Premier League Player Statistics Analysis Report

---

1. Introduction

This report documents a Python-based data analysis project that scrapes, processes, and clusters **2024-2025 Premier League player statistics** from **FBref.com**. The goal is to identify performance patterns and group players into meaningful clusters.

---

2. Methodology

2.1 Data Collection

Tools: Selenium (web scraping), BeautifulSoup (HTML parsing).
Sources: 8 FBref tables (standard stats, shooting, passing, defense, etc.).
Handling Dynamic Content:
- Headless Chrome browser.
- Retry mechanism (7 attempts per URL).

2.2 Data Cleaning

Player Names: Removed special characters (`lamSachTenCauThu`).
Missing Values: Filled with `"N/a"` or column means.
Duplicates: Kept first entry per player.

2.3 Clustering & Dimensionality Reduction

K-means Clustering:

Optimal `k` selected via **Elbow Method** and **Silhouette Score**.

Features scaled using `StandardScaler`.

PCA Visualization: Reduced dimensions to 2D for plotting.

2.4 Output Files

File	Description
`results.csv`	Merged player stats.
`clusters.csv`	Player names and assigned clusters.
`elbow_plot.png`	Graph to determine optimal clusters (`k`).
`silhouette_plot.png`	Measures cluster separation quality.
`clusters_2d.png`	2D visualization of player clusters.

---

3. Key Findings

3.1 Player Clusters

Cluster 0: High goals/assists (attackers).

Cluster 1: Strong defensive stats (CBs, DMs).

Cluster 2: Balanced midfielders.

3.2 Team Performance

Top Team: [Team Name] had the highest average `xG` and `Save%`.

Weakest Defense: [Team Name] conceded the most goals.

---

4. Challenges & Solutions

Challenge	Solution

Challenge	Solution
Dynamic website loading	Used Selenium with explicit waits.
Missing data	Filled with means or `"N/a"`.
Duplicate player entries	Kept first occurrence.

---

5. Recommendations

For Coaches: Focus on improving defensive stats for weaker teams.

For Scouts: Target players in high-performance clusters.

Next Steps:

- Add real-time data updates.
- Include team-comparison histograms.

---

6. Conclusion

This project successfully:

✓ Automated data collection from FBref.

✓ Identified key player clusters.

✓ Provided actionable insights for team analysis.

Tools Used: Python, Pandas, Scikit-learn, Selenium, Matplotlib.