The Components of MLB Success

Defining How Stats Correlate to Wins in the Majors

Nolan Lowe
School of Management
Wentworth Institute of Technology
Boston, MA, USA
lowen1@wit.edu

ABSTRACT

The following report details an investigation into the relationships between various MLB stats (divided into the areas of batting, pitching, and fielding) and winning. It aims to determine which of these areas and which of the considered statistics are most highly correlated with winning and how these relationships have changed over time. The determination of the correlations was aided by the use of Python, along with all data handling and processing. In ultimately defining the strength of these relationships, MLB personnel can make better informed decisions about which tangibles to prioritize in player acquisition and development.

KEYWORDS

Baseball Analytics

Linear Regression

Python

1 Introduction

Baseball has often been referred to as "America's Pastime", largely earning this connotation through its combination of historical significance and its slow, segmented play style. Its significance within American culture and the sporting world gives rise to a multitude questions that can be answered with the help of data science practices. Among these questions is the focus of this report, namely, which aspects of MLB gameplay are most closely related with team success?

This question can be broken down into four secondary questions, which are as follows:

1) What is the relationship between batting and baserunning and winning in the regular season?
2) What is the relationship between fielding and winning in the regular season?
3) What is the relationship between pitching on winning in the regular season?
4) How have these relationships changed over time?

The first three inherently represent three different areas of play, with batting and baserunning being determined by offensive statistics and fielding and pitching being determined by defensive statistics. The fourth evidently references these same areas, but in the context of tracking trends over time. In understanding how these facets of the game impact winning both in the regular season and the playoffs, fans of the sports can gain greater appreciation for the statistics within baseball and team officials can make educated decisions on how to structure their rosters for the most success.

2 Data

2.1 Source of dataset

All datasets for this investigation were sourced from Baseball Reference. Baseball Reference is the primary resource for all baseball statistics, featuring all traditional statistics found on MLB.com, as well as more advanced, sabermetric stats for deeper level analysis on player and team performance. Datasets for this project focus on the 2009, 2013, 2017, 2021, and 2025 seasons and were generated within their respective years. The generation of these datasets is the result of data merging from several sources, including: MLB.com for current data, the Lahman database for historical data, and SABR for more advanced metrics.

2.2 Characters of the datasets

25 different datasets were used in this investigation, with five datasets coming from each of the five previously indicated MLB seasons. Each dataset includes statistics for all 30 MLB franchises, with the number of variables ranging from 18 to 32 between datasets. Each season has the same five datasets, batting, fielding, starting pitching, relief pitching, and league standings. Due to the number of variables in each dataset, only the variables included in the analysis will be identified.

2.2.1 Batting

Investigating batting and baserunning is done with the intention of determining how a team's offensive output correlates to winning games. The specific statistics that will be used to represent this component of play are runs scored per game, total bases stolen, batting average, on base percentage, slugging percentage, and home runs. Breaking this category down into individual statistics not only provides quantitative values to perform analysis on, but also allows differing components of offense to be examined and

compared. For example, there has long been a debate on whether contact hitting (hitting for average) or slugging (hitting for power) is a better strategy for team success. Such a debate will be settled within the contents of this report.

2.2.2 Fielding

Fielding represents the core component of a team's defense. Being able to turn plays to get outs, particularly those that may be considered more difficult, can be the difference between winning and losing teams. The statistics to be examined here are defensive efficiency, defensive runs saved, and errors.

2.2.3 Pitching

Pitching represents a secondary, yet arguably more important aspect of team defense. For the purpose of this investigation, the pitching category will capture both starting pitcher and bullpen statistics. These statistics are inclusive of quality start percentage, average game score, innings pitched per game started (on average), and save percentage.

2.2.3 Variable Summary

In addition to the variables discussed prior, the standings datasets were used the pull the response variable, wins, for each team in the indicated seasons. Table 1 offers a summary of variables included in the analysis.

Table 1: Summary of variables used from each dataset for each year.

Batting	Fielding	Starting Pitching	Relief Pitching	Standings
Runs/Game	Errors	Innings/ Game	Save %	Wins
Bases Stolen	Def. Efficiency	Quality Start %		
Batting Avg.	Def. Runs Saved	Avg. Game Score		
On-Base %				
Slugging %				
Home Runs				

It should be noted that the number of variables chosen from each dataset was reflective of the statistics that could intuitively have the greatest impact on winning games. Some areas of baseball naturally yield more statistical categories for consideration. While that gives rise to some unbalance, there are roughly an equal amount of offensive and defensive statistics.

2.2.4 Cleaning and Processing

Given the credibility of the source, the datasets did not require cleaning, but they did need to be combined. For each year considered, the five datasets from that season were merged into one complete dataset based on team name, while dropping all unnecessary variables in the process. Python's Pandas library was instrumental here, allowing for dataset merging and feature engineering.

3 Methodology

To determine the relationships between the various components of baseball and winning, a simple linear regression approach was considered. Such an approach was conducted by computing the correlation coefficients between each statistic presented in Table 1 and winning for the 2025 season. These coefficients were then averaged for each area (batting, fielding, pitching). This was then repeated for the 2021, 2017, 2013, and 2009 seasons to get an idea of how these correlations have changed over time. The primary assumption of this model is that the segmented, almost chess-like nature of baseball limits the amount of collinearity or the significance of the relationships between offensive and defensive variables. Such a model can be advantageous as it provides highly interpretable numeric constants portraying the relationship between a specific statistic or area of baseball (batting, fielding, pitching) and winning in isolation. The disadvantage to this approach is that there may be unforeseen interactions among variables which heighten or lessen an individual predictor's correlation to winning when all are considered together. Thus, the reason for this model's selection resides in its simple interpretability, with the assumption of limited collinearity. To execute this approach, the Pandas correlation function, corr(), was used. Additionally, Matplotlib and Seaborn were used for plotting, along with NumPy for various numerical operations. The only alteration that was made to this model to improve its validity was the exclusion of the runs per game statistic in the mean of the batting stats. This was done because runs are evidently based upon hitting stats.

4 Results

In the results portion of this report, the correlations among the selected predictors and wins will be presented for each of the three categories for 2025, followed by a comparison of the categories overall. After which, the changes in the overall correlations over time will be presented as well as some trends of interest.

4.1 Batting and Baserunning Results

In the 2025 season, the most significant predictor of wins among batting stats was on-base percentage. This is plotted visually in Figure 1. In fact, on-base percentage (0.724) had a higher correlation to winning than runs scored per game (0.712), which is

evidently interesting because ultimately, you need runs to win games. A potential explanation for this is that although scoring is important, runs per game can be inflated by teams that put up large numbers in some games, but few in others, while on-base percentage may reflect a more consistent performance. On the other side of the spectrum, bases stolen had a low correlation to winning (0.172). This was a consistent theme over the seasons investigated. It should also be noted that batting average appears to have a greater impact on winning than home runs. This is notable as there is a constant debate on whether teams should hit for contact (average) or for power (home runs). Table 2 summarizes all the correlation coefficients for the investigated batting stats for 2025.

Table 2: Summary of all batting coefficients MLB 2025

Batting and Baserunning

Statistic	Correlation Coefficient
Runs per Game	0.712
Total Bases Stolen	0.172
Batting Average	0.557
On Base Percentage	0.724
Slugging Percentage	0.500
Total Home Runs	0.342
Adjusted Average	**0.459**

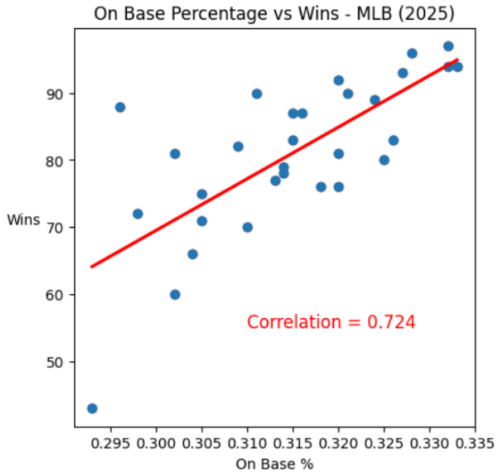


Figure 1: On-base percentage plotted against wins for the 2025 MLB seasons for all 30 teams.

4.2 Fielding Results

For 2025, defensive efficiency (0.592) appeared to have the greatest impact on winning among the considered fielding stats. This is visualized in Figure 2. Considering defensive efficiency reflects a team's ability to limit runners reaching base on balls put into play, this seems to be a reasonable conclusion. Defensive runs saved and errors sat just below, each at relatively similar values, 0.324 and -0.363, respectively. Note that the errors coefficient is negative because making more errors results in fewer wins.

Table 3: Summary of all fielding coefficients MLB 2025

Fielding

Statistic	Correlation Coefficient
Defensive Efficiency	0.592
Defensive Runs Saved	0.324
Errors	-0.363
Average	**0.426**

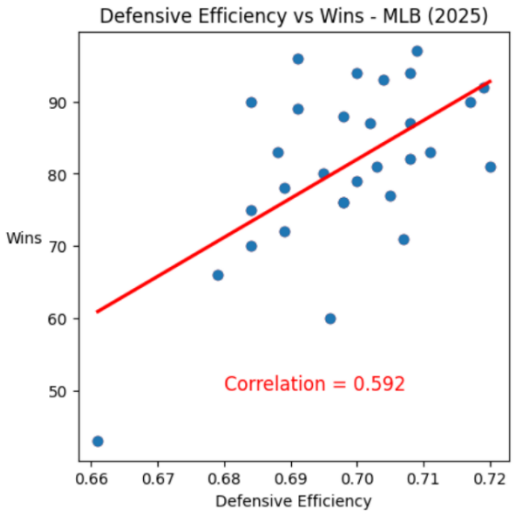


Figure 2: Defensive efficiency plotted against wins for the 2025 MLB season for all 30 teams.

4.3 Pitching Results

Pitching is of particular interest because two camps exist within this category, starting pitching and relief pitching. Based on the pitching correlation coefficients, average game score appears to be the greatest predictor of wins, which intuitively makes sense, as it wraps a few statistics (not included in this analysis) such as walks, hits, and runs allowed into one number. In fact, average game score was the greatest indicator of a winning team, with a correlation coefficient of 0.783. It should be noted that save percentage trails closely behind at 0.637, which reflects reliever importance. The other coefficients are summarized in Table 4.

Table 4: Summary of all pitching coefficients MLB 2025

Pitching

Statistic	Correlation Coefficient
Quality Start %	0.501
Average Game Score	0.783
Innings Pitched per Start	0.454
Save %	0.637
Average	**0.594**

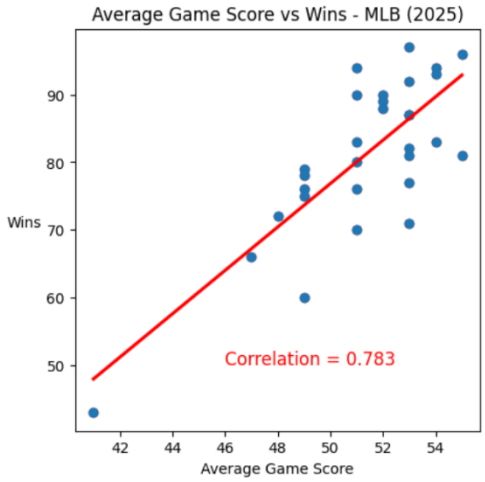


Figure 3: Average game score plotted against wins for the 2025 MLB seasons for all 30 teams.

4.4 Overall Correlations

In observing Tables 2-4 the relative weighting of each category's importance to winning becomes apparent. For the 2025 season, pitching had the highest correlation coefficient, followed by batting and fielding. This seems to echo the prevailing attitudes towards the importance of each, with pitching continuously being seen as a separator for bad, good, and great teams. Figure 4 represents these correlation coefficients in the form of a bar chart.

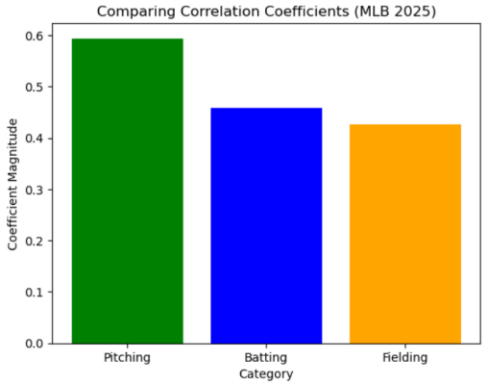


Figure 4: Correlation coefficients by category MLB 2025

4.5 Gameplay Change over Time

One of the cornerstones of this investigation was attempting to understand how these correlations have changed over time. As previously indicated, a similar analysis to 2025 was conducted for the 2021, 2017, 2013, and 2009 seasons. Over this period, as it was in 2025, pitching remained the most important aspect of MLB play. Additionally, batting remained a close second, while fielding remained the least important of the three. These trends are illustrated in Figure 5.

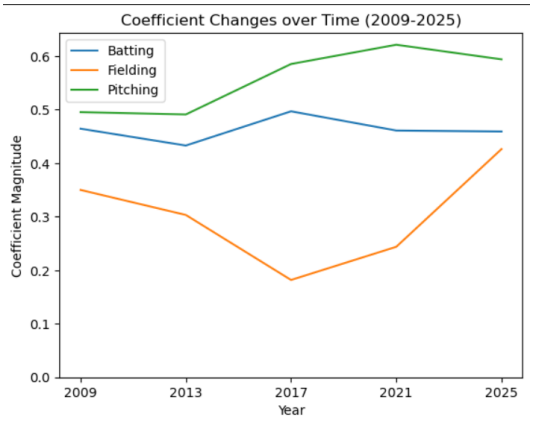
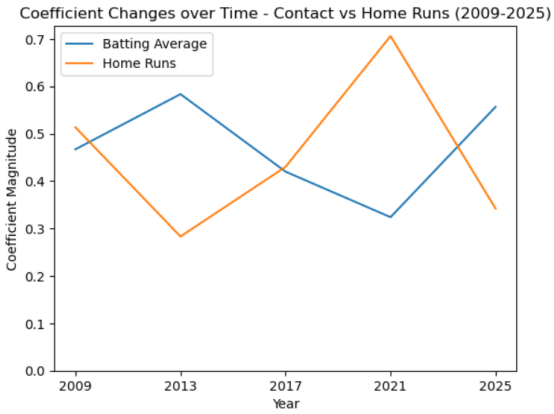


Figure 5: Changes in category correlation coefficients MLB (2009-2025)

In observing Figure 5, it is clear that pitching has seen an increase in importance over the 16 years, fielding saw an initial decline, followed by a sharp uptick, and batting has remained relatively constant. Based on the most recent trends, fielding may soon overtake batting in terms of relative importance, and pitching may even regress closer to batting, similar to 2009. In addition to the categorical trends, it may also be of interest to see how specific stats have evolved in terms of their importance to winning. Recall from section 4.1, where the debate between hitting for contact versus hitting for power was discussed. Figure 6 visualizes the change in relative importance between both home runs and batting average to winning.



As can be seen, the two approaches to gameplay have varied in their effectiveness in terms of earning wins in recent years. A potential reason for the relevance of home runs in the late 2010's could be related to what is referred to as the "juiced ball era". The juiced ball era references a period where the properties of MLB baseballs had changed to result in more aerodynamic characteristics [1]. In 2022 however, the MLB changed their baseball manufacturing process, which could explain the recent resurgence of hitting for average [2].

5 Discussion

The primary weaknesses of this project reside in two areas. The first is whether or not the assumptions made by the model are adequate. The nature of the sport and the selection of the stats suggests that the assumptions should hold to a reasonable degree, but regardless, the analysis could be improved. These improvements could take the form of a multi-variable linear regression to consider any potential interactions between variables. The other primary weakness of this investigation is the limited number of years investigated for trends over time. Since each year required the blending of 5 different datasets, it became costly timewise to add more seasons. Evidently though, the quality of the analysis would only improve in adding more years.

6 Conclusion

In investigating the relationships between different MLB statistical areas and winning, valuable conclusions and insight can be gathered. Among these conclusions are the statistics that are most highly correlated with winning, which are on-base percentage, runs per game, and average game score. In addition, different and competing approaches to offense can be compared and explored. Notably, as of the 2025 season, teams that hit well for contact generally outperformed teams that prioritized hitting home runs. Overall, though, one of the largest takeaways is the importance of pitching relative to other areas of the game.

Developing an understanding for these relationships and how they have evolved over time has many real-world implications. Aside from being of interest to the average MLB fan or sports supporter, they may provide a strategical advantage for team managers and owners. By analyzing the data provided in this report, team officials can make informed decisions about where to spend their money and what stats to consider when signing and developing different players. Further, they may be able to estimate what will become more important in the future by analyzing the trends among the coefficients over time. Ideally, this all combines to create a better on-field product. All things considered, the MLB and sports as a whole are becoming increasingly reliant on data-driven analysis. This report serves as a prime example of what kinds of valuable insights can be obtained through a data science approach.

ACKNOWLEDGMENTS

Acknowledgements for this report are made to Professor Weijie Pang as a primary resource for assistance with Python code through the 2025 Fall semester.

REFERENCES

[1] MLB. (2023, February 21). MLB receives report on Increased Home Run Rate. MLB.com. https://www.mlb.com/press-release/major-league-baseball-receives-report-on-increased-home-run-rate-278136100

[2] Davis, B. W. (2022, December 6). *Major League Baseball used at least two types of balls again this year, and evidence points to a third*. Business Insider. https://www.businessinsider.com/mlb-used-two-balls-again-this-year-and-evidence-points-to-a-third-2022-12