OnSports: Data-Driven Player Clustering for Fantasy Pricing Strategy

Premier League Performance Segmentation Using Unsupervised Learning

By: Larriet Poteat
11/30/2025

BUSINESS PROBLEM

How can OnSports segment players based on past season performance to define fair and competitive starting prices for the new fantasy season?

OnSports is preparing pricing for Premier League players for the upcoming fantasy season. Player prices must reflect real-world performance to maintain competitive balance, drive user engagement, and incentivize strategic play. However, player performance varies greatly across multiple statistics (goals, assists, influence, threat, clean sheets, etc.), making manual pricing subjective and inconsistent.

1. Player valuation currently lacks a consistent, data-driven approach
2. Performance metrics differ in scale and impact, making comparison difficult
3. Elite performers are undervalued while mid-tier players may be overpriced
4. A systematic segmentation is required to classify players into pricing tiers

We apply unsupervised machine learning to uncover natural player groups based on multivariate performance signal to address this.

SOLUTION APPROACH

We apply **Unsupervised Learning** to segment Premier League players into meaningful performance-based groups. These clusters enable OnSports to assign initial prices aligned with real-world contributions.

Methods Used

1. **K-Means Clustering** to identify natural player segments
2. **Hierarchical Clustering** to validate and compare grouping structures
3. **Feature Scaling** and **EDA** to prepare and standardize data before clustering

WE USE UNSUPERVISED LEARNING BECAUSE THERE ARE NO PREDEFINED PRICING LABELS AND OUR OBJECTIVE IS TO DISCOVER PATTERNS HIDDEN IN PERFORMANCE DATA

ATTACKING PERFORMANCE METRICS

Elite attacking players are scarce, yet they generate outsized value through goals, assists, and scoring opportunities. These players naturally belong to the **highest fantasy pricing tier**, as their contributions significantly influence match outcomes and user decision-making.

Key Observations

Goals Scored, **Assists**, and **Threat** all show **strong right-skew** distributions.

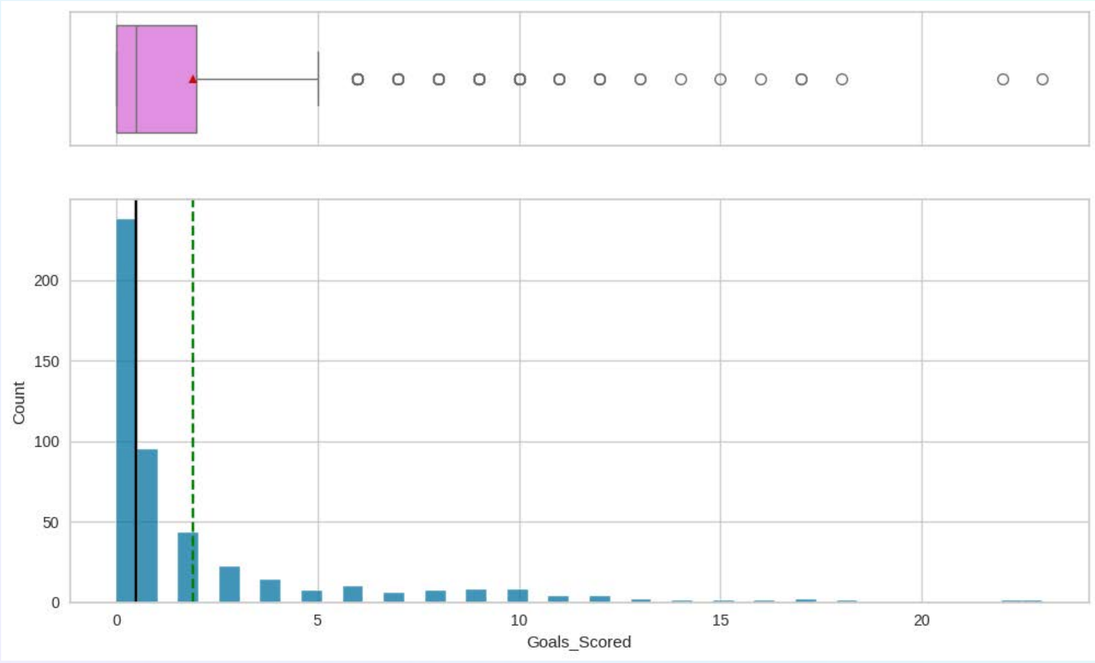
The **majority of players contribute minimally** to attacking outcomes:

- Most score **0–2 goals**
- Most provide **0–2 assists**
- Most have **low Threat scores**

A **small elite subset** demonstrates **high Threat and scoring potential**, extending far beyond the median and forming clear performance outliers.

Why This Matters for OnSports

- Attack-driven players create the **largest scoring differential** in fantasy outcomes
- Mispricing elite attackers leads to **competitive imbalance** and **user frustration**
- Identifying this group early ensures:
  ✔ Fair, skill-based gameplay
  ✔ Strong engagement
  ✔ Strategic team building



The skewed distribution of attacking contributions indicates natural player segmentation, providing a strong foundation for clustering-based pricing.

PLAYMAKING & INFLUENCE METRICS

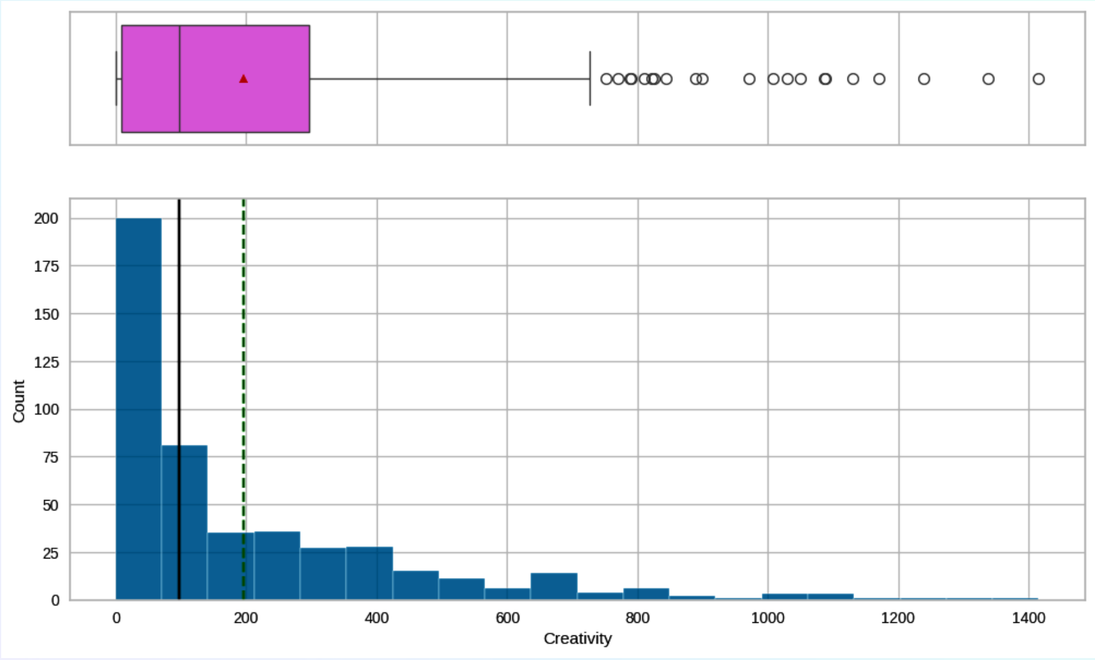
Playmakers and high-influence players are set apart and rare yet have an outsized impact on fantasy outcomes. Their ability to generate chances, control matches, and accumulate bonus points strengthens them into a distinct premium pricing tier, separate from general attackers.

Key Observations

- **Creativity**, **Influence**, and **Bonus** metrics demonstrate **significant right-skew**, similar to attacking metrics but with **even more extreme outliers**.
- Only a **small number of players** operate at exceptionally high levels of creativity and influence.
- **Bonus points** accumulate disproportionately among players who already excel in other offensive or creative dimensions.

Why This Matters for OnSports

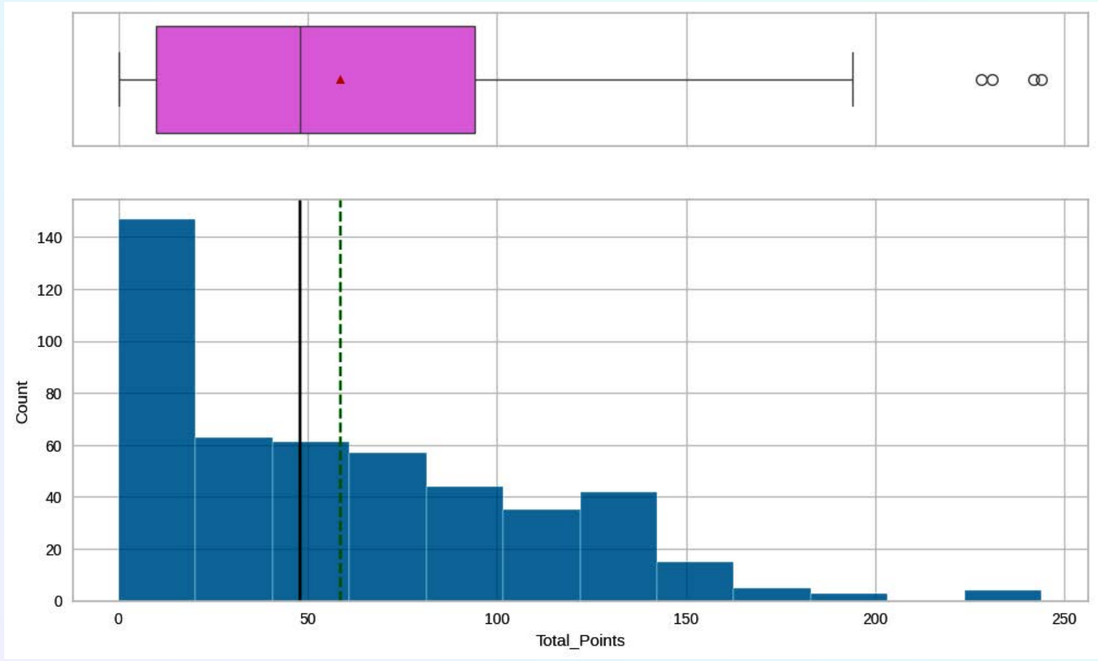
- Players with high creativity and influence produce **consistent non-goal contributions**, which fantasy users rely on for steady scoring
- Pricing must account for **both direct (goals)** and **indirect (chance creation, match impact)** contributions
- Undervaluing elite playmakers distorts user strategy and rewards luck over skill



The presence of extreme outliers across playmaking metrics further reinforces the need for clustering to differentiate elite creators from standard midfield contributors.

DEFENSIVE & PARTICIPATION METRICS

Defensive performance and consistent match participation form the backbone of mid-tier pricing. However, players with high Total Points , driven by both minutes and impactful actions, emerge as a separate premium tier, confirming the need to differentiate players based on cumulative fantasy contribution.



Key Observations

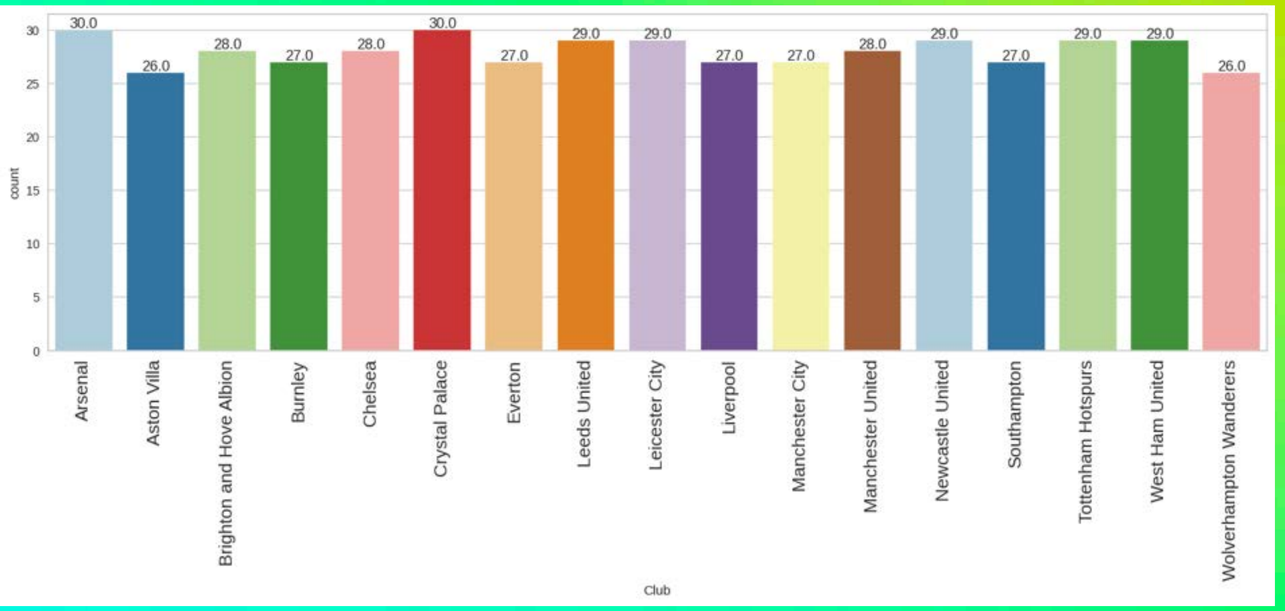
- **Clean Sheets** show a moderate right-skew, reflecting stable defensive contributors, mostly defenders and goalkeepers.
- **Goals Conceded** varies widely by club and position, identifying defenses that perform at different tiers.
- **Minutes Played** exhibits a broad distribution — some players are regular starters, while others are peripheral impact players.
- **Total Points** displays a **long-tailed distribution** with a small number of players accumulating significantly more points than the average.

Why This Matters for OnSports

- Regular starters offer **predictable returns**, shaping the **middle pricing tier**
- High clean sheet players provide value independent of scoring
- High Total Points players consistently outperform peers — undervaluing them would distort competitive fairness
- Understanding defensive and participation patterns prevents **overpricing flashy attackers** and **underpricing reliable performers**

Together, attacking, playmaking, and defensive metrics reveal naturally distinct player groups, making clustering the optimal approach for defining data-driven pricing tiers.

PLAYER DISTRIBUTION OVERVIEW



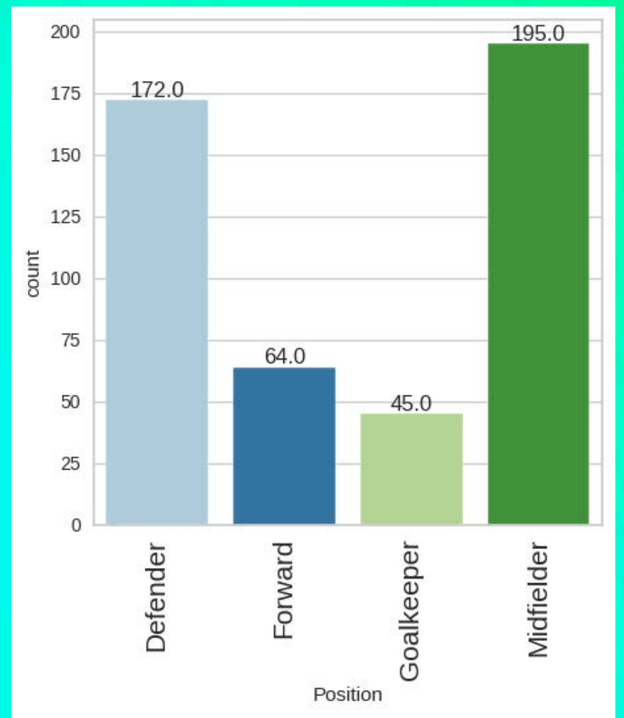
Understanding player distribution by club and position provides essential context for pricing. Since midfielders and defenders dominate the player pool, pricing strategies must reflect positional depth and role-based impact, ensuring balanced selection options for fantasy users.

KEY OBSERVATIONS

The dataset covers **476 Premier League players** across **17 clubs**, with representation varying slightly by team. *Arsenal, Crystal Palace, Leicester City, Tottenham Hotspurs, and West Ham United each contribute ~29–30 players.*

Midfielders are the largest position group (**195 players**), followed by **Defenders** (**172 players**), while **Goalkeepers** are the least represented (**45 players**).

The high share of midfielders aligns with fantasy gameplay, where midfield roles accumulate points from both scoring and creative metrics.



Narrative Link to Modeling

- Uneven club representation suggests **team performance effects** may influence metrics like clean sheets and goals conceded
- Positional breakdown explains **metric-driven segmentation**, supporting clustering later

With a clear understanding of player roles and representation, we now explore how key performance variables relate to each other to uncover hidden patterns.

BIVARIATE RELATIONSHIPS AND KEY PERFORMANCE PATTERNS

Player value in fantasy football is not determined by a single statistic, but by the interaction of multiple performance metrics. These relationships reveal how distinct attributes like scoring threat, creativity, and consistent minutes, collectively shape a player's fantasy pricing potential.

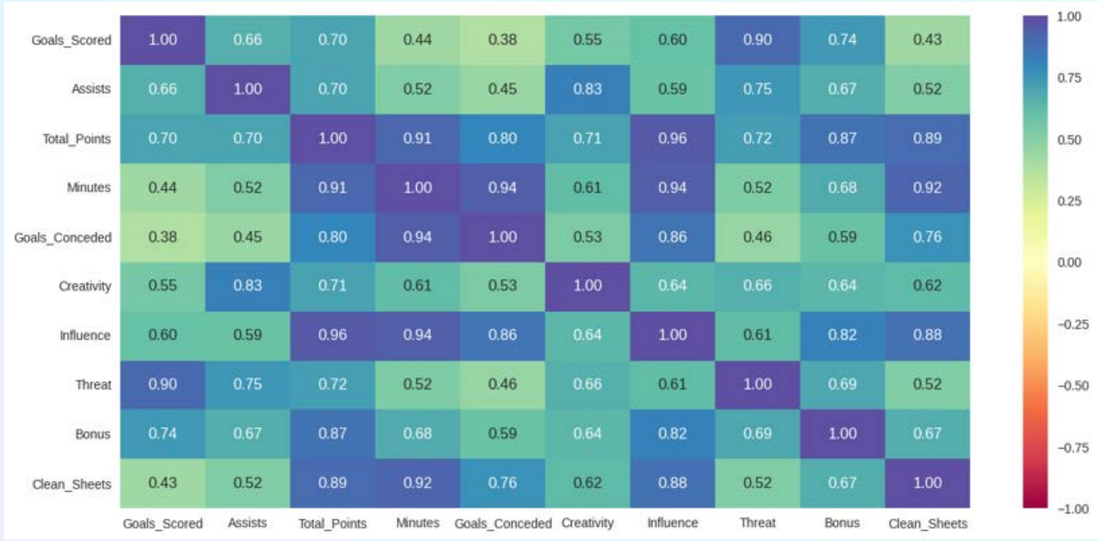
Key Insights from Variable Relationships

Threat shows a strong positive correlation with **Goals Scored**
→ Players who frequently attempt shots are more likely to convert them into goals.

Minutes Played strongly correlates with **Total Fantasy Points**
→ Consistent participation drives predictable scoring returns, independent of position.

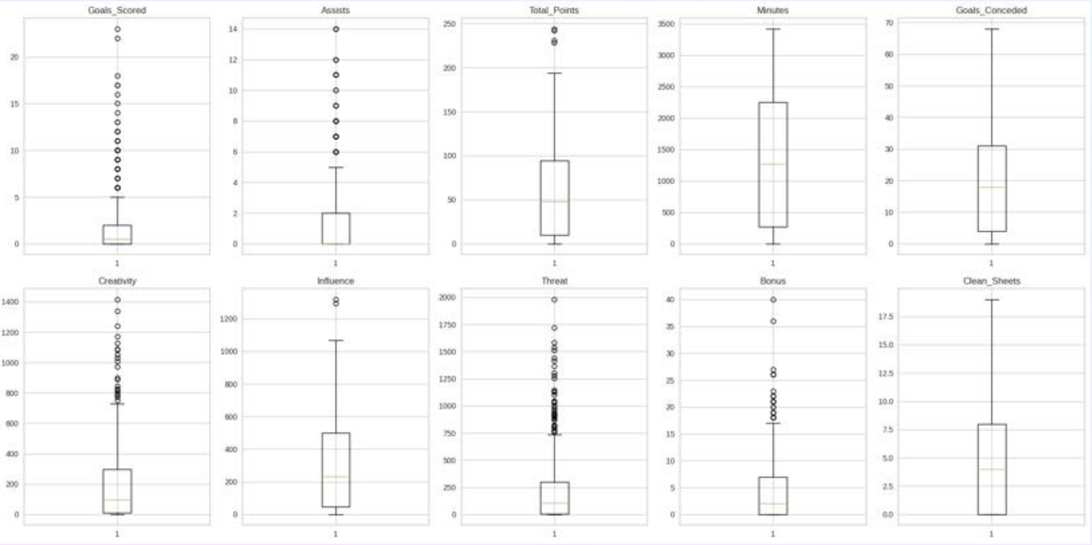
Creativity and **Assists** exhibit high correlation
→ Creative midfielders generate chance creation value that translates directly to fantasy scoring.

Clean Sheets correlate with reduced **Goals Conceded**
→ Defensive stability contributes meaningfully to fantasy value beyond scoring metrics.



With clear evidence of interconnected performance metrics, we now prepare the dataset for clustering by addressing outliers and scaling features.

DATA PREPROCESSING: PREPARING FEATURES FOR CLUSTERING



Checks Performed

Step	*Result*	*Interpretation*
Missing Values	❌ None detected	Dataset is complete — no imputation required
Duplicate Records	❌ None found	Each player is uniquely represented
Outliers	✔ Present across metrics	These represent *elite players* and **must be kept** because removing them would eliminate the very pricing tiers we want to identify
Feature Types	10 numerical features selected	All relevant performance metrics included for clustering

Business Logic

Outliers are not noise — they **create meaningful separations** between:

- Rotational players
- Consistent starters
- Elite premium-tier fantasy assets

Therefore, **we preserve outliers** to allow clusters to reveal pricing tiers naturally.

PREPARING DATA FOR CLUSTERING: FEATURE SELECTION, OUTLIERS, AND SCALING

✗ Outliers exist: e.g. top scorers like **Bruno Fernandes**, **Harry Kane**, **Martínez**, **Dallas**
We did **not** remove them
Elite players create the price brackets and removing them defeats the purpose of clustering.

If removed, player tiers collapse, elite players look average, and can create poor pricing signals. If kept, clusters naturally form performance tiers, top performers get premium pricing, and users see transparent data driven logic.

Why scaling matters

Without scaling:

- Features like **Minutes (~3500 max)** would dominate
- Features like **Bonus (~40 max)** would be ignored
- Clusters would reflect magnitude, not player value

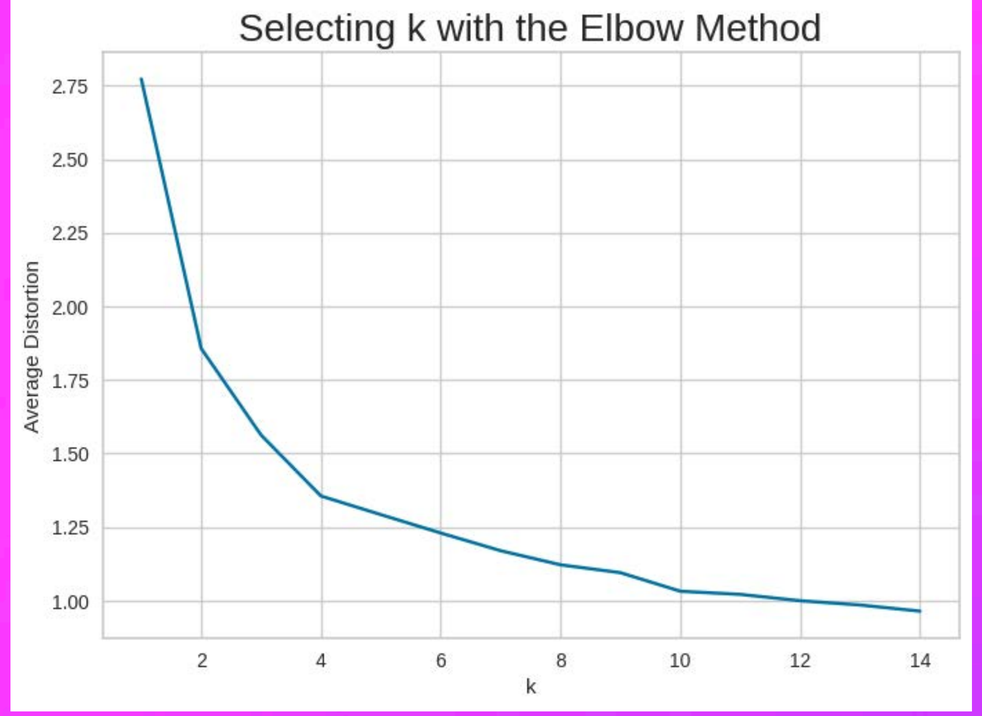
With scaling:

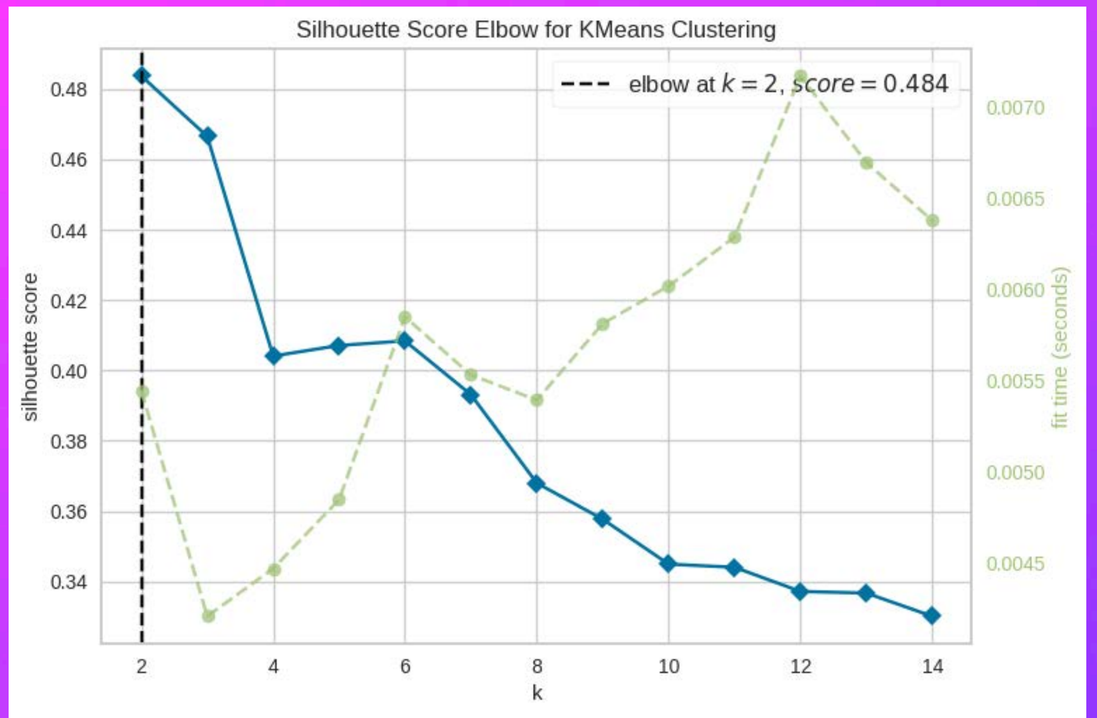
- Every metric contributes equally
- Clusters reflect **true player profiles**
- No attribute unfairly drives segmentation

Business Translation:
Scaling ensures pricing tiers are based on **skill and contribution**, not unit size.

Performance driven features that directly influence fantasy valuation.
Why these features?
They represent **how often players play**, **how much impact they have**, and **how that translates into fantasy points** : the three pillars of pricing.

Category		Features	
Attacking Output		Goals_Scored, Assists, Threat	
Creative Influence		Creativity, Influence, Bonus	
Defensive Metrics		Goals_Conceded, Clean_Sheets	
Opportunity / Playtime		Minutes	
Overall Performance		Total_Points	





OPTIMAL NUMBER OF PLAYER CLUSTERS

Why Select an Optimal k?

To determine how many meaningful player segments exist in the dataset, we evaluated different values of **k** using two industry-standard validation techniques:

- **Elbow Method** measures reduction in distortion
- **Silhouette Score** measures cluster separation quality

Both metrics help ensure clusters are **interpretable, stable, and actionable** for pricing decisions.

Elbow Method Insight

- Distortion decreases sharply until **k = 2**
- Beyond **k = 2**, improvements taper off
  Indicates **diminishing returns** past two clusters

Interpretation:
There are only **two natural performance tiers** among players. Adding more clusters would create artificial splits without business value.

Silhouette Score Validation

- The **highest silhouette score occurs at k = 2**
- Silhouette plot shows **clear separation** between two groups

Interpretation:
Players form **two behaviorally distinct clusters** based on performance metrics.

OnSports will adopt a 2-cluster segmentation model
representing **two pricing tiers** that reflect how players contribute in fantasy scoring.

Business Impact

✔️ Simple and intuitive for users
✔️ Supports transparent and fair pricing
✔️ Aligns with real-world player value distribution
✔️ Reduces cognitive load and increases engagement

K-MEANS CLUSTERING OVERVIEW

Group Premier League players into **distinct performance-based segments** to enable data-driven, tiered pricing for the upcoming fantasy season.

Model Used: K-Means Clustering

Algorithm Choice Rationale:

- Designed for **high-dimensional quantitative performance metrics**
- Efficient for large datasets like player statistics
- Ideal for uncovering **natural groupings** based on contribution patterns

Variables Included in Clustering

K-Means was trained using **scaled numerical indicators** of fantasy value:

- Goals Scored, Assists, Total Points, Minutes Played
- Creativity, Influence, Threat
- Goals Conceded, Clean Sheets, Bonus

These metrics jointly capture how players contribute to **scoring, influence, consistency, and defensive value**.

Final Model:

KMeans(n_clusters=2, random_state=1)

Interpretation:

K-Means has successfully identified **two distinct tiers of players** based on their historical fantasy performance:

- S**egment 0** – Baseline/regular contributors
  Segment 1 – High-impact / premium players

These clusters will now be analyzed to determine **how they differ in value and how pricing can leverage those differences**.

K-MEANS CLUSTER PROFILES

K-Means identified **two distinct tiers of Premier League players** based on last season's fantasy performance

Cluster 0 : *High-Impact Fantasy Performers*

(199 players | Premium Tier)

Players in this group:

- Play significantly more minutes (avg. **2,399**)
- Score and assist far more (avg. **4 goals & 3.5 assists**)
- Drive match outcomes through Creativity and Influence
- Earn high Total Points (**110+ average**) and Bonus awards
- Form the **core of fantasy team selection and pricing strategy**

Examples:
Mohamed Salah, Harry Kane, Bruno Fernandes, Son Heung-Min, Kevin De Bruyne, Jamie Vardy

Business Meaning:
These are **your highest-value assets**.

They justify premium pricing, captain multipliers, and marketing features.

Cluster 1 : *Supporting / Rotational Contributors*

(277 players | Value Tier)

Players in this group:

- Play limited minutes (**~573 avg.**)
- Minimal direct scoring contribution
- Low fantasy Influence, Threat, and Creativity
- Often substitutes, squad players, recovering injuries, or defensive depth

Examples:
Bench defenders, second-choice goalkeepers, youth prospects, short-term transfers

Business Meaning:
These players fit **entry-level pricing** and enable budget formations,

differential tactics, and user strategy variety.

K-Means reveals **two clear performance-driven player segments**. This segmentation forms the foundation for **transparent, data-driven pricing**, ensuring fairness and strategic depth for OnSports users.

HIERARCHICAL CLUSTERING & DENDROGRAM ANALYSIS

Hierarchical clustering doesn't just group players, it exposes the *true outliers* who move the Premier League.

Selected Method: Weighted Linkage + Euclidean Distance
Reason:
It achieved the **highest cophenetic correlation (0.862)**, meaning it preserves player similarity patterns **more accurately than any other method**.

Linkage Method		Cophenetic Correlation	
Weighted		0.862(Best	
Average		0.848	
Single		0.843	
Centroid		0.807	
Complete		0.741	
Ward		0.578	

Methodology

- **Model Used:** Agglomerative Clustering
  linkage = 'average', **Euclidean** distance
- **Why Hierarchical?**
  Unlike K-Means, this approach:
  - Does **not require specifying k** upfront
  - Reveals **data inheritance structure** through dendrograms
  - Helps evaluate which linkage method best preserves data similarity

Cluster Outcome

Hierarchical clustering produced **2 distinct segments**:

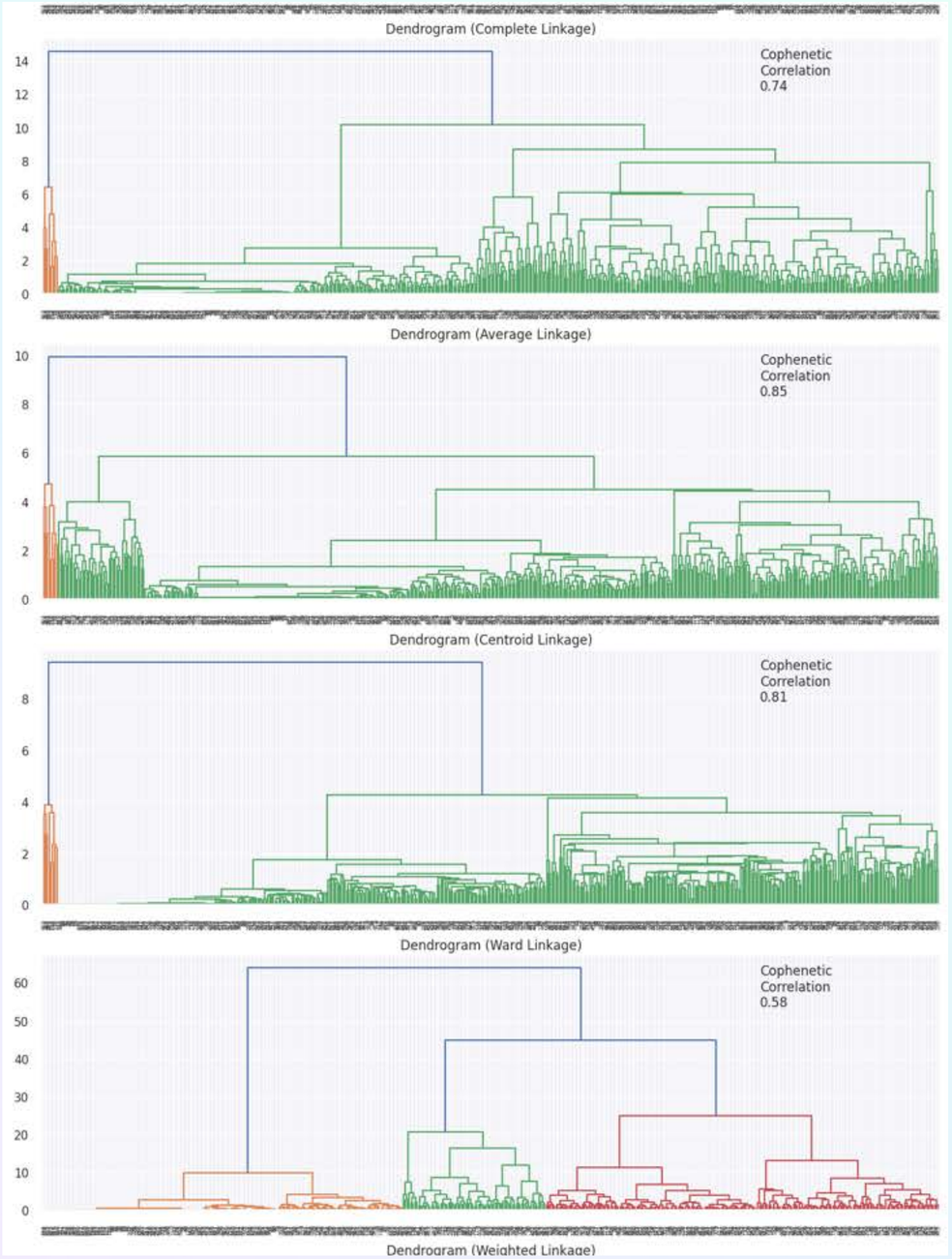
Segment	Description	Players	
HC Segment 0	Majority cluster, solid contributors with moderate stats	468 Players	
HC Segment 1	**Ultra-elite performers** with exceptional scoring, assists, influence, and minutes	8 players (e.g., Kane, Salah, Son, Fernandes, Vardy, Bamford, Watkins, Calvert-Lewin)	

Key Insight:
Hierarchical clustering **isolates only the most impactful players**,

enabling a premium-tier pricing and marketing strategy.

Business Interpretation

- Hierarchical clustering **validates** the presence of a **small, high-impact elite group**
- These players directly influence **fantasy engagement**, pricing premiums, and user acquisition
- The model reveals a **revenue opportunity** in distinguishing **premium superstars** from the broader player base



HIERARCHICAL CLUSTER PROFILES

Hierarchical clustering reveals who truly changes matches.

Cluster Interpretation

Cluster 0: Regular Squad Contributors

- Large group of rotational / supporting players
- Provide moderate points, inconsistent minutes, and limited offensive threat
- Ideal for **budget pricing tiers** and **bench depth strategies**

Cluster 1: Elite Match Winners

- Extremely high fantasy output across **every performance metric**
- Central to scoring, creativity, and match influence
- Includes the league's most valuable assets:

Examples:
Harry Kane, Bruno Fernandes, Mohamed Salah, Heung-Min Son, Jamie Vardy, Patrick Bamford, Ollie Watkins, Dominic Calvert-Lewin

These players standout as **premium-tier purchases** in any fantasy platform.

Business Insight

This segmentation **automatically isolates elite performers** without manual intervention, enabling:

✓ Dynamic **player pricing based on value, not hype**
✓ Tiered engagement strategies (casual vs competitive users)
✓ Enhanced fairness and credibility of the fantasy platform

Feature	Cluster 0 (Majority Players)	Cluster 1 (Elite Performers)
Count	468 Players	8 Players
Totals Points	55.97	207.38
Goals Scored	1.64	17.74
Assists	1.60	10.63
Minutes Played	1,307	3,059
Creativity	187.37	699.45
Influence	282.63	995.83
Threat	203.50	1,480.25
Bonus	4.35	26.38
Clean Sheets	4.63	11.50

K-MEANS VS HIERARCHICAL CLUSTERING

WE COMPARE BOTH TO TO VALIDATE CLUSTER ROBUSTNESS AND SELECT THE APPROACH THAT BEST SUPPORTS AUTOMATED PRICING AND SCALABLE PLAYER SEGMENTATION.

Key Observations

K-Means Results

- Groups players based on **overall performance patterns**
- Separates **consistent starters** from **rotational contributors**
- Best suited for **broad pricing tiers and scalable user interactions**

Hierarchical Results

- Identifies a **micro-segment of elite match deciders**
- Produces **clear talent stratification**
- Provides high-confidence clustering validated by the **highest cophenetic correlation (0.86)** using **Weighted Linkage**

Business Verdict

Automated & scalable pricing engine use K-Means
Detecting elite / premium-tier talent use Hierarchical
Hybrid pricing + player valuation model use Both

K-Means scales the fantasy economy; Hierarchical identifies the players who reshape it.
Together, they create a pricing system that is **fair, data-driven, and strategically defensible.**

Criteria	K-Clustering	Hierarchical Clustering
Cluster Shape Assumptions	Spherical clusters	No redefined cluster shape
Number of clusters	Requires manual selection (k=2)	Determined visually and cophentic correlation
Best Metric Indicator	Elbow and Silhouette Score	Cophentic Correlation
Performance	Fast, scalable, to thousands of players	slower with large dataset, best for validation
Interpretability	Moderate	Very high, clear dendrogram splits
Final Output	2 clusters: Core starters vs. squad depth	2 clusters: Elite performers vs. everyone else
Business Fit	Pricing automation, platform-wide scalability	Identifying true premium players with certainty

BUSINESS ACTIONS AND RECOMMENDATIONS

FROM PERFORMANCE DATA TO PRICING STRATEGY : ONSPORTS CAN NOW LEVERAGE CLUSTERING INSIGHTS TO CREATE A TRANSPARENT, DATA-DRIVEN FANTASY PRICING MODEL THAT ENHANCES COMPETITIVENESS, USER ENGAGEMENT, AND REVENUE PREDICTABILITY.

Key Actions

Implement Tiered Player Pricing

Use cluster assignments to define pricing levels:

- **Elite Tier** *(Hierarchical Cluster 1)*
  Premium pricing for match-deciding superstars
  → Higher entry cost, captain multipliers, promotional visibility

- **Starter Tier** *(K-Means Cluster 0)*
  Fair pricing tied to reliable, high-minute contributors
  → Core roster players users depend on

- **Value / Bench Tier** *(K-Means Cluster 1)*
  Low-cost options with situational upside
  → Enables lineup diversity and strategic roster building

Enhance User Engagement

- Highlight elite-tier players in matchday banners and UX touchpoints
- Offer weekly “value alerts” identifying rising players moving across clusters
- Gamify decisions via **cluster badges** (Elite, Core, Upside)

Monetize Through Predictability

- Reduce subjective/hype-based pricing errors
- Align costs with **actual fantasy impact metrics**
- Increase user trust and retention by making pricing **explainable and fair**

Platform Growth Opportunities

Cluster insights support:

- Dynamic pricing updates each gameweek
- Partnership with clubs, sponsors, or broadcasters around elite talent
- Player performance forecasts, trade suggestions, and personalized recommendations

Strategic Payoff

OnSports transforms from a fantasy game into a data-validated decision engine.
Users gain clarity. Prices reflect performance. Elite players become premium assets , not guesses.

Clustering doesn't just categorize players, t powers a pricing system that matches the way football is played: the few redefine the many.