A Multivariate Analysis of KPIs Affecting Wins and Losses in Professional Baseball

Jaden Hu

Abstract

Wins and losses are the culmination of many different events in baseball, and there are many factors that go into a win or loss. A single win or loss column can only give so much information, and this research paper will measure which metrics are the most telling in measuring a team's success, defined by a team's winning percentage. Whether it is statistics, the time of day, or the strength of a team's opponents, each factor has a different effect on wins and losses. By delving into these factors, we hope to explain what determines success or failure in the captivating world of baseball.

Introduction

Baseball, often referred to as America's pastime, has evolved far beyond its origins into a data-driven sport where every pitch, swing, and defensive play is meticulously analyzed. In this era of modern baseball, the quest for a competitive edge and enhanced understanding of player performance has led to the exploration of a simple but intriguing question: What factors predict baseball wins and losses? The stats that appear in the box score such as batting average and earned run average tell only half of the story. More advanced stats such as fielding independent pitches may tell a more complete story, even though there are often intangible factors such as luck or randomness that can contribute to wins and losses. However, this paper will only explore the effects of what we can tangibly measure using statistics.

Background

The idea of analyzing baseball teams using statistics to analyze players and teams was first popularized in 1971 by Bill James. The idea of a science behind baseball statistics is called "sabermetrics," which is a term coined by Bill James. This concept was further revolutionized with Michael Lewis's book, *Moneyball*, (Losak 2016). Although the ideology for sabermetrics has stayed the same, the methods used to determine metrics have evolved over the years. There have been a number of new statistics used to measure and predict performance along with growing technology. Modern technology has expanded the horizons of how baseball can be analyzed beyond what anyone thought was possible.

Baseball has many different terms and statistics, and analyzing all of these numbers are a basic part of a front office executive or scout. The numbers all make up a comprehensive evaluation of a team's season, and if they overperformed or underperformed. We will use the stats from the 2023 MLB season to explore their impacts on wins and losses, and the stats are all from Baseball Reference. The stats this research paper will cover are batting average, slugging percentage, on base plus slugging plus (OPS+), earned run average (ERA), earned run average

plus (ERA+), and fielding independent pitching (FIP). Batting average, which is the number of hits divided by the total number of plate appearances. Slugging percentage measures the amount of power a hitter has. Slugging percentage is calculated by dividing the total number of bases reached by the total number of at bats. OPS+ takes the sum of the on base percentage, the percentage that a hitter reaches base, and the slugging percentage and adjusts it according to league and ballpark conditions. A 100 OPS+ is average, and a 120 OPS+ means that a hitter is 20% better than the average player in the MLB in offensive performance. ERA is a pitching stat, which measures the average number of runs a pitcher gives up per 9 innings. A lower ERA indicates a better pitching performance. ERA+ is the same idea as OPS+, but it uses ERA. ERA+ takes the ERA and adjusts it for ballpark and league conditions. An ERA+ of 100 is average, just like for OPS+. FIP purely measures pitching performance, and excludes uncontrollable factors such as defensive errors.

Comparison of How Different Stats Affect Wins/Winning Percentage

First, we will explore the batting statistics. Batting average, slugging percentage, and OPS+ are all respected statistics in baseball that are commonly used to measure a player's success. Transferring these stats to a team as a whole should also measure a team's success offensively. Having higher offensive team stats should signify that a team will score more runs, which will lead to more wins. However, in our analysis, the opposite is true. When these stats are plotted against wins, there is very little to no correlation to wins. The correlation coefficient in each of the graphs is close to 0, which means there is no correlation.The data showcases a perplexing scenario where teams boasting low batting averages are capable of amassing a high number of wins, while conversely, teams with lofty batting averages may find themselves in possession of a meager win count.

In Figure 2, slugging percentage is also shown to have little to no correlation with wins. Teams boasting formidable slugging percentages were just as likely to achieve a meager win count as those with relatively lower SLG figures. Advanced offensive strategies may prioritize hitting home runs and accumulating extra-base hits. Consequently, a team with a high SLG may, in theory, be better equipped to score more runs. However, the reality portrayed in our analysis challenges this assumption, hinting that there may be more to a team's success than raw power-hitting capabilities. The negligible correlation between SLG and wins suggests that the ability to hit for power alone does not guarantee a team's triumphs.

The story remains consistent when we turn our attention to slugging percentage and OPS+. Contrary to conventional wisdom, a higher OPS+ exhibits a negative correlation with wins, as illustrated in Figure 3. These findings collectively suggest that a team's offensive statistics alone do not strongly correlate with their success in the win-loss column. From the graphs, we can conclude that a team's offensive statistics does not have a very strong correlation in wins and losses.

One explanation for this disconnect could be because teams may struggle to capitalize on their chances to drive runners in, which will lead to fewer runs scored and fewer wins. Even with the inclusion of more advanced statistics, there are many external factors such as the strength of opponents that have a much higher effect on wins and losses.

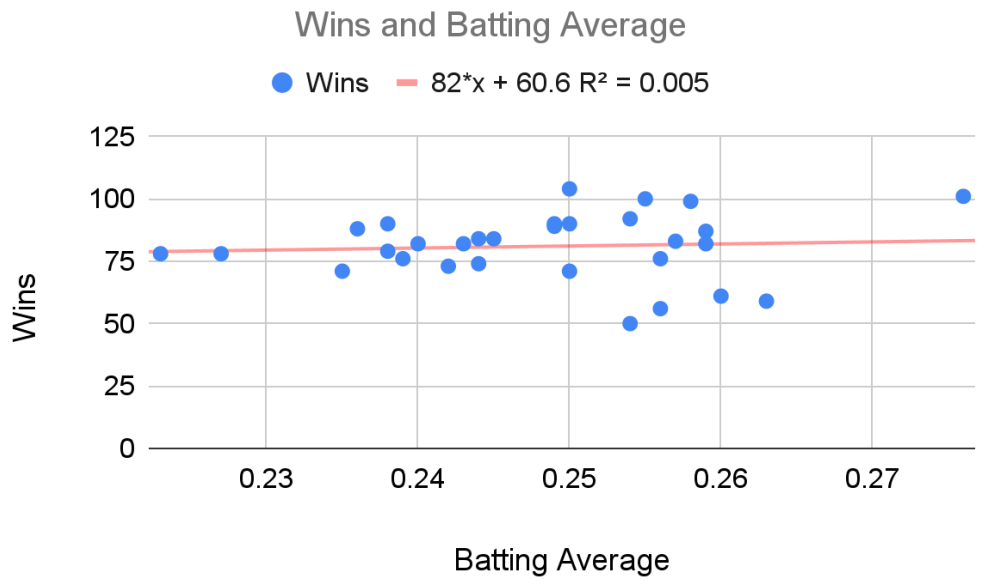

Figure 1, 2023 MLB Batting Average Vs Wins

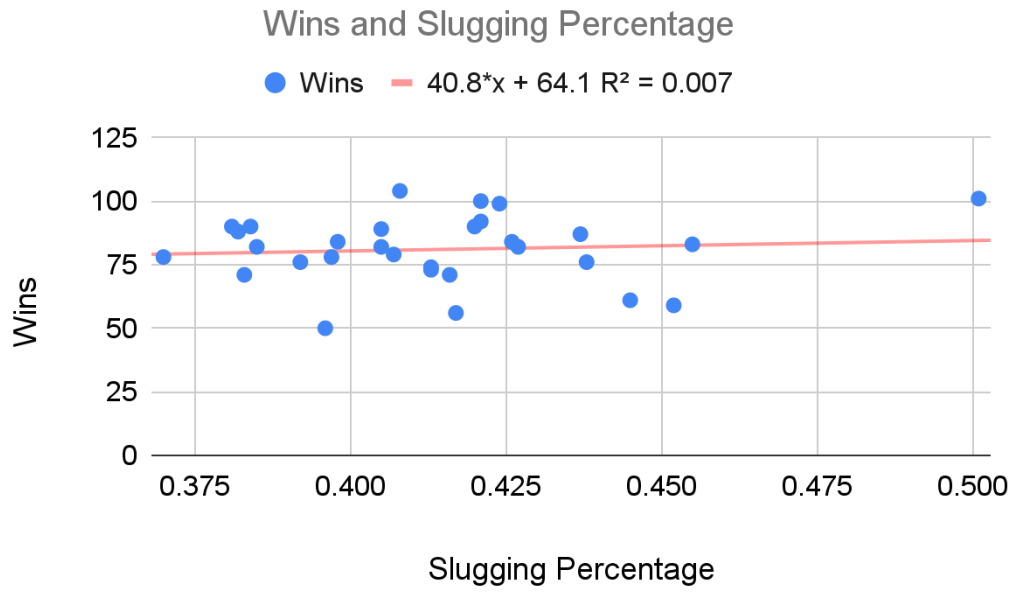

Figure 2, 2023 MLB Slugging Percentage vs Wins

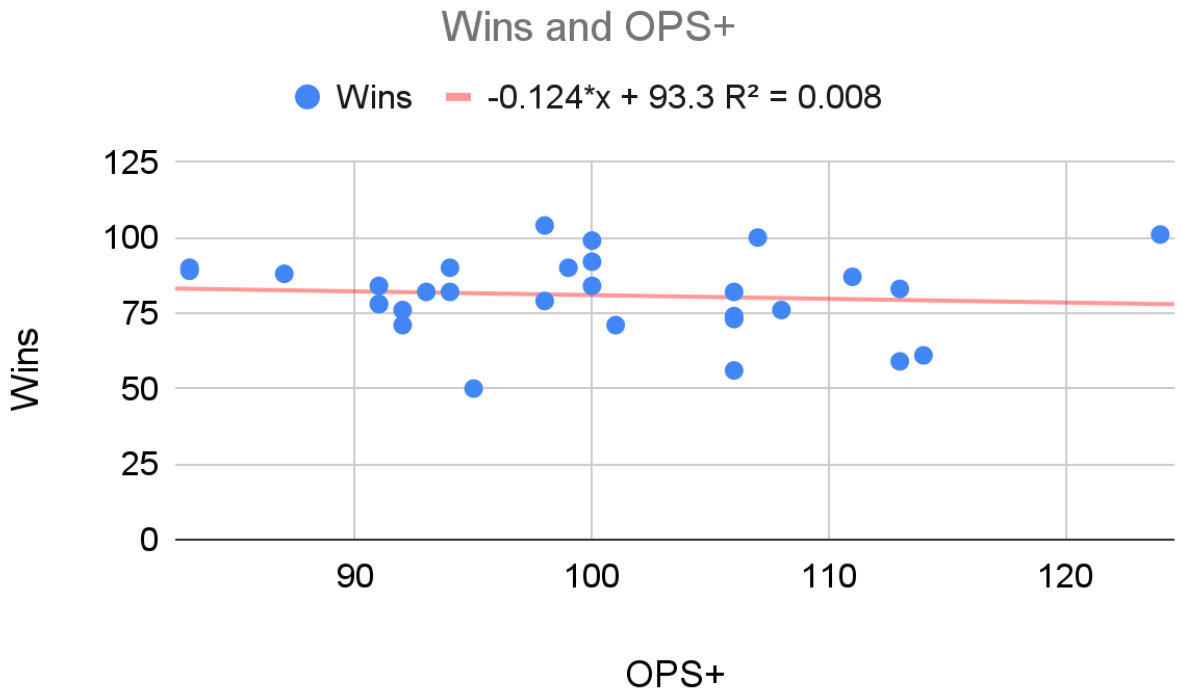

Figure 3, 2023 MLB OPS+ vs Wins

Continuing our exploration into the intricate web of baseball performance metrics, we turn our attention to team pitching statistics, anticipating insights into their relationship with wins and losses. Astonishingly, our findings reveal a remarkable similarity to our observations regarding offensive statistics – a lack of substantial correlation between these pitching metrics and the number of wins accrued by Major League Baseball teams. In our analysis, we have examined three key pitching statistics: ERA, ERA+, and FIP. ERA+. ERA shows the most correlation of the three metrics, with a correlation coefficient of 0.077. The correlation is negligible, and figure 4 shows that teams with low ERA's can have a small number of wins, while teams with higher ERA's can be one of the best teams in Major League Baseball.

ERA+, a metric designed to measure a pitcher's effectiveness while adjusting for external factors such as the home ballpark, showcases a correlation coefficient of 0 when plotted against wins. This value signifies an absence of any linear relationship between ERA+ and wins, an outcome that challenges conventional notions about the pivotal role of a strong pitching staff in securing victories.

Similarly, FIP also reveals a correlation coefficient of 0 when plotted against wins. Once again, this result signals a complete absence of a linear correlation between FIP and the number of wins a team accumulates throughout the season.

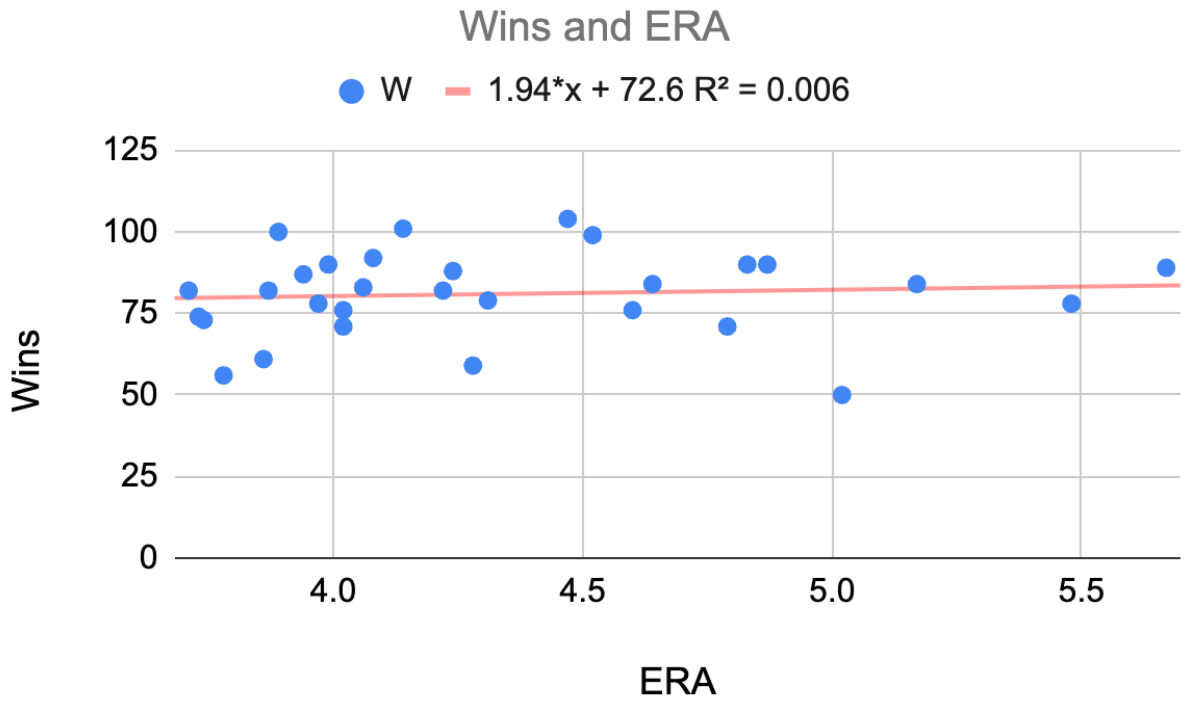

Figure 4, 2023 MLB ERA vs Wins

Such findings are not only surprising but also profoundly thought-provoking. They prompt us to reassess the traditional belief in the paramount significance of pitching prowess as a predictor of team success. The lack of a significant correlation between these pitching statistics and wins calls for a reevaluation of the factors that genuinely influence a team's performance in the win-loss column.

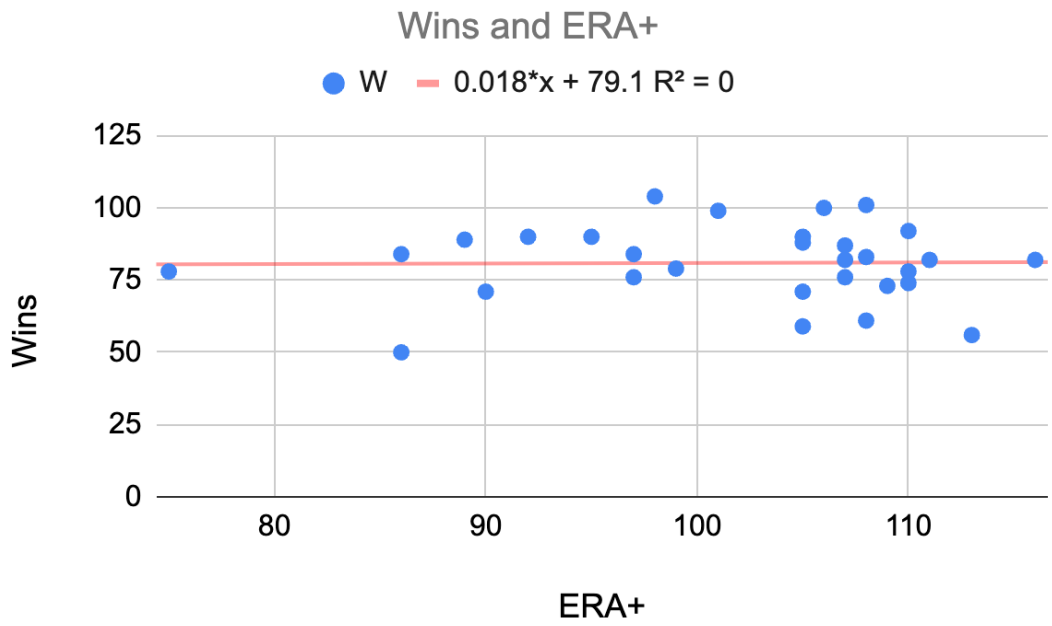

Figure 5, 2023 ERA+ vs Wins

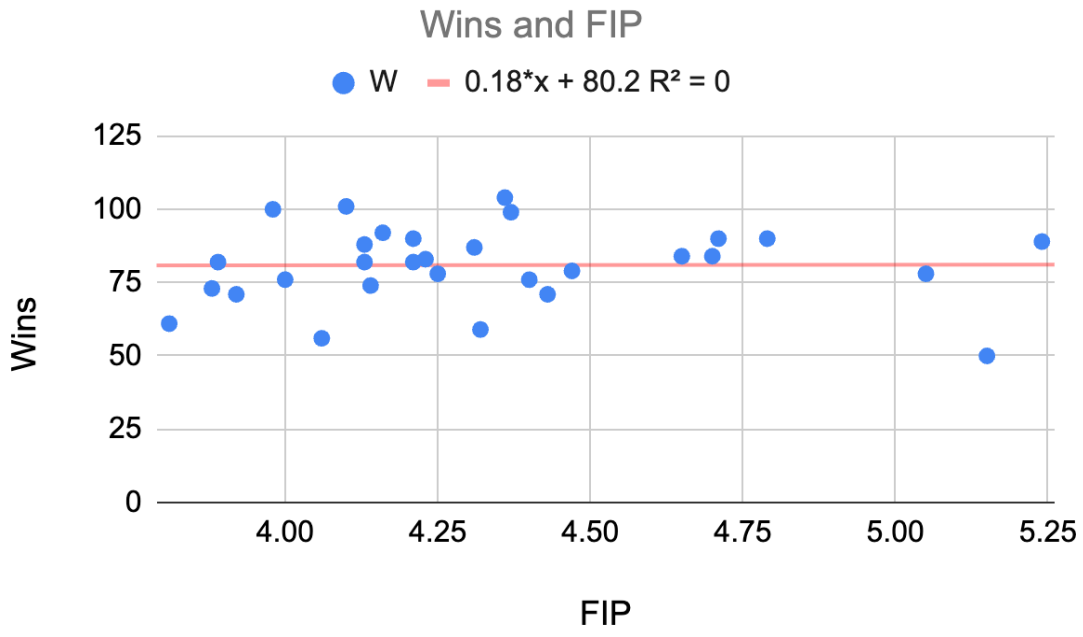

Figure 6, 2023 FIP vs Wins

Other Factors

Now that we have determined team statistics have a negligible correlation with wins, we will now explore other factors that may influence wins and losses. These factors include whether the game was played in the day or night. According to mlb.com, there was a high correlation between a high winning percentage and having a large number of home wins. The correlation coefficient of 0.894 shows that relationship, as seen in Figure 7, and confirms this claim. On the contrary, a high winning percentage and a large amount of day wins does not have a very high correlation coefficient. One reason for this could be because of the taxing nature of day games. Having less rest in between games has a significant impact on the body and performance, which could skew the results in favor of the night games. In addition, day games introduce new factors into games such as fielders having to deal with the sun.

Our analysis of game timing provides valuable insights into the influence of day and night games on team performance. While a high number of home wins exhibits a strong correlation with overall success, day game victories demonstrate a less pronounced link. This finding underscores the multifaceted nature of baseball, where external factors such as game timing and conditions can significantly affect a team's performance.

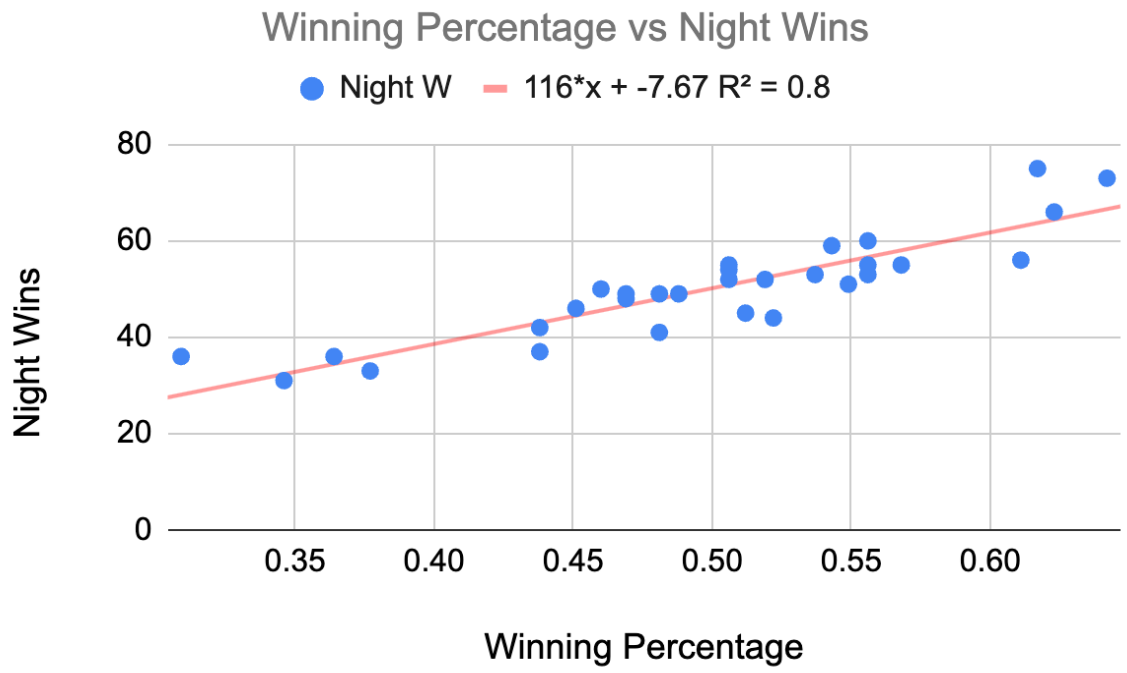

Figure 7, 2023 Winning Percentage vs Night Wins

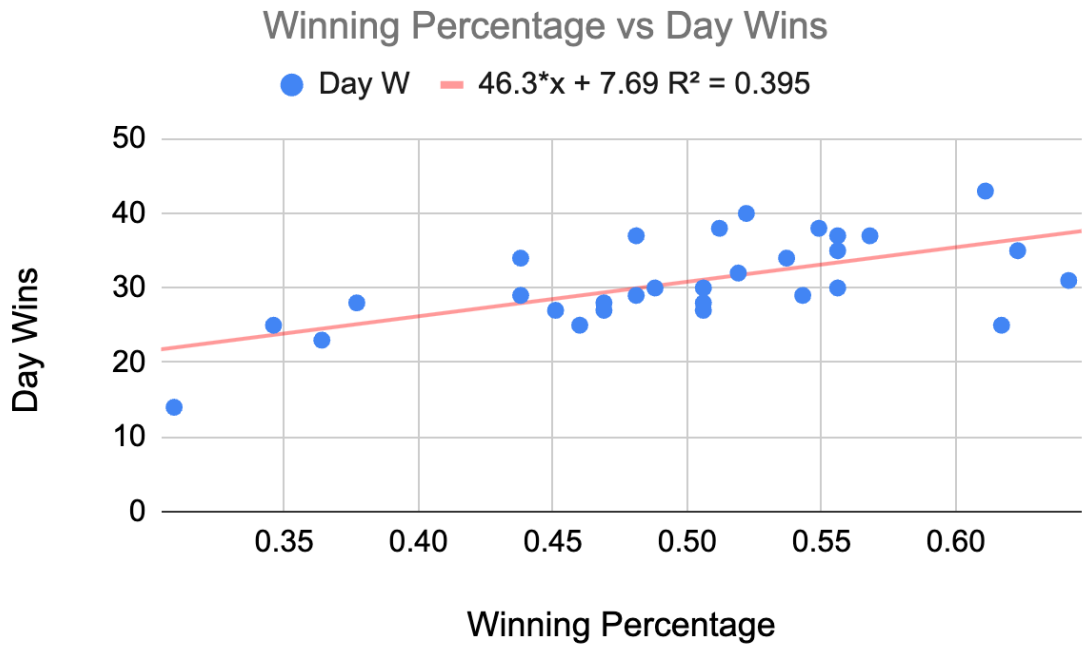

Figure 8, Winning Percentage vs Day Wins

Refer to the appendix for more graphs charting wins against other factors.

Pythagorean Win Percentage

While our exploration of various statistics has revealed intriguing insights into the factors that influence a team's wins and losses, there exists a powerful tool that captures the essence of baseball's success equation. This tool, known as the Pythagorean Win Percentage, reinforces the fundamental objective of baseball: scoring more runs than your opponents.

The basic formula for Pythagorean Win Percentage is

$$\text{Win Percentage} = \frac{\text{Runs Scored}^2}{\text{Runs Scored}^2 + \text{Runs Allowed}^2}$$

The Pythagorean Win Percentage directly takes into account the number of runs a team scores and allows, so it evaluates winning in baseball's purest form. Over the years, the exponent in the Pythagorean Win Percentage formula has been altered, but the base formula is still the same.

When plotting run differential against wins, we can see that there is a very strong correlation between run differential and wins. The correlation coefficient is 0.943, which signifies that run differential and wins move together very closely. Figure 9 displays that the higher the run differential for a given team, the higher the wins that team has. This correlation makes sense because in order for a team to win a baseball game, they must outscore their opponents. Run differential measures the difference between the number of runs a team scores and allows. A low run differential implies that a team scores more runs than they allow, leading to more wins.

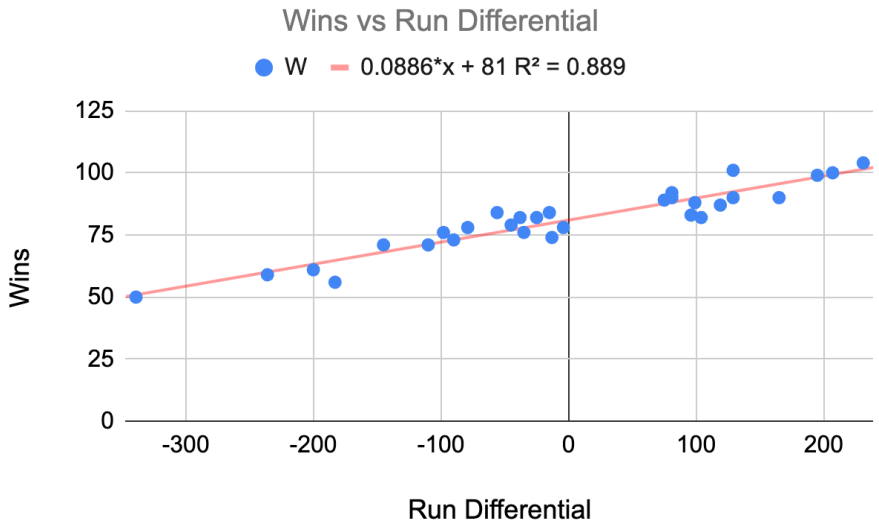

Figure 9, 2023 Wins vs Run Differential

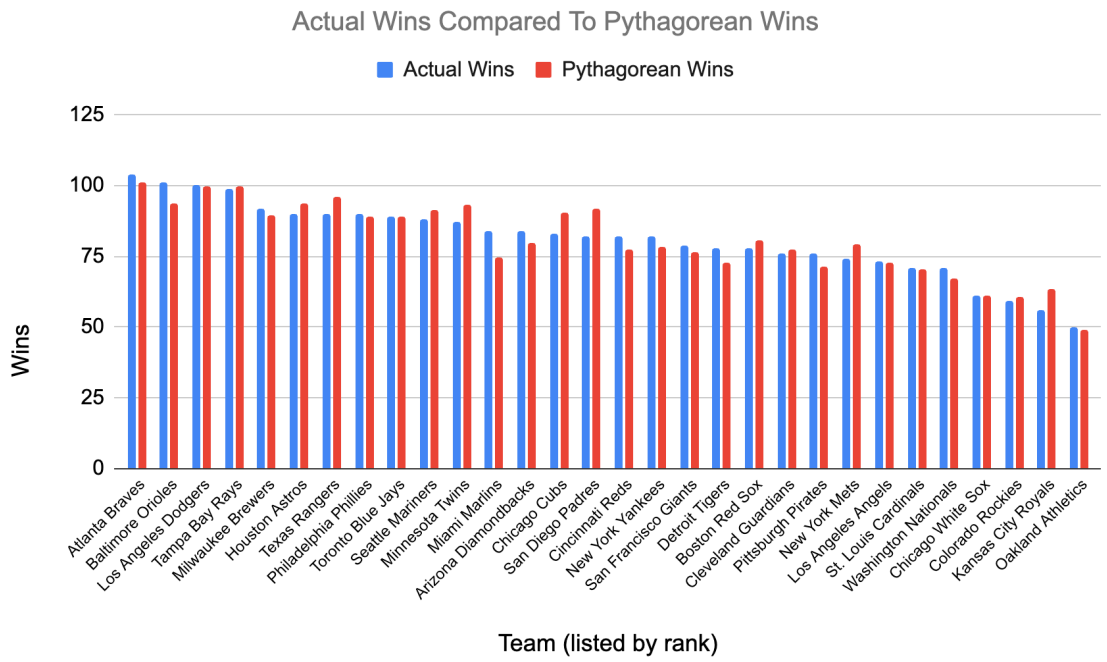

Figure 10, 2023 Actual Wins Compared To Pythagorean Wins

Future Technology With AI Algorithms/Current Direction in the Field

The landscape of baseball statistics is undergoing a transformative shift towards embracing the power of Artificial Intelligence (AI) and neural networks (Koseler 2018). Teams are increasingly using machine learning algorithms to analyze player statistics and injury data to make better informed decisions about player development and acquisitions. AI's predictive capabilities enable teams to identify emerging talent and fine-tune player strategies, ultimately enhancing a team's competitive edge. For example, there have been developed neural networks to predict catch probability (Koseler 2018). The neural network would be able to predict the coordinates and velocity of the ball in order to catch it, which would in turn lead to predicting catch probability. Neural networks like these can contribute to the rapidly changing world of sabermetrics. Using AI to predict wins and losses, statistics, and other nuances of the sport of baseball is rapidly evolving, and the sport will benefit greatly with the advancement of technology.

Conclusion

Our findings were surprising, and showed that statistics are often a very small part in the bigger picture of wins and losses. Run differential played a much larger role in wins and losses, and it showed a much higher correlation to winning. Having the best stats may not always translate to having the highest winning percentage due to the many factors that go into winning a baseball

game. Whether it's day and night games or the countless number of other uncontrollable external factors, baseball is a sport where there is not only one factor that goes into winning.

References

Huang, Mei-Ling, and Yun-Zhi Li. "Use of Machine Learning and Deep Learning to Predict the Outcomes of Major League Baseball Matches." Applied Sciences, vol. 11, no. 10, May 2021, p. 4499. DOI.org (Crossref), https://doi.org/10.3390/app11104499.

Koseler, Kaan, and Matthew Stephan. "Machine Learning Applications in Baseball: A Systematic Literature Review." Applied Artificial Intelligence, vol. 31, no. 9–10, Nov. 2017, pp. 745–63. DOI.org (Crossref), https://doi.org/10.1080/08839514.2018.1442991.

Lyle, Arlo. Baseball Prediction Using Ensemble Learning. 2007. University of Georgia, https://getd.libs.uga.edu/pdfs/lyle_arlo_m_200705_ms.pdf.

Paul, Rodney J., et al. "Managerial Decisions and Player Impact on the Difference between Actual and Expected Wins in Major League Baseball." Volume 7 Table of Contents 2016 (2016): 57. https://www.economics-finance.org/aefj/issues/AEFJ_volume7.pdf#page=60

Yang, Tae Young, and Tim Swartz. "A Two-Stage Bayesian Model for Predicting Winners in Major League Baseball." Journal of Data Science, vol. 2, no. 1, July 2021, pp. 61–73. DOI.org (Crossref), https://doi.org/10.6339/JDS.2004.02(1).142.

Baseball-Reference.com. "2023 Major League Baseball Team Statistics." https://www.baseball-reference.com/leagues/majors/2023.shtml.

"Baseball Standings." MLB.Com, www.mlb.com/standings.

Appendix

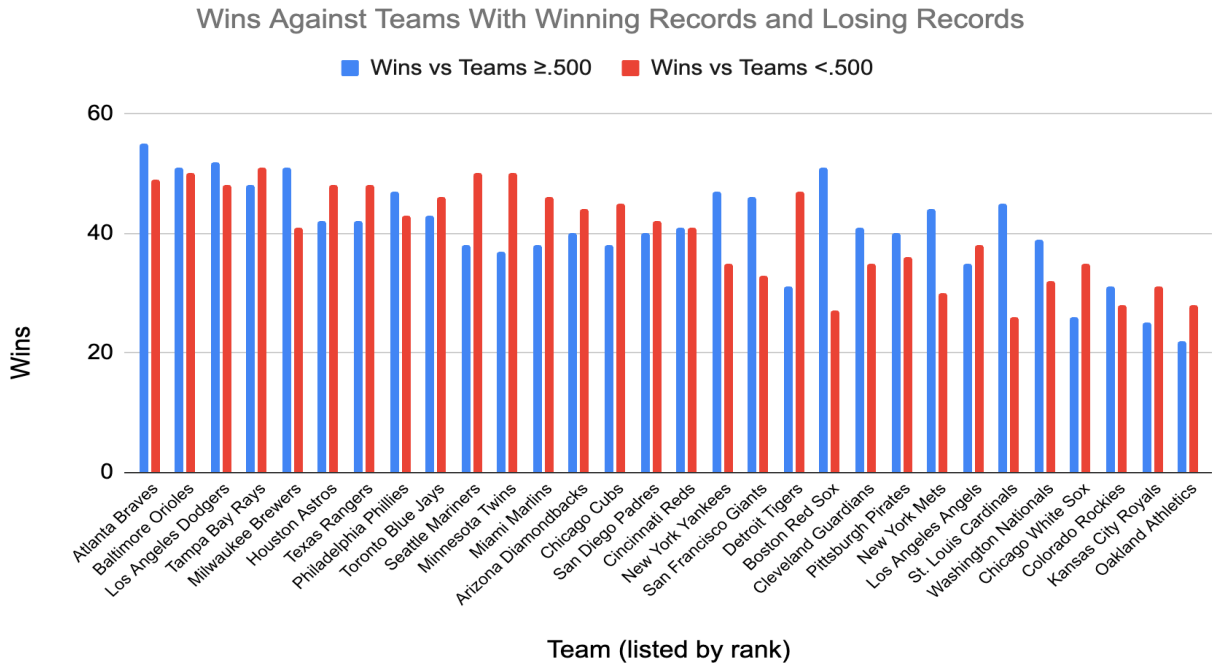

Figure 11, 2023 Wins Against Teams With Winning Records and Losing Records

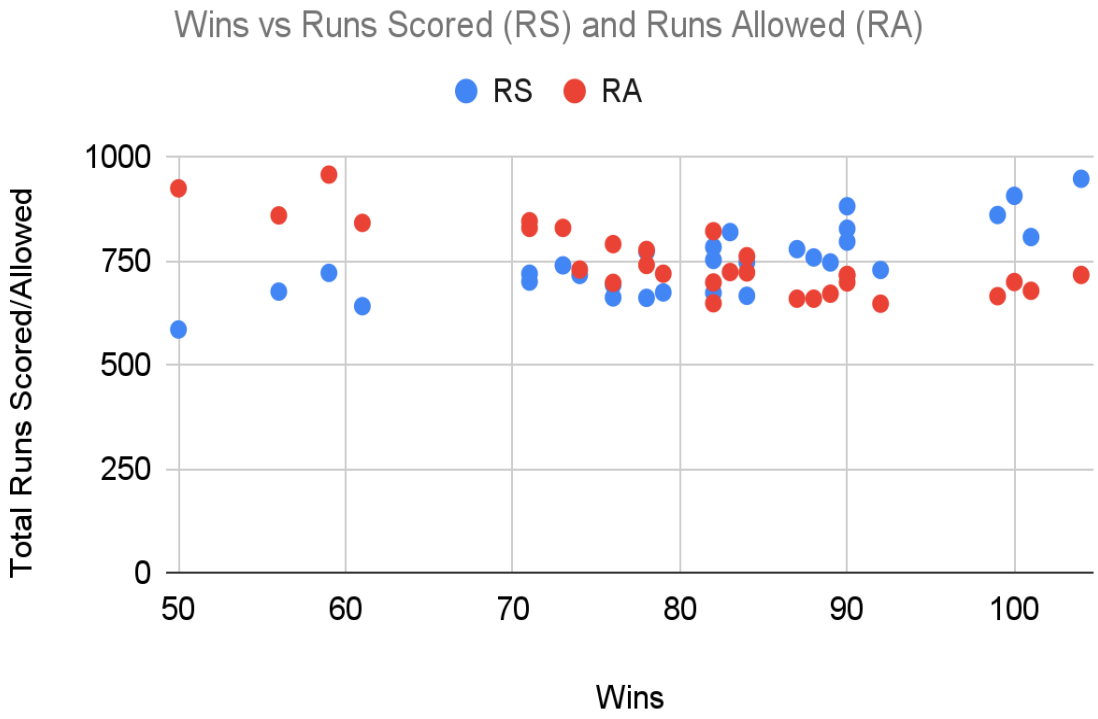

Figure 12, 2023 Wins vs Runs Scored and Runs Allowed

The graphs show other factors that can affect wins. Figure 11 shows the number of wins against teams with winning records and losing records.The majority of teams have more wins against worse teams, but that is not true for every team. We can see that the Atlanta Braves, the team with the best winning percentage in 2023, has more wins against teams with winning records than teams with losing records. This could be because they are more motivated to beat good teams or have a winning mentality that thrives in challenging situations. Conversely, some other teams with high winning percentages might struggle when facing better competition due to injuries or a weaker roster.

Furthermore, there are teams such as the St. Louis Cardinals and New York Mets who rank towards the bottom of the MLB, yet have more wins against teams with winning records. One reason for this could be because bottom ranked teams have an underdog mentality when going into games against top ranked opponents. They are more motivated, and could catch their opponents off guard. Playing top ranked opponents could also influence performance, and lead to better statistics in games against top ranked opponents. This could skew team stats when compared to wins, which may explain the correlation seen when plotting team stats to wins.

Figure 12 plots runs scored and runs allowed against wins. As mentioned before, we have found that run differential has a very strong correlation with wins, and Figure 12 shows that the teams with higher runs scored have a higher number of wins. On the contrary, teams with higher runs allowed have a higher number of losses. Figure 12 confirms our findings with how run differential has a strong correlation with wins.