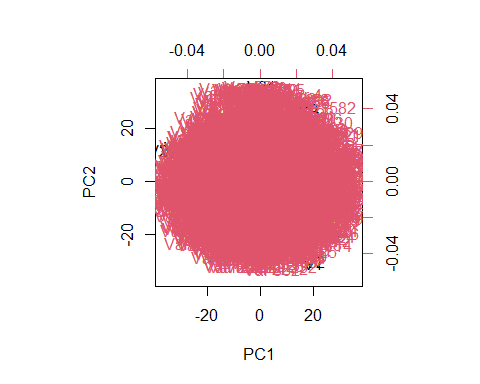
489 Homework 8

Nicholas Thompson

12/6/2021

set.seed(2401)  
#Question 2  
  
#a Perform PCA on the "xtrain" and plot the first two principal component score vectors  
  
pca2a=prcomp(Khan$xtrain, scale. = TRUE)  
biplot(pca2a, scale = 0)



#b Perform K-means clustering of the "xtrain" with K=4  
#How well do the cluster that you obtained in K-means clustering compare to the true class labels ("ytrain")?  
  
km2b=kmeans(Khan$xtrain,4)  
table(km2b$cluster, Khan$ytrain)

##   
## 1 2 3 4  
## 1 0 0 0 6  
## 2 4 8 3 6  
## 3 4 6 9 6  
## 4 0 9 0 2

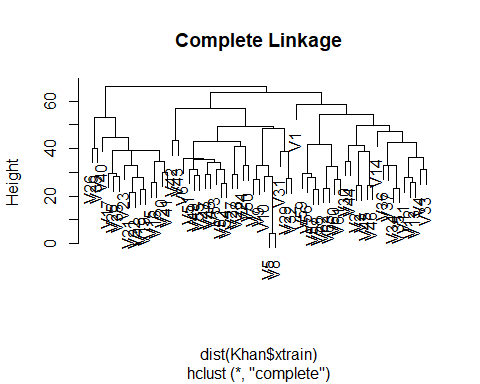
mean(km2b$cluster!=Khan$ytrain)

## [1] 0.6984127

km2b$betweenss/km2b$totss

## [1] 0.2627053

#The 4 means clustering is not a good fit and only predicted 30% correctly  
  
  
#c Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.  
hc.complete=hclust(dist(Khan$xtrain), method="complete")   
plot(hc.complete,main="Complete Linkage")



#d Cut the dendrogram at a height that results in 4 distinct clusters  
cutree(hc.complete,4)

## V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20   
## 1 2 2 2 3 3 3 3 3 3 2 2 2 2 4 4 4 4 4 4   
## V21 V22 V23 V24 V25 V26 V27 V28 V29 V30 V31 V32 V33 V34 V35 V36 V37 V38 V39 V40   
## 4 4 4 3 4 4 3 3 3 2 3 3 2 2 2 2 2 4 4 4   
## V41 V42 V43 V44 V45 V46 V47 V48 V49 V50 V51 V52 V53 V54 V55 V56 V57 V58 V59 V60   
## 4 3 3 2 2 2 3 3 3 3 3 3 3 3 3 2 2 2 2 2   
## V61 V62 V63   
## 2 2 2