# Data Science 2: Statistics for Data Science

## Report on the Analysis & Predictive Modelling of Strikeouts vs Home Runs in Major League Baseball Players 1871 - 2020

### 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

# Objectives

What are you setting out to prove or predict? What is your rationale for there being a correlation in the data that you’re looking to confirm and/or exploit?

# Data Preparation

We used an online baseball statistics database from <http://www.seanlahman.com/baseball-archive/statistics/>. The databases are free, with copyright 1996-2021 by Sean Lahman. The website contains complete pitching, batting, and fielding statistics from Major League Baseball from 1871 to 2020. It includes data from the two current leagues (American and National), the four other "major" leagues (American Association, Union Association, Players League, and Federal League), and the National Association of 1871-1875).

For our assignment, we used the batting statistics database (Batting.csv). There were 22 column variables, with 108,789 row entries. The column variables were defined as follows:

playerID Player ID code

yearID Year

stint Player's stint (order of appearances within a season)

teamID Team

lgID League

G Games

AB At Bats

R Runs

H Hits

2B Doubles

3B Triples

HR Homeruns

RBI Runs Batted In

SB Stolen Bases

CS Caught Stealing

BB Base on Balls

SO Strikeouts

IBB Intentional walks

HBP Hit by pitch

SH Sacrifice hits

SF Sacrifice flies

GIDP Grounded into double plays

For our assignment, we focused on data for 4 variables of interest, specifically playerID, yearID, HR, and SO.

Missing Values:

Overall the data quality was very good. However, there were 2,100 missing values for the SO variable (comprising 1.9% of the total 108,789 row entries). There were a few ways we could have dealt with those 2,100 missing values for SO:

a) We could fill in the missing values with the median or mean of the SO variable

b) We could impute a value based on grouping of similar characteristics for other variables

c) We could drop the rows with missing SO values

Because we wanted to have the most accurate estimate of the association between HR and SO, we wanted to minimize bias and confounding. We felt that trying to fill in missing values with the median or mean, or trying to impute values could have created some bias and confounding. We wanted to have the purest unaltered data available.

Therefore, we felt that it was most appropriate to drop the rows with missing SO values. 2,100 missing values out of a total of 108,789 total entries represented only 1.9% of the total sample. Dropping those rows still left 106,689 rows of data. Therefore, dropping the rows with missing values should have a negligible effect on the analyses.

Unique Players:

Inspecting the data frame further, we noted that the same playerID’s appeared in multiple rows. This represented the same player playing in multiple different seasons. After dropping the rows with missing SO values, there were 19,445 unique players among the 106,689 remaining row entries.

Since each observation in an analysis should be an independent observation, when conducting our analyses, we merged the data for rows with the same playerID (e.g., the same player but data for a different season).

Feature Engineering:

Since we wanted to explore the relationship between HR and SO over the years, we created different bins based on what era of baseball a player played in.

Based on this article from The Sport Journal (<https://thesportjournal.org/article/examining-perceptions-of-baseballs-eras/>), the different eras of baseball can be broken down generally as follows:

Era1 Pre Dead Ball Era (1870-1900)

Era2 Dead Ball Era (1901-1919)

Era3 Live Ball Era (1920-1941)

Era4 Integration Era (1942-1960)

Era5 Expansion Era (1961-1976)

Era6 Free Agency Era (1977-1993)

Era7 Long Ball/Steroid Era (1994-2005)

Era8 Post Steroid Era (2006-2020)

# Analysis or model

If you are conducting an inference test, explain the analysis you performed clearly and include well-labelled diagrams to make your points. If you chose to do a predictive model, explain the model, how you trained and tested it, and how well it works. How did you confirm that the data met the requirements for the test or modelling technique to be valid?

# Conclusions

Did you prove/disprove your hypothesis or create a useful model? What did you learn about your data set?