FIFA in depth analysis with Linear Regression

1. Introduction

In this project, we analyzed a dataset related to football (soccer) players to understand various aspects of player performance and characteristics. The goal was to preprocess the data, perform exploratory data analysis (EDA), and build a linear regression model to predict players' overall ratings based on various features.

2. Data Description

- **Dataset**: The dataset was read from a CSV file (`data.csv`) containing information on football players.
- **Initial Columns**: The dataset included columns such as `Unnamed: 0`, `Photo`, `Flag`, `Club Logo`, `Height`, `Weight`, `Value`, `Wage`, `Position`, and many performance metrics (e.g., `Crossing`, `Finishing`, `SprintSpeed`).
- **Data Types**: Various data types were present including numerical, categorical, and text fields.

3. Steps and Methodology

Data Cleaning

1. **Remove Unnecessary Columns**:
    - Columns like `Unnamed: 0`, `Photo`, `Flag`, `Club Logo`, `Height`, `Weight`, `Value`, `Wage`, and others were dropped as they were either redundant or not useful for the analysis.
2. **Handle Missing Values**:
    - Used `missingno` library to visualize missing values.
    - Identified and removed rows with missing values in critical columns like `Height` and `Weight`.
    - Further dropped columns with too many missing values (`Loaned From`, `Release Clause`, `Joined`).
3. **Data Transformation**:
    - Converted categorical values (like `Value` and `Wage`) into numerical values for better analysis.
    - Created new binary indicator variables for features such as `Real Face` and `Preferred Foot`.
    - Simplified position data into broader categories (`GK`, `DF`, `DM`, `MF`, `AM`, `ST`).

4. **Feature Engineering**:
    - Created binary indicators for major football nations.
    - Expanded `Work Rate` into two separate columns.

Data Analysis

1. **Exploratory Data Analysis (EDA)**:
    - Visualized the distribution of payment methods and player ratings.
    - Analyzed player demographics such as nationality and club.
    - Examined relationships between player attributes like `Age`, `Potential`, `SprintSpeed`, and `Dribbling`.
2. **Correlation Analysis**:
    - Used heatmaps to visualize correlations between features, noting high correlations among certain performance metrics.

Modeling

1. **Prepare Data**:
    - Dropped irrelevant columns and handled missing values.
    - Applied one-hot encoding to categorical features.
2. **Train-Test Split**:
    - Split data into training and testing sets (80% training, 20% testing).
3. **Build and Train Model**:
    - Used Linear Regression to predict the `Overall` rating of players.
    - Evaluated the model using R^2 score and Root Mean Squared Error (RMSE).
4. **Permutation Importance**:
    - Used `eli5` to determine the importance of features in the model, identifying `Potential`, `Age`, and `Reactions` as top contributors.

4. Results

- **Model Performance**:
    - **R^2 Score**: [Insert R^2 score]
    - **RMSE**: [Insert RMSE]
- **Important Features**:
    - The top features affecting the `Overall` rating were `Potential`, `Age`, and `Reactions`.
- **Visualizations**:
    - Joint plots and regression plots demonstrated relationships between `Age`, `Potential`, and other attributes.

5. Conclusion

The linear regression model successfully predicted players' overall ratings based on key performance metrics and attributes. Feature importance analysis highlighted the critical factors influencing player ratings, providing valuable insights for evaluating player performance.

6. Future Work

- **Enhanced Models**:
  - Explore more advanced models (e.g., decision trees, gradient boosting) to potentially improve prediction accuracy.
- **Additional Features**:
  - Incorporate more features or external data (e.g., player performance in different leagues) to enrich the model.
- **Data Collection**:
  - Update the dataset with more recent player data and include additional player statistics for a more comprehensive analysis.
- **Interactive Visualizations**:
  - Develop interactive visualizations using tools like Plotly to better explore the relationships between features and player ratings.

Feel free to adjust the specifics based on the actual results and insights you obtained from your project!