DATA 602 - Fall 2025

Data-Driven Soccer Scouting

Identifying Ball-Playing Center Backs

DATA 602 Final Project
João De Oliveira

CUNY SPS

Abstract

- Use player performance data to support defender scouting
- Focus on ball-dominant and vertical build-up systems
- Examine passing, ball control, and progression metrics
- Create a style-fit score and similarity-based shortlist
- Find full abstract by clicking here: **Abstract**

Research Question

- Which defender metrics reflect tactical adaptability?
- How can data analysis help identify suitable transfer targets?

Tactical Style of Interest

- Ball dominant build up
- Comfort under pressure
- Vertical progression from defense

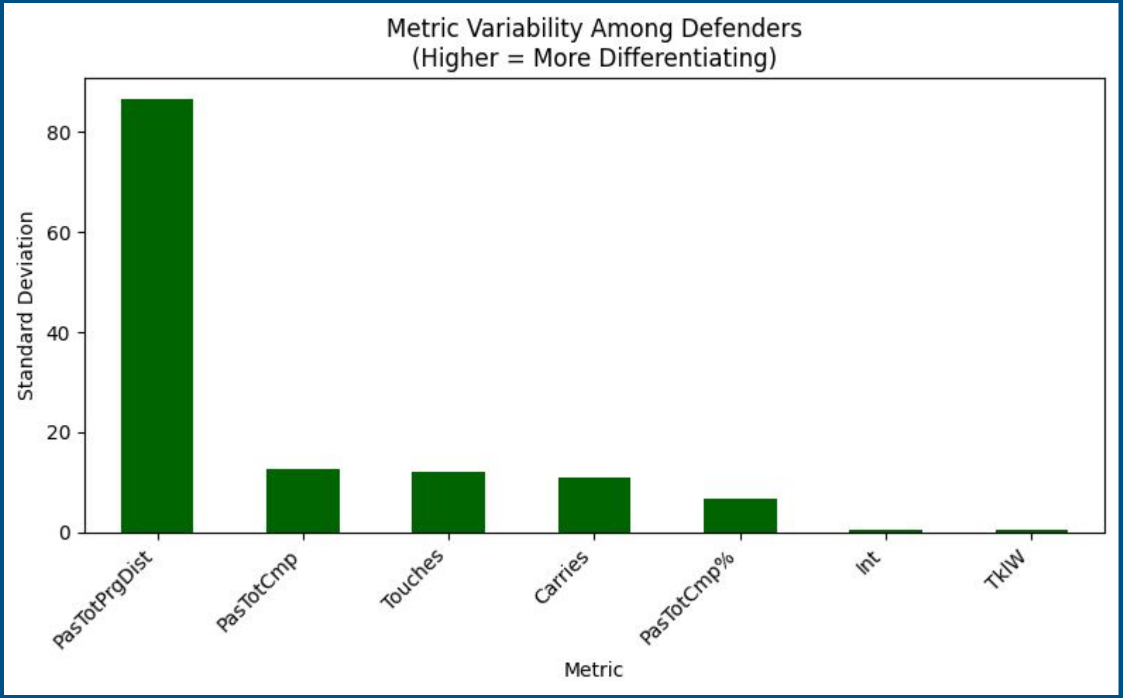
Data Overview

- 2021-2022 European player statistics
- Publicly available FBref data extracted from Kaggle
- Defenders with at least 900 minutes played
- Encoding cleanup and missing-value handling

Data: https://www.kaggle.com/datasets/vivovinco/20212022-football-team-stats

Which Metrics Matter?

- Progressive passing distance
- Passing accuracy and volume
- Ball carries and touches

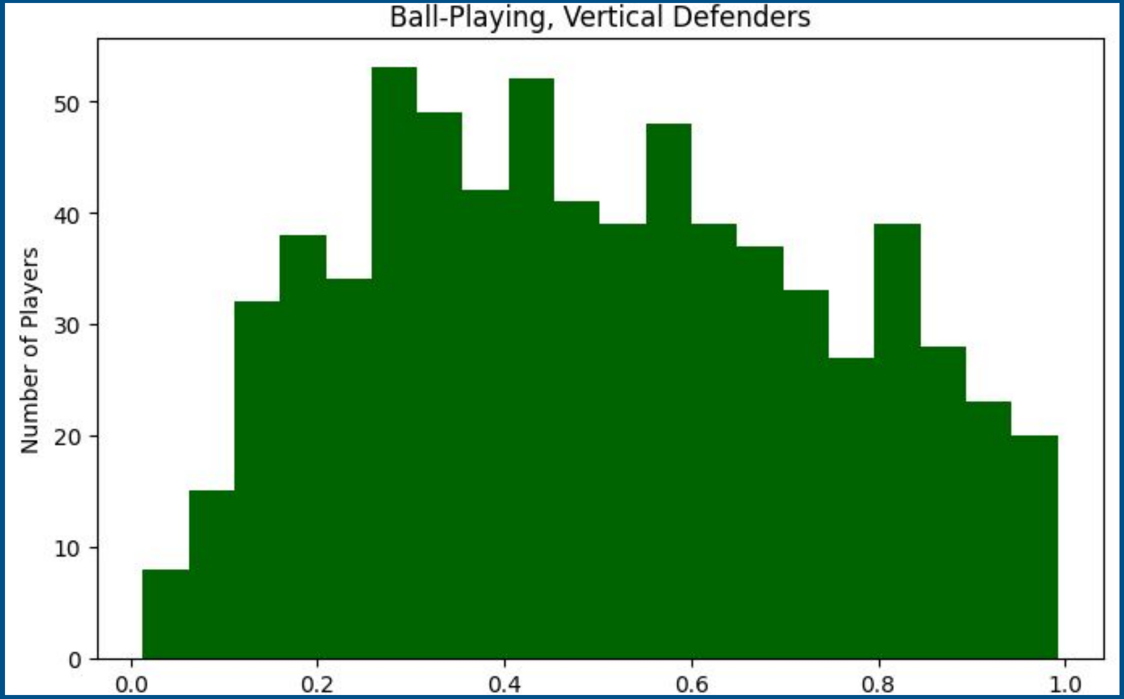


Style Fit Score

- Composite score using:
  - Progressive passing
  - Passing accuracy
  - Carries
  - Touches

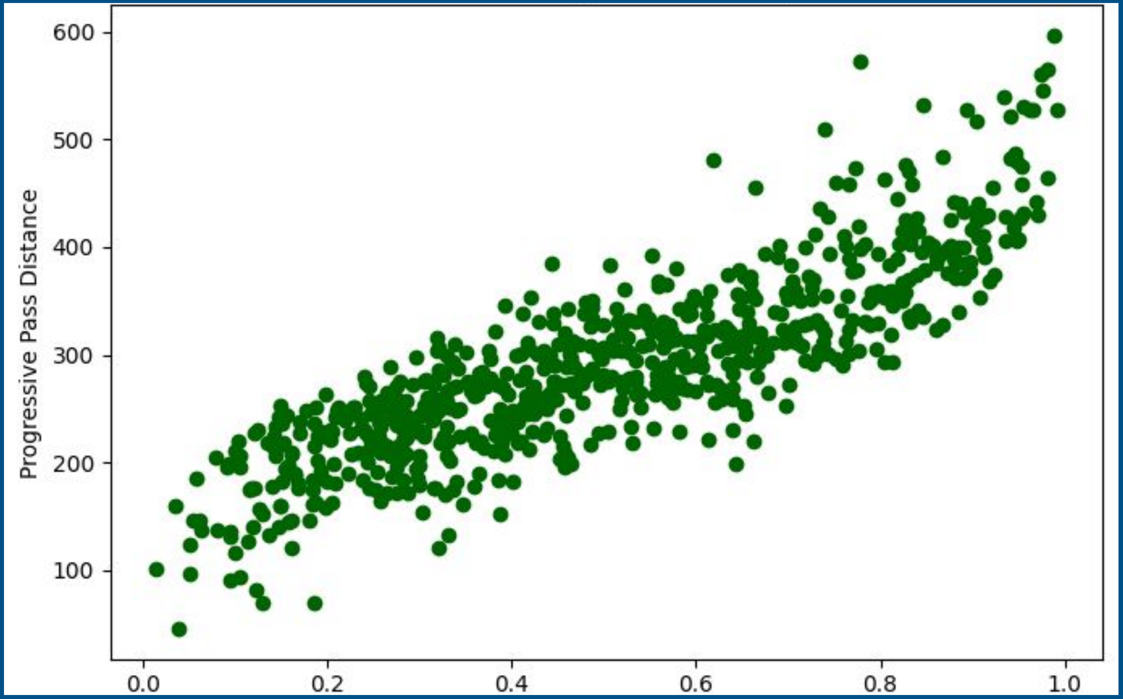
Distribution of Style Fit Scores

- Most defenders score moderately
- Few defenders strongly match the profile



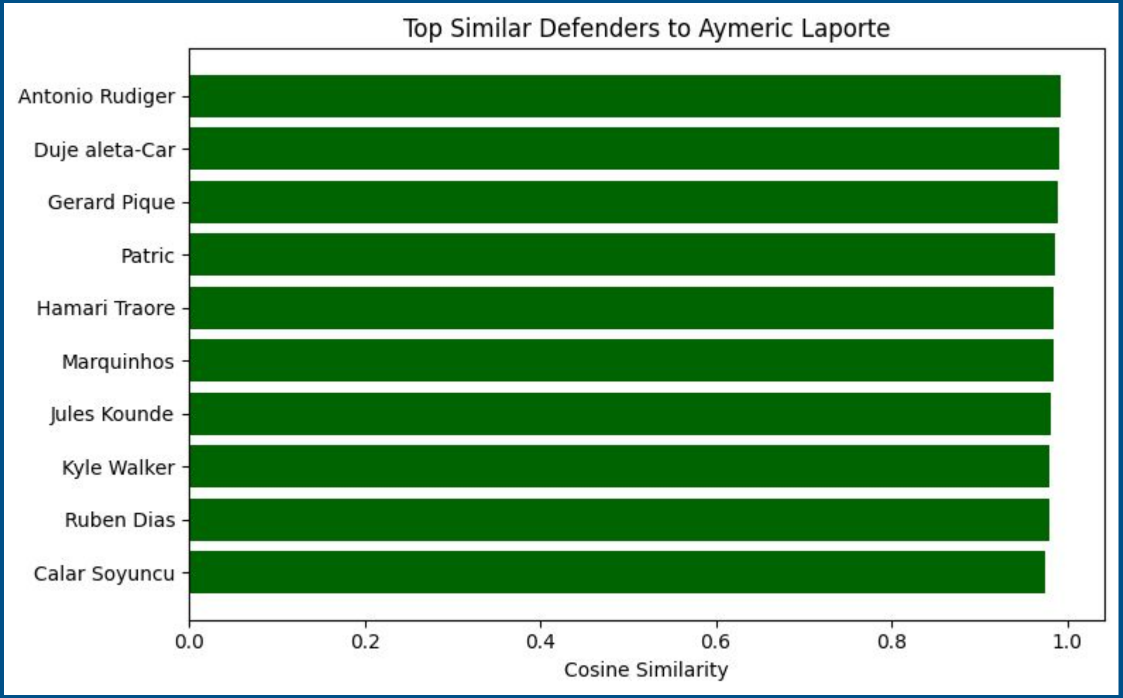
Validating the Style Fit Score

- Higher scores align with greater progressive passing
- Confirms metric selection



Similarity-Based Scouting

- Cosine similarity used
- Compared to a reference defender (Aymeric Laporte)
- Identifies comparable transfer targets



Conclusions

- Passing and ball-involvement metrics best distinguish defender styles
- Progressive passing is a strong indicator of tactical fit
- Style Fit Score helps narrow large defender pools to a focused shortlist
- Similarity analysis identifies realistic replacement or transfer options

Limitations

- Based on one season of data (no long-term consistency)
- Tactical fit inferred from statistics, not direct tactical data
- Does not include financial or squad-specific constraints

Thank you!

João De Oliveira

https://colab.research.google.com/drive/1LMwjKYCpbU7kKzz5iCgtAmGLH1DQg-oz?usp=sharing