Project documentation focused on team and player performances during the 2025/26 Big 5 season

Project introduction

The objective of this web application is to visualize the performances of both teams and players during the 2025/26 season. Player data come from Opta The Analyst, FBref, and Transfermarkt. This analysis therefore focuses on the 2025/26 season for the following competitions: Ligue 1, Bundesliga, Premier League, La Liga, and Serie A.
More concretely, this application allows users to evaluate each team's style of play and to identify a player of their choice based on their statistical performances throughout the current season.

Data extraction and data preprocessing

- Team

Given the data requirements related to team performances during the current season, we collected information from several data providers.

Regarding **Opta**, we scraped data from the five major leagues by retrieving the following tables: **Attack, Defense, Pressing, Sequence, and Other**. All statistics from these tables are collected, except for redundant ones, and renamed using the name of their respective table as a prefix. In addition, these data are **updated weekly** via GitHub Actions for each league.

We then structured our database using statistics available on **FBref,** which offers a wide range of advanced indicators covering offensive and defensive phases, possession, pressing, and goalkeeper actions. FBref organizes its data into multiple categories: **Standard Stats, Shooting, Passing, Pass Types, Goal and Shot Creation, Defensive Actions, Possession, Playing Time, Miscellaneous Stats**, among others.

To integrate these data into our model, we first retrieved team data via Python scraping. Historically, the entire collection process was automated, but we now rely on a semi-manual extraction assisted by a Python script in order to ensure better control over data cleaning and structuring. This process includes a cleaning phase during which we corrected league names, harmonized various text variables, and removed redundant columns such as Rk, Squad, Comp, etc.

We also integrated team salary data to add an economic dimension to the analysis, as well as each team's league position.

We then joined all these variables at the team level using a **mapping based on team names** in order to obtain a complete database. We also created several derived variables from the previously collected information, such as league rankings, the percentage of attacking actions by style of play for each team, and possession-adjusted defensive metrics.

From this dataset, we established a category-based scoring system according to the available statistics. These categories are divided into four distinct groups:

- **On-ball statistics**: Chances created, Finishing, Set pieces (offensive), Chance creation, Projection, Crosses, Dribbles.
- **Off-ball statistics**: Chances conceded, Defensive actions, Set pieces (defensive), Goalkeeper efficiency.
- **Style of play statistics**: Direct play, Counter-attacks, Possession, Pressing.
- **Other**: League position, Ground duels, Aerial duels, Fouls drawn, Fouls committed, Ball losses, Substitutions.

Each of these **subcategories** contains several associated statistics, each with a **coefficient** reflecting its importance in defining the group. It should be noted that the choice of these metrics and coefficients was made jointly by the two project members and is therefore entirely subjective.

Similarly, each statistical category has a coefficient used to compute an overall team rating between 0 and 100. This rating determines the Power Ranking among the 96 Big 5 teams. Finally, we accounted for league strength by applying a **slight penalty based on the league power index** calculated by Opta in May 2025.

- Player

For individual player analysis, we built our database using the dataset available on **Kaggle**: https://www.kaggle.com/datasets/hubertsidorowicz/football-players-stats-2025-2026

Previously, these data were also obtained via scraping, but this process is no longer used. We now rely directly on this dataset, ensuring greater consistency and reduced noise in the raw data.

An in-depth cleaning phase was then carried out to harmonize and validate the variables. The main operations included:

- **Position standardization**: abbreviations were replaced with their full forms to obtain clearer categories, for example:
  - DF -> Defender
  - MF -> Midfielder
  - FW -> Forward
  - GK -> Goalkeeper
- **Correction and standardization of country names**, replacing abbreviated forms with full names.
- **Duplicate removal**, keeping a single record per player.

In addition, we integrated salary data from **Capology**, enriching the database with a key economic dimension for performance and player value analysis.

To further complete our dataset, we also integrated data from **Transfermarkt** using the unofficial API available at:
https://github.com/felipeall/transfermarkt-api

Following the provided instructions, we installed **Docker** and configured the necessary environment to query the API. This process allows us to first retrieve the list of clubs and then the players belonging to each club. Additional requests are then performed to obtain detailed information on each player.

Transfermarkt data are particularly rich and provide a level of precision not offered by other sources. Among the most important variables integrated into our model are:

- The **player's natural position**, described in a granular manner (for example Left Winger, Right Back, Attacking Midfielder), rather than a general category.
- **Representation data** (agent or agency managing the player).
- **Injury** history and current injury status.
- A set of contextual information including:
    - player name (player_name)
    - position (position)
    - date of birth (dateOfBirth)
    - age (age)
    - nationality (nationality)
    - current club (currentClub)
    - height (height)
    - preferred foot (foot)
    - year joined current club (joinedOn / joined)
    - previous club (signedFrom)
    - contract end date (contract)
    - market value (marketValue)
    - status (indicating whether the player is injured, suspended, or captain)

Based on these data, we **aggregated certain statistics per 90 minutes and adjusted others according to team possession**. The goal of possession adjustment is to objectively analyze defensive actions as if the player's team had 50 percent possession. We then joined the two data providers using several strategies to match as many players as possible while minimizing errors.

The first step consisted of matching players with the same name, same league, same year of birth, and same position when the player was a goalkeeper, as this is the only position with no discrepancies between providers. This step matched most players, but name variations required further processing. We therefore used a textual similarity function (fuzzy matching) following this logic: if the similarity percentage is high enough, combined with the same league, same year of birth, and same position for goalkeepers, then we consider it to be the same player.

Finally, we manually matched players whose names differed significantly between providers but whose other information matched perfectly. It should be noted that the lower the name similarity percentage, the higher the risk of error. For this reason, we did not go below a 65 percent threshold to ensure reliable matching.

We then established a **player rating system** between 0 and 100, both per category and overall, based on the performances of other players in the same position and position group. A high rating reflects a strong level of performance during the 2025/26 season. For consistency, some variables were inverted, such as errors and ball losses. Each position also has its own coefficients adapted to its specific role. For example, a high-performing forward must show strong finishing and chance creation statistics. Similarly, each statistical category has a weight designed to highlight players performing well in key dimensions. Goalkeepers have their own dedicated rating system due to the specific nature of their position. Finally, penalties may be applied to the overall rating depending on league strength and the percentage of minutes played during the season.

List of pages

The application is divided into two main parts: team analysis and player analysis.

Team pages (6 pages)

- **Home**: Project presentation and data sources.
- **Team analysis**: Detailed exploration of the selected team using various statistics.
- **Team comparison**: Side-by-side comparison of two teams.
- Ranking - Basic statistics: Team rankings based on a selected simple statistic.
- **Ranking - Advanced statistics**: Team rankings based on a selected advanced statistic.
- **Power Ranking**: Overall ranking of the 96 Big 5 teams, with league-based rankings.

Player pages (6 pages)

- **Home**: Project presentation and data sources.
- **Player analysis**: Detailed analysis of the selected player using various statistics.
- **Player comparison**: Comparison of two players playing in the same position.
- **Ranking - Basic statistics**: Player rankings based on a selected simple statistic.
- **Ranking - Advanced statistics**: Player rankings based on a selected advanced statistic.
- **Scouting**: Search for players matching user-defined criteria (general information, basic statistics, advanced statistics).

Project structure for application development

```
Application
├── scripts (team, player, big_5_performance.py): All project scripts by type
├── data (team, player): All project data grouped by type
├── image: Images used in the project
├── README.md
├── .github: All workflows used for data automation
└── requirements.txt: Library dependency file
```

Application setup

It is important to note that the application will be available in three versions: French, English, and Spanish.

Home

As explained previously, the home page briefly presents the project components and provides access to various resources such as documentation and the project source code.

Analysis section

Project header

We will first create several functions (analysis tables, glossary, translations, etc.) as well as various variables (glossary, data lists by category) in order to avoid excessive project length, given the three available languages and the similarities across analysis types.

Application display

- **Team**

For **team analysis**, the user must first select a league in order to choose the team they wish to analyze. The analysis is structured into several sections: general team information (name, league, ranking, power ranking), display of the top 5 players of the season based on our performance criteria, list of the five most valuable players according to Transfermarkt, and several comparative charts (offensive, defensive, style of play, others) positioning the team relative to the Big 5 or its league. A list of five similar teams based on available statistics is also provided.

The **Duel** section allows the user to compare two selected teams by confronting their statistics to facilitate comparative analysis.

The **Stats+** section provides rankings based on aggregated statistics by selected category (chance creation, finishing, dribbles, etc.), with optional league filtering.

The **Stat** section follows the same logic but applies to all available raw statistics (xG, PPDA, counter-attacks per 90 minutes, etc.). The user must first select the statistic category to ease navigation. A glossary is also available to explain each statistic.

Finally, the **Power Ranking** section displays the ranking of the 96 teams, with optional league filtering and quick comparison of their respective statistics.

- **Player**

For **player analysis**, the user is first asked to select a player. This page displays the player profile with basic information (name, photo, position, club, etc.), a statistical radar, and a table of the five most statistically similar players. The radar can be adjusted by country, league, or age group. The statistics shown on the radar are those most relevant to the player’s position, displayed in blue, and compared with the median of the selected group, displayed in red. A statistics glossary is also available and expandable if needed.

The **player comparison** page follows the same logic, allowing the user to select two players in the same position. Their profiles are displayed side by side, followed by their respective radars in blue and red.

On the **basic statistics ranking** page, the user can obtain rankings by selected category (finishing, dribbling, etc.), with optional filters such as position, age group, or country. A podium of the best players according to these criteria is displayed, along with general information on all players matching the user's request.

Similarly, on the **advanced statistics ranking** page, the user selects the statistic to analyze. Based on this choice, a ranking of the best players according to this metric is displayed, with a podium for the top three players, followed by a table containing their basic information. The sidebar includes optional filters (position, club, league, age group, market value) as well as a statistics glossary. An image is displayed before the user's selection to avoid leaving the page empty.

Finally, the **Scout** page provides a list of players based on the criteria selected by the user (basic information, basic statistics, advanced statistics). A summary of the selected filters is available in the sidebar, along with a statistics glossary.