Pràctica 2 - Neteja i validació de dades

Jordi Puig Benages

3 de enero, 2022

1 - Conceptes prèvis

1.1 - Presentació

En aquesta pràctica s'elabora un cas pràctica orientat a aprendre a identificar les dades rellevants per un projecte analític i usar les eines d'integració, neteja, validació i anàlisi de les mateixes.

1.2 - Competències

En aquesta pràctica es desenvolupen les següents competències del Màster de Data Science:

- Capacitat d'analitzar un problema en el nivell d'abstracció adequat a cada situació i aplicar les habilitats i coneixements adquirits per abordar-lo i resoldre'l.
- Capacitat per aplicar les tècniques específiques de tractament de dades (integració, transformació, neteja i validació) per al seu posterior anàlisi.

1.3 - Objectius

Els objectius concrets d'aquesta pràctica són:

- Aprendre a aplicar els coneixements adquirits i la seva capacitat de resolució de problemes en entorns nous o poc coneguts dintre de contextos més amplis o multidisciplinaris.
- Saber identificar les dades rellevants i els tractaments necessaris (integració, neteja i validació) per dur a terme un projecte analític.
- Aprendre a analitzar les dades adequadament per abordar la informació continguda en les dades.
- Identificar la millor representació dels resultats per tal d'aportar conclusions sobre el problema plantejat en el procés analític.
- Actuar amb els principis ètics i legals relacionats amb la manipulació de dades en funció de l'àmbit d'aplicació.
- Desenvolupar les habilitats d'aprenentatge que els permetin continuar estudiant d'una manera que haurà de ser en gran manera autodirigida o autònoma.
- Desenvolupar la capacitat de cerca, gestió i ús d'informació i recursos en l'àmbit de la ciència de dades.

2 - Resolució

2.1 - Descripció del dataset

En aquesta pràctica s'elabora un projecte de tractament i anàlisi de dades a partir d'un dataset generat amb la informació disponible a https://fbref.com/en/comps/ (https://fbref.com/en/comps/)

El conjunt de dades shooting_stats_big5_leagues_21_22 conté informació d'una mostra de tots els jugadors que han participat en partits oficials de les 5 lligues europees de futbol més importants: la lliga espanyola, la lliga anglesa, la lliga francesa, l'alemanya i la italiana. En el conjunt de dades es recullen mètriques relacionades amb els tirs.

Els camps que trobem al dataset són els següents:

- Rk: identificador del jugador
- Player: nom del jugador i identificador únic de la web de fbref.com
- Nation: nacionalitat del jugador, primer en minúscules amb2/3 caràcters i després normalitzat en majúscules i 3 caràcters.
- Pos: Posició del jugador, pot ser una o dues concatenades entre GK, DF, MF i FW
- Competition: competició a la que juga entre les 5 principals europees
- Squad: equip en el que juga
- Age: edat, en format anys i dies
- Born: any de naixement
- noventas: número de blocs de 90 minuts jugats, fet servir per normalitzar després algunes mètriques per 90 minuts.
- Gls: número de gols marcats
- Sh: número de tirs
- SoT: número de tirs a porteria
- SoT%: percentatge de tirs a porteria
- Sh/90: tirs per 90 minuts
- SoT/90: tirs a porteria per 90 minuts
- G/Sh: gols per tir
- G/SoT: gols per tir a porteria
- Dist: distancia total dels tirs
- FK: tirs de falta
- PK: gols de penalti
- PKatt: tirs de penalti intentats
- xG: gols esperats
- npxG: gols esperats sense tenir en compte penaltis
- npxG/Sh: gols esperats per tir sense tenir en compte penaltis
- G-xG: diferència entre gols i gols esperats
- np:G-xG: diferència entre gols i gols esperats sense tenir en compte penaltis

2.2 - Integració i selecció de les dades d'interès a analitzar

- Es llegeix el fitxer shooting_stats_big5_leagues_21_22 i es guarden les dades a l'objecte shooting.
- Es verifica que les dades s'han carregat correctament.
- S'avalua que no hi hagi jugadors duplicats, en cas que fos així s'escolliria el registre amb més dades.
- Es fa un preprocessament d'alguns camps, eliminant de Player el codi a partir del caràcter '',

- S'agafa només els 3 últims caràcters de la nacionalitat.
- Respecte la posició, es genera un camp amb posició principal, amb els dos primers caràcters del camp Pos, i posició alternativa amb els altres caracters si existeixen.
- En quant a l'edat, s'agafa només els anys, descartant els dies (que són els caràcters a partir del guió '-')
- Un cop realitzats els canvis, l'objectiu és analitzar si hi ha alguna diferència important entre les grans competicions europees en termes de finalització. En una temporada en la que es parla molt de la baixada de nivell de la competició espanyola, es poden demostrar aquestes impressions amb dades? És cert que la Premier League té els millors davanters del mon?

```
shooting <- read_excel("shooting_stats_big5_leagues_21_22.xlsx") ##Lectura de l'arxiu Excel
head (shooting, 10) ##Es mostren els primers 10 registres de l'arxiu d'entrada
```

```
## # A tibble: 10 x 26
##       Rk Player  Nation Pos   Competition Squad Age    Born noventas   Gls    Sh
##    <dbl> <chr>   <chr>  <chr> <chr>       <chr> <chr> <dbl> <chr>    <dbl> <dbl>
##  1    77 "Aaron~ ie IRL FW    Premier Le~ Brig~ 21-3~  2000 1.7          0     2
##  2    83 "Aaron~ eng E~ DF    Premier Le~ West~ 32-0~  1989 12.7         1     9
##  3   623 "Aarón~ es ESP GK    La Liga     Gran~ 26-0~  1995 3.2          0     0
##  4  1658 "Aaron~ sct S~ DF    Serie A     Bolo~ 19-2~  2002 16.1         4    15
##  5   238 "Aaron~ eng E~ MFFW  Premier Le~ Burn~ 34-2~  1987 2.1          1     4
##  6  1263 "Aarón~ es ESP DF    Bundesliga  Main~ 24-2~  1997 11.5         0    11
##  7   345 "Aaron~ eng E~ GK    Premier Le~ Arse~ 23-2~  1998 17.0         0     0
##  8  1829 "Aaron~ wls W~ MF    Serie A     Juve~ 31-0~  1990 1.1          0     0
##  9   456 "Aaron~ eng E~ DF    Premier Le~ Manc~ 24-0~  1997 14.0         0     2
## 10  2215 "Abdel~ dz ALG DF    Ligue 1     Bord~ 24-1~  1997 3.0          0     2
## # ... with 15 more variables: SoT <dbl>, SoT% <chr>, Sh_noventa <chr>,
## #   SoT_noventa <chr>, G_Sh <chr>, G_SoT <chr>, Dist <chr>, FK <dbl>, PK <dbl>,
## #   PKatt <dbl>, xG <chr>, npxG <chr>, npxG_Sh <chr>, G_xG <chr>, npG-xG <chr>
```

```
tail(shooting, 10) ##Es mostren els darrers 10 registres de l'arxiu d'entrada
```

```
## # A tibble: 10 x 26
##       Rk Player  Nation Pos   Competition Squad Age    Born noventas   Gls    Sh
##    <dbl> <chr>   <chr>  <chr> <chr>       <chr> <chr> <dbl> <chr>    <dbl> <dbl>
##  1  2473 "Yusuf~ tr TUR MFFW  Ligue 1     Lille 24-3~  1997 5.7          1    10
##  2  2341 "Yvan ~ cm CMR MF    Ligue 1     Sain~ 24-3~  1997 9.9          0     6
##  3  2279 "Yvann~ fr FRA DF    Ligue 1     Sain~ 23-0~  1998 10.4         0     6
##  4    46 "Yves ~ ml MLI MF    Premier Le~ Brig~ 25-1~  1996 12.4         0    11
##  5   412 "Zack ~ us USA GK    Premier Le~ Manc~ 26-2~  1995 1.0          0     0
##  6  2475 "Zaydo~ fr FRA MF    Ligue 1     Sain~ 22-1~  1999 9.3          0    14
##  7  2063 "Zeki ~ tr TUR DF    Ligue 1     Lille 24-3~  1997 13.7         0    11
##  8  1388 "Zidan~ dk DEN FWMF  Bundesliga  Leve~ 16-3~  2005 0.1          0     0
##  9  1925 "Zinho~ be BEL DF    Serie A     Genoa 22-1~  1999 6.0          0     4
## 10  1664 "Zlata~ se SWE FW    Serie A     Milan 40-0~  1981 7.8          7    35
## # ... with 15 more variables: SoT <dbl>, SoT% <chr>, Sh_noventa <chr>,
## #   SoT_noventa <chr>, G_Sh <chr>, G_SoT <chr>, Dist <chr>, FK <dbl>, PK <dbl>,
## #   PKatt <dbl>, xG <chr>, npxG <chr>, npxG_Sh <chr>, G_xG <chr>, npG-xG <chr>
```

```
summary(shooting) ##S'extreu un primer anàlisi estadístic amb les característiques bàsiques d
el dataset
```

```
##        Rk            Player             Nation              Pos
##  Min.   :   1.0   Length:2436        Length:2436        Length:2436
##  1st Qu.: 614.8   Class :character   Class :character   Class :character
##  Median :1238.5   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1237.7
##  3rd Qu.:1855.2
##  Max.   :2478.0
##
##  Competition           Squad               Age                 Born
##  Length:2436        Length:2436        Length:2436        Min.   :1981
##  Class :character   Class :character   Class :character   1st Qu.:1992
##  Mode  :character   Mode  :character   Mode  :character   Median :1995
##                                                           Mean   :1995
##                                                           3rd Qu.:1998
##                                                           Max.   :2005
##
##    noventas              Gls               Sh              SoT
##  Length:2436        Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  Class :character   1st Qu.: 0.000   1st Qu.: 1.000   1st Qu.: 0.000
##  Mode  :character   Median : 0.000   Median : 6.000   Median : 1.000
##                     Mean   : 1.013   Mean   : 9.126   Mean   : 2.979
##                     3rd Qu.: 1.000   3rd Qu.:13.000   3rd Qu.: 4.000
##                     Max.   :19.000   Max.   :75.000   Max.   :37.000
##
##      SoT%            Sh_noventa        SoT_noventa           G_Sh
##  Length:2436        Length:2436        Length:2436        Length:2436
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     G_SoT               Dist                FK                PK
##  Length:2436        Length:2436        Min.   : 0.0000   Min.   :0.00000
##  Class :character   Class :character   1st Qu.: 0.0000   1st Qu.:0.00000
##  Mode  :character   Mode  :character   Median : 0.0000   Median :0.00000
##                                        Mean   : 0.3248   Mean   :0.09606
##                                        3rd Qu.: 0.0000   3rd Qu.:0.00000
##                                        Max.   :16.0000   Max.   :6.00000
##                                        NA's   :1
##      PKatt             xG               npxG             npxG_Sh
##  Min.   :0.0000   Length:2436        Length:2436        Length:2436
##  1st Qu.:0.0000   Class :character   Class :character   Class :character
##  Median :0.0000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :0.1154
##  3rd Qu.:0.0000
##  Max.   :7.0000
##
##      G_xG              npG-xG
##  Length:2436        Length:2436
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
```

```
##
##
```

```
#S'analitza els casos duplicats del camp Rk i es comprova que siguin 0
idx <- table(shooting$Rk)>1
count <- sum(idx)

#S'analitza els casos duplicats del camp Player i es comprova que siguin 0
noms_duplicats <- table(shooting$Player)>1
count2 <- sum(noms_duplicats)

#Es modifiquen alguns camps, extreient el codi del nom del player, eliminant els dies del cam
p edat, agafant el camp nacionalitat amb només 3 caràcters en majúsucles, i afegint la posici
ó principal i la secundària del camp pos.
shooting$Player <- str_sub(shooting$Player,1,nchar(shooting$Player)-9)
shooting$Age <- sapply(strsplit(as.character(shooting$Age),"-"), "[",1)
shooting$Nation <- sapply(strsplit(as.character(shooting$Nation)," "), "[",2)
shooting$Pos_main <- substr(shooting$Pos,1,2)
shooting$Pos_secondary <- substring(shooting$Pos, 3)

#Es mostren els primers 10 registres del dataset modificat
head(shooting,10)
```

```
## # A tibble: 10 x 28
##       Rk Player  Nation Pos   Competition Squad Age    Born noventas   Gls    Sh
##    <dbl> <chr>   <chr>  <chr> <chr>       <chr> <chr> <dbl> <chr>    <dbl> <dbl>
##  1    77 Aaron ~ IRL    FW    Premier Le~ Brig~ 21     2000 1.7          0     2
##  2    83 Aaron ~ ENG    DF    Premier Le~ West~ 32     1989 12.7         1     9
##  3   623 Aarón ~ ESP    GK    La Liga     Gran~ 26     1995 3.2          0     0
##  4  1658 Aaron ~ SCO    DF    Serie A     Bolo~ 19     2002 16.1         4    15
##  5   238 Aaron ~ ENG    MFFW  Premier Le~ Burn~ 34     1987 2.1          1     4
##  6  1263 Aarón ~ ESP    DF    Bundesliga  Main~ 24     1997 11.5         0    11
##  7   345 Aaron ~ ENG    GK    Premier Le~ Arse~ 23     1998 17.0         0     0
##  8  1829 Aaron ~ WAL    MF    Serie A     Juve~ 31     1990 1.1          0     0
##  9   456 Aaron ~ ENG    DF    Premier Le~ Manc~ 24     1997 14.0         0     2
## 10  2215 Abdel ~ ALG    DF    Ligue 1     Bord~ 24     1997 3.0          0     2
## # ... with 17 more variables: SoT <dbl>, SoT% <chr>, Sh_noventa <chr>,
## #   SoT_noventa <chr>, G_Sh <chr>, G_SoT <chr>, Dist <chr>, FK <dbl>, PK <dbl>,
## #   PKatt <dbl>, xG <chr>, npxG <chr>, npxG_Sh <chr>, G_xG <chr>, npG-xG <chr>,
## #   Pos_main <chr>, Pos_secondary <chr>
```

2.3 - Neteja de les dades

En aquest apartat s'analitzen els camps per comprovar com tractar els valors buits i/o nuls, que en aquest cas, en ser en camps que són conseqüència dels tirs, que són 0 en tots els casos, es decideix per omplir amb valors 0.

Seguidament s'analitza el tipus de dades dels camps, passant a numèric alguns dels camps que quedaven com a character de forma errònia (age i xG).

Per últim, es comprova que el rang de tots els valors dels camps està dins de l'acceptable i que no hi ha valors extrems que no siguin vàlids.

L'únic matís que es durà a terme més endavant serà filtrar aquells jugadors que no han arribat a 540 minuts (noventas >=6) en el transcurs de la temporada, doncs no seria del tot fiable tenir en compte tots els registres tot i que no hagin participat un mínim de partits.

```
rk_na <- sum(is.na(shooting$Rk)) ##No hi ha buits ni nuls
player_na <- sum(is.na(shooting$Player)) ##No hi ha buits ni nuls
nation_na <- sum(is.na(shooting$Nation)) ##No hi ha buits ni nuls
pos_main_na <- sum(is.na(shooting$Pos_main)) ##No hi ha buits ni nuls
pos_sec_na <- sum(is.na(shooting$Pos_secondary)) ##No hi ha buits ni nuls
comp_na <- sum(is.na(shooting$Competition)) ##No hi ha buits ni nuls
squad_na <- sum(is.na(shooting$Squad)) ##No hi ha buits ni nuls
age_na <- sum(is.na(shooting$Age)) ##No hi ha buits ni nuls
born_na <- sum(is.na(shooting$Born)) ##No hi ha buits ni nuls
goals_na <- sum(is.na(shooting$Gls)) ##No hi ha buits ni nuls
sh_na <- sum(is.na(shooting$Sh)) ##No hi ha buits ni nuls
sot_na <- sum(is.na(shooting$SoT)) ##No hi ha buits ni nuls
dist_na <- sum(is.na(shooting$Dist)) ## Hi ha 453 registres buits, es decideix substituir a 0 ja que són registres on tir =0
shooting$Dist [is.na(shooting$Dist)] <- 0

fk_na <- sum(is.na(shooting$FK)) ##1 registre buit, es passa a 0 doncs no té registres de tirs.
shooting$FK [is.na(shooting$FK)] <- 0

pk_na <- sum(is.na(shooting$PK))
pkatt_na <- sum(is.na(shooting$PKatt))

xg_na <- sum(is.na(shooting$xG)) ##1 registre buit, es passa a 0 doncs no té registres de tirs.
shooting$xG [is.na(shooting$xG)] <- 0

#Comprovació del data type d'alguns camps

class(shooting$Gls)
```

```
## [1] "numeric"
```

```
class(shooting$Sh)
```

```
## [1] "numeric"
```

```
class(shooting$SoT)
```

```
## [1] "numeric"
```

```
class(shooting$Dist)
```

```
## [1] "character"
```

```
shooting$Dist <- as.numeric(shooting$Dist)

class(shooting$Age)
```

```
## [1] "character"
```

```
shooting$Age <- as.numeric(shooting$Age)
class(shooting$Age)
```

```
## [1] "numeric"
```

```
class(shooting$xG)
```

```
## [1] "character"
```

```
shooting$xG <- as.numeric(shooting$xG)
class(shooting$npxG_Sh)
```

```
## [1] "character"
```

```
shooting$npxG_Sh <- as.numeric(shooting$npxG_Sh)
class(shooting$G_Sh)
```

```
## [1] "character"
```

```
shooting$G_Sh <- as.numeric(shooting$G_Sh)

class(shooting$noventas)
```

```
## [1] "character"
```

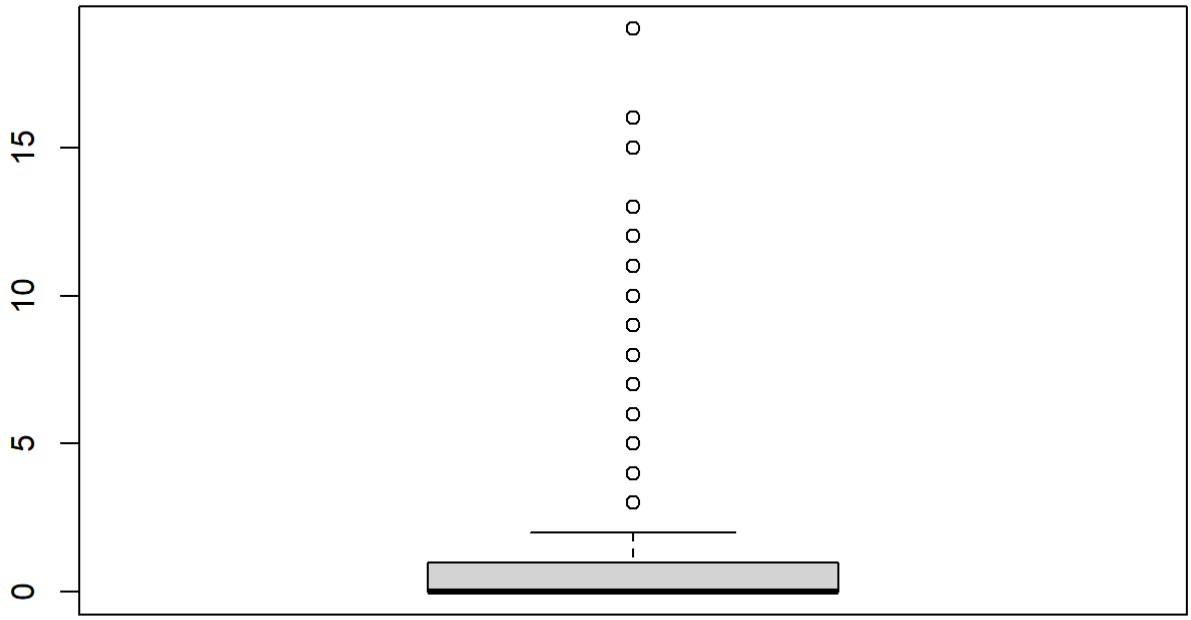
```
shooting$noventas <- as.numeric(shooting$noventas)

class(shooting$G_xG)
```

```
## [1] "character"
```

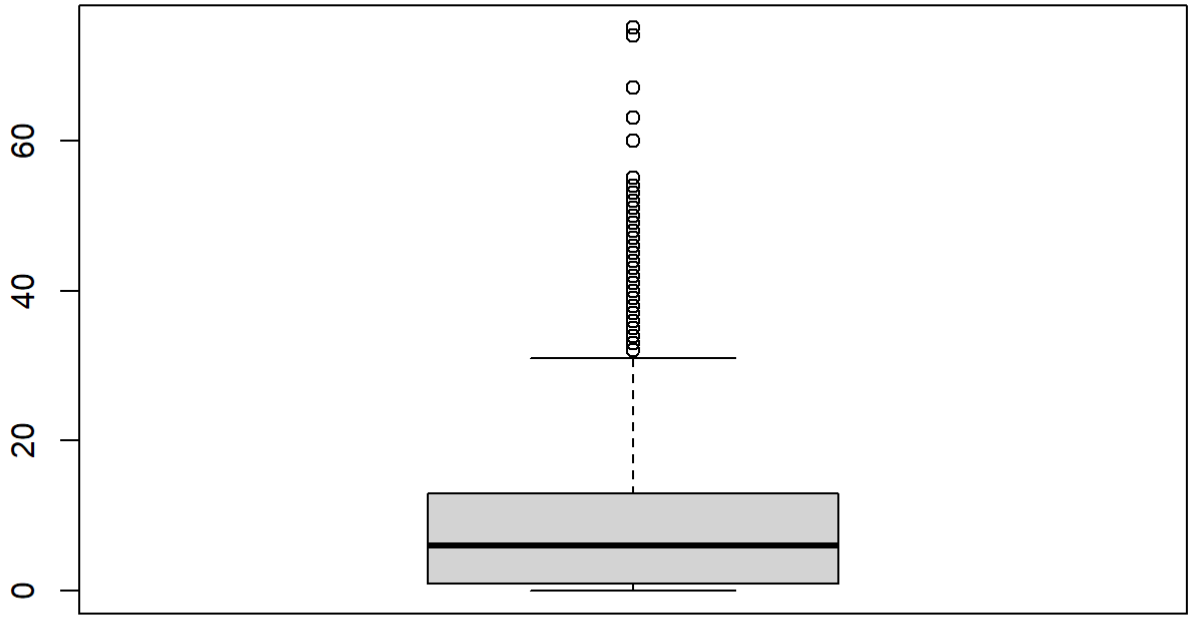
```
shooting$G_xG <- as.numeric(shooting$G_xG)

class(shooting$npxG)
```

```
## [1] "character"
```

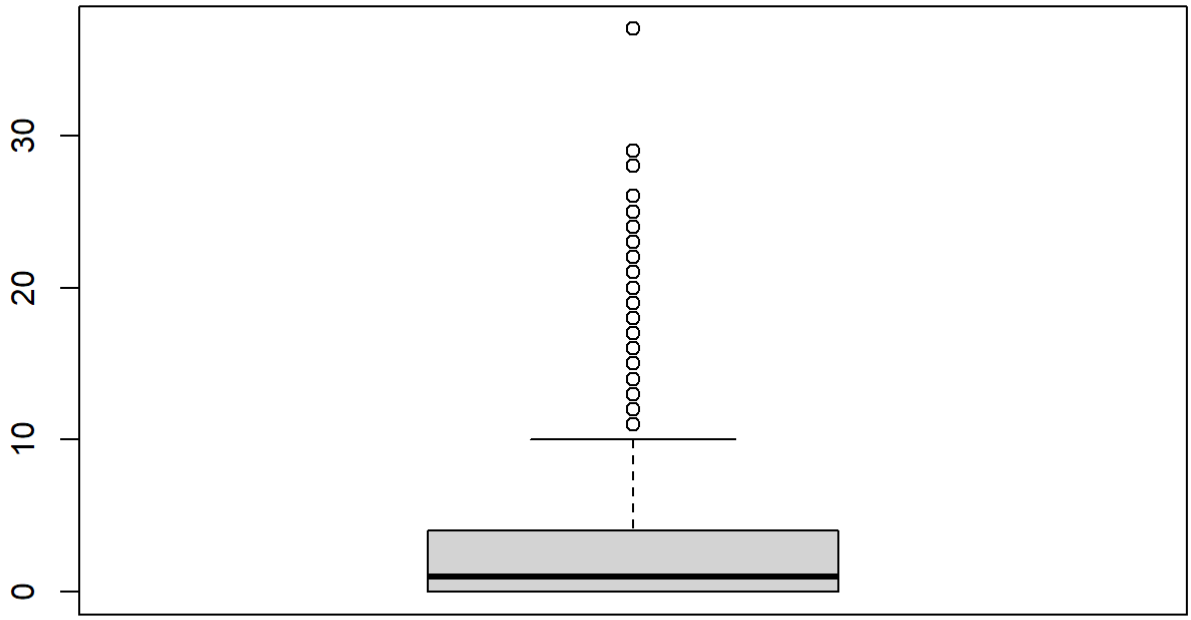
```
shooting$npxG <- as.numeric(shooting$npxG)

write.csv(shooting, "shooting_clean.csv", row.names=FALSE)


#Valors atípics amb boxplot: si no s'aplica cap segmentació és impossible analitzar els resul
tat, doncs lògicament els porters o els defenses hauran realitzat molts menys tirs que els da
vanters.

boxplot(shooting$Gls)
```



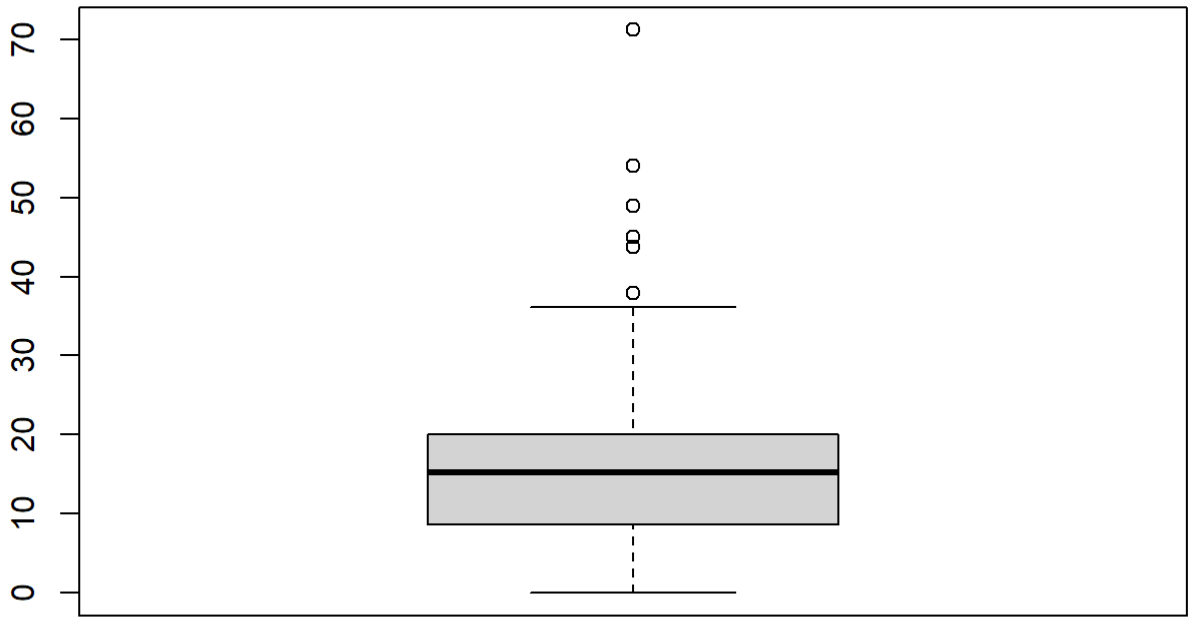

```
boxplot(shooting$Sh)
```



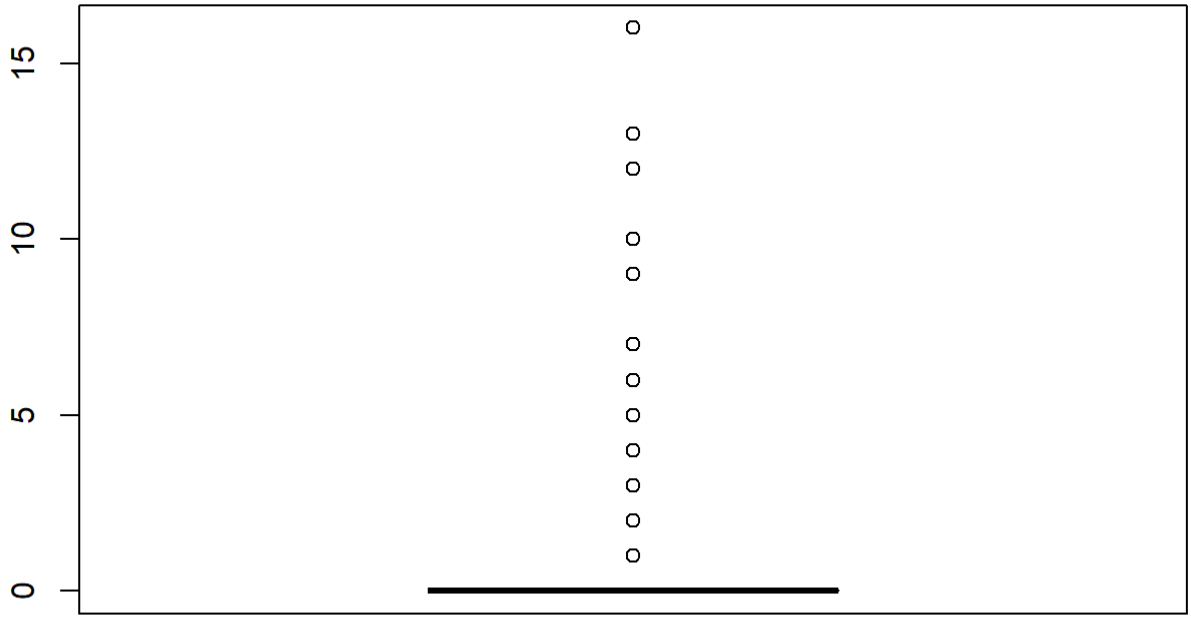

```
boxplot(shooting$SoT)
```



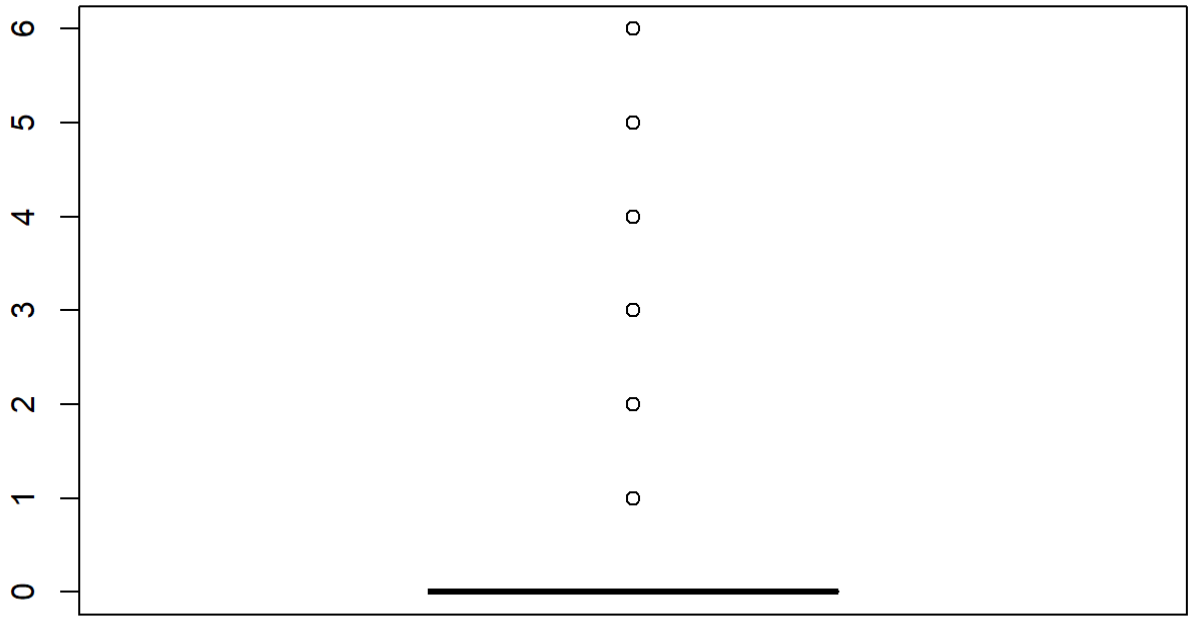
```
boxplot(shooting$Dist)
```



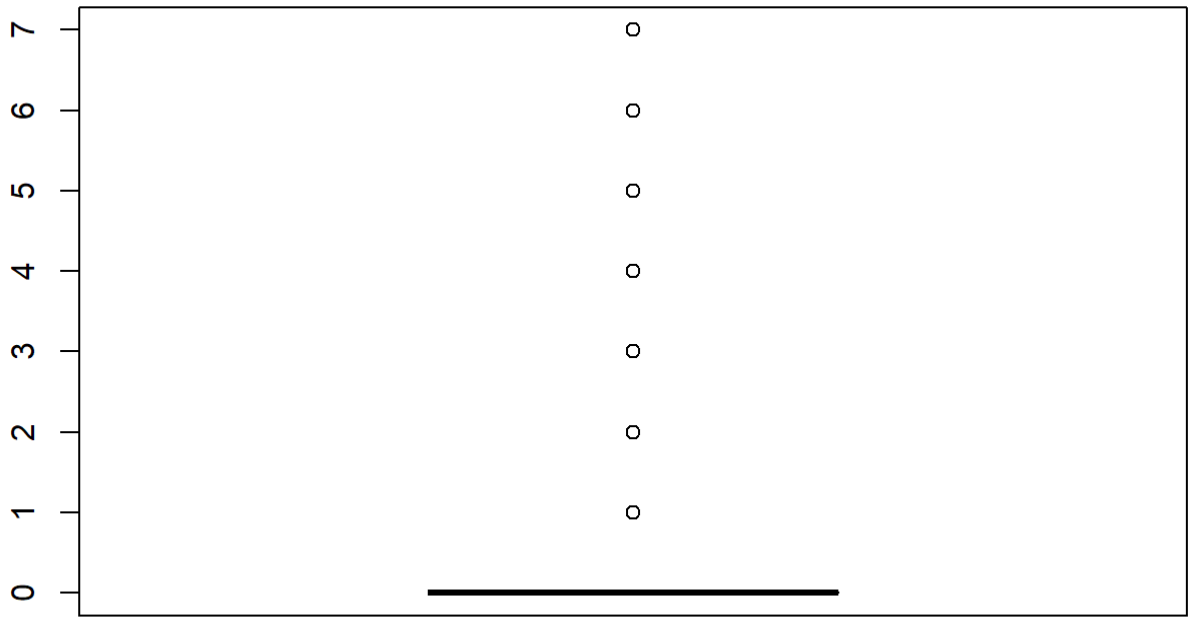

```
boxplot(shooting$FK)
```



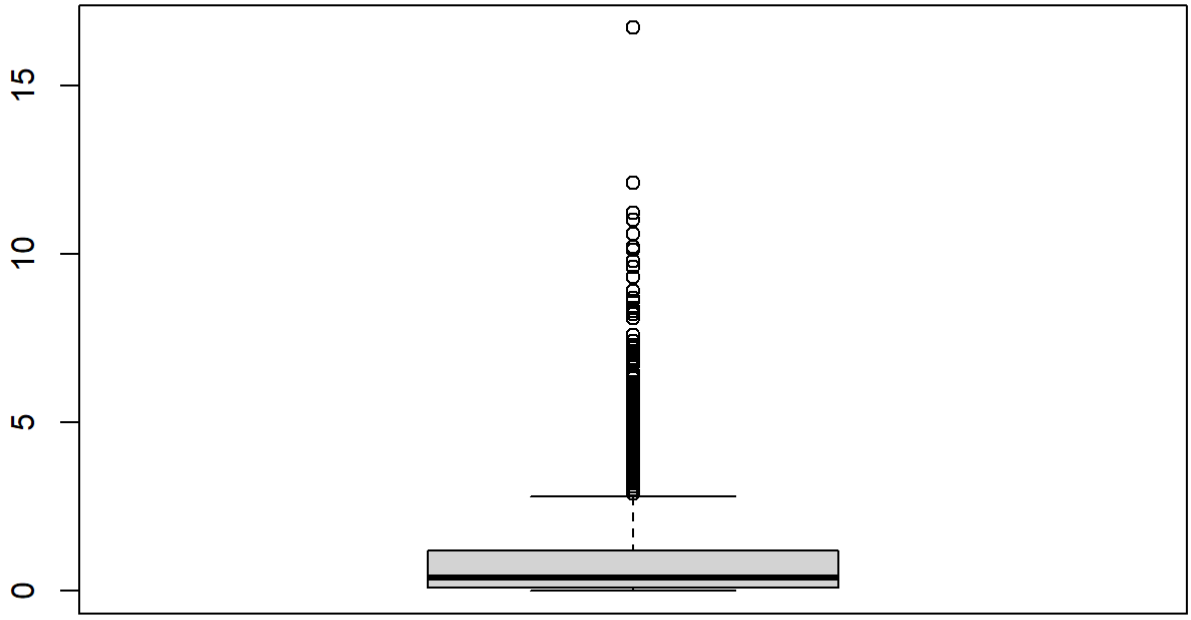

```
boxplot(shooting$PK)
```



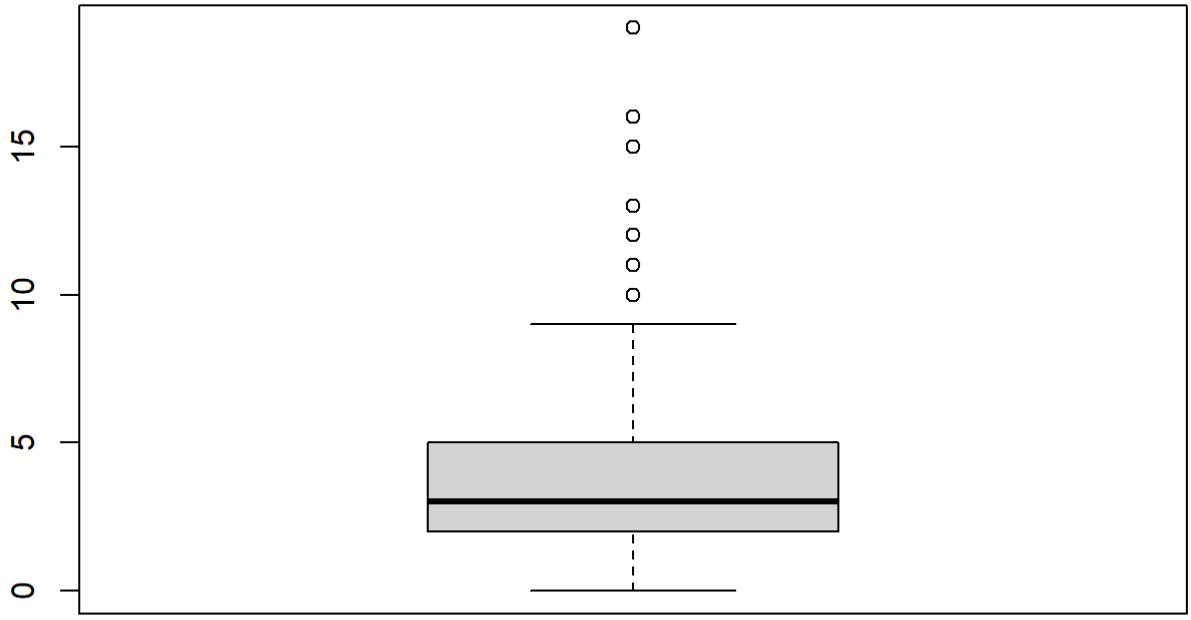
```
boxplot(shooting$PKatt)
```



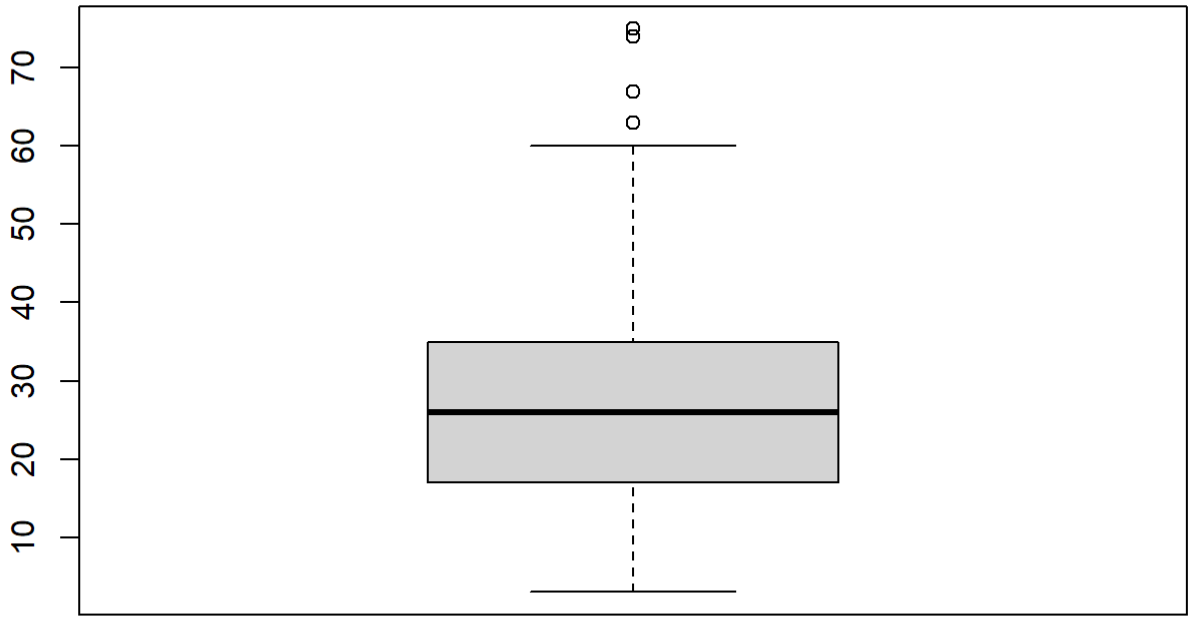

```
boxplot(shooting$xG)
```



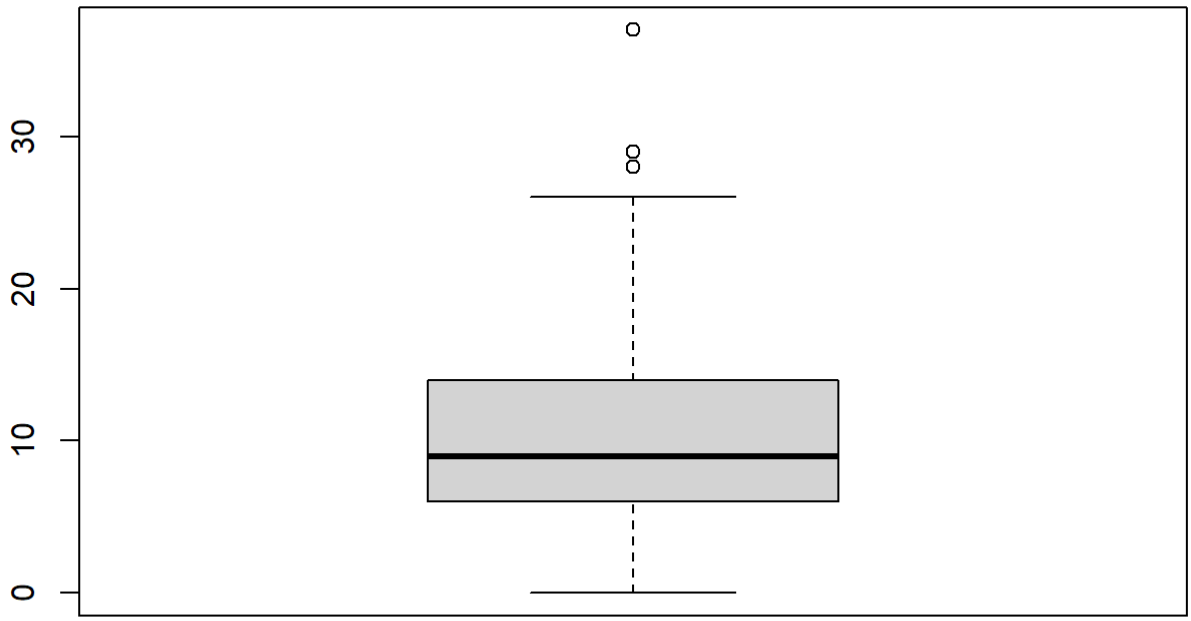

```
#Es generen filtres per posició principal i s'analitzen els resultats. Tot i les diferències,
no hi ha cap valor que s'estimi com a invàlid, doncs està dins del rang normal.

#Analitzant els davanters, que tenen els valors més alts de tirs i gols.

fwd <- filter(shooting, shooting$Pos_main=='FW' & shooting$noventas >=6)
boxplot(fwd$Gls)
```



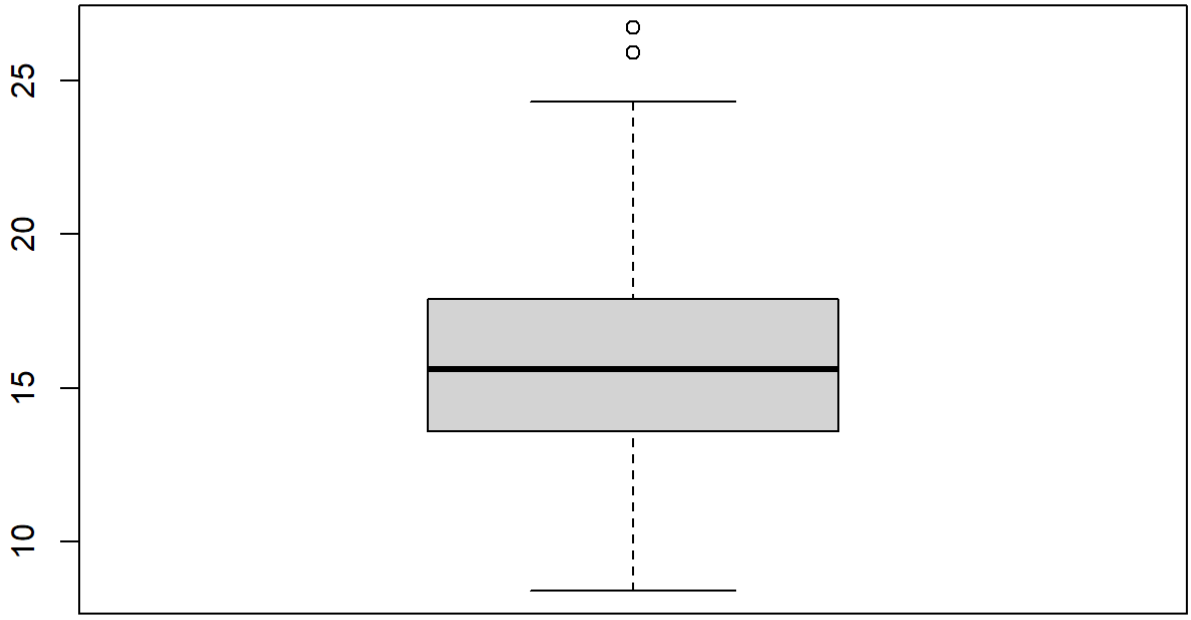

```
boxplot(fwd$Sh)
```



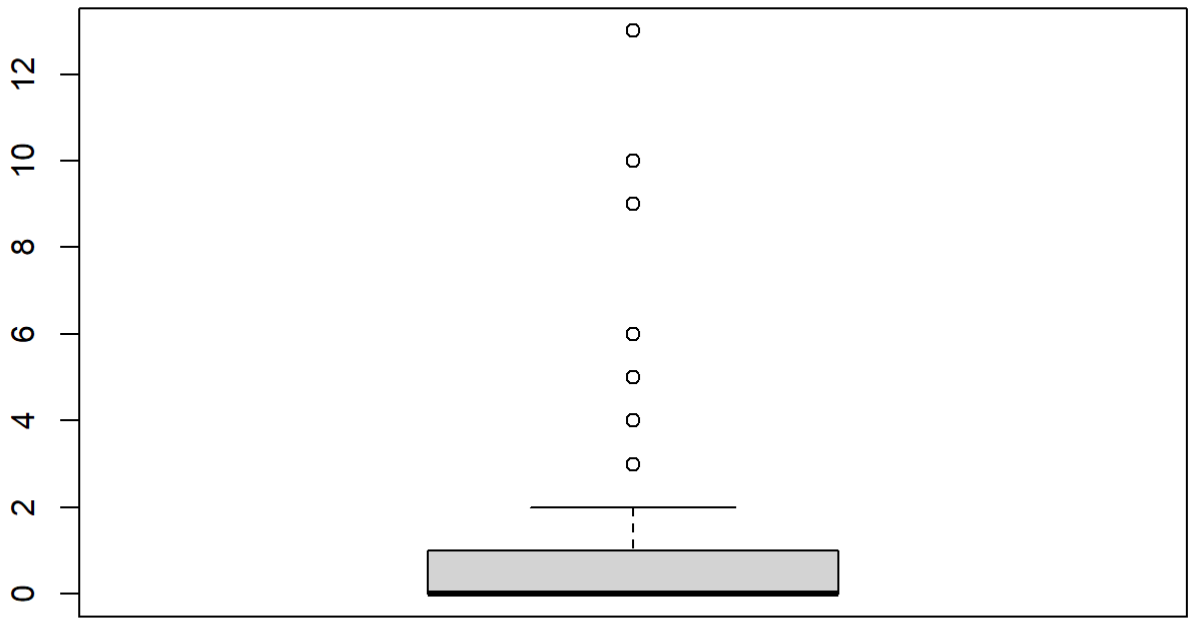
```
boxplot(fwd$SoT)
```



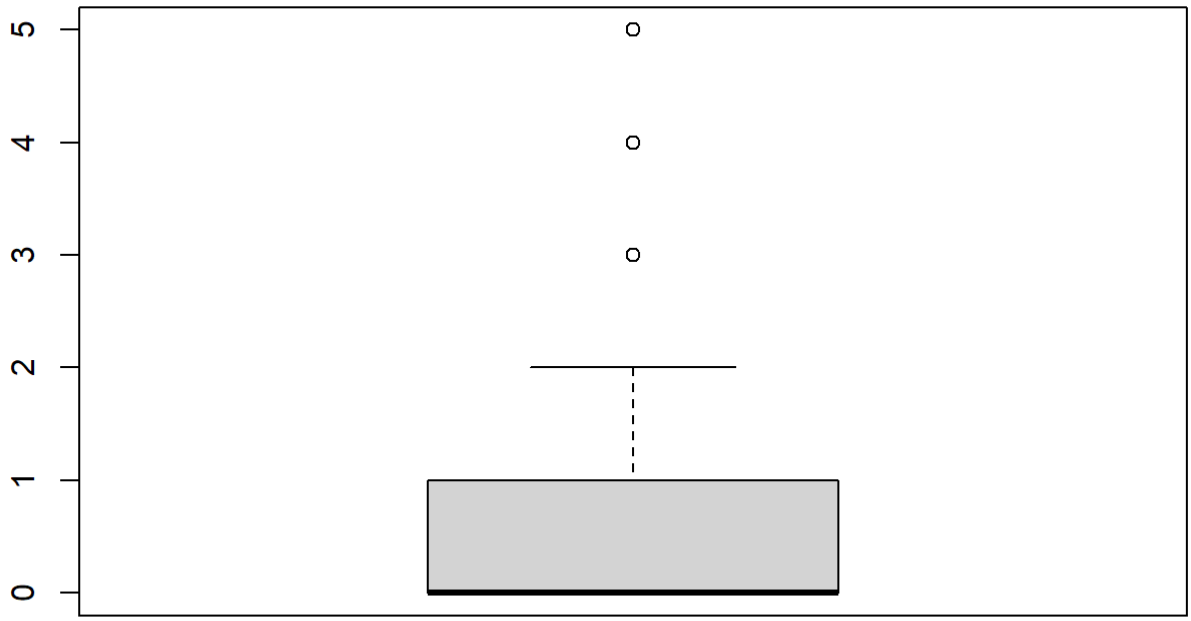

```
boxplot(fwd$Dist)
```



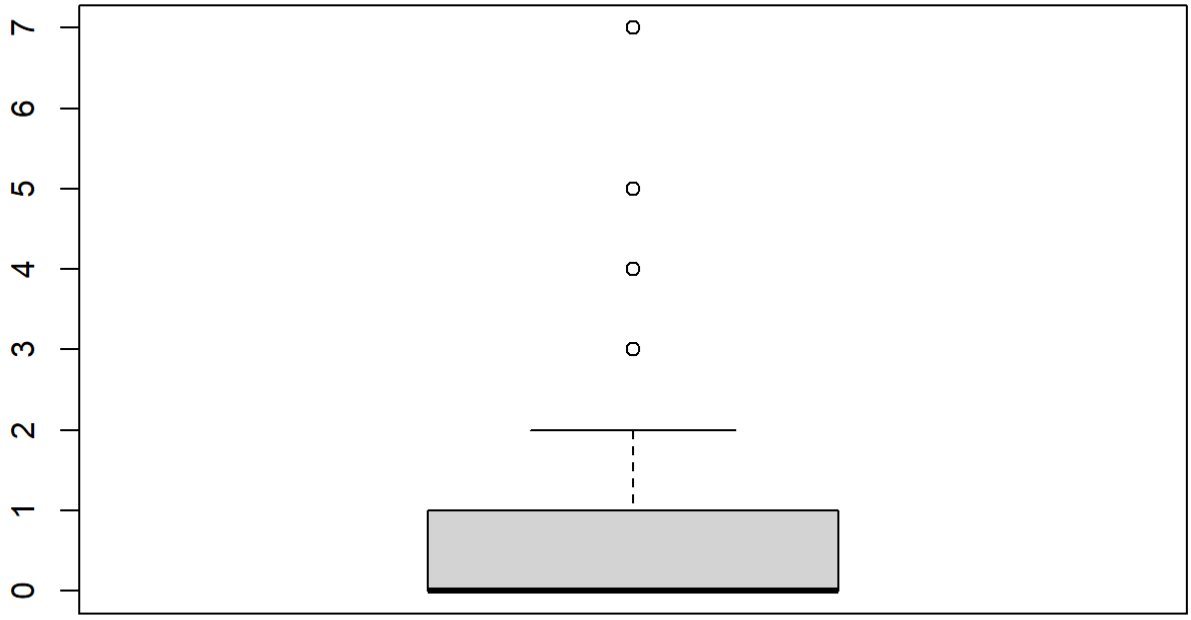

```
boxplot(fwd$FK)
```



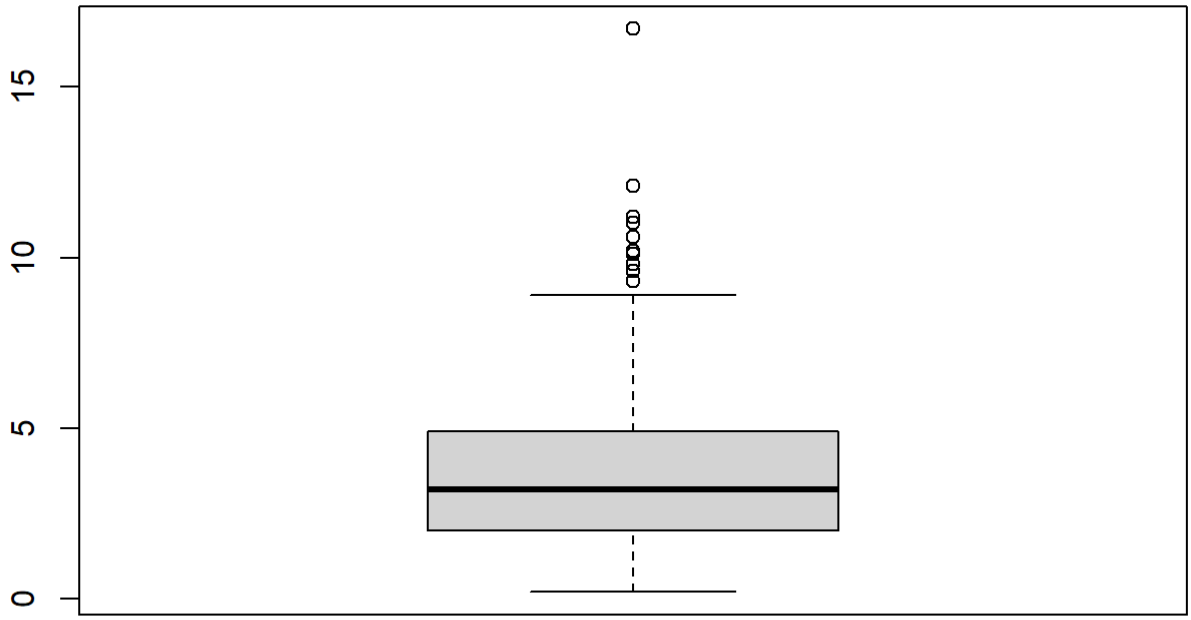
```
boxplot(fwd$PK)
```



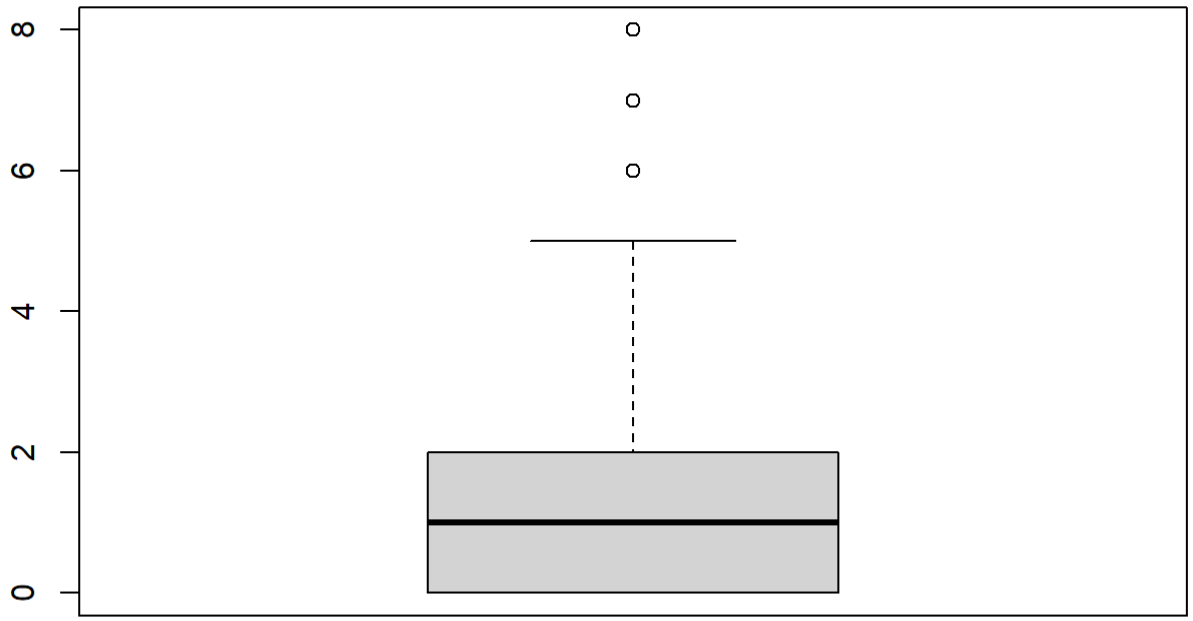

```
boxplot(fwd$PKatt)
```



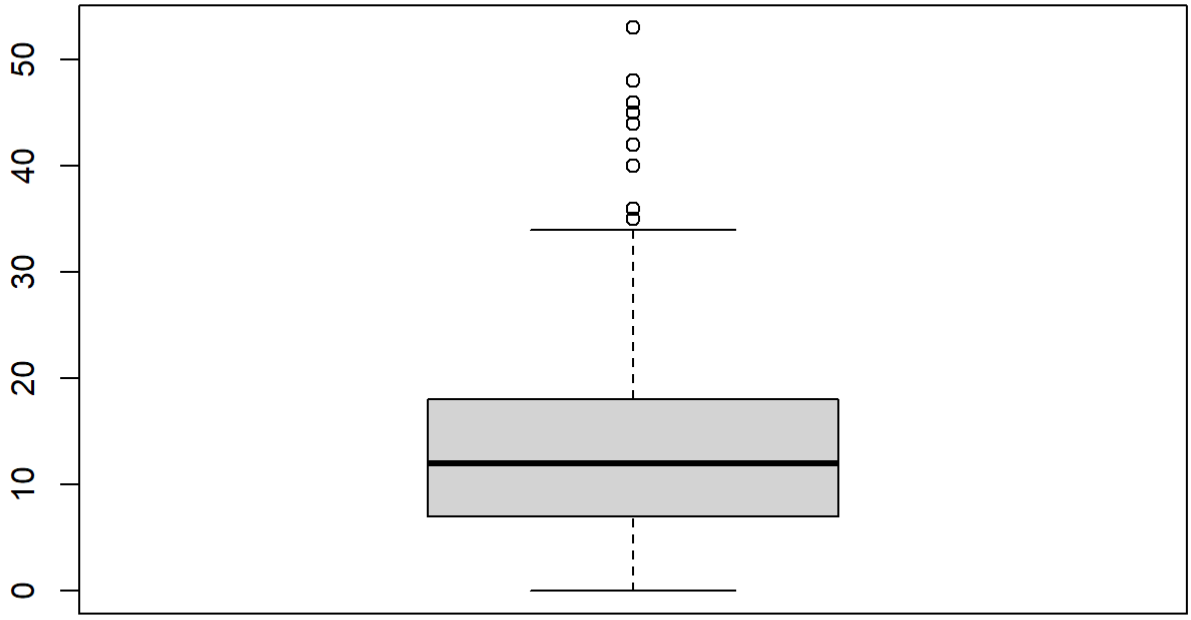

```
boxplot(fwd$xG)
```



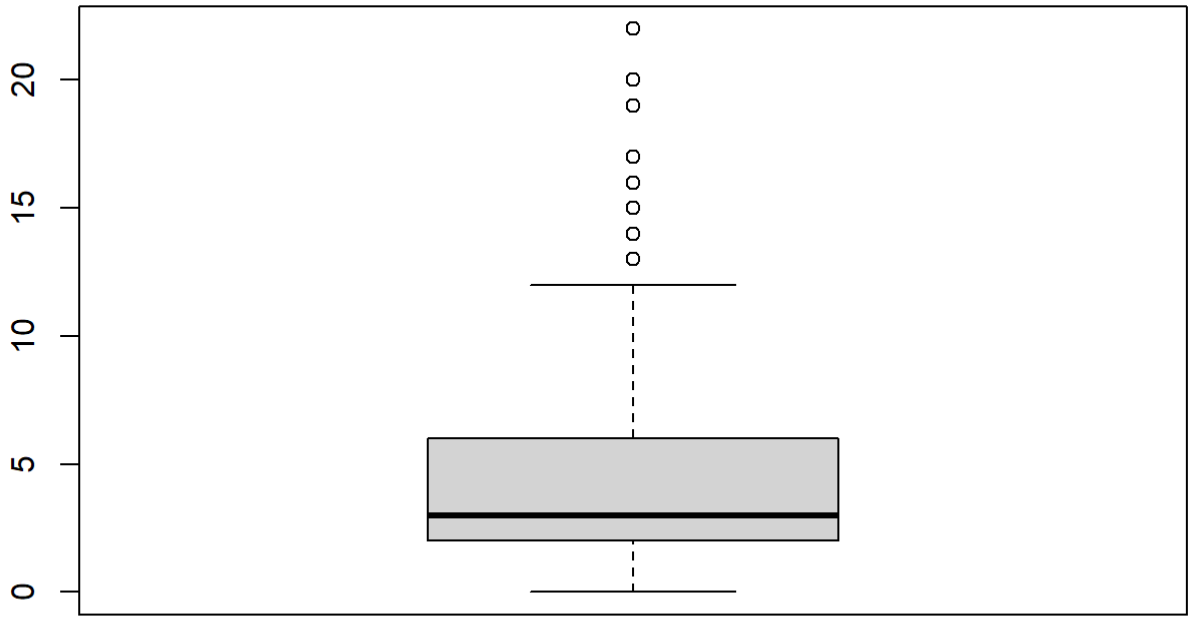
```
#Analitzant els centrecampistes

mf <- filter(shooting, shooting$Pos_main=='MF' & shooting$noventas >=6)
boxplot(mf$Gls)
```



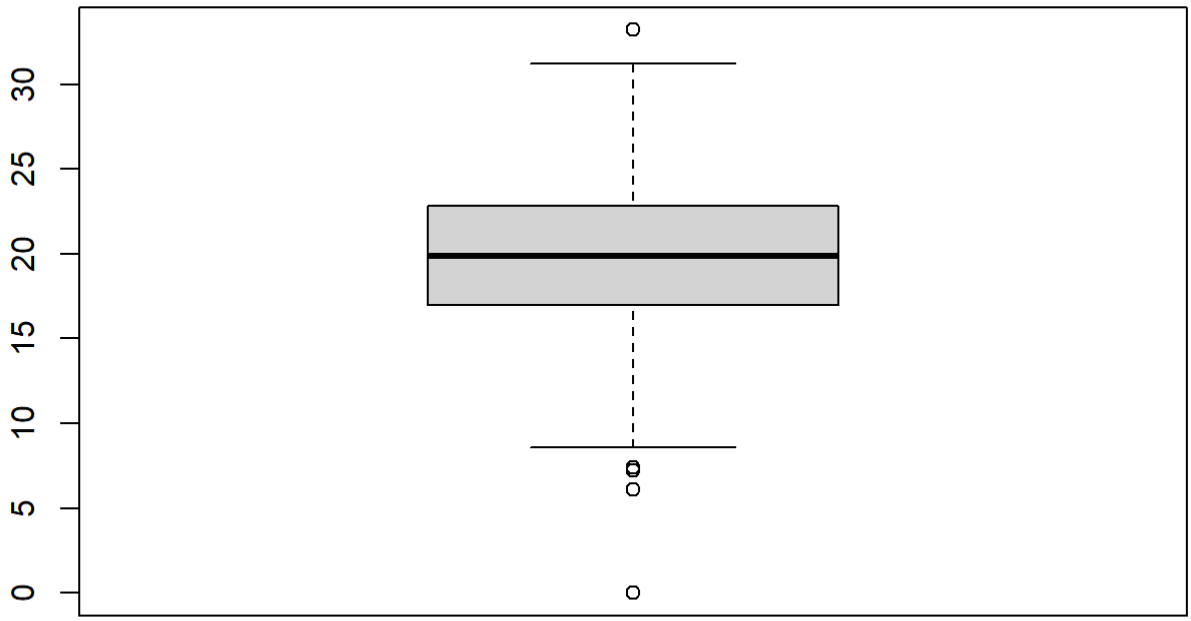

```
boxplot(mf$Sh)
```



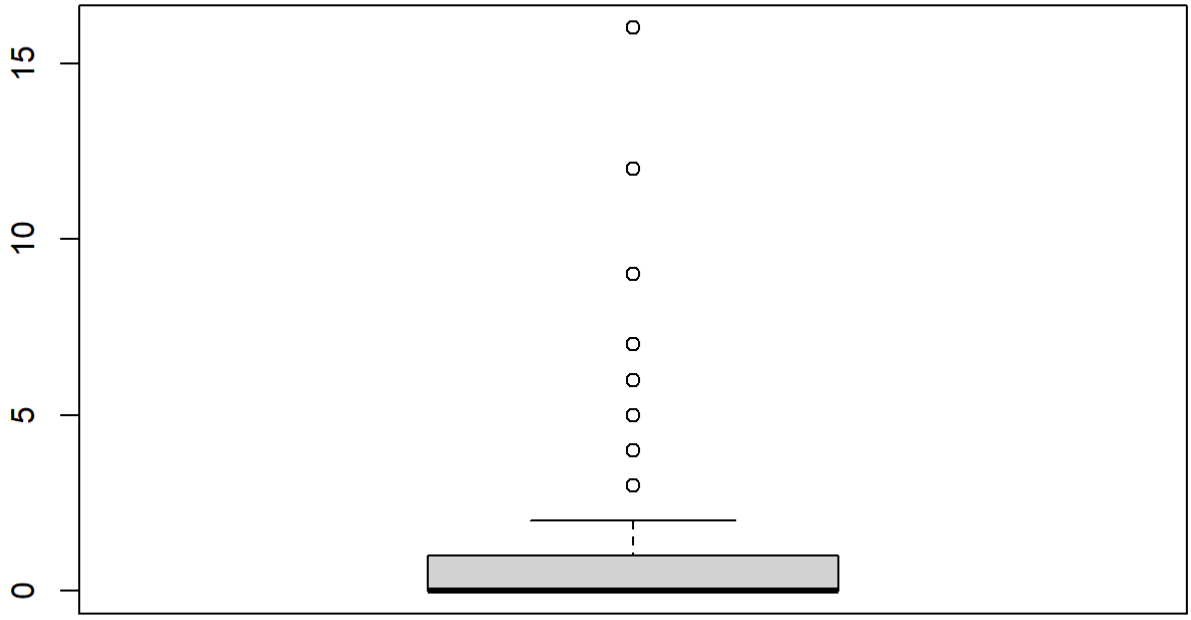

```
boxplot(mf$SoT)
```



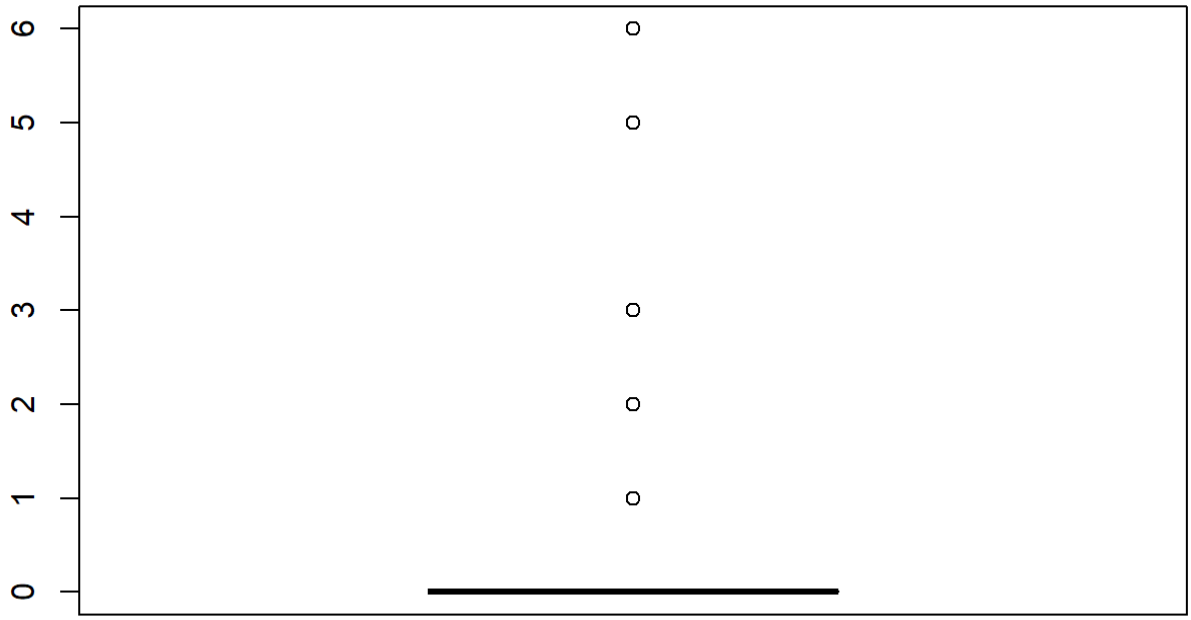
```
boxplot(mf$Dist)
```



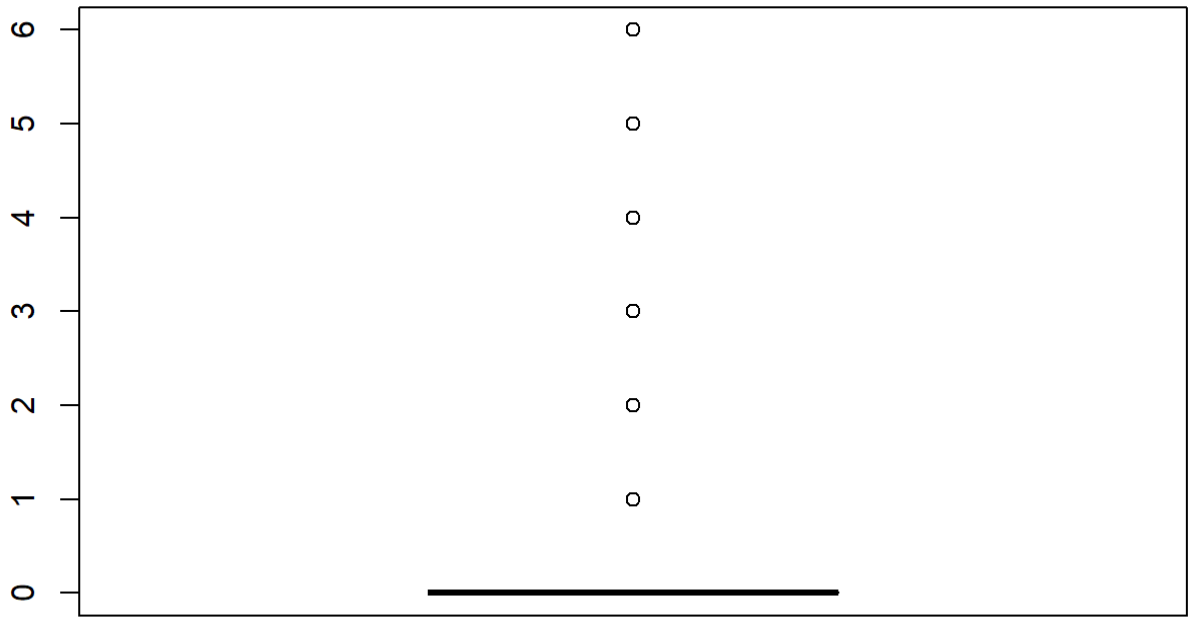

```
boxplot(mf$FK)
```



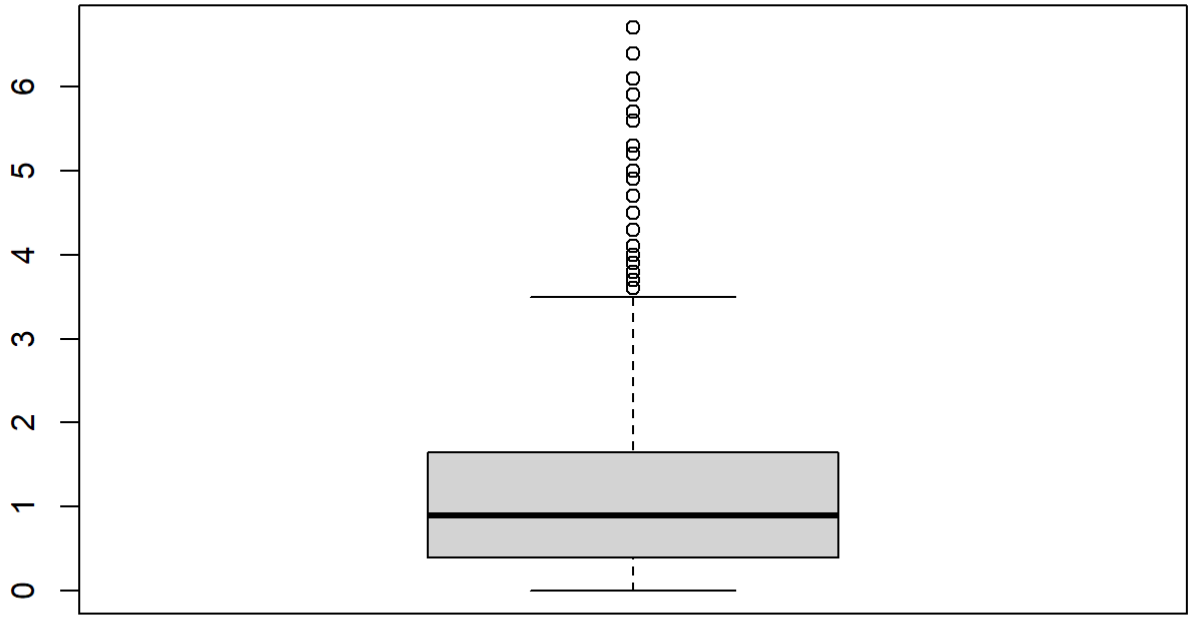

```
boxplot(mf$PK)
```



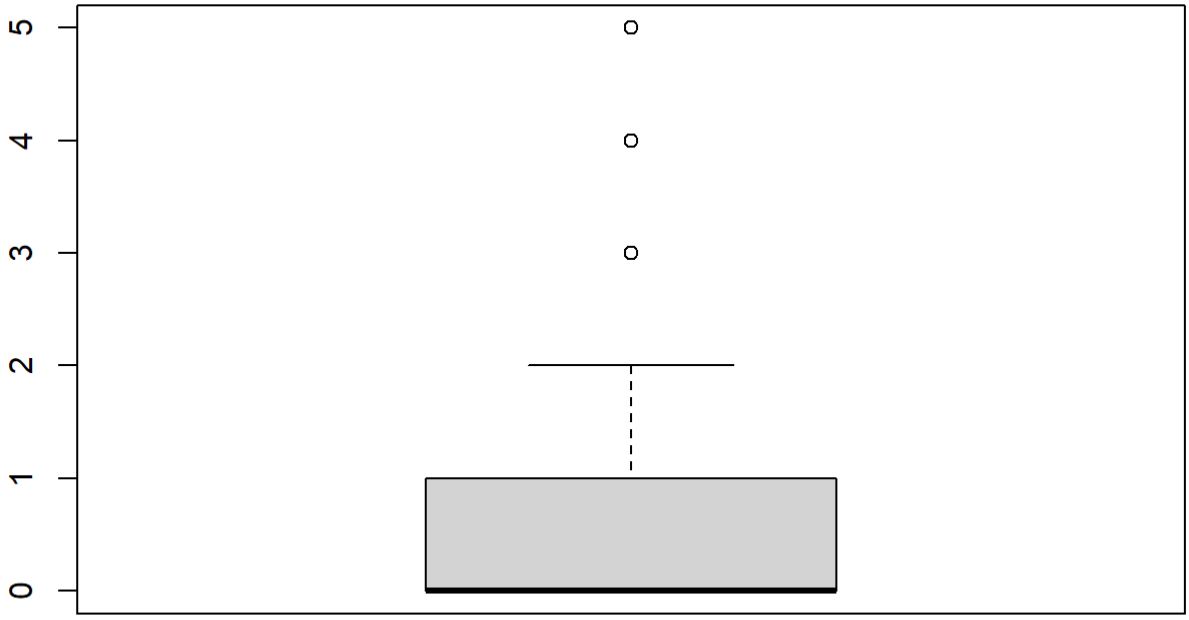
```
boxplot(mf$PKatt)
```



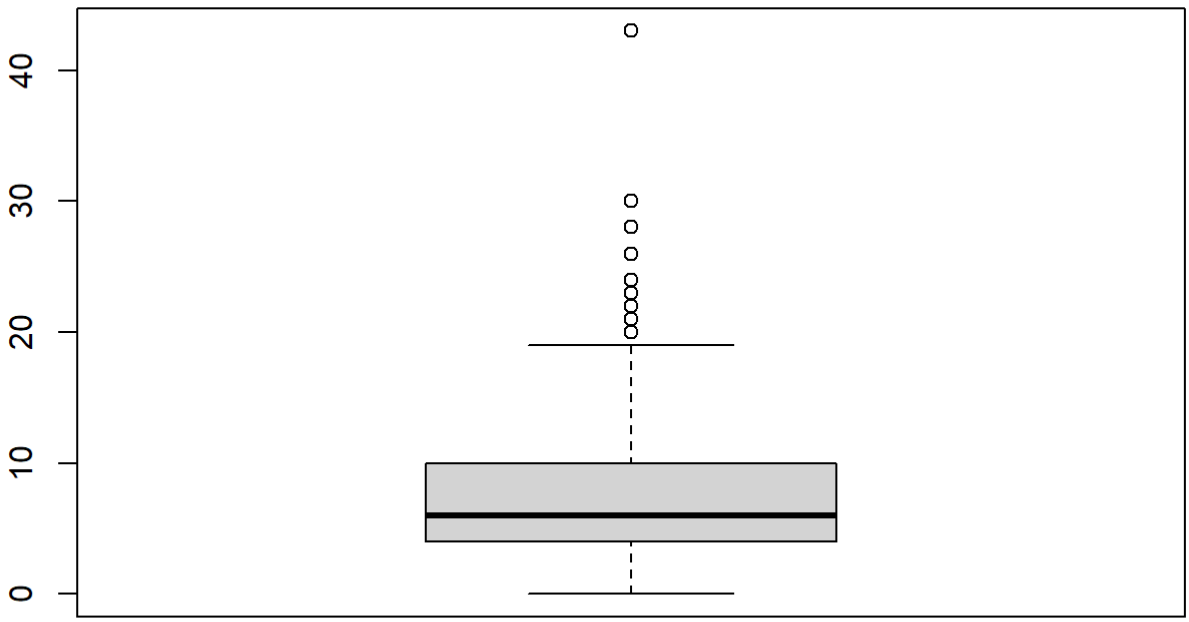

```
boxplot(mf$xG)
```



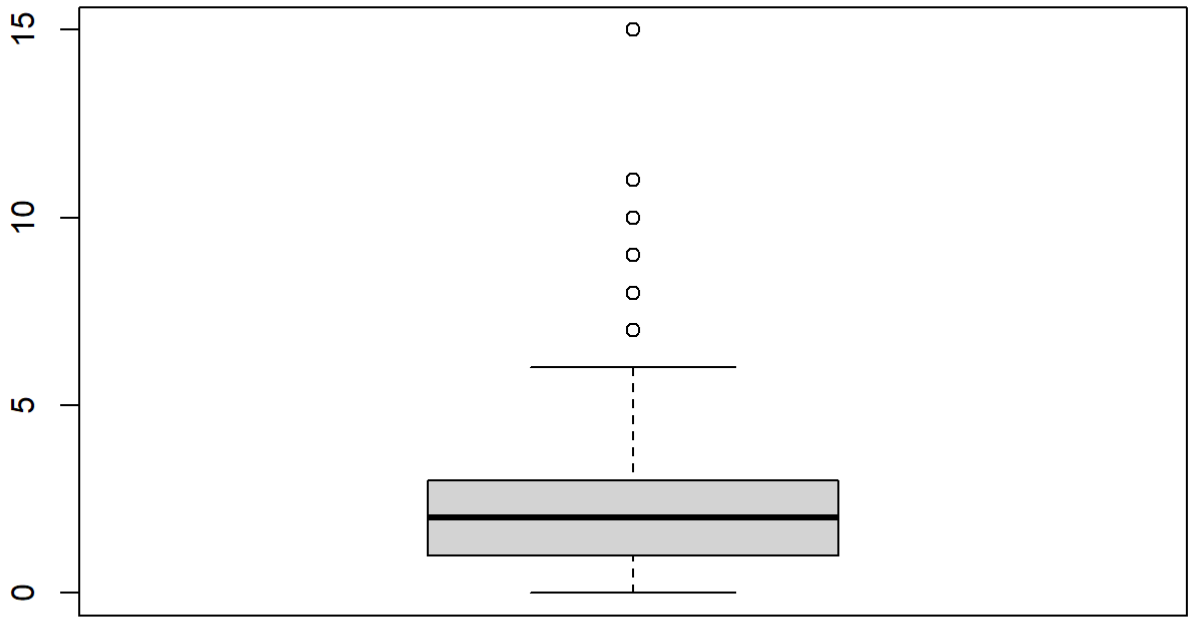

```
#Analitzant els defenses
df <- filter(shooting, shooting$Pos_main=='DF' & shooting$noventas >=6)
boxplot(df$Gls)
```



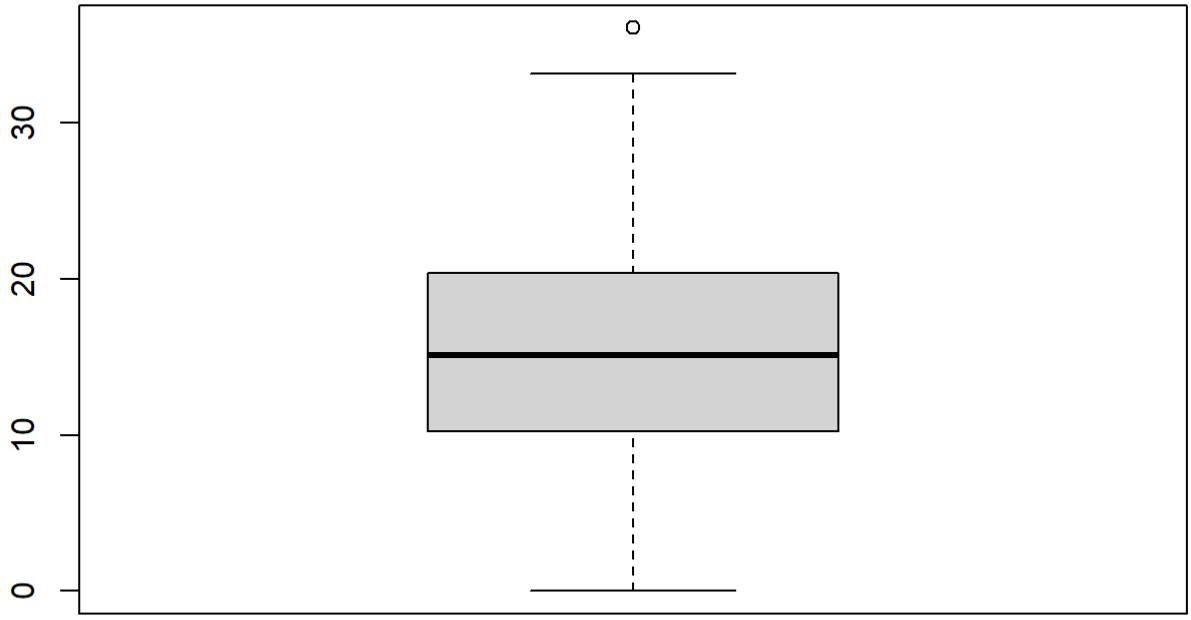

```
boxplot(df$Sh)
```



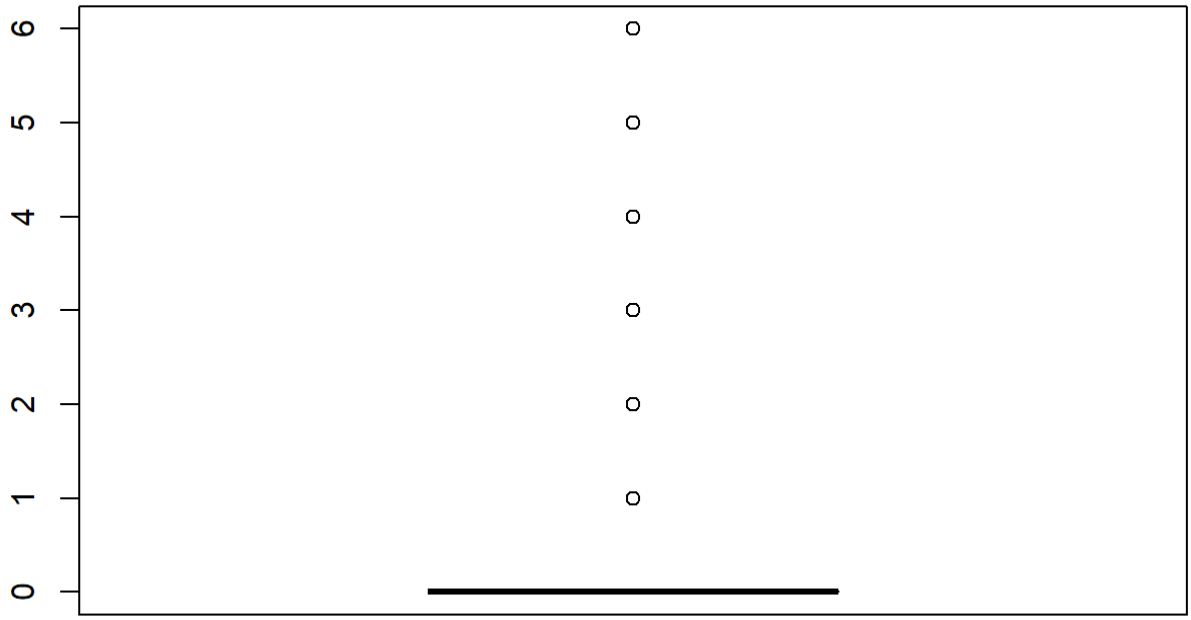
```
boxplot(df$SoT)
```



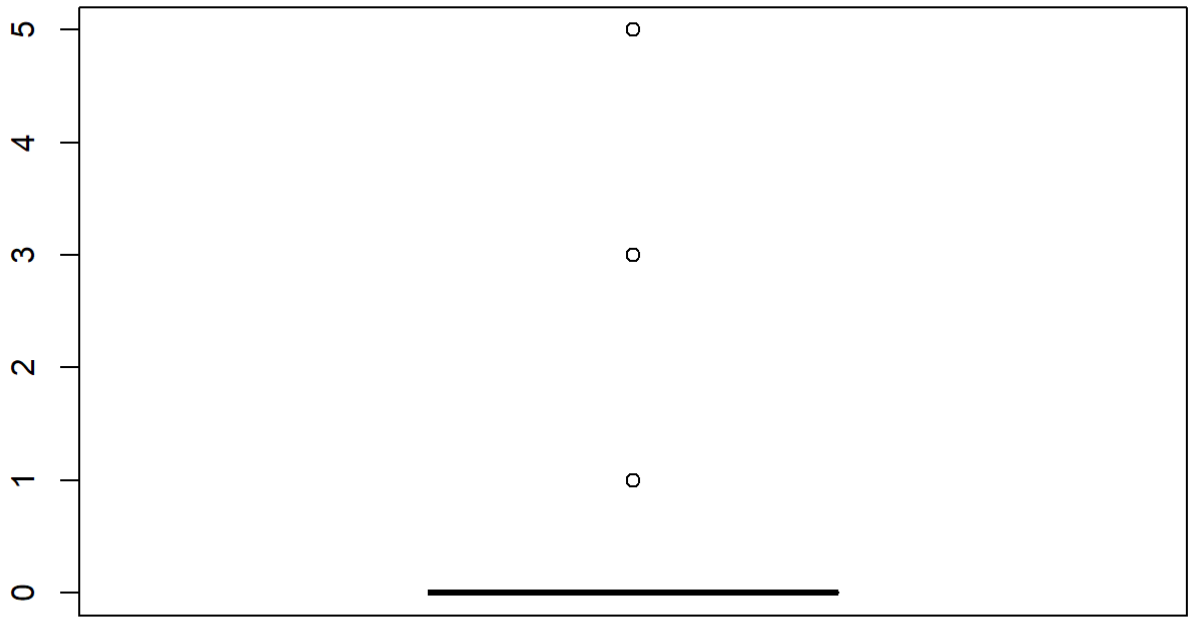

```
boxplot(df$Dist)
```



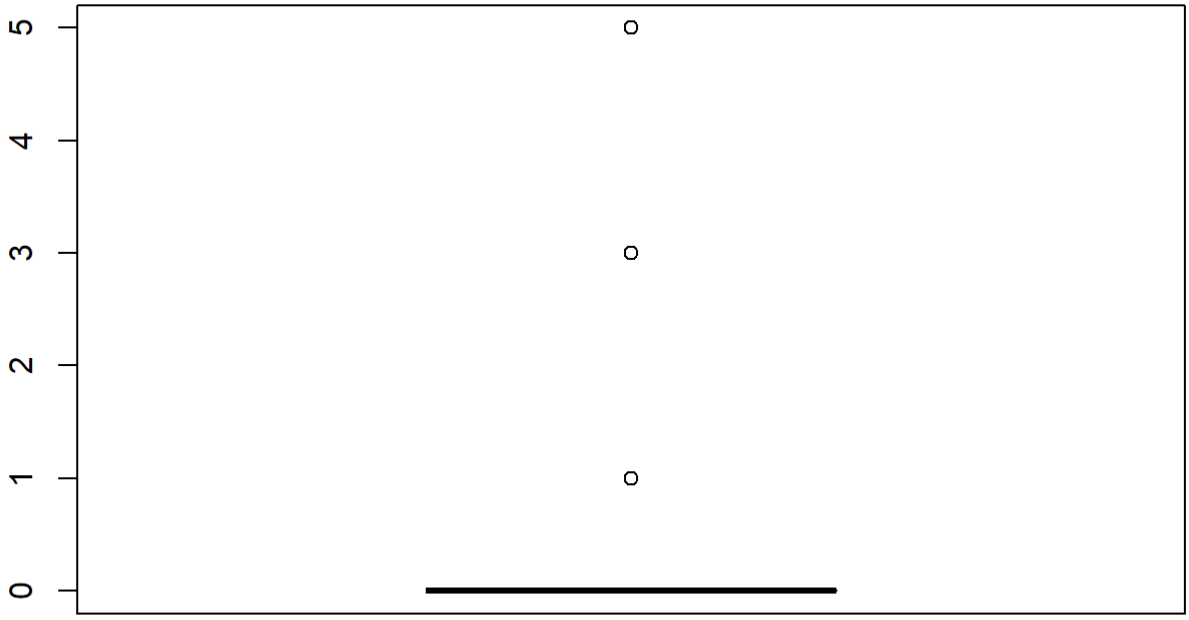

```
boxplot(df$FK)
```



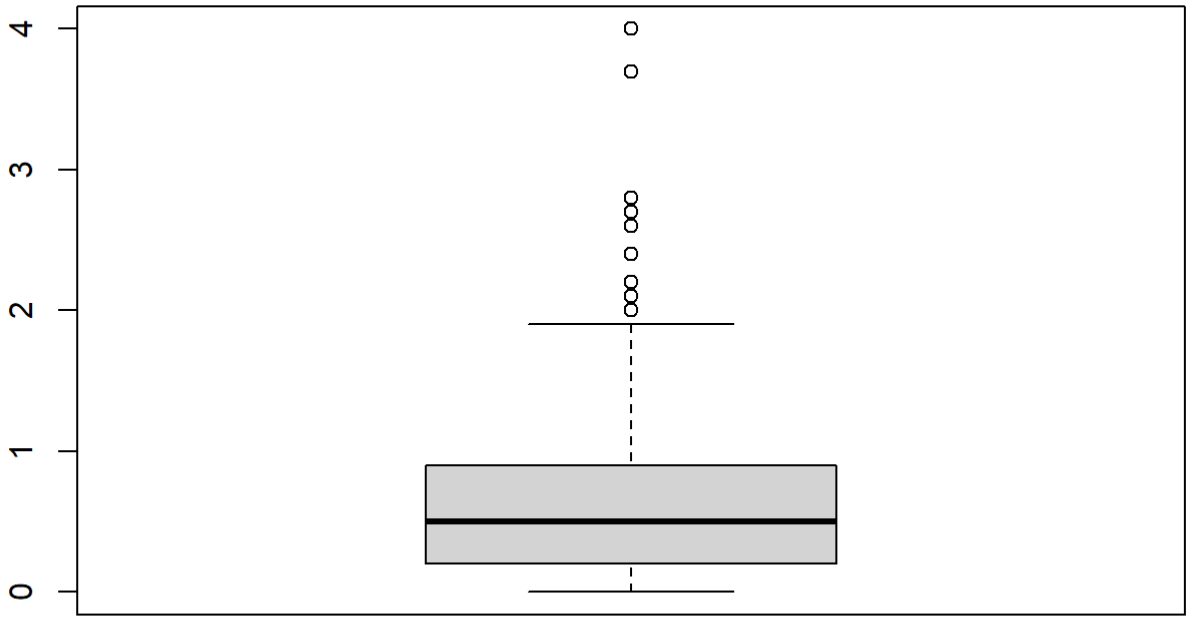
```
boxplot(df$PK)
```



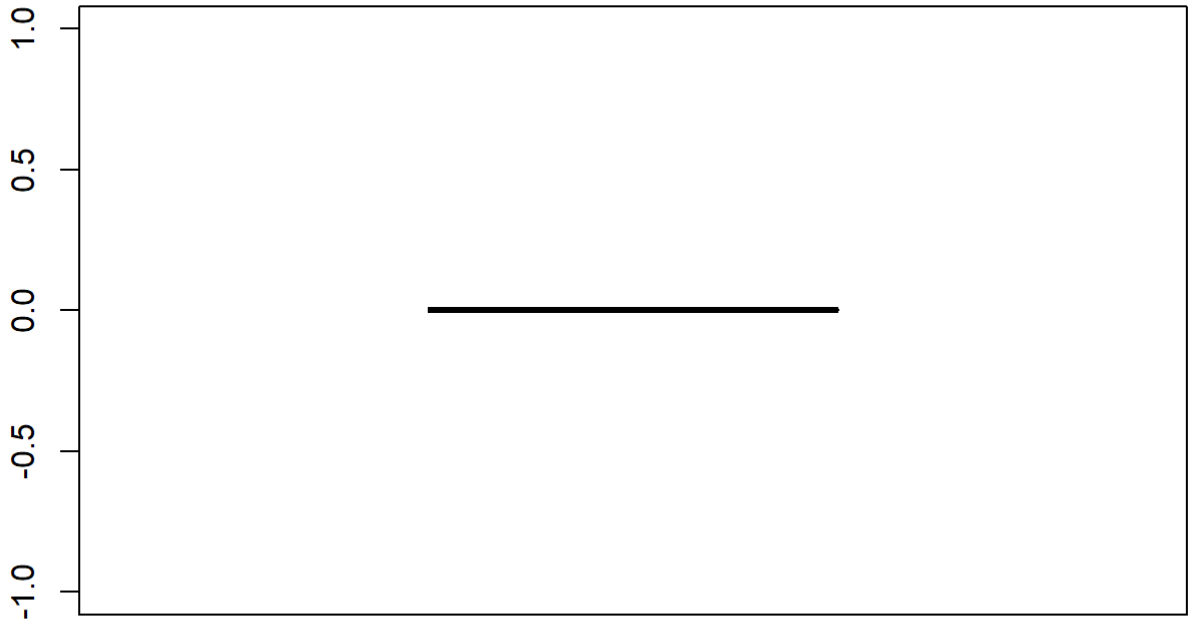

```
boxplot(df$PKatt)
```



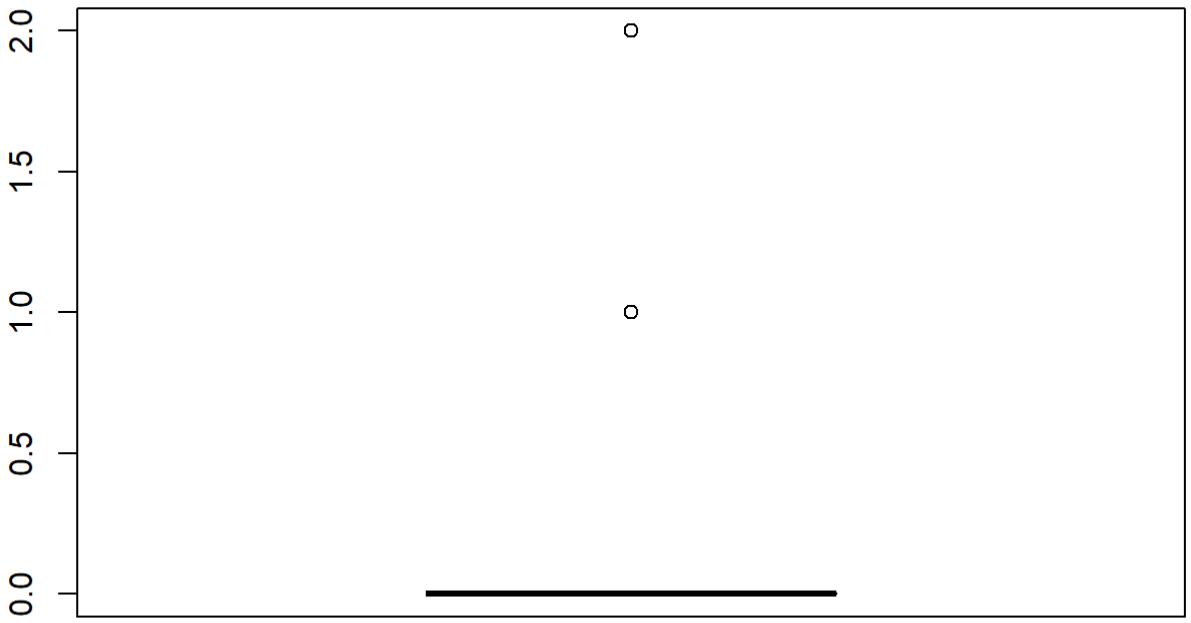

```
boxplot(df$xG)
```



```
#Analitzant els porters, curiós trobar un porter que ha marcat gol.
gk <- filter(shooting, shooting$Pos_main=='GK' & shooting$noventas >=6)
boxplot(gk$Gls)
```

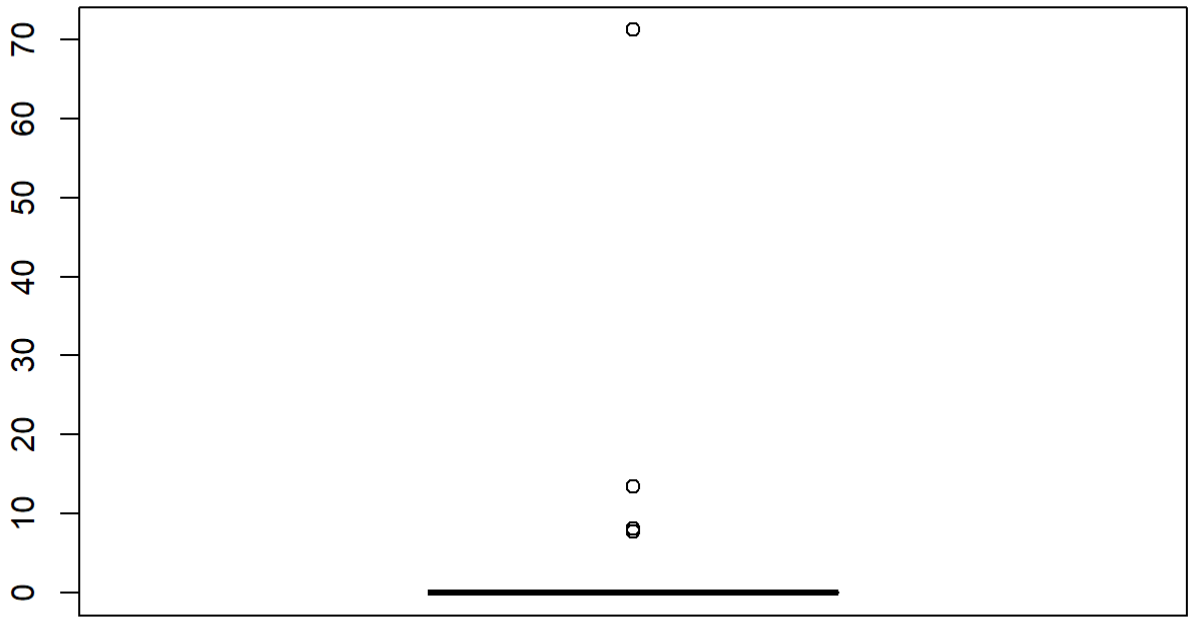



```
boxplot(gk$Sh)
```

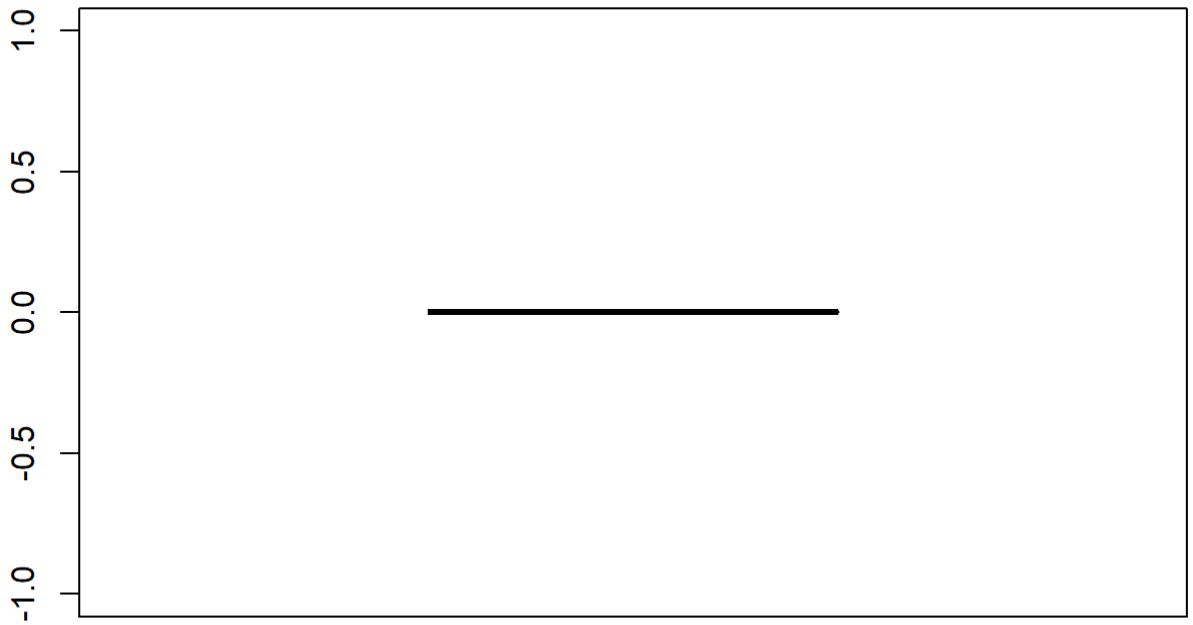



```
boxplot(gk$SoT)
```

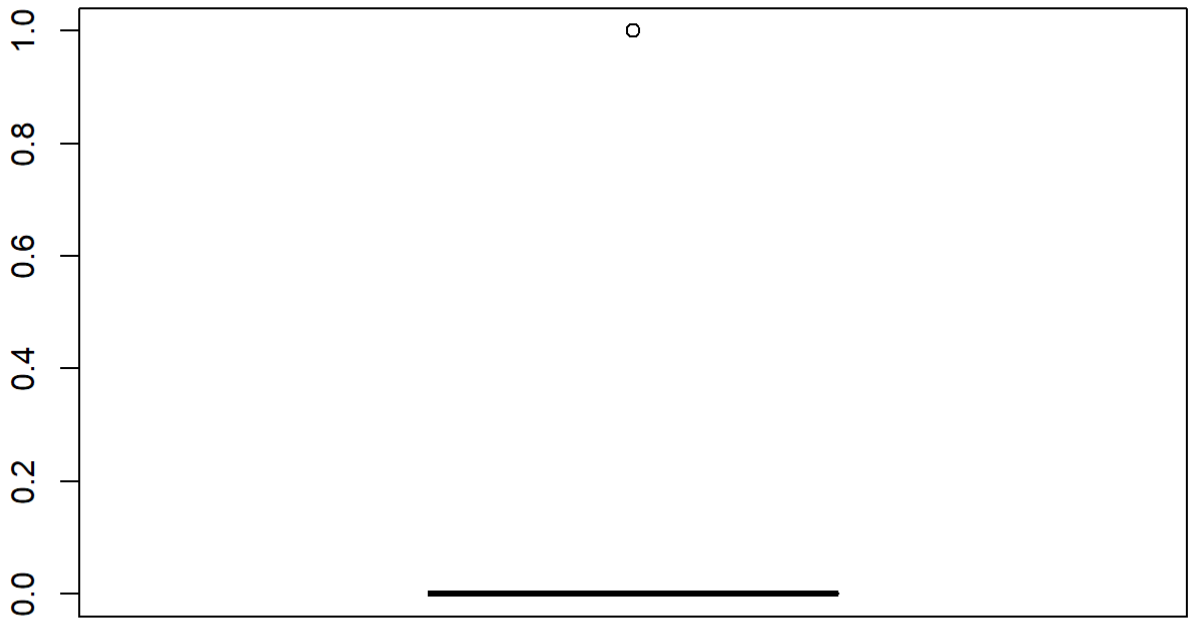
```
boxplot(gk$Dist)
```



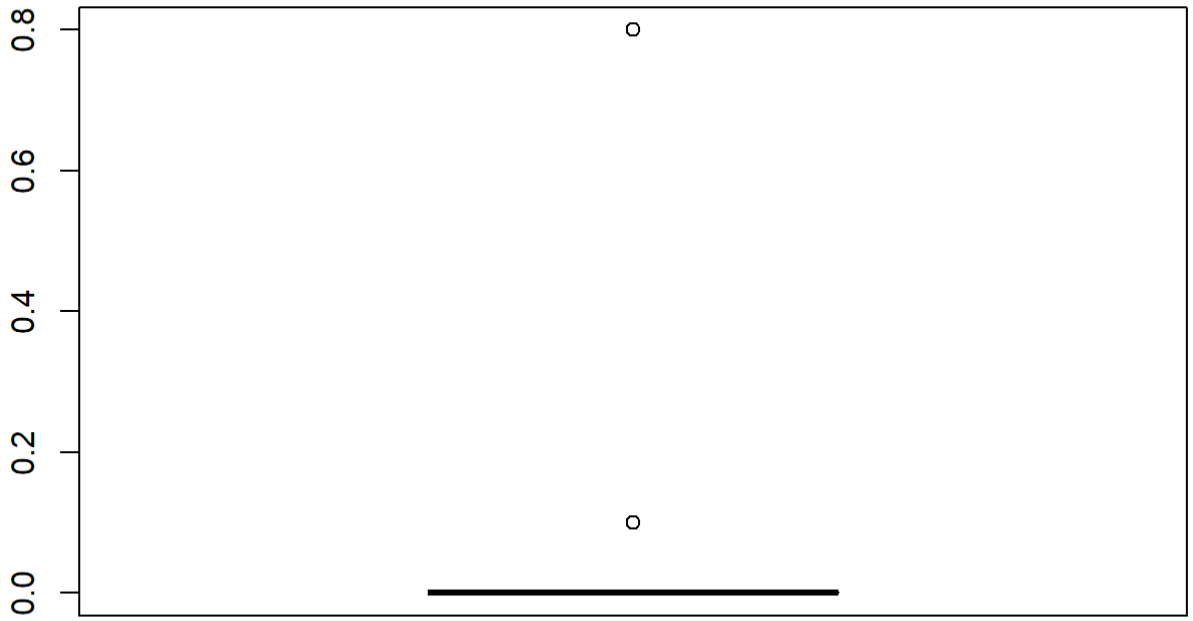

```
boxplot(gk$FK)
boxplot(gk$PK)
```




```
boxplot(gk$PKatt)
```



```
boxplot(gk$xG)
```



2.4 - Anàlisi de les dades

En aquest punt es creen els subdatasets d'estudi, que en aquest cas seran les mostres de davanters de cada competició, doncs com és un anàlisi centrat en la finalització a porteria, els que tindran més informació seran els davanters.

En particular, les mètriques que es consideren més interessant per realitzar les comparacions són: número de gols, número de gols esperats (amb i sense incloure els penaltis), número de tirs i les mètriques calculades a partir d'aquestes, G-xG, xG_Sh i npxG_Sh.

Es comprova s hi ha normalitat amb el test de Lilliefors, que demostra que no tenim normalitat en moltes de les mètriques, que és lògic pensant en que es tracta de mètriques que tindran moltes més repeticions de valors baixos que no d'alts. Es comprova també si hi ha igualtat de variàncies amb el test de Bartlett, que en el nostre cas és que sí.

Per a realitzar el test d'hipòtesis sobre les mitjanes s'escull la mètrica G-xG, que si que té normalitat en totes les mostres.

A més a més, es realitza un model lineal de variables quantitatives. El model lineal generat entre la variable npxG/Sh i la distància demostra no ser significatiu, com tampoc ho és amb la variable Sh.

Per altra banda, s'analitza si hi ha correlació entre parelles de mètriques, obtenint els valors de correlació mostrar, r, més alts en parelles de mètriques bastant intuitives. Per exemple, hi ha alta correlació entre el número de tirs i el número de gols esperats i gols realitzats, com és lògic.

Per últim també es genera un model de regressió múltiple amb valors quantitatius i qualitatius, intentant analitzar si la competició, la posició o l'edat dels jugadors és un element important a l'hora de predir el resultat de la mètrica npxG/Sh, i resulta en què la posició dels jugadors que són centrecampistes i davanters és important, així com la distància, el número de tirs i el número de xG esperats. No obstant, el valor de Rsquared és baix, amb la qual cosa el model no s'ajusta del tot bé i s'hauria de millorar.

Es mostren els dos gràfics, valors ajustats enfront dels residus i el gràfic quantil-quantil que compara els residus del model amb els valors d'una variable que es distribueix normalment.

S'observa com hi ha molts residus que es concentren entre els valors 5 i -5, però també existeix una quantitat considerable d'outliers.

També es veu que quan els valors ajustats es disparen, també s'incrementa el valor dels residus.

Amb el gràfic QQ es veu com la part central s'ajusta perfectament a la recta però els extrems es separen molt, el què vol dir que hi ha més valors extrems dels esperats si la distribució fos realment Normal.

```
##Creació dels subsets de dades per cada competició
fwd_laliga <- filter(fwd, fwd$Competition=='La Liga')
fwd_seriea <- filter(fwd, fwd$Competition=='Serie A')
fwd_ligue1 <- filter(fwd, fwd$Competition=='Ligue 1')
fwd_premier <- filter(fwd, fwd$Competition=='Premier League')
fwd_bundesliga <- filter(fwd, fwd$Competition=='Bundesliga')



##Anàlisi de normalitat de les mètriques: s'observa com no hi ha normalitat
#Gls
lillie.test(fwd$Gls)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd$Gls
## D = 0.13818, p-value = 7.028e-16
```

```
lillie.test(fwd_laliga$Gls)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_laliga$Gls
## D = 0.14793, p-value = 0.002916
```

```
lillie.test(fwd_ligue1$Gls)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_ligue1$Gls
## D = 0.1408, p-value = 0.001502
```

```
lillie.test(fwd_premier$Gls)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_premier$Gls
## D = 0.13714, p-value = 0.00491
```

```
lillie.test(fwd_seriea$Gls)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_seriea$Gls
## D = 0.15914, p-value = 0.0001449
```

```
lillie.test(fwd_bundesliga$Gls)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_bundesliga$Gls
## D = 0.19169, p-value = 4.36e-05
```

```
#xG
lillie.test(fwd$xG)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd$xG
## D = 0.12448, p-value = 9.743e-13
```

```
lillie.test(fwd_laliga$xG)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_laliga$xG
## D = 0.11633, p-value = 0.04899
```

```
lillie.test(fwd_ligue1$xG)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_ligue1$xG
## D = 0.11154, p-value = 0.03078
```

```
lillie.test(fwd_premier$xG)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_premier$xG
## D = 0.12475, p-value = 0.01623
```

```
lillie.test(fwd_seriea$xG)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_seriea$xG
## D = 0.20167, p-value = 1.704e-07
```

```
lillie.test(fwd_bundesliga$xG)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_bundesliga$xG
## D = 0.13764, p-value = 0.01374
```

```
#npxG
lillie.test(fwd$npxG)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd$npxG
## D = 0.13296, p-value = 1.221e-14
```

```
lillie.test(fwd_laliga$npxG)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_laliga$npxG
## D = 0.12037, p-value = 0.03581
```

```
lillie.test(fwd_ligue1$npxG)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_ligue1$npxG
## D = 0.12086, p-value = 0.01294
```

```
lillie.test(fwd_premier$npxG)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_premier$npxG
## D = 0.1352, p-value = 0.005981
```

```
lillie.test(fwd_seriea$npxG)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_seriea$npxG
## D = 0.15166, p-value = 0.0003919
```

```
lillie.test(fwd_bundesliga$npxG)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_bundesliga$npxG
## D = 0.16476, p-value = 0.001019
```

```
#Sh
lillie.test(fwd$Sh)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd$Sh
## D = 0.083176, p-value = 1.874e-05
```

```
lillie.test(fwd_laliga$Sh)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_laliga$Sh
## D = 0.086071, p-value = 0.3533
```

```
lillie.test(fwd_ligue1$Sh)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_ligue1$Sh
## D = 0.096746, p-value = 0.1077
```

```
lillie.test(fwd_premier$Sh)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_premier$Sh
## D = 0.067865, p-value = 0.6691
```

```
lillie.test(fwd_seriea$Sh)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_seriea$Sh
## D = 0.12931, p-value = 0.005458
```

```
lillie.test(fwd_bundesliga$Sh)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_bundesliga$Sh
## D = 0.12531, p-value = 0.03706
```

```
#G-xG
lillie.test(fwd$G_xG)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd$G_xG
## D = 0.075601, p-value = 0.0001791
```

```
lillie.test(fwd_laliga$G_xG) ##Normal
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_laliga$G_xG
## D = 0.10512, p-value = 0.1143
```

```
lillie.test(fwd_ligue1$G_xG) ##Normal
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_ligue1$G_xG
## D = 0.072114, p-value = 0.4901
```

```
lillie.test(fwd_premier$G_xG) ##Normal
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_premier$G_xG
## D = 0.052096, p-value = 0.9411
```

```
lillie.test(fwd_seriea$G_xG) ##Normal
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_seriea$G_xG
## D = 0.12457, p-value = 0.008933
```

```
lillie.test(fwd_bundesliga$G_xG) ##Normal
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_bundesliga$G_xG
## D = 0.1157, p-value = 0.07398
```

```
#npxG_Sh
lillie.test(fwd$npxG_Sh)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd$npxG_Sh
## D = 0.06725, p-value = 0.001607
```

```
lillie.test(fwd_laliga$npxG_Sh) ##Normal
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_laliga$npxG_Sh
## D = 0.080005, p-value = 0.4702
```

```
lillie.test(fwd_ligue1$npxG_Sh) ##Normal
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_ligue1$npxG_Sh
## D = 0.10133, p-value = 0.07193
```

```
lillie.test(fwd_premier$npxG_Sh)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_premier$npxG_Sh
## D = 0.13124, p-value = 0.008837
```

```
lillie.test(fwd_seriea$npxG_Sh) ##Normal
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_seriea$npxG_Sh
## D = 0.089588, p-value = 0.1776
```

```
lillie.test(fwd_bundesliga$npxG_Sh) ##Normal
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  fwd_bundesliga$npxG_Sh
## D = 0.1206, p-value = 0.05249
```

```
##Anàlisi de igualtat de variàncies per diferents mètriques
bartlett.test(Gls ~ Competition , data = fwd )
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Gls by Competition
## Bartlett's K-squared = 9.644, df = 4, p-value = 0.04687
```

```
bartlett.test(xG ~ Competition , data = fwd )
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  xG by Competition
## Bartlett's K-squared = 10.348, df = 4, p-value = 0.03496
```

```
bartlett.test(Sh ~ Competition , data = fwd )
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Sh by Competition
## Bartlett's K-squared = 4.6745, df = 4, p-value = 0.3224
```

```
bartlett.test(npxG ~ Competition , data = fwd )
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  npxG by Competition
## Bartlett's K-squared = 17.123, df = 4, p-value = 0.001829
```

```
bartlett.test(npxG_Sh ~ Competition , data = fwd )
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  npxG_Sh by Competition
## Bartlett's K-squared = 1.1039, df = 4, p-value = 0.8937
```

```
bartlett.test(G_xG ~ Competition , data = fwd )
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  G_xG by Competition
## Bartlett's K-squared = 2.1601, df = 4, p-value = 0.7063
```

```
##Generació d'un model de regressió lineal amb variables quantitatives
fwd_datos <- data.frame (fwd$npxG_Sh, fwd$Dist)
fwd_datos2 <- data.frame(fwd$npxG_Sh, fwd$Sh)


modelo_lineal <- lm (fwd.npxG_Sh ~ fwd.Dist, data= fwd_datos)
summary(modelo_lineal)
```

```
##
## Call:
## lm(formula = fwd.npxG_Sh ~ fwd.Dist, data = fwd_datos)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.080049 -0.020638 -0.002474  0.019097  0.086694
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.2734494  0.0083785   32.64   <2e-16 ***
## fwd.Dist    -0.0094691  0.0005162  -18.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02943 on 312 degrees of freedom
## Multiple R-squared:  0.5189, Adjusted R-squared:  0.5174
## F-statistic: 336.5 on 1 and 312 DF,  p-value: < 2.2e-16
```

```
modelo_lineal2 <- lm (fwd.npxG_Sh ~ fwd.Sh, data= fwd_datos2)
summary(modelo_lineal2)
```

```
##
## Call:
## lm(formula = fwd.npxG_Sh ~ fwd.Sh, data = fwd_datos2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.098485 -0.032581 -0.000393  0.026522  0.111066
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.1166902  0.0056486  20.658   <2e-16 ***
## fwd.Sh      0.0002244  0.0001879   1.194    0.233
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04233 on 312 degrees of freedom
## Multiple R-squared:  0.00455,    Adjusted R-squared:  0.001359
## F-statistic: 1.426 on 1 and 312 DF,  p-value: 0.2333
```

```
##Anàlisi de correlació entre diferents mètriques
cor.test(fwd$xG,fwd$Dist)
```

```
##
##  Pearson's product-moment correlation
##
## data:  fwd$xG and fwd$Dist
## t = -5.1242, df = 312, p-value = 5.247e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3776526 -0.1732741
## sample estimates:
##        cor
## -0.2786147
```

```
cor.test(fwd$Sh,fwd$Gls)
```

```
##
##  Pearson's product-moment correlation
##
## data:  fwd$Sh and fwd$Gls
## t = 16.912, df = 312, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6290484 0.7452204
## sample estimates:
##       cor
## 0.6915807
```

```
cor.test(fwd$xG, fwd$Sh)
```

```
##
##  Pearson's product-moment correlation
##
## data:  fwd$xG and fwd$Sh
## t = 26.482, df = 312, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7943806 0.8631254
## sample estimates:
##       cor
## 0.8319181
```

```
cor.test(fwd$Gls, fwd$Sh)
```

```
##
##  Pearson's product-moment correlation
##
## data:  fwd$Gls and fwd$Sh
## t = 16.912, df = 312, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6290484 0.7452204
## sample estimates:
##       cor
## 0.6915807
```

```
cor.test(fwd$Sh,fwd$G_Sh)
```

```
##
##  Pearson's product-moment correlation
##
## data:  fwd$Sh and fwd$G_Sh
## t = 1.611, df = 312, p-value = 0.1082
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.02005898  0.19950507
## sample estimates:
##       cor
## 0.09082669
```

```
cor.test(fwd$G_xG, fwd$Sh)
```

```
##
##  Pearson's product-moment correlation
##
## data:  fwd$G_xG and fwd$Sh
## t = 1.4378, df = 312, p-value = 0.1515
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.02982187  0.19010683
## sample estimates:
##       cor
## 0.08112994
```

```
cor.test(fwd$npxG_Sh,fwd$Dist)
```

```
##
##  Pearson's product-moment correlation
##
## data:  fwd$npxG_Sh and fwd$Dist
## t = -18.344, df = 312, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7696616 -0.6624788
## sample estimates:
##       cor
## -0.720343
```

```
cor.test(fwd$G_xG,fwd$Dist)
```

```
##
##  Pearson's product-moment correlation
##
## data:  fwd$G_xG and fwd$Dist
## t = 0.48313, df = 312, p-value = 0.6293
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.08359554  0.13760903
## sample estimates:
##       cor
## 0.02734145
```

```
cor.test(fwd$npxG_Sh,fwd$G_xG)
```

```
##
##  Pearson's product-moment correlation
##
## data:  fwd$npxG_Sh and fwd$G_xG
## t = 0.45994, df = 312, p-value = 0.6459
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.08489869  0.13632118
## sample estimates:
##       cor
## 0.02602992
```

```
cor.test(fwd$Gls,fwd$npxG_Sh)
```

```
##
##  Pearson's product-moment correlation
##
## data:  fwd$Gls and fwd$npxG_Sh
## t = 7.3204, df = 312, p-value = 2.115e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2842169 0.4734768
## sample estimates:
##       cor
## 0.3828569
```

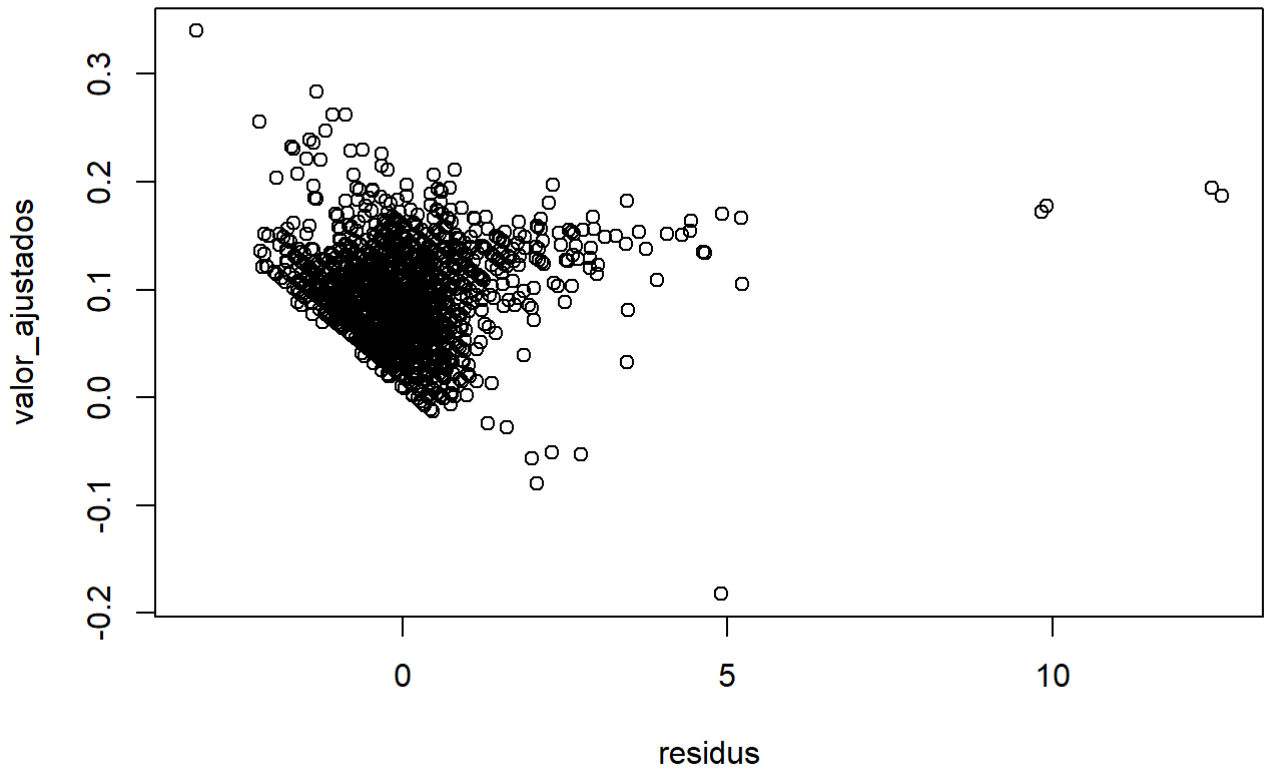
```
##Generació d'un model de regressió múltiple amb variables quantitatives i qualitatives
modelo_multiple <- lm(npxG_Sh ~ Competition + Age + Pos_main + Sh + Dist + xG, data = shootin
g)
summary(modelo_multiple)
```

```
##
## Call:
## lm(formula = npxG_Sh ~ Competition + Age + Pos_main + Sh + Dist +
##     xG, data = shooting)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.14985 -0.02486 -0.00426  0.01775  0.61300
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.1736782  0.0079303  21.901  < 2e-16 ***
## CompetitionLa Liga        -0.0015396  0.0035320  -0.436    0.663
## CompetitionLigue 1        -0.0017504  0.0035431  -0.494    0.621
## CompetitionPremier League -0.0001116  0.0035744  -0.031    0.975
## CompetitionSerie A        -0.0054557  0.0035313  -1.545    0.123
## Age                       -0.0003253  0.0002643  -1.231    0.219
## Pos_mainFW                 0.0247657  0.0031193   7.940 3.38e-15 ***
## Pos_mainGK                 0.0024873  0.0246152   0.101    0.920
## Pos_mainMF                 0.0198069  0.0027625   7.170 1.06e-12 ***
## Sh                        -0.0033230  0.0002194 -15.143  < 2e-16 ***
## Dist                      -0.0047858  0.0002001 -23.914  < 2e-16 ***
## xG                         0.0277595  0.0014610  19.001  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04884 on 1971 degrees of freedom
##   (453 observations deleted due to missingness)
## Multiple R-squared:  0.438,  Adjusted R-squared:  0.4349
## F-statistic: 139.6 on 11 and 1971 DF,  p-value: < 2.2e-16
```

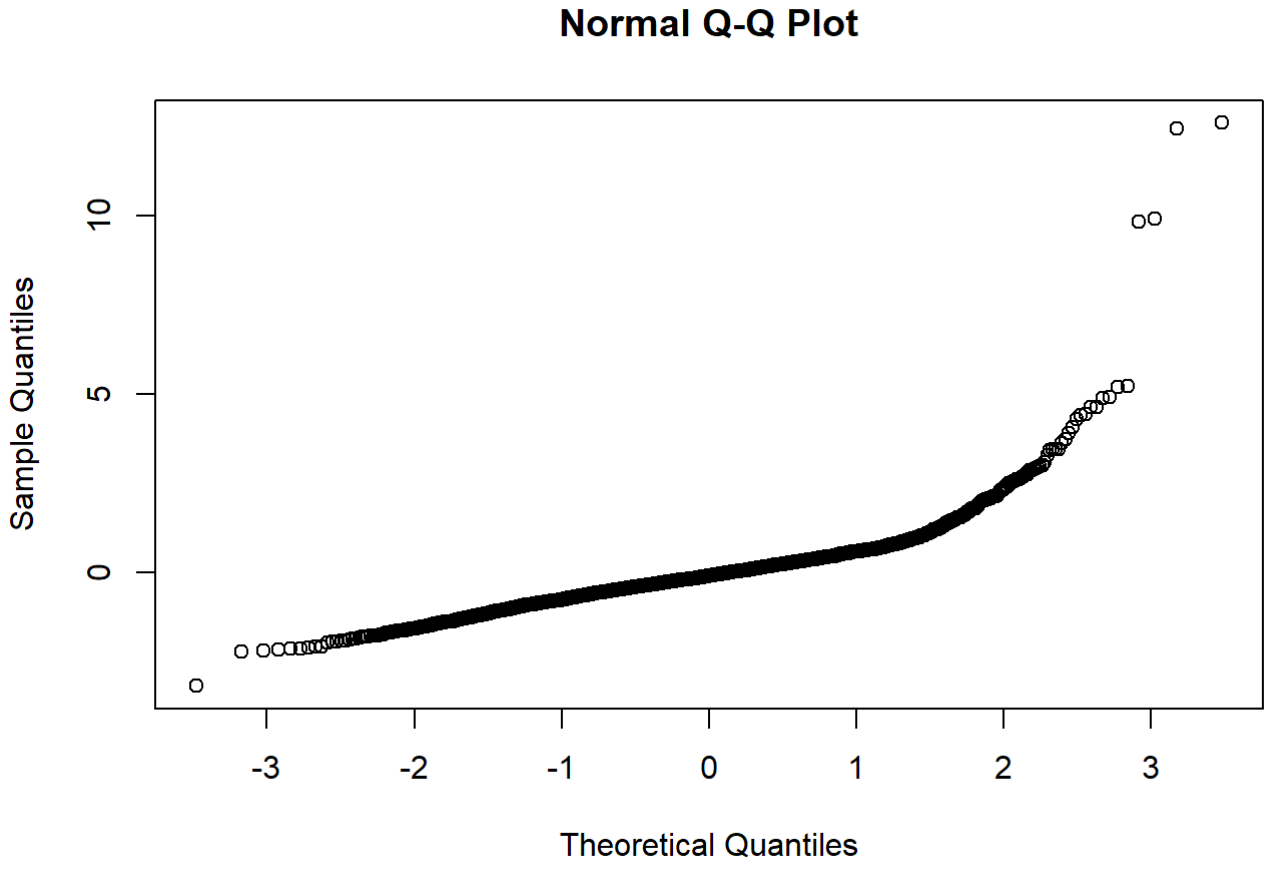
```
exp(coefficients((modelo_multiple)))
```

```
##               (Intercept)        CompetitionLa Liga        CompetitionLigue 1
##                 1.1896727                 0.9984616                 0.9982511
## CompetitionPremier League        CompetitionSerie A                       Age
##                 0.9998884                 0.9945592                 0.9996747
##                Pos_mainFW                Pos_mainGK                Pos_mainMF
##                 1.0250749                 1.0024903                 1.0200043
##                        Sh                      Dist                        xG
##                 0.9966826                 0.9952256                 1.0281484
```

```
residus <- rstandard(modelo_multiple)
valor_ajustados <- fitted (modelo_multiple)
plot(residus, valor_ajustados)
```




```
qqnorm(residus)
```



```
##Test d'igualtat de variàncies per parelles
var.test(fwd_laliga$G_xG, fwd_premier$G_xG) #El test no troba diferències significatives a les variàncies
```

```
##
##  F test to compare two variances
##
## data:  fwd_laliga$G_xG and fwd_premier$G_xG
## F = 1.1751, num df = 57, denom df = 62, p-value = 0.5331
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7051345 1.9694160
## sample estimates:
## ratio of variances
##           1.175137
```

```
var.test(fwd_seriea$G_xG, fwd_premier$G_xG) #El test no troba diferències significatives a les variàncies
```

```
##
##  F test to compare two variances
##
## data:  fwd_seriea$G_xG and fwd_premier$G_xG
## F = 1.3174, num df = 69, denom df = 62, p-value = 0.271
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8050374 2.1420237
## sample estimates:
## ratio of variances
##           1.317421
```

```
var.test(fwd_ligue1$G_xG, fwd_premier$G_xG) #El test no troba diferències significatives a les variàncies
```

```
##
##  F test to compare two variances
##
## data:  fwd_ligue1$G_xG and fwd_premier$G_xG
## F = 1.3798, num df = 69, denom df = 62, p-value = 0.1988
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8431574 2.2434525
## sample estimates:
## ratio of variances
##           1.379804
```

```
var.test(fwd_bundesliga$G_xG, fwd_premier$G_xG) #El test no troba diferències significatives a les variàncies
```

```
##
##  F test to compare two variances
##
## data:  fwd_bundesliga$G_xG and fwd_premier$G_xG
## F = 1.3593, num df = 52, denom df = 62, p-value = 0.246
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8078403 2.3156271
## sample estimates:
## ratio of variances
##           1.359341
```

```
##Test d'hipòtesi sobre les mitjanes, amb mètriques i mostres que compleixen normalitat i hom
ocedasticitat

t.test(fwd_laliga$G_xG,fwd_premier$G_xG, conf.level=0.95) #s'obté un valor inferior a 0.05, p
er tant s'accepta la hipòtesi nul·la, es pot concloure amb un 95% de nivell de confiança que
 la mitjana d'efectivitat G-xG és la mateixa a la Lliga Espanyola que a la Premier.
```

```
##
##  Welch Two Sample t-test
##
## data:  fwd_laliga$G_xG and fwd_premier$G_xG
## t = 1.9853, df = 115.9, p-value = 0.04947
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.001413889 1.191744294
## sample estimates:
##  mean of x  mean of y
##  0.4568966 -0.1396825
```

```
t.test(fwd_laliga$G_xG,fwd_seriea$G_xG, conf.level=0.95) #s'obté un valor superior a 0.05, pe
r tant es rebutja la hipòtesi nul·la, es pot concloure amb un 95% de nivell de confiança que
 la mitjana d'efectivitat G-xG és diferent a la Lliga Espanyola que a la Serie A.
```

```
##
##  Welch Two Sample t-test
##
## data:  fwd_laliga$G_xG and fwd_seriea$G_xG
## t = 0.58434, df = 123.82, p-value = 0.5601
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4359337  0.8011553
## sample estimates:
## mean of x mean of y
## 0.4568966 0.2742857
```

```
t.test(fwd_laliga$G_xG,fwd_bundesliga$G_xG, conf.level=0.95) #s'obté un valor superior a 0.0
5, per tant es rebutja la hipòtesi nul·la, es pot concloure amb un 95% de nivell de confiança
que la mitjana d'efectivitat G-xG és diferent a la Lliga Espanyola que a la Bundesliga.
```

```
##
##  Welch Two Sample t-test
##
## data:  fwd_laliga$G_xG and fwd_bundesliga$G_xG
## t = 1.1484, df = 106.17, p-value = 0.2534
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2825315  1.0604755
## sample estimates:
##  mean of x  mean of y
## 0.45689655 0.06792453
```

```
t.test(fwd_laliga$G_xG,fwd_ligue1$G_xG, conf.level=0.95) #s'obté un valor superior a 0.05, pe
r tant es rebutja la hipòtesi nul·la, es pot concloure amb un 95% de nivell de confiança que
 la mitjana d'efectivitat G-xG és diferent a la Lliga Espanyola que a la Bundesliga.
```

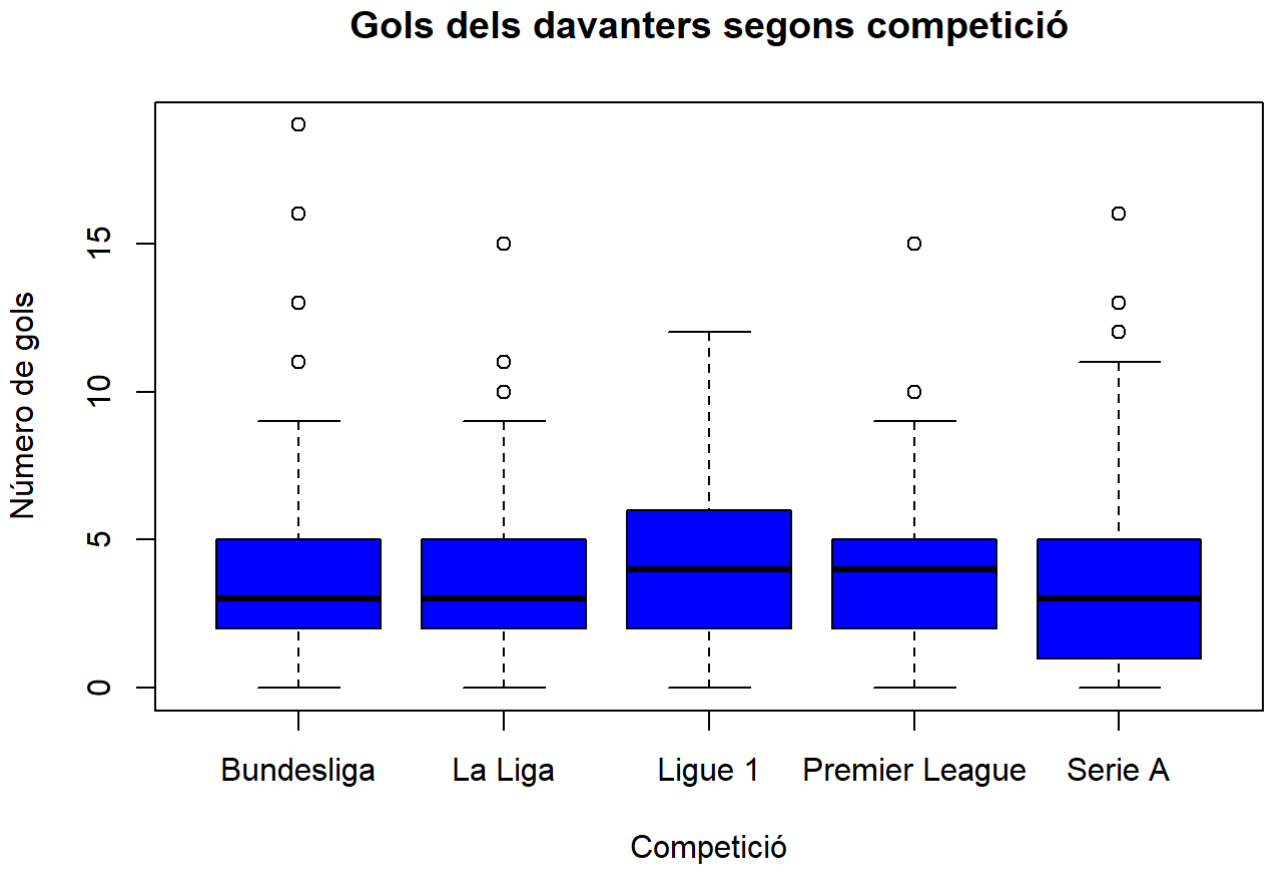
```
##
##  Welch Two Sample t-test
##
## data:  fwd_laliga$G_xG and fwd_ligue1$G_xG
## t = 1.1247, df = 124.51, p-value = 0.2629
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.270055  0.980991
## sample estimates:
## mean of x mean of y
## 0.4568966 0.1014286
```

2.5 - Representació gràfica

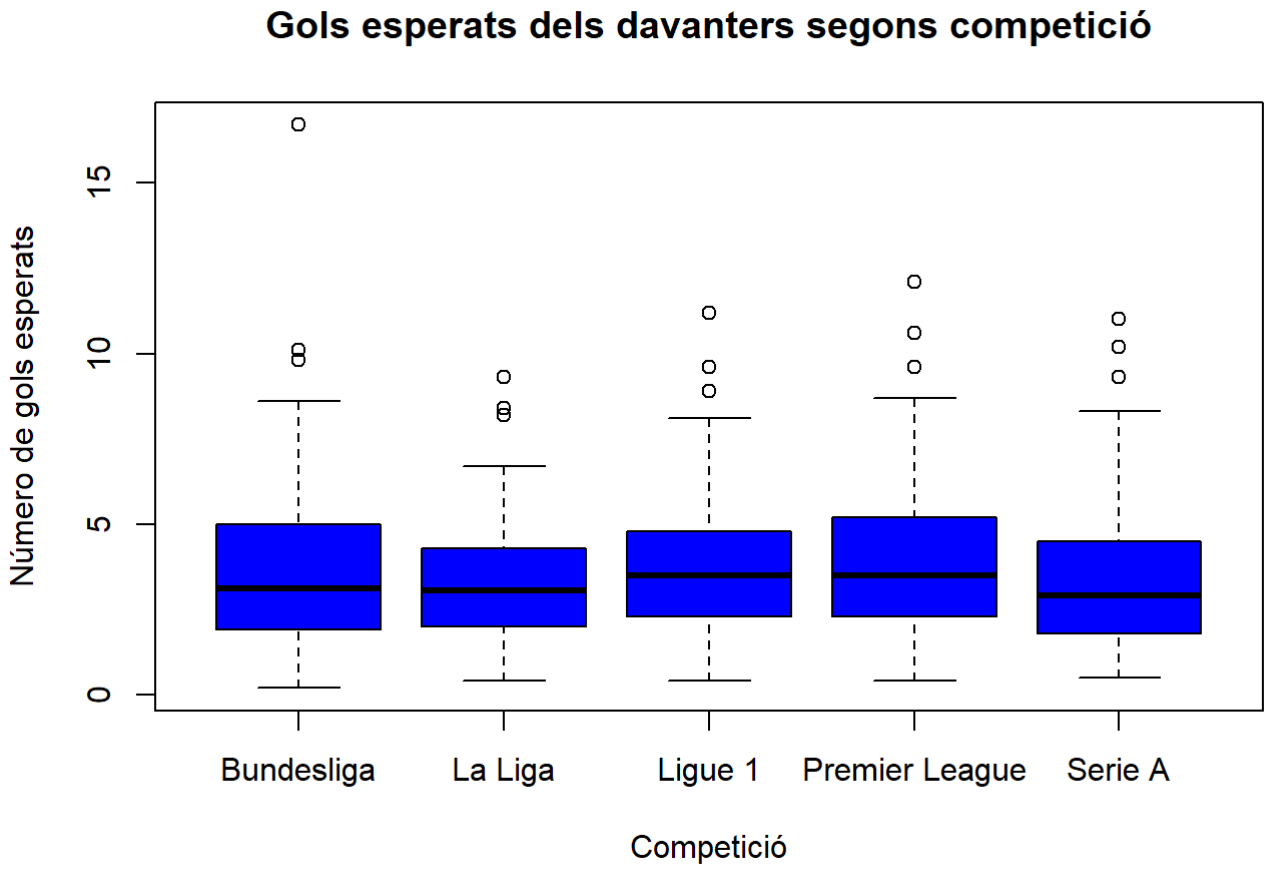
Es mostra una comparació amb boxplots sobre algunes de les mètriques analitzades anteriorment. Dels boxplots podem concloure que les grans diferències entre les finalitzacions de les 5 grans lligues són conseqüència més aviat de la qualitat dels tirs que no del rendiment d'ells. Per un costat, es veu com a la Ligue 1 i a la Premier league el valors dels gols és superior, però si s'analitza la mètrica G-xG, que calcularia l'efectivitat o rendiment, s'observa com està en valors similars.No obstant, a la Premier League els davanters generen més número de tirs i , sobretot, cada tir té un valor de xG més alt.Amb el boxplot de la distància, tenint en compte que la Premier League té més tirs, s'observa com la distància dels tirs és menor, i és això el que fa que els tirs tinguin un valor de xG més alt, doncs quan més a prop de porteria sigui el tir, generalment, més probable serà que acabi en gol.

Amb els histogrames s'ajuda a observar aquestes tendències, especialment amb les diferències entre els G-xG, que es veuen millor que als boxplots, i les distàncies dels tirs i el npxG/Sh.

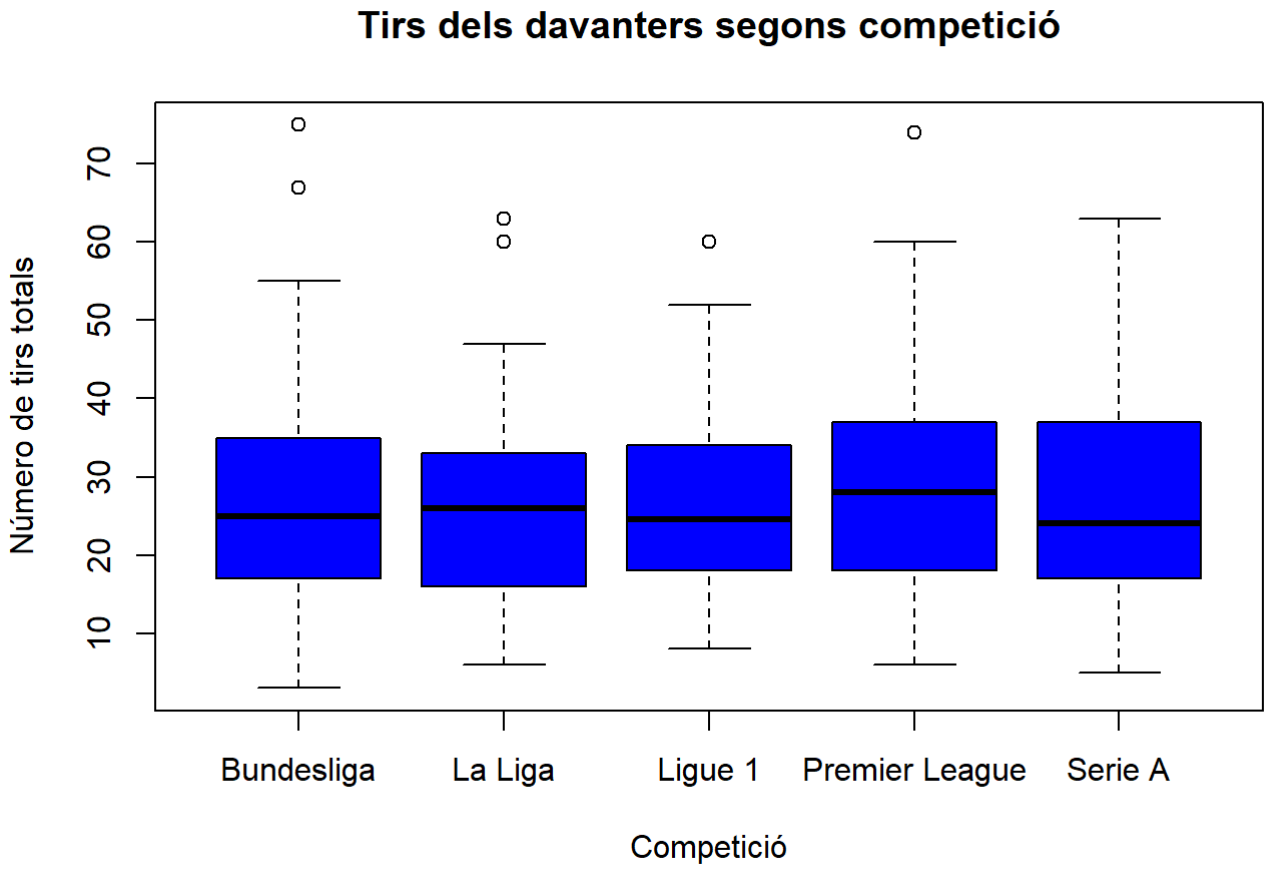
```
boxplot(fwd$Gls ~ fwd$Competition,
data=fwd,
main="Gols dels davanters segons competició",
xlab="Competició",
ylab="Número de gols",
col="blue",
border="black"
)
```



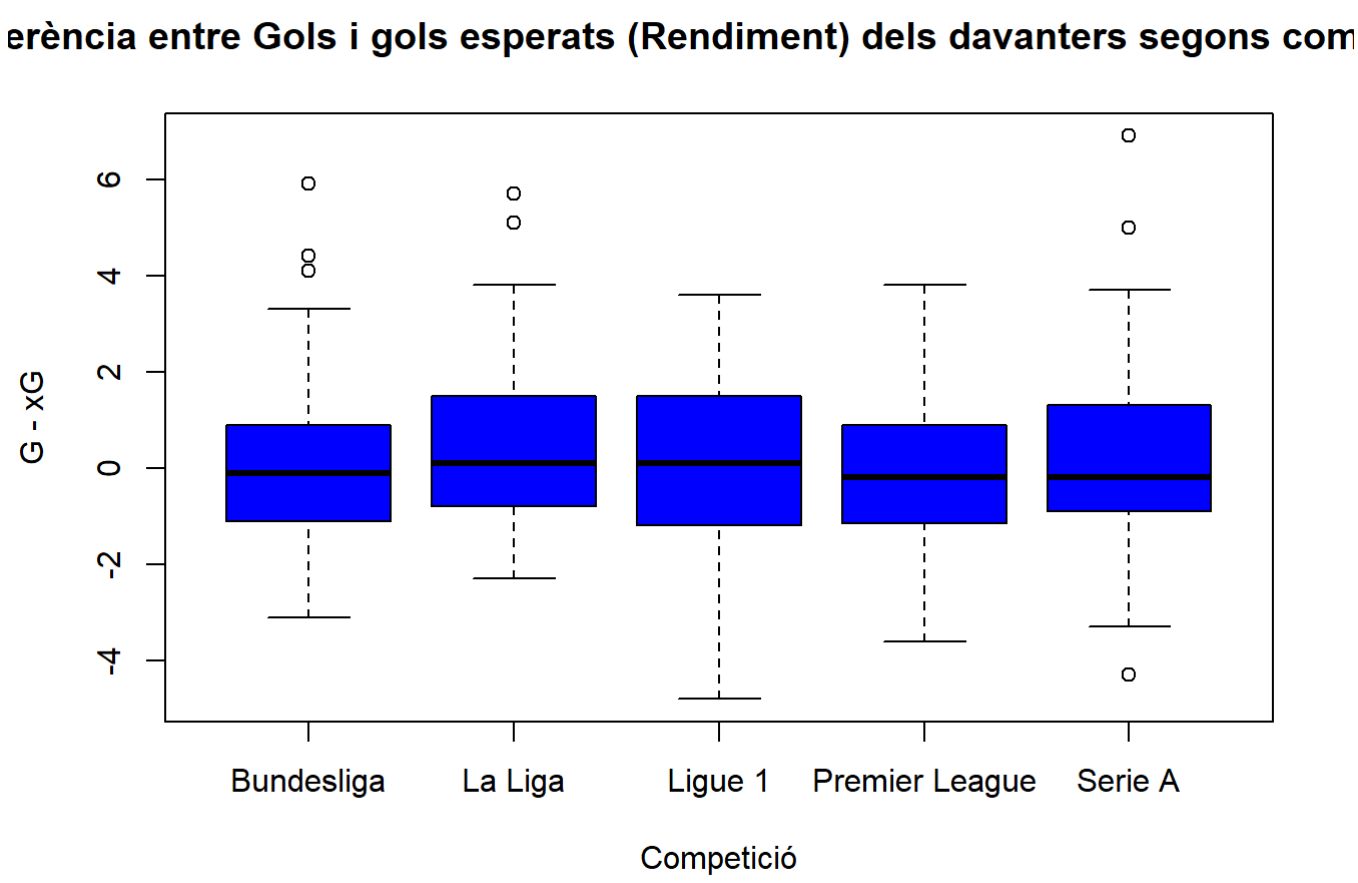

```
boxplot(fwd$xG ~ fwd$Competition,
data=fwd,
main="Gols esperats dels davanters segons competició",
xlab="Competició",
ylab="Número de gols esperats",
col="blue",
border="black"
)
```



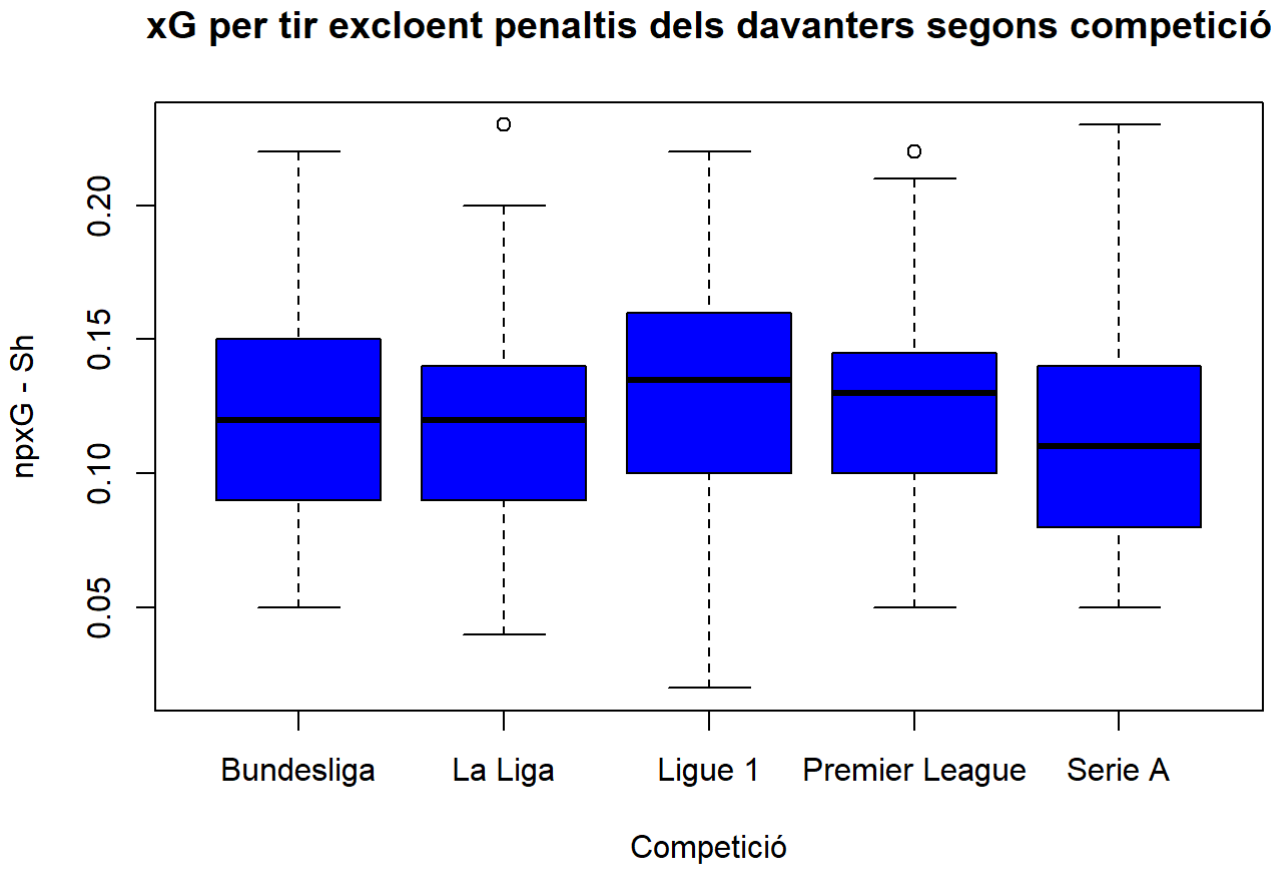

```
boxplot(fwd$Sh ~ fwd$Competition,
data=fwd,
main="Tirs dels davanters segons competició",
xlab="Competició",
ylab="Número de tirs totals",
col="blue",
border="black"
)
```



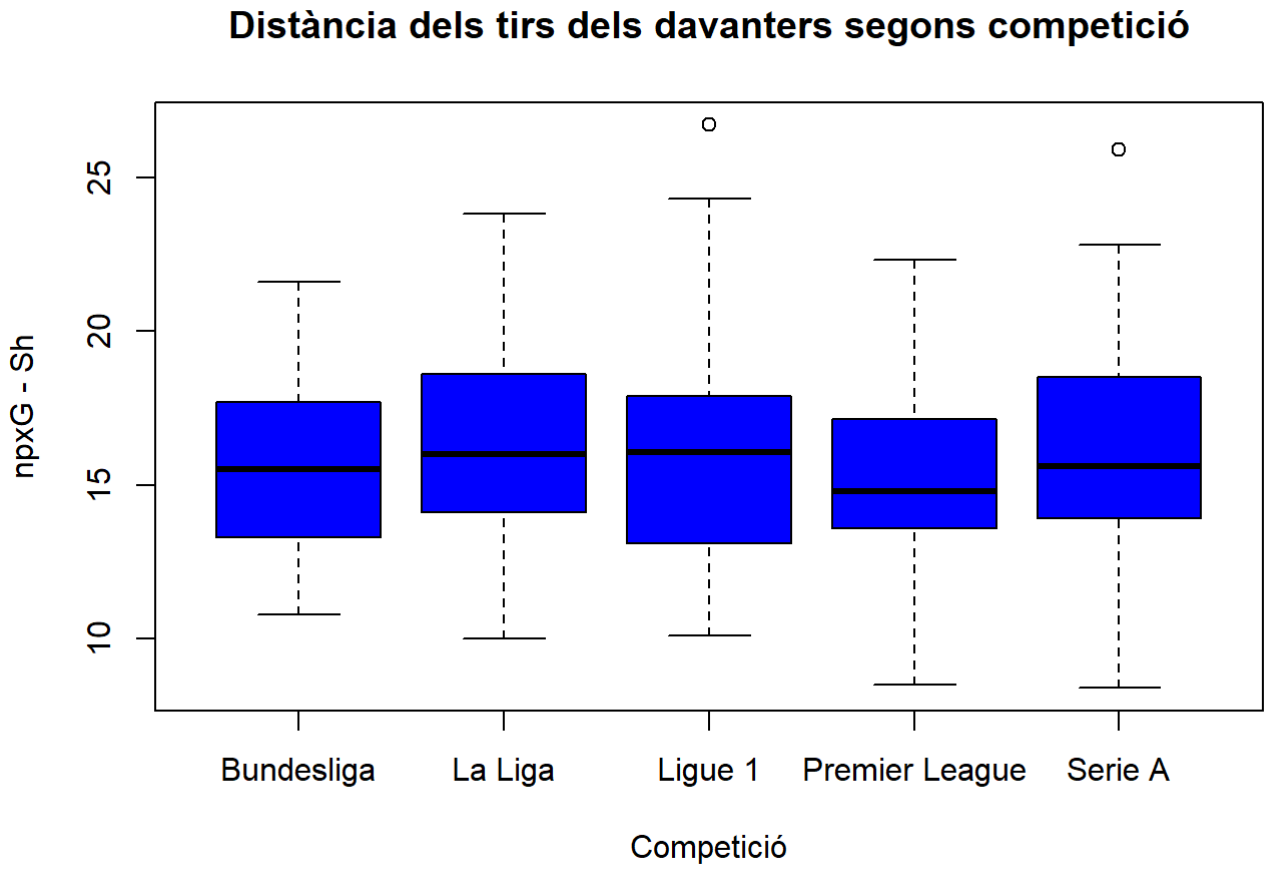

```
boxplot(fwd$G_xG ~ fwd$Competition,
data=fwd,
main="Diferència entre Gols i gols esperats (Rendiment) dels davanters segons competició",
xlab="Competició",
ylab="G - xG",
col="blue",
border="black"
)
```



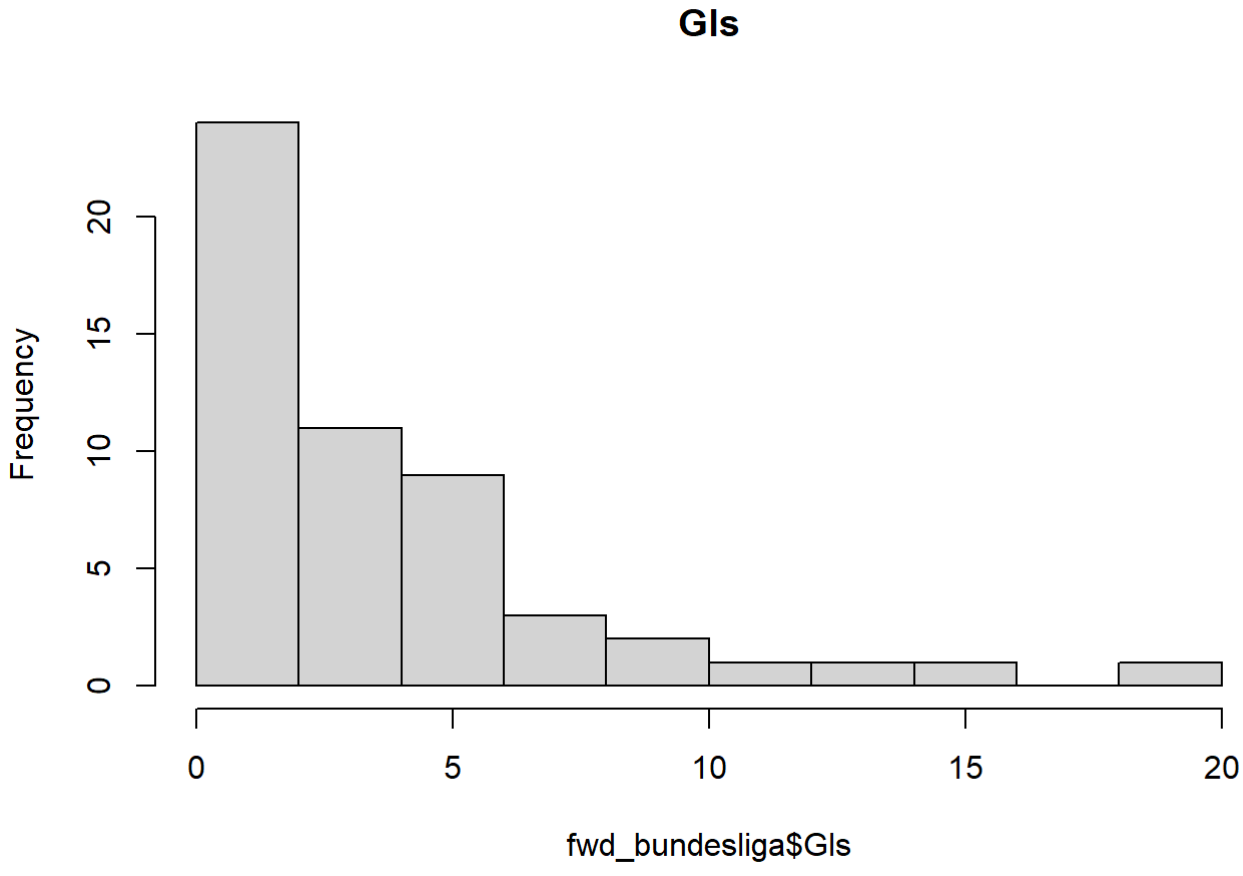

```
boxplot(fwd$npxG_Sh ~ fwd$Competition,
data=fwd,
main="xG per tir excloent penaltis dels davanters segons competició",
xlab="Competició",
ylab="npxG - Sh",
col="blue",
border="black"
)
```



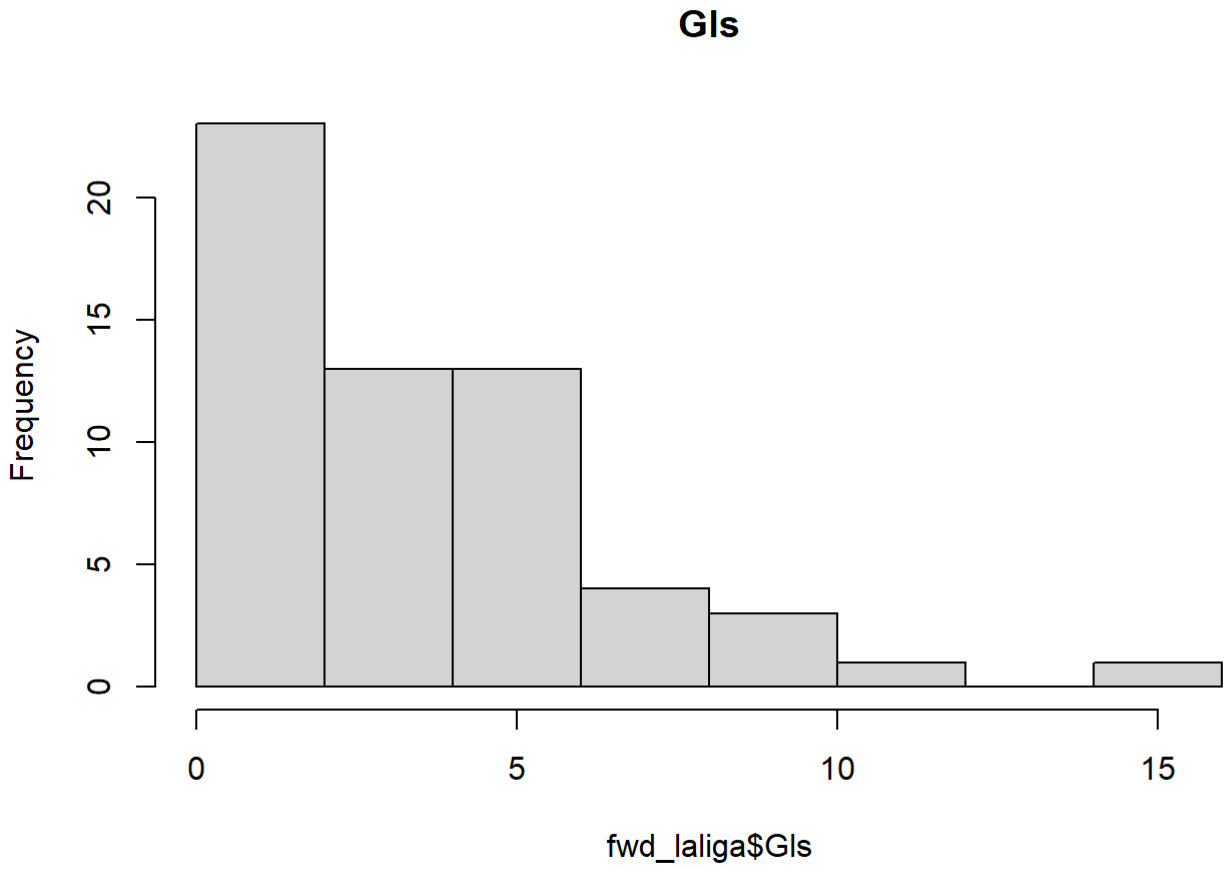

```
boxplot(fwd$Dist ~ fwd$Competition,
data=fwd,
main="Distància dels tirs dels davanters segons competició",
xlab="Competició",
ylab="npxG - Sh",
col="blue",
border="black"
)
```



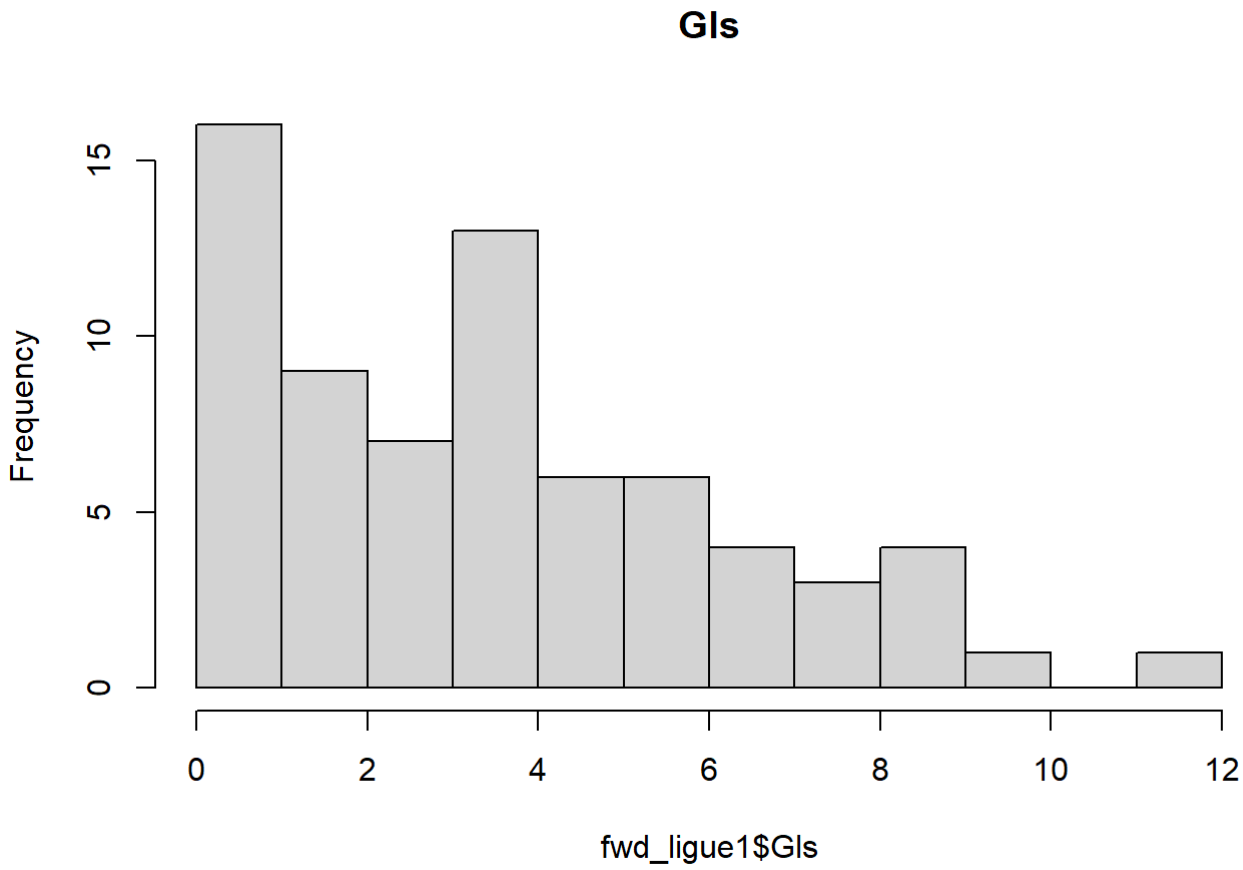

```
hist(fwd_bundesliga$Gls, main='Gls', breaks=10)
```



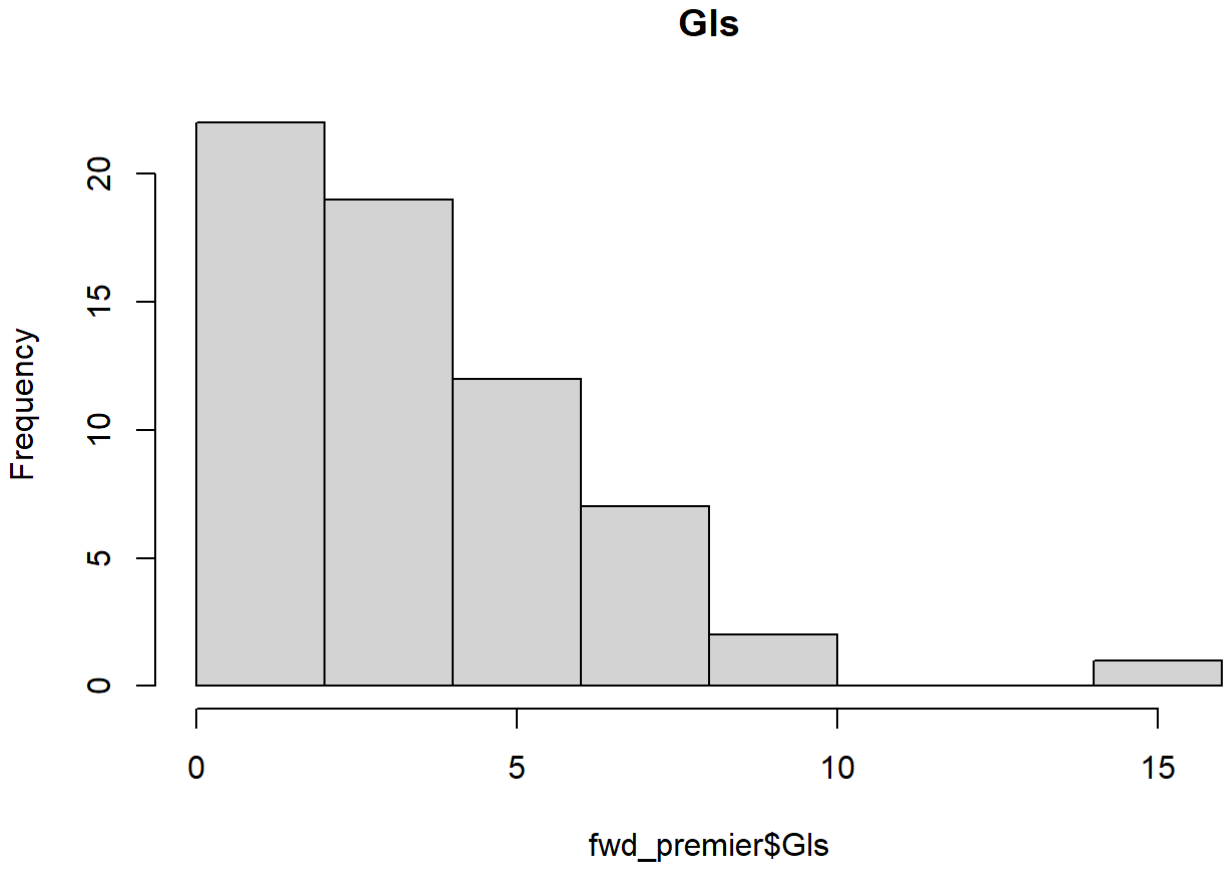
```
hist(fwd_laliga$Gls, main='Gls', breaks=10)
```



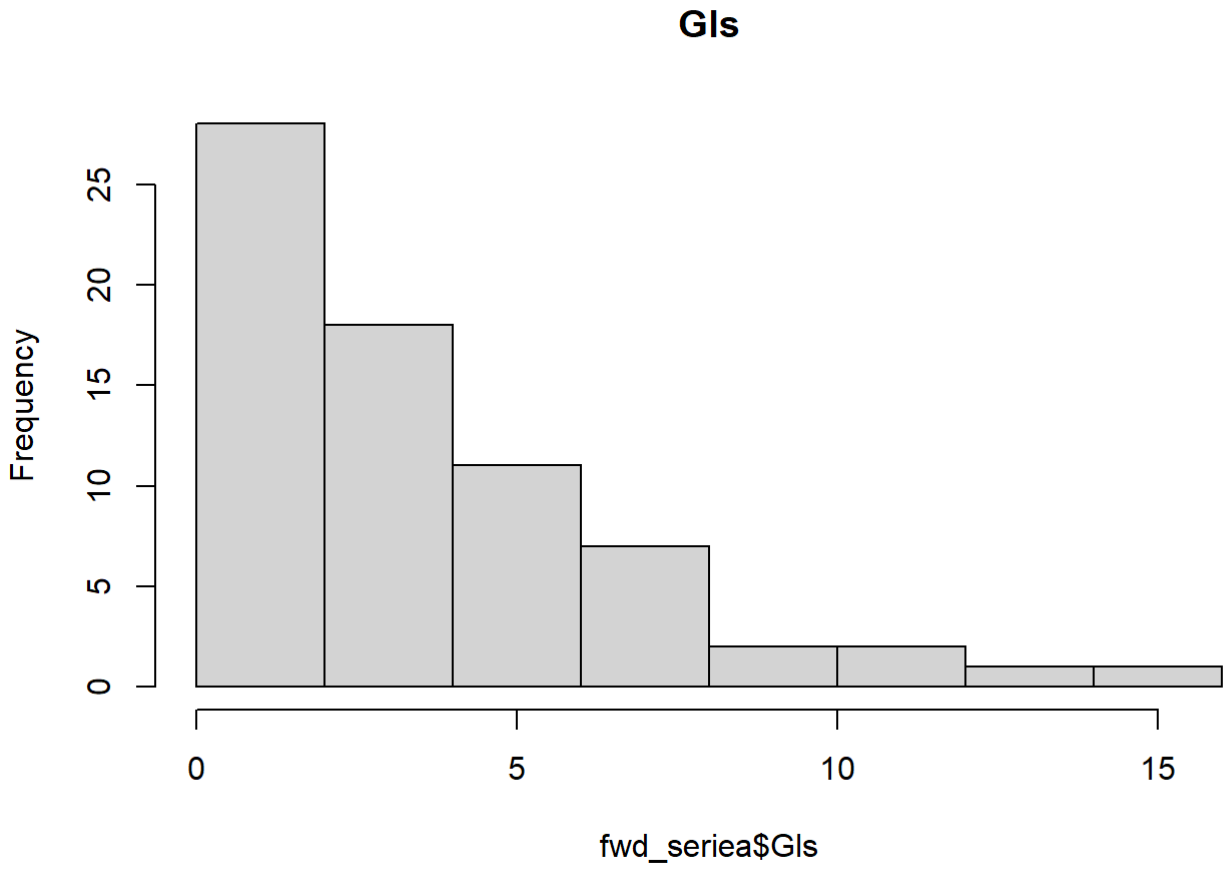

```
hist(fwd_ligue1$Gls, main='Gls', breaks=10)
```



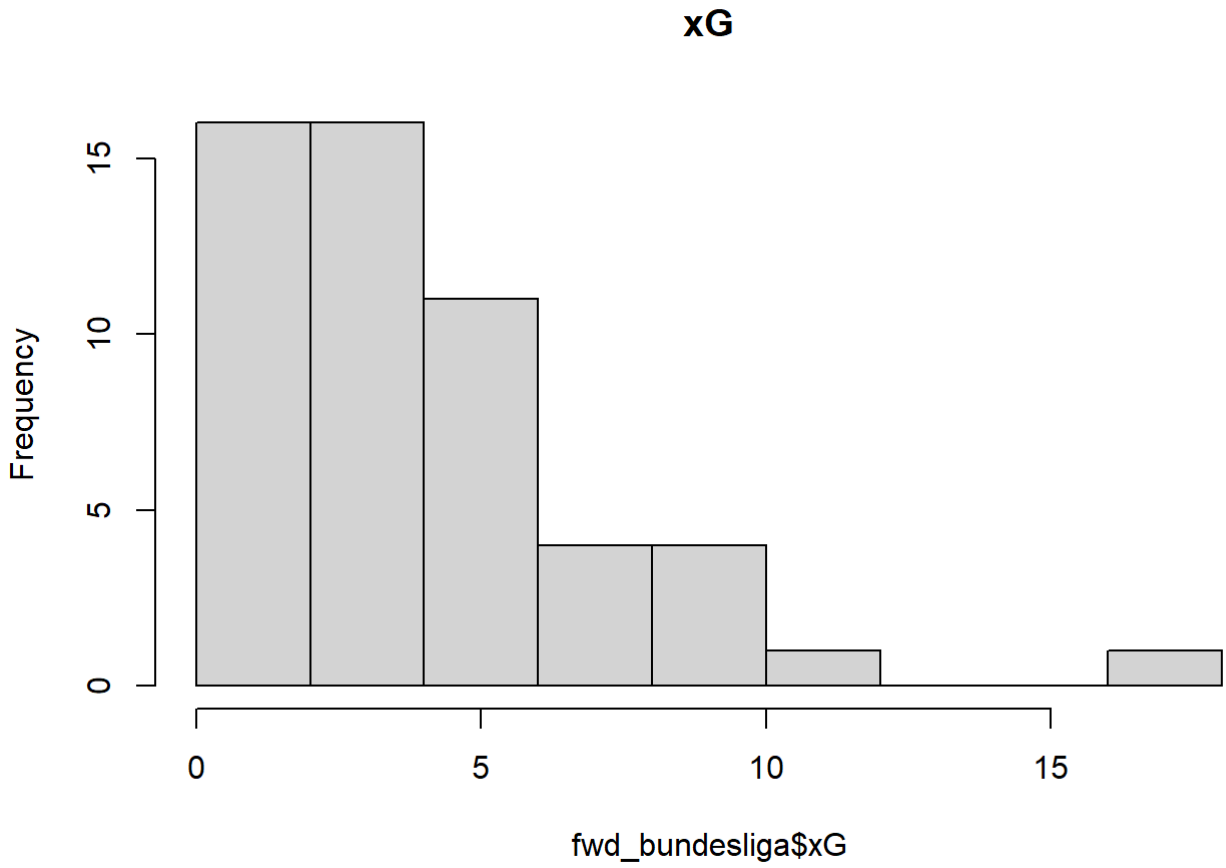

```
hist(fwd_premier$Gls, main='Gls', breaks=10)
```



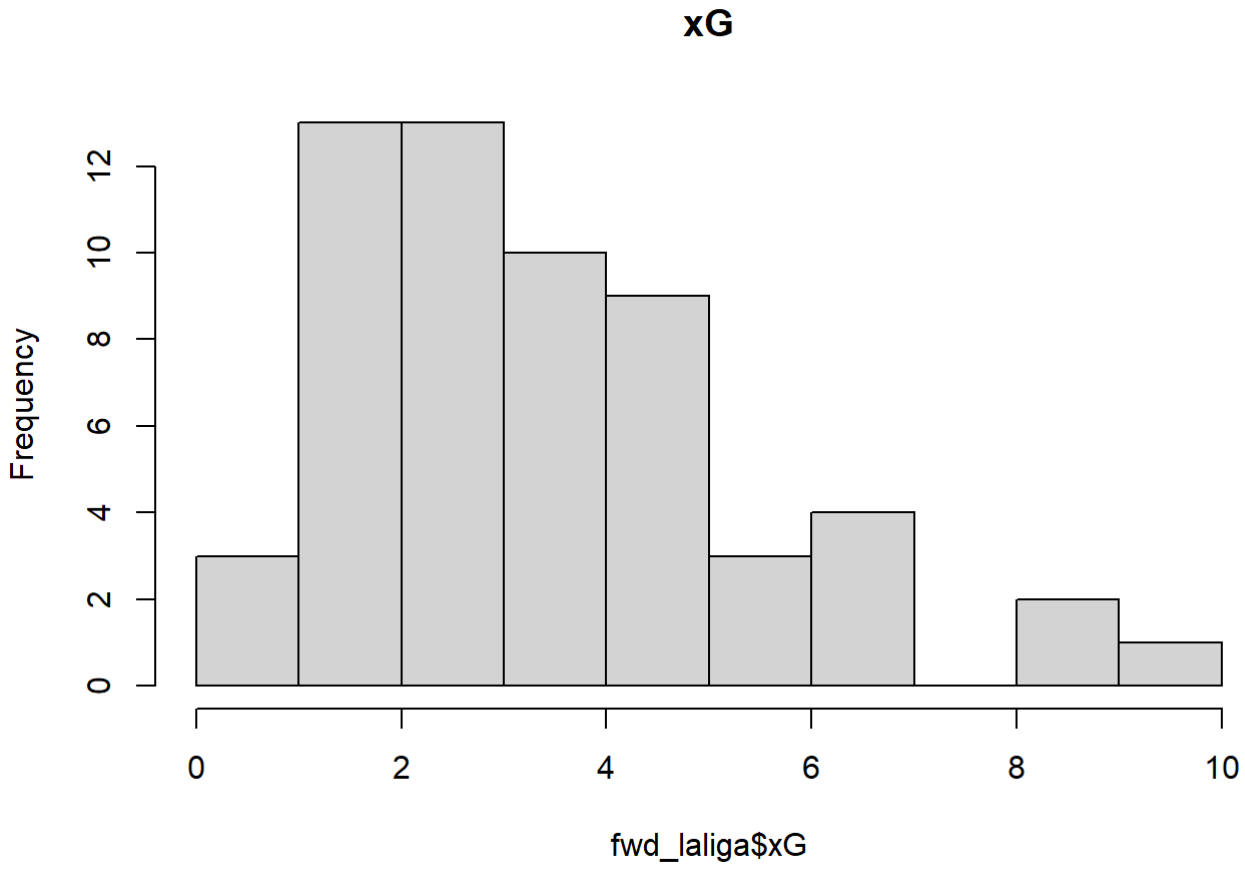
```
hist(fwd_seriea$Gls, main='Gls', breaks=10)
```



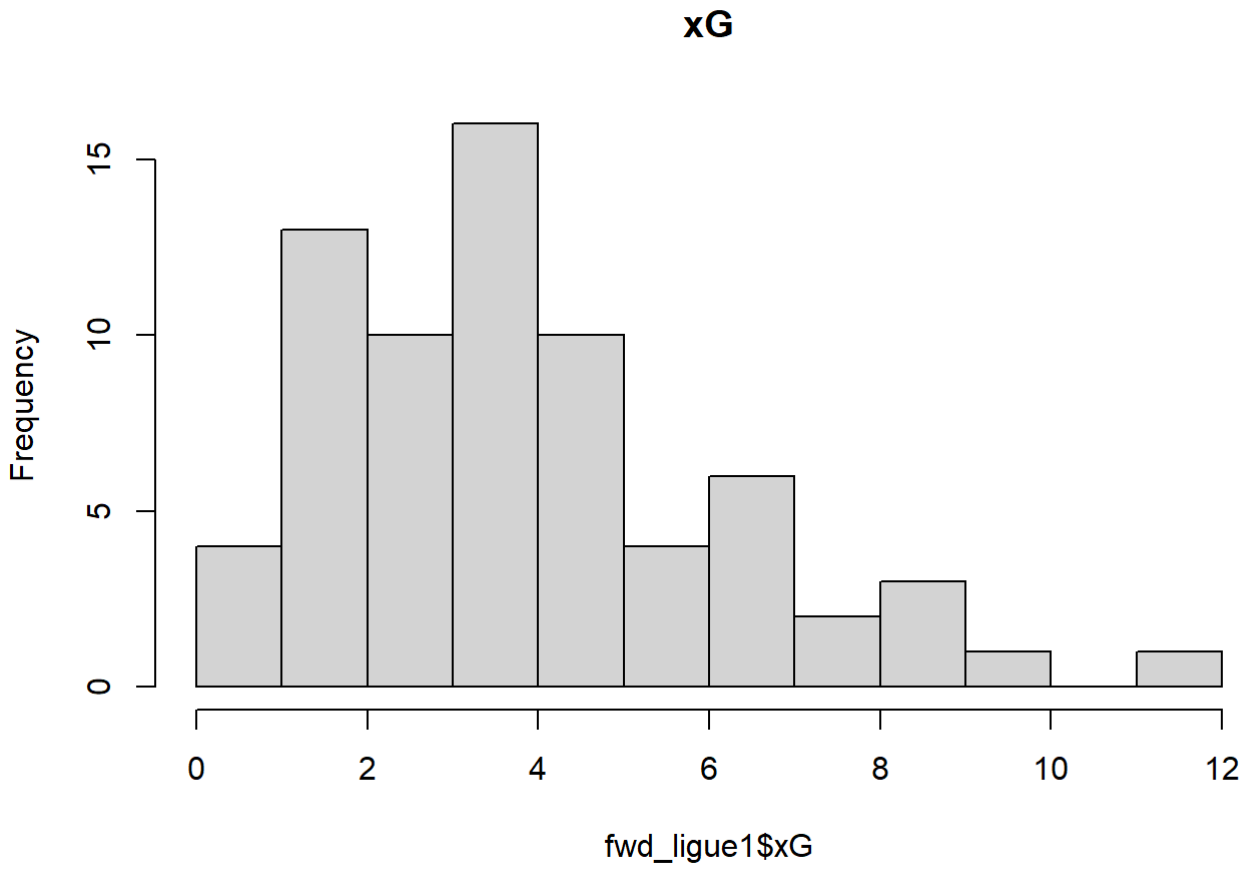

```
hist(fwd_bundesliga$xG, main='xG', breaks=10)
```



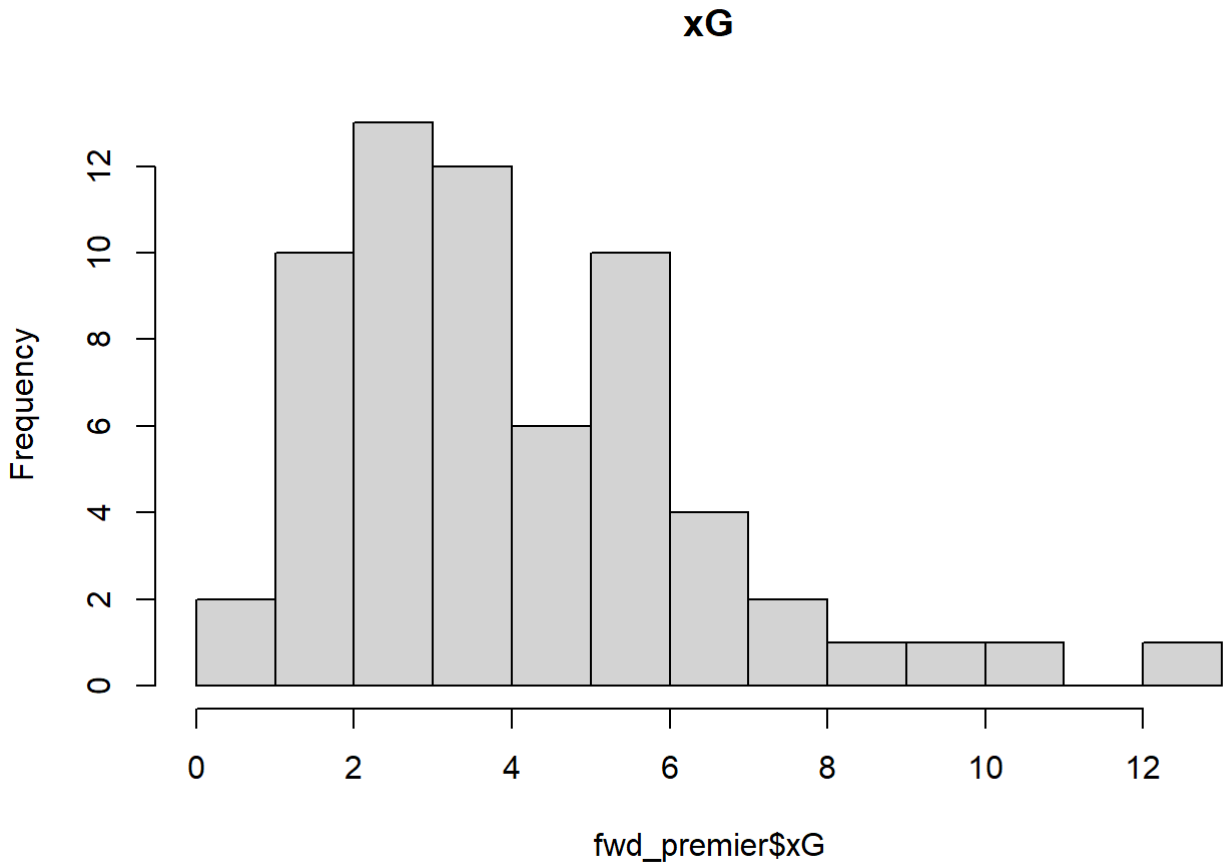

```
hist(fwd_laliga$xG, main='xG', breaks=10)
```



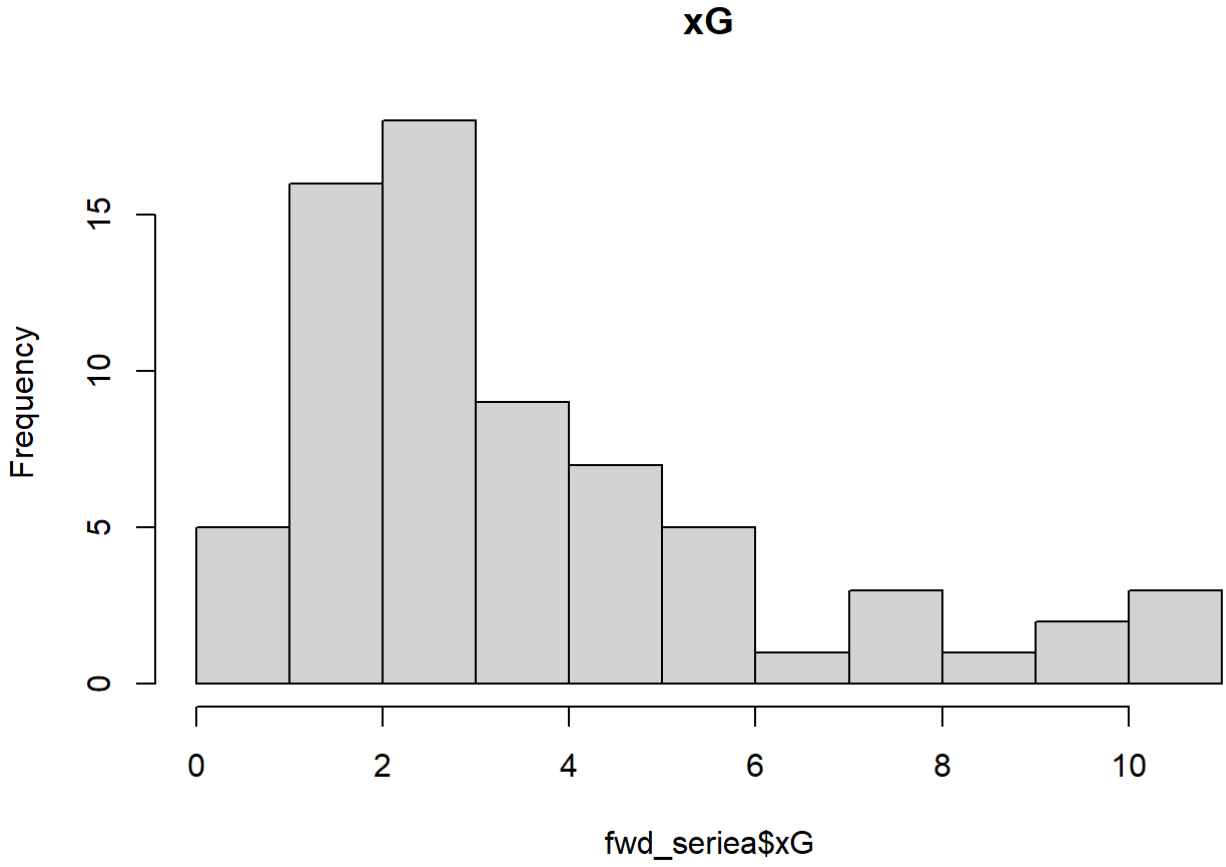
```
hist(fwd_ligue1$xG, main='xG', breaks=10)
```



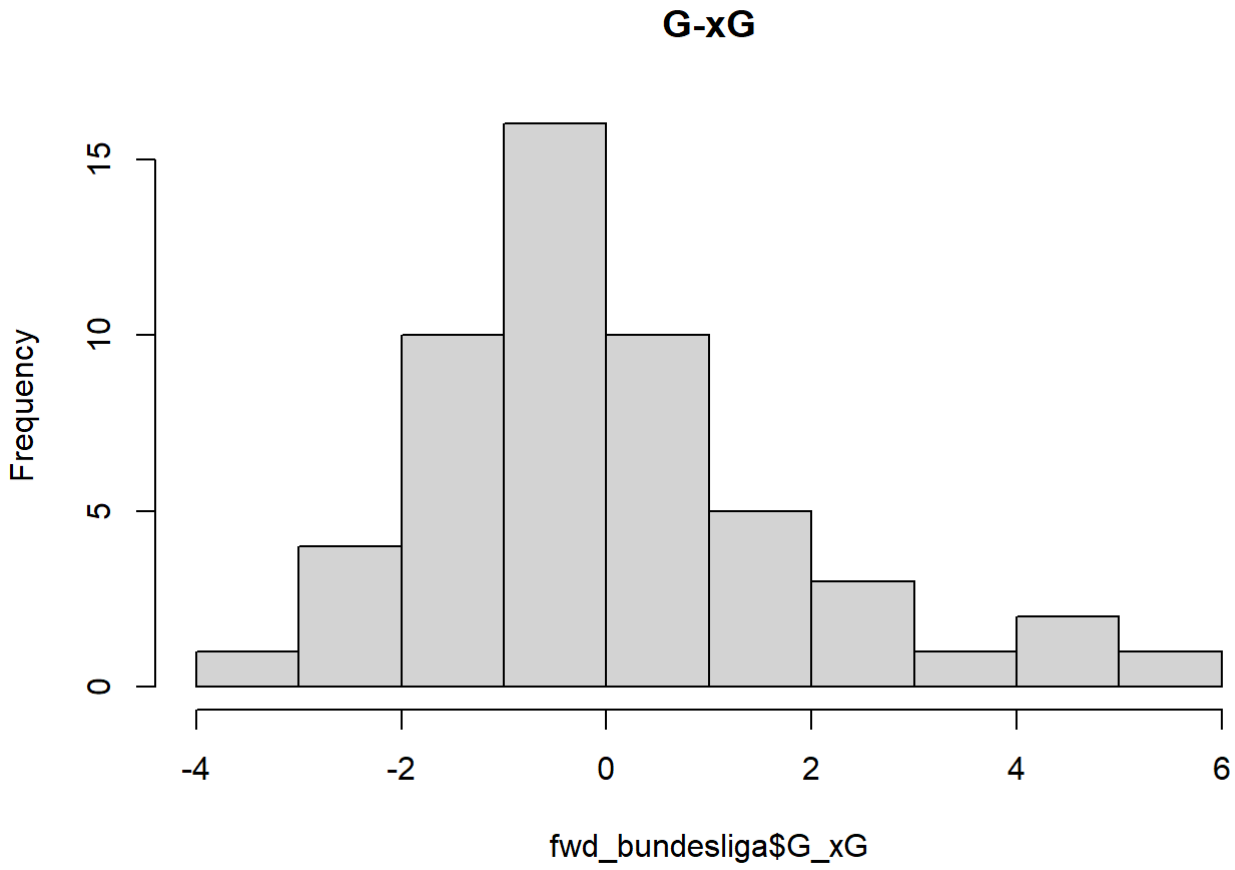

```
hist(fwd_premier$xG, main='xG', breaks=10)
```



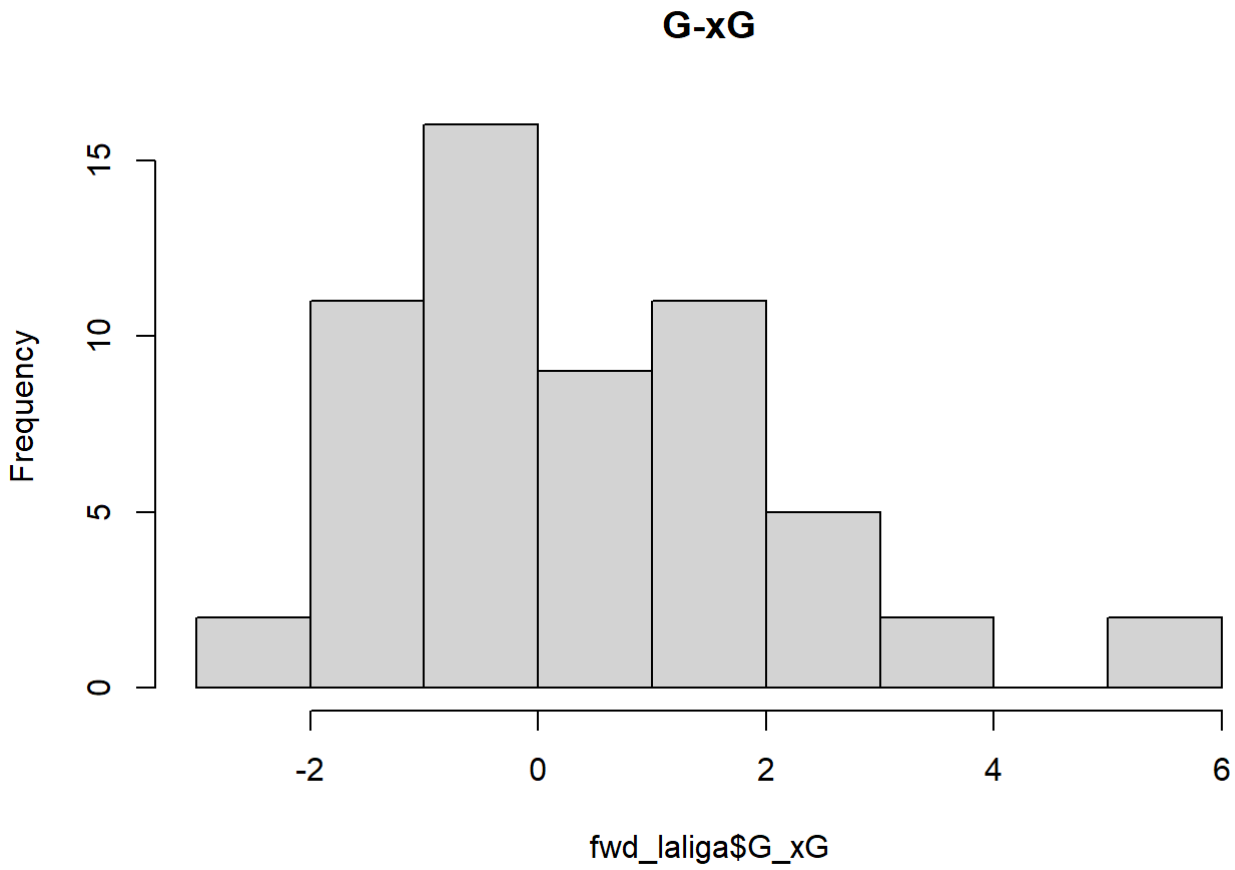

```
hist(fwd_seriea$xG, main='xG', breaks=10)
```



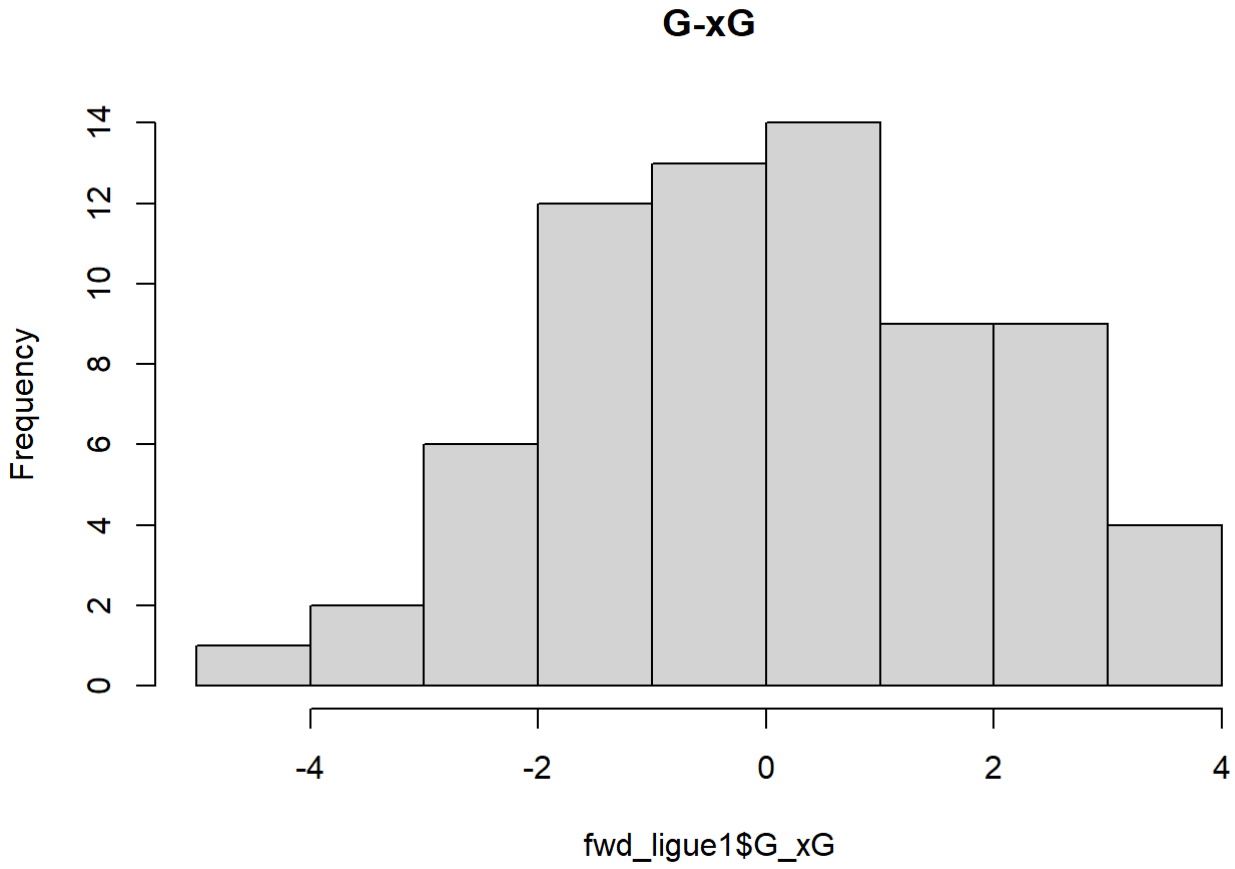
```
hist(fwd_bundesliga$G_xG, main='G-xG', breaks=10)
```



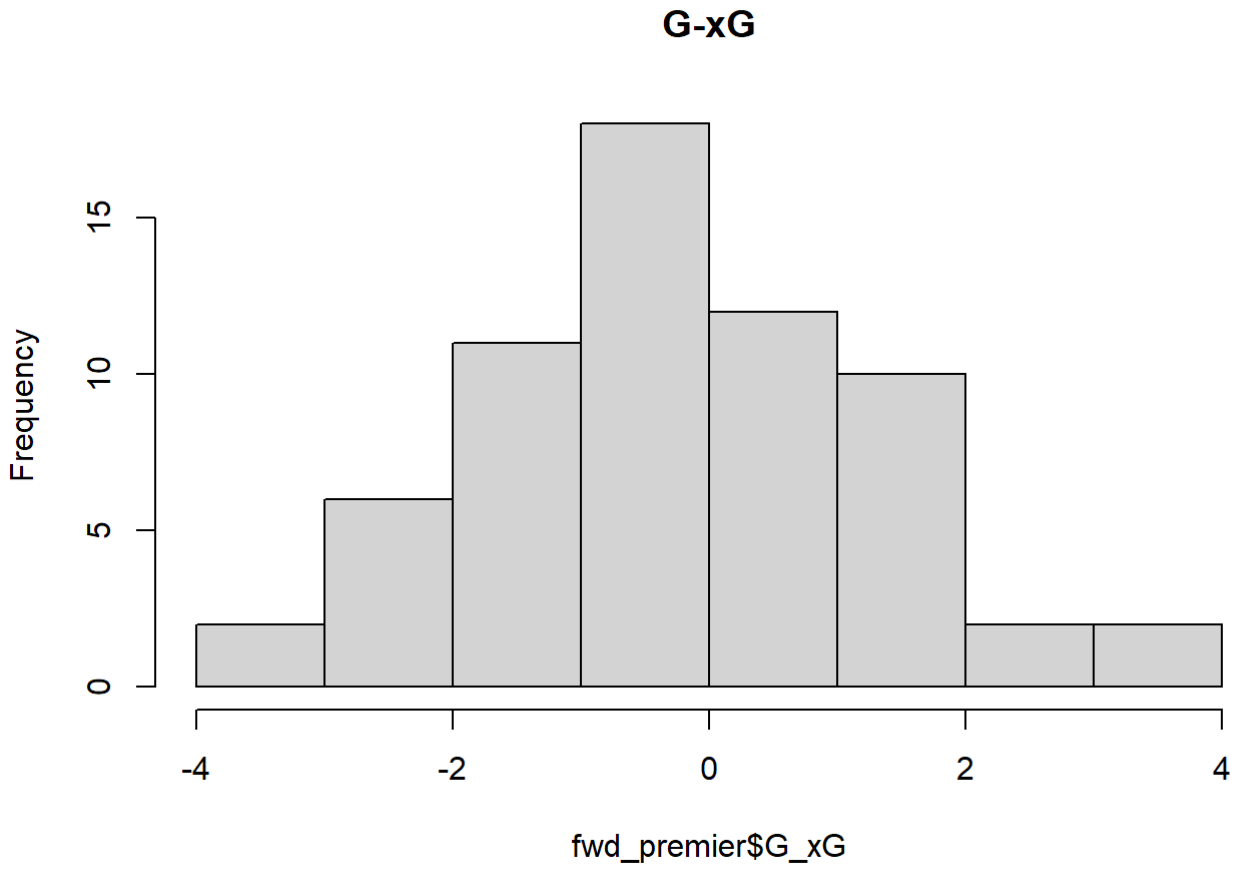

```
hist(fwd_laliga$G_xG, main='G-xG', breaks=10)
```



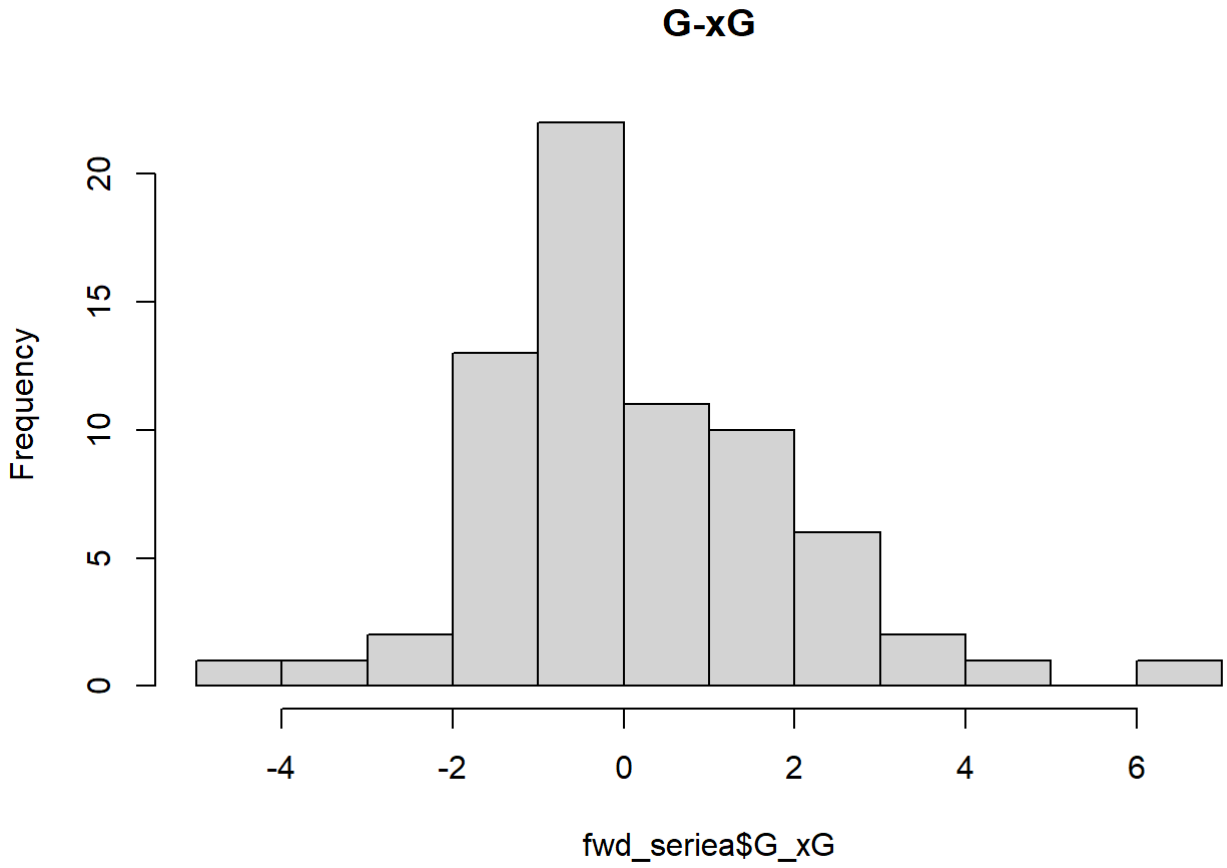

```
hist(fwd_ligue1$G_xG, main='G-xG', breaks=10)
```



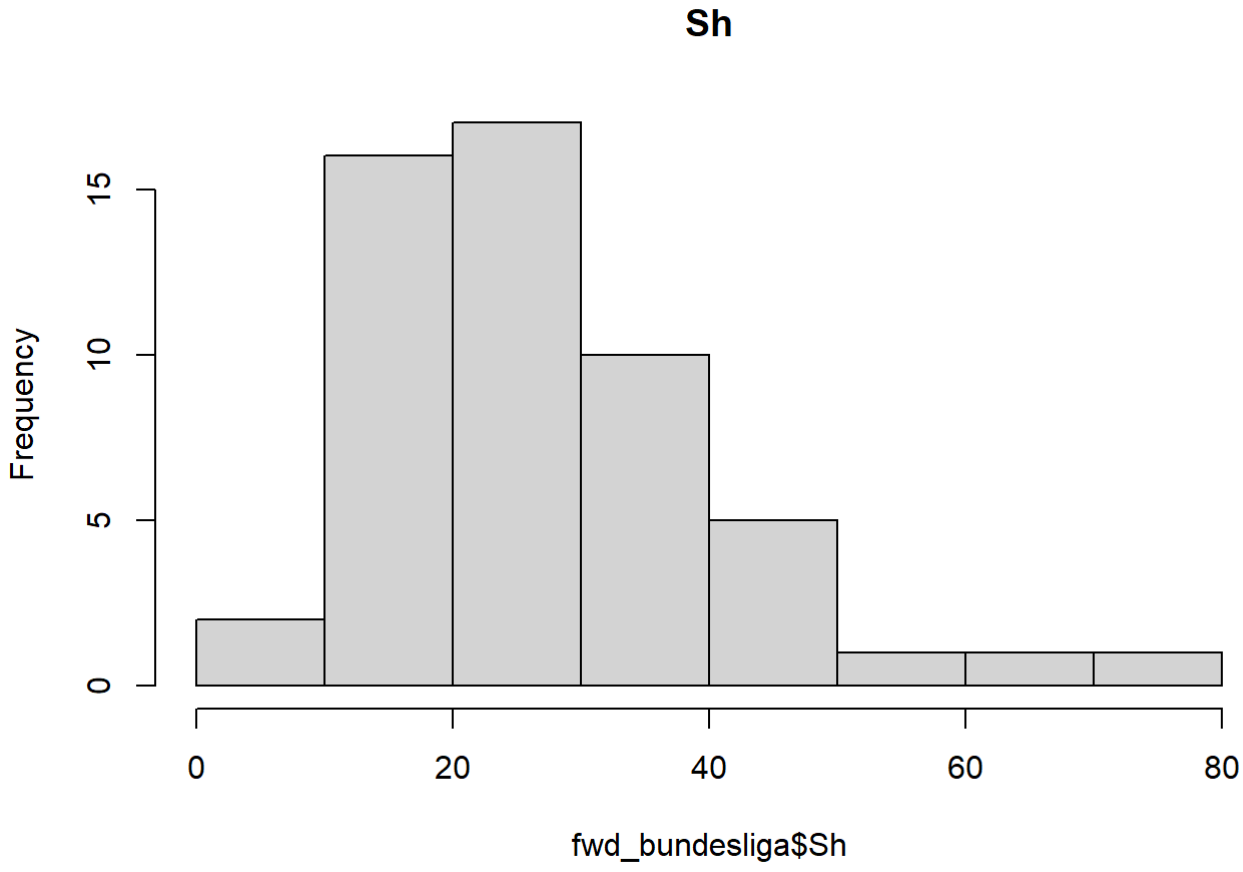
```
hist(fwd_premier$G_xG, main='G-xG', breaks=10)
```



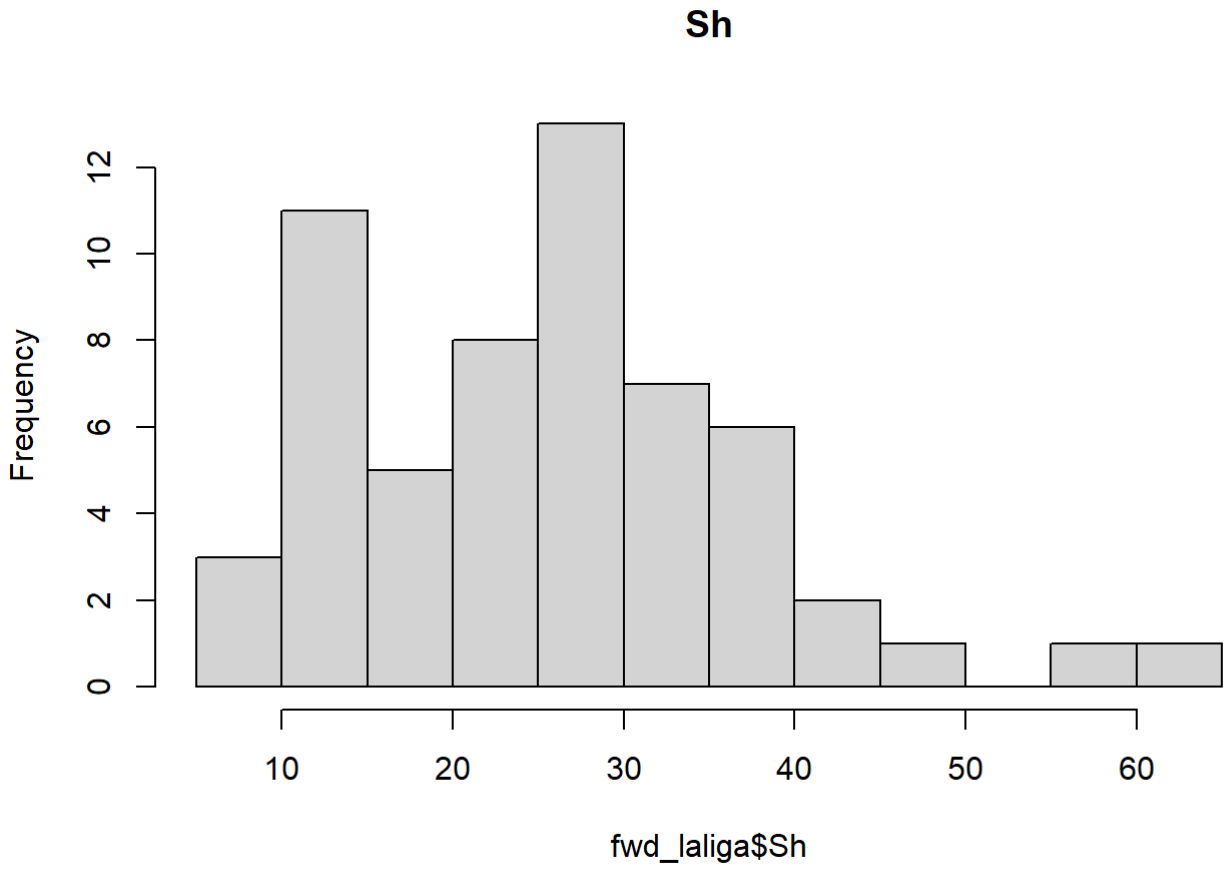

```
hist(fwd_seriea$G_xG, main='G-xG', breaks=10)
```



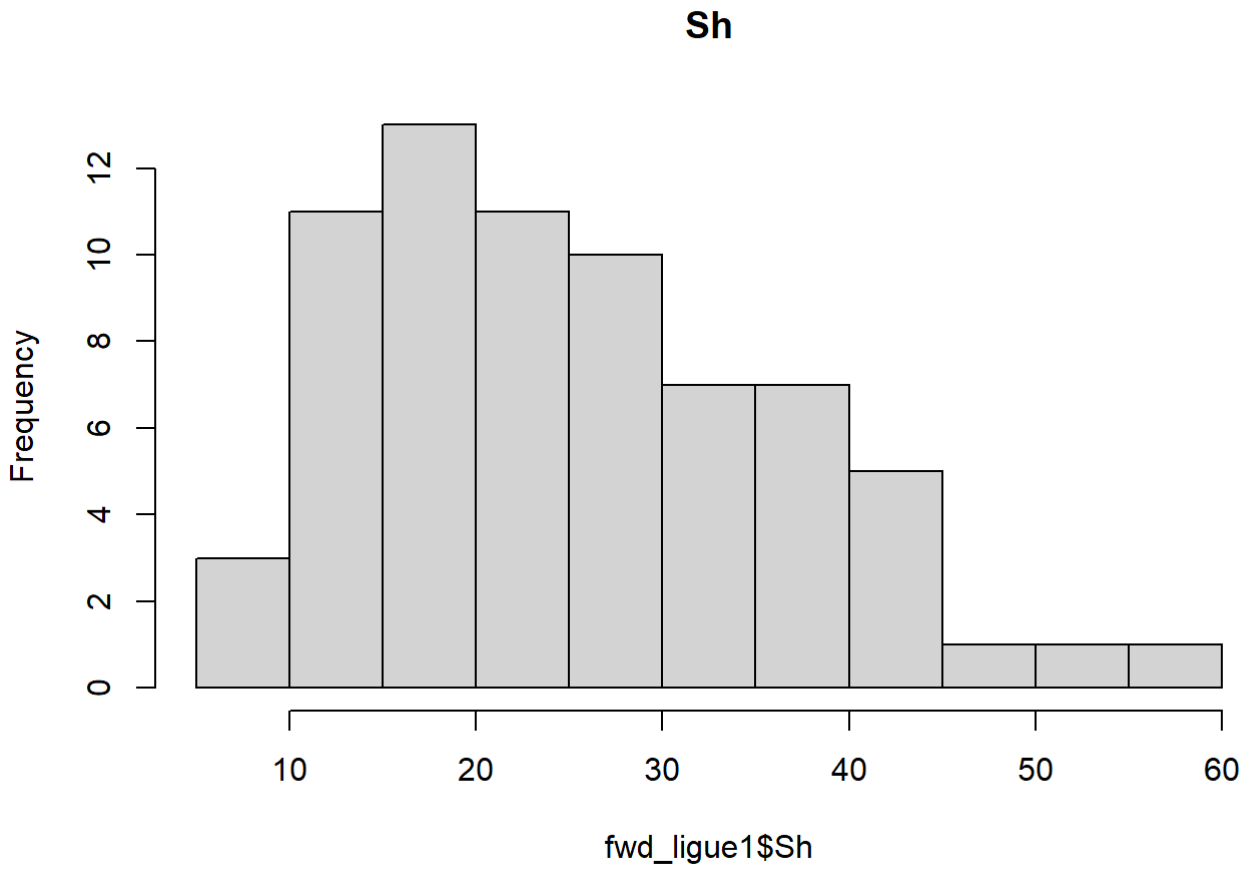

```
hist(fwd_bundesliga$Sh, main='Sh', breaks=10)
```



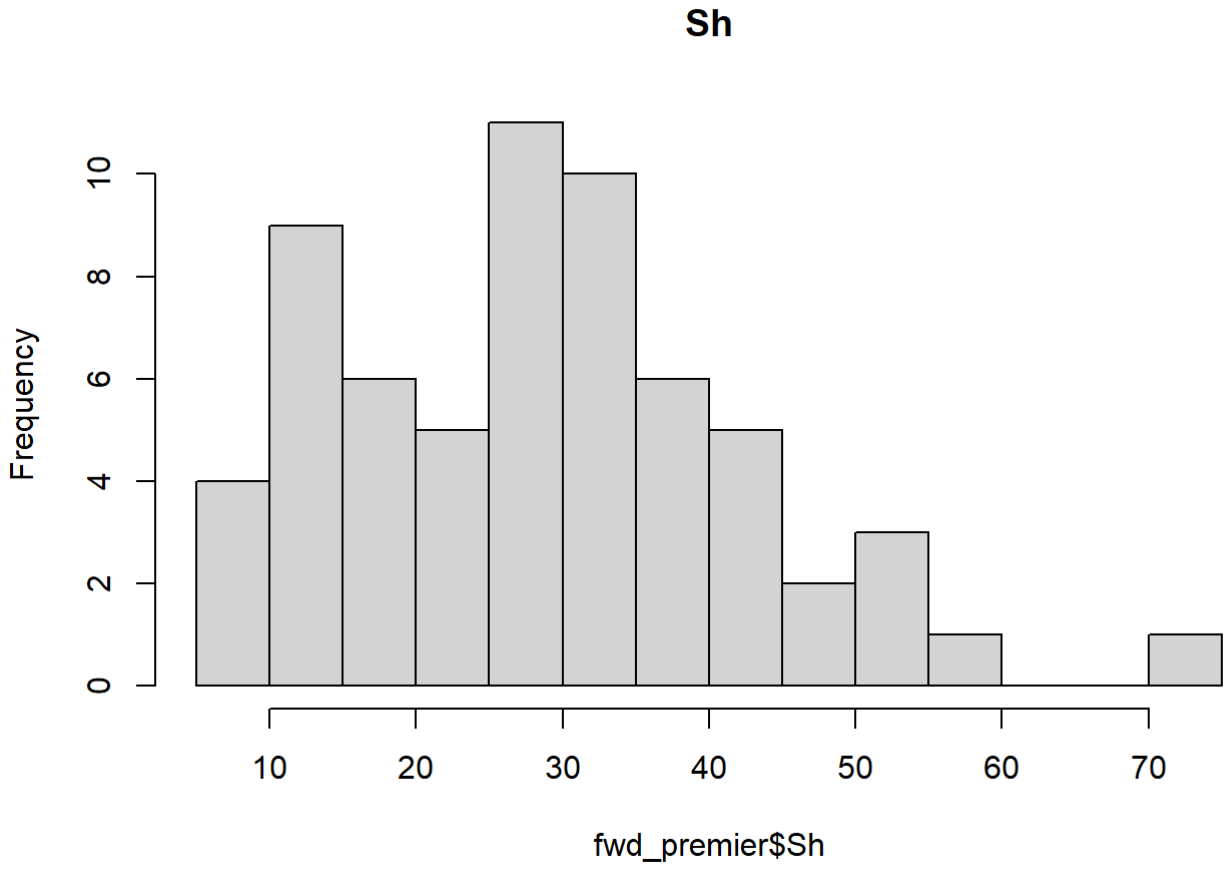
```
hist(fwd_laliga$Sh, main='Sh', breaks=10)
```



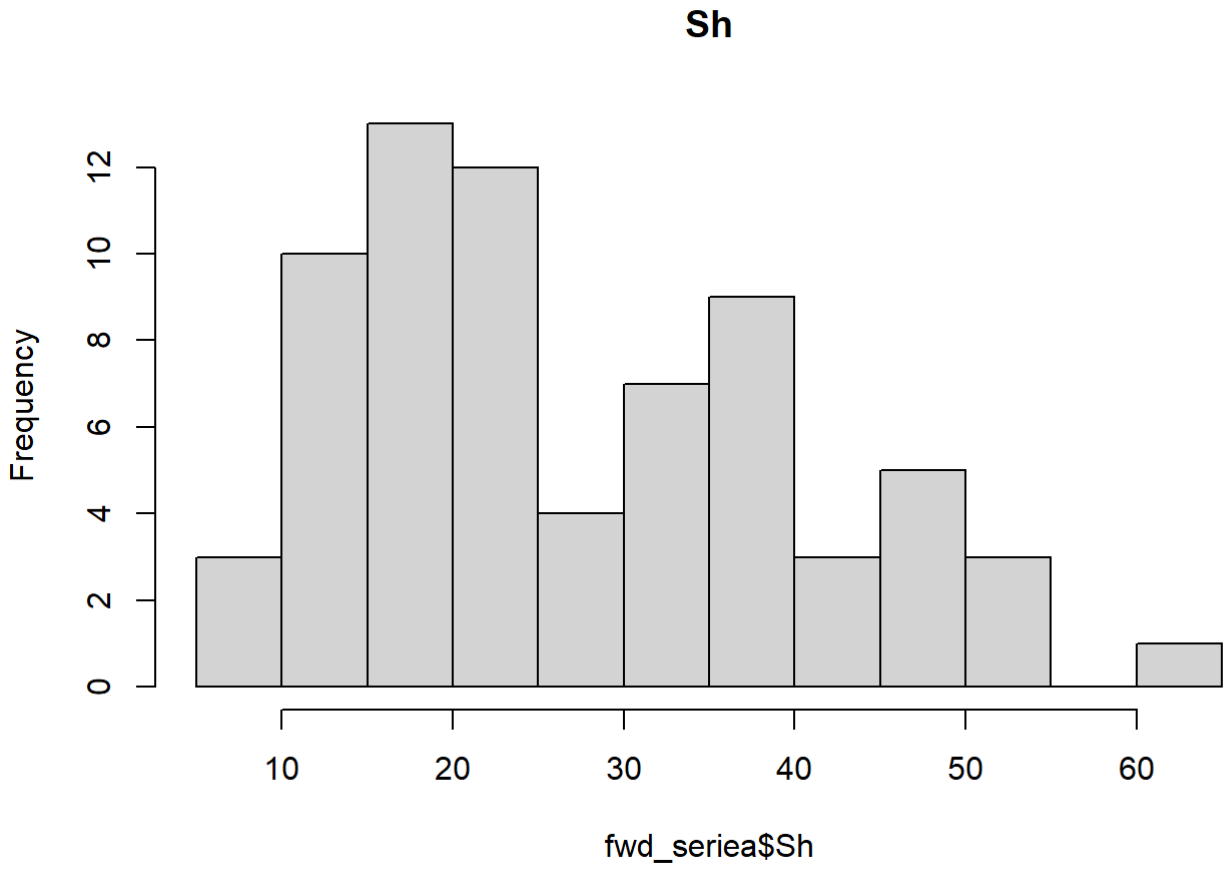

```
hist(fwd_ligue1$Sh, main='Sh', breaks=10)
```



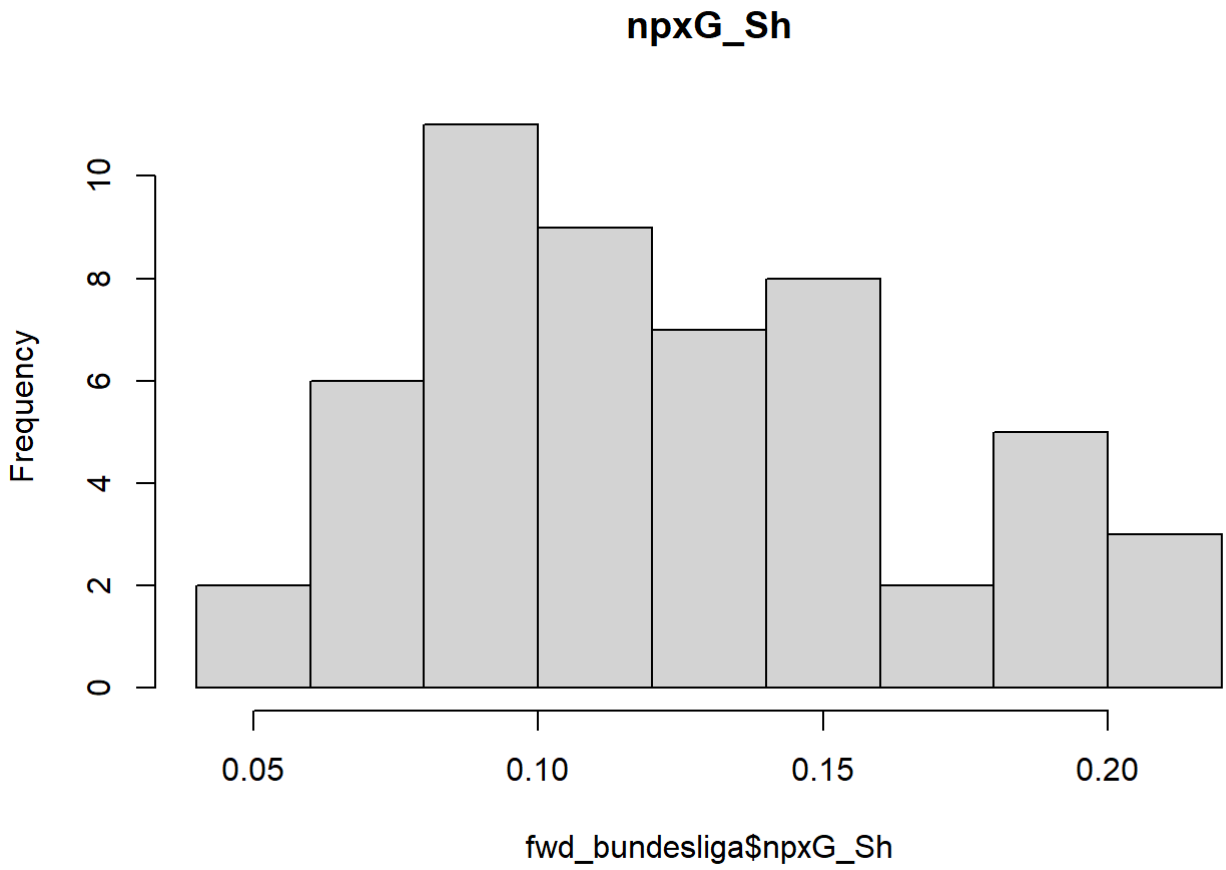

```
hist(fwd_premier$Sh, main='Sh', breaks=10)
```



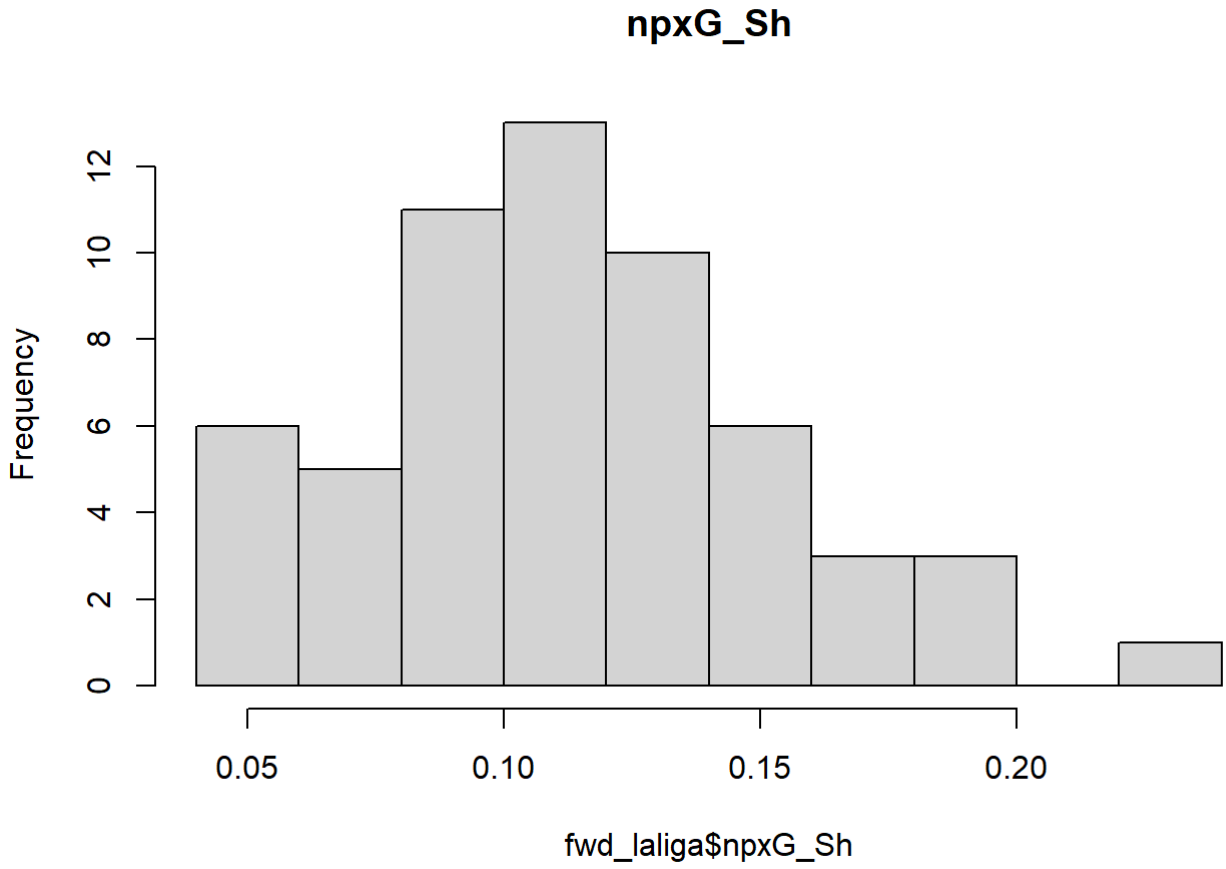
```
hist(fwd_seriea$Sh, main='Sh', breaks=10)
```



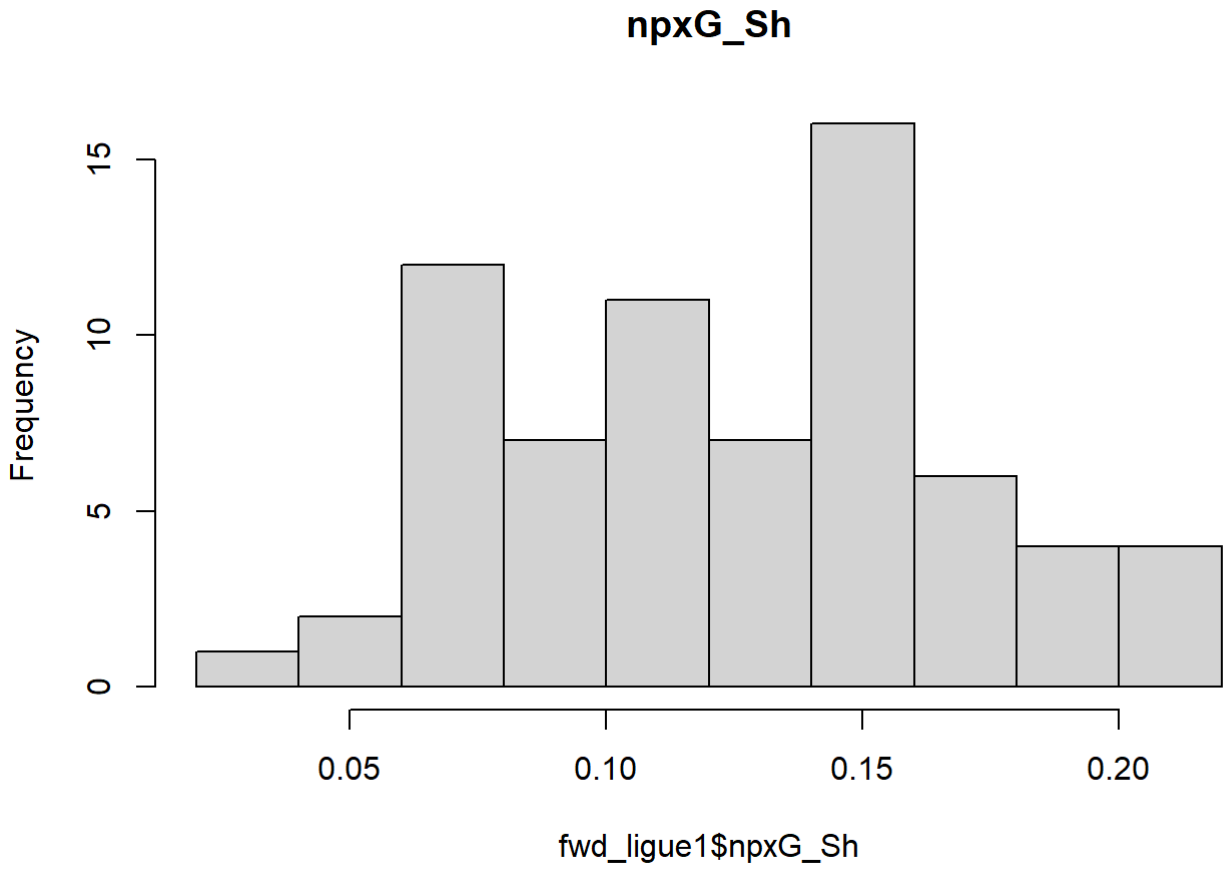

```
hist(fwd_bundesliga$npxG_Sh, main='npxG_Sh', breaks=10)
```



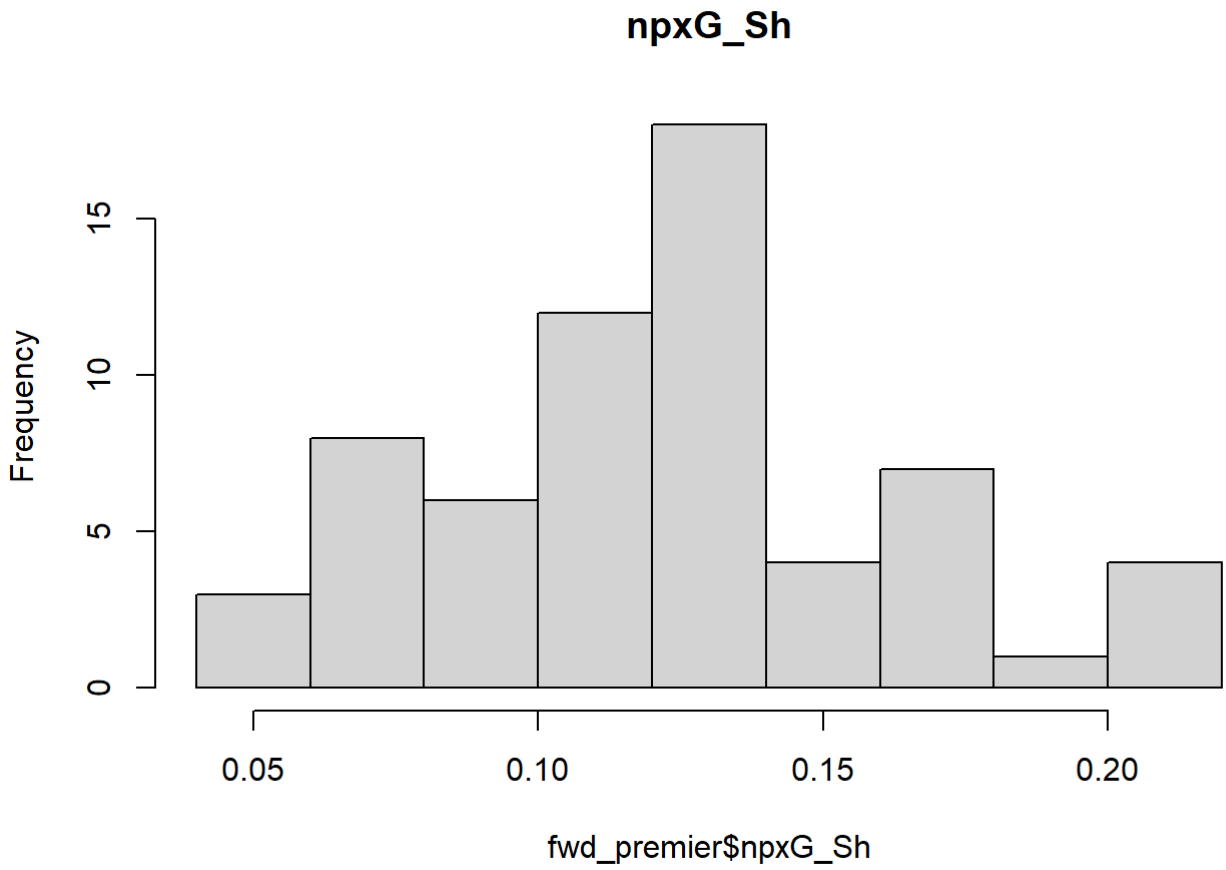

```
hist(fwd_laliga$npxG_Sh, main='npxG_Sh', breaks=10)
```



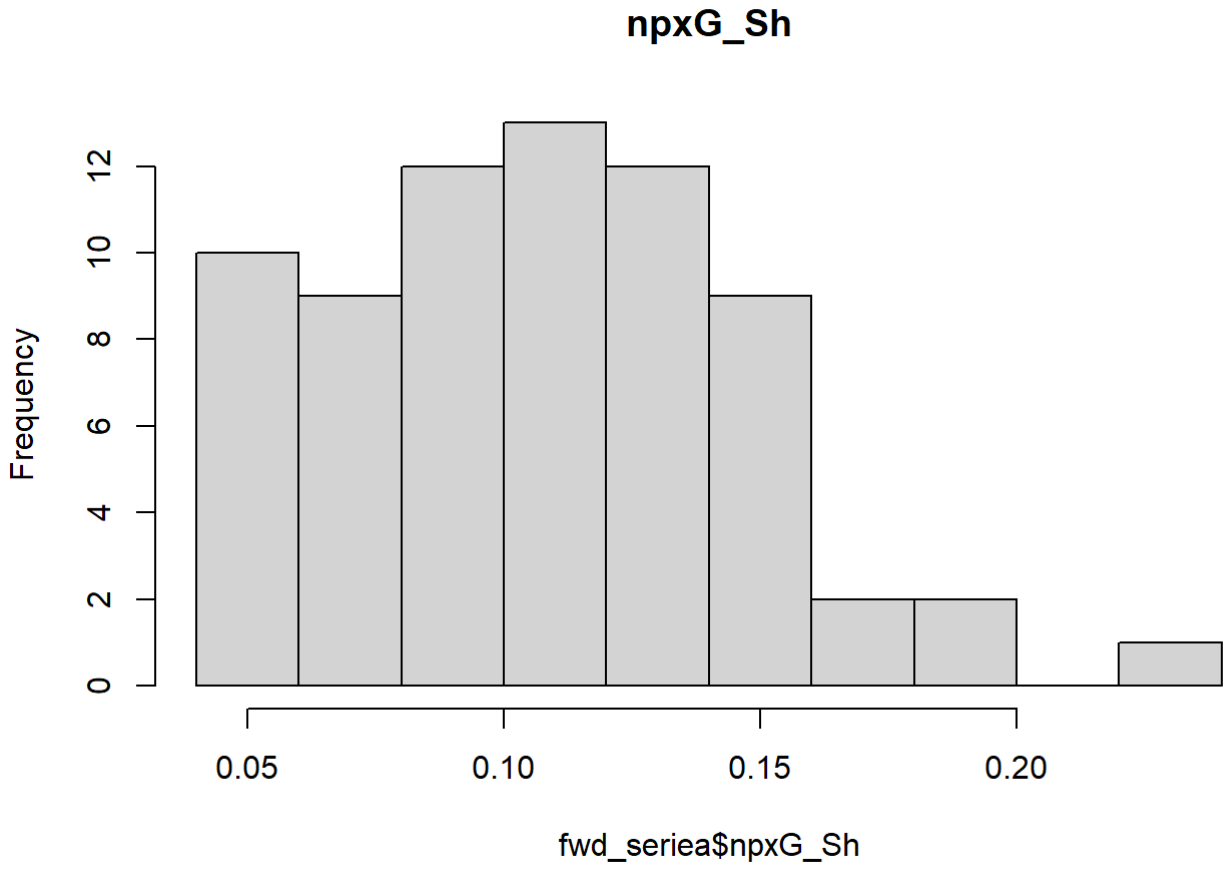
```
hist(fwd_ligue1$npxG_Sh, main='npxG_Sh', breaks=10)
```



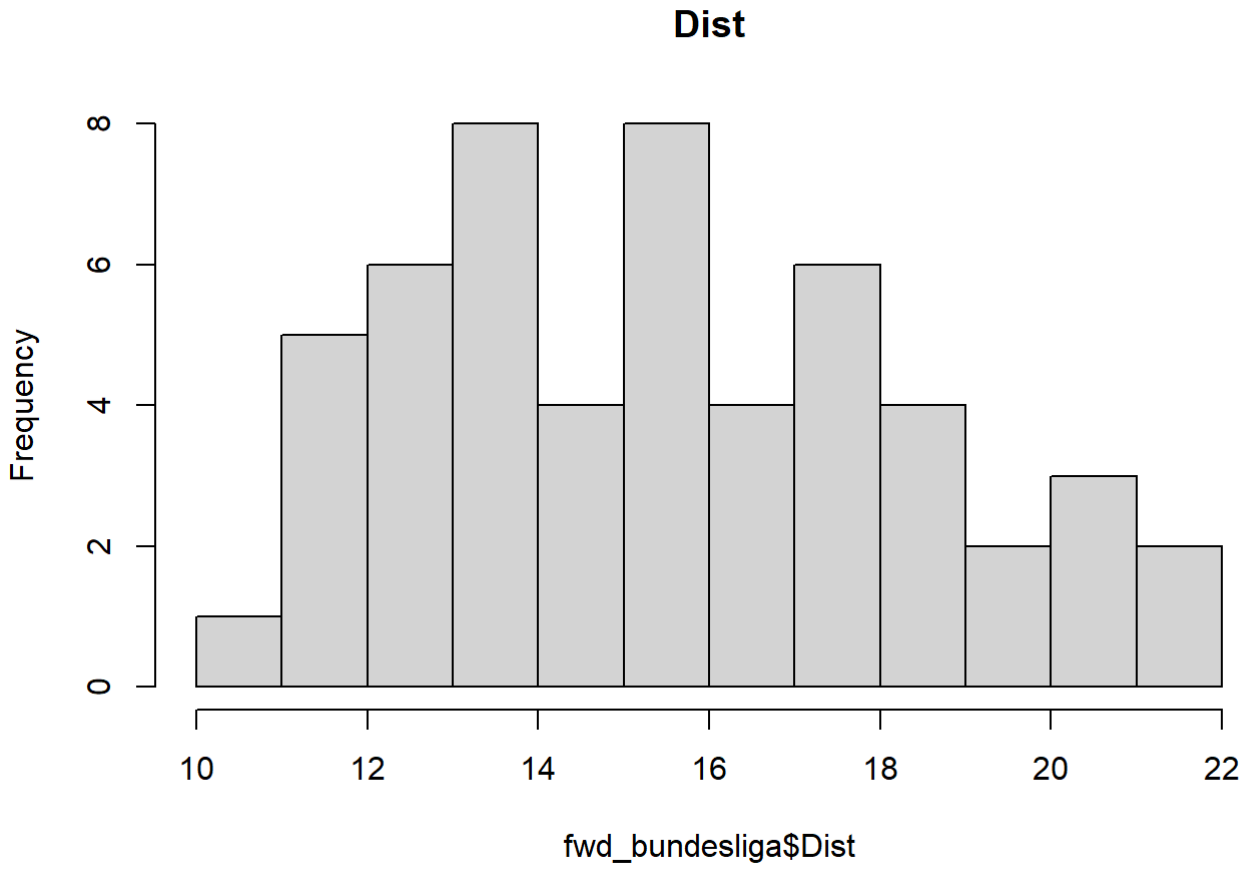

```
hist(fwd_premier$npxG_Sh, main='npxG_Sh', breaks=10)
```



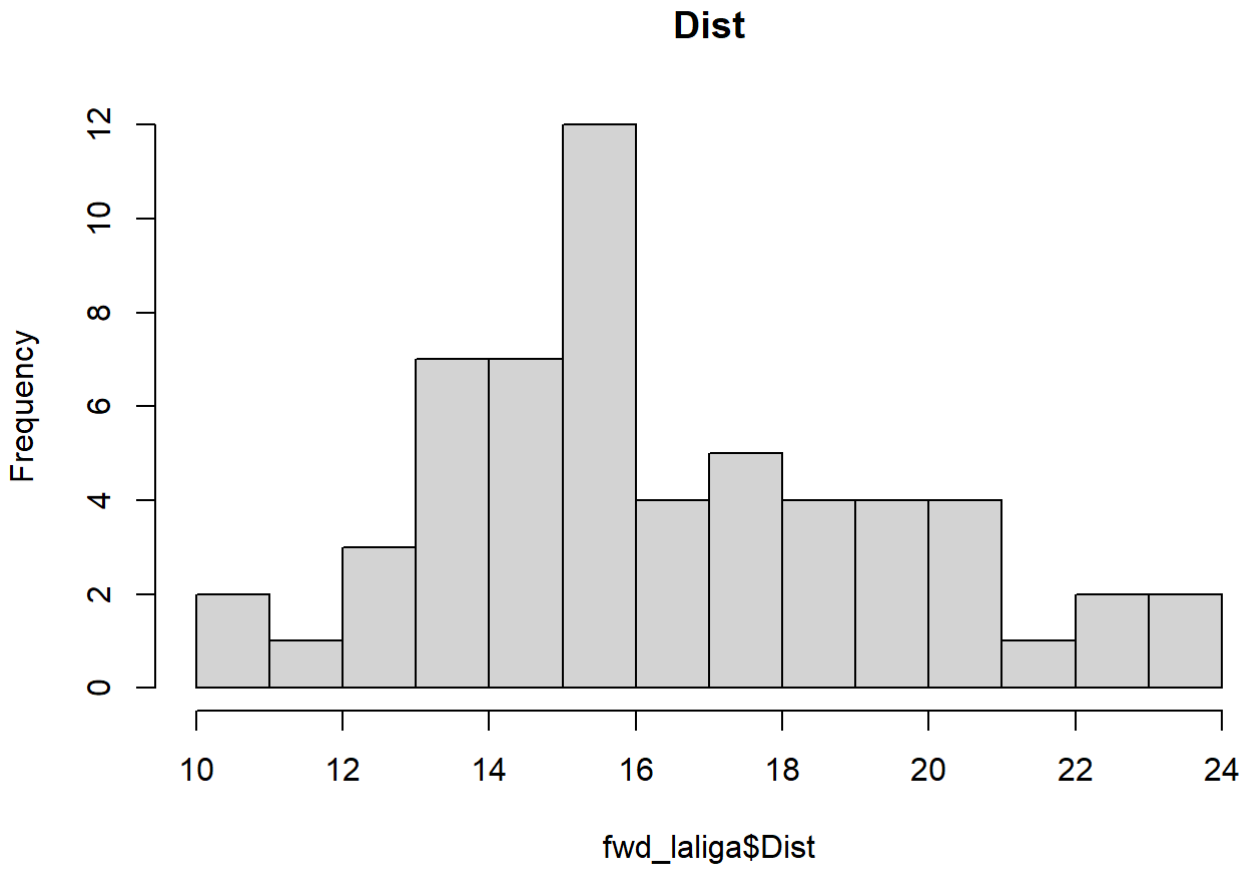

```
hist(fwd_seriea$npxG_Sh, main='npxG_Sh', breaks=10)
```



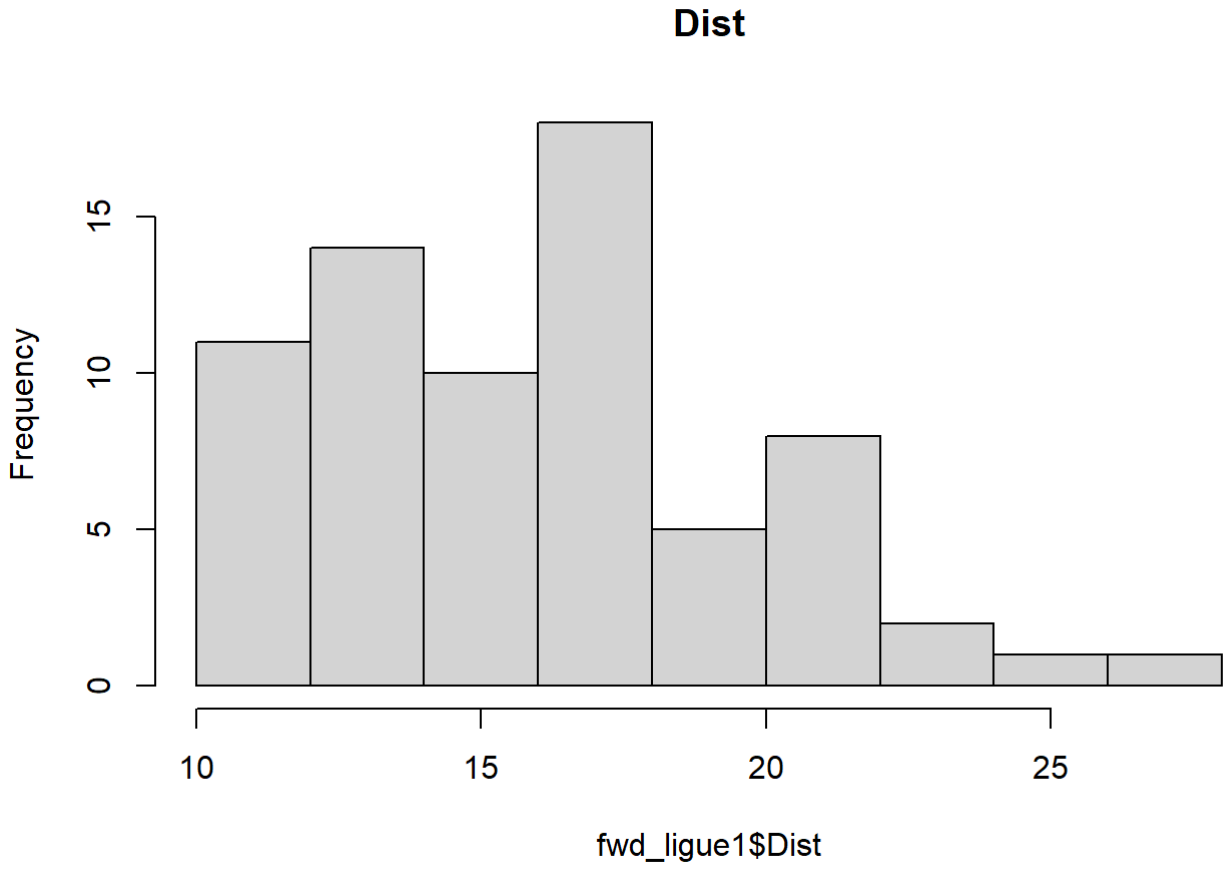
```
hist(fwd_bundesliga$Dist, main='Dist', breaks=10)
```



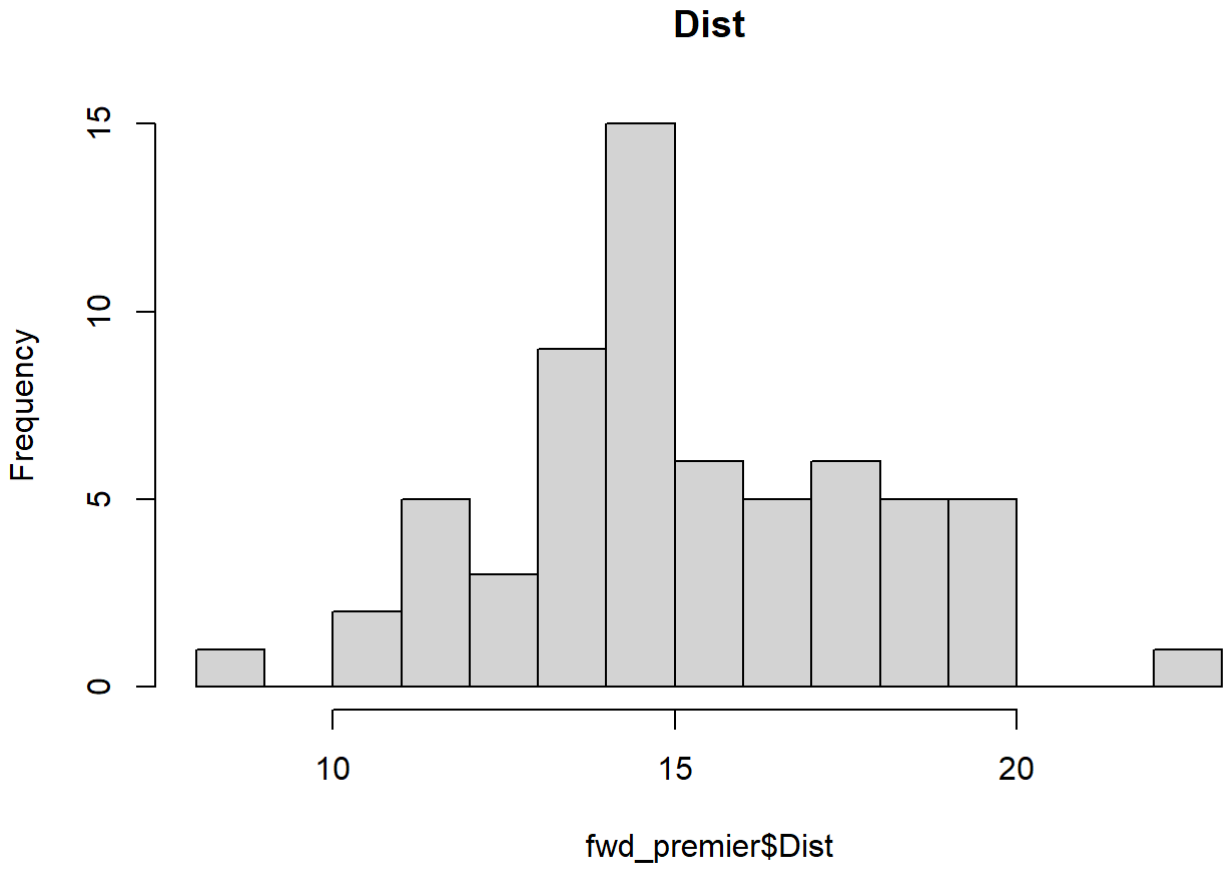

```
hist(fwd_laliga$Dist, main='Dist', breaks=10)
```



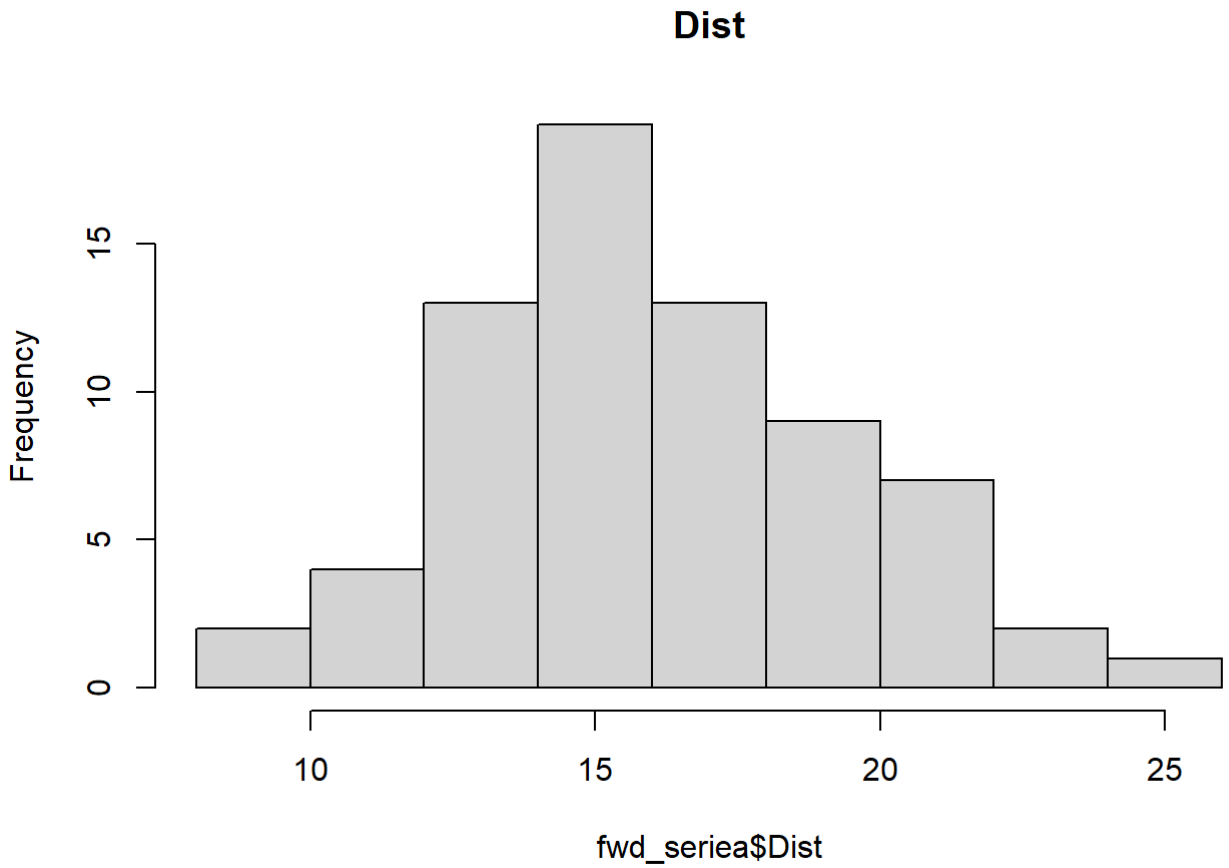

```
hist(fwd_ligue1$Dist, main='Dist', breaks=10)
```



```
hist(fwd_premier$Dist, main='Dist', breaks=10)
```




```
hist(fwd_seriea$Dist, main='Dist', breaks=10)
```



2.6 - Resolució del problema

Tal com s'ha comentat als apartats anteriors, el projecte ha servit per analitzar en profunditat les mètriques de finalització dels davanters de les 5 grans lligues europees. Mitjançant un preprocessament d'un dataset, s'ha pogut obtenir un arxiu "clean" per a poder realitzar l'anàlisi posterior. Amb l'anàlisi, s'han mostrat les diferències més importants de la lliga espanyola amb les altres lligues, una disminució en la quantitat dels tirs, però sobretot en l'eficàcia i el nivell d'aquests tirs. Es tira en pitjors condicions (menor npxG/Sh) i es té menor qualitat (pitjor G-xG), fet que confirma els temors que hi ha que la lliga espanyola hagi perdut part del seu nivell amb la marxa d'alguns dels grans davanters en els darrers anys.

Com a passos següents, es recomana incloure dades de temporades anteriors per poder fer una evolució històrica i saber si la tendència és només d'aquest any o ja es porta alguna temporada, a més de comprovar si en temproades passades era la lliga espanyola la que tenia els millors davanters.

Contribucions

Aquesta pràctica ha estat realitzada de forma individual per l'estudiant Jordi Puig Benages. Així doncs, tant la investigació prèvia, com la redaacció de les respostes i el desenvolupament del codi ha estat realitzat per aquest mateix estudiant.