Determinants of Soccer Players' Market Values in Major European League(s)*

Goals, Assists, and Age Dominate, While League-Specific Patterns Add Complexity

John Zhang

December 3, 2024

The global soccer transfer market is shaped by diverse factors influencing player valuations across leagues. This study analyzes determinants of market value in five major European leagues using data from the 2023/24 season. Goals and assists consistently emerge as the strongest predictors, with the Premier League valuing goals highest and the Bundesliga emphasizing assists, while age negatively affects value across all leagues, particularly in La Liga. These observations highlight universal trends and league-specific valuation dynamics, offering actionable guidance for clubs, agents, and analysts.

Table of contents

*Code and data are available at: https://github.com/Clearsky21z/Player_Market_Value_Analysis

1 Introduction

The global soccer transfer market is a high-stakes, multi-billion-Euro industry where decisions on player valuations can significantly influence team-building strategies, investment decisions, and negotiations. Player market values are determined by a mix of tangible performance metrics, such as goals scored or assists provided, and intangible factors like potential, reputation, and team context. While certain predictors are universally recognized, such as offensive contributions and age, the relative importance of these factors varies widely across leagues with differing financial capabilities, tactical preferences, and competitive levels. Understanding the dynamics behind these variations is important for clubs, agents, and analysts aiming to optimize their strategies in this complex market.

This study examines the determinants of soccer players' market values across five major European leagues: Premier League(England), La Liga (Spain), Serie A (Italy), Bundesliga (Germany), and Ligue 1 (France). These leagues not only dominate the global football landscape but also present unique economic and competitive ecosystems, making them ideal for a comparative analysis. Using player data from the 2023/24 season, sourced from platforms such as Transfermarkt (2024), Stathead (2024), FIFA (2024), and FootballDatabase (2024), this research addresses a vital gap: how do the drivers of player market value differ across these leagues?

The estimand in this study is the expected market value of a soccer player as a function of performance metrics (e.g., goals, assists), contextual variables (e.g., club and national team rankings), and demographic factors (e.g., age, position), while accounting for league-specific differences. This estimand represents a linear relationship between player attributes and market value, enabling a comparison of the impact of various predictors across leagues.

To conduct this analysis, the study employs league-specific linear regression models with player market value as the dependent variable. Key predictors include performance metrics (e.g., goals, assists, and minutes played), contextual factors (e.g., club and national team rankings), and demographic variables (e.g., age and position). The study relies heavily on the statistical computing environment R Core Team (2023) and the programming language Python Software Foundation (2023), employing a suite of packages to ensure methodological rigor and transparency:

- **Data Retrieval, Cleaning and Validation**: The Python library `BeautifulSoup` was used to scrape market value and player performance data from Transfermarkt and other sources. In R, the `arrow` package by Richardson et al. (2024) facilitated seamless data reading and processing, the `tidyverse` collection by Wickham et al. (2019), including `dplyr` package by Wickham et al. (2023) for data manipulation and `ggplot2` package by Wickham (2016) for data visualization, formed the backbone of the analysis, while `validate` package by van der Loo and de Jonge (2021) and `testthat` package by Wickham (2011) ensured the integrity and accuracy of the dataset. Additional tools like

`styler` package by Müller and Walthert (2024) enhanced code readability and reproducibility.

- **Modeling and Summaries**: Regression models were built using the `caret` package by Kuhn and Max (2008) for efficient training and testing, while the `broom` package by Robinson, Hayes, and Couch (2024) was employed to tidy model outputs, extract coefficients, and augment datasets with residuals and predictions, enabling detailed validation and diagnostic analysis. The `modelsummary` package by Arel-Bundock (2022) provided clear, publication-ready tables summarizing model outputs.
- **Dynamic Report Generation**: The `knitr` package by Xie (2024) and `kableExtra` package by Zhu (2024) were integral to generating dynamic, visually appealing reports that integrated results, tables, and figures.

The results show both universal and league-specific patterns in player valuation. Offensive contributions, such as goals and assists, consistently emerged as significant predictors, with goals having the strongest influence in the Premier League and assists being most impactful in the Bundesliga. Age also played an important role, with younger players commanding higher valuations, particularly in La Liga. Contextual variables like club and national team rankings demonstrated notable variation, reflecting the differing emphasis on team prestige across leagues.

This study contributes to the literature on soccer economics by uncovering how market dynamics vary across Europe's top leagues. For clubs, the findings provide actionable insights for tailoring recruitment and investment strategies. For agents, the results highlight the most marketable player attributes for negotiations. Methodologically, the integration of elevated statistical and programming tools offers a replicable framework for analyzing market dynamics in soccer and other sports.

The paper is structured as follows: Section 2 describes the dataset and variable construction processes, emphasizing the integration of data from multiple sources. Section 3 details the modeling approach and validation techniques. Section 4 presents the empirical results, identifying key differences across leagues. Section 5 discusses the implications of these findings, highlights the limitations of the study, and outlines future research directions. Section A provides an in-depth examination of data cleaning methodologies, in-depth documentation of observational data collection processes, datasheets for each dataset, and full model results. By bridging methodological rigor with actionable insights, this study advances our understanding of the determinants of soccer players' market values.

2 Data

2.1 Data Sources and Construction

The dataset was created by merging multiple data sources using player names as unique identifiers. **Transfermarkt** provided player market values in Euros, which serve as the dependent variable in this study. Performance metrics such as goals and assists were sourced from **Stathead**, a platform that aggregates detailed statistics on football players. To ensure consistency and comparability, rankings were applied to national teams and clubs. National team rankings (1 to 210) were extracted from **FIFA**, while club rankings were sourced from **Footballdatabase**, reflecting both domestic and international performances (See Section A.1 for a detailed walkthrough of retrieving data).

After merging these sources, high-level cleaning was conducted to address missing data and standardize variable formats, and the players who are classified as goalkeepers are dropped (See Section A.3). The final cleaned datasets for each league include seven variables, as outlined below:

- **Market Value**: The player's estimated market value, measured in Euros, as provided by Transfermarkt. This serves as the target variable for the analysis.
- **Age**: The player's age at the start of the season, reflecting their stage in the career lifecycle.
- **Goals and Assists**: Performance metrics capturing offensive contributions, key determinants of player valuation.
- **Club Ranking**: The global ranking of the player's club, reflecting both domestic and international performance.
- **National Team Ranking**: The ranking of the player's national team, highlighting their exposure and representation at the international level.
- **Minutes Played**: The total minutes played by the player during the season, used as a proxy for fitness and importance to the team.
- **Position**: Players are categorized into one of six roles: Defender (DF), Forward (FW), Midfielder (MF), and hybrids such as FWDF, FWMF, and MFDF.

2.2 Data Summary

2.2.1 England Dataset

The England dataset provides a thorough view of player characteristics in the English Premier League. As illustrated in Figure 1, the market value distribution is heavily right-skewed, with the majority of players concentrated in the lower market value brackets (below €50 million). This skewness reflects a superstar effect, where a small subset of players, often international stars or highly sought-after talent, command disproportionately higher valuations.

Figure 2 explores the distributions of key predictor variables, including age, goals, assists, national team ranking, club ranking, and minutes played. These distributions show important patterns in the dataset:

- **Age**: The age distribution peaks around 24-25 years, indicating that the league predominantly features players in their athletic prime. The range extends from 16 years (likely youth players) to 38 years for veteran professionals.
- **Goals and Assists**: Both offensive contributions exhibit a heavily right-skewed pattern, with most players scoring fewer than 5 goals or assists in a season. This suggests that only a minority of players serve as primary offensive contributors.
- **Club Ranking**: Club ranking is inversely related to club performance, with lower values indicating higher-performing teams. The wide range (from 1 to 838) reflects the diversity of clubs in the dataset, from global giants to smaller teams.
- **National Team Ranking**: Similarly, the national team ranking varies significantly, from top-tier nations to lower-ranked teams. Players from lower-ranked national teams may reflect either emerging talent or under-scouted regions.
- **Minutes Played**: Minutes played are distributed across a broad spectrum, with many players logging significant game time (over 1,000 minutes), indicative of their importance to team dynamics.

Figure 3 focuses on the percentage distribution of player positions. Defenders (DF) account for the largest share (32.6%), followed by midfielders (MF) and midfield-forward hybrids (FWMF). Pure forwards (FW) make up 16.4% of the players, with hybrid positions like FWDF forming a smaller subset. This distribution aligns with the league's tactical emphasis on strong defensive and midfield structures, with forwards occupying more specialized roles.

Table 1 provides a detailed statistical summary of the numeric variables in the dataset. The mean market value of players is €20.99 million, but the median is much lower at €14.00 million, reaffirming the skewness of the market value distribution. Other variables, such as assists and goals, also display low averages relative to their maximum values, highlighting the rarity of elite-level contributions.

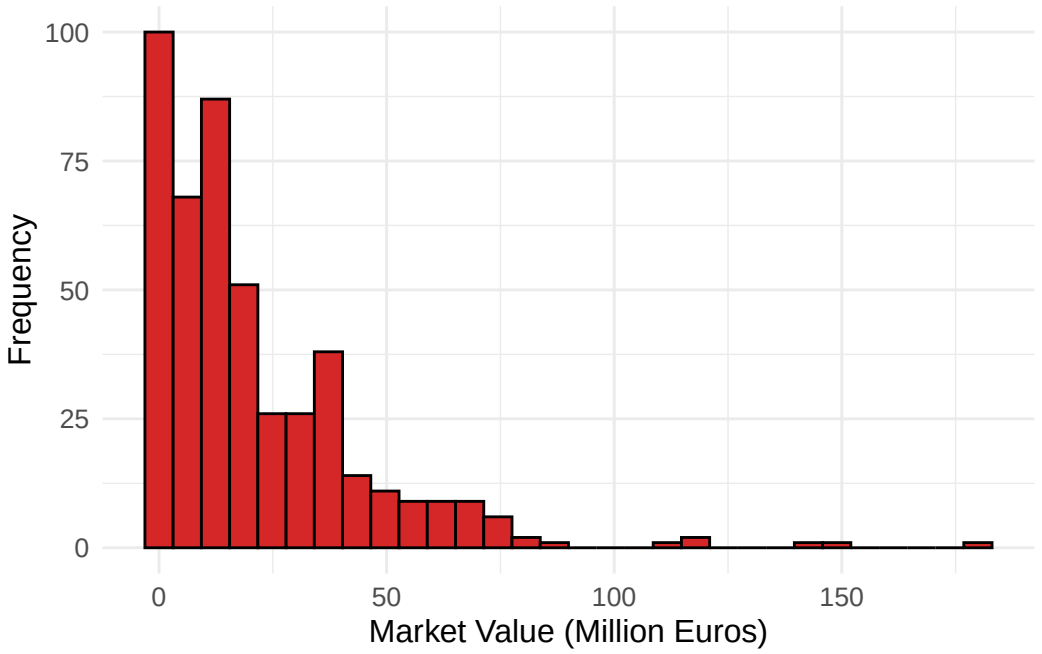


Figure 1: Distribution of Market Value of the England Dataset (in Million Euros)

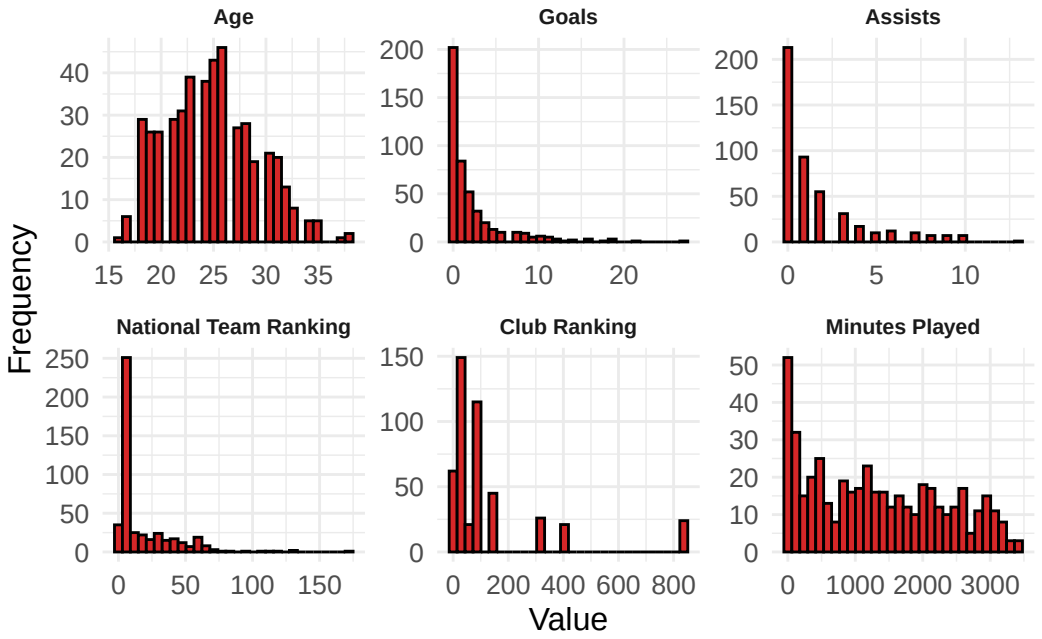


Figure 2: Distribution of Predictor Variables of the England Dataset

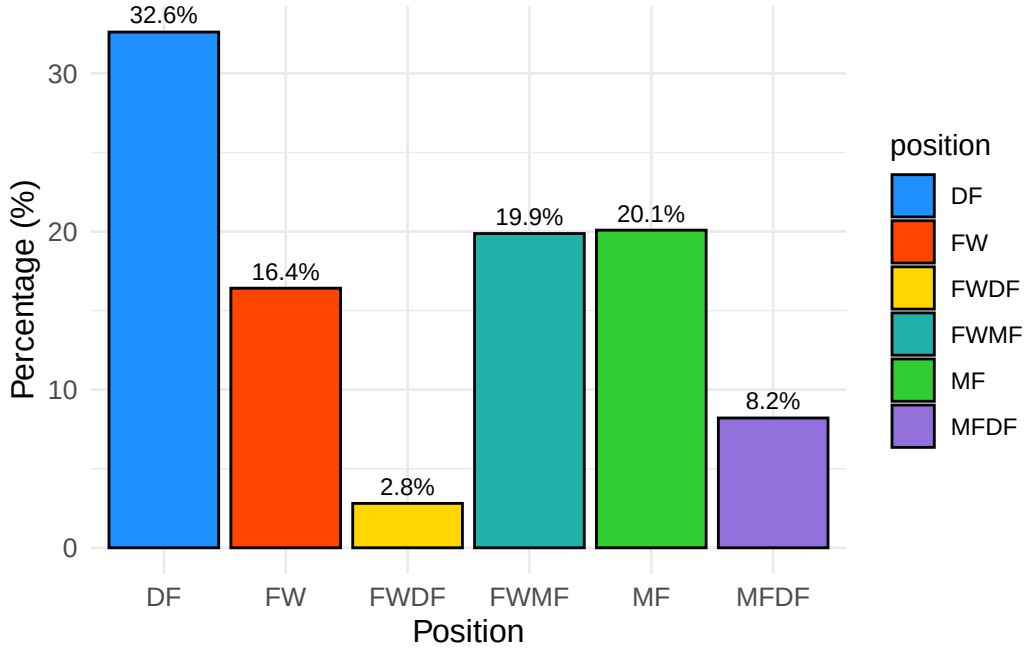


Figure 3: Percentage Distribution of Player Positions of the England Dataset

Table 1: Summary Statistics of the England Dataset

Variable	Mean	Median	SD	Min	Max
Age	24.79	25.00	4.38	16.00	38.00
Assists	1.64	1.00	2.41	0.00	13.00
Club Ranking	130.05	72.00	193.08	1.00	838.00
Goals	2.24	1.00	3.68	0.00	27.00
Market Value (in Million Euros)	20.99	14.00	23.23	0.10	180.00
Minutes	1310.96	1190.00	1001.09	1.00	3420.00
National Team Ranking	17.60	5.00	22.69	1.00	173.00

2.2.2 France Dataset

The France dataset offers interpretation into player profiles from Ligue 1, one of Europe's top football leagues. Similar to other leagues, the distribution of market values in the French dataset, as depicted in Figure 4, is heavily right-skewed. Most players have market values below €50 million, with only a few elite players commanding significantly higher valuations. This reflects the trend where top-tier talents, often representing national teams or excelling in international competitions, are valued much higher than their peers.

Figure 5 illustrates the distributions of predictor variables in the dataset:

- **Age**: The age distribution in Ligue 1 peaks around 24 years, with a majority of players between 20 and 30 years old. This highlights the league's balance between developing young talent and retaining experienced professionals. The maximum age in the dataset is 39 years, underscoring the presence of veteran players in the league.
- **Goals and Assists**: The distributions for goals and assists are highly right-skewed, with most players contributing fewer than 5 goals or assists in a season. This indicates that offensive production is concentrated among a small group of players, typically forwards or attacking midfielders.
- **Club Ranking**: Club rankings span a wide range, from 11 to 568, reflecting the diversity in club performance levels within the league.
- **National Team Ranking**: National team rankings also show significant variation, with players representing countries ranked in the top 10 as well as those from lower-ranked national teams. This diversity emphasizes Ligue 1's role as a platform for both elite and emerging talent.
- **Minutes Played**: The distribution of minutes played indicates that a significant proportion of players log substantial game time (over 1,000 minutes), with some exceeding 3,000 minutes, indicating their vital role in team strategies.

Figure 6 examines the percentage distribution of player positions in Ligue 1. The largest proportion of players are defenders (DF) at 29.1%, followed by midfielders (MF) at 22.3% and midfield-forward hybrids (FWMF) at 24.3%. Forwards (FW) constitute a smaller proportion (13.1%), while hybrid positions such as FWDF and MFDF account for the remaining distribution. This positional breakdown reflects a balanced tactical structure, with a strong emphasis on defensive and midfield roles.

Table 2 provides a detailed statistical summary of the numeric variables in the dataset. The mean market value stands at €7.98 million, significantly lower than the Premier League, indicating the varying financial dynamics across leagues. The average age is 23.91 years, reinforcing the league's focus on younger talent. Other variables such as goals and assists demonstrate low means relative to their maximum values, reflecting the rarity of exceptional performances.

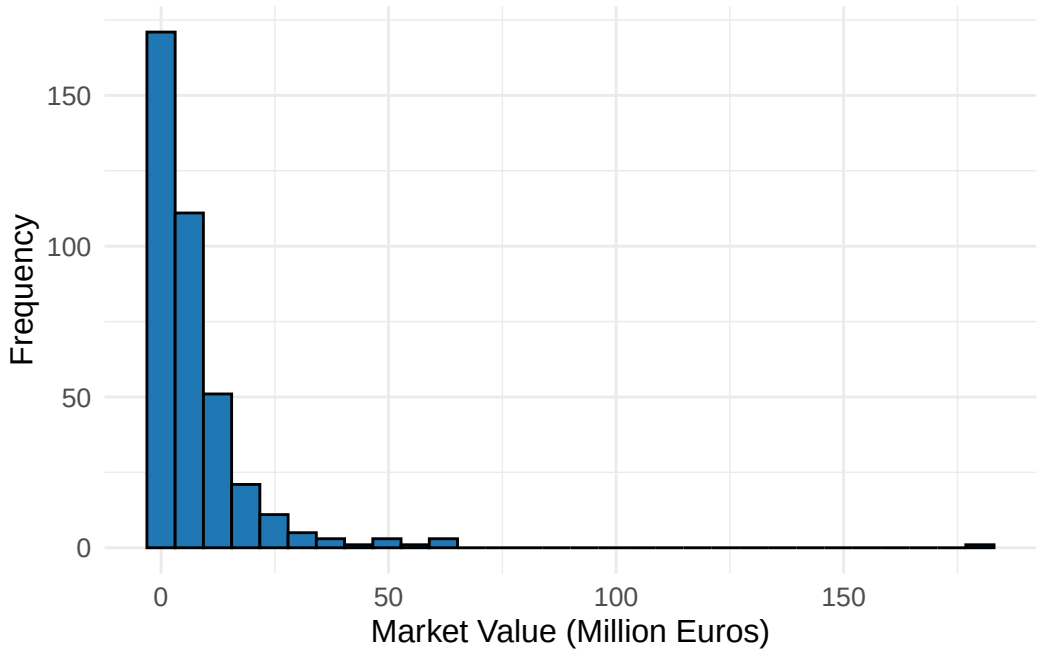


Figure 4: Distribution of Market Value of the France Dataset (in Million Euros)

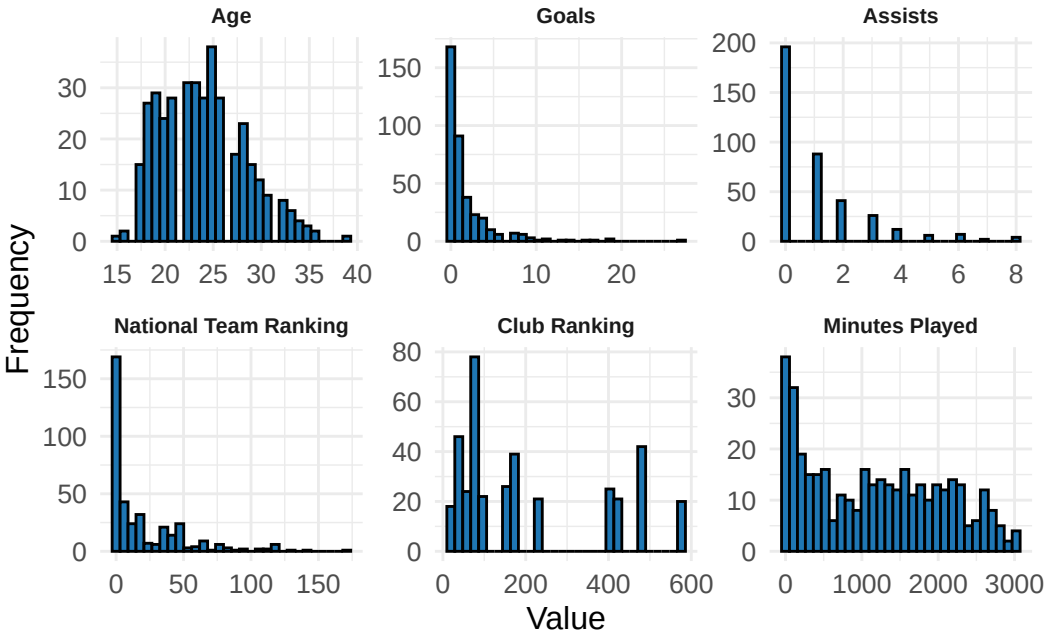


Figure 5: Distribution of Predictor Variables of the France Dataset

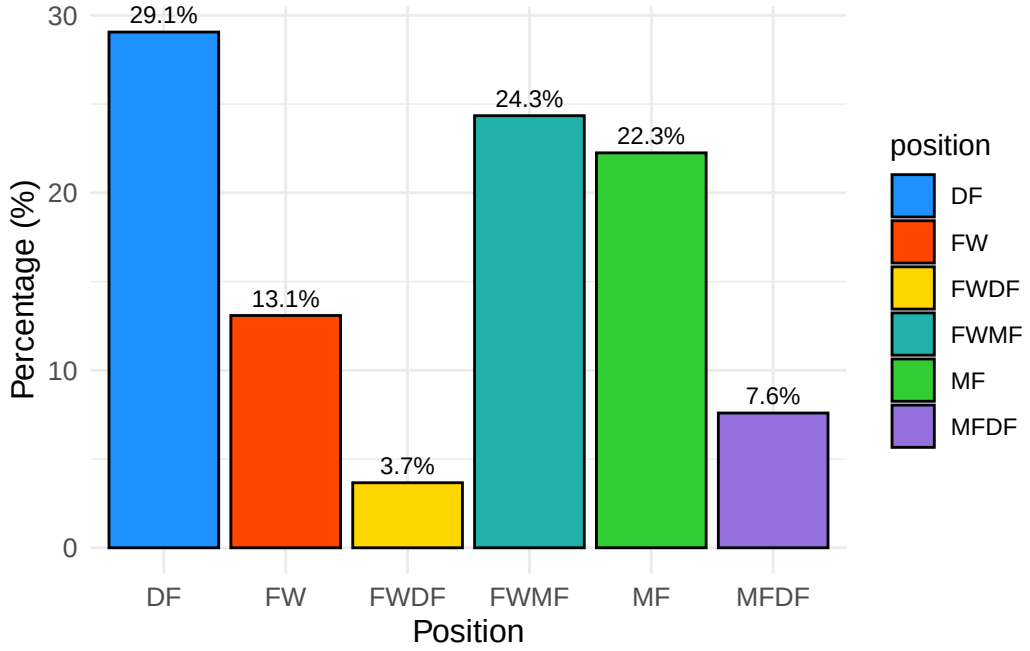


Figure 6: Percentage Distribution of Player Positions of the France Dataset

Table 2: Summary Statistics of the France Dataset

Variable	Mean	Median	SD	Min	Max
Age	23.91	24.00	4.47	15.00	39.00
Assists	1.08	0.00	1.61	0.00	8.00
Club Ranking	202.65	148.00	176.00	11.00	568.00
Goals	1.79	1.00	3.11	0.00	27.00
Market Value (in Million Euros)	7.98	4.00	13.25	0.05	180.00
Minutes	1166.79	1121.00	883.42	1.00	3022.00
National Team Ranking	20.68	4.50	28.82	1.00	172.00

2.2.3 Germany Dataset

The Germany dataset captures player characteristics across teams in the Bundesliga. As illustrated in Figure 7, the market value distribution is heavily right-skewed, with the majority of players concentrated in the lower market value brackets (below €50 million). This skewness reflects a superstar effect, with a small subset of players driving disproportionately higher market valuations.

Figure 8 explores the distributions of predictor variables, highlighting key patterns:

- **Age**: The age distribution peaks around 25 years, indicating that the league features a significant concentration of players in their athletic prime. The range extends from 17 years (likely youth players) to 39 years for veteran professionals.
- **Goals and Assists**: Both offensive contributions display a heavily right-skewed pattern, with most players contributing fewer than 5 goals or assists. This highlights the rarity of elite offensive contributors in the league.
- **Club Ranking**: Club ranking values range from 5 to 750, reflecting the diverse composition of the league, from globally dominant clubs to lower-performing teams.
- **National Team Ranking**: National team rankings show significant variation, with players representing both top-ranked and lower-ranked nations, capturing the Bundesliga's international player diversity.
- **Minutes Played**: The distribution of minutes played indicates substantial variability, with many players exceeding 1,000 minutes. This suggests that a majority of players are regular starters or receive substantial playing time.

The position distribution, depicted in Figure 9, demonstrates that defenders (DF) comprise the largest share of players (31.7%), followed by midfielders (MF) and forward-midfield hybrids (FWMF). Pure forwards (FW) form 9.6% of the players, while hybrid positions like FWDF make smaller contributions. This composition underscores the league's focus on tactical balance, with strong emphasis on midfield and defensive strategies.

Table 3 provides a detailed statistical summary of the numeric variables in the dataset. The mean market value is €10.20 million, while the median is much lower at €4.00 million, highlighting the skewed distribution. Similar trends are observed for assists and goals, where the mean is low compared to the maximum values, further emphasizing the role of standout performers in shaping league outcomes.

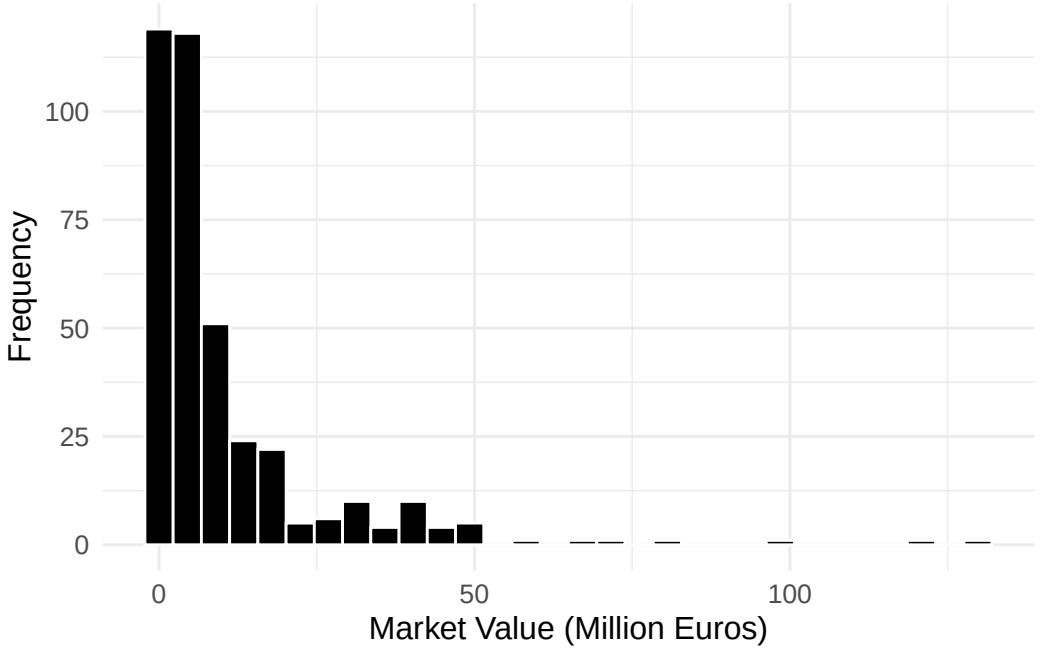


Figure 7: Distribution of Market Value of the Germany Dataset (in Million Euros)

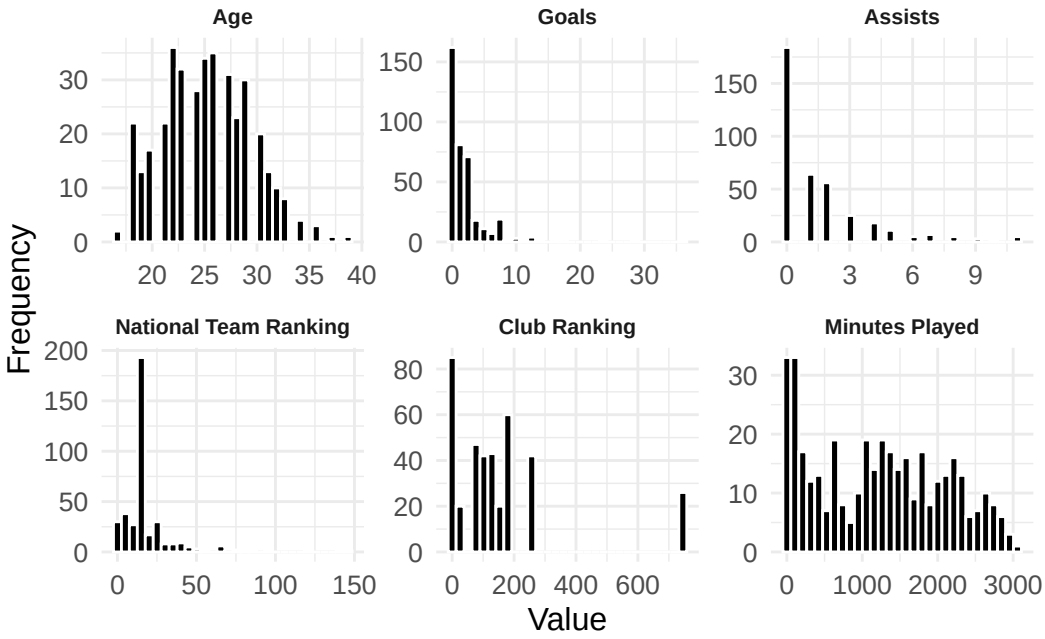


Figure 8: Distribution of Predictor Variables of the Germany Dataset

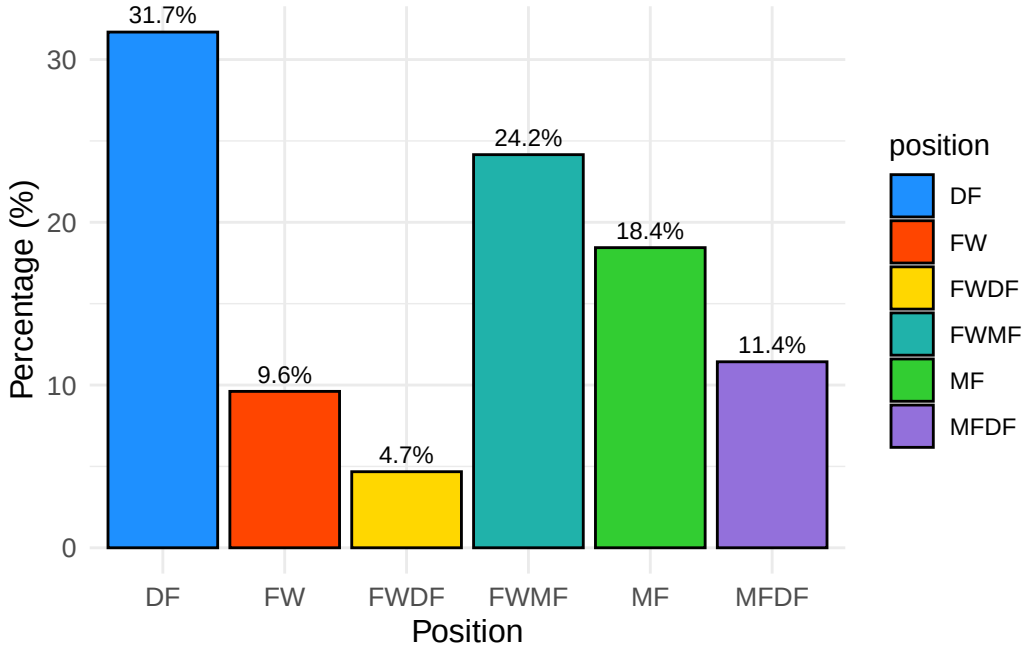


Figure 9: Percentage Distribution of Player Positions of the Germany Dataset

Table 3: Summary Statistics of the Germany Dataset

Variable	Mean	Median	SD	Min	Max
Age	25.19	25.00	4.20	17.00	39.00
Assists	1.56	1.00	2.30	0.00	11.00
Club Ranking	149.44	101.00	179.86	5.00	750.00
Goals	2.10	1.00	3.73	0.00	36.00
Market Value (in Million Euros)	10.20	4.00	15.79	0.20	130.00
Minutes	1214.80	1217.00	868.45	2.00	3060.00
National Team Ranking	19.22	16.00	17.63	1.00	147.00

2.2.4 Italy Dataset

The Italy dataset provides an detailed analysis of player characteristics within the Serie A league. As shown in Figure Figure 10, the market value distribution is right-skewed, with a majority of players concentrated in lower valuation brackets (below €30 million). This distribution is characteristic of the league, where the presence of a few superstar players significantly elevates the upper tail.

Figure Figure 11 highlights the distributions of key predictor variables:

- **Age**: The age distribution centers around 25 years, reflecting a mix of players in their early careers and those in their prime. The range extends from 15 to 36 years.
- **Goals and Assists**: These variables exhibit heavy skewness, with most players scoring fewer than 5 goals or assists in a season, underscoring the prominence of specialized attacking roles.
- **Club Ranking**: With values ranging from 3 to 681, the club ranking illustrates the diverse spectrum of team performance within Serie A.
- **National Team Ranking**: The national team ranking shows a range from 1 to 141, capturing the representation of players from both elite and emerging national teams.
- **Minutes Played**: The distribution of minutes played indicates significant variability, with several players logging substantial game time (over 1,000 minutes) and a few recording minimal participation.

Figure Figure 12 depicts the percentage distribution of player positions. Defenders (DF) form the largest proportion (35.4%), followed by midfielders (MF) at 24%. Forward-midfield hybrids (FWMF) and forwards (FW) are well-represented, emphasizing the league's balanced approach to attacking and defensive strategies. Hybrid positions such as FWDF and MFDF form smaller shares of the dataset.

Table Table 4 presents summary statistics for the numeric variables in the Italy dataset. The mean market value is €9.37 million, with a median of €4.00 million, demonstrating a significant skew. Goals and assists have low averages relative to their maximum values, aligning with the skewed distributions observed in their histograms.

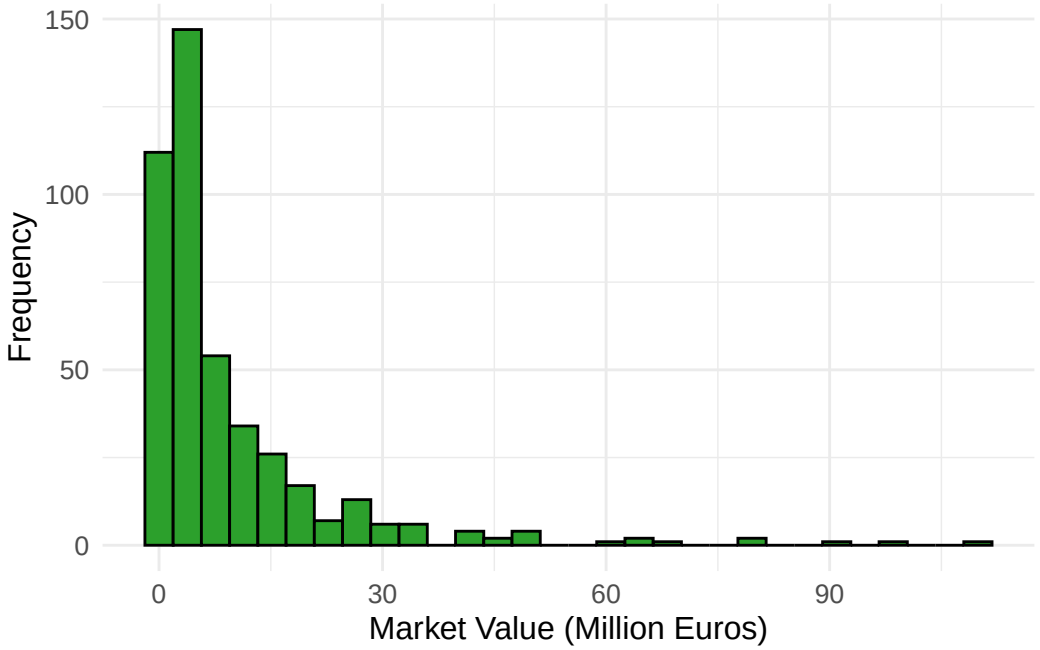


Figure 10: Distribution of Market Value of the Italy Dataset (in Million Euros)

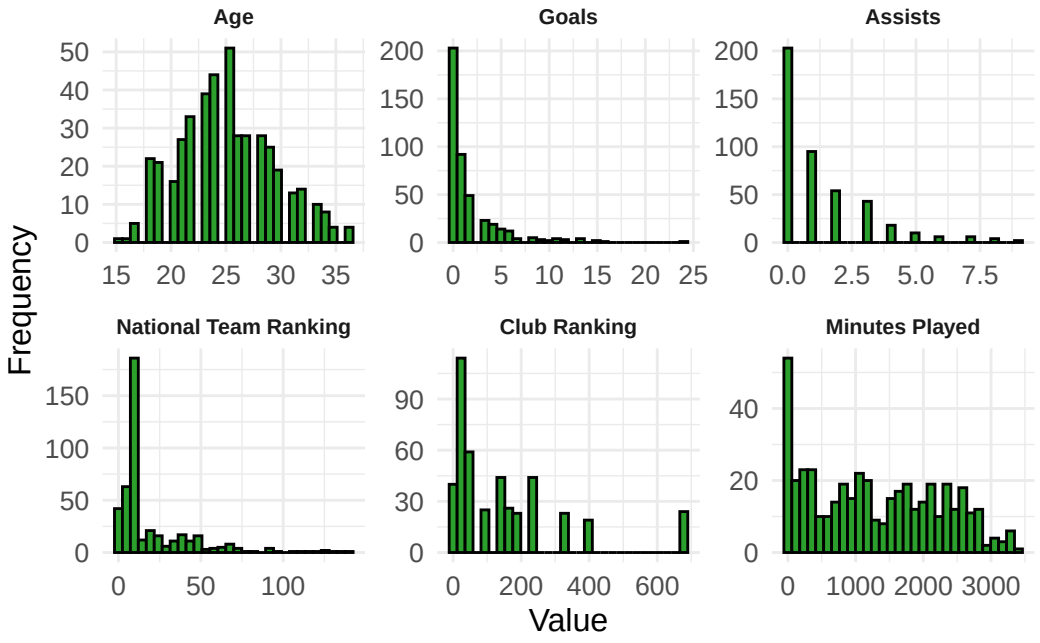


Figure 11: Distribution of Predictor Variables of the Italy Dataset

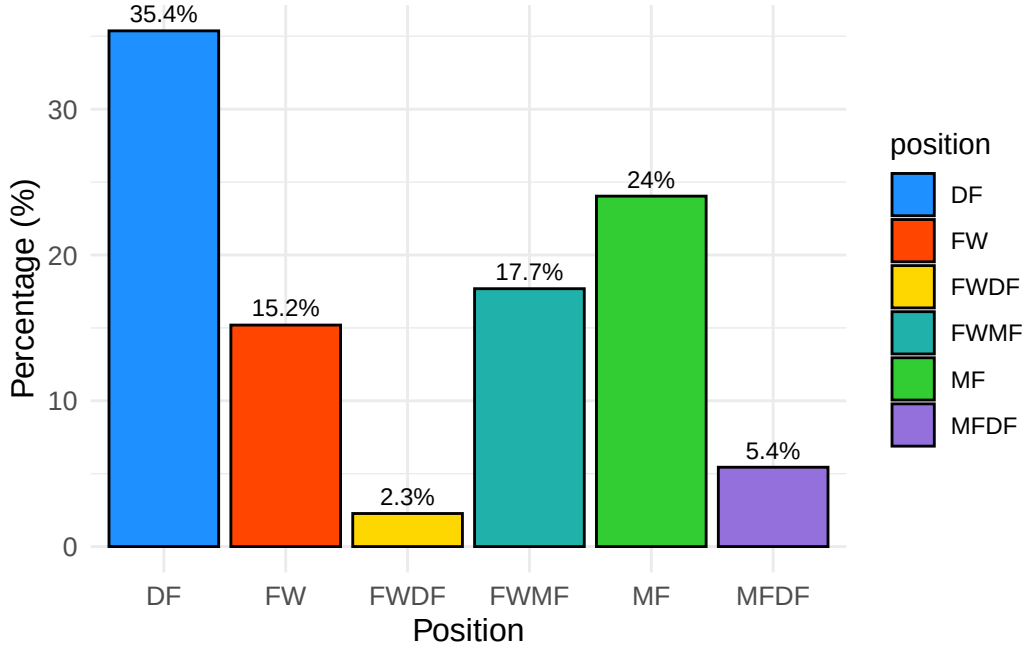


Figure 12: Percentage Distribution of Player Positions of the Italy Dataset

Table 4: Summary Statistics of the Italy Dataset

Variable	Mean	Median	SD	Min	Max
Age	25.06	25.00	4.33	15.00	36.00
Assists	1.32	1.00	1.78	0.00	9.00
Club Ranking	147.59	90.00	166.87	3.00	681.00
Goals	1.80	1.00	3.01	0.00	24.00
Market Value (in Million Euros)	9.37	4.00	14.11	0.10	110.00
Minutes	1290.32	1187.00	953.51	1.00	3406.00
National Team Ranking	20.24	10.00	23.95	1.00	141.00

2.2.5 Spain Dataset

The Spain dataset offers a detailed analysis of player characteristics in La Liga. As illustrated in Figure 13, the market value distribution is heavily right-skewed, with the majority of players concentrated in the lower valuation brackets (below €30 million). This trend reflects the league's structure, where a few superstar players dominate the upper tail of the distribution, commanding significantly higher valuations.

Figure 14 shows the distributions of key predictor variables:

- **Age**: The age distribution peaks at 26 years, showcasing a blend of players in their prime years alongside younger emerging talents and experienced veterans. The range spans from 16 to 37 years.
- **Goals and Assists**: These offensive metrics display a pronounced skewness, with most players contributing fewer than 5 goals or assists in a season. This underscores the rarity of elite-level offensive production within the league.
- **Club Ranking**: The club ranking ranges from 2 to 464, highlighting the diversity in team performance levels across La Liga, from top-tier teams to mid-table and lower-ranked sides.
- **National Team Ranking**: National team rankings range from 1 to 150, reflecting the presence of players from elite international teams as well as those from emerging footballing nations.
- **Minutes Played**: The minutes played distribution demonstrates significant variability, with a substantial number of players logging over 1,000 minutes, emphasizing their pivotal roles in team dynamics.

Figure 15 illustrates the percentage distribution of player positions. Defenders (DF) comprise the largest group (32.3%), followed by midfielders (MF) at 24.4%. Hybrid roles such as forward-midfield (FWMF) and forward (FW) are well-represented, reflecting La Liga's tactical diversity and emphasis on attacking football. Smaller proportions of hybrid defensive positions (FWDF, MFDF) further highlight the league's strategic aspect.

Table 5 summarizes the numeric variables in the Spain dataset. The mean market value stands at €9.91 million, with a median of €3.50 million, indicating a substantial skew. Goals and assists exhibit low means relative to their maximums, consistent with their heavily skewed distributions. Similarly, club rankings demonstrate a wide range, reflecting the competitive disparity across the league. Minutes played show a significant average of 1,316 minutes, emphasizing the league's reliance on core players for sustained performance.

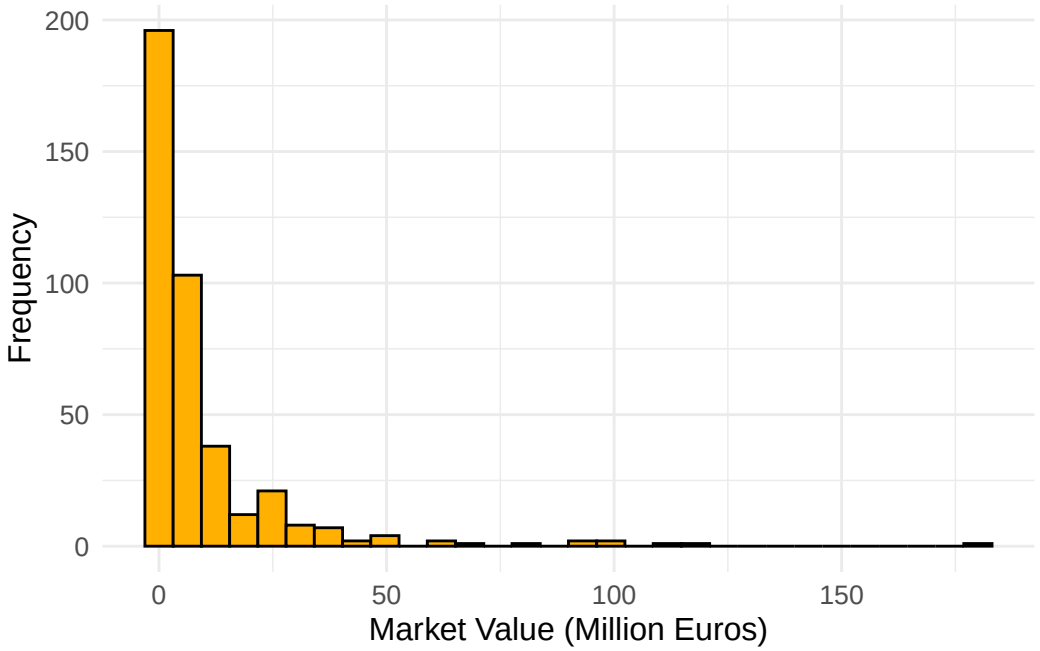


Figure 13: Distribution of Market Value of the Spain Dataset (in Million Euros)

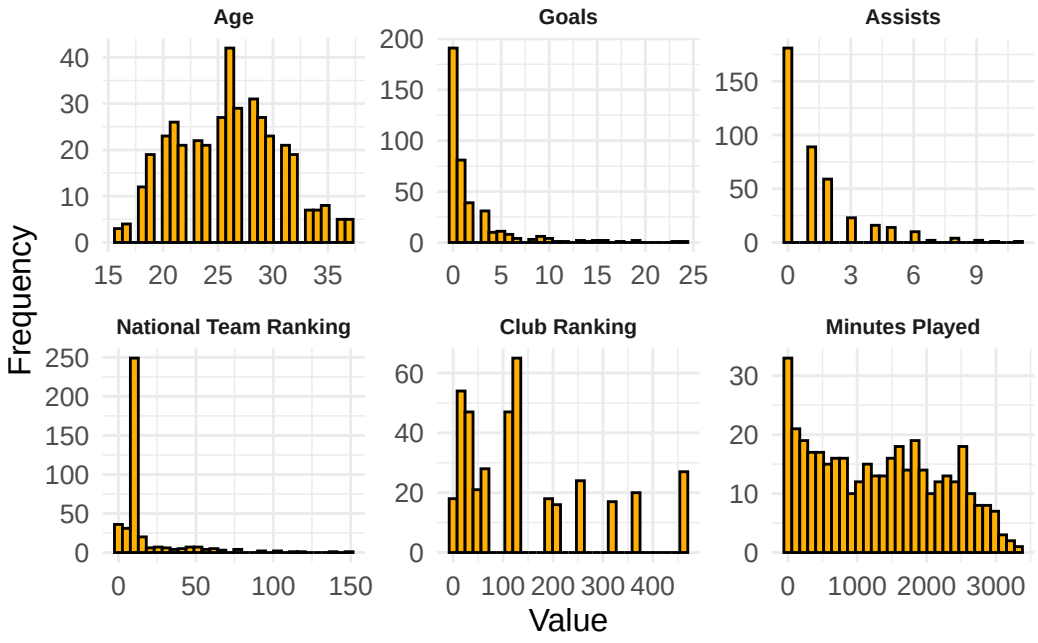


Figure 14: Distribution of Predictor Variables of the Spain Dataset

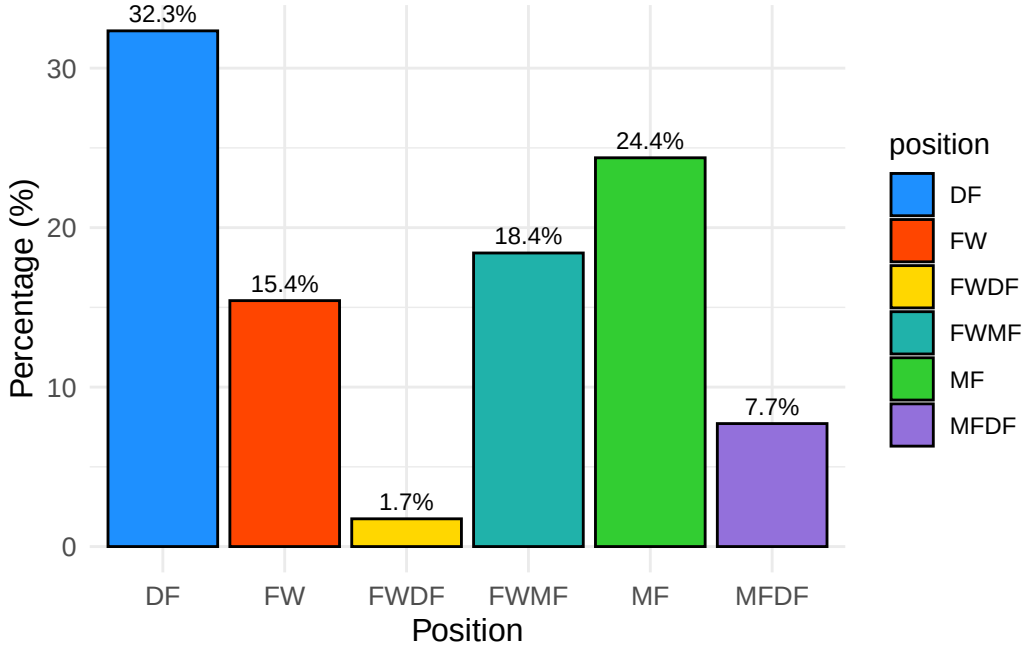


Figure 15: Percentage Distribution of Player Positions of the Spain Dataset

Table 5: Summary Statistics of the Spain Dataset

Variable	Mean	Median	SD	Min	Max
Age	25.96	26.00	4.73	16.00	37.00
Assists	1.38	1.00	1.91	0.00	11.00
Club Ranking	138.90	115.00	131.52	2.00	464.00
Goals	1.91	1.00	3.50	0.00	24.00
Market Value (in Million Euros)	9.91	3.50	18.22	0.05	180.00
Minutes	1315.79	1294.00	912.96	1.00	3327.00
National Team Ranking	14.97	8.00	20.56	1.00	150.00

2.3 Measurement

The construction of variables in this study ensures their relevance and accuracy. Player market values were obtained from Transfermarkt, a widely regarded source for estimating market valuations. These values are based on a combination of player performance, market demand, and subjective expert evaluation, making them a holistic representation of player worth. Performance metrics, such as goals and assists, were sourced from Stathead, a trusted repository of sports data. National team and club rankings were quantified to enable comparison across leagues. National team rankings were sourced from FIFA's official ranking system, which reflects international match results, strength of opposition, and match importance. Club rankings were sourced from Footballdatabase, which evaluates team performance using a consistent point-based methodology (See Section A.2 for a thorough description of data collection processes)

High-level cleaning included reconciling discrepancies in player names across sources, imputing missing data where appropriate, and ensuring categorical variables like position were consistently coded. While these steps enhance data usability, it is worth noting potential limitations in data accuracy, such as subjective biases in market valuations or inconsistencies in performance data collection.

2.4 Broader Context and Alternative Datasets

The selected dataset is uniquely suited for this analysis due to its extensive scope and integration of market, performance, and contextual data. While alternative sources such as Opta or WyScout provide more granular match-level data, their focus on in-game metrics (e.g., pass accuracy, expected goals) lacks the broader valuation context provided by Transfermarkt. Moreover, FIFA rankings and Footballdatabase rankings are well-established benchmarks for quantifying national and club performance, making them preferable over less transparent systems.

In summary, the dataset offers a holistic view of player characteristics and market values, enabling robust analyses of league-specific dynamics. The careful construction and validation of variables ensure their reliability and relevance, providing a solid foundation for the subsequent analyses presented in this paper.

3 Model

In this study, we constructed a linear regression model for each of the five major European football leagues to analyze the determinants of player market value. The dependent variable, market value (Market Value$_i$), was scaled to millions of Euros for easier interpretability. The model incorporates a wide-ranging set of predictors, including player performance metrics, contextual factors, and positional variables, to capture the multi-dimensional aspects influencing market valuation.

The general form of the model is as follows:

$$\begin{aligned}\text{Market Value}_i = \beta_0 &+ \beta_1(\text{Age}_i) + \beta_2(\text{Goals}_i) \\ &+ \beta_3(\text{Assists}_i) + \beta_4(\text{Club Ranking}_i) \\ &+ \beta_5(\text{National Team Ranking}_i) + \beta_6(\text{Minutes Played}_i) \\ &+ \beta_7(\text{Position}_i) + \epsilon_i\end{aligned}$$

where:

- β_0: The intercept, representing the baseline market value when all predictors are zero.
- β_1: The coefficient for age, showing how market value changes with each additional year of age.
- β_2: The coefficient for goals, representing the increase in market value per additional goal scored.
- β_3: The coefficient for assists, reflecting the market value impact of providing assists.
- β_4: The coefficient for club ranking, with better (lower) rankings associated with higher market values.
- β_5: The coefficient for national team ranking, highlighting the effect of international representation.
- β_6: The coefficient for minutes played, indicating the impact of game time on market value.
- β_7: The set of coefficients for positional categories, capturing positional differences in market valuation relative to a baseline.

- ϵ_i: The residual error, accounting for unobserved factors affecting market value.

3.1 Modeling Process

The models were implemented using the R programming language, specifically utilizing the `lm()` function for linear regression. For each league, data were split into training and testing sets to validate model performance. A stratified train-test split (80% training, 20% testing) was performed using the `caret` package to ensure balanced representation of market value distributions across the sets.

3.2 Model Validation

To evaluate model performance, predictions were generated on the test set, and Mean Squared Error (MSE) was computed for each league. MSE provides an average measure of prediction error, quantifying the extent to which observed market values deviate from predicted values. The MSE values across the five leagues are as follows:

- England: 265.27
- France: 100.49
- Germany: 215.51
- Italy: 109.95
- Spain: 122.17

The variation in MSE reflects differences in data variability and the complexity of valuation processes in each league. For instance, the relatively high MSE in England suggests greater variability in market values, likely influenced by the financial power and diversity of clubs in the Premier League.

3.3 Alternative Models and Future Directions

While linear regression provides a transparent and interpretable framework, alternative approaches could enhance predictive accuracy:

- **Ridge Regression**: To address multicollinearity, ridge regression could be employed to shrink coefficient estimates, reducing variance.
- **Random Forests or Gradient Boosting**: Nonlinear methods could capture complex interactions between predictors, potentially improving fit for leagues with high variability.
- **Bayesian Regression**: A Bayesian approach would allow for the incorporation of prior knowledge and explicit uncertainty quantification for coefficients.

4 Results

This section summarizes the key findings from the league-specific regression analyses, emphasizing only significant predictors of market value. Insignificant estimates, such as national team rankings and most position-specific coefficients, are excluded from this section but are included in the appendix.

The intercepts, which represent the baseline market value for a player with average characteristics, vary significantly across leagues. The Premier League (England) exhibits the highest intercept (41.057 million Euros), as shown in Figure 16, highlighting the league's financial dominance. Conversely, Ligue 1 (France) has the lowest intercept (22.946 million Euros), reflecting its relatively limited financial capacity.

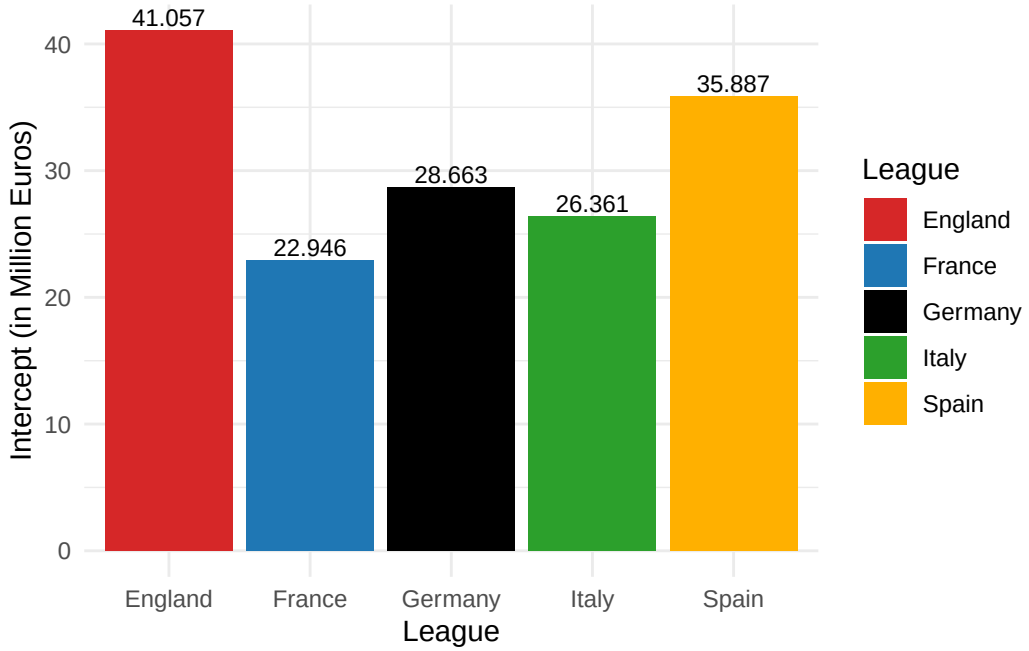


Figure 16: Intercept Estimates Across Leagues

Age negatively affects market value across all leagues, reflecting the premium placed on youth. Figure 17 demonstrates that La Liga (Spain) places the strongest emphasis on younger players, with a coefficient of -1.179 million Euros, followed by the Premier League (-1.137 million Euros). Ligue 1 (-0.593 million Euros) places the least emphasis on age.

Goals and assists are consistent predictors of market value across leagues. As shown in Figure 18, the impact of goals is highest in the Premier League (2.993 million Euros per goal) and Serie A (Italy) (2.489 million Euros). Figure 19 shows that assists have the strongest effect in the Bundesliga (Germany) (1.828 million Euros), while Serie A demonstrates the small-

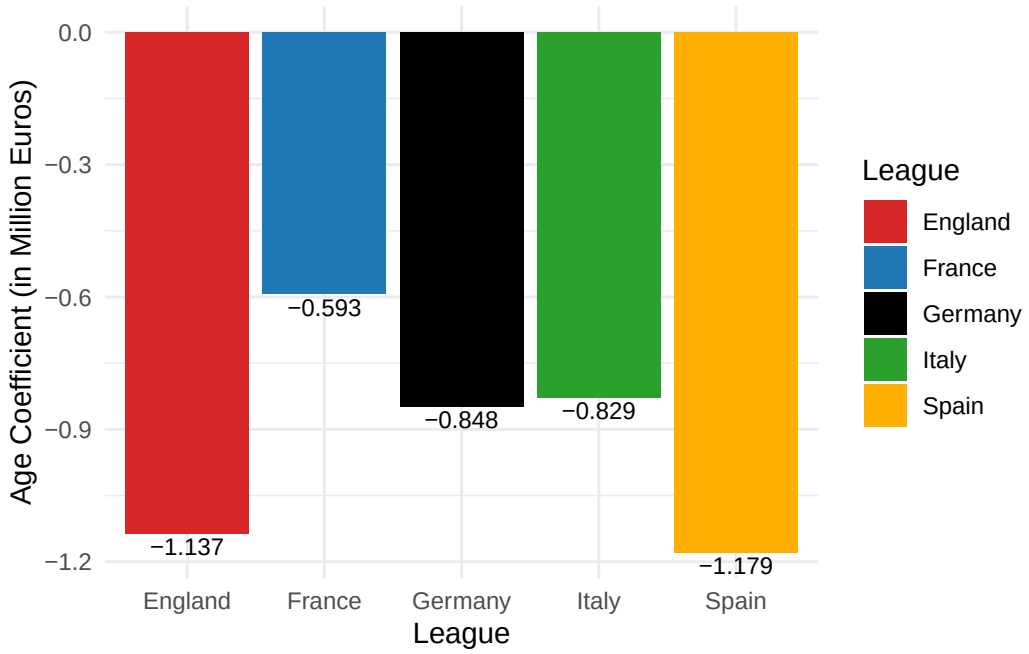


Figure 17: Age Coefficient Estimates Across Leagues

est effect (0.820 million Euros). These findings emphasize the financial rewards for offensive contributions, particularly in leagues like England and Germany.

Club ranking negatively influences market value in all leagues, as better club performance (lower numerical ranking) increases valuations. Figure 20 illustrates that the strongest effect is observed in La Liga (-0.034 million Euros per rank), followed by the Premier League (-0.027 million Euros). This highlights the importance of team performance in player valuation, particularly in Spain.

Minutes played is a significant, though modest, predictor of market value in England, Italy, and Spain. As shown in Figure 21, the coefficients range from 0.004 million Euros in England and Spain to 0.002 million Euros in Italy. These results suggest that playing time moderately impacts market value in these leagues.

Most position-specific coefficients are insignificant across leagues. However, Table 6 highlights some notable findings. In England, forwards (FW) have a significant negative effect on market value (-9.608 million Euros). Similarly, hybrid forward-midfielders (FWMF) have significant negative coefficients in France (-5.011 million Euros), Germany (-5.238 million Euros), and Italy (-4.877 million Euros). In Spain, none of the position-specific estimates are significant, indicating minimal variation in market valuation by position.

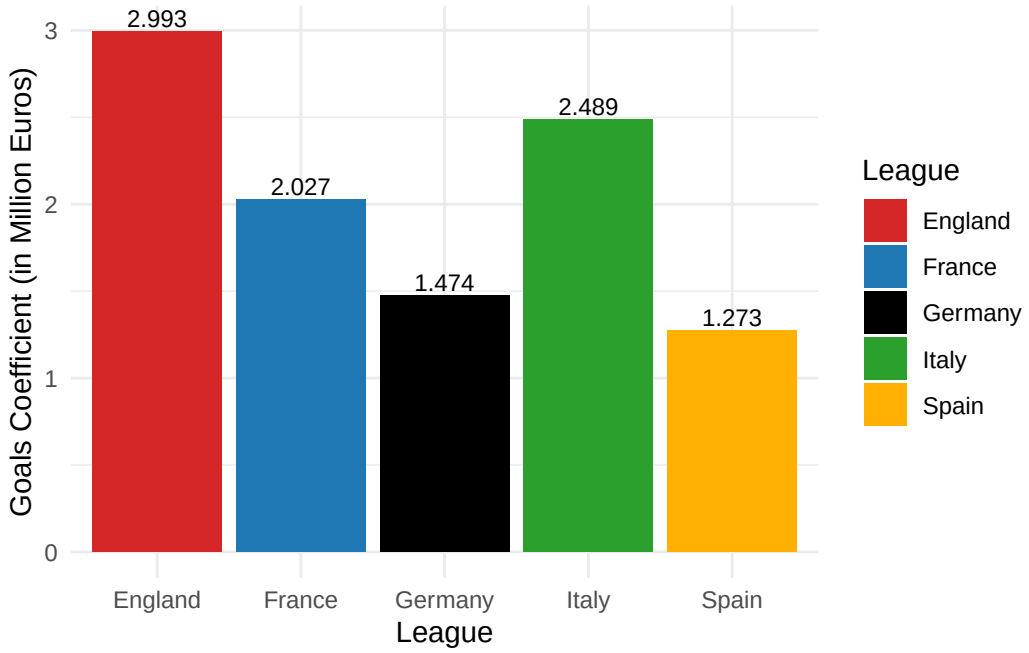


Figure 18: Goal Coefficient Estimates Across Leagues

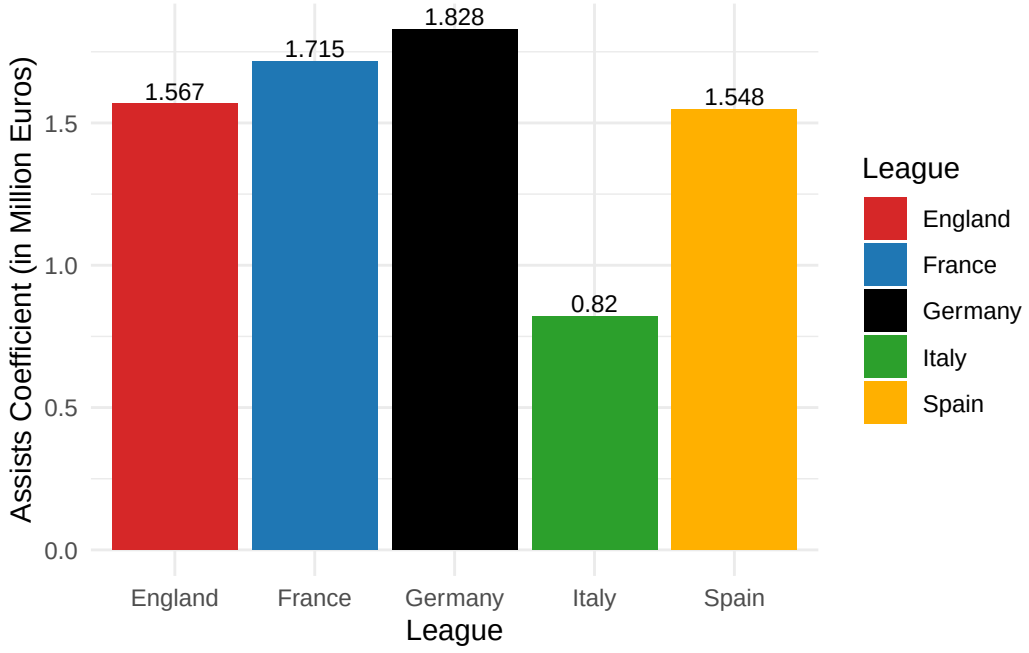


Figure 19: Assists Coefficient Estimates Across Leagues

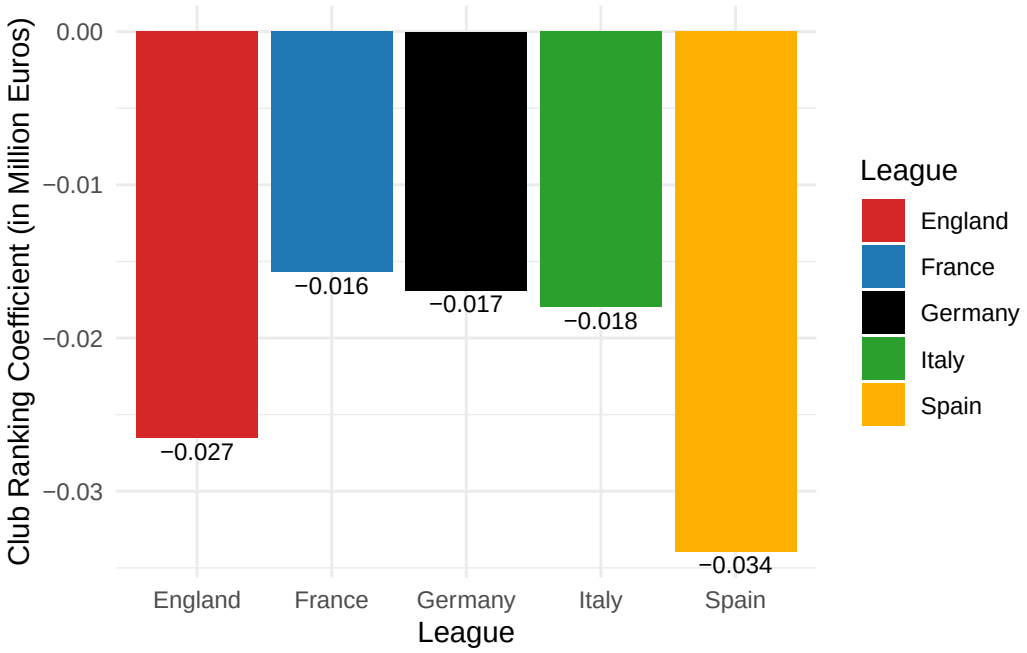


Figure 20: Club Ranking Coefficient Estimates Across Leagues

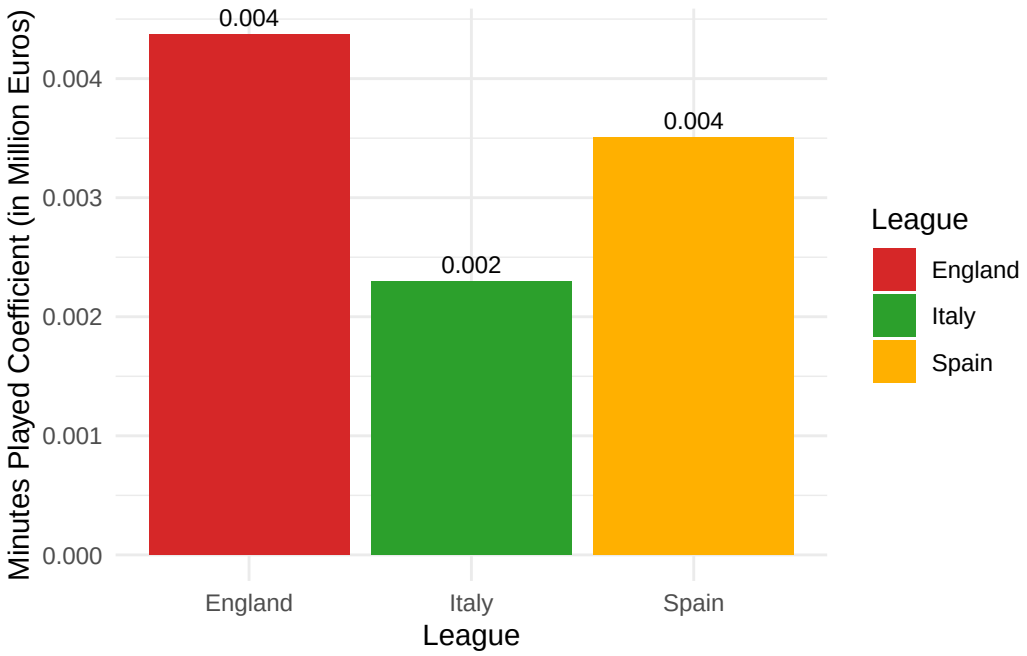


Figure 21: Minutes Played Coefficient Estimates Across Selected Leagues

Table 6: Position Specific Estimates in Selected Leagues

Position	League	Estimate
positionFW	England	-9.608
positionFW	France	-7.041
positionFWMF	France	-5.011
positionFWDF	Germany	-7.835
positionFWMF	Germany	-5.238
positionFWMF	Italy	-4.877

5 Discussion

5.1 Cross-League Differences in Market Value Baselines

One of the most striking findings is the variation in the intercept values across leagues, which reflect the baseline market value for players with average characteristics (or zero predictors). The Premier League (England) has the highest baseline market value (41.057 million Euros), followed by La Liga (Spain) and the Bundesliga (Germany). In contrast, Ligue 1 (France) exhibits the lowest intercept (22.946 million Euros). These results align with existing research, which highlights the financial strength and global appeal of the Premier League compared to other leagues. The disparity in intercepts underscores the role of financial resources, international exposure, and league-specific brand value in shaping player valuations.

5.2 Universal Importance of Goals and Assists

Goals and assists emerged as the most consistent predictors of market value across leagues. The Premier League shows the strongest relationship between goals and market value, followed by Serie A. This finding reflects the premium placed on offensive contributions, particularly in leagues that emphasize high-scoring games or where attacking players are highly valued in the transfer market. Similarly, assists play a significant role in Germany's Bundesliga, likely reflecting the league's focus on team-oriented and high-tempo playing styles. These results confirm prior research, which identifies offensive output as a key driver of player valuations in soccer.

5.3 Age and the Youth Premium

Age negatively impacts market value across all leagues, with the largest effect observed in La Liga (-1.179 million Euros per year) and the smallest in Ligue 1 (-0.593 million Euros per year). This pattern demonstrates the universal preference for younger players, who are

perceived to have greater potential for development and resale value. However, the magnitude of this effect varies by league, possibly due to differences in scouting practices and player development pipelines. For example, Spain's emphasis on nurturing young talent aligns with its stronger age effect, while Ligue 1, known for exporting young players, may rely on other metrics for valuation.

5.4 Club Ranking and Market Value

Club ranking, a proxy for team reputation and performance, shows a significant but modest negative relationship with market value. This finding indicates that players from higher-performing clubs (lower-ranked) tend to have higher valuations. Spain exhibits the strongest club ranking effect (-0.034 million Euros per rank improvement), which may reflect the emphasis placed on club prestige in La Liga's valuation processes. However, the relatively small magnitude across leagues suggests that individual performance metrics, such as goals and assists, are weighted more heavily than team success.

5.5 Position-Based Valuations

The role of player position in determining market value varied significantly across leagues, with most position-specific coefficients being statistically insignificant. However, some notable exceptions were observed. Forwards (FW) in England and France exhibited significant negative effects on market value, with the Premier League showing the strongest effect (-9.608 million Euros). Hybrid positions, such as forward-midfielders (FWMF), were significant in France, Germany, and Italy, indicating that players with versatile roles are valued differently depending on the league. Interestingly, none of the position-specific coefficients in Spain were significant, suggesting that La Liga places less emphasis on positional roles in player valuation and more on overall talent and individual performance. These findings highlight the complexity of valuing players based on position, as tactical demands and league-specific preferences play a significant role.

5.6 Weaknesses of the Study

While the findings provide useful results, the study has several limitations. First, it does not account for external factors such as player injuries, media influence, and agent negotiations, all of which are known to influence market value. Second, the analysis focuses on a single season (2023/24), limiting its ability to capture temporal trends and variations caused by external events such as the COVID-19 pandemic or changes in league regulations. Third, the use of linear regression assumes a straightforward relationship between predictors and market value, which may oversimplify complex interactions or nonlinear effects (e.g., diminishing returns on goals). Finally, the study's focus on Europe's top five leagues excludes emerging leagues, such

as Major League Soccer (MLS) or the Saudi Pro League, where market dynamics may differ significantly.

5.7 Future Directions

The findings of this study have important implications for clubs, agents, and analysts in the soccer industry. Clubs can use these observations to refine their scouting and recruitment strategies by focusing on metrics that are most valued in their specific league, such as goals in the Premier League or assists in the Bundesliga. Agents can utilize the results to highlight the most marketable aspects of their clients during contract negotiations, emphasizing factors like offensive contributions or club prestige. Additionally, analysts and researchers can build on these findings to explore broader trends in player valuations and market dynamics.

Future research could address the limitations of this study by incorporating external variables such as media coverage, player injuries, and sponsorship deals. Expanding the analysis to include data from multiple seasons would provide a more in-depth understanding of long-term trends and the impact of external shocks, such as economic crises or regulatory changes. Furthermore, the use of cutting edge machine learning models could capture more complex relationships between predictors and market value, offering a deeper understanding of the factors that drive player valuations. Including emerging leagues and non-European markets would also provide a more global perspective on the determinants of market value, highlighting how valuations differ between established and developing soccer ecosystems.

A Appendix

A.1 Data Retrieval

This section provides a detailed explanation of how the data for this analysis was collected and processed. The data spans several categories, including market values, performance metrics, national team rankings, and club rankings. All data was gathered from reliable sources, as described below.

A.1.1 Market Value Data

The market value data for players in the five major European leagues was retrieved from Transfermarkt.com, a widely recognized platform for football statistics and player valuations. The Python script, located at `scripts/02-scrape_data.py`, was used to scrape data from Transfermarkt. The script utilizes the `requests` library for fetching webpage content and `BeautifulSoup` for parsing HTML.

The following steps were followed:

1. **Target URLs**: Separate lists of URLs for the club pages of each league (England, Germany, Italy, Spain, and France) were prepared. Each URL corresponds to the squad page of a club for the 2023 season.
2. **Scraping Logic**:
    - Player names were extracted from table cells with the class `hauptlink`.
    - Market values were parsed from cells with the class `rechts hauptlink`.
    - Rows corresponding to players were identified using row classes `odd` and `even`.
3. **Data Storage**: The scraped data for each league was saved as a CSV file in the directory `data/01-raw_data/raw_market_value_data/`.
4. **Script Features**:
    - It includes headers to mimic browser requests and avoid potential blocking.
    - A function, `scrape_transfermarkt`, was developed to extract player information from a single URL, while `scrape_and_save` automates the process for multiple URLs.

The output consists of player names and their estimated market values, providing a foundational dataset for this study.

A.1.2 Performance Data

Performance data, such as goals, assists, and minutes played, was sourced from **Stathead.com**, a service linked to the FBref database.

To retrieve the data:

1. Navigate to the "Season Finder" tool under the football section.
2. Set the filters, including season (2023–2024), league, and performance metrics such as goals and assists and click on "Get Results" to generate the dataset.
3. Export the results as a CSV file using the "Export Data" > "Get Table as CSV" option.
4. Copy and paste the produced dataset.

The downloaded data contains essential player performance statistics, formatted in a tabular structure, which was later cleaned and merged with other datasets.

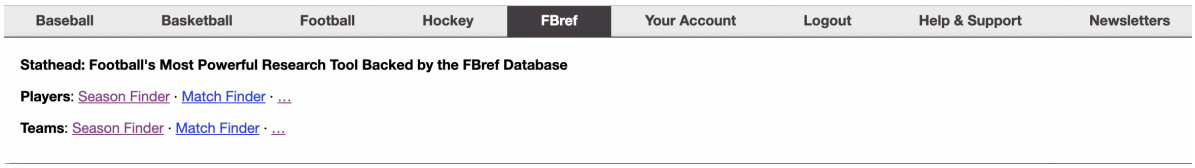


Figure 22: Step 1

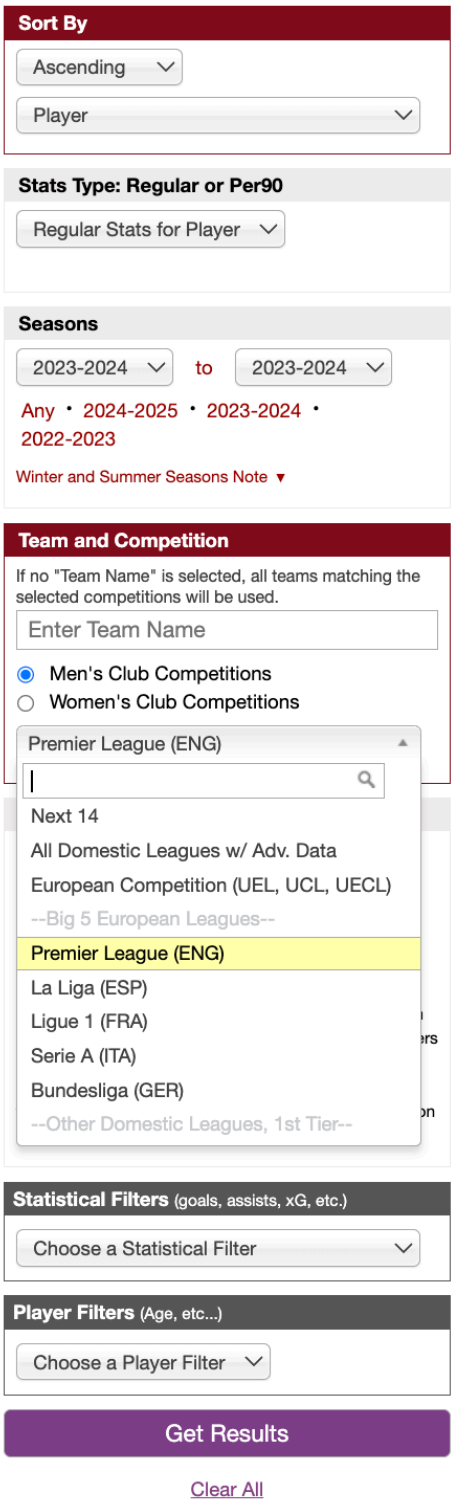
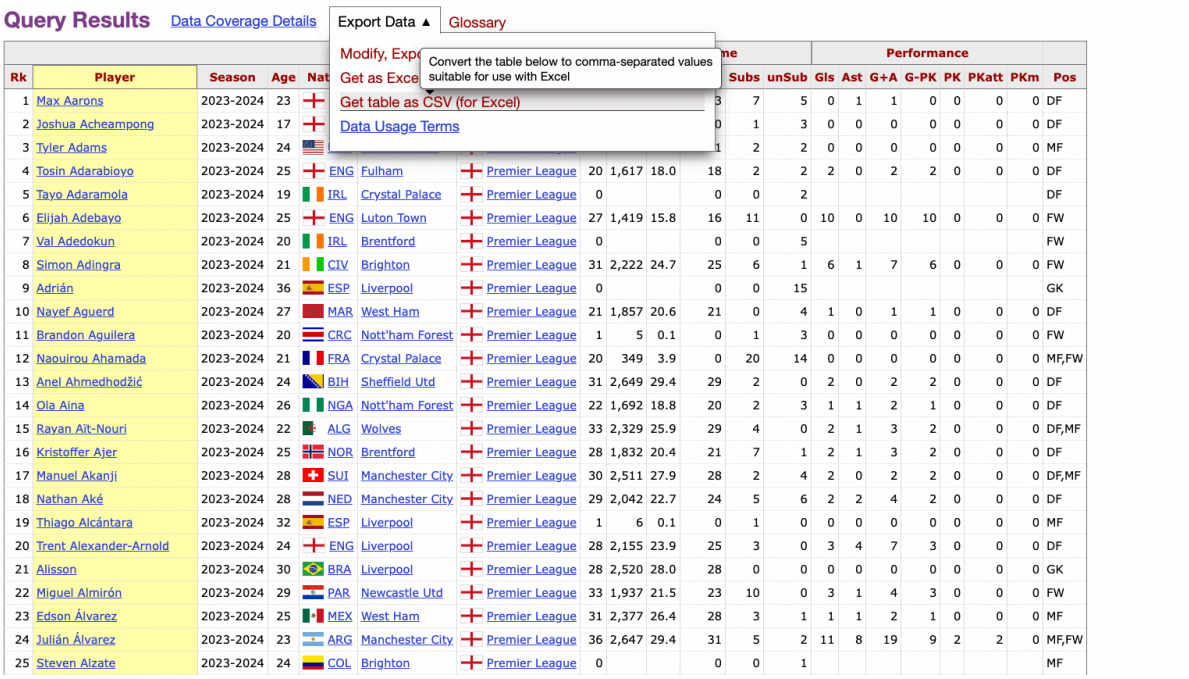


Figure 23: Step 2



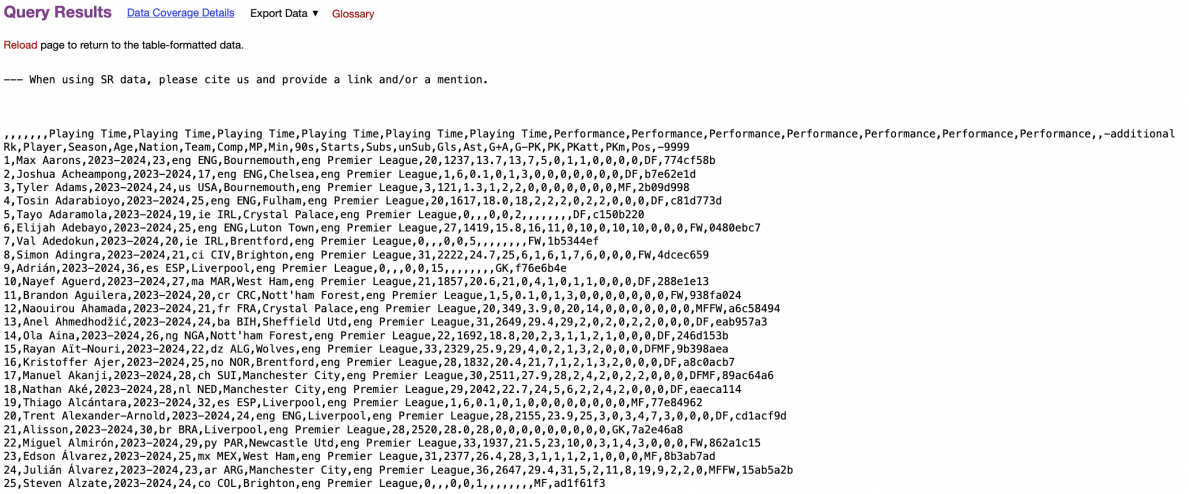
Query Results Data Coverage Details | Export Data ▲ | Glossary

Modify, Expo… | Get as Exce… | Get table as CSV (for Excel) | Data Usage Terms

Convert the table below to comma-separated values suitable for use with Excel

															Performance							
Rk	Player	Season	Age	Nat								Subs	unSub	Gls	Ast	G+A	G-PK	PK	PKatt	PKm	Pos
1	Max Aarons	2023-2024	23									7	5	0	1	1	0	0	0	0	DF
2	Joshua Acheampong	2023-2024	17									1	3	0	0	0	0	0	0	0	DF
3	Tyler Adams	2023-2024	24									2	2	0	0	0	0	0	0	0	MF
4	Tosin Adarabioyo	2023-2024	25	ENG	Fulham	Premier League	20	1,617	18.0	18	2	2	2	0	2	2	0	0	0	0	DF
5	Tayo Adaramola	2023-2024	19	IRL	Crystal Palace	Premier League	0			0	0	2									DF
6	Elijah Adebayo	2023-2024	25	ENG	Luton Town	Premier League	27	1,419	15.8	16	11	0	10	0	10	10	0	0	0	0	FW
7	Val Adedokun	2023-2024	20	IRL	Brentford	Premier League	0			0	0	5									FW
8	Simon Adingra	2023-2024	21	CIV	Brighton	Premier League	31	2,222	24.7	25	6	1	6	1	7	6	0	0	0	0	FW
9	Adrián	2023-2024	36	ESP	Liverpool	Premier League	0			0	0	15									GK
10	Nayef Aguerd	2023-2024	27	MAR	West Ham	Premier League	21	1,857	20.6	21	0	4	1	0	1	1	0	0	0	0	DF
11	Brandon Aguilera	2023-2024	20	CRC	Nott'ham Forest	Premier League	1	5	0.1	0	1	3	0	0	0	0	0	0	0	0	FW
12	Naouirou Ahamada	2023-2024	21	FRA	Crystal Palace	Premier League	20	349	3.9	0	20	14	0	0	0	0	0	0	0	0	MF,FW
13	Anel Ahmedhodžić	2023-2024	24	BIH	Sheffield Utd	Premier League	31	2,649	29.4	29	2	0	2	0	2	2	0	0	0	0	DF
14	Ola Aina	2023-2024	26	NGA	Nott'ham Forest	Premier League	22	1,692	18.8	20	2	3	1	1	2	1	0	0	0	0	DF
15	Rayan Aït-Nouri	2023-2024	22	ALG	Wolves	Premier League	33	2,329	25.9	29	4	0	2	1	3	2	0	0	0	0	DF,MF
16	Kristoffer Ajer	2023-2024	25	NOR	Brentford	Premier League	28	1,832	20.4	21	7	1	2	1	3	2	0	0	0	0	DF
17	Manuel Akanji	2023-2024	28	SUI	Manchester City	Premier League	30	2,511	27.9	28	2	4	2	0	2	2	0	0	0	0	DF,MF
18	Nathan Aké	2023-2024	28	NED	Manchester City	Premier League	29	2,042	22.7	24	5	6	2	2	4	2	0	0	0	0	DF
19	Thiago Alcántara	2023-2024	32	ESP	Liverpool	Premier League	1	6	0.1	0	1	0	0	0	0	0	0	0	0	0	MF
20	Trent Alexander-Arnold	2023-2024	24	ENG	Liverpool	Premier League	28	2,155	23.9	25	3	0	3	4	7	3	0	0	0	0	DF
21	Alisson	2023-2024	30	BRA	Liverpool	Premier League	28	2,520	28.0	28	0	0	0	0	0	0	0	0	0	0	GK
22	Miguel Almirón	2023-2024	29	PAR	Newcastle Utd	Premier League	33	1,937	21.5	23	10	0	3	1	4	3	0	0	0	0	FW
23	Edson Álvarez	2023-2024	25	MEX	West Ham	Premier League	31	2,377	26.4	28	3	1	1	1	2	1	0	0	0	0	MF
24	Julián Álvarez	2023-2024	23	ARG	Manchester City	Premier League	36	2,647	29.4	31	5	2	11	8	19	9	2	2	0	MF,FW	
25	Steven Alzate	2023-2024	24	COL	Brighton	Premier League	0			0	0	1									MF

Figure 24: Step 3



Query Results Data Coverage Details | Export Data ▾ | Glossary

Reload page to return to the table-formatted data.

```
--- When using SR data, please cite us and provide a link and/or a mention.


,,,,,,,Playing Time,Playing Time,Playing Time,Playing Time,Playing Time,Playing Time,Performance,Performance,Performance,Performance,Performance,Performance,Performance,,-additional
Rk,Player,Season,Age,Nation,Team,Comp,MP,Min,90s,Starts,Subs,unSub,Gls,Ast,G+A,G-PK,PK,PKatt,PKm,Pos,-9999
1,Max Aarons,2023-2024,23,eng ENG,Bournemouth,eng Premier League,20,1237,13.7,13,7,5,0,1,1,0,0,0,0,0,DF,774cf58b
2,Joshua Acheampong,2023-2024,17,eng ENG,Chelsea,eng Premier League,1,6,0.1,0,1,3,0,0,0,0,0,0,0,0,DF,b7e62e1d
3,Tyler Adams,2023-2024,24,us USA,Bournemouth,eng Premier League,3,121,1.3,1,2,2,0,0,0,0,0,0,0,0,MF,2b09d998
4,Tosin Adarabioyo,2023-2024,25,eng ENG,Fulham,eng Premier League,20,1617,18.0,18,2,2,2,0,2,2,0,0,0,0,DF,c81d773d
5,Tayo Adaramola,2023-2024,19,ie IRL,Crystal Palace,eng Premier League,0,,,0,0,2,,,,,,,,DF,c150b220
6,Elijah Adebayo,2023-2024,25,eng ENG,Luton Town,eng Premier League,27,1419,15.8,16,11,0,10,0,10,10,0,0,0,0,FW,0480ebc7
7,Val Adedokun,2023-2024,20,ie IRL,Brentford,eng Premier League,0,,,0,0,5,,,,,,,,FW,1b5344ef
8,Simon Adingra,2023-2024,21,ci CIV,Brighton,eng Premier League,31,2222,24.7,25,6,1,6,1,7,6,0,0,0,0,FW,4dcec659
9,Adrián,2023-2024,36,es ESP,Liverpool,eng Premier League,0,,,0,0,15,,,,,,,,GK,f76e6b4e
10,Nayef Aguerd,2023-2024,27,ma MAR,West Ham,eng Premier League,21,1857,20.6,21,0,4,1,0,1,1,0,0,0,0,DF,288e1e13
11,Brandon Aguilera,2023-2024,20,cr CRC,Nott'ham Forest,eng Premier League,1,5,0.1,0,1,3,0,0,0,0,0,0,0,0,FW,938fa024
12,Naouirou Ahamada,2023-2024,21,fr FRA,Crystal Palace,eng Premier League,20,349,3.9,0,20,14,0,0,0,0,0,0,0,0,MFFW,a6c58494
13,Anel Ahmedhodžić,2023-2024,24,ba BIH,Sheffield Utd,eng Premier League,31,2649,29.4,29,2,0,2,0,2,0,2,2,0,0,0,DF,eab957a3
14,Ola Aina,2023-2024,26,ng NGA,Nott'ham Forest,eng Premier League,22,1692,18.8,20,2,3,1,1,2,1,0,0,0,0,DF,246d153b
15,Rayan Aït-Nouri,2023-2024,22,dz ALG,Wolves,eng Premier League,33,2329,25.9,29,4,0,2,1,3,2,0,0,0,0,DFMF,9b398aea
16,Kristoffer Ajer,2023-2024,25,no NOR,Brentford,eng Premier League,28,1832,20.4,21,7,1,2,1,3,2,0,0,0,0,DF,a8c0acb7
17,Manuel Akanji,2023-2024,28,ch SUI,Manchester City,eng Premier League,30,2511,27.9,28,2,4,2,0,2,2,2,0,0,0,DFMF,89ac64a6
18,Nathan Aké,2023-2024,28,nl NED,Manchester City,eng Premier League,29,2042,22.7,24,5,6,2,2,4,2,0,0,0,0,DF,eaeca114
19,Thiago Alcántara,2023-2024,32,es ESP,Liverpool,eng Premier League,1,6,0.1,0,1,0,0,0,0,0,0,0,0,0,MF,77e84962
20,Trent Alexander-Arnold,2023-2024,24,eng ENG,Liverpool,eng Premier League,28,2155,23.9,25,3,0,3,4,7,3,0,0,0,0,DF,cd1acf9d
21,Alisson,2023-2024,30,br BRA,Liverpool,eng Premier League,28,2520,28.0,28,0,0,0,0,0,0,0,0,0,0,GK,7a2e46a8
22,Miguel Almirón,2023-2024,29,py PAR,Newcastle Utd,eng Premier League,33,1937,21.5,23,10,0,3,1,4,3,0,0,0,0,FW,862a1c15
23,Edson Álvarez,2023-2024,25,mx MEX,West Ham,eng Premier League,31,2377,26.4,28,3,1,1,1,2,1,0,0,0,0,MF,8b3ab7ad
24,Julián Álvarez,2023-2024,23,ar ARG,Manchester City,eng Premier League,36,2647,29.4,31,5,2,11,8,19,9,2,2,0,MFFW,15ab5a2b
25,Steven Alzate,2023-2024,24,co COL,Brighton,eng Premier League,0,,,0,0,1,,,,,,,,MF,ad1f61f3
```

Figure 25: Step 4

A.1.3 National Team Ranking Data

National team rankings were sourced from **FIFA World Ranking**, a globally recognized metric for evaluating the performance of national football teams based on international match results. These rankings provide useful context for assessing the market values of players, particularly those participating in international competitions or representing highly ranked national teams.

To retrieve the data:

1. Visit **FIFA World Ranking: Men**.
2. Set the **year** to **2024** in the available filters.
3. Choose the **date** as **June 20, 2024**, which aligns with the end of the 2023/2024 football season.
4. Manually extract the rankings for all national teams listed. Each team is assigned a rank between 1 and 210, with 1 being the highest-performing team.

These rankings were then matched to players in the dataset based on their national team affiliations. The June 2024 rankings ensure that the data captures the latest team standings before the next season begins, offering a reliable indicator of the competitive strength of each player's national team. The rankings were stored and merged as a numeric variable in the dataset to analyze their influence on player market values.

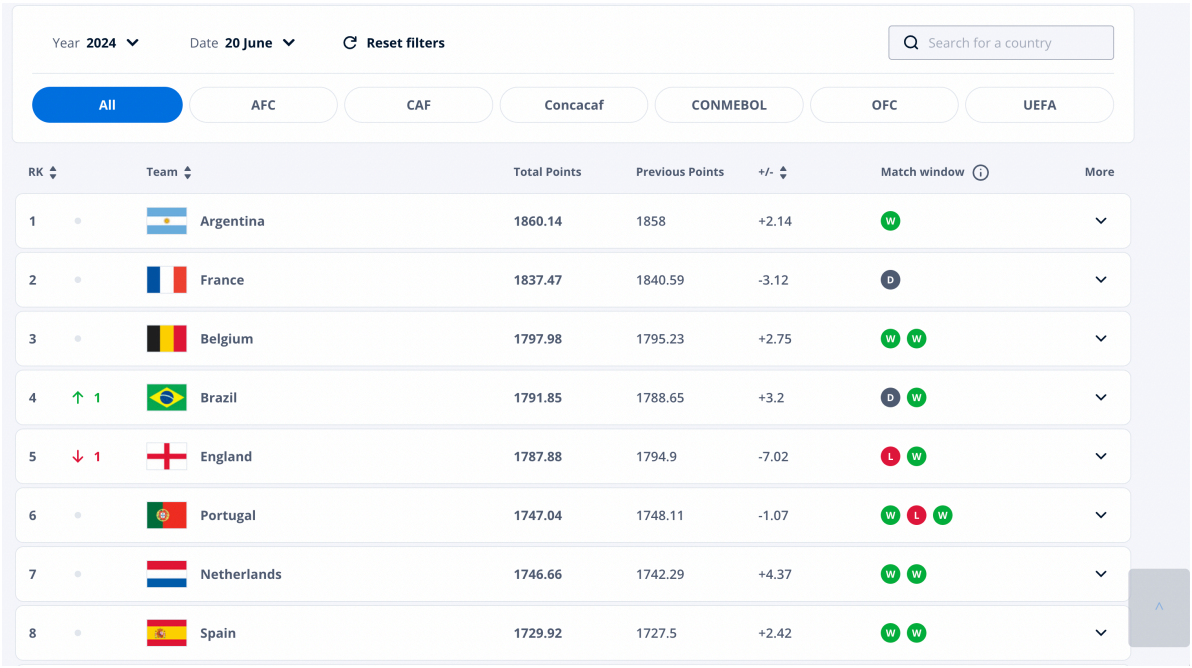


Figure 26: Example to Get National Team Ranking Data

A.1.4 Club Ranking Data

Club rankings were collected from **Footballdatabase.com**, a platform for club-level football statistics and rankings. The rankings reflect club performances based on both domestic league and international competition results. These rankings are essential for understanding the relative strength of the clubs in which players are involved, providing a significant explanatory variable for player market values.

To retrieve the data:

1. Visit **Footballdatabase.com**.
2. Navigate to the page of the club of interest by using the search bar or browsing the league standings.
3. On the club's profile page, locate the **line chart** displaying the club's ranking history over time.
4. Click on the point corresponding to **June 2024**, marking the end of the 2023/2024 season. This ensures the ranking reflects the club's performance for the entire season.
5. Record the club's ranking from the chart and compile the rankings for all clubs in the dataset.

These rankings were manually extracted for each club and merged into the dataset by matching player club affiliations. This method provided an accurate and up-to-date measure of club performance, offering pivotal context for analyzing player market values.

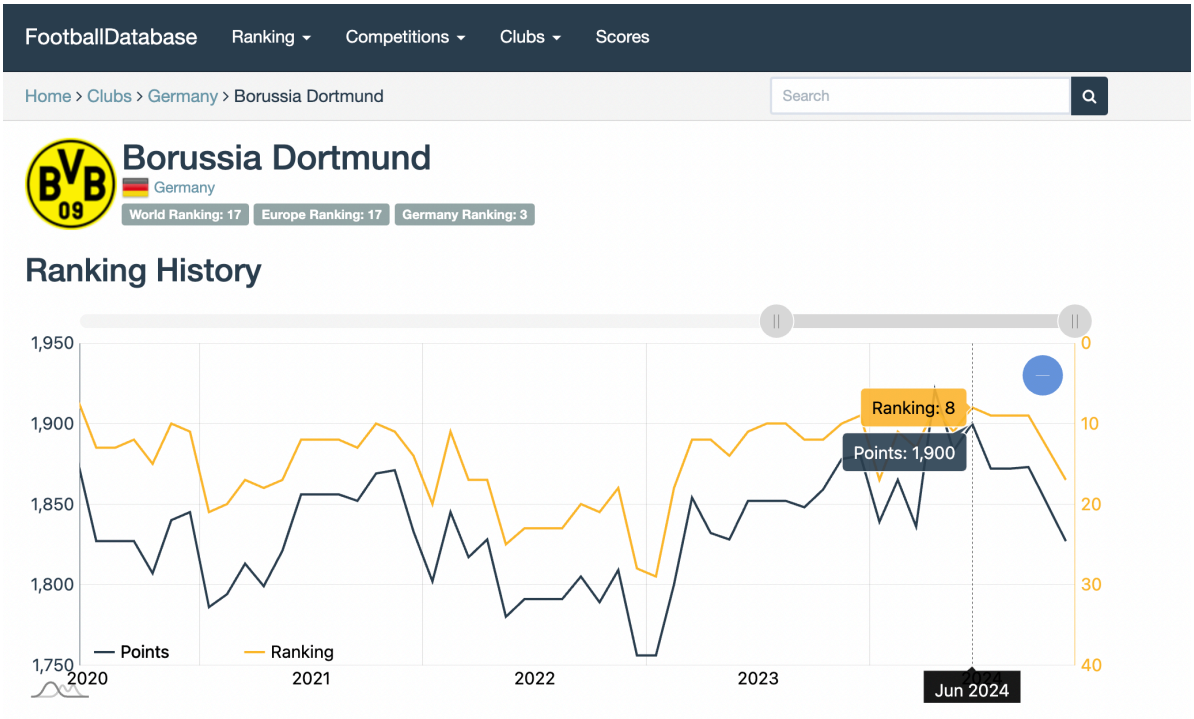


Figure 27: Example to Get Club Ranking Data

A.2 Methodology and Observational Data Collection Processes

This appendix provides an in-depth examination of the methodologies employed by prominent soccer-related data platforms—Transfermarkt, Stathead FBref, FIFA, and Footballdatabase—in collecting, curating, and analyzing observational data. These platforms are integral to understanding player valuations, performance metrics, national team rankings, and club performance, offering detailed information into the global soccer ecosystem.

A.2.1 Transfermarkt: A Community-Driven Approach to Market Valuation

Transfermarkt operates as a unique amalgamation of crowd-sourced information and expert validation, establishing itself as a leading authority on player market valuations. The platform collects observational data from official league statistics, contractual details, and biographical data provided by club announcements and news outlets. This raw data is refined and augmented through a global network of registered users who actively contribute to player profiles, transfer histories, and performance metrics (Transfermarkt 2024).

The market valuation process on Transfermarkt is particularly notable for its participatory approach. Registered users submit valuation estimates for players, which are then debated in

community forums. These discussions include consideration of a player's recent performance, age, injury history, positional role, and market trends, such as positional scarcity or demand from specific clubs. While the initial valuation is user-driven, Transfermarkt employs a team of moderators and experts who finalize these values by synthesizing community input with objective performance indicators. This collaborative approach ensures valuations remain dynamic and responsive to real-world developments (Transfermarkt 2024).

Transfermarkt's emphasis on transparency extends to its valuation methodology. Key determinants of market value include performance metrics like goals, assists, and appearances, as well as contextual factors such as a player's league, club reputation, and international exposure. However, despite its robust framework, Transfermarkt acknowledges that market valuations are inherently subjective, particularly for players in underrepresented leagues or emerging football markets. This highlights a potential limitation in its community-driven model, as expert consensus in these areas may be less reliable (Transfermarkt 2024).

Additionally, Transfermarkt actively updates its database to account for real-time changes, such as transfers, injuries, or tactical shifts in a player's role. This ensures that the platform remains relevant and up-to-date, making it a preferred choice for analysts and clubs alike (Transfermarkt 2024).

A.2.2 Stathead FBref: Detailed Analytics and Historical Depth

Stathead FBref, managed by Sports Reference LLC, is a platform that specializes in providing both traditional and modern soccer statistics. The platform's data collection process is rigorous, relying on official league partnerships, historical archives, and real-time match feeds supplied by industry leaders such as StatsBomb and Opta. These partnerships allow FBref to maintain a dataset that spans over 300 leagues worldwide, encompassing both historical and contemporary records (Stathead 2024).

Stathead FBref offers an extensive range of statistics that can be divided into traditional and modern metrics. Traditional metrics include goals, assists, appearances, fouls, and shots, while modern analytics examine measures such as Expected Goals (xG), Expected Assists (xA), Shot-Creating Actions (SCAs), and progressive passes. These modern metrics are designed to provide deeper awareness into a player's contribution beyond surface-level statistics, offering a detailed understanding of performance. For instance, xG and xA quantify the quality of scoring opportunities created and converted, accounting for factors such as shot angle, distance, and defensive pressure (Stathead 2024).

The platform is particularly strong in contextualizing player performance within tactical frameworks. Positional data, for example, helps users understand a player's role in various formations, while heatmaps and passing networks demonstrate spatial tendencies and team dynamics. This level of detail is invaluable for analysts seeking to evaluate players in specific systems or roles (Stathead 2024).

To ensure the reliability of its data, Stathead FBref employs a robust validation process. Match data is cross-referenced with multiple sources, and discrepancies are resolved through manual audits and automated algorithms. Proprietary tools are used to process raw match feeds into user-friendly formats, ensuring that the data presented is both accurate and accessible. Despite this rigor, Stathead FBref acknowledges the challenges of maintaining consistency across leagues with varying levels of data availability, which can introduce gaps in the dataset (Stathead 2024).

A.2.3 FIFA: Elo-Based Rankings for National Teams

FIFA's ranking system for men's national teams represents a methodologically robust application of the Elo model, originally developed for chess. The rankings are designed to provide a quantitative measure of team strength based on match outcomes, with adjustments for factors such as opponent strength, match importance, and confederation weightings. This approach allows FIFA to rank teams in a way that reflects both current form and historical performance (FIFA 2024).

The ranking process begins with the collection of match results, which are submitted by FIFA's member associations through standardized reports. These reports include detailed information on match outcomes, player statistics, and disciplinary records, ensuring consistency across competitions. FIFA's database is updated regularly, incorporating results from international friendlies, qualifiers, and tournament matches (FIFA 2024).

The Elo-based formula used by FIFA calculates the points gained or lost by a team after each match (FIFA 2024). This formula accounts for:

1. **Match Result (M)**: Teams earn more points for a win and lose fewer points for a draw or defeat.
2. **Match Importance (I)**: Matches in the FIFA World Cup carry higher weight compared to friendlies or regional qualifiers.
3. **Opponent Strength (T)**: The strength of the opposing team is derived from its current FIFA ranking.
4. **Regional Strength (C)**: Confederations are assigned weightings to account for the relative competitiveness of different regions.

The formula is expressed as:

$$P = M \times I \times T \times C \times 100$$

While FIFA's methodology is widely regarded as transparent and systematic, it has faced criticism for favoring teams that play more matches or those in competitive confederations. FIFA has addressed these concerns by periodically refining the algorithm and ensuring that its assumptions remain aligned with the evolving dynamics of international football.

A.2.4 FootballDatabase: Club and Player Ratings and Rankings

FootballDatabase employs a similar Elo-based approach to assess the performance of clubs and players worldwide. The platform aggregates data from official league sources, media reports, and proprietary tracking systems, enabling it to maintain a database that spans thousands of matches and hundreds of leagues (FootballDatabase 2024).

The rating system used by FootballDatabase assigns numerical scores to clubs and players based on match outcomes, with higher scores awarded for victories against stronger opponents. Factors considered in the Elo formula include the type of match (e.g., domestic league, continental competition, or friendly), the relative strength of the opponent, and the context of the match (e.g., finals vs. qualifiers) (FootballDatabase 2024).

FootballDatabase also incorporates historical data to provide a longitudinal perspective on performance trends. This allows users to analyze changes in team and player ratings over time, offering revelation into development trajectories and competitive dynamics (FootballDatabase 2024).

One of the platform's strengths is its commitment to transparency. FootballDatabase provides detailed documentation of its methodology, ensuring that users can interpret ratings within the appropriate competitive context. The platform's regular updates ensure that its ratings remain reflective of current performance, while its historical depth makes it a useful resource for longitudinal analyses.

A.3 Detailed Process of Data Cleaning

The data cleaning process is essential to prepare the raw datasets for accurate and meaningful analysis. This involves reading raw data files, merging datasets, handling missing or inconsistent values, transforming and standardizing data, and saving the cleaned data for future use. Below is a detailed description of each step involved in cleaning the player data from two sources: market value data and performance data.

1. Workspace Setup

The first step is setting up the environment by loading the necessary libraries:

- `tidyverse`: A collection of R packages for data science that provides tools for data manipulation and visualization.
- `dplyr`: Part of the tidyverse, it offers a grammar for data manipulation.
- `arrow`: Provides support for reading and writing Parquet files, which are efficient for storage and analysis.

The data files are organized in a structured directory, with raw data located in `data/01-raw_data/` and cleaned data to be saved in `data/02-analysis_data/`.

2. Reading and Merging Datasets

The cleaning function begins by reading the raw market value data and performance data from their respective CSV files using `read.csv()`. To ensure consistency:

- In the performance data, the "Player" column is renamed to "Name" to match the market value data.
- Unnecessary columns such as "Season" and "Comp" are removed from the performance data.

An inner join is performed on the "Name" column using `inner_join()`, which merges the datasets and retains only the players present in both datasets. This ensures the analysis focuses on players with complete information.

3. Filtering and Excluding Irrelevant Data

To enhance data quality, several filters are applied:

- **Removing Empty Strings**: Rows where any column has an empty string are removed using `filter_all(all_vars(. != ""))`.
- **Excluding Missing Market Values**: Rows where "Market.Value" is "-" are removed, as these entries lack valid market value data.
- **Excluding Goalkeepers**: Rows where the "Pos" (position) is "GK" are filtered out using `filter(Pos != "GK")`, since goalkeepers may require separate analysis due to their unique role.
- **Removing Ambiguous Team Entries**: Rows where "Team" is "2 Team" or "2 Teams" are excluded to avoid ambiguity in team identification.

4. Extracting ISO Country Codes

The "Nation" column may contain additional information along with the country code. To standardize this:

- A regular expression is used to extract the three-letter ISO country code from the "Nation" column: `str_extract(Nation, "\\b[A-Z]{3}\\b")`.
- Rows with missing ISO codes are dropped using `drop_na(Nation)` to ensure subsequent transformations have valid data.

5. Transforming ISO Codes to Full Country Names

The extracted ISO codes are mapped to full country names to standardize the data. This is achieved using the `recode()` function, which replaces each ISO code with its corresponding country name. This step is important for:

- **Data Consistency**: Ensures all country references are uniform across the dataset.
- **Facilitating Merging with External Data**: Full country names are necessary to merge with datasets like national rankings.

6. Mapping Country Names to National Team Rankings

After standardizing country names, each country is mapped to its national team ranking:

- A predefined mapping of country names to their FIFA national team rankings is created.
- The `recode()` function is used to replace country names in the "Nation" column with their corresponding rankings.
- This converts the "Nation" column into a numeric variable "national_team_ranking", representing each player's national team's standing.

7. Standardizing Club Names and Mapping to Club Rankings

Similarly, club names in the "Team" column are standardized and mapped to their club rankings:

- A mapping of club names to their global rankings is established.
- The `recode()` function replaces club names with their rankings, transforming the "Team" column into a numeric "club_ranking" variable.
- This step ensures that each player's club performance level is quantitatively represented.

8. Removing Unnecessary Columns

Columns that are not required for the analysis are removed to streamline the dataset:

- Columns such as "MP", "X90s", "Starts", "Subs", "unSub", "G.A", "G.PK", "PK", "PKatt", "PKm", and "X.9999" are dropped using `select(-c(...))`.
- This focuses the dataset on variables relevant to player valuation and performance.

9. Converting Market Values to Numeric Format

The "Market.Value" column contains values with currency symbols and units (e.g., "€50m", "€500k"). To make this data usable:

- The "€" symbol is removed using `gsub("€", "", Market.Value)`.
- A conditional transformation is applied:
  - Values ending with "m" (millions) are converted by removing the "m" and multiplying by 1,000,000.
  - Values ending with "k" (thousands) are converted by removing the "k" and multiplying by 1,000.
  - Values without "m" or "k" are treated as numeric.
- This ensures all market values are numeric and represent the actual amounts in Euros.

10. Renaming and Formatting Columns

To enhance clarity and consistency:

- Columns are renamed using `rename()`:

  - "Market.Value" to "market_value"
  - "Age" to "age"
  - "Nation" (now containing rankings) to "national_team_ranking"
  - "Team" (now containing rankings) to "club_ranking"
  - "Min" to "minutes_played"
  - "Gls" to "goals"
  - "Ast" to "assists"
  - "Pos" to "position"

- Data types are adjusted using `mutate()` to ensure all numeric columns are correctly typed:
  - "national_team_ranking", "age", "club_ranking", "minutes_played", "goals", and "assists" are converted to numeric.

- Position labels are standardized:
  - Dual positions like "DFMF" are recoded to "MFDF" for consistency.

11. Handling Missing Values and Duplicates

- Rows with any missing values are removed using `drop_na()` to ensure data completeness.
- Duplicate entries are handled by grouping by "Name" and filtering:
  - `group_by(Name)` groups the data by player name.
  - `filter(n() == 1)` keeps only those players who appear exactly once, removing duplicates.

12. Finalizing the Dataset

At this stage, the dataset is clean, standardized, and ready for analysis:

- All variables are correctly formatted and named.
- Only relevant data is retained.
- Players are uniquely represented, with no duplicates or missing values.

13. Saving the Cleaned Data

The cleaned dataset is saved in Parquet format using `write_parquet()`:

- Parquet is chosen for its efficient storage and quick read/write capabilities, which is beneficial for large datasets.
- The cleaned data is saved to the `data/02-analysis_data/` directory with a filename that includes the country name (e.g., "cleaned_england_data.parquet").

14. Applying the Cleaning Function to Each Country

The cleaning process is encapsulated in a function `clean_data()` to ensure consistency across different datasets. This function is applied to each country's data:

- **England**:
  - Market value data: "raw_england_market_value_data.csv"
  - Performance data: "raw_england_performance_data.csv"
  - Output: "cleaned_england_data.parquet"
- **Germany**, **Italy**, **Spain**, and **France**:
  - Similar file naming conventions are used for each country.
  - The function is called with the respective file paths for each country's datasets.

By using a function, we ensure that the same cleaning steps are consistently applied to all datasets, which is important for comparative analysis across countries.

This careful data cleaning process ensures that the datasets are accurate, consistent, and enriched with useful contextual information like national team and club rankings. By standardizing country and club names, transforming and formatting variables, and handling missing data and duplicates, we prepare the datasets for robust and reliable analysis. The cleaned data sets the foundation for meaningful observations into player valuations and performance across different leagues.

A.3.1 Reflection on Challenges in Dataset Integration

During the integration of soccer datasets, a major challenge encountered was the inconsistency in naming conventions across different datasets. Examples include:

- **Team Names:** Variants such as "Tottenham Hotspurs" and "Tottenham," or "Manchester United" and "Man Utd," require thorough mapping and standardization.
- **Country Names:** Differences like "N. Ireland" versus "Northern Ireland" and "Turkey" versus "Turkiye" created difficulties in ensuring accurate joins.
- **Standardization Gaps:** The lack of universal naming conventions in soccer data necessitated the creation of custom dictionaries and mappings, increasing the complexity and time required for data cleaning.

The effort to merge and clean inconsistent datasets underscores the need for standardized naming conventions in soccer data. If universally accepted standards for player, team, and country names are adopted across platforms, it would significantly ease data analysis and enhance cross-platform comparability for soccer enthusiasts and researchers. Establishing a global soccer data protocol could streamline workflows and foster more accessible observations into the sport.

A.4 Datasheets

Datasheets for the cleaned datasets used in this study are available in the directory `other/datasheet`. These datasheets are modeled on the questions and structure proposed by Gebru et al. (2021) in their seminal work on datasheets for datasets, ensuring transparency and accountability in data use.

Each datasheet provides detailed documentation for the respective dataset (e.g., England, France, Germany, Italy, and Spain), covering the motivation, composition, collection process, preprocessing/cleaning, uses, distribution, and maintenance. The following is an overview of the key aspects included in the datasheets:

- **Motivation**: Discusses the purpose of dataset creation, including its intended tasks such as analyzing soccer players' market value and identifying key determinants.
- **Composition**: Describes the data's structure, including player records, variables such as performance metrics, demographic features, and target attributes like market value.
- **Collection Process**: Details the methods and sources used to compile the dataset, including web scraping and manual curation from Transfermarkt, Stathead, FIFA, and Football-database.
- **Preprocessing/Cleaning**: Outlines the steps undertaken to clean and preprocess the data for analysis, such as unifying ISO country codes, mapping country names to rankings, and normalizing team names.
- **Uses and Limitations**: Identifies current and potential use cases for the dataset, as well as tasks for which the dataset might not be suitable.
- **Distribution and Maintenance**: Provides information on how the dataset is distributed (e.g., open-access via GitHub), licensing details, and mechanisms for updates or contributions.

For full details, refer to the individual datasheets, accessible in PDF format in the `other/datasheet` directory. Each datasheet ensures that the datasets are well-documented, promoting transparency and encouraging ethical usage.

A.5 Model Results

Table 7: Regression Results for Premier League (England)

Variable	Estimate	Standard Error	P-Value
(Intercept)	41.057	4.663	0.000
age	-1.137	0.173	0.000
goals	2.993	0.292	0.000
assists	1.567	0.408	0.000
club_ranking	-0.027	0.004	0.000

Variable	Estimate	Standard Error	P-Value
national_team_ranking	-0.065	0.032	0.044
minutes_played	0.004	0.001	0.000
positionFW	-9.608	2.661	0.000
positionFWDF	-7.253	4.575	0.114
positionFWMF	-3.811	2.300	0.098
positionMF	2.019	2.083	0.333
positionMFDF	-2.042	2.822	0.470

Table 8: Regression Results for Ligue 1 (France)

Variable	Estimate	Standard Error	P-Value
(Intercept)	22.946	3.193	0.000
age	-0.593	0.128	0.000
goals	2.027	0.237	0.000
assists	1.715	0.430	0.000
club_ranking	-0.016	0.003	0.000
national_team_ranking	-0.036	0.018	0.051
minutes_played	0.001	0.001	0.429
positionFW	-7.041	2.085	0.001
positionFWDF	-4.372	2.938	0.138
positionFWMF	-5.011	1.642	0.002
positionMF	-2.938	1.500	0.051
positionMFDF	-1.905	2.148	0.376

Table 9: Regression Results for Bundesliga (Germany)

Variable	Estimate	Standard Error	P-Value
(Intercept)	28.663	3.917	0.000
age	-0.848	0.150	0.000
goals	1.474	0.227	0.000
assists	1.828	0.356	0.000
club_ranking	-0.017	0.004	0.000
national_team_ranking	0.012	0.035	0.726
minutes_played	0.001	0.001	0.152
positionFW	-2.081	2.557	0.416
positionFWDF	-7.835	3.077	0.011
positionFWMF	-5.238	1.845	0.005
positionMF	-1.469	1.817	0.419

Variable	Estimate	Standard Error	P-Value
positionMFDF	-2.433	2.127	0.253

Table 10: Regression Results for Serie A (Italy)

Variable	Estimate	Standard Error	P-Value
(Intercept)	26.361	3.038	0.000
age	-0.829	0.113	0.000
goals	2.489	0.216	0.000
assists	0.820	0.353	0.021
club_ranking	-0.018	0.003	0.000
national_team_ranking	-0.014	0.020	0.487
minutes_played	0.002	0.001	0.001
positionFW	-2.247	1.694	0.185
positionFWDF	-2.327	3.227	0.471
positionFWMF	-4.877	1.510	0.001
positionMF	-1.911	1.268	0.132
positionMFDF	-1.933	2.183	0.376

Table 11: Regression Results for La Liga (Spain)

Variable	Estimate	Standard Error	P-Value
(Intercept)	35.887	4.505	0.000
age	-1.179	0.160	0.000
goals	1.273	0.296	0.000
assists	1.548	0.502	0.002
club_ranking	-0.034	0.006	0.000
national_team_ranking	-0.020	0.037	0.593
minutes_played	0.004	0.001	0.001
positionFW	-4.474	2.788	0.109
positionFWDF	0.062	5.845	0.992
positionFWMF	2.013	2.393	0.401
positionMF	1.170	2.029	0.565
positionMFDF	6.327	3.004	0.036

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

FIFA. 2024. "FIFA Men's World Ranking." https://inside.fifa.com/fifa-world-ranking/men.

FootballDatabase. 2024. "FootballDatabase." https://footballdatabase.com/.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

Kuhn, and Max. 2008. "Building Predictive Models in r Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26. https://doi.org/10.18637/jss.v028.i05.

Müller, Kirill, and Lorenz Walthert. 2024. *Styler: Non-Invasive Pretty Printing of r Code*. https://CRAN.R-project.org/package=styler.

Python Software Foundation. 2023. *Python: A Powerful Programming Language*. https://www.python.org/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Robinson, David, Alex Hayes, and Simon Couch. 2024. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Stathead. 2024. "Stathead Soccer - Comprehensive Soccer Stats from FBref.com." https://stathead.com/fbref/.

Transfermarkt. 2024. "Transfermarkt." https://www.transfermarkt.com/.

van der Loo, Mark P. J., and Edwin de Jonge. 2021. "Data Validation Infrastructure for R." *Journal of Statistical Software* 97 (10): 1–31. https://doi.org/10.18637/jss.v097.i10.

Wickham, Hadley. 2011. "Testthat: Get Started with Testing." *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

———. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2024. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.