ANALYZING BASEBALL STATISTICS ACROSS CULTURES

: A STUDY OF THE KBO

Boston University
Junho Eum

Table of Contents

Abstract *2*

Introduction *2*

Baseball Statistics and Analytical Approaches 2

Adaptations and Extensions 3

Limitations and Scope 3

Objectives of the Current Study 3

Methodology *4*

Preprocessing & parameter selection 4

Regression Analysis to Understand Influence of Batting and Pitching Variables 7

Principal component analysis with Gaussian Mixture Model 8

Optimization and Interaction Analysis of Batting and Pitching Variables Using XGBoost Models.... 22

Lasso Regression for variable selection 25

Ridge Regression 27

Pythagorean Expectation Modeling 29

Conclusion *29*

Abstract

The Korean Baseball Organization (KBO), established in 1982, presents a unique confluence of influences from the significantly older professional baseball leagues like the Nippon Professional Baseball (NPB) in Japan (1936) and Major League Baseball (MLB) in the United States (1876). The KBO has absorbed influences from Nippon Professional Baseball (NPB) in Japan (1936) and Major League Baseball (MLB) in the United States (1876). This historical context has engendered a unique playing style in the KBO. This research paper focuses on analysis of baseball statistics in KBO, particularly focusing on the application of the Pythagorean Expectation formula: a tool for estimating a team's win rate.

Introduction

Baseball Statistics and Analytical Approaches

Numerous studies have investigated a multitude of facets in baseball statistics, including player performance, game strategies, and team dynamics across various leagues. In particular, the **Pythagorean Expectation**, developed by Bill James, is a simple but powerful method to predict expected wins for a team in a single season. The Pythagorean Expectation formula offers a foundational perspective on a team's performance based on the runs they score and allow. However, it is not without its deviations and nuances.

$$\text{Win Ratio} = \frac{\text{runs scored}^2}{\text{runs scored}^2 + \text{runs allowed}^2} = \frac{1}{1 + (\text{runs allowed}/\text{runs scored})^2}$$

Additionally, Bruce Bukiet and Elliotte Harold introduced the **Markov chain approach model** to predict the expected number of wins for baseball teams per season, utilizing batter transition matrix P and run distribution to measure the influence of players at each state of play (Bukiet & Harold, 1997).

Adaptations and Extensions

First, the original formula often exhibits discrepancies when applied to actual game data, leading to the development of the **Pythagenport formula** by Clay Davenport, a baseball sabermetrician and the co-founder of Baseball Prospectus (Davenport & Woolner, 1999). This modified version, utilizing empirical exponents, provides a more accurate reflection of a team's win-loss record. David Smyth's Pythagenport adapts the Pythagorean Expectation by employing an exponent dependent on the scoring environment, thereby allowing for greater precision in various league contexts (Smyth, 2016).

Second, the concepts of **second-order and third-order teams** add complexity to the Pythagorean framework. Second-order teams refer to the expected win-loss record based on underlying performance statistics, while third-order teams consider the quality of the opposition, providing a more nuanced understanding of a team's true performance level. These adaptations of the Pythagorean Expectation illustrate the multifaceted nature of baseball analytics and affirm the importance of considering multiple variables and perspectives in assessing and predicting team success.

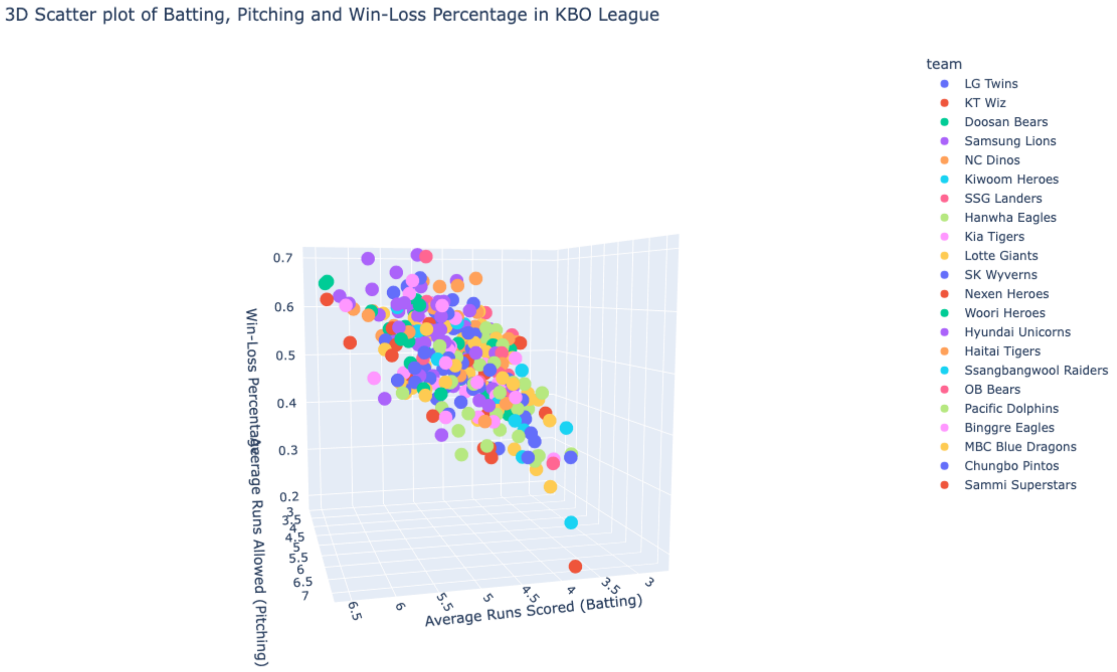
Limitations and Scope

Despite significant strides in baseball analytics, the statistics in baseball had mostly been derived from Major League Baseball data. For example, Steven Miller used 2004 American league data to validate the optimal exponent to predict expected wins, not for KBO teams (Miller, 2007). Furthermore, the best fit exponent for MLB has been identified as approximately 1.74, in excellent agreement with the observation that 1.82 is the optimal exponent for the Pythagorean Formula (Miller, 2007).

Objectives of the Current Study

This paper aims to bridge this gap by introducing a win prediction model for KBO team's seasonal data. 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. However, its 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 these factors and attempt to explain these anomalies. Through comparative analysis across different baseball leagues, we aim to gain insights into the distinctiveness of each league, which might contribute 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, I began by importing two primary datasets: specifically, the pitching and batting records. Upon conducting an initial data integrity check, I identified the presence of missing data points in several columns due to the lack of a comprehensive stat recording system in the historical context. Given the study's objective to investigate statistical characteristics in the

Korean Baseball Organization (KBO) playstyle, I removed duplicates and rows containing null values to uphold data quality and validity.

Subsequently, I implemented a correlation analysis to identify variables that significantly correlate with the dependent variable: win ratio. The outcome of this analysis was visually represented in a correlation heatmap (See Figure 1); additionally, I created a focused heatmap to explicitly illuminate the correlations between win ratio 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 for 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, and 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, I took an approach and split the KBO league's features into two categories: pitcher and batter features. This allowed me to dive deeper into the data and gain better insights.

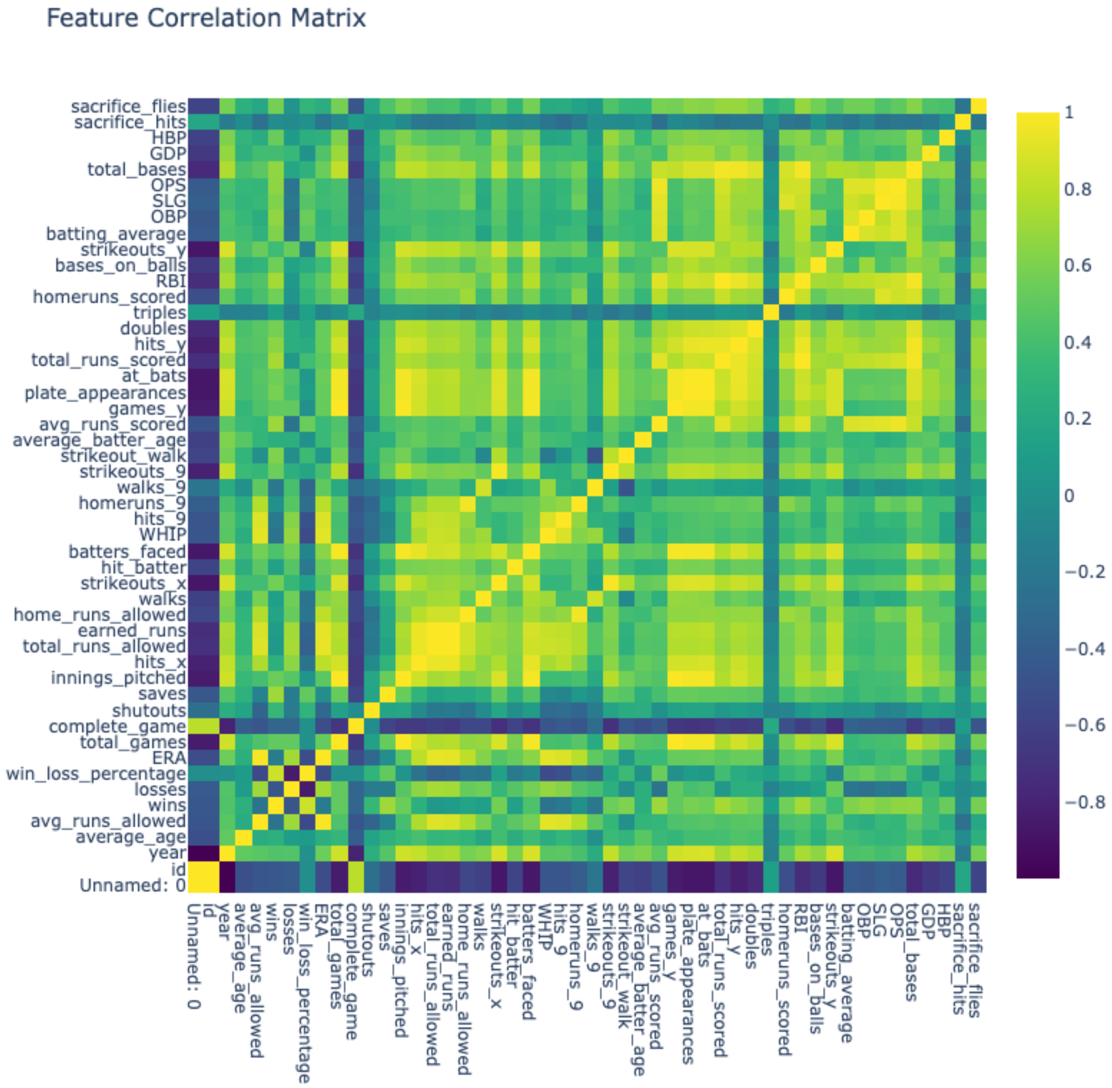


Figure 1. Correlation heatmap for dataset parameters

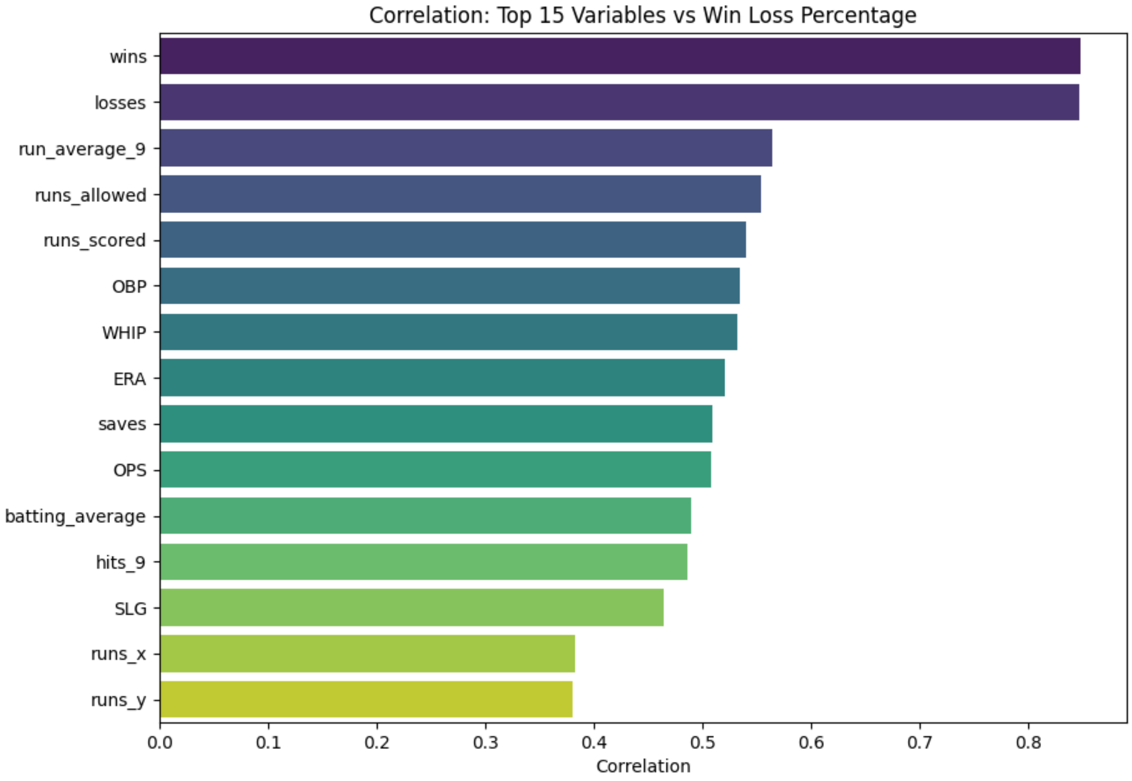


Figure 2. Specified heatmap for parameter selection

Regression Analysis to Understand Influence of Batting and Pitching Variables

I 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. For the pitching-related variables, the sum was approximately 1.42; this suggests that for a one-unit change in the normalized pitching-related features, we can expect an average change of approximately 1.42 in the win-loss percentage. Conversely, 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 can expect an average change of approximately 2.98 in the win-loss percentage.

I 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. For the pitching variables, the decrease in R-squared was approximately

0.417, indicating that about 41.7% of the variability in the win-loss percentage could be explained by these variables; for the batting variables, the decrease was approximately 0.284, meaning about 28.4% of the variability could be explained by these 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.

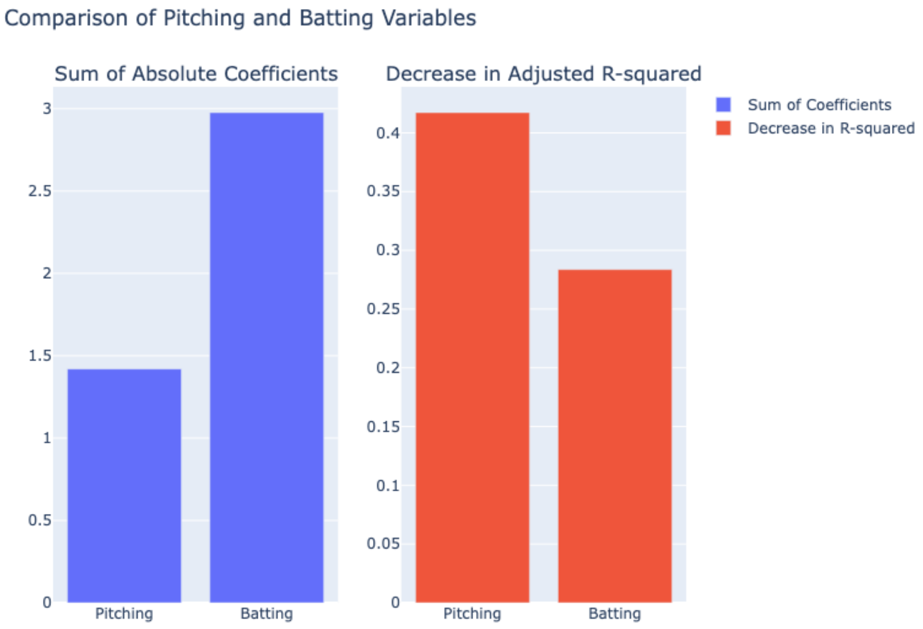


Figure 3. Comparative Influence and Explained Variability of Pitching and Batting Variables on Win-Loss Percentage

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 a linear combination of the original features; each component represented a certain amount of the total variance in the dataset.

From the correlation analysis, I identified KBO team features that is highly correlated with win ratio. This process first filtered out variables that has low impact on team's wins. 29 variables were left - representing the team's offensive and defensive capabilities - with 25 variables being filtered out. I decided number of components to retain in a PCA analysis by observing a scree plot and considering the cumulative explained variance. I settled on eight components as they captured about 89.2% of the variation in the dataset (See Figure panel 4).
After getting the PC loadings for team feature variables, I interpreted these loadings so that it represents the team's characteristic.

Here's an interpretation of each principal components:
Component 1: "Balanced Offense and Dominant Pitching"

- Offense:
    - Power Hitting: This component highlights a balanced offensive strategy with significant positive correlations for metrics like 'Total Bases', 'RBI', 'Hits by Batter', 'Doubles', 'Homeruns Scored', 'OPS', 'SLG', and 'Batting Average'. These may reflect a team that combines power hitting with consistent batting and run production.
    - Plate Discipline**:** The positive loadings on 'Bases on Balls', 'HBP', and 'OBP' may indicate a team that is patient at the plate and capable of getting on base through various means.
- Defense:
    - Dominant Presence: Positive correlations with 'batters faced', 'strikeouts by pitcher', 'innings pitched', 'homeruns_9', and 'ERA' may reflect a pitching style that emphasizes aggressiveness and dominance on the mound. This could include striking out batters and not shying away from challenging hitters.

Component 2: "Offensive Shortcomings and Aggressive Strike Zone Control "

- Offense:
    - Lack of Power and Production: The lack of correlations with traditional power-hitting metrics like 'homeruns scored', 'OPS', 'SLG', 'OBP', and 'total bases' may

    underline weaknesses in hitting for power, driving in runs, and producing extra-base hits. This may reflect a team that struggles with offensive production.
  - On-Base Struggles: The lack of positive correlations with 'bases on balls', 'batting average', and other on-base metrics could indicate difficulties in getting on base, further compounding offensive challenges.
- Defense:
  - Aggressive Approach: The positive correlation with 'hit batter' and negative correlation with 'walks_9' could signify a more aggressive pitching approach. This may include challenging hitters in the strike zone, potentially leading to hitting more batters but limiting walks.
  - Late-Game Execution: Saves are used as a metric to evaluate the effectiveness of a team's bullpen. Negative loadings for this component reflect struggles to close out games, indicating a lack of reliability in the team's relief pitching.

Component 3: "Contact-Oriented Offense and Pitching Adaptability"

- Offense:
  - Contact-First Philosophy: Positive loadings on 'batting average', 'SLG', 'OPS', and 'OBP' signify a focus on making contact and hitting for average, without necessarily emphasizing power, setting it apart from the other components.
  - Reduced Focus on Power: The limited emphasis on home runs ('homeruns scored') and the negative loading on 'bases on balls' contrasts with Component 2's offensive struggles, *highlighting a purposeful approach that doesn't rely on power or drawing walks.*
- Defense:
  - Adaptive Pitching: Negative correlations with 'innings pitched', 'batters faced', 'shutouts', and 'strikeouts' depict a pitching staff that doesn't rely on dominance but rather *focuses on adaptability and utilizing the defense*.
  - Strategic Balance: Unlike Components 1 and 2, this component may represent a team that focuses on balanced, flexible gameplay, with neither a dominant pitching staff nor a power-driven offense. The strategy may involve winning through consistent performance across all aspects of the game.

Component 4: "Strategic Small Ball and Grounded Pitching"

- Offense:
    - Small Ball Strategy: Emphasizes speed and agility with positive correlations to 'triples' and 'sacrifice_flies'. A focus on drawing walks ('bases_on_balls') and a reduced reliance on power hitting ('homeruns_scored', 'SLG') sets it apart from the power-centric offenses of other components.
- Defense:
    - Control Over Dominance: With negative correlations to 'strikeouts_x', and 'hit_batter', the pitching strategy seems focused on control and efficiency rather than dominance, possibly inducing ground balls and fly balls. This contrasts with the aggressive or efficient pitching styles of the other components.

Component 5: "Tactical Offense and Disciplined Pitching"

- Offense:
    - Tactical Execution: Focuses on small ball tactics ('sacrifice_hits', 'triples', 'doubles') and avoids power hitting ('homeruns_scored', 'OPS'). This contrasts with the aggressive or balanced offenses in other components.
- Defense:
    - Controlled Approach: Emphasizes control and limiting damage ('walks_9', 'HBP', 'homeruns_9'), differing from the dominant or aggressive pitching in other components.
    - Balanced Strategy: Balanced approach prioritizing situational tactics and disciplined pitching, including a minor focus on defense with 'GDP'.

Component 6: "Defensive Dominance and Team Defense"

- Defense:
    - Defensive Dominance: Emphasizes shutouts and limiting home runs ('shutouts', 'homeruns_9'), setting it apart with a focus on keeping opponents scoreless.
    - Cautious Approach: Contrasts with other components by minimizing hitting batters and accepting a higher walk rate.

    - Reliance on Team Defense: Focuses on defensive strategies like double plays ('GDP'), and less on traditional closers ('saves').

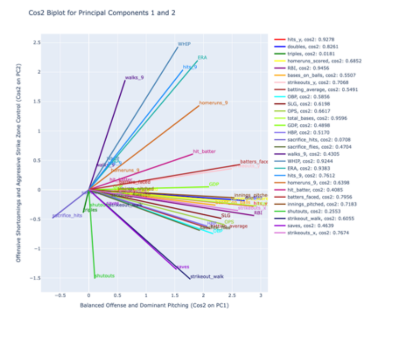
Component 7: "Risk-Embracing Pitching and Power Hitting"

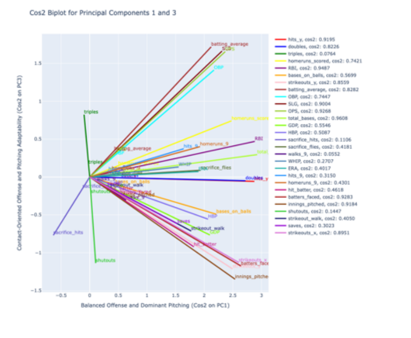
- Offense:
    - Aggressive Pitching: Positive correlations with 'saves', 'walks_9', 'HBP', and 'hit_batter' contrast with other components by highlighting an aggressive, risk-taking pitching style that yields effectiveness ('negative correlations with 'hits_9', 'ERA', 'WHIP').
    - Power over Small Ball: Focuses on power hitting ('homeruns_scored', 'SLG') rather than small ball ('negative correlation with 'sacrifice_hits'), indicating a preference for runs through hits, not sacrifices.
- Defense:
    - Controlled Aggression: Though walks may occur ('positive correlation with 'walks_9'), negative correlations with 'bases_on_balls' and 'strikeout_walk' suggest a controlled strategy, not a lack of control.

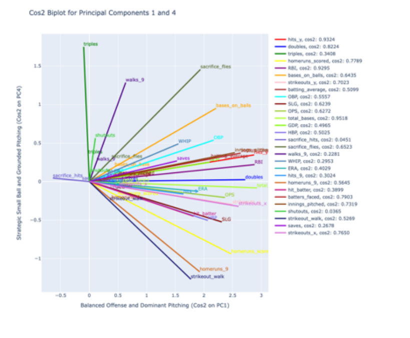
Component 8: "Robust Bullpen and Balanced Offense"

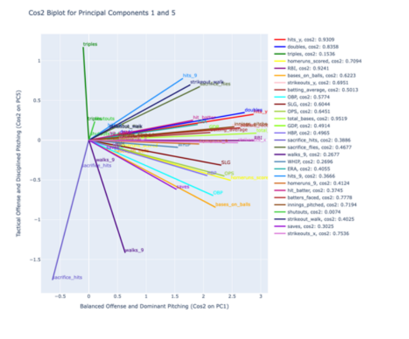
- Offense:
    - Balanced Offense: Prefers a tactical approach over power hitting ('negative correlations with 'batting_average', 'OPS', 'SLG', 'triples'), setting it apart from power or small ball-focused components.
- Defense:
    - Defensive Excellence: Emphasizes effective defense ('negative correlations with 'shutouts', 'hit_batter', 'HBP') and the ability to execute double plays ('positive correlation with 'GDP'), highlighting a strategy different from more aggressive or offensive-centric components.
    - Relief Pitching: Focuses on bullpen effectiveness and game closing ('positive correlation with 'saves'), underscored by controlled pitching ('negative correlation with 'walks_9', 'WHIP').

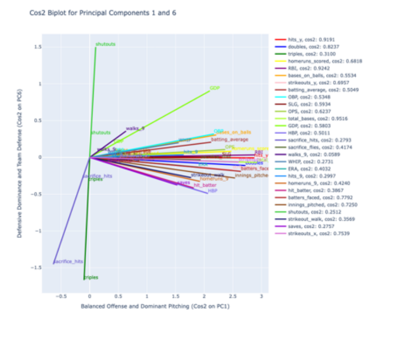
To understand the relationship between the original variables and the derived PCA components, a cos2 biplot was generated. This visualization maps the loading of each variable onto the principal components, highlighting their contributions to the components. The x, and y axis were set as PC component interpretation from my analysis.

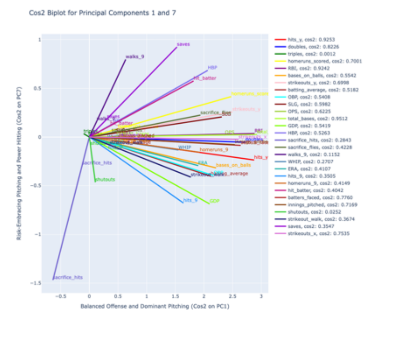

Cos2 Biplot for Principal Components 1 and 2

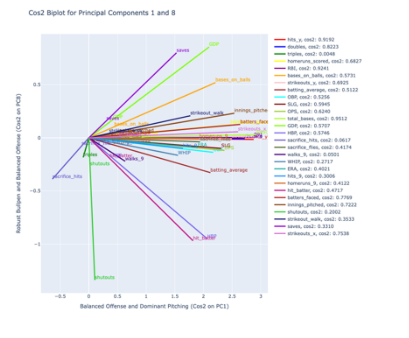

Cos2 Biplot for Principal Components 1 and 3

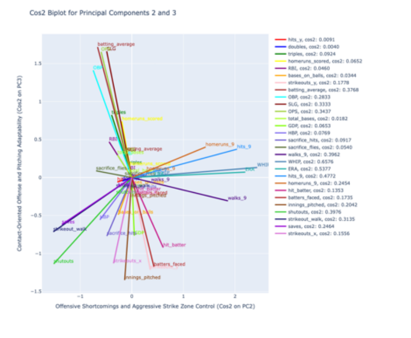

Cos2 Biplot for Principal Components 1 and 4

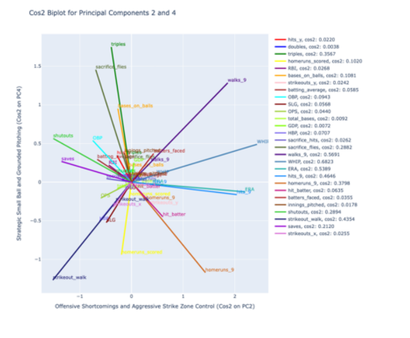

Cos2 Biplot for Principal Components 1 and 5

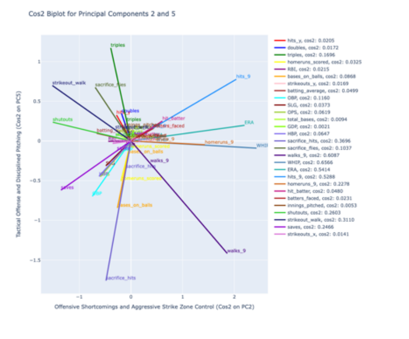

Cos2 Biplot for Principal Components 1 and 6

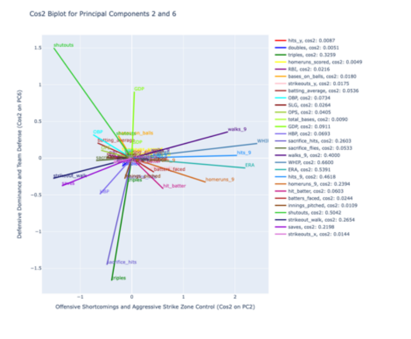

Cos2 Biplot for Principal Components 1 and 7

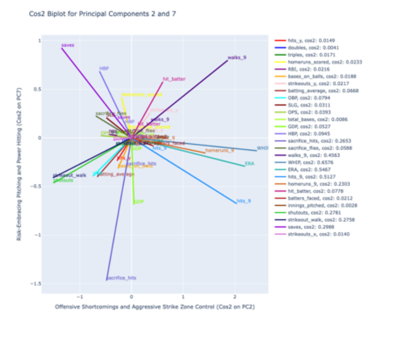

Cos2 Biplot for Principal Components 1 and 8

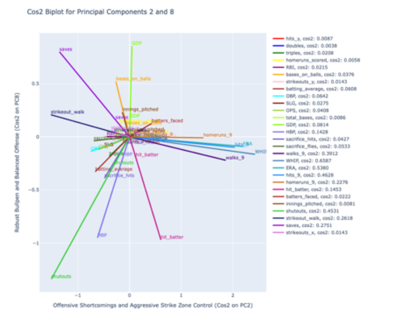

Cos2 Biplot for Principal Components 2 and 3

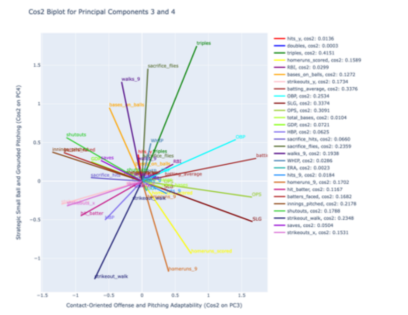

Cos2 Biplot for Principal Components 2 and 4

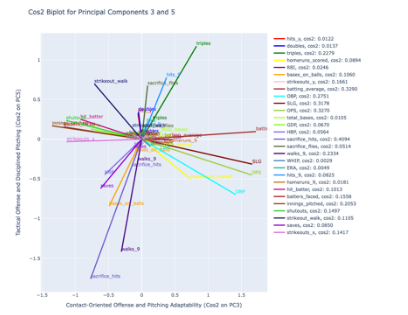

Cos2 Biplot for Principal Components 2 and 5

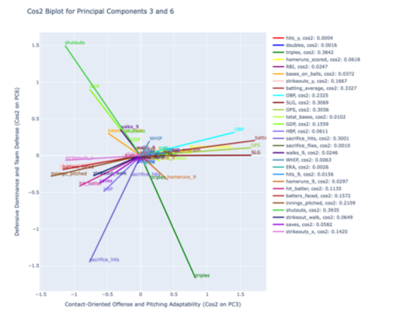

Cos2 Biplot for Principal Components 2 and 6

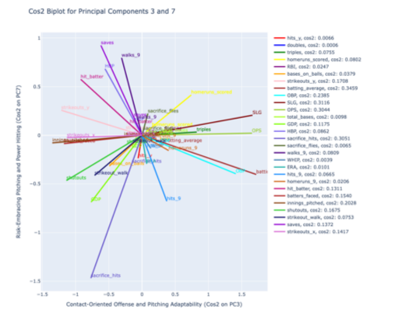

Cos2 Biplot for Principal Components 2 and 7

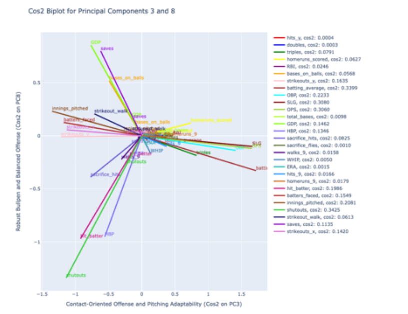

Cos2 Biplot for Principal Components 2 and 8

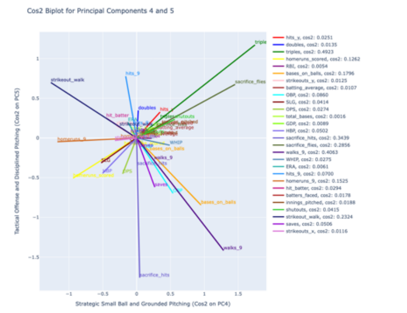

Cos2 Biplot for Principal Components 3 and 4

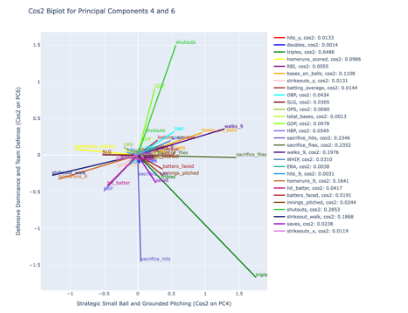

Cos2 Biplot for Principal Components 3 and 5

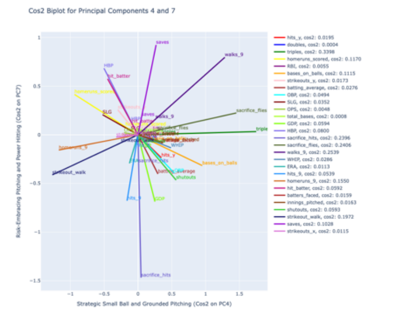

Cos2 Biplot for Principal Components 3 and 6

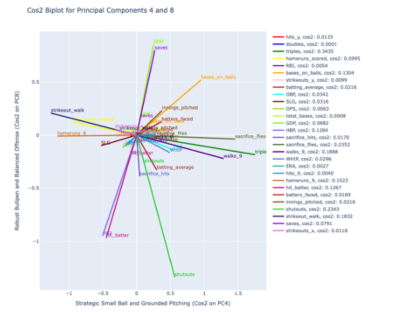

Cos2 Biplot for Principal Components 3 and 7

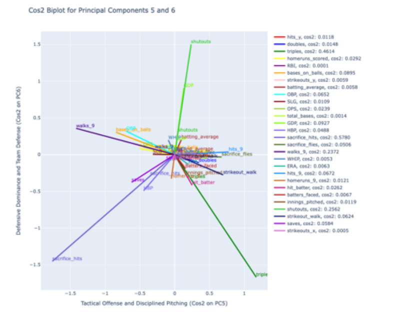

Cos2 Biplot for Principal Components 3 and 8

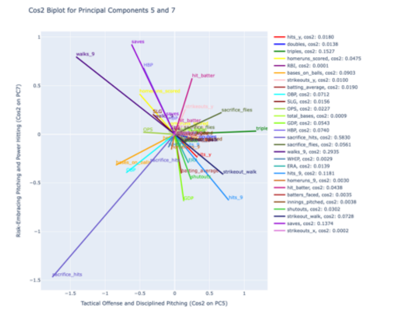

Cos2 Biplot for Principal Components 4 and 5


Cos2 Biplot for Principal Components 4 and 6


Cos2 Biplot for Principal Components 4 and 7


Cos2 Biplot for Principal Components 4 and 8


Cos2 Biplot for Principal Components 5 and 6


Cos2 Biplot for Principal Components 5 and 7

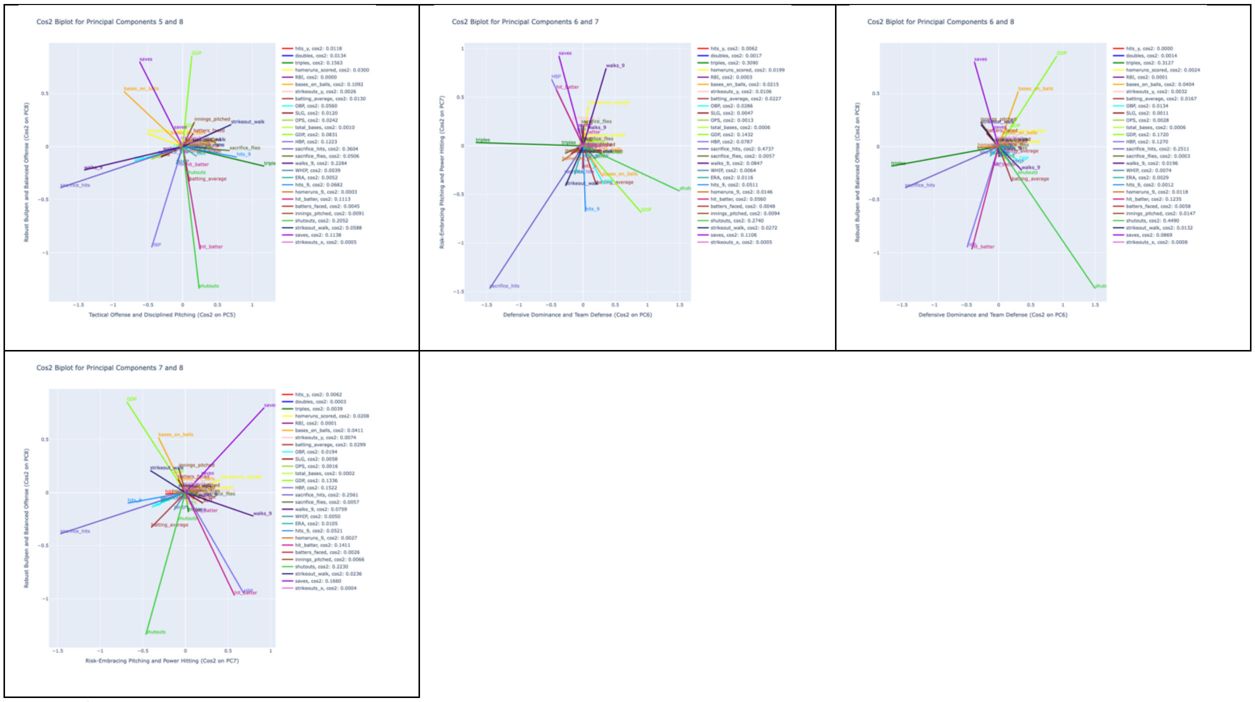


Figure Panel 4. ***Cos2 biplot from every pair of PC component interpretation set as axis:***

'Balanced Offense and Dominant Pitching',

'Offensive Shortcomings and Aggressive Strike Zone Control',

'Contact-Oriented Offense and Pitching Adaptability',

'Strategic Small Ball and Grounded Pitching',

'Tactical Offense and Disciplined Pitching',

'Defensive Dominance and Team Defense',

'Risk-Embracing Pitching and Power Hitting',

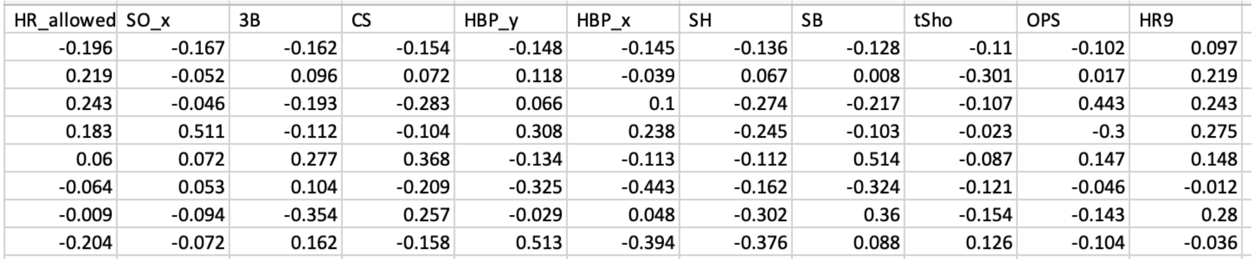
'Robust Bullpen and Balanced Offense'

Summary Interpretation of PC1 vs PC3

Alternatively, a team emphasizing the "Aggressive Pitching" dimension may focus on robust defensive strategies, such as:

1. ***Emphasis on Strikeouts****: With variables like* ***strikeouts*** *(cos^2 = 0.8951), this team likely has pitchers who prioritize striking out opponents, limiting their ability to get on base.*
2. ***Controlled Pitching Stats****: Metrics such as* ***WHIP*** *(cos^2 = 0.2707),* ***ERA*** *(cos^2 = 0.4017), and* ***homeruns_9*** *(cos^2 = 0.4301) suggest a controlled pitching environment where hits and runs are minimized.*
3. ***Innings Pitched and Shutouts****: Strong values in* ***innings pitched*** *(cos^2 = 0.9184) and* ***shutouts*** *(cos^2 = 0.1447) may indicate effective use of pitchers over the course of a game and a focus on completely shutting down the opposing offense.*

The analysis of principal components within the KBO data uncovers a multitude of strategies that teams may embrace, ranging from a balanced offense to masterful defense. Uniquely, Components 3, 4, and 5 illuminate specific aspects of success in the KBO league, concentrating on the art of small ball and astute offense. Together, they account for nearly 15% of the total variance, marking a significant departure from prevailing MLB tactics that often favor power hitting and specialized pitching roles.



HR_allowed	SO_x	3B	CS	HBP_y	HBP_x	SH	SB	tSho	OPS	HR9
-0.196	-0.167	-0.162	-0.154	-0.148	-0.145	-0.136	-0.128	-0.11	-0.102	0.097
0.219	-0.052	0.096	0.072	0.118	-0.039	0.067	0.008	-0.301	0.017	0.219
0.243	-0.046	-0.193	-0.283	0.066	0.1	-0.274	-0.217	-0.107	0.443	0.243
0.183	0.511	-0.112	-0.104	0.308	0.238	-0.245	-0.103	-0.023	-0.3	0.275
0.06	0.072	0.277	0.368	-0.134	-0.113	-0.112	0.514	-0.087	0.147	0.148
-0.064	0.053	0.104	-0.209	-0.325	-0.443	-0.162	-0.324	-0.121	-0.046	-0.012
-0.009	-0.094	-0.354	0.257	-0.029	0.048	-0.302	0.36	-0.154	-0.143	0.28
-0.204	-0.072	0.162	-0.158	0.513	-0.394	-0.376	0.088	0.126	-0.104	-0.036

Table 5. PCA Loadings on MLB data

In a parallel analysis conducted over three decades of MLB team data (as presented in Table 5), a contrasting pattern emerges. The primary focus within MLB's major PCA components centers on a twofold strategy: on one hand, suppressing home runs when on defense, and on the other, aggressively scoring them when on offense. This is exemplified by a loading of -0.196 on Component 1, explaining 17% of the variance for home runs allowed, and a 0.219 loading on Component 2, accounting for 12% of the variance for Homeruns per 9 innings.
This comparative exploration highlights the distinct characteristics and philosophies between the KBO and MLB leagues. While the MLB leans heavily on powerful hitting and a robust defense against home runs, the KBO embraces more nuanced and diverse pathways to victory, including a more strategic and adaptable approach to the game. These findings not only underscore the individuality of the KBO league but also contribute valuable insights into the underlying structures that shape winning teams within the league.

The subsequent analysis focused on KBO batting features, utilizing Principal Component Analysis (PCA) and Gaussian Mixture Model (GMM) clustering. By observing a scree plot, I decided to retain six components in the PCA, capturing about 92.3% of the dataset's variation, thereby condensing 16 variables into six orthogonal components (See Figure 5). This selection made the subsequent GMM clustering more manageable and insightful.

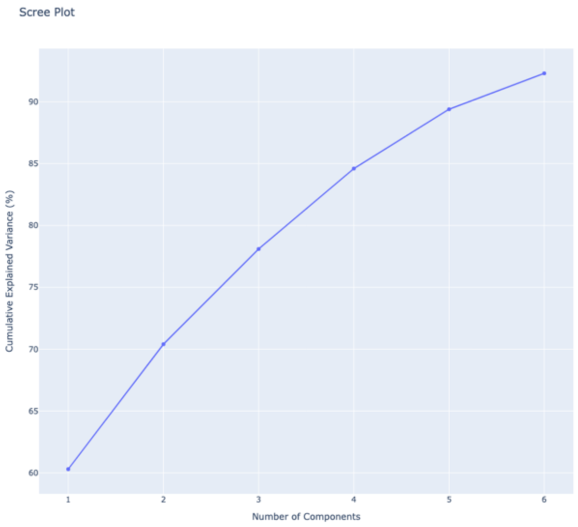


Figure 4. Scree Plot Illustrating the Optimal Number of Principal Components and Explained Variance for PCA Analysis for KBO batting features.

The first stage identified the most significant components of team performance, providing a condensed view of influential batting metrics. Deciding on the optimal number of clusters for the GMM was crucial. By examining the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC), I determined the most suitable number of clusters for the data (Figure 6). Teams were then assigned to clusters based on Gaussian distribution alignment. The results of the PCA-GMM analysis, including unique PCA component loadings and final cluster assignments, were visually depicted through scatter plots and histograms (Figures 7 and 8). These visual representations allowed me to delve into the traits that define 'winning' KBO teams, leading to the identification of distinct clusters marked by specific traits such as "Offensive Powerhouses," "Disciplined and Fast Teams," and others. These designations were derived from the PCA loadings, underscoring how each batting metric contributed to performance factors.

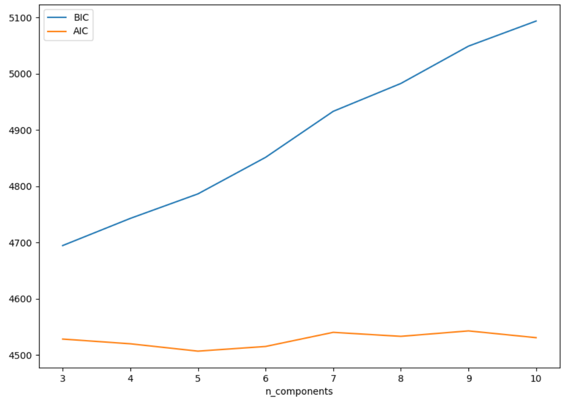


Figure 5. Comparative Analysis of BIC and AIC Metrics to Determine the Optimal Number of Clusters for GMM Clustering

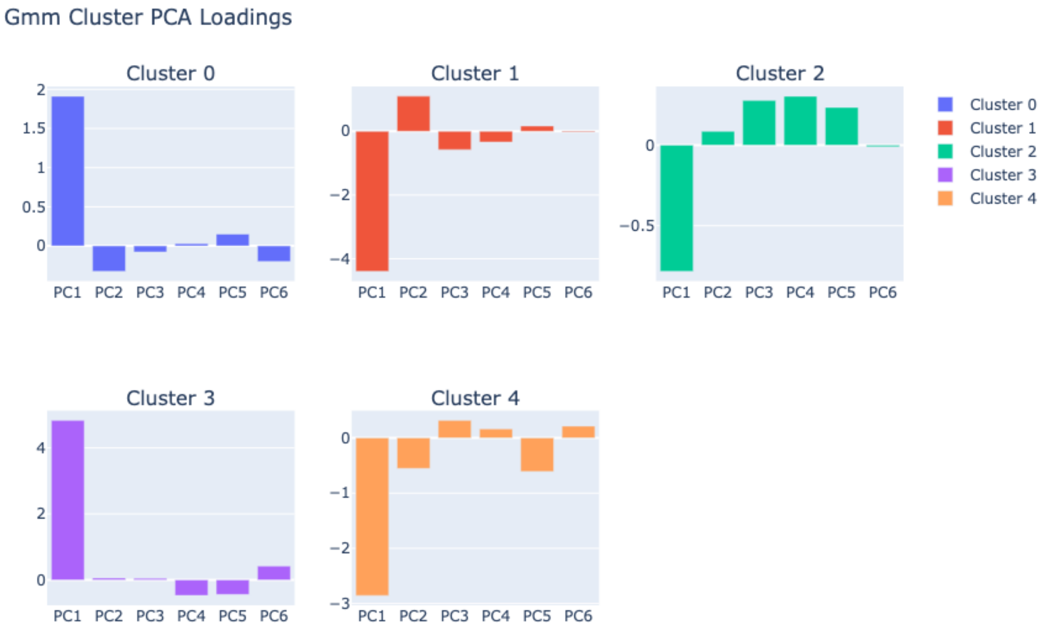


Figure 6. Presentation of Principal Component Analysis (PCA) Loadings for Clusters Identified via Gaussian Mixture Model (GMM)

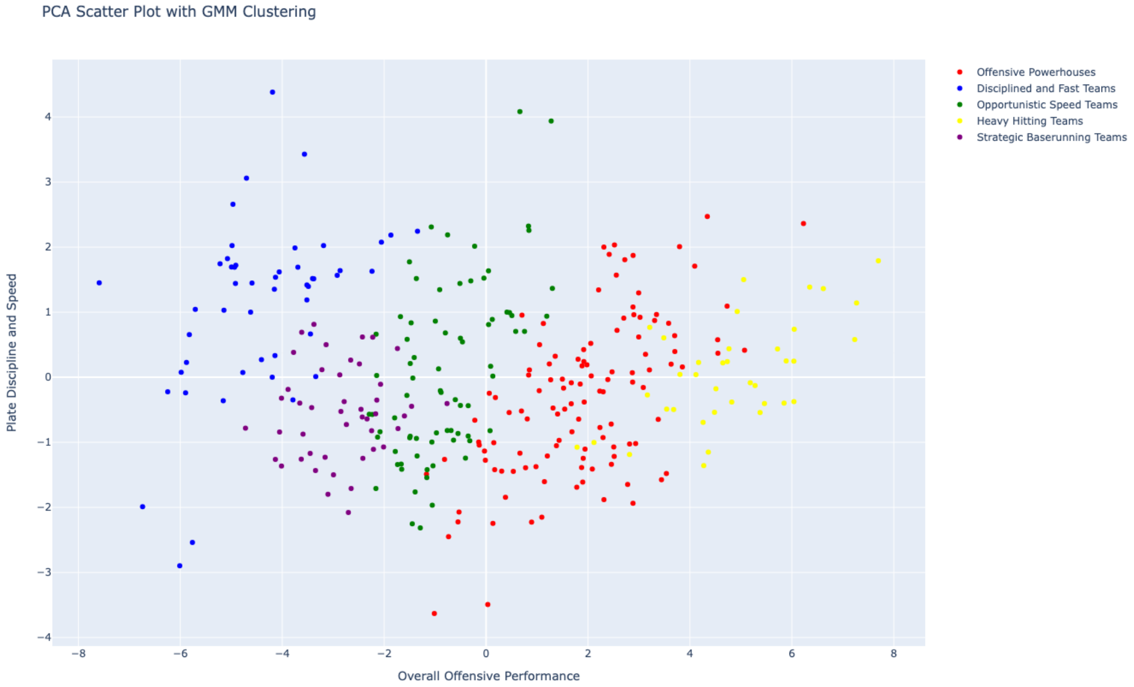


Figure 7. Scatter Plot Representation of Principal Component Analysis (PCA) Clusters, Generated using a Gaussian Mixture Model (GMM), for Korea Baseball Organization (KBO) Batter Data

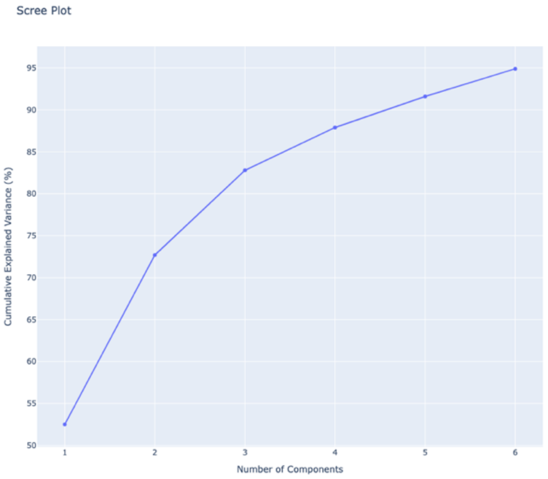


Figure 8. Scree Plot Illustrating the Optimal Number of Principal Components and Explained Variance for PCA Analysis for KBO pitching features.

Turning our attention to pitcher metrics within the KBO, I employed a scree plot to select five components that encompass 91.5% of the variance, with the first component strongly reflecting a pitcher's workload and effectiveness (See Figure 9). By analyzing the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC), I determined three optimal components for pitching variables (Figure 11).

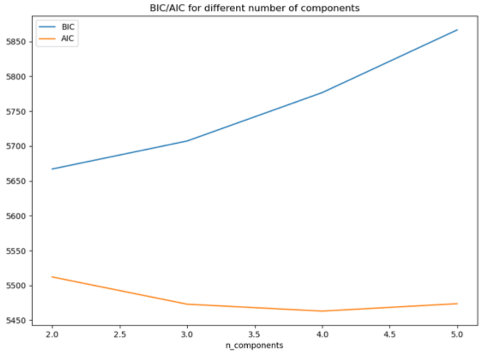


Figure 9. Comparative Analysis of BIC and AIC Metrics to Determine the Optimal Number of Clusters for GMM Clustering for KBO pitching features

Teams were assigned to clusters based on their alignment with the closest Gaussian distribution. The results of the PCA-GMM analysis, including unique PCA loadings and final cluster assignments, are visually presented through scatter plots and histograms (Figures 11, 12).

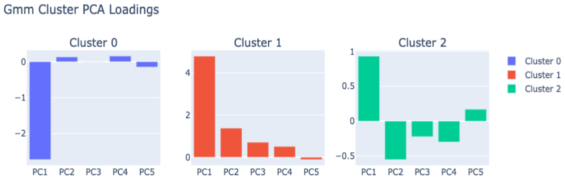


Figure 10. Presentation of Principal Component Analysis (PCA) Loadings for Clusters Identified via Gaussian Mixture Model (GMM) for pitching features.

The analysis uncovered distinct team clusters, represented by specific traits:

1. **Cluster 0: "Struggling Pitching Performance."** With the highest negative value for Component 1, "General Pitching Performance," this cluster indicates poor performance in hits allowed, facing batters, home runs allowed, innings pitched, and higher ERA.
2. **Cluster 1: "Highly Controlled Efficiency."** Exhibiting the highest positive value for Component 2, "Control and Efficiency," this cluster reveals outstanding control and efficiency, leading to fewer walks and lower WHIP.
3. **Cluster 2: "Mixed Performance."** Displaying mixed results across components, this cluster neither outperforms nor underperforms distinctly, justifying its name as "Mixed Performance."

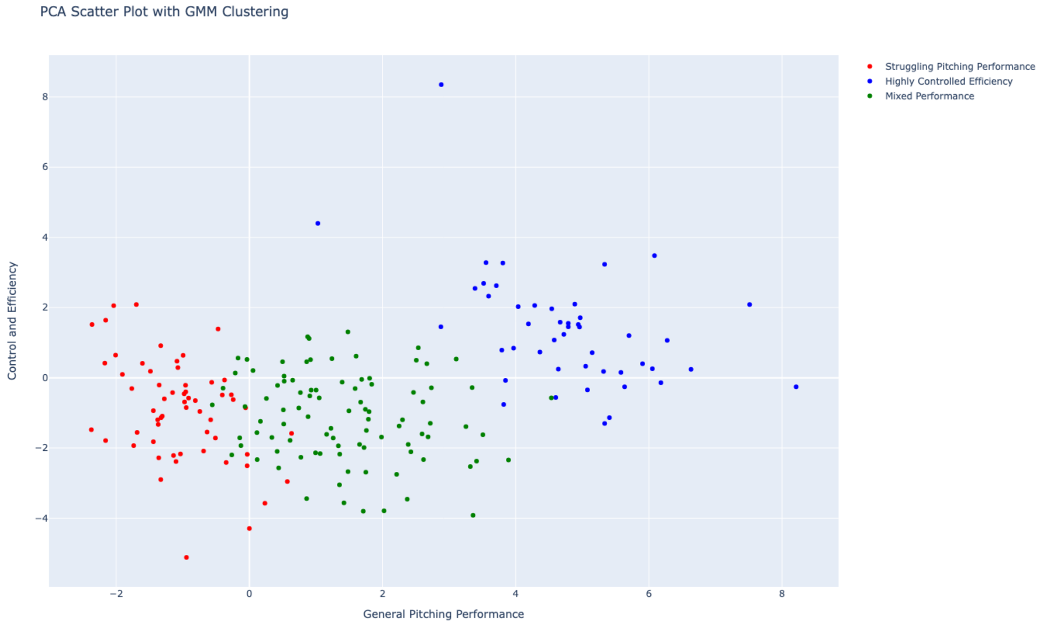


Figure 11. Scatter Plot Representation of Principal Component Analysis (PCA) Clusters, Generated using a Gaussian Mixture Model (GMM), for Korea Baseball Organization (KBO) Pitcher Data

These three components capture essential characteristics of pitching performance, simplifying the dataset for further analyses such as clustering.

By interpreting the PCA-GMM clustering, this analysis offers an enlightening view into the elements that define success in the KBO's pitching realm. The insights gleaned can guide teams in refining strategies to improve their performance, shedding light on key areas for enhancement.

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 batting data and total runs allowed being the target feature for pitcher data.

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.

Analyzing pitching statistics, a notable interaction of 1.91 emerged between 'ERA' (Earned Run Average) and 'batters_faced', suggesting a pitcher's ERA tends to rise with the number of batters faced. This likely stems from increased scoring opportunities for the opposition. We also detected significant interactions between 'hits_x' and 'ERA', 'home_runs_allowed' and 'ERA', and 'walks' and 'home_runs_allowed', demonstrating how different pitching scenarios and strategies can impact game results. For example, allowing more walks can indirectly result in more home runs, underlining the importance of control in a pitcher's skills. The SHAP interaction plots provide a visual representation of these dynamics and can be found in Figure 15.

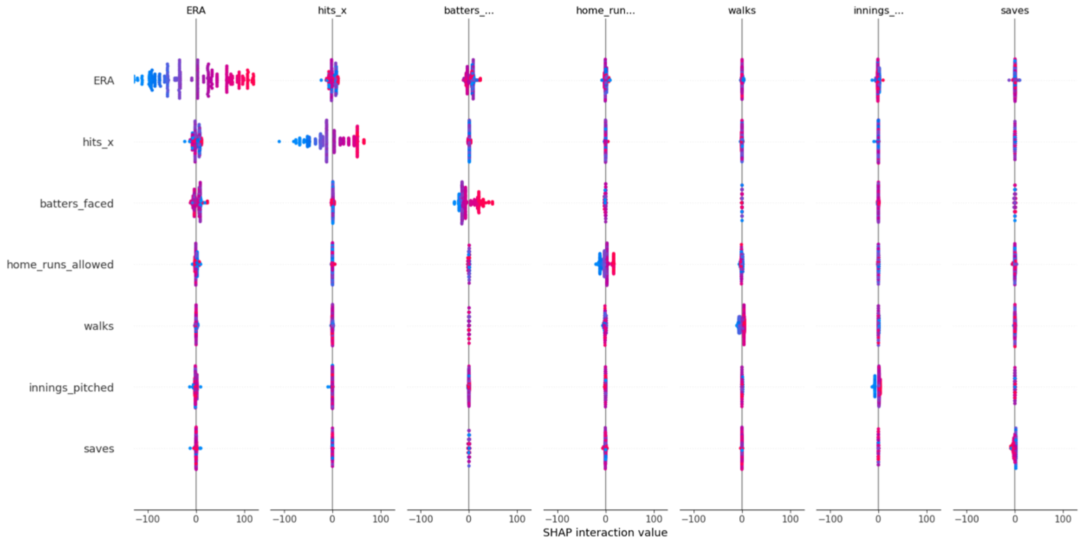


Figure 12. Summary of Interaction plot of KBO pitcher features

Focusing on batting features, we discovered a coefficient of 0.32 between 'homeruns_scored' and 'doubles', suggesting that teams proficient in doubles often score more home runs. This could point to a synergistic relationship between these two hitting strategies. A detailed summary plot of SHAP interaction for batting features is displayed in Figure 14.

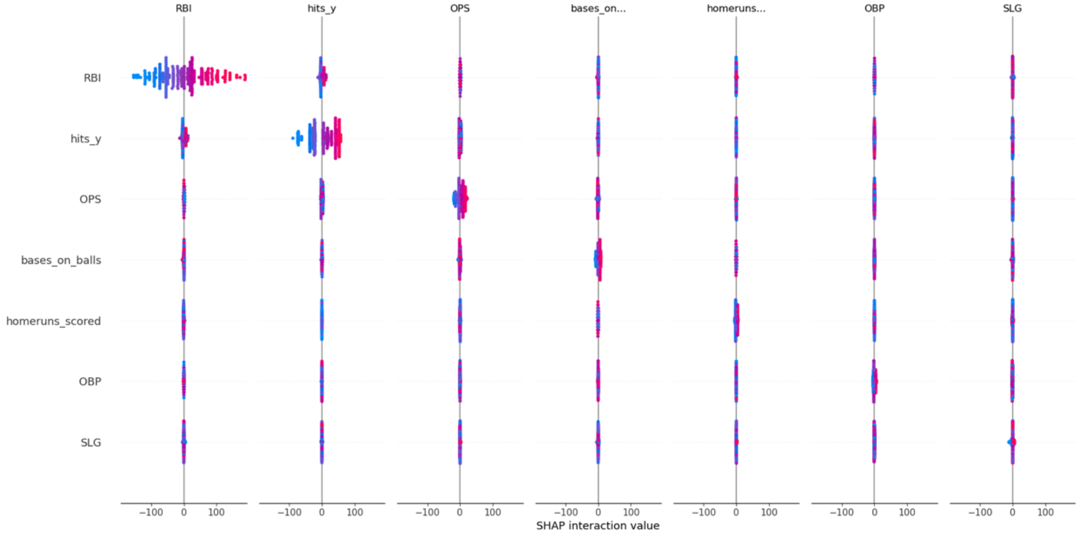


Figure 13. Summary of Interaction plot of KBO batter features

In summary, the significant interaction effects observed across both batting and pitching features offer insightful perspectives into KBO league dynamics. By integrating these insights into future predictive models, we may enhance their performance, providing a more detailed understanding of the factors that contribute to a team's success.

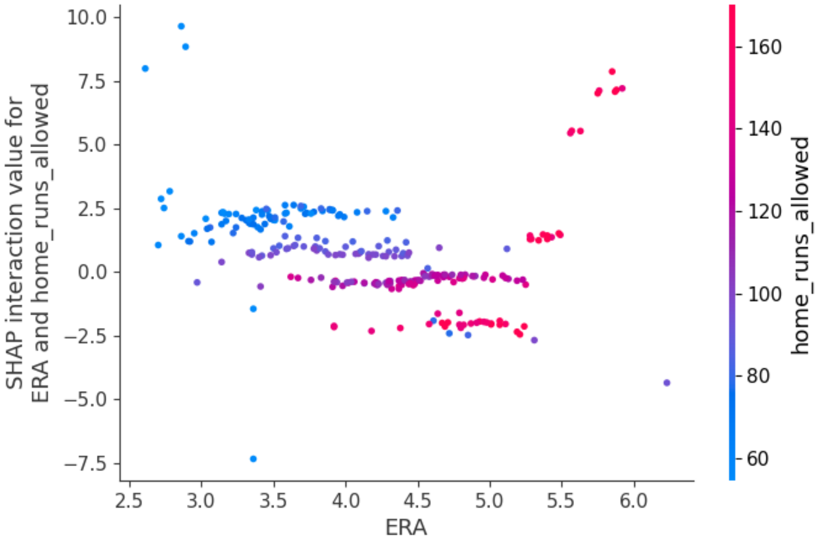
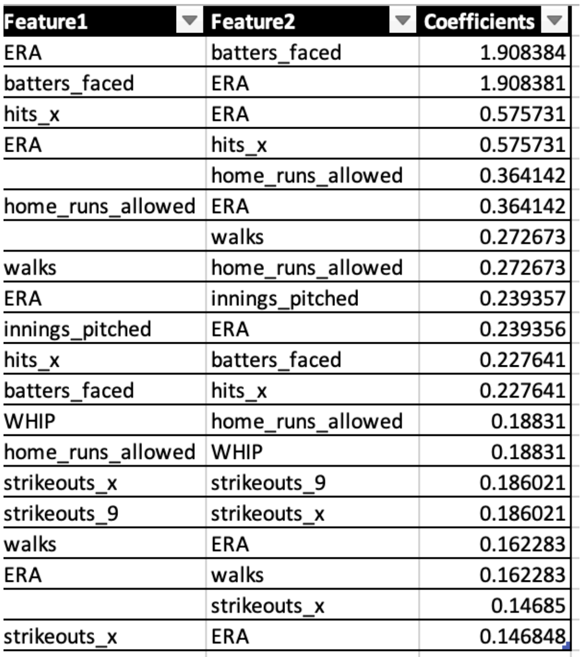
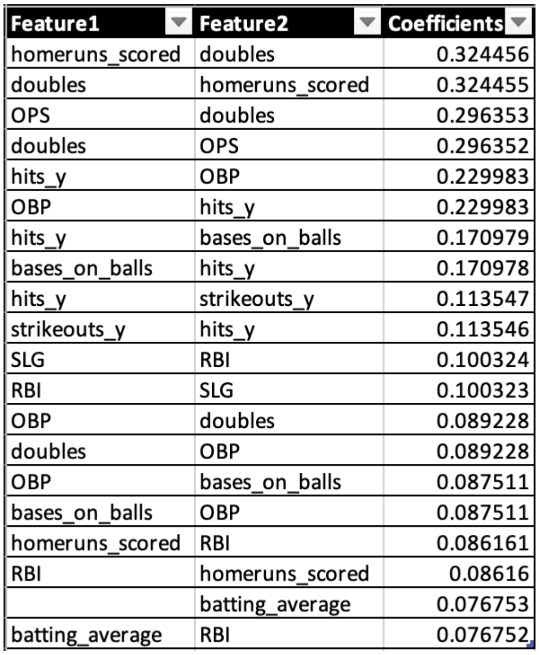


Figure 14. Detailed view of interaction between ERA and homeruns allowed.



Feature1	Feature2	Coefficients
ERA	batters_faced	1.908384
batters_faced	ERA	1.908381
hits_x	ERA	0.575731
ERA	hits_x	0.575731
	home_runs_allowed	0.364142
home_runs_allowed	ERA	0.364142
	walks	0.272673
walks	home_runs_allowed	0.272673
ERA	innings_pitched	0.239357
innings_pitched	ERA	0.239356
hits_x	batters_faced	0.227641
batters_faced	hits_x	0.227641
WHIP	home_runs_allowed	0.18831
home_runs_allowed	WHIP	0.18831
strikeouts_x	strikeouts_9	0.186021
strikeouts_9	strikeouts_x	0.186021
walks	ERA	0.162283
ERA	walks	0.162283
	strikeouts_x	0.14685
strikeouts_x	ERA	0.146848

Figure 15. KBO Pitcher Feature Interaction Coefficients



Feature1	Feature2	Coefficients
homeruns_scored	doubles	0.324456
doubles	homeruns_scored	0.324455
OPS	doubles	0.296353
doubles	OPS	0.296352
hits_y	OBP	0.229983
OBP	hits_y	0.229983
hits_y	bases_on_balls	0.170979
bases_on_balls	hits_y	0.170978
hits_y	strikeouts_y	0.113547
strikeouts_y	hits_y	0.113546
SLG	RBI	0.100324
RBI	SLG	0.100323
OBP	doubles	0.089228
doubles	OBP	0.089228
OBP	bases_on_balls	0.087511
bases_on_balls	OBP	0.087511
homeruns_scored	RBI	0.086161
RBI	homeruns_scored	0.08616
	batting_average	0.076753
batting_average	RBI	0.076752

Figure 16. KBO Batter Feature Interaction Coefficients

Lasso Regression for variable selection

To enhance the accuracy of our predictive model for the target variable 'Win-Loss Probability', a Lasso regression technique was deployed for effective variable selection. Lasso regression, a regularization method, is particularly useful in scenarios like ours due to its inherent capability to shrink the coefficients of less important variables towards zero. This essentially simplifies the model while retaining the influential variables, larger coefficients of which imply a greater impact on the outcome.

Starting with a total of 54 KBO independent features, I utilized dimensionality reduction techniques to refine these down to 18 key features (See Figure 19). An intriguing finding from this process was the prevalence of batting features among those with the highest absolute Lasso coefficients. This concentration suggests that batting has a more significant impact on team victories in the KBO than pitching does.

To identify these key features, I employed Lasso regularization, which minimizes the sum of the absolute values of the coefficients. This approach shrinks some coefficients to zero, effectively performing feature selection by keeping only the most influential features in the model. The selected features, which had the highest absolute coefficients, included both batting and pitching features. The former category suggests a team's offensive strength, while the latter refers to the team's defensive abilities.

After establishing these optimal features, I then employed findings from PCA and interaction analysis to synthesize these features, aiming to enhance our model's predictive power. For example, from the PCA, we found significant components from both batting and pitching features, which were named based on their characterizing qualities.

From the pitching side, we had "General Pitching Performance", "Control and Efficiency", "Walks and Home Runs Allowed", and "Shutouts vs. Saves", “Aggressiveness and Save Situations”. It was observed that 'home_runs_allowed' and 'ERA', both indicative of the third component, interacted to affect the team's performance. Thus, they were combined into a single feature to directly compare the fraction of earned runs due to home runs.

Similarly, from the batting side, we found ‘Sacrifice Play’, ‘Aggressive Play and Efficiency’, ‘Walks and Home Runs Allowed’.

'OBP', 'SLG', and 'OPS' features, integral to "Overall Offensive Performance" and demonstrating high correlation with one another, were integrated into a single feature. **However, to avoid overrepresentation of this information in the model, the differences: OPS - (OBP + SLG) was used instead.** This approach captures the essence of a team's overall batting performance more accurately.

Finally, 'walks_9' and 'strikeout_walk', core features from the second pitching component, "Control and Efficiency", were combined due to their high level of interaction. A feature like strikeout_walk / walks_9 was created.

Through this approach, we created composite features such as (OBP, SLG, OPS), and (walks_9, strikeout_walk), which integrated the knowledge from our PCA and interaction analysis. These composite features, which encapsulate the most salient aspects of both batting and pitching performance, are expected to contribute to a more effective predictive model for KBO team wins.

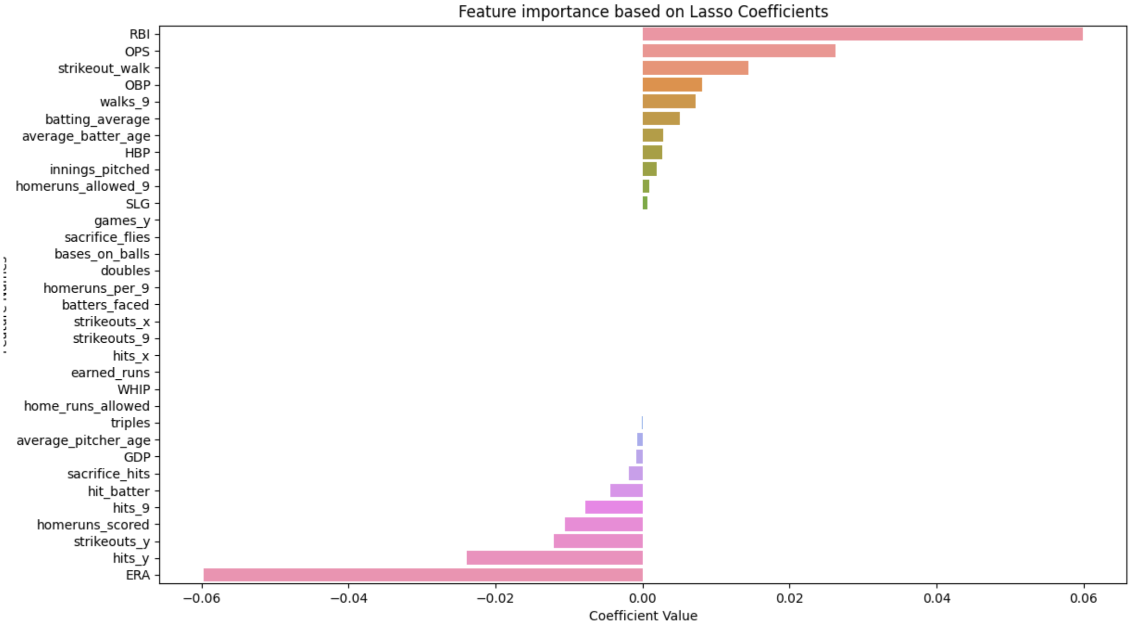


Figure 17. Lasso coefficient values for kbo team features

Ridge Regression

To address potential multicollinearity issues in our data, we employed Ridge Regression, a model that introduces a degree of bias into the regression estimates, thereby stabilizing them. Ridge Regression is suitable for this study, where composite features, integrated from prior analyses using Principal Component Analysis (PCA) and interaction analysis, are used to construct the model. Given that these composite features potentially contain information from overlapping domains, Ridge Regression gives more reliable interpretation of their relationships with the dependent variable.

Below is a more detailed description of how these composite features were combined.

1. **OPS_difference**: On-base percentage (OBP), Slugging percentage (SLG), and On-base Plus Slugging (OPS) are all advanced metrics that capture different aspects of offensive performance. OPS is already a combination of OBP and SLG, so adding them all together again may not be beneficial and could give extra weight to OBP and SLG. So I took an approach to look at the difference between OPS and the sum of OBP and SLG, which might capture how much 'extra' a team is getting beyond just the sum of their on-base and slugging performance.

'OPS_difference' = 'OPS' - 'OBP' + 'SLG'

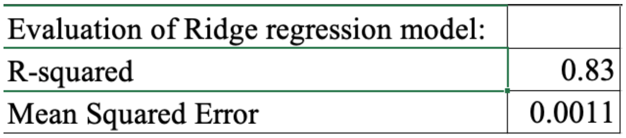
2. **Normalized strikeout walk**: These are two measures of a pitcher's control. walks_9 is the number of walks allowed per 9 innings, and strikeout_walk is the ratio of strikeouts to walks. Adding these together might not make much sense because they're different types of measures (a rate and a ratio).So I combined these to create a composite metric like strikeout_walk divided by walks_9, which could give a sense of a team's control relative to how many walks they give up.

   'Normalized strikeout walk' = 'strikeout walk' / 'walks 9'

3. **Unproductive outs**: In baseball, both strikeouts and grounding into double plays (GDP) are generally considered unproductive outs, as they do not contribute to advancing runners already on base. A high number of strikeouts indicates a failure to put the ball in play, while a high GDP indicates that when the ball is put into play, it often leads to more than one out. Consequently, adding strikeouts and GDPs together can provide an aggregate measure of 'unproductive outs'.

   'Unproductive outs' = 'GDP' + 'Strikeouts'

The model parameters selected for the ridge regression analysis included innings_pitched, hit_batter, hits_9, OPS_difference, homeruns_allowed_9, ERA, unproductive_outs, average_batter_age, hits_y, triples, normalized_strikeout_walk, homeruns_scored, RBI, HBP, and sacrifice_hits.



Evaluation of Ridge regression model:	
R-squared	0.83
Mean Squared Error	0.0011

Figure 18. Table Summary of Ridge Regression

As shown in the Figure 20, The MSE value of 0.0011 indicates that, on average, the predicted win-loss percentage from the ridge regression model deviates from the actual win-loss percentage by 0.0011 units. Furthermore, the R-squared value of 0.83 suggests that the ridge regression model can explain approximately 83% of the variation in win probability for a team.

Based on these results, it can be inferred that the ridge regression model is a strong predictor for the KBO dataset. The low MSE value indicates that the model's predictions closely align with the actual win-loss percentages, while the high R-squared value signifies that the model captures a significant portion of the variability in win probability.

Pythagorean Expectation Modeling

In this study, a Pythagorean Expectation model was developed using a multi-polynomial regression approach to predict the win-loss percentages in the Korea Baseball Organization (KBO). This model, although simple and based on run differentials, proved to be impressively robust, achieving an adjusted R-squared value of 0.885. This suggests that the model can account for approximately 88.5% of the variation in win-loss percentages. Moreover, it demonstrated a high level of accuracy, as evidenced by its Mean Squared Error (MSE) of 0.00087, indicating the model's predictions deviated from the actual win-loss percentage by about 0.00087 on average.

Conclusion

My in-depth investigation using various statistical techniques and machine learning models has revealed critical insights into the principal factors impacting win-loss percentages in the Korea Baseball Organization (KBO). I commenced the research with comprehensive preprocessing and parameter selection for my dataset comprising 54 independent variables. A robust cleaning process was applied to ensure data quality, and preliminary correlations were assessed to facilitate the initial understanding of the relationships among variables. Subsequently, I conducted a regression analysis to discern the influence of batting and pitching variables on team victories. Following the regression analysis, I employed Principal Component Analysis (PCA) coupled with Gaussian Mixture Models. PCA transformed the refined feature set into orthogonal components, enabling me to reveal the underlying structure and relationships within the data. Key components such as "General Pitching Performance", "Control and Efficiency", "Sacrifice Play", and "Aggressive Play and Efficiency" were identified in the pitching and batting dimensions, respectively. To further optimize and examine the interaction of batting and pitching variables, I implemented XGBoost models. The results from these models led to the creation of composite features that integrated the most influential aspects of batting and pitching performances. This step effectively encapsulated insights from both PCA and interaction analysis, laying the groundwork for subsequent modeling efforts. To refine the feature set further and enhance model simplicity, I employed Lasso Regression for variable selection. The technique narrowed down the features to the most influential 18. From these features, composite features were selected based on the previous analysis and reduced the dimensionality of the model. Building upon these findings, I constructed a Ridge Regression model that displayed impressive predictive power, evidenced by a mean square error (MSE) of 0.0011 and an R-squared value of 0.83. These metrics indicated a strong alignment between the model's predictions and actual win-loss percentages and an ability to explain approximately 83% of the variation in a team's win probability. Alongside Ridge Regression, I also developed a Pythagorean Expectation model. This model, despite its simplicity, demonstrated significant robustness. It explained about 88.5% of the variation in win-loss percentages and maintained high accuracy, as indicated by its low MSE of 0.0009. Collectively, my findings underscore the value of rigorous data preprocessing, thoughtful feature selection, and sophisticated modeling approaches in uncovering the determinants of team success in KBO. These insights can be utilized to aid team strategists in focusing their efforts on enhancing key areas that contribute to improved win-loss percentages. Future research should consider delving deeper into the identified composite features, incorporating additional variables such as team payroll or fan attendance, or widening the scope of analysis to other baseball leagues to facilitate a comparative analysis of factors influencing team success.