2023 MLB MVP Prediction

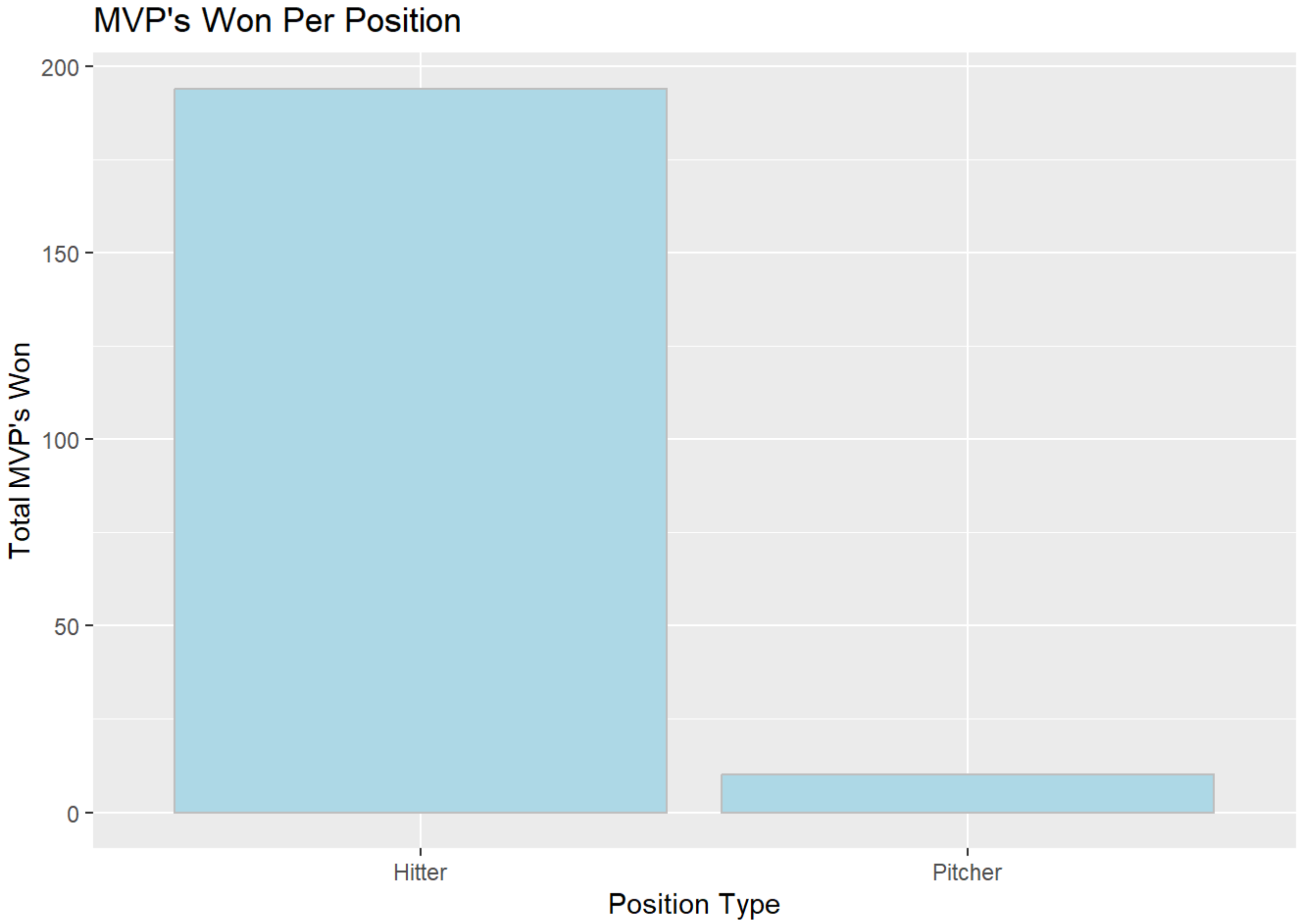
Author: Colin Macy

Summary:

This paper will analyze MLB player statistics over the last 112 years in order to predict the Most Valuable Player for the 2023 season. Using Lahman's Baseball Database we will track previous MVP awards and player statistics to form our predictions.

Positions:

We first want to look at the position types of the MVP award to see if we can narrow down a probable position for the award. Using some filtering, join functions, and ggplot, we are able to produce the graph below from our data.



The graph shows that hitters are significantly more likely to receive the MVP award. This is most likely due to the fact pitchers have their own CY Young award.

Consistency:

Now that we can narrow out the fact that the MVP will be a position player, we can dive into the stats of what makes someone qualified to win the award. First off, to be the Most Valuable Player you likely need to be getting consistent at bats

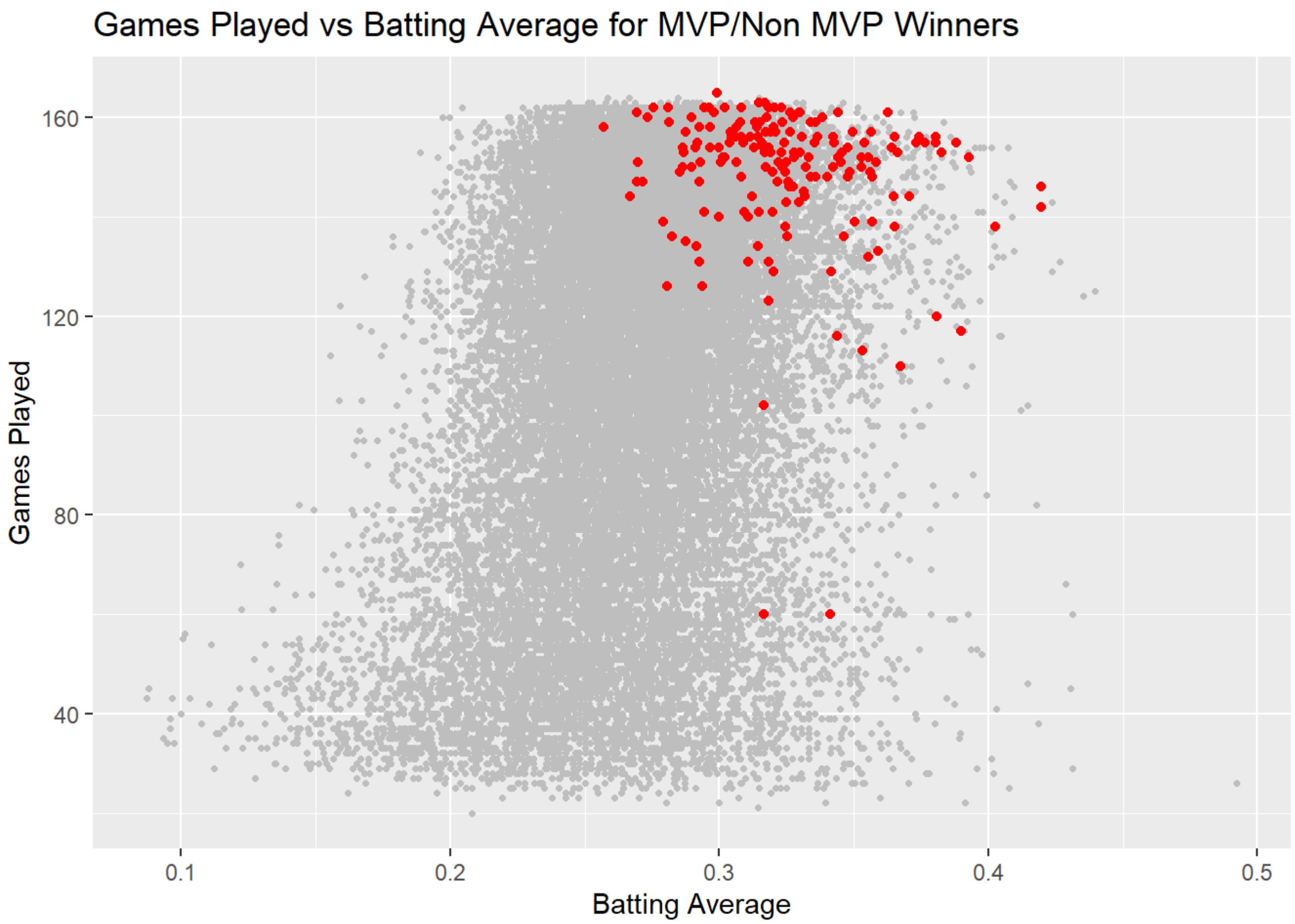
```
mvp_batting = mvp_list %>%
  filter(position != 'Pitcher') %>%
  summarize(avg_ab = mean(AB), avg_hit = mean(H))
```

When calculating the mean at bat count for all MVP’s, we can observe that they obtained, on average, 522 at bats per season, and an average of 168 hits. That comes out to a .322 batting average.

```
average_batting = batting %>%
  filter(AB/G >= 3.1) %>%
  summarize(avg_ab = mean(AB), avg_hit = mean(H))
```

To be a statistically qualified batter, one must obtain a 3.1 AB / game average throughout the season. Amongst those hitters, the average player has about 348 at bats per season, and 95 hits. This comes out to be a .273 batting average.

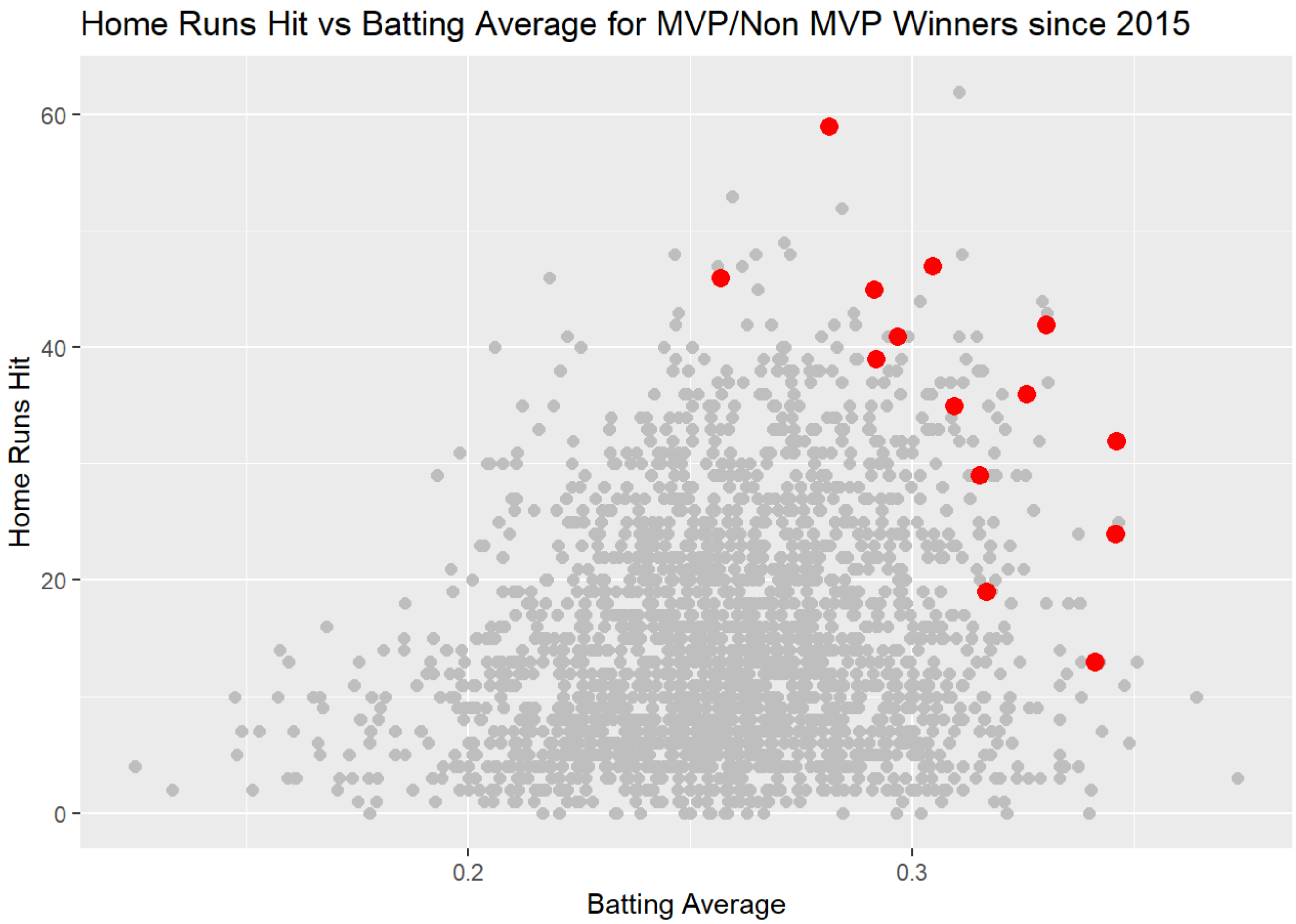
When plotted amongst all qualified hitters, a graph can be depicted with red dots signifying MVP winning seasons.



From this data, we can observe a very clear correlation between the number of games played and a higher batting average resulting in a more favorable chance to win MVP.

Distribution

Lets try to narrow down some of this data into more recent years. After all, the game is consistently changing so lets get an idea of how more recent MVP’s have been performing. We can also compare the new averages to home runs.



Once again, we have used the red points to mark MVP winners since 2015.

We can now take these averages of MVP vs non MVP winners and use a Binomial Distribution to determine the probability of obtaining the given stats within a season.

Probability:

The average home run count for MVP’s since 2015 is 36 with a batting average of .310. For non-mvp’s it rests around 15 home runs with a .259 batting average. Using this information with our at bat size and rate for how often a home run is hit per at bat, we can calculate the probability of achieving MVP level stats in a given season.

```
n = 506

non_mvp_atbats = 382
non_mvp_mean_hr = 15
non_mvp_avg = .259
mvp_avg = .310
mvp_hr_total = 36

non_mvp_hr_rate = non_mvp_mean_hr/non_mvp_atbats
mvp_hit_total = n*mvp_avg


(1 - pbinom(mvp_hr_total, n, non_mvp_hr_rate))*100
```

```
## [1] 0.02771998
```

```
(1 - pbinom(mvp_hit_total, n, non_mvp_avg))*100
```

```
## [1] 0.5533046
```

Using pbinom, if an average player was given a full workload in terms of at bats as previous MVP winners were given, they would, on average, put up MVP like HR numbers .02% of the time and MVP like AVG numbers .55% of the time. As we can see from this distribution, the power numbers in terms of HR is what mostly sets the MVP apart from the rest of the competition.

Data Collection:

Now we can look through the data we have gathered and analyze what it takes to be MLB’s most valuable player.

1. It more than likely MUST be a position player/hitter.
2. Batting average and home run numbers must be near .310 and 36 respectfully.

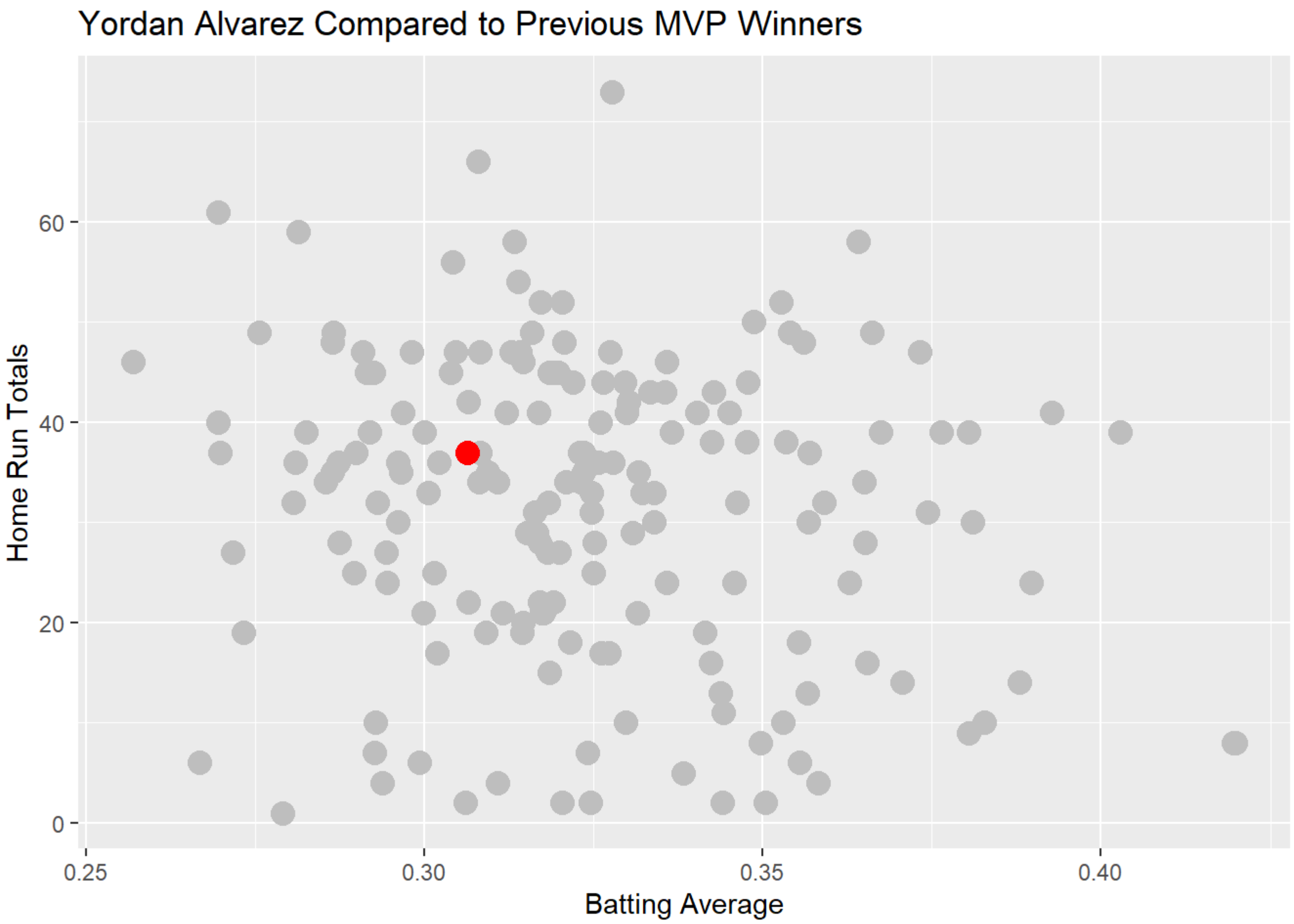
We can now look at players who have put up similar numbers within the last 5 years to see who is on trend to break out this next year.

```
possible_mvps = non_mvp_ba %>%
  filter(yearID >= 2018, G > 100, avg >= .300, HR >= 30)
```

When sorting players by individuals who have put up above a .300 average and above 32 home runs, the only three that come up are Yordan Alvarez, Aaron Judge, and Paul Goldschmidt.

Aaron Judge and Paul Goldschmidt are coming off a 2022 campaign where they both won the MVP award. Having said that, Yordan Alvarez is much younger and coming off his best season yet. If there were to be a new MVP award winner, it is my assumption that **Yordan Alvarez** would be the one to take it. Let’s look at his stats from his 2022 campaign amongst previous MVP winners to see how he would stack up.

```
ggplot(NULL, aes(x = avg, y = HR)) + geom_point(data = mvps, color = 'grey', size = 4) + geom_point(data = yordan, color = 'red', size = 4) + xlab('Batting Average') + ylab('Home Run Totals') + ggtitle('Yordan Alvarez Compared to Previous MVP Winners')
```



According to the comparisons, we can observe that Yordan Alvarez did indeed put up MVP type numbers last season. Because of this trend and based on data collected, he will win the AL MVP award in 2023.

alt text here