INDIAN INSTITUTE OF INFORMATION TECHNOLOGY
DHARWAD

Lab Report

Course: Data Analytics in R (CS6E23L)

Course Instructor: Dr. Kavi Mahesh

Lab Instructor: Pragya Verma

By:

Paka Sravan Kumar Yadav

6th Semester

3rd Year

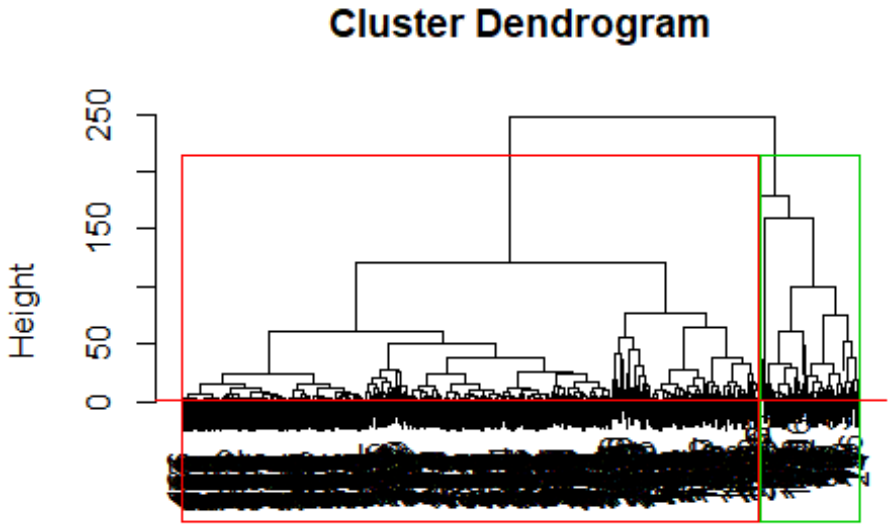
16CS11

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY
DHARWAD

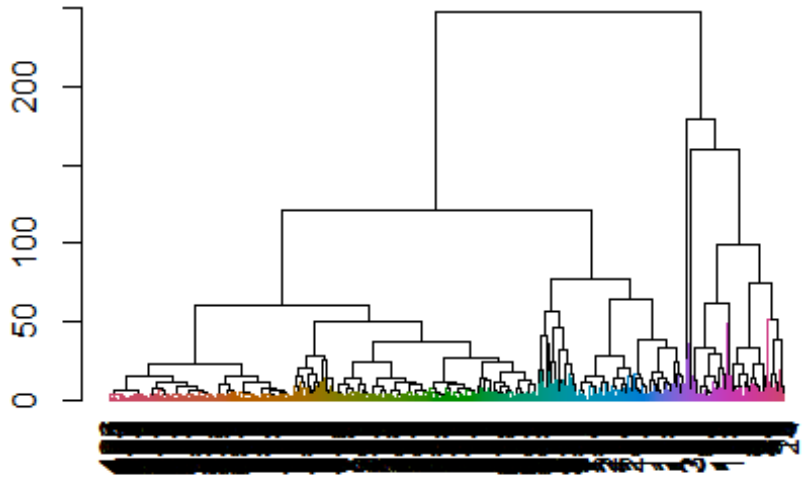
Lab – 02

```
file_loc <- read.csv("G:\\Required\\6th Sem\\DA\\Lab\\Lab2\\Cricket Dat
a Set\\Player Ratings\\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))
```



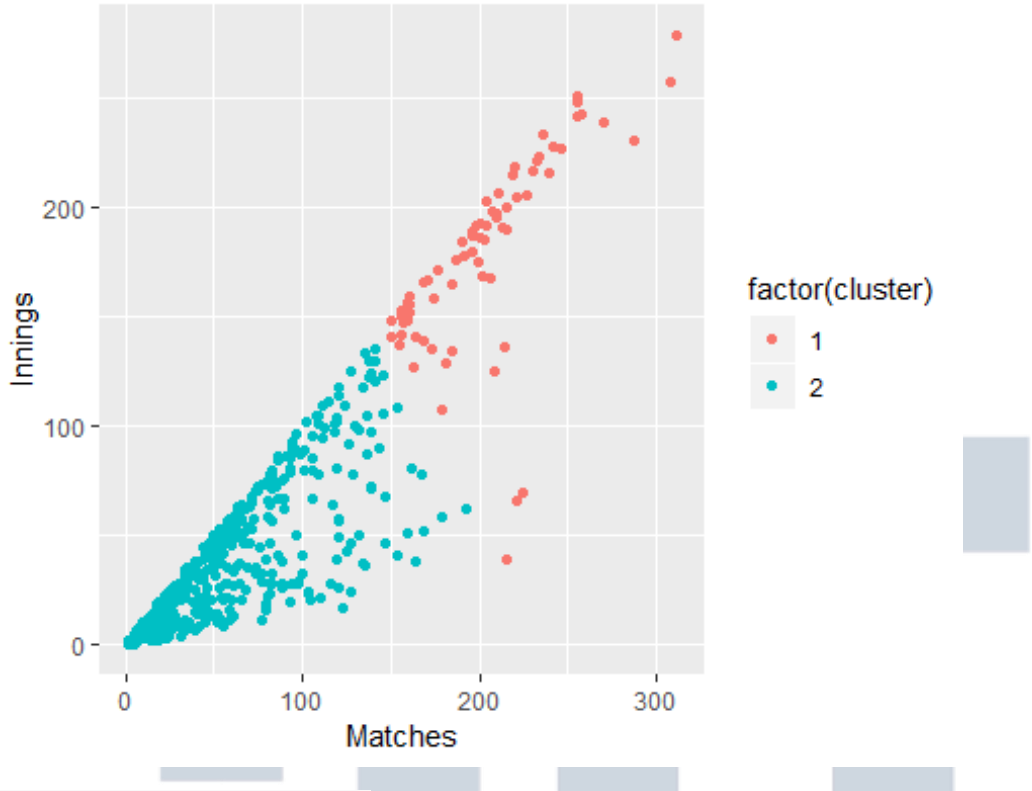

```
avg_dend_obj <- as.dendrogram(hclust_avg)
avg_col_dend <- color_branches(avg_dend_obj, h = 3)
plot(avg_col_dend)
```




```
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))) + geo
m_point()
```



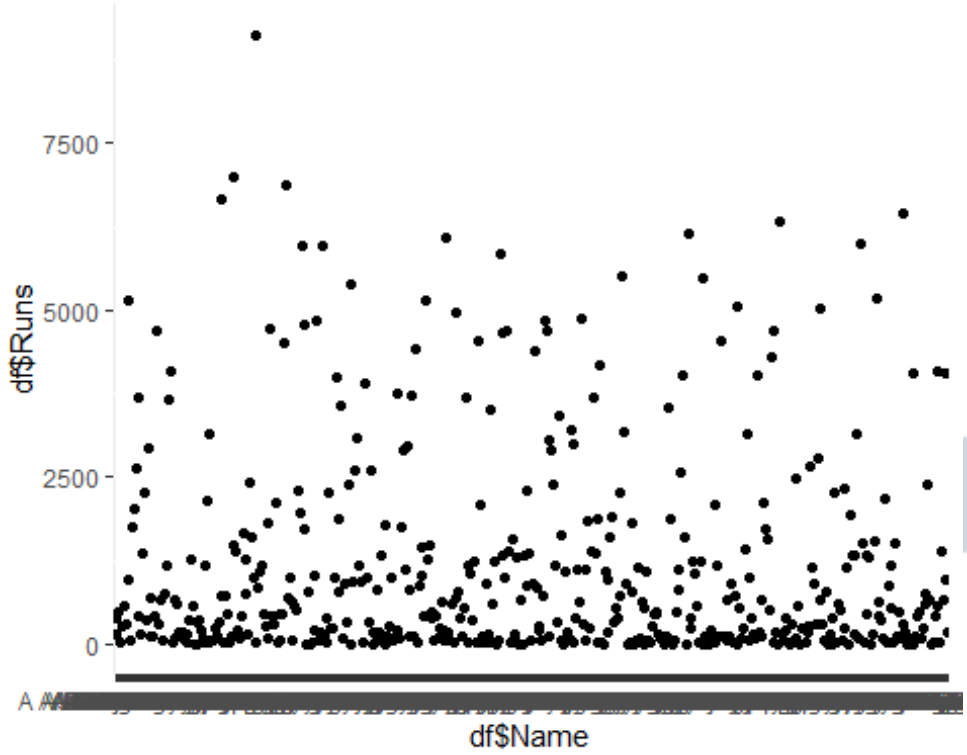

```
table(file_loc_cl$cluster)
```

```
##
##   1   2
##  72 414
```

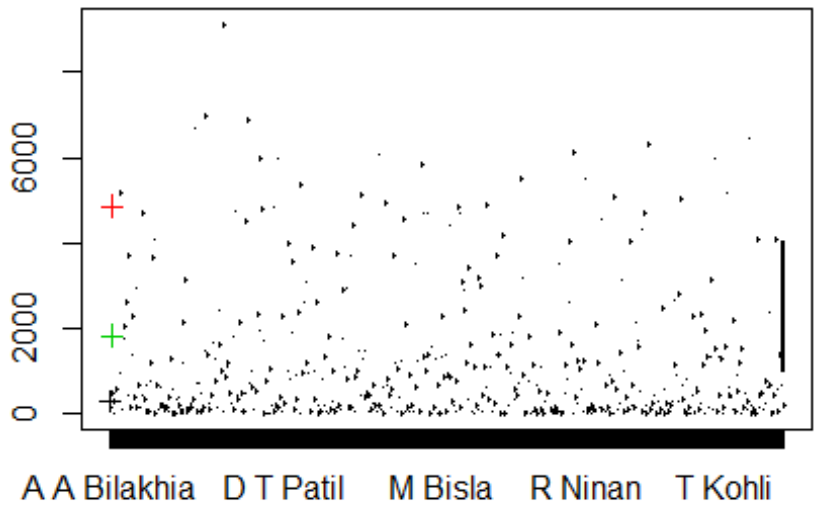
```
library(tidyverse)  # data manipulation

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

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

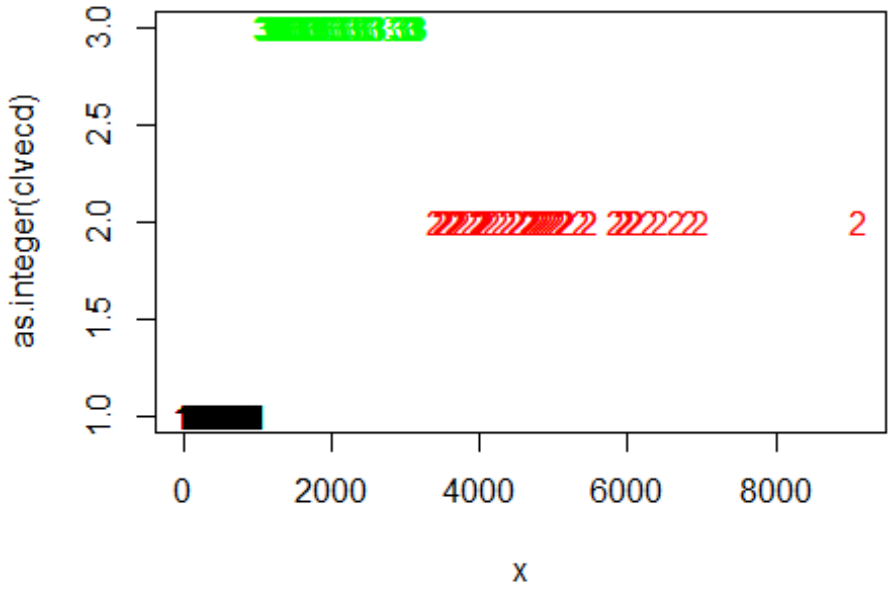
```
df <- read.csv("D:\\Programming\\DA\\Lab 4\\domestict20careerbattingrating_mod.csv")
df[is.na(df)] <- 0
ggplot(df, aes(df$Name, df$Runs)) + geom_point()
```




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

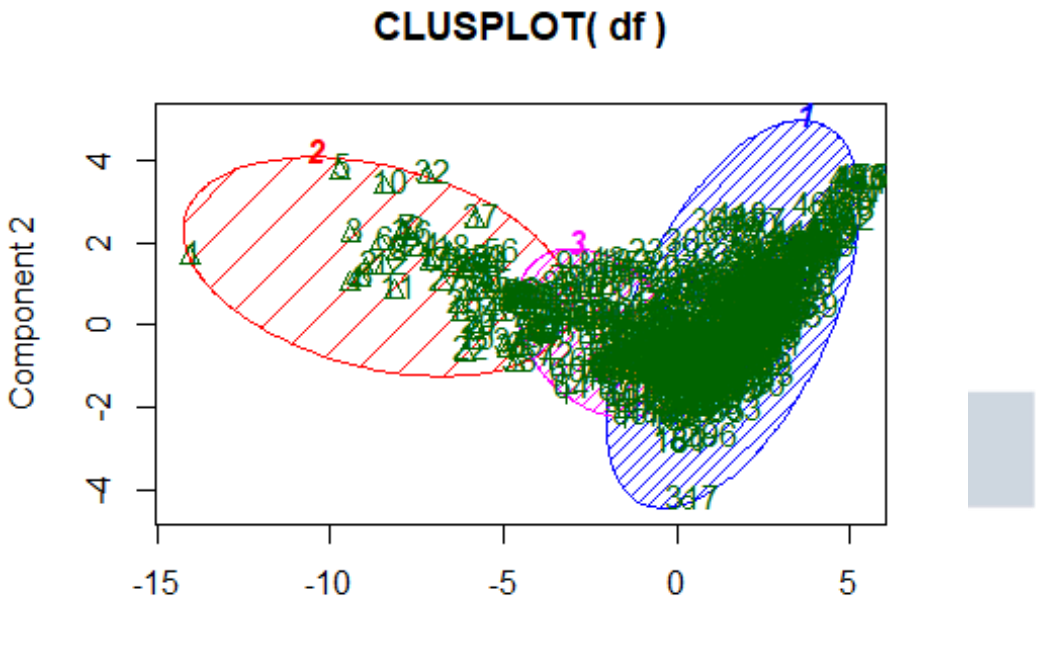


```
library(cluster)
library(fpc)
plotcluster(df$Runs,fit$cluster)
points(fit$centers,col=1:8,pch=16)
```

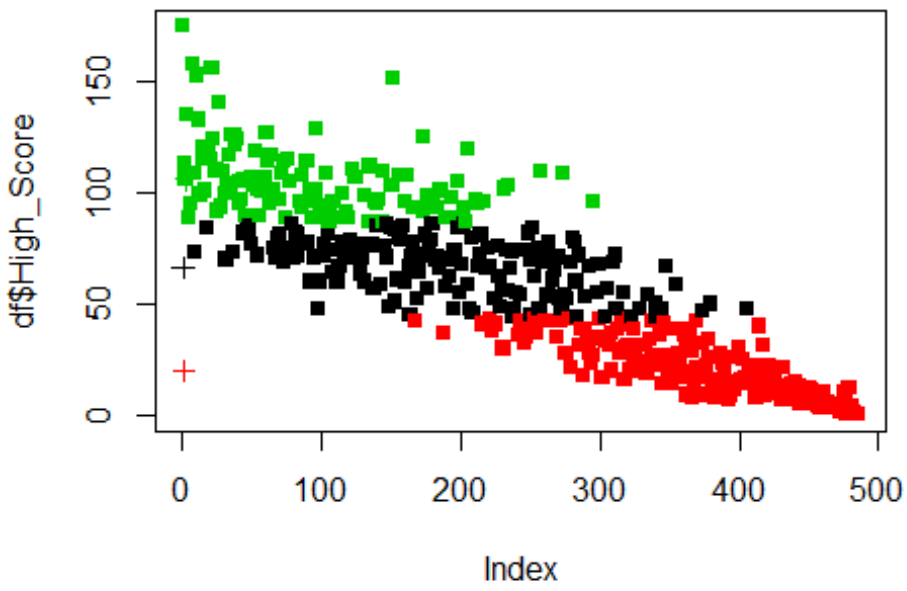



```
clusplot(df, fit$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)
```

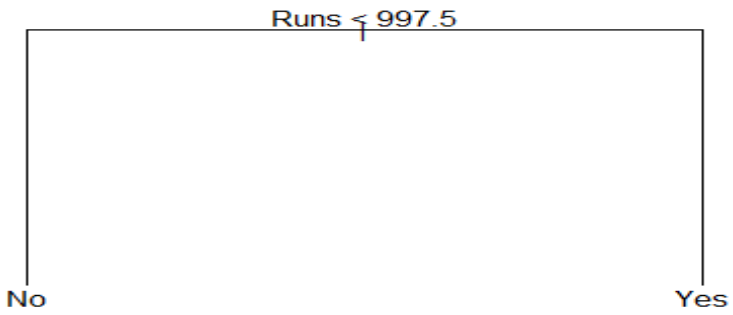
INDIAN INSTITUTE OF INFORMATION TECHNOLOGY
DHARWAD




```
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)
```