STRATEGIC DIVERGENCE IN GLOBAL BASEBALL: A Comparative Analysis of KBO and MLB Game Strategies Through Principal Component Analysis

Junho Eum

Boston University

Commonwealth Ave, Boston, MA 02215

Phone: 305.250.8674

E-mail: jxe282@gmail.com

Major Category: Sports Studies

STRATEGIC DIVERGENCE IN GLOBAL BASEBALL: A Comparative Analysis of KBO and MLB Game Strategies Through Principal Component Analysis

Table of Contents

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. Previous studies have focused on using team statistics to explain league characteristics within a single league (Attarian et al., 2013; McShane et al., 2011). However, there is a gap in research comparing the tactics between baseball leagues in different cultural settings. To bridge this gap, I utilized Principal component analysis (PCA) to uncover the team tactics between the Korean Baseball League (KBO) and Major League Baseball (MLB) teams. Then, I used thresholding to categorize the impact of the variable to principal components for clearer comparison. The result showed that KBO teams under component 1 emphasize balanced offensive production. In contrast, MLB teams focus on high-impact power hits at key moments. However, both leagues demonstrate strong defensive skills. Additionally, KBO teams under component 2 face challenges in scoring and late-game defenses. Meanwhile, MLB teams under component 2 exhibit characteristics like cautious offense, struggles with increased pitching problems, and defense in tense game situations. 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.

Keywords: Baseball, KBO, MLB, PCA, Comparative Analysis, Tactic

1. Literature Review

Baseball spans continents, bringing with it a multitude of strategies and styles that reflect diverse cultural landscapes. The Korean Baseball Organization (KBO) and Major League Baseball (MLB) exemplify the global fascination with baseball. Previous studies have leveraged statistical analyses to extract insights within a single league. Attarian et al. (2013) proposed an MLB pitch classification model using Principal Component Analysis (PCA) for dimensionality reduction to improve efficiency. Similarly, McShane et al. (2011) utilized PCA for feature validation in evaluating the predictive capability of offense metrics on MLB player performance. Expanding beyond MLB, Bae, Lee, and Lee (2012) harnessed PCA to forecast final rankings of teams in the KBO league. Smart and Wolfe (2003) took a qualitative approach in assessing the contribution of leadership and human resources to MLB team performance. Additionally, Panda (2014) combined PCA and penalized linear regression to predict MLB player performance over a career. While these studies have generated meaningful findings within MLB or KBO in isolation, there remains a gap in comparative analysis between the two high-profile leagues to uncover distinct tactical attributes. This study aims to bridge this gap by leveraging PCA to extract unique strategic characteristics from KBO and MLB game statistics from 1982-2021 and 2002-2023, respectively. Contrasting the principal components between the leagues will provide novel insights into the cultural and systemic differences in baseball tactics. This study aims to expand the literature by providing a data-driven comparative analysis of contemporary baseball leagues on the global stage.

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, 23 key variables from both datasets were selected for further analysis. The feature selection process was inspired from the work of Bae, and Lee and Lee (2012, p.369) on predicting Korea Pro-baseball rankings by principal component regression analysis. The study mentioned their approach on selection on variables subject for the reduction of PCA for the purpose of increasing the accuracy of the performance prediction model. Moreover, to capture the performance of teams in MLB context, I incorporate the concept of game state changes proposed by Heaton and Mitra (2021, p.5) to evaluate offensive metrics. This approach defines the outcome of each pitch in terms of changes in ball-strike count base occupancy, number of outs, and score. Features that had significant impact on the game state change were also selected for the comparative analysis. As a result, I composed a comprehensive dataset that captures a team's offensive and defensive performance. This dataset captures:

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.

Defensive Statistics: Central to the pitching metrics are Earned Run Average (ERA), Saves, Innings Pitched, and Homeruns Allowed. Additionally, the statistics highlight the number of Batters Faced, Hit Batters, 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.

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 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 are 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.

2.1. Table 1: Variable Notations and Descriptions

Offensive Features (X1-X12)		Defensive Features (X13-X22)	
Batter Age	The average batter age of a team.	ERA (Earned Run Average)	The average number of earned runs a pitcher allows over nine innings pitched, indicating a pitcher's effectiveness.
At Bats	The number of times a player has been at bat, not including walks, hit by pitch, or sacrifices.	Shutouts	Number of games pitcher does not allow the opposing team to score any runs.
Strikeouts by Batter	The number of times a batter is called out on strikes.	Saves	Awarded to a pitcher who finishes a game for the winning team under certain conditions, typically preserving a lead.
Runs Batted In (RBI)	The number of runs a player has driven in with hits.	Innings Pitched	The cumulative number of innings a pitcher has thrown, with one inning equaling three outs recorded.
Hits	The number of times a batter has hit and reaches the first base	WHIP	Number of batters reaching base per inning.
Doubles	The number of times a batter has hit and reaches the second base	Homeruns Allowed	The count of home runs that a pitcher has conceded to opposing batters.
Triples	The number of times a batter has hit and reaches the third base	Base on Balls	The number of times a pitcher has issued a walk by allowing four pitches outside the strike zone.
Homeruns Scored	The count of homeruns hit by a batter.		
Stolen Bases	The number of bases stolen by a runner.	Strikeouts by Pitcher	The total number of batters a pitcher has retired by strikeout.

Caught Stealing	The number of bases stolen by a runner.	Batters Faced	The total number of batters who have had at least one plate appearance against the pitcher.
SLG (Slugging Percentage)	Measure of player's batting power by calculating the total bases achieved per at-bat, with more weight given to extra-base hits.	Innings Pitched	The total number of innings pitched by the pitcher.
Sacrifice Hits	The number of sacrifice bunts which advance runners while the batter is put out.	Ground Double Play (GDP)	The number of times a pitcher induced a ground ball play that results in two outs being recorded on the play

2.2. Data Collection & Feature Selection

The MLB team data was collected from the Baseball Reference website, which contains comprehensive statistics spanning from 2002-2023. For the KBO league, team statistics from 1982-2021 were gathered from the official KBO website, which houses the official statistical records. Given the interconnected nature of gameplay events in baseball, an initial correlation analysis was conducted between all variables to assess collinearity within the datasets. Variables demonstrating high correlation run the risk of providing redundant information in further analyses. Initial examination showed a strong correlation between the 'Hit Batter' and 'WHIP' (Walks and Hits Per Inning Pitched) statistics. To resolve this collinearity, the 'WHIP' metric was retained for its more comprehensive representation of defensive pitching effectiveness. The finalized sets of variables selected for MLB and KBO contained batting statistics such as 'Home Runs' and 'Stolen Bases' as measures of offensive production, pitching metrics like 'ERA' (Earned Run Average) to quantify defensive success, and team statistics including 'Errors' to represent broader gameplay competencies. Through this feature selection procedure, two refined datasets were constructed containing informative variables that could elucidate strategic variations between the leagues. Limiting collinearity facilitated more accurate data reduction using PCA while retaining variables with tactical relevance strengthened interpretability. This

process was then streamlined datasets tailored to contrast MLB and KBO team strategies through subsequent PCA and thresholding techniques.

2.3. Team Characteristic Identification Using Principal Component Analysis

2.3.1. Data Standardization

Before the application of PCA, I standardized the dataset to minimize the impact of different scale in the feature to the result of the PCA.

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

2.3.2. 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 S_{ij} represents the covariance between feature i and feature j, and the diagonal elements S_{ii} represent the variance of feature i (Jorge & Ihaka, 2019).

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

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

2.3.3. Eigen Decomposition

Proceeding with the PCA, the next 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. 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 (Kirschvink, 1980). 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 1.1. The eigenvalues (λ) indicate the amount of variance captured by each principal component showing the importance of each principal component in representing the data. 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 (Abdi & Williams, 2010, p.438).

2.3.4. 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 (Kirschvink, 1980, p.705). Specifically, multiplying the standardized data matrix by A results in matrix Z. This matrix encapsulates the data in terms of its principal components, offering a clearer representation of its inherent structure (Wold et al., 1987, p.40). This transformation not only makes the data more interpretable but also often achieves dimensionality reduction without significant loss of information.

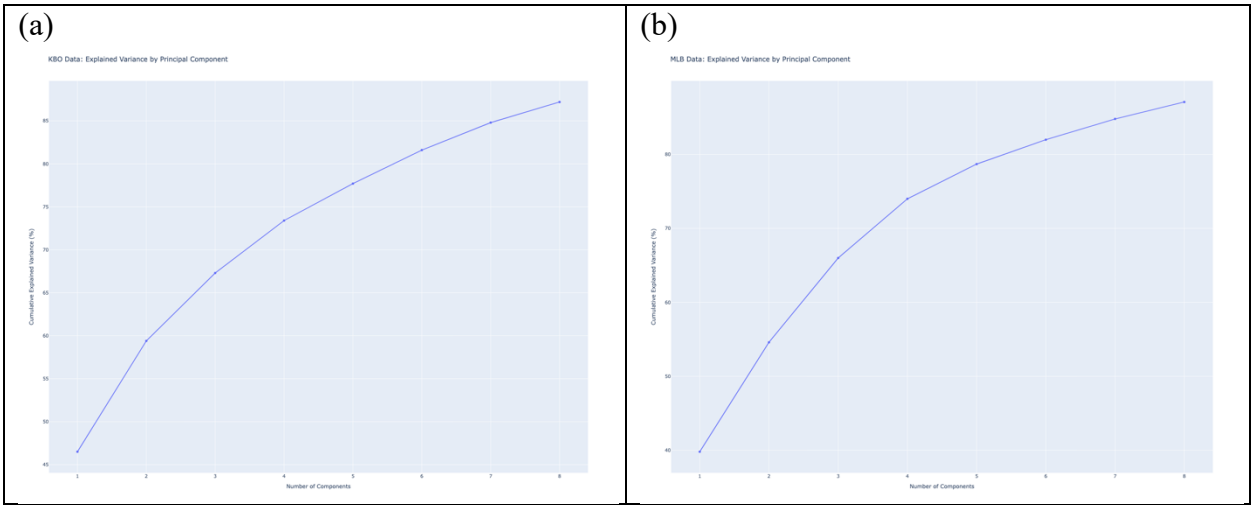
$$Z = P_{std}A$$

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 (Viola et al., 2020, p.55).

2.3.4. Interpreting Principal Components

After identifying the principal components, they are assessed based on the explained variance 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.

2.4. Figure Panel 2: Scree plot and Table for Explained Variance for Principal Components



Screeplot of explained variance ratio of principal components on (a) KBO and (b) MLB

Principal Component	Cumulative Explained Variance KBO (%)	Additional Explained Variance KBO (%)	Cumulative Explained Variance MLB (%)	Additional Explained Variance MLB (%)
1	48.9	48.9	41.3	41.3
2	57.9	9.0	53.6	12.3
3	65.9	8.0	65.4	11.8
4	72.4	6.5	73.7	8.3
5	77.0	4.6	78.5	4.8
6	81.1	4.1	82.0	3.5
7	84.5	3.4	84.8	2.8
8	87.1	2.6	87.2	2.4

Table of explained variance of principal components from 1-8

(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 (David & Jacobs, 2014, p.197). The scree plot for the KBO data shows the explained variance for each principal component. The first component accounts for 48.9% of the total variance. This is followed by the second component, explaining an additional 9% variance. 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 41.3% variance. The second component accounts for an additional 12.3% variance. 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 65.4% of the total variance 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.

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 data. Wedding et al.'s examination of team and player performance in elite rugby league adopted a methodology that systematically interprets performance indicators based on their factor loadings. In their study, indicators with values greater than 0.60 from the rotated component matrix were considered significant contributors to identified playing styles (CJ Wedding et al., 2022, p.140). 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. 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 to understand 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 $x_{ij} < 0$)

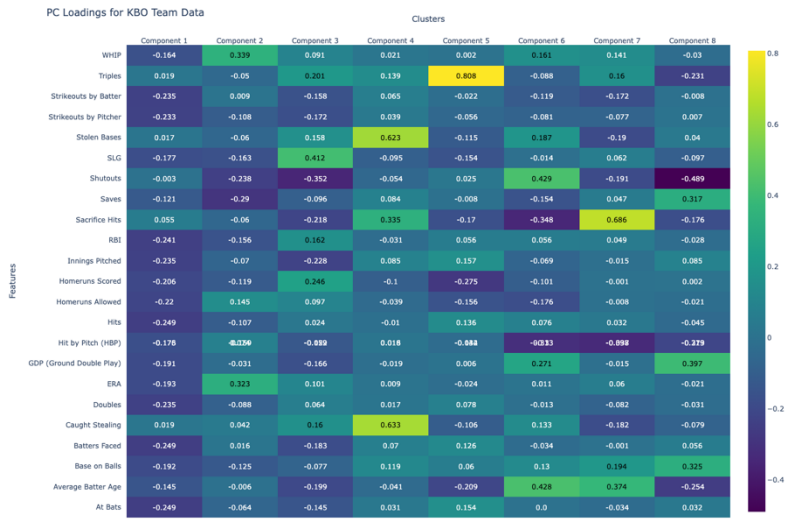
Low Negative: Loadings that are less than one standard deviation above the mean, negative, and higher than one standard deviation below the mean are classified as low negative ($\mu_i - \sigma_i < x_{ij} < \mu_i + \sigma_i$ and $x_{ij} < 0$)

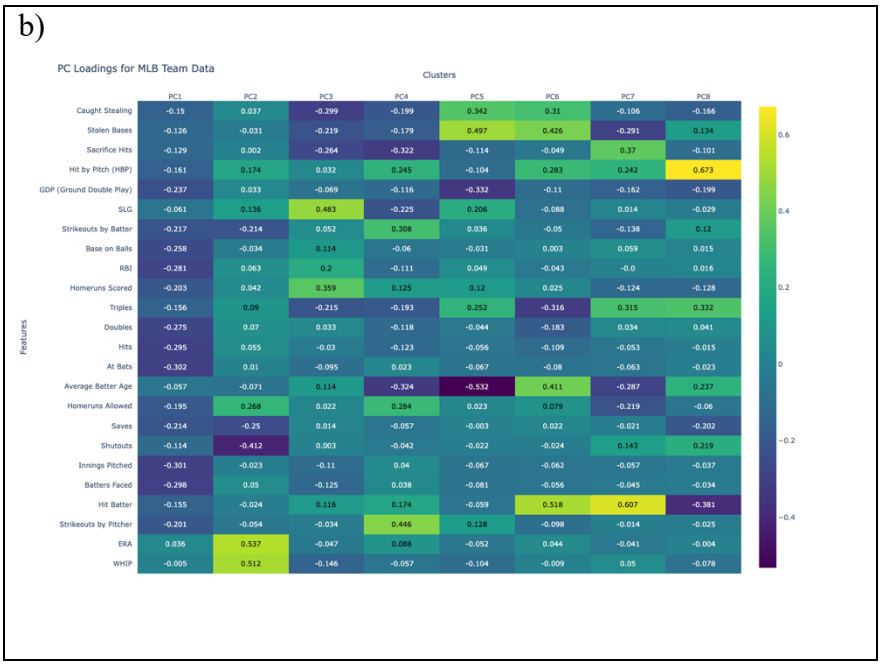
3.1.2. Detailed Explanation of Key Performance Variables

To enable a detailed discussion on the strategic orientation of the teams in the KBO league based on various performance metrics, I identify variables within components that show 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 heatmaps on Figure Panel 3 with the loadings associated with each variable across different clusters.

3.1.3. Figure Panel 3: Heatmap Representation of PC loadings

a)





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

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

Based on the heatmap representation in Figure Panel 3 and Table 6.1, 6.2. I will explain the implication behind the numeric loading features within the principal components and the categorized label from thresholding. This implication will later be used when analyzing the components using the distribution of these variables.

WHIP (Walks plus Hits per Innings Pitched). Component 2 (0.339): High Positive Loading - 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. Component 1 (-0.164): Low Negative Loading

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.

ERA (Walks plus Hits per Innings Pitched). 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.

Average Batter Age. 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.

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. However, potential challenges for such teams might include issues related to physical fitness, agility, or speed when compared to younger players.

Caught Stealing. 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. 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.

GDP (Ground Double Play). Component 6: High Positive Loading (0.271) - Teams under component 6 exhibit high positive loadings for GDP. 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.

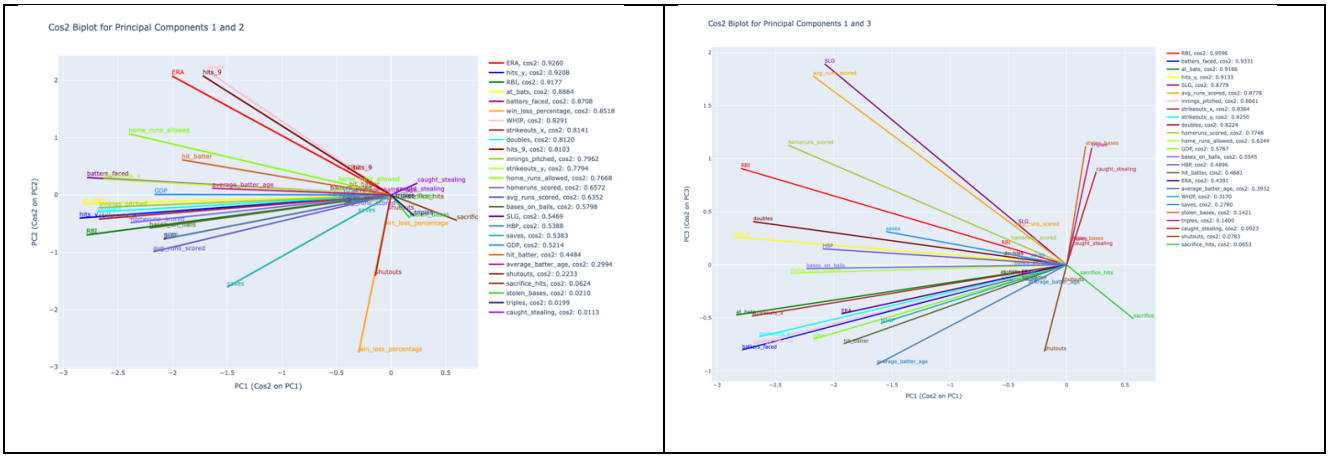
Homeruns Scored. Component 3: High Positive Loading (0.246) - On the flip side, teams within Component 3 demonstrate a high positive loading. This indicates an above-average homerun-scoring capability. Teams in this bracket likely possess strong power hitters who can capitalize on pitcher mistakes and deliver game-changing plays.

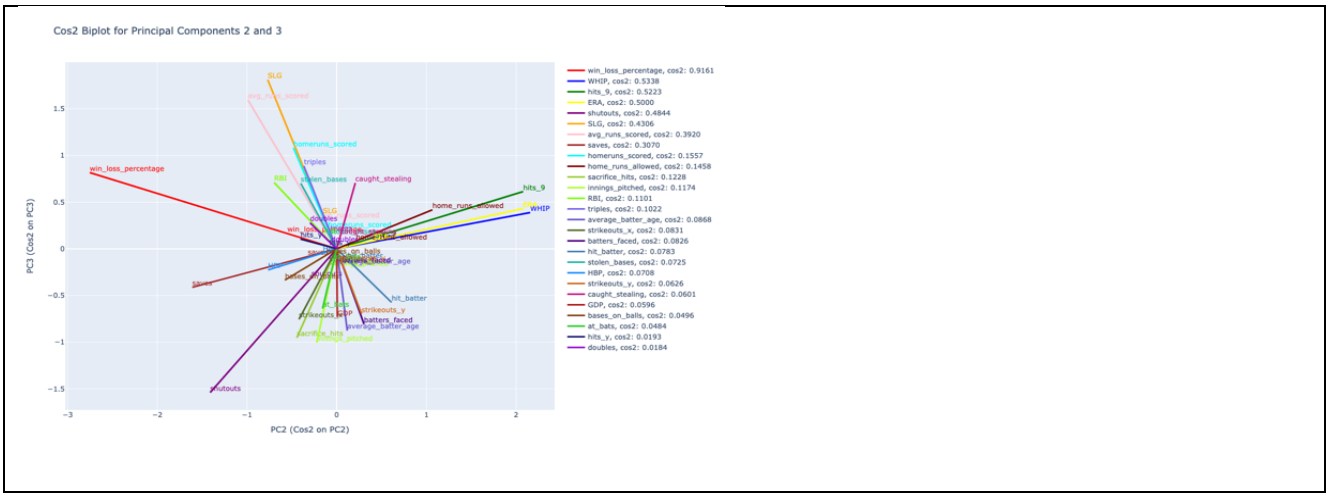
Component 5: High Negative Loading (-0.275) - Teams in Component 5 exhibit a high negative loading, suggesting they score significantly fewer homeruns compared to most teams in the league. Such a pronounced low score indicates potential areas of concern in the strength aspect of the team's offensive lineup.

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.

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 often bet 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 indicates a team with more power hitters or one that's confident in its ability to string together hits.

3.1.4. Figure Panel 4: Comparative Biplots Highlighting Principal Strategies in 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)

3.1.5. Biplot Analysis

Biplots play a central role in clarifying the correlations between original variables, making them essential for visualizations within principal component analysis. My analysis leveraged these two-dimensional plots to concurrently represent both samples and variables of the dataset following the PCA application. 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 after its PCA treatment.

Direction and Magnitude of Vectors: The orientation of vectors within a biplot offers an 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.

Proximity to Axes: A vector's nearness to a principal component axis signifies its relevance or contribution to that 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.

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.

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 crucial 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 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. Korean Baseball League Principal Components

By aggregating my findings from the biplot analysis and the categorization of numeric loading on each principal components, I derived the strategical characteristics and a brief title of Korean Baseball League teams. The interpretation was divided into offensive and defensive categories.

Component 1: "Controlled Approach and Defensive Prowess" - Measure Approach (Offense): Metrics such as 'At Bats', 'Doubles', 'Hits', 'Homeruns Scored', 'RBI', and 'SLG' all show 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 (Offense): 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.

Pitching Mastery (Defense): 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 " - Hitting Hurdles (Offense): 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 (Offense): 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.

Late-Game Difficulties (Defense): 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" - Power-Driven Base Strategy (Offense): 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.

Aggressive but Effective (Defense): 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.

4.2. Major League Baseball Principal Components

In the Major League Baseball (MLB) data, my PCA revealed 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” - Precision at the Plate (Offense): 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. Emphasis on Explosive Hits (Offense): 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.

Maximizing Impact (Offense): 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.

Component 2: "Controlled Offense with Defensive Dilemmas" - Restrained Baserunning (Offense): 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.

Trouble on the Mound (Defense): 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.

Power Struggles (Defense): 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.

Missed Opportunities (Defense): 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" - Power-centric Approach (Offense): 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 lack this level of power-hitting emphasis.

Selective Baserunning (Offense): High negative values for 'Triples', 'Stolen Bases', and 'Caught Stealing' depict teams that seem to be selective or cautious in their baserunning endeavors. These teams aren't risking much on the base paths, unlike the KBO's aggressive style.

Reduced At-bats but Effective Scoring (Offense): 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.

Focused Pitching (Defense): 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 see pitchers grinding through more innings regardless of the matchup.

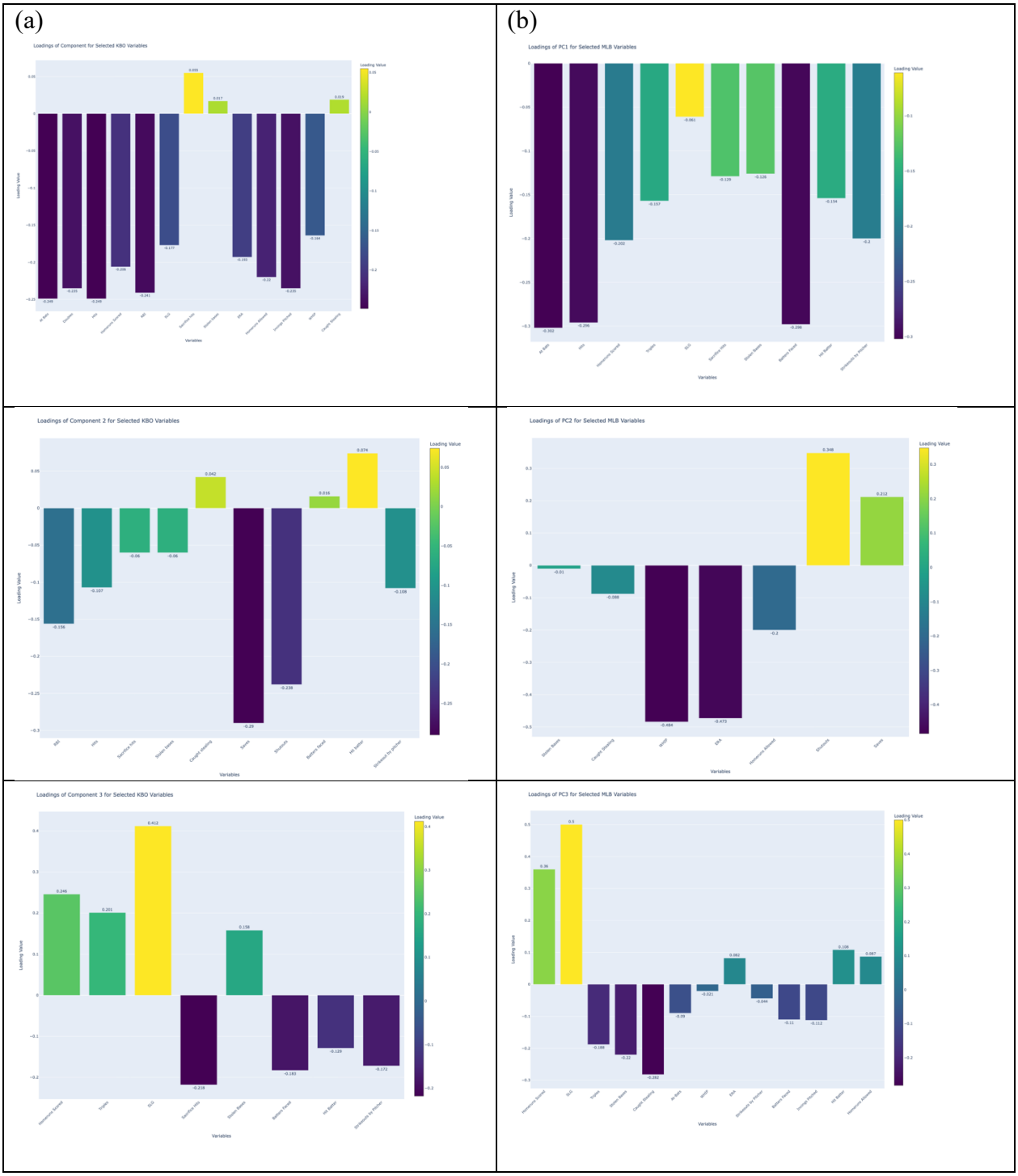
Pitching Control but with Vulnerabilities (Defense): While the low positive 'Hit Batter' and 'Homeruns Allowed' point towards pitchers occasionally losing control, the overall low negative values in 'WHIP' and 'ERA' suggest that these instances are not frequent enough to significantly mar their performance.

Securing Games (Defense): 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.

4.3. 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, showing significant variables influencing the principal components in both KBO and MLB leagues. To summarize this analysis and to visually show how teams in each component represent these loadings,

histograms were crafted for each principal component. These histograms underscore the influential variables, offering a consolidated view of their impact across the components.



(a) Subsection 4.1 analyzed that PC1 captures the team strategy of strategic yet not always successful baserunning. 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.

4.4. 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.

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.

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

5. Conclusion

This comparative analysis of the Korean Baseball Organization (KBO) and Major League Baseball (MLB) leveraged Principal Component Analysis (PCA) and statistical techniques like thresholding to uncover strategic differences between the leagues over the past four decades. The PCA revealed dimensions representing critical tactical aspects, including batting prowess, pitching control, and baserunning approaches. A key finding was that KBO teams emphasize balanced offensive production across all hitters, while MLB teams specialize in power-hitting at critical moments to maximize run potential. Additionally, the analysis uncovered contrasting pitching and baserunning philosophies – KBO teams take more risks on the basepaths but combine it with pitching skills to suppress opponent run potential. Meanwhile, MLB teams exhibit more patient batting with defensive vulnerabilities in high-pressure situations. The study also utilized biplots to showcase evolving team tactics. As KBO teams on component 2 shifted focus from defense to power hitting, metrics like ERA and WHIP decreased in precedence compared to offensive metrics like HRs and RBIs. By leveraging PCA to extract tactical insights from KBO and MLB statistics, this research expands the baseball analytics knowledge base and provides a framework for contrasting contemporary baseball leagues.

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

References

Attarian, A., Danis, G., Gronsbell, J., Iervolino, G., Layne, L., Padgett, D., & Tran, H. (2013). Baseball pitch classification: A Bayesian method and dimension reduction investigation. In *IAENG Transactions on Engineering Sciences: Special Issue of the International MultiConference of Engineers and Computer Scientists 2013 and World Congress on Engineering* (pp. 393).

McShane, B. B., Braunstein, A., Piette, J., & Jensen, S. T. (2011). A hierarchical Bayesian variable selection approach to Major League Baseball hitting metrics. *Journal of Quantitative Analysis in Sports*, 7(4), Article 2. https://doi.org/10.2202/1559-0410.1323

Bae, Jae Young, Jae Myung Lee, and Jung Yoon Lee. (2012). Predicting Korea Pro-baseball rankings by principal component regression analysis. *Communications for Statistical Applications and Methods*, 19(3), 367-379. https://doi.org/10.1016/j.heliyon.2023.e23231.

Panda, M. L. (2014). Penalized regression models for major league baseball metrics (Doctoral dissertation, University of Georgia).

Heaton, C., & Mitra, P. (2021, September). Learning to describe player form in the MLB. In *International Workshop on Machine Learning and Data Mining for Sports Analytics* (pp. 93-102). Cham: Springer International Publishing. https://doi.org/10.48550/arXiv.2109.05280.

Bro, Rasmus, and Age K. Smilde. (2014). Principal component analysis. *Analytical Methods*, 6(9), 2812-2831. https://doi.org/10.1039/C3AY41907J.

Cadima, Jorge, & Ihaka. (2019). Data Standardization in Principal Component Analysis: One Step Further. *Journal of Data Science*, 27(4), 421-439.

David, C. C., & Jacobs, D. J. (2014). Principal component analysis: a method for determining the essential dynamics of proteins. *Methods in Molecular Biology*, 1084, 193–226.

Gower, J. C., & Hand, D. J. (1996). *Biplots*. Boca Raton: CRC Press.

Huamin, L., et al. (2017). Algorithm 971: An implementation of a randomized algorithm for principal component analysis. *ACM Transactions on Mathematical Software*, 43(3), Article 28.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202.

Kirschvink, J. L. (1980). The least-squares line and plane and the analysis of palaeomagnetic data. *Geophysical Journal International*, 62(3), 699–718.

Rojas-Valverde, D., Pino-Ortega, J., Gómez-Carmona, C. D., & Rico-González, M. (2020). 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(23), 8712. https://doi.org/10.3390/ijerph17238712.

Viola, I., Chen, M., & Isenberg, T. (2020). Visual abstraction. In M. Chen, H. Hauser, P. Rheingans, & G. Scheuermann (Eds.), *Foundations of Data Visualization* (pp. 55-77). Cham: Springer.

Wedding, C. J., et al. (2022). Operational insights into analysing team and player performance in elite rugby league: A narrative review with case examples. *Sports Medicine - Open*, 8(1), 140. https://doi.org/10.1186/s40798-022-00408-1.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3), 37-52.