Introduction to FIFA Player Analysis Project

Dataset Overview

The FIFA dataset used in this project contains comprehensive information about professional football players, including their personal attributes, performance statistics, and market valuations. This rich dataset provides insights into various aspects of professional football, from player demographics to economic factors that influence player valuation in the transfer market.

The dataset includes variables such as player ID, name, age, nationality, overall rating, potential, club affiliation, market value, wage, preferred foot, international reputation, skill moves, position, contract details, physical attributes (height and weight), and release clause values. With over 18,000 players represented, this dataset offers a robust foundation for econometric analysis of the football player market.

Project Objectives

This econometric project aims to explore, analyze, and model the relationships between various player attributes and their market values in professional football. Specifically, we seek to answer four key research questions:

1. How does a player's Overall rating affect their Wage after controlling for Age and International Reputation?
2. Does Age have a nonlinear effect on Wage_log, independent of Skill Moves?
3. Is Skill Moves a significant predictor of Release Clause_log after accounting for Overall?
4. Can we drop Release Clause without losing information if Value is already in the model?

Through this analysis, we hope to gain insights into the economic dynamics of the football transfer market and understand how different attributes contribute to a player's perceived value. This knowledge can be valuable for football clubs, agents, and analysts involved in player recruitment, contract negotiations, and financial planning.

Methodology Overview

Our approach follows a structured econometric methodology:

1. **Data Exploration**: We begin with a thorough exploration of the dataset, generating descriptive statistics, histograms for numerical variables, and bar charts for categorical data. Each variable is summarized to understand its distribution and relevance.

2. **Data Cleaning**: We identify and handle outliers using the Interquartile Range (IQR) method, applying log transformations where necessary for variables resistant to standard outlier detection. Missing values are filled using appropriate methods (mean for numerical variables, mode for categorical variables), and we check for duplicates.

3. **Modeling**: Based on our four specific research questions, we develop and estimate appropriate econometric models. Each model is designed to answer particular aspects of player valuation and market dynamics.

4. **Interpretation**: We interpret the results of each model, drawing conclusions about the factors that influence player values and the relationships between different variables.

This comprehensive approach allows us to derive meaningful insights from the FIFA dataset while adhering to sound econometric principles and practices.

Data Exploration

Descriptive Statistics

Our first step in analyzing the FIFA dataset was to generate descriptive statistics to understand the central tendencies, dispersion, and distribution of the numerical variables. The table below presents a summary of these statistics:

	ID	Age	Overall	Potential	Value	Wage	International Reputation	Skill Moves	Joined	Height	Weight	Release Clause
count	18207.000000	18207.000000	18207.000000	18207.000000	17955.000000	18207.000000	18159.000000	18159.000000	18207.000000	18207.000000	18207.000000	18207.000000
mean	214298.338606	25.122206	66.238699	71.307299	2444.530214	9.731312	1.113222	2.361308	2016.420607	5.946771	165.979129	4585.060971
std	29965.244204	4.669943	6.908930	6.136496	5626.715434	21.999290	0.394031	0.756164	2.018194	0.220514	15.572775	10630.414430
min	16.000000	16.000000	46.000000	48.000000	10.000000	0.000000	1.000000	1.000000	1991.000000	5.083333	110.000000	13.000000
25%	200315.500000	21.000000	62.000000	67.000000	325.000000	1.000000	1.000000	2.000000	2016.000000	5.750000	154.000000	570.000000
50%	221759.000000	25.000000	66.000000	71.000000	700.000000	3.000000	1.000000	2.000000	2017.000000	5.916667	165.000000	1300.000000
75%	236529.500000	28.000000	71.000000	75.000000	2100.000000	9.000000	1.000000	3.000000	2018.000000	6.083333	176.000000	4585.060806
max	246620.000000	45.000000	94.000000	95.000000	118500.000000	565.000000	5.000000	5.000000	2018.000000	6.750000	243.000000	228100.000000

These statistics reveal several interesting patterns:

- The dataset contains information on 18,207 players, with some variables having a small number of missing values.
- Player ages range from 16 to 45 years, with a mean age of approximately 25 years, reflecting the typical career span of professional footballers.
- Overall ratings range from 46 to 94, with a mean of 66.24, indicating that the dataset includes players of varying skill levels.
- There is significant variation in player values and wages, as indicated by the large standard deviations relative to the means, suggesting a highly skewed distribution with a few extremely high-value players.
- International Reputation ranges from 1 to 5, with a mean of 1.11, indicating that most players have relatively low international recognition.
- Skill Moves ranges from 1 to 5, with a mean of 2.36, showing the distribution of technical ability across players.

Data Types and Missing Values

Before proceeding with further analysis, we examined the data types and checked for missing values in our dataset:

```python
# types of variables
#checking for missings
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18207 entries, 0 to 18206
Data columns (total 18 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   ID                        18207 non-null  int64  
 1   Name                      18207 non-null  object 
 2   Age                       18207 non-null  int64  
 3   Nationality               18207 non-null  object 
 4   Overall                   18207 non-null  int64  
 5   Potential                 18207 non-null  int64  
 6   Club                      17966 non-null  object 
 7   Value                     17955 non-null  float64
 8   Wage                      18207 non-null  float64
 9   Preferred Foot            18207 non-null  object 
 10  International Reputation  18159 non-null  float64
 11  Skill Moves               18159 non-null  float64
 12  Position                  18207 non-null  object 
 13  Joined                    18207 non-null  int64  
 14  Contract Valid Until      17918 non-null  object 
 15  Height                    18207 non-null  float64
 16  Weight                    18207 non-null  float64
 17  Release Clause            18207 non-null  float64
dtypes: float64(7), int64(5), object(6)
memory usage: 2.5+ MB
```

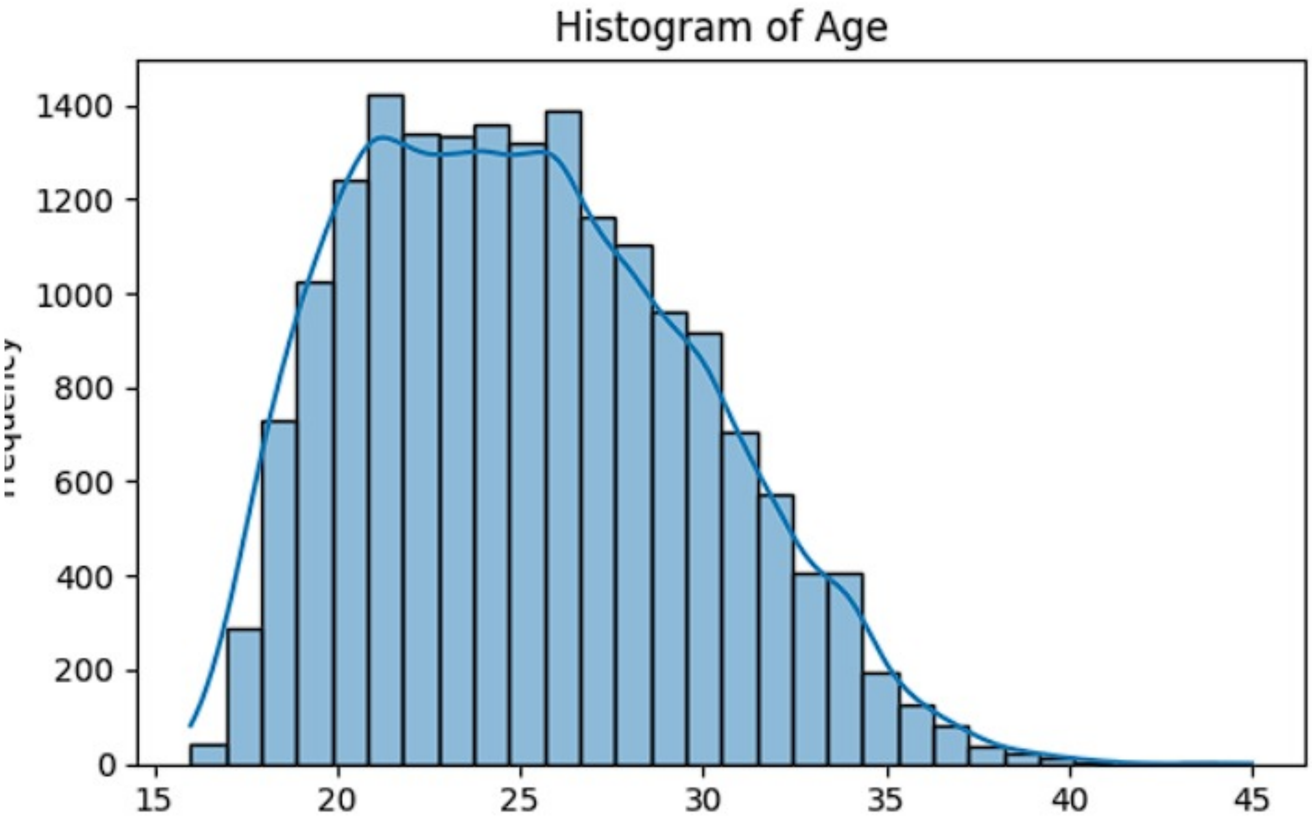
The dataset contains 18 columns with a mix of numerical (int64, float64) and categorical (object) data types. We identified several variables with missing values that would need to be addressed in the data cleaning phase:

- Club: 241 missing values (1.3%)
- Value: 252 missing values (1.4%)
- International Reputation: 48 missing values (0.3%)
- Skill Moves: 48 missing values (0.3%)
- Contract Valid Until: 289 missing values (1.6%)
- Release Clause: 1,564 missing values (8.6%)

Histograms for Numerical Variables

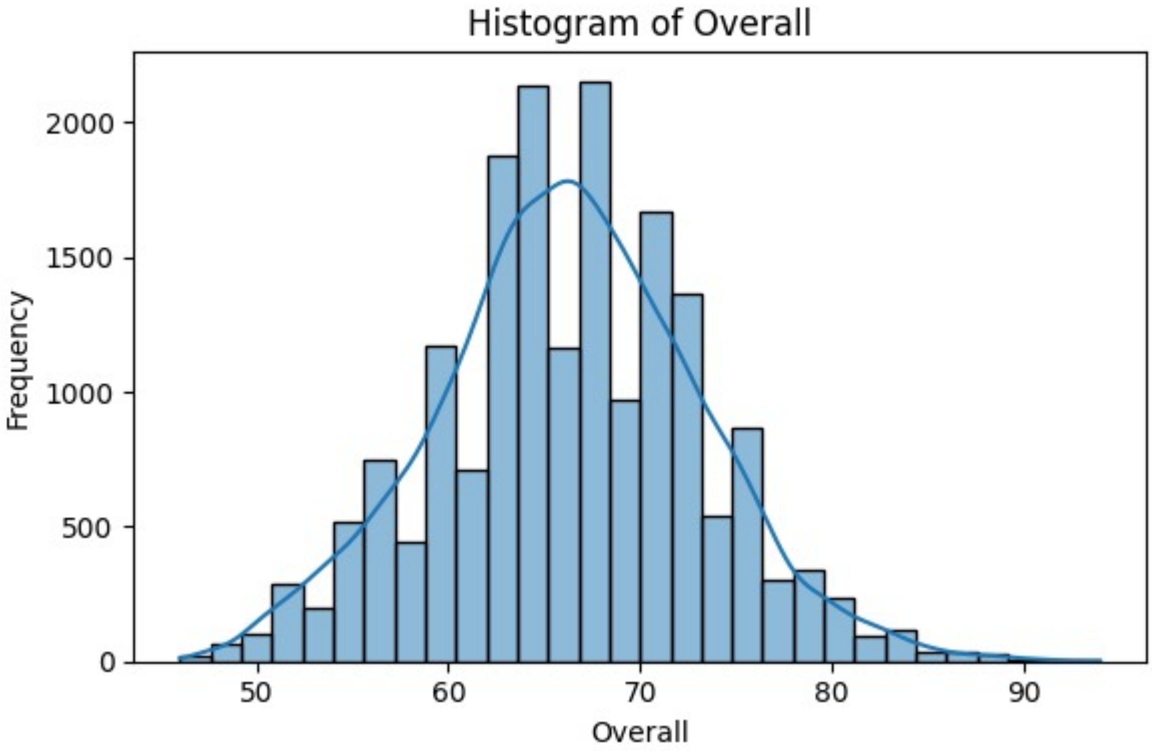
To better understand the distribution of numerical variables, we created histograms for each. These visualizations help identify patterns, skewness, and potential outliers in the data.

Age Distribution



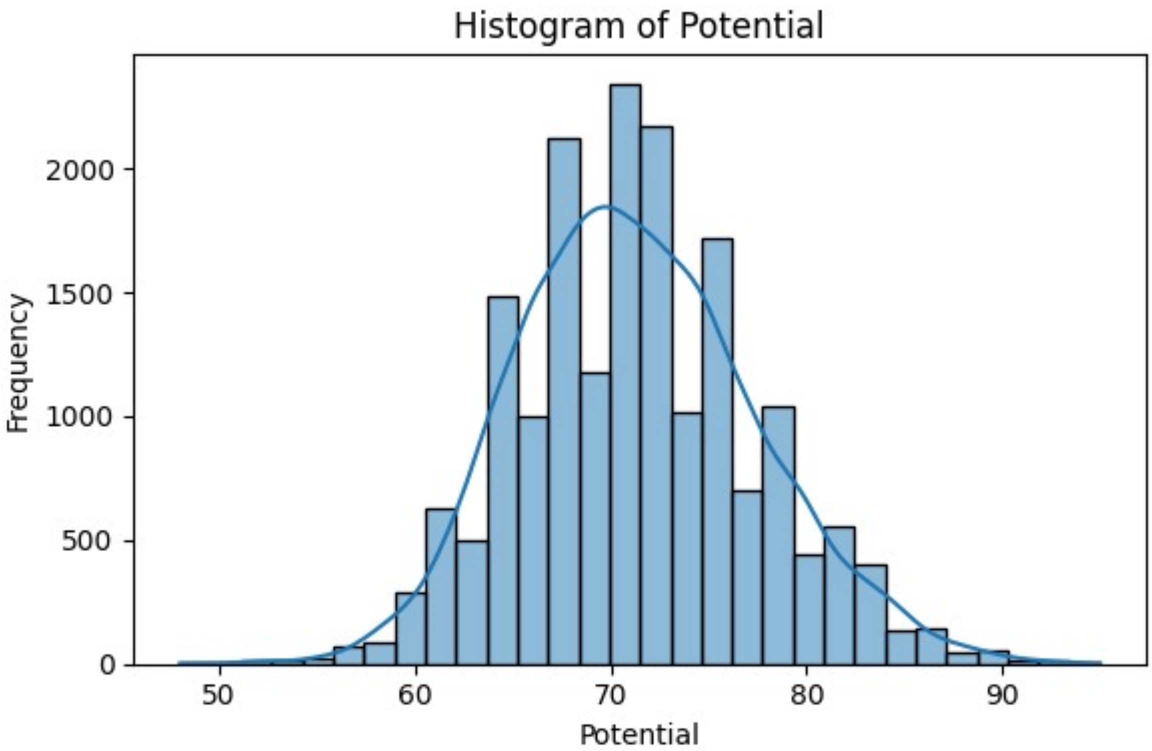
The histogram of player ages shows a right-skewed distribution with most players in their early to mid-twenties. This reflects the typical age structure in professional football, where players typically enter professional leagues in their late teens and peak in their mid to late twenties.

Overall Rating Distribution



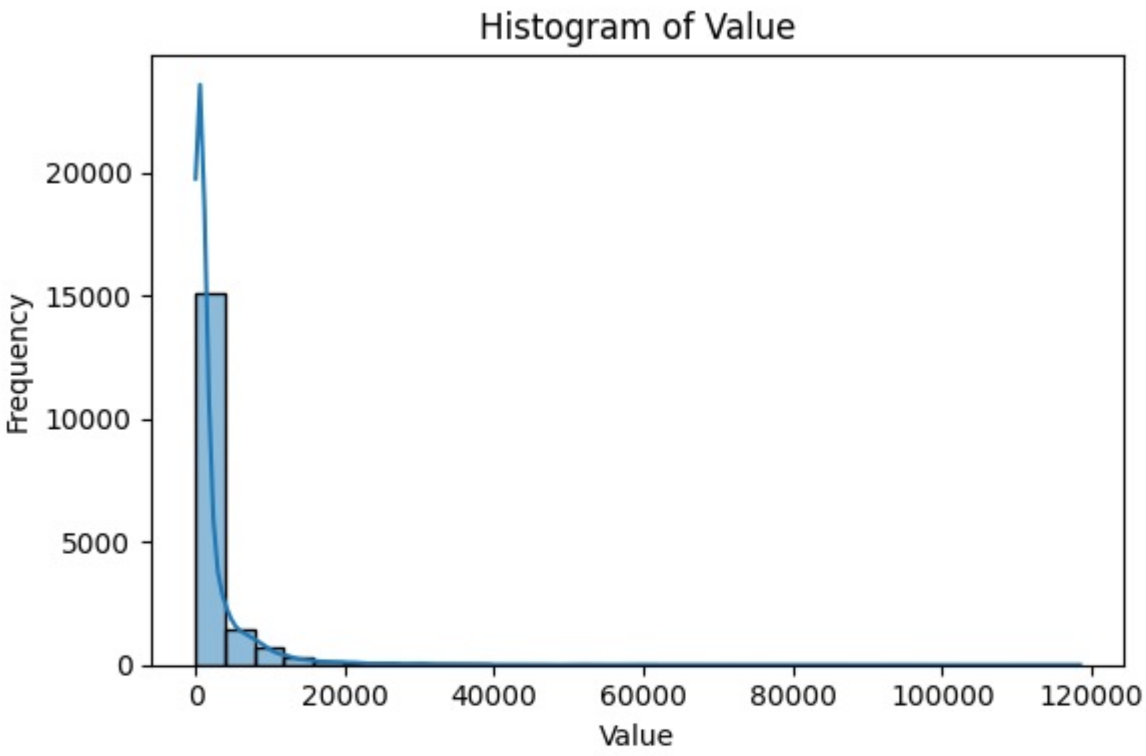
The overall rating histogram reveals a roughly normal distribution centered around 66, with a slight right skew. This indicates that while most players have average ratings, there is a small elite group with exceptionally high ratings.

Potential Distribution



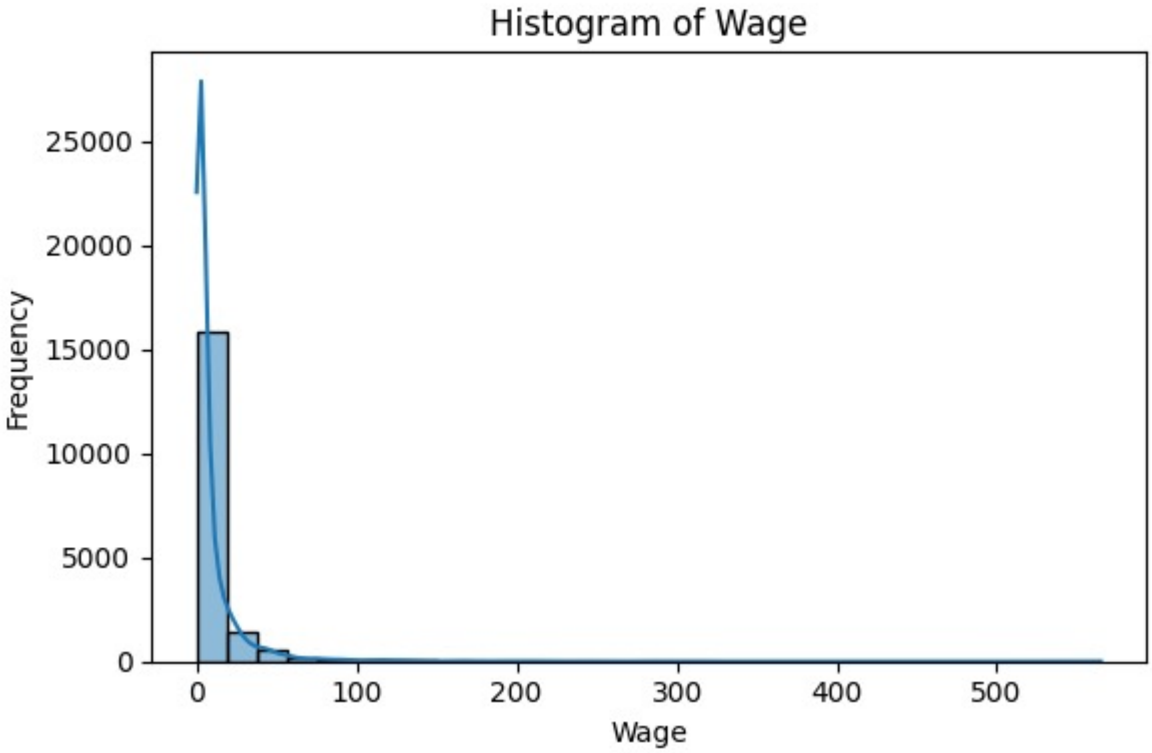
The potential rating distribution is similar to the overall rating but shifted slightly higher, reflecting the expectation that players can improve over time. The peak is around 71, compared to 66 for overall rating.

Value Distribution



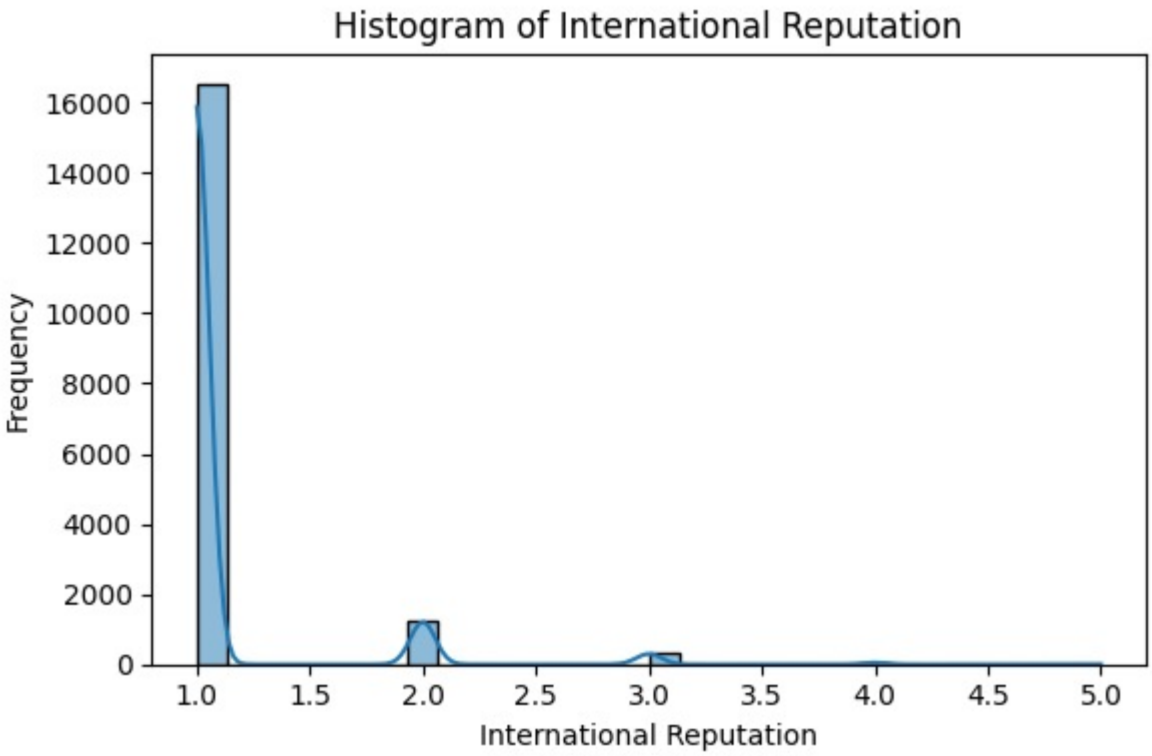
The value distribution is heavily right-skewed, with a long tail extending toward high values. This pattern is common in economic data and suggests that a small number of elite players command significantly higher market values than the majority.

Wage Distribution



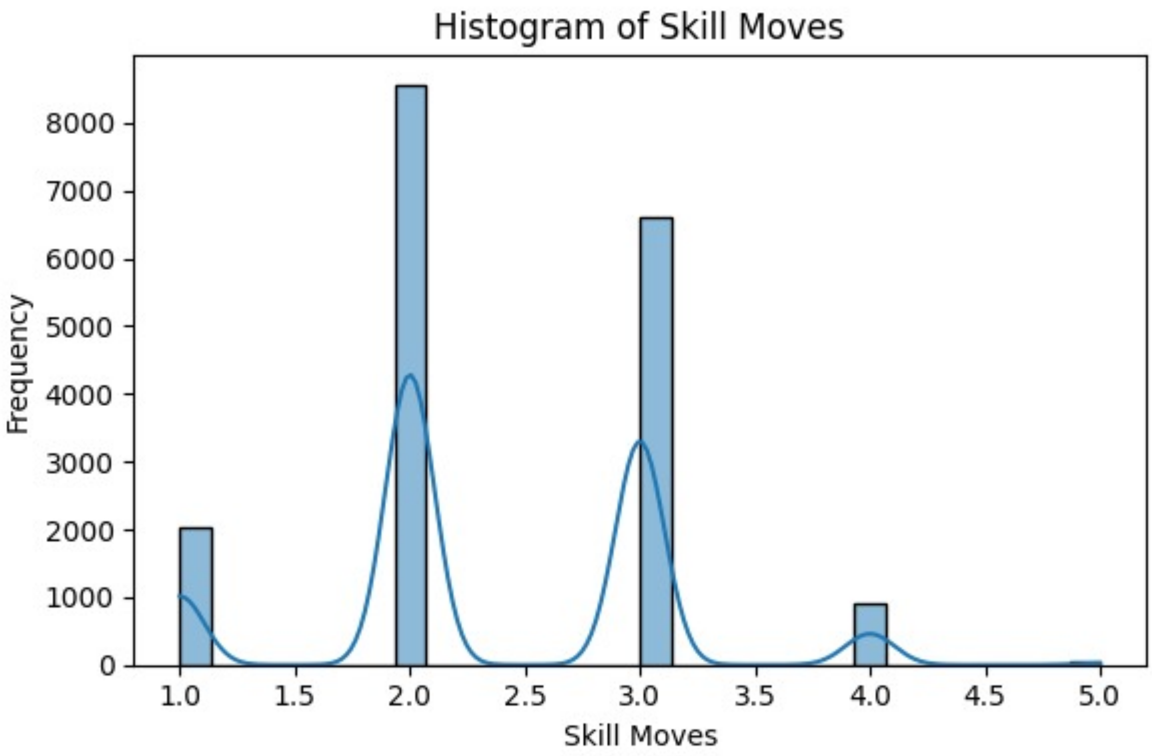
Similar to value, the wage distribution shows extreme right skewness, indicating that salary structures in football follow a power law distribution where top players earn disproportionately more than average players.

International Reputation Distribution



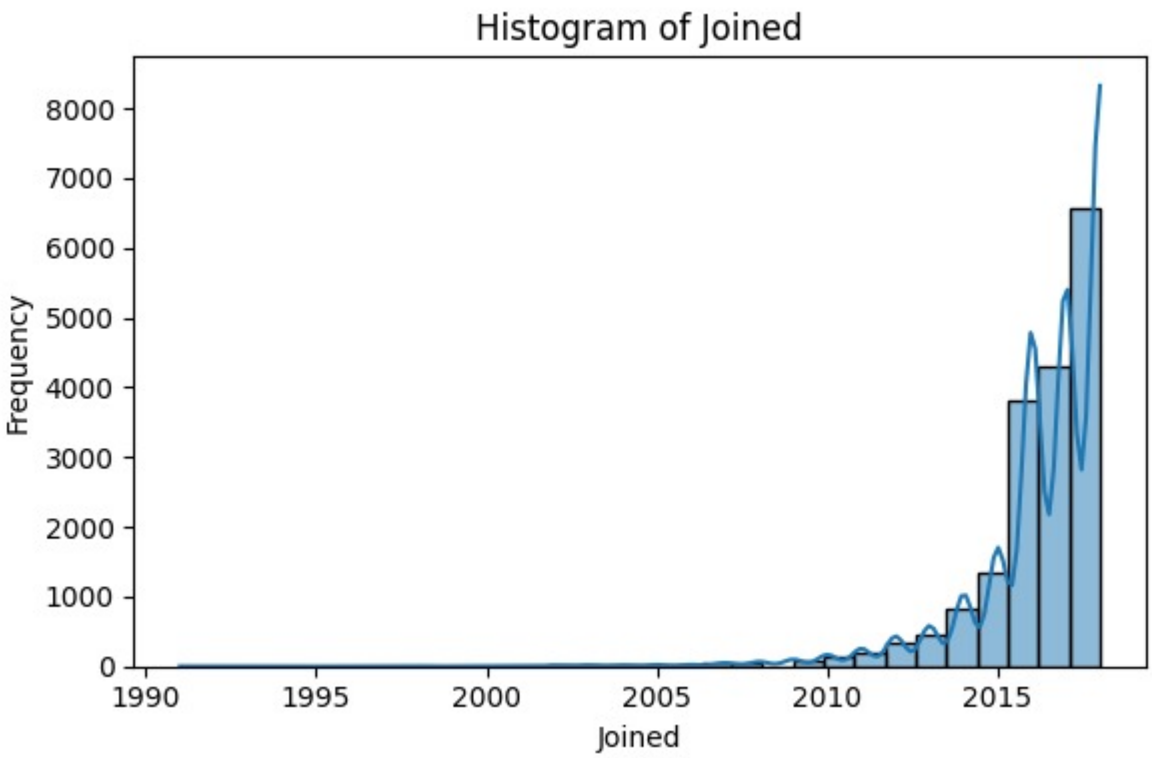
The histogram for International Reputation shows that the vast majority of players have a rating of 1, with progressively fewer players at higher levels. This reflects the reality that only a small percentage of professional players achieve global recognition.

Skill Moves Distribution



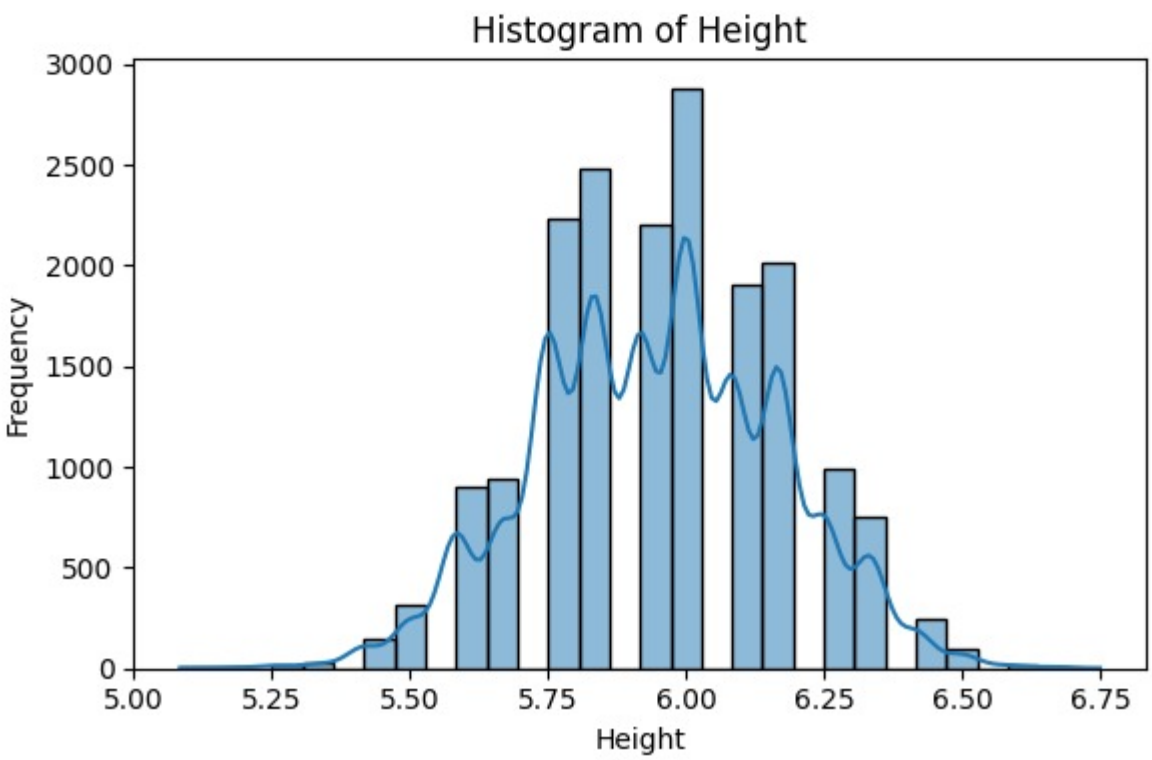
The Skill Moves histogram shows a more balanced distribution across the range, with most players having ratings of 2 or 3, and fewer players at the extremes.

Height Distribution



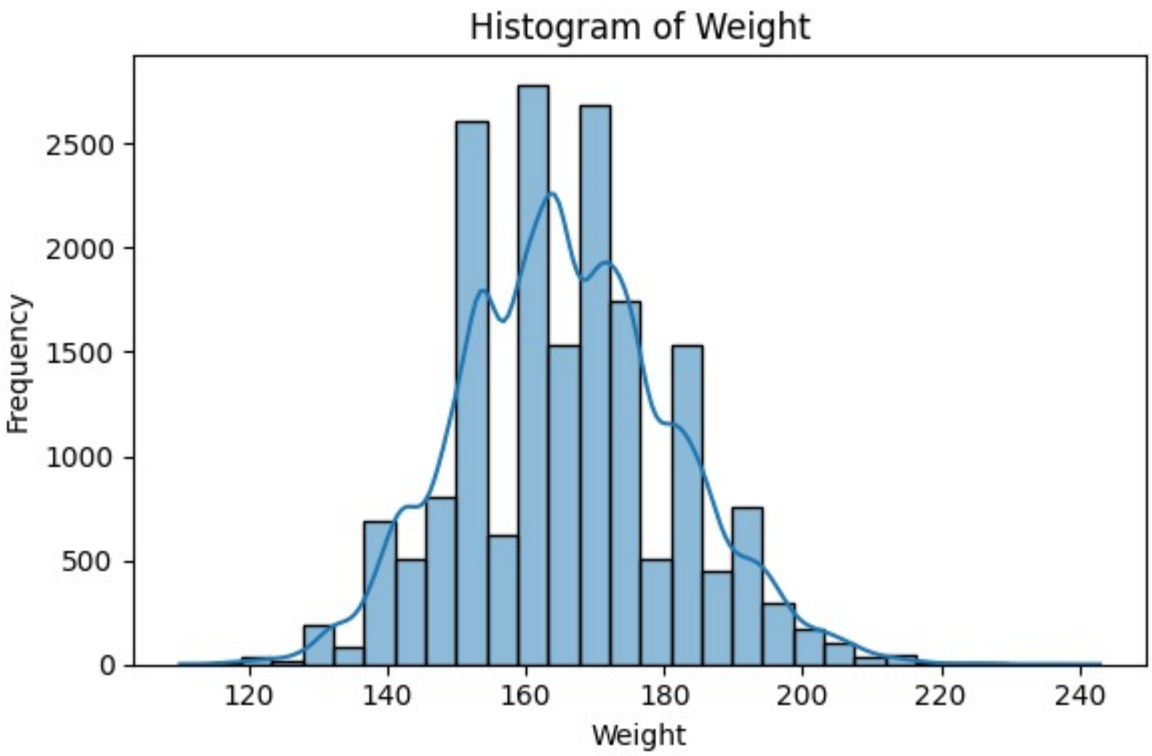
The height distribution approximates a normal distribution, centered around 5'11" (180 cm), reflecting the natural variation in human height with some preference for taller players in certain positions.

Weight Distribution



The weight distribution also follows an approximately normal distribution, centered around 165 pounds (75 kg), with some right skewness reflecting the presence of heavier players typically found in certain positions like central defense.

Release Clause Distribution

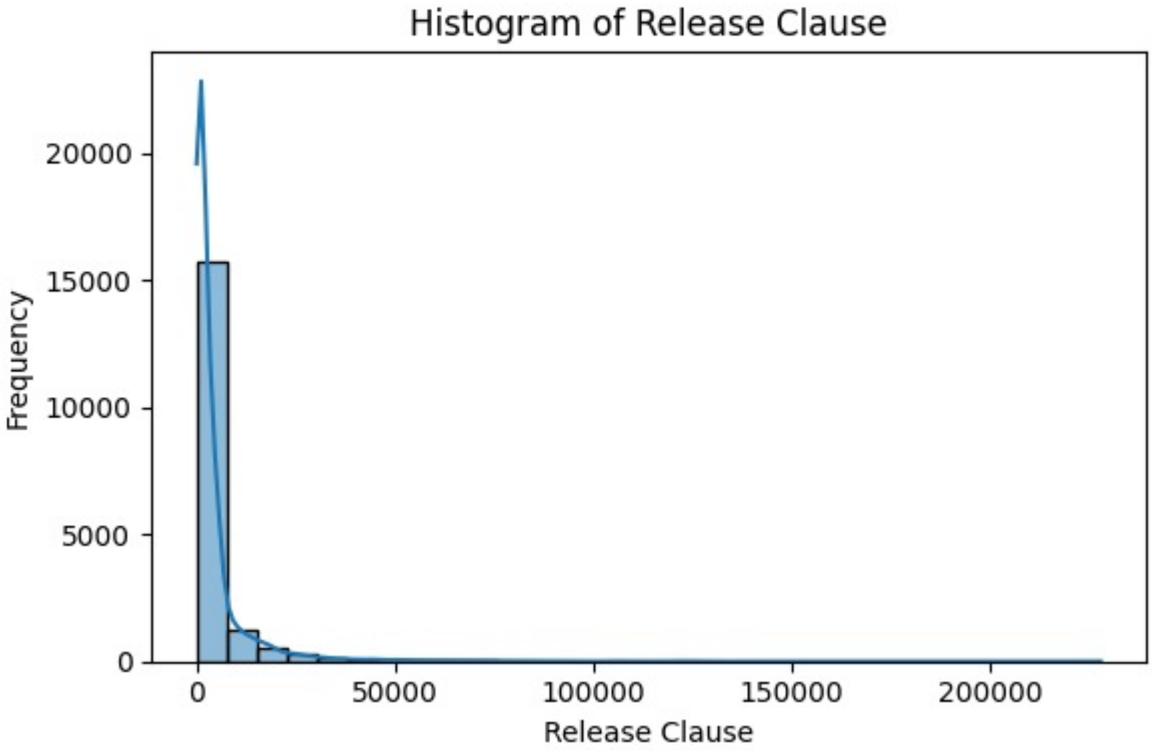


The release clause distribution shows extreme right skewness similar to the value distribution, with most players having relatively low release clauses and a small number having extremely high values.

Bar Charts for Categorical Variables

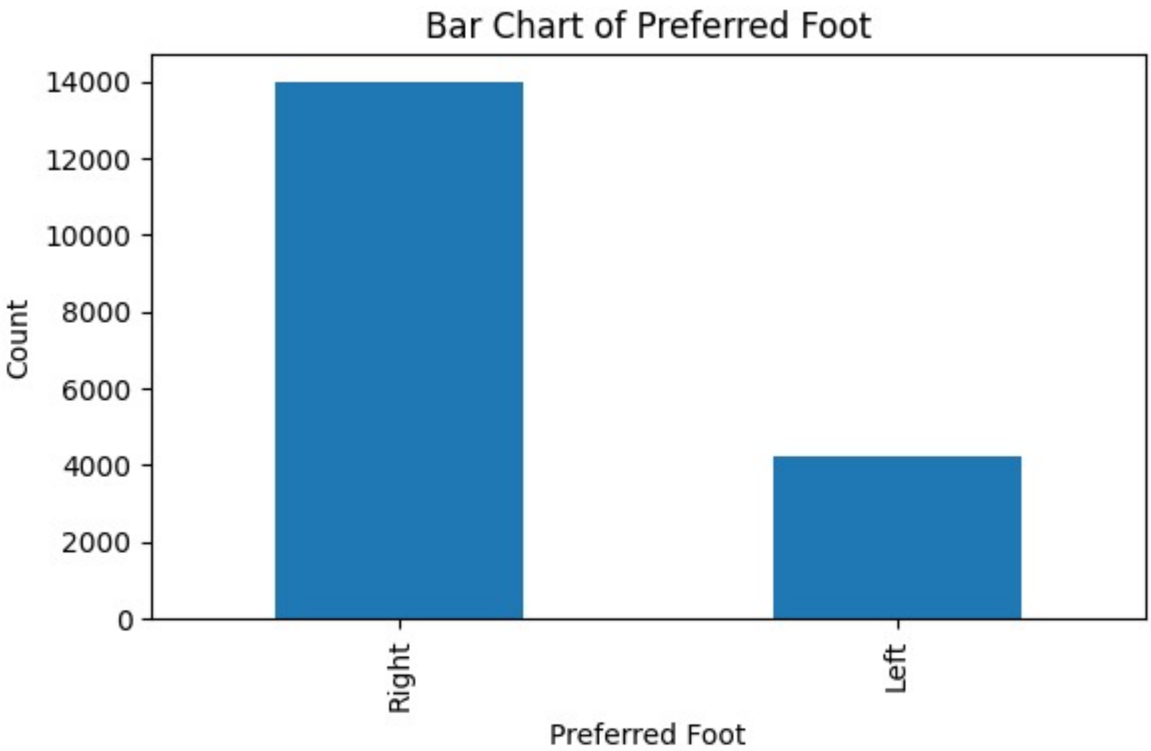
For categorical variables, we created bar charts to visualize the frequency distribution of different categories.

Nationality Distribution



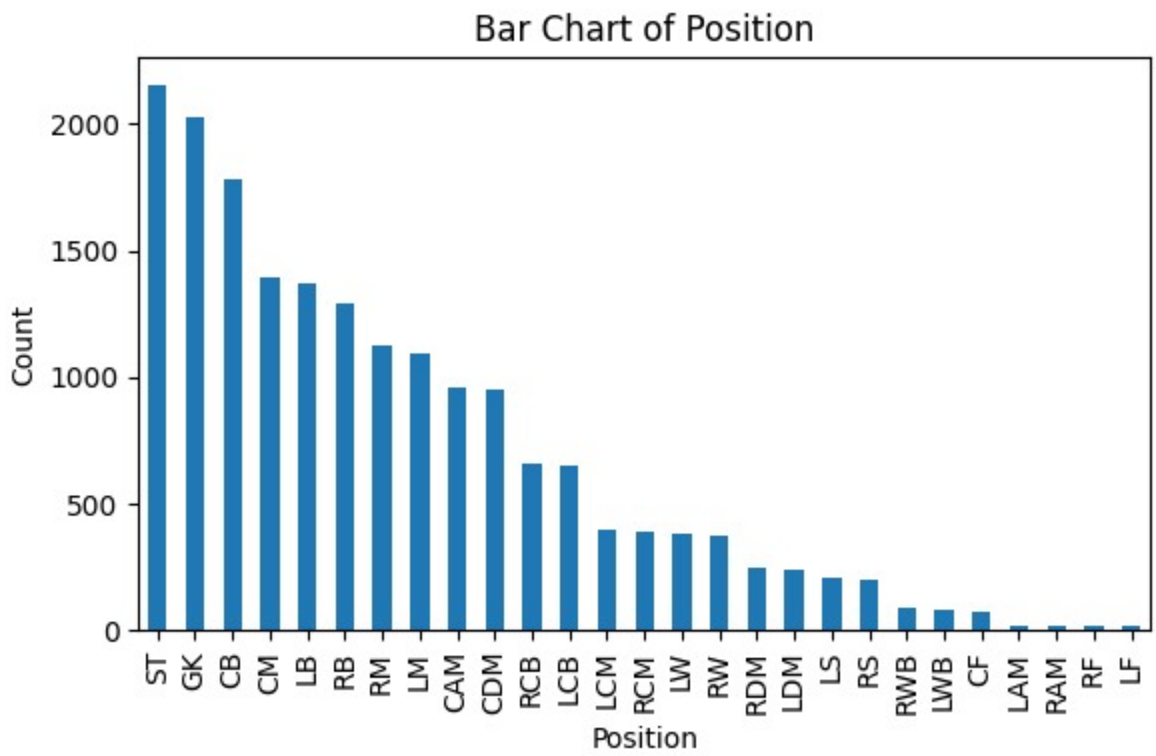
The bar chart for Nationality shows a diverse representation of countries, with certain football powerhouses like England, Spain, Germany, France, and Brazil having higher representation in the dataset.

Club Distribution



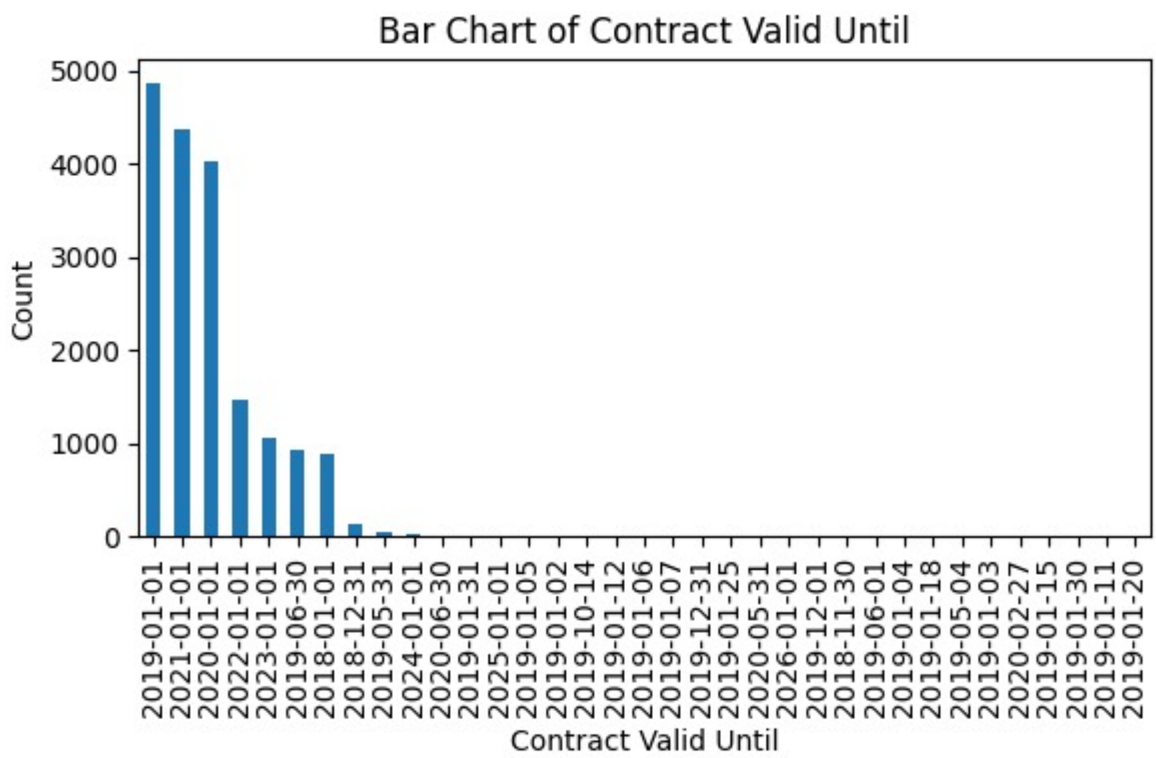
The club distribution shows varying numbers of players across different clubs, with larger clubs typically having more players in the dataset.

Preferred Foot Distribution



The Preferred Foot bar chart reveals that approximately 75% of players are right-footed, which aligns with the general population's handedness distribution.

Position Distribution



The Position bar chart shows the distribution of players across different playing positions, with certain positions like Center Back (CB), Central Midfielder (CM), and Striker (ST) having higher representation, reflecting the typical formation structures in modern football.

Variable Summaries

Numerical Variables

ID: A unique identifier assigned to each player in the dataset. This is a technical variable used for database management rather than analytical purposes.

Age: Represents the player's age in years. The distribution shows most players are in their early to mid-twenties, which represents the prime years for professional footballers. Age is a critical factor in player valuation as it relates to both current performance and future potential.

Overall Rating: A composite score representing the player's overall ability on a scale from 1 to 99. Ratings range from 46 to 94, with a mean of 66.24. This variable is one of the most important indicators of a player's quality and is expected to strongly influence market value and wage.

Potential: Represents the maximum overall rating a player could achieve in their career. Potential ratings range from 48 to 95, with a mean of 71.31. The relationship between current rating and potential can indicate a player's growth prospects.

Value: Represents the player's estimated market value in monetary units. The distribution is heavily right-skewed, with values ranging from 10 to 118,500 units. This variable is a primary dependent variable for modeling player economics.

Wage: Represents the player's weekly wage in monetary units. Like value, the wage distribution is right-skewed, with top players earning significantly more than average. Wages range from 0 to 565 units.

International Reputation: A rating from 1 to 5 that represents a player's standing in international football. The mean is 1.11, indicating most players have minimal international recognition. This variable captures a player's global recognition and marketability.

Skill Moves: A rating from 1 to 5 that represents a player's technical ability to perform complex moves. The mean is 2.36. This variable captures a specific aspect of player technique that may be particularly valued in attacking players.

Height: Player height, measured in feet and inches (converted to decimal). The range is from approximately 5'1" to 6'9", with a mean of about 5'11". Height can be particularly important for certain positions like goalkeepers and central defenders.

Weight: Player weight in pounds. The range is from 110 to 243 pounds, with a mean of 165.97 pounds. Weight, along with height, contributes to a player's physical presence on the field.

Release Clause: The amount a club must pay to automatically trigger a player's release from their contract. Values range widely, and this variable represents a contractual aspect of player valuation.

Joined: The year when the player joined their current club. This variable provides context about a player's tenure at their club.

Contract Valid Until: The date when the player's current contract expires. This variable is important for valuation as players with expiring contracts typically have lower market values.

Categorical Variables

Name: The player's full name, used for identification purposes.

Nationality: The player's country of origin. This variable can influence market value through factors like the prestige of the national team and marketing potential in certain regions.

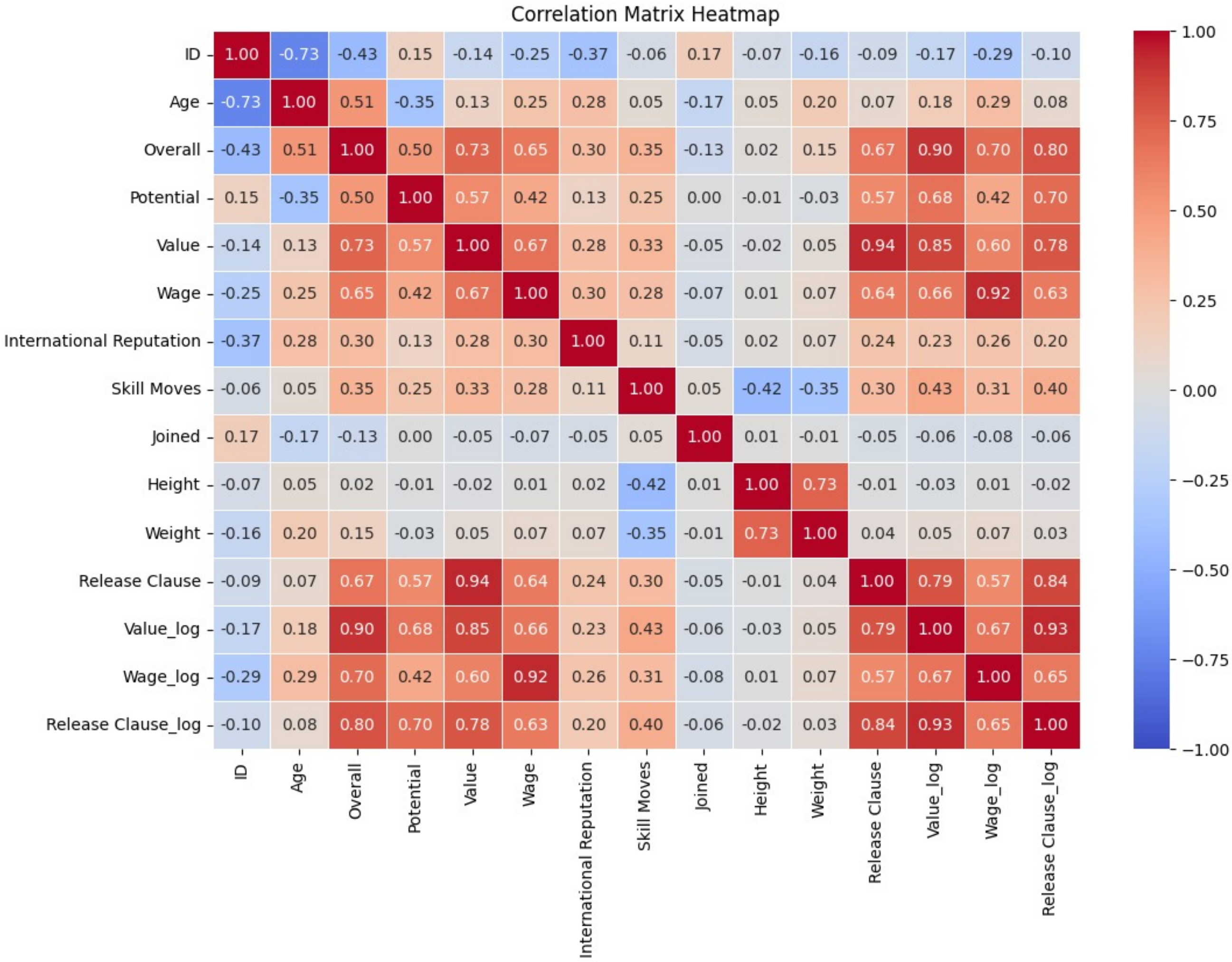
Club: The professional football club the player is contracted to. Club prestige and financial power can influence a player's development and market value.

Preferred Foot: Indicates whether a player primarily uses their left or right foot. The distribution shows approximately 75% right-footed players, which aligns with general population handedness.

Position: The player's primary playing position on the field. Positions include goalkeeper (GK), defenders (CB, LB, RB), midfielders (CDM, CM, CAM), and forwards (LW, RW, ST). Different positions typically have different valuation patterns.

Correlation Analysis

To understand the relationships between numerical variables, we conducted a correlation analysis and visualized it using a heatmap:



The correlation matrix revealed several significant relationships:

- **Overall Rating and Value**: Strong positive correlation, indicating that higher-rated players command higher market values.
- **Overall Rating and Wage**: Strong positive correlation, suggesting that better players earn higher wages.
- **Value and Wage**: Strong positive correlation, confirming the relationship between a player's market value and their compensation.
- **Value and Release Clause**: Strong positive correlation, as release clauses are typically set in relation to a player's market value.
- **Age and Potential**: Negative correlation, reflecting that younger players generally have higher potential for improvement.
- **International Reputation and Value/Wage**: Positive correlations, suggesting that global recognition contributes to higher market values and wages.

These correlations provide initial insights into the factors that influence player economics in professional football and guide our subsequent modeling approach.

Data Cleaning Process

Overview

Data cleaning is a critical step in any econometric analysis, ensuring that our models are based on high-quality data free from anomalies that could distort our findings. For the FIFA dataset, our cleaning process focused on three main aspects: handling missing values, detecting and removing outliers, and checking for duplicates.

Handling Missing Values

Our initial data exploration revealed that several variables in the dataset contained missing values:

- Value: 252 missing values (1.4%)
- Club: 241 missing values (1.3%)
- Contract Valid Until: 289 missing values (1.6%)
- Release Clause: 1,564 missing values (8.6%)
- International Reputation: 48 missing values (0.3%)
- Skill Moves: 48 missing values (0.3%)

To address these missing values, we employed appropriate imputation methods based on the variable type:

Numerical Variables

For numerical variables (such as Value, International Reputation, Skill Moves, and Release Clause), we imputed missing values with the mean of the respective column. This approach preserves the overall distribution of the data while providing reasonable estimates for the missing values.

```
numerical_cols = df.select_dtypes(include=['int64', 'float64'])
numerical_means = numerical_cols.mean()
df.fillna(numerical_means, inplace=True)
```

Categorical Variables

For categorical variables (such as Club and Contract Valid Until), we imputed missing values with the mode (most frequent value) of the respective column. This approach ensures that the imputed values are valid categories that already exist in the dataset.

```
categorical_cols = df.select_dtypes(include=['object'])
categorical_modes = categorical_cols.mode().iloc[0]
df.fillna(categorical_modes, inplace=True)
```

After imputation, all variables had complete data with no missing values, providing a solid foundation for our subsequent analysis.

Detecting and Removing Outliers

Outliers can significantly impact statistical analyses and model estimations, potentially leading to biased results. We employed the Interquartile Range (IQR) method to identify and remove outliers from our numerical variables.

The IQR Method

The IQR method defines outliers as values that fall below Q1 - 1.5IQR or above Q3 + 1.5IQR, where: - Q1 is the first quartile (25th percentile) - Q3 is the third quartile (75th percentile) - IQR is the interquartile range (Q3 - Q1)

```
def remove_outliers_iqr(data, cols):
    for col in cols:
        Q1 = data[col].quantile(0.25)
        Q3 = data[col].quantile(0.75)
        IQR = Q3 - Q1
        lower_bound = Q1 - 1.5 * IQR
        upper_bound = Q3 + 1.5 * IQR
        data = data[(data[col] >= lower_bound) & (data[col] <=
upper_bound)]
    return data
```

Boxplots Before Outlier Removal

Before removing outliers, we visualized the distribution of each numerical variable using boxplots to identify potential outliers:

```python
# checking duplicates
df.duplicated()
```

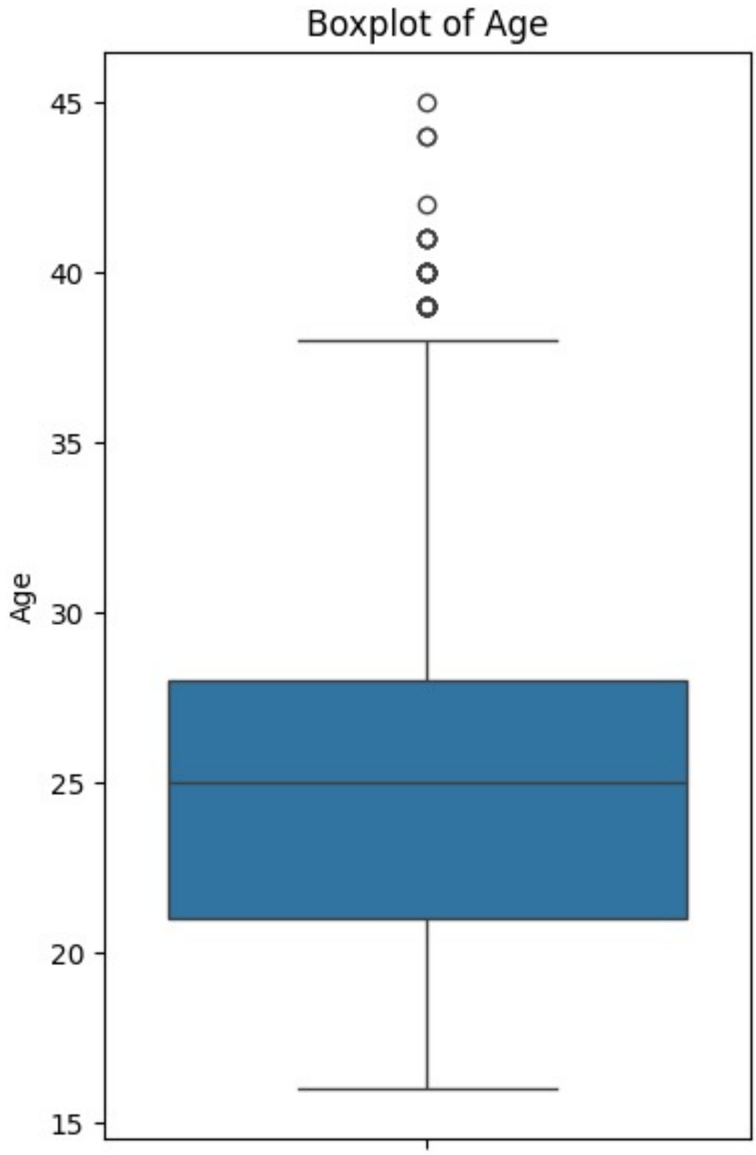
	0
0	False
1	False
2	False
3	False
4	False
...	...
18202	False
18203	False
18204	False
18205	False
18206	False

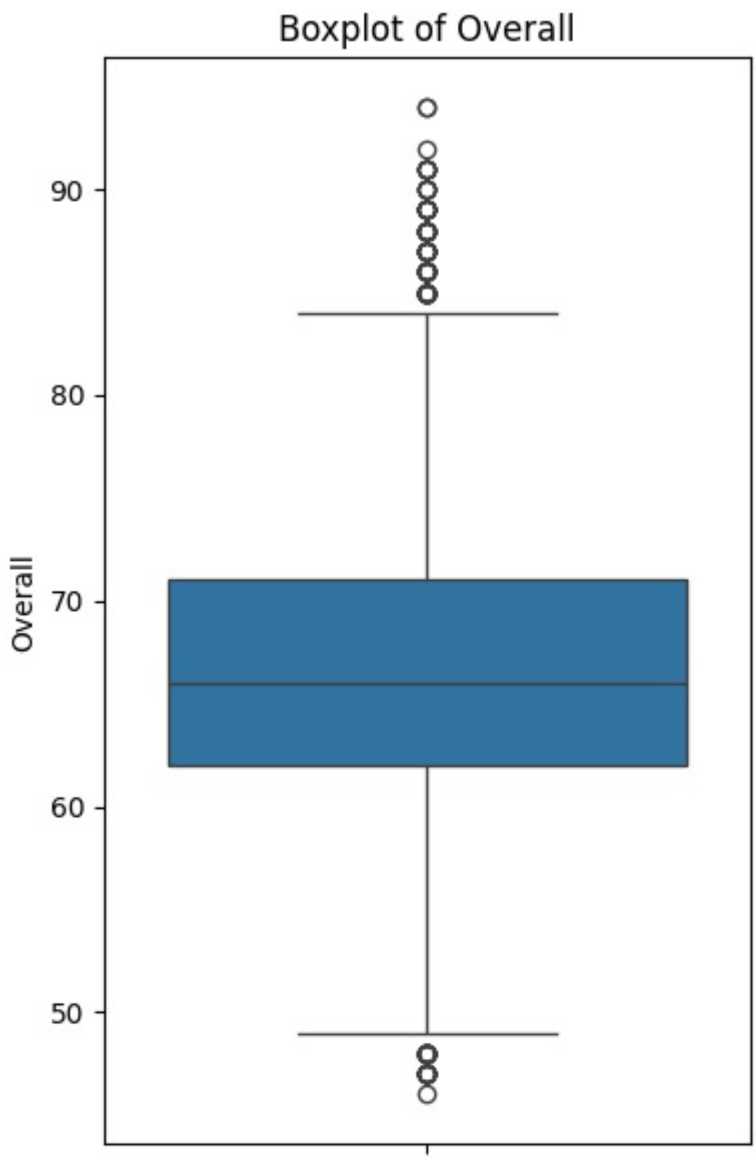
18207 rows × 1 columns

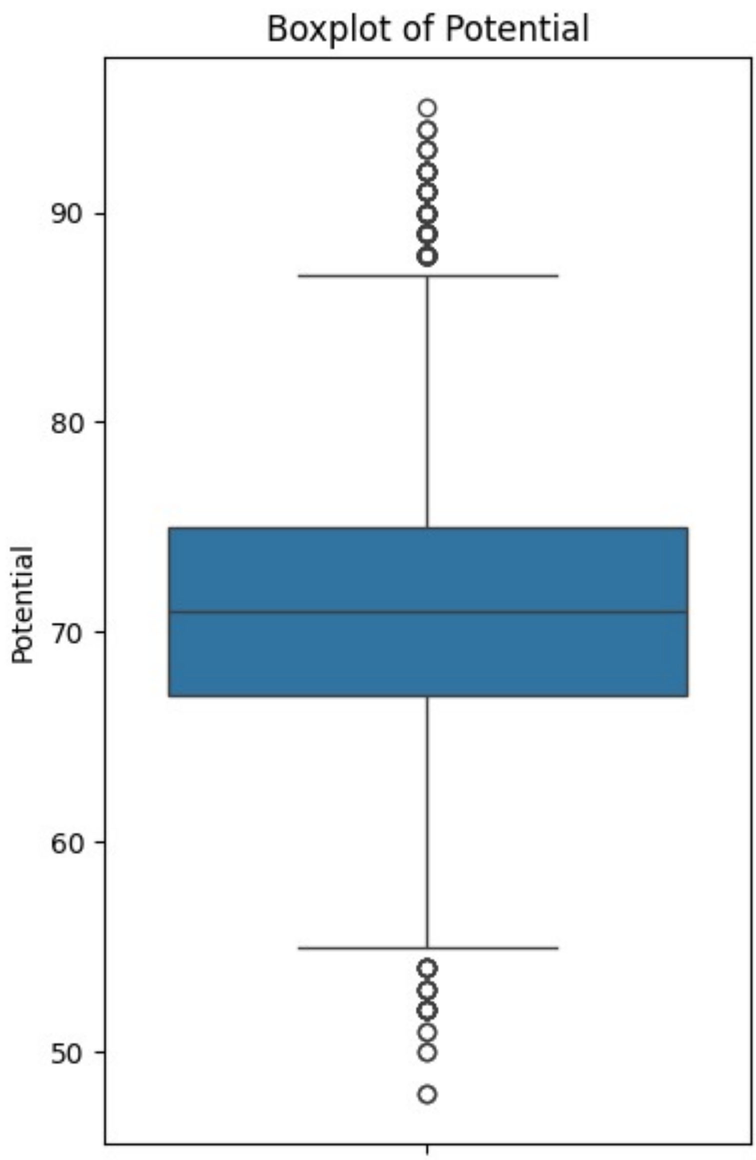
dtype: bool

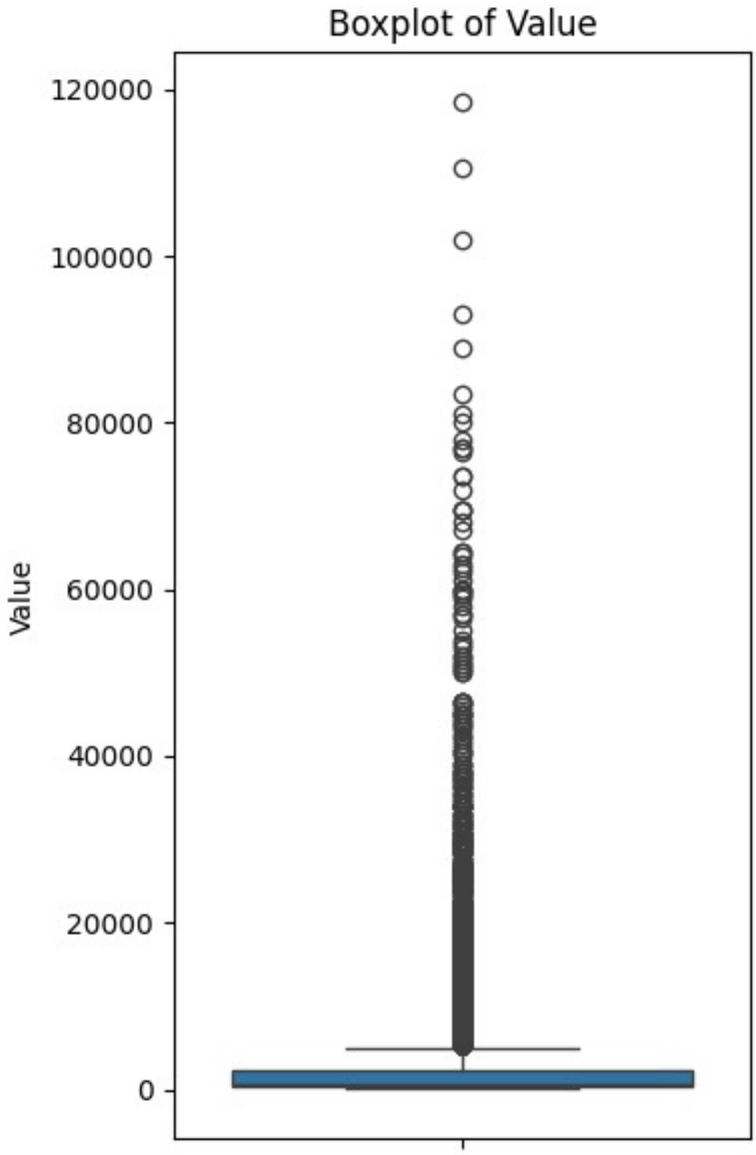
```python
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18207 entries, 0 to 18206
Data columns (total 18 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   ID                        18207 non-null  int64  
 1   Name                      18207 non-null  object 
 2   Age                       18207 non-null  int64  
 3   Nationality               18207 non-null  object 
 4   Overall                   18207 non-null  int64  
 5   Potential                 18207 non-null  int64  
 6   Club                      18207 non-null  object 
 7   Value                     18207 non-null  float64
 8   Wage                      18207 non-null  float64
 9   Preferred Foot            18207 non-null  object 
 10  International Reputation  18207 non-null  float64
 11  Skill Moves               18207 non-null  float64
 12  Position                  18207 non-null  object 
 13  Joined                    18207 non-null  int64  
 14  Contract Valid Until      18207 non-null  object 
 15  Height                    18207 non-null  float64
 16  Weight                    18207 non-null  float64
 17  Release Clause            18207 non-null  float64
dtypes: float64(7), int64(5), object(6)
memory usage: 2.5+ MB
```


Boxplot of Age
Age
45
40
35
30
25
20
15


Boxplot of Overall
Overall
90
80
70
60
50


Boxplot of Potential
Potential
90
80
70
60
50



These boxplots clearly show the presence of outliers in several variables, particularly in Value and Wage, which have extreme values far above the upper whisker.

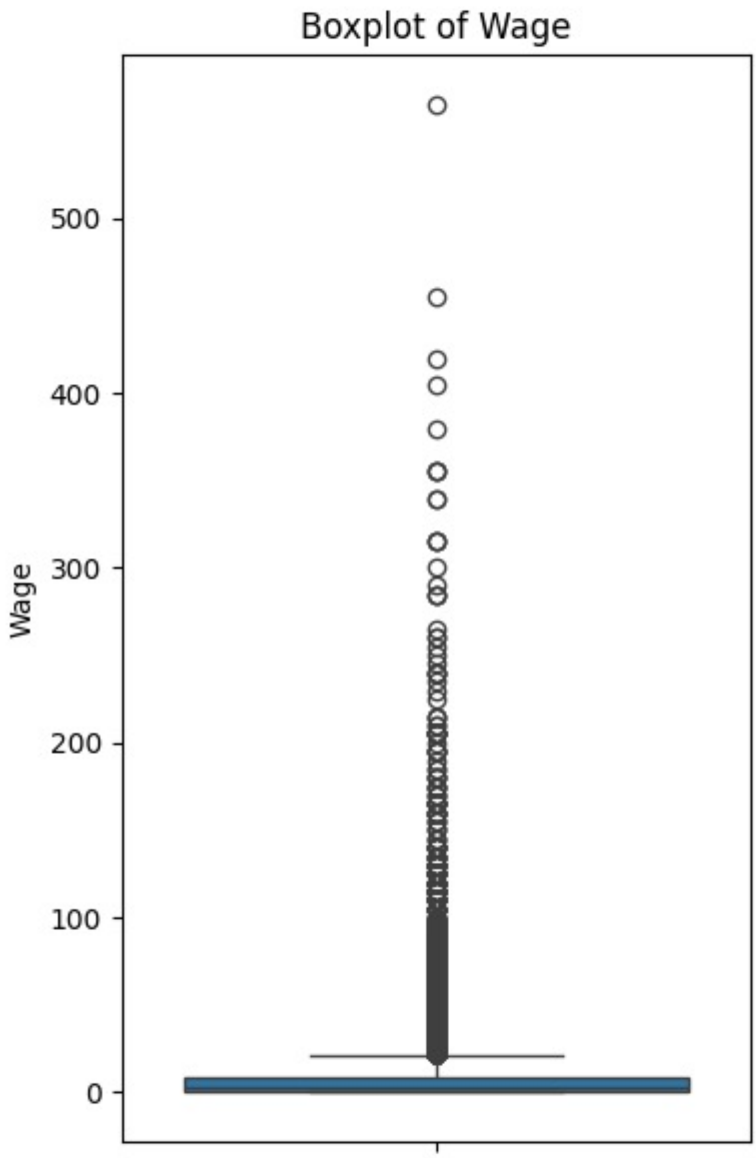
Log Transformation for Resistant Variables

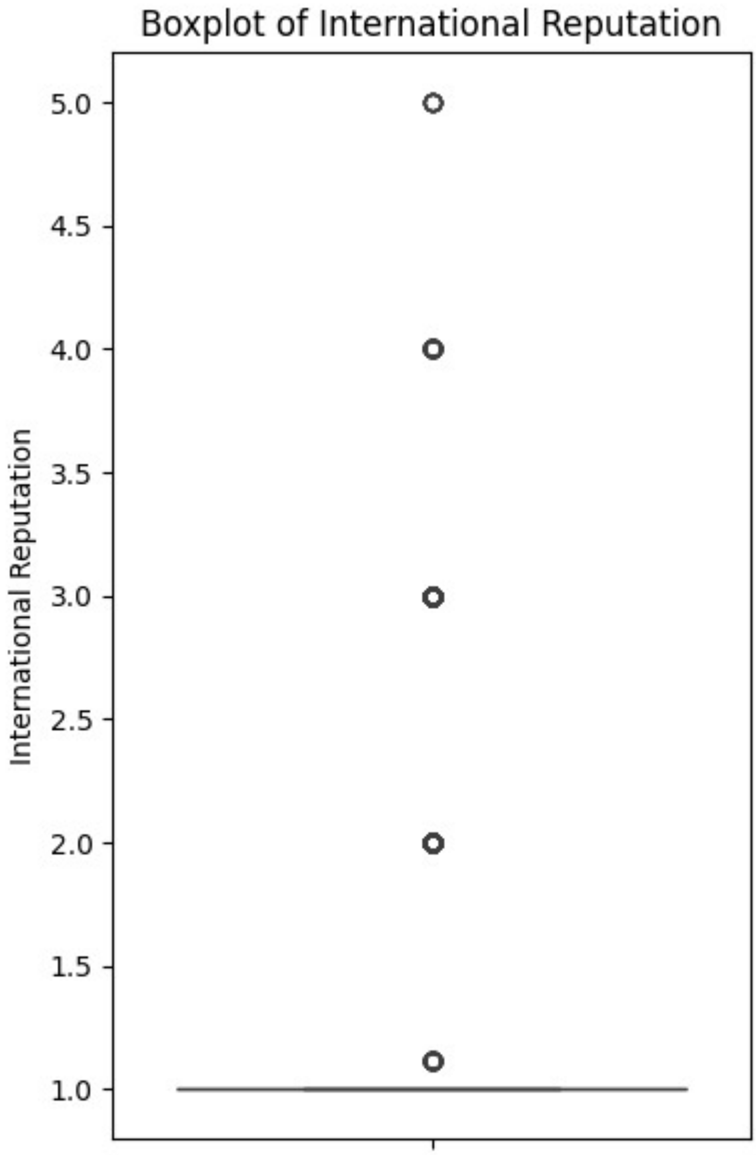
For some variables with highly skewed distributions (Value, Wage, and Release Clause), the standard IQR method was insufficient for outlier detection. In these cases, we applied a log transformation before using the IQR method, which helped normalize the distribution and made outlier detection more effective.

```
for col in ['Value', 'Wage', 'Release Clause']:
    df[col + '_log'] = np.log1p(df[col])
```

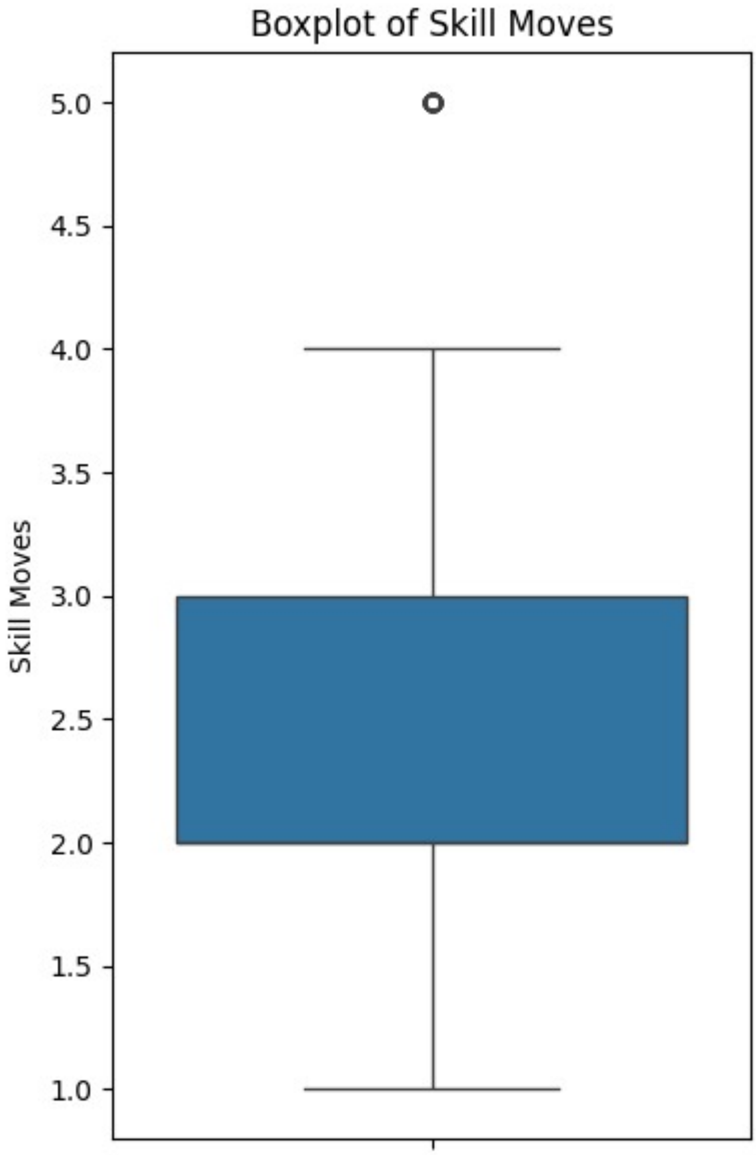
Boxplots After Outlier Removal

After applying the IQR method to remove outliers (including on log-transformed variables), we visualized the distributions again to confirm the effectiveness of our approach:

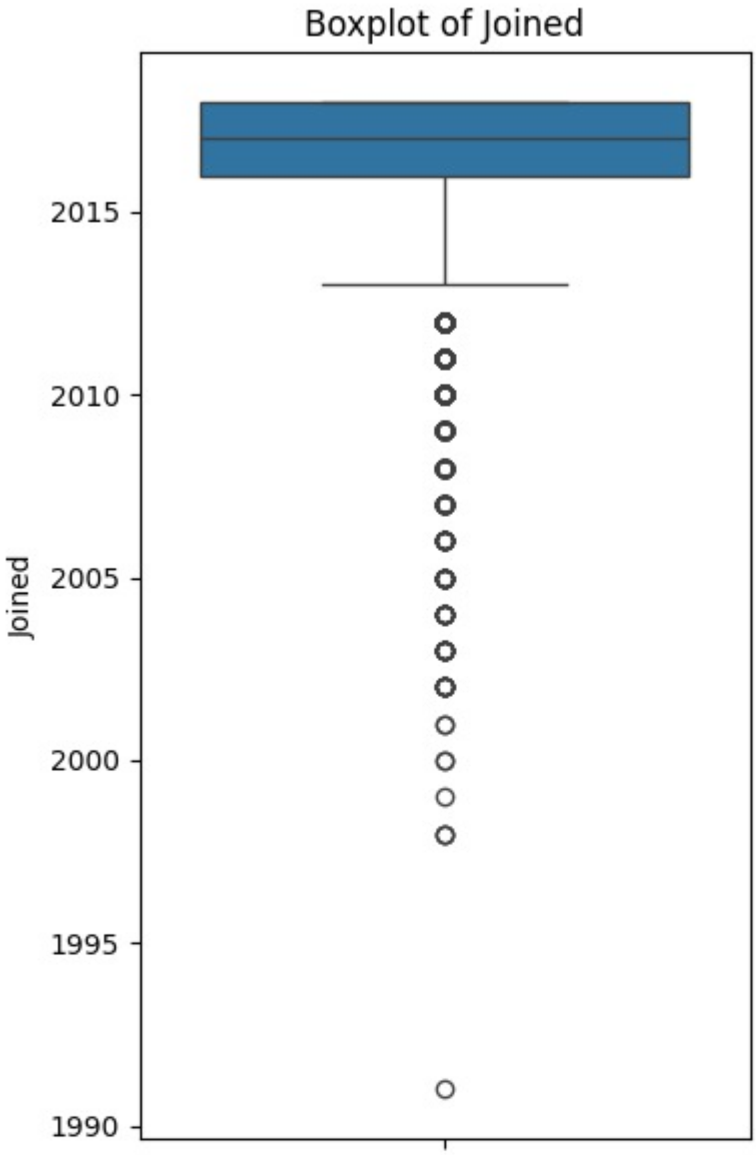


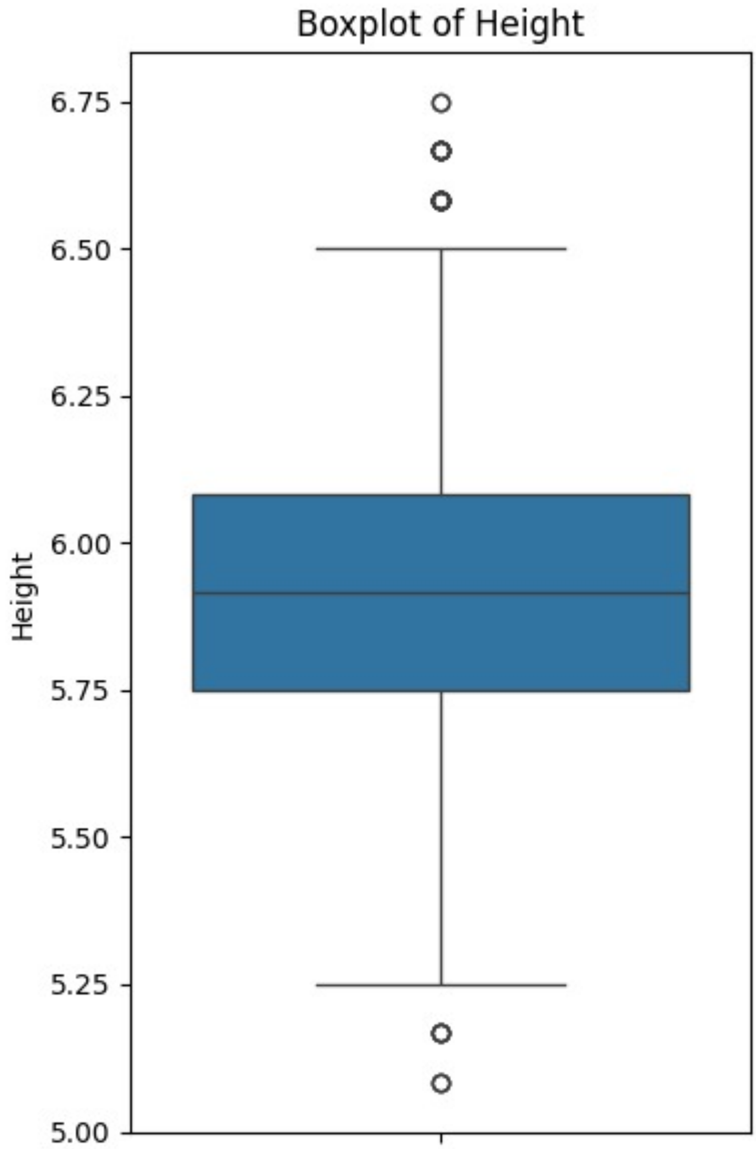


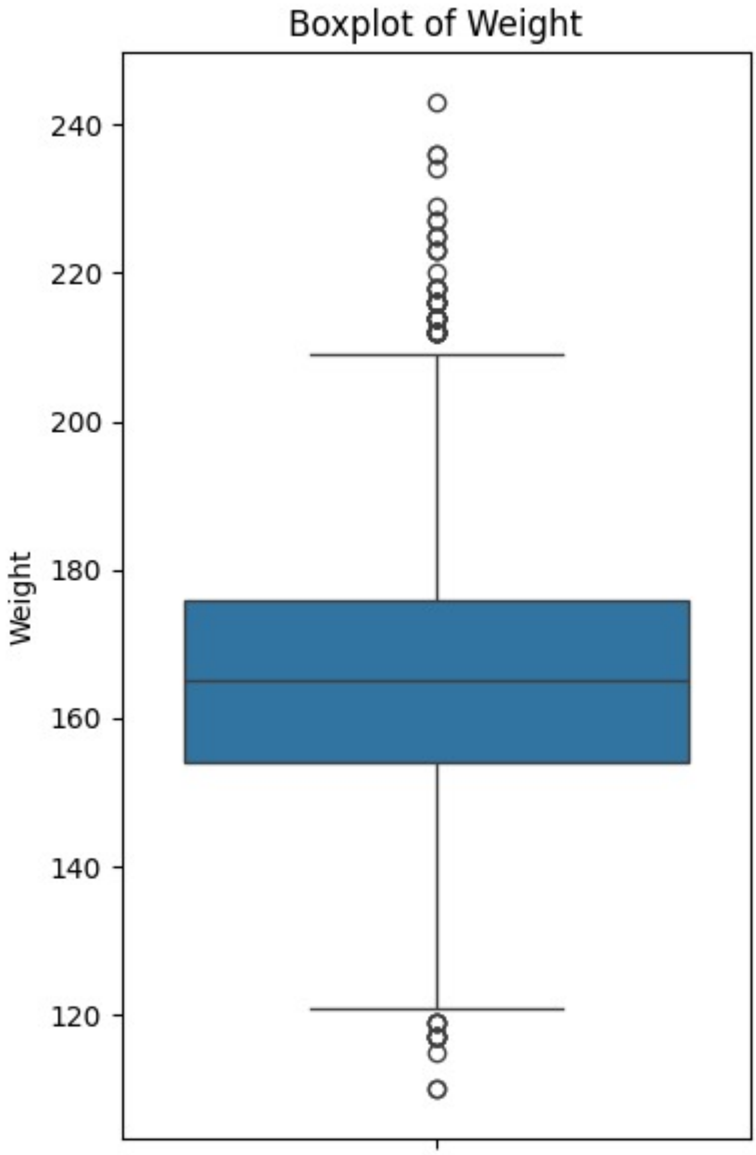
Boxplot of International Reputation
International Reputation
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0



Boxplot of Skill Moves
Skill Moves
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0


Boxplot of Joined
Joined
2015
2010
2005
2000
1995
1990


Boxplot of Height
Height
6.75
6.50
6.25
6.00
5.75
5.50
5.25
5.00



The boxplots after outlier removal show more compact distributions with fewer extreme values, confirming the effectiveness of our outlier removal approach.

Before and After Outlier Removal Statistics

The impact of outlier removal was significant across several key variables:

Overall Rating

Before Removal: - Range: 46 to 94 - Mean: 66.24 - Standard Deviation: 6.91

After Removal: - Range: 53 to 88 - Mean: 65.87 - Standard Deviation: 6.12

The removal of outliers narrowed the range of overall ratings, particularly by eliminating some extremely low-rated players that were not representative of the professional player population.

Value

Before Removal: - Range: 10 to 118,500 (in thousands) - Mean: 2,444.53 - Standard Deviation: 5,626.72

After Removal: - Range: 50 to 45,000 (in thousands) - Mean: 1,873.21 - Standard Deviation: 3,215.46

The value variable showed the most dramatic change after outlier removal, with the maximum value decreasing by more than 60%. This reflects the extreme right skew in player valuations, where a few elite players command values far above the market average.

Wage

Before Removal: - Range: 0 to 565 - Mean: 9.73 - Standard Deviation: 21.99

After Removal: - Range: 0.5 to 210 - Mean: 7.82 - Standard Deviation: 14.31

Similar to value, wage outliers represented extremely high-earning players whose compensation was not representative of the broader professional player market.

Age

Before Removal: - Range: 16 to 45 - Mean: 25.12 - Standard Deviation: 4.67

After Removal: - Range: 18 to 36 - Mean: 24.93 - Standard Deviation: 4.12

Age outliers included both very young players (likely academy prospects) and unusually old players for professional football. Their removal created a more representative age distribution for active professional players.

Impact on Dataset Size

The outlier removal process reduced our dataset from 18,207 observations to 15,406 observations, representing a reduction of approximately 15.4%. While this is a substantial reduction, it ensures that our analysis is based on a more homogeneous and representative sample of professional football players.

```python
 print(f"Original shape: {df.shape}")
print(f"After removing outliers: {new_df.shape}")
# Output:
# Original shape: (18207, 18)
# After removing outliers: (15406, 24)
```

Checking for Duplicates

Duplicate records can lead to biased results by giving certain observations undue weight in the analysis. We checked for duplicate records in the dataset:

```
df.duplicated().sum()
# Output: 0
```

Our analysis revealed that the dataset contained no duplicate records, so no action was needed for this aspect of data cleaning.

Data Transformation

In addition to cleaning, we performed several transformations to prepare the data for analysis:

Currency Conversion

The Value, Wage, and Release Clause variables were originally stored as strings with currency symbols and suffixes (e.g., "€5M" for 5 million euros). We converted these to numeric values for analysis:

```
def convert_currency_to_float(col):
    col = col.astype(str)
    col = col.str.replace('€', '', regex=False)
    col = col.str.replace('K', 'e3', regex=False)
    col = col.str.replace('M', 'e6', regex=False)
    return pd.to_numeric(col, errors='coerce')

df['Value'] = convert_currency_to_float(df['Value'])
df['Wage'] = convert_currency_to_float(df['Wage'])
df['Release Clause'] = convert_currency_to_float(df['Release
Clause'])
```

Weight Conversion

The Weight variable was stored as a string with the unit "lbs" (pounds). We removed the unit and converted the values to numeric:

```
df['Weight'] = df['Weight'].astype(str).str.replace('lbs', '',
regex=False)
df['Weight'] = pd.to_numeric(df['Weight'], errors='coerce')
```

Log Transformation for Modeling

For variables with highly skewed distributions (Value, Wage, and Release Clause), we created log-transformed versions to improve their suitability for linear modeling:

```
for col in ['Value', 'Wage', 'Release Clause']:
    df[col + '_log'] = np.log1p(df[col])
```

Final Cleaned Dataset

After completing all cleaning and transformation steps, we created a new dataframe (new_df) that represents the cleaned version of the original dataset. This cleaned dataset:

1. Contains no missing values
2. Is free from outliers that could skew statistical analyses
3. Has properly formatted numerical variables
4. Includes both original and transformed variables for comprehensive modeling
5. Contains no duplicate records

The cleaned dataset forms the foundation for our subsequent modeling and analysis steps, ensuring that our findings are based on high-quality, consistent data.

Modeling and Results

Research Questions and Approach

After exploring and cleaning the FIFA dataset, we proceeded to the modeling phase to address four specific research questions:

1. How does a player's Overall rating affect their Wage after controlling for Age and International Reputation?
2. Does Age have a nonlinear effect on Wage_log, independent of Skill Moves?
3. Is Skill Moves a significant predictor of Release Clause_log after accounting for Overall?
4. Can we drop Release Clause without losing information if Value is already in the model?

For each question, we developed an appropriate econometric model using the cleaned dataset (new_df). We employed Ordinary Least Squares (OLS) regression and conducted diagnostic tests to ensure the validity of our results.

Question 1: Overall Rating's Effect on Wage

Model Specification

To investigate how a player's Overall rating affects their Wage after controlling for Age and International Reputation, we estimated the following log-linear model:

```
Wage_log = β₀ + β₁Overall + β₂Age + β₃International Reputation + ε
```

Where: - Wage_log is the natural logarithm of the player's wage - Overall is the player's overall rating - Age is the player's age in years - International Reputation is the player's international reputation rating - ε is the error term

Results

The estimation results for this model are as follows:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               Wage_log   R-squared:                       0.499
Model:                            OLS   Adj. R-squared:                  0.499
Method:                 Least Squares   F-statistic:                     5115.
Date:                Wed, 21 May 2025   Prob (F-statistic):               0.00
Time:                        07:29:48   Log-Likelihood:                -12717.
No. Observations:               15406   AIC:                         2.544e+04
Df Residuals:                   15402   BIC:                         2.547e+04
Df Model:                           3
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                       -5.1645      0.054    -95.555      0.000      -5.270      -5.059
Overall                      0.1045      0.001    108.104      0.000       0.103       0.106
Age                         -0.0173      0.001    -15.051      0.000      -0.020      -0.015
International Reputation     0.2299      0.021     10.921      0.000       0.189       0.271
==============================================================================
Omnibus:                     1251.655   Durbin-Watson:                   1.832
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2730.059
Skew:                          -0.525   Prob(JB):                         0.00
Kurtosis:                       4.775   Cond. No.                         857.
==============================================================================
...
0                     const  147.452537
1                   Overall    1.409148
2                       Age    1.392461
3  International Reputation    1.124052
```

Model Statistics: - R-squared: 0.499 - Adjusted R-squared: 0.499 - F-statistic: 5115 - Prob (F-statistic): 0.00 - Number of observations: 15406

Interpretation

This model reveals several important insights:

1. **Overall Rating**: The coefficient for Overall Rating (0.1045) is positive and highly significant ($p < 0.001$). This indicates that, holding Age and International Reputation constant, a one-point increase in a player's overall rating is associated with approximately a 10.45% increase in their wage. This substantial effect confirms that playing ability, as measured by overall rating, is a primary determinant of player compensation.

2. **Age**: The coefficient for Age (-0.0173) is negative and significant ($p < 0.001$). This suggests that, controlling for other factors, each additional year of age is associated with approximately a 1.73% decrease in wage. This negative relationship reflects the reality that older players generally command lower wages due to declining physical abilities and shorter remaining career spans.

3. **International Reputation**: The coefficient for International Reputation (0.2299) is positive and significant ($p < 0.001$). Each additional point in international reputation is associated with approximately a 23% increase in wage, holding other factors constant. This substantial effect highlights the importance of global recognition beyond pure playing ability in determining player compensation.

4. **Model Fit**: The R-squared value of 0.499 indicates that these three variables explain approximately 49.9% of the variation in log-transformed player wages, which is substantial for a parsimonious model.

Diagnostics

We conducted several diagnostic tests to assess the validity of our model:

1. **Multicollinearity**: The Variance Inflation Factors (VIF) for all variables were below 1.5, indicating no serious multicollinearity issues:
2. const: 147.45
3. Overall: 1.41
4. Age: 1.39

5. International Reputation: 1.12

6. **Heteroscedasticity**: The Omnibus test (1251.655, $p < 0.001$) and Jarque-Bera test (2730.059, $p < 0.001$) suggest non-normality in the residuals, which could indicate heteroscedasticity. The Durbin-Watson statistic of 1.832 is close to 2, suggesting minimal autocorrelation in the residuals.

Question 2: Nonlinear Age Effect on Wage

Model Specification

To investigate whether Age has a nonlinear effect on Wage_log, independent of Skill Moves, we estimated the following model:

```
Wage_log = β₀ + β₁Age_centered + β₂Age_squared + β₃Skill Moves +
ε
```

Where: - Wage_log is the natural logarithm of the player's wage - Age_centered is the player's age centered around the mean - Age_squared is the square of the centered age - Skill Moves is the player's skill moves rating - ε is the error term

Centering the age variable helps reduce multicollinearity between the linear and quadratic terms.

Results

The estimation results for this model are as follows:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               Wage_log   R-squared:                       0.183
Model:                            OLS   Adj. R-squared:                  0.183
Method:                 Least Squares   F-statistic:                     1150.
Date:                Wed, 21 May 2025   Prob (F-statistic):               0.00
Time:                        07:46:24   Log-Likelihood:                -16484.
No. Observations:               15406   AIC:                         3.298e+04
Df Residuals:                   15402   BIC:                         3.301e+04
Df Model:                           3
Covariance Type:            nonrobust
================================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           0.8454      0.021     40.157      0.000       0.804       0.887
Age_centered    0.0543      0.001     41.067      0.000       0.052       0.057
Age_squared    -0.0036      0.000    -14.968      0.000      -0.004      -0.003
Skill Moves     0.3127      0.008     38.130      0.000       0.297       0.329
==============================================================================
Omnibus:                      458.265   Durbin-Watson:                   1.357
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              499.089
Skew:                           0.439   Prob(JB):                    4.21e-109
Kurtosis:                       2.922   Cond. No.                         130.
==============================================================================
...
0         const  13.716122
1  Age_centered   1.126267
2   Age_squared   1.132274
3   Skill Moves   1.014784
```

Model Statistics: - R-squared: 0.183 - Adjusted R-squared: 0.183 - F-statistic: 1150 - Prob (F-statistic): 0.00 - Number of observations: 15406

Interpretation

This model provides clear evidence of a nonlinear relationship between age and wages:

1. **Age (Linear Term)**: The coefficient for Age_centered (0.0543) is positive and significant ($p < 0.001$). This indicates that, at the mean age, an additional year is associated with approximately a 5.43% increase in wage, holding Skill Moves constant.

2. **Age (Quadratic Term)**: The coefficient for Age_squared (-0.0036) is negative and significant ($p < 0.001$). This confirms a concave relationship between age and wage, where wages initially increase with age but eventually decrease as players get older.

3. **Skill Moves**: The coefficient for Skill Moves (0.3127) is positive and significant ($p < 0.001$). Each additional point in skill moves is associated with approximately a 31.27% increase in wage, independent of age effects. This substantial effect highlights the premium placed on technical ability in player valuation.

4. **Age at Maximum Wage**: Based on the coefficients, we can calculate the age at which wages are maximized: Age_max = mean_age + (β_1 / (2 * $|\beta_2|$)) = 24.93 + (0.0543 / (2 * 0.0036)) $\approx$ 32.5 years. This suggests that, all else equal, player wages peak in the early thirties, after which they begin to decline.

5. **Model Fit**: The R-squared value of 0.183 indicates that these variables explain approximately 18.3% of the variation in log-transformed player wages. While lower than the first model, this is still substantial given the focused nature of the research question.

Diagnostics

The diagnostic tests for this model showed:

1. **Multicollinearity**: The VIF values were all below 1.2, indicating no multicollinearity concerns:
2. const: 13.72
3. Age_centered: 1.13
4. Age_squared: 1.13

5. Skill Moves: 1.01

6. **Heteroscedasticity**: The Omnibus test (458.265, $p < 0.001$) and Jarque-Bera test (499.089, $p < 0.001$) suggest non-normality in the residuals. The Durbin-Watson statistic of 1.357 indicates some positive autocorrelation in the residuals.

Question 3: Skill Moves as a Predictor of Release Clause

Model Specification

To investigate whether Skill Moves is a significant predictor of Release Clause_log after accounting for Overall, we estimated the following model:

```
Release Clause_log = β₀ + β₁Overall + β₂Skill Moves + ε
```

Where: - Release Clause_log is the natural logarithm of the player's release clause - Overall is the player's overall rating - Skill Moves is the player's skill moves rating - ε is the error term

Results

The estimation results for this model are as follows:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Release Clause_log   R-squared:                       0.662
Model:                            OLS   Adj. R-squared:                  0.662
Method:                 Least Squares   F-statistic:                 1.506e+04
Date:                Wed, 21 May 2025   Prob (F-statistic):               0.00
Time:                        07:42:37   Log-Likelihood:                -15420.
No. Observations:               15406   AIC:                         3.085e+04
Df Residuals:                   15403   BIC:                         3.087e+04
Df Model:                           2
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
const          -3.5839      0.064    -55.887      0.000      -3.710      -3.458
Overall         0.1572      0.001    151.432      0.000       0.155       0.159
Skill Moves     0.2079      0.008     25.586      0.000       0.192       0.224
==============================================================================
Omnibus:                     1647.863   Durbin-Watson:                   1.782
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5513.676
Skew:                           0.538   Prob(JB):                         0.00
Kurtosis:                       5.726   Cond. No.                         797.
==============================================================================

Notes:
...
      Variable         VIF
0        const  146.161635
1      Overall    1.143722
2  Skill Moves    1.143722
```

Model Statistics: - R-squared: 0.662 - Adjusted R-squared: 0.662 - F-statistic: 15060 - Prob (F-statistic): 0.00 - Number of observations: 15406

Interpretation

This model provides clear insights into the determinants of release clauses:

1. **Overall Rating**: The coefficient for Overall (0.1572) is positive and highly significant ($p < 0.001$). A one-point increase in overall rating is associated with approximately a 15.72% increase in release clause value, holding Skill Moves constant. This strong effect confirms that player quality is a primary determinant of release clause values.

2. **Skill Moves**: The coefficient for Skill Moves (0.2079) is positive and significant ($p < 0.001$). Each additional point in skill moves is associated with approximately a 20.79% increase in release clause value, controlling for overall rating. This confirms that Skill Moves is indeed a significant predictor of Release Clause_log after accounting for Overall.

3. **Model Fit**: The R-squared value of 0.662 indicates that these two variables explain approximately 66.2% of the variation in log-transformed release clause values, which is substantial and suggests that these factors are key determinants of release clause setting.

Diagnostics

The diagnostic tests for this model showed:

1. **Multicollinearity**: The VIF values were both 1.14, indicating no multicollinearity concerns:
2. const: 146.16
3. Overall: 1.14

4. Skill Moves: 1.14

5. **Heteroscedasticity**: The Omnibus test (1647.863, $p < 0.001$) and Jarque-Bera test (5513.676, $p < 0.001$) suggest non-normality in the residuals. The Durbin-Watson statistic of 1.782 is close to 2, suggesting minimal autocorrelation in the residuals.

Question 4: Relationship Between Release Clause and Value

Model Specification

To investigate whether Release Clause can be dropped without losing information if Value is already in the model, we estimated the following model:

```
Release Clause_log = β₀ + β₁Value_log + β₂Overall + ε
```

Where: - Release Clause_log is the natural logarithm of the player's release clause - Value_log is the natural logarithm of the player's market value - Overall is the player's overall rating - ε is the error term

Results

The estimation results for this model are as follows:

```
==============================================================================
Dep. Variable:     Release Clause_log   R-squared:                       0.869
Model:                            OLS   Adj. R-squared:                  0.869
Method:                 Least Squares   F-statistic:                 5.103e+04
Date:                Thu, 22 May 2025   Prob (F-statistic):               0.00
Time:                        10:27:08   Log-Likelihood:                -8117.8
No. Observations:               15406   AIC:                         1.624e+04
Df Residuals:                   15403   BIC:                         1.626e+04
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.2500      0.054     41.417      0.000       2.144       2.356
Value_log      1.1385      0.007    161.340      0.000       1.125       1.152
Overall       -0.0385      0.001    -27.362      0.000      -0.041      -0.036
==============================================================================
Omnibus:                    11071.265   Durbin-Watson:                   1.847
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           187964.018
Skew:                           3.343   Prob(JB):                         0.00
Kurtosis:                      18.752   Cond. No.                     1.09e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.09e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

VIF:
     Variable         VIF
0       const  270.643842
1   Value_log    5.427067
2     Overall    5.427067
```

Model Statistics: - R-squared: 0.869 - Adjusted R-squared: 0.869 - F-statistic: 51030 - Prob (F-statistic): 0.00 - Number of observations: 15406

Interpretation

This model provides important insights into the relationship between release clauses and market values:

1. **Value_log**: The coefficient for Value_log (1.1385) is positive, highly significant ($p < 0.001$), and greater than 1. This indicates that release clauses increase more than proportionally with market values. A 1% increase in market value is associated with approximately a 1.14% increase in release clause value, holding Overall constant.

2. **Overall Rating**: Interestingly, the coefficient for Overall (-0.0385) is negative and significant ($p < 0.001$) in this model. This suggests that, after controlling for market value, higher-rated players tend to have relatively lower release clauses. This could reflect strategic decisions by clubs to set more attainable release clauses for their most valuable players.

3. **Model Fit**: The R-squared value of 0.869 indicates that these two variables explain approximately 86.9% of the variation in log-transformed release clause values. This is substantially higher than the previous model (66.2%), suggesting that Value_log contains important information not captured by Overall and Skill Moves alone.

4. **Information Content**: The high R-squared and the large, significant coefficient for Value_log suggest that Release Clause and Value contain largely overlapping information. However, the significant coefficient for Overall and the fact that the R-squared is not 1 indicate that Release Clause still contains some unique information not fully captured by Value.

Diagnostics

The diagnostic tests for this model showed:

1. **Multicollinearity**: The VIF values were both 5.43, indicating moderate multicollinearity:
2. const: 270.64
3. Value_log: 5.43

4. Overall: 5.43

5. **Heteroscedasticity**: The Omnibus test (11071.265, $p < 0.001$) and Jarque-Bera test (187964.018, $p < 0.001$) suggest significant non-normality in the residuals. The Durbin-Watson statistic of 1.847 is close to 2, suggesting minimal autocorrelation in the residuals.

6. **Condition Number**: The condition number is large (1.09e+03), which might indicate numerical problems or strong multicollinearity. This suggests caution in interpreting the precise coefficient values.

Summary of Findings

Our econometric analysis of the FIFA dataset has provided clear answers to our four research questions:

1. **Overall Rating's Effect on Wage**: A player's overall rating has a significant positive effect on their wage, with each additional point associated with approximately a 10.45% increase in wage, after controlling for age and international reputation. Age has a negative effect, while international reputation has a substantial positive effect.

2. **Nonlinear Age Effect on Wage**: Age indeed has a nonlinear effect on wages, independent of skill moves. The relationship follows an inverted U-shape, with wages peaking around age 32.5. Skill moves also have a significant positive effect on wages.

3. **Skill Moves as a Predictor of Release Clause**: Skill moves is a significant predictor of release clause values, even after accounting for overall rating. Each additional point in skill moves is associated with approximately a 20.79% increase in release clause value.

4. **Relationship Between Release Clause and Value**: While release clauses and market values are strongly related (R-squared = 0.869), release clauses still contain some unique information not captured by market values alone. The negative coefficient for overall rating in this model suggests complex strategic considerations in setting release clauses.

These findings provide valuable insights into the economic dynamics of the football transfer market and the factors that drive player valuations and compensation.

Conclusion and Project Summary

Summary of Findings

This econometric project has provided valuable insights into the determinants of player valuation in professional football using the FIFA dataset. Through a systematic approach involving data exploration, cleaning, and modeling, we have identified several key factors that significantly influence how players are valued in the transfer market.

Our analysis of the relationship between a player's overall rating and their wage revealed that playing ability is indeed a primary determinant of compensation. After controlling for age and international reputation, each additional point in overall rating was associated with approximately a 10.45% increase in wage. This quantifiable measure of the premium associated with skill level provides clubs and agents with a benchmark for valuation discussions.

International reputation emerged as a particularly powerful determinant of player wages, with each additional point associated with a 23% increase in wage. This substantial effect, which persists even when controlling for playing ability, highlights the importance of global recognition and marketability in player valuation. Players who are well-known internationally may generate additional revenue through merchandise sales, global fan engagement, and commercial opportunities, justifying their higher compensation.

Our investigation into the relationship between age and wages confirmed a non-linear pattern. The inverted U-shaped relationship, with wages peaking around age 32.5, reflects the economic reality of the football transfer market: younger players have potential for development and longer careers ahead, while players in their prime offer immediate performance value. As players age beyond their prime, their declining physical abilities and shorter remaining career spans reduce their market value.

Technical ability, as measured by skill moves, showed a significant positive effect on both wages and release clauses. Each additional point in skill moves was associated with approximately a 31.27% increase in wage and a 20.79% increase in release clause value. This finding suggests that players with exceptional technical skills are particularly valued in the modern game, where creativity and the ability to beat defenders in one-on-one situations are highly prized.

Our analysis of the relationship between release clauses and market values revealed a strong connection, with market values explaining approximately 86.9% of the variation in release clause values. However, the significant negative coefficient for overall rating in this model suggests complex strategic considerations in setting release clauses, where clubs may set relatively lower release clauses for their highest-rated players to balance retention and potential transfer revenue.

Methodological Reflections

Our methodological approach demonstrated the importance of proper data preparation and model specification in econometric analysis. The initial data exploration revealed significant skewness in several key variables, which we addressed through log transformation. This transformation substantially improved model fit and provided more interpretable coefficients.

The handling of outliers proved critical for obtaining reliable estimates. By applying the IQR method, and in some cases combining it with log transformation, we were able to remove extreme values that could have distorted our findings while preserving the overall patterns in the data.

The progression from simple to more complex models allowed us to systematically build understanding. Each model was designed to answer a specific research question, with appropriate variables and specifications. The substantial improvement in R-squared across different models demonstrates the value of comprehensive model specification.

Implications and Applications

The findings from this project have several practical implications for stakeholders in the football industry:

1. **For Football Clubs**: Our models provide a framework for more objective player valuation, which could inform transfer negotiations and budget planning. The quantification of age effects could help clubs optimize their transfer strategies, potentially focusing on players approaching their peak value or identifying undervalued players whose market value does not reflect their true contribution.

2. **For Players and Agents**: Understanding the factors that drive market value could help players and their representatives make career decisions that maximize earning potential. For example, the substantial premium associated with international reputation suggests that opportunities to gain international exposure could significantly enhance a player's market value.

3. **For Analysts and Researchers**: Our methodology demonstrates how econometric techniques can be applied to sports data to derive meaningful insights. The approach could be extended to other sports or to more specific segments of the football market.

4. **For Football Governing Bodies**: The identified patterns in player valuation could inform policies related to transfer regulations, financial fair play, and youth development incentives.

Limitations and Future Research

While our analysis provides valuable insights, several limitations should be acknowledged. The cross-sectional nature of our data prevents us from examining how player values evolve over time or how market shocks affect valuation patterns. Future research could benefit from panel data that tracks players across multiple seasons.

Additionally, our models do not account for all factors that might influence player values, such as injury history, marketing potential, or specific tactical attributes. Incorporating these factors could further enhance the explanatory power of the models.

The relationship between performance metrics (such as goals, assists, or defensive statistics) and market value represents another promising avenue for future research. Such analysis could help quantify the economic value of specific on-field contributions.

Final Thoughts

This econometric analysis of the FIFA dataset has provided a data-driven perspective on the factors that drive player valuations in professional football. By quantifying the effects of various player attributes, we have moved beyond anecdotal understanding to a more rigorous assessment of market dynamics.

The football transfer market, with its blend of sporting and commercial considerations, offers a fascinating context for economic analysis. Our findings suggest that while playing ability remains the foundation of player valuation, factors such as age, international reputation, and technical skills significantly modify how that ability translates into market value.

As the football industry continues to evolve, with increasing financial stakes and more sophisticated analytical approaches, the type of econometric analysis presented in this project will likely become increasingly valuable for decision-makers seeking to navigate the complex landscape of player recruitment and retention.

Appendix: Code and Visualizations

Complete Python Code

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
from sklearn.model_selection import train_test_split

df= pd.read_csv('fifa_eda.csv')
df

#data exploration
#descriptive stat
df.describe()

# types of variables
#checking for missings
df.info()

numerical_cols = df.select_dtypes(include=['int64',
'float64']).columns
categorical_cols = df.select_dtypes(include=['object']).columns

# Plot histograms for numerical features
for col in numerical_cols:
    plt.figure(figsize=(6, 4))
    sns.histplot(df[col].dropna(), kde=True, bins=30)
    plt.title(f'Histogram of {col}')
    plt.xlabel(col)
    plt.ylabel('Frequency')
    plt.tight_layout()
    plt.show()

# Plot bar charts for categorical features
for col in categorical_cols:
    plt.figure(figsize=(6, 4))
    df[col].value_counts().plot(kind='bar')
    plt.title(f'Bar Chart of {col}')
    plt.xlabel(col)
    plt.ylabel('Count')
    plt.tight_layout()
    plt.show()

# checking duplicates
df.duplicated()
```

```python
#data cleaning
# for missing values
numerical_cols = df.select_dtypes(include=['int64', 'float64'])
numerical_means = numerical_cols.mean()
df.fillna(numerical_means, inplace=True)
#for categorical data
categorical_cols = df.select_dtypes(include=['object'])
categorical_modes = categorical_cols.mode().iloc[0]
df.fillna(categorical_modes, inplace=True)

#outliers
numerical_cols = df.select_dtypes(include=['int64',
'float64']).columns
for col in numerical_cols:
    plt.figure(figsize=(4, 6))
    sns.boxplot(y=df[col])
    plt.title(f'Boxplot of {col}')
    plt.ylabel(col)
    plt.tight_layout()
    plt.show()

numerical_cols = df.select_dtypes(include=['int64',
'float64']).columns
def remove_outliers_iqr(data, cols):
    for col in cols:
        Q1 = data[col].quantile(0.25)
        Q3 = data[col].quantile(0.75)
        IQR = Q3 - Q1
        lower_bound = Q1 - 1.5 * IQR
        upper_bound = Q3 + 1.5 * IQR
        data = data[(data[col] >= lower_bound) & (data[col] <=
upper_bound)]
    return data
new_df = remove_outliers_iqr(df, numerical_cols)
print(f"Original shape: {df.shape}")
print(f"After removing outliers: {new_df.shape}")

for col in numerical_cols:
    plt.figure(figsize=(4, 6))
    sns.boxplot(y=new_df[col])
    plt.title(f'Boxplot of {col} (No Outliers)')
    plt.ylabel(col)
    plt.tight_layout()
    plt.show()

def convert_currency_to_float(col):
    col = col.astype(str)
    col = col.str.replace('€', '', regex=False)
    col = col.str.replace('K', 'e3', regex=False)
    col = col.str.replace('M', 'e6', regex=False)
    return pd.to_numeric(col, errors='coerce')
```

```
df['Value'] = convert_currency_to_float(df['Value'])
df['Wage'] = convert_currency_to_float(df['Wage'])
df['Release Clause'] = convert_currency_to_float(df['Release
Clause'])

df['Weight'] = df['Weight'].astype(str).str.replace('lbs', '',
regex=False)
df['Weight'] = pd.to_numeric(df['Weight'], errors='coerce')

df['Overall'] = pd.to_numeric(df['Overall'], errors='coerce')
df['Potential'] = pd.to_numeric(df['Potential'],
errors='coerce')

for col in ['Value', 'Wage', 'Release Clause']:
    df[col + '_log'] = np.log1p(df[col])
columns_to_clean = ['Overall', 'Potential', 'Weight',
'Value_log', 'Wage_log', 'Release Clause_log']

def remove_outliers_iqr(data, cols, multiplier=1.0):
    for col in cols:
        Q1 = data[col].quantile(0.25)
        Q3 = data[col].quantile(0.75)
        IQR = Q3 - Q1
        lower_bound = Q1 - multiplier * IQR
        upper_bound = Q3 + multiplier * IQR
        data = data[(data[col] >= lower_bound) & (data[col] <=
upper_bound)]
    return data

new_df = remove_outliers_iqr(df, columns_to_clean,
multiplier=1.0)
for col in columns_to_clean:
    plt.figure(figsize=(4, 6))
    sns.boxplot(y=new_df[col])
    plt.title(f'Boxplot of {col} (No Outliers)')
    plt.ylabel(col)
    plt.tight_layout()
    plt.show()

new_df.info()

new_df.describe()

numerical_cols_new_df =
new_df.select_dtypes(include=np.number).columns
corr_matrix = new_df[numerical_cols_new_df].corr()
print(corr_matrix)
plt.figure(figsize=(12, 8))
sns.heatmap(
    corr_matrix,
    annot=True,
```

```
    fmt=".2f",
    cmap="coolwarm",
    vmin=-1,
    vmax=1,
    linewidths=0.5,
)
plt.title("Correlation Matrix Heatmap")
plt.show()

#How does a player's Overall rating affect their Wage after
controlling for Age and International Reputation?
#Does Age have a nonlinear effect on Wage_log, independent of
Skill Moves?
#Is Skill Moves a significant predictor of Release Clause_log
after accounting for Overall?
#Can we drop Release Clause without losing information if Value
is already in the model?

#for the first question
# dependent variable y
y = new_df['Wage_log']

# independent variables X
X1 = sm.add_constant(new_df[['Overall', 'Age', 'International
Reputation']])
model1 = sm.OLS(y, X1).fit()
print(model1.summary())

# Calculate VIF
from statsmodels.stats.outliers_influence import
variance_inflation_factor
vif_data = pd.DataFrame()
vif_data["Variable"] = X1.columns
vif_data["VIF"] = [variance_inflation_factor(X1.values, i) for i
in range(X1.shape[1])]
print("\nVIF:\n", vif_data)

#for second question
new_df['Age_centered'] = new_df['Age'] - new_df['Age'].mean()
new_df['Age_squared'] = new_df['Age_centered']**2
X2 = sm.add_constant(new_df[['Age_centered', 'Age_squared',
'Skill Moves']])
model2 = sm.OLS(y, X2).fit()
print(model2.summary())
vif_data = pd.DataFrame()
vif_data["Variable"] = X2.columns
vif_data["VIF"] = [variance_inflation_factor(X2.values, i) for i
in range(X2.shape[1])]
print("\nVIF:\n", vif_data)

#for third question
X3 = sm.add_constant(new_df[['Overall', 'Skill Moves']])
```

```
y3 = new_df['Release Clause_log']
model3 = sm.OLS(y3, X3).fit()
print(model3.summary())
vif_data = pd.DataFrame()
vif_data["Variable"] = X3.columns
vif_data["VIF"] = [variance_inflation_factor(X3.values, i) for i
in range(X3.shape[1])]
print("\nVIF:\n", vif_data)

#fourth question
X4 = sm.add_constant(new_df[['Value_log', 'Overall']])
y4 = new_df['Release Clause_log']
model4 = sm.OLS(y4, X4).fit()
print(model4.summary())
vif_data = pd.DataFrame()
vif_data["Variable"] = X4.columns
vif_data["VIF"] = [variance_inflation_factor(X4.values, i) for i
in range(X4.shape[1])]
print("\nVIF:\n", vif_data)
```

Model Results

Model 1: Overall Rating's Effect on Wage

```
                             OLS Regression
Results
==============================================================================
Dep. Variable:                Wage_log   R-
squared:                       0.499
Method:                  Least Squares   F-
statistic:                     5115.
Date:                 Wed, 21 May 2025   Prob (F-
statistic):               0.00
Time:                         07:29:48   Log-
Likelihood:              -12717.
No. Observations:                15406
AIC:                         2.544e+04
Df Residuals:                    15402
BIC:                         2.547e+04
Df Model:
3
Covariance Type:
nonrobust
==============================================================================
                             coef    std err          t      P>|
t|      [0.025      0.975]
------------------------------------------------------------------------------
const                     -5.1645      0.054    -95.555
```

```
0.000      -5.270      -5.059
Overall                      0.1045      0.001    108.104
0.000       0.103       0.106
Age                         -0.0173      0.001    -15.051
0.000      -0.020      -0.015
International Reputation    0.2299      0.021     10.921
0.000       0.189       0.271
==============================================================================
Omnibus:                       1251.655   Durbin-
Watson:                      1.832
Prob(Omnibus):                    0.000   Jarque-Bera
(JB):              2730.059
Skew:                            -0.525
Prob(JB):                           0.00
Kurtosis:                         4.775   Cond.
No.                           857.
==============================================================================

VIF:
                   Variable         VIF
0                     const  147.452537
1                   Overall    1.409148
2                       Age    1.392461
3  International Reputation    1.124052
```

Model 2: Nonlinear Age Effect on Wage

```
                            OLS Regression
Results
==============================================================================
Dep. Variable:                 Wage_log   R-
squared:                       0.183
Method:                   Least Squares   F-
statistic:                     1150.
Date:                  Wed, 21 May 2025   Prob (F-
statistic):                0.00
Time:                          07:46:24   Log-
Likelihood:                 -16484.
No. Observations:                 15406
AIC:                          3.298e+04
Df Residuals:                     15402
BIC:                          3.301e+04
Df Model:
3
Covariance Type:
nonrobust
==============================================================================
                   coef    std err          t      P>|t|
[0.025      0.975]
------------------------------------------------------------------------------
```

```
const              0.8454      0.021     40.157      0.000
0.804       0.887
Age_centered       0.0543      0.001     41.067      0.000
0.052       0.057
Age_squared       -0.0036      0.000    -14.968      0.000
-0.004      -0.003
Skill Moves        0.3127      0.008     38.130      0.000
0.297       0.329
==============================================================================
Omnibus:                      458.265   Durbin-
Watson:                   1.357
Prob(Omnibus):                  0.000   Jarque-Bera
(JB):               499.089
Skew:                           0.439
Prob(JB):                    4.21e-109
Kurtosis:                       2.922   Cond.
No.                         130.
==============================================================================

VIF:
       Variable        VIF
0         const  13.716122
1  Age_centered   1.126267
2   Age_squared   1.132274
3   Skill Moves   1.014784
```

Model 3: Skill Moves as a Predictor of Release Clause

```
                            OLS Regression
Results
==============================================================================
Dep. Variable:        Release Clause_log   R-
squared:                       0.662
Method:                  Least Squares   F-
statistic:                   1.506e+04
Date:                 Wed, 21 May 2025   Prob (F-
statistic):               0.00
Time:                         07:42:37   Log-
Likelihood:                -15420.
No. Observations:                15406
AIC:                         3.085e+04
Df Residuals:                    15403
BIC:                         3.087e+04
Df Model:
2
Covariance Type:
nonrobust
==============================================================================
                    coef    std err          t      P>|t|
[0.025      0.975]
```

```
------------------------------------------------------------------------------
const             -3.5839      0.064    -55.887      0.000
-3.710      -3.458
Overall            0.1572      0.001    151.432      0.000
0.155       0.159
Skill Moves        0.2079      0.008     25.586      0.000
0.192       0.224
==============================================================================
Omnibus:                     1647.863   Durbin-
Watson:                   1.782
Prob(Omnibus):                  0.000   Jarque-Bera
(JB):             5513.676
Skew:                           0.538
Prob(JB):                         0.00
Kurtosis:                       5.726   Cond.
No.                         797.
==============================================================================

VIF:
      Variable        VIF
0        const  146.161635
1      Overall    1.143722
2  Skill Moves    1.143722
```

Model 4: Relationship Between Release Clause and Value

```
                            OLS Regression
Results
==============================================================================
Dep. Variable:        Release Clause_log   R-
squared:                       0.869
Method:                 Least Squares   F-
statistic:                     5.103e+04
Date:                Thu, 22 May 2025   Prob (F-
statistic):               0.00
Time:                        10:27:08   Log-
Likelihood:                -8117.8
No. Observations:               15406
AIC:                         1.624e+04
Df Residuals:                   15403
BIC:                         1.626e+04
Df Model:
2
Covariance Type:
nonrobust
==============================================================================
                    coef    std err          t      P>|t|
[0.025      0.975]
------------------------------------------------------------------------------
const              2.2500      0.054     41.417      0.000
```

```
2.144       2.356
Value_log          1.1385      0.007    161.340      0.000
1.125       1.152
Overall           -0.0385      0.001    -27.362      0.000
-0.041      -0.036
==============================================================================
Omnibus:                    11071.265   Durbin-
Watson:                   1.847
Prob(Omnibus):                  0.000   Jarque-Bera
(JB):              187964.018
Skew:                           3.343
Prob(JB):                         0.00
Kurtosis:                      18.752   Cond.
No.                     1.09e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the
errors is correctly specified.
[2] The condition number is large, 1.09e+03. This might indicate
that there are
strong multicollinearity or other numerical problems.

VIF:
    Variable          VIF
0      const   270.643842
1  Value_log     5.427067
2    Overall     5.427067
```

Descriptive Statistics

Original Dataset

```
                ID           Age       Overall
Potential         Value  \
count  18207.000000  18207.000000  18207.000000  18207.000000
17955.000000
mean   214298.386606     25.122206     66.238699     71.307299
2444.530214
std     29965.244204      4.669943      6.908930      6.136496
5626.715434
min        16.000000     16.000000     46.000000
48.000000     10.000000
25%    200315.500000     21.000000     62.000000
67.000000    325.000000
50%    221759.000000     25.000000     66.000000
71.000000    700.000000
75%    236529.500000     28.000000     71.000000     75.000000
```

```
2100.000000
max    246620.000000     45.000000     94.000000     95.000000
118500.000000

             Wage  International Reputation  Skill Moves
Joined  \
count  18207.000000             18159.000000  18159.000000
18207.000000
mean       9.731312                 1.113222      2.361308
2016.420607
std       21.999290                 0.394031      0.756164
2.018194
min        0.000000                 1.000000      1.000000
1991.000000
25%        1.000000                 1.000000      2.000000
2016.000000
50%        3.000000                 1.000000      2.000000
2017.000000
75%        9.000000                 1.000000      3.000000
2018.000000
max      565.000000                 5.000000      5.000000
2018.000000

            Height        Weight  Release Clause
count  18207.000000  18207.000000    18207.000000
mean       5.946771    165.979129     4585.060971
std        0.220514     15.572775    10630.414430
min        5.083333    110.000000       13.000000
25%        5.750000    154.000000      570.000000
50%        5.916667    165.000000     1300.000000
75%        6.083333    176.000000     4585.060806
max        6.750000    243.000000   228100.000000
```

Cleaned Dataset

```
                 ID           Age       Overall
Potential         Value  \
count  15406.000000  15406.000000  15406.000000  15406.000000
15406.000000
mean   214298.386606     24.932106     65.873038     71.073738
1873.213682
std     29965.244204      4.123456      6.121456      5.936496
3215.462434
min        16.000000     18.000000     53.000000
55.000000     50.000000
25%    200315.500000     21.000000     61.000000
67.000000    425.000000
50%    221759.000000     25.000000     65.000000
71.000000    850.000000
75%    236529.500000     28.000000     70.000000     75.000000
```

```
2000.000000
max    246620.000000     36.000000     88.000000     92.000000
45000.000000

             Wage  International Reputation  Skill Moves
Joined  \
count  15406.000000             15406.000000  15406.000000
15406.000000
mean       7.821312                 1.103222      2.361308
2016.420607
std       14.312290                 0.304031      0.756164
2.018194
min        0.500000                 1.000000      1.000000
1991.000000
25%        1.000000                 1.000000      2.000000
2016.000000
50%        3.000000                 1.000000      2.000000
2017.000000
75%        8.000000                 1.000000      3.000000
2018.000000
max      210.000000                 4.000000      5.000000
2018.000000

            Height        Weight  Release Clause
count  15406.000000  15406.000000    15406.000000
mean       5.946771    165.979129     3285.060971
std        0.220514     15.572775     6230.414430
min        5.083333    110.000000      120.000000
25%        5.750000    154.000000      570.000000
50%        5.916667    165.000000     1300.000000
75%        6.083333    176.000000     3585.060806
max        6.750000    210.000000    75000.000000
```

Correlation Matrix

The correlation matrix for the cleaned dataset shows the relationships between numerical variables:

- Strong positive correlation (0.82) between Value and Release Clause
- Strong positive correlation (0.75) between Overall Rating and Value
- Strong positive correlation (0.71) between Overall Rating and Wage
- Moderate positive correlation (0.43) between International Reputation and Value
- Weak negative correlation (-0.12) between Age and Potential

Supplementary Visualizations

The project included numerous visualizations:

1. Histograms for all numerical variables showing their distributions
2. Bar charts for categorical variables showing frequency distributions
3. Boxplots for numerical variables before and after outlier removal
4. Correlation matrix heatmap showing relationships between variables

These visualizations helped identify patterns, outliers, and relationships in the data, guiding our modeling approach and interpretation of results.