**Perform Clustering for the crime data and identify the number of clusters formed and draw inferences.**

**Data Description:**

**Murder -- Muder rates in different places of United States**

**Assualt- Assualt rate in different places of United States**

**UrbanPop - urban population in different places of United States**

**Rape - Rape rate in different places of United States**

crime\_data<-read.csv(file.choose(),header=T)

View(crime\_data)

Head(crime\_data)

X Murder Assault UrbanPop Rape

1 Alabama 13.2 236 58 21.2

2 Alaska 10.0 263 48 44.5

3 Arizona 8.1 294 80 31.0

4 Arkansas 8.8 190 50 19.5

5 California 9.0 276 91 40.6

6 Colorado 7.9 204 78 38.7

**#Normalize the data**

normalized\_crime\_data<-scale(crime\_data[2:5])

View((normalized\_crime\_data))

Head(normalized\_crime\_data)

Murder Assault UrbanPop Rape

[1,] 1.24256408 0.7828393 -0.5209066 -0.003416473

[2,] 0.50786248 1.1068225 -1.2117642 2.484202941

[3,] 0.07163341 1.4788032 0.9989801 1.042878388

[4,] 0.23234938 0.2308680 -1.0735927 -0.184916602

[5,] 0.27826823 1.2628144 1.7589234 2.067820292

[6,] 0.02571456 0.3988593 0.8608085 1.86496720

**#here I am using hierarchical clustering**

**#create distance matrix using Euclidean distance**

distance<-dist(normalized\_crime\_data,method = "euclidean")

**#create dendrogram using different linkage, and choose better linkage for clustering**

**# create dendrogram using single linkage**

fit\_crime\_data1 <- hclust(distance, method="single")

fit\_crime\_data1

Call:

hclust(d = distance, method = "single")

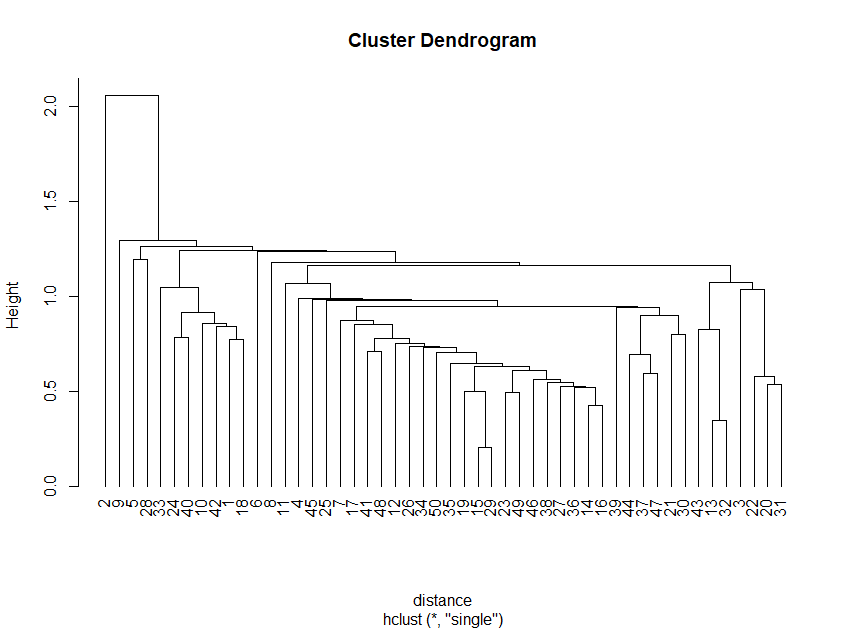
Cluster method : single

Distance : euclidean

Number of objects: 50

**#plot the dendrogram**

plot(fit\_crime\_data1, hang=-1)



**# using single, it is very difficult to interpret, so we would try different**

**#linkages for better dendrogram**

**## create dendrogram using complete linkage**

**# create dendrogram using single linkage**

fit\_crime\_data2 <- hclust(distance, method="complete")

fit\_crime\_data2

Call:

hclust(d = distance, method = "complete")

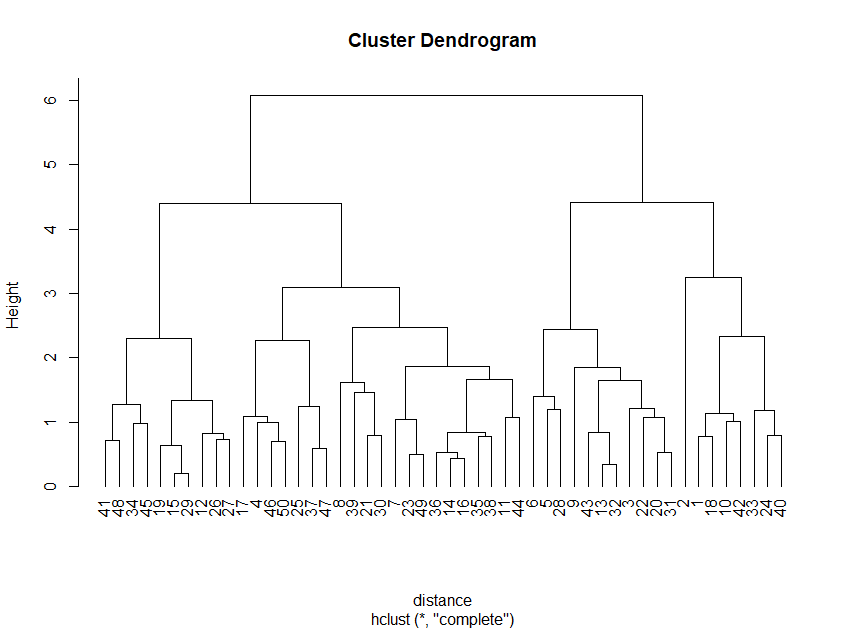
Cluster method : complete

Distance : euclidean

Number of objects: 50

**#plot the dendrogram**

plot(fit\_crime\_data2, hang=-1)



**#using complete linkage we got better dendrogram**

**#lets try centroid linkage for creating dendroid**

fit\_crime\_data3 <- hclust(distance, method="centroid")

fit\_crime\_data3

Call:

hclust(d = distance, method = "centroid")

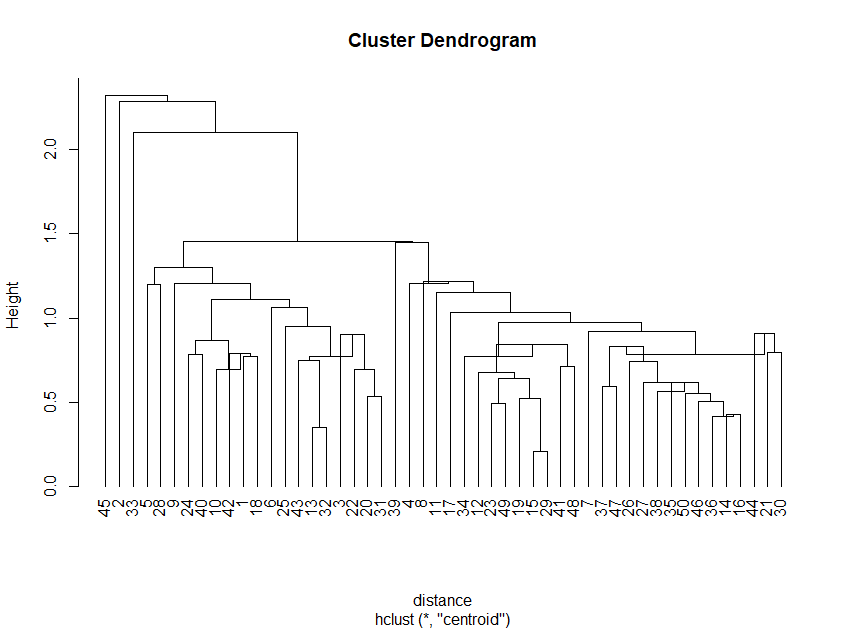
Cluster method : centroid

Distance : euclidean

Number of objects: 50

**#plot the dendrogram**

plot(fit\_crime\_data3, hang=-1)



**# using average linkages**

fit\_crime\_data4 <- hclust(distance, method="average")

fit\_crime\_data4

Call:

hclust(d = distance, method = "average")

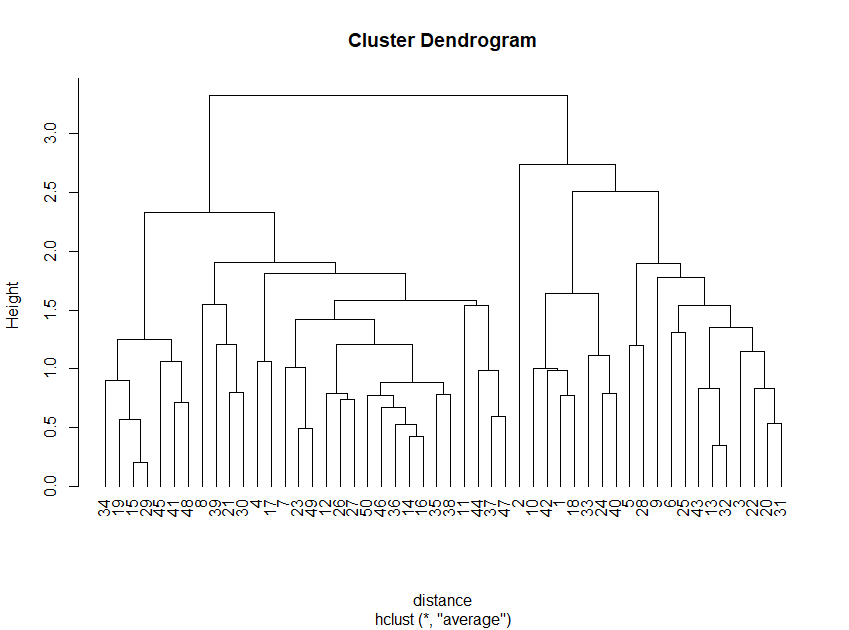
Cluster method : average

Distance : euclidean

Number of objects: 50

**#plot the dendrogram**

plot(fit\_crime\_data4, hang=-1)



**#here complete linkage provide a simple dendrogram**

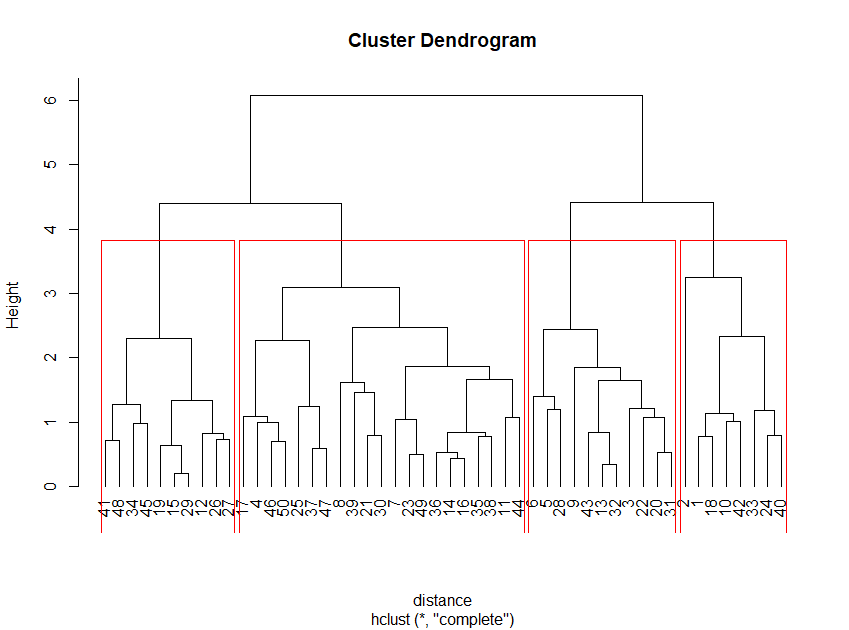
**#so we can choose fit\_crime\_data2 for further processing**

**#next step is to cut the dedrogram**

**#here i am going to cut dendrogram into four cluster**

crime\_groups <- cutree(fit\_crime\_data2, k=4)# cut tree into 4 clusters

rect.hclust(fit\_crime\_data2, k=4, border="red")



**#convert groups information into a matrix for better understanding**

crime\_cluster<-as.matrix(crime\_groups)

View(crime\_cluster)

**#create dataframe to combine crime\_cluster and original data**

data <- data.frame(crime\_data, crime\_cluster)

View(data)

Head(data)

X Murder Assaul UrbanPop Rape crime\_cluster

1 Alabama 13.2 236 58 21.2 1

2 Alaska 10.0 263 48 44.5 1

3 Arizona 8.1 294 80 31.0 2

4 Arkansas 8.8 190 50 19.5 3

5 California 9.0 276 91 40.6 2

6 Colorado 7.9 204 78 38.7 2

**#here i am going to change the position of the column crime\_cluster in to first**

final\_data <- data[,c(ncol(data),1:(ncol(data)-1))]

View(final\_data)

Head(final\_data)

crime\_cluster X Murder Assault UrbanPop Rape

1 1 Alabama 13.2 236 58 21.2

2 1 Alaska 10.0 263 48 44.5

3 2 Arizona 8.1 294 80 31.0

4 3 Arkansas 8.8 190 50 19.5

5 2 California 9.0 276 91 40.6

6 2 Colorado 7.9 204 78 38.7

**#here I am changing second column name as State for better analysis**

names(final\_data)[2]<-"State"

head(final\_data)

crime\_cluster State Murder Assault UrbanPop Rape

1 1 Alabama 13.2 236 58 21.2

2 1 Alaska 10.0 263 48 44.5

3 2 Arizona 8.1 294 80 31.0

4 3 Arkansas 8.8 190 50 19.5

5 2 California 9.0 276 91 40.6

6 2 Colorado 7.9 204 78 38.7

**#display the states of each cluster**

crime\_cluster1 <-subset(final\_data,crime\_cluster==1)

crime\_cluster1$State

[1] "Alabama" "Alaska" "Georgia" "Louisiana" "Mississippi"

[6] "North Carolina" "South Carolina" "Tennessee"

crime\_cluster2 <-subset(final\_data,crime\_cluster==2)

crime\_cluster2$State

[1] "Arizona" "California" "Colorado" "Florida" "Illinois" "Maryland"

[7] "Michigan" "Nevada" "New Mexico" "New York" "Texas"

crime\_cluster3 <-subset(final\_data,crime\_cluster==3)

crime\_cluster3$State

[1] "Arkansas" "Connecticut" "Delaware" "Hawaii" "Indiana"

[6] "Kansas" "Kentucky" "Massachusetts" "Minnesota" "Missouri"

[11] "New Jersey" "Ohio" "Oklahoma" "Oregon" "Pennsylvania"

[16] "Rhode Island" "Utah" "Virginia" "Washington" "Wisconsin"

[21] "Wyoming"

crime\_cluster4 <-subset(final\_data,crime\_cluster==4)

crime\_cluster4$State

[1] "Idaho" "Iowa" "Maine" "Montana" "Nebraska"

[6] "New Hampshire" "North Dakota" "South Dakota" "Vermont" "West Virginia"

**#display the average murder rate of each cluster**

tapply(final\_data$Murder,crime\_cluster,mean)

1 2 3 4

14.087500 11.054545 5.871429 3.180000

**#here we can see that cluster 1 has highest murder rate**

**#display average Assualt rate of each cluster**

tapply(final\_data$Assault,crime\_cluster,mean)

1 2 3 4

252.7500 264.0909 134.4762 78.7000

**#here cluster2 has highest assult rate and cluster 4 has lowest assault rate**

**#display average urban population of each cluster**

tapply(final\_data$UrbanPop,crime\_cluster,mean)

1 2 3 4

53.50000 79.09091 70.76190 49.30000

**#here cluster 2 has highest urben population and cluster 4 has lowest**

**#display average Rape rate in different cluster**

tapply(final\_data$Rape,crime\_cluster,mean)

1 2 3 4

24.53750 32.61818 18.58095 11.63000

**##here cluster 2 has highest rape rate**