Football Data Analysis Report: English Premier League (2025/2026 Season)

---

1. Introduction

This report outlines the analysis of English Premier League (EPL) football teams for the 2025/2026 season. It involved gathering data from the football-data.org API (team details) and BBC Sport (league standings via web scraping), followed by data cleaning, merging, exploratory analysis, and visualization of key metrics.

---

2. Data Sources

- **Football Data API (football-data.org):**
  - **Endpoint:** /v4/competitions/PL/teams
  - **Data:** Team Name, Short Name (TLA), Founded Year, Stadium, Country. Saved as Premier_League_Teams.csv.
- **BBC Sport Website (Web Scraping):**
  - **URL:** /sport/football/premier-league/table
  - **Data:** League table (Team, P, W, D, L, GF, GA, GD, Pts, Form). Saved as premier_league_table.csv.

---

3. Methodology

1. **Data Acquisition:** Fetched JSON data from the API and scraped/parsed HTML table data from BBC Sport using requests and BeautifulSoup.
2. **Data Cleaning & Preparation:**
   - Removed duplicates from both datasets.
   - Renamed columns for consistency (ShortName to Team, Goals For/Against to GF/GA).
   - Cleaned team names scraped from BBC (removed leading position numbers).
   - Filled potential missing values (NaN) in the scraped table with 0.
   - Converted relevant columns in the scraped table to integer types.
3. **Data Merging:** Performed a left merge (pd.merge) of the scraped table (df2) with the API data (df) on team names. Handled potential NaN values from merge mismatches by filling numeric columns with the median and object columns with "Unknown". The final cleaned DataFrame was named clean.
4. **Exploratory Data Analysis (EDA):** Calculated descriptive statistics (describe()) and computed a correlation matrix for key numerical metrics (Points, Wins, GF, GA, GD, Losses, etc.).
5. **Data Visualization:** Created a correlation heatmap, box plots for metric distributions, a scatter plot (GF vs. GA), and a bar plot (Points per team) using matplotlib and seaborn.

---

4. Technologies Used

- **Language/Environment:** Python 3, Jupyter Notebook
- **Libraries:** pandas, numpy, requests, beautifulsoup4 (bs4), matplotlib, seaborn

5. Key Findings and Visualizations

1. **Correlation Analysis:**
   - Strong positive correlations were observed between **Points** and key performance indicators like **Wins**, **Goals For (GF)**, and **Goal Difference (GD)**.
   - Strong negative correlations were found between **Points** and **Losses** / **Goals Against (GA)**.
   - The **founding year** showed negligible correlation with current performance.

- *Visualization: Correlation Heatmap.*

2. **Columns Distributions:**
   - Box plots for Points, GF, GA, GD, Wins, and Losses showed reasonable distributions without significant outliers, suggesting consistent data.
   - *Visualization: Box Plots.*
3. **Attack vs. Defense (GF vs. GA):**
   - The scatter plot visually confirmed that teams with high GF and low GA (top-left quadrant) generally rank higher in the league.
   - *Visualization: Scatter Plot GF vs GA.*
4. **Team Standings:**
   - The bar plot provided a clear ranking of teams by points, reflecting the current league table. Arsenal led the table based on the scraped data.
   - *Visualization: Bar Plot of Points per Team.*

---

6. Conclusion

This project successfully integrated data from an API and web scraping to analyze EPL team performance. The EDA confirmed expected correlations between match outcomes, goals scored/conceded, and overall points. The visualizations provide a clear snapshot of team performance and league standings for the 2025/2026 season as per the data collected. The workflow demonstrates a practical approach to combining and analyzing data from varied web sources.