



Proyecto en R

Juan Nocetti y Ionas Josponis

Universidad Católica del Uruguay
Programación para el análisis de datos

Montevideo, Uruguay

27 de Julio de 2025

Serie A Análisis

Dataset

Como fanáticos del fútbol, decidimos explorar Kaggle en busca de algún dataset centrado en este deporte. Dimos con un dataset de la Serie A italiana que recopila, para cada partido desde la temporada 2020/21 hasta la 2024/25, información como la fecha, equipos, resultado, goles, estadísticas de posesión y entre otras más. El dataset cuenta con 3800 filas y 29 columnas. Cada fila corresponde a un partido y cada columna a los atributos del mismo.

Link: <https://www.kaggle.com/datasets/marcelbiezunski/serie-a-matches-dataset-2020-2025>

Preguntas elegidas

1. ¿La vuelta del público post-pandemia influye en los resultados de los equipos locales?
2. ¿Los clásicos Milan–Inter, Roma–Lazio y Torino–Juventus atraen más público y son más parejos que el resto de los partidos?

Análisis de los atributos del dataset

- **Date:** Fecha en la que se jugó el partido. (tipo: categórica ordinal)
- **Time:** Hora de inicio del partido. (tipo: categórica ordinal)
- **Comp:** Nombre de la competición (siempre “Serie A”). (tipo: categórica nominal)
- **Round:** Jornada o número de la fecha del campeonato. (tipo: categórica ordinal)
- **Day:** Día de la semana en que tuvo lugar el partido (ej. Dom, Sáb). (tipo: categórica nominal)
- **Venue:** “Home” si el equipo (columna Team) jugó en su estadio; “Away” en caso contrario. (tipo: categórica binaria)
- **Result:** Resultado para el equipo de la columna Team: W (Win) = Victoria, D (Draw) = Empate, L (Loss) = Derrota. (tipo: categórica nominal)
- **GF:** Goles a favor – número de goles marcados por el equipo. (tipo: numérica discreta)
- **GA:** Goles en contra – número de goles encajados frente al rival. (tipo: numérica discreta)
- **Opponent:** Nombre del equipo rival. (tipo: categórica nominal)

- **xG**: Goles esperados – estimación de la calidad de las oportunidades creadas por el equipo. (tipo: numérica continua)
- **xGA**: Goles esperados en contra – estimación de la calidad de las oportunidades concedidas. (tipo: numérica continua)
- **Poss**: Porcentaje de posesión que tuvo el equipo durante el partido. (tipo: numérica continua)
- **Attendance**: Número de espectadores presentes en el estadio. (tipo: numérica discreta)
- **Captain**: Nombre del capitán del equipo en ese partido. (tipo: categórica nominal)
- **Formation**: Formación táctica utilizada por el equipo (p.ej. 4-3-3). (tipo: categórica nominal)
- **Opp Formation**: Formación táctica utilizada por el rival. (tipo: categórica nominal)
- **Referee**: Nombre del árbitro del partido. (tipo: categórica nominal)
- **Match Report**: Enlace o referencia al informe oficial del partido. (tipo: categórica nominal)
- **Notes**: Observaciones o eventos especiales adicionales (p.ej. aplazado, abandonado). (tipo: categórica nominal)
- **Sh**: Total de disparos (shots) realizados por el equipo. (tipo: numérica discreta)
- **SoT**: Disparos a puerta (shots on target) – disparos que fueron entre los tres palos. (tipo: numérica discreta)
- **Dist**: Distancia media recorrida por el equipo (en kilómetros). (tipo: numérica continua)
- **FK**: Número de tiros libres (free kicks) a favor del equipo. (tipo: numérica discreta)
- **PK**: Número de penaltis concebidos y ejecutados por el equipo. (tipo: numérica discreta)
- **PKatt**: Número de intentos de penalti realmente lanzados. (tipo: numérica discreta)
- **Season**: Temporada (p.ej. 2020/21 se etiqueta como “2021”, 2021/22 como “2022”, etc.). (tipo: categórica ordinal)
- **Team**: Nombre del club al que corresponden los datos de esa fila. (tipo: categórica nominal)

Preparación de datos

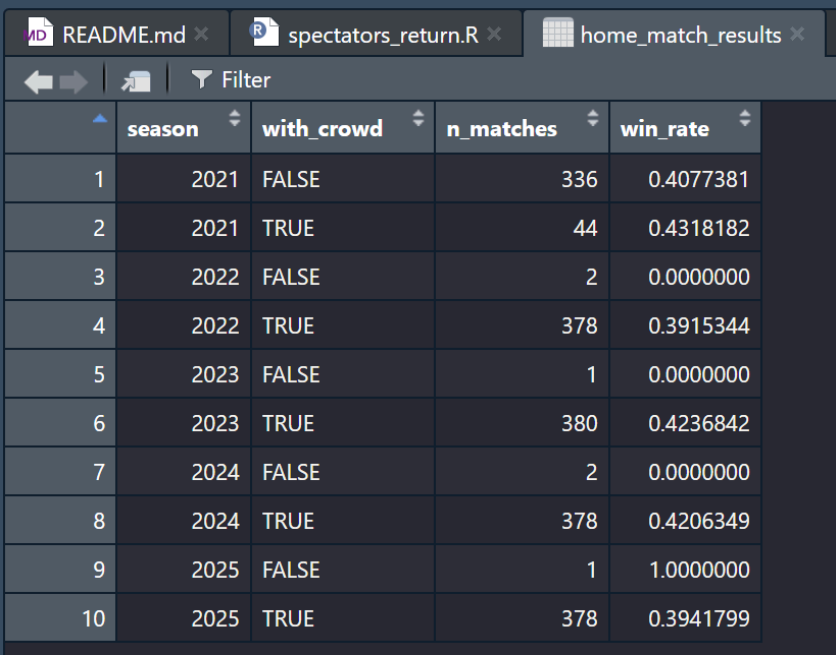
Se realizó una limpieza inicial que incluyó la selección de las variables necesarias en función de las preguntas a responder, descartando aquellas columnas que eran irrelevantes para los objetivos planteados. Además, se verificó la existencia de valores faltantes (NA), eliminando aquellas filas incompletas para evitar distorsiones en los análisis estadísticos. También se detectaron y removieron registros duplicados. Cada uno de los scripts desarrollados para las preguntas del trabajo realizó una limpieza independiente pero siguiendo estos mismos criterios, asegurando consistencia y calidad en los datos utilizados para los gráficos, cálculos de correlación, tablas de frecuencias y comparaciones entre partidos.

Pregunta 1

¿La vuelta del público post-pandemia influye en los resultados de los equipos locales?

Primero para responder esta pregunta transformamos el data frame original y generamos uno nuevo que sólo conserva cinco columnas. Convertimos el texto Date en un objeto de fecha (date) y a partir de este calculamos la temporada(season). Luego eliminamos valores faltantes y duplicados.

Posteriormente agrupamos cada partido de local según la temporada y si hubo público (with_crowd). Para cada combinación calculamos el número total de encuentros (n_matches) y la tasa de victorias de equipos locales (win_rate). Si vemos el dataset obtenemos.



	season	with_crowd	n_matches	win_rate
1	2021	FALSE	336	0.4077381
2	2021	TRUE	44	0.4318182
3	2022	FALSE	2	0.0000000
4	2022	TRUE	378	0.3915344
5	2023	FALSE	1	0.0000000
6	2023	TRUE	380	0.4236842
7	2024	FALSE	2	0.0000000
8	2024	TRUE	378	0.4206349
9	2025	FALSE	1	1.0000000
10	2025	TRUE	378	0.3941799

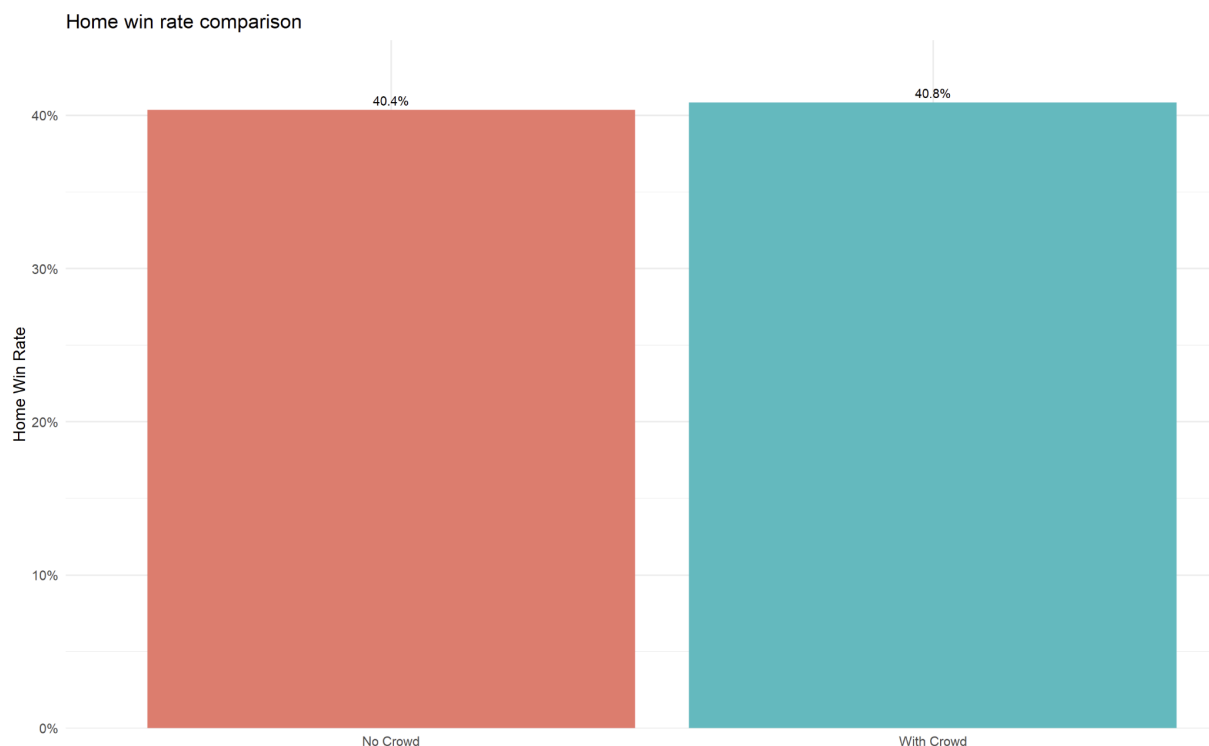
La tabla muestra para cada temporada cuántos encuentros en casa se disputaron sin público frente a con público y la tasa de victorias locales en cada grupo. Llama la atención que incluso tras el fin de las restricciones sanitarias, hay varios partidos a puertas cerradas en las campañas 2022–2025, lo que apunta a sanciones disciplinarias o problemas de registro de datos. Al extraer las fechas e investigar, aparecen:

24 de febrero de 2024: Genoa - Udinese se jugó sin público tras un caso de abuso racista hacia el portero Mike Maignan. Udinese fue sancionado con cierre de grada (ESPN, 2024).

16 de septiembre de 2024: Genoa - Juventus se disputó a puertas cerradas tras graves disturbios entre aficiones en un partido de Copa Italia (Reuters, 2024).

Para el resto de fechas (10-12-2021, 27-4-2022, 11-6-2023, 9-3-2024) no se hallaron comunicados de sanción. Es posible que diciembre de 2021 aún refleje aforos muy reducidos por una nueva variante del COVID y que los ceros de 2022/23 fueran partidos aplazados o datos no volcados correctamente.

A continuación representamos gráficamente la tasa de victorias en casa con público frente a sin público:



Con esta gráfica podemos concluir que la presencia de público apenas incrementó la tasa de victorias locales (40,4 % sin afición contra 40,8 % con ella), por lo que no hay evidencia de un efecto relevante de la vuelta de los espectadores en el desempeño del equipo.

Conclusión

¿La vuelta del público tras la pandemia influyó en los resultados en casa?

La vuelta del público tras la pandemia no tuvo un impacto significativo en los resultados en casa: la tasa de victorias locales apenas se incrementó en un 0,4 %, pasando de un 40,4 % sin espectadores a un 40,8 % con ellos.

Elementos solicitados en el análisis exploratorio de datos

Cuadro de frecuencias relativas y absolutas de una variable categórica

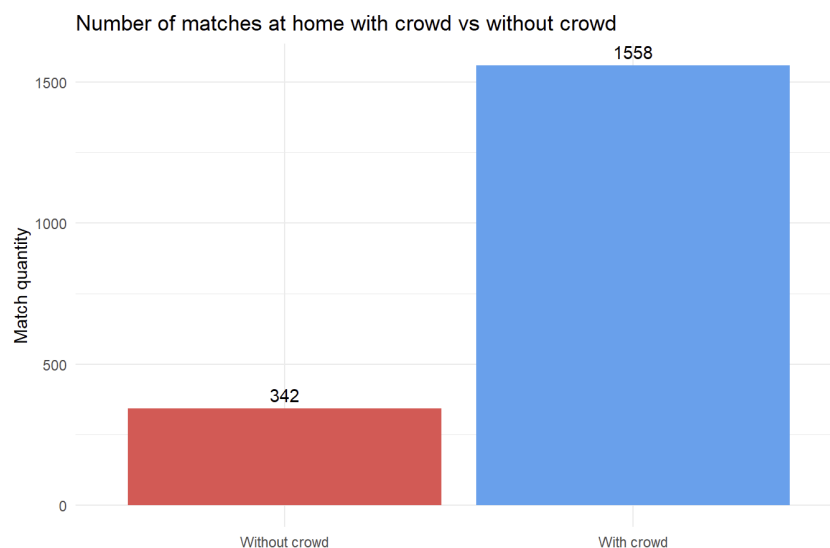
A continuación se muestra un cuadro que muestra cuántos partidos en casa se jugaron sin público (`with_crowd = FALSE`) frente a con público (`with_crowd = TRUE`), tanto en valores absolutos como relativos.

	with_crowd	n	relative_frequency
	<lgl>	<int>	<dbl>
1	FALSE	342	0.18
2	TRUE	1558	0.82

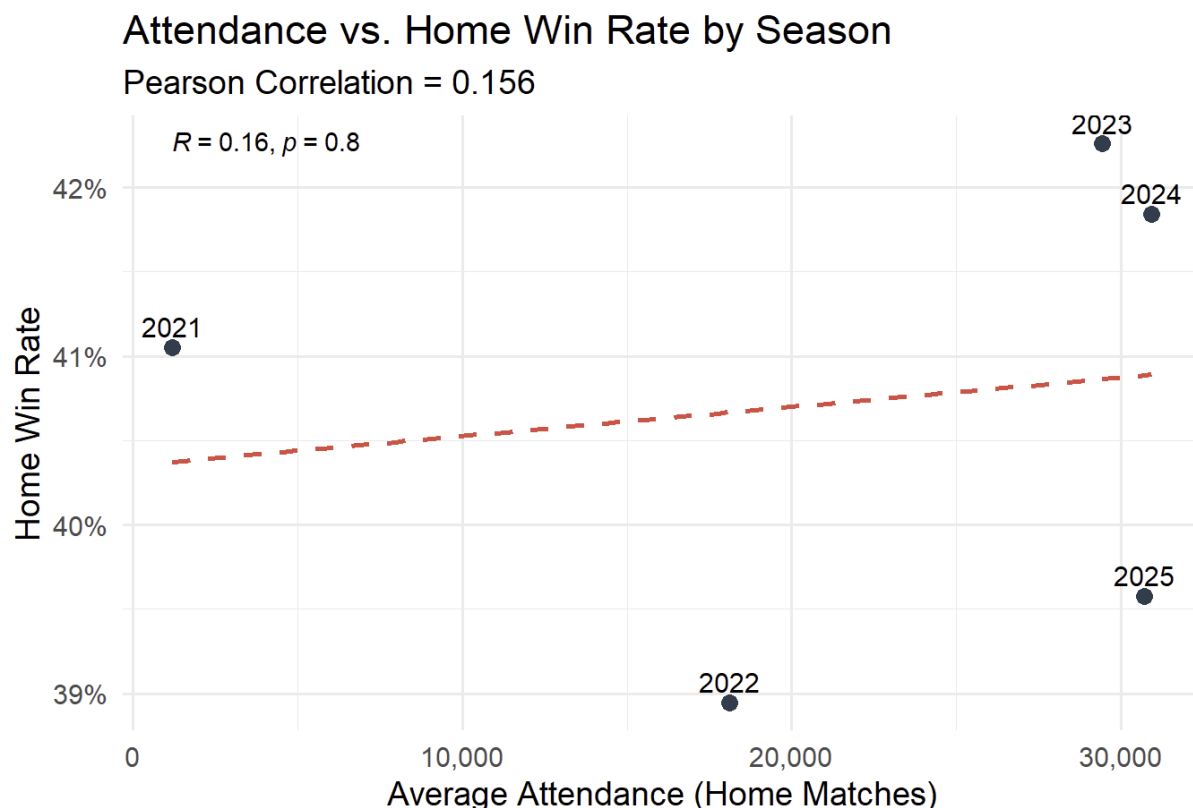
En total hubo 1.900 partidos en casa, de los cuales 342(18%) fueron con puertas cerradas y 1.558(82 %) con espectadores. Esto nos permite ver de forma rápida que la gran mayoría de los encuentros del dataset se disputaron con público.

Gráficos en ggplot2

1. **Gráfico de barras de una variable categórica (`with_crowd`):** Este gráfico muestra el número de partidos en casa sin público (342) frente a con público (1.558). La altura de cada barra evidencia que la gran mayoría de los encuentros se disputó con público.

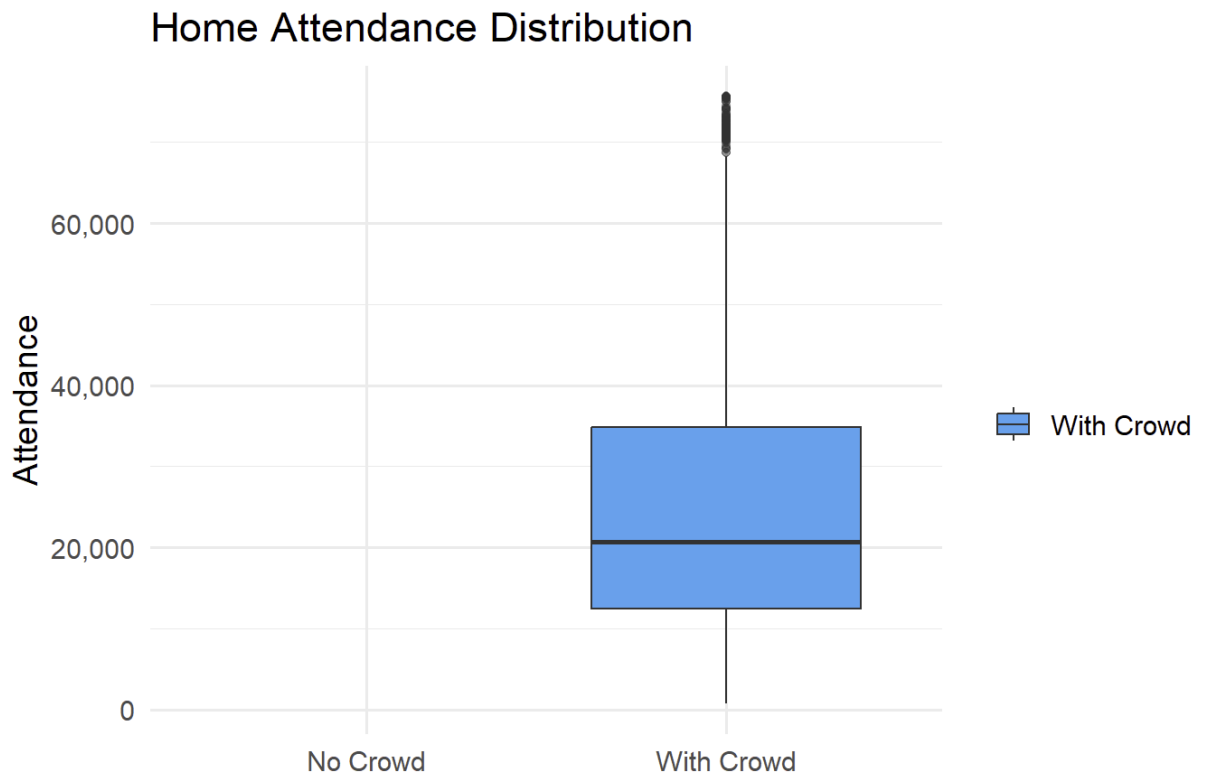


2. **Gráfico de dispersión entre dos variables continuas:** Este gráfico compara la asistencia media en casa con la tasa de victorias locales, temporada a temporada. Aunque la línea de tendencia es ligeramente ascendente, el coeficiente indica una correlación muy débil y no significativa. En definitiva, el volumen de público no parece estar ligado de forma consistente a mejores resultados en casa.



3. **Boxplot de una variable continua desagregado por una variable categórica:**
A continuación observamos que la categoría “No Crowd” carece de variabilidad (todos los valores están en cero), por lo que únicamente se dibuja el boxplot de los partidos con público. Ahí la mediana de asistencia se ubica en torno a los 20.000 espectadores, con un rango intercuartílico que va aproximadamente de 13.000 a 34.000. Los números aislados por encima de los 60.000 reflejan la existencia de encuentros con estadios de alta capacidad, mientras que el extremo inferior del bigote casi toca el cero, indicando alguna excepción de muy baja concurrencia.

En conjunto, el gráfico confirma una amplia dispersión de la asistencia cuando hay público, concentrándose la mayoría de los partidos entre los 20.000 y 30.000 asistentes.



Pregunta 2

¿Los clásicos Milan–Inter, Roma–Lazio y Torino–Juventus atraen más público y son más parejos que el resto de los partidos?

El análisis comenzó con la carga del dataset completo y la selección de las columnas relevantes para responder la pregunta, estas fueron:

- Team

- Opponent
- Result
- Attendance
- Poss
- xG
- xGA

Luego se eliminaron las filas con valores faltantes y duplicados para asegurar la calidad de los datos.

Esta pregunta es posible separarla en dos partes:

La primera parte corresponde a que si los clásicos atraen más público que el resto del partido, nos enfocamos en comparar las distribuciones completas de la asistencia del público en clásicos vs el resto de partidos (Attendance).

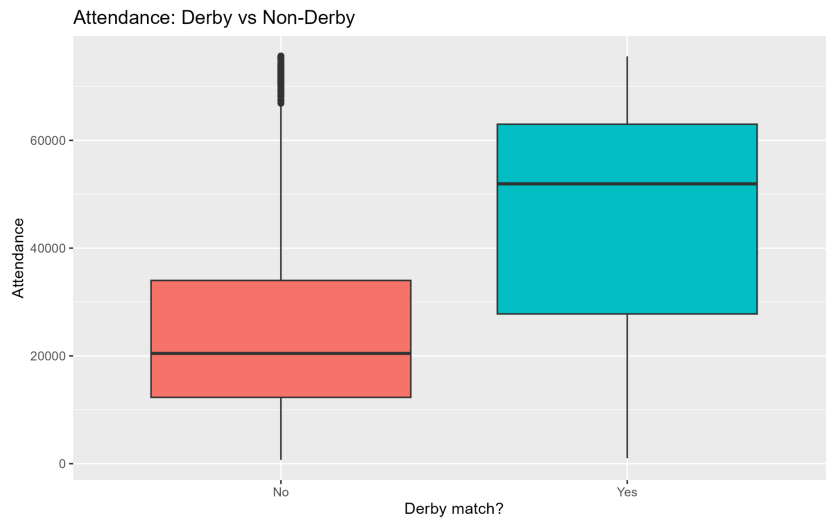
Finalmente para responder si los clásicos fueron más parejos que el resto de los partidos, se construyeron tres métricas específicas que permiten evaluar si el partido fue parejo. La primera métrica fue el balance de posesión, calculado como el valor absoluto de la diferencia entre el porcentaje de posesión y un reparto equitativo del balón (es decir 50%). Una posesión del balón cercana al 50% sugiere que ninguno de los dos equipos dominó ampliamente el juego. La segunda métrica fue la diferencia de goles esperados que compara la calidad de las oportunidades generadas por el equipo con la calidad de las que tuvo en contra. Cuanto menor es esta diferencia, más similar fue el rendimiento ofensivo de ambos equipos, lo que nuevamente indica mayor equilibrio del partido. Por último, se analizó la proporción de partidos que terminaron en empate a través de una variable lógica (Result == "D"), considerando que un resultado final como empate es el mayor reflejo de un partido parejo a pesar de que como es fútbol no siempre se da así.

Se eligió utilizar gráficos tipo boxplot para comparar las variables entre clásicos y no clásicos porque permiten visualizar de forma clara la distribución, mediana y la presencia de valores atípicos (caso que no se encontró).

Gráficos

En los gráficos representaremos a los clásicos en color verde y el resto de los partidos en color rojo.

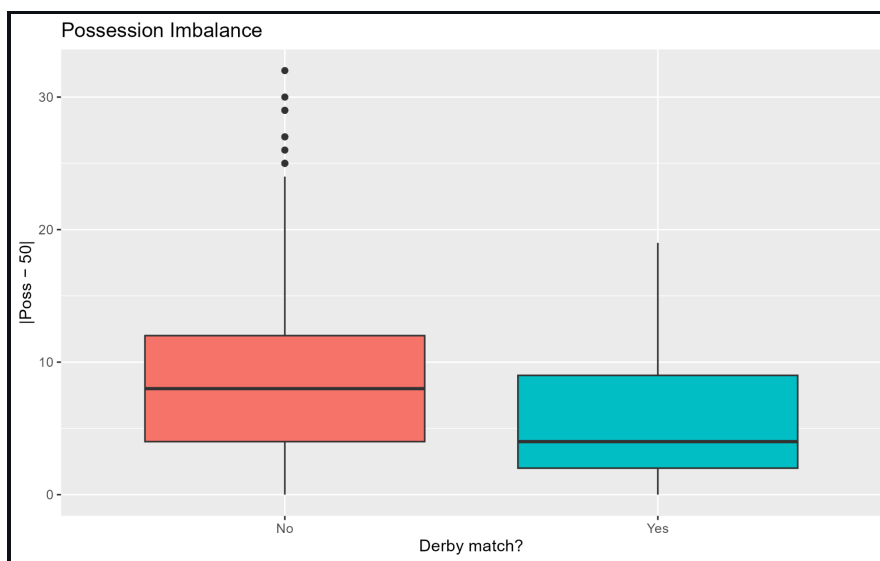
Distribución de asistencia de público en clásicos vs. no clásicos



El gráfico muestra claramente que los clásicos presentan una mediana de asistencia considerablemente más alta, además de una distribución desplazada hacia valores mayores.

Además, en los partidos no clásicos se observan muchos outliers en el extremo superior lo que representa encuentros puntuales con alta asistencia. Esto también se refleja en una mayor dispersión, lo que indica que el número de espectadores varía mucho entre partidos. En cambio, los clásicos tienen una distribución más concentrada en valores altos con menos variabilidad, lo que sugiere una asistencia más constante y elevada en este tipo de encuentros.

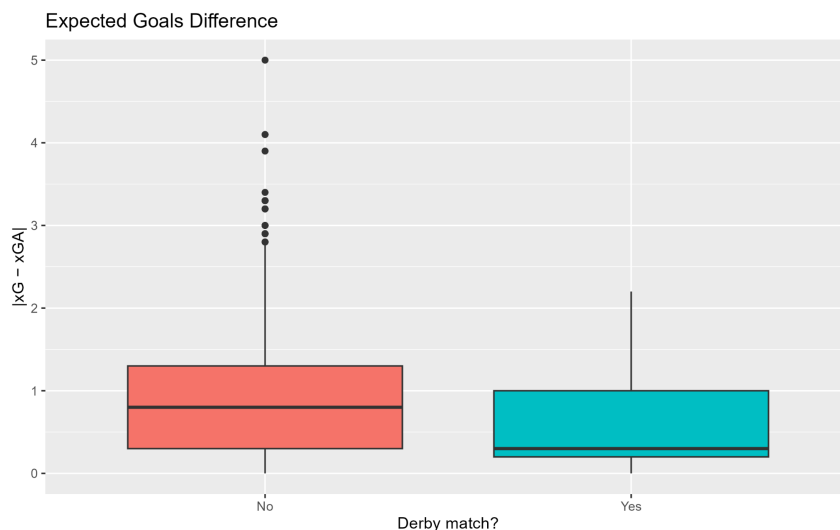
Posesión equilibrada



En este caso, cuanto más cerca del 0 están los boxplots más parejo en posesión fue el partido. Por ejemplo, en caso que la posesión sea 50% sería lo más parejo ya que la diferencia sería igual a 0 (50-50), por lo tanto, lo que se calcula es el desbalance que mide lo lejos que estuvo la posesión de 50%.

Tal como se aprecia en la gráfica, en el caso de los clásicos se tiene menos dispersión (viendo el área de las cajas) lo que nos indica menos variabilidad en los resultados obtenidos y una mediana más baja lo que nos sugiere que estos partidos son más equilibrados en términos de posesión de balón.

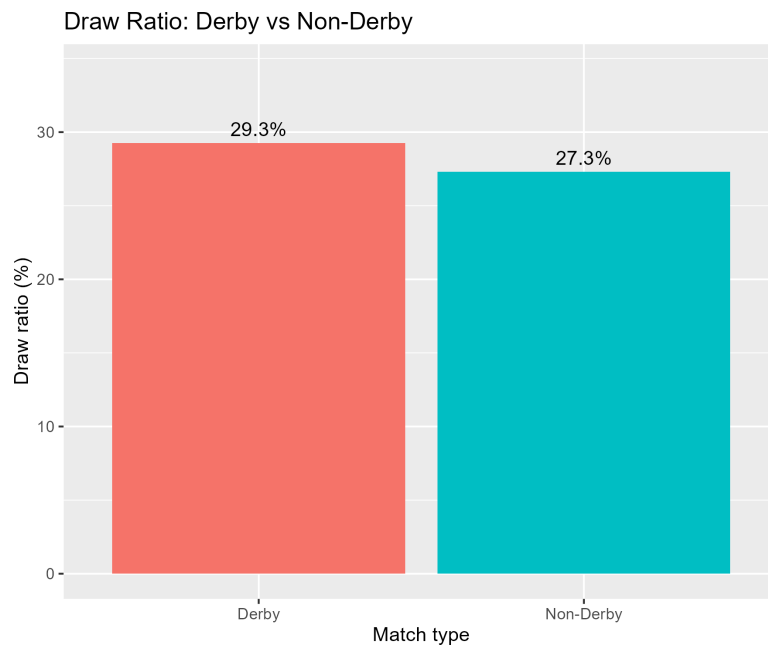
Diferencia de ocasiones de gol esperadas de ambos equipos



Los partidos no-derbi tienen una mayor mediana de diferencia esperada de goles y una dispersión más grande, con varios outliers (casos donde un equipo dominó mucho más).

En cambio, los derbis tienen una mediana más baja y están más concentrados cerca del 0, lo que indica que las oportunidades fueron más similares entre ambos equipos.

Porcentaje de empates



Los derbis tienen un 29.3% de empates, mientras que los partidos comunes tienen un 27.3%. Aunque la diferencia no es muy grande, los clásicos tienden ligeramente más al empate, lo cual puede interpretarse como un signo de mayor paridad o equilibrio competitivo.

Conclusión

Los análisis realizados muestran evidencia clara de que los clásicos entre Milan–Inter, Roma–Lazio y Torino–Juventus atraen más público que el resto de los partidos de la Serie A. Esto se refleja en el gráfico de asistencia del público, donde la mediana y la distribución en los clásicos son notablemente más altas que en los demás encuentros. Además, al evaluar si estos partidos son más parejos, se observa que las métricas de equilibrio como la diferencia de goles esperados, el desbalance de posesión respecto al 50%, y la proporción de empates tienden a ser más equilibradas en los clásicos. Estas diferencias indican que los clásicos no solo convocan a más espectadores, sino que también suelen ser enfrentamientos más peleados lo cual refuerza su carácter competitivo e impredecible.

Elementos solicitados en el análisis exploratorio de datos

i. Cuadro de frecuencias relativas y absolutas de una variable categórica

Se construyó una tabla de frecuencias absolutas y relativas para la variable categórica **Result**, que indica el resultado del partido desde la perspectiva del equipo (**Team**): victoria (**W**), empate (**D**) o derrota (**L**). Se observa que tanto las victorias como las derrotas

representan un 36.3% de los casos cada una, mientras que los empates constituyen el 27.3%. Esto permite visualizar la distribución de resultados a lo largo del dataset.

	Result	Absolute	Relative
	<chr>	<int>	<dbl>
1	D	852	0.273
2	L	1132	0.363
3	W	1132	0.363

ii. Estadísticos descriptivos de la variable Attendance

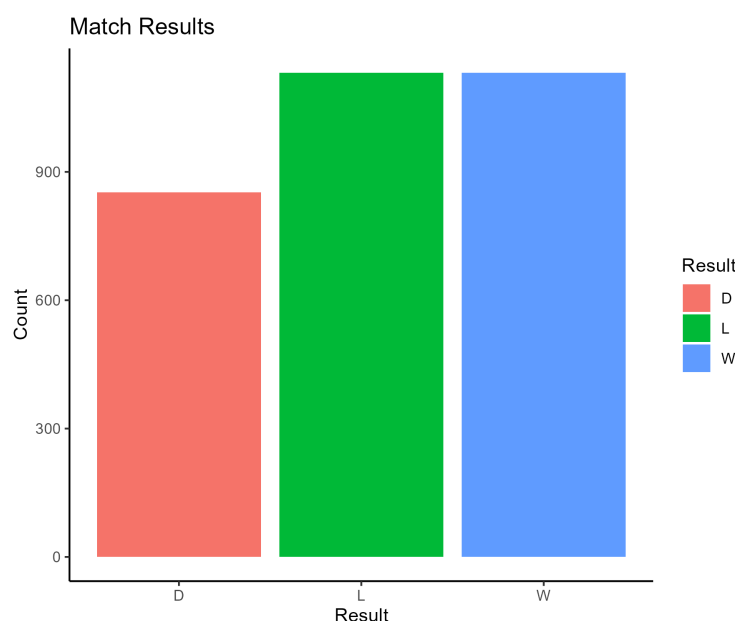
Se calcularon tres estadísticos descriptivos para la variable **Attendance**, que representa la cantidad de público presente en cada partido. La **media** fue de 26.560 espectadores, mientras que la **mediana** fue de 20.708, al ser la media mayor a la mediana nos indica que hay partidos con asistencias muy altas que elevan el promedio (outliers). La **desviación estándar** fue de 19.449, lo cual muestra una alta variabilidad en la asistencia entre distintos encuentros.

	Mean	Median	Std_dev
	<dbl>	<dbl>	<dbl>
1	26560.	20708.	19449.

iii. Gráficos en ggplot2

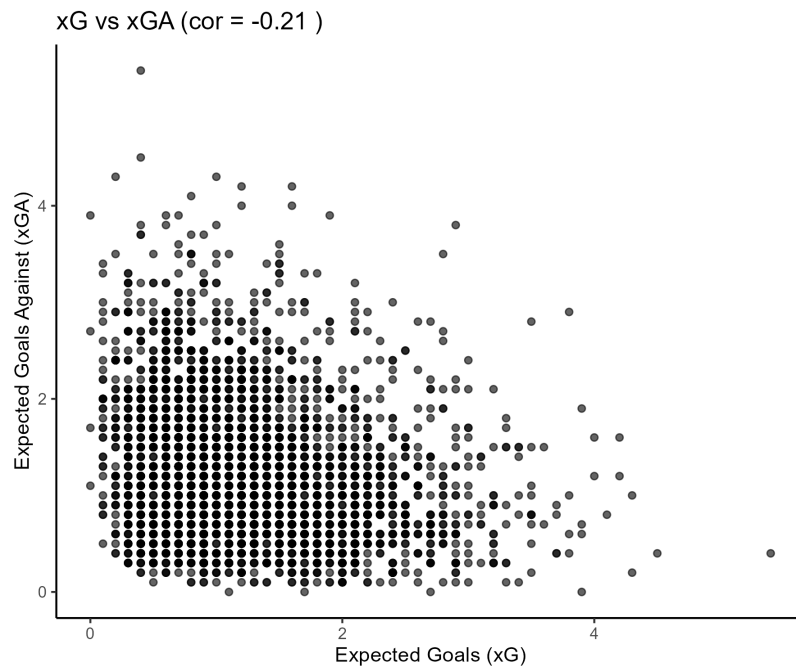
1. Gráfico de barras de una variable categórica (**Result**)

El gráfico de barras muestra la distribución de la variable categórica **Result**. Se observa que las victorias y derrotas tienen una frecuencia similar, ambas superiores a la de los empates. Esto refleja que la mayoría de los partidos tienen un ganador, y que los empates son menos comunes en la Serie A.



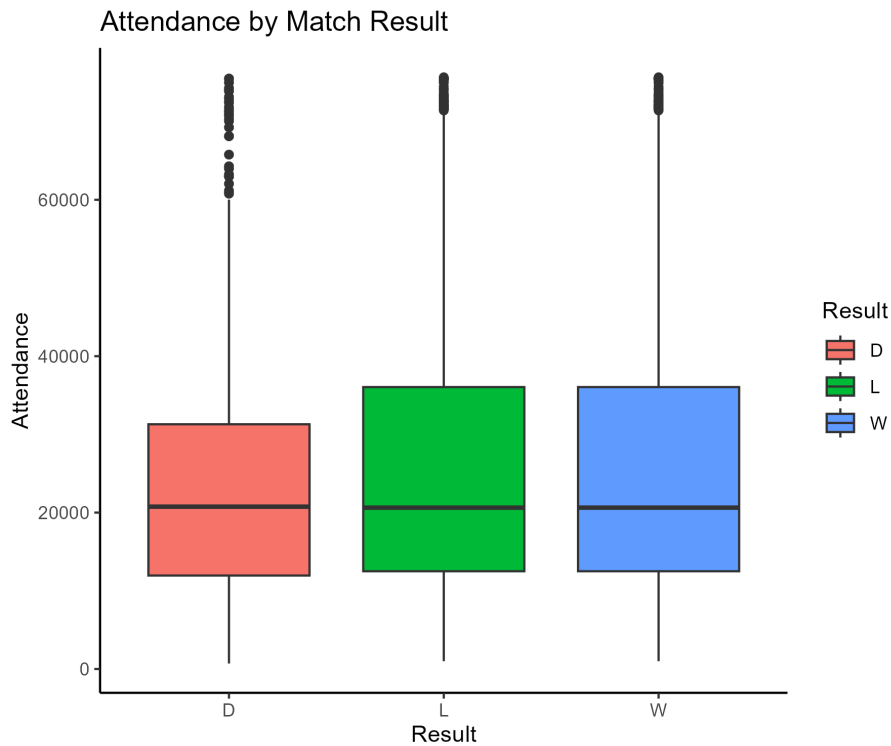
2. Gráfico de dispersión entre dos variables continuas (**xG** vs **xGA**)

Este gráfico muestra la relación entre los goles esperados a favor (**xG**) y en contra (**xGA**) para cada partido. Visualmente, los puntos están bastante dispersos, y el coeficiente de correlación calculado es -0.21, indicando una relación débil y negativa. Esto sugiere que generar más ocasiones de gol no implica necesariamente ser menos atacado por el rival.



3. Boxplot de una variable continua desagregado por una variable categórica (Attendance según Result)

En este gráfico se compara la distribución de asistencia de público (**Attendance**) en función del resultado del partido (**Result**). Se observa gran variabilidad en todos los casos, con muchos valores atípicos. Las medianas son similares entre resultados, aunque los partidos ganados y perdidos parecen atraer levemente más público que los empatados.



Referencias:

ESPN. (2024, 26 de febrero). Udinese to play behind closed doors after racist abuse of Maignan. ESPN.com.

https://www.espn.com/soccer/story/_/id/39369058/udinese-play-closed-doors-maignan-racist-abuse

Reuters. (2024, 27 de septiembre). Genoa home match against Juventus to be played behind closed doors. Reuters.

<https://www.reuters.com/sports/soccer/genoa-home-match-against-juventus-be-played-behind-closed-doors-2024-09-27/>