ANALYZING BASEBALL STATISTICS ACROSS CULTURES

: A STUDY OF THE KBO

Boston University
Junho Eum

Table of Contents

Abstract

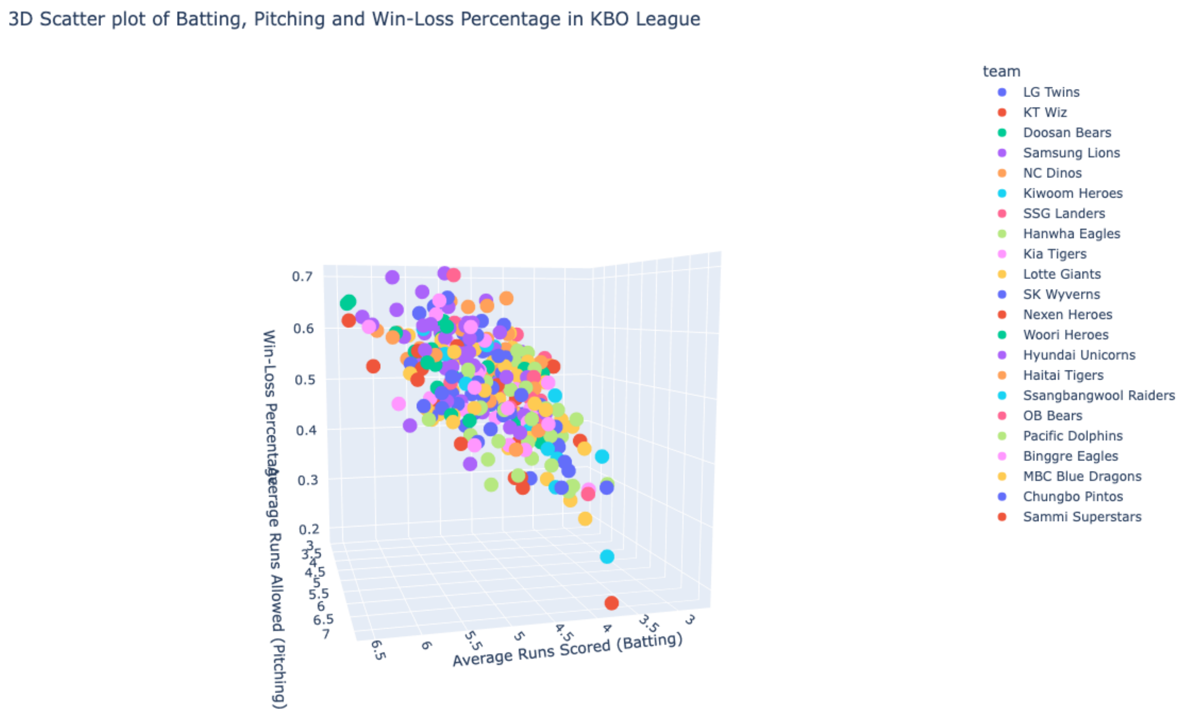
The Korean Baseball League (KBO) was established in 1982, significantly later than other professional baseball leagues such as Nippon Professional Baseball (NPB) in Japan and Major League Baseball (MLB) in the United States which were founded in 1936 and 1876, respectively. Consequently, the KBO has been influenced by both the Japanese and American baseball cultures, resulting in the development of its unique playing style. This research paper aims to analyze and compare the playing style of the KBO with that of the MLB and NPB by studying their respective statistical data. The research question for this study is: "How does the KBO differ in its statistical characteristics compared to the MLB and NPB, and how do these differences contribute to its distinct playing style?"

Introduction

Numerous studies have investigated a multitude of baseball statistics facets, including player performance, game strategies, and team dynamics, in various leagues. Yet, there is an unaddressed need for a comprehensive analysis across different professional baseball leagues, particularly between the Korean Baseball Organization (KBO), Major League Baseball (MLB), and Nippon Professional Baseball (NPB). This paper aims to bridge this gap by implementing a cross-cultural analysis of baseball statistics, leveraging the Pythagorean Expectation formula. Conceived by sabermetrics pioneer Bill James, this formula estimates a team's win rate based on the ratio of the square of runs scored to the square of runs scored and runs allowed: Pythagorean Expectation = (RS^2) / (RS^2 + RA^2).

In the scope of this paper, we propose to construct a model centered around the Pythagorean Expectation formula. We will employ diverse parameter selection methodologies using pitching and batting datasets. This approach will facilitate a comparative analysis that could illuminate distinctive patterns or divergences across the leagues.

The Pythagorean Expectation proposes a direct proportionality between a team's win-loss ratio and the ratio of their runs scored to the combined squares of runs scored and allowed. Consequently, an increase in runs scored (or a decrease in runs allowed) should lead to an elevation in win-loss probability. However, the theory's simplistic nature doesn't account for the complexity and variability inherent in the sport. As observed from a 3D plot depicting KBO data, several outliers do not adhere to the Pythagorean Expectation (Figure 1-1).



These outliers could represent teams with extraordinary clutch performance, exceptional pitching prowess, or teams that starkly deviate from their predicted win-loss ratio due to unforeseen factors or pure chance.

This paper seeks to explore the factors that might contribute to these deviations and attempt to explain these anomalies. Through a comparative analysis across different baseball leagues, we aim to gain insights into the distinctiveness of each league, which might be contributing to these outliers. The objective is to provide a more nuanced understanding of team performance that extends beyond the simplicity of the Pythagorean Expectation.

Methodology

Preprocessing & parameter selection

In this study, we began by importing two primary datasets, specifically the pitching and batting records. Upon conducting an initial data integrity check, we identified the presence of missing data points in several columns. We attribute these inconsistencies to the lack of a comprehensive stat recording system in the historical context. Given the study's objective to investigate recent advancements and differences in the Korean Baseball Organization (KBO) playstyle, we made a strategic decision to remove duplicates and rows containing null values, to uphold data quality and validity.

Subsequently, we implemented a correlation analysis to identify variables that significantly correlate with the dependent variable, i.e., the win-loss percentage. The outcome of this analysis was visually represented in a correlation heatmap (See Figure 1). To further refine our exploration, we created a focused heatmap to explicitly illuminate the correlations between win-loss percentage and other associated variables (Figure 2). This process helped underscore variables that significantly contributed to the dependent variable.

However, it is important to acknowledge the limitations of such an approach. While it effectively highlighted variables like 'wins' and 'losses' that exhibited high collinearity with the dependent variable 'win-loss percentage,' this technique of feature selection may not be sufficient in isolation.

Selecting features based solely on their correlation with the dependent variable can lead to model errors due to several reasons. Firstly, it overlooks the possibility of multicollinearity, a situation where independent variables are highly correlated with each other. In such a case, the model's interpretability suffers as it becomes challenging to distinguish the individual effects of predictors on the response variable. Secondly, correlation does not imply causation. High correlation may simply result from lurking or confounding variables. Finally, this approach is more suitable for linear relationships and may miss out on important non-linear relationships that exist in the data.

To avoid any possible issues, we took a smart approach and split the KBO league's features into two categories: pitcher and batter features. This allowed us to dive deeper into the data and gain better insights.

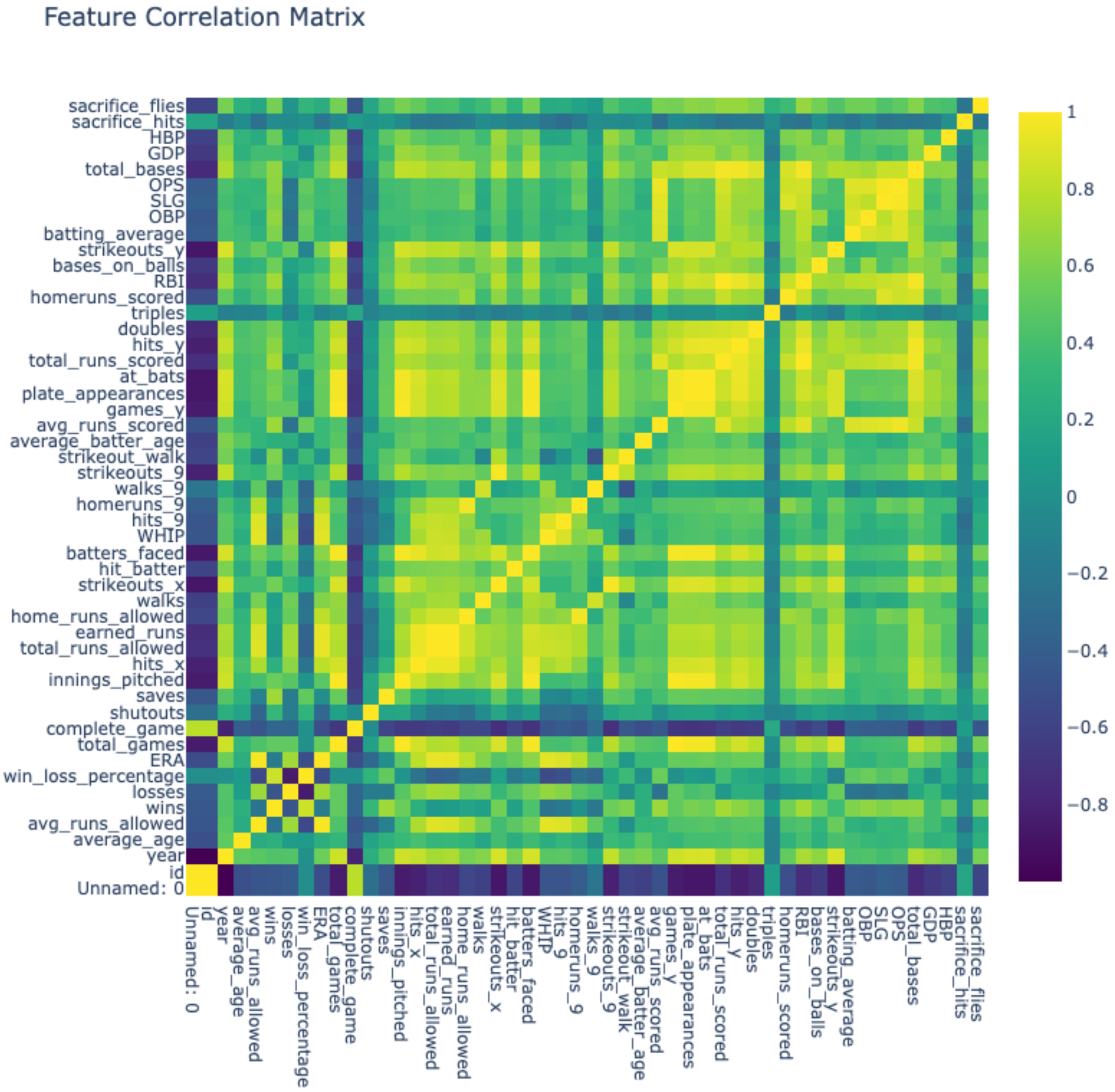


Figure 1. Correlation heatmap for dataset parameters

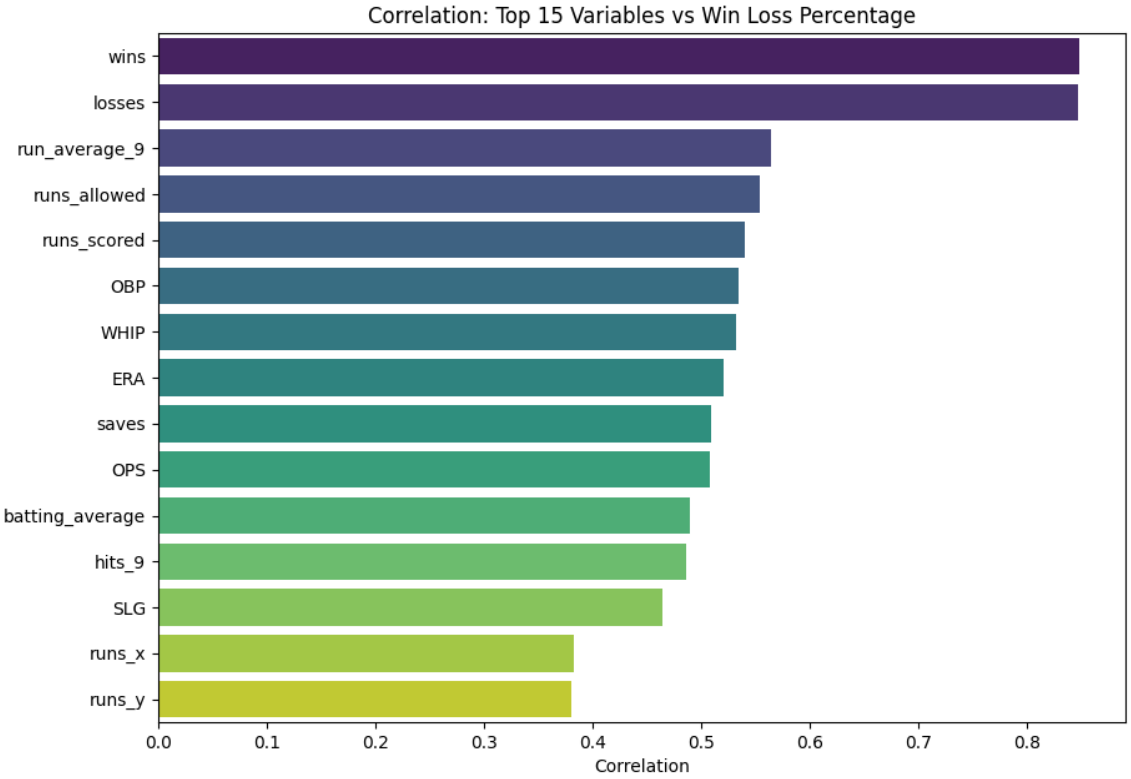


Figure 2. Specified heatmap for parameter selection

Regression Analysis to Understand Influence of Batting and Pitching Variables

We next performed regression analyses to understand the influence of batting and pitching variables on the win-loss percentage.

The sum of the absolute values of the coefficients indicated the total influence of each group of variables, assuming all variables are on the same scale. The sum for the pitching-related variables was approximately 1.42, suggesting that for a one-unit change in the normalized pitching-related features, we expect an average change of approximately 1.42 in the win-loss percentage. The sum for the batting-related variables was approximately 2.98, implying that for a one-unit change in the normalized batting-related features, we expect an average change of approximately 2.98 in the win-loss percentage.

We also examined the decrease in the adjusted R-squared when removing a group of variables. This showed how much of the variance in the win-loss percentage that group of variables explains. The decrease in R-squared when removing the pitching variables was approximately 0.417, indicating that about 41.7% of the variability in the win-loss percentage could be explained by the pitching variables. Conversely, the decrease in R-squared when removing the batting variables was approximately 0.284, meaning about 28.4% of the variability in the win-loss percentage could be explained by the batting variables (See Figure 3).

These analyses revealed that batting and pitching variables have different levels of influence on the win-loss percentage, depending on the criteria used to evaluate their importance. This necessitates further investigations into potential multicollinearity and interaction effects among variables.

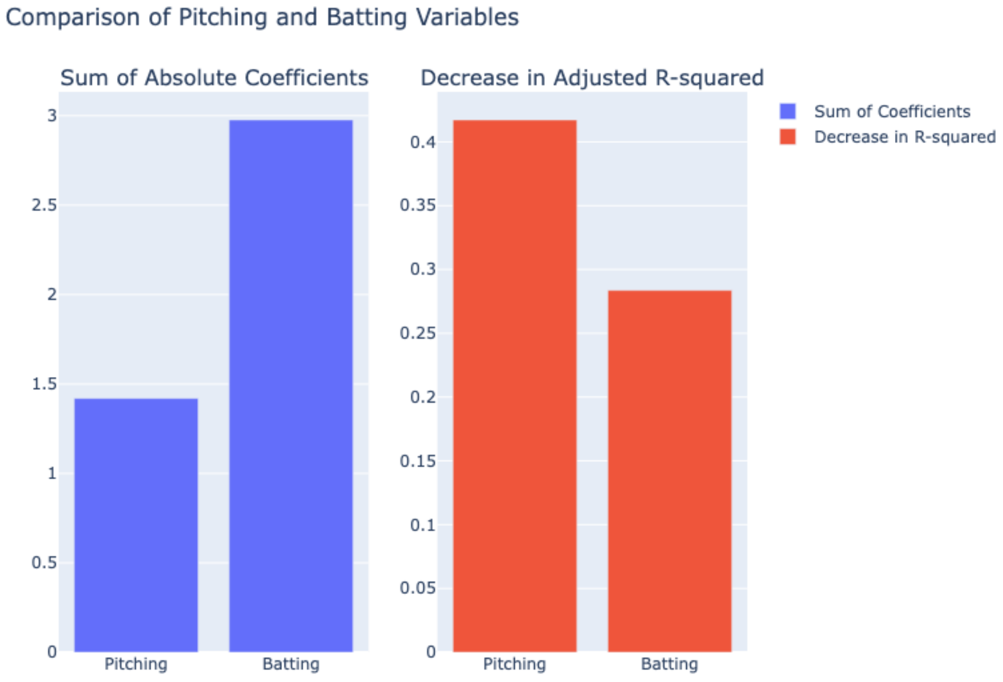


Figure 3. Comparative Influence and Explained Variability of Pitching and Batting Variables on Win-Loss Percentage

Principal component analysis with Gaussian Mixture Model

The methodology employed Principal Component Analysis (PCA) in conjunction with Gaussian Mixture Models (GMM) clustering on both batter and pitcher data. The primary goal of this methodological step was to reduce the dataset's dimensionality and identify key features that could be leveraged to enhance the precision of a win percentage prediction model.

To manage the high dimensionality of the dataset, PCA was utilized initially. PCA transformed the dataset into a new set of orthogonal features, or "components", each of which was a linear combination of the original features. Each component represented a certain amount of the total variance in the dataset.

The selection of features was carried out with care to ensure the quality of the PCA and subsequent GMM. We focused on the following KBO batting features: 'hits_y', 'doubles', 'triples', 'homeruns_scored', 'RBI', 'bases_on_balls', 'strikeouts_y', 'batting_average', 'OBP', 'SLG', 'OPS', 'total_bases', 'GDP', 'HBP', 'sacrifice_hits', and 'sacrifice_flies'. These features were carefully chosen based on a correlation heatmap analysis, as they demonstrated strong interactions and were crucial in determining the offensive performance of a team.

Deciding on the number of components to retain in a PCA analysis is a crucial step. In our case, the decision was made by observing a scree plot and considering the cumulative explained variance. We settled on six components as they captured about 92.3% of the variation in the dataset (See Figure 4). This selection allowed us to condense the essential information of 16 variables into 6 orthogonal components, making the subsequent GMM clustering more manageable and insightful.

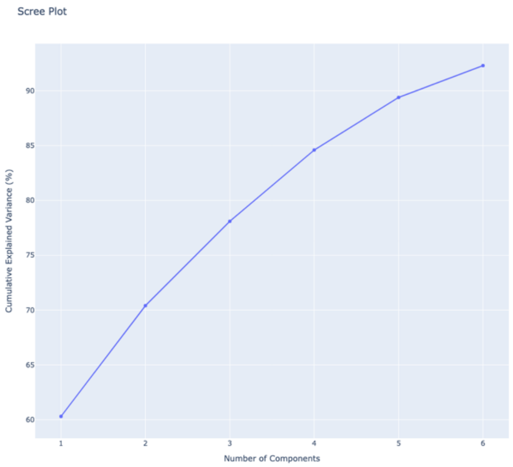


Figure 4. Scree Plot Illustrating the Optimal Number of Principal Components and Explained Variance for PCA Analysis for KBO batting features.

1. **Component 1: "Overall Offensive Performance"** - This component captures general offensive performance as it includes a number of batting metrics. Features like 'RBI', 'total_bases', 'hits_y', and 'OPS' have high positive loadings, implying that players who

score high in this component are strong overall offensive performers. 'sacrifice_hits' is negatively correlated with this component, suggesting that players who score high here do not tend to make many sacrifice hits.

2. **Component 2: "Plate Discipline and Speed"** - This component might be capturing a mixture of plate discipline and speed, considering the negative correlation with 'strikeouts_y' and the positive correlation with 'triples' and 'batting_average'. This suggests that players who score high on this component have good plate discipline (do not strike out frequently), are good at making contact with the ball, and may also be fast runners (as evidenced by the high number of triples).
3. **Component 3: "Speed and Opportunism"** - This component might represent a tendency to use speed and take advantage of situational play, with 'triples' and 'sacrifice_flies' having high positive loadings and 'homeruns_scored' having a negative loading. Players with high scores on this component might be fast and good at seizing scoring opportunities but do not hit many home runs.
4. **Component 4: "Sacrifice Play"** - This component has a high positive loading for 'sacrifice_hits', which indicates that it might be capturing players who are often asked to sacrifice for the good of the team. This means that players who score high on this component often execute sacrifice hits, potentially moving other runners around the bases at the expense of their own at-bat.
5. **Component 5: "Aggressive Play and Efficiency"** - This component appears to be capturing a mixture of aggressive play (as indicated by the negative loading on 'HBP' or Hit By Pitch) and efficient hitting (positive loadings on 'GDP' and 'bases_on_balls', negative loadings on 'doubles' and 'total_bases'). Players scoring high in this component could be those who tend to get hit by pitches often, draw walks, and are efficient in hitting for extra bases.
6. **Component 6: "Tactical Baserunning"** - This component seems to reflect a player's tactical approach to advancing runners and gaining bases, with 'sacrifice_flies' and 'sacrifice_hits' having notable loadings. In contrast, 'bases_on_balls' and 'HBP' have negative correlations, suggesting that players with high scores in this component may not walk or get hit by pitches often, but instead contribute to their team's offense through strategic baserunning.

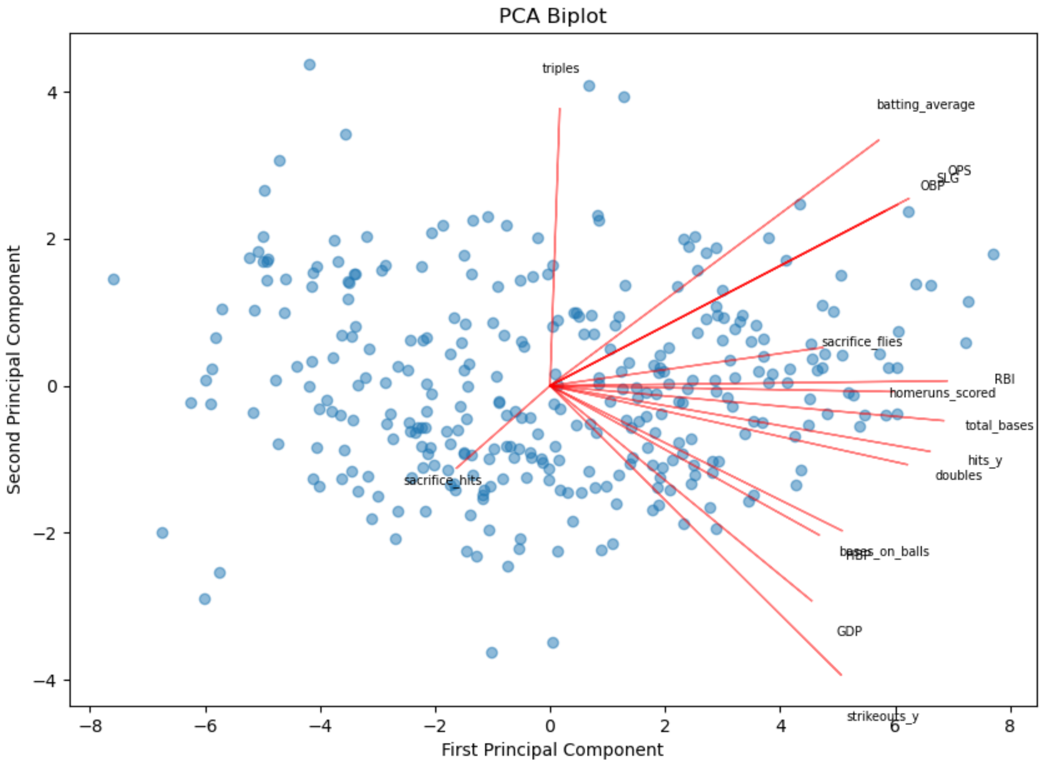


Figure 5. Principal Component Analysis (PCA) Biplot: Batting Features

To understand the relationship between the original variables and the derived PCA components, a biplot was generated. This visualization maps the loading of each variable onto the principal components, highlighting their contributions to the components. The first principal component, dubbed **"Overall Offensive Performance"**, and the second, **"Plate Discipline and Speed"**, serve as the axes (See Figure 5).

Our findings indicate that the variables 'OBP', 'SLG', and 'OPS' display robust positive loadings on the first principal component, interpreted as 'Overall Offensive Performance'. This suggests that a player's On-Base Percentage (OBP), Slugging Percentage (SLG), and On-Base Plus Slugging (OPS), which are key metrics in evaluating a player's offensive capabilities, significantly reinforce this component. Accordingly, a rise in these metrics aligns with an enhancement in a player's overall offensive performance.

Conversely, the first principal component is adversely affected by the variables 'RBI', 'GDP', 'bases_on_balls', and 'strikeouts_y', with 'strikeouts_y' demonstrating the most substantial negative loading. This finding suggests a significant inverse relationship between the number of strikeouts and a player's overall offensive performance, which aligns with common baseball wisdom that higher strikeouts generally signify poorer offensive success. Similarly, increments in RBI, GDP, and bases on balls seem to lower the score on the 'Overall Offensive Performance' component, hinting at less fruitful offensive outcomes.

Lastly, the 'sacrifice_hits' variable is projected towards the third quadrant, which implies a dual effect: a negative influence on 'Overall Offensive Performance' and a positive influence on 'Plate Discipline and Speed'. This trend may suggest that while sacrifice hits tend to negatively impact a player's aggregate offensive statistics (given a sacrifice hit often results in an out), they may be indicative of superior plate discipline and speed. These insights gleaned from the biplot afford us a better understanding of the interplay between player metrics and their contributions to the winning probabilities in the KBO league.

In the first stage, PCA was used to identify the components or factors that most significantly contributed to team performance. This gave us a condensed view of the batting metrics, highlighting the most influential ones. Following this, we used the GMM to cluster the teams based on these principal components.

Deciding on the optimal number of clusters for the GMM was a crucial step. For this, we examined both the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC), as these provide a statistical measure of the model's goodness of fit while factoring in its complexity. By choosing the model with the lowest BIC and AIC values, we were able to determine the most appropriate number of clusters for our data (as shown in Figure 6).

We assigned teams to their respective clusters according to the Gaussian distribution they most closely aligned with. The results of the PCA-GMM analysis, which included unique PCA component loadings and final cluster assignments for each team, were visually represented through scatter plots and histograms. In particular, the histograms in Figure 7 give a graphical view of the distribution of the principal component loadings across the different clusters.

By analyzing the original batting metrics in the context of the PCA component loadings and GMM cluster assignments, we were able to delve into the traits that define a 'winning' team in the KBO league. As depicted in Figure 8, our exploration led to the identification of distinct team clusters, each marked by specific traits: "Offensive Powerhouses", "Disciplined and Fast

Teams", "Opportunistic Speed Teams", "Heavy Hitting Teams", and "Strategic Baserunning Teams". These designations were derived from the PCA loadings, which indicated how much each batting metric contributed to a specific factor of performance.

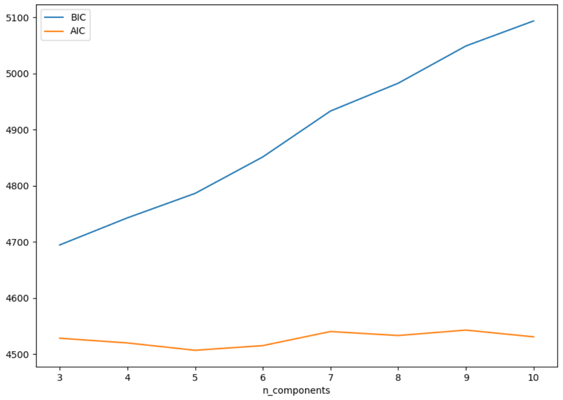


Figure 6. Comparative Analysis of BIC and AIC Metrics to Determine the Optimal Number of Clusters for GMM Clustering

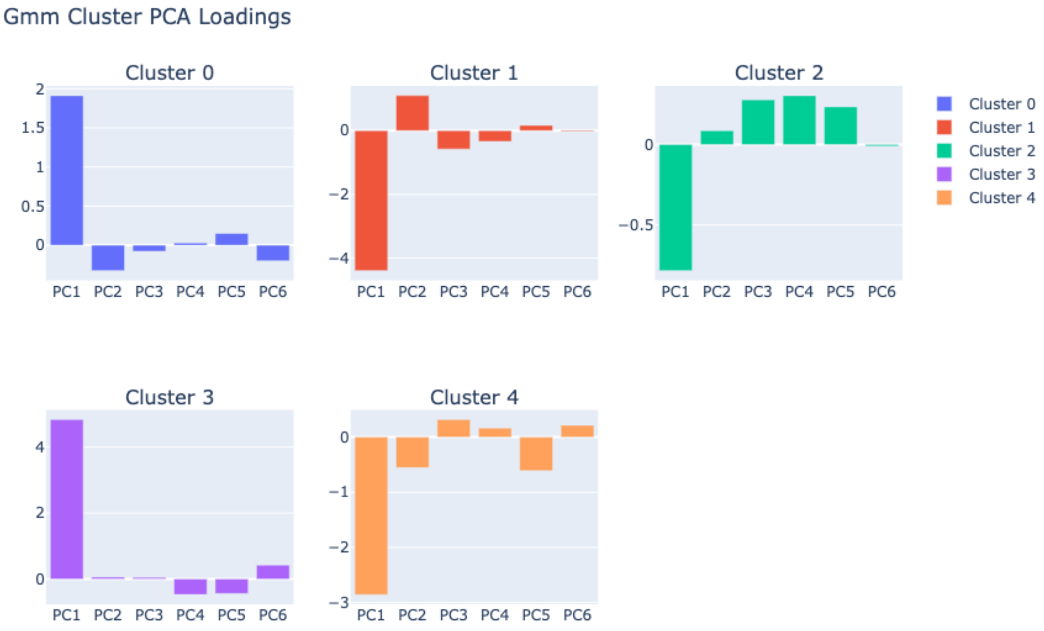


Figure 7. Presentation of Principal Component Analysis (PCA) Loadings for Clusters Identified via Gaussian Mixture Model (GMM)

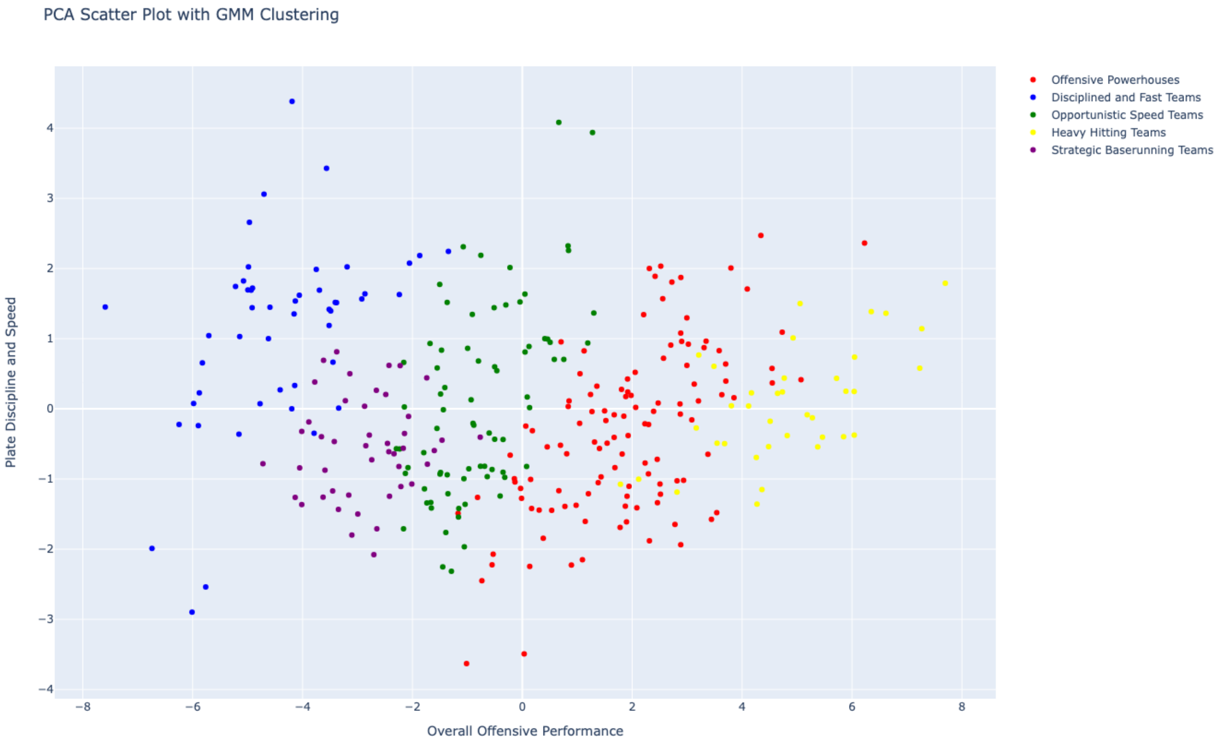


Figure 8. Scatter Plot Representation of Principal Component Analysis (PCA) Clusters, Generated using a Gaussian Mixture Model (GMM), for Korea Baseball Organization (KBO) Batter Data

Shifting to the pitcher metrics, number of components were selected using scree plot after observing the proportion of explained variance reaching 91.5% for 5 components (See Figure 9).

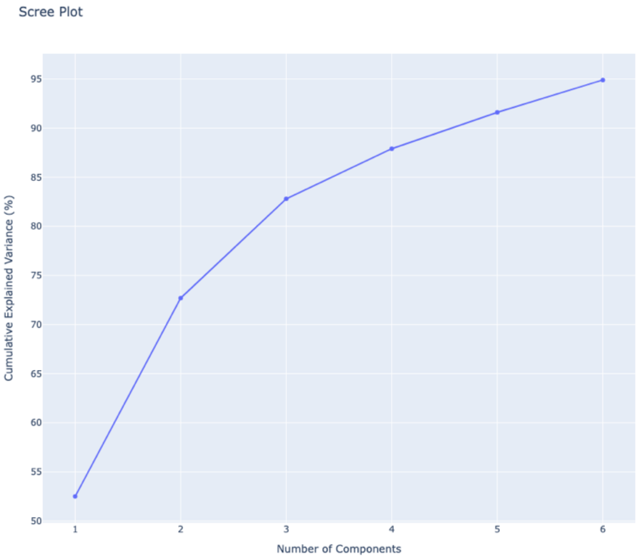


Figure 9. Scree Plot Illustrating the Optimal Number of Principal Components and Explained Variance for PCA Analysis for KBO pitching features.

The first component was strongly associated with variables relating to a pitcher's overall workload and effectiveness.

Component 1: "General Pitching Performance" - High negative loadings on metrics like 'hits_x', 'batters_faced', 'home_runs_allowed', 'innings_pitched', 'strikeouts_x', and 'ERA' suggest this component represents overall pitching performance. Teams with a high score in this component likely have strong pitching staff, allowing fewer hits, striking out more batters, and having a lower ERA. 'Shutouts' has a small positive correlation, implying that teams scoring high on this component may perform shutouts less frequently.

Component 2: "Control and Efficiency" - The negative loading on 'strikeout_walk' and positive loadings on 'WHIP' and 'walks_9' suggests this component may represent control and efficiency in pitching. High scores in this component would correspond to teams whose pitchers have good control, leading to fewer walks and hits per innings pitched (lower WHIP).

Component 3: "Walks and Home Runs Allowed" - This component might represent a tendency to allow walks and home runs, given the negative loadings on 'walks_9' and 'walks', and positive loadings on 'homeruns_9' and 'home_runs_allowed'. Teams scoring high on this component might have pitchers who are prone to giving up walks and home runs.

Component 4: "Shutouts vs. Saves" - The high positive loading for 'shutouts' and high negative loading for 'saves' in this component indicate a balance between shutouts and saves. High scoring teams on this component often perform shutouts and less frequently have saves.

Component 5: "Aggressiveness and Save Situations" - This component appears to capture a mix of aggressive play (as indicated by the negative loading on 'hit_batter') and performance in save situations (positive loading on 'saves'). High scores in this component could be associated with teams whose pitchers tend to hit batters less often and perform well in save situations.

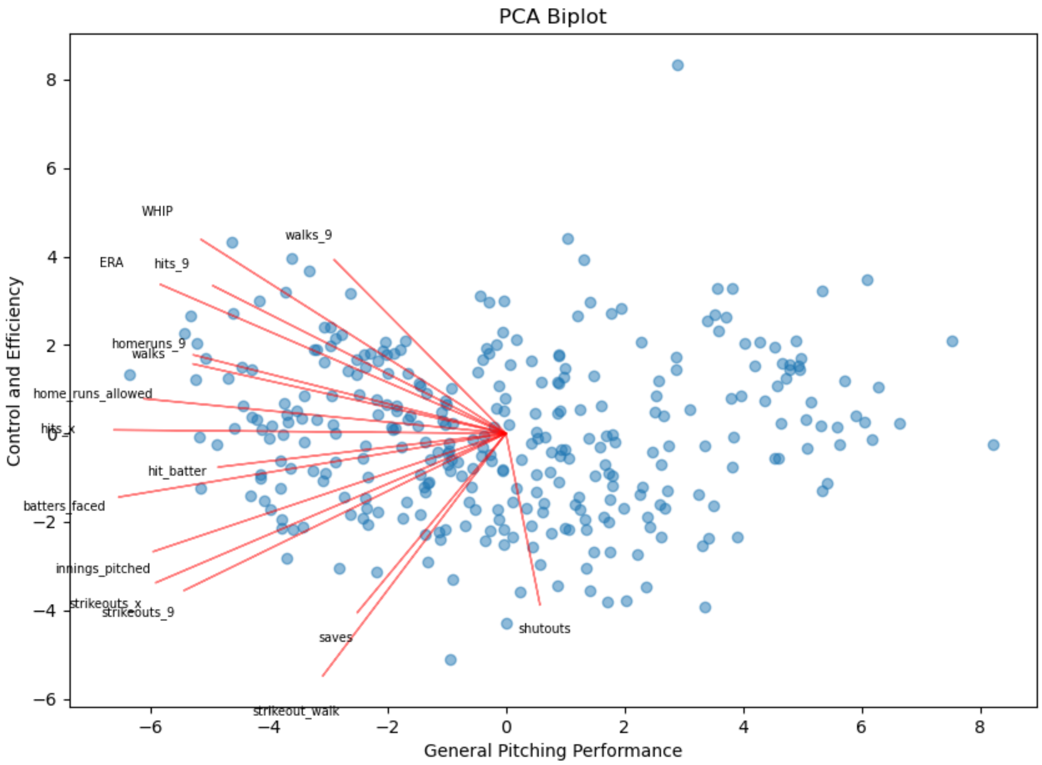


Figure 10. Principal Component Analysis (PCA) Biplot: Pitching Features

To illustrate the relationships between our pitching variables and the principal components, we developed a biplot. This visualization, presented in Figure 10, maps each variable's contribution to the principal components, labeled as "General Pitching Performance" and "Control and Efficiency".

Our observations reveal that 'Walks9', 'WHIP', 'hits_9', 'ERA', and 'homeruns_9' exhibit significant negative loadings on the "Control and Efficiency" component. Particularly, 'Walks9' presents the steepest negative slope, suggesting that its increase, indicating decreased control, negatively impacts the "Control and Efficiency" score. The remaining variables have less steep slopes, pointing out that while they inversely affect "Control and Efficiency", their influence is not as profound as 'Walks9'.

On the other hand, 'strikeouts_x', 'innings pitched', and 'saves' show robust positive loadings on the "General Pitching Performance" component. This implies that the increase in these variables, reflecting superior pitching performance, boosts the score on "General Pitching Performance".

Interestingly, 'shutout', despite its positive interpretation in pitching performance, appears to be negatively associated with the fourth quadrant. This suggests a mixed effect: while an increase in shutouts might dampen the "General Pitching Performance", it could hint at better "Control and Efficiency". This trade-off may originate from a player's decision to adopt more conservative strategies in pursuit of a shutout.

These insights, extracted from the biplot, enhance our comprehension of the complex interaction between various pitching performance indicators and their relative significance to the league's overall performance.

We examined both the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC) for pitching variables as well which showed 3 components as optimal (See Figure 11).

We assigned teams to their respective clusters according to the Gaussian distribution they most closely aligned with. The outcomes of the PCA-GMM analysis, encompassing unique PCA loadings and team cluster assignments, were illustrated with scatter plots and histograms (See Figure 11, 12).

As shown in Figure 13, we identified distinct team clusters, such as "Struggling Pitching Performance", "Highly Controlled Efficiency", and “Mixed Performance”.

Cluster title and interpretations

- Cluster 0: "Struggling Pitching Performance" This cluster has the highest negative value for Component 1, which represents "General Pitching Performance." The higher negative value suggests that these teams perform poorly in terms of allowing hits, facing batters, home runs allowed, innings pitched, and having a higher ERA. Therefore, the name "Struggling Pitching Performance" seems to capture this characteristic.
- Cluster 1: "Highly Controlled Efficiency" This cluster has the highest positive value for Component 2, "Control and Efficiency," indicating that these teams perform exceptionally well in terms of control and efficiency, leading to fewer walks and lower WHIP (Walks plus Hits per Inning Pitched). Hence, the name "Highly Controlled Efficiency" is proposed.

- Cluster 2: "Mixed Performance" This cluster seems to show mixed results across components, with no one component standing out as particularly strong or weak. This is why it is referred to as "Mixed Performance" - there isn't a clear area where this cluster outperforms or underperforms others.

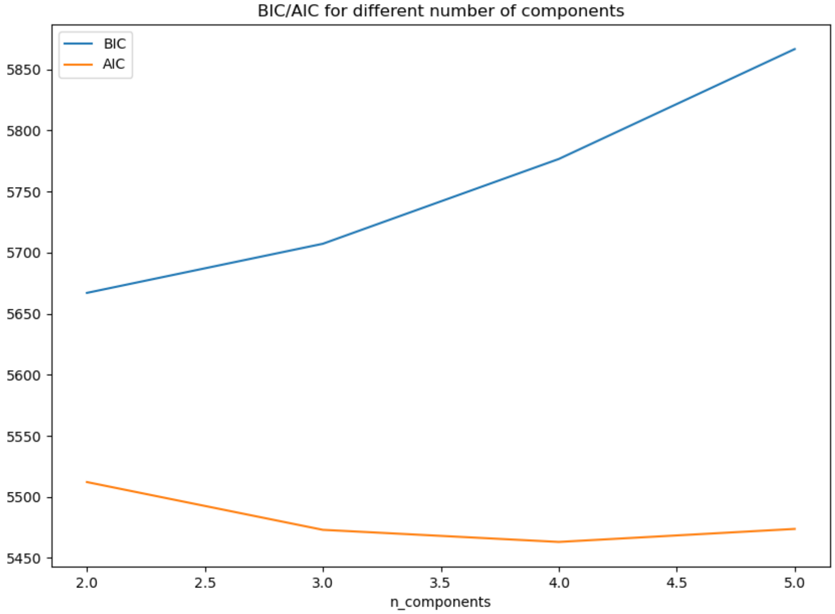


Figure 11. Comparative Analysis of BIC and AIC Metrics to Determine the Optimal Number of Clusters for GMM Clustering for KBO pitching features

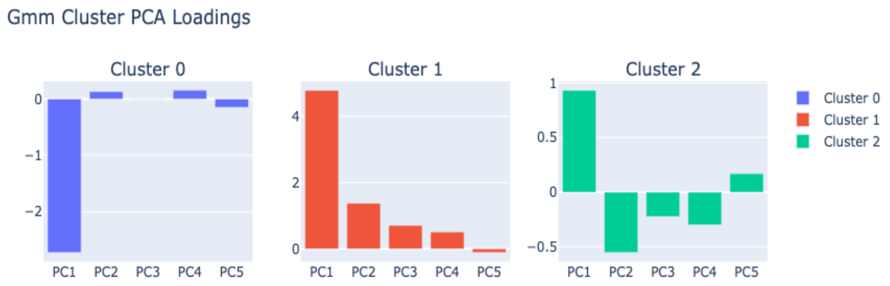


Figure 12. Presentation of Principal Component Analysis (PCA) Loadings for Clusters Identified via Gaussian Mixture Model (GMM) for pitching features.

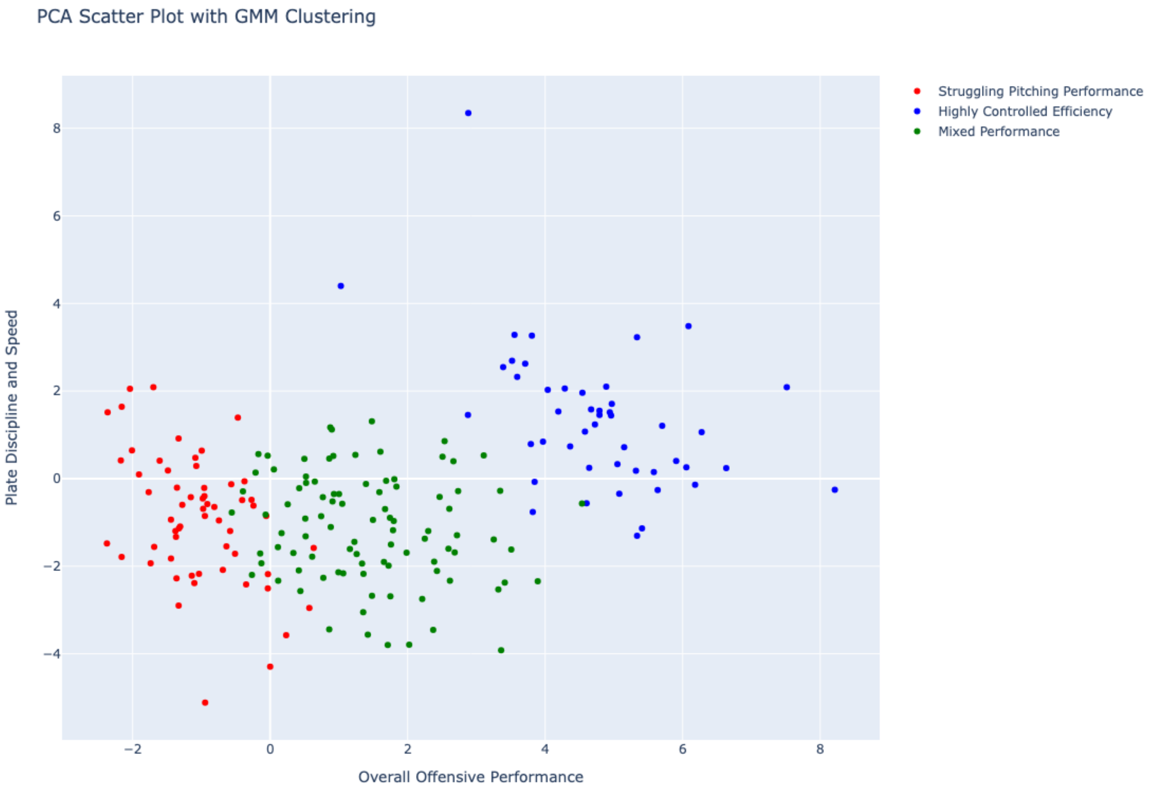


Figure 13. Scatter Plot Representation of Principal Component Analysis (PCA) Clusters, Generated using a Gaussian Mixture Model (GMM), for Korea Baseball Organization (KBO) Batter Data

These three components allow us to capture the critical characteristics of pitching performance with fewer variables, reducing complexity and making further analysis such as clustering more manageable.

This interpretation of the PCA-GMM clustering offers an insightful perspective into what defines a successful team in the KBO. It provides valuable strategic insights, aiding teams to focus on the right areas to improve their win-loss percentages.

Optimization and Interaction Analysis of Batting and Pitching Variables Using XGBoost Models

In the pursuit of comprehensive understanding and capturing interactions between variables, we constructed two XGBoost models. These were developed separately for batter and pitcher data, with the target variable being the total runs scored for both. We categorized the feature variables into two distinct sets, one each for pitcher and batter data, which include respective variables such as Earned Run Average (ERA) and Runs Batted In (RBI).

To identify the optimal parameters for the XGBoost models, we deployed a Grid Search methodology. This approach yielded promising results, with model accuracy scores of 0.9837 for the batter data and 0.9796 for the pitcher data.

Upon further analysis of the batter data, we noted significant interaction between the features RBI and hits, as visualized in the SHAP summary plot (Figure 14). Therefore, these two features were consolidated and considered as a single variable in subsequent models, with the aim of evaluating whether this aggregation improves the model's performance.

We applied similar investigative techniques to the pitcher data in our study. During this process, we identified key pairs of variables - namely (hits allowed, ERA), (ERA, batters faced), and (homeruns allowed, ERA) - which showed notable interaction effects, as illustrated in Figure 15. We combined these pairs of variables to create new, composite features, thereby introducing a more sophisticated layer of analysis to our model.

Additionally, we delved deeper into the specific interaction between homeruns allowed and ERA, as displayed in Figure 16. Our findings revealed a noteworthy dynamic: when the ERA is low, a significant interaction with low homerun counts is evident. However, as the ERA rises, we notice a strong interaction with higher homerun counts. This can be interpreted in such a way that when teams allow more homeruns, it generally coincides with an ERA greater than 4.5. This observation could suggest that teams which concede higher scores tend to allow more homeruns. These insights are of value as we aim to refine our model further. By incorporating this nuanced understanding of the variable interactions into future model iterations, we anticipate enhancing our model's predictive performance, and thereby providing a more detailed understanding of KBO league dynamics.

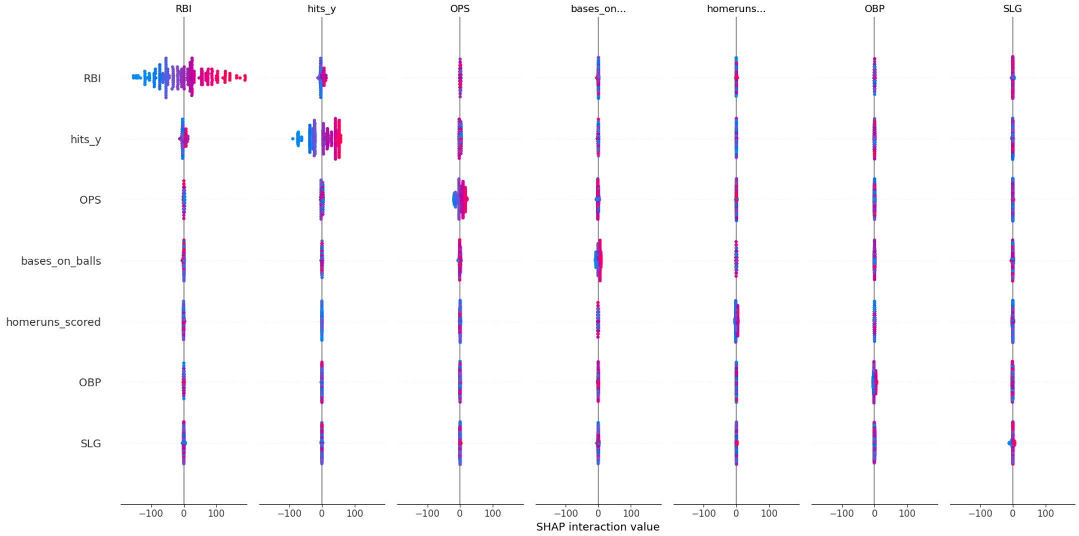


Figure 14. Summary of Interaction plot of KBO batter features

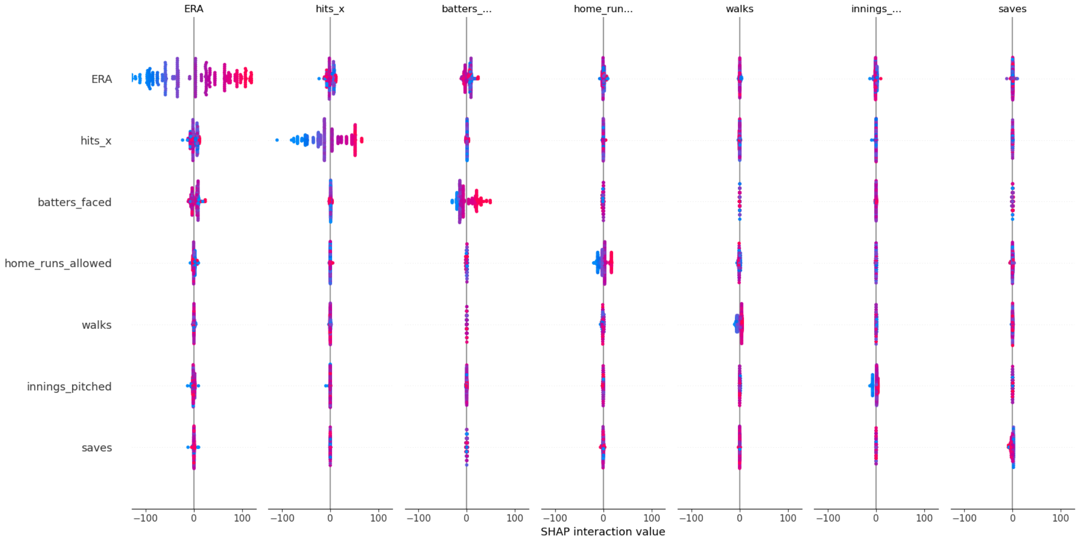


Figure 15. Summary of Interaction plot of KBO pitcher features

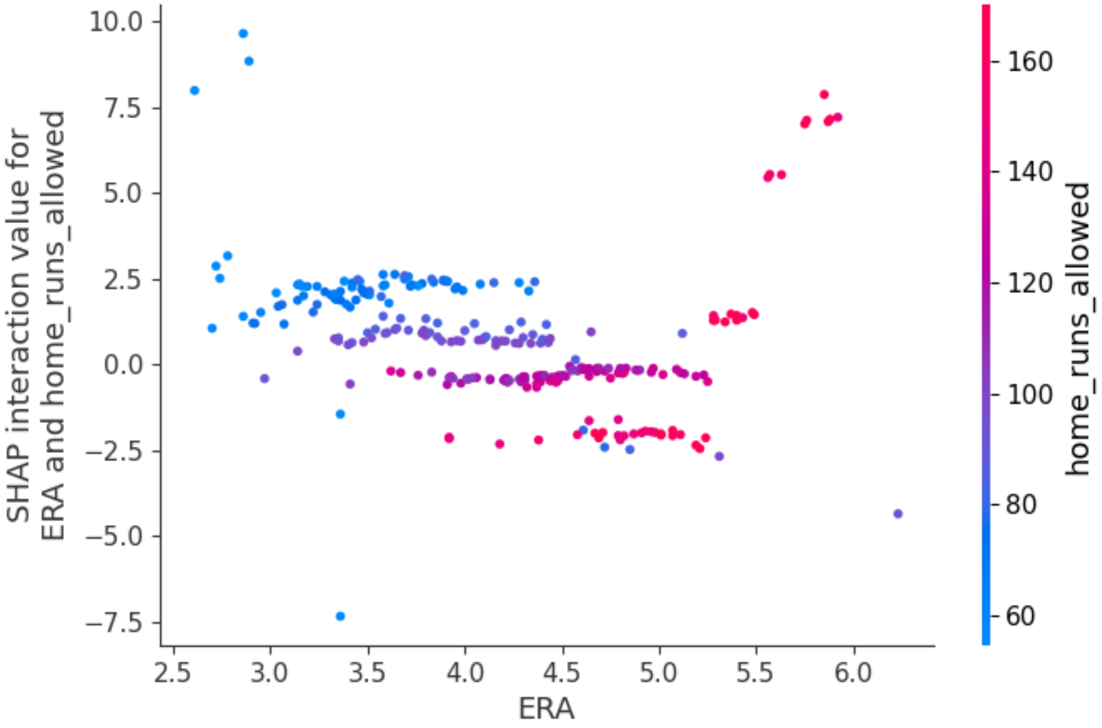


Figure 16. Detailed view of interaction between ERA and homeruns allowed.