

Unsupervised-HW

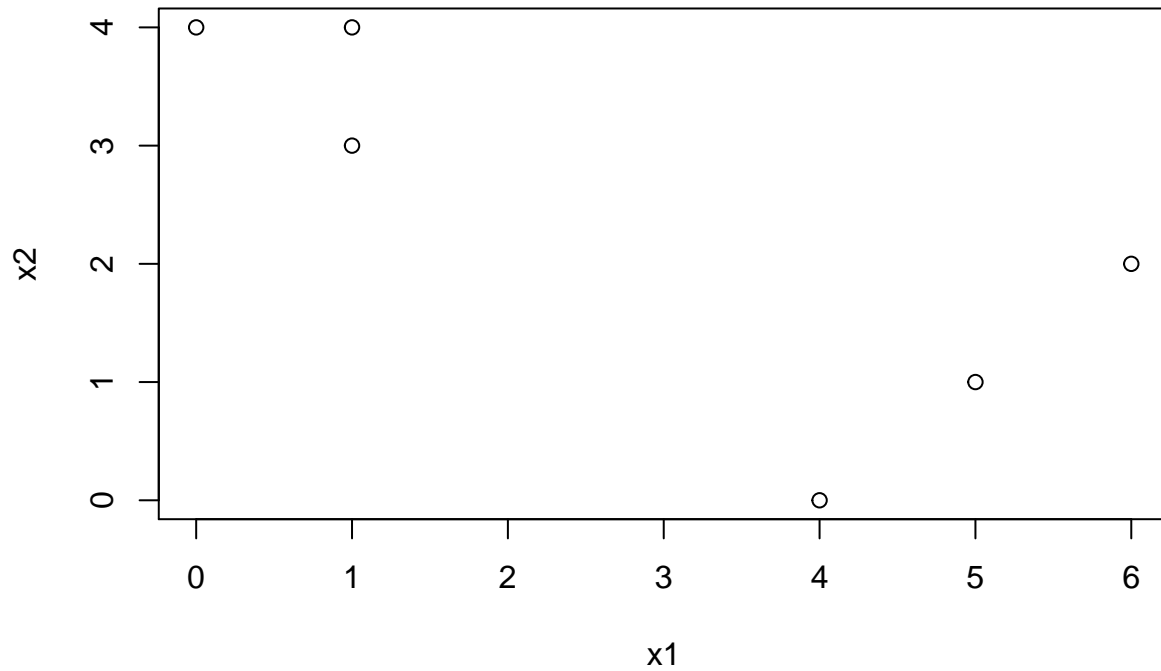
Ziyi Bai

2021/3/19

10.3

##(a)

```
x1 <- c(1,1,0,5,6,4)
x2 <- c(4,3,4,1,2,0)
df <- data.frame(x1,x2)
plot(df)
```

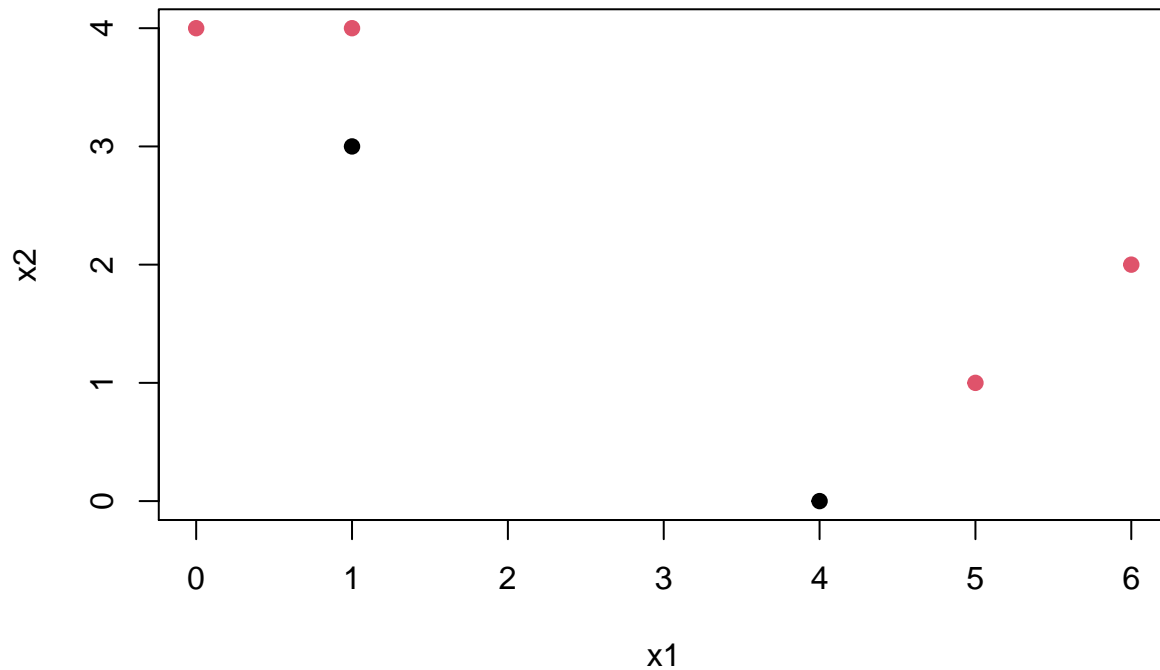


##(b)

```
set.seed(99)
in_clusters <- sample(2,nrow(df),replace=T)
in_clusters
```

```
## [1] 2 1 2 2 2 1
```

```
plot(x1,x2,col=in_clusters,pch=20,cex=1.5)
```



```
df1 <- data.frame(df,in_clusters)
```

```
##(c)
```

```
cent_1 <- c(mean(df1[df1$in_clusters==1,1]),mean(df1[df1$in_clusters==1,2]))
```

```
cent_2 <- c(mean(df1[df1$in_clusters==2,1]),mean(df1[df1$in_clusters==2,2]))
```

```
print(paste0("Centroid for cluster 1 is: ", "(", cent_1[1], ",", cent_1[2], ")"))
```

```
## [1] "Centroid for cluster 1 is: (2.5,1.5)"
```

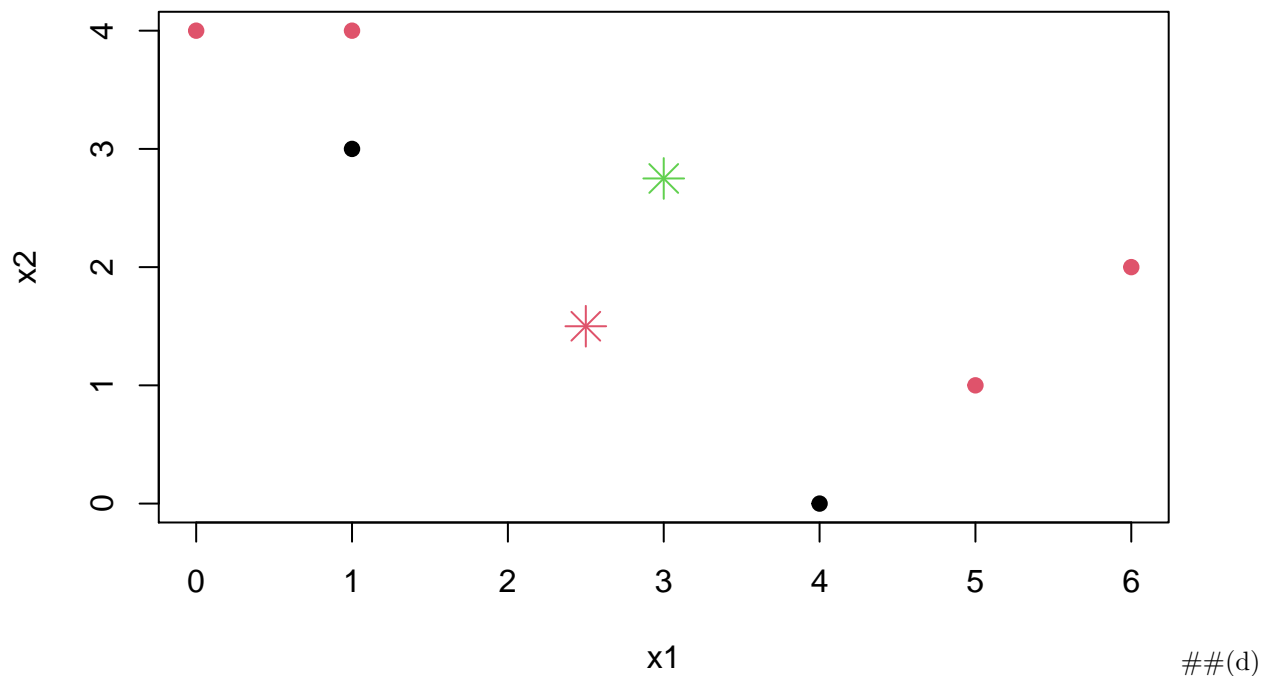
```
print(paste0("Centroid for cluster 2 is: ", "(", cent_2[1], ",", cent_2[2], ")"))
```

```
## [1] "Centroid for cluster 2 is: (3,2.75)"
```

```
plot(x1, x2, col = in_clusters, pch = 20, cex = 1.5)
```

```
points(cent_1[1], cent_1[2], pch = 8, cex = 2, col = 2)
```

```
points(cent_2[1], cent_2[2], pch = 8, cex = 2, col = 3)
```



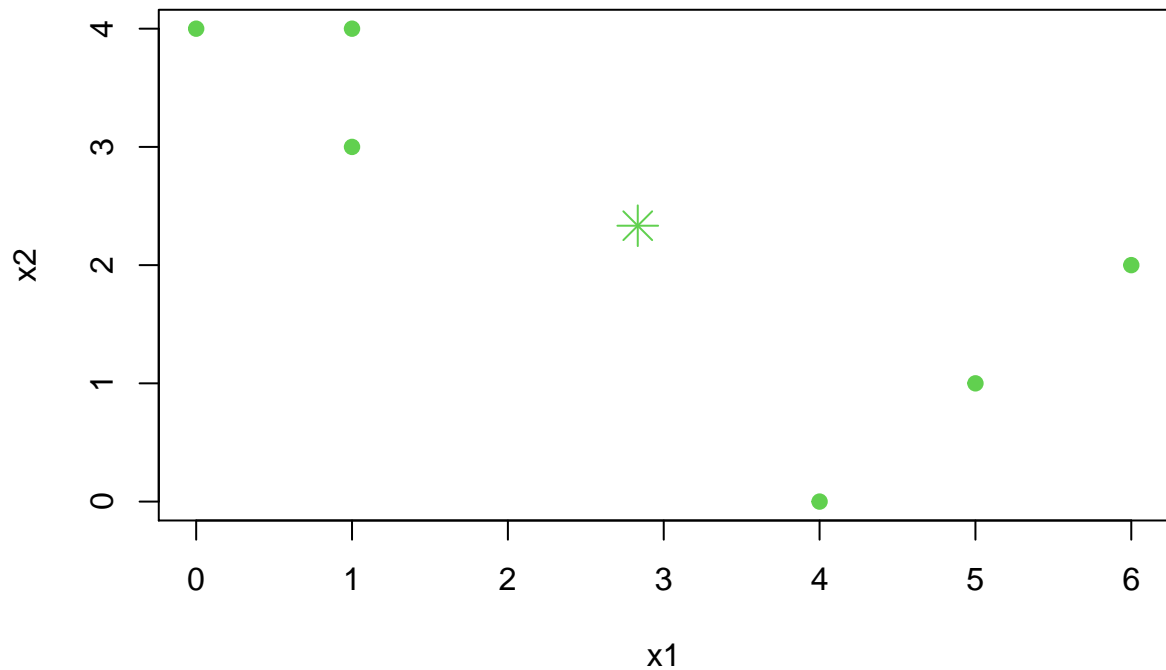
```
euc_dist <- function(v,z){
  sqrt(sum(v-z)^2)
}

df1$updated_cluster <- c()

for (i in 1:nrow(df1)) {
  d1 <- euc_dist(c(df1[i,1],df1[i,2]),c(cent_1[1],cent_1[2]))
  d2 <- euc_dist(c(df1[i,1],df1[i,2]),c(cent_2[1],cent_2[2]))

  if (d1<=d2){
    df1$update_cluster[i] <- 1
  }else{
    df1$updated_cluster[i] <- 2
  }
}

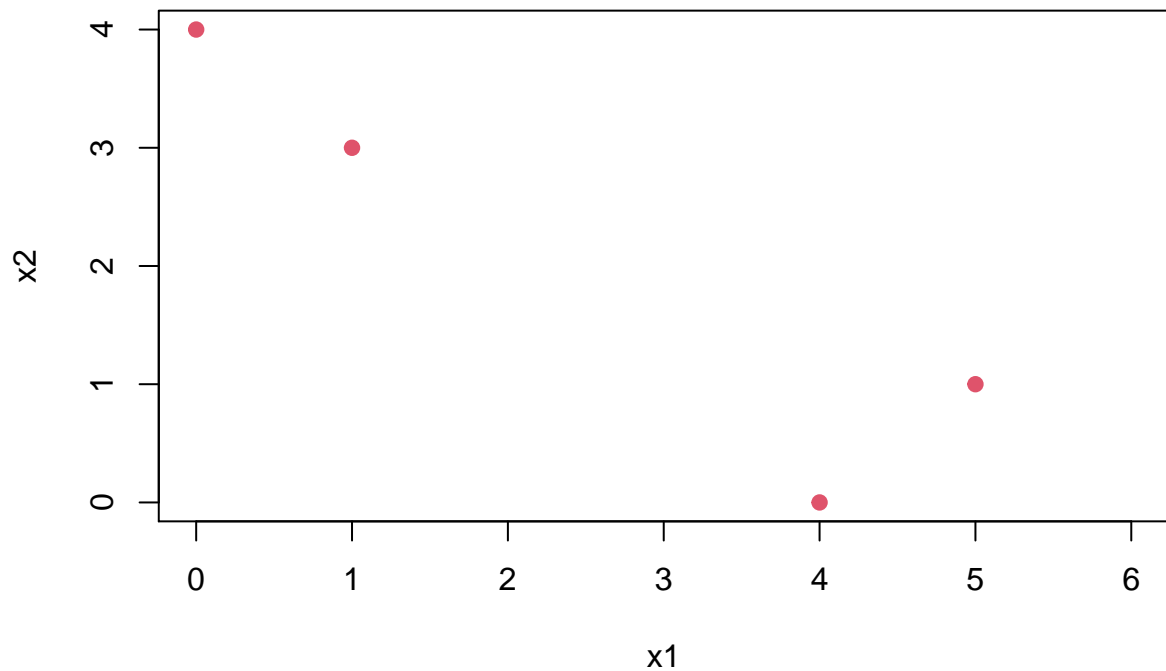
updated_centroid1 <- c(mean(df1[df1$updated_cluster == 1,1]), mean(df1[df1$updated_cluster == 1,2]))
updated_centroid2 <- c(mean(df1[df1$updated_cluster == 2,1]), mean(df1[df1$updated_cluster == 2,2]))
plot(x1, x2, col = df1$updated_cluster+1, pch = 20, cex = 1.5)
points(updated_centroid1[1], updated_centroid1[2], pch = 8, cex = 2, col = 2)
points(updated_centroid2[1], updated_centroid2[2], pch = 8, cex = 2, col = 3)
```



we don't have any changes.

##(f)

```
plot(x1,x2,col=df1$update_cluster+1,pch=20,cex=1.5)
```



10.5

Left: The left side plot shows unscaled variables. In this case, the number of socks becomes more important than the number of computers. With $K = 2$, the K-Means clustering will result in two clusters separately for socks and computer purchases. Center: Since the variables are scaled, in this case, the purchase of computers becomes as important as socks. Here, the K-Means clustering will produce two clusters - one of people who have purchased a computer and

the other of people who haven't. Right: In this case, K-Means clustering will produce clusters of socks purchases and computer purchases separately because there is a huge difference in the price of socks and computers.

10.6

##(a) 90% of the information in the original data is lost in projecting the tissue sample observations onto the first principle component. Or, 90% of the original data is not contain the first priciple component.

##(b) Since each patient sample was rin on either of the machine A and B, the machine used could be used as a feature in the PCA. We check if there is an improvment in the PVE after adding the machine used as a feature.

##(c)

```
set.seed(9)
control <- matrix(rnorm(50*1000),ncol=50)
treatment <- matrix(rnorm(50*1000),ncol=50)

x <- cbind(control,treatment)
x[1,] <- seq(-18,18 -.36,.36)
pca <- prcomp(scale(x))
summary(pca)$importance[,1]

##      Standard deviation Proportion of Variance Cumulative Proportion
##      3.159783                0.099840          0.099840

X <- rbind(x,c(rep(0,50),rep(10,50)))
pca_out <- prcomp(scale(X))
summary(pca_out)$importance[,1]

##      Standard deviation Proportion of Variance Cumulative Proportion
##      3.391937                0.115050          0.115050
```

The proportion of variance explained by the first principle component is 9.98%. Including the machine used as a feature, coding 0 for A and 10 for B, the PVE increased to 11.5%.

10.8

##(a)

```
# the sdev approach to PVE
data("USArrests")
pca_usa <- prcomp(USArrests,scale. = T)
pca_usa$sdev

## [1] 1.5748783 0.9948694 0.5971291 0.4164494

# variance
pca_var <- pca_usa$sdev^2
pca_var

## [1] 2.4802416 0.9897652 0.3565632 0.1734301

#PVE
pve <- pca_var/sum(pca_var)
pve

## [1] 0.62006039 0.24744129 0.08914080 0.04335752
```

##(b)

```
# the prcomp PVE approach
usa_scaled <- scale(USArrests)
loadings <- pca_usa$rotation
sum_var <- sum(apply(as.matrix(usa_scaled)^2,2,sum))
apply((as.matrix(usa_scaled)%*%loadings)^2,2,sum)/sum_var
```

```
##          PC1          PC2          PC3          PC4
## 0.62006039 0.24744129 0.08914080 0.04335752
```

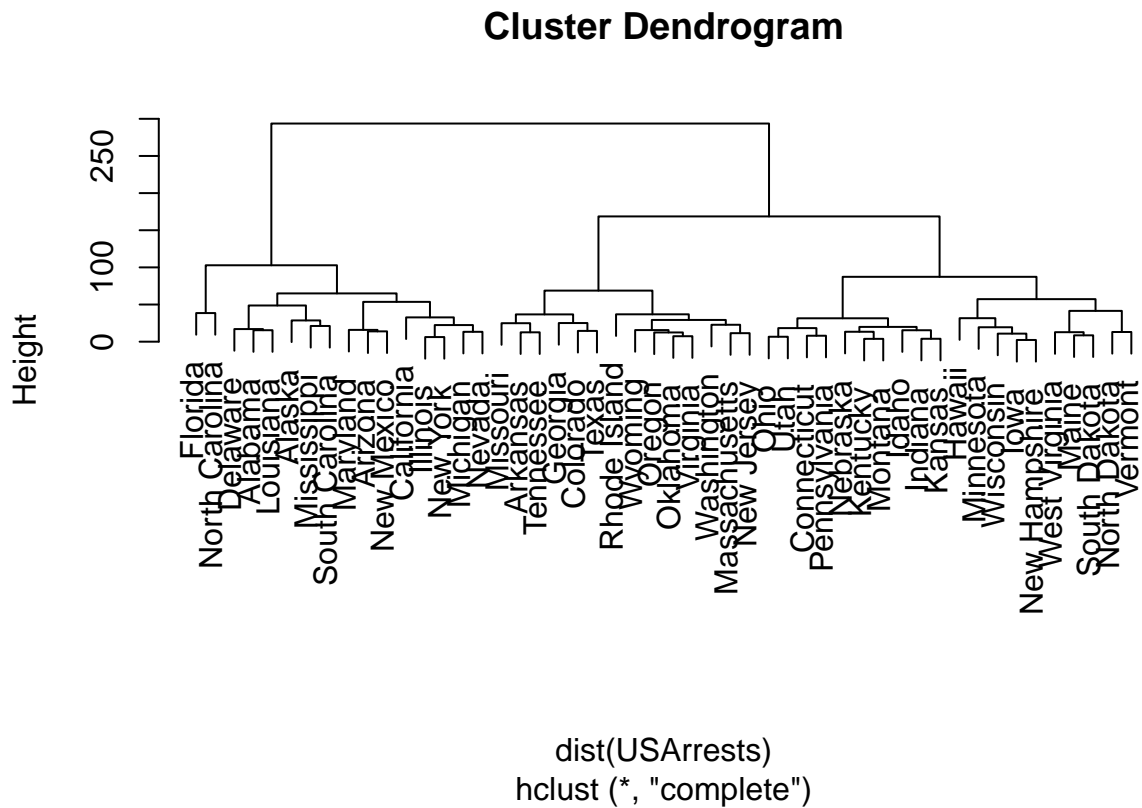
The PVE for each principal component from both the approaches is the same.

10.9

##(a)

```
set.seed(9)

hc.complete <- hclust(dist(USArrests),method = "complete")
plot(hc.complete)
```



##(b)

```
hc_cut <- cutree(hc.complete,3)
clusters <- split(data.frame(names(hc_cut),hc_cut),as.factor(hc_cut))
clusters
```

```
## $`1`
##          names.hc_cut. hc_cut
## Alabama          Alabama      1
## Alaska           Alaska       1
## Arizona          Arizona       1
```

## California	California	1
## Delaware	Delaware	1
## Florida	Florida	1
## Illinois	Illinois	1
## Louisiana	Louisiana	1
## Maryland	Maryland	1
## Michigan	Michigan	1
## Mississippi	Mississippi	1
## Nevada	Nevada	1
## New Mexico	New Mexico	1
## New York	New York	1
## North Carolina	North Carolina	1
## South Carolina	South Carolina	1

##

\$`2`

##	names.hc_cut.	hc_cut
## Arkansas	Arkansas	2
## Colorado	Colorado	2
## Georgia	Georgia	2
## Massachusetts	Massachusetts	2
## Missouri	Missouri	2
## New Jersey	New Jersey	2
## Oklahoma	Oklahoma	2
## Oregon	Oregon	2
## Rhode Island	Rhode Island	2
## Tennessee	Tennessee	2
## Texas	Texas	2
## Virginia	Virginia	2
## Washington	Washington	2
## Wyoming	Wyoming	2

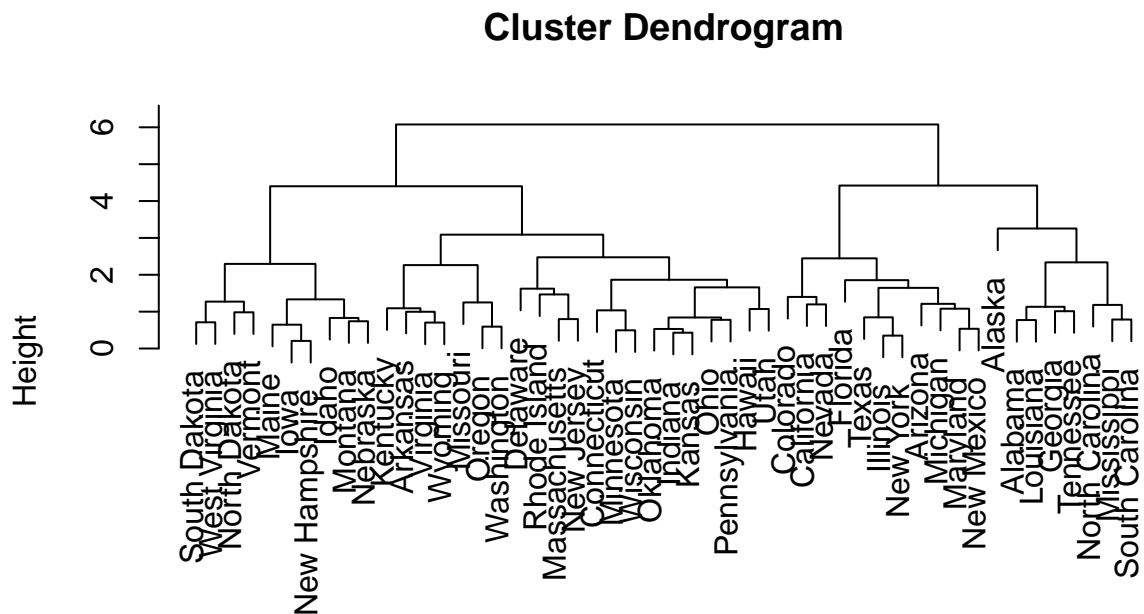
##

\$`3`

##	names.hc_cut.	hc_cut
## Connecticut	Connecticut	3
## Hawaii	Hawaii	3
## Idaho	Idaho	3
## Indiana	Indiana	3
## Iowa	Iowa	3
## Kansas	Kansas	3
## Kentucky	Kentucky	3
## Maine	Maine	3
## Minnesota	Minnesota	3
## Montana	Montana	3
## Nebraska	Nebraska	3
## New Hampshire	New Hampshire	3
## North Dakota	North Dakota	3
## Ohio	Ohio	3
## Pennsylvania	Pennsylvania	3
## South Dakota	South Dakota	3
## Utah	Utah	3
## Vermont	Vermont	3
## West Virginia	West Virginia	3
## Wisconsin	Wisconsin	3

```
##(c)
```

```
hc_scaled <- hclust(dist(scale(USArrests)),method = "complete")  
plot(hc_scaled)
```



```
dist(scale(USArrests))  
hclust (*, "complete")
```

```
##(d)
```

```
hc_scaled_cut <- cutree(hc_scaled,3)  
clusters_scaled <- split(data.frame(names(hc_scaled_cut),hc_scaled_cut),as.factor(hc_scaled_cut))  
clusters_scaled
```

```
## $`1`  
##          names.hc_scaled_cut. hc_scaled_cut  
## Alabama          Alabama          1  
## Alaska           Alaska           1  
## Georgia          Georgia          1  
## Louisiana        Louisiana        1  
## Mississippi      Mississippi      1  
## North Carolina   North Carolina   1  
## South Carolina   South Carolina   1  
## Tennessee        Tennessee        1  
##  
## $`2`  
##          names.hc_scaled_cut. hc_scaled_cut  
## Arizona          Arizona          2  
## California       California       2  
## Colorado         Colorado         2  
## Florida          Florida          2  
## Illinois         Illinois         2  
## Maryland         Maryland         2  
## Michigan         Michigan         2
```



```
## Nevada Nevada 2
## New Mexico New Mexico 2
## New York New York 2
## Texas Texas 2
##
## $`3`
## names.hc_scaled_cut. hc_scaled_cut
## Arkansas Arkansas 3
## Connecticut Connecticut 3
## Delaware Delaware 3
## Hawaii Hawaii 3
## Idaho Idaho 3
## Indiana Indiana 3
## Iowa Iowa 3
## Kansas Kansas 3
## Kentucky Kentucky 3
## Maine Maine 3
## Massachusetts Massachusetts 3
## Minnesota Minnesota 3
## Missouri Missouri 3
## Montana Montana 3
## Nebraska Nebraska 3
## New Hampshire New Hampshire 3
## New Jersey New Jersey 3
## North Dakota North Dakota 3
## Ohio Ohio 3
## Oklahoma Oklahoma 3
## Oregon Oregon 3
## Pennsylvania Pennsylvania 3
## Rhode Island Rhode Island 3
## South Dakota South Dakota 3
## Utah Utah 3
## Vermont Vermont 3
## Virginia Virginia 3
## Washington Washington 3
## West Virginia West Virginia 3
## Wisconsin Wisconsin 3
## Wyoming Wyoming 3
```

```
table(hc_cut, hc_scaled_cut)
```

```
## hc_scaled_cut
## hc_cut 1 2 3
## 1 6 9 1
## 2 2 2 10
## 3 0 0 20
```

Scaling the variable affect the clusters obtained. It is better to scale the variable because they are measured on different units.

10.10

```
##(a)
```

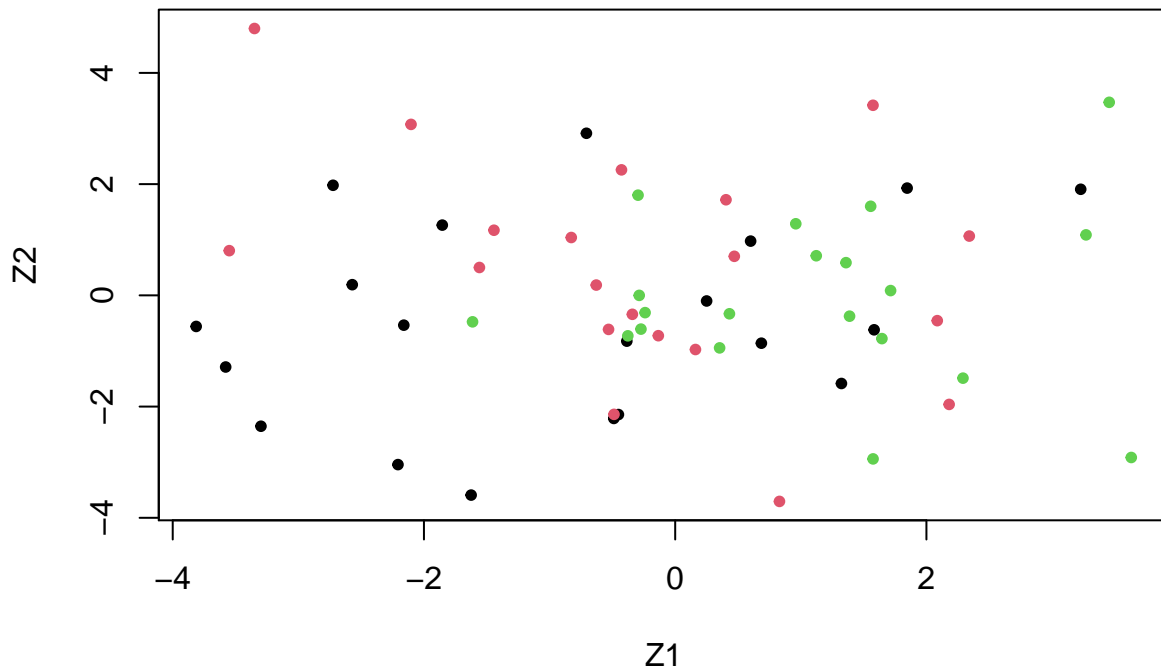
```
set.seed(9)
```

```
sim_data <- matrix(sapply(1:3, function(x) rnorm(20*50,mean = 0,sd=0.001)),ncol=50)
class <- unlist(lapply(1:3,FUN=function(x) rep(x,20)))

sim_data <- data.frame(sim_data)
sim_data$true_labels <- c(rep(1,20),rep(2,20),rep(3,20))
```

##(b)

```
sim_pca <- prcomp(sim_data,scale. = T,center = T)
plot(sim_pca$x[,1:2],col=class,xlab = "Z1",ylab="Z2",pch=20)
```



##(c)

```
set.seed(9)

sim_kmeans <- kmeans(sim_data,3)
table(sim_data$true_labels,sim_kmeans$cluster)
```

```
##
##      1  2  3
##  1  0  0 20
##  2  0  0 20
##  3 10 10  0
```

Cluster the observations correctly.

##(d)

```
set.seed(9)

sim_kmeans2 <- kmeans(sim_data,2)
table(sim_data$true_labels,sim_kmeans2$cluster)
```

```
##
##      1  2
##  1  0 20
##  2  0 20
```

```
##    3 20  0
```

```
##(e)
```

```
set.seed(9)
```

```
sims_kmeans4 <- kmeans(sim_data,4)
table(sim_data$true_labels,sims_kmeans4$cluster)
```

```
##
```

```
##      1  2  3  4
```

```
##    1  0  0 20  0
```

```
##    2  0  0  0 20
```

```
##    3 10 10  0  0
```

This one doesn't perform as well as the above two.

```
##(f)
```

```
set.seed(9)
```

```
km_out <- kmeans(sim_pca$x[,1:2],3)
table(sim_data$true_labels,km_out$cluster)
```

```
##
```

```
##      1  2  3
```

```
##    1  4  4 12
```

```
##    2  7  4  9
```

```
##    3  7  8  5
```

many miss classified observations.

```
##(g)
```

```
set.seed(9)
```

```
km_out_1 <- kmeans(scale(sim_data),3)
table(sim_data$true_labels,km_out_1$cluster)
```

```
##
```

```
##      1  2  3
```

```
##    1  8  4  8
```

```
##    2 10  3  7
```

```
##    3  9  8  3
```