Clustering

This project involves Hitters dataset from the ISLR package in R. It consists of 20 variables measured on 263 major league baseball players (after removing those with missing data). Salary is the response variable and the remaining 19 are predictors. Some of the predictor variables are categorical with two classes. Our goal is to clustering the players.

```
library(ISLR)
str(Hitters)
```

```
'data.frame':   322 obs. of  20 variables:
 $ AtBat    : int  293 315 479 496 321 594 185 298 323 401 ...
 $ Hits     : int  66 81 130 141 87 169 37 73 81 92 ...
 $ HmRun    : int  1 7 18 20 10 4 1 0 6 17 ...
 $ Runs     : int  30 24 66 65 39 74 23 24 26 49 ...
 $ RBI      : int  29 38 72 78 42 51 8 24 32 66 ...
 $ Walks    : int  14 39 76 37 30 35 21 7 8 65 ...
 $ Years    : int  1 14 3 11 2 11 2 3 2 13 ...
 $ CAtBat   : int  293 3449 1624 5628 396 4408 214 509 341 5206 ...
 $ CHits    : int  66 835 457 1575 101 1133 42 108 86 1332 ...
 $ CHmRun   : int  1 69 63 225 12 19 1 0 6 253 ...
 $ CRuns    : int  30 321 224 828 48 501 30 41 32 784 ...
 $ CRBI     : int  29 414 266 838 46 336 9 37 34 890 ...
 $ CWalks   : int  14 375 263 354 33 194 24 12 8 866 ...
 $ League   : Factor w/ 2 levels "A","N": 1 2 1 2 2 1 2 1 2 1 ...
 $ Division : Factor w/ 2 levels "E","W": 1 2 2 1 1 2 1 2 2 1 ...
 $ PutOuts  : int  446 632 880 200 805 282 76 121 143 0 ...
 $ Assists  : int  33 43 82 11 40 421 127 283 290 0 ...
 $ Errors   : int  20 10 14 3 4 25 7 9 19 0 ...
 $ Salary   : num  NA 475 480 500 91.5 750 70 100 75 1100 ...
 $ NewLeague: Factor w/ 2 levels "A","N": 1 2 1 2 2 1 1 1 2 1 ...
```

```
Hitters <- na.omit(Hitters)
dim(Hitters)
```

```
[1] 263  20
```

```
Hitters.X<-Hitters[-19]
Hitters.X$League<-ifelse(Hitters.X$League=="A",0,1)
Hitters.X$Division<-ifelse(Hitters.X$Division=="E",0,1)
Hitters.X$NewLeague<-ifelse(Hitters.X$NewLeague=="A",0,1)
```

a) Standardizing depends on the given application. It may well be that some variables are intrinsically more important than others in a particular application, and then the assignment of weights should be based on subject-matter Knowledge and in that case standardizing is not recomended. In some applications, changing the measurement units may even lead one to see a very different clustering structure. To avoid this dependence on the choice of measurement units, one has the option of standardizing the data. This converts the original measurements to unitless variables.In the given dataset Hitters, the variables are measured in different scales. Hence it is a good idea to standardize the variables.

b) Most of the features in Hitters data are not highly correlated. Therefore metric – based distance measurements are more suitable for this data.

c) Complete linkage. We selected 2 clusters and cluster 1 has 233 players and cluster 2 has 30 players. The second cluster means of the variables are higher than the first cluster means of the variables. The mean salary of the second cluster is higher than the first cluster Except Assist and Error.

```
xsc <- scale(Hitters.X)
xsc.hc.complete <- hclust(dist(xsc), method = "complete")
cut.tree<-cutree(xsc.hc.complete, 2)
```

```
plot(xsc.hc.complete, main = "", xlab = "", sub = "", cex = 0.3)
```

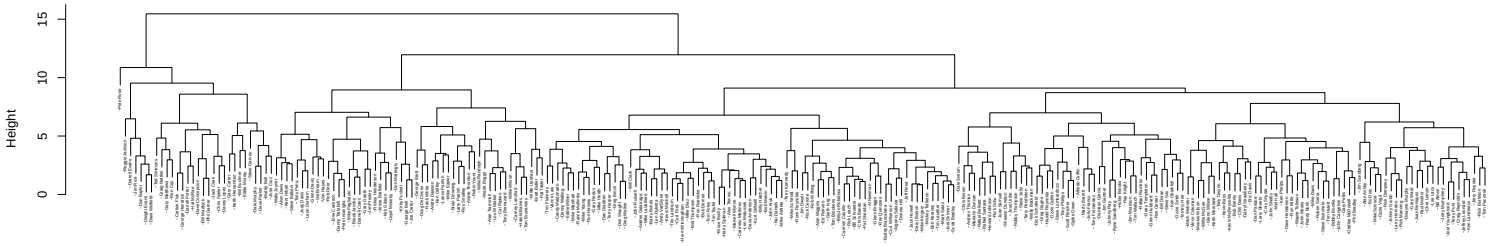

Figure 1: Hierarchical Clustering for Hitters Data

```
library(xtable)
print(xtable(tab.mu[,1:10]),table.placement="H")
```

% latex table generated in R 4.3.0 by xtable 1.8-4 package % Sat Nov 18 02:17:14 2023

	AtBat	Hits	HmRun	Runs	RBI	Walks	Years	CAtBat	CHits	CHmRun
cluster1.mu	402.37	107.61	11.07	54.48	49.86	39.97	6.21	2068.36	555.89	47.49
cluster2.mu	413.50	109.57	15.87	56.83	64.13	50.03	15.83	7233.50	2013.73	238.17

```
print(xtable(tab.mu[,11:19],caption = "Cluster means"),table.placement="H")
```

% latex table generated in R 4.3.0 by xtable 1.8-4 package % Sat Nov 18 02:17:14 2023

	CRuns	CRBI	CWalks	League	Division	PutOuts	Assists	Errors	NewLeague
cluster1.mu	275.52	239.76	191.93	0.45	0.51	282.51	125.10	8.75	0.45
cluster2.mu	1026.87	1034.53	791.03	0.60	0.53	354.40	69.50	7.40	0.57

Table 1: Cluster means

```
# mean salary of the players in the two clusters
cluster1.sal<-mean(Hitters$Salary[cut.tree==1])
cluster2.sal<-mean(Hitters$Salary[cut.tree==2])
```

```
tab.salary<-rbind(cluster1.sal,cluster2.sal)
colnames(tab.salary)<-"Mean Salary"
rownames(tab.salary)<-c("Cluster 1","Cluster 2")
print(xtable(tab.salary,caption = "Mean salary of the players in the two clusters"),table.placement="H")
```

% latex table generated in R 4.3.0 by xtable 1.8-4 package % Sat Nov 18 02:17:14 2023

	Mean Salary
Cluster 1	479.95
Cluster 2	970.68

Table 2: Mean salary of the players in the two clusters

d) K-means clustering. We selected K=2 clusters and cluster 1 has 189 players and cluster 2 has 74 players. The second cluster means of the variables are higher than the first cluster means of the variables. The mean salary of the second cluster is higher than the first cluster except League, Division, Assists,Errors and NewLeague.

```
# K-means with K = 2
set.seed(1)
km.out <- kmeans(xsc, 2, nstart = 20)
cut.km<-km.out$cluster

# Cluster means of the variables
# 1st cluster means of the variables
km.cluster1<-apply(Hitters.X[cut.km==1,],2,mean)

# 2nd cluster means of the variables
km.cluster2<-apply(Hitters.X[cut.km==2,],2,mean)
```

```
km.mean.tab<-rbind(km.cluster1,km.cluster2)
print(xtable(km.mean.tab[,1:10]),table.placement="H")
```

% latex table generated in R 4.3.0 by xtable 1.8-4 package % Sat Nov 18 02:17:14 2023

	AtBat	Hits	HmRun	Runs	RBI	Walks	Years	CAtBat	CHits	CHmRun
km.cluster1	377.54	99.49	9.41	49.53	44.81	35.96	5.30	1571.89	414.74	31.31
km.cluster2	470.31	129.14	17.26	68.05	68.53	54.28	12.45	5430.36	1507.43	166.12

```
print(xtable(km.mean.tab[,11:19],caption = "Cluster means of the variables (K-means)"),table.placement="H")
```

% latex table generated in R 4.3.0 by xtable 1.8-4 package % Sat Nov 18 02:17:14 2023

	CRuns	CRBI	CWalks	League	Division	PutOuts	Assists	Errors	NewLeague
km.cluster1	199.68	171.35	135.61	0.53	0.55	265.82	125.26	8.92	0.51
km.cluster2	773.80	736.68	578.64	0.32	0.41	354.28	102.15	7.77	0.35

Table 3: Cluster means of the variables (K-means)

```
# mean salary of the players in the two clusters
km.cluster1.sal<-mean(Hitters$Salary[cut.km==1])
km.cluster2.sal<-mean(Hitters$Salary[cut.km==2])
```

```
tab.salary1<-rbind(km.cluster1.sal,km.cluster2.sal)
colnames(tab.salary1)<-"Mean Salary"
rownames(tab.salary1)<-c("Cluster 1","Cluster 2")
print(xtable(tab.salary1,caption = "Mean salary of the players in the two clusters"),table.placement="H")
```

% latex table generated in R 4.3.0 by xtable 1.8-4 package % Sat Nov 18 02:17:14 2023

	Mean Salary
Cluster 1	368.55
Cluster 2	963.40

Table 4: Mean salary of the players in the two clusters

e) According to the above results both clustering methods give relatively similar results. Therefore identifying the better algorithm that gives more sensible results would be a difficult choice.