A MULTIVARIATE ANALYSIS OF BASKETBALL PLAYER ATTRIBUTES

Anum Damani
STATS 411
Winter 2025

Motivation & Research Question

- Motivation:
  - In their careers, athletes are evaluated and recruited based on their physical and performance attributes.
    - It is well known that the height, weight, and skills of basketball players have a crucial role in their success on the court.
- Research Questions:
  - How effectively can Principal Component Analysis (PCA) reduce the dimensionality of the data while preserving as much variance as possible, and what underlying relationships are revealed by the principal components?
  - How effectively can Canonical Correlation Analysis (CCA) reduce the dimensionality of the data while maximizing correlation, and what underlying relationships can be determined through the canonical variates?

DESCRIPTION OF THE DATASET

About The Data

- Dataset obtained from Kaggle & contains basketball player statistics across several NBA seasons, from 1996-2023.
    - Each observation represents an individual basketball player.
- After data cleaning, data contains 12,844 observations and 10 numeric variables.
- Dataset contains no missing values.

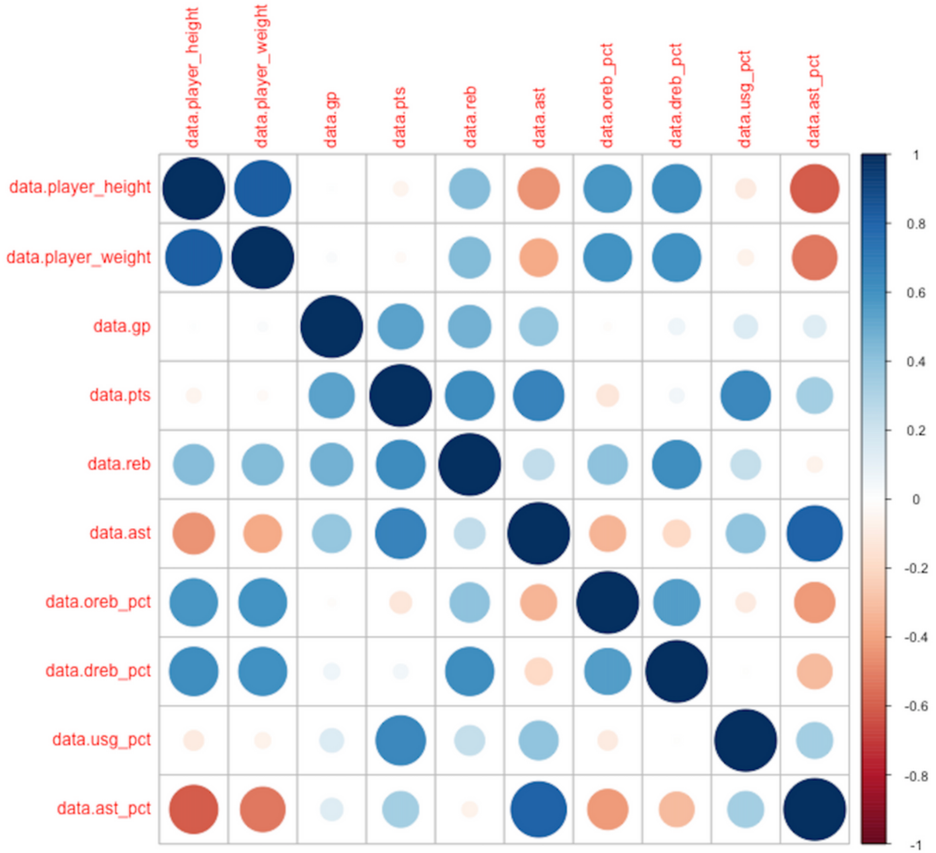

Figure 1: Correlation Plot of Variables

About The Data

- Variable descriptions:
  - player_height: Height of player (in centimeters)
  - player_weight: Weight of player (in kilograms)
  - gp: Number of games played in the season
  - pts: Average number of points scored
  - reb: Average number of rebounds
  - ast: Average number of assists
  - oreb_pct: Percentage of offensive rebounds
  - dreb_pct: Percentage of defensive rebounds
  - usg_pct: Percentage of team plays utilized by player
  - ast_pct: Percentage of goals assisted

PRINCIPAL COMPONENT ANALYSIS

Analysis

- Principal Component Analysis (PCA) is a dimensionality-reduction method that aims to preserve as much variation in the dataset as possible.
  - It transforms the original, correlated variables into a smaller set of uncorrelated variables (principal components), which capture the variance-covariance structure of the original data.
  - This allows for more interpretability of relationships between variables.
- It is appropriate for this dataset because there are strong-moderate correlations among the 10 variables.
- These 10 variables can be condensed into a smaller set (components) using this method.
- Therefore, PCA is helpful for determining underlying relationships within the data.

Results

- Scree plot (Figure 2) shows clear elbow formed by PC1 and PC2, suggesting that the first 2 PCs explain the most variation in the data.
    - PC1 explains 39.15% of the variance, and PC2 explains 29.44% of the variance (68.58% together).
    - Based on this, proceed with first two PCs.
- Plot of First 2 PCs (Figure 3) showcases elliptical shape.

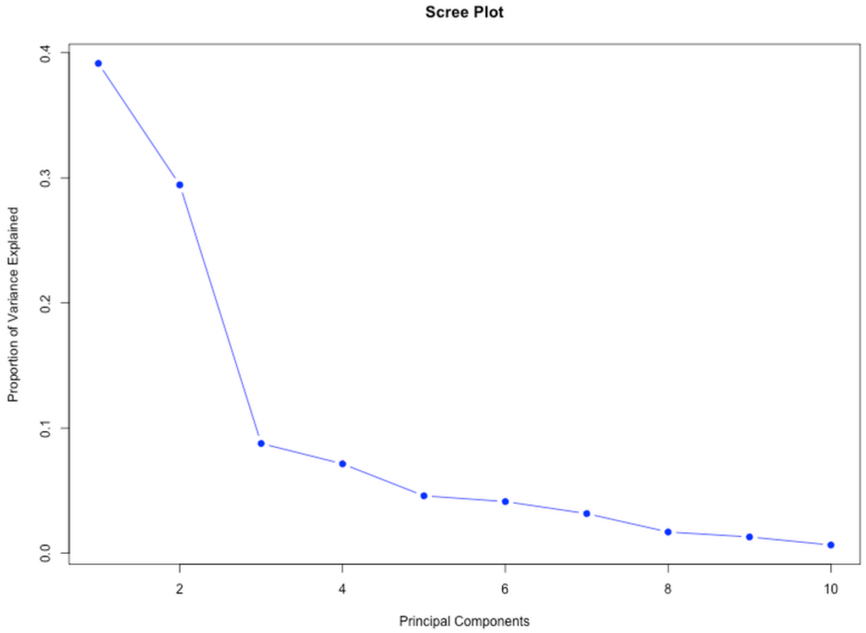

Figure 2: Scree Plot

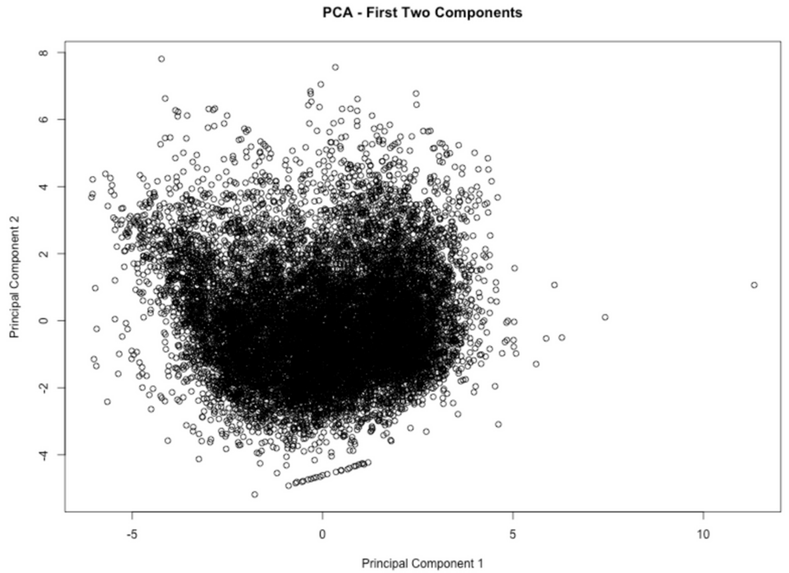

Figure 3: Plot of First 2 PCs

Results

- Figure 4 shows the PC Loadings for PC1 and PC2, and Figure 5 shows the correlations between PCs and the original variables.

Variable	PC1 Loadings	PC2 Loadings
player_height	**-0.4398**	-0.1235
player_weight	**-0.4216**	-0.1518
gp	0.0604	-0.3724
pts	0.1536	**-0.5089**
reb	-0.1782	**-0.4923**
ast	0.3464	-0.3418
oreb_pct	-0.3781	-0.1104
dreb_pct	-0.3521	-0.2402
usg_pct	0.1552	-0.3335
ast_pct	0.3986	-0.1543

Figure 4: PC Loadings for PC 1 & 2

Variable	PC1 Corr	PC2 Corr
player_height	**-0.8701**	-0.2119
player_weight	**-0.8343**	-0.2605
gp	0.1196	-0.6389
pts	0.3040	**-0.8730**
reb	0.3526	**-0.8447**
ast	0.6853	-0.5864
oreb_pct	-0.7481	-0.1894
dreb_pct	-0.6966	-0.4121
usg_pct	0.3071	-0.5722
ast_pct	0.7886	-0.2649

Figure 5: Correlations between PCs & Original Variables

Interpretation (PC1)

- PC1:
  - Player height and weight contribute most to PC1. They have the highest absolute loadings (around -0.4) and strongest correlations (around -0.8).
    - More negative PC1 values correspond to taller and heavier players. Conversely, more positive PC1 values correspond to shorter and less heavy players.
  - Other variables:
    - ast_pct has a moderate absolute loading (0.3986) and a strong, positive correlation with PC1 (0.7886).
      - Higher ast_pct values contribute to higher, positive PC1 values.
    - oreb_pct and dreb_pct have moderate PC1 loadings (-0.3781 and -0.3521, respectively) with strong, negative correlations with PC1 (-0.7481 and -0.6966, respectively).
      - Higher rebound percentages correspond to lower PC1 values.
    - ast has a moderate PC1 loading (0.3464) with a moderate, positive correlation (0.6853).
      - Higher number of average assists correspond with higher PC1 values.

Interpretation (PC2)

- PC2:
  - Points and rebounds contribute most to PC2.
    - pts and reb have the highest absolute loadings (-0.5089 & -0.4923, respectively) with strong, negative correlations (-0.8730 & -0.8447, respectively).
      - More negative PC2 values correspond to basketball players that are successful at scoring points and rebounding.
      - More positive PC2 values correspond to basketball players that are less successful at scoring points and rebounding.
  - It can be noted that the variable gp has a moderate loading (-0.3724) with moderate, negative correlation with PC2 (-0.6389).
    - Higher amounts of games played corresponds with lower PC2 values.

CANONICAL CORRELATION ANALYSIS

Analysis

- Canonical Correlation Analysis (CCA) is a dimensionality reduction method that is helpful for analyzing the relationships between two sets of variables.
  - Dataset includes 2 types of variables: physical attributes and performance attributes.
- CCA provides summaries for the relationship between physical attributes and performance attributes, while preserving the key information.
  - It finds canonical variate pairs (linear combinations of original variables) that maximize correlation between two variable sets.

Results

- 2 sets of variables:
  - player_height & player_weight in first set, while the remaining 8 are in the second set.
- p=2 and q=8 so there are min(p,q) = 2 sample canonical correlations and sample canonical coefficient vectors.

$$Z^{(1)} = \begin{bmatrix} Z_1^{(1)} \\ Z_2^{(1)} \end{bmatrix} = \begin{bmatrix} \text{player\_height} \\ \text{player\_weight} \end{bmatrix}, Z^{(2)} = \begin{bmatrix} Z_1^{(2)} \\ Z_2^{(2)} \\ Z_3^{(2)} \\ Z_4^{(2)} \\ Z_5^{(2)} \\ Z_6^{(2)} \\ Z_7^{(2)} \\ Z_8^{(2)} \end{bmatrix} = \begin{bmatrix} \text{gp} \\ \text{pts} \\ \text{reb} \\ \text{ast} \\ \text{oreb\_pct} \\ \text{dreb\_pct} \\ \text{usg\_pct} \\ \text{ast\_pct} \end{bmatrix}$$

Results

- Canonical correlations: ρ1* = 0.8001144, ρ2* = 0.1565184
  - First pair of canonical variates conveys a substantial amount of information about the relationship between the two sets, while the second pair does not do this as much.
- Below is the first pair of canonical variates (U1, V1):

$$\hat{U}_1 = -0.6686568z_1^{(1)} - 0.3749867z_2^{(1)},$$
$$\hat{V}_1 = 0.03252165z_1^{(2)} - 0.03547987z_2^{(2)} - 0.21348676z_3^{(2)} - 0.09957098z_4^{(2)} - 0.27273937z_5^{(2)} - 0.35273886z_6^{(2)} - 0.03752319z_7^{(2)} + 0.45272971z_8^{(2)}$$

  - Coefficients of U1 reveals that height contributes more to U1, with player height being slightly more dominant than player weight.
  - Coefficients of V1 reveals it is clear that ast_pct, dreb_pct, oreb_pct, and reb contribute most to V1.

Results

- Figure 6 shows correlations between Z^(1) and U1.
  - Physical attributes, player_height and player_weight, are strongly correlated with U1, with correlations -0.98 and -0.93, respectively.
    - These strong correlations between the physical attributes and U1 align with the interpretation of the coefficients stated previously.
    - So, these variables strongly impact U1, the canonical variate representing physical attributes.

$Z^{(1)}$ Variables	$\hat{U}_1$	$\hat{V}_1$
$z_1^{(1)}$ $(player\_height)$	-0.98	-0.78
$z_2^{(1)}$ $(player\_weight)$	-0.93	-0.74

Figure 6: Correlations between Z^(1) and (U1, V1)

Results

- Figure 7 shows correlations between Z^(2) and V1.
  - dreb_pct, oreb_pct, and reb are moderately to strongly, negatively correlated with V1, with correlations -0.80, -0.77, and -0.60, respectively.
  - ast_pct is strongly, positively correlated with V1, with correlation 0.75.
  - This aligns well with the interpretation of the coefficients previously.
  - So, these variables strongly impact V1, the canonical variate representing performance attributes.

$Z^{(2)}$ Variables	$\hat{U}_1$	$\hat{V}_1$
$z_1^{(2)}$ (gp)	-0.01	-0.01
$z_2^{(2)}$ (pts)	0.05	0.06
$z_3^{(2)}$ (reb)	-0.45	-0.60
$z_4^{(2)}$ (ast)	0.44	0.54
$z_5^{(2)}$ $(oreb\_pct)$	-0.62	-0.77
$z_6^{(2)}$ $(dreb\_pct)$	-0.64	-0.80
$z_7^{(2)}$ (usg_pct)	0.09	0.12
$z_8^{(2)}$ (ast_pct)	0.60	0.75

Figure 7: Correlations between Z^(2) and (U1, V1)

Results

- Matrices of errors of approximation for first canonical pair effectively summarizes (reproduces) the intra-set correlations in R12 due to small error values (closer to zero).
  - However, there are larger values for the errors for R11 and R22, which means that the first canonical pair do not effectively summarize (reproduce) the sampling variability in the original Z^(1) and Z^(2) variable sets.
- The proportion of total standardized sample variance in Z^(1) explained by U1 is 90.48%. The proportion of total standardized sample variance in Z^(2) explained by V1 is 27.85%.

$$R_{11} - \text{sampleCov}(\tilde{z}^{(1)}) = \begin{bmatrix} 0.046 & -0.081 \\ -0.081 & 0.145 \end{bmatrix}$$

$$R_{22} - \text{sampleCov}(\tilde{z}^{(2)}) = \begin{bmatrix} 0.00 & -0.01 & 0.01 & 0.00 & 0.03 & 0.01 & 0.01 & 0.03 \\ -0.01 & 0.03 & -0.02 & -0.01 & -0.10 & -0.02 & -0.03 & -0.08 \\ 0.01 & -0.02 & 0.02 & 0.01 & 0.09 & 0.02 & 0.02 & 0.08 \\ 0.00 & -0.01 & 0.01 & 0.00 & 0.03 & 0.01 & 0.01 & 0.03 \\ 0.03 & -0.10 & 0.09 & 0.03 & 0.36 & 0.07 & 0.09 & 0.31 \\ 0.01 & -0.02 & 0.02 & 0.01 & 0.07 & 0.01 & 0.02 & 0.06 \\ 0.01 & -0.03 & 0.02 & 0.01 & 0.09 & 0.02 & 0.02 & 0.08 \\ 0.03 & -0.08 & 0.08 & 0.03 & 0.31 & 0.06 & 0.08 & 0.26 \end{bmatrix}$$

$$R_{12} - \text{sampleCov}(\tilde{z}^{(1)}, \tilde{z}^{(2)}) = \begin{bmatrix} 0.00 & -0.01 & 0.01 & 0.00 & 0.02 & 0.00 & 0.01 & 0.02 \\ -0.00 & 0.01 & -0.01 & -0.00 & -0.04 & -0.01 & -0.01 & -0.03 \end{bmatrix}$$

Results

- 2 hypothesis tests were conducted.
  - Result of first test: Since the Bartlett's test statistic (13440.38) is greater than the critical value (26.29623), we reject H0 at the α = 0.05 significance level.
    - This suggests that at least one canonical correlation is not equal to zero. We can proceed to the second hypothesis test.
  - Result of second test: The test statistic (318.4094) is greater than the critical value (14.06714), which means we reject H0 at the α = 0.05 significance level.
    - This indicates that the second canonical correlation is statistically significant.
- Therefore, the rejection of both null hypotheses suggests that both ρ1∗ and ρ2∗ are significant (nonzero). Both ρ1∗ and ρ2∗ capture the relationship between the physical attributes and performance attributes.
  - The relationship between physical attributes and performance attributes is statistically significant.

	Test 1	**Test 2**
H_0	$\Sigma_{12} = 0, (\rho_1^* = \rho_2^* = 0)$	$\rho_1^* \neq 0, \rho_2^* = 0$
H_1	$\Sigma_{12} \neq 0$	$\rho_2^* \neq 0$

Figure 8: Hypothesis Tests

CONCLUSION

Conclusion

- Similarities:
  - Both PCA and CCA have a common goal of dimensionality reduction:
    - PCA reduces data from 10 dimensions to 2 dimensions (principal components), explaining 68.58% of the variance.
    - CCA reduces data from 10 dimensions to 2 dimensions (2 canonical variate pairs), with one pair having a stronger canonical correlation than the other.
  - Both PCA and CCA have a common goal of enhancing interpretability:
    - PCA transforms the original, correlated variables into a smaller set of uncorrelated variables (principal components), which capture the variance-covariance structure of the original data.
    - CCA finds canonical variate pairs (linear combinations of original variables) that maximize correlation between two variable sets.
  - Both PCA and CCA transform original variables into linear combinations of uncorrelated variables, which are called PCs in PCA and canonical variates in CCA.

Conclusion

- Differences:
  - PCA:
    - Goal is to preserve variance.
    - PC1 explains 39.15% of the variance, and PC2 explains 29.44% of the variance (68.58% together).
    - The following variables contribute most to PC1 and PC2, respectively:
      - PC1: **player_height, player_weight**, ast_pct, oreb_pct, dreb_pct, ast
      - PC2: **pts, reb**, gp
  - CCA:
    - Goal is to maximize correlation.
    - Canonical correlations: ρ1* = 0.8001144, ρ2* = 0.1565184
    - The strongest correlations between the original variables and canonical variates:
      - Z^(1) and U1: **player_height, player_weight**
      - Z^(2) and V1: **dreb_pct, oreb_pct, ast_pct**, reb
    - The proportion of total standardized sample variance in Z^(1) explained by U1 is 90.48%, while the proportion in Z^(2) explained by V1 is 27.85%.
    - Relationship between physical and performance attributes is statistically significant.

Reference

- Dataset link: https://www.kaggle.com/datasets/justinas/nba-players-data

Thank You!