# Data Science 2: Statistics for Data Science

## Summary Paper of the Analysis & Predictive Modelling of Strikeouts vs Home Runs in Major League Baseball Players 1871 - 2000

### Submitted by Group 8

Han Chen

Andrew Lagree

Rebecca Manning

Charlie Wang

Nathan Ye

Johanna Yu

### Supervised By:

Dr. Sergiy Nokhrin

Dr. Amanda Easson

Dr. Prakriti Sodhi

December 2021

# Data Source

Statistical analysis has been embedded in the fabric of baseball since the inception of Major League Baseball (MLB). A hundred years before it reached public consciousness, what would later be coined “sabermetrics” were being used in clubhouses to make key decisions about which players to draft, how to best utilize their skill set, and when to let them go. [[1]](#footnote-0)

[Lahman’s Baseball Database](http://www.seanlahman.com/baseball-archive/statistics/), which consists of pitching, hitting, and fielding data for Major League Baseball (MLB) from 1871 to 2020, is among the most widely used, accurate sources of MLB statistics on the internet. We selected from [this comprehensive database](https://github.com/chadwickbureau/baseballdatabank/archive/master.zip) the “Batting.csv” file.

We selected this particular data set because (1) batting statistics are of particular interest in MLB, as these metrics are fundamental to whether a team wins or loses games, (2) it is a rich data set, containing 108,789 lines of data, with batting statistics of 19,898 players across 149 years, sufficient for analysis and building a predictive model, and (3) the popularity of sabermetrics will allow for further research and comparison to reference models against which to evaluate our own.

# Analysis & Model

This data set captures individual batting data across 22 categories, including player ID, team, year, games played, number of at bats, runs, and hits. Of these, we will analyze to determine if there is a relationship between the number of home runs (HR) and number of strikeouts (SO) that a MLB player has over a given period of time. Our hypothesis is that players who hit more HRs swing the bat more often and will therefore also have more SOs, such that the number of HRs can be predicted based on the number of SOs.

1. Analysis or Model: Explain the model/analysis and how you confirmed the validity of the approach. What were the challenges? How did you overcome them?

# Conclusions

1. Conclusions: What did you learn about your dataset?

1. A guide to sabermetric research. Society for American Baseball Research. (n.d.). Retrieved December 5, 2021, from https://sabr.org/sabermetrics. [↑](#footnote-ref-0)