**MLB Salary vs Performance Report**

Jacob Nevarez, Jason Shanker, Dinel Alic, Dylan Leigh, Sam Powell

**Overview**

Gerit Cole recently signed a contract to make $324 million over the next nine seasons from the New York Yankees. This contract comes one offseason after Bryce Harper and Manny Machado also signed contracts worth over $300 million while Mike Trout signed an extension worth over $400 million. MLB contracts are especially large for two reasons: it is an un-salary-capped sport where general managers are free to spend as much as they see fit, and a 162-game season makes for some very lucrative TV contracts. Nonetheless, the numbers on these contracts are staggering and fans and MLB execs alike can’t help but wonder: do the players live up to these enormous salaries in terms of on field performance? Are they really worth that much more than the average players in the league?

This analysis will utilize statistics and salary figures of MLB players to find to what extent an above average contract is fulfilled with above average performance, or vice versa.

**Data Preparation**

Our project consists of two different data sets from two different locations, one regarding the salary of MLB hitters and pitchers from Spotrac.com of all 30 MLB Teams (need to switch different team by clicking team tab and clicking go) ,<https://www.spotrac.com/mlb/boston-red-sox/payroll/2019/> which we web scraped and the other is about performance metrics of hitters and pitchers from MLB.com which we pasted into two different CSV files. The CSV files are called MLB.com\_Pitching\_data and MLB.com\_Hitting\_data.

2.1 MLB Salary

To measure the monetary value of MLB Players, we web scraped Spotac.com’s MLB payroll website. The code in Rstudio consisted of a loop that navigated to each team’s individual page to retrieve MLB player’s name, position, and salary. The salary consists of base salary, incentives, and a signing bonus.

The data set had the salaries of both pitchers and hitters, but we separated them since our analysis uses a mean salary for the positions to determine if a player is considered above/below average. Using a subset function, we categorized hitters as: 1B, 2B, 3B, C, CF, LF, RF, SS, OF, INF, and DH, (MLB\_Spotrac\_df\_hitting). For pitchers, we subsetted the position SP to its own data frame MLB\_Spotrac\_df\_pitching).

2.2 MLB Hitting Data

To measure the performance value of the MLB hitters, we used an external data source from MLB.com that retrieved performance metrics such as, rank, runs, hits, home runs, RBIs, batting average, on-base percentage, slugging, and on base plus slugging.

2.3 MLB Pitching

Like with MLB hitting data, we retrieved performance metrics on MLB.com, but the metrics differ since hitters and pitchers are evaluated differently. The metrics we pulled are, rank, wins, loses, era, games, games started, innings pitched, hits, home runs allowed, bases on balls, strikeouts, and opponents’ batting average.

2.4 Data Cleaning and Integration

Since both the performance data sets come from the same source, MLB.com, they are formatted the same. Therefore, cleaned the same as each other. This produce included separating the first and last name from each player column and normalizing the column fields (Player\_Last\_Name to Lastname\_plater) to merge smoothly with our Spotrac dataset.

Our integration created two different data tables, one for hitters and pitchers with their salary; we merged the data with an inner join based on the last name of the player and first initial. After doing this we did have some duplicates in our data. For example, Miguel Cabrera and Melky Cabrera have the same last name and first initial so it would create duplicates by storing statics in both when it only belonged to one player. However, we felt this wouldn’t impact our results that much. We could also have merged by positions for hitters to distinguish which stats went with what players however, the players positions for hitters were different for some players in spotrac than they were in MLB.com. For example, Jose Abreu plays 1st base sometimes but, mostly DH’s. On MLB.com Abreu position is listed as 1st base where’s in Spotrac it was listed as DH. So that wouldn’t have worked merging by position as well because than the salary amount would have been NA from spotrac for where those positions didn’t matchup with MLB.com Positions.

Data dictionary for Hitters Dataset

|  |  |  |
| --- | --- | --- |
| Column | Type | Description |
| Lastname\_Player | cha | Players Last Name |
| FirstInital | cha | Players first name initial |
| Rk | int | Rank among other players for same POS |
| Team | chr | Team the player plays for |
| POS | chr | Position the player plays |
| R | int | Integer of runs scored |
| H | int | Integer of hits |
| HR | int | Integer of HR |
| RBI | num | Number of runs batted in |
| AVG | num | Batting average |
| OBP | num | On base percentage |
| SLG | num | Slugging: Bases covered over total at bats |
| OPS | num | On base Plus Slugging: OPS plus Slugging |
| Players | chr | Players first and last name |
| Age | chr | Players age |
| Salary | num | Recent Salary contract |
| Salary Classification | chr | High or low salary based on average salary threshold we created |
| AVG Classification | Chr | High batting average or low batting average based on average threshold we created. |

Data dictionary for Pitchers Dataset

|  |  |  |
| --- | --- | --- |
| Column | Type | Description |
| Lastname\_Player | Chr | Pitchers Last Name |
| RK | int | Pitchers ranking |
| Team | chr | Team Pitchers Plays for |
| W | int | Amount of wins Pitcher has in season |
| L | int | Amount of losses Pitcher has in Season |
| ERA | num | Earned runs allowed |
| G | int | Games Pitched |
| IP | num | Innings Pitched |
| H | int | Hits allowed |
| R | int | Runs allowed |

|  |  |  |
| --- | --- | --- |
| ER | int | Earned Runs allowed |
| HR | int | Homeruns allowed |
| BB | int | Walks allowed |
| SO | int | Amount of strikeouts |
| Players | chr | Players first and last name |
| Age | chr | Pitchers age |
| Position | chr | Position |
| Salary | int | Salary amount |
| Salary\_Classification | chr | Low or high salary based on average salary threshold for pitchers. |
| Salary\_Classification2 | num | Salary as Binary 0 = low salary 1=high salary |

|  |  |  |
| --- | --- | --- |
| Firstname\_Player | chr | Pitchers first name |
| ERAClassification | chr | High or low ERA |

**Analysis**

The goal of this analysis is to determine whether there is a significant relationship between a player’s salary and their on-field performance: more specifically, do players who make an above average salary also perform above the average in their meaningful statistics?

*1. Hitters*

Who are the that are hitters paid above average? Below average?

We calculated the mean salary from the salary column in our dataframe for hitters: $3,638,907.

Then we calculated the average batting average from our list of hitters and found 0.2725926.

New columns were then added to the dataframe to denote whether the player was above or below each of these averages, seen below as “High Salary/Low Salary” and “High Average/Low Average”.

A screen shot of a computer

Description automatically generated

Now we can compare the two:

It can already be seen in these prop tables that there likely is not a significant correlation between salary level and average level, as they are relatively evenly distributed. However, high average/high salary has the largest distribution as we would expect, so chi squared test is performed. A screenshot of a cell phone

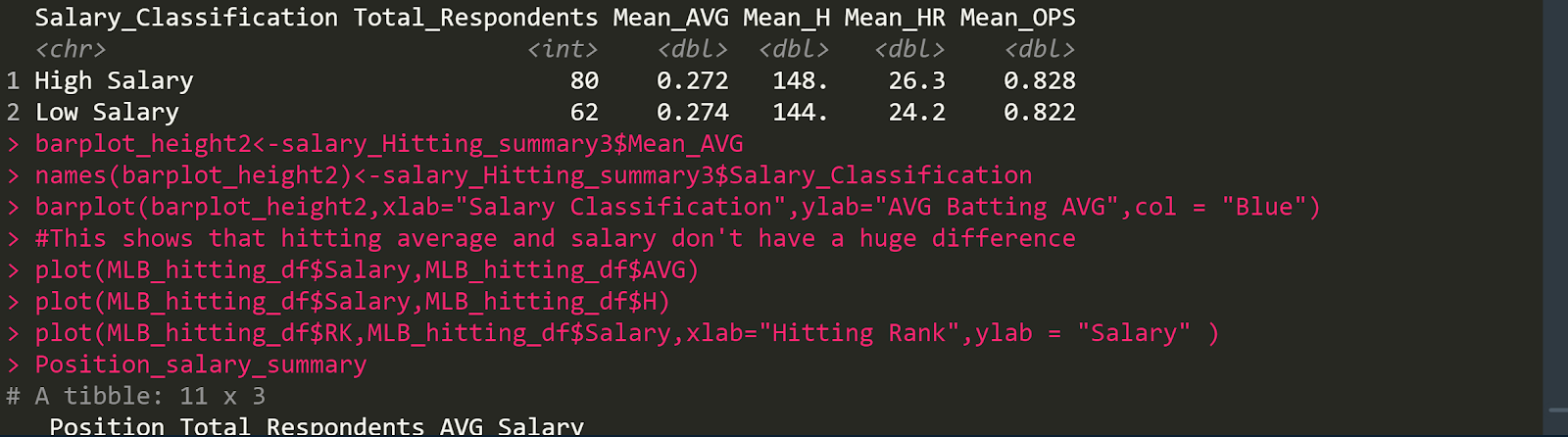
Description automatically generatedA screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

We get a very large p-value, which confirms our hypothesis after seeing the prop table, and we cannot conclude that salary level and batting average level are related. To further investigate we created a summary table for other batting statistics based on salary level:



A screenshot of a cell phone

Description automatically generated

The summary table shows that, with roughly similar sample sizes, the low salary players in the data set actually have a higher mean batting average than the high salary players do, and have very comparable averages for mean hits, home runs, and OPS. The scatter plot is a good visualization to see that hitters, ranked by batting average, fall all across the board when compared with salary, in fact the number one ranked hitter is almost near the bottom of the salary scale.

Some additional insight we gained from the data is that designated hitters have the highest average salary, followed by first baseman, and right fielders. They earn on average $16.6 million, $10.3 million, and $9.3 million per season respectively. When comparing the positions by hitting statistics, we see these salary distributions are reasonable as DH, 1B, and RF are near the top of every statistic, although we only have one player listed as DH in our dataset.

**A screenshot of a cell phone

Description automatically generated**

*2. Pitchers*

The same analysis was used for pitchers.

The average salary for pitchers in our dataset is $5,386,657 and the average ERA was found to be 3.88. Columns were added to the pitching dataset to denote salary level and ERA level as low and high.

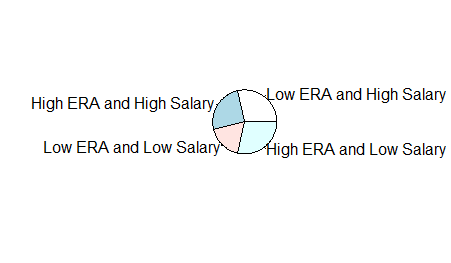
A screenshot of a computer screen

Description automatically generated

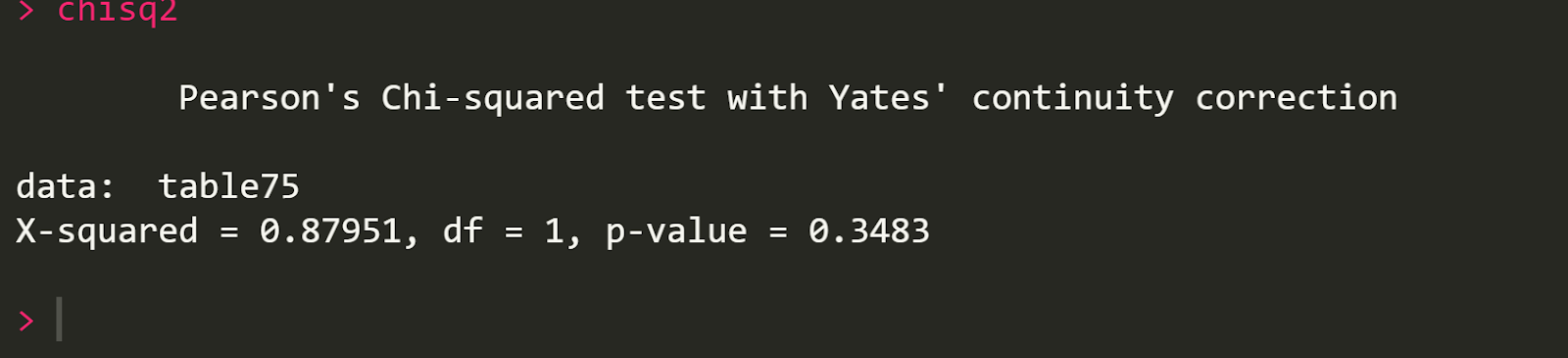
It is important to note that in baseball, high ERA is undesirable while low ERA is the desirable option and should theoretically be accompanied by a high salary. In the table below, as expected we see that the two highest distributions are high ERA/low salary and low ERA/high salary; however, there is nearly an equal number of pitchers who have a high ERA/high salary, which is contradictory to the intuition. The pie chart reinforces this visually.

**A screenshot of a cell phone

Description automatically generated**



We perform a chi square test to determine if there is a relationship between salary and ERA, and the test again gives us a value far larger than .05 and thus we cannot determine that there is a relationship between salary level and ERA.

****

We do some further analysis to see that the high salary group did outperform the low salary group in every statistic: ERA, Wins, Losses, innings pitches, hits and walks (fewer walks and hits is better for pitchers). And when we analyze the scatter plot with salary and pitchers ranked by ERA we see that the highest paid pitchers do fall in the general area that we expect them too.

**A close up of text on a black background

Description automatically generated**

**A screenshot of a cell phone

Description automatically generated**

**A close up of a piece of paper

Description automatically generated**

**Conclusion**

This data contains player statistics and salaries scraped from spotrac.com/mlb. We attempted to use this information to answer whether players who make larger salaries also perform better on the field. We performed chi square tests to investigate whether hitters making above the league average in salary also hit above the league batting average, and whether pitchers making an above average salary had an ERA below the average for pitchers included in the dataset. The analysis provided mixed results: both chi square tests failed, and we could not determine that salary was related to either ERA or batting average.

Further analysis of the hitters did show, however, that the three highest paid positions on average also performed the best in the hitting statistics. This finding gives hope that there is still meaning to be found in the data with a little tweaking.

Further analysis of the pitchers showed that many of the highest paid pitchers do have low ERAs and the high salary class performed better than the low salary class in every pitching statistic. The pitcher’s analysis provided findings more synonymous with what we expected going in, and there is strong reason to believe a better designed test would provide statistically meaningful results.

The main limitations of this analysis:

* The statistics we used for this analysis are too narrow to give a true idea whether a player’s performance is synonymous with their salary. For example, a player can have a batting average of only .230 (18 points below league average in 2018) but he can hit 35 home runs (almost double the league average of 18 in 2018) and clearly the amount of home runs he hit is much more valuable than if he had merely hit above the league batting average.
* The non-salary-capped nature of the MLB means that not all salaries are created equal. In the NFL, for example, all teams have a maximum amount of money to spend on contracts, which means that all salaries in the league exist on a scale of 0 to the salary cap. In the MLB, the teams in the largest markets have larger TV contracts and more money to spend. Teams like the New York Yankees can afford to pay 8 different players over $10 million in a season while a small market team like the Kansas City Royals can only afford to pay a few players that much money, last year they had 4. The implication is that a player is essentially paid by what they’re worth to their specific team and not necessarily their value compared to the rest of the league.

To better design this test we would create a dataset of all players in the league, rescale all of the salaries with the lowest and highest as 0 and 1, and compare salaries to [a more broadly meaningful statistic](https://library.fangraphs.com/misc/war/) like WAR or WAR+. Another way we could have calculated player performance to salary could have been by creating a column where we created our own made up performance metric. Example, for hitters we could have taken instead of just batting average hits, runs scored, RBI’s, Home runs, and walks and made a weighted average out of this by combing those statics together and then use this metric we created to compare to the players’ salaries. We could have also done this with our pitchers’ data where we combined a pitcher’s number of runs given up, walks, homeruns and so forth together. This could have helped us get better results instead of using one performance metric to compare to players salary like we did in our project.

Our Final datasets are called MLB\_pitching\_Final\_Data2.csv and MLB\_hitting\_Final\_Data.csv