A Regression Analysis of Baseball Performance Metrics and its Effects on Win Percentage*

Runs Per Game and ERA are Critical while Stolen Bases and Hits-Against Rate are Minor

Sean Chua

November 30, 2024

This paper examines the factors influencing baseball win percentage through a multiple regression analysis of key performance metrics, such as runs per game, earned run average (ERA), hits against per nine innings, and total stolen bases. Using a multiple regression prediction model, findings reveal that scoring more runs significantly enhances a team's chances of winning, while higher ERA negatively impacts win percentage; interestingly, the effects of hits allowed and stolen bases are comparatively minor. These results highlight the critical importance of offensive production and pitching effectiveness in determining team success in baseball. Understanding these dynamics not only informs team strategies and player evaluations but also contributes to a broader understanding of baseball performance metrics overall.

Table of contents

*Code and data are available at: https://github.com/jfhasj/mlb_lahman_analysis.

1 Introduction

In the highly competitive landscape of Major League Baseball (MLB), understanding the factors that influence a team's performance is crucial for players, coaches, analysts, and fans alike. Winning percentage serves as a key indicator of a team's success during the regular season, reflecting not only the outcomes of games but also overall team performance and the relationships between key success factors. Thanks to the growth of machine learning and the rise of sports analytics, teams are able to better assess potential outcomes and optimize performance based on available historical and current data. This paper aims to predict MLB teams' winning percentages during the regular season using various performance metrics, including run differential, runs scored per game, earned run average (ERA), hits allowed per game, and stolen bases.

The primary estimand for this analysis is the predicted winning percentage of each MLB team, which serves as a holistic measure of how well they performed during the regular season. This estimand will be calculated based on historical team data, specifically focusing on the

aforementioned variables, that will help estimate the winning percentage of a team given specific values of each variable.

Data from the most recent iteration of the Lahman Baseball Database (Friendly et al. 2024) was used, focusing on the Teams table to be able to extract only team statistics. It is worth noting that despite existing research on baseball analytics, more research has been geared toward determining which success factors correlate more or less strongly with a team's winning percentage. In light of this, this paper includes a statistical analysis using data from recent MLB seasons was conducted to build a predictive model for investigating how different aspects of team performance correlate with winning percentage.

Results show the relatively large impact between the average number of runs scored per game, as well as earned run average (ERA) and winning outcomes, as opposed to the minor impact between the number of hits a team gets against them (per 9 innings), as well as the total number of stolen bases, highlighting the importance of run differential and runs scored per game as primary predictors of success. By examining these dynamics, we aim to illuminate how teams can optimize their performance through effective run production and quality pitching and defense. Ultimately, this analysis seeks to provide valuable insights into the fundamental elements that drive success in baseball. The rest of the paper is structured as follows: Section 2 details the data and measurement process. Section 3 presents the model and justifies the choices made in the building of the chosen model. Section 4 presents the results, highlighting the relationship between different variables and the winning percentage, and Section 5 discusses the implications of the findings for this as well as areas of future research to improve predictive analytics in baseball.

2 Data

2.1 Overview

The dataset used for this paper is the most recently released version of the Lahman Baseball Database, first created in 1996 by Sean Lahman, which contains "contains complete batting and pitching statistics back to 1871, plus fielding statistics, standings, team stats, managerial records, postseason data, and more (Lahman 2024)." This dataset offers a comprehensive resource for baseball statistics, serving as an extensive collection of historical data on MLB players, teams, and games. In doing so, the dataset serves as an invaluable tool for analysts, teams, and casual fans alike.

The Lahman Database is only one of many baseball databases, though each of them vary in depth. These include data found on MLB's own website, Baseball Reference, and Retrosheet (Society for American Baseball Research (SABR) 2024). Aside from providing pitching, hitting, and fielding data, the Lahman database also includes miscellaneous data such as those about awards, Hall of Fame voting, salaries, All-Star games, and the like. The dataset covers available statistics from 1871 (the year the first ever major league game was played) to

2023, the most recent season in which complete data is available. While there are other baseball datasets available, such as the aforementioned Baseball Reference (Sports Reference LLC 2024) and Retrosheet (Retrosheet, Inc. 2024), the Lahman dataset was selected due to its overall completeness with season data and its ease of use; for the purposes of this analysis, the statistics found in this dataset suffice. Baseball Reference and Retrosheet were not used due to their extreme granularity; Baseball Reference allows for the ability to break statistics down by numerous criteria such as batter handedness, time of day, month, and so on which is not necessary for our purposes. Similarly, the Retrosheet database contains play-by-play data which is not of current interest for this analysis. As such, the aforementioned features of the Lahman database provide a sufficient balance between complexity and convenience.

All the data analysis was done through `R` (R Core Team 2024), along with a variety of external R packages used for specific tasks. For data manipulation and wrangling, the following packages were used: `tidyverse` (Wickham et al. 2019), `dplyr` (Wickham et al. 2023), `readr` (Wickham, Hester, and Bryan 2024), and `reshape2` (Wickham 2007). Modelling-related tasks were supported by `modelsummary` (Arel-Bundock 2022) and `tidymodels` (Kuhn and Wickham 2020). Data visualization was carried out using `ggplot2` (Wickham 2016), `ggpubr` (Kassambara 2023), and `DiagrammeR` (Iannone and Roy 2024), while `kableExtra` (Zhu 2024) was used for generating tables. For handling external data formats, the `arrow` (Richardson et al. 2024) and `rvest` (Wickham 2024) packages were used for reading and scraping data, respectively. Additionally, `here` (Müller 2020) helped manage project file paths, and `png` (Urbanek 2022) was used for saving images, while `rsvg` (Ooms 2024) and `DiagrammeRsvg` (Iannone 2016) were leveraged to properly render SVG diagrams (DAG) in the paper.

In this paper, it is important to note that team data was filtered to include those from only the 2014, 2017, 2022, and 2023 seasons, as these are seasons where all teams played exactly 162 games. Moreover, the past decade (and the statistics therein) most closely reflects the current state of baseball. The game of baseball has evolved significantly over the decades and even within the past few years, resulting in substantial changes in play and strategy.

2.2 Measurement

The process of converting real-world baseball phenomena into data involves careful measurement of team statistics and other performance metrics in the regular season, among many others (such as postseason, all-star, and Hall of Fame data).

The measurement for this conversion of data, and hence the creation of the overall dataset, is built from a variety of data sources. Lahman has attributed the source of raw data used in his database to statistician Pete Palmer, responsible for numerous baseball encyclopedias published in the past 5 decades. Over the years, Lahman and his team of researchers have constantly maintained and updated the various tables in the database using data from Retrosheet. Note that the database has also undergone various overhauls and redesigns since its

inception, and in October 2024, Lahman officially gave the Society of American Baseball Research (SABR) Committee full responsibility of managing his database (Lahman 2024). Note that prior to Lahman's donation of his database to SABR, many people contributed to the database's updating and management; for example, Lahman acknowledges that "Ted Turocy has done the lion's share of the work to updating the main data tables since 2012" while "the 2023 version of the Lahman Database was updated and generated by Bryan Walko." Consequently, some tables contain missing data as a result of human error, although the Teams table — the table pertintent for this analysis — is not one of them.

Currently, the Lahman database is maintained in a relational database format, with such formats including Microsoft Database files (.mdb) through Microsoft Access, a comma-delimited version (.csv), and MS SQL (.mssql). This variety in file formats allows for flexibility, as well as easy integration with statistical software and other database management systems.

Thus, each entry in the dataset is a result of careful validation, ongoing maintenance, and collaborative efforts by researchers and experts of compiled statistics from various sources. This is then separated into tables within the database, which is then organized into a relational database format, thereby allowing the public to conveniently access and analyze baseball data. This makes it possible to easily examine factors that affect teams' winning percentages.

2.3 Outcome Variables

The objective is to predict the outcome variable, `win_pct` which is a continuous variable which takes values between 0 and 1 inclusive. Specifically, `win_pct` indicates a team's winning percentage given a certain amount of wins and losses. The visualization of this outcome variable is presented below:

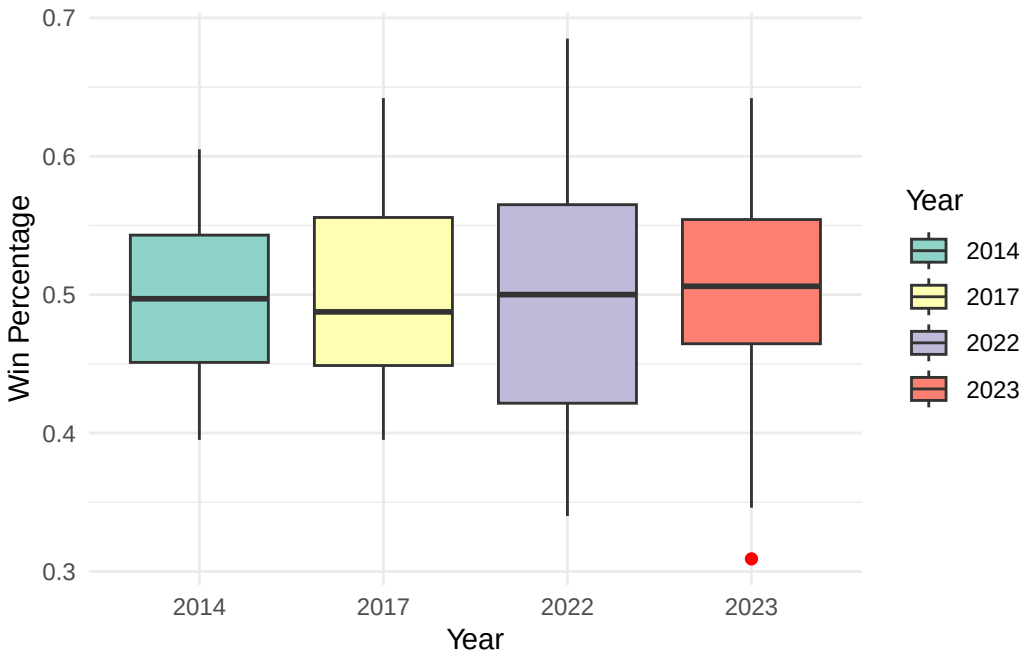

Figure 1: Distribution of Team Win Percentage in 2014, 2017, and 2022

Figure 1 illustrates the distribution of team winning percentage in the 2014, 2017. and 2022 regular seasons. We can see that the median winning percentage is relatively stable, hovering slightly lower or above 0.5 (50%) indicating that average teams tend to win about 50% of theur games. However, there are slight discrepancies in the winning percentages of the top 25% and 75% of teams, where the 25th percentile of teams won about 40% and 60% of their games respectively in 2014, while this range in win percentage further increases in 2017 and 2022. This may suggest that there was less and less parity within the MLB through the past few years. This being said, however, we can see that the median team winning percentage is relatively stable over a 162-game MLB regular season.

2.4 Predictor Variables

Runs per Game (`runs_per_game`): This continuous variable indicates the average number of runs a team scores in a game. It is important for analyzing team winning percentage since teams who score more runs tend to win more games (and vice-versa). Adding this variable to our analysis helps in making more accurate predictions about winning percentage, as this serves as one measure of offensive output by a team that contributes to winning games.

Earned Run Average (`ERA`): This continuous variable indicates the average number of earned runs given up by a team in a game. It is important for analyzing team winning percentage

since teams who give up less runs tend to win more games (and vice-versa). Similar to `runs_per_game`, adding this variable to our analysis helps in making more accurate predictions about winning percentage, as this serves as one measure of the quality of team that contributes to winning games.

Number of Hits Against per 9 Innings (`HA_9`): This continuous variable represents the average number of hits given up by a team over the course of 9 innings (the standard length of an MLB game). Similar to `ERA`, this metric is one measure used to evaluate a team's defensive effectiveness. A lower HA_9 indicates better performance, as teams are effectively limiting the number of hits allowed, which go hand-in-hand with giving up less runs.

Number of Stolen Bases (`SB`): This discrete variable indicates the total number of stolen bases by a team over the course of the season. Stealing bases increase teams' run-scoring opportunities and so serve as a factor in `runs_per_game`.

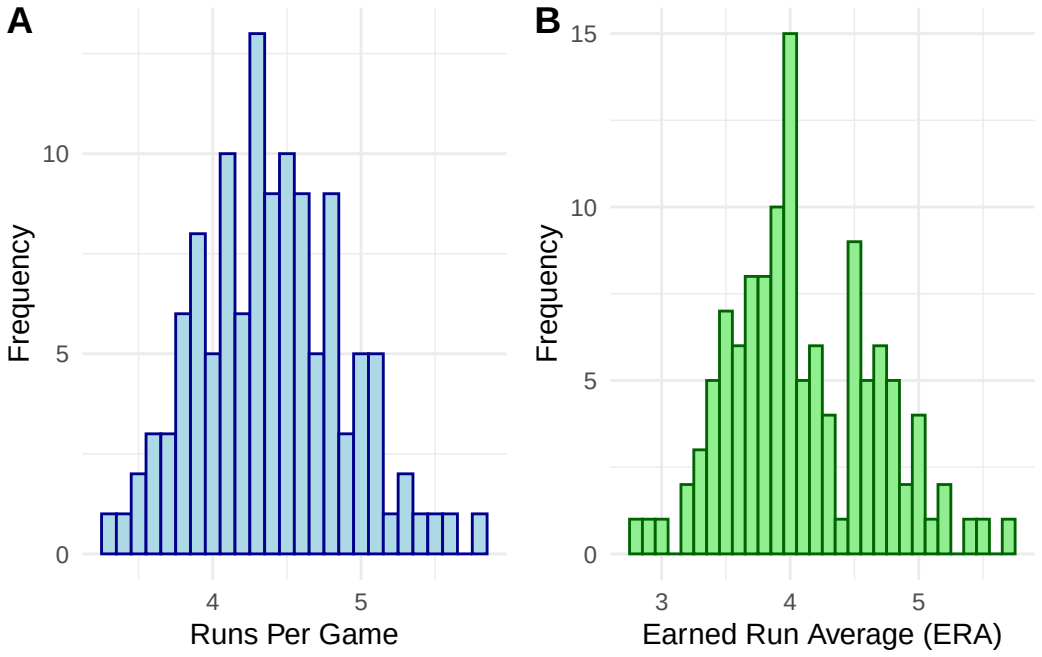

Figure 2: Distributions and Data for runs_per_game and ERA for the 3 Seasons

For Plot A in Figure 2, the distribution of `runs_per_game` reveals that most teams score between 4 and 5 runs a game, with the mode being about 4.3 runs per game. This suggests that average teams tend to score this amount of runs, with poor-performing and achieving teams scoring closer to an average of about 3.5 and 5.5 runs per game. The concentration of scores in this range indicates a typical performance level for MLB teams, highlighting the competitive nature of scoring.

Plot B in Figure 2 presents a histogram of `ERA`, showing that most teams give up between 3.7 and 5 earned runs a game, with significantly fewer teams giving up less than 3 runs or more than 5.5 runs. This distribution underscores the importance of pitching and defense in baseball; teams that can maintain an ERA below 4 are likely to be more competitive. The tails distribution suggest that while exceptional pitching exists, it is relatively rare.

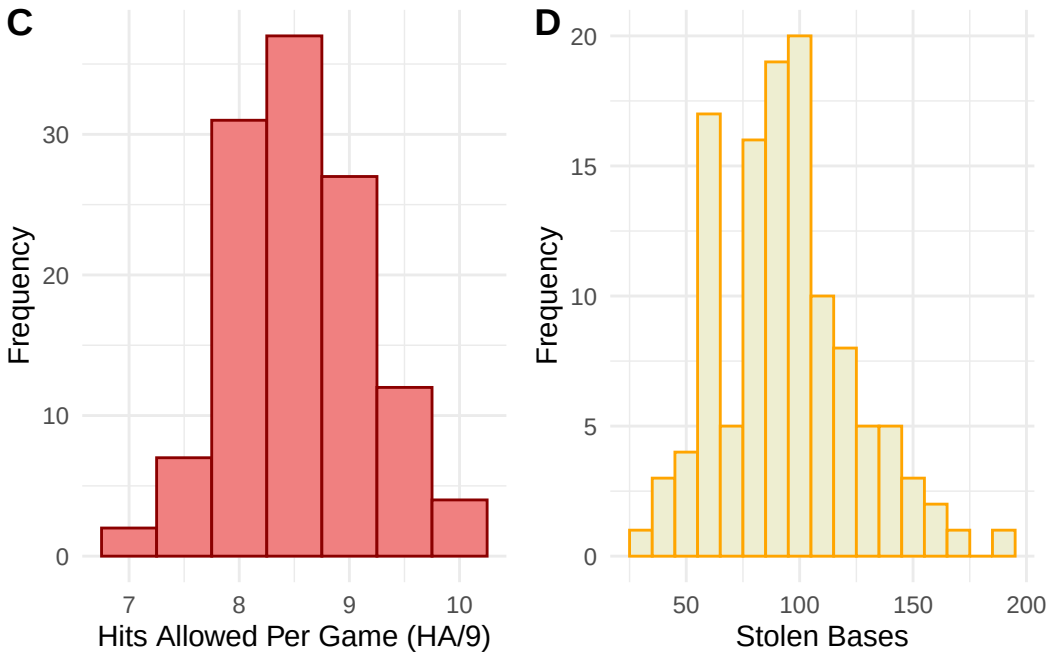

Figure 3: Distributions and Data for HA_9 and SB for the 3 Seasons

Plot C in Figure 3 shows the distribution of `HA_9` looks almost like a normal distribution, where a large majority of teams give up between 8 and 9 hits per game, while it is quite rare to give up 7 or 10 hits per game, indicating that these outcomes are outliers in the context of overall team performance.

In Figure 3, Plot D illustrates the distribution of `SB` among teams; we see that teams tend to record about 60-110 stolen bases per season. This range highlights a common level of aggressiveness on the base paths among teams, suggesting that stealing bases is a strategic element utilized by many clubs to enhance their offensive capabilities. We notice that very few teams record more than 150 stolen bases; this usually indicates a team having many speedy batters and better base-running strategies.

3 Model

3.1 Overview

This section describes a multiple regression model for predicting teams' win percentages over a typical 162-game MLB season. The model estimates win percentage given several variables — namely average runs scored per game, team ERA, the number of hits against per 9 innings, and the total number of stolen bases — that may or may not significantly affect a team's chances of winning.

3.2 Model Structure

The model is represented by the following equation:

$$\text{predicted\_win\_pct} = \beta_0 + \beta_1 \cdot \text{runs\_per\_game} + \beta_2 \cdot \text{ERA} + \beta_3 \cdot \text{HA\_9} + \beta_4 \cdot \text{SB}$$

Here, β_0, β_1, β_2, β_3, and β_4 are the model coefficients and are described as follows:

- β_0 is the intercept or the baseline win percentage
- β_1 is the expected increase in win percentage for each additional (average) run scored per game, holding all other variables constant
- β_2 is the expected increase in win percentage for each additional unit increase in ERA, holding all other variables constant
- β_3 is the expected increase in win percentage for each additional hit allowed per nine innings, holding all other variables constant
- β_4 is the expected increase in win percentage for each additional stolen base, holding all other variables constant.

3.3 Model Predictors

- **Runs Per Game (`runs_per_game`):** This metric is a direct measure of a team's offensive effectiveness. Historical data shows that higher scoring teams tend to win more games, making this a critical predictor.
- **Team ERA (`ERA`):** ERA is a standard measure of pitching effectiveness and was included to assess how well a team prevents runs. Strong pitching is essential for winning games, thus making this variable vital for the model.
- **Hits Against per 9 Innings (`HA_9`):** This metric measures defensive capabilities and pitching efficiency. Understanding how many hits a team allows can provide insights into overall performance.

- **Total Stolen Bases (`SB`):** Stolen bases reflect a team's aggressiveness and ability to capitalize on scoring opportunities. This variable accounts for strategic elements of gameplay that could influence outcomes.

3.4 Assumptions and Limitations

The model assumes that relationship between predictors (`runs_per_game`, `ERA`, `HA_9`, and `SB`) and the outcome variable (win percentage) is linear. A limitation of the model is that it assumes that the relationships between predictors and win percentage remain constant, but these relationships actually evolve over time due to changes in rules or league dynamics.

3.5 Software Implementation

The model was developed in R (R Core Team 2024) using the `tidymodels` (Kuhn and Wickham 2020) package for multiple regression and the extraction of coefficients.

3.6 Model Validation

Model validation and performance metrics ($R^2 \sim 0.900$, R^2 Adj. ~ 0.897, RMSE ~ 0.02) are available in `scripts/modeling.R`. These metrics indicate that the model is robust, appropriate for the model's simplicity and data limitations, and performs well in prediction based on the chosen predictors.

3.7 Alternative Models

Alternative models considered included some machine learning models like neural networks or gradient boosting machines, but these are not easily interpretable and risk overfitting. Additionally, simple linear regression is not suitable for this analysis, as it is too simplistic to generalize well with multiple predictor variables. The final model balances simplicity with predictive accuracy.

3.8 Model Justification

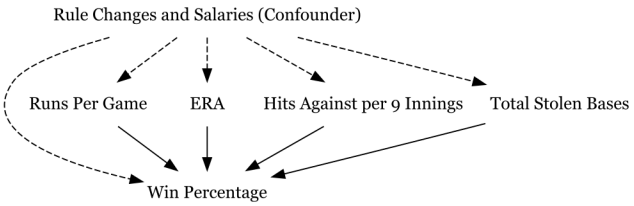

Figure 4: Causal Relationship between Predictor Variables and Win Percentage

The multiple regression model describes the following causal relationship (Figure 4) where some key performance metrics are able to predict winning percentage, excluding baseball rule changes and difference in team salaries. Note that we assume that there exists a relationship between these metrics and winning percentage in the first place. In Figure 4, baseball rule changes and salary difference amongst teams serve as confounding variables since rule changes affect the way baseball is played and the consequent offensive and defensive strategies teams employ, which then have a corresponding impact on the aforementioned variables affecting win percentage over time. In addition, teams willing to spend more on salaries are more often able to sign better players that help improve the team. However, we are unable to observe the precise effects of these directly given our data. These are further expounded on in Section 5.

4 Results

4.1 Interpretation of Coefficients

The model predicting team win percentage (in 2024) features several coefficients that illustrate the influence of various factors as seen in Table 1. The intercept term β_0 establishes the baseline win percentage when all predictors are set to zero, serving as a reference for interpreting other coefficients (although this does not have a viable practical interpretation). The values of β_1 and β_2 imply that the expected increase in win percentage for each additional (average) run scored per game and additional unit increase in ERA are about 9.3% and -9.3% respectively. An increase of -9.3% is equivalent to a decrease in win percentage of 9.3%. The values of β_3 and β_4 show that the expected increase in win percentage for each additional hit a team gets against them (per 9 innings) and additional stolen base are about 0.2% and 0.0% respectively, implying that these two measures may not be significant for team success.

Furthermore, from Table 1 we note that the model explains about 90.0% of the total variability in the data, with 89.7% of the variance in `win_pct` explained by the predictor variables included in the model, suggesting strong explanatory power. The effects of `runs_per_game` and `ERA` on in percentage relatively large compared to that of `HA_9` and `SB`. The very low value of the

Table 1: Model Summary of Multiple Regression Model

	(1)
(Intercept)	0.487
	(0.049)
runs_per_game	0.093
	(0.005)
ERA	−0.093
	(0.007)
HA_9	−0.002
	(0.007)
SB	0.000
	(0.000)
Num.Obs.	120
R2	0.900
R2 Adj.	0.897
AIC	−545.2
BIC	−528.5
RMSE	0.02

RMSE provides measures of the model's predictive accuracy and suggests that its predictions are quite accurate.

4.2 Margin of Error on 2024 Data Model Predictions

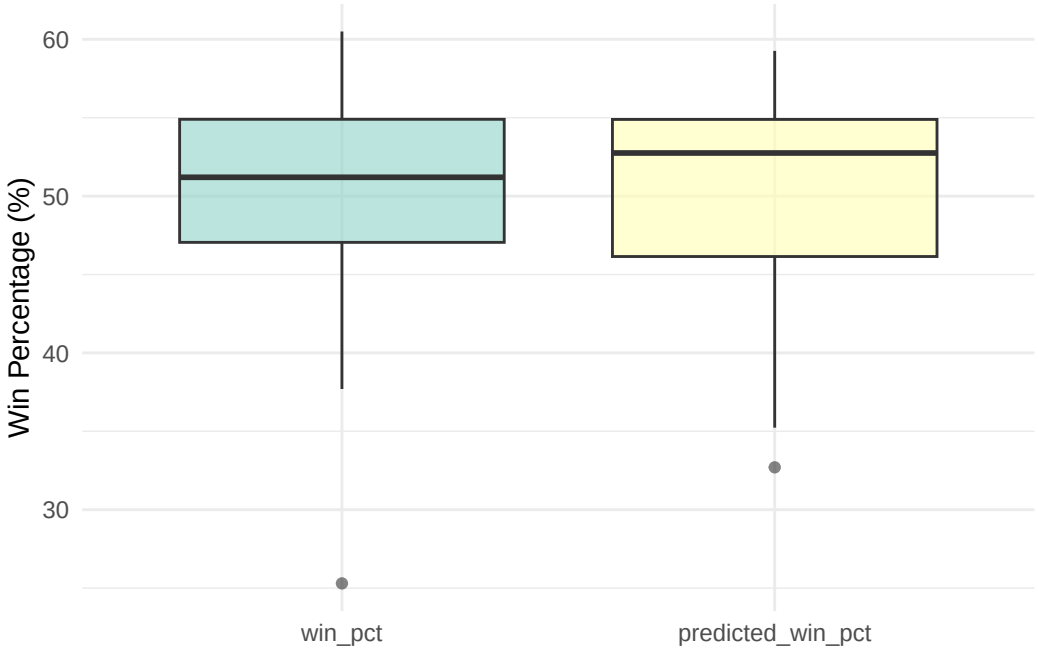

Figure 5: Distribution of Actual vs Predicted Team Win Percentage in 2024

Figure 5 shows the distribution (using boxplots) of the actual win percentage (`win_pct`) and the win percentage predicted by the model (`predicted_win_pct`). We see that the range of values for `predicted_win_pct` are tighter, and the median win percentage in `predicted_win_pct` is slightly higher than that of `win_pct`. We can delve into these results further by looking at Figure 6.

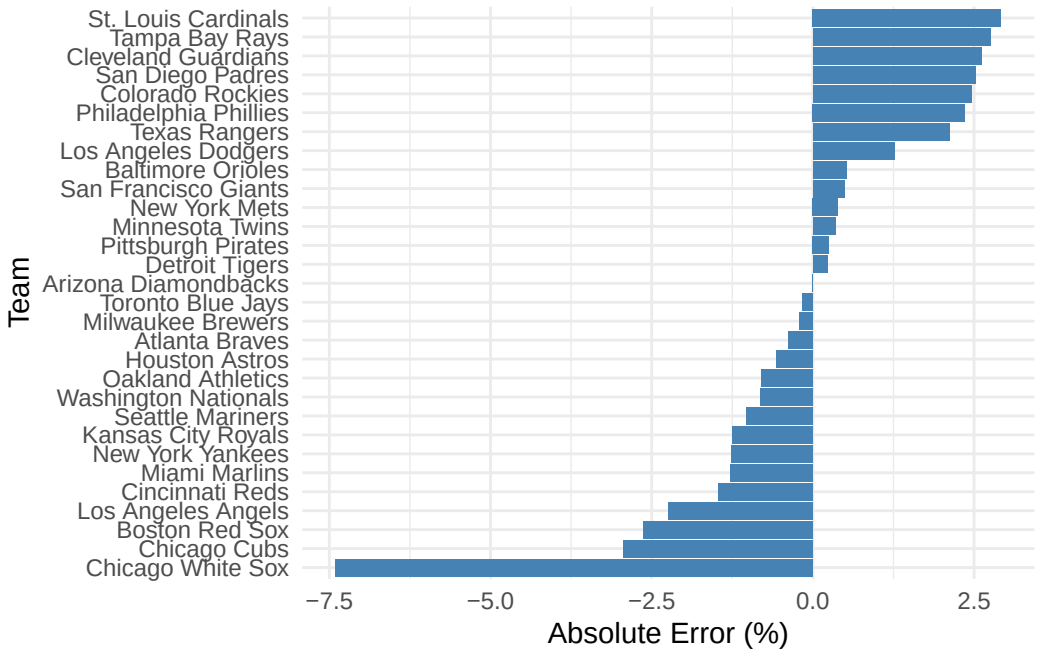

Figure 6: Distribution of Errors in Actual vs Predicted Team Win Percentage in 2024

Figure 6 shows the absolute errors between the actual and predicted win percentage. At first glance, we see that the difference in win percentages range from about -7.4% to about +2.9%. We note that the largest error comes from the Chicago White Sox who had the worst season in the Modern Era (since 1901) (Close 2024). Removing this outlier, the range of absolute errors decrease to about -2.94% to +2.9% (the full table can be found in Section A in Table 2) equivalent to about ± 4.7 wins.

In addition, it is interesting to note that along with the almost symmetrical range of absolute errors, Figure 6 shows that the model over-estimates and under-estimates the win percentages of almost the same number of teams (14 and 15 respectively), with only the Arizona Diamondbacks being the only team that the model predicts exactly right (error of 0.002%). This suggests the model is not exhibiting any bias towards consistently over- or under-predicting team performance; the errors appear to be randomly distributed around the true values. This symmetry in errors indicates that the model captures overall trends and relationships in the data reasonably well. However, the fact that there are still notable deviations points to the inherent challenges in predicting the unpredictable nature of baseball performance.

5 Discussion

5.1 Runs Per Game and ERA are Highly Important for Winning

From the coefficients in Table 1, we can see that run scoring and team ERA highly determine winning percentage. The (absolute) values of `runs_per_game` and `ERA` suggest that improving either run scoring or run prevention (or both) by can have an relatively great positive impact on a team's expected win percentage. This has important implications for team building and roster construction; it suggests that teams should prioritize investing in offensive talent that can consistently produce runs over the course of a season, as well as developing and acquiring high-quality starting and relief pitchers with low ERA. to maximize a team's winning chances. There do exist factors that affect the number of runs per game and ERA over a season, namely park factors (whether a ballpark is "hitter-friendly" or "pitcher-friendly"), opposing offenses, and the like. However, ERA is ultimately a reflection of a team's defense and overall pitching staff quality.

As a matter of fact, the correlation between `runs_per_game` and `ERA` has been studied for many decades. In an article from the 1976 Baseball Research Journal, we see that using games from 1920-1959, the correlation of the number of runs scored (which is runs_per_game multiplied by the number of games played) and ERA to winning percentage was found to be .737 and .743 respectively (Wiley 1976); note that some rules and the way baseball was played was different then than it is now). Of course, there are many other factors that also contribute to team success, but all in all the dominant influence of runs scored and run prevention, as reflected in Table 1, underscores their great importance in determining which teams see the most regular season success.

5.2 Hits Against per 9 Innings and Stolen Bases Are Not as Essential as We Think

From Table 2, we can see that the Washington Nationals boast the highest number of stolen bases in 2024, while the Atlanta Braves, with the third-lowest stolen bases, boast an 11% higher win percentage compared to the Nationals. This contrast highlights that while stolen bases can enhance offensive potential, they do not directly translate to wins without effective run conversion. This is corroborated by an article from Samford University, where it states that stolen bases have a very weak — almost negligible — negative correlation with the number of wins, with $R^2 = 0.003$ (Murray 2022). In addition, a lower hits against per nine innings (`HA_9`) is generally advantageous, but it does not tell the full story; teams must convert those hits into runs, which can be challenging due to factors like double plays and defensive plays.

For example, take the situation of the 2020 New York Mets. They had the fifth best wRC+ (which stands for Weighted Runs Created +, a statistic which evaluates how effective a team creates runs relative to league average) in MLB history (Foolish Baseball 2022). However, the

Mets were statistically worse with runners on base (with offensive production regressing back to about league average compared to when there were no runners on base). As such, even if they were able to get a lot of hits against teams, they were not able to bring baserunners (and more importantly runners in scoring position) home to actually score runs. It is important to note, however, that teams can still score runs even if they allow fewer hits by utilizing walks, hit-by-pitches (HBP), and other non-hit methods to get runners on base. Additionally, stolen base effectiveness can be influenced by various factors including the catcher's arm strength, pitchers' delivery times, and managerial philosophies (some managers and hitters are more aggressive in their base-stealing tendencies).

5.3 Winning Together: ERA and Scoring Runs are Interdependent

So far, from our previous discussions, we see that low ERA and scoring as many runs as possible are of paramount importance to winning. However, to win games throughout the season, teams must be able to do both at the same time; teams cannot solely focus on offensive or pitching (defensive) firepower.

To see why this is the case, we first see that run support, despite high ERA, can win teams more games, and pitchers with low ERA cannot be the sole engine to winning games throughout the season. In 2018, Jacob deGrom pitched at an elite level with a 1.70 ERA, but finished with a subpar win-loss record of 10-9 due to poor run support. In contrast, in the 2015 season Toronto Blue Jays pitcher Drew Hutchison had a win-loss record of 13-5 despite a 5.57 ERA due to receiving strong run support (Foolish Baseball 2021). While pitcher win-loss records do not paint the complete picture for the quality of pitchers, these examples illustrate that while individual pitching performance is critical, it is ultimately the ability to convert opportunities into runs that determines success on the field.

On the other hand, scoring runs at a high rate might not always translate to success. In 2014, the Colorado Rockies scored an average of 4.66 runs per game, good for third-best in the league. However, the team also had a league-worst 4.84 ERA. Consequently, the team had the second-worst record in the league with a 66-96 record (equivalent to about a 40.7% win rate).

5.4 Limitations and Areas for Future Research

A limitation of this analysis lies in the nature of the data used. As previously mentioned in Section 3, we are unable to directly observe the precise effects of rule changes. The offensive and defensive statistics of teams must ideally be adjusted for some other unseen factors affecting the data itself, such as teams' strength of schedule, as teams do not play each other the same number of times especially for teams within the same division. In 2022, the MLB revised the schedule where more interleague games were played (Feinsand 2022). In 2023, the MLB changed the structure of the schedules again where "teams will be playing 24 fewer games

against their own divisions than they did before" (Petriello 2023). This can lead to disparities in the quality of opponents faced, which could impact the observed relationships between the predictors and win percentage.

Additionally, the lack of a salary cap in Major League Baseball introduces another potential confounding factor. Big-spending teams, such as the 2024 Los Angeles Dodgers who signed Shohei Ohtani to a 10-year, $700 million contract (Wexler 2023), have a clear advantage in their ability to recruit and retain top talent compared to small-market franchises. However, the Lahman Database only includes salary information up to 2016; this limits the ability to directly analyze and account for payroll and consequent roster quality differences in the current analysis.

Another limitation is the inability to directly observe the precise and immediate effects of baseball rule changes on team statistics and win percentage. For example, the ban on defensive shifts, larger bases, and new disengagement rules introduced in 2023 are believed to have contributed to a league-wide increase in runs scored by approximately 0.3 per game compared to 2022 (Clemens 2023). Incorporating the necessary adjustments to data from the all these rule changes era could provide valuable insights into how alterations to the game's rules and regulations impact the relative importance of various performance metrics in predicting team success.

Future research could explore these limitations in several ways. Incorporating adjustments for schedule strength, accessing more recent salary data, and studying the effects of rule changes could lead to a more comprehensive understanding of the complex factors driving team win percentages in MLB. Overall, addressing these limitations could provide valuable insights to enhance the predictive models and understanding of the key drivers of success in professional baseball.

A Appendix

A.1 Datasheet

Accompanying the paper, a datasheet that contains an in-depth discussion of the nature of the Lahman Baseball Database can be found in `other/datasheet/datasheet_lahman.pdf`.

A.2 Additional Tables and Data Details

Table 2 below shows a complete table containing the outcome variable and predictor variables used for the model in this paper, as well as the model's predictions.

Table 2: 2024 Team Data with Model Predictions

Team Name	ERA	HA/9	Runs Per Game	SB	Win %	Predicted Win %
Arizona Diamondbacks	4.62	9.2	5.47	119	54.9	54.898
Atlanta Braves	3.49	8.0	4.35	69	54.9	55.287
Baltimore Orioles	3.94	8.1	4.85	98	56.2	55.673
Boston Red Sox	4.04	8.4	4.64	144	50.0	52.627
Chicago Cubs	3.78	8.2	4.54	143	51.2	54.144
Chicago White Sox	4.68	8.9	3.13	90	25.3	32.703
Cincinnati Reds	4.09	8.2	4.31	207	47.5	48.968
Cleveland Guardians	3.61	7.7	4.40	148	57.1	54.488
Colorado Rockies	5.47	10.1	4.21	85	37.7	35.234
Detroit Tigers	3.61	7.9	4.21	76	53.1	52.871
Houston Astros	3.74	7.8	4.60	93	54.7	55.265
Kansas City Royals	3.76	8.2	4.54	134	53.1	54.352
Los Angeles Angels	4.56	8.4	3.92	133	38.9	41.136
Los Angeles Dodgers	3.90	7.9	5.20	136	60.5	59.236
Miami Marlins	4.73	9.0	3.93	125	38.3	39.574
Milwaukee Brewers	3.65	8.0	4.80	217	57.4	57.613
Minnesota Twins	4.26	8.3	4.58	65	50.6	50.245
New York Mets	3.96	7.7	4.74	106	54.9	54.511
New York Yankees	3.74	7.9	5.03	88	58.0	59.261
Oakland Athletics	4.37	8.7	3.97	98	42.6	43.402
Philadelphia Phillies	3.85	8.4	4.84	148	58.6	56.239
Pittsburgh Pirates	4.15	8.6	4.10	106	46.9	46.648
San Diego Padres	3.86	8.1	4.69	120	57.4	54.872
San Francisco Giants	4.10	8.4	4.28	68	49.4	48.915
Seattle Mariners	3.49	7.4	4.17	140	52.5	53.532
St. Louis Cardinals	4.03	8.5	4.15	91	51.2	48.280
Tampa Bay Rays	3.77	8.0	3.73	178	49.4	46.648
Texas Rangers	4.35	8.3	4.22	97	48.1	45.981
Toronto Blue Jays	4.29	8.3	4.14	72	45.7	45.857
Washington Nationals	4.30	9.0	4.07	223	43.8	44.616

A.3 Additional Model Details (Prediction Error and Pythagorean Expectation)

As previously discussed in Section 3, the model evaluation relies on R^2 and RMSE as these metrics suggest that our model predictions align closely with the actual win percentage of each team. However, some statisticians have come up with various ways to estimate a teams win percentage without the need for prediction models. Here, we explore a formula by Bill James, an influential baseball writer and statistician, known as Pythagorean Expectation, as well as how well it aligns with the actual win percentage of each team. The formula is as follows:

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

Applying this to 2024 data, we can see Table 3 below.

Table 3: Comparing 2024 Team Actual Win Percentage with Pythagorean W-L%

teamID	Runs Scored	Runs Against	Actual Win %	Pythagorean Win %	Error
Arizona Diamondbacks	886	788	54.9	55.83425	-0.9342
Atlanta Braves	704	607	54.9	57.35865	-2.4586
Baltimore Orioles	786	699	56.2	55.83855	0.3615
Boston Red Sox	751	747	50.0	50.26702	-0.2670
Chicago Cubs	736	669	51.2	54.75786	-3.5579
Chicago White Sox	507	813	25.3	28.00043	-2.7004
Cincinnati Reds	699	694	47.5	50.35893	-2.8589
Cleveland Guardians	708	621	57.1	56.51834	0.5817
Colorado Rockies	682	929	37.7	35.02005	2.6800
Detroit Tigers	682	642	53.1	53.01839	0.0816
Houston Astros	740	649	54.7	56.52348	-1.8235
Kansas City Royals	735	644	53.1	56.57037	-3.4704
Los Angeles Angels	635	797	38.9	38.83010	0.0699
Los Angeles Dodgers	842	686	60.5	60.10411	0.3959
Miami Marlins	637	841	38.3	36.45560	1.8444
Milwaukee Brewers	777	641	57.4	59.50355	-2.1036
Minnesota Twins	742	735	50.6	50.47392	0.1261
New York Mets	768	697	54.9	54.83506	0.0649
New York Yankees	815	668	58.0	59.81589	-1.8159
Oakland Athletics	643	764	42.6	41.46328	1.1367
Philadelphia Phillies	784	671	58.6	57.71976	0.8802
Pittsburgh Pirates	665	739	46.9	44.74395	2.1561
San Diego Padres	760	669	57.4	56.34237	1.0576
San Francisco Giants	693	699	49.4	49.56897	-0.1690
Seattle Mariners	676	607	52.5	55.36251	-2.8625
St. Louis Cardinals	672	719	51.2	46.62499	4.5750
Tampa Bay Rays	604	663	49.4	45.35341	4.0466
Texas Rangers	683	738	48.1	46.13528	1.9647
Toronto Blue Jays	671	743	45.7	44.92123	0.7788
Washington Nationals	660	764	43.8	42.73538	1.0646

Since we are interested in the distribution of errors, we see Figure 7 below.

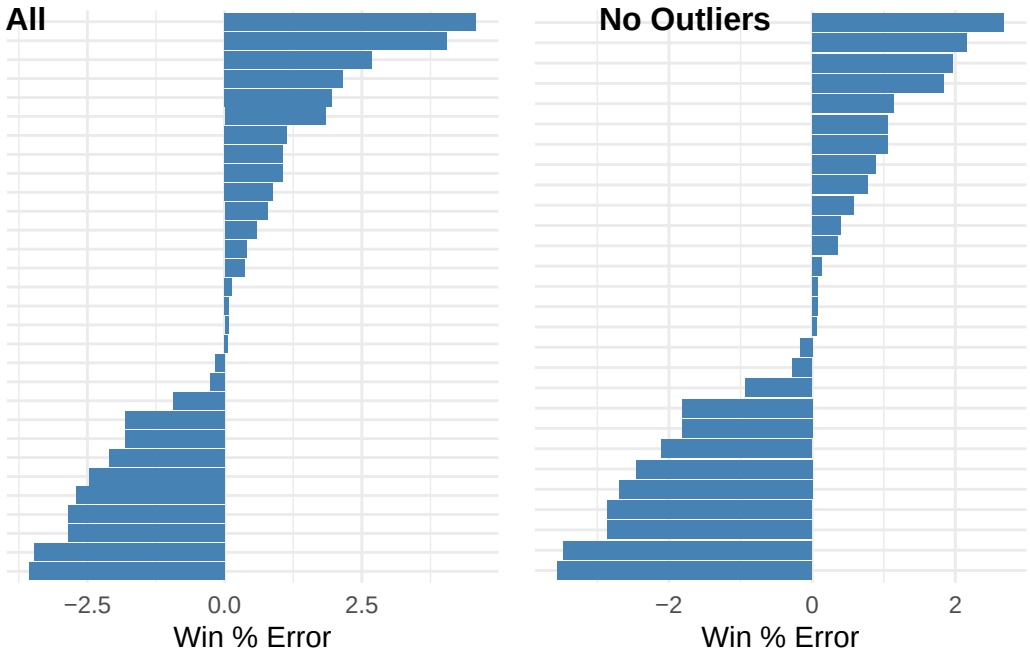

Figure 7: Distribution of Errors in Actual vs Pythagorean Team Win Percentage in 2024

Notice that compared to Figure 6, the error margins are more significant, especially when it comes to underestimating win percentage. We see that 2 teams were predicted to have a win percentage more than 4% less than what they actually achieved. That being said, removing these 2 teams, it can be inferred that the distribution looks very similar to Figure 6. However, notice that Bill James' formula underestimates the winning percentage of more teams compared to the predictive model. That being said, we can see the significant impact of scoring and preventing runs in both cases.

A.4 A Short Discussion on Observational Data and the Lahman Database

Throughout this analysis, we have been using team data from the Lahman database, which is observational in nature. From QuestionPro (2024), a definition for observational data could be data that is "collected by observing and recording events, behaviors, or phenomena as they naturally occur without interference or manipulation." In fact, this is exactly the kind of data contained in the database; it is collected by observing baseball games and seasons and recording the necessary statistics, and MLB game do occur without interference or manipulation of any kind. This is in contrast with experimentally-sourced data, which is data "collected through controlled experiments, where variables are manipulated to observe their

effects on other variables (Easily 2024)." For example, a sports-related experiment might involve simulating games under controlled conditions — altering player lineups, changing park dimensions, or introducing new rules. By isolating variables, experiments can better identify causal relationships. However, experiments may not always replicate the complexity of natural situations, and discrepancies can arise due to inherent variability.

The observational nature of the Lahman database has both strengths and limitations. On one hand, since the data is comprehensive and exact, it allows better analysis of trends and patterns without introducing potential accuracy errors and biases which may otherwise arise if data were experimentally manipulated instead.

On the other hand, the reliance on observational data also means that confounding factors must be carefully considered. In the context of our analysis, changes in league structure, rule changes, or other external influences may impact team performance and winning percentage over time. It is difficult to contextualize and explain these changes by using models alone. Unlike experimental data, observational data does not allow researchers to control for these variables directly, which makes isolating causal relationships difficult. For example, an increase in team payroll may correlate with higher winning percentages, but this relationship could also be influenced by other factors like management quality or market size.

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Clemens, Ben. 2023. "How Have the New Rules Changed the Game?" FanGraphs. https://blogs.fangraphs.com/how-have-the-new-rules-changed-the-game/.

Close, David. 2024. "White Sox Set Record for Most Losses in Modern MLB History." 2024. https://www.cnn.com/2024/09/27/sport/white-sox-most-losses-modern-mlb-spt-intl/index.html.

Easily, Statistics. 2024. "What Is Experimental Data? Explained in Detail." 2024. https://statisticseasily.com/glossario/what-is-experimental-data-explained-in-detail/.

Feinsand, Mark. 2022. "Balanced Schedule to Bring More Interleague Games." https://www.mlb.com/news/more-interleague-games-on-balanced-schedule.

Foolish Baseball. 2021. "The Anti-deGrom | Baseball Bits." YouTube. https://www.youtube.com/watch?v=kBXoIU_YhqQ.

———. 2022. "The 2020 Mets Were Elite, but They Weren't Clutch | Baseball Bits." YouTube. https://www.youtube.com/watch?v=s_3-nAoa1QE.

Friendly, Michael, Chris Dalzell, Martin Monkman, and Dennis Murphy. 2024. *Lahman: Sean 'Lahman' Baseball Database*. https://CRAN.R-project.org/package=Lahman.

Iannone, Richard. 2016. *DiagrammeRsvg: Export DiagrammeR Graphviz Graphs as SVG*. https://CRAN.R-project.org/package=DiagrammeRsvg.

Iannone, Richard, and Olivier Roy. 2024. *DiagrammeR: Graph/Network Visualization*. https://CRAN.R-project.org/package=DiagrammeR.

Kassambara, Alboukadel. 2023. *Ggpubr: 'Ggplot2' Based Publication Ready Plots*. https://CRAN.R-project.org/package=ggpubr.

Kuhn, Max, and Hadley Wickham. 2020. *Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles*. https://www.tidymodels.org.

Lahman, Sean. 2024. "Sean Lahman Donates Lahman Baseball Database to SABR." Society for American Baseball Research (SABR). https://sabr.org/latest/sean-lahman-donates-lahman-baseball-database-to-sabr/.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Murray, Braden. 2022. "MLB Winning Percentage Breakdown: Which Statistics Help Teams Win More Games?" https://www.samford.edu/sports-analytics/fans/2022/MLB-Winning-Percentage-Breakdown-Which-Statistics-Help-Teams-Win-More-Games.

Ooms, Jeroen. 2024. *Rsvg: Render SVG Images into PDF, PNG, (Encapsulated) PostScript, or Bitmap Arrays*. https://CRAN.R-project.org/package=rsvg.

Petriello, Mike. 2023. "How 2023 Balanced Schedule Could Affect Playoff Races." https://www.mlb.com/news/how-2023-mlb-balanced-schedule-could-affect-playoff-races.

QuestionPro. 2024. "What Is Observational Data?" 2024. https://www.questionpro.com/blog/observational-data.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna,

Austria: R Foundation for Statistical Computing. https://www.R-project.org/.
Retrosheet, Inc. 2024. "Retrosheet: Historical Baseball Game Data and Statistics." https://www.retrosheet.org/.
Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.
Society for American Baseball Research (SABR). 2024. "SABR Sabermetrics: Data Resources." https://sabr.org/sabermetrics/data/.
Sports Reference LLC. 2024. Baseball-Reference.com - Major League Statistics; Information. https://www.baseball-reference.com/.
Urbanek, Simon. 2022. *Png: Read and Write PNG Images*. https://CRAN.R-project.org/package=png.
Wexler, Sarah. 2023. https://www.mlb.com/news/shohei-ohtani-contract-with-dodgers.
Wickham, Hadley. 2007. "Reshaping Data with the reshape Package." *Journal of Statistical Software* 21 (12): 1–20. http://www.jstatsoft.org/v21/i12/.
———. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.
———. 2024. *Rvest: Easily Harvest (Scrape) Web Pages*. https://CRAN.R-project.org/package=rvest.
Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.
Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.
Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.
Wiley, George T. 1976. "Computers in Baseball Analysis." *1976 Baseball Research Journal* First Published in 1976, Now Available Online. https://sabr.org/journal/article/computers-in-baseball-analysis/.
Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.