Clustering Analysis

Blake Pappas

2023-12-17

Clustering Analysis in R

Load the following packages:

```r
library(stats)
library(dplyr)
library(cluster)
```

In this exercise, we use the “iris.csv” file.

P1: Import the dataset

```r
iris = read.csv("iris.csv")
```

P2: Normalize the data. Note that only the first four columns are relevant for clustering. Run the following lines to set up the normalization function:

```r
normalize = function(x) {
  return ((x - min(x)) / (max(x) - min(x)))
  }
```

Next, use this function to normalize the data. Think about which columns to normalize.

```r
iris_normalized = iris %>%
  mutate_at(c(1:4), normalize)
```

P3: Get the distance matrix

Use Euclidean distance.

```
distance_matrix = dist(iris_normalized[, 1:4], method = "euclidean")
```

P4: Apply Hierarchical Clustering

Use Ward's method to measure distance between clusters.

```
hierarchical = hclust(distance_matrix, method = "ward.D")
```

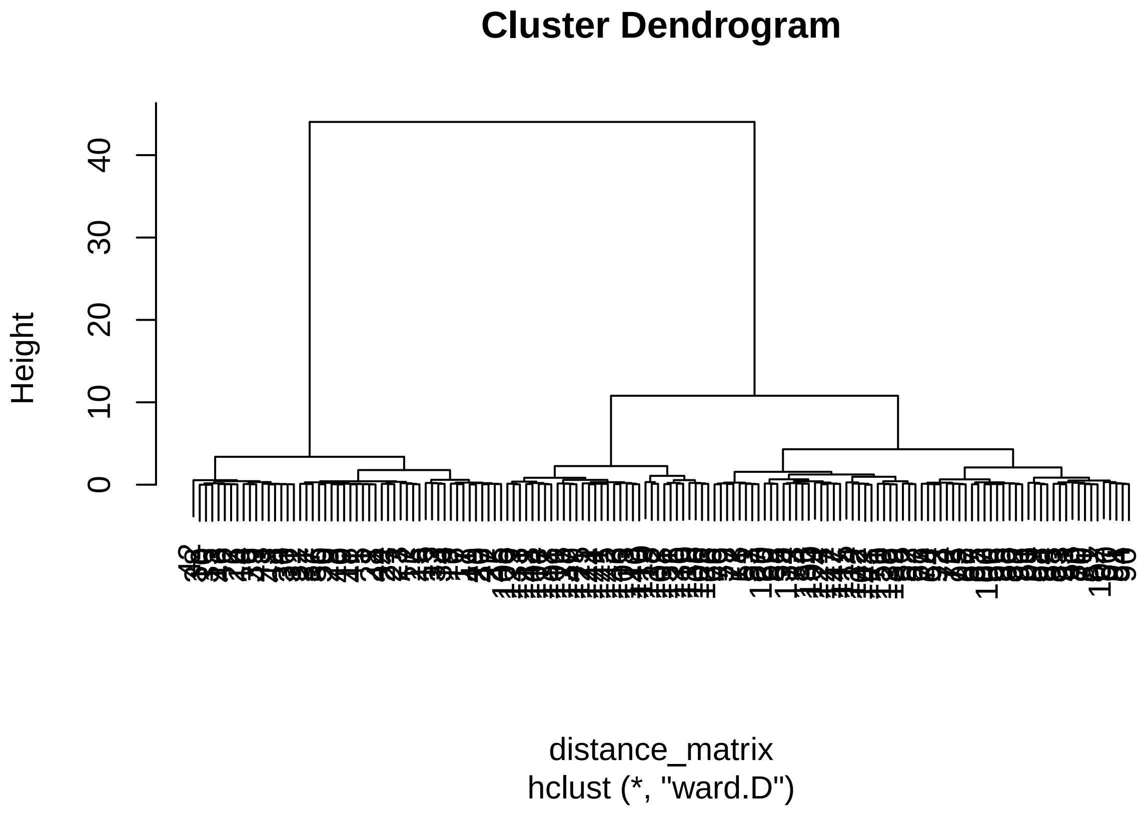
P5: Plot the dendrogram

No need to specify the label parameter.

How many clusters do you think is appropriate?

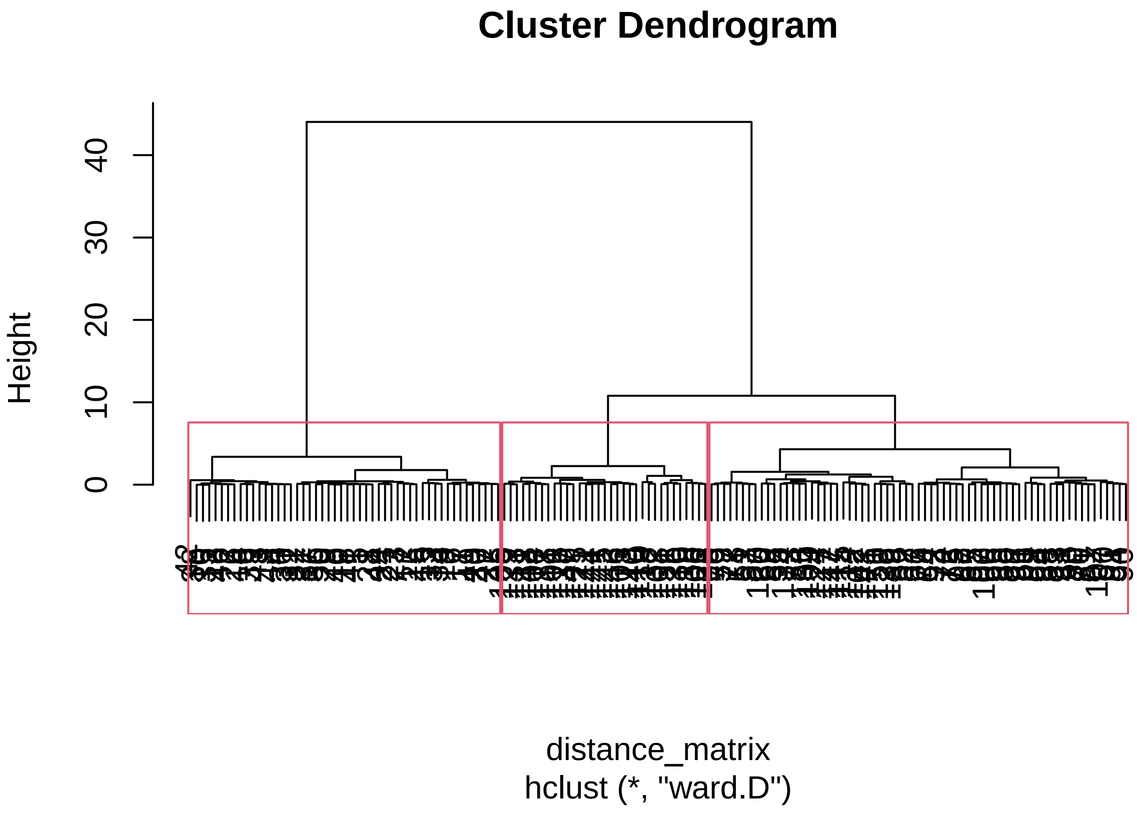
Answer: I think that three clusters are appropriate.

```
plot(hierarchical, labels = iris_normalized$Name)
```



P6: Mark the 3-cluster solution on the dendrogram

```
plot(hierarchical, labels = iris_normalized$Name)
rect.hclust(hierarchical, k = 3)
```



P7: Take the 3-cluster solution based on hierarchical clustering and add it to the original dataframe.

```
iris_normalized$cluster = cutree(hierarchical, k = 3)
```

P8: Apply K-Means Clustering

Choose 3 as the number of clusters.

```
kcluster = kmeans(iris_normalized[, 1:4], centers = 3)
```

P9: Report cluster centroids

```
kcluster$centers
```

```
##   sepal_length sepal_width petal_length petal_width
## 1    0.4412568   0.3073770   0.57571548   0.5491803
## 2    0.1961111   0.5908333   0.07864407   0.0600000
## 3    0.7072650   0.4508547   0.79704476   0.8247863
```

P10: Based on cluster centroids, interpret andcharacterize each cluster.

Answer: Cluster 1 represents the biggest iris flowers, as is evidenced by the larger sepal lengths (0.7072650) and widths (0.4508547) and petal lengths (0.79704476) and widths (0.8247863). Cluster 2 represents the mid-sized iris flowers, as is evidenced by the more moderately-sized sepal lengths (0.4412568) and widths (0.3073770) and petal lengths (0.57571548) and widths (0.5491803). Cluster 3 represents the smallest iris flowers, as is evidenced by the smaller sepal lengths (0.1961111) and widths (0.5908333) and petal lengths (0.07864407) and widths (0.0600000).

P11: Is 3 the most natural cluster number?

Answer: No, three is not the most natural cluster number.

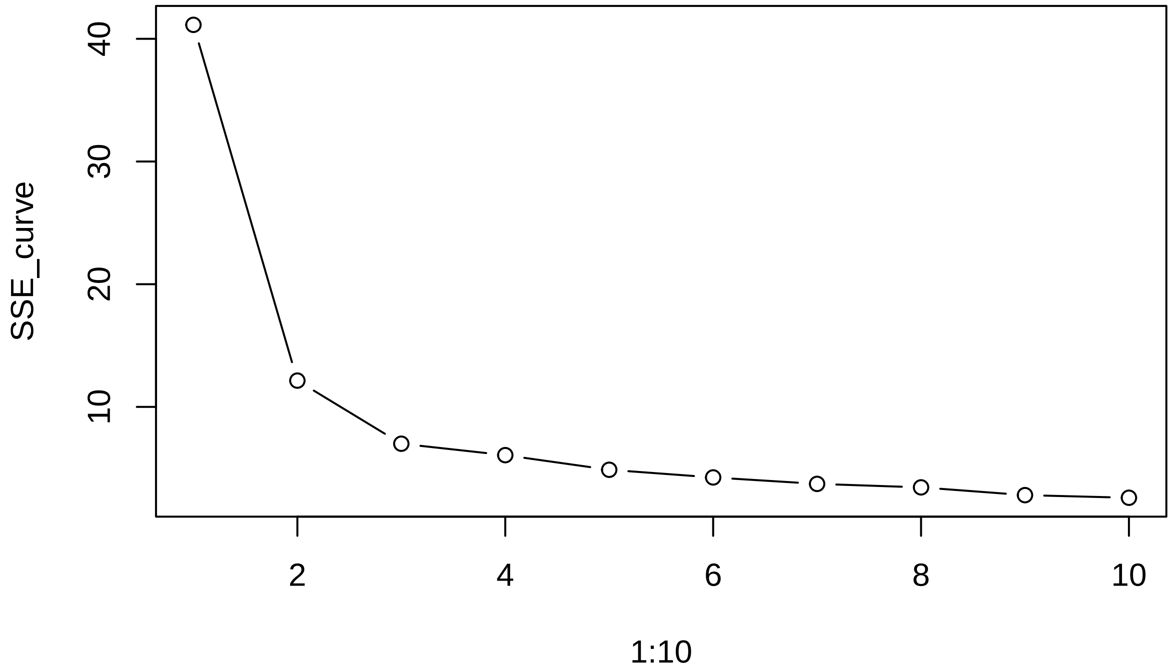
Make a plot of SSE against 2 - 10 clusters.

Based on the SSE plot, how many clusters do you think there are?

Answer: Based on the SSE plot, I think there are two clusters.

```
SSE_curve <- c()
for (n in 1:10) {
  kcluster = kmeans(iris_normalized[, 1:4], n)
  sse = kcluster$tot.withinss
  SSE_curve[n] = sse
  }

plot(1:10, SSE_curve, type = "b")
```


SSE_curve
40
30
20
10
2
4
6
8
10
1:10