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

library(readr)

## importing the data set

data <- read.csv(choose.files())

View(data)

ncol(data)

##Removing the state column because it does not effect our cluster results##

newdata <- data[,2:5]

##As we performing clustering we should perform normalization of the Data set because the model depends on the euclineary distance##

nordata <- scale(newdata)

##Now we should find distance##

d <- dist(nordata,method = "euclidean")

str(d)

##Model##

crimerate <- hclust(d,method = "complete")##H clustering model

plot(crimerate,hang = -1)

rect.hclust(crimerate,plot(crimerate,hang = -1),k=4,border = "red")## rectangular cluster

groups <- cutree(crimerate,k=4)##cutting the tree into 4 clusters

crime <- as.matrix(groups)##coverting the groups as the matrix form

crimefinal <- cbind(data,crime)

View(crimefinal)

write.csv(crimefinal, file="crimefinal.csv",row.names = F)

aggregate(data[,-1],by=list(crimefinal$crime),FUN=mean)

|  |
| --- |
| Group.1 Murder Assault UrbanPop Rape  1 1 14.087500 252.7500 53.50000 24.53750  2 2 11.054545 264.0909 79.09091 32.61818  3 3 5.871429 134.4762 70.76190 18.58095  4 4 3.180000 78.7000 49.30000 11.63000 |
|  |
| |  | | --- | | > | |

Inference from the model

* From the dendogram we can say the group two has large number data point in it
* Has the group 2 has the large data points the mean of the variable are large compare to other groups
* Considering the data points size the group one has highest number of crimes committed according to mean
* The mean of murder is highest in the group 1
* Group 4 has less number of crimes committed