Basketball performance analysis

# **Dataset information**

## **Source:** <https://www.kaggle.com/datasets/ziya07/basketball-player-performance>

The dataset is designed to simulate the performance and training data of basketball players in order to develop personalized training plans using AI. It contains various physical performance metrics and calculates the overall training effectiveness of each player. The goal is to predict how effective a training regimen is based on individual player characteristics, such as speed, endurance, jump height, and strength.

## **Description of Variables (Why This Data?)**

The basketball training dataset contains various player performance metrics that are crucial for understanding player characteristics and patterns. These variables were chosen because:

* Statistical Performance Metrics: The dataset includes fundamental basketball statistics like points, assists, rebounds, and shooting percentages. These metrics provide quantifiable measures of player performance across different aspects of the game
* Physical Attributes: Variables like height, weight, and age offer essential context about players' physical capabilities and development stage. These attributes often correlate with playing styles and roles
* Efficiency Metrics: Advanced statistics such as player efficiency rating helps evaluate players' overall effectiveness beyond basic statistics
* The dataset comprises metrics collected from basketball players during training sessions. Each variable serves a specific purpose in understanding player performance:
  + heart\_rate:
    - Type: Integer
    - Range: 120-180 beats per minute (bpm)
    - Purpose: Heart rate is a critical indicator of the intensity of physical activity and recovery, helping to understand the cardiovascular demands of training
  + speed:
    - Type: Float
    - Range: 8-15 meters per second (m/s)
    - Purpose: This measures the player's agility and quickness, essential for various in-game movements, such as breaking away from defenders or fast breaks
  + jump\_height:
    - Type: Float
    - Range: 20-40 inches
    - Purpose: Jump height is a key indicator of a player's explosiveness, particularly important in basketball for actions like rebounding and shooting
  + endurance:
    - Type: Float
    - Range: 10-40 minutes
    - Purpose: Measures the player's stamina and ability to maintain performance over time, particularly important for sustained effort during games
  + strength:
    - Type: Float
    - Range: 50-150 kilograms (kg)
    - Purpose: Strength influences a player's ability to compete physically, particularly in post play, defense, and maintaining body control in high-contact situations
  + player\_efficiency:
    - Type: Float
    - Range: 10-30 (scaled index)
    - Purpose: A measure that combines various on-court statistics to provide an overall idea of a player's performance efficiency
  + training\_effectiveness:
    - Type: Categorical (integer)
    - Values:

**0**: Low Effectiveness – Training plan is not effective in improving performance

**1**: Moderate Effectiveness – Training plan has a moderate impact on the player's improvement

**2**: High Effectiveness – Training plan significantly improves performance

* + - Purpose: This target variable is used to assess how well the AI-generated personalized training plans improve player performance, based on the collected data

## **Description of Observations (What Are We Looking At?)**

The dataset presents a comprehensive view of basketball players with the following characteristics:

* Sample Size and Scope: The dataset includes multiple players from various teams, providing a broad representation of basketball talent
* Data Quality: After cleaning (removing missing values), we maintained data integrity while ensuring sufficient sample size for meaningful analysis
* Variable Distribution: Initial data exploration revealed natural variations in player statistics, suggesting distinct player types and roles
* Time Period: The data represents a specific season or period, allowing for consistent comparison across players
* Total observations: 500 players
* Key statistics:
  + heart\_rate: Mean = 149 bpm, range = 120-180 bpm
  + speed: Mean = 11.44 m/s, range = 8 -15 m/s
  + jump\_height: Mean = 30.09 cm, range = 20- 40 cm
  + endurance: Mean = 25.51, range = 10 - 40 minutes
  + strength: Mean = 100.53, range = 50 -150 kilograms
  + player\_efficiency: Mean = 20.21, range = 10 - 30(scaled index)
  + training\_effectiveness: Binary, 70.4% effective sessions

## 

## **Data Analysis Approach**

Our analysis followed a structured methodology:

* Data Preprocessing:
  + Removal of missing values to ensure data quality
  + Selection of numerical variables for clustering
  + Feature scaling using StandardScaler to normalize all variables
* Dimensionality Analysis:
  + Use of Principal Component Analysis (PCA) for visualization
  + Retention of key features that explain the majority of variance
* Cluster Analysis:
  + Implementation of agglomerative hierarchical clustering analysis.

## 

## **Motivation for Chosen Analysis Methods**

The analysis was done using two main methods, namely Principal Component Analysis (PCA) and Hierarchical Clustering Analysis (HCA)

### **Principal component analysis(PCA)**

Principal Component Analysis (PCA) is a statistical technique used to simplify a dataset by reducing its dimensions while preserving as much variability (information) as possible.

* **Dimensionality Reduction**: The dataset includes several interrelated variables. PCA simplifies the analysis by reducing the number of dimensions while retaining the maximum variance in the data
* **Identifies Key Factors**: PCA might show that certain training metrics are more critical for predicting training effectiveness, helping you prioritize these variables
* **Reveals Patterns**: By plotting players in the space of the first two principal components, you can identify clusters of players with similar performance or training effectiveness
* **Improves Interpretability**: PCA provides insights into how variables like "jump height," "endurance," and "effectiveness" contribute to overall performance differences among players

### **Hierarchical Clustering Analysis**

Agglomerative(‘bottom-up’) hierarchical clustering was selected as the primary analysis method for several reasons:

* **Pattern Discovery**: Agglomerative hierarchical clustering is excellent for uncovering natural groupings in the data without requiring a predefined number of clusters, making it ideal for analyzing the performance metrics of basketball players and identifying distinct player profiles.
* **Interpretability**: The hierarchical structure and dendrogram visualization provide an intuitive way to explore the relationships and similarities between players at different levels of granularity, allowing for a deeper understanding of the data.
* **Flexibility**: Unlike algorithms like k-means, agglomerative clustering can capture clusters of varying shapes and sizes, which is crucial for datasets like ours, where player performance metrics may not form regular patterns.
* **Scalability with Metrics**: The algorithm works well with a variety of distance metrics, ensuring it adapts to the nature of the basketball statistics and captures the nuances in the dataset.
* **Outlier Detection**: Agglomerative clustering highlights outliers naturally as single clusters in the early stages, helping to identify unique players with exceptional or unusual performance traits.

# **PCA Analysis**

# **Presentation of Results**

### 1. **Correlogram of Factor Loadings**

**Component 1:**

* Shows a very strong positive correlation (0.94) between strength and training\_effectiveness
* This suggests that strength training is a crucial factor in overall training effectiveness

**Component 2:**

* Shows positive correlations with jump\_height (0.67) and player\_efficiency (0.69)
* This component seems to capture athletic explosiveness and its relationship to player performance

**Component 3:**

* Has a strong positive correlation with speed (0.74)
* Shows some positive correlation with endurance (0.62)
* This component represents the speed-endurance relationship in player performance

**Component 4:**

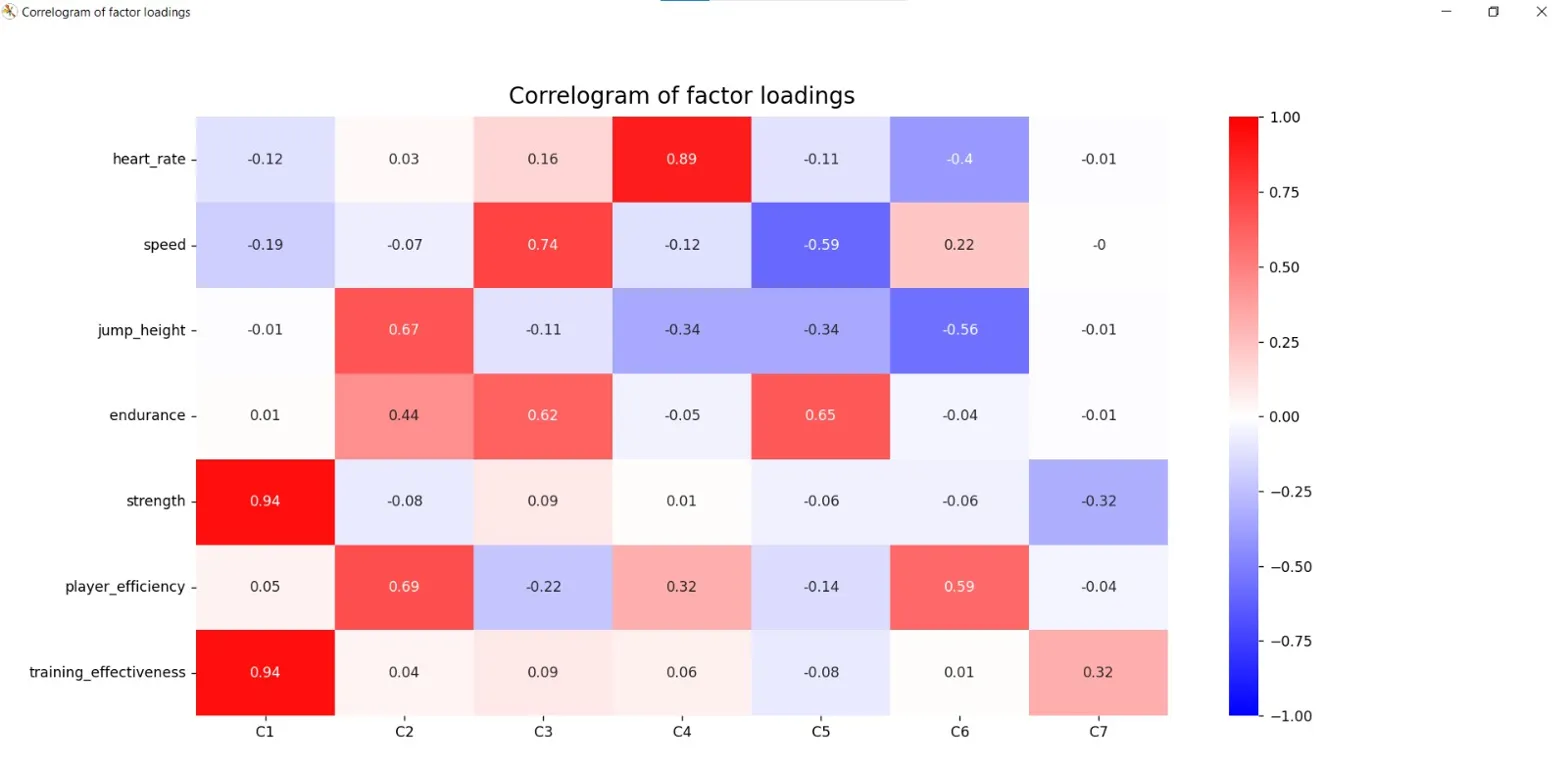
* Shows strong correlation with heart\_rate (0.89)
* This component likely represents cardiovascular fitness

**Component 5:**

* Shows moderate positive correlation with endurance (0.65)
* Shows negative correlation with speed (-0.59)
* This might represent the trade-off between endurance and explosive speed

**Component 6:**

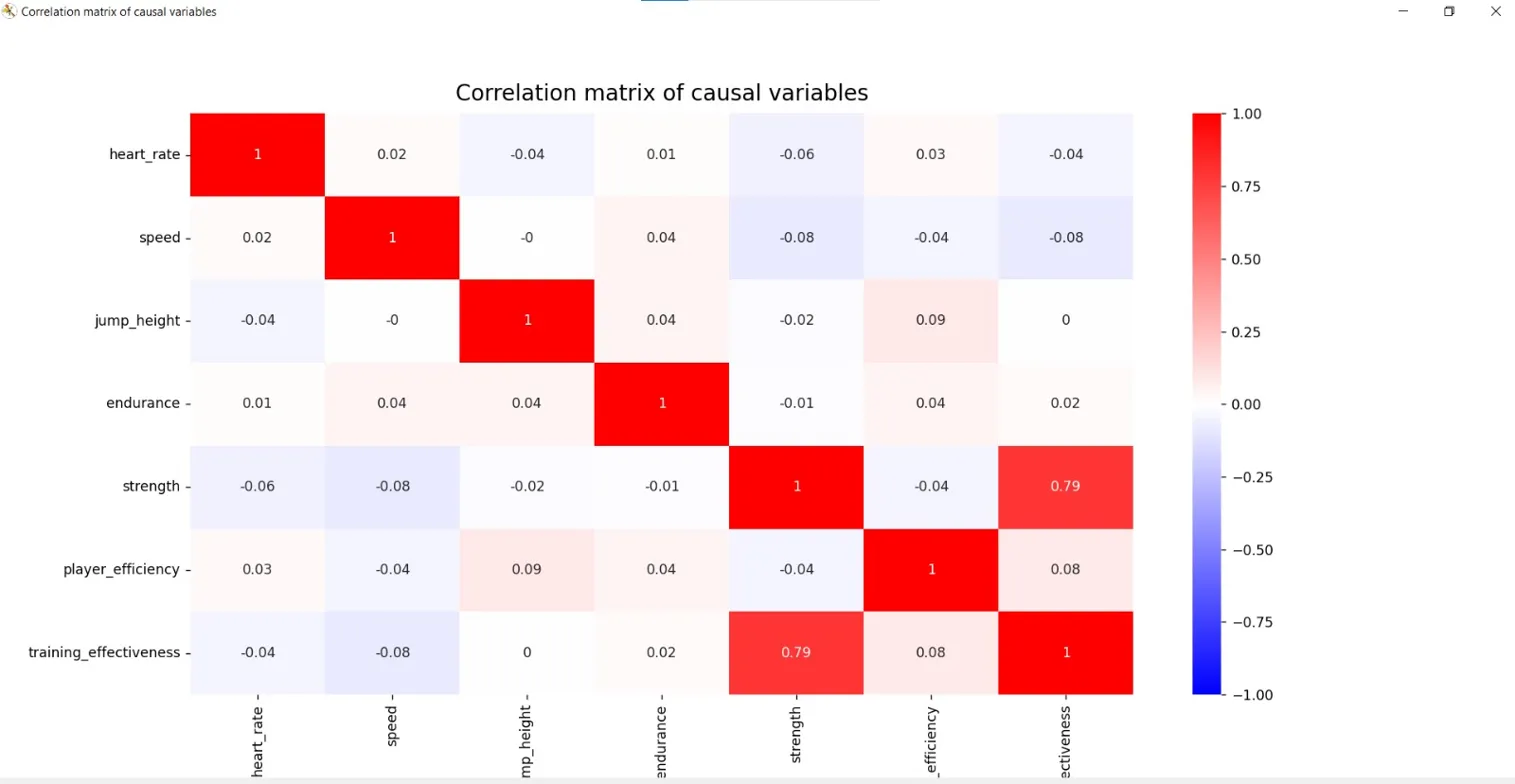
* Shows positive correlations with player\_efficiency (0.59)
* This component seems to capture pure player efficiency independent of physical metrics



2. **Correlation Matrix of Causal Variables**

This is vital because it shows direct relationships between variables:

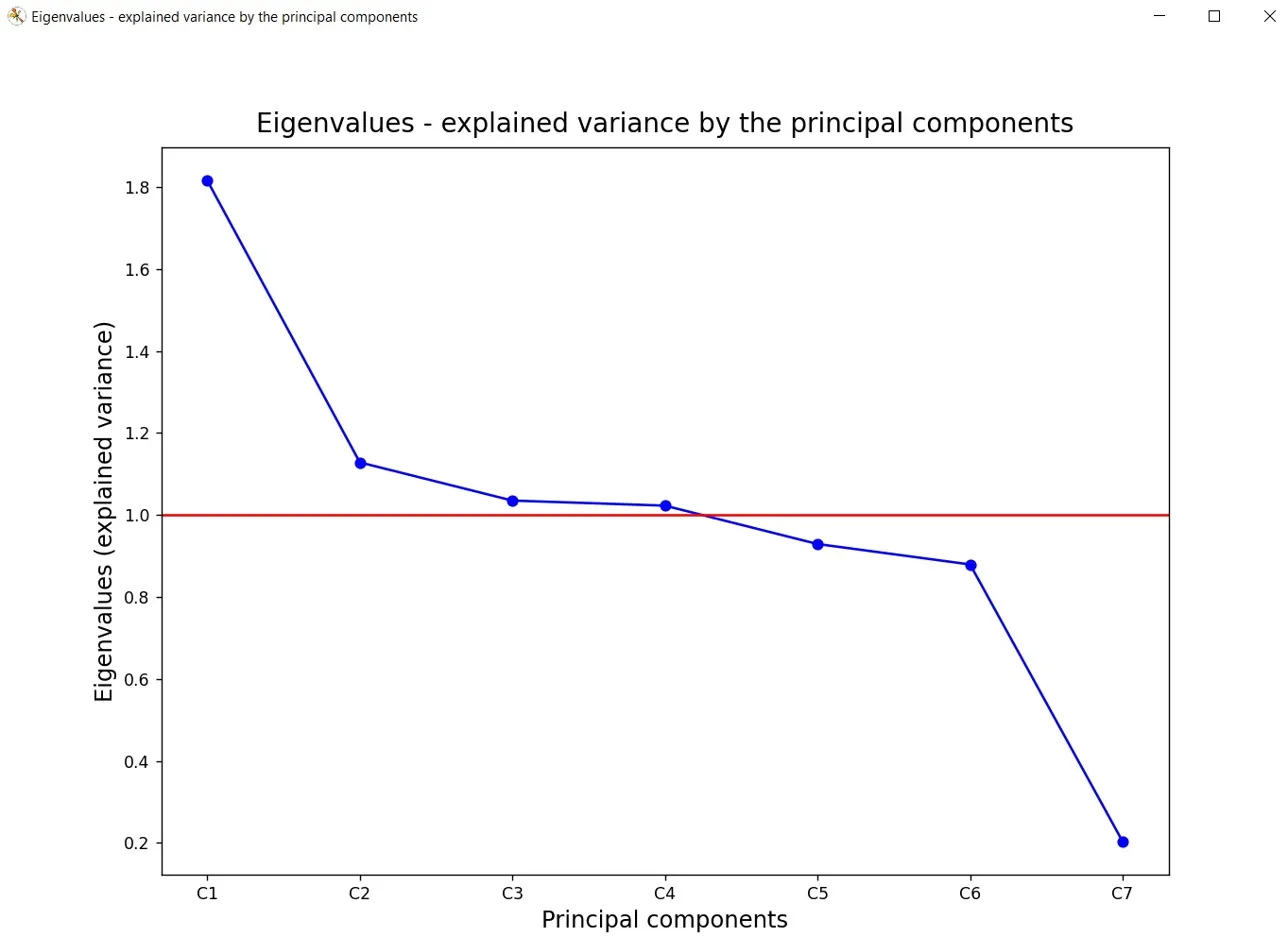
* Strong positive correlation (0.79) between strength and training\_effectiveness
* Most physical attributes (speed, heart\_rate, jump\_height, endurance) have low correlations with each other. This suggests that these athletic attributes are relatively independent of each other
* The diagonal is always 1.0 (red) as variables perfectly correlate with themselves
* Most other variables show weak correlations (light colors near zero)



3. **Eigenvalues - Explained Variance**

This line plot shows the amount of variance explained by each principal component:

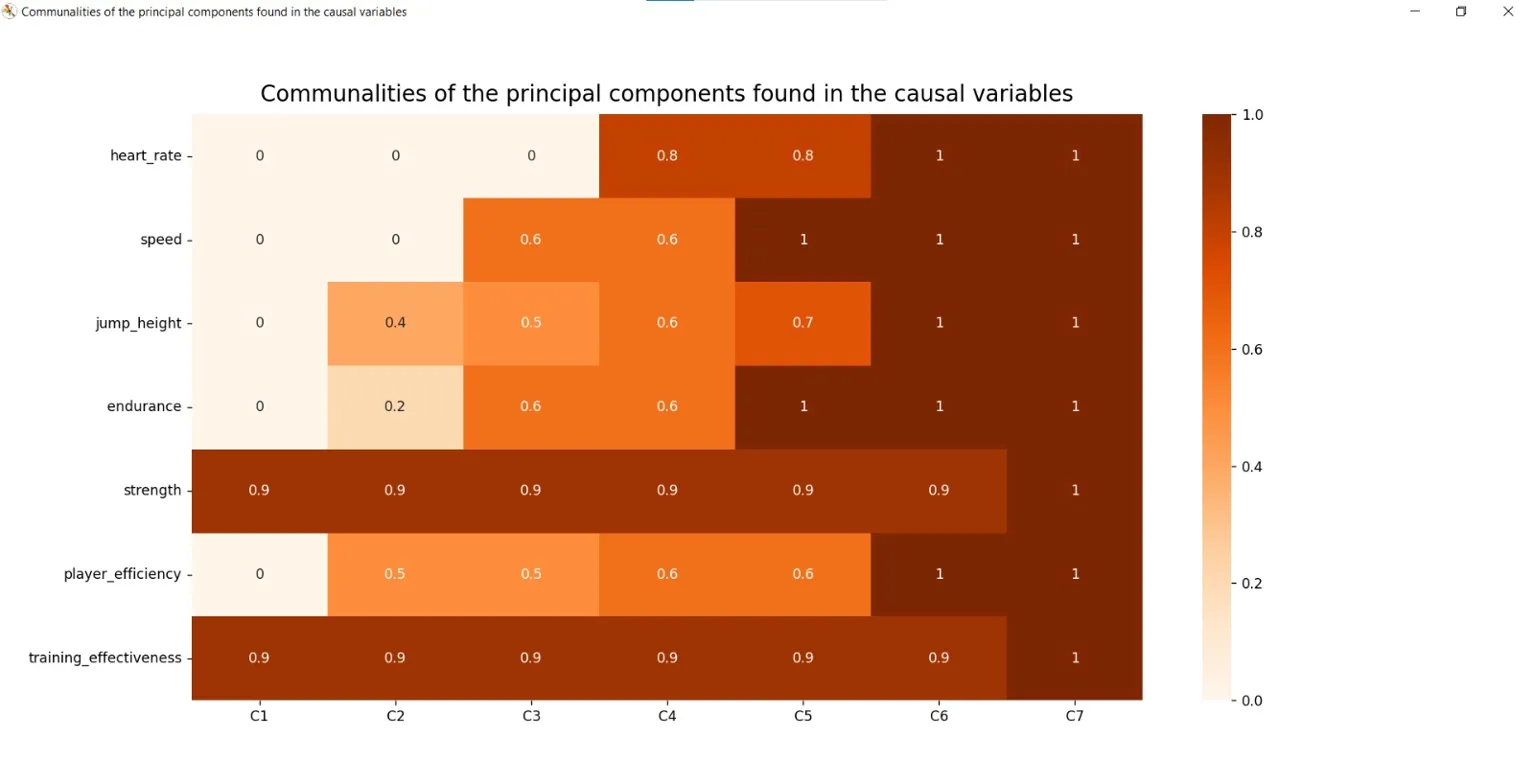
* C1 has eigenvalue ~1.8 (highest variance explained)
* Red horizontal line at 1.0 represents the Kaiser criterion
* Components 1 - 4 are above the Kaiser criterion
* Sharp drop after C1 indicates it's the most important component
* C7 explains very little variance (eigenvalue ~0.2)



4. **Communalities of Principal Components**

This heatmap is valuable because it shows:

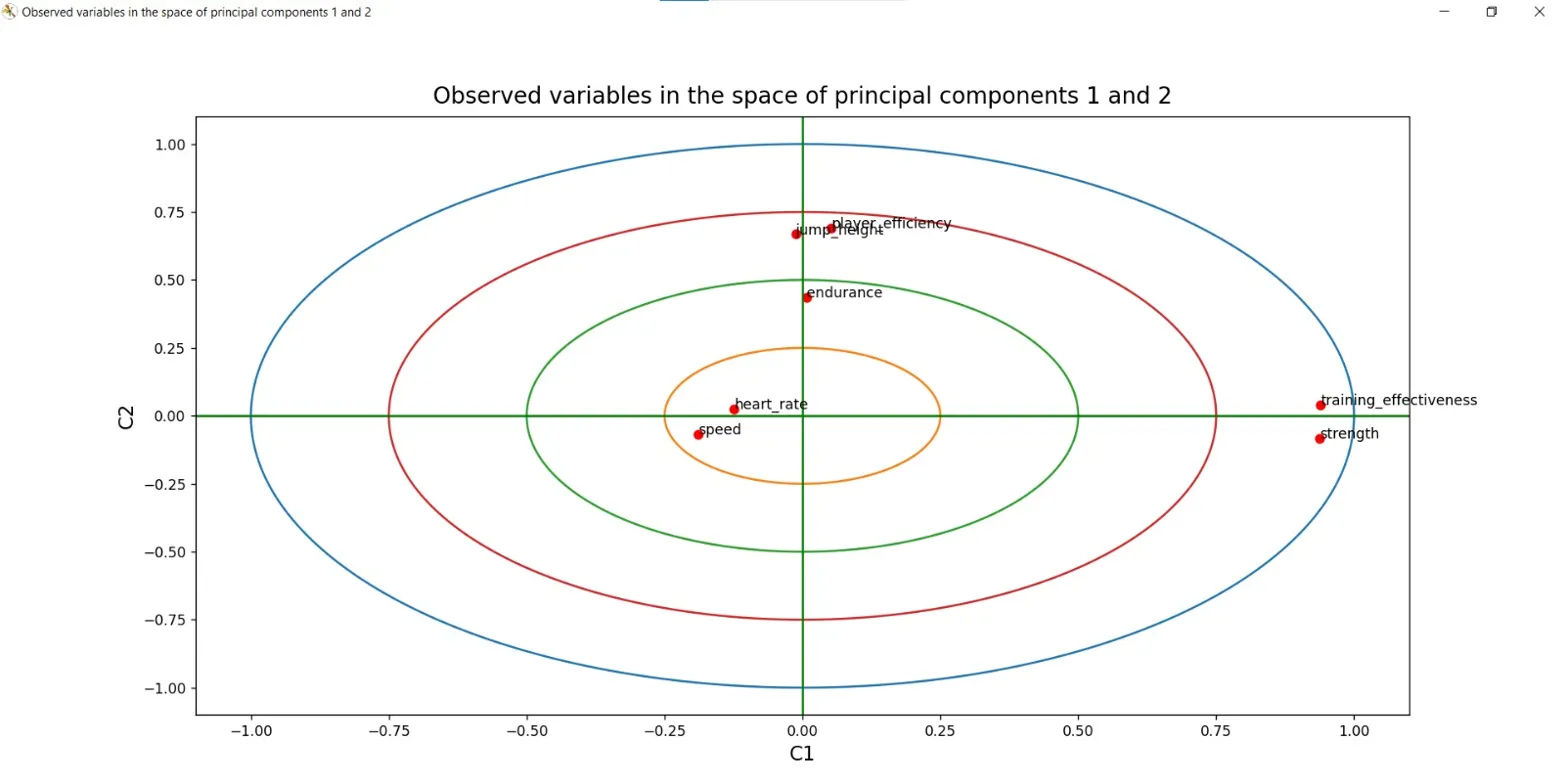
* Heart\_rate and speed have high communalities in later components
* Strength and training\_effectiveness show consistently high values (0.9) across early components. Training\_effectiveness shows pattern similar to strength
* Jump\_height and player\_efficiency show moderate communalities
* Darker colors indicate higher communalities
* Values range from 0 to 1, with 1 indicating perfect representation
* The progression from light to dark shows how additional components improve variable representation



5. **Observed Variables in PCA Space**

This biplot shows how variables relate in the space of the first two principal components:

* Strength and training\_effectiveness cluster together on the right (high C1)
* Jump\_height and player\_efficiency cluster together in the upper portion (high C2)
* Heart\_rate and speed are closer to the center
* The concentric circles represent different levels of correlation
* Variables far from the center have stronger relationships with PC1 and PC2
* The angle between variables indicates their correlation (smaller angles = stronger correlation)



**Additional insights from combining these visualizations:**

* Your training program's effectiveness is most strongly tied to strength development
* Cardiovascular factors (heart\_rate, endurance) form a distinct aspect of performance
* Athletic metrics (jump\_height, speed) form another distinct component
* The data has a clear hierarchical structure with strength/training effectiveness at the top

This analysis suggests several practical applications:

1. Focus strength training programs as they most strongly correlate with effectiveness

2. Monitor heart rate and speed as independent performance indicators

3. Consider player efficiency and jump height as related metrics

# **Clustering Analysis**

## **Presentation of Results**

The analysis produced several key outputs:

* **Cluster Visualization**:The dendrogram for **Observations Clustering** shows two major clusters: a large group containing over 450 players and a smaller group of outliers.
* **Cluster Characteristics**:

**Cluster 0 (Majority)**: Represents the majority of players with similar physical attributes, but not as exceptional in specific performance metrics (endurance, speed, strength).

**Cluster 1 (Minority)**: A smaller group of players who are significantly different from the majority, possibly representing exceptional players or outliers in specific attributes (speed, jump height, strength, etc.).

## **Interpretation**

The clustering analysis reveals several significant insights:

**Player Archetypes**:

* **Cluster 0**: Likely represents average-performing players who are neither exceptional nor underperforming in any specific area, but share common physical traits.
* **Cluster 1**: This smaller group potentially contains exceptional players with unique combinations of attributes (for example, a mix of high speed, jump height, or strength), or they could be underperforming players who stand out in some specific metrics.

**Limitations and Considerations**:

* **Cluster Representation**: While the analysis helps in identifying broad player archetypes, it may not account for the unique traits of every individual player. Some players may not fit neatly into a single cluster.
* **Over-simplification**: The results might simplify player performance, and thus, should be complemented by traditional scouting methods for a more nuanced understanding of player capabilities.
* **Outliers**: A small group of outliers (Cluster 1) may distort the overall analysis. It’s important to further explore whether this cluster truly represents exceptional players or simply players with unique combinations of attributes.

The **variable clustering** analysis produced several key outputs:

* **Cluster Visualization**: The dendrogram for **Variable Clustering** reveals the relationships between the different performance attributes (e.g., speed, endurance, strength). Variables that are closely related are grouped together, forming distinct clusters.
* **Cluster Characteristics and Interpretation**: there are 5 clusters, Cluster 0 being formed the “earliest”, having together the strength and training\_effectiveness, which shows that these 2 have a strong correlation. The other variables are not as compatible, which is shown by the “late” bonding.

## **Conclusion**

The analysis provides a deeper understanding of how different performance metrics relate to each other. This knowledge can help refine training programs, recruitment strategies, and overall team development:

* **Targeted Training**: By understanding which attributes are strongly correlated (e.g., speed and strength), training can focus on improving multiple areas simultaneously.
* **Personalized Development**: Players who exhibit strengths in one cluster but weaknesses in another can receive personalized development plans to enhance their overall performance.

This analysis is valuable in ensuring that a player’s development is aligned with the team’s needs and that recruitment decisions are based on a comprehensive understanding of both physical and performance-related metrics. Future studies could explore more complex relationships between these attributes and how they evolve over time.