Clustering NBA players based on perfomance

Alfonso Marino

2024-04-08

The NBA has a rich history spanning decades, with players of various skills, styles, and physical attributes gracing the courts. In this project, we leverage data spanning from 1950 to 2021 to gain deeper insights into player performance. Unlike traditional approaches that categorize players based solely on their positions (such as point guards, shooting guards, etc.), we employ cluster analysis to group players based on their overall performance metrics. This allows us to uncover hidden similarities and differences among players that may not be evident when considering positions alone.

Data Preparation

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(measurements)
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
nba = read.csv("/Users/alfonsomarino/Desktop/Progetti/nba/seasons_stats.csv")
player_stat = read.csv("/Users/alfonsomarino/Desktop/Progetti/nba/player_data.csv")
```

```
nba %>%
  glimpse()
```

```
## Rows: 28,057
## Columns: 51
## $ X      <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1~
```

```
## $ Year   <int> 1950, 1950, 1950, 1950, 1950, 1950, 1950, 1950, 1950, 1950, 195~
## $ Player <chr> "Curly Armstrong", "Cliff Barker", "Leo Barnhorst", "Ed Bartels~
## $ Pos    <chr> "G-F", "SG", "SF", "F", "F", "F", "G", "G-F", "F-C", "F-C", "F-~
## $ Age    <int> 31, 29, 25, 24, 24, 24, 22, 23, 28, 28, 28, 25, 22, 22, 24, 27,~
## $ Tm     <chr> "FTW", "INO", "CHS", "TOT", "DNN", "NYK", "INO", "TRI", "TOT", ~
## $ G      <int> 63, 49, 67, 15, 13, 2, 60, 3, 65, 36, 29, 57, 60, 59, 62, 61, 4~
## $ GS     <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ MP     <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ FG     <int> 144, 102, 174, 22, 21, 1, 340, 5, 226, 125, 101, 80, 88, 204, 2~
## $ FGA    <int> 516, 274, 499, 86, 82, 4, 936, 16, 813, 435, 378, 248, 305, 600~
## $ FG.    <dbl> 0.279, 0.372, 0.349, 0.256, 0.256, 0.250, 0.363, 0.313, 0.278, ~
## $ X3P    <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ X3PA   <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ X3P.   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ X2P    <int> 144, 102, 174, 22, 21, 1, 340, 5, 226, 125, 101, 80, 88, 204, 2~
## $ X2PA   <int> 516, 274, 499, 86, 82, 4, 936, 16, 813, 435, 378, 248, 305, 600~
## $ X2P.   <dbl> 0.279, 0.372, 0.349, 0.256, 0.256, 0.250, 0.363, 0.313, 0.278, ~
## $ eFG.   <dbl> 0.279, 0.372, 0.349, 0.256, 0.256, 0.250, 0.363, 0.313, 0.278, ~
## $ FT     <int> 170, 75, 90, 19, 17, 2, 215, 0, 209, 132, 77, 82, 78, 204, 240,~
## $ FTA    <int> 241, 106, 129, 34, 31, 3, 282, 5, 321, 209, 112, 131, 117, 267,~
## $ FT.    <dbl> 0.705, 0.708, 0.698, 0.559, 0.548, 0.667, 0.762, 0.000, 0.651, ~
## $ ORB    <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ DRB    <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ TRB    <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ AST    <int> 176, 109, 140, 20, 20, 0, 233, 2, 163, 75, 88, 46, 40, 95, 137,~
## $ STL    <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ BLK    <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ TOV    <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ PF     <int> 217, 99, 192, 29, 27, 2, 132, 6, 273, 140, 133, 97, 111, 203, 2~
## $ PTS    <int> 458, 279, 438, 63, 59, 4, 895, 10, 661, 382, 279, 242, 254, 612~
## $ PER    <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ TS.    <dbl> 0.368, 0.435, 0.394, 0.312, 0.308, 0.376, 0.422, 0.275, 0.346, ~
## $ X3PAr  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ FTr    <dbl> 0.467, 0.387, 0.259, 0.395, 0.378, 0.750, 0.301, 0.313, 0.395, ~
## $ ORB.   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ DRB.   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ TRB.   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ AST.   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ STL.   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ BLK.   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ TOV.   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ USG.   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ OWS    <dbl> -0.1, 1.6, 0.9, -0.5, -0.5, 0.0, 3.6, -0.1, -2.2, -0.7, -1.5, 0~
## $ DWS    <dbl> 3.6, 0.6, 2.8, -0.1, -0.1, 0.0, 1.2, 0.0, 5.0, 2.2, 2.8, 1.3, 1~
## $ WS     <dbl> 3.5, 2.2, 3.6, -0.6, -0.6, 0.0, 4.8, -0.1, 2.8, 1.5, 1.3, 1.8, ~
## $ WS.48  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ OBPM   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ DBPM   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ BPM    <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ VORP   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
```

```
player_stat %>%
  glimpse()
```

```
## Rows: 4,979
## Columns: 8
## $ Player     <chr> "Alaa Abdelnaby", "Zaid Abdul-Aziz", "Kareem Abdul-Jabbar*"~
## $ From       <int> 1991, 1969, 1970, 1991, 1998, 1997, 1977, 1957, 1947, 2017,~
## $ To         <int> 1995, 1978, 1989, 2001, 2003, 2008, 1981, 1957, 1948, 2019,~
## $ Pos        <chr> "F-C", "C-F", "C", "G", "F", "F", "F", "G", "F", "G-F", "F"~
## $ Ht         <chr> "6-10", "6-9", "7-2", "6-1", "6-6", "6-9", "6-7", "6-3", "6~
## $ Wt         <int> 240, 235, 225, 162, 223, 225, 220, 180, 195, 200, 225, 185,~
## $ Birth.Date <chr> "June 24 1968", "April 7 1946", "April 16 1947", "March 9 1~
## $ Colleges   <chr> "Duke", "Iowa State", "UCLA", "LSU", "Michigan San Jose Sta~
```

When analyzing the null values present in *player_stat*, note how there are only 5 records corresponding to the weight column, which are removed because their presence historically has been marginal.

```
player_stat %>%
  select(everything()) %>%
  summarise_all(list(~sum(is.na(.))))
```

```
##   Player From To Pos Ht Wt Birth.Date Colleges
## 1      0    0  0   0  0  5          0        0
```

```
player_stat %>%
  filter(is.na(Wt) == T)
```

```
##            Player From   To Pos   Ht Wt       Birth.Date
## 1        Dick Lee 1968 1968   F  6-6 NA
## 2 Murray Mitchell 1950 1950   C  6-6 NA    March 19 1923
## 3      Paul Nolen 1954 1954   C 6-10 NA September 3 1929
## 4      Ray Wertis 1947 1948   G 5-11 NA     July 30 1923
## 5        Bob Wood 1950 1950   G 5-10 NA   October 7 1921
##                       Colleges
## 1                   Washington
## 2 Sam Houston State University
## 3                   Texas Tech
## 4                   St. John's
## 5            Northern Illinois
```

```
player_stat = player_stat %>%
  drop_na(Wt)
```

Any duplicates are removed.

```
player_stat[duplicated(player_stat),]
```

```
##                Player From   To Pos   Ht  Wt        Birth.Date
## 1320    Ledell Eackles 1989 1998 G-F  6-5 220  November 24 1966
## 1321        Jim Eakins 1969 1978   C 6-11 215       May 24 1946
## 1322         Acie Earl 1994 1997 F-C 6-10 240      June 23 1970
## 1323          Ed Earle 1954 1954   F  6-3 190     April 28 1927
## 1324  Cleanthony Early 2015 2016   F  6-8 210     April 17 1991
## 1325       Penny Early 1969 1969   G  5-3 114       May 30 1943
```

```
## 1326        Mark Eaton 1983 1993   C  7-4 275   January 24 1957
## 1327       Jerry Eaves 1983 1987   G  6-4 180   February 8 1959
## 1328      Devin Ebanks 2011 2013   F  6-9 215   October 28 1989
## 1329        Bill Ebben 1958 1958   G  6-4 190    October 7 1935
## 1330       Al Eberhard 1975 1978   F  6-6 225       May 10 1952
## 1331         Ndudi Ebi 2004 2005   F  6-9 200      June 18 1984
## 1332         Roy Ebron 1974 1974   C  6-9 220    August 31 1951
## 1333   Jaime Echenique 2022 2022   C 6-11 258     April 27 1997
## 1334      Jarell Eddie 2016 2018 G-F  6-7 218   October 30 1991
## 1335     Patrick Eddie 1992 1992   C 6-11 240  December 27 1967
## 1336     Dike Eddleman 1950 1953 F-G  6-3 189  December 27 1922
## 1337     Kenton Edelin 1985 1985   F  6-8 205       May 24 1962
## 1338      Charles Edge 1974 1975   F  6-6 210  February 23 1950
## 1339     Bobby Edmonds 1968 1970   F  6-6 220      March 8 1941
## 1340    Keith Edmonson 1983 1984   G  6-5 195 September 28 1960
## 1341        Tyus Edney 1996 2001   G 5-10 152  February 14 1973
## 1342   Anthony Edwards 2021 2022   G  6-4 225     August 5 2001
## 1343      Bill Edwards 1994 1994   F  6-8 215 September 22 1971
## 1344      Blue Edwards 1990 1999 G-F  6-4 200   October 31 1965
## 1345    Carsen Edwards 2020 2022   G 5-11 200     March 12 1998
## 1346   Corsley Edwards 2005 2005   F  6-9 275      March 5 1979
## 1347      Doug Edwards 1994 1996   F  6-7 220   January 21 1971
## 1348  Franklin Edwards 1982 1988   G  6-1 170   February 2 1959
## 1349     James Edwards 1978 1996 C-F  7-0 225  November 22 1955
## 1350       Jay Edwards 1990 1990   G  6-4 185    January 3 1969
## 1351      John Edwards 2005 2006   C  7-0 275      July 31 1981
## 1352   Kessler Edwards 2022 2022   F  6-8 215     August 9 2000
## 1353     Kevin Edwards 1989 2001   G  6-3 190   October 30 1965
## 1354       Rob Edwards 2022 2022   G  6-5 205   January 20 1997
## 1355     Shane Edwards 2014 2014   F  6-7 220       May 31 1987
## 1356     Vince Edwards 2019 2019   F  6-8 225      April 5 1996
## 1357       Johnny Egan 1962 1972   G 5-11 180   January 31 1939
## 1358  Lonnie Eggleston 1949 1949   G  6-0 170       June 8 1918
## 1359      Bulbs Ehlers 1948 1949 F-G  6-3 198     March 10 1923
## 1360        Craig Ehlo 1984 1997 G-F  6-6 180    August 11 1961
## 1361    Rich Eichhorst 1962 1962   G  6-3 200   October 21 1933
## 1362     Howard Eisley 1995 2006   G  6-2 177   December 4 1972
## 1363     Obinna Ekezie 2000 2005 F-C  6-9 270    August 22 1975
## 1364    Khalid El-Amin 2001 2001   G 5-10 200     April 25 1979
## 1365       Don Eliason 1947 1947   F  6-2 210      July 24 1918
## 1366        Mario Elie 1991 2001 G-F  6-5 210  November 26 1963
## 1367         CJ Elleby 2021 2022   F  6-6 200      June 16 2000
## 1368      Ray Ellefson 1949 1951   C  6-8 230  November 18 1922
## 1369    Henry Ellenson 2017 2021   F 6-10 240   January 13 1997
## 1370   Wayne Ellington 2010 2022   G  6-4 207  November 29 1987
## 1371       Bob Elliott 1979 1981 C-F  6-9 225    August 18 1955
## 1372      Sean Elliott 1990 2001   F  6-8 205   February 2 1968
## 1373          Bo Ellis 1978 1980   F  6-9 197     August 8 1954
## 1374         Boo Ellis 1959 1960   F  6-5 185  February 11 1936
## 1375        Dale Ellis 1984 2000 G-F  6-7 205     August 6 1960
## 1376      Harold Ellis 1994 1998   G  6-5 200    October 7 1970
## 1377         Joe Ellis 1967 1974 F-G  6-6 175        May 3 1944
## 1378    LaPhonso Ellis 1993 2003   F  6-8 240        May 5 1970
## 1379       LeRon Ellis 1992 1996 F-C  6-9 225     April 28 1969
```

```
## 1380       Leroy Ellis 1963 1976 C-F 6-10 210     March 10 1940
## 1381       Monta Ellis 2006 2017   G  6-3 185   October 26 1985
## 1382    Pervis Ellison 1990 2001 F-C  6-9 210      April 3 1967
## 1383        Len Elmore 1975 1984 C-F  6-9 220     March 28 1952
## 1384   Francisco Elson 2004 2012   C  7-0 235  February 28 1976
## 1385    Darrell Elston 1975 1977   G  6-4 190    August 15 1952
## 1386        Melvin Ely 2003 2014   C 6-10 260        May 2 1978
## 1387       Joel Embiid 2017 2022   C  7-0 280     March 16 1994
## 1388      Wayne Embry* 1959 1969 C-F  6-8 240     March 26 1937
## 1389      Andre Emmett 2005 2012   F  6-5 230    August 27 1982
## 1390       Ned Endress 1947 1947 F-G  6-2 200      March 2 1918
## 1391   Wayne Engelstad 1989 1989   F  6-8 245   December 6 1965
## 1392      Chris Engler 1983 1988   C 6-11 245      March 1 1959
## 1393      A.J. English 1991 1992   G  6-3 175      July 11 1967
## 1394     Alex English* 1977 1991   F  6-7 190    January 5 1954
## 1395    Claude English 1971 1971   F  6-4 185  December 26 1946
## 1396     Jo Jo English 1993 1995   G  6-4 195   February 4 1970
## 1397       Kim English 2013 2013   G  6-6 200 September 24 1988
## 1398     Scott English 1973 1975   F  6-6 205   October 20 1950
## 1399      Gene Englund 1950 1950 F-C  6-5 205   October 21 1917
## 1400   James Ennis III 2015 2022   F  6-6 215       July 1 1990
## 1401       Tyler Ennis 2015 2018   G  6-3 194    August 24 1994
## 1402          Ray Epps 1979 1979   F  6-6 195    August 20 1956
## 1403       Semih Erden 2011 2012   C  7-0 240      July 28 1986
## 1404          Bo Erias 1958 1958   F  6-3 220      July 30 1932
## 1405    Keith Erickson 1966 1977 F-G  6-5 195     April 19 1944
## 1406    Julius Erving* 1972 1987 F-G  6-7 210  February 22 1950
## 1407    Evan Eschmeyer 2000 2003   C 6-11 255       May 30 1975
## 1408     Jack Eskridge 1949 1949 C-F  6-5 200   January 21 1924
## 1409 Vincenzo Esposito 1996 1996   G  6-3 198      March 1 1969
## 1410      Drew Eubanks 2019 2022   F  6-9 245   February 1 1997
## 1411       Billy Evans 1970 1970   G  6-0 170      March 3 1947
## 1412         Bob Evans 1950 1950   G  6-2 175       May 31 1925
## 1413       Brian Evans 1997 1999   F  6-8 220 September 13 1973
## 1414        Earl Evans 1980 1980   F  6-8 202  November 11 1955
## 1415       Jacob Evans 2019 2020 G-F  6-4 210      June 18 1997
## 1416       Jawun Evans 2018 2019   G  6-0 185      July 26 1996
## 1417      Jeremy Evans 2011 2018   F  6-9 200   October 24 1987
## 1418     Maurice Evans 2002 2012   G  6-5 220   November 8 1978
## 1419        Mike Evans 1980 1988   G  6-1 170     April 19 1955
## 1420      Reggie Evans 2003 2015   F  6-8 245       May 18 1980
## 1421      Tyreke Evans 2010 2019 G-F  6-6 220 September 19 1989
## 1422      Daniel Ewing 2006 2007   G  6-3 185     March 26 1983
## 1423    Patrick Ewing* 1986 2002 C-F  7-0 240     August 5 1962
## 1424     Patrick Ewing 2011 2011   F  6-8 235       May 20 1984
## 1425        Dante Exum 2015 2021   G  6-5 214      July 13 1995
## 1426  Christian Eyenga 2011 2012   F  6-5 210      June 22 1989
## 1427     Festus Ezeli 2013 2016   C 6-11 255   October 21 1989
## 1428    Johnny Ezersky 1948 1950 F-G  6-3 175     March 21 1922
##                                    Colleges
## 1320                           New Orleans
## 1321                                   BYU
## 1322                                  Iowa
## 1323                         Loyola Chicago
```

```
## 1324                            Wichita State
## 1325
## 1326                                    UCLA
## 1327                               Louisville
## 1328                            West Virginia
## 1329                            Detroit Mercy
## 1330                                 Missouri
## 1331
## 1332                                Louisiana
## 1333         Trinity Valley CC Wichita State
## 1334                            Virginia Tech
## 1335       Arkansas State University Ole Miss
## 1336                                Illinois
## 1337                                Virginia
## 1338                     LeMoyne-Owen College
## 1339                          Tennessee State
## 1340                                  Purdue
## 1341                                    UCLA
## 1342                                 Georgia
## 1343                  Wright State University
## 1344                East Carolina University
## 1345                                  Purdue
## 1346     Central Connecticut State University
## 1347                              Florida State
## 1348              Cleveland State University
## 1349                                Washington
## 1350                                   Indiana
## 1351                    Kent State University
## 1352                                Pepperdine
## 1353                                   DePaul
## 1354 Cleveland State University Arizona State
## 1355                              Little Rock
## 1356                                  Purdue
## 1357                              Providence
## 1358                            Oklahoma State
## 1359                                  Purdue
## 1360                          Washington State
## 1361                         SE Missouri State
## 1362                            Boston College
## 1363                                 Maryland
## 1364                                    UConn
## 1365                          Hamline University
## 1366           American International College
## 1367                          Washington State
## 1368                West Texas A&M University
## 1369                                Marquette
## 1370                                      UNC
## 1371                                  Arizona
## 1372                                  Arizona
## 1373                                Marquette
## 1374                       Niagara University
## 1375                                Tennessee
## 1376                        Morehouse College
## 1377                            San Francisco
```

```
## 1378                            Notre Dame
## 1379                       Kentucky Syracuse
## 1380                              St. John's
## 1381
## 1382                              Louisville
## 1383                                Maryland
## 1384                              California
## 1385                                     UNC
## 1386                            Fresno State
## 1387                                  Kansas
## 1388                          Miami University
## 1389                              Texas Tech
## 1390                      University of Akron
## 1391          University of California Irvine
## 1392                      Minnesota Wyoming
## 1393            Virginia Union University
## 1394                           South Carolina
## 1395                             Rhode Island
## 1396                           South Carolina
## 1397                                 Missouri
## 1398                           Texas-El Paso
## 1399                               Wisconsin
## 1400                     Cal State Long Beach
## 1401                                 Syracuse
## 1402                           Norfolk State
## 1403
## 1404                         Niagara University
## 1405                                      UCLA
## 1406                                    UMass
## 1407                              Northwestern
## 1408                                   Kansas
## 1409
## 1410                              Oregon State
## 1411                            Boston College
## 1412                                    Butler
## 1413                                  Indiana
## 1414                                USC UNLV
## 1415                               Cincinnati
## 1416                            Oklahoma State
## 1417                          Western Kentucky
## 1418                        Wichita State Texas
## 1419                              Kansas State
## 1420                                      Iowa
## 1421                                   Memphis
## 1422                                      Duke
## 1423                                Georgetown
## 1424                         Indiana Georgetown
## 1425
## 1426
## 1427                              Vanderbilt
## 1428                            Rhode Island
```

```
player_stat = player_stat %>%
  distinct(Player, .keep_all = T)
```

The columns for weight and height are measured in pounds and feet, respectively; for convenience of use we convert the units to kilograms and centimeters.

```
convertHt <- function(x) {
  heights <- as.character(x)
  heights_split <- strsplit(heights, "-")
  feet <- as.numeric(sapply(heights_split, `[`, 1))
  inches <- as.numeric(sapply(heights_split, `[`, 2))
  heights_cm <- round(conv_unit(feet, "ft", "cm") + conv_unit(inches, "inch", "cm"), 0)

  return(heights_cm)
}

player_stat <- player_stat %>%
  rowwise() %>%
  mutate(Ht = convertHt(Ht), Wt = round(conv_unit(Wt, "lbs", "kg")))

player_stat %>%
  select(c(Wt, Ht)) %>%
head(10)
```

```
## # A tibble: 10 x 2
## # Rowwise:
##       Wt    Ht
##    <dbl> <dbl>
##  1   109   208
##  2   107   206
##  3   102   218
##  4    73   185
##  5   101   198
##  6   102   206
##  7   100   201
##  8    82   190
##  9    88   190
## 10    91   198
```

Regarding the *nba* dataset, we remove the rows for the year 2022 because it is the same as the year 2021.

```
nba = nba %>%
  filter(Year<2022)
```

Some players' names are flanked by an asterisk to symbolize their presence in the Hall of Fame. For convenience of use, we remove the asterisks and add an appropriate column.

```
nba %>%
  filter(str_detect(nba$Player, ".*\\*$")) %>%
  select(Player) %>%
  head(10)
```

```
##              Player
## 1         Al Cervi*
## 2       Bob Davies*
```

```
## 3        Joe Fulks*
## 4   Harry Gallatin*
## 5      Alex Hannum*
## 6      Red Holzman*
## 7  Buddy Jeannette*
## 8      Ed Macauley*
## 9    Slater Martin*
## 10    Dick McGuire*
```

```
nba = nba %>%
  mutate(HallOfFame = if_else(str_detect(Player, ".*\\*$"), "Yes", "No"))

nba$Player = gsub("\\*$", "", nba$Player)
player_stat$Player <- gsub("\\*$", "", player_stat$Player)
```

Per-game statistics and FTr are added.

```
nba = nba %>%
  mutate(MpG = round(MP/G,3), PpG = round(PTS/G,3), ApG = round(AST/G,3),
         RpG = round(TRB/G,3), TOpG = round(TOV/G,3), BpG = round(BLK/G,3),
         SpG = round(STL/G,3),FpG = round(PF/G, 3), .before = 9)

nba = nba %>%
  mutate(FTr = round(FT/FGA,3), .before = 30)
```

Positions are redefined.

```
nba <- nba %>%
  mutate(Pos = case_when(
    Pos == "PF-C" ~ "PF",
    Pos == "C-F" ~ "C",
    Pos == "SF-SG" ~ "SF",
    Pos == "C-PF" ~ "C",
    Pos == "SG-SF" ~ "SG",
    Pos == "PF-SF" ~ "PF",
    Pos == "SF-PF" ~ "SF",
    Pos == "SG-PG" ~ "SG",
    Pos == "SF-PG" ~ "SF",
    Pos == "C-SF" ~ "C",
    Pos == "PG-SG" ~ "PG",
    Pos == "PG-SF" ~ "PG",
    Pos == "SG-PF" ~ "SG",
    Pos == "SF-C" ~ "SF",
    Pos == "F-C" ~ "PF",
    Pos == "F-G" ~ "SF",
    Pos == "G-F" ~ "SF",
    Pos == "F" ~ "PF",
    Pos == "G" ~ "SG",
    TRUE ~ Pos
  ))

table(nba$Pos)
```

```
##
##    C   PF   PG   SF   SG
## 5351 5786 5194 5372 5649
```

At this point the two datasets can be merged to get a comprehensive overview of the information. In addition to that, we also extract two datasets namely *nba_withTOT* and *nba_performance*. The former excludes partial statistics for those players who changed teams in the middle of the season; while the latter is filtered by minutes played.

```
player_stat = player_stat %>%
  select(c("Player","Ht", "Wt"))
player_stat = as.data.frame(player_stat)

nba = left_join(nba, player_stat, by = "Player", relationship = "many-to-many")

nba = nba %>%
  select(Year:Tm, HallOfFame:Wt, everything())

nba = nba %>%
  distinct(Player,Year, Tm, Age, .keep_all = T)

nba_withTOT = nba %>%
  group_by(Year, Player) %>%
  mutate(count_tot = sum(Tm == "TOT")) %>%
  filter(count_tot == 0 | (count_tot > 0 & Tm == "TOT")) %>%
  select(-count_tot)

nba_withTOT = as.data.frame(nba_withTOT)

nba_performance = nba_withTOT %>%
  filter(MpG > mean(MpG, na.rm = T) & Year > 1999)

nba_performance = as.data.frame(nba_performance)
```

The full dataset contains several null values since not all statistics date back to the same historical period, so from time to time we are going to handle them without removing them. Also in the column for colleges, there are also blank spaces, these may be due to lack of information or because non-American players actually did not belong to any college. The NA values related to the Ht, Wt columns are due to the fact that those players are not present in the player_stat dataset.

```
nba_performance %>%
  summarise_all(list(~sum(is.na(.))))
```

```
##   Year Player Pos Age Tm HallOfFame  Ht  Wt X G GS MpG PpG ApG RpG TOpG BpG SpG
## 1    0      0   0   0  0          0 416 416 0 0  0   0   0   0   0    0   0   0
##   FpG MP FG FGA FG. X3P X3PA X3P. X2P X2PA X2P. eFG. FT FTA FT. ORB DRB TRB AST
## 1   0  0  0   0   0   0    0  252   0    0    0    0  0   0   4   0   0   0   0
##   STL BLK TOV PF PTS PER TS. X3PAr FTr ORB. DRB. TRB. AST. STL. BLK. TOV. USG.
## 1   0   0   0  0   0   0   0     0   0    0    0    0    0    0    0    0    0
##   OWS DWS WS WS.48 OBPM DBPM BPM VORP
## 1   0   0  0     0    0    0   0    0
```

```
nba_performance$X3P. = replace(nba_performance$X3P., is.na(nba_performance$X3P.), 0)
nba_performance$FT. = replace(nba_performance$FT., is.na(nba_performance$FT.), 0)

nba_performance = nba_performance %>%
  group_by(Pos) %>%
  mutate(MeanWt = mean(Wt, na.rm = TRUE),
         MeanHt = mean(Ht, na.rm = TRUE)) %>%
  ungroup()

nba_performance$Wt = ifelse(is.na(nba_performance$Wt), nba_performance$MeanWt, nba_performance$Wt)
nba_performance$Ht = ifelse(is.na(nba_performance$Ht), nba_performance$MeanHt, nba_performance$Ht)

nba_performance <- nba_performance %>% select(-c(MeanWt, MeanHt))

nba_performance = as.data.frame(nba_performance)
```

Physique evolution

In recent decades, the evolution of NBA players' physiques has reflected a significant transformation in the game. Different positions on the court have shown distinctive variations in players' physiques over time. This evolution of physique reflects not only changes in the game itself, but also the training strategies and demands of the modern NBA style of play.

```
PosColorCode <- c("C"="red", "PF"="orange",
                  "SF"="yellow" ,"SG"="blue", "PG"="green")
```

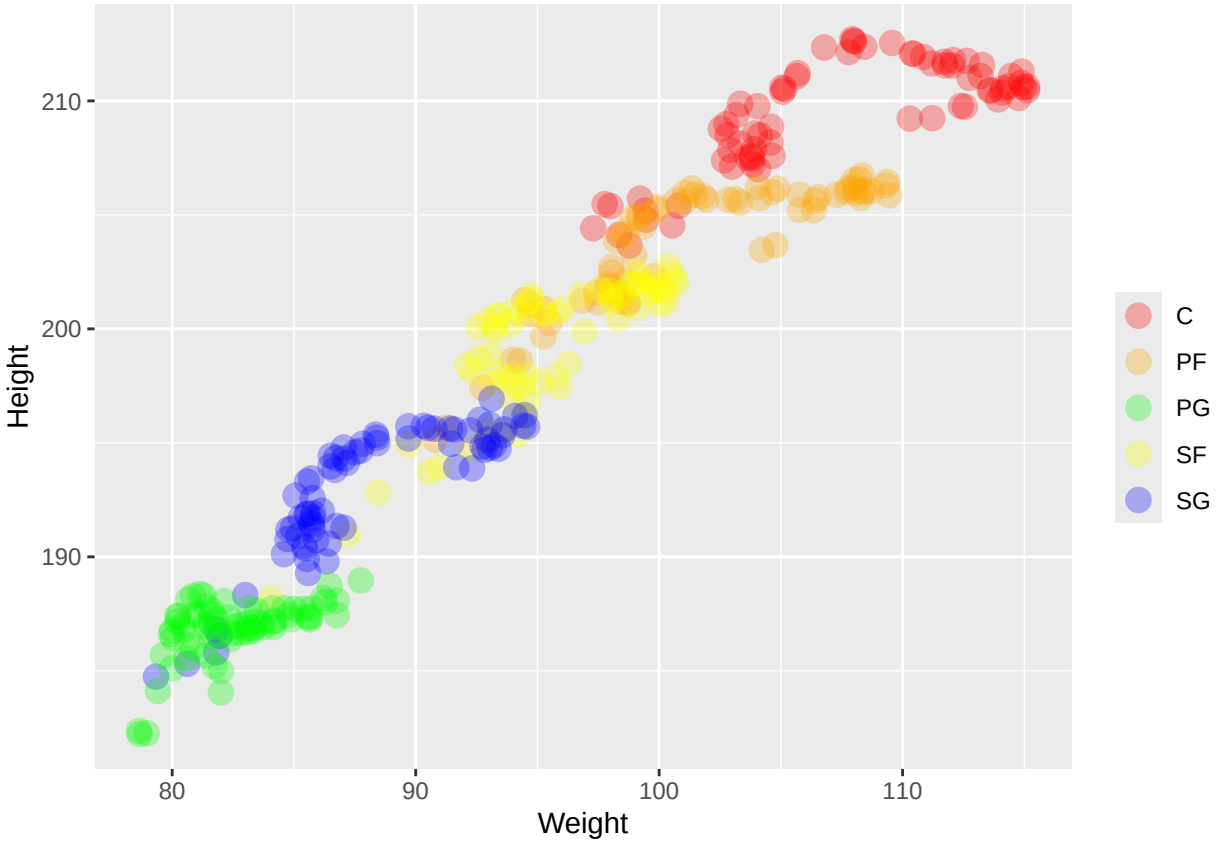
Note how centers and power forward are the strongest athletes physically, while point guards and small forwards are the lightest. Shooting guards have a very wide distribution, so they can be defined, in terms of physical stature, as the ideal NBA player prototype.

```
physique <- nba_withTOT %>%
  group_by(Year, Pos) %>%
  summarise("Height" = mean(Ht, na.rm = T), "Weight"= mean(Wt, na.rm = T))
```

```
## `summarise()` has grouped output by 'Year'. You can override using the
## `.groups` argument.
```

```
#physique

ggplot(physique, aes(x=Weight, y=Height, color=Pos)) +
  geom_point(size=4, alpha = 0.3)+
  scale_color_manual(values = PosColorCode, name = "")
```



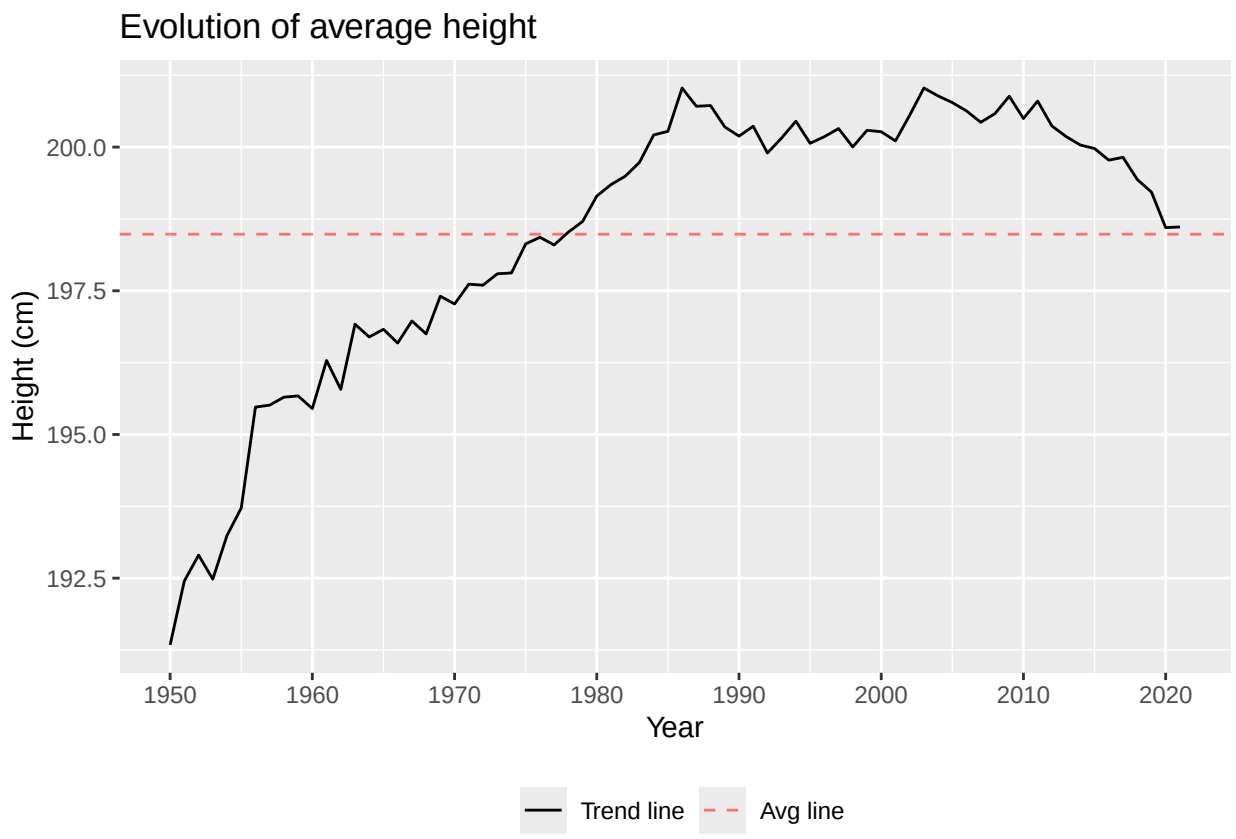
The game of basketball has inevitably changed a great deal since its inception, therefore, even the same conventional positions have adapted to the modern NBA. The most significant change came with the introduction of the three-point shot in the early 1980s, which radically changed the way the game was played. The new rule increased the importance of shooting skills from long distance, giving more responsibility for scoring points to players who, as opposed to pivots, played away from the basket. Traditionally, having a dominant center was fundamental to the structure of a team, but, however, in recent years, many teams have embraced a playing philosophy that emphasizes speed, mobility, and versatility, preferring lineups without a traditional center. This significant evolution in the position concept is called **"small ball"**. This has resulted in taller players also moving to the perimeter, with shorter players filling more interior roles. These new centers, referred to as **"big stretch"**, in addition to dominating in the painted area, are also able to shoot long range, pushing opponents out of their defensive comfort zone. This has opened up spaces for teammates and created new offensive opportunities.

The evolution of the game led to a consequent transformation of the players, both in terms of individual technique and from the point of view of physical structure. It can be seen that, until the early 1980s, growth in terms of stature was stable. From there on there is no clear positive or negative trend, but it is evident how the last decade was the first in history in which NBA players became shorter than the previous decade. This decline is also tangible for individual positions, except for point guards who have reached their maximum height in recent seasons, the other positions have become shorter and shorter.

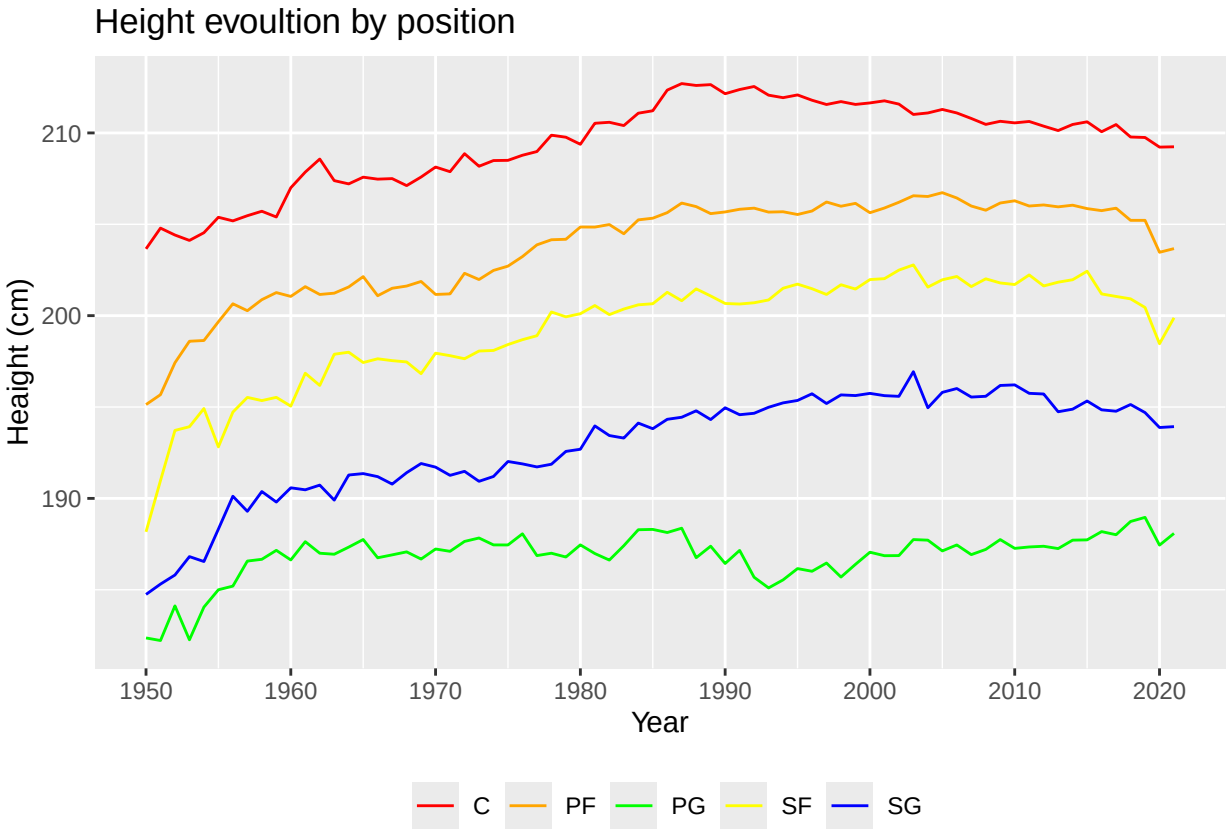
```
avg <- nba_withTOT %>%
  group_by(Year) %>%
  summarise("Height"=mean(Ht, na.rm = T), "Weight"=mean(Wt, na.rm = T))

ggplot(avg, aes(x=Year, y=Height, linetype = "Trend line")) +
  geom_line()+
  labs(x="Year", y="Height (cm)", title = "Evolution of average height")+
```

```
  geom_hline(aes(yintercept = mean(Height), linetype = "Avg line"), col = "red", alpha = 0.5) +
  scale_x_continuous(breaks = seq(1950, 2021, 10)) +
  scale_linetype_manual(name = "", values = c(2, 1), guide = guide_legend(reverse = TRUE))+
  theme(legend.position = "bottom")
```

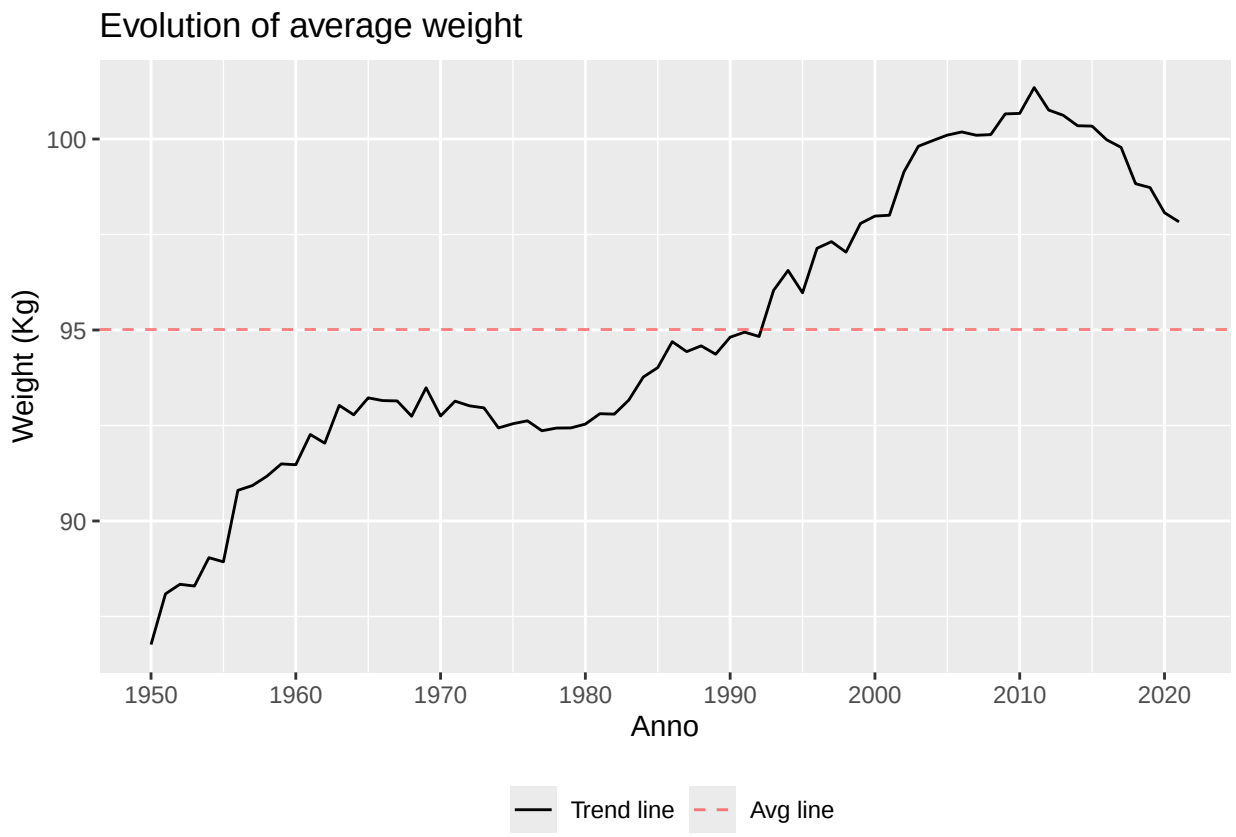



```
physique %>%
  ggplot( aes(x=Year, y=Height, group=Pos, color=Pos)) +
  geom_line()+
  labs(x="Year", y = "Heaight (cm)", title = "Height evoultion by position")+
  theme(legend.position = "bottom")+
  scale_color_manual(values = PosColorCode , name = "")+
  scale_x_continuous(breaks = seq(1950, 2021, 10))
```

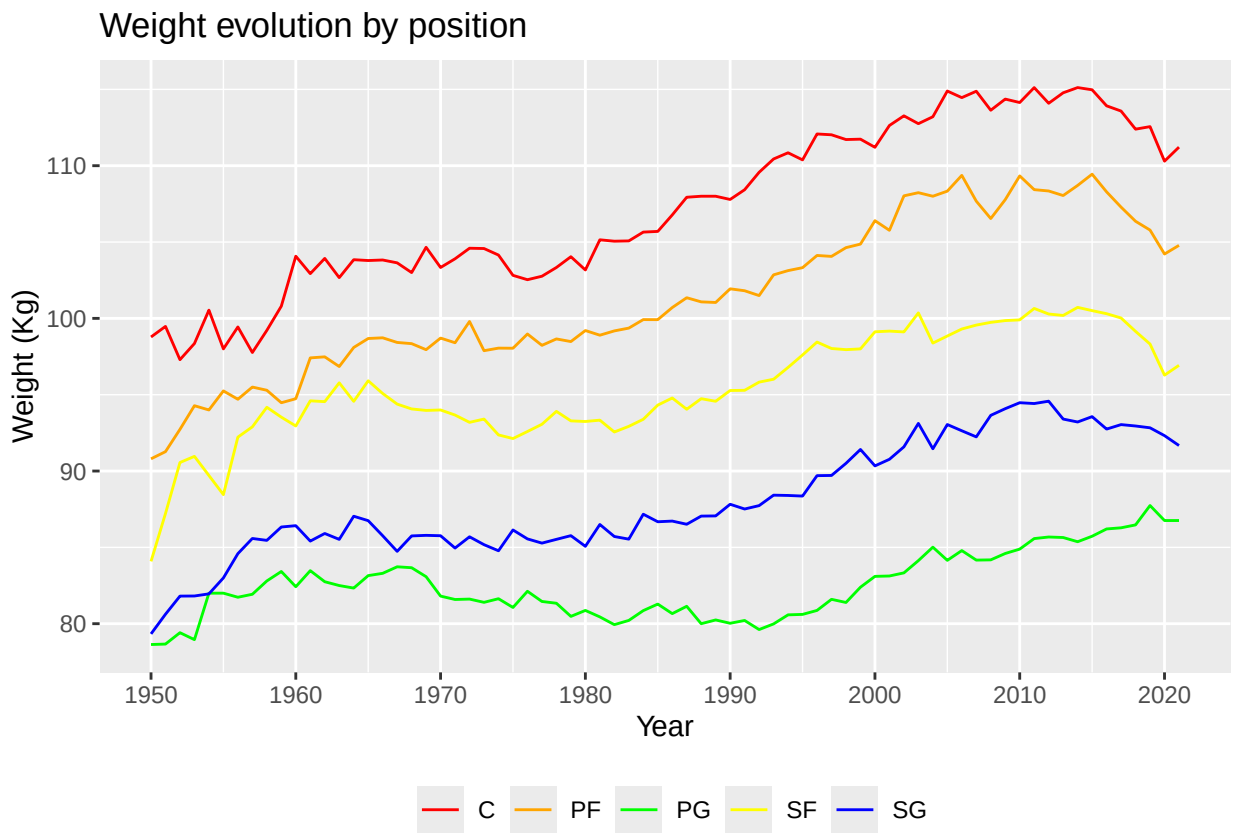


In terms of weight, its average grew steadily until the 1970s, peaked in the 2010/11 season, and has been declining ever since, in fact the players are found to be among the lightest in the 21st century. It can be seen that since 2010 height and weight have undertaken a decreasing trend. This phenomenon can be attributed to the fact that NBA athletes, particularly the **"big men"**, have had to become faster and leaner to adapt to the perimeter-oriented game. This is why NBA centers and forwards are facing the greatest decline in their physical stature. Similar to the evolution of height, point guards are among the heaviest they have ever been, while all others have become lighter.

```
ggplot(avg, aes(x=Year, y=Weight, linetype = "Trend line")) +
  geom_line()+
  labs(x="Anno", y="Weight (Kg)", title = "Evolution of average weight")+
  geom_hline(aes(yintercept = mean(Weight), linetype = "Avg line"), col = "red", alpha = 0.5) +
  scale_x_continuous(breaks = seq(1950, 2021, 10)) +
  scale_linetype_manual(name = "", values = c(2, 1), guide = guide_legend(reverse = TRUE))+
  theme(legend.position = "bottom")
```




```
physique %>%
  ggplot( aes(x=Year, y= Weight, group=Pos, color=Pos)) +
  geom_line()+
  labs(x="Year", y = "Weight (Kg)", title = "Weight evolution by position")+
  theme(legend.position = "bottom")+
  scale_color_manual(values = PosColorCode , name = "")+
  scale_x_continuous(breaks = seq(1950, 2021, 10))
```

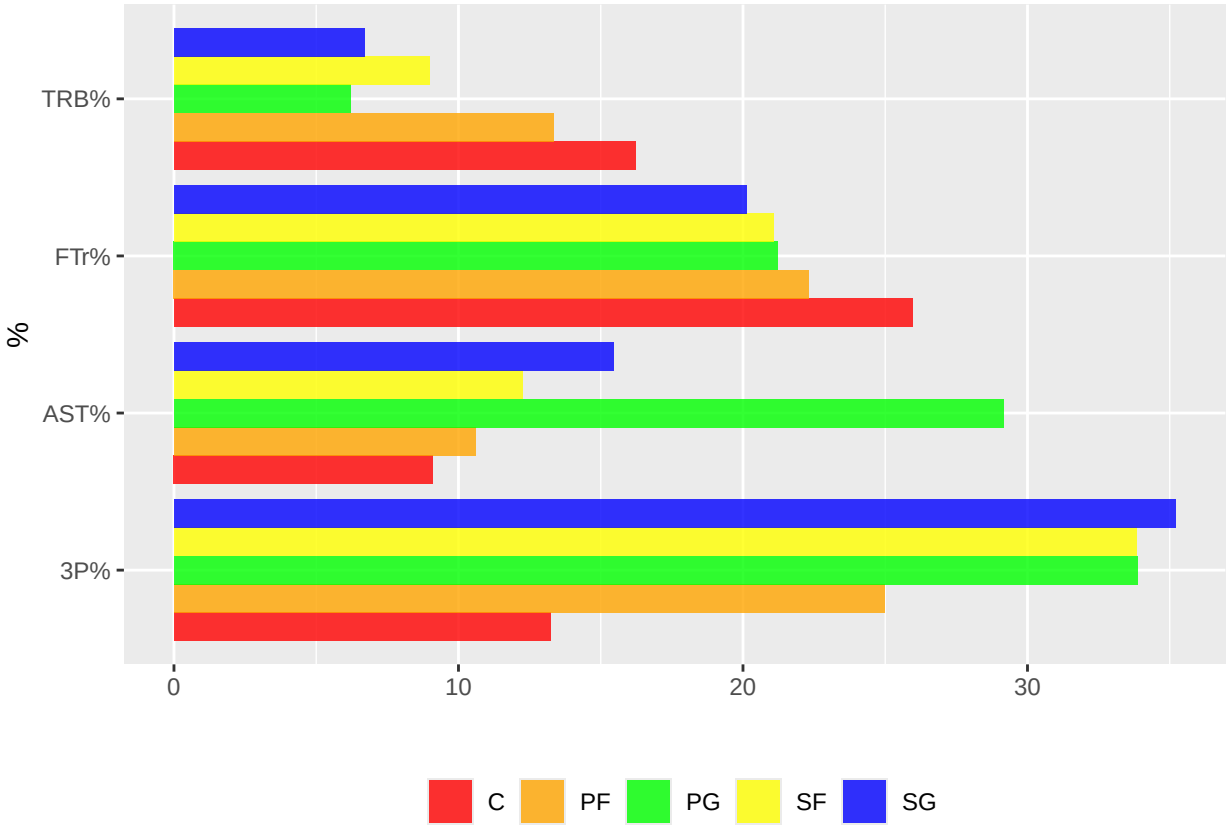


These technical and physical differences between positions are made explicit by some advanced statistics of the game represented by Oliver's four factors. It can be seen that centers and power forwards, taking advantage of their physicality, record high values for the percentage of rebounds collected per game and the rate of free throws obtained, the latter quantifying the player's aggressiveness to attack the basket and thus to obtain fouls; the point guards, being the players from whom each team's offensive actions start, are those who have the greatest ability to distibute smarcating passes to the forwards and thus simplify their finalization; demonstrating shooting skills, especially from behind the arc, is the percentage of three-point shots made in which the shooting guards excel.

```
x = nba_performance %>%
  group_by(Pos) %>%
  summarise("TRB%"= mean(TRB., na.rm = T), "3P%"=mean(X3P., na.rm = T)*100,
            "AST%"=mean(AST., na.rm = T), "FTr%" = mean(FTr, na.rm = T)*100)


df <- data.frame(stats = rep(c("AST%", "FTr%", "3P%", "TRB%"), each = 5),
                 position = rep(x$Pos, times = 4),
                 value = c(x$`AST%`,x$`FTr%`, x$`3P%`, x$`TRB%`))

ggplot(df, aes(fill = position, y = value, x = stats))+
  geom_bar(position = "dodge", stat = "identity", alpha = 0.8)+
  labs(x="%", y="")+
  coord_flip()+
  theme(legend.position = "bottom")+
  scale_fill_manual(values = PosColorCode, name = "")
```



Cluster analysis

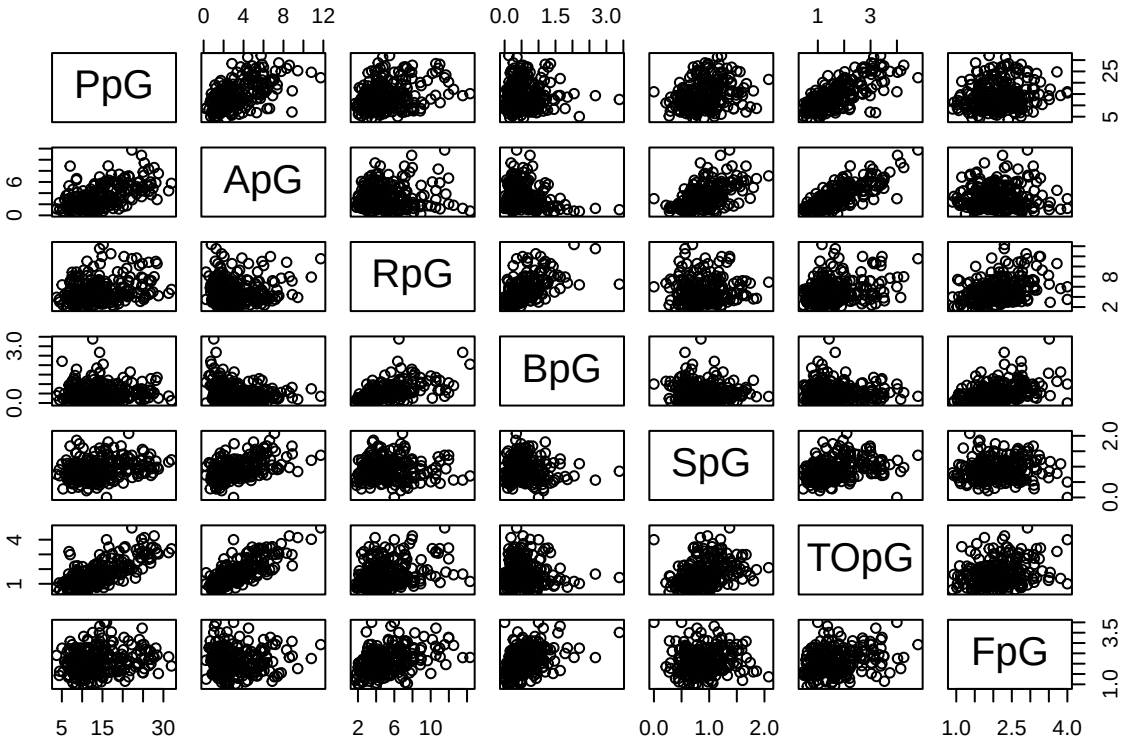
In the following section, the non-hierarchical k-means technique is applied to a group of NBA players related to the 2020/2021 season.

In line with Alagappam's study, the variables considered in clustering will be: * points per game (PpG) * assists per game (ApG) * rebounds per game (RpG) * blocked shots per game (BpG) * stolen balls per game (SpG) * possessions lost per game (TOpG) * fouls per game (FpG)

```
players <- nba_performance %>%
  filter(Year == 2021) %>%
  mutate("FpG" = round(PF/G,3)) %>%
  select(Player, Pos, PpG, ApG, RpG, BpG, SpG, TOpG, FpG)
table(players$Pos)
```

```
##
##  C PF PG SF SG
## 41 48 51 44 67
```

```
players_data <- players[3:9]
plot(players_data)
```

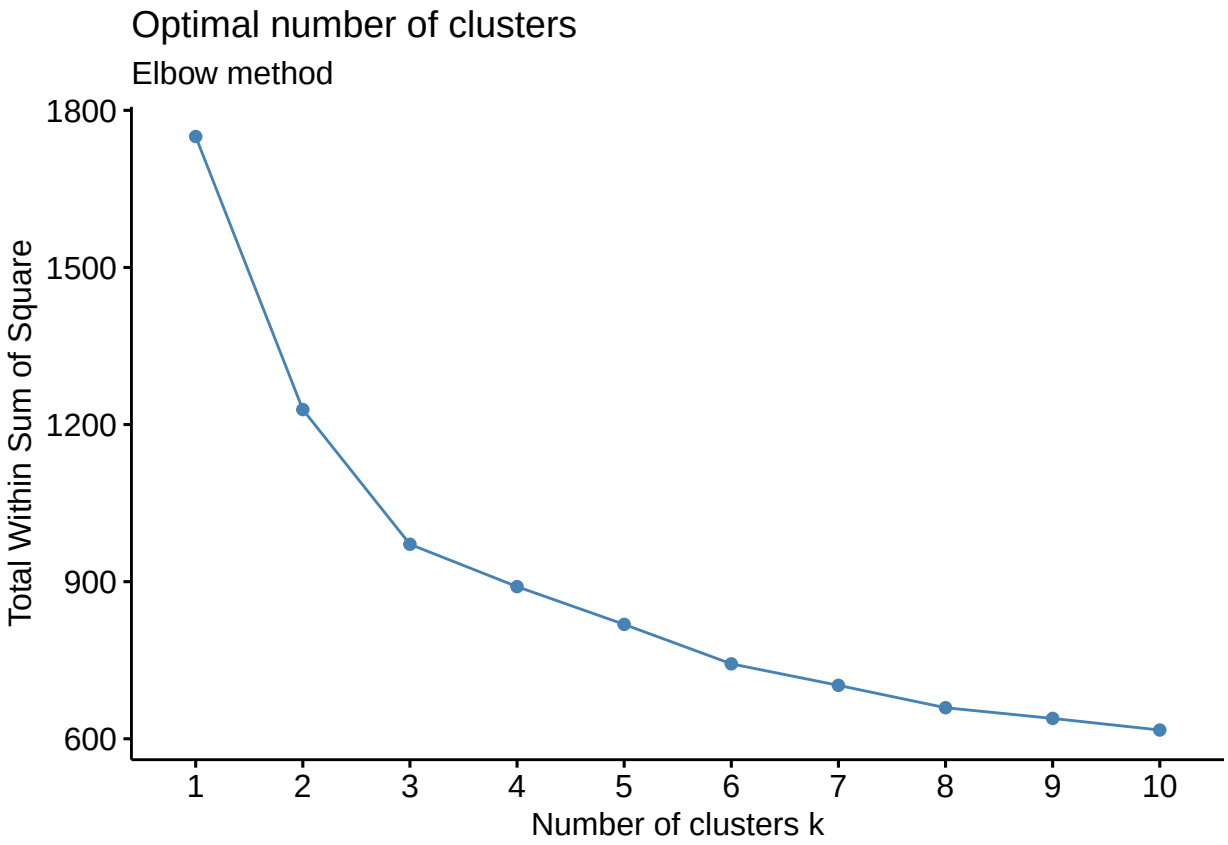


Before beginning the analysis, it should be pointed out how the use of k-means on raw data could provide results that are not very useful because they are highly variable. Therefore, as with hierarchical clustering procedures, we chose to normalize each variable, that is, transform them so that they have a mean of zero and a standard deviation of one, through the scale() function.

```
players_data_scale <- scale(players_data)
```

The methodology involves a preliminary hierarchical analysis to determine the appropriate number of partitions. The technique is that of the elbow method, which is based on analyzing the variation of the within-cluster sum-of-squares index, also known as WSS (Within-Cluster Sum of Squares). The main idea of the elbow method is to identify the point at which the WSS index curve exhibits an "elbow" or an abrupt reversal of its trend. This point indicates the optimal number of clusters to be used.

```
fviz_nbclust(players_data_scale, kmeans, method = "wss")+
  labs(subtitle = "Elbow method")
```

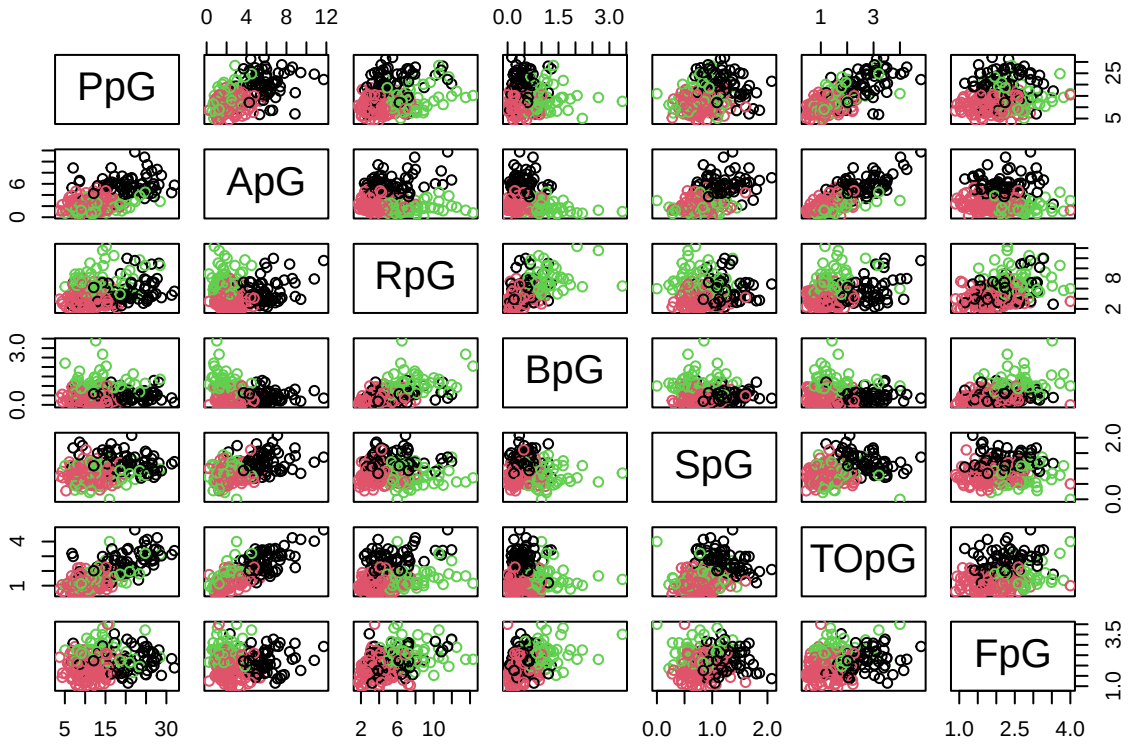



```r
set.seed(123)
km.out <- kmeans(players_data_scale, centers = 3)
print(km.out)
```

```
## K-means clustering with 3 clusters of sizes 66, 136, 49
##
## Cluster means:
##           PpG        ApG        RpG        BpG        SpG        TOpG        FpG
## 1  1.07137699  1.3164141  0.1541463 -0.1973861  1.0076705  1.2002542  0.1964302
## 2 -0.51233373 -0.4086120 -0.4972922 -0.4023371 -0.3255658 -0.5427195 -0.4076230
## 3 -0.02109171 -0.6390224  1.1726139  1.3825579 -0.4536592 -0.1103453  0.8667825
##
## Clustering vector:
##   [1] 3 1 2 2 2 3 2 1 2 2 1 2 2 3 2 3 1 1 2 2 2 2 2 2 1 2 2 2 2 2 2 2 1 3 2 2
##  [38] 3 2 1 2 1 2 1 3 2 2 3 3 2 2 1 2 2 3 3 2 2 2 2 3 1 2 3 2 2 1 3 2 1 2 2 2 1
##  [75] 2 2 2 3 1 1 2 3 2 2 2 1 3 1 2 1 2 1 1 3 2 2 2 3 2 1 2 2 2 2 2 1 2 2 2 3 2
## [112] 1 1 2 2 2 2 1 2 3 2 2 2 2 3 2 2 1 1 3 3 2 1 2 2 2 1 2 3 2 2 2 2 1 1 1 1 3
## [149] 2 1 2 2 2 2 1 1 2 2 1 2 2 1 2 1 2 2 1 1 1 2 3 2 3 2 2 2 2 1 3 3 2 3 1 2 3
## [186] 3 2 1 3 2 3 2 2 1 2 2 2 2 3 2 2 2 1 1 1 1 2 1 2 1 2 1 1 1 2 2 3 3 1 2 3 1
## [223] 3 2 3 2 2 3 3 1 3 1 2 1 2 3 1 2 2 3 2 2 2 1 3 3 2 1 1 2 3
##
## Within cluster sum of squares by cluster:
## [1] 347.6019 344.2336 279.5614
##  (between_SS / total_SS =  44.5 %)
##
```

```
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
plot(players_data, col = km.out$cluster)
```

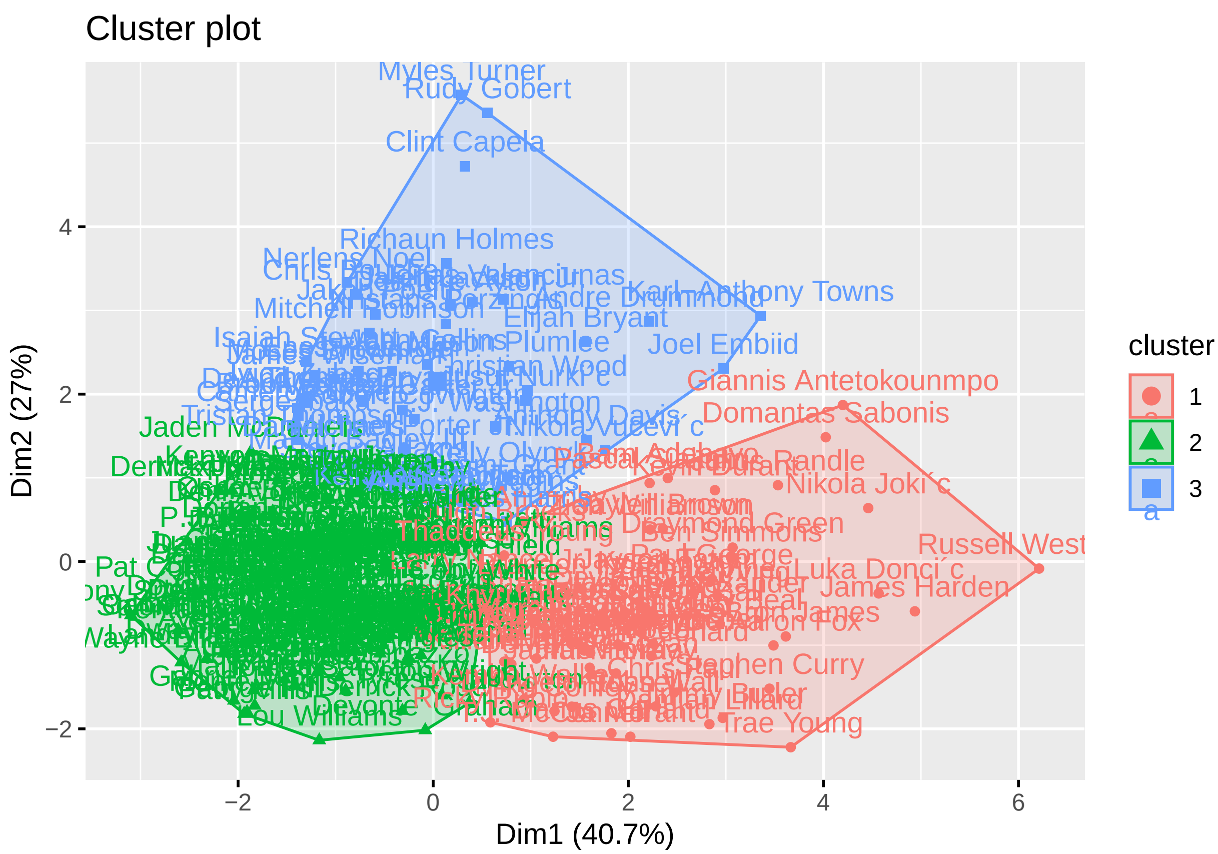


The main output produced by the algorithm is as follows:

```
km.clusters <- km.out$cluster

rownames(players_data_scale) <- players$Player

fviz_cluster(list(data = players_data_scale, cluster= km.clusters))
```




```r
table(km.clusters, players$Pos)
```

```
##
## km.clusters  C PF PG SF SG
##           1  2 10 30  9 15
##           2  6 29 21 30 50
##           3 33  9  0  5  2
```

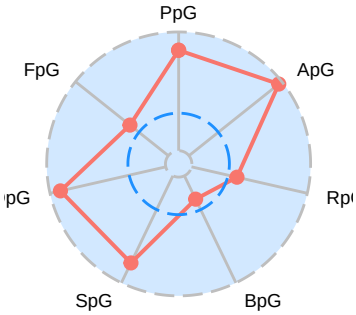
```r
library(BasketballAnalyzeR)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

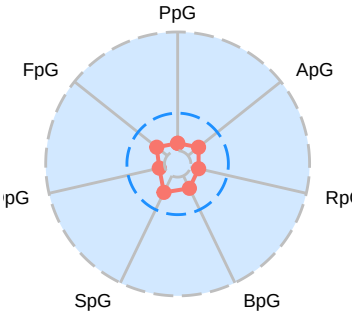
```
##
## If you want to reproduce the figures contained in the book of
## Zuccolotto and Manisera (2020) and
## if the version of your R machine is >= 3.6.0, you need to type
## RNGkind(sample.kind = "Rounding")
## at the beginning of your working session
```

```r
set.seed(123)
kc <- kclustering(players_data, k=3, labels = players$Player)
plot(kc, profiles = F, title = c("Scoring Leaders","Perimeter Facilitator","Painted Protectors"))
```

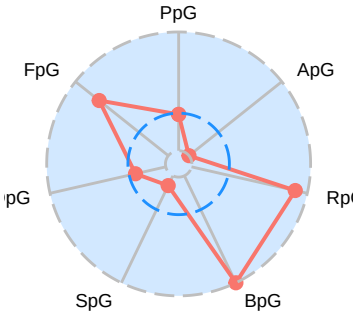
Scoring Leaders



Perimeter Facilitator



Painted Protectors




```
players %>%
  mutate(Cluster = km.clusters) %>%
  group_by(Cluster) %>%
  select(-Player, -Pos) %>%
  summarise_all(mean)
```

```
## # A tibble: 3 x 8
##   Cluster   PpG   ApG   RpG   BpG   SpG  TOpG   FpG
##     <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1       1  20.4  5.85  5.49 0.463 1.22   2.66  2.23
## 2       2  11.0  2.31  3.96 0.370 0.777  1.18  1.88
## 3       3  13.9  1.84  7.89 1.18  0.734  1.55  2.62
```