SMDE Laboratory Assignment
First Assignment Q4

,
Aftab Ahmed Choudhry, aftab.ahmed.choudhry@estudiantat.upc.edu
Roger Canal, roger.canal@estudiantat.upc.edu
Miquel Sala Franci, miquel.sala.franci@estudiantat.upc.edu

January 2026
Course 2025/2026

Abstract

This study applies multivariate statistical techniques to analyze basketball player performance and predict player efficiency (EFF). Using Principal Component Analysis (PCA), we reduced the dimensionality of the dataset, identifying three key latent structures: "Offensive Volume," "Inside/Outside Playstyle," and "Playmaking Responsibility." These three components cumulatively explain 73.26% of the total variance in player statistics.

Subsequently, a Principal Component Regression (PCR) was conducted to model efficiency based on these extracted dimensions. Backward elimination revealed that the first dimension (Volume) is the primary determinant of efficiency, yielding a final model with an adjusted R^2 of approximately 94.5%. Diagnostic tests confirmed the assumptions of normality and homoscedasticity; however, the Durbin-Watson test indicated significant autocorrelation, suggesting the presence of latent clustering effects (e.g., team dynamics) within the data. These findings demonstrate that while player style varies significantly, efficiency is overwhelmingly driven by overall offensive production volume.

Contents

Chapter 1

QUESTION 4: PCA and Regression Analysis of Eurobasket 2025

1.1 Section A: Data Preparation and Structure

The "Eurobasket 2025" dataset was preprocessed for multivariate analysis. A critical step involved reassigning the `PLAYER` variable as row identifiers; this excludes text from mathematical computations while retaining player labels for graphical interpretation. Finally, the data structure was verified using summary statistics to ensure variable integrity prior to analysis.

1.2 Section B: Variable Classification and Preprocessing

The dataset structure was verified to ensure correct data typing. The categorical variables `Team` and `Position` were converted from character strings to factors to function as supplementary qualitative variables in the PCA. Notably, `Position` consolidates players into three distinct roles (**Center**, **Forward**, and **Guard**), which facilitates the analysis of stylistic clustering within the projection space.

1.3 Section C: Correlation Analysis and Feasibility of PCA

To isolate inter-variable relationships, a subset (`df_pca`) was created containing only the relevant numerical metrics and player position, preparing the data for dimensionality reduction.

1.3.1 Visual Inspection of Correlations

Linearity and association strength were assessed using a scatterplot matrix (Figure 1.1) and a Pearson correlation heatmap (Figure 1.2). Visual inspection confirms distinct patterns of multicollinearity across the dataset:

- **Minutes and Volume:** Playing time (`MIN`) exhibits a broad positive correlation with `DREB`, `AST`, `EFF`, and `PTS`, indicating that minutes played serves as a proxy for overall statistical accumulation.
- **Scoring and Efficiency:** The Efficiency metric (`EFF`) acts as a central hub, correlating strongly with `PTS`, `REB` (and `DREB`), `AST`, and even `TO`. Furthermore, `PTS` correlates with `FT`, `DREB`, and `REB`, reinforcing the link between scoring volume and general productivity.

- **Inside/Outside Dynamics:** Field Goals (`FG`) correlate with `2PT FG` and `BLK`, reflecting inside play, whereas `3PT FG` shows a specific relationship with `FT`.
- **Rebounding and Playmaking:** Rebounding metrics are highly interconnected; `OREB` correlates with `DREB`, `REB`, and `EFF`. Similarly, playmaking metrics show an expected trade-off, where `AST` correlates with both `TO` and `EFF`.

Figure 1.1: Scatterplot Matrix showing linearity between variables.

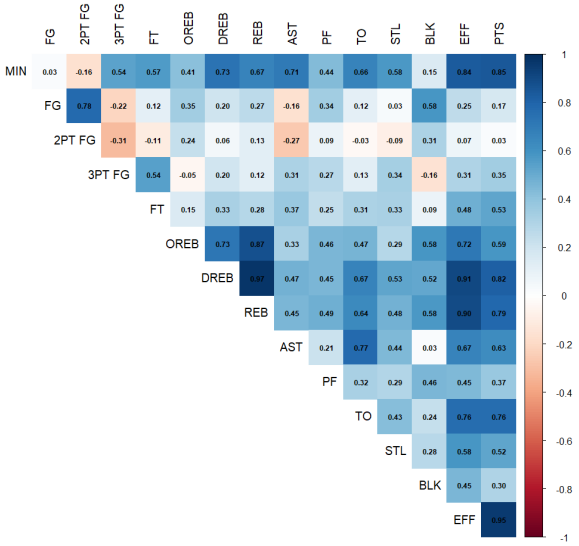


Figure 1.2: Correlation Matrix Heatmap quantifying relationship strength.

These strong multicollinearities indicate significant redundancy in the dataset, suggesting that Principal Component Analysis (PCA) is an appropriate method to reduce dimensionality without significant information loss.

1.3.2 Statistical Tests for Sampling Adequacy

To formally verify the suitability of the data for factor analysis, we performed two key diagnostic tests:

- **Bartlett's Test of Sphericity:** The test yielded a statistically significant result ($p < 2.2 \times 10^{-16}$), rejecting the null hypothesis of an identity matrix. This confirms the presence of sufficient inter-variable correlations to justify the application of PCA.
- **Kaiser-Meyer-Olkin (KMO) Test:** The initial test yielded an MSA of 0.65. However, removing the composite variable `EFF` to avoid artificial redundancy improved the statistic to **0.73**. This result falls within Kaiser's "good" classification, confirming the dataset's suitability for dimensionality reduction.

1.4 Section D: PCA Specification and Variable Roles

With the suitability of the data confirmed, Principal Component Analysis (PCA) was performed on the dataset using the `FactoMineR` package. A crucial step in this process was correctly assigning

the roles of the active and supplementary variables to ensure the resulting dimensions purely reflect player performance style rather than position or efficiency formulas. The PCA was specified with the following structure:

- **Active Variables:** The 14 remaining numerical performance metrics (e.g., `PTS`, `AST`, `REB`, `MIN`) were used as active variables. These are the inputs that mathematically construct the principal components (axes) and determine the eigenvalues.
- **Supplementary Quantitative Variable:** The Efficiency metric (`EFF`) was set as a *quantitative supplementary variable* (`quanti.sup`). This means it did not influence the calculation of the dimensions but was projected onto the final space. This allows us to visually assess how "overall efficiency" correlates with the extracted playing styles without artificially driving the correlations.
- **Supplementary Qualitative Variable:** The `Position` variable was set as a *qualitative supplementary variable* (`quali.sup`). This prevents the player's position from altering the statistical construction of the map while allowing us to color-code and group individuals later to interpret the biological meaning of the dimensions (e.g., verifying if "Centers" naturally cluster in specific areas).

By segregating these variables, the analysis ensures that the extracted dimensions represent organic statistical variations in performance, which can then be cross-referenced with external labels (Position) and composite scores (Efficiency).

1.5 Section E: Eigenvalue Analysis and Component Selection

To determine the optimal number of principal components to retain, we examined the eigenvalues and the proportion of variance explained by each dimension. The scree plot (Figure 1.3) and the numerical eigenvalue table were utilized to apply standard selection criteria.

1.5.1 Selection Criteria

We employed two primary methods for component selection:

- **Kaiser's Criterion:** This rule suggests retaining only components with an eigenvalue greater than 1, as they explain more variance than a single standardized variable. As shown in Table 1.1, the first three components meet this threshold ($\lambda_1 = 6.25$, $\lambda_2 = 2.76$, $\lambda_3 = 1.25$), while the fourth drops below 1 ($\lambda_4 = 0.91$).
- **Scree Plot Inspection:** Figure 1.3 displays the "elbow" of the graph. A distinct inflection point is observed after the third component, indicating diminishing returns in explained variance for subsequent dimensions.

1.5.2 Variance Explained

Based on these criteria, we decided to retain the first **three principal components**. These three dimensions cumulatively explain **73.26%** of the total inertia in the dataset. This substantial coverage ensures that the reduced dimensional space accurately represents the underlying structure of player performance with minimal loss of information.

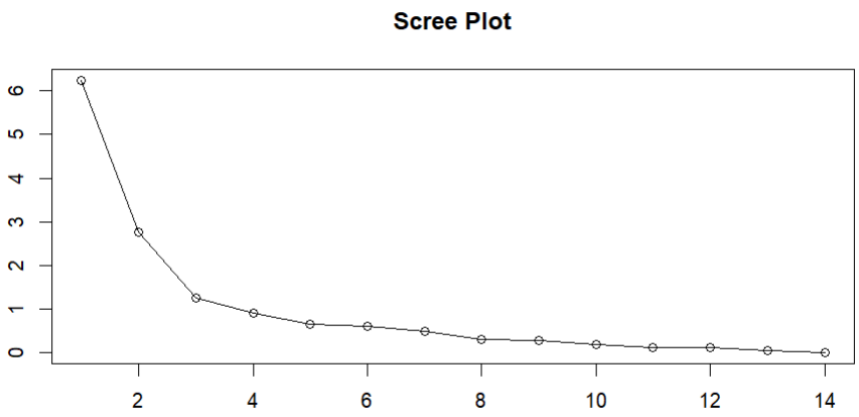


Figure 1.3: Scree Plot showing the eigenvalues of the principal components. The "elbow" confirms the retention of 3 dimensions.

Component	Eigenvalue	% Variance	Cumulative %
Comp 1	6.25	44.62	44.62
Comp 2	2.76	19.71	64.32
Comp 3	1.25	8.94	**73.26**
Comp 4	0.91	6.52	79.78

Table 1.1: Eigenvalues and variance explained by the first four components.

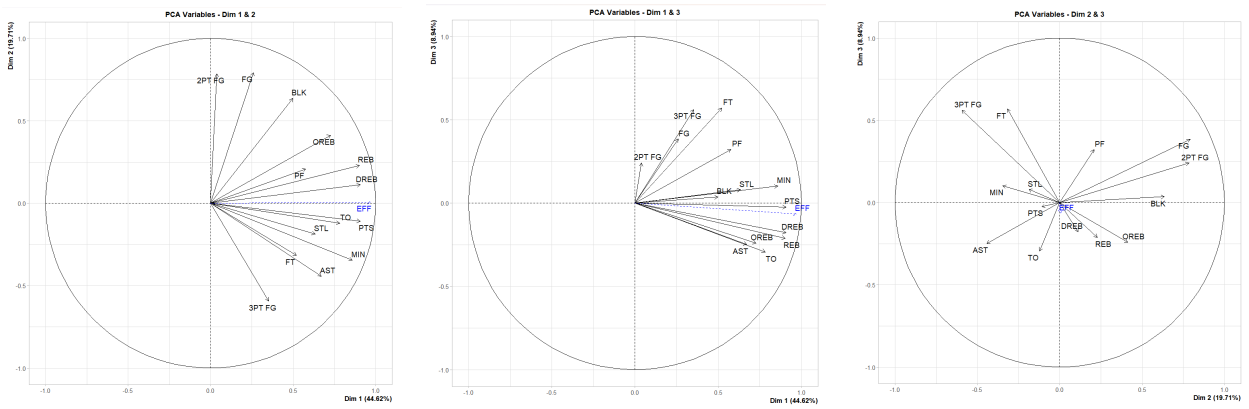
1.6 Section F: Interpretation of Principal Components

The interpretation of the extracted dimensions was conducted by examining the factor loadings, which represent the correlations between the original variables and the principal components. This analysis combines visual inspection of the correlation circles (Figure 1.4) with a precise numerical assessment of the coordinate matrix to define the underlying structure of each axis.

1.6.1 Visual Analysis of Dimensions

Figure 1.4 displays the projection of variables onto the three combinations of the extracted axes, providing a comprehensive view of the variable relationships:

- **Dim 1 vs. Dim 2 (Left):** This plot illustrates the primary separation in the dataset. Most variables point to the right (Dim 1), indicating performance volume, while the vertical axis (Dim 2) clearly opposes "inside" stats (top) against "perimeter" stats (bottom).
- **Dim 1 vs. Dim 3 (Center):** Here, we observe a nuanced separation where "shooting" metrics like FT and 3PT FG pull upwards, distinguishing them from the playmaking metrics (AST, TO) which pull downwards.
- **Dim 2 vs. Dim 3 (Right):** By removing the performance volume (Dim 1), this plot isolates pure playstyle, showing an orthogonal relationship between the "Inside/Outside" role and the "Playmaking/Shooting" responsibility.



Dim 1 vs. Dim 2 Dim 1 vs. Dim 3 Dim 2 vs. Dim 3

Figure 1.4: Correlation circles for the three retained dimensions, visualizing the variable loadings.

1.6.2 Dimension Naming

Based on the numerical coordinates (`$var$coord`), the dimensions are interpreted as follows:

Dimension 1: "Overall Performance / Offensive Volume"

This dimension explains **44.62%** of the variance. It acts as a measure of general productivity, showing very strong positive correlations with nearly all offensive volume metrics: `PTS` ($r = 0.91$), `DREB` ($r = 0.91$), `REB` ($r = 0.90$), and `MIN` ($r = 0.86$). This axis effectively separates high-performing starters (positive values) from bench players (negative values).

Dimension 2: "Inside vs. Outside Playstyle"

Explaining **19.71%** of the variance, this dimension captures the structural role of the player on the court.

- *Positive Direction (Inside):* Strongly driven by "Big Man" statistics such as `FG` (0.79), `2PT FG` (0.79), and `BLK` (0.64).
- *Negative Direction (Outside):* Driven by perimeter statistics like `3PT FG` (-0.59) and `AST` (-0.44).

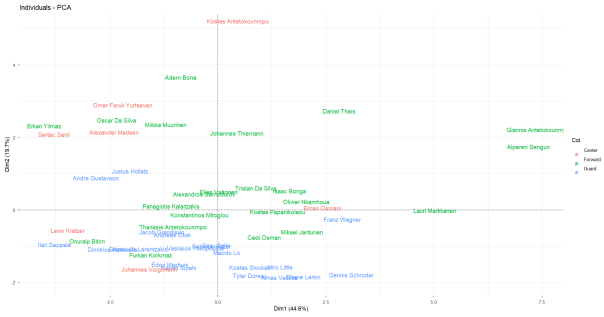
Dimension 3: "Shooting vs. Playmaking Responsibility"

Explaining **8.94%** of the variance, this axis refines the guard roles by distinguishing between scorers and facilitators.

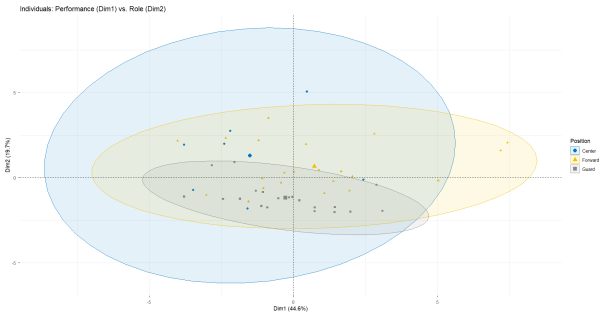
- *Positive Direction (Shooting):* Correlated with `FT` (0.57) and `3PT FG` (0.56), indicating pure scoring or finishing actions.
- *Negative Direction (Playmaking):* Correlated with `TO` (-0.29) and `AST` (-0.25), indicating ball-handling duties that carry turnover risk.

1.7 Section G: Analysis of Individual Patterns and Positional Clustering

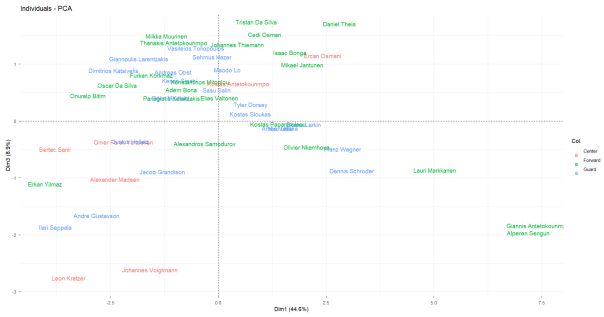
Having defined the statistical dimensions, we projected the individuals (players) onto the new subspace to identify performance hierarchies and stylistic clusters. Figure 1.5 presents the individual maps, enriched by the supplementary categorical variable `Position`.



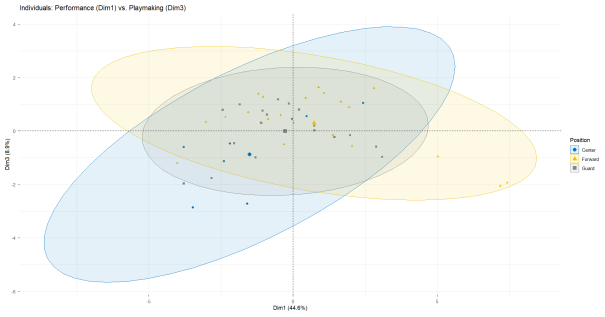
A. Individuals (Dim 1 vs. Dim 2)



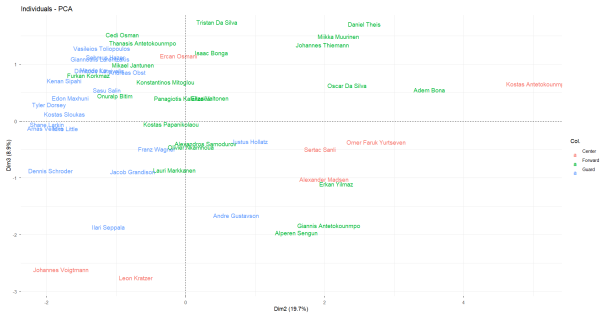
B. Clustering by Position (Dim 1 vs. Dim 2)



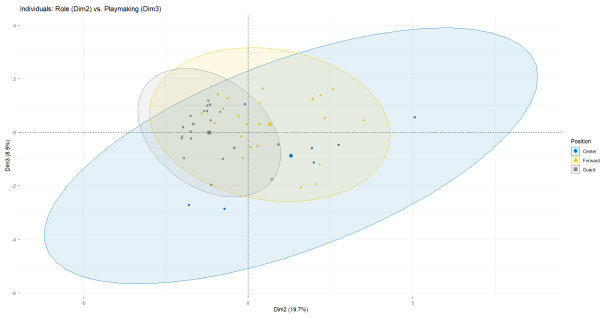
C. Individuals (Dim 1 vs. Dim 3)



D. Clustering by Position (Dim 1 vs. Dim 3)



E. Individuals (Dim 2 vs. Dim 3)



F. Clustering by Position (Dim 2 vs. Dim 3)

Figure 1.5: Projection of individuals onto the principal planes. Left column displays player names; Right column displays concentration ellipses grouped by Position.

1.7.1 Interpretation of Player Positions

The analysis of the individual coordinates (`$ind$coord`) reveals distinct hierarchies and groupings that validate our dimension naming.

Performance Hierarchy (Dimension 1)

The first dimension acts as a clear "Superstar Index."

- **The Elite:** Players with the highest positive coordinates are the statistical leaders of the tournament. **Giannis Antetokounmpo** (7.42) and **Alperen Sengun** (7.18) are extreme outliers on the far right, reflecting their dominance in volume metrics (Points, Rebounds, Minutes). They are followed by **Lauri Markkanen** (5.02).
- **Role Players:** Players with negative coordinates on Dim 1 (e.g., Erkan Yilmaz at -4.02) correspond to bench players with limited minutes and statistical output.

Role Differentiation (Dimension 2)

The second dimension provides a near-perfect separation of playing positions, validating the "Inside vs. Outside" interpretation.

- **Centers (Top):** Players with high positive scores are exclusively "Paint" operators. **Kostas Antetokounmpo** (5.06) and **Adem Bona** (3.51) appear at the top, driven by high blocks and FG% without perimeter play. The blue ellipses in Plot B clearly occupy the upper region.
- **Guards (Bottom):** Players with high negative scores are perimeter handlers. **Shane Larkin** (-2.00), **Tyler Dorsey** (-1.97), and **Dennis Schroder** (-1.95) are located at the bottom, reflecting high assist and 3-point volumes. The grey ellipses in Plot B cluster in this negative region.

Stylistic Nuances (Dimension 3)

Dimension 3 separates players based on their "Playmaking vs. Finishing" responsibility.

- **Pure Finishers (Positive):** Players like **Tristan Da Silva** (1.64) and **Daniel Theis** (1.61) appear on the positive side. These players contribute via efficient scoring (FT, FG) rather than ball dominance.
- **Ball Dominant/Grinders (Negative):** Interestingly, both heavy playmakers and physical interior players appear on the negative side, likely due to the correlation between `TO` (Turnovers) and `AST`/`REB`. **Leon Kratzer** (-2.86) and **Johannes Voigtmann** (-2.72) show high negative scores, suggesting roles that involve "dirty work" (screens, rebounding battles) or high-usage playmaking that accrues turnovers.

1.7.2 Positional Clustering

The confidence ellipses (Right Column of Figure 1.5) confirm that player roles are statistically distinct:

- **Guards (Grey)** are tightly clustered at the bottom of the maps, defined by perimeter play.
- **Centers (Blue)** are clustered at the top, defined by interior defense and rebounding.
- **Forwards (Yellow)** occupy the middle ground, overlapping with both groups. This accurately reflects modern basketball, where forwards often share duties (shooting like guards or rebounding like centers).

1.8 Section H: Principal Component Regression (PCR) Modeling

Principal Component Regression (PCR) was employed to predict Efficiency (`EFF`), utilizing uncorrelated principal components to eliminate multicollinearity and ensure stable coefficient estimates.

1.8.1 Model Selection Process

A backward elimination strategy was employed to identify the most parsimonious model, using Adjusted R^2 to compare fit quality across models with different numbers of predictors.

1. **Model 1 (PC1 + PC2 + PC3):** The initial model achieved a high Adjusted R^2 of 0.9443. However, Dimension 2 (Playstyle) was statistically insignificant ($p = 0.857$).
2. **Model 2 (PC1 + PC3):** Removing PC2 improved the fit slightly (Adj $R^2 = 0.9454$), but Dimension 3 (Playmaking) remained marginally insignificant ($p = 0.061 > 0.05$).
3. **Model 3 (PC1 only):** The final model retains only Dimension 1. Despite the simplification, the model explains **94.23%** of the variance in Efficiency (Adj $R^2 = 0.9423$), confirming that the excluded dimensions contributed negligible predictive power.

1.8.2 Final Model Equation

The relationship between Efficiency and Dimension 1 is described by the following linear equation:

$$\widehat{EFF} = 9.52 + 2.62(PC1) \tag{1.1}$$

The coefficient (2.62) indicates a strong positive relationship: as a player's overall statistical volume increases, their efficiency rating rises proportionally.

1.8.3 Verification of Assumptions

The validity of the final model was assessed using standard diagnostic tests:

- **Normality and Homoscedasticity:** The Shapiro-Wilk ($p = 0.39$) and Breusch-Pagan ($p = 0.08$) tests were both non-significant. We conclude that residuals are normally distributed and variance is constant.
- **Independence of Errors:** The Durbin-Watson test yielded a significant result ($p < 0.05$), indicating a violation of the independence assumption. This suggests autocorrelation, likely arising from latent clustering (e.g., players on the same team influencing each other's stats). While this violation implies that standard errors may be underestimated (making p-values overly optimistic), the extremely high R^2 (> 0.94) confirms that the predictive power of the model remains robust despite this limitation.

1.9 Conclusion

This study applied multivariate analysis to the Eurobasket 2025 dataset, identifying three latent performance structures: **Offensive Volume**, **Positional Role**, and **Responsibility**. Principal Component Regression revealed that while player roles vary significantly, they do not intrinsically determine efficiency. The final model demonstrated that **94.2%** of the variance in Player Efficiency is explained solely by the first dimension (Volume). This confirms that the "Efficiency" metric functions primarily as a proxy for total statistical accumulation, favoring high-usage players regardless of their specific playstyle or position.