

Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Estudios de Postgrado
Maestría en Ingeniería para la Industria con
Especialización en Ciencias de la Computación
Introducción Data Analytics



ESCUELA DE ESTUDIOS DE
POSTGRADO
FACULTAD DE INGENIERÍA

Análisis de Datos - Jugadores de Fútbol - 2015 al 2020

Proyecto Inicial

16 de noviembre del 2,024

Análisis de Datos de Jugadores de Fútbol (2015-2020)

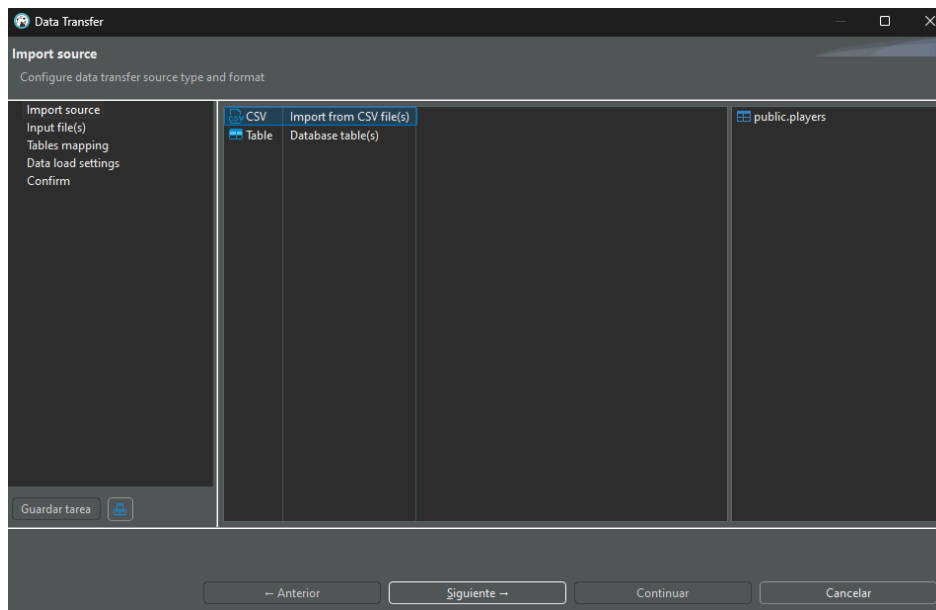
Introducción

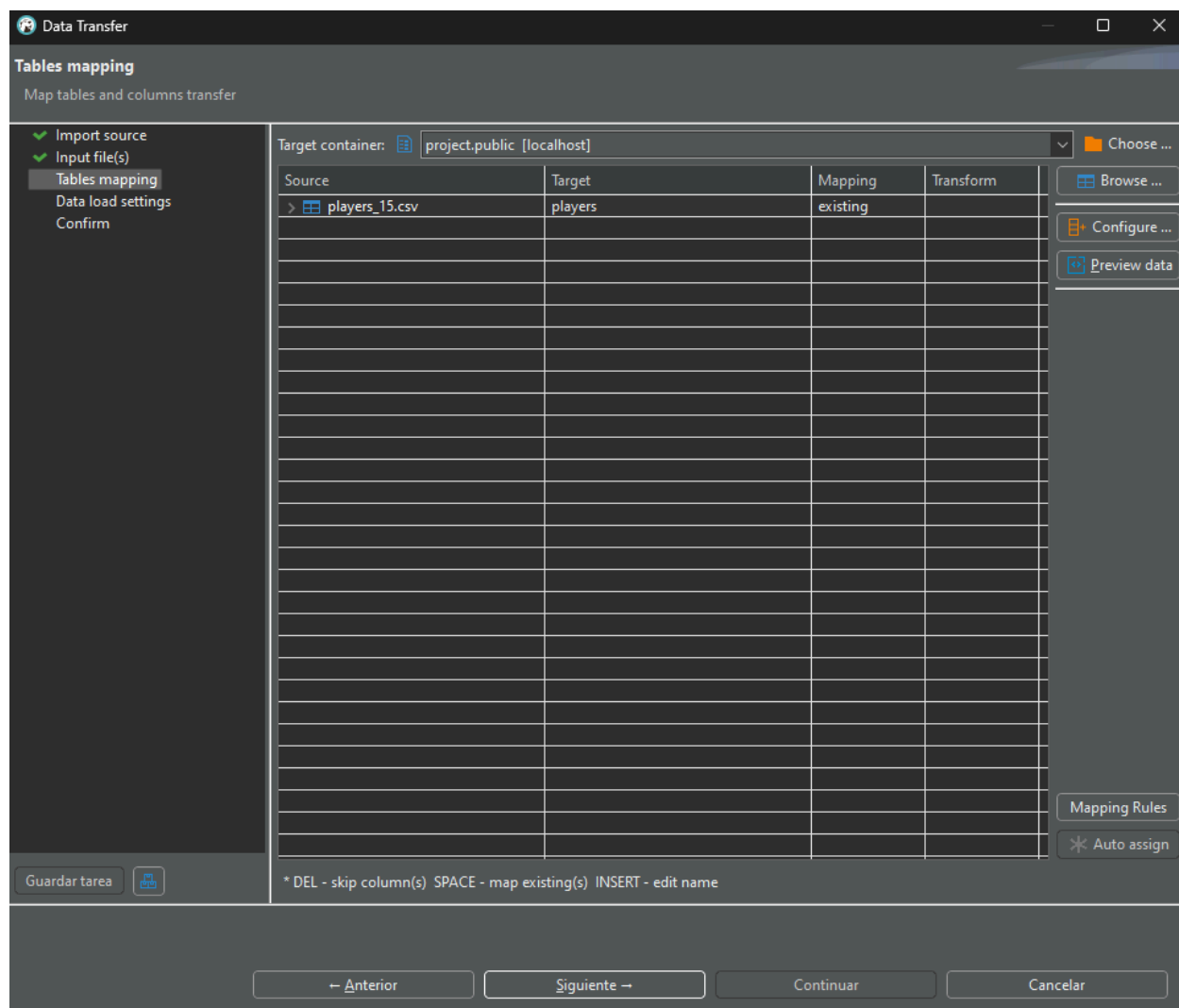
El análisis se centró en un conjunto de datos de jugadores de fútbol recopilados entre 2015 y 2020. El objetivo principal fue explorar y comprender mejor las características de los jugadores mediante la manipulación de datos y análisis exploratorio. Para ello, se utilizaron herramientas de ciencia de datos como Python, junto con bibliotecas como pandas, matplotlib, y seaborn. También se integraron bases de datos PostgreSQL para la gestión de los datos.

Metodología

El trabajo se desarrolló en varias fases:

1. **Extracción de Datos:** Se realizó una conexión a una base de datos PostgreSQL donde se realizó el esquema de la información y se importó de manera manual los archivos .csv, para acceder a la tabla de jugadores. Utilizando consultas SQL, se recuperó un conjunto de datos que incluía variables relevantes como el potencial, las habilidades técnicas, y otros atributos importantes. Se usó DBeaver como gestor de base de datos, realizando el siguiente proceso para importar, el archivo de creación de la DB se encuentra en el github.





Se escoge el csv de inicio y se da siguiente hasta que termine de importar la información

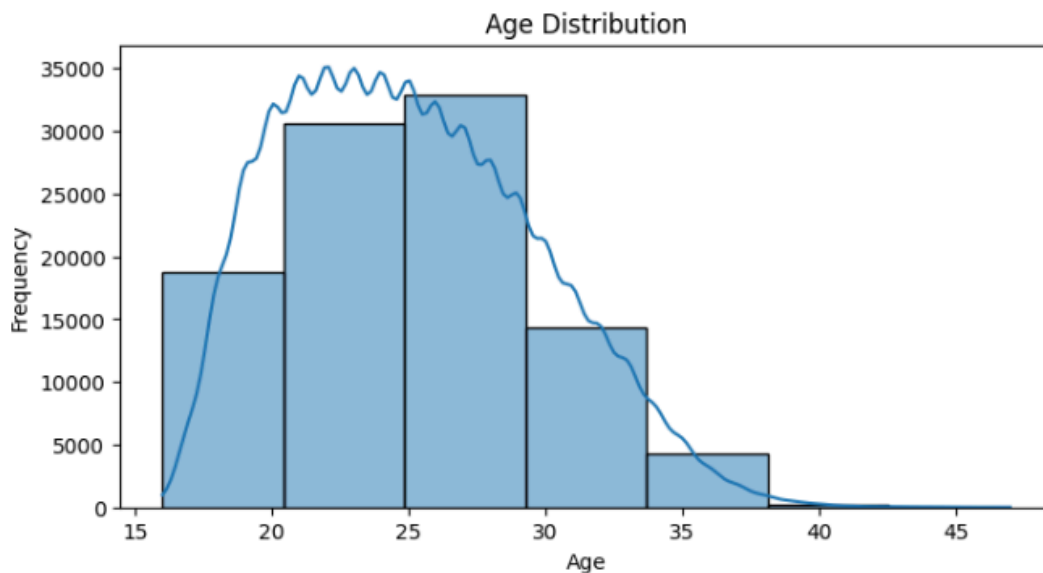
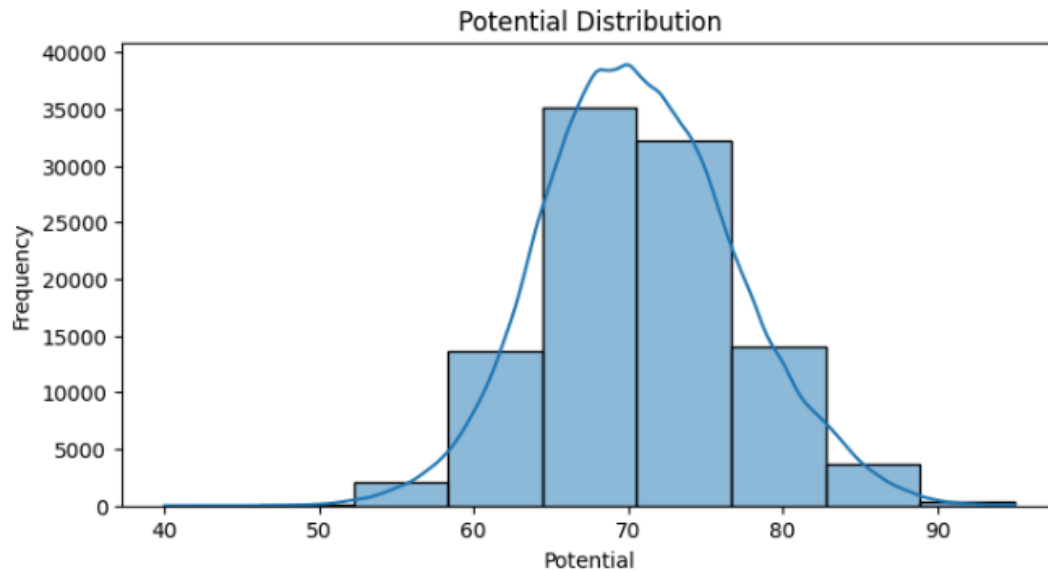
- 2. Limpieza de Datos:** Antes de proceder con los análisis, se llevaron a cabo tareas de preprocesamiento. Esto incluyó:
 - Eliminación de columnas irrelevantes, como fechas de ingreso que no aportan valor analítico.
 - Revisión de valores faltantes o inconsistentes en las columnas clave.
 - Estandarización de formatos para facilitar la manipulación de los datos.

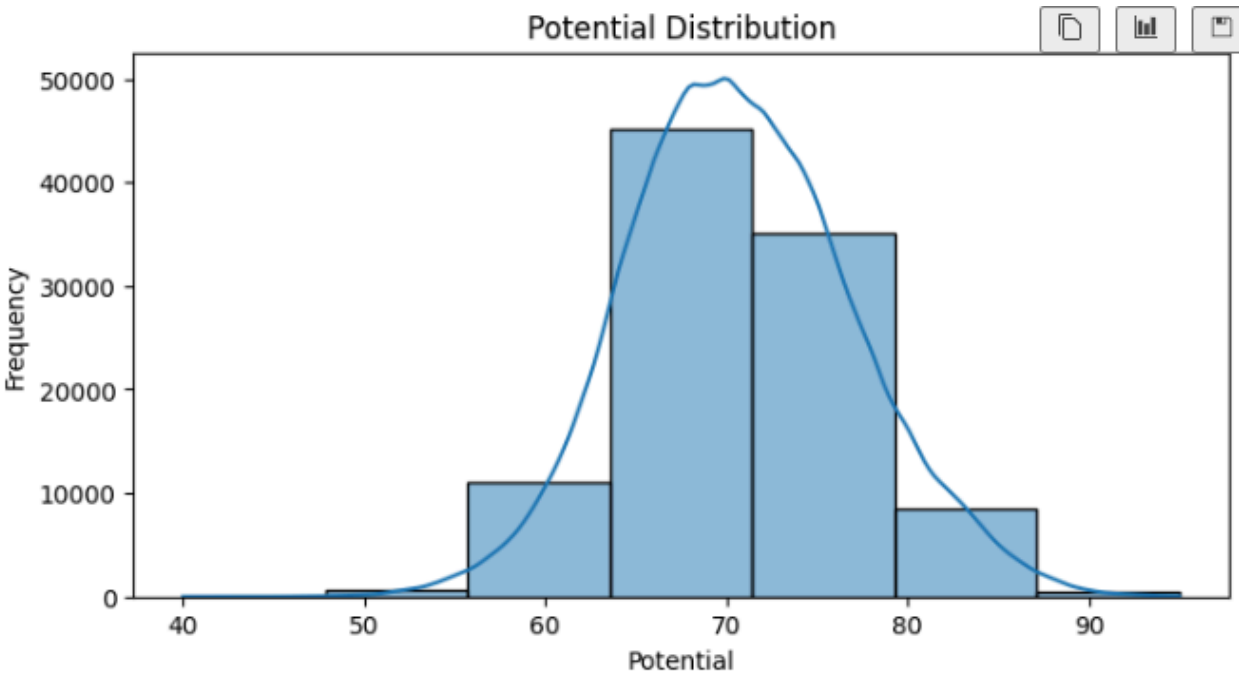
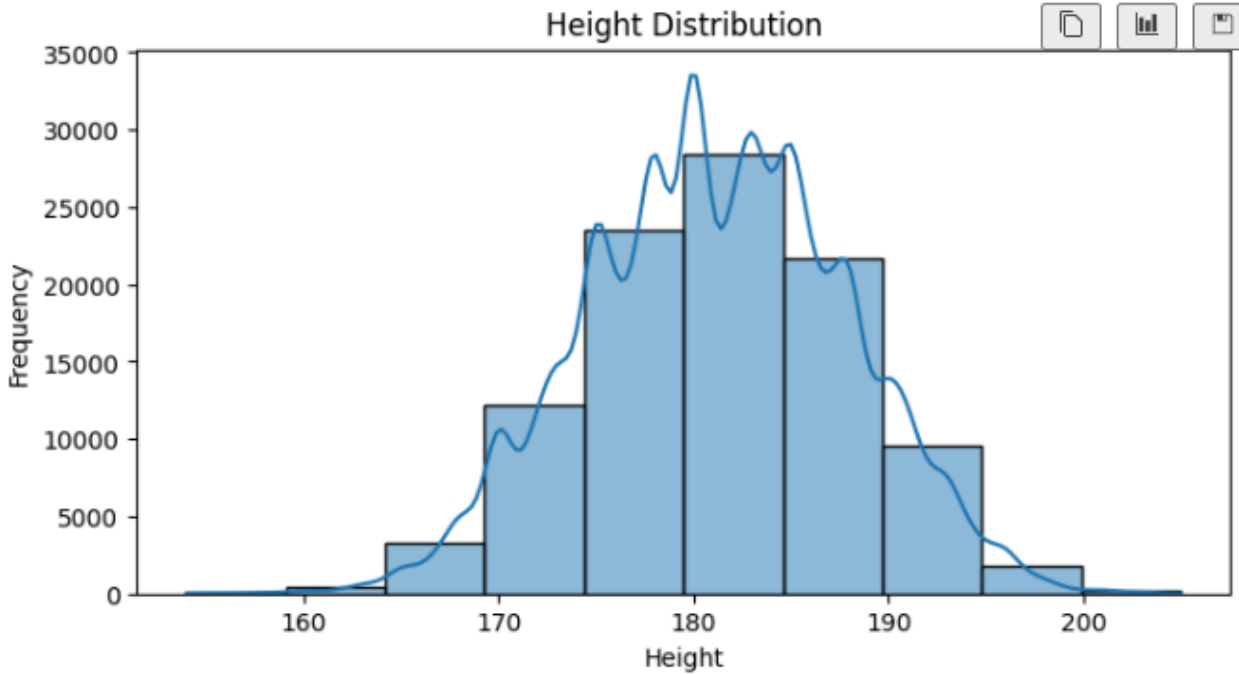
Todo este procedimiento se encuentra en el cuaderno de trabajo en python.

3. **Análisis Exploratorio de Datos (EDA):** Para comprender la distribución y relación entre las variables, se generaron visualizaciones estadísticas. Estas ayudaron a identificar patrones en los datos, como:

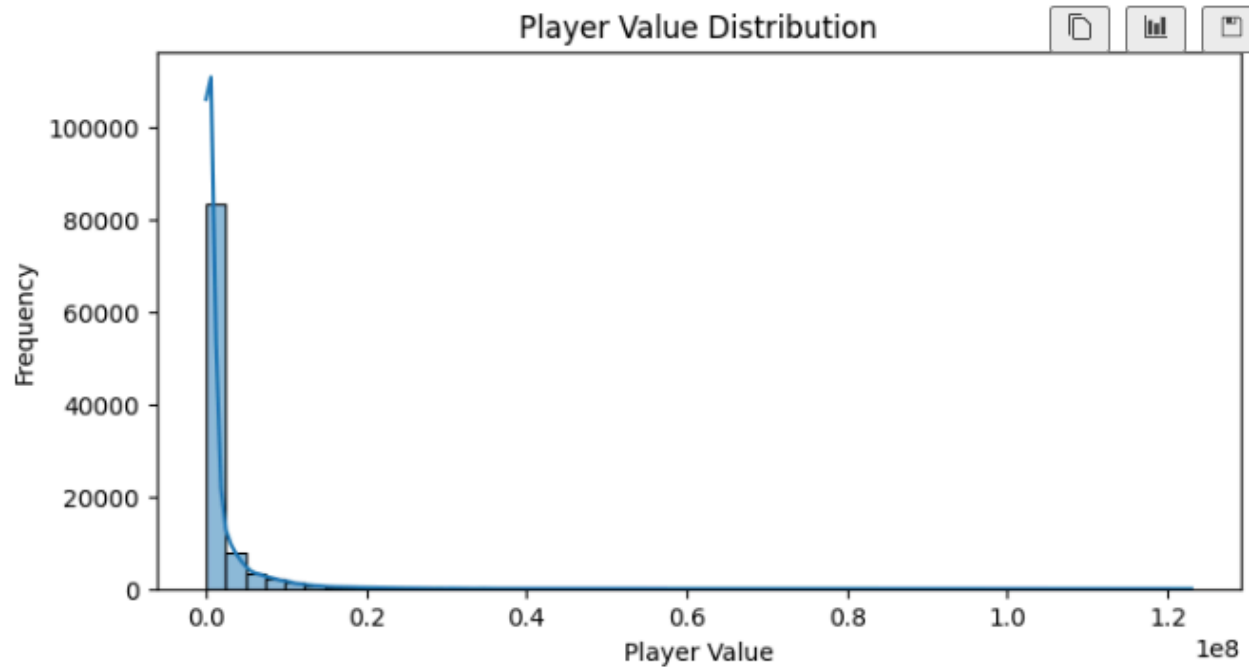
- Distribución del potencial de los jugadores.
- Comparaciones entre atributos técnicos y rendimiento.

Casi todos los datos técnicos del físico, lb, potencial y entre otros funcionan de la misma manera, hay una amplia mayoría alrededor de un rango y outsiders por los extremos

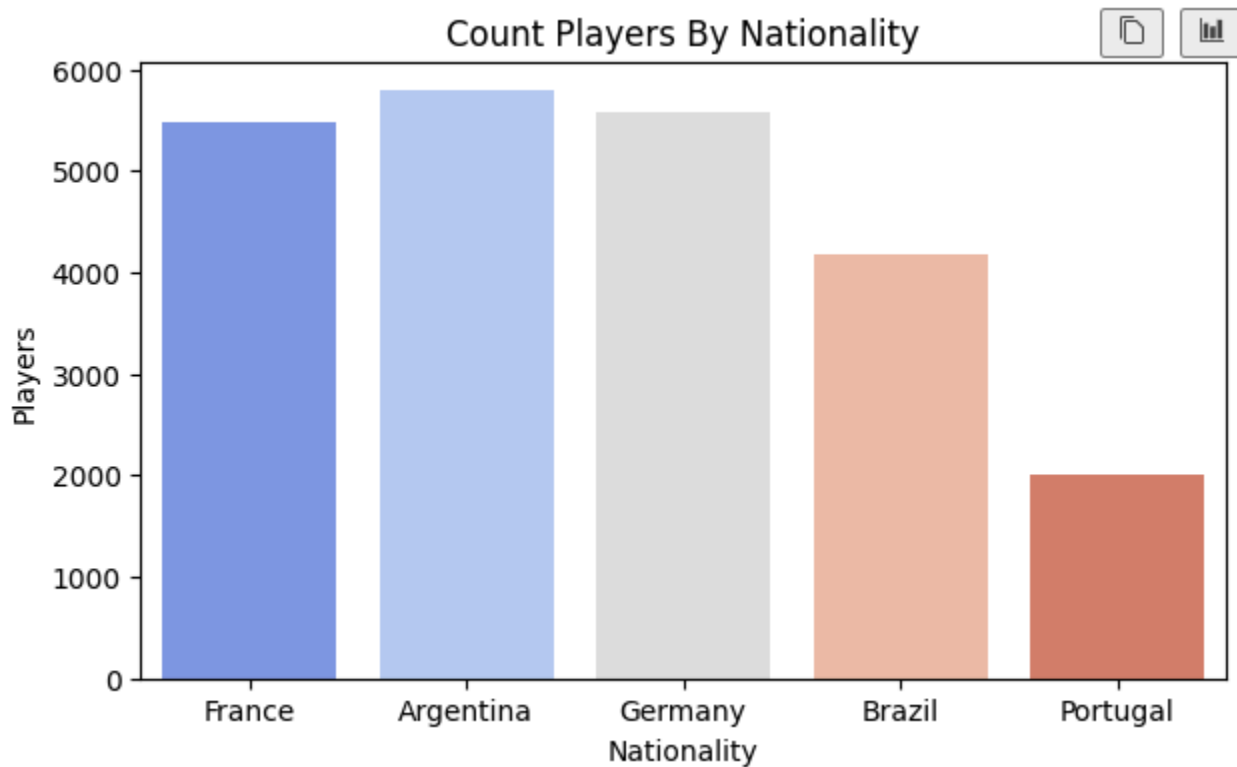




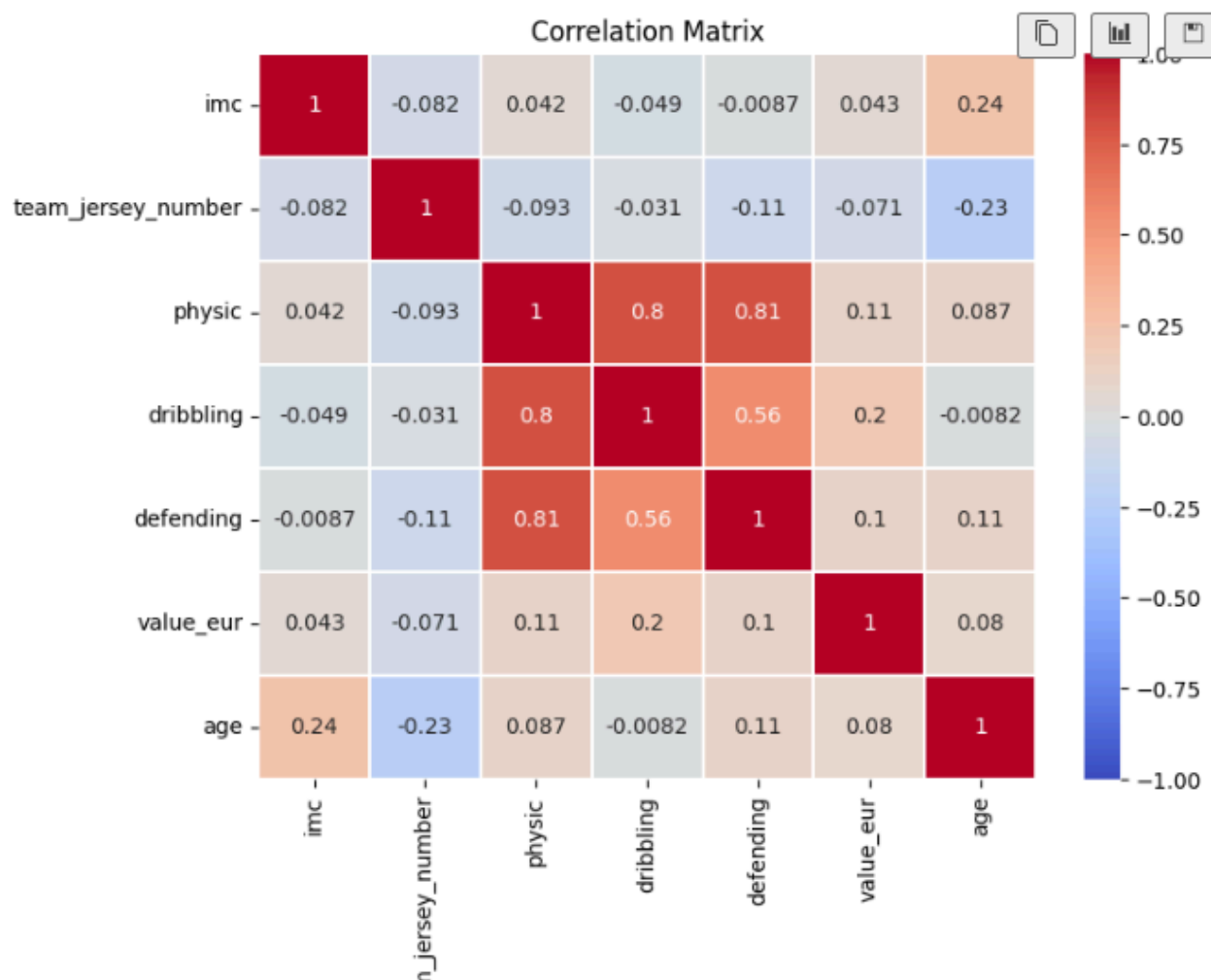
Solo con el valor de los jugadores encontramos una alta mayoría en el rango inferior y disminuye rápidamente, hasta tener outsiders muy alejados de la amplia mayoría.



También como medio interesante se escogieron unos países para ver qué cantidad de jugadores han sido trackeados, y es muy distinto entre países.



4. **Data Cleaning y Data Wrangling:** A través de varias técnicas y análisis de información se limpió la data recibida y se agregaron nuevas variables para obtener información mucho más completa y bien implementadas para hacer el análisis posterior
5. **Data transformation y Visualización de Resultados:** De último ya se realizó la transformación de una data que estaba en rangos y se necesitaba un valor numérico para calcular en estos datos de manera más sencilla. También a través de gráficos como histogramas, gráficos de dispersión y mapas de calor, se presentaron insights clave relacionados con las habilidades y características de los jugadores. Esto permitió observar tendencias y agrupamientos dentro de los datos.



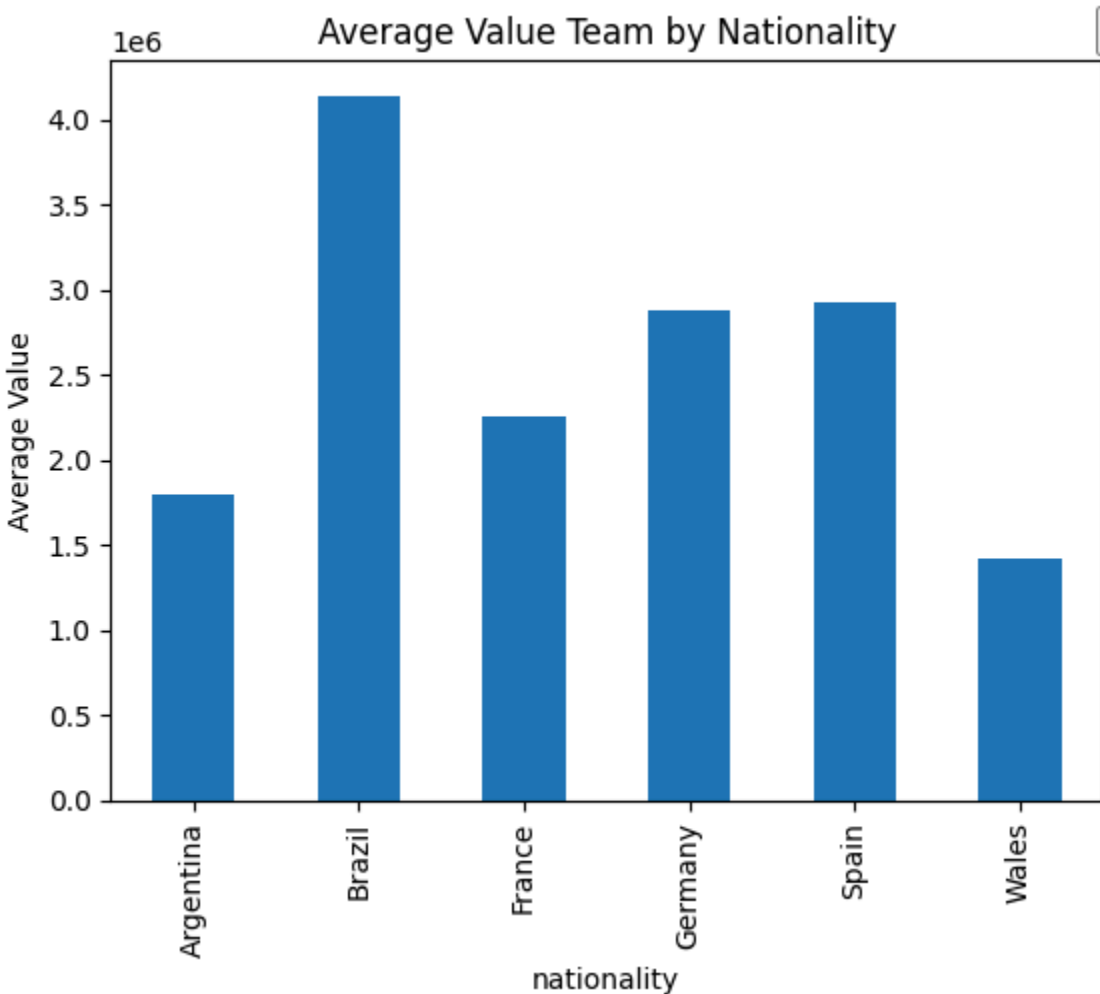
Podemos ver que del IMC con el que tiene más relación es con la edad, tiene sentido, ya que la edad muchas veces afecta al peso.

También vemos algo interesante y es que el físico está altamente relacionado con lo que es el dribleo y la defensa, mientras que la defensa y el dribleo también están relacionados pero en menor medida.

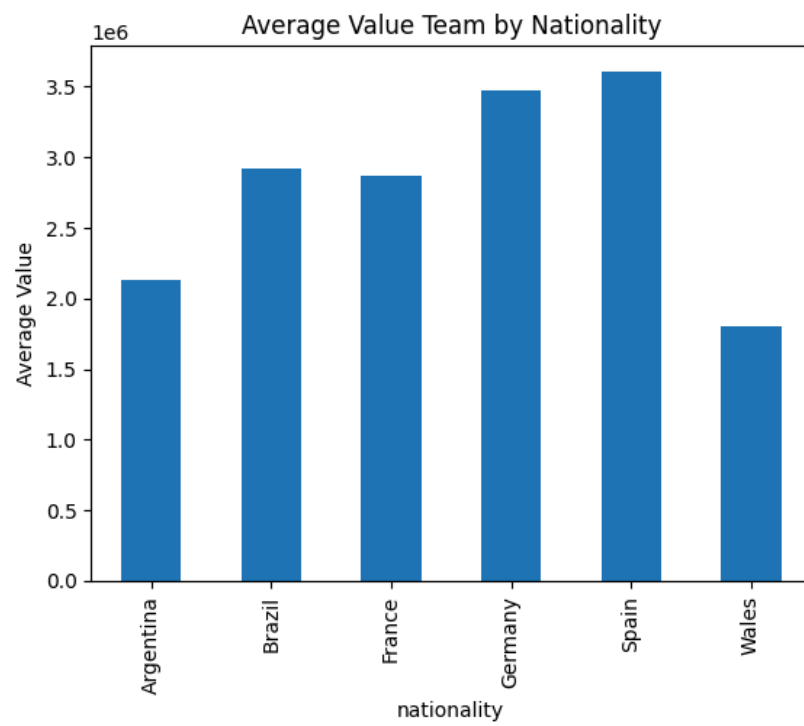
Lo más interesante de esta tabla de correlación es la relación de valor en euros que con el que tiene mayor relación es con el dribleo, osea estrellas de ataque como Cristiano, Messi, y la edad como el físico o defensa son casi la mitad de importantes que el dribleo.

A continuación se mostrará a través de los años la media de valor que hay de jugadores en diferentes países, 5 potencias en el fútbol y una más pequeña en logros históricos.

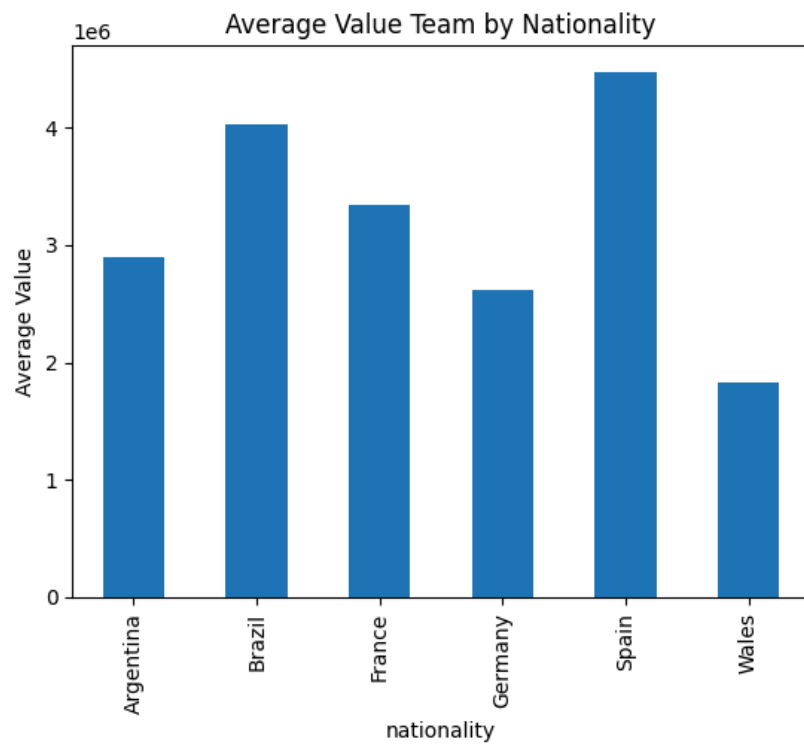
2016



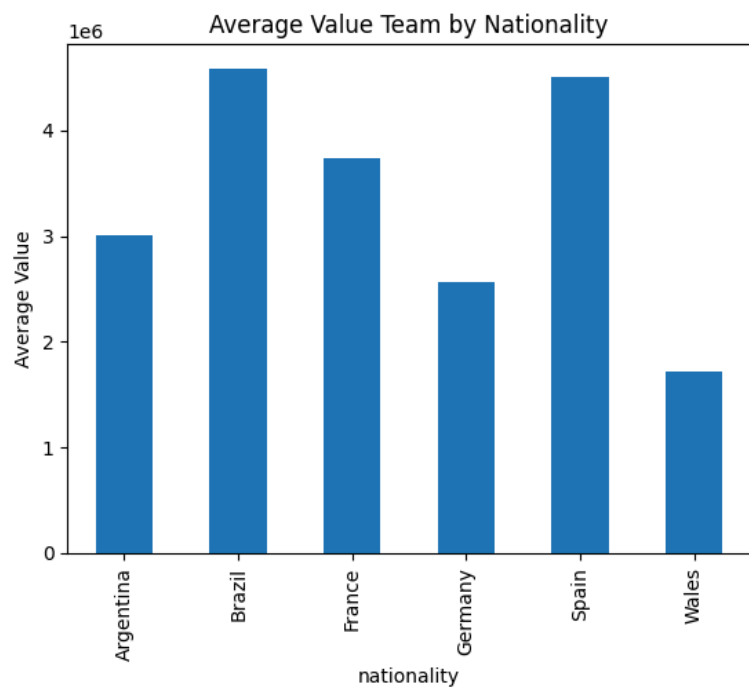
2017



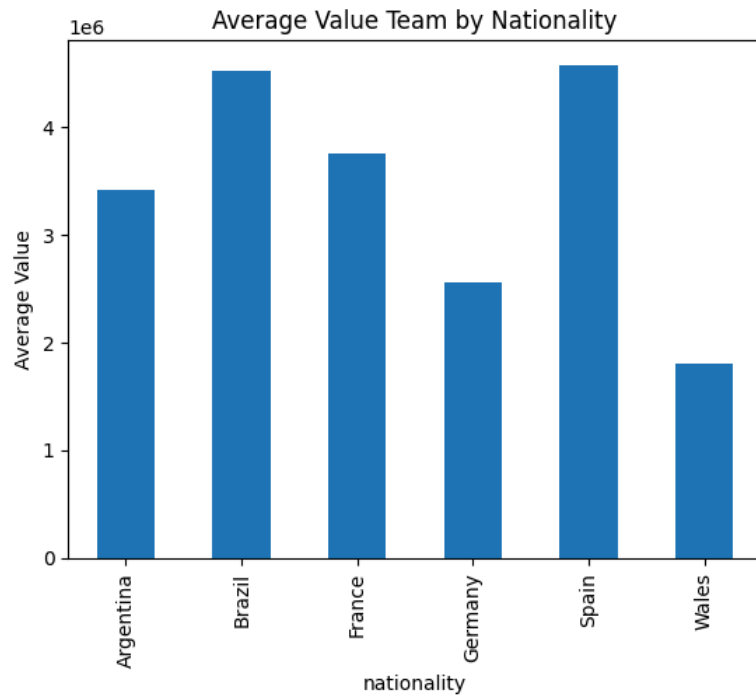
2018



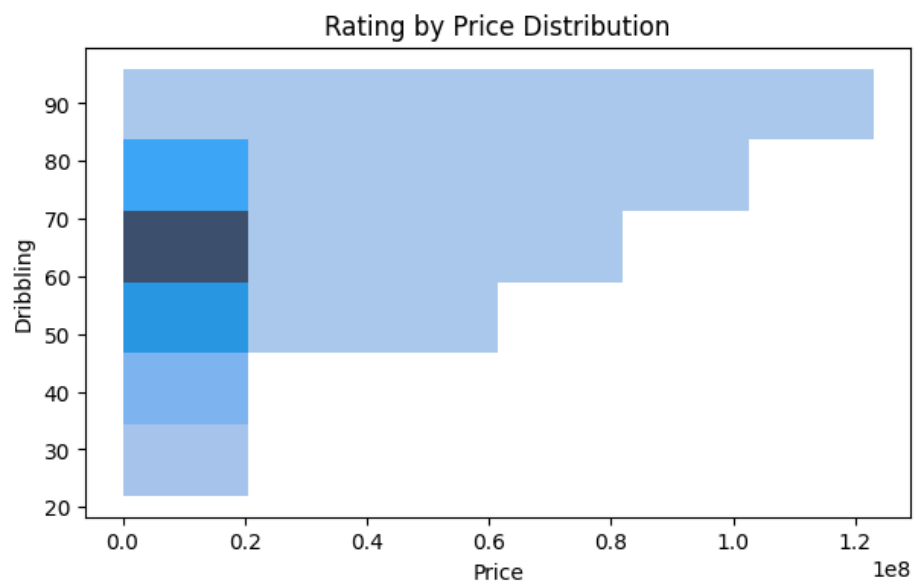
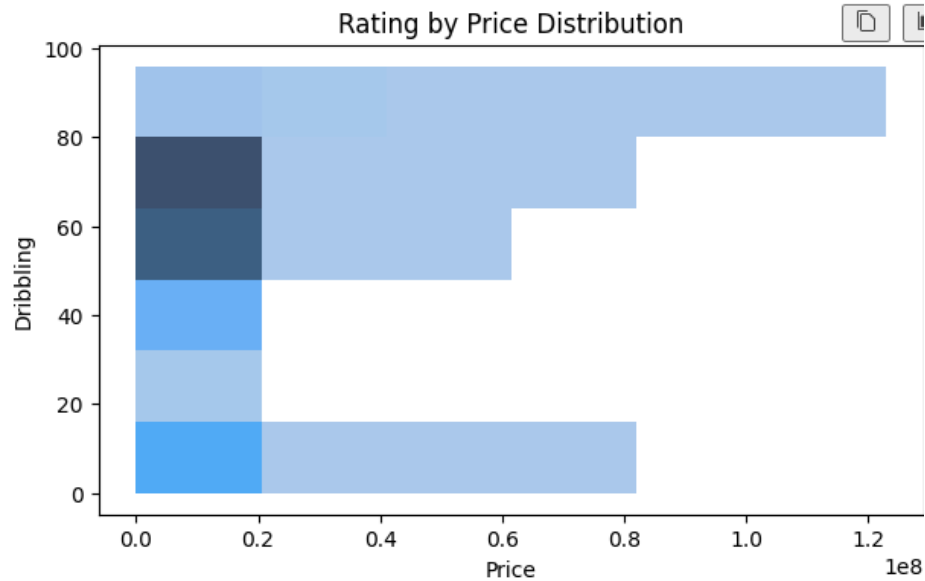
2019



2020

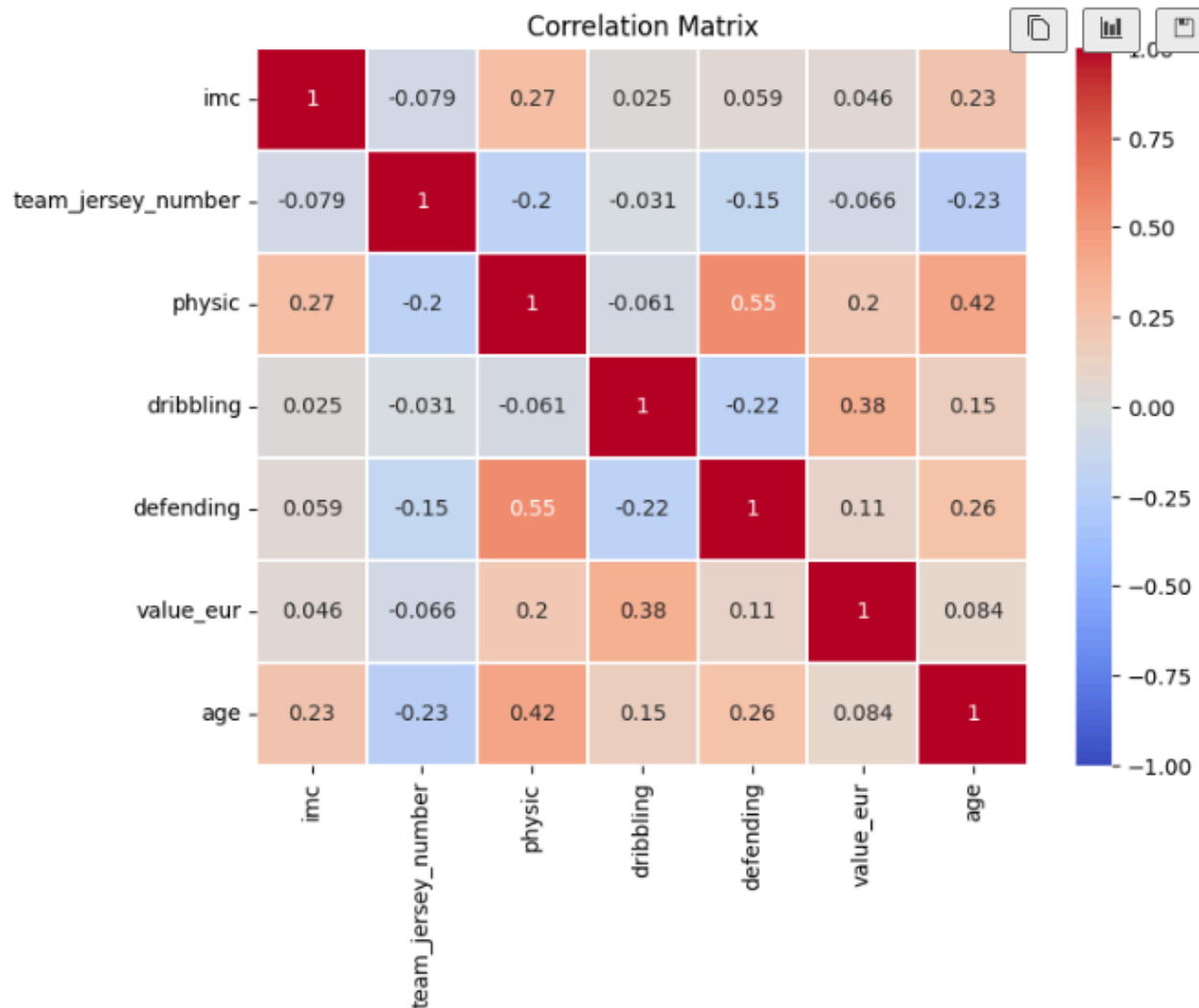


Nos damos cuenta que la nación más pequeña en la historia del fútbol, tiene en promedio jugadores de menor valor en el mercado, lo cuál puede mantener esta ventaja estructural de los países que ya tienen muchos jugadores en grandes ligas.



En este último es sobre el valor de los jugadores vs dribbling que vimos en la matriz de correlación que tienen una correlación significativa, la primera es contando a porteros, y la segunda es sin porteros, lo cuál muestra que en los porteros para el precio que tienen en el mercado no es importante esta característica, pero en los demás jugadores parece ser muy importante sin importar la posición.

Dando como matriz de correlación con solo jugadores de campo lo siguiente, que hace que el dribbling esté aún más relacionado con el precio de los jugadores.



Resultados

Durante el análisis, se observaron los siguientes aspectos destacados:

- **Distribución del Potencial:** Los jugadores mostraron una distribución variada de su potencial, con picos en ciertos rangos específicos. Esto sugiere una concentración de talento en categorías particulares.
- **Correlaciones Entre Variables:** A través de un análisis de correlación, se detectaron relaciones significativas entre habilidades técnicas y el rendimiento potencial.
- **Características Notables:** Se identificaron subgrupos de jugadores con habilidades específicas, útiles para segmentar y clasificar a los atletas en base a sus fortalezas.

Conclusión

Este análisis proporcionó una visión general sobre las características de los jugadores de fútbol entre 2015 y 2020. Los resultados obtenidos pueden servir de base para investigaciones más detalladas, como predicciones de rendimiento, análisis de reclutamiento o estudios de evolución de habilidades a lo largo del tiempo. Además, el enfoque utilizado garantiza la reproducibilidad del análisis para futuros conjuntos de datos similares.