clustering.R

Suraj Durgesht

16EC06

Fri Mar 8 13:35:17 2019

file\_loc <- read.csv("D:\\Programming\\DA\\Lab 4\\domestict20careerbattingrating\_mod.csv")[1:3]   
View(file\_loc)  
  
str(file\_loc)

## 'data.frame': 486 obs. of 3 variables:  
## $ Name : Factor w/ 483 levels "A A Bilakhia",..: 83 458 70 100 193 386 433 62 333 271 ...  
## $ Matches: int 256 204 256 220 312 242 233 236 234 240 ...  
## $ Innings: int 251 192 242 219 279 228 221 233 223 216 ...

summary(file\_loc)

## Name Matches Innings   
## A Singh : 2 Min. : 1.00 Min. : 0.00   
## S Sharma : 2 1st Qu.: 30.00 1st Qu.: 17.00   
## Yuvraj Singh : 2 Median : 60.00 Median : 38.00   
## A A Bilakhia : 1 Mean : 80.75 Mean : 59.58   
## A A Chavan : 1 3rd Qu.:119.00 3rd Qu.: 80.00   
## A A Jhunjhunwala: 1 Max. :312.00 Max. :279.00   
## (Other) :477

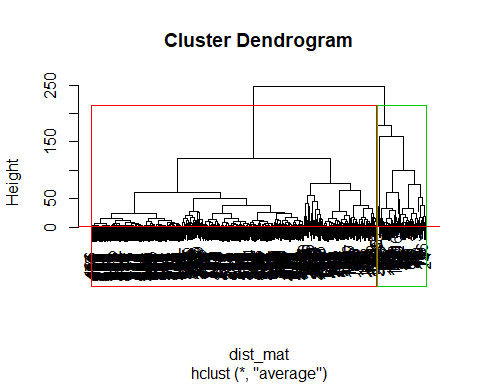
file\_loc[is.na(file\_loc)] <- 0  
str(file\_loc)

## 'data.frame': 486 obs. of 3 variables:  
## $ Name : Factor w/ 483 levels "A A Bilakhia",..: 83 458 70 100 193 386 433 62 333 271 ...  
## $ Matches: int 256 204 256 220 312 242 233 236 234 240 ...  
## $ Innings: int 251 192 242 219 279 228 221 233 223 216 ...

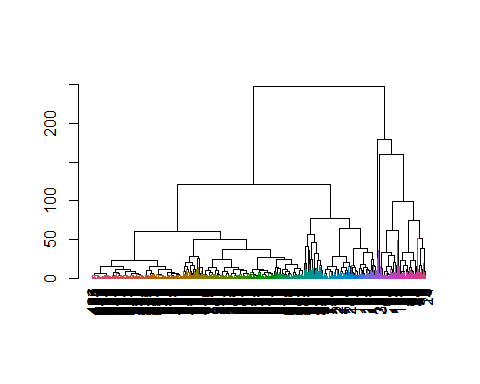
View(file\_loc)  
  
file\_loc\_sc <- as.data.frame(file\_loc)  
dist\_mat <- dist(file\_loc\_sc, method = 'euclidean')

## Warning in dist(file\_loc\_sc, method = "euclidean"): NAs introduced by  
## coercion

hclust\_avg <- hclust(dist\_mat, method = 'average')  
plot(hclust\_avg)  
cut\_avg <- cutree(hclust\_avg, k = 2)  
plot(hclust\_avg)  
rect.hclust(hclust\_avg , k = 2, border = 2:6)  
abline(h = 2, col = 'red')  
suppressPackageStartupMessages(library(dendextend))



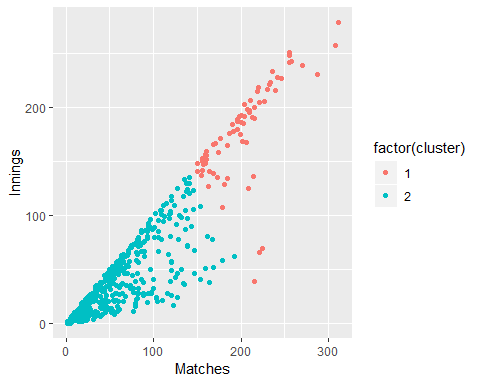
avg\_dend\_obj <- as.dendrogram(hclust\_avg)  
avg\_col\_dend <- color\_branches(avg\_dend\_obj, h = 3)  
plot(avg\_col\_dend)



suppressPackageStartupMessages(library(dplyr))  
file\_loc\_cl <- mutate(file\_loc, cluster = cut\_avg)  
count(file\_loc\_cl,cluster)

## # A tibble: 2 x 2  
## cluster n  
## <int> <int>  
## 1 1 72  
## 2 2 414

suppressPackageStartupMessages(library(ggplot2))  
ggplot(file\_loc\_cl, aes(x=Matches, y = Innings, color = factor(cluster))) + geom\_point()



table(file\_loc\_cl$cluster)

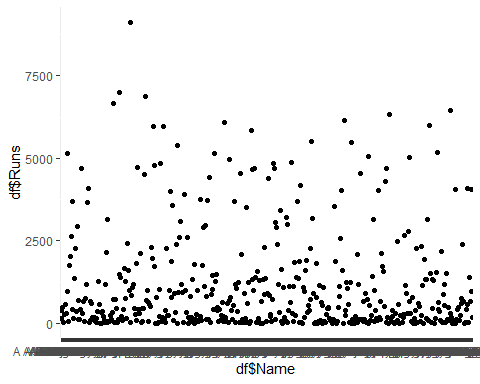
##   
## 1 2   
## 72 414

library(tidyverse) # data manipulation

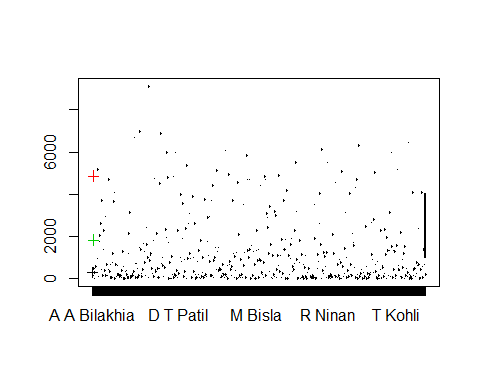
library(cluster) # clustering algorithms  
library(factoextra) # clustering algorithms & visualization

## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ

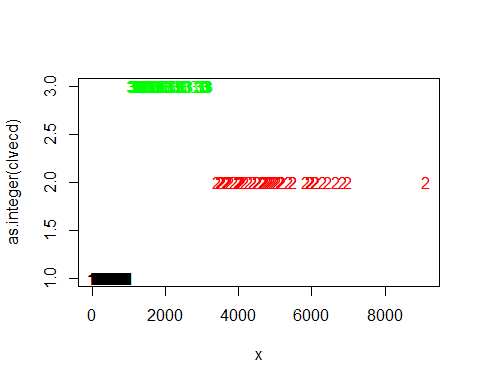
df <- read.csv("D:\\Programming\\DA\\Lab 4\\domestict20careerbattingrating\_mod.csv")  
df[is.na(df)] <- 0  
ggplot(df, aes(df$Name, df$Runs)) + geom\_point()



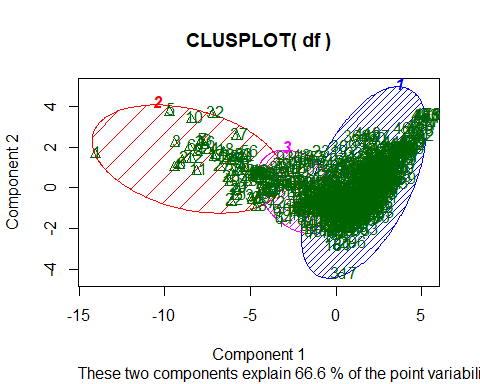
fit <- kmeans(df$Runs,3)  
plot(df$Name,df$Runs,col=fit$cluster,pch=1)  
points(fit$centers,col=1:8,pch=3)



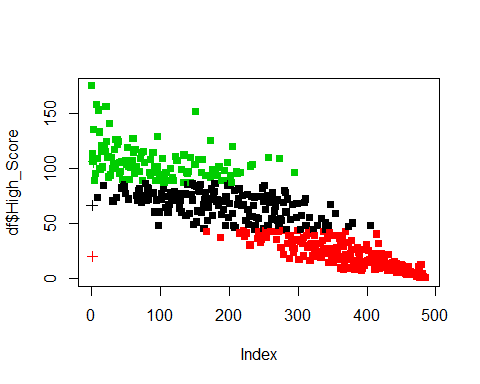
library(cluster)  
library(fpc)  
plotcluster(df$Runs,fit$cluster)  
points(fit$centers,col=1:8,pch=16)



clusplot(df, fit$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)



fit <- kmeans(df$High\_Score,3)  
plot(df$High\_Score,col=fit$cluster,pch=15)  
points(fit$centers,col=1:8,pch=3)



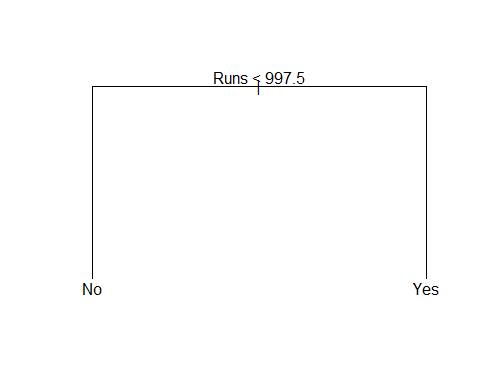
####################################################  
  
library(tree)  
df <- read.csv("D:\\Programming\\DA\\Lab 4\\domestict20careerbattingrating\_mod.csv")  
df[is.na(df)] <- 0  
High = ifelse(df$Runs < 1000, "No", "Yes")  
df = data.frame(df, High)  
tree.df = tree(High~.-df$Runs, data = df[,c(5,17)])  
summary(tree.df)

##   
## Classification tree:  
## tree(formula = High ~ . - df$Runs, data = df[, c(5, 17)])  
## Number of terminal nodes: 2   
## Residual mean deviance: 0 = 0 / 484   
## Misclassification error rate: 0 = 0 / 486

tree.df

## node), split, n, deviance, yval, (yprob)  
## \* denotes terminal node  
##   
## 1) root 486 629.2 No ( 0.6502 0.3498 )   
## 2) Runs < 997.5 316 0.0 No ( 1.0000 0.0000 ) \*  
## 3) Runs > 997.5 170 0.0 Yes ( 0.0000 1.0000 ) \*

plot(tree.df)  
text(tree.df, pretty = 0)



tree.df

## node), split, n, deviance, yval, (yprob)  
## \* denotes terminal node  
##   
## 1) root 486 629.2 No ( 0.6502 0.3498 )   
## 2) Runs < 997.5 316 0.0 No ( 1.0000 0.0000 ) \*  
## 3) Runs > 997.5 170 0.0 Yes ( 0.0000 1.0000 ) \*

set.seed(101)  
train=sample(1:nrow(df), 250)  
tree.df = tree(High~.-df$Runs,df[,c(5,17)], subset=train)  
plot(tree.df)  
text(tree.df, pretty=0)

