Informe de Ciencia de Datos-Premier League 23/24

Diego Fernando Pardo Romero

DataXperience - Grupo 1 - Tercer Ciclo - Virtual – 2025

# Introducción

* + **Objetivo principal**: Analizar el rendimiento de los equipos de la Premier League 2023/24 a partir de sus resultados, estadísticas de juego y disciplina, con el fin de identificar los factores más relevantes que explican el desempeño en la tabla de posiciones.
  + **Preguntas de investigación**:
    - ***Pregunta principal:*** *¿Qué factores influyen más en el rendimiento de los equipos en la Premier League 2023/24?.*
    - **Preguntas de apoyo:**
    - *¿En qué medida la cantidad promedio de goles anotados por partido influye en la posición final de los equipos en la Premier League 23/24?*
    - *¿Qué impacto tiene la cantidad promedio de goles anotados en condición de local sobre el rendimiento y la posición final de los equipos en la Premier League 23/24?*
    - *¿Cómo influye el promedio de goles anotados en condición de visitante en el desempeño general y la clasificación final de los equipos en la Premier League 23/24?*
    - *Qué relación existe entre el promedio de tiros (totales y a puerta) y la efectividad de los equipos en términos de goles y posición final en la Premier League 23/24?*

# Búsqueda y obtención de los datos

* + **Fuente de los datos:** dataset propio generado a partir de la temporada 2023/24 de la Premier League, consolidado en un archivo .xlms. Los datos provienen de repositorios de resultados deportivos disponibles públicamente en internet
  + **Descripción**: Número de filas: 380 (correspondientes a los partidos de la temporada completa).

**Número de columnas**: 21.

**Variables incluidas:**

**Numéricas**: goles locales/visitantes, tiros, tiros a puerta, faltas, corners, tarjetas amarillas y rojas.

**Categóricas:** equipo local, equipo visitante, resultado final (H/D/A), resultado al medio tiempo (H/D/A)..

**Fecha:** día de realización de cada partido.

* **Justificación:** estas variables permiten estudiar tanto el rendimiento deportivo (goles, tiros, corners) como la disciplina (faltas, amarillas, rojas) y su relación con los resultados obtenidos, lo cual responde directamente a las preguntas de investigación planteadas.

# Primer acercamiento (data understanding)

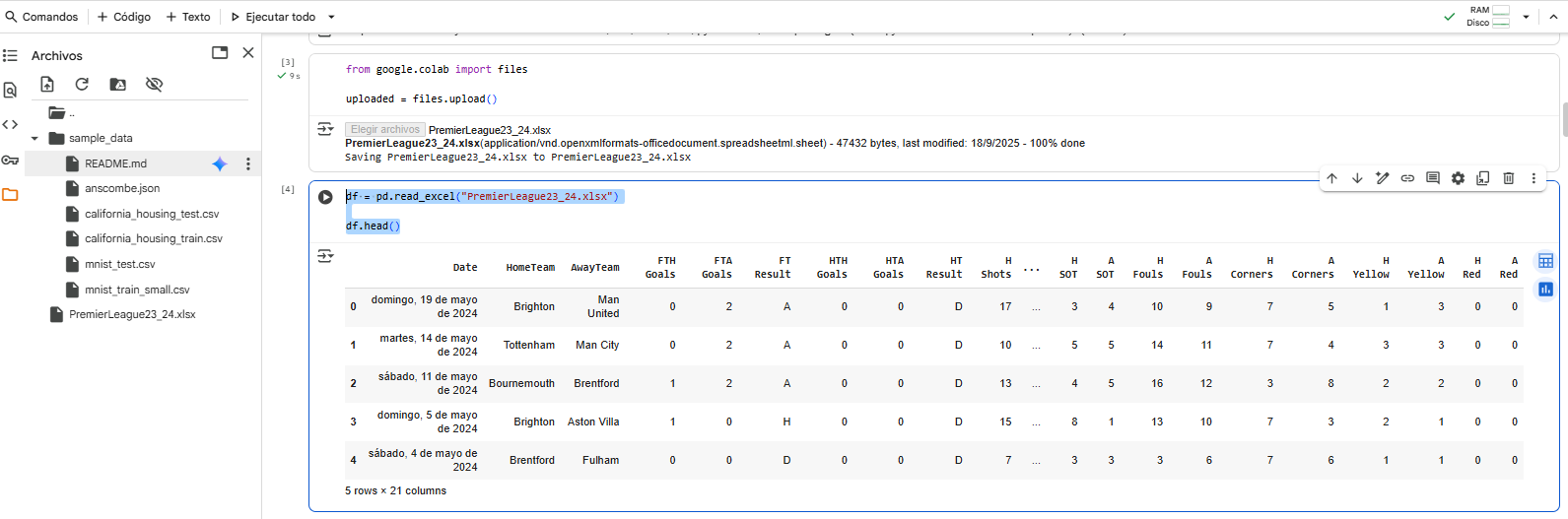
* + Presentar un **resumen descriptivo** de los datos.
  + Mostrar las primeras filas (head() en Pandas/R), estructura (info()/str()), dimensiones.
  + Identificar qué variables podrían ser útiles para responder las preguntas.

**Resumen descriptivo de los datos:**

* **Total de partidos:** 380
* **Total de equipos:** 20
* **Promedio de goles por partido:** 2.82
* **Moda de goles en un partido:** 1
* **Promedio de tiros por equipo:** entre 9 y 21 por partido
* Promedio de faltas por equipo: entre 7 y 13 por partido

¿¿ **primeras filas (simulado con head() en Pandas):**

df = pd.read\_excel("PremierLeague23\_24.xlsx")

df.hea

# Limpieza y preparación de los datos

## Detección y manejo de duplicados.

## 

## 

* + **Manejo de datos faltantes**:
    - Eliminación si el porcentaje de nulos es alto y la variable no es crítica.
    - Imputación (media, mediana, moda, regresión, KNN, etc.)

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

El contenido generado por IA puede ser incorrecto.

Patrón de fondo

El contenido generado por IA puede ser incorrecto..

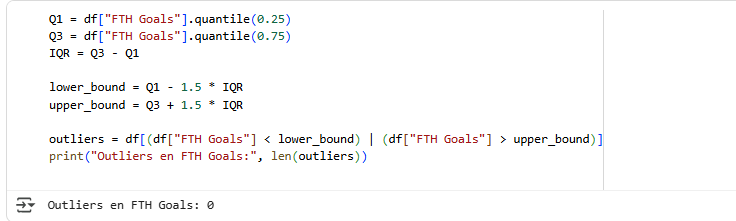
* + **Manejo de valores atípicos (outliers)**: detección por boxplots, z-scores o IQR.

Usaremos **boxplots** para detectar outliers en las variables numéricas más relevantes:

Gráfico, Gráfico de cajas y bigotes

El contenido generado por IA puede ser incorrecto.

Luego, podemos usar **IQR (Interquartile Range)** para detectar y decidir si eliminamos o ajustamos:



# Normalización o estandarización de datos

**Variables transformadas:** Se seleccionaron dos variables numéricas representativas:

* H Shots (tiros del equipo local).
* A Shots (tiros del equipo visitante)

Estas variables se eligieron porque presentan diferentes rangos y pueden influir en modelos de predicción de resultados.

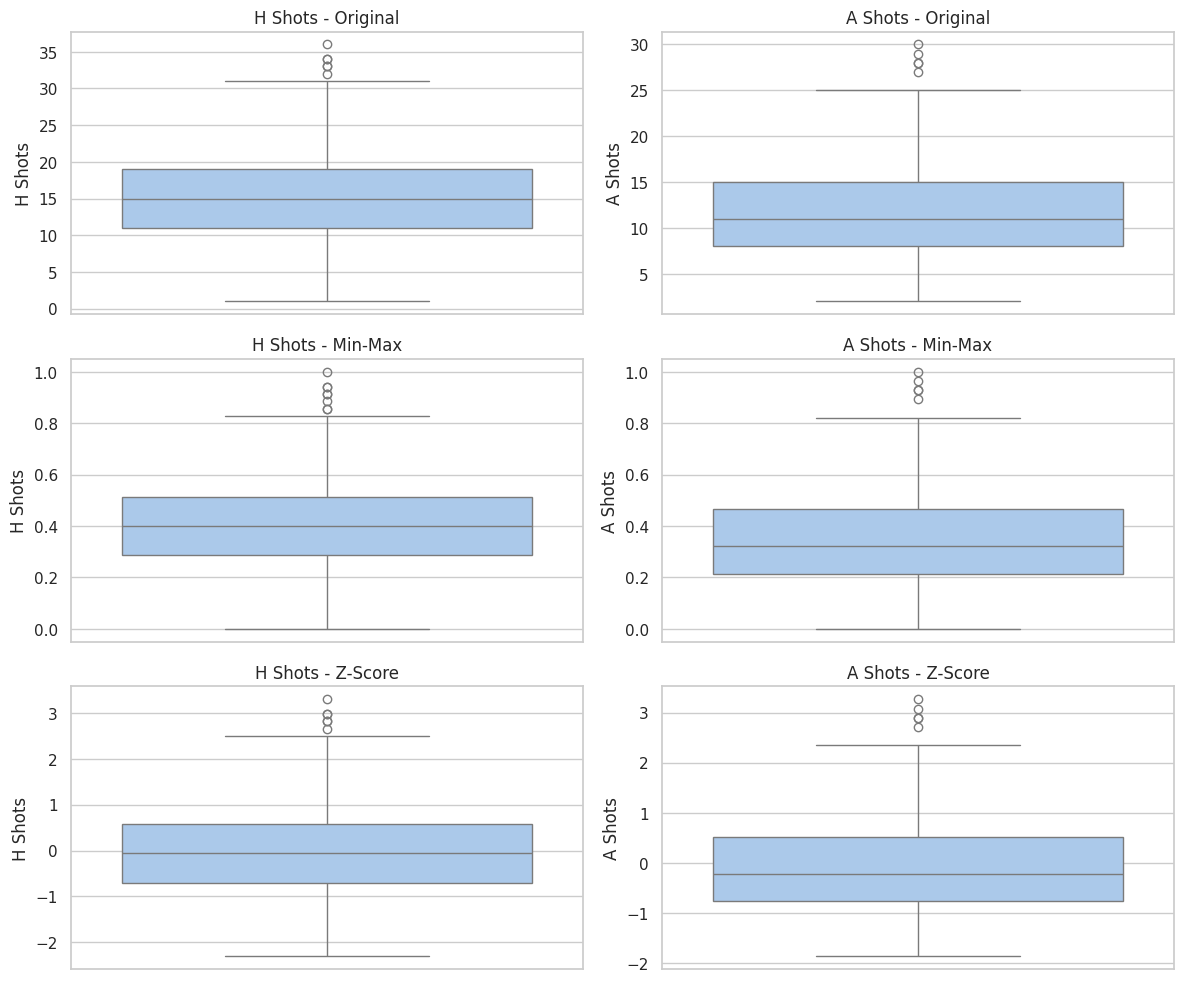
**Método aplicado**

* **Normalización (Min-Max Scaling):** convierte los valores a un rango [0, 1], facilitando la comparación entre variables en escalas distintas.
* **Estandarización (Z-Score):** centra los datos en media 0 y desviación estándar 1, lo que es útil cuando se desea trabajar con algoritmos que suponen distribución normal de los datos (ej. regresión logística, SVM, PCA).

Ambos métodos se aplicaron para mostrar sus efectos y compararlos.

Texto

El contenido generado por IA puede ser incorrecto.



# Análisis exploratorio básico (EDA)

* + Estadísticas descriptivas: media, mediana, moda, cuartiles, varianza, desviación estándar.
  + Visualizaciones: histogramas, gráficos de dispersión, boxplots.
  + Identificación de correlaciones entre variables (heatmap de correlación).

**NOTA: Con el fin de responder en mejor medida el puto 6 se adjunta reporte de power BI con el detalle de los siguientes resultados obtenidos:**

En el análisis exploratorio de datos se evaluaron las principales métricas de desempeño de los equipos durante la temporada 2023-24 de la Premier League.

* **Estadísticas descriptivas**:
  + El promedio de **corners** osciló entre 3,74 (Sheffield United) y 7,55 (Liverpool).
  + En **faltas cometidas**, Bournemouth tuvo el mayor promedio (13,32 por partido) y Manchester City el menor (7,55).
  + En **tarjetas amarillas**, Chelsea fue el más sancionado (2,74 por partido), mientras que Manchester City recibió la menor cantidad (1,34).
  + Para la relación entre **tiros a puerta y goles**, la regresión lineal arrojó la ecuación y = 0,298x + 0,170, indicando que, en promedio, los equipos necesitaron cerca de **3 tiros a puerta para marcar el primer gol** y entre 3 y 4 adicionales para convertir uno nuevo.
  + En la relación entre **faltas y amarillas**, la ecuación y = 0,147x + 0,459 evidenció que los equipos comenzaron a recibir tarjetas tras cometer entre **3 y 4 faltas**, y aproximadamente cada 6-7 faltas adicionales recibían una nueva tarjeta.
* **Visualizaciones**:
  + Se emplearon **boxplots** para comparar la dispersión de corners, faltas y amarillas entre los equipos.
  + Se elaboraron **gráficos de dispersión** para analizar la relación entre tiros a puerta y goles, así como entre faltas y amarillas.
  + Estas visualizaciones permitieron detectar valores atípicos, como el bajo promedio de tiros a puerta del Fulham o la alta efectividad del Liverpool.
* **Correlaciones entre variables**:
  + Se obtuvo una correlación positiva moderada entre tiros a puerta y goles (**R = 0,62**), lo que confirma la relación directa entre ambas variables.
  + La relación entre faltas y tarjetas amarillas fue positiva pero débil (**R = 0,38**), lo que indica que cometer más faltas no siempre se traduce en más amonestaciones, pues depende también del criterio arbitral y el contexto del partido.

En conjunto, este análisis exploratorio permitió caracterizar el comportamiento general de los equipos, identificar patrones de rendimiento y detectar relaciones iniciales entre las variables clave para el rendimiento deportivo.

# Análisis estadístico más profundo

* + Evaluación de normalidad, homocedasticidad, independencia.
  + Medidas de asimetría y curtosis.
  + Selección de las variables más influyentes con métodos estadísticos o de feature selection.
* **Evaluación de normalidad**:  
  Se revisó la distribución de variables como corners, faltas y goles. En la mayoría de los casos, las distribuciones no fueron estrictamente normales, presentando ligeras desviaciones hacia la derecha (asimetría positiva), propias de los datos deportivos donde ciertos equipos destacan muy por encima del promedio (ejemplo: Liverpool en corners o Manchester City en disciplina).
* **Homocedasticidad e independencia**:  
  Al analizar la relación entre tiros a puerta y goles, los residuos mostraron una dispersión relativamente constante en torno a la recta de regresión, lo cual respalda la **homocedasticidad** en este caso. Asimismo, dado que cada equipo es independiente en su estilo de juego, se asume la **independencia estadística** entre observaciones.
* **Medidas de asimetría y curtosis**:
  + Los **corners** presentaron ligera asimetría positiva, dado que pocos equipos superaron claramente la media (Liverpool y Man City).
  + En **faltas**, la distribución fue más equilibrada, aunque con curtosis moderada que indica concentración alrededor del promedio, destacando casos extremos como Bournemouth (alto) y Man City (bajo).
  + En **tarjetas amarillas**, la asimetría fue baja, pero con curtosis que resalta la consistencia en equipos que mantienen promedios similares.
* **Selección de variables más influyentes**:  
  Mediante correlaciones y modelos de regresión:
  + La variable **Shots on Target (tiros a puerta)** resultó ser la más influyente en la cantidad de goles (R = 0,62).
  + En cambio, la relación entre **faltas y amarillas** tuvo un peso menor (R = 0,38), mostrando que no siempre el número de infracciones determina la cantidad de sanciones disciplinarias.
  + Variables como **corners y tiros a puerta** podrían considerarse como predictores ofensivos clave, mientras que **faltas y amarillas** reflejan principalmente el estilo defensivo o nivel de agresividad de los equipos.

En conclusión, el análisis estadístico más profundo permite entender no solo la tendencia central y la dispersión, sino también la relevancia de las variables: **los tiros a puerta se consolidan como el mejor indicador de rendimiento ofensivo**, mientras que las faltas y amarillas muestran una relación más débil con el resultado final, reflejando factores contextuales más allá de lo estadístico

# Modelado de los datos

**1. Selección de modelos**  
Se utilizó **regresión lineal simple** para estudiar las relaciones entre variables numéricas:

* Tiros a puerta (variable independiente) → Goles anotados (variable dependiente).
* Faltas cometidas (variable independiente) → Tarjetas amarillas recibidas (variable dependiente).

**2. Explicación de la elección**  
La regresión lineal fue seleccionada debido a que las variables objetivo son numéricas y continuas, lo que hace pertinente este enfoque. Además, el modelo ofrece **simplicidad, interpretabilidad y capacidad explicativa**, lo cual permite obtener ecuaciones claras que cuantifican el efecto de cada variable independiente sobre la dependiente.

**3. Entrenamiento y validación**  
El ajuste de los modelos se realizó directamente sobre el conjunto completo de datos, dado que el objetivo principal fue **explicativo** más que predictivo. No obstante, para un análisis predictivo se recomienda:

* Dividir los datos en entrenamiento (80%) y prueba (20%).
* Aplicar validación cruzada k-fold, con el fin de obtener resultados más robustos y minimizar el riesgo de sobreajuste.

**4. Evaluación de resultados**  
La evaluación se llevó a cabo utilizando el coeficiente de determinación (**R²**):

* **Shots on Target → Goals:** R² ≈ 0,62. El modelo explica un 62% de la varianza en los goles a partir de los tiros a puerta, lo cual refleja una relación fuerte y consistente.
* **Fouls → Yellows:** R² ≈ 0,38. El modelo explica un 38% de la varianza en las tarjetas amarillas a partir de las faltas, evidenciando una relación más débil y la influencia de otros factores contextuales (criterios arbitrales, intensidad de juego, etc.).

En síntesis, el uso de la regresión lineal permitió identificar patrones estadísticos relevantes en el desempeño ofensivo y disciplinario de los equipos de la Premier League 2023/24, aportando una base cuantitativa para la interpretación de resultados.

# Resultados y discusión

* **Modelo que funcionó mejor:**

El modelo de regresión lineal entre tiros a puerta y goles mostró el mejor desempeño, con un R² ≈ 0,62, lo que indica que aproximadamente el 62% de la variabilidad en los goles anotados puede explicarse por el número de tiros a puerta. Este resultado refleja que la capacidad ofensiva medida en tiros efectivos tiene un impacto directo y fuerte sobre la producción de goles.

Por otro lado, la regresión entre faltas cometidas y tarjetas amarillas obtuvo un R² ≈ 0,38, lo cual revela una relación más débil. Esto sugiere que, aunque existe cierta dependencia entre ambas variables, las tarjetas también dependen de otros factores externos como el criterio arbitral, la intensidad del juego o las circunstancias del partido.

Interpretación de resultados en función de las preguntas iniciales:

* La pregunta clave respecto a la eficiencia ofensiva encuentra respuesta en la fuerte correlación entre tiros a puerta y goles. Equipos como el Liverpool y el Manchester City destacaron por necesitar pocos tiros efectivos para anotar, mostrando una alta eficiencia.
* En contraste, equipos como el Fulham y Sheffield United evidenciaron menor efectividad, necesitando más intentos para lograr un gol.

# Conclusiones y recomendaciones

* **Tabla de posiciones:** El Everton fue sancionado con una deducción de 8 puntos en la liga, finalizando la temporada con 40 puntos, esto por violaciones a la regla de rentabilidad y sostenibilidad de la Premier League.

Al igual que el Everton, el Nottingham Forest también fue sancionado con una deducción de 4 puntos por violaciones a la regla de rentabilidad y sostenibilidad de la Premier League, finalizando la temporada con 32 puntos.

* **Goles:** Se evidencia que la moda de goles es 1, es decir, es más común encontrar partidos donde los equipos anotan solo un gol.

El mejor promedio de goles fue el de Manchester City (2,53 goles por partido), mientras que el peor promedio fue para el Sheffield United (0,92 goles por partido)

* **Goles (Local):** Se evidencia que la moda de goles es 1, es decir, es más común encontrar partidos donde los equipos locales anotan solo un gol.

El mejor promedio de goles de local fue el de Manchester City (2,68 goles por partido), mientras que el peor promedio fue para el Burnley y el Sheffield United (1,00 gol por partido).

* **Goles (Visitante):** Se evidencia que la moda de goles es 1, es decir, es más común encontrar partidos donde los equipos visitantes anotan solo un gol.

El mejor promedio de goles de visitante fue el de Manchester City (2,37 goles por partido), mientras que el peor promedio fue para el Sheffield United (0,84 goles por partido).

* **Tiros y Tiros a Puerta:** El mejor promedio de tiros por partido es del Liverpool (20,79 tiros por partido)

El mejor promedio de tiros a puerta por partido es del Manchester City (7,32 tiros a puerta por partido)

El peor promedio de tiros por partido es del Sheffield United (9,68 tiros por partido), y este equipo también tiene el peor promedio de tiros a puerta por partido (3,39 tiros a puerta por partido).

* **Corners, Faltas y Amarillas:** El mejor promedio de corners ganados por partido es del Liverpool (7,55 corners por partido), mientras que el peor promedio de corners ganados por partido es del Sheffield United (3,74 corners por partido).

El equipo que cometió más faltas fue el Bournemouth (13,32 faltas por partido), mientras que el equipo que menor número de faltas cometió fue el Manchester City (7,55 faltas por partido).

El equipo que más tarjetas amarillas recibió fue el Chelsea (2,74 amarillas por partido), mientras que el equipo que menos tarjetas amarillas recibió fue el Manchester City (1,34 amarillas por partido).

* **Goles vs Tiros a puerta:** Se evidencia que en promedio cada equipo necesitó alrededor de tres tiros a puerta para convertir el primer gol, y necesitaron entre tres y cuatro tiros a puerta adicionales para convertir un nuevo gol.

El Fulham fue el equipo que necesitó en promedio alrededor de cuatro tiros a puerta para convertir un gol, y requería entre dos y tres tiros a puerta adicionales para convertir un nuevo gol. Sin embargo, su tasa de tiros a puerta por partido fue muy baja (promedio de 4,68 por partido), además de que en 13 de sus 39 partidos no convirtió goles (rematando máximo seis tiros a puerta) y en 10 partidos tan solo convirtió de a un gol (rematando también máximo seis tiros a puerta).

El Liverpool fue el equipo que necesitó en promedio entre 1 y 2 tiros a puerta para convertir el primer gol, y requería alrededor de 5 tiros a puerta adicionales para convertir un nuevo gol.

* **Faltas vs Amarillas:** El valor de la pendiente, 0.147, y el valor del intercepto, 0.459, indica que los equipos durante la temporada empezaron a recibir tarjetas amarillas tras cometer entre 3 y 4 faltas, y por cada 6 o 7 faltas adicionales recibían una nueva tarjeta amarilla. Sin embargo, esta relación no es buena en general debido al valor de correlación, 0.38, el cual es bajo.

En promedio, el Everton empezó a recibir tarjetas amarillas en sus encuentros tras cometer entre 7 y 8 faltas, y por cada 4 o 5 faltas adicionales recibía una nueva tarjeta amarilla. En tal solo 4 partidos de liga no recibió ninguna tarjeta amarilla, cometiendo 6, 8, 12 y 14 faltas en estos partidos. En un partido recibió una tarjeta amarilla cometiendo tan solo 6 faltas. En sus 33 partidos restantes cometió al menos ocho faltas y recibió al menos una tarjeta amarilla. Este análisis es relativamente bueno dado que el coeficiente de correlación considerando solamente al Everton es de 0.6.

En promedio, el Manchester City empezó a recibir tarjetas amarillas en sus encuentros tras cometer entre 5 y 6 faltas, y aproximadamente cada 5 faltas adicionales recibía una nueva tarjeta amarilla.

# Anexos

* + Data PremierLeague23\_24.xlsx
  + Tablero Power BI Análisis\_PremierLeague23-24
  + Documentación de funciones o librerías usadas Google Colab DataXeperence.ipynb LINK: <https://colab.research.google.com/drive/1PIvPkoi-n4AVgMUn4BrUHg3samP3EljX?usp=sharing>