Benjamin Schieber
Data Wrangling Final Project Report
Professor Redmond
5/9/2025

Evaluating MLB Player's Value: 2024 Offensive Performance vs Salary

1. Introduction

Major League Baseball(MLB) is one of the most statistically driven sports in the world. Over the course of the long 162 game season, there are mountains of statistics ready to be analyzed. Nearly every moment is captured in a statistic, which makes baseball an ideal domain for data-driven analysis, especially when it comes to player value and performance.

In recent years, fans and front offices alike have increasingly turned to advanced metrics such as OPS (On-base Plus Slugging) to assess a player's contribution to their team. However, the relationship between these performance metrics and a player's salary is often less transparent. With some players earning tens of millions of dollars each year, MLB front offices are under tons of pressure to give their players accurate and fair salaries that reflect performance. Do higher paid players really produce the best hitting stats like they should? Are there younger players who have more team friendly contracts outperforming older, higher paid veterans? These are the types of questions that inspired me to do this project.

This project explores these questions in the context of the 2024 MLB season. By examining offensive performance across players of different ages and comparing results to salary levels, we can gain a clearer understanding of how value is distributed—and possibly identify inefficiencies in how teams pay for performance. The findings have implications not only for front office decision-making but also for fans and analysts interested in how modern baseball balances talent and cost.

2. Data

This project uses two datasets. One includes hitting statistics for each player from baseball-reference.com, and player salary data from a public kaggle dataset which includes salaries from 2011-2024.

2.1 Batting Statistics

Batting statistics for the 2024 MLB season were collected from baseball-refrence.com, a trusted database for current and historical stats. Using selenium web scraper, a table of standard batting statistics were scraped for all qualified MLB players. The raw table included over 300 player rows and 33 columns of different statistics. The table includes key offensive performance metrics, such as games played, at bats, hits, home runs, walks, strikeouts, on base percentage (OBP), slugging percentage (SLG), and on base plus slugging.

Players who had zero plate appearances were removed from my final dataframe because they had no valid stats relating to this project. Only columns relevant to offensive output were retained for further analysis. All numeric columns were converted from a string to float, and player names were cleaned to remove asterisks and whitespace.

2.2 Salary Data

The salary dataset was downloaded from Kaggle and includes player salaries from 2011-2024. Each row includes the players name, salary, team, and season year. For this project, the only data used was from the 2024 season to match with the batting statistics data.

The “Salary” column originally contained dollar signs and commas (e.g., "$2,000,000") and was cleaned using regex replacement to convert values into numerical float types. Player names were also trimmed to enable merging with batting statistics data.

2.3 Merging and Final Dataset Construction

The two datasets were merged using both player name and team abbreviation as join keys. Only players who appeared in both datasets and had at least one plate appearance in 2024 were included in the final analysis.

https://www.baseball-reference.com/leagues/majors/2024-standard-batting.shtml
https://www.kaggle.com/datasets/christophertreasure/mlb-player-salaries-2011-2024

The final DataFrame, analysis_df, contains:

- 319 players
- 12 columns
- Offensive stats, age, team, and salary data from the 2024 season

2.4 Data Dictionary

Column	Type	Description
Player	string	Full name of the player
Age	float	Player's age during the 2024 season
Team	string	3-letter abbreviation of the MLB team (e.g., NYY, LAD, ATL)
HR	float	Home runs hit by the player
BB	float	Walks (bases on balls) drawn by the player
SO	float	Number of times the player struck out
BA	float	Batting average (hits divided by at-bats)
OBP	float	On-base percentage — how often the player reached base
SLG	float	Slugging percentage — total bases per at-bat
OPS	float	On-base Plus Slugging (OBP + SLG) — a key measure of total offensive output
RBI	float	Runs batted in — how often the player drove in runs
Salary	float	Player's actual 2024 salary in U.S. dollars

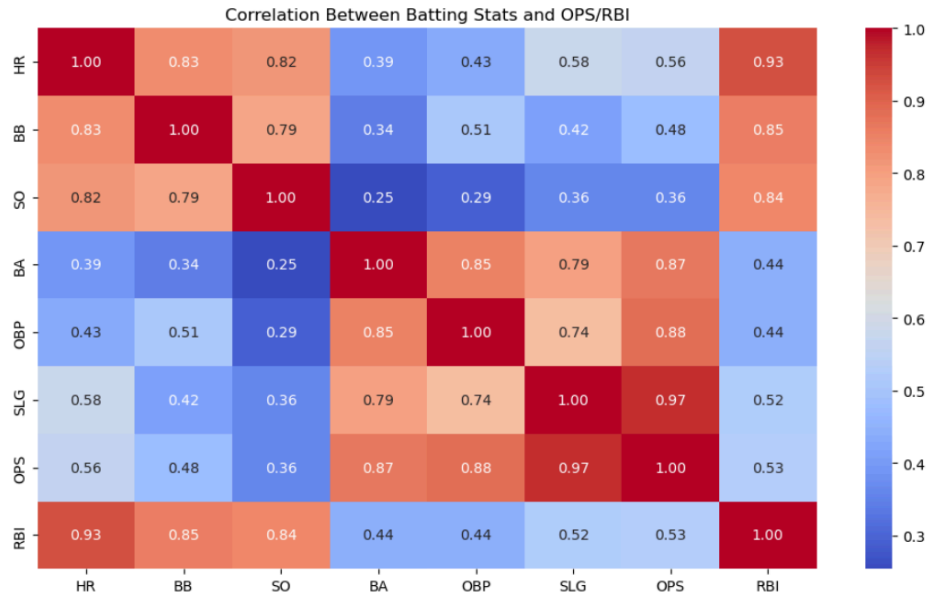
3. Analysis

3.1 Predictive Batting Statistics for OPS and RBI

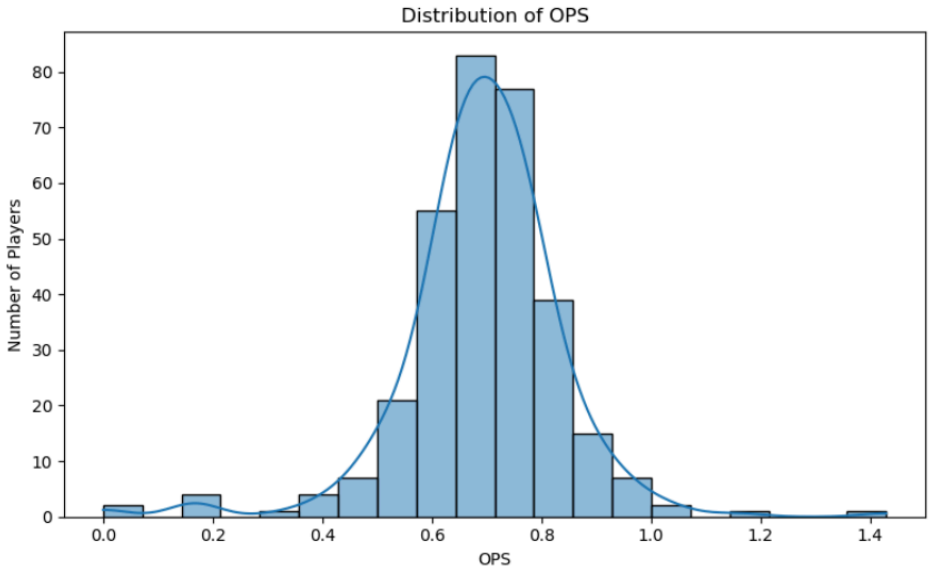
To start, I wanted to see how key batting statistics correlate with other hitting outcomes. The correlation heatmap shows that OPS is strongly driven by SLG (0.97) and OBP (0.88), which is expected since OPS is the sum of those two metrics. However, the biggest contributor to RBI were HR (0.93), BB (0.850, and SO (0.84). This suggests that players who hit for power and

https://www.baseball-reference.com/leagues/majors/2024-standard-batting.shtml
https://www.kaggle.com/datasets/christophertreasure/mlb-player-salaries-2011-2024

those who are disciplined at the plate to get walked tend to drive in more runs, even if strikeouts are high.



The univariate analysis below shows OPS is normally distributed, centered around 0.7 average for the league. This gives us a good baseline for interpreting performance.

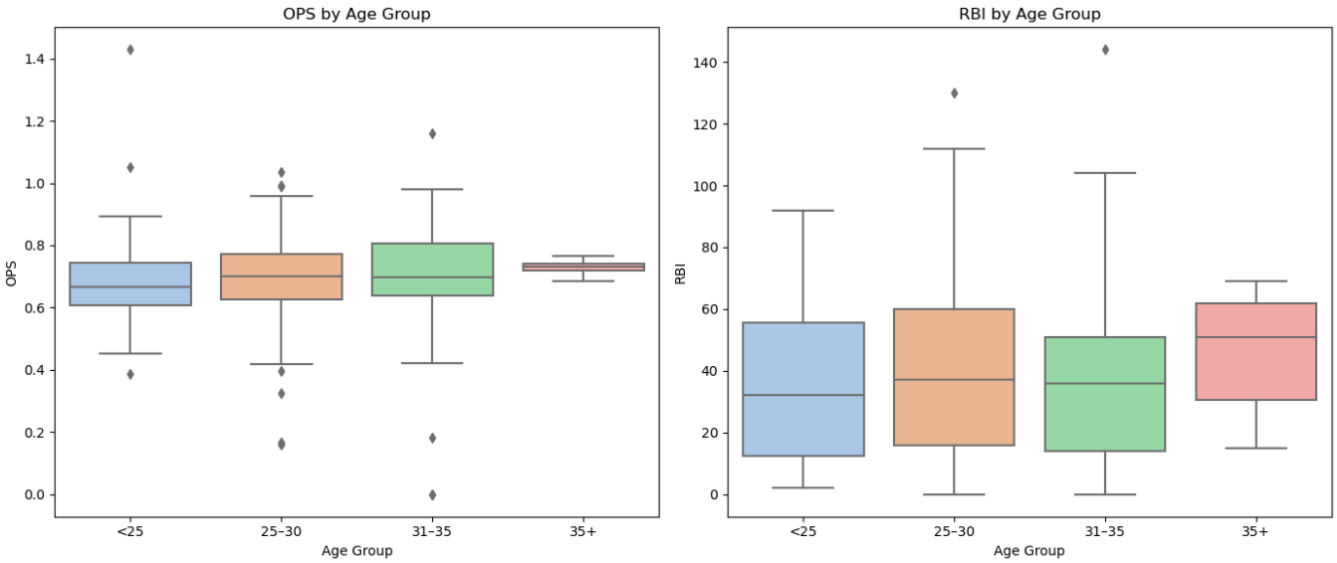


Next, I built a linear regression model using HR,BB, SO and SLG to predict RBI. The model gave an R^2 of 0.88, meaning these stats explain 88% of the variance in RBI across players, further confirming the predictive strength of plate discipline and power.

https://www.baseball-reference.com/leagues/majors/2024-standard-batting.shtml
https://www.kaggle.com/datasets/christophertreasure/mlb-player-salaries-2011-2024

3.2 Age and Offensive Performance

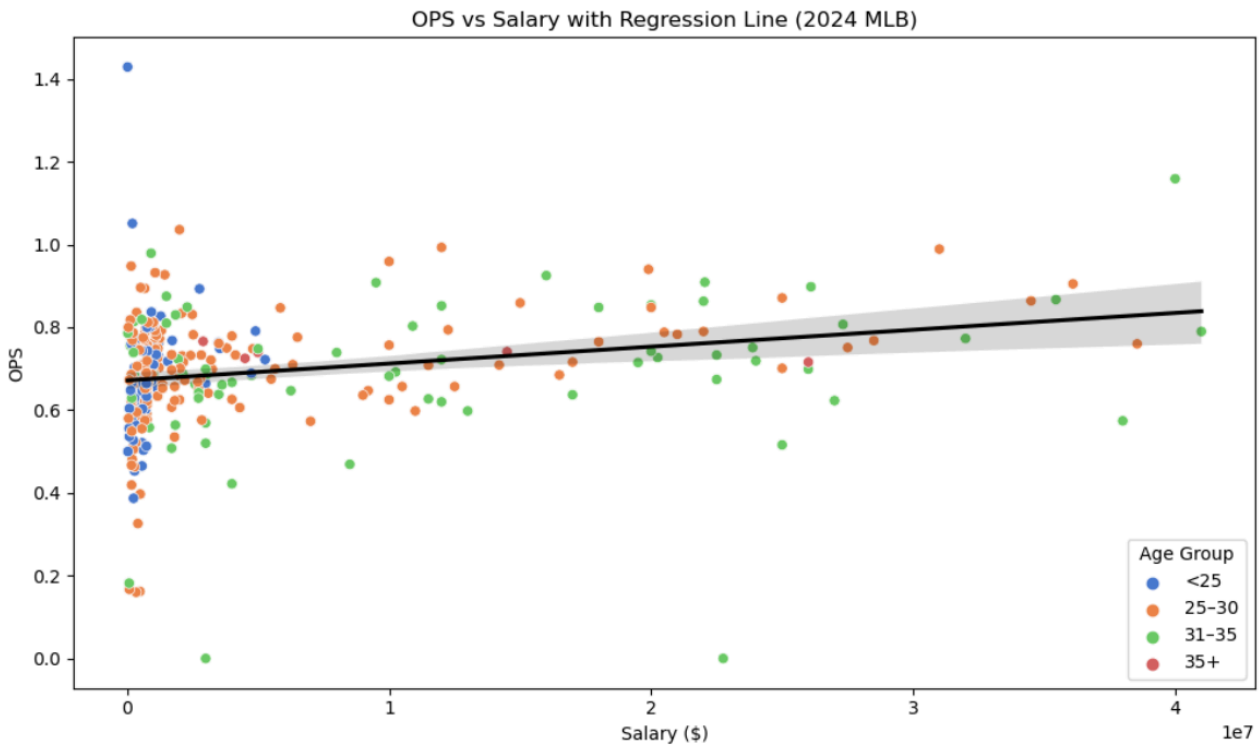
To see how age affects offensive performance, I grouped players into four different age bins: <25, 25-30, 31-35, and >35. The boxplots below of OPS and RBI by age group show a fairly even distribution of performance, the 35+ bins have the highest OPS but there are much less players in this group. 25-30 year olds having a slightly higher OPS average than the other two age age groups.. The spread within each group is wide and outliers exist in all age bins except 35+. These findings suggest that the best hitters in the MLB can be any age.



To further test the statistical significance of these findings, I ran a two sample t-test comparing OPS between players under 30 and over 30 years old. The test resulted in a t-statistic of -1.03 and a p-value of 0.3, indicating there is no statistically significant difference in OPS between the two groups. This means that age alone is not a reliable predictor of offensive performance. Conventionally, it's thought that younger players in their prime will produce much better than their older counterparts, but veterans can still perform at just as a high level, so experience can lead to better results even if their bodies are not as powerful as they used to be.

3.3 Salary vs. Performance

Lastly, I analyzed the relationship between offensive performance and compensation. I created a scatter plot comparing OPS to salary, color coded by age group, as well as a regression line.



This visual above shows us two key findings. First, the regression line shows a weak upward trend, showing that players with a higher OPS tend to earn slightly more, but the relationship is not linear or consistent. The R^2 from a corresponding linear regression model is only 0.36, meaning just 36% of salary variation can be explained by offensive stats like OPS, HR, and RBI. This means there are also many other factors that influence player salaries, like contract timing, defensive skill, the team's payroll constraints, player popularity, or past accomplishments.

Secondly, the visual above helps identify underpaid and overpaid players. Players below the regression line have lower OPS than expected for their salary, suggesting they may have been overpaid relative to their 2024 performance. Players above the line outperformed what their salary predicted, suggesting they may have been underpaid for the 2024 season. There is a dense cluster of low salary players with widely varying OPS's, many younger players still on their rookie contracts. High salary players tend to hover near the line, with many players aged 31-35 making more than $10 million falling under the line, highlighting inefficiencies in salary to offensive output. Past performance and long guaranteed contracts may lead to higher salaries even when the player's stats don't back it up.

These patterns were confirmed by a regression model I ran to analyze players' value gap to estimate what a player should be earning based on offensive output. Younger players like Shohei Ohtani, Brent Rooker, and Gunnar Henderson are among the most underpaid players in the MLB with salaries predicted between $13-18 million, but actually earning under $3 million. This further confirms that younger players on team friendly contracts provide great benefit to their teams.

Most Underpaid Players:

	Player	Team	Salary	PredictedSalary	SalaryDifference
1	Shohei Ohtani	LAD	2000000.0	1.849398e+07	-1.649398e+07
43	Brent Rooker	OAK	1437804.0	1.436387e+07	-1.292606e+07
2	Gunnar Henderson	BAL	2763378.0	1.316107e+07	-1.039769e+07
56	Jake Burger	MIA	760000.0	1.079873e+07	-1.003873e+07
77	Shea Langeliers	OAK	1000500.0	1.097600e+07	-9.975503e+06
112	Mark Vientos	NYM	927690.0	1.004376e+07	-9.116070e+06
21	Spencer Steer	CIN	750000.0	9.317355e+06	-8.567355e+06
52	Alec Burleson	STL	750050.0	9.002586e+06	-8.252536e+06
49	Zach Neto	LAA	1192551.0	9.409880e+06	-8.217329e+06
55	Riley Greene	DET	1284146.0	9.457135e+06	-8.172989e+06

Most Overpaid Players:

	Player	Team	Salary	PredictedSalary	SalaryDifference
214	Anthony Rendon	LAA	38000000.0	1.743356e+06	3.625664e+07
15	Jose Altuve	HOU	41000000.0	8.203914e+06	3.279609e+07
259	Mike Trout	LAA	35450000.0	3.785032e+06	3.166497e+07
70	Will Smith	LAD	38550000.0	8.621381e+06	2.992862e+07
155	Carlos Correa	MIN	36100000.0	6.224922e+06	2.987508e+07
252	Kris Bryant	COL	27000000.0	2.200984e+06	2.479902e+07
78	Corey Seager	TEX	34500000.0	1.084085e+07	2.365915e+07
109	Giancarlo Stanton	NYY	32000000.0	1.015526e+07	2.184474e+07
317	Luis Castillo	SEA	22750000.0	1.863245e+06	2.088676e+07
193	Javier Báez	DET	25000000.0	4.116971e+06	2.088303e+07

Players like Anthony Rendon, Jose Altuve, and Mike Trout are making $30 million plus, but their offensive production predicts values under $4 million. This is because they all signed very long legacy contracts after peak seasons when they were in their primes. MLB teams could benefit from more shorter, production based contracts that reflect actual performance, instead of past success.

4. Conclusion

In this project, I analyzed three key questions related to MLB hitting performance and player salary value during the 2024 season. The study combined a scraped table on hitting data and player salary records to explore how production, age, and salary interact with one another. In summary, below are the findings of each question in my proposal:

1. *Which batting statistics are most predictive of OPS and RBI?*
   There is a very strong relationship between OPS and both OBP and SLG, which is expected since OPS is a sum of both statistics. When predicting RBI, HR, BB, and SLG had the strongest correlation, with a multiple linear regression model achieving an R^2 score of 0.88. This shows that players who hit for power and the ability to get on base are the best indicators of overall run production

2. *How does age impact offensive performance for the 2024 season?*
   Younger players are expected to be more productive, but the data shows that OPS and RBI are evenly distributed across age groups. A two sample t-test comparing OPS between players under and over 30 resulted in a p-value of 0.3, showing no statistical significance. Boxplots also show performance does tend to peak in the 25-30 age range, older players can still produce at a high level.

3. *Are higher paid players producing better results than lower paid players?*
   Results show that salary is weakly correlated with offensive production. A regression model using OPS, HR, and RBI to predict salary resulted in an R^2 score of 0.36, indicating that the majority of salary variation is not explained by offensive performance. When I compared predicted vs. actual salaries, many players stood out as highly underpaid(Shohei Ohtani) or overpaid(Anhony Rendon), suggesting that contracts are based off many other factors than just recent offensive performance alone.

The project has several limitations, including it only used single season stats, which may not accurately reflect trends. Some players simply have down years. Also, there are many market fluxes and contract timing that determine player salary like the status of the free agent market, or the budget of the team. Lastly, the regression model did not account for non-performance factors like player popularity they bring to the team, or the veteran presence older players can provide for a team. I think for future improvements to this project I would use more years to evaluate player trends and consistency over a longer period of time, and try to find data includes other metrics that are not only offensive because some players are great at defense, but not so good at hitting, so they are still being paid high salaries with below average offensive performance.