

Quantifying the Success of Long-Term Contracts in Major League Baseball

CSC-475: Seminar in Computer Science
Spring 2024

Caden Parry
Department of Computer Science
Furman University
caden.parry@furman.edu

Abstract

Major League Baseball players are making more money than ever. In recent years, players like Shohei Ohtani, Mookie Betts, and Mike Trout have received record-breaking contracts worth hundreds of millions of dollars. These contracts raise the question of whether player production continues steadily throughout these contracts, or if it tapers off once they enter their largest contract. To answer this question, a data set of all MLB hitters from 1985 to 2023 was collected. Next, a statistical analysis was conducted to determine how often players' *Wins Above Replacement* (WAR) significantly varies from their career norms during their largest contract. The final phase of this project was to create a website that visualizes individual player WAR over time. This website can be found at: <https://cadenparry.github.io/MLB-WAR-Analysis/website/>. This study found that 87.5% of players who signed a contract worth over \$100,000,000 did not see a statistically significant deviation in WAR during this contract. These findings indicate that current MLB trends to sign players to long-term contracts are likely optimal because there is only a 13% chance that players' WAR will deviate significantly over during their largest contract.

1 Introduction

In December of 2023, Japanese superstar Shohei Ohtani signed a massive 10-year 700 million dollar contract with the Los Angeles Dodgers, the largest in sports history. While Shohei Ohtani's contract is a definitive outlier, the baseball world has become increasingly accustomed to seeing contracts totaling hundreds of millions of dollars. As a result of these lucrative signings many fans and executives alike are tasked with answering the question, "Was all of that money worth it?" In the mid to late 1900's teams were far more risk-averse than they are today, straying away from "mega deals", like Shohei Ohtani's contract. But in recent years, more MLB teams are opening up their pocketbooks to secure the best of the best, long term. Is this new-age thinking truly the right choice, or were the more frugal, short-term deals of the past more optimal?

This research will work to provide quantitative measures to determine how often MLB players' production deviates from their career norms during their largest contract. This question is important because it works to quantify the risk that teams face when they sign a player long-term. In practice, this project will work to build an all-encompassing database of players from 1985 to 2023 and then compute how often players' performance fluctuates during different contract lengths. A database or easily accessible reference tool housing this data could serve as a beneficial resource for fans, journalists, and MLB executives.

To answer this question, this project will leverage *Wins Above Replacement* (WAR), a statistic that attempts to quantify how many more wins a player generates than a "replacement" level player. The term "replacement level player" encompasses all players who fall slightly below average production

marks in the MLB. These players were coined as "replacement" level because it is accepted that they can easily be "replaced" by a variety of Minor League players. *WAR* is very complicated to compute and can be considered a black box stat, meaning we only care about its output, not the computation it took to achieve it. For the purposes of this project, one only needs to understand the scale on which a player *WAR* is measured. A *WAR* of 0 or lower is considered a "replacement level." A *WAR* from 0-2 represents a player who can act as a substitute but is likely not an everyday starter. A *WAR* of 2-5 deems a player a likely everyday starter. A *WAR* of 5-8 deems a player a likely all-star. And, finally, a *WAR* of 8 or above deems a player a likely MVP candidate.

To achieve this goal, a data set of all MLB hitters from 1985-2023 with more than 7.6 career *WAR* was created. Unfortunately, pitchers were not able to be included in the data set due to time constraints and unforeseen implementation issues. All players with less than 7.6 career *WAR* also had to be filtered out of the data due to these time constraints and implementation issues. All data for this project was collected from Baseball-Reference and Spotrac.com. The data obtained from Baseball-Reference was obtained manually, but a web scraper was implemented to obtain contract data from Spotrac.com. Once the data collection and cleaning were complete, this project leveraged *Scipy stats* and *matplotlib*, two powerful Python libraries, to conduct t-tests and other further analyses on this data. This study found that only 15% of players, on average, exhibited statistically significant deviations in their *WAR* production during their largest contract.

2 Related Work

The concept of MLB free agency itself is not as old as one might expect. Prior to 1975, if a player signed an initial contract with a team they were virtually bound to that team for life unless they were traded or released. It was not until December of 1975 that MLB players were granted the right to enter the free market after the conclusion of their contract [1]. This seismic shift in power from owners to players created a new competitive market that drives the game of baseball to this day. In this ever-evolving landscape, the dynamics of player contracts and their subsequent performance evaluation have garnered increasing attention. Recent groundbreaking deals like Shohei Ohtani's 10-year, 700 million-dollar contract with the Los Angeles Dodgers, have prompted discussions on the successes and failures of large MLB contracts. The literature surrounding labor relations, contracts, and player evaluation provides valuable insight into the complex dynamics between players, team owners, and agents. Understanding this body of work is essential for comprehending the context that governs player contracts and free agency in MLB.

The introduction of free agency changed the landscape of MLB and created contractual trends that are still present today. Understanding which players are able to demand these contracts, how they pursue them, and how these transactions differ from traditional economic markets is pivotal for understanding free agency as a whole. Kahn explains that both arbitration and free agency eligibility lead to higher annual compensation for players. However, it is noteworthy that only free agency eligibility is associated with an increase in contract duration [2]. The findings of this paper indicate that free agency allows players to negotiate more favorable terms. This work solidifies the colloquial knowledge that more successful MLB players use their privilege to demand longer-term contracts and that MLB teams are inclined to engage in uniquely large contracts to court highly talented players. In tandem with this, Maxcy's research highlights the nuanced relationship between economic incentives, risk perception, and contractual arrangements in professional sports labor markets, adding depth to our understanding of player contracts and labor relations in MLB. Maxcy explains that large contracts are predominantly held by star players rather than marginal ones. This greatly differs from traditional economic models that suggest that marginal workers seek long-term contracts to mitigate income risk, while high-producing workers can secure economic stability without engaging in long-term contracts [3]. These findings highlight *who* is signing these contracts, but an important factor has yet to be discussed.

A key factor in the contractual process is that of the player's agents. A player agent's role in the contractual process is to aid a given player in the process of communicating with teams and signing contracts. In entering into this agreement, the player places a large amount of faith in their agent to field offers from other teams that are representative of their desires, but this is not always the case. Agents representing a large number of clients are more likely to set a large minimum-acceptable-salary for their client. While this *may* benefit their client, an agent's lack of risk in any given singular contract may adversely affect their client as lower offers the player may have accepted will not be

heard by the agent [4]. These behaviors create unique opportunities where a player may be signed to considerably less or more lucrative contracts than the market would suggest. An analysis of player contracts can aid teams and players in the decision-making process to avoid the pitfalls of rogue agents.

Equally as important as who is signing these contracts, is how players are performing throughout them. A specific point of focus for this topic is how players perform based on the construction of a contract. Although fans do not always remember this, players are people too and they can be greatly affected by factors outside of simply playing the game of baseball. Players who are less likely to secure subsequent contracts exhibit a significant reduction in performance compared to expectations, highlighting the influence of contract uncertainty on player output. However, the incentive to perform in the final year of a player's given contract often leads to performance expectations being exceeded in hopes of securing a new contract [5]. Furthermore, even if a player is able to sign a large contract researchers have identified production patterns that show that player outputs are likely to decrease at the beginning of a long contract and increase toward the end [6].

This phenomenon is known as shirking, which means to avoid or neglect something. In the case of baseball, researchers assert that players are more likely to shirk or slack off at the beginning of a major contract because of the large amount of guaranteed money in their contract [6] [7]. Although there is a bit of conflicting evidence with regard to shirking, Krautmann and Donley found no detection of shirking when testing with performance metrics but did detect shirking when conducting tests on the marginal revenue product produced by a player [6]. Conversely, O'Neill and Deacle were able to detect shirking in performance metrics that had been standardized based on the total offensive and defensive production through the league [7]. Krautmann also published two responses to the creator of MRP for baseball players [8], critiquing [9] and completing [10] the research. Although there are numerous disagreements, the academic community is in agreement that shirking is actively occurring in MLB.

With a better understanding of shirking in MLB in place, researchers devised a method of quantifying if a player met, fell below, or exceeded expectations in their contract. Again, led by Krautmann, researchers estimated economic impact on a team is often not realized and MLB teams are prone to overpay players on long-term contracts [11]. This highlights the variability of player production and strengthens the body of work surrounding viable prediction and evaluation models.

With these previous works and their issues addressed, *Quantifying the Success of Long-Term Contracts in Major League Baseball* will work to hone in on specific players and streamline the findings to the public. Although these academic papers present graphic visualization of their data, many are difficult to quickly comprehend or gain access to if you are not associated with an academic institution. This research will work to lower the barrier of entry to the data encompassed in previous work, conduct case studies on specific players' performance and trends over time, as well as make the findings publicly available and easily comprehensible.

3 Data

All of the data gathered for this project came from two sources: Baseball-Reference [12] and Spotrac.com [13]. This project used Baseball-Reference to collect WAR values for every MLB hitter from 1985 to 2023 and Spotrac.com to collect each individual player's contract data. I specifically collected the *Player Value* [12] data tables for the years 1985 to 2023 from Baseball-Reference and the *Transaction* [13] data tables for each player from Spotrac.



Baseball-Reference relies on MLB's officially published play-by-play data, as well as *Retrosheet.org*. *Retrosheet.org* primary use case is for data collection prior to 1984, as many games prior to this data are not as readily available. Baseball-Reference is forthcoming about the lack of available data for players and teams pre-1984 and displays a useful graphic on their website (found at <https://www.baseball-reference.com/about/coverage.shtml>) that details what year, players, data values, etc. that may be incomplete. After conducting a review of Baseball-Reference's data, it was concluded that there were no missing data entries for MLB hitters from the years 1953 to 2023, and thus, missing data from Baseball-Reference would not affect this project. Spotrac.com relies on *Cot's Baseball Contracts*, *USA Today*, and *mlbtraderumors.com* to populate their data set. Gathering accurate contract data for any sport can be quite complicated, as the first announced duration, amount, or even specified team of any given contract may be incorrect. Though not explicitly stated, it is

Figure 1: Baseball-Reference and Spotrac Web Data

Player Value [WAR Explained \(v2.2\)](#): 8+ MVP, 5+ A-S, 2+ Starter, 0-2 Sub, < 0 Repl [Share & Export](#) [Glossary](#)

Rk	Name	Age	Tm	G	PA	Rbat	Rbaser	Rdp	Rfield	Rpos	RAA	WAA	Rrep	RAR	WAR	waaWL%	162WL%	oWAR	dWAR	oRAR
1	CJ Abrams*	22	WSN	151	614	-5	4	3	4	9	14	1.4	21	35	3.4	.509	.509	3.0	1.3	31
2	José Abreu	36	HOU	141	594	-8	-2	-1	-1	-7	-19	-1.9	20	1	0.0	.486	.488	0.1	-0.8	2
3	Wilmer Abreu*	24	BOS	28	85	3	0	0	2	0	5	0.5	3	8	0.8	.517	.503	0.6	0.2	6
4	Ronald Acuña Jr.	25	ATL	159	735	63	7	-1	-2	-5	63	6.1	23	86	8.2	.538	.538	8.5	-0.7	88
5	Willy Adames	27	MIL	149	638	-5	-3	0	8	9	10	1.0	21	31	3.0	.507	.506	2.2	1.7	23
6	Jordyn Adams	23	LAA	17	40	-6	-1	0	-1	0	-8	-0.9	1	-7	-0.7	.450	.495	-0.6	-0.1	-6
7	Riley Adams	27	WSN	44	158	3	-1	-1	1	3	5	0.5	5	11	1.0	.512	.503	1.0	0.4	10
8	Ty Adcock	26	SEA	1	0	0	0	0	0	0	0	0.0	0	0	0.0	.500	.500	0.0	0.0	0
9	Jo Adell	24	LAA	17	62	-2	0	0	3	0	1	0.1	2	3	0.3	.507	.501	0.0	0.3	0
10	Ehire Adrianza#	33	ATL	5	11	-2	0	0	-1	0	-3	-0.3	0	-3	-0.3	.432	.498	-0.2	-0.1	-2
11	Jesús Aguilar	33	OAK	36	115	-3	-1	-1	-3	-2	-9	-0.9	4	-5	-0.6	.474	.494	-0.3	-0.5	-2
12	Nick Ahmed	33	ARI	72	210	-14	1	1	1	3	-8	-0.8	7	-1	-0.1	.489	.495	-0.2	0.4	-2
13	Hanser Alberto	30	CHW	30	90	-3	0	1	1	1	0	0.0	3	3	0.3	.499	.500	0.2	0.2	2
14	Ozzie Albies#	26	ATL	148	660	16	5	1	0	5	27	2.7	21	49	4.7	.518	.517	4.7	0.5	49
15	Scott Alexander*	33	SFG	1	0	0	0	0	0	0	0	0.0	0	0	0.0	.500	.500	0.0	0.0	0
16	Jorge Alfaro	30	2TM	18	52	-5	0	0	0	0	-6	-0.6	2	-4	-0.4	.466	.498	-0.5	0.0	-4
17	Greg Allen#	30	NYY	22	28	1	0	0	-2	-1	-1	-0.1	1	0	0.0	.495	.499	0.2	-0.3	2
18	Nick Allen	24	OAK	106	329	-17	1	-1	3	5	-8	-0.8	11	3	0.3	.492	.495	0.0	0.9	0
19	Abraham Almonte#	34	NYM	8	16	-3	0	0	-1	0	-4	-0.4	1	-3	-0.3	.450	.498	-0.2	-0.1	-2
20	Pete Alonso	28	NYM	154	658	16	-2	-1	6	-8	11	1.1	22	33	3.2	.507	.507	2.6	-0.2	27

(a) Baseball-Reference *Player Value* Data

 Mike Trout CENTER FIELD (CF) Age: 32-261d Exp: 12.07 Years Bat/Throw: Right/Right		Drafted: Round 1 (#27 overall), 2009 Country: United States Agent(s): Craig Landis (LSW Baseball)		EMBED THIS  - More Angels -
Contract Details	Career Earnings	Transactions	Injuries	Statistics
Transactions				
MAR 19 2019 Signed a 12 year \$426.5 million contract extension with Los Angeles (LAA)				
MAR 28 2014 Signed a 6 year \$144.5 million extension with Los Angeles (LAA)				
FEB 26 2014 Signed a 1 year \$1 million contract with Los Angeles (LAA)				
MAR 3 2013 Signed a 1 year \$510,000 contract with Los Angeles (LAA)				
MAR 3 2012 Signed a 1 year \$482,500 contract with Los Angeles (LAA)				
JUL 5 2009 Signed a contract with Los Angeles (LAA)				
JUN 24 2009 Drafted by Los Angeles (LAA): Round 1 (#27 overall)				

(b) Spotrac *Contract Transaction* Web Data

likely that Spotrac relies on multiple sources to gather its contract data to build redundancy into its data collection process. Spotrac’s MLB contract dataset is considerably smaller and less complete than that of Baseball-References as it only contains fully accurate records from 1985 to the present. Given this fact, the availability of Spotrac’s contract data was the limiting factor on this project’s final scope. Even with their limitations, Baseball-Reference.com and Spotrac.com are accepted by the baseball community as best practice data collection sites due to the fact that all data is publicly available, updated daily, and extremely reliable.

Prior to the application of any constraints on the data, the sample size of the Baseball-Reference data was 51297, as in there were 51297 individual rows. This number includes every MLB player and every season they played from 1985 to 2023. That is, if a player plays 10 years they will be counted ten times. The sample size of the scraped Spotrac data was 9239, as in there were 9239 individual rows. This number reflects the number of unique MLB players from 1985 to 2023. The initial Baseball-Reference data contained a variety of data columns but this project was only concerned with player identification information like name, age, etc., and a player’s WAR for every season. The initial Spotrac contract data that was collected was curated specifically for this project so only useful data was included in the table and no initial data cleaning was needed. The columns that were included in the contract data included player, and player-id, which were used to match players to their stats in the Baseball-Reference data, as well as the amount, length, year, team, and position played during a given player’s largest contract.

Table 1: Baseball-Reference Data

player_id	yearID	Name	Age	teamID	WAR
troutmi01	2023	Mike Trout	31	LAA	2.9
bettsmo01	2023	Mookie Betts	29	LAD	6.7
judgeaa01	2023	Aaron Judge	30	NY Yankees	10.5
machama01	2023	Manny Machado	30	SDP	2.9
lindofr01	2023	Francisco Lindor	28	NYM	5.6
...

Table 2: Players Largest Contract Data

player	player_id	amount	length	year	team	position
Mike Trout	troutmi01	\$426,500,000	12	2019	LAA	hitter
Mookie Betts	bettsmo01	\$365,000,000.00	12	2020	LAD	hitter
Aaron Judge	judgeaa01	\$360,000,000.00	9	2023	NY Yankees	hitter
Manny Machado	machama01	\$350,000,000.00	11	2023	SD	hitter
Francisco Lindor	lindofr01	\$341,000,000.00	10	2021	NYM	hitter
...

The data previously described was the *premerged*. We will now discuss how that data was cleaned, merged, and filtered to create a working data set for the analysis and visualization. First, the two data tables were merged together based on the *player_id* column. This merge created some redundancies in the data because a player’s salary information would be listed in every season row. Next, a *out_of_contract* and *total_WAR* column was created. The *out_of_contract* column was created to signify which year(s) a given player was in their largest contract. The *total_WAR* columns depict the total amount of WAR that a player accumulated through their entire career. This column would be used later to filter out players who did not have more than 7.6 total career WAR. Players with less than 7.6 total career WAR needed to be removed because there were a large number of pitchers that appeared in the *Batter Value* table, and thus, needed to be removed. While there are other solutions to this issue, this solution seemed based on the time constraints of the project. Finally, all nulls and unnecessary columns were removed. All of the removed columns contained statistics that were not necessary for the analysis and visualization of this project. The final table, *player_visualization* has 9 columns and 5807 rows. The *Player*, *player_id*, and *yearID* were used to identify players and any given year that they played a game in MLB. The *WAR* and *out_of_contract* column tracked a given player’s WAR for each year and the years that they were or were not playing during their largest contract. The *length*, *amount*, and *total_WAR* columns were constant for any given player as they did

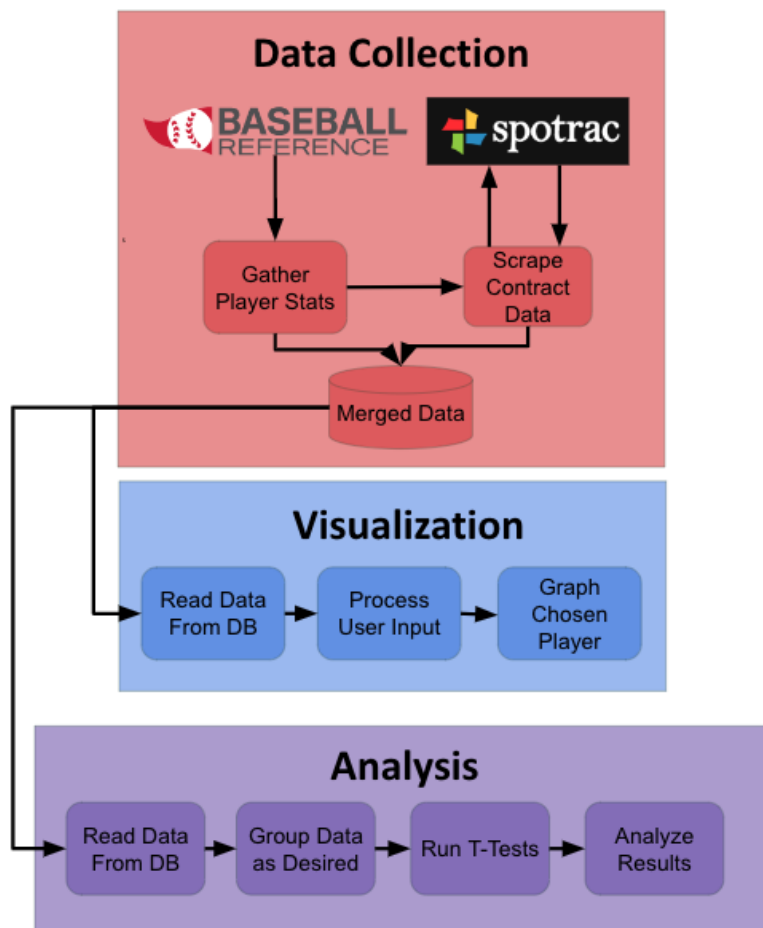
not change based on the year but they were still pivotal for the visualization. The number of rows, 5807, encompasses the WAR, salary info, etc. for every season for every player in the data set.

Table 3: Final Player and Contract Data

player	yearID	WAR	length	amount	out_of_contract	total_WAR	
Mike Trout	2023	2.9	12	\$426,500,000	False	85.1	
Mike Trout	2022	6.2	12	\$426,500,000	False	85.1	
Mike Trout	2021	1.8	12	\$426,500,000	False	85.1	
Mike Trout	2020	1.8	12	\$426,500,000	False	85.1	
Mike Trout	2019	7.9	12	\$426,500,000	False	85.1	
...

4 Methods

The workflow pipeline for this project included three main phases: Data Collection, Visualization, and Analysis. Both the Visualization and Analysis phases are dependent on the completion of the Data Collection phase, but it should be noted that the Visualization and Analysis are not dependent on one another. That is to say, that the Visualization and Analysis phases can run independently of one another. A graphical representation of this pipeline is depicted below. The *start* of the pipeline is the Baseball-Reference node and the subsequent arrow(s) denote the next phase in the pipeline. Each stage in the pipeline presented unique challenges, but clear implementation goals and a variety of code libraries were used to overcome these issues.



4.1 Data Collection

The goal of the Data Collection phase was to create a single data set that encompassed all 1985 to 2023 MLB player's season WAR and largest contract data. Although all of the data is publicly available, the data collection process for this project was quite time-intensive. After much trial and error with data sets from *Kaggle*, and websites like *baseballsavant.com* and *fangrpahs.com* it was concluded that Baseball-Reference and Spotrac were the best options.

All of the data collection from Baseball-Reference was done manually as there were only 39 individual tables that needed to be collected (1985 to 2023). For this process, the number of players was insignificant because each season's data contained all of the players for that year. Thus, only the number of years was taken into account when analyzing the most efficient way to collect data. To collect the Baseball-Reference data, each year's respective table was manually downloaded from Baseball-Reference [12] as a csv and combined into an all-encompassing csv that holds data for all years from 1985 to 2023. After all 39 years of data were collected, some simple data cleaning was done to prepare the data to be used by the Spotrac contract scraper.

The data collection process for Spotrac was dependent on the total number of unique players spanning from 1985 to 2023. The input size for this process is far larger given that there are thousands of unique baseball players from the years 1985 to 2023. Due to this, manual data collection from Spotrac was deemed inadequate. To remedy this issue, a web scraper was implemented to automate the data collection process. The input list of players for the scraper was all unique players from 1985 to 2023. To create this list, each unique player ID from the previously gathered Baseball-Reference data was added to a list and passed to the scraper. This web scraper primarily leveraged the *Pandas*, *requests*, and *BeautifulSoup Python* libraries. The algorithm for this web scraper is depicted below.

Algorithm 1 Spotrac Contract Scraper

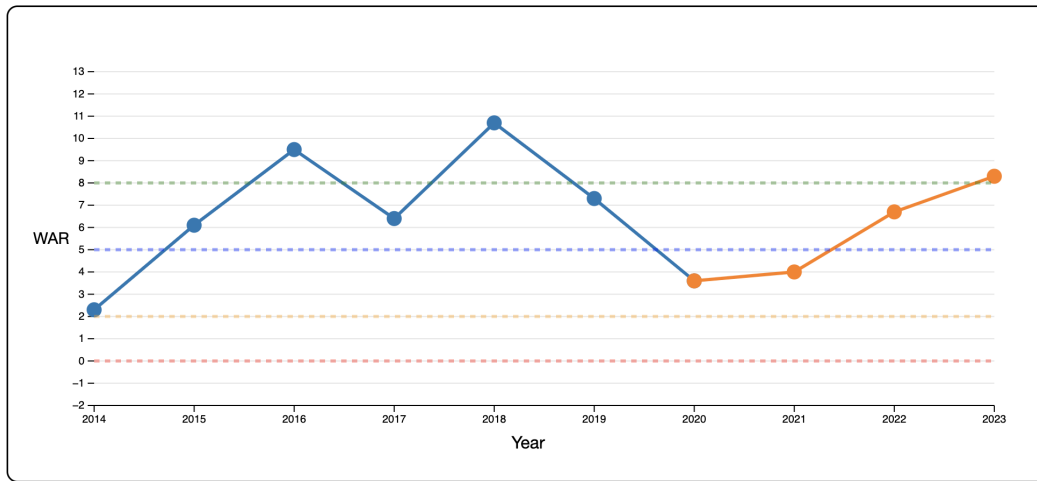
```
0: for player in all_unique_players do
0:   Create player URL
0:   Request base HTML
0:   if HTML is null:
0:     break
0:   Get player WAR info and largest contract info
0:   Add information csv file
```

This process took over 2.5 hours as there were close to 10,000 players whose contract data needed to be requested and each request took 1 second. After this process was complete, the Spotrac data was merged with the Baseball-Reference data as previously described.

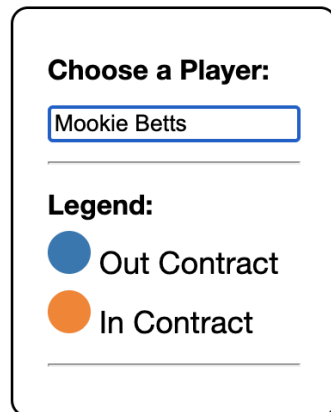
4.2 Visualization

The goal of the *Visualization* phase was to create an interactive and user-friendly website that visualizes a given player's WAR throughout their career, specifically fighting their WAR during the years of their largest contract. To achieve this goal, HTML, JavaScript, and CSS were leveraged, alongside the previously mentioned data to create a website. The bulk of the work for this website was implemented in JavaScript, specifically the D3.js Library. This library allows you to create dynamic graphs that can change based on the player who is selected. An example output for Mookie Betts is depicted below.

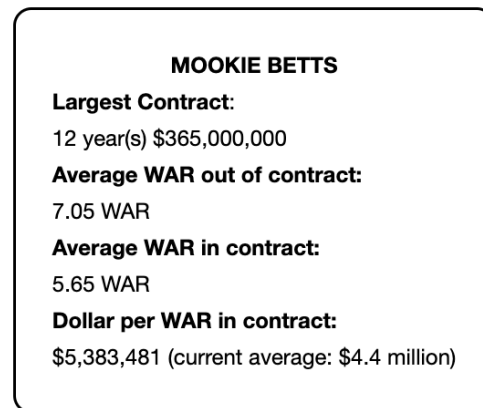
Figure 2: Mookie Betts Website Visualization



(a) Mookie Betts WAR Graph



(b) Mookie Betts Graph Legend



(c) Mookie Betts Contract Info

Figure 2a. depicts Mookie Betts' WAR for each year of his career. The color of graph denotes whether Betts' was or was not playing during his largest contract. Figure 2b. depicts the Legend for Figure 2a. As described by the Legend, dots and lines that are blue mean Betts' was *not* playing in his largest contract, and dots and lines that are orange mean Betts' *was* playing in his largest contract. Figure 3a. provides a few interesting data points that attempt to quickly describe Betts' earnings and production throughout his career. *Largest Contract* displays the details Betts' largest contract. *Average WAR out of contract* and *Average WAR in of contract* display the Betts' average WAR prior to and during his largest contract. *Dollar per WAR in Contract* depicts the amount of money that Bett's earned for every point of WAR that he accumulated. The *current average* side note describes the average amount that an MLB team spends for one WAR. This value can be used to determine how over or underpaid a player is based on league average WAR per dollar.

4.3 Analysis

The goal of the *Analysis* phase was to determine if MLB players' production changes when they enter their largest contract. To achieve this goal, a T-Test was conducted on two different groups from the merged Baseball-Reference and Spotrac data. To conduct the T-test in a Python environment, the *Scipy stats* library was used because it has a variety of t-test functions.

$$t - test = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (1)$$

The two primary groups were: *Contract Length* and *Contract Amount*. The *Contract Length* group encompassed all players in the data set but players were grouped based on the length of the contract. The final subgroups that were used were All Contract Lengths, 1-3 year contracts, 4-6 year contracts, and 7+ year contracts. Grouping players by the length of their contract provided further insights into the relationship of different contract lengths and WAR. The *Contract Amount* group encompassed all players in the data set but players were grouped by the amount of their contract. The final subgroups that were used were All players, Contracts less than \$100,000,000, Contracts between \$100,000,000 and \$200,000,000, and Contracts greater than \$200,000,000.

To conduct these T-Tests, the necessary data was read into a Python file using *Pandas*. After the data was read into the file, the necessary testing groups were created based. Then, for each of the previous groups, WAR in the largest contract and WAR out of contract were separated and the T-test was run on the two samples. The output of each t-test was sorted into 4 categories: Not significant, Significant, Very significant, and Highly significant.

All code, analyses, and visualizations for this project are publicly available on GitHub at:

- **Website Visualization:** <https://cadenparry.github.io/MLB-WAR-Analysis/website/>
- **Code, Data, and Analyses:** <https://cadenparry.github.io/MLB-WAR-Analysis>

5 Results

Given that the website visualization and data analysis were conducted independently of one another, the results of this project will be separated into two sections.

5.1 Results: Data Analysis

The results of the t-tests pointed to the fact that the majority of MLB players' WAR does not significantly deviate from the norm during their largest contract. Based on the results of the analysis of this project, the sweet spot for MLB contracts is contracts of 4-6 years in length that are worth between 100 and 200 million because these categories of players exhibit the lowest rates of statistically significant deviance in WAR.

For all contracts spanning 4-6 years, roughly 89% of players did not see a significantly significant in WAR during their largest contract. This means that only around 10% of players whose largest contract was 4-6 years saw a significant change in production, positive or negative, during their largest contract. MLB players who signed contracts that were 7+ years or less than 4 years in length did exhibit a slightly higher rate of deviation, with 12-18% having significant deviances from career norms. This, though, is to be expected as short-term deals are usually given to less productive players who have yet to prove themselves to MLB teams. Conversely, long-term contracts of 7+ years are all but exclusively reserved for the best players in all of MLB, and while this group certainly encompasses the best player, these deals take players into their mid to late 30s where production inevitably drops off due to age.

The full results of the analysis are depicted at the bottom of this section. The columns denote the span of contract lengths or the total amount of a contract that was analyzed and the rows denote the statistical significance of the WAR values. As an example examine *Table 4: Statistical Significance of WAR Deviations by Contract Length*, let's take the *1-3 Years* column and the *% Not Significant* row. The meeting of these points is 88.46% and should be interpreted to say *88.4% of players whose largest contract lasted 1-3 years will not play significantly better or worse than they did when they were not in their largest contract.*

These findings align with the current trend of MLB signing more players to lucrative long-term deals. Had the results of this project, for example, found that 50% of MLB players deviate from their career norms it would have been quite puzzling as to why so many teams are willing to lock up players long-term. It is virtually impossible to truly predict how well a player will perform in the

next 3-5 years, but it seems quite logical that teams would be willing to commit large sums of money to individual players if they have, so this study found, that around an 85% chance that that player will not deviate from their career norms. It should be noted though, that the bounds of what defines *statistically significant* in this study were not adjusted to "baseball terms." Seeing a deviance of 1-2 WAR may not show up as statistically significant in this study, but many baseball fans would assert that this is a sizeable enough deviation to be considered *statistically significant*. Regardless of that side note, the results of this study were quite positive as they align with common trends in the MLB.

Table 4: Statistical Significance of WAR Deviations by Contract Length

	All Contract Lengths	1-3 Years	4-6 Years	7+ Years
% Not Significant	88.08%	88.46%	89.73%	82.54%
% Significant	3.84%	3.15%	3.42%	7.94%
% Very Significant	1.82%	0.35%	4.11%	3.17%
% Highly Significant	0.92%	0.35%	1.37%	1.59%
Total Players	468	264	244	60

Table 5: Statistical Significance of WAR Deviations by Contract Amount

	All Contracts Amounts	>\$100 Mil	\$100 - \$200 Mil	<\$200 Mil
% Not significant	88.08%	87.74%	91.97%	82.61%
% Significant	3.84%	4.25%	4.17%	13.04%
% Very Significant	1.82%	1.65%	2.08%	0%
% Highly Significant	0.92%	0.94%	0%	0%
Total Players	468	401	47	22

5.2 Results: Website Visualization

The "results" of the website visualization, though not far less quantitative than the data analysis portion of this project, were very positive. The major goals for this website were to:

1. Create a tool that is easy to use and visually appealing
2. Create an auto-fill-in search bar to ease the burden of typing player names
3. Have access to an accurate and all-encompassing list of MLB players from 1985 to 2023
4. Display all collected data accurately as described in the data set

For the remainder of this section, we will refer to each goal by its number, as in, 1. will represent "Create a tool that is easy to use and visually appealing" and so on. Goals 1, 2, and 4, were the greatest successes. The website format and color scheme present an inviting environment for a variety of users to interact with the website. The auto-filling search bar was a particularly great addition, as many test users took notice of the ease that it proved when using the website. All user interactions were overwhelmingly positive and the website received praise from baseball and non-baseball fans alike. Along with the visual appeal and ease of use, the website also accurately depicts all as it is intended to. The most glaring shortcoming of the web visualization is the completeness of the final set of players. In theory, the website should be able to display all pitchers and hitters, regardless of their WAR. Sadly, this goal was not achieved as the final data set did not contain any pitchers and only held hitters with more than 7.6 career WAR. While this failure certainly affects the visualization tool, it is not the website's fault that the data set it is reading is incomplete. Overall, the web visualization was a great success that provides a launch pad for future work.

6 Future Work

The primary limitation of this research was the accuracy of the final data set that was used for visualization and analysis. Although Baseball-Reference and Spotrac provided accurate data, the act of scraping, cleaning, and merging the data proved to be more difficult than expected. The primary issues that were encountered were the accuracy of contract data that was scraped from Spotrac, lack

of contract data prior to 1985 on Spotrac, and the inclusion of pitchers in the Batter Value database of Baseball-Reference. As it stands, the best resolution to the contract scraper's issues would be to take into account the month that the contract was signed. A regular MLB season usually runs from late March to early November, so any contract that was signed prior to March should mark the start date to be the given year, but a contract that is signed in or after March should mark the start date of that contract to be the following year. For example, if a player signs a contract on December 10th, 1999, a date that falls after the current year's season has ended, the first year that the player will play under the given contract will be 2000. Conversely, if a player signs a contract on January 10th, 1999, a date that falls before the start of the current year's season, the first year that the player will play under the given contract will be 1999. To remedy the issue of missing contract data pre-1985, other data sources should be investigated. It should be noted though, that pre-1985 contract data was not collected as rigorously as it is today and may not even be available. To remedy the issue of pitchers appearing in the Batting Value data, Pitcher Value data could be collected to act as a cross-reference for which players were primary pitchers, and which players were primary hitters. With fully-fledged Hitter Value and Pitcher Value data, one could search the tables for duplicate players and assign that duplicate player as "hitter" or "pitcher" based on the position in which they accumulated the most WAR.

Working with these constraints, the current data analysis and visualization was done *only* on players that appeared in the Batter Value data with more than 7.6 total career WAR. This solution was chosen because it successfully removed all pitchers from the data, although it also removed a large number of hitters from the data as many players never accumulated 7.5 career WAR. Specifically, this constraint affected the inclusion of players who have only played a few MLB seasons. These constraints undoubtedly skewed the results of this research as the dataset did not fully encompass the 1985 to 2023 MLB seasons.

The natural next steps for this research would be to incorporate accurate pitcher data into the dataset, create a more scalable system that would allow future seasons to be added seamlessly into the current dataset, and to continue innovating upon the current analysis and visualization. The inclusion of accurate pitcher data would allow the findings from any data analysis to be extended to the entirety of the MLB, instead of just hitters. Furthermore, it would also greatly improve the use cases for the website visualization tool as more players would be included. The process of creating a scalable system would yield immense gains. The primary purpose of building a scalable system would be to allow new MLB season data to be added to the current dataset. This would vastly improve upon the current work because it would allow the visualization and analysis to grow in accuracy with each new MLB season. To implement this system, one would need to create a program that accurately inserts new player data into the current working player data. A foreseen issue for this process would be updating players' largest contracts based on new signings. In its current implementation, the scraper code took multiple hours to run, so the new solutions to obtaining player signings should be used. Finally, continued data analysis could yield new information about MLB, as well as provide a launch pad for more features in the visualization. Interesting further analysis may include but is not limited to: how well MLB players play the year before their largest contract, how many players have their best or worst year during their largest contract, and how player WAR changes over time based upon their age. Incorporating age into all analyses would be particularly helpful because the current analysis method may disproportionately fault older players for having lower production, when in fact, they are performing better than their age would predict. A variety of useful visualizations of player production drop-off due to age can be found on Google and can act as a guiding wind for this investigation.

References

- [1] History of free agency in the mlb, 2023. February 20, 2024.
- [2] Lawrence M. Kahn. Free agency, long-term contracts and compensation in major league baseball: Estimates from panel data. *The Review of Economics and Statistics*, 75(1):157–164, 1993.
- [3] Joel Maxcy. Motivating long-term employment contracts: risk management in major league baseball. *Managerial and Decision Economics*, 25(2):109–120, 2004.

- [4] Anthony Krautmann, Peter von Allmen, and Stephen Walters. Should players trust their agents? portfolio size and agency behavior in major league baseball. *Journal of Sport Management*, 32:1–12, 12 2017.
- [5] Anthony C. Krautmann and John L. Solow. The dynamics of performance over the duration of major league baseball long-term contracts. *Journal of Sports Economics*, 10(1):6–22, 2009.
- [6] Anthony C. Krautmann and Thomas D. Donley. Shirking in major league baseball revisited. *Journal of Sports Economics*, 10(3):292–304, 2009.
- [7] Heather M. O'Neill and Scott Deacle. All out, all the time? evidence of dynamic effort in major league baseball. *Applied Economics*, 51(38):4191–4202, 2019.
- [8] Gerald W. Scully. Pay and performance in major league baseball. *The American Economic Review*, 64(6):915–930, 1974.
- [9] Anthony Krautmann. What's wrong with scully—estimates of a player's marginal revenue product. *Economic Inquiry*, 37:369–81, 02 1999.
- [10] Anthony Krautmann. What is right with scully estimates of a player's marginal revenue product: Reply. *Journal of Sports Economics*, 14:97–105, 01 2011.
- [11] John L. Solow and Anthony C. Krautmann. Do you get what you pay for? salary and ex ante player value in major league baseball. *Journal of Sports Economics*, 21(7):705–722, 2020.
- [12] Sports Reference LLC. "Batter Value, 1985-2023." Baseball-Reference.com - Major League Statistics and Information. <https://www.baseball-reference.com/leagues/majors/year-value-batting.shtml>. 04/19/2024.
- [13] Spotrac LLC. "Player Transactions, 1985-2023." Spotrac.com - Major League Contract Data and Transactions. <https://www.spotrac.com/mlb/team/player-name/transactions/>. 04/19/2024.