Tipologia i cicle de vida de les dades: Pràctica 2: Com realitzar la neteja i la anàlisi de dades?

Autor: Adrià Setó Balcells i Miquel Arisa Fuente

Gener 2023

Contents

Descripció del dataset

Actualment, al món dels esports, explotar la informació et pot donar un immens avantatge. Com més informació tenen els equips tant d'ells mateixos com del seu rival, de forma col · lectiva o de forma individual per jugador, més fàcil és per ells preparar els partits i les competicions i augmenten molt més les possibilitats d'aconseguir la victòria. Els mètodes de scouting dins del món del futbol estan molt avançats i personalitzats per a cada un dels equips i jugadors, abarcant un mercat molt gran tenint dades fins i tot d'equips de categories inferiors. Això és una cosa que ens crida molt l'atenció, degut a la nostra passió pels esports i per les dades. Pel desenvolupament d'aquesta pràctica hem decidit escollir un dataset diferent al generat a la pràctica 1, on feiem un scrapping de les apostes juntament amb les seves quotes de diferents cases d'apostes. En aquest cas hem agafat varis datasets de la pàgina web Data Hub. Aquestes són dades dels partits de futbol jugats a tres grans lligues (Premier, La Liga, Bundesliga) durant quatre temporades (des de la 2018-2019 fins a la 2021-2022). Aquests datasets recullen tot de dades rellevants dins dels partits de futbol com ara córners, targetes, xuts a porteria... Aquestes dades estan sota la llicencia Open Data Commons i són subministrades per la web www.football-data.co.uk/ Ajuntem aquests dotze datasets per formar un gran dataset amb tota l'informació i d'on poder realitzar un bon estudi.

```
require(data.table)
library(datasets)
library(tidyverse)

# Carreguem el joc de dades

premier_1819 <- read_csv('dataset_lligues/Anglesa/premier_18-19.csv')
premier_1920 <- read_csv('dataset_lligues/Anglesa/premier_19-20.csv')
premier_2021 <- read_csv('dataset_lligues/Anglesa/premier_20-21.csv')
premier_2122 <- read_csv('dataset_lligues/Anglesa/premier_21-22.csv')

laliga_1819 <- read_csv('dataset_lligues/Espanyola/laliga_18-19.csv')
laliga_1920 <- read_csv('dataset_lligues/Espanyola/laliga_19-20.csv')
laliga_2021 <- read_csv('dataset_lligues/Espanyola/laliga_20-21.csv')
laliga_2122 <- read_csv('dataset_lligues/Espanyola/laliga_21-22.csv')

bundesliga_1819 <- read_csv('dataset_lligues/Alemana/bundesliga_18-19.csv')
bundesliga_1920 <- read_csv('dataset_lligues/Alemana/bundesliga_19-20.csv')
bundesliga_2021 <- read_csv('dataset_lligues/Alemana/bundesliga_20-21.csv')
bundesliga_2122 <- read_csv('dataset_lligues/Alemana/bundesliga_21-22.csv')
```

Veiem que tenim 306 registres a cada una de les temporades de la Bundesliga i 380 registres a cada una de les temporades de la Premier League i de La Liga. Això és degut al fet que la lliga alemana consta de 18 equips a la competició, mentre que a les altres dues en són 20 els que la formen. Veiem també que les tres últimes temporades de la Premier tenen 106 variables, mentre que l'altre en té 62, això passa exactament de la mateixa manera amb les altres dues lligues, tenint-ne 105 a les tres últimes i 61 a la restant. Hi ha una diferència d'una variable entre la lliga anglesa i les altres dues, això és perquè als datasets de la Premier hi ha la variable Referee, que guarda l'àrbitre principal del partit, mentre que la lliga alemana i la lliga espanyola aquesta variable no la guarden.

Objectiu de l'anàlisi

Comportament dels àrbitres davant el públic local

És conegut que la majoria d'equips prefereixen jugar els partits als seu estadi, de forma local, que al del seu rival, de forma visitant. Això és degut a la comoditat del camp, però també a l'efecte públic que efecta sobre el rival i sobre la presió a l'àrbitre.

Els àrbitres són educats i entrenats per suportar la presió, de la mateixa manera que els jugadors, però pensem que com a persones també els afecta. Així volem comprovar si els àrbitres tenen un comportament que tendeixi a afavorir més a l'equip local que al visitant, i si l'experiència d'aquests juga a favor o en contra.

Restriccions COVID-19 als estadis

Durant el març de 2020 es van implementar estrictes mesures arreu del món, posades en marxa per intentar evitar la propagació de la COVID-19 i els seus efectes, que principalment minimitzaven el contacte entre persones. Això va comportar l'aturada temporal de totes les lligues europees, algunes de les quals fins i tot van donar la temporada per finalitzada tot i tenir partits pendents per jugar, com en el cas de la Ligue 1.

Entre maig i juny de 2020, les competicions de futbol professional que no van cancel · lar la resta de partits pendents, van tornar a posar-se en marxa, però amb restriccions totals d'aforament. Així doncs, el tram final de la temporada 2019-2020 es va jugar sense públic en les tres lligues analitzades (La Liga, Premier League, Bundesliga).

La temporada següent, 2020-2021, va tenir unes restriccions similars, tot i que es va permetre un aforament bastant reduït en un nombre molt limitat de jornades.

Finalment, durant la temporada 2021-2022 es va anar normalitzant la situació i, tot i que hi havia un límit d'aforament que variava segons el país i la regió, la majoria dels partits es van jugar ja amb una assistència de públic de milers de persones.

Degut a l'impacte que se li atribueix als aficionats al rendiment dels equips tot i no participar directament en el joc, s'analitzaran els resultats dels partits (victòria local/visitant o empat) segons múltiples variables, per poder comprovar si les restriccions d'aforament per la COVID-19 han pogut influir en els resultats dels partits.

Integració i selecció

```
# Seleccionem les columnes que ens interessen

premier_1819 <- na.omit(premier_1819, row.names=NULL)[,2:26]
premier_1920 <- na.omit(premier_1920, row.names=NULL)[,2:27]
premier_2021 <- na.omit(premier_2021, row.names=NULL)[,2:27]
premier_2122 <- na.omit(premier_2122, row.names=NULL)[,2:27]

laliga_1819 <- na.omit(laliga_1819, row.names=NULL)[,2:25]
laliga_1920 <- na.omit(laliga_1920, row.names=NULL)[,2:26]
laliga_2021 <- na.omit(laliga_2021, row.names=NULL)[,2:26]
laliga_2122 <- na.omit(laliga_2122, row.names=NULL)[,2:26]

bundesliga_1819 <- na.omit(bundesliga_1819, row.names=NULL)[,2:25]
bundesliga_1920 <- na.omit(bundesliga_1920, row.names=NULL)[,2:26]
bundesliga_2021 <- na.omit(bundesliga_2021, row.names=NULL)[,2:26]
```

```r
bundesliga_2122 <- na.omit(bundesliga_2122, row.names=NULL)[,2:26]


# Afegim variables omplertes com a NA (valors perduts) als datasets que els n'hi falten algunes (Time i

premier_1819$Time <- NA
laliga_1819$Time <- NA
bundesliga_1819$Time <- NA

laliga_1819$Referee <- NA
laliga_1920$Referee <- NA
laliga_2021$Referee <- NA
laliga_2122$Referee <- NA

bundesliga_1819$Referee <- NA
bundesliga_1920$Referee <- NA
bundesliga_2021$Referee <- NA
bundesliga_2122$Referee <- NA


# Afegim variables (Season i Competition) per reconeixer la temporada i la competició

premier_1819$Season <- "2018-2019"
premier_1920$Season <- "2019-2020"
premier_2021$Season <- "2020-2021"
premier_2122$Season <- "2021-2022"

laliga_1819$Season <- "2018-2019"
laliga_1920$Season <- "2019-2020"
laliga_2021$Season <- "2020-2021"
laliga_2122$Season <- "2021-2022"

bundesliga_1819$Season <- "2018-2019"
bundesliga_1920$Season <- "2019-2020"
bundesliga_2021$Season <- "2020-2021"
bundesliga_2122$Season <- "2021-2022"

premier_1819$Competition <- "Premier League"
premier_1920$Competition <- "Premier League"
premier_2021$Competition <- "Premier League"
premier_2122$Competition <- "Premier League"

laliga_1819$Competition <- "La Liga"
laliga_1920$Competition <- "La Liga"
laliga_2021$Competition <- "La Liga"
laliga_2122$Competition <- "La Liga"

bundesliga_1819$Competition <- "Bundesliga"
bundesliga_1920$Competition <- "Bundesliga"
bundesliga_2021$Competition <- "Bundesliga"
bundesliga_2122$Competition <- "Bundesliga"
```

```r
# Unim tots els datasets en un de sol anomenat football_matches

football_matches <- rbind(
  premier_2122, premier_2021, premier_1920, premier_1819,
  laliga_2122, laliga_2021, laliga_1920, laliga_1819,
  bundesliga_2122, bundesliga_2021, bundesliga_1920, bundesliga_1819
)
```

Neteja de les dades

Les dades contenen zeros o elements buits? Gestiona cadascun d'aquests casos

Veiem un resum dels valors zero o elements buits del joc de dades amb el qual treballarem.

```r
print('NA all_matches')
```

```
## [1] "NA all_matches"
```

```r
colSums(is.na(football_matches))
```

```
##        Date        Time    HomeTeam    AwayTeam        FTHG        FTAG
##           0        1065           0           0           0           0
##         FTR        HTHG        HTAG         HTR     Referee          HS
##           0           0           0           0        2707           0
##          AS         HST         AST          HF          AF          HC
##           0           0           0           0           0           0
##          AC          HY          AY          HR          AR       B365H
##           0           0           0           0           0           0
##       B365D       B365A      Season Competition
##           0           0           0           0
```

```r
print('Blancs all_matches')
```

```
## [1] "Blancs all_matches"
```

```r
colSums(football_matches=="")
```

```
##        Date        Time    HomeTeam    AwayTeam        FTHG        FTAG
##           0          NA           0           0           0           0
##         FTR        HTHG        HTAG         HTR     Referee          HS
##           0           0           0           0          NA           0
##          AS         HST         AST          HF          AF          HC
##           0           0           0           0           0           0
##          AC          HY          AY          HR          AR       B365H
##           0           0           0           0           0           0
##       B365D       B365A      Season Competition
##           0           0           0           0
```

Podem veure com només tenim variables on no hi ha valors a les variables Time i Referee, que són les dues a les quals els hi hem ficat nosaltres per tal de que tots els datasets tinguessin les mateixes columnes i poder-los unir en un sol dataset gran. A la resta de columnes veiem on no hi ha cap valor zero o element buit, cosa positiva per al nostre anàlisi.

Identifica i gestiona els valors extrems

Generarem un resum de tot el dataset complet per tenir una primera impresió d'aquest i poder veure si hi ha algun valor que ens faci sospitar d'algun error. Per això veurem el màxim, mínim, mitjana, mediana... de cada variable.

```
summary(football_matches)
```

```
##      Date               Time            HomeTeam           AwayTeam
##  Length:4222        Length:4222       Length:4222        Length:4222
##  Class :character   Class1:hms        Class :character   Class :character
##  Mode  :character   Class2:difftime   Mode  :character   Mode  :character
##                     Mode  :numeric
##
##
##       FTHG            FTAG           FTR                HTHG
##  Min.   :0.000   Min.   :0.000   Length:4222        Min.   :0.000
##  1st Qu.:1.000   1st Qu.:0.000   Class :character   1st Qu.:0.000
##  Median :1.000   Median :1.000   Mode  :character   Median :0.000
##  Mean   :1.528   Mean   :1.251                      Mean   :0.667
##  3rd Qu.:2.000   3rd Qu.:2.000                      3rd Qu.:1.000
##  Max.   :9.000   Max.   :9.000                      Max.   :6.000
##       HTAG            HTR              Referee                HS
##  Min.   :0.0000   Length:4222        Length:4222        Min.   : 0.00
##  1st Qu.:0.0000   Class :character   Class :character   1st Qu.:10.00
##  Median :0.0000   Mode  :character   Mode  :character   Median :13.00
##  Mean   :0.5616                                        Mean   :13.36
##  3rd Qu.:1.0000                                        3rd Qu.:17.00
##  Max.   :6.0000                                        Max.   :36.00
##        AS              HST              AST               HF
##  Min.   : 0.00   Min.   : 0.000   Min.   : 0.000   Min.   : 0.00
##  1st Qu.: 8.00   1st Qu.: 3.000   1st Qu.: 2.000   1st Qu.: 9.00
##  Median :11.00   Median : 4.000   Median : 4.000   Median :12.00
##  Mean   :11.15   Mean   : 4.678   Mean   : 3.959   Mean   :11.89
##  3rd Qu.:14.00   3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.:14.00
##  Max.   :32.00   Max.   :17.000   Max.   :20.000   Max.   :28.00
##        AF              HC               AC               HY
##  Min.   : 1.00   Min.   : 0.000   Min.   : 0.000   Min.   :0.000
##  1st Qu.: 9.00   1st Qu.: 3.000   1st Qu.: 3.000   1st Qu.:1.000
##  Median :12.00   Median : 5.000   Median : 4.000   Median :2.000
##  Mean   :11.94   Mean   : 5.323   Mean   : 4.487   Mean   :1.911
##  3rd Qu.:14.00   3rd Qu.: 7.000   3rd Qu.: 6.000   3rd Qu.:3.000
##  Max.   :30.00   Max.   :19.000   Max.   :16.000   Max.   :8.000
##        AY              HR               AR             B365H
##  Min.   :0.000   Min.   :0.00000   Min.   :0.0000   Min.   : 1.050
##  1st Qu.:1.000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.: 1.660
##  Median :2.000   Median :0.00000   Median :0.0000   Median : 2.250
##  Mean   :2.061   Mean   :0.06869   Mean   :0.0874   Mean   : 2.863
##  3rd Qu.:3.000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.: 3.200
##  Max.   :8.000   Max.   :2.00000   Max.   :2.0000   Max.   :23.000
##      B365D            B365A            Season          Competition
##  Min.   : 2.750   Min.   : 1.070   Length:4222        Length:4222
##  1st Qu.: 3.400   1st Qu.: 2.300   Class :character   Class :character
##  Median : 3.700   Median : 3.300   Mode  :character   Mode  :character
```

```
##  Mean   : 4.138   Mean   : 4.529
##  3rd Qu.: 4.330   3rd Qu.: 5.250
##  Max.   :17.000   Max.   :41.000
```

Veiem com al resum de totes les variables, el qual ens deixa veure els màxims, mínims, mitjana i mediana de cada una de les variables, no apreciem cap dada que s'escapi de la normalitat del que és el món del futbol. Veiem molts mínims a zero i alguns valors màxims alts que a vegades pot costar de veure al món del futbol, però no impossibles.

Anàlisi de les dades

Selecció dels grups de dades que es volen analitzar/comparar (p. e., si es volen comparar grups de dades, quins són aquests grups i quins tipus d'anàlisi s'aplicaran?)

```
# Creem un dataset amb tot el recull de dades dels àrbitres, partits jugats, tagetes ensenyades, faltes

total_referee_faults <- aggregate(HF+AF ~ Referee, football_matches, sum)
total_referee_yellow_cards <- aggregate(HY+AY ~ Referee, football_matches, sum)
total_referee_red_cards <- aggregate(HR+AR ~ Referee, football_matches, sum)
mean_referee_faults <- aggregate(HF+AF ~ Referee, football_matches, mean)
mean_referee_yellow_cards <- aggregate(HY+AY ~ Referee, football_matches, mean)
mean_referee_red_cards <- aggregate(HR+AR ~ Referee, football_matches, mean)

referee_result <- dcast(setDT(football_matches), Referee~FTR, length)
referee_result = referee_result[-1,]

count_referee <- count(football_matches, Referee)
count_referee <- head(count_referee, - 1)

referee <- mean_referee_yellow_cards
names(referee)[names(referee) == 'Referee'] <- 'name'
names(referee)[names(referee) == 'HY + AY'] <- 'avg_Y'

referee['avg_R'] <- mean_referee_red_cards[2]
referee['avg_F'] <- mean_referee_faults[2]
referee['total_Y'] <- total_referee_yellow_cards[2]
referee['total_R'] <- total_referee_red_cards[2]
referee['total_F'] <- total_referee_faults[2]

referee['total_games'] <- count_referee$n
referee['H'] <- referee_result$H
referee['D'] <- referee_result$D
referee['A'] <- referee_result$A

# Definim l'esperiència dels arbitres:  10 partits o menys -> Poca experiència
#                                       11 - 89 partits    -> Mitja experiència
#                                       90 partits o més   -> Molta experiència

referee$experience <- ifelse(referee$total_games <= 10, "Low", ifelse(referee$total_games >= 90, "High"
```

```r
# Creem un dataset segons el nivell d'experiència i el percentatge de victories locals, visitants i empa

experience_total_games <- aggregate(total_games ~ experience, referee, sum)

experience_total_home <- aggregate(H ~ experience, referee, sum)

experience_total_draw <- aggregate(D ~ experience, referee, sum)

experience_total_away <- aggregate(A ~ experience, referee, sum)

experience_stats = experience_total_games

experience_stats['H'] <- experience_total_home$H/experience_total_games$total_games
experience_stats['D'] <- experience_total_draw$D/experience_total_games$total_games
experience_stats['A'] <- experience_total_away$A/experience_total_games$total_games

experience_stats <- experience_stats[,!names(experience_stats) %in% c("total_games")]

# S'afegeixen columnes addicionals necessàries en l'anàlisi posterior
football_matches <- football_matches %>%
  mutate(Result = case_when(FTR == "A" ~ "Away",
                            FTR == "D" ~ "Draw",
                            FTR == "H" ~ "Home"))
# Conversió de data
football_matches$Date <- as.Date(football_matches$Date , format = "%d/%m/%y")

# Restriccions covid (Temporada 19-20)
football_matches$CovidRestrictions <- "No"
football_matches$CovidRestrictions[football_matches$Date > as.Date("31/03/2020", format = "%d/%m/%y") &
                                   football_matches$Season == "2019-2020"] <- "Yes"

# Resultats per temporada
result_by_season <- count(football_matches, Season, Result, name = "NumMatches") %>%
  group_by(Season) %>% mutate(percent = 100*NumMatches/sum(NumMatches))

# Resultats per temporada i competició
result_by_competition_season <- count(football_matches, Competition, Season, Result, name = "NumMatches
  group_by(Competition, Season) %>% mutate(percent = 100*NumMatches/sum(NumMatches))

# Resultats per Post/pre covid i competició (2019-2020)
result_by_competition_covid <- football_matches %>% filter(Season == "2019-2020") %>%
  count(Competition, CovidRestrictions, Result, name = "NumMatches") %>%
  group_by(Competition, CovidRestrictions) %>% mutate(percent = 100*NumMatches/sum(NumMatches))
```

Comprovació de la normalitat i homogeneïtat de la variància

S'analitzaran els valors de targetes grogues i vermelles per comprovar si es tracta d'una variança normal i homegenïa.

```r
shapiro.test(football_matches$HY)
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  football_matches$HY
## W = 0.91663, p-value < 2.2e-16
```

```
shapiro.test(football_matches$AY)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  football_matches$AY
## W = 0.92759, p-value < 2.2e-16
```

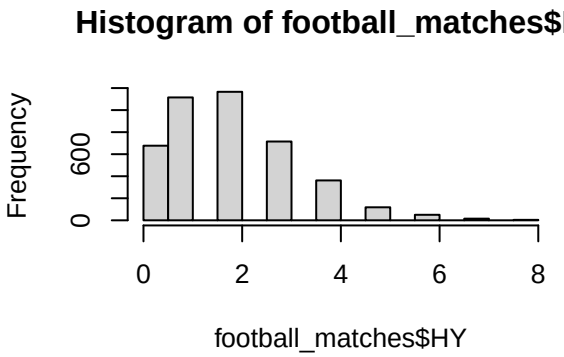
```
shapiro.test(football_matches$HR)
```

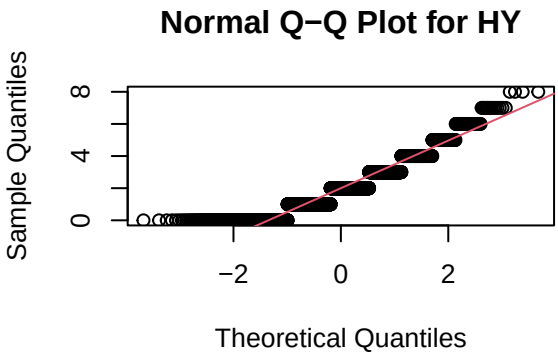
```
##
##  Shapiro-Wilk normality test
##
## data:  football_matches$HR
## W = 0.2699, p-value < 2.2e-16
```

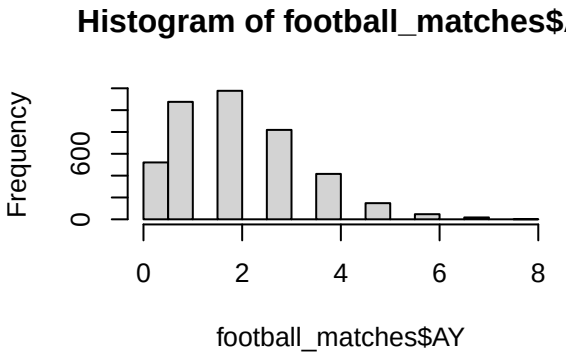
```
shapiro.test(football_matches$AR)
```

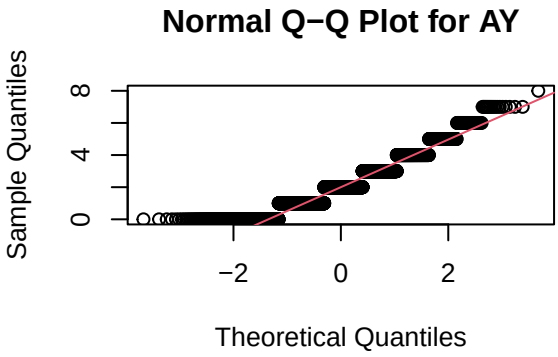
```
##
##  Shapiro-Wilk normality test
##
## data:  football_matches$AR
## W = 0.31216, p-value < 2.2e-16
```

```
par(mfrow=c(2,2))
hist(football_matches$HY)
qqnorm(football_matches$HY, main="Normal Q-Q Plot for HY")
qqline(football_matches$HY,col=2)
hist(football_matches$AY)
qqnorm(football_matches$AY, main="Normal Q-Q Plot for AY")
qqline(football_matches$AY,col=2)
```

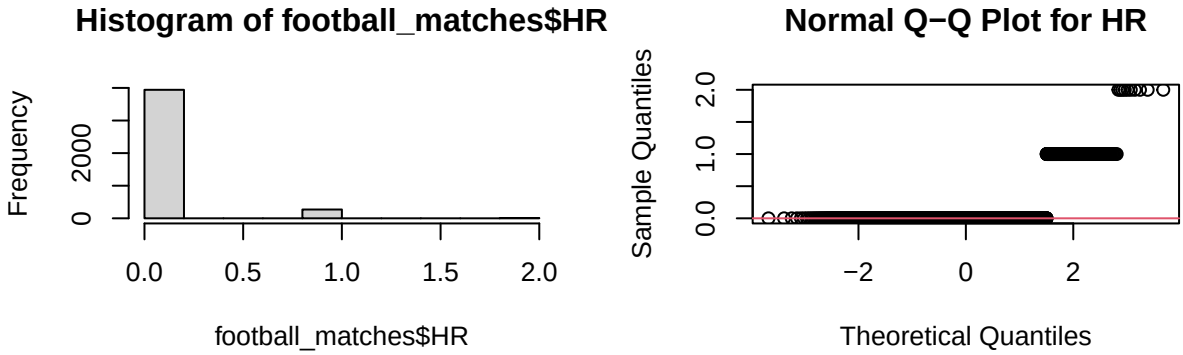


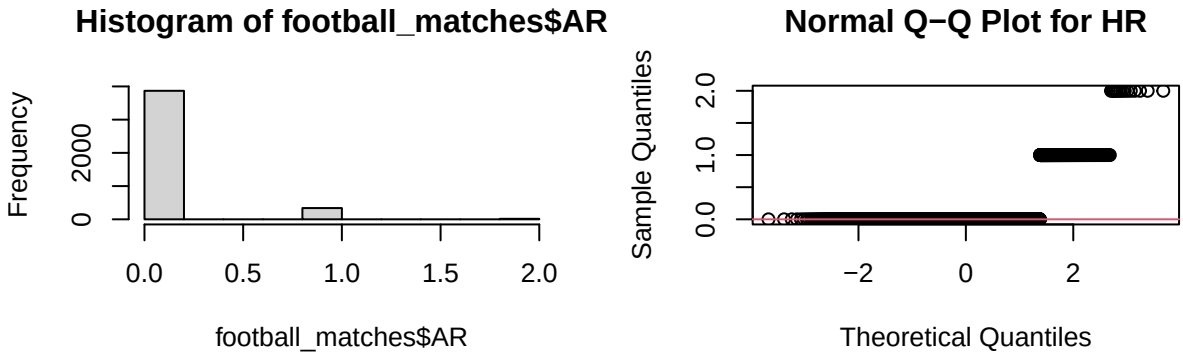







```r
hist(football_matches$HR)
qqnorm(football_matches$HR, main="Normal Q-Q Plot for HR")
qqline(football_matches$HR,col=2)
hist(football_matches$AR)
qqnorm(football_matches$AR, main="Normal Q-Q Plot for HR")
qqline(football_matches$AR,col=2)
```





En el cas de la comprovació per test de Shapiro-Wilk, s'observa en totes les quatre variables un p-valor molt més petit que el nivell de significació de 0.05, per la qual cosa es podria rebutjar la hipòtesi de normalitat i concloure que les dades no compten amb una distribució normal.

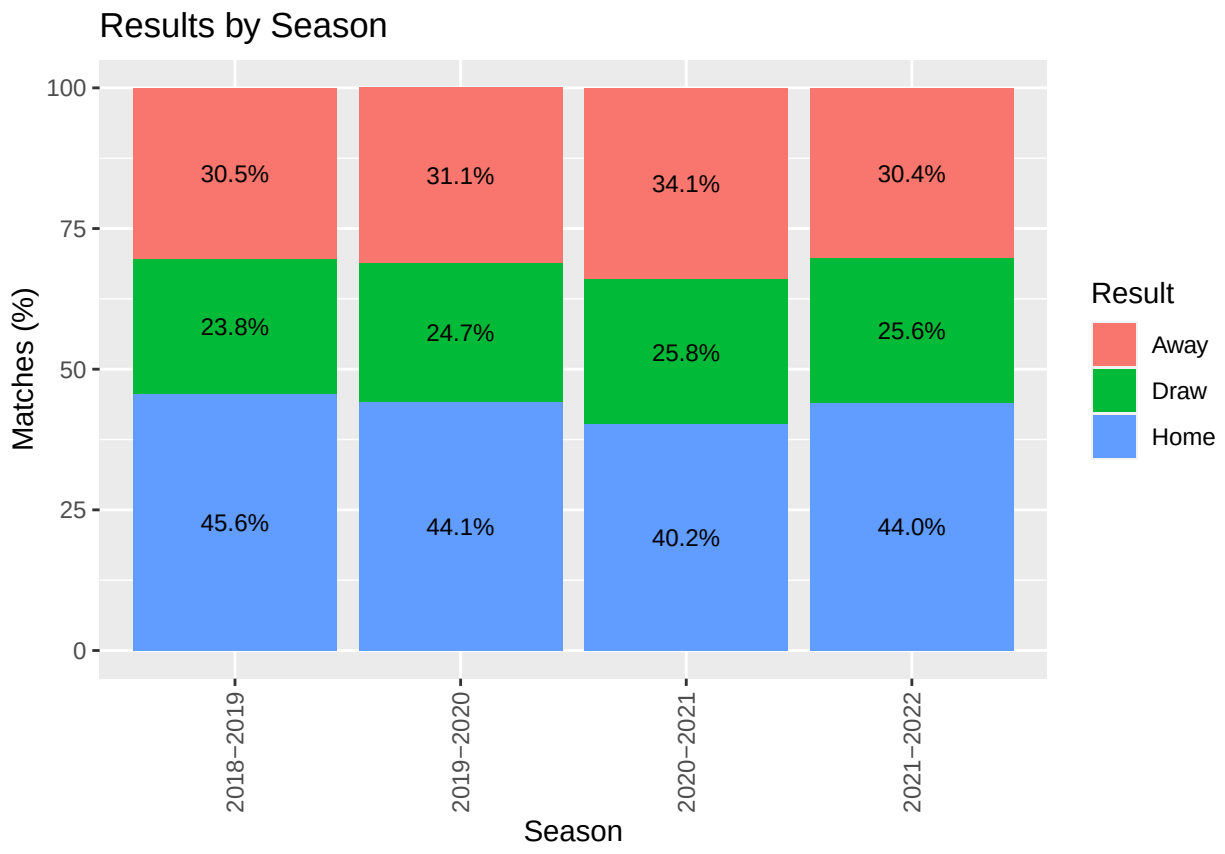
No obstant, si es realitza un anàlisi per gràfics Q-Q (gràfics de quantils teòrics), sí que s'observa una possible normalitat en les targetes grogues, tant locals com visitants. Així doncs, s'hauria de fer un anàlisi estadístic més a fons en aquestes dues variables per determinar si la hipòtesi de normalitat es compleix.

Aplicació de proves estadístiques i representació dels resultats

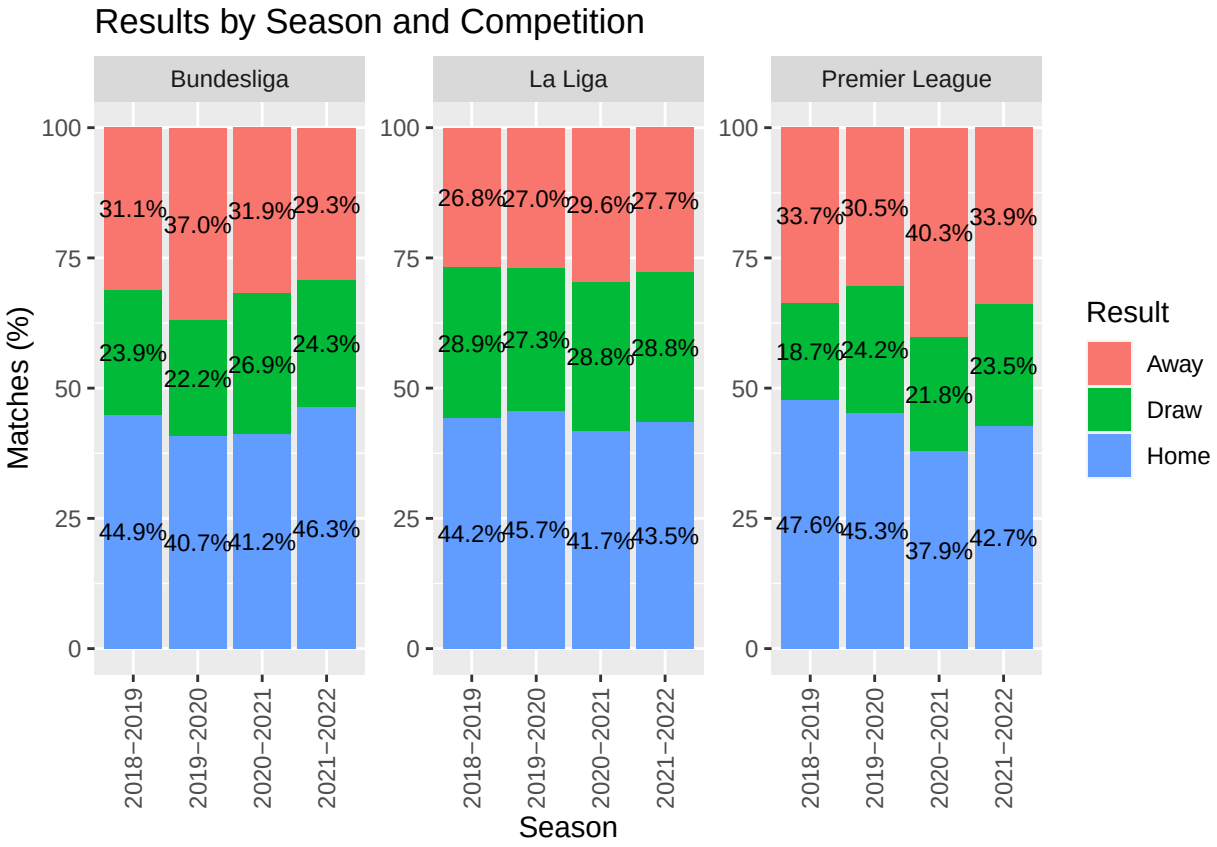
Restriccions COVID-19 als estadis

En aquest apartat es representaran els resultats de l'anàlisi dels resultats dels partits (victòria local/visitant o empat) segons múltiples variables. La representació d'aquestes dades es realitzarà mitjançant gràfiques de barres apilades, amb el percentatge de partits sobre el total.

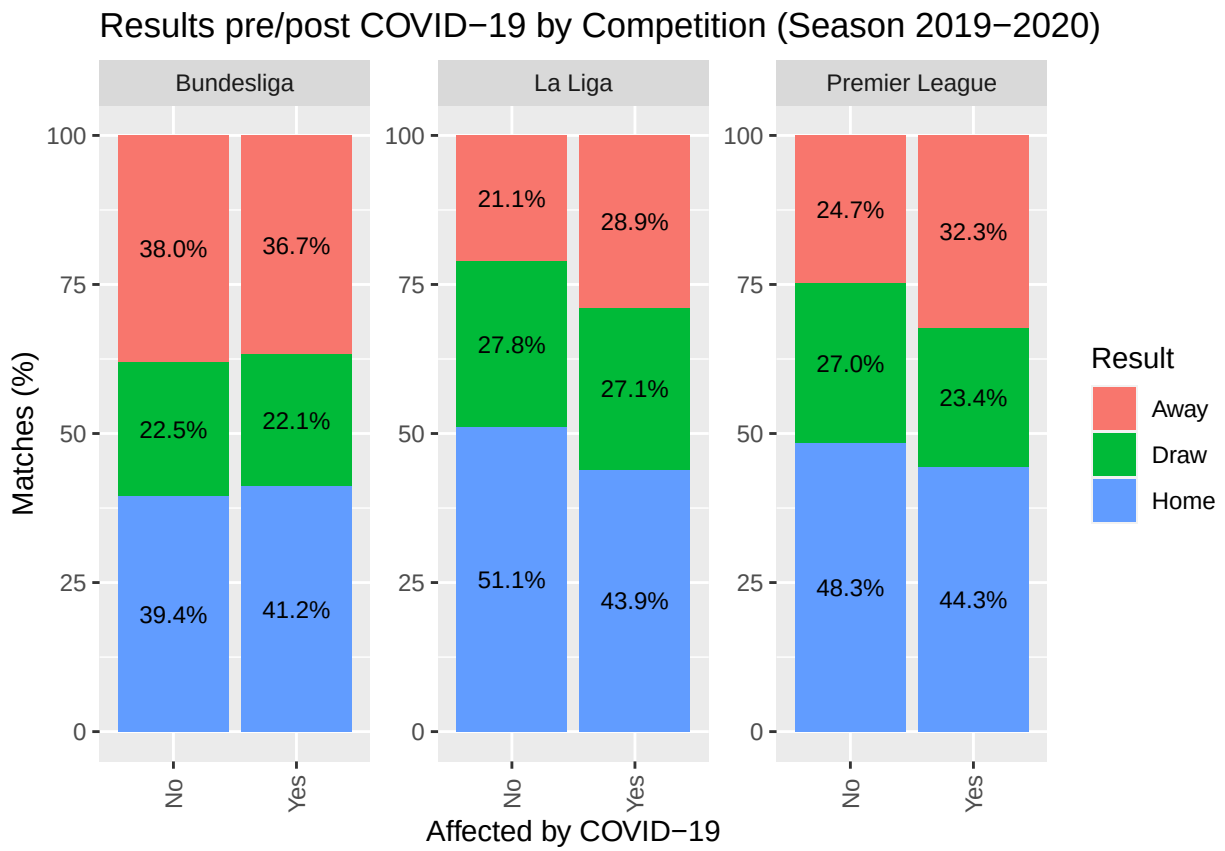
```
# Resultats per temporada
ggplot(result_by_season, aes(x = Season, y = percent, fill = Result, label = paste0(sprintf("%1.1f", pe
  geom_bar(stat = "identity", width = 0.8) +
  geom_col() +
  labs(y = "Matches (%)", fill = "Result", title = "Results by Season") +
  geom_text(size = 3, position = position_stack(vjust = 0.5)) +
  theme(axis.text.x = element_text(angle = 90,
                                   vjust = 0.5,
                                   hjust = 0.5))
```




```r
# Resultats per temporada i competició
ggplot(result_by_competition_season, aes(x = Season, y = percent, fill = Result, label = paste0(sprintf
  facet_wrap(~Competition, scales="free") +
  geom_bar(stat = "identity", width = 0.8) +
  geom_col() +
  labs(y = "Matches (%)", fill = "Result", title = "Results by Season and Competition") +
  geom_text(size = 3, position = position_stack(vjust = 0.5)) +
  theme(axis.text.x = element_text(angle = 90,
                                   vjust = 0.5,
                                   hjust = 0.5))
```




```r
# Resultats per Post/pre covid i competició (2019-2020)
ggplot(result_by_competition_covid, aes(x = CovidRestrictions, y = percent, fill = Result, label = paste
  facet_wrap(~Competition, scales="free") +
  geom_bar(stat = "identity", width = 0.8) +
  geom_col() +
  labs(x = "Affected by COVID-19", y = "Matches (%)", fill = "Result", title = "Results pre/post COVID-
  geom_text(size = 3, position = position_stack(vjust = 0.5)) +
  theme(axis.text.x = element_text(angle = 90,
                                   vjust = 0.5,
                                   hjust = 0.5))
```



A simple vista, a la gràfica per temporada (Results by Season) ja s'observa una variació a la temporada posterior a l'aparició de la COVID-19 respecte als resultats de les altres temporades analitzades, un augment de les victòries visitants a costa de les locals. En canvi, durant la temporada 2019-2020 no s'aprecia una gran diferència, probablement pel fet que s'havien disputat més de la meitat dels partits abans de l'aturada.

També s'aprecia un augment menys pronunciat dels empats, però degut al fet que tampoc és tan gran i que a l'any següent es manté pràcticament constant tot i haver tornat a la normalitat, no es pot determinar que sigui degut a les restriccions. Es podria fer un estudi per buscar una relació amb la introducció del videoarbitratge (VAR) a les principals lligues europees als últims anys, ja que aquest sí que podria ser el causant d'aquest augment.

A l'anàlisi per temporada i competició (Results by Season and Competition) s'observa un augment important de les victòries visitants tant a la Bundesliga com a la Premier, tot i que en temporades diferents, en el primer cas és durant l'aparició de la pandèmia i, en el segon, a la temporada següent. A la lliga espanyola s'aprecia un increment menys pronunciat durant la temporada 2020-2021.

Finalment, a l'anàlisi de la temporada 2019-2020 abans i després de l'aparició de la COVID-19, s'observa que el resultat anterior de la Bundesliga probablement no sigui causat per la COVID, ja que és bastant constant abans i després d'aquesta. En canvi, a la lliga espanyola i la Premier sí que s'observa aquest augment destacable de les victòries locals.

Com a conclusió, tot sembla apuntar que l'efecte dels aficionats de l'estadi en els resultats del seu equip és real i apreciable, tot i que cal destacar que, en una lliga com la Bundesliga que sempre té uns bons números d'assistència de públic que alhora genera un gran ambient a l'estadi, sorprén que no es vegui l'efecte de les restriccions d'aforament en els resultats.

Comportament dels àrbitres segons la seva experiència als terrenys de joc

```
# Models de regressió lineal

# Targetes vermelles

model1 = lm(avg_R ~ total_games, data=referee, na.action=na.exclude)

summary(model1)
```

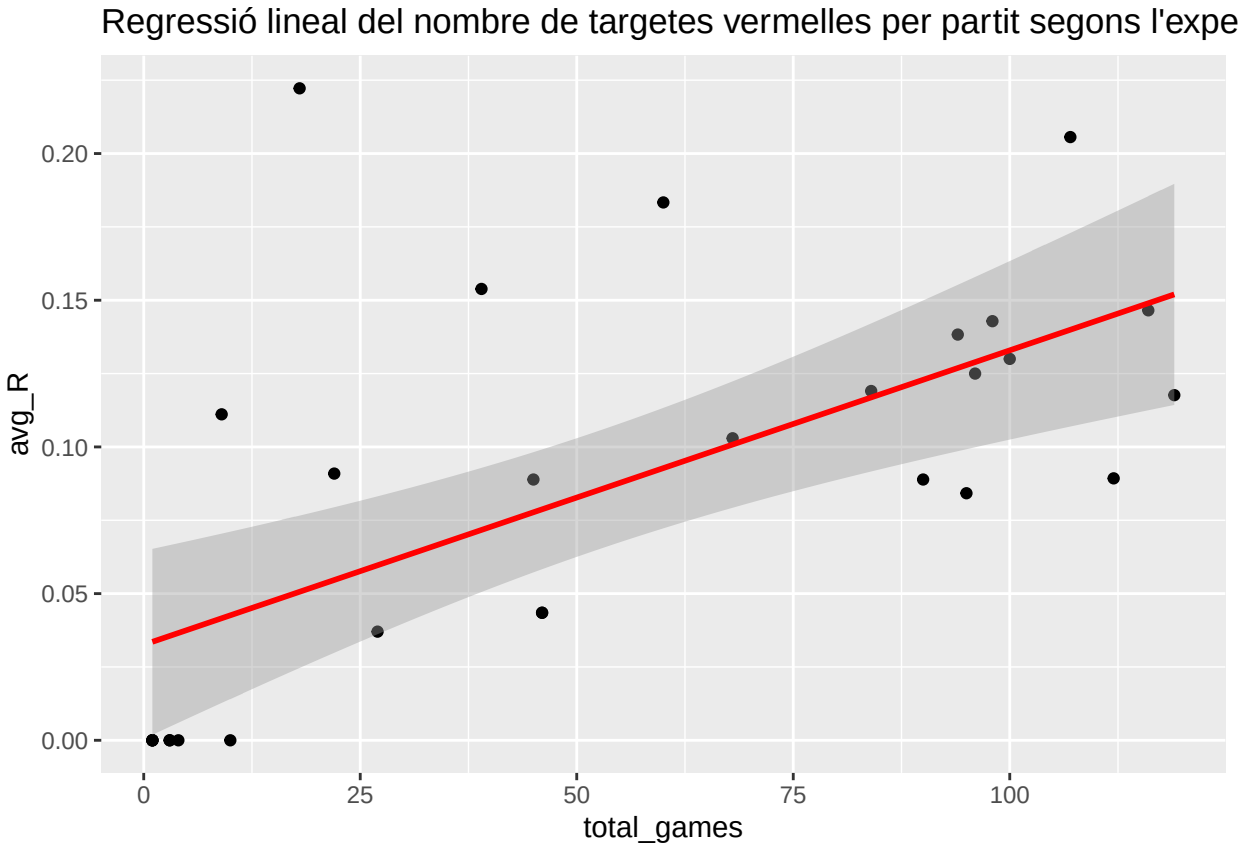
```
##
## Call:
## lm(formula = avg_R ~ total_games, data = referee, na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.05567 -0.03525 -0.02262  0.01140  0.17160
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.0325610  0.0156220   2.084 0.046722 *
## total_games 0.0010035  0.0002323   4.321 0.000188 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05299 on 27 degrees of freedom
## Multiple R-squared:  0.4088, Adjusted R-squared:  0.3869
## F-statistic: 18.67 on 1 and 27 DF,  p-value: 0.0001885
```

```
corr_pearson = sqrt(0.4088)
corr_pearson
```

```
## [1] 0.6393747
```

```
ggplot(referee, aes(total_games, avg_R)) + geom_point() + geom_smooth(method = "lm", colour = "Red") +
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# Targetes grogues

model2 = lm(avg_Y ~ total_games, data=referee, na.action=na.exclude)

summary(model2)
```

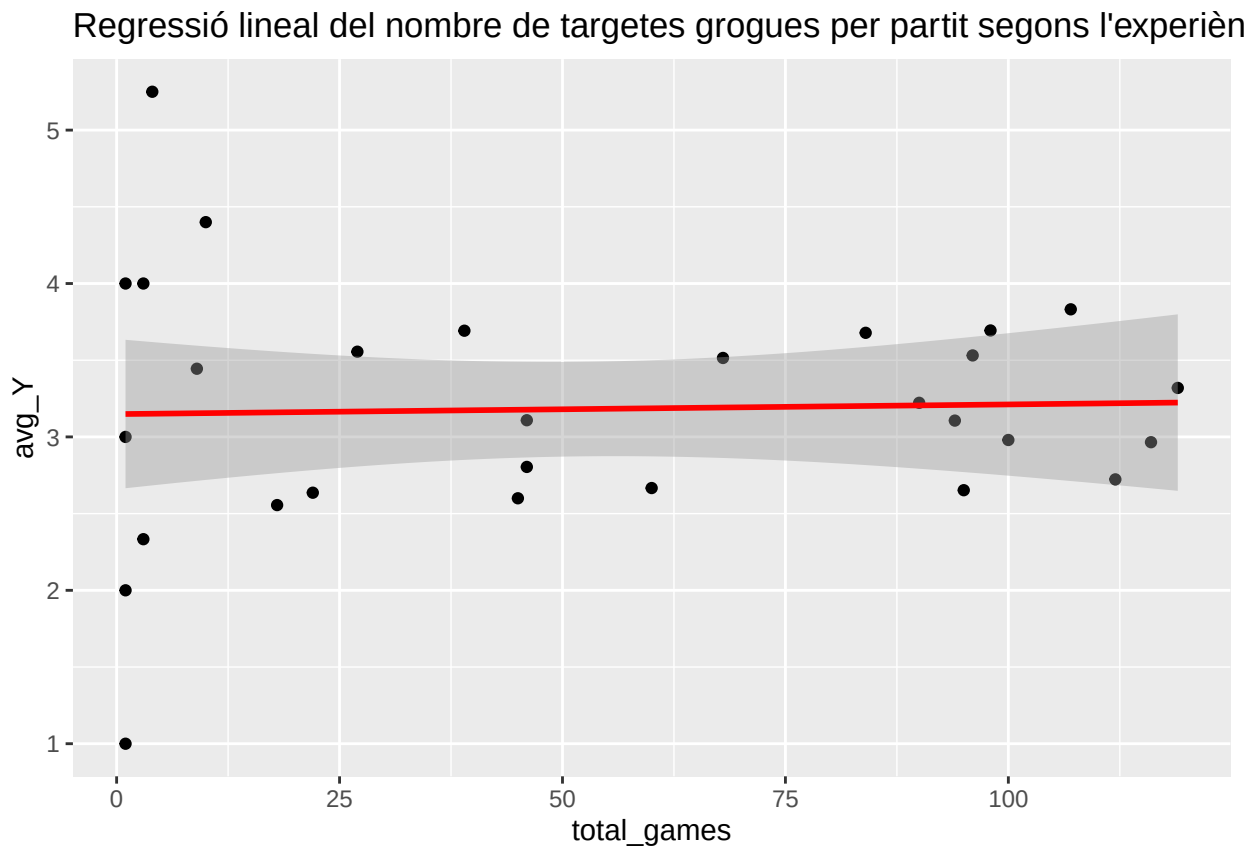
```
##
## Call:
## lm(formula = avg_Y ~ total_games, data = referee, na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.14927 -0.51984 -0.06898  0.47691  2.09884
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.1486392  0.2384768  13.203  2.7e-13 ***
## total_games 0.0006312  0.0035454   0.178     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8089 on 27 degrees of freedom
## Multiple R-squared:  0.001172,   Adjusted R-squared:  -0.03582
## F-statistic: 0.03169 on 1 and 27 DF,  p-value: 0.86
```

```r
corr_pearson2 = sqrt(0.001172)
corr_pearson2
```

```
## [1] 0.03423449
```

```r
ggplot(referee, aes(total_games, avg_Y)) + geom_point() + geom_smooth(method = "lm", colour = "Red") +
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```r
# Faltes

model3 = lm(avg_F ~ total_games, data=referee, na.action=na.exclude)

summary(model3)
```

```
##
## Call:
## lm(formula = avg_F ~ total_games, data = referee, na.action = na.exclude)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.439 -1.072 -0.082  1.254  5.561
##
```

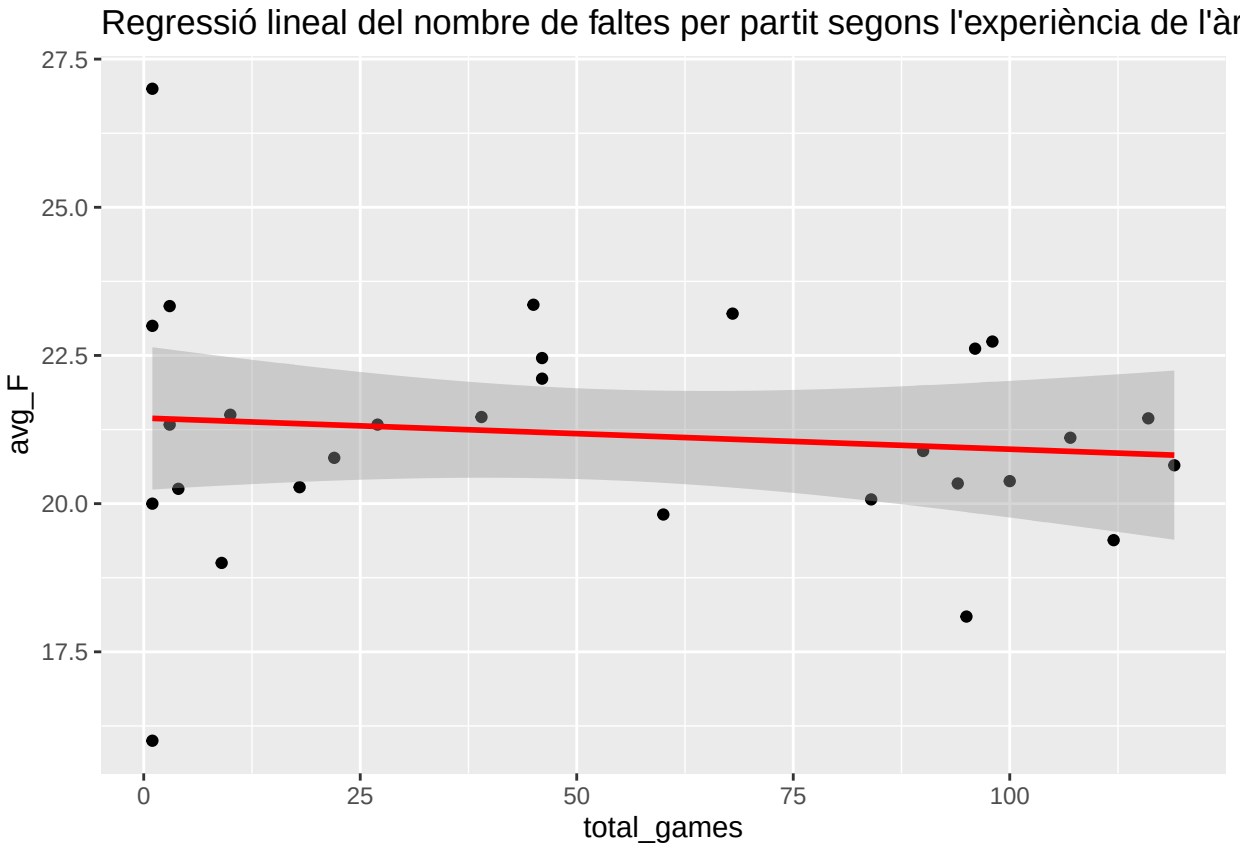
```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.444025   0.591923  36.228   <2e-16 ***
## total_games -0.005257   0.008800  -0.597    0.555
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.008 on 27 degrees of freedom
## Multiple R-squared:  0.01304,    Adjusted R-squared:  -0.02351
## F-statistic: 0.3568 on 1 and 27 DF,  p-value: 0.5553
```

```
corr_pearson3 = sqrt(0.01304)
corr_pearson3
```

```
## [1] 0.1141928
```

```
ggplot(referee, aes(total_games, avg_F)) + geom_point() + geom_smooth(method = "lm", colour = "Red") +
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



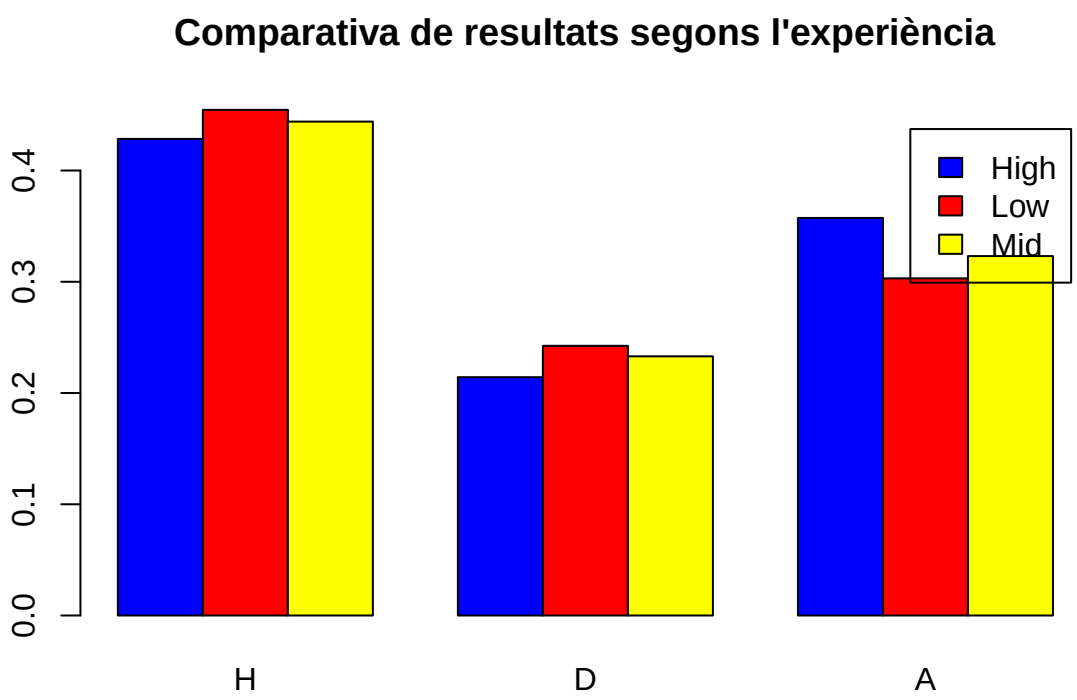
Podem observar com els coeficients de la correlació de pearson són bastant baixos, cosa que indica que no hi ha molta dependència a les correlacions lineals. Sí que amb la comparativa de nombre de partits i ratgetes vermelles el nombre és una mica més elevat i podem treure la conclusió que els àrbitres amb menys experiència treuen menys targetes vermelles. Això pot ser degut a la presió del primer partit i al no voler quedar assenyalat fent un partit sense prendre riscos.

```r
H <- c(experience_stats$H)
D <- c(experience_stats$D)
A <- c(experience_stats$A)

games <- cbind(H, D, A)

rownames(games) <- c("High", "Low", "Mid")

barplot(games, beside = T, col = c("blue","red","yellow"), main = "Comparativa de resultats segons l'exp
        legend.text = rownames(games))
```



A l'hora de mirar les diferències de resultats segons si l'experiència de l'àrbitre es veu com hi ha una diferència entre les victòries locals i visitants. S'observa com als partits arbitrats per àrbitres novells hi ha un percentatge de victòries i d'empats superior als arbitrats per àrbitres experimentats, i tot el contrari amb les victòries visitants. Això pot ser degut a la presió efectuada del públic sobre l'àrbitre i la manera que té l'àrbitre de gestionar-la. També considerem que les temporades estudiades no acaben de reflectir al 100% la presió rebuda pels àrbitres degut a que la meitat d'aquests partits es van jugar a porta tancada o sense públic, tot i així ja es veuen reflectits uns indícis del comportament dels àrbitres davant la presió efectuada pel públic local.

Bibliografia

Dades d'aforament extretes de FBREF:

https://fbref.com/en/comps/12/2019-2020/schedule/2019-2020-La-Liga-Scores-and-Fixtures

https://fbref.com/en/comps/12/2020-2021/schedule/2020-2021-La-Liga-Scores-and-Fixtures
https://fbref.com/en/comps/20/2019-2020/schedule/2019-2020-Bundesliga-Scores-and-Fixtures
https://fbref.com/en/comps/20/2020-2021/schedule/2020-2021-Bundesliga-Scores-and-Fixtures
https://fbref.com/en/comps/9/2019-2020/schedule/2019-2020-Premier-League-Scores-and-Fixtures
https://fbref.com/en/comps/9/2020-2021/schedule/2020-2021-Premier-League-Scores-and-Fixtures

Contribucions

Investigació prèvia: Adrià Setó, Miquel Arisa
Redacció de les respostes: Adrià Setó, Miquel Arisa
Desenvolupament del codi: Adrià Setó, Miquel Arisa
Participació al vídeo: Adrià Setó, Miquel Arisa

Vídeo

https://drive.google.com/file/d/1i8nl0oREUSPbGcPLKCCXEC6IaKHS5DEU/view?usp=sharing