SoccerIndex

title: “FIFA 22 Clustering – Identifying Player Archetypes”

description: “How can we use clustering to uncover different types of professional soccer players?” categories: - Clustering - K-means - Sports Analytics

## Module Overview

This SCORE module explores how we can use **unsupervised learning** — specifically **K-means clustering** — to group soccer players based on their physical and technical attributes. Using FIFA 22 player statistics, students identify player archetypes such as defenders, playmakers, and goal scorers based purely on data. The project culminates with visual interpretations using PCA, radar charts, and nationality breakdowns of cluster types.

## Learning Goals

By the end of this module, students will be able to:

* Use dplyr and ggplot2 to explore and summarize player data
* Calculate and interpret descriptive statistics and pairwise correlations
* Standardize variables and apply K-means clustering in R
* Use PCA and radar plots to visualize and interpret cluster groupings
* Discuss the implications of cluster characteristics in a real-world context (e.g., scouting, team strategy)

## Sports Content Question

**What types of players exist in FIFA 22, and how can we group them based on their core skills using data alone?**

## Dataset

This module uses player-level data from **FIFA 22**, publicly available via [Kaggle](https://www.kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset). The dataset includes attributes for each player such as pace, shooting, dribbling, defending, and more.

We filtered to only include players with complete data for the following attributes:

* pace, shooting, passing, dribbling, defending, physic

These were used as the input features for clustering.

## Variable Descriptions

| Variable | Description |
| --- | --- |
| short\_name | Player’s commonly known name |
| nationality\_name | Country the player represents |
| club\_name | Team the player is contracted to |
| overall | Overall rating (0–100) |
| potential | Maximum potential rating (0–100) |
| age | Player age |
| pace | Speed and acceleration |
| shooting | Finishing and shot power |
| passing | Accuracy and vision |
| dribbling | Ball control and skill moves |
| defending | Tackling and marking |
| physic | Strength, stamina, and aggression |

## Materials

* players\_22.csv — original full FIFA 22 dataset from Kaggle
* fifa\_clean.csv — cleaned version with relevant columns and no missing data
* worksheet.qmd — guided student worksheet with embedded R code chunks and questions
* worksheet\_key.qmd — instructor version with answers and explanation
* Radar plots and PCA cluster visuals are generated with fmsb, cluster, and factoextra.

## Notes

This module is suitable for undergraduate courses covering **clustering**, **exploratory data analysis**, or **data visualization**. It can be completed as a 1–2 day class activity or a short project. Students should have prior exposure to data wrangling and basic plotting in R.