Baseball Salary Analysis - Project Proposal

Edgar Hernandez, Ryan Sims, Jessica Tomas

2022-07-26

Contents

Description of Dataset

This dataset contains pitching, hitting, and fielding statistics for Major League Baseball from 1871 through 2021. 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.

```
library(Lahman)
```

Details:

- The relevant portion of the dataset that we will be using contains 26 variables. Some of the more critical variables are `Salary`, `Team`, `Player`, `Wins`, `Losses`, `Position`, `Games Started`, `Runs`, `Hits`, and `Homeruns`.

```
summary(salary_data)
```

```
##    playerID            yearID         teamID      lgID
##  Length:33937       Min.   :1985   SLN    : 1327   AL:16499
##  Class :character   1st Qu.:1993   LAN    : 1259   NL:17438
##  Mode  :character   Median :2001   SDN    : 1258
##                     Mean   :2001   OAK    : 1238
##                     3rd Qu.:2009   PIT    : 1237
##                     Max.   :2016   NYN    : 1202
##                                    (Other):26416
##      salary            POS                 GS            InnOuts
##  Min.   :       0   Length:33937       Min.   :  0.00   Min.   :   0.0
##  1st Qu.:  275000   Class :character   1st Qu.:  1.00   1st Qu.: 113.0
##  Median :  522500   Mode  :character   Median : 14.00   Median : 339.0
##  Mean   : 1886801                      Mean   : 34.09   Mean   : 908.2
##  3rd Qu.: 2000000                      3rd Qu.: 45.00   3rd Qu.:1206.0
##  Max.   :33000000                      Max.   :162.00   Max.   :4388.0
##
##        PO               A                E                DP
##  Min.   :   0.0   Min.   :  0.00   Min.   : 0.000   Min.   :  0.000
##  1st Qu.:   3.0   1st Qu.:  2.00   1st Qu.: 0.000   1st Qu.:  0.000
##  Median :  13.0   Median :  9.00   Median : 1.000   Median :  1.000
```

```
##  Mean   : 102.3   Mean   : 39.18   Mean   : 2.524   Mean   :  9.566
##  3rd Qu.: 107.0   3rd Qu.: 27.00   3rd Qu.: 3.000   3rd Qu.:  4.000
##  Max.   :1597.0   Max.   :570.00   Max.   :42.000   Max.   :176.000
##
##        AB               R                H               X2B
##  Min.   :  0.0   Min.   :  0.00   Min.   :  0.00   Min.   : 0.000
##  1st Qu.:  5.0   1st Qu.:  0.00   1st Qu.:  1.00   1st Qu.: 0.000
##  Median :130.0   Median : 14.00   Median : 30.00   Median : 5.000
##  Mean   :194.9   Mean   : 25.93   Mean   : 51.17   Mean   : 9.882
##  3rd Qu.:347.0   3rd Qu.: 44.00   3rd Qu.: 91.00   3rd Qu.:17.000
##  Max.   :716.0   Max.   :152.00   Max.   :262.00   Max.   :59.000
##
##       X3B               HR              RBI               BB
##  Min.   : 0.000   Min.   : 0.000   Min.   :  0.00   Min.   :  0.00
##  1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.:  0.00   1st Qu.:  0.00
##  Median : 0.000   Median : 1.000   Median : 13.00   Median : 10.00
##  Mean   : 1.092   Mean   : 5.455   Mean   : 24.53   Mean   : 18.75
##  3rd Qu.: 2.000   3rd Qu.: 8.000   3rd Qu.: 40.00   3rd Qu.: 30.00
##  Max.   :23.000   Max.   :73.000   Max.   :165.00   Max.   :232.00
##
##        SO              IBB              HBP               SH
##  Min.   :  0.00   Min.   :  0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:  2.00   1st Qu.:  0.000   1st Qu.: 0.000   1st Qu.: 0.000
##  Median : 25.00   Median :  0.000   Median : 0.000   Median : 0.000
##  Mean   : 35.99   Mean   :  1.505   Mean   : 1.738   Mean   : 1.799
##  3rd Qu.: 58.00   3rd Qu.:  2.000   3rd Qu.: 2.000   3rd Qu.: 3.000
##  Max.   :223.00   Max.   :120.000   Max.   :35.000   Max.   :39.000
##
##        SF              GIDP
##  Min.   : 0.00   Min.   : 0.000
##  1st Qu.: 0.00   1st Qu.: 0.000
##  Median : 1.00   Median : 2.000
##  Mean   : 1.65   Mean   : 4.437
##  3rd Qu.: 3.00   3rd Qu.: 7.000
##  Max.   :17.00   Max.   :35.000
##
```

Dataset Background Information

The data is copyright 1996-2022 by Sean Lahman, licensed under a Creative Commons Attribute-ShareAlike 3.0 Unported License. Chris Dalzell and his team maintain the R package and library that we will leverage in this project. This dataset was originally created in July 1995 by Sean Lahman, who was at the timeSoftware Developer for Eastman Kodak, but later worked as a Digital Publishing Director for Team Sports Publishing, a sports columnist for the New York Sun, and now is the Data Projects Manager at the Society for American Baseball Research (SABR). SABR is infact the organization from which the term Sabermetrics (originally SABRmetrics) is derived, which is the empirical analysis of baseball, especially the statistics thereof.

This dataset contains data going back as far as 1871. We will however only be looking back so far as 1985 and only up to 2016, because the salary data only covers that range, and we are interested in the salaries of players. We might consider also limiting the scope to include only players who are not pitchers or catchers, as the statistical metrics used to measure their performance are different than the metrics used for all of the other positions in baseball.

Dataset Interest

We will be examining what statistics related to a baseball player's performance have a significant effect on the salary of a player. This is particularly interesting as it is sort of the flip-side of the coin to the traditional Sabermetrics employed by the Oakland Atheletics in the 1990's to analyze player statistics to try to assemble a winning team. We will be instead be leveraging the dataset to instead work for the players, hopefully determining which statistic(s) a player should focus on improving in order to increase their compensation.

Evidence of Data

This dataset can be accessed directly in `R` through the `Lahman` library. However, we will also be providing the extracted `.csv` file for the data.