Predicting the Market Value of Soccer Players based on Performance Metrics*

Rahil Chadha, Harnehmat Kaur, Tanmay Shinde

October 10, 2024

Contents

1 Contributions

Rahil Chadha: In this project, I came up with the research question and found the dataset, ensuring it met the project requirements. I conducted a comprehensive review of peer-reviewed literature, summarizing key findings that informed our predictors. I also outlined the justification for using linear regression, explaining its suitability for answering the research question. Additionally, I contributed to the data analysis plan and ensured the correct citation of data sources.

Harnehmat Kaur: Conducted analysis and visualization of the response variable and key predictors for estimating football player market values. Explored attributes such as age, sprint speed, agility, and release clauses to determine their impact on market value. Created visualizations to identify patterns in the data and evaluate the suitability of these predictors for inclusion in the regression model based on their distributions. Contributed to the research component by finding and reviewing relevant research papers for the project.

Tanmay Sachin Shinde: Alongside my teammates, I contributed to cleaning and preparing the data for analysis. I worked on writing the code for fitting the model and performing residual analysis to assess model assumptions and summarized the conclusion made for each assumption and condition. I interpreted the results from the preliminary model in context of the research question and summarized the effect of predictors on the response variable. Additionally, I ensured the trustworthiness and ethical use of the dataset, incorporating the insights from the embedded ethics module held during our lecture. Finally, I formatted the R Markdown (.rmd) file for our submission, ensuring it was well-organized.

*Code and data are available at: PredictingMarketValueOfSoccerPlayers

2 Introduction

The European football transfer market is a multi-billion-euro industry, and accurately assessing a players market value is crucial for clubs aiming to make strategic decisions around team building. This project aims to predict a football players' market value based on key attributes such as the players age, on-field position, his release clause value, and performance attributes such as sprint speed, and dribbling ability.

By understanding the impact of key variables like age, position, and performance metrics on market value, clubs can strategically allocate resources and negotiate better deals. Key findings from three peer-reviewed articles in the field of sports analytics and market value determination suggest that the age of a player and market value are negatively correlated, noting that forwards peak in value between the **age** of 20-25, defenders around 27, and goalkeepers at 33 (Rong, Wang, and Xie 2024). This paper also notes that players playing in forwards tend to have higher market values due to their goal-scoring roles, whereas midfielders and defenders are valuated less (**position**). Apart from these, the transfer fee (**release clause value**) of a player is a crucial predictor of demand and market value (Prayoga, Sudrajat, and Azhar 2023). Lastly, inherent physical attributes and skills such as **sprint speed**, **dribbling skills**, and agility are also considered to be key predictors of players performance and market value (Poza 2020).

Using a linear regression model enables us to quantify how changes in performance indicators, such as age, sprint speed, dribbling skills, and release clauses impact a player's market value. Estimating a linear trend is particularly useful for our research question as it allows us to ascertain the direction and strength of the relationships between the variables. By fitting a linear model, we can derive coefficients that represent the expected change in market value associated with a one-unit change in each performance metric. In this analysis, the primary focus will be on interpretability. While precise predictions of market value are valuable, by interpreting the coefficients of the linear regression model, we can explain how specific aspects of player performance influence market valuations.

3 Data Description

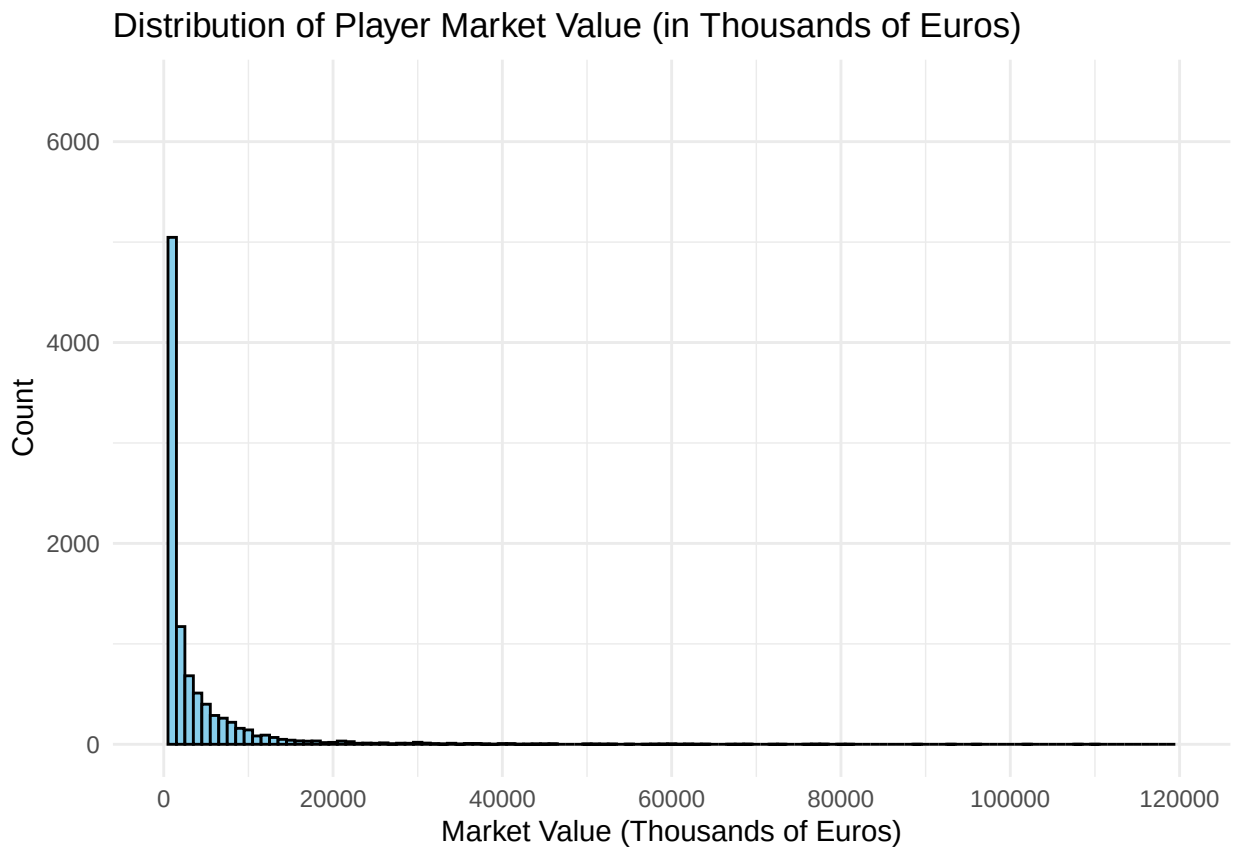
3.1 Source and Data Collection

The dataset for this project was obtained from Kaggle, and can be found here Kaggle - Football Players Data (Ahmed 2023). The original curators of the dataset obtained the data from SoFIFA (SoFIFA 2024), a website that provides detailed player information for FIFA video game enthusiasts. The data reflects real world statistics and performance metrics of players and was originally collected to help players build teams in the FIFA video game. In our project, we use this extensive data to predict football players' market value in the real transfer market, shifting the focus from gaming to real-world applications.

3.2 Response Variable Summary

Table 1: Market Value Statistics (Values in Thousands of Euros)

Mean Market Value	Median Market Value	Standard Deviation	Min Market Value	Max Market Value
2465.353	675	5828.476	10	110500



Description: The distribution of player market values is highly right-skewed, with most players having low values and only a few with extremely high values. This distribution suggests the presence of a few outliers or highly valued players that skew the data.

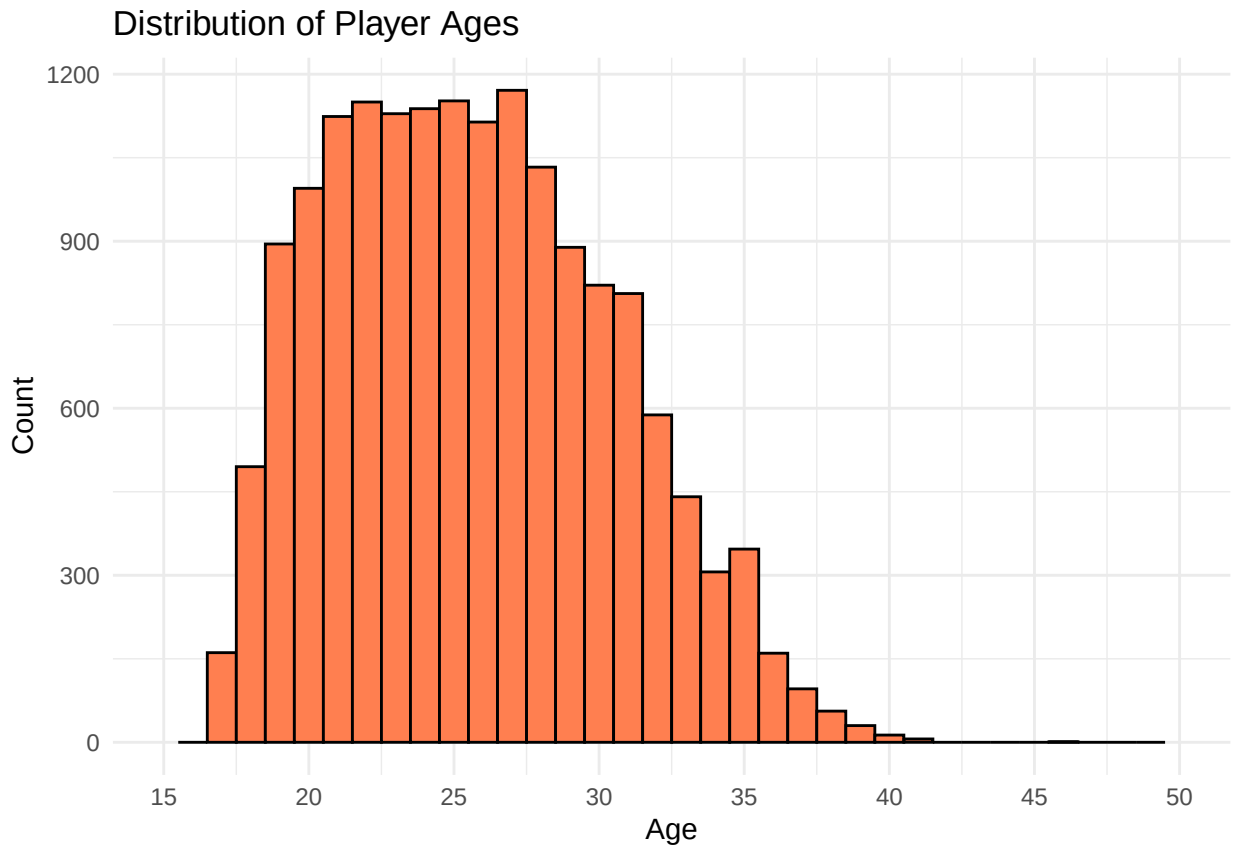
Relevance: This response variable is appropriate for linear regression, as market value is a central measure of player importance and performance in the football world. As seen in *papers 2 and 3*, a log transformation might stabilize the variance, making the relationship between predictors and market value clearer. Log transformations are commonly used in economic and financial data to handle skewed distributions and meet regression assumptions. The transformed variable is expected to have a more normal distribution, improving the model fit.

3.3 Predictor Variable Summary

3.3.1 Age

Table 2: Age Summary Statistics

Mean Age	Median Age	Standard Deviation	Min Age	Max Age
25.67333	25	4.773	17	46



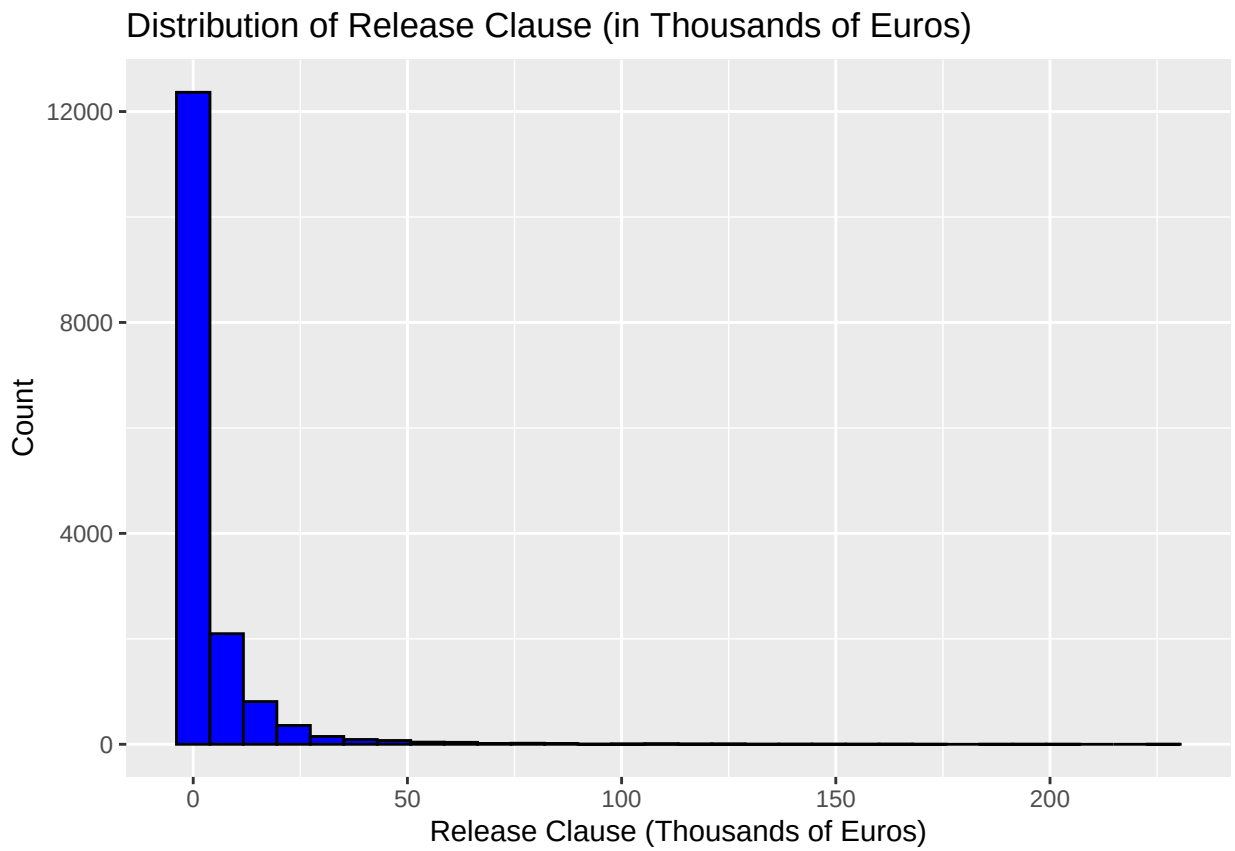
Description: The age distribution is right-skewed, with most players between 20 and 30 years old.

Relevance: Age is an important predictor because younger players are often valued for their potential, while older players may have experience that increases their market value. *All of the research papers* conclude that there is a relationship between age and market value, and *paper 2* suggests that it is mostly a negative relationship. This variable warrants closer examination in the modelling phase.

3.3.2 Release Clause Value

Table 3: Release Clause Summary Statistics (Values in 1000s of Euros)

Mean Release Clause	Median Release Clause	Standard Deviation	Min Release Clause	Max Release Clause
4622.522	1200	11290.77	13	226500



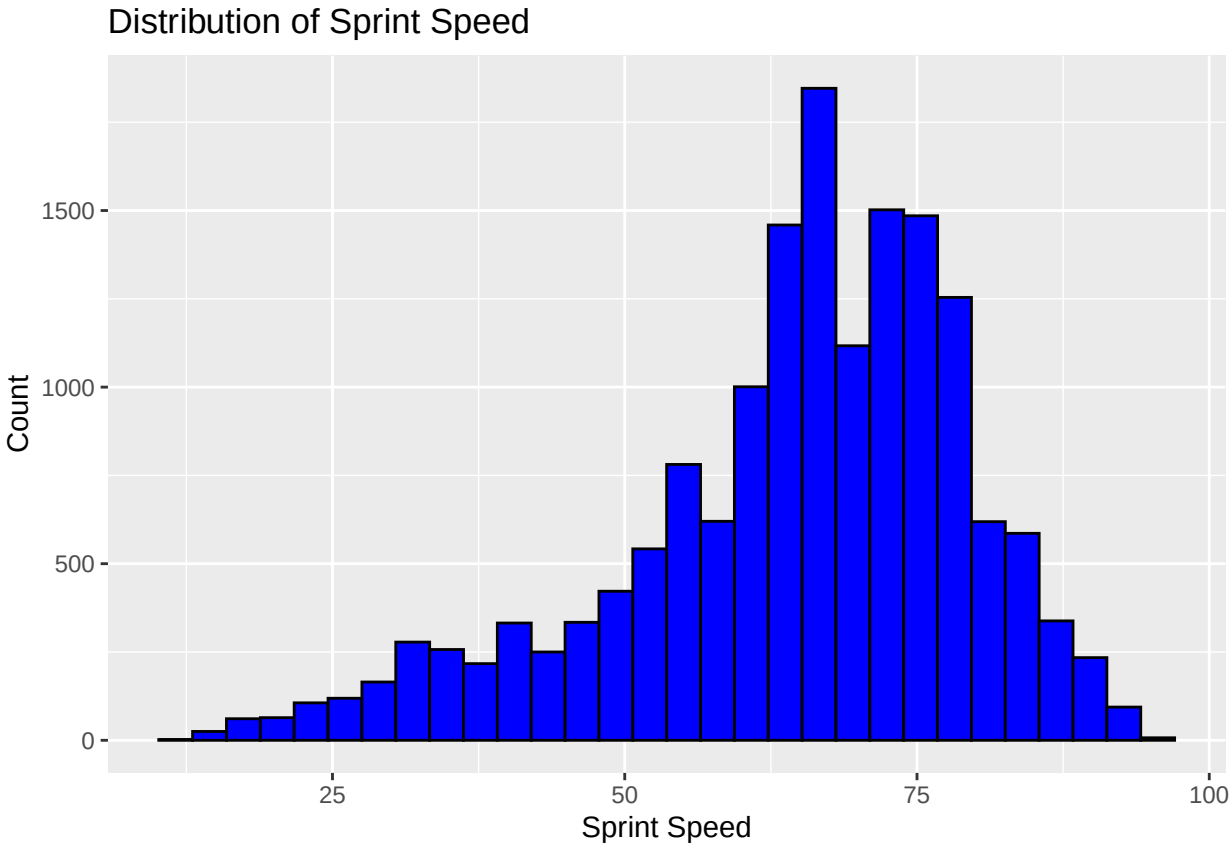
Description: The distribution of release clause values is highly right-skewed, with a vast majority of players having a low release clause, while a few players have extremely high clauses. The values are now expressed in thousands of euros, which makes it easier to interpret the data, but there is still a long tail of higher values.

Relevance: The release clause is a direct monetary representation of the player's perceived market worth, so it's highly relevant for predicting the market value. This variable's relevance to our analysis is also highlighted in *paper 2*.

3.3.3 Sprint Speed

Table 4: Sprint Speed Summary Statistics

Mean Sprint Speed	Median Sprint Speed	Standard Deviation	Min Sprint Speed	Max Sprint Speed
64.48675	67	14.90737	12	96



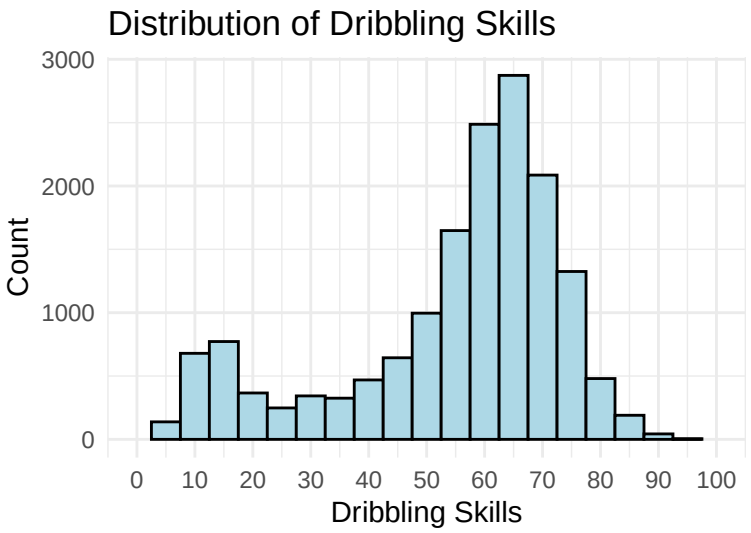
Description: The sprint speed distribution is approximately normal but only slightly left skewed, peaking between 70 and 80. Most players have sprint speeds between 50 and 80, with very few players at the extremes. The smooth distribution makes this a well-behaved variable for linear regression. There is no obvious skewness, so transformation is likely unnecessary.

Relevance: Sprint speed is a key performance metric that influences player value, especially for forwards and wingers. Faster players are often more valuable as they can offer more offensive opportunities. This makes sprint speed a strong predictor for market value, as seen in *paper 3* as well.

3.3.4 Dribbling

Table 5: Dribbling Summary Statistics

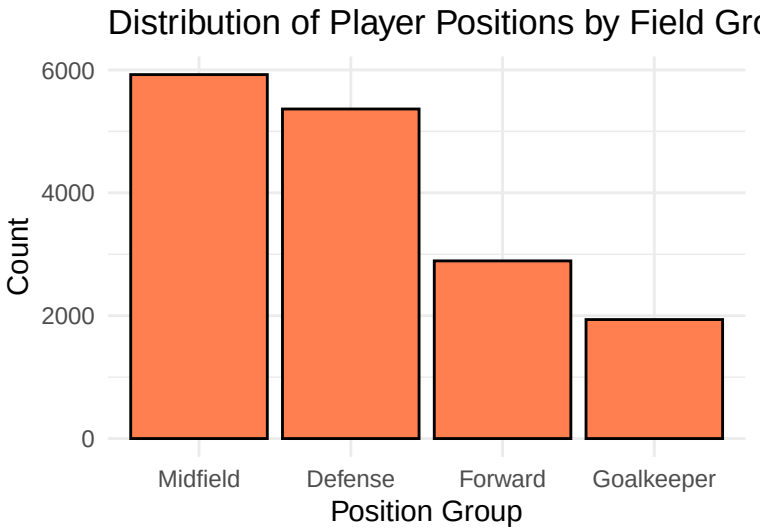
Mean Dribbling	Median Dribbling	Standard Deviation	Min Dribbling	Max Dribbling
54.75876	61	19.29033	4	97



Description: The distribution of dribbling skills has a peak around the 60-70 range. Most players have a dribbling score between 60 and 70, while fewer players have lower (below 50) or higher (above 80) scores. There are fewer players with very high scores (80+). A small number of players have very low scores, which could represent either younger, less-skilled players, or goalkeepers and defenders.

Relevance: Dribbling is a critical skill, especially for forwards and midfielders, and is likely to have a positive influence on market value, as seen with other similar physical attributes being used in *paper 1*. This predictor is suitable for linear regression and does not require transformation.

3.3.5 Position (Categorical Variable)



Description: Position data was grouped into four main categories (Midfielder, Forward, Defense, Goalkeeper) to simplify the analysis and ensure that each group has enough representation for meaningful comparisons. We can see that Midfielders are the largest group of players, followed by Defenders, Forwards, and finally Goalkeepers.

Relevance: This modification *aligns with the literature in papers 1 and 3*, where similar grouping strategies are used to highlight the differential impact of positions on market value.

4 Ethics Discussion

The dataset used in this project is deemed trustworthy based on several criteria discussed in the ethics module. Firstly, the data was collected from real-world statistics and performance metrics, rather than being simulated, ensuring its relevance and applicability to our study. The metadata is adequately filled out on Kaggle, providing essential context for understanding the variables involved. Furthermore, the dataset's source, SoFIFA, is well-documented and recognized in the football community, indicating its reliability. The dataset has gained popularity and is vetted by third parties on platforms like Kaggle, enhancing its trustworthiness.

From an ethical standpoint, the data pertains to public figures (football players), and informed consent is taken from these players to include their names, personal data, and performance ratings in such datasets. Thus, the dataset's creation and use align with ethical guidelines regarding privacy and consent. The SoFIFA data is already open-source and available on the website, hence the curators that prepared the data on Kaggle have adhered to the ownership rules surrounding the dataset by giving due credit to the SoFIFA website.

5 Preliminary Results

5.0.1 Preliminary Analysis:

Now, we will fit a multiple linear model on the data, using `value_euro` as the response variable, which is the market value of a player in Euros, and `age`, `sprint_speed`, `dribbling`, `position`, and `release_clause_euro` as the predictors. The results of the model are as follows:

Table 6: Regression Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-686.8433504	43.2992629	-15.8627031	0.0000000
age	20.0375584	1.0589523	18.9220588	0.0000000
sprint_speed	1.3893824	0.5090120	2.7295673	0.0063486
dribbling	3.0183487	0.5777461	5.2243519	0.0000002
release_clause_euro	0.5111803	0.0004630	1104.0812062	0.0000000
positionForward	29.0847722	15.4329051	1.8845948	0.0595026
positionGoalkeeper	127.8235843	25.6436817	4.9846035	0.0000006
positionMidfield	-1.7294585	13.3551164	-0.1294978	0.8969654

From fitting a preliminary model, we can see that the expected market value of a player is -686,843.4 Euros (the market value column is in 1000 of Euros) when his age, sprint speed, dribbling skills rating, release clause value are all 0 and the player plays in a Defensive position.

Similarly, we can see that the expected market value of a player who is a Goalkeeper is -559,019.8 Euros (the market value column is in 1000 of Euros) when his age, sprint speed, dribbling skills rating, release clause value are all 0.

The model also shows that holding all other variables constant (age, sprint_speed, release_clause_euro, and position), for each one-unit increase in the dribbling score, the player's value in euros.

From the preliminary results, we can make the following conclusions:

Age: The positive coefficient of 20.04 suggests that for each additional year of a player's age, the market value increases by approximately 20,037.6 Euros. This aligns with findings in the literature, where age is often positively correlated with market value *up to a certain point in a player's career*. However, as players age beyond their peak performance years, the value may decline, indicating a nonlinear relationship not captured in our linear model.

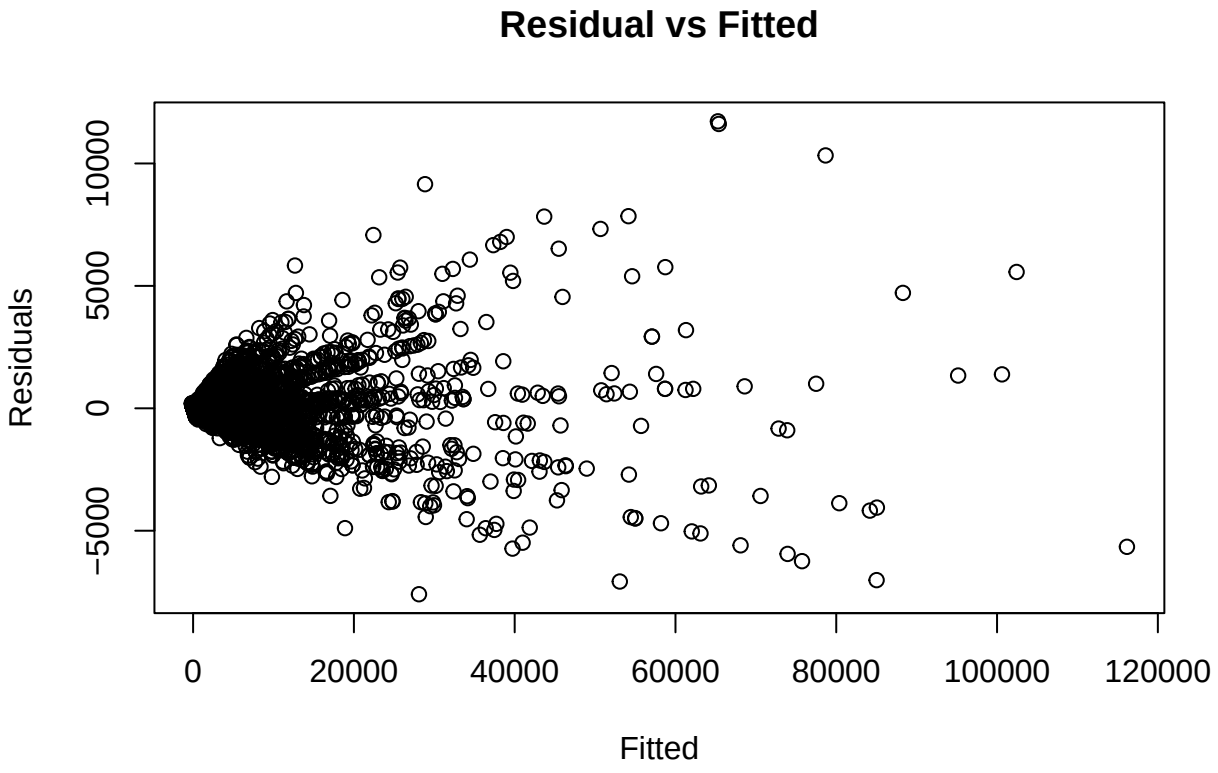
Sprint Speed and Dribbling: The coefficients of 1.39 and 3.02 respectively indicate that an increase in either corresponds to a rise in market value. Our literature review corroborates this, suggesting that speed and dribbling skills are crucial attributes for players, particularly in attacking positions for maintaining possession and creating scoring opportunities.

Release Clause: The extremely high coefficient (0.51) for release clause value (since both are in 1000s of euros) indicates a strong correlation with market value, suggesting that players with higher release clauses are perceived as more valuable in the market. This finding is consistent with the concept that a player's transfer fee often reflects their on-field performance and potential, as discussed in the literature.

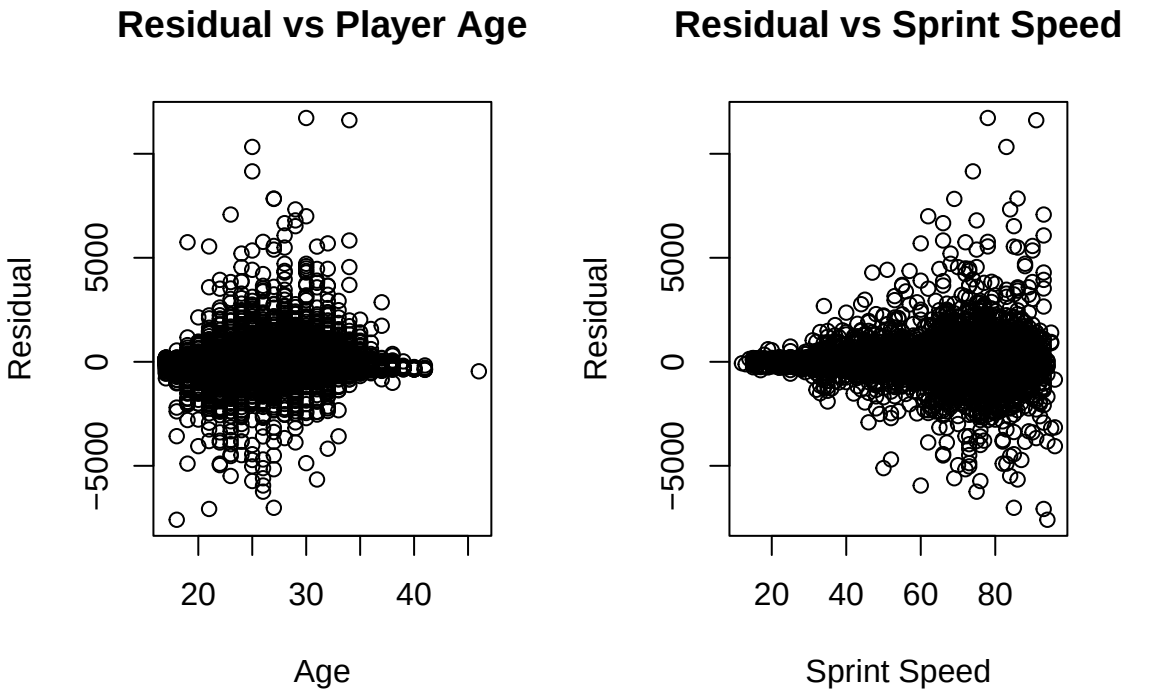
Position: The results for different positions reveal that forwards and goalkeepers have positive coefficients, indicating higher expected market values compared to players in defensive positions. This aligns with the literature, where attacking players often command higher market values due to their direct contribution to goal-scoring, while the value of goalkeepers has gained recognition in recent years.

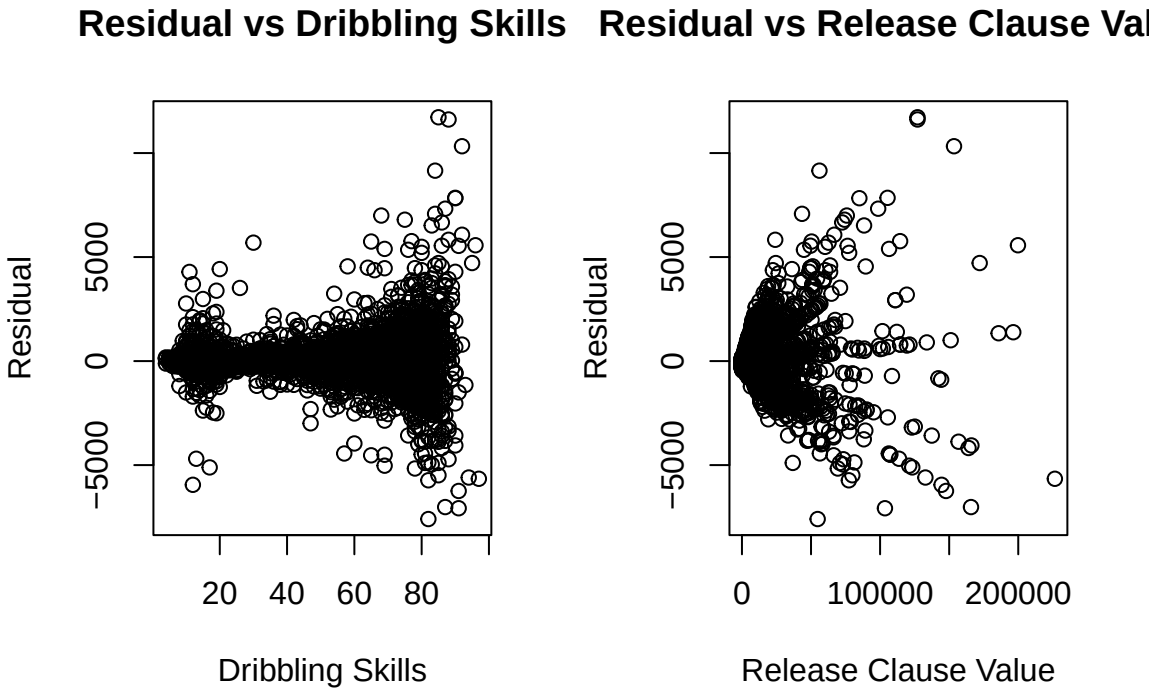
5.0.2 Checking for Violated Assumptions:

First, we analyze the Residual vs Fitted plot to check if any assumptions are violated in the model.



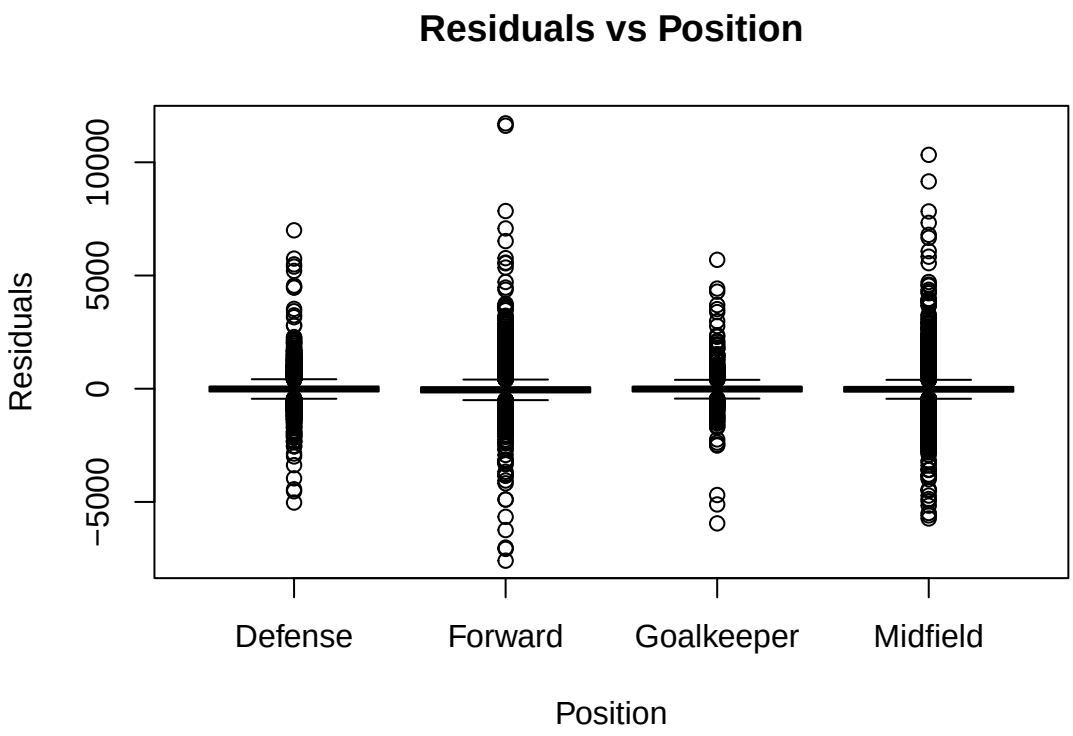
We see a fanning pattern in the Residual vs. Fitted plot, which indicates that there is likely a violation of the constant variance assumption.

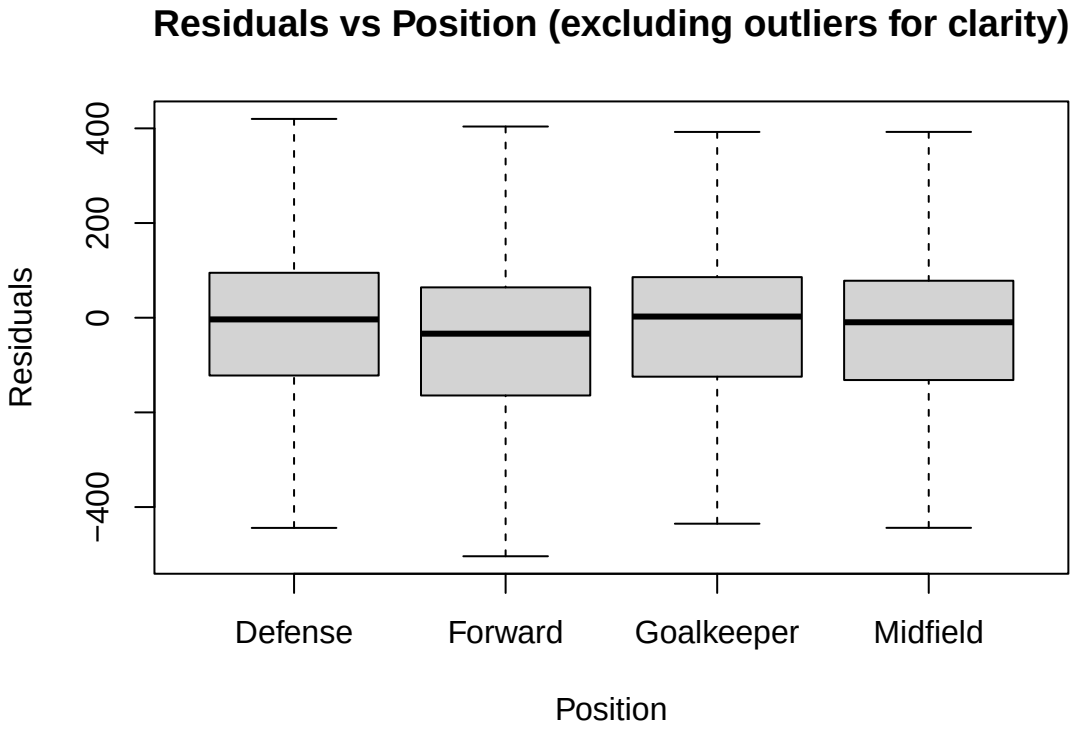


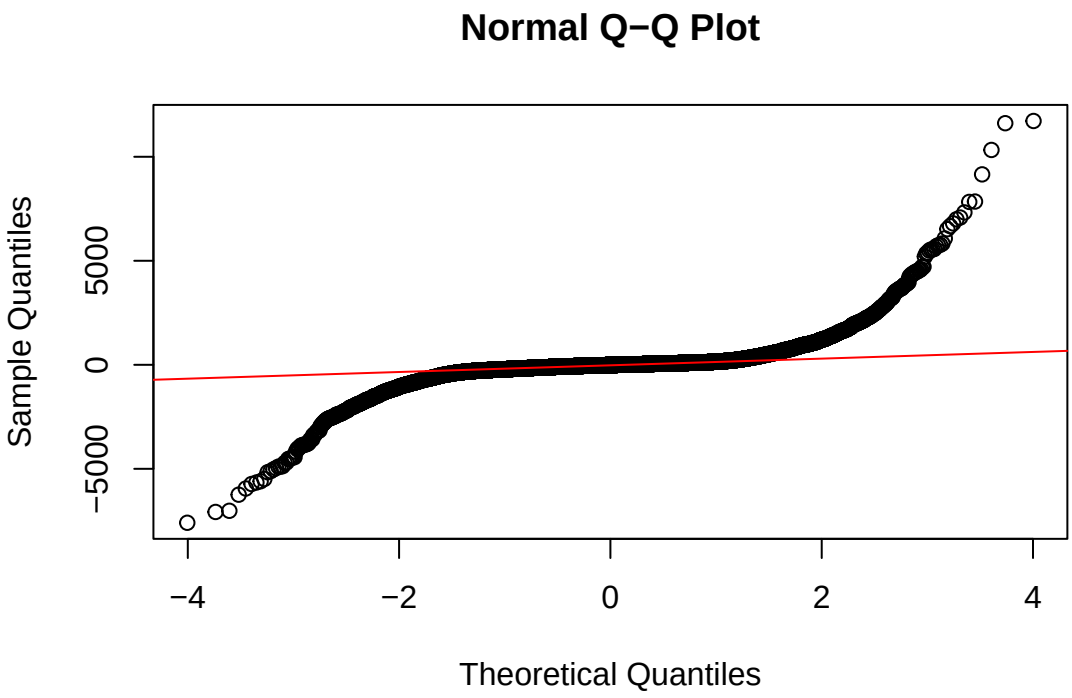


From the above plots, we can see that all the residuals vs. predictor plots show a fanning pattern indicating a violation of constant variance assumption.

Now, we plot the boxplots for residuals vs. our categorical variable - player position.

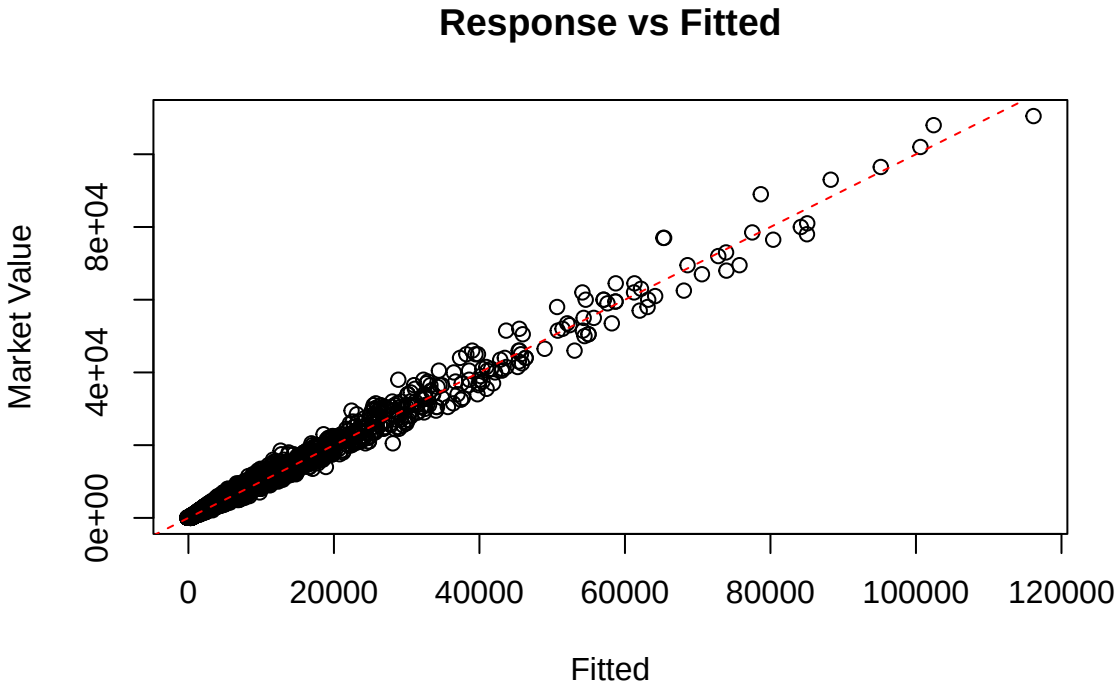

Residuals vs Position
Residuals
−5000
0
5000
10000
Defense
Forward
Goalkeeper
Midfield
Position


Residuals vs Position (excluding outliers for clarity)
Residuals
−400
0
200
400
Defense
Forward
Goalkeeper
Midfield
Position

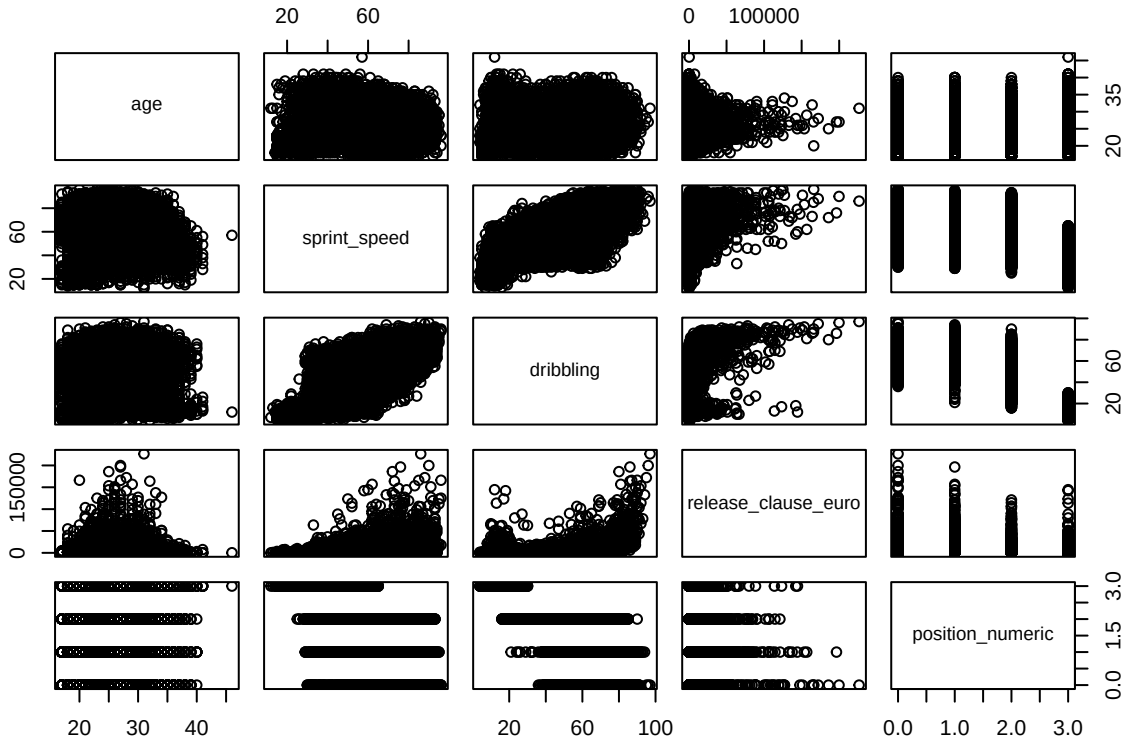


From the normal QQ-plot, we can see that the normality assumption is also violated.

Now, we plot response vs. fitted to check if the conditional mean response condition holds.



From the response vs. fitted graph, we see a fairly random diagonal scatter in the plot and thus our conditional mean response condition holds. We must check the conditional mean predictor condition to ensure that our residual plots are reliable to check for violations in our assumptions.



Most of these plots have a random scatter while some show either a systemic trend/pattern. Overall, from the preliminary examination and exploratory analysis of data, we can see that some of the predictor variables have a heavy skew and certain assumptions have been violated in our model. These violations will need to be corrected before proceeding with further analysis or making any conclusions from the model.

References

Ahmed, Masood. 2023. "Football Players Data." Kaggle. https://doi.org/10.34740/KAGGLE/DSV/6960429.

Poza, Carlos. 2020. "A Conceptual Model to Measure Football Player's Market Value. A Proposal by Means of an Analytic Hierarchy Process. [Un Modelo Conceptual Para Medir El Valor de Mercado de Los Futbolistas. Una Propuesta a Través de Un Proceso Analítico Jerárquico]." *RICYDE. Revista Internacional de Ciencias Del Deporte* 16 (January): 24–42. https://doi.org/10.5232/ricyde2020.05903.

Prayoga, Nanak Andrean, Sudrajat, and Rialdi Azhar. 2023. "The Influence of Performance on the Market Value of Professional Football Players at Football Clubs in Europe 2021/2022 Season." *International Journal of Research Publication and Reviews* 4 (February): 169–76. https://doi.org/https://doi.org/10.55248/gengpi.2023.4206.

Rong, Zhangyi, Lujie Wang, and Shengting Xie. 2024. "Factors That Influence Player Market Value in Different Position: Evidence from European Leagues." *Advances in Economics, Management and Political Sciences* 82 (May): 50–63. https://doi.org/10.54254/2754-1169/82/20230718.

SoFIFA. 2024. "SoFIFA - FIFA 21 Player Ratings Database." https://sofifa.com/.