FOOTBALL COLOSSEUM

A R-Shiny App for Football Data Representation and Visualization

Ashmit Bathla, Piyush Kumar, Tattwa Shiwani, Ved Patil Vinay

The shiny app developed in the project provides a one stop comprehensive place to easily view and visualize Both players specific and club specific football statistics From the 2020 to 2024 seasons of the five major European leagues, namely: - English Primer League - Serie A - Bundesliga - Laliga - French League 1

Insights to the Application

The R-shiny app serves multiple purposes, and offers the following features:

- **Squad/Player Analytics** : The most fundamental feature, this part of the app allows a user to view data for a particular club (player) for a specific season. The detailed data is segregated according to the possibles roles a personal takes on the field, these include '**Attack**', '**Defense**', '**Advance**' (general term for what is company called midfield) '**Fair Play**', '**General**' (Only for player specific analysis, includes statistics of the players appearances and time spent on the field.) and '**Win Loss**' (Only for club specific analysis, mainly includes the final league standing of the league for that perticular season.) The user will be prompted to select a season, league, a club or a player and then the kind of statistics he is interested in.
- **Player Comparison** The app facilitates visual comparison of up to 3 players for a given season. The comparisons can be again made according to the the roles which a player can take up, as detailed above. An interesting note to make is that the since the app allows a user to choose any player Irrespective of his position in the squad and the kind of statistic to compare the players, It is possible to make Unorthodox comparisons, like comparing the defensive actions of 2 center forwards, Since they could also be expected to perform defensive duties in case the they are playing with 10 pairs of foots on the field.

Dataset Description

What the Dataset Represents

This dataset captures detailed football statistics from the five major European leagues—Serie A, Bundesliga, LaLiga, Ligue 1, and the English Premier League—for each season from 2020 to 2024. It includes information for all participating clubs and players, allowing for in-depth analysis of performance, attributes, and trends across leagues, clubs, and individual players over these seasons.

Variables in the Dataset

The dataset includes several key variables, organized into categories that reflect the league, club, player, and specific player attributes:

League and Season Information

- **League:** Identifies the league each entry belongs to, including Serie A, Bundesliga, LaLiga, Ligue 1, or the English Premier League.
- **Season:** The season year, ranging from 2020 to 2024, indicating the specific timeframe for each entry.

Club and Player Information

- **Club:** The name of the club to which the player is associated within the specified season.
- **Player:** The name of the player for whom the data is recorded.
- **Position:** The primary playing position of each player (e.g., Forward, Midfielder, Defender, Goalkeeper).

Player Attributes

Each player's performance is captured across various attribute categories, which include:

- **Advanced Attributes:** These are specialized metrics such as passes completed, distance covered, and other advanced statistics that capture a player's overall impact on the game.
- **Attack Attributes:** These attributes represent offensive performance, including statistics such as goals scored, assists, shots taken, and successful dribbles.
- **Defence Attributes:** These variables cover defensive contributions, like tackles, interceptions, clearances, and blocked shots.

- **Goalkeeping Attributes:** Specific to goalkeepers, this includes metrics such as saves made, clean sheets, and goals conceded.
- **Fair Play Attributes:** Attributes that track disciplinary metrics, such as yellow and red cards, fouls committed, and fouls drawn.
- **General Attributes:** General statistics that provide an overall summary of the player's contributions or characteristics.
- **Market Value:** The market value of each player across seasons, providing insight into valuation trends over time and allowing for comparative analysis of player worth.

Purpose of the Dataset

This dataset is designed for use in performance analysis, player comparison, and trend visualization. Specifically, it allows for:

- Comparative analysis of player performance across different leagues and seasons.
- In-depth examination of individual player attributes for up to three players across various clubs and seasons.
- Market value trend analysis, showing how the value of players changes over time and across different leagues.
- Evaluation of performance trends by league, club, or player position from 2020 to 2024.

Overall, this dataset provides a robust foundation for examining European football performance data, player valuation trends, and statistical comparisons across multiple dimensions.

We have got 96 club for a season across 5 seasons. Also, there are a total of 13.5k players across all the clubs in these 5 seasons

Data Scraping Overview

This project involves scraping data from the FBRER website and converting it into various `.csv` and `.RData` files. The main script, `script.R`, is responsible for scraping and collecting the data, while additional scripts handle file conversions for further processing.

We also used data available on transfermarkt website to obtain the market value of each player across all the seasons for all the football players. This website doesn't allow scraping so, we had to get market values manually for all the players

Scripts Overview

1. **script.R** - This is the main data scraping script. It accesses the target website, retrieves the necessary data, and saves it as `.csv` files in the `/Data` directory. This script ensures the data is collected in a structured format suitable for conversion.
2. **AdvanceScript.R**, **AttackScript.R**, **DefenceScript.R**, **GoalKeepingScript.R** - These scripts process the `.csv` files generated by `script.R`. Each script converts specific `.csv` files into `.RData` files, as required for different analyses. This modular approach ensures each data set is available in the R environment in the desired format, organized by metric (e.g., Advanced, Attack, Defense, Goalkeeping).

Script Details

1. `script.R` - Main Data Scraping Script

- **Objective**: Scrapes data from FBRER for five major European leagues (e.g., Bundesliga, La Liga) and saves the data in structured `.csv` files.
- **Steps**:
    1. **Loading Required Packages**: Utilizes `tidyverse`, `rvest`, `dplyr`, `httr`, and `curl`.
    2. **Directory Setup**: Establishes directories based on league and season to organize the scraped data.
    3. **Data Extraction**: Loops through seasons, sends HTTP requests to retrieve HTML content, and extracts tables for each league-season combination.
    4. **Data Cleaning and Saving**: Cleans and saves the tables as `.csv` files.

2. `AttackScript.R` - Attack Metric Processing

- **Objective**: Processes `.csv` files to extract and structure attacking metrics, such as goals, assists, shot accuracy, and shot-creating actions, saving the data in `.RData` format for advanced analytics.
- **Steps**:
    1. **Loading Packages**: Uses `dplyr` for data manipulation.
    2. **Directory and File Setup**: Defines helper functions to locate player data files across leagues and seasons.
    3. **Metric Selection**: Selects columns related to attacking performance, such as "Goals+Assists," "Shots on Target Percentage," and "Shot Creating Actions."
    4. **Data Aggregation**: Iterates over leagues and seasons, reads each `.csv` file, selects relevant columns, and appends them to a combined data frame.

5. **Saving in `.RData` Format**: The combined data frame is saved as an `.RData` file, allowing for fast loading and compatibility within the R environment.

3. Additional Metric Scripts (`AdvanceScript.R`, `DefenceScript.R`, `GoalKeepingScript.R`)

- **Objective**: Similar to `AttackScript.R`, these scripts handle metrics specific to other aspects of performance, ensuring that each data set is saved in `.RData` format.
- **Functionality**:
  - **AdvanceScript.R**: Focuses on advanced metrics such as progressive passes, ball carries, and dribbles.
  - **DefenceScript.R**: Concentrates on defensive metrics like tackles, interceptions, and clearances.
  - **GoalKeepingScript.R**: Handles goalkeeping stats, such as saves, clean sheets, and save percentage.
- **Approach**: Each script follows the same general process of reading `.csv` files, selecting relevant columns, and saving the processed data as `.RData`. # Identifying any Biases in the Data

In analyzing data from the 2020 to 2024 seasons of the top five European football leagues—**English Premier League**, **Serie A**, **Bundesliga**, **LaLiga**, and **French Ligue 1**—several potential biases could affect the accuracy and fairness of insights. Below are some key biases to consider:

1. Selection Bias

- **Description**: The data only covers the most recent five years, potentially excluding historical trends, long-term player performance, or older market influences.
- **Impact**: Insights may be less applicable to players whose peak years or significant impacts occurred before 2020.

2. Survivorship Bias

- **Description**: Focusing on players who were active from 2020 to 2024 may overlook those who transferred to other leagues or retired during this period.
- **Impact**: This could skew the analysis by making some players look better or worse due to natural player turnover not reflected in the data.

3. Market Value Prediction Bias

- **Description**: Market values are influenced by factors beyond performance (e.g., media attention, club reputation, and market trends), which might not be fully captured by performance metrics.
- **Impact**: Predictions could be biased toward high-profile players or leagues, potentially inflating or undervaluing players based on external factors not reflected in the data.

4. Position-Specific Bias

- **Description**: Comparing players across all positions can inherently favor certain positions (e.g., forwards who score more goals) over others (e.g., defenders or goalkeepers).
- **Impact**: This could lead to skewed analysis if players in less "visible" roles are not compared within their specific positions.

5. League-Specific Bias

- **Description**: Each league has unique styles of play, competitive levels, and tactical approaches, impacting player and team statistics.
- **Impact**: Direct comparisons across leagues may introduce bias, as player stats might vary due to league

Interesting Questions to ask from the Data

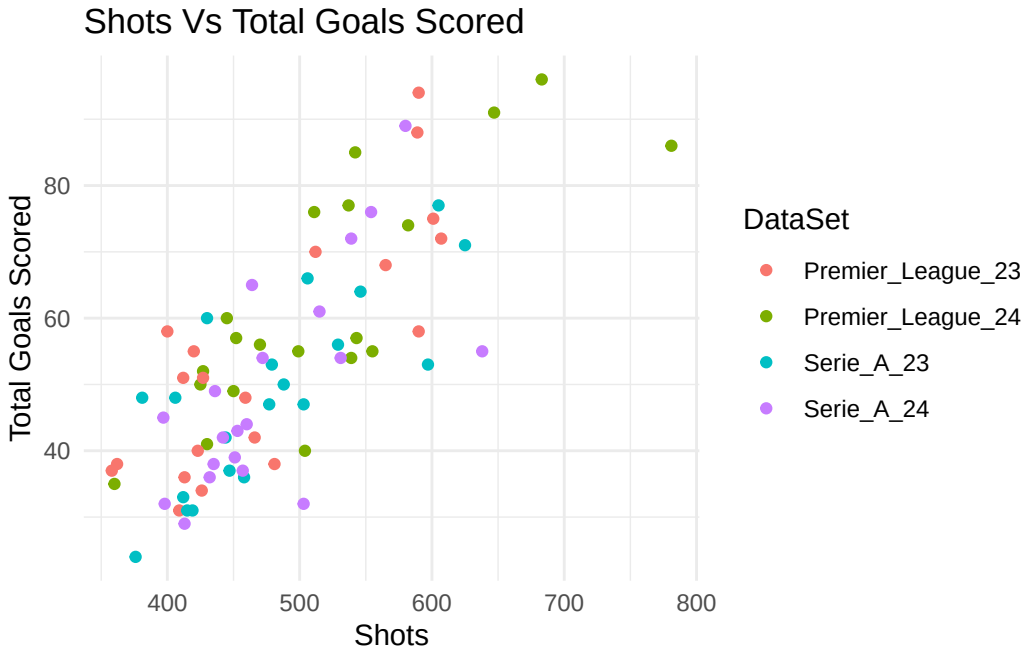
Analyzing data from the 2020 to 2024 seasons of the major European leagues provides insights into various aspects of team and player performance. Below are some key questions this data can help answer:

1. **Player Performance**
    - Who are the top-performing players across different positions in each league?
    - Which players have shown the most improvement over the last five seasons?
2. **Team Dynamics and Strength**
    - Which teams have consistently ranked at the top across the five seasons?
    - How do teams compare in terms of offensive and defensive statistics?
3. **Market Value Trends**
    - How has the market value of individual players evolved over the past five years?
    - Are there noticeable trends in market values for certain positions or nationalities?

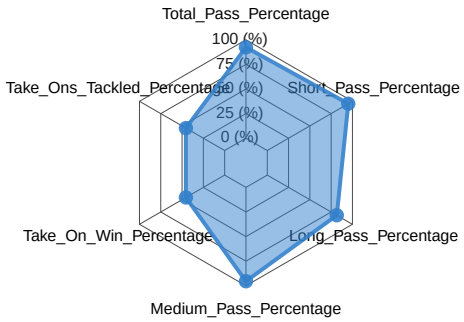
4. **League Comparisons**
    - How do player and team statistics vary across the five major leagues?
    - Which league demonstrates the highest average player performance or team efficiency?
5. **Player Comparison**
    - How do individual players compare in terms of performance metrics such as goals, assists, and defensive actions?
    - Which players have the highest influence on their team's success?

Visualizations

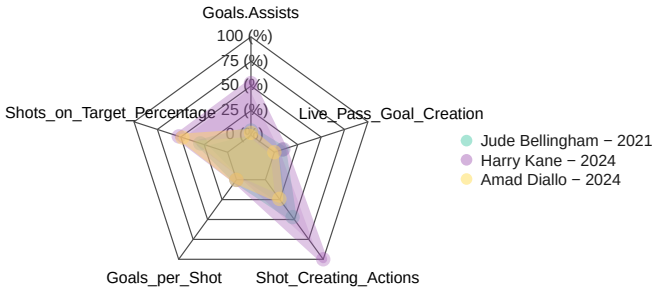
A scatter plot visualizes statistics from the past five league seasons, allowing for clearer visual interpretation of trends and patterns. Also, we can adjust the variables at X and Y-axis so that we can observe the patterns and relation between any two stats for any season and any League



A spider chart allows us to understand the attributes of a player for any player of any Club for any Season and League in a very efficient way

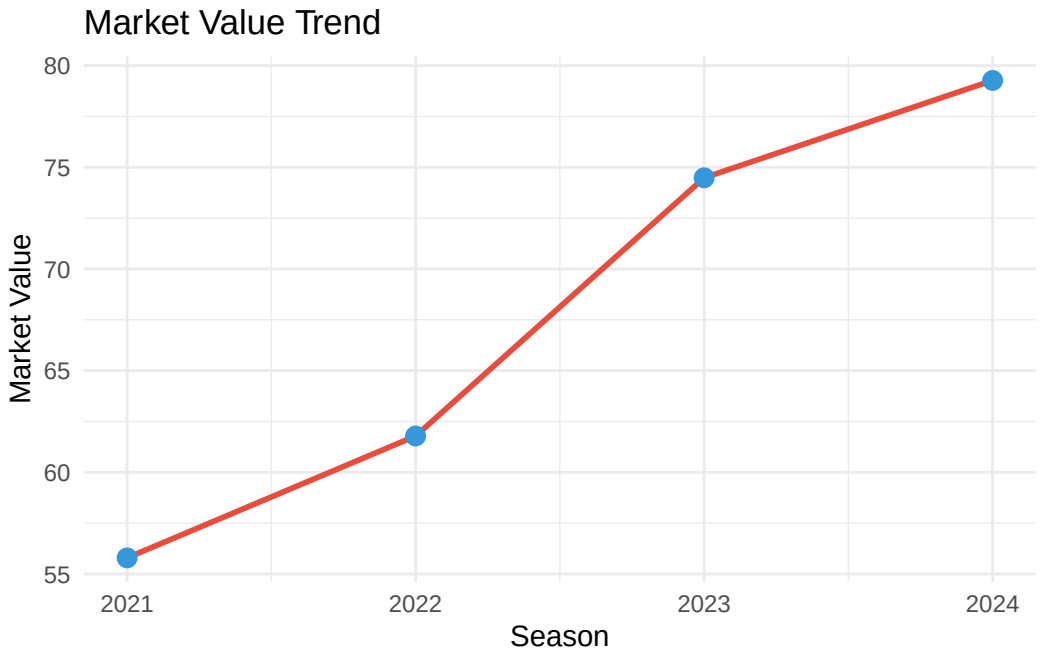


We can also compare the performance of any three players from matches across the past five seasons and from any club



We also have plotted the market value of each Football player across the past five seasons,

which helps us predict the future market value of the player by the next or following seasons. It also helps us in better understanding how players have performed and how their market value has affected in the past seasons



Final Conclusions

Summary of Insights

Our analysis of the 2020 to 2024 seasons across the five major European football leagues—**English Premier League**, **Serie A**, **Bundesliga**, **LaLiga**, and **French Ligue 1**—has provided valuable insights into team performance, player dynamics, and market value trends. Through various visualizations and statistical comparisons, we gained a clearer understanding of:

- **Player Performance**: Identifying standout players in each position, as well as those with consistent improvement over multiple seasons.
- **Team Comparisons**: Highlighting teams with strong offensive and defensive records and analyzing their consistency across seasons.
- **Market Value Trends**: Observing the factors influencing player market values, including position, nationality, and league, and predicting future trends.

Future Directions

This project opens up several avenues for further analysis and potential improvements, including:

- **Longitudinal Analysis**: Expanding the dataset to include more seasons could help capture long-term trends and reduce recency bias.
- **Advanced Predictive Models**: Refining prediction models for player market value using machine learning techniques could enhance accuracy and reveal new influencing factors.

Conclusion

Overall, this project highlights the power of data-driven analysis in sports, providing both fans and analysts with a deeper understanding of European football leagues. Our findings can aid teams, scouts, and investors in making informed decisions and may also serve as a foundation for further research into the evolving world of football.

As football continues to grow and evolve, data analysis will play an increasingly critical role in understanding player and team performance, enhancing the fan experience, and driving strategic decisions across the industry.

References

1. FBREF website - this source contains all the data of past seasons of the five major European leagues of Football
2. transfermarkt website - this source contains data required to predict the market value of the players
3. For Shiny App, we used information available on these sites : https://shiny.posit.co/py/components/, https://shiny.posit.co/py/templates/ and https://shiny.posit.co/py/layouts/