CA4010 Project - Group 35

'Sabermetrics' by Connell Kelly and Patrick Gildea

Contents

Declaration on Plagiarism

Name(s): Patrick Gildea, Connell Kelly

Programme: Computer Applications (CASE4)

Module Code: CA4010

Assignment Title: Sabermetrics

Submission Date: 22/11/2020

Module Coordinator: Mark Roantree

I/We declare that this material, which I/We now submit for assessment, is entirely my/our own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my/our work.

I/We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying.

I/We have read and understood the Assignment Regulations.

I/We have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the source cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me/us or any other person for assessment on this or any other course of study.

Name(s): Patrick Gildea, Connell Kelly

Date: 22/11/20

Section 1 - Outline

Sabermetrics (SABRmetrics)

The application of statistical analysis to baseball records, especially in order to evaluate and compare the performance of individual players.

In this project, our goal was to use sabermetric analysis to evaluate classifications and predictions associated with Major League Baseball (MLB) players between the years 2016 and 2019. Our classification experiment involves accurately defining the different players based on their performance during the 2016-2018 MLB seasons. In turn, we will use this and further relevant data to estimate their performance in the following 2019 season.

We plan on using our analysis to classify players and predict potential MLB award winners and nominees for the 2019 season with the appropriate real-world results to compare them to. We plan to utilise a number of tried and tested sabermetric algorithms to analyse a player's offensive output such as home runs and Slugging Percentage (SLG%). The quality of play is different for each season, so relevant batting trends will need to be considered for improved accuracy, for example, the average Slugging Percentage in 2016 was different to that in 2017.

Data scientists are paid thousands every year to predict hundreds of facets of Major League Baseball, so to prevent potential scope creep, we've narrowed our goals down to a few achievable results with ample room for experimentation. In our 'Sabermetrics' project, we intend to get reasonably accurate predictions on potential home run totals, slugging percentages and award winners for the 2019 MLB seasons.

1.1 - Dataset

Our dataset comprise that of all relevant batting information recorded during the 2016-2018 seasons, courtesy of the Lahman Database. It is maintained by Sean Lahman, Sean Forman and Ted Turocy and has been operational since 1994. It was created with the express intention of making baseball statistics more freely available to the general public.

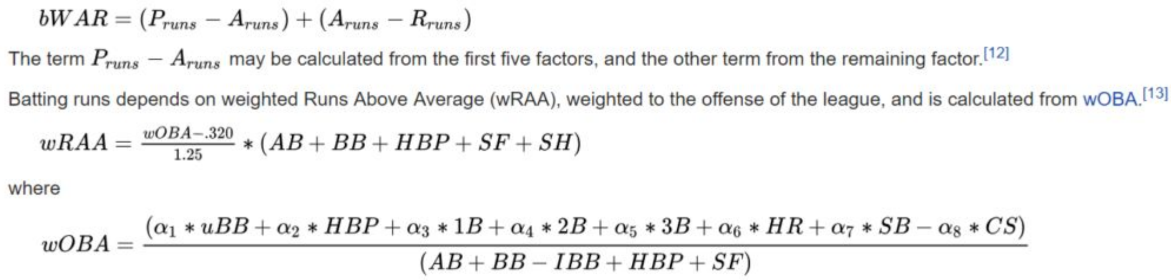
In regards to the data, players who are below a certain threshold of playtime will not be considered due to their inherent lack of longstanding and accurate data. These limitations will also prevent fluky outliers from offsetting any of our results.

If there are aspects of baseball's rule set that are unclear, please watch this three minute video on how the sport functions: **The Rules of Baseball - EXPLAINED!**

Section 2 - Analysis

2.1 Introduction

Initially we had planned to develop a much larger dataset containing many different sabermetrically determined values for each player in order to project a variety of possible elements of the 2019 season, for example, using WAR (Wins Above Replacement) to determine the exact amount of wins a single player contributed to their team for during a season. Statistics and algorithms like these, while well suited to projection, proved to be highly complex, weighted against league averages and prone to being offset by positive or negative outliers.



$$bWAR = (P_{runs} - A_{runs}) + (A_{runs} - R_{runs})$$

The term $P_{runs} - A_{runs}$ may be calculated from the first five factors, and the other term from the remaining factor.[12]

Batting runs depends on weighted Runs Above Average (wRAA), weighted to the offense of the league, and is calculated from wOBA.[13]

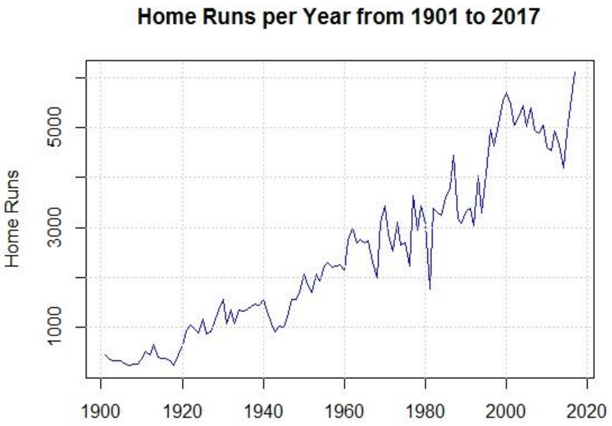
$$wRAA = \frac{wOBA-.320}{1.25} * (AB + BB + HBP + SF + SH)$$

where

$$wOBA = \frac{(\alpha_1 * uBB + \alpha_2 * HBP + \alpha_3 * 1B + \alpha_4 * 2B + \alpha_5 * 3B + \alpha_6 * HR + \alpha_7 * SB - \alpha_8 * CS)}{(AB + BB - IBB + HBP + SF)}$$

bWAR, wRAA and wOBA, several sabermetric algorithms we considered.
'b' meaning Baseball Reference (source of data) and w meaning 'weighted'.

To avoid scope creep, we decided to work with more compact and comprehensible statistics that we could apply sabermetric sensibilities to, shift our focus to more tried and true statistics such as home runs and slugging percentages.



To better prepare the data, we will need to consider several aspects regarding the state of baseball up-to and during the period of our analysis and how it will support our projections going forward. Given that our project is centered around home run hitting, it's important to note that the amount of home runs seen per year has been steadily rising ever since the inception of Major League Baseball. There have been occasional dips in production, for example, during World War 2, but as time passed, the sport modernised and in turn, batters were taught improved ways to hit better and harder.

Rampant steroid use caused a spike in the 1990's which then dipped after a crackdown in 2004. However, the 2010's have seen home run production climb to the highest it's been in history thanks to advanced training, player specialisation and increased pitching velocity. Our data will need to reflect the recent surge and recent years in a player's career will be weighted accordingly due to their recency and how it could be affected by new home run production standards.

2.2 Data Cleaning

In order to accrue enough quality data with which we can make accurate and informed projections, the cleansing of incomplete data is necessary. While there are well over a thousand individual position players who played Major League Baseball between the 2015 and 2018 seasons, only roughly a quarter of these players played enough to provide valid projection information.

The players we intend to work with must be separate from the following:

- Players who have only played one or two seasons will also be identified and removed from our dataset. Our projection requires three years of consistent data and any missing seasons will hamper its accuracy.
- Starters who were injured early in a season and haven't contributed enough data, making them ineligible for our projection.
- Players who were traded to another team midway through one of the three seasons. The abruptness of a trade can often throw a player off their game and lead to them underperforming for the rest of the season as they adapt.
- Utility players who haven't had enough at bats to be eligible for our analysis.

NOTE: Minimum At Bats

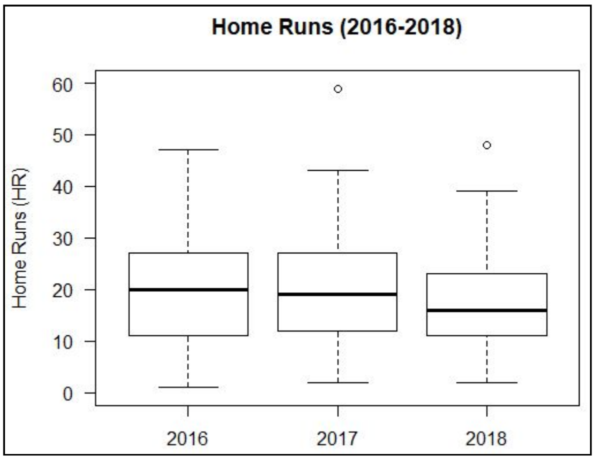
In Major League Baseball, an award named the Batting Title is given away at the end of the season to the player with the highest batting average (ie. most successful balls hit into play). Players need a minimum of 502 at bats (AB) in a season to qualify for it's batting title. MLB starters average at 550 AB per season, so for the sake of providing more data and to accommodate for consistent platoon players (ie. players play half a season), the minimum AB needed to qualify for our analysis is 275. This will also help exclude pitchers from our analysis. Pitchers are offensive outliers who specialise only in defensive play, but are often forced to bat when their turn arrives. This means that their offensive output always correlates very negatively and could offset otherwise reliable data from regular position players (batters).

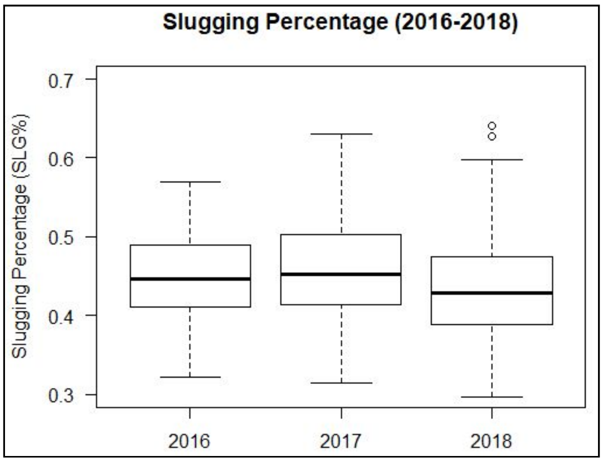
What remains is three datasets comprising **147 players each** over **three years** for a total of **441 rows of data.**

In order to determine players we intend to analyse and make projections for, we will be constructing our own dataset containing the relevant information. Our dataset will be based on information gathered from The Lahmen Database Archive. Along with using AB to qualify players we'll also be using the algorithm Slugging Percentage (SLG%) to determine the consistency of a player which we can use to bolster our home run projections. This will be detailed further in Section 3.

2.3 Data Dispersal

Analysing the dispersion of data regarding our datasets will allow us to better anticipate and validate the results of our projections, as well as paint a picture of the state of offense in baseball. Home runs and slugging percentages will be analysed. By plotting the home run and SLG% data from each season onto a series of graphs, we're able to make a number of observations regarding home run production.

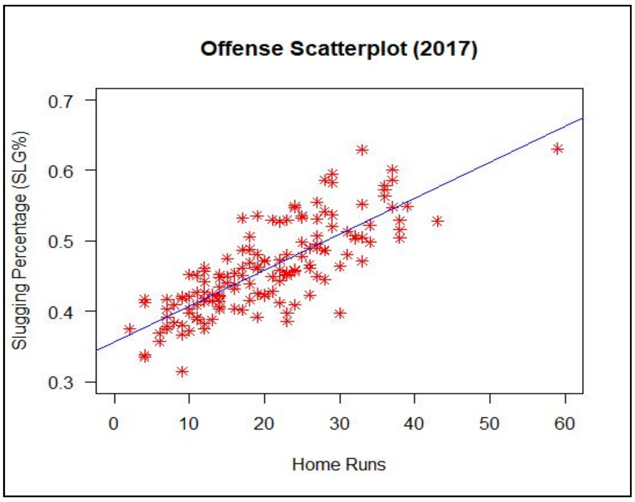




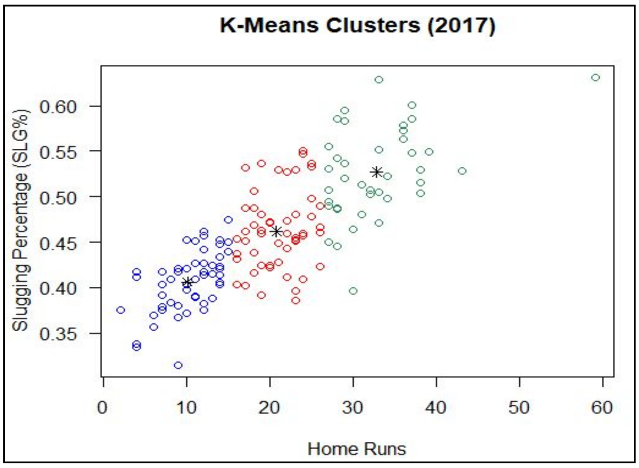
2017 was a record breaking season for total home runs scored and this correlates well with the positive skews found on its boxplot. SLG% was also higher on average in 2017 than the other seasons which supports the HR result. Let's look at 2017.

$$SLG\% = \frac{TB}{AB} = \frac{1B + (2 \times 2B) + (3 \times 3B) + (4 \times HR)}{AB}$$

Slugging percentage (SLG%), a numerical evaluation of a player's offensive output.



Given that home runs are weighted heavily in a player's SLG%, it's no surprise that, with **Pearson's Correlation**, the linear associations between the two numeric variables is fairly strong **(0.7928044)**. From this, we can assert that players we project to have the most home runs are likely to have higher SLG%'s and vice versa.



We ran several K-Means clustering algorithms using RStudio on all our datasets and successfully identified 3 clusters in each. **Cluster 1 (Blue)** clustered most players who performed below average this season and stood no chance of receiving any awards or distinctions. **Cluster 2 (Red)** clustered average and above average players who performed well and could be considered for several awards. **Cluster 3 (Green)** top performing players who led in SLG% and home runs and would be clear choices for awards and distinctions.

2017 and 2018 proved to be highly similar and with 2016, will prove to be a useful point of comparison for what we produce with our projection algorithm.

Section 3 - Algorithm

$$\frac{\sum w \cdot x}{\sum w}$$

For our project, we decided to implement a weighted mean algorithm called **miniMarcel**. It's an algorithm we adapted from the Marcel Baseball Projection Algorithm by Tom Tango. We rewrote it to better fit the datasets with which we were working with.

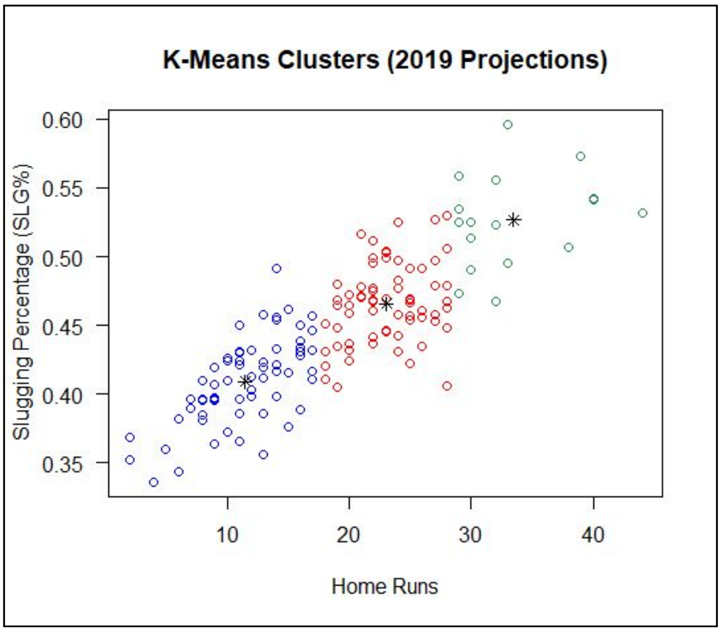
miniMarcel takes a single player's home runs over three years and multiplies them by descending series of weights that account for age and experience starting at their most recent year.

Example:
Nelson Cruz hit 43 home runs in 2016, 39 in 2017 and 37 in 2018.
Projected 2019 HR = ((43 * 5) + (39 *4) + (37 * 3)) / 12 = <u>40</u>
Real 2019 HR = <u>41</u> (Individual confidence of 0.975)

Example:
Nolan Arenado slugged .570 in 2016, .586 in 2017 and .561 in 2018.
Projected 2019 HR = ((.570 * 5) + (.586 *4) + (.561 * 3)) / 12 = <u>0.572</u>
Real 2019 HR = <u>.583</u> (Individual confidence of 0.981)

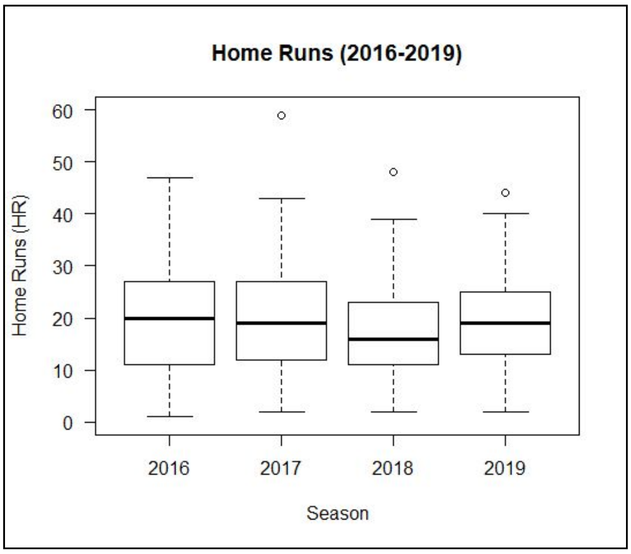
We developed a Python program that would take every player offensive numbers, calculate their slugging percentages and then use this our miniMarcel formula to calculate their potential home run total and slugging percentage in 2019.

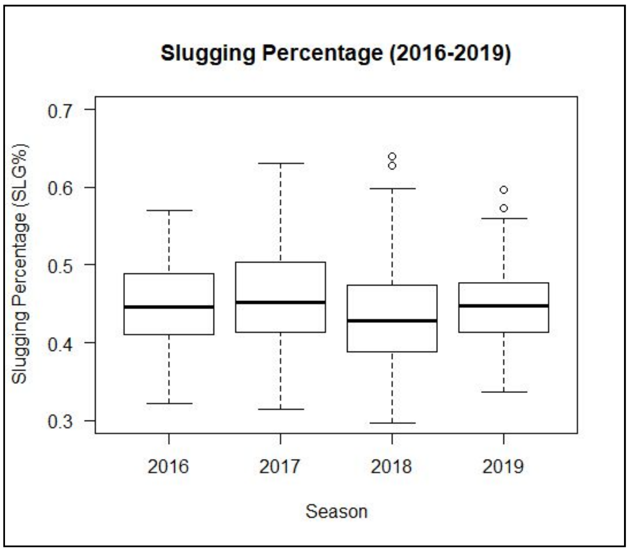
With the help of RStudio, we calculated several iterations of K-Means clusters for the 2019 season the same way we did for the 2016-2018 seasons to determine the three clusters of below average, average and above average players.



The results showed similar trends to the other seasons, correlating high slugging percentages with home run production. With this information we could make estimations about statistic leaders for the 2019 season and players with a chance of winning awards or distinction. We will compare these results from our projection algorithm to their real-life equivalents in the following section below.

Section 4 - Results

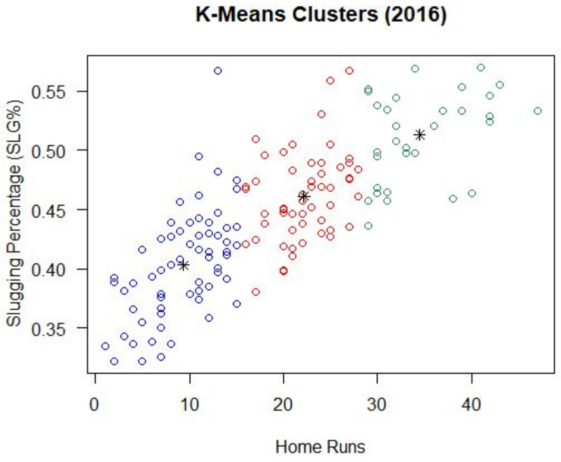


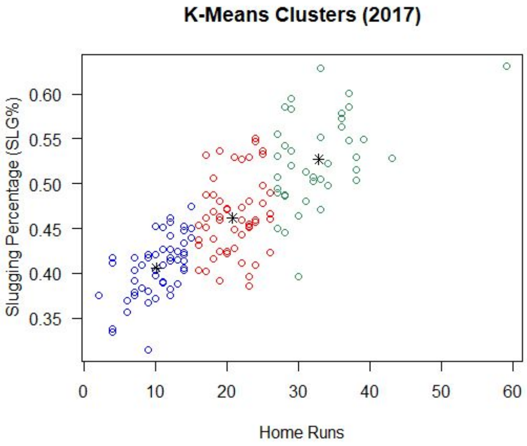


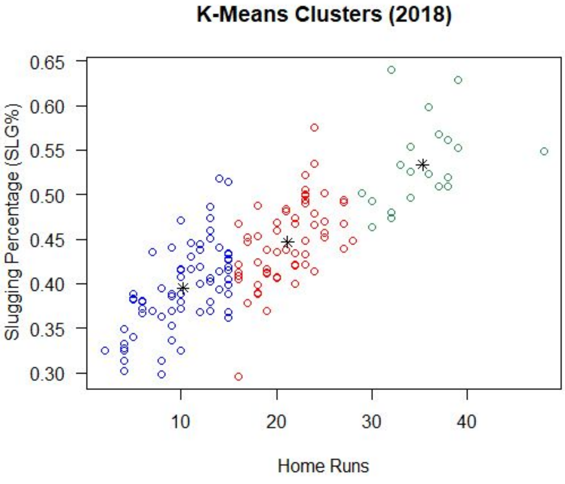
The results from our projections were just as we had hoped.

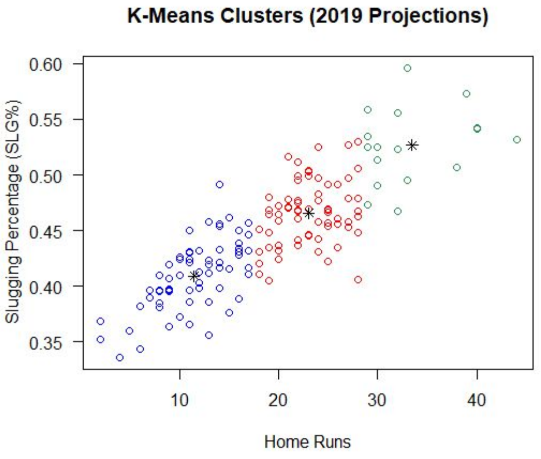
Notes:

- 2019 appeared to follow suit with the 2017 season and re emerge from the offensive slump in 2018 and average out with a mean similar to 2017 and 2016 seasons.
- In regards to home runs, 2019 did not skew as positively as 2017 and slugging percentages skewed negatively in comparison to 2017 as well.
- Results were more concentrated than any of the other years in both categories with less extreme max and minimum values.









As seen above, the clusters accrued in all four of our K-Means Cluster graphs move towards defining the same kinds of players in the same kinds of clusters, 2019 included. Three clusters was the ideal amount for defining players who fall under average, are average and rise above average and our 2019 projections play to this format as well to project the highest performing players offensively in 2019.

Below we will arbitrarily select 5 players from our dataset and determine the accuracy of their 2019 projections based on our miniMarcel algorithm.

Carlos Correa
Projected Home Runs: **20**
Projected Slugging Percentage: **0.472**
Real Home Runs: **21**
Real Slugging Percentage: **.568**

Mike Trout
Projected Home Runs: **33**
Projected Slugging Percentage: **0.596**
Real Home Runs: **45**
Real Slugging Percentage: **0.645**

Kolten Wong
Projected Home Runs: **6**
Projected Slugging Percentage: **0.382**
Real Home Runs: **11**
Real Slugging Percentage: **0.423**

Eduardo Núñez
Projected Home Runs: **9**
Projected Slugging Percentage: **0.419**
Real Home Runs: **2**
Real Slugging Percentage: **0.305**

Jose Altuve
Projected Home Runs: **21**
Projected Slugging Percentage: **0.516**
Real Home Runs: **31**
Real Slugging Percentage: **0.550**

After calculating the confidence of projected home runs compared to the real home runs for each player we received **0.80.** After doing the same for the slugging percentages, we received **0.95.** While our slugging percentages were close and our home runs passed the typical minimum confidence threshold of 0.80, the simplistic nature of our projection algorithm holds it back from closer accuracy with home runs.

NOTE: MLB Leagues

MLB players are split between a pair of 15 team leagues known as the American League (AL) and the National League (NL). They're in place to organise teams into regional sub-divisions and to separate teams which play in the same city, for example the New York Yankees and the New York Mets. For the interest of this project, each league determines player standings, awards and distinctions to their specific league where they play most of their games, for example there'll always be 2 of each award given to the appropriate player in each league

In regards to predicting awards, our top players went as follows:

MLB HR Leaders

1. Khris Davis
2. Nelson Cruz
3. Giancarlo Stanton

MLB SLG Leaders

1. Mike Trout
2. Nolan Arenado
3. Freddie Freeman

Of these top players our algorithm projected, the following would go on to win:

- Nelson Cruz (Silver Slugger Award, ie. best hitter in his position)
- Mike Trout (MVP, Silver Slugger, Aaron Award, and selected to be an All-Star)
- Freddie Freeman (MVP Candidate, Silver Slugger and All-Star)

Not every top player we predicted would go on to win an award, but the subjective distinctions like these are much more difficult to predict than a player's potential statistics. However ranking an MVP candidate so high was an accomplishment as it is considered the most difficult to predict. The 0.50 confidence score here proves that our algorithm has room to improve. It's simplicity, while effective at predicting the statistics of players performing consistently, is unable to properly take into account sudden improvements and breakout seasons.

Overall, we are pleased with our results as our projections proved to be close to accurate more often than not and miniMarcel has proven to be a compact and effective projection algorithm for quick player predictions. We will be researching more ways of adding complexity for miniMarcel such as weighting it against league averages, taking different ballpark layouts into effect and expanding to defensive projections as well. While the pool of players we were analysing was small in comparison to some datasets, the data contained for each was rich and well suited to the algorithms we implemented in Python and R. We have attached a dataset containing the 2016-2018 seasons including the 2019 home run and SLG% projections with each row matching accordingly. Thank you for your consideration.

Section 4 Glossary

Plate Appearances (PA) *A batter's turn batting against a pitcher.*

At Bats (AB) *A batter's turn batting which results in a fielder's choice, hit or an error or when a batter is put out on a non-sacrifice.*

Runs Scored (R) *A point assigned to a given player's team when they have advanced around first, second and third base and safely returned to home plate.*

Base Hits (H) *An at-bat that results in the baseball being struck into fair territory and batter reaching base without doing so via an error or a fielder's choice*

Extra Base Hits (2B/3B/HR) *A hit where the batter reached a base beyond that of first, i.e. doubles, triples and home runs.*

Home Runs (HR) *A hit that makes the ball inaccessible to the defending team and allows the batter to make a complete circuit of the bases and score a run.*

Strikeout (SO) *The result of an AB ending in failure for the batter as a result of them receiving three strikes from the opposing pitcher, contributing to an out.*

Base on Balls/Walks (BB) *Four pitches thrown out of the strike zone, none of which are swung at by the batter, allowing them to advance to first base.*

Sacrifice Hits (SH) *The act of a batter grounding out, before there are two outs, in a manner that allows a baserunner to advance to another base.*

Sacrifice Flies (SF) *A ball hit into the air with less than two outs that, when caught, allows a base runner the opportunity to score a run before being put out.*

Times Hit by Pitch (HBP) *An event in which a batter is struck directly by a pitch from the pitcher, allowing them to immediately advance to first base.*

Total Bases (TB) *Total bases refer to the number of bases gained by a batter.*