Data Cleaning and Exploratory Data Analysis on the players_21.csv Dataset

Abstract

This report presents a comprehensive analysis of the FIFA players_21 dataset. The study covers
data cleaning procedures—including handling missing values, duplicate record removal, outlier
detection and treatment, and categorical value standardization—as well as extensive exploratory data analysis (EDA). Univariate, bivariate, and multivariate analyses are performed
to understand the distribution of player attributes, relationships between performance metrics and market-related features, and clustering effects within different player groups. The insights
gathered provide a foundation for further predictive modeling and advanced sports analytics.

1. Introduction

The players_21.csv dataset is extracted from sofifa and contains detailed information on football
players such as age, height, weight, overall rating, potential, market value, wage, playing positions, and a wide range of skill attributes. In addition to numerous numerical and categorical
variables, the dataset also includes missing values, duplicates, and outliers that must be addressed before meaningful statistical or machine learning analyses can be performed.
The primary objectives of this analysis are to:
Prepare the dataset through extensive cleaning.
Explore the underlying distributions and relationships among various attributes.
Draw actionable insights that could support player valuation models and further research in sports analytics.

2. Data Cleaning

2.1 Loading and Initial Inspection

The dataset is first loaded into the environment. An inspection of the first few rows and overall
structure reveals multiple numerical variables (such as age, height_cm, weight_kg, overall, potential, value_eur, wage_eur, pace, shooting, passing, dribbling, defending, and physic) along
with categorical variables (such as short_name, nationality, club_name, league_name, player_positions, and preferred_foot). The initial investigation highlights issues such as missing
values and duplicate entries that must be handled.

2.2 Handling Missing Values

For numerical features (e.g., age, overall rating, wage) that may be affected by outliers, the missing values were imputed using the median. For categorical features (like club_name and player_positions), the mode was used as the imputation method. This approach minimizes distortions from extreme values while preserving the central tendency of variables.

2.3 Removing Duplicate Records

The dataset was checked for duplicate rows. Any found duplicates were removed so that each
player record is unique. This step is essential to prevent bias in the summary statistics and visualizations.

2.4 Outlier Detection and Treatment

Outliers were identified using the Interquartile Range (IQR) method, especially on highly skewed
numerical columns such as wage_eur, value_eur, and overall ratings. Outlying values were then
capped at the lower and upper bounds (Q1 – 1.5·IQR and Q3 + 1.5·IQR) to reduce distortion in
the analysis while maintaining important trends.

2.5 Standardizing Categorical Values

To ensure consistency, categorical variables were standardized by converting text to lowercase
and trimming white spaces. This process normalized entries (such as club names and player positions) so that similar values are not split into distinct categories due to formatting differences.

3. Exploratory Data Analysis (EDA)

The EDA section is divided into univariate, bivariate, and multivariate analyses to understand individual distributions, pairwise relationships, and complex interactions.

3.1 Univariate Analysis

Numerical Variables

Summary Statistics:

Descriptive measures (mean, median, standard deviation, skewness,
variance) for attributes such as age, overall rating, wage_eur, and market value highlighted that while many players cluster around typical age and rating ranges (for example, many players are in their mid-20s to early 30s), monetary attributes show a long tail indicating a few very expensive players dominate the averages.

Histograms and Box Plots:

Histograms for features like overall rating and numerical player skills revealed multimodal and skewed distributions.
Box plots exposed the presence of extreme values even after treatment, with some variables exhibiting slight asymmetry or remaining long tails.

Categorical Variables

Frequency Distributions:

Bar plots and pie charts for variables such as nationality, club_name, league_name, and player_positions indicate that a limited set of clubs and leagues accounts for the majority of records. For instance, dominant leagues like the English Premier League and Spanish Primera Division appear more frequently, and positions such as
“RW” and “ST” occur in clusters.
These univariate analyses establish the base level of understanding regarding the data distribution and variability across both player performance and market-related features.

3.2 Bivariate Analysis

Correlation Analysis:

A correlation matrix of numerical features revealed strong positive correlations between overall rating and potential, as well as between wage_eur and overall rating. These associations suggest that higher-rated players tend to have higher potentials and command larger wages.
Some other skill-based metrics (for example, passing, dribbling, and shooting) revealed moderate correlations with overall performance, which supports their potential importance in predictive models.

Scatter Plots:

Scatter plots such as overall rating vs. potential show near-linear relationships, reinforcing that a player’s current performance is strongly related to his future potential.
Other scatter plots (e.g., between wage_eur and overall rating) illustrate that while the relationship is positive, there is noticeable scatter due to factors like market negotiation and club budgets.

Categorical vs. Numerical Comparisons:

Box and violin plots comparing overall ratings across different leagues or clubs indicate that players in top leagues (e.g., the English Premier League, Spanish Primera Division) generally

have higher medians and a narrower interquartile range, though some leagues do exhibit a wide spread due to differing club strategies.
These bivariate analyses help in identifying which variables have the strongest relationships, setting a foundation for feature selection in further modeling work.

3.3 Multivariate Analysis

Pair Plots:

Pair plots of selected numerical features reveal clusters of players who share similar attributes. Such clustering is observable among players in similar positions or with similar performance and market ratings. This confirms the hypothesis that players tend to group together based on shared skill profiles and positions.

Grouped Comparisons:

For example, when grouping by primary playing positions (extracted from the player_positions field), the average wage analysis shows that forwards and attacking midfielders tend to have higher wages compared to defenders.
These visualizations provide insights into how different combinations of attributes influence player valuation.
The multivariate analysis thus uncovers complex interactions in the dataset and suggests that
player performance and financial metrics are interdependent on positional roles and league-specific factors.

4. Inferences and Conclusions

1. Data Preparation Impact:

Rigorous handling of missing values, duplicate records, and outliers was crucial. Median imputation for numerical and mode imputation for categorical features ensured robustness while capping outliers reduced distortion. The standardized categorical values enhanced consistency, which is vital for reliable grouping and comparisons.

2. Univariate Insights:

Numerical distributions, particularly for market-related features like wage and value, are heavily skewed by a few high-value players. Most player ages cluster between 20 and 35, and overall ratings generally concentrate in the upper 70s to mid-90s.
Categorical frequency charts reveal that the dataset is dominated by players from a few top clubs and leagues, which may have implications for club-specific or league-specific predictive models.

3. Bivariate Relationships:

The strong correlation between overall ratings and potential, and between wage and overall rating, supports the notion that a player's performance profile is tightly linked totheir market value.
Scatter plots confirm a positive trend, even as some level of variance suggests additional influencing factors (e.g., market conditions and club financial strength).

4. Multivariate Patterns:

Pair plots and grouped bar charts highlight that players tend to form clusters based on positional roles and performance metrics. Attacking players, for example, often show both higher wages and superior skill metrics compared to defenders.
The multivariate visualization emphasizes that complex interactions (including the effect of league and club competition) are at play in determining player value and performance.

5. Final Remarks

This analysis of the players_21.csv dataset provides a solid foundation for further work in sports
analytics. By thoroughly cleaning the dataset and exploring its multidimensional structure, the
study reveals clear patterns between player attributes and market valuations. Such insights can
be leveraged in predictive modeling, talent scouting, and strategic team-building efforts.
For future work, integration of additional external data (such as match performance metrics or
injury history) could further enhance these insights. Moreover, using advanced analytical methods like clustering or supervised learning may help in segmenting players based on performance and market trends.