Moneyball Revisited: Analyzing the Relationship Between Performance and Pay in MLB

Jaxon Bennett and Luke Hamm

2024-04-29

Abstract

Each year, Major League Baseball (MLB) witnesses a flurry of free agent activity, with players commanding vastly different salaries based on their perceived value to teams. Our study revisits the principles of "Moneyball" by exploring the intricate relationship between player performance metrics, personal attributes, and compensation in MLB. We aim to identify the key statistical indicators that significantly influence player salaries, focusing particularly on the 2022 season and subsequent 2023 free agency deals.

Introduction

Every year there are hundreds of free agents in Major League Baseball (MLB). Some make hundreds of millions of dollars in free agency, while others make just a small fraction of that. Our project aims to explore the correlation between MLB players' salaries, performance metrics, and other relevant information. The objective is to identify the key statistics that significantly influence a player's earning potential, especially as they approach their contract year. Given that front offices tend to keep their player analysis internal and inaccessible to the public, we intend to employ linear regressions and other analytical tools to uncover the metrics that carry the most weight in the evaluation of potential free agents by these front offices. Additionally, we intend to employ statistical tests to compare the salaries of right-handed hitters with those of left-handed hitters. Given the common belief that left-handed hitters hold greater value than their right-handed ones, we aim to use the findings from our linear regression to investigate whether two players with identical statistics and metrics, but differing dominant hands, would command equal compensation upon entering free agency, or if left-handed hitters would indeed earn more.

Our analysis focuses on players' statistics from the 2022 season and their salaries for the subsequent year, 2023. This timeframe is significant because players typically negotiate contracts during the offseason following the 2022 season, and the terms of these contracts are often influenced by their performance in the preceding season. In this analysis, we utilize two primary datasets. First the Lahman Baseball Dataset, compiled by Sean Lahman, which provides comprehensive historical data on Major League Baseball (MLB) players, teams, and seasons, and is included in R. The second dataset is The Cot's Baseball Contracts dataset (https://legacy.baseballprospectus.com/compensation/cots/), which offers detailed information on player contracts, salaries, and transactions in MLB.

There are a number of different statistics that we will refer to in throughout this paper, the following table can be refered to for clarity:

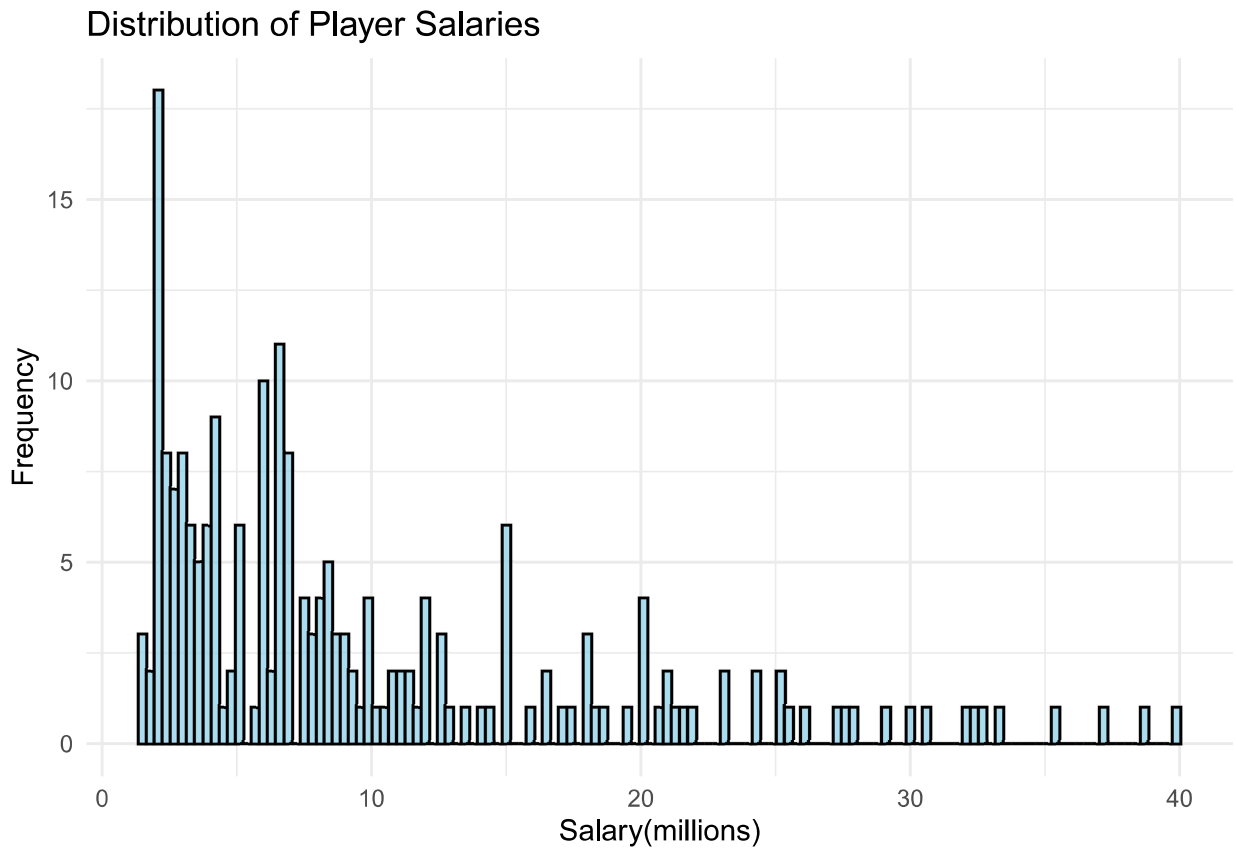
Table 1: Hitting statistics and player information

Name of Stat	Description
Salary	Players Salary for the 2023 season
Age	Players age when they signed their contract (2022)
OPS	Players On Base plus Slugging Percentage
HR per AB	Players home runs divided by their at bats
RBI	Number of runs the player drove in
BB per SO	Players walks divided by their strikeouts
SB	Stolen Bases
MLS	Major League Service time (Number of years in the MLB)
BatsL	Player bats Left Handed
BatsR	Player bats Right Handed

Exploratory Data Analysis

Our analysis began by merging two datasets: the Lahman Baseball dataset and Cot's Baseball Salaries dataset. This merger provided us with a comprehensive dataset containing information on every player who signed a free agent deal during the 2022-2023 offseason. To focus solely on hitters for this analysis, we excluded all pitchers from the dataset. Additionally, we filtered out players with fewer than 50 at-bats during the season, as this criterion often indicates injury while still resulting in substantial contracts the following year. As a result, our final dataset comprises 211 players, each with statistics from the 2022 season and corresponding salaries for the 2023 season. Below shows the distribution of the player salaries for the 2023 season from our dataset.

Figure 1



We can see that Figure 1 shows the distribution of the salaries of mlb hitters from the 2023 season, and that it follows a skewed right distribution. The right skew can be explained by the high percentage of players making league minimum. Below we can notice the correlation between some of the more common hitting statistics and how they correlate to player salary before we run the regression.

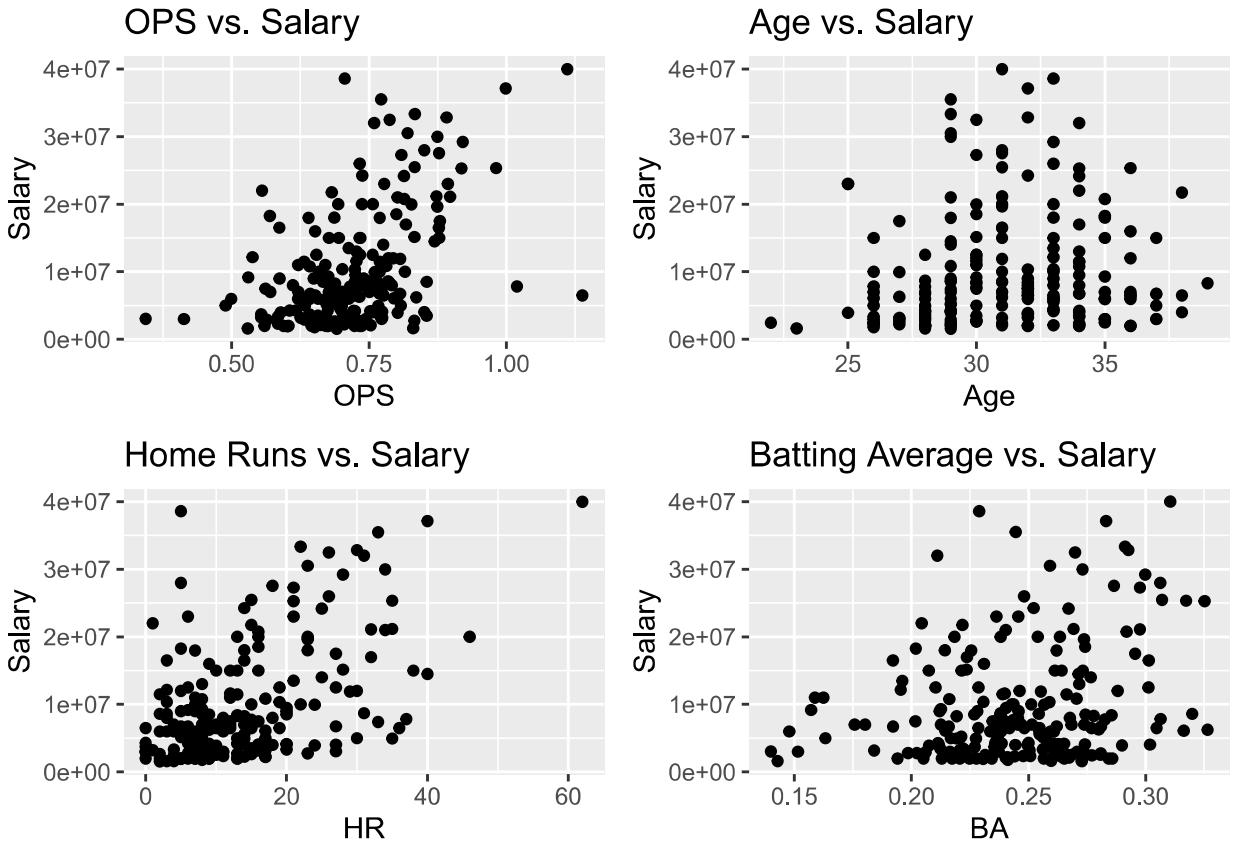


Figure 2

We can see from the above plots that there is varying correlation between the different variables and player salaries. We can begin to see how certain variables have greater correlation to player salary than others. As our dataset has many more variables than are plotted above, we created a regression to find which stats have the highest correlation to player salary.

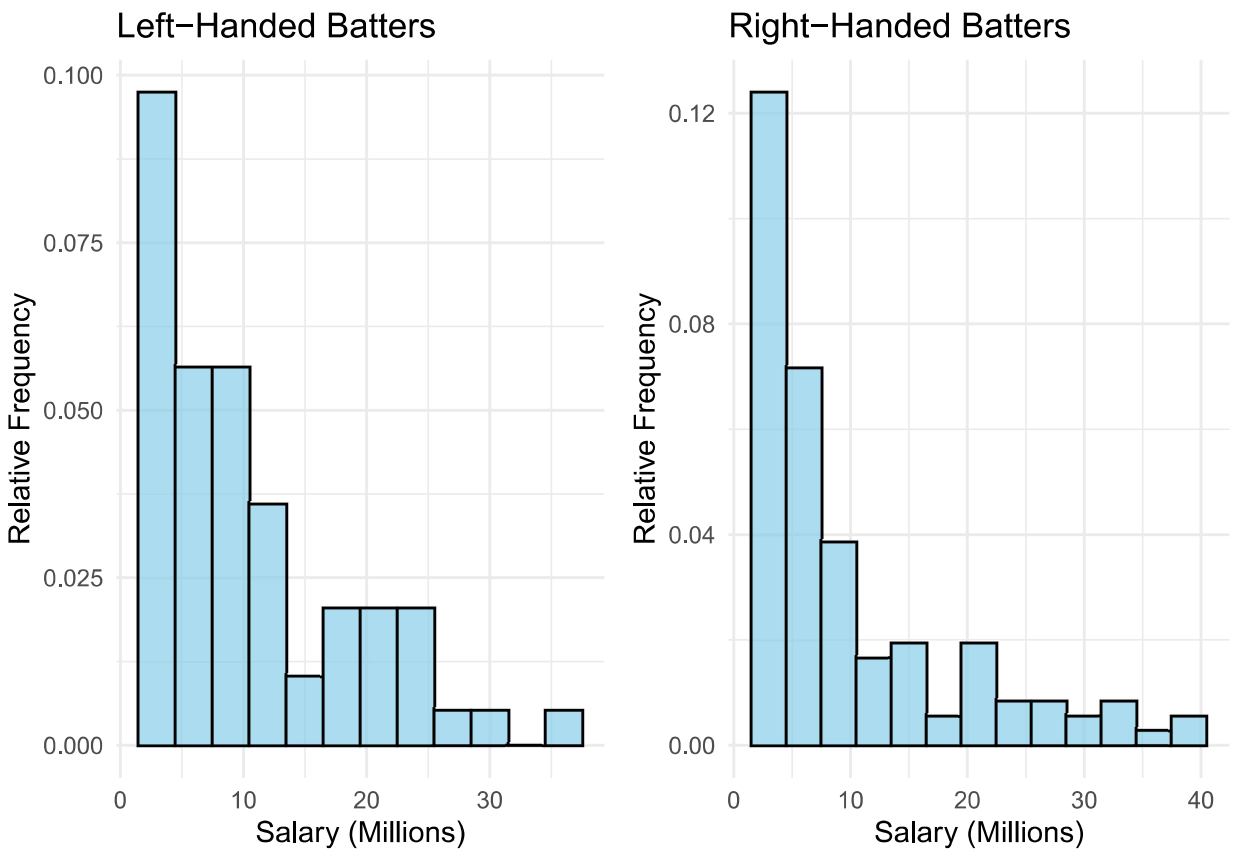


Figure 3

```
## Call:
##    aov(formula = salary_m ~ bats, data = d10)
##
## Terms:
##                       bats Residuals
## Sum of Squares     91.586 15375.264
## Deg. of Freedom         2       208
##
## Residual standard error: 8.597647
## Estimated effects may be unbalanced
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## bats          2     92   45.79   0.619  0.539
## Residuals   208  15375   73.92
```

As the second part of this report is exploring if left handed batters are more valuable than right handed batters, we first created of the distributions of the side by side of player salaries for left and right handed hitters seen above in Figure 3. We used relative frequency as a way to compare to show they they have similar distributions, because there are a greater number of right handed hitters in the MLB than left handed hitters. We also preformed a ANOVA test to confirm what we saw visually. The high P-value indicates that at the 5% significance level, we fail to reject the null hypothesis that the mean salary in millions of left handed batters is equal to that of right handed batters. Subsequently, in our Statistical Analysis section, we preform a hypothesis test to discern whether left-handed hitters indeed see higher salaries compared to their right-handed counterparts.

Statistical Analysis

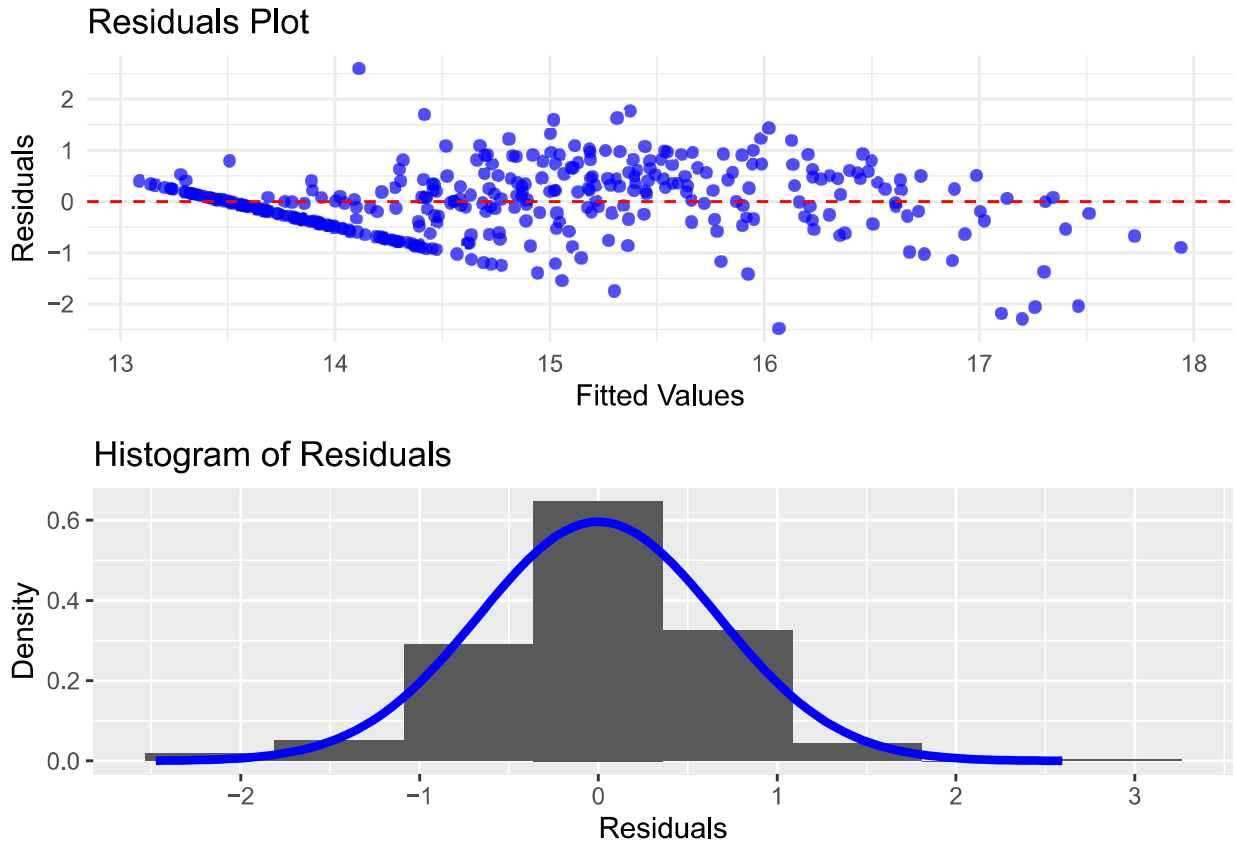
Linear Model

```
##
## Call:
## lm(formula = salary_ln ~ age + OPS + HR_per_AB + RBI + BB_per_SO +
##     MLS + bats, data = d9)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.47505 -0.38188  0.00457  0.41368  2.59556
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.976262   0.562667  24.839  < 2e-16 ***
## age         -0.041482   0.018715  -2.216 0.027315 *
## OPS         -0.240248   0.516813  -0.465 0.642323
## HR_per_AB    3.276514   2.719269   1.205 0.229063
## RBI          0.014691   0.001753   8.382 1.36e-15 ***
## BB_per_SO    0.729585   0.196980   3.704 0.000247 ***
## MLS          0.270646   0.019957  13.561  < 2e-16 ***
## batsL        0.155299   0.122679   1.266 0.206409
## batsR        0.181371   0.119870   1.513 0.131184
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6769 on 343 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.7133, Adjusted R-squared:  0.7067
## F-statistic: 106.7 on 8 and 343 DF,  p-value: < 2.2e-16
```

```
##    Variable Multiplier Percent_Change
## 1       Age  0.9593666          -4.1%
## 2       OPS  0.7864328         -21.4%
## 3 HR_per_AB 26.4832900       2,548.3%
## 4       RBI  1.0147990           1.5%
## 5 BB_per_SO  2.0742200         107.4%
## 6       MLS  1.3108110          31.1%
## 7     BatsL  1.1680070          16.8%
## 8     BatsR  1.1988600          19.9%
```

The linear model above is the reduced linear regression we created with the natural log of salary as the predictor variable. Based on the data, the variables with the highest correlation to salary are: Age, RBI, BB_per_SO, and MLS, as seen above in the regression. To determine this, we looked at the statistical significance of each of the beta coefficients by examining the individual standard errors and doubling them. If they were larger than the coefficient, then we knew it was not statistically significant. We then cross-checked these findings with the individual P-values and confirmed that the four variables were statistically significant. Since we used a log-linear regression model, our interpretation of the beta coefficients looked a little different than normal. Using age as an example variable, we can tell that roughly for every year increase in a player's age, the predicted salary will decrease by about 4.1%. This is true for all the variables; however, we also found the actual estimates by using the percent change formula and found the percentage changes of all the estimates, which are included in the table below. The two variables with the most impact were MLS at a 31.1% change

and BB_per_SO at 107.4%. Both of these make sense; for Major League service time (MLS), as a player spends more time in the MLB and demonstrates value, they will naturally be worth more money. For walks per strikeout (BB_per_SO), this stat has a high predictive percentage because the state will be a low number less than one, so a player will most likely never have one unit of change in this variable in a given year. This means the real percentage change in one year will be much lower than 107.4%. Home runs per at-bat (HR_per_AB) was another variable like BB_per_SO that had a massive percentage effect on salary per one unit of change, but this would never be realized because of the tiny number it would be in one year. This variable was statistically insignificant, so we did not include it as one of our noted predictors of salary. **Figure 4**



There are a few assumption we need to make. We first have to assume linearity in the regression. In the Figure 4 above, there does seem to be a clear linear pattern in the residuals. We think this pattern exists due to the league minimum salary of $700,000. We thought it might be colinearity as well and tried dropping various variables from the regression however, this failed to correct the problem. We next have to assume nearly normal residuals. In Figure 4, the plot of the residuals roughly follows a normal distribution, so we can confirm this assumption. The final assumption is that the variability is constant. In Figure 4 we can see that excluding the league minimum line, the data is spread fairly evenly in the plot of the residuals.

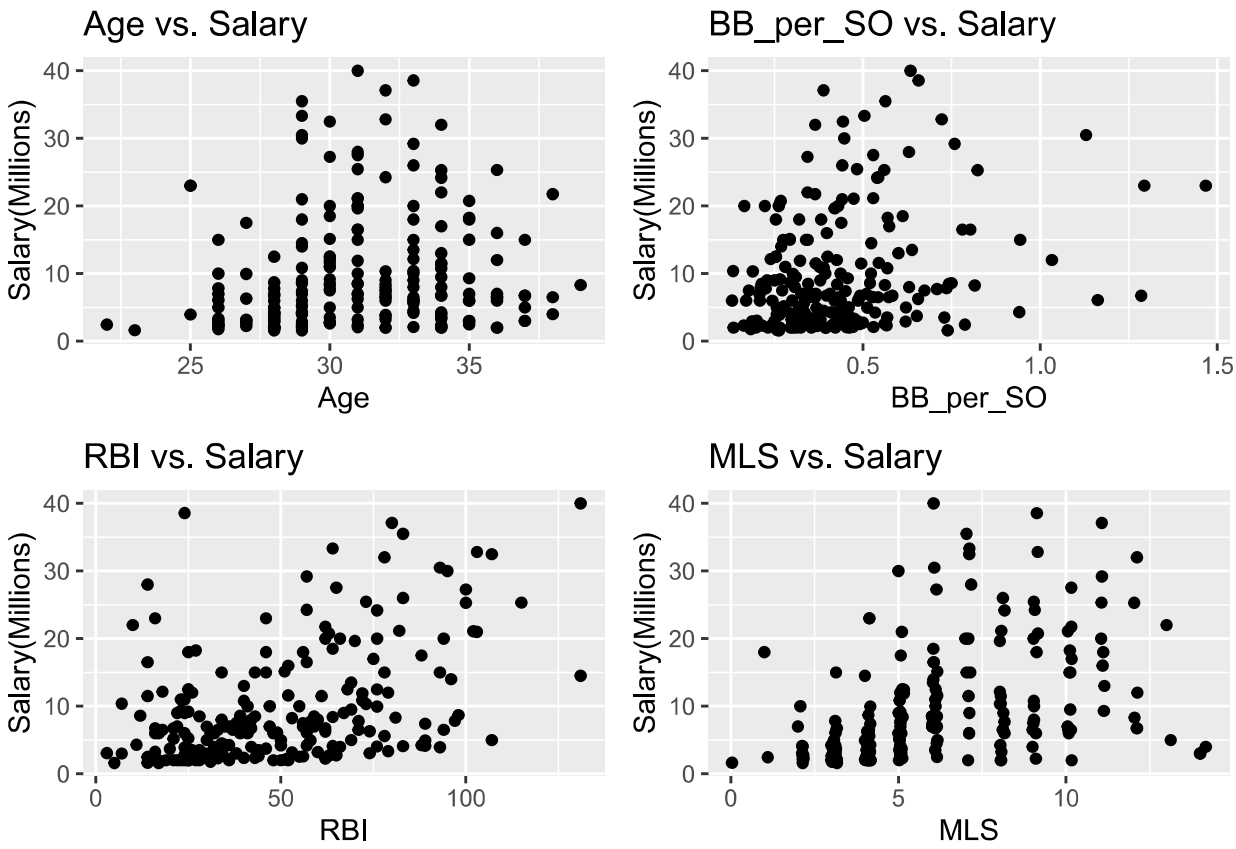


Figure 5

The scatterplot on the bottom left in Figure 5 illustrates the correlation between OPS and Salary, revealing a notable relationship between a player's offensive performance and their subsequent salary. Furthermore, the data reflects an intuitive trend: as a player's RBI count increases, their salary for the following season tends to increase as well. The same trend can be seen in the top right for BB_per_So compared to Salary. Similarly, the scatterplot depicting age versus salary in the top left aligns with expectations, as it showcases a common career trajectory in Major League Baseball. Typically, players secure their initial substantial contracts around the age of 30, following the expiration of their rookie deals, a pattern also portrayed by this plot.

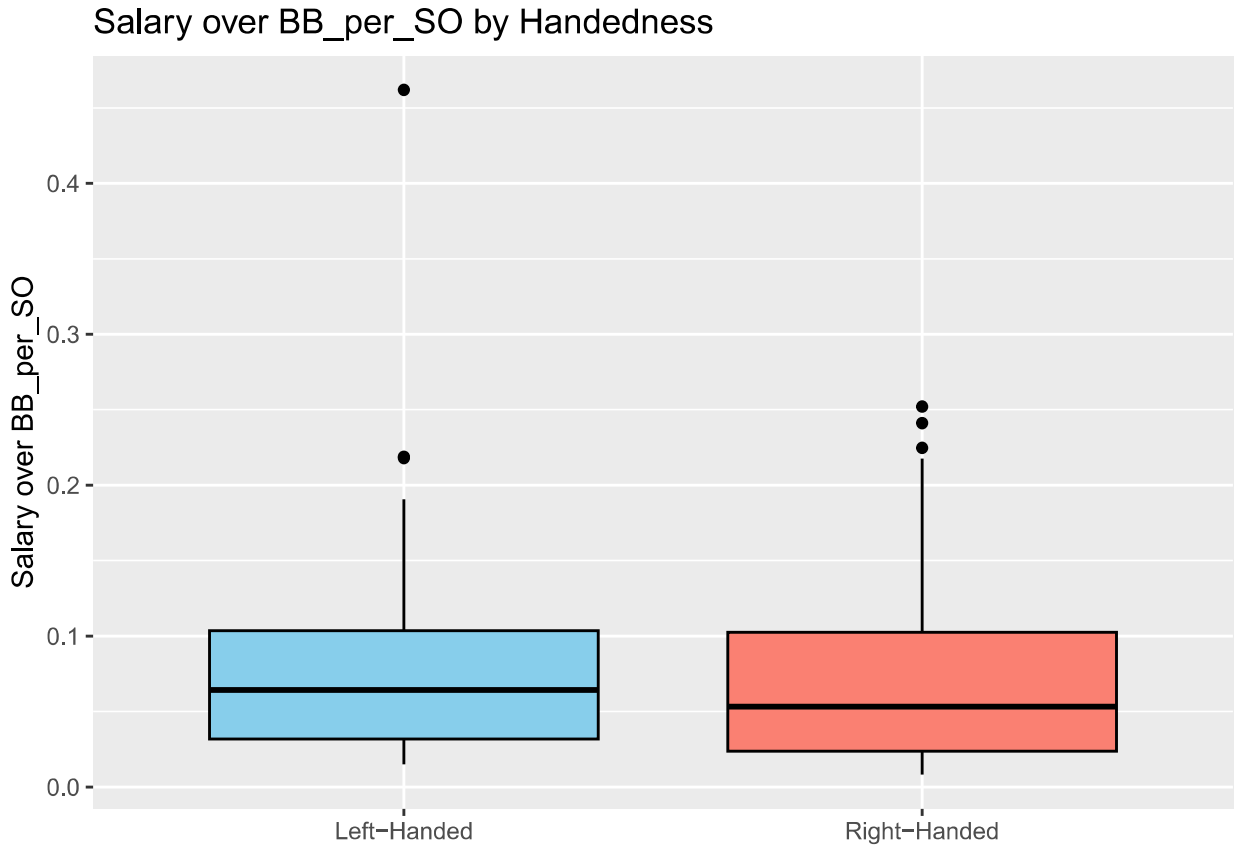
Hypothesis Test for left vs right handed hitters Now that we have found that BB_per_SO is the most significant statistical predictor for player salaries, we now want to test to see if we should expect to see a difference in salary for two players with identical stats, one being left handed, the other being right handed. Our Hypothesis are as follow: Null hypothesis (H_0): There is no difference in the average salary based on BB_per_SO for left-handed and right-handed hitters. Alternative hypothesis (H_a): The average salary based on BB_per_SO, is higher for left-handed vs. right-handed hitters.

In order to test this, we used a Welch's Two Sample t-test with a 95% confidence interval.

```
##
##  Welch Two Sample t-test
##
## data:  left_salaries$salary_per_BB_per_SO and right_salaries$salary_per_BB_per_SO
## t = 0.59785, df = 116.31, p-value = 0.2756
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.01095693         Inf
## sample estimates:
##  mean of x  mean of y
```

```
## 0.08045979 0.07428127
```

Figure 7



Based on the output of our t-test, the p-value is very large at .276 which is significantly larger than our significance level at .05. Based on this, we fail to reject the null hypothesis. The data suggests that there is not a difference in the average salary based on BB_per_SO between left-handed and right-handed hitters. In Figure 7, we can see in the side-by-side boxplots above that the mean salary over BB_per_SO is nearly identical for both, confirming our hypothesis test.

Conclusion

Based on our analysis, we have gained valuable insights into the factors influencing Major League Baseball (MLB) player salaries. Our findings shed light on the complex interplay between player performance metrics, personal attributes, and compensation.

Firstly, through our linear regression analysis, we identified MLS, BB_per_SO, Age, and RBI as the most significant predictors of player salaries. This suggests that both experience and offensive contribution play crucial roles in determining player compensation. Interestingly, while age exhibited a negative correlation with salary, indicating that younger players tend to command higher salaries, RBI and BB_per_SO showed a positive correlation, implying that players who drive in more runs, strike out less, and hit more home runs are rewarded with higher salaries in the subsequent season.

Furthermore, our hypothesis test comparing the salaries of left-handed and right-handed hitters yielded intriguing results. Despite the common perception that left-handed hitters are more valuable, our analysis found no significant difference in average salaries based on BB_per_SO between the two groups. This

challenges conventional wisdom and suggests that, from a financial standpoint, both left-handed and right-handed hitters are equally valued in the MLB.

However, we do acknowledge several weaknesses in this test. Firstly, there is a limitation in the available data. While the Lahman Baseball Dataset offers valuable hitting statistics for our analysis, MLB front offices have access to a much broader array of data. Their datasets encompass a multitude of statistics beyond what we can access, which may explain why our R-squared values are lower than optimal. Additionally, our analysis is confined to players' data from the 2022 season for predicting their 2023 free agency deals. In reality, front offices consider a player's entire career, though they undoubtedly place significant emphasis on the most recent season, such as 2022. Future research could benefit from considering players' entire careers and exploring more granular data to gain a deeper understanding of the factors influencing MLB player salaries.

In conclusion, our project contributes to the ongoing conversation surrounding MLB player compensation by providing valuable insights into the relationship between player performance metrics and salaries. By uncovering the key factors driving player compensation, our analysis offers valuable implications for players, agents, team managers, and MLB stakeholders navigating the complex landscape of professional baseball.