**Assignment 2**

Objective: Students will employ techniques for data cleansing, preprocessing (numeric, textual, and time-series), missing value treatment, and outlier detection to a dataset in order to conduct Exploratory Data Analysis (EDA). The objective of the assignment is to enhance their comprehension of the best practices for optimizing performance in Python and the management of unprocessed data from multiple sources.

Date on which assignment was provided: 25/9/2024

Marks: 35 Marks

Last Date of Submission: 25/10/2024

Assignment Questions and Tasks

Prepare assignments based on the questions. The utilization of a same dataset and comparable work by multiple groups will result in zero marks. Use Python programming whenever it is necessary. As it is a collaborative activity, all students must participate equally.

The questions below will help you look into it. You can add to this list of questions to help you explain your answer. Also, add another set of data which works along with previous dataset and merge them and work on the following tasks.

1. Implications of Raw Data   
  
- Question: What are the primary obstacles to utilizing unprocessed data in EDA? What methods can be employed to identify prospective issues in unprocessed data?   
  
- Objective:   
  
- Load the original dataset and provide a summary of its structure, including data types, missing values, and so forth.   
  
- Evaluate the potential consequences of employing this unprocessed data for analysis.   
  
2. Data from a Variety of Sources   
  
- Inquiry: How do you manage the assimilation of data from multiple sources? What obstacles arise during this process?   
  
- Objective:   
  
- If applicable, load and merge multiple datasets. Identify any discrepancies or duplicates.   
  
- Standardize the merged data to facilitate additional analysis.   
  
3. Data Cleaning   
  
- Question: What is the significance of data cleansing, and what are the most common methods used?   
  
- Objective:   
  
- Identify and eliminate duplicates.   
  
- Manage data formats that are inconsistent or inaccurate.   
  
4. Data Preprocessing: Numeric Data   
  
- Question: What is the procedure of preprocessing numeric data? What is the necessity of normalization or scaling?   
  
- Objective:   
  
- Implement normalization and/or standardization methods for the numerical columns.   
  
  
5. Textual Data Preprocessing   
  
- Question: What are the primary procedures for preprocessing textual data?   
  
- Objective:   
  
- Erase stopwords, punctuation, and perform tokenization to clean and preprocess textual data.   
  
- Convert the text to numerical representation using TF-IDF or another vectorization technique.   
  
6. Data Preprocessing: Time-Series Data   
  
- Inquiry: What is the process for preprocessing time-series data? What is the function of resampling?   
  
- Objective:   
  
Resample the time-series data to ensure that it is consistent in frequency.   
  
- Utilize forward or reverse fill to address absent values in time-series.   
  
7. The Significance, Detection, and Treatment of Missing Values

**The dataset “concatenated.df” has ……. missing values**   
  
- Inquiry: What is the significance of managing missing values? What are the most frequently employed methods for identifying and managing them?   
  
- Objective:   
  
- Apply imputation techniques (e.g., mean, median, or mode) to identify absent values.   
  
  
  
8. Significance, Types, Detection, and Treatment of Anomalous/Outlier Data   
  
- Inquiry: What are the different categories of outliers? What is the process for identifying and managing outliers in a dataset?   
  
- Objective:   
  
- Utilize the IQR or Z-score method to identify outliers.   
  
- Determine whether outliers should be eliminated or addressed in alternative manners.   
  
  
9. Optimal Performance Best Practices   
  
- In order to guarantee optimal performance, what are the most effective EDA practices?   
  
- Objective:   
  
- Employ techniques for feature selection.   
  
- Illustrate the primary relationships and correlations in the data.   
  
  
Report Format   
  
Students are required to submit a comprehensive report in PDF format, which should include:   
  
1. Introduction: A summary of the project's objectives and the dataset.   
  
2. Raw Data Analysis: A discussion of the framework and problems with the raw dataset.   
  
3. Data Cleaning and Preprocessing: The procedures implemented for data cleaning, missing value management, and preprocessing.   
  
4. Outlier Detection: Techniques employed to identify and address outliers.   
  
5. Visualizations: Charts and visualizations that illustrate data trends, correlations, and insights.   
  
6. Conclusion: Summary of the results and concluding reflections on the ways in which EDA enhanced the dataset to facilitate further analysis.   
  
  
Submission Requirements   
  
1. Please submit the Python code and explanations in your Jupyter Notebook (.ipynb) file.   
  
2. Submit a PDF report that provides a concise summary of your work.   
  
3. Deadline: By [Insert Deadline], submit all files through the university portal.   
  
4. Adhere to appropriate file naming conventions, such as `EDA\_Project\_StudentName.ipynb` and `EDA\_Project\_StudentName.pdf`.

**STARTING OF THE REPORT:**

1. Implications of Raw Data   
  
- Question: What are the primary obstacles to utilizing unprocessed data in EDA? What methods can be employed to identify prospective issues in unprocessed data?

Utilizing unprocessed data in Exploratory Data Analysis (EDA) can present several obstacles. Here are some primary challenges along with methods to identify potential issues in unprocessed data:

* Primary Obstacles

1. \*Data Quality Issues\*:

- \*Missing Values\*: Unprocessed data often contains gaps that can affect analysis.

- \*Inconsistent Formatting\*: Data might come from various sources, leading to different formats and units.

- \*Outliers\*: Extreme values may distort statistical analysis and visualizations.

2. \*Noise and Redundancy\*:

- \*Irrelevant Features\*: Unprocessed data may include features that do not contribute to the analysis, complicating interpretation.

- \*Duplicate Entries\*: Redundant records can skew results and lead to erroneous conclusions.

3. \*Scalability\*:

- \*Volume of Data\*: Large datasets can be cumbersome and time-consuming to analyze without proper preprocessing.

- \*Computational Constraints\*: Limited processing power can hinder the analysis of complex, unprocessed data.

4. \*Understanding Context\*:

- \*Lack of Metadata\*: Unprocessed data may lack documentation or metadata, making it difficult to understand its origins and meanings.

- \*Domain Knowledge\*: Without sufficient domain knowledge, it can be challenging to interpret data correctly.

* Methods to Identify Issues in Unprocessed Data

1. \*Data Profiling\*:

- Conduct statistical summaries (mean, median, mode, range) and visualizations (histograms, box plots) to assess distributions and identify anomalies.

2. \*Missing Value Analysis\*:

- Use tools to check for missing values and assess their patterns and impacts on analysis.

3. \*Outlier Detection\*:

- Apply statistical tests (like Z-scores, IQR) and visualization methods (scatter plots, box plots) to detect and analyze outliers.

4. \*Consistency Checks\*:

- Verify data consistency across different features and sources. This can involve checking formats, ranges, and logical relationships.

5. \*Duplicate Detection\*:

- Use algorithms to identify and remove duplicate entries to ensure data integrity.

6. \*Data Visualization\*:

- Utilize visual tools (scatter plots, heat maps, pair plots) to explore relationships and distributions, which can help identify trends, correlations, and anomalies.

7. \*Documentation Review\*:

- Check any available metadata or documentation to understand the data's context, origin, and intended use.

8. \*Exploratory Data Techniques\*:

- Apply techniques like principal component analysis (PCA) to reduce dimensionality and highlight relationships between features, which can reveal hidden issues.

By systematically addressing these obstacles and employing these methods, analysts can better prepare unprocessed data for meaningful exploration and analysis.

* Potential consequences of employing this unprocessed data for analysis:

Using unprocessed data for analysis can lead to several potential consequences that might impact both the quality of insights and decision-making outcomes. Here are some key potential consequences:

1. \*Misleading Insights\*

- \*Biased Results\*: Unprocessed data may contain biases (e.g., missing or outlier-heavy values) that skew the findings. Insights drawn from biased data can lead to incorrect assumptions about trends and patterns.

- \*Inaccurate Conclusions\*: Without addressing inconsistencies or outliers, the analysis may yield results that don't accurately represent the underlying data, leading to erroneous conclusions.

2. \*Faulty Predictive Models\*

- \*Poor Model Performance\*: Unprocessed data often lacks the standardization, normalization, and feature engineering necessary for machine learning models. Models trained on such data can perform poorly or fail to generalize well to new data.

- \*Overfitting or Underfitting\*: Noise and irrelevant features in unprocessed data can cause models to overfit (learn from noise) or underfit (fail to capture patterns), reducing their predictive accuracy.

3. \*Increased Computational Costs\*

- \*Longer Processing Times\*: Analyzing and modeling large, raw datasets require more computational resources. Noise and redundancy increase processing time and may strain storage and memory resources.

- \*Higher Costs\*: More complex computations can translate to higher costs, especially when scaling analyses across multiple projects or in resource-intensive environments.

4. \*Decision-Making Risks\*

- \*Poor Business Decisions\*: Insights based on unprocessed data may lead to poor strategic choices. For example, in finance, incorrect analysis of data may lead to wrong investment decisions; in healthcare, it could result in incorrect patient care guidelines.

- \*Reputational Risks\*: Erroneous conclusions can harm an organization’s credibility, especially if decisions are made based on inaccurate findings.

5. \*Compliance and Ethical Concerns\*

- \*Privacy Issues\*: Unprocessed data might contain sensitive or personally identifiable information (PII) that should be masked or anonymized. Failure to do so can violate privacy laws and regulations.

- \*Ethical Concerns\*: Analysis based on incomplete or biased datasets could perpetuate harmful biases, leading to unfair or unethical outcomes, especially in areas like hiring, lending, and social services.

6. \*Difficulty in Reproducibility\*

- \*Inconsistent Findings\*: Unprocessed data might yield non-replicable results due to uncontrolled variations. Analysis may change each time it’s performed if different cleaning steps or methods are applied inconsistently, reducing reproducibility.

- \*Documentation Gaps\*: Using unprocessed data without proper documentation or preprocessing can hinder collaboration, as others might struggle to understand or replicate the analysis without standardization steps.

Mitigating These Consequences

- \*Data Cleaning and Preprocessing\*: Standardize and clean data to improve consistency, accuracy, and overall quality.

- \*Feature Engineering\*: Select and transform relevant features to reduce noise and redundancy, enhancing model performance.

- \*Data Documentation\*: Properly document data sources, cleaning methods, and assumptions made during preprocessing.

- \*Compliance Checks\*: Ensure data complies with legal and ethical standards, particularly for privacy and fairness.

SUMMARY OF DATASETS:

2021:

For 2022 and 2019 its the same as 2021 as it follows the same features.

2)MULTIPLE SOURCE:

3) DATA CLEANING:

Based on the analysis, here are the main inconsistencies and inaccuracies in the data:

Missing Values:

Significant missing data in loaned\_from (18,186 missing entries)

player\_tags (17,536 missing entries)

nation\_position (17,817 missing entries)

Club-related information (225 missing entries each in club\_name, league\_name, league\_rank, team\_position, team\_jersey\_number)

Contract information (release\_clause\_eur has 995 missing values, joined has 983 missing entries)

Inconsistent Data Types:

Date fields (dob, joined) are stored as timestamps which can lead to precision issues

Mixed numerical and string values in player position ratings (e.g., "89+3", "61+2")

Some numerical fields that should be integers contain floating-point values

Data Quality Issues:

Duplicate player names exist (e.g., multiple "Liam Kelly" and "Matthew Smith")

Some player names contain non-standard characters or different character encodings (especially visible in Asian names)

Inconsistent formatting in work rates and player positions

Player ratings contain modifiers (e.g., "61+2") instead of clean numerical values

Potential Accuracy Issues:

Release clause values have many missing entries which could affect financial analysis

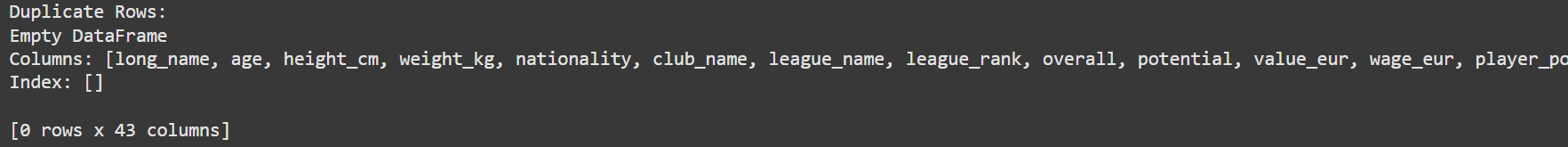
National team information (position and jersey numbers) is largely incomplete

Contract information has gaps that could affect player availability analysis

These inconsistencies could significantly impact any analysis performed on the dataset, especially when dealing with financial values, player positions, or team-related statistics.

NUMERIC PREPROCESSING:

1. Duplicates: After merging the 3 datasets, there were no duplicates as atleast the feature ‘year’ would be different, hence giving no duplicates.



1. Irrelevant Columns Removal

Removed Columns: The columns sofifa\_id, url, short\_name, real\_face, loan\_amount, team\_jersey\_number, and national\_jersey\_number were dropped. These features are irrelevant to our analysis, as they do not contribute to player performance or skill. Including them could introduce noise or unnecessary complexity, potentially leading to overfitting.

* Position-Specific Features Consolidation

Dropped Position Columns: Detailed columns describing specific player positions, such as from "ls" to "rb," were removed. Since we already have a feature called player\_position, retaining these extra features could result in an overly complex model with redundant information. Consolidating these avoids overfitting and keeps the dataset focused.

* .Age vs. Date of Birth (DOB)

DOB Removal: The dob (Date of Birth) column was removed since we already have age. The use of the dob feature is only to get the age and since we already that feature. Having both would introduce redundancy, as age is calculated directly from the DOB, making DOB unnecessary for our analysis.

* Goalkeeper-Specific Features Removal

Focus on Outfield Players: We removed all features specific to goalkeeping (e.g., gk\_diving, gk\_handling, etc.) to focus solely on outfield players. This ensures the analysis and model are streamlined for attributes relevant to non-goalkeeper roles.

5. Feature Imputation with Averages

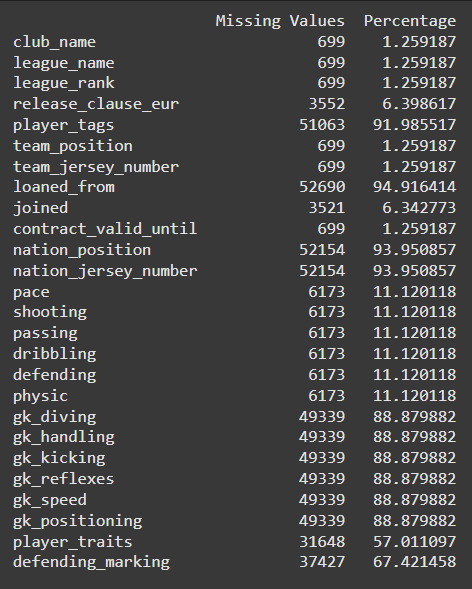
Attacking, Skill, Movement, Power Features: Instead of dropping features with missing values, we applied imputation by averaging attributes within specific skill categories, such as attacking (attack\_avg), skill (skill\_avg), movement (movement\_avg), and power (power\_avg). This approach maintains the value of each group of features by capturing their overall contribution without losing specific performance data. For instance, a high attack\_avg suggests strong attacking ability, allowing for meaningful analysis while reducing feature count.

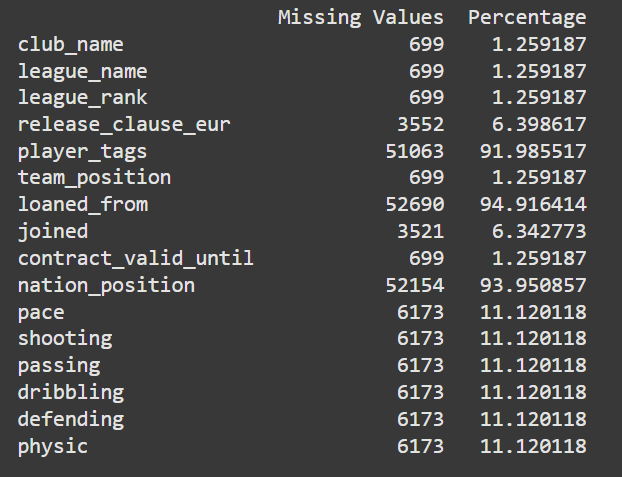
6. Inconsistency Handling in body\_type Feature

Normalization of Body Type: The body\_type feature was inconsistent, with top players such as Ronaldo, Messi, Neymar, and Salah labeled by name rather than physique descriptors like lean, stocky, or normal, as with other players. To resolve this inconsistency and prevent potential bias from named categorizations, we standardized this feature by assigning "other" to all unique or ambiguous entries.

7. Missing values

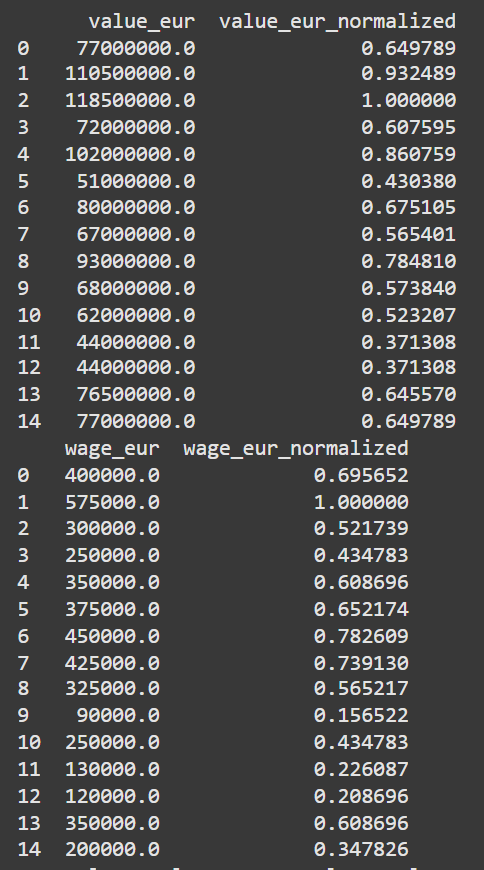
We had many missing values as seen in the fig() To address missing values in our dataset, we applied median imputation for numerical features and mode imputation for categorical features. Using the median for numerical data preserves the central tendency without being skewed by extreme values, ensuring robust fill-ins. For categorical data, the mode maintains the most common category, preserving the feature’s original distribution and minimizing bias introduced through imputation.

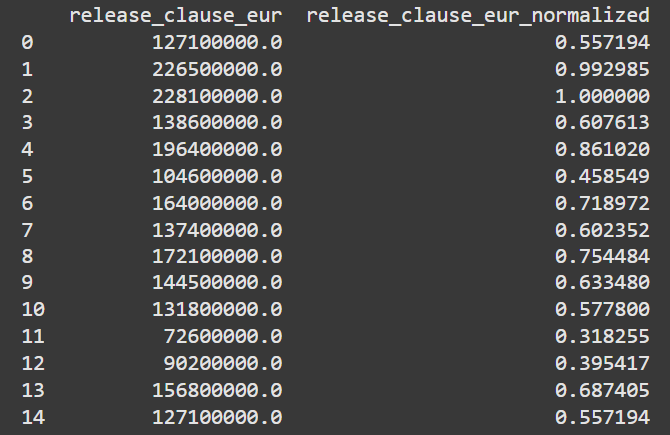
missing values before dropping the features

missing values before handling.

 missing values after handling

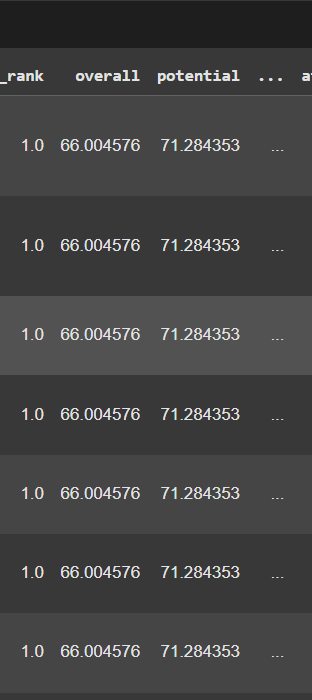
7. Normalization:

We normalized the features `wage\_eur`, `value\_eur`, and `release\_clause\_eur` because they contain large values, often in the 8-digit range. Without normalization, the model might mistakenly prioritize these features due to their scale rather than their importance. To address this, we chose normalization (Z-score), so that the values could be in the range 0 to 1, to ensure that the model interprets these monetary features appropriately.



8. Handling outliers:

Given that the outliers in our dataset represent high-performing, notable players, it’s essential to approach them carefully. These players naturally exhibit standout characteristics having higher skill metrics, physical attributes, or performance stats. Placing them outside the typical range of values, Hence making them outliers. Dropping these outliers would mean losing critical data on some of the most influential players, which would fundamentally alter the dataset's composition, introducing bias, will represent top-tier athletes like average players. Even if we handle these outliers by substituting their values with the mean or with the median would dilute the unique, defining attributes of elite players. For instance, players like Messi and Ronaldo would appear more average, with significant reductions in their characteristic features,(refer to the image Fig. ). This approach would effectively normalize exceptional traits, rendering the dataset misleading as it doesn’t reflect the real range of player abilities. Instead, it's essential to retain these outliers to preserve the dataset’s true variance and maintain the integrity of player comparisons and analyses.



After handling outlier with mean or with median: changes the meaning of dataset as seen in the image

TEXTUAL PREPROCESSING

Columns which are not changed: Long\_name, Club\_name, Nationality, League\_name, Body\_type, Player\_position, Team\_position. (Because of to many unique values)

Preferred\_foot = vectorization (Right =0 , Left =1 )

work\_rate = Special character removal, Created separate column for each type (One hot Encoding). After creating columns dropped work rate.

Player\_tags= Special character removal, Removing punctuation, Convert the TF-IDF matrix to a DataFrame and concatenate it with the original data, Drop the original 'player\_tags' column as we no longer need it.

Dropped = Nation\_position as we already have team\_position and player\_position

Textual Preprocessing