**EDA: ANÁLISIS DE PARTIDOS INTERNACIONALES DE FÚTBOL**



MEMORIA

MARCOS ALFONSO MARTÍNEZ

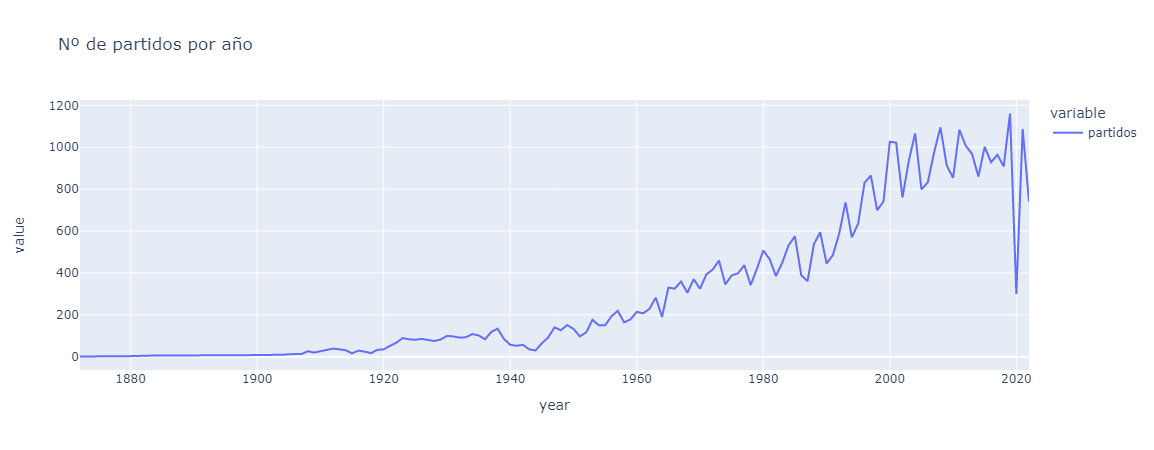
**Introducción.**

Para este EDA hemos utilizado los siguientes [dataset](https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017). Hemos elegido este en concreto por varias razones. La primera es porque el fútbol es un tema que me interesa bastante. La segunda es que tiene una cantidad de datos considerable y se puede sacar bastante información. La última es que los datos están actualizados y bastante limpios. Salvo por algún dato puntual todos los demás están bien, por lo que no hemos tenido que realizar mucha limpieza. Uno de los dataset complementa al otro, ya que en el de resultados globales no incluye el resultado de las tandas de penaltis en caso de empate, así que se ha hecho un merge sobre las columnas comunes y sacar la nueva, que hemos tenido que rellenar realizando un apply con una función que saque los ganadores o los empates sin tanda de penaltis. También hemos realizado algún retoque como por ejemplo pasar la fecha a formato pd.DateTime, sacar año por separado, y una columna con los goles totales.

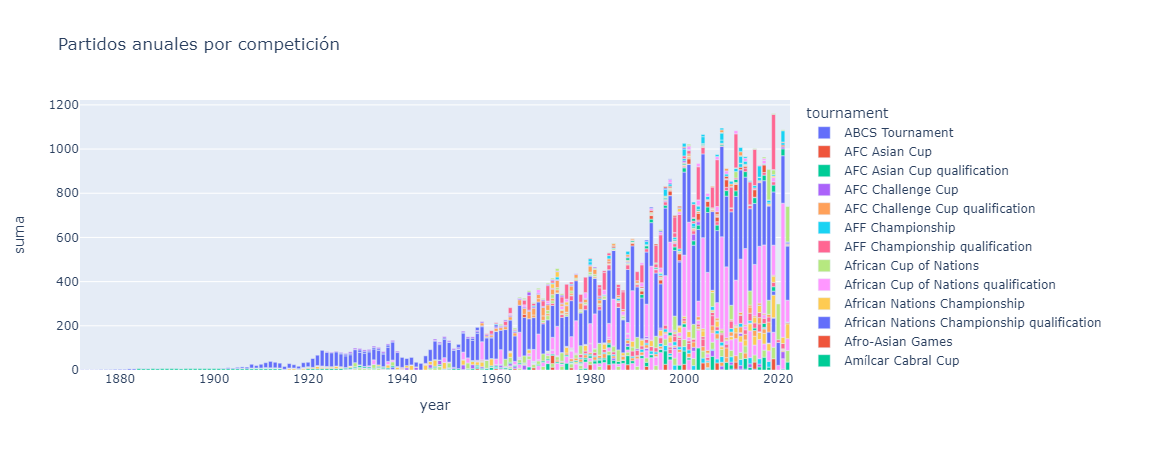
**Evolución del fútbol internacional.**

Toda esta parte ha consistido en hacer groupby y colocar bien los datos.

Lo primero que hacemos es ver los partidos que se jugaron cada año y lo representamos de 2 formas, mediante un line plot y luego mediante un scatter plot para ver bien representados los datos.



Lo siguiente ha sido ver las competiciones que corresponden a esos partidos cada año para poder comprobar la causa de los picos periódicos. Como hemos dicho anteriormente, los picos son causados por la presencia del mundial, pero ahora está confirmado. La gráfica puede parecer que tiene demasiada información (muchas competiciones), pero si quitábamos competiciones con pocos partidos se nos iban algunas importantes. Para poder usar las gráficas interactivas y ayudar a su mejor interpretación hemos usado plotly. Después de esta ha venido una gráfica de un intervalo de 4 años de la misma para poder ver mejor los datos.



**Cuándo se empezó a jugar.**

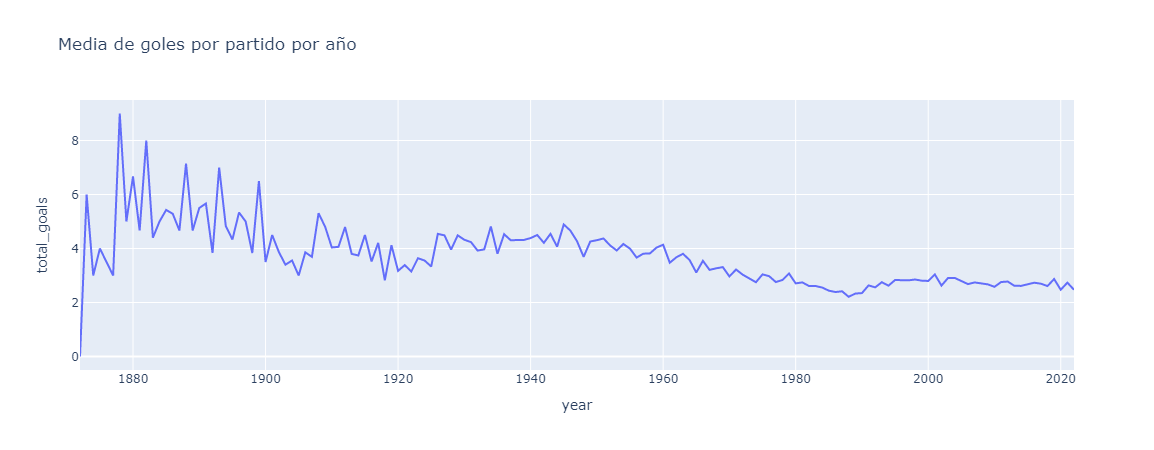
Después de unos groupby conseguimos sacar el año en el que cada selección jugó su primer partido. En la presentación lo diremos de palabra, ya que no merece mucho la pena gastar demasiado tiempo explicando estas gráficas. Las primeras selecciones fueron la actual Gran Bretaña, seguidas de Norteamérica y Canadá.

Siguiendo en la misma tónica hacemos una suma acumulativa para ver cuántos equipos iba habiendo cada año. De nuevo no pondremos esto en la presentación porque habría poco que explicar nuevo.

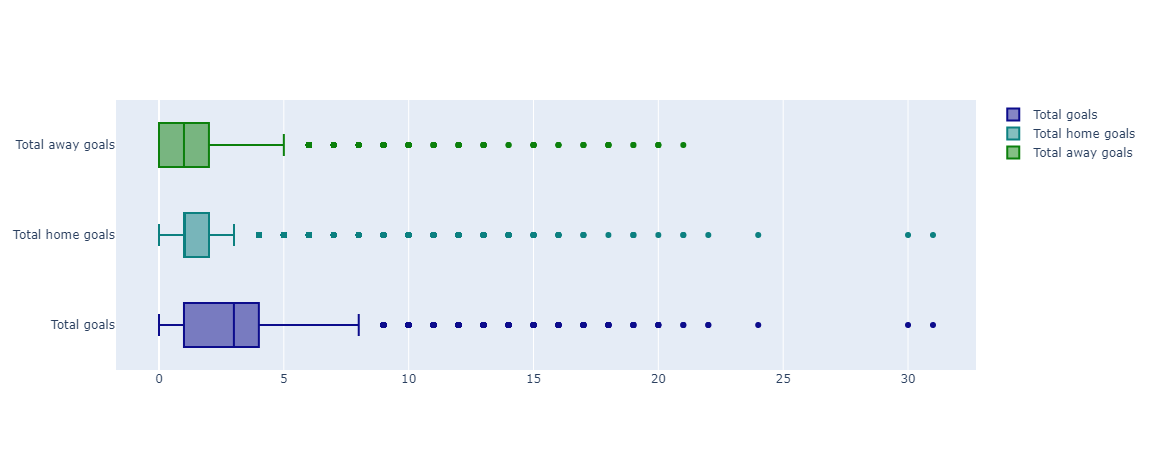
Para terminar este apartado echamos un vistazo a los equipos que más partidos han jugado. Hemos creado una serie que junte los equipos locales y los visitantes (ningún equipo juega contra sí mismo) y podemos ver cuántos hay con un groupby.

**Goles.**

En este apartado usamos la columna de goles totales creada al principio para agrupar por año y realizar la media para ver como ha variado en el tiempo este valor.

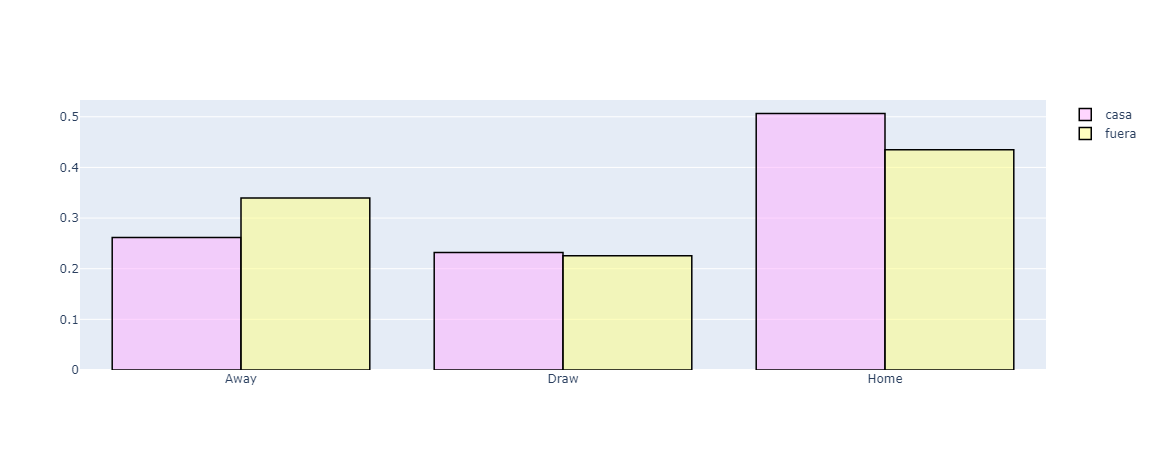


A continuación vemos como se distribuyen los valores con un boxplot y ver las diferencias entre los goles que se marcan en casa y los goles fuera. Directamente usamos las columnas pertinentes.

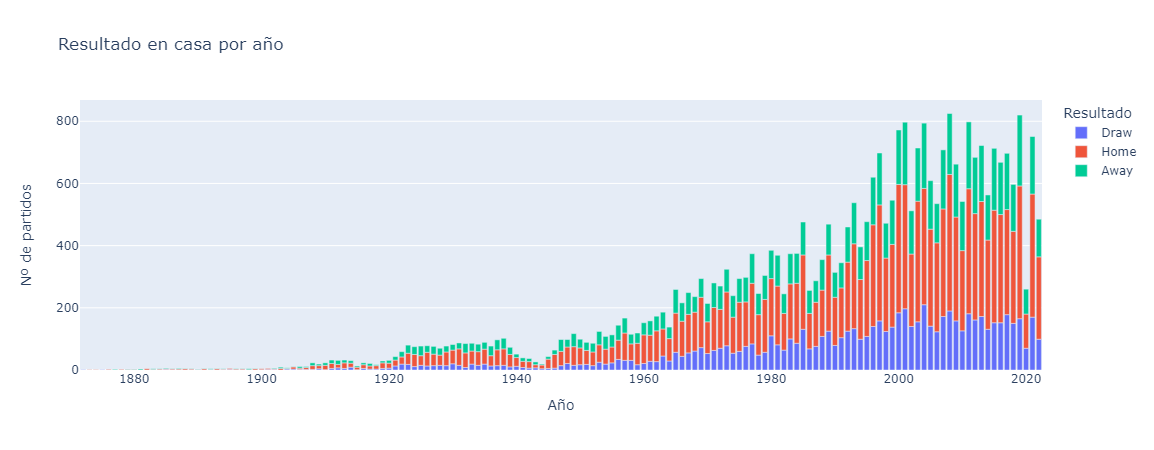


**Tasa de victorias.**

Hemos creado una función que compara 2 columnas, la del equipo local y la de los ganadores. Si ambos nombres coinciden nos da un return diciendo que gana el local, si el ganador es distinto al local nos sacará que el ganador es el visitante, y si lee que el resultado fue empate, obviamente nos dará un empate. Todo esto lo ponemos en una columna nueva para poder analizarlo. Simplemente agrupamos en esta nueva columna y contamos las veces que aparece cada valor.



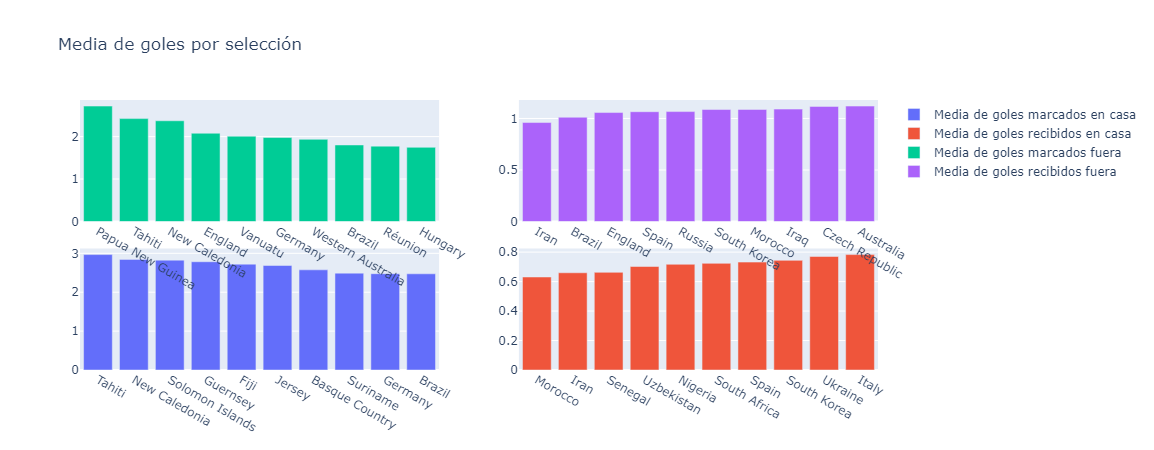
Para terminar con este apartado hemos analizado lo mismo de antes, pero esta vez el número total por año. El resultado es un stacked bar chart con la misma forma de los partidos jugados por año, pero poniendo por colores el resultado del número de partidos correspondiente dentro de cada barra.



**Análisis de selecciones.**

En esta sección hemos querido analizar 4 casos: media de goles marcados en casa, media de goles marcados fuera, media de goles encajados en casa y media de goles encajados fuera. La intención era comparar estos 4 gráficos y ver si había alguna relación entre ellos, y de ser así, si esto podría significar ser una selección top.

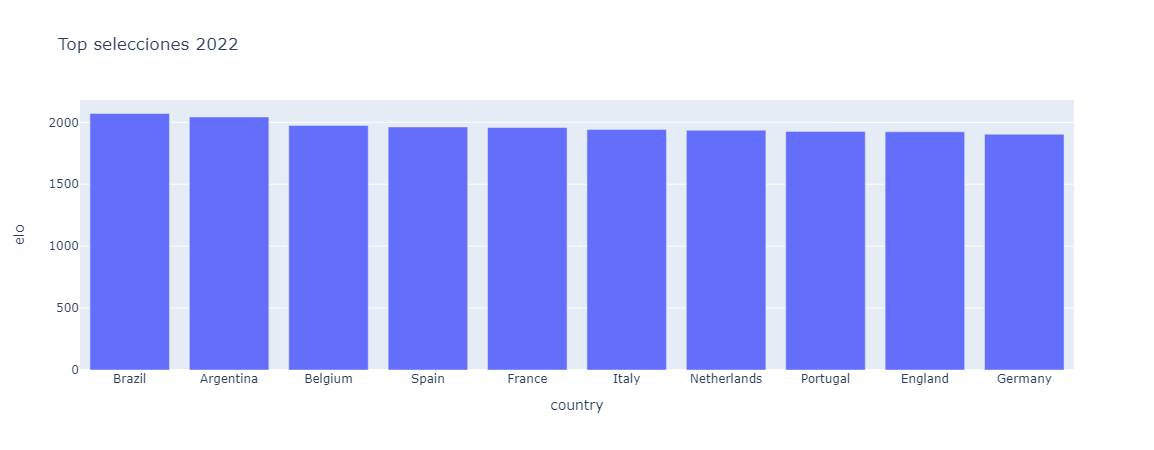
Para conseguir esto se han hecho varios groupby con las mismas columnas pero en diferente orden para representar cada cosa hasta al final obtener la siguiente gráfica. Nos interesa representar los que más goles marcan y los que menos goles reciben en ambos casos. Para tener algo coherente hemos eliminado selecciones con menos de 30 partidos jugados, ya que esto hacía que saliesen cosas poco coherentes.

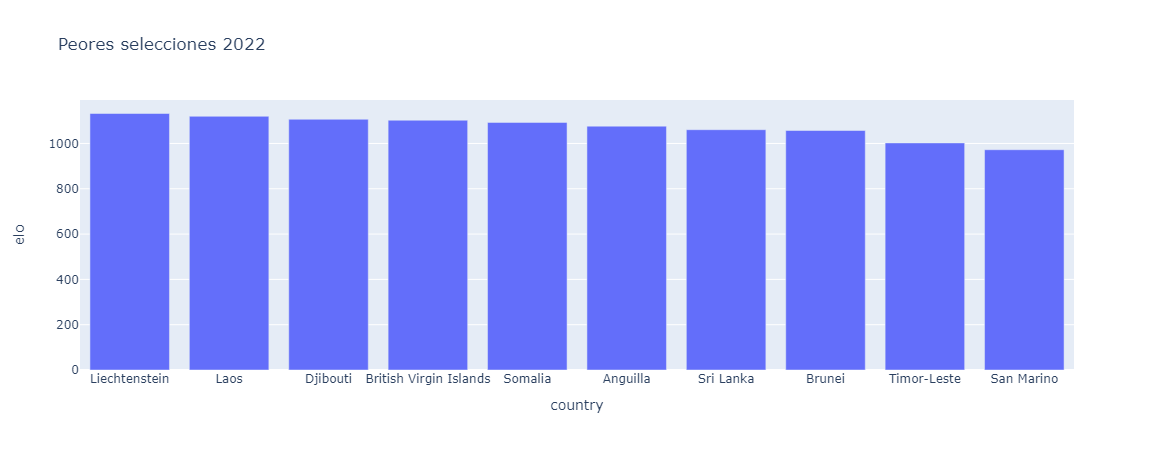


**Análisis por ranking.**

En este último apartado hemos utilizado la fórmula de cálculo de ELO que usa la FIFA para hacer su ranking. Básicamente ha sido implementar [esto](https://en.wikipedia.org/wiki/World_Football_Elo_Ratings) a nuestro dataset para poder tener algo en lo que basarnos.

La función funciona de maravilla, y saca todo lo esperado. El problema viene a la hora de representar los datos (fallo del humano, no de la máquina). Si filtramos un año concreto, agrupamos y ordenamos para representar tenemos unas gráficas bastante buenas y fieles a la realidad, lo que nos indica que esto funciona.





El problema viene a la hora de querer hacerlo no para un año, si no queriendo ver los máximos, con lo que sacamos lo siguiente:



Sólo salen los últimos años. En el momento de hacerlo no me pareció tan mal porque tiene un poco de sentido, ya que de las que aparecen ahí 3 ganaron el mundial en su año, España aparece otros 3 años porque dominó esa era, y los demás son recientes. La verdadera interpretación es que, aunque España sí que aparecería en nuestra intención, otras no. Esto se debe a que hemos creado un sistema que es muy bueno para comparar selecciones en el mismo año, pero no selecciones de diferentes años. El cálculo que realiza es un valor al que suma o resta otro dependiendo del resultado. Si una selección continúa ganando muchos partidos menores, al final sube inevitablemente y eclipsa a otras que en su año lo hicieron muy bien pero el valor cambió después teniendo un rendimiento peor. Esto hizo que se cambiara el camino de buscar las mejores selecciones de la historia, a las mejores recientemente.

**Comentarios adicionales.**

La intención es seguir trabajando en este proyecto, añadiendo o cambiando cosas. Sobre todo, cambiar algunas gráficas y añadir muchas otras. Pero lo principal es intentar retomar la idea inicial de buscar las mejores selecciones de la historia y representar la jerarquía de cada año, y también intentar realizar todo esto en Streamlit.