Q45

John Bute

2024-11-29

R Markdown

2. Cluster the Wine datasets using k−means, for various distance metrics and algorithm parameters. What is your best estimation for the number of clusters in each case? Validate your results.

```
data <- read.csv("C:/Users/johnb/Desktop/Machine Learning/data/wine.csv")
data <- scale(data)

set.seed(1)

kmeans_clustering <- kmeans(data, centers = 3, nstart = 25)

print(kmeans_clustering)
```

```
## K-means clustering with 3 clusters of sizes 61, 49, 68
##
## Cluster means:
##         Class    Alcohol Malic.acid        Ash Alcalinity.of.ash
Magnesium
## 1 -1.16822514  0.8756272 -0.3037196  0.3180446        -0.6626544
0.56329925
## 2  1.34366784  0.1860184  0.9024258  0.2485092         0.5820616 -
0.05049296
## 3  0.07973544 -0.9195318 -0.3778231 -0.4643776         0.1750133 -
0.46892793
##   Total.phenols  Flavanoids Nonflavanoid.phenols Proanthocyanins
## 1    0.87403990  0.94098462         -0.583942581      0.58014642
## 2   -0.98577624 -1.23271740          0.714825281     -0.74749896
## 3   -0.07372644  0.04416309          0.008736157      0.01821349
##   Colour.intensity        Hue OD280.OD315    Proline
## 1        0.1667181  0.4823674   0.7648958  1.1550888
## 2        0.9857177 -1.1879477  -1.2978785 -0.3789756
## 3       -0.8598525  0.4233092   0.2490794 -0.7630972
##
## Clustering vector:
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1
##  [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3
3 3 1
##  [75] 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3
## [112] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2
```

```
2 2 2
## [149] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##
## Within cluster sum of squares by cluster:
## [1] 350.5475 304.6223 623.1702
##  (between_SS / total_SS =  48.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

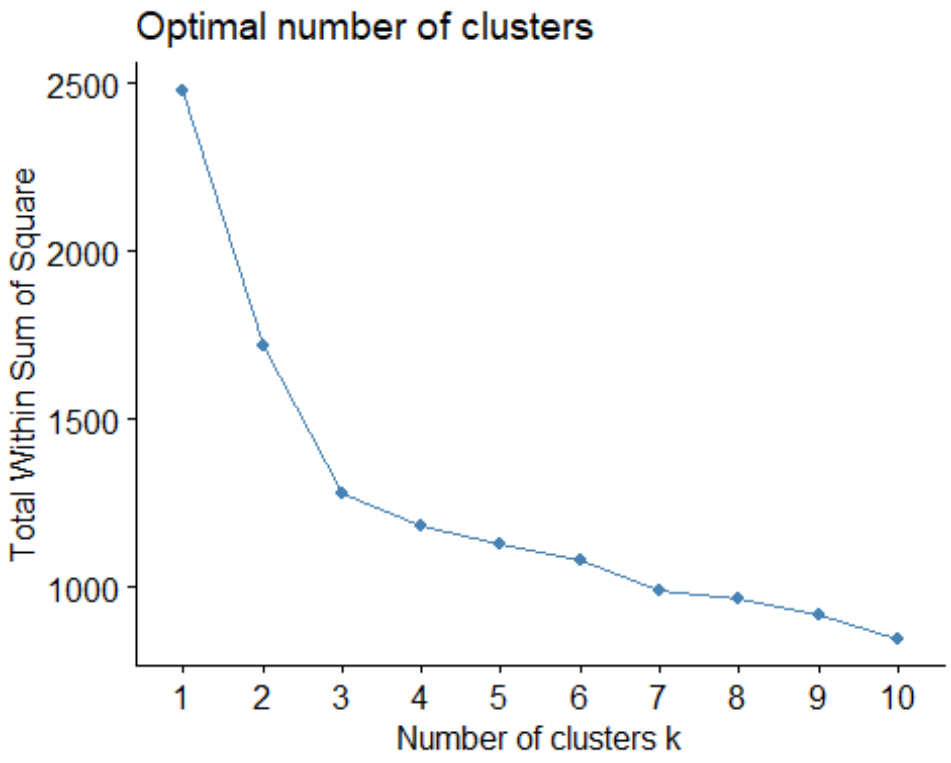
```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.4.2
```

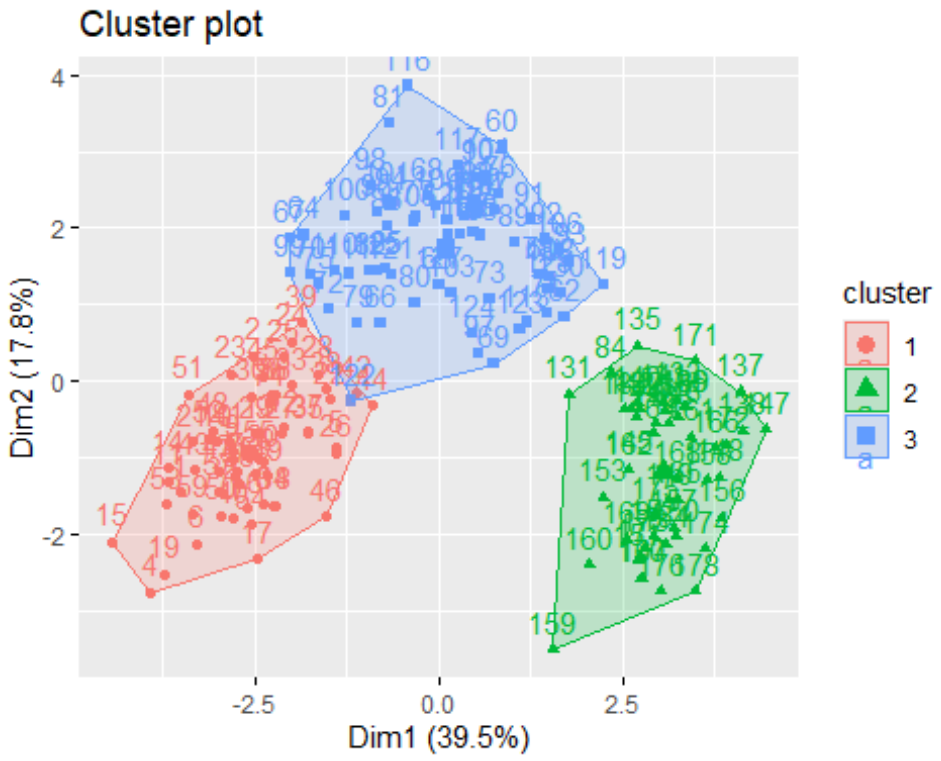
```
## Loading required package: ggplot2
```

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

```
fviz_nbclust(data, kmeans, method = "wss")
```




```
fviz_cluster(kmeans_clustering, data = data)
```



According to the elbow method, it appears to bend at 3 clusters, therefore 3 is the optimal number.

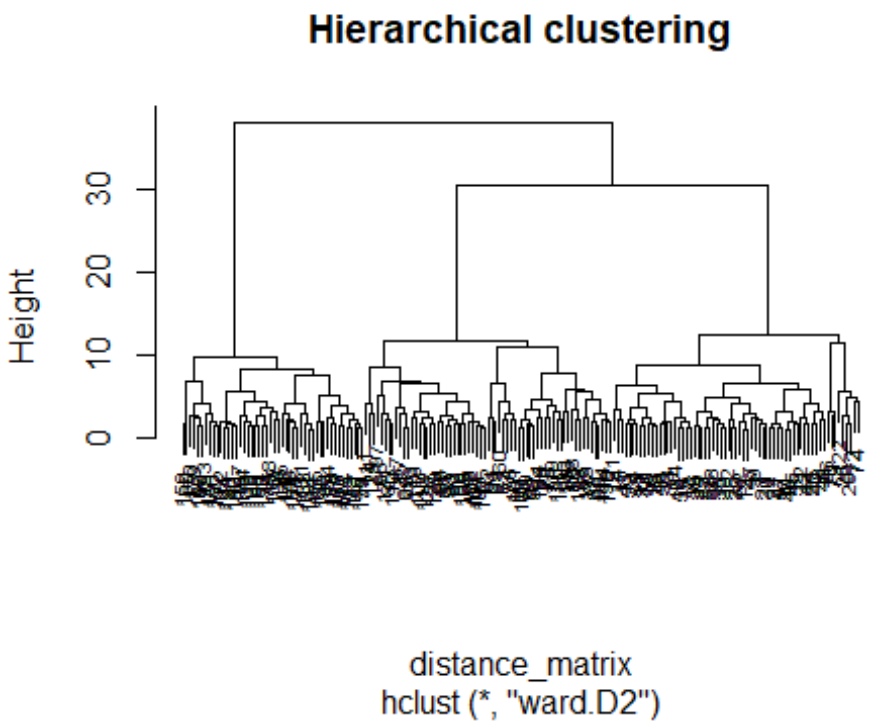
3. Cluster the Wine datasets using hierarchical clustering, for various algorithm parameters. Validate your results.

```r
set.seed(1)

distance_matrix <- dist(data, method = "euclidean")

hc_clustering <- hclust(distance_matrix, method = "ward.D2")

plot(hc_clustering, main = "Hierarchical clustering", cex = 0.6)
```




```
hc_clusters <- cutree(hc_clustering, k = 3)


cluster_sizes <- table(hc_clusters)

print(cluster_sizes)
```

```
## hc_clusters
##  1  2  3
## 65 65 48
```