# Using Historical Data to Predict Batting Success

Donna J. Harris (994042890)

**Abstract**

This project aims to determine various ways in which statistical Major League Baseball batting data can be used to predict batting success in current players. The primary approach is to study the data of highly productive batters historically and discover patterns that successfully identify players on track to attain similar successes.

## 1. Description of Applied Problem

Baseball is a sport that is driven by data and data analysis. The movie Moneyball starring Brad Pitt brought public attention to how data analytics are being used in Major League Baseball to make personnel decisions, with teams aiming to maximize wins while minimizing costs (Wikipedia contributors, 2022b). The field of sabermetrics has developed over the decades, from the most basic statistics of batting averages and hits to increasingly comprehensive statistics that attempt to bring more meaning to the data and value to both the business and fans of baseball (Wikipedia contributors, 2022a).

It is a fascinating idea to use data in order to predict and maximize results in baseball. The goal of this investigation is to identify markers within historical batting data based on regular season plate appearances. The markers will interpret and ultimately help to predict the effectiveness of a batter. By looking at decades of data and thousands of plate appearances, the desired outcome is to devise and validate one or more models that predict effective batting in Major League Baseball and also to propose players who show evidence of becoming the game's next heavy hitting stars, based on the findings.

Many factors exist which are beyond the scope of traditional statistics - and therefore this project. The context of a player's and a team's situation, both on and off the field, bears significant weight on the development and success of an individual player. Disregarding those factors, for the purposes of this work, I will use traditional baseball statistics and metrics to determine the overall efficacy of a batter at the plate. Does a player get on base? Do they stay on base? Does the player score and also help their teammates to score? When a batter does these things, they are contributing to their team scoring runs. While their team still might not win (since the other team can always score more runs), the impact that individual player makes is highly valued by the business, teammates, and fans alike.

## 2. Description of Available Data

The primary data source for this project is a Kaggle dataset called "MLB Batting Stats By Game 1901-2021" (HugeQuiz.com, 2021). It includes every regular season game's batting statistics for all players with at least one plate appearance from 1901 to 2021. Its data was collected by HugeQuiz.com from the well respected data website Baseball-Reference.com. The original dataset contains plate appearances for 15985 unique players across 120 years of baseball data and is described in Table 1.

*Table 1.* Description of "MLB Batting Stats By Game 1901-2021" dataset.

| Feature | Description |
|---|---|
| ID | Player ID - Unique to each player |
| Player | Player Name |
| Date | Date of Game[1] |
| Tm | Player's Team |
| Opp | Opponent |
| Rslt | Game Result |
| PA | Plate Appearances |
| AB | At Bats |
| R | Runs |
| H | Hits |
| 2B | Doubles |
| 3B | Triples |
| HR | Home Runs |
| RBI | RBIs (Runs Batted In) |
| BB | Base On Balls (Walks) |
| IBB | Intentional Base On Balls |
| SO | Strikeouts |
| HBP | Hit By Pitch |
| SH | Sacrifice Hits |

[1] "Date" may also include an appended value, such as (1) or (2), which indicates the first, second, etc. game of multiple games played that day.

| Feature | Description |
|---|---|
| SF | Sacrifice Flies |
| ROE | Reached On Error |
| GDP | Grounded Into Double Play |
| SB | Stolen Bases |
| CS | Caught Stealing |
| WPA | Win Probability Added |
| RE24 | Base-Out Runs Added |
| aLI | Average Leverage Index |
| BOP | Batting Order Position |
| Pos Summary | Positions Played |
| DFS(DK) | Daily Fantasy Sports Points - Draftkings |
| DFS(FD) | Daily Fantasy Sports Points - Fanduel |

It is anticipated that additional statistical and historical data will be used to help interpret this dataset and make more informed decisions relating to the markers of batting success. Any supplemental information will come from authoritative sources, such as Major League Baseball (MLB.com) and the Baseball-Reference.com website.

## 3. Analysis and Visualization Techniques

### 3.1. Pre-Processing Phase

Older data records do not contain certain statistical details that more modern data includes. Part of pre-processing will involve identifying different eras of data recording and deciding how to best categorize and use (or not use) their less complete data.

Checks for data corruption will also be executed, looking for data values that do not make sense for the context that may have been incorrectly inputted or otherwise accidentally altered. Also, a number of fields appear to be data typed improperly. For instance, RBI (runs batted in) can never be a partial value and yet is stored as a real number (float) in this dataset. Setting features like these features to the proper type will be a part of pre-processing.

For the purposes of this project, I will not be considering any fantasy sports points. Those columns will be removed. Additionally, I will expand some of the existing details from the Date and Rslt columns into their own columns, isolating the season year, win/loss result, and the final score information.

### 3.2. Analysis Phase

In order to help determine batting success, first a number of established and advanced baseball metrics will be calculated from the cleaned data, including metrics such as batting average (AVG), on-base percentage (OBP), and slugging percentage (SLG). These statistics are very commonly portrayed in broadcasting as a measure of batting success (Arth & Billings, 2021). Digging deeper into the data will provide truer insights into the actual productivity of hitters and show which statistics bring value with them.

*Curve Ball: Baseball, Statistics, and the Role of Chance in the Game* is a book which will prove to be a great resource of ideas and knowledge on the quest to find useful models for batting success prediction (Albert & Bennett, 2003). Experimenting with the ideas in this book and applying my own ideas as I discover them will be a critical part of this project. Analysis will involve examining the greater dataset but also looking at individual, successful players of interest. I will be looking for patterns during specific segments within their batting history and looking to find if there are correlations between their identified patterns and those of other players.

### 3.3. Visualization

Visualizations will be used to help find patterns and make sense of this large dataset. Scatter plots will be an important visual tool for seeing the relationships between variables and for predicting future trends. Bar graphs may also be employed when classifying aspects of the data.

## 4. References

Albert, J., & Bennett, J. (2003). Curve Ball: Baseball, Statistics, and the Role of Chance in the Game (Softcover reprint of the original 1st ed. 2001 ed.). Copernicus.

Arth, Z. W., & Billings, A. C. (2021). Batting Average and Beyond: The Framing of Statistics Within Regional Major League Baseball Broadcasts. International Journal of Sport Communication, 14(2), 212–232. https://doi.org/10.1123/ijsc.2020-0112

HugeQuiz.com. (2021, October 28). MLB Batting Stats By Game 1901–2021 [Dataset]. Baseball-Reference.com. https://www.kaggle.com/datasets/darinhawley/mlb-batting-stats-by-game-19012021

Wikipedia contributors. (2022a, May 7). Sabermetrics. Wikipedia. Retrieved May 26, 2022, from https://en.wikipedia.org/wiki/Sabermetrics

Wikipedia contributors. (2022b, May 19). Moneyball (film). Wikipedia. Retrieved May 26, 2022, from https://en.wikipedia.org/wiki/Moneyball_(film)