DATA 602 Project Proposal

Football Player Scouting and Style Adaptability Analysis

Joao De Oliveira

1. Research Question

Which player performance metrics best predict adaptability to a team's tactical style, and how can data analysis help identify suitable transfer targets for specific positions?

2. Justification

Modern football (soccer) scouting increasingly relies on analytics to supplement traditional human-based observation. Clubs aim to sign players whose attributes align with their tactical philosophy, style of play, and role requirements. This project simulates a data-driven scouting model to help identify players who statistically fit a target playing style (for example, full-backs suited to possession-based systems or forwards effective in transition). From an industry perspective, this kind of analysis replicates how professional recruitment departments and platforms like Wyscout, StatsBomb, and FBref use machine learning and similarity metrics to shortlist ideal players. This project is personally meaningful because it combines my interests in data science and football analysis in a practical, real-world application.

3. Data Sources

My plan is to explore various data sources to gather player performance data from different professional leagues.

- FBref (https://fbref.com/en/) is an open data platform that provides teams and players' performance data.

- Kaggle football datasets (https://www.kaggle.com/datasets/?search=football) https://www.kaggle.com/datasets/vivovinco/20222023-football-player-stats

- Football-Data.org for live and recent player stats.

Among these, I will likely focus on FBref or the Kaggle dataset as the primary data source for consistency and availability, while the Football-Data API will probably be used for comparison and data enrichment.

4. Libraries to Be Used

pandas, numpy, matplotlib, seaborn, scikit-learn, requests (API data), plotly (for interactive visuals), and possibly sqlalchemy or sqlite3 for local database storage.

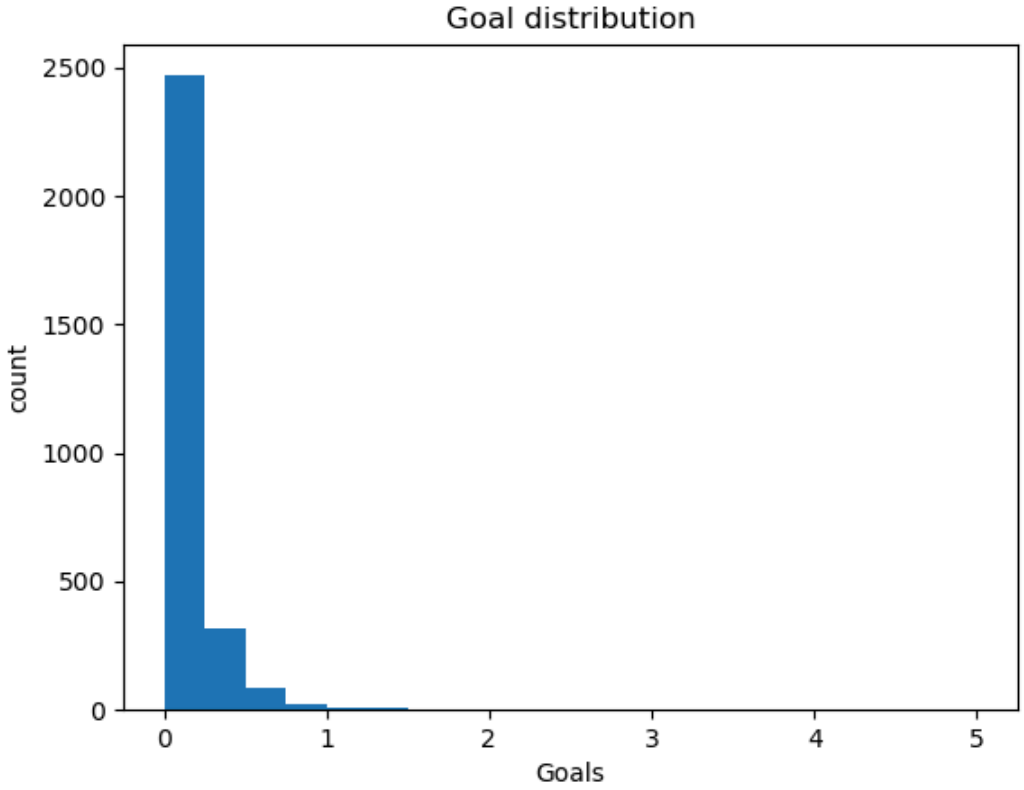
5. EDA & Summary Statistics (illustrative)

project_proposal.py

```python
import pandas as pd
import matplotlib.pyplot as plt

raw_url = "https://raw.githubusercontent.com/JDO-MSDS/DATA602/refs/heads/main/Project/2021-2022%20Football%20Player%20Stats.csv"
df = pd.read_csv(raw_url, encoding="latin1", sep=";")

print(df.head())

# metrics
num_df = df.select_dtypes(include=["number"])
num_df.describe().T


if "Goals" in df.columns:
    plt.figure()
    df["Goals"].dropna().plot(kind="hist", bins=20)
    plt.title("Goal distribution")
    plt.xlabel("Goals")
    plt.ylabel("count")
    plt.show()
```

Run: project_proposal

```
C:\Users\joao\python.exe C:/Users/jpfer/PycharmProjects/python_course_2/project_proposal.py
   Rk             Player Nation Pos  ... Recov AerWon  AerLost  AerWon%
0   1         Max Aarons    ENG  DF  ...  5.53   0.47     1.59     22.7
1   2   Yunis Abdelhamid    MAR  DF  ...  6.77   2.02     1.36     59.8
2   3  Salis Abdul Samed    GHA  MF  ...  8.76   0.88     0.88     50.0
3   4    Laurent Abergel    FRA  MF  ...  8.87   0.43     0.43     50.0
4   5        Charles Abi    FRA  FW  ...  4.00   2.00     0.00    100.0

[5 rows x 143 columns]
```



Preliminary observation:
Players in full-back and midfield positions exhibit the highest variance between progressive passing and defensive actions, suggesting these attributes could be key differentiators for identifying adaptability to possession-heavy systems. One analysis that makes sense would be a cluster analysis grouping players by style profiles. Another approach is to focus on computing distances between player attribute vectors to find replacements for specific players/positions. A predictive model might also make sense for estimating fit scores based on the team's playing style.