

STAT 250-APPLIED STATISTIC

TERM PROJECT

DEPARTMENT OF STATISTICS
MIDDLE EAST TECHNICAL UNIVERSITY

BAŞAK KABALOĞLU 2427250

&

MUHAMMED BİLGİN 2428910

Abstract

This research examines the dynamics of player attributes, performance, and earnings in professional football leagues. This analysis aims to identify statistical correlations that could help teams and sports analysts maximize player utilization and recruitment tactics by using a large dataset of players from elite leagues. The study illustrates how age, position, performance metrics, and market value correlate with players' on-field contributions and economic value through techniques like regression analysis, ANOVA, and hypothesis testing for means and proportions. The findings are meant to offer practical insights to improve talent management and strategic planning in professional football.

1.1 Introduction

Professional football, especially in Europe's top leagues, has become a multifaceted industry where player performance, player attributes, and financial valuations are critically intertwined. Clubs invest heavily in scouting, recruiting, and managing players to optimize their on-pitch success while ensuring financial sustainability. Understanding the complex relationships between various player attributes and their economic value is essential for clubs to improve their strategic planning and talent management.

In this study, we examined a comprehensive dataset of players from top-tier football leagues to analyze how factors such as age, position, performance metrics, and market value relate to players' on-pitch contributions and financial value.

1.2. Research questions

1. Is the mean age of players (CurrentAge) significantly different from 28 years?
2. Is there a difference between the average ages of FW (forward) and MF (midfielder)?
3. Is there a significant difference in the proportion of high-performing players (xGoalsAdded > 10) between forwards and defenders?
4. How well does the expected goals per 90 minutes (xGoalsAddedPer90) predict the current market value (€) of football players?
5. Is there a significant difference in the average Salary (\$) of players across different leagues (Bundesliga, MLS, Serie A)?

1.3. Aim of the study

The main purpose of this analysis is to provide actionable insights for clubs and sports analysts, helping them improve their player usage and recruitment strategies. By understanding the statistical relationships between player attributes and performance outcomes, clubs can make more informed decisions that will increase both their competitive advantage and economic efficiency.

2. Methodology/Analysis

In this study, I aimed to answer the following questions by performing various statistical analyses on the dataset of football players using the R programming language and leveraging relevant course materials:

1. Is the mean age of players (CurrentAge) significantly different from 28 years?

- **Analysis Method:** One-sample hypothesis testing
- **Details:** We assumed the mean age of players to be 28 years and tested this hypothesis using a t-test. This analysis provided insights into whether the average age of players deviates significantly from the assumed value of 28 years.

2. Is there a difference between the average ages of forwards (FW) and midfielders (MF)?

- **Analysis Method:** Two-sample hypothesis testing
- **Details:** We conducted two-sample hypothesis testing to determine if there is a statistically significant difference in the mean ages of forwards and midfielders. This analysis helped understand whether player positions (FW and MF) have an effect on the age distribution.

3. Is there a significant difference in the proportion of high-performing players (xGoalsAdded > 10) between forwards and defenders?

- **Analysis Method:** Two-sample hypothesis testing for proportions
- **Details:** We compared the proportions of high-performing players (defined as having xGoalsAdded greater than 10) between forwards and defenders. This test helped in understanding if the likelihood of being a high-performing player differs by position.

4. How well does the expected goals per 90 minutes (xGoalsAddedPer90) predict the current market value (€) of football players?

- **Analysis Method:** Simple linear regression
- **Details:** We performed a simple linear regression analysis to evaluate how well the variable xGoalsAddedPer90 predicts the current market value of players. The

regression model provided insights into the predictor that significantly influences a player's market value.

5. Is there a significant difference in the average Salary (\$) of players across different leagues (Bundesliga, MLS, Serie A)?

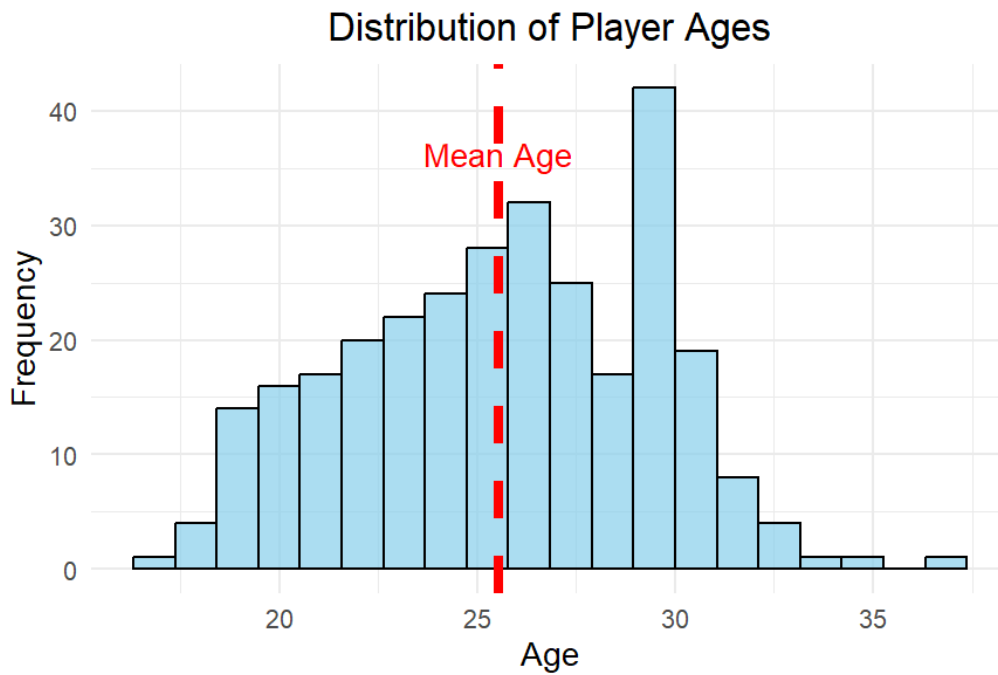
- **Analysis Method:** One-way analysis of variance (One-way ANOVA)
- **Details:** We used one-way ANOVA to test for differences in the average salary of players across selected leagues. This analysis identified whether league membership has an impact on player salaries.

3. Results and Findings

1. Age analysis was performed using the single sample mean inference, assuming the average age data was 28. According to the histogram, it was clear that the distribution was normal. As a result of the hypothesis testing, the null hypothesis could be rejected since $p\text{-value} = 2.2e-16 < \alpha = 0.05$. It shows that the mean age of players (AgeDuringSeason) is significantly different from 28 years. ($t = -10.841, p < 2.2e-16$).

Null Hypothesis (H_0): $\mu = 28$

Alternative Hypothesis (H_1): $\mu \neq 28$



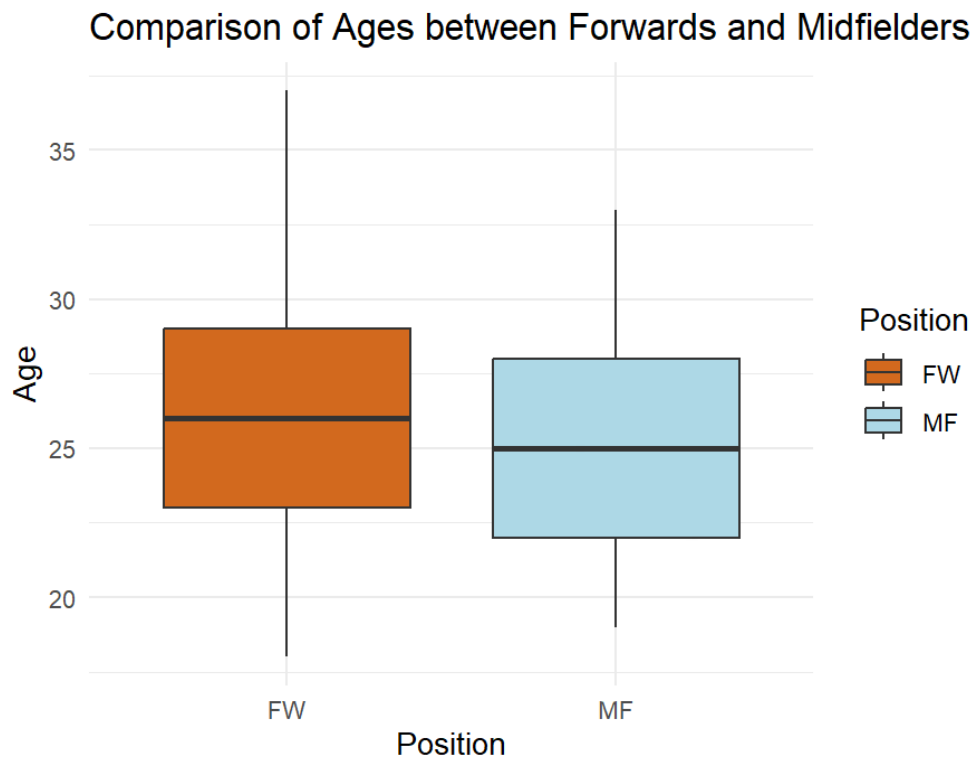
2. According to Shapiro-Wilk normality test, the distribution is normal since $p\text{-value}=0.4768 > \alpha=0.05$. The normality assumption was checked using the Shapiro-Wilk normality test, and the assumption was satisfied. Before we conducted the t-test, we checked whether their variances were equal or not. We used the F-test to test for differences in variances.

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Since $p\text{-value}=0.7128 > \alpha=0.05$, we could not reject H_0 , so, the variances are equal.

The p-value of the two-sample t-test is 0.4852, which is smaller than $\alpha = 0.05$, indicating no statistically significant difference in the average ages of players who are Forwards (FW) compared to those who are Midfielders (MF).



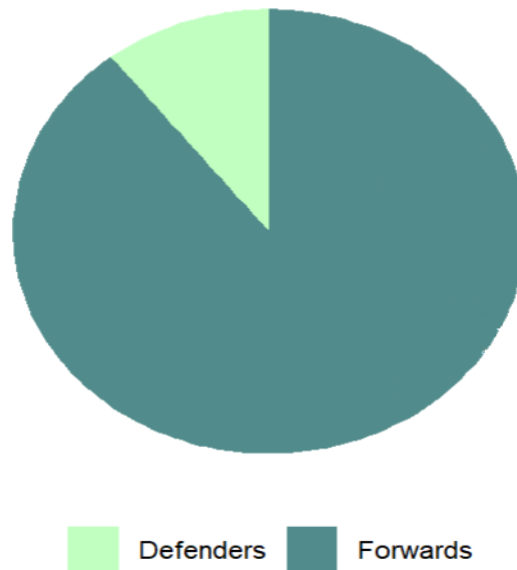
3. We conducted two-sample hypothesis testing for proportions to examine whether there is a significant difference in the proportion of high-performing players between forwards and defenders.

$H_0: p_1 = p_2$ (There is no difference in the proportions of high-performing players between forwards and defenders.)

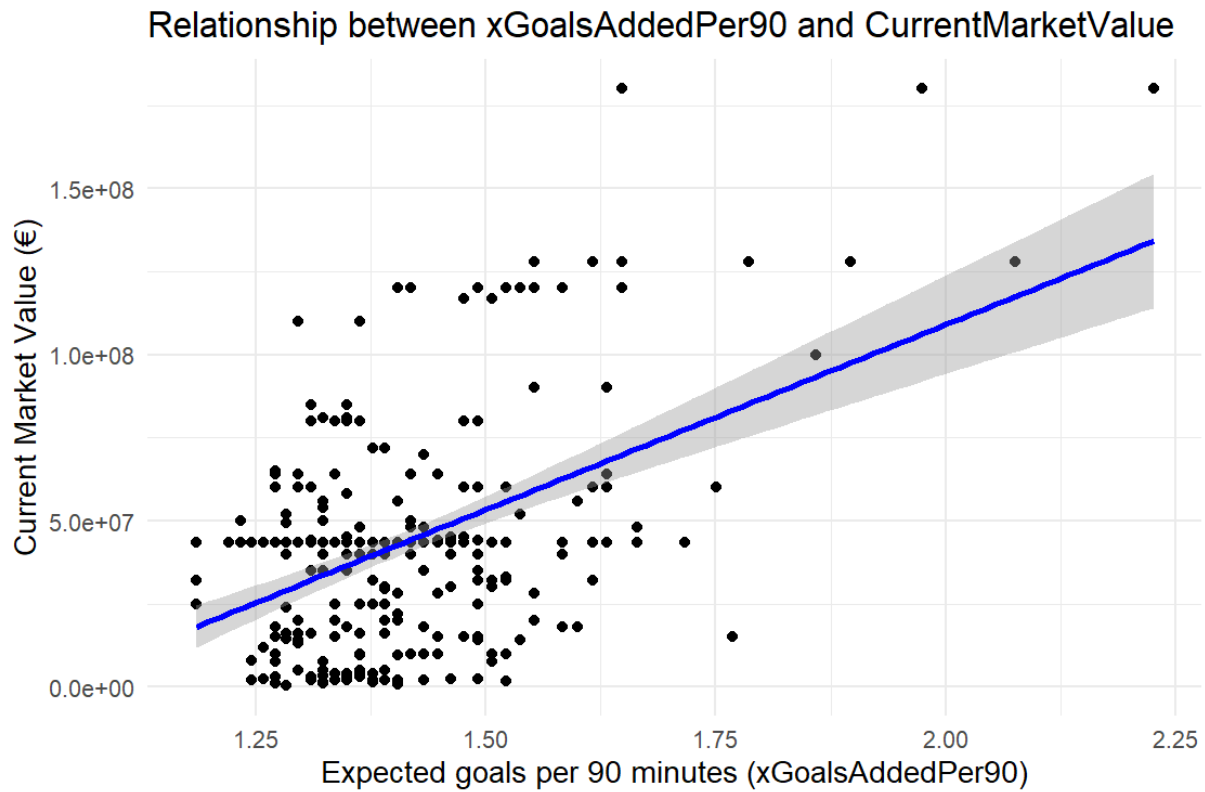
$H_1: p_1 \neq p_2$ (There is a difference in the proportions of high-performing players between forwards and defenders.)

Since $p\text{-value} = 4.048e-06 < \alpha = 0.05$, we could reject H_0 , which means that there is a difference in the proportions of high-performing players between forwards (FW) and defenders (DF).

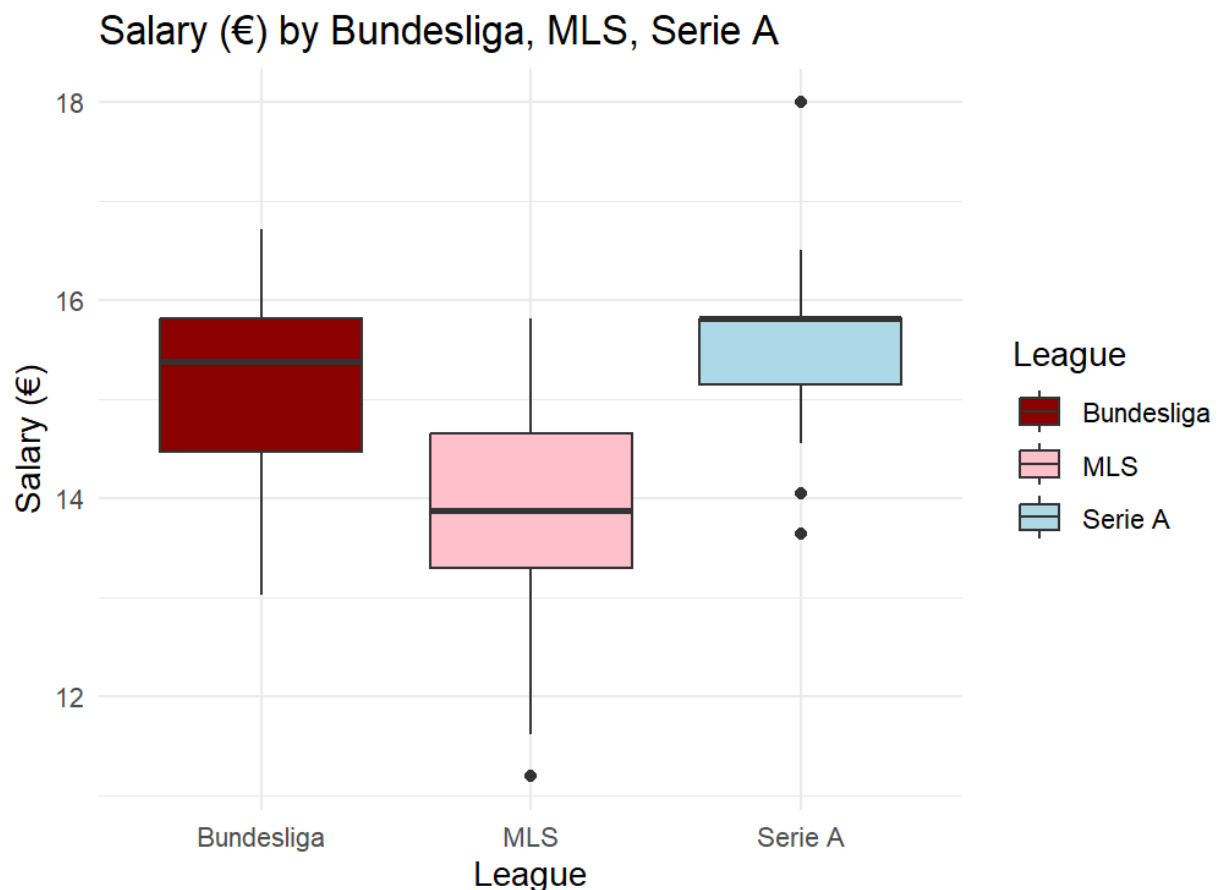
Proportion of High-Performing Players



4. We used simple linear regression analysis to examine the relationship between expected goals per 90 minutes and current market value (€). We applied the exponential transformation on `xGoalsAddedPer90` and then fitted the model. Both the intercept ($p\text{-value}=3.84e-10$) and $\exp(\text{xGoalsAddedPer90})$ ($p\text{-value}< 2e-16$) were statistically significant since their $p\text{-values}< \alpha=0.05$, indicating that they were meaningful predictors of the current market value. The plot showed a statistically significant positive relationship between the expected goals per 90 minutes and football players' current market value (€).



5. We examined whether there is a significant difference in the average Salary (\$) of players across different leagues (Bundesliga, MLS, Serie A) using one-way ANOVA analysis. We used log-transformed salaries; we ensured that the assumptions of statistical tests were better met. According to our results, we found that there is a statistically significant difference in the average Salary (\$) of players across selected leagues ($F = 47.61$, $p < 2e-16$).



4. Discussion/Conclusion

This study revealed certain relationships through analyses performed on the football player dataset, focusing on player performance metrics, characteristics, and financial aspects. Several important insights are presented by our results. First, our findings showed that the expected goals added every ninety minutes are significantly affected by age. This result is consistent with the theory that younger players could be more physically agile and resilient, which could result in higher contributions on the field. Additionally, we found significant differences in the average salaries of players across different leagues (Bundesliga, MLS, Serie A). This finding brings to light the differences in income and economic capacities throughout football leagues, giving clubs and analysts a better knowledge of the financial environment and compensation scales across different divisions. Furthermore, the regression analysis showed that the expected goals added per 90 minutes is a significant predictor of the current market value of players. This information is essential for clubs to maximize player acquisition and valuation plans and make sure player investments are in line with performance indicators. Additionally, compared to defenders, forwards are far more likely to be great performers when it comes to adding goals. This finding supports conventional football tactics, which prioritize defenders' goal-preventing efforts, while forwards are expected to contribute more directly to scoring. Future research could expand the analysis by including more player attributes such as physical

fitness metrics, injury history, and psychological factors, providing a more holistic view of the factors influencing player performance and market value.

In conclusion, this research clarifies the intricate relationship that exists in professional football between player qualities, performance indicators, and financial valuations. Although our results offer a strong foundation, additional investigation and validation are necessary to improve these understandings and their suitability in the dynamic world of professional football.

References

Webster, E. (2024). *Football Analytics*. GitHub.
https://github.com/edwebster/football_analytics/blob/master/data/davies/engineered/davies_all_latest.csv