# Ejercicio premier league

Joaquin Bravo

2024-01-04

# R Markdown

Como me gusta el fútbol y conozco la liga, voy a analizar el dataset de partidos en la premier league (https://github.com/rfordatascience/tidytuesday/blob/master/data/2023/2023-04-04/readme.md#soccer21-22csv (https://github.com/rfordatascience/tidytuesday/blob/master/data/2023/2023-04-04/readme.md#soccer21-22csv)). Para ello voy a intentar responder dos preguntas:

1. ¿Cuál es la tendencia de victorias locales vs. visitantes a lo largo de la temporada?
2. ¿Cómo se correlacionan las estadísticas del partido con ganar uno? Para ver si depende de los tiros, córners o faltas.

# 1. Tendencia de victorias locales vs. visitantes

```
library(DBI)
library(dbplyr)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.4.4     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::ident()  masks dbplyr::ident()
## ✖ dplyr::lag()    masks stats::lag()
## ✖ dplyr::sql()    masks dbplyr::sql()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to be
come errors
```

```
library(lubridate)
library(ggplot2)
library(RSQLite)

# Conectar a la base de datos
con <- dbConnect(RSQLite::SQLite(), ":memory:")

# Leer los datos en R y escribirlos en la base de datos
soccer_data <- read.csv("soccer21-22.csv")
dbWriteTable(con, "soccer_data", soccer_data)

# Crear tablas temporales y realizar consultas

dbExecute(con, "CREATE TABLE temp_results AS
          SELECT
          Date,
          CASE WHEN FTR = 'H' THEN 1 ELSE 0 END AS HomeWin,
          CASE WHEN FTR = 'A' THEN 1 ELSE 0 END AS AwayWin
          FROM
          soccer_data")
```

```
## [1] 0
```

```
home_away_wins <- dbGetQuery(con, "SELECT
                             SUM(HomeWin) AS TotalHomeWins,
                             SUM(AwayWin) AS TotalAwayWins
                             FROM
                             temp_results")

dbExecute(con, "
  CREATE TABLE temp_stats AS
  SELECT
    Date,
    HomeTeam,
    AwayTeam,
    FTR,
    HS,
    \"AS\" AS AwayShots,  -- Cambiando el nombre de la columna para evitar conflictos
    HF,
    AF,
    HC,
    AC
  FROM
    soccer_data
")
```

```
## [1] 0
```

```
game_stats <- dbGetQuery(con, "SELECT
                         *,
                         CASE WHEN FTR = 'H' THEN 1 ELSE 0 END AS HomeWin,
                         CASE WHEN FTR = 'A' THEN 1 ELSE 0 END AS AwayWin
                         FROM
                         temp_stats")

# Cerrar la conexión
dbDisconnect(con)

# Cargar datos
soccer_data <- read.csv("soccer21-22.csv")

# Convertir fechas
soccer_data$Date <- dmy(soccer_data$Date)

# Calcular victorias en casa y fuera
soccer_data <- soccer_data %>%
  mutate(HomeWin = if_else(FTR == 'H', 1, 0),
         AwayWin = if_else(FTR == 'A', 1, 0))

# Agregar victorias por mes
monthly_wins <- soccer_data %>%
  group_by(Month = floor_date(Date, "month")) %>%
  summarize(TotalHomeWins = sum(HomeWin),
            TotalAwayWins = sum(AwayWin))
```

# Visualización

```
# Visualización
ggplot(monthly_wins, aes(x = Month)) +
  geom_line(aes(y = TotalHomeWins, color = "Victorias locales")) +
  geom_line(aes(y = TotalAwayWins, color = "Victorias visitantes")) +
  labs(title = "Tendencia de victorias en casa vs. victorias fuera",
       x = "Mes",
       y = "Total de victorias") +
  scale_color_manual(values = c("Victorias locales" = "blue", "Victorias visitantes" = "re
d")) +
  theme_minimal()
```


Notamos que hay una tendencia más alcista de las victorias locales, mientras que las visitantes se mantienen en un rango.

Ahora exploro cómo se correlacionan las estadísticas de juego con ganar un partido

# 2. ¿Cómo se correlacionan las variables?

Armaremos una matriz de correlación para explorar las relaciones de todas las variables.

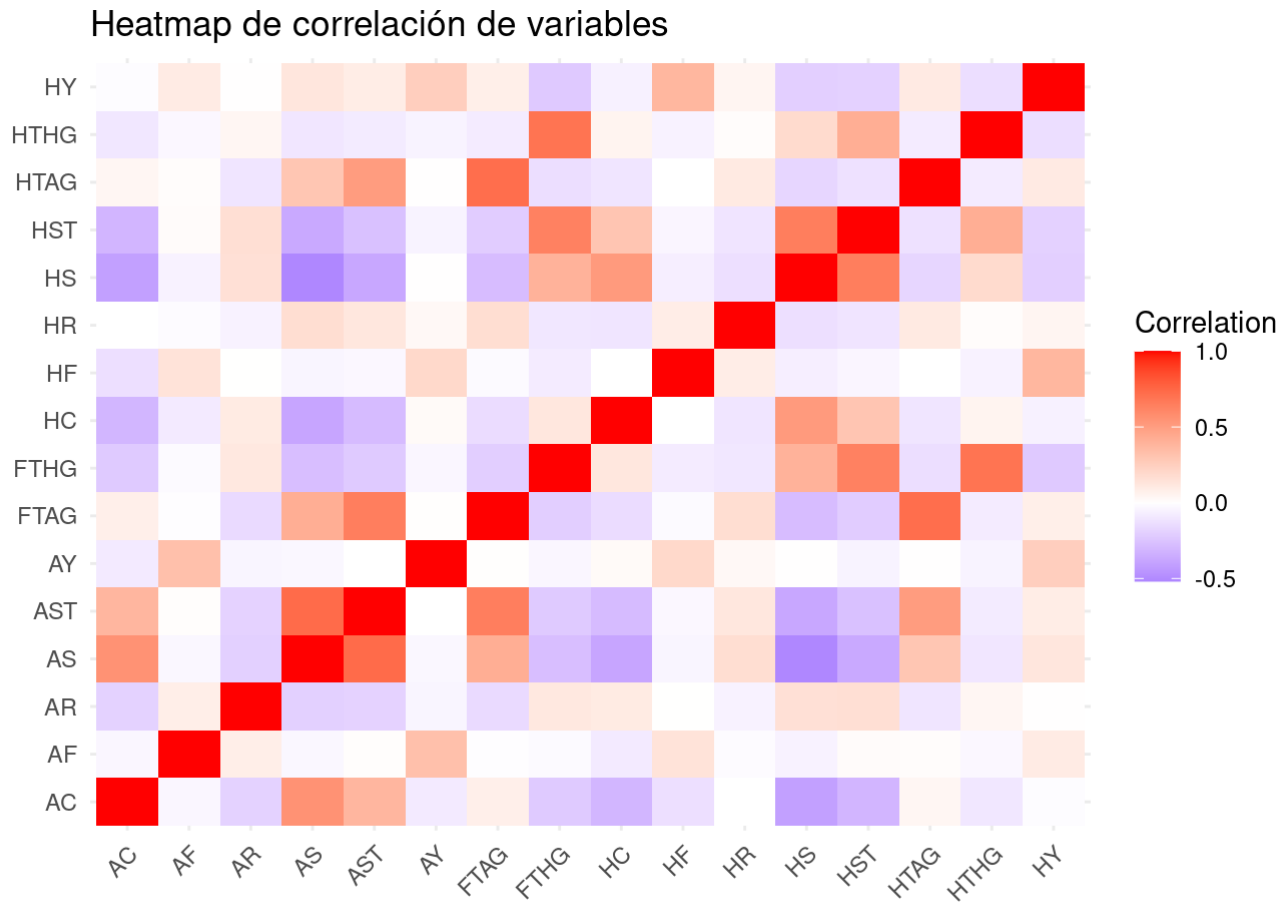
```
# Seleccionar solo las columnas numéricas relevantes
numeric_data <- soccer_data %>%
  select(FTHG, FTAG, HTHG, HTAG, HS, AS, HST, AST, HF, AF, HC, AC, HY, AY, HR, AR) %>%
  na.omit()  # Eliminar filas con valores faltantes

# Calcular la matriz de correlación
cor_matrix <- cor(numeric_data)
```

# Visualización

```
# Convertir la matriz de correlación a un formato largo para ggplot
cor_data <- cor_matrix %>%
  as.data.frame() %>%
  rownames_to_column("Variable1") %>%
  pivot_longer(cols = -Variable1, names_to = "Variable2", values_to = "Correlation")

# Gráfico de calor de las correlaciones
ggplot(cor_data, aes(x = Variable1, y = Variable2, fill = Correlation)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Heatmap de correlación de variables", x = "", y = "")
```




Y con la matriz de correlación podemos ver todas, pero podemos quedarnos sólo con las correlaciones más fuertes (mayores a 0.7)

```
# Encontrar pares con alta correlación
high_correlation_pairs <- cor_data %>%
  filter(abs(Correlation) > 0.5, Variable1 != Variable2)

high_correlation_pairs
```

```
## # A tibble: 20 × 3
##    Variable1 Variable2 Correlation
##    <chr>     <chr>           <dbl>
##  1 FTHG      HTHG            0.698
##  2 FTHG      HST             0.636
##  3 FTAG      HTAG            0.722
##  4 FTAG      AST             0.650
##  5 HTHG      FTHG            0.698
##  6 HTAG      FTAG            0.722
##  7 HTAG      AST             0.510
##  8 HS        AS             -0.518
##  9 HS        HST             0.654
## 10 HS        HC              0.519
## 11 AS        HS             -0.518
## 12 AS        AST             0.735
## 13 AS        AC              0.557
## 14 HST       FTHG            0.636
## 15 HST       HS              0.654
## 16 AST       FTAG            0.650
## 17 AST       HTAG            0.510
## 18 AST       AS              0.735
## 19 HC        HS              0.519
## 20 AC        AS              0.557
```

## Relación entre los goles en los 90 minutos y primer tiempo (FTHG/FTAG y HTHG/HTAG):

FTHG y HTHG tienen una correlación de 0.698, lo que indica una fuerte relación positiva. Esto sugiere que los equipos que anotan más goles en jugando de local en el primer tiempo tienden también a anotar más en el segundo tiempo.

FTAG y HTAG tienen una correlación de 0.722, similar a la anterior, pero para los equipos visitantes. Los equipos visitantes que hacen más goles en el primer tiempo tienden a hacer lo mismo en segundo.

## Relación entre los goles y los remates al arco (FTHG/FTAG y HST/AST):

FTHG y HST tienen una correlación de 0.636, y FTAG y AST una de 0.650. Esto indica que hay una relación positiva significativa entre los remates al arco de un equipo (tanto local como visitante) y la cantidad de goles que hacen. Cuantos más tiros al arco, más probable es que sea gol.

## Relación negativa entre los remates de los equipos locales y visitantes (HS y AS):

HS y AS tienen una correlación de -0.518. Esto podría sugerir que cuando un equipo (local en este caso) tiene muchos tiros, el equipo visitante tiende a tener menos, lo que podría reflejar un juego más defensivo o una mayor poseción del equipo local.

## Relación entre remates y remates al arco (HS/HST y AS/AST):

HS y HST tienen una correlación de 0.654, y AS y AST de 0.735. Esto indica que hay una fuerte relación positiva entre los tiros que realiza un equipo y los tiros que realmente van al arco, tanto para equipos locales como visitantes.

## Relación entre remates y córners (HS/HC y AS/AC):

HS y HC (0.519), así como AS y AC (0.557), muestran una correlación moderadamente positiva. Esto sugiere que los equipos que rematan más también tienden a tener más córners.