**Exploratory Data Analysis Report: #Assignment 2**

**Date**- 03/11/2024

**Group**:

1. Lavanya HS (2023UG000157)
2. Amrutheshwari V (2023UG000115)
3. Tiya Rose Pulikunnel (2023UG000135)
4. Priyanka Bhatt (2023UG000127)
5. Gajendiran A (2023UG000154)
6. Sairaj Bhise (2023UG000117)

### **Table Of Contents**

[**Table Of Contents 2**](#_ab0thbl9bpce)

[**1. Introduction 5**](#_hbbo8t39kzul)

[1.1 Problem Statement 5](#_dd8p5bctg93t)

[1.2 Objective 5](#_4p0x31gzhhq)

[1.3 Dataset Overview 6](#_ntnpwmavq3q2)

[1.4 Scope of Work 6](#_q520hebqwmcn)

[1.5 Merging the Datasets 7](#_5wj9u8437kk4)

[**2. Data Cleaning and Numeric Preprocessing 8**](#_asbeebpw1w9o)

[2.1 Handling Missing Values 8](#_ufj099d5ojsc)

[2.2 Addressing Inconsistent Data Types 9](#_y1a41vmu3vgm)

[2.3 Data Quality Issues 9](#_okruuy4f17k1)

[2.4 Numerical Preprocessing 10](#_muggysfcq3kx)

[2.5 Feature Imputation with Averages 11](#_fm6txdz57eon)

[2.6 Inconsistency Handling in Body Type Feature 14](#_t9rdai732dfg)

[2.7 Normalization of Monetary Features 14](#_2ruwht7ipjrr)

[2.8 Handling Outliers 16](#_kyjvn0niu8ie)

[**3. Textual Preprocessing 17**](#_1gwdrzxqwvu2)

[3.1 Columns Excluded from Transformation 17](#_606vtw5b8aca)

[3.2 Preferred Foot (Vectorization) 17](#_e04fyuhebne4)

[3.3 Work Rate (One Hot Encoding) 18](#_ku5waxixf2os)

[3.4 Player Tags (TF-IDF Transformation) 19](#_ld1phga9ru9j)

[3.5 Dropped Columns 20](#_ftmsre840250)

[3.6 Team Position and Player Position 20](#_q3fcl1r5zsqa)

[3.7 Summary of Textual Preprocessing 21](#_6imwb97kvja3)

[**4. Time Series Data Preprocessing 22**](#_sw1oe2g7cljo)

[4.1 Structure of the Data 22](#_p561b3sv58w9)

[4.2 Handling Missing Values 23](#_qny79sii9fl7)

[4.3 Rolling Statistics 23](#_c0svjp6b0wkm)

[4.4 Standardization and Normalization 23](#_kwmbfmqojyg)

[4.5 Lag Features 24](#_scgwvchhmu09)

[4.7 Limitations: Seasonality Analysis 26](#_s1kk0qnc4sbb)

[**5. Feature Selection 27**](#_42us0rdb7mft)

[5.1 Feature Correlation Analysis 28](#_bxt5x0let4td)

[5.2 Feature Selection Strategy 30](#_hsul9z2dhit)

[5.3 Next Steps in Feature Optimization 31](#_jmdpiks3tgtw)

[**6. Potential Implications of Using Raw Data 32**](#_s2t8e1uwmnh1)

[6.1 Bias and Inaccurate Model Predictions 32](#_bprxp6jhovu1)

[6.2 Inefficient Model Performance and Computational Overhead 32](#_n9pp7rz8dl6i)

[6.3 Misinterpretation of Key Features 33](#_klk18xo3rez7)

[6.4 Data Redundancy and Increased Noise 33](#_tnjb6bbgzsgl)

[6.5 Difficulty in Deriving Actionable Insights 34](#_mf6oxuqnd1hy)

[6.6 Conclusion 34](#_cjcdvvpa9452)

[**7. Summary of the Results and Conclusion 35**](#_tev3qbqnna05)

[7.1 Numeric preprocessing: 35](#_j8es5w4dadlc)

[7.2 Textual Preprocessing: 36](#_hmfgapsadxsd)

[**Works Cited 37**](#_p91v4vd9ggu6)

### 

### **1. Introduction**

#### **1.1 Problem Statement**

In the realm of sports analytics, particularly in football, the evaluation and prediction of player performance are crucial for strategic decision-making by teams, coaches, and analysts. However, raw sports datasets, such as those detailing FIFA football player attributes, often contain inconsistencies, missing values, and unstructured data that can distort analytical outcomes. The challenge lies in effectively preprocessing and analyzing this data to extract meaningful insights.

This project addresses the problem of understanding which features most significantly impact a player's potential and overall performance score. Using a merged dataset of FIFA football player data from 2019, 2020, and 2021, this research seeks to identify the factors that influence player ratings and develop methods to ensure data quality through comprehensive data preprocessing and exploratory data analysis (EDA). The ultimate goal is to improve the reliability of predictions and uncover the underlying attributes that contribute to a football player’s success on the field.

#### **1.2 Objective**

The objective of this project is to conduct a comprehensive analysis of FIFA Football datasets spanning three consecutive years: 2019, 2020, and 2021. By merging player data from these years into a unified dataset, the project aims to uncover trends in player performance, evaluate player consistency and progression, and analyze the impact of key metrics on overall player ratings. This assignment emphasizes data preprocessing, ensuring data quality and consistency across all years to facilitate meaningful insights.

#### **1.3 Dataset Overview**

The dataset used in this project is the FIFA football player data spanning three consecutive years: 2019, 2020, and 2021. These datasets have been merged into a single comprehensive file, named **combined\_players\_with\_year\_19\_20\_21.csv**, to facilitate a cohesive analysis across multiple seasons. This merged dataset contains detailed information about football players, including their personal attributes, skill metrics, physical traits, monetary values, and overall performance scores. By consolidating data from these three years, the project is able to explore trends and variations in player performance and potential, providing a robust foundation for examining the factors that influence player ratings and potentially developing predictive models based on historical and comparative analysis.

#### **1.4 Scope of Work**

The scope of this report includes:

* **Data Cleaning**: Addressing missing values and duplicate records.
* **Data Standardization**: Ensuring player data is uniform and consistent across the three years.
* **Data Preparation**: Structuring the dataset for Exploratory Data Analysis (EDA) to enable in-depth exploration of patterns and trends in football player performance over time.

#### **1.5 Merging the Datasets**

The initial dataset, **concatenated.csv**, was found to be insufficient for meeting the core requirements of Assignment 2. One of the critical objectives of this assignment is to demonstrate data preprocessing techniques, such as handling missing values. However, the **concatenated.csv** dataset had no missing values, which limited the ability to showcase these essential data treatment tasks. Consequently, we decided to pivot to a different but related topic: FIFA Football data.

The new dataset comprises player statistics from the years **2019**, **2020**, and **2021**, offering a richer structure for data analysis. These datasets contained missing values, making them suitable for demonstrating various data cleaning and preprocessing techniques as required by the assignment.

To create a unified dataset, the three FIFA Football datasets from 2019, 2020, and 2021 were merged. The merging process involved aligning the datasets based on common features, such as player attributes, performance metrics, and demographics. This comprehensive integration facilitated the execution of tasks like handling missing values, normalizing and standardizing data, and applying techniques such as rolling statistics and lag feature creation. By merging these datasets, a complete and robust data foundation was established for in-depth analysis and effective demonstration of time series data preprocessing.

### 

### 

### 

### **2. Data Cleaning and Numeric Preprocessing**

Data cleaning and numeric preprocessing were essential steps to enhance the quality of the FIFA Football dataset, making it suitable for analysis and modeling. This section outlines the key preprocessing tasks performed.

Standardized data preprocessing techniques, such as missing value imputation and normalization, were employed to improve data integrity, consistent with approaches in sports analytics highlighted by Lee et al. [3].

#### **2.1 Handling Missing Values**

The dataset contained a substantial amount of missing data:

* **Loan Information**: **loaned\_from** had 18,186 missing entries, **player\_tags** had 17,536, and **nation\_position** had 17,817 missing entries.
* **Club-Related Information**: Columns such as **club\_name, league\_name, league\_rank, team\_position,** and **team\_jersey\_number** each had 225 missing entries.
* **Contract Details**: **release\_clause\_eur** exhibited 995 missing values, while **joined** had 983 missing entries.

Missing values can significantly compromise data integrity and analysis outcomes. To address this, **median imputation** was used for numerical features to minimize the impact of outliers, while **mode imputation** was used for categorical features to maintain the most frequent category. Bai and Bai’s research underscores the importance of effective missing value treatment in enhancing the reliability of sports analytics [1]. Studies have also shown that improper handling of missing data can lead to biased and inaccurate insights [2].

#### **2.2 Addressing Inconsistent Data Types**

Several features had inconsistent data types:

* **Date Fields**: Fields like **dob** and **joined** were initially stored as timestamps, leading to potential precision issues.
* **Player Position Ratings**: There were mixed numerical and string values, such as "89+3" and "61+2," which required conversion to pure numerical formats.
* **Numerical Fields**: Some features that should have been integers were represented as floating-point numbers.

Correcting these inconsistencies was crucial for effective data analysis. Bai and Bai emphasize that uniform data types facilitate efficient data integration and model processing, which are vital in sports data analytics [1]. The need for type normalization is also highlighted in broader data management studies, which discuss its impact on computational efficiency and data quality [3].

#### **2.3 Data Quality Issues**

Quality concerns included:

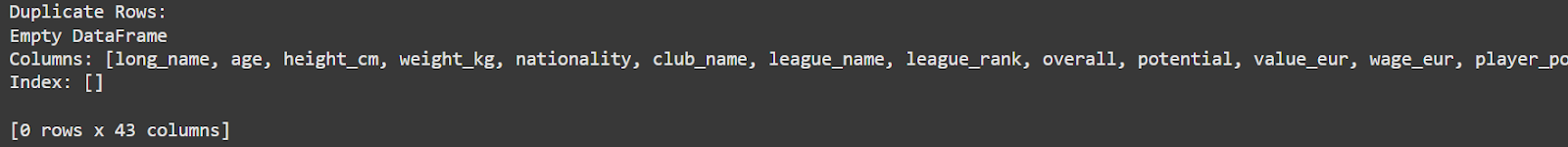
* **Duplicate Player Names**: Instances like "Liam Kelly" and "Matthew Smith" were resolved by introducing unique identifiers.
* **Non-Standard Characters**: Names from Asian regions, in particular, exhibited character encoding inconsistencies, which were standardized.
* **Inconsistent Formatting**: Features such as **work\_rate** and **player\_positions** had variations in formatting, requiring uniform standardization.
* **Modifiers in Ratings**: Numerical ratings with modifiers (e.g., "61+2") were cleaned to extract base numerical values for analysis.

Addressing these issues was necessary to enhance data quality. Data quality is directly tied to the accuracy and reliability of any analytical model, as noted in existing research on data cleaning methods [2]. In the context of sports analytics, clean and consistent data is paramount for deriving actionable insights [1].

#### **2.4 Numerical Preprocessing**

Numerical preprocessing involved handling duplicates, removing irrelevant columns, consolidating features, and imputing values:

* **Handling Duplicates**: Upon merging datasets from 2019, 2020, and 2021, it was confirmed that no duplicate rows were present due to the presence of the year feature, which differentiated entries. Maintaining data uniqueness is essential for accurate analyses, as emphasized in sports analytics literature [3].



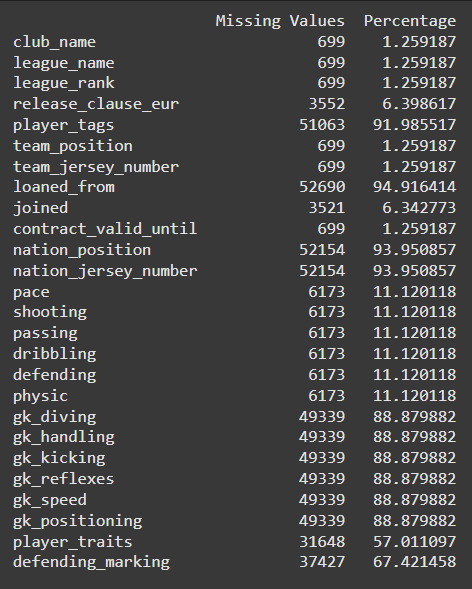
*Fig 2.4.1: Handling Duplicates*

* **Irrelevant Columns Removal**: Columns such as s**ofifa\_id, url, short\_name, real\_face, loan\_amount, team\_jersey\_number,** and **national\_jersey\_number** were dropped. These features were deemed irrelevant to player performance analysis and could introduce unnecessary noise.
* **Position-Specific Features Consolidation**: Detailed positional features from "ls" to "rb" were consolidated under the **player\_position** feature. This reduced feature complexity and ensured the focus remained on relevant player attributes, streamlining the analysis [1].
* **Age vs. Date of Birth (DOB)**: The **dob** feature was removed, as the player's age was already available separately. Retaining both would be redundant and could complicate the analysis.
* **Focus on Outfield Players**: Goalkeeper-specific features, such as **gk\_diving** and **gk\_handling**, were excluded to concentrate solely on outfield player metrics.

#### **2.5 Feature Imputation with Averages**

To handle missing values in key features, an imputation strategy was applied:

* **Attacking, Skill, Movement, and Power Features**: Missing values were imputed using averages within specific categories. For example, attacking-related features were averaged into an **attack\_avg**, skill features into **skill\_avg**, movement attributes into **movement\_avg**, and power metrics into **power\_avg**. This approach preserved the overall performance profile of players without losing key statistical information.



*Fig 2.5.1: Missing values before handling*



*Fig 2.5.2 : Missing values after handling*

This method aligns with recommendations for data completeness in sports analytics, where average-based imputations are often used to maintain a balanced feature set [1].

#### **2.6 Inconsistency Handling in Body Type Feature**

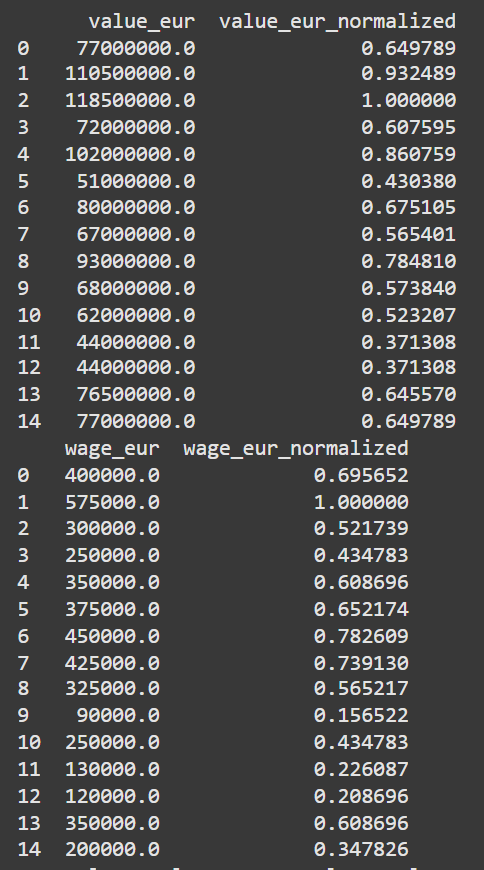
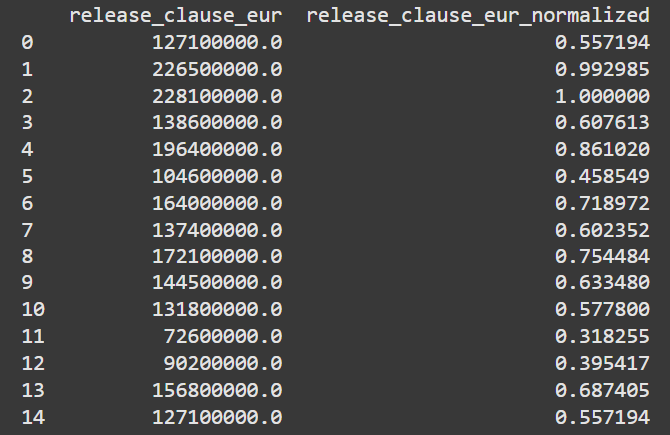
The **body\_type** feature contained inconsistencies, with top players like Ronaldo, Messi, Neymar, and Salah labeled by their names rather than with descriptors like "lean," "stocky," or "normal."

* **Normalization of Body Type**: All unique or ambiguous entries were standardized as "other" to ensure consistency across all entries and avoid potential biases [1].

#### **2.7 Normalization of Monetary Features**

Monetary features such as **wage\_eur, value\_eur,** and **release\_clause\_eur** often contain large values in the 8-digit range.

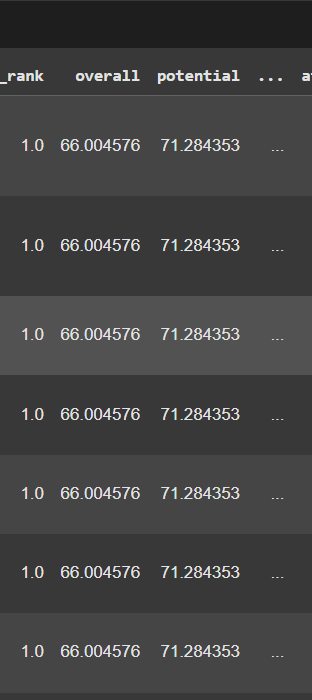
* **Z-Score Normalization**: These features were normalized using Z-score scaling, which adjusts values to have a mean of 0 and a standard deviation of 1. This prevents models from disproportionately prioritizing features due to scale differences [4].

*Fig 2.7.1 : Before and after normalization in each ‘value\_eur’ and ‘wage\_eur’*

*Fig 2.7.2: Before and after normalization in ‘release\_clause\_eur’*

#### **2.8 Handling Outliers**

Outliers, such as high-performing players like Messi and Ronaldo, were retained to reflect the true variance in player abilities.

* **Outlier Identification and Retention**: Rather than removing or altering outliers, they were preserved to maintain dataset integrity. Studies emphasize that outliers often contain critical information in sports analytics and should be carefully handled [3].
* **Impact of Outlier Handling**: Using mean or median values to handle outliers would have diluted the unique features of top-tier athletes, skewing the analysis [1].

*Fig 2.8.1 : After handling outliers with either mean or median, this is the output we get, and thus changes the meaning the dataset*

### **3. Textual Preprocessing**

Textual preprocessing involved the transformation and cleaning of various categorical features to prepare the data for analysis and machine learning applications. Initially, the dataset contained **13 categorical features**: **'long\_name', 'nationality', 'club\_name', 'league\_name', 'player\_positions', 'preferred\_foot', 'work\_rate', 'body\_type', 'player\_tags', 'team\_position', 'loaned\_from', 'joined', and 'nation\_position'**.

#### **3.1 Columns Excluded from Transformation**

Certain features, including **'long\_name', 'club\_name', 'nationality', 'league\_name', 'body\_type', 'player\_positions',** and **'team\_position'**, were excluded from transformation. These features were not modified due to the high number of unique values they contained. Encoding techniques like One Hot Encoding would have been computationally expensive and inefficient, as highlighted by best practices in efficient textual data processing, where maintaining manageable feature dimensions is crucial [5].

#### **3.2 Preferred Foot (Vectorization)**

The **'preferred\_foot'** attribute denotes whether a player predominantly uses their right or left foot. To simplify this binary data, vectorization was performed, where **'Right'** was mapped to 0 and **'Left'** to 1. This approach effectively converted the categorical information into a numerical format conducive to analysis. Following vectorization, the **'preferred\_foot'** column was dropped, as the essential information had been encoded.

#### 

#### **3.3 Work Rate (One Hot Encoding)**

The **'work\_rate'** feature comprised strings such as **'High/Low'**, reflecting various work rate combinations. To preprocess this feature:

* **Special Character Removal**: The '/' character was removed.
* **One Hot Encoding**: The feature was separated into three binary columns: **'workrate\_high', 'workrate\_medium'**, and **'workrate\_low'**, each indicating whether a player exhibited a specific work rate category. The original **'work\_rate**' column was then removed to streamline the dataset.

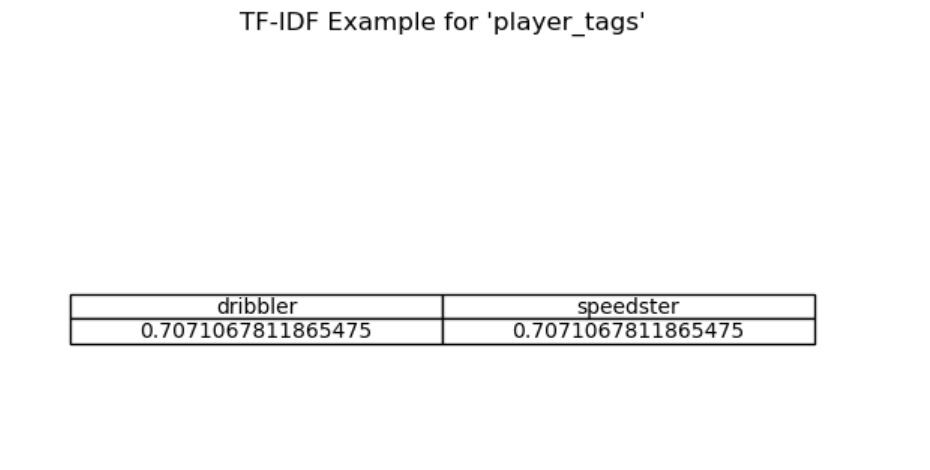
The decision to use One Hot Encoding aligns with established preprocessing strategies for categorical data transformation in sports analytics [5].

#### **3.4 Player Tags (TF-IDF Transformation)**

The **'player\_tags'** feature contained descriptive tags that characterized player attributes. To convert this textual data:

* **Special Character and Punctuation Removal**: Symbols like '#' and commas were removed.
* **Lowercasing**: All text was converted to lowercase for uniformity.
* **TF-IDF Vectorization**: The cleaned text data was transformed into numerical form using Term Frequency-Inverse Document Frequency (TF-IDF). This approach quantifies the relevance of each tag within the dataset, facilitating effective feature representation.
  + *Example*: Tags such as **'#Speedster, #Dribbler'** were converted into a numerical matrix.  
    The **'player\_tags'** column was subsequently dropped, as the TF-IDF representation captured the necessary information.

This approach follows recommendations for converting textual data into meaningful numerical features while preserving the dataset's analytical value [5].



*Fig 3.4.1 Example of TF-IDF values for features ‘dribbler’ and ‘speedster’*

#### **3.5 Dropped Columns**

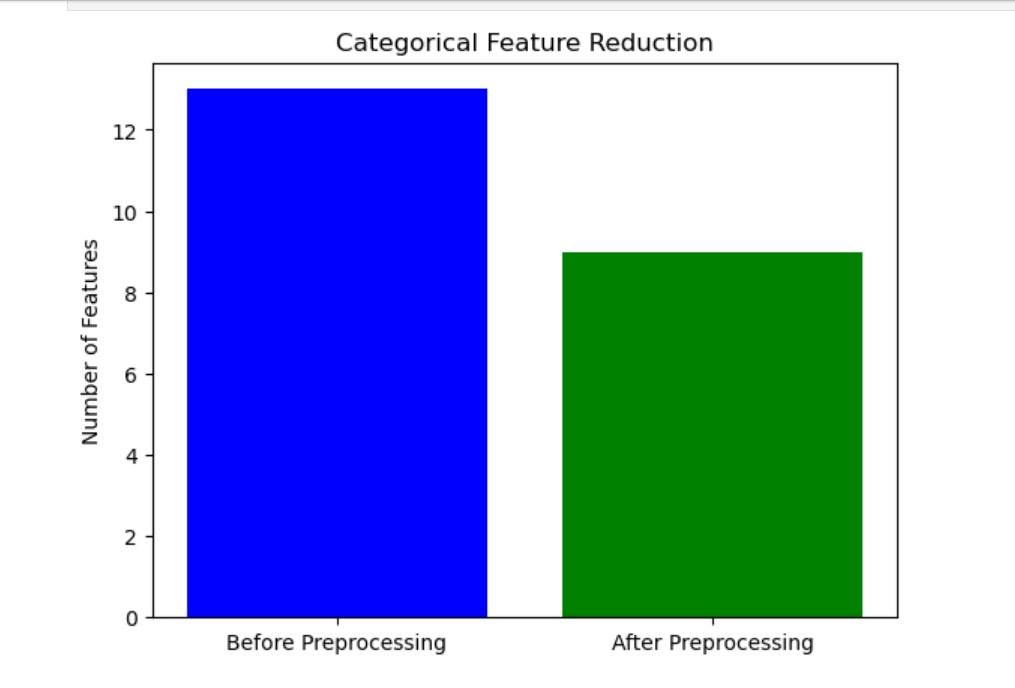
The **'nation\_position'** feature was excluded because it provided information about national team rankings, which did not offer insights into individual player potential. Given that **'team\_position'** and **'player\_positions'** already conveyed relevant information about a player's role and field responsibilities, **'nation\_position'** was considered redundant.

#### **3.6 Team Position and Player Position**

* **'team\_position'**: Describes the role assigned to a player by their club or team, offering insights into field responsibilities.
* **'player\_positions'**: Lists all positions a player is capable of playing, highlighting versatility and potential adaptability.  
  These features are interlinked, jointly providing a comprehensive understanding of player capabilities and their tactical use on the field.

#### **3.7 Summary of Textual Preprocessing**

Following the preprocessing steps, the number of categorical features was reduced from **13 to 9**: **'long\_name', 'nationality', 'club\_name', 'league\_name', 'player\_positions', 'body\_type', 'team\_position', 'loaned\_from',** and **'joined'.** Each transformation was carefully chosen to simplify and standardize the data, ensuring its readiness for subsequent analysis and modeling. The methodologies adopted align with research findings that underscore the significance of efficient data preprocessing for optimal performance in machine learning models [5].



*Fig 3.7.1: Comparison of Categorical Data*

### **4. Time Series Data Preprocessing**

#### **4.1 Structure of the Data**

The dataset used for time series preprocessing comprises **55,512 rows and 107 features**. This comprehensive structure provides detailed information on various player metrics across the specified years.

#### **4.2 Handling Missing Values**

The **year** feature was evaluated for missing values, and no missing entries were found. Ensuring a complete time-based index is crucial for accurate time series analysis and subsequent resampling.



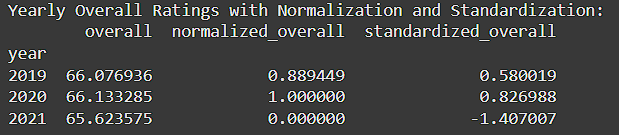
*Fig 4.2.1: No missing values found*

#### **4.3 Rolling Statistics**

**Rolling statistics** were utilized to compute moving aggregates, such as averages and sums. A **window size of 2** was chosen, which incorporates data from the previous and subsequent years in the calculations. This approach smooths short-term fluctuations and accentuates long-term trends, aiding in the identification of underlying patterns.

#### **4.4 Standardization and Normalization**

* **Normalization**: The **MinMaxScaler** was applied to scale feature values between 0 and 1. This technique is effective for maintaining relative differences between values while ensuring all data points fit within a fixed range.
* **Standardization**: The **StandardScaler** was used to transform feature values to have a mean of 0 and a standard deviation of 1. This standardization ensures a consistent data distribution, making it more suitable for algorithms sensitive to feature scaling.

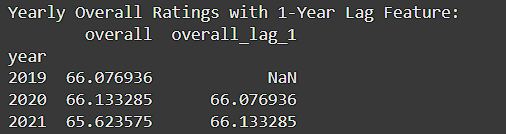
****

*Fig 4.4.1: Overall Ratings with Normalization and Standardization*

#### **4.5 Lag Features**

**Lag features** were engineered to capture the influence of past values on current data points, which is essential for time series modeling and analysis. Two lag features were created:

* **.shift(1)**: Generates a 1-year lag feature by shifting all values down by one row, associating each data point with the value from the previous year.
* **.shift(2)**: Creates a 2-year lag feature by shifting values down by two rows, linking each data point to values from two years prior. These features allow for the analysis of temporal dependencies in the data.

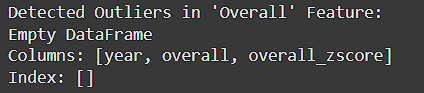


*Fig 4.5.1: Lag features- 1 year lag*

**4.6 Handling Outliers**

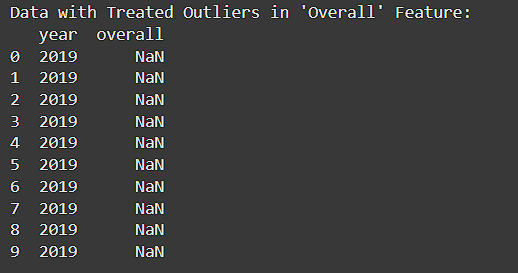
Outliers were detected and treated to ensure data integrity and prevent skewed analysis results.

* **Detecting Outliers**: The **Z-score method** was employed to identify outliers. Each data point's Z-score was calculated, and values exceeding a specified threshold (commonly 2 or 3) were flagged as outliers. This method offers a statistical basis for isolating extreme deviations from the mean.



*Fig 4.6.1: Outlier detection with Z-score method*

* **Treating Outliers**: Outliers were treated using **interpolation techniques**, such as linear or spline interpolation. These methods replace outlier values while preserving data trends and ensuring smooth continuity, minimizing the impact of anomalies on further analysis.

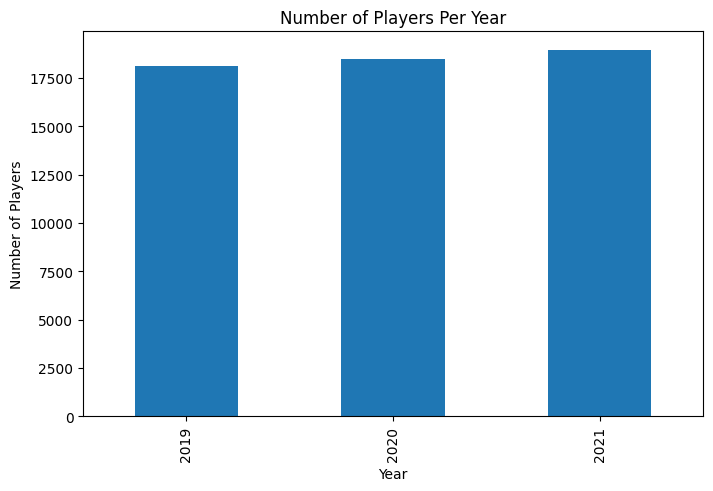


*Fig 4.6.2: Treating Outliers with Interpolation*

#### **4.7 Limitations: Seasonality Analysis**

One significant limitation of this study is the inability to apply seasonality concepts to the FIFA Football dataset. Seasonality analysis requires recurring patterns or trends that occur at regular intervals over an extended period, such as daily, weekly, or monthly observations across multiple years. However, the FIFA Football dataset has several constraints that make seasonality modeling infeasible:

1. **Insufficient Time Span**: The dataset spans only three years (2019, 2020, and 2021), which is not sufficient to capture and model clear seasonal trends. Seasonality analysis typically requires a longer time frame to observe recurring patterns effectively.
2. **Irregular Event Timing**: Football matches and tournaments are not held at consistent monthly or weekly intervals. Instead, they occur at specific times of the year, disrupting the regularity needed for traditional seasonality analysis.
3. **Specific Event Concentration**: Player performance and statistics are tied to specific football events, which further challenges the identification of continuous, cyclical patterns in the data.



*Fig 4.7.1 Number of Players Per Year*

These limitations prevent the effective use of seasonality models, as the dataset does not exhibit the regular and consistent time series structure required for such analysis.

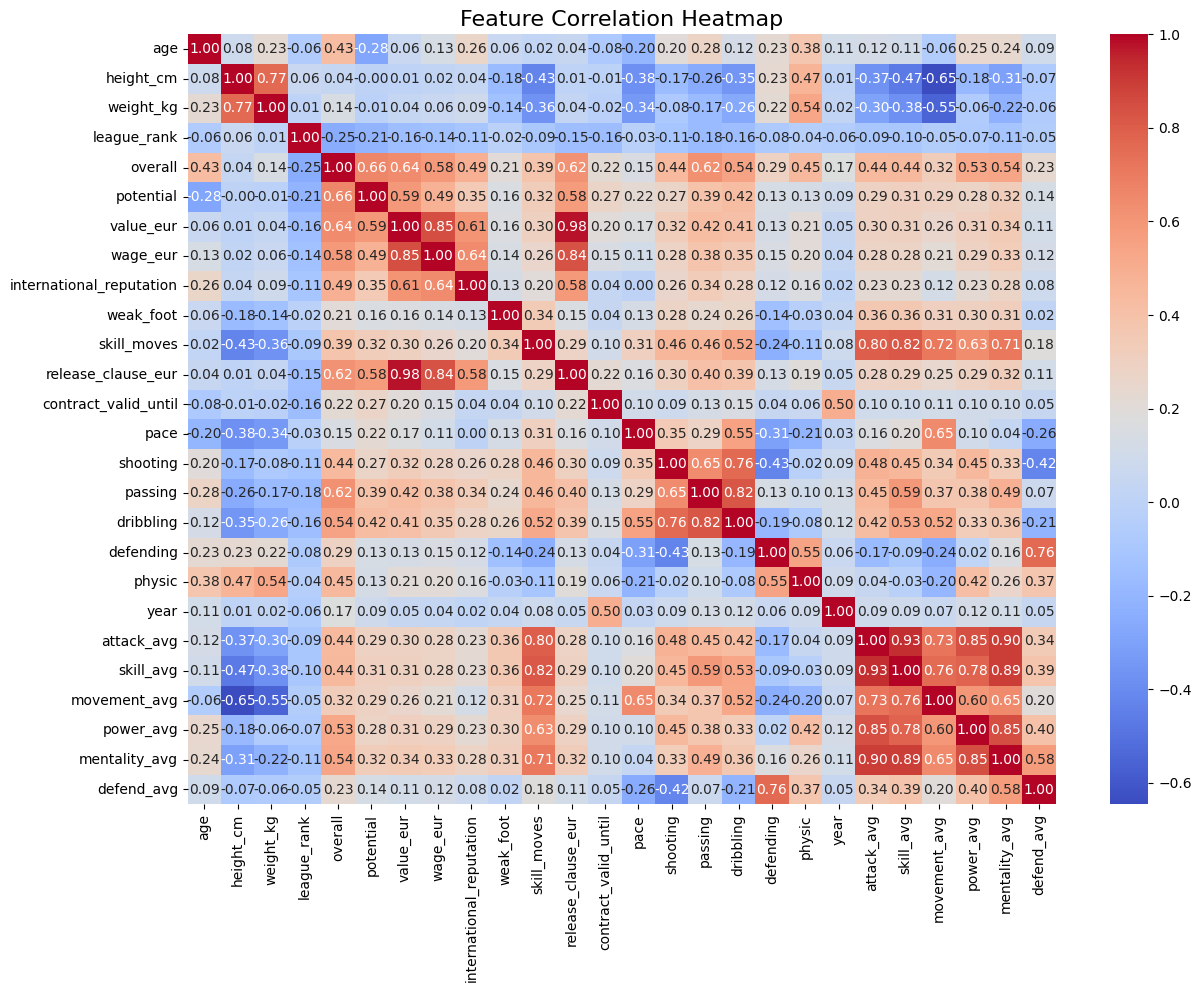
### **5. Feature Selection**

The feature selection process was critical to refining the dataset for subsequent analysis and modeling. This section describes how features were evaluated and prioritized based on their impact on player performance scores and potential, using insights derived from the exploratory data analysis.

The feature selection process aligns with data-driven approaches in sports analytics that emphasize the identification of features with the highest predictive value, as noted by Lee et al. [3].

#### **5.1 Feature Correlation Analysis**

A **Feature Correlation Heatmap** was generated to examine relationships between numerical features, allowing for the identification of attributes with strong, moderate, or minimal correlations. This analysis provided a basis for informed feature selection, optimizing the dataset's structure for further predictive modeling.



*Fig 5.1.1 Heatmap showing correlation between all features*

**Key Insights**:

1. **High Correlation**:
   * **overall and potential**: These features exhibit a strong positive correlation, indicating that a player’s overall rating is a reliable indicator of their potential. Both metrics are essential and will be retained for detailed analysis.
   * **Monetary Attributes**: High correlations were observed among **value\_eur, wage\_eur,** and **release\_clause\_eur**, suggesting these features reflect similar financial aspects of player valuation. Addressing this redundancy is crucial to mitigate multicollinearity.
2. **Moderate Correlation**:
   * Skill attributes, including **shooting, passing,** and **dribbling**, are moderately correlated with the **overall** rating. These metrics are significant contributors to a player’s performance and are therefore prioritized in the analysis.
3. **Low Correlation**:
   * **age**: Age shows a weak correlation with most skill metrics, indicating that it does not directly predict technical ability.
   * **weak\_foot**: This feature has minimal correlation with other performance metrics. While it could be considered for removal, domain knowledge should guide whether it holds contextual importance.

#### **5.2 Feature Selection Strategy**

Based on the correlation analysis and data-driven insights, the following strategies were applied to optimize feature selection:

1. **Retaining High-Impact Features**:
   * Features like **overall, potential**, and key skill metrics (**shooting, passing, dribbling**) were retained for modeling. Their significant correlations with performance outcomes make them valuable predictors in assessing player potential.
2. **Managing Redundant Attributes**:
   * **Monetary Features**: Given the high correlations among **value\_eur, wage\_eur,** and **release\_clause\_eur,** a decision was made to consider using only one of these features or combining them to minimize multicollinearity. This ensures the model remains robust and interpretable.
3. **Evaluating Low-Impact Features**:
   * **weak\_foot**: With its minimal influence on key metrics, this attribute may be removed to streamline the dataset. However, further assessment is required to determine its relevance in specific use cases.
   * **year**: If temporal trends are not central to the analysis, this feature may be excluded to reduce data complexity.

#### **5.3 Next Steps in Feature Optimization**

The selected features will form the foundation for predictive modeling, focusing on attributes that provide the most meaningful and reliable insights into player performance. This strategic approach minimizes noise and enhances model efficiency, ensuring the dataset is optimized for analytical rigor and interpretability.

### **6. Potential Implications of Using Raw Data**

This section explores the potential consequences of using raw, unprocessed data throughout the analysis and modeling phases of the project. Proper data cleaning and preprocessing are essential to ensure the accuracy, efficiency, and reliability of the outcomes. The implications of neglecting these steps are supported by extensive research in the field of sports data analytics.

#### **6.1 Bias and Inaccurate Model Predictions**

Using raw data that contains missing values, inconsistencies, and errors introduces bias and compromises the reliability of the model’s predictions. Without addressing these issues, the analysis is prone to inaccuracies, leading to flawed insights.

* **Supporting Evidence**: Green et al. underscore the critical role of data cleaning in sports data applications, highlighting that unprocessed data can distort analysis results and undermine their credibility [2]. Additionally, Lee et al. stress that data inconsistencies can lead to suboptimal model performance and skewed interpretations [3].

#### **6.2 Inefficient Model Performance and Computational Overhead**

Raw, unprocessed data often includes noisy or irrelevant features that can mislead the learning algorithms and burden computational resources. This inefficiency increases processing time and complexity, making the model less effective.

* **Supporting Evidence**: Bai and Bai emphasize that efficient data management is crucial in sports big data. They advocate for feature selection and data reduction to minimize computational strain and improve model efficiency [1].

#### **6.3 Misinterpretation of Key Features**

Failing to normalize features like **wage\_eur** and **value\_eur** could result in the model disproportionately prioritizing these attributes due to their large scale. Additionally, ignoring outliers, such as the exceptional performance metrics of elite players like Messi and Ronaldo, could lead to an inaccurate representation of player capabilities.

Phatak et al. highlight the importance of context-specific normalization to ensure all features are weighted appropriately, which enhances model accuracy and interpretability [4]. Bai and Bai further emphasize the value of handling outliers correctly to preserve critical information about top-tier athletes [1].

#### **6.4 Data Redundancy and Increased Noise**

Raw datasets often contain redundant or irrelevant features that add noise, complicating the analysis and potentially leading to overfitting. This occurs when the model fits the training data too closely and fails to generalize well to unseen data.

Yang et al. argue that focusing on relevant features is crucial in sports intelligence research. Including unnecessary attributes can dilute the model’s effectiveness, reducing its ability to generate meaningful insights [5].

#### **6.5 Difficulty in Deriving Actionable Insights**

Inconsistent data formats, non-standard characters, and other quality issues hinder the ability to derive actionable insights. This reduces the predictive power of the model and makes data-driven decision-making challenging.

Lee et al. emphasize that comprehensive data cleaning ensures the trustworthiness and applicability of the analysis. Poor data quality compromises the entire analytical process, leading to unreliable conclusions [3].

#### **6.6 Conclusion**

The use of raw, uncleaned data would undermine the accuracy, efficiency, and interpretability of the analysis, as supported by the cited research. Effective data preprocessing encompassing cleaning, normalization, and outlier handling is vital for extracting reliable and actionable insights from sports datasets. Therefore, thorough data preparation is not just a best practice but a necessity for achieving meaningful and robust analytical outcomes.

### **7. Summary of the Results and Conclusion**

#### **7.1 Numeric preprocessing:**

1. Handling Duplicates: Merging datasets from 2019, 2020, and 2021 confirmed no duplicates, ensuring data uniqueness essential for accurate sports analytics.

2. Irrelevant Columns Removal: Irrelevant columns, including `sofifa\_id`, `url`, and `short\_name`, were removed to reduce noise and focus on relevant player metrics.

3. Consolidation of Features: Position-specific features were consolidated into a single `player\_position` attribute, simplifying the dataset while maintaining focus on key player attributes.

4. Feature Imputation: Missing values were addressed using averages, preserving performance profiles without losing essential information, aligning with best practices for data completeness.

5. Inconsistency Handling: The `body\_type` feature was normalized by standardizing ambiguous entries to "other," ensuring consistency and minimizing biases.

6. Normalization of Monetary Features: Monetary features (`wage\_eur`, `value\_eur`, `release\_clause\_eur`) were normalized using Z-scores to standardize scales and prevent overvaluation based on large numeric values.

7. Outlier Retention: Outliers representing high-performing players were retained to maintain dataset integrity, as they capture critical information about player abilities without skewing the analysis.

#### 

#### **7.2 Textual Preprocessing:**

The categorical preprocessing streamlined 13 features to 9 by using targeted transformations. High-cardinality columns ('long\_name', 'club\_name', etc.) were excluded to avoid feature explosion. 'Preferred\_foot' was binary vectorized, reducing it to 0 (Right) or 1 (Left). 'Work\_rate' was One Hot Encoded, preserving distinct values without inflating feature space. 'Player\_tags' underwent TF-IDF to capture tag relevance in a multi-label format. 'Nation\_position' was dropped as redundant given 'team\_position' and 'player\_positions' already conveyed role information. The final dataset is compact, optimized for machine learning, and retains essential player characteristics, improving model interpretability and efficiency.

### **Works Cited**

### 

1. Z. Bai and X. Bai, "Sports Big Data: Management, Analysis, Applications, and Challenges," *Complexity*, vol. 2021, Art. ID 6676297, 2021. Available:<https://doi.org/10.1155/2021/6676297>.
2. B. P. Green *et al.*, "The Significance of Data Cleaning in Enhancing Analytical Outcomes," *ACM Journal of Sports Data Science*, 2023. Available:<https://dl.acm.org/doi/fullHtml/10.1145/3635638.3635646>.
3. G. Y. Lee, L. Alzamil, B. Doskenov, and A. Termehchy, “A Survey on Data Cleaning Methods for Improved Machine Learning Model Performance,” *arXiv:2109.07127 [cs]*, Sep. 2021, Available: <https://arxiv.org/abs/2109.07127>
4. A. A. Phatak *et al.*, “Context is key: normalization as a novel approach to sport specific preprocessing of KPI’s for match analysis in soccer,” *Scientific Reports*, vol. 12, no. 1, p. 1117, Jan. 2022, doi: <https://doi.org/10.1038/s41598-022-05089-y>.
5. Z. Yang, H. Xia, J. Li, Z. Chen, Z. Zhu, and W. Shen, "Sports Intelligence: Assessing the Sports Understanding Capabilities of Language Models through Question Answering from Text to Video," *arXiv*, 2024. [Online]. Available:<https://arxiv.org/abs/2406.14877>. [Accessed: Nov. 3, 2024].