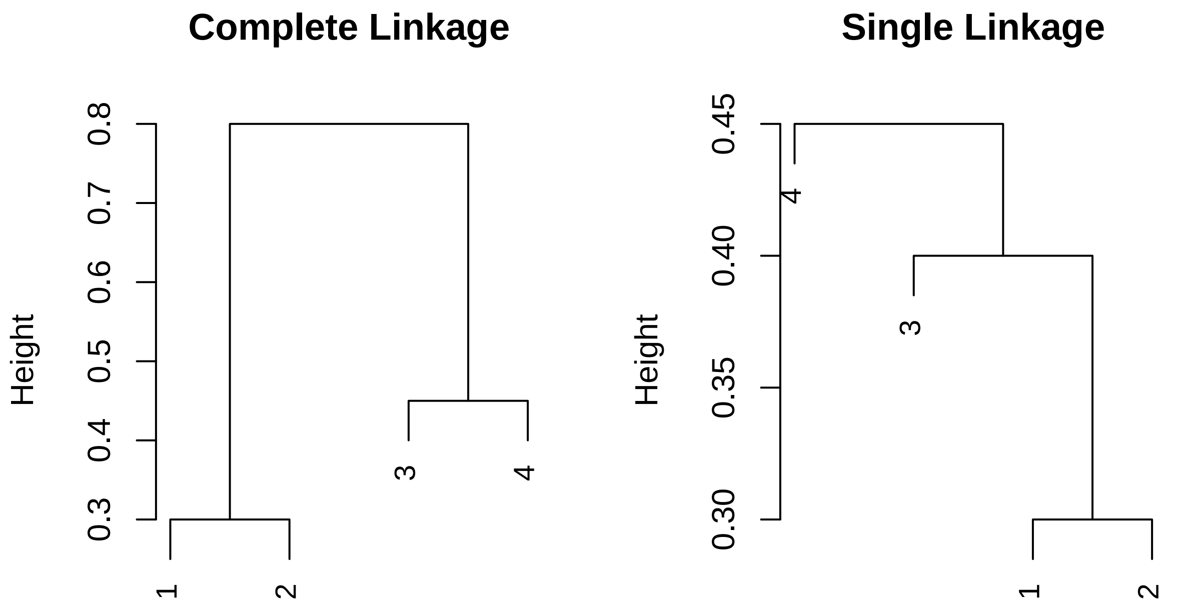
STATS 415 hw11 solution

Weijing Tang

April 7, 2018

Question 1

(a) & (b)



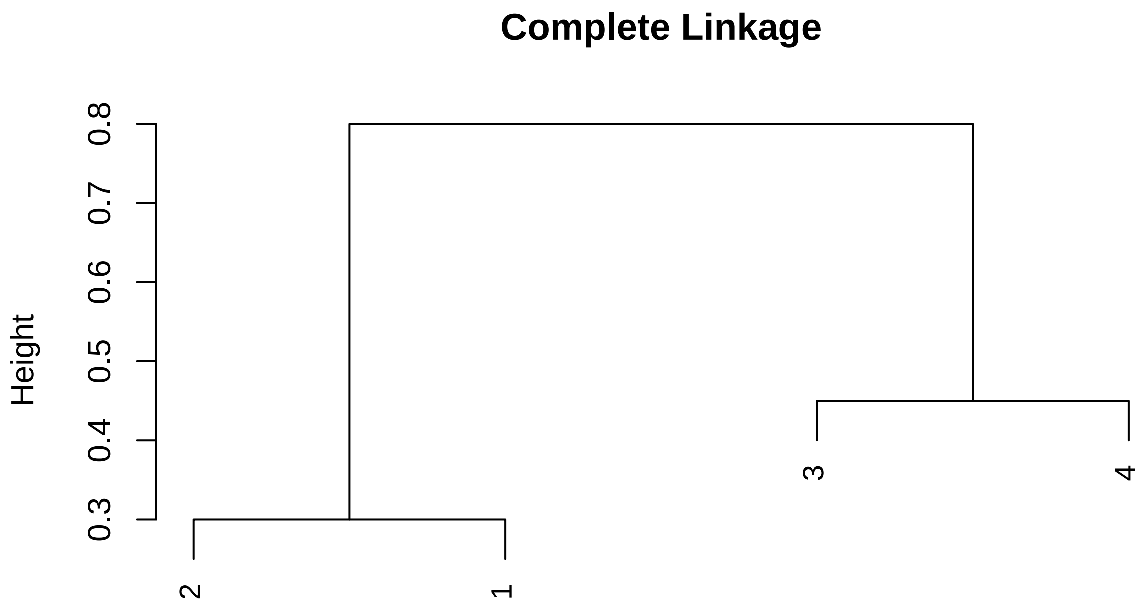
(c)

Cluster1: 1,2; Cluster2:3,4.

(d)

Cluster1:4; Cluster2:2,3,4.

(e)



Question 2

(a)

We load the package ISLR so to access the dataset USArrests and then use the `hclust()` function to perform hierarchical clustering.

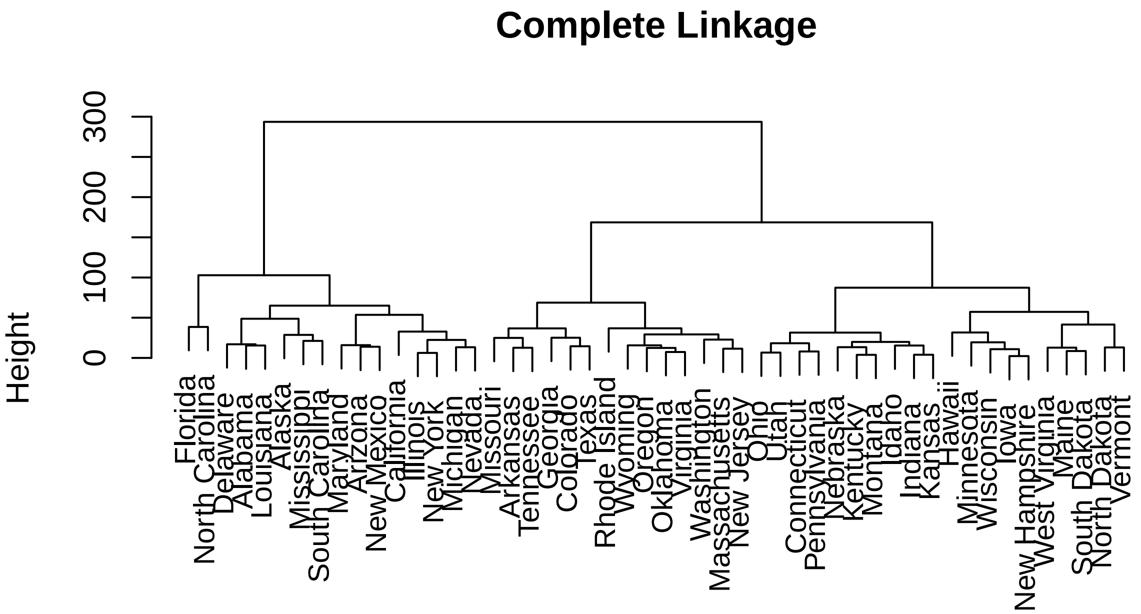
```
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 3.3.3
```

```
library(cluster)
```

```
## Warning: package 'cluster' was built under R version 3.3.3
```

```
data("USArrests")
hc.complete=hclust(dist(USArrests),method='complete')
plot(hc.complete,main="Complete Linkage", xlab="", sub="", cex=.9)
```



(b)

```
cutree(hc.complete,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
```

```
names(which(cutree(hc.complete,1)==1))
```

```
##  [1] "Alabama"        "Alaska"         "Arizona"        "Arkansas"
##  [5] "California"     "Colorado"       "Connecticut"    "Delaware"
##  [9] "Florida"        "Georgia"        "Hawaii"         "Idaho"
## [13] "Illinois"       "Indiana"        "Iowa"           "Kansas"
## [17] "Kentucky"       "Louisiana"      "Maine"          "Maryland"
## [21] "Massachusetts"  "Michigan"       "Minnesota"      "Mississippi"
## [25] "Missouri"       "Montana"        "Nebraska"       "Nevada"
## [29] "New Hampshire"  "New Jersey"     "New Mexico"     "New York"
## [33] "North Carolina" "North Dakota"   "Ohio"           "Oklahoma"
## [37] "Oregon"         "Pennsylvania"   "Rhode Island"   "South Carolina"
## [41] "South Dakota"   "Tennessee"      "Texas"          "Utah"
## [45] "Vermont"        "Virginia"       "Washington"     "West Virginia"
## [49] "Wisconsin"      "Wyoming"
```

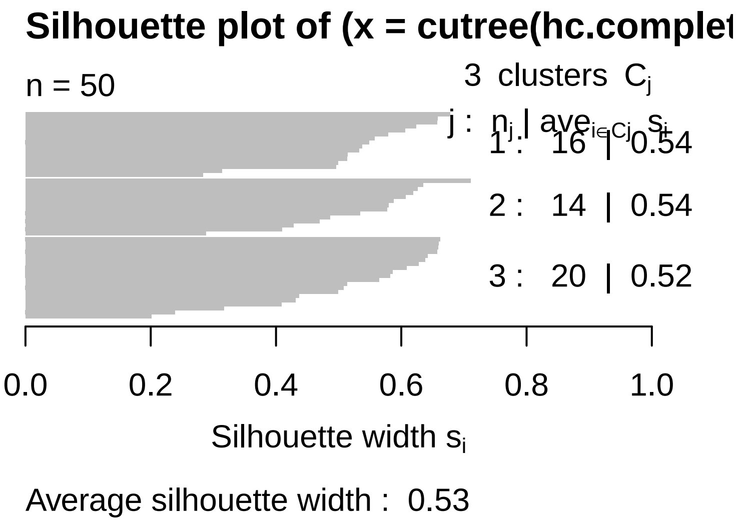
```
names(which(cutree(hc.complete,2)==2))
```

```
##  [1] "Arkansas"      "Colorado"      "Connecticut"   "Georgia"
##  [5] "Hawaii"        "Idaho"         "Indiana"       "Iowa"
##  [9] "Kansas"        "Kentucky"      "Maine"         "Massachusetts"
## [13] "Minnesota"     "Missouri"      "Montana"       "Nebraska"
## [17] "New Hampshire" "New Jersey"    "North Dakota"  "Ohio"
## [21] "Oklahoma"      "Oregon"        "Pennsylvania"  "Rhode Island"
## [25] "South Dakota"  "Tennessee"     "Texas"         "Utah"
## [29] "Vermont"       "Virginia"      "Washington"    "West Virginia"
## [33] "Wisconsin"     "Wyoming"
```

```
names(which(cutree(hc.complete,3)==3))
```

```
##  [1] "Connecticut"   "Hawaii"        "Idaho"         "Indiana"
##  [5] "Iowa"          "Kansas"        "Kentucky"      "Maine"
##  [9] "Minnesota"     "Montana"       "Nebraska"      "New Hampshire"
## [13] "North Dakota"  "Ohio"          "Pennsylvania"  "South Dakota"
## [17] "Utah"          "Vermont"       "West Virginia" "Wisconsin"
```

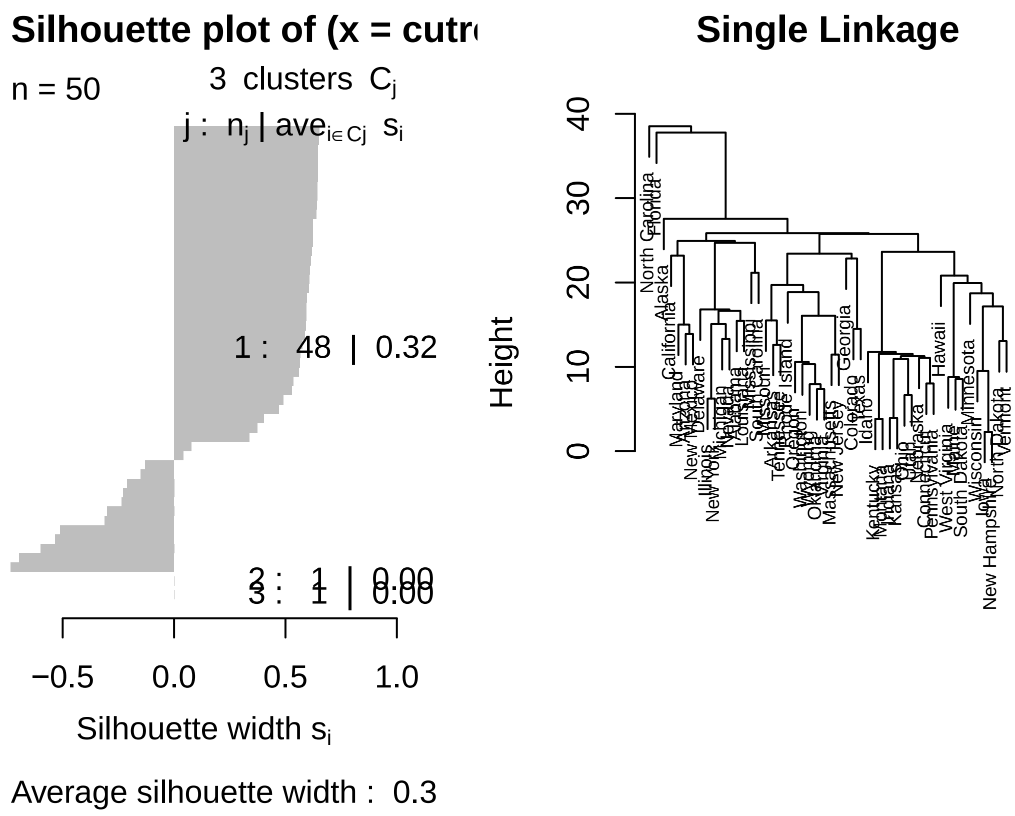
```
plot(silhouette(cutree(hc.complete,3),dist(USArrests)))
```



The average silhouette width is 0.53. There is no observation with nagetive silhouette score which indicates a good clustering result.

(c)

```r
hc.single=hclust(dist(USArrests),method='single')
par(mfrow=c(1,2))
plot(silhouette(cutree(hc.single,3),dist(USArrests)))
plot(hc.single,main="Single Linkage", xlab="", sub="", cex=.6)
```



```r
names(which(cutree(hc.single,1)==1))
```

```
##  [1] "Alabama"        "Alaska"         "Arizona"        "Arkansas"
##  [5] "California"     "Colorado"       "Connecticut"    "Delaware"
##  [9] "Florida"        "Georgia"        "Hawaii"         "Idaho"
## [13] "Illinois"       "Indiana"        "Iowa"           "Kansas"
## [17] "Kentucky"       "Louisiana"      "Maine"          "Maryland"
## [21] "Massachusetts"  "Michigan"       "Minnesota"      "Mississippi"
## [25] "Missouri"       "Montana"        "Nebraska"       "Nevada"
## [29] "New Hampshire"  "New Jersey"     "New Mexico"     "New York"
## [33] "North Carolina" "North Dakota"   "Ohio"           "Oklahoma"
## [37] "Oregon"         "Pennsylvania"   "Rhode Island"   "South Carolina"
## [41] "South Dakota"   "Tennessee"      "Texas"          "Utah"
## [45] "Vermont"        "Virginia"       "Washington"     "West Virginia"
## [49] "Wisconsin"      "Wyoming"
```

```r
names(which(cutree(hc.single,2)==2))
```

```
## [1] "North Carolina"
```

```r
names(which(cutree(hc.single,3)==3))
```

```
## [1] "North Carolina"
```

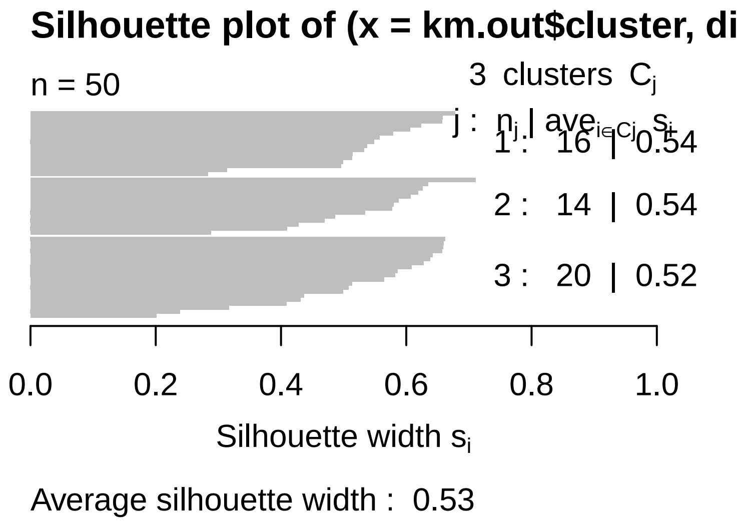
There are 48 states are clustered in one group. Compared with complete linkage, single linkage tends to yield trailing clusters. The average silhouette width is 0.3 which is much poorer than complete linkage.

(d)

```r
set.seed(1111)
km.out=kmeans(USArrests,3,nstart=20)
km.out$cluster
```

```
##        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
```

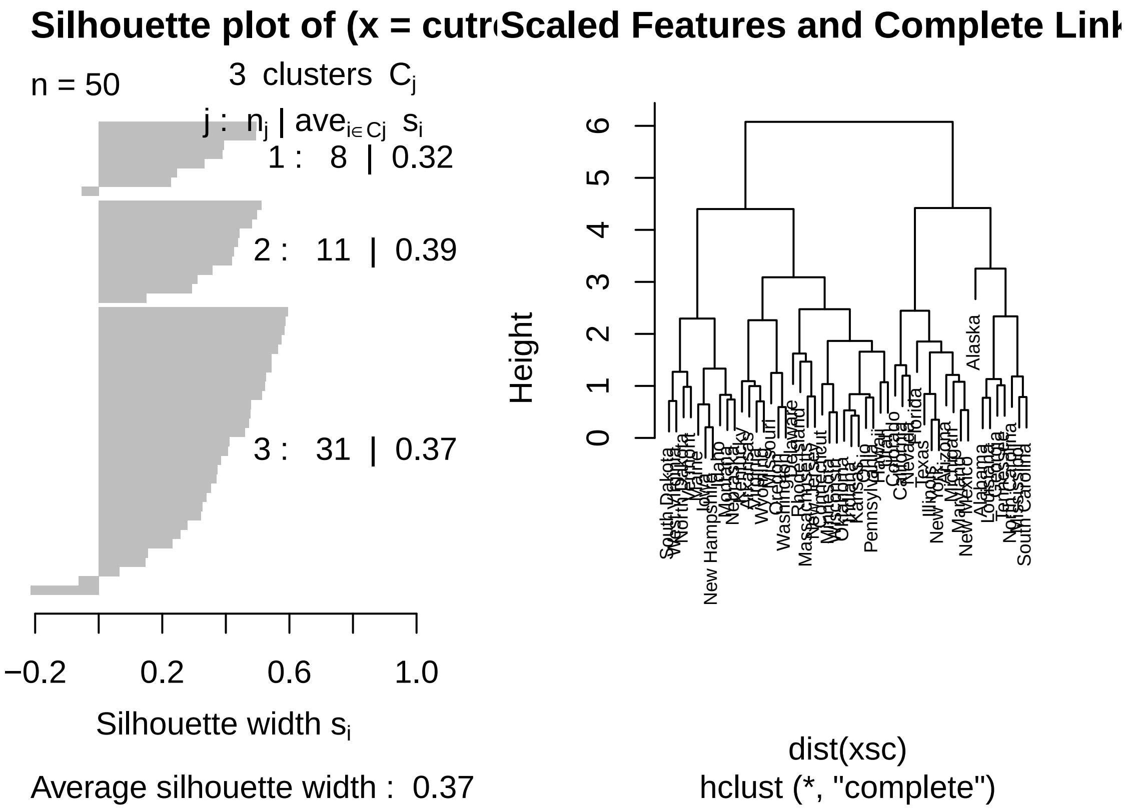
```r
par(mfrow=c(1,1))
plot(silhouette(km.out$cluster,dist(USArrests)))
```



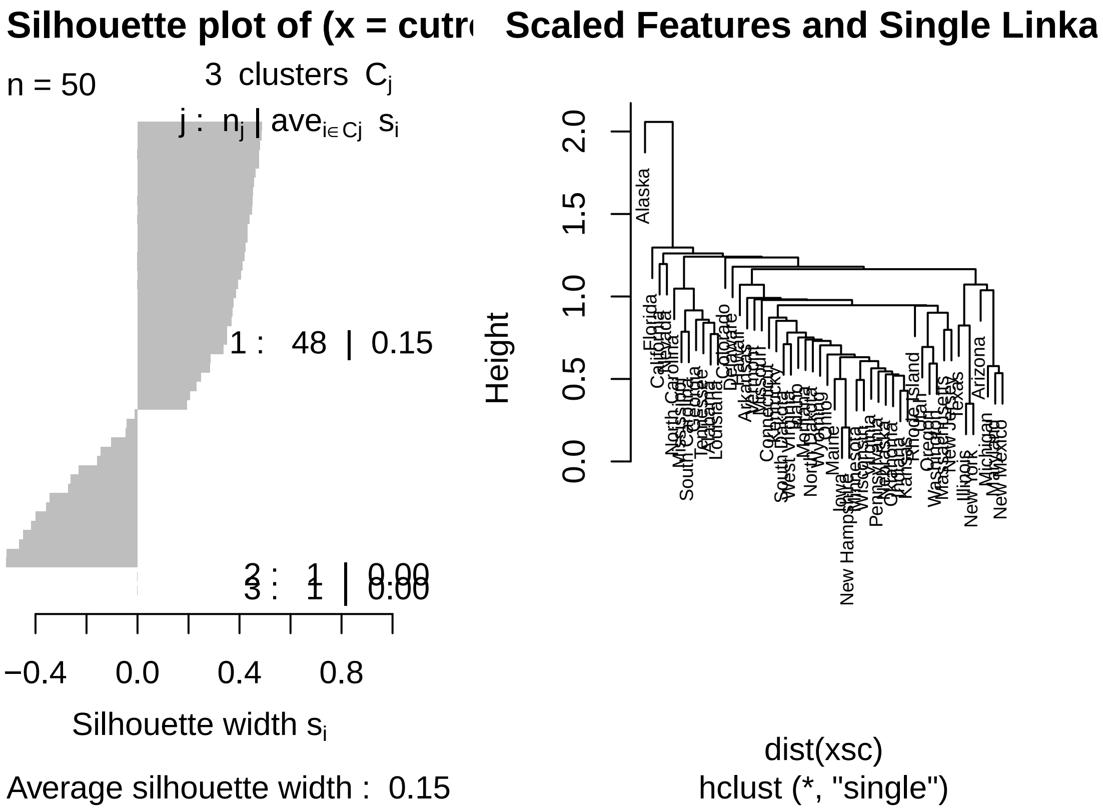
We set 20 random initial cluster assignments and report the best result. Kmeans gives us a similar clustering result as hierarchical clustering with complete linkage.

(e)

```
# hierarchical clustering
xsc=scale(USArrests)
hc.complete.sd=hclust(dist(xsc),method="complete")
hc.single.sd=hclust(dist(xsc),method='single')
par(mfrow=c(1,2))
plot(silhouette(cutree(hc.complete.sd,3),dist(xsc)))
plot(hc.complete.sd,main="Scaled Features and Complete Linkage",cex = 0.6)
```



```r
plot(silhouette(cutree(hc.single.sd,3),dist(xsc)))
plot(hc.single.sd,main="Scaled Features and Single Linkage", cex=.6)
```



```
cutree(hc.complete.sd,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
```

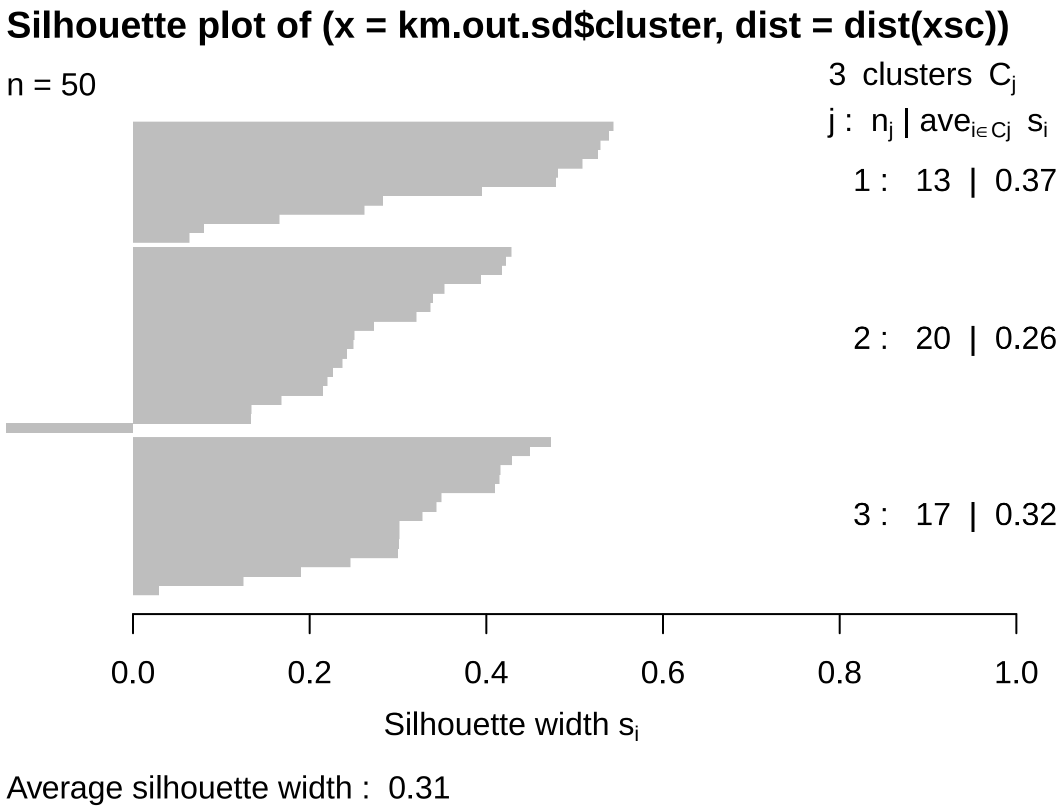
```
cutree(hc.single.sd,3)
```

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

```
# kmeans
set.seed(1111)
km.out.sd=kmeans(xsc,3,nstart=20)
km.out.sd$cluster
```

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

```
par(mfrow=c(1,1))
plot(silhouette(km.out.sd$cluster,dist(xsc)))
```



The average silhouette width of scaled data are smaller in all three cases than those of unscaled data.

(f)

We use the `table()` function to compare the results

```r
# compare HC
hc = cutree(hc.complete,3)
hc.sd = cutree(hc.complete.sd,3)
table(hc,hc.sd)
```

```
##     hc.sd
## hc    1  2  3
##    1  6  9  1
##    2  2  2 10
##    3  0  0 20
```

```r
km = km.out$cluster
km.sd = km.out.sd$cluster
table(km,km.sd)
```

```
##     km.sd
## km    1  2  3
##    1  0 15  1
##    2  0  5  9
##    3 13  0  7
```

We see that the three clusters obtained using original variables and scaled variables are somewhat different. For hierarchical clustering, cluster 3 obtained using scaled variables mainly combines the cluster 2 and 3 obtained using original variables. Regardless of permutation of cluster number, the difference between k-means clustering results is smaller than hierarchical clustering.

Note that the variables “Murder”, “Assault” and “Rape” are the number of corresponding arrests per 100,000 while the variable “UrbanPop” is the percentage of urban population. These variables have different units but all represent a ratio. The range of them don’t differ too much. The results from unscaled data also give better clustering result. Therefore, we prefer not to scaling data.