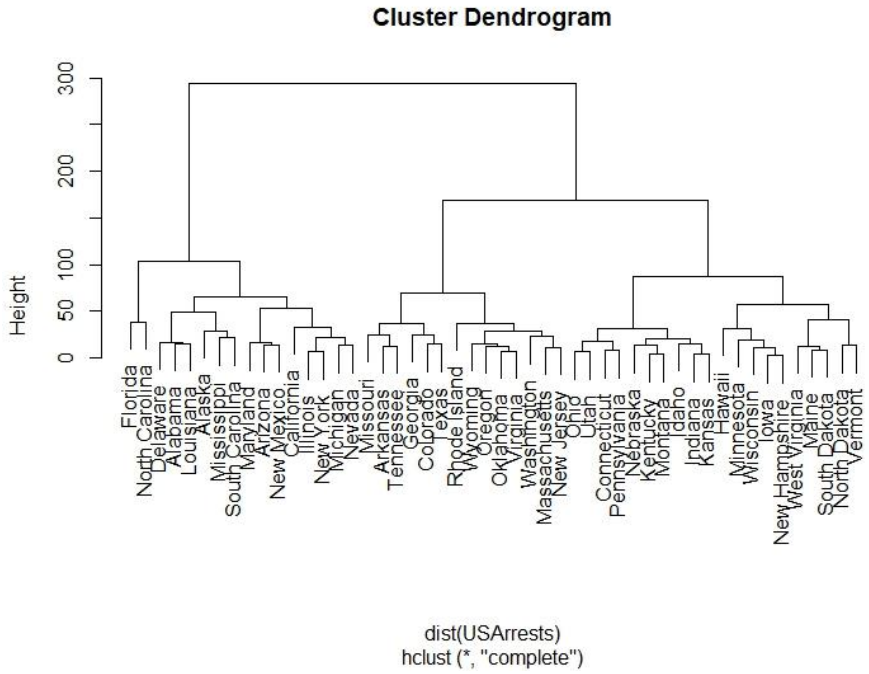
STA Home work - 2

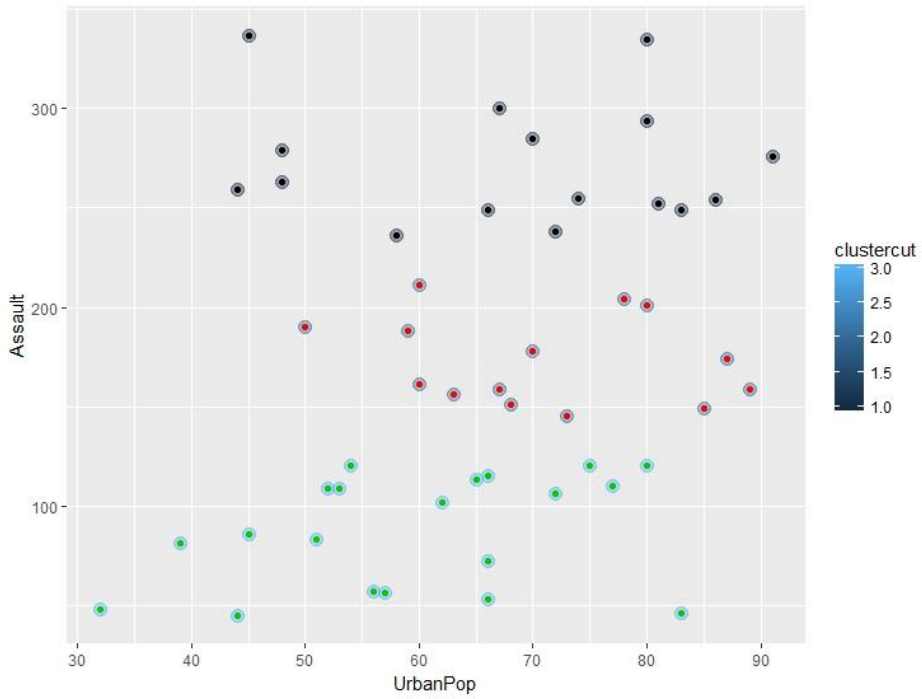
9)

a) The tree structure obtained after performing hierarchical clustering on the USArrests data based on calculating Euclidean distance and Complete linkage.



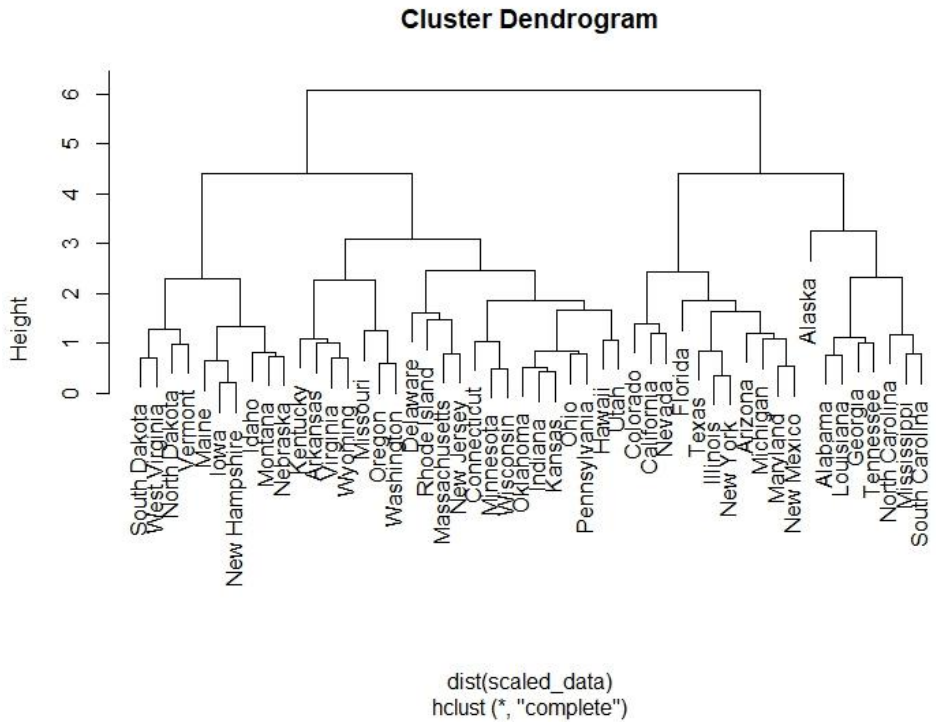
b) After cutting the dendogram at a height that divides the whole structures into three clusters. These are the states that belong to three different clusters.

```
> names(clust_1)
 [1] "Alabama"        "Alaska"         "Arizona"        "California"     "Del
aware"        "Florida"        "Illinois"
 [8] "Louisiana"      "Maryland"       "Michigan"       "Mississippi"    "Nev
ada"          "New Mexico"     "New York"
[15] "North Carolina" "South Carolina"
> names(clust_2)
 [1] "Arkansas"      "Colorado"      "Georgia"       "Massachusetts" "Missour
i"       "New Jersey"    "Oklahoma"      "Oregon"
 [9] "Rhode Island"  "Tennessee"     "Texas"         "Virginia"      "Washing
ton"     "Wyoming"
> names(clust_3)
 [1] "Connecticut"   "Hawaii"        "Idaho"         "Indiana"       "Iowa"
"Kansas"        "Kentucky"      "Maine"
 [9] "Minnesota"     "Montana"       "Nebraska"      "New Hampshire" "North D
akota"  "Ohio"          "Pennsylvania"  "South Dakota"
[17] "Utah"          "Vermont"       "West Virginia" "Wisconsin"
```



From the above plot, it can be observed that Assaults level are divided into three clusters and assaults are more where urban population is high.

c) Scaling the variables to Standard Deviation one using scale function and then the obtained dendogram is



d)

Unscaled:

```
clustercut
 1  2  3
16 14 20
```

Scaled:

```
           cluster_sd_cut
clustercut  1  2  3
         1  6  9  1
         2  2  2 10
         3  0  0 20
```

The main effect of scaling the variables on clustering is that there is no longer a logical point. In the scale d version, it would appear that only few clusters that are more logical. We can decrease this by the length of the branches in the dendogram, but an analysis on the included in cluster is showing a migration of stat es from original cluster to another cluster, if we cut the tree and make it three clusters.

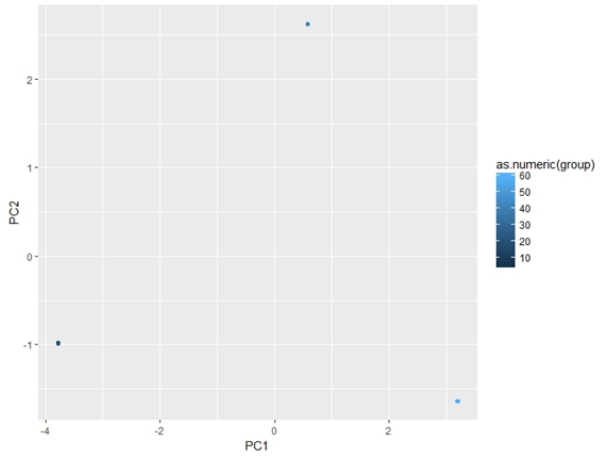
We can see that, in unscaled clustering that 8 of the 16 states in cluster 1 moved to cluster 3 of the scaled version and 3 moved from cluster 2 to cluster 3. Another look at the scaled dendrogram will show that clu ster 3 can easily be split into at least two separate clusters.

Here, in this situation scaling is appropriate since the ranges of Murder, Assault, Rape and Urban Pop are in a different unit of measurement. We will find more meaningful results when these in a same unit or me asurement or proportional.

10)

a) Generate 20 observations for each of the three classes and used rnorm function to create a matrix of 50 columns and 60 rows. Add a mean shift to the observations in each class so there will be three different classes.

b)



Inference: The three classes appear separated after plotting the first two principal component vectors.

c)

clusters	class labels 1	2	3
1	0	20	0
2	20	0	0
3	0	0	20

We can see that the K-means clustering with order 3 separates the three of the Classes perfectly into 3 distinct Clusters.

d)

clusters	class labels 1	2	3
1	20	0	0
2	0	20	20

Here also, we can see that the K-means clustering with order 2 separates the three of the Classes perfectly into 3 distinct Clusters.

e)

clusters	class labels 1	2	3
1	12	0	0
2	0	0	20
3	8	0	0
4	0	20	0

This is K-means clustering with order 4. Majority of Class 1 is assigned to Cluster 1. However k-means seemed to have forced some observations from Class 1 into Cluster 3. And class 2, class 3 are entirely assigned to clusters 4 and 2.

f)

clusters	class labels 1	2	3
1	0	0	20
2	0	20	0
3	20	0	0

We can see that the K-means clustering with order 3 on the first two principal component score vectors perfectly separated the observations into three clusters.

g)

clusters	class labels 1	2	3
1	7	5	2
2	4	8	11
3	9	7	7

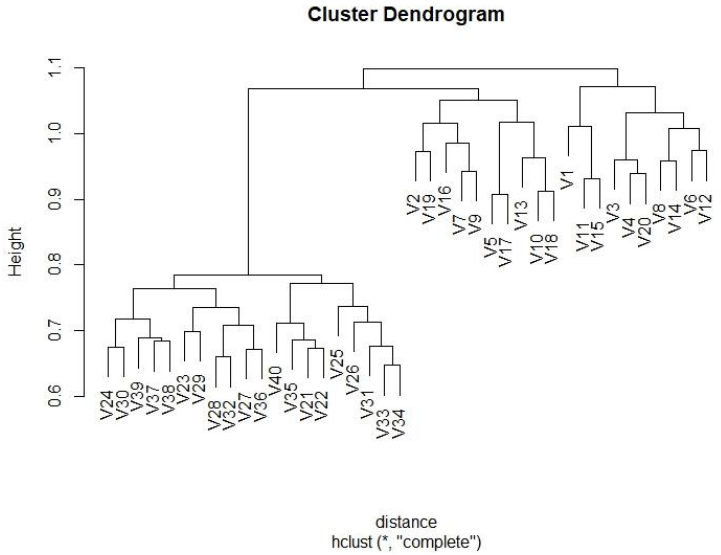
We can see that the K-means clustering with order 3 on the scaled data didn't divide the observations into clusters compared to unscaled data. Because scaling effects the distance between the observations.

11)

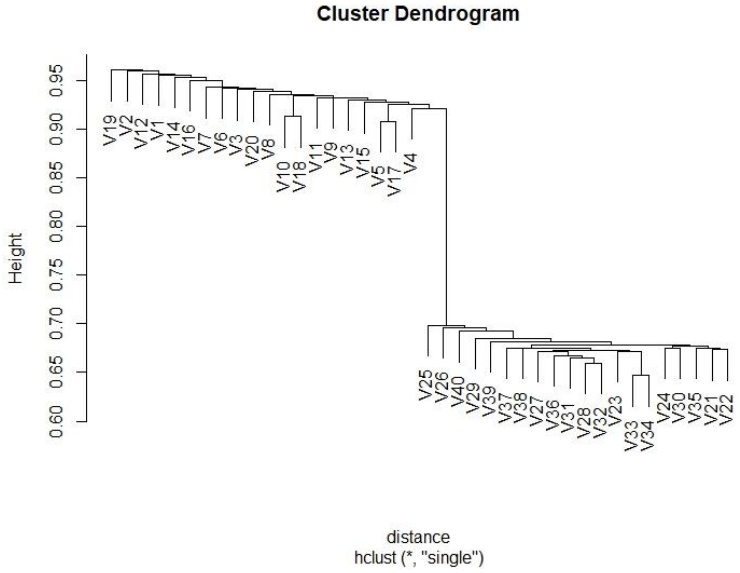
a) Load the genes data using read.csv keeping header as FALSE cause we don't have any gene names.

b) Apply hierarchical clustering on the genes data using complete, single and average linkage.

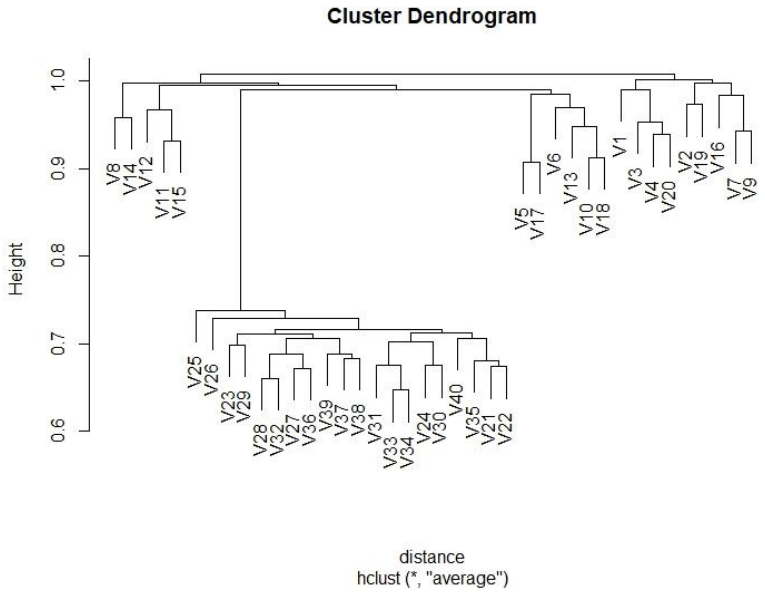
-> Complete Linkage



-> Single Linkage



-> Average Linkage



Inference:

The choice of linkage does affect the results obtained. From the above plots, we can observe that results obtained are different for linkages. Complete and single linkage have three clusters, whereas average linkage has three clusters. Average and complete linkage tend to balance the clusters, and single cluster tend to yield trailing clusters where observations are attached one by one to a large cluster. This is why complete and average linkages are preferred more.

c) Perform PCA on the data to see which genes differ among the healthy and diseased. By giving the scale as TRUE, the variables and scaled to standard deviation one and by prcomp() mean becomes zero for the variables.

The rotation matrix has the principal component loadings, each column contains principal component vector. The loadings of this can be considered as the weight of each gene in both the groups (healthy and diseased).