

# Baseball Hall of Fame Report

Andy, Dante, Kai, and Kobe

2023-03-02

## Introduction

The data we are working with is career baseball statistics. The data was collected from official scorekeepers who went to every Major League Baseball game and tracked various outcomes players had over the course of each game. For our finished data set, those statistics were added up at the end of each baseball season and also at the end of each player's career. The motivation for the scorekeepers to track the data was that they were paid, either by the home team, the league as a whole, or the local newspaper to track these statistics to keep baseball fans informed.

One motivation for our research question was to better understand which variables are associated with a successful hall of fame induction. Another motivation for our research is an extension of the first, which is for us to predict which players will be inducted into the hall of fame in the future.

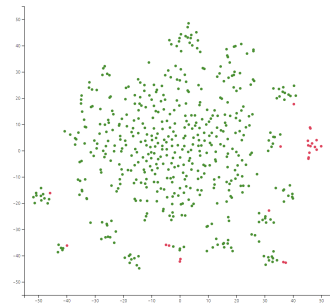
As such, our first research question is what factors are associated with a position player being inducted into the hall of fame and whether we can create a model that uses these factors to predict what players will make the hall of fame. Our other research question is what factors are associated with whether a pitcher is inducted into the hall of fame and whether we can create a model that uses these factors to predict what pitchers will make the hall of fame.

## Data Description

Our data consists of mostly quantitative variables mixed in with some categorical variables. Most baseball metrics come in quantitative form so it is often hard to obtain categorical measurements unless they come directly from the quantitative variables itself. However, some categorical variables we may use include factors such as year, league, team, stint, and possibly some other categorical variables that are computed from other quantitative variables. One other potential categorical variable would be a binary variable indicating whether a player was known to use steroids or not.

Furthermore, we are splitting this project into two parts: batters and pitchers. Thus, there is a need to split the data; in other words, we'll have two different data sets, since metrics for batters and pitchers are vastly different. However, we'll still be predicting the same variable: hall of fame. Hall of fame is a binary categorical variable (with the two values being 1 and 0), indicating whether a player was inducted into the hall of fame or not. In total, we have around 750 observations in the position players training data set and 400 observations in the pitching training data set,.

# Exploratory Data Analysis



## Steroids Clustering

## Methods

## Results

Our final models are as follows:

- Position Players:

```
wein ~ all_star + Pos + HR + nice_guy_awards + Steroids + votedBy + RBI + AVG +
most_valuable_player + R + SB + Pos:HR + all_star:most_valuable_player + Steroids:SLG
```

- Pitchers:

```
wein ~ W + L + SO + SV + Steroids + most_valuable_player + all_star + cy_young_award
+ rolaids_relief_man_award + votedBy + StarterReliever + Steroids:most_valuable_player
```

## Coefficients

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-26.466814	6.872037	-3.850787	0.000187
W	0.138425	0.038255	3.618862	0.000468
L	-0.054886	0.031140	-1.765460	0.078242
SO	0.007885	0.009810	1.254148	0.210778
SV	0.0188757	0.011945	1.582873	0.112287
Steroids	-0.128868	0.06322162	-2.050726	0.041184
most_valuable_player	0.007217	0.002548	2.832029	0.004377
all_star	1.178346	0.242601	4.857023	0.000024
cy_young_award	0.046729	1.093014	0.043029	0.970574
rolaids_relief_man_award	1.702841	0.853307	1.994881	0.046885
votedBy	0.007689	2.484506	0.003124	0.997182
StarterReliever	-0.154227	3.444706	-0.044603	0.963891
Steroids:most_valuable_player	-0.003837	0.12341497	-0.034474	0.984278

## QQ-Plots:

