

## Table of Contents

<i>Abstract</i> .....	1
<b>1. Introduction</b> .....	2
<b>1.1. Current Study Objectives</b> .....	3
<b>2. Data and Methods</b> .....	5
2.1. Figure Panel 1: Top Correlated Features Visual Representation / Variable Notations and Descriptions	
6	
<b>2.2. Team Characteristic Identification using Principal Component Analysis</b> .....	6
2.2.1. Data Standardization .....	7
2.2.2. Computing the Covariance Matrix .....	8
2.2.3. Eigen Decomposition .....	8
2.2.4. Finding Principal Components .....	8
2.2.5. Interpreting Principal Components .....	9
<b>2.3. Figure Panel 2: Scree plot for Explained Variance for each Principal Component</b> .....	9
<b>3. Result</b> .....	10
<b>3.1. Establishing Baseline Rules for PC Component Interpretation</b> .....	10
3.1.1. Detailed Explanation of Key Performance Variables .....	10
3.1.2. Figure Panel 3: Heatmap Representation of PC loadings .....	14
3.1.3. Figure Panel 4: Comparative Biplots Highlighting Principal Strategies in KBO League .....	15
<b>4. Discussion</b> .....	19
<b>4.1. Exploring Team Archetypes through PCA: A Biplot Analysis</b> .....	19
4.1.1. Biplots as Interpretative Tools .....	20
4.1.2. Interpretation and Title of PC components .....	20
4.1.3. Figure Panel 4: Histogram Distributions of Principal Component Loadings Across KBO and MLB Leagues .....	25
<b>4.2. Comparative Analysis on KBO and MLB datasets</b> .....	27
<b>5. Conclusion</b> .....	29
<b>6. Appendix</b> .....	30
Table 6.1. Threshold-Categorized PCA Loadings for KBO dataset .....	30
Table 6.2. Threshold-Categorized PCA Loadings for MLB dataset .....	31

## 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 shedding light on 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 is a globally celebrated sport that has been subject to a wide array of analyses, encompassing aspects such as player performance, game strategies, and team dynamics. A prominent tool in this domain is the Pythagorean Expectation, formulated by Bill James, which is pivotal in predicting a team's seasonal wins using the runs scored and allowed by a team as indicators of its performance.

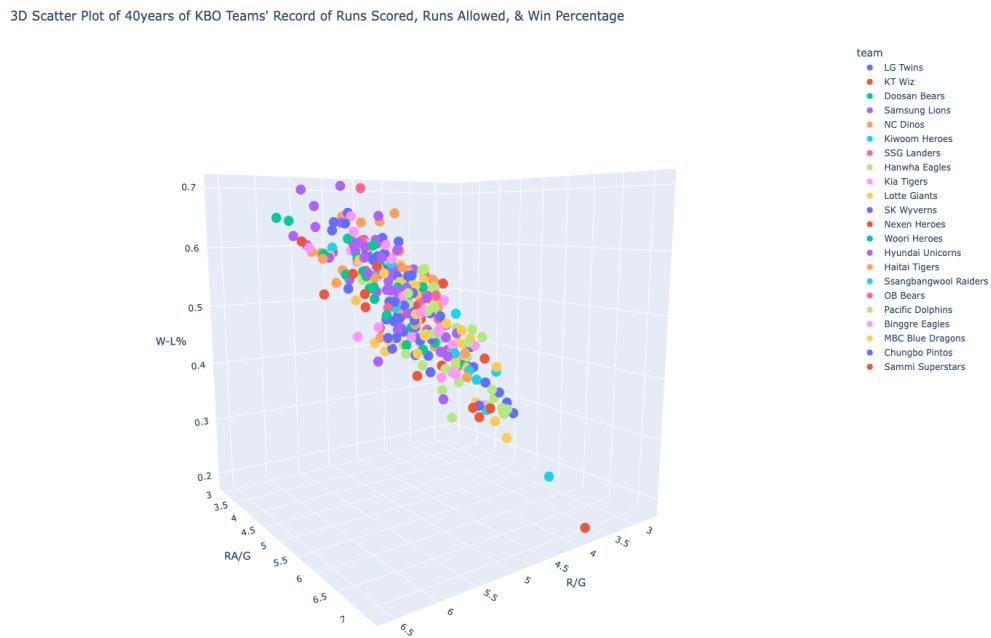
In addition to the Pythagorean Expectation, other analytical models have been developed, such as the Markov chain approach by Bruce Bukiet and Elliotte Harold, which predicts seasonal wins for baseball teams using batter transition matrices and run distributions (Bukiet & Harold, 1997). Moreover, adaptations like the Pythagagenport formula by Clay Davenport and David Smyth have been introduced to address the variances observed when applying the original Pythagorean Expectation to actual game data (Davenport & Woolner, 1999; Smyth, 2016). These adaptations highlight the multifaceted nature of baseball analytics and the importance of considering a wide array of variables and perspectives when assessing and predicting team success.

Although significant strides have been made in baseball analytics, most of the statistical analyses have been based on Major League Baseball (MLB) data. For example, Steven Miller used the 2004 American League data to determine the optimal exponent for predicting expected wins, but this was not specifically for Korean Baseball Organization (KBO) teams (Miller, 2007). This approach resulted in identifying 1.74 as the best fit exponent for MLB, closely aligning with the observation that 1.82 is the optimal exponent for the Pythagorean formula (Miller, 2007).

### 1.1. Current Study Objectives

The primary aim of this study is to find distinctive aspects in Korean Baseball League by reducing dimensionality of the team features, thus broadening the analytical perspective beyond existing frameworks. While the widely accepted Pythagorean Expectation suggests a linear relationship between a team's win-loss ratio and the ratio of their runs scored to the total squares of runs scored and allowed, a 3D plot of KBO data exhibits notable outliers that do not fully align with this expectation (Figure 1). These outliers could represent teams with exceptional clutch performances or outstanding pitching abilities, or those whose actual win-loss ratio drastically diverges from the predicted ratio due to unpredictable elements or sheer randomness. This research endeavors to decode the internal structure of KBO teams by analyzing the results of a principal component analysis (PCA), aiming to pinpoint the unique statistical attributes of the KBO league. In the end, this study aims to offer a more well-rounded and detailed analysis of team performance that goes beyond the basic principles of the Pythagorean Expectation.

(a)



(b)

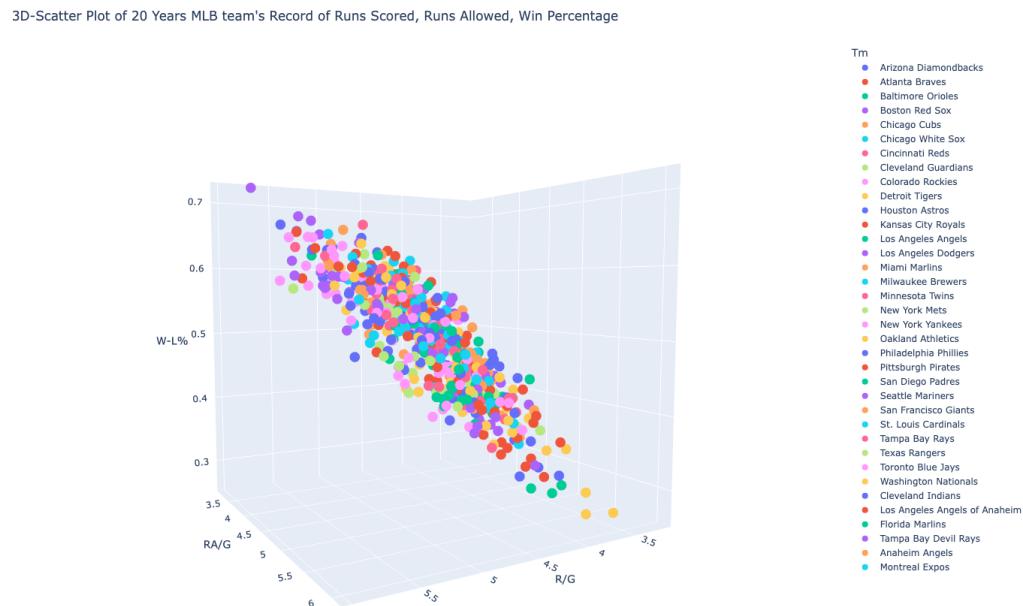


Figure 1. 3D plot of Runs Scored, Runs Allowed, Win Ratio of (a) KBO (b) MLB

## 2. Data and Methods

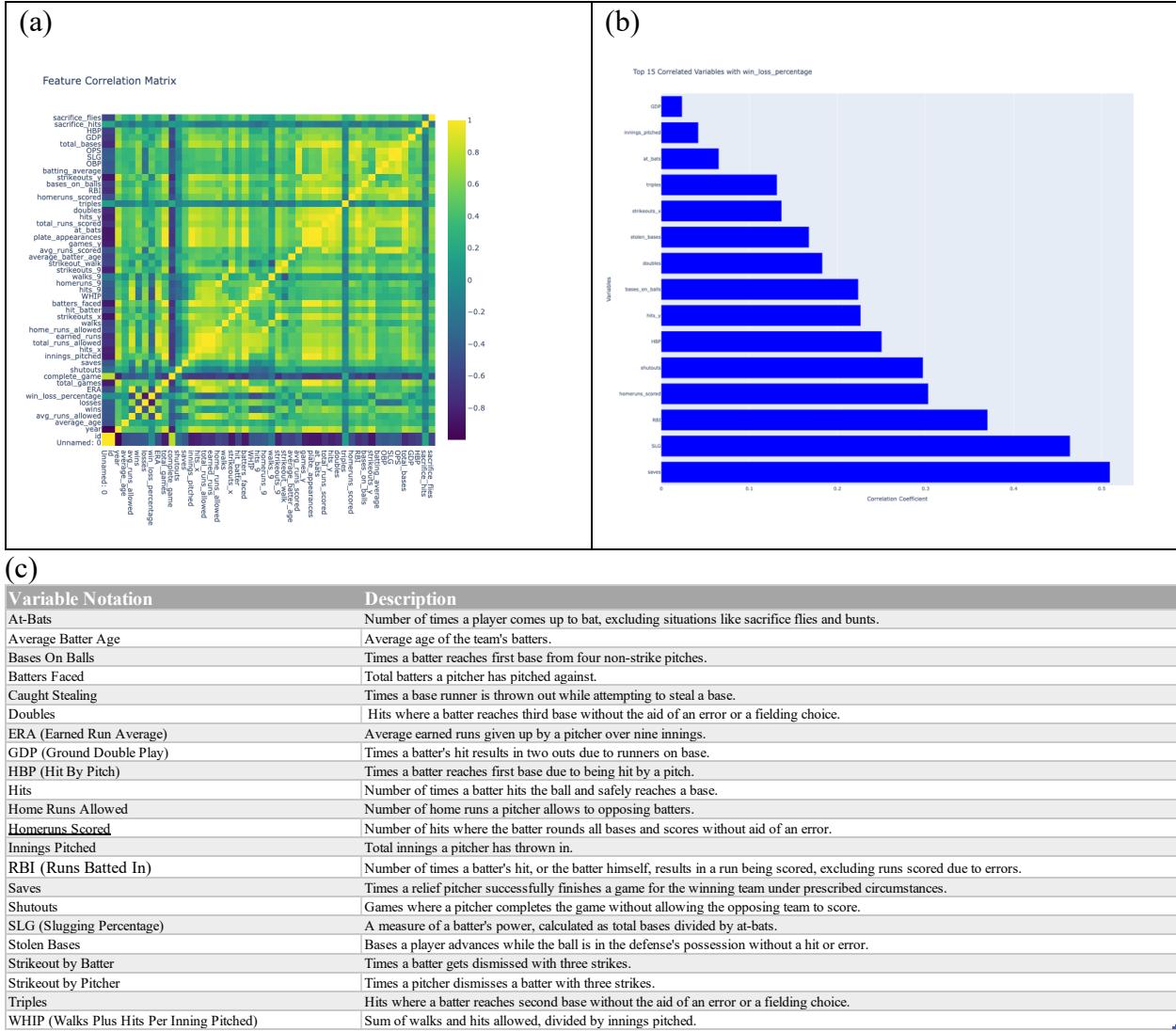
In this study, I utilized KBO league's team statistic from 1982 to 2021 to investigate statistical characteristics in the Korean Baseball Organization (KBO) playstyle. Moreover, for comparative analysis on both leagues comprehensive dataset encompassing three decades of MLB team data was sourced from Baseball Reference, specifically spanning the years 2002 to 2023. The dataset underwent preliminary processing, where variables were scrutinized through correlation analysis. This step was essential to filter out redundant or less impactful variables, leaving behind 28 crucial variables primed for deeper analysis for both datasets.

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 home runs allowed. Additionally, the statistics shed light on the number of batters faced, earned runs, hit batters (HBP), 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 accentuated 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, home runs 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 in a heatmap (Figure 2) and a bar chart (Figure 3) using R and Python.

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



(a) Correlation heatmap for dataset parameters (b) Specified bar chart for parameter selection (c)Notations 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 (2012) 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. Such an

initial step was pivotal in preparing the dataset for PCA's detailed analysis. In the modified dataset, now termed matrix  $P$ , variables were selected judiciously. This choice was made to capture essential team attributes, filtering out potential noise. By emphasizing the retention of significant variables that majorly contribute to the uncovering of core patterns and omitting redundant ones, a refined dataset was created, denoted as matrix  $P$  with dimensions (323, 28).

### 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's rigorous analysis. 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 we 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, 28).

$$P = \begin{bmatrix} p_{11} & p_{12} & \dots & p_{1,28} \\ p_{21} & p_{22} & \dots & p_{2,28} \\ \vdots & \vdots & \ddots & \vdots \\ p_{323,1} & p_{323,2} & \dots & p_{323,28} \end{bmatrix}$$

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. This standardization step holds significant importance as PCA is inherently sensitive to the scales of the different variables in play. Discrepancies in the scales can markedly influence the outcomes of PCA, thereby skewing the interpretations of the principal components.

Standardization is carried out using the formula:

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

Here,

- $P$  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 our 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.

### 2.2.2. Computing the Covariance Matrix

The covariance matrix, denoted as  $S$ , is computed using the standardized data matrix  $P_{\text{std}}$ . This matrix encapsulates the pairwise relationships between different variables in the dataset, providing a mathematical representation of the dataset's variance and covariance. 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:

$$S = \frac{1}{322} P_{\text{std}}^T P_{\text{std}}$$

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

### 2.2.3. Eigen Decomposition

Next, we proceed to find the eigenvalues and eigenvectors of the covariance matrix, a pivotal step in identifying the principal components. These can be obtained using established numerical algorithms such as the power iteration method or more commonly, the Singular Value Decomposition (SVD) approach. Though a deep dive into these algorithms is beyond the scope of this subsection, it is notable that these methods are optimized to find the best possible representation of the data in fewer dimensions.

The eigenvalues and eigenvectors satisfy the following characteristic equation:

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

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

In this equation, the eigenvalues  $(\lambda)$  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.

On the other hand, eigenvectors provide the coefficients of the linear combinations of the original variables to obtain the principal components. Importantly, these eigenvectors are orthogonal to each other, implying that the principal components are uncorrelated, and hence, each component brings forth new, unique information, facilitating the reduction of dimensionality without significant loss of information.

### 2.2.4. Finding Principal Components

The principal components are derived using the eigenvalues and eigenvectors obtained from the eigen decomposition. The eigenvectors form a transformation matrix ( $A$ ), and we find the matrix of principal components, denoted as ( $Z$ ), through the relation:

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

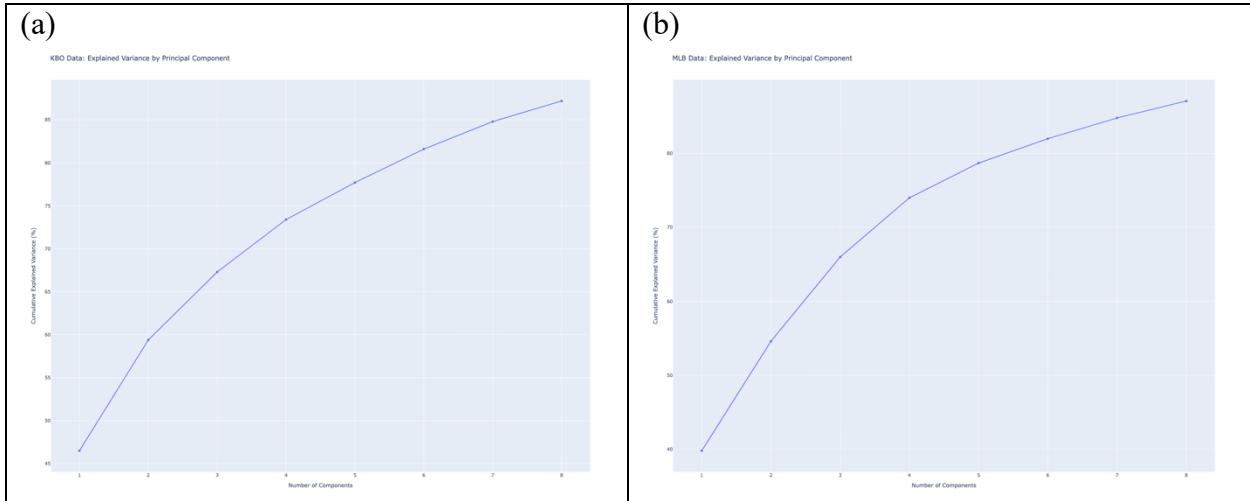
where ( $A$ ) is the matrix of eigenvectors, and ( $P_{\text{std}}$ ) is the standardized data matrix.

### 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 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. Based on the "elbow" of the scree plot, the first 3 principal components collectively capture around 67% of the total variance. (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 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.

**Utilizing rules set for the variables - WHIP, ERA, strikeouts by pitchers, hit batters, win-loss percentage, home runs allowed, and total runs allowed - we aim to provide a grounded analysis backed by data-driven insights.**

##### 3.1.1. Detailed Explanation of Key Performance Variables

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

Below, we dissect each variable, delving into what the different loadings in various clusters imply, guided visually through a comprehensive heatmap and representation that portrays the eigenvalues associated with each variable across different clusters (See Figure 6):

###### 3.1.1.1. Thresholding

To categorize the PC loadings derived from the KBO data, I utilized a thresholding strategy that is rooted in the distribution of each component's values. This strategy seeks to consider not only the magnitude of each loading but also its sign, which is vital to interpreting the data accurately. Firstly, I computed the mean ( $\mu_i$ ) and standard deviation ( $\sigma_i$ ) for each component individually, facilitating a nuanced understanding of the distribution of loadings in each component.

Next, we 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  $x_{ij} < 0$ )

### 3.1.1.2. 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.

### 3.1.1.3. 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.

### 3.1.1.4. 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 seasoned 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.

#### *3.1.1.5. 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. To improve, these teams might consider investing in base running drills, enhancing player decision-making on the base paths, or re-evaluating their aggressive base-stealing strategies against teams with strong defensive catchers.
- *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.

#### *3.1.1.6. 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 potential initial lapses 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.

#### *3.1.1.7. Homeruns Scored*

##### *PC Component Analysis and Interpretation:*

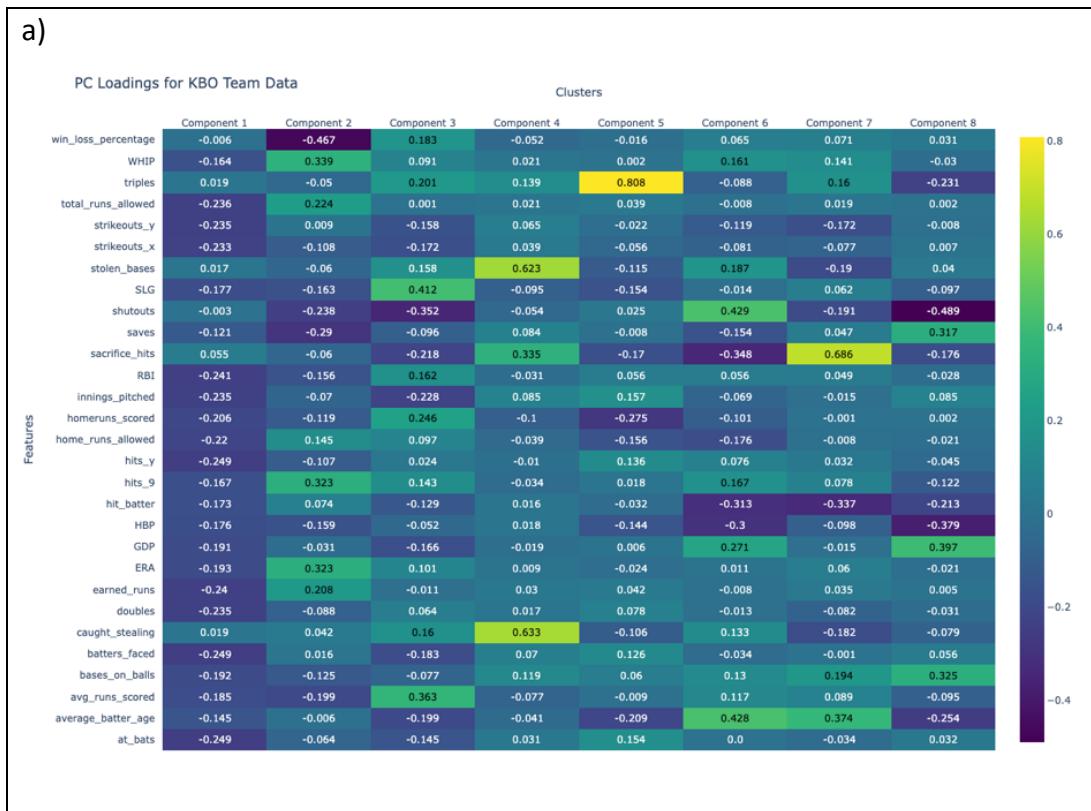
- *Component 1: Low Negative Loading (-0.206)*
  - Teams under this component display a low negative loading for the "Homeruns Scored" variable. This implies that while these teams may not be the lowest scoring in terms of homeruns, they certainly fall below the league's average. A low homerun scoring 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 clustered 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.

#### *3.1.1.8. 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 cluster 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.

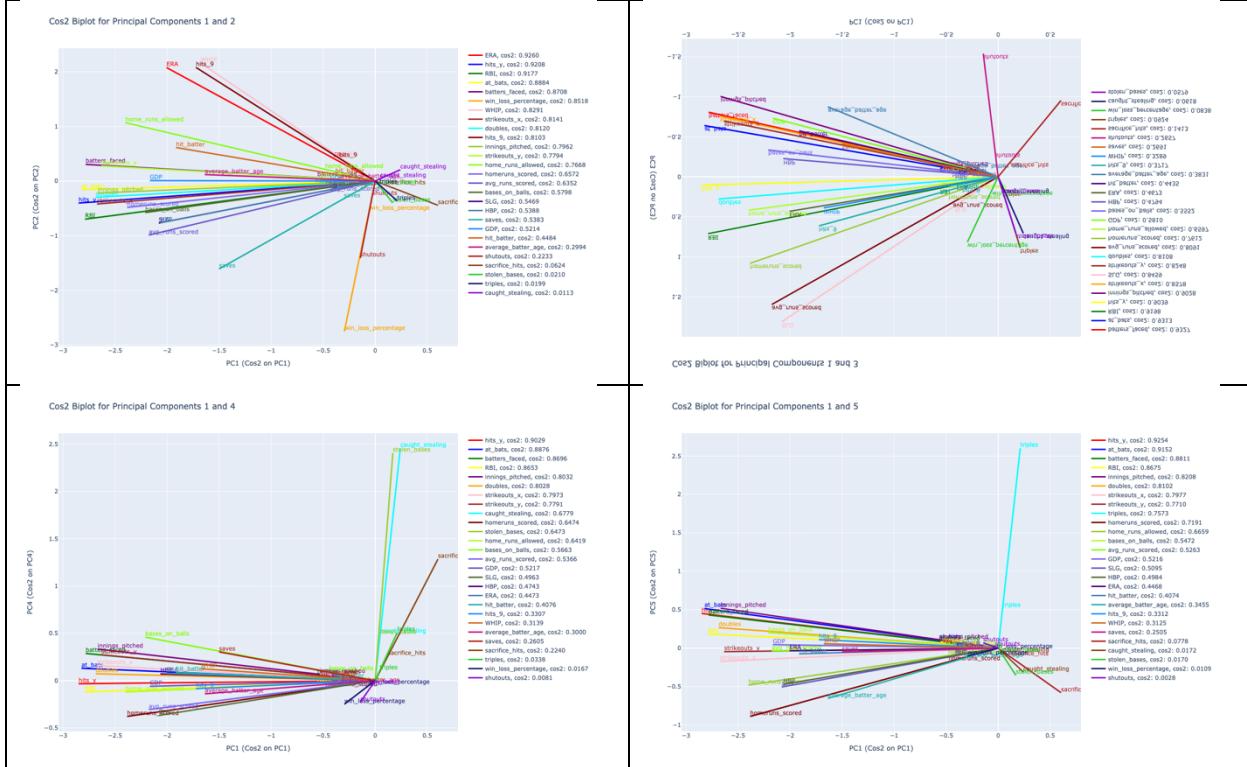
### 3.1.2. Figure Panel 3: Heatmap Representation of PC loadings

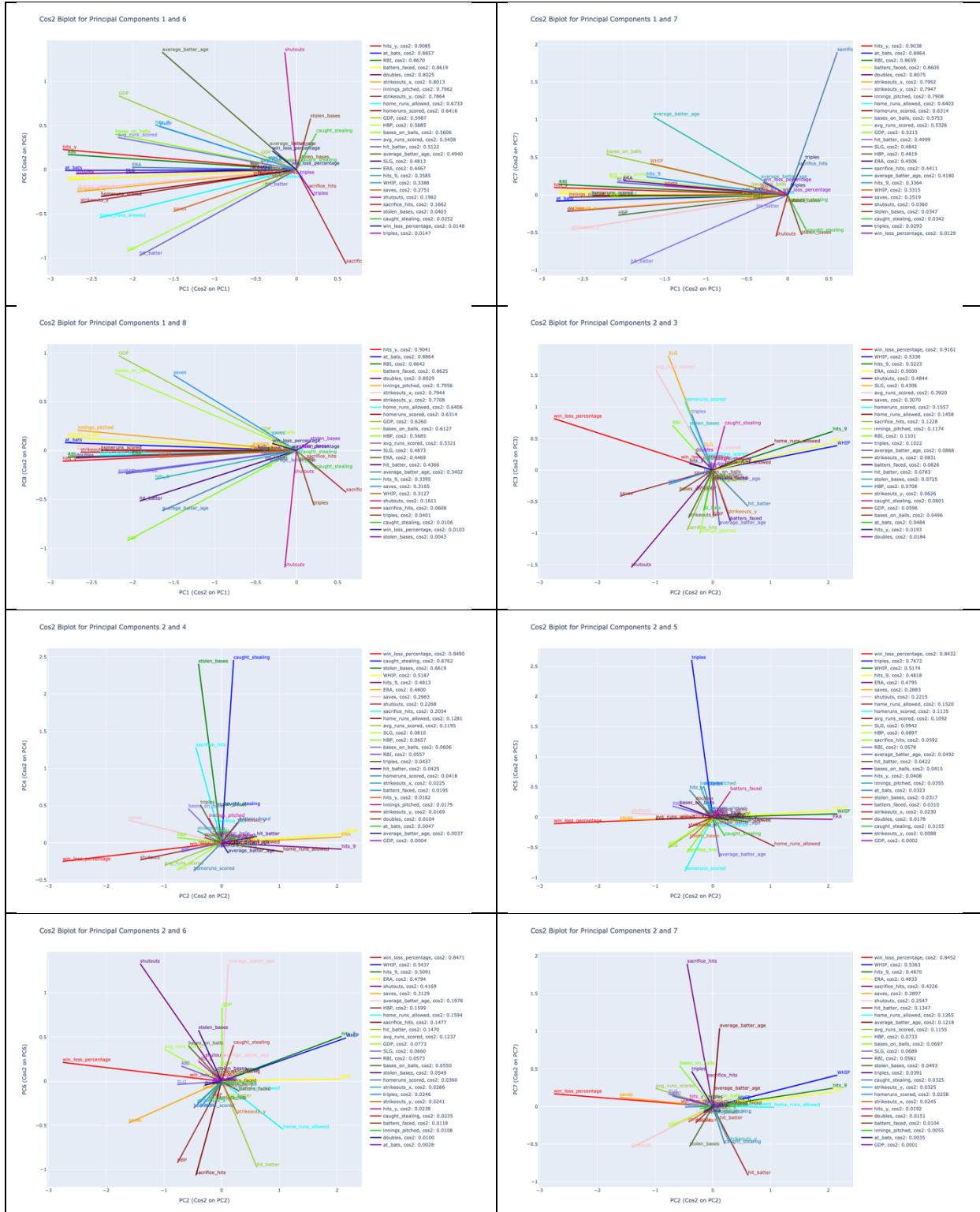




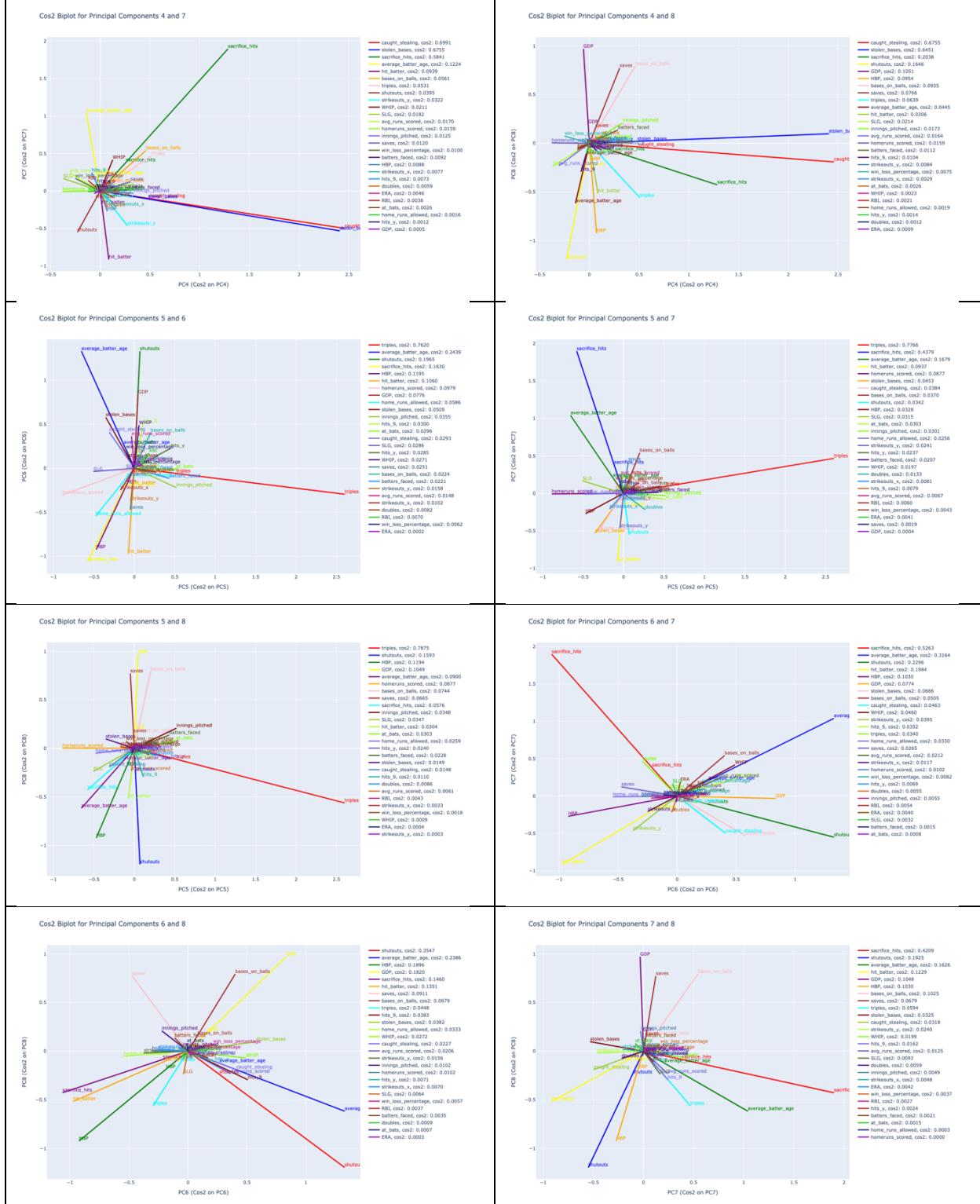
Heatmap representation of eigenvalues derived from Principal components: a) Principal component loading of KBO dataset. b) Principal component loading of MLB dataset

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









**Figure Panel 2.** 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  $\cos^2$  values essentially mirror the square of the correlation between the variables and the component, elucidating the extent to which these variables shape the component's nature. In other words, 'ERA' and 'WHIP', with their high  $\cos^2$  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' serve as double-edged swords: they attest to defensive challenges and echo the struggles 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 grapple with during the tail-end 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, as utilized in our Figure Panel 3, are two-dimensional plots that showcase both the samples and variables of a dataset after it has been subjected to PCA. They are particularly useful in visualizing how original variables correlate with one another and how they relate to the identified principal components.

1. Direction and Magnitude of Vectors: The orientation of vectors within a biplot offers a window 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 pivotal 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: Highlighted prominently in the biplot,  $\cos^2$  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 robustness and 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.

#### 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.

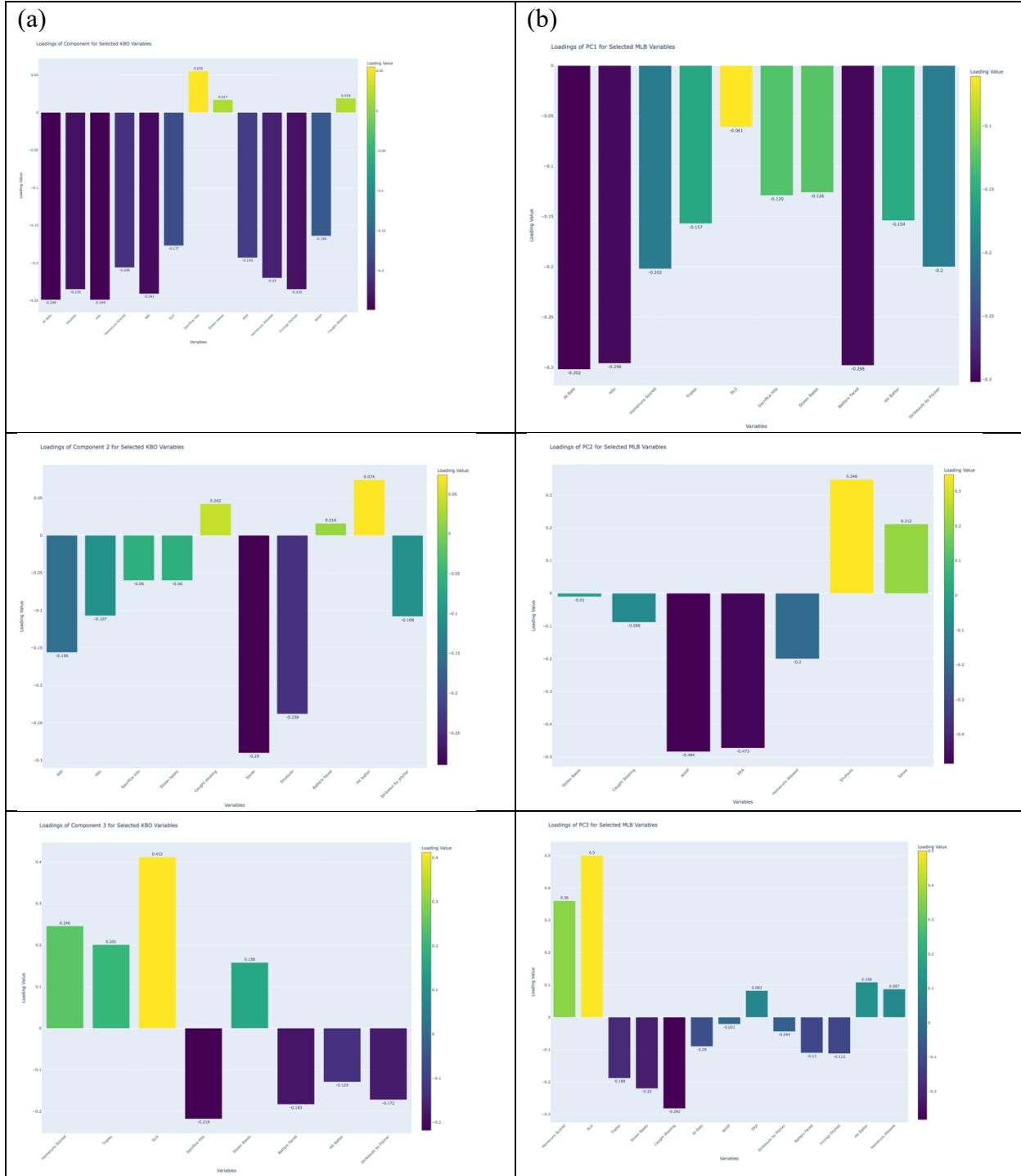
In summary, MLB's Component 3 teams are characterized by their heavy reliance on power hitting while maintaining a pitching lineup that exudes discretion. Unlike KBO's potential balance in batting and bowling, these teams are making their mark through explosive offense and careful, strategic defense.

#### 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 synthesize this analysis and visually portray how teams in each component embody these eigenvalues, histograms were meticulously crafted for each principal component. These histograms underscore the most 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) manifests 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 elevated loading for 'Sacrifice Hits' implies that teams resonating 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 testament to a team's audacity, also comes with inherent risks. The prominence 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 reduced focus on frequent hits like 'At Bats' and 'Hits', but elevated loadings for impactful plays such as 'Homeruns Scored' and notably, 'SLG', underpin this philosophy.

Through its variable loadings, the histogram elucidates 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 palpable impact. It's about ensuring every offensive maneuver, however limited, resonates throughout the innings.

#### 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 baserunning, 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 to examine the statistical properties and internal dynamics of KBO teams. The results revealed key differences between KBO and MLB teams, highlighting the unique style of play and strategy in the KBO league.

The PCA yielded 5 principal components that accounted for over 70% of the variance in the data. These components represented batting prowess, pitching control, fielding capability, baserunning aggressiveness, and hitting consistency. The subsequent GMM clustering showed that KBO teams can be categorized into three archetypes: slugging teams, pitching-focused teams, and balanced teams.

Notably, batting power as measured by component 1 was the prime differentiator between KBO and MLB teams. KBO teams scored higher on power hitting statistics like home runs, RBI, and slugging percentage. This indicates a greater emphasis on offensive play in the KBO compared to MLB, which focuses more on pitching and defense.

The biplot analysis also revealed the importance of individual player capabilities in the KBO, while MLB teams are more defined by their overall team dynamics. This further highlights the reliance on star players for strong offensive performance in the KBO league.

By moving beyond existing MLB-centric models like the Pythagorean expectation, this study provides new insights into the statistical properties and team strategies that characterize the KBO league. The findings can inform team management and help shape data-driven decision making tailored to the KBO's distinctive brand of baseball. Further research can build on these results to develop predictive models and performance evaluation tools customized for the KBO.

## 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
<b>At-Bats</b>	Low Negative	Low Negative	Low Negative	Low Positive	Low Positive	Low Positive	Low Negative	Low Positive
<b>Average Batter Age</b>	Low Negative	Low Negative	High Negative	Low Negative	High Negative	High Positive	High Positive	High Negative
<b>Bases On Balls</b>	Low Negative	Low Negative	Low Negative	Low Positive	Low Positive	Low Positive	Low Positive	High Positive
<b>Batters Faced</b>	Low Negative	Low Positive	Low Negative	Low Positive	Low Positive	Low Negative	Low Negative	Low Positive
<b>Caught Stealing</b>	High Positive	Low Positive	Low Positive	High Positive	Low Negative	Low Positive	High Negative	Low Negative
<b>Doubles</b>	Low Negative	Low Negative	Low Positive	Low Positive	Low Positive	Low Negative	Low Negative	Low Negative
<b>Earned Runs</b>	Low Negative	High Positive	Low Negative	Low Positive	Low Positive	Low Negative	Low Positive	Low Positive
<b>ERA</b>	Low Negative	High Positive	Low Positive	Low Positive	Low Negative	Low Positive	Low Positive	Low Negative
<b>GDP</b>	Low Negative	Low Negative	Low Negative	Low Negative	Low Positive	High Positive	Low Negative	High Positive
<b>HBP</b>	Low Negative	Low Negative	Low Negative	Low Positive	Low Negative	High Negative	Low Negative	High Negative
<b>Hit by Pitch</b>	Low Negative	Low Positive	Low Negative	Low Positive	Low Negative	High Negative	High Negative	Low Negative
<b>Hits</b>	Low Negative	Low Negative	Low Positive	Low Negative	Low Positive	Low Positive	Low Positive	Low Negative
<b>Home Runs Allowed</b>	Low Negative	Low Positive	Low Positive	Low Negative				
<b>Homeruns Scored</b>	Low Negative	Low Negative	High Positive	Low Negative	High Negative	Low Negative	Low Negative	Low Positive
<b>Innings Pitched</b>	Low Negative	Low Negative	High Negative	Low Positive	Low Positive	Low Negative	Low Negative	Low Positive
<b>RBI</b>	Low Negative	Low Negative	Low Positive	Low Negative	Low Positive	Low Positive	Low Positive	Low Negative
<b>sacrifice hits</b>	High Positive	Low Negative	High Negative	High Positive	Low Negative	High Negative	High Positive	Low Negative

<b>saves</b>	Low Negative	High Negative	Low Negative	Low Positive	Low Negative	Low Negative	Low Positive	High Positive
<b>shutouts</b>	Low Negative	High Negative	High Negative	Low Negative	Low Positive	High Positive	High Negative	High Negative
<b>SLG</b>	Low Negative	Low Negative	High Positive	Low Negative	Low Negative	Low Negative	Low Positive	Low Negative
<b>Stolen bases</b>	High Positive	Low Negative	Low Positive	High Positive	Low Negative	Low Positive	High Negative	Low Positive
<b>Strikeout by Batter</b>	Low Negative	Low Negative	Low Negative	Low Positive	Low Negative	Low Negative	Low Negative	Low Positive
<b>Strikeout by Pitcher</b>	Low Negative	Low Positive	Low Negative	Low Positive	Low Negative	Low Negative	High Negative	low negative
<b>Triples</b>	High Positive	Low Negative	High Positive	Low Positive	High Positive	Low Negative	Low Positive	high negative
<b>WHIP (Walks Plus Hits Per Inning Pitched)</b>	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

	<b>PC1</b>	<b>PC2</b>	<b>PC3</b>	<b>PC4</b>	<b>PC5</b>	<b>PC6</b>	<b>PC7</b>	<b>PC8</b>
<b>At-Bats</b>	High Negative	Low Positive	Low Negative	Low Positive	Low Negative	Low Negative	Low Negative	Low Negative
<b>Average Batter Age</b>	Low Negative	Low Negative	Low Positive	High Negative	High Negative	High Positive	High Negative	High Positive
<b>Bases On Balls</b>	Low Negative	Low Positive	High Positive	High Negative	Low Positive	Low Positive	Low Positive	Low Positive
<b>Batters Faced</b>	Low Negative	Low Negative	Low Positive	Low Negative	Low Negative	Low Positive	Low Positive	Low Positive
<b>Caught Stealing</b>	High Negative	Low Positive	Low Negative	Low Positive	Low Negative	Low Negative	Low Negative	Low Negative
<b>Doubles</b>	Low Negative	Low Positive	High Negative	Low Negative	High Positive	High Positive	Low Negative	Low Negative
<b>Earned Runs</b>	High Negative	Low Positive	Low Positive	Low Negative	Low Negative	High Negative	Low Positive	Low Positive
<b>ERA</b>	High Positive	High Positive	Low Negative	Low Positive	Low Negative	Low Positive	Low Negative	Low Negative
<b>GDP</b>	Low Negative	Low Positive	Low Negative	Low Negative	High Negative	Low Negative	Low Negative	High Negative
<b>HBP</b>	Low Negative	Low Positive	Low Positive	High Positive	Low Negative	High Positive	High Positive	High Positive
<b>Hit by Pitch</b>	Low Negative	Low Negative	Low Positive	Low Positive	Low Negative	High Positive	High Positive	High Negative
<b>Hits</b>	High Negative	Low Positive	Low Negative					
<b>Home Runs Allowed</b>	Low Negative	High Positive	Low Positive	High Positive	Low Positive	Low Positive	High Negative	Low Negative
<b>Homeruns Scored</b>	Low Negative	Low Positive	High Positive	Low Positive	Low Positive	Low Positive	Low Negative	Low Negative
<b>Innings Pitched</b>	High Negative	Low Negative	Low Negative	Low Positive	Low Negative	Low Negative	Low Negative	Low Negative

<b>RBI</b>	High Negative	Low Positive	Low Positive	Low Negative	Low Positive	Low Negative	Low Negative	Low Positive
<b>sacrifice hits</b>	Low Negative	Low Positive	High Negative	High Negative	Low Negative	Low Negative	High Positive	Low Negative
<b>saves</b>	Low Negative	High Negative	Low Positive	Low Negative	Low Negative	Low Positive	Low Negative	High Negative
<b>shutouts</b>	Low Negative	High Negative	Low Positive	Low Negative	Low Negative	Low Negative	Low Positive	High Positive
<b>SLG</b>	Low Negative	Low Positive	High Positive	Low Negative	Low Positive	Low Negative	Low Positive	Low Negative
<b>Stolen bases</b>	Low Negative	Low Negative	High Negative	Low Negative	High Positive	High Positive	High Negative	Low Positive
<b>Strikeout by Batter</b>	Low Negative	Low Negative	Low Negative	High Positive	Low Positive	Low Negative	Low Negative	Low Negative
<b>Strikeout by Pitcher</b>	Low Negative	High Negative	Low Positive	High Positive	Low Positive	Low Negative	Low Negative	Low Positive
<b>Triples</b>	Low Negative	Low Positive	High Negative	Low Negative	High Positive	High Negative	High Positive	High Positive
<b>WHIP (Walks Plus Hits Per Inning Pitched)</b>	Low Negative	High Positive	Low Negative	Low Negative	Low Negative	Low Negative	Low Positive	Low Negative

### *References and Citations*

- 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>
-