WAR Dollars: Winning the Budget Game in Baseball Management

Graeme Ashley, Matthew Kassiman, Jose Smith, Daniel Wong Ortiz & Yusef Trad

Professor Lorena Martin

February 22, 2024

1. Abstract

This study delves into the intricate relationship between Major League Baseball (MLB) player salaries and Wins Above Replacement (WAR) during the period spanning 2019 through 2022. A particular emphasis is placed on discerning the distinctions between position players and pitchers concerning their respective WAR, as defined by FanGraphs, and salaries. The primary objective is to determine whether position players or pitchers offer superior value in terms of WAR per dollar spent, providing valuable insights for teams aiming to optimize their payroll allocation for the maximization of both WAR and overall wins. Our hypotheses is as follows:

H_0: There is no difference in WAR between a pitcher and position player's based on salary

H_a: Pitchers are overpaid relative to their contributions to WAR where Position players are underpaid relative to their contributions to WAR

To execute this analysis, three datasets are imperative. Firstly, the Wins Above Replacement (WAR) data for each year from 2019 to 2022 was collected using the `pybaseball` Python package, which scrapes data from FanGraphs. Secondly, a player positions dataset was obtained, containing a list of player names and their primary positions spanning the same time frame. This dataset provides valuable information on player positions throughout the years under consideration. Finally, an MLB player salary dataset spanning the years 2019 to 2022 was acquired from Spottrac, the largest online sports team and player contract resource for American sports. The integration of these datasets will be facilitated by matching the 'player' and 'year' fields, ensuring that all common rows are aligned within the shared timeframe. It's important to note that salary figures will not be adjusted for inflation since our dataset spans only from 2019

to 2022, which is relatively recent. Therefore, inflation adjustments are not necessary for this analysis, given the data's recency.

The core focus of our investigation is to determine if a difference exists in WAR, holding salary constant, between MLB position players and pitchers, and within position players which positions provide the most value to teams in terms of WAR. Ultimately, this research contributes to the understanding of resource allocation strategies within MLB teams, offering evidence-based recommendations for optimizing team performance and financial efficiency. The findings are anticipated to have implications for team management, player contract negotiations, and overall decision-making processes within the dynamic landscape of professional baseball.

2. Introduction

In the ever-evolving landscape of baseball analytics, one metric has risen to prominence as a measure of a player's true impact on the field—Wins Above Replacement (WAR). This comprehensive statistic represents a significant shift in the evaluation of player performance, transcending traditional metrics and providing a nuanced understanding of a player's overall contribution to their team (Slowinski). At its core, WAR seeks to break down a player's all-around skills into a single numerical value, reflecting the number of wins they contribute above a replacement-level player. The replacement level serves as a benchmark, representing the expected performance from a readily available minor league player. By considering numerous factors, including hitting, fielding, baserunning, and positional adjustments, WAR encapsulates the holistic nature of a player's impact.

The calculation of WAR involves sophisticated statistical modeling, leveraging advanced metrics to capture the nuanced aspects of a player's performance. Hitters are assessed on everything from their ability to hit for power, contact, speed, and more. Defensive contributions are evaluated through metrics gauging range, arm strength, and overall fielding prowess. Baserunning skills and positional adjustments further enrich the calculation, providing a more nuanced and accurate representation of a player's on-field contributions. Pitching WAR is calculated differently, but captures a player’s overall ability to lead their team to wins on the mound.

As the baseball landscape continues to shift towards data-driven decision-making, the significance of understanding WAR becomes increasingly apparent. This metric not only aids in player evaluation but also plays a crucial role in the strategic decisions made by front-office executives and general managers. The relationship between a player's WAR and their salary has become a focal point of analysis, as teams strive to build competitive rosters while managing financial constraints.

Traditionally, player salaries were determined by basic descriptive statistics, such as batting average, or counting statistics, like total doubles and home runs. However, the ascendancy of advanced analytics has ushered in a new era where WAR is considered a more accurate measure of a player's impact. The theory proposes that a player with a higher WAR should command a higher salary, reflective of their more significant contribution to their team's success. However, the negotiation process is complex, influenced by factors beyond on-field performance, including a player's age, injury history, market demand, and the financial landscape of the team.

The correlation between financial investments in free agency and on-field success is far from a guaranteed formula for victory. While it may seem intuitive that splurging on high-profile free agents should directly translate into more wins, the reality of the sport often defies such straightforward expectations. History is replete with instances where teams boasting the most significant free-agent signings have fallen short of expectations. The intricacies of baseball, from the role of pitching to defensive prowess and situational hitting, demand a comprehensive understanding of team dynamics beyond a simple financial or statistical metric.

This introduction sets the stage for a comprehensive exploration of Wins Above Replacement in baseball, delving into its intricate calculation, role in player evaluation, and the intriguing dynamics between WAR and player salaries. As we delve deeper into the world of baseball analytics, the interplay between on-field performance and the metrics that seek to quantify it will undoubtedly shape the strategies of teams striving for success in the modern era of the game.

3. Hypothesis

So far the literature has taken into account the seismic shift that WAR has had on baseball front offices and how the cost of a win can be quantified. However, what has not been identified is which position per dollar is worth more in team wins. We aim to address this gap in the literature with our research. As such, contrary to what would be traditional expectations, hitters, we believe by virtue of their consistent offensive contributions and broader impact on the game, will be valued more highly than pitchers in terms of WAR. Our hypothesis suggests that MLB front offices should invest more in position players as their cumulative offensive and defensive skills, both at the plate and in the field, will yield a more significant and consistent impact on overall team success per dollar adjusted for inflation compared to the singular, albeit crucial, role of pitchers.

4. Methods

We leveraged the pybaseball package to retrieve pertinent statistics for every player across the seasons spanning from 2019 to 2022. Notably, this package efficiently categorized baseball players into two main groups: pitchers and position players. Additionally, we employed a web scraping tool to compile player salary data for all MLB teams from 2019 to 2022, sourced from Spotrac, a leading online resource for team and player contract information in American sports. Lastly, we acquired player positional information from another module in the pybaseball package that grabs data from Statcast and merged it with our player statistics and salaries datasets.

Given the recency in our dataset, we elected not to adjust our salary data for inflation. This decision was based on the fact that the majority of players included in the dataset are still active and contributing to their respective teams, thus rendering inflation adjustments unnecessary at this juncture. By integrating information from these sources, we constructed a well-rounded perspective to aid our analysis of positional WAR per dollar spent and the implications it has on the MLB.

Initially, we generated histograms illustrating the distributions of WAR and salaries for pitchers and position players. This initial exploration aimed to provide a preliminary understanding of the relationship between these variables. Furthermore, we incorporated additional visualizations to scrutinize key performance indicators for any potential skewness within the data. To ensure the integrity of our dataset, we subsequently applied logarithmic transformations to the relevant variables, aiming to mitigate any biases and evaluate their impact on skewness. Subsequently, we partitioned the player salary data from the pitchers and positional player datasets for streamlined analysis. Additionally, we merged the player positions dataset with the positional player dataframe and encoded binary values to represent player positions. Our initial regression analysis was conducted separately for pitchers and positional players, focusing on crucial metrics such as R-squared, Durbin-Watson, Jarque-Bera values, coefficients, and p-values to assess model effectiveness. Furthermore, decision trees were employed, and k-fold cross-validation techniques were utilized to gauge model accuracy and identify significant factors contributing to a player's WAR. Given the extensive array of performance statistics across both datasets, we employed factor analysis to discern distinctive attributes within each player position.

Anticipating suboptimal results in our initial run, given the novelty of our model, we planned to conduct a thorough analysis to identify potential issues. Once we identified areas for improvement, we implemented various data transformations, as noted in our discussion section.

To address any multicollinearity, We conducted a Variance Inflation Factor (VIF) analysis to detect potential multicollinearity issues within the dataset, and found none. However, during our examination of the positional player regression, we observed heteroscedasticity. In an effort to address this issue, we attempted a logarithmic transformation of the WAR variable. Despite this transformation, we observed no significant impact on the variance in the data points. Consequently, we decided to retain the original regression model without transforming the dependent variable.Additional transformations under consideration include transforming significant KPIs that we thought were relevant in the regression, namely: Strikeout to Walk Ratio (K/BB) for pitchers and the Walk to Strikeout Ratio (BB/K) for batters.

Following numerous regressions which incorporated these transformations, we ran a final regression to assess the significance and effectiveness of R-squared, Durbin-Watson, and Jarque-Bera values, as well as the significance of coefficients. A final residual analysis was conducted to ensure compliance with assumptions of Linearity, Homoscedasticity, Normality, and Autocorrelation. To facilitate this analysis, we employed JMP software, which allowed for a diligent examination of Key Performance Indicators (KPIs) and the ability to amend multicollinearity issues.

Our model for batters considered various predictors including the logarithms of On-base Plus Slugging (OPS), the Walks to Strikeouts ratio, and salary, in addition to age and specific fielding positions (2B, 3B, CF, SS, and Catcher, with 1B, LF, RF and DH lumped into the intercept, as they were not statistically significant on their own). Each of these variables demonstrated significance at a 0.05 level, suggesting a noteworthy influence on a player's WAR above or below the reference point from the intercept. Notably, the model achieved an R-squared value of 0.55, indicating that approximately 55% of the variability in WAR can be explained by our selected variables.

Our findings verified previous insights, particularly highlighting the strategic value of investing in "up-the-middle" positions (CF, SS, C), which tend to yield higher WAR contributions at equivalent salary levels while avoiding non-premium defensive positions like first base, left field, and DH. This underscores the importance of defensive prowess alongside offensive capabilities and advises against overinvestment in the latter positions, as they often accommodate players with strong offensive skills but limited defensive contributions and therefore less total wins contributed to the team.

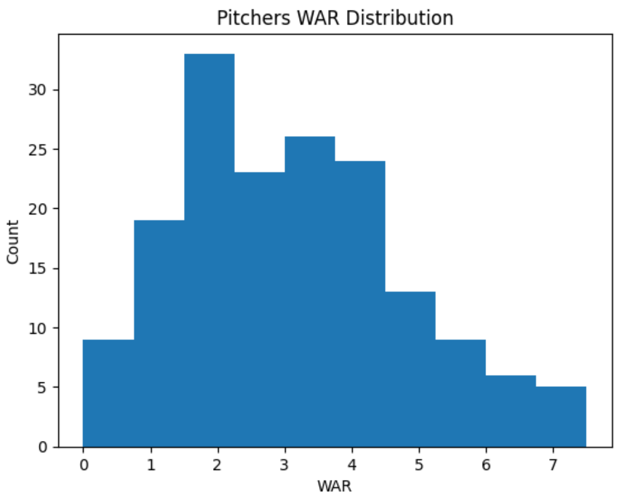
In the analysis of our pitchers' model, we observed a remarkably high R-squared value of 0.80, indicating that our model explains a substantial 80% of the variance in Wins Above Replacement (WAR) for pitchers. This high level of explanatory power is attributed to the more homogeneous nature of our dataset, which exclusively comprised starting pitchers. It is well-documented that proficient starting pitchers often command high salaries due to their significant impact on games, making them highly coveted in the free market. This trend was reflected in our findings, where both high salaries and high WAR values were characteristic of top-tier starting pitchers, highlighting their market value and the propensity of teams to invest heavily in this segment. Our findings also showed the significant influence of total innings pitched (IP) on a pitcher's WAR. This relationship is intuitive, given that WAR is an accumulative metric, and pitchers who log more innings have more opportunities to contribute positively to their team's success. This insight suggests that teams might benefit from targeting durable pitchers capable of "eating innings" in the open market, as their ability to shoulder a heavy workload can be invaluable over the course of a grueling season. Furthermore, the analysis highlighted the detrimental effect of a high Home Runs per Nine Innings (HR/9) ratio on a pitcher's WAR. This finding emphasizes the importance of a pitcher's ability to minimize home runs, as allowing fewer home runs is critical for attaining a high WAR. This insight can guide teams in scouting and developing pitchers with a propensity to keep the ball in the park, further enhancing their pitching staff's effectiveness.

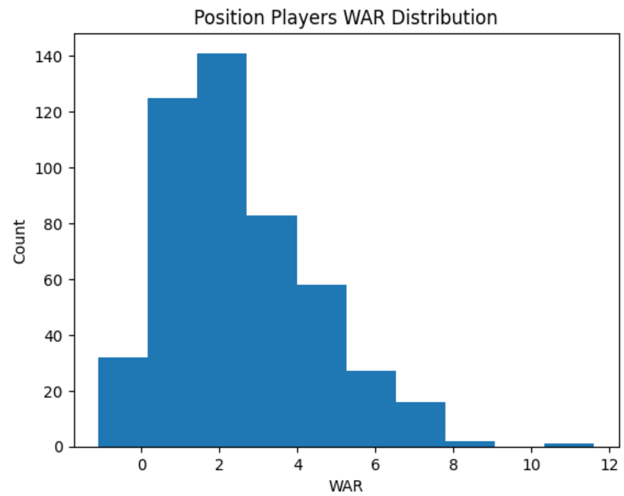
When examining the impact of salary on WAR, the coefficient for the natural logarithm of salary (ln(salary)) stood at 0.16 for pitchers, in stark contrast to 0.48 for hitters. This disparity suggests that while higher salaries are correlated with higher WAR values for both groups, the return on investment, in terms of WAR, is more pronounced for hitters than for pitchers. Consequently, when faced with the decision of allocating substantial contracts, teams might find greater value in investing in hitters, contingent on offseason needs. Some other notable differences in the batting and pitching models are that the analysis revealed a relatively modest decline in WAR with age for pitchers, at a rate of 0.04 per year. This gradual decrease contrasts with the steeper decline observed in hitters, supporting the notion that pitchers, particularly starters, may experience a slower rate of performance degradation. Notably, some pitchers do not reach their peak performance until their 30s, further illustrating the unique career trajectory of pitchers compared to position players. Furthermore, the significance of the salary variable in our model challenges the extreme "Moneyball" philosophy of minimizing player expenditure. It suggests that while seeking value is crucial, entirely forgoing higher salaries may not necessarily translate to substantial WAR gains. The model also emphasizes the value of a higher BB/K ratio, advocating for players who maintain discipline at the plate and minimize strikeouts—a trait that significantly contributes to team success and is often cited as one of the main metrics that scouts look for in collegiate and minor league players, since plate discipline tends to translate across competition levels.

Interestingly, some positions did not warrant statistically significant coefficients. We hypothesize that this could be due to positional fluidity, with some players playing multiple positions, thus muddying our coefficients for those particular positions. The decline in WAR with age, at a rate of 0.2 per year holding other stats constant, further informs contract negotiation strategies, cautioning against overvaluing players in the twilight of their careers.

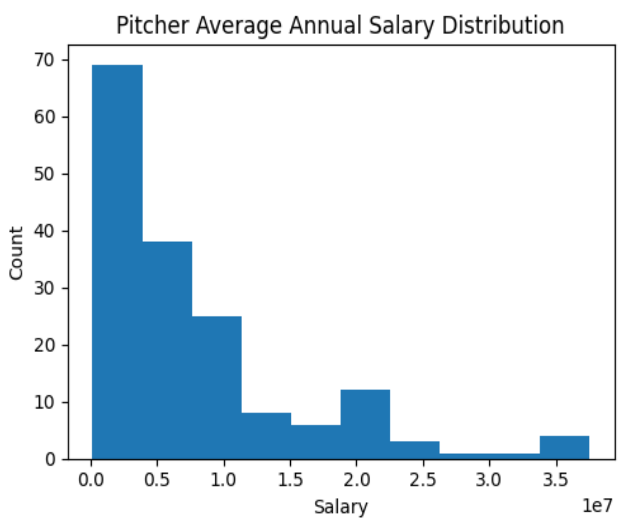
5. Results

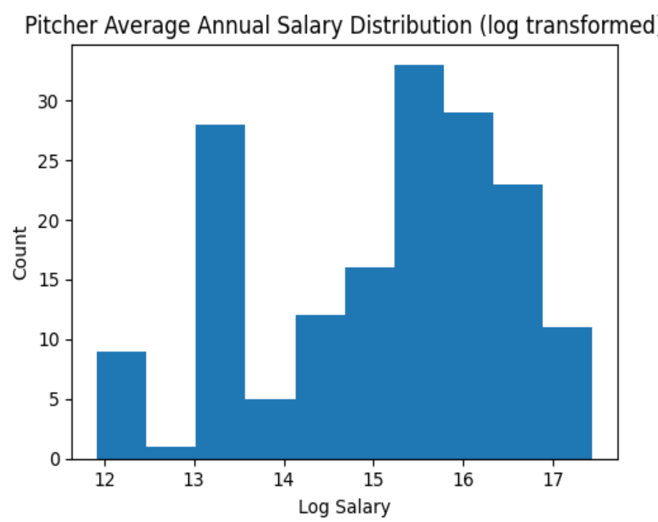
5.1 Histograms

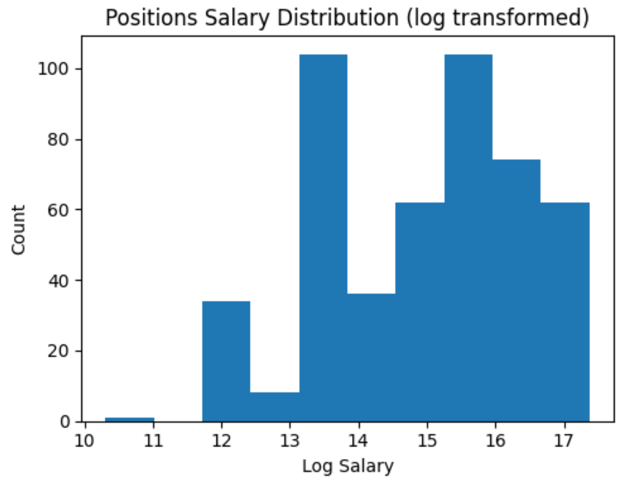




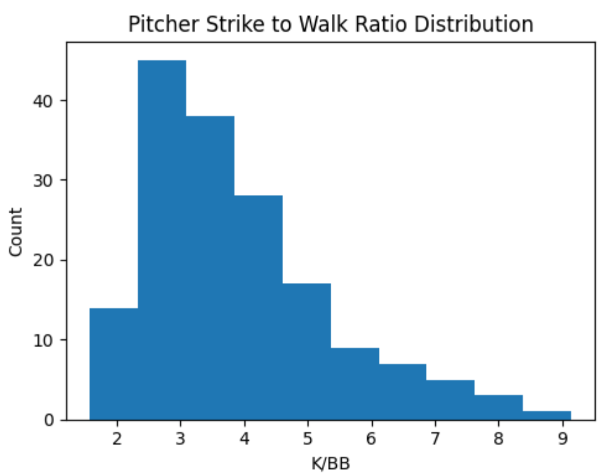
For WAR distribution the graphs suggest that pitchers and position players most frequently have a WAR between 0 and 4, with fewer players achieving higher WAR values

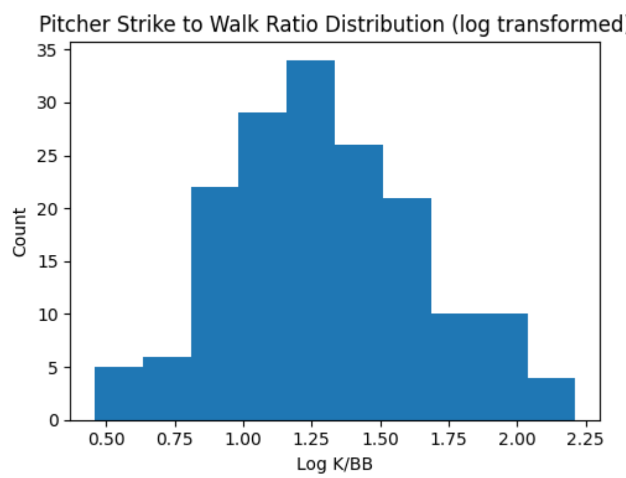




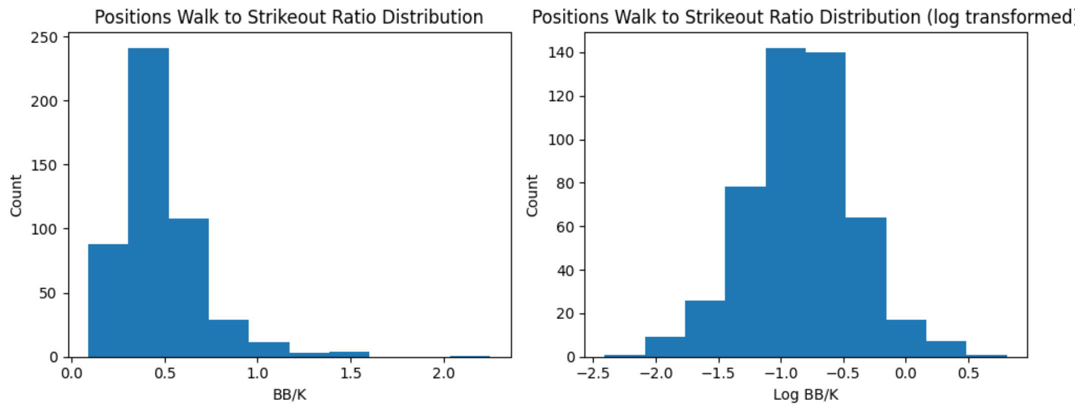


For salary distribution most pitchers and positions players fell into the lowest salary bin, with the frequency drastically declining as salary increased.





In our regression analysis, we included key performance indicators such as the strikeouts to walks ratio for pitchers and the walks to strikeouts for positional players. Upon plotting the distribution of these variables, we observed some skewness. To address this, we applied a logarithmic transformation to the variables. While the log transformations resulted in improvements for K/BB and BB/K, the distribution of innings pitched remained largely unchanged. Considering that our dataset primarily consisted of starting pitchers, where innings pitched typically exhibit minimal variance, we decided to retain the transformed variable.



5.2 Final Regression Results

5.2.1 - Pitchers

5.2.1.1 Pitchers Regression

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    WAR   R-squared:                       0.800
Model:                            OLS   Adj. R-squared:                  0.793
Method:                 Least Squares   F-statistic:                     106.9
Date:                Fri, 23 Feb 2024   Prob (F-statistic):           2.23e-53
Time:                        19:58:28   Log-Likelihood:                -187.05
No. Observations:                 167   AIC:                             388.1
Df Residuals:                     160   BIC:                             409.9
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -6.4737      0.970     -6.673      0.000      -8.390      -4.558
log_salary     0.1619      0.062      2.627      0.009       0.040       0.284
Age           -0.0474      0.021     -2.220      0.028      -0.090      -0.005
log_K/BB       1.7500      0.191      9.141      0.000       1.372       2.128
log_IP         2.0473      0.154     13.325      0.000       1.744       2.351
H/9           -0.2150      0.056     -3.829      0.000      -0.326      -0.104
HR/9          -2.0747      0.200    -10.382      0.000      -2.469      -1.680
==============================================================================
Omnibus:                        5.636   Durbin-Watson:                   1.227
Prob(Omnibus):                  0.060   Jarque-Bera (JB):                5.587
Skew:                           0.448   Prob(JB):                       0.0612
Kurtosis:                       2.991   Cond. No.                         578.
==============================================================================
```

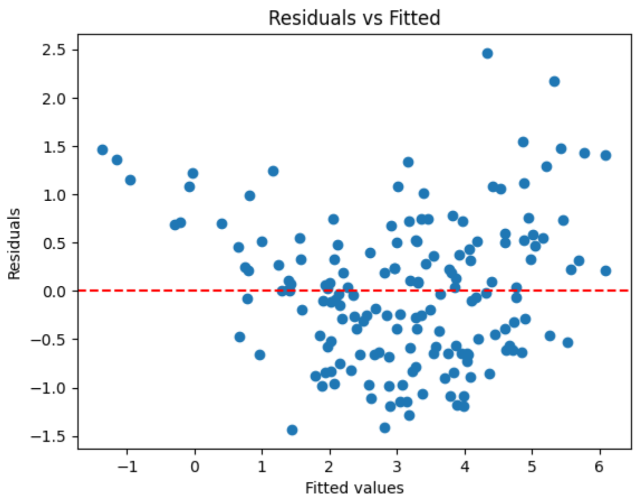
The regression analysis results reveal that the model has an R-squared value of 80%, suggesting that it can explain a substantial portion of the variability in WAR. Additionally, the F-statistic, with a very low probability, indicates the model's overall significance. We can deduce that log(salary) has a significant positive effect on WAR, and that a player's WAR decreases at a rate of 0.0474 per every year a player ages, holding other variables constant. Additionally, a player's ability to pitch lots of innings and not give up home runs has a large effect on a high WAR.

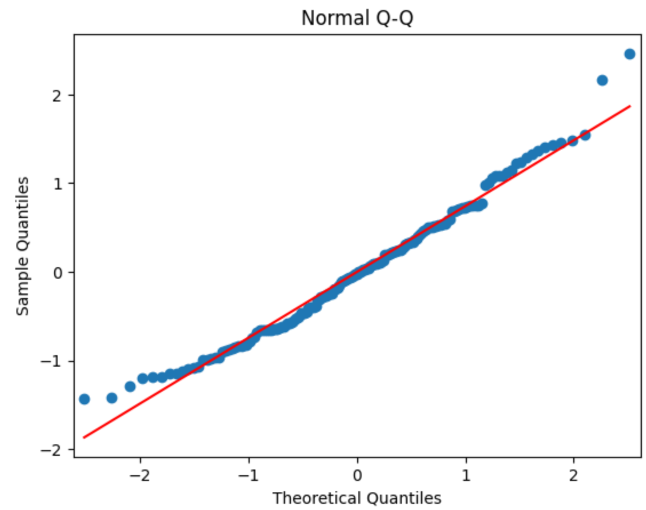
5.2.1.2 Pitchers Regression VIF

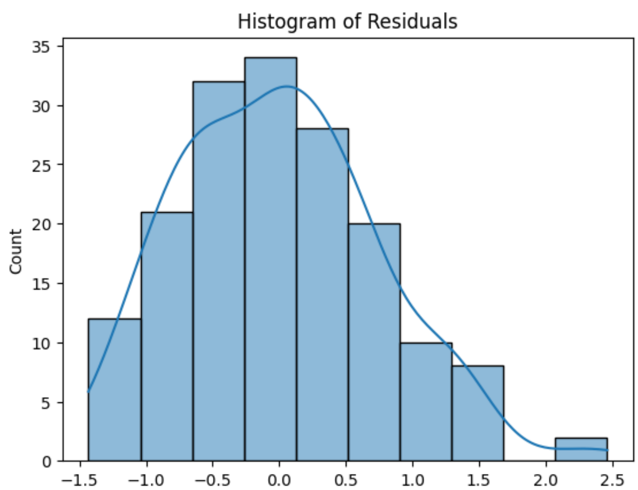
Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-6.473679	0.970101	-6.67	<.0001*	.
Log[salary]	0.1619327	0.061635	2.63	0.0094*	2.0797552
Age	-0.047405	0.021351	-2.22	0.0278*	1.8321918
Log[K/BB]	1.7500181	0.191439	9.14	<.0001*	1.3738722
Log[IP]	2.0472705	0.153641	13.33	<.0001*	1.1624603
H/9	-0.214983	0.056151	-3.83	0.0002*	1.4447906
HR/9	-2.074711	0.199836	-10.38	<.0001*	1.2014735

Based on the output above from JMP, there is no sign of multicollinearity between the independent variables.

5.2.1.3 Pitchers Regression Residuals







The Residuals vs Fitted plot exhibits a relatively random scatter of points, indicating consistent variance, or homoscedasticity, in the residuals. While there are some minor fluctuations in the scatter, overall, the data points achieve a satisfactory level of homoscedasticity. The Normal Q-Q plot compares the distribution of residuals to a normal distribution, with most points aligning closely with the red line. This alignment suggests that the residuals are approximately normally distributed, although there are some deviations in the tails. Furthermore, the Histogram of Residuals displays a bell-shaped distribution with a slight right skew, characteristic of a normal distribution. In summary, these plots suggest that the regression model meets several assumptions: the residuals are mostly normally distributed and exhibit consistent variance. However, there are indications of potential outliers and slight deviations from normality, particularly in the tails of the distribution.

5.2.1.4 Cross Validation

We conducted k-fold cross-validation by partitioning our dataset into training and testing subsets. We explored different training/testing split ratios, such as 75/25, 80/20, and 70/30, and assessed the model's performance across varying fold counts (5 and 10). Remarkably, the cross-validation scores exhibited consistency across these different configurations, indicating the stability of our model's performance. Specifically, the Mean Squared Error on the testing set remained consistent at 0.516, while the average cross-validation scores for Mean Squared Error (MSE) were 1.138 for 5 folds and 0.825 for 10 folds. These results underscore the reliability and generalizability of our model across different training/testing splits and fold counts.

5.2.1.5 Decision Trees

We conducted an in-depth analysis using decision tree regression and various testing/training splits to evaluate the effectiveness of our variables. Our regression models delivered promising results, as evidenced by a Mean Squared Error (MSE) of approximately 1.18, indicating a relatively low level of prediction error. Through hyperparameter tuning, we determined the optimal maximum depth of the decision tree to be 6, leading to a best score of around 0.602. Furthermore, both the Bagging ensemble model and the XGBoost regressor demonstrated impressive accuracy on the test data, achieving scores of about 0.802 and 0.798, respectively. Similarly, the Random Forest model achieved an accuracy of approximately 0.799

on the test data. These outcomes collectively suggest that our regression models perform admirably in predicting the dependent variable, with ensemble methods showcasing particularly robust performance on the test data.

Upon scrutinizing the decision tree, we observed that the root node, log_innings pitched, plays a pivotal role in determining a pitcher's WAR. Specifically, higher innings pitched correlate with higher WAR values. Additionally, our analysis revealed that distinguishing factors for pitchers with high WARs include a superior strike-to-walkout ratio and a lower hits-per-9-innings ratio. These insights shed light on the key metrics influencing a pitcher's performance and WAR, providing valuable insights for player evaluation and team strategy.

5.2.2 Position Players

5.2.2.1 Position Players regression

OLS Regression Results

Dep. Variable:	WAR	R-squared:	0.550
Model:	OLS	Adj. R-squared:	0.541
Method:	Least Squares	F-statistic:	64.44
Date:	Fri, 23 Feb 2024	Prob (F-statistic):	1.13e-76
Time:	19:58:40	Log-Likelihood:	-809.70
No. Observations:	485	AIC:	1639.
Df Residuals:	475	BIC:	1681.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.5779	0.698	3.691	0.000	1.206	3.950
log_OPS	9.6264	0.571	16.844	0.000	8.503	10.749
log_BB/K	0.4498	0.150	2.998	0.003	0.155	0.745
log_salary	0.3478	0.050	6.896	0.000	0.249	0.447
Age	-0.1122	0.021	-5.405	0.000	-0.153	-0.071
2B	1.0036	0.189	5.308	0.000	0.632	1.375
3B	1.0088	0.185	5.441	0.000	0.644	1.373
CF	0.8627	0.214	4.025	0.000	0.442	1.284
SS	1.3539	0.182	7.449	0.000	0.997	1.711
C	1.2109	0.329	3.678	0.000	0.564	1.858

Omnibus:	3.937	Durbin-Watson:	1.069
Prob(Omnibus):	0.140	Jarque-Bera (JB):	3.750
Skew:	0.176	Prob(JB):	0.153
Kurtosis:	3.248	Cond. No.	389.

This regression analysis endeavors to discern the factors influencing Wins Above Replacement (WAR) in position baseball players, employing various independent variables. The model's statistical evaluation provides insight into its effectiveness in explaining the variability observed in WAR. The R-squared value, a measure of how well the independent variables explain the variance in the dependent variable, stands at 0.55. This indicates that approximately 55% of the variability in WAR can be accounted for by the factors included in the model. Moreover, the adjusted R-squared, which considers the number of predictors and adjusts the R-squared value accordingly, is calculated at 0.541, reflecting a slight adjustment for the complexity of the model. The F-statistic, a test assessing the overall significance of the model, yields a value of 64.44, accompanied by an exceedingly low p-value. This outcome strongly suggests that the model is statistically significant, implying that at least one of the independent variables has a non-zero effect on the dependent variable. When examining the coefficients, each variable's impact on WAR becomes clearer. The intercept, representing the expected WAR when all independent variables are zero, stands at 2.5779. Meanwhile, key predictors such as log-transformed On-Base Plus Slugging (OPS), log-transformed Walks per Strikeout (BB/K), and log-transformed Salary exhibit notable impacts on WAR. For instance, a one-unit increase in log_OPS corresponds to a substantial increase of 9.6264 in WAR, highlighting the significance of offensive performance metrics. Conversely, older age demonstrates a negative association with WAR, indicating a decline in performance as players age.

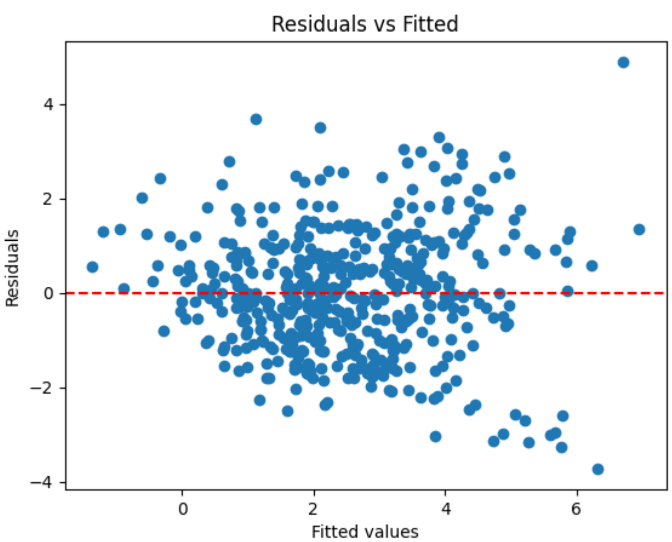
Furthermore, the inclusion of player position dummy variables adds depth to the analysis, elucidating the differential effects of various positions on WAR. We note that players who are shortstops and catchers give a premium to a baseball team in contributions to WAR, followed closely by second base and thirdbasemen. Centerfielders offer slightly less WAR at a constant salary level than the aforementioned positions, however more than the positions lumped into the intercept (LF, RF, 1B, and DH).

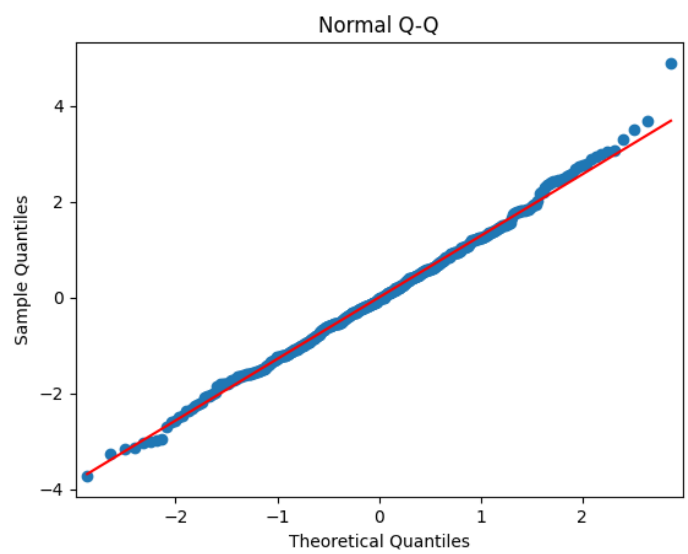
5.2.2.2 Position Player VIF

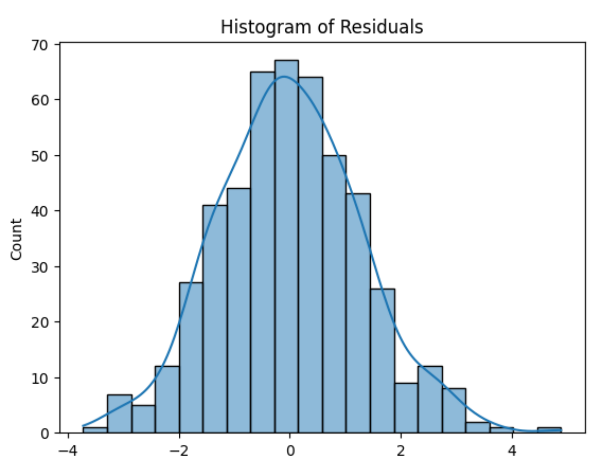
Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	2.5779007	0.698383	3.69	0.0002*	.
Log[OPS]	9.6263619	0.571498	16.84	<.0001*	1.2671908
Log[BB/K]	0.4498191	0.150041	3.00	0.0029*	1.212918
Log[salary]	0.3478469	0.050442	6.90	<.0001*	1.5537783
Age	-0.112165	0.020752	-5.41	<.0001*	1.5997787
2B	1.0036398	0.189094	5.31	<.0001*	1.2097871
3B	1.0087525	0.185406	5.44	<.0001*	1.1331932
CF	0.8626802	0.214319	4.03	<.0001*	1.1348551
SS	1.3538753	0.181754	7.45	<.0001*	1.2697418
C	1.2108777	0.329265	3.68	0.0003*	1.0553111

There is no sign of multicollinearity between the independent variables.

5.2.2.3 Position Player Regression Residuals







The Residuals vs Fitted plot illustrates a relatively random scatter of points, indicating consistent variance, or homoscedasticity, in the residuals. Although we observed some fluctuations in the scatter plot, we attempted to address this by transforming the dependent variable (WAR). However, upon examining the resulting residual plot, we noted a lack of homoscedasticity relative to the current model, leading us to retain the original untransformed variables. The Normal Q-Q plot compares the distribution of residuals to a normal distribution, with most points aligning closely with the red line, suggesting approximate normality. Similarly, the Histogram of Residuals exhibits a bell-shaped distribution, characteristic of a normal distribution. Overall, these plots suggest that the regression model meets several assumptions: the residuals are mostly normally distributed and exhibit consistent variance. However, there are indications of potential outliers and slight deviations from normality, particularly in the tails of the distribution.

5.2.2.4 Cross Validation

Similar to pitchers, we partitioned the dataset into training and testing subsets, aimed to assess the robustness and generalizability of our regression model. Across various training/testing split ratios (75/25, 80/20, and 70/30) and fold counts (5 and 10), our model exhibited consistent performance, as evidenced by stable cross-validation scores.

The Mean Squared Error (MSE) on the testing set, a measure of the average squared difference between the predicted and actual values, was the lowest at 1.743 at a training and test split ratio of 75/25. We do note that when adjusting the splits to 80/20 and 7030, the MSE on the testing set was 1.918 and 1.818 respectively. The 0.1 variation suggests that our model's predictive accuracy is reliable when applied to unseen data.

Furthermore, the average cross-validation scores for MSE, calculated across 5 folds and 10 folds, further validate the model's stability. Despite slight variations, the average MSE scores remained within a narrow range, with values of 2.710 for 5 folds and 2.218 for 10 folds. This consistency underscores the model's ability to generalize well to new data, regardless of the specific training/testing split ratio or the number of folds used in the cross-validation process.

5.2.2.5 Decision Tree

We conducted an extensive analysis employing decision tree regression alongside varied testing/training splits to assess the efficacy of our variables. Despite encountering fluctuations, our regression models yielded promising outcomes. Notably, the Mean Squared Error (MSE) stood at approximately 2.84, indicating a moderate level of prediction error. Through meticulous hyperparameter tuning, we identified the optimal maximum depth of the decision tree at 4, resulting in a best score of around 0.249. Impressively, both the Bagging ensemble model and the XGBoost regressor exhibited commendable accuracy on the test data, achieving scores of approximately 0.580 and 0.569, respectively. Similarly, the Random Forest model demonstrated respectable accuracy of about 0.578 on the test data. These collective findings underscore the potential of our regression models in predicting the dependent variable, while also suggesting avenues for further refinement to enhance individual model accuracy.

A comprehensive analysis of the decision tree reveals a clear delineation between players with superior on-base plus slugging (OPS) metrics and higher salaries, correlating with elevated levels of Wins Above Replacement (WAR), in contrast to their peers categorized as average or below average. Additionally, noteworthy observations emerge regarding the age and positional roles of players. Specifically, players aged over 33.5 years who do not occupy shortstop positions demonstrate diminished WAR values. This nuanced finding underscores the enduring prowess of older shortstops, who not only maintain their on-field performance but also contribute invaluable veteran leadership within the team dynamic.

6. Discussion

Our findings resonate with the widely accepted principles within baseball management circles, underscoring the strategic importance of allocating resources towards players in pivotal positions known colloquially as "up the middle." These positions, comprising second baseman, shortstop, centerfield, and catchers, are traditionally recognized for their critical defensive and strategic roles on the field while also being expected to contribute offensively. While our analysis reveals that second basemen exhibited a slightly lesser impact compared to the other three positions, it is noteworthy that both first basemen, designated hitters, right, and left fielders demonstrated statistically lower contributions to Wins Above Replacement (WAR). This observation aligns with prevailing baseball wisdom, which acknowledges that players in left field and first base typically prioritize offensive prowess over defensive excellence, thus rendering their overall WAR contribution relatively less significant. The success of the Rangers in 2023, securing their first-ever World Series victory after investing heavily in middle-infield players Marcus Semien and Corey Seager, further supports this strategy. Semien and Seager secured contracts worth a combined $500 million, the most for a middle infield duo ever. They also finished the season ranked sixth and seventh in WAR, demonstrating that the Rangers' investment paid off in wins.

Our findings underscore the importance of pitchers who consistently "eat innings," serving as reliable starters capable of enduring extended periods of play while maintaining commendable performance levels. This logical association is rooted in Wins Above Replacement

(WAR) accumulation, where sustained on-field presence emerges as a crucial determinant. Consequently, pitchers logging substantial innings tend to amass higher WAR totals over time. This observation holds strategic significance for Major League Baseball (MLB) front offices, emphasizing the need to prioritize the identification and acquisition of pitchers adept at handling significant workloads. Effectively leveraging such pitchers enables teams to optimize bullpen usage, mitigate strain on relief pitchers, and preserve pitching resources throughout the season. Additionally, our results indicate a noteworthy resilience in the face of aging among pitchers, suggesting that age may not significantly diminish their production levels.

The study's scope is constrained by several limitations inherent in the dataset utilized, primarily centered around its temporal and contextual constraints. Firstly, the dataset is confined to the salary and performance metrics of active baseball players exclusively spanning the 2019-2022 seasons. Consequently, this limited timeframe may fail to capture the longitudinal evolution and progression of players' skills and performance over extended periods. Moreover, the dataset incorporates data from the truncated 2020 season, significantly impacted by the COVID-19 pandemic, resulting in a reduced number of games played (60 as opposed to the standard 162). Notably, player salaries during this aberrant season were prorated in accordance with the abbreviated schedule, potentially skewing the salary distribution and introducing bias into our analyses due to incomplete and atypical data representations. These limitations underscore the need for cautious interpretation and generalization of findings derived from the dataset.

Another limitation lies in our position dataset, which currently only designates one "primary position" for each player, overlooking the reality that many players often excel in multiple positions. Factoring in players' secondary and tertiary positions could yield more substantial and nuanced results.

7. References

Clemens, Ben. "What Are Teams Paying per War in Free Agency?" *FanGraphs Baseball*, 3 Mar. 2022, blogs.fangraphs.com/what-are-teams-paying-per-war-in-free-agency/. Accessed 19 Nov. 2023.

"Pitcher War Calculations and Details." *Baseball*, www.baseball-reference.com/about/war_explained_pitch.shtml. Accessed 11 Sep. 2023.

"Position Player War Calculations and Details." *Baseball*, www.baseball-reference.com/about/war_explained_position.shtml. Accessed 11 Sep. 2023.

Slowinski, P. (2010, February 15). *What is WAR?*. Sabermetrics Library. https://library.fangraphs.com/misc/war/. Accessed 11 Sep. 2023.