Homework 1: Principal Component Analysis (PCA) and Multidimensional Scaling (MDS)

Group 13 : Anna Monsó Rodríguez, Walter José Troiani Vargas, Joan Acero Pousa

2024-10-07

0. Import the data set "euroleague_23_24.csv": the player statistics of four teams taken part in Final Four of Euro League 2023-2024.

```
rm(list=ls())
euroleague_23_24 <- read.csv2("euroleague_23_24.csv")
```

1. Exploratory data analysis

```
# a) Discard the variable "No" from the data set.

df <- subset(euroleague_23_24, select = -No)
```

```
# b) Split variable "Min" using strsplit() function. Give the name "aux" to the output.
#The first element of each row will show the mean minutes that the player played in total. (1p)

aux <- sapply(strsplit(df$Min, split=':'), function(x) x[1])
```

```
# c) Add a numerical variable to the data set named "Min 2" which shows on average
# how many minutes each player played in the game.

df$Min2 <- as.numeric(aux)
```

```
# d) Check the structure of the data and assign correct type to each variable considering
# whether it is a categorical or numerical variable.

type_info <- str(df)
```

```
## 'data.frame':    64 obs. of  22 variables:
##  $ TEAM    : chr  "PANATHINAIKOS" "PANATHINAIKOS" "PANATHINAIKOS" "PANATHINAIKOS" ...
##  $ PLAYER  : chr  "PANAGIOTIS KALAITZAKIS " "LUCA VILDOZA" "KYLE GUY" "DIMITRIS MORAITIS" ...
##  $ POSITION: chr  "Guard" "Guard" "Guard" "Guard" ...
##  $ GP      : int  30 28 8 7 24 34 1 16 41 35 ...
##  $ GS      : int  0 5 1 0 9 15 0 4 34 27 ...
##  $ Min     : chr  "5:56:00" "14:56:00" "10:38:00" "2:25:00" ...
##  $ PTS     : num  2.1 5.7 4 1.6 2.8 12.7 3 5.6 8.6 16 ...
##  $ X2P.    : num  69 42 71.4 25 62.9 59.1 0 46.9 49.7 46.6 ...
##  $ X3P.    : num  25 36.6 31.6 75 11.1 41.5 100 51.6 41.6 41 ...
```

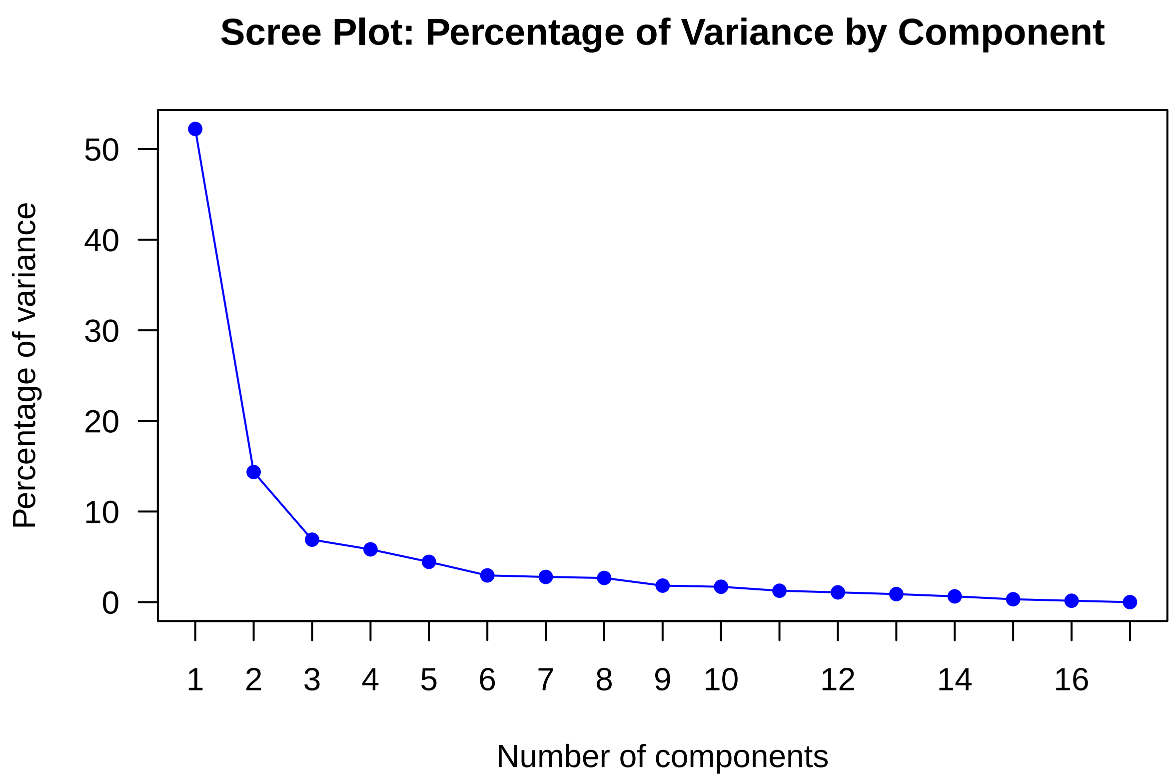
```
##  $ FT.      : num  100 76.2 80 0 70 85.3 0 80 86.1 95.9 ...
##  $ OR       : num  0.3 0.4 0 0 0.6 0.6 0 0.4 0.5 0.4 ...
##  $ DR       : num  0.6 1.1 0.9 0.3 0.8 2.6 0 1.6 1.8 2.3 ...
##  $ TR       : num  0.9 1.5 0.9 0.3 1.3 3.2 0 2 2.3 2.7 ...
##  $ AST      : num  0.2 1.5 0.8 0.7 0.3 5.6 1 0.7 3.5 3 ...
##  $ STL      : num  0.2 0.6 0.2 0.3 0.2 0.8 0 0.2 1.5 0.9 ...
##  $ TO       : num  0.2 1 1 0.3 0.3 2.4 0 0.4 1.1 3.1 ...
##  $ BLK      : num  0 0 0.1 0 0.4 0 0 0.2 0.1 0.1 ...
##  $ BLKA     : num  0 0.2 0 0.1 0.1 0.4 0 0.2 0.1 0.8 ...
##  $ FC       : num  0.8 0.8 1.2 0.1 1.5 1.8 0 1.4 2.3 2.2 ...
##  $ FD       : num  0.4 0.6 0.6 0 1.2 3 0 0.9 2.1 2.7 ...
##  $ PIR      : num  2.1 4.6 2.4 1.7 3.1 16.1 3 5.4 10.9 11.7 ...
##  $ Min2     : num  5 14 10 2 7 26 1 17 27 27 ...
```

```
df$TEAM <- as.factor(df$TEAM)
df$PLAYER<- as.factor(df$PLAYER)
df$POSITION <- as.factor(df$POSITION)
df$GP <- as.numeric(df$GP)
df$GS <- as.numeric(df$GS)
df <- subset(df, select = -Min)
```

2. Application of PCA

```
# a) Apply PCA on all the scaled numerical variables in the data set by using PCA()
# function in FactoMineR package. Treat the categorical variables and the variable "PIR"
# as suplemantary variables using arguments quali.sup and quanti.sup correctly. (3p)

library(FactoMineR)

pca_res <- PCA(df, scale.unit = TRUE, graph = FALSE,
           quali.sup = which(names(df) %in% c("TEAM","PLAYER","POSITION")),
           quanti.sup = which(names(df) == "PIR"))
eig_info <- pca_res$eig
```

```
# b) How many components should be extracted? Decide on the number of components considering eigenvalue

#Compute Scree plot
plot(pca_res$eig[,2], type="o", main="Scree Plot: Percentage of Variance by Component", xlab="Number of
axis(1, at=1:length(pca_res$eig[,2]), labels=1:length(pca_res$eig[,2]), las=1)
axis(2, las=1)
```




```r
# INTERPRETATION IN THE FILE: Interpretations.pdf, considering pca_res$eig and the scree plot
```

```r
# c) Interpret the loadings/correlations of variables at each dimension (3p).

pca_res$var$coord[, 1:3]
```

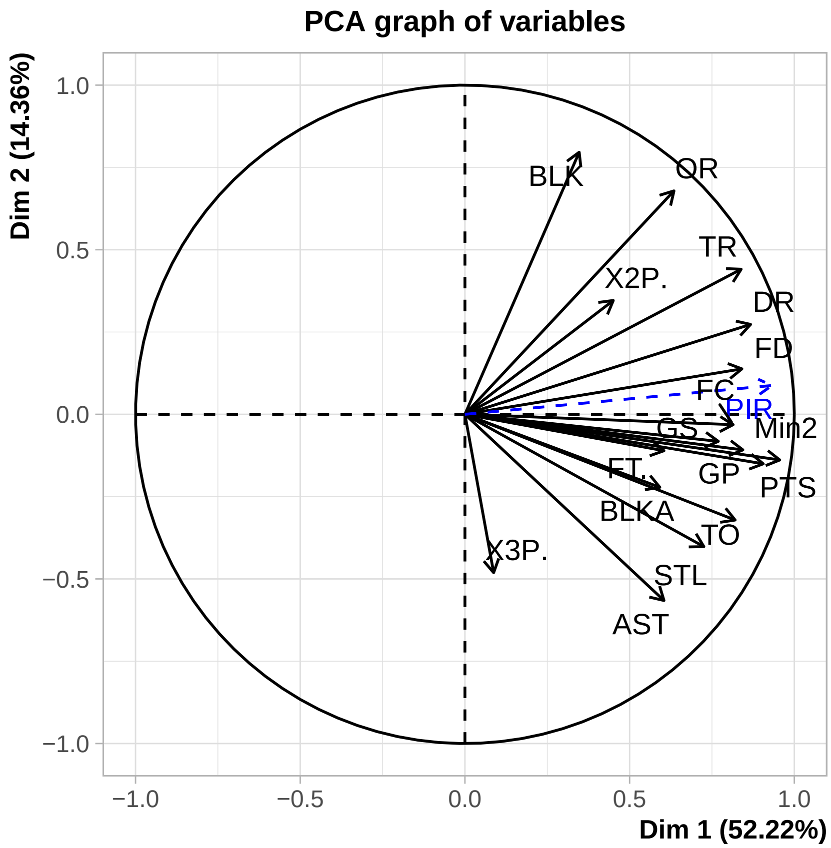
```
##            Dim.1       Dim.2       Dim.3
## GP    0.84393671 -0.10804438  0.22566679
## GS    0.76933470 -0.08208798 -0.09472409
## PTS   0.90572767 -0.15035128 -0.09308367
## X2P.  0.45039272  0.34583077  0.61213753
## X3P.  0.08717723 -0.48097648  0.14826155
## FT.   0.60436397 -0.11054957  0.61436928
## OR    0.63470909  0.67874950 -0.12507297
## DR    0.86731109  0.27317391 -0.06270111
## TR    0.83862833  0.44075406 -0.08684766
## AST   0.60438499 -0.56552953 -0.08274957
## STL   0.72547288 -0.40191934  0.04941070
## TO    0.82030853 -0.32115744 -0.19550367
## BLK   0.34688099  0.79614512 -0.08186885
## BLKA  0.59171534 -0.22162286 -0.39640555
## FC    0.81433984 -0.03140147  0.17134774
## FD    0.84087555  0.13801928 -0.24591607
## Min2  0.95586845 -0.13867651  0.01898908
```

```r
pca_res$var$cor[, 1:3]
```

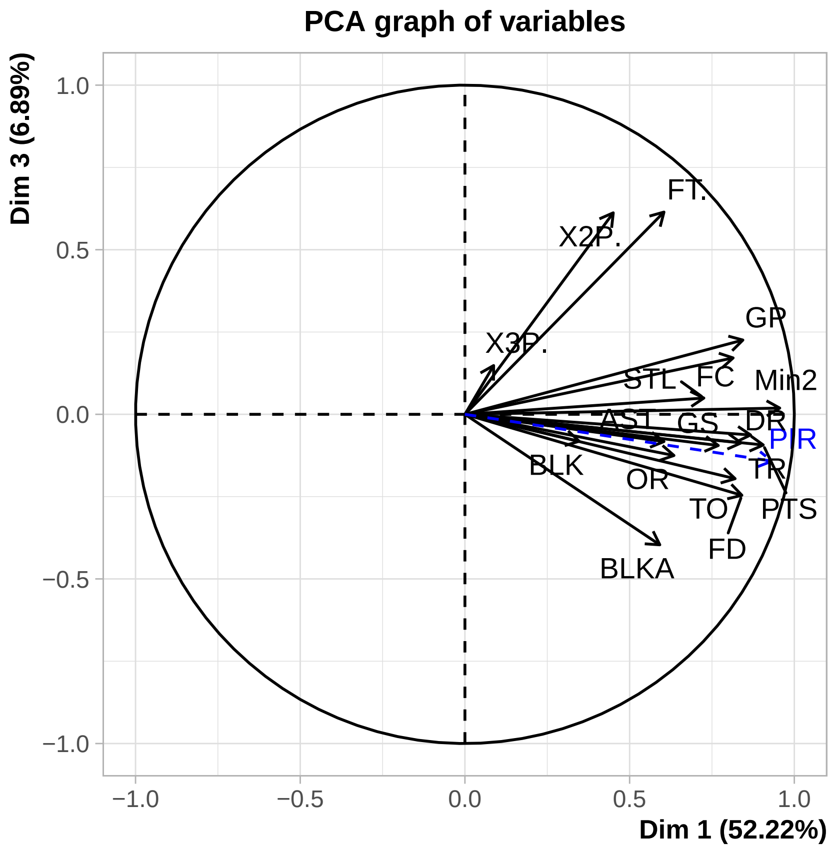
```
##           Dim.1       Dim.2       Dim.3
## GP   0.84393671 -0.10804438  0.22566679
## GS   0.76933470 -0.08208798 -0.09472409
## PTS  0.90572767 -0.15035128 -0.09308367
## X2P. 0.45039272  0.34583077  0.61213753
## X3P. 0.08717723 -0.48097648  0.14826155
## FT.  0.60436397 -0.11054957  0.61436928
## OR   0.63470909  0.67874950 -0.12507297
## DR   0.86731109  0.27317391 -0.06270111
## TR   0.83862833  0.44075406 -0.08684766
## AST  0.60438499 -0.56552953 -0.08274957
## STL  0.72547288 -0.40191934  0.04941070
## TO   0.82030853 -0.32115744 -0.19550367
## BLK  0.34688099  0.79614512 -0.08186885
## BLKA 0.59171534 -0.22162286 -0.39640555
## FC   0.81433984 -0.03140147  0.17134774
## FD   0.84087555  0.13801928 -0.24591607
## Min2 0.95586845 -0.13867651  0.01898908
```

```r
# INTERPRETATION IN THE FILE: Interpretations.pdf
```

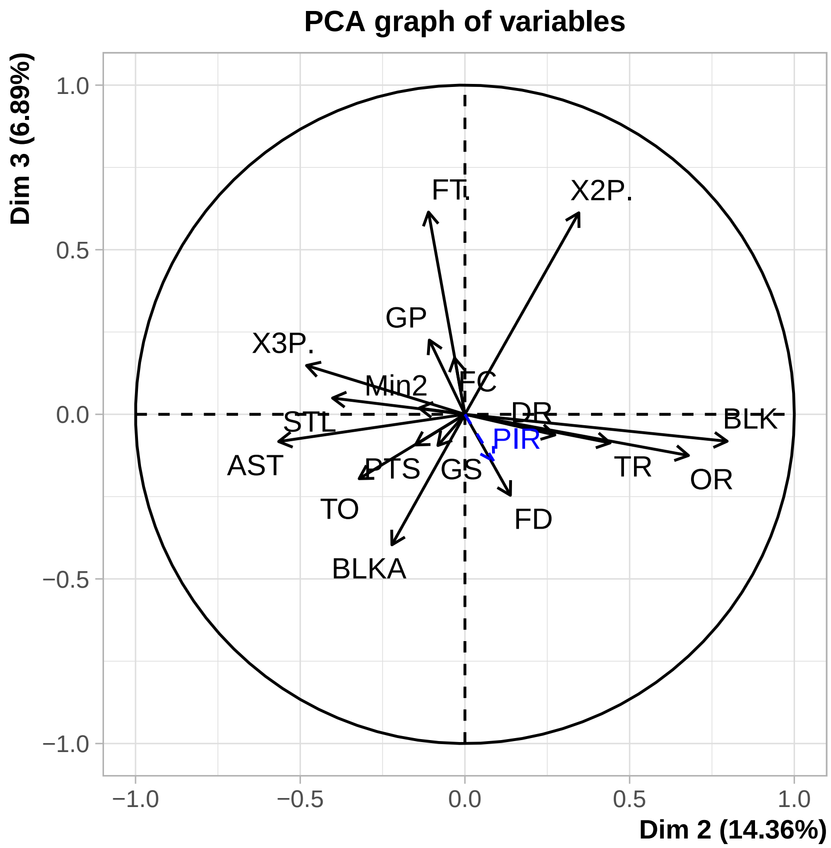
```r
# d) Use plot.PCA() function to show correlations between variables and the extracted
# dimensions. (For the variables you should use the argument choix = "var"). Plot all
# the extracted dimensions changing argument "axes".(3p)

scores <- pca_res$ind$coord

plot.PCA(pca_res, choix = "var", axes = c(1, 2))
```




```
plot.PCA(pca_res, choix = "var", axes = c(1, 3))
```



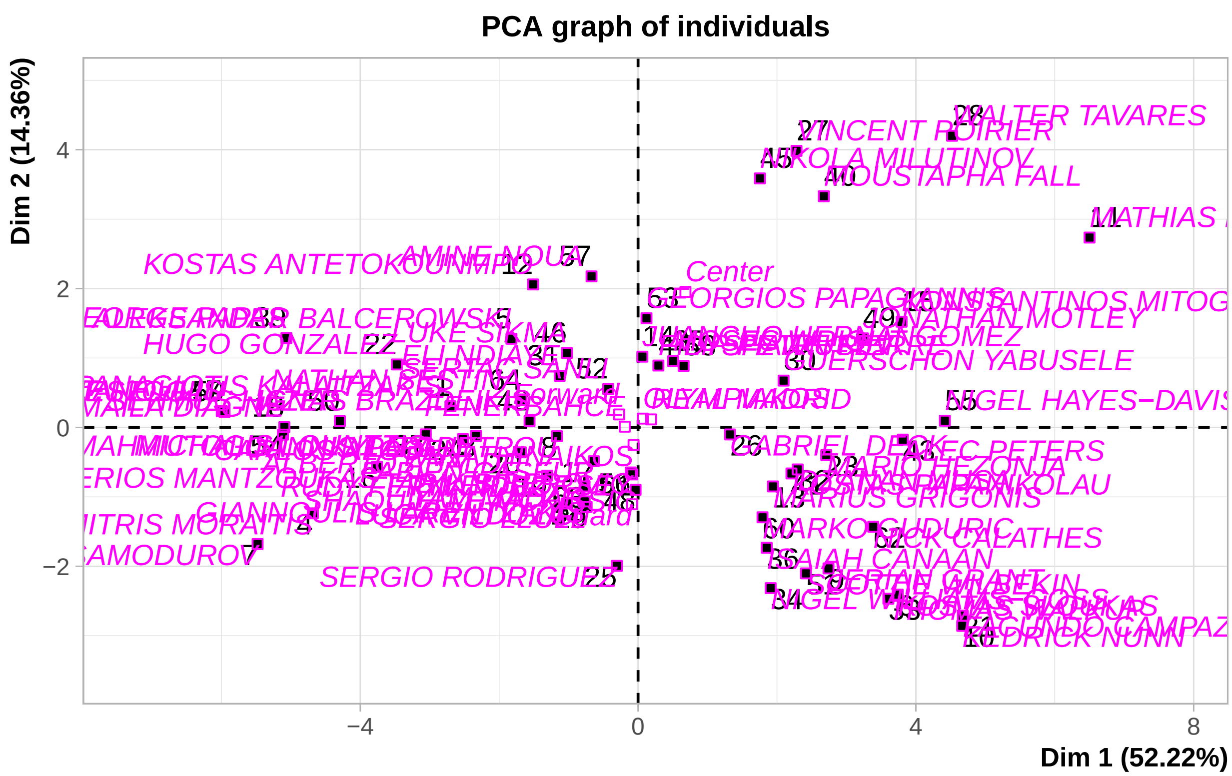

```
plot.PCA(pca_res, choix = "var", axes = c(2, 3))
```



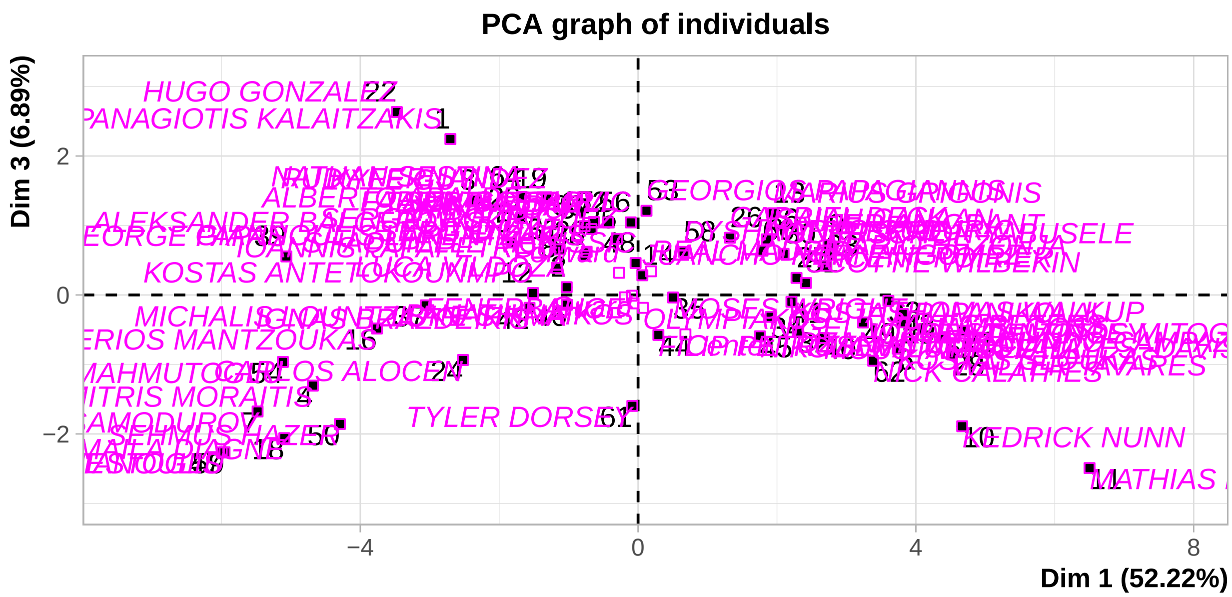

```
#plot.PCA(pca_res, choix = "var", axes = c(2, 1)) # same as 1,2
#plot.PCA(pca_res, choix = "var", axes = c(3, 1)) # same as 1,3
#plot.PCA(pca_res, choix = "var", axes = c(3, 2)) # same as 2,3
```

```
# e) Interpret variable plots. How can each dimension be named? (5p)

# INTERPRETATION IN THE FILE: Interpretations.pdf
```

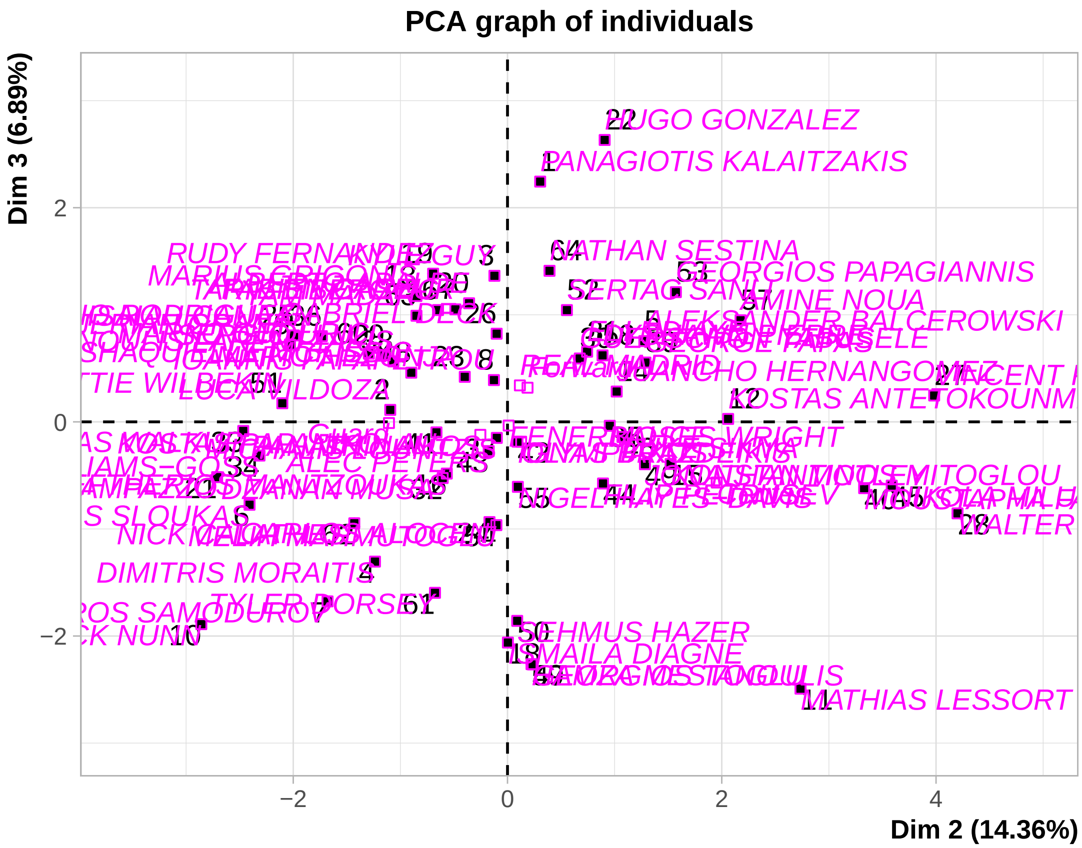
```
# f) Show individual plots for the extracted dimensions changing argumennt
# choix="ind" in plot.PCA() function. (2p)

plot.PCA(pca_res, choix = "ind", axes = c(1, 2))
```




```r
plot.PCA(pca_res, choix = "ind", axes = c(1, 3))
```




```
plot.PCA(pca_res, choix = "ind", axes = c(2, 3))
```




```r
#plot.PCA(pca_res, choix = "ind", axes = c(2, 1)) # same as 1,2
#plot.PCA(pca_res, choix = "ind", axes = c(3, 1)) # same as 1,3
#plot.PCA(pca_res, choix = "ind", axes = c(3, 2)) # same as 2,3
```

```r
# g) Interpret the individual plots. (3p)

# INTERPRETATION IN THE FILE: Interpretations.pdf (we used the plots and the
# following function to support our insights)

filter_players <- function(df, compare_players, compare_vars) {
  compare_players <- trimws(compare_players)
  #Trim white spaces to avoid mismatches
  filtered_df <- df[trimws(df$PLAYER) %in% compare_players, compare_vars, drop = FALSE]

  return(filtered_df)
}

#Compare 2 opposite players of dimension 1 (1 vs 2 dim)
filter_players(df, c("ALEXANDROS SAMODUROV", "MATHIAS LESSORT"), names(df))
```

```
##             TEAM               PLAYER POSITION GP GS  PTS X2P. X3P.  FT.  OR DR
## 7  PANATHINAIKOS ALEXANDROS SAMODUROV  Forward  1  0  3.0  0.0  100  0.0 0.0  0
## 11 PANATHINAIKOS      MATHIAS LESSORT   Center 41 29 13.9 62.6    0 60.7 2.3  4
##     TR AST STL  TO BLK BLKA  FC  FD  PIR Min2
## 7  0.0 1.0   0 0.0 0.0  0.0 0.0 0.0  3.0    1
```

```
## 11 6.3 1.4   1 1.8 0.9  0.5 2.7 6.7 19.6   29
```

```r
#Compare 2 close players of dimension 2 (1 vs 2 dim)
filter_players(df, c("MOUSTAPHA FALL", "NIKOLA MILUTINOV"), names(df))
```

```
##          TEAM           PLAYER POSITION GP GS PTS X2P. X3P.  FT.  OR  DR  TR
## 40 OLYMPIAKOS   MOUSTAPHA FALL   Center 36 34 7.2 78.9    0 43.9 1.5 3.2 4.7
## 45 OLYMPIAKOS NIKOLA MILUTINOV   Center 27  4 7.8 61.0    0 82.2 2.5 3.1 5.6
##    AST STL  TO BLK BLKA  FC  FD  PIR Min2
## 40 2.4 0.2 1.1 1.2  0.1 1.8 2.9 13.4   22
## 45 0.9 0.3 1.0 0.6  0.1 0.9 3.4 14.3   18
```

```r
#Compare 2 opposite players of dimension 2 (1 vs 2 dim)
filter_players(df, c("MATHIAS LESSORT", "FACUNDO CAMPAZZO"), names(df))
```

```
##             TEAM           PLAYER POSITION GP GS  PTS X2P. X3P.  FT.  OR  DR
## 11 PANATHINAIKOS MATHIAS LESSORT    Center 41 29 13.9 62.6  0.0 60.7 2.3 4.0
## 21   REAL MADRID FACUNDO CAMPAZZO    Guard 37 36 11.5 57.8 32.2 86.6 0.6 2.3
##     TR AST STL  TO BLK BLKA  FC  FD  PIR Min2
## 11 6.3 1.4 1.0 1.8 0.9  0.5 2.7 6.7 19.6   29
## 21 2.9 6.5 1.3 2.4 0.0  0.2 2.4 4.5 16.7   25
```

```r
#Compare 2 close players in dimension 1 and 2 (1 vs 2 dim)
filter_players(df, c("MATHIAS LESSORT ", "WALTER TAVARES"), names(df))
```

```
##             TEAM           PLAYER POSITION GP GS  PTS X2P. X3P.  FT.  OR  DR
## 11 PANATHINAIKOS MATHIAS LESSORT    Center 41 29 13.9 62.6    0 60.7 2.3 4.0
## 28   REAL MADRID   WALTER TAVARES   Center 34 33  9.4 60.7    0 73.8 2.3 4.2
##     TR AST STL  TO BLK BLKA  FC  FD  PIR Min2
## 11 6.3 1.4 1.0 1.8 0.9  0.5 2.7 6.7 19.6   29
## 28 6.5 1.4 0.7 1.7 1.5  0.1 2.6 3.0 14.9   22
```

```r
#Compare 2 close players in dimension 3 (1 vs 3 dim)
filter_players(df, c("HUGO GONZALEZ", "PANAGIOTIS KALAITZAKIS"), names(df))
```

```
##             TEAM                 PLAYER POSITION GP GS PTS X2P. X3P. FT.  OR
## 1  PANATHINAIKOS PANAGIOTIS KALAITZAKIS     Guard 30  0 2.1   69   25 100 0.3
## 22   REAL MADRID          HUGO GONZALEZ  Forward  6  1 0.7  100    0 100 0.0
##     DR  TR AST STL  TO BLK BLKA  FC  FD  PIR Min2
## 1  0.6 0.9 0.2 0.2 0.2   0    0 0.8 0.4  2.1    5
## 22 0.2 0.2 0.3 0.0 0.3   0    0 1.0 0.5 -0.3    4
```

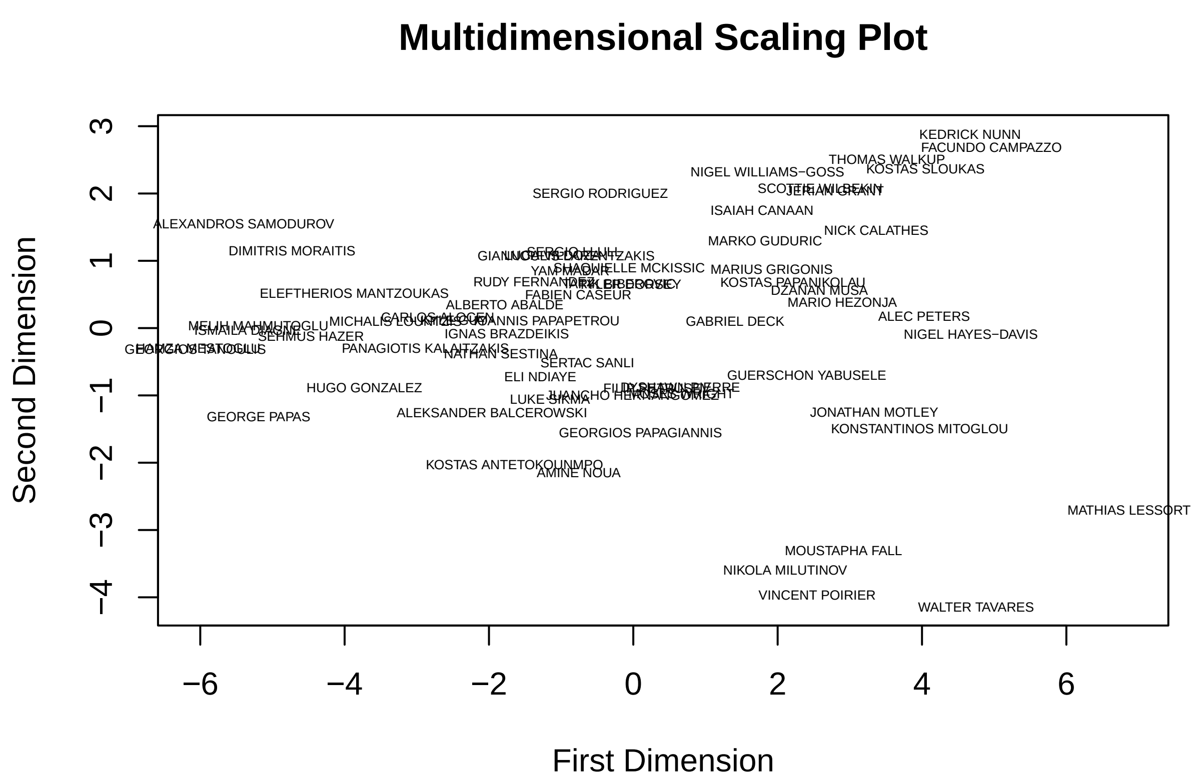
```r
#Compare 2 far players in dimension 3 (1 vs 3 dim)
filter_players(df, c("MATHIAS LESSORT", "HUGO GONZALEZ"), names(df))
```

```
##             TEAM           PLAYER POSITION GP GS  PTS  X2P. X3P.   FT.  OR  DR
## 11 PANATHINAIKOS MATHIAS LESSORT    Center 41 29 13.9  62.6    0  60.7 2.3 4.0
## 22   REAL MADRID    HUGO GONZALEZ  Forward  6  1  0.7 100.0    0 100.0 0.0 0.2
##     TR AST STL  TO BLK BLKA  FC  FD  PIR Min2
## 11 6.3 1.4   1 1.8 0.9  0.5 2.7 6.7 19.6   29
## 22 0.2 0.3   0 0.3 0.0  0.0 1.0 0.5 -0.3    4
```

3. Application of MDS

```r
# a) Apply metric MDS using Euclidean distance on scaled numerical variables. (2p)

player_names <- df$PLAYER
scaled_numerical_df <- as.data.frame(scale(df[sapply(df, is.numeric)]))
rownames(scaled_numerical_df) <- player_names
dist <- dist(scaled_numerical_df, method = "euclidean");

res_mds <- cmdscale(dist,eig=TRUE)
```

```r
# b) Plot the data using the points on the first two coordinates using players
# names as label. (2p)

plot(res_mds$points[,1], res_mds$points[,2], main = "Multidimensional Scaling Plot",
     type = "n", xlab = "First Dimension", ylab = "Second Dimension");
text(res_mds$points[,1], res_mds$points[,2], labels(dist), cex = 0.4, xpd = TRUE)
```



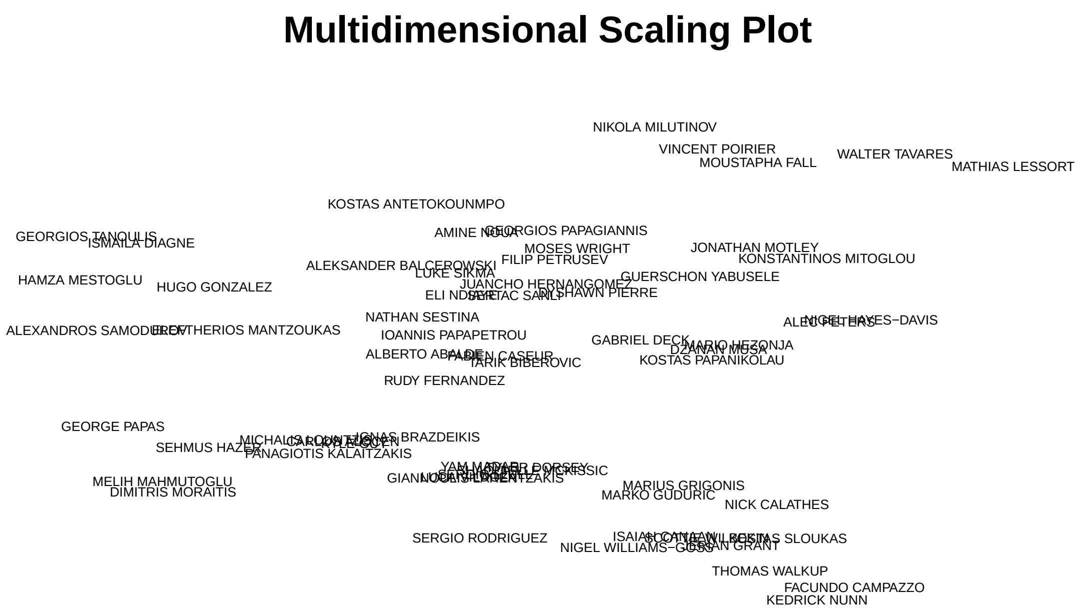

```r
# c) Interpret the plot.(3p)

# INTERPRETATION IN THE FILE: Interpretations.pdf
```

```r
# d) Calculate gower distance including variable "POSITION" to the data matrix. (3p)
```

```r
library(cluster)

new_df <- as.data.frame(scale(df[sapply(df, is.numeric)]))
new_df$POSITION <- df$POSITION
rownames(new_df) <- player_names
gowerdist <- daisy(new_df,metric = "gower")
```

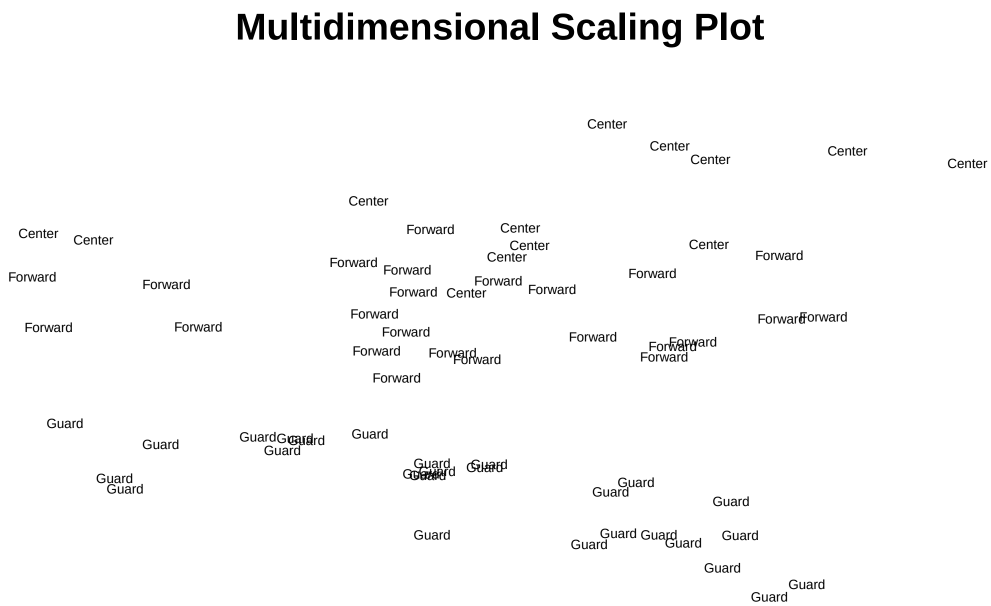
```r
# e) Apply metric MDS on gower distance matrix. (2p)

res_mds <- cmdscale(gowerdist,eig=TRUE)
```

```r
# f) Plot individual plots on the first two coordinates (2p).

plot(res_mds$points[,1], res_mds$points[,2], main = "Multidimensional Scaling Plot",
     type = "n", xlab = "", ylab = "", axes = FALSE,); text(res_mds$points[,1],
                                                            res_mds$points[,2],
                                                            labels(gowerdist),
                                                            cex = 0.4, xpd = TRUE)
```



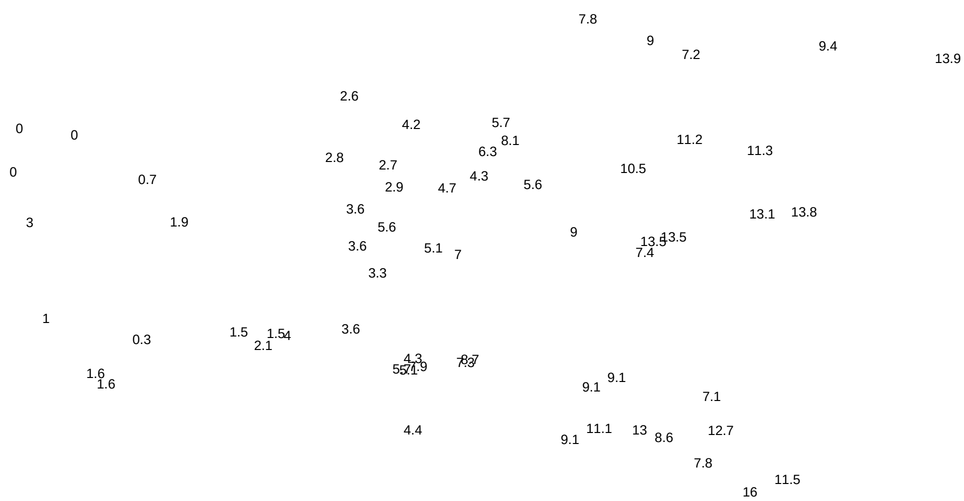

```r
# g) Use different categorical and numerical variables as labels so as to explain
# clusters that are constructed. (5p)

# Plot individuals by position
plot(res_mds$points[,1], res_mds$points[,2], main = "Multidimensional Scaling Plot",
```

```
     type = "n", xlab = "", ylab = "", axes = FALSE,); text(res_mds$points[,1],
                                                            res_mds$points[,2],
                                                            labels = df$POSITION,
                                                            cex = 0.4, xpd = TRUE)
```

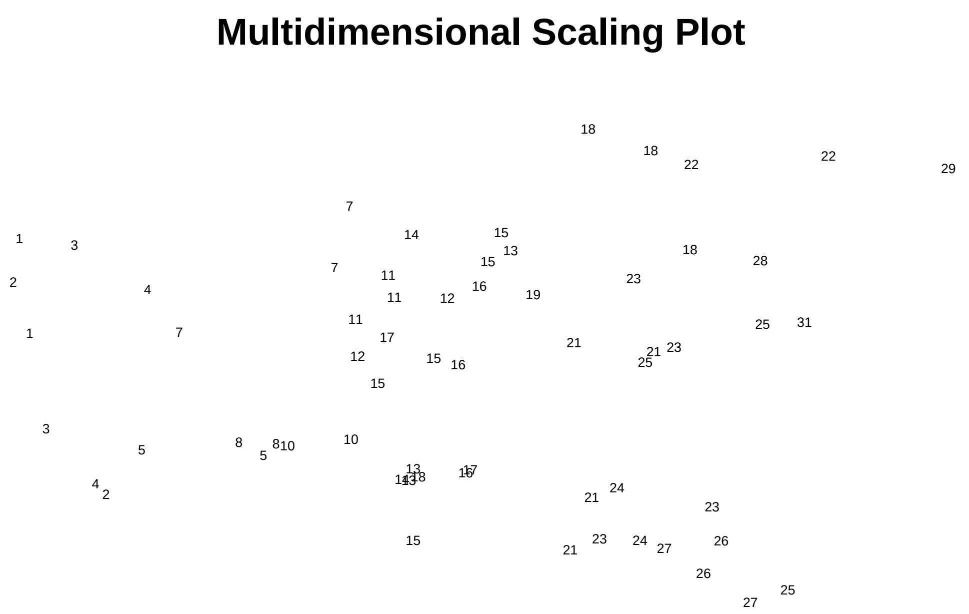



```
# Plot individuals by points
plot(res_mds$points[,1], res_mds$points[,2], main = "Multidimensional Scaling Plot",
     type = "n", xlab = "", ylab = "", axes = FALSE,); text(res_mds$points[,1],
                                                            res_mds$points[,2],
                                                            labels = df$PTS,
                                                            cex = 0.4, xpd = TRUE)
```

Multidimensional Scaling Plot

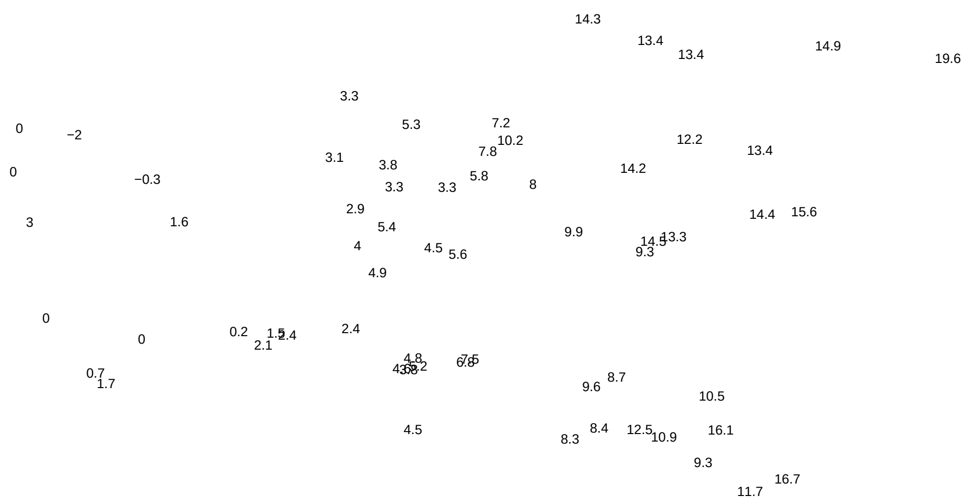



```
# Plot individuals by minutes played
plot(res_mds$points[,1], res_mds$points[,2], main = "Multidimensional Scaling Plot",
     type = "n", xlab = "", ylab = "", axes = FALSE,); text(res_mds$points[,1],
                                                              res_mds$points[,2],
                                                              labels = df$Min2,
                                                              cex = 0.4, xpd = TRUE)
```




```r
# Plot individuals by PERSONAL INDEX RATING
plot(res_mds$points[,1], res_mds$points[,2], main = "Multidimensional Scaling Plot",
     type = "n", xlab = "", ylab = "", axes = FALSE,); text(res_mds$points[,1],
                                                            res_mds$points[,2],
                                                            labels = df$PIR,
                                                            cex = 0.4, xpd = TRUE)
```

Multidimensional Scaling Plot



```
# h) Which MDS do you think better group the individuals? Why? (3p)

# INTERPRETATION IN THE FILE: Interpretations.pdf
```