Problem 1:

a)

Cluster Dendrogram
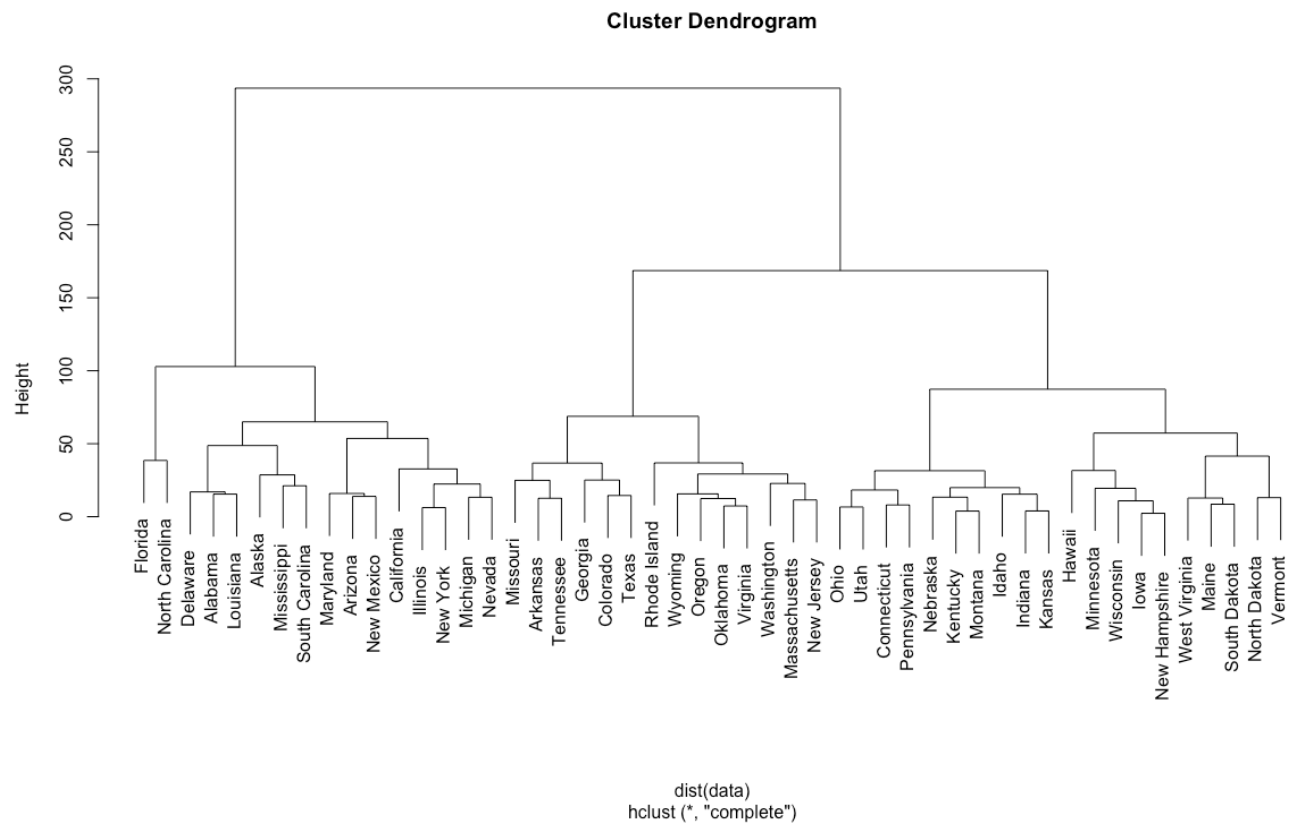

Cluster Dendrogram after using hierarchical clustering with complete linkage and Euclidean distance.

b) Cutting the dendrogram at a height that results in three distinct clusters:

```
> cutree(hierclust.compl, 3)
       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

> table(cutree(hierclust.compl, 3))

 1  2  3
16 14 20
```

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

c) Scaling the variables to have standard deviation one:

Cluster Dendrogram

dist(dsc)
hclust (*, "complete")
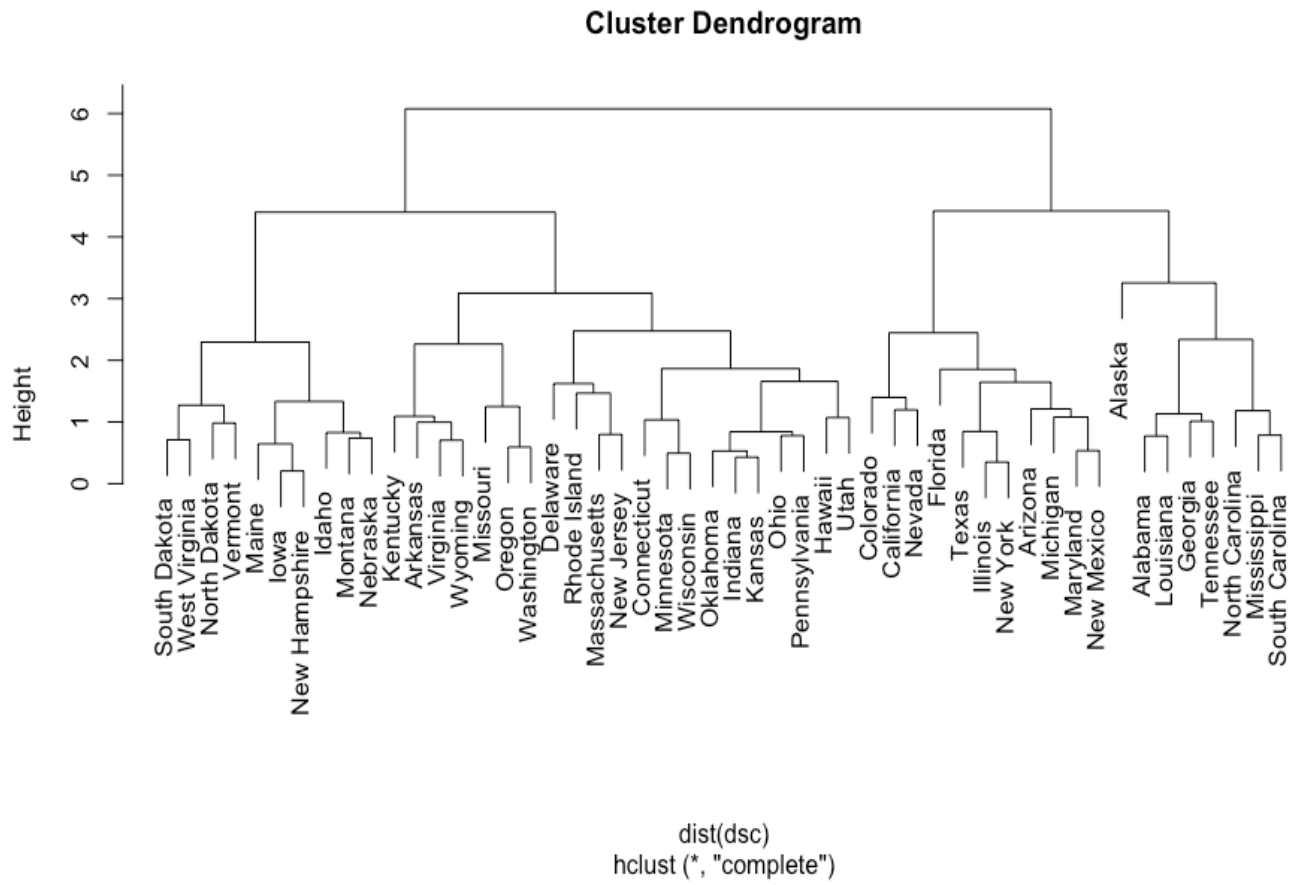
d)

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

 1  2  3
 8 11 31
```

- Although the trees are similar, scaling the variables affects the clusters.

- Maximum height of the dendrogram obtained from hierarchical clustering is affected by scaling the variables.

- Just looking at this, it seems as if it doesn't affect the bushiness of the tree obtained.

- However, clusters obtained from cutting the dendrogram into 3 clusters are affected.

- Scaling the covariates and centering changes the structure of the tree. It is no longer possible to group the states into just 3 clusters.

- It also results in better grouping of similar states together and should always be done for this data set.

- In my opinion, variables should be scaled before the inter-observation dissimilarities are computed and as the data measured has different units, the data should be standardized for this dataset.