Decision Analysis Research Project: Finding the Highest Contributing Statistic in Major League Baseball Wins Using Top-Down Decision Modeling

Hannah Sutton

Southern New Hampshire University

DAT 520: Decision Methods and Modeling

Prof. Jason Eborn

25 May 2025

I. Research Question and Data Set Evaluation

"Are winning baseball teams built based on pitching or hitting?" The age-old debate on whether pitching or hitting leads to success in a baseball team is one I am quite familiar with. Whether it's looking back to the "steroid-era" of baseball where Barry Bonds and Mark McGuire would frequently crush 500ft home runs, or the modern-day travesty of rookie Paul Skenes dominating the Pirate's game with Ks, only for their closer to come in and blow it; the question runs rampant through baseball fans across the world. By using Power BI to build a decision analysis, I am hoping to once and for all settle the debate.

The data set I am using is The Lahman Baseball Database 1871-2023, created by Sean Lahman. The database contains statistics for Major League Baseball teams from 1871 through 2023, including pitching, hitting, and fielding data. The data comes from the two current leagues (American and National), and four other "major" leagues (American Association, Union Association, Players League, and Federal League), and the National Association of 1871-1875 (Dropbox, n.d.). The intent of this database is to make baseball statistics free to the public, and with the help of researchers, the largest and most accurate source of baseball statistics anywhere (Dropbox, n.d.). From the data set, we will use the team's table (wins, losses), batting table, and pitching table for most of this analysis.

In the Power BI space, I imported each data set, starting with the batting table. I used the "transform data" button to start preparing the data for analysis. To clean the data, I first started with the batting spreadsheet. I made sure to remove the errors and duplicates in each column as well. I repeated this process for the pitching and teams tables. I tried to merge the queries, but since the data set is so large, this took a lot of time. I wanted to do a full outer join of teams and batting, as well as teams and pitching. This is because if you were to only allow matching

columns, you would truly have little data to work with. The statistics are different in each column, so a full outer join allows us to compare each table correctly.

By cleaning the data, we transformed our data for analysis. I chose a full outer join because the data sets had little matching data as they provided completely different statistics. It will be important to see, later, all the available data and how it contributes to wins or losses. Clean data allows for accurate results. Duplicates and errors may skew the predictability of the analyses, forecasting inaccurate probabilities. This would be a grave error for any analyst. Merging the queries allows us to build our analyses with Power Bi, adding values that we can compare. This will be especially important when making data visualizations for understanding.

For data visualization, one possibility could be a stacked bar chart, with the x-axis being team rank and the y-axis including variables from pitching statistics like ERA to batting statistics like HRs. This way you have a side by side visual of where one statistic is higher or lower, rank is higher or lower. There are many other visualizations that can be included in this report as well, as there is a plethora of data to sort through. In the competitive world of Major League Baseball (MLB), the ability to make strategic decisions based on data is crucial. With rising player salaries, advanced statistics, and a growing emphasis on analytics, team managers and analysts must understand what truly drives success on the field. The central research question of this project is: Does batting or pitching contribute more significantly to a team's success in terms of wins? This analysis looks to address this question using a top-down decision model within Power BI, allowing for an exploratory and intuitive breakdown of the factors contributing to team victories. The insights derived from this model can influence team-building strategies, resource allocation, and performance evaluation.

II. Conducting a Decision Analysis

In the competitive world of Major League Baseball (MLB), the ability to make strategic decisions based on data is crucial. With rising player salaries, advanced statistics, and a growing emphasis on analytics, team managers and analysts must understand what truly drives success on the field. The central research question of this project is: Does batting or pitching contribute more significantly to a team's success in terms of wins? This analysis seeks to address this question using a top-down decision model within Power BI, allowing for an exploratory and intuitive breakdown of the factors contributing to team victories. The insights derived from this model can influence team-building strategies, resource allocation, and performance evaluation.

To explore the relationship between team performance and game outcomes, I used the "Teams" dataset in the Lahman Baseball Database, containing season-level statistics for all MLB teams from 1871 to 2021. By using the "Teams" dataset, I did not have to conduct a full outer join. This led to smoother analysis, easier cleaning and mining, and less statistical input to be worried about.

A top-down model was chosen for this analysis, using the decomposition tree in Power BI. This approach begins with the final outcome, team wins, and breaks it down by contributing factors. The top-down model is appropriate because the research question focuses on what drives an end result, rather than aggregating inputs from individual players. "In the top-down model, an overview of the system is formulated without going into detail for any part of it. Each part of it is then refined into more details, defining it in yet more details until the entire specification is detailed enough to validate the model (GeeksforGeeks, 2025b)." The top-down model allows us to answer our question by starting with the statistic we are questioning and expanding down with different variables to answer ourselves.

In Power BI, the decomposition tree visual was configured as follows:

- Target Variable (Analyze): Winning Percentage (WPct) - (formula: W/ (W + L)), derived manually, WPct > 0.50 indicates winning season
- Input Predictor Variables (Explain By):
    - Batting Average (BA) - (formula: H/AB) total hits divided by total at-bats, derived manually
    - Slugging Percentage (SLG) - (formula: (1B + 2Bx2 + 3Bx3 + HRx4)/AB)), derived manually
    - Earned Run Average (ERA) - derived from data set
    - Walks Hits Per Innings Pitched (WHIP) - (formula: walks (BB) + hits allowed (HA) / total innings pitched (IPouts / 3)), derived manually (Glossary | MLB.com, n.d.)

Upon expanding the decomposition tree, Power BI found which variables had the most significant impact on Winning Percentage above 50%. In the dataset, ERA and WHIP emerged as stronger contributors to winning percentage than BA and SLG. This suggests that effective pitching metrics correlate more closely with team success than batting performance.

This visual representation allows decision-makers to interactively explore how different combinations of metrics impact the outcome. For instance, a team with high offensive output (BA and SLG) but poor WHIP may underperform compared to a team with superior pitching stats.

Although batting average was the statistics first pulled by Power BI, it has too much variance to allow it to be a significant statistic. We know this because the slugging percentage variable did not pull as a high contributor as well.

Figure 1:

Decomposition Tree Model showing order of metrics that highly correlate to Winning Percentage:

III. Revision and Evaluation of the Decision Tree Model

After conducting the decision analysis, I revised and evaluated my top-down decision model developed in Power BI, which investigates whether batting or pitching contributes more significantly to a baseball team's success. The original model used a decomposition tree to analyze average Winning Percentage (WPct), with Batting Average (BA), Slugging Percentage (SLG), Earned Run Average (ERA), and Walks plus Hits per Inning Pitched (WHIP) as explanatory variables. Based on feedback and discussions from peers using the same research question, I focused on testing the model's structure, evaluating the strength of its results, and performing a sensitivity analysis to determine how key variables influence team performance outcomes.

The decomposition tree visual in Power BI was configured with WPct as the analyzed metric and BA, SLG, ERA, and WHIP as the "Explain by" inputs. Upon running the model, Power BI consistently prioritized ERA and WHIP as the most influential splits, often appearing early in the tree's decision path. This suggests that pitching metrics were more predictive of a team's win percentage than batting statistics.

To validate and challenge the model, I restructured the tree in several ways. First, I manually removed ERA to observe the effect of its exclusion. Without ERA, Power BI was less capable of producing clear splits or meaningful differentiation in WPct. I then excluded SLG and retained the two pitching metrics and BA. The model maintained a strong explanatory structure, reinforcing that SLG may be redundant or less impactful in this context. Additionally, when both pitching metrics were removed, the tree struggled to explain variation in WPct effectively. This iterative testing helped confirm that ERA and WHIP are not only impactful but also essential to the model's explanatory power.

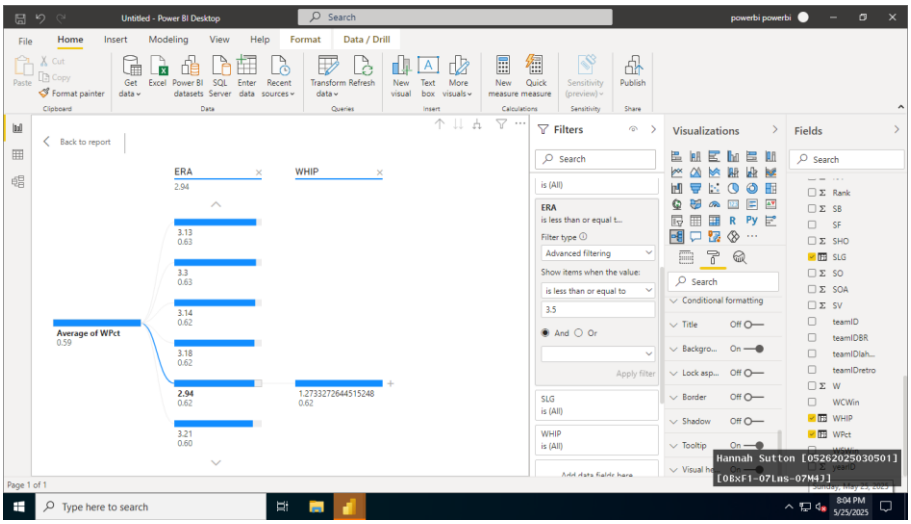
To further evaluate the model, I conducted a sensitivity analysis focusing on threshold behavior. I found that teams with an ERA below 3.50 often achieved WPct values above 0.600, while those with WHIP values above 1.35 were often below 0.500. These metrics appear to serve as informal thresholds where a minor change can significantly shift the model's outcome path. When these thresholds were crossed in the tree, the model "flipped," resulting in distinct performance groups. This insight highlights how certain performance metrics serve as tipping points and can be used to forecast or diagnose a team's likelihood of winning.

Figure 2:

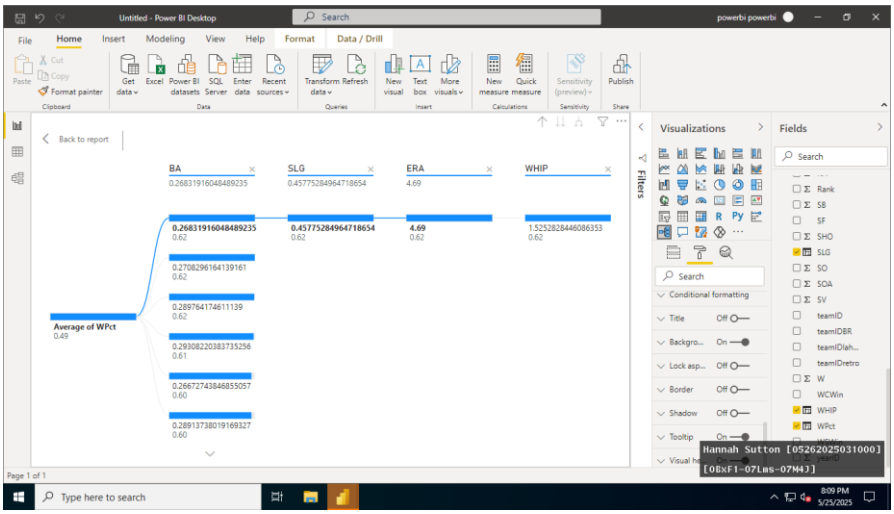
Analysis of ERA< 3.50

Figure 3:

Flipped model after applying thresholds



in conducting sensitivity analysis



When I excluded key pitching metrics like ERA or WHIP from the model, the tree's ability to segment WPct meaningfully diminished, with BA and SLG offering weaker explanations and less dramatic WPct separation. This suggests that while batting metrics contribute, they lack the consistent predictive power seen in pitching metrics. On the other side, when batting metrics were excluded and only pitching data was retained, the model continued to perform strongly, reinforcing that pitching is the primary driver of team success in this dataset.

Common modeling errors were also considered throughout the revision. One frequent mistake in performance modeling is using raw win totals instead of WPct, which fails to normalize for the number of games played. Another error is aggregating WPct using summation rather than averaging, which can distort outcomes. I avoided these by properly calculating WPct as W / (W + L) and setting Power BI's aggregation to average. I validated each derived column against sample calculations to avoid misinterpretation by the decomposition tree.

Overall, the revised model performs well in explaining the variation in team success and supports the conclusion that pitching is a more consistent predictor of winning than batting. The decomposition tree structure is well-suited for this type of exploratory decision modeling, and the insights drawn are both actionable and intuitive. The results suggest that team managers and

analysts should prioritize improvements in pitching when aiming to boost overall team performance.

The top-down decomposition tree provided an effective framework for analyzing baseball team performance. The visual structure, combined with sensitivity testing and diagnostic evaluation, supported a clear understanding. The model demonstrates that metrics related to pitching, particularly ERA and WHIP, are critical factors in winning MLB games.

One of the strengths of the top-down decomposition model is its agility. The structure can easily accommodate new data from different seasons or leagues, enabling ongoing analysis and adaptation. This model could be extended to forecasting future team performance, supporting decisions on player trades or acquisitions, or informing budget allocation between offensive and defensive talent.

The results from this analysis can be directly applied by team managers, coaches, and analysts. For example, a front office deciding whether to invest in a star pitcher or a power hitter could use insights from this model to prioritize pitching if it has historically contributed more to wins. Visual reports generated in Power BI can be easily shared across departments, supporting data-driven decision-making at all levels. To move from analysis to action, teams could develop performance benchmarks based on key pitching and batting metrics. These benchmarks can then inform scouting, player development, and game-day strategies.

The dataset used in this analysis is composed of comprehensive, publicly available statistics, which minimizes concerns related to data privacy or ethical misuse. However, it is still essential to cite data sources properly and avoid drawing unjustified conclusions that could influence individual player assessments unfairly. “When data are directly connected with publications and vice versa, they become easier to find, thereby increasing research transparency

and reproducibility while also encouraging data reuse for new research and synthesis (Brown, 2021).” So, ethically and legally, it is of best interest to cite any data source to avoid legal implications in the research process.

Data-driven decision-making challenges the traditional scouting method and player evaluation (Popescu, 2025). This could potentially pose risks to team ethics, taking the “human” out of the game and relying solely on numbers. However, in my opinion, major league baseball is a performance-based job at the end of the day. If there is a way to measure a team, player, or coach’s performance in relation to team outcomes, then it should be implemented.

This project demonstrates how a top-down decision model in Power BI can be used to find the relative impact of batting and pitching on a baseball team's success. By focusing on team-level outcomes and breaking them down into measurable components, the decomposition tree provides a powerful tool for uncovering actionable insights. While pitching showed a stronger correlation with wins in this dataset, the flexible structure of the model allows for ongoing exploration and refinement. This approach helps bridge the gap between data analysis and strategic execution in the ever-evolving landscape of professional baseball.

Works Cited

Dropbox. (n.d.). https://www.dropbox.com/scl/fi/5y1fts4i2ubhn65f0plmj/readme2023.txt?rlkey=l9wpr7qwaqcqsdhpyftw7xyuw&st=g0j9yg4c&dl=0

Brown, R. F. (2021). The importance of data citation. *BioScience*, *71*(3), 211. https://doi.org/10.1093/biosci/biab012

GeeksforGeeks. (2025b, April 5). *Difference between BottomUp Model and TopDown Model*. GeeksforGeeks. https://www.geeksforgeeks.org/difference-between-bottom-up-model-and-top-down-model/

Glossary | MLB.com. (n.d.). MLB.com. https://www.mlb.com/glossary

Popescu, B. (2025, March 17). Moneyball and the Sabermetrics Revolution: How Data Changed Baseball Forever. *SuchBaseball*. https://suchbaseball.com/moneyball-and-the-sabermetrics-revolution/