3080 Project Part 1

Adam Chow

2023-03-28

Introduction

In this report, we are studying the impact of various factors on a baseball player's contribution to winning in the 2022 MLB season. For aggregating how much a play contributes to winning baseball games, I will be using the statistic Wins Above Replacement (WAR) for this calculation. WAR has been one of the premier statistics in recent years that has been taken under consideration by franchises and managers around the world to analyze a player's performance. Ever since its addition into statistics in 1982, WAR has been a mainstay in assessing whether a player's production is leading to wins or losses. The name, Wins Above Replacement, describes the meaning of this statistic: the stat measures a player's value in all facets of the game, largely offense since it is very rare to make such a significant impact defensively, by analyzing how many more wins he would be worth over a replacement-level player at the same position - a Minor League replacement or a readily available fill-in free agent. However, each position is weighed differently. This is because the "replacement-level" of each position varies. For example, if a shortstop and a first baseman contribute equal overall production for their respective teams, the shortstop will have a better WAR because that position generally sees a lower level of production from replacement-level players. WAR has gotten its supreme reputation as a statistic because it quantifies each player's value in terms of a specific number of wins. WAR also differentiates itself from other major statistics because it factors in a positional adjustment; it is well suited for comparing players who play multiple defensive positions.

In the scope of this project, I will be assessing different factors and statistics by qualified players during the 2022 MLB season. A qualified player for this dataset is a player with at least 400 Plate Appearances (PA). I chose this number because it signifies players that were considered "everyday players" by their managers and did not sustain major injuries that caused them to miss more than one third of the overall season. Using this dataset, I want to deduce whether there are trends in certain players that cause them to contribute to more winning - higher WAR - than others.

First, I want to take a look and see if weighted Runs Created plus (wRC+), one of the newest statistics developed by baseball analysts, is actually a solid indicator of a player's value to winning. Very new stats like wRC+ have begun to become a controversial topic across sports due to many players and coaches calling these stats misleading. Where some athletes may fill up the spreadsheet with gaudy numbers, many often do not lead to championships - a team's ultimate goal. Many athletes that have played the game insist that novel, formulaic stats veer athletes from the intangible parts of the game that made sports what "they used to be back in the day." I want to test whether new statistics promote winning or fail to account for the intangible qualities that a player might possess.

With the surge of home runs and strikeouts in today's MLB, the importance of stealing bases and baserunning in general has seemed to completely diminish, since both home runs and strikeouts involve no speed at all. Unlike many other North American sports like hockey and basketball, it is not uncommon to see a baseball player that looks to be completely out of shape and weighing abnormally large for a professional athlete. Seeing these player's poses the question: Can you be a successful baseball player while still being out of shape, and does speed really matter in baseball?

Lastly, baseball is known for being the one sport where height does not matter. Unlike sports like basketball and football, you can be a successful baseball player playing any position at any height. Despite this notion,

just in the past five years, researchers have discovered that the optimal height for pitchers to get batters out is 6'2". This has led to the median height in the MLB ballooning to 6'2" in 2023. Out of the nine positions, only one of the positions has an average height of under 6'. With all of these things being considered, can players under the median height still produce value for a team at the same rate as taller players?

Through the study of different factors and variables, we can observe which of these factors correlate with WAR in a constantly evolving sport. In pursuit of their championship aspirations, this study can be used by teams, coaches, and organizations to improve their performance on the field by pinpointing which attributes in a player contribute to more winning. With this in mind, the questions we are focusing on are as follows:

1. The Major League average wRC+ is 100. Do players above this mark have an overall higher WAR than hitters lower than this mark?
2. The Major League average Spd index is 4.3. Do players above this mark have an overall higher WAR than hitters lower than this mark?
3. The Major League median height is 6'2". Do players over 6'2" contribute to a higher WAR than players shorter than 6'2"?

For the owners and affiliates of these professional sports teams, the sport of baseball is a business to them. Each year, a team spends hundreds of millions of dollars on their players, coaches, stadium, etc. Creating an atmosphere that attracts viewers in the stadium seats as well as on national television is at the utmost priority of these owners. One surefire way to do this is to win games and play in the postseason, pursuing the World Series trophy. In this light, it is in these owners' best interest to decipher the best course of action to generate business.

Data Summary

This data from FanGraphs was collected on all players who played and got plate appearances in the 2022 MLB season. The data I collected are a combination of baseball total numbers, qualitative descriptors, and calculated statistics, depicting who the person is as a professional baseball player. Due to FanGraphs's pliability, I was able to choose which statistics and counting stats I thought were relevant enough to include in my dataset, since it is useless to include statistics that have no relevance to winning. For example, I included statistics like OPS, ISO, and wOBA, because these statistics help accurately analyze a player's offensive production in areas of power and on base percentage. On the other hand, I did not include common counting statistics like Hits, Singles, Walks, and Strikeouts. One example of how these counting numbers can be misleading is if one player possibly had 200 more at bats than another. I also had to import another dataset to incorporate both the WAR and physical attributes of the players - height and weight - since FanGraphs did not provide either of this information. Much of these statistics could have been chosen from a multitude of different websites because this data is very readily available, but FanGraphs, ESPN, and Baseball Reference were the three sources that I chose to compile the data from.

As for choosing the players to be in my dataset, I set the minimum number of at bats to be at 400. This way, the statistics had a large enough sample size to normalize itself and not be too skewed by either an uncharacteristic hot streak or cold streak by any given player. A minimum number of 400 at bats also guarantees that the player played at least two thirds of the season: meaning they were considered to be an everyday player by the manager and did not sustain any major injuries that caused them to miss substantial time. Because there were only 206 players that fit this criteria, I decided to use all of them in the dataset. This allows us to calculate population parameters. For example, we can calculate the true mean WAR of the players in the2022 MLB season instead of having to estimate it using a sample. This was something we took into consideration when deciding how to assess a MLB player's WAR. Because there are so many players in the MLB, our data set is still susceptible to variance.

In the future, I may consider taking a larger sample from the MLB across multiple seasons. Getting data from multiple seasons would account for some error in a possible outlier year of 2022, the first full season of Major League Baseball due to the shortages of Covid-19. As for the credibility of my sources, the data was

collected by the MLB by recording every outcome of every plate appearance, required by every major league team. This information is then posted daily on their official public scorebook, making this information very accessible and hard to defraud.

Data Dictionary

The variables I am including in my report are as follows -

1. Team (Qualitative variable) - The abbreviation of the affiliated team of the player
2. Height (Quantitative variable) - A players' listed height (feet, inches)
3. Weight (Quantitative variable) - A players' listed weight (pounds)
4. Age (Quantitative variable) - Age of all players (in years)
5. AB (Quantitative variable) - A players' recorded number of at bats (Plate
6. Appearances (PA) - Base on Balls (BB) - Hit by Pitch (HBP))
7. PA (Quantitative variable) - A players' recorded number of times they came up to bat
8. BB% (Quantitative variable) - Percentage of plate appearances that resulted in a BB
9. K% (Quantitative variable) - Percentage of plate appearances that resulted in a K
10. BB/K (Quantitative variable) - Ratio of the number of BBs over the number of Ks
11. AVG (Quantitative variable) - Ratio of the number of hits over the total number of at bats
12. OBP (Quantitative variable) - Ratio of the number of hits, BBs, and HBPs over the total number of at bats, BBs, HBPs, and sacrifice flies
13. WAR (Quantitative variable) - Since all observations in this dataset are position players, the formula for WAR is, (The number of runs above average a player is worth in his batting, baserunning and fielding + adjustment for position + adjustment for league + the number of runs provided by a replacement-level player) / runs per win (per. mlb.com) - will generally be the response variable for much of the EDA

Exploratory Data Analysis

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.1      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.4      v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(pander)
```

Numerical Summaries

```
mlb_data <- data.frame(read.csv("mlb_data.csv"))
order <- c("5-6", "5-7", "5-8", "5-9", "5-10", "5-11", "6-0",
           "6-1", "6-2", "6-3", "6-4", "6-5", "6-6", "6-7")
mlb_data$Height <- factor(mlb_data$Height, levels = order)
num_sum1 <- group_by(mlb_data, Height) %>%
  summarize(avg.Speed=mean(Spd), avg.exitvelo=mean(EV),
            avg.wRC=mean(wRC.), avg.WAR=mean(WAR))
pander(num_sum1)
```

Summary of Important Statistics grouped by Height

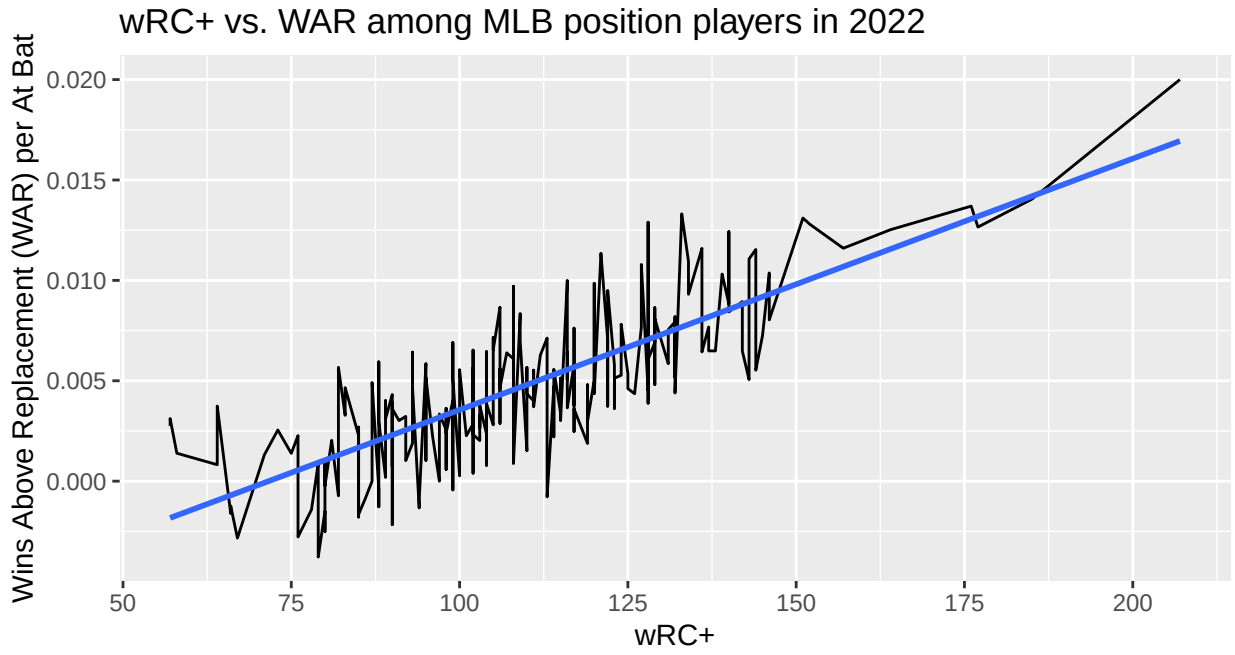
Height	avg.Speed	avg.exitvelo	avg.wRC	avg.WAR
5-6	5.05	85.15	127.5	4.05
5-7	6.1	87	117	2.5
5-8	2.55	88.95	113.5	2.6
5-9	4.971	87.99	119.9	4.1
5-10	4.286	87.77	96.67	1.943
5-11	3.8	88.17	102.5	1.909
6-0	3.858	88.77	105.8	2.197
6-1	4.303	89.29	105.5	2.309
6-2	3.461	89.04	112.7	2.503
6-3	3.736	89.95	122.2	3.036
6-4	3.214	89.6	108.9	1.862
6-5	2.85	91.17	117.5	2.413
6-6	0.9	95	115	1.2
6-7	4	95.8	207	11.4

Graphical Summaries

```
library(ggplot2)
mlb_data$WARperAB = mlb_data$WAR / mlb_data$AB
ggplot(mlb_data, aes(x=wRC., y=WARperAB)) +
  geom_line() +
  geom_smooth(method="lm", se=FALSE) +
  labs(title="wRC+ vs. WAR among MLB position players in 2022",
       x="wRC+", y="Wins Above Replacement (WAR) per At Bat") +
  scale_x_continuous(breaks=seq(0, 250, 25)) +
  scale_y_continuous(breaks=seq(-0.05, 0.020, 0.005))
```

Relationship between wRC+ and WAR among MLB position players in 2022

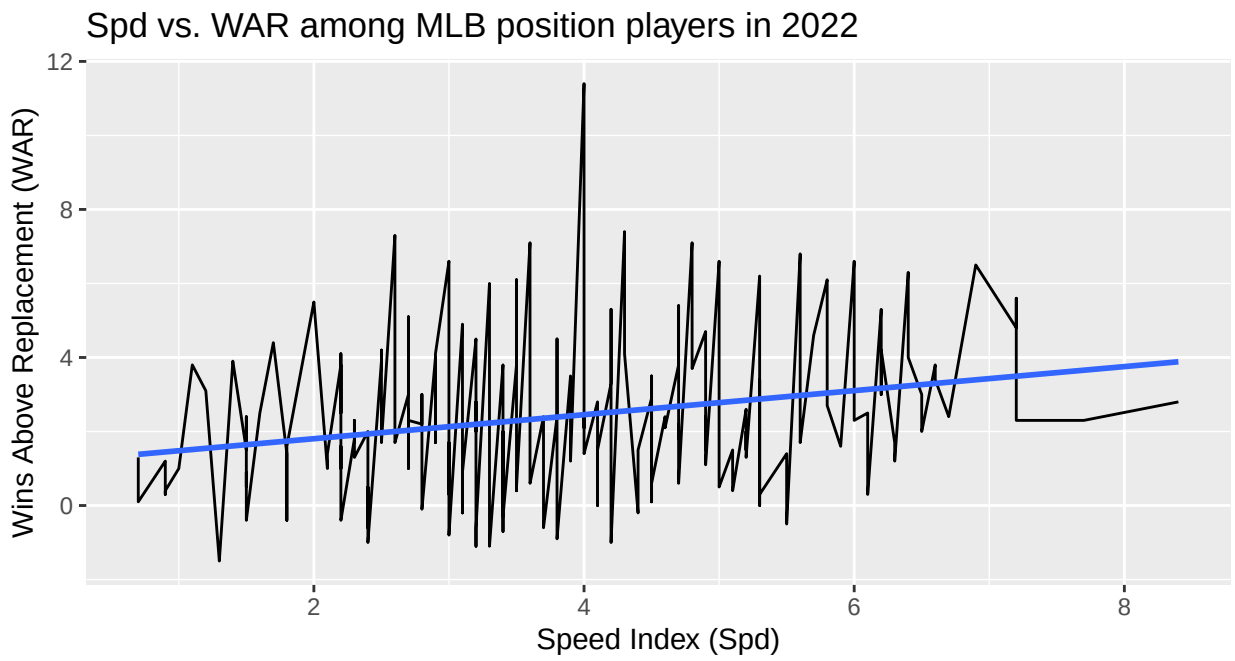
```
## `geom_smooth()` using formula = 'y ~ x'
```



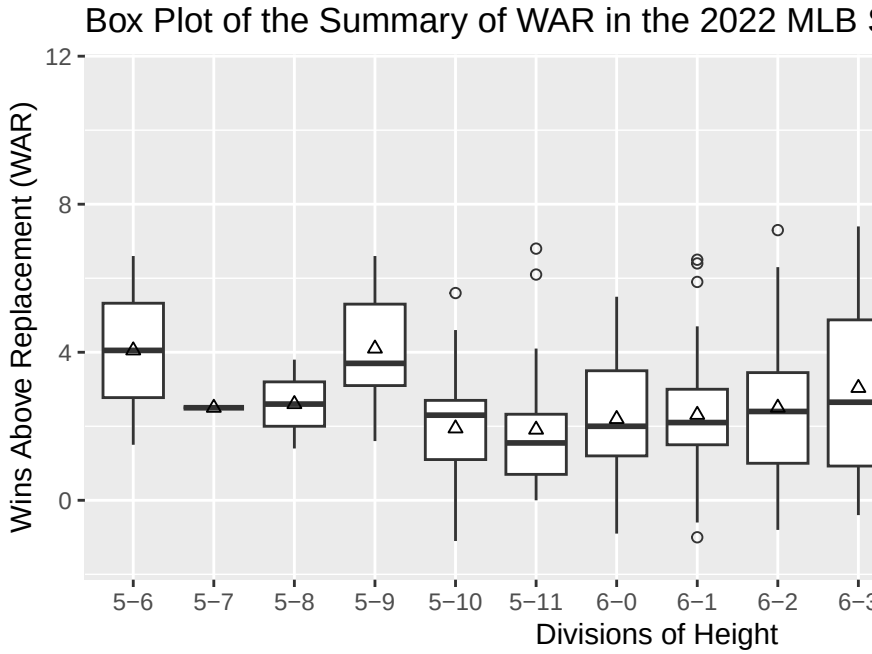

```
ggplot(mlb_data, aes(x=Spd, y=WAR)) +
  geom_line() +
  geom_smooth(method="lm", se=FALSE) +
  labs(title="Spd vs. WAR among MLB position players in 2022",
       x="Speed Index (Spd)", y="Wins Above Replacement (WAR)")
```

Relationship between Spd and WAR among MLB position players in 2022

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
ggplot(mlb_data, aes(x=Height, y=WAR)) + geom_boxplot(outlier.shape=1) +
  stat_summary(fun="mean", geom="point", shape=2) +
  labs(title="Box Plot of the Summary of WAR in the 2022 MLB Season by Height",
       x="Divisions of Height", y="Wins Above Replacement (WAR)")
```



Box Plot of wRC+ by Height in the MLB

Conclusion

Starting with the first numerical summary: columns representing the division of height, the number of players per height, average number of at bats, average isolated power, average on-base plus slugging, average speed index, average exit velocity, average weighted runs created plus, and average wins above replacement. As for my most important statistic, WAR, there is no clear trend or correlation between height and WAR from heights 5-10 to 6-4, there is a fairly steady slight increase in WAR, but then suddenly drops from 3.04 to 1.86 from 6-4 to 6-5. I believe the values of heights 5-6 to 5-9 and heights 6-5 to 6-7 to all generally be considered outliers because there are such a small sheer number of them. Heights 5-6, 5-9, and 6-7 all have noticeably high WARs compared to all of the other heights because they are skewed due to the small sample and the elite all-star caliber players that are in these height categories. The only category that developed a noticeable correlation with height was the exit velocity, or how hard the ball leaves the bat. Starting with 5-9, the avg.exitvelo statistic incrementally increases steadily. While it does not increase by much, it illustrates a more consistent strong correlation than any other statistic. However, this statistic does not directly correlate to wRC+ or WAR, two of the statistics that we are looking at to help assess value to winning.

From the scatterplots, I was able to conclude that there is a positive correlation with wRC+ and the amount of WAR per At Bat. While there is a large amount of variance between each level of wRC+ - generally ranging about 0.010 WAR/AB - there is a noticeable correlation between the two variables. However, when assessing the Spd index, there is almost no correlation at all between the Spd index and the amount of WAR a player has. There is a large outlier at the 4.0 mark - slightly above league average - which happens to be the 2022 AL MVP. Overall, for every integer increase in the Spd index, the average added WAR to a player is about 0.4 games. This conclusion makes me question whether the Spd index and WAR contribute any information to find an answer to what variables are the best indicators of winning.

For the set of box plots that are a summary of the WAR for each height. Overall, all of the median values of each height are all hovering between 1.8 and 2.8. Except for heights 5-6 and 5-9 which are around 4 because of the abnormally small sample size that was discussed in the first numerical summary. One other noticeable attribute of the box plots was that on every height except for 5-10, the mean of the WAR was larger than the median of the WAR. This reveals that there are more outliers in the upper half of the WAR per age and more than 50% of the players in the MLB actually fall under the 1.2 average WAR. This is due to the different amount of all-star/hall of fame caliber players that skew the averages upwards.

All together, I can conclude that the only real correlation so far of WAR is wRC+. However, the statistic is still flawed because wRC+ solely accounts for hitting while WAR is an encompassing statistic that includes hitting, baserunning, and fielding, hitting just being the largest contributor of all. While there are substantially more players in the MLB that are taller than 5-9, much of this data concludes that there is not a large statistical disadvantage in terms of adding winning value to a team by being shorter than the median MLB height of 6'2". Our data also indicates that the Spd index of a player has an extremely weak correlation to WAR. While the correlation is positive, it is extremely small and weak, indicating that there are plenty of successful and winning players that are not very fast.

Appendix

```
pander(head(mlb_data, 15))
```

Table 2: Table continues below

Name	playerid	Team	Height	Weight	Age	AB	PA
Aaron Judge	15640	NYY	6-7	282	30	570	696
Yordan Alvarez	19556	HOU	6-5	225	25	470	561
Paul Goldschmidt	9218	STL	6-3	220	34	561	651
Mike Trout	10155	LAA	6-2	235	30	438	499
Jose Altuve	5417	HOU	5-6	166	32	527	604
Freddie Freeman	5361	LAD	6-5	220	32	612	708
Manny Machado	11493	SDP	6-3	218	29	578	644
Nolan Arenado	9777	STL	6-2	215	31	557	620
Austin Riley	18360	ATL	6-3	240	25	615	693
Juan Soto	20123	- - -	6-2	224	23	524	664
Mookie Betts	13611	LAD	5-9	180	29	572	639
Joc Pederson	11899	SFG	6-1	220	30	380	433
Rafael Devers	17350	BOS	6-0	240	25	555	614
Shohei Ohtani	19755	LAA	6-4	210	27	586	666
Bryce Harper	11579	PHI	6-3	210	29	370	426

Table 3: Table continues below

BB.	K.	BB.K	AVG	OBP	SLG	OPS	ISO	Spd	EV
15.9%	25.1%	0.63	0.311	0.425	0.686	1.111	0.375	4	95.8
13.9%	18.9%	0.74	0.306	0.406	0.613	1.019	0.306	3	95.2
12.1%	21.7%	0.56	0.317	0.404	0.578	0.981	0.26	3.6	90.7
10.8%	27.9%	0.39	0.283	0.369	0.630	0.999	0.347	3.3	91.6
10.9%	14.4%	0.76	0.3	0.387	0.533	0.921	0.233	5	85.9
11.9%	14.4%	0.82	0.325	0.407	0.511	0.918	0.186	4.8	91.3
9.8%	20.7%	0.47	0.298	0.366	0.531	0.898	0.234	4.3	91.5
8.4%	11.6%	0.72	0.293	0.358	0.533	0.891	0.241	2.6	88.7
8.2%	24.2%	0.34	0.273	0.349	0.528	0.878	0.255	3	92.5
20.3%	14.5%	1.41	0.242	0.401	0.452	0.853	0.21	3.5	91
8.6%	16.3%	0.53	0.269	0.34	0.533	0.873	0.264	6	90.5
9.7%	23.1%	0.42	0.274	0.353	0.521	0.874	0.247	4	93.1
8.1%	18.6%	0.44	0.295	0.358	0.521	0.879	0.225	3.1	93.1
10.8%	24.2%	0.45	0.273	0.356	0.519	0.875	0.246	4.7	92.9
10.8%	20.4%	0.53	0.286	0.364	0.514	0.877	0.227	4.6	92.1

LA	HardHit.	wRAA	wOBA	wRC.	WAR	WARperAB
14.9	60.9%	82.1	0.458	207	11.4	0.02
12.3	59.8%	52.1	0.427	185	6.6	0.01404
15.7	46.9%	56.7	0.419	177	7.1	0.01266
24.6	50.3%	42.8	0.418	176	6	0.0137
16.1	29.5%	41.8	0.397	164	6.6	0.01252

LA	HardHit.	wRAA	wOBA	wRC.	WAR	WARperAB
13.6	48.0%	46.6	0.393	157	7.1	0.0116
16	49.0%	36.8	0.382	152	7.4	0.0128
21.7	38.9%	35.1	0.381	151	7.3	0.01311
12.9	50.8%	36.8	0.377	142	5.5	0.008943
9.1	47.3%	34.9	0.376	145	3.8	0.007252
18.6	44.7%	32.3	0.373	144	6.6	0.01154
14.8	51.8%	21.8	0.373	144	2.1	0.005526
11.3	50.9%	30.8	0.373	140	4.9	0.008829
12.1	49.8%	31.7	0.37	142	3.8	0.006485
11.8	47.9%	20.2	0.369	138	2.4	0.006486

Works Cited

Research Background Citations:

Goldstein, H. (2016, April 6). The Increasing Importance of Pitcher Height. The Hardball Times. https://tht.fangraphs.com/the-increasing-importance-of-pitcher-height/

Zaino, J. (2018, April 5). Does Height Matter in Baseball? Sports Warrior 365. https://sportswarrior365.com/does-height-matter-in-baseball/

Moore, D. (2019, June 20). Manager's View: What Role Does Speed Play in Today's Game? FanGraphs Baseball. https://blogs.fangraphs.com/managers-view-what-role-does-speed-play-in-todays-game/

Tennyson, J. (2022, March 25). Average Height of MLB Players 2023. Baseball Bible. https://www.baseballbible.net/mlb-player-heights/#:~:text=Average%20Height%20of%20MLB%20Players%202023,-The%20average%20height&text=The%20data%20for%20the%202023,being%20for%206%272%E2%80%B3>

Data Source Citations:

ESPN. (2022). MLB WAR Leaders. https://www.espn.com/mlb/war/leaders

FanGraphs. (2022). 2022 Batting Statistics. https://www.fangraphs.com/leaders.aspx?pos=all&stats=bat&lg=all&qual=y&type=1&season=2022&month=14&season1=2022&ind=0&team=0&rost=0&age=0&filter=&players=0&startdate=2022-01-01&enddate=2022-12-31

Baseball Reference. (n.d.). Baseball Reference. https://www.baseball-reference.com/