# A Bayesian Approach to Baseball

By: Tucker Saland

# Introduction

- The aim of this project is two-fold: (1) I want to project the aging curve of MLB players over time, and (2) I want to determine whether there are certain playstyles and characteristics that lend themselves to player longevity.

- More precisely, I want to project players' WAR (Wins Above Replacement) values over time to construct probability distributions for player performance rather than the point estimate predictions that are currently in use.

| CALIBER OF PLAYER | WINS ABOVE REPLACEMENT |
|---|---|
| BENCH GUY | 0-1 WAR |
| ROLE PLAYER | 1-2 WAR |
| SOLID STARTER | 2-3 WAR |
| ABOVE-AVERAGE | 3-4 WAR |
| ALL-STAR | 4-5 WAR |
| SUPERSTAR | 5-6 WAR |
| MVP | 6+ WAR |

# Hypothetical Example – Josh Donaldson Trade

Suppose you are the Yankees' GM in February 2022. You get a call from the Twins, and they are ready to part with Josh Donaldson with two years left on his contract in exchange for three prospects.

The consensus is that over the lifetime of the trade, you would expect to make off slightly better in the long run with the prospects than you would with two years of Donaldson. But at the same time, your team is built to win now, and a 3.0 WAR 2022 season from Donaldson should bring you from a projected wildcard team to legitimate World Series contenders.

Donaldson would be going into his age 36 season in 2022, and in the previous season, he had 3.0 WAR. Do you take the trade?

Expected WAR = $(1 + \mu) \times$ Previous Year WAR = 2.60

Probability of WAR > 3.0 = $1 - \Phi(-\mu/\sigma)$ = 39.78% → In other words, probability that percent change is greater than 0

# Data

1. A list of all baseball games from 2016-2022 with aggregate box scores, location, odds, and team record scraped from ESPN.

| Game | away | away-record | awayaway-record | home | home-record | homehome-record |
|---|---|---|---|---|---|---|
| <dbl> | <chr> | <chr> | <chr> | <chr> | <chr> | <chr> |
| 360403123 | STL | 0-1 | 0-1 Away | PIT | 1-0 | 1-0 Home |
| 360403130 | TOR | 1-0 | 1-0 Away | TB | 0-1 | 0-1 Home |
| 360403107 | NYM | 0-1 | 0-1 Away | KC | 1-0 | 1-0 Home |
| 360404108 | SF | 1-0 | 1-0 Away | MIL | 0-1 | 0-1 Home |
| 360404101 | MIN | 0-1 | 0-1 Away | BAL | 1-0 | 1-0 Home |
| 360404113 | SEA | 0-1 | 0-1 Away | TEX | 1-0 | 1-0 Home |

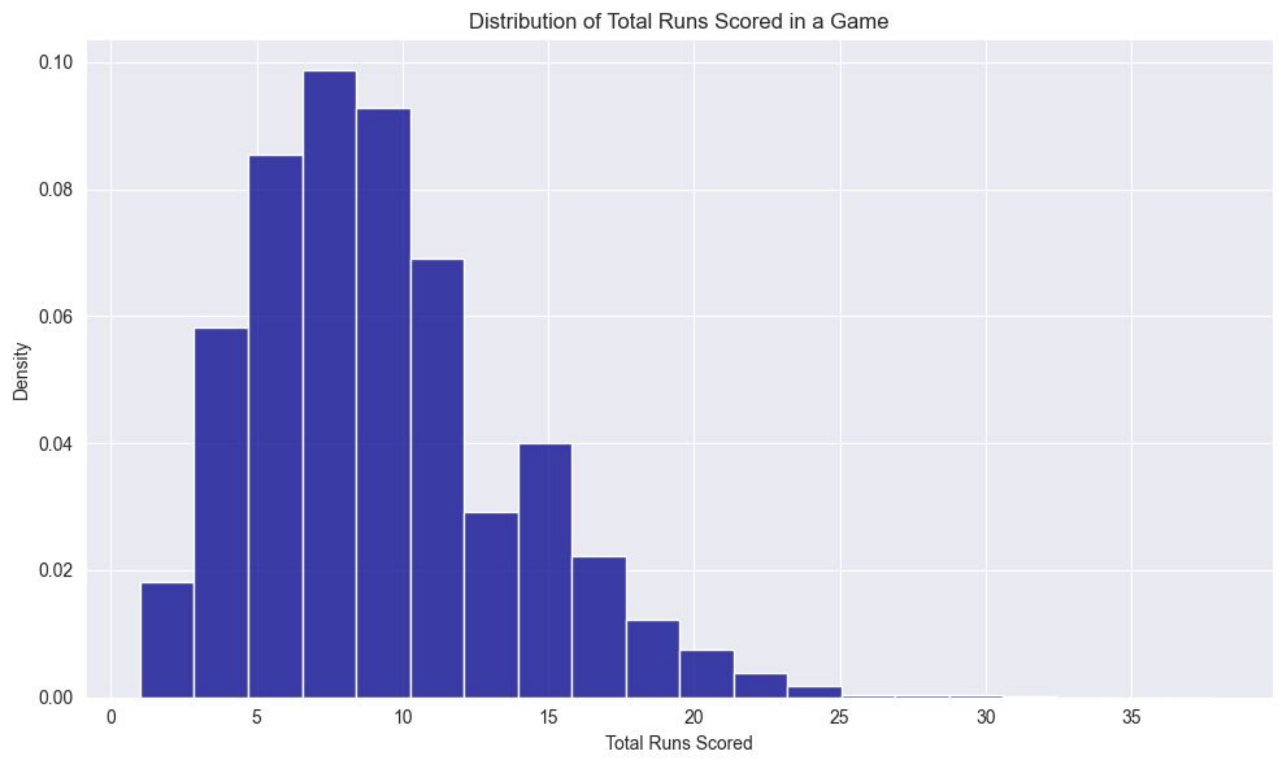
2. Hitter-specific game logs from 2016-2022 from Fangraphs scraped using the baseballr package. This dataset includes traditional hitting metrics, advanced hitter statistics (e.g., exit velocity, launch angle, wRC+), and WAR (Wins Above Replacement).

| PlayerName | playerid | Date | Te... | Opp | season | Age | BatOrder | Pos | G | AB | PA | H |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| <chr> | <int> | <chr> | <chr> | <chr> | <int> | <int> | <chr> | <chr> | <dbl> | <dbl> | <dbl> | <dbl> |
| Bryce Harper | 11579 | 2022-10-04 | PHI | @H... | 2022 | 29 | 1 | DH | 1 | 4 | 4 | 1 |
| Bryce Harper | 11579 | 2022-10-03 | PHI | @H... | 2022 | 29 | 3 | DH | 1 | 4 | 4 | 1 |
| Bryce Harper | 11579 | 2022-10-02 | PHI | @... | 2022 | 29 | 3 | DH | 1 | 4 | 4 | 1 |
| Bryce Harper | 11579 | 2022-10-01 | PHI | @... | 2022 | 29 | 3 | DH | 1 | 4 | 5 | 0 |
| Bryce Harper | 11579 | 2022-10-01 | PHI | @... | 2022 | 29 | 3 | DH | 1 | 3 | 4 | 0 |
| Bryce Harper | 11579 | 2022-09-30 | PHI | @... | 2022 | 29 | 3 | DH | 1 | 4 | 5 | 1 |

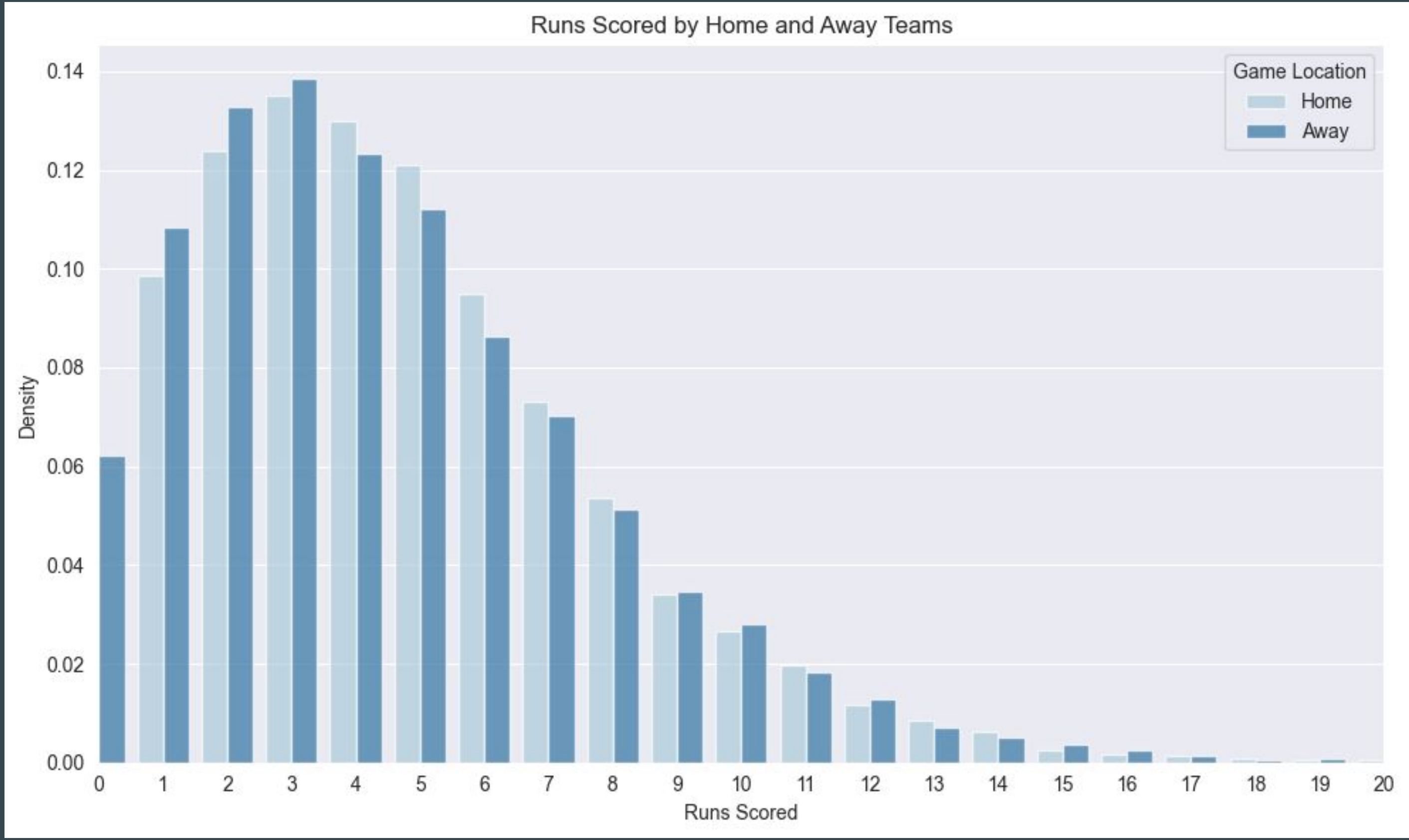
3. Pitcher-specific game logs from 2016-2022 from Fangraphs scraped using the baseballr package. This dataset includes advanced pitching metrics (e.g., FIP, HardHit%, pitch mix) along with traditional statistics.

| PlayerName | playerid | Date | Opp | teamid | season | Te... | HomeA... | Age | W | L | ERA |
|---|---|---|---|---|---|---|---|---|---|---|---|
| <chr> | <int> | <chr> | <chr> | <int> | <int> | <chr> | <chr> | <int> | <dbl> | <dbl> | <dbl> |
| Zack Wheeler | 10310 | 2021-09-28 | @ATL | 26 | 2021 | PHI | A | 31 | 0 | 1 | 2.571429 |
| Zack Wheeler | 10310 | 2021-09-22 | BAL | 26 | 2021 | PHI | H | 31 | 0 | 0 | 1.500000 |
| Zack Wheeler | 10310 | 2021-09-17 | @N... | 26 | 2021 | PHI | A | 31 | 1 | 0 | 1.800000 |
| Zack Wheeler | 10310 | 2021-09-11 | COL | 26 | 2021 | PHI | H | 31 | 1 | 0 | 1.350000 |
| Zack Wheeler | 10310 | 2021-09-06 | @MIL | 26 | 2021 | PHI | A | 31 | 1 | 0 | 0.000000 |
| Zack Wheeler | 10310 | 2021-08-30 | @W... | 26 | 2021 | PHI | A | 31 | 1 | 0 | 6.000000 |

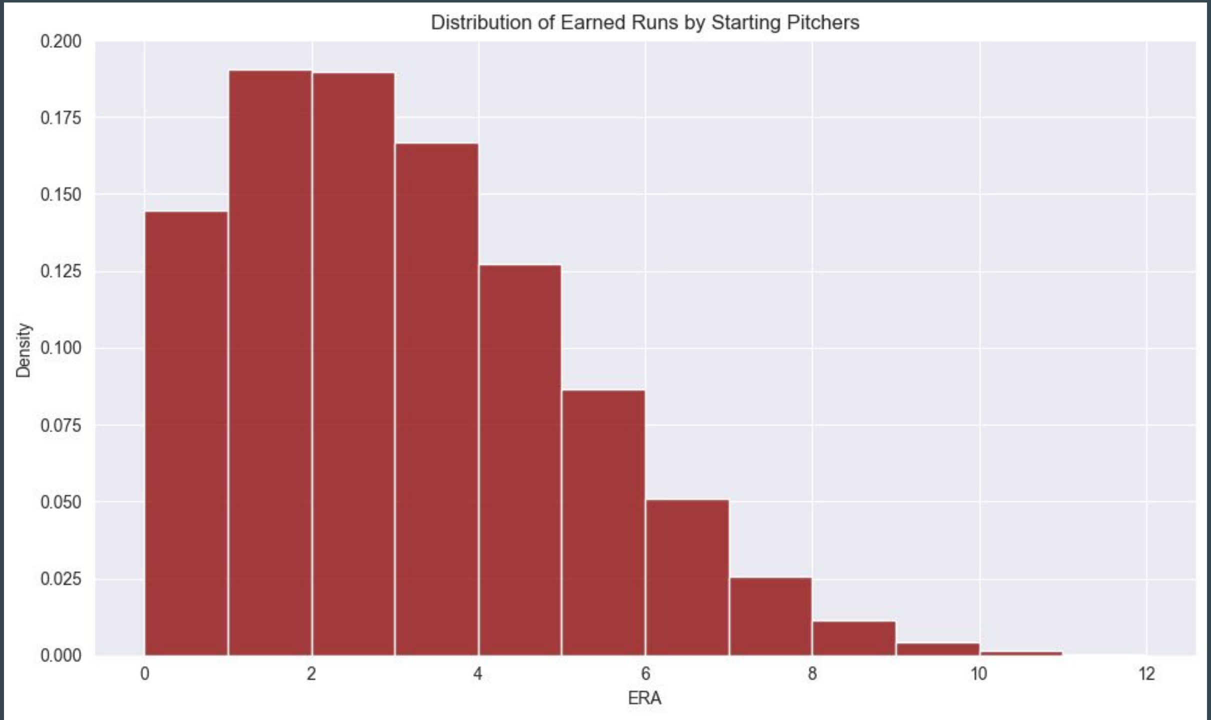
# Run Distribution



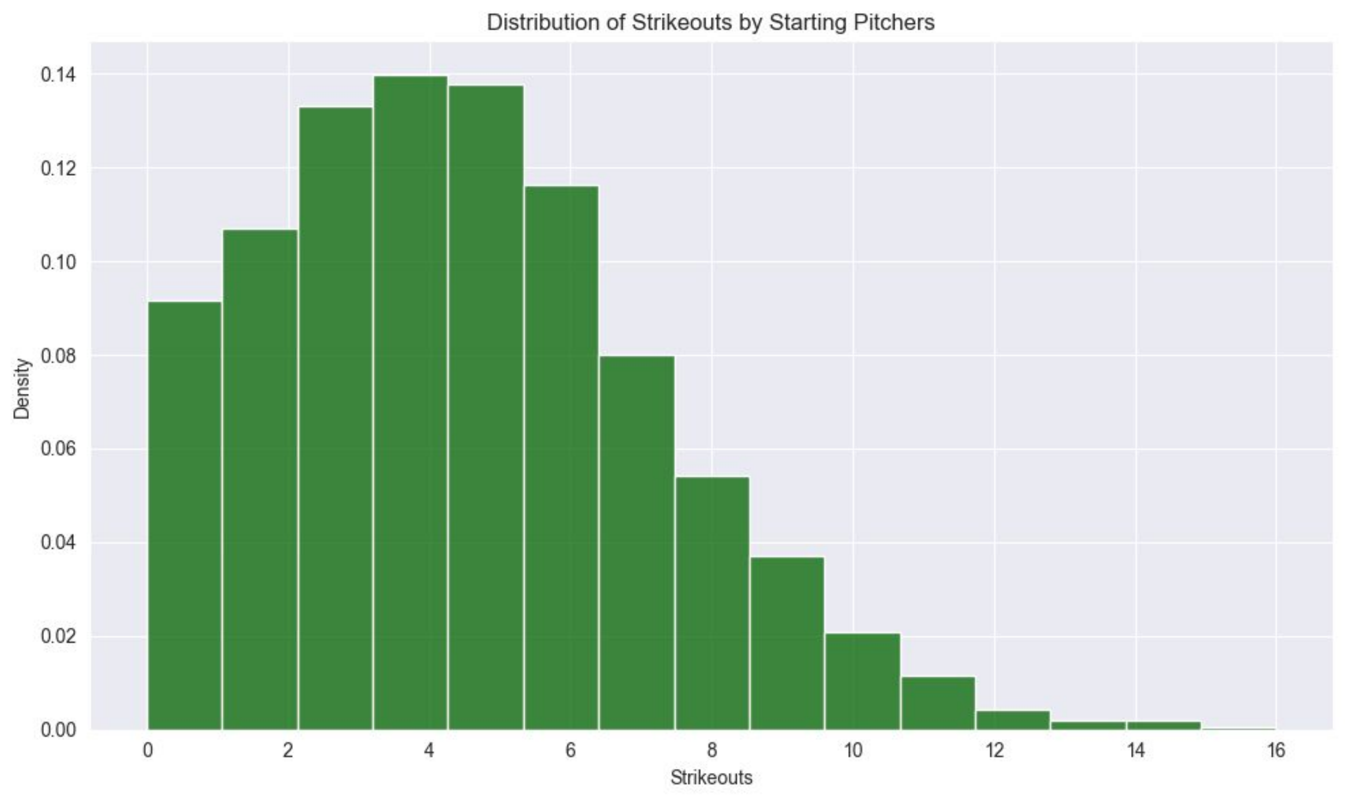
# Home vs. Away Run Distribution



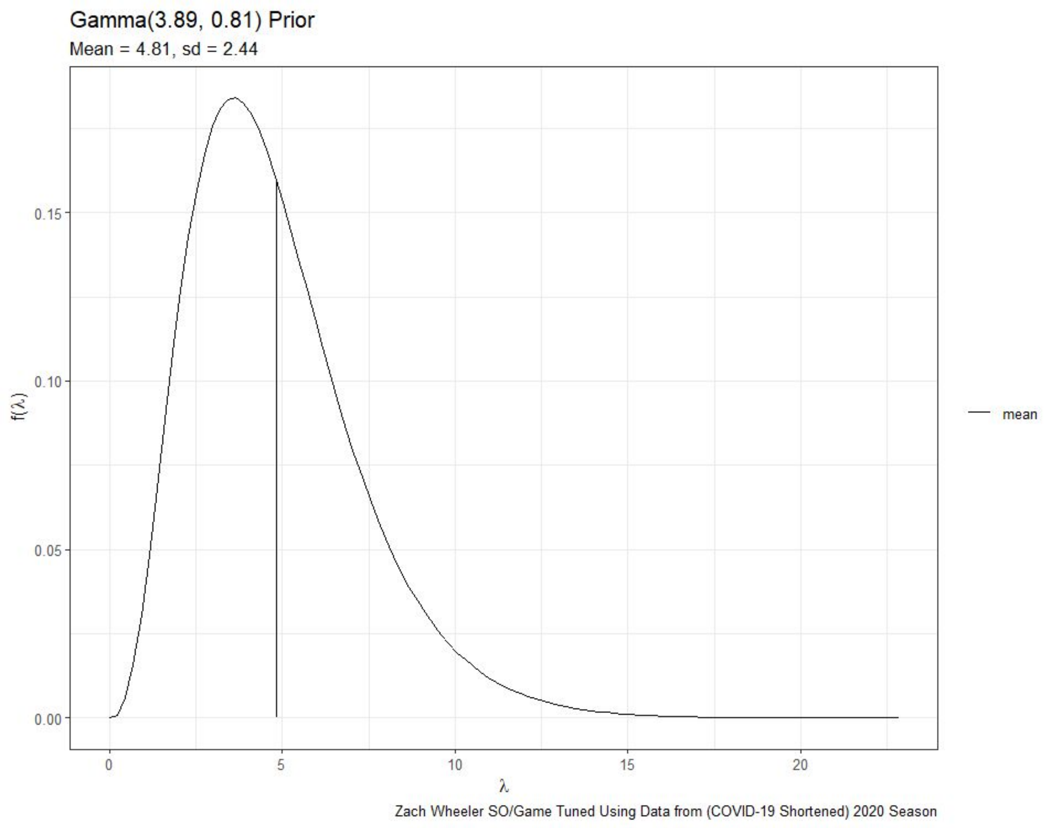
# Earned Runs


Distribution of Earned Runs by Starting Pitchers
Density
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
2
4
6
8
10
12
ERA

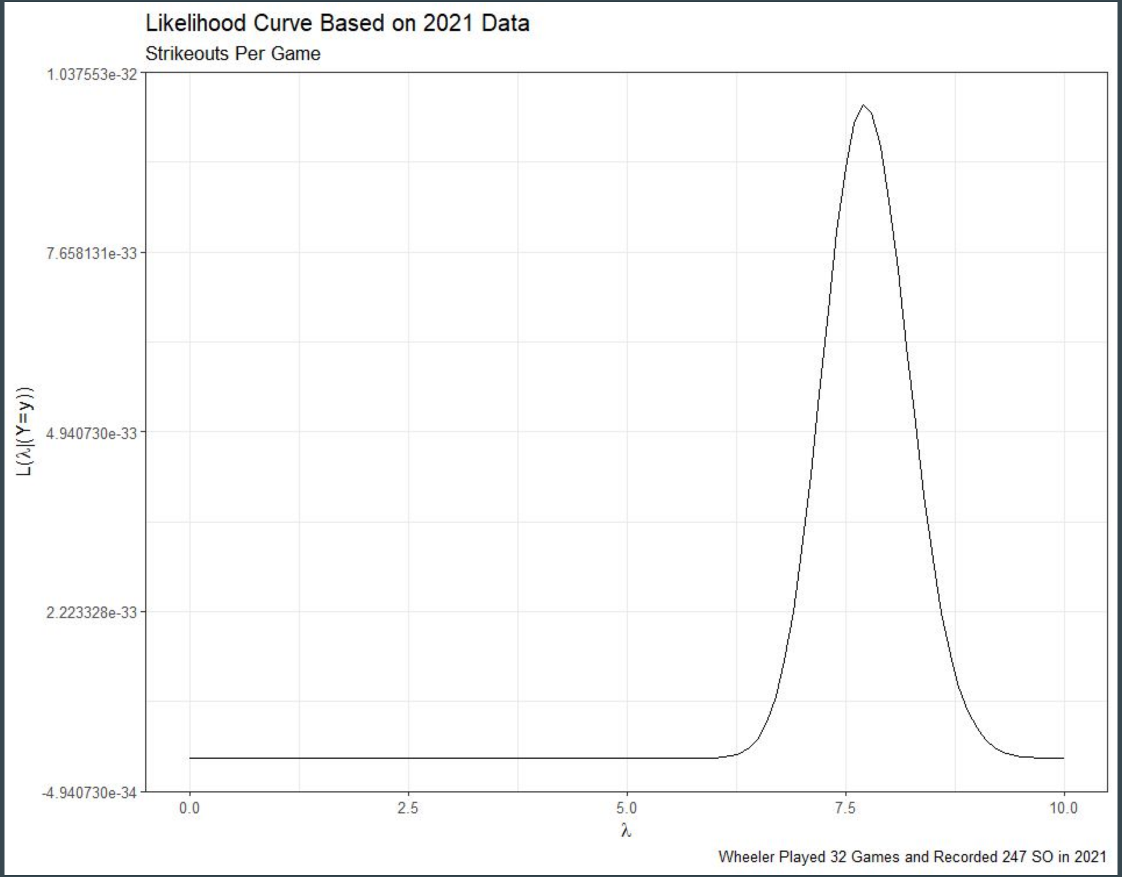
# Strikeouts


Distribution of Strikeouts by Starting Pitchers
Density
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
2
4
6
8
10
12
14
16
Strikeouts

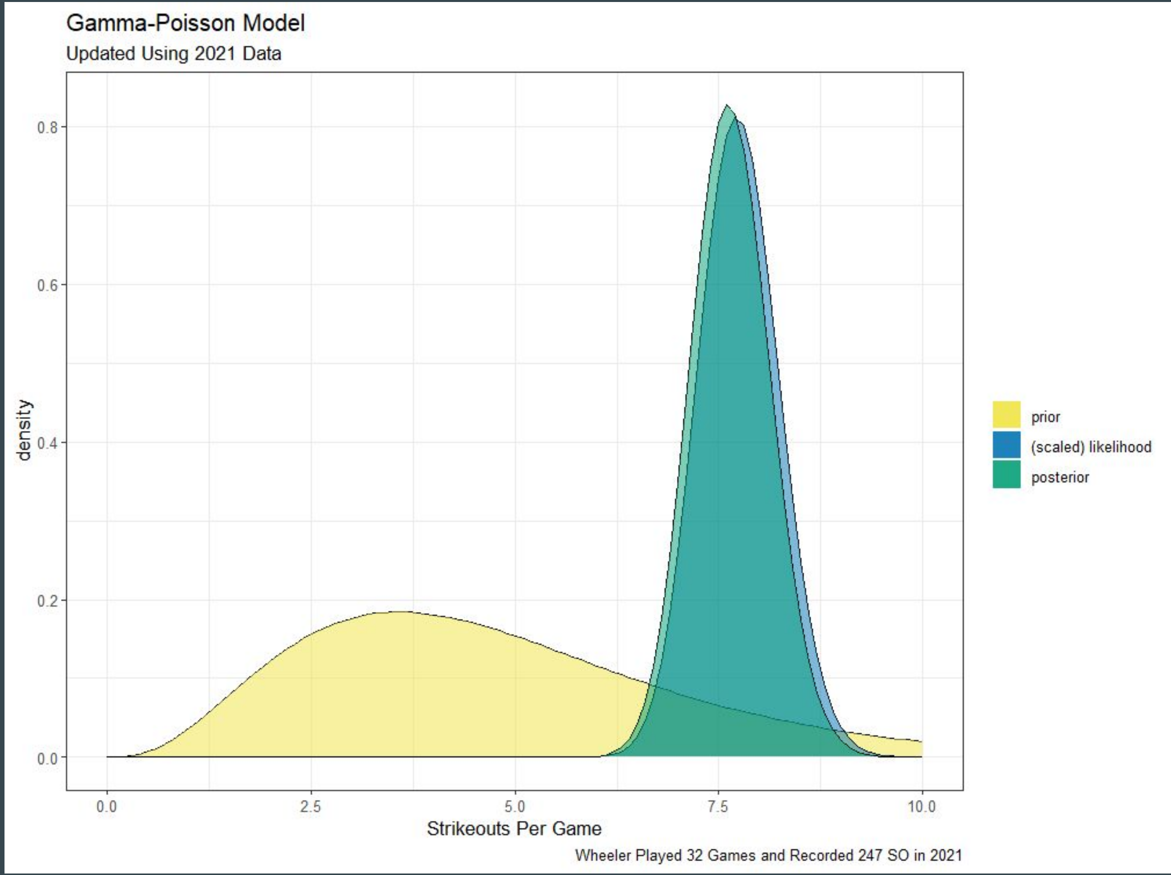
# Zach Wheeler K/Game: A Gamma-Poisson Example



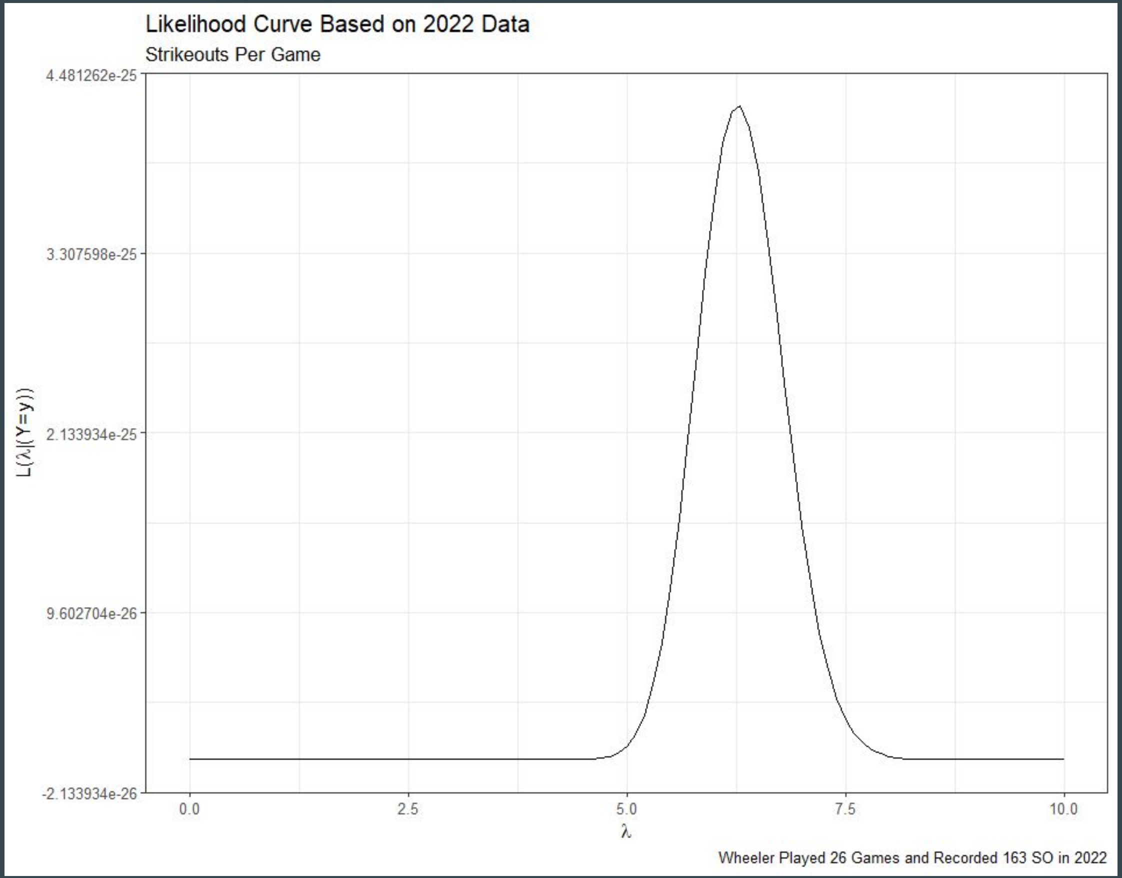
# 2021 Likelihood



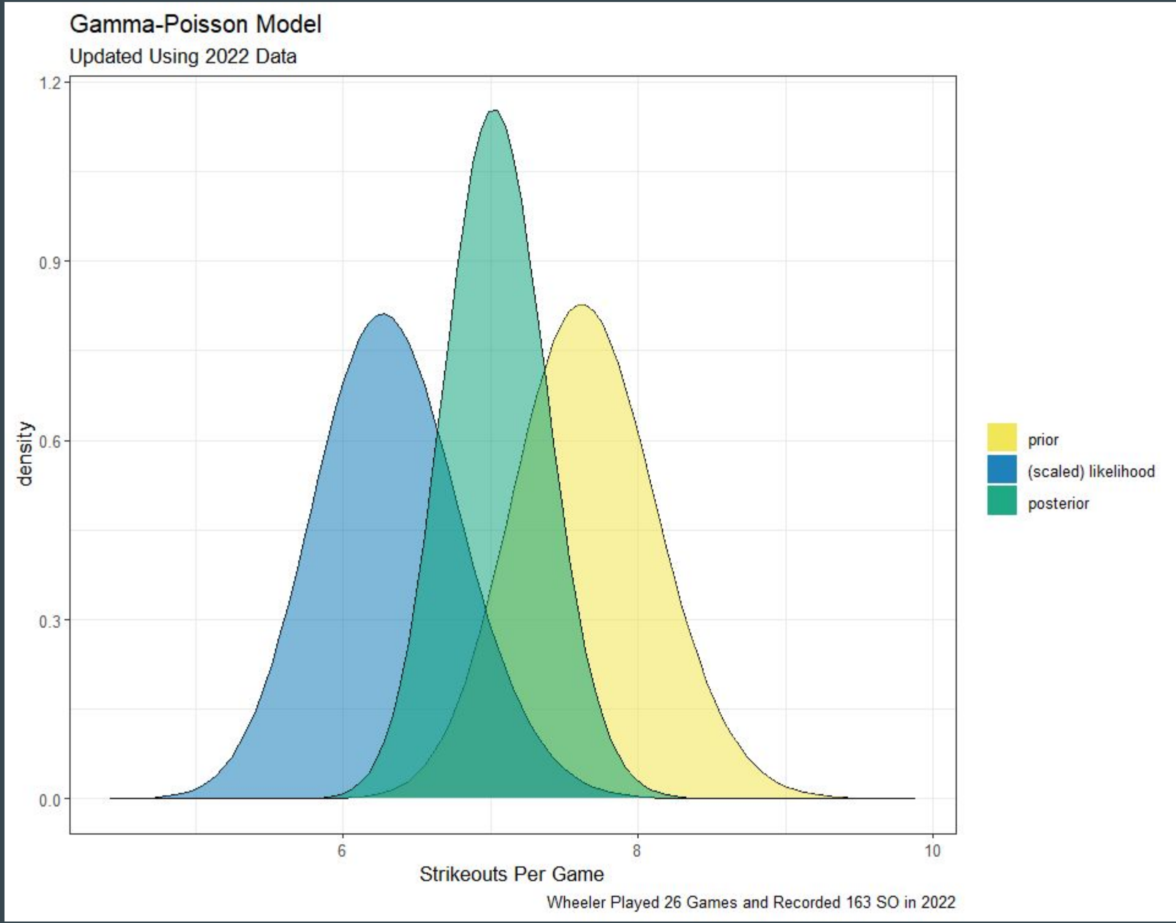
# Update Prior Using 2021 Data



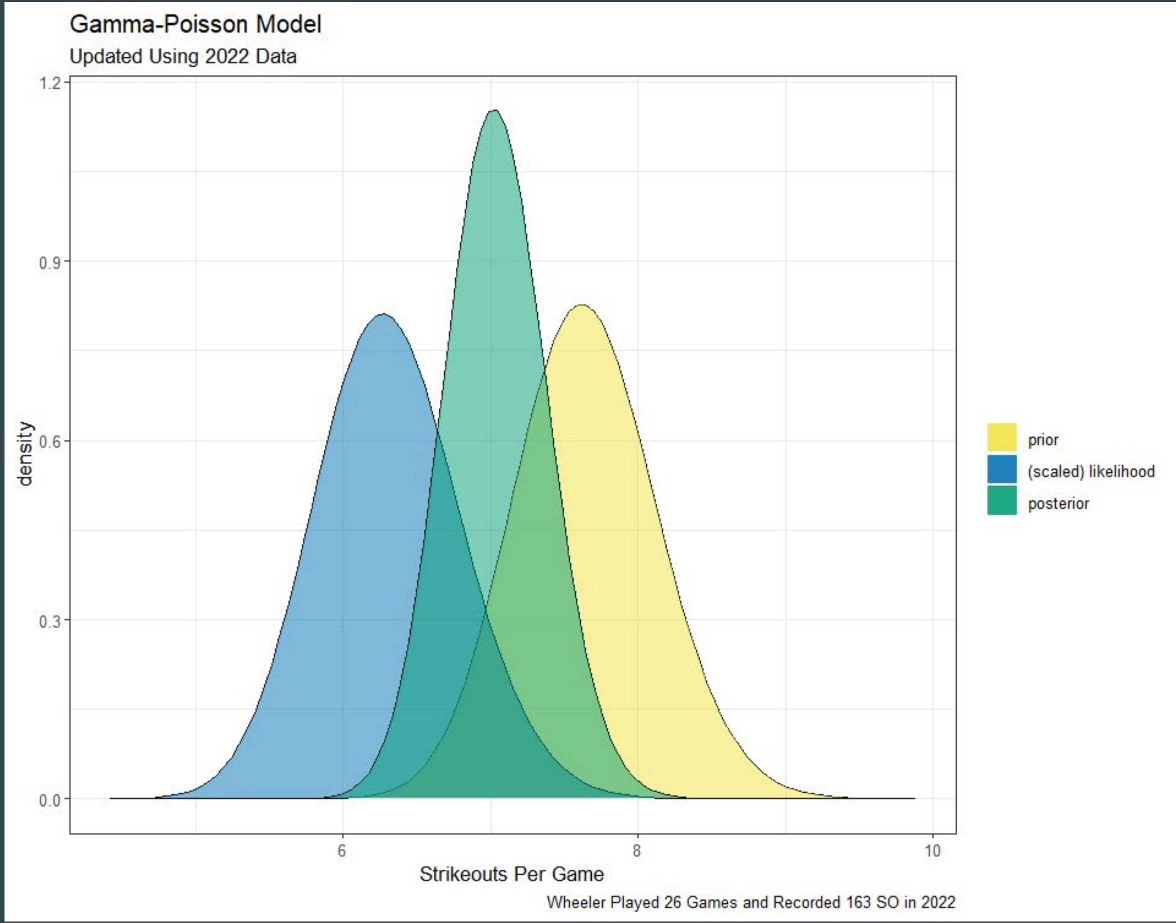
# 2022 Likelihood



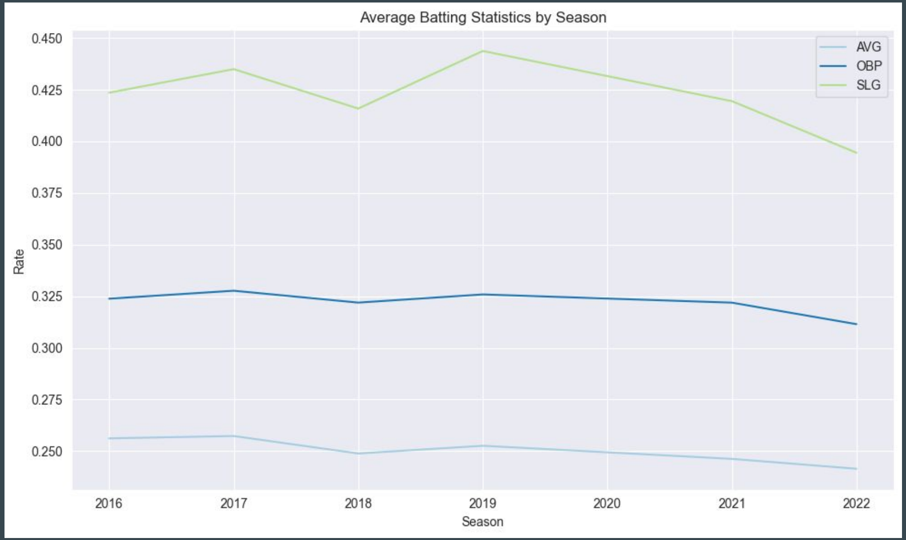
# Update Beliefs Using 2022 Data

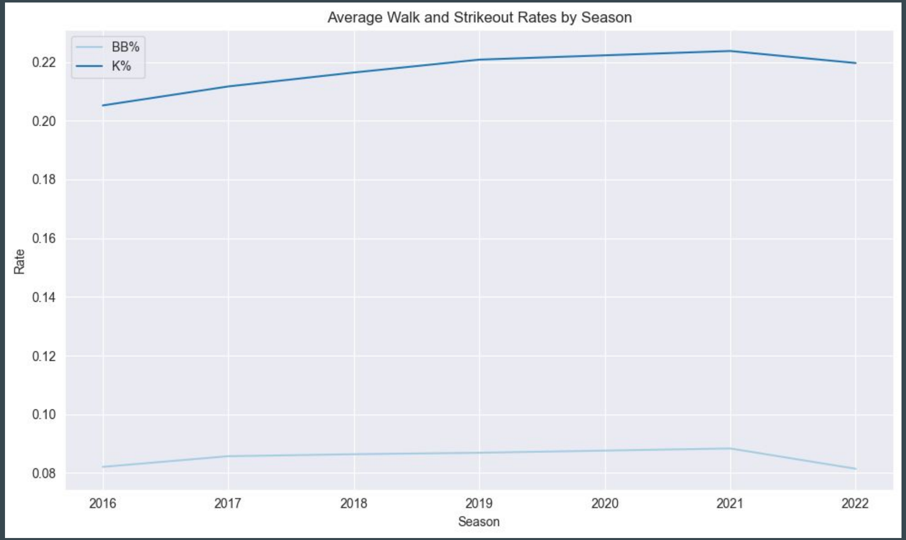


# Update Beliefs Using 2022 Data

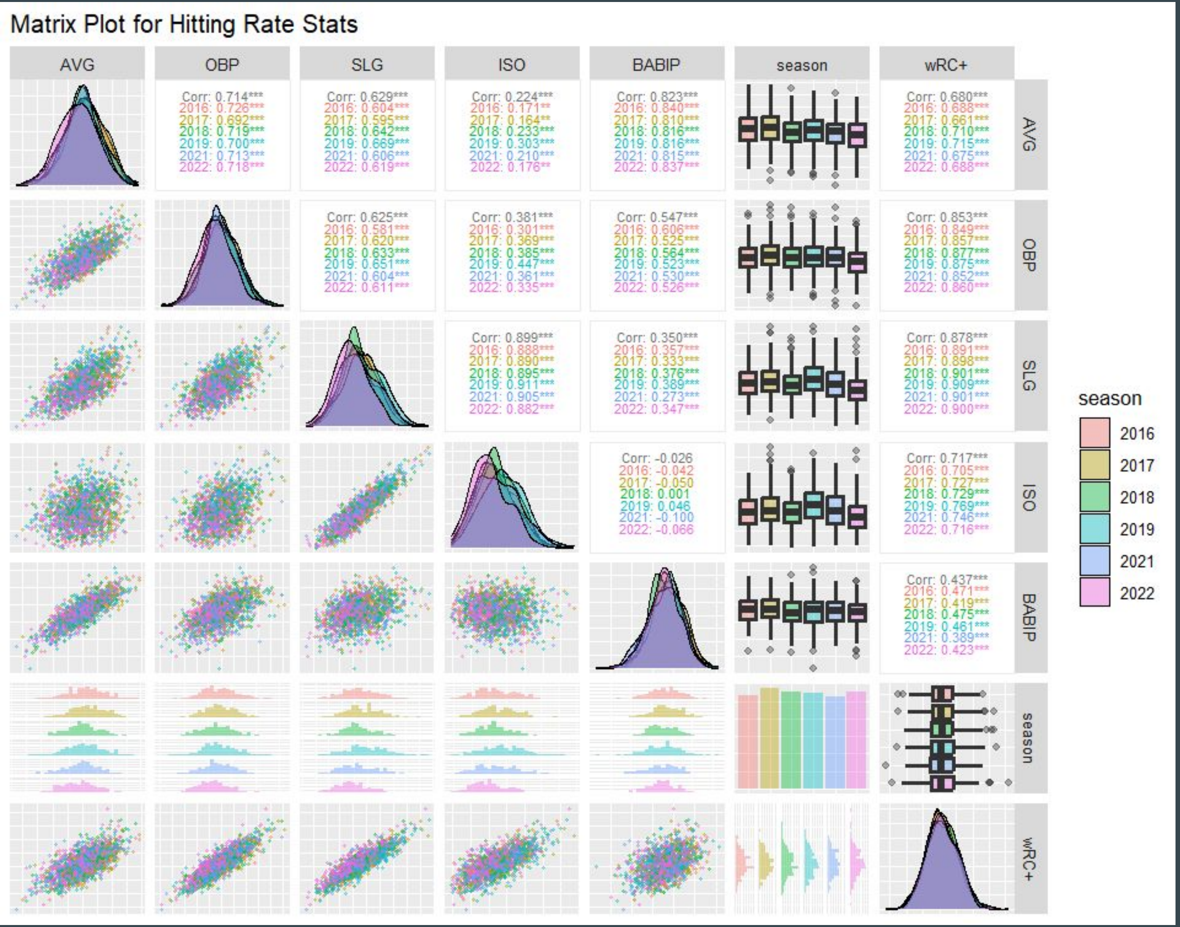


# Batting Trends

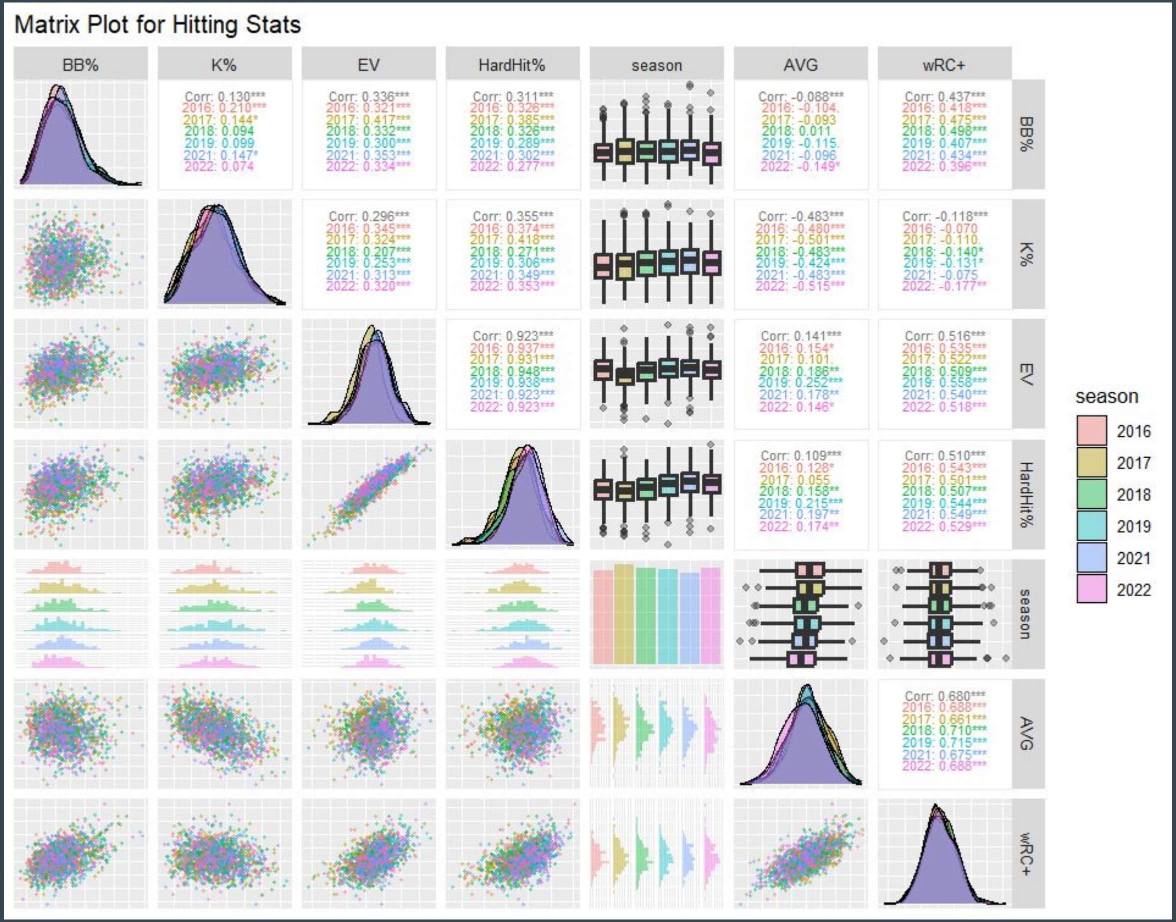


# Hitting Rate Stats

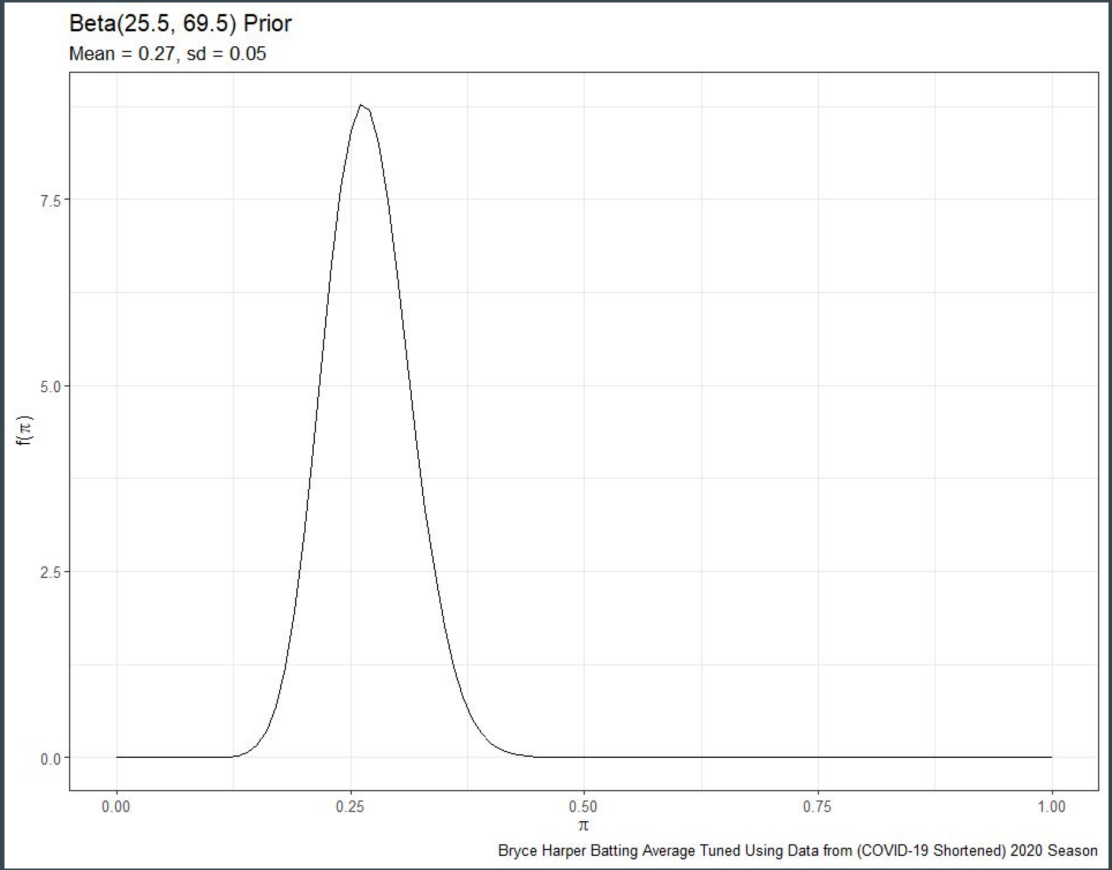


# Hitting Characteristics

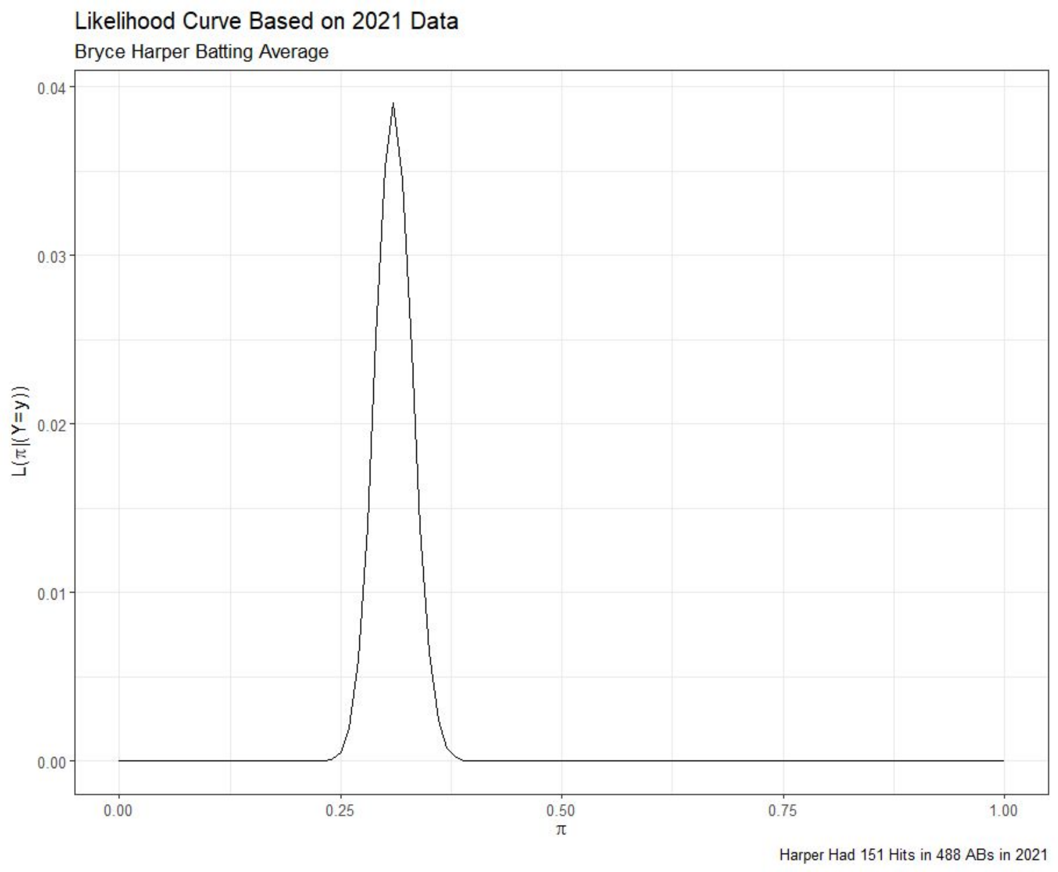


# wRC+ and OPS Correlation

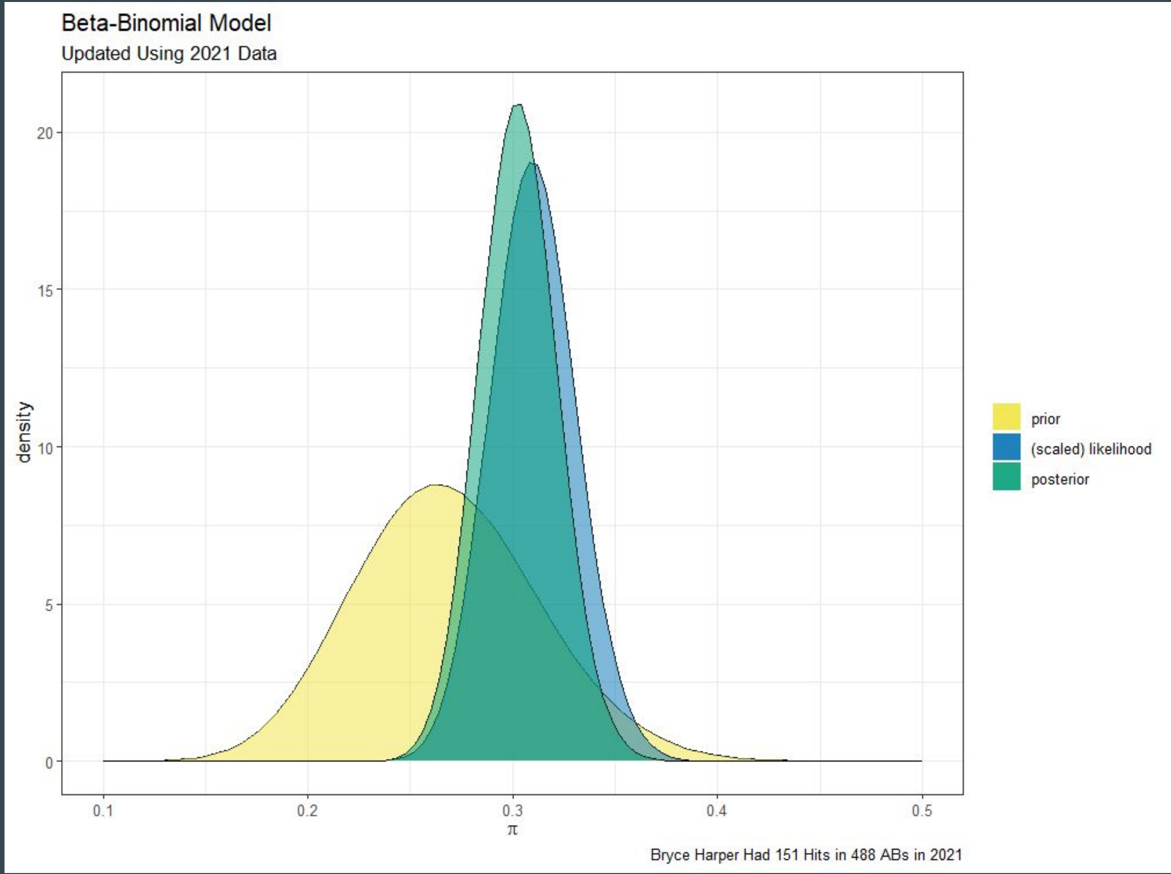
# Bryce Harper Batting Average: A Beta-Binomial Example



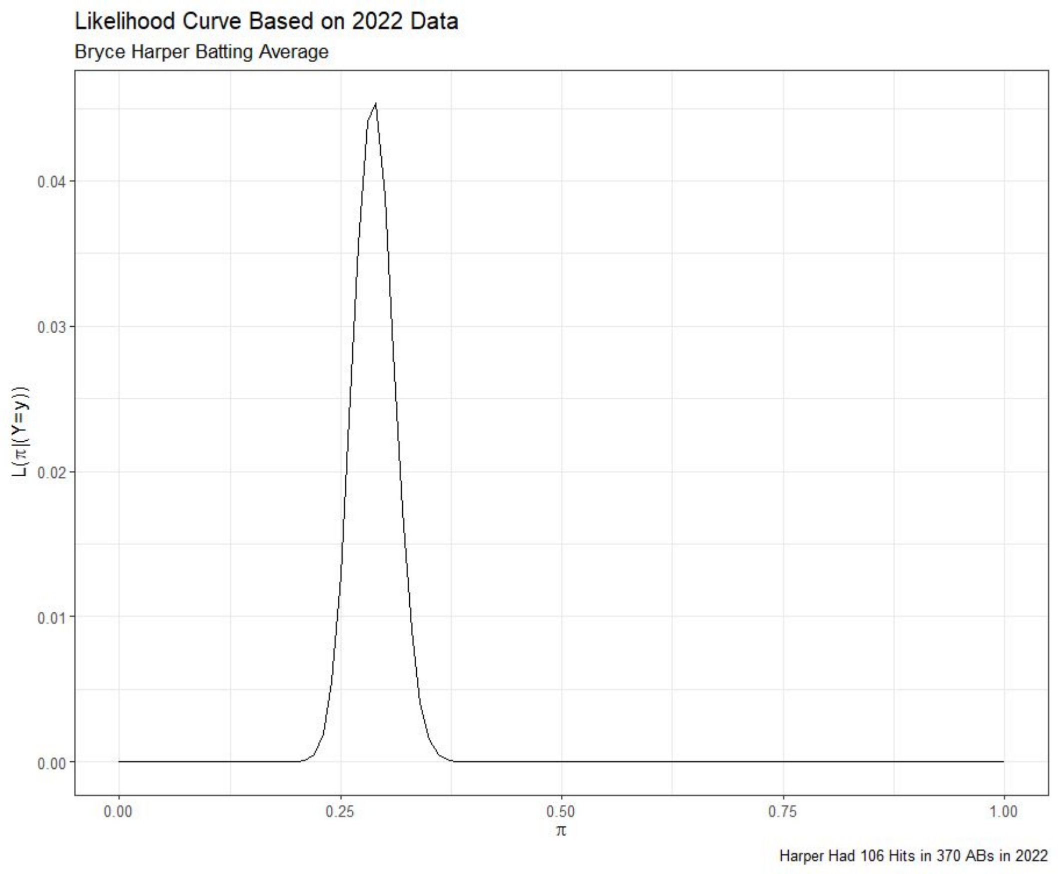
# 2021 Likelihood



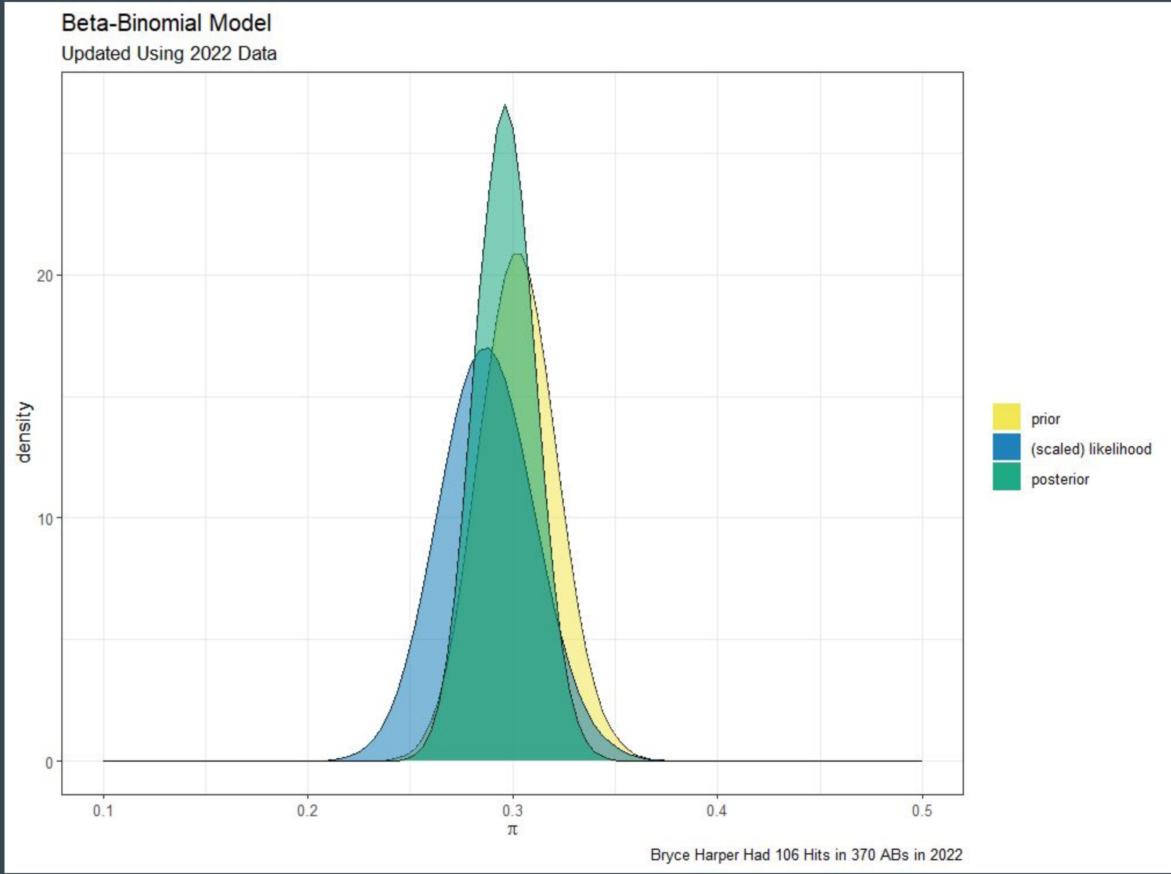
# Update Prior Using 2021 Data



# 2022 Likelihood

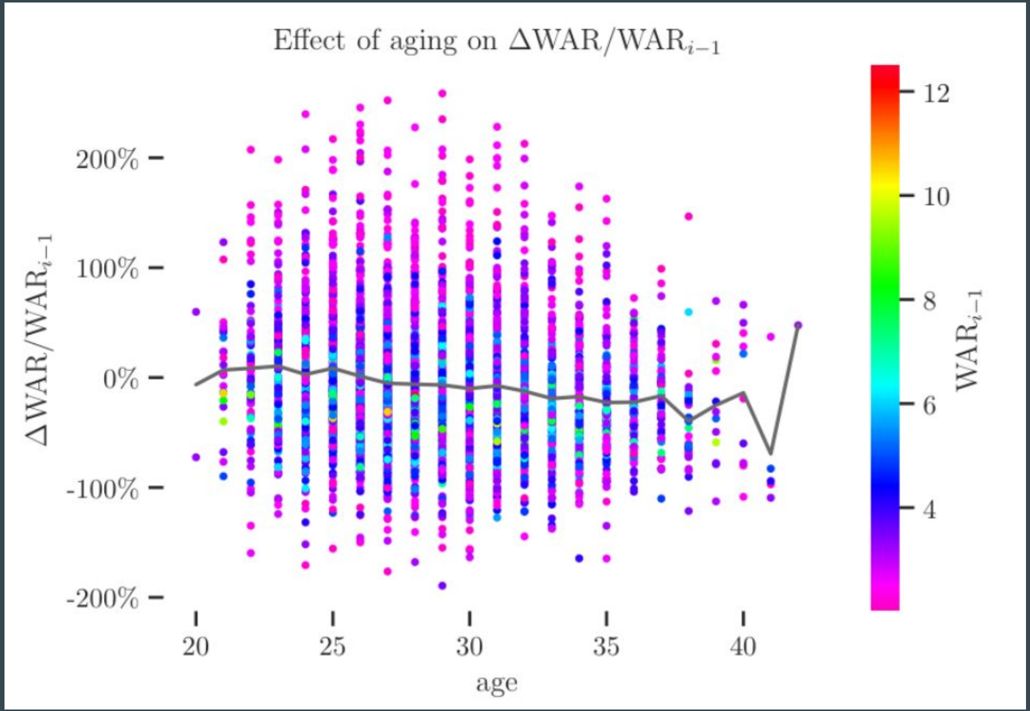


# Update Beliefs Using 2022 Data



# Traditional Methods for Predicting the Aging Curve

- Projecting the process of aging, in terms of WAR (Wins Above Replacement), is essential for executives looking to make contract decisions.
- Traditionally, this is done in one of two ways: averaging across all players of a given age or looking at aging patterns of players that are “nearest” to the player of interest.
- These methods are **not ideal** when it comes to predicting the performance of any given player.
- In the case of the average curve, the aging pattern for the average player does not map work well for outliers.
- The k-nearest neighbors approach is complicated by the problem of quantifying “nearness” between players.
- Importantly, neither method can give a probability distribution over next season’s performance for a given player.




Adapted from Cory Frontin (Baseball Prospectus, 2019)

# Proposed Prior Model

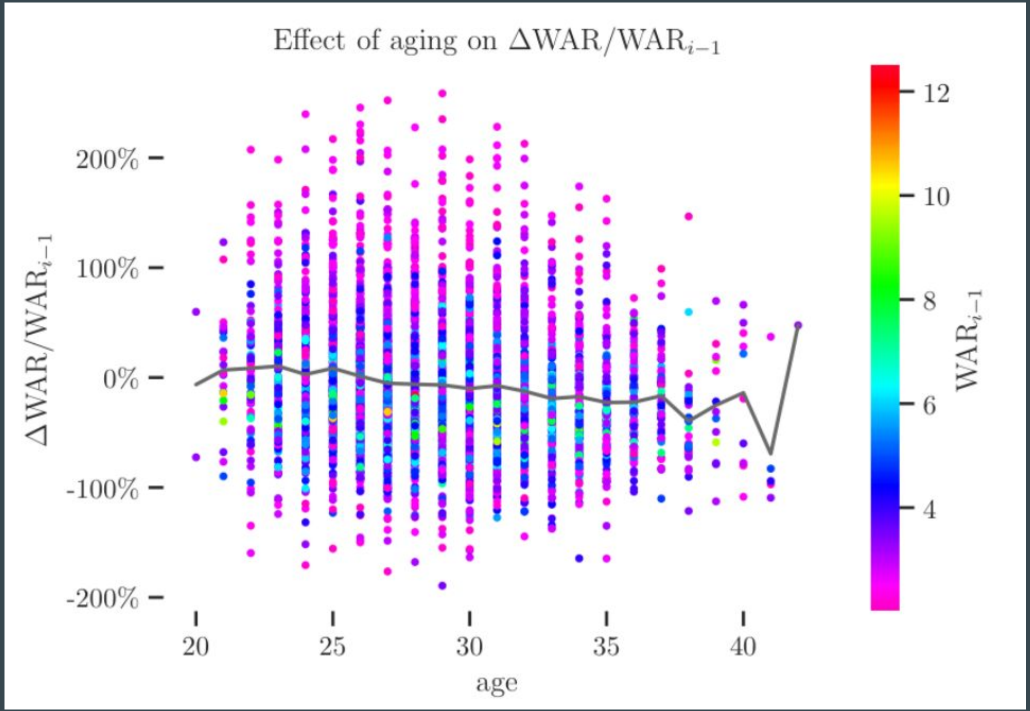
$$\left(\frac{\Delta\text{WAR}}{\text{WAR}_{i\text{-}1}}\right)_i \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mu_{\text{model}} = \left(\alpha_0 + e^{-\alpha_1 WAR_{i-1} - \alpha_2 WAR_{i-1}^2}\right) \times \left(\alpha_3 \text{age}^2 + \alpha_4 \text{age} + \alpha_5\right)$$

$$\sigma^2_{\text{model}} = \left(\beta_0 + e^{-\beta_1 WAR_{i-1} - \beta_2 WAR_{i-1}^2}\right) \times \left(\beta_3 \text{age}^2 + \beta_4 \text{age} + \beta_5\right)$$

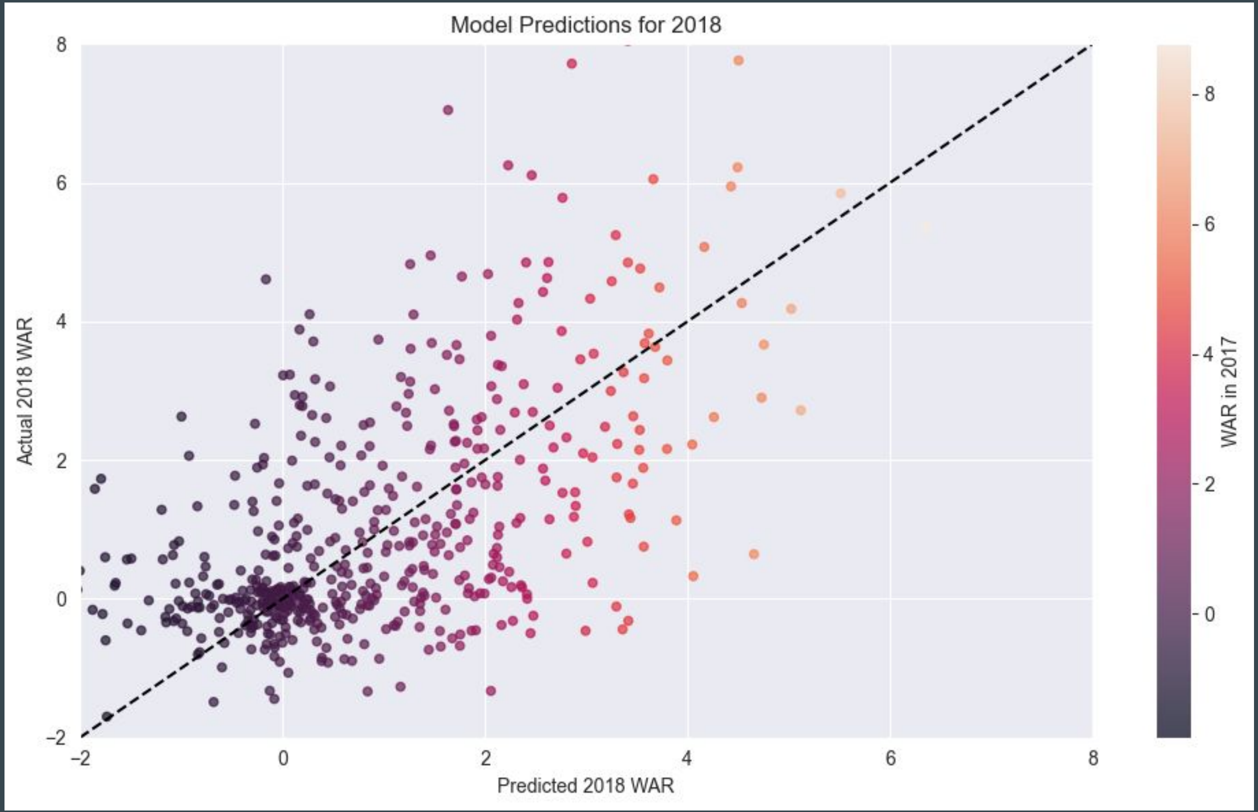
Functional form proposed by Cory Frontin (Baseball Prospectus, 2019) based on observed patterns in aging trends

Parameters were estimated using MLE from data on qualified hitters from 1955 to 2018



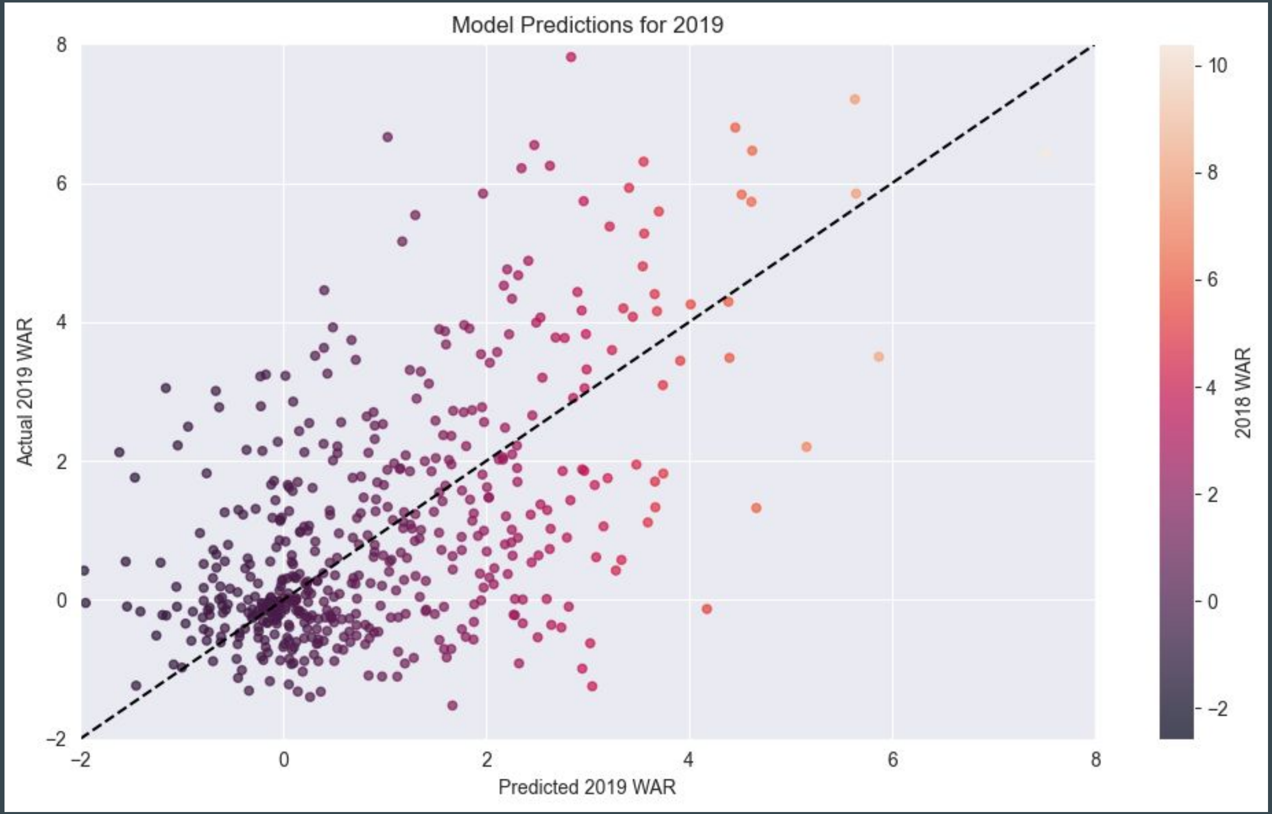

Adapted from Cory Frontin (Baseball Prospectus, 2019)

# Model Predictions for 2018




Data from Fangraphs

# Model Predictions for 2019



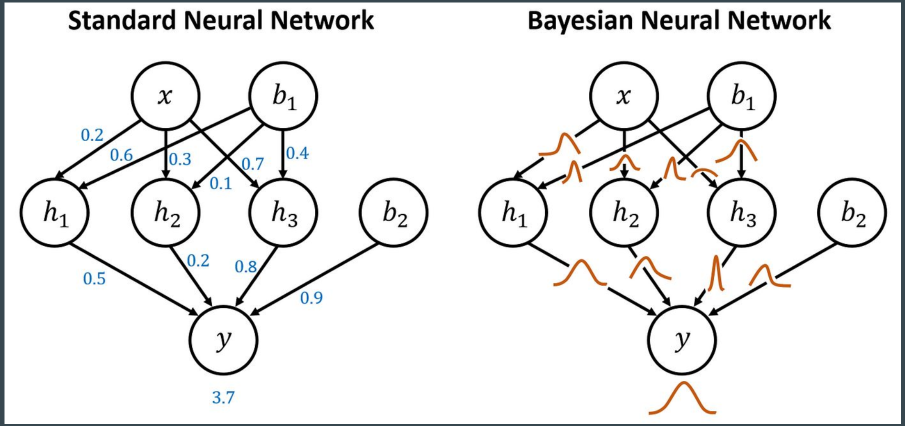

Data from Fangraphs

# Next Steps?

1. Use BNNs to project oWAR for players over time using historical player data to derive probability distributions.

2. Possibly use hierarchical models or poisson regression to account for differences in injury rates due to factors such as player position, age, and playing style.

3. Consider this analysis to also include pitchers and perform factor analysis.



# Questions?

# Thanks!