Week 7 assignment

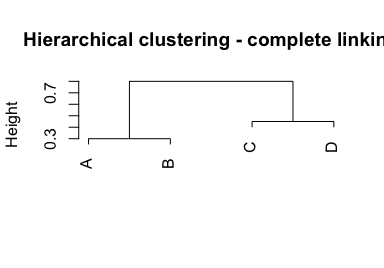
Kai Li

2/22/2017

## Question 1

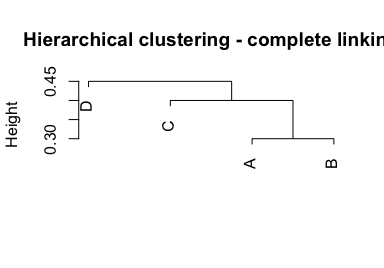
#### (a) Complete linking

data <- data.frame(A = c(0, 0.3, 0.4, 0.7),  
 B = c(0, 0, 0.5, 0.8),  
 C = c(0, 0, 0, 0.45),  
 D = c(rep(0, 4)))  
  
distance <- as.dist(data)  
  
plot(hclust(distance,   
 method = "complete"),  
 main = "Hierarchical clustering - complete linking",  
 xlab = "", sub = "")



#### (b) Single linking

plot(hclust(distance,   
 method = "single"),  
 main = "Hierarchical clustering - complete linking",  
 xlab = "", sub = "")

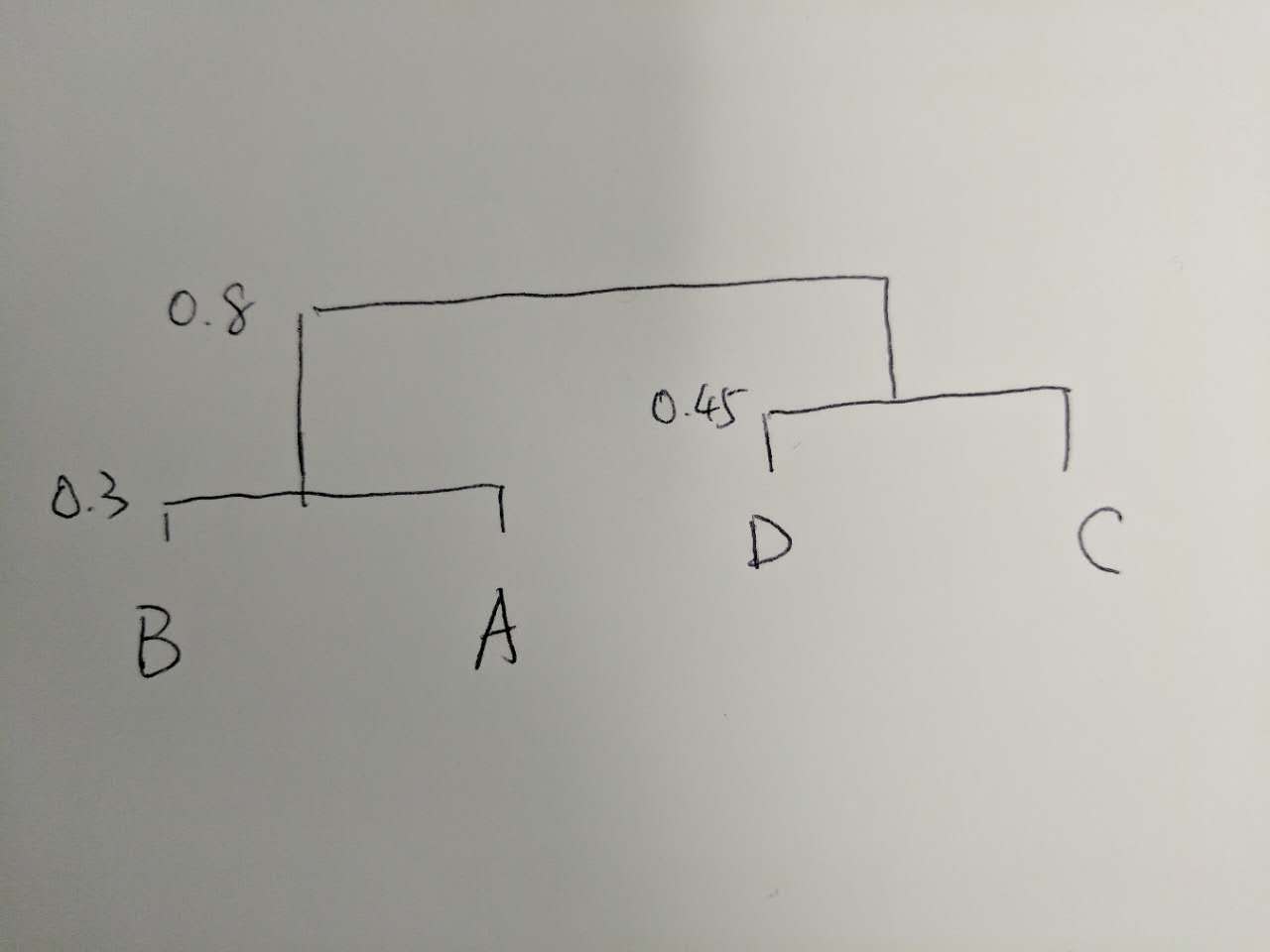


#### (c) & (d) Cutting dendogram

For the first dendogram, A & B are in the first cluster, C & D the second.

For the second dedogram, D is in its own cluster, while the other three in another cluster.

#### (e)



## Question 2

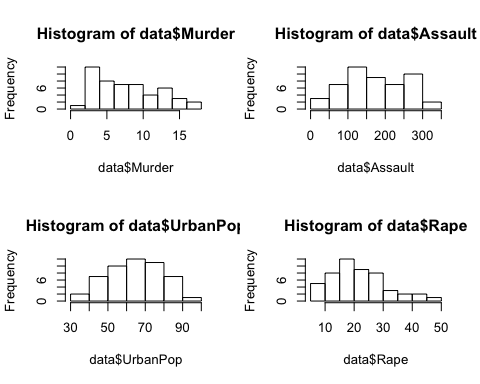
Below are some descriptive analysis of the data set. It seems that the variable of Rape is highly skewed.

data <- USArrests  
describeBy(data)

## Warning in describeBy(data): no grouping variable requested

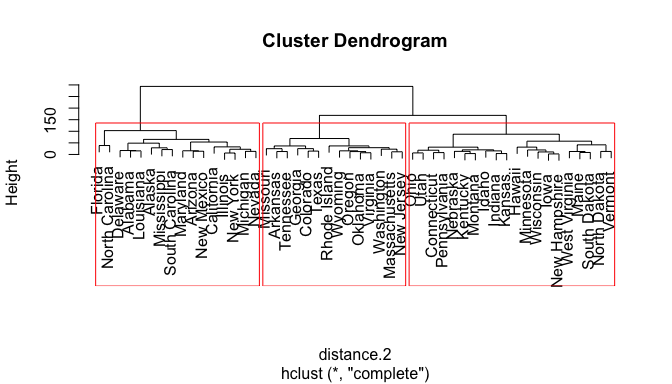
## vars n mean sd median trimmed mad min max range skew  
## Murder 1 50 7.79 4.36 7.25 7.53 5.41 0.8 17.4 16.6 0.37  
## Assault 2 50 170.76 83.34 159.00 168.47 110.45 45.0 337.0 292.0 0.22  
## UrbanPop 3 50 65.54 14.47 66.00 65.88 17.79 32.0 91.0 59.0 -0.21  
## Rape 4 50 21.23 9.37 20.10 20.36 8.60 7.3 46.0 38.7 0.75  
## kurtosis se  
## Murder -0.95 0.62  
## Assault -1.15 11.79  
## UrbanPop -0.87 2.05  
## Rape 0.08 1.32

par(mfrow = c(2,2))  
hist(data$Murder)  
hist(data$Assault)  
hist(data$UrbanPop)  
hist(data$Rape)



#### (a) & (b) complete linkage and cut into three clusters

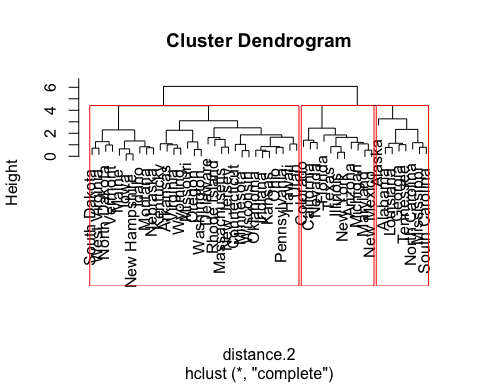
distance.2 <- dist(data)  
fit <- hclust(distance.2,  
 method = "complete")  
plot(fit)  
rect.hclust(fit, k = 3)



Based on the results, the first category includes: Florida, N Carolina, Delaware, Alabama, Louisiana, Alaska, Mississippi, S Carolina, Maryland, Arizona, New Mexico, California, Illinois, New York, Michigan, and Nevada. The second category includes Missouri, Arkansas, Tennessee, Georgia, Colorado, Texas, Rhode Island, Wyoming, Oregon, Oklahoma, Virginia, Washington, Mass., and New Jersey. The last category includes Ohio, Utah, Conn., Penn., Nebraska, Kentucky, Montana, Idaho, Indiana, Kansas, Hawaii, Minnesota, Wisconsin, Iowa, New Hampshire, W Virginia, Maine, S Dakota, N Dakota, and Vermont.

#### (c) Hierarchical clustering based on scaled data

data[,1:4] <- scale(data[,1:4])  
distance.2 <- dist(data)  
fit <- hclust(distance.2,  
 method = "complete")  
plot(fit)  
rect.hclust(fit, k = 3)



#### (d) Reasoning scale

Comparing the different trees acquired in section b and c, scaling the variables caused different clusters. In this case, scaling the variables is necessary, because the different variables have different units.