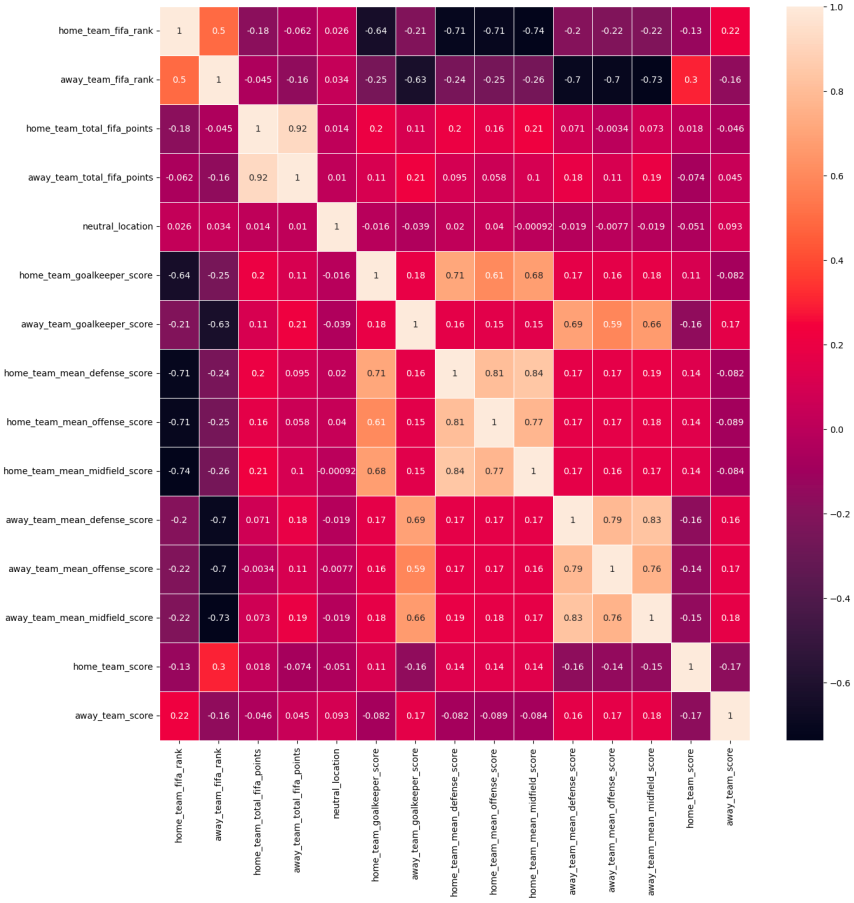
El Mundial de la FIFA es el torneo de fútbol más prestigioso del mundo. El campeonato se ha disputado cada cuatro años desde el inicio del torneo en 1930. El formato actual implica una fase de clasificación, que tiene lugar durante los tres años anteriores, para determinar qué equipos califican para el torneo. En el torneo, 32 equipos, incluido el país anfitrión, compiten por el título en diferentes estadios del país anfitrión.

En esta práctica se van a desarrollar modelos de regresión, con los que vamos a intentar predecir el resultado de todos los partidos que se van a disputar. El dataset que se va a usar se ha obtenido de la plataforma Kaggle y cuenta con datos de todos los partidos jugados desde 1993.

Como variables objectivo, se han elegido las variables **home_team_score** y **away_team_score**. A partir de estas dos variables, se puede predecir el ganador del partido y, mediante la ejecución de cuadro del torneo, llegar a predecir el ganador del mundial. Por lo tanto, diremos que es un problema de regresión lineal sobre dos variables.

De este dataset de 23921 filas, hay muchos equipos que no participan en el mundial, pero se van a usar indistintamente para entrenar los modelos de regresión. Cabe destacar, que algunas de las columnas tienen algunos valores que no existen y, por lo tanto, vamos a hacer una primera limpieza de estos elementos, y al final, nos quedaremos con un total de 4303 elementos. Podemos encontrar el dataset en Kaggle.com.

La matriz de correlación de los datos es la siguiente:



Y la comparación de las variables con las variables objetivos se puede representar gráficamente en: