Cluster Practice

library(tidyverse)

dt<- read.csv("C:/temp/MSA/University.csv")  
head(dt)

## University SAT Top10 Accept SFRatio Expenses Grad  
## 1 Harvard 14.00 91 14 11 39.525 97  
## 2 Princeton 13.75 91 14 8 30.220 95  
## 3 Yale 13.75 95 19 11 43.514 96  
## 4 Stanford 13.60 90 20 12 36.450 93  
## 5 MIT 13.80 94 30 10 34.870 91  
## 6 Duke 13.15 90 30 12 31.585 95

dim(dt)

## [1] 25 7

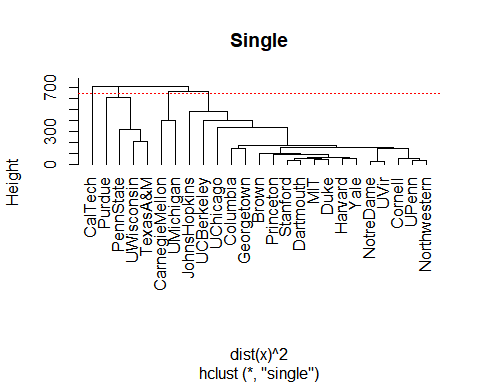
x<-dt[,2:7]  
  
######1#####  
h1<-hclust(dist(x)\*\*2,method="single")  
plot(h1,labels=dt$University,hang=-1,main="Single")  
cutree(h1,k=4)

## [1] 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 3 3 1 4 4 4 4

cutree(h1,h=650)

## [1] 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 3 3 1 4 4 4 4

abline(h=650,col="red",lty=3)



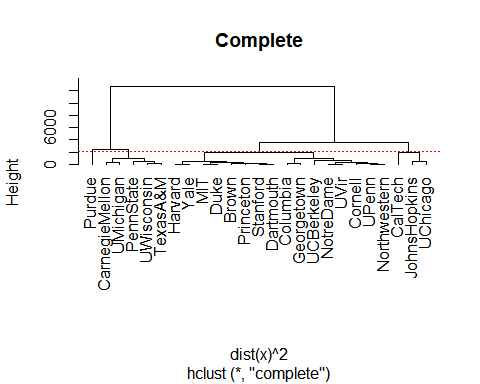
######2#####  
h2<-hclust(dist(x)\*\*2,method="complete")  
plot(h2,labels=dt$University,hang=-1,main="Complete")  
cutree(h2,k=4)

## [1] 1 1 1 1 1 1 2 1 1 2 2 1 1 1 1 1 1 1 3 3 1 3 3 4 3

cutree(h2,h=2050)

## [1] 1 1 1 1 1 1 2 1 1 2 2 1 1 1 1 1 1 1 3 3 1 3 3 4 3

abline(h=2050,col="red",lty=3)



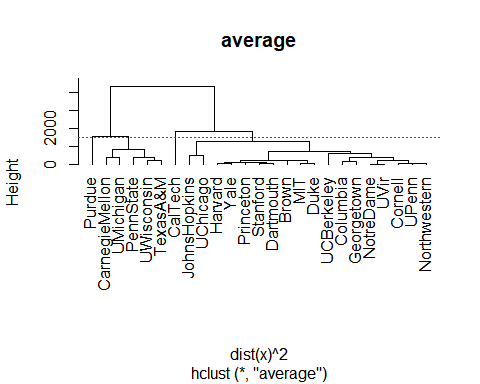
######3#####  
h3<-hclust(dist(x)\*\*2,method="average")  
plot(h3,labels=dt$University,hang=-1,main="average")  
cutree(h3,k=4)

## [1] 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 3 3 1 3 3 4 3

cutree(h3,h=1500)

## [1] 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 3 3 1 3 3 4 3

abline(h=1500,col="red",lty=3)



######4#####  
ck<-kmeans(x,centers=4)  
ck

## K-means clustering with 4 clusters of sizes 9, 2, 8, 6  
##   
## Cluster means:  
## SAT Top10 Accept SFRatio Expenses Grad  
## 1 13.51667 89.44444 21.77778 11.0 33.61556 93.77778  
## 2 13.60000 87.50000 34.50000 6.5 61.13300 84.00000  
## 3 12.61250 81.25000 38.50000 13.0 22.44825 89.00000  
## 4 11.14333 47.00000 67.83333 17.0 13.38467 74.00000  
##   
## Clustering vector:  
## [1] 1 1 1 1 1 1 2 1 1 2 3 3 3 3 1 3 3 3 4 4 3 4 4 4 4  
##   
## Within cluster sum of squares by cluster:  
## [1] 876.0280 524.0317 1444.7549 2385.9341  
## (between\_SS / total\_SS = 79.7 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

str(ck)

## List of 9  
## $ cluster : int [1:25] 1 1 1 1 1 1 2 1 1 2 ...  
## $ centers : num [1:4, 1:6] 13.5 13.6 12.6 11.1 89.4 ...  
## ..- attr(\*, "dimnames")=List of 2  
## .. ..$ : chr [1:4] "1" "2" "3" "4"  
## .. ..$ : chr [1:6] "SAT" "Top10" "Accept" "SFRatio" ...  
## $ totss : num 25792  
## $ withinss : num [1:4] 876 524 1445 2386  
## $ tot.withinss: num 5231  
## $ betweenss : num 20562  
## $ size : int [1:4] 9 2 8 6  
## $ iter : int 3  
## $ ifault : int 0  
## - attr(\*, "class")= chr "kmeans"

table(ck$cluster)

##   
## 1 2 3 4   
## 9 2 8 6

######5#####  
h1.clust<-cutree(h1,k=4)  
h2.clust<-cutree(h2,k=4)  
h3.clust<-cutree(h3,k=4)  
ck.clust<-ck$cluster  
  
dt.c<-cbind(dt,h1.clust,h2.clust,h3.clust,ck.clust)  
  
p1<-ggplot(dt.c,aes(SAT,Accept))+geom\_text(aes(label=University))+  
 geom\_point(aes(color=factor(h1.clust)),size=10,alpha=0.2)+ggtitle("Single Linkage")  
  
p2<-ggplot(dt.c,aes(SAT,Accept))+geom\_text(aes(label=University))+  
 geom\_point(aes(color=factor(h2.clust)),size=10,alpha=0.2)+ggtitle("Complete Linkage")  
  
p3<-ggplot(dt.c,aes(SAT,Accept))+geom\_text(aes(label=University))+  
 geom\_point(aes(color=factor(h3.clust)),size=10,alpha=0.2)+ggtitle("Average Linkage")  
  
p4<-ggplot(dt.c,aes(SAT,Accept))+geom\_text(aes(label=University))+  
 geom\_point(aes(color=factor(ck.clust)),size=10,alpha=0.2)+ggtitle("K-means")  
  
  
#install.packages("ggpubr")  
ggpubr::ggarrange(p1,p2,p3,p4,ncol=2,nrow=2)

지도, 텍스트이(가) 표시된 사진

자동 생성된 설명