Group 13 : Anna Monsó Rodríguez, Walter José Troiani Vargas, Joan Acero Pousa

Homework 1: Principal Component Analysis and Multidimensional Scaling

The first step was to import the dataset euroleague_23_24.csv, which includes the player statistics of four teams that participated in the Final Four of the Euro League 2023-2024. The homework statement described the variables; therefore, we just want to point out that the variable “Min” expresses the average minutes played and not the total number of minutes played (as stated in last week's theory lecture).

1. Exploratory Data Analysis

This part consists of performing the exploratory data analysis.

a. First, we deleted the “No.” variable or column from the data set since it provides no information about the observations themselves but just indicates the row number. Moreover, we can keep track of the observations using the variable “Player”, which can be used as the row identifier.

b/c. The variable “Min” expresses the average time played per player in the following format: “minutes:seconds:milliseconds”. Thus, we applied the strsplit() function to split the "Min" column into separate components based on the ":" character, resulting in a list containing three elements for each row: minutes, seconds, and milliseconds. From these components, we extracted the first element (representing the minutes) to create a new variable named "Min 2." This new variable captures the average number of minutes each player spent on the court. Note that to facilitate further analysis, we added "Min 2" as a numerical data type, allowing it to be used in mathematical calculations and statistical operations.

d. In this part, we changed the data type of the variables we considered incorrectly assigned. The variables “team”, “position”, and “player” were described as *char*, and we converted their type into *a factor* to convert them to categorical variables.

The new categorical variable “team” has four levels (Fenerbache, Olympiakos, Panathinaikos, Real Madrid) with 16 observations each. The variable “Position” counts with three levels: Center, with 13 cases, Forward, with 24 cases, and Guard, with 27 cases. Note that even though the variable “Player” has one level per row, we changed its type to *factor* in order to be able to apply specific pre-defined functions (in PCA and MDS sections) which do not allow for *char* variables.

Furthermore, we converted the variables "GP" and "GS," which initially represented "Games Played" and "Games Started" as *integer* data types, into *numeric* variables. This allows us to include them when performing further analysis of the data or applying additional methods.

Lastly, we removed the original "Min" variable from the data frame, as its valuable information is now fully captured in the new "Min 2" variable.

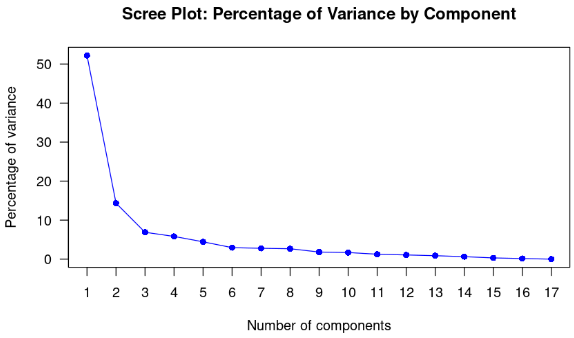
2. Principal Component Analysis

In the second part of the study, we applied Principal Component Analysis (PCA) to the dataset and extracted insights and interpretations from the results.

a. After loading the FactoMineR package, the PCA() function was used to perform Principal Component Analysis (PCA) on the dataset. The argument scale.unit = TRUE was specified to ensure that all numerical variables were standardized. Additionally, the categorical variables "Team", "Player", and "Position" were treated as qualitative supplementary variables using the quali.sup argument. This means that these variables did not directly influence the computation of the principal components but were projected onto the PCA space for interpretation. Similarly, the quantitative variable "PIR" was included as a supplementary variable using the quanti.sup argument. This allowed "PIR" to be represented in the PCA results without influencing the principal component calculations.

b. The following table and scree plot show valuable information regarding the decision on how many components to extract.

Comp.	Eigenvalue	% of variance	Cumulative % of variance
1	8.8773957830	52.21997519	52.21998
2	2.4405002344	14.35588373	66.57586
3	1.1713691822	6.89040695	73.46627
4	0.9896750398	5.82161788	79.28788
5	0.7556824940	4.44519114	83.73307



Since the PCA was performed on scaled variables, the analysis was based on the correlation matrix. This ensures that all variables, regardless of their original scales, contribute equally to the principal component calculations. Consequently, the Kaiser criterion can be used to determine the number of components to retain. According to this criterion, components with an eigenvalue of 1 or greater should be kept.

Applying the Kaiser criterion, the first three components should clearly be retained, as they all have eigenvalues above 1. The decision to keep the fourth component is slightly more subjective; while its eigenvalue is slightly below 1, it is close enough to consider it.

We applied the elbow rule to the scree plot, which showed a significant drop in explained variance after the third component, with the curve flattening beyond that point. Thus, we decided to exclude the fourth dimension, achieving a balance between dimensionality reduction and data representation. Retaining three components allows us to capture approximately

73.47% of the total variance, providing a compact yet informative representation of the original data.

c. The following table shows which five variables contribute more to each dimension, either in a positive or negative direction:

Dimension 1	Min2 = 0.956	PTS = 0.906	DR=0.867	GP= 0.844	FD = 0.841
Dimension 2	BLK = 0.796	OR = 0.679	AST = -0.566	X3P = -0.481	TR = 0.441
Dimension 3	FT = 0.614	X2P = 0.612	BLKA = -0.396	FD = -0.246	GP = 0.226

The first dimension is defined by the average minutes played, the points scored, defensive rebounds, games played, and personal faults drawn. This establishes a dimension which is strongly influenced by minutes/games played and performance metrics, where the best players will probably have larger values.

The second dimension is influenced positively by blocks, offensive rebounds, and total rebounds and negatively by assists and the percentage of three points. This dimension should differentiate the players by their playing styles/roles, separating more physical/defensive (Interior players like centre/power-forwards tend to be more defensive or Outer players like guard/wings are far more offensive) and technical/offensive players.

The percentage of free throws and two-point throws positively influences the third dimension, while the blocks and faults received negatively impact it. Since the positive loadings are almost double the negative, this dimension primarily defines the capacity of throwing from medium/short distances.

d. Using the function plot.PCA(), we plotted all the possible combinations of 2D plots, changing the dimensions in the x and y axes in order to show correlations between variables and the extracted dimensions.

e. Considering both the explanation in part c and the plots provided in the R markdown, we propose the following names for the dimensions:

- **Dimension 1:** Overall Performance and Playing Time
- **Dimension 2:** Physical vs. Technical Play
- **Dimension 3:** Mid-Range and Free-Throw Shooting Ability

f. Using the function plot.PCA(), with the argument choix set as "ind", we plotted all the individuals in each different combination of dimensions in the axes.

g. Individual plots allow us to extract some insights from the basketball players. On the one hand, players like Mathias Lessort and Walter Tavares have high overall performance and playing time and have a physical playing style. At the same time, Facundo Campazzo (Guard) and Kendrick Nunn also have high performance and playing time but are more technical.

Moreover, focusing on Plot 1 (Dimensions 1 vs. 2), we observed distinct patterns that support our hypotheses regarding the interpretation of these dimensions.

Comparing two players who are far apart along Dimension 1, such as Matthias and Alexandros Samodurov, we found substantial differences in various performance metrics. This is most notably seen in their PIR (Performance Index Rating), with Matthias scoring 19.6 compared to Alexandros's 3.0, indicating significant performance disparities. Conversely, when examining two players positioned closely on Dimension 1, such as Mustapha Fall and Nikola Milutinov, we observed remarkably similar statistics across critical metrics (e.g., X2P, X3P, DR, STL, TO, PIR), which aligns with our initial hypothesis that Dimension 1 primarily captures variations in overall player performance.

Applying a similar approach to Dimension 2, we analysed players who are close together, such as Mathias Lessort and Walter Tavares, and those who are far apart, like Mathias Lessort and Facundo Campazzo. Lessort and Tavares, both positioned similarly along Dimension 2, share a common playing style as centres, exhibiting strong performance in similar metrics (with high values on Dimension 1) and similar roles on the court. On the other hand, Lessort and Campazzo show a stark contrast in playing styles despite being high performers. This is evident when considering that Lessort, a physical centre, is vastly different from Campazzo, a more technical guard. These observations reinforce our interpretation of Dimension 2 as representing a gradient of playing style, with physicality on one end and technical skill on the other.

Furthermore, examining Plot 2 (Dimensions 1 vs. 3), we applied the same comparative approach to Dimension 3. Players positioned closely together, such as Hugo Gonzalez and Panagiotis Kalaitzakis, share similar statistics in terms of free-throw and mid-range shooting capabilities, both maintaining a perfect 100% free-throw record and above-average X2P rates with no blocked attempts (BLKA/BLK). In contrast, the differences between Mathias Lessort and Hugo Gonzalez are less pronounced along Dimension 3, reflecting discrepancies in free-throw percentage, mid-range accuracy, and blocked shots. These minor variations can be attributed to the fact that Dimension 3 explains a smaller portion of the overall variance compared to Dimensions 1 and 2.

Regarding the final plot (Dimensions 2 vs. 3), we observe distinct player profiles that highlight the differences in playing style and shooting ability. Amine Noua stands out as one of the more balanced players, demonstrating a combination of physicality and proficiency in mid-range shooting, positioning him towards the centre of both dimensions. This suggests that he can contribute effectively to various aspects of the game.

Conversely, Kendrick Nunn is located at the opposite end of these dimensions, indicating a different playing style. He appears to be a more technical player (as stated before) with less emphasis on physicality and comparatively lower performance in free-throw and mid-range shooting.

3. Multidimensional Scaling

In this last part, we applied the Multidimensional Scaling (MDS) method and extracted findings from the results obtained.

a. We started standardising the data by scaling its numerical features and then computed the Euclidean distance between each pair of players to understand their relative similarities. Then, we applied MDS to represent these distances in a lower-dimensional space, making it easier to visualise patterns and relationships between players.

b/c. The multidimensional scaling plot visually represents the relationships between players based on their standardised numerical characteristics from the players. In this plot, players positioned closer together share more similar statistical profiles, while those farther apart exhibit distinct differences in their attributes. The axes themselves represent the first two dimensions derived from the multidimensional scaling, which reflect underlying patterns in the data.

Several clusters are apparent, indicating groups of players. For example, clusters of players who play as guards or centres may emerge due to shared statistical features such as scoring efficiency, defensive statistics, or physical attributes. These clusters suggest that players within the same group may perform similarly or fulfil similar functions on their teams.

Notably, players situated far from any central cluster, such as Mathias Lessort or Walter Tavares, tend to have more extreme or unique characteristics. Their positioning suggests that they excel in particular aspects of the game or possess distinctive playing styles that set them apart from their peers. For instance, Lessort and Tavares may stand out due to exceptional physical presence or rebounding abilities, which makes them dissimilar from the majority of other players in the dataset.

d/e. We calculated the Gower distance (which also allows computing distances between non-numerical variables), including the variable “POSITION” to the data matrix, and applied metric MDS on the Gower distance matrix

f. After plotting individual player positions based on the first two dimensions, we observed noticeable differences compared to the initial plot. In the original plot, players were distributed solely based on their standardised numerical characteristics without incorporating positional information. However, in this second plot, the inclusion of the player position as a categorisation factor helps to group the players better, revealing distinct clusters or patterns that correspond to their roles on the court (e.g., guards, forwards, centres). By integrating positional information, the plot allows us to observe more meaningful patterns and better understand the underlying structure of the data, showing how various positions might impact players' statistical profiles.

g. We used various categorical and numerical variables as labels to help explain the constructed clusters and gain a deeper understanding of the plots. The selected variables for labelling were:

1. Position (categorical): When we labelled players by their positions (Guard, Forward, Center), distinct clusters emerged. Guards tended to cluster toward the bottom, Center players at the top, and Forwards occupied the middle section. This distribution provides insight into how different playing positions relate to the second dimension (y-axis).
2. Points Scored (numerical): Using total points scored as labels revealed a distribution along the first dimension (x-axis). Players with lower point totals appeared on the left side of the plot, while higher-scoring players were located on the right. This pattern indicates that Dimension 1 captures a gradient of scoring ability.
3. Minutes Played (numerical): When minutes played were used as a label, a similar trend was observed along the x-axis. Players with fewer minutes clustered on the left, whereas those with more playing time appeared on the right. This suggests that Dimension 1 may also be associated with players' overall contribution or involvement in games, as players with more minutes are likely to have a more significant impact.
4. Personal Index Rating (PIR) (numerical): Finally, plotting with PIR as the label reinforced the interpretation of the x-axis. The higher-performing players, with higher PIR values, were found on the right, while those with lower ratings were positioned on the left. This confirms that Dimension 1 captures aspects of player performance, including scoring, playing time, and overall impact.

h. Which MDS do you think better group the individuals? Why?

In a general setting, the choice between metric MDS with Euclidean distance and metric MDS using Gower's distance depends on the nature of the data and the characteristics of the variables being analysed.

In this case, Gower's metric MDS groups better the individuals, as it accounts for the mixed nature of the dataset. The presence of categorical variables like player positions or roles can have a significant impact on the players' similarity profiles. Gower's distance captures these categorical relationships along with numerical performance metrics, resulting in more meaningful clusters that reflect both statistical performance and categorical attributes. Therefore, if the data includes both numerical and categorical information, non-metric MDS with Gower's distance is likely to provide a more accurate and interpretable grouping of players.