STRATEGIC DIVERGENCE IN GLOBAL BASEBALL:

A Comparative Analysis of KBO and MLB Game Strategies Through Principal Component Analysis

Eum, Junho

Boston University

Table of Contents

Abstract 1

1. Introduction 2

2. Data and Methods 4

2.1. Figure Panel 1: Top Correlated Features Visual Representation / Variable Notations and Descriptions 5

2.2. Team Characteristic Identification using Principal Component Analysis 5

2.2.1. Data Standardization 6

2.2.2. Computing the Covariance Matrix 7

2.2.3. Eigen Decomposition 7

2.2.4. Finding Principal Components 8

2.2.5. Interpreting Principal Components 9

2.3. Figure Panel 2: Scree plot for Explained Variance for each Principal Component 9

3. Result 10

3.1. Establishing Baseline Rules for PC Component Interpretation 10

3.1.1. Figure Panel 3: Heatmap Representation of PC loadings 10

3.1.2. Detailed Explanation of Key Performance Variables 11

3.1.3. Figure Panel 4: Comparative Biplots Highlighting Principal Strategies in KBO League 15

4. Discussion 19

4.1. Exploring Team Archetypes through PCA: A Biplot Analysis 19

4.1.1. Biplots as Interpretative Tools 20

4.1.2. Interpretation and Title of PC components 21

4.1.3 Figure Panel 4: Histogram Distributions of Principal Component Loadings Across KBO and MLB Leagues 26

4.2. Comparative Analysis on KBO and MLB datasets 27

5. Conclusion 29

6. Appendix 30

Table 6.1. Threshold-Categorized PCA Loadings for KBO dataset 30

Table 6.2. Threshold-Categorized PCA Loadings for MLB dataset 31

6.3. References and Citations 32

Abstract

The strategic differences between the Korean Baseball Organization (KBO) and Major League Baseball (MLB) are the focus of this in-depth analysis. The research uses data from 1982-2021 for the KBO and 2002-2023 for MLB to highlight the unique tactics of each league. Principal component analysis (PCA) was the main analytical method, providing valuable insights into their contrasting approaches.

Specifically, KBO teams under Component 1 emphasize balanced offensive production. In contrast, MLB teams focus on high-impact power hits at key moments, even with fewer players on base. However, both leagues demonstrate strong defensive skills. Further, KBO teams

under Component 2 face challenges in scoring and fixing late-game defenses. But, MLB teams lean towards cautious offense, struggling with increased pitching problems, especially when limiting batters and defending in tense game situations.

My methodological approach was comprehensive, starting with the visual representation of top correlated features. The core of the PCA involved steps like data standardization, computing the covariance matrix, eigen decomposition, and identifying and interpreting principal components. This was supported by additional analyses like establishing baseline rules for interpreting the principal components, which included detailed explanations of performance variables, heatmap representations of PC loadings, and comparative biplots. Biplots, in particular, were key in gaining insights, serving as important interpretive tools, and revealed strategic archetypes across the leagues.

In summary, this study expands the baseball analytics framework by comparing the KBO and MLB, revealing a complex variety of strategies and performances within these two top baseball leagues.

1. Introduction

Baseball, a sport that spans continents, brings with it a multitude of strategies and styles that reflect the cultures and nuances of the leagues around the world. The Korean Baseball Organization (KBO) and Major League Baseball (MLB) are two representations of this global fascination. Both leagues have witnessed copious individual studies focused on player performance predictions and game strategies. For instance, research by Bae, Lee, and Lee leveraged Principal Component Analysis to predict the team's final ranking in KBO league.[1] On a different spectrum, Benavidez utilized cutting-edge machine learning models, namely LSTM/RNN, to predict MLB player performances.[2] Yet, there is a gap in literature apt for a thorough comparative exploration between the KBO and MLB, delving into their distinct tactical attributes via PCA.

This research explores the strategies of both leagues, using data from 1982-2021 for the KBO and 2002-2023 for MLB. The findings highlight distinct strategies between the KBO and MLB. The unique strategic differences between the two leagues have been revealed through an in-depth analysis. For instance, while KBO teams predominantly emphasize a well-rounded offensive strategy, MLB teams often rely on high-impact hits, especially during pivotal game moments, even if the bases aren't fully loaded.

From a methodological perspective, the study adopts a meticulous approach. It starts with visualizations of top-correlated features and then delves into the complexities of PCA. It starts with visualizations of top-correlated features and then moves into PCA's complexities. The entire

[1] Jae Young Bae, Jae Myung Lee, and Jung Yoon Lee, "Predicting Korea Pro-baseball rankings by principal component regression analysis," Communications for Statistical Applications and Methods 19, no. 3 (2012): 372-374.

[2] Jose Benavidez, Daniel Cervone, and Daniel M. Russell, "Forecasting MLB Player Performance Using Neural Networks," in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (New York: Association for Computing Machinery, 2017), 4-6.

analytical process, right from data standardization to eigen decomposition, is navigated. Supplementary evaluations are represented with heatmap visualizations and biplots, with the latter offering vivid glimpses into each league's strategic underpinnings. In summation, this research shows the different strategies that characterize these leagues and, in doing so, provides a clear view of the varied world of international baseball.

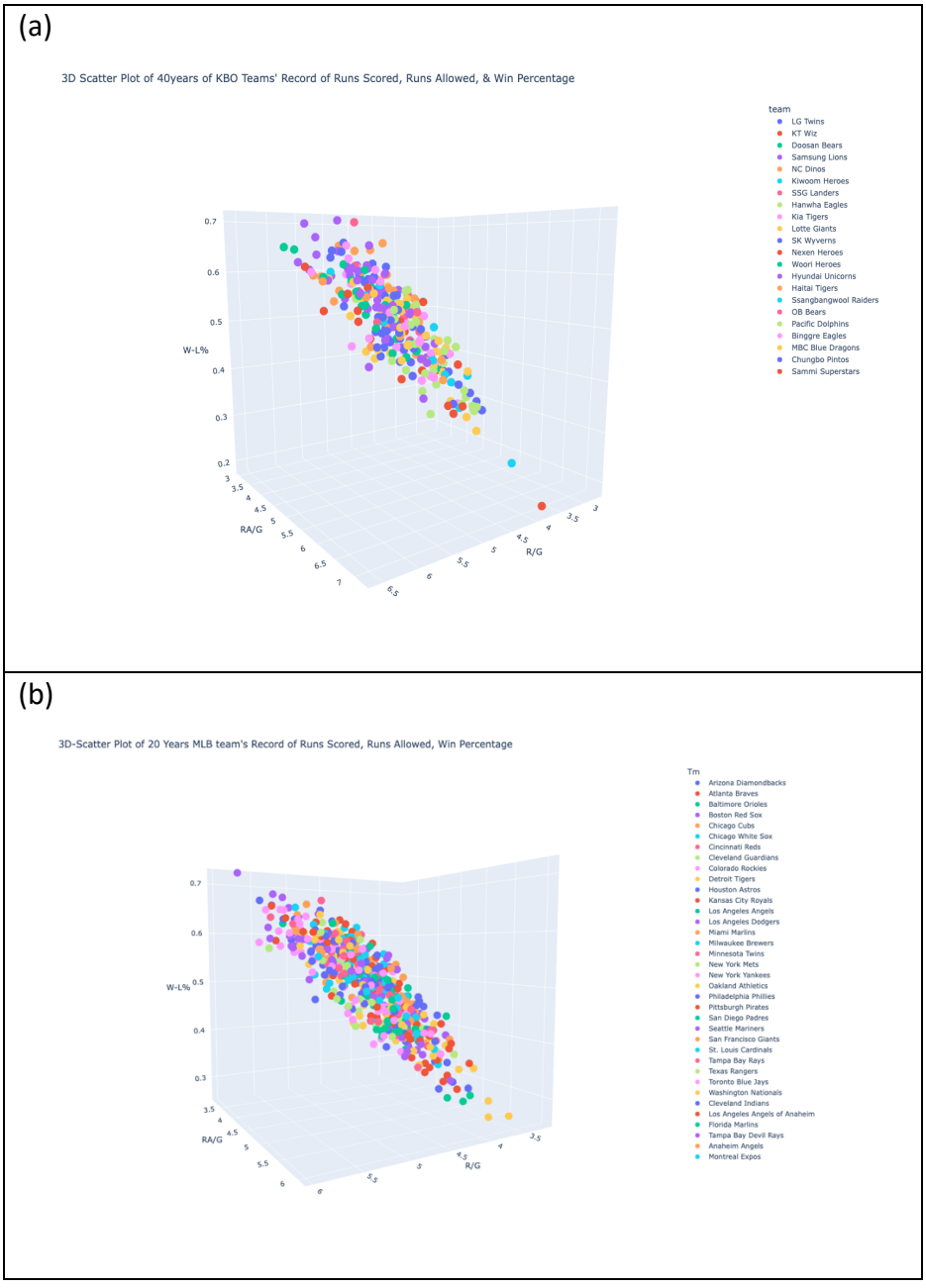


Figure 1. 3D plot of Runs Scored, Runs Allowed, Win Ratio of (a) KBO (b) MLB

2. Data and Methods

In this study, I examined KBO team data from 1982-2021 and compared it with MLB data from 2002-2023 from Baseball Reference. After a thorough review, 24 key variables from both datasets were selected for further analysis.

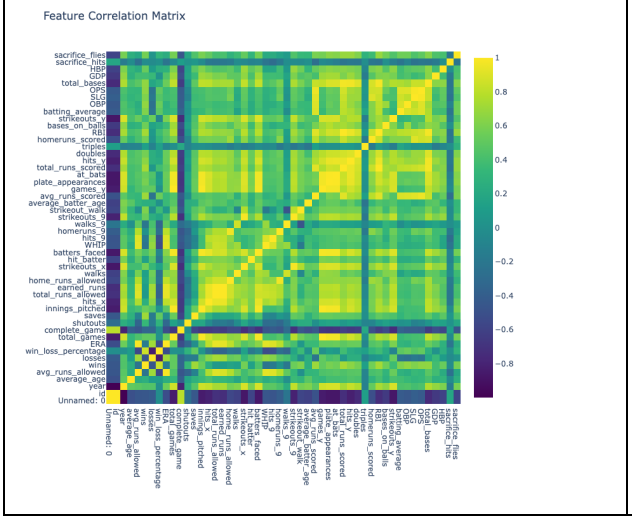
This dataset captures:

1. *Team and Player Details:* Each record is uniquely identified by an ID and encapsulates team information, the year the data pertains to, and the average age of players.
2. *Defensive Statistics:* Central to the pitching metrics are Earned Run Average (ERA), Saves, Innings Pitched, and Homeruns Allowed. Additionally, the statistics Highlights on the number of Batters Faced, Hit Batter, and Shutouts achieved. Important metrics that offer insights into a pitcher's efficiency include Walks and Hits per Innings Pitched (WHIP) and Strikeouts achieved by the pitcher. Defensive play is further emphasized by stats like grounded into double plays (GDP) and caught stealing, underscoring the team's defensive prowess beyond the mound. Together, these metrics offer a comprehensive understanding of a team's defensive capabilities and the effectiveness of its pitchers.
3. *Offensive Statistics*: From a batting standpoint, the dataset encompasses metrics such as At Bats, Hits, Doubles, Triples, Homeruns Scored, Runs Batted In (RBI), and Bases on Balls. A batter's prowess and strategy can be further understood through statistics like Slugging Percentage (SLG), Strikeouts by the Batter, and Sacrifice hits. Other relevant metrics that provide insights into a player's approach and risks they're willing to take include Stolen bases and being Hit by Pitch (HBP). While not strictly an offensive metric, the dataset also notes the average age of the batters, providing a dimension of experience and potential style variations across different age groups.

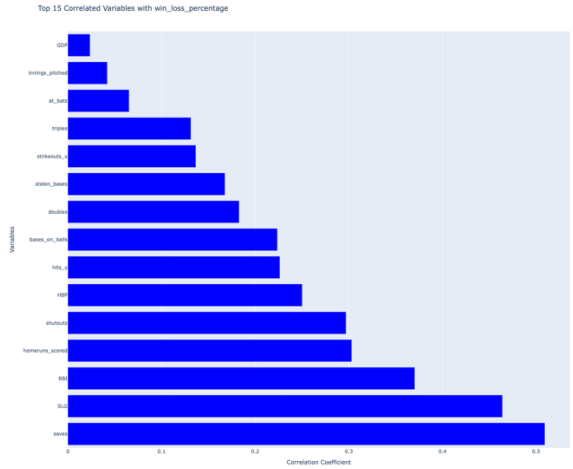
Ensuring data quality, duplicates and rows with null values were systematically filled with mean values or eliminated with the context of the baseball dataset. After this, a correlation analysis was executed to spotlight variables with significant correlation to the win ratio. Visual representations of this analysis were created using R and Python to generate a heatmap and bar chart, as displayed in Figure Panel 1.

2.1. Figure Panel 1: Top Correlated Features Visual Representation / Variable Notations and Descriptions

(a)



(b)



(c)

Variable Notation	Description
At-Bats	Number of times a player comes up to bat, excluding situations like sacrifice flies and bunts.
Average Batter Age	Average age of the team's batters.
Bases On Balls	Times a batter reaches first base from four non-strike pitches.
Batters Faced	Total batters a pitcher has pitched against.
Caught Stealing	Times a base runner is thrown out while attempting to steal a base.
Doubles	Hits where a batter reaches third base without the aid of an error or a fielding choice.
ERA (Earned Run Average)	Average earned runs given up by a pitcher over nine innings.
GDP (Ground Double Play)	Times a batter's hit results in two outs due to runners on base.
Hit Batter	Times a pitcher throws the ball that hits a batter, often resulting in the batter being awarded first base.
HBP (Hit By Pitch)	Times a batter reaches first base due to being hit by a pitch.
Hits	Number of times a batter hits the ball and safely reaches a base.
Home Runs Allowed	Number of home runs a pitcher allows to opposing batters.
Homeruns Scored	Number of hits where the batter rounds all bases and scores without aid of an error.
Innings Pitched	Total innings a pitcher has thrown in.
RBI (Runs Batted In)	Number of times a batter's hit, or the batter himself, results in a run being scored, excluding runs scored due to errors.
Sacrifice Hits	Plays in which the batter purposefully makes an out, typically through a bunt, to advance a runner on base at the expense of his own at-bat.
Saves	Times a relief pitcher successfully finishes a game for the winning team under prescribed circumstances.
Shutouts	Games where a pitcher completes the game without allowing the opposing team to score.
SLG (Slugging Percentage)	A measure of a batter's power, calculated as total bases divided by at-bats.
Stolen Bases	Bases a player advances while the ball is in the defense's possession without a hit or error.
Strikeout by Batter	Times a batter gets dismissed with three strikes.
Strikeout by Pitcher	Times a pitcher dismisses a batter with three strikes.
Triples	Hits where a batter reaches second base without the aid of an error or a fielding choice.
WHIP (Walks Plus Hits Per Inning Pitched)	Sum of walks and hits allowed, divided by innings pitched.

(a) Correlation heatmap for dataset parameters (b) Specified bar chart for parameter selection (c)Notations and descriptions for variables used in the Principal Component Analysis

2.2.Team Characteristic Identification using Principal Component Analysis

Drawing inspiration from the work of Bae, Lee, and Lee on predicting Korea Pro-Baseball rankings using Principal Component Regression Analysis, this study recognized the importance

of initiating the PCA process with preliminary preprocessing on the dataset.[3] Such an initial step was crucial in preparing the dataset for PCA's detailed analysis. Moreover, careful variable selection, as done in this study, is important to retain only significant contributors and remove redundancy.[4]

To capture essential team attributes and filter out potential noise, variables in the dataset were meticulously chosen. By emphasizing the retention of significant variables that majorly contribute to the uncovering of core patterns and omitting redundant ones, a refined dataset was formed, denoted as matrix P with dimensions (323, 24).

2.2.1. Data Standardization

Before proceeding with Principal Component Analysis (PCA), it was crucial to conduct preliminary preprocessing of our dataset to render it apt for PCA, aligning with the notion that PCA's main context is to work on a dataset with observations on several numerical variables, defining a data matrix. [5] In this reformatted dataset, denoted as matrix P, the variables were carefully chosen to capture the essential characteristics of teams without unnecessary noise. The selection ensured that I retained only those variables that significantly contribute to identifying underlying patterns while discarding redundant ones, thus forming a refined dataset represented by matrix P of shape (323, 24).

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1,24} \\ p_{21} & p_{22} & \cdots & p_{2,24} \\ \vdots & \vdots & \ddots & \vdots \\ p_{323,1} & p_{323,2} & \cdots & p_{323,24} \end{bmatrix}$$

Before the application of PCA, I standardized the dataset to ensure each variable exhibited a mean value of 0 and a standard deviation of 1. Standardization is crucial for PCA because different scales can greatly affect its results and alter the interpretation of the principal components.

Standardization is carried out using the formula:

$$P_{\text{std}} = \frac{P - \bar{P}}{\sigma(P)}$$

Here,

[3] Jae Young Bae, Jae Myung Lee, and Jung Yoon Lee, "Predicting Korea Pro-baseball rankings by principal component regression analysis," Communications for Statistical Applications and Methods 19, no. 3 (2012): 367-368.

[4] D. Rojas-Valverde et al., "A Systematic Review of Methods and Criteria Standard Proposal for the Use of Principal Component Analysis in Team's Sports Science," International Journal of Environmental Research and Public Health 17, no. 23 (2020): 3.

[5] Ian T. Jolliffe and Jorge Cadima, "Principal component analysis: a review and recent developments," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374, no. 2065 (2016): 2.

- P denotes the data matrix
- $\bar{P}$ symbolizes the mean value derived from the data matrix
- $\sigma(P)$ stands for the standard deviation of the data matrix.

In the context of my research utilizing a baseball dataset, the choice of standard deviation ($\sigma(P)$) was due to the dynamic nature of the game where a wide array of situations and strategies can significantly vary the individual data points. These fluctuations manifest as large variations in the data, thus justifying the use of standard deviation to bring all variables to a comparable scale, reducing the influence of high-variance features in the PCA and facilitating the extraction of the most meaningful components.[6]

2.2.2. Computing the Covariance Matrix

The covariance matrix, denoted as *S*, is computed using the standardized data matrix P_{std}.This matrix encapsulates the pairwise relationships between different variables in the dataset, providing a mathematical representation of the dataset's variance and covariance.
Each element of the matrix Sij represents the covariance between feature *i* and feature *j*, and the diagonal elements *Sii* represent the variance of feature *i*.[7]

The covariance matrix is not only central in understanding the relationships and affinities between the different variables, but it also has important mathematical properties; it is symmetric and positive semi-definite. These properties ensure that the eigenvalues derived in the subsequent step are non-negative, facilitating a real-valued solution to the eigen decomposition problem. The mathematical expression for computing the covariance matrix is given by the formula:

$$S = \frac{1}{322} P_{\text{std}}^{T} P_{\text{std}} \quad (1.1)$$

where the (322) is derived from $(323 - 1)$, accounting for degrees of freedom, and (P_{std}^{T}) is the transpose of the standardized data matrix.

2.2.3. Eigen Decomposition

Proceeding with the PCA, the next critical step is to determine the eigenvalues and eigenvectors of the covariance matrix. These eigenvalues and eigenvectors form the core of PCA, and their values are obtained through the eigen decomposition of the covariance matrix S.[8]. While various numerical algorithms exist for this purpose, like the power iteration method, the Singular Value Decomposition (SVD) method is often favored in practical implementations of PCA due to its computational efficiency and numerical stability.[9]

---

[6] Rasmus Bro and Age K. Smilde, "Principal component analysis," Analytical Methods 6, no. 9 (2014): 2817.
[7] Jorge Cadima and Ihaka, "Data Standardization in Principal Component Analysis: One Step Further," Journal of Data Science 27, no. 4 (2019): 423.
[8] J.L. Kirschvink, "The Least-Squares Line and Plane and the Analysis of Palaeomagnetic Data," Geophysical Journal International 62, no. 3 (September 1980): 702.
[9] Huamin Li et al., "Algorithm 971: An Implementation of a Randomized Algorithm for Principal Component Analysis," ACM Trans. Math. Softw. 43, no. 3 (September 2017): 3, https://doi.org/10.1145/3004053.

Eigenvectors point out the principal directions in the feature space, and the eigenvalues indicate the magnitude of variance along these directions. This relationship can be mathematically described through the characteristic equation:

$$\det(S - \lambda I) = 0$$

where (det) denotes the determinant, (λ) are the eigenvalues, (I) is the identity matrix, and (S) is the covariance matrix derived in equation 1.1.

In this equation, the eigenvalues (λ) indicate the amount of variance captured by each principal component, hence playing a critical role in determining the importance of each principal component in representing the data. The eigenvalues (λ) signify the variance captured by each principal component, making them instrumental in determining the significance of each principal component. The larger the eigenvalue, the more variance it accounts for in the dataset.[10] In contrast, eigenvectors provide the coefficients for forming the principal components from linear combinations of the original features. Their orthogonality ensures that principal components remain uncorrelated, thus each component offers unique information. This property allows PCA to effectively reduce dimensionality by eliminating redundant information while preserving the essential data structure.[11]

2.2.4. Finding Principal Components

The principal components, derived from the eigen decomposition of the covariance matrix, reveal the directions of maximal data variance. The transformation matrix, A, formed by the eigenvectors, enables the projection of our standardized data onto this new subspace.[12] Specifically, multiplying the standardized data matrix, P_{std} by A results in the matrix Z. This matrix encapsulates the data in terms of its principal components, offering a clearer representation of its inherent structure.[13] This transformation not only makes the data more interpretable but also often achieves dimensionality reduction without significant loss of information.

Mathematically, this projection is represented as:

$$Z = P_{\text{std}}A$$

[10] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," Chemometrics and Intelligent Laboratory Systems 2, no. 1-3 (1987): 41.
[11] H. Abdi and L.J. Williams, "Principal component analysis," Wiley Interdisciplinary Reviews: Computational Statistics 2, no. 4 (2010): 438.
[12] J.L. Kirschvink, "The Least-Squares Line and Plane and the Analysis of Palaeomagnetic Data," Geophysical Journal International 62, no. 3 (September 1980): 705.
[13] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," Chemometrics and Intelligent Laboratory Systems 2, no. 1-3 (1987): 40.

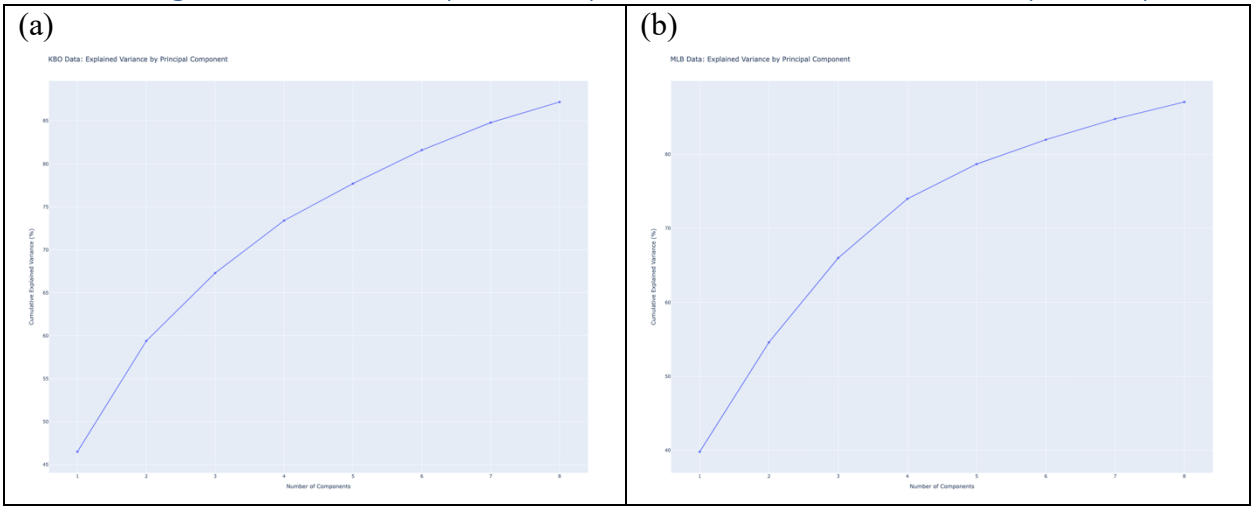
where A is the matrix of eigenvectors and P_{std} is the standardized data matrix. This transformed data, Z, encapsulates the essence of the original data but is now oriented along the axes of maximal variance.[14]

2.2.5. Interpreting Principal Components

After identifying the principal components, they are assessed based on the explained variance, which is derived from the eigenvalues. Typically, a few principal components account for a substantial amount of the total variance in the dataset, allowing for a reduced-dimensional representation without significant loss of information.

The principal components can be used to understand underlying patterns in the dataset, giving a deep insight into the latent variables governing the observed variables. Through this analysis, we can effectively reduce the dimensionality of our dataset, aiding in more efficient data analysis and visualization.

2.3. Figure Panel 2: Scree plot for Explained Variance for each Principal Component



(a) The scree plot for the KBO data displays the explained variance captured by each principal component. Scree plots depict the eigenvalues of the correlation matrix, with the steep slope on the left indicating the most informative components.[15] The scree plot for the KBO data shows the explained variance for each principal component. The first component accounts for 46.5% of the total variance. This is followed by the second component explaining an additional 12.9% variance (59.4% - 46.5%). The plot begins to taper off after the first two components, with diminishing variance explained by each subsequent component. (b) The MLB data scree plot also displays an initial steep drop-off, with the first principal component explaining 39.8% variance. The second component accounts for an additional 14.8% variance (54.6% - 39.8%). The first two components together explain over 50% of the total variance. After this point, the plot slowly levels off with each further component adding a smaller proportion of explained variance. The first 3 components collectively explain approximately 66% of the total variance

[14] I. Viola, M. Chen, and T. Isenberg, “Visual Abstraction,” in Foundations of Data Visualization, eds. M. Chen et al. (Cham: Springer, 2020), 55.

[15] C.C. David and D.J. Jacobs, "Principal component analysis: a method for determining the essential dynamics of proteins," Methods in Molecular Biology 1084 (2014): 197.

based on the elbow point of the plot. In both cases, we see that the first few (2-3) components explain a substantial portion of the variance, after which additional components contribute diminishing returns. This suggests that a reduced dimensional representation using just the top couple of components can effectively summarize most of the information in the high-dimensional dataset. The scree plot helps identify the optimal trade-off between dimensionality reduction and retention of variance.

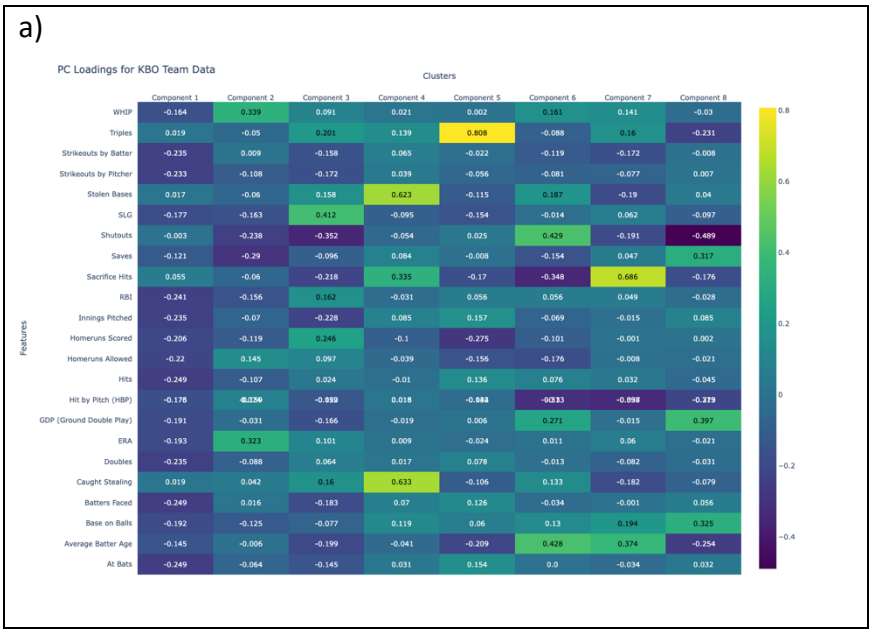
3. Result

3.1. Establishing Baseline Rules for PC Component Interpretation

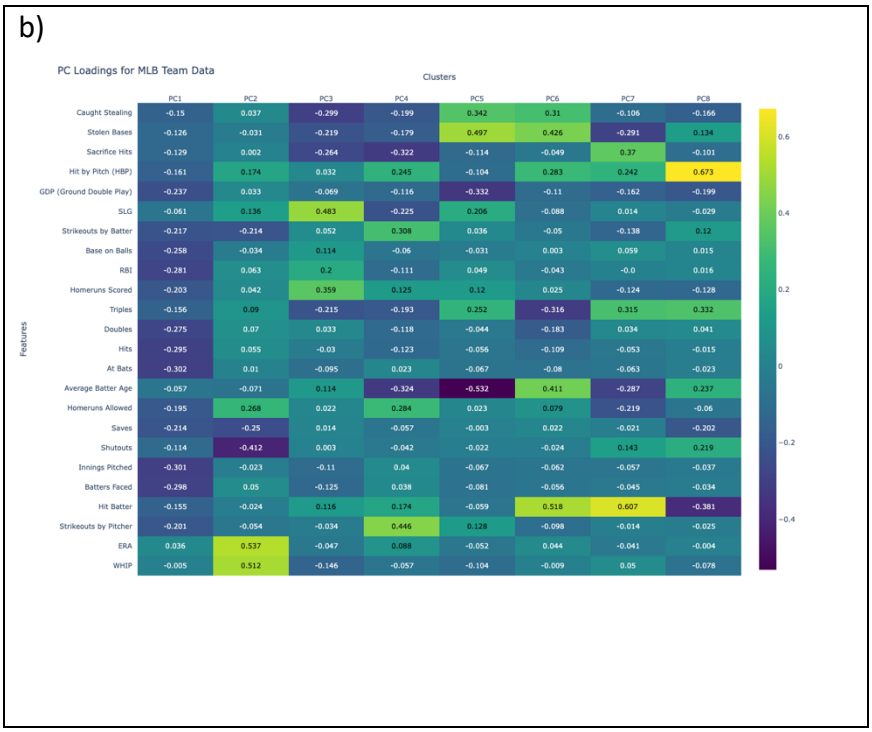
In this subsection, I establish a structured pathway to interpret the data derived from the principal component analysis of KBO combined data. Drawing inspiration from Wedding et al.'s examination of team and player performance in elite rugby league, I adopted a methodology that systematically interprets performance indicators based on their factor loadings.[16] In their study, indicators with values greater than 0.60 from the rotated component matrix were considered significant contributors to identified playing styles. Similarly, for the KBO data, I've devised a set of rules to interpret the contribution and significance of various performance metrics.

3.1.1. Figure Panel 3: Heatmap Representation of PC loadings

a)



[16] CJ Wedding et al., "Operational Insights into Analyzing Team and Player Performance in Elite Rugby League: A Narrative Review with Case Examples," Sports Medicine - Open 8, no. 1 (December 3, 2022): 140.



Heatmap representation of eigenvalues derived from Principal components: a) Principal component loading of KBO dataset. b) Principal component loading of MLB dataset

3.1.2. Detailed Explanation of Key Performance Variables

To enable a detailed discussion on the strategic orientation of the teams in the KBO league based on various performance metrics, I identify components with high positive and negative loadings for each variable. The understanding and analysis of these loadings are important in extracting meaningful insights from the PCA clustering analysis.

Below, I dissect each variable, delving into what the different loadings in various components imply, guided visually through a comprehensive heatmap and representation that portrays the eigenvalues associated with each variable across different clusters (See Figure Panel 3):

3.1.2.1. Thresholding

To categorize the PC loadings derived from the KBO data, I employed a thresholding strategy grounded in the distribution of each component's values. This approach emphasizes the importance of both the magnitude and direction (sign) of each loading for precise data interpretation. Firstly, I computed the mean (μ_i) and standard deviation (σ_i) for each component individually, facilitating a nuanced understanding of the distribution of loadings in each component.

Next, I defined four categories based on the distance of each loading from the mean, in terms of standard deviations, as well as the sign of the loading (Refer to Table 6.1, 6.2 to see complete categorized label):

- *High Positive*: A loading is classified as high positive if it is greater than or equal to one standard deviation above the mean and is non-negative *($x_{ij} \geq \mu_i + \sigma_i$ and $x_{ij} \geq 0$)*

- *Low Positive:* A loading falls into this category if it is less than one standard deviation above the mean but still non-negative, and higher than one standard deviation below the mean($\mu_i - \sigma_i \leq x_{ij} \leq \mu_i + \sigma_i \; and \; x_{ij} \geq 0$)
- *High Negative:* This category includes loadings that are greater than or equal to one standard deviation above the mean but are negative ($x_{ij} \geq \mu_i + \sigma_i$ and $x_{ij} < 0$)
- *Low Negative:* Loadings that are less than one standard deviation above the mean, negative, and higher than one standard deviation below the mean are classified as low negative ($\mu_i - \sigma_i < x_{ij} < \mu_i + \sigma_i$ and $x_{ij} < 0$)

3.1.2.2. WHIP (Walks plus Hits per Innings Pitched)

PC Component Analysis and Interpretation:

- *Component 2 (0.339)*
- *High Positive Loading (WHIP)*
- Teams in this component exhibit high positive loadings for the WHIP variable, indicating a less effective pitching strategy characterized by a high number of walks and hits per innings pitched. A high WHIP value signifies that the pitchers allow too many batters to reach base, thus reflecting a potential area of concern and a possible indicator of poor pitching performance. Strategies for improvement might include enhancing pitch selection, improving control to reduce walks, or working on defensive strategies to reduce hits allowed.

- *Component 1 (-0.164)*
- *Low Negative Loading (WHIP)*
- Contrarily, principal component 1 represent teams maintaining a strong pitching performance, as illustrated by the low negative loading on the WHIP variable. These teams have successfully controlled the number of walks and hits per innings pitched, achieving a lower WHIP value and, consequently, a stronger defense strategy. Teams in these clusters are likely characterized by skilled pitchers who can maintain a low WHIP, possibly resulting in fewer scoring opportunities for opposing teams. To continue this trend, teams might focus on nurturing pitcher talent and adopting strategies that favor a low WHIP.

3.1.2.3. ERA (Walks plus Hits per Innings Pitched)

PC Component Analysis and Interpretation:

- *Component 2 (0.323)*
  - *High Positive Loading (ERA)*
  - Teams within this component showcase high positive loadings for the ERA variable. A heightened ERA signals that teams are giving up a significant number of earned runs over their innings pitched. Essentially, it paints a picture of a weaker pitching performance.

3.1.2.4. Average Batter Age

PC Component Analysis and Interpretation:

- *Component 3 (-0.199)*
- *High Negative Loading (Average Batter Age)*
- Teams within this component show high negative loadings for the "Average Batter Age" variable. This indicates that these teams have a younger average batter age relative to others. Such teams might benefit from the agility, speed, and potential for long-term player development that younger batters typically bring. However, they might also face challenges related to inexperience, especially when competing against teams with experienced players. To optimize performance, these teams might consider blending their squads with a mix of young talent and experienced players to benefit from both agility and expertise.

- *Component 6 (0.428)*
- *High Positive Loading*
- This component reveals high positive loadings for the "Average Batter Age" variable, suggesting that these teams have an older average batter age. Older batters often come with a wealth of experience, which can be advantageous in high-pressure situations and can serve as a guiding force for younger players. Their strategic understanding of the game can be a crucial asset. However, potential challenges for such teams might include issues related to physical fitness, agility, or speed when compared to younger players.

3.1.2.5. Caught Stealing

PC Component Analysis and Interpretation:

- *Component 1 (0.019): High Positive Loading*
- Teams categorized within this component exhibit high positive loadings for the "Caught Stealing" variable. This suggests that these teams experience a relatively higher number of instances where their base runners are caught stealing bases. While attempting to steal bases can demonstrate an aggressive offensive strategy, being frequently caught can impede momentum and waste scoring opportunities. It may also be indicative of either poor base running decisions or exceptional defense by the opposing team.

- *Component 7 (-0.182)*
- *High Negative Loading*
- Teams associated with this component reflect high negative loadings for the "Caught Stealing" variable. This indicates that these teams have fewer instances of their players being caught while attempting to steal bases. Such teams either have exceptional base runners, adopt a conservative base-stealing approach, or both. This can be a significant asset as it allows teams to advance players into scoring positions without surrendering outs needlessly. To maintain this edge, teams could continue to emphasize the importance of reading pitchers, ensuring base runners are well-aware of situational baseball scenarios, and working with coaches to identify ideal base-stealing opportunities.

3.1.2.6. GDP (Ground Double Play)

PC Component Analysis and Interpretation:

- *Component 6: High Positive Loading (0.271)*

- Teams under component 6 exhibit high positive loadings for GDP (Ground Double Play). Ground double plays are pivotal moments in a game, often diffusing potentially high-scoring situations for the opposition. For teams with this trait, their defense shines by minimizing threats even when runners are on base, especially on first base, leading to fewer scoring opportunities for the opposing teams. However, it's also worth noting that consistently relying on double plays might suggest that these teams often find themselves in situations with runners on base, indicating occasional weaknesses in their pitching or defense. Thus, while inducing double plays is a strength, it might be beneficial for these teams to also address and reduce situations where they're vulnerable to offensive threats. Continued success in this metric might involve focusing on drills that emphasize quick infield reactions and precise throws, as well as pitchers mastering low and outside pitches that are more likely to result in grounders.

3.1.2.7. Homeruns Scored

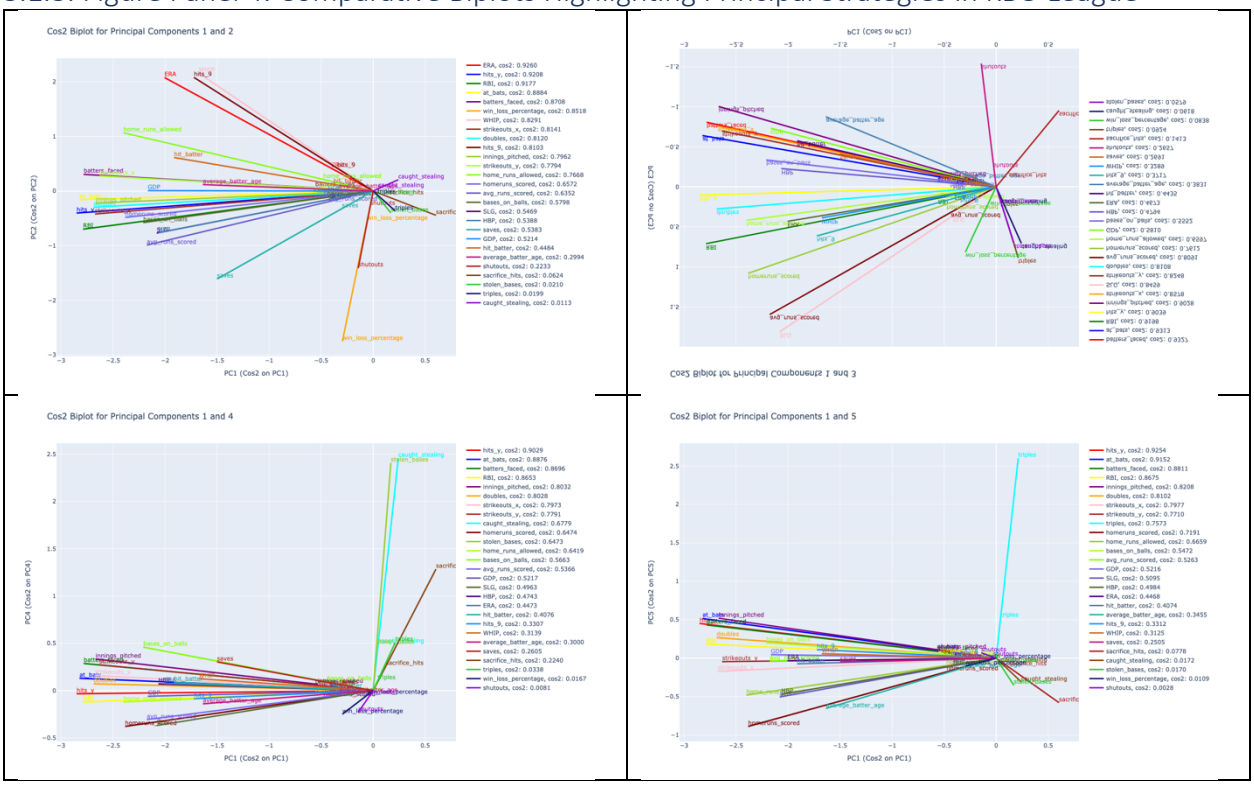
PC Component Analysis and Interpretation:

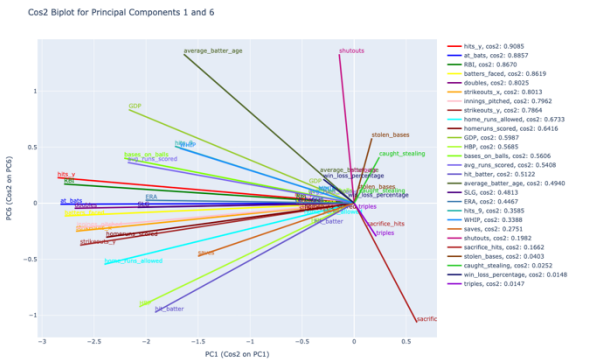
- *Component 1: Low Negative Loading (-0.206)*
- Teams under this component display a low negative loading for the "Homeruns Scored" variable. This implies that while these teams may not be the lowest scoring in terms of homeruns, they certainly fall below the league's average. A low homerun scoring capability can be a consequence of a more strategic, base-hit oriented approach or potentially due to the lack of power hitters in the lineup. Such teams might need to focus on alternative strategies for run production, including base-stealing, bunting, or situational hitting.
- *Component 3: High Positive Loading (0.246)*
- On the flip side, teams within Component 3 demonstrate a high positive loading. This indicates an above-average homerun-scoring capability. Teams in this bracket likely possess strong power hitters who can capitalize on pitcher mistakes and deliver game-changing plays. Such a lineup can be intimidating for the opposition, providing a psychological edge. For these teams, maintaining this momentum might involve investing in strength and conditioning programs tailored for power hitters and perhaps even hiring specialized coaches to further hone their skills.
- *Component 5: High Negative Loading (-0.275)*
- Teams in Component 5 exhibit a high negative loading, suggesting they score significantly fewer homeruns compared to most teams in the league. Such a pronounced low score might indicate potential areas of concern in the team's offensive lineup. This could be due to a lack of power hitters, ineffective hitting strategies, or facing particularly strong pitching opponents. Teams within this cluster may need to reevaluate their roster and offensive strategies. Scouting for emerging power hitters or trading for established ones, as well as leveraging advanced analytics to study pitcher-batter matchups, might be potential avenues for improvement.

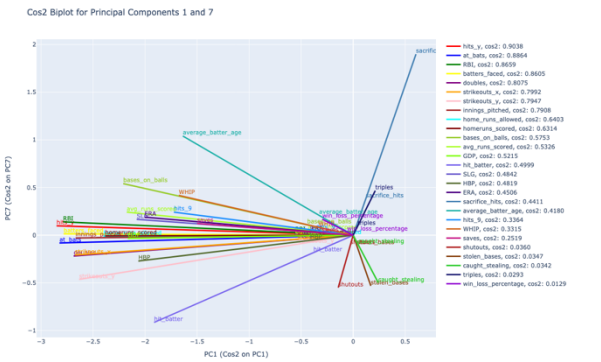
3.1.2.8. Sacrifice Hits

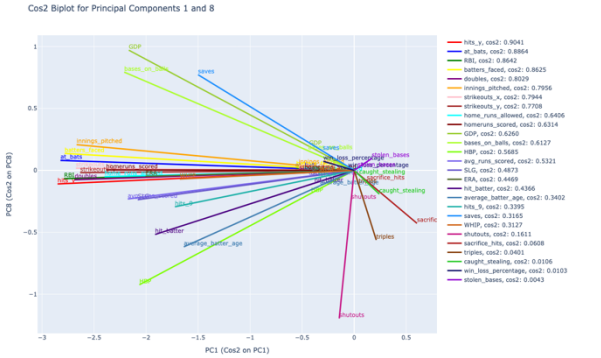
- *Component 1: High Positive Loading (0.055)*
- Teams under this component demonstrate a high positive loading for the "Sacrifice Hits" variable. This suggests that these teams frequently employ strategies that prioritize advancing runners at the expense of an out, often in tight situations where a strategic play might lead to a scoring opportunity in the subsequent at-bats. Sacrifice hits, including bunts, play a pivotal role in situations where manufacturing a run becomes crucial, like in closely contested games. Teams in this component might be more traditionally inclined, valuing small-ball tactics to scratch across runs. For them, ensuring that their players are well-drilled in bunting techniques and understanding game situations where a sacrifice might be beneficial will be essential.
- *Component 3: High Negative Loading (-0.218)*
- On the other hand, teams under Component 3 showcase a high negative loading for the "Sacrifice Hits" metric. This can imply a different offensive strategy where these teams might prefer letting their batters swing away, rather than giving up an out to advance a runner. Teams with this trait could be betting on their batting lineup's strength and capability to hit for extra bases or bring runners home without the need for sacrifices. Such an approach might indicate a team with more power hitters or one that's confident in its ability to string together hits.

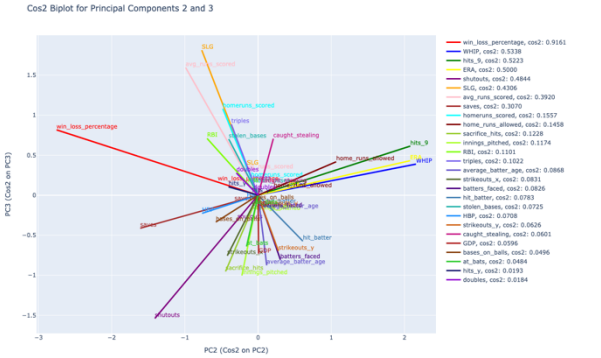
3.1.3. Figure Panel 4: Comparative Biplots Highlighting Principal Strategies in KBO League

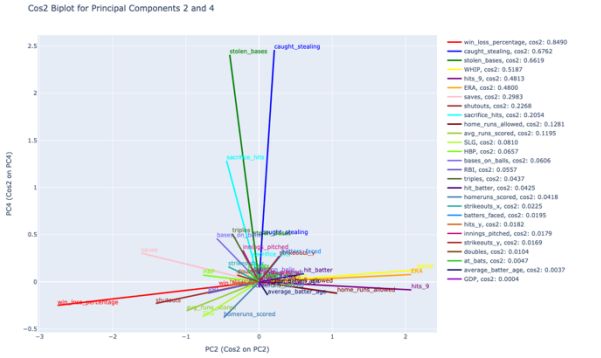


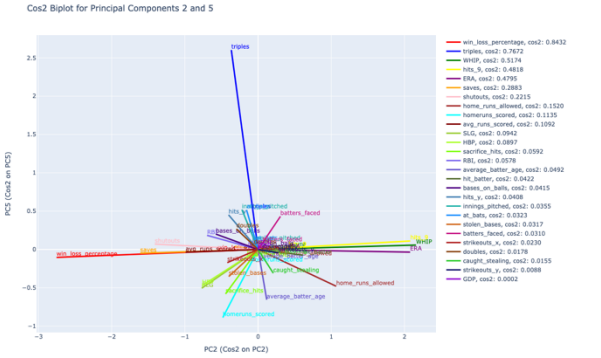

Cos2 Biplot for Principal Components 1 and 6
PC1 (Cos2 on PC1)
PC6 (Cos2 on PC6)
hits_y, cos2: 0.9085
at_bats, cos2: 0.8857
RBI, cos2: 0.8670
batters_faced, cos2: 0.8619
doubles, cos2: 0.8025
strikeouts_x, cos2: 0.8013
innings_pitched, cos2: 0.7962
strikeouts_y, cos2: 0.7864
home_runs_allowed, cos2: 0.6733
homeruns_scored, cos2: 0.6416
GDP, cos2: 0.5987
HBP, cos2: 0.5685
bases_on_balls, cos2: 0.5606
avg_runs_scored, cos2: 0.5408
hit_batter, cos2: 0.5122
average_batter_age, cos2: 0.4940
SLG, cos2: 0.4813
ERA, cos2: 0.4667
hits_9, cos2: 0.3585
WHIP, cos2: 0.3388
saves, cos2: 0.2751
shutouts, cos2: 0.1982
sacrifice_hits, cos2: 0.1662
stolen_bases, cos2: 0.0403
caught_stealing, cos2: 0.0252
win_loss_percentage, cos2: 0.0148
triples, cos2: 0.0147

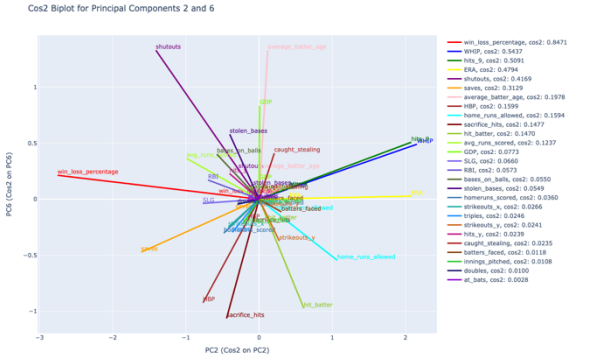

Cos2 Biplot for Principal Components 1 and 7
PC1 (Cos2 on PC1)
PC7 (Cos2 on PC7)
hits_y, cos2: 0.9038
at_bats, cos2: 0.8864
RBI, cos2: 0.8659
batters_faced, cos2: 0.8605
doubles, cos2: 0.8075
strikeouts_x, cos2: 0.7992
strikeouts_y, cos2: 0.7947
innings_pitched, cos2: 0.7908
home_runs_allowed, cos2: 0.6403
homeruns_scored, cos2: 0.6314
bases_on_balls, cos2: 0.5753
avg_runs_scored, cos2: 0.5326
GDP, cos2: 0.5215
hit_batter, cos2: 0.4999
SLG, cos2: 0.4842
HBP, cos2: 0.4819
ERA, cos2: 0.4506
sacrifice_hits, cos2: 0.4411
average_batter_age, cos2: 0.4180
hits_9, cos2: 0.3364
WHIP, cos2: 0.3315
saves, cos2: 0.2519
shutouts, cos2: 0.0360
stolen_bases, cos2: 0.0347
caught_stealing, cos2: 0.0342
triples, cos2: 0.0293
win_loss_percentage, cos2: 0.0129

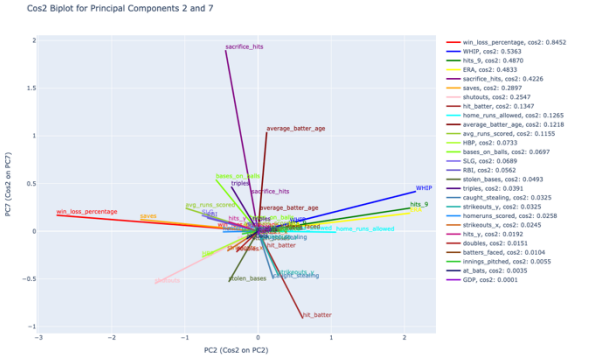

Cos2 Biplot for Principal Components 1 and 8
PC1 (Cos2 on PC1)
PC8 (Cos2 on PC8)
hits_y, cos2: 0.9041
at_bats, cos2: 0.8864
RBI, cos2: 0.8642
batters_faced, cos2: 0.8625
doubles, cos2: 0.8029
innings_pitched, cos2: 0.7956
strikeouts_x, cos2: 0.7944
strikeouts_y, cos2: 0.7708
home_runs_allowed, cos2: 0.6406
homeruns_scored, cos2: 0.6314
GDP, cos2: 0.6260
bases_on_balls, cos2: 0.6127
HBP, cos2: 0.5685
avg_runs_scored, cos2: 0.5321
SLG, cos2: 0.4873
ERA, cos2: 0.4469
hit_batter, cos2: 0.4366
average_batter_age, cos2: 0.3402
hits_9, cos2: 0.3395
saves, cos2: 0.3165
WHIP, cos2: 0.3127
shutouts, cos2: 0.1611
sacrifice_hits, cos2: 0.0608
triples, cos2: 0.0401
caught_stealing, cos2: 0.0106
win_loss_percentage, cos2: 0.0103
stolen_bases, cos2: 0.0043


Cos2 Biplot for Principal Components 2 and 3
PC2 (Cos2 on PC2)
PC3 (Cos2 on PC3)
win_loss_percentage, cos2: 0.9161
WHIP, cos2: 0.5338
hits_9, cos2: 0.5223
ERA, cos2: 0.5000
shutouts, cos2: 0.4844
SLG, cos2: 0.4306
avg_runs_scored, cos2: 0.3920
saves, cos2: 0.3070
homeruns_scored, cos2: 0.1557
home_runs_allowed, cos2: 0.1458
sacrifice_hits, cos2: 0.1228
innings_pitched, cos2: 0.1174
RBI, cos2: 0.1101
triples, cos2: 0.1022
average_batter_age, cos2: 0.0868
strikeouts_x, cos2: 0.0831
batters_faced, cos2: 0.0826
hit_batter, cos2: 0.0783
stolen_bases, cos2: 0.0725
HBP, cos2: 0.0708
strikeouts_y, cos2: 0.0626
caught_stealing, cos2: 0.0601
GDP, cos2: 0.0596
bases_on_balls, cos2: 0.0496
at_bats, cos2: 0.0484
hits_y, cos2: 0.0193
doubles, cos2: 0.0184

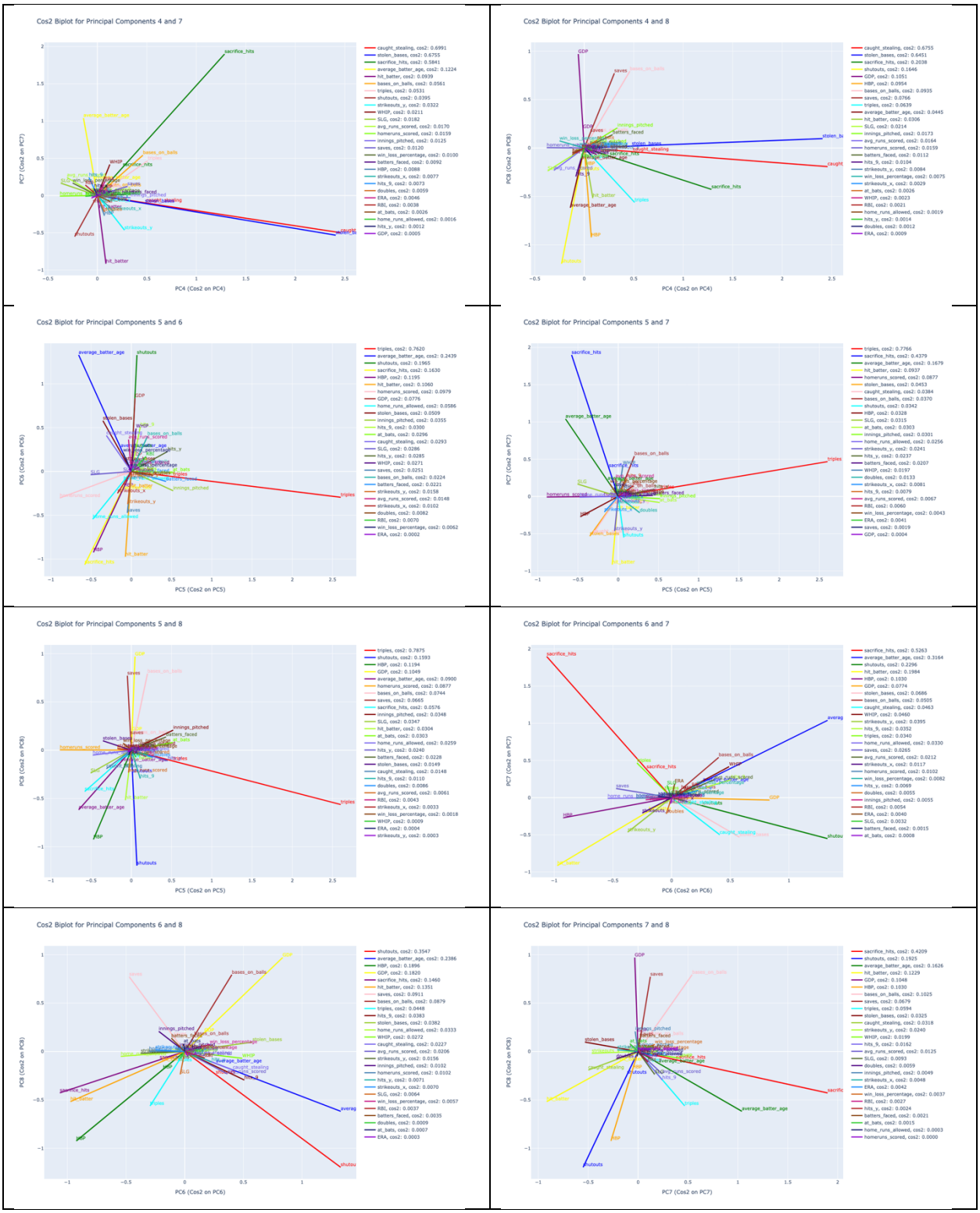

Cos2 Biplot for Principal Components 2 and 4
PC2 (Cos2 on PC2)
PC4 (Cos2 on PC4)
win_loss_percentage, cos2: 0.8490
caught_stealing, cos2: 0.6762
stolen_bases, cos2: 0.6619
WHIP, cos2: 0.5187
hits_9, cos2: 0.4813
ERA, cos2: 0.4800
saves, cos2: 0.2983
shutouts, cos2: 0.2268
sacrifice_hits, cos2: 0.2054
home_runs_allowed, cos2: 0.1281
avg_runs_scored, cos2: 0.1195
SLG, cos2: 0.0810
HBP, cos2: 0.0657
bases_on_balls, cos2: 0.0606
RBI, cos2: 0.0557
triples, cos2: 0.0437
hit_batter, cos2: 0.0425
homeruns_scored, cos2: 0.0418
strikeouts_x, cos2: 0.0225
batters_faced, cos2: 0.0195
hits_y, cos2: 0.0182
innings_pitched, cos2: 0.0179
strikeouts_y, cos2: 0.0169
doubles, cos2: 0.0104
at_bats, cos2: 0.0047
average_batter_age, cos2: 0.0037
GDP, cos2: 0.0004


Cos2 Biplot for Principal Components 2 and 5
PC2 (Cos2 on PC2)
PC5 (Cos2 on PC5)
win_loss_percentage, cos2: 0.8432
triples, cos2: 0.7672
WHIP, cos2: 0.5174
hits_9, cos2: 0.4818
ERA, cos2: 0.4795
saves, cos2: 0.2883
shutouts, cos2: 0.2215
home_runs_allowed, cos2: 0.1520
homeruns_scored, cos2: 0.1135
avg_runs_scored, cos2: 0.1092
SLG, cos2: 0.0942
HBP, cos2: 0.0897
sacrifice_hits, cos2: 0.0592
RBI, cos2: 0.0578
average_batter_age, cos2: 0.0492
hit_batter, cos2: 0.0422
bases_on_balls, cos2: 0.0415
hits_y, cos2: 0.0408
innings_pitched, cos2: 0.0355
at_bats, cos2: 0.0323
stolen_bases, cos2: 0.0317
batters_faced, cos2: 0.0310
strikeouts_x, cos2: 0.0230
doubles, cos2: 0.0178
caught_stealing, cos2: 0.0155
strikeouts_y, cos2: 0.0088
GDP, cos2: 0.0002


Cos2 Biplot for Principal Components 2 and 6
PC2 (Cos2 on PC2)
PC6 (Cos2 on PC6)
win_loss_percentage, cos2: 0.8471
WHIP, cos2: 0.5437
hits_9, cos2: 0.5091
ERA, cos2: 0.4794
shutouts, cos2: 0.4169
saves, cos2: 0.3129
average_batter_age, cos2: 0.1978
HBP, cos2: 0.1599
home_runs_allowed, cos2: 0.1594
sacrifice_hits, cos2: 0.1477
hit_batter, cos2: 0.1470
avg_runs_scored, cos2: 0.1237
GDP, cos2: 0.0773
SLG, cos2: 0.0660
RBI, cos2: 0.0573
bases_on_balls, cos2: 0.0550
stolen_bases, cos2: 0.0549
homeruns_scored, cos2: 0.0360
strikeouts_x, cos2: 0.0266
triples, cos2: 0.0246
strikeouts_y, cos2: 0.0241
hits_y, cos2: 0.0239
caught_stealing, cos2: 0.0235
batters_faced, cos2: 0.0118
innings_pitched, cos2: 0.0108
doubles, cos2: 0.0100
at_bats, cos2: 0.0028


Cos2 Biplot for Principal Components 2 and 7
PC2 (Cos2 on PC2)
PC7 (Cos2 on PC7)
win_loss_percentage, cos2: 0.8452
WHIP, cos2: 0.5363
hits_9, cos2: 0.4870
ERA, cos2: 0.4833
sacrifice_hits, cos2: 0.4226
saves, cos2: 0.2897
shutouts, cos2: 0.2547
hit_batter, cos2: 0.1347
home_runs_allowed, cos2: 0.1265
average_batter_age, cos2: 0.1218
avg_runs_scored, cos2: 0.1155
HBP, cos2: 0.0733
bases_on_balls, cos2: 0.0697
SLG, cos2: 0.0689
RBI, cos2: 0.0562
stolen_bases, cos2: 0.0493
triples, cos2: 0.0391
caught_stealing, cos2: 0.0325
strikeouts_y, cos2: 0.0325
homeruns_scored, cos2: 0.0258
strikeouts_x, cos2: 0.0245
hits_y, cos2: 0.0192
doubles, cos2: 0.0151
batters_faced, cos2: 0.0104
innings_pitched, cos2: 0.0055
at_bats, cos2: 0.0035
GDP, cos2: 0.0001


Cos2 Biplot for Principal Components 2 and 8
PC2 (Cos2 on PC2)
PC8 (Cos2 on PC8)
win_loss_percentage, cos2: 0.8426
WHIP, cos2: 0.5175
hits_9, cos2: 0.4901
ERA, cos2: 0.4796
shutouts, cos2: 0.3798
saves, cos2: 0.3543
HBP, cos2: 0.1599
home_runs_allowed, cos2: 0.1267
avg_runs_scored, cos2: 0.1150
bases_on_balls, cos2: 0.1070
GDP, cos2: 0.1047
SLG, cos2: 0.0720
hit_batter, cos2: 0.0714
RBI, cos2: 0.0545
triples, cos2: 0.0500
average_batter_age, cos2: 0.0439
sacrifice_hits, cos2: 0.0423
homeruns_scored, cos2: 0.0258
strikeouts_x, cos2: 0.0197
hits_y, cos2: 0.0194
stolen_bases, cos2: 0.0189
batters_faced, cos2: 0.0124
doubles, cos2: 0.0104
innings_pitched, cos2: 0.0102
caught_stealing, cos2: 0.0089
strikeouts_y, cos2: 0.0086
at_bats, cos2: 0.0035
Cos2 Biplot for Principal Components 3 and 4
PC3 (Cos2 on PC3)
PC4 (Cos2 on PC4)
caught_stealing, cos2: 0.7268
stolen_bases, cos2: 0.6988
SLG, cos2: 0.3800
avg_runs_scored, cos2: 0.2934
sacrifice_hits, cos2: 0.2843
shutouts, cos2: 0.2693
homeruns_scored, cos2: 0.1458
innings_pitched, cos2: 0.1245
triples, cos2: 0.1162
average_batter_age, cos2: 0.0874
batters_faced, cos2: 0.0814
win_loss_percentage, cos2: 0.0810
strikeouts_x, cos2: 0.0663
strikeouts_y, cos2: 0.0624
GDP, cos2: 0.0600
RBI, cos2: 0.0577
at_bats, cos2: 0.0475
hits_9, cos2: 0.0426
hit_batter, cos2: 0.0375
bases_on_balls, cos2: 0.0361
saves, cos2: 0.0292
ERA, cos2: 0.0213
home_runs_allowed, cos2: 0.0210
WHIP, cos2: 0.0185
doubles, cos2: 0.0092
HBP, cos2: 0.0063
hits_y, cos2: 0.0013
Cos2 Biplot for Principal Components 3 and 5
PC3 (Cos2 on PC3)
PC5 (Cos2 on PC5)
triples, cos2: 0.8397
SLG, cos2: 0.3933
avg_runs_scored, cos2: 0.2831
shutouts, cos2: 0.2640
homeruns_scored, cos2: 0.2175
innings_pitched, cos2: 0.1420
sacrifice_hits, cos2: 0.1381
average_batter_age, cos2: 0.1329
batters_faced, cos2: 0.0930
win_loss_percentage, cos2: 0.0753
at_bats, cos2: 0.0752
stolen_bases, cos2: 0.0685
strikeouts_x, cos2: 0.0667
caught_stealing, cos2: 0.0660
GDP, cos2: 0.0599
RBI, cos2: 0.0598
strikeouts_y, cos2: 0.0543
home_runs_allowed, cos2: 0.0450
hits_9, cos2: 0.0432
hit_batter, cos2: 0.0373
HBP, cos2: 0.0304
hits_y, cos2: 0.0239
ERA, cos2: 0.0208
saves, cos2: 0.0192
WHIP, cos2: 0.0171
bases_on_balls, cos2: 0.0170
doubles, cos2: 0.0165
Cos2 Biplot for Principal Components 3 and 6
PC3 (Cos2 on PC3)
PC6 (Cos2 on PC6)
shutouts, cos2: 0.4594
SLG, cos2: 0.3650
avg_runs_scored, cos2: 0.2977
average_batter_age, cos2: 0.2815
sacrifice_hits, cos2: 0.2265
hit_batter, cos2: 0.1420
homeruns_scored, cos2: 0.1401
GDP, cos2: 0.1370
innings_pitched, cos2: 0.1174
HBP, cos2: 0.1005
triples, cos2: 0.0970
stolen_bases, cos2: 0.0918
win_loss_percentage, cos2: 0.0791
caught_stealing, cos2: 0.0740
batters_faced, cos2: 0.0737
hits_9, cos2: 0.0705
strikeouts_x, cos2: 0.0704
strikeouts_y, cos2: 0.0696
RBI, cos2: 0.0593
home_runs_allowed, cos2: 0.0524
at_bats, cos2: 0.0456
saves, cos2: 0.0438
WHIP, cos2: 0.0435
bases_on_balls, cos2: 0.0304
ERA, cos2: 0.0207
doubles, cos2: 0.0088
hits_y, cos2: 0.0070
Cos2 Biplot for Principal Components 3 and 7
PC3 (Cos2 on PC3)
PC7 (Cos2 on PC7)
sacrifice_hits, cos2: 0.5014
SLG, cos2: 0.3679
shutouts, cos2: 0.2971
avg_runs_scored, cos2: 0.2895
average_batter_age, cos2: 0.2055
homeruns_scored, cos2: 0.1298
hit_batter, cos2: 0.1297
innings_pitched, cos2: 0.1121
triples, cos2: 0.1116
stolen_bases, cos2: 0.0862
caught_stealing, cos2: 0.0830
strikeouts_y, cos2: 0.0780
win_loss_percentage, cos2: 0.0772
batters_faced, cos2: 0.0723
strikeouts_x, cos2: 0.0683
GDP, cos2: 0.0598
RBI, cos2: 0.0583
hits_9, cos2: 0.0484
at_bats, cos2: 0.0464
bases_on_balls, cos2: 0.0451
WHIP, cos2: 0.0361
ERA, cos2: 0.0246
saves, cos2: 0.0206
home_runs_allowed, cos2: 0.0194
HBP, cos2: 0.0139
doubles, cos2: 0.0139
hits_y, cos2: 0.0023
Cos2 Biplot for Principal Components 3 and 8
PC3 (Cos2 on PC3)
PC8 (Cos2 on PC8)
shutouts, cos2: 0.4223
SLG, cos2: 0.3711
avg_runs_scored, cos2: 0.2889
GDP, cos2: 0.1643
homeruns_scored, cos2: 0.1298
average_batter_age, cos2: 0.1276
triples, cos2: 0.1225
sacrifice_hits, cos2: 0.1211
innings_pitched, cos2: 0.1168
HBP, cos2: 0.1005
saves, cos2: 0.0852
bases_on_balls, cos2: 0.0825
win_loss_percentage, cos2: 0.0747
batters_faced, cos2: 0.0744
hit_batter, cos2: 0.0665
strikeouts_x, cos2: 0.0635
caught_stealing, cos2: 0.0594
RBI, cos2: 0.0566
stolen_bases, cos2: 0.0557
strikeouts_y, cos2: 0.0541
hits_9, cos2: 0.0515
at_bats, cos2: 0.0463
ERA, cos2: 0.0209
home_runs_allowed, cos2: 0.0197
WHIP, cos2: 0.0173
doubles, cos2: 0.0092
hits_y, cos2: 0.0025
Cos2 Biplot for Principal Components 4 and 5
PC4 (Cos2 on PC4)
PC5 (Cos2 on PC5)
triples, cos2: 0.7811
caught_stealing, cos2: 0.6821
stolen_bases, cos2: 0.6579
sacrifice_hits, cos2: 0.2207
homeruns_scored, cos2: 0.1036
average_batter_age, cos2: 0.0498
SLG, cos2: 0.0436
innings_pitched, cos2: 0.0425
at_bats, cos2: 0.0315
batters_faced, cos2: 0.0298
bases_on_balls, cos2: 0.0280
home_runs_allowed, cos2: 0.0272
HBP, cos2: 0.0252
hits_9, cos2: 0.0228
avg_runs_scored, cos2: 0.0106
saves, cos2: 0.0106
strikeouts_y, cos2: 0.0086
doubles, cos2: 0.0085
win_loss_percentage, cos2: 0.0081
shutouts, cos2: 0.0064
strikeouts_x, cos2: 0.0062
RBI, cos2: 0.0054
hits_y, cos2: 0.0022
WHIP, cos2: 0.0021
hit_batter, cos2: 0.0014
ERA, cos2: 0.0008
GDP, cos2: 0.0006
Cos2 Biplot for Principal Components 4 and 6
PC4 (Cos2 on PC4)
PC6 (Cos2 on PC6)
caught_stealing, cos2: 0.6901
stolen_bases, cos2: 0.6812
sacrifice_hits, cos2: 0.3092
shutouts, cos2: 0.2018
average_batter_age, cos2: 0.1984
hit_batter, cos2: 0.1062
HBP, cos2: 0.0954
GDP, cos2: 0.0777
bases_on_balls, cos2: 0.0415
triples, cos2: 0.0385
saves, cos2: 0.0351
home_runs_allowed, cos2: 0.0346
hits_9, cos2: 0.0294
WHIP, cos2: 0.0285
homeruns_scored, cos2: 0.0262
avg_runs_scored, cos2: 0.0251
strikeouts_y, cos2: 0.0239
innings_pitched, cos2: 0.0179
SLG, cos2: 0.0153
win_loss_percentage, cos2: 0.0120
batters_faced, cos2: 0.0106
strikeouts_x, cos2: 0.0098
hits_y, cos2: 0.0059
RBI, cos2: 0.0049
at_bats, cos2: 0.0019
doubles, cos2: 0.0008
ERA, cos2: 0.0008



The Cos2 biplot captures the intricate relationships among varied baseball strategies in the KBO league. (a) The biplot focuses on two principal components: *'Controlled Approach' (PC1) and 'Scoring Struggles and Late-Game Difficulties' (PC2). (b) 'Controlled Approach' (PC1) and ''Power-Driven Base Strategy and Controlled Aggression' (PC3)*

In the biplot representation, the x-axis, which corresponds to PC1, highlights the defensive mastery that some KBO league teams exhibit. Mainly, the defensive metrics 'ERA' and 'WHIP' emerge as essentials of this theme, showing high cos^2 values of 0.937 and 0.844, respectively, in plot (a). These values suggest that both metrics possess strong magnitudes in the x-direction, indicating their substantial impact on the component emphasizing defensive finesse. The high cos2 values effectively reflect the square of the correlation between the variables and the component, clarifying the degree to which these variables shape the component's nature. In other words, 'ERA' and 'WHIP', with their high cos2 values, play a pivotal role in interpreting the defensive strengths and strategies of teams in the KBO league.

Furthermore, the y-axis delves into the theme of 'Scoring Struggles and Late-Game Difficulties', where the multidimensional influence of 'ERA' and 'WHIP' comes to the shine again. Their strong magnitudes in both x and y directions validate their underlying impact across the two principal components. While their presence along the x-axis articulates the narrative of defensive acumen, along the y-axis, they highlight the in-game challenges teams face. Specifically, elevated values of 'ERA' and 'WHIP' are akin to a two-faced coin: they confirm defensive challenges and resonate with the difficulties teams confront during late-game situations.

Supplementing this narrative, the 'Saves' metric, with its moderate cos^2 value of 0.552, reveals an inclination in the negative y-direction. This signifies the hurdles teams associated with this component contend with in the closing moments of games. Normally, a higher count of saves indicates an ability to seal victories effectively; however, this orientation implies certain teams' vulnerabilities in those pivotal moments.

In moving to plot (b), I see a change in the KBO league strategies. While the y-axis of plot (a) mostly focused on 'Scoring Struggles and Late-Game Difficulties', in plot (b) it shifts to highlight 'Power-Driven Base Strategy and Controlled Aggression', represented by PC3.
The metrics 'ERA' and 'WHIP', which were important in their effect before, now show a clear drop in their cos^2 values, settling at 0.47 and 0.513 respectively. This change points to a reduced importance of these defensive metrics in the context of PC3. The controlled aggression seen in this dimension seems to lessen the impact of 'ERA' and 'WHIP' in team strategies. Basically, as teams focus more on power and controlled aggression, the metrics like 'ERA' and 'WHIP' become less central in shaping the game strategy.

On the other hand, the increased cos^2 values of 'RBI' and 'Homeruns Scored'—0.945 and 0.775, respectively—emphasize the core of the 'Power-Driven Base Strategy'. These metrics show the teams' ability to not only score runs but to do so with strong hitting power. These high values indicate the teams' tendency to rely on powerful hits as a main tactic, showing a clear difference from the earlier focus on defense. In short, plot (b) highlights a change in strategy towards a more offensive approach, where hitting hard becomes key, and some defensive metrics become less important in the overall strategy of the KBO league.

4. Discussion

4.1. Exploring Team Archetypes through PCA: A Biplot Analysis

4.1.1. Biplots as Interpretative Tools

Biplots play a central role in clarifying the correlations between original variables, making them essential for visualizations within principal component analysis (PCA). Drawing inspiration from Gabriel's foundational work on biplots, my analysis leveraged these two-dimensional plots to concurrently represent both samples and variables of the dataset following PCA application.[17] Historically, biplots have been instrumental in illustrating the intricate relationships present in the correlation matrix of original variables, particularly in their interactions with the identified principal components.[18]

Specifically, in the context of my research, cos^2 biplots, as depicted in Figure Panel 4, serve to provide a visual representation of both the samples and variables from the dataset, post its PCA treatment.

1. Direction and Magnitude of Vectors: The orientation of vectors within a biplot offers a n insight into the correlations among the dataset's variables. To elaborate, vectors that align in similar directions suggest a positive correlation. On the flip side, vectors that diverge in opposing directions indicate negative correlations. Moreover, the vector's magnitude or length offers a measure of the strength or intensity of said correlation.
2. Proximity to Axes: A vector's nearness to a principal component axis signifies its relevance or contribution to that particular component. In my analysis, this metric was crucial for comprehending the influence and importance of metrics like WHIP, Home Runs Allowed per 9 Innings, and Strikeouts by Pitchers within our established PC dimensions.
3. Cos^2 Values: Cos^2 values elucidate the quality of a variable's representation on the principal components. When a variable boasts a high cos^2 value, it indicates that the variable is depicted effectively and accurately by the principal component.
4. Measurement and Interpretation of PC Components: To understand the relationship of each variable to the PC component axis, I adopted a nuanced method. By iterating and setting the x-axis as PC1 and toggling the y-axis between PC2 and PC3, I was able to extract the cos^2 values.
    - For example, in plot (a), the variable “At Bats” showcased a cos^2 value of 0.9040. However, when transitioning to plot (b) with the y-axis set to PC3, the cos^2 value for 'At Bats' dropped to 0.9138. This discrepancy reveals that 'At Bats' has a moderate impact on PC1, PC2, and PC3 since the decrease in its cos^2 value was marginal.
    - A contrasting example can be seen in the “ERA” variable. In plot (a), “ERA” displayed a cos^2 value of 0.9367, which dramatically plummeted to 0.4609 in plot (b) when the y-axis was set to PC3. This stark reduction suggests that ERA's relationship with the PC1-PC3 combination is considerably weaker compared to its bond with the PC1-PC2 pair. Essentially, this observation cements the notion that the ERA metric is less accurately represented or influenced by the PC1-PC3 plane in comparison to the PC1-PC2 plane.

[17] K. Ruben Gabriel, "The biplot graphic display of matrices with application to principal component analysis," Biometrika 58 (1971): 454.
[18] J.C. Gower and D.J. Hand, Biplots (Boca Raton: CRC Press, 1996), 23.

4.1.2. Interpretation and Title of PC components

4.1.2.1. KBO PC Component Interpretation

Here is the interpretation and title for PC components until PC6 which holds the most variation of the dataset based on previously set rules:

Component 1: "*Controlled Approach and Defensive Prowess*"

• Offense:

- *Measure Approach*: Metrics such as 'At Bats', 'Doubles', 'Hits', 'Homeruns Scored', 'RBI', and 'SLG' all demonstrate a low negative correlation. This suggests that teams high on Component 1 may not be the most explosive offensively, instead opting for a more controlled approach. The consistent negative correlations across these batting metrics indicate teams that may prioritize placement and strategy over outright power.
- *Strategic yet Unsuccessful Baserunning:* The high positive correlations for 'Sacrifice hits', and 'Stolen bases' underline a team that's strategic in its base-running decisions. 'Triples' further exemplify this team's audacity in capitalizing on gaps in the opponent's defense. However, high positive loading for 'Caught Stealing' indicates failed execution of aggressive baserunning tactics.

• Defense:

- *Pitching Mastery*: Key metrics, including 'ERA', Homeruns Allowed, 'Innings Pitched', and 'WHIP', have a low negative correlation, reflecting dominant pitching performances. Teams excelling in these metrics usually maintain a fortified defense, crucial for reducing the number of earned runs and ensuring an overall superior pitching performance.

Component 2: "*Scoring Struggles and Late-Game Difficulties* "

• Offense:

- *Hitting Hurdles*: Metrics such as 'RBI', 'Hits' show a high negative correlation. These correlations suggest significant challenges in offensive production for teams aligned with this component. They may struggle to both hit for power and get on base consistently.
- *Base-Running Reversals*: The low negative correlations in 'Sacrifice hits' and 'Stolen Bases', paired with the high positive for 'Caught Stealing', signal a potential overemphasis on aggressive base-running, often leading to missed opportunities.

• Defense:

- *Late-Game Difficulties*: The high negative correlations for 'Saves' and 'Shutouts' together depict teams that experience difficulties during critical game moments. Their inability to close out games effectively, as evidenced by the struggles in 'saves', coupled with their rarity in completely dominating the opposition through 'shutouts', showcases a defensive vulnerability especially during pressure-cooker situations.
- *Bold Pitching*: The low positive correlations for metrics like 'Batters Faced', 'Hit Batter', and 'Strikeouts by Pitcher' paint a picture of a team that might take risks on the mound. This could lead to occasional mistakes but can also result in high-reward situations, such as striking out key batters.

Component 3: " *Power-Driven Base Strategy and Controlled Aggression*"

• Offense:

- *Power-Driven Base Strategy*: The high positive values in metrics like 'Homeruns Scored, 'Triples', and 'SLG' highlight the team's powerful hitting approach. Pairing this with the high negative values in 'Sacrifice Hits' and 'Stolen Bases', it seems the team tends to

prioritize powerful hits over conventional base-running techniques. This suggests they may rely less on sacrifices and stolen bases, focusing instead on utilizing their power to drive in runs. When they do engage in base-running, it's likely a calculated risk, leveraging their powerful batting backdrop to distract or pressure the defense, ultimately aiming to capitalize on scoring opportunities.

• *Defense:*

- *Aggressive but Effective*: The low positive values for ‘Batters Faced’ and 'Hit Batter' combined with the high positive value for 'Strikeouts by Pitcher' reveal an aggressive approach. While pitchers might occasionally make mistakes or take some calculated risks leading to more batters faced or hits, they counterbalance this with a strong ability to strike out opponents.

Component 4: "Contact-Focused Offense and Risky Runners"

• *Offense:*

- *High-Stakes Decisions*: A high positive correlation in 'Caught Stealing', paired with low negative values in metrics related to 'Stolen Bases', suggests that teams might be more daring on the bases. They might frequently take risks, leading to both successful steals and getting caught in the act, hinting at an all-or-nothing base-running approach.
- *Contact Prioritization:*
    - *Avoiding the K*: The low 'Strikeouts by Batter' loading indicates batters are adept at making contact and avoiding strikeouts. This shows an offensive strategy that values bat-on-ball skills and keeping the ball in play.

• *Defense:*

- *Frequent Player Engagement: The low positive values for 'GDP' (Grounded into Double Play) and 'WHIP' suggest that while they engage with opposing batters quite frequently, they're moderately successful in pulling off key defensive plays and maintaining a manageable number of runners on base.*

Component 5: "Balanced Plate Discipline and Inconsistent Defense"

- *Offense:*
- *Selective Aggressiveness:* Metrics that exhibit high negative correlations, such as those linked to ‘Homeruns Scored’. However, 'RBI' has low positive correlations, which suggest that teams in alignment with this component are choosy about when they exhibit power. These teams may be waiting for the right pitch to drive, indicating a disciplined approach at the plate.
- *Making the Most of Opportunities:* With metrics like ‘Stolen Bases’ and 'Triples' falling into the low negative and high positive spectrum respectively, it's clear that this team leverages their speed and base-running intelligence. Moreover, low negative correlation on ‘Caught Stealing’ adds strength to the strategic gameplay. They are proficient at taking the extra base when available and understanding when to hold, highlighting an optimal blend of speed and strategy.
- *Defense:*
- *GDP Anomaly:* Unlike other components which tend to exhibit a low negative correlation for Grounded Double Plays (GDP), Component 5 displays a low positive value. This

suggests that teams aligning with this component are slightly more prone to getting their batters out through double plays. Such an occurrence may indicate a defensive style where pitchers aim to induce ground balls in double-play situations, possibly to offset other vulnerabilities or as part of a strategic trade-off. While this approach can be effective, it also comes with the risk of being predictable, especially if opponents are aware of this tendency. The uniqueness of this GDP profile within Component 5 suggests a team that often walks the fine line between strategic genius and potential pitfall.

Component 6: " Safeguarding the Out and Double Play Mastery"
Offense:

- *Safeguarding the Out:* Their strategic prowess is further highlighted by the high negative value in 'Sacrifice Hits'. Instead of sacrificing outs to move runners forward, they utilize their speed and reading of the game. This approach underscores a tactic where they take the risk of stealing bases but prioritize not giving away easy outs, aiming for a balance between aggression and preservation.
- *Audacious Base-Running:* The low positive correlations in metrics such as 'Stolen Bases' and 'Caught Stealing' illustrate a team willing to take assertive chances on the bases.
- *Power Struggles*: A low negative association with 'Home Runs Scored' indicates that, notwithstanding their base-running skill, the team may not consistently demonstrate power hitting. Their batting strategy may focus more on contact and speed rather than sheer hitting power.

Defense:

- *Double Play Mastery:* The high positive loadings for 'GDP' (Grounded into Double Play) suggests that this team's infield is adept at turning potential threats into double plays. This indicates a combination of sharp reflexes and coordinated team play to quickly get two outs, potentially thwarting the momentum of the opposing team.
- *Economic Pitching:* The low negative value for ‘Batters Faced’ combined with 'Innings Pitched' suggests that their pitchers tend to get outs without facing too many batters in each inning. This might be indicative of a pitching strategy that emphasizes efficiency, conserving energy, and minimizing risks.

4.1.2.2. MLB PC Component Interpretation

In the Major League Baseball (MLB) data, my PCA revealed intriguing team dynamics and strategies that mirror some facets of the KBO yet diverge in others. The following components shed light on these distinct aspects:

Component 1: “Slugging Over Singles”

Contrast Note: The KBO's Component 1 interpretation painted a picture of teams valuing a measured offensive approach, highlighting finesse, positioning, and thought-out strategy. On the other hand, the MLB dataset emphasizes a different narrative, indicating an intriguing contrast in offensive maneuvers between the two leagues.
Offense:

1. *Precision at the Plate:* The high negative correlations for 'At Bats' and 'Hits' insinuate that teams in this component are less focused on merely getting the ball into play or accumulating base hits. Instead, their tactics lean towards a discerning approach at the plate. The emphasis is likely on awaiting the right pitch, capitalizing on opportune moments, and making every swing count, even if it translates to fewer overall hits.
2. *Emphasis on Explosive Hits:* A pronounced low negative value for 'Homeruns Scored' bolsters this narrative. These numbers hint that even if there are fewer at-bats and hits, when these teams do strike the ball, it's often with profound impact. The lineup likely comprises batters endowed with the prowess to dispatch the ball beyond the boundary with ease. This is a stark divergence from the KBO's interpretation, where there might be a harmonious blend of power and tactic. Here, the MLB data underscores teams that are significantly invested in the power-hitting dimension.
3. *Maximizing Impact:* This strategy resonates with a clear philosophy - it's less about sheer frequency of reaching the base and more about optimizing the impact once they do. Instead of relentlessly trying to land on base, there's a conspicuous thrust on amplifying scoring avenues whenever the opportunity arises. The overarching goal seems to revolve around efficacy; ensuring each hit, though fewer, holds a substantial impact.

To encapsulate, Component 1 from the MLB dataset sketches a landscape where teams appear to be emphasizing explosive hits, primarily home runs, over the steady accumulation of base hits. This is a fascinating deviation from the KBO's Component 1 and emphasizes the tactical variety inherent in baseball, molded distinctly across varied leagues and contexts.

Component 2: "Controlled Offense with Defensive Dilemmas"

Contrast Note: The previous interpretation emphasized the offensive might and tact of teams, focusing on the power game. However, Component 2 shines a light on the defensive challenges some teams might be facing, highlighting the multifaceted nature of baseball where a strength in one area might coincide with vulnerabilities in another.

Offense:

1. *Restrained Baserunning:* The low negative value for 'Stolen Bases' combined with the low positive 'Caught Stealing' indicates a cautious approach to base-running, as opposed to KBO's more adventurous baserunners.

Defense:

1. *Trouble on the Mound:* High positive values for 'WHIP' and 'ERA' underscore a team's difficulties in pitching. A higher WHIP points to more batters reaching base, be it through hits, walks, or being hit by a pitch. Simultaneously, a high ERA implies these base runners are capitalizing by scoring runs, marking pronounced defensive vulnerabilities.
2. *Power Struggles:* A high positive correlation for 'Homeruns Allowed' showcases the team's susceptibility to the long ball. The data indicates that pitchers are frequently getting taken deep, suggesting potential challenges in pitch selection, delivery, or even overall strategy.
3. *Missed Opportunities:* High negative correlations in 'Shutouts' and 'Saves' suggest that these teams rarely dominate games by completely nullifying the opposition or saving

tight games from the brink of defeat. This further emphasizes the struggles on the defensive side.

Component 3: "Power-Hitting Offense with Discerning Pitching"

Contrast Note: MLB's Component 3 delves into teams that lean heavily on their power-hitting capacities while balancing it with a discerning approach to pitching, emphasizing quality over quantity in terms of innings pitched. This stands in contrast to the KBO's Component 3, which paints a picture of teams that use their powerful hitting to steer their base-running strategies, often leveraging their dominant offensive profile to apply pressure on opposing defenses. Additionally, while KBO's defensive strategy suggests aggressive yet effective pitching, MLB's approach seems to blend raw power from the batters with a more judicious pitching strategy, illuminating the diversity of gameplay strategies between the leagues.

Offense:

1. *Power-centric Approach:* The high positive correlations in 'Homeruns Scored', and 'SLG' unmistakably paint a picture of teams that are geared towards power-hitting. They're not just content with getting on base; they're aiming for the fences. This differs from the KBO's balanced batting strategy which might lack this level of power-hitting emphasis.
2. *Selective Baserunning:* High negative values for 'Triples', 'Stolen Bases', and 'Caught Stealing' depict teams that seem to be selective or cautious in their base-running endeavors. These teams aren't risking much on the base paths, unlike the KBO's aggressive style.
3. *Reduced At-bats but Effective Scoring:* The low negative 'At-Bats' combined with high positive 'Average Runs Scored' suggests efficient scoring, possibly due to the emphasis on home runs over singles or doubles.

Defense:

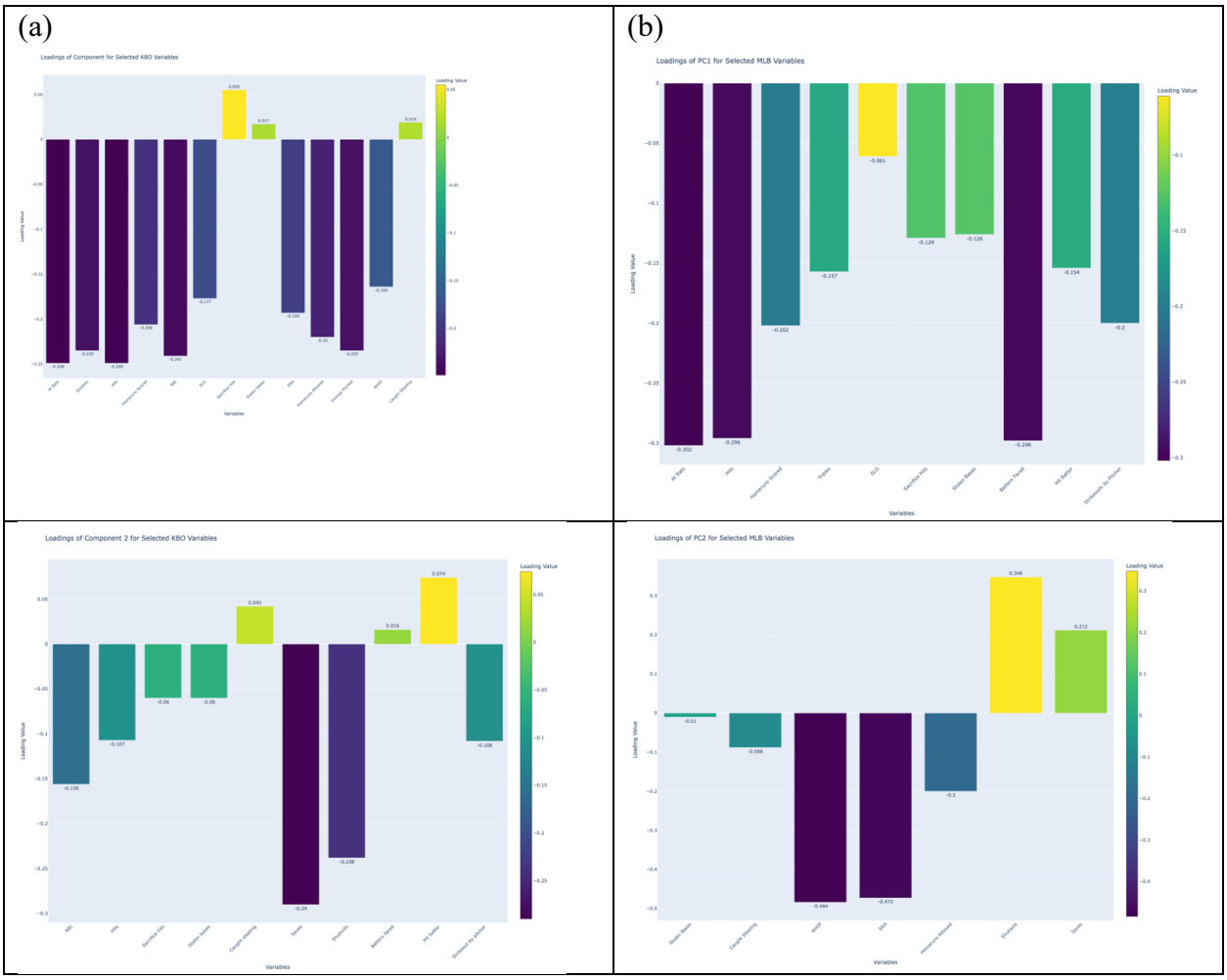
1. *Focused Pitching:* Low negative values in metrics such as 'WHIP', 'ERA', 'Strikeouts by Pitcher', 'Batters Faced', and 'Innings Pitched' highlight a pitching roster that is discerning and perhaps strategic about when they pitch and whom they pitch against. This contrasts with KBO's teams which might see pitchers grinding through more innings regardless of the matchup.
2. *Pitching Control but with Vulnerabilities:* While the low positive 'Hit Batter' and 'Homeruns Allowed' point towards pitchers occasionally losing control, the overall low negative values in 'WHIP' and 'ERA' suggest that these instances are not frequent enough to significantly mar their performance.
3. *Securing Games:* The low positive correlations in 'Shutouts' and 'Saves' indicate teams that can, often, hold onto their lead when they secure one. Unlike KBO's potential roller-coaster finishes, MLB's Component 3 teams are more adept at closing out games.

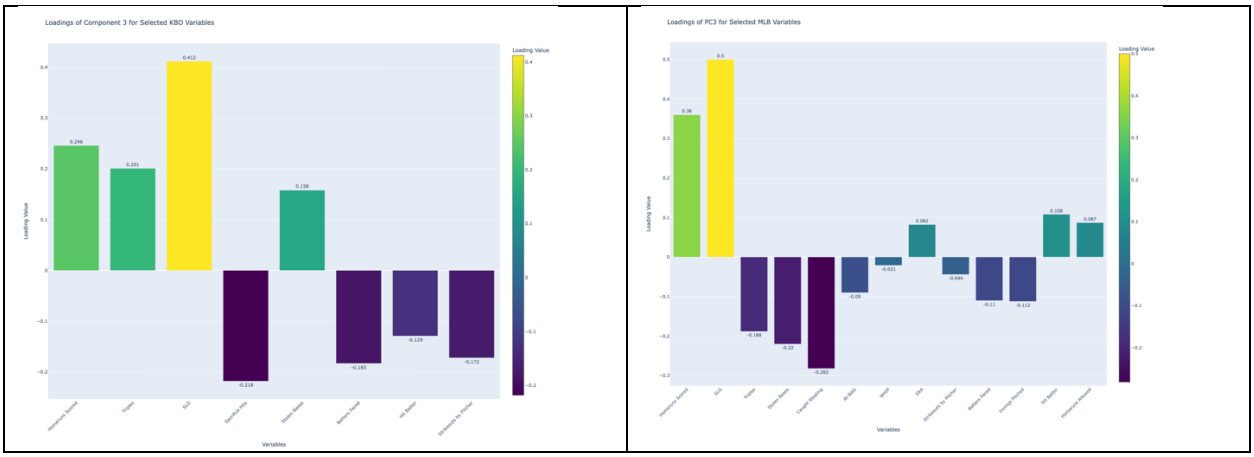
In summary, MLB's Component 3 teams are characterized by their heavy reliance on power hitting while maintaining a pitching lineup that exudes discretion. Unlike KBO's potential balance in batting and bowling, these teams are making their mark through explosive offense and careful, strategic defense.

4.1.3 Figure Panel 4: Histogram Distributions of Principal Component Loadings Across KBO and MLB Leagues

Figure Panel 4 presents histograms detailing the distribution of PC loadings, capturing the essence of significant variables influencing the principal components in both KBO and MLB leagues.

The biplot analysis provided a comprehensive lens to discern the relative significance of each variable. This was achieved by examining the cos^2 values and keenly observing variations when different combinations of PC components were juxtaposed on the x and y axes. To summarize this analysis and visually show how teams in each component represent these eigenvalues, histograms were crafted for each principal component. These histograms underscore the influential variables, offering a consolidated view of their impact across the components.





(a) PC1 captures the essence of strategic yet not always successful baserunning (refer to subsection 4.1.2.1). The variable 'Sacrifice Hits' in plot (a) appears with a positive loading, hinting at a team that deploys a tactical approach to their offense. Sacrifice hits are deliberate plays, typically orchestrated to advance runners even if it means the batter must sacrifice their chance of getting on base. The high value for 'Sacrifice Hits' implies that teams aligning with this component tend to make these sacrificial plays, underscoring a commitment to team progression over individual stats.

On a related note, 'Stolen Bases' holds a positive loading in PC1, indicating a team's aggressive intent on the basepaths. However, the intriguing part lies in the accompanying positive loading for 'Caught Stealing'. Stealing bases, while a showcasing to a team's audacity, also comes with inherent risks. The visibility of 'Caught Stealing' in PC1 reinforces this idea, suggesting that the same teams that frequently attempt to steal bases also find themselves caught in the act more often. It's a reflection of their daring approach: while they are willing to seize every opportunity to advance, it does sometimes lead to them being thrown out, portraying the gamble involved in aggressive baserunning.

Looking at plot (b), Among the metrics represented in the histogram, 'SLG' (Slugging Percentage) stands out with the most positive loading. This emphasizes its significance in this component. A higher SLG indicates a player's ability to achieve bases per at-bat, with a larger weight given to extra-base hits. Essentially, it underscores a batter's power-hitting prowess. The dominant positive loading for SLG in the histogram implies that teams aligning with this component are often producing powerful hits, not just settling for singles but aiming for doubles, triples, or even home runs. The loadings suggest that for some teams, it's not just about how often they get on base but the quality of those base-reaching moments. A lessened emphasis on frequent stats like 'At Bats' and 'Hits', with stronger values for significant actions like ‘Homeruns Scored’ and 'SLG', support this mindset.

Through its variable loadings, the histogram conveys a clear philosophy: effectiveness. Whenever a player from these teams’ steps to the plate, the objective isn't just to make it to base, but to make a significant impact. It's about making sure every offensive action, no matter how few, is felt throughout the game.

4.2. Comparative Analysis on KBO and MLB datasets

A comparative look at the principal components derived from KBO and MLB data reveals intriguing contrasts in team dynamics and playing styles between the two leagues.

Component 1: "Controlled Approach and Defensive Prowess" (KBO) vs "Slugging Over Singles" (MLB)

The KBO teams positioned high on Component 1 underscore a meticulous offensive strategy. They don't rely on outright firepower but emphasize a measured approach to batting, focusing on placement, and well-calculated decisions. This deliberate strategy is evident in their prioritization of doubles and strategy-driven baserunning, even though the latter doesn't always translate to success as indicated by their caught stealing rates.

Contrastingly, MLB teams high on Component 1 depict an inclination towards a high-impact, power-driven batting style. They might not be the most frequent in terms of hits or at-bats, but their strategy is tuned towards maximizing the impact of every successful connection with the ball. These teams likely have batters who patiently wait for the right pitch, aiming to hit it hard. Instead of trying to get on base frequently, they focus on making big, impactful plays with their hits.

When it comes to baserunning, while KBO teams seem to take calculated risks, aiming to exploit gaps in defenses and use strategic plays, MLB teams appear to adopt a more conservative strategy. This could be to prevent the loss of their precious power hitters on the bases.
Defensively, KBO teams exhibit strong prowess, especially in their pitching, a strategy aimed at suppressing opponent scoring opportunities. Their mastery in this area is underlined by their impressive performances in metrics like ERA, WHIP, and Innings Pitched. The MLB counterpart, given their power-hitting emphasis, likely tailors their defensive approach to counteract similar strategies from opposing teams. This involves being particularly cautious with pitch locations and counts, and potentially having outfielders ready for deeper hits.
In essence, while KBO teams in this component emphasize a balance between strategic offensive production and robust defensive mastery, their MLB counterparts seem to accentuate the significance of explosive batting, albeit with fewer base hits. Both leagues, however, demonstrate commendable defensive capabilities, albeit with varied focus and strategies.

In summary, KBO teams under Component 1 favor a well-balanced offensive production while MLB teams specialize in explosive power hitting at key moments despite fewer baserunners. However, both leagues boast strong defensive capabilities from pitching and fielding.

Component 2: "Scoring Hurdles and Late-Game Vulnerabilities" (KBO) vs "Restrained Offense with Pitching Pitfalls" (MLB)

KBO's Component 2 brings to light teams that grapple with offensive challenges. The noticeable negative correlations in metrics like 'Average Runs Scored' and 'Hits' elucidate their offensive conundrums, hinting at difficulties in consistently producing runs and achieving base hits. Their baserunning strategy, although aggressive, appears to backfire often. The heightened 'Caught Stealing' rate juxtaposed with decreased 'Sacrifice hits' and 'Stolen Bases' implies an overaggressive baserunning tactic, leading to more wasted opportunities. Defensively, their most pressing issue seems to revolve around late-game scenarios. Their struggles in 'Saves' combined with a dearth in 'Shutouts' indicate their inability to maintain control in game-deciding moments.

Furthermore, their pitching strategy, while bold and perhaps high reward at times, manifests its downsides in terms of mistakes on the mound and, possibly, overworking their pitchers.

In contrast, *MLB's Component 2* outlines teams that adopt a more restrained offensive approach, especially concerning baserunning. The diminished emphasis on 'Stolen Bases', paired with relatively fewer instances of 'Caught Stealing', suggests that they prefer not to gamble much on the bases. Defensively, their major area of concern is undeniably their pitching. The elevated 'WHIP' and 'ERA' metrics signify a recurrent problem of allowing batters on base and, worse, letting them score. Their susceptibility to power hits, evident from the 'Homeruns Allowed' metric, points towards possible misjudgments in pitch delivery, selection, or even broader strategic flaws. The pronounced negative correlations in 'shutouts' and 'saves' reinforce the narrative of their defensive woes, especially in tight situations or when aiming for game dominance.

In summary, while KBO teams under Component 2 grapple with scoring and late-game defensive challenges, exacerbated by their overaggressive baserunning, MLB teams exhibit a cautious offense and significant pitching concerns, reflecting in their struggles to keep batters at bay and defend tight game situations.

Component 3: "Focused Power Hitting with Strategic Baserunning" (KBO) vs "Explosive Power Plays with Deliberate Pitching" (MLB)

KBO's Component 3 paints a portrait of teams that lean heavily into a power-driven batting strategy. Their enhanced values in metrics like 'Homeruns Scored', 'Triples', and 'SLG' depict a penchant for making big plays with the bat. While they seem to de-emphasize traditional baserunning methods such as 'Sacrifice Hits' and 'Stolen Bases', their strategy leans towards maximizing run-scoring opportunities through power hits. It's a fusion of aggressive hitting with a controlled approach to baserunning. Their defense adopts an aggressive yet effective stance. Their propensity to face more batters or allow hits is counteracted by their considerable ability to strike out the opposition.

Contrarily, *MLB's Component 3* teams are a force of power-hitting prowess, constantly seeking to dispatch the ball over the fence. Metrics like 'Homeruns Scored' and 'SLG' illuminate their hard-hitting mentality. However, they temper this aggressive batting with a more selective approach on the base paths, evident from their restrained values in 'Triples', 'Stolen Bases', and 'Caught Stealing'. While they may not frequently engage in at-bats, their scoring efficiency is commendable, likely due to the predominant emphasis on home runs. On the defensive side, their pitching approach is marked by discernment. They focus on quality over quantity, ensuring that their pitches are deliberate and tailored for the situation. Their occasional lapses, such as allowing hits or home runs, don't derail their overall performance, as evidenced by decent 'WHIP' and 'ERA' metrics. Furthermore, their adeptness at sealing games, demonstrated by their 'shutouts' and 'saves', stands in contrast to KBO's more volatile end-game scenarios.

To encapsulate, KBO teams in this component combine power hitting with strategic base-running, backed by an aggressive yet proficient defensive play. In contrast, MLB teams harness sheer batting power, supplemented by a careful pitching approach, ensuring that they maintain an edge both offensively and defensively. This contrast between the two leagues showcases the intricate tapestry of strategies that can lead to success in the sport of baseball.

5. Conclusion

This study utilized principal component analysis and statistical techniques like thresholding to conduct a comparative analysis of the Korean Baseball Organization and Major League Baseball over multiple decades.

The PCA uncovered dimensions representing key strategic aspects like batting prowess, pitching control, and baserunning. A salient finding that key differences, with KBO teams focused on balanced offensive production and MLB teams specializing in power hitting at critical moments.

The biplot analysis provided intriguing insights into changes in team strategies. As KBO teams shifted focus from defense to power hitting, metrics like ERA and WHIP became less important compared to offensive metrics like HRs and RBIs. This demonstrates how PCA and biplots can reveal nuances in strategic orientations.

By moving beyond existing MLB-centric models, this study provides data-driven strategic archetypes tailored specifically for the KBO league, enabling informed management decisions. Overall, this research highlights the potential of advanced analytical techniques like PCA and biplots for uncovering the intricacies differentiating baseball leagues globally. The findings pave the way for developing enhanced predictive models leveraging the KBO's unique attributes and performance patterns. They also address gaps in literature centered on MLB data, contributing more holistically to baseball and sports analytics.

In summary, by comprehensively analyzing decades of KBO data, this study expands the baseball analytics framework, reveals contrasts with MLB strategies, and emphasizes the global diversity in baseball gameplay and statistics.

6. Appendix

Table 6.1. Threshold-Categorized PCA Loadings for KBO dataset

	Component 1	Component 2	Component 3	Component 4	Component 5	Component 6	Component 7	Component 8
At-Bats	Low Negative	Low Negative	Low Negative	Low Positive	Low Positive	Low Positive	Low Negative	Low Positive
Average Batter Age	Low Negative	Low Negative	High Negative	Low Negative	High Negative	High Positive	High Positive	High Negative
Bases On Balls	Low Negative	Low Negative	Low Negative	Low Positive	Low Positive	Low Positive	Low Positive	High Positive
Batters Faced	Low Negative	Low Positive	Low Negative	Low Positive	Low Positive	Low Negative	Low Negative	Low Positive
Caught Stealing	High Positive	Low Positive	Low Positive	High Positive	Low Negative	Low Positive	High Negative	Low Negative
Doubles	Low Negative	Low Negative	Low Positive	Low Positive	Low Positive	Low Negative	Low Negative	Low Negative
Earned Runs	Low Negative	High Positive	Low Negative	Low Positive	Low Positive	Low Negative	Low Positive	Low Positive
ERA	Low Negative	High Positive	Low Positive	Low Positive	Low Negative	Low Positive	Low Positive	Low Negative
GDP	Low Negative	Low Negative	Low Negative	Low Negative	Low Positive	High Positive	Low Negative	High Positive
HBP	Low Negative	Low Negative	Low Negative	Low Positive	Low Negative	High Negative	Low Negative	High Negative

Hit by Pitch	Low Negative	Low Positive	Low Negative	Low Positive	Low Negative	High Negative	High Negative	Low Negative
Hits	Low Negative	Low Negative	Low Positive	Low Negative	Low Positive	Low Positive	Low Positive	Low Negative
Home Runs Allowed	Low Negative	Low Positive	Low Positive	Low Negative	Low Negative	Low Negative	Low Negative	Low Negative
Homeruns Scored	Low Negative	Low Negative	High Positive	Low Negative	High Negative	Low Negative	Low Negative	Low Positive
Innings Pitched	Low Negative	Low Negative	High Negative	Low Positive	Low Positive	Low Negative	Low Negative	Low Positive
RBI	Low Negative	Low Negative	Low Positive	Low Negative	Low Positive	Low Positive	Low Positive	Low Negative
sacrifice hits	High Positive	Low Negative	High Negative	High Positive	Low Negative	High Negative	High Positive	Low Negative
saves	Low Negative	High Negative	Low Negative	Low Positive	Low Negative	Low Negative	Low Positive	High Positive
shutouts	Low Negative	High Negative	High Negative	Low Negative	Low Positive	High Positive	High Negative	High Negative
SLG	Low Negative	Low Negative	High Positive	Low Negative	Low Negative	Low Negative	Low Positive	Low Negative
Stolen bases	High Positive	Low Negative	Low Positive	High Positive	Low Negative	Low Positive	High Negative	Low Positive
Strikeout by Batter	Low Negative	Low Negative	Low Negative	Low Positive	Low Negative	Low Negative	Low Negative	Low Positive
Strikeout by Pitcher	Low Negative	Low Positive	Low Negative	Low Positive	Low Negative	Low Negative	High Negative	low negative
Triples	High Positive	Low Negative	High Positive	Low Positive	High Positive	Low Negative	Low Positive	high negative
WHIP (Walks Plus Hits Per Inning Pitched)	Low Negative	High Positive	Low Positive	Low Positive	Low Positive	Low Positive	Low Positive	low negative

Table 6.2. Threshold-Categorized PCA Loadings for MLB dataset

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**
At-Bats	High Negative	Low Positive	Low Negative	Low Positive	Low Negative	Low Negative	Low Negative	Low Negative
Average Batter Age	Low Negative	Low Negative	Low Positive	High Negative	High Negative	High Positive	High Negative	High Positive
Bases On Balls	Low Negative	Low Positive	High Positive	High Negative	Low Positive	Low Positive	Low Positive	Low Positive
Batters Faced	Low Negative	Low Negative	Low Positive	Low Negative	Low Negative	Low Positive	Low Positive	Low Positive
Caught Stealing	High Negative	Low Positive	Low Negative	Low Positive	Low Negative	Low Negative	Low Negative	Low Negative
Doubles	Low Negative	Low Positive	High Negative	Low Negative	High Positive	High Positive	Low Negative	Low Negative
Earned Runs	High Negative	Low Positive	Low Positive	Low Negative	Low Negative	High Negative	Low Positive	Low Positive

ERA	High Positive	High Positive	Low Negative	Low Positive	Low Negative	Low Positive	Low Negative	Low Negative
GDP	Low Negative	Low Positive	Low Negative	Low Negative	High Negative	Low Negative	Low Negative	High Negative
HBP	Low Negative	Low Positive	Low Positive	High Positive	Low Negative	High Positive	High Positive	High Positive
Hit by Pitch	Low Negative	Low Negative	Low Positive	Low Positive	Low Negative	High Positive	High Positive	High Negative
Hits	High Negative	Low Positive	Low Negative	Low Negative	Low Negative	Low Negative	Low Negative	Low Negative
Home Runs Allowed	Low Negative	High Positive	Low Positive	High Positive	Low Positive	Low Positive	High Negative	Low Negative
Homeruns Scored	Low Negative	Low Positive	High Positive	Low Positive	Low Positive	Low Positive	Low Negative	Low Negative
Innings Pitched	High Negative	Low Negative	Low Negative	Low Positive	Low Negative	Low Negative	Low Negative	Low Negative
RBI	High Negative	Low Positive	Low Positive	Low Negative	Low Positive	Low Negative	Low Negative	Low Positive
sacrifice hits	Low Negative	Low Positive	High Negative	High Negative	Low Negative	Low Negative	High Positive	Low Negative
saves	Low Negative	High Negative	Low Positive	Low Negative	Low Negative	Low Positive	Low Negative	High Negative
shutouts	Low Negative	High Negative	Low Positive	Low Negative	Low Negative	Low Negative	Low Positive	High Positive
SLG	Low Negative	Low Positive	High Positive	Low Negative	Low Positive	Low Negative	Low Positive	Low Negative
Stolen bases	Low Negative	Low Negative	High Negative	Low Negative	High Positive	High Positive	High Negative	Low Positive
Strikeout by Batter	Low Negative	Low Negative	Low Negative	High Positive	Low Positive	Low Negative	Low Negative	Low Negative
Strikeout by Pitcher	Low Negative	High Negative	Low Positive	High Positive	Low Positive	Low Negative	Low Negative	Low Positive
Triples	Low Negative	Low Positive	High Negative	Low Negative	High Positive	High Negative	High Positive	High Positive
WHIP (Walks Plus Hits Per Inning Pitched)	Low Negative	High Positive	Low Negative	Low Negative	Low Negative	Low Negative	Low Positive	Low Negative

6.3. References

- Bae, Jae Young, Jae Myung Lee, and Jung Yoon Lee. "Predicting Korea Pro-baseball rankings by principal component regression analysis." *Communications for Statistical Applications and Methods* 19, no. 3 (2012): 367-379.
- Bro, Rasmus, and Age K. Smilde. "Principal component analysis." *Analytical Methods* 6, no. 9 (2014): 2812-2831.
- Cadima, Jorge, and Ihaka. "Data Standardization in Principal Component Analysis: One Step Further." *Journal of Data Science* 27, no. 4 (2019): 421-439.

- David, C.C., and D.J. Jacobs. "Principal component analysis: a method for determining the essential dynamics of proteins." *Methods in Molecular Biology* 1084 (2014): 193–226.
- Gower, J.C., and D.J. Hand. *Biplots*. Boca Raton: CRC Press, 1996.
- Huamin Li, et al. "Algorithm 971: An Implementation of a Randomized Algorithm for Principal Component Analysis." *ACM Transactions on Mathematical Software* 43, no. 3 (September 2017): Article 28.
- Jolliffe, Ian T., and Jorge Cadima. "Principal component analysis: a review and recent developments." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, no. 2065 (2016): 20150202.
- Kirschvink, J.L. "The Least-Squares Line and Plane and the Analysis of Palaeomagnetic Data." *Geophysical Journal International* 62, no. 3 (September 1980): 699–718.
- Rojas-Valverde, D., J. Pino-Ortega, C.D. Gómez-Carmona, and M. Rico-González. "A Systematic Review of Methods and Criteria Standard Proposal for the Use of Principal Component Analysis in Team's Sports Science." *International Journal of Environmental Research and Public Health* 17, no. 23 (2020): 8712.
- Viola, I., M. Chen, and T. Isenberg. "Visual Abstraction." In *Foundations of Data Visualization*, edited by M. Chen, H. Hauser, P. Rheingans, and G. Scheuermann, 55-77. Cham: Springer, 2020.
- Wedding, CJ, et al. "Operational Insights into Analysing Team and Player Performance in Elite Rugby League: A Narrative Review with Case Examples." *Sports Medicine - Open* 8, no. 1 (December 2022): 140.
- Wold, S., K. Esbensen, and P. Geladi. "Principal component analysis." *Chemometrics and Intelligent Laboratory Systems* 2, no. 1-3 (1987): 37-52.