# **Milestone 1: Data Understanding, Cleaning and Integration**

## **1. Data Collection and analysis**

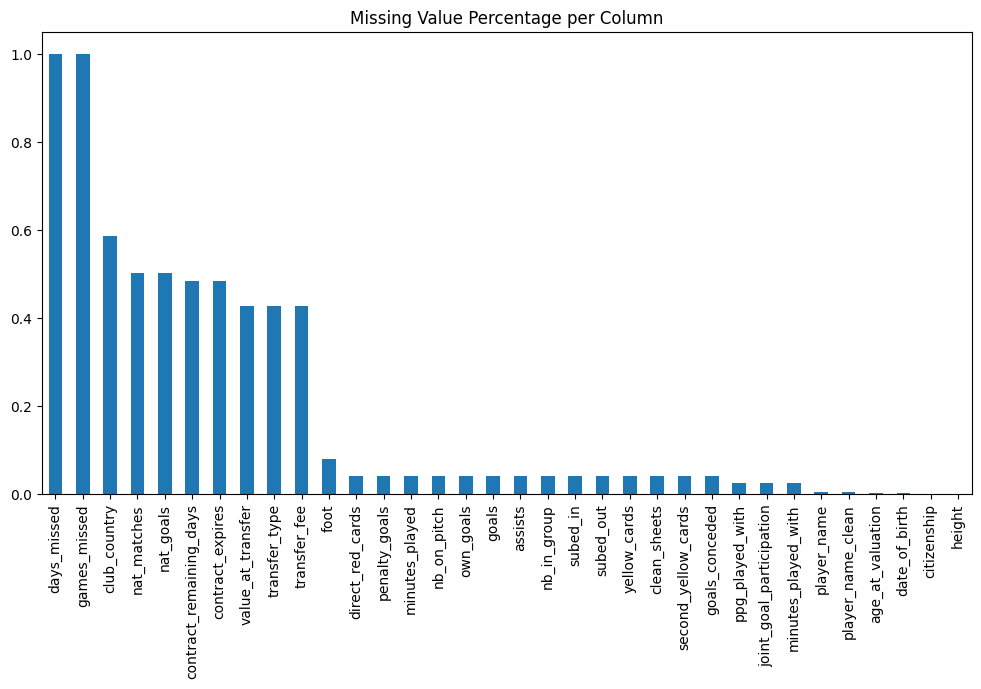
Initially, multiple real-world football datasets were collected to capture the diverse factors that influence a player’s market value. These datasets include “player profiles, match performance statistics, injury records, transfer history, historical and latest market values, national team appearances, teammate interactions, and team-level competition information.” Each dataset was explored individually to understand its structure, scale, and relevance. This step helped in identifying overlaps between datasets, confirming player coverage across sources, and gaining an overall understanding of how different football-related attributes are distributed and recorded.

## **2. Exploratory Data Analysis and Quality Assessment**

An initial exploratory data analysis was carried out to evaluate the overall quality, completeness, and reliability of the collected datasets. The combined raw data consisted of **11** independent sources, covering more than **92,000** unique players in the player profile datasets and approximately **79,000 players** with available market value information. During early joins, the total number of rows exceeded **900,000**, primarily due to the presence of match-level, season-level, and event-level records for individual players.

A detailed missing-value analysis revealed that several columns contained **90–100% null values**, particularly URL-based attributes, social media links, and secondary club information. In contrast, core numerical features such as goals, assists, minutes played, injury counts, and transfer fees showed relatively low missing rates, confirming their reliability for modeling. Distribution analysis also highlighted strong right-skewness in financial variables such as market value and transfer fees, indicating the presence of extreme outliers and large inter-player variability.

Significant data inconsistencies were identified across datasets. Season information appeared in multiple formats (e.g., *22/23*, *98/99*, *03-Apr*, *2024*), while critical date fields such as injury periods, transfer dates, and market value timestamps were stored as strings rather than standardized date types. Additionally, multiple duplicated player entries were observed after naïve merging, caused by repeated performance records, injury events, and historical market value snapshots.



## **3. Identifying Structural Inconsistencies**

One of the major issues identified was the lack of structural consistency across datasets. Season information appeared in multiple formats, and important date fields such as transfer dates, injury periods, and market value timestamps were stored as strings. These inconsistencies limited the ability to perform time-based calculations. To resolve this, all date-related fields were converted into standardized datetime formats, and season representations were normalized into a single numeric year format. This ensured temporal consistency across datasets and enabled reliable trend analysis.

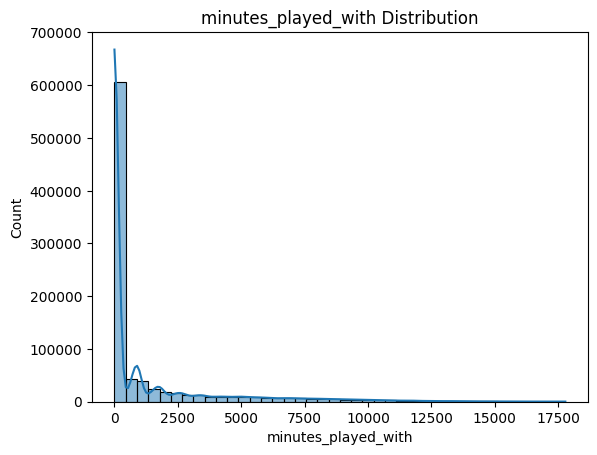
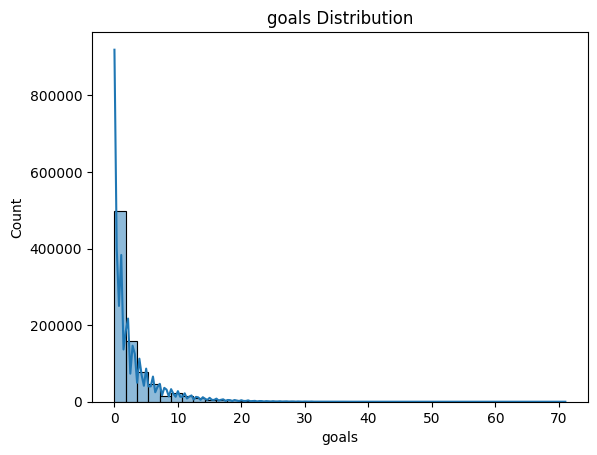


Fig 1.a. Goal Distribution Fig 1.b. Minutes Played

## **4. Data Cleaning and Feature Refinement**

Data cleaning focused on removing noise and improving dataset quality. Columns with extremely high missing values or those carrying only descriptive or visual information, such as URLs and image links, were eliminated. Partially available URL-based fields were converted into binary indicators to retain minimal informational value. Additionally, non-contributing attributes such as player name variants and redundant metadata were removed. Numeric and categorical columns were standardized into appropriate data types to ensure computational efficiency and consistency during integration.

## **5. Aggregation from Event-Level to Player-Level Data**

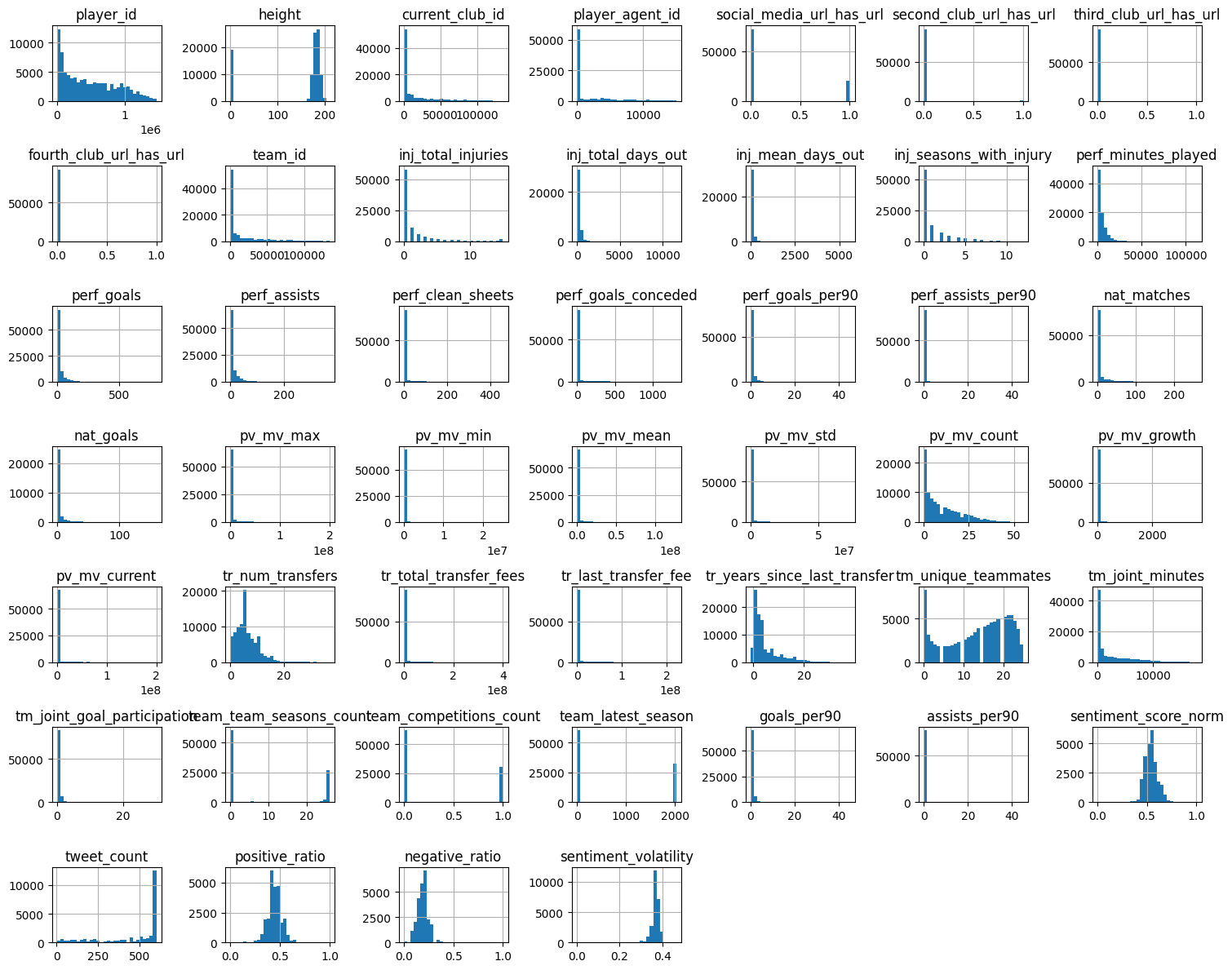
Since most datasets recorded events at match or season level, direct merging resulted in an unmanageable increase in rows. To address this, a player-centric aggregation strategy was adopted. Performance statistics were summarized using cumulative and per-90 metrics, injury data was condensed into total and average measures, and transfer history was aggregated to reflect career movement and regency. Market value history was summarized through statistical and trend-based features. This transformation ensured that each player was represented by a single, comprehensive record.

## **6. Integration of Multi-Source Football Data**

After cleaning and aggregation, all datasets were integrated using player identifiers as the primary key. Special care was taken to preserve meaningful relationships between player-level, team-level, and competition-level information. Team reputation, competition exposure, and teammate interactions were incorporated as contextual features. This integration process resulted in a unified dataset that captures both individual performance and environmental factors affecting a player’s market value.

## **7. Final Dataset Preparation and Readiness for Modeling**

The final outcome of this milestone is a consolidated, machine-learning-ready dataset containing approximately 92,000 players with over 300 engineered features. Each row represents a unique player, ensuring suitability for supervised learning tasks. The dataset was saved in compressed and sample formats for efficient storage and inspection. By resolving inconsistencies, eliminating redundancy, and aligning multi-season data, this milestone establishes a robust foundation for feature engineering, model development, and deployment in subsequent phases of the project.



The figure presents univariate histograms for the numerical features in the football player dataset, providing insight into their underlying distributions. Most performance, transfer, injury, and market value–related variables exhibit strong right skewness, with a large concentration of players at lower values and a small number of extreme outliers representing elite players.

## **Outcomes**

Through systematic data cleaning, normalization, rolling analysis, risk modeling, sentiment integration, and efficient encoding, a robust feature engineering pipeline was developed. This pipeline significantly enhances the predictive capability of machine learning models for football transfer value estimation taking market value as the target feature.

### 