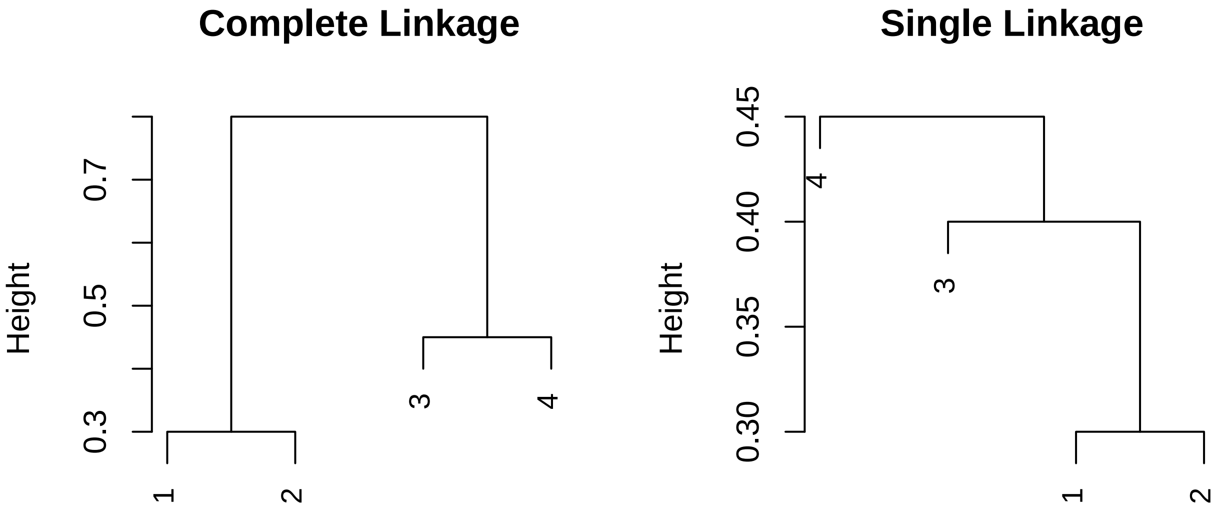
STATS 415 HW11

Wenfei Yan (wenfeiy). GSI: April Cho.

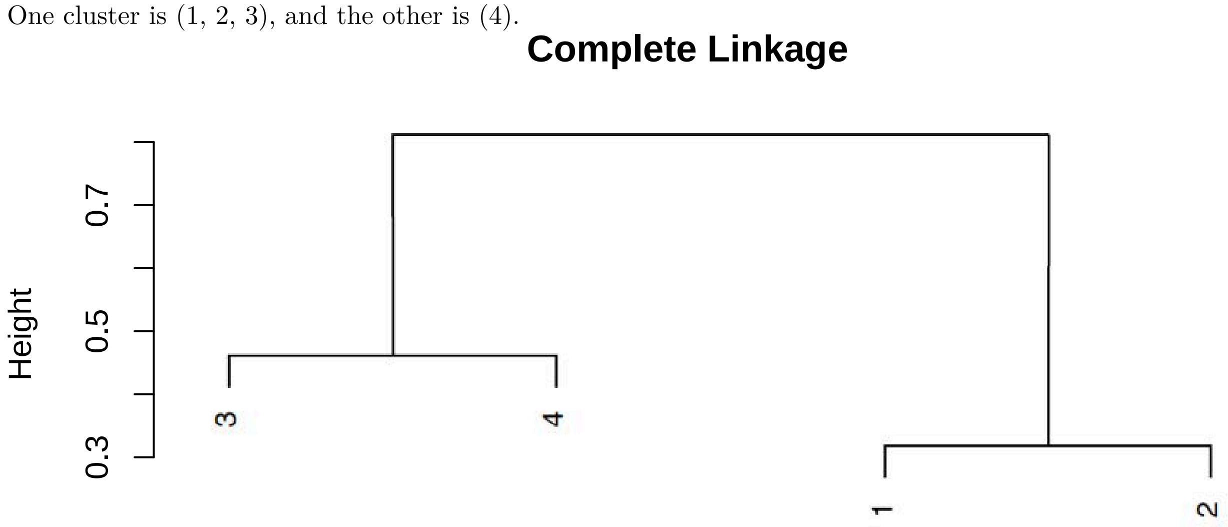
1. (a)(b) Denote the four observations by 1, 2, 3, 4 from left to right in the dissimilarity matrix. The dendrograms using complete linkdage and single linkage are shown in the following graph.

```r
distance = matrix(c(0, 0.3, 0.4, 0.7, 0.3, 0, 0.5, 0.8, 0.4, 0.5, 0, 0.45, 0.7, 0.8, 0.45, 0), nrow = 4)
hc.complete=hclust(as.dist(distance), method="complete")
hc.single=hclust(as.dist(distance), method="single")
par(mfrow=c(1,2))
plot(hc.complete,main="Complete Linkage", xlab="", sub="", cex=.9)
plot(hc.single, main="Single Linkage", xlab="", sub="", cex=.9)
```



(c) One cluster is (1, 2), and the other is (3, 4).

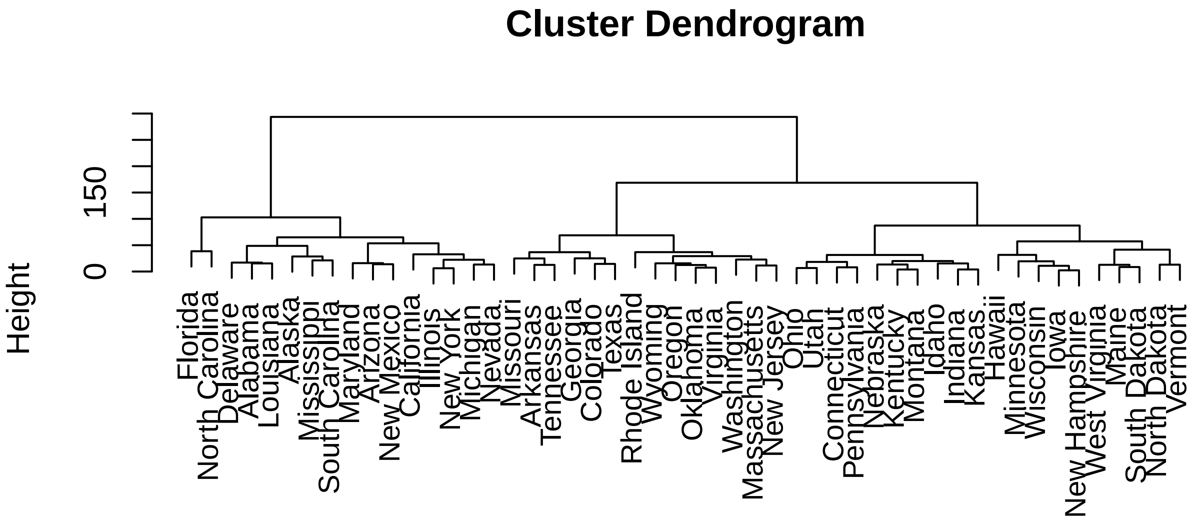
(d) One cluster is (1, 2, 3), and the other is (4).



(e)

2. (a)

```r
library(ISLR)
X = as.matrix(USArrests)
hc.complete=hclust(dist(X), method="complete")
plot(hc.complete, xlab="", sub="", cex=.9)
```

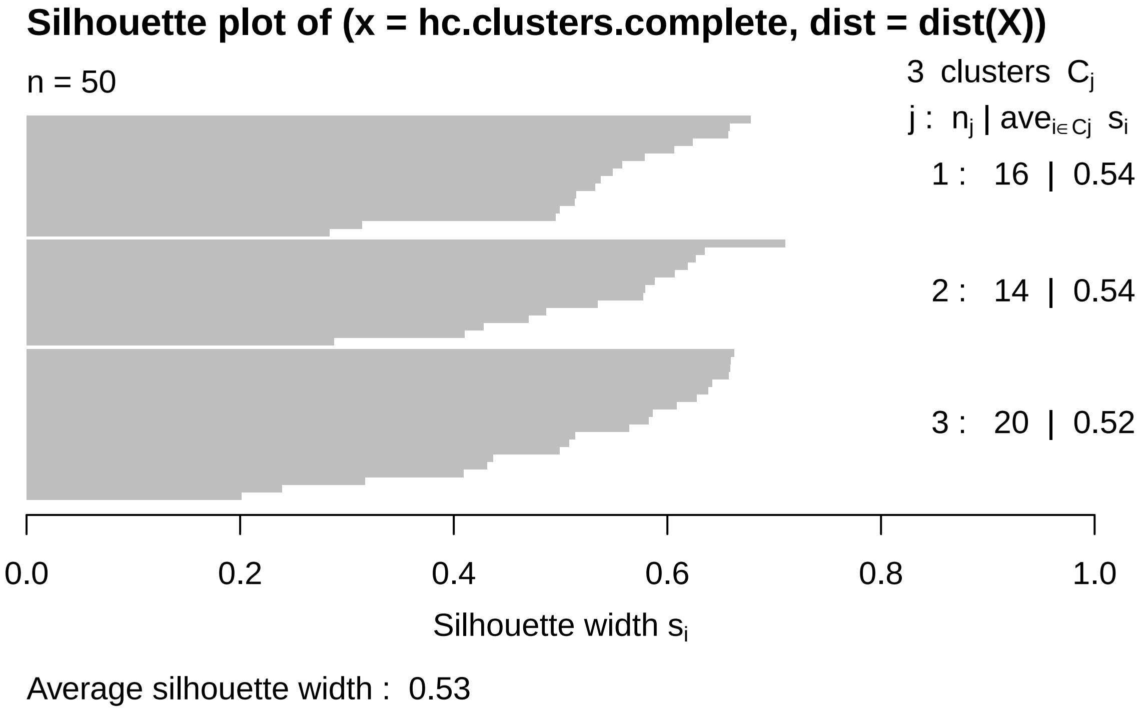


(b)

```r
hc.clusters.complete = cutree(hc.complete,3)
table(hc.clusters.complete, rownames(USArrests))
```

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

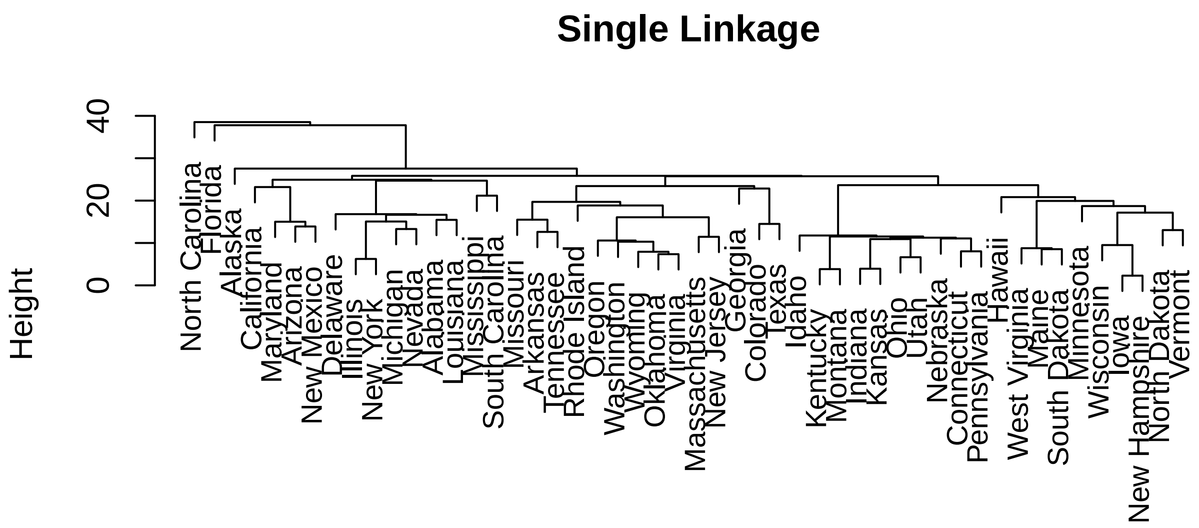
```r
library(cluster)
plot(silhouette(hc.clusters.complete, dist = dist(X)))
```



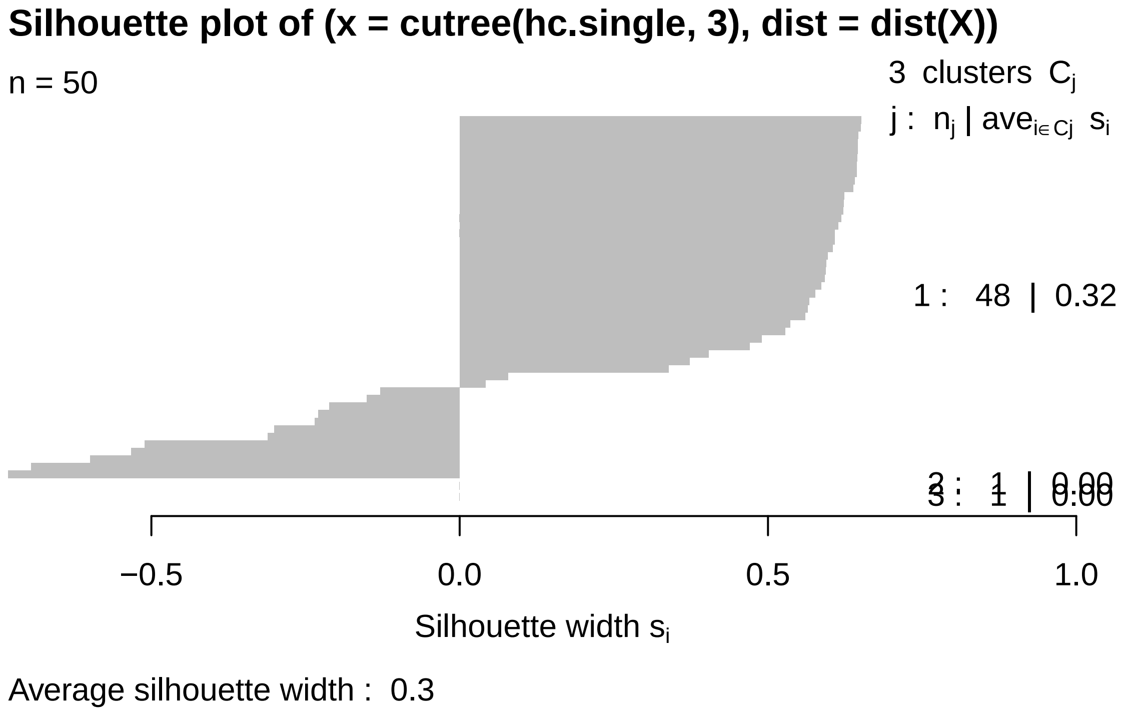
The silhouette widths of each cluster are very close and the sizes of each cluster are also similar, which means the qualities of each cluster are similar. Also there's no point with negative silhouette coefficients, which means the clustering is pretty good in terms of the silhouette measure.

(c)

```
hc.single=hclust(dist(X), method="single")
plot(hc.single,main="Single Linkage", xlab="", sub="", cex=.9)
```



```
plot(silhouette(cutree(hc.single, 3), dist = dist(X)))
```



```
table(cutree(hc.single, 3), rownames(USArrests))
```

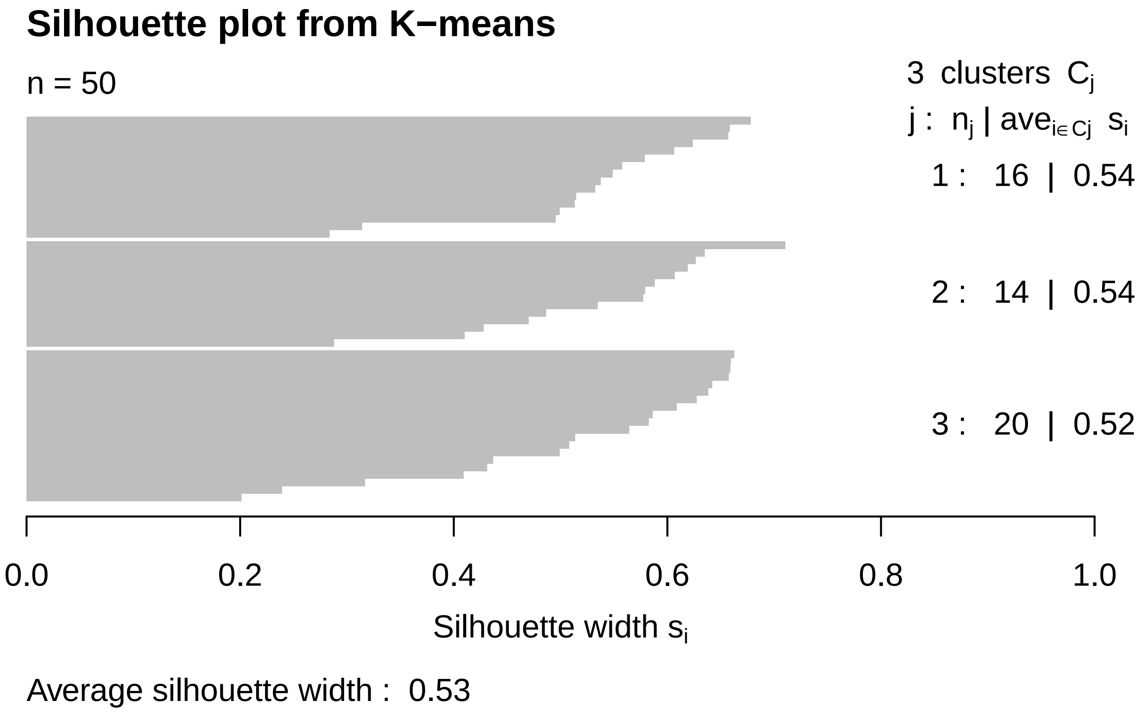
Cluster 1: Alabama, Alaska, Arizona, California, Delaware, Illinois, Louisiana, Maryland, Michigan, Mississippi, Nevada, New Mexico, New York, South Carolina, Arkansas, Colorado, Georgia, Massachusetts, Missouri, New Jersey, Oklahoma, Oregon, Rhode Island, Tennessee, Texas, Virginia, Washington, Wyoming, Connecticut, Hawaii, Idaho, Indiana, Iowa, Kansas, Kentucky, Maine, Minnesota, Montana, Nebraska, New Hampshire, North Dakota, Ohio, Pennsylvania, South Dakota, Utah, Vermont, West Virginia, Wisconsin. Cluster 2: Florida. Cluster 3: North Carolina.
The clusters are extremely skewed. Most points are in cluster 1, and there's only one in cluster 2 and another one in cluster 3. The clustering result with single linkage is not very ideal.
(d)

```
set.seed(1111)
km.out=kmeans(dist(X), 3, nstart=20)
km.clusters=km.out$cluster
table(km.clusters, hc.clusters.complete)
```

```
##             hc.clusters.complete
## km.clusters  1  2  3
##           1 16  0  0
##           2  0 14  0
##           3  0  0 20
```

Here I assign nstart=20 to use multiple random assignments for initial clusters, and the kmeans() function will report only the best result with smallest within cluster sum of squares. The clustering results of the k-means clusteirng and the hierarchical clustering with complete linkage are exactly the same, so I omit the result of "reporting which states belong to which clusters is".

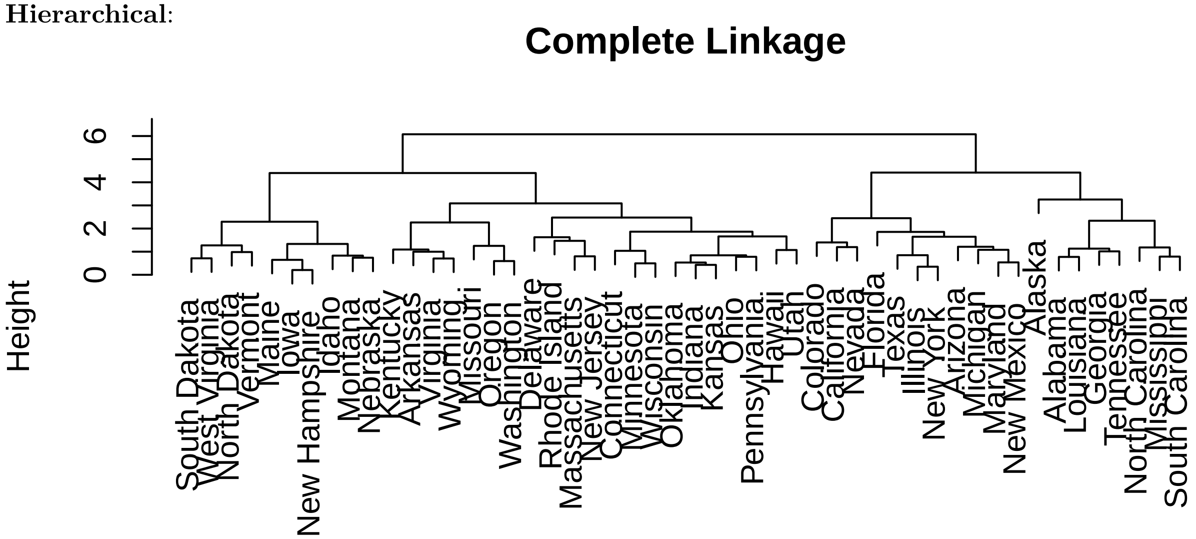
```
plot(silhouette(km.clusters, dist(X)), main="Silhouette plot from K-means")
```



(e) I scale the variables by the scale() function, and I omit other codes, which is the same as the previoius, except that I use dist(sdX) instead of dist(X).

```
sdX = scale(X)
```

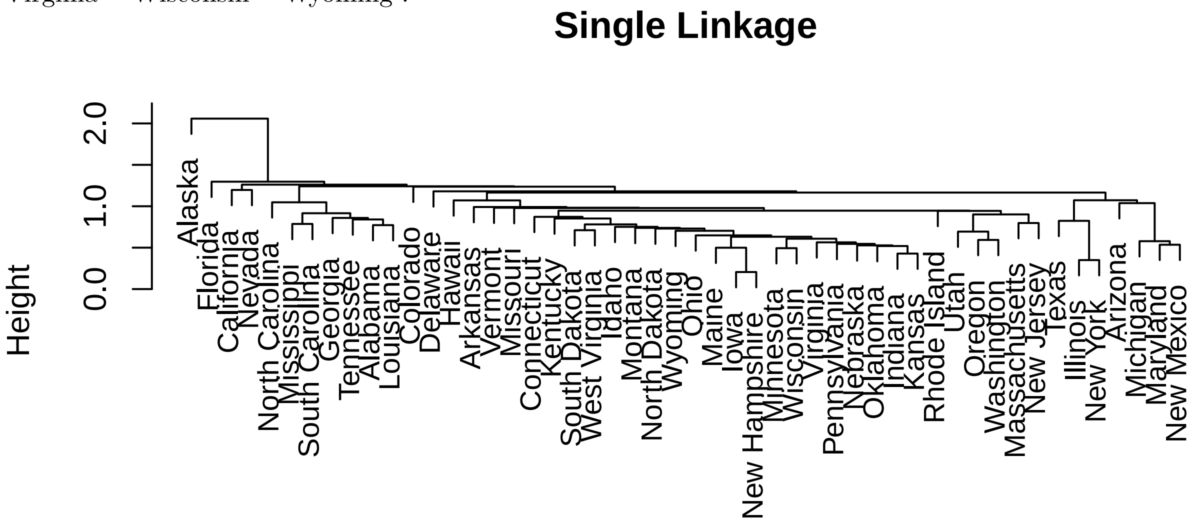
Hierarchical:



Complete linkage clusters are
Cluster 1: Alabama, Alaska, Georgia, "Louisiana" "Mississippi" "North Carolina" "South Carolina" Tennessee.
Cluster 2: "Arizona" "California" "Colorado" "Florida" "Illinois" "Maryland" "Michigan" "Nevada" "New Mexico" "New

York” “Texas”.

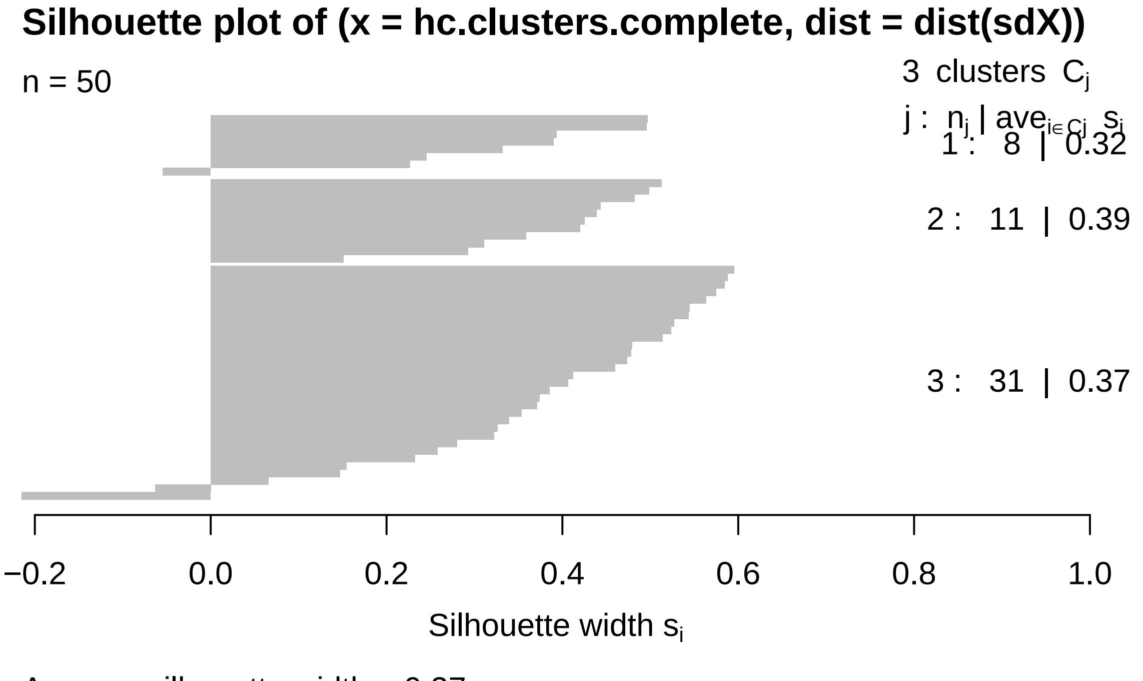
Cluster 3: “Arkansas” “Connecticut” “Delaware” “Hawaii” “Idaho” “Indiana” “Iowa” “Kansas” “Kentucky” “Maine” “Massachusetts” “Minnesota” “Missouri” “Montana” “Nebraska” “New Hampshire” “New Jersey” “North Dakota” “Ohio” “Oklahoma” “Oregon” “Pennsylvania” “Rhode Island” “South Dakota” “Utah” “Vermont” “Virginia” “Washington” “West Virginia” “Wisconsin” “Wyoming”.

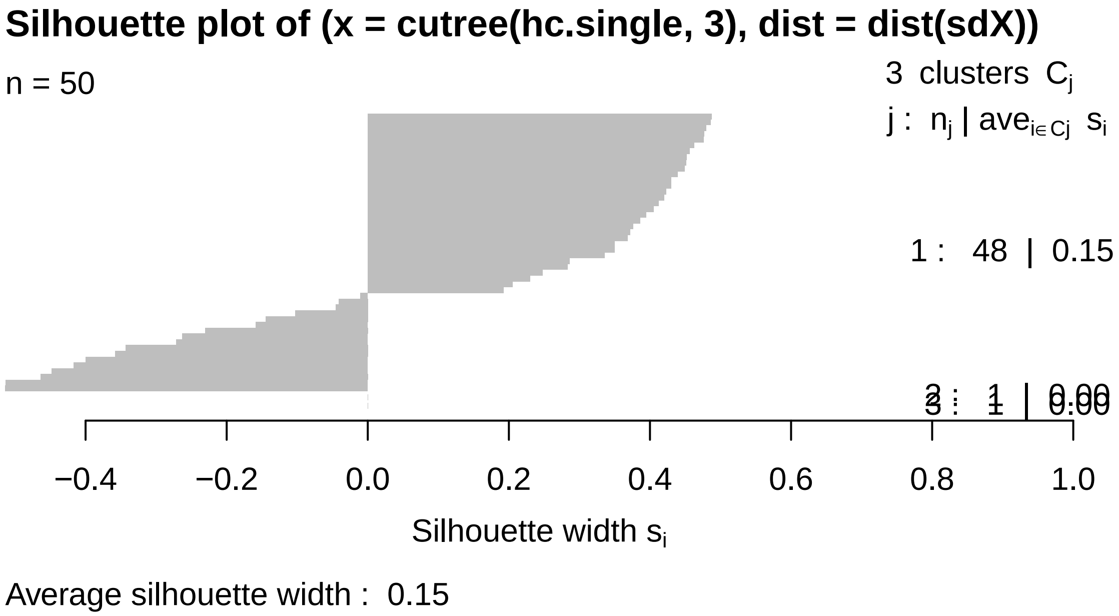


Single linkage clusters are

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

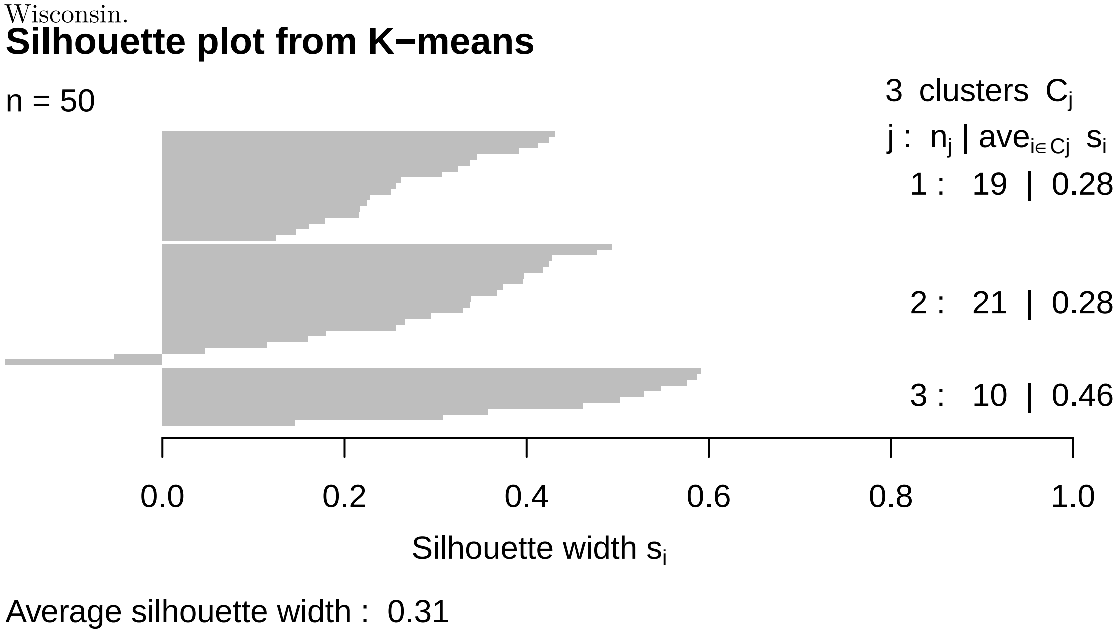
Cluster 2: Alaska. Cluster 3: Florida.





K-means:

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



(f) The scaling changes the cluster assignments of each points. For hierarchical clustering with complete linkaage, the sizes of each cluster becomes different, and there also appears points with negative silhouette coefficients, while previous there're no such points. The quality of the clustering seems to be worse in terms of the silhouette measure. For hierarchical clustering with single linkaage, the clustering becomes even worse. The average Silhouette width decreases, and there's more points with negative silhouette coefficients. For k-means clustering, the result is no longer the same as the hierarchical clustering using complete linkage. Also, there appear two points wiht negative sihouette width. The clustering quality also drops.
Yes, if we initially don't have any preference on any varaibles. If we don't scale the varaibles, the number of assualts is likely to domain the clustering.