

The Big Six or the Big One? Defining the True Hierarchy of the Premier League through Financial and Sporting Metrics

Oscar Barnes

Contents

Introduction: The Quantitative Anatomy of the Elite	2
Aims and Objectives	2
Report Roadmap	2
Data: Defining the Variables and Scopes of Competitive Success	3
Variable Domains	3
Rationale: The Pillars of Dominance	4
Data Preparation	4
Analytical Scopes: Sensitivity and Robustness	4
Methodology: From Raw Data to Competitive Tiers	5
1. Data Standardisation and Multicollinearity Check	5
2. Dimension Reduction: Principal Component Analysis (PCA)	5
3. Cluster Validation and K-Means Implementation	6
4. Proximity Audit and The “Chasm”	6
Results and Analysis	7
1. Initial Data Diagnostics and Preparation	7
2. Validation of Dimensionality and Structure (PCA Findings)	8
3. Determination of Competitive Tiers (Clustering Validation)	10
4. Identification and Analysis of Competitive Tiers	12
5. Discussion: The Mechanics of Stratification	14
6. The Challenger Audit: Quantifying the Glass Ceiling	15
Conclusion: The Structural Calculus of Success	18
Synthesis of Findings	18
The Manchester United Paradox and Structural Weight	18
Model Limitations	18
Opportunities for Further Analysis	18
Final Outlook	19

Introduction: The Quantitative Anatomy of the Elite

The term ‘Big Six’ is the latest iteration of a shifting power structure that has evolved alongside the Premier League’s commercial growth. This informal designation—comprising Arsenal, Chelsea, Liverpool, Manchester City, Manchester United, and Tottenham Hotspur—superseded previous eras of concentration, including the 1980s “Big Five,” the late-90s “Big Two” duopoly of Arsenal and Manchester United, and the mid-2000s “Big Four.” Since 2004, these six clubs have typically accounted for over half of the league’s total annual revenue. Their sporting dominance is equally stark; since the league’s inception in 1992, a “Big Six” member has won the title in all but two seasons (1994/95 and 2015/16) (Wikipedia 2024).

Despite this historical permanence, the “Big Six” hierarchy now faces unprecedented pressure from massive external investment (e.g. Newcastle United) and the rising sporting efficiency of ambitious challengers like Brighton and Aston Villa. This report moves beyond media-driven narratives by employing multivariate statistical techniques, **Principal Component Analysis (PCA)** (Jolliffe 2002) and **K-Means Clustering** (MacQueen 1967), to objectively map the Premier League’s structure. By analysing 11 variables across **Sporting Output**, **Financial Scale**, and **Market Reach** over a five-year window, I quantify the precise statistical distance between the elite and their challengers to determine if the “Big Six” remains a closed shop.

Aims and Objectives

The primary objective of this project is to use **Principal Component Analysis (PCA)** and **K-Means Clustering** to identify and profile the underlying structure of the Premier League, thereby establishing a new, data-driven hierarchy. Specifically, this analysis aims to:

1. **Dimension Reduction (PCA):** Uncover and name the **three core, uncorrelated pillars** (Absolute Scale, Competitive Stability, and Infrastructure/Pedigree) that collectively explain over 92% of the variance observed across the 11 club metrics.
2. **Cluster Validation:** Statistically test the hypothesis of the traditional ‘Big Six’ by determining if they truly occupy a unique, isolated statistical cluster.
3. **Hierarchy Definition (K-Means):** Determine the optimal grouping of clubs ($k = 4$ clusters) to establish a clear stratification of the league, moving from the elite powerhouses to the “Yo-Yo” tier.
4. **Challenger Identification (The Vanguard):** Identify the specific clubs (the “Vanguard”) that the model naturally separates from the middle-tier, positioning them as the most likely disruptors of the established order.
5. **Proximity Audit:** Conduct a high-resolution distance analysis to quantify the mathematical “gap” between the Elite center and the chasing pack, determining how close the Vanguard actually is to breaking the glass ceiling.

By successfully executing these aims, this report provides a quantitative anatomy of the Premier League, detailing the measurable **Financial Chasm** that separates the top tier from the rest and highlighting the cost of maintaining elite status.

Report Roadmap

The remainder of this report proceeds as follows: the **Data** section details the variables used; the **Methodology** section walks through the standardisation process and the application of PCA and Hierarchical Clustering; the **Results** section presents the outcome of the primary analysis, including the **Component Loadings**, the final **Cluster Profiles**, the **Spatial Biplot**. The report concludes with the **Challenger Audit**, which measures the statistical proximity of Tier 2 clubs to the Elite cluster, and a discussion on the implications of the identified structure.

Data: Defining the Variables and Scopes of Competitive Success

The analysis is primarily based on a structured dataset comprising 27 current and recent Premier League clubs. This dataset was compiled from authoritative financial accounts (e.g. club annual reports, publicly available data from **(Ramble 2024)**, comprehensive sporting statistics (**FotMob 2025**)/(**UEFA 2025**) spanning the five seasons leading up to and including the 2024/25 season (financial data was collected from the 2023/24 season) as well as **Wikipedia** (Wikipedia 2025) and **Instagram** (Instagram 2025) for stadium capacity and followers respectively.

Each club serves as a single observation, and the data includes 11 key quantitative variables chosen to encapsulate the fundamental drivers of club success across three distinct domains:

Variable Domains

1. Sporting Output and Efficiency

These variables measure the club's on-pitch performance, tactical style, and historical European pedigree.

Table 1: Sporting Output and Efficiency Variables

Variable Name	Description
Avg_Pts	Average Points per Season
Avg_xGD	Average Expected Goal Difference
Avg_Poss	Average Possession Percentage
Avg_UEFA_Coeff	Average UEFA Club Coefficient
Szns	Seasons Present (Out of 5)

2. Financial and Market Scale

These variables quantify the absolute economic and global size of the club. All financial figures represent the most recently available accounting period for consistent comparison.

Table 2: Financial and Market Scale Variables

Variable Name	Description
Sqd_Mkt_Val	Squad Market Value (£)
Wage_Bill	Total Player Wage Bill (£)
Com_Rev	Commercial Revenue (£)
Tot_Rev	Total Revenue (£)

3. Fan engagement

These variables provide context on the club's physical assets and digital brand footprint.

Table 3: Fan Engagement Variables

Variable Name	Description
Std_Cap	Stadium Capacity
Ig_Fol	Instagram Followers

Rationale: The Pillars of Dominance

The selection of these 11 variables is predicated on the "Cycle of Elite Status" within modern football. In the Premier League's current climate, sporting success is rarely an isolated event; it is the result of a feedback loop between Financial Scale, Infrastructure, and On-Pitch Dominance.

- **From Scale to Success:** High Commercial and Total Revenues (Financial Scale) allow for the acquisition of elite talent (**Squad Market Value**) and the payment of competitive salaries (**Wage Bill**).
- **From Success to Dominance:** Sustained on-pitch performance (**Average Points, Expected Goal Difference**) leads to European pedigree (**UEFA Coefficient**), which secures global attention (**Instagram Followers**) and justifies massive physical assets (**Stadium Capacity**).

Ultimately, this cycle creates a "**Financial moat**" around the established elite. Because the data shows a near-perfect correlation between financial input and sporting output, the model can now clearly identify which clubs have truly crossed the threshold into permanent elite status and which are still struggling to build a sustainable moat.

Data Preparation

Prior to analysis, the raw data required a critical preprocessing step. All 11 quantitative variables were subjected to **Standardisation (Z-score transformation)**. This ensures that variables with large magnitudes (e.g. Revenue in the hundreds of millions) do not artificially dominate the PCA and clustering results over variables with smaller scales (e.g. average points).

Analytical Scopes: Sensitivity and Robustness

The analysis is conducted across a **Full League Scope**, incorporating all 27 clubs that have competed in the Premier League within the five-year window. This comprehensive approach is designed to:

1. **Establish a Baseline:** To identify the primary "Financial Chasm" between the established global brands and the rest of the domestic competition.
2. **Ensure Resolution:** By including a wide spectrum of clubs—from perennial title contenders to recently promoted sides—the model is able to distinguish between different tiers of stability.
3. **Identify the Vanguard:** This scope allows the clustering algorithm to naturally separate "Vanguard" challengers (clubs with high-growth financial and sporting trajectories) from the "Established Middle", providing a high-resolution map of the league's shifting power dynamics.

By analysing the full dataset simultaneously, the report ensures that the identified tiers are mathematically stable and reflect the organic, multi-layered structure of the modern Premier League.

Methodology: From Raw Data to Competitive Tiers

The analytical procedure was executed in a three-stage process to transform 11 complex variables into a clear, multidimensional map of the Premier League.

1. Data Standardisation and Multicollinearity Check

Prior to any analysis, the data was prepared to ensure mathematical parity across all variables.

1.1 Z-Score Transformation (Standardisation)

The initial dataset contained variables measured on vastly different scales. To prevent variables with the largest magnitudes from “bullying” the model and artificially dominating the results, all 11 variables were subjected to Z-score transformation. This centers the data by subtracting the mean and scaling by the standard deviation:

$$Z_i = \frac{X_i - \mu_X}{\sigma_X}$$

This places every club on the same “yardstick”, where a score of 0 represents the league average, and positive or negative numbers represent how far they sit above or below that average.

1.2 Multicollinearity Assessment

In football, many metrics are redundant; for instance, a club with high revenue almost always has high wages. I generated a **Correlation Matrix** to identify these overlapping relationships. The high degree of correlation between financial inputs and sporting outputs (often exceeding $r = 0.90$) justifies the use of PCA to boil the data down into its most essential, unique factors.

2. Dimension Reduction: Principal Component Analysis (PCA)

PCA was employed to transform the 11 original variables into 3 uncorrelated, composite variables known as **Principal Components (PCs)**. This reduces the complexity of the data while keeping the “big picture” intact.

- **PCA Execution and Selection:** To determine the optimal number of components, I employed a dual-criteria approach. First, I utilised the ‘Scree Plot Test’ to visually inspect for an ‘elbow’, the point where diminishing returns set in. However, to ensure the model captured the full complexity of the league, I also applied the 90% rule-of-thumb (keep the PCs that retain 90% of the total variance). Consequently, I chose to retain three components (PC1 to PC3), which together explain 92.3% of the total variance. In statistics, variance represents the total amount of ‘information’; ensuring we cross the 90% threshold guarantees a nearly complete picture of what differentiates these clubs, rather than relying solely on the most dominant factor.”.
- **Component Interpretation and Rotation:** The components were subjected to "orthogonal rotation" (Varimax). This is essentially a mathematical sharpening tool that ensures each variable clearly belongs to only one factor. This allowed us to assign clear, real-world labels to the components: **Absolute Scale** (PC1), **Competitive Stability** (PC2), **Infrastructure & Pedigree** (PC3).

3. Cluster Validation and K-Means Implementation

The final phase involved identifying the natural “tribes” within the league using a combination of hierarchical and partitioning clustering.

3.1 Hierarchical Clustering (The Exploratory Phase)

I first used Hierarchical Clustering (Sneath and Sokal 1973) to visualise the natural groupings. This produces a Dendrogram—a “family tree” of the league. Using Ward’s Minimum Variance Method, the clusters are ensured to be compact and clearly separated. This provided the visual evidence that the league naturally divides into four distinct competitive tiers.

3.2 K-Means Clustering (The Definitive Phase)

K-Means was used to assign each club to its final, definitive cluster ($k = 4$). The algorithm places a “center point” (centroid) for each group and assigns clubs to the one they are closest to. To ensure the results were mathematically stable, the process was run with 25 random starts ($nstart = 25$). To determine the most appropriate number of clusters (k), the Elbow Method was utilised (Kaufman and Rousseeuw 1990). By plotting the Total Within-Cluster Sum of Squares (WSS) against the number of clusters, the ‘elbow’ point is identified where the marginal gain in variance explanation begins to plateau.

4. Proximity Audit and The “Chasm”

The final step involved interpreting the results to define the actual distance between the clusters:

- **Cluster Profiling:** I calculated the average stats for each group (e.g. "The average Tier 1 club has £600m in revenue compared to £189m in Tier 2") to define the "character" of each tier.
- **Euclidean Distance:** I calculated the mathematical "straight-line distance" from the mathematical center (Centroid) of the Elite cluster to every other team. This provides a hard, objective number to describe the "Financial Chasm"—quantifying exactly how far a "Vanguard" club like Aston Villa or Newcastle remains from the established "Big Six."

All data processing, dimensionality reduction, and clustering were performed using the R statistical programming language (R Core Team 2024). Visualisations were generated using ggplot2 (from tidyverse), corrplot and factoextra packages to ensure high-fidelity spatial mapping of the results. Many other packages (knitr, rio, psych,forcats) were used to tidy, manipulate and structure data.

Results and Analysis

1. Initial Data Diagnostics and Preparation

This section validates the structure of the dataset and provides the quantitative justification for using **dimension reduction (PCA)** to simplify the complex relationships between the 11 variables.

1.1 Standardised Data Structure

The analysis began by applying a Z-score transformation to all 11 quantitative variables. This process “levels the playing field” by ensuring every variable is measured on a uniform scale with a **mean (μ) of 0** and a **standard deviation (σ) of 1**.

As shown in the sample below, elite clubs like **Arsenal** sit significantly above average in every category, while clubs in lower tiers, such as **Burnley**, sit consistently below. Without this step, variables with large numerical ranges (like Revenue in millions) would “bully” the model, drowning out smaller but equally important metrics (like Points per Game).

Table 4: Head of the standardised dataset (Z-scores)

	Avg_Pts	Avg_xGD	Avg_Poss	Avg_UEFA_Coef	Szns	Sqd_Mkt_Val
Arsenal	1.58	1.48	1.15	1.47	0.83	2.21
Aston Villa	0.71	0.37	0.22	1.39	0.83	0.09
Bournemouth	0.11	0.14	-0.51	-0.76	-0.45	0.00
Brentford	0.24	0.45	-0.38	-0.76	0.19	-0.08
Brighton	0.37	0.74	1.06	0.70	0.83	0.09
Burnley	-0.68	-0.63	-0.68	-0.76	-0.45	-0.64

1.2 Multicollinearity Assessment

In the modern Premier League, variables rarely move in isolation; a club’s wage bill, squad value, and points tally usually move in tandem. I used a **Correlation Matrix** (Figure 1) to quantify these relationships.

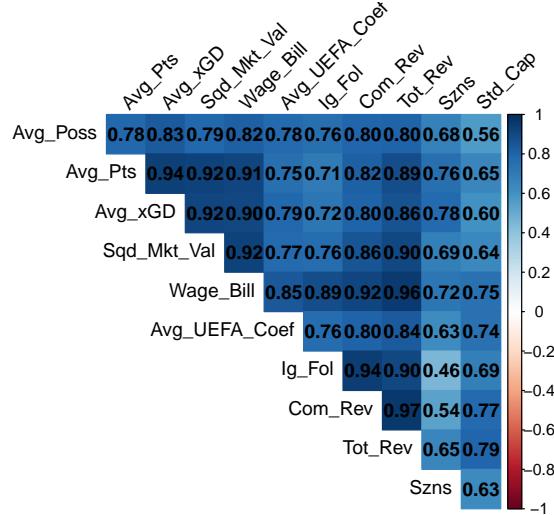


Figure 1: Heatmap of the Correlation Matrix

The heatmap reveals overwhelmingly positive correlations (darker shades of blue) across several key domains:

- **The Performance-Finance Link: Average Points and Expected Goal Difference** show a near-perfect correlation ($r = 0.94$). This proves that on-pitch performance is almost entirely synchronised with underlying quality; there is very little "statistical noise" or luck separating how a team plays from the points they earn.
- **Financial Scale:** A dense block of variables, including Squad Value, Wages, and Total Revenue, shows correlations consistently exceeding $r = 0.85$. This confirms that these metrics are highly redundant: if you know a club's revenue, you can predict their wage bill and squad market value with extreme accuracy.
- **The "Closed Loop" of Dominance:** A nearly perfect link exists between financial input and sporting output (e.g. Wage Bill and Average Points exceed $r > 0.9$). This demonstrates that the Premier League is a "pay-to-play" environment where financial scale is the primary gatekeeper of success.

This pervasive pattern of high multicollinearity confirms that the 11 variables are heavily overlapping. This provides the ultimate justification for using **Principal Component Analysis (PCA)** to boil the data down into its three most essential, uncorrelated pillars.

2. Validation of Dimensionality and Structure (PCA Findings)

PCA was employed to simplify the 11 variables into a concise set of three underlying factors that define the Premier League's competitive structure.

2.1 Component Selection: The $k = 3$ Decision

To determine the optimal number of factors, the Eigenvalues and the cumulative variance were both examined. While the Scree Plot (Figure 2) displays a sharp 'elbow' after the first component, reflecting the sheer dominance of Financial Scale, relying solely on this visual would discard critical information regarding stability and infrastructure. Therefore, the **90% Rule-of-Thumb** was applied. Since PC1 and PC2 combined explain only 87.47% of the variance, retaining the third component (PC3) was essential to cross the **90% threshold** (reaching 92.30%). This ensures the model accounts for the nuanced structural drivers that the primary financial dimension obscures.

Table 5: PCA Component Variance and Eigenvalues

	Eigenvalue	Variance Explained (%)	Cumulative Variance (%)
PC1	8.87	80.68	80.68
PC2	0.75	6.79	87.47
PC3	0.53	4.83	92.30
PC4	0.31	2.86	95.16
PC5	0.21	1.94	97.11

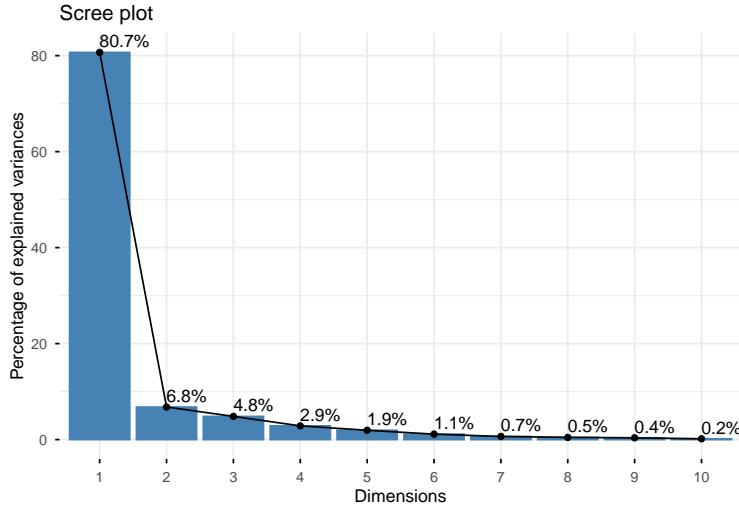


Figure 2: Variance Explained by each Principal Component

2.2 Interpretation of Principal Components

To make these factors easy to interpret, I applied a “Varimax rotation”, a mathematical sharpening tool that ensures each variable clearly belongs to one specific factor. This allowed us to name the four dimensions:

Table 6: Rotated Component Loadings for 11 Variables.

	PC1: Scale and Finance	PC2: Competitive Entrenchment	PC3: Physical and Continental Pedigree
Avg_Pts	0.57	0.74	0.22
Avg_xGD	0.59	0.77	0.15
Avg_Poss	0.66	0.60	0.13
Avg_UEFA_Coef	0.62	0.45	0.46
Szns	0.11	0.87	0.43
Sqd_Mkt_Val	0.68	0.65	0.19
Wage_Bill	0.72	0.56	0.35
Com_Rev	0.85	0.32	0.37
Tot_Rev	0.78	0.45	0.40
Std_Cap	0.41	0.29	0.85
Ig_Fol	0.89	0.20	0.33

- **PC1: Absolute Scale and Finance (The "Ceiling"):** Defined by **Commercial Revenue (0.85)** and **Instagram Followers (0.89)**. This factor sets the competitive ceiling—the financial and global brand gravity that separates the "Global Giants" from domestic participants.
- **PC2: Competitive Entrenchment (The "Stability"):** Overwhelmingly defined by **Premier League Tenure (0.87)** and high **Points/xGD**. This factor distinguishes the "Established Vanguard" from transient or "Yo-Yo" clubs.
- **PC3: Physical and Continental Pedigree:** Driven by **Stadium Capacity (0.85)** and **UEFA History**. This dimension captures the physical infrastructure and historical prestige of a club.

2.3 Visualising the Competitive Landscape (The Biplot)

The relationship between the clubs and the newly defined factors is visually mapped using a **Biplot** (Figure 3). This chart displays the clubs' coordinates based on their scores for **Dim1 (Absolute Scale)** and **Dim2 (Competitive Entrenchment)**.

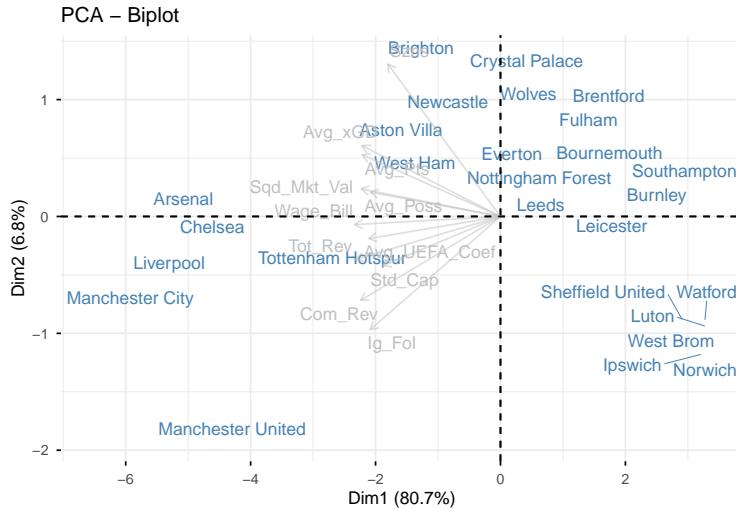


Figure 3: Competitive Landscape Biplot of Premier League Clubs

Key Spatial Observations:

- **The Horizontal Axis (PC1 - 80.7%):** This represents a club's financial and brand "gravity." The further to the left a club sits, the higher their revenue, global following, and squad value.
- **The Vertical Axis (PC2 - 6.8%):** This represents stability. Clubs positioned higher on this axis possess greater longevity and consistent sporting output.
- **The Elite Chasm:** The "Big Six" are pulled to the extreme left. Interestingly, **Manchester United** shows a distinct vertical deviation from the elite group; while their scale (PC1) remains massive, their position is driven heavily by an outlier status in **Instagram Followers**, pulling them further down the axis compared to the sporting-led positions of Manchester City and Arsenal.
- **The Challenger Vanguard:** A group of "Vanguard" clubs, led by **Brighton, Newcastle, and Aston Villa**, are clearly pulled upward and toward the center. This group has successfully distanced itself from the "Yo-Yo" cluster in the bottom-right (containing **West Brom, Norwich, and Luton**), who currently lack both the financial scale and the tenure to move toward the center.

3. Determination of Competitive Tiers (Clustering Validation)

This stage identifies the optimal number of competitive “groups” (k) within the Premier League. To ensure these groups are meaningful, clubs are clustered based on their scores across the three **Master Factors** (PC1–PC3) rather than raw, noisy stats.

3.1 Hierarchical Clustering: The “Family Tree”

Hierarchical Clustering was employed to visualise the natural “tribes” within the league. The resulting **Dendrogram** (Figure 4) functions as a family tree of competitive profiles; clubs connected by lower “branches” share a more similar statistical DNA.

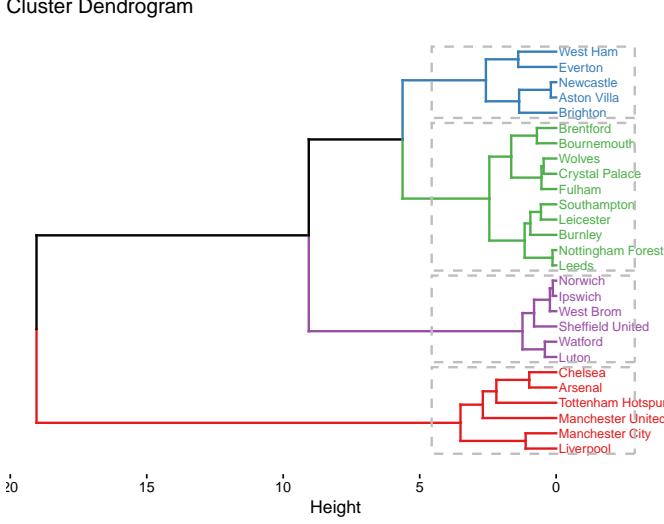


Figure 4: Dendrogram of Hierarchical Clustering on Three Principal Components

The dendrogram reveals a highly structured league. The primary split separates the global elite from the rest of the competition. Critically, the branches now show a clear secondary division: the “**Vanguard**” (Newcastle, Aston Villa, Brighton, etc.) forms a distinct sub-branch, while the bottom-tier clubs, including West Brom, Ipswich and Norwich, are grouped logically with other “Yo-Yo” sides like Watford and Luton. This structure suggests that a four-tier model captures the nuances of the league’s competitive tiers with high precision.

3.2 The Elbow Criterion: Finding the “Sweet Spot”

To move from visual estimation to mathematical certainty, the **Elbow Method** was used. This calculates the “Within-Cluster Sum of Squares” (WSS), essentially a measure of how “tight” and cohesive the groups are.

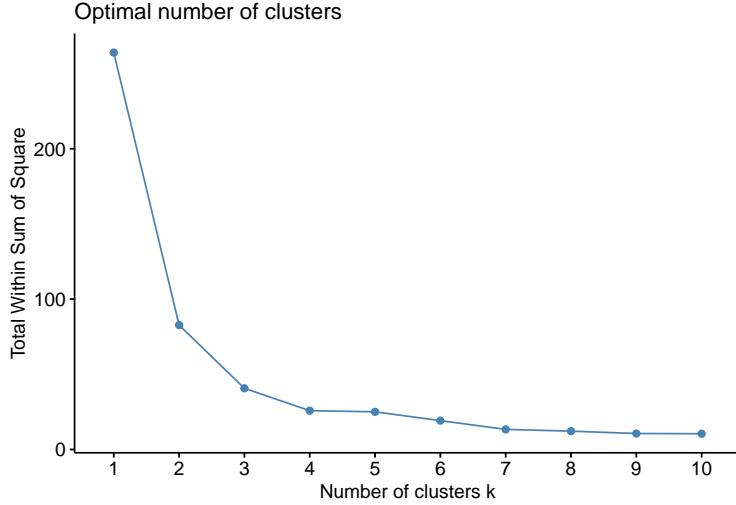


Figure 5: Elbow Plot for the Optimal Number of Clusters

As shown in the plot, a clear “elbow” appears at $k = 4$. While $k = 5$ is statistically plausible and provides a slightly “tighter” model, I have chosen $k = 4$ based on the **Principle of Parsimony**. This principle

suggests that the simplest explanation that fits the data is usually the best; a four-tier structure provides the most practical and clear framework for defining the league's primary hierarchy without over-complicating the narrative.

Observation on Higher Resolutions ($k = 5$ to $k = 8$)

While $k = 4$ is the most robust baseline, the **Dendrogram** reveals a telling sequence of “power splits” as the resolution increases:

- **The "Super-Elite" ($k = 5$):** The first major refinement splits the Big Six, isolating **Manchester City** and **Liverpool** into their own tier—a mathematical nod to the “90-point era” they have defined.
- **The Brand Outlier ($k = 6$):** At six clusters, **Manchester United** splits from the remaining giants. This likely reflects their unique statistical profile: elite global reach (PC1) decoupled from recent sporting stability (PC2).
- **The Vanguard Stratification ($k = 7$):** Interestingly, the next split occurs within Tier 2, separating **Newcastle, Aston Villa, and Brighton** from West Ham and Everton. This identifies the “Inner Vanguard”—the clubs currently closest to the elite center.
- **Bottom-Tier and Spurs Splits ($k = 8+$):** Further splits eventually separate the “Stable Middle” from the “Extreme Yo-Yos” (West Brom, Luton, Norwich) and isolate **Tottenham Hotspur** from the remaining London giants.

Despite these nuances, $k = 4$ remains the optimal choice for this report, as it most cleanly identifies the “**Financial Chasm**” and the “**Vanguard**” challengers.

4. Identification and Analysis of Competitive Tiers

Using our validated $k = 4$, the final **K-Means Clustering** algorithm was applied. This officially assigns every club into one of four distinct tiers based on their scores across the three Master Factors: Absolute Scale (PC1), Competitive Stability (PC2), and Pedigree (PC3).

4.1 The Competitive Profile of Each Tier

To understand the “character” of these tiers, the average scores across our four Master Factors were examined. Figure 6 visualises these profiles, acting as a “DNA strip” for each group.

Note: These bars represent deviations from the league average (0.0). A bar pointing away from the center indicates a tier is significantly higher or lower than the rest of the league in that specific dimension.

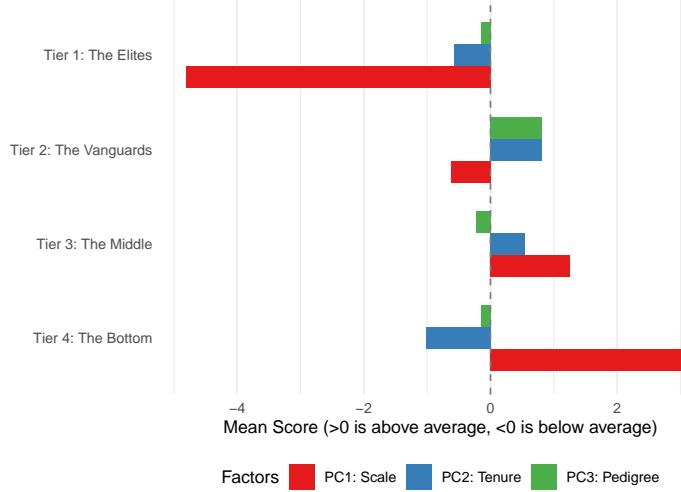


Figure 6: Competitive Profile of Each Tier

The bar chart (Figure 6) reveals the distinct identities of each group. While **Tier 1 (The Elites)** shows the most extreme financial scale (PC1), their performance stability (PC2) is more varied than one might expect, reflecting the high-pressure, high-volatility nature of competing for titles and the ‘Manchester United effect’ of high spend but fluctuating results.”

Tier 2 (The Vanguards) emerges as the efficiency leader of the league. Their position at the positive pole of PC2 indicates they are the most ‘over-indexed’ for stability—they maximize their sporting output relative to their financial footprint. In contrast, **Tier 3 and Tier 4** show progressively lower scores across all dimensions, with Tier 4 defined by a near-total lack of the financial or historical ‘Scale’ required to break out of the bottom quartile.

Spatial Validation (The Colored Biplot)

By re-plotting our “Map of the Premier League” and coloring the clubs by their assigned tier, we can visualise the physical “**Chasm**” in the league’s structure.

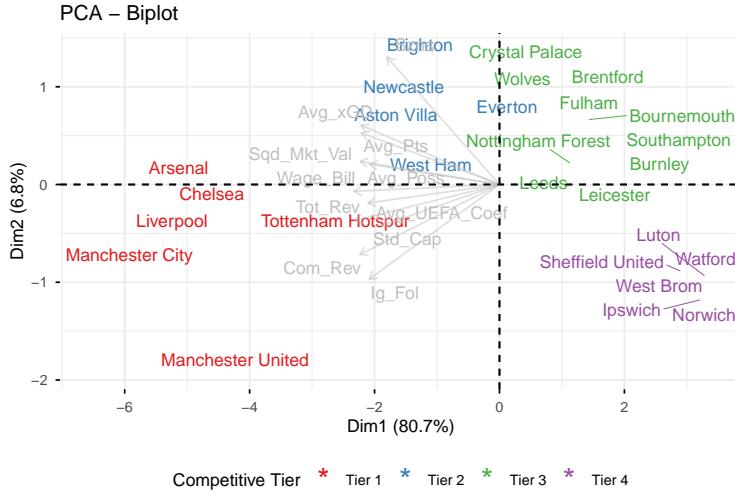


Figure 7: Competitive Landscape Biplot with Clubs Colored by Final Competitive Tier ($k=4$)

As the map illustrates, the **Tier 1 clubs (The Big Six)** are completely isolated on the far left. **Tier 2 (The Vanguard)** occupies the “upper-center” ground, pulling away from the general population toward the elite. **Tier 3 (The Middle)** clusters in the center, while **Tier 4** occupies the high-risk “bottom-right” quadrant. This confirms a strictly stratified hierarchy where a club’s movement between tiers is a result of long-term structural shifts rather than transient form.

4.3 Final Tier Membership

The following membership reflects the structure of the league over the five-season window:

- **Tier 1 (The Elite):** Arsenal, Chelsea, Liverpool, Manchester City, Manchester United, Tottenham Hotspur.
- **Tier 2 (The Vanguard):** Aston Villa, Newcastle United, Brighton, West Ham United, Everton.
- **Tier 3 (The Middle):** Crystal Palace, Wolves, Brentford, Fulham, Bournemouth, Nottingham Forest, Leicester City, Leeds United, Southampton, Burnley.
- **Tier 4 (The Bottom):** West Bromwich Albion, Norwich City, Sheffield United, Watford, Luton Town, Ipswich Town.

5. Discussion: The Mechanics of Stratification

This section interprets the structural findings to address the central question of this report: how “closed” is the Big Six, and what are the specific mechanics that prevent Tier 2 clubs from bridging the gap?

5.1 The Brand-Performance Paradox: The Manchester United Case

The most fascinating spatial discovery in the current model is the vertical isolation of **Manchester United**. While the other “Elite” members cluster relatively tightly on the left, United drifts significantly downward on the vertical axis (PC2: Stability/Tenure).

This suggests that United’s “Elite” status is currently sustained almost entirely by **PC1 (Absolute Scale and Brand)**, driven by their massive commercial revenue and global Instagram following, rather than the

high-efficiency sporting stability seen in **Manchester City or Liverpool**. United serves as the primary evidence that a “Financial Moat” can protect a club’s status even when sporting output begins to fluctuate toward mid-table norms.

5.2 The “Financial Chasm” and the Commercial Moat

The primary barrier to entry remains the extreme disparity in global brand scale. The transition from Tier 2 to Tier 1 is not a gradual slope but a statistical “No-Man’s Land.”

Table 7: Mean Financial and Market Scale Metrics: Tier 1 vs 2

Tier	Avg. Wage Bill	Avg. Total Revenue	Avg. Commercial Rev	Avg. Squad Value	Avg. Instagram
Tier 1	341850000	600883333	275800000	1049608333	44016667
Tier 2	186900000	255020000	54720000	505260000	3500000

The data confirms that while Vanguard clubs can compete in terms of squad value or occasional points tallies, they lack the **Global Brand Footprint (PC1)** required to sustain elite spending. A Tier 2 club is essentially playing a domestic game, while Tier 1 clubs are playing a global one.

5.2 Dimensional Distance and Elite Homogeneity

To measure this gap precisely, the **Mapping Distance** (mathematically known as Euclidean distance) of each club was calculated from the “Elite Center”—the average score of the Big Six. This reveals how remarkably consistent the Big Six are.

5.3 The Vanguard: Stability as a Launchpad

The model naturally identifies a “Vanguard” class (Aston Villa, Newcastle, Brighton, West Ham, and Everton) that has successfully separated itself from the Tier 3 “Middle”.

- **Aston Villa and Newcastle:** These clubs are utilising a "Dual-Track" strategy—leveraging significant historical infrastructure (PC3) while aggressively scaling their financial inputs to move horizontally toward the Elite.
- **Brighton & Hove Albion:** Brighton represents the "Efficiency" model. They occupy the highest vertical position in the Vanguard, suggesting that their status is built on **Competitive Stability (PC2)**. They have optimised their "Tenure" and "xGD" to such an extent that they sit closer to the Elite on the Biplot than clubs with significantly higher revenues.

5.4 The “Yo-Yo” Gravity of Tier 4

Tier 4 (The Bottom), including West brom, Norwich, and Sheffield United suffers from a lack of “Competitive Entrenchment” (PC2). Without the foundation of long-term Premier League tenure, these clubs cannot accumulate the commercial revenue necessary to build a “moat,” leaving them perpetually vulnerable to the “gravity” of the Championship.

6. The Challenger Audit: Quantifying the Glass Ceiling

While clustering identifies the competitive groups, the **Euclidean Distance Audit** quantifies the exact mathematical effort required for a Tier 2 club to “break” the Big Six.

6.1 The Stability of the Elite Core

To establish a baseline, the distance of each “Big Six” club from the **Elite Centroid** (the group average) was calculated. This reveals a defined elite structure where all six member clubs are contained within a 2.0-unit radius of the group average, creating a distinct statistical island compared to the rest of the league.

Table 8: Internal Cohesion of the Elite Tier

Team	Distance To Mean
Arsenal	0.62
Liverpool	0.89
Chelsea	1.02
Manchester United	1.78
Manchester City	1.98
Tottenham Hotspur	1.99

6.2 Ranking the Vanguard: The “Tier 1.5” Leaderboard

The distances were measured from the Elite Centroid to the top challengers in Tier 2. The results show a significant “Structural Chasm”. **Aston Villa (3.90)** and **Newcastle (4.04)** are almost mathematically twice as far from the Elite center as the most distant Elite club (Tottenham, 1.99) is from its own group.

Table 9: Distance of Challengers to the Elite Cluster

Team	Distance To Elite
Aston Villa	3.90
Newcastle	4.04
Brighton	4.63
West Ham	4.69
Everton	5.66

6.3 Diagnostic Audit: Why the Gap Persists

By decomposing these distances into our three Master Factors, we can identify the specific “structural wall” each club is hitting.

Table 10: Percentage Contribution of Each Factor to the Elite Gap

Team	PC1: Scale Gap	PC2: Stability Gap	PC3: Pedigree Gap
Aston Villa	83.59	13.36	3.04
Brighton	82.11	17.81	0.08
Everton	88.82	4.69	6.49
Newcastle	83.17	12.63	4.19
West Ham	77.35	3.79	18.86

Key Audit Findings:

- **The Commercial Moat (PC1):** For every major challenger, **Absolute Scale (PC1)**, encompassing commercial revenue and global brand reach, is responsible for over 80% of their distance from the elite. This confirms that the Big Six is not a sporting monopoly, but a financial and global brand monopoly.

- **The Newcastle/Villa Paradox:** Both clubs show very low **Pedigree Gaps** (<5%), meaning their historical status and infrastructure are already at an elite level. Their only remaining barrier is the horizontal "Scale" axis.
- **The Brighton Efficiency Limit:** Brighton has almost zero **Pedigree Gap** (0.08%) relative to the elite, yet they suffer from a high **Stability Gap** (17.81%). This suggests that while they are historically "pedigreed" enough to belong, their lack of long-term elite tenure (PC2) and global scale (PC1) creates a combined structural barrier that scouting efficiency alone cannot solve.
- **West Ham's Infrastructure Wall:** Unlike the others, West Ham faces a notable **Pedigree/Infrastructure Gap** (18.86%). While they have scaled their finances, their path to the elite is uniquely hindered by the physical and historical variables measured in PC3.
- **Everton's Scale Barrier:** Everton possesses the highest **Scale Gap** (88.82%) in the audit. Despite their high stability and pedigree (lower gaps in PC2 and PC3), their inability to convert history and tenure into elite-level commercial revenue keeps them statistically tethered to Tier 2.

Conclusion: The Structural Calculus of Success

The objective of this report was to move beyond the subjective “Big Six” narrative and apply a multi-dimensional statistical framework to the hierarchy of the Premier League. Through Principal Component Analysis (PCA) and K-Means Clustering, this study has successfully quantified the structural barriers that define modern English football.

Synthesis of Findings

The primary finding confirms that the “Big Six” is not merely a media construct, but a statistically fortified elite. The analysis revealed a significant **1.91-unit Euclidean chasm** between the most distant elite club (Tottenham) and the strongest challenger (Aston Villa). This gap is primarily driven by a **“Commercial Moat” (PC1)**—a self-sustaining cycle where global brand reach and non-matchday revenue provide a financial floor that short-term sporting failure cannot easily collapse.

The model naturally identified the emergence of a **“Vanguard” (Tier 2)**. Clubs such as **Aston Villa, Newcastle, Brighton, West Ham, and Everton** have effectively separated themselves from the “Stable Middle”. This group represents a distinct competitive tier that sits at the threshold of the elite, defined by high stability and historical pedigree.

The Manchester United Paradox and Structural Weight

A notable finding in the final model is the vertical isolation of Manchester United. Despite sporting volatility, United’s massive scores in **Absolute Scale (PC1)** provide enough “structural weight” to keep them firmly within the Elite cluster. This reinforces the theory that the Premier League hierarchy is built on “slow-moving” variables; global brand power acts as a safety net, making decline a multi-year process regardless of on-pitch results.

Model Limitations

While the PCA-K-Means approach provides a rigorous objective mapping, certain limitations remain:

- **Smoothing of Volatility:** By using a five-year aggregate, the model "smoothes" rapid spikes or drops in form. While excellent for identifying long-term power, it can mask "Black Swan" events, such as Leicester City's Premier League title or Newcastle United's sudden change in ownership.
- **Variable Correlation:** The model relies on the high correlation between factors like “Commercial Revenue” and “Instagram Followers”. While effective for clustering, it assumes these variables will continue to grow in tandem in a shifting digital landscape.

Opportunities for Further Analysis

The findings of this report provide a foundation for more granular temporal studies. Future analysis should consider:

- **Temporal Momentum Studies:** Re-running this methodology on a 24-month “Rolling Window” would allow for a Recency Bias Analysis. This would highlight whether the **Vanguard** is currently accelerating toward the elite or merely stagnating at the “glass ceiling”.
- **Weighted-Decay Models:** Applying a “Time-Decay” factor—where recent seasons are weighted more heavily—would offer a more accurate “Predictive Hierarchy” to better capture the rapid rise of clubs like **Newcastle** or **Brighton**.

Final Outlook

In conclusion, while the Premier League remains competitive on a match-to-match basis, it is structurally stratified. The **Big Six** remain isolated on a high plateau, protected by a decade of commercial accumulation. However, the identification of a distinct **Vanguard** proves that the hierarchy is not static. For the first time in the modern era, a “Base Camp” has been established at the foot of the elite mountain, providing a mathematical blueprint for how clubs might eventually bridge the structural chasm.

References

- FotMob. 2025. “Premier League Player and Club Statistics.” <https://www.fotmob.com>.
- Instagram. 2025. “Official Club Profile Follower Counts.” <https://www.instagram.com>.
- Jolliffe, Ian T. 2002. *Principal Component Analysis*. 2nd ed. Springer.
- Kaufman, Leonard, and Peter J Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons.
- MacQueen, James. 1967. “Some Methods for Classification and Analysis of Multivariate Observations.” *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1: 281–97.
- R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Ramble, Swiss. 2024. “Premier League Financial Analysis: 2023/24 Season Review.” <https://swissramble.substack.com>.
- Sneath, Peter HA, and Robert R Sokal. 1973. *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. W. H. Freeman.
- UEFA. 2025. “UEFA Association Club Coefficients.” <https://www.uefa.com/nationalassociations/uefarankings/club/?year=2025>.
- Wikipedia. 2024. “Big Six (Premier League).” [https://en.wikipedia.org/wiki/Big_Six_\(Premier_League\)](https://en.wikipedia.org/wiki/Big_Six_(Premier_League)).
- . 2025. “List of Premier League Stadiums.” https://en.wikipedia.org/wiki/List_of_Premier_League_stadiums.