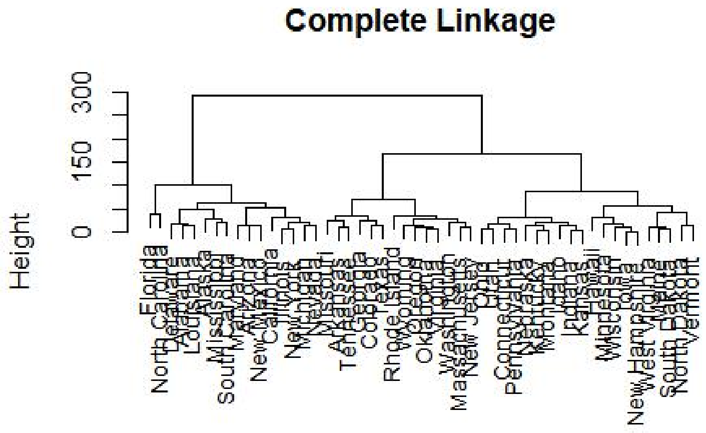
Stat415homework11

2. Consider the USArrests data from the textbook.

(a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states. Plot the dendrogram.

```
data1 = USArrests
summary(data1)

hc.complete = hclust(dist(data1), method="complete")
plot(hc.complete,main="Complete Linkage", xlab="", sub="", cex=.9)
```



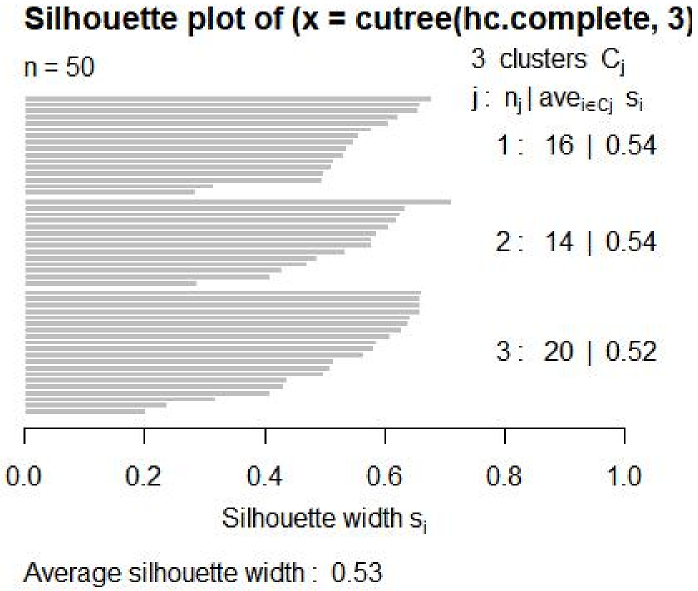
Comment: clustering process and dendrogram have been shown above.

(b) Cut the dendrogram at a height that results in three distinct clusters. Report the states belonging to each of the three clusters. Make a silhouette coefficient plot and comment on any interesting features.

```
# three distinct clusters
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

# silhouette coefficient
library(cluster)

sil.complete = silhouette(cutree(hc.complete,3),dist = dist(data1))
```

Comment: The states and their corresponding labels have been shown above.

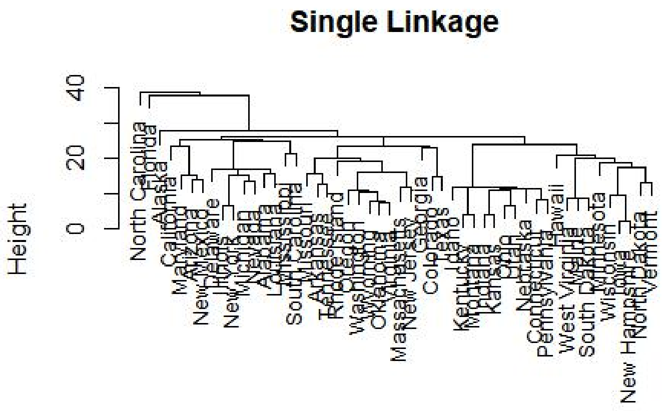
```
plot(sil.complete)
```



Comment: According to this plot, most of the silhouette coefficients are between 0.4 and 0.6 and almost all the coefficients are above 0.2. The average silhouette width is 0.53. The clustering is more effective if the coefficient is closer to 1. The plot indicates that although the result is not so perfect, it is still acceptable. Since there are no large negative values, there are no poor clustering points.

(c) Repeat questions (a) and (b) using single linkage instead.

```
hc.single = hclust(dist(data1), method="single")
plot(hc.single, main="Single Linkage", xlab="", sub="", cex=.9)
```



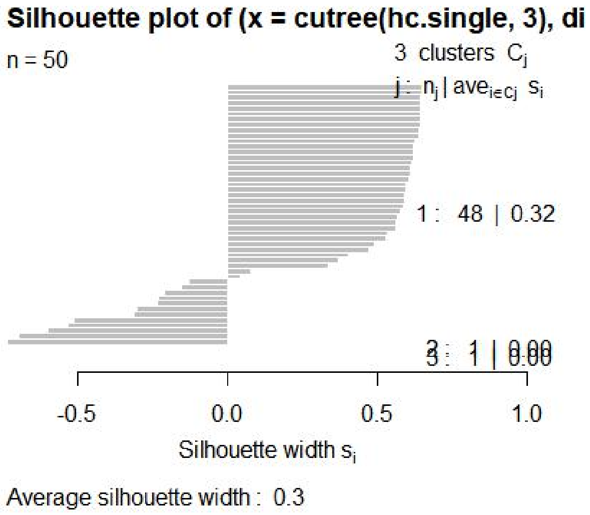
```r
cutree(hc.single, 3)
```

```
##  Alabama         Alaska        Arizona       Arkansas     California
##       1              1              1              1              1
## Colorado    Connecticut       Delaware        Florida        Georgia
##       1              1              1              2              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              3              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
```

```r
sil.single = silhouette(cutree(hc.single,3),dist = dist(data1))
```

Comment: The states and their corresponding labels have been shown above.

```r
plot(sil.single)
```



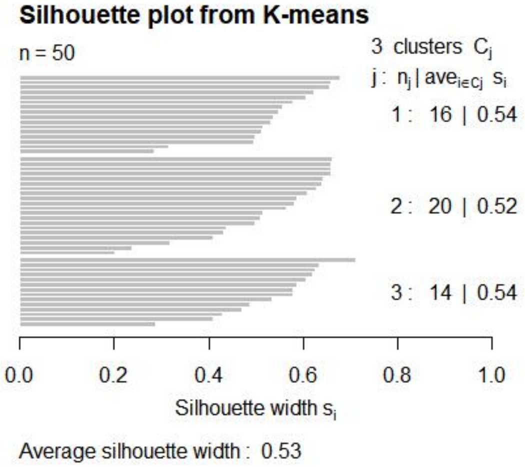
Comment: According to this plot, the silhouette coefficient is 0.32 for the first group and 0 for the other two groups. The average silhouette width is 0.3. Also, there are many large negative values, which indicates that many poor clustering points exist. The plot indicates that single linkage result is quite bad for this data set. Compared with complete linkage, single linkage is not suitable here.

(d) Perform K-means clustering on the data with K = 3 and report which states belong to which clusters. Report how you initialized the algorithm. Make a slihouette coefficient plot and comment on any interesting features.

```
set.seed(45678)
# multiple initial cluster assignments
value =rep(0,20)
set.seed(45678)
for (i in 1:20){
  km.out=kmeans(data1,3,nstart=i)
  value[i] = km.out$tot.withinss
}
value

##  [1] 47964.27 47964.27 47964.27 47964.27 47964.27 47964.27 47964.27
##  [8] 47964.27 47964.27 47964.27 47964.27 47964.27 47964.27 47964.27
## [15] 47964.27 47964.27 47964.27 47964.27 47964.27 47964.27

# K-mean clustering
km.out=kmeans(data1,3,nstart=20)
km.out

## K-means clustering with 3 clusters of sizes 14, 16, 20
## Cluster means:
##      Murder  Assault UrbanPop     Rape
## 1  8.214286 173.2857 70.64286 22.84286
## 2 11.812500 272.5625 68.31250 28.37500
## 3  4.270000  87.5500 59.75000 14.39000
## Clustering vector:
## Alabama         Alaska        Arizona       Arkansas     California
##      2              2              2              1              2
##Colorado    Connecticut       Delaware        Florida        Georgia
##      1              3              2              2              1
## Hawaii          Idaho       Illinois        Indiana           Iowa
##      3              3              2              3              3
## Kansas       Kentucky      Louisiana          Maine       Maryland
##      3              3              2              3              2
##  Massachusetts Michigan    Minnesota    Mississippi       Missouri
##      1              2              3              2              1
## Montana       Nebraska         Nevada  New Hampshire     New Jersey
##      3              3              2              3              1
## New Mexico    New York North Carolina   North Dakota           Ohio
##      2              2              2              3              3
## Oklahoma      Oregon   Pennsylvania   Rhode Island South Carolina
##      1              1              3              1              2
## South Dakota    Tennessee     Texas           Utah        Vermont
##      3              1              1              3              3
## Virginia     Washington  West Virginia      Wisconsin     Wyoming
##     1              1              3              3              1
```

Comment: we decide to use nstart=20 and the result indicates that the sum of squares is equal. The states and their corresponding labels have been shown above.

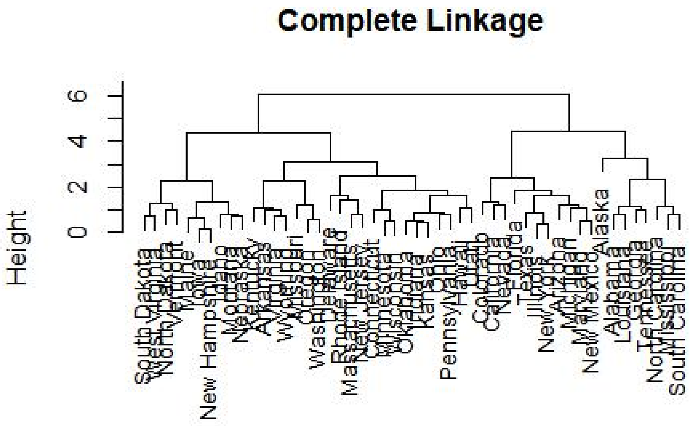
```
# use the solution from some hierarchical algorithm as initial value
cutree(hc.complete, 3)

hcmean1 = colMeans(data1[which(cutree(hc.complete, 3)==1),])
hcmean2 = colMeans(data1[which(cutree(hc.complete, 3)==2),])
hcmean3 = colMeans(data1[which(cutree(hc.complete, 3)==3),])
hcmean= rbind(hcmean1,hcmean2,hcmean3)
hcmean

##            Murder  Assault UrbanPop     Rape
## hcmean1 11.812500 272.5625 68.31250 28.37500
## hcmean2  8.214286 173.2857 70.64286 22.84286
## hcmean3  4.270000  87.5500 59.75000 14.39000

km.hc = kmeans(data1,centers = hcmean)
km.hc$iter

## [1] 1

# coefficient plot
set.seed(45678)
km.out=kmeans(data1, 3, nstart=20)
km.clusters = km.out$cluster
data.dist=dist(data1)
plot(silhouette(km.clusters,data.dist),main="Silhouette plot from K-mea
ns")
```



Comment: the algorithm stops after only one iteration. According to this plot, most of the silhouette coefficients are between 0.4 and 0.6 and almost all the coefficients are above 0.2. The average silhouette width is 0.53. The clustering is more effective if the coefficient is closer to 1. The plot indicates that although the result is not so perfect, it is still acceptable. Since there are no large negative values, there are no poor clustering points.

(e) Scale all the variables to have mean 0 and standard deviation 1. Repeat questions (a)-(d) using the scaled data.

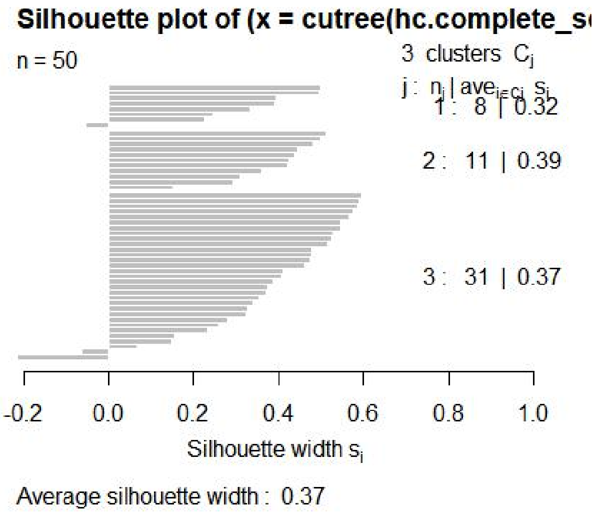
```r
# scale data
sd.data=scale(data1)
# a)
hc.complete_sd = hclust(dist(sd.data), method="complete")
plot(hc.complete_sd,main="Complete Linkage", xlab="", sub="", cex=.9)
```



```
# b)
# three distinct clusters
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
```

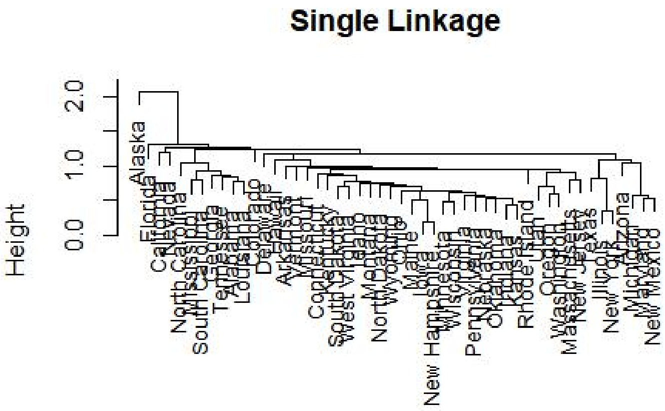
```
# silhouette coefficient
sil.complete_sd = silhouette(cutree(hc.complete_sd,3),dist = dist(sd.da
ta))
plot(sil.complete_sd)
```



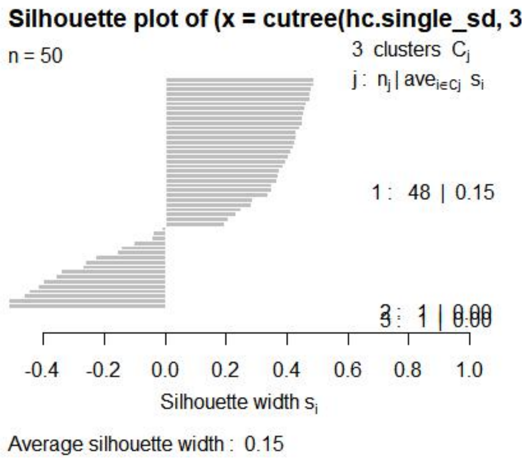
Comment: According to this plot, most of the silhouette coefficients are between 0.2 and 0.6 and almost all the coefficients are above 0.2. The average silhouette width is

0.37. The plot indicates that the result is not good since the silhouette coefficient is quite small. Also, there exists several poor clustering points.

```
# c)
hc.single_sd = hclust(dist(sd.data), method="single")
plot(hc.single_sd, main="Single Linkage", xlab="", sub="", cex=.9)
```

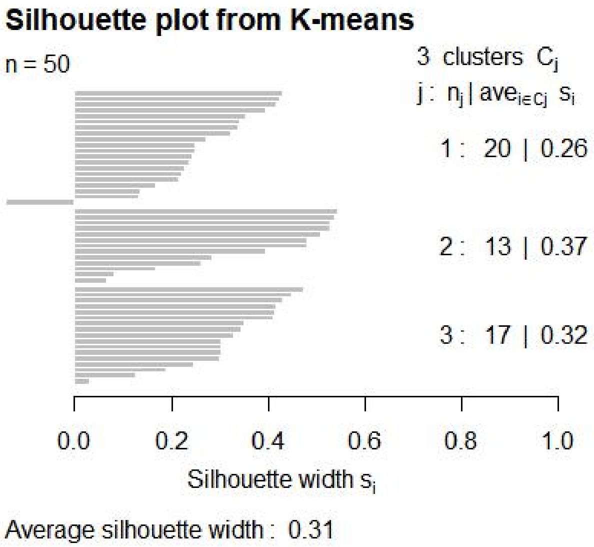


```
cutree(hc.single_sd, 3)

sil.single_sd = silhouette(cutree(hc.single_sd,3),dist = dist(sd.data))

plot(sil.single_sd)
```



Comment: According to this plot, the silhouette coefficient is 0.15 for the first group and 0 for the other two groups. The average silhouette width is 0.15, which is the worst in these operations. Also, there are many large negative values, which indicates that many poor clustering points exist. The plot indicates that single linkage result is quite bad for this data set.

```
# d)
set.seed(45678)
# multiple initial cluster assignments
value =rep(0,20)
set.seed(45678)
for (i in 1:20){
  km.out_sd=kmeans(sd.data,3,nstart=i)
  value[i] = km.out_sd$tot.withinss
}
value

##  [1] 85.11049 81.59329 81.59329 78.32327 78.32327 78.32327 78.32327
##  [8] 78.32327 78.32327 78.32327 78.32327 78.32327 78.32327 78.32327
## [15] 78.32327 78.32327 78.32327 78.32327 78.32327 78.32327

# K-mean clustering
km.out_sd=kmeans(sd.data,3,nstart=20)
km.out_sd

## K-means clustering with 3 clusters of sizes 13, 20, 17
##
## Cluster means:
##       Murder    Assault   UrbanPop       Rape
## 1 -0.9615407 -1.1066010 -0.9301069 -0.9667633
## 2  1.0049340  1.0138274  0.1975853  0.8469650
## 3 -0.4469795 -0.3465138  0.4788049 -0.2571398
## Within cluster sum of squares by cluster:
## [1] 11.95246 46.74796 19.62285
##  (between_SS / total_SS =  60.0 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"

# use the solution from some hierarchical algorithm as initial value
cutree(hc.complete_sd, 3)

hcmean1_sd = colMeans(data1[which(cutree(hc.complete_sd, 3)==1),])
hcmean2_sd = colMeans(data1[which(cutree(hc.complete_sd, 3)==2),])
hcmean3_sd = colMeans(data1[which(cutree(hc.complete_sd, 3)==3),])
hcmean_sd = rbind(hcmean1_sd,hcmean2_sd,hcmean3_sd)
hcmean_sd

##               Murder  Assault UrbanPop     Rape
## hcmean1_sd 14.087500 252.7500 53.50000 24.53750
## hcmean2_sd 11.054545 264.0909 79.09091 32.61818
## hcmean3_sd  5.003226 116.4839 63.83871 16.33871

# coefficient plot
set.seed(45678)
km.out_sd=kmeans(sd.data, 3, nstart=20)
km.clusters_sd = km.out_sd$cluster
data.dist_sd=dist(sd.data)
plot(silhouette(km.clusters_sd,data.dist_sd),main="Silhouette plot from
 K-means")
```



Comment: According to this plot, most of the silhouette coefficients are between 0.2 and 0.6 and almost all the coefficients are above 0.1. The average silhouette width is 0.31. The plot indicates that the result is worse than the result of un-scaled data.And there is also one large negative value.

(f) What effect does scaling the variables have on hierarchical clustering? On K-means clustering? In your opinion, should the variables be scaled before clustering in this example? Explain your reasoning. Comment: On both hierarchical clustering and K-means clustering, the effect of scaled data is worse than the effect of the un-scaled data. There are more poor clustering points and the silhouette coefficient is smaller.Thus the variables should not be scaled in this data set. Since the UrbanPop is a percent variable but not a numeric variable, then scaling the data with well-defined meaning will cause distortion. Thus scaling data will not improve the clustering effect.