Brendan Egan

Seth Berry

Unstructured Data Analytics

February 26th, 2026

PAY THAT MAN!!!!!

What is the Project:

The goal of the project is to scrape through Understat.com using the soccer portion of the website and find the ‘best’ teams across European soccer. I wanted to look through the ‘top 5 European Leagues: England’s Premier League, France’s Ligue 1, Spain’s La Liga, Germany’s Bundesliga, and Italy’s Serie A. I collected each season’s data including individual player statistics and put those into a .csv file that was then used to create visualizations. Once I had this data scraped across eight seasons, I made visualizations showing the most dominant teams and players.

What is the Problem:

The issue was that Understat hated me attempting to scrape through its individual pages on the site. That was a huge problem, but after some research I was able to find the UnderstatAPI to fix that issue. Another major issue I ran into was that the individual player data does not contain contract valuations and amounts, so I was unable to complete the original goal of the project. The only publicly available data are from players’ agents who announce those figures or clubs

revealing it for FFP (Financial Fair Play) purposes. So, that ruined my chances of pulling all that info in one swoop. The final issue I had was when I scraped the 2014/2015 season of data, I learned that xG, xA as well as advanced statistics like that were not tracked, which ruined an entire session of scraping data.

How was it solved:

The original issue of not being able to complete a full scrape was solved with learning and using UnderstatAPI and a few tips I learned from the documentation (it was La_Liga not La_liga) in order to scrape all the leagues. My first attempt was to scrape FBRef for the player contract amounts on top of stats, but oh my goodness CloudFlare does not play around. I could not get a single thing to scrape without it screaming that I was a bot doing nothing good. Being forced to pivot to scraping Understat.com was a blessing in disguise, because I got to learn how to use the understat package in python and got to create a brand new scraper. I removed the 2014 set of data and changed my scraping boundaries to the 2015-2023 seasons (2015/2016 - 2023/2024) so I would have data in all my columns and rows.

Why does it matter:

I want to use this project in the future to determine if a player is worth a new contract or determine if they are 'falling off' in their ability. Players who exceed their Expected Goal and Assist rates (xG & xA) will often be overpaid on the premises that they played well one season, but the very next will regress to the expected rate of their performance. Scouts and soccer front office staff can use this model to predict how players will perform in the future. My plan is to work with this data I have collected and begin to compare stats to contract valuations and current

wages. I still need to scrape/collect data for individual contracts, but I want to be able to predict when a player deserves their big pay day or when a player is being overpaid.