tharunte_Homework1

2024-09-12

Q1. Consider the USArrests data. We will now perform hierarchical clus- tering on the states. (a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states. (b) Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters? 12.6 Exercises 551 (c) Hierarchically cluster the states using complete linkage and Eu- clidean distance, after scaling the variables to have standard de- viation one. (d) What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer.

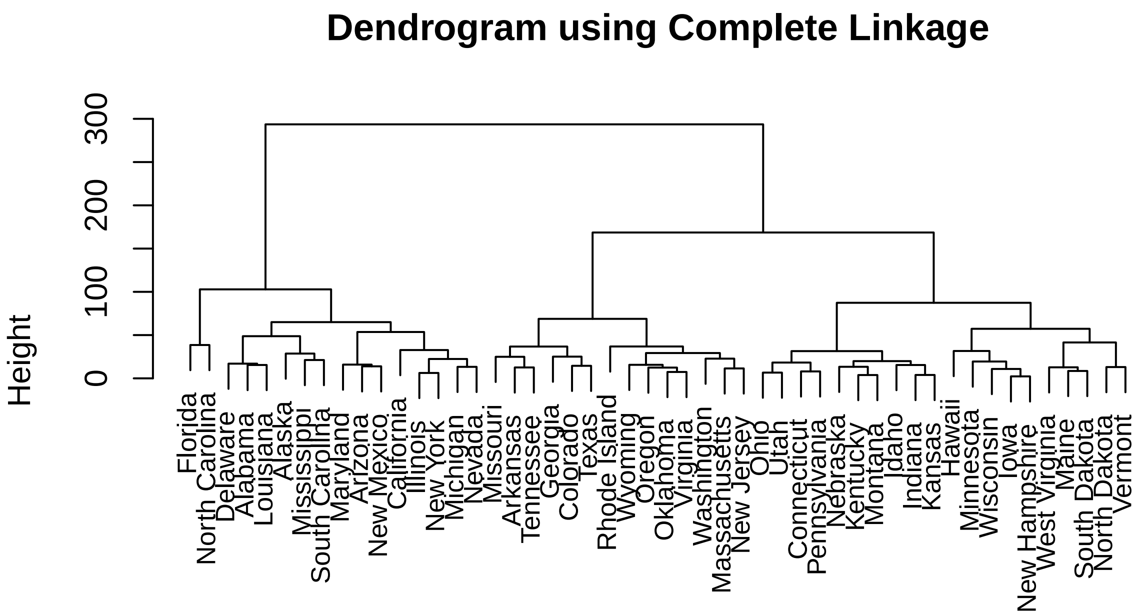
Load and view the dataset

```
data("USArrests")
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape
## 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
```

Without scaling the data

```
hc.complete <- hclust(dist(USArrests), method = "complete")
plot(hc.complete, main = "Dendrogram using Complete Linkage", xlab ="", sub = "", cex =0.8)
```



Cutting the dendrogram to get 3 clusters

```
clusters <- cutree(hc.complete,3)  # cuts the dendrogram to get 3 clusters
USArrests$Cluster <- clusters# adds a column to the table
clusters
```

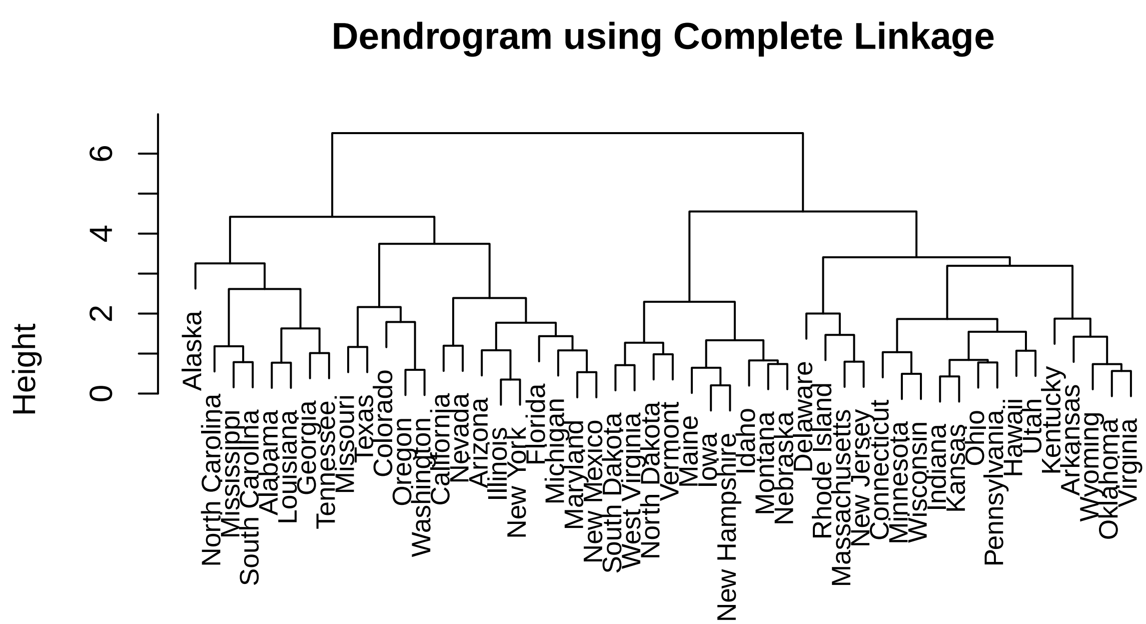
```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              1              2              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              3              1              1              2
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              3              1              3              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              3              3              1              3              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              2              1              3              1              2
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              3              3              1              3              2
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              1              3              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              2              2              3              2              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              3              2              2              3              3
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              2              2              3              3              2
```

```r
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape Cluster
## Alabama      13.2     236       58 21.2       1
## Alaska       10.0     263       48 44.5       1
## Arizona       8.1     294       80 31.0       1
## Arkansas      8.8     190       50 19.5       2
## California    9.0     276       91 40.6       1
## Colorado      7.9     204       78 38.7       2
```

After Scaling the data

```r
# Scaling the data to have mean 0 and standard deviation 1
scaled_USArrests <- scale(USArrests)
hc.complete_sc = hclust(dist(scaled_USArrests), method = "complete")
plot(hc.complete_sc, main ="Dendrogram using Complete Linkage", xlab = "", sub = "", cex = 0.8)
```



Cutting the dendrogram to get 3 clusters

```r
clusters_sc <- cutree(hc.complete_sc,3)
USArrests$Cluster <- clusters_sc
clusters_sc
```

```
##        Alabama         Alaska        Arizona       Arkansas     California 
##              1              1              1              2              1 
##       Colorado    Connecticut       Delaware        Florida        Georgia 
##              1              2              2              1              1 
##         Hawaii          Idaho       Illinois        Indiana           Iowa 
##              2              3              1              2              3 
##         Kansas       Kentucky      Louisiana          Maine       Maryland 
##              2              2              1              3              1 
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri 
##              2              1              2              1              1 
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey 
##              3              3              1              3              2 
##     New Mexico       New York North Carolina   North Dakota           Ohio 
##              1              1              1              3              2 
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina 
##              2              1              2              2              1 
##   South Dakota      Tennessee          Texas           Utah        Vermont 
##              3              1              1              2              3 
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming 
##              2              1              3              2              2
```

```r
head(USArrests)
```

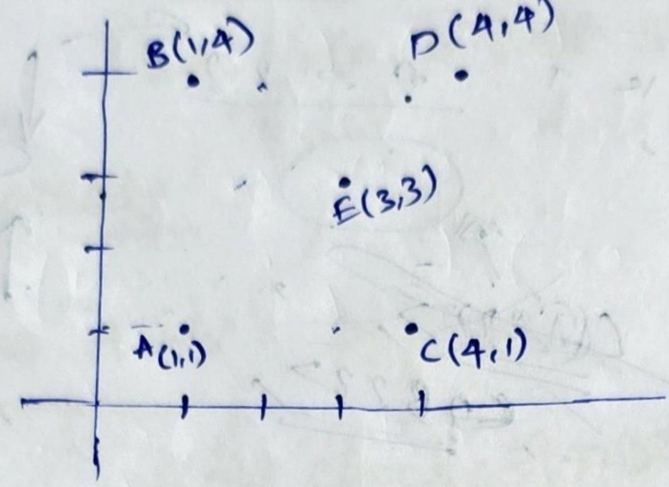
```
##            Murder Assault UrbanPop Rape Cluster
## Alabama      13.2     236       58 21.2       1
## Alaska       10.0     263       48 44.5       1
## Arizona       8.1     294       80 31.0       1
## Arkansas      8.8     190       50 19.5       2
## California    9.0     276       91 40.6       1
## Colorado      7.9     204       78 38.7       1
```

d) Scaling variables before computing hierarchical clustering is crucial because it ensures that all features contribute equally to the distance measurements. Without scaling, variables with larger ranges dominate the distance calculations, which can skew the clustering results. Scaling brings all features to the same scale, so each one has an equal impact on the clustering process. This leads to more meaningful and balanced clusters, reflecting true similarities between observations rather than being biased by the scale of individual features.

Homework:-

data points:- A(1,1), B(1,4), C(4,1), D(4,4), E(3,3)

Graph



→ Inital Random Centroids:-

Centroid 1: (3,1)

Centroid 2: (2,4)

distance of each data point from centroids:-

Euclidean distance:- $d(p,q) = \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$

Distance from Centroid 1 (3,1):

→ Distane A(1,1) = $\sqrt{(3-1)^2+(1-1)^2} = \sqrt{4} = 2$

→ Distance B(1,4) = $\sqrt{(3-1)^2+(1-4)^2} = \sqrt{4+9} = \sqrt{13} = 3.61$

→ Distance C(4,1) = 1

→ Distance D(4,4) = 3.16

→ Distance E(3,3) = 2

Distane from Centroid 2 (2,4):

→ Distance A(1,1) = $\sqrt{(2-1)^2+(4-1)^2} = \sqrt{1+9} = \sqrt{10} = 3.16$

→ Distance B(1,4) = $\sqrt{(2-1)^2+(4-4)^2} = \sqrt{1} = 1$

→ Distanee C(4,1) = $\sqrt{(2-4)^2+(4-1)^2} = 3.61$

→ Distance D(4,4) = 2

→ Distance E(3,3) = 1.41.

Clusters 1:- A, C,

Cluster 2: B, D, E

⇒ New centroid 1:- A(1,1) and C(4,1)

$$\left(\frac{1+4}{2}, \frac{1+1}{2}\right) = (2.5, 1)$$

New centroid 2:- B(1,4) D(4,4) and E(3,3)

$$\left(\frac{1+4+3}{3}, \frac{4+4+3}{3}\right) = (2.67, 3.67)$$

Final cluster are A, C and B, D, E.

Even after 2nd itteration the cluster remanis same. Hence we reached the ~~conclusion~~ convergence.