Project Requirement Document

Project Title: FIFA 21 Data Cleaning Task

1. Introduction

1.1 Purpose

The **FIFA 21 Data Cleaning Task** aims to preprocess and refine the FIFA 21 dataset to ensure accuracy, consistency, and usability for further analysis. The raw dataset contains inconsistencies, missing values, incorrect data types, and redundant information. Cleaning this data enhances its reliability for statistical analysis and machine learning applications.

1.2 Scope

This project focuses on transforming the raw FIFA 21 dataset into a well-structured format suitable for data analysis and predictive modeling. The process involves handling missing values, standardizing data formats, removing duplicates, and eliminating outliers.

2. Objectives

2.1 Primary Objectives

1. **Handle Missing Values** – Identify and manage missing values using appropriate strategies (e.g., deletion, imputation).
2. **Correct Data Types** – Convert improperly formatted columns (e.g., numerical values stored as text) to their correct types.
3. **Remove Duplicates** – Ensure each row represents a unique player.
4. **Standardize Data** – Ensure uniform formats (e.g., currency symbols, height/weight units, categorical values).
5. **Drop Irrelevant Features** – Remove unnecessary columns that do not contribute to analysis.
6. **Fix Structural Inconsistencies** – Resolve variations in names, categories, and special characters.
7. **Detect and Handle Outliers** – Identify and address extreme values that may distort analysis.
8. **Validate Cleaned Data** – Recheck for missing values, incorrect data types, and duplicates to ensure integrity.

2.2 Success Criteria

- The dataset is free of missing values and duplicates.
- All numerical and categorical values have consistent formatting.
- Standardized height, weight, and currency values.

- Structural inconsistencies (e.g., country and club names) are resolved.
- Outliers are properly handled using appropriate statistical methods.
- Final dataset is ready for analysis and machine learning applications.

3. Technology Stack

- **Programming Language**: Python 3.x
- **Libraries**:
  - **Data Cleaning**: Pandas, NumPy.
- **Version Control**: Git
- **Development Environment**: Jupyter Notebook

4. Data Cleaning Steps

4.1 Load the Dataset

- Read the dataset from a CSV file.
- Display the first few rows to verify successful loading.
- Use .info() and .describe() to understand the structure.

4.2 Handle Missing Values

- Identify columns with missing data using .isnull().sum().
- Drop columns with excessive missing data (e.g., ‘Loaned From’, ‘Marking’).
- Impute missing numerical values using the mean.

4.3 Remove Duplicates

- Detect duplicate rows using .duplicated() and remove them.

4.4 Convert Data Types

- Identify incorrect data types.
- Convert text-based numerical values (e.g., salaries, heights) to numeric format.
- Format categorical variables appropriately.

4.5 Standardize Formatting

- Convert height values (e.g., “5’9”) to centimeters.
- Convert weight values (e.g., “180lbs”) to kilograms.
- Standardize currency values in wage and market value columns.

4.6 Drop Unnecessary Columns

- Remove irrelevant columns such as "ID", "Flag", and "Club Logo".

4.7 Fix Structural Inconsistencies

- Standardize country, club, and league names.
- Ensure consistent formatting for player positions.

4.8 Detect and Handle Outliers

- Use **Interquartile Range (IQR)** to detect extreme values.
- Apply appropriate capping methods to handle outliers.

4.9 Validate Cleaned Data

- Recheck missing values, data types, and duplicates.
- Ensure dataset integrity for further analysis.

5. Deliverables

1. **Cleaned FIFA 21 dataset** stored as a CSV file.
2. **Jupyter Notebook** containing the complete data cleaning process with code and explanations.
3. **README.md** file explaining the project, objectives, and execution steps.

6. Conclusion

The FIFA 21 Data Cleaning Task ensures that the dataset is clean, structured, and ready for analysis. The processed dataset can be used for statistical modeling, machine learning applications, and player performance analysis.