STONY BROOK UNIVERSITY

AMS 572 PROJECT

QUANTITATIVE ANALYSIS OF SOCCER PLAYER WAGES

By: Group 20 (Aditya Rana, Ananya Sadana, Avdhoot Patil, Krishna Bhuma)

ABSTRACT

This project uses R for data analysis to examine key factors affecting soccer player wages. It tests two hypotheses: one, whether players with long-term contracts (till 2026 or later) have different average wages compared to those with shorter contracts; and two, the influence of age, international reputation, and contract length on wages. Statistical methods like T-Tests, U-Tests and Generalized Linear Models are applied to a detailed dataset, offering insights into the complex factors determining professional football players' compensation.

INTRODUCTION

This project analyzes factors affecting soccer player wages, emphasizing contract length and player attributes. We first examine whether long-term contracts lead to different wages compared to shorter ones using t-tests. Then, we explore the combined effect of player age, international reputation, and contract length on wages through a Generalized Linear Model (GLM). Our goal is to reveal key influences on player compensation using robust statistical methods, offering insights for strategic decision-making in the football industry

BACKGROUND INFORMATION

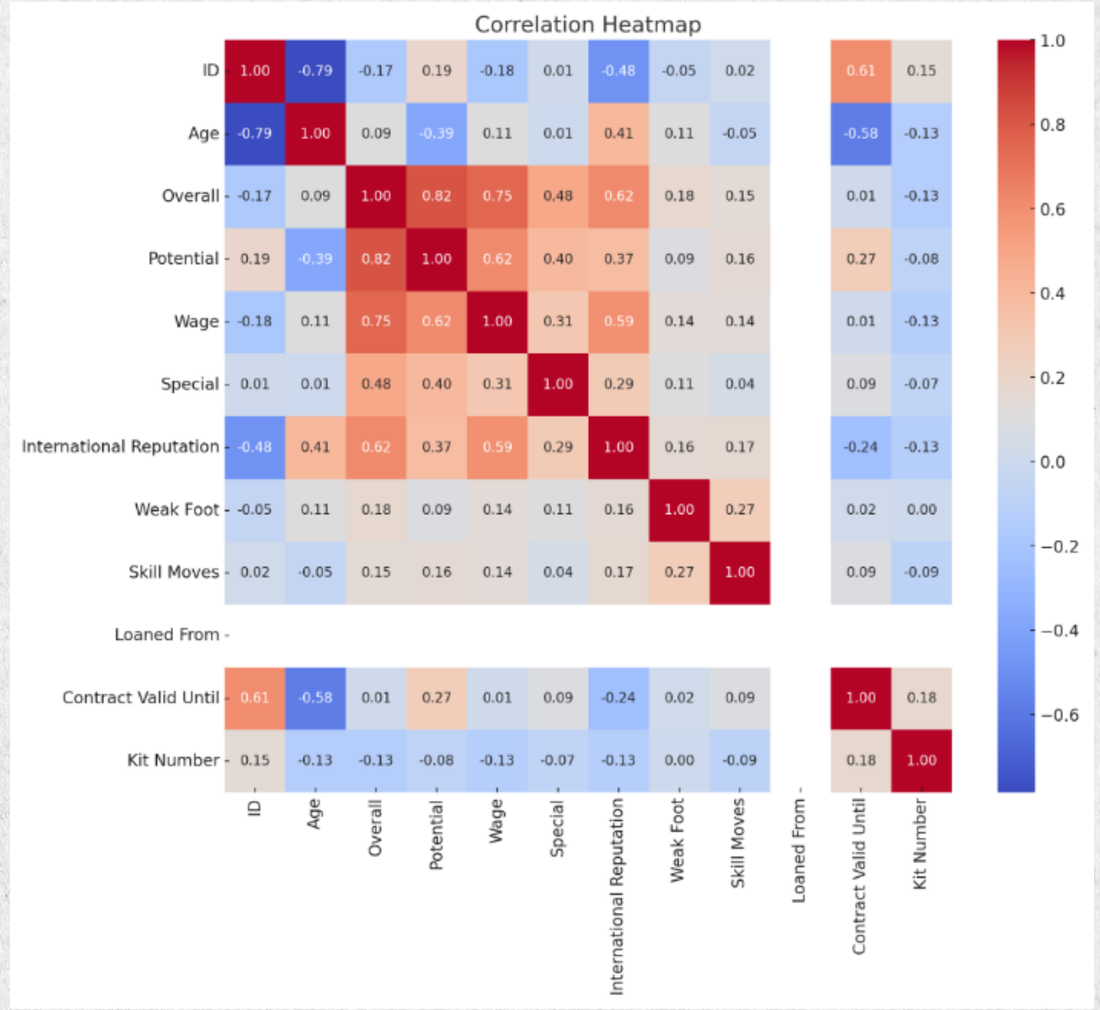
Our dataset, derived from the FIFA video game series, offers detailed insights into the world of professional soccer. It includes key information such as players' names, nationalities, physical attributes (age, height, weight), and professional details like club affiliations and positions. Significantly, it provides players' performance metrics, including their current skills and potential growth, and financial data like wages. Additionally, it covers contract lengths and other specifics, serving as a valuable resource for analyzing the diverse factors influencing a soccer player's professional journey.

SOME DATASET OVERVIEW PLOTS

This histogram shows the distribution of wages among the players. The plot indicates how the wages are spread and highlights the frequency of different wage ranges.

SOME DATASET OVERVIEW PLOTS

This heatmap can be used to identify potential relationships between various attributes.



HYPOTHESIS 1

Null Hypothesis (H0): The average wage of players with contracts valid until 2026 or later (long contracts) is equal to the average wage of players with contracts ending before 2026 (short contracts).

Alternate Hypothesis (H1): The average wage of players with long contracts is not equal to the average wage of players with short contracts.

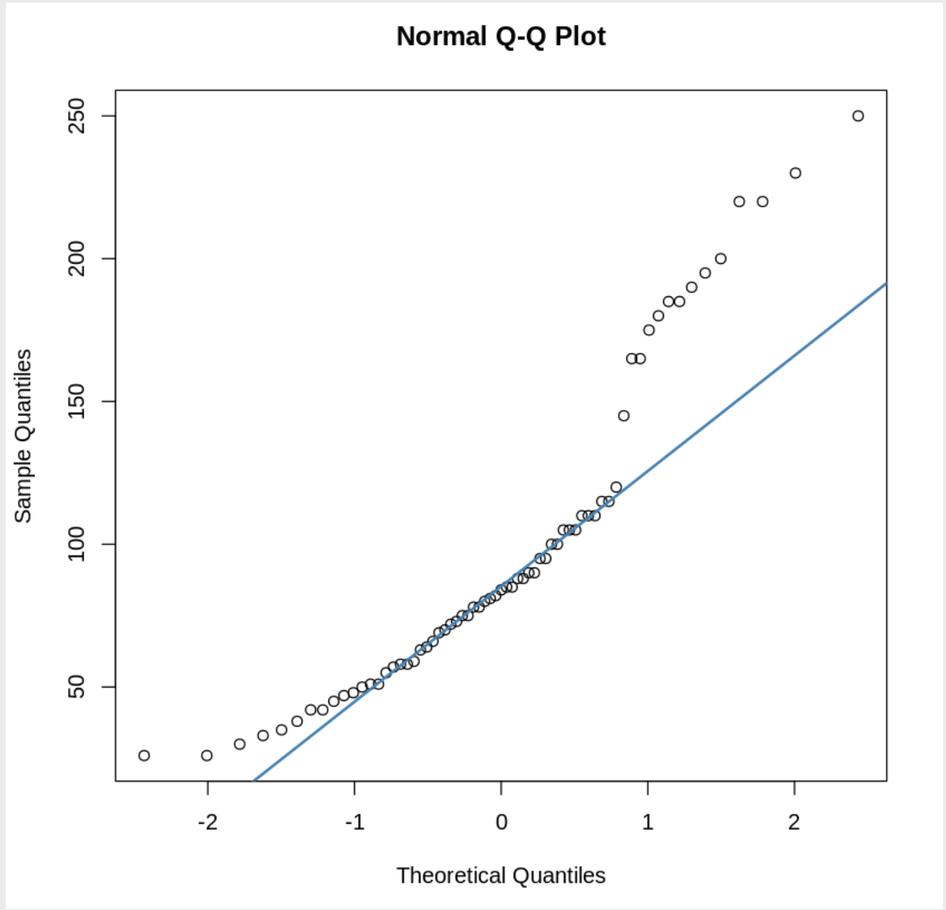
In our analysis of Hypothesis 1, we used a combination of statistical methods: Shapiro-Wilk test for normality, logarithmic transformations to stabilize variances, Welch's Two Sample T-Test for mean comparison on log-transformed data, and the Mann-Whitney U Test for median wage comparisons between long-term and short-term contract groups.

NORMALITY TEST

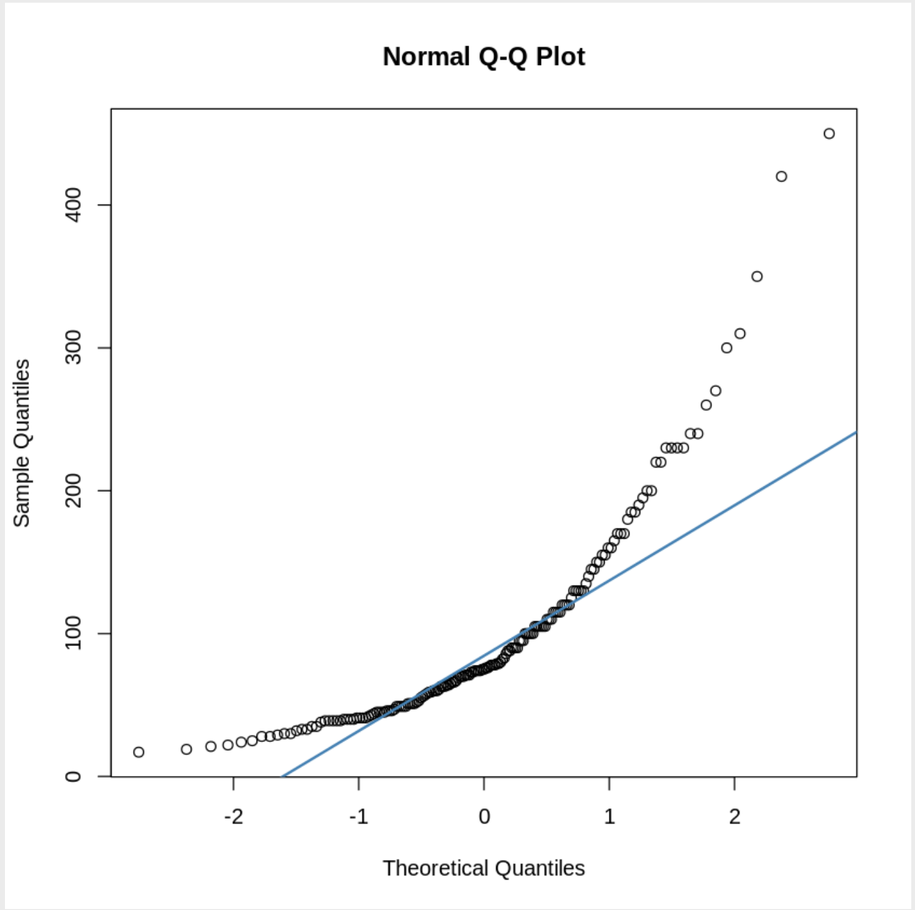
In the initial Shapiro-Wilk Test, the untransformed wage data showed non-normality, indicated by low p-values (1.93e-05 for long contracts, 1.375e-13 for short contracts). After logarithmic transformation, the retest yielded a Shapiro-Wilk statistic (W) of 0.97835 with a p-value of 0.2926 for long contracts, and a W of 0.99218 with a p-value of 0.4832 for short contracts, confirming a return to normal distribution and validating the use of parametric tests like the T-Test.

In our QQ plot analysis, the initial plots indicated skewness and outliers in wage data for both long and short contracts. After logarithmic transformations, the re-assessed QQ plots showed that the data for both groups aligned closely with normal distribution, indicating effective normalization and mitigation of skewness.

QQ PLOTS FOR UNTRANSFORMED DATA

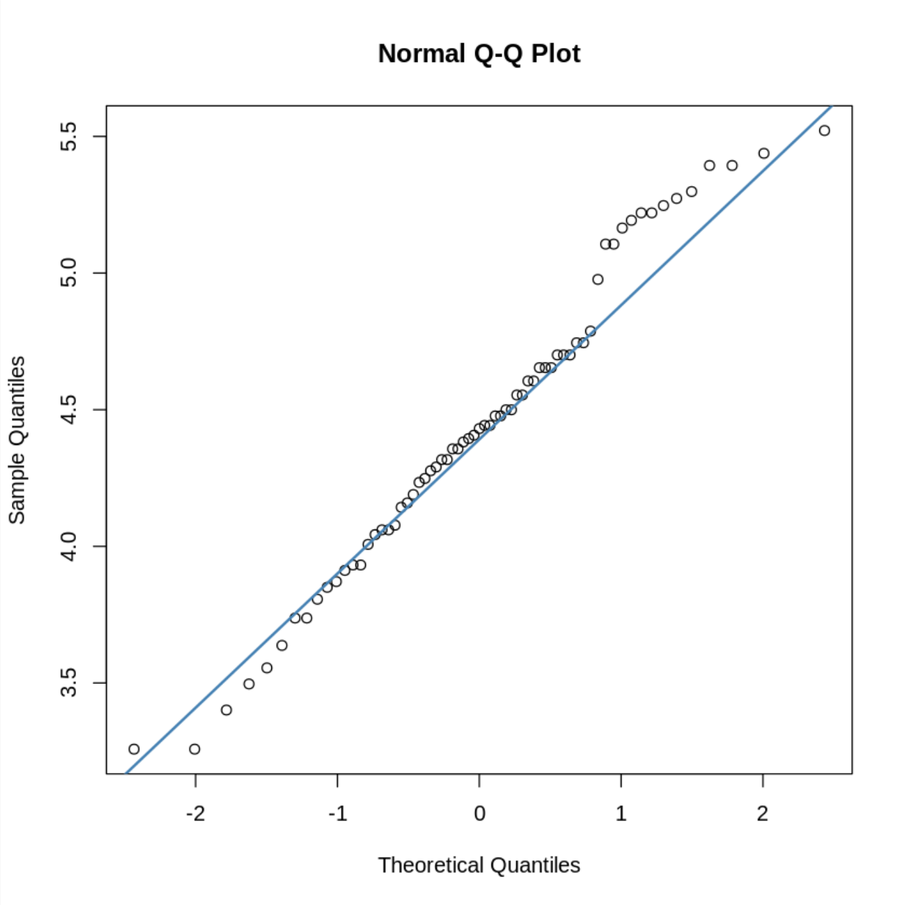


Long contract wages

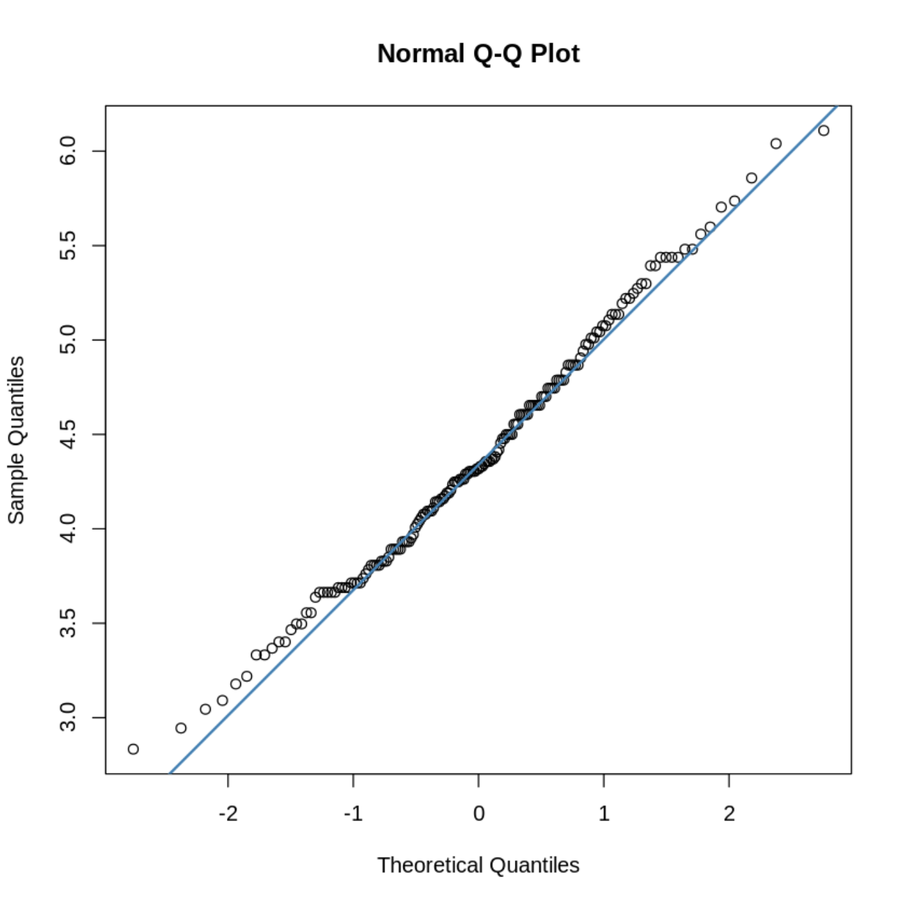


Short contract wages

QQ PLOTS FOR LOG-TRANSFORMED DATA



Long contract wages



Short contract wages

F-TEST, T-TEST AND U-TEST

Variance Test (F-test): Used to assess the homogeneity of variances between wages of players with long-term and short-term contracts. The test showed a significant difference in variances, with an F-value of 0.54928 and a p-value of 0.003619, indicating unequal variances between the two groups

Welch's Two Sample T-Test: Conducted on the log-transformed wage data to investigate average wage differences between long-term and short-term contracts. The transformation addressed skewness, allowing for a parametric hypothesis test. The test produced a t-value of 0.69276 with 140 degrees of freedom

Mann-Whitney U-Test: A non-parametric test used to compare the medians of wages between the two contract groups, compensating for the non-normal characteristics of the original data. The test's outcomes showed a statistic (W) of 6125.5 with a p-value of 0.4064, indicating no significant difference in median wages between the groups

CONCLUSION

Both the Welch's T-Test and the non-parametric Mann-Whitney U-Test yielded p-values well above the conventional threshold for significance, suggesting the differences in mean and median wages, respectively, could be attributed to random variation rather than a true effect of contract length. Thus, we don't have enough evidence to reject the null hypothesis and the conclusion drawn is that contract length does not have a statistically significant impact on players' wages.

HYPOTHESIS 2

Null Hypothesis (H0): The player's age, overall rating, contract length, Potential and values have no effect on their wage.

Alternate Hypothesis (H1): At least one of the player's age, overall rating, contract length and Potential and values has a significant effect on their wage.

For Hypothesis 2, we utilized Generalized Linear Models (GLMs) to examine the effect of players' age, overall rating, and contract length on their wages. GLMs are versatile regression models suitable for exploring both linear and non-linear relationships between variables, allowing for a more flexible analysis compared to traditional linear regression. This approach enabled us to investigate the complex interplay between these key factors and their impact on player wages.

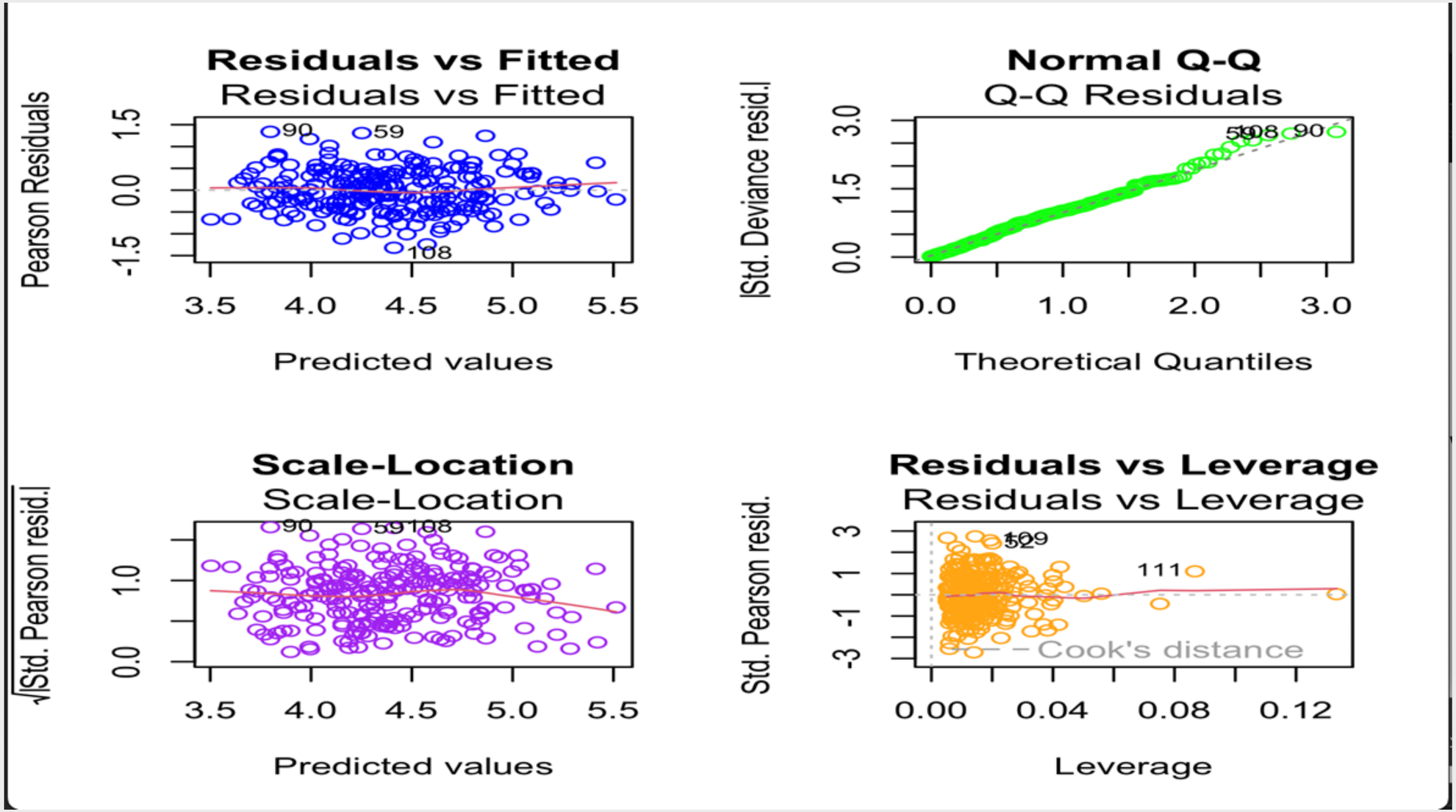
GLM ANALYSIS

In our GLM analysis for Hypothesis 2, we found:

- Age negatively impacts wages (coefficient: -0.03348, $p = 0.00249$).
- International reputation has a positive effect (coefficient: 0.45373, $p < 2e\text{-}16$).
- Contract length does not significantly affect wages (coefficient: 0.03009, $p = 0.20354$).

Model fit was assessed using null and residual deviance, with lower residual deviance indicating a better fit. The Akaike Information Criterion (AIC) was used for model comparison. The results suggest that while age and international reputation significantly influence wages, contract length does not.

VISUALIZING ASSUMPTIONS AND PERFORMANCE: GRAPHS FOR MODEL OUTPUTS



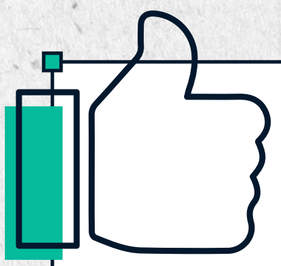
CONCLUSION

In conclusion, this GLM analysis supports the hypothesis that a player's wage is notably influenced by their age and international reputation. The non-significant effect of contract length suggests that, in the context of this model, contract length does not play a substantial role in determining player wages. These findings provide valuable insights for decision-makers in the sports industry, enhancing our understanding of the factors contributing to player wages.

MISSING VALUES ANALYSIS

In our analysis, we applied a Generalized Linear Model (GLM) to a dataset with 20% of player wage data intentionally set as missing. We explored two imputation methods: replacing missing values with the mean and the median wage of the non-missing data. This approach allowed us to evaluate the impact of different imputation strategies on our model's performance.

Additionally, we conducted a targeted analysis on 'old' players, introducing missingness selectively to 20% of this group's wages. For these missing values, we again used mean and median imputations, aiming to assess how specific missing data handling affects the model outcomes and the robustness of the findings.



THANK YOU