Sports Analytics

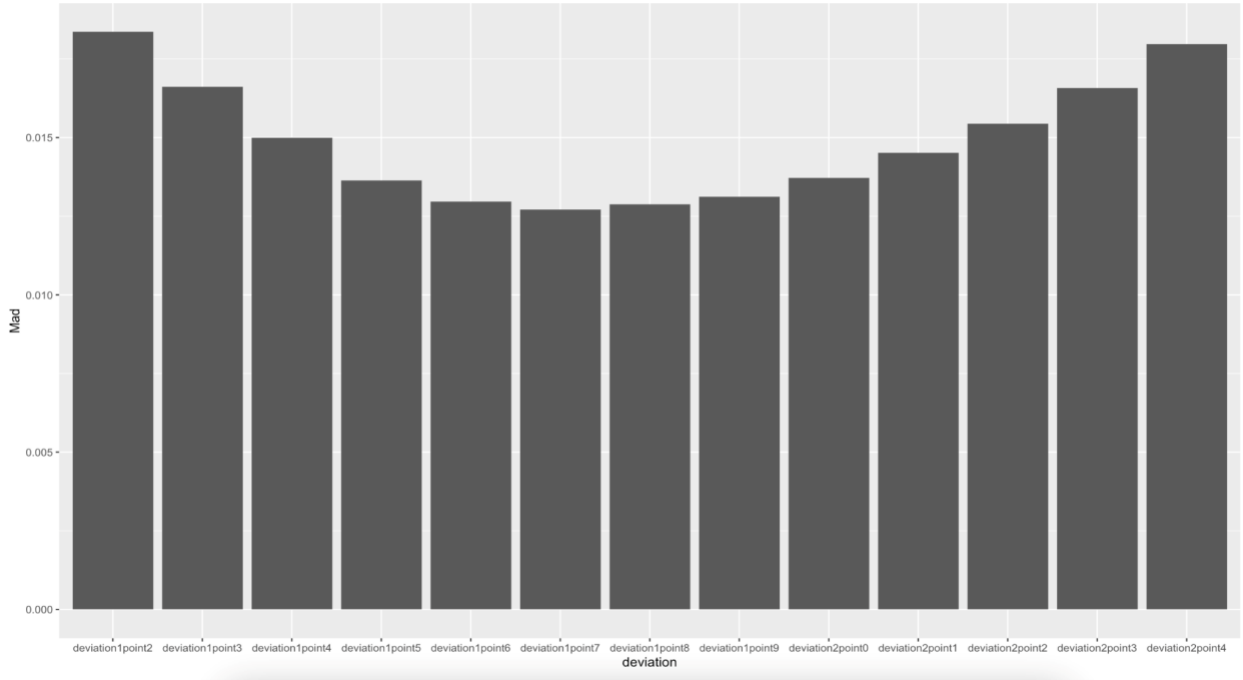
8/24

Team used for example is the Seattle Mariners

1. Determine the exponent in the Pythagorean Wins Formula that results in the lowest prediction error for games in recent history (you can decide how many years to include, last 5 years or 10 years, but include all major league games in that time frame). Are we performing better or worse than expected?

A: To determine the appropriate Pythagorean Win Formula exponent, data was scraped using the baseballR package, with the data coming directly from MLB. The final standings of the 2018, 2019, 2021, 2022, and 2023 season were used, with the 2020 season being excluded due to only being 60 games. Once team results from each year were scrapped, the data frame was reduced to the needed information: year, team, wins, losses, winning percentage, runs scored, and runs allowed. The equation runs scored/runs allowed was calculated and labeled as “R” from the data. With all the appropriate data, the Pythagorean Win Formula of R^x/((R^x)+1) was executed from numbers 1.2-2.4, the typical range for the exponent to fall between. Absolute error was found by subtracting the actual win percentage of a team and subtracting it from the result of every R^x/((R^x)+1) outcome for each exponent applied to each team's record between 2018-2023. From the data of absolute error, we could see the average deviation the absolute error from each yearly team result was from the median absolute deviation of that exponents results, in other words, we could see how big the variance was on average for each result and find which

R^x/((R^x)+1 led to the least amount of deviation from predicted team results to actual team results as a whole.



The graph above shows the Mean Absolute Deviation for each exponent, with the 1.7 exponent resulting in the lowest Mean Absolute Deviation. Thus, the appropriate Pythagorean Win Formula to use is R^1.7/((R^1.7)+1. With the MAD for R^1.7/((R^1.7)+1 being 0.01287426, that is about how much variation we can predict from actual team performance to predicted team performance on average.

The past five years have displayed a trend of the Mariners overall is that they have overperformed their Pythagorean Win Percentage 3 of the 5 years sampled, but underperformed it last year. The last two years saw the most variance, with the team underperforming by

0.089493113 from expected win percentage last year and over performing it by 0.072258751 the year prior. However, they do not even make the top 50 in teams who've experienced the most variance in any of the last five seasons. In comparison, the Mariners divisional rivals of the

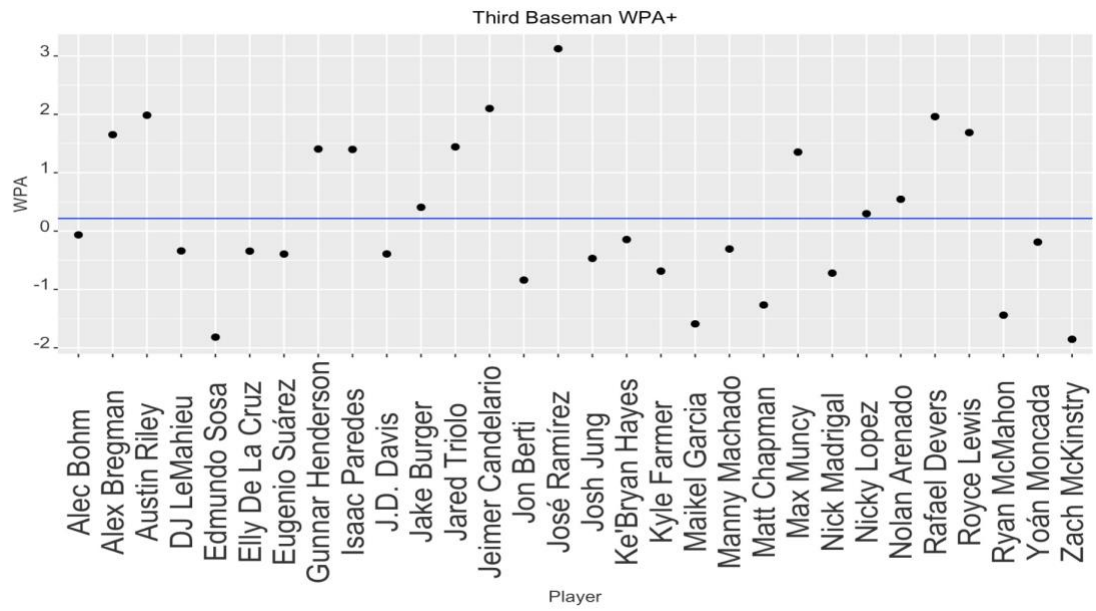
Houston Astros have three seasons in the top ten variance where they overperformed in 2018, 2019, 2022. Drastic variance typically comes from a team overperform expected numbers rather than underperform, so it is not entirely surprising to see it for a team that has finished first in the division to be consistently over performing while the Mariners, who have finished mainly second and third the past few years, to finish as expected.

1. Pick the three players with the most at-bats this year and determine where they rank at their position compared to the rest of the league. Pick 4 metrics to discuss with at least one coming from seasonal data, one coming from play-by-play data, and one coming from pitch-by-pitch data

Using data scraped by BaseballR from Fangraphs Batting Leaderboard, The three Mariners players with the most at bats in 2023 were Julio Rodriguez, Eugenio Suarez, and Teoscar Hernandez. Looking at the seasonal data Weighted Runs Created (wRC+), Weighted On Base Average (wOBA), Win Probability Added (WPA), and Contact Percentage. Weighted Runs Created gives an idea of how many runs a player contributed compared to the rest of the league, with 100 being average, and accounts for outside factors like park factors and league offense. It works close to a percentage, with a player with a 100 wRC+ performing exactly at average, one

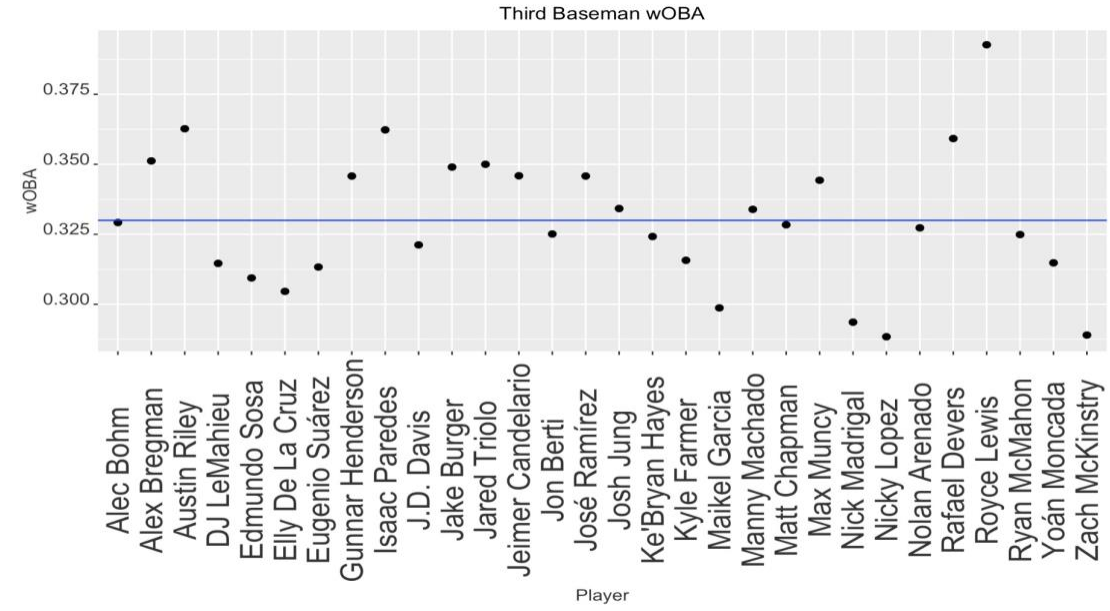
at 110 wRC+ performing 10% better, and one at 90 wRC+ performing 10% worse. The inclusion of park factor is especially an important reason to look at wRC+, as Seattle is an area greatly affected by park factor. Weighted Runs Created takes a lot from wOBA, which uses linear weights to assign offensive values to different ways to score based on how often the league converts each action to a run. This statistic gives us a full picture of the offensive contributions of a player. The play by play statistic selected is Win Probability Added, which helps us understand how said offensive contributions converted to actually helping the team. Taking into account each plate appearance, we know how their offense or lack of it helped or hurt the team and by how much. Finally, as a pitch by pitch statistic, we have contact percentage. This tells us how often a player makes contact with the ball both in and out of the zone. While contact doesn't necessarily have a 1:1 relationship with better offense and shouldn't be used to individually evaluate a player, it can help give context to a player's success or struggles to see how they compare with their counterparts.

Here is how the top thirty third baseman stacked up in the four selected stats.

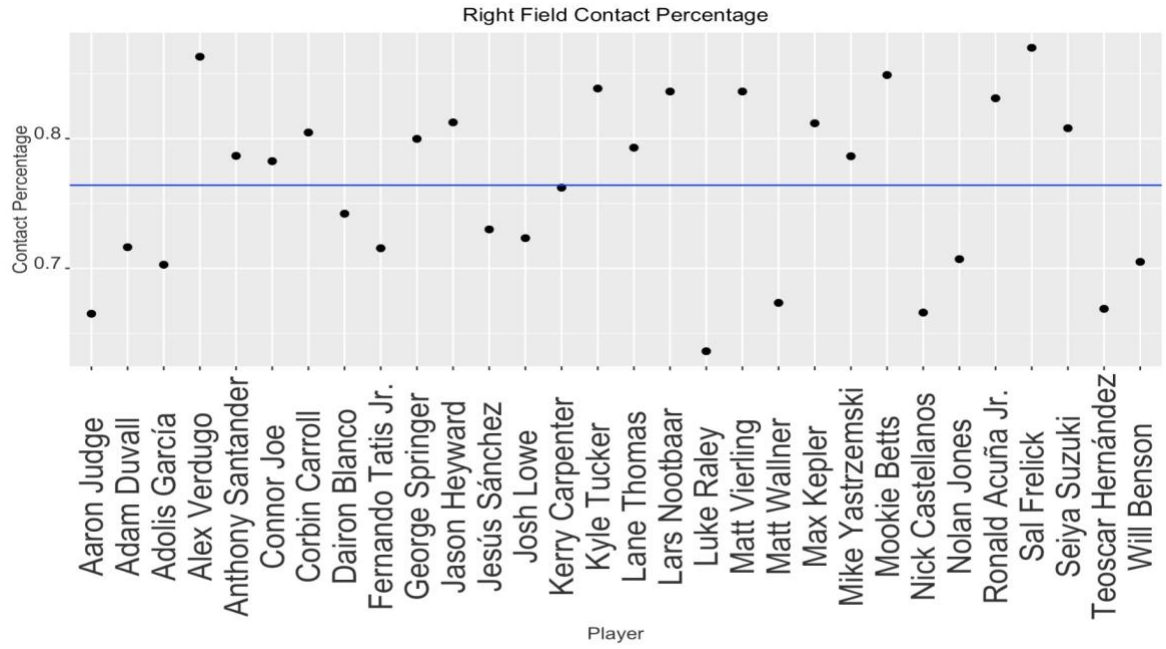

Third Baseman WPA+
WPA
3
2
1
0
-1
-2
Alec Bohm
Alex Bregman
Austin Riley
DJ LeMahieu
Edmundo Sosa
Elly De La Cruz
Eugenio Suárez
Gunnar Henderson
Isaac Paredes
J.D. Davis
Jake Burger
Jared Triolo
Jeimer Candelario
Jon Berti
José Ramírez
Josh Jung
Ke'Bryan Hayes
Kyle Farmer
Maikel Garcia
Manny Machado
Matt Chapman
Max Muncy
Nick Madrigal
Nicky Lopez
Nolan Arenado
Rafael Devers
Royce Lewis
Ryan McMahon
Yoán Moncada
Zach McKinstry
Player

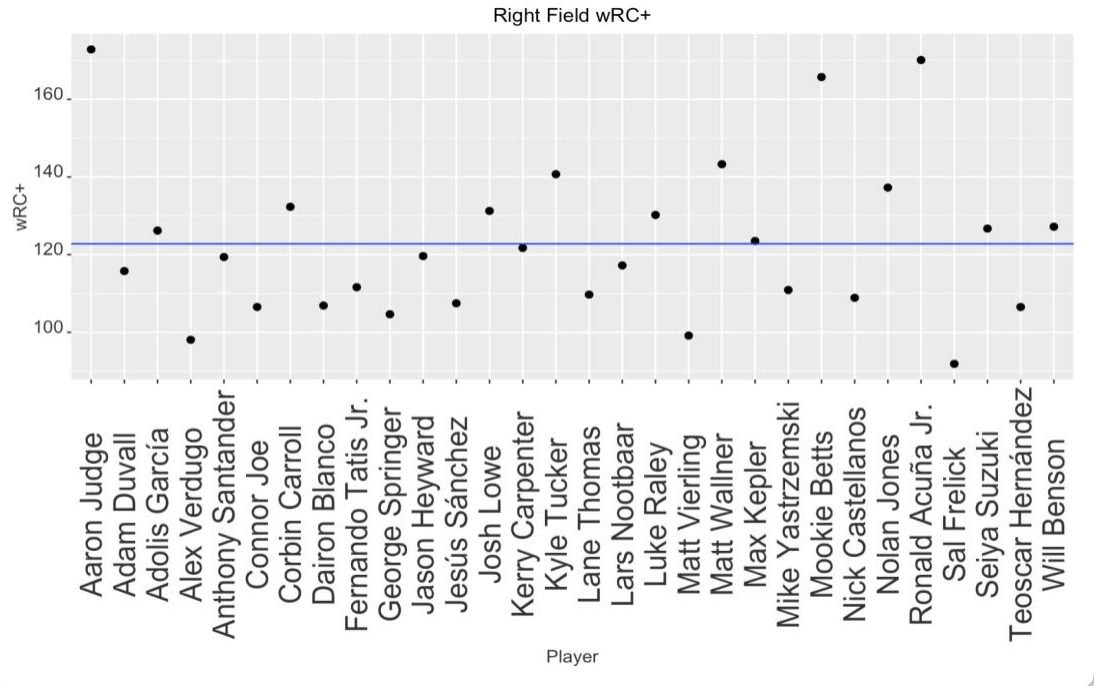
Contact Percentage
Contact Percentage
0.90
0.85
0.80
0.75
0.70
Alec Bohm
Alex Bregman
Austin Riley
DJ LeMahieu
Edmundo Sosa
Elly De La Cruz
Eugenio Suárez
Gunnar Henderson
Isaac Paredes
J.D. Davis
Jake Burger
Jared Triolo
Jeimer Candelario
Jon Berti
José Ramírez
Josh Jung
Ke'Bryan Hayes
Kyle Farmer
Maikel Garcia
Manny Machado
Matt Chapman
Max Muncy
Nick Madrigal
Nicky Lopez
Nolan Arenado
Rafael Devers
Royce Lewis
Ryan McMahon
Yoán Moncada
Zach McKinstry
Player

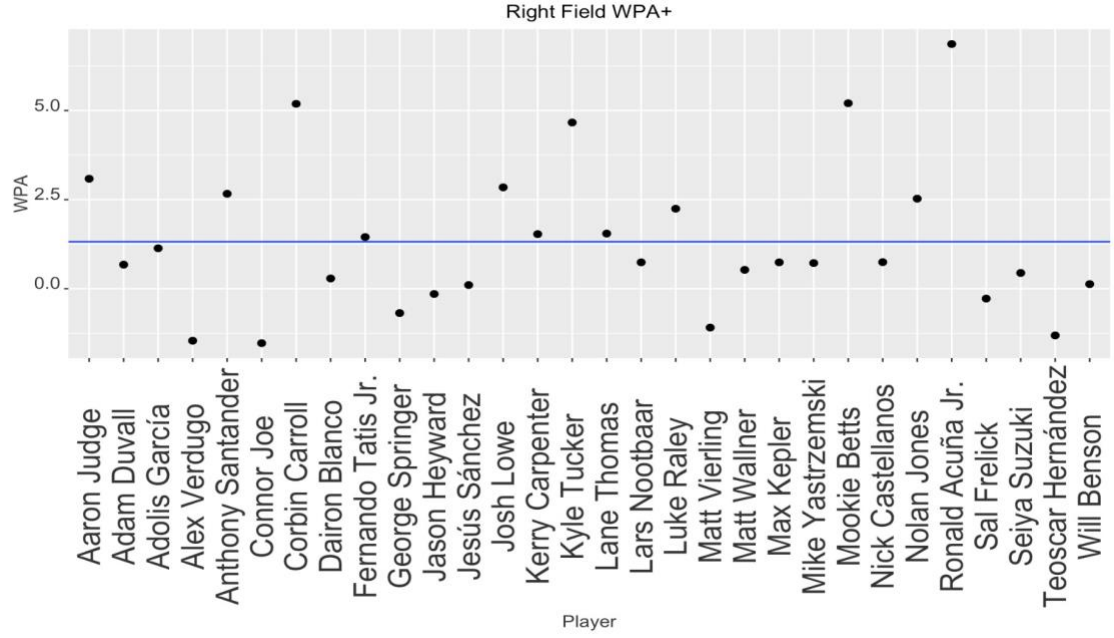
Third Baseman wRC+
wRC+
140
120
100
80
Alec Bohm
Alex Bregman
Austin Riley
DJ LeMahieu
Edmundo Sosa
Elly De La Cruz
Eugenio Suárez
Gunnar Henderson
Isaac Paredes
J.D. Davis
Jake Burger
Jared Triolo
Jeimer Candelario
Jon Berti
José Ramírez
Josh Jung
Ke'Bryan Hayes
Kyle Farmer
Maikel Garcia
Manny Machado
Matt Chapman
Max Muncy
Nick Madrigal
Nicky Lopez
Nolan Arenado
Rafael Devers
Royce Lewis
Ryan McMahon
Yoán Moncada
Zach McKinstry
Player

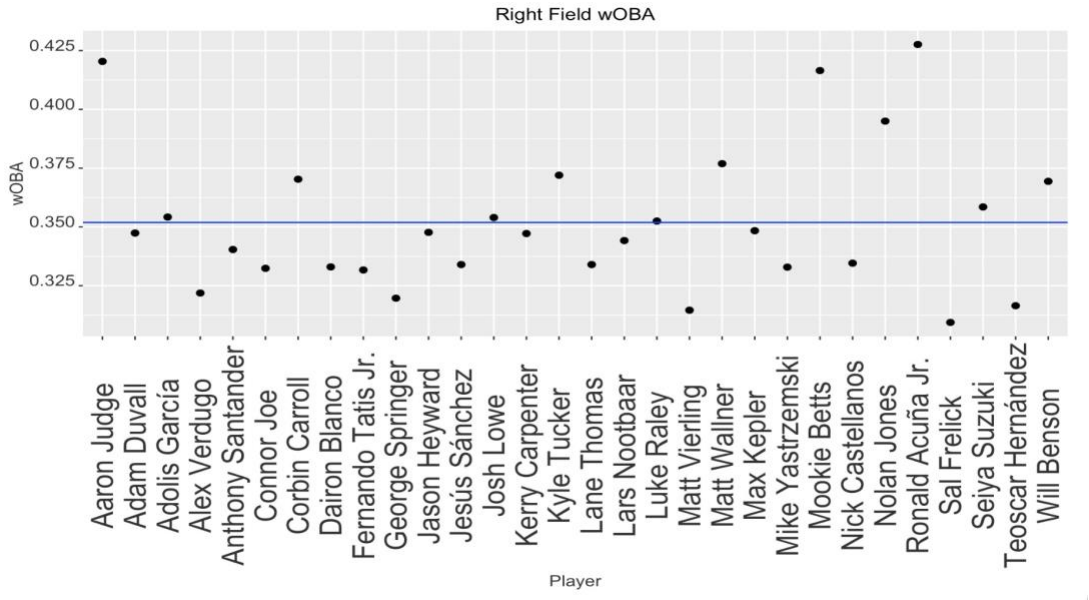


Here are the top thirty right fielder stacked up in the four selected stats.

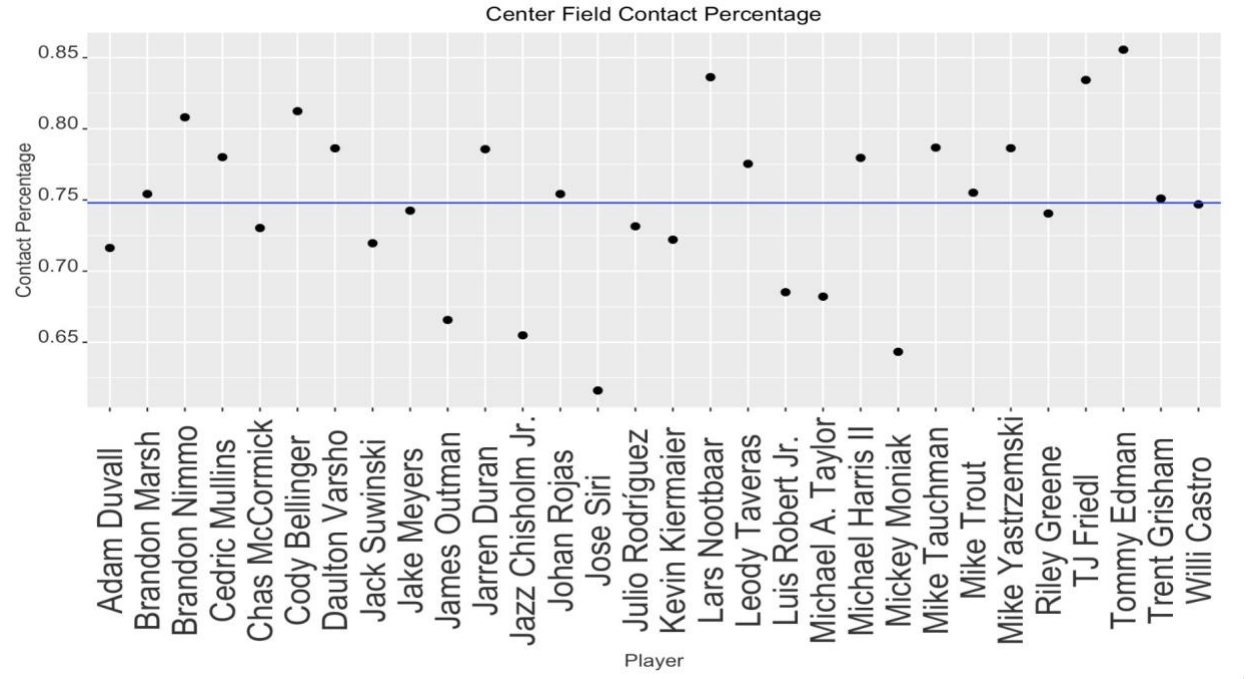



Right Field wRC+
wRC+
160
140
120
100
Aaron Judge
Adam Duvall
Adolis García
Alex Verdugo
Anthony Santander
Connor Joe
Corbin Carroll
Dairon Blanco
Fernando Tatis Jr.
George Springer
Jason Heyward
Jesús Sánchez
Josh Lowe
Kerry Carpenter
Kyle Tucker
Lane Thomas
Lars Nootbaar
Luke Raley
Matt Vierling
Matt Wallner
Max Kepler
Mike Yastrzemski
Mookie Betts
Nick Castellanos
Nolan Jones
Ronald Acuña Jr.
Sal Frelick
Seiya Suzuki
Teoscar Hernández
Will Benson
Player


Right Field WPA+
WPA
5.0
2.5
0.0
Aaron Judge
Adam Duvall
Adolis García
Alex Verdugo
Anthony Santander
Connor Joe
Corbin Carroll
Dairon Blanco
Fernando Tatis Jr.
George Springer
Jason Heyward
Jesús Sánchez
Josh Lowe
Kerry Carpenter
Kyle Tucker
Lane Thomas
Lars Nootbaar
Luke Raley
Matt Vierling
Matt Wallner
Max Kepler
Mike Yastrzemski
Mookie Betts
Nick Castellanos
Nolan Jones
Ronald Acuña Jr.
Sal Frelick
Seiya Suzuki
Teoscar Hernández
Will Benson
Player

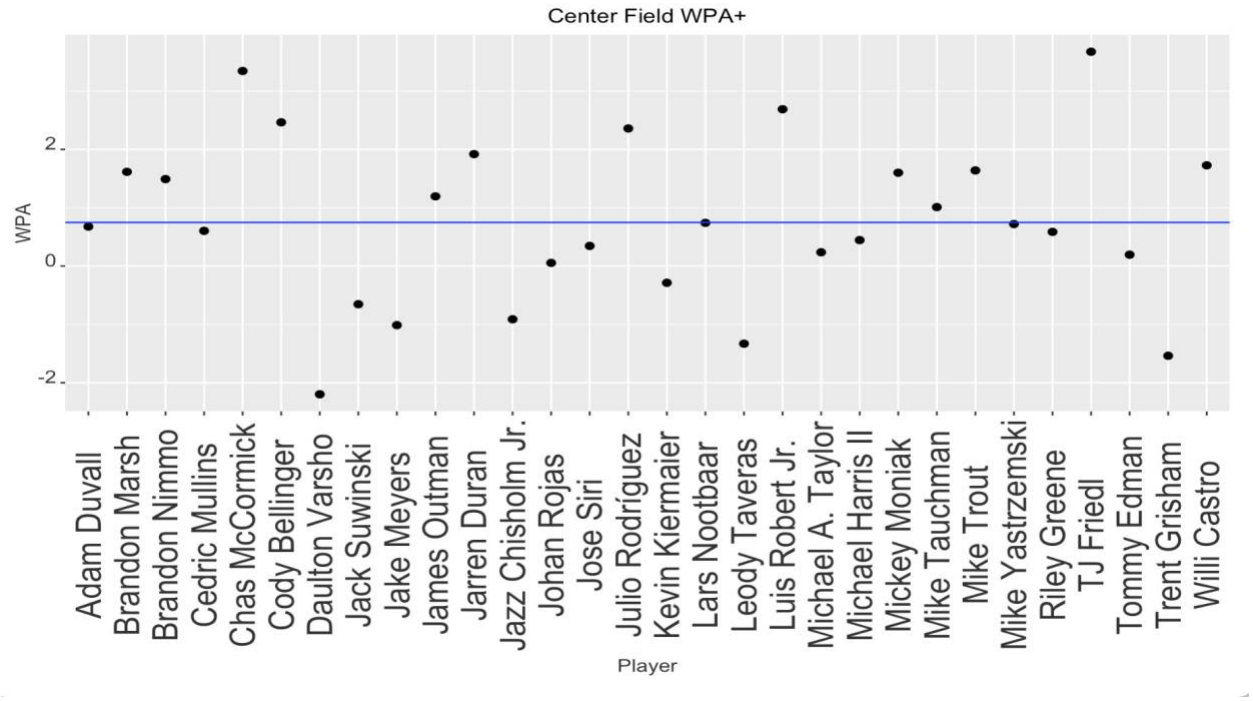


Here is how the top thirty center fielders for each of the selected stats.



Center Field wOBA
wOBA
0.37
0.35
0.33
0.31
0.29
Adam Duvall
Brandon Marsh
Brandon Nimmo
Cedric Mullins
Chas McCormick
Cody Bellinger
Daulton Varsho
Jack Suwinski
Jake Meyers
James Outman
Jarren Duran
Jazz Chisholm Jr.
Johan Rojas
Jose Siri
Julio Rodríguez
Kevin Kiermaier
Lars Nootbaar
Leody Taveras
Luis Robert Jr.
Michael A. Taylor
Michael Harris II
Mickey Moniak
Mike Tauchman
Mike Trout
Mike Yastrzemski
Riley Greene
TJ Friedl
Tommy Edman
Trent Grisham
Willi Castro
Player

Center Field wRC+
wRC+
130
120
110
100
90
Adam Duvall
Brandon Marsh
Brandon Nimmo
Cedric Mullins
Chas McCormick
Cody Bellinger
Daulton Varsho
Jack Suwinski
Jake Meyers
James Outman
Jarren Duran
Jazz Chisholm Jr.
Johan Rojas
Jose Siri
Julio Rodríguez
Kevin Kiermaier
Lars Nootbaar
Leody Taveras
Luis Robert Jr.
Michael A. Taylor
Michael Harris II
Mickey Moniak
Mike Tauchman
Mike Trout
Mike Yastrzemski
Riley Greene
TJ Friedl
Tommy Edman
Trent Grisham
Willi Castro
Player



The charts show that compared to his counterparts, center field is one of our strongest positions with well above average wRC+, wOBA, and WPA. Right fielder is the position that is struggling the most, nearly last in the league in all selected statistics. All three of our Mariners players taking the most at bats struggle with contact and are well below average compared to the rest of the league. While contact itself is not the end-all be-all evaluator for a productive offense, it gives some idea of an area the team as a whole may be lacking in if all three players are below average at contact and two are performing below league average in offensive production.

3. From here we are trying to determine the impact of adding a new player. Based on the findings in (2) pick a player to replace in our current lineup with another player from the same position group. Determine the additional runs we could expect based on seasonal data and possibly

through simulation. You can follow the approach outlined in Mathletics (Chapters 3 and 4) or Chapter 9 of Analyzing Baseball Data with R. You may also attempt to come up with a novel approach on your own.

Using the above findings, we will look for a right fielder replacement. To search for a replacement, the main evaluation criteria used was wOBA. The player we are looking at will be Nolan Jones, in part as a thought experiment to see how a player with an archetype similar to Teoscar Hernandez may compare in terms of run created. Like Hernandez, Jones is a high-strikeout, high-power corner outfielder, though he is a few years younger.

The data used is still the Fangraphs batting leaderboard, but the data will have to be manipulated somewhat. Total bases is essential for the formula needed for runs created, so a column with the formula 1B + (2B' * 2) + (3B' * 3) + (HR * 4) is used to calculate it for each player. After, we can calculate the total outs both Hernandez and Jone made with the formula ((0.982)* AB) Hits+ GDP + SF + SB + CS) / 26.72. Hernandez made 17.77 outs, whereas Jones made 10.38.
After, runs created can be calculated with (H + BB + HBP) * (TB)) / ((AB) + BB + HBP). Hernandez created 83.91 and Jones created 77.62 runs. We can divide these numbers by outs and see that per game, Jones is creating 7.48 runs per game, compared to Hernandez, who was creating 4.72 runs per game. Even though Hernandez is creating more runs as a whole, he is getting out significantly more than Nolan, which is why Nolan would create 2.76 runs a game

more for the Mariners as their right fielder. Thus, if the 2023 season was played with Nolan in right field instead of Hernandez, we could expect 2.76 runs a game more.

4. Describe what additional inputs or adjustments that you could make to help our department in the future. For instance, if we were facing a particular pitcher, what would you like to account for in your simulation that would provide a more accurate prediction?

A few major pieces of information should be considered before acquiring a new player. One of the major ones in this particular simulation is the park factor. The Seattle Mariners play in the toughest ballpark in terms of park factor, whereas a player like Nolan Jones plays in the most hitter friendly park. While there are currently stats that can take into account park factors (such as wRC+), fully understanding the negative effects that playing in a low run environment may due to certain archetypes of players is important for teams who experience such drastic circumstances as Seattle. Finally, another important factor to consider with a player like Nolan Jones who is younger is running a simulation with more data such as his zone heat map, runs created against specific types of pitches, and platoon splits to understand what situations he best thrives in and if that is an area of need for the team.