Clustering Methods

Contents

1 K-means **2**

2 Hierarchical Clustering **6**

3 DBSCAN **12**

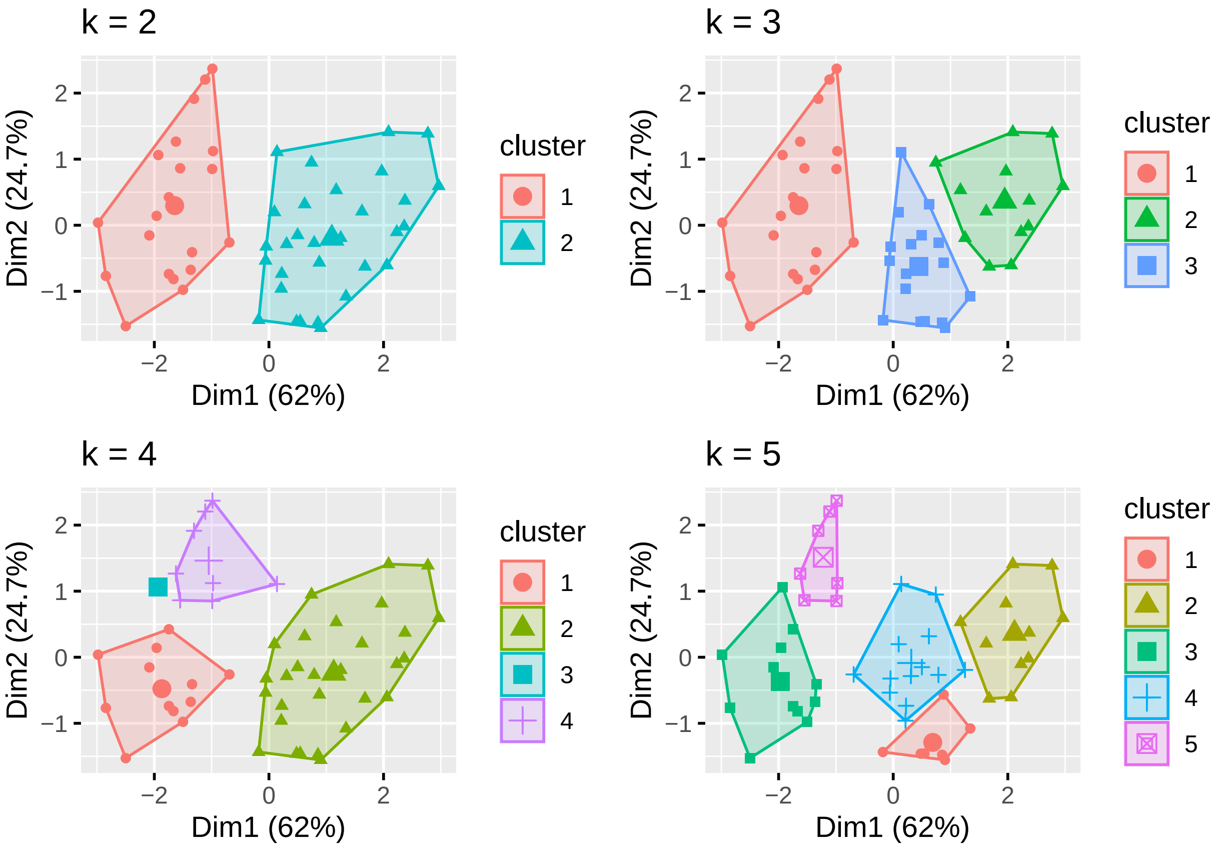
4 References **14**

```
library(dplyr)
library(cluster)
library(factoextra)
library(gridExtra)

# loading the data
data('USArrests')
df <- na.omit(USArrests)  # remove observations with missing values
df <- scale(df) # normalize the data before clustering
head(df)
```

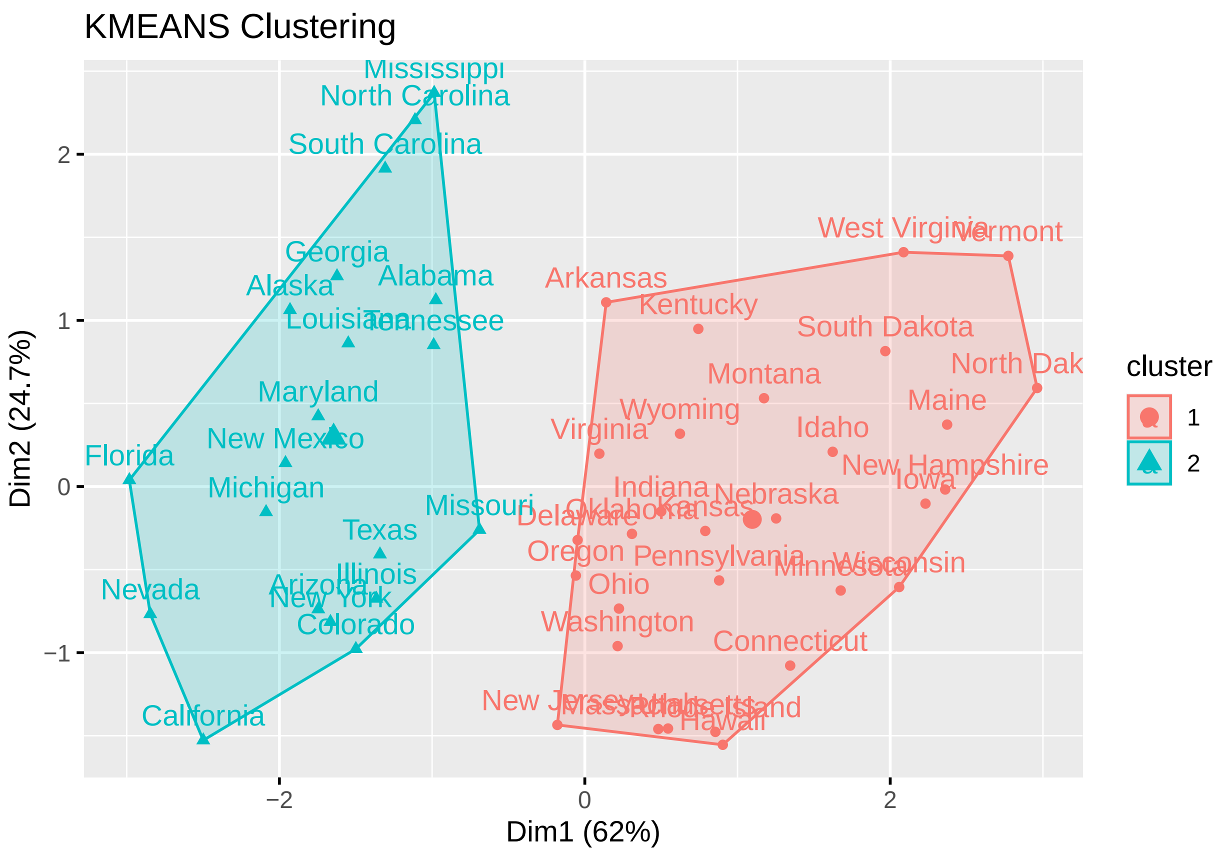
```
##                 Murder   Assault   UrbanPop         Rape
## Alabama    1.24256408 0.7828393 -0.5209066 -0.003416473
## Alaska     0.50786248 1.1068225 -1.2117642  2.484202941
## Arizona    0.07163341 1.4788032  0.9989801  1.042878388
## Arkansas   0.23234938 0.2308680 -1.0735927 -0.184916602
##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

1 K-means

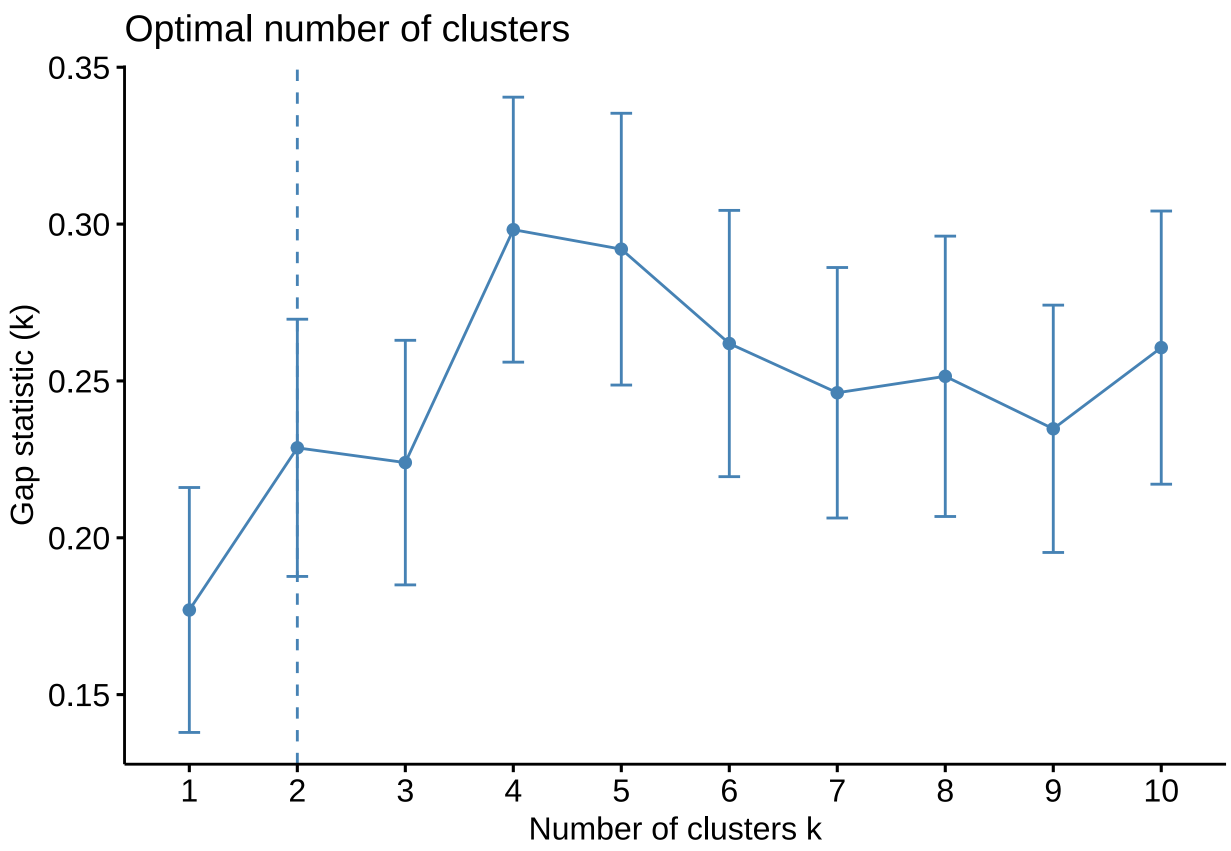
```
# different setting about the number of clusters K
set.seed(1)
kmeans2 <- kmeans(df, centers = 2)
kmeans3 <- kmeans(df, centers = 3)
kmeans4 <- kmeans(df, centers = 4)
kmeans5 <- kmeans(df, centers = 5)

# visualization
plot1 <- fviz_cluster(kmeans2, geom = "point", data = df) + ggtitle("k = 2")
plot2 <- fviz_cluster(kmeans3, geom = "point", data = df) + ggtitle("k = 3")
plot3 <- fviz_cluster(kmeans4, geom = "point", data = df) + ggtitle("k = 4")
plot4 <- fviz_cluster(kmeans5, geom = "point", data = df) + ggtitle("k = 5")
grid.arrange(plot1, plot2, plot3, plot4, nrow = 2)
```



```
# enhanced k-means clustering
res.km <- eclust(df, "kmeans")
```

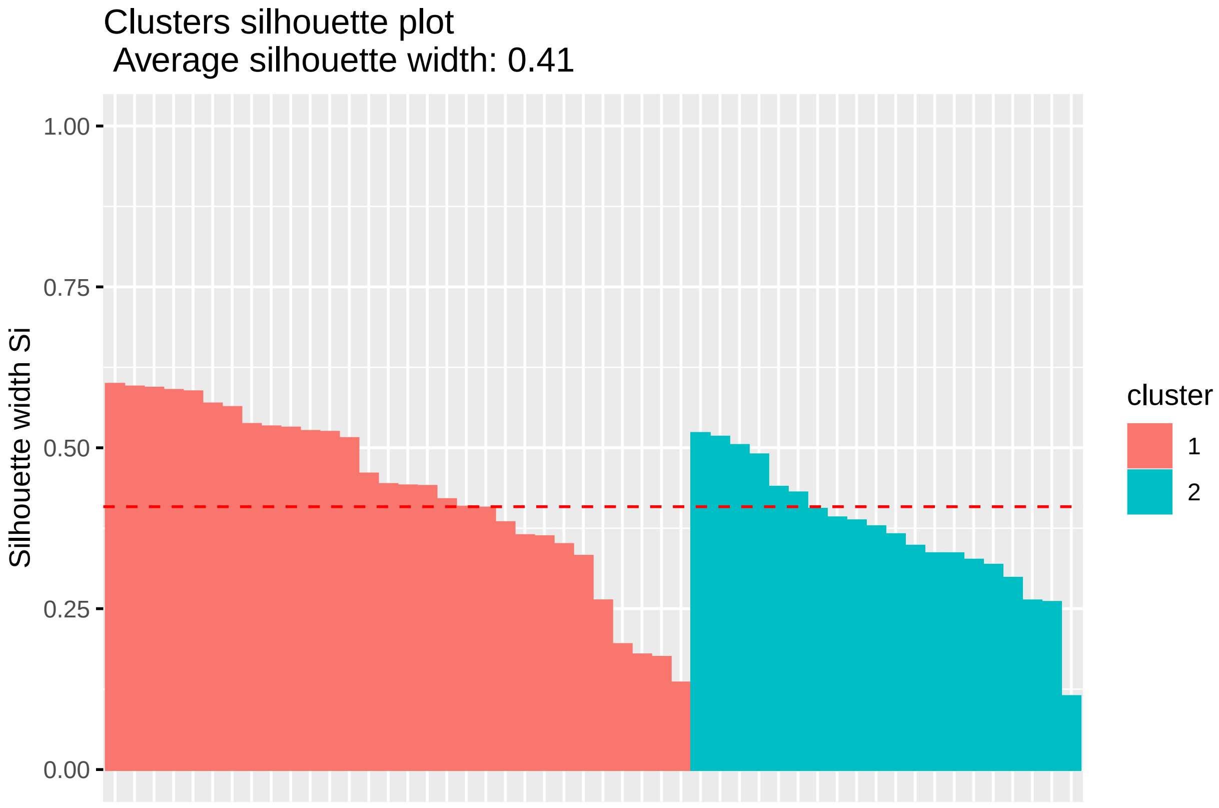


```
# Gap statistic plot
fviz_gap_stat(res.km$gap_stat)
```



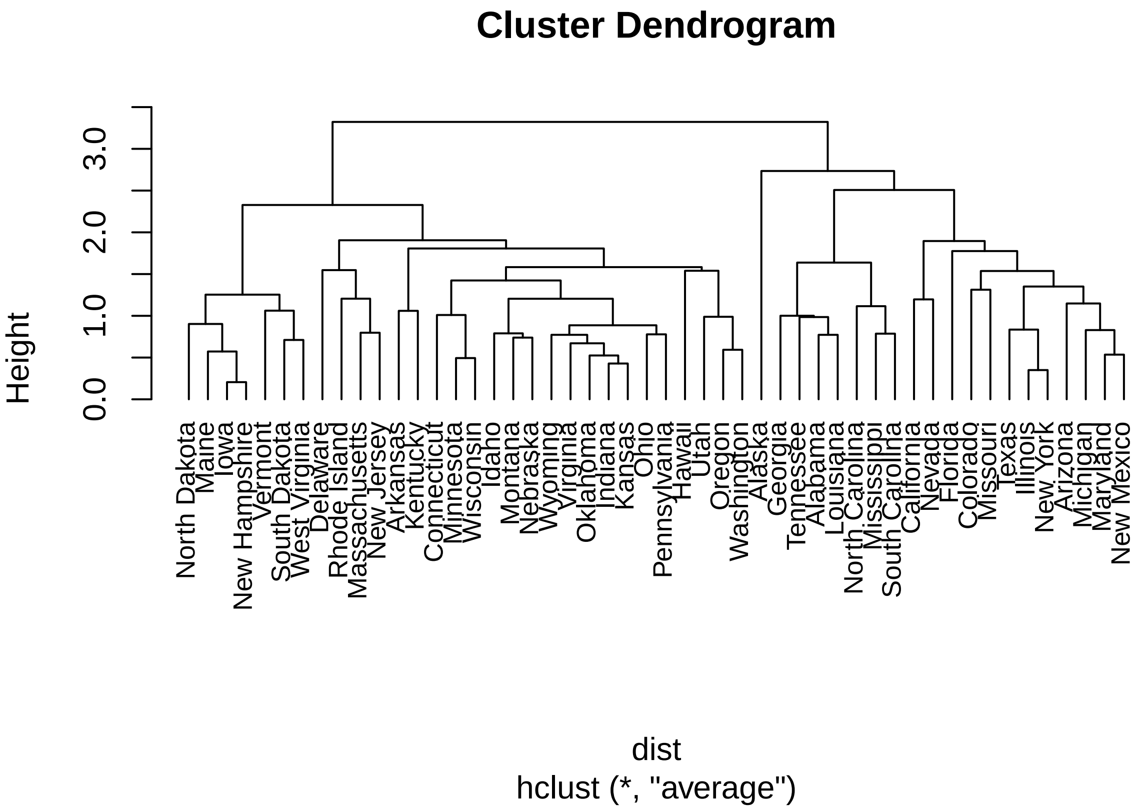
```
# Silhouette plot
fviz_silhouette(res.km)
```

```
##   cluster size ave.sil.width
## 1       1   30          0.43
## 2       2   20          0.37
```



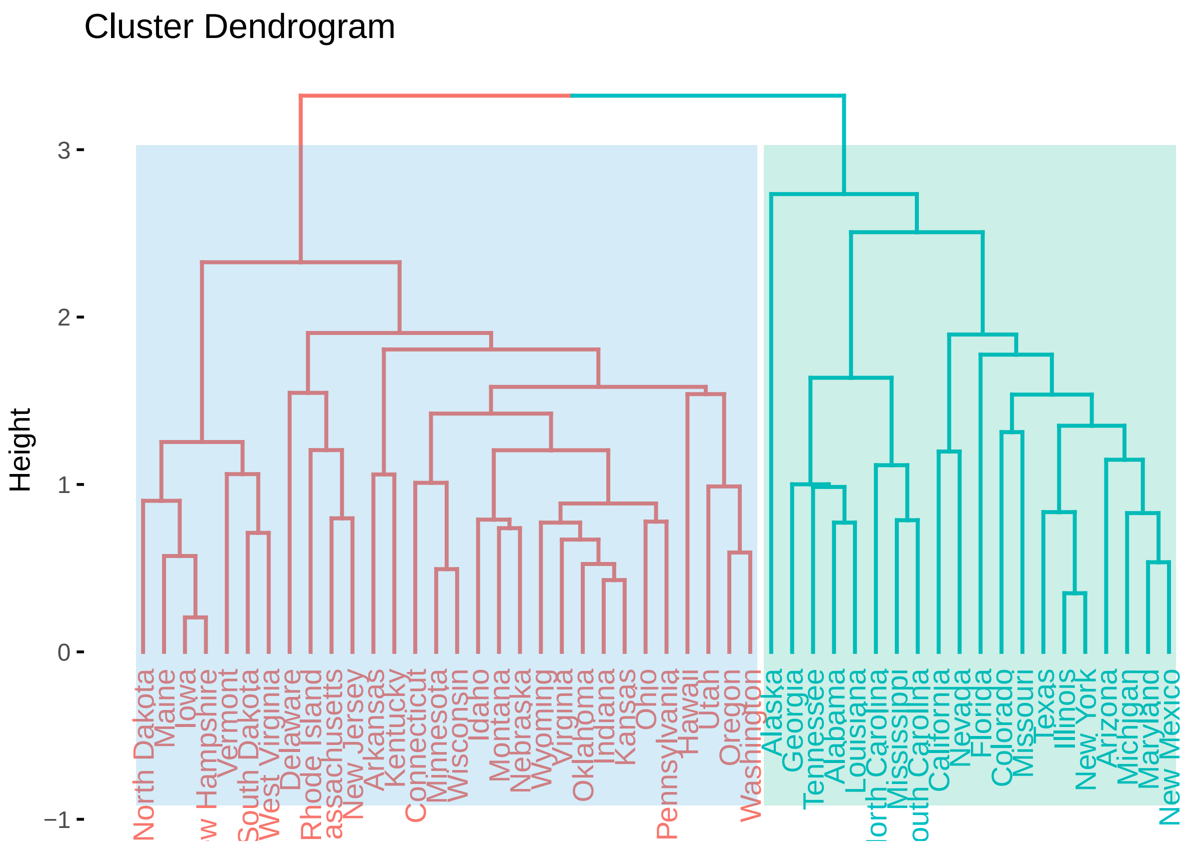
2 Hierarchical Clustering

```
# distance matrix
dist <- dist(df)
# fitting hierarchical clustering model
hc <- hclust(dist, method = "average")
plot(hc, hang = -1, cex = 0.8)
```

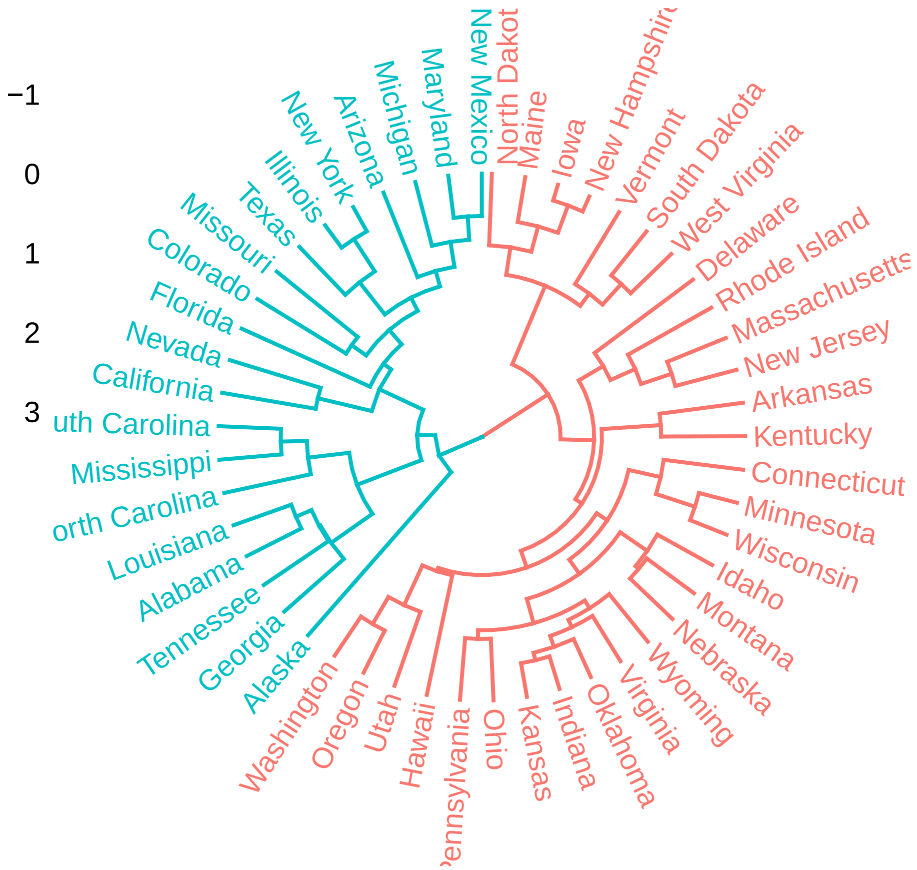


```
# different  visualization methods
fviz_dend(hc, k = 2, rect = TRUE, rect_fill =  TRUE, rect_border = c("#2E9FDF", "#00AF88"))
```

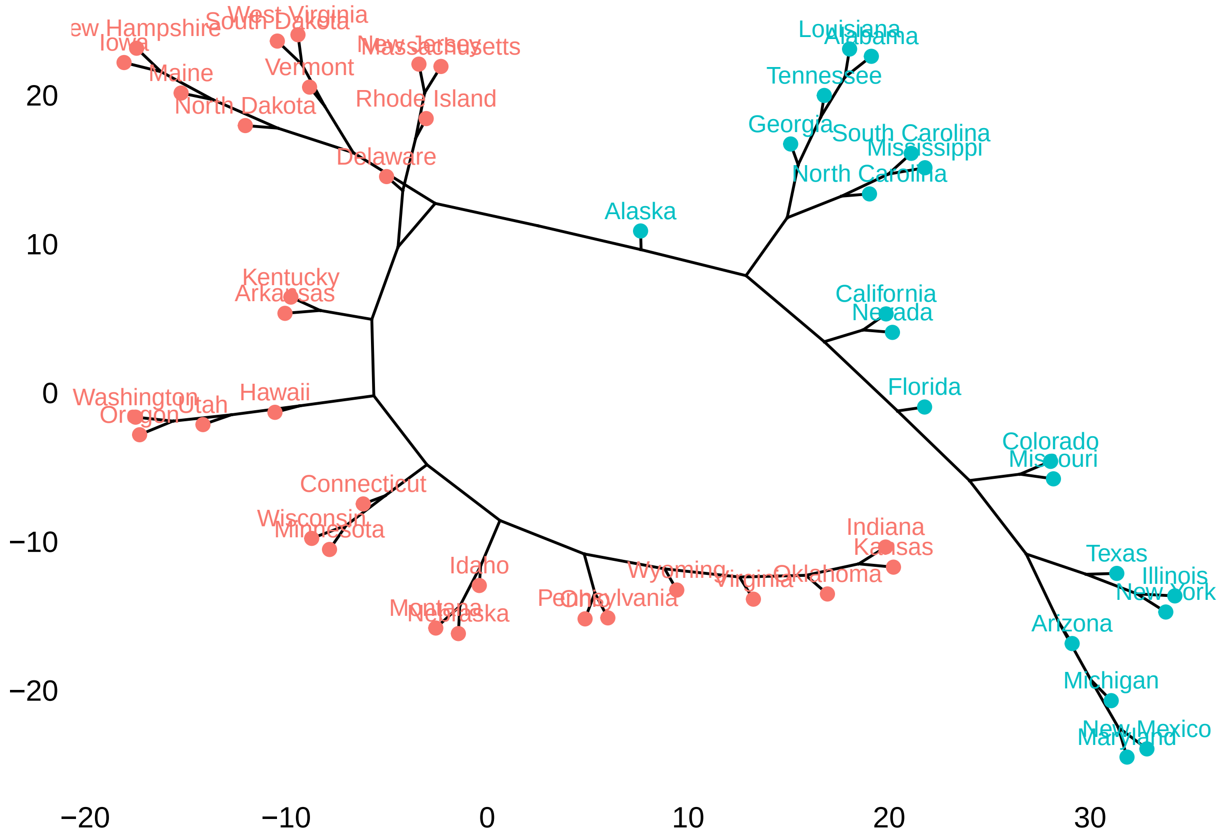
```
## Warning in if (color == "cluster") color <- "default":
##     €
```



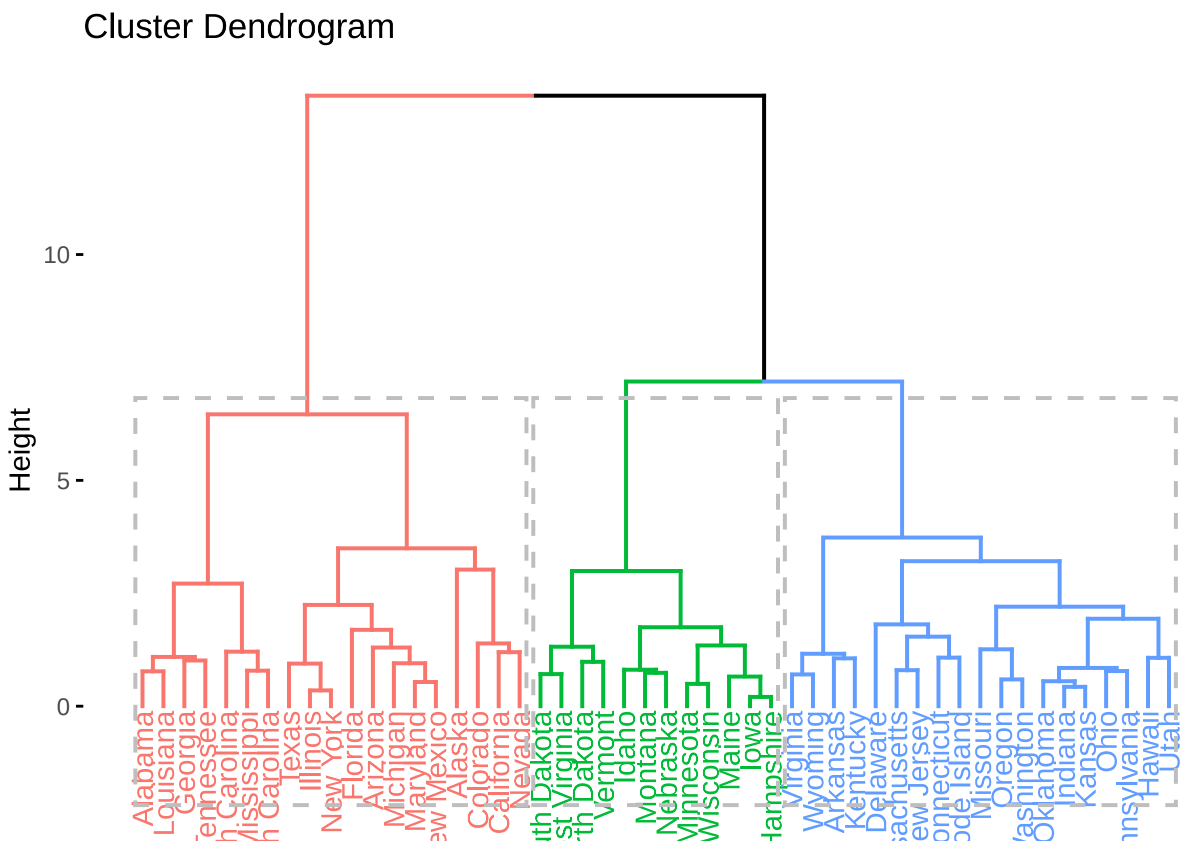
```
fviz_dend(hc, k = 2, rect = TRUE, rect_fill =  TRUE, type = 'circular', rect_border = c("#2E9FDF", "#00A
```



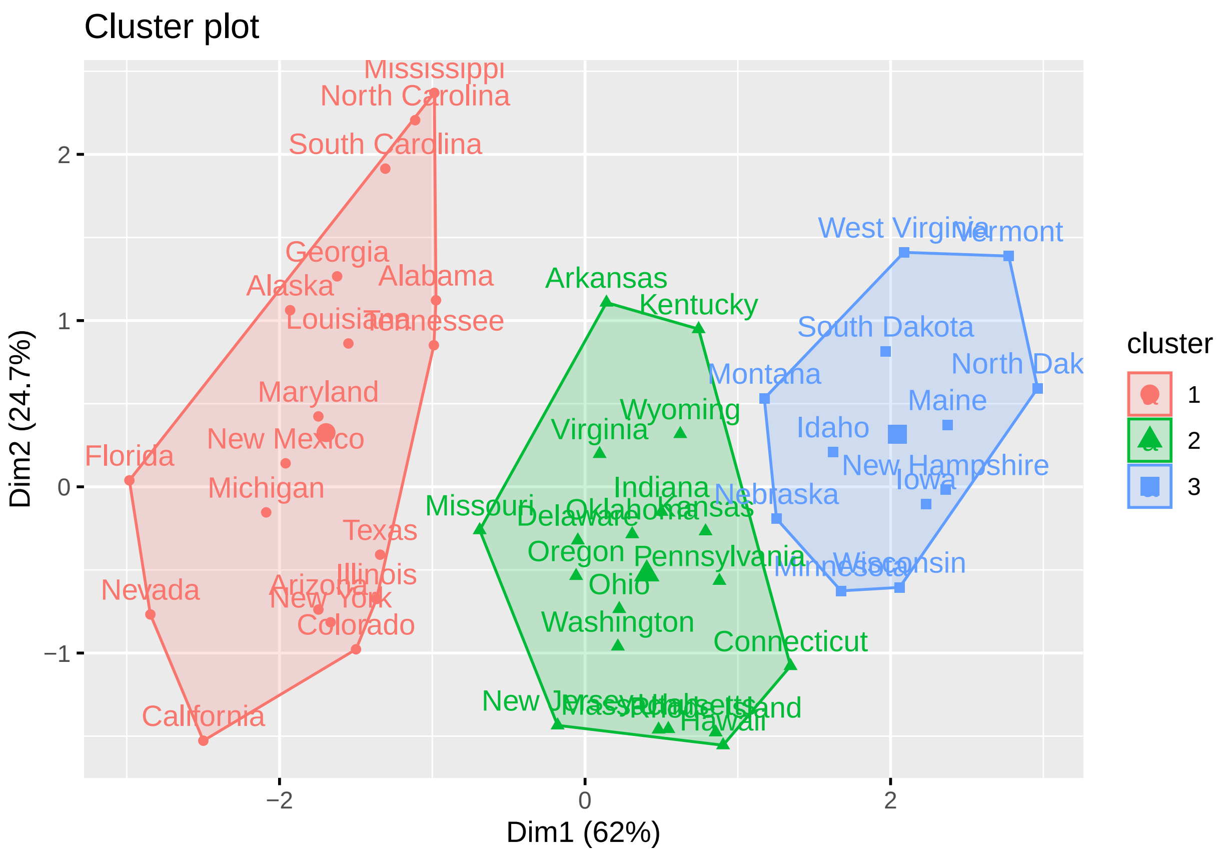
```
fviz_dend(hc, k = 2, rect = TRUE, rect_fill =  TRUE, type = 'phylogenic', rect_border = c("#2E9FDF", "#
```



```
# enhanced hierarchical clustering
res.hc <- eclust(df, "hclust")
fviz_dend(res.hc, rect = TRUE)
```

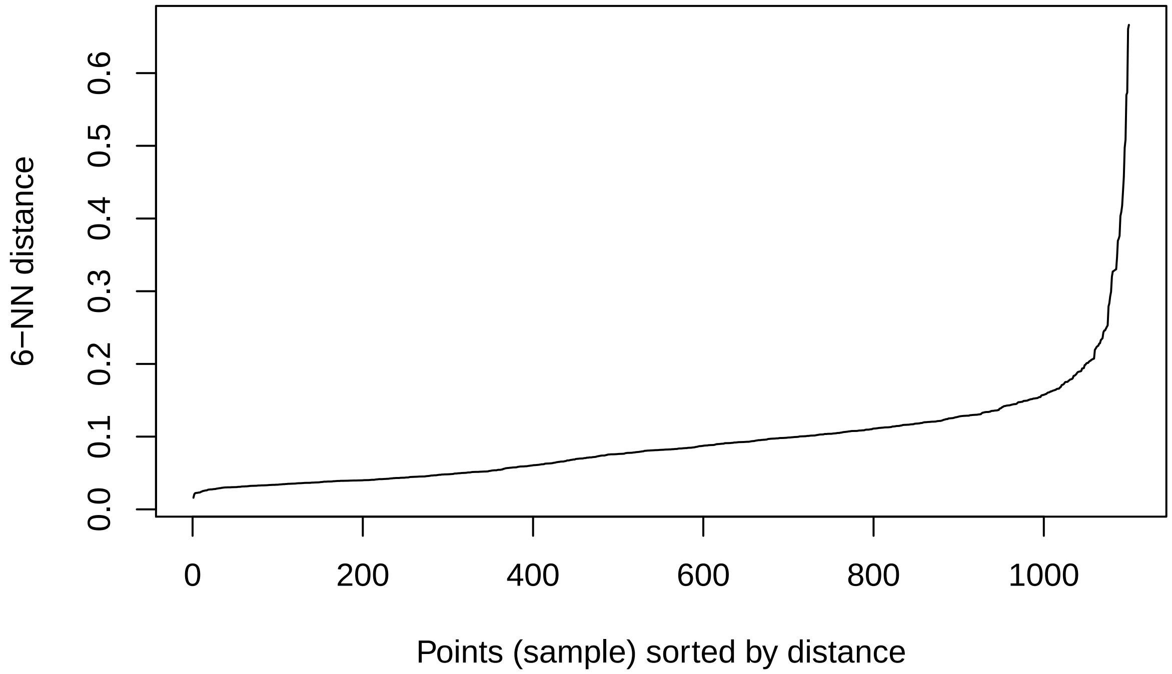


```
fviz_cluster(res.hc)
```



3 DBSCAN

```
data("multishapes")
df1 <- multishapes[,1:2]

library("dbscan")
kNNdistplot(df1, k = 6)
```

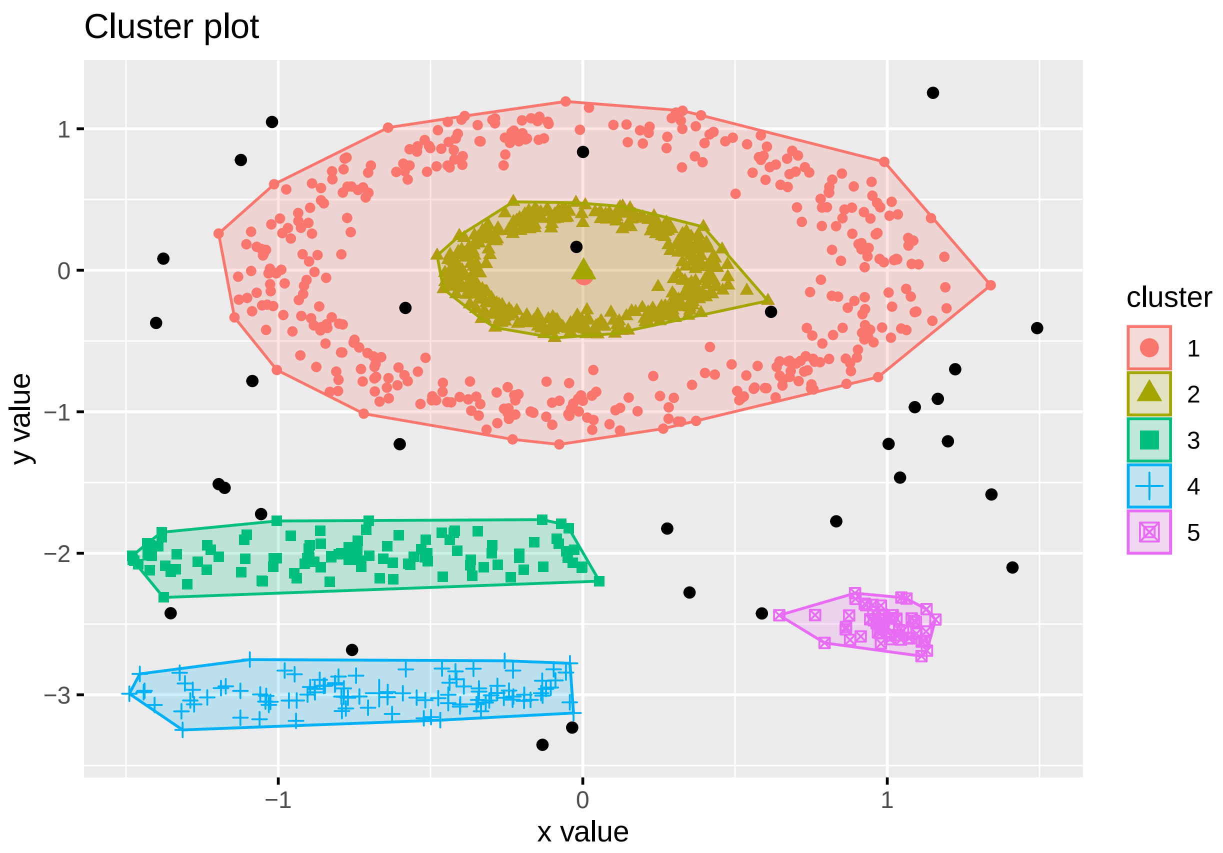


```
library(fpc)
```

```
##
## Attaching package: 'fpc'

## The following object is masked from 'package:dbscan':
##
##     dbscan
```

```
db <- dbscan(df1, eps = 0.15, MinPts = 5)
fviz_cluster(db, data = df1, stand = FALSE, frame = FALSE, geom = "point")
```

```
## Warning: argument frame is deprecated; please use ellipse instead.
```



4 References

https://data-flair.training/blogs/clustering-in-r-tutorial/

https://blog.csdn.net/dege857/article/details/116697417

https://zhuanlan.zhihu.com/p/30890984