

Homework1

2024-05-03

Principal Component Analysis (PCA) and Multidimensional Scaling (MDS)

The homework should not exceed 6 pages. Preferably, you can use R Markdown.

Import the data set “euroleague_22_23.csv” the player statistics of four teams taken part in Final Four of Euro League 2022-2024.

The variables in the data set are defined as follows:

No - Row Number

TEAM -Team of the Player

PLAYER - Player name

POSITION - Position of the player

GP - Games played

GS - Games started

Min - Minutes played

PTS - Points scored

X2P. - Percentage of Two-points

X3P. - Percentage of Three-points

FT. - Percentage of Free-throws

OR - Offensive rebounds

DR - Defensive rebounds

TR - Total rebounds

AST - Assists

STL - Steals

TO - Turnovers

BLK - Blocks

BLKA - Blocks against

FC - Personal fouls committed

FD - Personal fouls drawn

PIR - Performance Index Rating

1. First do the exploratory data analysis.

- a) Discard the variable “No” from the data set. (1p)
- b) Split variable “Min” using `strsplit()` function. Give the name `aux` to the output. The first element of each row will show the minutes that the player played in total. (1p)
- c) Add a numerical variable to the data set named “Min 2” which shows how many minutes each player played in the game. (2p)
- d) Check the structure of the data and assign correct type to each variable considering whether it is a categorical or a numerical variable. (2p)

2. Application of PCA.

- a) Apply PCA on all the scaled numerical variables in the data set by using `PCA()` function in `FactoMineR` package. Treat the categorical variables and the variable “PIR” as supplementary variables using arguments `quali.sup` and `quant.sup` correctly. (3p)
- b) How many components should be extracted? Decide on the number of components considering eigenvalues. (3p)
- c) Use `plot.PCA()` function to show correlations between variables and the extracted dimensions by changing “axes” according to the dimension of interest. (For the variables you should use the argument `choix = “var”`). Interpret the loadings/correlations of variables at each dimension (6p).
- d) Interpret variable plots. How can each dimension be named? (5p)
- e) Show individuals plots changing argument `choix=“ind”` in `plot.PCA()` function and interpret the important findings. (5p)

3. Application of MDS.

- a) Apply metric MDS using Euclidean distance on scaled numerical variables. (2p)
- b) Plot the data using the points on the first two coordinates using players positions as label. (2p)
- c) Interpret the plot.(3p)
- d) Calculate gower distance including variable “POSITION” to the data matrix. (3p)
- e) Apply metric MDS on gower distance matrix. (2p)
- f) Represent individuals plot on the first two coordinates (2p).
- g) Use different categorical and numerical variables as labels so as to explain clusters that are constructed. (5p)
- h) Which MDS do you think better group the individuals? Why? (3p)

Recommendations for further analysis:

A sensitivity analysis can be done to check the robustness of the conclusions. This is an optional analysis and will not be evaluated as a part of homework. In case you would like to try, following steps can be applied:

- the players who played less than 5 games or 5 minutes can be removed.
- For “0” values for the variables that represent percentages can be imputed by using a proper method.
- After that PCA and MDS analyses can be performed again.