Universidad Icesi

Análisis Exploratorio de Datos

Proyecto Final

Autores:

María Paula Fernández

Cristian Palechor

Jorge Antonio Morales

Facultad de Ingeniería

Maestría en Ciencia de Datos

Marzo de 2022

Índice

Introducción

Debido a la pandemia del COVID-19 que asotó al mundo en el 2020, muchos eventos masivos se han visto afectados, entre ellos el futbol. Es por esto que se ha decidido realizar un estudio con el propósito de saber el impacto que tiene el público en los jugadores que juegan de local a la hora del partido cuando éste se encuentra en la Premier League.

Debido a lo anterior, el objetivo principal para esta primera parte del informe es realizar el análisis exploratorio de los partidos de la Premier League, describiéndolos con el propósito de encontrar relación entre el rendimiento de los equipos locales y el aforo en los estadios durante la pandemia del Covid-19.

Se recopilaron datos históricos de partidos de la Premier League desde el 2018 hasta mediados del 2021, estos datos fueron recolectados por la empresa propietaria del proyecto.

La estructura del presente documento consta en: 1) introducir las variables clave para el cumpliemiento del objetivo del proyecto, analizándolas y limpiándolas de cualquier tipo de anomalías, 2) realizar un análisis exploratorio de los datos, empezando por hacer un análisis univariado de todas las variables y después un multivariado teniendo en cuenta nuestra variable dependiente con relación a las demás variables explicativas, y, por último, 3) un aparto de conclusiones y recomendaciones del respectivo informe.

Sobre los datos

El repositorio de los datos cuenta con tres bases de datos independientes que corresponden a los partidos disputados por la liga inglesa entre los años 2018 - 2019 (primera base de datos con 380 registros), 2019 - 2020 (segunda base de datos con 383 registros) y 2020 - 2021 (tercera base de datos con 385 registros).

La primera base de datos posee 62 variables que representan a los datos más característicos del partido y las probabilidades según diferentes casas de apuestas. Por otra parte, tanto la segunda y tercera base de datos cuenta con 106 variables idénticas que muestran también los datos más característicos del juego y las diferentes probabilidades según las casas de apuestas.

Con el propósito de cumplir con los objetivos del proyecto, se seleccionaron las variables de interés que tienen que ver directamente con el rendimiento de los equipos durante ciertos periodos de tiempo (antes, durante y después de la pandemia). En cuanto a las columnas que representan las probabilidades de las casas de apuestas, se llegó a la conclusión de omitirlas debido a que éstas no influyen en el impacto que tiene en los equipos que juegan de local cuando hay puerta cerrada.

Después de haber hecho la selección de las variables de interés y unir las tres bases de datos, el dataset general cuenta con 19 columnas y 1148 registros. Es con esta información que se procederá a trabajar a lo largo de este informe.

Además, se agregaron 3 variables nuevas para el cumplimiento del objetivo del proyecto:

-**Pandemic**, describe si el partido se jugó sin restricciones de público, a puerta cerrada o con aforo limitado.

-**HomeTeamSize**, esta variable define si el equipo local es grande, mediano o pequeño:

– Los equipos grandes se clasificaron teniendo en cuenta la sumatoría de sus titulos y número de veces subcampeón es mayor o igual que veinte y haya quedado campeón en los últimos veinte años.

– Los equipos medianos se clasificaron teniendo en cuenta la sumatoría de sus titulos y número de veces subcampeón está entre diez y diecienueve.

– Los equipos pequeños se clasificaron teniendo en cuenta la sumatoría de sus titulos y número de veces subcampeón es menor que diez y no haya quedado campeón en los últimos veinte años.

-**AwayTeamSize**, esta variable conserva la misma clasificación que para los equipos locales.

Cuadro 1: Tabla de variables definitivas

Variable	Tipo de variable	Rango o categoría
Date	Cuantitativa - Discreta	Fecha en el que se jugó el partido
HomeTeam	Cualitativa - Nominal	Equipo local

continúa en la siguiente página

Cuadro 1: Tabla de variables definitivas continuación

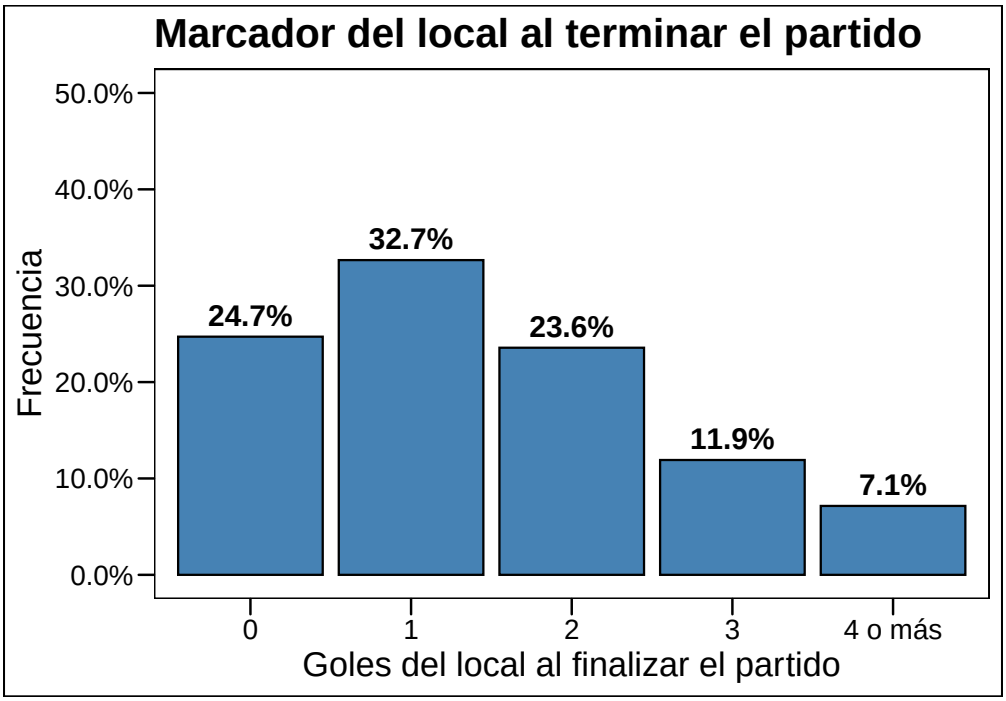
Variable	Tipo de variable	Rango o categoría
AwayTeam	Cualitativa - Nominal	Equipo visitante
FTHG	Cualitativa - Ordinal	Goles del local local al finalizar el partido
FTAG	Cualitativa - Ordinal	Goles del visitante al finalizar el partido
FTR	Cualitativa - Nominal	Resultado final del partido
HTHG	Cualitativa - Ordinal	Goles del local el medio tiempo
HTAG	Cualitativa - Ordinal	Goles del visitante al medio tiempo
HTR	Cualitativa - Nominal	Resultado final del primer tiempo
HST	Cuantitativa - Discreta	Tiros al arco por parte del equipo local
AST	Cuantitativa - Discreta	Tiros al arco por parte del equipo visitante
HF	Cuantitativa - Discreta	Faltas por parte del equipo local
AF	Cuantitativa - Discreta	Faltas por parte del equipo visitante
HC	Cuantitativa - Discreta	Tiros de esquina por parte del equipo local
AC	Cuantitativa - Discreta	Tiros de esquina por parte del equipo visitante
HY	Cualitativa - Ordinal	Tarjetas amarillas por parte del equipo local
AY	Cualitativa - Ordinal	Tarjetas amarillas por parte del equipo visitante
HR	Cualitativa - Ordinal	Tarjetas rojas por parte del equipo local
AR	Cualitativa - Ordinal	Tarjetas rojas por parte del equipo visitante
HomeTeamSize	Cualitativa - Nominal	Pequeño - Mediano - Grande
AwayTeamSize	Cualitativa - Nominal	Pequeño - Mediano - Grande
Pandemic	Cualitativa - Nominal	Sin restricciones - Aforo Limitado - Puerta cerrada

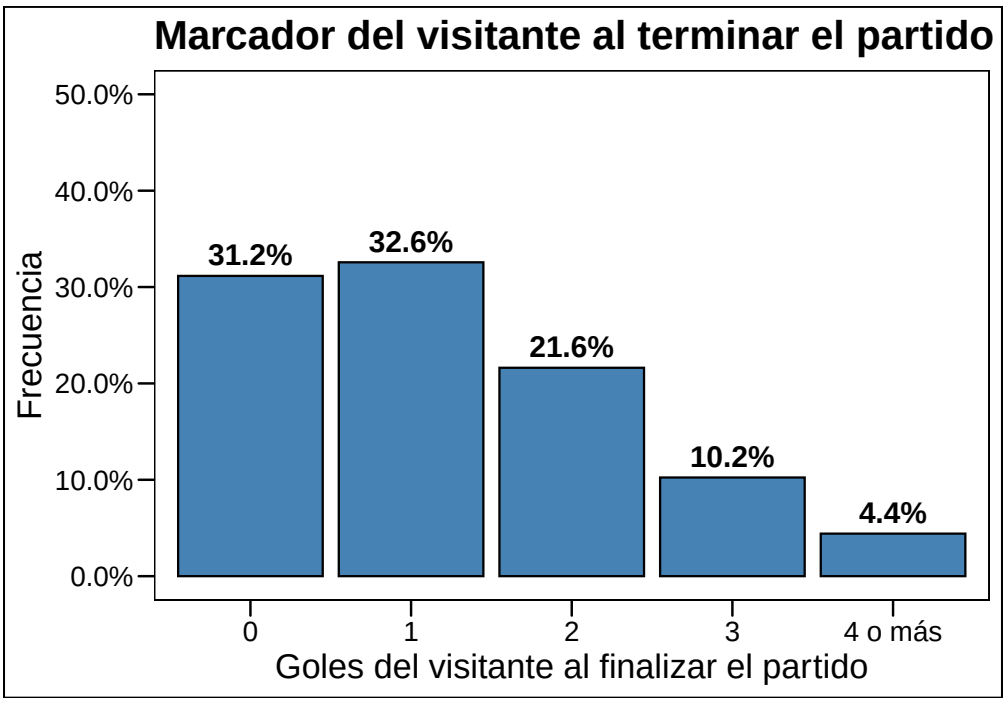
Con respecto a la detección y limpieza de anomalías de la base de datos, se halló una perdida aleatoria de datos (MCaR) con los resultados del marcador para algunos partidos, sin embargo, siendo información fácil de acceder, se lograron obtener los resultados y adicionarlos en los faltantes. Por otra parte, se encontró una perdida de datos en las variables de descripción de tiros al arco tanto para los locales como para los visitantes del 0.61 %, por lo que se optó por realizar un listwise deletion debido a que dichas filas contenían campos vacíos en la mayoría de sus columnas. En resumen, de las 37 anomalías encontradas, 14 fueron imputadas mediante corroboración con demás medios oficiales y 23 fueron perdidas por Listwise Deletion debido a un MCaR.

Resultados de la exploración de datos

Análisis Univariado

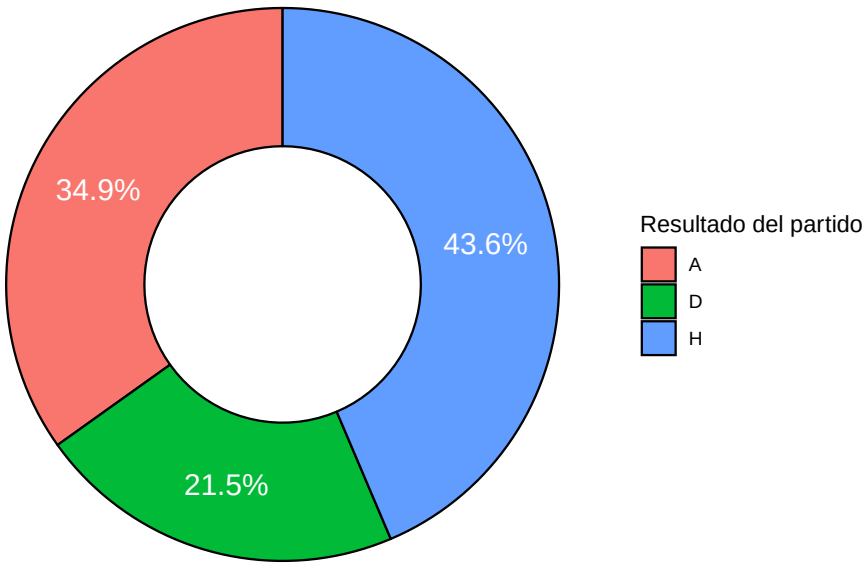
Goles del local y goles del visitante al terminar el partido





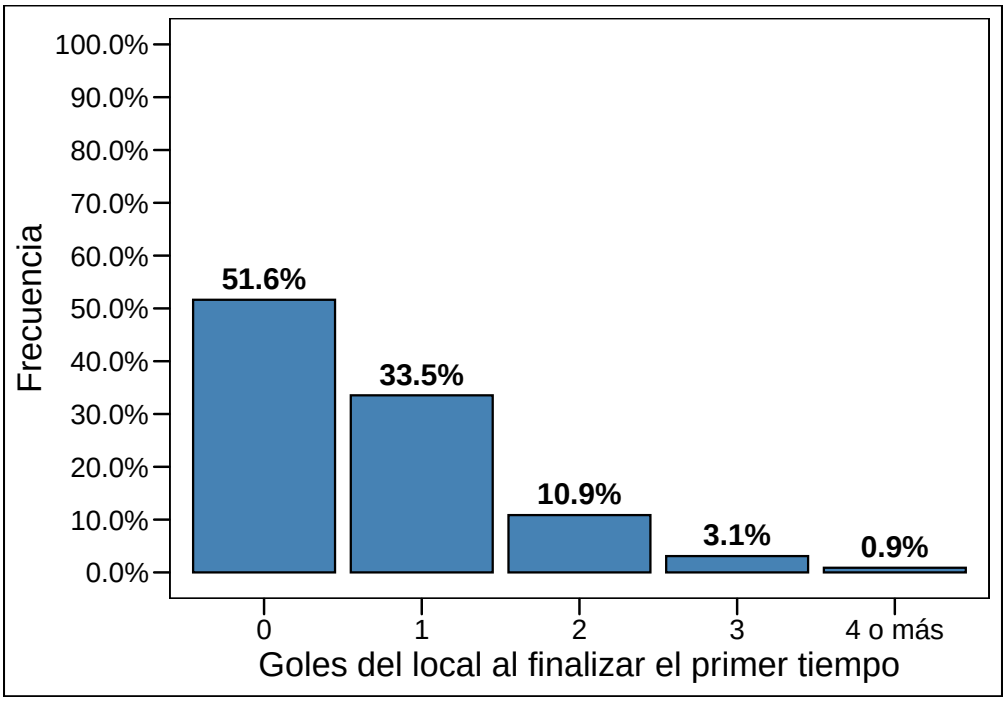
Tanto para los equipos locales y visitantes, la mayoría hicieron un gol al finalizar el partido y la minoría hicieron cuatro o más goles. Por otra parte, los visitantes tienden a quedar más con un marcador de 0 goles.

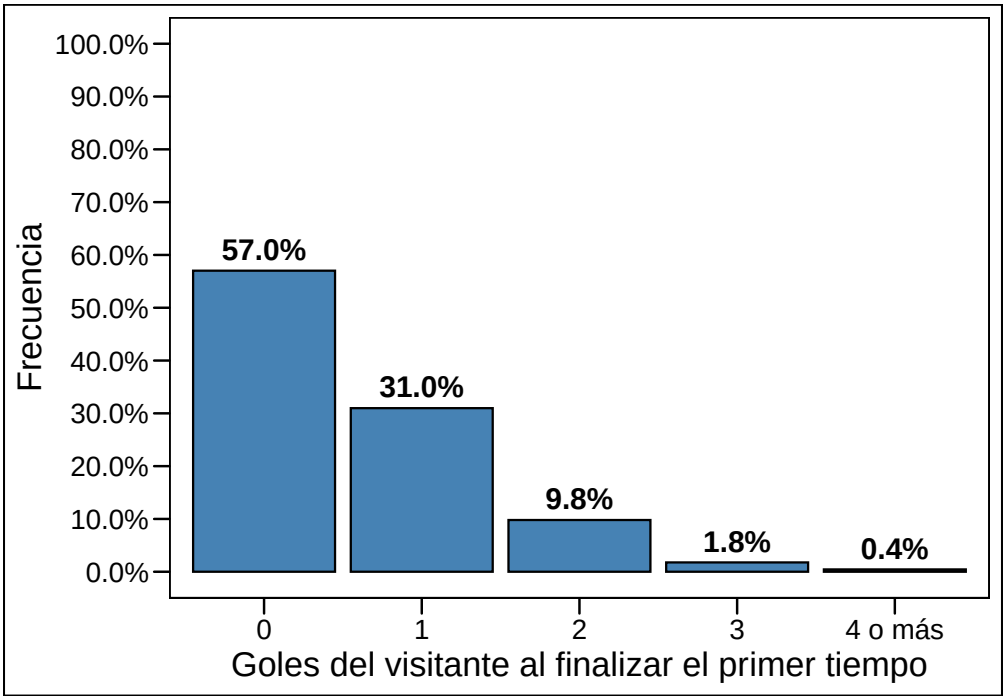
Marcardo final del partido Cabe resaltar que **A** es la victoria del visitante, **D** representa el empate y **H** es la vicoria del local.



De este gráfico podemos observar que los equipos que juegan de local tienden a ganar más partidos que los que juegan de visitante.

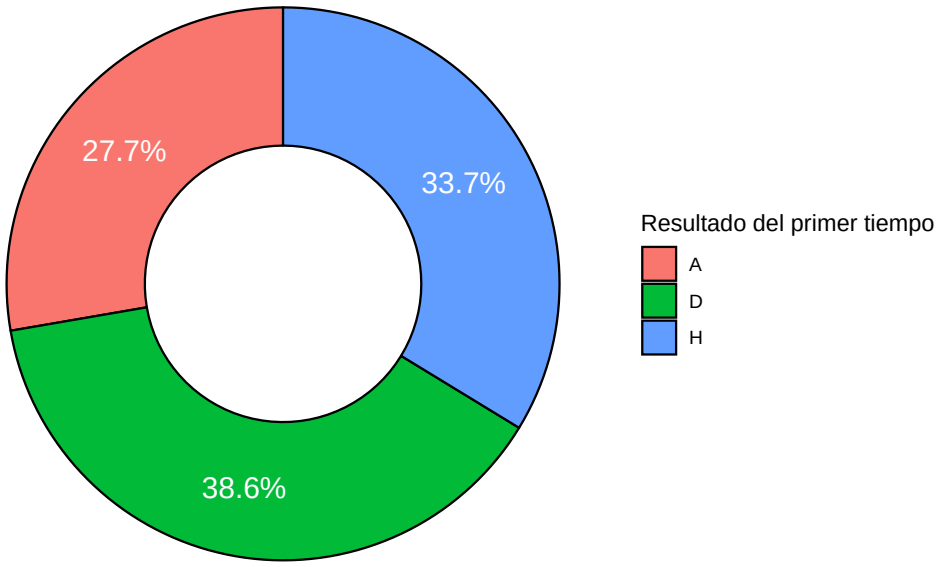
Goles del local y visitante al medio tiempo





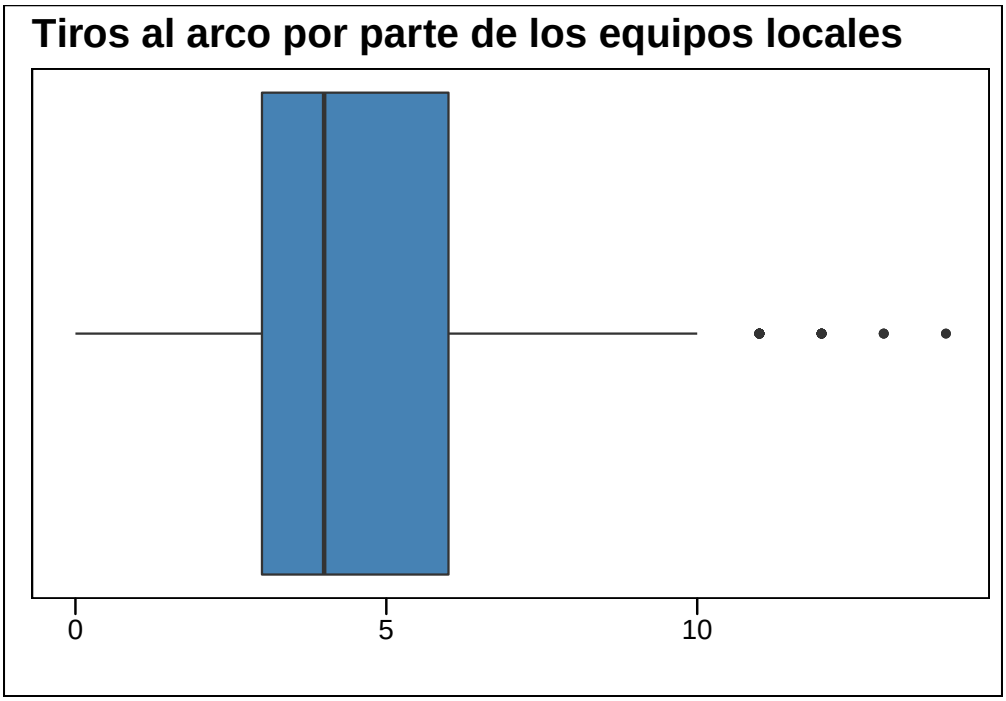
De acuerdo con ambas gráficas, tanto para el local como para el visitante, la mayoría terminan el primer tiempo con cero goles y la minoría con cuatro más goles.

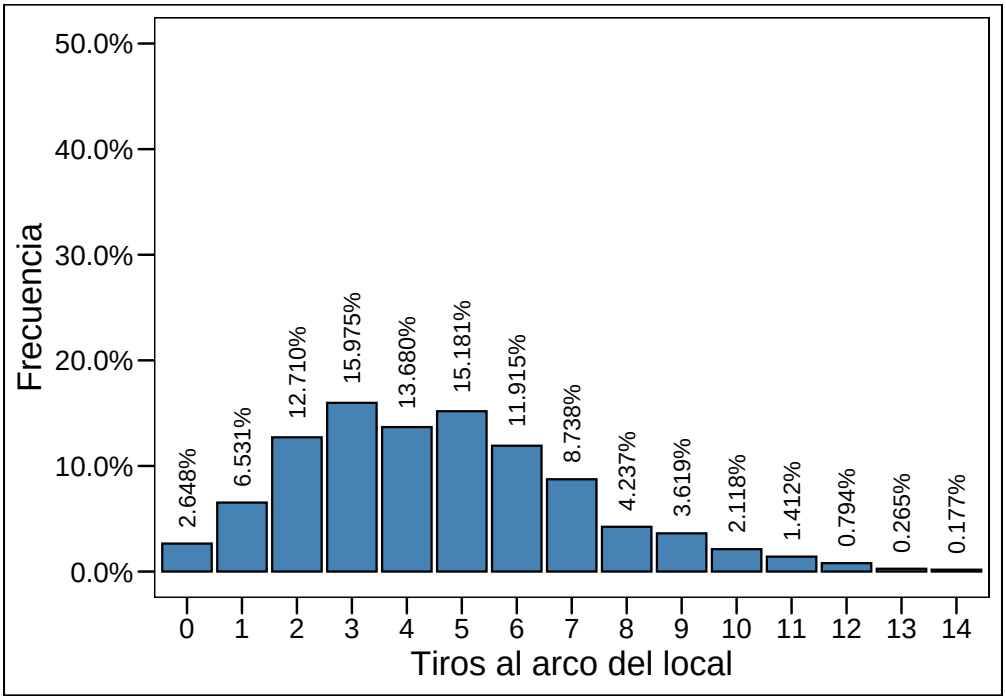
Marcador del medio tiempo



De este gráfico podemos observar que al finalizar el primer tiempo, la mayoría de los equipos quedan en empate y que la minoría resulta en una ventaja para los equipos visitantes.

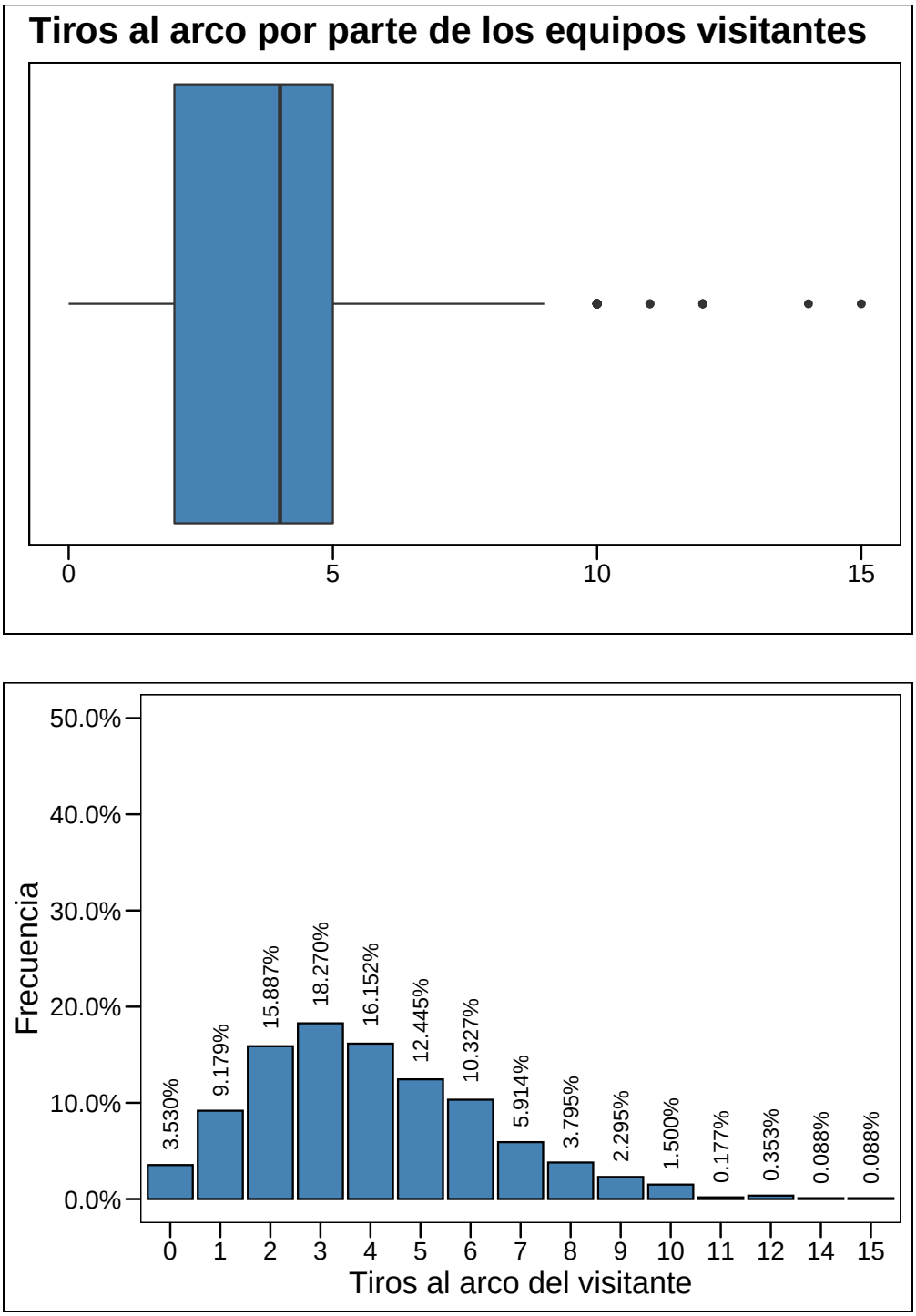
Tiros al arco del local





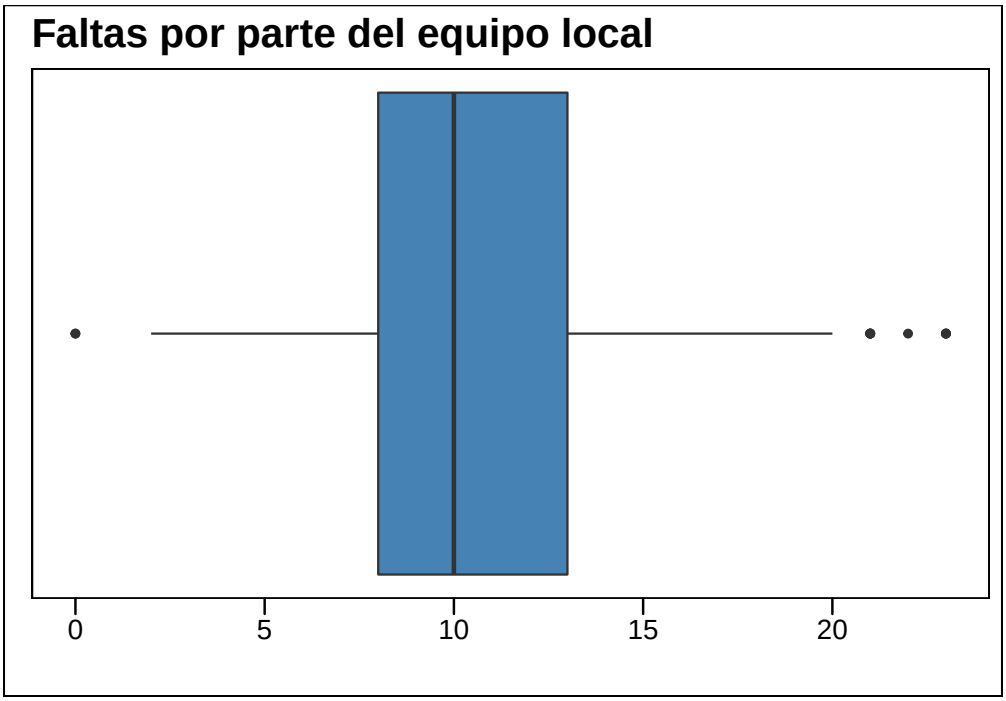
Apartir del diagrama de cajas, se puede observar que los equipos locales tienden a realizar 4.6 tiros al arco por partido. Además, según el diagrama de barras, los tiros al arco poseen una distribución asimétrica positiva.

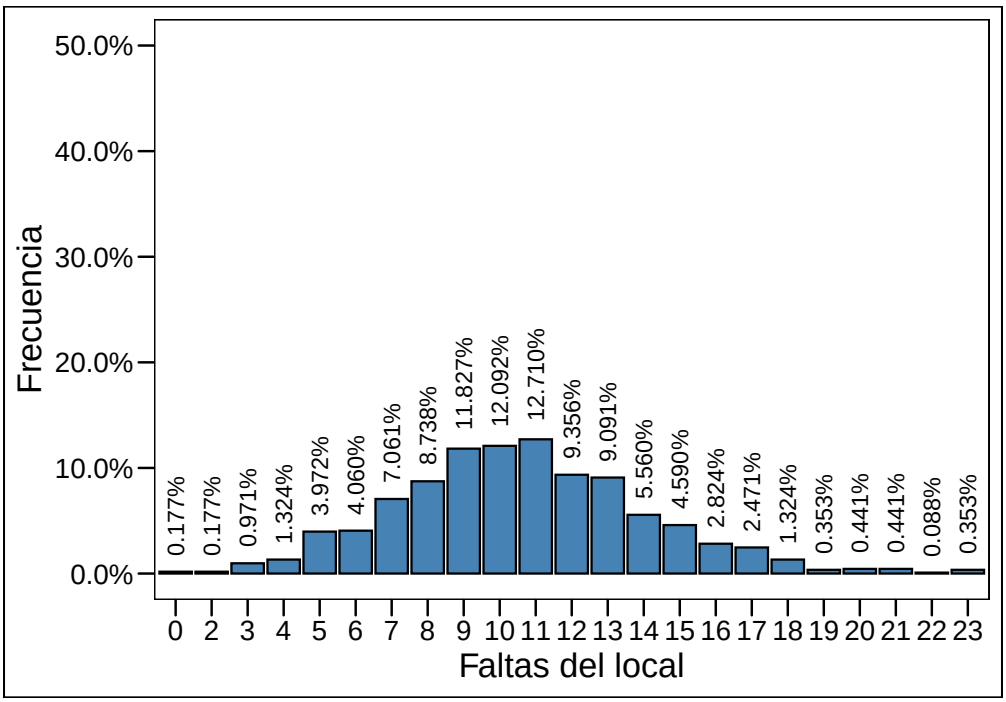
Tiros al arco del visitante



Apartir del diagrama de cajas, se puede observar que los equipos visitantes tienden a realizar 4 tiros al arco por partido. Además, según el diagrama de barras, los tiros al arco poseen una distribución asimétrica positiva.

Faltas del local

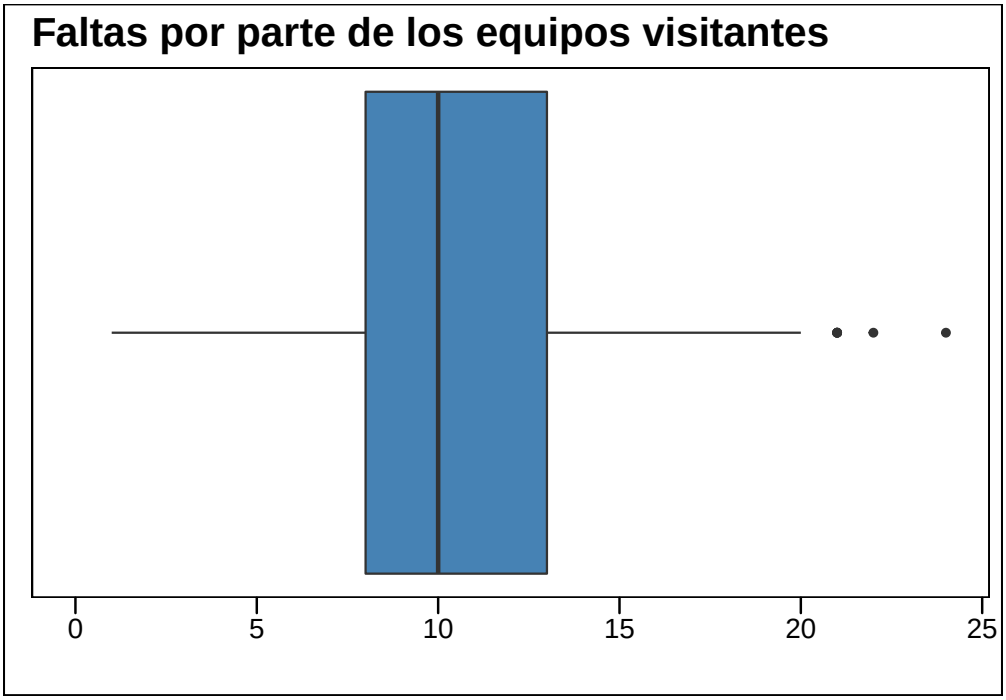


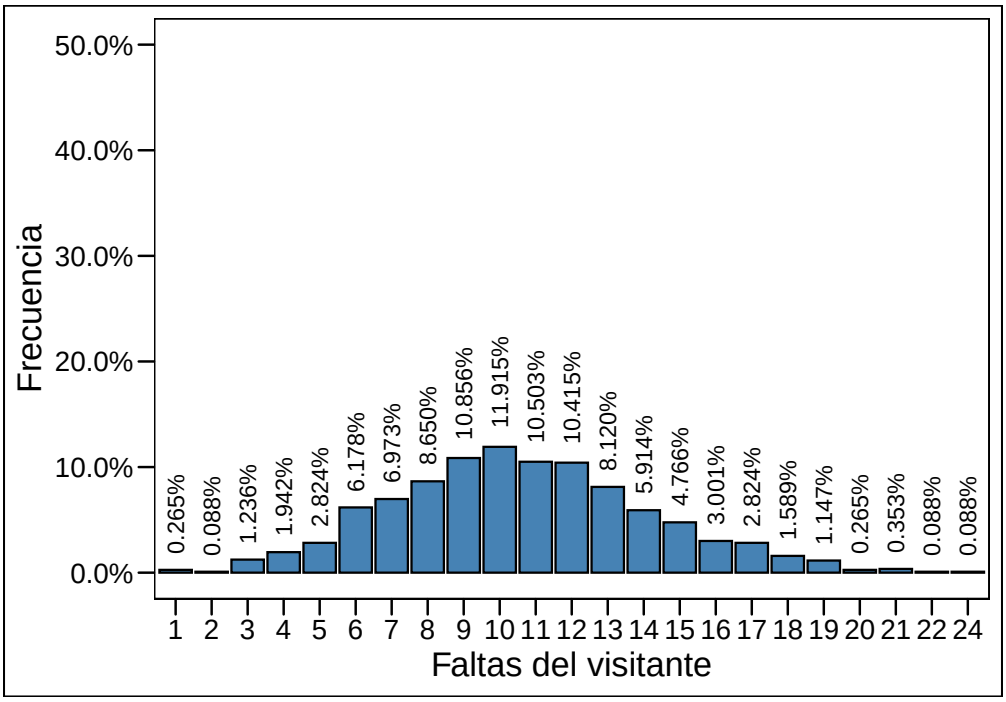


Teniendo en cuenta el diagrama de cajas, se puede ver que los equipos locales cometen aproximadamente 10.6 faltas por juego. Por otro lado, de acuerdo con la gráfica de barras, se muestra la frecuencia de faltas de los jugadores locales, por lo que es probable que los puntajes se rigen bajo una distribución normal, para asegurarse, al realizar la prueba de

normalidad Anderson-Darling, no se puede rechazar la hipótesis nula y concluimos que **no se tiene evidencia suficiente para decir que el puntaje global sigue una distribución normal** debido a que el valor p es menor que nuestro alfa (0.05).

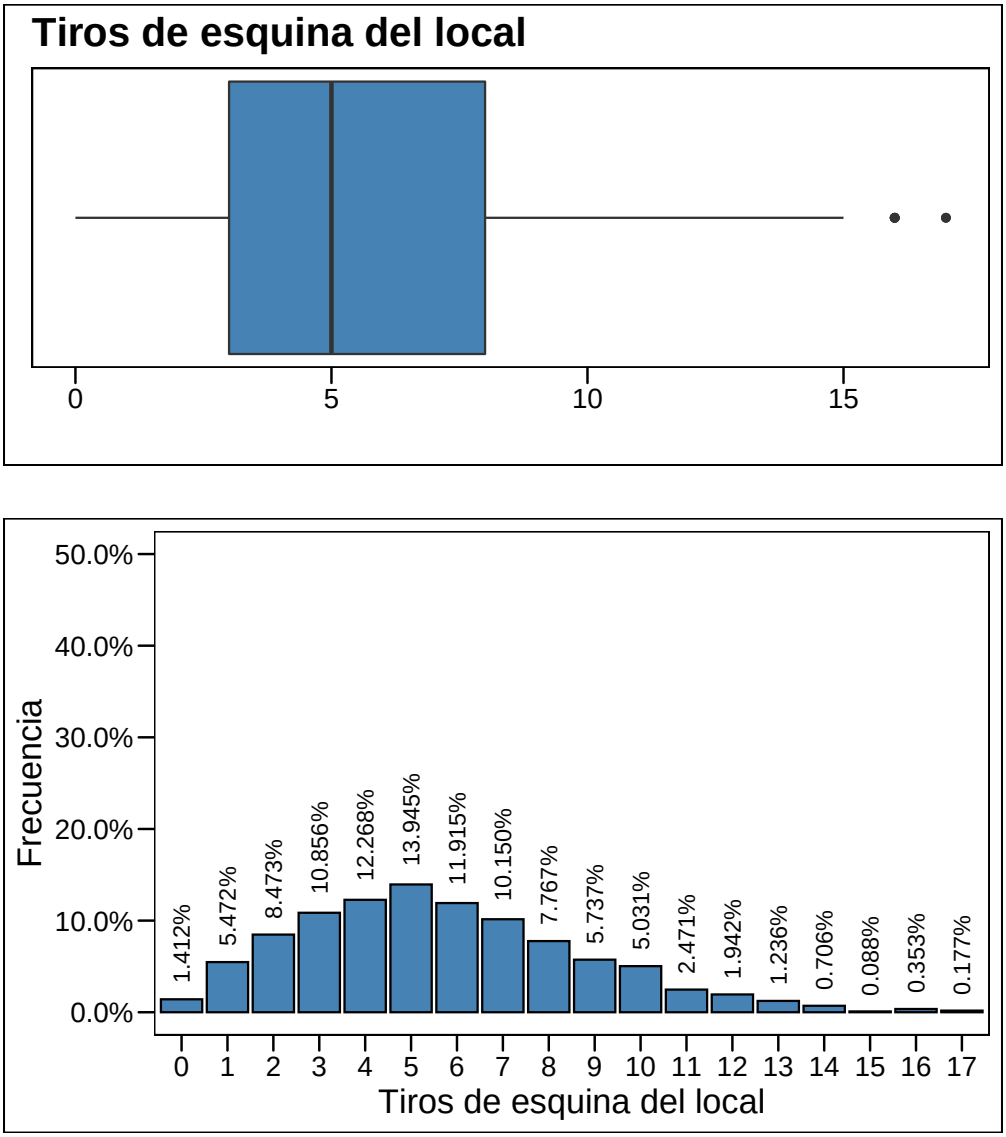
Faltas del visitante





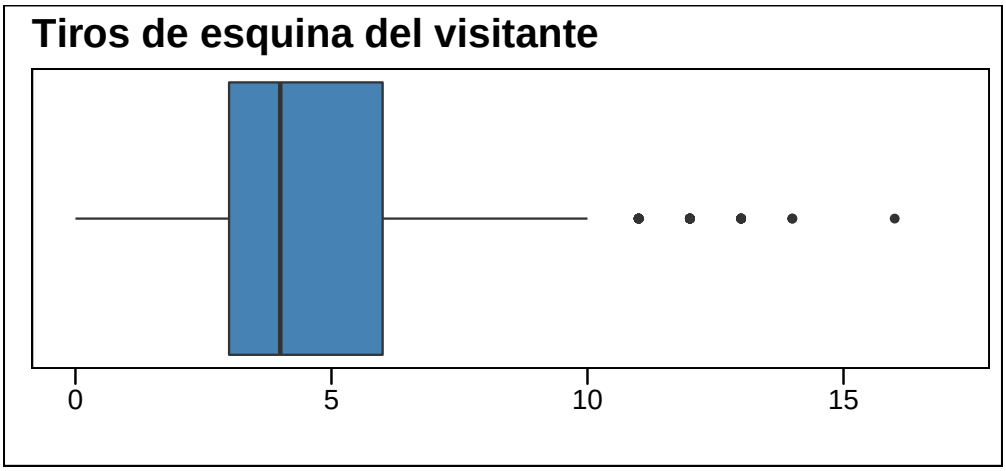
Teniendo en cuenta el diagrama de cajas, se puede ver que los equipos visitantes cometen aproximadamente 10.6 faltas por juego. Por otro lado, de acuerdo con la gráfica de barras, se muestra la frecuencia de faltas de los jugadores visitantes, por lo que es probable que los puntajes se rigen bajo una distribución normal, para asegurarse, al realizar la prueba de normalidad Anderson-Darling, no se puede rechazar la hipótesis nula y concluimos que **no se tiene evidencia suficiente para decir que el puntaje global sigue una distribución normal** debido a que el valor p es menor que nuestro alfa (0.05).

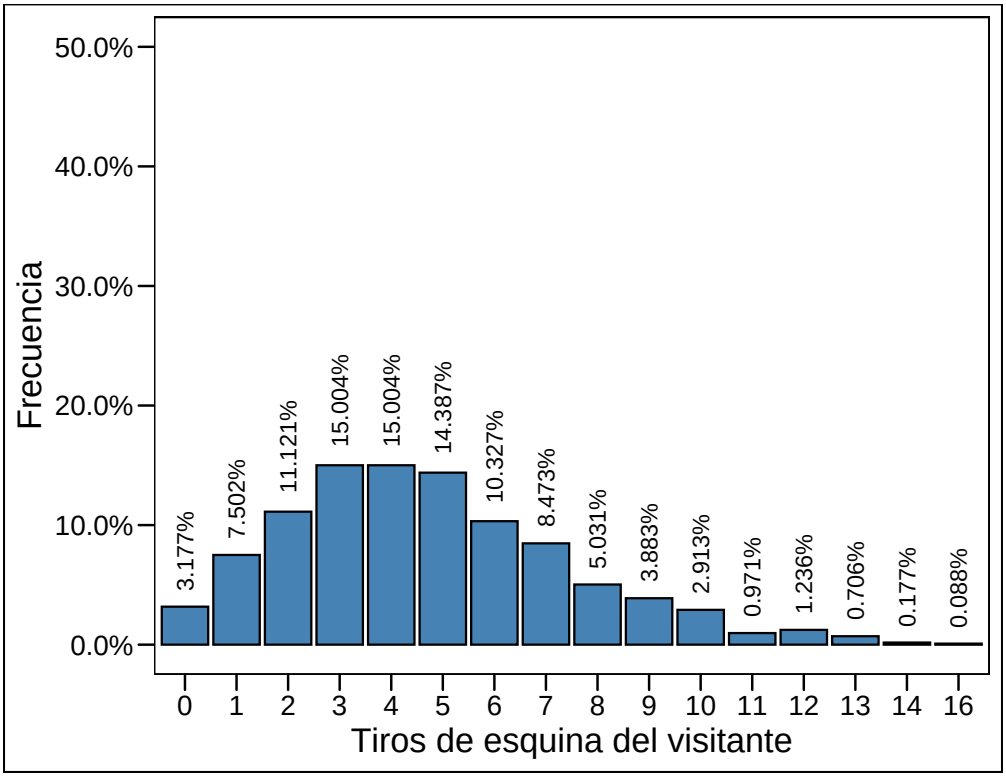
Tiros de esquina del local



Apartir del diagrama de cajas, se puede observar que los equipos locales tienden a realizar 5.6 tiros de esquina por partido. Además, según el diagrama de barras, los tiros de esquina de los locales poseen una distribución asimétrica positiva.

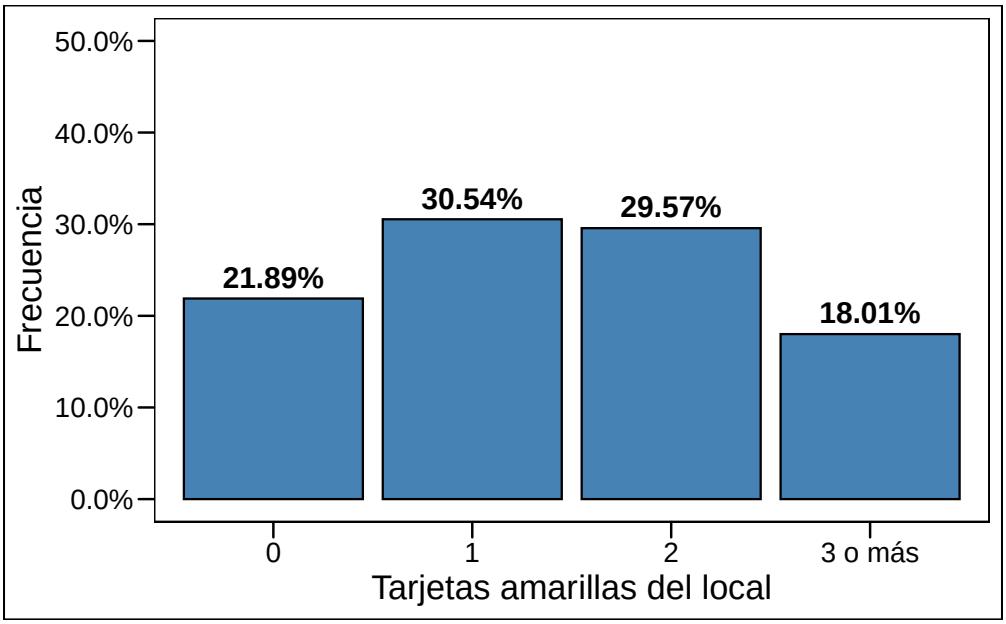
Tiros de esquina del visitante

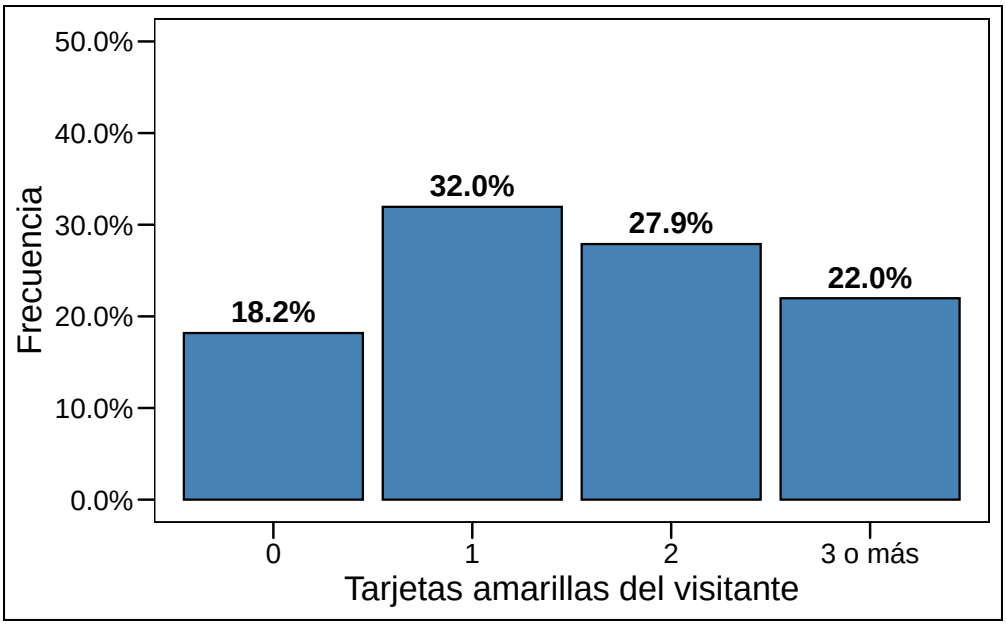




Apartir del diagrama de cajas, se puede observar que los equipos visitantes tienden a realizar 4.7 tiros de esquina por partido. Además, según el diagrama de barras, los tiros de esquina de los visitantes poseen una distribución asimétrica positiva.

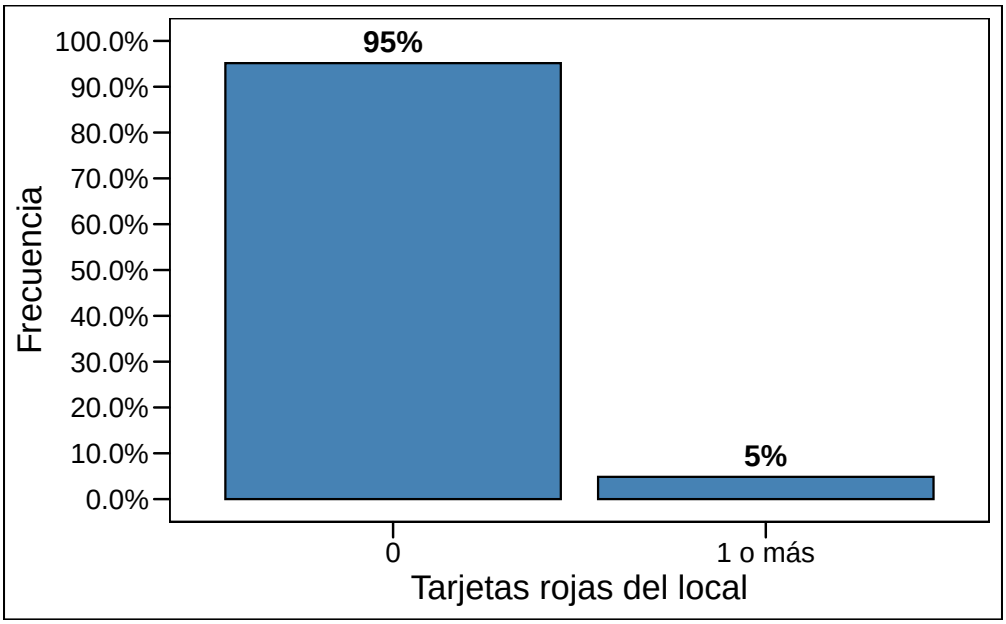
Tarjetas amarillas del local y del visitante

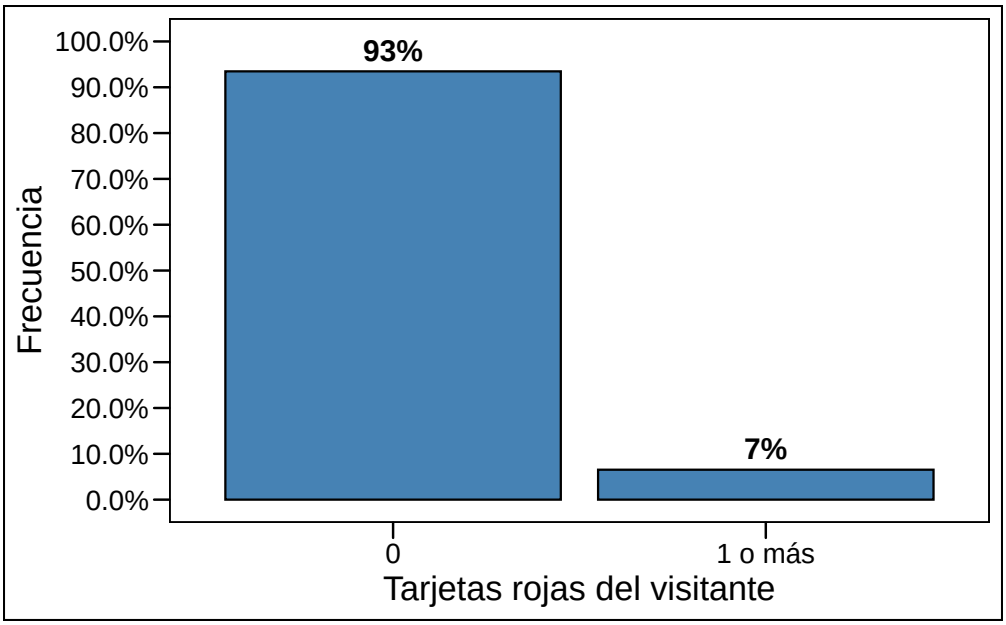




Según el gráfico de barras de las frecuencias de tarjetas amarillas, tanto para equipos locales como visitantes, es mucho más frecuente que haya una tarjeta amarilla. Sin embargo, para los locales es menos común que se cometan tres o más faltas, en cambio para los visitantes, es menos común que no se cometa ninguna falta.

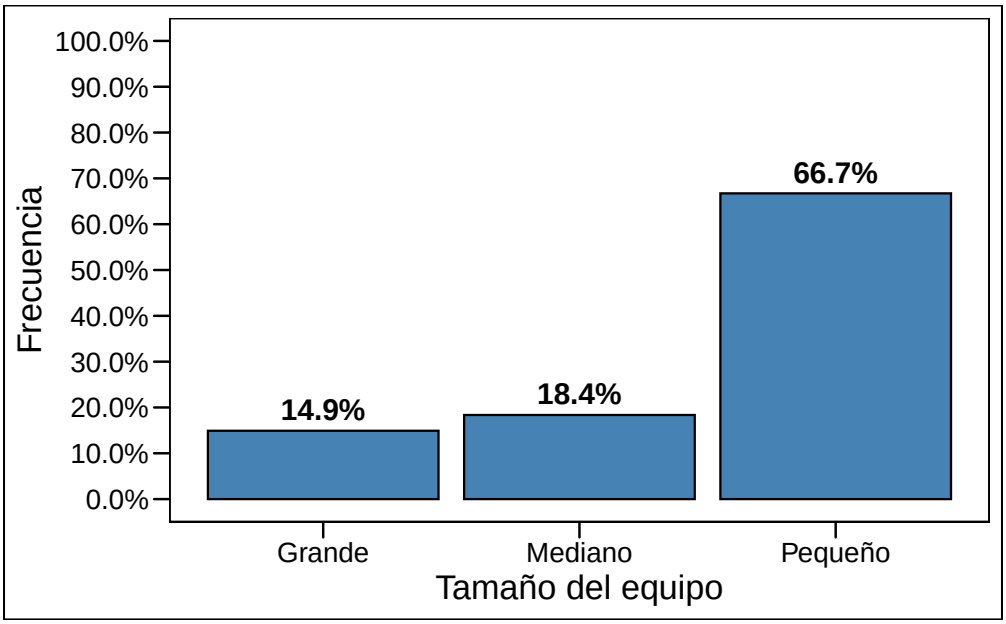
Tarjetas rojas de los equipos locales y visitantes





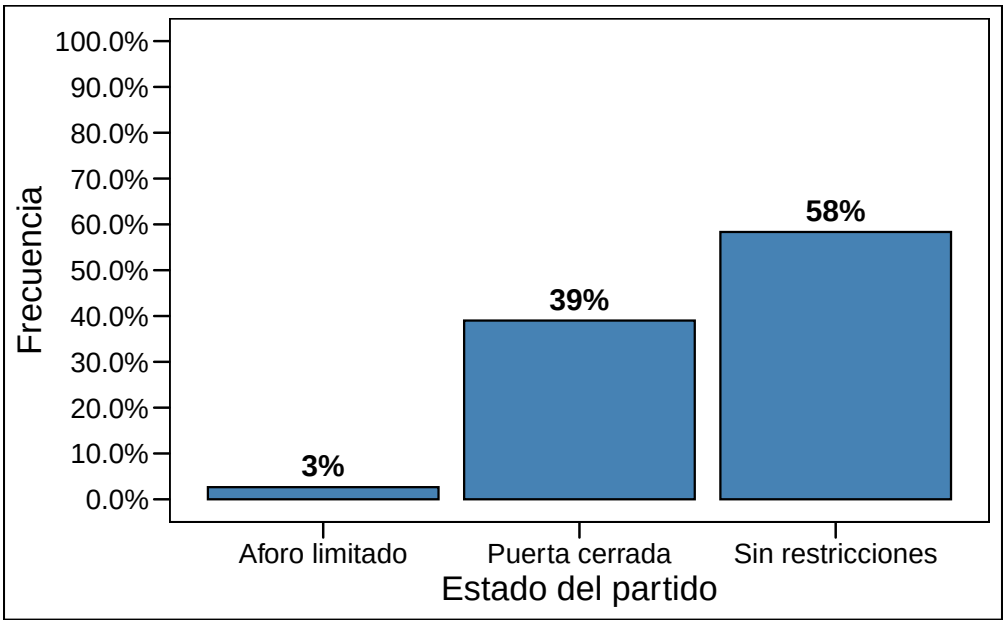
Para ambos equipos, local y visitante, la gran mayoría no cometen tarjetas rojas. Sin embargo, los equipos visitantes, son algo más propensos a cometer una o más tarjetas rojas.

Tamaño del equipo



De acuerdo con la gráfica de barras, la mayoría de los equipos son equipos pequeños y la minoría son los equipos grandes.

Pandemia



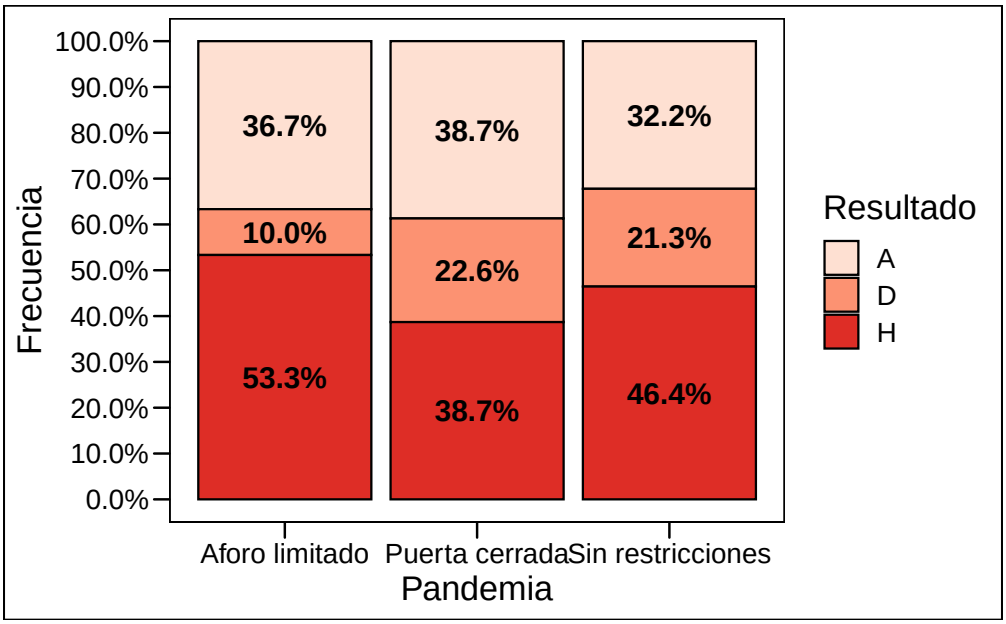
Según la gráfica, la mayoría de los partidos se jugaron sin restricciones y la minoría se jugaron con aforo limitado.

Análisis Bivariado

Para este caso, inicialmente se escogerá la variable dependiente que son todas aquellas que puedan medir el rendimiento de los equipos locales: **FTR, FTHG, HTHG, HTR, HST, HF, HC y AY** y las relacionaremos con las variables explicativas: **Pandemic** y **TeamSize** con el propósito de encontrar una solución al objetivo del proyecto de investigación.

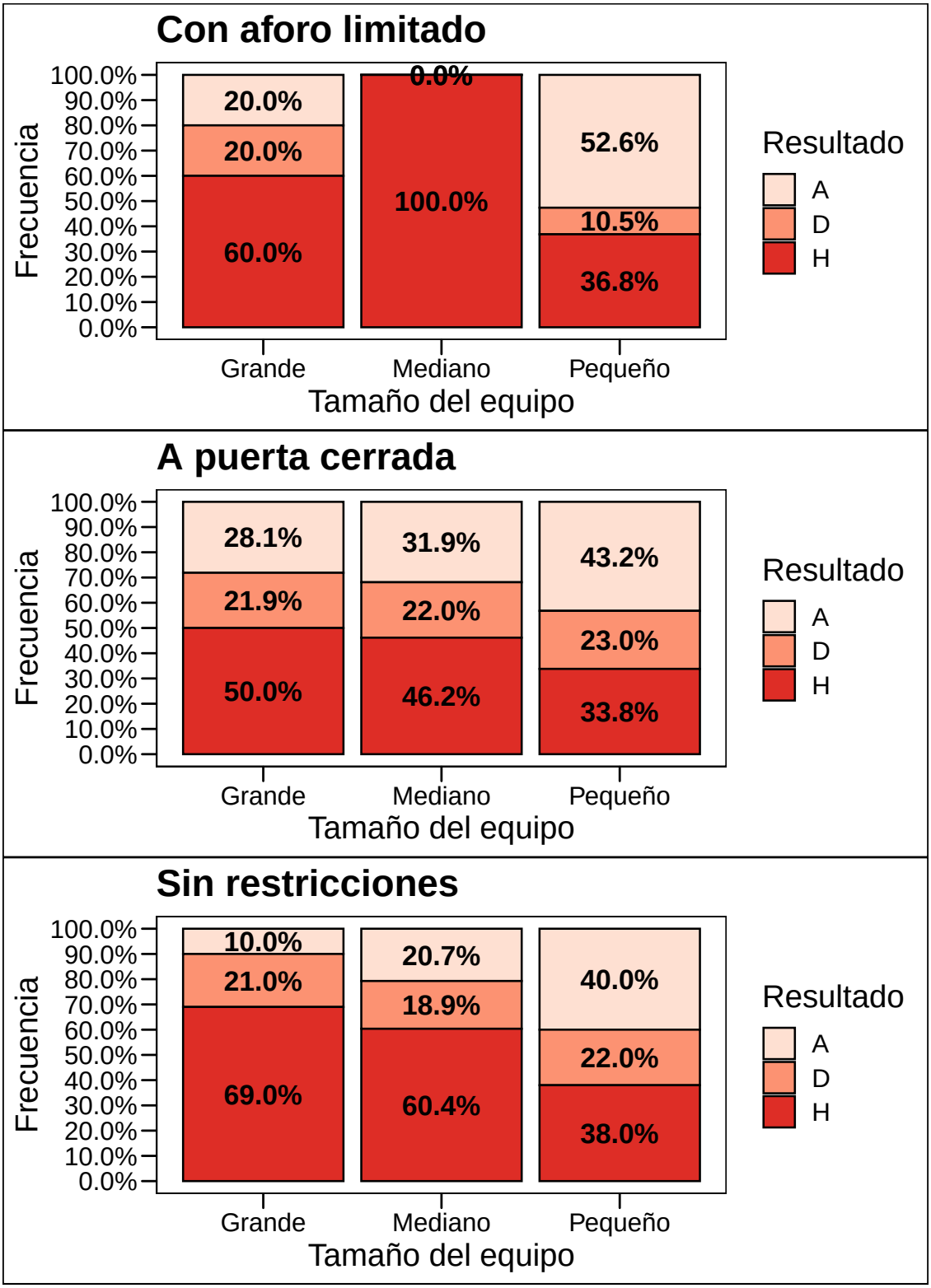
NOTA: Debido a que la muestra de partidos con aforo limitado es muy pequeña (30 datos) en comparación con puerta cerrada y sin restricciones, puede que la toma de decisiones se vea afectada y por ende, los resultados para esta variable no sean del todo representativos, sin embargo, se mantienen para revisar la tendencia de desempeño de los equipos a medida que sigan jugando más partidos en las temporadas.

FTR y Pandemic



Se puede observar que las victorias del equipo local disminuyen al cerrarse los estadios, por otro lado, cuando se reactiva la entrada del público, estos equipos locales subieron su tasa de victorias significativamente.

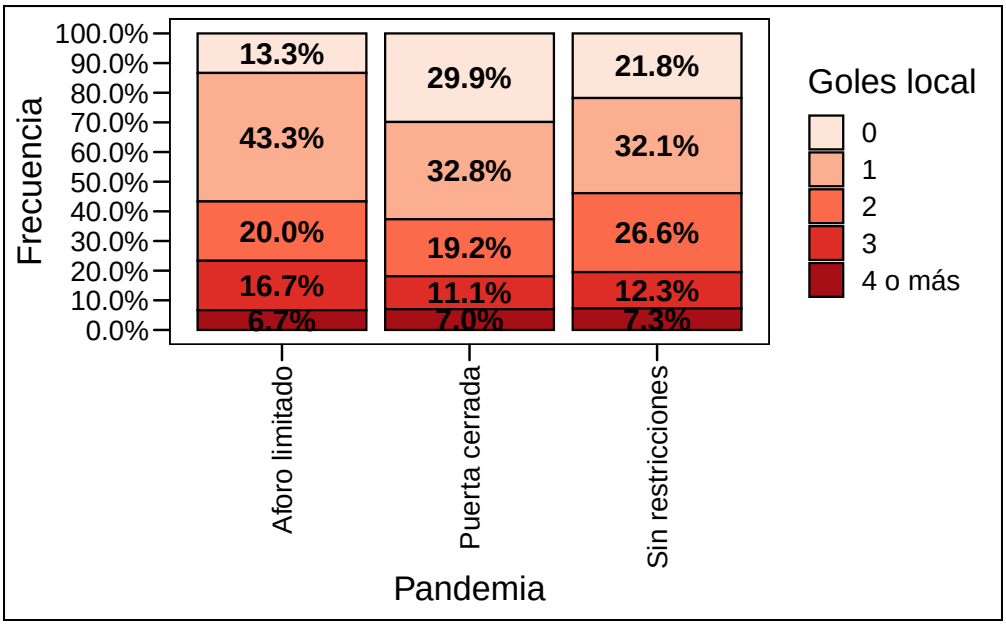
FTR vs TeamSize



La tasa de victoria de los equipos locales medianos al jugar con aforo limitado se ve aumentada drásticamente, por otra parte, al jugarse con puerta cerrada o sin restricciones, se

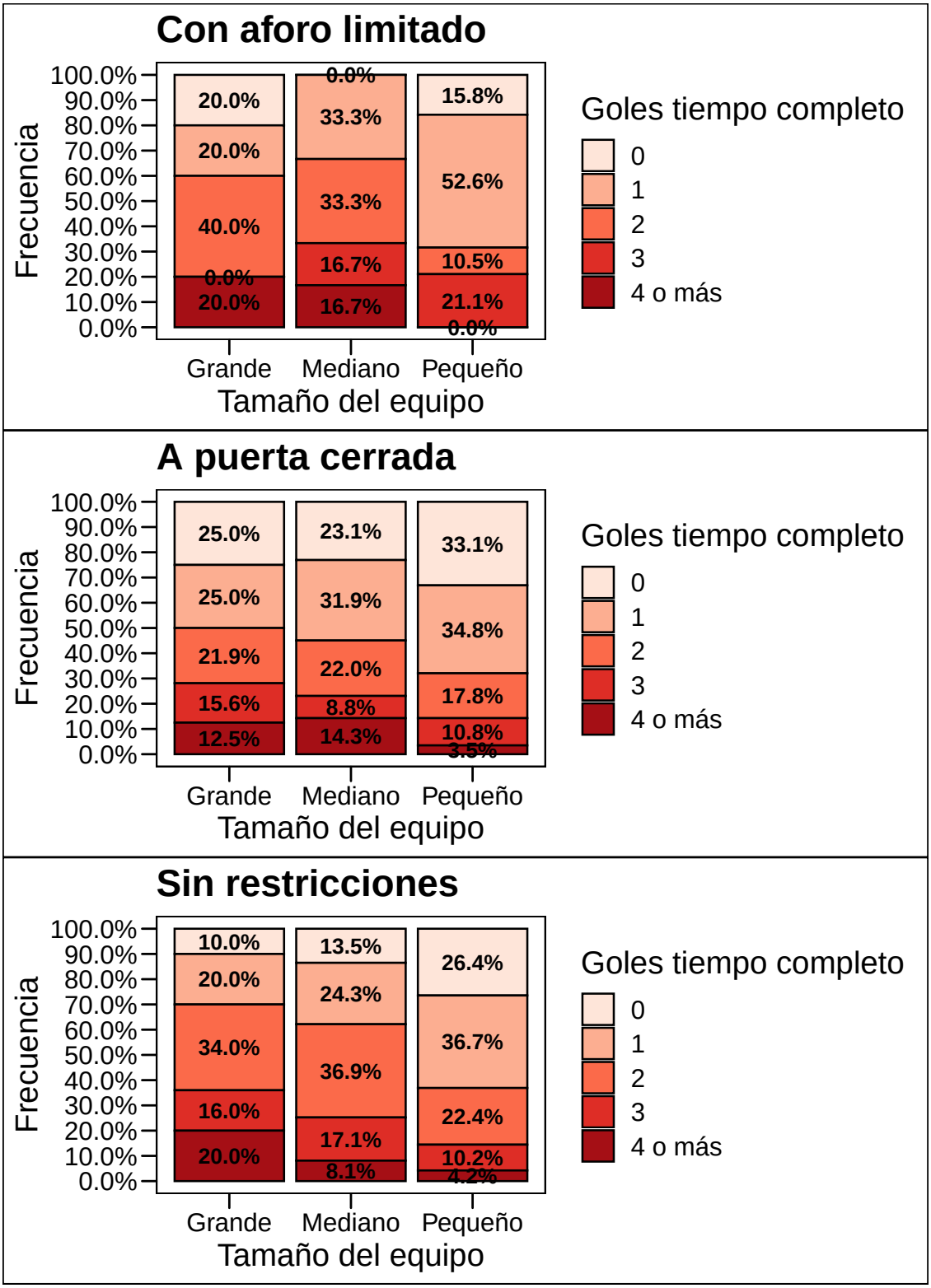
puede observar que se mantiene la proporción de mientras más grande el equipo local, más victorias obtendrá.

FTHG y Pandemic



Se puede observar que la cantidad de goles que anotaron los equipos locales se vió afectada al cerrar los estadios, además, al reabrir los estadios con aforos limitados, se vió un incremento considerable de goles, esto al parecer debido al impacto que tiene el público en los jugadores.

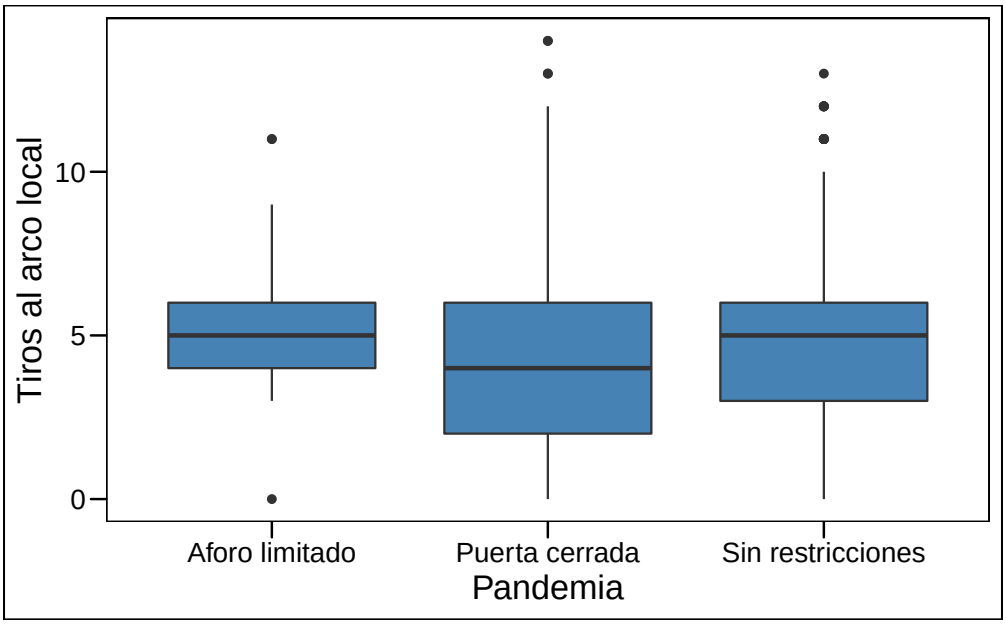
FTHG vs TeamSize vs Pandemic



Se puede decir que, cuando no había restricciones, los equipos grandes anotaban más goles y los pequeños menos, sin embargo, al cerrar los estadios, la cantidad de goles para cualquier

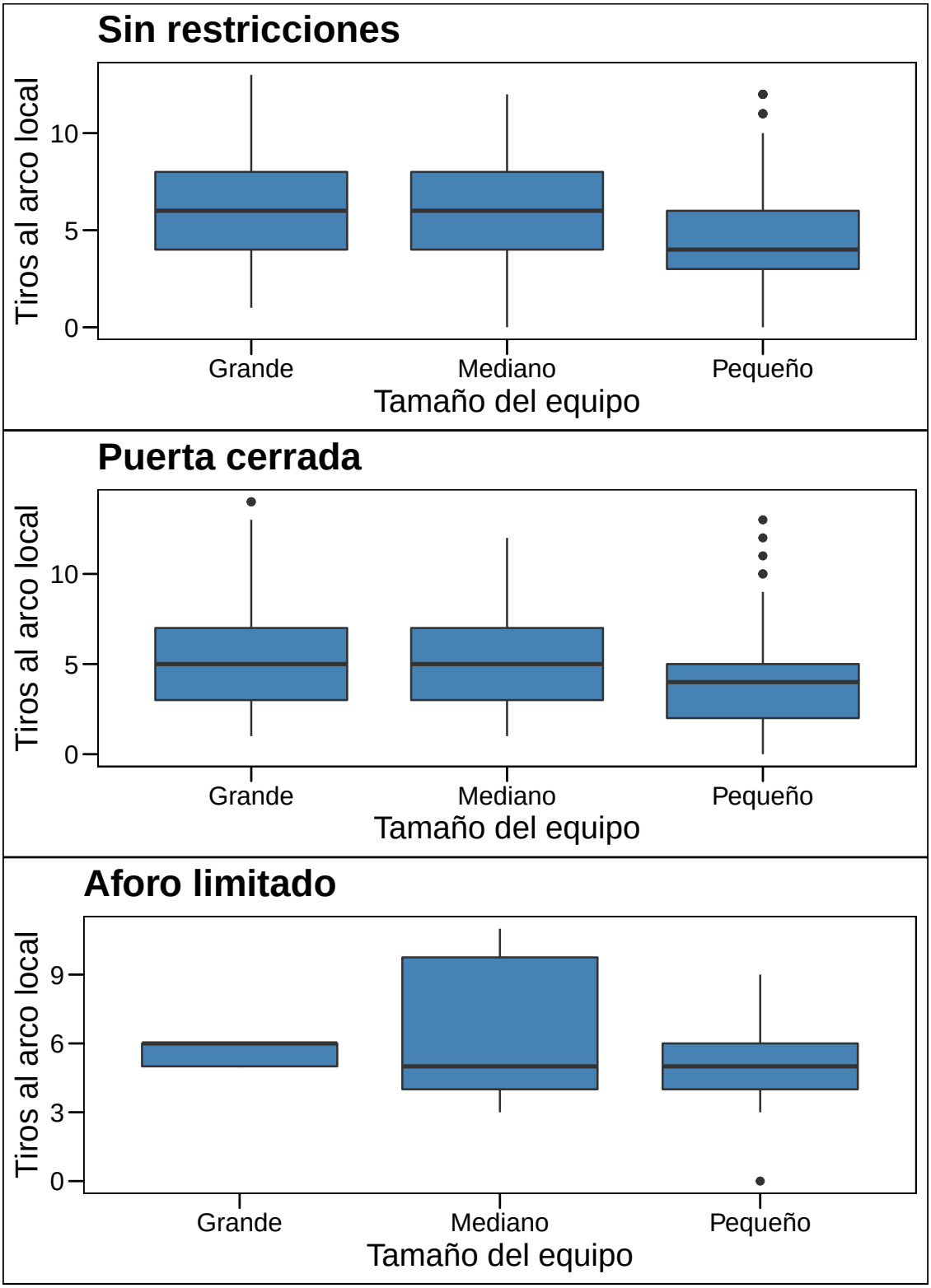
equipo se vio afectada en una disminución de goles totales, los equipos pequeños se vieron menos comprometidos. Se supone que la hinchada de los equipos pequeños es mucho menor que los demás, por ende no fueron alterados tanto su cantidad de goles al cerrar sus estadios. Además, al reabrir los estadios con aforo limitado, todos los equipos aumentaron su cantidad de goles considerablemente.

HST vs Pandemic



Teniendo en cuenta la gráfica, los tiros al arco de los equipos locales se vieron reducidos al cerrar los estadios, no obstante, al reabrirlos se nota una mejoría considerable con respecto a los partidos que jugaron a puerta cerrada y sin restricciones.

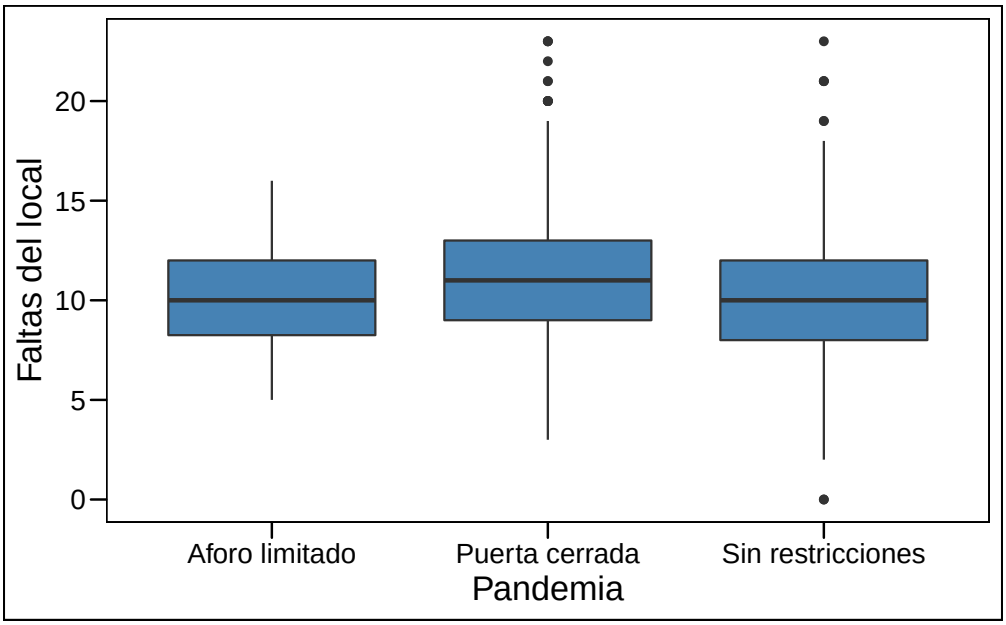
HST vs TeamSize vs Pandemic



Se observa que, sin restricciones, los equipos grandes hacían más tiros al arco y los pequeños menos, sin embargo, al cerrar los estadios, todos se vieron afectados en la misma proporción.

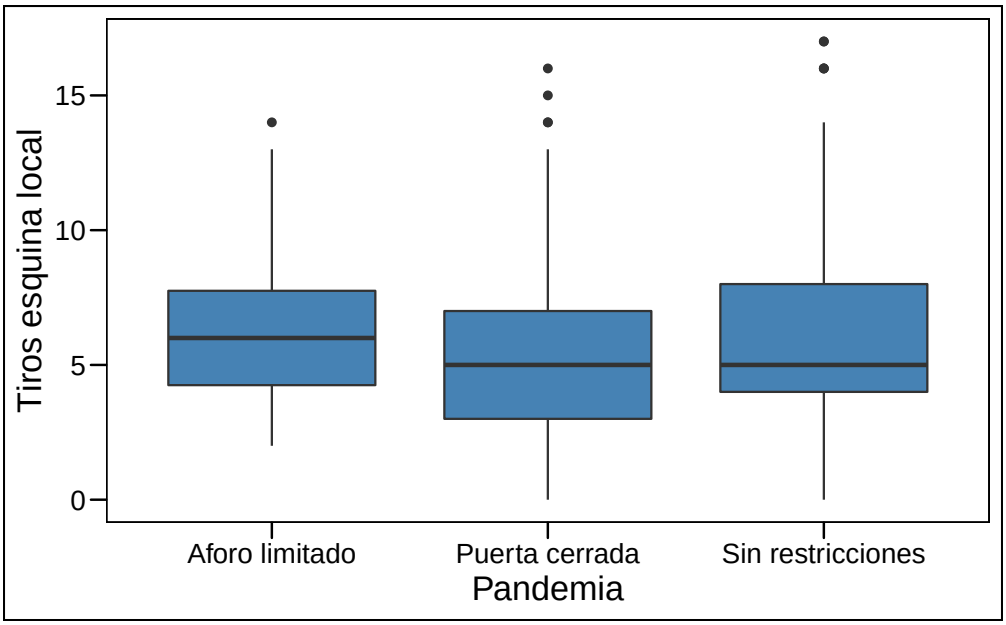
Al abrir nuevamente los estadios, todos los equipos subieron la cantidad de tiros al arco, sobre todo los equipos medianos.

HF vs Pandemic



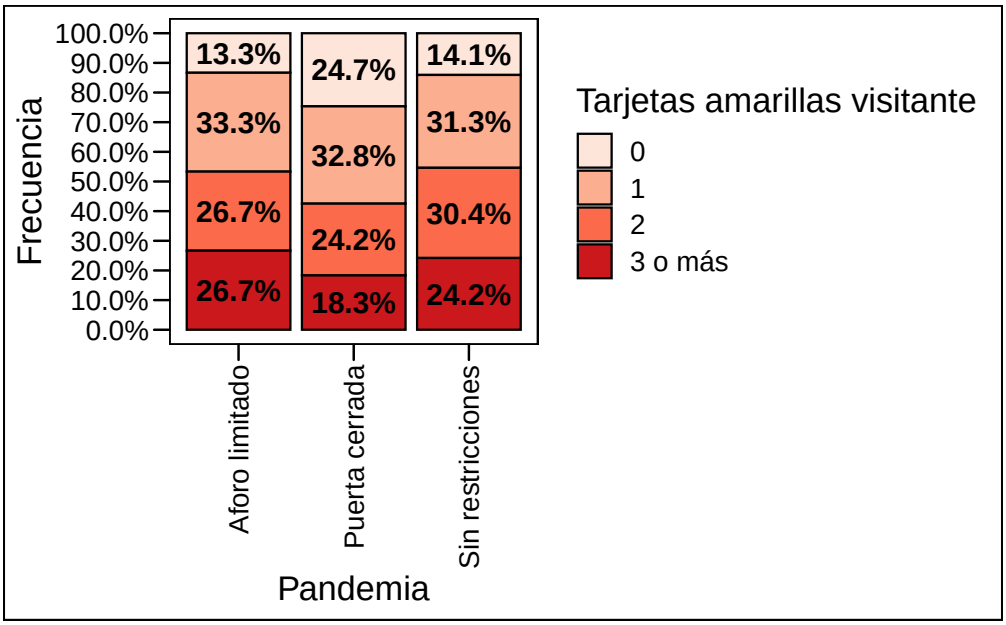
Apartir del diagrama de cajas, se observa que la cantidad de faltas aumentó cuando cerraron los estadios y volvió a su estado inicial cuando reabrieron los estadios.

HC vs Pandemic



Observamos que hubo una disminución de tiros de esquina al cerrar los estadios y un aumento levemente al reabrirlos con aforo limitado con respecto cuando no habían restricciones.

AY vs Pandemic



Se observa que la cantidad de tarjetas amarillas de los equipos visitantes tuvo una disminución cuando se cerraron los estadios y cuando se reabrieron volvieron a la normalidad. Esto pudo haber ocurrido debido a la constante presión de la hinchada del local al arbitro para que éste los sancionara con más frecuencia.

Conclusiones y recomendaciones

Lo desarrollado en este informe es un proyecto de investigación sobre el análisis exploratorio de los partidos de la Premier League, describiéndolos con el propósito de encontrar relación entre el rendimiento de los equipos locales y el aforo en los estadios durante la pandemia del Covid-19. Se realizó un proceso de detección y limpieza de anomalías sobre la base de datos suministrada, de las 37 anomalías encontradas, se lograron imputar por medios oficiales 14 de ellas y el resto se procedio a eliminarlas tomándolas como una MCaR. Después, se procedió a realizar un análisis univariado para todas aquellas variables de interés y se procedió a implementar un análisis bivariado, relacionando nuestras variables de explicativas (Pandemia y Tamaño del equipo) con las demás dependietes que lograban medir el rendimiento de los equipos locales.

- **Para el análisis univariado:** La mayoría de los locales de la Premier League, al terminar el partido, hacen un gol y quedan victorios, por otro lado, no hacen ninguno y quedan en empate cuando finaliza el primer tiempo. Además, tienen una media de 4.6 tiros al arco, cometen 10.6 faltas por juego, realizan 5.6 tiros de esquina y obtienen al menos una tarjeta amarilla. Por otra parte, predominan los equipos pequeños y la mayoría de los partidos se jugaron sin restricciones durante las temporadas.
- **Para el análisis bivariado:** Tanto las victorias, los goles, tiros al arco y tiros de esquina de los equipos locales, tuvieron una disminución al verse cerrado los estadios y, para todos los casos, se recuperaron al reabrirse y ver a su hinchada otra vez. Por otra parte, las faltas cometidas por los locales aumentaron cuando no había público y disminuían cuando si. Por último, las tarjetas amarillas obtenidas por los equipos visitantes disminuyeron cuando no existía presencia del público, en especial, la hinchada del local. También, se revisó cómo el tamaño del equipo también reaccionó a todas las variables dependientes teniendo como respuesta que, sin importar el tamaño del equipo, se tuvo una disminución en las victorias y demás variables de rendimiento con el cierre de los estadios, y que considerablemente se notó el cambio cuando se reapertura los estadios aumentando así los resultados de ganadores de partidos por parte del equipo local, así como sus tiros de esquina y al arco y número de goles, además de una disminución en las faltas y tarjetas amarillas.

Teniendo en cuenta todo lo anterior, se puede afirmar que **SI** hubo una afectación de la hinchada en el rendimiento de los equipos locales al cerrarse los estadios independientemente del tamaño del equipo.