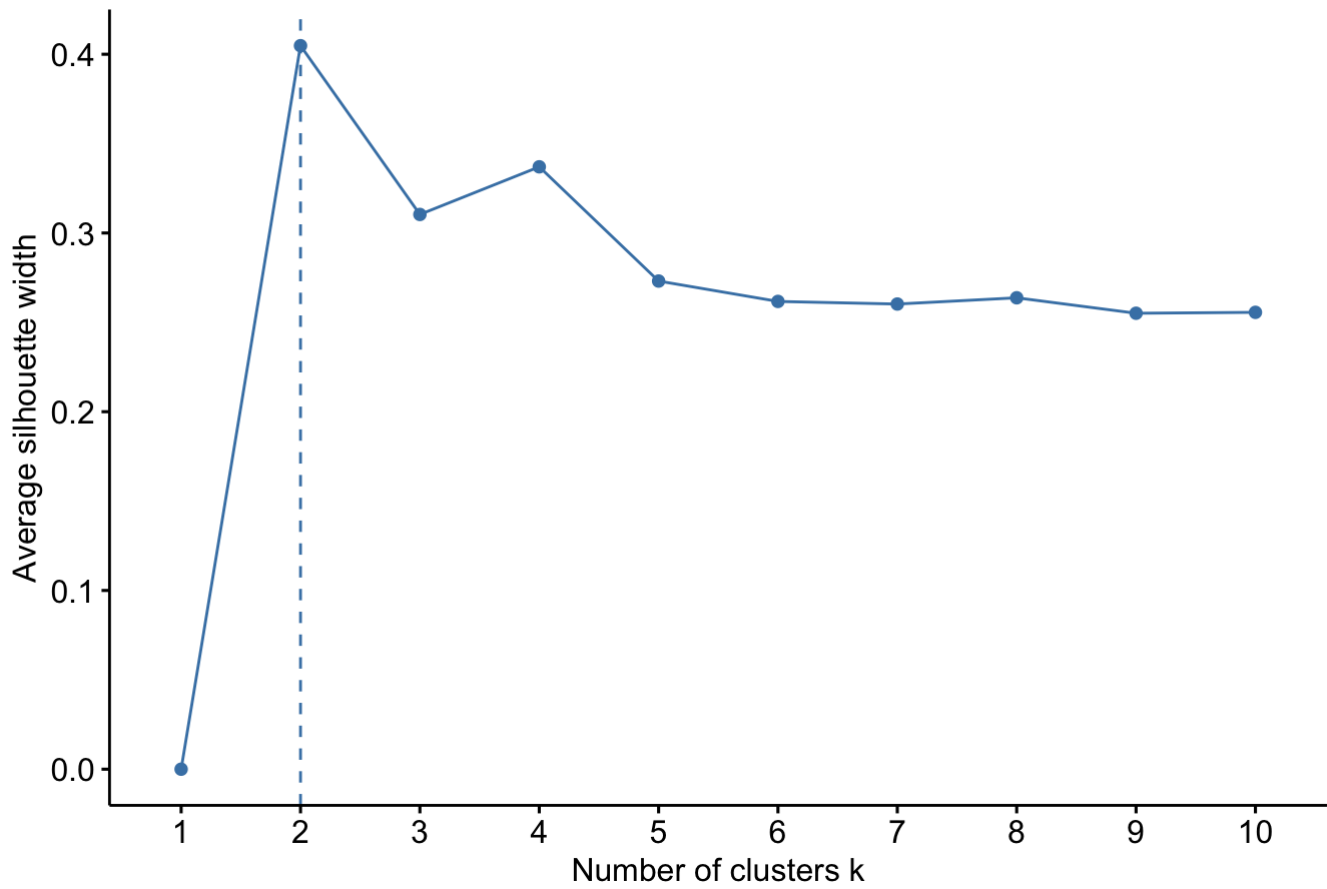

Optimal number of clusters



```
# Average Silhouette Method
```

```
fviz_nbclust(df, FUN = hcut, method = "silhouette")
```

Optimal number of clusters



```
# Gap Statistic Method
```

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

Optimal number of clusters

