UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

HW 1: Principal Components Analysis and Multidimensional Scaling

Group 8

Multivariate Analysis

Ercolani L., Bakas A. K. , Selis R.

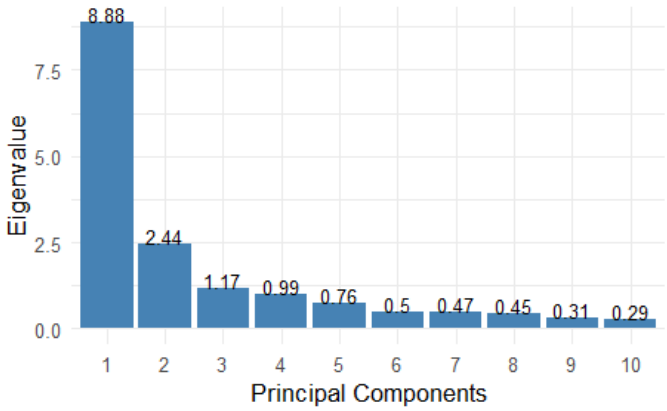
Summary: The homework focuses on analyzing EuroLeague 2023-2024 Final Four player statistics using PCA and MDS techniques. After preprocessing the dataset to correct errors, PCA was used to reduce dimensionality, capturing around 79% of the data variance using four components. The loadings showed that overall player performance was driven by metrics like minutes played and points scored. MDS highlighted the similarity between players based on their roles (Center, Forward, Guard), rather than their teams, with key performance metrics driving clustering in the reduced space.

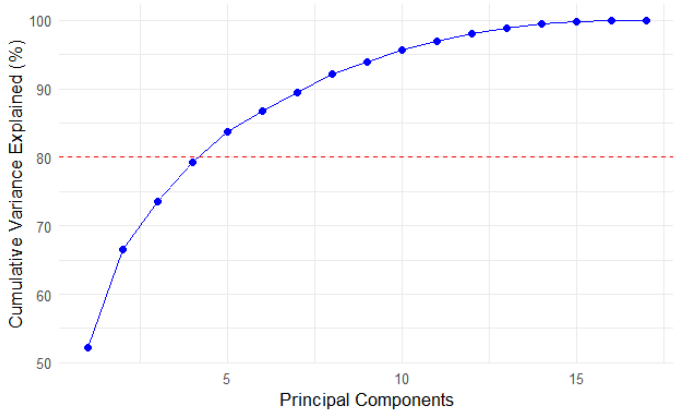
1. Preprocessing and Dataset Interpretation

The dataset consists of player statistics from the four teams that participated in the EuroLeague 2023-2024 Final Four. It contains 64 observations with 22 variables, including player names, positions, team, playing time, and a range of performance metrics such as points, assists, rebounds, and shooting percentages. Some values are recorded as percentages (e.g. shooting percentages),some are average values while others capture raw statistics.
In the exploratory data analysis (EDA), we first removed the unnecessary "No" column, which served as a redundant index in the dataset. Initially, the "Min" column, representing playing time, was interpreted incorrectly as hours:minutes:seconds, which led to unrealistic game times. Upon review, we recognized the correct format as minutes:seconds and made the necessary adjustments. The column was split, and the minutes component was used for further calculations. A new variable, “Min 2,” was created to represent the average minutes played per game, derived by dividing total minutes by the number of games played. Lastly, the datatype for categorical and numerical variables was correctly assigned. Lastly, categorical and numerical variables were properly assigned.

2. Principal Component Analysis

To visualize the variance explained by the principal components, two separate plots were created: one for the **eigenvalues** (representing the raw variance explained) and another for the **percentage of variance explained**. The first component has an eigenvalue of **8.88**, explaining **52.24%** of the variance, followed by the second component with an eigenvalue of **2.44**, accounting for **14.35%**. Together, these two components explain a total of **66.59%** of the dataset’s variability. The third and fourth components, with eigenvalues of **1.17** and **0.99**, explain **6.89%** and **5.82%** of the variance, respectively. By retaining the first **four principal components**, the cumulative explained variance reaches **79.29%**, which is close to the desired **80%** threshold. Therefore, We chose to retain four components to capture the majority of the dataset’s variability. This selection balances the need for interpretability and ensures that enough variance is captured while keeping the model manageable.





The loadings for Principal Component 1 (PC1) highlight that variables such as `Min total`, `PTS`, and `DR` have the highest positive contributions. This indicates that PC1 is primarily driven by overall player activity, including the total minutes played, points scored, and defensive rebounds, capturing a strong aspect of general player performance. Variables such as `GP`, `TR`, and `FD` also contribute notably to this component, reinforcing its focus on active participation in the game. For Principal Component 2 (PC2), the top loadings come from `BLK`, `OR`, and `AST`. These suggest that PC2 is more representative of specific game actions like blocking shots, offensive rebounds, and assists. Negative loadings for variables such as `AST` and `3P%` highlight inverse relationships in this dimension, where players with higher assist counts or shooting percentages may contribute less to other factors influencing PC2. Overall, PC2 captures a mix of offensive and defensive gameplay specificities.

Variable	Loadings PC1	Loadings PC2	Loadings PC3	Loadings PC4
Min_total	0.9578	-0.1349	0.0205	0.0664
PTS	0.9054	-0.1507	0.1487	-0.0804
DR	0.8675	0.2728	-0.1257	0.1257
GP	0.8438	-0.1085	0.2255	0.3039
FD	0.8408	0.1375	-0.2462	-0.1753
TR	0.8389	0.4403	-0.0884	0.1842
TO	0.8201	-0.3217	-0.1208	-0.2054
FC	0.8144	-0.0320	0.0137	0.1199
GS	0.7690	-0.0822	0.0948	0.1934
STL	0.7256	-0.4025	-0.0940	-0.0530
OR	0.6350	0.6784	-0.1257	0.1257
FT%	0.6046	-0.1111	0.6142	-0.7711
AST	0.6040	-0.5659	0.1503	-0.2111
BLKA	0.5919	-0.2222	0.2984	0.0219
2P%	0.4507	0.3454	-0.1696	-0.1697
BLK	0.3476	0.7958	-0.3183	-0.4095
3P%	0.0871	-0.4811	0.1487	0.4815

Table 1: Loadings for Principal Components 1, 2, 3, and 4

For Principal Component 3 (PC3), the top contributors include `FT%`, `GP`, and `PTS`, suggesting that PC3 emphasizes shooting efficiency and game involvement. Players with high free throw percentages and those who play more games tend to load highly on this component, pointing to a specialized focus on scoring efficiency and presence in games. Finally, Principal Component 4 (PC4) is driven by variables such as `GP`, `GS`, and `3P%`. This component appears to capture starting participation and three-point shooting performance, indicating that PC4 reflects players' consistency in starting games and their ability to score from long range.

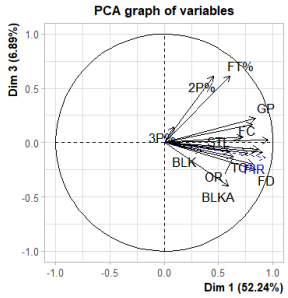


Figure 1: PC1-PC2

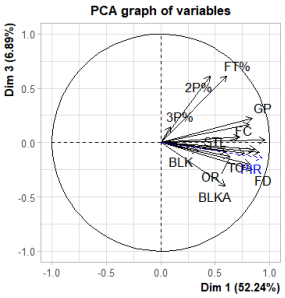


Figure 2: PC1-PC3

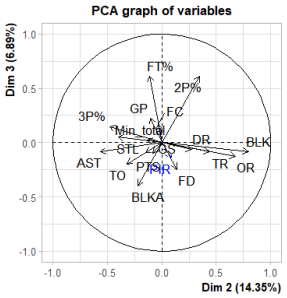


Figure 3: PC2-PC3

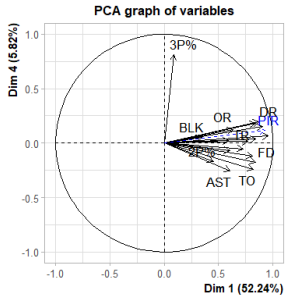


Figure 4: PC1-PC4

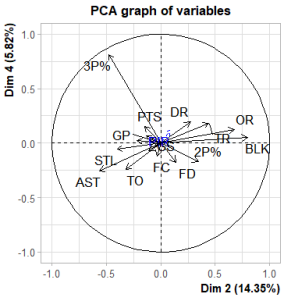


Figure 5: PC2-PC4

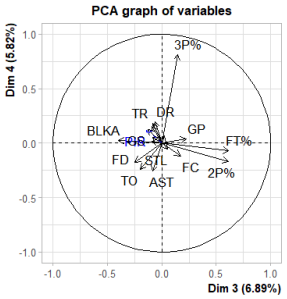


Figure 6: PC3-PC4

Figure 7: Comparison of principal components. Each subplot represents different combinations of the principal components

The PCA plots displayed above represent the relationships between variables based on the loadings of the first three principal components (PC1, PC2, and PC3). Statistically, these principal components explain a significant portion of the variance in the dataset. **PC1**, accounting for over 52% of the variance, is primarily driven by variables like **Min_total**, **PTS**, and **DR**, which represent overall player performance. In the **PC1-PC2 plot (Figure 1)**, geometrically, the strong vectors aligned with **PC1** indicate that these variables contribute heavily to this component. **PC2** accounts for around 14% of the variance, and variables such as **BLK** and **OR** contribute to specific game actions. In **PC1-PC3** (Figure 2), **FT%** and **PTS** show high loadings on **PC3**, indicating a focus on shooting efficiency. **PC3** adds around 6.9% variance, complementing the other components. In the **PC2-PC3** plot (Figure 3), the spread of variables highlights different offensive and defensive contributions, with **AST** and **BLK** showing opposing loadings. Geometrically, the closer a variable is to the edge of the circle, the stronger its contribution to the component.

The PCA individual plots, colored by team, demonstrate the projection of individual players from different teams (Fenerbahce, Olympiakos, Panathinaikos, and Real Madrid) onto the principal component space. Upon examining the plots, it becomes apparent that there is no clear clustering of players based on team, indicating that players from different teams exhibit a broad distribution of performance characteristics. This suggests that, at least based on the principal components being analyzed (PC1, PC2, PC3, and PC4), the teams show a similar level of performance overall.

The fact that players from each team are distributed relatively uniformly across the PCA space, rather than forming distinct groups, implies that the teams do not have highly distinct or unique playing styles, at least within the dimensions considered. The ellipses, if present, would also likely overlap significantly, further confirming that performance traits are fairly similar across these teams, making it difficult to distinguish one team's players from another based on these components alone.

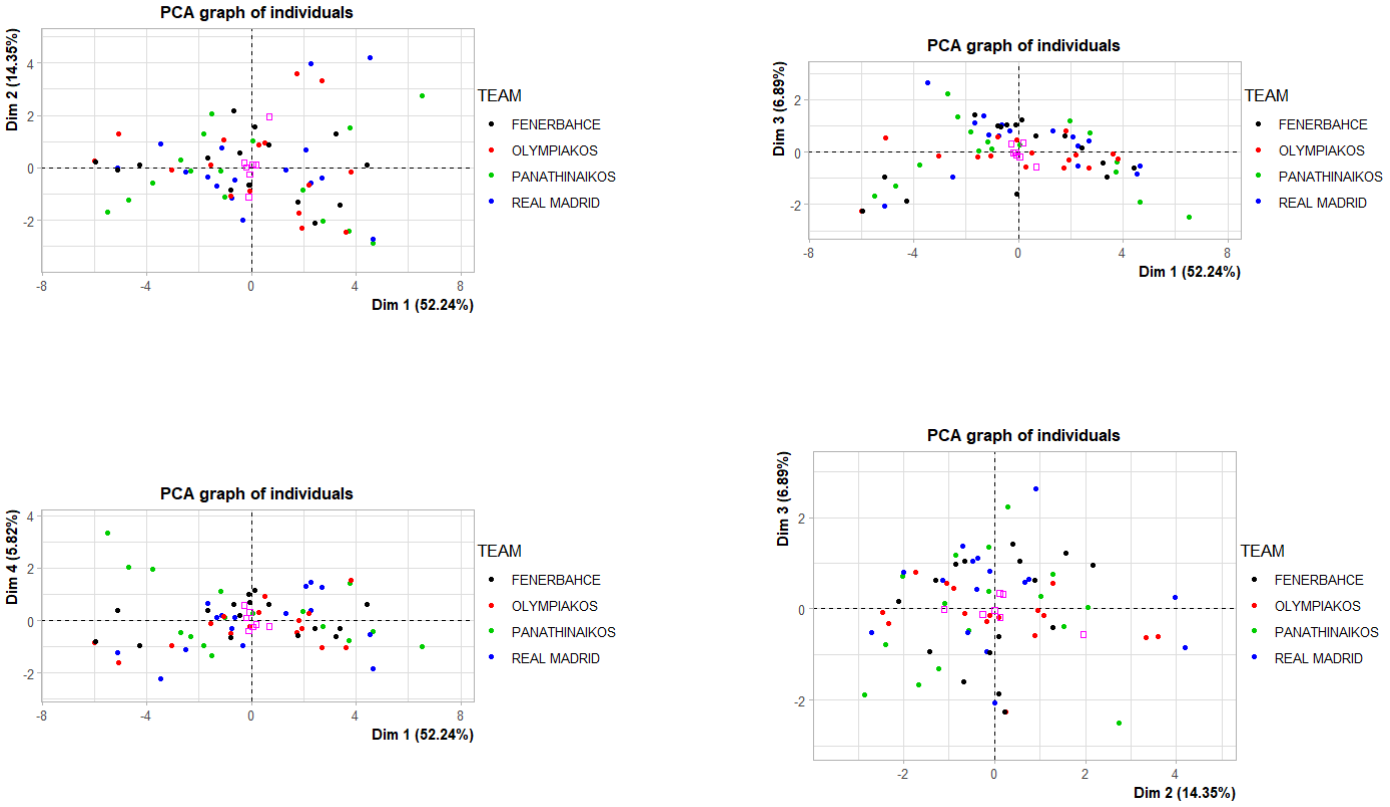


Figure 8: Comparison of PCA results. Each subplot shows different combinations of the principal components

In this analysis, we plotted the PCA results for individuals categorized by their **Position** (Center, Forward, Guard) using **PC1** and **PC2**, which explain 52.24% and 14.35% of the variance, respectively. These two dimensions provide the clearest representation of performance differences among positions. The plot shows that **Centers** dominate in overall performance, as they cluster in the upper right, with high values in general metrics like minutes played and points scored. Their moderate spread along **PC2** suggests variability in specific game actions. **Forwards** are more evenly distributed, indicating balanced performance, with moderate scores across both dimensions. **Guards** are positioned primarily in the lower left, indicating lower overall performance in **PC1**, but they likely excel in specialized roles such as playmaking or defense, as suggested by their negative values in **PC2**. Overall, the plot highlights performance distinctions between positions, with some overlap in roles.

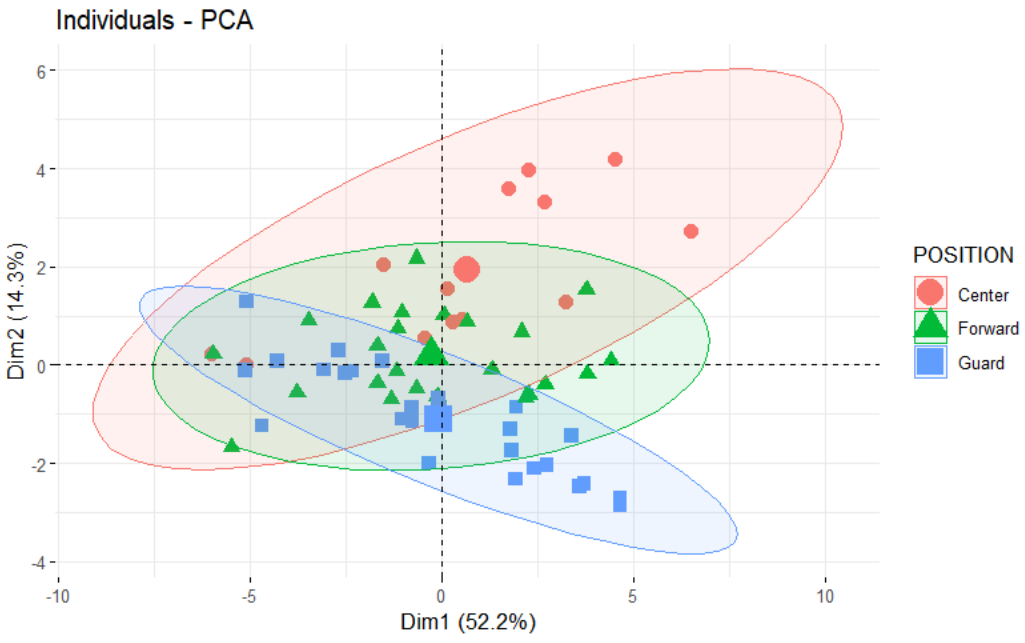
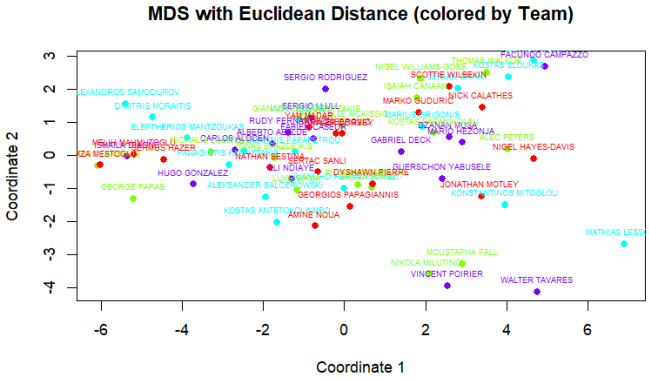
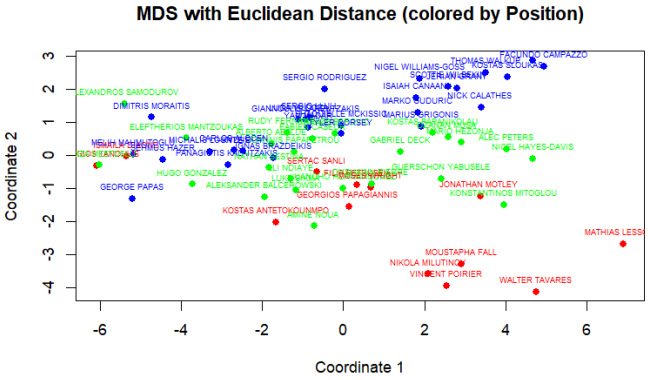


Figure 9: Clustering based on PCA

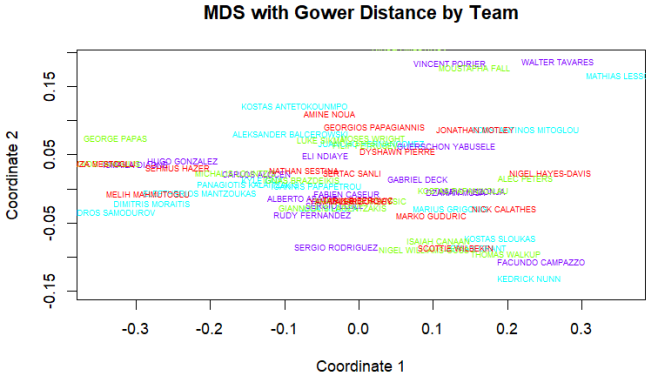
3. Multidimensional Scaling

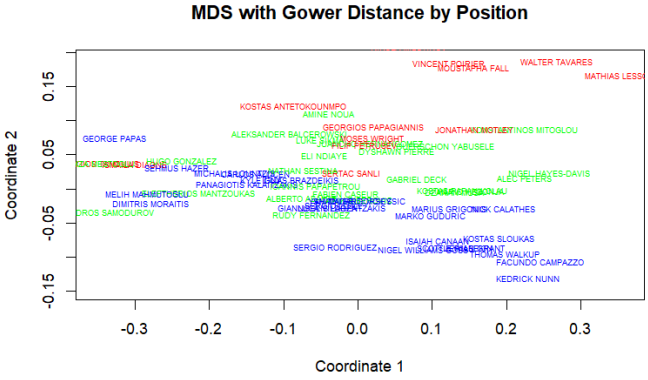
Multidimensional Scaling (MDS) is a technique used to visualize the level of similarity or dissimilarity between objects in a lower-dimensional space. It aims to preserve the pairwise distances between objects as much as possible in the new space. In this case, metric MDS is applied using **Euclidean distance**, a measure of straight-line distance between points in multidimensional space. By scaling the numerical variables and calculating the Euclidean distances, MDS allows us to visualize player performance similarity in two dimensions, helping to reveal any natural groupings or patterns in the data.

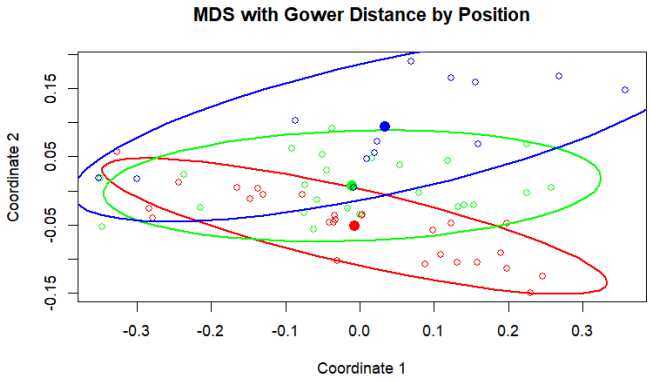


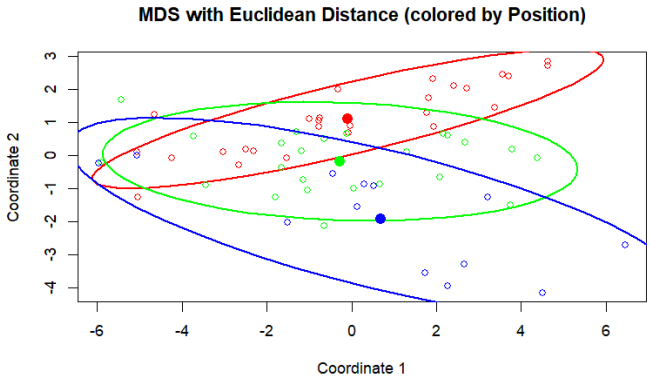


In the MDS plot colored by **Position**, a clear separation between player roles (Centers, Forwards, and Guards) is visible, indicating that position plays a significant role in differentiating player performance characteristics. However, the MDS plot colored by **Team** shows no distinct clustering, suggesting that players from different teams have similar performance traits. While position significantly influences performance, with clearer boundaries in the MDS plot, team affiliation does not lead to distinct groups. This reinforces the idea that player roles, rather than teams, are more relevant in explaining performance variability across dimensions, in this context where the final stage of the trophy is short, the Teams play the same number of matches and for example, the second win the same number of matches than the third. Interestingly, two of the tournament's top players, Facundo Campazzo and Thomas Walkup, both positioned as Guards, are closely located in the plot, reflecting their high performance and similar playstyle. This reinforces the idea that position plays a more crucial role than team affiliation in explaining performance variability.
To account for both numerical and categorical data, **Gower distance** was calculated, incorporating the "position" variable into the distance matrix. By applying "metric MDS" to the Gower distance matrix, the expected result is a more comprehensive representation of player similarity, accounting for mixed data types.

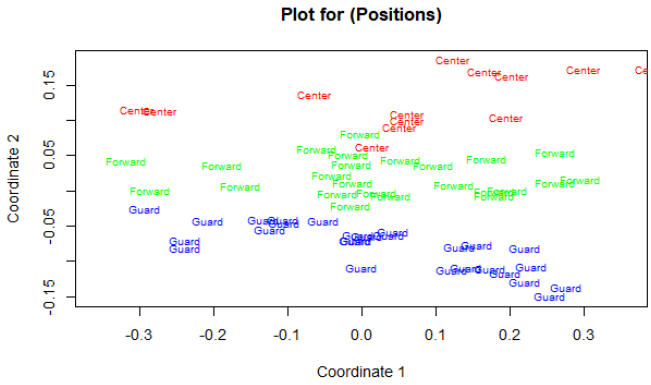


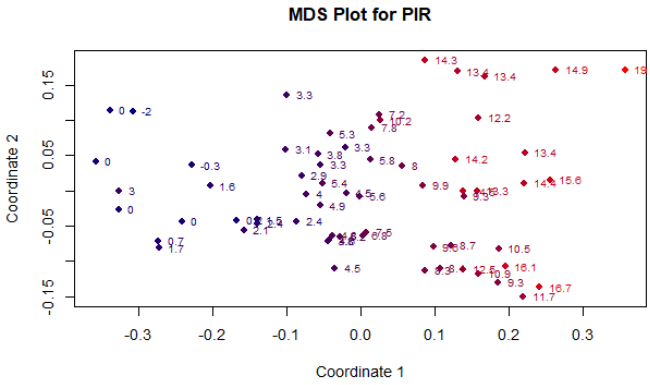




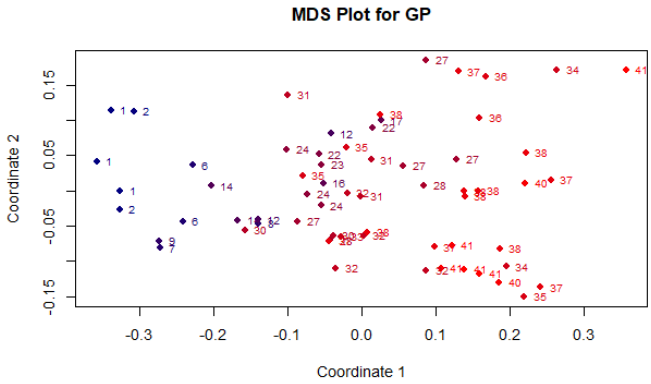


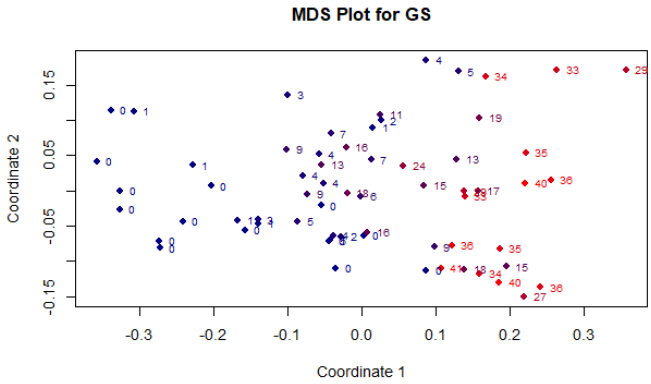
To conclude the analysis, we utilized the Gower distance matrix to create scatter plots that highlight various numerical and categorical variables. Notably, the plots featuring the variables "Positions" and "PIR" emerged as crucial indicators for the first two coordinates. Specifically, a higher "PIR" corresponds to an increase in coordinate 1, while the "Position" variable exhibits a similar pattern, with values distinctly separated along the second axis.

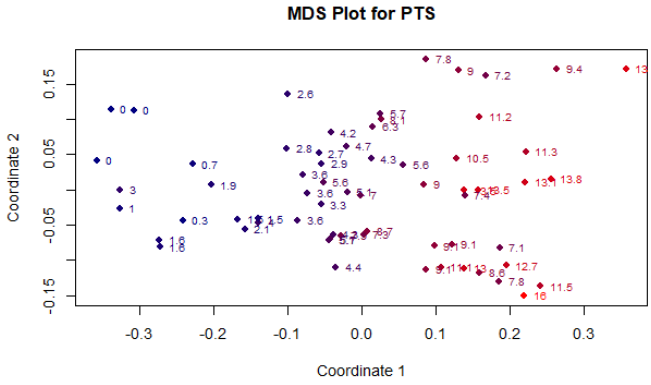


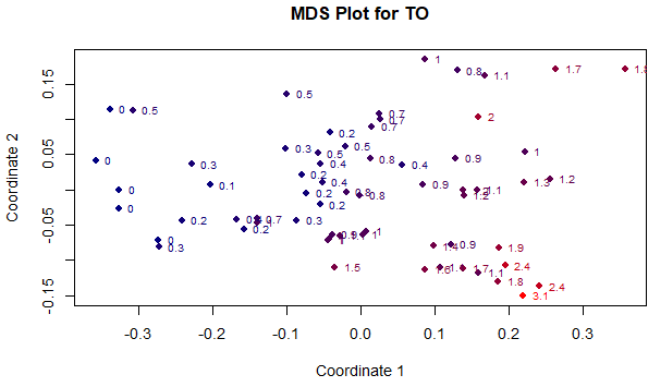


Further exploration of additional variables like "GP" games played, "GS" games started, "PTS" points scored and "TO" turnovers supports the assertion that performance is the primary factor measured by coordinate 1.

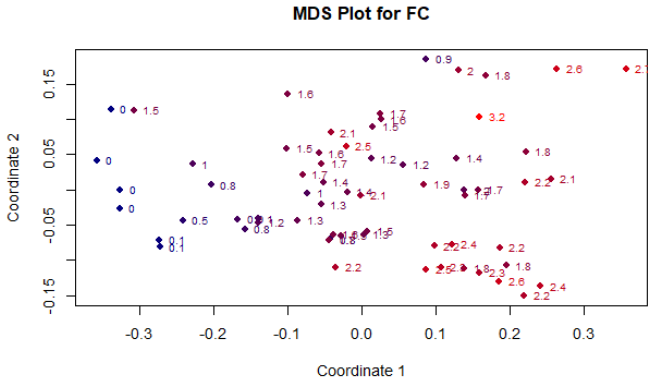




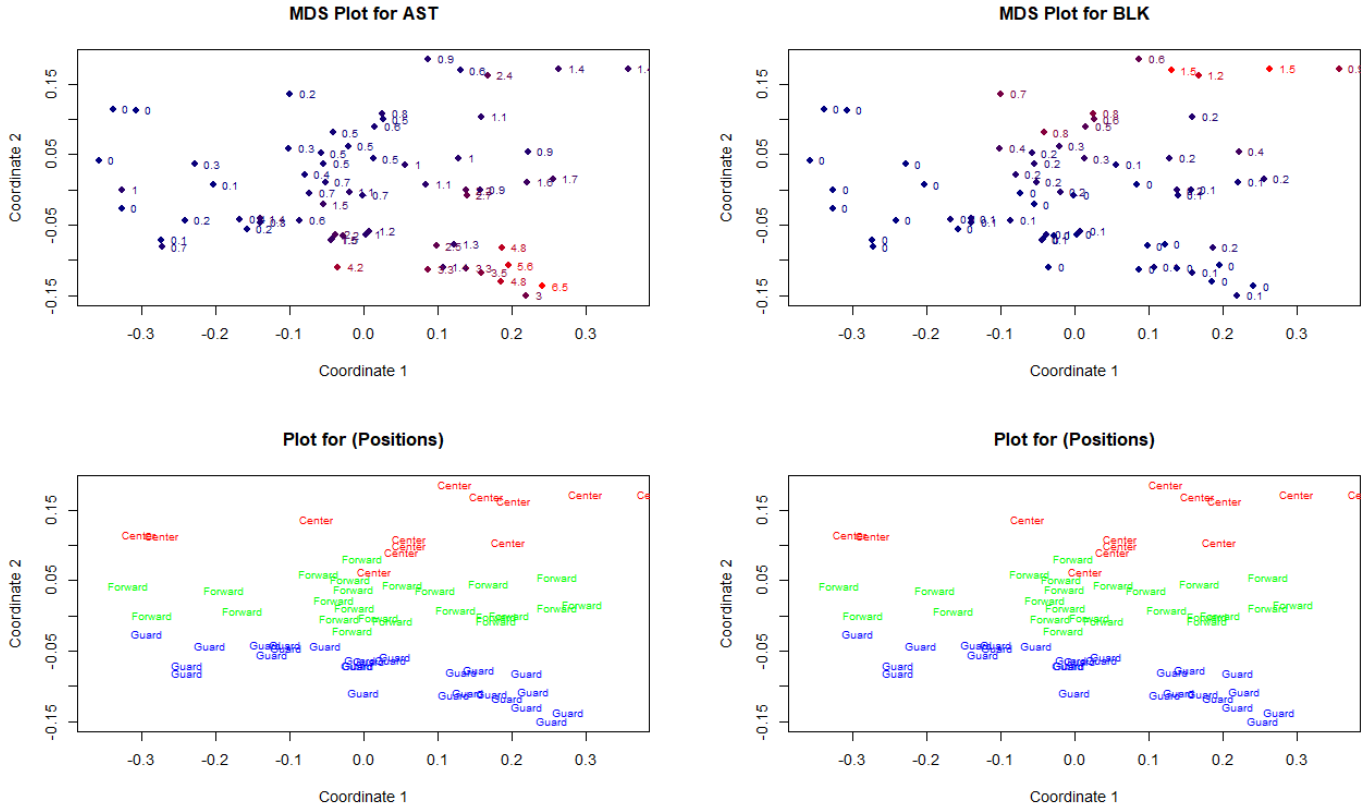




Interestingly, the variable "FC" (fouls committed) also shows an increasing trend with coordinate 1. While one might argue that better players tend to commit more fouls, we believe this correlation is more closely related to playtime rather than skill level. It suggests that players who are more actively involved in the game, whether due to higher performance or greater minutes on the court, are more likely to accumulate fouls.



In contrast, coordinate 2 seems to reflect the influence of position and play style. For example, the variable AST, which counts the number of assists, demonstrates higher values in the lower region of the plot, where guards are predominantly located. This indicates a strong correlation between assist rates and player positions, reinforcing the significance of playstyle in the analysis.
Similarly, players in the center position are more likely than those in other positions to block scoring attempts. This tendency is clearly illustrated in the plot for the variable "BLK" (blocks), where the frequency of blocked shots is notably higher among (good performing) centers.



In summary, our results indicate that using Gower distance is superior to Euclidean distance for powering Multidimensional Scaling (MDS) for this dataset. Gower distance effectively accounts for the variable of position, which is crucial in understanding how playstyle influences all other variables.