# Project Documentation

## English Premier League Performance Analysis (2018–2023)

---

## DEPI – Data Analysis Program

## Project Overview:

This project provides a comprehensive analysis of English Premier League (EPL) performance data covering the seasons from 2018 to 2023. It aims to uncover meaningful KPIs, trends, and performance patterns that support informed decision-making for coaches, analysts, and sporting directors. By examining both team-level and player-level metrics, the project evaluates offensive and defensive performance across multiple seasons to highlight areas of strength, weakness, and potential improvement.

The analysis is designed to transform raw match and player data into actionable insights through statistical exploration, comparative performance assessment, and time-series analysis. The final output will include fully interactive dashboards and well-structured data reports that present key findings, visual summaries, and strategic recommendations to enhance overall team performance and support data-driven sports management.

---

## Project Scope:

This project focuses on delivering a full end-to-end analytical solution for evaluating English Premier League performance trends between the 2018 and 2023 seasons. The scope defines the boundaries, activities, and deliverables required to address the management's need for data-driven insights into player and team evolution over time.

### In-Scope Activities

1. **Data Acquisition & Preparation**
    - Collecting historical player and team data from 2018–2023.
    - Cleaning and preprocessing datasets, resolving missing values, duplicates, and inconsistencies.
    - Structuring datasets into unified formats suitable for comparative analysis.
2. **Exploratory & Comparative Analysis**
    - Conducting exploratory data analysis to identify distributions, trends, and anomalies.

- Comparing player performance across seasons to track progression, regression, and consistency.

- Analyzing offensive and defensive metrics across teams, clubs, and player positions.

3. **KPI Development & Insights Generation**

    - Calculating key metrics such as goals, assists, clean sheets, player rankings, and team performance indicators.

    - Identifying top scorers, key playmakers, reliable defenders, and emerging talents.

    - Mapping business questions to relevant data points and generating analytical insights to answer them.

4. **Dashboard & Visualization Design**

    - Developing interactive dashboards using Power B

    - Visualizing trends, KPIs, season-to-season comparisons, and club/player performance insights.

    - Ensuring dashboards are intuitive, fast-loading, and suitable for non-technical decision-makers.

5. **Reporting & Documentation**

    - Creating structured reports summarizing findings, interpretations, and recommendations.

    - Documenting the entire workflow including data sources, methods, KPI formulas, and analytical logic.

    - Delivering final documentation and presentation materials for stakeholders.

## Project Deliverables

- Clean and structured datasets (2018–2023).

- Exploratory data analysis summary.

- Season-by-season performance comparison reports.
- KPI tables and ranking models for players and teams.
- Interactive dashboards (player-level & club-level).
- Final comprehensive report with actionable recommendations.
- Full project documentation and a quality-checked final presentation.

### Stakeholders

- Team Management
- Coaches
- Performance Analysts
- Sporting Directors
- Data Analysts / BI Specialists

### Success Criteria

- 100% dataset integrity ensured after cleaning and preprocessing.
- At least 90% of business questions answered through analytical insights.
- Dashboards load in under 3 seconds and are fully navigable for at least 80% of users.
- A minimum of three actionable recommendations grounded in data.
- Complete, clear, and professional final documentation delivered by the deadline.

---

## Project Requirements:

The requirements define what is needed to successfully execute the Premier League performance analysis project and deliver high-quality insights, dashboards, and documentation.

## 1. Business Requirements

1. **Performance Understanding**
   Management must gain a clear, data-driven understanding of how player and team performance evolved between 2018 and 2023.

2. **Decision-Making Support**
   The system should provide insights to guide decisions related to:

   - Player transfers
   - Training priorities
   - Tactical evaluations
   - Performance monitoring across seasons

3. **KPI Identification**
   Extract key indicators such as goals, assists, clean sheets, and defensive metrics to highlight top performers and performance patterns.

4. **Insight Delivery**
   Create visual and written outputs that are easy for non-technical stakeholders (coaches, directors) to understand and use.

## 2. Functional Requirements:

These specify what the system **must do**:

1. **Data Collection & Preparation**

   - Gather player and team data for all seasons from 2018 to 2023.
   - Clean, merge, and standardize datasets to ensure consistency across seasons.

2. **Exploratory Data Analysis**

   - Identify trends, distributions, and anomalies in team and player performance.
   - Generate season-by-season comparisons.

3. **KPI Calculation**

   - Compute offensive KPIs: goals, assists, shots, goal contributions.

- Compute defensive KPIs: clean sheets, tackles, interceptions, defensive errors.
- Generate player rankings and team-level performance metrics.

4. **Insight Generation**
    - Identify top scorers, playmakers, and consistent defenders.
    - Detect performance trends across players, clubs, and positions.
    - Answer all business questions through structured analysis.
5. **Dashboard Development**
    - Build interactive dashboards using Power BI
    - Display KPIs, trends, club comparisons, and player statistics.
    - Ensure dashboards are intuitive and load in under 3 seconds.
6. **Reporting & Documentation**
    - Produce a final analytical report summarizing key findings.
    - Provide actionable recommendations based on data insights.
    - Document the full methodology, data sources, and KPI definitions.

## 3. Technical Requirements:

1. **Tools & Languages**
    - **Python (NumPy, Pandas)** for data processing.
    - **Excel, CSV, SQL** for data storage and organization.
    - **Power BI** for visualization.
    - **DAX, Matplotlib, Seaborn** for analytical computations and visual analysis.
    - **GitHub** for version control.
2. **Hardware/Software Environment**
    - Stable Python environment with required libraries.
    - BI tools installed (Power BI Desktop or Tableau Public).

## 4. Data Requirements:

1. **Historical Dataset Coverage**
    - Player-level and team-level data for seasons 2018–2023.
    - Includes offensive, defensive, positional, and match-level statistics.
2. **Data Quality Standards**
    - Unified naming conventions.
    - Standardized fields across seasons.
    - Clean, complete, and validated datasets.

---

# Data Sources:

The data used in this project was collected from **FBref**, a reputable and publicly accessible football statistics platform that provides detailed player and team performance metrics for major leagues worldwide. FBref is widely recognized for its comprehensive coverage of match statistics, advanced metrics, and season-by-season breakdowns.

## Primary Data Source

- **FBref – English Premier League Statistics (2018–2023)**
  The dataset includes historical records for players, teams, clubs, and match events across six seasons. FBref offers structured, season-based tables covering offensive, defensive, passing, goalkeeping, and possession metrics.

## Data Acquisition Method

To gather the required data efficiently and accurately, a **Python-based Web Scraping pipeline** was developed. The scraping process involved:

1. **Extracting season-level tables** for:
    - Player statistics
    - Team performance metrics
    - Offensive and defensive KPIs
    - Match-by-match data

2. **Using Python libraries** such as:
    - **Requests** – for sending HTTP requests
    - **BeautifulSoup / lxml** – for parsing and extracting HTML table content
    - **Pandas** – for structuring, cleaning, and exporting the data into CSV/Excel files
3. **Automating multi-season extraction**, ensuring consistent formatting across:
    - 2018–2019
    - 2019–2020
    - 2020–2021
    - 2021–2022
    - 2022–2023
4. **Validating the scraped data** by checking:
    - Column consistency across seasons
    - Accurate KPI calculations
    - Removal of duplicates and missing values

### Final Data Storage

After extraction and cleaning, the datasets were stored in:

- **CSV files** for raw and preprocessed data
- **Excel files** for manual review and KPI checks
- **SQL tables** for structured querying and season-based comparisons

---

## Methodology / Approach:

The project follows a structured, data-driven methodology designed to transform raw Premier League statistics into actionable insights for decision-makers. The approach combines systematic data collection, rigorous analysis, and intuitive visualization to ensure accuracy, clarity, and high-value outcomes.

## 1. Data Collection

- Gathered comprehensive Premier League data (2018–2023) from **FBref** using a Python-based web scraping pipeline.
- Extracted season-level, player-level, and team-level tables, ensuring coverage of offensive, defensive, passing, and match statistics.
- Stored raw data in CSV and Excel formats for traceability and version control.

## 2. Data Cleaning & Preprocessing

- Standardized dataset formats to ensure consistency across all seasons.
- Handled missing values, removed duplicates, and resolved formatting inconsistencies.
- Normalized column names and merged multi-season datasets into a unified analytical structure.
- Created data quality checks to confirm integrity and completeness.

## 3. Exploratory Data Analysis (EDA)

- Conducted initial data exploration to understand distributions, patterns, and anomalies.
- Generated descriptive statistics and visual summaries to identify early trends.
- Explored relationships between player performance metrics and team outcomes.
- Highlighted outliers and season-specific shifts in performance.

## 4. Business Question Mapping

- Translated business objectives into measurable analytical questions.
- Linked each question to the relevant dataset attributes or KPIs (e.g., goals, assists, clean sheets).
- Created a clear analytical roadmap to ensure that all management needs are addressed.

## 5. Feature Engineering & KPI Development

- Calculated key offensive KPIs (goals, assists, goal contributions, shots).
- Calculated defensive KPIs (tackles, interceptions, clearances, clean sheets).
- Built ranking frameworks for players and teams using standardized scoring logic.
- Generated season-over-season comparison metrics to assess performance changes.

## 6. Trend & Performance Analysis

- Compared player and team performance across six seasons to identify improvements, declines, and stability.
- Analyzed positional trends (defenders, midfielders, forwards) and club-level patterns.
- Identified top performers: scorers, playmakers, reliable defenders, emerging talents.
- Extracted actionable insights based on statistical evidence.

## 7. Dashboard Design & Visualization

- Designed interactive dashboards using **Power BI / Tableau** for intuitive decision support.
- Visualized KPIs using charts, scorecards, heatmaps, and season comparison visuals.
- Ensured fast load times (<3s) and user-friendly navigation for non-technical audiences.
- Built both player-level and team-level dashboards for multi-layer insights.

## 8. Reporting & Documentation

- Consolidated all findings into a comprehensive analytical report.
- Included visual summaries, interpretations, and data-backed recommendations.
- Documented the full workflow: data sources, cleaning steps, models, KPI formulas, and dashboard logic.

- Conducted a final quality review to ensure accuracy, coherence, and professional presentation.

### 9. Final Review & Delivery

- Validated all insights and visualizations against original data sources.
- Performed peer review and refinement to ensure clarity and reliability.
- Delivered final dashboards, datasets, and documentation to stakeholders.
- Prepared the project for future scalability (adding new seasons, integrating more metrics).

---

## Tools Used:

A diverse set of tools and technologies were utilized throughout the project to ensure efficient data collection, processing, analysis, and visualization. These tools supported each stage of the workflow, from data acquisition to dashboard development and documentation.

### 1. Data Extraction & Processing

#### Python

Used extensively for automating data collection and transforming raw datasets into structured analytical tables.

- **Libraries:**
    - **Requests** – Sending HTTP requests for web scraping
    - **BeautifulSoup / lxml** – Parsing HTML tables from FBref
    - **Pandas** – Data cleaning, merging, KPI calculations
    - **NumPy** – Mathematical operations and performance optimization

### 2. Data Storage & Management

#### Excel / CSV

- Used for storing raw and cleaned datasets.
- Enabled quick data validation and manual quality checks.

**SQL**

- Used for structured storage and running performance comparisons through SQL queries.
- Facilitated efficient filtering, joining, and aggregation of multi-season datasets.

## 3. Data Visualization & Dashboarding

**Power BI**

- Primary BI tool for creating interactive dashboards.
- Used for visualizing KPIs, trends, and season-over-season performance.

**Visualization Libraries (for EDA)**

- **Matplotlib** – Line charts, histograms, and distribution plots
- **Seaborn** – Heatmaps, correlation matrices, and advanced visualizations

## 4. Analytical & Computation Tools

**DAX (Power BI)**

- Used for creating calculated measures, KPIs, and time-intelligence functions within dashboards.

**Python Analytics**

- Custom scripts for generating advanced metrics, normalization, and trend analysis.

## 5. Collaboration & Version Control

**GitHub**

- Hosted all scripts, datasets, and documentation.
- Ensured version control, change tracking, and collaborative workflow.
- Provided a centralized repository for reproducibility and transparency.

## 6. Documentation & Reporting

**Microsoft Word / Google Docs**

## Results / Findings:

The analysis of Premier League data (2018–2023) revealed key insights into player and team performance:

1. **Player Performance Trends**
    - Identified top scorers, assist leaders, and consistent defenders across seasons.
    - Highlighted players who showed improvement or decline in performance metrics over multiple seasons.
    - Detected emerging talents with consistent contributions relative to team performance.
2. **Team-Level Insights**
    - Ranked teams based on offensive and defensive KPIs, such as goals scored, clean sheets, and goal difference.
    - Tracked performance trends over time, identifying clubs with stable performance versus volatile seasons.
3. **Positional Analysis**
    - Offensive and defensive metrics revealed key patterns by position, including high-impact forwards and reliable defensive lineups.
    - Midfielders with consistent playmaking contributions were highlighted for strategic decisions.
4. **Interactive Dashboard Outcomes**
    - Dashboards provided intuitive visualizations for decision-makers.
    - Users can easily compare player and team performance across seasons.
    - Enabled quick identification of top performers, performance trends, and key metrics for strategic planning.

## Challenges & Solutions:

**Challenge:** Inconsistent data formats across seasons.
**Solution:** Standardized column names and merged datasets into a unified structure using Python and Pandas.

**Challenge:** Missing or incomplete data for certain players and matches.
**Solution:** Implemented data cleaning procedures including imputation, removal of duplicates, and cross-verification with secondary sources.

**Challenge:** Large volume of multi-season data.
**Solution:** Optimized data storage using CSV, Excel, and SQL, and automated processing pipelines to reduce manual work.

**Challenge:** Translating raw data into actionable insights for non-technical stakeholders.
**Solution:** Designed interactive dashboards with clear visualizations, KPIs, and season comparisons for intuitive decision-making.

**Challenge:** Ensuring dashboard performance and responsiveness.
**Solution:** Optimized queries and visuals to achieve load times under 3 seconds and simplified navigation for users.

---

## Conclusion & Recommendations:

### Conclusion

The project successfully transformed historical Premier League data into actionable insights for management, coaches, and analysts. By systematically collecting, cleaning, and analyzing data from 2018–2023, the project:

- Provided a clear understanding of player and team evolution.
- Identified key performers and emerging trends.
- Delivered interactive dashboards that summarize KPIs, trends, and comparisons across seasons.

### Recommendations

1. **Strategic Player Decisions:** Use the insights to guide player transfers, contract renewals, and development programs.

2. **Training & Development:** Focus training on areas where performance trends indicate weaknesses or inconsistencies.

3. **Recruitment & Scouting:** Leverage data-driven rankings to identify emerging talents or undervalued players.

4. **Continuous Monitoring:** Update dashboards annually to track ongoing performance and maintain historical comparability.

5. **Enhanced Metrics:** Consider integrating additional performance indicators such as xG, expected assists, or advanced positional data for deeper analysis.

---

## Thank You

**We would like to sincerely thank the DEPI Data Analysis Program for providing us with this invaluable learning opportunity. The guidance, knowledge, and resources offered through the program have greatly enhanced our skills in data collection, analysis, visualization, and reporting.**

**This project would not have been possible without the structured support and mentorship provided by DEPI, which enabled us to apply practical data analysis techniques to real-world datasets and gain hands-on experience in building actionable insights.**

**We are truly grateful for the opportunity to grow as data professionals and for the encouragement to explore, learn, and deliver high-quality analytical work.**

**Thank you for your trust, support, and inspiration throughout this journey.**