Feasibility Study on Football Transfer Data

Group 103

Samuel Amuzu, Julian Purtschert, Nick Schaufelberger

Introduction

As passionate sports enthusiasts, our team has chosen to focus our CIP project on sports data analytics. Specifically, our project aims to analyze trends and patterns within the European football leagues and transfer markets to better understand player movements, transfer values, spending patterns, and win ratios in relation to the final standings of each season.

We plan to collect data from various publicly available online sources through web scraping and publicly available APIs. This will include league results, club performance statistics (such as goals scored, goals conceded, matches played, and total points per season, we are planning on the seasons from 2009 to 2021).

To enhance our analysis, we plan to supplement this dataset with additional data. We will utilize a publicly available dataset from Kaggle, which contains detailed transfer records from major European leagues between 2009 and 2021, including:

- English Premier League
- La Liga
- Serie A
- Bundesliga
- French Ligue 1
- Liga Portugal Bwin
- Dutch Eredivisie

This project is feasible due to the availability of comprehensive, open-source datasets, the accessibility of data through web scraping, and the team's combined experience in data analysis and sports analytics tools.

Research Questions

To give us some more structure and guide us through our project, we determined the following 3 research questions, which we will be answering:

Transfer patterns throughout the leagues in general (All leagues)

This part examines player transfers across multiple leagues and seasons, focusing on player origins, transfer fees, club and leagues spending behavior. It explores patterns in nationality and position dominance and analyzes spending trends across different transfer windows and years. (focus: Top 4 Position, CB, CF, CM, GK, amount of money spent in each League and their differences)

League characteristics and transfer spending

This research question investigates how transfer market dynamics such as player age, transfer type, and league characteristics influence transfer spending patterns and their relationship with team performance across major European football leagues with a focus on mainly the Premier League, La Liga and Italian League across seasons from 2009 to 2021. The research will examine how different leagues balance between developing, selling and acquiring players of different age groups, and how these trends evolve over time. The study further explores which factors, such as player age, market value, and transfer type,

explain variations in the likelihood and value of player age groups (mainly prime and youth players) transfers. Finally, the analysis assesses the broader impact of player transfer activity on league development, focusing on whether higher player transfer rates contribute to greater competitive performance across leagues and player groups over time.

How do La Liga (Spain) and the Premier League (UK) differ in terms of transfer spending, and how does this impact the final league standings?

In this research question, we examine two of Europe's most prominent football leagues, Spain's La Liga and the UK's Premier League, to compare how their clubs perform, how transfer spending varies between them, and how these spending patterns relate to the final league standings. For this analysis, we used two datasets, one static dataset obtained from Kaggle, and another scraped dataset containing the final standings for each season along with the corresponding numbers of transfers and total amounts spent for the seasons 2013-2015.

Description of key data sources we plan on using

Our dataset, which is in addition to the scraping part, is the Football Transfer dataset from Kaggle. It contains detailed information on player transfers, fees, clubs, seasons, and positions for the biggest leagues in Europe from 2009 to 2021. The dataset consists of 23 columns and more than 70'000 entries.

To enrich the dataset, we plan to scrape additional player and club information from different websites or use the one that proves to be the most suitable. At the moment, we have identified three potential sources to scrape from: Football365, Wikipedia (which would require scraping each season individually), and FBref. For this, we plan to use tools such as BeautifulSoup or Selenium, which have been presented in class.

This approach provides a dynamic foundation with individually selected and enriched data for further analysis. Through scraping, we aim to collect information about each club's season, including possible attributes such as Rank, Team, Matches Played, Wins, Draws, Losses, Goals For, Goals Against, Goal Difference, Points, Points Per Match, Expected Goals, Expected Goals Against, Expected Goal Difference, Expected Goal Difference per 90 Minutes, Average Attendance, Top Goal Scorer, Goalkeeper.

Potential risks relating data collection & quality

A key risk in the data collection process relates to data quality, consistency, and accessibility. Differences in structure, formatting, and data availability across sources such as Football365, Wikipedia, and FBref may introduce inconsistencies or gaps in the resulting dataset. Additionally, changes to website layouts or access restrictions can disrupt automated scraping pipelines.

To mitigate these risks, we plan to perform data extraction in a single run, minimizing dependence on repeated scraping. Since our dataset spans static historical data (2009–2021), this one-time collection is sufficient and appropriate. Furthermore, we will review each site's robots.txt file to ensure compliance with ethical and legal data collection practices.

To further improve reliability and consistency, we will validate and clean the data according to generally known standards, cross-verify and cross-check the data to ensure consistency, document the extraction process to ensure reproducibility and transparency, as well as store backups of the raw data scraped, in case we need it to re-use if any issues occur in a later stage of the project.