Matt Gurgiolo May 20, 2024

# The Top 5 European Leagues:

# Exploring Player Value and Team Performance Through Clustering

**Abstract**

In this study, we examine a novel approach to understanding soccer player roles by clustering players based on their functional contributions rather than traditional positions. Utilizing comprehensive player statistics from the top five European leagues during the 2022-23 season, we identified distinct player clusters. Our analysis revealed a significant correlation between the composition of these clusters within a team's top players and the team's league performance. Employing regression analysis, we demonstrated that teams with certain player clusters performed better, introducing a different technique for analyzing team and player performance. In addition, we attempted to analyze the wages of players based on their cluster assignment, but little evidence was found.

## Introduction

European soccer's top five leagues are renowned for their high level of play and competitiveness. Because of this, they also maintain highly detailed and accurate records for all player actions in all games. This study leverages the detailed player and team data from these leagues' 2022-2023 season to explore how teams might optimize performance not by traditional positions but through functional roles of players. By clustering players based on diverse statistics reflecting their on-field actions, we aim to redefine how strategic team compositions can be analyzed and understood. In other words, are positions are the best way to define players, or can we understand their contributions to their team detached from where on the field they initially line up?

**Data**

The primary data was sourced from fbref.com, which provided a rich set of player performance metrics aggregated over the season. Additional data on player positions and league standings were acquired from Fotmob.com and fbref.com, respectively. This was accomplished using the *pandas* "read_html" function and the *selenium* webscraping package. The dataset encompassed various performance indicators such as goals, assists, defensive actions, and more detailed metrics like progressive passes and shot-creating actions, covering nearly every measurable aspect of a player's game during the season. Below are all the different statistics with descriptions that I used in my analysis.

| Stat Name | Description |
|---|---|
| Name | Player name |
| Squad | Team played for |
| Pos | Position |
| Min | Minutes played |
| Tkl | Tackles |
| Tkl% | Tackle Win Percentage |
| Blk | Blocks |
| blkSh | Blocked Shots |
| blkPass | Blocked Passes |
| Int | Interceptions |
| Clr | Clearances |
| Gls | Goals |
| Ast | Assists |
| npxG | Non-Penalty Expected Goals |
| xAG | Expected Assists |
| PrgC | Progressive Carries |
| PrgR | Progressive Receptions |
| Fls | Fouls |
| Fld | Fouls Drawn |
| Recov | Ball Recoveries |
| aerWon | Aerial Duels Won |
| aerLost | Aerial Duels Lost |
| passAtt | Passes Attempted |
| Cmp% | Pass Completion Percentage |
| KP | Key Passes |

| | |
|---|---|
| pass1/3 | Passes into the Final Third |
| PPA | Passes into the penalty Aerial |
| CrsPA | Crosses into the Penalty Aerial |
| PrgP | Progressive Passes |
| Sh | Shots |
| SoT% | Shot on target Percentage |
| Dist | Average Shot Distance |
| FK | Number of Free Kicks taken |
| SCA | Shot Creating Actions |
| PassLive | SCA from live passes (not FK's) |
| scaTO | SCA from Take-ons |
| scaDef | SCA from Defensive Actions |
| Touches | Touches |
| Attdrib | Attempted Dribbles |
| Succdrib | Successful Dribbles |
| Carries | Carries |
| CPA | Carries into the Penalty Area |
| Rec | Receptions |

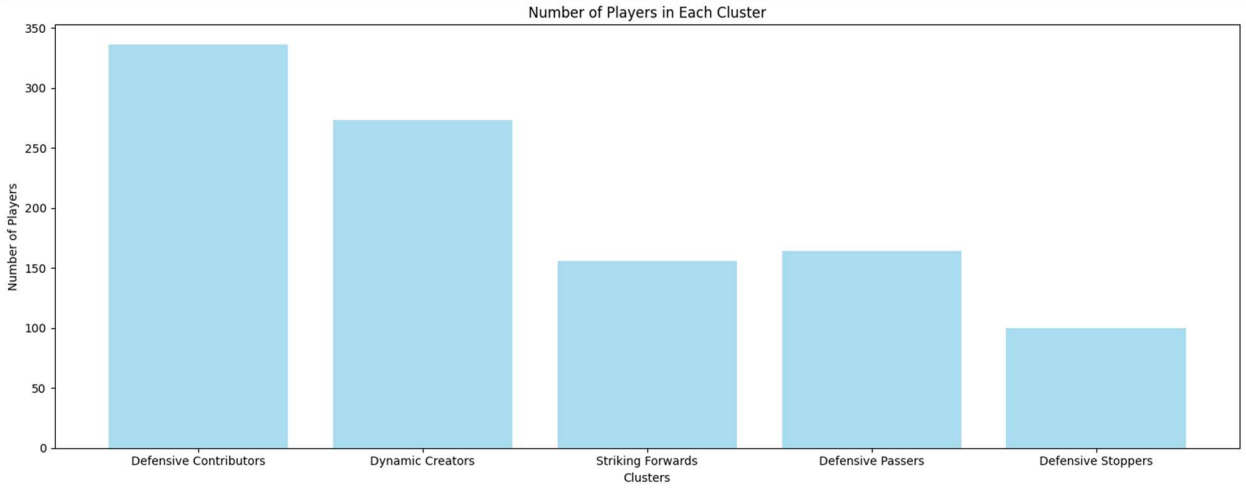
To facilitate better analysis, I cleaned the data, removing all players with less than 900 minutes, the equivalent of 10 games. This is a common threshold in soccer statistics. I also removed all goalkeepers since their role is so different from other players, and I removed all players who transferred to a new team mid-season because this could show a change in role under a different manager with different teammates. This left me with 1029 players to cluster.

**Analysis**

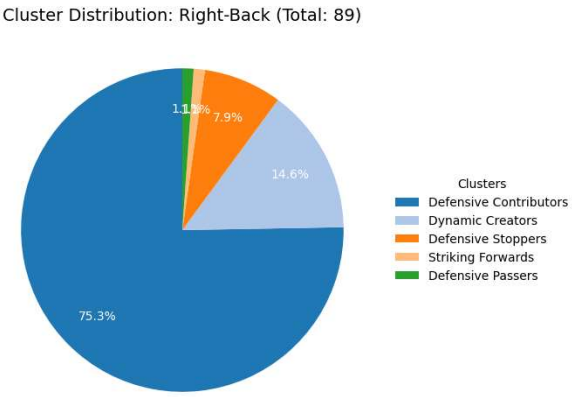
The analysis was conducted using Python within a Jupyter Notebook environment. Data manipulation and analysis were primarily handled using Pandas and NumPy. The *sklearn* library facilitated principal component analysis (PCA), though we did not end up clustering using the data transformed by the principal components. Using K-means clustering from *sklearn* as well, we categorized players into 5 functional groups (k = 5). The players were clustered based on their stats as percentiles across all players, normalized within each player. This allowed me to adjust for how

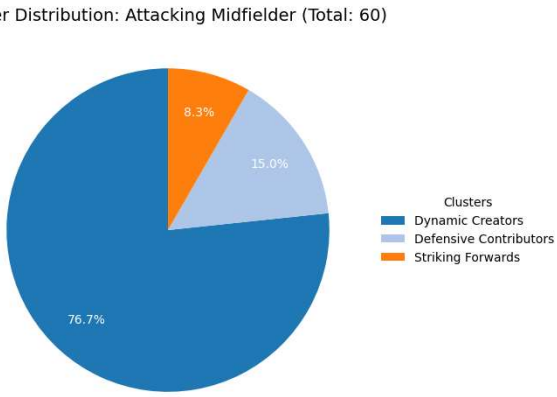
good players were. I assigned names to each cluster and described their distributions of stats in the table below, with the number of players in each cluster below that.

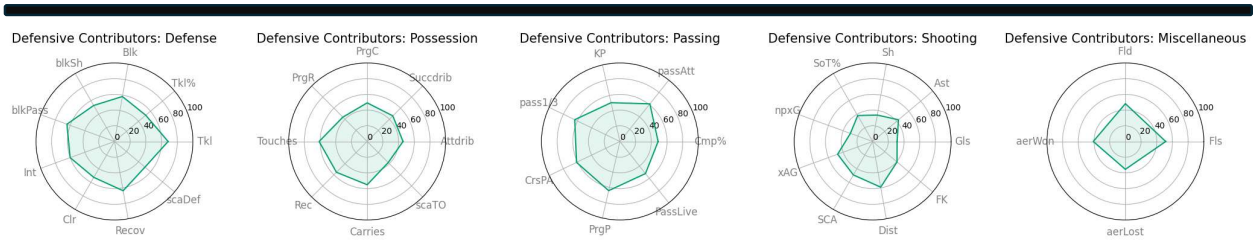
| **Cluster Name** | **Description** |
|---|---|
| Defensive Stoppers | Lots of defensive actions (tkl, blkShot, clr) and little passing numbers. |
| Dynamic Creators: | Progress the ball a lot, and create chances to score. |
| Striking Forwards: | Lots of shots, Goals, and attacking dribbles. |
| Defensive Contributors: | Does a bit of everything, with an emphasis on well rounded defensive stats. |
| Defensive Passers: | Similar to Defensive Stoppers, with much more progression and passing. |

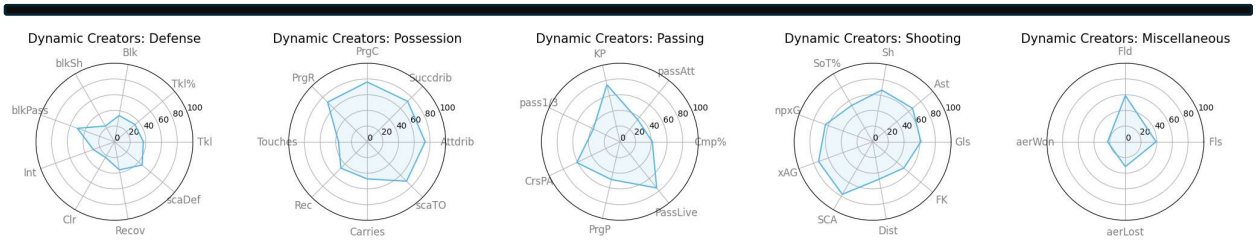


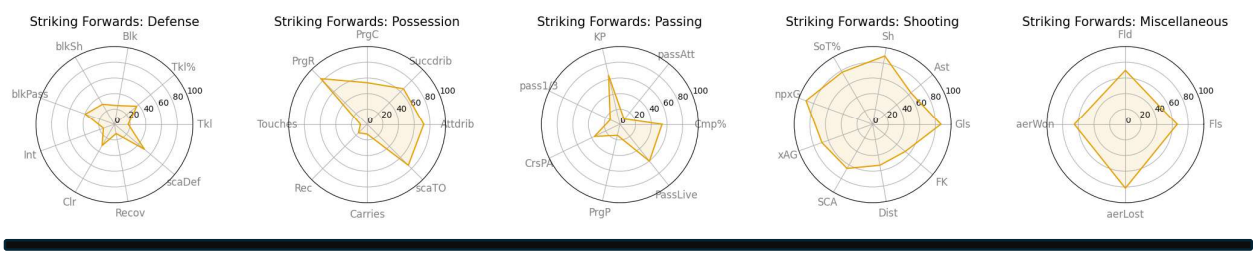

As you can see below in the charts for Right Back and Attacking Midfielder, the distribution of role is varied within specific positions, leading me to believe that there is validity in the idea of positionless play. Beneath that, you can seen the average statitistics for each cluster, in the form of percentiles of per 90 minute values compared to all players.

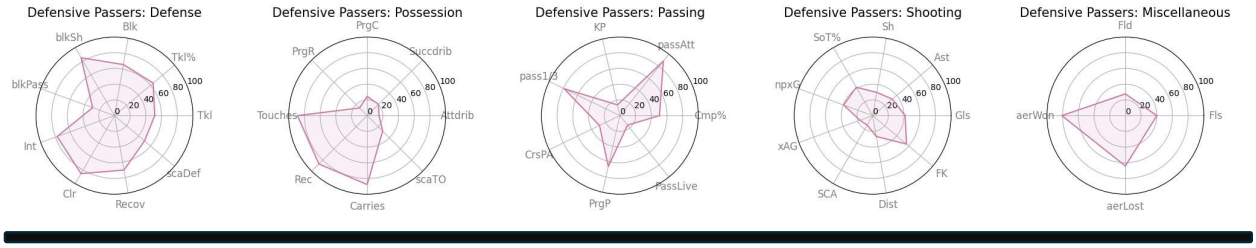

Cluster Distribution: Right-Back (Total: 89)
1.1%
1.1%
7.9%
14.6%
75.3%
Clusters
Defensive Contributors
Dynamic Creators
Defensive Stoppers
Striking Forwards
Defensive Passers


Cluster Distribution: Attacking Midfielder (Total: 60)
8.3%
15.0%
76.7%
Clusters
Dynamic Creators
Defensive Contributors
Striking Forwards


Defensive Contributors: Defense
Blk
Tkl%
Tkl
scaDef
Recov
Clr
Int
blkPass
blkSh
Defensive Contributors: Possession
PrgC
Succdrib
Attdrib
scaTO
Carries
Rec
Touches
PrgR
Defensive Contributors: Passing
KP
passAtt
Cmp%
PassLive
PrgP
CrsPA
pass1/3
Defensive Contributors: Shooting
Sh
Ast
Gls
FK
Dist
SCA
xAG
npxG
SoT%
Defensive Contributors: Miscellaneous
Fld
Fls
aerLost
aerWon

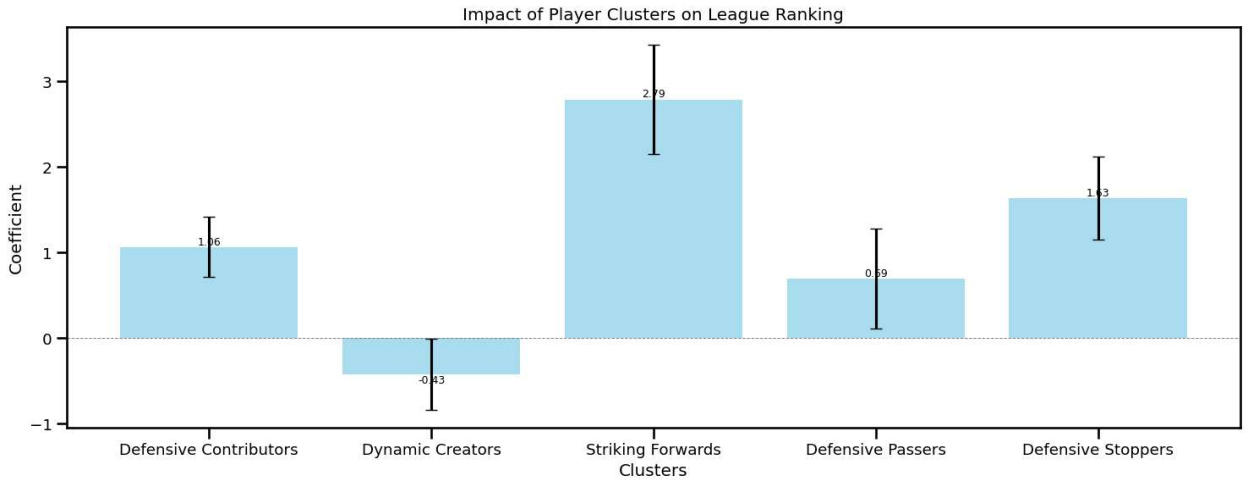

Dynamic Creators: Defense
Blk
Tkl%
Tkl
scaDef
Recov
Clr
Int
blkPass
blkSh
Dynamic Creators: Possession
PrgC
Succdrib
Attdrib
scaTO
Carries
Rec
Touches
PrgR
Dynamic Creators: Passing
KP
passAtt
Cmp%
PassLive
PrgP
CrsPA
pass1/3
Dynamic Creators: Shooting
Sh
Ast
Gls
FK
Dist
SCA
xAG
npxG
SoT%
Dynamic Creators: Miscellaneous
Fld
Fls
aerLost
aerWon


Striking Forwards: Defense
Blk
Tkl%
Tkl
scaDef
Recov
Clr
Int
blkPass
blkSh
Striking Forwards: Possession
PrgC
Succdrib
Attdrib
scaTO
Carries
Rec
Touches
PrgR
Striking Forwards: Passing
KP
passAtt
Cmp%
PassLive
PrgP
CrsPA
pass1/3
Striking Forwards: Shooting
Sh
Ast
Gls
FK
Dist
SCA
xAG
npxG
SoT%
Striking Forwards: Miscellaneous
Fld
Fls
aerLost
aerWon

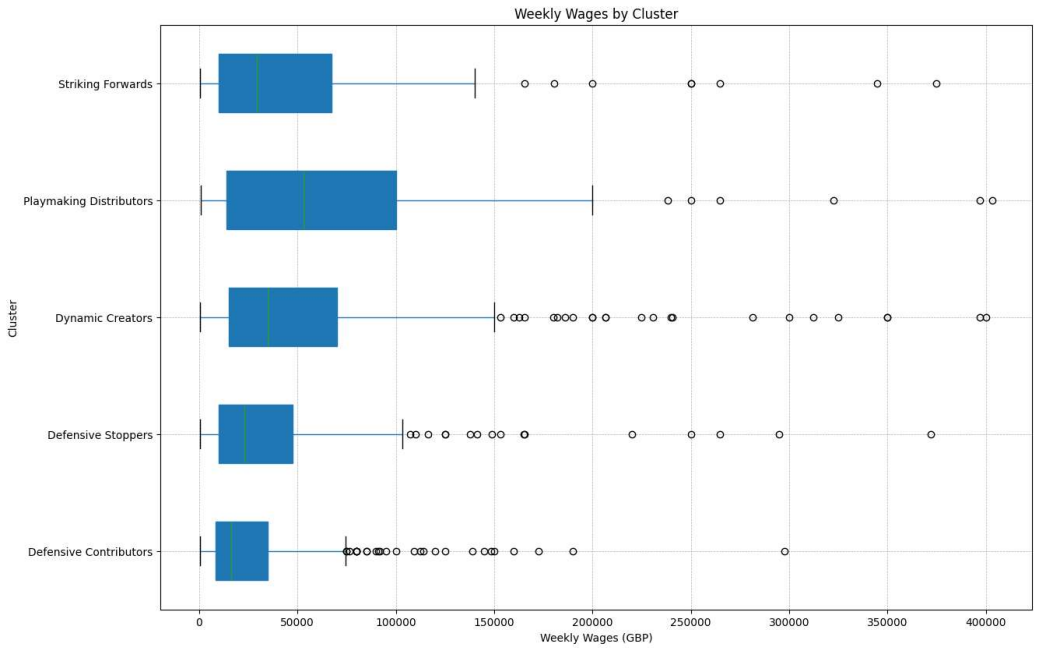

Defensive Passers: Defense
Blk
Tkl%
Tkl
scaDef
Recov
Clr
Int
blkPass
blkSh
Defensive Passers: Possession
PrgC
Succdrib
Attdrib
scaTO
Carries
Rec
Touches
PrgR
Defensive Passers: Passing
KP
passAtt
Cmp%
PassLive
PrgP
CrsPA
pass1/3
Defensive Passers: Shooting
Sh
Ast
Gls
FK
Dist
SCA
xAG
npxG
SoT%
Defensive Passers: Miscellaneous
Fld
Fls
aerLost
aerWon


Defensive Stoppers: Defense
Blk
Tkl%
Tkl
scaDef
Recov
Clr
Int
blkPass
blkSh
Defensive Stoppers: Possession
PrgC
Succdrib
Attdrib
scaTO
Carries
Rec
Touches
PrgR
Defensive Stoppers: Passing
KP
passAtt
Cmp%
PassLive
PrgP
CrsPA
pass1/3
Defensive Stoppers: Shooting
Sh
Ast
Gls
FK
Dist
SCA
xAG
npxG
SoT%
Defensive Stoppers: Miscellaneous
Fld
Fls
aerLost
aerWon

Having established the clusters, I sought to determine whether the composition of teams with respect to the number of players from each cluster correlated to their performance in their respective league. Using the *statsmodels.api* Ordinary Least Squares class, I found the relationship between the number of players from each cluster in each team's top 11 players (by minutes) and the teams league rank from the 22-23 season. The coefficients for each cluster were as follows:




I also graphed the wages of players based on clusters but was unable to find much useful insight from this, indicating that player quality is more important than player role when it comes to wage.



As you can see, the median wages of players are fairly similar, with all clusters having significant outliers on the higher end. Further analysis of individual positions and cluster wages did not yield any interesting results either.

**Conclusions**

This study's aim was to analyze soccer player roles by clustering players based on their functional contributions rather than traditional positions. By examining comprehensive player statistics from the 2022-23 season across the top five European leagues, we successfully identified five distinct player clusters: Defensive Stoppers, Dynamic Creators, Striking Forwards, Defensive Contributors, and Defensive Passers. The findings suggest that studying players by positions is potentially limiting and that a more nuanced understanding based on contributions could provide a better picture of team dynamics and player effectiveness. The correlation between team composition, in terms of these clusters, and league performance was significant, indicating that teams with a higher proportion of Dynamic Creators generally performed better in their respective leagues. In addition, an over-reliance on Striking Forwards could negatively impact team performance, suggesting that a balanced team structure with varied functional roles contributes more effectively to winning outcomes. As for the wages of players, there is not significant insight to be gleaned from the clustering of players and leads me to believe that wages are determined more based on quality compared to others than overall role. Further research could explore the optimal number of clusters to better represent all players and potential better ways to account for player quality. This study paves the way for more specific coaching and recruiting, driven by data and a deeper understanding of player capabilities and team needs, though not necessarily with how to better align finances and pay structures for players.

**Citations**

Data was sourced from fbref.com and Fotmob.com. The analysis utilized various Python libraries: Pandas, NumPy, Selenium, sklearn, Matplotlib, and statsmodels.