Analyzing Transfer Patterns and Spending Behavior in European Football Leagues (2009-2021)

A Data Analytics Approach to Understanding Transfer Market Dynamics

https://wettbonus.net/sportwetten-magazin/fussball-ligen-europa/, 2nd of November 2025

Group 103: Samuel Amuzu – Julian Purtschert – Nick Schaufelberger

Course: W.MSCIDS_CIP02.H25 – Data Collection, Integration and Preprocessing

Submission Date: 7th of November 2025

Semester: Autumn Semester 2025, Lucerne, Switzerland

Lecturer / Supervision: Andreas Melillo

GitHub: https://github.com/S-amuzu/CIP_HS2025_103.git

GitHub URL: https://github.com/S-amuzu/CIP_HS2025_103

Comment on the use of AI

AI tools, such as ChatGPT, were used to support various parts of this project, mainly for refining Python code related to data cleaning, visualizations and statistical testing, as well as for improving the readability and structure of the report. The visual and statistical analyses themselves were fully conducted and interpreted by the project team. All AI-assisted outputs were critically reviewed, adjusted to fit the use-case, and validated to ensure accuracy and alignment with our own findings.

1 Introduction

1.1 Overview of the project

Our project focuses on analyzing trends and patterns within the European football transfer market between 2009 and 2021. Using a combination of publicly available dataset and web-scraped data, we examine how transfer activities, player movements, and spending behaviors vary across leagues and influence club success. The analysis centers on identifying key differences between leagues, exploring spending patterns of Premier League clubs, and understanding the relationship between transfer investments and team performance. Through this data-driven approach, we aim to uncover insights into how financial strategies shape outcomes in European football.

1.2 Motivation

As passionate football and data enthusiasts, we wanted to combine our interests in sports and analytics to better understand the dynamics behind football transfers. The football transfer market offers rich, publicly accessible data that allows for deep exploration of patterns and behaviors across clubs and leagues.

2 Methods

2.1 Data source and creation

The primary dataset used in this project is a publicly available football transfer dataset sourced from Kaggle. It contains detailed transfer records from Europe's top leagues (Premier League (UK), La Liga (Spain), Serie A (Italy), Bundesliga (Germany), Ligue 1 (France), Primeira Liga (Portugal), Eredivisie (Netherlands)) between the years 2009 to 2021.

2.2 Transformation and cleaning steps

2.2.1 Checking for missing data and data types

The first step we inspected the structure of the dataset and identified missing values and incorrect data types. Five columns were found to contain missing values. These columns include player_age, player_nation, player_nation2, transfer_fee_amnt,and market_val_amnt. Since many players did not have secondary nationalities or were transferred for free or on loan, it resulted in missing transfer fees and market values. These gaps were considered valid. However, the missing values in key analytical variables such as player_age and player_nation, we decided to remove the rows that were missing for these features because they were considered important to us for further analysis. The player_nation2 column was not interesting for us, so it was removed. Data types were then converted to the required data types for analysis. We converted categorical fields such as league, team_name and player_pos to the *category* type and numerical fields such as player_age and counter_team_id to integer types. To enhance dataset consistency, missing values in transfer_fee_amnt and market_val_amnt were filled with zeros. The dataset was also checked for duplicates which contained none.

2.2.2 Inspecting Unique Values and Correcting Inconsistencies

Next, categorical variables were inspected to identify any inconsistencies or spelling mistakes. The player_pos column showed irregular entries. We were expecting specific player positions (CB, RB, CM) but received a general position for some of the players (attack, midfield, defence) so we decided to remove these rows from the data. These accounted for less than 0.4% of the dataset and were mostly from the Italian league. Our primary focus for our analysis will be based on specific positions. Therefore, these rows were safely removed without affecting representativeness. Unused categories were also dropped to keep the variable space clean and to ensure efficient use of memory.

2.2.3 Handling Outliers and Anomalies

Several numeric columns such as player_age, transfer_fee_amnt, and market_val_amnt were assessed for outliers. Using the describe() method and boxplots, unrealistic values were identified. One clear anomaly was a player recorded as 1,775 years old, which was obviously incorrect. This entry was removed. Transfer fees exceeding €250 million were also flagged as outliers because according to our football knowledge, Neymar has been the player with the highest transfer fees. Hence, rows with identical inflated values of €500 million were concluded as errors and deleted. The top legitimate transfer, Neymar's €220 million move, was retained as a realistic benchmark. Similarly, for market value, no extreme anomalies were detected after cross-checking top players (e.g., Kylian Mbappé, Lionel Messi) against known football knowledge.

3 Findings and Discussions of results

3.1 Research Question 1

"What patterns can be observed in transfer spending across Europe's top football leagues from 2009 to 2021, and how do nationality and player position influence transfer activity?"

To examine how nationality and player position influence transfer patterns, we analyzed the distribution of incoming players across Europe's top leagues and visualized the relationship between nationalities and playing positions. The results reveal clear structural tendencies in how talent moves across various football markets.

Across all major leagues, domestic players make up the majority of transfers, reflecting strong reliance on national development systems and local recruitment networks. For example, Spanish, Italian, English, and German players are the dominant groups within their respective leagues. However, a consistent cross-league trend emerges with Brazilian players ranking among the top five nationalities in almost every league, underscoring Brazil's role as one of the world's leading exporters of football

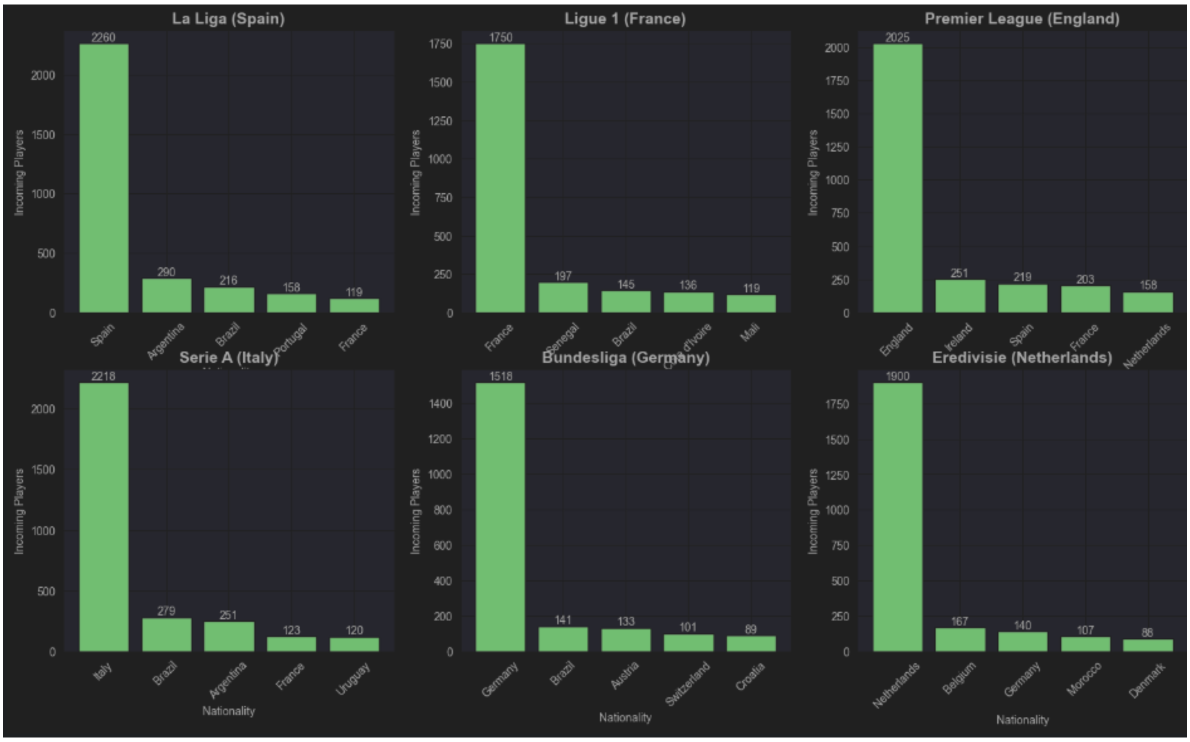

Figure 1: Top 5 player nationalities across European major leagues

talent worldwide. Other frequent contributors include Argentina, Portugal, and France, highlighting potential regional and linguistic ties between European and South American football markets.

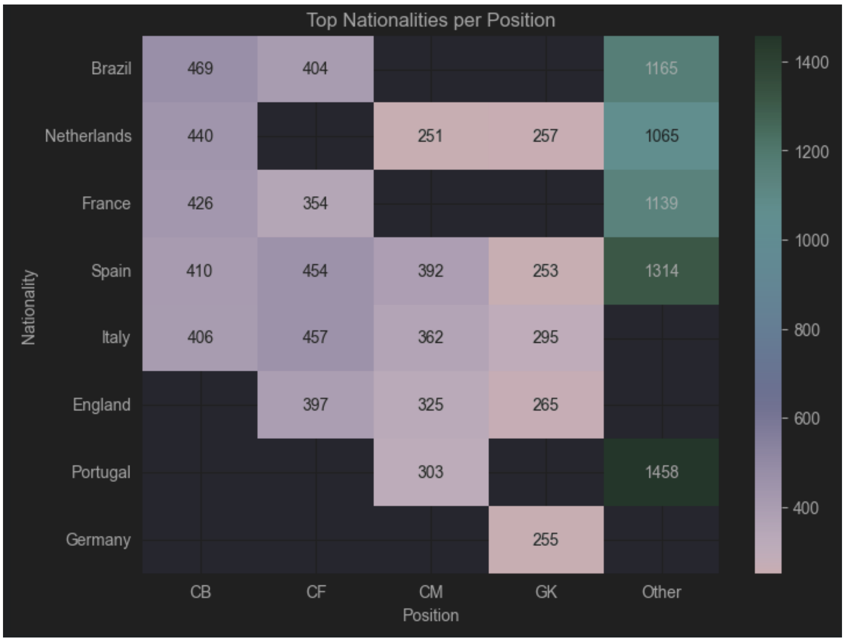

Figure 2: Heatmap showing top 5 nationalities by player position

When examining these nationalities by position, further points become visible. After determining the top four incoming transferred positions (CB, CF, CM, GK and Other) we created a heatmap to showcase the combination in both positions and nationality.

Brazilian and French players appear especially strong in both defensive and attacking roles (CB, CF), indicating their versatility overall. Italy and Spain exhibit a balanced distribution across nearly all positions. Germany shows a specific strength in developing Goalkeepers. Portugal appears very strong in the "Other" category. This category contains positions not accounted for in the top four, such as Wingers and Fullbacks for example. In this heatmap, empty cells indicate nationality-positions that did not rank among the top five nationalities for that position. They do not imply zero players but simply fall outside the top five threshold. Overall, the combined findings suggest that national football cultures and developmental strengths shape positional export patterns.

3.2 Research Question 2

"How do transfer market dynamics particularly player age, transfer type, and league characteristics influence the number of transfer activity, transfer spending patterns and their relationship with team performance across major European football leagues and seasons, and how do these patterns differ across age groups and transfer directions?"

The analysis reveals that transfer market behavior in European football is heavily influenced by player age, transfer direction (either incoming or outgoing) and whether it is a domestic or foreign transfer across different leagues. The results show that the Premier League and Serie A operate differently in the transfer market.

In the Premier League, transfer activity is dominated by incoming prime-age players, reflecting a strategy built on acquiring young, experienced players (prime) rather than focusing on long-term development on youth players. The Premier League is backed by vast financial resources and global revenue streams; this could be a potential factor why they are able to purchase players in their peak years at inflated market prices. This approach drives transfer fees, particularly for foreign players, where competition for talent is high. The league's weak positive relationship between spending and performance suggests that investment in the transfer market leads to better performance results. However, further analysis should be done across many leagues to draw stronger conclusions from this hypothesis.

In contrast, as shown in Figure 3 and 4, Serie A shows the highest number of outgoing players in the transfer market of both youth and prime players as a financial sustainability mechanism. Italian clubs often rely on selling youth prospects but marketable prime players to balance budgets, illustrating a “selling league” dynamic. Although Serie A exports a lot of players during the transfer window, despite this, the league shows lower average transfer fees compared to the Premier League. This shows the financial strategy in Serie A focuses more on maintaining liquidity and squad depth over high-cost acquisitions. The Linear model visualization (shown in the Jupyter Notebook) also revealed that heavy spending in Serie A was not necessarily associated with improved performance, possibly due to the Leagues' focus on domestic transfers rather than foreign transfers and less focus on acquiring prime players at higher transfer fees.

Across leagues, youth outgoing transfers were consistently more valuable than incoming youth transfers, suggesting that developing and selling young talent represents a major source of profit for smaller clubs. These players were probably acquired for minimal fees or developed through the academies and later sold to wealthier leagues like the Premier League for higher transfer fees than their market values. This shows a financial reliance of smaller leagues on bigger.

In contrast, The Poisson regression model showed a different pattern for prime-age players compared to youth players. Incoming prime transfers were more expensive, while outgoing prime transfers were relatively stable or even undervalued, reflecting the limited resale potential of players at or near their peak. Leagues buying prime players tend to overpay for performance certainty, experience or competition from other leagues. The regression model confirms these findings. However, further analysis may be required to determine to draw solid conclusions on the league level since different clubs operate differently. Another limitation is that other factors such as contract lengths are not considered.

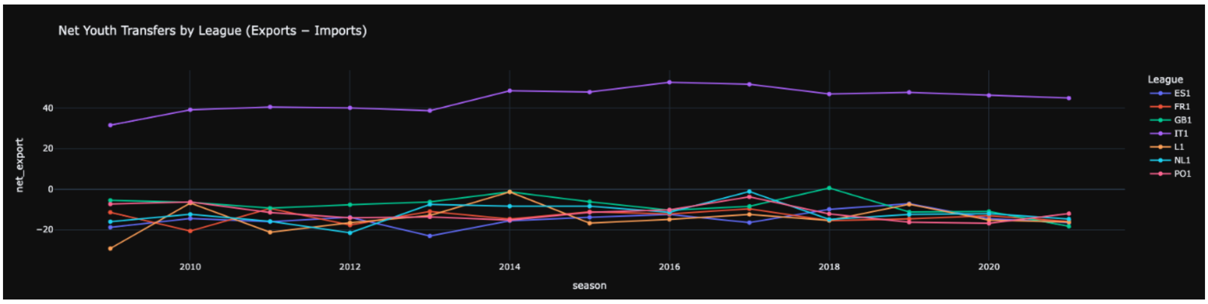

Figure 3: Net Balance in % (difference between transfer exports and imports for youth players across leagues over time)

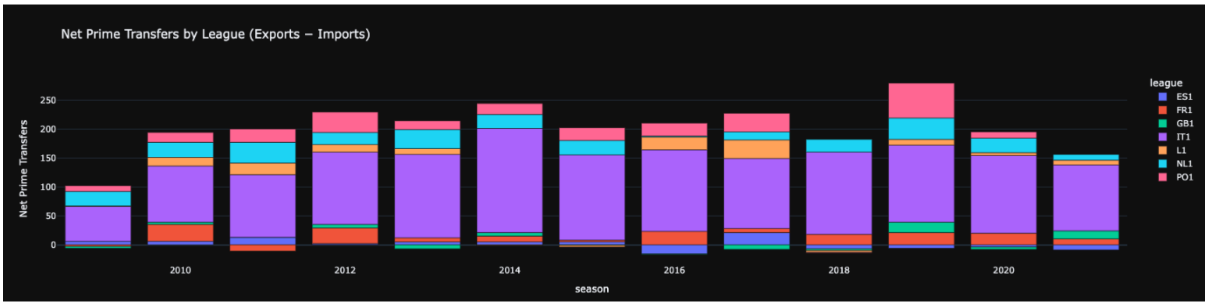

Figure 4: Net Balance (difference between transfer exports and imports for prime players across leagues over time)

3.3 Research Question 3

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

The analysis of scatterplots (Figure 5) from Spain's top football league reveals a consistent pattern. Typically, three to four clubs spend substantially more on player transfers than the rest. These higher-spending teams generally achieve better league positions, confirming a possible positive relationship between spending and final rank. The 2015 season stands out as an exception, where the top-spending club finished around 12th place.

To quantify this relationship, Spearman's rank correlation was used to measure the association between total transfer spending and final league ranking for the 2013-2015 seasons. This non-parametric method is appropriate because it assesses rank-based relationships. The results indicate that higher spending tends to correspond with better rankings, though the method does not account for the actual magnitude of spending differences or imply causation. It should also be kept in mind that the sample size is quite small, covering only three seasons. For future evaluations, an analysis including all seasons would be more insightful.

The poisson model analysis shows a strong negative correlation between transfer spending and league ranking, indicating that higher-spending clubs generally finish higher in the table. However, this relationship is not causal, factors such as squad quality, management, and injuries also influence outcomes, and some overdispersion and limited data across seasons warrant cautious interpretation.

Looking at the Premier League plots, our first impression is that the points are more scattered, with no clear indication that the clubs spending the most money consistently finish at the top, at least in the three seasons we looked at. What we can say is that, compared to the Spanish league, the Premier League features more clubs that spend large amounts of money, rather than just four dominant teams as in Spain. Overall, the patterns differ noticeably between the two leagues. What makes the conclusions difficult is that we can't really compare the different leagues with each other. There are many factors that differ, and the Spearman correlation and the results from the Poisson model also cannot be compared one-to-one in this case.

The premier league results show that the relationship between total transfer spending and league ranking varies across seasons. In some years, higher spending is linked to better league performance, while in others, there is no clear connection. An exceptional case, such as Leicester City's performance, further illustrates this inconsistency.

In the two scatterplots, you can see how the final standings compare to the teams' spending over the three seasons from 2013 to 2015. The Premier League plot on the left appears more scattered than the La Liga plot on the right, indicating that mid-table La Liga clubs spend less and that the league might be dominated by three or four big clubs.

Overall, although greater spending often coincides with stronger results, this pattern is not consistent over time. The findings do not suggest a causal relationship, as success may also depend on factors like squad quality, management, or injuries. Moreover, the limited sample size adds uncertainty to the strength and generalizability of these conclusions. Ultimately, sport remains an inherently unpredictable activity in which outcomes are influenced by numerous factors beyond financial considerations.

Extraordinary events such as Leicester City's 2015 Premier League season demonstrate the potential for unexpected success and highlight the intrinsic beauty and analytical appeal of sport.

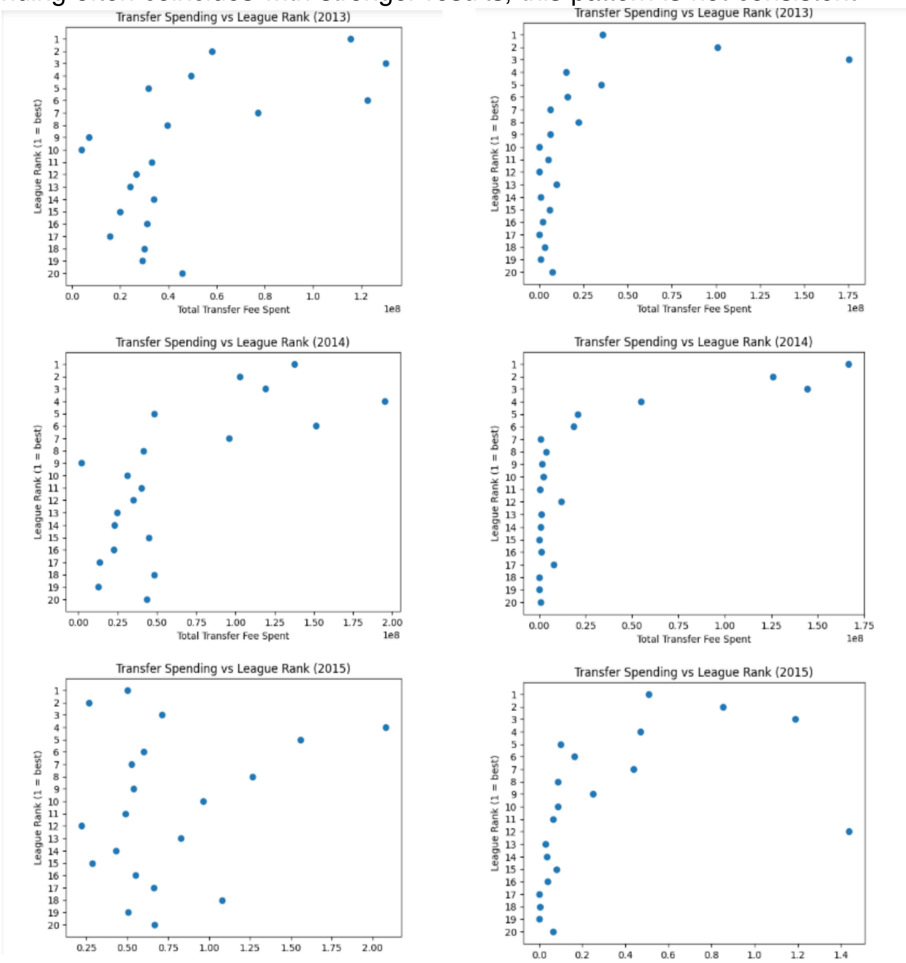

Figure 5: Scatterplots comparing club spending and final standings (2013-2015)

4 Conclusion

Through our analysis of transfer data from 2009 to 2021, we gained valuable insights into how spending behaviors, player nationalities, and positional trends shape the European football market. We learned that while financial investment often correlates with performance, it does not guarantee success, as other factors like player development, club strategy, and market timing play equally important roles. Limitations of our study include mainly restricted data coverage on contract details and various performance metrics, as well as potential inconsistencies in web-scraped data. For future work, expanding the dataset to include variables such as player salaries, contract lengths, and performance indicators could deepen the understanding of transfer market efficiency throughout seasons, leagues and standings during the season and enable predictive modeling of transfer success. We also encountered several difficulties while programming together and working with web scraping. When combining our code and using different IDEs and package versions, we ran into dependency issues and inconsistencies in output formats. Although the scraping worked well until the last few days, we ultimately had to implement an emergency solution to avoid changing too much of the existing code. Overall, this project gave us valuable insight into the challenges of completing a full data science project and workflow from start to finish.