UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Big Data aplicado al fútbol internacional de selecciones: Un análisis histórico y un enfoque en la Eurocopa 2024

Trabajo Fin de Grado

Grado en Ingeniería Informática

AUTOR/A: Saborit Báguena, José

Tutor/a: García Ortega, Beatriz

Cotutor/a: Galán Cubillo, Javier

CURSO ACADÉMICO: 2023/2024

Resumen

En este Trabajo de Fin de Grado se aplican técnicas de Big Data para realizar un análisis del fútbol internacional de selecciones, combinando una perspectiva histórica con un enfoque detallado en la Eurocopa 2024. En la primera parte del trabajo se realizará un análisis histórico de los datos de partidos internacionales a lo largo de la historia, con el que se obtendrán patrones significativos en estadísticas varias como los goles por partido o el rendimiento de ciertas selecciones con el paso de los años. La segunda parte del trabajo se enfoca en la Eurocopa 2024 disputada este año en Alemania, aplicando técnicas de análisis de datos para ofrecer una visión detallada del torneo, así como el rendimiento de algunos de los equipos y jugadores participantes en ella. Junto al análisis, se explica como distintos tipos de visualizaciones pueden ser útiles a la hora de analizar equipos y futbolistas.

Palabras clave: Big Data, fútbol, análisis de datos, visualización.

Abstract

In this final degree project, Big Data techniques are applied to analyze international football at national teams' level, combining a historical perspective with a detailed focus on the UEFA Euro 2024. The first part of the project involves a historical analysis of international match data over time, aiming to uncover significant patterns in various statistics, such as goals per match and the performance of certain national teams over the years. The second part of the project focuses on the UEFA Euro 2024, hosted by Germany this summer, using data analysis techniques to provide a detailed view of the tournament, as well as the performance of some of the participating teams and players. Alongside the analysis, the project explains how different types of visualizations can be useful when analyzing teams and footballers.

Keywords : Big Data, football, data analysis, visualization.

Tabla de contenidos

Índice de gráficos

Índice de tablas

1. Introducción

El deporte es una parte fundamental de nuestras vidas, trae grandes beneficios tanto físicos como mentales a la gente que lo practica, pero esa no es la única manera de disfrutarlo. No es necesario estar dentro del campo o de la pista para apreciarlo, hay quien siente entusiasmo entendiendo cómo funcionan los deportes y observándolos desde la distancia en vez de participar de manera directa en ellos.

El fútbol es conocido como el deporte rey. Su enorme popularidad es la que le ha hecho ganarse este apodo, siendo el deporte más popular a nivel mundial, por encima de otros como el baloncesto, el tenis o los deportes de motor. La FIFA, entidad reguladora del futbol mundial, cuenta actualmente con 211 federaciones afiliadas como representantes de sus respectivos países. Sin embargo, si se observa la cantidad de jugadores, clubes o incluso ligas, estas ascienden a los millones.

Dentro del mundo del fútbol, son las competiciones de clubes las que reciben una atención de manera más constante, debido a una mayor frecuencia de partidos y a las grandes rivalidades que se forman entre equipos de todo el mundo. Los equipos más grandes, cuentan con millones de aficionados que se reparten por todos los rincones del planeta. Adicionalmente seguidores de clubes de menor tamaño también cuentan con grandes rivalidades con equipos cercanos de su nivel, generando grandes ambientes de emoción a pesar de jugarse trofeos de menor importancia.

A pesar de ello, el fútbol de selecciones cuenta con un poder único para unir a las personas de una nación bajo una misma bandera. Torneos como la Copa del Mundo o la Eurocopa, que ocurren cada pocos años, generan una expectación especial en muchos de los países, registrando las audiencias más altas del mundo de los deportes y llegando a paralizar países enteros para apoyar a su selección nacional.

En la era moderna, el análisis de datos se ha convertido, como en muchos otros ámbitos, en una herramienta esencial en el mundo del fútbol. Tanto clubes como selecciones cuentan con grandes equipos de analistas, que les permiten gozar de ventajas competitivas. Ya sean decisiones estratégicas o tácticas, u otras menos relacionadas de manera directa con el juego como la prevención de lesiones. El análisis de datos permite una toma de decisiones informadas con una base sólida.

1.1. Motivación

La realización de este trabajo se ve motivada por un fuerte interés personal por el mundo del deporte y del fútbol en concreto, así como el deseo de unir ese interés con una aplicación práctica de los conocimientos adquiridos durante la carrera.

La ciencia de datos, introducida en la rama de sistemas de la información dentro de la carrera, resulta de gran utilidad en el contexto tecnológico actual, independientemente del sector en el que se aplique.

El interés despertado por esta parte de la informática unido a la oportunidad que proporciona el verano del año en que se realiza este trabajo, con importantes competiciones de fútbol a nivel de selecciones como la Eurocopa o la Copa América, unidas a otros grandes eventos deportivos como los Juegos Olímpicos celebrados en París, son lo que generaron la motivación para llevar a cabo este proyecto de final de carrera. Poder comparar datos históricos con datos actuales resulta de gran valor a la hora de realizar un análisis de estas características.

1.2. Estructura de la memoria

Esta memoria está compuesta por nueve apartados, teniendo algunos de estos subapartados internos.

En el primer apartado se pretende introducir al lector al contexto del trabajo, con un breve resumen de la temática de este, así como explicar la motivación que llevó a la realización del trabajo, los objetivos que se pretenden conseguir con este y la metodología que se seguirá para asegurar el cumplimiento de estos objetivos y la correcta realización del trabajo, así como un diagrama de Gantt orientativo de la duración del trabajo.

A continuación, el segundo punto se utiliza para explicar el contexto tecnológico actual del análisis de datos, así como su relación con el mundo del deporte y más concretamente con el mundo del fútbol.

El tercer y cuarto punto conforman la parte principal del trabajo, en estos apartados primero se realizará una breve introducción al mundo del fútbol de selecciones, así como al torneo que se utilizará para un análisis más detallado, la Eurocopa 2024. Seguidamente se explicará en cada uno de estos apartados cómo se han obtenido los datos y los posibles tratamientos a los que se han sometido estos previos a su análisis. Por último, se realizará el propio análisis, se tratará de un análisis histórico en el caso

del tercer punto, buscando tendencias en el fútbol internacional, y un análisis más detallado en el caso del punto 4, sobre la Eurocopa 2024.

El quinto punto contendrá las conclusiones que se han obtenido durante la etapa de análisis, así como el cumplimiento de los distintos objetivos marcados al principio de este.

En el sexto punto se explicará la relación que tiene el trabajo con los estudios cursados durante la carrera, destacando las áreas de estudio con las que más se relaciona el trabajo.

En el séptimo punto se explicarán las posibles áreas de trabajo futuras que pueden derivar de este trabajo pero que no se han realizado por no entrar dentro de las dimensiones de este.

Por último, se incluirá un apartado dedicado a las referencias que se ha utilizado durante el desarrollo del proyecto (apartado 8) , así como los anexos que sean necesarios en este (apartado 9).

1.3. Objetivos

Durante la realización de este trabajo de fin de grado se realizará un análisis genérico del fútbol internacional y uno exhaustivo de la Eurocopa 2024. Mediante este análisis se pretende:

- Desarrollar los conocimientos obtenidos durante la carrera.
- Introducirse en el mundo del Big Data y comprobar que el fútbol es un deporte influenciado por este sector.
- Aprender sobre técnicas de limpieza y procesado de datos.
- Realizar un análisis histórico del fútbol internacional para identificar posibles tendencias en este.
- Analizar en profundidad los datos de la Eurocopa 2024 para observar diferentes rendimientos y extraer conclusiones.
- Analizar la influencia de jugadores en el desempeño de la selección mediante el uso de estadísticas avanzadas.
- Demostrar la utilidad de distintos tipos de representaciones de datos a la hora de analizar un deporte como el fútbol.
- Explicar como el análisis de datos puede suponer una influencia positiva para distintos equipos o selecciones que utilicen estas herramientas.

1.4. Metodología

Para poder llevar a cabo los análisis propuestos en este trabajo, en primer lugar, se deberá de realizar una búsqueda de bases de datos con registros históricos sobre el fútbol internacional al igual que estadísticas más avanzadas sobre la Eurocopa 2024. En el caso de este trabajo, se realizan búsquedas en las webs de Kaggle, StatsBomb y FBRef en búsqueda de las mejores fuentes de datos para los análisis a realizar.

Una vez encontradas las tablas con los datos sobre los que se realizará el estudio, estas se tendrán que cargar en un programa de análisis estadístico. En este caso el programa elegido es RStudio, se trata de un programa que actúa como un entorno de desarrollo integrado (IDE) sobre el lenguaje de programación R y que está diseñado específicamente para el análisis de datos y la estadística.

Antes de empezar con en análisis, estos datos pasarán por una etapa de preprocesamiento, donde pasarán un proceso de limpieza y transformación que sea necesario para un estudio satisfactorio sobre ellos.

En el análisis histórico del fútbol de selecciones se buscarán tendencias más generales mientras que en el análisis específico de la Eurocopa 2024 se indagará más en los datos, buscando utilizar estadísticas más avanzadas para determinar la influencia de ciertos aspectos del juego o jugadores en determinadas selecciones que hayan destacado durante el transcurso de este torneo. Aspectos del juego ofensivos y defensivos, así como el rendimiento individual de distintos jugadores en distintas partes del campo darán una visión completa del transcurso del torneo.

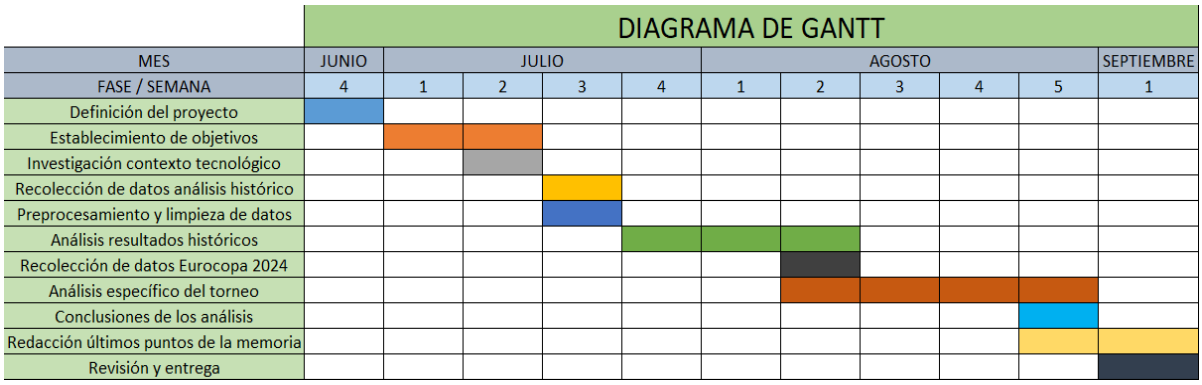
Se aprovechará el primer análisis histórico para dar entrada a la Eurocopa, explicando el contexto del fútbol internacional actual y su influencia en este torneo.

Durante todo el trabajo, se mantendrá una documentación de los pasos realizados durante el análisis que se verá plasmada en esta memoria.

El programa elegido, RStudio, nos permitirá, gracias a su gran versatilidad, tanto un tratamiento de datos cómodo, como un buen análisis de estos, así como una visualización variada gracias a distintas librerías disponibles para instalar.

A continuación, se muestra un diagrama de Gantt con las distintas fases del trabajo y su duración aproximada, el trabajo fue completado en 11 semanas:

Gráfico 1: Diagrama de Gantt del trabajo



Fuente: Elaboración propia

2. Contexto tecnológico

El análisis de datos es el proceso de examinar, limpiar, transformar y modelar datos con el objetivo de extraer información que resulte útil para llegar a conclusiones y respaldar la toma de decisiones. Durante este proceso, es necesario aplicar distintas técnicas y métodos para extraer patrones, tendencias o correlaciones del conjunto de datos estudiado.

El análisis de datos ha ganado relevancia durante los últimos años gracias a su capacidad de apoyar a las empresas a la hora de tomar decisiones más informadas, basadas en los datos. Esta ciencia permite a las empresas mejorar su rendimiento y eficiencia, identificando las áreas que pueden ser mejoradas y tomar medidas para mejorar la productividad y la rentabilidad.

Dentro del mundo del deporte, el Big Data ha revolucionado la manera en la que se enfocan muchos deportes, como el fútbol, el baloncesto, el béisbol o el tenis, empleando el análisis de datos para optimizar entrenamientos y diseñar estrategias. Encontramos dos aplicaciones principales de estas tecnologías: El análisis del rendimiento, permitiendo estudiar detalladamente el rendimiento de los atletas; y la creación de estrategias de juego, utilizando los datos para diseñar estrategias tácticas y estudiar al oponente para encontrar posibles debilidades, así como mejorar la toma de decisiones durante el juego.

Uno de los mejores ejemplos de la aplicación del análisis estadístico en el mundo del deporte es el caso de los Oakland Athletics, un equipo de la liga de béisbol americana. Este equipo comenzó a usar una técnica llamada SABRmetrics, basada en el análisis empírico de las estadísticas del béisbol, durante la década de 1990. Su principal promotor fue Billy Beane, quien asumió el cargo de gerente general del equipo en 1997 y utilizó los principios de esta técnica para fichar jugadores infravalorados, gracias a la ayuda de Paul DePodesta, graduado en economía de Harvard. Este equipo hizo historia en 2002, siendo el primer equipo en la historia de la MLB en ganar 20 partidos consecutivos. Adicionalmente, bajo el liderazgo de Beane, los Athletics se convirtieron en uno de los equipos más eficientes respecto a sus costes de la liga. Un ejemplo de esto es la temporada 2006, cuando fueron 5º en la temporada regular, sin embargo, se encontraban en la 24º posición respecto al ranking de salarios en la liga de 30 equipos. Los logros de Beane fueron plasmados en la obra titulada “Moneyball: The Art of

Winning an Unfair Game" (Lewis, 2003), que posteriormente fue adaptada a la gran pantalla en la película "Moneyball", dirigida por Bennet Miller y protagonizada por Brad Pitt.

Este es solo uno de los ejemplos de cómo el análisis de datos puede mejorar la eficiencia de un equipo. En el caso del fútbol, el enfoque en el análisis de datos se ha ido popularizando con el tiempo (Memmert & Raabe, 2018). Hoy en día se poseen grandes cantidades de datos, tanto propios como del rival, en distintas áreas: Datos biométricos, muchas veces recolectados gracias a equipamiento específico que cargan los jugadores durante los entrenamientos y los partidos; datos técnicos, como goles y asistencias o algunos más detallados como las intercepciones o los metros recorridos por un jugadores durante un partido; y datos tácticos, sobre cómo afectan a los resultados de un equipo la alineación escogida por el entrenador o si juega de local o de visitante.

Toda esta información está al alcance de los clubes profesionales hoy en día, pero tan importante es disponer de estos datos como tener a alguien capaz de interpretarlos en el cuerpo técnico de un club o selección nacional. Es por ello que hoy en día prácticamente todos los grandes equipos de fútbol y selecciones nacionales cuentan con un jefe de análisis de datos. Uno de los ejemplos más recientes en el mundo del fútbol es el del Liverpool, equipo de la primera división inglesa, que cuenta con uno de los departamentos de analítica más avanzados del mundo del deporte y está aliado con el departamento de inteligencia artificial de Google, creando un modelo que les llevó a la consecución de la Champions League en 2019 y a alzarse con su primer título de Premier League (habían ganado previamente la primera división inglesa, pero no lo conseguían desde el cambio de formato de la competición en la temporada 1992/1993). Otro caso es el del Barcelona, equipo que ha lanzado recientemente el "Barça Innovation Hub", un centro global para el conocimiento, la analítica y la innovación en el deporte.

Como se ha destacado, es de vital importancia para los clubes y selecciones contar con un equipo de analistas que permitan obtener ventajas competitivas sobre sus rivales, sin embargo, este sector no solo ha ganado importancia dentro de los cuerpos técnicos. El fútbol es el deporte más seguido del planeta, y los torneos internacionales en especial consiguen las mayores audiencias del mundo del deporte. Cada vez más gente está interesada en el fútbol, pero ese interés no se queda únicamente en ver los partidos de tu equipo favorito o de la selección de tu país. En los últimos años han surgido numerosas páginas web dedicadas específicamente a la analítica de datos en el mundo del fútbol, páginas como Opta Analyst, FBRef, WhoScored o StatsBomb son las más

conocidas y cuentan con grandes repositorios de datos disponibles para realizar análisis sobre el deporte rey.

Para el público más general, interesado en las estadísticas, pero en menor medida, los organismos encargados del fútbol profesional también ponen a su disposición nuevos servicios innovadores en el mundo del deporte. Uno de los casos más importantes en este sentido es el acuerdo que alcanzaron La Liga de Futbol Profesional española con Microsoft, para convertir a la marca tecnológica en su “Tech and Innovation Partner” a nivel global en 2021.

Gracias a este acuerdo, ambas marcas buscaban crear una nueva experiencia para el usuario, gracias a la creación de soluciones tecnológicas innovadoras en la industria. Estas nuevas soluciones están enfocadas tanto para los aficionados, con retransmisiones enriquecidas con datos y opciones de streaming personalizadas, así como nuevas experiencias de realidad virtual y aumentada; como para los titulares de los derechos de retransmisión y los gestores de los recintos deportivos, permitiéndoles una mejor organización y una experiencia reforzada con contenidos adicionales y otros servicios como las redes sociales. Todas estas soluciones informáticas, basadas en los datos, son impulsadas por las tecnologías de Microsoft Azure, PowerBI y Microsoft 365.

Esta alianza también llevó a la creación de “Beyond Stats”, un proyecto de análisis futbolístico avanzado que pretende profundizar en el juego de cada equipo. Este se trata de un portal web disponible al público a nivel global, presentándole todo tipo de estadísticas avanzadas sobre esta competición.

A nivel de fútbol internacional, la UEFA comenzó en 2016 a presentar de cara al público los “Technical Reports” sobre sus competiciones, documentos donde detallaba a nivel técnico el transcurso de estas. La FIFA por su parte, buscó implementar nuevas mejoras en este ámbito en el último mundial, celebrado en Catar en 2022. Para ello, contaron con un equipo de 25 analistas de datos por partido atentos a cada acción que ocurrió en el campo, permitiendo aumentar el promedio de eventos que se registran en un partido de unos 2.500 hasta alrededor de los 15.000. Con esto, la FIFA buscaba ofrecer distintos niveles de datos, tanto para los medios encargados de la retransmisión de los partidos, como para los espectadores, tanto en análisis posteriores al partido, como en tiempo real.

3. Análisis de resultados históricos

3.1. Introducción al fútbol internacional de selecciones

El origen del fútbol como lo conocemos hoy en día se remonta al año 1863, cuando se fundó en el Reino Unido “The Football Association”, organismo encargado de oficializar las primeras reglas del deporte.

Sin embargo, no fue hasta el año 1872 en el que se disputó el primer partido entre selecciones de distintos países, enfrentando, como no podía ser de otra manera, a las selecciones de Inglaterra y Escocia. Estas selecciones, junto al resto de países británicos, fueron los pioneros de este deporte, disputando entre ellos los primeros amistosos entres selecciones y creando la primera competición entre selecciones nacionales, la “British Home Championship”.

A principio del siglo XX, el fútbol comenzaba a ganar popularidad y se comenzaban a formar asociaciones nacionales de fútbol por el mundo, llegando a ser reconocido por el Comité Olímpico Internacional como deporte olímpico para los Juegos Olímpicos (JJOO) del 1900, aunque los partidos disputados solo eran para jugadores aficionados.

Sin embargo, no fue hasta la creación de la FIFA, organismo encargado desde su fundación en el 1904 hasta la actualidad de la organización del fútbol a nivel global, que las selecciones de otros países comenzaron a disputar partidos profesionales entre ellas, a excepción de un par de amistosos disputados en Norteamérica entre Estados Unidos y Canadá y en Sudamérica entre Uruguay y Argentina en los años anteriores.

Tras esto, el fútbol se expandió rápidamente por todos los rincones del planeta, llevando a la FIFA a la creación de su competición más exitosa, y el evento deportivo que más espectadores atrae hasta hoy en día, la Copa Mundial de Fútbol, disputada por primera vez en el año 1930 en Uruguay, debido a que fue el país ganador de las últimas dos ediciones de los JJOO. La creación de esta competición llegó tras numerosos desacuerdos entre la FIFA y el Comité Olímpico Internacional, especialmente por las reglas sobre el profesionalismo de los jugadores partícipes y la exclusión del deporte en el programa de los juegos de 1932, debido a la baja popularidad del deporte en Estados Unidos, donde se disputaban.

Debido a la gran cantidad de federaciones nacionales afiliadas a la FIFA, este organismo decidió en 1953 dar luz verde a la creación de confederaciones continentales de fútbol. Tan solo un año después, se creó la UEFA, organismo encargado del control del fútbol en Europa y que creó en esa misma década dos de las competiciones más exitosas: La Copa de Europa para los clubes, que posteriormente pasaría a llamarse la Champions League y está considerada hoy en día como la competición de clubes más prestigiosa del mundo; y el Campeonato de Europa para selecciones nacionales, a la que se conoce hoy en día como la Eurocopa.

Tras la UEFA, se fundaron confederaciones en el resto de los continentes: Se fundó la Conmebol para el fútbol sudamericano, la Concacaf para el futbol norteamericano y centroamericano, la CAF para el futbol africano, la AFC para el futbol asiático y la OFC para el futbol de Oceanía, todas ellas avaladas por la FIFA y con sus propias competiciones continentales.

Gráfico 2: Confederaciones filiales de la FIFA

Fuente: elordenmundial.com

Hoy en día, tanto la FIFA como sus subdivisiones continentales se encargan de dar apoyo al desarrollo de este deporte, con ayudas financieras y logísticas para sus países afiliados, a cambio de que estos respeten sus estatutos y actúen de acuerdo con los objetivos del organismo internacional.

3.2. Obtención y preprocesado de los datos

Para dar comienzo a la etapa de análisis de este apartado, primero se debe encontrar un conjunto de datos que sea útil y que se considere fiable. Tras realizar una búsqueda por distintas páginas dedicadas a almacenar repositorios de datos se recurre a Kaggle, una plataforma web donde se reúnen miles de usuarios con experiencia en el análisis de datos para plantear sus análisis y problemas.

Dentro de esta plataforma, se encuentra un dataset muy completo llamado “International football results from 1872 to 2024” creado por el usuario Mart Jurisoo, que se encarga de su mantenimiento, con una frecuencia de actualización aproximada de un mes. En este dataset se incluyen, según el autor, 47.126 resultados de partidos disputados entre selecciones nacionales masculinas. Se incluyen tanto encuentros amistosos como partidos de las competiciones internacionales más importantes como la Eurocopa o la Copa del Mundo. Por otro lado, se excluyen los partidos de los Juegos Olímpicos (sólo se incluyen las ediciones de 1924 y 1928, reconocidas por la FIFA como los primeros mundiales al seguir sus reglas de competición), partidos de categorías inferiores como juveniles o Sub-23, o partidos en los que al menos uno de los equipos fuera el equipo reserva o un club en lugar de una selección nacional.

El dataset está compuesto por tres archivos csv: *results* (con los resultados de los partidos), *shootouts* (en esta hoja se incluyen las tandas de penaltis de aquellos partidos que se decidieron mediante este método) y *goalscorers* (con los autores de cada gol marcado en partidos de competiciones oficiales, sin incluir amistosos ni competiciones de menor nivel).

Las columnas de cada tabla son las siguientes:

- results:
    - date: fecha en la que se disputó el partido
    - home_team: selección que jugaba como local
    - away_team: selección que jugaba como visitante
    - home_score: goles marcados por el equipo que actuaba como local, incluyendo prórrogas y sin incluir tandas de penaltis
    - away_score: goles marcados por el equipo que actuaba como visitante, incluyendo prórrogas y sin incluir tandas de penaltis
    - tournament: nombre de la competición en la que se disputaba el encuentro, incluyendo amistosos

    - city: nombre de la ciudad donde se disputaba el encuentro
    - country: nombre del país donde se disputaba el encuentro
    - neutral: columna de tipo booleano que indica si el partido se jugaba en un estadio neutral

- shootouts:
    - date: fecha del partido
    - home_team: selección que jugaba como local
    - away_team: selección que jugaba como visitante
    - winner: ganador de la tanda de penaltis
    - first_shooter: equipo que empezó lanzando en la tanda de penaltis

- goalscorers:
    - date: fecha del partido donde se metió el gol
    - home_team: selección que jugaba como local
    - away_team: selección que jugaba como visitante
    - team: equipo que metió el gol
    - scorer: nombre del jugador que metió el gol
    - own_goal: booleano que indica si el gol fue en propia puerta o no
    - penalty: booleano que indica si el gol fue de penalti

El autor destaca que para la columna *country* de la hoja *results*, se utiliza el nombre del país al que pertenecía esa ciudad en el momento en el que se disputó el partido, por lo que se dan instancias en las que los nombres de *home_team* y *country* no coinciden, a pesar de ejercer este de local, por lo que se indica en la columna *neutral* como FALSE.

Una vez se ha decidido los datos sobre los que se va a trabajar en este caso, el siguiente paso es cargarlos a un programa adecuado para un análisis estadístico de estos. Como se mencionó anteriormente, para esta primera parte se va a utilizar exclusivamente el programa RStudio, ya que esta parte no se centrará en la visualización interactiva de los datos, si no en buscar tendencias en la historia del futbol internacional.

En primer lugar, debemos configurar el directorio de trabajo de RStudio, este debe ser la carpeta en donde se encuentran los archivos csv con los datos. Esto se realiza mediante el comando setwd(“ruta del directorio”).

Para ejecutar comandos en RStudio hay varias opciones, una de ellas es escribirlos directamente en la consola, disponible en una de las ventanas del programa. Sin embargo, para este trabajo crearemos un nuevo archivo de tipo R Script, donde introduciremos todos los comandos que vayamos utilizando, ya que este archivo facilita la lectura del código y se puede guardar.

Una vez configurado, es hora de cargar los tres archivos en R, esto se realiza con el comando read.csv("nombre del archivo", stringsAsFactors = FALSE). Ejecutaremos esta instrucción tres veces, una por cada archivo, asignándole a cada una un nombre, en este caso el mismo que tenía el archivo original. El último argumento asegura que las columnas de tipo string se importen como caracteres en lugar de factores o categorías, evitando futuros problemas de manipulación con estos datos. Tras la ejecución de esta instrucción, ya tendremos nuestros datos cargados en el entorno de programación.

Una vez cargados los datos, estos deben de pasar por una etapa de preprocesado, que resulta de vital importancia, donde se debe comprobar la estructura de estos para que esta no de problemas posteriormente en la etapa de análisis. Durante el preprocesado, se deberá establecer el manejo de datos faltantes en las hojas de datos, decidiendo que hacer con ellos: omitirlos, rellenarlos con valores como la media o la mediana o dejarlos como están si estos valores nulos son significativos de algo en particular. También corresponderá a esta etapa la creación de nuevas variables que puedan aportar valor al análisis realizado, así como posibles conversiones de tipos de datos, manejo de datos duplicados, corrección de valores o la estandarización de estos si así lo requiere el estudio.

Para empezar esta etapa, primero vamos a comprobar si la estructura de los datos es correcta. Para ello utilizaremos la función str(nombre), que nos devuelve la estructura de cada hoja de datos, por ejemplo, para str(results) obtenemos la siguiente estructura:

Tabla 1: Estructura de la hoja results

```
> str(results)
'data.frame':   47379 obs. of  9 variables:
 $ date      : chr  "1872-11-30" "1873-03-08" "1874-03-07" "1875-03-06" ...
 $ home_team : chr  "Scotland" "England" "Scotland" "England" ...
 $ away_team : chr  "England" "Scotland" "England" "Scotland" ...
 $ home_score: int  0 4 2 2 3 4 1 0 7 9 ...
 $ away_score: int  0 2 1 2 0 0 3 2 2 0 ...
 $ tournament: chr  "Friendly" "Friendly" "Friendly" "Friendly" ...
 $ city      : chr  "Glasgow" "London" "Glasgow" "London" ...
 $ country   : chr  "Scotland" "England" "Scotland" "England" ...
 $ neutral   : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

Fuente: Elaboración propia

Como podemos observar tras la ejecución de esta instrucción, esta hoja contiene 47.379 observaciones (partidos), con 9 variables. Si nos fijamos, todas parecen tener la estructura correcta a excepción de *date,* la cual es de tipo char mientras que debería de ser de tipo date. Esto ocurre en las tres hojas de datos, por lo que vamos a intentar solucionarlo. Para ello, ejecutaremos la siguiente instrucción:

```
results$date <- as.Date(results$date, format="%Y-%m-%d")
```

Esta función se ejecutará tres veces, una por cada hoja, es importante especificar el formato en el que viene la fecha para una correcta conversión de tipo. Como se puede observar en la Tabla 1, esta viene en formato YY-MM-DD. Una vez ejecutada se puede volver a ejecutar la instrucción `str` para comprobar que el cambio de tipo se ha completado correctamente.

Una vez comprobados los tipos de datos, se va a continuar con la identificación de valores nulos. Para ello se va a utilizar la siguiente instrucción: `colSums(is.na(nombre))` . Tras su ejecución se obtiene la siguiente tabla:

Tabla 2: Valores nulos en los datasets

```
> colSums(is.na(results))
      date  home_team  away_team home_score away_score tournament       city    country    neutral
         0         25         25         84         84          0          0          0          0
> colSums(is.na(shootouts))
         date     home_team     away_team        winner first_shooter
            0             0             0             0             0
> colSums(is.na(goalscorers))
     date home_team away_team      team    scorer    minute  own_goal   penalty
        0         0         0         0        49       259         0         0
```

Fuente: Elaboración propia

Varios de estos valores llaman la atención. En primer lugar, resulta extraño que en la hoja *results* coincidan los valores nulos tanto de los equipos como los goles locales y visitantes. En segundo lugar, en la revisión de la estructura de *shootouts*, se observó que la columna *first_shooter* se encontraba vacía en muchos casos, sin embargo, RStudio no ha detectado estas celdas como valores nulos. Por último, en la tabla *goalscorers* encontramos valores nulos en las columnas *scorer* y *minute*.

Para arreglar estos problemas, se van a inspeccionar los distintos datasets. En primer lugar, vamos a observar las filas en las que se encuentran estos valores nulos. Para ello, utilizaremos la siguiente instrucción, por ejemplo, para visualizar los datos nulos de la columna *home_score* de *results*: `results[is.na(results$home_score), ]`

Una vez ejecutada esta instrucción, se observa que los datos faltantes en la hoja *results* corresponden a los partidos disputados durante el último mes, en las competiciones de la Eurocopa y la Copa América disputadas en Julio. Esto se debe a que el dataset que

se estaba usando se descargó al inicio del trabajo y estos partidos estaban programados y, por lo tanto, incluidos en la hoja de datos. Para solucionarlo se vuelven a descargar y cargar en RStudio los archivos de datos y se les aplica el cambio de tipo a las columnas *date*.

Una vez actualizados los datasets, ya no se encuentran valores nulos en la hoja de *results.* A continuación, se va a investigar lo ocurrido con la columna *first_shooter* de la hoja de *shootouts*, ya que con un simple comando como head(shootouts) , que muestra las primeras cinco filas de la tabla, se puede observar cómo hay bastantes valores nulos en esta. Esto seguramente esté causado por que el programa no detecte una cadena vacía como valor nulo, por lo que vamos a sustituir para la tabla *shootouts* todos las cadenas vacías por valores nulos con la instrucción: shootouts[shootouts == ""] <- NA

Tras la ejecución de esa instrucción, ahora nos encontramos que la columna cuenta con 414 valores nulos, valor considerablemente alto para las 644 filas de la tabla. No eliminaremos esta columna debido a que las observaciones que si que contienen valor pueden resultarnos de utilidad a la hora de analizar, por ejemplo, si el primer lanzador en una tanda de penaltis tiene más posibilidades de ganar la tanda que el que lanza segundo. Pese a ello, debemos remarcar cuando hagamos el análisis la gran cantidad de datos nulos de esta columna.

Por último, observamos que los valores nulos de la hoja de *goalscorers* corresponden a partidos de entre 1960 y 1980 disputados entre selecciones africanas y asiáticas de menor nivel. Seguramente estos valores nulos sean causados por la dificultad del autor original del dataset de encontrar los valores exactos, pero la falta de estos no debería de suponer un gran obstáculo en el análisis al ser un número relativamente bajo frente a las observaciones totales. Se decide no eliminar estas filas ya que el resto de sus atributos resultan útiles para el análisis.

Cabe destacar tras una observación inicial de los datos, que las 3 tablas de datos no empiezan desde el mismo partido. La tabla *results* cuenta con todos los partidos disputados desde el inicio del futbol de selecciones en 1872, sin embargo, la tabla *goalscorers* comienza en 1916, con los partidos de la primera Copa América (competición entre las selecciones de Sudamérica, equivalente hoy en día a la Eurocopa) debido a que el autor del dataset no incluye amistosos ni competiciones de menor nivel, como las disputadas durante los primeros años del fútbol de selecciones. La tabla *shootouts* comienza en 1967 (a pesar de que la FIFA no aprobara este método de desempate hasta 1970) en un torneo asiático en Malasia.

Por último, durante esta etapa resulta interesante la creación de nuevas variables que puedan resultar útiles para su posterior análisis. A continuación, se detallan las nuevas variables creadas, lo que representan y la instrucción para su creación:

- *goal_difference*: Diferencia de goles anotados entre el equipo local y el visitante. results$goal_difference <- results$home_score - results$away_score
- *match_result*: Indica si el resultado del partido fue victoria local, visitante o empate. results$match_result <- ifelse(results$goal_difference > 0, "Home Win", ifelse(results$goal_difference < 0, "Away Win", "Draw")).
- *confederation*: Confederación de la FIFA a la que corresponde cada selección. Esta variable puede resultar de gran utilidad a la hora de comprar el rendimiento entre confederaciones. Para ello, primero hay que comprobar cuantas selecciones distintas hay en el dataset, se agrupan juntando en un vector todos los nombres de los equipos locales y visitantes y eliminando valores duplicados con las siguientes instrucciones:

  all_teams <- c(results$home_team, results$away_team)

  unique_teams <- unique(all_teams)

  Una vez se tienen todos los nombres de las selecciones que aparecen, se procede a crear un Excel con dos columnas: El nombre, exactamente como está escrito en el dataset *results,* y la confederación a la que pertenece dicho país. Cabe destacar que se obtuvieron 336 nombres distintos, debido a que el dataset incluye partidos entre selecciones no reconocidas por la FIFA y que no son miembros de ninguna confederación, muchas de ellas por ser selecciones que representan a una sola región de un país, como la selección catalana o la andaluza en el caso de España. Este proceso se lleva a cabo manualmente, ya que no se encuentra un archivo con estos datos y para asegurar que la semántica sea exactamente igual a la del dataset con el que se está trabajando.

  Una vez obtenido el Excel, este se importa a RStudio, se procede a crear un dataset de *results* filtrado para que solo aparezcan partidos disputados entre selecciones reconocidas por la FIFA o alguna de sus confederaciones. Adicionalmente, se creará en este dataset una columna para la confederación de la selección local y otra para la de la visitante.

  A continuación, se explica el proceso para llevar a cabo lo explicado anteriormente:

```
# Filtrar results para mostrar partidos entre equipos de la FIFA

filtered_results <- results[results$home_team %in% Confederaciones$country &
results$away_team %in% Confederaciones$country, ]

# Añadir confederación para la seleccion local

filtered_results <- merge(filtered_results, Confederaciones, by.x = "home_team",
by.y = "country", all.x = TRUE)

names(filtered_results)[names(filtered_results) == "confederation"] <-
"home_confederation"

# Añadir confederación para la seleccion visitante

filtered_results <- merge(filtered_results, Confederaciones, by.x = "away_team",
by.y = "country", all.x = TRUE)

names(filtered_results)[names(filtered_results) == "confederation"] <-
"away_confederation"
```

El resultado es un nuevo dataset, con 44.238 variables y 13 columnas.

- *decade*: Indica la década en la que se disputó el partido, se realiza para todos los datasets con los que se trabaja, un ejemplo con la tabla *results* es:

```
results$year <- format(results$date, "%Y")
results$decade <- paste0(substr(results$year, 1, 3), "0s")
results$year <- NULL
```

Una vez obtenidas estas nuevas variables, que nos serán útiles en la siguiente etapa de análisis, se va a dar por finalizada la etapa de preprocesado de los datos. Esto no significa que posteriormente durante el análisis, no se creen o eliminen columnas y filas según los objetivos que se pretendan alcanzar con este.

3.3. Análisis histórico

Da comienzo ahora la siguiente etapa, donde se pretende realizar un análisis basado en los resultados de los partidos disputados entre selecciones de distintos países. Con este análisis se busca encontrar tendencias que expliquen la evolución del fútbol internacional, así como poner ejemplos de las distintas funcionalidades que puede ofrecer un análisis de estas características.

Durante esta etapa, vamos a trabajar principalmente con la tabla que se creó en el apartado anterior *filtered_results*, ya que nos permitirá una mayor precisión en el análisis que la tabla *results*, al no contar con partidos de selecciones no afiliadas a la FIFA. Cabe destacar que la tabla *golascorers* y la tabla *shootouts* solo tienen datos sobre competiciones FIFA, por lo que no ha sido necesario filtrarlas.

Antes de empezar esta etapa, vamos a instalar la librería de R ggplot2, que nos permitirá utilizar funciones útiles para una visualización simple de los datos analizados.

Una vez instalada, para empezar con el análisis, vamos a repasar los datasets un poco por encima, para ello empezaremos utilizando la instrucción summary(dataset), esta instrucción nos devuelve un resumen de cada columna de la tabla seleccionada, es útil especialmente para las columnas de tipo int, date o logical, al devolver datos como la media, mediana o valores máximos y mínimos de cada una. Vamos a observar el resumen de los datos de *filtered_results:*

Tabla 3: Resumen columnas filtered_results

```
     date             home_score       away_score
Min.   :1872-11-30   Min.   : 0.000   Min.   : 0.000
1st Qu.:1979-04-15   1st Qu.: 1.000   1st Qu.: 0.000
Median :1999-03-17   Median : 1.000   Median : 1.000
Mean   :1993-01-15   Mean   : 1.739   Mean   : 1.166
3rd Qu.:2011-11-10   3rd Qu.: 2.000   3rd Qu.: 2.000
Max.   :2024-07-14   Max.   :31.000   Max.   :21.000
 neutral        goal_difference    match_result
Mode :logical   Min.   :-21.0000   Length:44238
FALSE:32797     1st Qu.: -1.0000   Class :character
TRUE :11441     Median :  0.0000   Mode  :character
                Mean   :  0.5729
                3rd Qu.:  2.0000
                Max.   : 31.0000
```

Fuente: Elaboración propia

Como se puede apreciar, las columnas de tipo char no fan mucha información más allá del número de observaciones (solo se ha incluido una de este tipo como ejemplo). Sin

embargo, esta instrucción nos resulta útil para ver rápidamente datos como el mayor número de goles anotados por un equipo local (31) o por un equipo visitante (21), la media de goles anotados por cada equipo (se puede observar que la media de goles del equipo local es considerablemente mayor que la del equipo visitante), con la que se puede intuir que la media de la diferencia de goles es a favor del equipo local, como se puede observar en la tabla. También nos permite ver rápidamente el número de veces que una variable de tipo lógico ha resultado verdadera o falsa, viendo que alrededor de un cuarto de los partidos se disputaron en un campo neutral.

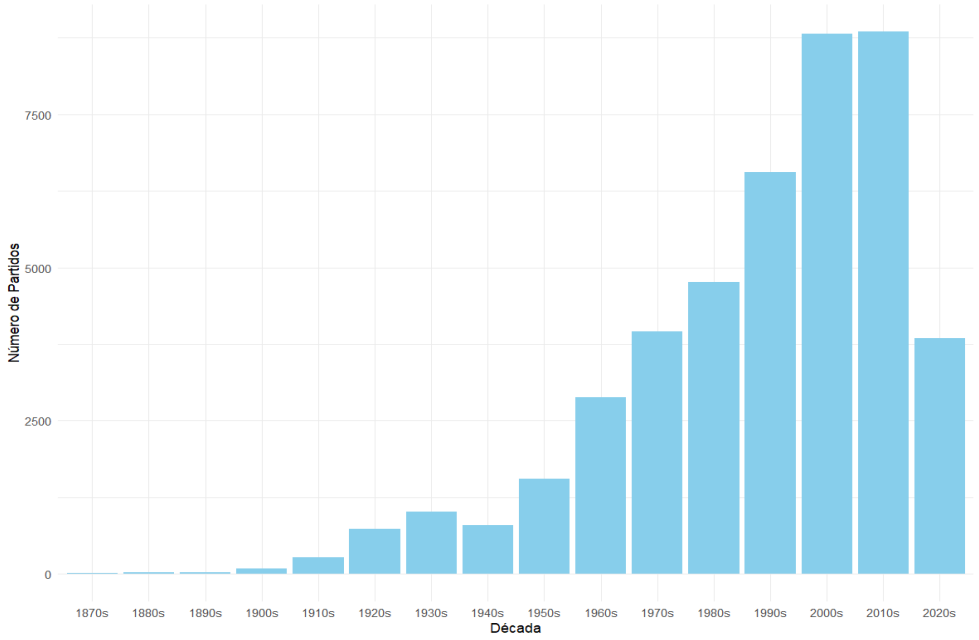
Ahora vamos a observar el número de resultados de los que disponemos por década. En primer lugar, creamos una tabla con la suma de partidos de cada década y luego creamos un gráfico de barras para facilitar su visualización.

Las instrucciones para realizar lo descrito anteriormente son las siguientes:

```
matchs_per_decade <- table(filtered_results$decade)

ggplot(data = as.data.frame(matchs_per_decade), aes(x = Var1, y = Freq)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(x = "Década", y = "Número de Partidos")
```

El gráfico resultante es el siguiente:

Gráfico 3: Numero de resultados por década



Fuente: Elaboración propia

Como podemos observar, el número de partidos internacionales disputados ha ido creciendo con el paso de los años, siguiendo la tendencia actual del mundo del fútbol, donde cada vez se disputan más partidos y los jugadores tienen calendarios más congestionados. Cabe destacar que el resultado de la década actual es menor al no haber llegado ni a la mitad de la década en el momento en el que se realiza este trabajo.

A continuación, vamos a ver la distribución de los resultados, para ello utilizaremos la variable creada durante el preprocesado *match_result* , esta variable nos permite comprobar si un equipo tiene ventaja al jugar como local, como suele ser común en el mundo del deporte. Creamos la tabla *result_distribution* con:

```
result_distribution <- table(filtered_results$match_result)
```

Tras su ejecución nos da los siguientes valores:

Tabla 4: Distribución de resultados

Away Win	Draw	Home Win
12371	10216	21651

Fuente: Elaboración propia

Como podemos comprobar, jugar como local da una cierta ventaja a la selección, siendo el número de victorias locales casi el doble que las visitantes. También podemos observar que el resultado menos probable es un empate, pero, ¿Ha sido siempre así? Para contestar a esta pregunta podemos obtener la distribución de resultados por décadas. Creamos una nueva tabla con:

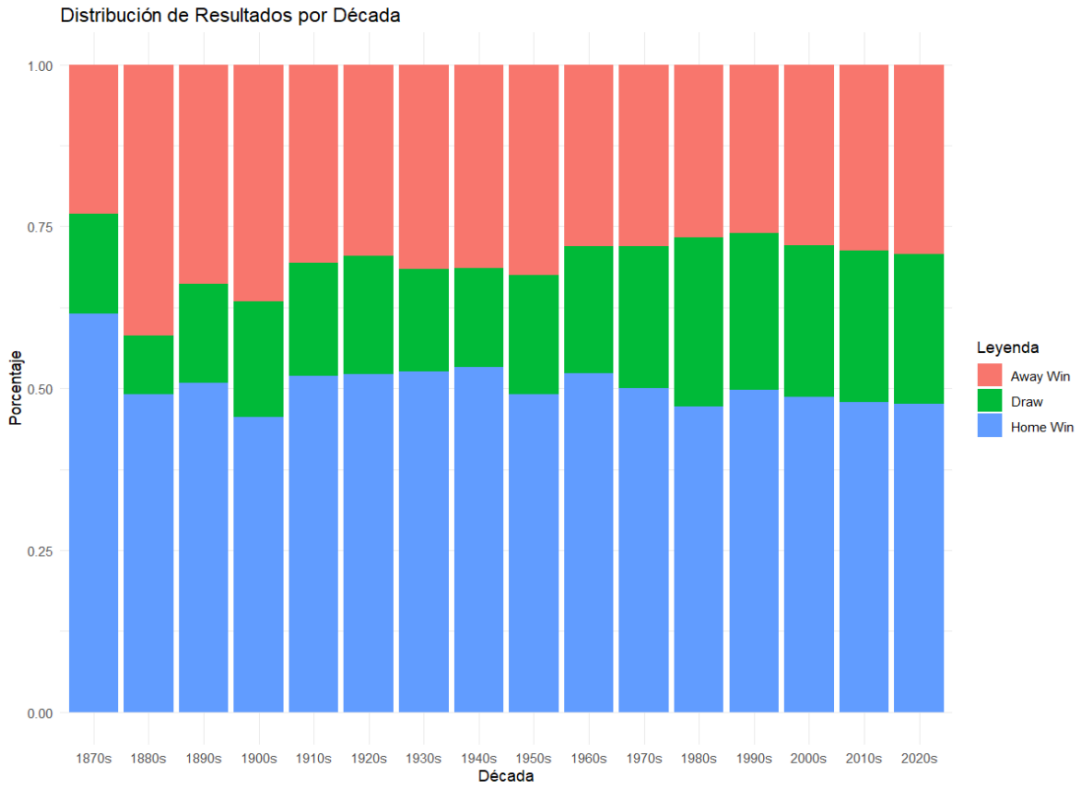
```
result_distribution_by_decade <- table(results$decade, results$match_result)
```

Esta tabla contiene más elementos, por lo que su comprensión es más difícil que la anterior, debido a esto vamos a crear un gráfico de líneas, utilizaremos las siguientes instrucciones:

```
ggplot(df_distribution, aes(x = Var1, y = Freq, fill = Var2)) +
  geom_bar(stat = "identity", position = "fill") +
  labs(title = "Distribución de Resultados por Década",  x = "Década",  y = "Porcentaje",
       fill = "Leyenda")
```

El gráfico resultante es el siguiente:

Gráfico 4: Distribución de resultados por década



Fuente: Elaboración propia

Como se puede observar, los resultados por décadas son más o menos similares, a excepción de las primeras décadas donde el número de observaciones es menor, por lo que la varianza de los datos también lo es. Las victorias locales suelen representar alrededor del 50% de los resultados, mientras que los empates siguen siendo el resultado más improbable durante todas las décadas. Las victorias visitantes fueron disminuyendo durante el siglo XX, sin embargo, con el comienzo de este siglo, estas comenzaron a comerle terreno al resto de resultados, sugiriendo que estos cada vez se igualan más y quizás el factor campo está perdiendo fuerza en el fútbol de selecciones.

Otro aspecto que considerar en nuestro análisis puede ser la competición en la que se disputa cada partido, vamos a observar la cantidad de partidos de cada competición en *filtered_results*. Para ello creamos una tabla de las competiciones que aparecen en el dataset, con la instrucción unique(filtered_results$tournament) que devuelve los valores únicos de las competiciones. A la hora de hacer el gráfico encontramos que hay demasiadas competiciones para que el gráfico sea legible. Si contamos el tamaño de la

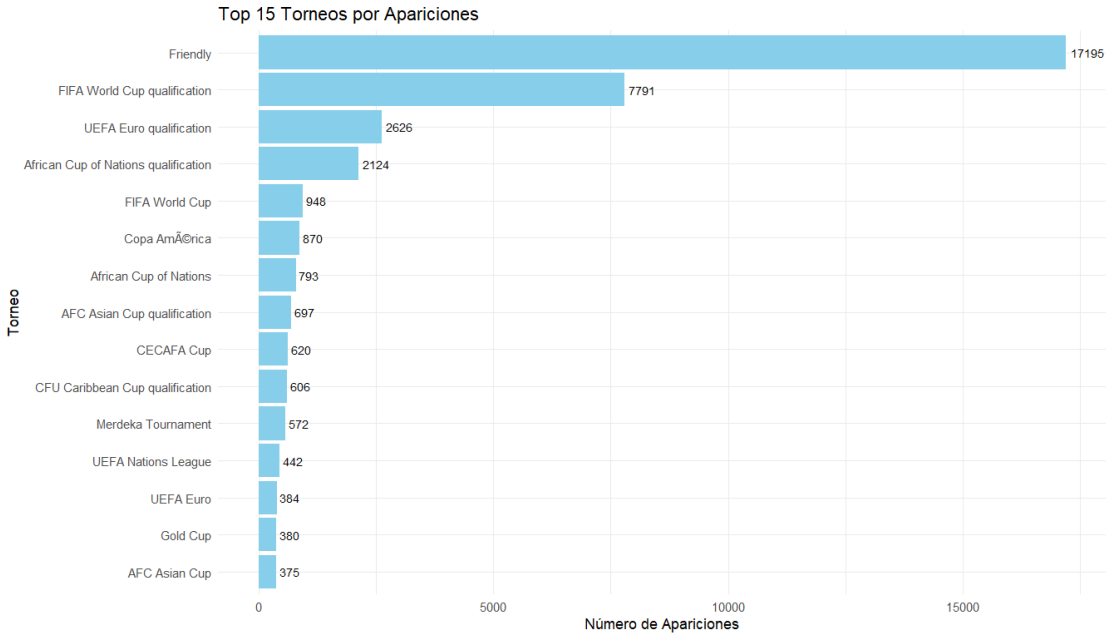
tabla con los torneos obtenemos que hay 156 competiciones distintas, por lo que vamos a hacer el gráfico únicamente con las 15 competiciones que más aparecen.

Para la creación de este gráfico utilizaremos las siguientes instrucciones:

```
top_15_competitions <- sort(all_competitions, decreasing = TRUE)[1:15]

top_15_df <- as.data.frame(top_15_competitions)

# Crear el gráfico de barras

ggplot(top_15_df, aes(x = reorder(Var1, Freq), y = Freq)) +

  geom_bar(stat = "identity", fill = "skyblue") +

  geom_text(aes(label = Freq), hjust = -0.15, size = 3) +

  coord_flip() +

  labs(x = "Torneo", y = "Número de Apariciones", title = "Top 15 Torneos por Apariciones")
```

A diferencia del primer gráfico de barras, usamos la línea coord_flip() cambiar las barras a horizontal para una mejor legibilidad de los nombres de las competiciones. Adicionalmente, usamos la línea geom_text() para añadir una etiqueta a cada barra con la frecuencia de cada competición. El gráfico obtenido es el siguiente:

Gráfico 5: Top 15 torneos por apariciones



Fuente: Elaboración propia

Como se puede observar en el gráfico, los amistosos internacionales son, con diferencia, la "competición" de los que más registros se tiene. Esto se debe a que este tipo de partidos se juegan entre selecciones de todo el planeta y se utilizan mucho para coger ritmo en las selecciones, ya que estas se tratan de equipos de jugadores que no suelen jugar juntos en el mismo equipo.

Adicionalmente, observamos que los partidos clasificatorios entre selecciones son, después de los amistosos, los más disputados. Encabezados por los clasificatorios para el Mundial, pues estos se disputan en todas las confederaciones, y seguidos por los de la Eurocopa, al tratarse la UEFA de la confederación con más selecciones nacionales. Respecto a torneos oficiales, observamos que los dos torneos de selecciones más antiguos que se siguen disputando hoy en día, la Copa América (1916) y la FIFA World Cup (Mundial, 1930), son las competiciones con más partidos disputados.

Observando este gráfico, se puede pensar que los partidos amistosos, donde muchas veces las selecciones no alinean a su equipo más fuerte, o donde la competición es menor, pueden afectar a los resultados de nuestros análisis, por lo que se va a crear un dataset prescindiendo de este tipo de partidos. Para la creación de este dataset utilizaremos la librería *dplyr,* que nos permite añadir filtros a un dataset. Al cargar esta librería, RStudio nos avisa que ciertas funciones de este paquete, como *filter*, que es la que vamos a utilizar, se encuentran cargadas con el mismo nombre por otras librerías, en este caso por el paquete *stats*, es por ello a la hora de utilizar esta función, se le especificará al programa de que librería la vamos a utilizar. Una vez aclarado esto usamos para crear el nuevo dataset la siguiente instrucción:

```
filtered_results_no_friendly <- dplyr::filter(filtered_results, tournament != "Friendly")
```

Observando el summary() de nuestro nuevo dataset, encontramos que los datos generales no cambian demasiado, sufriendo únicamente un ligero aumento a favor del equipo local en la media *goal_difference.* Resulta curioso observar que los resultados más abultados, tanto a favor de los locales como de los visitantes, siguen apareciendo en este dataset, por lo que estos partidos fueron disputados durante competiciones oficiales, y no en amistosos, donde hay menos en juego.

Otro aspecto interesante que podemos analizar para ver la evolución del fútbol de selecciones es el número de selecciones nacionales que disputaron su primer partido en cada década. Podemos, gracias a la variable *confederation,* segmentar el gráfico para observar la distribución de estas nuevas selecciones entre las distintas confederaciones. Veamos cómo podemos crear esta tabla y el gráfico:

Para la creación de la tabla, se instala la librería tidyr, ya que usaremos su función pivot_longer() para juntar las columnas *home_team* y *away_team.*

```
first_appearance <- filtered_results %>%
  select(date, decade, home_team, away_team, home_confederation, away_confederation) %>%
  pivot_longer(cols = c(home_team, away_team), names_to = "team_role", values_to = "team") %>%
  mutate(confederation = ifelse(team_role == "home_team", home_confederation, away_confederation)) %>%
  group_by(team) %>%
  summarize( first_match_date = min(date), first_decade = first(decade[date == min(date)]), confederation = first(confederation[date == min(date)]) ) %>%
  ungroup()
```

A continuación, contamos el número de nuevas selecciones agrupadas por década y confederación:

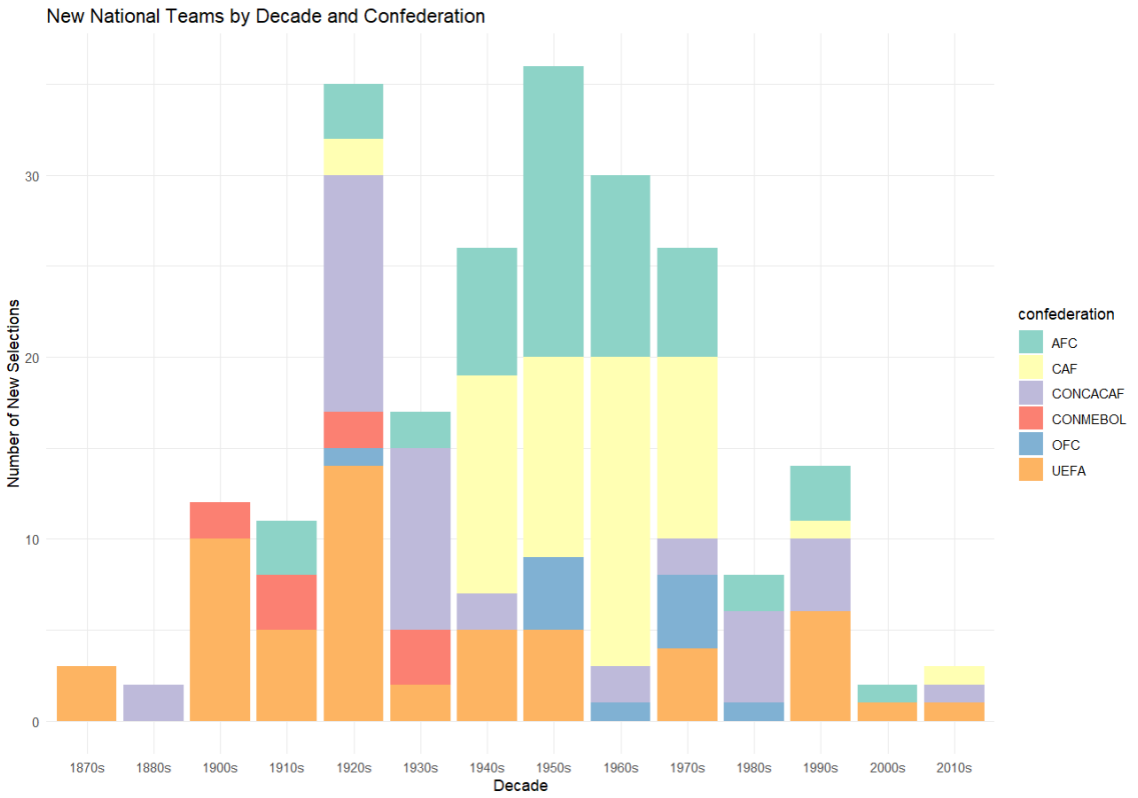
```
new_selections_by_decade <- first_appearance %>%
  group_by(first_decade, confederation) %>%
  summarise(num_teams = n()) %>%
  ungroup()
```

Y creamos el gráfico de barras apliadas con las siguientes instrucciones:

```
ggplot(new_selections_by_decade, aes(x = factor(first_decade), y = num_teams, fill = confederation)) +
  geom_bar(stat = "identity") +
  labs(x = "Decade", y = "Number of New Selections", title = "New National Teams by Decade and Confederation")
```

El gráfico resultante es el siguiente:

Gráfico 6: Nuevas selecciones nacionales por década y confederación



Fuente: Elaboración propia

Este gráfico resulta de gran utilidad a la hora de observar el desarrollo del fútbol de selecciones durante los años. Cabe recordar que el fútbol no se comenzó a dividir entre confederaciones de la FIFA hasta el 1953, pero este gráfico ayuda a visualizar el desarrollo de este deporte en los distintos continentes. Como se mencionó anteriormente, las primeras selecciones que disputan partidos entre sí son las de los países británicos en la década de los 70. Los países europeos son, en general, los que antes empiezan a desarrollar sus selecciones nacionales y no es hasta la década de los 20, cuando realmente el número de selecciones nacionales comienza a aumentar drásticamente, impulsado principalmente por la creación de numerosas selecciones centroamericanas. Podemos observar que Oceanía y África son los continentes donde este deporte evoluciona con menor ritmo, sufriendo este último un gran desarrollo durante las décadas a mitad de siglo, al igual que el futbol asiático. Observamos que, con la entrada del siglo XXI, el número de nuevas selecciones nacionales ha descendido drásticamente, esto se debe principalmente a que la gran mayoría de países del mundo ya cuentan con una selección nacional de este deporte, y solo se integran nuevas

selecciones con la formación de países nuevos, algo que no ocurre frecuentemente en la actualidad.

Vamos a crear ahora una nueva variable llamada *total_goals* que sume los goles locales y visitantes de cada partido de *filtered_results*. Con ella, podemos observar la evolución de la media de goles anotados por partido de cada década, para comprobar si antes el fútbol, al ser más caótico y menos ordenado, daba lugar a partidos con más goles o si, al perfeccionarse técnica y tácticamente con el paso de los años esto ha llevado a partidos más entretenidos.

En primer lugar, creamos la columna *total_goals* con:

```
filtered_results <- filtered_results %>%
  mutate(total_goals = home_score + away_score)
```

Acto seguido calculamos la media de goles totales por partido de cada década:

```
average_goals_by_decade <- filtered_results %>%
  group_by(decade) %>%
  summarize(avg_goals = mean(total_goals))
```

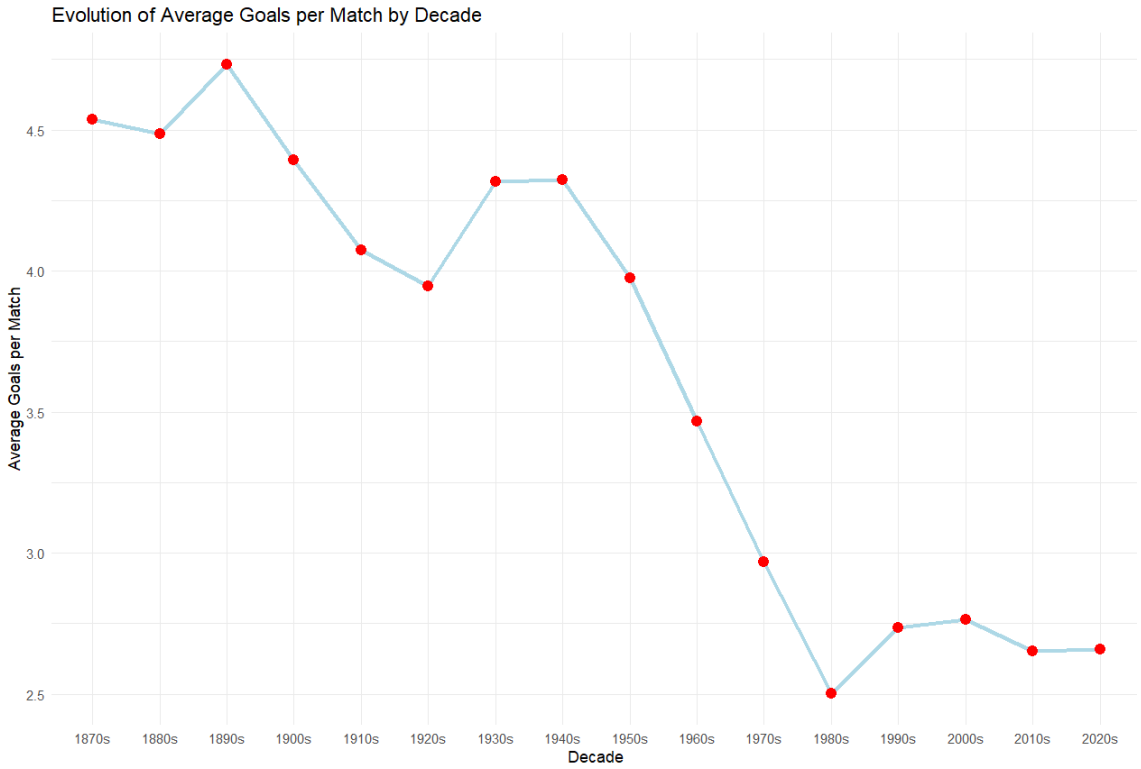
Y por último creamos el gráfico de puntos unidos por líneas para visualizar su evolución:

```
ggplot(average_goals_by_decade, aes(x = decade, y = avg_goals, group = 1)) +
  geom_line(color = "lightblue", linewidth = 1.2) +
  geom_point(color = "red", size = 3) +
  labs(x = "Decade", y = "Average Goals per Match", title = "Evolution of Average Goals per Match by Decade")
```

En el gráfico resultante (Ver Gráfico 6), se puede observar cómo, efectivamente, la media de goles por partido ha ido disminuyendo con el paso de los años. Esta caída en el promedio de goles anotados se fue acentuando con el paso del siglo XX, alcanzando su mínimo histórico en la década de los 80, cuando el fútbol italiano, famoso por su solidez defensiva, era el referente europeo y mundial a nivel de tanto de clubes como de selección, ganando el mundial de 1982. A partir de los años 90, la cifra de goles promedio por partido se ha estabilizado, aunque en valores mucho más bajos que los de la primera mitad del siglo XX. Esto no quiere decir que el fútbol se haya vuelto menos

entretenido, quizás para aquellos espectadores que solo quieren ver goles, si no que la evolución táctica de los equipos los ha vuelto menos caóticos y más sólidos.

Gráfico 7: Evolución de la media de goles por partido en cada década



Fuente: Elaboración propia

De momento, hemos realizado análisis generales sobre el dataset de resultados, así como el dataset ajustado tras eliminar selecciones regionales y no oficiales. Vamos a analizar rápidamente los otros dos datasets, *shootouts* y *goalscorers*.

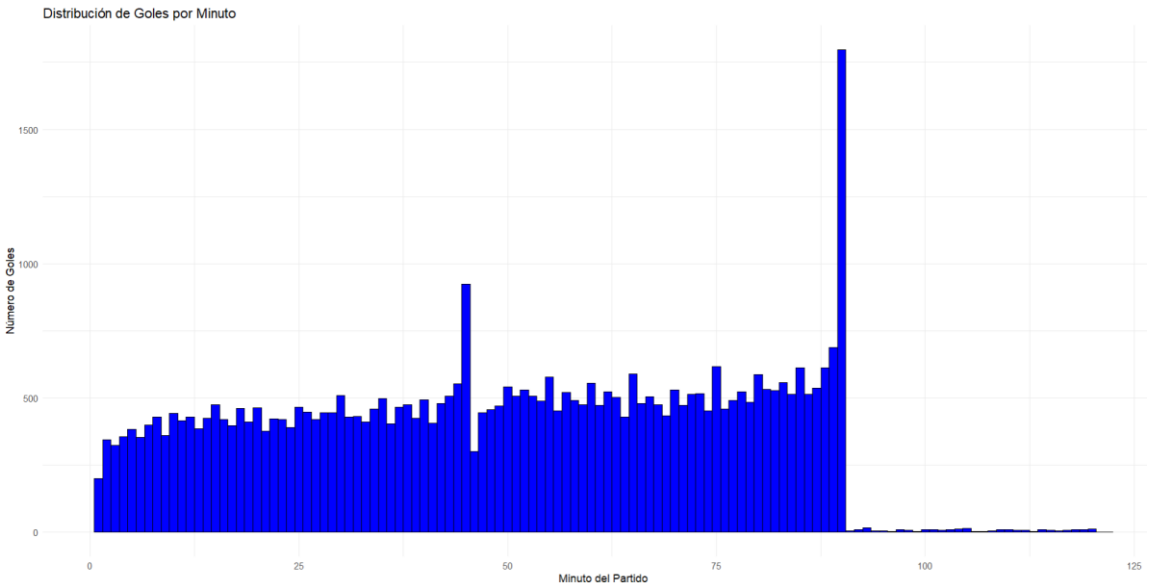
Empecemos por *shootouts*, que registra las tandas de penaltis a lo largo de la historia de los partidos de selecciones nacionales. Muchas veces se afirma, que las tandas de penaltis son una lotería, sin embargo, se ha demostrado que se pueden tomar decisiones para influir en ellas. Desde apuntarle al portero en una botella los lanzadores de penaltis rivales y sus lados preferidos, como hizo el portero de la selección inglesa este verano durante los octavos de final de la Eurocopa en la tanda de penaltis que enfrentaba a su selección contra Suiza, a entrenamientos específicos para lanzadores y porteros o las distracciones de los porteros momentos antes de la ejecución de la pena máxima, que hicieron famoso al portero de la selección de Argentina durante la consecución de su último mundial, hoy en día hay muchas maneras de influir en ellas.

Una de las mayores preguntas a la hora de elegir en una tanda de penaltis es: ¿Debería tirar mi selección primero o dejarles a los otros empezar? Antes de comenzar la tanda, se elige por sorteo, normalmente a cara o cruz entre los capitanes de cada equipo, lanzando la moneda el árbitro del encuentro, tanto el orden de lanzar como la portería donde se realizará la tanda de penaltis.

Desafortunadamente, no se puede estudiar la segunda variable, puesto que no se tienen registros sobre ella, a pesar de ser indudablemente una variable significativa si el estadio donde se disputa el partido es neutral, ya que las aficiones de los equipos suelen ocupar cada una un fondo de las gradas. Vamos a estudiar rápidamente la primera variable, que se encuentra en la columna *first_shooter* de esta tabla. Como se mencionó durante la etapa de preprocesado de los datos, esta columna contiene un elevado número de valores nulos, de hecho, casi 2/3 de las tandas registradas no cuentan con datos en esta columna, por lo que este estudio no se puede asegurar la efectividad de este estudio. De todas formas, se observa que de aquellas tandas de penaltis que contienen un valor en *first_shooter*, un 54% de las veces el equipo que comenzó tirando la tanda de penaltis resultó vencedor en esta, por lo que se podría intuir que algo de ventaja puede dar, aunque los resultados son inconcluyentes.

Vamos a realizar ahora un análisis general de la tabla *goalscorers,* en primer lugar, vamos a ver la distribución de goles durante los minutos del partido, creando un histograma:

Gráfico 8: Distribución de goles por minuto



Fuente: Elaboración propia

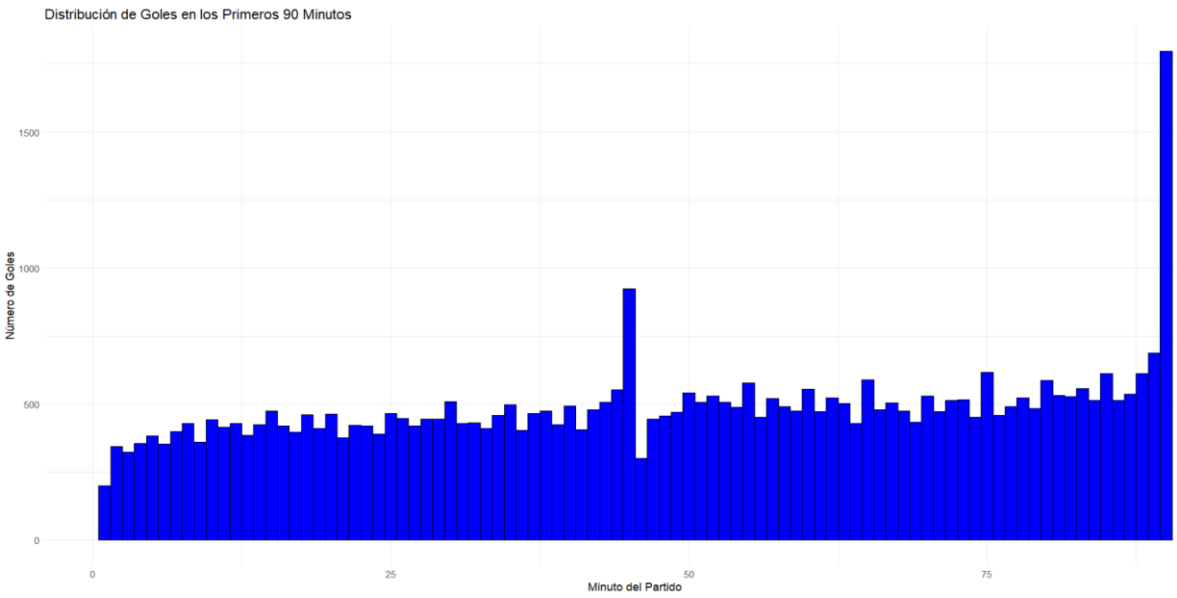
Este primer gráfico resulta algo confuso, a primera vista se puede observar que contiene datos de goles durante la prórroga, lo que dificulta su visión. Adicionalmente, observamos que los últimos minutos de cada parte, el minuto 45y el 90, contienen los goles anotados durante esos minutos y el tiempo añadido de cada mitad, por lo que tienen una mayor frecuencia, al contar con más tiempo disponible.

En primer lugar, vamos a separar los gráficos entre los goles durante los 90 minutos de un partido normal y los goles en la prórroga, esto lo haremos creando una columna llamada *period,* que indique si el gol es anotado en un minuto del tiempo regular o en uno de la prórroga:

```
goalscorers$period <- ifelse(goalscorers$minute <= 90, "Primeros 90 minutos", "Prórroga")

# Filtrar goles en los primeros 90 minutos

goals_first_90 <- goalscorers %>% filter(period == "Primeros 90 minutos")

# Filtrar goles en la prórroga

goals_extra_time <- goalscorers %>% filter(period == "Prórroga")
```

Una vez separados, volvemos a hacer los dos histogramas y obtenemos:

Gráfico 9: Distribución de goles en los primeros 90 minutos

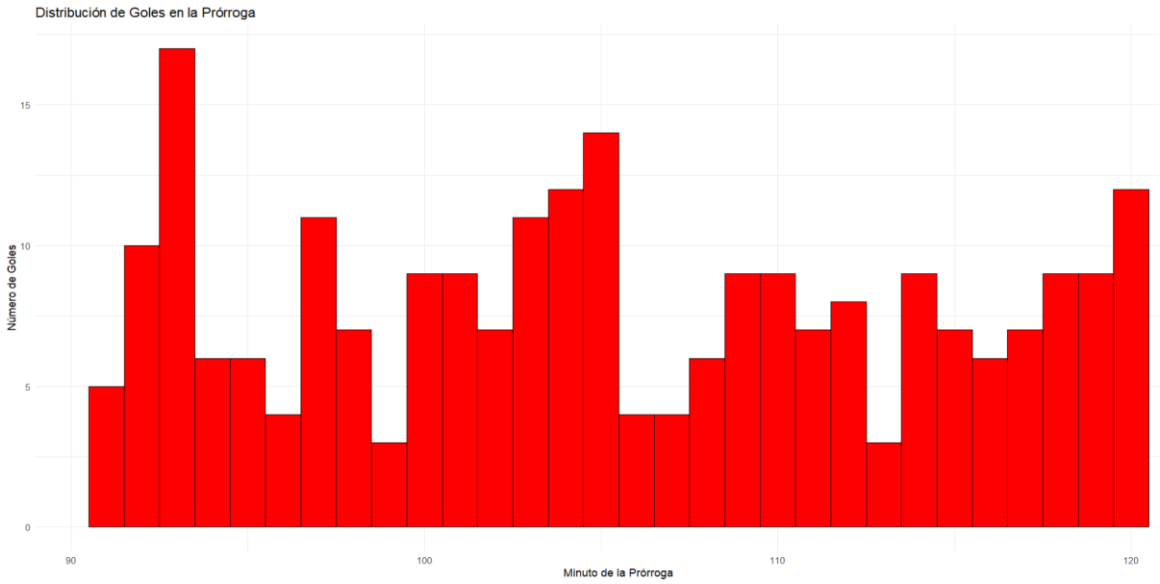


Fuente: Elaboración propia

En este gráfico podemos observar cómo, efectivamente, el último minuto de cada parte obtiene más frecuencia, por los motivos comentados anteriormente. Además, podemos observar cómo los primeros tienen frecuencias algo menores, especialmente los primeros minutos del partido, mientras que los últimos minutos de cada parte tienen una tendencia algo más alta, especialmente los de la segunda parte, donde se deciden muchos de los partidos y los equipos están más cansados, propiciando más errores defensivos.

Por otra parte, vamos a analizar los resultados de la distribución de goles en los minutos de la prórroga con el siguiente histograma:

Gráfico 10: Distribución de goles por minutos en las prórrogas



Fuente: Elaboración propia

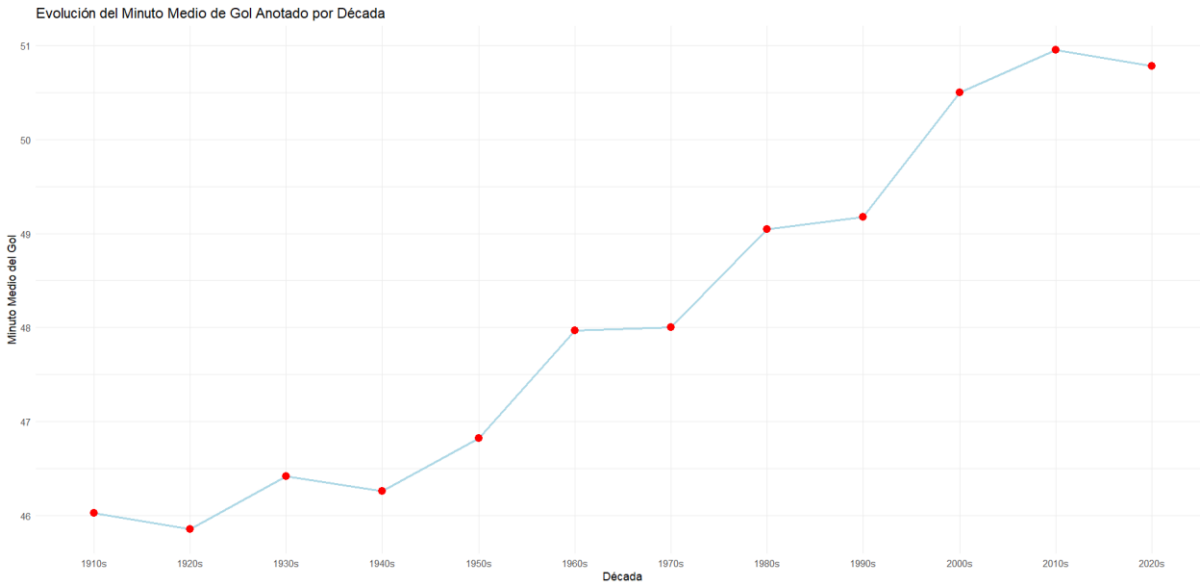
En este gráfico llama la atención el tercer minuto de la primer mitad de la prórroga, siendo el minuto en el que más goles se marcan. Se observan también un mayor número de goles en las primeras partes de las prórrogas que en las segundas, y se sigue la tendencia creciente del tiempo regular en los últimos minutos de cada mitad.

Puede resultar interesante también observar la evolución con el paso de las décadas del minuto medio del gol anotado, para ello utilizaremos solo los goles anotados en el tiempo regular, almacenados en la *tabla goals_first_90* descrita antes.

```
# Calcular el minuto medio del gol por década

average_minute_by_decade_90 <- goals_first_90 %>%

  group_by(decade) %>% summarise(mean_minute = mean(minute, na.rm = TRUE))
```

El gráfico resultante es el siguiente:

Gráfico 11: Evolución del minuto medio del gol anotado por década



Fuente: Elaboración propia

Resulta interesante observar esta evolución, que presenta una tendencia creciente con el paso de los años. Se puede afirmar que los goles en los partidos internacionales se marcan, de media, cada vez más tarde.

Hasta el momento, se ha realizado un análisis general sobre las tendencias del fútbol internacional de selecciones, sin embargo, el análisis de datos nos permite profundizar mucho más allá de tendencias generales. Esta disciplina de la informática nos permite extraer conclusiones específicas sobre la competición o la selección que sea de nuestro interés.

Vamos a empezar a profundizar en las selecciones nacionales observando cuales son, en general, las mejores selecciones tanto a nivel ofensivo como defensivo. Esto se puede observar en el número promedio de goles anotados y encajados por cada selección.

En primer lugar, vamos a calcular la media de goles anotados por cada equipo, tanto cuando juega de local, como de visitante:

```
mean_goals_scored <- filtered_results %>% mutate(team = home_team, goals_scored = home_score) %>%

select(team, goals_scored) %>%
```

```
bind_rows(filtered_results %>%

mutate(team = away_team, goals_scored = away_score) %>%

select(team, goals_scored)) %>%

group_by(team) %>%

summarise(mean_goals = mean(goals_scored))
```

De la misma manera podemos calcular la media de goles concedidos por cada equipo.

Una vez obtenidos, combinamos estos resultados en una misma tabla que llamaremos *team_stats:*

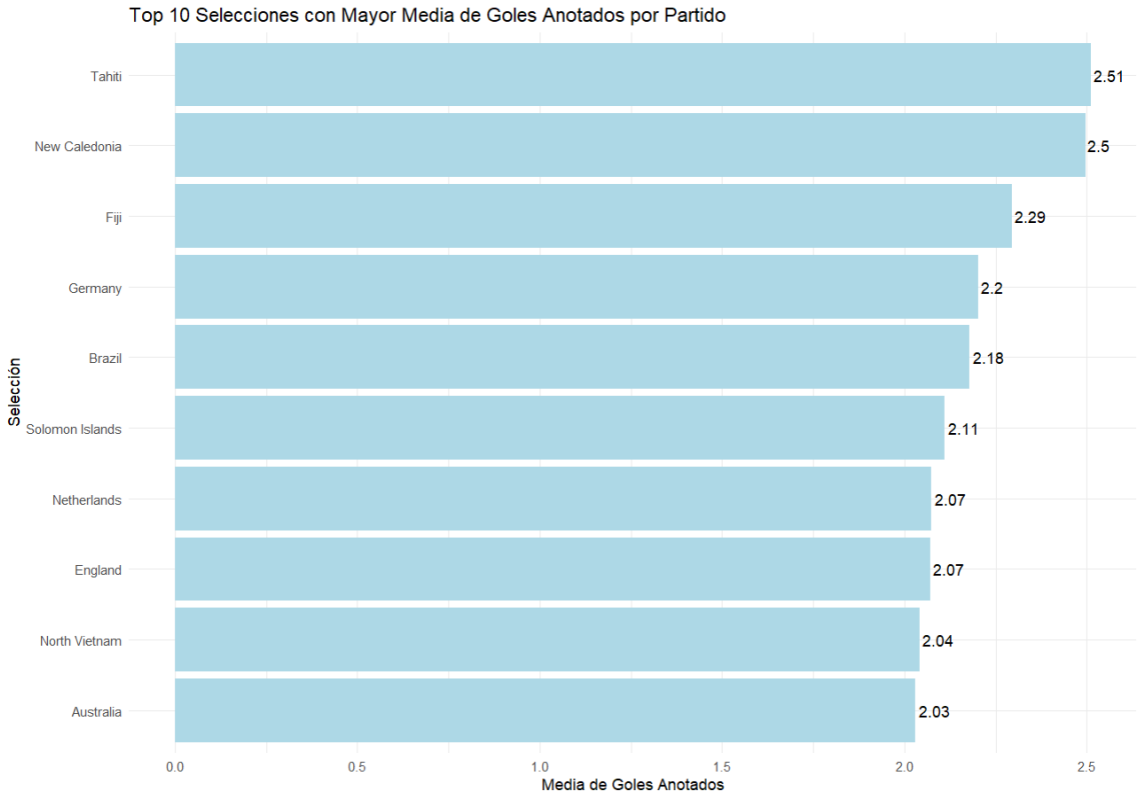
```
team_stats <- mean_goals_scored %>%

  inner_join(mean_goals_conceded, by = "team")
```

Y obtenemos el número de resultados que queremos para nuestra tabla, en este caso vamos a obtener el top 10, tanto para el rendimiento ofensivo, como el defensivo.

```
# Top 10 selecciones que más goles anotaron por partido

top10_scored <- team_stats %>%

  arrange(desc(mean_goals)) %>%

  top_n(10, mean_goals)

# Top 10 selecciones que menos goles concedieron por partido

top10_conceded <- team_stats %>%

  arrange(mean_goals_conceded) %>%

  top_n(10, -mean_goals_conceded)
```

Una vez ya tenemos nuestros resultados filtrados y ordenados, podemos crear el gráfico que ayude con la visualización de estos. Estos gráficos se crean de una manera similar a los creados anteriormente, gracias a la librería ggplot2. Vamos a observar los resultados:

Gráfico 12: Top 10 selecciones con mejor promedio goleador



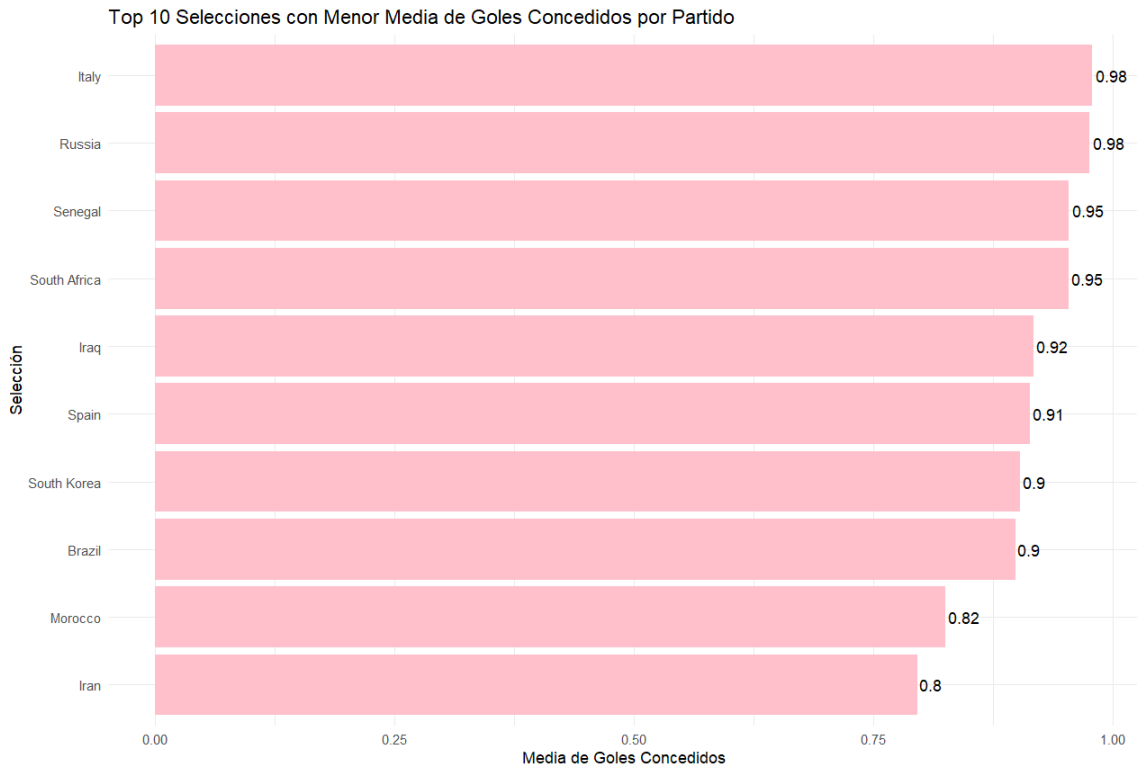
Fuente: Elaboración propia

En primer lugar, tenemos el gráfico a nivel ofensivo, resulta curioso encontrar en lo más alto a selecciones como Tahití o Nueva Caledonia, selecciones, a priori, de menor nivel. Observamos que las tres primeras selecciones pertenecen a la confederación de Oceanía, lo que puede indicar que esta sea una confederación donde los partidos disputados acaban con muchos goles. Son más reconocibles en este gráfico selecciones de mayor nivel a lo largo de la historia, como Alemania, Brasil o Inglaterra.

Si observamos el gráfico a nivel defensivo (Ver gráfico 12), encontramos aquí a selecciones reconocibles a nivel defensivo como Italia, que lidera el ranking. Resulta llamativa la aparición de tres selecciones africanas en este top, ya que, si analizamos conjuntamente ambos gráficos observamos que ninguna selección africana destaca a nivel ofensivo, lo que puede sugerir que los partidos de la confederación africana cuentan con menos goles. El caso contrario de lo que pasaba con Oceanía, donde no encontramos ninguna selección en el ranking del nivel defensivo.

Para apoyar estas hipótesis, vamos a calcular a continuación los goles medios anotados y encajados por cada confederación. La obtención de estos datos es muy similar a nivel de código que los anteriores, por lo que este no se incluye en la memoria.

Gráfico 13: Top 10 selecciones con mejor rendimiento defensivo



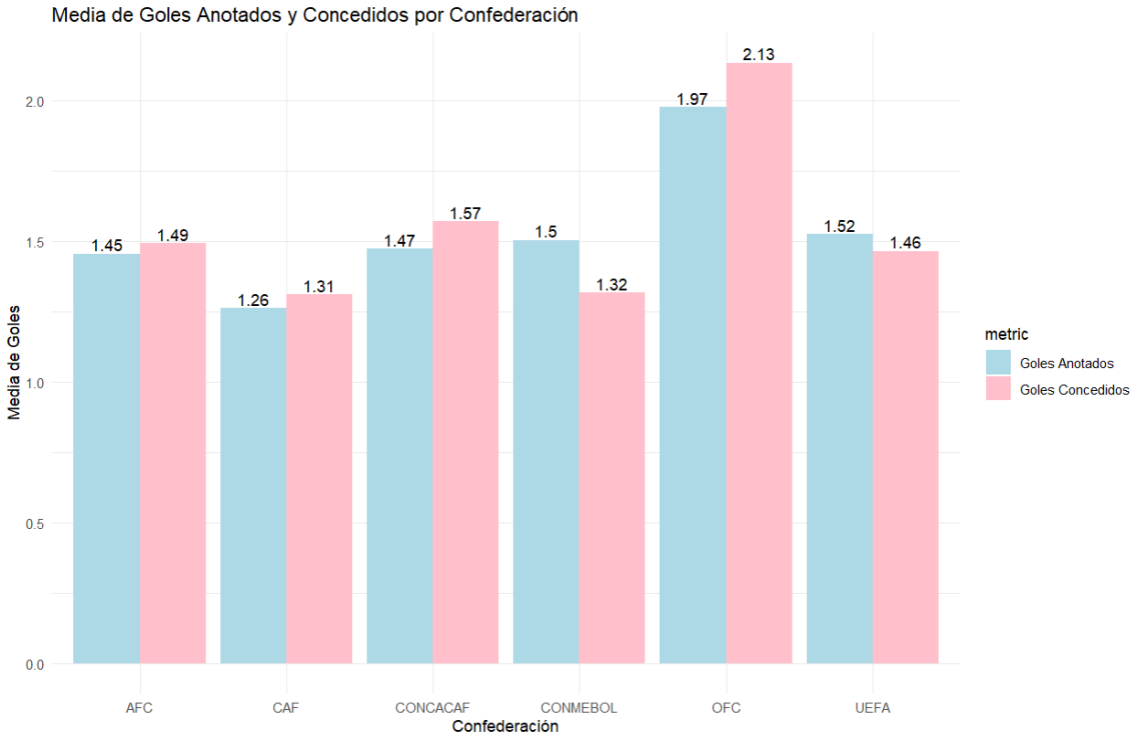
Fuente: Elaboración propia

Una vez obtenidos los datos del promedio de goles anotados y encajados por cada confederación, vamos a crear un gráfico de barras agrupadas para su visualización:

```
ggplot(confederation_stats_long, aes(x = confederation, y = mean_goals, fill = metric))
+
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = round(mean_goals, 2)),
            position = position_dodge(width = 0.9), vjust = -0.25) +
  labs(title = "Media de Goles Anotados y Concedidos por Confederación",
       x = "Confederación", y = "Media de Goles") +
  scale_fill_manual(values = c("mean_goals" = "blue", "mean_goals_conceded" = "red"),
                    labels = c("Goles Anotados", "Goles Concedidos"))
```

El gráfico resultante es el siguiente:

Gráfico 14: Media de goles anotados y concedidos por confederación



Fuente: Elaboración propia

Como se puede comprobar en este gráfico, efectivamente la confederación de Oceanía cuenta con una media, tanto de goles anotados como encajados, considerablemente superior al resto de confederaciones, por lo que los partidos disputados por selecciones de esta región cuentan con más goles que el resto. Se observa también que el promedio de goles en partidos donde participan equipos de la confederación africana es el más bajo de todas las confederaciones, apoyando la hipótesis de que estos partidos son con los que menos goles cuentan en promedio.

Por último, observamos en este gráfico que las únicas confederaciones con un promedio positivo de goles son la UEFA y la CONMEBOL, que han sido históricamente las confederaciones que han contado con las selecciones de mayor nivel, sugiriendo un mejor rendimiento de estas confederaciones frente al resto.

Para respaldar esta hipótesis, vamos a crear un gráfico que nos permita visualizar los rendimientos entre confederaciones. Para poder llevar a cabo esto, vamos a utilizar las columnas de *match_result*, *home_confederation* y *away_confederation* creadas anteriormente en la tabla *filtered_results.*

En primer lugar, deberemos asignar los resultados a cada confederación en una nueva tabla, filtrando los empates.

```
confederation_results <- filtered_results %>%
  mutate(winner_confederation = case_when(
    match_result == "Home Win" ~ home_confederation,
    match_result == "Away Win" ~ away_confederation,
    TRUE ~ NA_character_
  ))
confederation_wins <- confederation_results %>%
  filter(!is.na(winner_confederation))
```

Acto seguido calcularemos el número de victorias entre confederaciones y de esto extraeremos el porcentaje de victorias entre cada una:

```
confederation_vs_confederation <- confederation_wins %>%
  mutate(loser_confederation = case_when(
    match_result == "Home Win" ~ away_confederation,
    match_result == "Away Win" ~ home_confederation
  )) %>%
  group_by(winner_confederation, loser_confederation) %>%
  summarise(wins = n()) %>%
  ungroup()


total_matches <- filtered_results %>%
  filter(home_confederation != away_confederation) %>%
  mutate(home_vs_away = paste(home_confederation, away_confederation, sep = " vs
"),
```

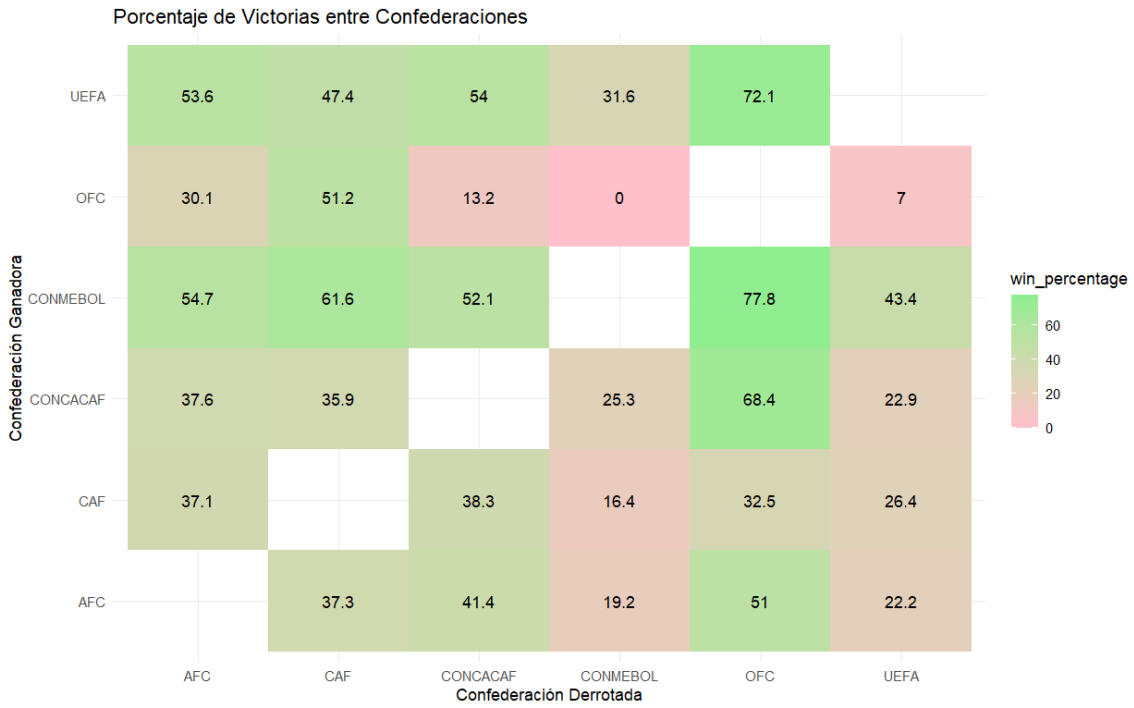
```
    away_vs_home = paste(away_confederation, home_confederation, sep = " vs ")) %>%
  pivot_longer(cols = c(home_vs_away, away_vs_home), names_to = "matchup_type", values_to = "matchup") %>%
  group_by(matchup) %>%
  summarise(total_matches = n()) %>%
  ungroup()

confederation_win_percentage <- confederation_vs_confederation %>%
  mutate(matchup = paste(winner_confederation, loser_confederation, sep = " vs ")) %>%
  inner_join(total_matches, by = "matchup") %>%
  mutate(win_percentage = (wins / total_matches) * 100)
```

Por último, vamos a crear un gráfico que nos ayude a visualizar los resultados, para ello elegiremos la opción de un mapa de calor, que nos permite observar rápidamente el porcentaje de victorias entre confederaciones.

```
ggplot(confederation_win_percentage, aes(x = loser_confederation, y = winner_confederation, fill = win_percentage)) +
  geom_tile() +
  geom_text(aes(label = round(win_percentage, 1))) +
  scale_fill_gradient(low = "pink", high = "lightgreen") +
  labs(title = "Porcentaje de Victorias entre Confederaciones",
       x = "Confederación Derrotada", y = "Confederación Ganadora")
```

El gráfico resultante es el siguiente:

Gráfico 15: Porcentaje de victorias entre confederaciones



Fuente: Elaboración propia

Como se puede observar, la elección de un mapa de calor es la más adecuada para esta visualización. En el eje vertical encontramos las confederaciones ganadoras sobre cada confederación de su eje horizontal. La diagonal que se encuentra vacía representa los cruces de cada confederación consigo misma, por lo que se deja vacía. Adicionalmente, el porcentaje faltante entre cada confederación representan los empates entre estas.

Llama mucho la atención el valor 0 en las victorias de la OFC sobre CONMEBOL, esto no se trata de un error, por increíble que parezca, nunca una selección absoluta de un país de la OFC ha vencido a una selección de la CONMEBOL, ni en partido oficial ni en amistosos.

Como era de esperar, este gráfico demuestra que la UEFA y la CONMEBOL son las dos confederaciones más fuertes históricamente. La CONMEBOL obtiene un resultado positivo contra todas las confederaciones mientras que la UEFA solo pierde contra los sudamericanos.

La OFC representa la confederación con peores resultados, aunque destaca su sorprendente buen papel ante selecciones africanas. Esta se ve lastrada por sus

pésimos resultados ante selecciones de las dos confederaciones más potentes, así como ante selecciones de la CONCACAF.

La CAF por otra parte no obtiene buenos resultados ante ninguna confederación, pero estos malos resultados no son tan drásticos como los de la confederación de Oceanía.

Los enfrentamientos entre selecciones resultan muy útiles para indagar en estos resultados, pero su visualización es más complicada al tener un número tan grande de selecciones, que resultaría en un gráfico demasiado grande e ilegible.

Para poder obtener los resultados de enfrentamientos directos (comúnmente conocidos como “head to head”) entre dos selecciones vamos a crear una función en RStudio que nos devuelva los valores de las victorias de cada equipo, los empates y los goles anotados por cada equipo, así como el total de partidos jugados entre ellos. Entre los argumentos de entrada de la función vamos a poder filtrar entre el dataset de resultados que queremos utilizar, los equipos que queremos seleccionar y la competición que queremos observar, pudiendo dejar este último argumento vacío en el caso de querer observar el “head to head” general entre dos equipos.

El código de la función se incluye en el Anexo 2 del trabajo debido a su gran longitud.

Esta función puede resultar de gran utilidad a la hora de comprar enfrentamientos entre dos equipos. La capacidad de poder incluir filtros como el dataset a utilizar o la competición específica que queremos observar puede resultar de gran ayuda para facilitar el estudio del rendimiento entre selecciones nacionales. Por ejemplo, si buscamos resultados exclusivos de competiciones oficiales, podemos utilizar el dataset que preparamos *filtered_results_no_friendly* que no contiene los partidos amistosos. Por otro lado, si quisiéramos observar el rendimiento entre selecciones no oficiales podríamos usar el dataset original *results* para realizar estas observaciones.

Para poner un ejemplo de la correcta ejecución de esta función vamos a observar los enfrentamientos entre la selección española y la selección inglesa en varios escenarios, enfrentamiento que se disputó por última vez en la final de la Eurocopa de este verano y que se estudiará más en profundidad en el siguiente punto del trabajo.

Tabla 5: Head to head entre la selección española y la inglesa

```
> h2h_summary <- head_to_head(filtered_results, "Spain", "England")
> print(h2h_summary)
  team1   team2 team1_wins team2_wins draws team1_goals team2_goals total_matches
1 Spain England         11         13     4          34          46            28
> h2h_summary <- head_to_head(filtered_results_no_friendly, "Spain", "England")
> print(h2h_summary)
  team1   team2 team1_wins team2_wins draws team1_goals team2_goals total_matches
1 Spain England          3          4     2           9          10             9
> h2h_summary <- head_to_head(filtered_results, "Spain", "England", "UEFA Euro")
> print(h2h_summary)
  team1   team2 team1_wins team2_wins draws team1_goals team2_goals total_matches
1 Spain England          1          1     1           3           3             3
```

Fuente: Elaboración propia

En la primera ejecución encontramos los resultados de los enfrentamientos directos entre ambas selecciones en el dataste *filtered_results.* Sin embargo, podría resultarnos útil observar estos resultados excluyendo partidos amistosos, donde la competitividad es menor y muchas veces los seleccionadores utilizan estos partidos para probar tácticas o jugadores nuevos. En este caso, los resultados no varían demasiado, al tratarse de un enfrentamiento bastante parejo, con una ligera inclinación hacia el lado de los ingleses, pero en otros casos podría influir. En la última ejecución de esta función añadimos el filtro por competición, centrándonos en la Eurocopa, puesto que es la competición que se estudiará más a fondo durante la realización de este trabajo. Si observamos los resultados esta vez, encontramos que la igualdad en este torneo es máxima entre estas dos selecciones, habiéndose enfrentado tres veces, con una victoria para cada una y un empate, e incluso con los mismos goles anotados.

Para terminar con este apartado del trabajo, vamos a indagar un poco más en las tendencias del torneo anteriormente mencionado, la Eurocopa. Sacar conclusiones interesantes de este análisis puede complementar el análisis más profundo que se realizará en el siguiente punto del trabajo sobre la última edición de este torneo.

Para este último análisis lo primero que haremos será crear un dataset que incluya únicamente los resultados de los partidos disputados en esta competición, llamaremos a este *euro_results*. Adicionalmente, será interesante crear versiones de *shootouts* y *goalscorers* que incluyan solo las instancias de partidos de la Eurocopa, aunque esto resulta algo más complicado al no tener estos datasets una columna que almacene el torneo donde se disputaba el partido en el que ocurrieron cada una de sus entradas. Para solucionar esto, filtraremos los goles y las tandas de penaltis en base a la coincidencia de los equipos que actúan como local y visitante sumado a la fecha en la que se disputa el partido. Las instrucciones para este filtrado son las siguientes:

```
euro_shootouts <- shootouts %>%
```

```
  filter(date %in% euro_results$date & home_team %in% euro_results$home_team &
           away_team %in% euro_results$away_team)
euro_goalscorers <- goalscorers %>%
  filter(date %in% euro_results$date & home_team %in% euro_results$home_team &
           away_team %in% euro_results$away_team)
```

En este caso, resulta interesante añadir una columna que llamaremos *edición*, con el año en que se disputó cada torneo. Adicionalmente, se convertirá esta columna de tipo numérico a factor, para facilitar la segmentación de los datos a la hora de analizarlos.

```
euro_results$edicion <- as.factor(euro_results$edicion)
```

Una vez contamos con estos tres datasets, que únicamente tienen datos sobre las distintas ediciones de las Eurocopas, podemos empezar el análisis de este torneo.

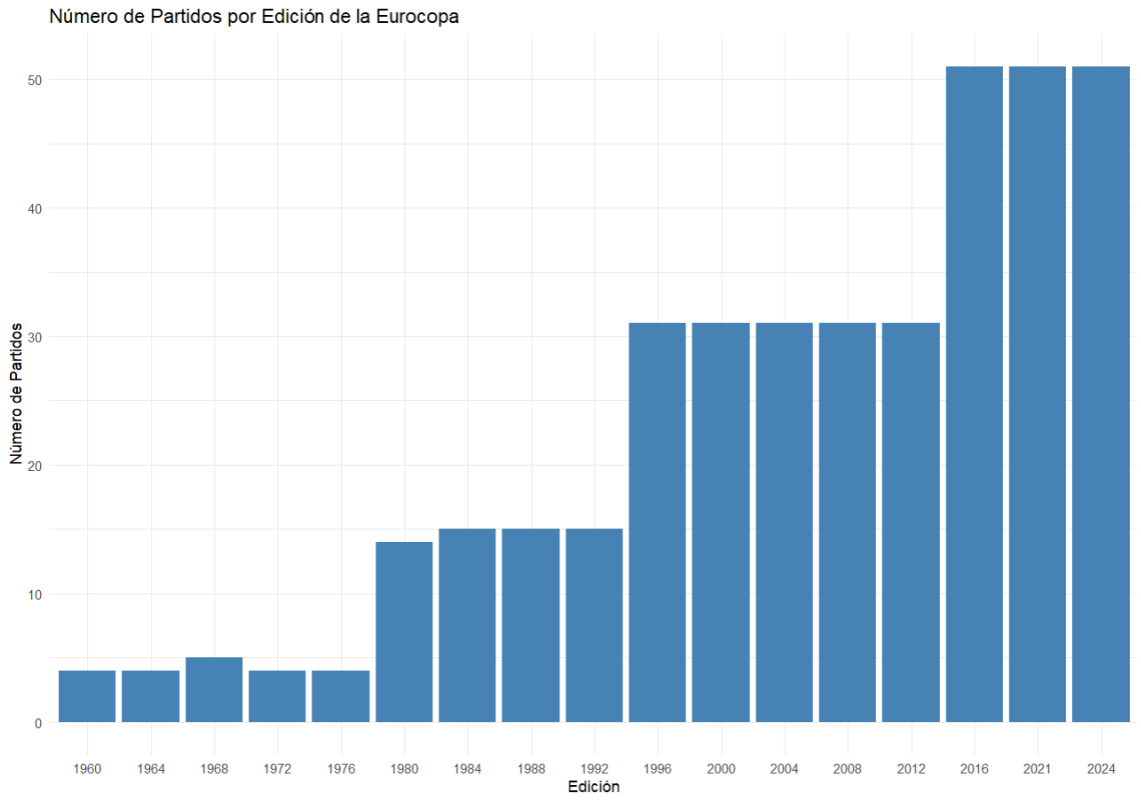
Con el summary() de estos datasets encontramos ya algunos datos interesantes. Se han disputado un total de 338 partidos de esta competición, en 17 ediciones celebradas de este torneo desde su fundación en 1960. Durante estos encuentros, se han anotado 948 goles y tan solo 25 partidos se han decantado con una tanda de penaltis.

La mayor diferencia de goles entre dos equipos ha sido de cinco goles, esto ha ocurrido en cinco ocasiones diferentes, incluyendo un Holanda – Serbia disputado en la edición del año 2000, cuando Holanda anotó seis goles, el mayor número de goles anotado por una selección en esta competición.

De una manera similar a lo aplicado con *filtered_results*, podemos obtener estadísticas generales del torneo como el número de partidos disputados o el número de goles promedio anotado en cada edición.

Utilizaremos gráficos para analizar estos datos:

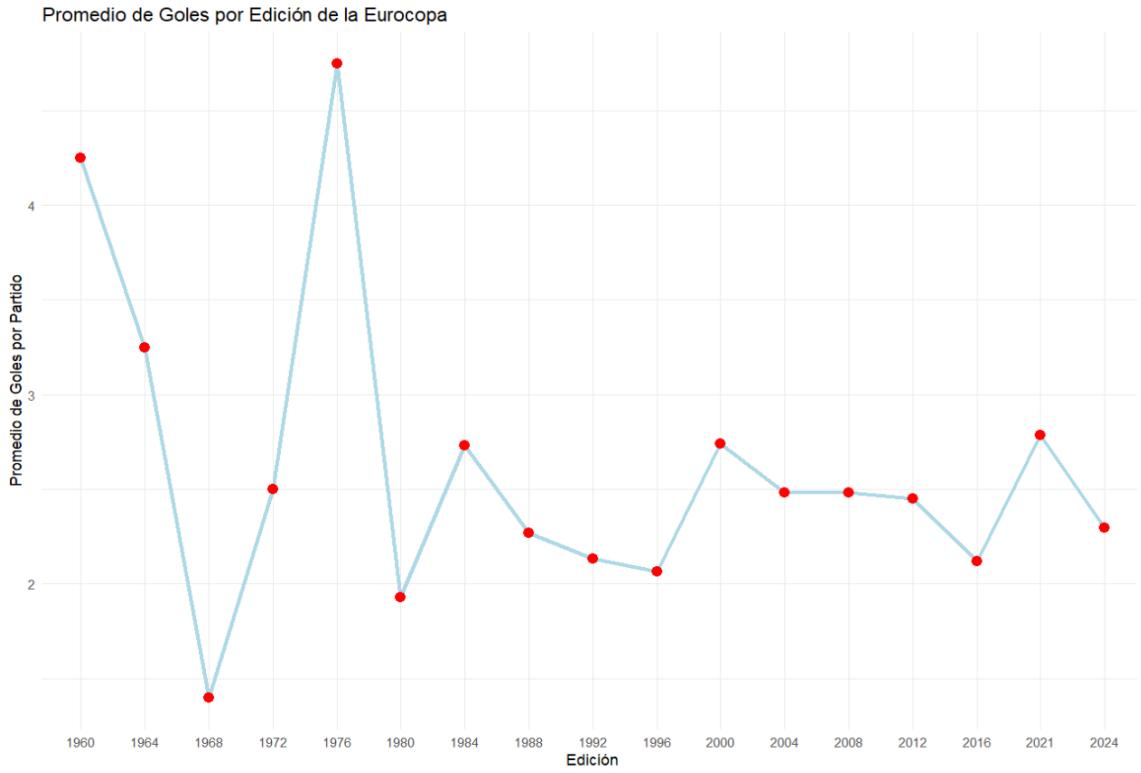
Gráfico 16: Numero de partidos disputados en cada edición de la Eurocopa



Fuente: Elaboración propia

Como podemos observar en este gráfico, a medida que ha ido cambiando el formato del torneo se han ido incrementando los partidos disputados en cada edición. En las primeras ediciones del torneo sólo se disputaban las rondas de semifinales, así como la final y el partido por el tercer puesto, el resto de los partidos se consideraban clasificatorios. En las primeras ediciones tampoco se habían introducido las tandas de penaltis, por lo que se jugaba un partido extra en caso de empate tras la prórroga (como la final de la Eurocopa de 1968 entre Italia y Yugoslavia). No fue hasta la edición de 1996 donde encontramos un formato similar al actual, con fase de grupos y clasificación a cuartos de final. En el año 2016, se cambió de nuevo el formato para dar entrada a más equipos en la competición y una ronda extra de eliminación. Esto se suma a la tendencia creciente en el mundo del fútbol de disputar cada vez más encuentros y que está empezando a generar descontento entre los jugadores profesionales, que cada vez sufren más físicamente.

Gráfico 17: Promedio de goles anotados por edición de la Eurocopa



Fuente: Elaboración propia

En este gráfico, que muestra el promedio de goles anotados en cada edición de la Eurocopa, observamos como los pocos partidos disputados durante las primeras ediciones causan una gran varianza entre los resultados y a medida que se cambia el formato, este valor se estabiliza. Llama la atención que la tendencia en este valor no disminuye, como sí lo hacía en los resultados generales observados anteriormente.

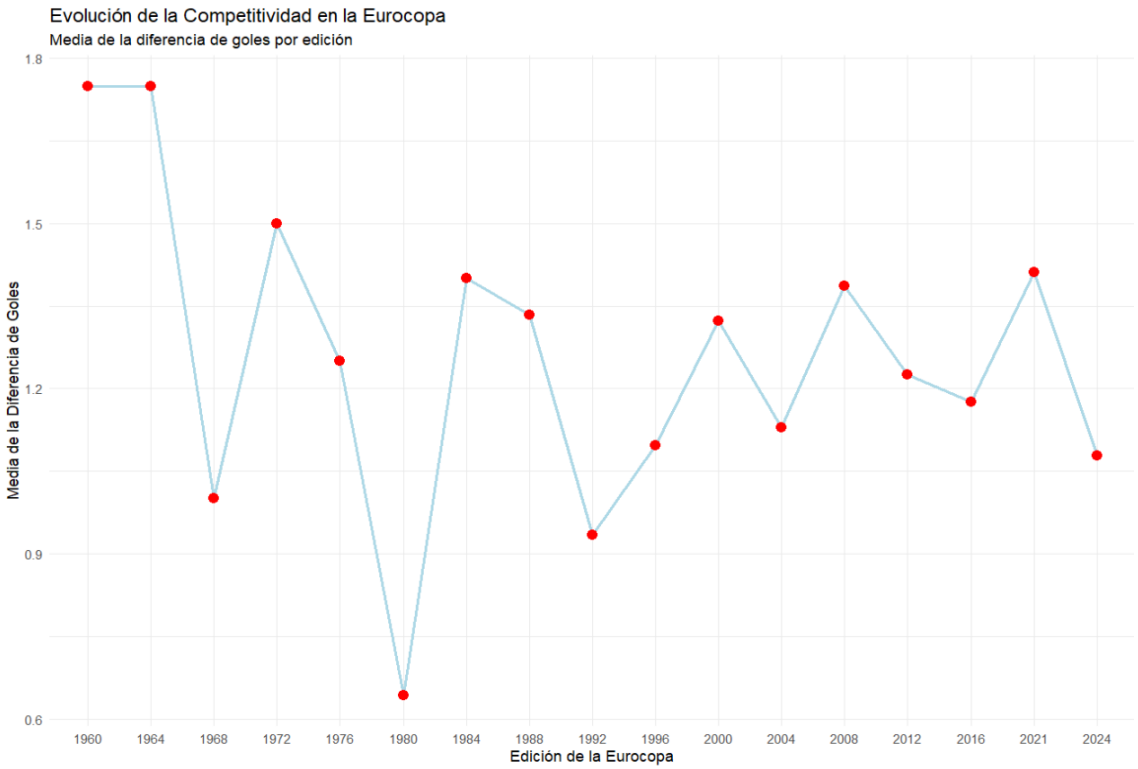
También resulta interesante analizar cómo se ha desarrollado la competitividad entre las distintas ediciones del torneo, es decir, que ediciones fueron las más igualadas. Para ello utilizaremos la columna *goal_difference* de la tabla de resultados, aunque hay que tener en cuenta que, hasta ahora, la tabla contenía valores positivos para la diferencia a favor de los equipos locales y negativos a favor de los equipos que actuaban como visitantes. Para corregir esto, en primer lugar, convertiremos los valores de esta columna a valores absolutos y luego calcularemos la media de estos en cada edición.

```
euro_results$goal_difference_abs <- abs(euro_results$goal_difference)

goal_diff_by_edition <- euro_results %>%

  group_by(edicion) %>%
```

```
summarize(mean_goal_difference = mean(goal_difference_abs, na.rm = TRUE))
```

A continuación, creamos un gráfico para interpretar los resultados:

Gráfico 18: Evolución de la competitividad en la Eurocopa



Fuente: Elaboración propia

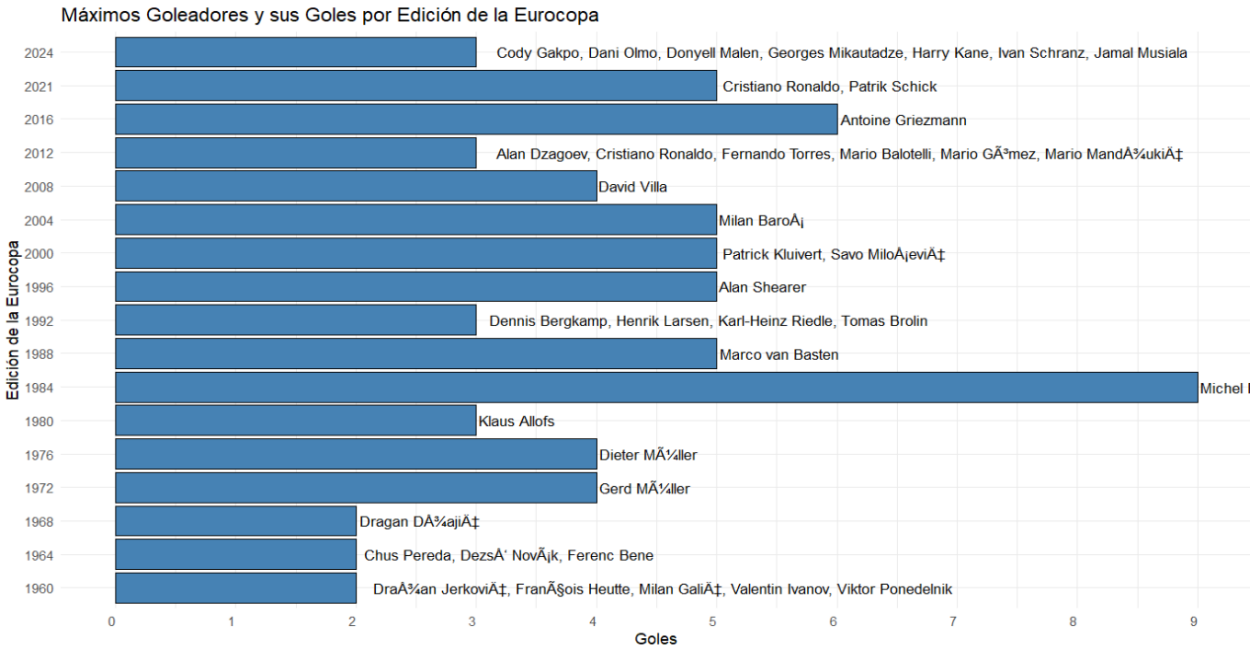
Como se destacó anteriormente, el menor número de partidos de las primeras ediciones del torneo causa una gran varianza entre estas, aunque se observa que en las dos primeras ediciones la diferencia de goles fue la mayor de la historia del torneo. Destaca también la poca diferencia de goles en los partidos de la Eurocopa de 1980, la primera en introducir una fase de grupos, con un valor cercano a los 0,6 goles de diferencia, por lo que se tuvieron que dar un alto número de empates para que esta cifra fuera posible. Durante las últimas ediciones del torneo, este valor se ha estabilizado, se observa que el equipo vencedor del partido lo hace por algo más de un gol de ventaja, aunque la Eurocopa disputada este verano ha sido la más igualada de lo que va de siglo, con un valor muy cercano a uno.

En el caso de las tandas de penaltis disputadas en las distintas ediciones de este torneo, encontramos que los datos de la columna *first_shooter* no registran valores nulos, por lo que un análisis centrado en este torneo podría resultar más fiable que el ejecutado anteriormente. En los resultados, encontramos que, de 25 tandas, 13 victorias son para el equipo que lanzó primero y 12 para el que fue en segundo lugar, por lo que de nuevo,

no se puede concluir que el equipo que lanza primero en una tanda obtenga una ventaja significativa.

Profundizando un poco más en los datos, se pueden extraer grandes cantidades de información según vaya cambiando el objetivo del estudio. Estadísticas como qué países han disputado más partidos en este torneo (Alemania con 58), países que más partidos han hospedado (Francia con 70) o quién es el máximo goleador en la historia de este torneo (Cristiano Ronaldo con 14), incluso en cada edición (Ver gráfico 18), son extraíbles gracias al análisis de datos y a las herramientas de las que disponemos hoy en día en cuestión de segundos.

Gráfico 19: Máximos goleadores de la Eurocopa por edición



Fuente: Elaboración propia

Nota: El máximo goleador de la edición de 1984 fue Michel Platini, cuya increíble actuación, con 9 goles en un solo torneo, le valió para conquistar el título con Francia

4. Análisis de la Eurocopa 2024

4.1. Introducción a la Eurocopa 2024

La Eurocopa es un torneo creado por la UEFA que se organiza cada cuatro años, con el objetivo de determinar cuál es la selección nacional más fuerte de Europa. Su primera edición se celebró en 1960 y desde entonces este torneo se ha consolidado como una de las competiciones de selecciones más prestigiosas del mundo, superada únicamente con claridad por la Copa del Mundo, organizada por la FIFA y en la que se enfrentan las selecciones más fuertes de todo el mundo del fútbol.

Este año 2024 se disputó la 17º edición de este torneo durante los meses de junio y julio. El país anfitrión fue Alemania, que contó con diez sedes en ciudades repartidas por todo el país, alojando estas los partidos en los estadios más emblemáticos del país germano. Esta era la segunda vez que Alemania era elegida anfitriona de este torneo, siendo la anterior la edición del 1988, adicionalmente, este país fue el anfitrión del Mundial en el año 2006.

Esta edición del torneo continuó con el formato elegido por la UEFA desde la edición de 2016, donde 20 equipos se clasifican para el torneo tras unas fases de grupos eliminatorias, la anfitriona, Alemania, se clasifica automáticamente para el torneo y las tres últimas selecciones, que se hayan quedado fuera en la fase de grupos eliminatorias, lo hacen a través de unos play-offs en función del rendimiento de selecciones en la UEFA Nations League, otro torneo organizado por esta confederación.

En total, 24 equipos disputan esta edición del torneo, separados en seis grupos distintos formados por cuatro equipos cada uno. Durante la fase de grupos, cada selección se enfrenta una vez contra cada rival de su grupo, clasificándose las dos primeras de cada uno para la siguiente ronda, y las cuatro mejores terceras de grupo. Una vez finalizada la fase de grupos, las selecciones clasificadas entran en el cuadro de las fases eliminatorias, empezando por los octavos de final. Este cuadro se forma en base a los resultados obtenidos durante la fase de grupos, asegurándose que dos primeras de grupo no puedan coincidir en la primera ronda.

En total, se disputaron 51 encuentros durante la realización del torneo, desde el primer partido inaugural entre la selección anfitriona y Escocia, hasta la final, sin disputarse un partido por el tercer y cuarto puesto, como en otros torneos internacionales como el

Mundial o la Copa América (equivalente a la Eurocopa, pero organizada por la CONMEBOL, con la inclusión de selecciones de la Concacaf en su última edición, disputada también durante este verano).

Gráfico 20: Las 10 sedes de la Eurocopa 2024

Fuente: Sportytrader.es

Respecto a la previa del torneo, Alemania y España se presentaban como las selecciones que más veces habían levantado este trofeo (3). Sin embargo, estas no partían como principales favoritas a ganarlo. A continuación, se presentan las cuotas que ofrecían las principales casas de apuestas internacionales a una semana de empezar el torneo:

Tabla 6: Cuotas de selección ganadora de la Eurocopa 2024 previas al torneo

	Team	Best Odds	Implied Probability
1	England	16/5	24%
2	France	17/4	19%
3	Germany	11/2	15.38%
4	Portugal	8/1	11%
5	Spain	9/1	10%
6	Italy	18/1	5.26%
7	Netherlands	18/1	5.25%
8	Belgium	22/1	4.35%
9	Croatia	45/1	2.04%
10	Denmark	55/1	1.79%

Fuente: Oddschecker.com

Como se puede observar, la selección inglesa era la que partía como favorita, con Francia en segundo lugar y la selección anfitriona en el tercer puesto. Un escalón por debajo se encontraban las selecciones de Portugal y España, que cerraban el top 5 de selecciones que más probabilidades tenían de ganar el torneo antes del comienzo de este, según las casas de apuestas.

Si comparamos esta tabla con una que contenga el valor de las plantillas que viajaban convocadas al torneo (Ver tabla 7), podemos observar una fuerte correlación entre el valor de la plantilla y las cuotas ofertadas por las casas de apuestas. Selecciones que se salen algo de este patrón, como la alemana o la croata, se explican por el rendimiento

en los partidos previos al torneo, la posible ventaja de actuar como local o los resultados históricos de estas selecciones en las últimas ediciones de torneos internacionales.

Tabla 7: Resumen de plantillas selecciones participantes en la Eurocopa 2024

CLUBS STARTING INTO TOURNAMENT AT A LATER POINT

Club	Squad	ø-Age	EURO participations	Foreigners	Market Value	ø-Market Value
England	26	26.3	11	11.5 %	€1.52bn	€58.46m
France	24	26.6	11	75.0 %	€1.23bn	€51.29m
Portugal	25	26.6	9	84.0 %	€1.06bn	€42.56m
Spain	26	27.3	12	34.6 %	€1.04bn	€39.83m
Netherlands	25	26.2	11	76.0 %	€817.00m	€32.68m
Germany	22	27.6	14	18.2 %	€804.00m	€36.55m
Italy	26	26.7	11	15.4 %	€725.50m	€27.90m
Belgium	24	26.6	7	95.8 %	€583.00m	€24.29m
Denmark	26	28.0	10	96.2 %	€415.50m	€15.98m
Ukraine	26	26.5	4	50.0 %	€379.00m	€14.58m
Türkiye	26	26.0	6	42.3 %	€352.90m	€13.57m
Serbia	28	24.7	1	82.1 %	€324.20m	€11.58m
Croatia	24	26.7	7	83.3 %	€310.50m	€12.94m
Switzerland	24	27.4	6	95.8 %	€286.00m	€11.92m
Austria	26	27.1	4	88.5 %	€237.00m	€9.12m
Poland	26	28.0	5	88.5 %	€210.30m	€8.09m
Scotland	26	28.5	4	73.1 %	€207.40m	€7.98m
Czech Republic	26	25.6	8	42.3 %	€187.90m	€7.23m
Georgia	23	26.7	1	100.0 %	€181.40m	€7.89m
Hungary	26	27.5	5	76.9 %	€165.45m	€6.36m
Slovakia	26	27.4	3	84.6 %	€158.40m	€6.09m
Slovenia	26	27.4	2	92.3 %	€148.25m	€5.70m
Albania	26	27.4	2	100.0 %	€113.10m	€4.35m
Romania	26	27.2	6	73.1 %	€95.53m	€3.67m

Fuente: Transfermarkt.com

Otras columnas interesantes en esta tabla son el tamaño de la plantilla, la edad media de esta, el número de participaciones en este torneo o el porcentaje de jugadores de la plantilla que juegan en una liga extranjera.

Otras selecciones que destacar, a parte de las favoritas, eran Serbia y Georgia, las cuáles se consiguieron clasificar por primera vez a este torneo, y además contando la primera de ellas con la plantilla más joven del torneo. Selecciones como Ucrania, Hungría o Escocia venían de rendir a un buen nivel en partidos previos al torneo.

Sin embargo, los favoritismos no importan en el mundo del fútbol una vez empieza a rodar el balón y algunas selecciones demostraron su nivel mejor que otras durante la fase de grupos. A continuación, se presentan los resultados de esta fase:

Tabla 8: Clasificación fase de grupos Eurocopa 2024

Group A

		Played	Won	Drawn	Lost	For	Against	Goal difference	Points	Form guide
1	Germany	3	2	1	0	8	2	6	7	W W D
2	Switzerland	3	1	2	0	5	3	2	5	W D D
3	Hungary	3	1	0	2	2	5	-3	3	L L W
4	Scotland	3	0	1	2	2	7	-5	1	L D L

Group B

		Played	Won	Drawn	Lost	For	Against	Goal difference	Points	Form guide
1	Spain	3	3	0	0	5	0	5	9	W W W
2	Italy	3	1	1	1	3	3	0	4	W L D
3	Croatia	3	0	2	1	3	6	-3	2	L D D
4	Albania	3	0	1	2	3	5	-2	1	L D L

Group C

		Played	Won	Drawn	Lost	For	Against	Goal difference	Points	Form guide
1	England	3	1	2	0	2	1	1	5	W D D
2	Denmark	3	0	3	0	2	2	0	3	D D D
3	Slovenia	3	0	3	0	2	2	0	3	D D D
4	Serbia	3	0	2	1	1	2	-1	2	L D D

Group D

		Played	Won	Drawn	Lost	For	Against	Goal difference	Points	Form guide
1	Austria	3	2	0	1	6	4	2	**6**	L W W
2	France	3	1	2	0	2	1	1	**5**	W D D
3	Netherlands	3	1	1	1	4	4	0	**4**	W D L
4	Poland	3	0	1	2	3	6	-3	**1**	L L D

Group E

		Played	Won	Drawn	Lost	For	Against	Goal difference	Points	Form guide
1	Romania	3	1	1	1	4	3	1	**4**	W L D
2	Belgium	3	1	1	1	2	1	1	**4**	L W D
3	Slovakia	3	1	1	1	3	3	0	**4**	W L D
4	Ukraine	3	1	1	1	2	4	-2	**4**	L W D

Group F

		Played	Won	Drawn	Lost	For	Against	Goal difference	Points	Form guide
1	Portugal	3	2	0	1	5	3	2	**6**	W W L
2	Türkiye	3	2	0	1	5	5	0	**6**	W L W
3	Georgia	3	1	1	1	4	4	0	**4**	L D W
4	Czechia	3	0	1	2	3	5	-2	**1**	L D L

Fuente: UEFA.com

En concreto, España y Alemania salieron reforzados de esta fase inicial al ofrecer el mejor fútbol, así como los mejores resultados. En el caso de la selección española, fue la única capaz de ganar sus tres partidos en la fase de grupos (en el grupo más complicado del torneo basándose en el ranking de selecciones) sin conceder un solo gol, mientras que la selección alemana sólo empató el último partido, que les enfrentaba contra Suiza, estando ya clasificados para la siguiente ronda.

Destacó mucho, pero negativamente, la fase de grupos realizada por las dos selecciones, a priori, favoritas para llevarse el torneo. La selección francesa sólo consiguió anotar dos goles durante esta fase del torneo, uno de penalti y otro en propia meta, adicionalmente, sólo consiguió una victoria, por la mínima ante Austria en el primer partido del grupo y acabó pasando a la siguiente ronda como segunda de grupo, por detrás de precisamente la selección austriaca. Por su parte, la selección inglesa tampoco consiguió mejores resultados, con los mismos puntos tras una victoria y dos empates, desplegando además un juego muy pobre, sobre todo a nivel ofensivo, que resultó en numerosas críticas por parte de la prensa del país, especialmente dirigidas hacia el seleccionador inglés Gareth Southgate. Si algo positivo pudo sacar la selección inglesa, es que, debido a otros resultados en su grupo, todos empates excepto su victoria por la mínima ante Serbia, pasó como primera de grupo, lo que resultó ser muy valioso posteriormente respecto al lado del cuadro eliminatorio en el que acabó, como se observará posteriormente.

Respecto al resto de la fase de grupos, cabe destacar el buen papel presentado por Suiza, que consiguió asegurarse el segundo puesto, dejando fuera a Hungría que venía en una racha muy positiva previa al torneo. Italia consiguió clasificarse como segunda de grupo gracias a un gol en el último minuto de descuento que mandó a la selección croata a casa, quizás de manera algo injusta viendo el fútbol desplegado por ambas selecciones. Austria y Rumanía cuajaron una buena fase de grupos, ambas clasificadas como primeras de grupo, por encima de selecciones a priori de mayor nivel como Francia, Países Bajos o Bélgica. Ucrania no demostró el buen nivel presentado con anterioridad, quedando última en un grupo que tuvo un cuádruple empate, teniendo todos sus participantes una victoria, un empate y una derrota, y quedando fuera por su peor diferencia de goles. En el último grupo, destacó la clasificación de Georgia como tercera de grupo tras una sorprendente victoria frente a Portugal en el último partido, la cual ya estaba clasificada, así como un buen papel de Turquía, que presentó un fútbol atrevido que hizo que sus partidos tuvieran muchos goles y fueran entretenidos para el espectador.

Tras la finalización de la fase de grupos, todos los primeros y segundos de cada grupo, así como los cuatro mejores equipos que habían quedado terceros se clasificaron para las fases eliminatorias, el cuadro de estas fases quedó así:

Gráfico 21: Cuadro fases finales Eurocopa 2024

Fuente: As.com

Algo que destacó bastante respecto a este cuadro es el desbalance que había entre ambos lados. En el lado izquierdo del cuadro, se encontraban Alemania, España, Portugal y Francia, cuatro de las cinco selecciones que partían como favoritas para ganar el torneo al principio de este. Por el otro lado, la selección inglesa se topó con un lado del cuadro en el que era altamente favorita a pesar del pobre juego demostrados sobre el campo. Esto se vio reflejado en las cuotas ofrecidas por las casas de apuestas, las cuales, a pesar de lo visto en la fase de grupos siguieron apostando por Inglaterra como la selección favorita para ganar el torneo.

No hubo grandes sorpresas en los octavos de final, al menos por el lado izquierdo del cuadro, donde las cuatro selecciones favoritas pasaron y se vieron las caras en unos cuartos de final que prometían mucho. Por el otro lado del cuadro, las sorpresas fueron las victorias de Turquía sobre Austria, la cual no fue capaz de mostrar el mismo nivel que durante su excelente fase de grupos, y la victoria de Suiza sobre Italia, que fue sorprendente sobre el papel, pero no para aquellos que vieron jugar a estas selecciones durante la fase de grupos, pues Suiza jugó a un nivel considerablemente superior al de una selección italiana que demostró muy poco. La selección inglesa por su parte, estuvo muy cerca de ser eliminada por una selección eslovaca que era a priori muy inferior, necesitando los ingleses de un gol de chilena de Jude Bellingham en los últimos minutos del tiempo reglamentario para llevar el partido a la prórroga, en la cual los ingleses sí que fueron superiores.

Los cuartos de final empezaron con el partido que fue, para muchos, la final anticipada del torneo, y es que las dos selecciones que mejor papel habían mostrado hasta el momento, Alemania y España, se vieron las caras en el primer partido de cuartos. Este partido se saldó con una victoria al final de la prórroga para la selección española, gracias a un gran cabezazo de Mikel Merino, que decantó un partido muy igualado a favor de La Roja. Fue tras este partido cuando la selección favorita para ganar el torneo cambió gracias a la buena imagen del combinado español y a superar a otra de las grandes favoritas. Por otro lado, Francia venció en penaltis a Portugal, tras un partido sin goles.

En el otro lado del cuadro, la selección inglesa y la holandesa superaron a sus rivales en partidos muy ajustados, necesitando Inglaterra la tanda de penaltis para superar a Suiza, en otro partido no muy convincente de los británicos.

Las semifinales del torneo fueron igual de ajustadas que la ronda anterior, con resultados igualados, 2-1 en ambos casos, en partidos que dispusieron de ocasiones para ambos equipos y que no necesitaron de prórroga ni penaltis para resolverse. España remontó el gol inicial de Francia, gracias en parte a uno de los mejores goles del torneo a manos de Lamine Yamal, mientras que los ingleses disputaron el que seguramente fue su partido más convincente hasta la fecha.

Se presentó entonces una final España contra Inglaterra, que brindaba la oportunidad al combinado español de consagrarse como la selección con más títulos de este torneo, mientras que los ingleses soñaban con levantar este trofeo por primera vez en su historia. La final fue un partido igualado y entretenido, con diferencias claras entre los estilos de juego presentados por cada seleccionador, que se acabó llevando la selección española por 2-1 gracias a un tardío gol de Mikel Oyarzabal que permitió a La Roja levantar su cuarta Eurocopa.

Se dio por finalizado así un torneo destacado por la gran igualdad entre selecciones de mayor y menor nivel, siendo, como se observó en el análisis del apartado anterior del trabajo, la edición del torneo con menor diferencia de goles entre equipos de lo que va de siglo, síntoma de esta igualdad.

El cuadro de las fases finales y sus resultados es el siguiente:

Gráfico 22: Resultados cuadro fases finales Eurocopa 2024

Fuente: UEFA.com

El once ideal del torneo, elegido por la UEFA tras la finalización de este fue el siguiente:

Gráfico 23: Once ideal de la Eurocopa 2024

Fuente: UEFA.com

4.2. Obtención y preprocesado de datos

Para realizar un análisis más en profundidad de este torneo, se van a abandonar los datasets trabajados en el apartado de análisis histórico, en búsqueda de conjuntos de datos que tengan estadísticas más especializadas. Para ello, se recurre a la página StatsBomb, que surge como un blog de analítica de fútbol en el año 2013 y se ha desarrollado hasta ser hoy en día una de las mayores fuentes de datos sobre este deporte del mundo.

Como parte de su iniciativa por apoyar al desarrollo de analíticos de datos, esta página web tiene a disposición del público una gran cantidad de datos sobre diferentes competiciones y temporadas. Entre ellos, publicaron al finalizar el torneo paquetes disponibles en R y Python con alrededor de 3.400 eventos sobre cada uno de los 51 partidos disputados durante el torneo.

Gracias a esta página y al soporte que da a futuros analistas que se está iniciando en este ámbito, es posible extraer de su base de datos varios datasets que nos resulten útiles para llevar a cabo el análisis más detallado deseado para este apartado.

En primer lugar, crearemos un nuevo proyecto en RStudio, en el que instalaremos debido a la indicación de StatsBomb, los paquetes ggplot2 (que ya teníamos instalado, útil para la visualización de datos), tidyverse (útil para manipular datos) y devtools (que nos permite instalar paquetes desde Github). Gracias a estos paquetes podremos instalar el paquete específico creado por esta empresa para R, con la instrucción:

```
devtools::install_github("statsbomb/StatsBombR")
```

Una vez instalado, siguiendo las indicaciones disponibles para este paquete, ejecutaremos las siguientes instrucciones para obtener un dataset con los datos generales de todos los partidos y otro con todos los eventos de estos partidos disponibles de forma gratuita:

```
Comp <- FreeCompetitions() %>%

  filter(competition_id==55 & season_id==282)

Matches <- FreeMatches(Comp)

StatsBombData <- free_allevents(MatchesDF = Matches, Parallel = T)

StatsBombData = allclean(StatsBombData)
```

El dataset *Matches* cuenta con información más detallada sobre los resultados de los partidos que los utilizados en el anterior apartado. Por otro lado, el dataset *StatsBombData* contiene 181 variables posibles, cada una utilizada para un tipo de evento registrado distinto, por lo que resultará de vital importancia conocer los distintos tipos de evento que se registran, las variables que estos utilizan y quedarnos con aquellos que nos resulten interesantes a la hora de realizar nuestro análisis.

Las tablas con las indicaciones sobre cada variable se pueden encontrar en el documento de guía que ofrece la propia empresa en su página web o en su página de GitHub.

Cabe destacar que, en este caso, no se llevarán a cabo tantas acciones durante la etapa de preprocesado de los datos más allá de la carga de estos en el programa, ya que estos vienen bien procesados y limpiados por parte de StatsBomb y aquí se filtran ya exclusivamente los partidos del torneo a analizar.

Se podría separar cada tipo de evento en el dataset *StatsBombData* y eliminar columnas para reducir el número de valores nulos en el dataset, sin embargo, esto se realizará conforme se necesiten los eventos durante la etapa de análisis.

4.3. Análisis del torneo

Se va a dar comienzo ahora a la fase del análisis de la Eurocopa 2024. Las posibilidades con un conjunto de datos tan completo como el que se dispone son infinitas, sin embargo, debido a la longitud del trabajo, dividiremos este apartado dos partes. Como se ha mencionado en el apartado anterior, en primer lugar, se realizará un análisis de las estadísticas generales del torneo, para posteriormente profundizar algo más en otras estadísticas para analizar el rendimiento individual de los jugadores en distintas alturas del campo, desde los delanteros hasta los porteros.

En primer lugar, vamos a comenzar analizando estadísticas generales sobre los equipos que disputaron este torneo. Empecemos con algunas estadísticas ofensivas como los goles marcados o los tiros intentados.

Para ello vamos a crear una tabla llamada *shot_goals*, con los tiros y los goles de cada selección durante el torneo. Resultará interesante añadir a esta tabla una nueva columna que indique la efectividad de cada selección, es decir, cuantos goles anotó cada equipo por disparo intentado. Esta métrica nos permitirá ver que selecciones fueron mejores de cara a puerta respecto al número de ocasiones de las que dispusieron.

Por último, para eliminar el efecto que conlleva que unas selecciones lleguen más lejos en el torneo que otras y, por lo tanto, disputen más partidos, se va a crear otra tabla shot_goals_p90 que índice el número de tiros y de goles de cada selección por partido disputado.

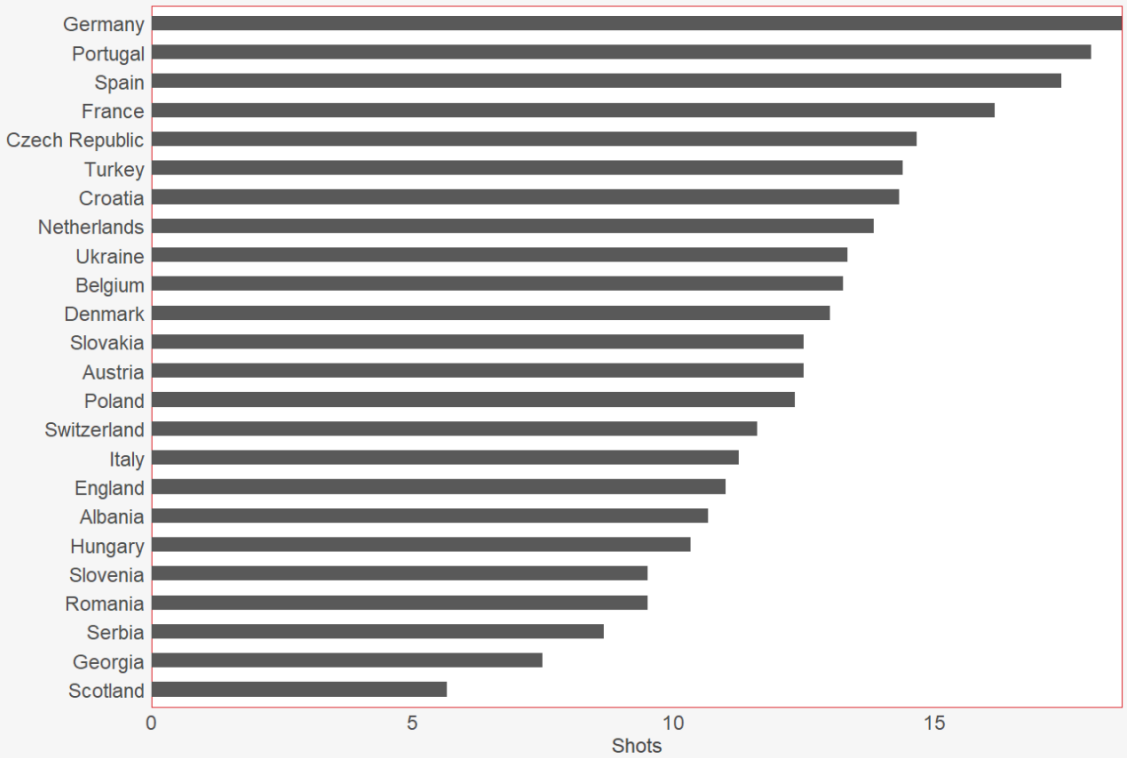
```
#Tiros y goles totales

shots_goals = StatsBombData %>%

  group_by(team.name) %>%

  summarise(shots = sum(type.name=="Shot", na.rm = TRUE),

            goals = sum(shot.outcome.name=="Goal", na.rm = TRUE))

#Efectividad

shots_goals$efectiveness <- shots_goals$goals / shots_goals$shots
```

```
#Tiros y goles por partido

shots_goals_p90 = StatsBombData %>%

  group_by(team.name) %>%

  summarise(shots = sum(type.name=="Shot", na.rm = TRUE)/n_distinct(match_id),

            goals = sum(shot.outcome.name=="Goal", na.rm = TRUE)/n_distinct(match_id))
```

Cabe destacar que tras trabajar con los datos se descubre que las tandas de penaltis se contabilizan en el dataset como tiros, sin embargo, no nos interesan para este análisis, por lo que se debe incluir un filtro que indique al programa que ignore aquellos tiros efectuados tras la finalización de la prórroga, ya que nuestros resultados se verían afectados, esto se realizará usando la variable *period,* que indica el tramo del partido en la que ocurrió el evento.

Una vez disponemos de estas estadísticas, podemos realizar una serie de gráficos para facilitar su visualización:

Gráfico 24: Tiros por partido Eurocopa 2024



Fuente: Elaboración propia

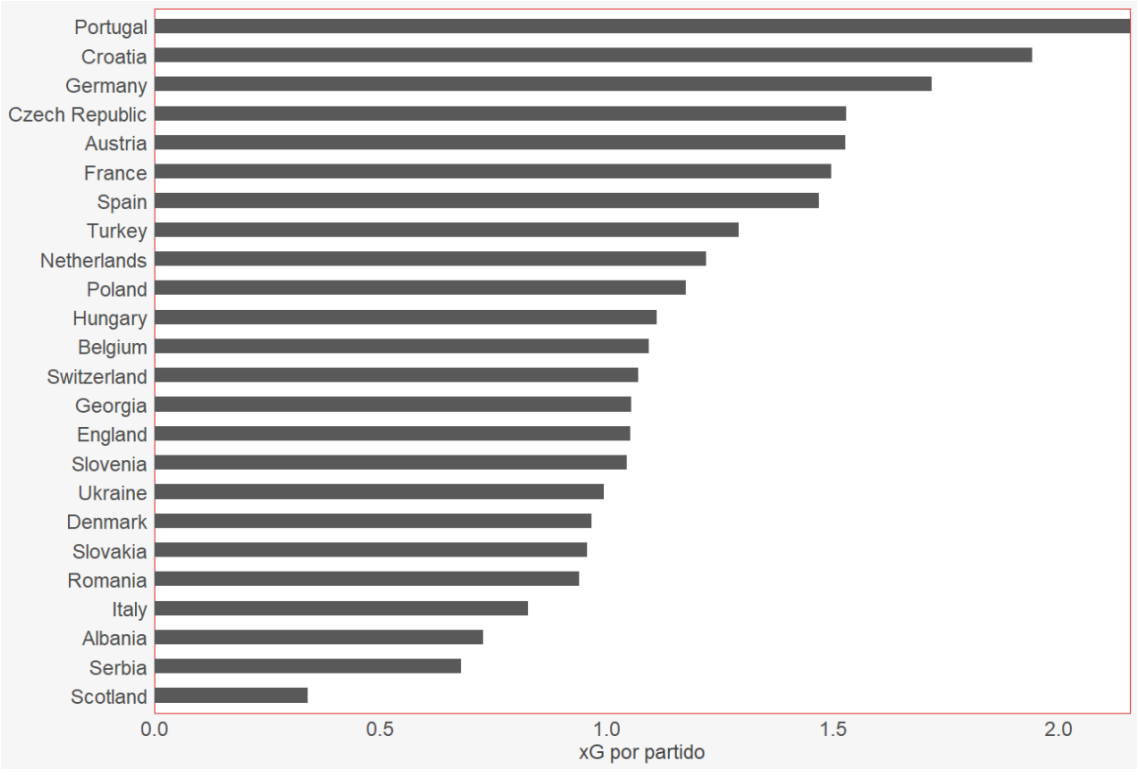
Como podemos observar en este gráfico, que representa los tiros por partido de cada selección, cuatro de las cinco favoritas a ganar el título ocupan las primeras posiciones. Ya se puede empezar a observar que la que falta, Inglaterra, no fue una selección demasiado productiva en ataque.

Destacan en este gráfico respecto a las selecciones que se clasificaron en la fase de grupos y las que no, la aparición de la República Checa, Croacia o Ucrania entre las mejores selecciones, a pesar de no clasificarse para la siguiente ronda. Lo contrario se observa en Georgia, Rumanía o Eslovenia, que sí que se clasificaron a pesar de ocupar posiciones bajas en esta tabla.

Y es que tirar mucho no siempre se traduce en meter muchos goles. Hace unos años se introdujo en el mundo del fútbol una nueva estadística llamada goles esperados (xG), que mide la calidad de los tiros efectuados durante un partido, asignándoles una probabilidad de que acaben en gol respecto a los tiros similares efectuados bajo circunstancias similares en el pasado.

Conseguimos los xG por cada partido disputado de cada selección de manera similar a como conseguimos los tiros y los goles y obtenemos el siguiente gráfico:

Gráfico 25: xG medio por partido de cada selección Eurocopa 2024



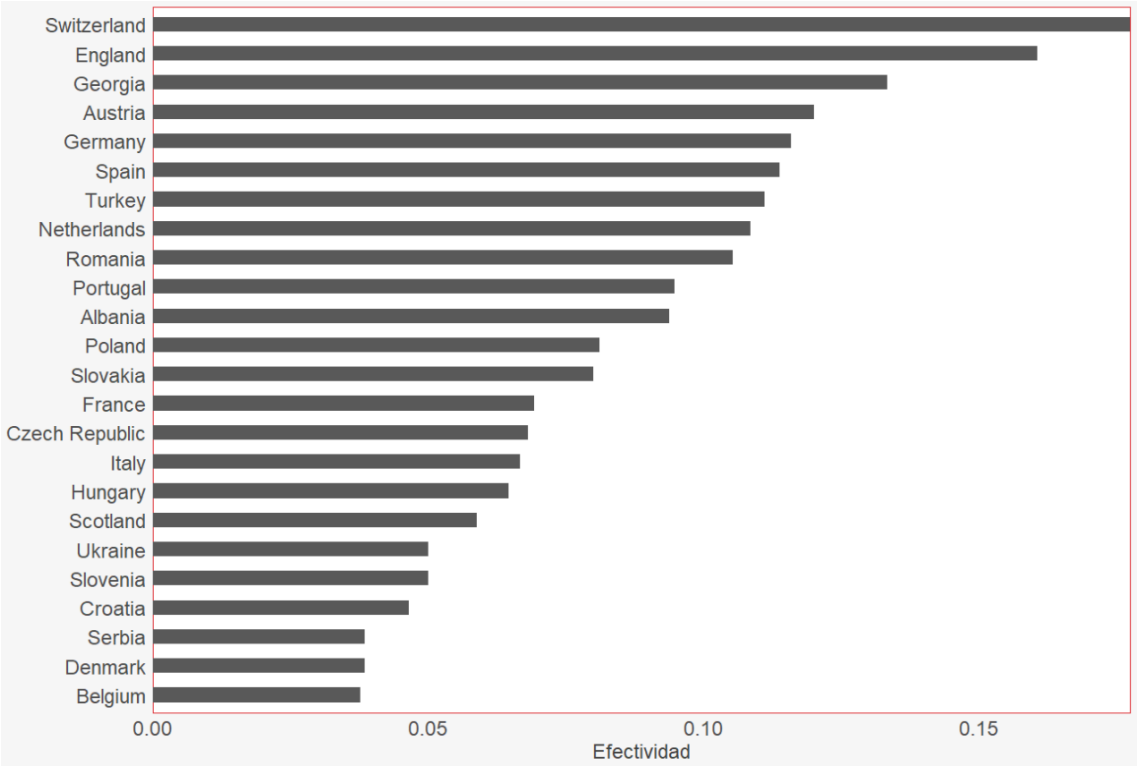
Fuente: Elaboración propia

Esta tabla guarda cierta correlación respecto a la anterior, al fin y al cabo, cuantos más tiros realices, más veces sumas xG al total, aunque ciertas selecciones pueden realizar muchos tiros de poco valor, como tiros lejanos o muy escorados.

Sigue sorprendiendo la posición de Croacia, eliminada como una de las peores terceras de grupo, especialmente si la comparamos con la selección que la dejó fuera en el último minuto, Italia, la cual baja aún mas de posición en este ranking, evidenciando la poca producción ofensiva de esta selección. No sorprende que las tres últimas selecciones en este ranking cayeran eliminadas en la fase de grupos, y como últimas de grupo, al producir muy pocas ocasiones durante sus partidos.

Por último, vamos a observar una gráfica más, que mide la efectividad de las selecciones durante el torneo, observamos el grafico sobre la variable *efectiveness* creada anteriormente:

Gráfico 26: Efectividad de cada selección según sus goles por tiro Eurocopa 2024



Fuente: Elaboración propia

Como podemos observar, algunas selecciones como la suiza, la inglesa o la georgiana aprovecharon muy bien sus tiros, con un ratio de conversión alto, lo que podría explicar en parte su clasificación a las fases finales a pesar de no destacar en los anteriores rankings. Por otro lado, observamos como Portugal, selección que más tiros y xG por

partido tenía, no tuvo una efectividad tan alta. Destacan Bélgica y Dinamarca, últimas en este ranking a pesar de clasificarse para la siguiente ronda y la gran caída de Croacia, que no supo aprovechar sus oportunidades a pesar de tener un gran número de ellas, lo que explica su temprana eliminación del torneo.

Una vez observadas las estadísticas generales ofensivas, vamos a dar un repaso a las estadísticas defensivas. En primer lugar, vamos a observar el número de goles recibidos en contra. Para ello utilizaremos el dataset *Matches* y le asignaremos a cada equipo el valor de goles marcado por el equipo rival, posteriormente sumaremos el número de goles recibidos agrupando los datos por equipo. Resultará interesante extraer los goles recibidos de media por partido, pues no todas las selecciones disputaron el mismo número de partidos y esto podría llevar a que selecciones que cayeron eliminadas en las primeras rondas contaran con un mejor registro defensivo.

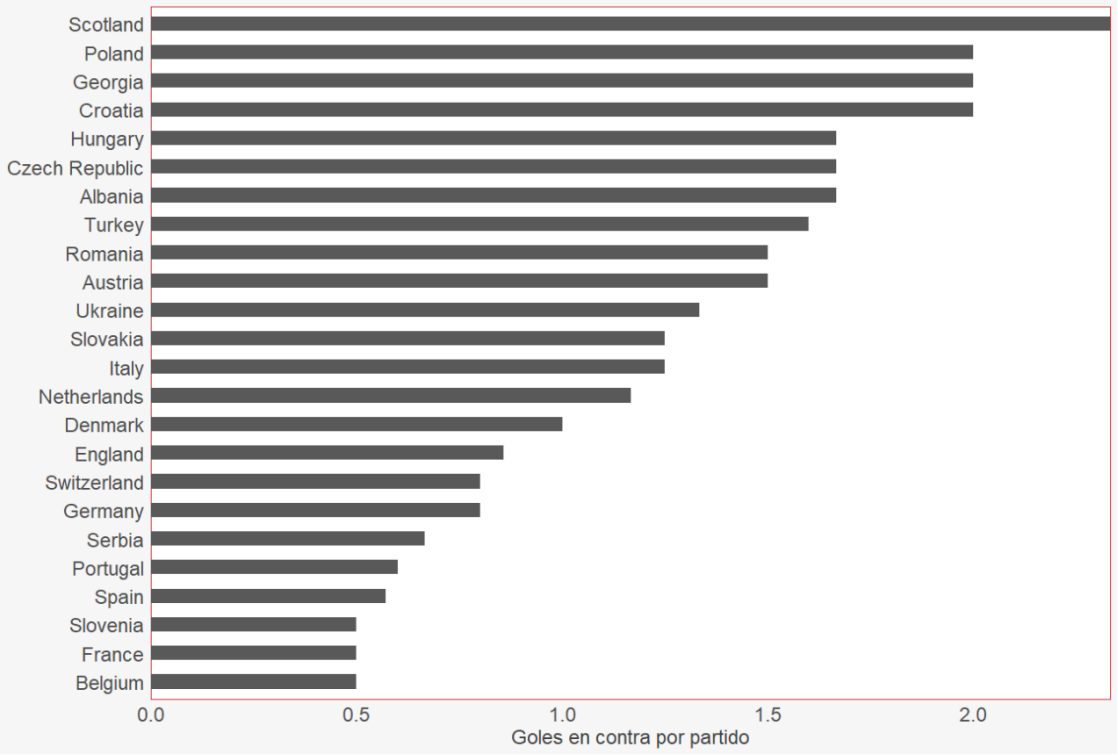
El código usado en este paso es el siguiente:

```
# Calcular los goles en contra por equipo
goles_contra <- Matches %>%
  select(team.name = home_team.home_team_name, goles_contra = away_score) %>%
  bind_rows(
    Matches %>% select(team.name = away_team.away_team_name, goles_contra = home_score)
  ) %>%
  group_by(team.name) %>%
  summarise(
    goles_en_contra = sum(goles_contra, na.rm = TRUE),
    goles_contra_por_partido = mean(goles_contra, na.rm = TRUE)
  )
```

```
# Gráfico de barras para goles en contra promedio

ggplot(goles_contra, aes(x = reorder(team.name, goles_contra_por_partido), y = goles_contra_por_partido)) +

  geom_bar(stat = "identity", width = 0.5) +

  labs(y="Goles en contra por partido") +

  theme(axis.title.y = element_blank()) +

  scale_y_continuous( expand = c(0,0)) +

  coord_flip() +

  theme_SB()
```

Obtenemos el siguiente gráfico:

Gráfico 27: Media de goles en contra por partido Eurocopa 2024



Fuente: Elaboración propia

Como podemos observar en este gráfico, gran parte de las selecciones que promedian más goles en contra fueron eliminadas durante la fase de grupos. En primer lugar, se encuentra Escocia, la cual recibió durante el partido inaugural el mayor número de goles encajados en un único partido de todo el torneo (cinco) de manos de la anfitriona.

Destaca también el elevado número de goles encajados por Georgia al conseguir esta pasar la fase de grupos, aunque su partido de octavos de final contra la selección española aumenta su media considerablemente, al encajar cuatro goles en ese partido.

Por la parte positiva destacan las selecciones belga y francesa, las cuales explican su continuidad en el torneo pasada la fase de grupos a pesar de contar con unas cifras ofensivas más bajas, especialmente en el caso de Bélgica. Sin embargo, la mayor sorpresa de este gráfico es la selección eslovena, la cual obtiene el promedio más bajo junto a las dos mencionadas anteriormente, a pesar de clasificarse como tercera en un grupo complicado y de enfrentarse a Portugal en octavos de final, selección con buenas estadísticas ofensivas, pero que tuvo que recurrir a los penaltis para eliminar a Eslovenia tras un resultado sin goles al terminar la prórroga. Eslovenia consigue además el menor número total de goles encajados durante el torneo (dos), junto a Bélgica y Serbia, la cual es eliminada en fase de grupo a pesar de su buen hacer defensivo. La selección española por su parte, cuenta con un buen trabajo defensivo, que junto a su buen desempeño en el otro lado del campo explican los buenos resultados obtenidos durante el torneo.

Sin embargo, los goles encajados no siempre cuentan la historia al completo, como se explicó en el apartado ofensivo, los equipos pueden disponer de grande oportunidades, pero no aprovecharlas. Es por ello que volveremos a utilizar los goles esperados para poder observar a que selecciones les crearon las mayores oportunidades.

Para ello, utilizaremos ahora el dataset *StatsBombData*, ya que ahí se encuentra el valor de xG asociado a cada acción de tiro. Estos se sumarán para cada equipo y partido y se unirán al dataset *Matches* para poder asignar estos valores al equipo rival, ya que buscamos xG en contra. De nuevo, habrá que filtrar los tiros para que no se contabilicen las aquellos efectuados durante las tandas de penaltis, ya que estas no tienen por qué significar un mal rendimiento defensivo y contaminarían los resultados. Una vez asignados los valores al rival, se crearán dataframes con agrupaciones según la selección para poder obtener los valores de xG en contra promedios de cada una. Cabe destacar que la dificultad de este paso aumenta considerablemente al no tener los eventos asociados a los partidos ninguna variable que los relacione con el equipo rival, por lo que es necesaria el dataset de *Matches* para asociarlos entre sí.

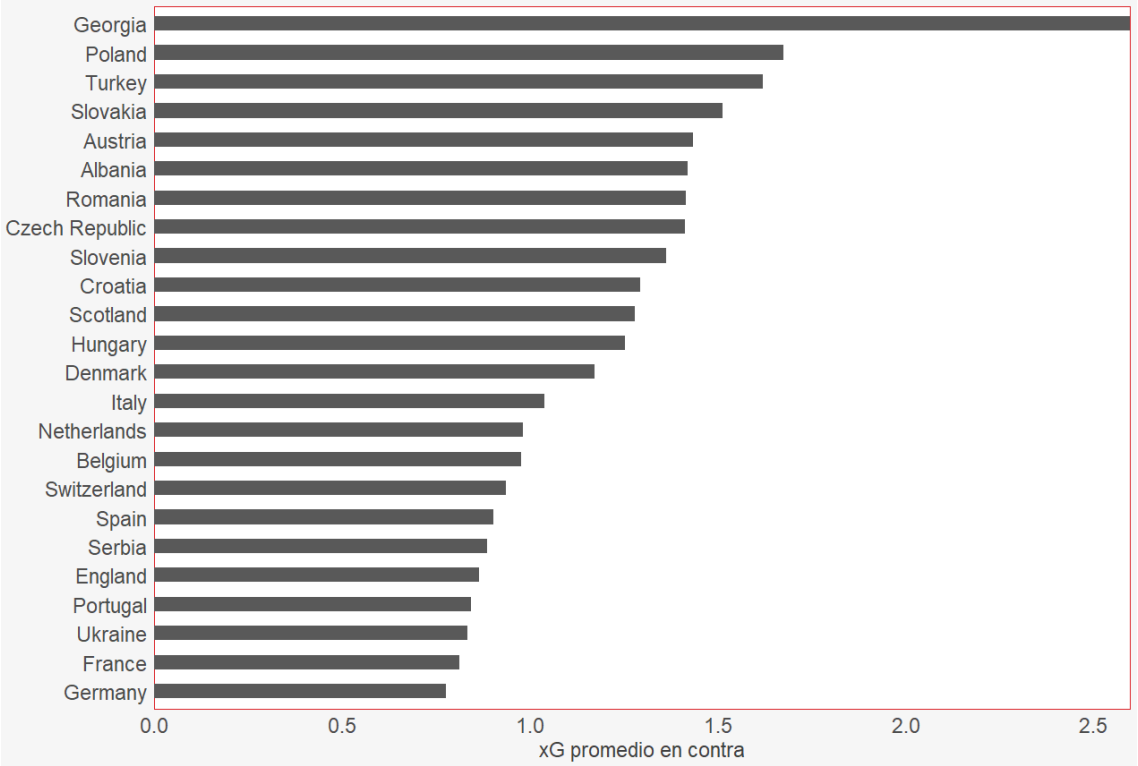
Las instrucciones utilizadas para llevar a cabo todo esto son las siguientes:

```
# Filtrar los eventos de tipo 'Shot' y sumar los xG
xg_por_equipo <- StatsBombData %>%
  filter(type.name == "Shot" & period < 5) %>%
  group_by(match_id, team.name) %>%
  summarise(xG = sum(shot.statsbomb_xg, na.rm = TRUE)) %>%
  ungroup()
# Unir xG con partidos para obtener xG en contra
xg_contra <- Matches %>%
  select(match_id, home_team.home_team_name, away_team.away_team_name) %>%
  left_join(xg_por_equipo, by = c("match_id", "home_team.home_team_name" = "team.name")) %>%
  rename(xG_home = xG) %>%
  left_join(xg_por_equipo, by = c("match_id", "away_team.away_team_name" = "team.name")) %>%
  rename(xG_away = xG) %>%
  mutate(xG_home = ifelse(is.na(xG_home), 0, xG_home),
         xG_away = ifelse(is.na(xG_away), 0, xG_away))
# Crear un dataframe con equipo y xG en contra
xG_against <- xg_contra %>%
  select(match_id, home_team.home_team_name, xG_away) %>%
  rename(team = home_team.home_team_name, xG_against = xG_away) %>%
  bind_rows(
    xg_contra %>%
```

```
    select(match_id, away_team.away_team_name, xG_home) %>%
    rename(team = away_team.away_team_name, xG_against = xG_home)
)
#Crear dataframe agrupado por selecciones para xG promedio en contra
xG_against_team90 = xG_against %>%
 group_by(team) %>%
 summarise(xG_against_p90 = sum(xG_against, na.rm = TRUE)/n_distinct(match_id))
```

El resultado de todo esto se muestra en la siguiente gráfica:

Gráfico 28: xG promedio en contra Eurocopa 2024



Fuente: Elaboración propia

En primer lugar, destaca el elevado valor de la selección georgiana, ya que conseguir pasar la fase de grupos cuando te crean tantas ocasiones y no eres el mejor equipo ofensivamente resulta difícil de creer. Uno podría pensar que su partido de octavos de final contra España tiene una gran influencia en este valor, sin embargo, durante este

partido la selección española le generó unos goles esperados de 2,76 siendo este un valor que no se aleja mucho de la media. Todos los equipos que jugaron contra la selección de Georgia le generaron más de 2,35 goles esperados, por lo que es realmente sorprendente su clasificación a octavos de final.

En el otro lado del ranking, encontramos a las selecciones que menos le permitieron generar a sus rivales, aquí encontramos entre todas las favoritas a ganar el torneo y clasificadas pasada la fase de grupos a Ucrania y Serbia. Destaca especialmente la primera de estas, ya que no se encontraba cerca de los mejores puestos en el ranking de goles en contra, es decir, esta selección encajó considerablemente más goles de los que debería según el sistema de goles esperados. Esto se puede deber a que sus rivales aprovechasen muy bien sus ocasiones, o que la defensa y en especial, el portero, al cual se le suelen atribuir estas diferencias entre goles esperados y encajados, no estuvieran a su mejor nivel durante este torneo, lo que podría explicar su pronta eliminación. En el caso opuesto encontramos a Eslovenia, la cual no destaca en este ranking, pero sí que obtiene una media de goles encajados por partido muy inferior a los esperados, pudiendo atribuir a su capitán y estrella del equipo, el portero Jan Oblak, gran parte del mérito de su clasificación a octavos de final.

Hasta ahora hemos analizado estadísticas generales de las distintas selecciones que participaron en el torneo, pero estos datos nos permiten profundizar mucho más, gracias a ellos podemos adentrarnos en los eventos de un partido o de un jugador en específico.

Vamos a analizar por ejemplo el rendimiento de los jugadores ofensivos, para ello, podemos utilizar dos métricas muy comunes en un análisis de este tipo, los tiros realizados y los pases claves efectuados con éxito. Se entiende por pase clave aquel que genera una ocasión manifiesta de gol.

Para ello, crearemos un nuevo dataframe que contenga estas dos métricas al que llamaremos *player_shots_keypasses*. Adicionalmente, crearemos otro dataframe con los minutos jugados por cada jugador durante el torneo, gracias a la función del paquete de StatsBombR get.minutesplayed(), que devuelve los minutos de cada jugador durante cada partido y agrupándolos, lo llamaremos *player_minutes*. Esto será necesario para estandarizar los valores, ya que no todos los jugadores disputaron el mismo número de partidos. Adicionalmente, esto nos permitirá aplicar un filtro a nuestro ranking, ya que encontraremos jugadores que disputaran muy pocos minutos durante el torneo, pero consiguieron algún tiro o algún pase clave, por lo que encontramos valores muy altos pero que no son representativos, se decide que el filtro serán 180 minutos jugados, es decir, la duración de dos partidos enteros.

Una vez dispongamos de ambos dataframes, los podremos juntar en uno para obtener los tiros y pases claves de cada jugador por cada 90 minutos (duración de un partido estándar), que ordenaremos de manera descendente y filtraremos para los 20 mejores jugadores en esta métrica para poder visualizar un gráfico que no se encuentre demasiado poblado, dificultando su lectura.

El código utilizado para esto se muestra a continuación:

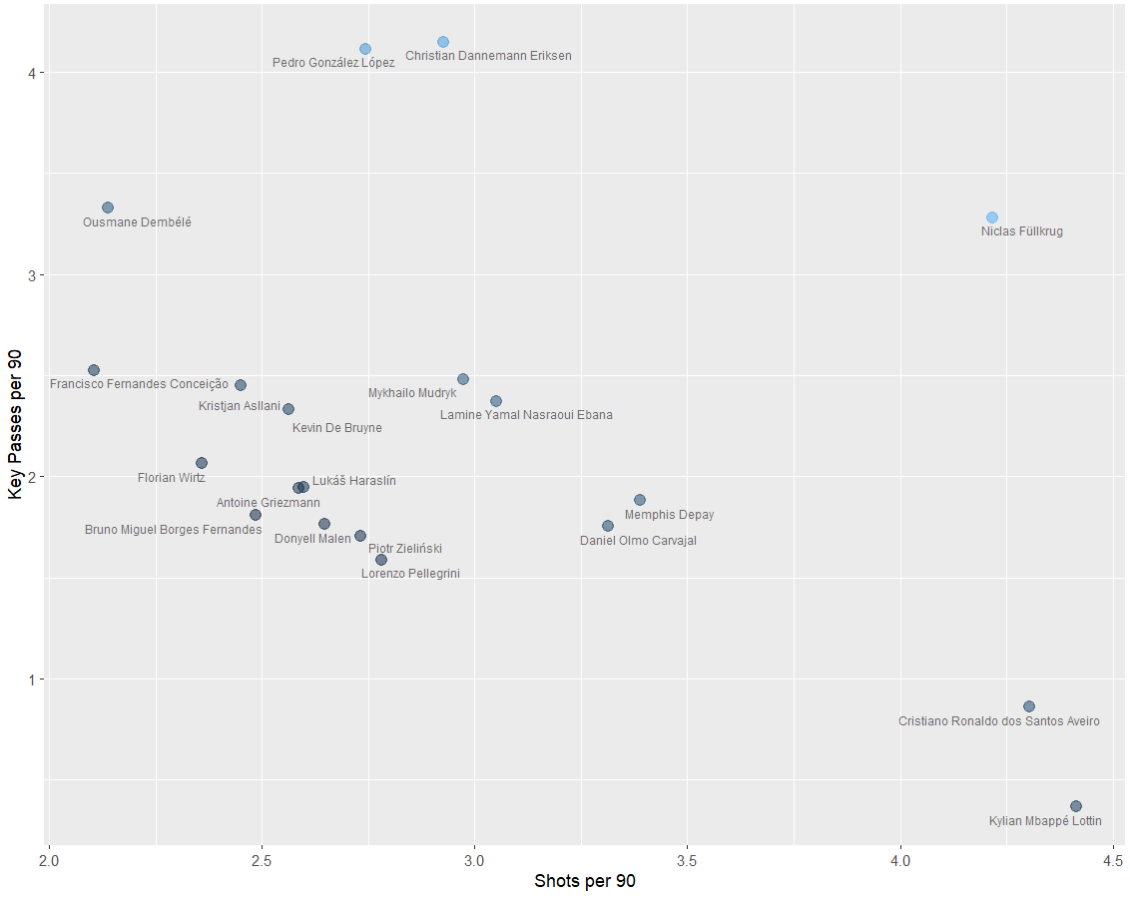
```
#Tiros y pases clave por jugador
player_shots_keypasses = StatsBombData %>%
  group_by(player.name, player.id) %>%
  summarise(shots = sum(type.name=="Shot", na.rm = TRUE),
            keypasses = sum(pass.shot_assist==TRUE, na.rm = TRUE))
#Minutos disputados por jugador
player_minutes = get.minutesplayed(StatsBombData)
player_minutes = player_minutes %>%
  group_by(player.id) %>%
  summarise(minutes = sum(MinutesPlayed))
player_shots_keypasses = left_join(player_shots_keypasses, player_minutes)
player_shots_keypasses = player_shots_keypasses %>% mutate(ninties = minutes/90)
player_shots_keypasses = player_shots_keypasses %>% mutate(shots_p90 = shots/ninties,  kp_p90 = keypasses/ninties, shots_kp_p90 = shots_p90 + kp_p90)
player_shots_keypasses = player_shots_keypasses %>% filter(minutes>180) %>%
  arrange(desc(shots_kp_p90)) %>% head(n=20)
```

Una vez disponemos de los datos en un único dataframe, pasaremos a crear el gráfico. En este caso se ha optado por un gráfico de dispersión, que permita observar la posición de cada jugador según sus tiros y sus pases clave por cada 90 minutos jugados. Cabe destacar que se instala el paquete de R *ggrepel*, que permite poner etiquetas en los gráficos sin que estas se interpongan entre sí, facilitando así la lectura.

El código utilizado y el gráfico resultante se muestran a continuación:

```
ggplot(player_shots_keypasses, aes(x=shots_p90, y = kp_p90, colour = shots_kp_p90, alpha = 90))+
  labs(x= "Shots per 90", y = "Key Passes per 90") +
  geom_point(size = 3, show.legend = FALSE) +
  geom_text_repel(data = player_shots_keypasses, aes(label = player.name), colour = "black", size = 2.5, vjust =0 )+
  guides(alpha = "none")
```

Gráfico 29: Top 20 jugadores según tiros y pases clave por partido Eurocopa 2024



Fuente: Elaboración propia

En este gráfico destaca en primer lugar la posición de Niclas Fülkrug, delantero de la selección alemana, que a pesar de no jugar de titular ningún partido durante el torneo fue un habitual revulsivo para el seleccionador alemán durante las segundas partes de los partidos. Este jugador cuenta con menos minutos que la mayoría de los otros que aparecen en esta lista, siendo estos titulares en sus selecciones, sin embargo, sus cifran

no dejan de ser llamativas. Se trata de un delantero puro, por lo que no sorprendería tanto que tuviera un buen registro respecto a los tiros intentados, pero también obtiene un sorprendente buen registro respecto a los pases clave.

No es de extrañar que en los mejores valores respecto a tiros intentados por partido encontremos a dos delanteros como Kylian Mbappé o Cristiano Ronaldo, ambos siendo la referencia ofensiva de sus respectivas selecciones, aunque quizás sorprende algo el bajo registro respecto a los pases claves del primero, al que se le conoce como un jugador algo más asociativo que a un Ronaldo que a sus 39 años, centra más su juego en la finalización en el área.

Por otro lado, encontramos a dos centrocampistas de carácter ofensivo como Pedri y Christian Eriksen, expertos en pases de carácter ofensivo, tanto durante el juego abierto como en jugadas a balón parado.

La aparición de Lamine Yamal en entre los mejores jugadores ofensivos de esta Eurocopa, un chico que cumplió 17 años un día antes de disputar la final del torneo, la cual acabaría ganando con la selección española siendo además importante en el equipo, sorprende por su gran precocidad y augura un futuro brillante para él en la selección.

Otra manera de medir el rendimiento ofensivo sería volviendo a utilizar los goles esperados, solo que esta vez también vamos a utilizar los goles esperados asistidos (xGA), esta medida se refiere al valor de goles esperados que dio un jugador directamente con sus pases, es decir, que su pase resultó en un tiro con ese valor de xG. Juntando estas dos métricas podemos sacar aquellos jugadores que más valor de goles esperados generaron, tanto con sus tiros como con sus pases. El valor xGA no se encuentra en el dataset, sin embargo, se puede extraer de los tiros realizados por otro jugador viendo de quien recibió el pase previo a ese tiro. Adicionalmente, se excluyen los xG generados por penaltis y una vez se tienen los valores xG y xGA por cada 90 minutos, se elige el top 15 de jugadores respecto a la suma de estas métricas. Una vez obtenemos el top, se vuelven a separar ambas métricas para diferenciarlas en la tabla, siendo posible visualizar en el gráfico la parte en la que contribuyen cada una.

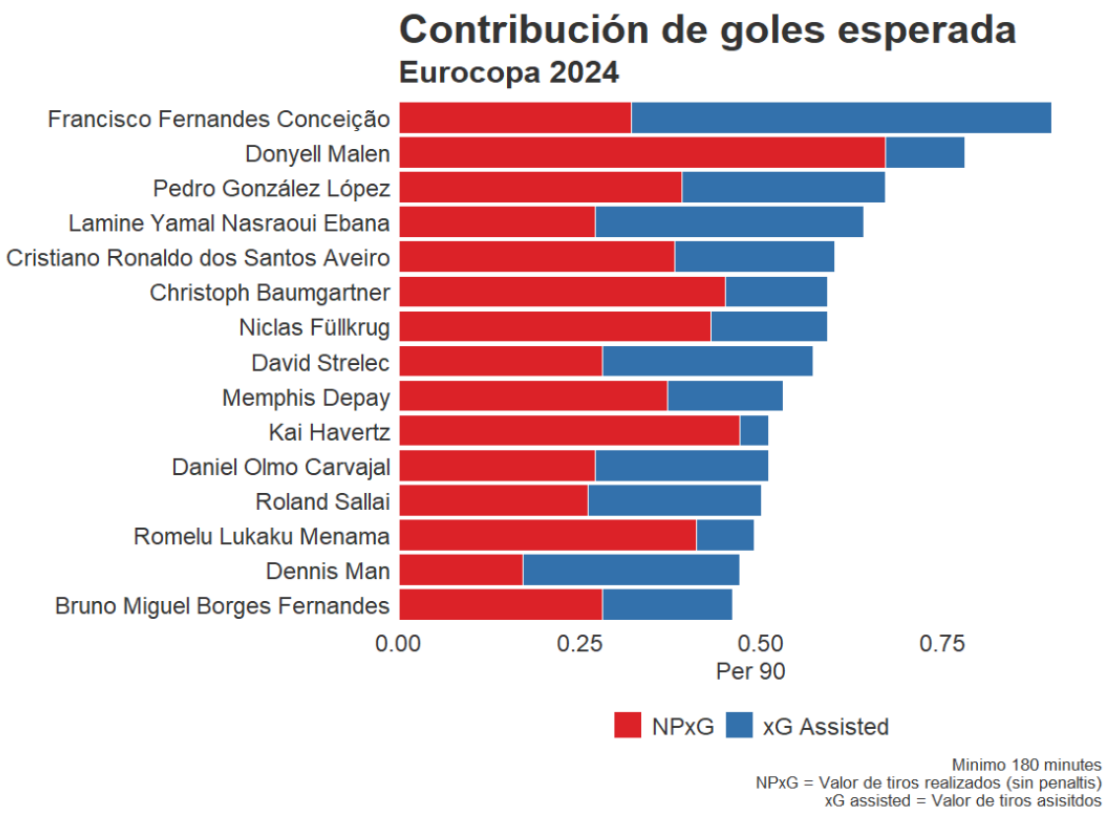
El código utilizado es el siguiente:

```
#Creación xGA
xGA = StatsBombData %>% filter(type.name=="Shot") %>%
  select(shot.key_pass_id, xGA = shot.statsbomb_xg)
shot_assists = left_join(StatsBombData, xGA, by = c("id" = "shot.key_pass_id")) %>%
  select(team.name, player.name, player.id, type.name, pass.shot_assist,
         pass.goal_assist, xGA ) %>%
 filter(pass.shot_assist==TRUE | pass.goal_assist==TRUE)
player_xGA = shot_assists %>%
  group_by(player.name, player.id, team.name) %>%
  summarise(xGA = sum(xGA, na.rm = TRUE))
#xG de cada jugador
player_xG = StatsBombData %>%  filter(type.name=="Shot") %>%
  filter(shot.type.name!="Penalty" | is.na(shot.type.name)) %>%
  group_by(player.name, player.id, team.name) %>%
  summarise(xG = sum(shot.statsbomb_xg, na.rm = TRUE)) %>%
  left_join(player_xGA) %>% mutate(xG_xGA = sum(xG+xGA, na.rm =TRUE) )
#Tabla xG + xGA
player_xG_xGA = left_join(player_xG, player_minutes) %>%
  mutate(nineties = minutes/90,
         xG_90 = round(xG/nineties, 2),
         xGA_90 = round(xGA/nineties,2),
         xG_xGA90 = round(xG_xGA/nineties,2) )
player_xG_xGA = player_xG_xGA %>%
filter(minutes>180)%>%arrange(desc(xG_xGA90))%>% head(n=15)
```

```
chart<-player_xG_xGA %>%
  select(1, 9:10)%>%
  pivot_longer(-player.name, names_to = "variable", values_to = "value") %>%
  filter(variable=="xG_90" | variable=="xGA_90")
```

La gráfica resultante es la siguiente:

Gráfico 30: Top 15 jugadores según contribución de goles esperada Eurocopa 2024



Fuente: Elaboración propia

Como se puede observar, la gran mayoría de nombre en este ranking se comparten con el anterior, ya que los tiros realizados son los que generan xG y los pases clave suelen ser los que más xGA generan. Aun así, resulta interesante esta visualización, ya que permite observar de manera clara la contribución de los tiros y los pases de cada jugador a su creación ofensiva.

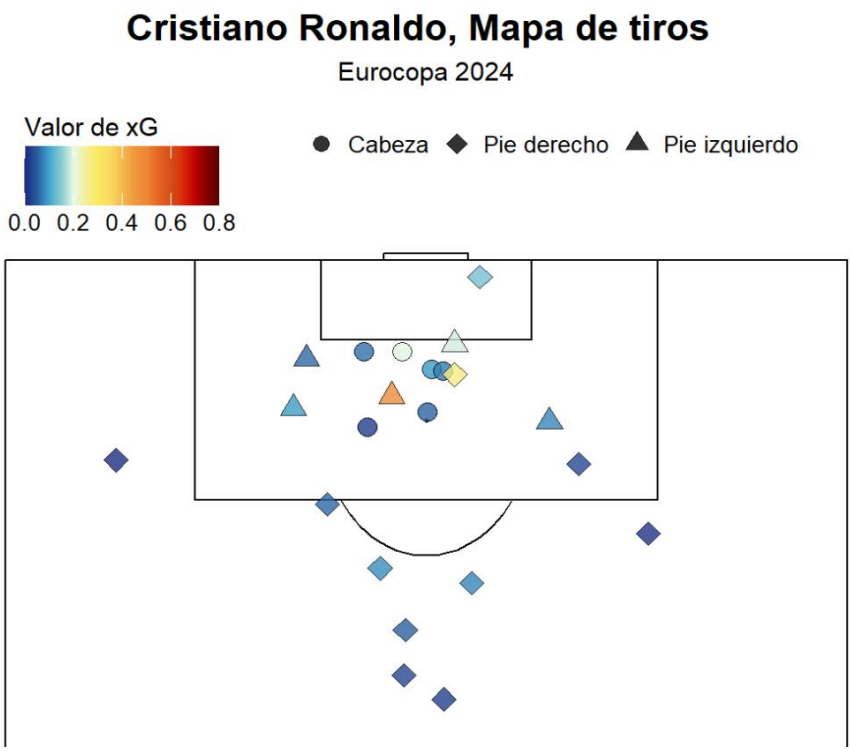
Otro gráfico que podemos utilizar para analizar el rendimiento ofensivo, especialmente de los delanteros, es lo que se conoce como un mapa de tiros. En él, se ven

representados en el campo todos los tiros intentados por un jugador, pudiendo adjuntar a ellos su valor de xG así como la pierna con los que fueron ejecutados o si estos fueron remates de cabeza. Gracias a este gráfico, se puede analizar el rendimiento de cara a puerta de un jugador, la zona desde donde suele rematar y otros aspectos relacionados con su tiro.

Para realizar este gráfico, primero almacenamos todos los tiros realizados por un jugador en un dataframe (excluimos los lanzamientos de penalti) y posteriormente creamos el gráfico correspondiente. En nuestro caso, vamos a analizar el mapa de tiros de Cristiano Ronaldo, uno de los jugadores que más tiros intentó por partido, pero que se fue de la competición sin conseguir anotar. Por otro lado, analizaremos el mapa de tiros de Jamal Musiala, media punta de la selección alemana que, pese a no aparecer en la tabla anterior, acabó el torneo como uno de los máximos goleadores con tres tantos. El código utilizado se encuentra en el Anexo 3 al tener una longitud considerable.

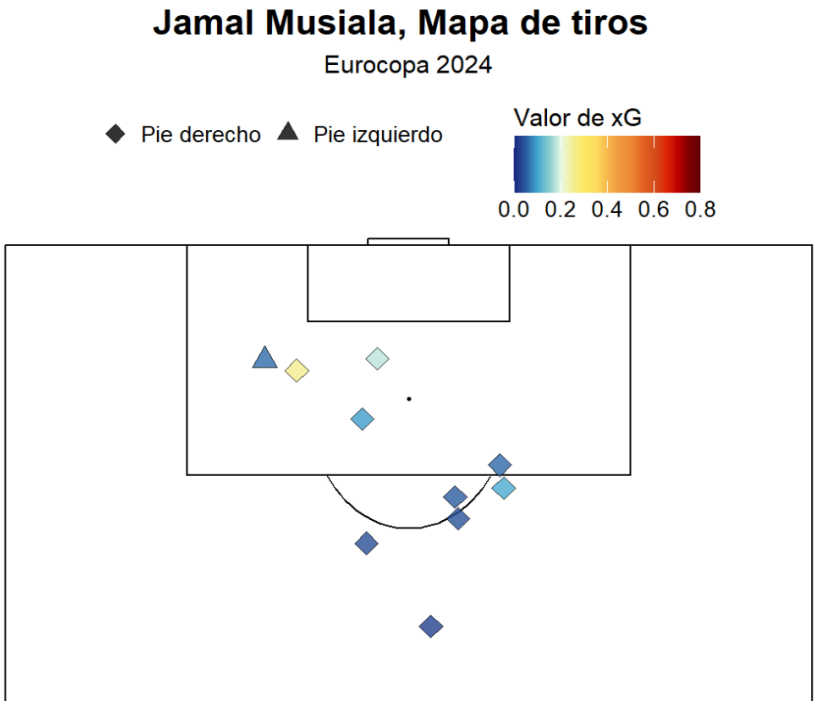
Los gráficos resultantes son los siguientes:

Gráfico 31: Mapa de tiros de Cristiano Ronaldo Eurocopa 2024



Fuente: Elaboración propia

Gráfico 32: Mapa de tiros de Jamal Musiala Eurocopa 2024



Fuente: Elaboración propia

A simple vista destaca la gran diferencia entre ambos gráficos, en el primero, observamos una gran variedad de disparos, ya sean con la diestra, la zurda o de cabeza, mientras que en el gráfico de Musiala la gran mayoría son con la diestra, sin tener este remates de cabeza. También destaca la diferencia entre el número de disparos intentados, destaca la gran cantidad de tiros lejanos intentados por Cristiano Ronaldo, con un bajo valor de xG, además de dos remates de alto valor alrededor de la zona del punto de penalti que no acabaron en gol. También destaca mucho la gran efecitvidad que tuvo el mediapunta alemán, al conseguir tres goles con tan solo diez remates durante el torneo.

Si bajamos algo más en el campo podemos centrarnos en el rendimiento de centrocampistas o defensas que, a pesar de no generar tanto a nivel ofensivo, siguen contando con una importancia vital dentro del equipo.

Por ejemplo, otros jugadores deben encargarse de hacerle llegar el balón a los jugadores de ataque, una métrica que puede ayudarnos a comprender quienes consiguieron suministrar con más balones a sus jugadores de ataque podría ser los pases realizados al último tercio del campo. Estos son aquellos pases con éxito que llegan al último tercio del campo (dirección hacia la que ataca el equipo) desde lugares más atrasados.

Para analizar esta métrica hay que tener en cuenta un par de cosas que no se habían mencionado hasta ahora. El dataset utilizado de StatsBomb contiene entre sus variables un sistema de coordenadas para guardad dónde ocurre cada evento registrado en él. En el dataset encontramos unas longitudes de 120 metros de largo (eje x) por 80 metros de ancho (eje y) para un campo de fútbol. Sabiendo esto, los pases al último tercio serán aquellos cuya coordenada x de origen sea inferior a los 80 metros y la de su destino sea igual o mayor. Adicionalmente, cabe mencionar que los pases completados, aquellos que llegan con éxito a su destino, se registran en la variable *pass.outcome.name* como valores nulos, pues son los más frecuentes y esto facilita al equipo de recolección de datos a la hora de registrarlos, por otro lado, los pases fallados si tienen valores en este atributo. Es por ello que a la hora de analizar esto, utilizaremos estos dos filtros para diferenciar los pases completados al último tercio, filtrando además de nuevo entre los jugadores que hayan disputado más de 180 minutos durante el torneo.

El código utilizado es el siguiente:

```
passes = StatsBombData %>%
  filter(type.name=="Pass" & is.na(pass.outcome.name)) %>%
  filter(location.x<80 & pass.end_location.x>=80) %>%
  group_by(player.name,player.id) %>%
  summarise(f3_passes = sum(type.name=="Pass"))
passes = left_join(passes, player_minutes)
passes = passes %>% mutate(ninties = minutes/90)
passes = passes %>% mutate(f3_passes_p90 = f3_passes/ninties)
passes = passes %>% filter(minutes>180) %>%
  arrange(desc(f3_passes_p90))
```

Como se puede observar en la última línea del código, ordenamos la tabla para que muestre a los jugadores según los valores más altos de pases al tercio final por cada 90 minutos de juego (almacenados en la columna *f3_passes_p90*).

La tabla resultante es la siguiente (se muestran sólo los primeros 15 resultados):

Tabla 9: Top 15 jugadores según sus pases al tercio final por partido Eurocopa 2024

	player.name	player.id	f3_passes	minutes	ninties	f3_passes_p90
1	Luka Modrić	5463	29	248.2356	2.758173	10.514206
2	Toni Kroos	5574	59	510.8775	5.676417	10.393880
3	Aymeric Laporte	4353	64	567.3548	6.303942	10.152377
4	Aurélien Djani Tchouaméni	10481	55	502.3875	5.582084	9.852951
5	Florian Grillitsch	11396	21	202.9587	2.255096	9.312240
6	Antonio Rüdiger	3167	51	524.4995	5.827772	8.751201
7	Milos Veljkovic	6321	27	289.8523	3.220581	8.383580
8	Granit Xhaka	3500	48	518.9395	5.765994	8.324670
9	Mateo Kovačić	5456	20	238.4778	2.649753	7.547873
10	Marcelo Brozović	5469	20	244.7323	2.719248	7.354976
11	Robin Aime Robert Le Normand	22128	38	473.4338	5.260375	7.223819
12	Alessandro Bastoni	7480	31	391.8113	4.353459	7.120775
13	Arthur Theate	43913	21	267.4563	2.971736	7.066576
14	Nuno Mendes	41092	35	450.1094	5.001216	6.998299
15	Fabián Ruiz Peña	6655	45	579.2927	6.436586	6.991284

Fuente: Elaboración propia

Si observamos la lista, encontramos en ella como era de esperar en gran medida a centrocampistas de diferentes selecciones. No sorprende demasiado encontrar encabezando la lista a la dupla legendaria de centrocampistas, compañeros en el Real Madrid, aunque de distintas selecciones, formada por Luka Modric (Croacia) y Toni Kroos. Puede sorprender algo más la presencia de Aymeric Laporte, defensa central francés pero nacionalizado español y que fue una parte indispensable para que la selección española se llevara el trofeo. Su presencia en el podio de esta métrica se puede explicar por la alta línea de presión practicada por su selección, además de su gran capacidad de conducción a la hora de sacar el balón jugado desde atrás, lo que le permitió llegar en numerosas ocasiones hasta la línea de medio campo y realizar pases al último tercio.

Para confirmar esto vamos a realizar en esta ocasión un gráfico distinto de los utilizados hasta ahora, que nos permita visualizar el origen y destino de estos pases dentro del propio campo de fútbol. Para ello, haremos uso de una nueva librería, desarrollada por el usuario de GitHub “FCrStats” llamada *SBpitchs,* con la intención de facilitar la

visualización de estadísticas de fútbol, permitiendo crear una gráfica en forma de campo de manera mucho más sencilla a la utilizada anteriormente.

En primer lugar, crearemos un nuevo dataframe que almacene todos los pases que cumplan los requisitos establecidos de un jugador específico. Posteriormente, crearemos el gráfico del campo de fútbol y meteremos en él los pases completados.

El código para la creación del dataframe y el gráfico correspondiente se muestra a continuación:

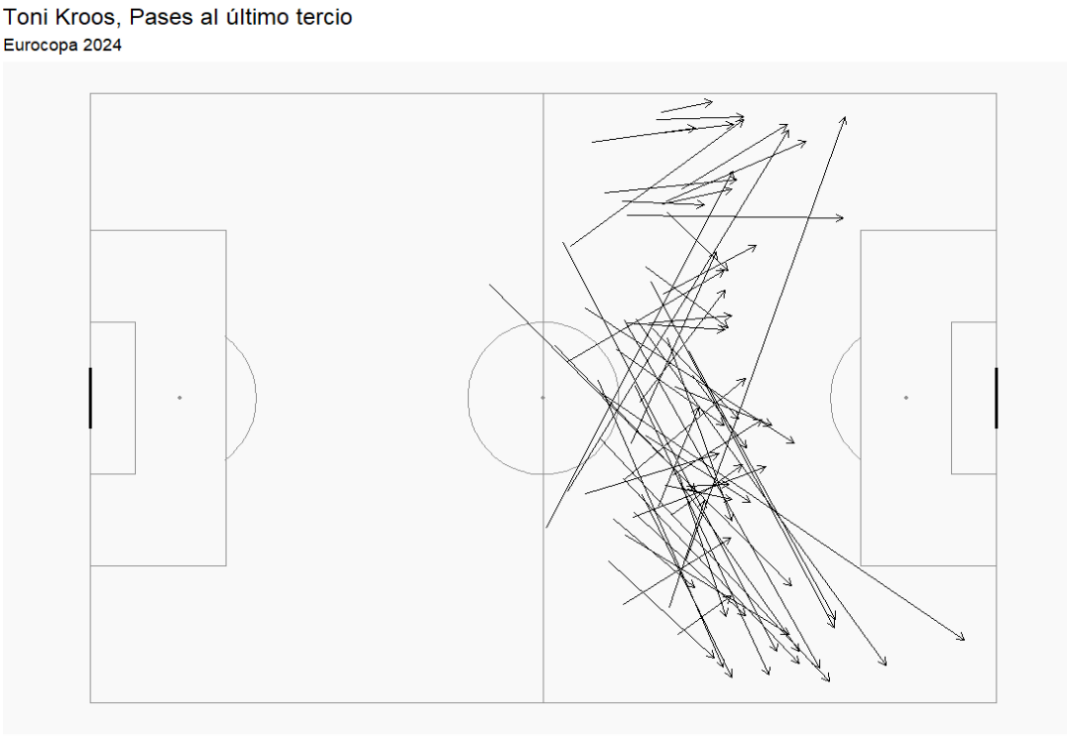
```
#Creación del dataframe

player_passes = StatsBombData %>%

 filter(type.name=="Pass" & is.na(pass.outcome.name) & player.name=="NOMBRE DEL JUGADOR") %>%

 filter(location.x<80 & pass.end_location.x>=80)

#Creación del gráfico

create_Pitch() +

 geom_segment(data = player_passes, aes(x=location.x, y=location.y, xend = pass.end_location.x, yend = pass.end_location.y),

          lineend = "round", linewidth = 0.5, colour = "black", arrow = arrow(length = unit(0.07, "inches"),ends = "last", type = "open"))+

 labs(title = "NOMBRE DEL JUGADOR, Pases al último tercio", subtitle = "Eurocopa 2024") +

 scale_y_reverse() +

 coord_fixed(ratio = 105/100)
```

Tocará sustituir en este código el *player.name* del jugador que queramos analizar.

En este caso vamos a elegir al segundo y al tercero del ranking anterior, Toni Kroos y Aymeric Laporte, ya que la selección croata de Luka Modric cayó eliminada en fase de grupos. En los gráficos de pases al último tercio vamos a observar si se encuentran diferencias entre ellos, al jugar en posiciones del campo claramente distintas.

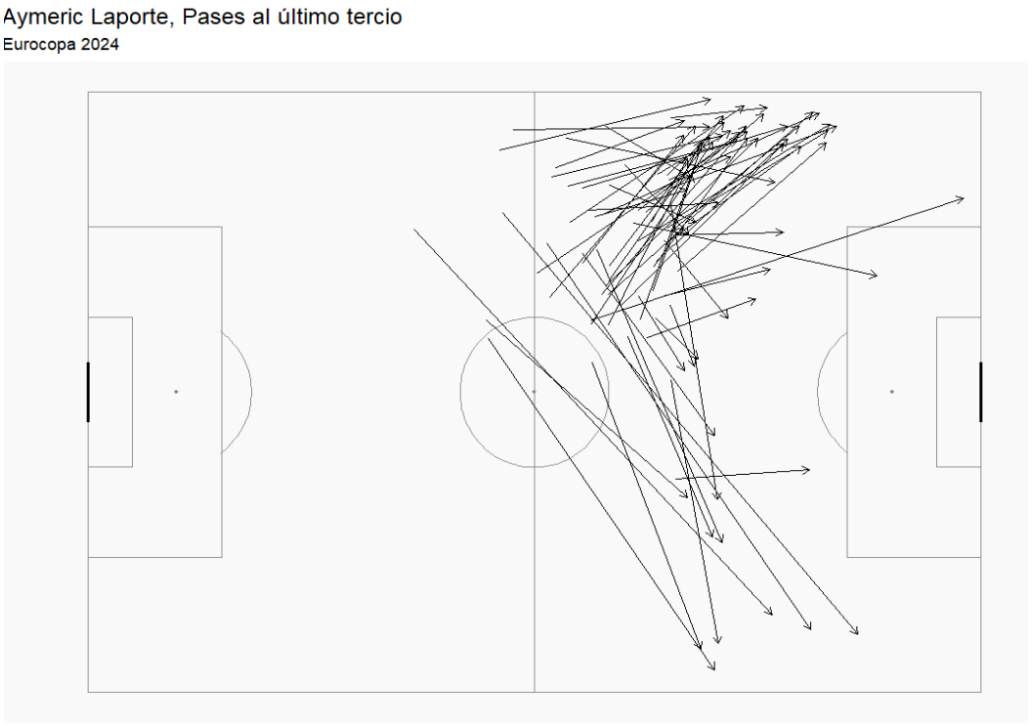
Los gráficos se muestran a continuación:

Gráfico 33: Pases al último tercio de Toni Kroos Eurocopa 2024



Fuente: Elaboración propia

Gráfico 34:Pases al último tercio de Aymeric Laporte Eurocopa 2024



Fuente: Elaboración propia

A primera vista se observan claras diferencias entre los mapas de pases de ambos jugadores. Por un lado, Kroos juega en una posición mucho más central, repartiendo pases tanto a las bandas como a la zona central con una frecuencia similar, con una ligera mayor presencia en el lado derecho. Si observamos el mapa de pases de Laporte, observamos una clara tendencia hacia el lado izquierdo, ya que ejerce en la posición de defensa central izquierdo, aunque llama la atención lo pegado a la banda que se encuentra en muchos de sus pases. El central de la selección española realiza pocos pases a la zona central del campo, siendo estos pases de mayor riesgo que un defensa central no toma con tanta frecuencia, al ser él la última línea de jugadores y un fallo puede causar un contraataque peligroso del rival. En general, observamos que el central también realiza más cambios de banda que el centrocampista alemán, abriendo el juego con balones en largo a la banda contraria, saltándose el pase fácil y buscando mover el balón más rápido con un desplazamiento directo al lado derecho del campo.

Gracias a estas gráficas, podemos observar cómo se comportan jugadores en distintas posiciones de distintos esquemas tácticos. Estos resultan de gran ayuda a la hora de analizar la posesión de distintos equipos, y como estos mueven el balón para generar peligro. Cabe destacar que estos gráficos no se limitan a los pases al último tercio del campo, y que podrían usarse para analizar más en profundidad el juego de un equipo, analizando por ejemplo como salen jugando desde atrás o la distribución de los porteros a la hora de sacar de portería.

Otra forma de analizar la influencia de los distintos jugadores en un sistema de juego sería observar la red de pases de una selección durante un partido específico o un torneo. Este gráfico representa la posición media en la que los jugadores dieron y recibieron pases durante el partido, adicionalmente, crea pares entre jugadores para contabilizar el número de pases que dos jugadores se dieron entre ellos. Para ello, se va a hacer uso de el código creado para el usuario de GitHub “pavibear” en su repositorio de “football-analytics”, en concreto de la función *passing_network*. Cabe destacar que el gráfico ignora los pases fallados, así como a aquellos jugadores que no dieron pases exitosos durante el partido (normalmente suplentes que no juegan todo el partido). Los jugadores vienen representados en el gráfico por su dorsal.

Para este caso, vamos a analizar la red de pases de la final del torneo, que enfrentó a España e Inglaterra, dos selecciones que habían tenido grandes diferencias respecto a su estilo de juego durante el transcurso de la competición. Para ello, encontraremos en nuestro dataset el *match_id* de la final y lo sustituiremos en el código.

Para poner en contexto las alineaciones iniciales, así como los cambios, se presentan estas a continuación:

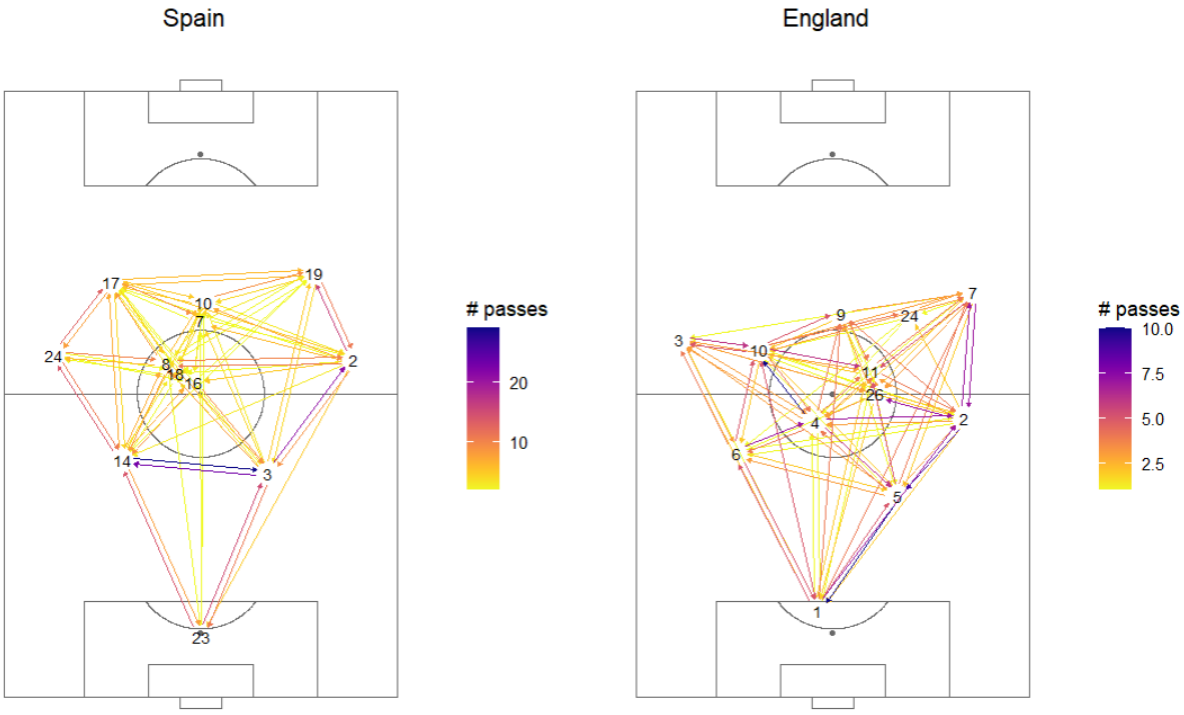
Gráfico 35: Alineaciones final Eurocopa 2024

Fuente: UEFA.com

Estas alineaciones presentan una forma sobre el papel, pero el gráfico de la red de pases nos permitirá visualizar cómo los jugadores realmente interactuaron entre sí durante el transcurso del encuentro.

Veamos ahora la red de pases de ambas selecciones:

Gráfico 36: Red de pases final de la Eurocopa 2024



Fuente: Elaboración propia

En primer lugar, respecto a la alineación, se observa en el combinado español una formación mucho más simétrica, con la línea de cuatro defensas, los dos pivotes titulares (más Martín Zubimendi, dorsal 18, que sustituyó a Rodri, dorsal 16, al descanso por lesión), el media punta y el delantero en zonas muy parejas y los dos extremos en las bandas como los jugadores más adelantados. Por su parte Inglaterra presenta una alineación asimétrica, empezando por la línea defensiva, donde su central derecho en la alineación de la UEFA actúa casi como lateral y su carrilero derecho (Bukayo Saka, dorsal 7) actúa como un extremo puro, siendo el jugador más adelantado del equipo, por otro lado, destaca la gran altura de su lateral izquierdo, cubriendo una banda que no presenta extremo, ya que Jude Bellingham (dorsal 10) actuó por zonas más centrales. Se puede observar también que Declan Rice (dorsal 4) actuó como único pivote defensivo, con una posición mucho más retrasada que el resto de centrocampistas, mientras que la posición cercana de Phil Foden (dorsal 11) y Mainoo (dorsal 26) llama la atención ya que al primer se le presupone un jugador mucho más ofensivo que el segundo.

Por otro lado, hay que destacar la gran diferencia entre el número de pases efectuado por cada equipo, si nos fijamos en la leyenda de cada una observamos que el mayor número de pases entre dos jugadores (los dos defensas centrales) fue casi tres veces mayor en el caso de la selección española. El mapa de pases de Inglaterra es bastante más caótico que el de España, empezando por la distribución del portero, donde el inglés distribuyó pases a un gran número de compañeros, mientras que el portero español principalmente jugó en corto con sus centrales. Se puede observar que la selección española distribuyó su juego de forma equitativa tanto por las bandas como por la zona central, mientras que la selección inglesa se centró principalmente en mover el balón por la banda derecha.

Para acabar con este análisis sobre estadísticas de jugadores se va a poner el foco ahora en estadísticas defensivas. Para ello, vamos a hacer uso de cuatro métricas: recuperaciones de balón, intercepciones, despejes y blocajes.

A la hora de evaluar estas estadísticas hay que tener en cuenta que estas varían mucho según la posición y el estilo de juego que elija el entrenador para su selección. Por ejemplo, un equipo que ejerza una presión muy alta, si esta es exitosa, tendrá un mayor número de recuperaciones de sus jugadores ofensivos, o un equipo que mantenga un bloque bajo y espere al rival, tendrá más despejes y disparos bloqueados por sus defensores.

Es por eso que a la hora de evaluar estas estadísticas se va a utilizar un sistema de percentiles, que compare a un jugador seleccionado con todos los jugadores de su misma posición que hayan disputado al menos 90 minutos durante el torneo. Para ello, en primer lugar, habrá que simplificar las posiciones que se tienen en el dataset, ya que en la variable *player.position* encontramos 25 valores diferentes, y por ejemplo encontramos distinciones entre defensas centrales que hayan jugado en el lado izquierdo, en el centro de una línea de tres o en la derecha, mientras que si queremos ver los percentiles de un central lo querremos comparar con todos los centrales del torneo, sin importar donde jugaran estos dentro del esquema táctico.

Para esto, vamos a realizar un mapeo de las posiciones para dividirlas en cinco valores más simples: porteros, centrales, laterales, centrocampistas y delanteros. Después obtendremos un dataframe con las estadísticas defensivas mencionadas anteriormente y las posiciones de los jugadores. A este se le unirá el dataframe creado en fases del análisis anteriores que contiene los minutos disputados por cada jugador para obtener

las estadísticas por cada 90 minutos jugados y eliminar la diferencia que crean en los resultados los efectos de disputar un distinto número de partidos.

El código utilizado para esto es el siguiente:

```
#Mapeo posiciones

position_map <- c(

  "Goalkeeper" = "Goalkeeper",

  "Center Forward" = "Forward", "Left Back" = "Full Back", "Right Back" = "Full Back", "Center Back" = "Center Back",

  "Left Defensive Midfield" = "Midfield", "Left Wing" = "Forward", "Right Wing" = "Forward", "Left Center Back" = "Center Back",

  "Right Defensive Midfield" = "Midfield", "Right Center Back" = "Center Back", "Right Wing Back" = "Full Back",

  "Center Attacking Midfield" = "Midfield", "Left Wing Back" = "Full Back", "Left Center Forward" = "Forward",

  "Left Center Midfield" = "Midfield", "Center Defensive Midfield" = "Midfield", "Right Center Midfield" = "Midfield",

  "Right Center Forward" = "Forward", "Left Attacking Midfield" = "Midfield", "Right Attacking Midfield" = "Midfield",

  "Left Midfield" = "Midfield", "Right Midfield" = "Midfield", "Substitute" = "Substitute")


StatsBombData = StatsBombData %>% mutate(position_simple = recode(position.name, !!!position_map))

#DEFENSIVAS

defensive = StatsBombData %>%

  filter(type.name %in% c("Ball Recovery", "Interception", "Clearance", "Block")) %>%

  group_by(player.name, player.id, position_simple) %>%

  summarise(
```

```
    recoveries = sum(type.name == "Ball Recovery", na.rm = TRUE),
    interceptions = sum(type.name == "Interception", na.rm = TRUE),
    clearances = sum(type.name == "Clearance", na.rm = TRUE),
    blocks = sum(type.name == "Block", na.rm = TRUE))
defensive_p90 = left_join(defensive, player_minutes)
defensive_p90 = defensive_p90 %>% filter(minutes>90)
defensive_p90 = defensive_p90 %>% mutate(ninties = minutes/90)
defensive_p90 = defensive_p90 %>% mutate(recoveries_p90 = recoveries/ninties,
                                         interceptions_p90 = interceptions/ninties,
                                         clearances_p90 = clearances/ninties,
                                         blocks_p90 = blocks/ninties)
```

Una vez tenemos todo esto preparado es hora de crear la función. Los argumentos de esta serán el nombre del jugador elegido, el dataset donde se encuentren los datos que se van a comparar y un array con las columnas de las que se quiere obtener el percentil. Cabe destacar que en este caso sólo se va a usar esta función para encontrar los percentiles de las estadísticas mencionadas anteriormente, pero esta se podría utilizar para encontrar otros percentiles si se prepararan dataset con los datos agrupados por jugador y las columnas que se buscan se indicaran en el tercer argumento de la función.

El código de la función es el siguiente y se añade también un ejemplo de uso:

```
percentiles_defensivos <- function(player_name, dataset, stats_columns) {
  # Verificar que el jugador esté en el dataset
  if (!(player_name %in% defensive_p90$player.name)) {
    stop("El jugador no se encuentra en el dataset.") }
  # Obtener la posición del jugador
  player_position <- dataset %>%
```

```
  filter(player.name == player_name) %>%
  select(position_simple) %>%
  pull()
# Filtrar el dataset para incluir solo jugadores en la misma posición
same_position_data <- dataset %>%
  filter(position_simple == player_position)
# Filtrar los datos del jugador específico
player_data <- defensive_p90 %>%
  filter(player.name == player_name)
# Calcular percentiles para cada estadística
percentiles <- sapply(stats_columns, function(stat) {
  player_value <- player_data[[stat]]
  all_values <- dataset[[stat]]
  # Calcular el percentil del jugador
  percentile <- ecdf(all_values)(player_value) * 100
  return(percentile) })
# Crear un dataframe para mostrar los resultados
result <- data.frame(
  Statistic = stats_columns,
  Percentile = percentiles )
return(result) }
# Ejemplo de uso
stats_columns <- c("recoveries_p90", "clearances_p90", "interceptions_p90", "blocks_p90")
```

player_percentiles <- percentiles_defensivos(“Marc Cucurella Saseta", defensive_p90, stats_columns)

En el ejemplo de uso encontramos al lateral de la selección española Marc Cucurella, nombrado en el mejor once del torneo por la UEFA tras un torneo brillante a la vez que sorprendente para muchos.

Si observamos sus percentiles encontramos los siguientes resultados:

Tabla 10: Percentiles defensivos Marc Cucurella Eurocopa 2024

	Statistic	Percentile
recoveries_p90	recoveries_p90	42.85714
clearances_p90	clearances_p90	61.47186
interceptions_p90	interceptions_p90	95.23810
blocks_p90	blocks_p90	95.23810

Fuente: Elaboración propia

Como se puede ver, no obtiene malos resultados en ninguna estadística, siendo la más baja las recuperaciones, que teniendo en cuenta el contexto donde jugaba, con un centro del campo en la selección española que recuperaba la gran parte de los balones, no es tan mala siendo comparado con los otros laterales del torneo. Sin embargo, defensivamente hablando se ve su gran papel a la hora de interceptar pases y bloquear disparos, superando al 95% de los jugadores del torneo con los que comparte posición (se recuerda que no se distingue entre laterales izquierdos y derechos). Estos percentiles sumados a su aporte ofensivo son los que llevaron a Cucurella a aparecer en el mejor once del torneo.

Para acabar con este punto de análisis, vamos a ponerle un poco de atención a los porteros. Muchas veces es difícil juzgar esta posición, pero es sin duda una de las más importantes en el mundo del futbol. Para analizarla en este caso vamos a elegir el número de paradas por partido de cada portero, esta métrica es buena para saber lo bueno que es un portero parando, aunque suele favorecer a aquellos porteros de selecciones inferiores que reciben muchos más disparos por partido que los de selecciones de mayor nivel, a las que les suelen generar menos ocasiones. Es por eso que, para acompañarla, también se va a incluir el porcentaje de paradas sobre el total de disparos a puerta recibidos (los disparos que no van a portería no requieren ser parados), para observar a aquellos porteros que recibieron pocas ocasiones, pero las resolvieron con solvencia. Se tendrán en cuenta solo los porteros que disputaron más

de 180 minutos, equivalentes a dos partidos, para tener en cuenta solamente a los porteros titulares de cada selección.

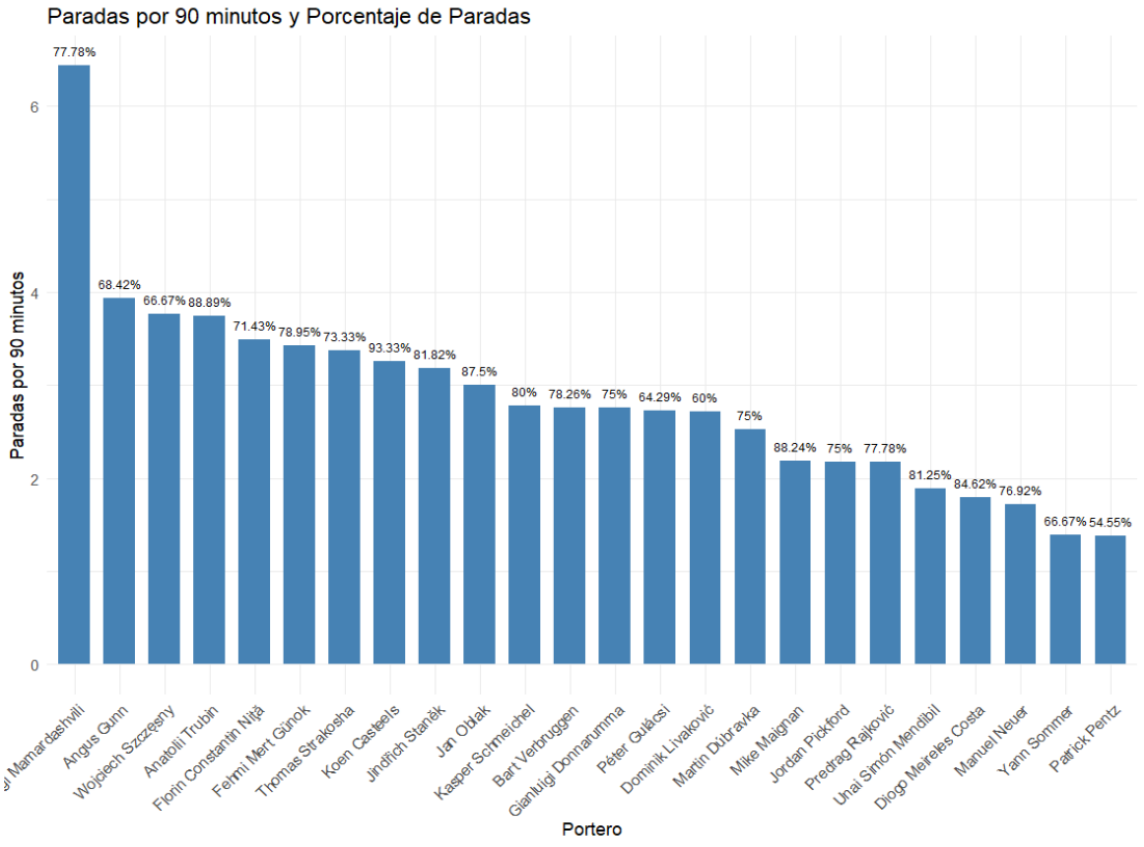
Las instrucciones para la creación del dataframe y el correspondiente gráfico se muestran a continuación:

```
#Dataframe porteros
goalkeepers = StatsBombData %>%
  filter(goalkeeper.type.name %in% c("Shot Saved", "Goal Conceded")) %>%
  group_by(player.name, player.id) %>%
  summarise(saves = sum(goalkeeper.type.name == "Shot Saved"),
            total = saves + sum(goalkeeper.type.name == "Goal Conceded"))
goalkeeper_stats = left_join(goalkeepers, player_minutes)
goalkeeper_stats = goalkeeper_stats %>% filter(minutes>180)
goalkeeper_stats = goalkeeper_stats %>% mutate(ninties = minutes/90)
goalkeeper_stats = goalkeeper_stats %>% mutate(saves_p90 = saves/ninties,
                                               save_percentage = (saves/total)*100)
#Creacion gráfico
ggplot(goalkeeper_stats, aes(x = reorder(player.name, -saves_p90), y = saves_p90)) +
  geom_bar(stat = "identity", fill = "steelblue", width = 0.7) +
  geom_text(aes(label = paste0(round(save_percentage, 2), "%")),
            vjust = -1, size = 2.5, color = "black") +
  labs(
    title = "Paradas por 90 minutos y Porcentaje de Paradas",
    x = "Portero",
    y = "Paradas por 90 minutos"
  ) +
```

```
theme_minimal() +

theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

El gráfico resultante es el siguiente:

Gráfico 37: Paradas por 90 minutos y porcentaje de paradas Eurocopa 2024



Fuente: Elaboración propia

En primer lugar, destaca el altísimo valor de paradas por partido del portero de la selección georgiana Giorgi Mamardashvili, el cual acumula más de dos paradas adicionales por partido al siguiente portero respecto a este apartado. Esto podría atribuirse a la gran cantidad de ocasiones que recibió la selección de Georgia, pero además, su porcentaje de paradas no es para nada malo comparado con el resto de porteros, especialmente aquellos que ocupan las primeras posiciones respecto a paradas por partidos, lo que certifica el gran torneo que disputo Mamardashvili. Otros porteros como el escocés Gunn o el polaco Szczesny tuvieron un gran número de paradas por partido, sin embargo, su porcentaje de paradas no fue tan alto, encajando estos muchos goles.

Respecto al porcentaje de paradas, destacan el ucraniano Trubin, que no partió como titular durante el primer partido del torneo, pero luego le quitó el puesto a su compañero Lunin en los dos siguientes partidos, siendo esta una decisión acertada según los datos. También destaca en primer lugar en este apartado el portero belga Koen Casteels, sustituto de un lesionado Thibaut Courtois y que a pesar de que su selección no realizara un gran torneo, tiene gran parte de la culpa de que esta llegara a donde llegó.

Por último, hay que destacar a Mike Maignan, portero de la selección de Francia que no necesitó realizar un gran número de paradas por partido gracias a la sólida defensa francesa, pero que mantuvo un alto porcentaje de paradas cuando se le necesitó. Esto le llevó a ser seleccionado como el mejor portero del torneo y entrar en el once ideal, por delante de los porteros finalistas Jordan Pickford y Unai Simón, a los cuales supera en ambas estadísticas.

Con esta comparación de porteros se va a dar por finalizada esta etapa de análisis de la Eurocopa 2024, habiendo repasado tanto el desempeño general de las distintas selecciones, así como el rendimiento individual de los jugadores de las distintas zonas del campo.

5. Conclusiones

Durante la realización de este trabajo se han ido buscando cumplir los distintos objetivos marcados al principio de este. Entrando en la recta final de este, se puede afirmar que gracias a realizar un análisis general del fútbol de selecciones y uno más específico sobre la Eurocopa 2024, se han podido aplicar diferentes conocimientos adquiridos durante el curso del grado, destacando la estadística, la programación, el tratamiento de datos o la creación de diferentes métodos de visualización.

A medida que se ha trabajado con los datos, profundizando cada vez más en estos, se han ido aprendiendo nuevos aspectos del programa utilizado, así como nuevas funcionalidades de distintas librerías instaladas para poder llevar a cabo satisfactoriamente el desarrollo del trabajo.

Gracias al apartado del contexto tecnológico se ha podido observar la influencia del área del Big Data en el mundo del deporte en general, y como el fútbol es uno de los deportes que más fuerte está apostando por una evolución en conjunto al análisis de datos, atrayendo esta combinación cada vez a más personas. El uso de estas herramientas permite a los equipos una preparación cada vez mejor, tanto dentro como fuera del campo.

Tras la obtención de los datos, ha sido de vital importancia someter a estos a unas etapas de limpieza y preprocesado, especialmente en el caso del análisis histórico del futbol de selecciones, al provenir la fuente de datos de un único usuario no profesional. Eliminar las selecciones no afiliadas a confederaciones de la FIFA, así como poder crear datasets que excluyeran partidos de menor importancia como los amistosos, o elegir dentro de la gran base de datos de StatsBomb únicamente aquellos de la competición que interesaba estudiar ha resultado vital para el correcto desarrollo del trabajo.

Dentro del análisis histórico del fútbol de selecciones, se han podido extraer conclusiones y tendencias interesantes para el futuro. Comenzando por el gran desarrollo que experimentó el fútbol de selecciones durante el siglo pasado, y que se llevó a cabo a distintos ritmos en cada parte del planeta, empezando a disputar partidos primero selecciones europeas y americanas, con un fuerte desarrollo entre las décadas de los cuarenta a los setenta del futbol africano y asiático. Se ha visto el efecto del factor campo y su ligera disminución desde la década de los noventa o la tendencia

creciente respecto al número de partidos disputados, que preocupa especialmente a jugadores y técnicos.

Otra observación interesante es la disminución del promedio de goles por partido, que durante el siglo pasado vio una gran caída, alcanzando su mínimo en la década de los ochenta, para estabilizarse en valores algo superiores a los de esta hasta el día de hoy. Se han observado las diferencias entre los goles anotados y recibidos según las confederaciones, así como la distribución por minutos de los goles, tanto en el tiempo regular como en la prórroga, así como la evolución del minuto medio de goles anotados por década, pudiendo resultar todas estas observaciones útiles para cuerpos técnicos de diferentes equipos.

También se ha analizado el rendimiento entre confederaciones, obteniendo resultados interesantes, como el dominio de las selecciones de la CONMEBOL, que obtienen los mejores resultados frente al resto, incluyendo el sorprendente dato de que nunca una selección de la OFC ganó un partido no amistoso a alguna selección de esta confederación. Por último, se llevó a cabo la creación de una función que permite obtener el resumen de los resultados entre dos selecciones, permitiendo así una comparación rápida entre ellas.

Tras esto se cambió el foco de atención a la Eurocopa, uno de los torneos más prestigiosos de selecciones y organizado por la UEFA para que compitan entre ellas sus selecciones afiliadas. Se aprovecharon los datos históricos para hacer un breve resumen de este torneo y sus tendencias como paso introductorio antes del siguiente apartado. Se observaron los distintos formatos por los que ha pasado la competición, viendo la diferencia entre partidos disputados en cada edición, así como sus promedio de goles por partido, observando que este se ha mantenido relativamente estable tras el primer cambio de formato de la competición. Para terminar con este apartado se hizo un estudio sobre la evolución de la competitividad de este torneo, basándose en la diferencia de goles por partido de cada edición, donde se observó que se trata de un torneo bastante competitivo, con una diferencia de goles baja entre equipos, y que la edición a estudiar de 2024 fue la más competitiva en este aspecto de lo que va de siglo. Esto también se vio representado en la tabla de máximos goleadores de cada edición, donde este año hasta siete jugadores compartieron este galardón, el máximo número de jugadores de la historia de la competición.

Una vez terminado el análisis histórico se cambiaron los datos con los que se trabajaban para buscar estadísticas más avanzadas y específicas. Se recurrió a la base de datos

pública de StatsBomb, donde se filtraron los datos para buscar los partidos y los eventos registrados de la Eurocopa 2024.

En la introducción a este torneo se explicaron las selecciones favoritas a ganarlo antes del comienzo de este, y su relación directa con el valor de sus plantillas. También se observó la desigualdad que se encontró una vez finalizada la fase de grupos del torneo en ambos lados del cuadro de las fases finales, con cuatro de las cinco favoritas clasificadas en el mismo lado del cuadro.

Gracias a disponer de una base de datos mucho más completa y detallada, aunque más difícil de manejar que la anterior, se pudo poner el foco en distintas métricas avanzadas del fútbol para observar el rendimiento, tanto de selecciones como de jugadores, que pudiera explicar el transcurso del torneo.

Se analizaron estadísticas tanto ofensivas como defensivas de las selecciones, pudiendo explicar, por ejemplo, la eliminación de Croacia, una selección que creó oportunidades, pero no supo materializarlas, lo que sumado a su fragilidad defensiva hizo que cayeran eliminadas en la primera fase del torneo. Por otro lado, este análisis permitió encontrar sorpresas como la de la selección georgiana, la cual a pesar de contar con registros muy pobres tanto a nivel ofensivo, donde sólo les salvó su efectividad, como a nivel defensivo, fue capaz de clasificarse tras la fase de grupos en la que era su primera participación en el torneo.

Por último, se centró la atención en el rendimiento individual de algunos de los jugadores, intentando cubrir estadísticas representativas para las distintas posiciones del campo. Fue en este punto especialmente donde se demostró la gran cantidad de maneras que hay para visualizar datos futbolísticos, haciendo uso de distintos tipos de gráficos como mapas de pases o de tiros, o la red de pases de un partido, unidos a gráficos comunes a otras áreas como gráficos de barras o de dispersión, para extraer conclusiones útiles que permitan mejorar el rendimiento en el futuro. Se desarrolló en este apartado también una función que permite obtener los percentiles de un jugador frente a aquellos que actuaron en una posición similar a la suya, pudiendo elegir las estadísticas a observar, siendo el ejemplo utilizado las estadísticas defensivas de los laterales de la Eurocopa.

Todos los análisis realizados pueden resultar de gran interés tanto a aficionados como a los cuerpos técnicos y analistas de selecciones o clubes, que vean en el Big Data una oportunidad de mejorar su rendimiento, maximizando sus fuerzas e intentando mejorar sus debilidades.

6. Relación del trabajo con los estudios cursados

Durante la realización de este trabajo se han puesto en práctica una parte de los conocimientos adquiridos en el transcurso de la carrera y en particular de la rama de Sistemas de la Información.

El análisis de datos es una parte fundamental de la informática en el contexto actual, donde el dato como unidad de información ha ganado una importancia vital dentro de la sociedad. Durante el transcurso del trabajo se han utilizado herramientas como RStudio, que dan soporte a análisis estadísticos avanzados como el que se ha llevado a cabo en este trabajo sobre el fútbol internacional y la Eurocopa 2024.

El uso de esta herramienta ha permitido poner en práctica una parte de los conocimientos aprendidos durante la carrera, especialmente en áreas de estadística, bases de datos, programación y tratamientos de la información.

Este trabajo se enfoca principalmente en la rama de Sistemas de la Información, que describe cómo manejar, procesar y presentar la información para dar soporte a una toma de decisiones en base a unos resultados.

Este trabajo puede servir como introducción al mundo de la analítica deportiva, destacando algunas de las métricas clave que se pueden utilizar, en este caso en el deporte del fútbol, para poder observar las fortalezas y debilidades de distintos equipos y jugadores y poder trabajar en maximizar su rendimiento.

Durante las distintas etapas de análisis del trabajo se ha buscado transformar los datos en las visualizaciones más efectivas a la hora de comunicar las conclusiones de la información extraída. Gracias a esto se puede observar como distintos tipos de gráficos y tablas pueden permitir sacar conclusiones de gran interés a la hora de analizar información.

Con este trabajo se pretende demostrar como un uso eficiente y efectivo de la información puede servir para generar conocimientos y apoyar el proceso de la toma de decisiones.

7. Trabajos futuros

Durante la realización de este trabajo han surgido tres áreas principales que no se han podido desarrollar debido a que la extensión de este hubiera sobrepasado lo indicado en las guías de la escuela, así como requerir de más tiempo para un desarrollo correcto y completo de estas.

En primer lugar, se queda la idea de realizar un análisis al completo de la competición estudiada, ya que durante la realización de la fase de análisis se han ido exponiendo ejemplos de distintas métricas y gráficos que se pueden utilizar, pero su aplicación a todas las áreas del torneo habría supuesto una longitud mucho mayor a la disponible. No obstante, este análisis ha servido como introducción al mundo del análisis deportivo y con los conocimientos adquiridos durante la realización de este trabajo, es posible para el alumno extender su conocimiento sobre esta, u otras competiciones de las que se dispongan datos tan completos como los utilizados en este documento.

En segundo lugar, se descartó la creación de un documento que permitiera la visualización interactiva de la información extraída durante el transcurso del trabajo. Esta decisión fue tomada tras debatirse que información sería interesante representar en un documento como este, si se debía centrar este en una sola selección, con un análisis detallado de esta, un resumen más general de la competición elegida o una visualización según las distintas métricas con las que se ha trabajado. En un primer lugar se eligió el programa de Microsoft PowerBI para la creación de este documento, pero se descartó a finales del trabajo por motivos de extensión de este y que la visualización conseguida con RStudio se creyó suficiente para explicar la utilidad del trabajo. Su creación en un futuro no se descarta debido al interés despertado por esta parte de la informática y ver en este programa en específico una gran herramienta para visualizar la información.

Por último, no se llegó a plantear para este trabajo, pero se tiene que mencionar el área de análisis predictivo, con el cual crear modelos predictivos que, gracias a la información procesada y desarrollada durante el trabajo, se pudieran utilizar para competiciones futuras. Esta área es, desde luego, de gran interés y utilidad para el futuro, pero se creyó fuera del alcance inicial del trabajo.

8. Referencias

AS. (2024, 3 de agosto). *Octavos de final de la Eurocopa: selecciones clasificadas, cuadro, horarios, partidos y cuándo se juegan*. AS. https://as.com/futbol/eurocopa/octavos-de-final-de-la-eurocopa-selecciones-clasificadas-cuadro-horarios-partidos-y-cuando-se-juegan-n-2/

Big Data Sports. (2022, 22 de junio). *Algunas pistas sobre el análisis de datos y el nuevo lenguaje del fútbol que FIFA mostrará en Qatar 2022*. Big Data Sports. https://bigdatasports.media/2022/06/22/algunas-pistas-sobre-el-analisis-de-datos-y-el-nuevo-lenguaje-del-futbol-que-fifa-mostrara-en-qatar-2022/

Data Discovery Solutions. (2024, 1 de septiembre). *La importancia del análisis de datos*. LinkedIn. https://www.linkedin.com/pulse/la-importancia-del-an%C3%A1lisis-de-datos-data-discovery-solutions/

El Orden Mundial. (2020, 22 de diciembre). *Confederaciones internacionales de fútbol según la FIFA*. El Orden Mundial. https://elordenmundial.com/mapas-y-graficos/confederaciones-internacionales-futbol-fifa/

FBref. (2024). *Estadísticas de UEFA Euro*. https://fbref.com/es/comps/676/Estadisticas-de-UEFA-Euro

FCrSTATS. (2018). *SBpitch*. GitHub. https://github.com/FCrSTATS/SBpitch

FIFA. (n.d.). *Asociaciones*. Inside FIFA. https://inside.fifa.com/es/about-fifa/associations

Jones, I. (2024, 11 de julio). *Euro Championship odds*. Sports Betting Dime. https://www.sportsbettingdime.com/soccer/euro-championship-odds/

Lewis, M. (2003). *Moneyball: The art of winning an unfair game*. W.W. Norton & Company.

Mart, J. (2017). *International football results from 1872 to 2017* [Dataset]. Kaggle. https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017/data

Memmert, D., & Raabe, D. (2018). *Data analytics in football: Positional data collection, modelling and analysis*. Routledge.

Microsoft. (2021, 19 de mayo). *LaLiga se alía con Microsoft para transformar digitalmente el fútbol a nivel mundial y reimaginar una nueva era en el deporte*. Microsoft. https://news.microsoft.com/es-es/2021/05/19/laliga-se-alia-con-microsoft-para-transformar-digitalmente-el-futbol-a-nivel-mundial-y-reimaginar-una-nueva-era-en-el-deporte/

Microsoft. (2021, 6 de octubre). *LaLiga y Microsoft presentan Beyond Stats, un proyecto de análisis futbolístico avanzado que profundiza en el juego de cada equipo*. Microsoft. https://news.microsoft.com/es-es/2021/10/06/laliga-y-microsoft-presentan-beyond-stats-un-proyecto-de-analisis-futbolistico-avanzado-que-profundiza-en-el-juego-de-cada-equipo/

Nielsen. (2018). *Fan favorite: The global popularity of football is rising*. Nielsen. https://www.nielsen.com/es/insights/2018/fan-favorite-the-global-popularity-of-football-is-rising/

Oddschecker. (2024, 5 de junio). *Euro 2024 odds: the ten most likely tournament winners*. Oddschecker. https://www.oddschecker.com/insight/football/20240605-euro-2024-odds-the-ten-most-likely-tournament-winners

Pavibear. (2022). *football-analytics*. GitHub. https://github.com/pavibear/football-analytics

Primicias. (2023, 28 de junio). *Análisis de datos en el deporte: cómo el "moneyball" ha transformado el fútbol*. Primicias. https://www.primicias.ec/noticias/firmas/analisis-datos-deporte-futbol-moneyball/

SportyTrader. (2024, 3 de junio). *Estadios de la Eurocopa 2024 de fútbol*. SportyTrader. https://www.sportytrader.es/actualidad/estadios-eurocopa-2024-futbol/

StatsBomb. (2018). *StatsBombR*. GitHub. https://github.com/statsbomb/StatsBombR

StatsBomb. (2024). *StatsBomb release free Euro 2024 data*. https://statsbomb.com/news/statsbomb-release-free-euro-2024-data/

Transfermarkt. (2024). *Euro 2024: participantes*. Transfermarkt. https://www.transfermarkt.com/euro-2024/teilnehmer/pokalwettbewerb/EM24

Universidad Europea. (2024). *Big data deportivo: cómo está revolucionando el mundo del deporte*. Universidad Europea. https://universidadeuropea.com/blog/big-data-deportivo/

UEFA. (n.d.). *Technical reports*. UEFA. https://www.uefatechnicalreports.com/

UEFA. (2023, 25 de enero). *Fase de clasificación para la UEFA Euro 2024: todo lo que necesitas saber*. UEFA. https://es.uefa.com/european-qualifiers/news/0279-1635bc3cb147-fe3ba77a9459-1000--fase-de-clasificacion-para-la-uefa-euro-2024-todo-lo-que-/

UEFA. (2024, 28 de marzo). *Nuestra historia*. UEFA. https://es.uefa.com/news-media/news/028b-1a8435364b4d-7091d23dddcf-1000--nuestra-historia/

UEFA. (2024). *Euro 2024: fixtures & results – bracket*. UEFA. https://www.uefa.com/euro2024/fixtures-results/bracket/

UEFA. (2024). *Spain vs. England: line-ups*. UEFA. https://es.uefa.com/euro2024/match/2036211--spain-vs-england/lineups/

UEFA. (2024). *Euro 2024: standings*. UEFA. https://es.uefa.com/euro2024/standings/

UEFA. (2024). *Equipo del torneo de la UEFA Euro 2024*. UEFA. https://es.uefa.com/euro2024/news/028f-1b61a1c5a9c4-6b7d81c77bfe-1000--equipo-del-torneo-de-la-uefa-euro-2024/

9. Anexos

9.1. Anexo 1: ODS

OBJETIVOS DE DESARROLLO SOSTENIBLE

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenibles	Alto	Medio	Bajo	No Procede
ODS 1. **Fin de la pobreza.**				X
ODS 2. **Hambre cero.**				X
ODS 3. **Salud y bienestar.**			X	
ODS 4. **Educación de calidad.**			X	
ODS 5. **Igualdad de género.**				X
ODS 6. **Agua limpia y saneamiento.**				X
ODS 7. **Energía asequible y no contaminante.**				X
ODS 8. **Trabajo decente y crecimiento económico.**		X		
ODS 9. **Industria, innovación e infraestructuras.**	X			
ODS 10. **Reducción de las desigualdades**.				X
ODS 11. **Ciudades y comunidades sostenibles**.				X
ODS 12. **Producción y consumo responsables.**				X
ODS 13. **Acción por el clima.**				X
ODS 14. **Vida submarina.**				X
ODS 15. **Vida de ecosistemas terrestres.**				X
ODS 16. **Paz, justicia e instituciones sólidas.**				X
ODS 17. **Alianzas para lograr objetivos.**		X		

Reflexión sobre la relación del TFG/TFM con los ODS y con el/los ODS más relacionados.

- ODS 8. Trabajo decente y crecimiento económico:

El análisis de datos y el manejo de la información se han convertido hoy en día en habilidades indispensables en gran parte de las áreas del mercado laboral. Estos permiten el desarrollo de una economía más sostenible y una toma de decisiones basada en el conocimiento extraído.

- ODS 9. Industria, innovación e infraestructuras:

Este trabajo promueve el uso de tecnologías innovadoras en el mundo del fútbol. A través de la recolección y el análisis de datos se es capaz de procesar grandes volúmenes de información para mejorar el rendimiento deportivo y la comprensión sobre este deporte, impulsando la innovación en la industria del deporte. Gracias a esta innovación, el propio deporte es capaz de evolucionar tácticamente, generando nuevas ideas para los entrenadores y nuevas estrategias tanto dentro como fuera del campo.

- ODS 17. Alianzas para lograr objetivos:

El uso de datos disponibles para el público como es el caso de aquellos que se utilizan durante el desarrollo de este trabajo, así como analizar estos datos y compartir abiertamente las conclusiones, se pueden considerar como una cooperación entre distintos agentes, reforzando la idea de que las alianzas son claves para un desarrollo. Los análisis basados en datos que pueden suponer un apoyo para la toma de decisiones informadas promueven una cultura de colaboración necesaria para que estos objetivos se cumplan.

9.2. Anexo 2: Código función *head_to_head*

```
head_to_head <- function(data, team1, team2, competition = NULL) {
  if (!is.null(competition)) {
    data <- data %>%
      dplyr::filter(tournament == competition)
  } else {
    data <- data
  }
  # Filtrar los partidos donde los dos equipos se han enfrentado
  h2h_matches <- data[((data$home_team == team1 & data$away_team == team2) |
                       (data$home_team == team2 & data$away_team == team1)), ]
  # Inicializar contadores
  team1_wins <- 0
  team2_wins <- 0
  draws <- 0
  team1_goals <- 0
  team2_goals <- 0
  # Recorrer cada partido para contar resultados
  for (i in 1:nrow(h2h_matches)) {
    home_team <- h2h_matches$home_team[i]
    away_team <- h2h_matches$away_team[i]
    home_score <- h2h_matches$home_score[i]
    away_score <- h2h_matches$away_score[i]
```

```
# Sumar goles
if (home_team == team1) {
  team1_goals <- team1_goals + home_score
  team2_goals <- team2_goals + away_score
} else {
  team1_goals <- team1_goals + away_score
  team2_goals <- team2_goals + home_score
}
# Contabilizar resultados
if (home_score > away_score) {
  if (home_team == team1) {
    team1_wins <- team1_wins + 1
  } else {
    team2_wins <- team2_wins + 1
  }
} else if (home_score < away_score) {
  if (away_team == team1) {
    team1_wins <- team1_wins + 1
  } else {
    team2_wins <- team2_wins + 1
  }
} else {
  draws <- draws + 1
}
```

```
  }

  # Crear un data frame con el resumen del head-to-head
  result <- data.frame(
    team1 = team1,
    team2 = team2,
    team1_wins = team1_wins,
    team2_wins = team2_wins,
    draws = draws,
    team1_goals = team1_goals,
    team2_goals = team2_goals,
    total_matches = nrow(h2h_matches)
  )
  return(result)
}
```

9.3. Anexo 3: Código mapa de tiros

```
shots = StatsBombData %>%

 filter(type.name=="Shot" & (shot.type.name!="Penalty" | is.na(shot.type.name)) & player.name=="Cristiano Ronaldo dos Santos Aveiro") %>%

 mutate(shot.body_part_ESP.name = recode (shot.body_part.name, "Right Foot" = "Pie derecho", "Left Foot" = "Pie izquierdo", "Head" = "Cabeza"))

 #Asignación de colores según el xG generado por el tiro

 shotmapxgcolors <- c("#192780", "#2a5d9f", "#40a7d0", "#87cdcf", "#e7f8e6", "#f4ef95", "#FDE960", "#FCDC5F", "#F5B94D", "#F0983E", "#ED8A37", "#E66424", "#D54F1B", "#DC2608", "#BF0000", "#7F0000", "#5F0000")

#Creación del gráfico

ggplot() +

 annotate("rect",xmin = 0, xmax = 120, ymin = 0, ymax = 80, fill = NA, colour = "black", size = 0.6) +

 annotate("rect",xmin = 0, xmax = 60, ymin = 0, ymax = 80, fill = NA, colour = "black", size = 0.6) +

 annotate("rect",xmin = 18, xmax = 0, ymin = 18, ymax = 62, fill = NA, colour = "black", size = 0.6) +

 annotate("rect",xmin = 102, xmax = 120, ymin = 18, ymax = 62, fill = NA, colour = "black", size = 0.6) +

 annotate("rect",xmin = 0, xmax = 6, ymin = 30, ymax = 50, fill = NA, colour = "black", size = 0.6) +

 annotate("rect",xmin = 120, xmax = 114, ymin = 30, ymax = 50, fill = NA, colour = "black", size = 0.6) +

 annotate("rect",xmin = 120, xmax = 120.5, ymin =36, ymax = 44, fill = NA, colour = "black", size = 0.6) +

 annotate("rect",xmin = 0, xmax = -0.5, ymin =36, ymax = 44, fill = NA, colour = "black", size = 0.6) +
```

```
  annotate("segment", x = 60, xend = 60, y = -0.5, yend = 80.5, colour = "black", size = 0.6)+

  annotate("segment", x = 0, xend = 0, y = 0, yend = 80, colour = "black", size = 0.6)+

  annotate("segment", x = 120, xend = 120, y = 0, yend = 80, colour = "black", size = 0.6)+

  theme(rect = element_blank(),

        line = element_blank()) +

  annotate("point", x = 108 , y = 40, colour = "black", size = 1.05) +

  annotate("path", colour = "black", size = 0.6, x=60+10*cos(seq(0,2*pi,length.out=2000)), y=40+10*sin(seq(0,2*pi,length.out=2000)))+

  annotate("point", x = 60 , y = 40, colour = "black", size = 1.05) +

  annotate("path", x=12+10*cos(seq(-0.3*pi,0.3*pi,length.out=30)), size = 0.6, y=40+10*sin(seq(-0.3*pi,0.3*pi,length.out=30)), col="black") +

  annotate("path", x=107.84-10*cos(seq(-0.3*pi,0.3*pi,length.out=30)), size = 0.6, y=40-10*sin(seq(-0.3*pi,0.3*pi,length.out=30)), col="black") +

  geom_point(data = shots, aes(x = location.x, y = location.y, fill = shot.statsbomb_xg, shape = shot.body_part_ESP.name), size = 6, alpha = 0.8) +

  theme(axis.text.x=element_blank(),

        axis.title.x = element_blank(),

        axis.title.y = element_blank(),

        plot.caption=element_text(size=13,family="Source Sans Pro", hjust=0.5, vjust=0.5),

        plot.subtitle = element_text(size = 18, family="Source Sans Pro", hjust = 0.5),

        axis.text.y=element_blank(), legend.position = "top",

        legend.title=element_text(size=18,family="Source Sans Pro"),

        legend.text=element_text(size=16,family="Source Sans Pro"),
```

```
        legend.margin = margin(c(15, 30, -60, 10)), legend.key.size = unit(1.5, "cm"),

        legend.key.width = unit(0.95, "cm"),

        plot.title = element_text(margin = margin(r = 10, b = 10), face="bold",size = 26, family="Source Sans Pro", colour = "black", hjust = 0.5),

        legend.direction = "horizontal",

        axis.ticks=element_blank(),

        aspect.ratio = c(65/100),

        plot.background = element_rect(fill = "white"),

        strip.text.x = element_text(size=13,family="Source Sans Pro")) +

  labs(title = "Cristiano Ronaldo, Mapa de tiros", subtitle = "Eurocopa 2024") +

  scale_fill_gradientn(colours = shotmapxgcolors, limit = c(0,0.8), oob=scales::squish, name = "Valor de xG") +

  scale_shape_manual(values = c("Cabeza" = 21, "Pie derecho" = 23, "Pie izquierdo" = 24), name ="") +

  guides(fill = guide_colourbar(title.position = "top"),

         shape = guide_legend(override.aes = list(size = 5, fill = "black"))) +

  coord_flip(xlim = c(85, 125))
```