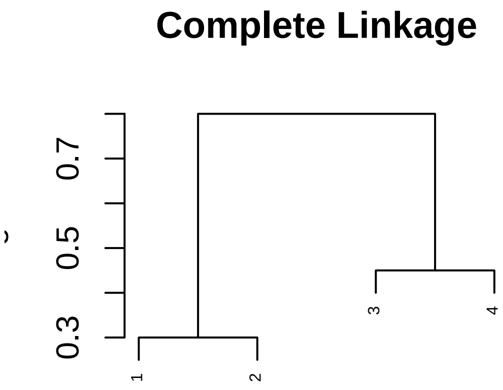
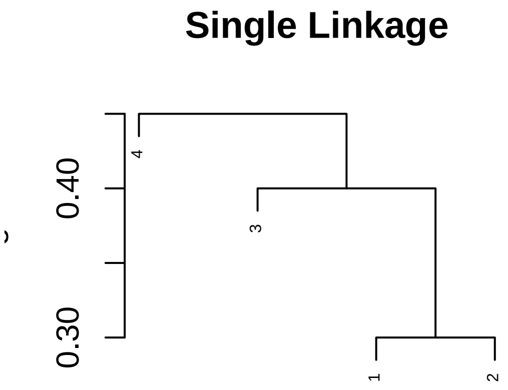
STATS415hw11

Name: Yunguo Cai uniqname: cyunguo lab section: 003

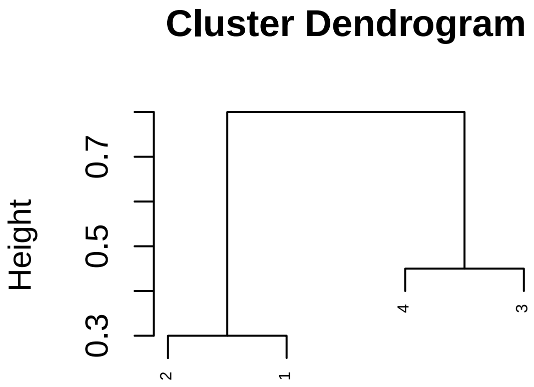




1.(a)(b)

(c) We have clusters (1,2) and (3,4).

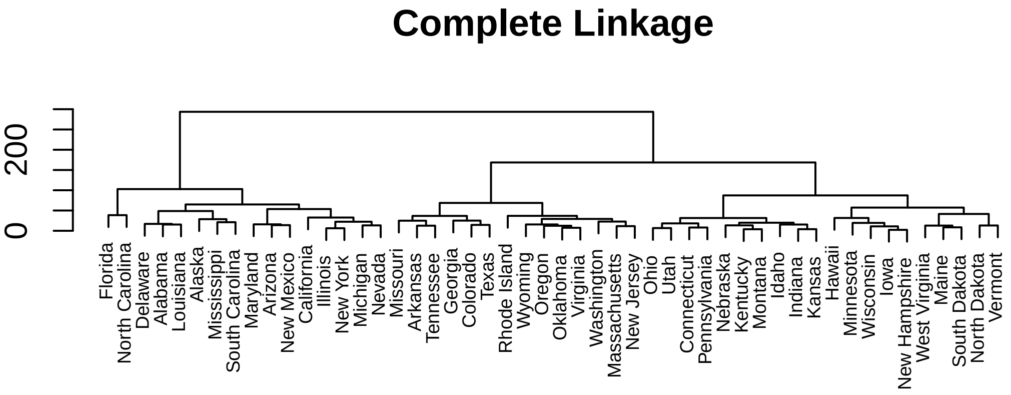
(d) We have clusters ((1,2),3) and (4).



(e)

2.(a)

```
attach(USArrests)
par(pin=c(5,1.5))
hc_complete = hclust(dist(USArrests, method = "euclidean"), method="complete")
plot(hc_complete,main="Complete Linkage",xlab="", sub="",ylab="",cex = 0.6)
```



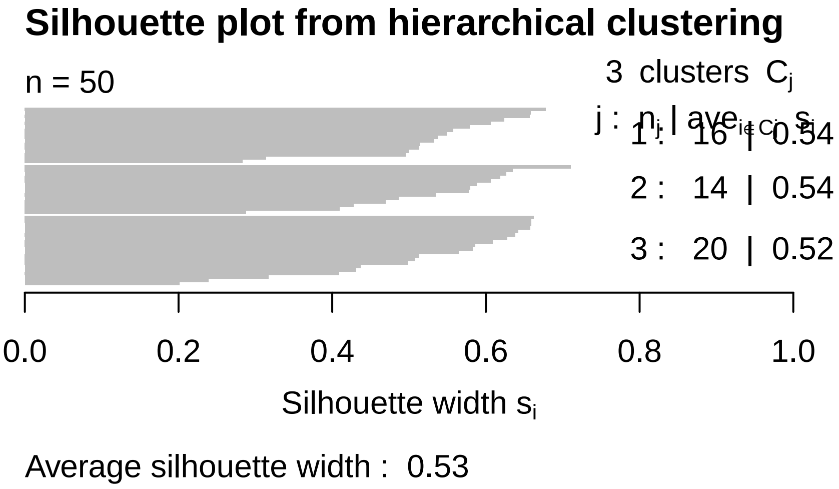
(b)

```
hc.clusters = cutree(hc_complete, 3)
```

```
clust_1 <- hc.clusters[hc.clusters == 1]
clust_2 <- hc.clusters[hc.clusters == 2]
clust_3 <- hc.clusters[hc.clusters == 3]
names(clust_1)
names(clust_2)
names(clust_3)
```

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.

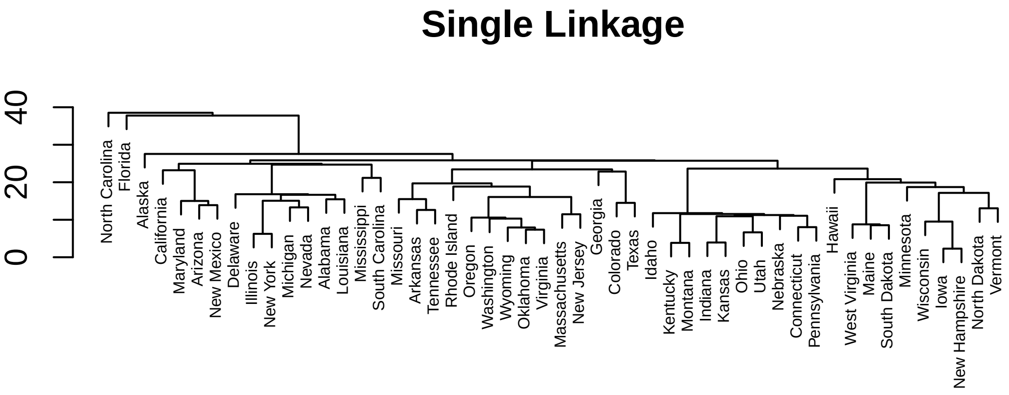
```
par(pin=c(4,1))
plot(silhouette(hc.clusters,dist(USArrests, method = "euclidean")), main = "Silhouette plot from hierarc
```



The silhouette widths of the three clusters are almost the same, size are also similar and there is no negative silhouette coefficient, so no state is misclustered and the clustering is good.

(c)

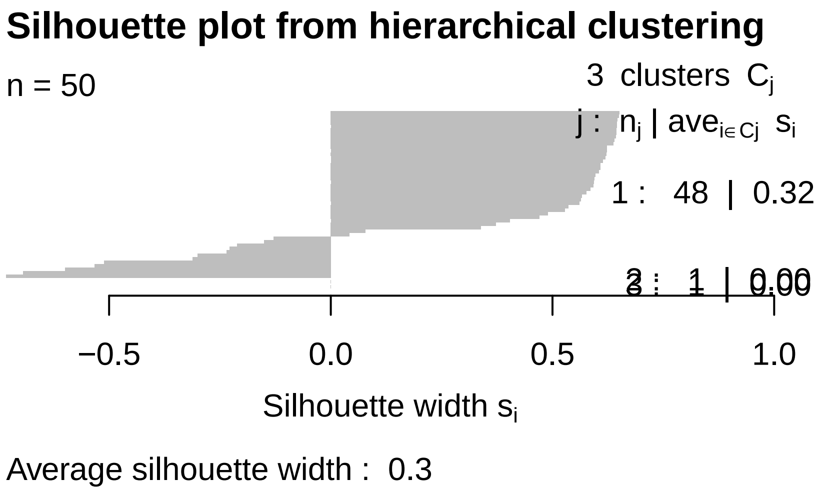
```
par(pin=c(5,1.5))
hc_single = hclust(dist(USArrests, method = "euclidean"), method="single")
plot(hc_single,main="Single Linkage",xlab="", sub="",ylab="",cex = 0.5)
```



```
hc.clusters2 = cutree(hc_single, 3)
```

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.

```
par(pin=c(4,1))
plot(silhouette(hc.clusters2,dist(USArrests, method = "euclidean")), main = "Silhouette plot from hiera
```



The silhouette of the three clusters are extremely skewed. There is only one point in cluster 2 and one point in cluster 3. There are negative silhouette widths, which means that misclustering might exist. The average silhouette width is 0.3, which is worse than complete linkage. The clustering with single linkage is not good.

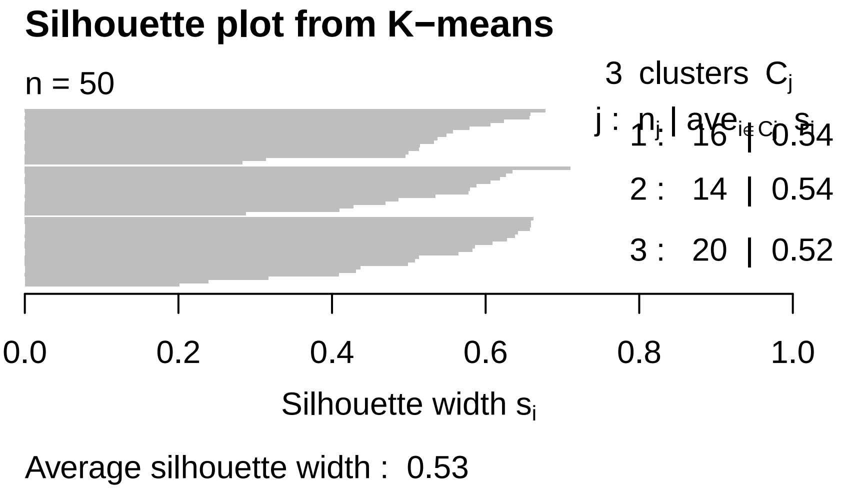
(d)

```
set.seed(1111)
km.out = kmeans(USArrests, 3, nstart = 50)
km.clusters=km.out$cluster
table(km.clusters, hc.clusters)
```

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

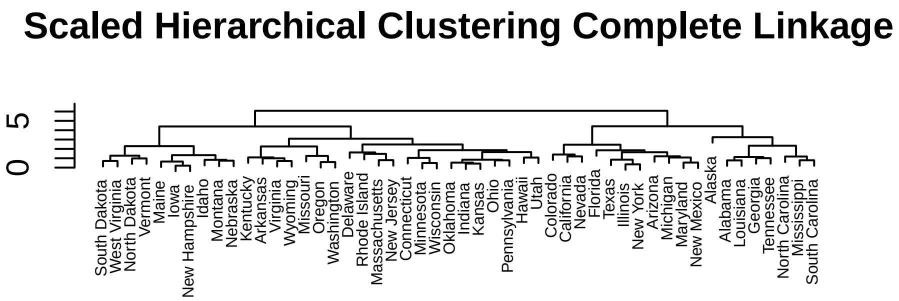
I set nstart=50 to random assign initial clusters for multiple times since I tried different values of nstart and they all converge to the same total within-cluster sum of squares and get the same clustering results. The clustering results of K-means clusteirng are the same as the hierarchical clustering, so the report of which states belong to which clusters and the comment on interesting features are omitted here.

```
par(pin=c(4,1))
plot(silhouette(km.out$cluster,dist(USArrests)),main="Silhouette plot from K-means")
```



(e) The codes are the same as before, so I omitted them.

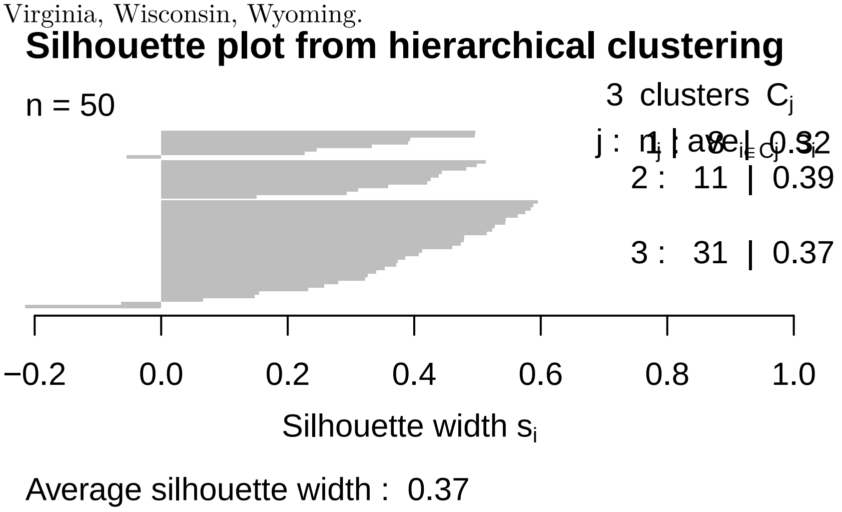
```r
set.seed(1111)
sd.arrests <- scale(USArrests)
```



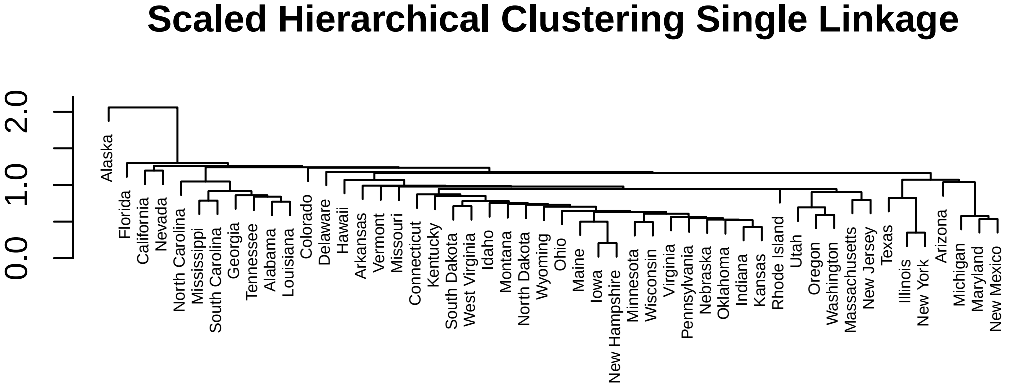
Hierarchical:

```r
hcsd.clusters = cutree(hcsd_complete, 3)
```

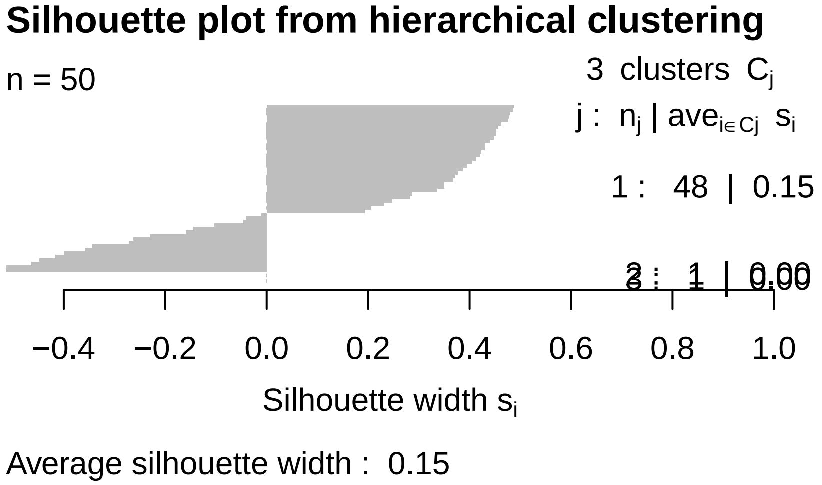
Cluster 1: Alabama, Alaska, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee.
Cluster 2: Arizona, California, Colorado, Florida, Illinois, Maryland, Michigan, Nevada, New Mexico, New York, Texas.
Cluster 3: Arkansas, Connecticut, Delaware, Hawaii, Idaho, Indiana, Iowa, Kansas, 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.



The average silhouette width is 0.37, smaller than 0.53 and there exist misclustering.The clusters are mixed with each other without a really clear boundary.



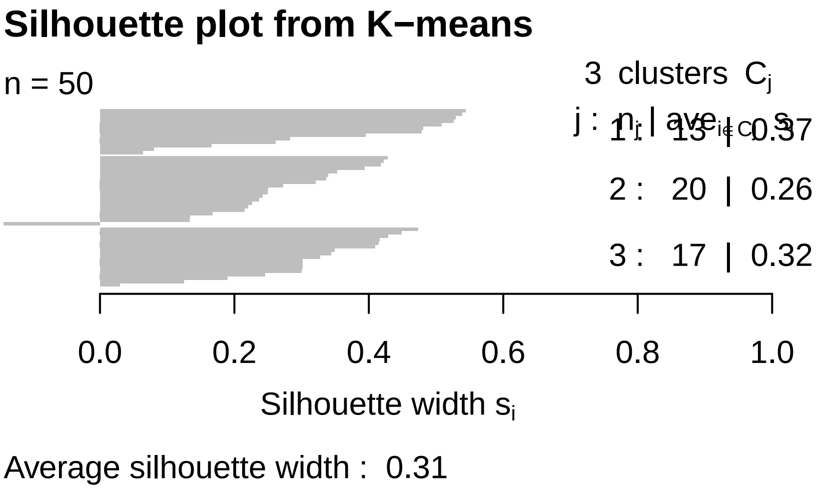
Cluster 1: Alabama, North Carolina, Arizona, California, Delaware, Illinois, Louisiana, Maryland, Michigan, Mississippi, Nevada, New Mexico, New York, South Carolina, Arkansas, Colorado, Georgia, Massachusetts, Missouri, New Jersey, Oklahoma, Oregon, Rhode Island, Tennessee, Texas, Virginia, Washington, 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, Wyoming.
Cluster 2: Alaska. Cluster 3: Florida.



The average silhoutte width is 0.15, smaller than 0.3. There exist about a half misclustering.

K-means:

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



The initialized algorithm is the same as (d). The average silhouette is 0.31, smaller than 0.53. Though, only one point has negative coefficient. In general it works well.

(f) Scaling the variables impacts the clusters assignments, the branch and height of the cluster dendrogram. For hierarchical clustering with complete linkage, the sizes of each cluster become different and points with negative silhouette coeffcients occur after scaling. The average silhouette width decreases away from 1, which indicates that the quality of the clustering seems to be worse in terms of the silhouette measure. For hierarchical clustering with single linkaage, the clustering also becomes worse. There are more points with negative silhouette coefficients and the average silhouette width decreases. For k-means clustering, two points wiht negative sihouette coefficients occur and the average silhouette width decreases. The quality of clustering seem to be worse. However, for the USArrests data, three variables, Murder, Assault and Rape, are reported on a per capita basis. The fourth variable UrbanPop is reported on a percentage basis, rather than in absolute numbers. Since the variables are reporting in different units, scaling is recommended. If we don't scale the varaibles, the number of assualts is likely to domain the clustering.