Finding alternative indicators for evaluation

- using regression model

Index

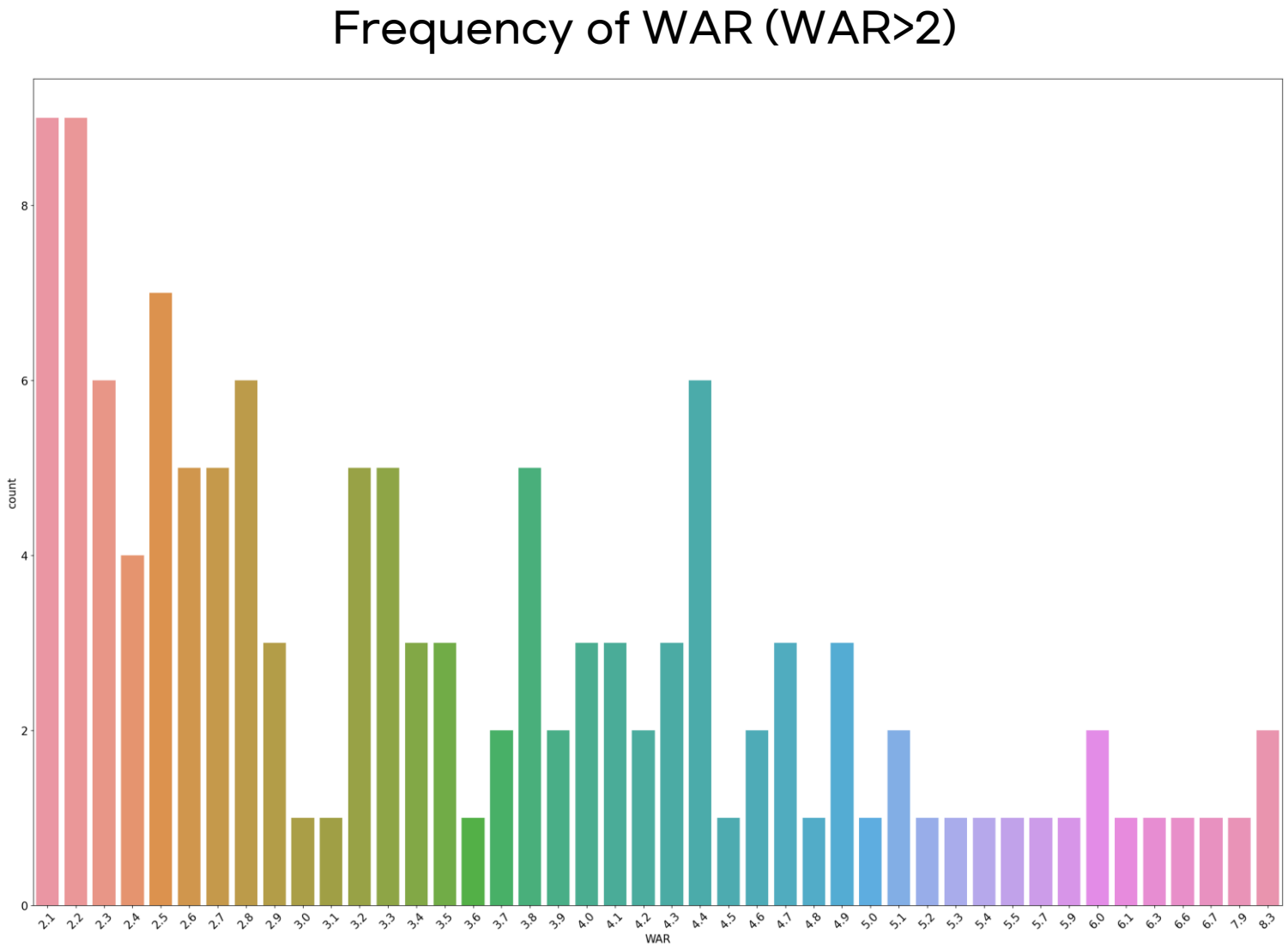
Definition of problem

WAR

: Win Above Replacement
(대체 선수 대비 득점 기여)

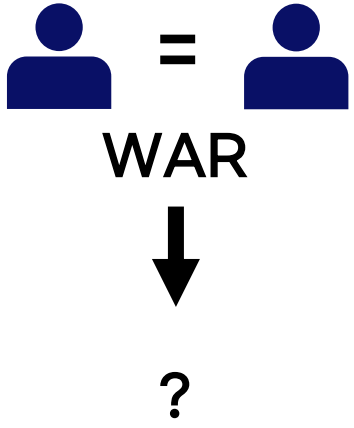
Q. If ‘WAR’ numbers of players are the same,
Which players should we recruit?

Definition of problem



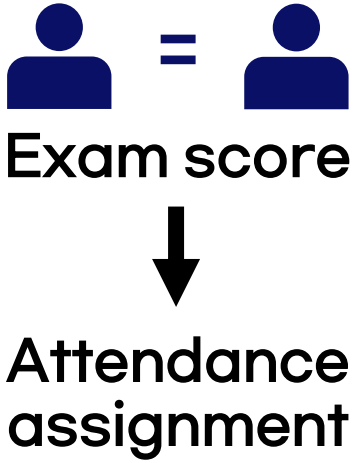
Definition of problem

If ‘WAR’ numbers of players are the same,
which players should we recruit?



Finding an indicator that are most
closely related to ‘WAR’

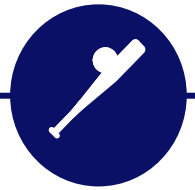
If students’ test scores are same,
Who will you give a higher grade to?



Finding an indicator that are
Most closely related to the exam score

Definition of problem

We think…



AVR
(batting average,
타율)

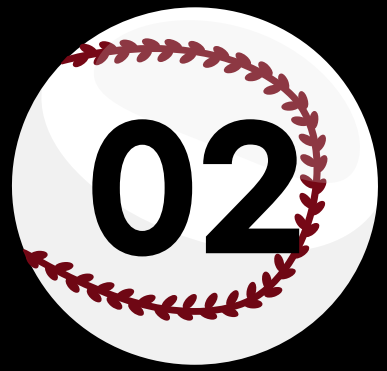
SLG
(slugging average,
장타율)

OBP
(on-base average,
출루율)

Will be closely related to ‘WAR’.
How?

Finding a correlation between WAR and all attributes in a data

Comparison in coefficient of attributes with high-correlation



Description of data

Description of data

Data attributes

- Recorded data
- Continuous
- Ordinal data
  (Games, scoring, …)
- Ratio data
  (Batting average, slugging, …)

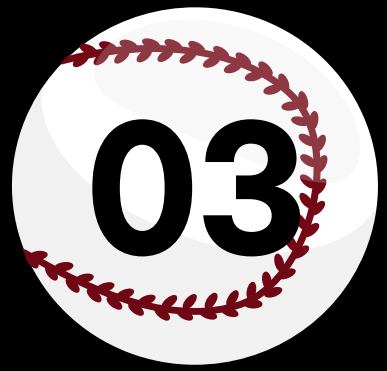
Description of data

Data attributes

- Recorded data
- Total 461 lows and 105 columns

```
df.info()
# df는 float형 77개, int형 26개, object형 2개로 이루어져 있다.
```

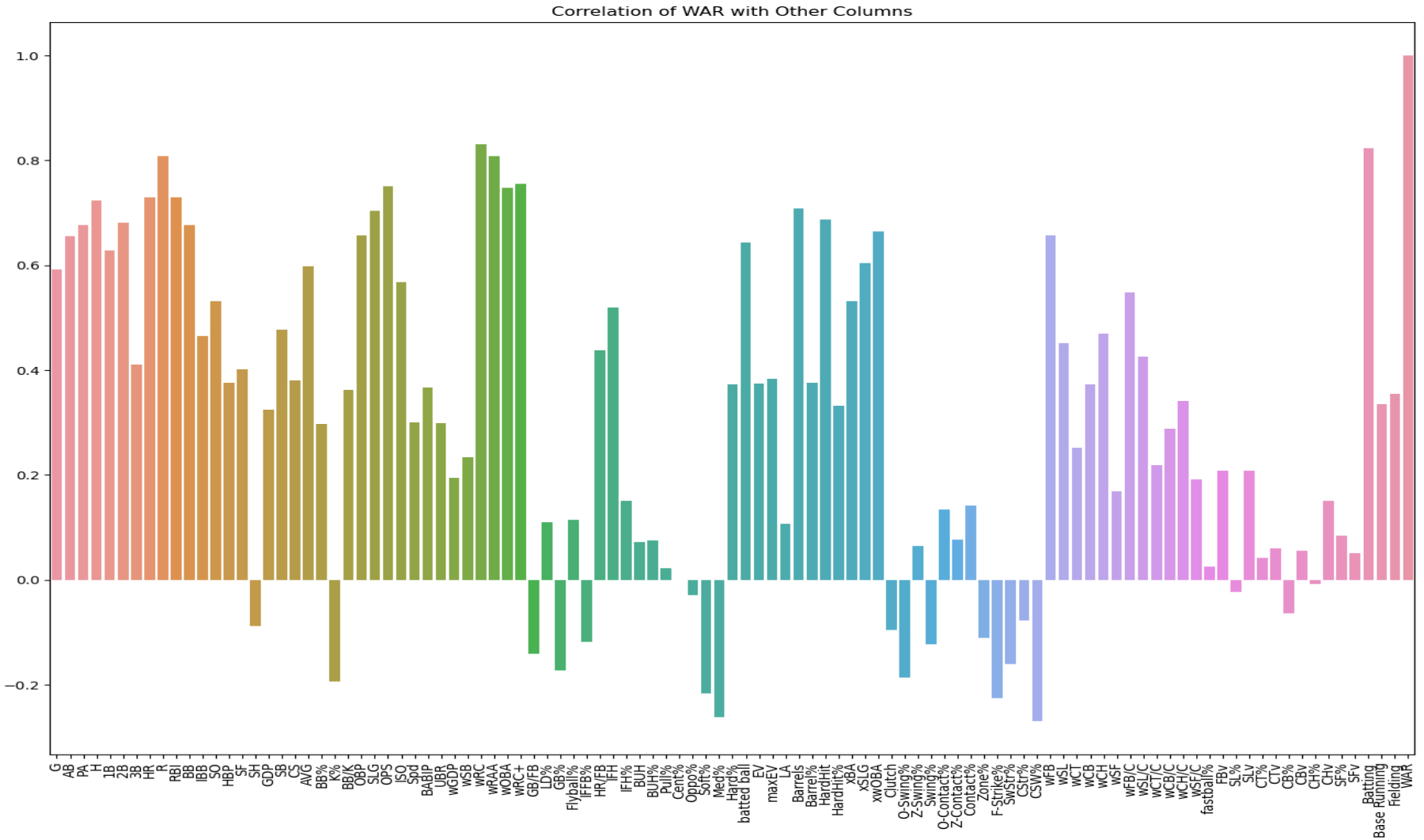
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 461 entries, 0 to 460
Columns: 105 entries, Name to WAR
dtypes: float64(77), int64(26), object(2)
memory usage: 378.3+ KB
```



Data exploration

Data exploration

First, we looked at the correlation with WAR.



Data exploration

	WAR
H	0.723380
HR	0.729618
R	0.807554
RBI	0.730223
SLG	0.704668
OPS	0.750966
wRC	0.830320
wRAA	0.808097
wOBA	0.747118
wRC+	0.755879
Barrels	0.707928
Batting	0.822758
WAR	1.000000

We decided to choose only columns with correlations Exceeding 0.7 with WAR.

Correlation codfficient	Degree of correlation
Less than +/- .2	Little correlation
Less than +/- .2 to .4	Low correlation
Less than +/- .4 to .7	Slightly high correlation
Less than +/- .7 to .9	High correlation
More than +/- .9	Very high correlation

Data exploration

SIMILAR

	H	HR	R	RBI	SLG	OPS	wRC	wRAA	wOBA	wRC+	Barrels	Batting	WAR
count	461.000000	461.000000	461.000000	461.000000	461.000000	461.000000	461.000000	461.000000	461.000000	461.000000	461.000000	461.000000	461.000000
mean	85.121475	12.368764	46.676790	45.028200	0.400998	0.714577	47.240781	0.863991	0.310679	94.952278	21.101952	0.955965	1.275488
std	45.980601	9.674632	27.268984	26.945462	0.076169	0.105525	30.426043	12.855483	0.041048	28.324016	16.007349	13.043989	1.762824
min	13.000000	0.000000	6.000000	5.000000	0.168000	0.322000	-3.000000	-26.200000	0.145000	-17.000000	0.000000	-30.600000	-2.000000
25%	47.000000	5.000000	24.000000	24.000000	0.351000	0.646000	22.000000	-7.200000	0.285000	78.000000	8.000000	-7.500000	0.000000
50%	78.000000	10.000000	43.000000	40.000000	0.400000	0.714000	40.000000	-1.900000	0.312000	96.000000	17.000000	-1.500000	0.900000
75%	120.000000	18.000000	65.000000	64.000000	0.453000	0.788000	68.000000	6.500000	0.340000	114.000000	30.000000	7.400000	2.200000
max	217.000000	54.000000	149.000000	139.000000	0.654000	1.066000	156.000000	66.700000	0.433000	180.000000	86.000000	64.100000	8.300000

- R(득점) : Record that counts when a runner enters a groove
- RBI(타점) : Record tallied by a batter who made it possible to score a goal

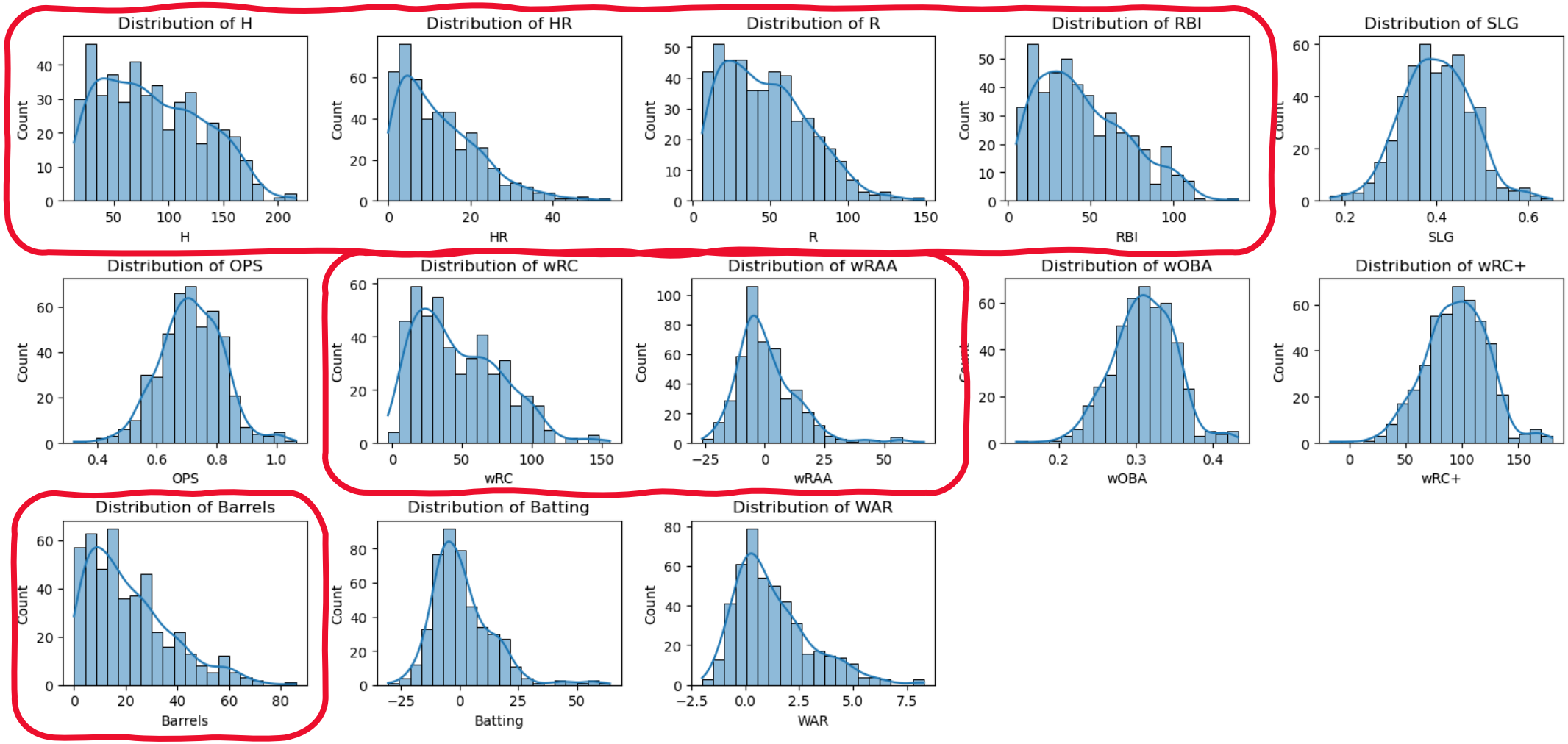
Data exploration

	H	HR	R	RBI	SLG	OPS	wRC	wRAA	wOBA	wRC+	Barrels	Batting	WAR
count	461.000000	461.000000	461.000000	461.000000	461.000000	461.000000	461.000000	461.000000	461.000000	461.000000	461.000000	461.000000	461.000000
mean	85.121475	12.368764	46.676790	45.028200	0.400998	0.714577	47.240781	0.863991	0.310679	94.952278	21.101952	0.955965	1.275488
std	45.980601	9.674632	27.268984	26.945462	0.076169	0.105525	30.426043	12.855483	0.041048	28.324016	16.007349	13.043989	1.762824
min	13.000000	0.000000	6.000000	5.000000	0.168000	0.322000	-3.000000	-26.200000	0.145000	-17.000000	0.000000	-30.600000	-2.000000
25%	47.000000	5.000000	24.000000	24.000000	0.351000	0.646000	22.000000	-7.200000	0.285000	78.000000	8.000000	-7.500000	0.000000
50%	78.000000	10.000000	43.000000	40.000000	0.400000	0.714000	40.000000	-1.900000	0.312000	96.000000	17.000000	-1.500000	0.900000
75%	120.000000	18.000000	65.000000	64.000000	0.453000	0.788000	68.000000	6.500000	0.340000	114.000000	30.000000	7.400000	2.200000
max	217.000000	54.000000	149.000000	139.000000	0.654000	1.066000	156.000000	66.700000	0.433000	180.000000	86.000000	64.100000	8.300000

- A big difference between the mean and the max

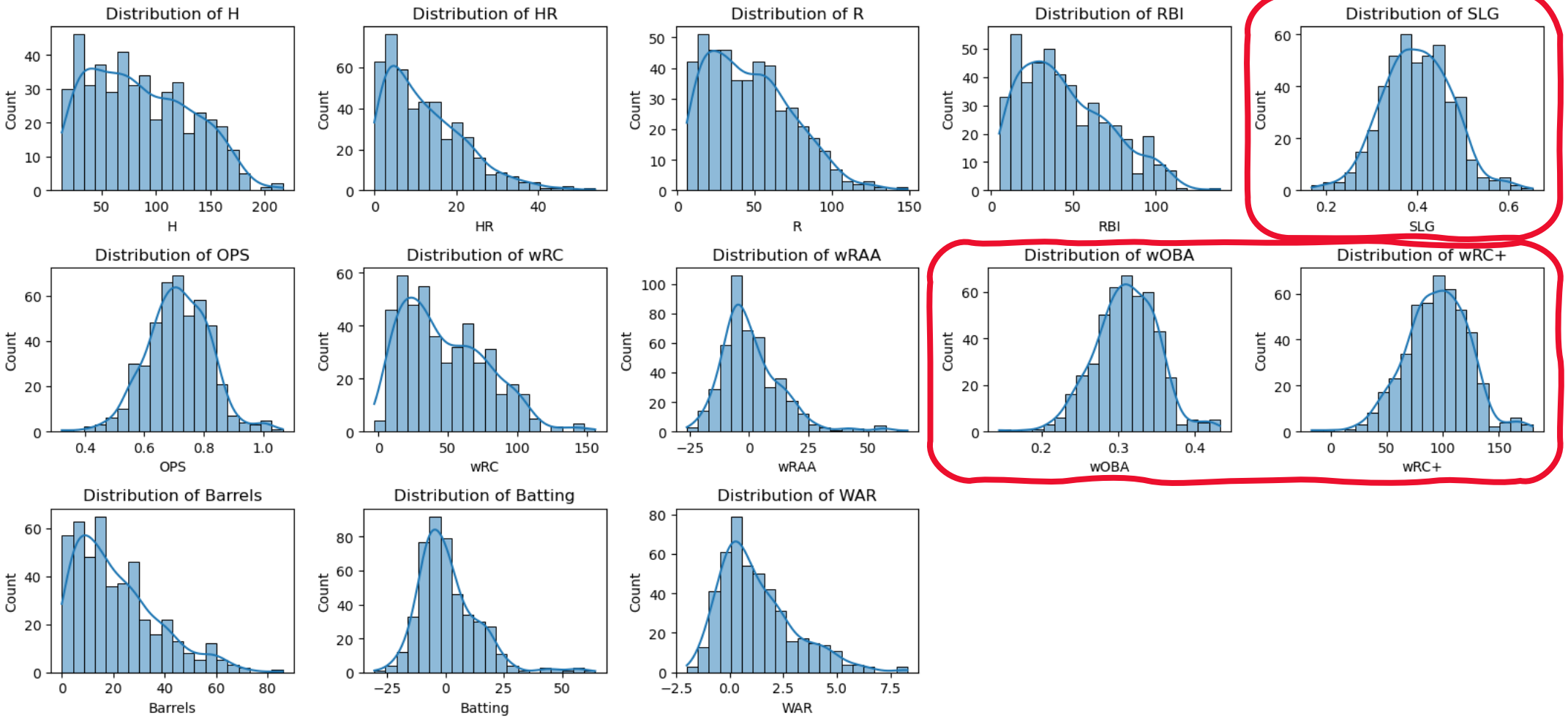
Data exploration

We looked at the histplot with the columns.



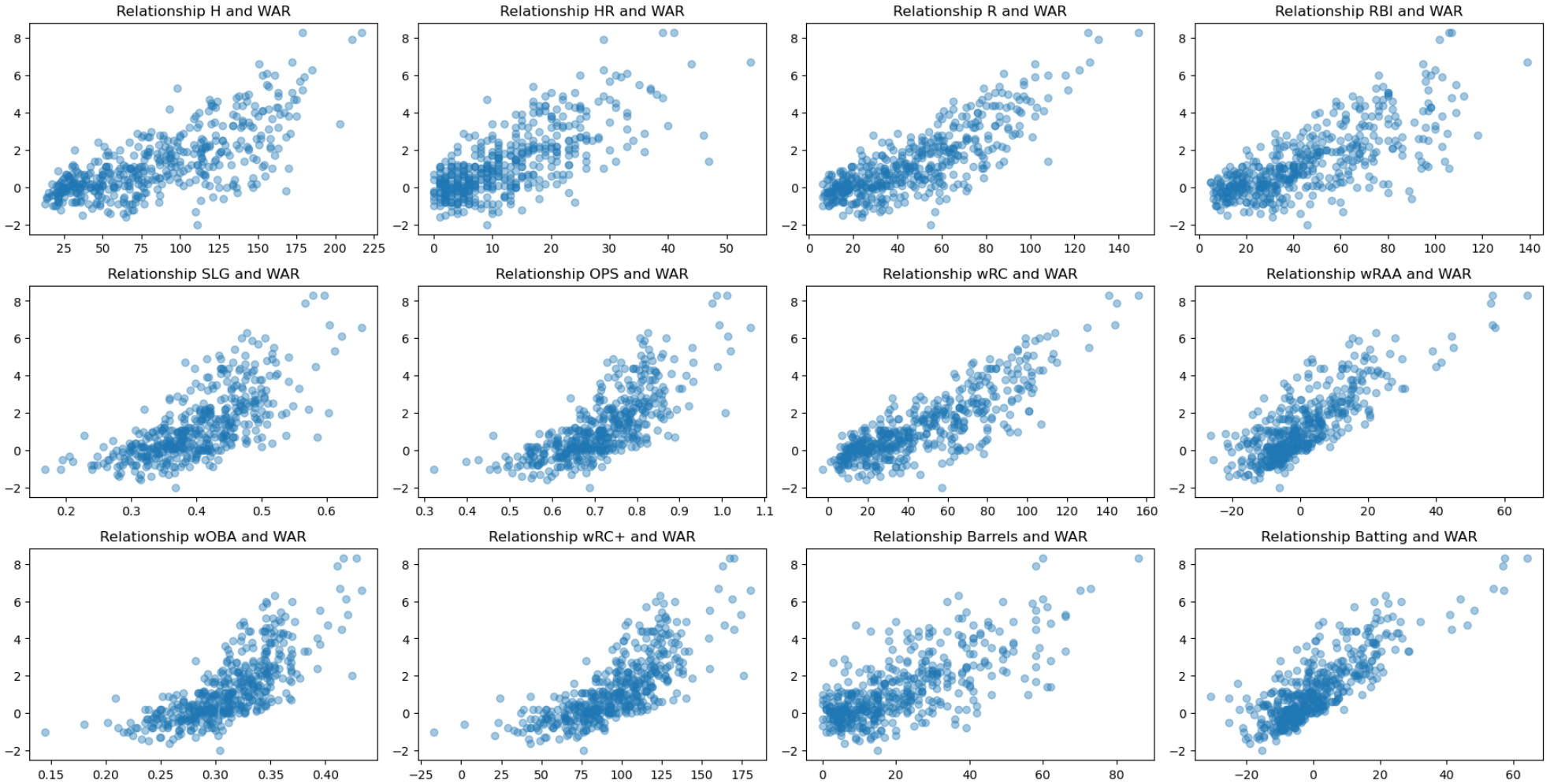
Data exploration

We looked at the histplot with the columns.



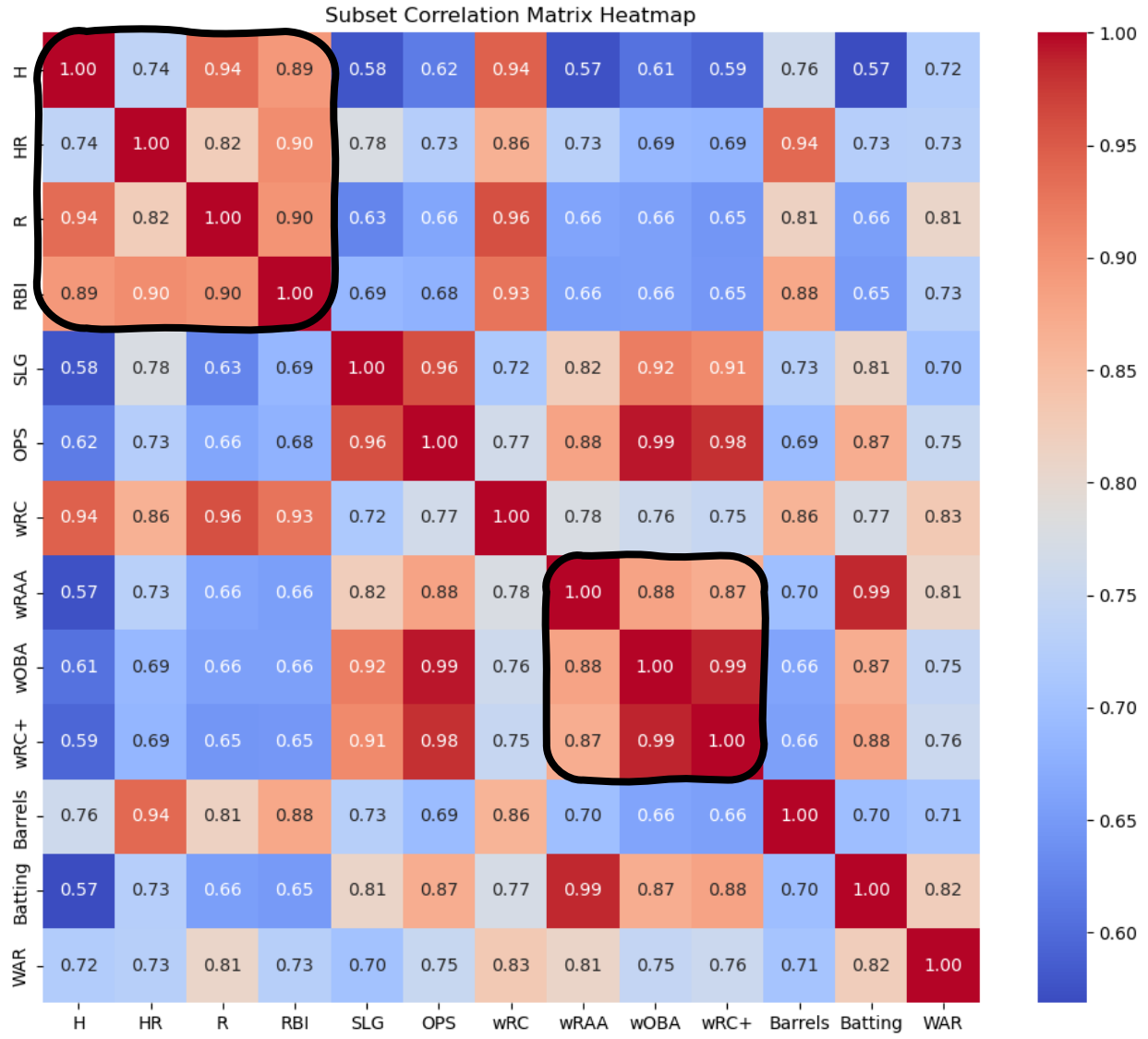
Data exploration

We looked at the relationship with WAR through a scatterplot.



Data exploration

We looked at the heatmap with columns.



Data analysis

Data analysis

Building the model

- 105 columns is a lot so features must be condensed to simplify model. Explore the low variance and low correlation columns to consider removing from model
- It looks like important metrics such as AVG and SLG appear to be low variance but still must be included in the model. We will not base our feature selection based on low variance.

Data analysis

1st model - Lasso

Results

Training accuracy score	0.822
Test accuracy score	0.769
Mean squared error	0.610
Mean absolute error	0.584
R squared score	0.769
Root mean squared error	0.781

Coefficients

Batting	0.0901080012455987 4
R	0.0456649179561784 9
wRC	-0.0311715207974260 4
H	0.0087969477554758 3
RBI	-0.0032877583563821 31
wRAA	-0.0032641937687765315
wRC+	0.0026377333704465187
Barrels	-0.0011100409631155355
HR	-0.0
SLG	0.0
OPS	-0.0
wOBA	-0.0

Data analysis

2nd model - Linear

Results

Training accuracy score	0.827
Test accuracy score	0.755
Mean squared error	0.645
Mean absolute error	0.588
R squared score	0.755
Root mean squared error	0.803

Coefficients

wOBA	30.65487407058719 4
OPS	-11.491822181736797
SLG	7.882623291564469
Batting	0.1818536028696 1263
wRAA	-0.089069072448 5791
R	0.0492929461200 7709
wRC	-0.031917324371600664
wRC+	-0.0195110407895 8332
HR	-0.014911433352540549
H	0.008920609728728488
Barrels	-0.00398485264094 3751
RBI	-0.002059890867092 4924

Data analysis

3rd model - Ridge

Results

Training accuracy score	0.826
Test accuracy score	0.757
Mean squared error	0.640
Mean absolute error	0.589
R squared score	0.755
Root mean squared error	0.780

Coefficients

SLG	1.490210265582046
OPS	0.5703542933467222
wOBA	0.2400322009001304
Batting	0.15259390054619842
wRAA	-0.05995301220905579
R	0.04979735036170926
wRC	-0.03688161866254072
H	0.010257816724129154
HR	-0.004609255237455193
Barrels	-0.0038966205540999057
wRC+	-0.003721278046407559
RBI	-0.0027241832093345213

Data analysis

We chose Ridge, because…

- In Lasso, the indicators we thought were important were zero. And the indicator that is not available in KBO is high.
- We thought that the value of Linear's OPS was outside the common sense of baseball.

Data analysis

Below are the results of test based on our previous data analysis.

Comparing real player data to our predicted 'WAR' results

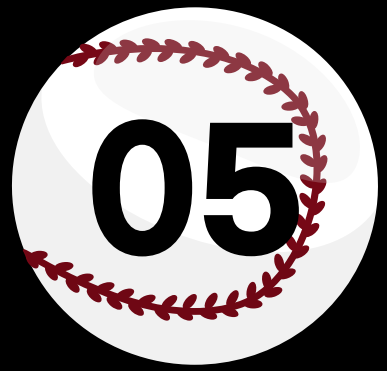
```
# Swap 'Name' value with any player we want to extract real data from
print(df[df['Name'] == 'Mookie Betts'][['H', 'HR', 'R', 'RBI', 'SLG', 'OPS', 'wRC', 'wRAA', 'wOBA', 'wRC+',
       'Barrels', 'Batting']])
```

```
       H  HR   R  RBI    SLG    OPS  wRC  wRAA   wOBA  wRC+  Barrels  Batting
136  179  39  126  107  0.579  0.987  141  56.5  0.416   167       60     57.3
```

```
# Swap 'Name' value with any player we want to extract real data from
print(df[df['Name'] == 'Ronald Acuña Jr.'][['H', 'HR', 'R', 'RBI', 'SLG', 'OPS', 'wRC', 'wRAA', 'wOBA', 'wRC+'
       'Barrels', 'Batting']])
```

```
      H  HR   R  RBI    SLG    OPS  wRC  wRAA   wOBA  wRC+  Barrels  Batting
84  217  41  149  106  0.596  1.012  156  66.7  0.428   170       86     64.1
```

	Name	WAR	Predicted WAR
0	Ronald Acuna Jr.	8.3	8.539526
1	Mookie Betts	8.3	7.341045



Conclusion

Conclusion

Coefficient about WAR : SLG > OPS > wOBA > Batting > R > H

Q. If 'WAR' numbers of players are the same, which player should we recruit?

A. We can recruit the players in the order shown above.

Ex) #1) 2 players with same WAR → recruit the player with higher SLG
if SLG is also same, then recruit the player with higher OPS

#2) 'A' player with higher record than 'B' in 4 parts, SLG, Batting, R and H
→ recruit A player

Conclusion

Apply to some cases - what should we consider?

1. Does ‘SLG’ figure have significant difference between 2 players?

2. How many parts(attributes) does a player record higher figures than the other player?

Coefficient about WAR : SLG > OPS > wOBA > Batting > R > H

Conclusion

Case in 2021 MLB - 1

Which player will you recruit?

name＼category	WAR	SLG	OPS	wOBA	Batting	R	H
Austin Riley	6.1	**0.531**	**0.898**	**0.379**	**30.5**	**91**	**179**
Brandon Crawford	6.1	0.522	0.877	0.377	27.7	79	144

War of **Austin Riley** in 2022 : **6.5**
War of Brandon Crawford in 2022 : 0.6

Coefficient about WAR : SLG > OPS > wOBA > Batting > R > H

Conclusion

Case in 2021 MLB - 2

Which player will you recruit?

name＼category	WAR	SLG	OPS	wOBA	Batting	R	H
Kyle Tucker	5.7	**0.557**	**0.916**	**0.383**	**31.8**	83	149
Cedric Mullins	5.7	0.518	0.878	0.372	29.4	**91**	**175**

War of **Kyle Tucker** in 2022 : **5.2**
War of Cedric Mullins in 2022 : 3.8

Coefficient about WAR : SLG > OPS > wOBA > Batting > R > H

Conclusion

Case in 2021 MLB - 3

Which player will you recruit?

name＼category	WAR	SLG	OPS	wOBA	Batting	R	H
Freddie Freeman	4.7	**0.503**	**0.896**	**0.379**	**31.8**	**120**	**180**
Jorge Polanco	4.7	**0.503**	0.826	0.349	18.7	97	158

War of **Freddie Freeman** in 2022 : **5.9**
War of Jorge Polanco in 2022 : 2.8

Coefficient about WAR : SLG > OPS > wOBA > Batting > R > H

Conclusion

Case in KBO - 1

In the 2017, Park & Choi's stats

NAME	WAR	SLG	OPS	wOBA	R
Park Gunwoo	**7.01**	**0.582**	1.006	0.433	91
Choi Hyungwoo	6.66	0.576	**1.026**	**0.442**	**98**

Park had a superior WAR, but Choi's every stats except SLG were better than Park's.
And the 2017 Golden Glove winner was Choi.

Conclusion

Case in KBO - 2

In the 2023 Jung & Oh's stats

NAME	WAR	SLG	OPS	wOBA	R
Jung Soobin	**3.89**	0.371	0.746	0.349	**75**
Oh Jihwan	**3.89**	**0.396**	**0.767**	**0.360**	65

Their WARs are same, but Oh is ahead in all stats except scoring.
And Oh is a strong candidate for the 2023 Golden Glove award.

Conclusion

Case in KBO - 3

In the 2020, there were two FA contract of shortstop players.
Below is the two-year average stats for the two players before FA contracts.

NAME	WAR	SLG	OPS	wOBA	R
Kim Sunbin	**3.12**	0.407	0.739	0.359	71
Oh Jihwan	2.78	0.394	0.734	0.349	68

Kim had a higher WAR,
But the rest of the stats had no significant difference between the two players.
LG Twins scouters thought Oh could play as well as Kim,
And signed FA contract for 4 billion won per 4 years, the same amount as Kim.

Conclusion

Case in KBO - 3

And these are their average stats of 3 years since their FA contrasts.

NAME	WAR	SLG	OPS	wOBA	R
Kim Sunbin	2.74	0.355	0.747	0.346	46
Oh Jihwan	**4.83**	**0.433**	**0.762**	**0.366**	**67**

If you look at this table, you can see that Oh is ahead of Kim in all stats.
Oh proved that the LG Twins scouters were right in their eyes.
Judging from this case,
It can be seen that SLG, OPS, wOBA, and R can be the criteria for judging players
As auxiliary indicators of WAR when scouting.

Thank you