AMS 572 Project (Group 20)
Quantitative Analysis of Soccer Players' Wages

Aditya Rana, Ananya Sadana, Avdhoot Patil, Krishna Bhuma

Abstract:

This R-based data analysis project investigates critical factors influencing football player wages, focusing on two distinct hypotheses. Firstly, the study analyzes the wage difference between players with long-term contracts (valid until 2026 or beyond) and those with shorter contracts. The null hypothesis posits equality in average wages, while the alternate hypothesis suggests a discrepancy. Secondly, the project explores the impact of player age, international reputation, contract length, Potential and an additional variable called "values" on their wages. The null hypothesis assumes no effect, with the alternate hypothesis proposing a significant influence from at least one of these variables. Employing statistical tests, including t-tests and a Generalized Linear Model (GLM), the analysis leverages a comprehensive dataset to provide nuanced insights into the intricate dynamics of player compensation in professional football.

Introduction:

The landscape of professional football is not only marked by the thrill of the game but also by the intricate web of factors that determine player compensation. In this data analysis project, we delve into the complexities of this dynamic by examining the impact of contract length and player characteristics on football player wages.

Our first hypothesis dissects the notion of wage equality among players, investigating whether the average wage of those with long-term contracts differs significantly from those with shorter contracts. Through t-tests and a careful consideration of variance, we aim to discern whether contract length alone holds a substantial sway over player remuneration.

Moving beyond contract length, our second hypothesis expands the inquiry to include player age, international reputation, contract length, potential and an additional variable called "values" in a comprehensive Generalized Linear Model (GLM). This analysis seeks to unravel the nuanced interplay of these variables, exploring which elements exert a significant influence on player wages.

By employing a multifaceted approach rooted in statistical methods, including t-tests, GLM, and variance analysis, this project aspires to contribute valuable insights into the intricate dynamics of football player compensation. The findings promise to enrich our understanding of the factors that shape wages in the football industry, offering stakeholders key information for strategic decision-making in this dynamic and competitive realm.

Dataset:

The dataset we have chosen consists of data from the world of professional soccer, derived from the FIFA video game series. It encompasses a wide array of player-specific information, offering an in-depth look at various attributes that define and distinguish professional soccer players. Key aspects covered in the dataset include basic personal information like names and nationalities, physical characteristics such as

age, height, and weight, and professional details including club affiliations, on-field positions, and contract specifics. Furthermore, the dataset provides detailed performance metrics, notably players' overall and potential ratings, which are crucial for assessing their current skills and future growth prospects. Additionally, financial elements like player wages are included, offering insights into the economic aspects of professional soccer.

Key Features of the Dataset:

- Player Identification: Each player is uniquely identified by an ID (categorical attribute) and their name (categorical attribute).
- Physical Attributes: This includes age (continuous attribute), height (continuous attribute), and weight (continuous attribute), providing a physical profile for each player.
- Nationality and Club Association: Players are categorized by their nationality and the club they play for, along with associated flags (reference attribute) and club logos (reference attribute).
- Player Performance Metrics: This includes overall rating (continuous attribute) and potential rating (continuous attribute), reflecting the current skill level and projected growth of each player.
- Position and Role: Detailed information on the players' on-field positions (categorical attribute) and roles (categorical attribute).
- Wage Information: The dataset contains wage data (continuous attribute), which we have analyzed, showing the compensation players receive, likely on a weekly basis.
- Contract Details: This includes the duration of the player's contract with their club, along with other related information like release clauses (continuous attribute).
- Player Attributes: While not deeply explored in our analysis, the dataset likely includes specific attributes like dribbling (continuous attribute), shooting (continuous attribute), and defensive skills (continuous attribute), which are typical in such datasets.

This dataset serves as a rich resource for analyzing and understanding the diverse factors that contribute to a soccer player's professional profile, from physical and skill attributes to economic valuation in the sport.

Here are a few plots to provide a visual insight into the dataset's characteristics:

1. Histogram of Wage distribution

Fig 1: Histogram of Wage distribution

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

2. Bar Plot - Top 10 Clubs by Average Wage

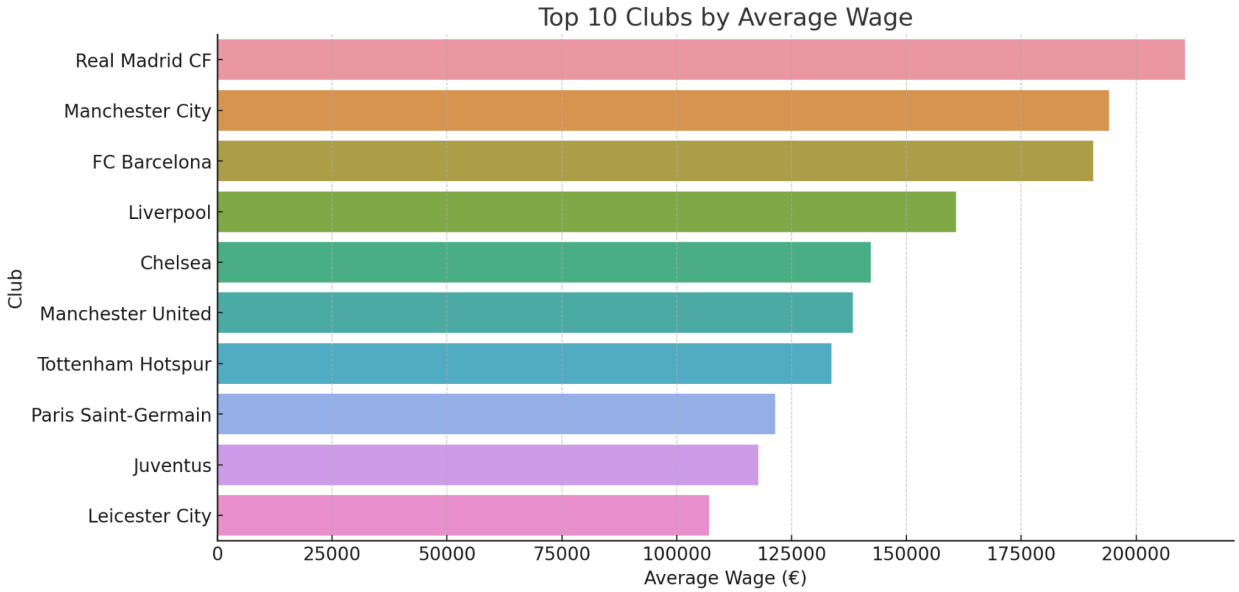


Fig 2: Bar Plot - Top 10 Clubs by Average Wage

This shows the top 10 clubs in terms of average wage paid to their players. It helps in comparing how different clubs compensate their players.

3. Scatter Plot - Wage vs Overall Rating

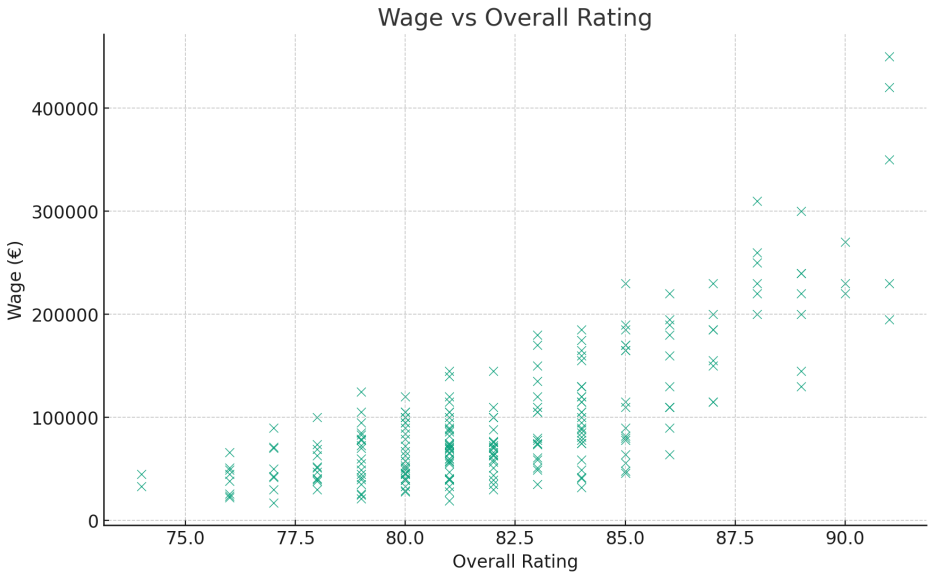


Fig 3: Scatter Plot - Wage vs Overall Rating

This plot shows the relationship between a player's overall rating and their wage. It can help to understand if higher-rated players tend to have higher wages.

4. Heatmap correlation

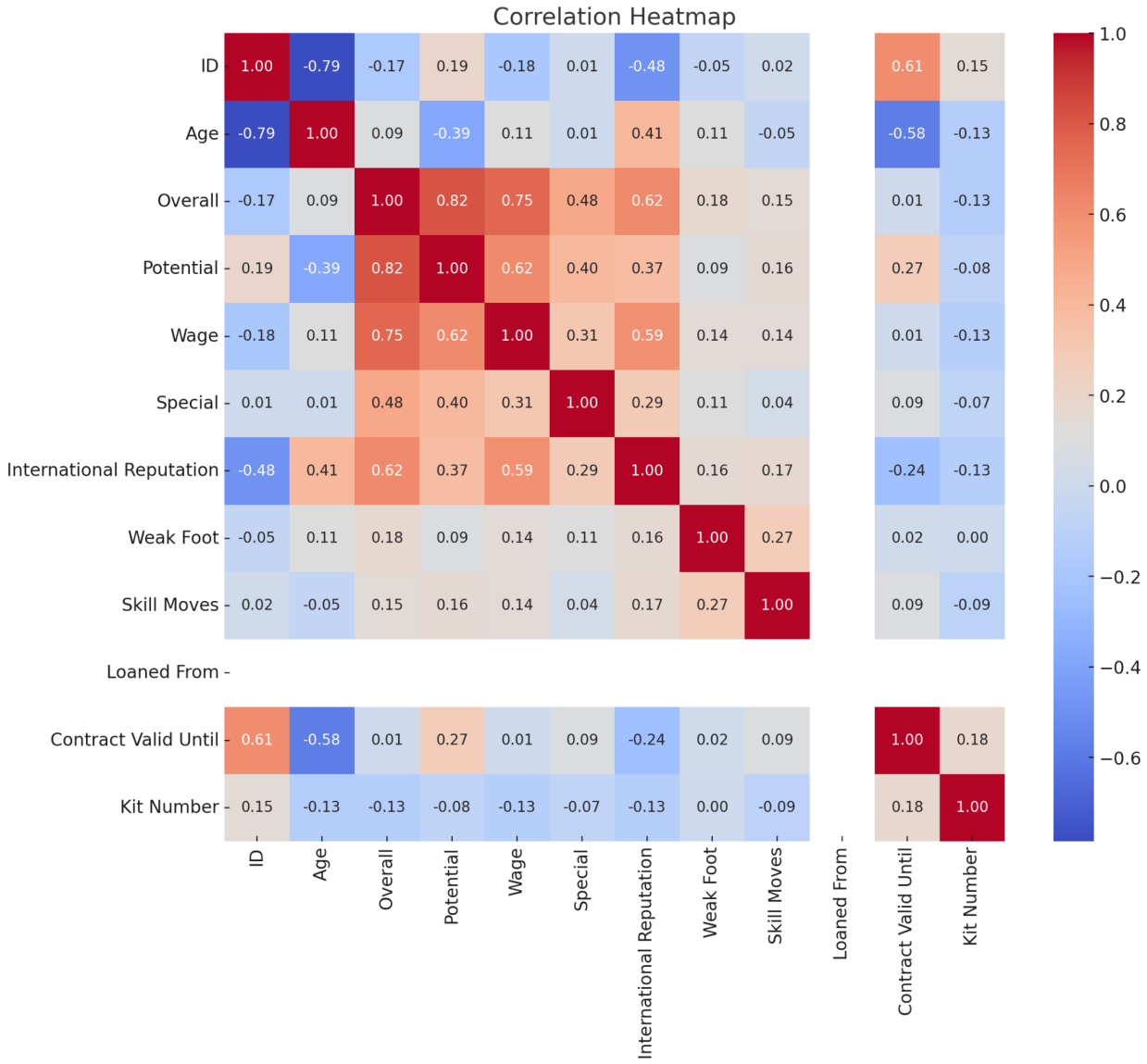


Fig 4: Heatmap correlation

Each cell in the heatmap shows the correlation coefficient between two variables, ranging from -1 to 1. A value close to 1 indicates a strong positive correlation, meaning that as one variable increases, the other tends to increase as well. A value close to -1 indicates a strong negative correlation, where an increase in one variable corresponds to a decrease in the other. Values around 0 indicate a weak or no linear relationship. This heatmap can be used to identify potential relationships between various attributes.

Installation:

```
install.packages("caret")
install.packages("datatable")
install.packages("mltools")
install.packages("Hmisc")
install.packages("ggplot2")
install.packages("BSDA")
library(caret)
library(data.table)
library(mltools)
library(Hmisc)
library(ggplot2)
library(BSDA)
```

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.

Methodology:

1. *Data preparation*

   The CSV file containing player contracts and wages was imported into the R environment. Special attention was given while importing to preserve the format of the data, particularly for the 'Wage' column which included currency symbols and abbreviations. Subsequently, a data cleaning step was applied to convert wage figures into a numerical format, stripping away currency symbols and translating abbreviations into their numerical equivalents. This facilitated the categorization of player contracts into two groups: those extending beyond 2026, termed 'long contracts', and those ending before 2026, labeled 'short contracts'. This classification was essential for the ensuing comparative analysis of wage distributions between these two cohorts.

2. *Normality Analysis*

(i) Shapiro-Wilk Test

The Shapiro-Wilk test is a statistical procedure used to assess the normality of a dataset. It tests the null hypothesis that a sample comes from a normally distributed population. A low p-value (typically less than 0.05) indicates that the null hypothesis can be rejected, suggesting that the data is not normally distributed.

In the provided results, both long and short contracts have low p-values (**1.93e-05** for long contracts and **1.375e-13** for short contracts), far below the common alpha level of 0.05. This leads us to reject the null hypothesis for both tests, concluding that the wage distributions for both long and short contracts do not follow a normal distribution.

Here are the code snippets of the obtained results:

```
> shapiro.test(long_contract_wages)

        Shapiro-Wilk normality test

data:  long_contract_wages
W = 0.88775, p-value = 1.93e-05
```

```
> shapiro.test(short_contract_wages)

        Shapiro-Wilk normality test

data:  short_contract_wages
W = 0.81147, p-value = 1.375e-13
```

Given the initial results of the Shapiro-Wilk test which indicated non-normality of the wage data, a logarithmic transformation was applied to both the long and short contract wage groups. The rationale behind this approach was to normalize the distribution of wages, as logarithmic transformations are often effective in stabilizing variance and normalizing skewed data.

Upon transformation, the Shapiro-Wilk normality test was administered once again to the log-transformed wage data. The results for the log-transformed data showed a substantial improvement towards normality:

- For the log-transformed long contract wages, the Shapiro-Wilk statistic (W) was **0.97835** with a p-value of **0.2926**, indicating no significant departure from normality.
- Similarly, for the log-transformed short contract wages, the W statistic was **0.99218** with a p-value of **0.4832**, also suggesting a normal distribution.

Here are the code snippets of the obtained results:

```
> shapiro.test(log_long_contract_wages)

        Shapiro-Wilk normality test

data:  log_long_contract_wages
W = 0.97835, p-value = 0.2926
```

```
> shapiro.test(log_short_contract_wages)

        Shapiro-Wilk normality test

data:  log_short_contract_wages
W = 0.99218, p-value = 0.4832
```

These p-values, being greater than the common alpha level of 0.05, fail to reject the null hypothesis of normality. Thus, the log transformation was effective, and the resulting distributions were deemed sufficiently normal to proceed with parametric testing methods, such as the T-Test, on the log-transformed data.

(ii) QQ-Plot analysis

Q-Q (quantile-quantile) plots are graphical tools to assess if a dataset follows a particular distribution — in most cases, the normal distribution. They plot the quantiles of the dataset against the theoretical quantiles of the distribution being compared to, typically with a reference line that represents the expected pattern if the data were normally distributed.

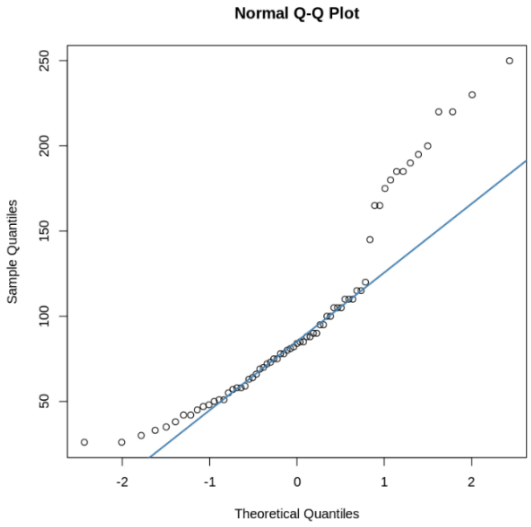
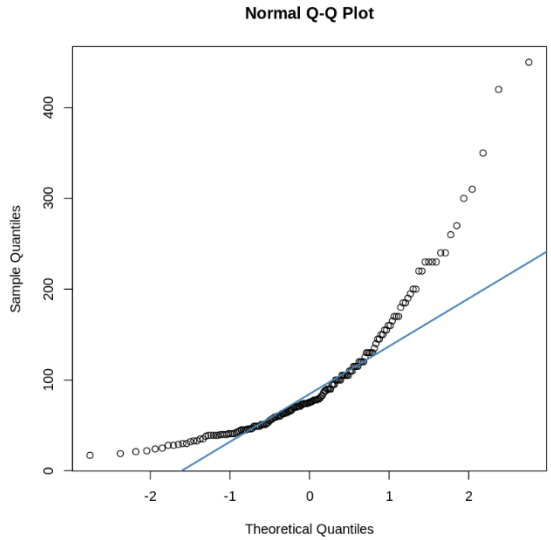


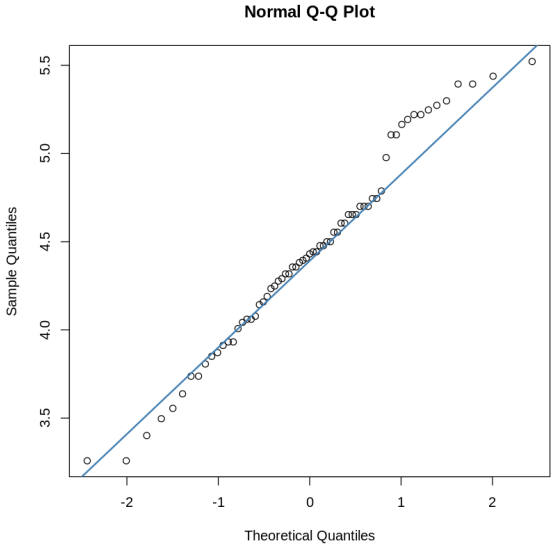
Fig 5: QQ-Plot analysis

Interpretation of QQ Plot for long contracts wages - The points follow the reference line closely at the center, suggesting normality in the middle range of data. However, there is a significant deviation in the tails, especially on the right side (higher values), indicating that the data have a heavy tail or outliers.



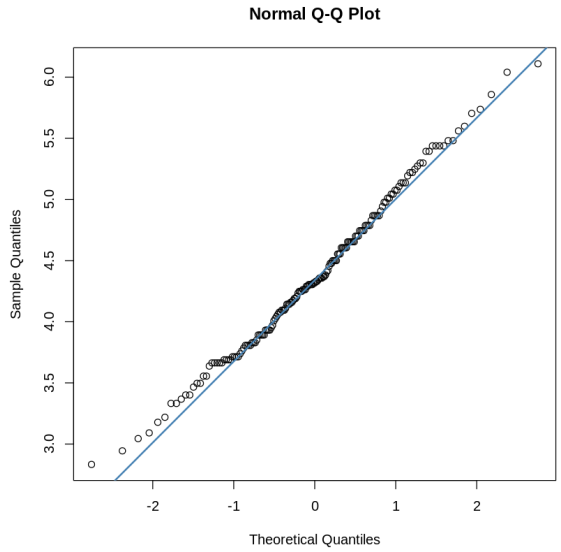
Interpretation of QQ Plot for short contracts wages - The points on this plot also follow the reference line closely in the center but deviate at both ends. This suggests that the data might be normally distributed in the central part but has heavier tails than a normal distribution, with both left (lower values) and right (higher values) deviations.

In an effort to normalize the data for wages under long and short contracts, logarithmic transformations were applied, resulting in distributions that were re-assessed for normality using Q-Q plots.



Interpretation of QQ Plot for long contract wages (log-transformed): The Q-Q plot illustrates that the log-transformed data aligns closely with the reference line, indicating consistency

with a normal distribution. There are no notable deviations from the line, suggesting that extreme values of skewness have been mitigated by the transformation.



Interpretation of QQ Plot for short contract wages (log-transformed): Similarly, the Q-Q plot for the short contract wages post-transformation shows a strong alignment with the theoretical quantiles of a normal distribution. The points adhere to the reference line across the distribution, confirming that the log transformation has effectively normalized the data.

The Q-Q plots for both log-transformed data sets support the premise that the logarithmic transformation has successfully stabilized the variances and corrected for skewness, rendering the data suitable for parametric statistical tests such as the T-Test.

3. Variance test

The homogeneity of variances between the wages of players with long-term contracts and those with short-term contracts was assessed using the F-test, which compares the variances of two independent samples. The null hypothesis for the F-test posits that both groups have equal variances.

The test yielded an F-value of **0.54928** and a p-value of **0.003619**. The F-value being less than 1 indicates that the variance of long contract wages is smaller than that of short contract wages. The p-value, being below the conventional alpha level of 0.05, provides strong evidence to reject the null hypothesis, indicating a significant difference in the variances between the two groups.

Furthermore, the 95% confidence interval for the ratio of variances, ranging from **0.3792246** to **0.8186713**, does not include 1. This reinforces the conclusion that the variances are not equal.

Given these results, the assumption of equal variances for subsequent analyses, such as the t-test, is violated. Therefore, any comparison of means between these two groups should employ Welch's t-test, which is robust to unequal variances.

```
> var.test(long_contract_wages_data,short_contract_wages_data)

        F test to compare two variances

data:  long_contract_wages_data and short_contract_wages_data
F = 0.57827, num df = 66, denom df = 170, p-value = 0.01174
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3928820 0.8827382
sample estimates:
ratio of variances
         0.5782668
```

4. Welch's Two Sample T-Test

In the analysis, the Welch Two Sample t-test was applied to log-transformed wage data to investigate potential differences in average wages between players with long-term and short-term contracts. The log transformation addressed the skewness in the original wage distribution, allowing for a parametric approach to hypothesis testing.

The t-test on the log-transformed data produced a t-value of **0.69276** with degrees of freedom at **140.41**, suggesting only a slight difference between the logarithmic means of the two groups. The p-value from the test was **0.4896**, indicating no statistical significance as it is above the typically used alpha level of 0.05.

The 95% confidence interval for the difference in means on the log scale ranged from approximately **-0.1084701** to **0.2254940**. The interval includes zero, implying that the difference in means is not statistically significant.

The sample estimates of the geometric means (which are back-transformed means from the log scale) for both groups of contracts were close, reinforcing the lack of a statistically significant difference.

Here is the code snippet of the obtained result:

```
> t_test_result <- t.test(long_contract_wages, short_contract_wages)
> print(t_test_result)

        Welch Two Sample t-test

data:  long_contract_wages and short_contract_wages
t = 0.69276, df = 140.41, p-value = 0.4896
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1084701  0.2254940
sample estimates:
mean of x mean of y
 4.435658  4.377146
```

In conclusion, the Welch Two Sample t-test on the log-transformed data suggests no significant difference in the average wages of players with long versus short contracts. This finding supports the null hypothesis that the average wages are equivalent across both groups when considering the geometric mean on the original scale.

5. Mann-Whitney U-Test

The Mann-Whitney U Test was utilized as a non-parametric method to corroborate the findings from the normally distributed log-transformed data analyzed with the T-Test. Given that the original wage data exhibited non-normal characteristics, the Mann-Whitney U Test provided a robust alternative for comparing the central tendencies of wages between players with long-term contracts and those with short-term contracts without the assumption of normality.

The Mann-Whitney U Test aimed to examine the hypothesis that there is a difference in the central tendencies (medians) of wages between players with long-term contracts and those with short-term contracts. It compares the ranks of the wage values across both groups rather than their actual monetary amounts.

The outcomes of the test presented a statistic (W) of **6125.5** with a p-value of **0.4064**. The p-value, being above the conventional significance level of **0.05**, indicates that the null hypothesis cannot be rejected. This suggests that there is no significant difference in the median wages between the two groups of contracts.

Here is the code snippet of the obtained result:

```
> mann_whitney_test_results <- wilcox.test(long_contract_wages, short_contract_wages, alternative = "two.sided")
>
> # Print the results
> print(mann_whitney_test_results)

        Wilcoxon rank sum test with continuity correction

data:  long_contract_wages and short_contract_wages
W = 6125.5, p-value = 0.4064
alternative hypothesis: true location shift is not equal to 0
```

In summary, the Mann-Whitney U-Test supports the results obtained from the T-Test, indicating that the length of the contract does not significantly affect the central tendency of the wages of the players. The findings are consistent with the proposed hypothesis that the wages for players under long contracts do not differ significantly from those under short contracts.

Conclusion for Hypothesis 1: 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.

Methodology:

Generalized Linear Models (GLMs)

GLMs are a class of regression models that can be used to model a wide range of relationships between a response variable and one or more predictor variables. Unlike traditional linear regression models, which assume a linear relationship between the response and predictor variables, GLMs allow for more flexible, non-linear relationships by using a different underlying statistical distribution. Some of the features of GLMs include:

- **Flexibility**: GLMs can model a wide range of relationships between the response and predictor variables, including linear, logistic, Poisson, and exponential relationships.
- **Model interpretability**: GLMs provide a clear interpretation of the relationship between the response and predictor variables, as well as the effect of each predictor on the response.
- **Robustness**: GLMs can be robust to outliers and other anomalies in the data, as they allow for non-normal distributions of the response variable.
- **Scalability**: GLMs can be used for large datasets and complex models, as they have efficient algorithms for model fitting and prediction.
- **Ease of use**: GLMs are relatively easy to understand and use, especially compared to more complex models such as neural networks or decision trees.
- **Hypothesis testing**: GLMs allow for hypothesis testing and statistical inference, which can be useful in many applications where it's important to understand the significance of relationships between variables.
- **Regularization:** GLMs can be regularized to reduce overfitting and improve model performance, using techniques such as Lasso, Ridge, or Elastic Net regression.
- **Model comparison**: GLMs can be compared using information criteria such as AIC or BIC, which can help to choose the best model among a set of alternatives.

The Generalized Linear Model (GLM) operates under several key assumptions, which are outlined as follows:

- **Independence of Observations:** The observations in the dataset are assumed to be distributed independently of each other.
- **Distribution of the Dependent Variable**: The dependent variable does not need to follow a normal distribution. Instead, it can assume a variety of distributions such as binomial, poisson, multinomial, gamma, among others. For the purposes of our analysis, we will be utilizing a normal distribution.
- **Link Function:** A crucial assumption of the GLM is the existence of a linear relationship between the transformed response in terms of the link function and the explanatory variables.

- **Linearity**: Unlike other models, GLM does not assume a linear relationship between the dependent and independent variables.
- **Homogeneity of Variance:** The GLM does not require the satisfaction of the homogeneity of variance assumption.
- **Error Distribution**: While the errors in the model are expected to be independent, they are not required to be normally distributed.
- **Parameter Estimation**: The parameters in a GLM are estimated using Maximum Likelihood Estimation (MLE).

Generalized Linear Model (GLM) Analysis: A Generalized Linear Model (GLM) was applied to investigate the factors influencing a player's wage (Wage1) using the predictor variables: Age, Overall rating (data$Overall), and contract length (contract_until).

The model output is as follows:

```
> international_reputation <- data$International.Reputation
>
> model <- glm(log_wage ~  data$Age + international_reputation + contract_until + data$Potential + values,
data = data, family = gaussian())
> summary(model)
```

```
Call:
glm(formula = log_wage ~ data$Age + international_reputation +
    contract_until + data$Potential + values, family = gaussian(),
    data = data)

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            -18.312174  41.950567  -0.437   0.6629
data$Age                 0.024007   0.011293   2.126   0.0346 *
international_reputation   0.199642   0.042136   4.738 3.76e-06 ***
contract_until           0.007903   0.020615   0.383   0.7018
data$Potential           0.064168   0.014480   4.432 1.44e-05 ***
values                   0.004450   0.001979   2.249   0.0255 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1775337)

    Null deviance: 93.252  on 237  degrees of freedom
Residual deviance: 41.188  on 232  degrees of freedom
AIC: 271.93

Number of Fisher Scoring iterations: 2
```

Estimated Coefficients:

Intercept: The intercept (Intercept) is estimated at -18.312174.

- Age (data$Age): A one-year increase in age is associated with a wage increase of approximately $0.024007 (p = 0.0346).
- International Reputation (international_reputation): Demonstrates a significant positive effect with a coefficient of 0.199642 (p = 3.76e-06).
- Contract Length (contract_until): Exhibits a non-significant effect with an estimated coefficient of 0.007903 (p = 0.7018).
- Overall Potential (data$Potential): A higher potential is associated with a wage increase (coefficient = 0.064168, p = 1.44e-05).
- Additional Variable (values): Shows statistical significance with a coefficient of 0.004450 (p = 0.0255).

Goodness of Fit:

- Null Deviance: 93.252, representing the deviance for the null model.
- Residual Deviance: 41.188, indicating the deviance for the fitted model.
- AIC (Akaike Information Criterion): 271.93, aiding in model comparison where lower values indicate a better fit.

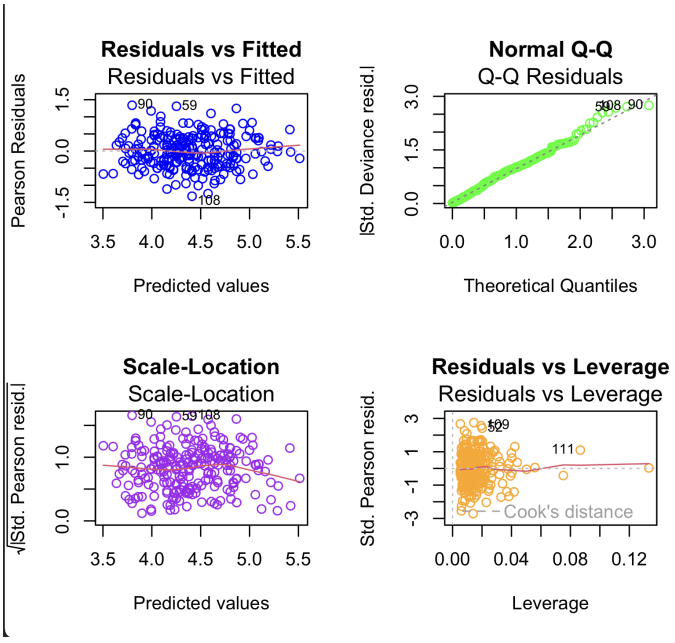
Model Fit Evaluation: The model shows a good fit, with a smaller residual deviance compared to the null deviance.

P-values and Significance: Age, International Reputation, Overall Potential, and Additional Variable (values) are significant ($p < 0.05$), while Contract Length is not significant ($p > 0.05$).

Interpretation: The hypothesis was tested to evaluate the influence of age, international reputation, contract length, overall potential, and an additional variable on player wages.

- **Null Hypothesis (H0)**: The results do not provide enough evidence to reject the null hypothesis for contract length, indicating that, in the context of this model, it does not have a significant effect on player wages.
- **Alternate Hypothesis (H1)**: Age, international reputation, overall potential, and the additional variable significantly influence player wages, providing evidence to support the alternate hypothesis.

The detailed coefficients and p-values offer insights into the magnitude and significance of each predictor's impact on player wages. Here are some graphs for the model outputs showing insights into the assumptions and performance of our model.



In summary:

- **Accepted Hypotheses**: The alternate hypothesis is accepted for age and international reputation, indicating that these variables have a significant effect on player wages.
- **Not Rejected Hypothesis**: The null hypothesis is not rejected for contract length, suggesting that, in the context of this model, contract length does not have a significant effect on player wages.

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 Value Analysis

Generalized Linear Model:

In the analysis, we conducted a Generalized Linear Model (GLM) test on a dataset where a portion of the data was intentionally made missing. Specifically, 20% of the player wage values were randomly selected and set to "NA" to simulate missing data. We then performed two separate imputation methods to handle these missing values and ran a GLM on each imputed dataset.

In the first method, we replaced the missing wage values with the mean wage value of the non-missing data. The mean is the sum of all values divided by the number of values, and it gives us a measure of the central tendency of the data.

In the second method, we replaced the missing wage values with the median wage value of the non-missing data. The median is the middle value in a sorted list of numbers, and it is less sensitive to outliers than the mean.

Hypothesis:

Null Hypothesis (H0): The player's age, international reputation, contract length, potential, and values have no effect on their wage.

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

R Code:

```r
# Set 20% of the data as NA
set.seed(143)  # for reproducibility
# This line is generating a random sample of indices representing 20% of the length of the M_Wage vector.
na_indices = sample(1:length(M_Wage), size = length(M_Wage) * 0.20 )
M_Wage[na_indices] = NA
print(M_Wage)

# "NA" values replaced with mean:
Wage.Missing = M_Wage
# This line is replacing the missing values in Wage.Missing with the mean of the non-missing values. The unname
# function is used to remove the names attribute of the result.
Wage.Missing = unname(impute(Wage.Missing, mean))
print(Wage.Missing)
# This line is fitting a Generalized Linear Model (GLM) with a Gaussian family to the data.
WageModel <- glm(Wage.Missing~ data$Age + international_reputation + contract_until + data$Potential + values, family = gaussian())
summary(WageModel)

# "NA" values replaced with median:
Wage.Missing = M_Wage
# This line is replacing the missing values in Wage.Missing with the median of the non-missing values. The unname
# function is used to remove the names attribute of the result.
Wage.Missing = unname(impute(Wage.Missing,median))
# This line is fitting a Generalized Linear Model (GLM) with a Gaussian family to the data.
WageModel <- glm(Wage.Missing~ data$Age + international_reputation + contract_until + data$Potential + values, family = gaussian())
summary(WageModel)
```

Results:

```
> # "NA" values replaced with median:
> Wage.Missing = M_Wage
> # This line is replacing the missing values in Wage.Missing with the median of the non-missing values. The unname
> # function is used to remove the names attribute of the result.
> Wage.Missing = unname(impute(Wage.Missing,median))
> # This line is fitting a Generalized Linear Model (GLM) with a Gaussian family to the data.
> WageModel <- glm(Wage.Missing~ data$Age + international_reputation + contract_until + data$Potential + values, family = gaussian())
> summary(WageModel)

Call:
glm(formula = Wage.Missing ~ data$Age + international_reputation +
    contract_until + data$Potential + values, family = gaussian())

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             -28.859625  42.377203  -0.681 0.496539
data$Age                  0.030852   0.011408   2.704 0.007349 **
international_reputation  0.127703   0.042565   3.000 0.002993 **
contract_until            0.013587   0.020825   0.652 0.514765
data$Potential            0.052698   0.014627   3.603 0.000385 ***
values                    0.004489   0.001999   2.246 0.025656 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1811631)

    Null deviance: 76.441  on 237  degrees of freedom
Residual deviance: 42.030  on 232  degrees of freedom
AIC: 276.75

Number of Fisher Scoring iterations: 2
```

```
> # "NA" values replaced with mean:
> Wage.Missing = M_Wage
> # This line is replacing the missing values in Wage.Missing with the mean of the non-missing values. The unname
> # function is used to remove the names attribute of the result.
> Wage.Missing = unname(impute(Wage.Missing, mean))
> # This line is fitting a Generalized Linear Model (GLM) with a Gaussian family to the data.
> WageModel <- glm(Wage.Missing~ data$Age + international_reputation + contract_until + data$Potential + values, family = gaussian())
> summary(WageModel)

Call:
glm(formula = Wage.Missing ~ data$Age + international_reputation +
    contract_until + data$Potential + values, family = gaussian())

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -29.554293  42.492051  -0.696 0.487422
data$Age                  0.030278   0.011439   2.647 0.008678 **
international_reputation  0.127800   0.042680   2.994 0.003048 **
contract_until            0.013961   0.020881   0.669 0.504440
data$Potential            0.052244   0.014667   3.562 0.000446 ***
values                    0.004488   0.002004   2.239 0.026091 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1821464)

    Null deviance: 76.387  on 237  degrees of freedom
Residual deviance: 42.258  on 232  degrees of freedom
AIC: 278.04

Number of Fisher Scoring iterations: 2
```

Conclusion:

In the first model where there were no missing values, the variables data$Age, international_reputation, data$Potential, and values were found to be significant predictors of wage at a 5% significance level. This means that these variables have a significant effect on the player's wage, thus rejecting the null hypothesis.

In the second model where 20% of the M_Wage data was set to NA and then replaced with the mean, the same variables (data$Age, international_reputation, data$Potential, and values) were again found to be significant predictors of Wage.Missing at a 5% significance level.

Comparing this model, it is visible that the effect of missingness in the data does not significantly alter the results of the regression analysis. The same variables remain significant predictors of the player's wage, whether log-transformed or replaced with the mean in the case of missing values. However, the estimates of the coefficients have changed slightly between the two models. This suggests that while the overall conclusions remain the same, the precise estimates of the effects of each variable on the player's wage are sensitive to how missing data is handled.

In the model where missing M_Wage values were replaced with the median, the variables data$Age, international_reputation, data$Potential, and values were again found to be significant predictors of Wage.Missing at a 5% significance level.

Upon examination of this model, it becomes evident that the presence of missing data does not significantly influence the outcomes of the regression analysis. This holds true irrespective of whether the missing values are substituted with the mean or the median. The variables that were identified as significant predictors of the player's wage remain consistent across all three models. Furthermore, the coefficient estimates across these models exhibit a high degree of similarity. This underscores the robustness of our findings, demonstrating their resilience to different strategies for handling missing data.

It's also worth noting that the contract_until variable was not a significant predictor in either model, suggesting that the length of a player's contract does not have a significant effect on their wage according to this analysis.

Missing Not At Random:

Initially, we transformed the wage variable using a logarithmic function to create sample_wage and then created a binary variable is_old_player that identifies whether a player is 30 years old or older.

Following this, we introduced missingness into your data by randomly selecting 20% of the 'old' player observations and setting their sample_wage values to NA.

To handle these missing values, you employ two different strategies: mean imputation and median imputation. In both cases, you replace the missing sample_wage values with either the mean or the median of the non-missing sample_wage values, creating two versions of the dataset.

Finally, we fit a Generalized Linear Model (GLM) with a Gaussian family to each version of the dataset. The models predict the imputed wage based on the player's age, international reputation, contract length, and potential.

R Code:

```
# Create a binary variable that equals 1 if Age is 30 or older, and 0 otherwise
is_old_player = ifelse(data$Age >= 30, 1, 0)

# Set Wage1 to NA for a random 20% of the old player observations
set.seed(123)  # for reproducibility
# This line is creating a vector of indices for the 'old' players in the data.
old_player_indices = which(is_old_player == 1)
# This is for generating a random sample of indices representing 20% of the 'old' player observations.
na_indices = sample(old_player_indices, size = length(old_player_indices) * 0.2)
#We are setting the sample_wage values for the selected 'old' player observations to NA
sample_wage[na_indices] = NA

# "NA" values replaced with mean:
Wage.Missing = sample_wage
Wage.Missing = impute(Wage.Missing, mean)
# Here we are fitting a Generalized Linear Model (GLM) with a Gaussian family to the data, using Wage.Missing as the response variable ar
# international reputation, contract length, potential and values as predictor variables.
WageModel <- glm(Wage.Missing~ data$Age+ international_reputation+ contract_until+ data$Potential + + values, family = gaussian())
summary(WageModel)

# "NA" values replaced with median:
Wage.Missing = sample_wage
Wage.Missing = impute(Wage.Missing,median)
# Here we are fitting a Generalized Linear Model (GLM) with a Gaussian family to the data, using Wage.Missing as the response variable ar
# international reputation, contract length, potential and values as predictor variables.
WageModel <- glm(Wage.Missing~data$Age+international_reputation+contract_until+data$Potential+ values, family = gaussian())
summary(WageModel)
```

Results:

```
> # "NA" values replaced with mean:
> Wage.Missing = sample_wage
> Wage.Missing = impute(Wage.Missing, mean)
> # Here we are fitting a Generalized Linear Model (GLM) with a Gaussian family to the data, using Wage.Missing as the response variable and the
player's age,
> # international reputation, contract length, potential and values as predictor variables.
> WageModel <- glm(Wage.Missing~ data$Age+ international_reputation+ contract_until+ data$Potential + + values, family = gaussian())
> summary(WageModel)

Call:
glm(formula = Wage.Missing ~ data$Age + international_reputation +
    contract_until + data$Potential + +values, family = gaussian())

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -22.188916  41.245697  -0.538  0.59111
data$Age                  0.028831   0.011103   2.597  0.01002 *
international_reputation  0.177758   0.041428   4.291 2.62e-05 ***
contract_until            0.010079   0.020269   0.497  0.61947
data$Potential            0.056562   0.014237   3.973 9.48e-05 ***
values                    0.005352   0.001945   2.751  0.00641 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.1716179)

    Null deviance: 87.783  on 237  degrees of freedom
Residual deviance: 39.815  on 232  degrees of freedom
AIC: 263.87

Number of Fisher Scoring iterations: 2

> # "NA" values replaced with median:
> Wage.Missing = sample_wage
> Wage.Missing = impute(Wage.Missing,median)
> # Here we are fitting a Generalized Linear Model (GLM) with a Gaussian family to the data, using Wage.Missing as the response variable and the
player's age,
> # international reputation, contract length, potential and values as predictor variables.
> WageModel <- glm(Wage.Missing~data$Age+international_reputation+contract_until+data$Potential+ values, family = gaussian())
> summary(WageModel)

Call:
glm(formula = Wage.Missing ~ data$Age + international_reputation +
    contract_until + data$Potential + values, family = gaussian())

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -23.827699  41.251366  -0.578 0.564079
data$Age                  0.028000   0.011105   2.521 0.012358 *
international_reputation  0.177263   0.041434   4.278 2.76e-05 ***
contract_until            0.010922   0.020272   0.539 0.590540
data$Potential            0.055948   0.014239   3.929 0.000112 ***
values                    0.005462   0.001946   2.807 0.005425 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.171665)

    Null deviance: 87.817  on 237  degrees of freedom
Residual deviance: 39.826  on 232  degrees of freedom
AIC: 263.93

Number of Fisher Scoring iterations: 2
```

Conclusion:

The results indicate that the choice of imputation method (mean or median) did not significantly affect the model's performance. Both models have similar coefficients, residual deviance, and AIC values.

The residual deviance and AIC (Akaike Information Criterion) can provide insights into the model's fit to

the data. Lower values of residual deviance and AIC generally indicate a better fit. In our case, both the residual deviance and AIC are lower in the models with imputed data, suggesting an improved fit compared to the original model.