Clustering US states on the basis of Arrests in each state

Saturday, March 7, 2020 7:53 PM

Data Source : USArrests data

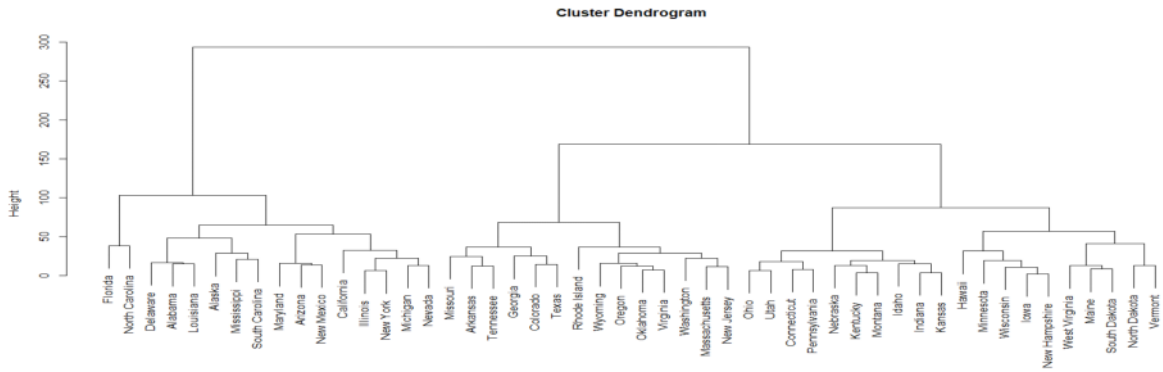
Observations :

- There are 4 different kinds of observations - murder, Assault, Urban population and Rape for all 50 states of USA.
- This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas

Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.

Steps :

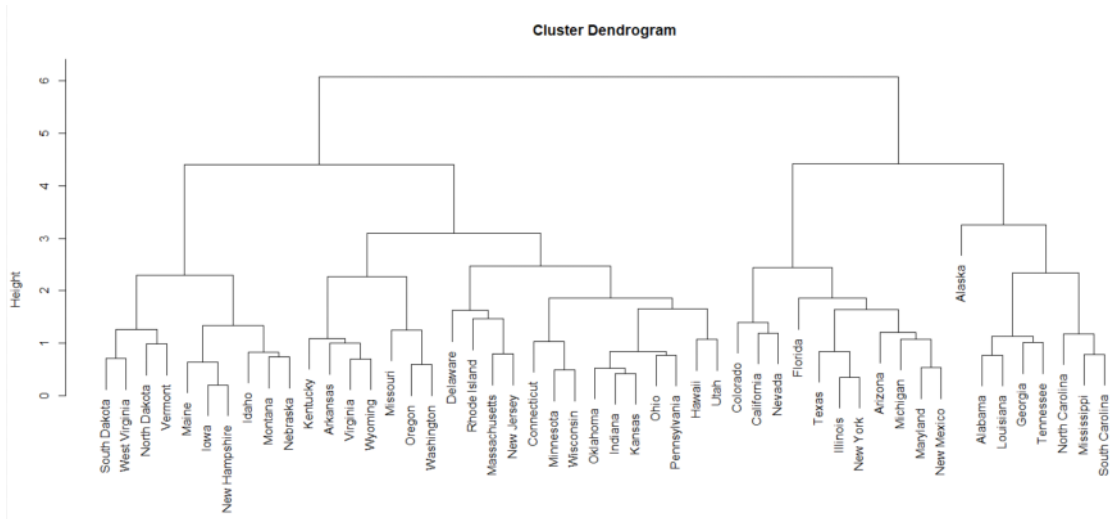
1. Find dissimilarity matrix for the dataset.
2. Clustering the US states using complete linkage and hierarchial clustering.



Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?

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

Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one.



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.

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

Inference :

- Scaling has reduced the height of the tree.
- The number of states included in each cluster as well as the clusters of the states have changed in the two cases
- It makes sense to scale and build the dendrogram as the Urban Population variables has a different scale. Urban Population variable is percent of urban population whereas other variables are numeric.