# The Sofifa.Com Scraper

## Overview

This project involves building a playwright scraper to extract information from sofifa.com website and deliver the dataset as standardized and cleaned. It involves extracting player information from the website, getting the required information, cleaning and de duplicating the dataset. The information required were

***Player Name, Player Fullname, Player Age, Player Birthdate, Player Height, Player Weight,***

***Player Position, Player OverallRating, Player Potential, Player Value, Player Wage, Player Club Name,***

***Player League, Player Nationality, Player NationalTeam***

## Process

The project was executed in two major phases, which were ***The Extraction Phase*** and ***The Standardization Phase.***

**The Extraction Phase:** This phase involves building the scraper, testing and using it to extract the required information. Network errors, missing information and changes in structure were effectively handled to ensure a smooth extraction process. This phase was mainly a programming and automation phase. The tools involved were python, using libraries like Playwright and BeautifulSoup.

**The Standardization phase:** This phase involves cleaning the extracted data and standardizing columns to match the data type. Every single columns was examined, new lines and space characters were properly handled. The phase was done using Jupiter notebook and python. Libraries used were Numpy and Pandas

## The Result

A cleaned and very presentable dataset exported to CSV format. The dataset was well standardized and can be used for further analysis or even easily inputting into any database. This project was completed in 2 days.