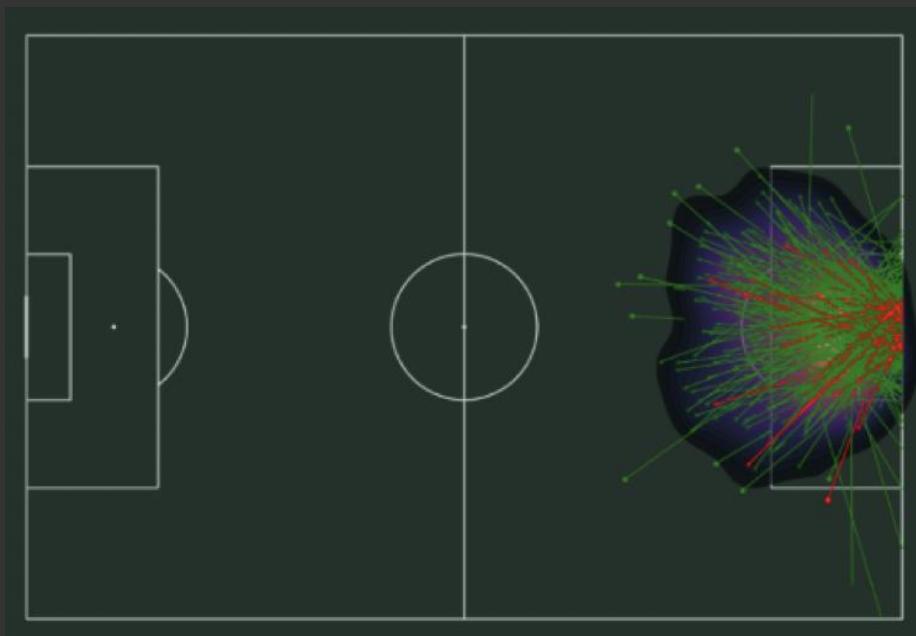


Máster en Big Data y Analytics

Encontrar las jugadas de más valor en un partido de fútbol



Entrega Final – 31/10/2023

Grupo 1

Matias Benjamin Arce Mones Ruiz

Daniel Áspera Dormido

Jose María Áspera Dormido

Miriam Iban Gil

Lluís Vega Roman



Encontrar las jugadas de más valor en un partido de fútbol del Manchester City Women's Football Club en la temporada 2020-2021 de la FA Women's Super League.

Grupo 1

Matias Benjamin Arce Mones Ruiz

Daniel Áspera Dormido

José María Áspera Dormido

Miriam Iban Gil

Lluís Vega Román

Con tutoría de

Felipe Vianna Juarez

Entrega Final

Octubre 2023

RESUMEN EJECUTIVO

El fútbol es un deporte de balón en equipo jugado entre dos conjuntos de 11 jugadores/as cada uno. Es considerado el deporte más popular del mundo, y según una encuesta realizada por la FIFA, existen alrededor de 26 millones de jugadoras en el mundo, algunas de ellas profesionales.

La Primera Guerra Mundial (1914-1918) fue clave en la masificación del fútbol femenino en Inglaterra, sin embargo, al fin de la guerra, la *football association* (FA) no reconoció al fútbol femenino a pesar del éxito de popularidad que alcanzó. Actualmente, el fútbol femenino está en auge y muchos clubes desean llevar al máximo nivel a sus equipos femeninos. El City Football Group Limited administra diferentes clubes de fútbol, entre ellos el Manchester City Women's Football Club (MCWFC) el cual disputa la FA Women's Super League. Esta empresa, está considerada Data Driven y pretende que el MCWFC obtenga las máximas victorias posibles, por lo que nuestro equipo de ciencia de datos, ha desarrollado este proyecto para encontrar las jugadas de más valor en la temporada 2020-2021 de este equipo en la FA Women's Super League.

Con el presente proyecto, pretendemos principalmente acercar al Manchester City Women's Football Club a la victoria de la liga 2023-2024, optimizando las estrategias para la creación de mejores jugadas en el campo. De ello se derivan objetivos secundarios que hacen referencia directamente a los datos como evaluar la calidad de estos, analizarlos para obtener *insights* y modelizar algoritmos de ML para encontrar inteligencia en estos datos. Todo ello, nos va a permitir alcanzar el quinto objetivo que va a ser crear entrenamientos específicos para mejorar el rendimiento del equipo aplicando todo aquel conocimiento que se ha extraído del proyecto.

En cuanto a la metodología usada ha sido CRISP-DM, donde en primer lugar llevamos a cabo un entendimiento del negocio para alinear nuestros objetivos con este. En segundo lugar, efectuamos el entendimiento de los datos del proveedor Statsbomb en su versión Open Data, realizando un data profiling. Seguidamente, hicimos la preparación de los datos y el EDA, lo que permitió la realización de un dashboard. En última instancia, llevar a cabo al modelado y la evaluación de este.

A la finalización del proyecto concluimos que los objetivos planteados fueron logrados, aunque la modelización del algoritmo no fue tan satisfactoria como esperábamos dado el carácter aleatorio de los datos.

ÍNDICE

1.	Introducción.....	5
2.	Definición del proyecto	7
2.1	Contexto y alcance	7
2.2	Organización y madurez de la empresa	8
2.3	Alcance del proyecto	10
2.4	Objetivos	12
2.5	Propuesta de valor	13
3.	Planificación	15
3.1	Roles y equipo.....	15
3.2	Tareas.....	16
3.3	Calendario.....	18
4.	Viabilidad del proyecto.....	19
4.1	Técnica.....	19
4.2	Económica	19
4.3	Valoración económica	22
4.4	Conclusión de viabilidad.....	22
5.	Desarrollo del proyecto.....	25
5.1	Metodología de trabajo.....	25
5.2	Arquitectura	26
5.3	Preparación de los datos y analítica	29
5.3.1	Descripción de la fuente de información.....	29
5.3.2	Data Quality	32
5.3.3	Reglas de negocio	43
5.3.4	Data profiling.....	46
5.3.5	EDA	49
5.3.6	Reconocimiento de patrones.....	53
5.4	Mapping de datos.....	65
5.5	Output del proyecto	67
6.	Casos de uso.....	75
6.1	Jugadora goleadora.....	75
6.2	Jugadora chutadora.....	76
6.3	Partido perdido	77
6.4	Rival directo.....	77
6.5	Técnica de disparo	77
6.6	Posiciones relevantes en el chute	78

6.7	Local vs Visitante	78
6.8	Penaltis	79
7.	Ampliación análisis (comparativa).....	80
7.1	Chelsea.....	81
7.2	Bristol City	81
7.3	Casos de uso	82
7.3.1	Jugadoras goleadoras y chuteadoras.....	82
7.3.2	Partidos contra el Manchester City.....	83
8.	Machine Learning	84
9.	Conclusiones	89
10.	Anexos	91
10.1	Anexo 1 (Crisp-DM).....	91
10.2	Anexo 2 (plantilla).....	92
10.3	Anexo 3 (goleadoras)	93
10.4	Anexo 4 (chutadora)	94
10.5	Anexo 5 (partido perdido)	94
10.6	Anexo 6 (rival directo).....	95
10.7	Anexo 7 (técnica)	95
10.8	Anexo 8 (posiciones)	96
10.9	Anexo 9 (local vs visitante)	97
10.10	Anexo 10 (penaltis).....	98
10.11	Anexo 11 (Chelsea y Bristol)	98
11.	Bibliografía	100

1. Introducción

El City Football Group Limited (CFG) es una sociedad de cartera que administra clubes de fútbol, donde el 78% es propiedad mayoritaria del Abu Dhabi United Group, el 10% de la firma estadounidense Silver Lake y el 12% de las firmas chinas China Media Capital y CITIC Group. Posee participaciones en clubes de Estados Unidos, Australia, India, Italia, Japón, España, Uruguay, China, Bélgica y Francia, aunque destaca el Manchester City Football club, el cual cuenta también con su equipo femenino Manchester City Women's Football Club el cual disputa la liga de la FA Women's Super League, donde destaca como uno de los mejores equipos.

Aunque el Manchester City femenino había existido desde 1988, existió únicamente como afiliado del club hasta agosto de 2012, y fue cuatro años del control del equipo masculino cuando el Abu Dhabi United Group, eligió traer al equipo femenino bajo su dominio, formando una asociación con el club que lo convertiría en un equipo del MC. El apoyo que recibió por parte del propietario mayoritario, fue retribuido rápidamente, ganando su primer gran trofeo en la FA Women's League Cup 2014.

A pesar de que el equipo femenino tiene varios triunfos importantes, lleva desde el año 2016 sin ganar la liga nacional, sumándose a esto que en la temporada 2020-2021 no obtuvo ningún trofeo, por lo que el CFG, una organización Data Driven que confía mucho en los datos, pretende identificar las jugadas de más valor de este equipo en la temporada 2020-2021 de la FA Women's Super League.

Para el desarrollo del proyecto hemos seguido la metodología CRISP-DM siguiendo las fases mencionadas pertinentes de dicha metodología, divididas en 3 entregas (principios de abril, principios de junio y principios de setiembre del año 2023). Esta metodología nos ha permitido iterar sobre las distintas fases a medida que íbamos desarrollando el proyecto. Las principales tareas llevadas a cabo han sido la propia definición del proyecto con el entendimiento del negocio y de los datos, seguido del diseño de la arquitectura que una vez desplegada nos permitió la realización de la ELT y el aprovisionamiento de la base de datos. En este punto pasamos a evaluar la calidad de datos, para poder extraer la información y utilizarla como conocimiento, y para visualizarlo hicimos un dashboard dinámico que permite visualizar todos los datos de la temporada mencionada. Por último, llevamos a cabo un proceso de modelización de algoritmos de ML para poder llegar a predecir cómo va a ser la finalización de un chute, ya que para nuestro equipo las jugadas de máximo valor en un partido son los chutes, y en su máxima expresión los goles.

Como conclusión, podemos afirmar que al Manchester City femenino tiene unas estadísticas que destacan positivamente en su principal competición, aunque hay distintos aspectos que pueden ser trabajados en los entrenamientos para aumentar la efectividad de los chutes y poder aumentar el número de goles, y sobretodo de victorias, como por ejemplo los disparos a puerta desde fuera del área.

En referencia a la modelización, es realmente complejo predecir cuál va a ser el resultado final de un chute, necesitaríamos muchas otras variables, y aún así, seguiría jugando un papel muy importante el azar.

2. Definición del proyecto

El presente proyecto se basa en analizar diferentes conjuntos de datos posicionales y eventing del **Manchester City Women's Football Club** en la temporada 2020-2021 de la **FA Women's Super League**.

El propósito principal del proyecto, es encontrar las jugadas más valiosas dentro de los distintos partidos disputados, mediante el análisis de datos proporcionados por la empresa Satstbomb (proveedor de datos que recoge acciones de presión a nivel tanto de equipos como de jugadores). Statsbomb, es líder en datos deportivos y muy relevante en el mundo de la analítica futbolística, ofreciendo diferentes productos a los analistas, aunque nosotros vamos a trabajar bajo el Opendata.

2.1 Contexto y alcance

El City Football Group Limited (CFG) es una sociedad de cartera que administra clubes de fútbol. Es propiedad de tres organizaciones; de los cuales el 78% es propiedad mayoritaria del Abu Dhabi United Group, el 10% de la firma estadounidense Silver Lake y el 12% de las firmas chinas China Media Capital y CITIC Group.

El club insignia del CFG, es el **Manchester City Football Club** de la ciudad de Mánchester, que juega en la Premier League. Se fundó el 23 de noviembre de 1880 bajo el nombre de St. Mark's, luego pasó a llamarse Ardwick Association Football Club en 1887 y finalmente, el 16 de abril de 1894, adquirió su denominación actual. Desde el 2003, el club disputa sus partidos en el Etihad Stadium.

El Manchester City ingresó a la Football League en 1892 y ganó su primer trofeo oficial con la FA Cup en 1904. El club disfrutó de su mayor período de éxito a finales de los años 60 y principios de los 70, cuando ganó el campeonato de Liga de First Division, FA Cup, Copa de la Liga y la Recopa de Europa bajo la dirección técnica primero de Joe Mercer y luego de Malcolm Allison. Después de perder la final de la FA Cup de 1981, el club pasó por un período de decadencia, que culminó con el descenso a la tercera división del fútbol inglés en 1998.

Después de recuperar su sitio en la Premier League en 2002, el 1 de septiembre de 2008 Abu Dhabi United Group for Development and Investment, un grupo inversor de los Emiratos Árabes con Mansour bin Zayed Al-Nahyan como máximo accionista, se hizo con el control del club por unos 250 millones de euros.

Desde la llegada de los nuevos dueños, el club ha obtenido un total de 17 títulos oficiales, entre los que se destacan seis Premier League (2012, 2014, 2018, 2019, 2021 y 2022).

Cabe destacar que cuatro de los diez clubes del CFG, tienen equipos femeninos, y se espera que la inversión tanto en el crecimiento como en la creación de nuevos equipos femenino siga creciendo, dado que varios clubes prometen invertir en el fútbol femenino de formas sin precedentes.

Así pues, el Manchester City cuenta también con un equipo femenino que destaca a nivel local desde el 2016; se trata del ***Manchester City Women's Football Club*** que juega en la ***FA Women's Super League***. Desde agosto del 2012, el equipo femenino Manchester City, comparte no sólo vínculos corporativos y recursos con el equipo masculino, sino también sus instalaciones de entrenamiento, además de ser incluido en el marketing y las redes sociales del lado de la Premier League.

2.2 ***Organización y madurez de la empresa***

El CFG, organiza cada club de fútbol de manera que se establece una clara estructura administrativa, donde existe un presidente que se encarga de las cuestiones puramente financieras, y un director de fútbol que se encarga de responsabilidades puramente deportivas.

Se conoce que el CFG es una de las organizaciones futbolísticas más innovadoras y emprendedoras a nivel tecnológico del mundo. Esta empresa presenta un estado de madurez en cuanto a datos entre el nivel de empresa y transformador, pudiendo considerarla una ***empresa Data Driven***. Esto es porque utiliza los datos como activo estratégico, tomando las decisiones en base a ellos. Algun ejemplo es que si algún rival localizado en Estados Unidos o en Australia, desarrolla una nueva estrategia para los tiros de esquina, videos de ello llegan inmediatamente a Manchester para ser analizados. También se hace uso de los datos para la selección de jugadores (scouting), para la mejora de la experiencia de sus fans o bien para operaciones de negocio como la venta de entradas.

A pesar de que el city posee una estrategia e inteligencia de negocio muy optimizada, y que parece estar en un nivel predictivo e incluso prescriptivo de toma de decisiones, esta información es la que se puede extraer de manera externa, aunque desde dentro de la organización, se debería de utilizar algún modelo que permita determinar objetivamente el nivel de madurez, como por ejemplo el [5x10 maturity model](#), para

clasificar a la empresa dentro de un nivel de madurez y poder poner el punto de mira en aquellas dimensiones en las que se deban implementar cambios.

En esta línea cabe destacar que el CFG firmó en 2015 un acuerdo con la empresa **tecnológica SAP**, la cual le proporciona herramientas para mejorar en todos los ámbitos, aspecto que también optimizó la comunicación interna y la trazabilidad. Este acuerdo se firmó con el pensamiento de que la unión del deporte y la tecnología puede optimizar los resultados. El City Football Group y sus clubes implementaron una gran variedad de soluciones basadas en la nube e impulsados por la plataforma SAP HANA, con el objetivo de simplificar sus operaciones mundiales, incrementar su negocio, aumentar la productividad y mejorar la experiencia de los aficionados. La adopción de esta tecnología, permite el análisis de los datos que generan los clubes del CFG de manera centralizada, tanto en lo deportivo como en el comercial.

En concreto, la aplicación dedicada exclusivamente al fútbol se llama **SAP Sports One**. Esta herramienta, basada en la nube, permite unificar la gestión del equipo, la planificación de los entrenamientos, la supervisión del estado de los jugadores y el análisis de los partidos.

En 2017 el Manchester City, comenzó a utilizar **SAP Challenger Insights** (una app basada en la nube) antes de los partidos para trazar su plan de juego, durante el descanso de los encuentros para realizar ajustes y después de los mismos para la planificación de futuros encuentros. Esta app, proporciona a entrenadores y analistas de datos información sobre las tácticas y características de un oponente, como las formaciones que emplea y las pautas ofensivas y defensivas que sigue.

Actualmente, el Manchester City no solo emplea la tecnología de SAP en modo batch antes del partido, en el descanso y una vez finalizado, sino en el banquillo mientras se está celebrando el encuentro, lo que significa, que también está adentrado en el **análisis de datos en tiempo real**.

En cuanto al análisis de datos, el software **SAP Predictive Analytics** y el solucionador inteligente para negocios **SAP BusinessObjects**, es el centro de la infraestructura de información de CFG, que pone al alcance todo tipo de información, desde venta de entradas hasta nuevos productos, y que permitirá tomar decisiones más rápidas y con mayor información.

Además, la empresa también adquirió la solución de plataforma de e-commerce basada en la nube de hybris, dentro de la cartera de SAP. Se trata de la [SAP Commerce Cloud](#),

que brinda una mejor experiencia a los aficionados, impulsando la fidelidad de los clientes.

Haciendo referencia a la gestión empresarial como tal, en el 2019, el Manchester City seleccionó a la empresa **Nouss Intelligence SL**, la cual ayuda a la mejor gestión global de los recursos requeridos para la organización, gestión y seguimiento de los acontecimientos que se llevan a cabo en las instalaciones gestionadas por el CFG. Nouss Intelligence proporciona servicios basados en la aplicación de técnicas de inteligencia artificial, que resuelven eficazmente complejos problemas de planificación de recursos, tareas y procedimientos de empresas de amplios sectores del mercado.

En cuanto a la seguridad y la protección de los datos, el CFG logró la asociación con **Acronis**, una empresa desarrolladora de software en las instalaciones y en la nube con una integración única de copia de seguridad, recuperación ante desastres, ciberseguridad y gestión de puntos finales.

Todo lo nombrado anteriormente, es un signo de que el CFG está embarcado desde hace varios años en una transformación digital férrea, adoptando nuevo software, digitalizando sus procesos, y basando la toma de decisiones en datos, lo que está ayudando al grupo a aumentar los beneficios, tanto a nivel deportivo como comercial.

2.3 Alcance del proyecto

El Manchester City football club es el club que más dinero factura con un total de 731 millones de euros anuales, importe que seguramente tenga una estrecha relación con la buena estrategia de datos con la que el club cuenta.

Por otro lado, actualmente estamos en un punto de inflexión en el que es evidente el auge del fútbol femenino, con cada vez más recopilación de datos y más interés por parte de los fans. Este hecho, podría sumar en la facturación, y aumentar aún más los beneficios del CFG. Así pues, este proyecto nace de la intención de invertir más recursos en el equipo femenino.

Centrándonos en el equipo femenino **Manchester City Women's Football Club**, fue ganador de la **FA Women's Super League** en el 2016, de la **Fa Cup** en 2016-17, 2018-19 y 2019-20, y de la **FA Women's League Cup** en 2014, 2016, 2018-19 y 2021-22.

Como se aprecia, el equipo tiene varios triunfos importantes, pero lleva desde el año 2016 sin ganar la liga nacional. Además, en la temporada 2020-2021 no quedó victorioso de ninguna competición, quedando en segunda posición de la **FA Women's Super**

League. El proyecto pretende identificar aquellas jugadas de más valor de la temporada 2020-2021, para focalizar los entrenamientos o bien invertir en los fichajes pertinentes, y terminar ganando la liga. En este caso y para focalizar el estudio, vamos a considerar jugadas de valor los **chutes y goles** ejecutados en dicha temporada, aunque en el horizonte del proyecto caben muchas otras posibilidades a considerar (asistencias, pases, paradas del portero, y un largo etcétera).

De esta manera, el proyecto puede tener una continuación futura, planteando otros análisis de estos datos para extraer nuevos insights, como podría ser:

- Analizar el tracking de jugadores
- Analizar por visualización de campo de juego las áreas de mayor y menor peligro
- Analizar la influencia e impacto de los partidos de local como visitante
- Analizar la correlación de Key players con los resultados de los partidos
- Encontrar las jugadas de más peligro en contra
- Analizar la efectividad en los pases y tipo (largos, cortos, rasos, elevados)
- Modelos predictivos y prescriptivos para la preparación de los partidos
- Análisis de diferentes tácticas ofensivas/defensivas y el impacto que tienen en generar o impedir jugadas de valor
- Análisis de qué jugadoras concretas generan acciones de más valor
- Analizar comportamiento de jugadores (patrones de movimiento, reconocimiento de patrones de ataque, colocación de los jugadores...)
- Analizar la tasa de posesión de balón
- Estimar los resultados (se pueden predecir los resultados de los partidos antes de que comiencen)

El súmmum sería desarrollar mediante **algoritmos de machine learning** un modelo prescriptivo, que se pudiera aplicar en tiempo real, para conseguir reacciones inmediatas.

Todas estas propuestas con el objetivo de optimizar el juego del equipo y aumentar el número de trofeos, repercutiendo positivamente en los beneficios del CFG. Estas múltiples opciones, muestran la polivalencia de los datos. La conclusión, es que los datos deben ser exprimidos al máximo, obteniendo valor para distintos proyectos, pudiéndolos simplificar o complicar según las necesidades.

Aún así, el alcance de cualquier proyecto de datos, va a depender de los procesos críticos que deban abordarse inicialmente para poder llevar a cabo el proyecto en sí.

Uno de los procesos críticos en este proyecto, es la **contratación de profesionales de los datos** que sean capaces de llevar a cabo el proyecto al máximo nivel. Otro proceso crítico es determinar si la fuente de **datos es fiable**, y en este caso, tal como hemos comentado, la empresa Statsbomb es líder en datos deportivos y cuenta con la confianza de más de 150 equipos y organizaciones deportivas en docenas de países de todo el mundo, aunque se tendrá en cuenta que vamos a trabajar con Opendata, lo que se traduce con un mayor esfuerzo en calidad de datos. La **calidad de datos** es uno de los procesos críticos clave, esta debe estar presente y monitorizada constantemente, con lo que deberemos establecer unos KPIs de calidad que nos proporcionen el estado de la calidad en todo momento.

Y para nosotros, el principal proceso crítico para el desarrollo del proyecto es la conformación de la **arquitectura de datos** con la cual se va a llevar a cabo el proyecto, establecer el software que va a ser necesario para todas las etapas del dato, desde la fuente de datos ya definida, pasando por la integración de datos y el almacenamiento, como finalmente el acceso y análisis de las jugadas de más valor, así como también disponer de la infraestructura necesaria acorde a las necesidades del proyecto.

El principal problema del proyecto es que no podemos obtener datos directamente del Manchester City, que seguramente dispongan de muchos más que los que le vende a Statsbomb. Además, como se ha mencionado anteriormente, vamos a trabajar con Opendata, lo que mengua la cantidad de datos a nuestro alcance.

2.4 Objetivos

El fútbol aún no se ha entregado por completo al análisis de datos, a diferencia de otros deportes. Así que nuestro equipo de Big Data, pretende aportar valor mediante los datos al **Manchester City Women's Football Club**, aprovechando el gran auge del fútbol femenino que hay actualmente. Los principales objetivos derivados del presente proyecto son los siguientes:

- Evaluar la calidad de datos disponibles
- Acercar al **Manchester City Women's Football Club** a la victoria de la liga 2023-2024, y de otras competiciones
- Aplicar algoritmos de ML para encontrar inteligencia en los datos
- Mejorar estrategias para la creación de jugadas en el campo
- Crear nuevos entrenamientos específicos para mejorar el rendimiento en jugadas que generan peligro

Estos objetivos, se pretenden asumir mediante el análisis de los datos disponibles de la sesión 2020-2021 teniendo en cuenta todos los partidos disputados por el **Manchester City Women's Football Club**.

Como conclusión, en este proyecto se pretenden desarrollar conocimientos de Big Data y Analytics aplicados, y para ello, se ha realizado un estudio previo de la empresa, que, en nuestro caso, es el club de fútbol **Manchester City Women's Football Club**. Nuestra pretensión es conformar una arquitectura de datos con las herramientas y softwares adecuados, que nos permitan obtener información de valor a través de los datos proporcionados por Statsbomb, que ayuden a la toma de decisiones para este equipo en cuanto a las jugadas de más valor en un partido. Todo ello, con el fin de optimizar los resultados tanto a nivel económico como profesional, y con el objetivo principal de aumentar el número de trofeos del equipo.

2.5 Propuesta de valor

El Manchester City tiene **gran cantidad de fans**, y como se ha comentado anteriormente el CFG apuesta por el fútbol femenino. El hecho de que el equipo femenino triunfe, seguramente vaya a contribuir a que estos fans se animen también al seguimiento del equipo femenino, y aumenten el número de ventas de entradas, camisetas, merchandising, suscripciones, y un largo etcétera; convirtiéndose en un ciclo virtuoso de beneficios.

Actualmente, la mayoría de los equipos analizan datos, por lo que es importante centrarse en los que de verdad aportan valor, para sacar ventaja en un entorno tan competitivo como este. En fútbol, la victoria suele venir determinada por detalles, por lo que, cualquier valor extra que aumente las posibilidades de éxito será clave. Así pues, los datos pueden aportar una variedad de beneficios, tanto a nivel individual como a nivel organizacional. A nivel individual, el despliegue y la consecución del proyecto, pueden ayudar a desarrollar habilidades profesionales, tales como la gestión del tiempo, la resolución de problemas y la toma de decisiones. Estas habilidades son esenciales para el éxito a largo plazo. A nivel organizacional, pueden mejorar la productividad, reducir costos, mejorar la calidad y la satisfacción de los clientes.

De esta manera, gracias a nuestro proyecto, y a extraer información de valor de los datos en cuanto a las jugadas de más valor del equipo femenino del City, se pueden preparar mejor los entrenos, mejorar el análisis pre y post partido, plantear nuevas jugadas para mejorar el desempeño de las jugadoras, o incluso hacer nuevos fichajes,

para aumentar el número de triunfos del **Manchester City Women's Football Club**, potenciando el ciclo virtuoso comentado.

Es decir, este proyecto pretende aventajarse respecto a otros clubs, para **promocionar su equipo femenino de fútbol**, disponer de otro equipo al máximo nivel, y seguir siendo la empresa futbolística que más dinero genera. De esta manera, el proyecto ofrece claramente una ventaja competitiva respecto a la competencia.

Además, es un empujón para todos aquellos clubes que no disponen de equipo femenino, mostrando que la inversión en la creación y crecimiento de este, puede aumentar los beneficios de la organización. Cada vez más niñas quieren jugar a fútbol, y proyectos como el nuestro, pueden mejorar y profesionalizar más a los equipos femeninos de referencia como el Manchester City, haciendo que el nivel y el interés de todos los clubes por el fútbol femenino incremente.

Por ende, nuestro equipo puede marcar la diferencia en el desarrollo de este proyecto, ya que contamos con un **equipo versátil y polivalente**, con diferentes perfiles profesionales pudiendo abordar el proyecto desde distintos puntos de vista, ideas y conocimientos, enriqueciendo los resultados. Además, somos un equipo **joven con ganas de aprender** y capacidad de **resiliencia**, fanáticos del fútbol y apostamos por el fútbol femenino, queriéndolo hacer de manera objetiva basada en datos, aportando valor a la empresa futbolística dominante, económicamente hablando.

3. Planificación

3.1 Roles y equipo

Data engineer / architect (Lluís Vega): este profesional es el responsable de **diseñar la arquitectura de datos**, como también crearla y hacerla operativa. Es el encargado de administrar, procesar y almacenar los datos para que puedan ser usados de forma accesible y fiable (calidad del dato). En este caso, principalmente se va a encargar de aprovisionar la infraestructura necesaria para analizar los datos.

Este profesional va a configurar y poner a disposición del resto del equipo todas las herramientas necesarias para el desarrollo del proyecto.

Data analyst (Matias Arce): es la figura encargada de **extraer y procesar** los datos para sacar conclusiones y ayudar a la toma de decisiones estratégicas. Va a ser el encargado de analizar los datos de la fuente, para poder extraer valor e identificar aquellas jugadas más valiosas. Además se va a encargar de la elaboración de los informes con los datos relevantes, y finalmente de mostrar los resultados obtenidos al equipo. La herramienta con la que principalmente va a trabajar es **Power BI**.

Data scientist (Daniel Áspera): este profesional va a complementar la función del data analyst, trabajando coordinada y paralelamente sobre los datos curados, aunque este se encarga de problemas más complejos, como la **identificación tendencias y patrones**; y en última instancia del desarrollo del modelo de ML. Las herramientas con las que principalmente va a trabajar son **Databricks Community** y **Power BI**.

CDO / Data strategy (José Áspera): esta figura es la principal **responsable del proyecto**, tiene comprensión profunda de la estrategia del negocio, conocimiento sobre el diseño y la gestión de estrategia de datos. Va a encargarse de establecer los objetivos del proyecto, alineándose con los del negocio, y se va a ocupar principalmente del gobierno, la seguridad y la privacidad de los datos.

La herramienta con la que principalmente va a trabajar es **Talend Open Studio for Data Quality**, a parte estar al corriente de los trabajos que se realizan en otras herramientas, ya que este profesional también debe tener una parte de conocimientos técnicos.

Project manager (Míriam Ibán): esta profesional se encarga principalmente de la **definición, planificación y ejecución** de un proyecto. También debe desarrollar tareas de coordinación y motivación del equipo para llevar a cabo las tareas planteadas, evaluar la calidad de los entregables, estimar los recursos que van a ser necesarios, gestión de problemas y cambios inesperados. Para ello, debe establecer una metodología de trabajo, que en este caso se trata de una metodología Agile.

La herramienta con la que principalmente va a trabajar es **Microsoft Excel** y **Lucidchart**.

Una vez tenemos asignados los roles y sus tareas, podemos pactar los honorarios. En la siguiente tabla se muestran las tarifas tanto por jornada como por hora, que corresponden a cada perfil profesional:

Rol	Tarifa jornada	Tarifa hora
Data Architect	320	40
Data Scientist	320	40
Data Analyst	280	35
CDO / Data Strategy	360	45
Project Manager	320	40

Fuente: Tabla de elaboración propia en Excel (Roles profesionales)

Aunque cada profesional utiliza la herramienta que más se adapta a la tarea que debe desarrollar, todos los integrantes se deberán familiarizar con todas las herramientas usadas en el proyecto, fomentando la transversalidad y comunicación entre todos.

3.2 Tareas

Tras la conformación del equipo y distribución de los roles, es necesario definir cuales son las tareas a realizar y cómo se van a repartir. Como se ha comentado anteriormente, el encargado de realizar esta repartición, coordinar al equipo y planificar el proyecto, es principalmente la **project manager**, con ayuda y aprobación del CDO.

Para poder empezar a trabajar con los datos, es esencial montar, configurar y desplegar la **arquitectura** con todas aquellas herramientas que cada uno de los profesionales va a necesitar para llevar a cabo sus tareas.

Así pues, la primera tarea que va a ser necesaria, es el desarrollo de la arquitectura de datos, que en este caso va a ser principalmente la configuración de **MongoDB Atlas** y **Compass**, el cuál va a ser el Data Lake, y en segunda instancia, el Data Warehouse, contando así con dos capas de datos, la cruda y la curada. De esta manera podemos considerar que nuestro proceso de ingestión y transformación de los datos va a ser de tipo ELT, ya que primero se van extraer los datos de Statsbomb y se van a cargar en el Data Lake, y posteriormente se van a transformar en función de las necesidades en las distintas etapas del proyecto.

El pipeline que van a seguir los datos es: desde la fuente de datos de origen (**Statsbomb**), al Data Lake (capa cruda de **MongoDB Atlas**), pasando por el ecosistema de **Databricks Community** para la transformación y la posterior conexión y

subida de datos al Data Warehouse (capa curada de **MongoDB Atlas**). Una vez los datos están transformados, se trata de la identificación de patrones mediante código Python en Databricks Community. Y finalmente, cuando ya se tenga la analítica a punto, se conectará a **Power BI** para la visualización de datos.

El arquitecto, también va a tener que configurar una herramienta que posibilite la realización del data quality y data profiling, que en este caso va a ser Talend Open Studio for Data Quality. El **CDO** va a encargarse de establecer las reglas de negocio, para monitorizar la calidad de los datos y realizar el data profiling.

Después del levantamiento de los servicios ELT, con la configuración de *MongoDB Atlas* y *Compass*, la conexión de esta con *Databricks Community* y la configuración de la plataforma de Power BI, se puede empezar propiamente con el **desarrollo** del proyecto.

Como se ha mencionado, en la etapa de analítica, se deberán reconocer patrones y tendencias dentro de los datos, que nos permitan sacar una conclusiones. Con estos datos, el **analista de datos** definirá las métricas para la creación del modelo de datos en Power Bi y el posterior diseño del informe. También, el **data scientist**, con estos datos analizados, podrá desarrollar modelos de ML para la predicción de gol de una determinada jugada, los cuales deberá entrenar y validar.

Y finalmente, cuándo todos los profesionales ya estén desarrollando sus tareas dentro del proyecto, el arquitecto deberá estar pendiente de la **puesta en producción**, aunque cabe destacar que el proyecto utiliza datos estáticos, de un período de tiempo pasado definido, y que por lo tanto los procesos ELT y la actualización de los informes no van a variar mucho; aún así, estará activo en cuanto a la seguridad, actualización, cambios o fallas de las distintas herramientas utilizadas.

Fase	Responsable	Rol	Tarea	Detalle	Días	Importe	Fecha inicio
Arquitectura	Lluís Vega / Miriam Ibán	Data Architect	Diseño arquitectura	Determinar herramientas tecnológicas	18	2880	4/4/23
Arquitectura	Lluís Vega	Data Architect	Configuración MongoDB Atlas y Conexión a Databricks Community	Configuración del Data Lake en MongoDB Atlas y conexión de Databricks Community a la BD	7	1120	22/4/23
Arquitectura	Lluís Vega	Data Architect	Transformación de datos, Configuración MongoDB (DW) y Power BI	Realizar las transformaciones necesarias y crear la capa de DW en MongoDB Atlas. Configuración de Power BI	9	1440	28/4/23
Arquitectura	Lluís Vega	Data Architect	Configuración herramienta Data Quality	Configuración de la herramienta de DQ para el monitoreo de la calidad y el data profiling	9	1440	28/4/23
Gobierno y Calidad	Jose María Áspera	CDO / Data Strategy	Data Quality y Data Profiling	Establecer reglas de negocio, realizar el data profiling de los datos y monitorizar la calidad	16	2880	22/4/23
Desarrollo	Matías Arce / Daniel Áspera	Data Scientist	Análisis de los datos y reconocimiento de patrones	Explorar los datos para poder sacar insights que sean de valor	16	2560	6/5/23
Desarrollo	Matías Arce	Data Analyst	Definición y diseño del informe	Definición de métricas, Diseño de informe y dashboard	16	2240	21/5/23
Desarrollo	Daniel Áspera	Data Scientist	Modelos de ML	Ánalisis avanzado de los datos, desarrollo y entrenamiento de modelos de ML	26	4160	12/6/23
Dirección y Gestión	Miriam Ibán	Project Manager	Cordinación de las tareas en el equipo	Definición, planificación y ejecución de un proyecto	171	27360	20/2/23
Go Live	Lluís Vega	Data Architect	Puesta en producción	Schedule de procesos de ETL/ELT, actualización de informes, configuración de la seguridad...	25	4000	10/7/23

Fuente: Tabla de elaboración propia en Excel (Distribución de tareas)

3.3 Calendario

La project manager ha establecido un cronograma para la optimización del tiempo y la entrega del proyecto a su debido tiempo. En el calendario, se distribuyen todas las tareas mencionadas en el apartado anterior, por orden cronológico dentro de los tiempos establecidos, que en nuestro caso se extiende desde el día 20/02/23 donde iniciamos el proyecto, hasta el día 29/08/23, fecha que establecemos como límite para la finalización del proyecto.

AÑO 2023							
TAREAS	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto
Definición del proyecto		20/02 - 28/02					
			ENTREGA 1 (28/03/23 - 3/04/23)				
Diseño de la arquitectura			4/04 - 21/04				
Extracción y Carga de datos, Configuración MongoDB (DL) y conexión con DataBricks				22/04-28/04			
Transformación de datos, Configuración MongoDB (DW) y Power BI					28/04-6/05		
Configuración herramienta Data Quality					28/04-6/05		
Data Quality y Data profiling de los datos				22/04 - 6/05			
Análisis de datos y reconocimiento de patrones					6/05 - 21/05		
Definición de métricas, Diseño de informe y dashboard					21/05 - 05/06		
					ENTREGA 2 (30/05/23 - 5/06/23)		
Análisis avanzado de datos y desarrollo de modelo ML						12/06 - 7/07	
Puesta en producción							10/07 - 4/08
							ENTREGA 3 (29/08/23 - 4/09/23)

Leyenda: Etapa 1 Etapa 2 Etapa 3

Fuente: Calendario de elaboración propia en Lucidchart

4. Viabilidad del proyecto

4.1 Técnica

A día de hoy, los datos que proporciona **Statsbomb** en su versión **Open Data**, son suficientes para abordar el proyecto, dado que recoge muchos tipos de datos específicos como serían: goles, asistencias, minutos jugados, posición del jugador en el campo, pases, centros, etcétera. Aunque es verdad que, pagando la versión premium de Statsbomb se pueden obtener muchos otros datasets de eventos, que podrían ser de interés y permitirían optimizar el proyecto.

Haciendo referencia a la calidad de datos, Statsbomb es líder en el mundo como proveedor de datos avanzados en el fútbol, se ha convertido en el recurso de referencia para el análisis de datos futbolísticos, y asegura tener un proceso de control de calidad minucioso y detallado, que garantiza tener a disposición los datos más precisos y de mayor calidad. Cabe destacar, que **gran cantidad de clubes confían** en los datos de Statsbomb lo que demuestra que brindan datos de alta calidad.

Aún así, deberemos de analizar y monitorizar la calidad de datos, fijando unos umbrales para asegurarnos que la calidad es la suficiente para llevar a cabo el proyecto de manera óptima. Tendremos en cuenta las principales dimensiones de calidad del dato (Consistencia, Completitud, Unicidad, Temporalidad, Validez y Precisión). Así pues, una vez definidos qué eventos son de importancia para nuestro proyecto, debemos asegurarnos de que los datos cumplen con los umbrales definidos para cada dimensión de la data.

Por otro lado, ante cualquier proyecto de datos, se puede hacer uso de los datos internos que posee la misma empresa, pero también cabe la posibilidad de obtener o comprar datos externos a la compañía. De este modo, a parte de Statsbomb existen otras empresas distribuidoras de datos como Stats Perform (Opta Data), BeSoccer, Opta o Wyscout, que también podrían proporcionar datasets adicionales y ayudarnos a alcanzar los objetivos.

Además de lo nombrado anteriormente, se podría complementar con otros conjuntos de datos a través del networking y comunidad de datos en el ambiente futbolístico, aunque con gran precaución en la calidad de datos.

4.2 Económica

El proyecto requiere de la contratación de una serie de profesionales de datos que mencionamos a continuación:

- CDO / Data strategy: al ser un proyecto pequeño, consideramos que estos dos perfiles profesionales los va a asumir una misma persona. Esta figura es la **responsable del proyecto**, debe de tener comprensión profunda de la estrategia del negocio, conocimiento sobre el diseño y la gestión de estrategia de datos, así como también liderar y gestionar el equipo. También va a encargarse de establecer los objetivos del proyecto, alineándose con los del negocio. Este profesional es el que se va a ocupar principalmente del gobierno, la seguridad y la privacidad de los datos.

El rango salarial de este perfil en España es de **80k** euros anuales de media.

- Project Manager: su función es **interaccionar** con todos los miembros del equipo, estableciendo una metodología de trabajo y asegurándose de que cada cuál realiza las tareas asignadas. Debe ser capaz de ofrecer soluciones a los problemas que vayan surgiendo.

El rango salarial de este perfil en España es de **45k** euros anuales de media.

- Data Architect / Engineer: este profesional es el **responsable de diseñar la arquitectura de datos**, como también crearla y hacerla operativa. Es el encargado de administrar, procesar y almacenar los datos para que puedan ser usados de forma accesible y fiable (calidad del dato).

El rango salarial de este perfil en España es de **46k** euros anuales de media.

- Data Analyst: es la figura encargada de recopilar datos para **identificar tendencias** que ayuden a la toma de decisiones estratégicas. Realizan análisis estadísticos para ayudar a responder preguntas y resolver problemas.

El rango salarial de este perfil en España es de **35k** euros anuales de media.

- Data Scientist: este perfil se complementa con el data analyst, pero desarrolla **herramientas y métodos avanzados**, para extraer la información que la organización necesita para resolver problemas más complejos.

El rango salarial de este perfil en España es de **44k** euros anuales de media.

Así pues, teniendo en cuenta que la duración del proyecto va a ser de aproximadamente un año, el rango monetario en el que nos movemos es desde unos **250k y 300k** anuales. Este rango es una estimación, dado que dependiendo de las capacidades de los profesionales elegidos, se van a invertir más o menos horas, requiriendo una mayor o menor inversión monetaria.

La literatura estima que dentro del coste del proyecto, se debe tener en cuenta la inversión en infraestructura, que aproximadamente se calcula que es un 20% del coste

total. Aún así, al ser un proyecto relacionado con el trabajo final de máster, vamos a trabajar con herramientas OpenSource.

En cuanto a la arquitectura de datos que prevemos la siguiente arquitectura:

1. **Fuentes de datos:** Statsbomb
2. **Herramientas ETL / ELT / Trazabilidad:** TOSDQ / Azure Data Factory
3. **Herramientas de almacenamiento / clusters:** Spark con Azure CosmosDB
4. **Herramientas de analítica:** Databricks
5. **Visualización:** Power BI
6. **Herramientas para Gobierno de datos / Seguridad:** Azure Data Explorer y Talend Open Studio for Data Quality

Así pues, considerando las cifras mencionadas, opinamos que dado el estatus económico del CFG, el proyecto es totalmente viable económicamente hablando, dado que como hemos comentado, la inversión que se va a realizar es únicamente en los recursos humanos esenciales para la realización del proyecto con éxito.

Al finalizar el proyecto, si este resulta exitoso, se identifican aquellas jugadas de mayor valor para el equipo, y al tomar decisiones en base a estos datos se consigue ganar la liga 2023-2024, ya haber un ROI muy interesante y positivo.

4.3 Valoración económica

Como en todo proyecto, es esencial la valoración económica para evaluar la viabilidad del proyecto, económicamente hablando.

Teniendo en cuenta el coste de cada profesional (tabla de tarifas) y también la extensión en el tiempo de cada tarea (cronograma), podemos determinar el gasto en recursos humanos del proyecto. En la tabla siguiente, podemos visualizar las jornadas que suponen cada tarea y su precio asociado, dando como resultado el coste total.

Tarea	Días	Precio / Jornada	Suma de Coste
Diseño arquitectura	18	320,00 €	5.760,00 €
Configuración MongoDB Atlas y Conexión a Databricks Community	7	320,00 €	2.240,00 €
Transformación de datos, Configuración MongoDB (DW) y Power BI	9	320,00 €	2.880,00 €
Configuración herramienta Data Quality	9	320,00 €	2.880,00 €
Data Quality y Data Profiling	16	360,00 €	5.760,00 €
Análisis de los datos y reconocimiento de patrones	16	320,00 €	5.120,00 €
Definición y diseño del informe	16	280,00 €	4.480,00 €
Modelos de ML	26	320,00 €	8.320,00 €
Cordinación de las tareas en el equipo	171	320,00 €	54.720,00 €
Puesta en producción	25	320,00 €	8.000,00 €
Total General	313	3.200,00 €	100.160,00 €

Fuente: Tabla de elaboración propia en Excel (Coste por tarea)

Tal y como se ha explicitado anteriormente, todas las herramientas que vamos a utilizar, tanto la fuente de datos como las distintas tecnologías, son **open data** y **open source** respectivamente. Por tanto, los recursos tecnológicos están exentos de costes.

Así pues, la única inversión que va a realizarse es propiamente en el trabajo que realiza cada profesional, es decir, únicamente en los recursos humanos esenciales para la realización del proyecto. De esta manera, con la identificación de las jugadas de valor y la identificación de patrones, se va a poder optimizar la toma de decisiones, por lo que el ROI va a ser positivo, y el proyecto va a resultar beneficioso para el CFG.

4.4 Conclusión de viabilidad

Como conclusión y teniendo en cuenta todo lo mencionado, entendiendo los objetivos del proyecto, el análisis preliminar de los datos y la estimación en cuanto a inversión, consideramos que es un proyecto totalmente viable y pertinente. Actualmente disponemos de conjuntos de datos y herramientas de análisis Opensource, que facilitan la viabilidad tanto técnica como económica del presente proyecto.

Para evaluar de manera general la viabilidad del proyecto y analizar la realidad, tanto interna como externa de este y poder tomar decisiones o estar prevenidos hemos llevado a cabo un **análisis DAFO** que nos permite saber las debilidades y amenazas, como también las fortalezas y oportunidades. Este análisis, nos ha ayudado a definir una estrategia potenciando las fortalezas, intentando superar las debilidades, controlar las amenazas y beneficiarnos de las oportunidades.

Análisis Interno	Análisis Externo
<p>DEBILIDADES</p> <p>Aspectos limitadores de la capacidad de desarrollo del proyecto, debido a características internas.</p> <ul style="list-style-type: none">• No disponemos datos suficientes de temporadas recientes.• La temporada que analizamos es un tanto antigua, pueden haber modificado las jugadoras y otros aspectos.• No disponemos de perfiles profesionales experimentados en la materia, nos situamos en el inicio de la curva de aprendizaje.• Falta de disponibilidad de recursos económicos.	<p>AMENAZAS</p> <p>Factores externos que pueden llegar a impedir la ejecución o poner en peligro la viabilidad del proyecto.</p> <ul style="list-style-type: none">• Dificultad de encontrar herramientas adecuadas que sean gratuitas o tengamos licencia del máster.• Problemas de conexión a internet.
<p>FORTALEZAS</p> <p>Conjunto de recursos internos, posiciones de poder y cualquier tipo de ventaja competitiva propia del proyecto.</p> <ul style="list-style-type: none">• Profesores expertos en todas las fases del proyecto.• Predisposición para adoptar nuevas herramientas tecnológicas.• Comunicación óptima entre los miembros del equipo.	<p>OPORTUNIDADES</p> <p>Factores ajenos al proyecto que favorecen su desarrollo o brindan la posibilidad de implantar mejoras.</p> <ul style="list-style-type: none">• El auge del fútbol femenino.• Varias empresas proveedoras proporcionan datos Opendata.• Nuevas herramientas disponibles para el análisis de datos.

No obstante, al igual que en cualquier otro proyecto, pueden surgir problemas no contemplados que entorpezcan el avance del proyecto, aunque a priori contamos con la viabilidad necesaria, tanto a nivel tecnológico como también económico.

5. Desarrollo del proyecto

5.1 Metodología de trabajo

El equipo se ha regido por la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining). Esta consta de 6 fases principales:

- Entendimiento de negocio: determinación de los objetivos de negocio. Se revisa la situación, los recursos, análisis de costo-beneficio, y se determinan los objetivos finales, que en este caso es principalmente reconocer aquellas jugadas que generan valor, y a partir de aquí se plantea la planificación del proyecto.

Realizamos una búsqueda exhaustiva del estado actual del CFG, y en concreto del **Manchester City Women's Football Club** para entender qué es lo que podíamos extraer de los datos que fuera útil para mejorar el rendimiento del equipo, y por ende, generar más beneficios (primera etapa).
- Entendimientos de los datos: recolectamos los datos, realizamos el data profiling y evaluamos la calidad. Se realizó el primer vistazo de los datos mediante la API de **Statsbombpy** para realizar una primera visualización de los datos que se facilitaba de manera Open Data. Y posteriormente, se descargaron los Json de los eventos para realizar el data profiling y evaluar la calidad (segunda etapa).
- Preparación de los datos: seleccionamos los datos diana, determinando que columnas y variables son necesarias, con el resultado final del dataset listo para analizar. Una vez realizado el data profiling y evaluado la calidad de estos datos, pasamos a la realización del **EDA** y de la analítica para reconocer patrones y tendencias en los datos (segunda etapa).
- Modelado: realización de varios supuestos y pruebas para definir parámetros y métricas interesantes y de valor. También se lleva a cabo la **selección del modelo** que más se adapta a los datos para determinar qué jugadas van a ser de valor para el equipo (segunda / tercera etapa).
- Evaluación: es donde, considerando los objetivos establecidos, decidimos si el proyecto es **exitoso**, o si por contra tiene **margen de mejora** y podemos retroceder a alguna fase para optimizarlo (tercera etapa).
- Despliegue: se genera un **informe** y un **dashboard** que plasman toda la analítica y el modelo llevado a cabo de forma visual, y esta se presenta a los interesados (tercera etapa).

Como podemos visualizar en el anexo 1, es un proceso iterativo, en el que en cualquier caso se puede retroceder y replantear la fase anterior. Dentro de esta metodología y acorde a la organización de Threepoints, dividimos esas 6 fases de la metodología en 3 grandes etapas por las que va a pasar el proyecto, que corresponden a las 3 entregas establecidas.

En cuanto a la coordinación del equipo de trabajo, se establece una reunión periódica cada sábado por la mañana a las 11am de España, mediante la plataforma Google Meet, con una extensión máxima de 2h. De esta manera, se establecen sprints de 1 semana, donde se consensúan las tareas a realizar hasta la siguiente reunión. Así pues, en cada reunión se realiza lo siguiente:

- Presentación del resultado de las tareas acordadas en el sprint anterior.
- Debate, opiniones y resolución de dudas.
- Correcciones pertinentes.
- Consenso de tareas para siguiente sprint.

Además, durante la semana se trabaja individualmente y se consultan dudas a través del grupo común en whatsapp.

5.2 Arquitectura

La arquitectura considerada desde un inicio fue dentro del **ecosistema cloud** que ofrece Microsoft denominados Microsoft Azure. Se trata de una colección de servicios informáticos integrados en la nube, y entre ellos disponemos de gran cantidad de herramientas destinadas a la ciencia de datos y Big Data.

Realmente es una arquitectura muy potente, dado que al trabajar en la nube brinda la posibilidad de crear un espacio colaborativo donde todos los integrantes del equipo podemos realizar de manera paralela nuestras tareas. **Microsoft Azure** ofrece la posibilidad de crear una cuenta gratuita que dispone de recursos gratuitos hasta 12 de meses. Además, las herramientas disponen de gran cantidad de documentación que hacen más fácil el manejo.

De esta manera, empezamos a montar la arquitectura sobre la nube de Microsoft, de manera que hicimos la ingestión de datos desde Statsbomb en formato Json hacia el **Data Lake Gen2**. Posteriormente conectamos al almacenamiento el **Azure Data Factory**, que es la herramienta que permite realizar las transformaciones necesarias, con la posibilidad de crear pipelines, data flows y integrations runtimes.

Una vez los datos los tenemos limpios, los subimos a **Azure CosmosDB** que es la base de datos para datos no estructurados o semi-estructurados, para conectarnos con Azure Databricks y realizar la analítica. También se dispone de **Azure Machine Learning** para crear los modelos de ML y el **Azure Data Explorer** para el Data Quality. Y finalmente la conexión a **PowerBI Embedded** para la visualización de los datos.

Tras montar y poner en marcha esta arquitectura, nos encontramos con un gran problema; el coste. Con la herramienta de análisis de costes dentro de la nube de Microsoft pudimos reportar un elevado coste proveniente principalmente del Azure Data Factory y de la conexión con PowerBI Embedded. De esta manera, decidimos retroceder y plantear la arquitectura de nuevo.

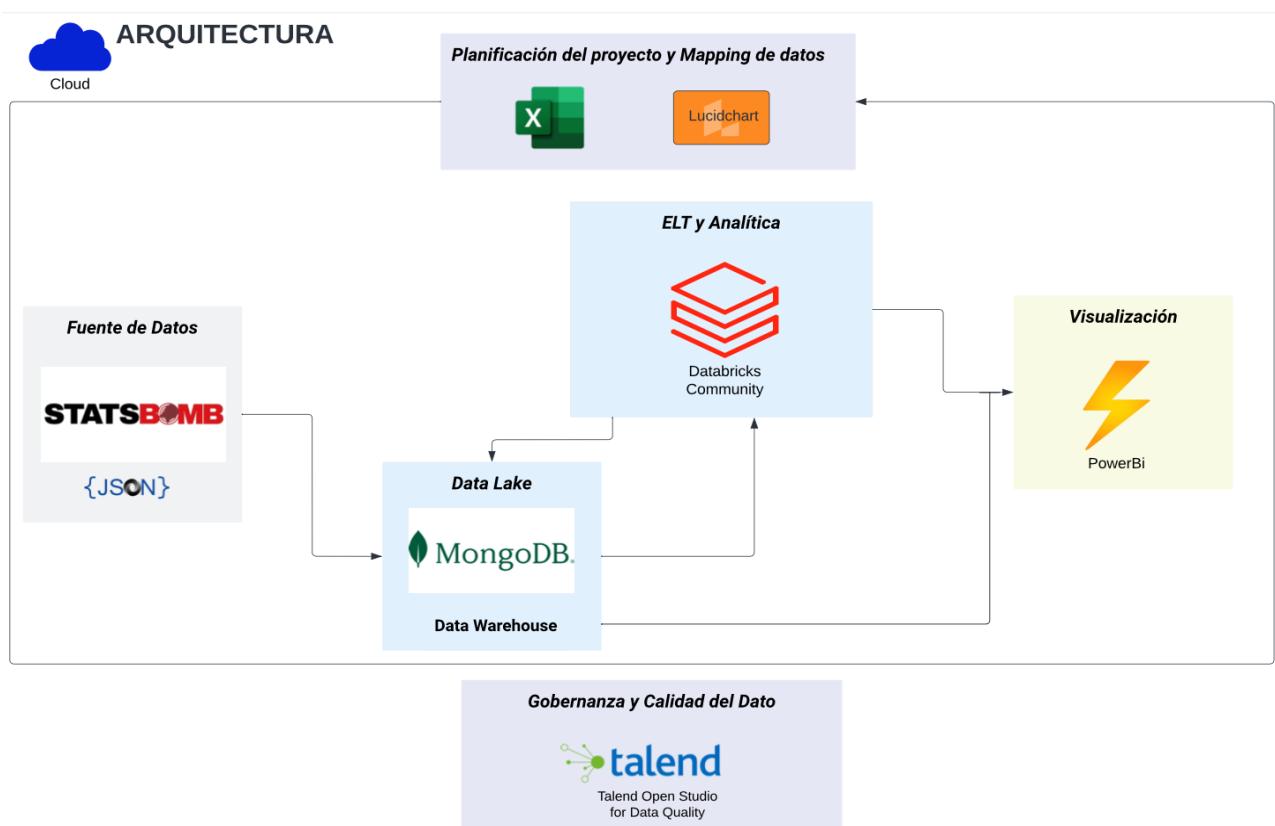
Por todo ello, la decisión de la arquitectura ha sido realmente compleja, dado que no encontrábamos las herramientas que cumplieran el principal requisito: la gratuidad. Es así, que se han acabado seleccionado las siguientes herramientas:

- **Fuente de datos:** *Statsbomb* (datos semi-estructurados), compañía externa. Proporciona en [Github](#) los datos en formato json, los cuales fueron descargados en local y subidos directamente al repositorio mediante el terminal.
- **Repositorios analíticos y operacionales:** *Mongodb Atlas* y *Compass*. Nos decantamos por esta base de datos ya que es basada en nube y está administrada por MongoDB, lo que disminuye el grado de complejidad técnico, además de un acceso único por todos los usuarios con una única cuenta compartida. Además es una base de datos escalable de manera horizontal, lo que permitiría la ampliación del proyecto en un futuro. Construimos una primera capa que corresponde a la capa Data Lake con los datos crudos, y otra capa Data Warehouse con los datos curados, con las transformaciones pertinentes.
- **Transformación, Analítica y valor:** *Databricks Community*. Esta herramienta también basada en la nube, permite transformaciones mediante código en Python, además de ser un entorno colaborativo. Se realizó la conexión con MongoDB Atlas a partir del enlace de conexión que esta facilita, para principalmente realizar el EDA, el proceso analítico y el reconocimiento de patrones.
- **Visualización:** **PowerBI**. La conexión de Databricks Community con PowerBI era previo pago, motivo por el cuál, ambas herramientas en nuestra arquitectura no tienen una conexión directa. Se descargaron los datos curados en formato

CSV, y tras el proceso analítico, se ha creado el cuadro de mandos con las métricas más representativas, y el respectivo informe.

- **Gobierno y Calidad:** *Talend Open Studio for Data Quality*. La versión open source de talend, no brinda la posibilidad de realizar el data quality en documentos con formato json. De esta manera, con los datos curados descargados desde MongoDB Atlas en formato CSV, se pudo realizar la calidad de datos con la herramienta Talend Open Studio for Data Quality.
- **Planificación del proyecto y Mapping de datos:** *Lucidchart y Excel*. Lucidchart es una herramienta de diagramación basada en la web, que permite a los usuarios colaborar y trabajar juntos en tiempo real, creando los diagramas de flujo, repartición de tareas y calendarios mostrados anteriormente, para la óptima planificación del proyecto. Excel ha sido la base para la estructura y organización, y en Lucidchart se ha conformado la visualización.

De esta manera, la arquitectura resultante tras varias pruebas, es la siguiente:



5.3 Preparación de los datos y analítica

5.3.1 Descripción de la fuente de información

En el proceso analítico, vamos a trabajar con documentos en formato JSON dado que la arquitectura desplegada, permite trabajar con este tipo de archivo, ya que si se estructuraran en formato tabla, quedarían muchos nulos y disminuiría la eficacia de la base de datos.

Statsbomb, proporciona una API Python para poder acceder a los datos de los documentos JSON. En primera instancia para hacer una vista previa de la fuente de datos, llevamos a cabo el siguiente proceso:

1. Competitions

In [5]: sb.competitions()						
13	16	44	Europe	Champions League	male	False
14	16	76	Europe	Champions League	male	False
15	37	90	England	FA Women's Super League	female	False
16	37	42	England	FA Women's Super League	female	False

2. Matches

```
In [6]: matches=sb.matches(competition_id=37, season_id=90)
```

3. Matches solo del City

```
In [7]: matches_City=matches.loc[(matches['home_team']=='Manchester City WFC') |  
(matches['away_team']=='Manchester City WFC')]
```

En este punto, hemos obtenido los partidos del Manchester City WFC de la FA Women's Super League del año 2020-2021. Proseguimos a obtener los eventos de todos estos partidos, que es lo que realmente nos interesa para poder llevar a cabo el análisis.

Cada partido, dispone de un JSON específico con los eventos, de manera que para visualizar todos los eventos de la temporada del City, se debe de generar un bucle a partir de los match_id de los partidos, para crear un dataframe con todos los eventos. Esto va a ser posible gracias a la librería Pandas. Así pues, creamos un dataframe con todos los eventos de los partidos del City en la temporada 2020-2021.

4. Dataframe con los eventos de los partidos del City temporada 2020-2021.

```
In [20]: dfs=[]  
for match_id in match_City_id:  
    df=sb.events(match_id=match_id)  
    dfs.append(df)
```

```
In [21]: #Juntamos los dataframes de la lista dfs
df_final=pd.concat(dfs)
```

Desde ya, tenemos todos los eventos de los partidos del City de la temporada 2020-2021. Una vez hecho este proceso, ya tenemos una vista preliminar de los datos con los que vamos a tratar, y de la cual se realizó el data profiling inicial.

A partir de aquí, y en vista de las características de los datos (semi-estructurados), creamos en MongoDB Atlas un clúster llamado “TFM” y una base de datos llamada “Statsbomb”, en la que crearemos la colección principal de “Eventos”. En esta colección subiremos todos los JSON en crudo, que Statsbomb proporciona sobre los eventos de la temporada de la **FA Women's Super League** 2020-2021.

Una vez tenemos cargados todos los eventos de la temporada en MongoDB, ya se puede realizar la conexión con Databricks community, desde la cuál realizaremos el proceso de transformación, y aplicaremos los filtros pertinentes.

```
5 #Conexión a la BD mongo
6 connectionString='mongodb+srv://miriamiban:passw0rd@tfm.syumvpa.mongodb.net/?retryWrites=true&w=majority'
7 client = pymongo.MongoClient(connectionString)
8
9 #Selección de la BD y colección
10 database = client['Statsbomb']
11 collection1 = database['Events']
```

El primer paso que vamos a realizar es filtrar esta colección en 3 pasos principales:

1. Obtener solo los **eventos del Manchester City**

```
1 # Filtrar eventos del Manchester City WFC
2 eventos_manchester_city_wfc = collection1.find({"team.name": "Manchester City WFC"})
3
4 # Convertir el cursor a una lista y obtener el número de documentos
5 documentos_filtrados = list(eventos_manchester_city_wfc)
6 num_eventos = len(documentos_filtrados)
7
8 # Verificar si hay documentos
9 if num_eventos > 0:
10     # Crear una nueva colección y borrar los documentos existentes
11     db_nueva_coleccion = client['Statsbomb']
12     nueva_coleccion = db_nueva_coleccion['eventos_manchester_city_wfc']
13     nueva_coleccion.delete_many({}) # Eliminar documentos existentes en la nueva colección
14
15     # Insertar los documentos filtrados en la nueva colección
16     nueva_coleccion.insert_many(documentos_filtrados)
17 else:
18     print("No se encontraron eventos del Manchester City WFC.")
```

Con este filtro reducimos de 78.953 documentos a 48.969, creando una nueva colección: “**eventos_manchester_City_wfc**”.

2. Filtrar los eventos del Manchester City para obtener solo los **chutes**

```
1 # Filtrar eventos del Manchester City WFC
2 chutes_manchester_city_wfc = collection2.find({"type.name": "Shot"})
3
4 # Convertir el cursor a una lista y obtener el número de documentos
5 documentos_filtrados1 = list(chutes_manchester_city_wfc)
6 num_eventos = len(documentos_filtrados1)
7
8 # Verificar si hay documentos
9 if num_eventos > 0:
10     # Crear una nueva colección y borrar los documentos existentes
11     db_nueva_coleccion = client['Statsbomb']
12     nueva_coleccion = db_nueva_coleccion['chutes_manchester_city_wfc']
13     nueva_coleccion.delete_many({})
14
15     # Insertar los documentos filtrados en la nueva colección
16     nueva_coleccion.insert_many(documentos_filtrados1)
17 else:
18     print("No se encontraron chutes del Manchester City WFC.")
```

La cantidad de disparos que el City realizó fueron 425 disparos, que corresponde al mismo número de documentos que contiene la colección obtenida con este filtro: **"chutes_manchester_city_wfc"**.

3. Filtrar los chutes del Manchester City para obtener solo los **goles**

```
1 # Filtrar eventos del Manchester City WFC
2 goals_manchester_city_wfc = collection3.find({"shot.outcome.name": "Goal"})
3
4 # Convertir el cursor a una lista y obtener el número de documentos
5 documentos_filtrados2 = list(goals_manchester_city_wfc)
6 num_eventos = len(documentos_filtrados2)
7
8 # Verificar si hay documentos
9 if num_eventos > 0:
10     # Crear una nueva colección y borrar los documentos existentes
11     db_nueva_coleccion = client['Statsbomb']
12     nueva_coleccion = db_nueva_coleccion['goals_manchester_city_wfc']
13     nueva_coleccion.delete_many({})
14
15     # Insertar los documentos filtrados en la nueva colección
16     nueva_coleccion.insert_many(documentos_filtrados2)
17 else:
18     print("No se encontraron goles del Manchester City WFC.")
```

Al filtrar por goles, obtenemos 62 documentos correspondientes a los 62 goles a favor del City en la temporada 2020-2021, creando otra colección: **"goals_manchester_city_wfc"**.

Una vez hemos realizado estas tres transformaciones, la base de datos en MongoDB Atlas queda de la siguiente manera:

Statsbomb

LOGICAL DATA SIZE: 84.8MB STORAGE SIZE: 26.21MB INDEX SIZE: 3.81MB TOTAL COLLECTIONS: 4

CREATE COLLECTION

Collection Name	Documents	Logical Data Size	Avg Document Size	Storage Size	Indexes	Index Size	Avg Index Size
Events	78953	51.31MB	682B	15.86MB	1	2.31MB	2.31MB
chutes_manchester_city_wfc	425	1.15MB	2.77KB	360KB	1	32KB	32KB
eventos_manchester_city_wfc	48969	32.18MB	690B	9.93MB	1	1.45MB	1.45MB
goals_manchester_city_wfc	62	163.59KB	2.64KB	68KB	1	20KB	20KB

Tenemos la base de datos llamada **Statsbomb** con la colección principal de **Eventos** de la temporada 2020-2021 de la liga femenina inglesa, y de esta extraemos tres colecciones más, relacionadas estrechamente con el Manchester City: eventos, chutes y goles.

En este punto ya tenemos la base de datos montada con toda la información necesaria para realizar los siguientes pasos: *data quality* y *EDA*.

5.3.2 Data Quality

Para la realización del Data Quality de los datos, tal y como ya se mencionó, optamos por la herramienta Talend Open Studio for Data Quality (TOSDQ).

En su versión open source no permite el análisis de JSON, motivo por el cual decidimos usar la librería Pandas para crear un dataframe de la colección de chutes, y poderlo transformar en CSV, formato que sí admite la versión **open source de TOSDQ**.

```
# Dataframe CHUTES
from pandas.io.json import json_normalize
import json
goals = list(collection3.find())
df_chutes = json_normalize(goals)
df_chutes
```

Tras la creación del dataframe, aunque se realiza la normalización de los json, quedan algunos campos importantes en formato array, como por ejemplo las columnas “location” y “related_events”, importantes para nuestro análisis. Así pues, se deben transformar para dividirlas ya que sino van a aparecer como un array en la misma columna.

```
# Dataframe CHUTES
df_chutes[['location_x', 'location_y']] = df_chutes['location'].apply(lambda loc: pd.Series(loc))
df_chutes[['shot_end_location_x', 'shot_end_location_y', 'shot_end_location_z']] = df_chutes['shot.end_location'].apply(lambda loc: pd.Series(loc))
df_chutes[['related_events1', 'related_events2', 'related_events3']] = df_chutes['related_events'].apply(lambda loc: pd.Series(loc))
df_chutes
```

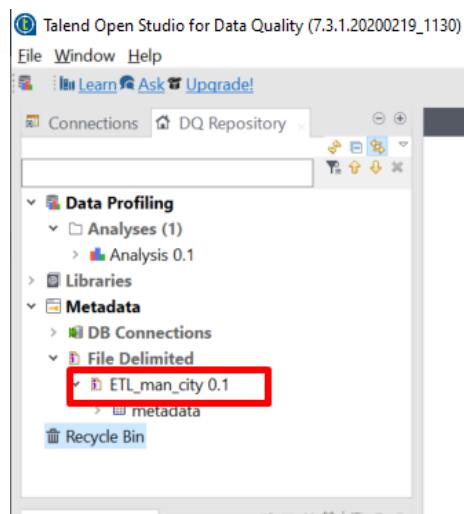
Columnas originales	Columnas transformadas
'location'	"location_x", "location_y"
'shot.end_location'	"shot_end_location_x", "shot_end_location_y", "shot_end_location_z"
'related_events'	"related_events1", "related_events2", "related_events3"

Automáticamente, tras la división, eliminamos las columnas originales para evitar redundancia de datos. El resultado del dataframe son 425 registros y 41 columnas que son las que se muestran a continuación:

```
(425, 41)
Index(['_id', 'id', 'index', 'period', 'timestamp', 'minute', 'second',
       'possession', 'duration', 'type.id', 'type.name', 'possession_team.id',
       'possession_team.name', 'play_pattern.id', 'play_pattern.name',
       'team.id', 'team.name', 'player.id', 'player.name', 'position.id',
       'position.name', 'shot.statsbomb_xg', 'shot.key_pass_id',
       'shot.outcome.id', 'shot.outcome.name', 'shot.type.id',
       'shot.type.name', 'shot.technique.id', 'shot.technique.name',
       'shot.body_part.id', 'shot.body_part.name', 'shot.freeze_frame',
       'under_pressure', 'location_x', 'location_y', 'shot_end_location_x',
       'shot_end_location_y', 'shot_end_location_z', 'related_events1',
       'related_events2', 'related_events3'],
      dtype='object')
```

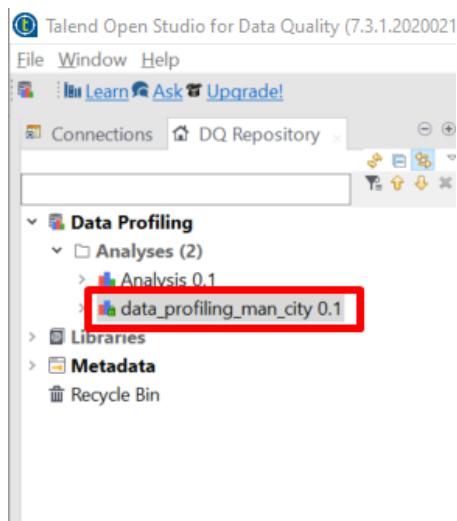
Este dataframe se va a exportar en formato CSV para poder cargarlo en TOSDQ. De esta manera, creamos una nueva colección de MongoDB (“**ETLchutes_Manchester_City_wfc**”) y la exportamos en CSV.

Dentro del TOSDQ hemos subido un “*File delimited*” con nuestro archivo csv “ETLchutes_manchester_City_wfc.csv” que los llamaremos “ETL_man_city”.



Fuente: Talend Open Studio for Data Quality (Repositorio)

Dentro del “*data profiling*” crearemos un análisis por columnas. El nombre del nuevo análisis lo llamaremos “**data_profiling_man_city**”.

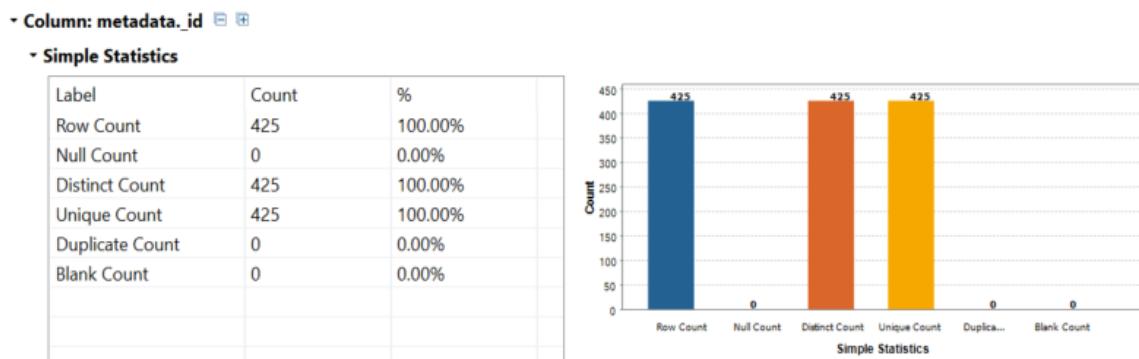


Fuente: Talend Open Studio for Data Quality (Data profiling)

Para determinar la calidad de los datos nos basamos principalmente en 6 dimensiones: completitud, unicidad, validez, consistencia, temporalidad y precisión, por orden de importancia.

Así pues, realizamos un análisis de calidad de los datos desglosado por cada columna seleccionada:

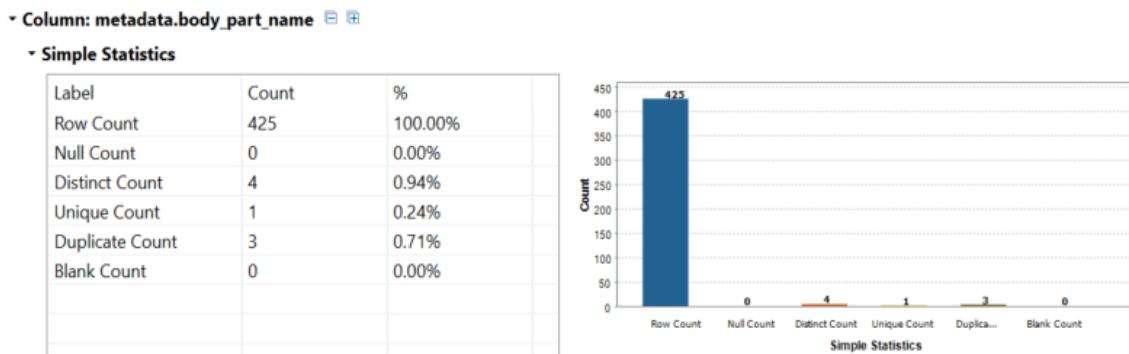
→ **ID**: Este es el “*id*” generado por “Statsbomb” al crear el archivo de eventos, y podemos observar desde el análisis de las columnas que no se repite ningún evento (unicidad) ni tenemos ningún blanco (completitud), por lo que es considerado una columna de calidad.



Fuente: Talend Open Studio for Data Quality (Estadísticas básicas “id”)

→ **Body_part_name**: En esta columna nos parece razonable que haya 3 etiquetas repetidas, ya que corresponde a la parte del cuerpo que la jugadora realizó el disparo.

Podemos observar que se encuentran como duplicados los disparos con el pie izquierdo, derecho y la cabeza. Y como etiqueta única se encuentra “other”, por lo que el disparo se realizó con una parte del cuerpo no habitual.



Fuente: Talend Open Studio for Data Quality (Estadísticas básicas “body_part_name”)

Duplicates Count:

body_part_name
Head
Left Foot
Right Foot

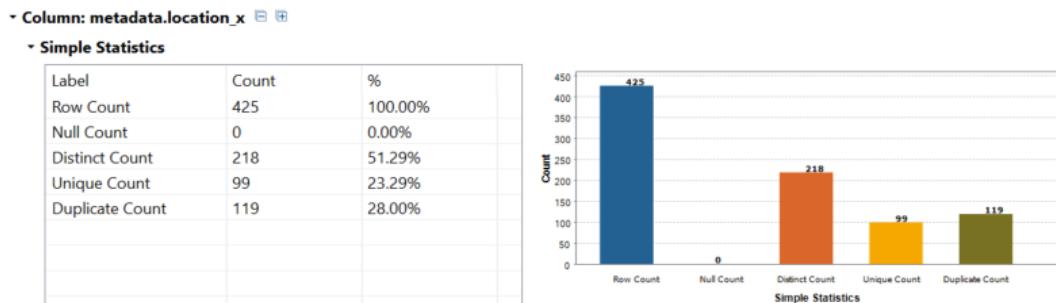
Unique count:

body_part_name
Other

Fuente: Talend Open Studio for Data Quality

→ **Location_x:** esta columna indica la posición de inicio que se encuentra la jugadora a la hora de realizar el disparo sobre el eje X. El valor máximo para el eje “X” es de 120, ya que es el límite del campo correspondiente a ese eje. Por lo que los valores se deben encontrar entre esos parámetros [0-120].

Se pueden observar datos duplicados, ya que existe la posibilidad de que distintos chutes se realicen desde la misma localización, por lo que no es un problema.



Fuente: Talend Open Studio for Data Quality (Estadísticas básicas “location_x”)

→ **Location_y:** esta columna indica la posición de inicio que se encuentra la jugadora a la hora de realizar el disparo dentro del eje Y. El valor máximo para el eje “Y” es de 80,

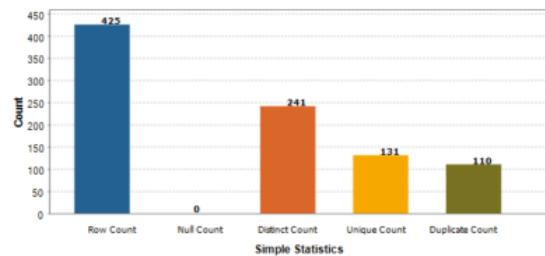
ya que es el límite del campo correspondiente a ese eje. Por lo que los valores se deben encontrar entre esos parámetros [0-80].

Sucede lo mismo que la columna anterior, se pueden observar datos duplicados, ya que existe la posibilidad de que distintos chutes se realicen desde la misma localización, por lo que no es un problema.

• Column: metadata.location_y

• Simple Statistics

Label	Count	%
Row Count	425	100.00%
Null Count	0	0.00%
Distinct Count	241	56.71%
Unique Count	131	30.82%
Duplicate Count	110	25.88%



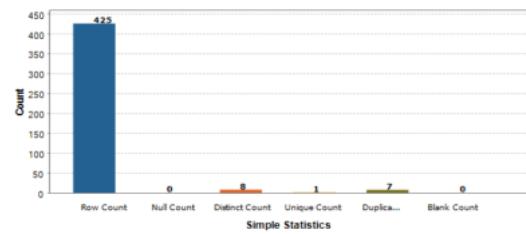
Fuente: Talend Open Studio for Data Quality (Estadísticas básicas “location_y”)

→ **Outcome_name**: En esta columna se encuentran etiquetas que corresponden al resultado del tiro a puerta, por lo que son atributos que pueden repetirse en diferentes eventos.

• Column: metadata.outcome_name

• Simple Statistics

Label	Count	%
Row Count	425	100.00%
Null Count	0	0.00%
Distinct Count	8	1.88%
Unique Count	1	0.24%
Duplicate Count	7	1.65%
Blank Count	0	0.00%



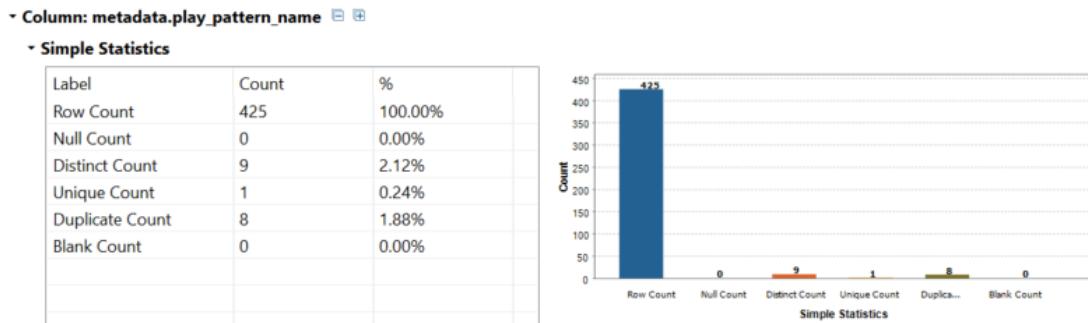
Fuente: Talend Open Studio for Data Quality (Estadísticas básicas “outcome_name”)

Podemos observar cuales pueden ser los posibles resultados de esta columna.

outcome_name
Blocked
Goal
Off T
Post
Saved
Saved Off Target
Saved to Post
Wayward

Fuente: Talend Open Studio for Data Quality

→ **Play_pattern_name:** En esta columna se encuentran etiquetas que corresponden al patrón de juego, por lo que son atributos que pueden repetirse en diferentes eventos.



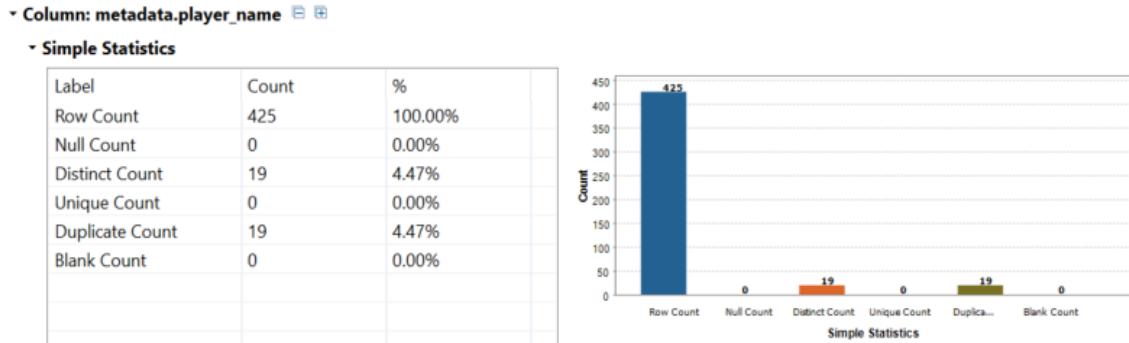
Fuente: Talend Open Studio for Data Quality (Estadísticas básicas “pattern_name”)

Podemos observar cuales pueden ser los posibles resultados de esta columna.

play_pattern_name
From Corner
From Counter
From Free Kick
From Goal Kick
From Keeper
From Kick Off
From Throw In
Other
Regular Play

Fuente: Talend Open Studio for Data Quality

→ **Player_name:** En esta columna observamos distintos duplicados, y es que una plantilla de fútbol puede estar conformada hasta por 25 jugadoras, por lo que es normal que se repitan 19 nombres de jugadoras en los distintos registros.



Fuente: Talend Open Studio for Data Quality (Estadísticas básicas “player_name”)

Podemos observar en “Distinct Count” que existen 19 jugadoras de la plantilla del Manchester City que realizaron por lo menos un disparo.

player_name
Abby Dahlkemper
Alex Greenwood
Caroline Weir
Chloe Kelly
Demi Stokes
Ellen White
Esme Beth Morgan
Gemma Bonner
Georgia Stanway
Janine Elizabeth Beckie
Jessica Park
Jill Scott
Keira Walsh
Laura Coombs
Lauren Hemp
Lucy Bronze
Rosemary Kathleen Lavelle
Samantha June Mewis
Stephanie Houghton

Fuente: Talend Open Studio for Data Quality

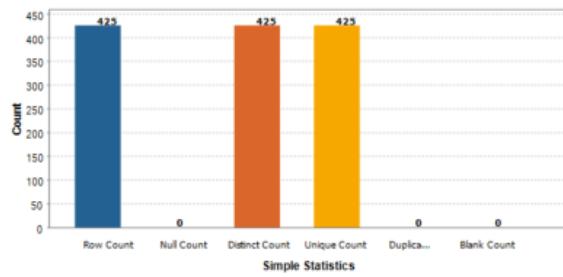
→ **Related_Events_1, Related_Events_2, Related_Events_3:** En estas columnas vamos a analizar el evento previo relacionado al disparo, es importante señalar que hay jugadas que tienen 1, 2, o incluso 3 eventos relacionados.

En este caso, no se debería de repetir ningún registro dado que cada uno de ellos es único en toda la temporada.

• Column: metadata.related_events_1 □ □

• Simple Statistics

Label	Count	%
Row Count	425	100.00%
Null Count	0	0.00%
Distinct Count	425	100.00%
Unique Count	425	100.00%
Duplicate Count	0	0.00%
Blank Count	0	0.00%



Fuente: Talend Open Studio for Data Quality (Estadísticas básicas “related_events_1”)

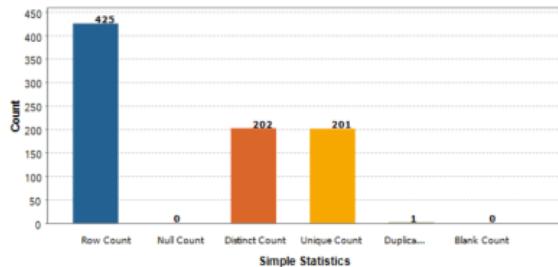
En el caso de las columnas de related_event_2 y related_event_3, el único valor que aparece como duplicado es el valor nulo “Nan”, ya que como se ha mencionado todas

los chutes tienen al menos 1 evento relacionado, pero hay algunos que tienen 2 o incluso 3, motivo por el cuál, van a haber algunos registros nulos, sobretodo en el 3.

- Column: metadata.related_events_2  

- Simple Statistics

Label	Count	%
Row Count	425	100.00%
Null Count	0	0.00%
Distinct Count	202	47.53%
Unique Count	201	47.29%
Duplicate Count	1	0.24%
Blank Count	0	0.00%

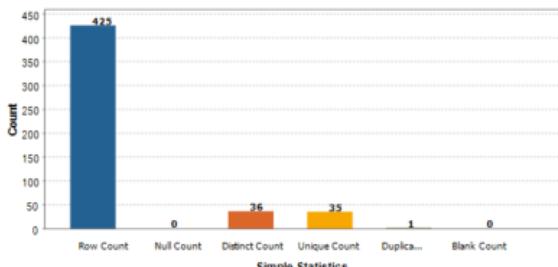


Fuente: Talend Open Studio for Data Quality (Estadísticas básicas “related_events_2”)

- Column: metadata.related_events_3  

- Simple Statistics

Label	Count	%
Row Count	425	100.00%
Null Count	0	0.00%
Distinct Count	36	8.47%
Unique Count	35	8.24%
Duplicate Count	1	0.24%
Blank Count	0	0.00%



Fuente: Talend Open Studio for Data Quality (Estadísticas básicas “related_events_3”)

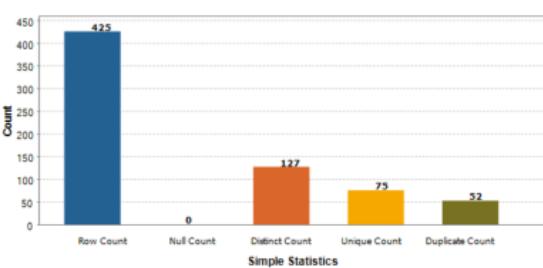
→ **Shot_end_location_x**: Esta columna indica la posición de final del disparo sobre el eje x. El valor máximo para el eje “x” es de 120, ya que es el límite del campo correspondiente a ese eje. Por lo que los valores se deben encontrar entre esos parámetros [0-120].

Se pueden observar datos duplicados, ya que existe la posibilidad de que distintos chutes terminen en la misma localización, por lo que no es un problema.

- Column: metadata.shot_end_location_x  

- Simple Statistics

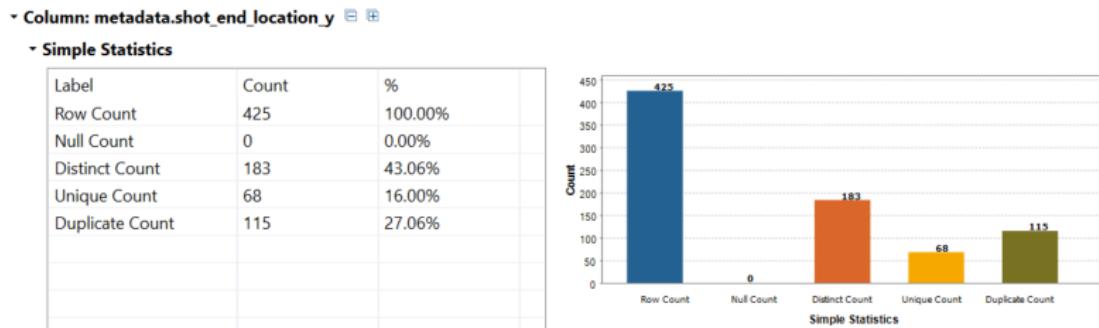
Label	Count	%
Row Count	425	100.00%
Null Count	0	0.00%
Distinct Count	127	29.88%
Unique Count	75	17.65%
Duplicate Count	52	12.24%



Fuente: Talend Open Studio for Data Quality (Estadísticas básicas “shot_end_location_x”)

→ **Shot_end_location_y**: esta columna indica la posición de final del disparo sobre el eje y. El valor máximo para el eje "y" es de 80, ya que es el límite del campo correspondiente a ese eje. Por lo que los valores se deben encontrar entre esos parámetros [0-80].

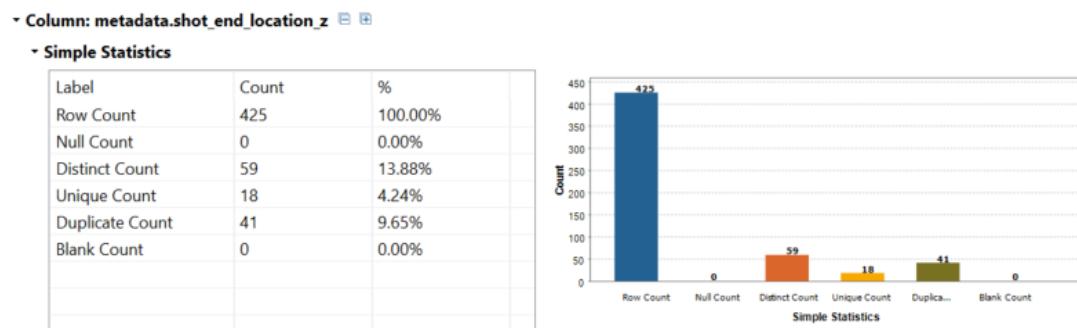
Sucede lo mismo que la columna anterior, se pueden observar datos duplicados, ya que existe la posibilidad de que distintos chutes terminen en la misma localización, por lo que no es un problema.



Fuente: Talend Open Studio for Data Quality (Estadísticas básicas “shot_end_location_y”)

→ **Shot_end_location_z**: esta columna indica la altura en la que termina el disparo, todos los disparos que van hacia el arco tiene que estar entre [0 – 2.67].

Se pueden observar datos duplicados, ya que existe la posibilidad de que la altura del disparo sea la misma.



Fuente: Talend Open Studio for Data Quality (Estadísticas básicas “shot_end_location_z”)

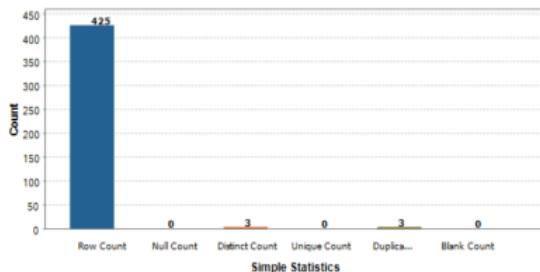
Para las 3 columnas de shot_end_location vamos a encontrar valores nulos dado que esta columna contiene un valor válido en el caso de que el shot_outcome sea “Goal”, para los demás resultados el valor será “NaN”.

→ **Shot_type_name**: En esta columna vamos a encontrar 3 únicos valores que se pueden repetir en diferentes registros ya que corresponden al tipo de chute realizado.

- Column: metadata.shot_type_name  

- Simple Statistics

Label	Count	%
Row Count	425	100.00%
Null Count	0	0.00%
Distinct Count	3	0.71%
Unique Count	0	0.00%
Duplicate Count	3	0.71%
Blank Count	0	0.00%



shot_type_name
Free Kick
Open Play
Penalty

Fuente: Talend Open Studio for Data Quality (Estadísticas básicas “shot_type_name”)

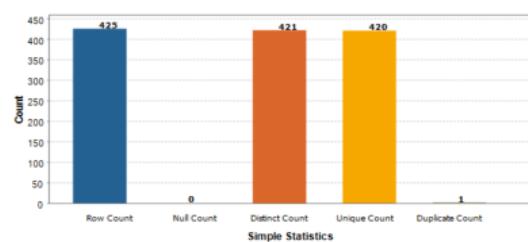
→ **Statsbomb_xG**: Esta columna, expected goal “**xG**” (goles esperados), muestra cuál es la probabilidad de que ese disparo termine en gol.

Cada modelo de xG tiene sus particularidades pero estos son los factores que se usan tradicionalmente para realizar el modelo: distancia a portería, ángulo respecto a portería, parte del cuerpo con la que se realiza el tiro y tipo de asistencia o acción previa (pase en profundidad, centro, balón parado, regate, etc). Con esta información alimentada con una gran cantidad de disparos históricos, el modelo atribuye a cada tiro un valor entre 0 y 1 que expresa la probabilidad de que termine en gol.

- Column: metadata.statsbomb_xg  

- Simple Statistics

Label	Count	%
Row Count	425	100.00%
Null Count	0	0.00%
Distinct Count	421	99.06%
Unique Count	420	98.82%
Duplicate Count	1	0.24%



Fuente: Talend Open Studio for Data Quality (Estadísticas básicas “statsbomb_xg”)

El único valor que se repite es el **xG= 0.76**, ya que corresponde a la probabilidad de gol cuando se ejecuta un penal.

A modo **resumen** sobre la calidad de los datos sobre el dataset de chutes:

- IDIOMA

Mantenemos el inglés de la fuente de datos y no se han detectado valores alfabéticos anómalos.

- DUPLOCADOS

No existen registros duplicados, aunque sí que vemos que existen etiquetas duplicadas en distintas columnas y esto es porque se trata de variables categóricas con valores pre-establecidos que pueden repetirse en los registros.

- COMPLETITUD

No existen registros nulos ni en blanco, lo que sí que observamos son algunos blancos en alguna columna dado el origen JSON de los datos, existen atributos que pertenecen a cada evento, por lo que es normal que nos encontramos con blancos en un registro del que no se pueden obtener valores de algún atributo (ejemplo: si es un chute fuera de portería, no tendremos valores en el atributo de shot_end_location_z).

- VALIDEZ Y CONSISTENCIA

Con las reglas de calidad establecidas anteriormente nos aseguramos que los valores con los que trabajamos sean válidos y consistentes en cuanto a tipo, rango (variables cuantitativas) o conjunto (variables cualitativas).

- TEMPORALIDAD

No se deben actualizar, son eventos puntuales en un momento pasado, por lo que esta condición se cumple.

- PRECISIÓN

La precisión no la podemos evaluar mediante los datos ya que no sabemos cuál es el resultado en la realidad, lo que sí que demostramos es que son válidos y eso nos da cierta confiabilidad.

Podemos observar que una herramienta de calidad de datos es muy útil tanto para evaluar la calidad, como también para conocer un poco mejor los datos con los que estamos trabajando. Algunos problemas que se han detectado, se resuelven inmediatamente al trabajar con los JSON en formato semi-estructurado, aunque como ya hemos comentado con anterioridad, por la condición de pago para el análisis de JSON en TOSDQ, lo convertimos puntualmente en CSV teniendo en cuenta los “falsos problemas” de calidad que nos puede mostrar la herramienta.

Una vez tenemos el análisis de la calidad de nuestros datos y vemos que tienen la calidad necesaria para poder llevar a cabo el proyecto, pasamos a realizar una ampliación del data profiling inicial incluyendo un análisis exploratorio de datos.

5.3.3 Reglas de negocio

En cuanto a los datos, necesitamos establecer unas reglas de negocio, de las cuales van a nacer nuevas necesidades de **transformación de los datos**.

→ Reglas de negocio IT

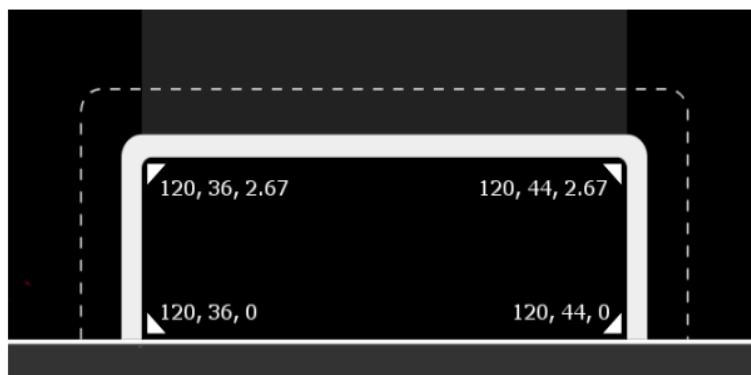
(ID Regla: 1) - Arquitectura: Se van a utilizar herramientas tecnológicas que estén disponibles de manera gratuita en todas las capas de la arquitectura.

(ID Regla: 2) - División de columnas: La localización de las jugadas y el mapa de calor son claves para determinar las jugadas de valor, motivo por el cual es esencial transformar la columna “*location*” en dos, *location_x* y *location_y*, y de igual manera la columna “*shot_end_location*” dividiéndola en tres *shot_end_location_x*, *shot_end_location_y*, *shot_end_location_z*.

→ Reglas de negocio analíticas

(ID Regla: 3) - Chutes que resultan en gol: Consideraremos de máximo valor, aquellos eventos que son chutes y el “*shot_outcome*” es “goal”, por lo que la **X** de la columna “*shot_end_location*” debe de ser 120.

Para profundizar hemos creado una regla de calidad para marcar los casos que el “*shot_outcome*” = “Goal” y la columna “*shot_end_location*” para “**X**” es = “120”, ademas la columna “*shot_end_location*” para “**Y**” debe estar entre “36-44” y la la columna “*shot_end_location*” para “**Z**” debe estar entre “0-2.67”. Caso contrario nos va a marcar los registros erróneos para su revisión.



Fuente: Statsbomb (Coordenadas portería)

```

1 # Crear una lista para almacenar los índices de las filas a excluir
2 indices_a_excluir = []
3
4 # Verificar si algún registro de la columna shot_end_location está fuera de rango.
5 if ((df_goles['shot_end_location_x'] != 120) |
6     (df_goles['shot_end_location_y'] < 36) |
7     (df_goles['shot_end_location_y'] > 44)).any():
8     # Almacenar los índices de las filas que no cumplen las condiciones
9     indices_a_excluir = df_goles[(df_goles['shot_end_location_x'] != 120) |
10                                (df_goles['shot_end_location_y'] < 36) |
11                                (df_goles['shot_end_location_y'] > 44)].index
12 else:
13     print("La columna 'shot_end_location' de todas las filas es válida.")
14 indices_a_excluir

```

La columna 'shot_end_location' de todas las filas es válida.
Out[83]: []

Vemos que todas las filas tienen una "shot_end_location" válida.

(ID Regla: 4) - Chutes entre los 3 palos: Consideraremos de valor, aquellos eventos que son chutes y el "shot_outcome" se sitúa entre los 3 palos de la portería, es decir, "Post", "Saved", "Saved to Post" o "Goal", por lo que la X de la columna "shot_end_location" debe de ser 120 y la Y debe estar entre 36-44.

```

1 # Filtrar eventos del Manchester City WFC
2 chutes_3_palos_manchester_city_wfc = collection3.find({'$or': [{'shot.outcome.name': 'Saved'},
3                                         {'shot.outcome.name': 'Goal'},
4                                         {'shot.outcome.name': 'Post'},
5                                         {'shot.outcome.name': 'Saved to Post'}]}))
6
7 # Convertir el cursor a una lista y obtener el número de documentos
8 documentos_filtrados4 = list(chutes_3_palos_manchester_city_wfc)
9 num_eventos = len(documentos_filtrados4)
10
11 # Verificar si hay documentos
12 if num_eventos > 0:
13     # Crear una nueva colección y borrar los documentos existentes
14     db_nueva_coleccion = client['Statsbomb']
15     nueva_coleccion = db_nueva_coleccion['chutes_3_palos_manchester_city_wfc']
16     nueva_coleccion.delete_many({})
17
18     # Insertar los documentos filtrados en la nueva colección
19     nueva_coleccion.insert_many(documentos_filtrados4)
20 else:
21     print("No se encontraron chutes entre los 3 palos.")

```

Para el posterior análisis de estos chutes, se crea una colección en mongoDB Atlas:

Statsbomb

LOGICAL DATA SIZE: 85.06MB STORAGE SIZE: 26.32MB INDEX SIZE: 3.83MB TOTAL COLLECTIONS: 5

CREATE COLLECTION

Collection Name	Documents	Logical Data Size	Avg Document Size	Storage Size	Indexes	Index Size	Avg Index Size
Events	78953	51.31MB	682B	15.86MB	1	2.31MB	2.31MB
chutes_3_palos_manchester_city_wfc	98	261.33KB	2.67KB	96KB	1	20KB	20KB
chutes_manchester_city_wfc	425	1.15MB	2.77KB	372KB	1	32KB	32KB
eventos_manchester_city_wfc	48969	32.18MB	690B	9.93MB	1	1.45MB	1.45MB
goals_manchester_city_wfc	62	163.59KB	2.64KB	68KB	1	20KB	20KB

(ID Regla: 5) - Location correcta: Verificamos que los valores de la columna “location” son válidos, es decir, que la localización se sitúe dentro del campo. Gracias a estas localizaciones correctas, podremos visualizar el mapa de calor determinando las zonas calientes del campo.

```
1 # Crear una lista para almacenar los indices de las filas a excluir
2 indices_a_excluir = []
3
4 # Verificamos que los valores de la columna “location” son válidos, es decir,
5 # que la localización se sitúe dentro del campo.
6 if ((df_goles['location_x'] < 0) |
7     (df_goles['location_x'] > 120) |
8     (df_goles['location_y'] < 0) |
9     (df_goles['location_y'] > 80)).any():
10    # Almacenar los índices de las filas que no cumplen las condiciones
11    indices_a_excluir = df_goles[(df_goles['location_x'] < 0) |
12                                (df_goles['location_x'] > 120) |
13                                (df_goles['location_y'] < 0) |
14                                (df_goles['location_y'] > 80)].index
15 else:
16    print("Todas las location son están dentro del terreno de juego.")
17 indices_a_excluir
```

```
Todas las location son están dentro del terreno de juego.
Out[88]: []
```

Vemos que todas las filas tienen una “location” válida.

Otras reglas de negocio que planteamos en la probable ampliación futura a corto plazo del proyecto son las siguientes:

(ID Regla: 6) - Pases previos a un chute: Los pases que sean previos a un chute se van a considerar de valor, obteniendo el valor máximo cuando este pase sea una asistencia de gol.

(ID Regla: 7) - Eventos relacionados: Aquellos eventos que se relacionen con el gol, es decir, los eventos que se incluyen dentro de la columna “related_events” se van a considerar como jugadas de valor.

```
1 related_events_ids=[]
2
3 col_chutes=client.Statsbomb.chutes_manchester_city_wfc
4
5 for doc in col_chutes.find():
6     related_events=doc.get('related_events', [])
7     #related_events=doc.related_events
8     related_events_ids.extend(related_events)
9
```

Primero obtenemos la lista de los ids de los eventos relacionados en una lista.

```
1 db = client.Statsbomb
2 original_collection = db.Events
3 new_collection = db.eventos_relacionados_chutes
4
5
6 for document in original_collection.find({'id': {'$in': related_events_ids}}):
7     new_collection.insert_one(document)
8
9
```

Después con esta lista, creamos una colección con todos los eventos relacionados.

(ID Regla: 8) - Proximidad a portería: Las jugadas que ocurren dentro de la location [60-120, 0-80], es decir, de medio campo para delante se consideran de valor de manera progresiva, contra la X se acerque más a 120 y la Y a valores centrales 60-20, más peligro se puede generar.

(ID Regla: 9) - Eventos aislados: Aquellos goles que sean aislados, con características muy diferentes a los demás, los consideraremos eventos aleatorios y no los consideraremos de valor.

(ID Regla: 10) - Pases exitosos: Aquellos pases que se produzcan de $\frac{3}{4}$ de campo hasta la portería, y que lleguen al receptor deseado, los vamos a considerar de alto valor.

5.3.4 Data profiling

En todo proyecto, es esencial desde un inicio realizar un Data Profiling para visualizar cuáles son los datos de los que disponemos y que calidad tienen estos. Realizar este esfuerzo desde el inicio, nos ayuda a determinar la viabilidad del proyecto y evitar sacar conclusiones inválidas por trabajar con datos incorrectos.

Tal como hemos comentado, el proyecto va a basarse en los datos que provee [Statsbomb](#) de manera OpenData, un conjunto de datos en batch retrospectivos, sin posibilidad de modificaciones. Los datos en Statsbomb son de tipo semi-estructurado, y se distribuyen en distintos documentos JSON. Para realizar la vista preliminar de los datos, hemos hecho uso de la librería de Statsbomb en Python importando el módulo sb.

El primer dataset que encontramos es “**Competition Data**” que contiene las siguientes variables:

- Competition_id (integer)
- Season_id (integer)

- Country_name (string)
- Competition_name (string)
- Competition_gender (string)
- Competition_youth (string)
- Competition_international (string)
- Season_name (string)
- Match_updated (datetime)
- Match_updated_360 (datetime)
- Match_available (datetime)
- Match_available_360 (datetime)

competitions.columns

```
Index(['competition_id', 'season_id', 'country_name', 'competition_name',
       'competition_gender', 'competition_youth', 'competition_international',
       'season_name', 'match_updated', 'match_updated_360',
       'match_available_360', 'match_available'],
      dtype='object')
```

En este dataset encontramos un total de 42 registros y 12 columnas, de las cuales las más significativas para nosotros son las 4 primeras y la season_name.

En este caso, como se ha mencionado anteriormente, se ha escogido la FA Women's Super League (competition_name), de la temporada 2020/2021 (season_name), que corresponde al 37 (competition_id) y 90 (season_id).

Una vez recopilada esta información, llamamos al segundo dataset que es el “**Match Data**” que contiene las siguientes variables:

- Match_id (integer)
- Match_date (date)
- Kick_off (time)
- Competition (integer/string)
- Season (integer/string)
- Home_team (integer/string)
- Away_team (integer/string)
- Home_score (integer)
- Away_score (integer)
- Match_status (string)
- Match_status_360 (string)
- Last_updated (datetime)
- Last_updated_360 (datetime)

- Match_week (integer)
- Competition_stage (integer/string)
- Stadium (integer/string)
- Referee (integer/string)
- Away_manager (string)
- Home_manager (string)
- Data_version (string)
- Shot_fidelity_version (integer)
- Xy_fidelity_version (integer)

matches.columns

```
Index(['match_id', 'match_date', 'kick_off', 'competition', 'season',
       'home_team', 'away_team', 'home_score', 'away_score', 'match_status',
       'match_status_360', 'last_updated', 'last_updated_360', 'match_week',
       'competition_stage', 'stadium', 'referee', 'home_managers',
       'away_managers', 'data_version', 'shot_fidelity_version',
       'xy_fidelity_version'],
      dtype='object')
```

En este dataset, existen un total de 131 registros y 22 columnas. En este caso, vamos a filtrar centrandonos únicamente en los partidos disputados por el Manchester City WFC, ya sea en casa o fuera. De esta manera, el dataset resultante contiene 22 registros, con las 22 columnas mencionadas.

De cada partido, también podemos obtener las alineaciones con el dataset “*Lineups*”.

Este dataset contiene las siguientes columnas:

- Player_id (integer)
- Player_name (string)
- PlayerNickname (string)
- Jersey_number (integer)
- Country (string)
- Cards (time/string)
- Positions (integer/string)

lineup.columns

```
Index(['player_id', 'player_name', 'playerNickname', 'jersey_number',
       'country', 'cards', 'positions'],
      dtype='object')
```

Existen un total de 18 registros (convocadas) y 7 columnas.

Y finalmente, otro de los datasets que contiene Statsbomb es el de “*Events*”. Este dataset contiene muchos más datos que cualquier otro ya que recoge todos los eventos

del partido. Es por ello, que pueden haber gran cantidad de registros en comparación con los anteriores (aproximadamente unas 3500-4000 filas). También contiene bastantes columnas (80-110), de las cuales muchas de ellas no van a ser útiles ya que son NaN.

eventos_MC_MU.columns

```
Index(['50_50', 'ball_receipt_outcome', 'ball_recovery_recovery_failure',
       'carry_end_location', 'clearance_aerial_won', 'clearance_body_part',
       'clearance_head', 'clearance_left_foot', 'clearance_right_foot',
       'counterpress', 'dribble_outcome', 'dribble_overrun', 'duel_outcome',
       'duel_type', 'duration', 'foul_committed_advantage',
       'foul_committed_card', 'foul_committed_type', 'foul_won_advantage',
       'foul_won_defensive', 'goalkeeper_body_part', 'goalkeeper_end_location',
       'goalkeeper_outcome', 'goalkeeper_position', 'goalkeeper_technique',
       'goalkeeper_type', 'id', 'index', 'interception_outcome', 'location',
       'match_id', 'minute', 'off_camera', 'out', 'pass_aerial_won',
       'pass_angle', 'pass_assisted_shot_id', 'pass_body_part', 'pass_cross',
       'pass_deflected', 'pass_end_location', 'pass_goal_assist',
       'pass_height', 'pass_inswinging', 'pass_length', 'pass_outcome',
       'pass_recipient', 'pass_shot_assist', 'pass_switch', 'pass_technique',
       'pass_through_ball', 'pass_type', 'period', 'play_pattern', 'player',
       'player_id', 'position', 'possession', 'possession_team',
       'possession_team_id', 'related_events', 'second', 'shot_aerial_won',
       'shot_body_part', 'shot_end_location', 'shot_first_time',
       'shot_freeze_frame', 'shot_key_pass_id', 'shot_one_on_one',
       'shot_open_goal', 'shot_outcome', 'shot_statsbomb_xg', 'shot_technique',
       'shot_type', 'substitution_outcome', 'substitution_replacement',
       'tactics', 'team', 'timestamp', 'type', 'under_pressure'],
      dtype='object')
```

Tal y como se ha comentado anteriormente, Statsbomb provee más datasets como por ejemplo **“Aggregated Stats”**, con las credenciales pertinentes mediante pago.

Si bien, este Data Profiling se ha realizado de manera precoz al iniciar el proyecto, es un proceso dinámico que se va realizando a lo largo del proyecto, para asegurarnos de que estos datos tienen la calidad que el proyecto requiere.

5.3.5 EDA

Previamente se había realizado un data profiling preliminar en Jupyter, el cuál mediante el análisis exploratorio de datos (EDA) realizado posteriormente, ha sido ampliado. Este EDA se ha realizado en Databricks Community mediante el lenguaje Python, lo que nos ha permitido extraer datos interesantes para el dashboard, así como también reconocer patrones, y empezar a pensar en el modelado.

Después de aplicar los filtros y transformaciones, disponemos de 425 documentos de disparos del Manchester City WFC durante la temporada 2020/21. En ellos, tenemos los siguientes campos y el tipo de datos que contienen:

```
Campo: _id, Tipo de dato: <class 'bson.objectid.ObjectId'>
Campo: id, Tipo de dato: <class 'str'>
Campo: index, Tipo de dato: <class 'int'>
Campo: period, Tipo de dato: <class 'int'>
Campo: timestamp, Tipo de dato: <class 'str'>
Campo: minute, Tipo de dato: <class 'int'>
Campo: second, Tipo de dato: <class 'int'>
Campo: type.id, Tipo de dato: <class 'int'>
Campo: type.name, Tipo de dato: <class 'str'>
Campo: possession, Tipo de dato: <class 'int'>
Campo: possession_team.id, Tipo de dato: <class 'int'>
Campo: possession_team.name, Tipo de dato: <class 'str'>
Campo: play_pattern.id, Tipo de dato: <class 'int'>
Campo: play_pattern.name, Tipo de dato: <class 'str'>
Campo: team.id, Tipo de dato: <class 'int'>
Campo: team.name, Tipo de dato: <class 'str'>
Campo: player.id, Tipo de dato: <class 'int'>
Campo: player.name, Tipo de dato: <class 'str'>
Campo: position.id, Tipo de dato: <class 'int'>
Campo: position.name, Tipo de dato: <class 'str'>
Campo: location, Tipo de dato: ["<class 'float'>", "<class 'float'>"]

Campo: duration, Tipo de dato: <class 'float'>
Campo: related_events, Tipo de dato: [<class 'str'>, <class 'str'>]
Campo: shot.one_on_one, Tipo de dato: <class 'bool'>
Campo: shot.open_goal, Tipo de dato: <class 'bool'>
Campo: shot.statsbomb_xg, Tipo de dato: <class 'float'>
Campo: shot.end_location, Tipo de dato: [<class 'float'>, <class 'float'>, <class 'float'>]
Campo: shot.outcome.id, Tipo de dato: <class 'int'>
Campo: shot.outcome.name, Tipo de dato: <class 'str'>
Campo: shot.technique.id, Tipo de dato: <class 'int'>
Campo: shot.technique.name, Tipo de dato: <class 'str'>
Campo: shot.body_part.id, Tipo de dato: <class 'int'>
Campo: shot.body_part.name, Tipo de dato: <class 'str'>
Campo: shot.type.id, Tipo de dato: <class 'int'>
Campo: shot.type.name, Tipo de dato: <class 'str'>
Campo: shot.freeze_frame, Tipo de dato: [<class 'dict'>, <class 'dict'>, <class 'dict'>, <class 'dict'>]
Campo: shot.key_pass_id, Tipo de dato: <class 'str'>
Campo: shot.first_time, Tipo de dato: <class 'bool'>
Campo: under_pressure, Tipo de dato: <class 'bool'>
Campo: shot.aerial_won, Tipo de dato: <class 'bool'>
Campo: shot.deflected, Tipo de dato: <class 'bool'>
```

Como podemos observar la mayoría de variables que vamos a utilizar en el proyecto son de tipo categóricas. Algunos de estos campos forman parte de otros, por lo que tenemos arrays y diccionarios.

Arrays:

- location: formado por el valor del eje x y del eje y.
- related_events: formado por los eventos relacionados.
- shot.end_location: formado por los valores del eje x, del eje y y del eje z.
- shot.freeze_frame: está formado por más diccionarios dentro de más diccionarios, pero es un dato que no utilizaremos en el proyecto.

Diccionarios:

- type: formado por id y name.

- possession_team: formado por id y name.
- play_pattern: formado por id y name.
- team: formado por id y name.
- player: formado por id y name.
- position: formado por id y name.
- shot: formado por one_on_one, open_goal, statsbomb_xg, end_location, outcome, technique, body_part, type y freeze_frame.

En función de nuestro objetivo, de todos estos campos, nosotros nos centraremos en el “*shot.outcome*” como variable dependiente y utilizaremos:

- Periodo (“period”)
- Localización inicial (“location”)
- Patrón de juego (“play_pattern.name”)
- Jugadora (“player.name”)
- Posición de juego (“position.name”)
- Localización final del chute (“shot_end_location”)
- Tipo de chute (“shot.type.name”)
- Técnica de chute (“shot.technique.name”)
- Campo (local o visitante) y Rival
- Parte del cuerpo involucrada (“shot.body_part.name”)

Desarrollamos una matriz de correlación para una vista rápida de la correlación de estas variables entre ellas:

	shot_outcome	period	location	play_pattern	player	position	shot_end_location	shot_type	shot_technique	shot_body_part
shot_outcome	1.0	0.0	0.0	0.0	0.0	0.0	0.129563	0.143094	0.0	0.143914
period	0.0	1.0	0.068761	0.065148	0.0	0.087979	0.027909	0.0	0.057498	0.0
location	0.0	0.068761	1.0	0.0	0.121047	0.098568	0.0	0.068843	0.0	0.043722
play_pattern	0.0	0.065148	0.0	1.0	0.081823	0.112696	0.042117	0.75635	0.052123	0.203129
player	0.0	0.0	0.121047	0.081823	1.0	0.732787	0.17017	0.194424	0.0	0.351608
position	0.0	0.087979	0.098568	0.112696	0.732787	1.0	0.124549	0.128372	0.0	0.243045
shot_end_location	0.129563	0.027909	0.0	0.042117	0.17017	0.124549	1.0	0.128793	0.098364	0.07799
shot_type	0.143094	0.0	0.068843	0.75635	0.194424	0.128372	0.128793	1.0	0.0	0.0
shot_technique	0.0	0.057498	0.0	0.052123	0.0	0.0	0.098364	0.0	1.0	0.112015
shot_body_part	0.143914	0.0	0.043722	0.203129	0.351608	0.243045	0.07799	0.0	0.112015	1.0

Fuente: Tabla de elaboración propia en Databricks Community (Matriz de Correlación)

A raíz de esta matriz, a priori, no se observan correlaciones muy fuertes entre las distintas variables, por tanto, **bajo riesgo de multicolinealidad**. Podemos concluir que:

- La columna “*play_pattern*” muestra una correlación moderada con la columna “*shot_type*” (0.75635), lo que indica una relación considerable entre el patrón de juego y el tipo de disparo realizado.
- La columna “*player*” tiene una correlación moderada con la columna “*position*” (0.732787), lo que indica una fuerte relación entre la jugadora y su posición en el campo (esperado).
- La columna “*shot_body_part*” muestra una correlación moderada con la columna “*player*” (0.351608), lo que indica cierta relación entre la parte del cuerpo utilizada para el disparo y la jugadora involucrada.

A continuación, tenemos el resumen de todas las columnas del dataframe de chutes, los campos completos y el tipo de dato que contiene cada una.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 425 entries, 0 to 424
Data columns (total 41 columns):
 #   Column            Non-Null Count  Dtype  
--- 
 0   _id                425 non-null    object 
 1   id                 425 non-null    object 
 2   index              425 non-null    int64  
 3   period             425 non-null    int64  
 4   timestamp          425 non-null    object 
 5   minute             425 non-null    int64  
 6   second             425 non-null    int64  
 7   possession          425 non-null    int64  
 8   duration            425 non-null    float64
 9   type.id            425 non-null    int64  
 10  type.name           425 non-null    object 
 11  possession_team.id 425 non-null    int64  
 12  possession_team.name 425 non-null    object 
 13  play_pattern.id    425 non-null    int64  
 14  play_pattern.name   425 non-null    object 
 15  team.id            425 non-null    int64  
 16  team.name           425 non-null    object 
 17  player.id           425 non-null    int64  
 18  player.name          425 non-null    object 
 19  position.id          425 non-null    int64  
 20  position.name         425 non-null    object 
 21  shot.statsbomb_xg    425 non-null    float64
 22  shot.key_pass_id     293 non-null    object 
 23  shot.outcome.id      425 non-null    int64  
 24  shot.outcome.name     425 non-null    object 
 25  shot.type.id          425 non-null    int64  
 26  shot.type.name         425 non-null    object 
 27  shot.technique.id      425 non-null    int64  
 28  shot.technique.name     425 non-null    object 
 29  shot.body_part.id      425 non-null    int64  
 30  shot.body_part.name     425 non-null    object 
 31  shot.freeze_frame      422 non-null    object 
 32  under_pressure          113 non-null    object 
 33  location_x             425 non-null    float64
 34  location_y             425 non-null    float64
 35  shot_end_location_x     425 non-null    float64
 36  shot_end_location_y     425 non-null    float64
 37  shot_end_location_z     293 non-null    float64
 38  related_events1         425 non-null    object 
 39  related_events2         201 non-null    object 
 40  related_events3         35 non-null     object 
dtypes: float64(7), int64(15), object(19)
memory usage: 136.3+ KB
```

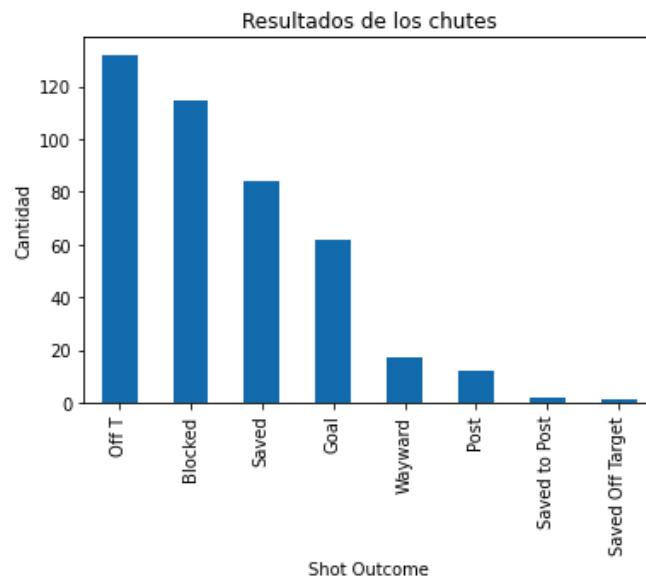
Además, para la obtención de la variable independiente campo y rival, creamos dos nuevas variables “campo” y “rival” en el archivo de matches_City, y seguidamente hacemos un merge con el archivo de los goles.

```
Goles_campoYrival=goles_City_limpios.merge(matches_City, on='match_id', how='left')  
Goles_campoYrival
```

En este punto del proyecto, nos encontramos en la fase final de la preparación de los datos, a punto de adentrarnos en el análisis profundo de los datos.

5.3.6 Reconocimiento de patrones

Uno de los primeros aspectos importantes en este caso, es saber cuantos chutes se han realizado en esta temporada, y cuáles de ellos se han materializado en gol. Para ello, primero hemos analizado cuales son las finalizaciones de los chutes.



Fuente: Gráfico de elaboración propia en Databricks Community (Resumen chutes)

shot_outcome	shot_outcome									Total
	Count	Blocked	Goal	Off T	Post	Saved	Saved Off Target	Saved to Post	Wayward	
Blocked	115.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	115.0
Goal	0.0	62.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	62.0
Off T	0.0	0.0	132.0	0.0	0.0	0.0	0.0	0.0	0.0	132.0
Post	0.0	0.0	0.0	12.0	0.0	0.0	0.0	0.0	0.0	12.0
Saved	0.0	0.0	0.0	0.0	84.0	0.0	0.0	0.0	0.0	84.0
Saved Off Target	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
Saved to Post	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	2.0
Wayward	0.0	0.0	0.0	0.0	0.0	0.0	0.0	17.0	0.0	17.0
Total	115.0	62.0	132.0	12.0	84.0	1.0	2.0	17.0	425.0	

Fuente: Tabla de elaboración propia en Databricks Community (Resumen chutes)

Aquí podemos observar que de los 425 chutes a puerta en toda la temporada, la gran mayoría finalizan fuera de portería, despejados por un defensa o parados por un portero; y a continuación les sigue la finalización en gol. En esta línea, si observamos cuál es la cantidad de disparos que van a puerta, es decir, que se sitúan entre los 3 palos, gracias a la colección creada con anterioridad, vemos que de los 425 chutes solo 160 se han dirigido a puerta, situándose entre los 3 palos.



Fuente: Gráfico de elaboración propia en Databricks Community (Chutes entre los 3 palos)

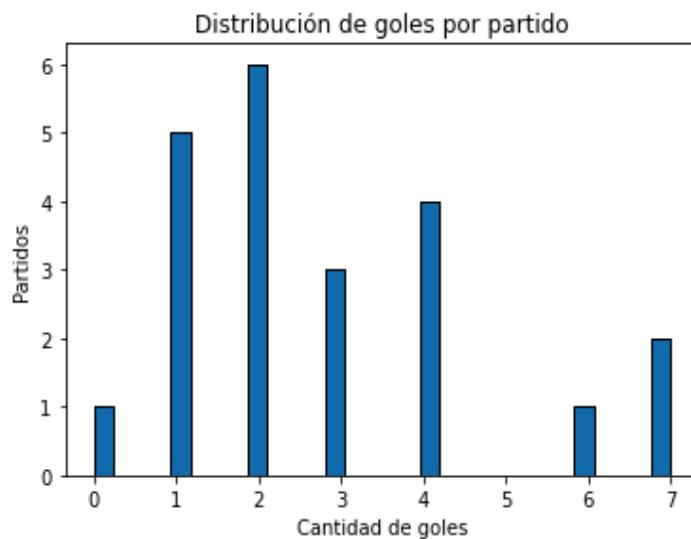
Dentro de esta pequeña proporción de chutes que van entre los 3 palos, solo 62 chutes se han materializado en gol. Así pues, tal y como se muestra, hay una baja efectividad del equipo en cuanto a gol. De manera general, en el fútbol la tasa de conversión varía mucho entre equipos y jugadores, por lo que no se puede establecer un valor estándar, pero consideramos que el valor obtenido es un valor medio tirando a bajo.



Fuente: Gráfico de elaboración propia en Databricks Community (Tasa de conversión a gol)

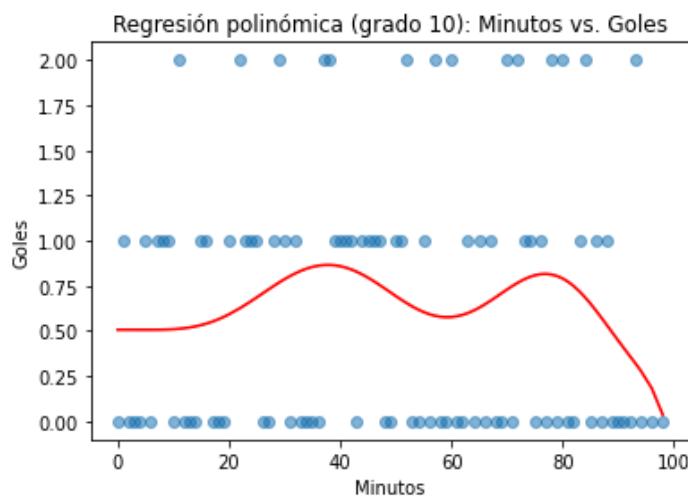
Una vez tenemos un poco de información sobre los chutes, creemos conveniente analizar las características de los goles.

Cabe mencionar que la media de goles por partido fue de 2,82 goles. En el siguiente gráfico podemos observar la distribución de goles por partido, donde vemos principalmente que lo más habitual es marcar de 1 a 4 goles por partido.



Fuente: Gráfico de elaboración propia en Databricks Community (Distribución goles por partido)

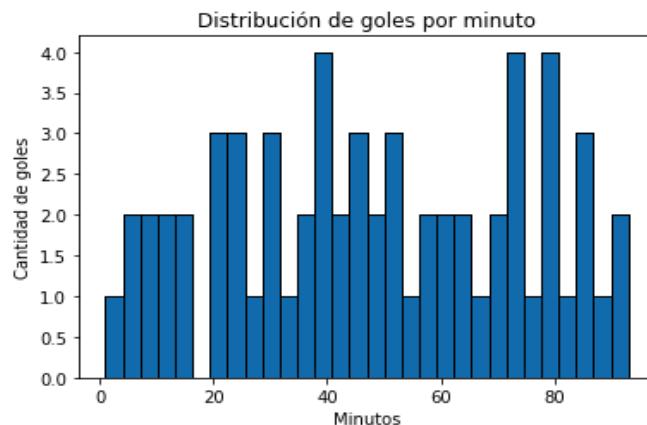
Un patrón interesante encontrado es la distribución de los goles por minuto. Para analizar esto, construimos una regresión polinómica de grado 10 con el siguiente resultado:



Fuente: Gráfico de elaboración propia en Databricks Community (Goles según el minuto)

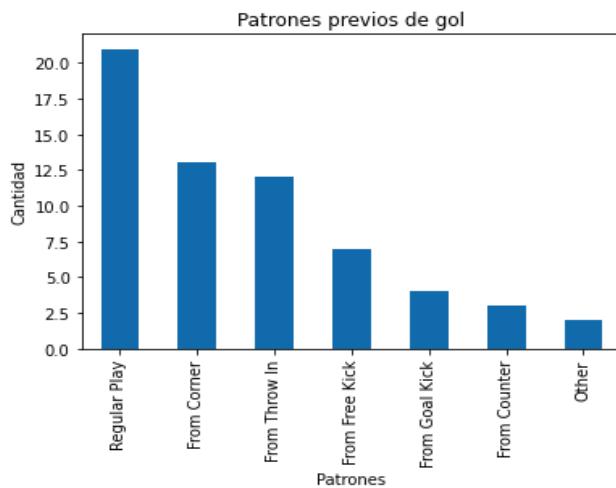
Observamos claramente que cerca del minuto 40 y del minuto 80 hay se anotan una mayor cantidad de goles; y estos minutos corresponden al final de los dos períodos de un partido, pudiendo deducir que la fatiga del rival cerca del final o bien la intensidad del equipo sabiendo que se acerca el descanso, juegan a favor del City.

En el siguiente gráfico se pueden observar la frecuencia de goles por minuto de manera exhaustiva:



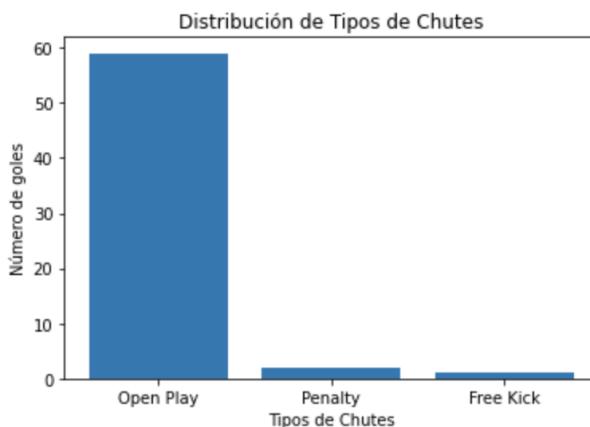
Fuente: Gráfico de elaboración propia en Databricks Community (Goles según el minuto)

→ **Patrones previos a gol:** podemos ver en el gráfico que el patrón que se repite con diferencia antes de un gol es el juego regular, seguido del córner y saque de banda.



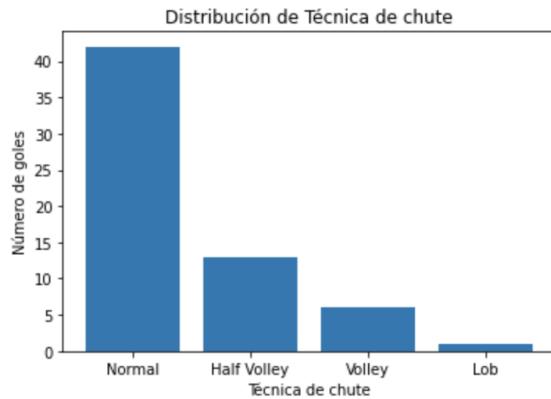
Fuente: Gráfico de elaboración propia en Databricks Community (Patrones previos de gol)

Tipo de chute: podemos ver en el gráfico cuáles son los tipos de chutes más comunes que finalizan en gol. La mayoría provienen de jugada abierta.



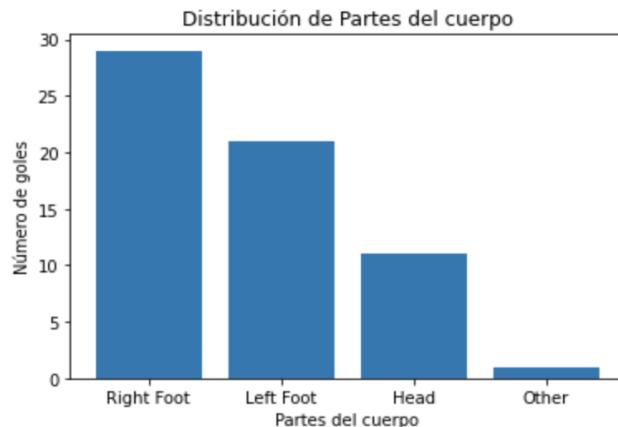
Fuente: Gráfico de elaboración propia en Databricks Community (Distribución por tipo de chute)

→ **Técnica de chute:** podemos ver en el gráfico cuáles son las técnicas de chutes más comunes que finalizan en gol. Lo más común son los chutes normales seguido de la media bolea.



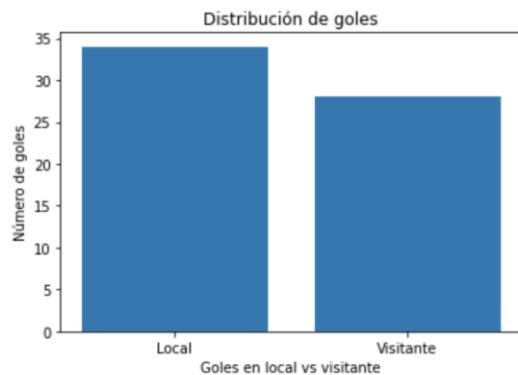
Fuente: Gráfico de elaboración propia en Databricks Community (Distribución por técnica de chute)

→ **Partes del cuerpo:** podemos ver en el gráfico cuáles son las partes del cuerpo más involucradas en los goles.

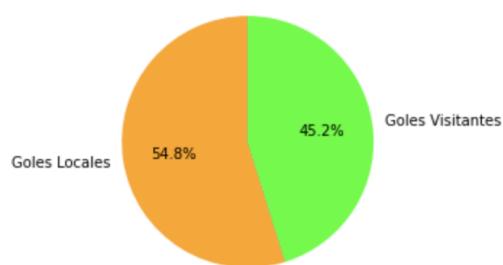


Fuente: Gráfico de elaboración propia en Databricks Community (Distribución por parte del cuerpo)

→ **Local vs Visitante:** podemos ver en el gráfico donde se marcan más goles, si el echo de jugar en casa puede influir, y vemos que sí que se aprecia una diferencia pero no es significativa.

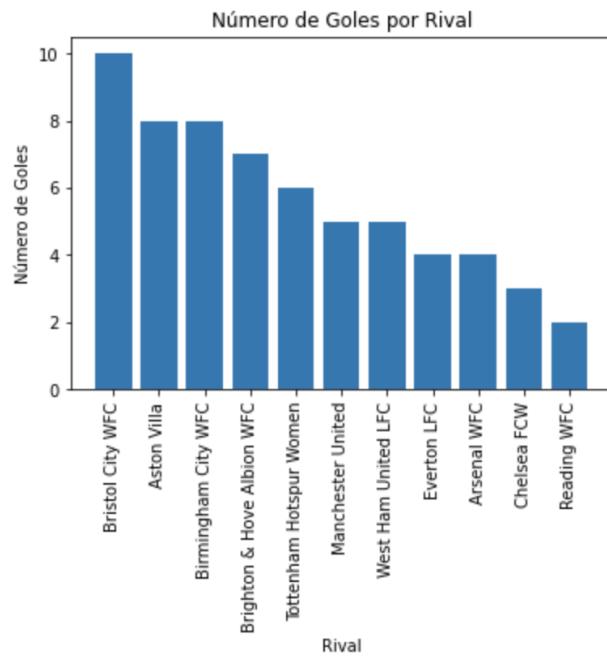


Distribución de Goles: Locales vs Visitantes



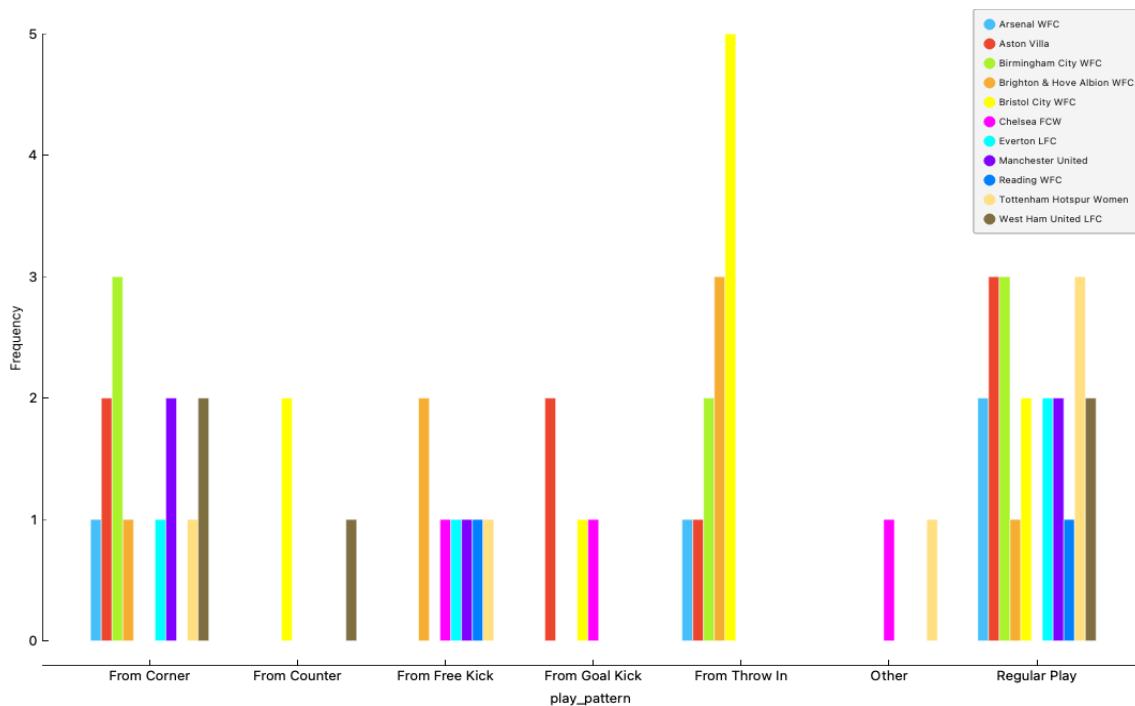
Fuente: Gráficos de elaboración propia en Databricks Community (Local vs Visitante)

→ **Equipos y goles:** podemos ver en el gráfico la cantidad de goles que se ha marcado a cada equipo rival. Esta distribución parece lógica ya que los menos goleados son rivales más directos, situados de media tabla hacia arriba, y los más goleados se sitúan de media tabla hacia abajo.



Fuente: Gráfico de elaboración propia en Databricks Community (Goles anotados por rival)

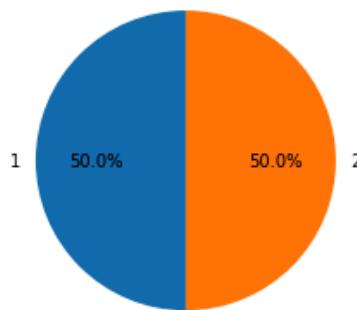
En el siguiente gráfico visualizamos los patrones de juego que más goles han proporcionado contra los rivales específicos.



Fuente: Gráfico de elaboración propia en Databricks Community (Goles y patrones de juego por rival)

→ **Período de partido:** curiosamente, en analizar la distribución de goles por período observamos que se marcan los mismos goles en la primera parte que en la segunda. De aquí podemos sacar conclusiones varias como por ejemplo que la fatiga de la segunda mitad no es muy elevada, o que los cambios producidos en la segunda parte no son muy relevantes en cuanto a goles.

Distribución de goles por período



Fuente: Gráfico de elaboración propia en Databricks Community (Goles por período)

→ **Jugadoras:** es de nuestro interés saber que jugadoras marcan más goles ya que es un aspecto importante. En las tablas siguientes vemos que no hay ninguna jugadora que destaque en cuanto a gol, por lo que no tenemos una “killer”. Y no solo es importante los goles que mete cada jugadora, sino también los chutes que realiza, porque la tasa de conversión de chute a gol, es lo que nos muestra la efectividad de las jugadoras.

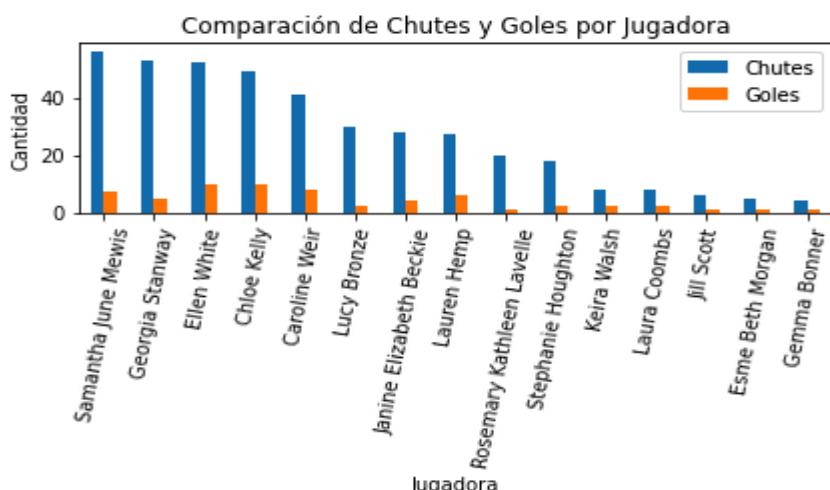
	player.name	chutes
17	Samantha June Mewis	56
8	Georgia Stanway	53
5	Ellen White	52
3	Chloe Kelly	49
2	Caroline Weir	41
15	Lucy Bronze	30
9	Janine Elizabeth Beckie	28
14	Lauren Hemp	27
16	Rosemary Kathleen Lavelle	20
18	Stephanie Houghton	18
1	Alex Greenwood	11
12	Keira Walsh	8
13	Laura Coombs	8
11	Jill Scott	6
6	Esme Beth Morgan	5
7	Gemma Bonner	4
4	Demi Stokes	4
10	Jessica Park	3
12	Abby Dahlkemper	2

	player.name	goles
1	Chloe Kelly	10
2	Ellen White	10
0	Caroline Weir	8
13	Samantha June Mewis	7
10	Lauren Hemp	6
5	Georgia Stanway	5
6	Janine Elizabeth Beckie	4
8	Keira Walsh	2
9	Laura Coombs	2
11	Lucy Bronze	2
14	Stephanie Houghton	2
3	Esme Beth Morgan	1
4	Gemma Bonner	1
7	Jill Scott	1
12	Rosemary Kathleen Lavelle	1

Fuente: Tablas de elaboración propia en Databricks Community (Goles por jugadora)

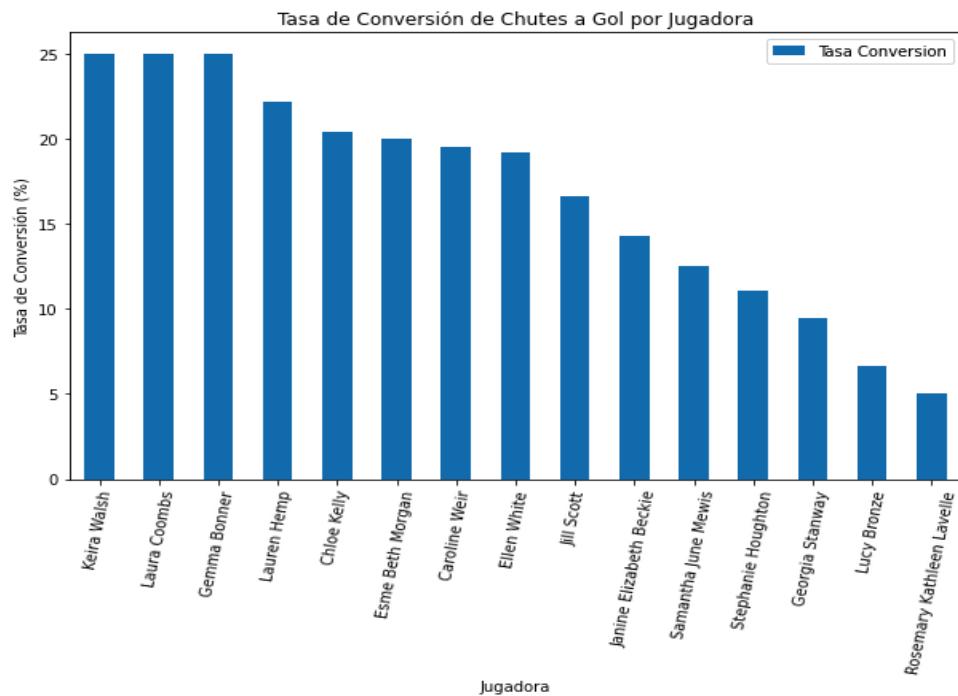
En el [anexo 2](#), podemos visualizar la plantilla completa con sus respectivas posiciones. En esta imagen podemos visualizar que jugadoras componían la plantilla del Manchester City WFC en la temporada 2020-2021, y sus respectivas posiciones. Esto nos será de utilidad para poder filtrar más específicamente en el Dashboard.

En el gráfico siguiente podemos ver que cantidad de chutes y cuales resultan en gol por cada jugadora.



Fuente: Gráfico de elaboración propia en Databricks Community (Chutes por jugadora)

De este gráfico podemos extraer la tasa de conversión de chutes a gol por jugadora, dando paso a observar claramente cuales son las jugadoras más efectivas de cara a gol.



Fuente: Gráfico de elaboración propia en Databricks Community (Tasa de conversión por jugadora)

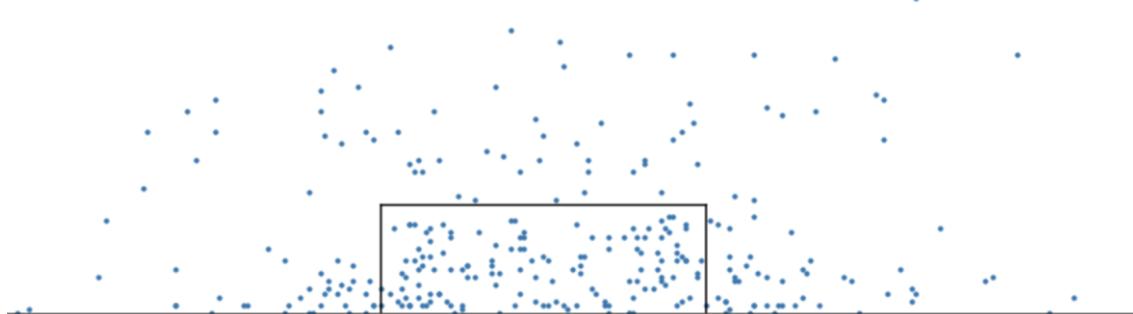
Según estos datos podemos concluir que las jugadoras que mayor tasa de conversión de chutes a gol tienen, realizan una pequeña cantidad de chutes en comparación a otras. Esto es porque estas jugadoras tienen un perfil y posición más defensiva. La conclusión que realmente extraemos de esto, es que las delanteras principales del equipo (Chloe Kelly, Ellen White, Janine Beckie, Georgina Stanway y Lauren Hemp) tienen una tasa de conversión más bien baja, por tanto, poca efectividad.

→ **Goles esperados por posición:** otro aspecto importante relacionado con las jugadoras es analizar qué posiciones tienen probabilidad de gol más alta.

	Posiciones	Promedio_xG
0	Right Midfield	0.287292
1	Right Wing	0.188242
2	Center Forward	0.175494
3	Left Wing	0.153868
4	Left Center Midfield	0.141141
5	Right Center Midfield	0.114905
6	Left Center Back	0.104418
7	Right Center Back	0.076987
8	Right Back	0.074802
9	Center Defensive Midfield	0.067901
10	Center Attacking Midfield	0.067355
11	Left Back	0.038781

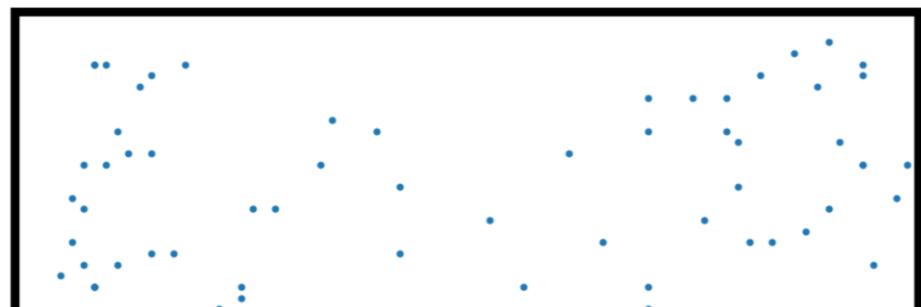
Fuente: Tabla de elaboración propia en Databricks Community (XG por posición)

→ **Localización final del chute:** nos interesa ver dónde acaban los chutes respecto a la portería, para determinar qué localizaciones de la portería nos llevan a marcar más goles.



Fuente: Gráfico de elaboración propia en Databricks Community (Localización de chutes)

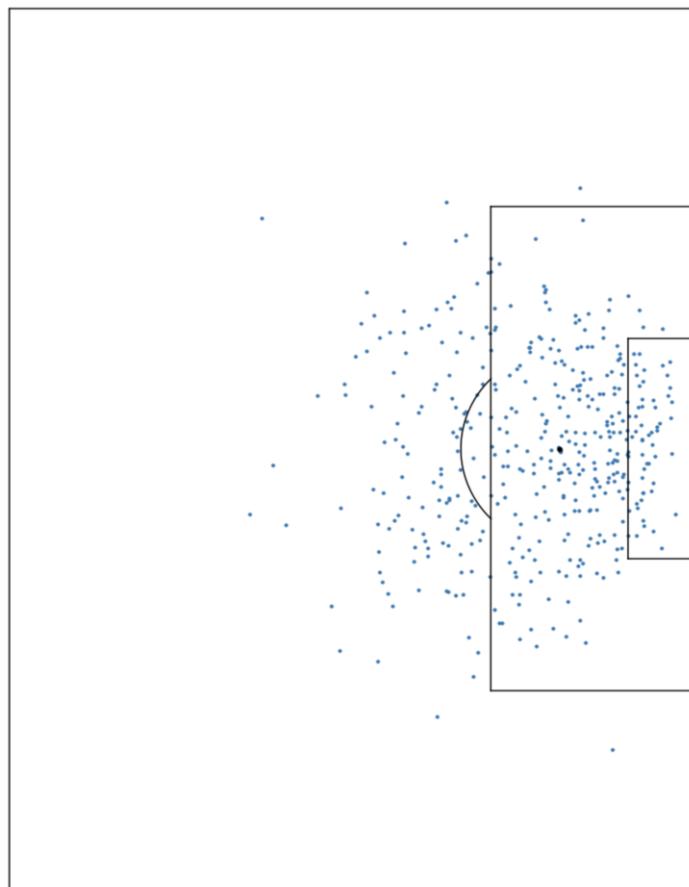
A simple vista vemos una distribución de los chutes bastante uniforme, pero si solo representamos los goles vemos el siguiente gráfico:



Fuente: Gráfico de elaboración propia en Databricks Community (Localización de goles)

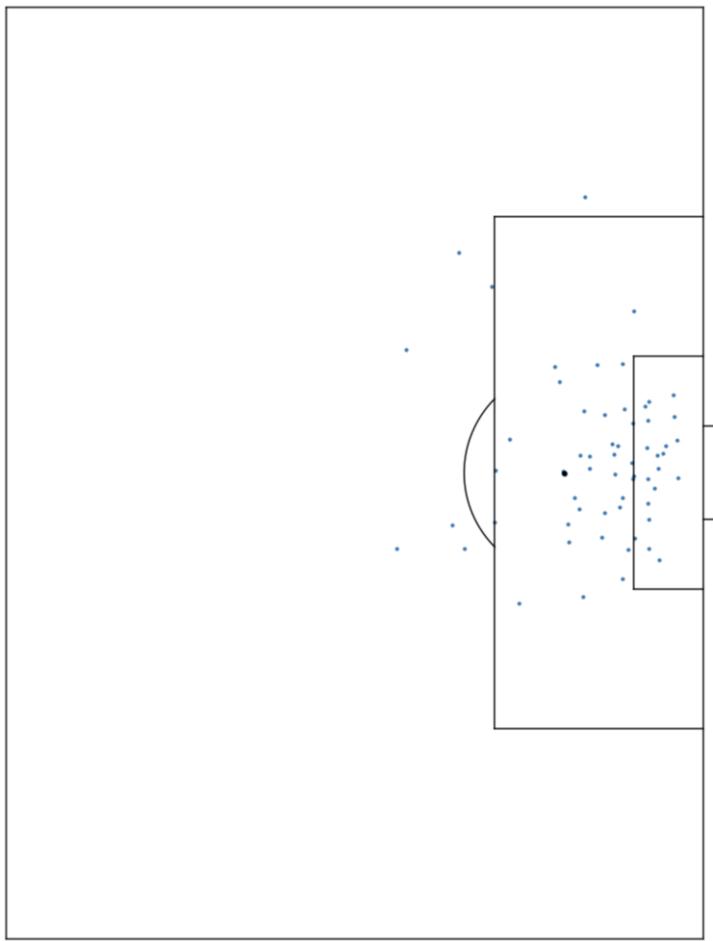
Como es normal, por el centro de la portería se visualizan menos goles que por los costados, al ser la posición principal de las porteras.

→ **Localización en el campo del chute:** también es de especial interés localizar las acciones en el campo de juego, para validar desde qué zonas son más peligrosos los chutes realizados:



Fuente: Gráfico de elaboración propia en Databricks Community (Posición inicial de chute)

Y aquí la localización de los goles:



Fuente: Gráfico de elaboración propia en Databricks Community (Posición inicial de los goles)

Se ve más concentración de goles en la zona central del área, cerca de la portería, con solo 8 goles fuera del área durante la temporada.

Como resumen de todas estas estadísticas, podemos extraer la siguiente información:

- Hemos visto un factor de conversión de chutes entre los tres palos del 37,6% y del 14,6% de cara a gol. Se debería buscar el total de la liga para tener un factor de referencia y poder determinar el valor de los chutes del equipo.
- Centrándonos en los goles, la jugada que determinamos de máximo valor, tenemos:
 - o Mayor cantidad de goles alrededor de los minutos 40 y 80 del partido. Se debería investigar qué ocurre en esos minutos.
 - o Se marcan los mismos goles en la primera que en la segunda parte, por lo que esta variable no parece significativa.
 - o Distribución ligeramente superior de goles como local.
 - o La mayoría de goles anotados son en jugada regular, con el tipo de chute en jugada abierta, con técnica de chute normal y con el pie derecho.

Estas 4 categorías entran dentro de lo esperado, al ser lo más repetido en un partido de fútbol. Para encontrar realmente qué variables nos aportan valor, se debería buscar el porcentaje relativo al total de acciones de cada variable. Por ejemplo, si tenemos el máximo de goles en jugada regular, pero con un % de conversión a gol bajo, quizá sería interesante potenciar la obtención de córner o faltas que pueden tener un % de conversión más alto.

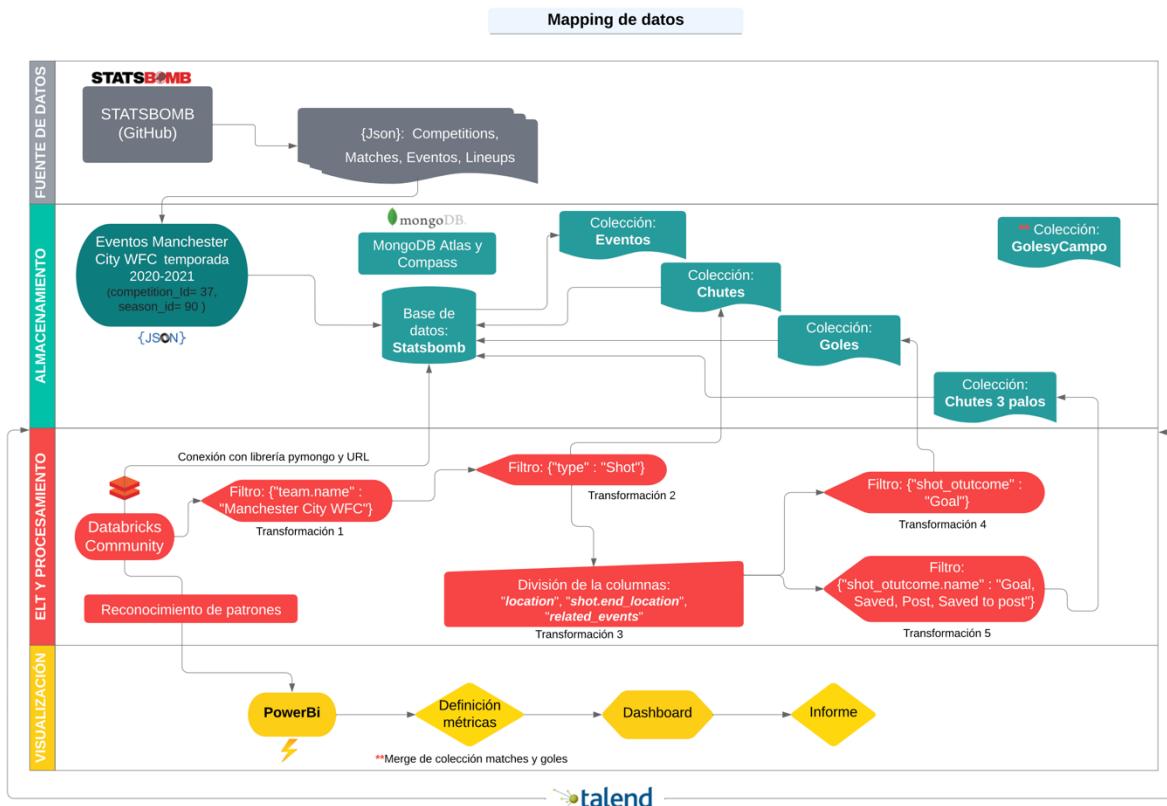
- Según el rival, se observa como los partidos contra equipos peor clasificados en la liga reciben más goles del City, mientras que los rivales directos de media tabla hacia arriba reciben menos.
- El gol parece muy repartido en el equipo, ninguna jugadora destaca en el aspecto goleador. Si que vemos más chutes de las delanteras del equipo, pero la conversión a gol no parece alta. Se debería investigar quizá con la efectividad media de la liga, aplicando un mínimo de chutes significativo, ya que las jugadoras más defensivas realizan muy pocos chutes.
- La posición con más goles esperados es la de centrocampista derecho, con un valor bastante más alto que las otras posiciones.
- Tenemos una distribución bastante uniforme en cuanto a la localización de los chutes respecto a la portería, mientras que en los goles predominan los costados de la portería.
- Tenemos una distribución de chutes a gol, en cuanto a la localización en el campo, cerca o dentro del área rival, concentrándose la gran mayoría en la parte central y dentro del área.

A partir de estos datos, sin aún haber aplicado ningún modelo avanzado, podemos descubrir pequeños patrones que nos guían para encarar el análisis profundo. Disponemos de muchas variables que influyen en los chutes a gol, por lo que determinar cuáles serán más importantes y cómo las relacionamos entre ellas, será determinante para obtener un modelo fiable y preciso.

Como ya hemos comentado, utilizamos una metodología iterable en cualquier parte del proceso, por lo que en este punto realizamos varios supuestos y pruebas para definir parámetros y métricas interesantes y de valor, las cuales vamos a especificar en el siguiente apartado.

5.4 Mapping de datos

Para nuestro proyecto es fundamental conocer los procesos de integración y mapeo de los orígenes de los datos, así como los destinos. Para ello hemos desarrollado con Lucidchart el siguiente diagrama:



Fuente: Gráfico de elaboración propia en Lucidchart (Mapping de datos)

A modo resumen explícito del diagrama, cabe comentar las herramientas y transformaciones principales por las que han pasado los datos:

- **Fuente origen:** GitHub de Statsbomb mediante los Json que facilitan.
- **Json “Events”:** los eventos de cada partido vienen almacenados en formato Json todos en una misma carpeta. Seleccionamos aquellos de competicion_id=30 y season_id=90 dado que son los que corresponden a los eventos del Manchester City femenino en la temporada 2020-2021.
- **Base de datos “Statsbomb”:** creamos en MongoDB una base de datos llamada *TFM* con una colección llamada *Eventos* en la cual almacenamos los Json de los partidos mencionados anteriormente.
- **Databricks Community:** creamos conexión directa a la base de datos Statsbomb, colección Eventos para hacer los siguientes filtros:
 - **Transformación 1 (“team.name”: “Manchester City WFC”):** filtramos los eventos para extraer solo los que pertenezcan al City.

- **Transformación 2** (“*type*”:”*Shot*”): filtramos los eventos para extraer solo los que han sido chute.
 - **Transformación 3** (*División de columnas*): para el data quality en TOSDQ creamos un dataframe con los chutes, que requerirá una división de las columnas location (2 columnas nuevas: *location_x*, *location_y*), *shot.end_location* (3 columnas nuevas: *shot.end_location_x*, *shot.end_location_y*, *shot.end_location_z*) y related_events (3 columnas nuevas: *related_events1*, *related_events2*, *related_events3*) para evitar que aparezca como arrays.
 - **Transformación 4** (“*shot.outcome*”:”*Goal*”): filtramos los eventos para extraer solo los que han finalizado en gol.
 - **Transformación 5** (“*shot.outcome*”:”*Goal, Saved, Post, Saved to Post*”): filtramos los eventos para extraer solo los que el chute finalice entre los 3 palos de la portería.
- Colecciones: de estas transformaciones surgen nuevas colecciones en MongoDB.
- **Chutes**: 425 documentos.
 - **Goles**: 62 documentos.
 - **Chutes 3 palos**: 160 documentos.
- Reconocimiento de patrones: con todas las transformaciones realizadas y las colecciones pertinentes, realizamos el proceso analítico que se muestra en el punto anterior.
- PowerBi: realizada la analítica y la extracción de patrones se definen unas métricas que van a ser graficadas en un dashboard, del cual se va a extraer un informe final del proyecto.
- **Merge Matches & Goles**: en la definición de métricas nos surgió una nueva necesidad; necesitábamos conocer los equipos rivales y en qué campo se jugaba el partido para obtener mayor información y poder extraer algún patrón más. Para ello, con el Json de Matches, creamos una nueva colección en MongoDB, la transformamos en Databricks a Dataframe, y por último hacemos un merge de la información del dataframe de matches con el de goles para poder obtener en una misma vista los goles, rivales y campos.

5.5 Output del proyecto

En un principio vamos a considerar de valor aquellas jugadas que hayan finalizado en gol, analizando los patrones y características de estas. De esta manera, el output inicial va a contener las siguientes métricas:

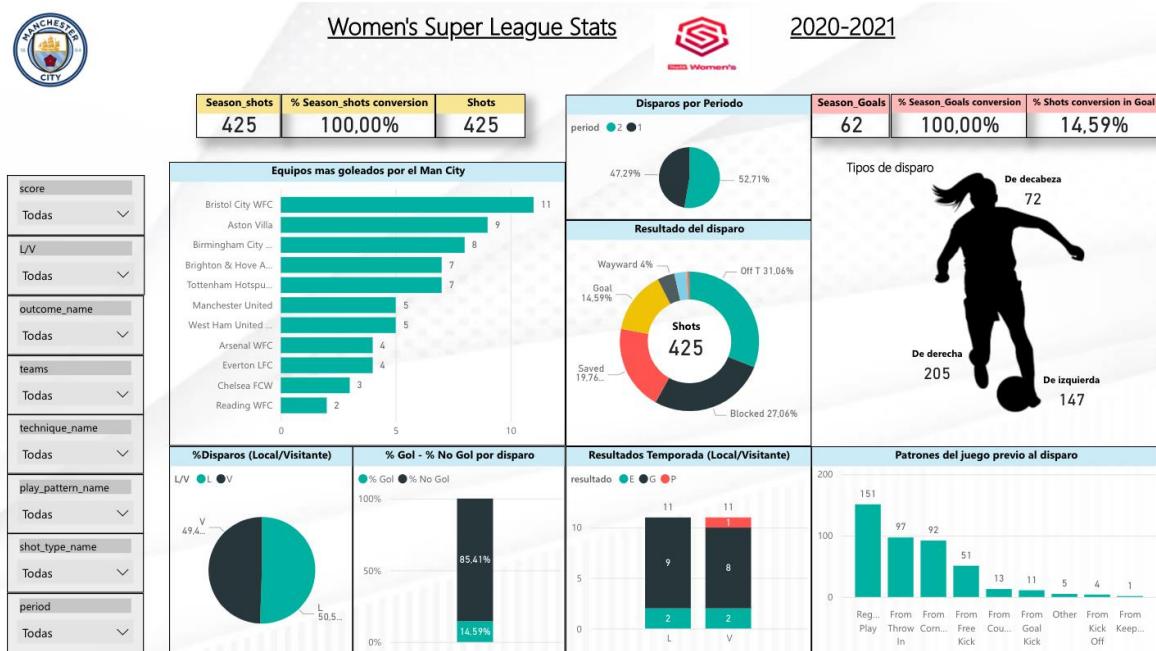
- Métrica 1: Número total de chutes y Tasa de conversión de estos a gol.
- Métrica 2: Número total de goles en toda la temporada.
- Métrica 3: Número total de goles por períodos (primera o segunda parte).
- Métrica 4: Mapa del campo con los chutes y goles.
- Métrica 5: Número total de chutes que se sitúan entre los 3 palos.
- Métrica 6: Número de chutes y goles por jugadora.
- Métrica 7: Partes del cuerpo que intervienen en los goles.
- Métrica 8: Patrones de juego previos al gol.
- Métrica 9: Tipo de finalización tras un chute.
- Métrica 10: Mapa de calor de chutes y goles.
- Métrica 11: Tasa de goles en casa vs goles en campo ajeno.
- Métrica 12: Goles marcados contra cada equipo rival.
- Métrica 13: Distribución de goles partidos.
- Métrica 14: Probabilidad de gol por jugada.
- Métrica 15: Promedio de xG por jugadora y posición.

Para una óptima visualización, dada la gran cantidad de métricas, decidimos realizar 2 páginas:

→ **Página 1 (Shots Overview)**: estadísticas generales del Manchester City durante la temporada 2020-2021. El [dashboard](#) se ha realizado con el dataset de chutes y se ha optado por una forma dinámica, por lo que se pueden ir aplicando los distintos filtros que aparecen a la izquierda, para visualizar con mayor detalle aquellos datos sobre los que nos interese profundizar.

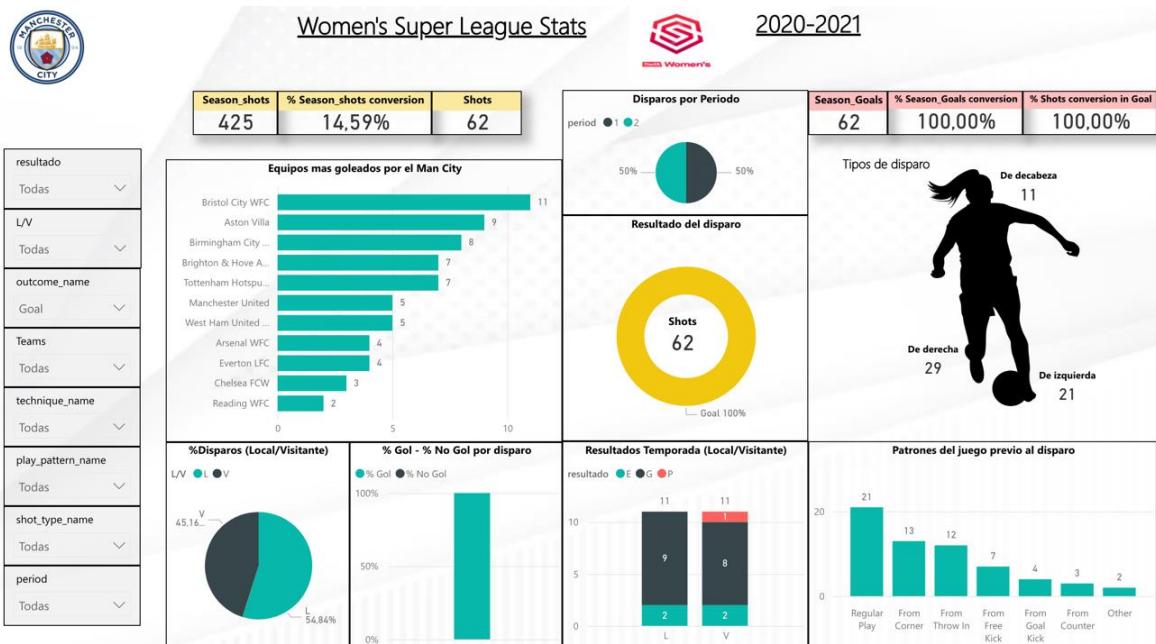
- Score: Empate / Ganado / Perdido
- L/V: Local / Visitante
- *Outcome_name*: Blocked / Goal / Off T / Post / Saved / saved Off Target / Saved To Post / Wayward
- *Teams*: equipos rivales
- *Technique name*: Backheel / Half Volley / Lob / Normal / Overhead Kick / Volley
- *Play_pattern*: From Corner / Counter / Free Kick / Goal Kick / Keeper / Kick Off / Throe In / Other / Regular play
- *Shot_type*: Free Kick / Open Play / Penalty

- Period: 1 o 2



Fuente: Dashboard de elaboración propia en Power BI (Estadísticas generales MC)

Por ejemplo, si filtramos por “**outcome_name**”, que se trata del resultado final del chute, podemos seleccionar aquellos que fueron goles y visualizar todas las estadísticas relacionadas a este filtro.

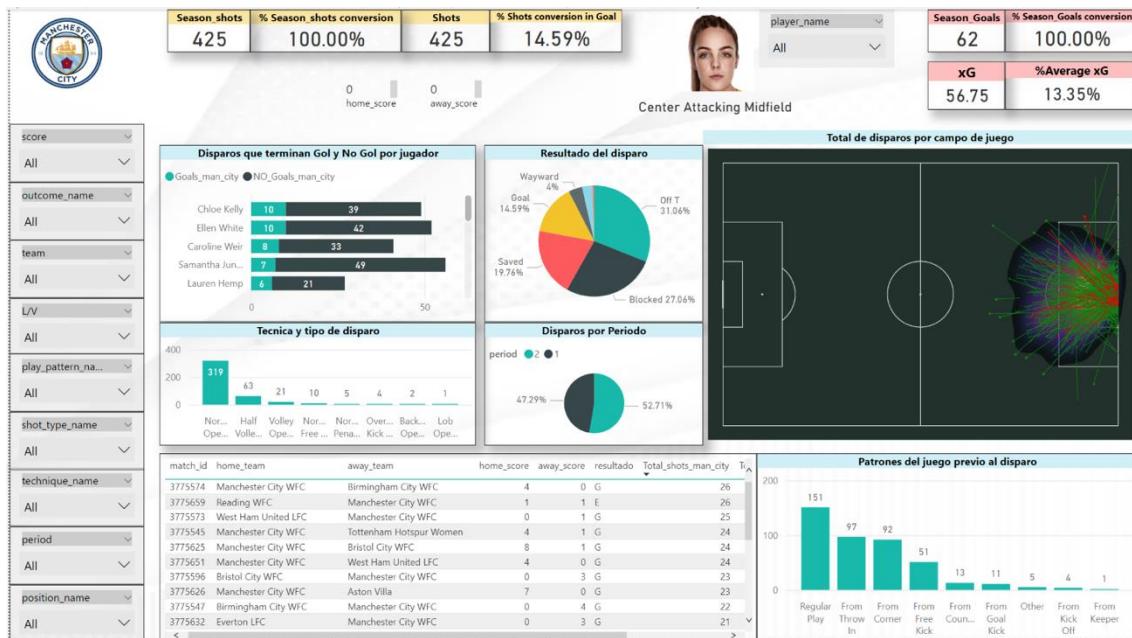


Fuente: Dashboard de elaboración propia en Power BI (Estadísticas de goles MC)

En esta visualización existen métricas extraídas mediante fórmulas y que les corresponde una breve explicación:

- **% Season_shots conversion:** corresponde a la tasa de chutes respecto al total. Es decir, si seleccionamos un filtro de outcome_name (Goal), como en la imagen anterior, este índice nos va a indicar que % del total de los 425, son goles. O poniendo otro ejemplo, si seleccionamos el filtro de technique_name (Half Volley), el índice nos indicará que % del total de los 425 chutes, fueron mediante la técnica de bolea media.
- **% Season_Goals conversion:** corresponde a la tasa de goles respecto al total. Es decir, si seleccionamos un filtro de technique_name (Half Volley), este índice va a indicar que % de los goles se han marcado mediante esta técnica.
- **% Shots conversión in Goal:** corresponde a la tasa de goles respecto al total. Es decir, si volvemos a seleccionar el filtro de technique_name (Half Volley), este índice va a indicar que % de los chutes mediante esa técnica han finalizado en gol.

→ **Página 2:** visualización dinámica con distintos filtros, con la posibilidad de filtrar por partido y jugadora, lo que nos permite visualizar de manera individual la performance de la jugadora durante un partido en particular.



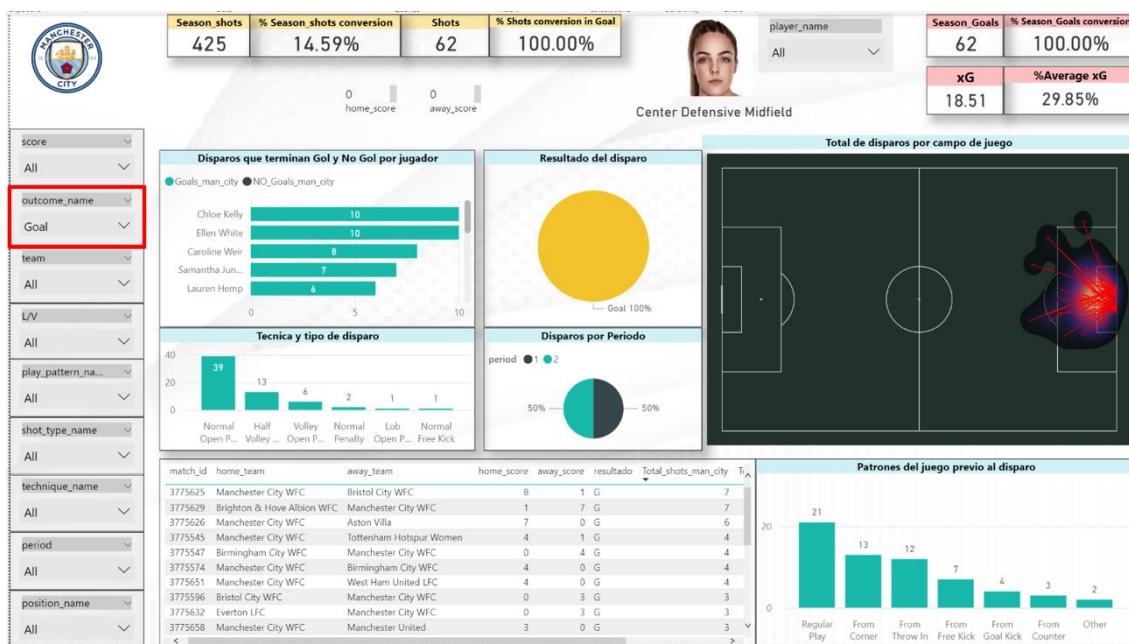
Fuente: Dashboard de elaboración propia en Power BI (Estadísticas generales MC)

Es importante destacar de esta visualización, que por defecto aparece la imagen de una jugadora y una posición aleatoria en el caso de que no se seleccione ningún filtro. De manera que, en términos generales la imagen y la posición no tienen importancia, hasta que no se empiece con la aplicación de filtros.

En esta segunda visualización existen otras métricas extraídas mediante fórmulas y que les corresponde una breve explicación:

- **xG**: corresponde a la sumatoria de la estadística “xG - expected goals”. Es decir, al seleccionar cualquier filtro, lo que se va a hacer es sumar todos los xG de cada jugada incluida dentro de este filtro. De esta manera, si se obtiene un xG de 18.51 por ejemplo, esto significa que con los filtros establecidos, se esperaban alrededor de 18 goles.
- **% Average xG**: corresponde a la media de la estadística “xG - expected goals”. Es decir, al seleccionar cualquier filtro, lo que se va a visualizar en esta estadística es el promedio del xG de las jugadas que se incluyen en el filtro seleccionado. Recordemos que el xG mide la probabilidad de que un disparo resulte en gol, teniendo en cuenta diferentes variables (distancia a portería, ángulo a portería, parte del cuerpo, tipo de asistencia o acción previa al disparo). De cada disparo se estima una probabilidad de gol entre 0 y 1.

Como ejemplo vamos a filtrar el **outcome_name=“Goal”**. En este caso visualizaremos todos los estadísticos de aquellos chutes que resultaron en gol.

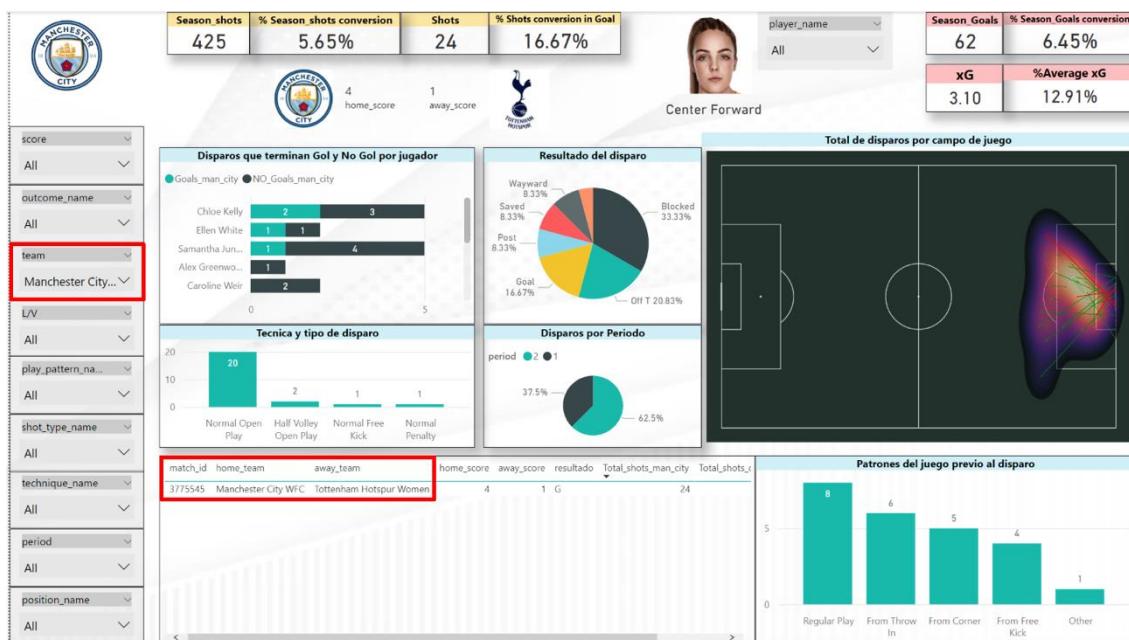


Fuente: Dashboard de elaboración propia en Power BI (Estadísticas por goles MC)

Como se ha mencionado, también podemos filtrar por equipo “**team**” o por jugadora “**player_name**”.

A modo ejemplo, vamos a filtrar en “team” el partido que jugó de local el “**Manchester City**” contra el “**Tottenham Hotspur**”.

Se puede apreciar en el campo de juego que los disparos marcados en color “rojo” representan los disparos que terminaron en “Gol” y los disparos en color “verde” los que “NO terminaron en gol”.

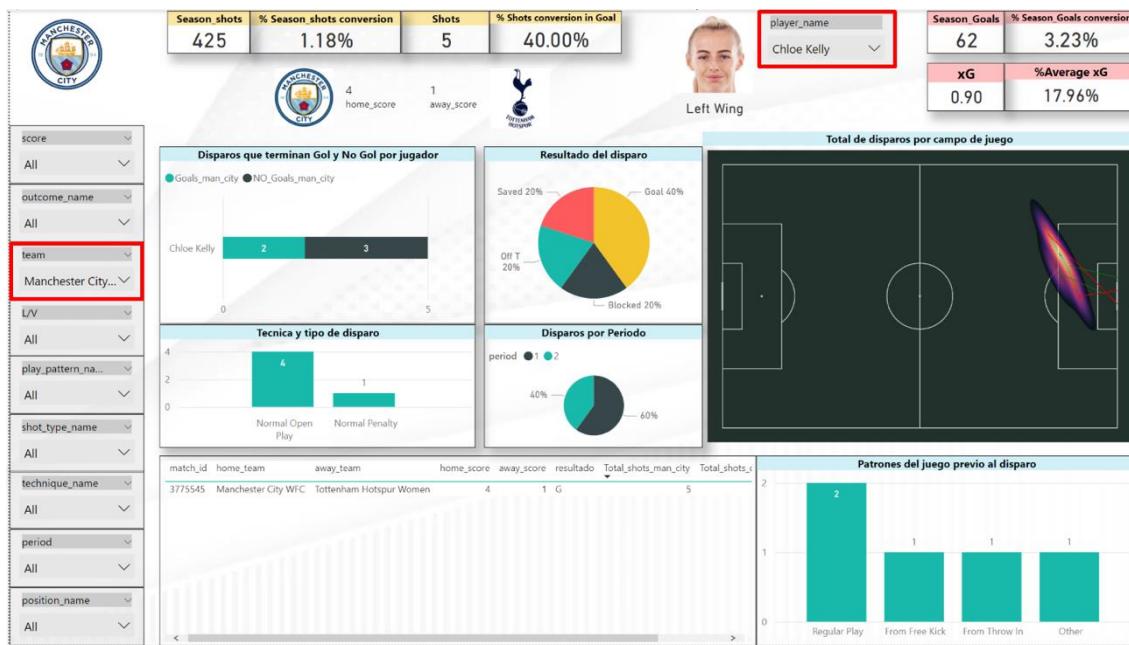


Fuente: Dashboard de elaboración propia en Power BI (Estadísticas MC contra Tottenham)

Para realizar el filtro de un partido concreto disponemos de dos opciones:

- **Tabla:** mediante la tabla que aparece abajo a la izquierda, podemos buscar el partido en cuestión y con un doble click seleccionarlo. Esto generara las estadísticas de ese partido concreto.
- **Team:** en el filtro “team” aparecerán todos los equipo de la liga inglesa en la temporada 2020-2021. El primer equipo corresponde al equipo local, y el que aparece debajo (con sangría) es el equipo visitante. De esta manera, si queremos el el Manchester City sea el equipo local, deberemos seleccionar este y aparecerá un desplegable con todos los otros equipos de la liga. Si por el contra, queremos que el Manchester City sea visitante, deberemos seleccionar en primera instancia el equipo que deseamos como local, y a continuación por defecto asignará al Manchester City como visitante.

Y como último ejemplo, vamos a filtrar en el mismo partido, por una de las jugadoras claves como es “**Chloe Kelly**”. Al seleccionar la jugadora, podemos visualizar la imagen de la jugadora y sus distintas estadísticas con las métricas establecidas.



Fuente: Dashboard de elaboración propia en Power BI (Estadísticas de Chloe Kelly)

En cualquier caso, para más información sobre cada categoría, Statsbomb dispone en [GitHub](#), de un documento PDF con una explicación exhaustiva de cada categoría y sus valores.

Modelado Power BI:

Como parte del modelo de diseño del reporte en Power BI hemos generado distintas tablas con las variables categóricas que aparecen en nuestro **dataset** “**ETL_chutes_manchester_city wfc**”.

Las tablas creadas son las siguientes: “**Team**” - “**Player_name**” - “**body_part_name**”, “**outcome_name**”, “**technique_name**”, “**position_name**”, “**play_pattern_name**”, “**shot_type_name**”, y “**match_id**”.

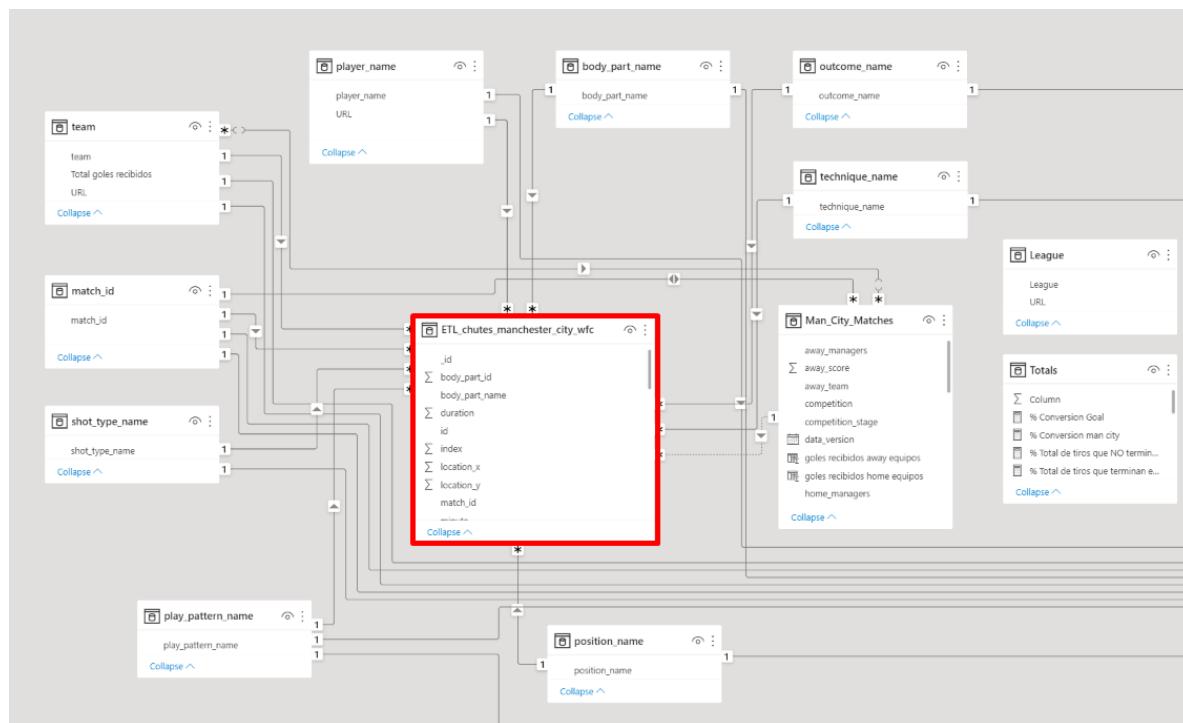
La idea de la generación de estas tablas, es para poder crear relaciones con otros datasets, y al ir aplicando los filtros correspondientes, los mismos trabajen de manera eficiente y coherente. En este caso, el tipo de relación aplicado es de “**uno a muchos** (1:*)”. Esto significa que, en una tabla voy a tener en una fila un valor único y ese valor único se puede relacionar con otro dataset que tiene ese valores múltiples.

Por ejemplo: en mi tabla “**match_id**” tengo un valor único por cada “**match_id**” como puede ser el “**3775545**” al relacionarlo con el dataset “**ETL_chutes_manchester_city wfc**”, puede haber varias filas en la tabla “**ETL_chutes_manchester_city wfc**” que están relacionadas con esa fila específica.

Después, hemos subido a nuestro modelo una tabla con los resultados de los partidos “**Man_city_matches**”, marcando si los mismos fueron de local o visitante, cada partido con su **match_id** en particular para poder crear una relación.

Por otro lado, para obtener el logo de la liga femenina de Inglaterra, hemos subido una tabla con el URL del logo.

Y por último, hemos creado una tabla de “**totales**” donde al ir aplicando distintas fórmulas, creamos medidas como el “total de disparos”, “total de disparos convertidos en gol como NO gol”, y las conversiones de los disparos en porcentaje % en relación al “**shot_outcome**”, además de la sumatoria del gol esperado (xG) y su promedio.



Fuente: Elaboración propia en Power BI (Modelado de datos)

6. Casos de uso

Tras el EDA y el reconocimiento de patrones, una vez ya hemos extraído e interpretado todas las estadísticas e ítems mostrados anteriormente, consideramos oportuno exemplificar con diversos casos de uso sobre el dashboard elaborado, para poder ver su funcionamiento y utilidad.

6.1 Jugadora goleadora

En este caso nos encontramos que el Manchester City tiene dos jugadoras empatadas a goles, que son Chloe Kelly y Ellen White. No nos sorprende que ambas son delanteras.

Chloe Kelly

En el filtro de “*player name*” seleccionamos la jugadora deseada, que en este caso es Chloe Kelly. Esta jugadora disputó 21 partidos de los 22 y sólo en 8 de ellos fue capaz de marcar un gol, con un máximo de 2 goles en un mismo partido; aún así, vamos a analizar en profundidad los datos obtenidos.

Vemos en los insights superiores que ha realizado 49 disparos a puerta, de los cuales 10 han resultado gol, con una tasa de conversión del 20,41%. Además, como podemos visualizar en el ítem “**% Season_Goals conversion**”, sabemos que sus goles representan un 16,13% del total del Manchester City. También cabe mencionar que el **xG** es de 8,62 lo que significa que se esperaban una media de 8-9 goles, y esta ha sido superada.

Un punto a destacar es que el Manchester City únicamente le marcó 3 goles a su rival directo y campeón de la liga (Chelsea), y de esos, 2 fueron de Chloe Kelly, por lo que consideramos que esta jugadora es una pieza clave en el equipo.

En cuanto a la **técnica de disparo**, un dato importante es que de los 6 disparos de volea alta, 4 resultan en gol con una muy buena tasa de conversión; el 40% de los goles de Chloe Kelly son de volea alta.

Otro aspecto interesante es que el 67,35% de los disparos se sitúan **entre los 3 palos** y solo el 32,65% se va fuera, aunque tampoco muy alejados, por lo que esta jugadora es efectiva en los chutes.

Haciendo referencia al campo de juego, cuando el partido se disputa en local, acumula un total de 33 goles, lo que representa el 67% del total de chutes y con una tasa de conversión que aumenta hasta el 21% anotando 7 goles de los 10 totales.

El **mapa de calor** de chutes y goles de esta jugadora nos muestra que la tendencia es a realizar chutes desde la frontal del área mayoritariamente, y también se deduce que es más efectiva por el lado izquierdo.

Ellen White

En el filtro de “*player name*” seleccionamos la jugadora deseada, que en este caso es Ellen White. Esta jugadora disputó 22 partidos de los 22 y sólo en 9 de ellos fue capaz de marcar un gol, con un máximo de 2 goles en un mismo partido; estadísticas muy similares a la jugadora anterior.

En los insights superiores, se puede ver que ha realizado 52 disparos a puerta, de los cuales 10 han resultado gol, con una tasa de conversión del 19,23%. Además, como podemos visualizar en el ítem “% *Season_Goals conversión*”, sabemos que sus goles representan un 16,13% del total del Manchester City. También cabe mencionar que el **xG** es de 7,79 lo que significa que se esperaban una media de 7-8 goles, y esta ha sido superada.

En cuanto a la **técnica de disparo**, no se destaca ninguna de ellas; la mayoría fue normal, sin categoría específica.

Otro aspecto interesante es que el 73,03% de los disparos se sitúan **entre los 3 palos** y solo el 26,92% se va fuera, por lo que esta jugadora es efectiva en los chutes, aún más que la anterior.

El **mapa de calor** de chutes y goles de esta jugadora nos muestra una clara tendencia de chutes en la frontal del área pequeña y en una posición bastante centrada.

6.2 Jugadora chutadora

El Manchester City no tiene una jugadora que destaque demasiado en cuanto a chutes, aunque es Samantha June Mewis la jugadora que más chutes acumuló en la temporada 2020-2021. La posición de mediocentro ofensiva, va acorde el mapa de calor y el cúmulo de chutes, predominantemente desde fuera del área que ejecutó esta jugadora.

Esta jugadora finalizó 56 **chutes**, representando un 13,18% del total de los chutes del Manchester City. De estos, solo 7 se convirtieron en **gol**, con una tasa de conversión del 12,5%, y con un **xG** muy ajustado.

Cabe destacar que a pesar de tener unas estadísticas similares a las jugadoras antes mencionadas, Sam Mewis disputó 17 partidos con un total de 1147 minutos acumulados, bastantes menos que Chloe y Ellen, y aún así consiguió estadísticas nada despreciables.

6.3 Partido perdido

El Manchester City solo perdió un partido en la temporada 2020-2021 y fue contra el Chelsea, equipo campeón de la liga inglesa femenina en dicha temporada.

El partido se perdió en el estadio del Chelsea, por tanto, jugando como visitantes. Únicamente fueron capaces de realizar 9 chutes, teniendo en cuenta que la media de chutes por partido se sitúa en torno a los 20. Además, solo se consiguió anotar un gol y este fue de penalti, aumentando la tasa de conversión de chutes a gol hasta el 11,11%, aunque bastante ajustada.

Aspecto a destacar es que el Chelsea realizó 18 chutes (el doble), de los cuales 3 resultaron en gol, con una tasa de conversión del 16,7%, y esto nos indica que fueron bastante superiores.

En el mapa de calor podemos ver que los **tiros** son **poco precisos**, solo el 33,3% se sitúa entre los 3 palos, teniendo en cuenta que en esta estadística se incluye el penalti.

La gran mayoría de empates y el partido perdido se sitúan entre los 7 primeros partidos de liga, lo que puede sugerirnos que la pretemporada no fue demasiado buena.

6.4 Rival directo

El gran rival del Manchester City en esta liga, fue el Chelsea, el único equipo con el que perdió y al que no fue capaz de ganar en ninguno de los encuentros.

En comparación con el caso anterior, vemos que las estadísticas mejoran cuando también contemplamos el partido de vuelta jugado en casa. Aún así, el total de chutes del Manchester City en los dos partidos contra el Chelsea, se parece a la media de un solo partido, lo que indica que generaron pocas ocasiones.

Un aspecto a destacar es que el **patrón de juego** previo al disparo más prevalente es mediante un **saque de banda**, por lo que podemos considerar que los saques de banda de $\frac{3}{4}$ de campo en adelante generan ocasiones de gol.

La jugadora que más valor aporta en cuanto a chutes fue Chloe Kelly, la cual ya hemos mencionado que es una pieza clave para el equipo.

6.5 Técnica de disparo

La técnica de disparo más frecuente con un 75,06% del total es la normal de juego abierto, es decir, un chute que se produce en juego abierto y sin ninguna característica especial.

En cuanto al patrón de juego previo al disparo, mencionar los más usuales más allá del juego abierto que fueron 97 disparos provenientes de saque de banda de los cuales 12 fueron gol, 92 provenientes de córner de los cuales 12 se materializaron en gol y 51 de tiro libre de los cuales 7 resultaron en gol.

Cabe destacar que **no existe una chutadora de tiros libres**, sino que varias de ellas fueron las pateadoras, siendo Samantha June Mewis la que más tiros libres disputó; aún así consideramos que deberíamos de darle más oportunidades a jugadoras como Keira Walsh o Laura Coombs, dado que de 2 tiros libres disputados cada una, ambas consiguieron 1 gol, por lo que deberíamos de poner a prueba el posible potencial de chute de estas jugadoras.

El Manchester City disfrutó de **5 penaltis a favor**, de los cuales **solo 2 terminaron en gol**, con una tasa de conversión del 40%, aspecto mejorable, dado que un penalti es una jugada de alta probabilidad de gol que hay que aprovechar.

6.6 *Posiciones relevantes en el chute*

No es sorprendente que los chutes se sitúan de $\frac{3}{4}$ del campo en adelante. La gran mayoría se realizan desde una posición bastante central y muchos de ellos cerca de la frontal del área.

Aún así, visualizamos claramente en el mapa de calor de los goles, que la mayoría son en el espacio **entre el punto de penalti y la frontal del área pequeña**.

Este echo nos indica dos cosas principales:

- El Manchester City es mucho más efectivo en chutes próximos a portería.
- Se deberían de entrenar en mayor medida los chutes desde la frontal y fuera del área para ser más efectivas.

6.7 *Local vs Visitante*

En cuanto a disparos no existe una gran diferencia entre partidos jugados en local o visitante, es decir, no parece tener especial relevancia el campo de juego en términos generales. En cambio, si se trata de los goles, podemos ver que sí que existe una ligera diferencia, teniendo una **mayor tasa de conversión de gol** cuando se juega en **campo propio**, obteniendo en casa un 15,81% y en siendo visitantes un 13,33%.

Un dato curioso en cuanto a los goles, es que el xG es mayor en visitante que en local, por lo que se intuye que a priori se deberían de anotar más goles en campos ajenos que en el propio campo del Manchester City.

Se anotaron 34 goles a favor en el Academy Stadium superando de casi 10 goles la estimación de entorno a los 26-27 goles esperados. Por el contrario, se anotaron 28 goles en campos ajenos, lo cual se acerca más a la cifra de 29-30 goles esperados. Deducimos pues, que **se esperaban menos goles de los realmente anotados**; fueron más efectivas de lo que se esperaba, sobre todo en los partidos locales.

En referencia a las jugadoras, la jugadora **Chloe Kelly**, en los partidos disputados en campo propio tiene las mejores estadísticas en cuanto a chutes y goles, sin embargo, en los partidos en campo ajeno cae hasta el 4º puesto reduciendo a la mitad tanto los goles como los disparos. Tendríamos que indagar qué le sucede a nuestra mejor jugadora en los partidos en los campos rivales, para poder incidir en ello y sacarle el máximo partido a nuestra jugadora estrella.

Acerca de la técnica y tipo de disparo, así como también el patrón de juego previo al disparo, **no existen diferencias significativas** entre los partidos jugados en local o visitante.

Brighton & Hove Albion WFC, sorprende que en local se quedara 0-0 y jugando como visitantes ganáramos 1-7. En nuestro campo se realizaron 15 chutes y en el suyo 19, por lo que podemos decir que jugando como locales fuimos bastante imprecisas. Además, sorprende ya que fue el **único partido que se empató en casa**, a parte de contra el líder de liga. Esto lo consideramos realmente importante dado que el resultado final en cuanto a puntos de esta temporada fue de 57 puntos para el Chelsea y 55 para el Manchester City, lo que significa que este partido empatado en casa contra el Brighton, perdiendo esos dos puntos, y teniendo en cuenta la sobrada victoria de la vuelta contra este equipo, pudo ser decisivo para no conseguir ganar la liga.

6.8 *Penaltis*

Tal y como se ha comentado anteriormente, el Manchester City tuvo la oportunidad de chutar **5 penaltis a favor**, con una conversión de gol del 40%, convirtiendo el gol solo 2 de ellos.

En el campo de juego, podemos observar una clara tendencia de lanzar los penaltis hacia el **lado derecho de la portería**, y este dato seguramente también esté en manos de los porteros rivales, siendo más fácil parar el penal. Es por ello, que se debe ir alternando el lado.

Queda claro que la jugadora que debería tirar los penales es Chloe Kelly por su efectividad, ya que ninguna otra jugadora ha sido capaz de materializar su oportunidad.

Los penaltis no marcados no han repercutido negativamente en el resultado global de la temporada, es decir, ninguno de ellos ha sido decisivo. Aún así, se debería entrenar ese chute a puerta desde los 11 metros para convertir en gol la gran mayoría de penaltis, ya que son una gran oportunidad de anotar a favor en el marcador.

7. Ampliación análisis (comparativa)

Para extraer valor de los datos tenemos que saber que estadísticas tienen los demás equipos, para realmente determinar si el Manchester City hizo una buena temporada, y cuáles aspectos necesitan mejorarse y entrenarse.

Los datos extraídos por si solos nos aportan información limitada, debemos de **contextualizarlos**, y para ello vamos a realizar el análisis del equipo que se posicionó primero (Chelsea femenino) y el que quedó último (Bristol City femenino). Para realizar tal análisis, debemos llevar a cabo el mismo proceso que se ha definido para el análisis de los datos del Manchester City femenino.

Obviamos el proceso, aunque iremos mencionando aquellas variaciones específicas que corresponden a las estadísticas de dichos equipos, aunque el proceso de preparación de datos y analítica va a ser el detallado en apartados anteriores. Así pues, procederemos directamente a la analítica de los datos.

El ID de la competiciones y de la temporada, obviamente van a ser los mismos (competition_id= 30 y season_id= 90). Una vez estamos dentro del dataset, seleccionamos los partidos creando dos datasets distintos, uno solo con los partidos del Chelsea y otro con los del Bristol City, tanto local como visitante. De ahí extraemos los eventos de todos los partidos, y posteriormente seleccionamos sólo aquellos eventos que sean chutes y goles, creando así las 3 principales colecciones en Mongo DB de Eventos, Chutes y Goles.

Una vez tenemos la base de datos montada con todas las colecciones necesarias, pasamos al proceso de Data Quality mencionado en el punto 5.3.2 utilizando la herramienta TOSDQ.

En este punto ya tenemos todo lo necesario para pasar al modelado en Power BI para obtener el mismo dashboard que confeccionamos con el Manchester City, y poder hacer una comparación de datos entre este y el primer y último equipo de liga.

7.1 Chelsea

En cuanto a las estadísticas generales vemos que el Chelsea ([dashboard](#)) realizó 430 chutes a portería en la temporada 2020-2021, 5 más que el Manchester City, diferencia poco relevante. Acerca de los goles, fue capaz de meter 67 goles, con una tasa de conversión del 15,58%, es decir, un 1% más efectivas que el Manchester City.

El Chelsea ejecutó 227 chutes con la pierna derecha, 114 con la izquierda y 86 remates de cabeza. Se diferencia del Manchester City dado que, el Chelsea tiene **menor cúmulo de chutes con la pierna izquierda y mayor cúmulo de remates de cabeza y chutes con la pierna derecha**. Este punto, es meramente informativo, no creemos que sea determinante en la diferencia de goles, aunque si que vemos que el Chelsea es un 6% más efectivo con la pierna izquierda que con la derecha y un 8% más de cabeza, a diferencia del Manchester City que no existen diferencias significativas en cuanto a la parte del cuerpo con la que se dispara.

Otro aspecto destacable es que el Chelsea perdió 1 partido, empató 3 y ganó 18, la diferencia es que el Manchester empató 4 y ganó 17, y esto constituyó la diferencia final de 2 puntos, que hizo ganar la liga al Chelsea.

Las chicas del Chelsea acumulan **muchos más chutes de juego regular abierto** que de todos los otros patrones de juego, en el caso del Manchester City se reparte de manera más homogénea.

Y por último, queda comentar la diferencia en cuanto a equipos goleados. Encontramos 2 grandes diferencias que nos llaman la atención:

- El Chelsea **goleó con 10 goles al Reading WFC** siendo el segundo equipo más goleado, a diferencia del Manchester City que solo le metió 2, siendo el equipo que menos goles recibió de parte de este.
- El Manchester City **goleó con 7 al Brighton & Hove Albion** siendo el cuarto equipo más goleado, a diferencia del Chelsea que solo le metió 2, siendo el equipo que menos goles recibió por parte de este.

7.2 Bristol City

En cuanto a las estadísticas generales vemos que el Bristol City ([dashboard](#)) realizó 188 chutes a portería en la temporada 2020-2021, 237 menos que el Manchester City, lo que supone menos de la mitad, siendo una diferencia muy importante. Acerca de los goles, fue capaz de meter 17 goles, con una tasa de conversión del 9,04%, es decir, casi un 6% menos efectivas que el Manchester City. No sorprende tanto la efectividad, sino la **poca cantidad de chutes realizados**.

El Bristol ejecutó 138 chutes con la pierna derecha, 35 con la izquierda y 14 remates de cabeza. No existen diferencias significativas en cuanto a la parte del cuerpo con la que se dispara, todas tienen una tasa de conversión en torno al 9%.

Este equipo perdió 14 partidos, empató 6 y ganó 2, **perdió más de la mitad de partidos** y solo fue capaz de ganar un par de partidos.

Un aspecto a destacar es que el **25% de los chutes fueron de falta**, es decir, a juego parado, y solo alrededor del 27% fueron de juego abierto. Esto significa, que son poco efectivas en la finalización, no consiguen realizar muchos disparos provenientes de juego abierto. Muestran una dificultad de llegar al área rival, por lo que deberían de trabajar la finalización desde fuera del área para aumentar las ocasiones de gol.

Y por último, queda comentar la diferencia en cuanto a equipos goleados. Sorprenden dos datos concretos:

- El Bristol **anotó 4 goles a favor contra el Reading WFC**, a diferencia del Manchester City que solo le metió 2. Esto nos da información de valor, dado que vemos que el Reading WFC encajó 4 goles contra el último de la liga, lo que significa que el Manchester City pinchó en los partidos contra este equipo, lo que influyó negativamente en su clasificación.
- El Bristol **anotó 4 goles a favor contra el Brighton & Hove Albion**, a diferencia del Chelsea que solo le metió 2, otro aspecto diferenciador que el Chelsea debería de tener en cuenta para su performance.

7.3 Casos de uso

En referencia a los casos de uso, vamos a plantear 2 casos de uso que creemos que pueden ser de utilidad para la performance del Manchester City.

7.3.1 Jugadoras goleadoras y chuteadoras

El Chelsea sí que dispone de una jugadora que podríamos llamar “killer” ya que destaca de manera importante frente al resto. **Samantha May Kerr** chutó 86 veces y anotó 21 goles, casi el doble que Chloe Kelly (la mejor jugadora del Manchester City). Sorprende que esta jugadora es medio-campo defensiva, aunque tuvo un gran número de ocasiones dentro del área rival y el xG mayor de todo el equipo con diferencia.

El Bristol no tiene una jugadora que destaque en cuanto a gol, de echo **Ebony Salmon** es la jugadora que más goles marcó con un recuento de solo 6 goles en toda la temporada, aún así, si que generó el casi el 32% de los chutes del Bristol con un total

de 60 chutes, aunque con una tasa de conversión del 10%. De echo, esta jugadora fue la que fue capaz de marcar el único gol al Manchester City. En el mapa de calor, podemos ver un aspecto comentado anteriormente, hay una gran cantidad de disparos realizados desde fuera del área, en este caso, con mayor efectividad desde el lado derecho del campo.

De esta manera, el Manchester City ya conoce cuales son las jugadoras claves de ambos equipos que más peligro generan, y que por tanto, más cubiertas y anuladas deberán de estar.

7.3.2 *Partidos contra el Manchester City*

El **Chelsea** ganó 1 partido y empató otro contra el Manchester City con una sumatoria de goles a favor de 5 en total. Se realizaron 34 chutes con una tasa de conversión de 14,71%, un tanto más baja que la general. En el partido como visitantes alcanzaron una tasa de conversión de solo el 12,50%, a diferencia del partido local donde subieron la tasa hasta el 16,67%, siendo mucho más decisivas y efectivas como locales. La gran mayoría de chutes provienen de juego regular, aunque hay un alto porcentaje (casi un 30%) que provienen de córner.

Otro aspecto destacable es que en la **segunda parte**, las chicas del Chelsea **mejoran sus estadísticas**, tanto en los chutes como en los goles, pudiendo ser gracias a los cambios o a la fatiga del equipo rival.

El 40% de los goles provienen de penalti. En cada partido les pitaron un penalti a favor que supieron aprovechar convirtiendo ambos en gol.

Todo esto nos indica que debemos tener cuidado con las faltas que realizamos dentro del área, además de intentar aumentar el número de chutes contra el Chelsea para crear más oportunidades de gol, sobretodo en jugadas abiertas en las que acumulamos muy pocos disparos. Realmente, la diferencia no es exagerada y podríamos haber empatado o incluso ganado ambos partidos si hubiéramos aprovechado mejor las ocasiones.

El **Bristol** perdió ambos partidos contra el Manchester City, acumulando 1 único gol en el partido disputado como visitante. Solo fueron capaces de realizar 9 chutes, y la mayoría de ellos provenientes de una jugada a balón parado.

En el partido como locales, a pesar de no marcar ningún gol, generaron más ocasiones; 7 de los 9 chutes ejecutados se hicieron en este partido y encajando únicamente 3 goles. Aún así, en el partido como visitantes solo ejecutaron 2 disparos, de los cuales 1 fue gol, por lo que podríamos decir que aprovecharon bien las oportunidades que tuvieron, aunque por el contrario, encajaron 8 goles, siendo menos efectivas en defensa.

En realidad este equipo **no genera ningún peligro**, aunque deberíamos analizar con más profundidad las causas de el gol encajado, dado que también nos interesa dejar la portería a 0 y encajar los menos goles posibles. Como ya habíamos mencionado con anterioridad, destacamos a la jugadora Ebony Salmon, a la cuál si hubiéramos puesto atención especial, seguramente podríamos haber evitado el gol.

En resumen, **realmente las estadísticas del Manchester City en esta temporada son bastante buenas** dado que realizando una pequeña investigación del resto de los equipos de esta misma liga, los datos son bastante positivos.

Es verdad que a simple vista, y acostumbrados al fútbol masculino, estas estadísticas pueden resultar mediocres, de ahí la importancia de realizar un trabajo de comparación para contextualizar los datos y poder extraer la información pertinente que nos aporte valor.

8. Machine Learning

Una vez ya hemos superado la analítica del dato, donde era indispensable presentar el dashboard, pasaremos a iterar sobre la etapa de modelado para la selección del modelo que más se adapta a los datos, para determinar qué jugadas van a ser de valor para el equipo con todas las variables analizadas anteriormente.

Para el desarrollo del modelo optamos por **Python** y las principales librerías utilizadas fueron **pandas, numpy, sklearn y matplotlib** con sus respectivas funciones de ciencia de datos y visualización.

Dadas las características de los datos disponibles, hemos optado por confeccionar un algoritmo de ML de clasificación que nos permita predecir qué resultado se va a obtener del chute en cuestión, siendo la columna '**shot_outcome**' la variable objetivo y el resto las predictoras. De manera que, el **objetivo del algoritmo** va a ser que mediante una entrada de datos nos prediga qué tipo de finalización va a resultar del chute introducido.

Lo primero que hicimos, fue seleccionar aquellas columnas que creemos irrelevantes para el modelo, eliminando las siguientes: '*duration*', '*id*', '*index*', '*location*', '*match_id*', '*possession_team_id*', '*related_events*', '*shot_end_location*', '*team*', '*timestamp*', '*type*', '*shot_end_location_z*', '*related_events1*', '*related_events2*', '*related_events3*', '*possession_team*'. El motivo por el que eliminamos estas variables fue básicamente porque o bien son números identificadores sin ningún valor, o bien son columnas que tienen muchos valores en blanco que no nos aportan demasiada información.

El siguiente paso, fue la transformación con el sistema de codificación **one hot encoding** de las siguientes columnas: 'play_pattern', 'player', 'position', 'shot_body_part', 'shot_technique', 'shot_type'.

Sobre la variable 'shot_outcome' también realizamos una codificación, pero esta vez con LabelEncoder clasificando de manera arbitraria del 0 al 7 los valores de la variable quedando los [datos](#) de la siguiente manera:

- 0: Blocked
- 1: Goal
- 2: Off T
- 3: Post
- 4: Saved
- 5: Saved Off Target
- 6: Saved to Post
- 7: Wayward

Será importante tener en cuenta esta clasificación ya que las predicciones se harán partiendo de la base de esta codificación. Realmente, el orden de la numeración no es importante, pero algunos algoritmos requieren que la variable sea numérica, por lo que se optó por hacer este tipo de codificación.

En este punto ya procedemos a la modelización como tal, en la que los modelos que vamos a comparar son:

- Random Forest
- Naive Bayes
- SVM
- XGBoost
- Gradient Boosting

Para proceder a entrenar los algoritmos, en primer lugar tenemos que **dividir** el conjunto de datos en X (todos los datos menos la variable objetivo) e y (variable objetivo). Una vez tenemos esto, volvemos a dividir los datos en entrenamiento y testeo dando como resultado: X_train, X_test, y_train, y_test. Los datos incluidos en "train" contienen el 80% y los contenidos en "test" el 20% restante.

En este punto ya podemos crear los modelos y entrenarlos, y una vez entrenados realizamos la predicción para posteriormente poder analizar las métricas y la validez de los modelos. Además, deberemos de realizar la **optimización** de los modelos, en nuestro caso utilizando la función GridSearchCV() que nos permite obtener, los mejores hiperparámetros para el conjunto de datos y modelo en cuestión.

En la matriz de correlación, **no observamos ninguna correlación que ponga en peligro los datos en cuanto a multicolinealidad**, aún así, pudimos observar a partir de distintas pruebas que si se omite la variable 'location_y' que creímos importante, las métricas del modelo mejoraban. Al usar la función `feature_importance()` que ofrece sklearn, dicha variable tiene un peso moderadamente importante en el modelo, aunque al entrenar todos los modelos sin esta variable todas las métricas aumentan, por lo que decidimos dejarla fuera del modelo. Lo que está claro es que la localización, tanto inicial como final del chute es importante en el resultado de este.

Las métricas por las que optamos para la comparación de los modelos fueron la exactitud (**accuracy**), el **R²**, el **F1-scoreMacro** (media no ponderada de las puntuaciones F1 calculadas por clase) y la **Matriz de confusión**. A partir de la comparación de estas métricas, el modelo de Random Forest fue el que mejores resultados dio con una *Accuracy* de 0.84, un *R²* de 0.23 y un *F1-ScoreMacro* de 0.56.

Además de estas métricas, también se realiza la **validación cruzada** tanto en los conjuntos de entrenamiento como de testeo de todos los modelos. Generalmente, los modelos que mejor exactitud presentan como son el **Random Forest** y el **Gradient Boosting**, son los que menor variación presentan en cuanto a la validación cruzada; contrariamente **SVM** y **Naive Bayes** tienen una mayor accuracy en el conjunto de test, y esto puede ser debido a que hay una menor cantidad de datos.

Y por último, también se ha analizado la Curva ROC – AUC de los algoritmos, y al ser un problema de clasificación multiclase, se ha arrojado una curva por cada clase que se detecta. En el caso del algoritmo que por su rendimiento hemos elegido (Random Forest), podemos ver que este modelo para la predicción de la clase 6 (saved to post), el **AUC** es cercano a 0.5 y cerca de la diagonal aleatoria, por lo que para esta clase no es muy efectivo el modelo. Por el contrario, si nos fijamos en la clase 0 (blocked) y la clase 7 (wayward) tienen un **AUC** muy cercano a 1 lo que indica que el modelo se adapta bastante bien. En general, este modelo se adapta bien a todas las clases menos a la 6 y a la 5. En este caso, debemos tomar esto con precaución ya que existen pocos valores y sesgan los resultados, motivo por el cual tampoco podemos ver una curva como tal.

En el siguiente cuadro resumen se sumarizan todos los algoritmos modelizados, las respectivas métricas y sus correspondientes outputs para poder hacer una comparación de forma directa:

Algoritmo	Métrica	Output
Random Forest	Accuracy	0,835
	R^2	0,232
	F1-ScoreMacro	0,563
	AUC (clase 1)	1
XGBoost	Accuracy	0,788
	R^2	0,196
	F1-ScoreMacro	0,509
	AUC (clase 1)	0,99
Naive Bayes	Accuracy	0,056
	R^2	-2,299
	F1-ScoreMacro	0,038
	AUC (clase 1)	0,41
SVM	Accuracy	0,376
	R^2	-0,0004
	F1-ScoreMacro	0,078
	AUC (clase 1)	-----
Gradient Boosting	Accuracy	0,8
	R^2	0,379
	F1-ScoreMacro	0,493
	AUC (clase 1)	0,98

Una de las métricas en la que nos basamos es en el **Accuracy**, que al final lo que evalúa es el total de predicciones correctas realizadas por el modelo en comparación con el total de predicciones arrojadas. En este caso, podemos observar que en el modelo de Random Forest, de cada 100 predicciones, 84 serán correctas, en comparación con el modelo Naive Bayes que de cada 100 predicciones, solo 6 serán correctas. En realidad, la exactitud del Random Forest no es excelente, pero en este caso, es aceptable. Aún así, tenemos en cuenta el desbalanceo de los datos, motivo por el cuál no consideramos ningún modelo de alta calidad.

Otra métrica que también nos aporta mucha información es el **F1-ScoreMacro** (incluso más que la anterior), la media no ponderada de las puntuaciones F1 calculadas por clase (ya que es un problema de clasificación multiclase), combinando las medidas de precisión y exhaustividad en un solo valor. La elección de esta es porque en comparación con la puntuación micro, la macro funciona mejor en conjuntos de datos desequilibrados como es el que nosotros trabajamos. Además, esta métrica también nos es útil dado que en nuestro problema nos importa de igual forma la precisión y la exhaustividad. Así pues, el valor más alto del F1-ScoreMacro es el del modelo Random Forest, el cual también dispone de la mejor diagonal en la matriz de confusión.

En algunos casos, dada la **baja cantidad de datos**, y por ende el **desbalance** entre clases, hay comportamientos sobretodo de validación cruzada y de AUC que pueden ser extraños, motivo por el cuál no las tenemos demasiado en cuenta. Este aspecto podría mejorarse incluyendo más datos y realizando un balanceo de estos, como por ejemplo ampliar el análisis a más temporadas, tal y como planteamos en la continuación futura del proyecto.

Realmente, considerando que el fútbol es un deporte altamente aleatorio donde en una situación prácticamente idéntica, el resultado puede ser muy distinto, y también teniendo en cuenta que el volumen de datos y las variables no son demasiadas, y además los datos están desbalanceados, podemos adoptar el modelo de **Random Forest** mostrado como válido teniendo en cuenta sus limitaciones, ya que a pesar de que la exactitud es aceptable, el F1-score macro es bajo, y además en el R² vemos que el modelo no explica bien la variabilidad de los datos, y esto es por el carácter aleatorio de este deporte. Y es que es muy difícil predecir con alta precisión y exactitud como va a finalizar una jugada de fútbol, ya que a comparación con otros deportes, el abanico de posibilidades es muy amplio y sujeto al azar.

En cuanto a las variables, para optimizar el modelo pensamos que disponer de variables como la potencia del chute, el ángulo de chute, el portero rival, condiciones climáticas y del campo, entre muchas otras variables, se pueden considerar importantes para determinar el resultado de un chute. En el siguiente enlace (https://github.com/lban99/TFM/blob/7e45ab1c831c7acef6b41150b5b6b63a17c90d7a/ML_CHUTES.ipynb) se puede visualizar el código y las distintas gráficas resultantes del proceso de modelización.

9. Conclusiones

De acuerdo con los objetivos planteados, consideramos que hemos cumplido en gran medida todos los propuestos.

Por un lado, en cuanto a la calidad de los datos utilizados, podemos afirmar que Statsbomb provee datos de calidad ya que estos son consistentes, están completos, no se duplican y son válidos. La precisión no la pudimos comprobar dado que no hemos visualizado todos los partidos por lo que no conocemos la realidad, aunque por las características anteriores, confiamos en que también son precisos. Este análisis de calidad, nos aporta confiabilidad en los datos y nos da seguridad en cuanto a la toma de decisiones basadas en estos.

Acerca del objetivo principal de acercar al Manchester City Women's Football Club a la victoria de la liga 2023-2024, creemos que nuestro proyecto puede contribuir a ello, aunque obviamente son muchos los factores que intervienen en dicho objetivo, por lo que se podría ampliar este proyecto al análisis de pases, asistencias, intervenciones del portero y otro tipo de jugadas que aporten valor al equipo.

Por lo que corresponde a la mejora de estrategia para la creación de jugadas en el campo y de nuevos entrenamientos específicos, hemos obtenido distintos patrones que nos permiten incidir sobre distintos aspectos del chute para aumentar la efectividad y la tasa de conversión a gol, y para ello destacamos los insights más importantes:

- Sólo 38% de los chutes se sitúan entre los 3 palos, por lo que un aspecto a mejorar sería direccionar el chute a portería y evitar que tantos disparos no tengan la posibilidad de finalizar en gol. Una posibilidad sería incluir más ejercicios de disparo en los entrenamientos.
- Alrededor del minuto 40 y del minuto 80 se anotan mayor cantidad de goles, por lo que deberíamos de aprovechar al máximo los últimos 5-10 minutos de ambas partes, y esto se puede lograr si las jugadoras tienen un buen nivel físico ya que esto constituye una ventaja ante el rival.
- Los saques de banda son un punto fuerte del equipo de cara al gol, así que se deberían de aprovechar al máximo dichas jugadas e intentar ser directas y finalizar con un disparo a portería, e incluso forzar saques de banda para generar más ocasiones de gol.
- Se disfrutó de 5 penaltis a favor, de los cuales solo 2 se materializaron en gol, y esto constituye una gran pérdida ya que consideramos que un penalti es una gran oportunidad para el equipo. Analizar las fragilidades de los porteros rivales

y dar consejos a las jugadoras que van a chutar penaltis pueden ayudar a mejorar este punto.

- El City no dispone de una jugadora que sea altamente destacable en cuanto al gol, pero si que Chloe Kelly, Ellen White y Caroline Weir son las jugadoras que sumaron más goles, y esto puede ayudarnos a direccionar el juego para que la finalización provenga de alguna de estas jugadoras.
- Los goles anotados provienen de chutes entre el área grande y el área pequeña, deduciendo dos aspectos: las jugadoras se deben aproximar a dicha localización para aumentar la tasa de goles, y por otro lado, se debe practicar el tiro lejano desde fuera del área para poder aumentar la efectividad de estos.

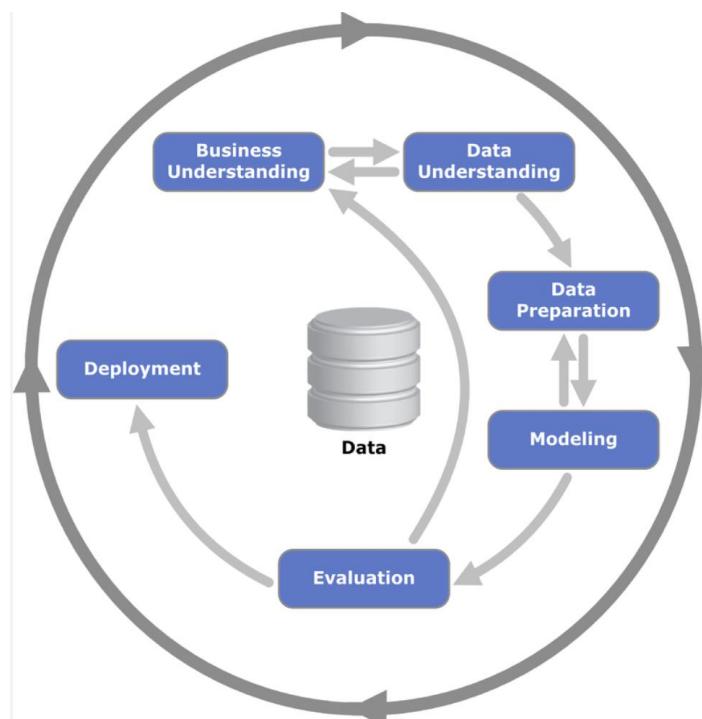
Si bien, las estadísticas del City nos parecían regulares dada la baja cantidad de disparos y goles en comparación con los equipos masculinos, tras la comparación con el Chelsea femenino que fue el líder de aquella temporada, pudimos comprobar que las estadísticas entre ambos equipos eran similares, con una única diferencia de 1 empate, lo que decantó la liga hacia el Chelsea. Esto nos muestra que en la parte alta de la tabla los equipos están bastante igualados y que un partido puede sentenciar la liga, por lo que se debe estar lo máximo precisas en cada partido y no desmerecer a ningún equipo, por muy mala clasificación que tengan.

Todos estos datos pueden ser visualizados en el dashboard desplegado en Power BI, y al ser de carácter dinámico y contener gran variedad de filtros, cabe la posibilidad de ante cualquier caso de uso poder aprovechar dicho cuadro de mandos para conocer los datos deseados.

Y finalmente, con relación al desarrollo de algoritmos de ML, consideramos que ha sido el objetivo que con más esfuerzo y menor calidad hemos alcanzado. El fútbol es un deporte altamente aleatorio, con una gran cantidad de variables influyentes que pueden modificar cualquier jugada, por lo que es realmente complejo predecir la finalización de un chute. El algoritmo que mejores métricas ha obtenido, ha sido el de random forest con una Accuracy de 0.84, un R2 de 0.23 y un F1-ScoreMacro de 0.56. Aún así, tal y como ya hemos mencionado, los datos están desbalanceados entre las clases de los chutes y además disponemos de una baja cantidad de datos, por lo que creemos podríamos optimizar los algoritmos teniendo en cuenta estos aspectos, y una posibilidad sería realizar dicho análisis de más temporadas y de otras competiciones, así como también obtener acceso a los datos de pago que proporciona Statsbomb.

10.**Anexos****10.1 Anexo 1 (Crisp-DM)**

Gráfico de la metodología CRISP-DM



Fuente: https://es.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

10.2 Anexo 2 (plantilla)

PLANTILLA

PORTERO

1		 Karen Bardsley 36 años	26		 Ellie Roebuck 21 años
34		 Karima Benameur 32 años	35		 Khiara Keating 17 años

DEFENSA

37		 Anna Phillips 18 años	20		 Lucy Bronze 29 años
2		 Aoife Mannion 25 años	4		 Gemma Bonner 29 años
6		 Steph Houghton 33 años	13		 Abby Dahlkemper 28 años
14		 Esme Morgan 20 años	3		 Demi Stokes 29 años
5		 Megan Campbell 28 años	27		 Alex Greenwood 27 años

CENTROCAMPISTA

36		 Alicia Window	39		 Millie Davies
38		 Millie Ravening 19 años	22		 Samantha Mewis 28 años
24		 Keira Walsh 24 años	7		 Laura Coombs 30 años
8		 Jill Scott 34 años	12		 Tyler Toland 19 años
19		 Caroline Weir 26 años	21		 Rose Lavelle 26 años

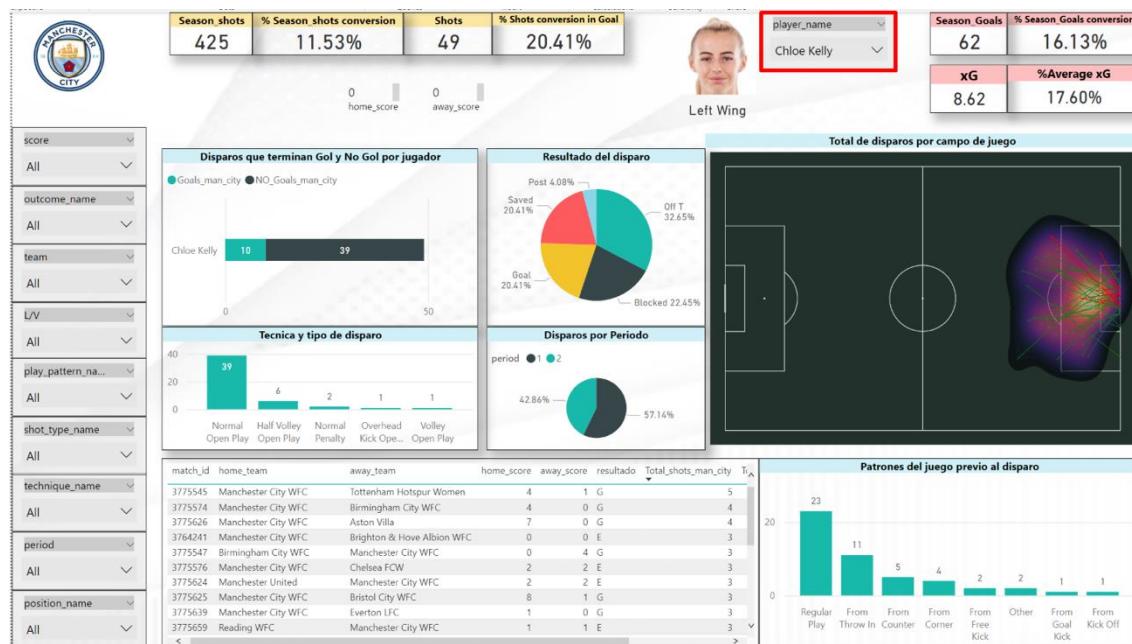
DELANTERO

16		 Jessica Park 19 años	10		 Georgia Stanway 22 años
11		 Janine Beckie 26 años	15		 Lauren Hemp 20 años
9		 Chloe Kelly 23 años	18		 Ellen White 32 años

Fuente: <https://www.livefutbol.com/equipos/manchester-city-wfc-frauen/2021/2/>

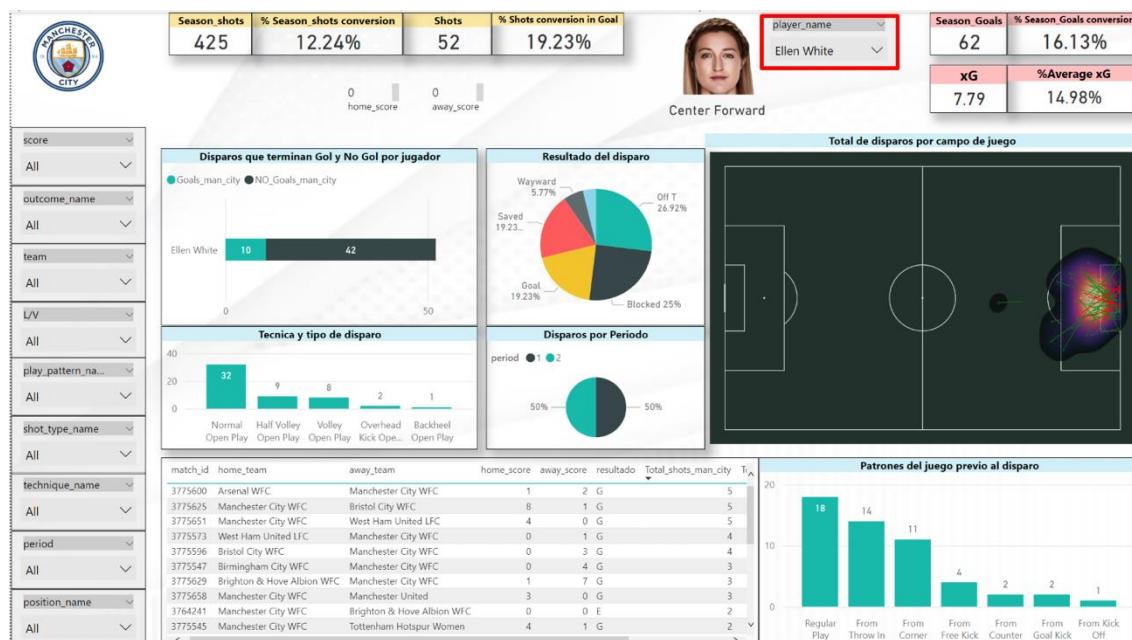
10.3 Anexo 3 (goleadoras)

Chloe Kelly



Fuente: Dashboard de elaboración propia en Power BI

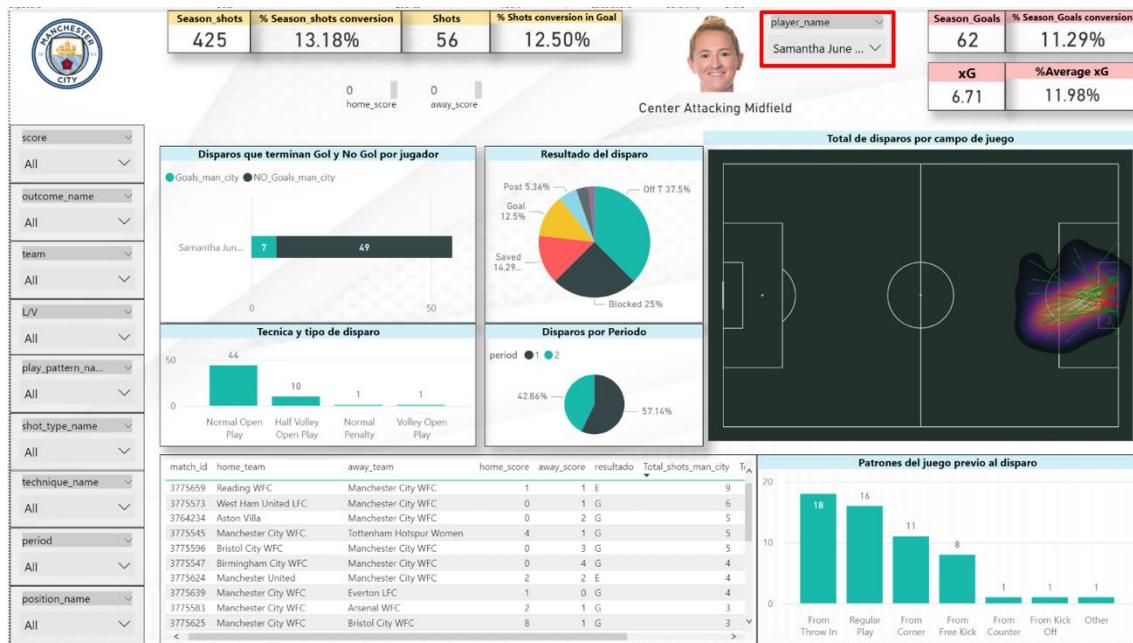
Ellen White



Fuente: Dashboard de elaboración propia en Power BI

10.4 Anexo 4 (chutadora)

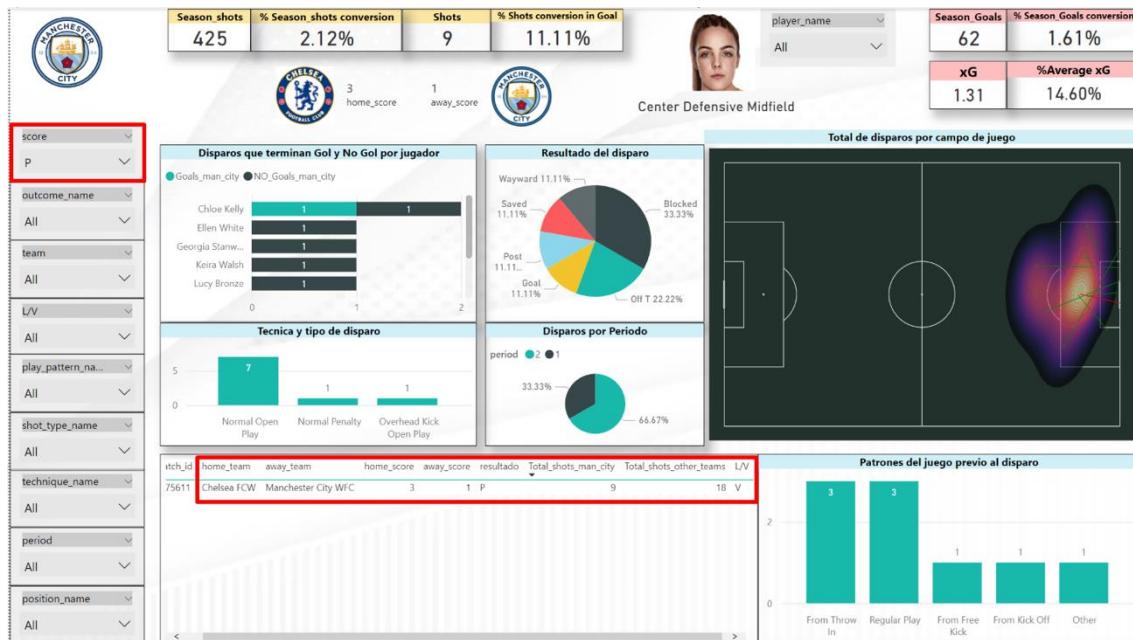
Samantha June Mewis



Fuente: Dashboard de elaboración propia en Power BI

10.5 Anexo 5 (partido perdido)

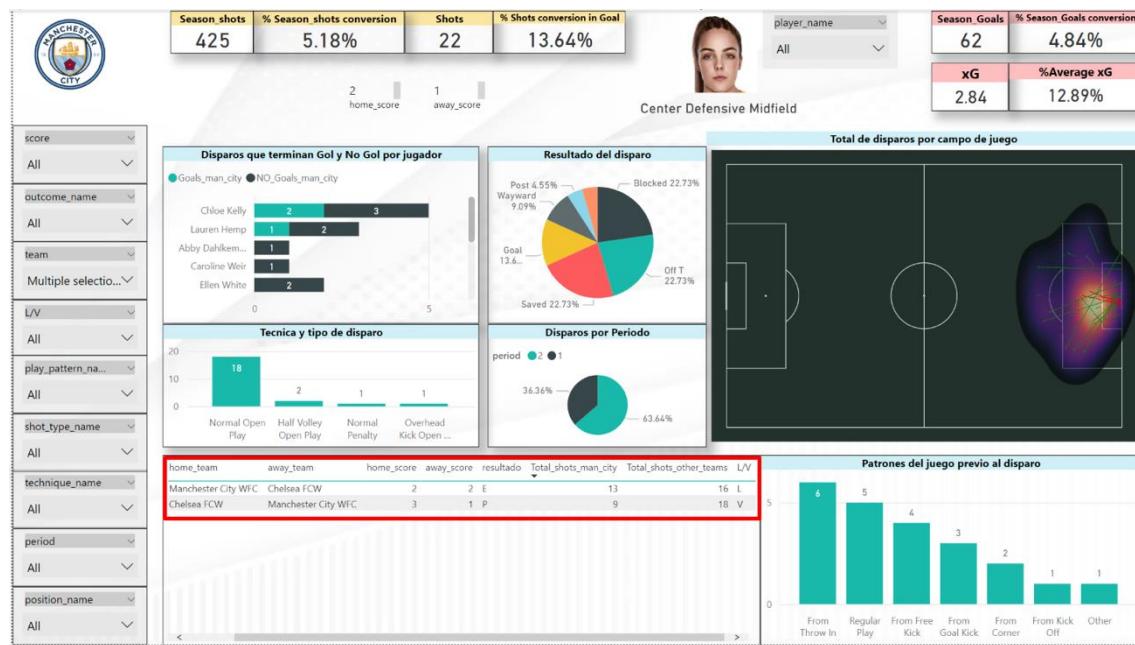
Partido perdido



Fuente: Dashboard de elaboración propia en Power BI

10.6 Anexo 6 (rival directo)

Rival directo



Fuente: Dashboard de elaboración propia en Power BI

10.7 Anexo 7 (técnica)

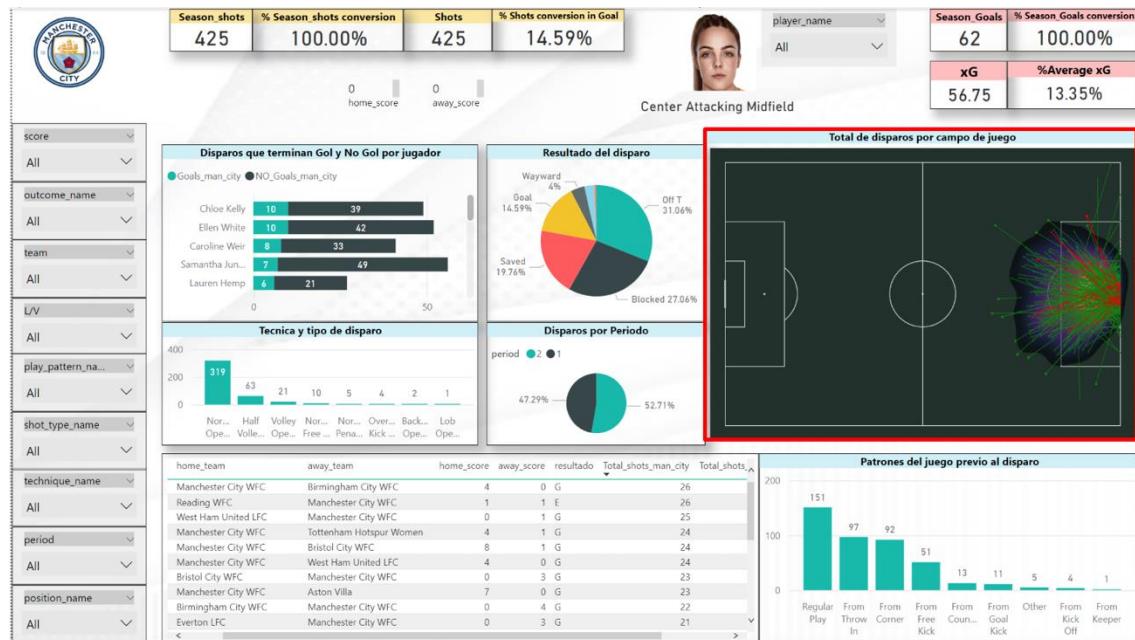
Técnica de disparo



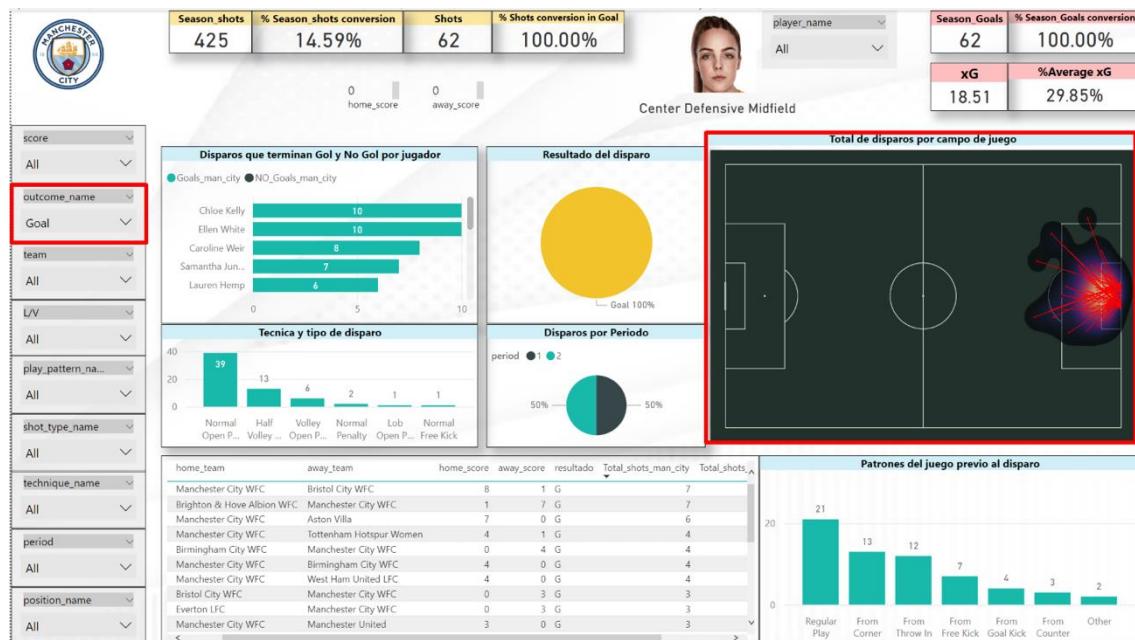
Fuente: Dashboard de elaboración propia en Power BI

10.8 Anexo 8 (posiciones)

Posiciones relevantes



Fuente: Dashboard de elaboración propia en Power BI



Fuente: Dashboard de elaboración propia en Power BI

10.9 Anexo 9 (local vs visitante)

Local



Fuente: Dashboard de elaboración propia en Power BI

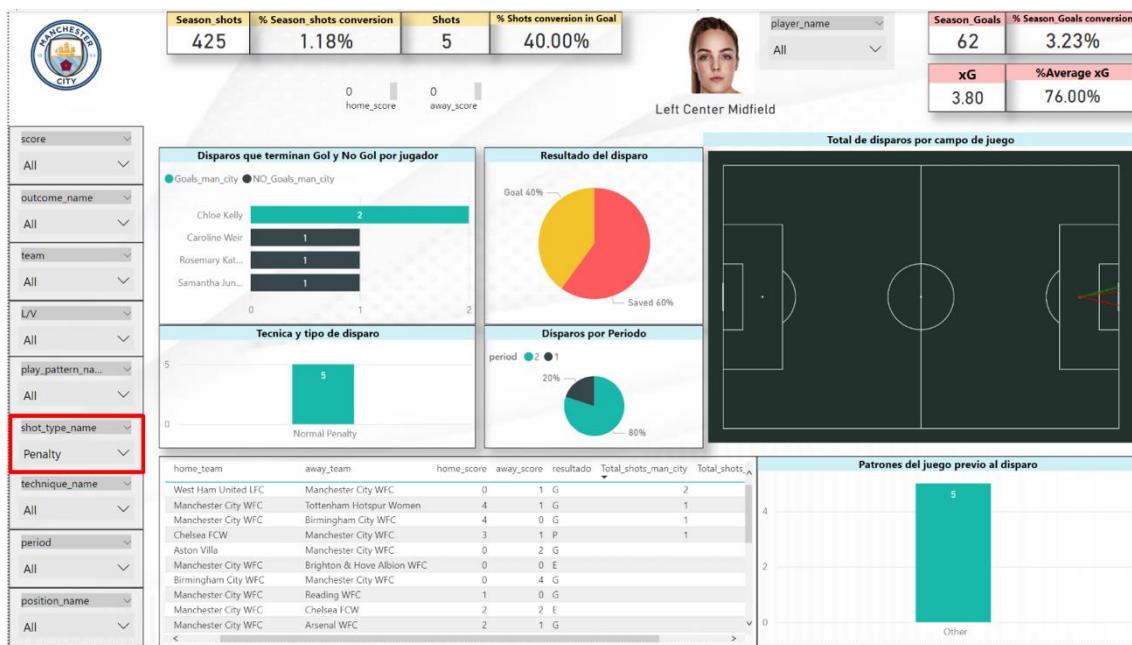
Visitante



Fuente: Dashboard de elaboración propia en Power BI

10.10 Anexo 10 (penaltis)

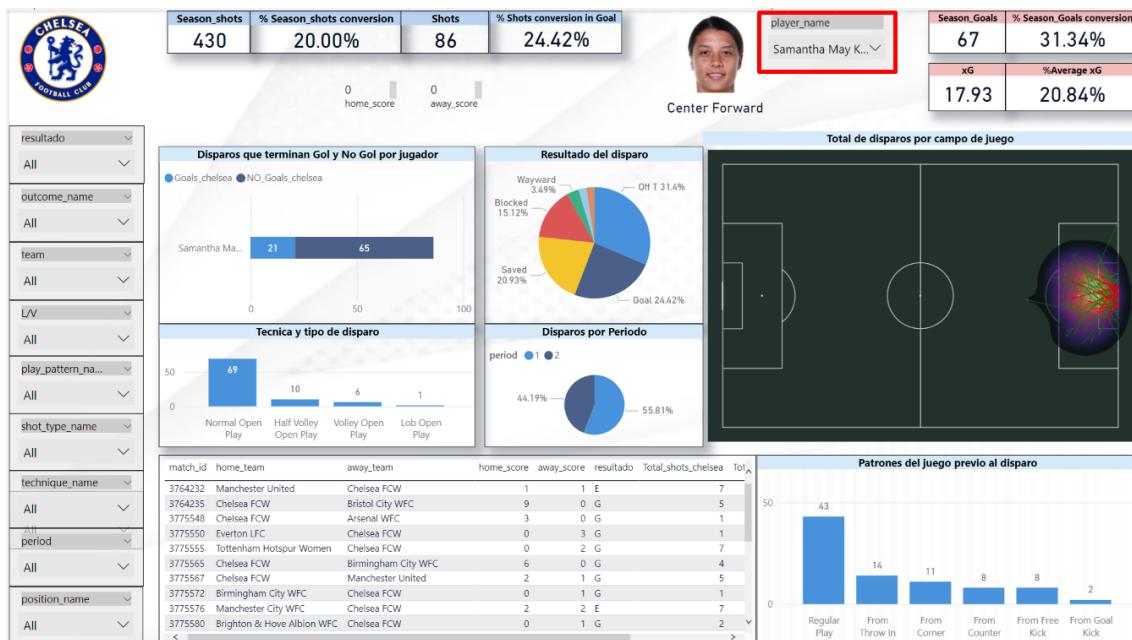
Penaltis



Fuente: Dashboard de elaboración propia en Power BI

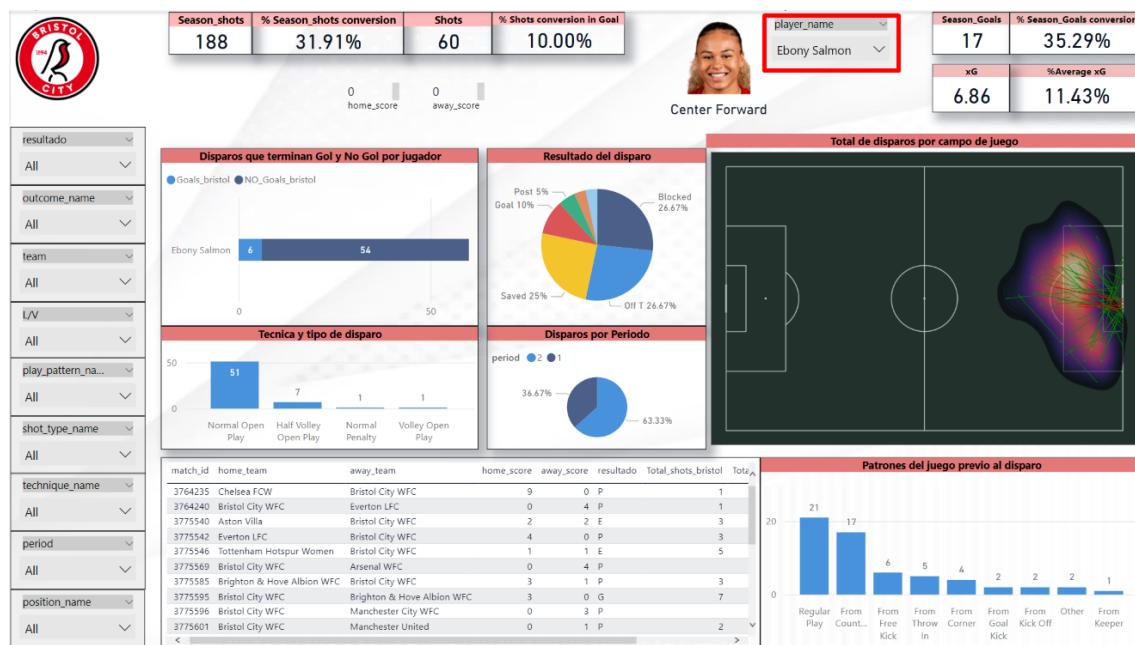
10.11 Anexo 11 (Chelsea y Bristol)

Chelsea



Fuente: Dashboard de elaboración propia en Power BI

Bristol



Fuente: Dashboard de elaboración propia en Power BI

11.**Bibliografía**

Estadísticas del Manchester City en la Women's Super League (2020-2021). FBREF. Recuperado de: <https://fbref.com/en/squads/9ce68f8a/2020-2021/Manchester-City-Women-Stats>

El City abraza el 'big data' para convertirse en uno de los mejores clubes del mundo. El confidencial. Recuperado de: https://www.elconfidencial.com/deportes/futbol/internacional/2015-07-13/manchester-city-sap-academia-etihad-campus_924913/

Premier League Title Holder Man City Uses Data To Improve Its Game. Forbes. Recuperado de: <https://www.forbes.com/sites/annatobin/2018/08/09/premier-league-title-holders-man-city-uses-data-to-improve-its-game/?sh=3ce9cf024d90>

SAP y CFG llevan 'el deporte rey' a la nube. Manchester City. Recuperado de: <https://es.mancity.com/noticias/club-news/club-news/2015/july/sap-and-city-football-group-take-the-beautiful-game-into-the-cloud>

Statsbomb Open Data. GitHub. Recuperado de: <https://github.com/statsbomb/open-data>

Explaining Expected Threat. Medium. Recuperado de: <https://soccermatics.medium.com/explaining-expected-threat-cbc775d97935>

Salarios en ciencia de datos. Glassdoor. Recuperado de:

- Chief data officer: https://www.glassdoor.es/Sueldos/chief-data-officer-sueldo-SRCH_K00,18.htm
- Project manager: https://www.glassdoor.com/Salaries/barcelona-project-manager-salary-SRCH_IL,0,9_IC2547194_KO10,25.htm
- Data engineer: https://www.glassdoor.com/Salaries/barcelona-data-engineer-salary-SRCH_IL,0,9_IC2547194_KO10,23.htm
- Data analyst: https://www.glassdoor.com/Salaries/barcelona-data-analyst-salary-SRCH_IL,0,9_IC2547194_KO10,22.htm
- Data scientist: https://www.glassdoor.com/Salaries/barcelona-data-scientist-salary-SRCH_IL,0,9_IC2547194_KO10,24.htm

Manchester City dashboard Power BI:

https://drive.google.com/file/d/1iY5rKvTpTZhJRy-Qt6yh1ywhWmWdjkWT/view?usp=drive_link

Chelsea dashboard Power BI:

https://drive.google.com/file/d/1V4Q3VdU83tNzN2_TrwUacHtJAhEsS5q_/view?usp=drive_link

Bristol dashboard Power BI:

https://drive.google.com/file/d/1N2bIWLkCXWurZBzL7K35uH4mf5Dxs2No/view?usp=drive_link