Kimchi and Curveballs: Six Weeks in South Korea Scouting Baseball Players

Leonardo C. Gentile

Williams College Class of 1945 Fellowship Recipeint

June 2024 - August 2024



Sajik Baseball Stadium, Busan - Home of the Lotte Giants

Contents

1 Introduction **2**

2 Literature Review **3**
2.1 WAR in Korean Baseball . . . 3
2.2 Pitching Metrics and Sequencing . . . 3
2.3 Park Factors in the KBO . . . 3

3 Methodology **4**
3.1 Data . . . 4
3.2 Exploratory Data Analysis . . . 4
3.3 Regression Analysis . . . 4
3.4 Model Selection . . . 5

4 Findings **6**
4.1 WAR Distribution and Summary . . . 6
4.2 Factor Analysis . . . 6
4.3 Correlation and Multicollinearity . . . 6
4.4 Regression Results . . . 6

5 Discussion and Conclusion **7**

6 References **10**

Appendix A: WAR Distribution Analysis **11**
Summary Statistics . . . 11
WAR Histogram . . . 12
Sampling Distribution . . . 13
WAR by Role . . . 14
WAR by Team . . . 16

Appendix B: Factor Analysis **18**
ANOVA Results . . . 18
WAR by Position . . . 19
WAR by Handedness . . . 20
WAR by Park Factor . . . 22
Top Pitchers by WAR . . . 23

Appendix C: Correlation and Multicollinearity **25**
Correlations with WAR . . . 25
Key Scatter Plots . . . 27
Highly Correlated Pairs . . . 31

Appendix D: Regression Analysis **32**
Yeo-Johnson Transformations . . . 32
Full Model . . . 33
Stepwise Model Selection . . . 34
Model Comparison . . . 35
VIF and Multicollinearity Reduction . . . 36
Final Model Diagnostics . . . 38
Top Predicted Pitchers . . . 40

1 Introduction

Pitching is the cornerstone of a successful baseball team, as preventing runs is critical to winning. While offense is necessary, acquiring consistent, long-term pitching depth is far more challenging. Teams that contend for the World Series typically feature above-average offenses, two to three ace-caliber starting pitchers, and a deep, reliable bullpen—what I define as "the blueprint to baseball success."

A prime example is the 2009 New York Yankees, who won the World Series behind a dominant rotation of CC Sabathia, Andy Pettitte, and A.J. Burnett, supported by closer Mariano Rivera. Similarly, the 2019 Washington Nationals rode elite pitching, notably from World Series MVP Stephen Strasburg, to secure a championship.

Scouting pitchers requires evaluating pitch mix, arm talent, injury history, competition level, and defense-independent pitching statistics. Given the growing international talent pool, particularly from East Asia, I have focused on the Korean Baseball Organization (KBO) to identify potential MLB-caliber pitchers.

Recent high-profile signings of Japanese pitchers Yoshinobu Yamamoto and Shōta Imanaga highlight the increasing value of international arms. The KBO has also produced MLB talent such as Ryu Hyun-jin and Kim Ha-seong. Using my background in Statistics and Computer Science, I aim to develop a linear regression model utilizing KBO player data to determine which pitching metrics correlate most strongly with WAR, a key indicator of player value. This analytical approach could uncover hidden gems in the KBO.

However, scouting extends beyond numbers. To enhance my analysis, I traveled to South Korea for six weeks, funded by the Williams Fellowship Office, to observe pitchers firsthand and refine my scouting methodology.

2 Literature Review

2.1 WAR in Korean Baseball

WAR—Wins Above Replacement—is a measure of a player's value represented by how many more wins he is worth than a standard, MLB-average replacement player. In Korean baseball, WAR has been increasingly adopted as a key indicator of pitcher value, allowing analysts and scouts to compare players across roles and teams within the KBO. My dataset tracks WAR for all qualified KBO pitchers, and the distribution of WAR across 162 pitchers yields a mean of 0.878 and a standard deviation of 1.120, with a range from -0.600 to 4.840. These figures reflect the wide variation in pitcher quality across the league and are consistent with the expectation that a replacement-level pitcher contributes near zero wins above the baseline.

2.2 Pitching Metrics and Sequencing

Scouting pitchers requires evaluating pitch mix, arm talent, injury history, competition level, and defense-independent pitching statistics. Key metrics include ERA (Earned Run Average), FIP (Fielding Independent Pitching), WHIP (Walks and Hits per Inning Pitched), SIERA (Skill-Interactive ERA), strikeout rate, and walk rate, among others. FIP focuses solely on the events a pitcher has the most control over—strikeouts, walks, hit-by-pitches, and home runs—removing results on balls hit into the field of play entirely. SIERA builds on FIP by also adjusting for the type of ball in play. The correlations between these metrics and WAR form the empirical foundation of the regression models constructed in this study.

In my analysis, strikeouts had the highest positive correlation with WAR ($r = 0.884$), followed by innings pitched ($r = 0.861$) and plate appearances ($r = 0.824$). ERA and WHIP had the highest negative correlations, at -0.604 and -0.669 respectively. These relationships confirm that high strikeout and inning totals are associated with higher WAR while higher ERAs and WHIPs are associated with lower WAR, which is all to be expected.

2.3 Park Factors in the KBO

Home Park Factor (HPF) measures how much a specific ballpark contributes to offensive production, calculated as:

$$\text{HPF} = 100 \times \frac{(\text{homeRS} + \text{homeRA})/\text{homeG}}{(\text{awayRS} + \text{awayRA})/\text{awayG}}$$

A park factor above 1000 favors hitters; below 1000 favors pitchers. In the KBO, park factors range from 940 to 1054 across the ten home stadiums. Although boxplots of WAR by home park factor did not reveal any statistically significant differences (ANOVA $p = 0.987$), pitchers playing in more hitter-friendly parks tend to show slightly higher mean WAR values—a counterintuitive pattern that may reflect the confounding influence of team quality and sample size. The park factor variable was nonetheless included as a categorical factor in the exploratory analysis for completeness.

---

3 Methodology

3.1 Data

My dataset consists of pitching data from July 23rd, including all KBO pitchers who had thrown at least 15 innings that season. I set this cutoff to minimize inflated statistical rates from low-inning totals, ensuring model accuracy. This resulted in 163 qualified pitchers out of 257. The dataset tracks 43 stats for each pitcher, with many definitions sourced from the official MLB Glossary.

Once all the data was loaded into RStudio, I first converted my categorical variables (Team, Handedness, Position, and Home Park Factor) into factors and saved my continuous variables as a new object. Next, I extracted the dependent variable, WAR, and computed its mean (0.8784568) and standard deviation (1.120296) so that I may better understand its shape and spread. Then, I computed the 99% confidence interval for the mean WAR to understand the range within which the true mean likely falls. I found this range to be [0.6517355, 1.105178] for all KBO pitchers. I also drew one thousand samples of size one hundred from the WAR data as a means of simulating the sampling distribution of the sample mean. I found the mean and standard deviation of sample means of WAR to be 0.8787952 and 0.1115167 respectively, so we know that the sample mean is a good estimator of the population mean.

3.2 Exploratory Data Analysis

I created boxplots to analyze the relationship between my factors and WAR. Although the plots did not reveal any significant differences in WAR between teams or home park factors, they did show that starting pitchers and left-handed pitchers did tend to have a higher average WAR. When I conducted an ANOVA test to examine if WAR significantly varies across different levels of each categorical factor, I found that only Position was statistically significant, a piece of information I would keep in mind when constructing my model.

Next, I plotted the scatter plots of WAR against each continuous variable as well as the residuals from linear regression models to assess the appropriateness of the linear fit for each continuous variable. When I calculated correlation coefficients between WAR and each continuous variable, I found that Strikeouts, Innings Pitched, Plate Appearances, Wins, and Games Started had the highest positive correlation while ERA and WHIP had the highest negative correlation.

I then checked for multicollinearity and found that hits were highly correlated with many other offensive statistics. I also found that innings pitched and plate appearances were almost perfectly positively correlated as the more innings a pitcher is on the mound strongly predicts the number of batters he faces. Fly-ball rate and ground-ball rate show a very high negative correlation, suggesting that as one decreases the other increases, which is to be expected. I also noticed that several pitching statistics are highly correlated with their respective rates. For example, walks are closely correlated with walk rate, and similarly walk rate is closely correlated with walks per nine innings, suggesting that these metrics are capturing the same aspect of a pitcher's performance. Understandably, ERA was also highly correlated with Earned Runs. Since many of the statistics are closely correlated, I know I can use many of them in my assessment, however I do not want to remove any variables just yet over a fear of "redundant information" as I want to capture the most variation in my models which I can only do with more variables.

3.3 Regression Analysis

Based on the residual plots, I transformed all my data using Yeo-Johnson transformations in order to make the data follow a more normal distribution. Yeo-Johnson transformations are an extension of the Box-Cox transformation, enabling power transformations on variables with zero and negative values, in addition to positive values. I used Yeo-Johnson instead of Box-Cox as many of my variables have data points of "0" or negative values that are meaningful and thus should not be removed.

After the data was transformed using Yeo-Johnson, it was time to fit some models and try to find the best one. First, I fit a preliminary regression model using all of the variables of interest. The residuals of my model were relatively small, which indicates that my model provided a reasonably good fit. The most significant

predictors (at the 0.05 level) included Innings Pitched, Hits, Strikeouts, ERA, and WHIP. Many variables, such as Age, Years of Service, and Holds had non-significant p-values (greater than 0.05), indicating that they had very little significance to the model. Overall, the model had an Adjusted R-squared of 0.8635. This indicates that approximately 86.35% of the variance in WAR is explained by the model. The F-statistic had a p-value less than 2.2e-16, indicating that the model is statistically significant overall.

Then, I fit the same model with the addition of the Position factor that we identified as significant when we ran the ANOVA test. The model with the Position factor came out almost identical to the model without it, with the only difference being that the model with the position factor no longer utilized ERA. This model suggested that while the Position factor does not drastically affect the overall model, it does indicate some influence, especially for starting pitchers.

3.4 Model Selection

I then checked two more models, one using forward stepwise selection and one using backward stepwise selection. In a forward stepwise selection model, we start with no variables in the model, and at each step, the variable that is added is the one that improves the model the most. We repeat this process until the results are optimal. In a backwards stepwise selection, we simply start with a full model and remove variables one by one until the model is optimal.

After all four models had been constructed, I compared them using their Adjusted R-squared values and found that the model that explained the most variation (had the highest Adjusted R-squared) was the backwards stepwise selection model. It provided me with a formula of:

$$\begin{aligned}\widehat{\text{WAR}} = 0.878 &- 0.314 \cdot \text{GamesRelieved} + 0.229 \cdot \text{Hits} - 0.052 \cdot \text{HomeRuns} \\ &+ 0.054 \cdot \text{Walks} + 0.309 \cdot \text{Strikeouts} - 0.358 \cdot \text{ERA} - 0.264 \cdot \text{WHIP} \\ &+ 0.133 \cdot \text{HR/9} + 0.144 \cdot \text{OBPAgainst} - 0.204 \cdot \text{OPSAgainst}\end{aligned}$$

This model had an Adjusted R-squared of 0.879 and an F-statistic with a p-value less than 2.2e-16.

Earlier, we had checked for multicollinearity and found that many of the variables in our backwards stepwise selection model were highly correlated. When I calculated the Variance Inflation Factors (VIFs), which measure the correlation between independent variables, I found that Innings Pitched and Plate Appearances had VIFs over 550, Earned Runs and ERA had VIFs over 50, and Hits and WHIP had VIFs over 27. When variables have high VIF (VIF over ten suggests very high correlation) normally you would remove them from the model, but I found that removing these variables actually lowered the Adjusted R-squared, meaning that the model is predicting less of the variation in WAR. This suggests that while being highly correlated, these variables are positively contributing to the model. I later learned that in straightforward predictive applications such as this, very high VIFs can be unproblematic. I then checked the model assumptions of the backwards stepwise selection model using a Residuals vs Fitted Plot, a Normal Q-Q Plot, a Scale-Location Plot, and a Residuals vs Leverage Plot, which all suggested that the model was an overall good fit.

---

4 Findings

4.1 WAR Distribution and Summary

The distribution of WAR across 162 qualified KBO pitchers is right-skewed, with a mean of 0.878 and a median of 0.515, reflecting the predominance of relievers and the relatively small number of elite starters who accumulate high WAR totals. The standard deviation is 1.120, and values range from −0.600 to 4.840. The 99% confidence interval for the mean WAR is [0.652, 1.105]. Bootstrap sampling (1,000 samples of n = 100) produced a mean of sample means of 0.879 and a standard deviation of 0.112, confirming that the sample mean is a reliable estimator of the population mean.

See Appendix A for all WAR distribution figures and tables.

4.2 Factor Analysis

Of the four categorical factors examined—Position, Team, Handedness, and Home Park Factor—only Position was statistically significant in predicting WAR (ANOVA, $p < 0.0001$). Starting pitchers accumulated substantially more WAR on average (mean = 1.708) than relievers (mean = 0.279) or closers (mean = 1.042), consistent with the greater innings workload of starters. Left-handed pitchers showed a higher mean WAR (1.124) than right-handed pitchers (0.787), though this difference fell short of statistical significance ($p = 0.089$). Neither team nor home park factor produced significant variation in WAR across the league.

See Appendix B for all factor analysis figures and tables.

4.3 Correlation and Multicollinearity

Strikeouts ($r = 0.884$), Innings Pitched ($r = 0.861$), Plate Appearances ($r = 0.824$), Wins ($r = 0.801$), and Games Started ($r = 0.731$) had the highest positive correlations with WAR. ERA ($r = -0.604$) and WHIP ($r = -0.669$) had the highest negative correlations. Extensive multicollinearity was identified: Plate Appearances and Innings Pitched were nearly perfectly correlated ($r = 0.996$), and 29 variable pairs exceeded $|r| > 0.8$. Despite this, high-VIF variables were retained in the stepwise models as their removal reduced predictive performance.

See Appendix C for all correlation figures and tables.

4.4 Regression Results

The final reduced model, selected via backward stepwise AIC and iterative VIF reduction (threshold = 10), achieved an Adjusted R-squared of 0.837 and a residual standard error of 0.452. All VIF values in the final model were below 10. The strongest significant predictors included ERA ($= -0.358$, $p < 0.001$), GamesRelieved ($= -0.314$, $p < 0.001$), Strikeouts ($= 0.309$, $p < 0.001$), WHIP ($= -0.264$, $p < 0.001$), and Hits ($= 0.229$, $p < 0.01$). Model diagnostics—Residuals vs Fitted, Normal Q-Q, Scale-Location, and Residuals vs Leverage plots—all indicated an overall good fit.

See Appendix D for all regression modeling figures and tables.

---

5 Discussion and Conclusion

Based on this model and the formula it provided, I printed the twenty best pitchers from my data set. These pitchers were Kyle Hart, James Naile, Ariel Jurado, Enmanuel De Jesus, Gwak Been, William Cuevas, Aaron Wilkerson, Dietrich Enns, Connor Seabold, Charlie Barnes, Daniel Castano, Wes Benjamin, Ryu Hyun-jin, Yang Hyeon-jong, Won Tae-in, Denyi Reyes, Um Sang-back, Park Se-woong, Brandon Waddell, and Casey Kelly. Based on these pitchers, I will narrow down the candidates by breaking down their stats as well as utilizing some very important metrics that I left out of my model for simplicity of computing. These metrics are the run value of individual pitches, pitch velocity, and age. While age was used as a variable of interest in our model, it wasn't used for narrowing down MLB candidates. Realistically, it makes more sense for older MLB pitchers, such as Johnny Cueto or Carlos Carrasco, to transition to KBO than the reverse, as no MLB team is going to want to take a flier on an aging foreign prospect. Keep in mind that while I am critical of many of these pitchers, they are still among the best in Korea, a nod to the overall lack of pitching depth on the peninsula.

Ryu Hyun-jin, 37, is the oldest pitcher that was produced by our model. Ryu Hyun-jin has enjoyed a long career, including ten years already in the MLB with the Los Angeles Dodgers and Toronto Blue Jays where he accumulated over 930 strikeouts, 20 WAR, an ERA title, an All-Star appearance, as well as two top-three finishes in the Cy Young Award voting, an award given annually to the best pitchers in Major League Baseball. He is also considered one of the best pitchers in KBO history. A return to MLB after already spending ten seasons in America at such an old age would make little sense for the decorated Ryu Hyun-jin.

Yang Hyeon-jong, another KBO veteran, has pitched 17 seasons thus far, all for the Kia Tigers, and will no doubt finish his baseball career in Korea, especially considering his four seam fastball tops out around 87 miles an hour with a pitch value below five.

Similarly, at 35 years old Aaron Wilkerson is just a little too old to be considered a realistic MLB target, despite his ability to limit walks. Casey Kelly and William Cuevas are two international pitchers who have pitched in the KBO since 2019. Both are over thirty three years old, well above the typical age a free agent pitcher from overseas would get signed by an MLB team. Kelly has the lowest strikeout rate and WAR as well as the highest FIP of any of our top pitchers and recently parted ways with the LG Twins after six seasons. While Cuevas has a fantastic slider that batters are hitting an embarrassing 0.102 against, his age and his high walk rate coupled with his fastball sitting at around ninety one miles an hour limit his MLB potential. I do not see Cuevas as a candidate for MLB, however seeing as Kelly is a current free agent, I could understand why an MLB team might take a flier on him for depth after a strong, successful stint in KBO.

Dietrich Enns, who also spent time in Japan with the Seibu Lions, limits walks and has a good cutter that can miss bats, however his age and less-than-ideal ERA make him a risky candidate for MLB especially considering his 2023 season in Japan was shaky to say the least, posting a 5.17 ERA in just fifty-four innings against worse competition than in MLB.

Both Park Se-woong and Um Sang-back have exceeded the period in KBO when a star pitcher would typically be asked to be posted (seven seasons) having both played at least nine seasons. Park Se-woong currently has the highest ERA and WHIP among our top twenty pitchers and is having his worst season since 2018. Um Sang-back has given up the second most home runs, has the second highest fly-ball rate, and the second highest ERA in our top twenty list, three stats that don't bode well for future success against better hitters, despite his high strikeout numbers. I do not see either as future MLB arms, especially considering their fastballs barely touch 90 miles per hour.

Daniel Castano has a high ERA, low strikeout totals, and the lowest accumulated WAR of any of our top pitchers thus far. Similar to Cuevas, his strong slider makes up for his sub-90 fastball. Denyi Reyes has put up good numbers for the Samsung Lions this year, but his general lack of velocity on his pitches limits his strikeout numbers and contributes to his high OPSAgainst. Still young, I would enjoy seeing Reyes add a few more miles per hour to his pitches to take away his reliance on his fastball and to give more effectiveness to his secondary arsenal.

Another Samsung Lion, Connor Seabold has shown his ability to miss bats with both four-seam fastball and

slider having a run value above ten and good strikeout numbers. When batters make contact, however, the ball tends to leave the park, seeing as Seabold has been hurt by the long ball in his lone KBO season.

Wes Benjamin is a KBO veteran of three years whose numbers have been steadily on the decline since 2022, his best year on paper. While his ERA and FIP are higher than most of our top twenty pitchers and it is clear that Benjamin has been hurt by the long ball, he has a solid strikeouts per nine and opponents can't seem to hit his slider nor his curve. Unlike many of our other pitchers, Benjamin has been known to get knocked out of games early and give up runs. He gave up eleven runs in just three innings against Hanwha on March 31st, probably the worst outing of any top pitcher on our list.

Brandon Waddell is a three year pitcher for the Doosan Bears who has missed some time over the last two seasons due to injury. He has pitched the least amount of innings of any one in our top twenty. While his outings are limited and he does give up a lot of hits, he has showcased his ability to limit walks having not walked more than two batters in a game this season thus far. He does lack velocity and none of his pitches are noticeably valuable, with his slider being his most effective pitch. Batters are hitting over 0.290 off his change-up and over 0.300 off his four seam fastball and his curveball.

Charlie Barnes, who came up as an influential point in our residual plots earlier, has a very low ERA, FIP, WHIP, and BAA combined with great strikeout numbers across the board. While his slider is extremely effective (and is thus his most used), the rest of his arsenal is shaky at best and often gets hit very hard due to his overall velocity deficiency.

Enmanuel De Jesus is an interesting arm that I was able to scout in person on multiple occasions. One of the harder throwers on our list, De Jesus has good numbers in his first season in KBO and has shown his ability to go deep into games while tallying up lots of strikeouts. While he has been hurt by walks, he also has had multiple games of nine plus strikeouts while only having three games where he did not last five innings, ranking in the top five in KBO in quality starts.

Another Kiwoom pitcher, Ariel Jurado, leads the KBO in quality starts and has almost identical numbers to De Jesus, having worked into the seventh inning in almost every start this year while also racking up strikeouts. While his fastball gets hit due to its lack of velocity, his arsenal of plus offspeed pitches helps him work out of trouble and gives him three pitches he can fall back on when his fastball isn't doing him any favors.

James Naile, a pitcher for the Kia Tigers, has the most valuable pitch among any of our top twenty pitchers, throwing a slider with a run value of thirty one which sits at around 83 miles per hour. Naile ranks higher than both Jurado and De Jesus in home runs, walks, strikeouts, ERA, and FIP while also lasting at least five innings in all twenty one starts thus far this season. While his slider is dominant, it is his only valuable pitch in his arsenal and he lacks velocity, with his fastball sitting at around 91.

Won Tae-in, who started this year's All-Star game, has been one of the best Korean-born starting pitchers in the league the past few years. While he has a good pitch mix, commanding multiple pitches for strikes with good movement, he lacks the velocity necessary to overpower hitters. Currently, I see him as a more likely candidate for NPB than MLB unless he can up his velocity. I wouldn't count him out just yet though, as he still has another year before he can officially ask to be posted.

Gwak Been, my favorite pitcher to watch while in Korea, has a similar scouting report to Won Tae-in with the added benefit of a few extra ticks on his fastball and more strikeouts. Having only played in KBO for five seasons, Gwak Been has two more years to work on his command as he has been hurt by walks and possibly add a few more miles per hour to his fastball. He has an excellent curveball that hitters are batting 0.154 against. A top KBO prospect and fan favorite, he was called the Shohei Otani of high school baseball, and I am interested to see how the next two years of his KBO career pan out.

Kyle Hart is probably the best pitcher in KBO this season. He currently leads the league in ERA and strikeouts. While he lacks the velocity of a major leaguer, his pitches' velocity and movement has proved enough to give him a trio of effective pitches in his four seam fastball, slider, and change-up which KBO hitters can't seem to touch. If he can increase his velocity, I could very well see him going down the path of Erick Fedde, who after a dominant tenure in KBO also with the Dinos, signed a contract with the Chicago

White Sox in the MLB.

By far the most compelling KBO pitching prospect is one that was not even included in our original dataset. An Woo-jin is a twenty-four year old right-handed flamethrower for the Kiwoom Heroes who has missed the entire 2024 season rehabbing from his September Tommy John surgery. Since becoming a full time starter in 2021, An Woo-jin has been without a doubt the best pitcher in South Korea. In 2022, he led KBO with a 2.11 ERA (winning the ERA title), 224 strikeouts (the most ever by a Korean-born pitcher in a season), WHIP, and WAR and was leading the league in strikeouts again in 2023 before requiring surgery. He has a filthy pitch mix, utilizing a slider which touches 93 miles per hour along with his dominant four seam fastball, which sits at 99 miles per hour and has a whopping pitch run value of 32.6 that completely overpowers hitters. He also ranks in the 100th percentile for WAR, ERA, FIP, WHIP, and strikeouts. Having already played six years in KBO, An Woo-jin will have the chance to be posted after one more season and I would be shocked if there weren't many big market teams in MLB checking in with the Heroes on a posting fee to negotiate with the twenty-four-year-old superstar. It'll be very interesting seeing how An Woo-jin bounces back from TJS, however I expect him to debut in MLB in 2025 following a return to form regardless.

Based on all my scouting and models, I love An Woo-jin most immediately as an MLB prospect and will be keeping an eye out for Ariel Jurado, Enmanuel De Jesus, James Naile, and Kyle Hart to see if they can up their velocity and finish this current KBO season on a high note. I will also be watching Gwak Been and Won Tae-in as they continue to develop as top KBO arms.

At the end of the day, there is a limited pool of pitching talent coming out of South Korea, particularly due to the lack of depth in relievers and the generally lower velocity across most pitchers. This presents a challenge when transitioning from the KBO to the MLB, where the standards are significantly higher, especially in terms of pitching. With the KBO's smaller player pool, it is easier for players to get signed professionally, and many opt to stay in South Korea, where they enjoy celebrity status, comfortable living, and a loyal fanbase. The language barrier and the difficult road through the minor leagues make it less enticing for players with a marginal shot at cracking the MLB roster.

However, as baseball evolves beyond the "Moneyball" era and into one that focuses on advanced player development through technology and physics, the transition may be less daunting than it seems. With the right tools, a pitcher—even a homer-prone finesse type in their thirties—can still make waves in the MLB. The lack of "set" rates in today's game and the unpredictable nature of injuries make room for players to reinvent themselves and carve out a role, especially if they possess the right pitching mechanics and arm talent.

Having spent time in South Korea, I've been able to sharpen my understanding of player evaluation and scouting. Despite limited access to advanced metrics and technology, I've developed a keen eye for assessing pitchers by focusing on the fundamentals and watching games closely. My proficiency in building predictive models will certainly prove valuable going forward, as I plan to create a machine learning model using Random Forests to predict the potential value of Japanese pitchers transitioning to MLB.

---

6 References

Albert, J., & Marchi, M. (2014). *Analyzing baseball data with R.* CRC Press.

Baseball-Reference.com. (n.d.). *Baseball statistics and history.* Retrieved May 31, 2026, from https://www.baseball-reference.com

Friendly, M., & Meyer, D. (2016). *Discrete data analysis with R: Visualization and modeling techniques for categorical and count data.* CRC Press.

FanGraphs. (n.d.). *FanGraphs baseball statistics and analysis.* Retrieved May 31, 2026, from https://www.fangraphs.com

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R.* Springer.

KBO League. (n.d.). *KBO statistics portal.* Retrieved May 31, 2026, from https://www.koreabaseball.com

MLB.com. (n.d.). *Official MLB statistics definitions.* Retrieved May 31, 2026, from https://www.mlb.com/glossary

Tango, T., Lichtman, M., & Dolphin, A. (2007). *The book: Playing the percentages in baseball.* Potomac Books.

Yeo, I.-K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika, 87*(4), 954–959. https://doi.org/10.1093/biomet/87.4.954

Appendix A: WAR Distribution Analysis

Sections: Summary Statistics | WAR Histogram | Sampling Distribution | WAR by Role | WAR by Team

Summary Statistics

Table 1: Summary statistics for KBO pitcher WAR.

Statistic	Value
N	162.000
Mean WAR	0.878
Median WAR	0.515
SD	1.120
Min	-0.600
Max	4.840
99% CI Lower	0.652
99% CI Upper	1.105

WAR Histogram

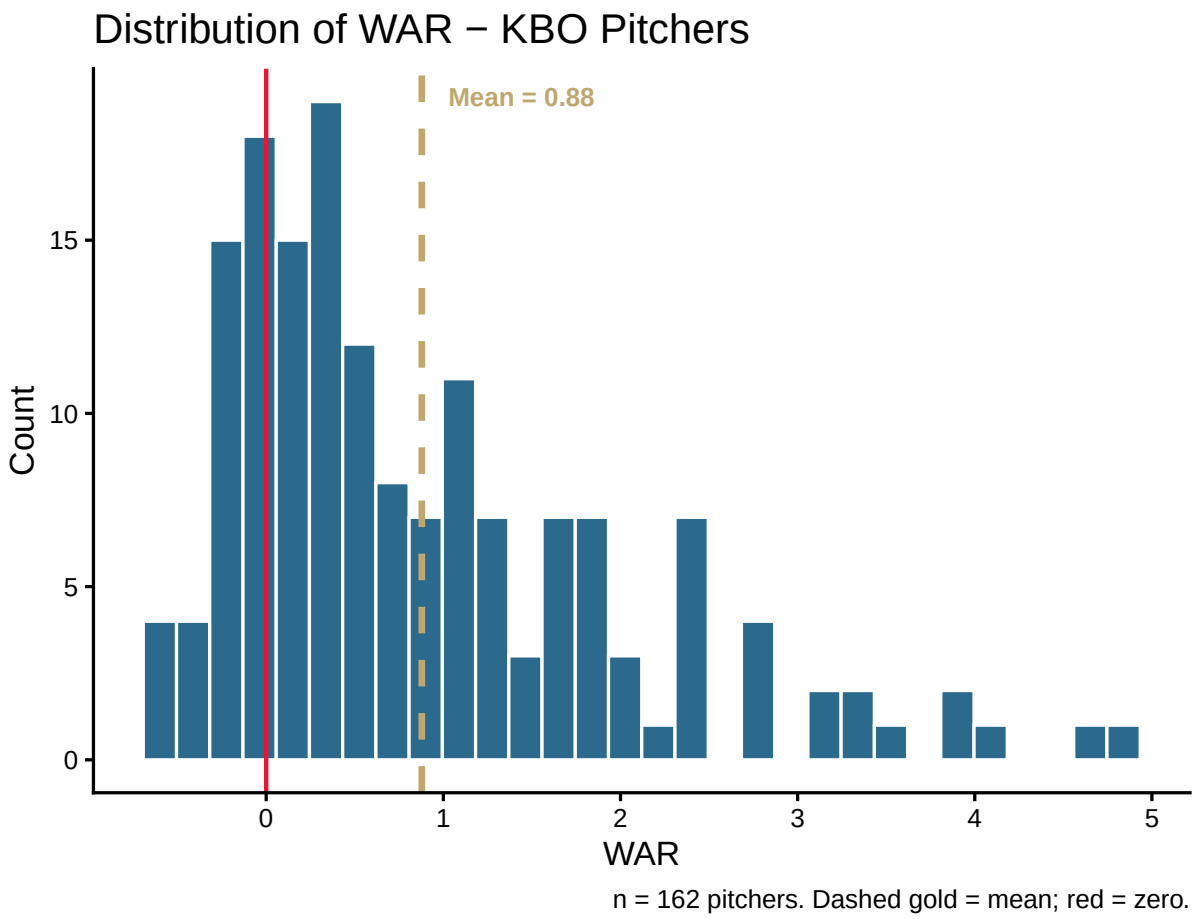


Figure 1: Distribution of pitcher WAR across all 162 KBO pitchers. Vertical dashed line = mean; solid line = 0.

Sampling Distribution

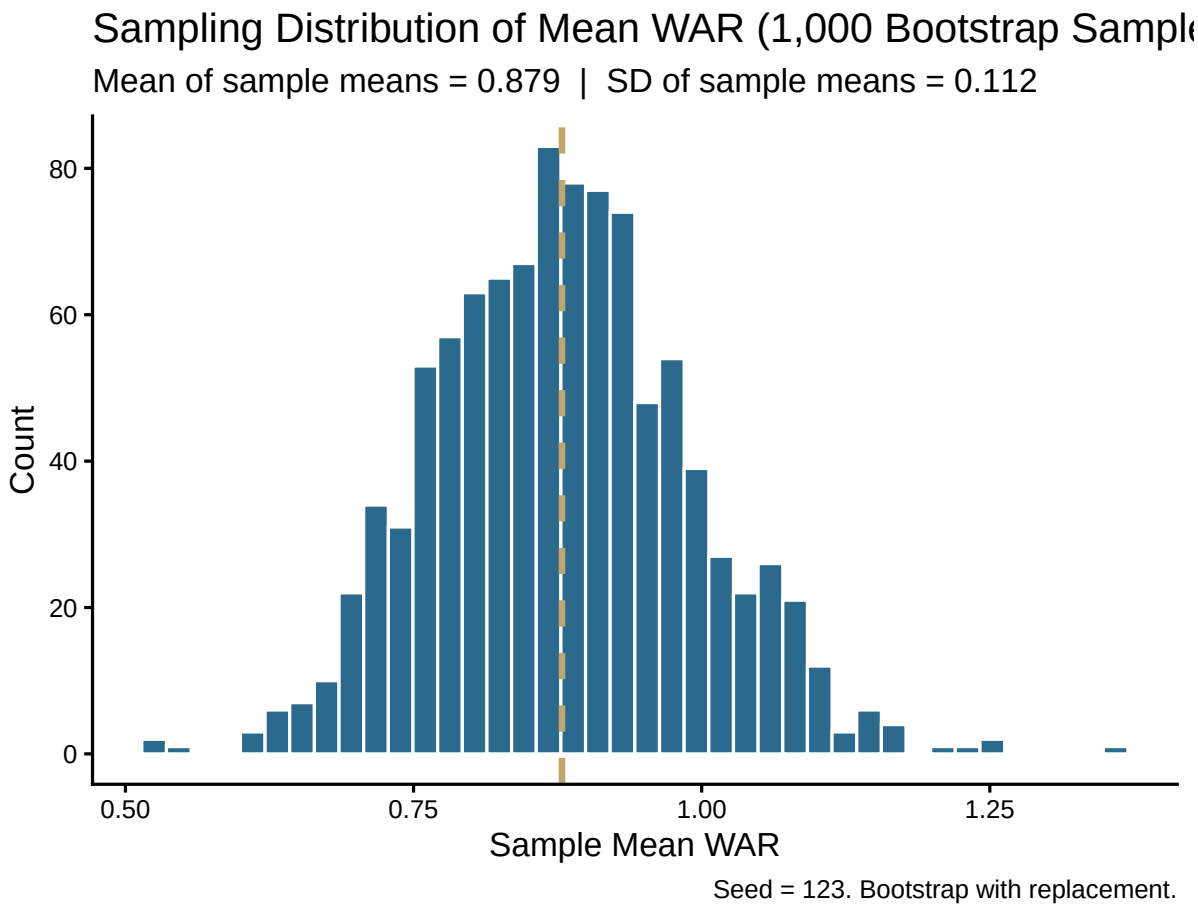


Figure 2: Histogram of 1,000 sample means drawn from bootstrap samples of size 100. Demonstrates the Central Limit Theorem applied to KBO WAR.

WAR by Role

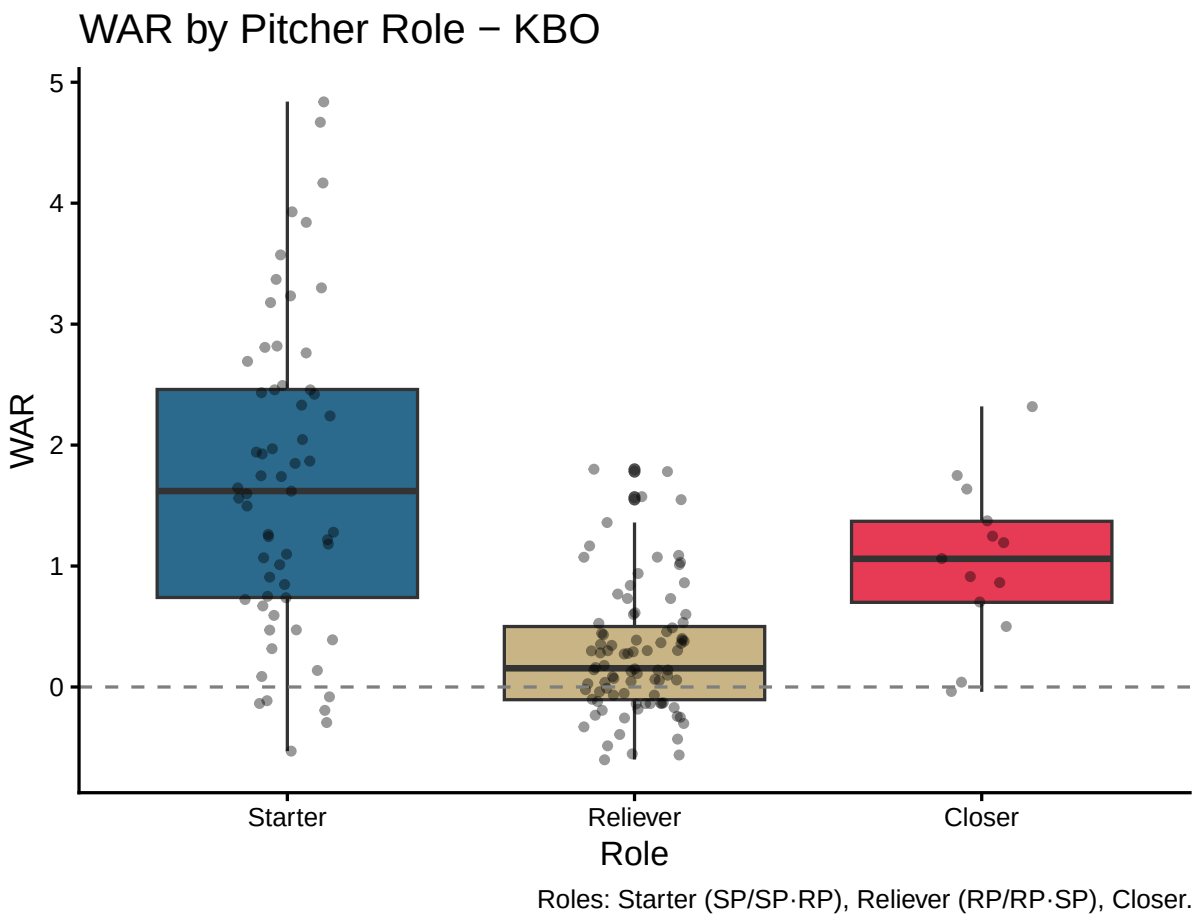


Figure 3: WAR distribution by pitcher role. Closers and starters tend to accumulate more WAR than relievers.

Table 2: WAR summary statistics by pitcher role.

role	N	Mean	Median	SD	Min	Max
Starter	61	1.708	1.620	1.285	-0.53	4.84
Reliever	88	0.279	0.155	0.524	-0.60	1.80
Closer	13	1.042	1.060	0.665	-0.04	2.32

WAR by Team

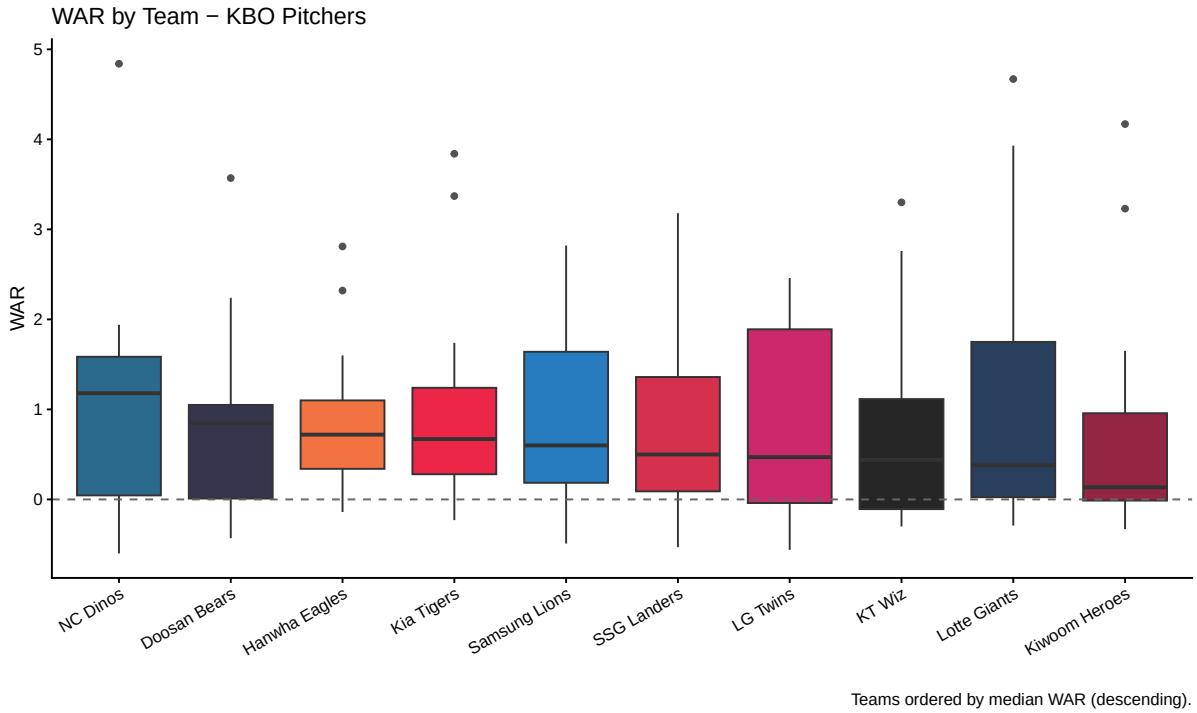


Figure 4: WAR distribution by KBO team. Team colors reflect official KBO branding.

Table 3: WAR summary statistics by KBO team, sorted by total WAR.

Team	N	Mean	Median	SD	Total WAR
Lotte Giants	15	1.112	0.380	1.542	16.68
Kia Tigers	17	0.965	0.670	1.147	16.41
NC Dinos	15	1.031	1.180	1.364	15.46
Hanwha Eagles	17	0.902	0.720	0.794	15.33
Doosan Bears	19	0.752	0.850	0.978	14.28
Samsung Lions	15	0.941	0.600	1.096	14.11
SSG Landers	17	0.778	0.500	0.995	13.23
LG Twins	15	0.837	0.470	1.038	12.56
KT Wiz	16	0.767	0.440	1.141	12.28
Kiwoom Heroes	16	0.748	0.135	1.284	11.97

Appendix B: Factor Analysis

Sections: ANOVA Results | WAR by Position | WAR by Handedness | WAR by Park Factor | Top Pitchers by WAR

ANOVA Results

Table 4: One-way ANOVA results for each categorical factor predicting WAR. * = p < 0.05.

Factor	Df	Sum Sq	Mean Sq	F	p-value	Sig
Position	4	92.235	23.059	32.962	0.0000	*
Team	9	2.332	0.259	0.197	0.9942	
Handedness	1	3.631	3.631	2.928	0.0890	
HomeParkFactor	8	2.270	0.284	0.217	0.9874	

WAR by Position

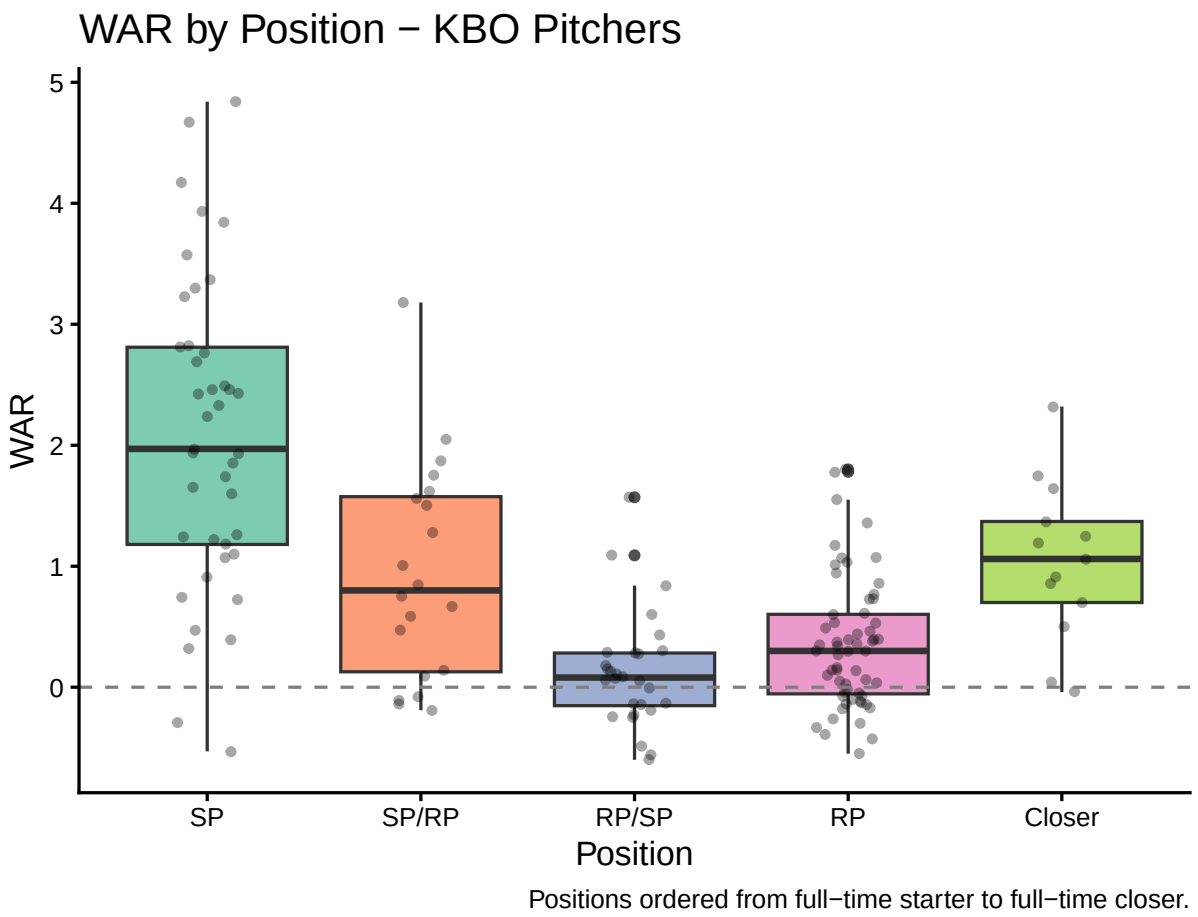


Figure 5: WAR by fielding position type. SP = Starting Pitcher; RP = Relief Pitcher; Closer = dedicated closer role.

WAR by Handedness

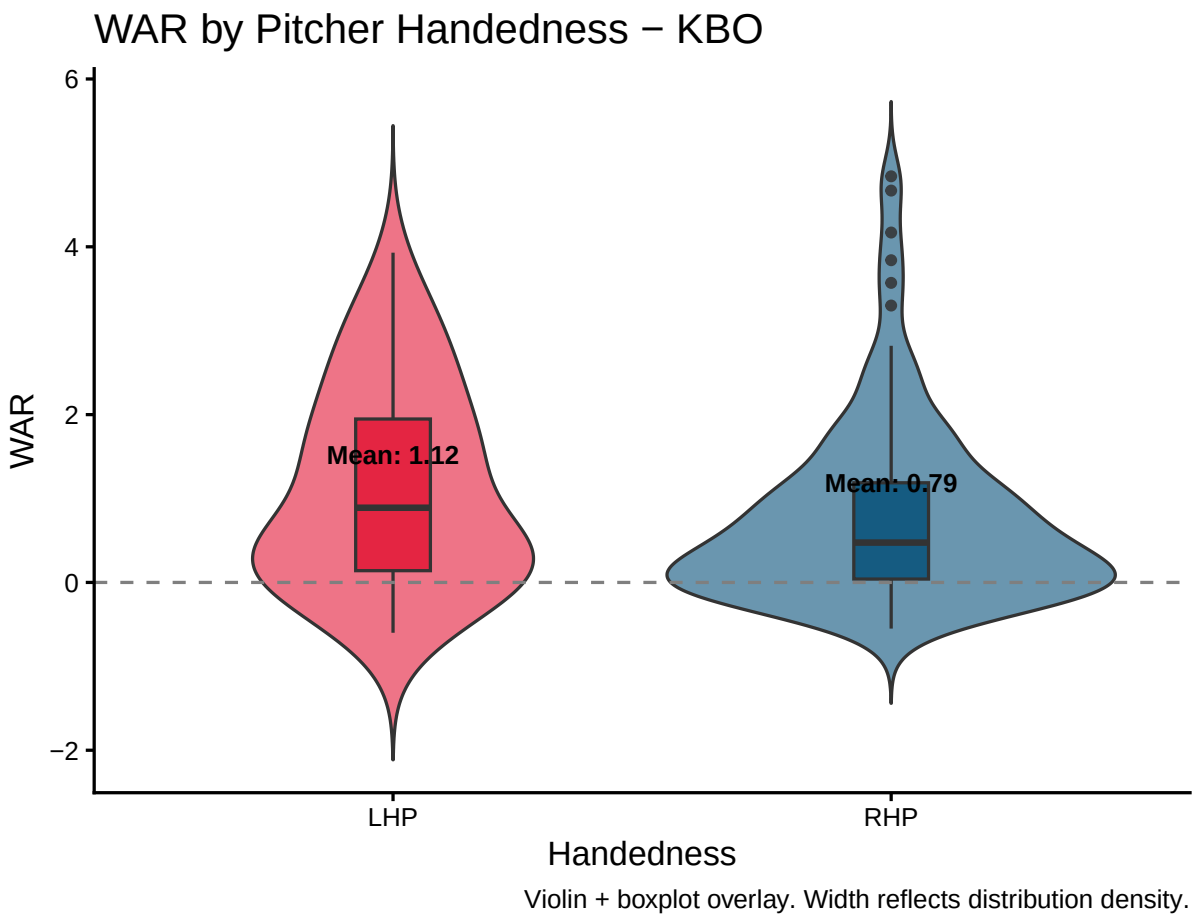


Figure 6: WAR distribution for right-handed pitchers (RHP) vs. left-handed pitchers (LHP) in the KBO.

Table 5: WAR summary statistics by pitcher handedness.

Handedness	N	Mean	Median	SD
LHP	44	1.124	0.890	1.195
RHP	118	0.787	0.475	1.082

WAR by Park Factor

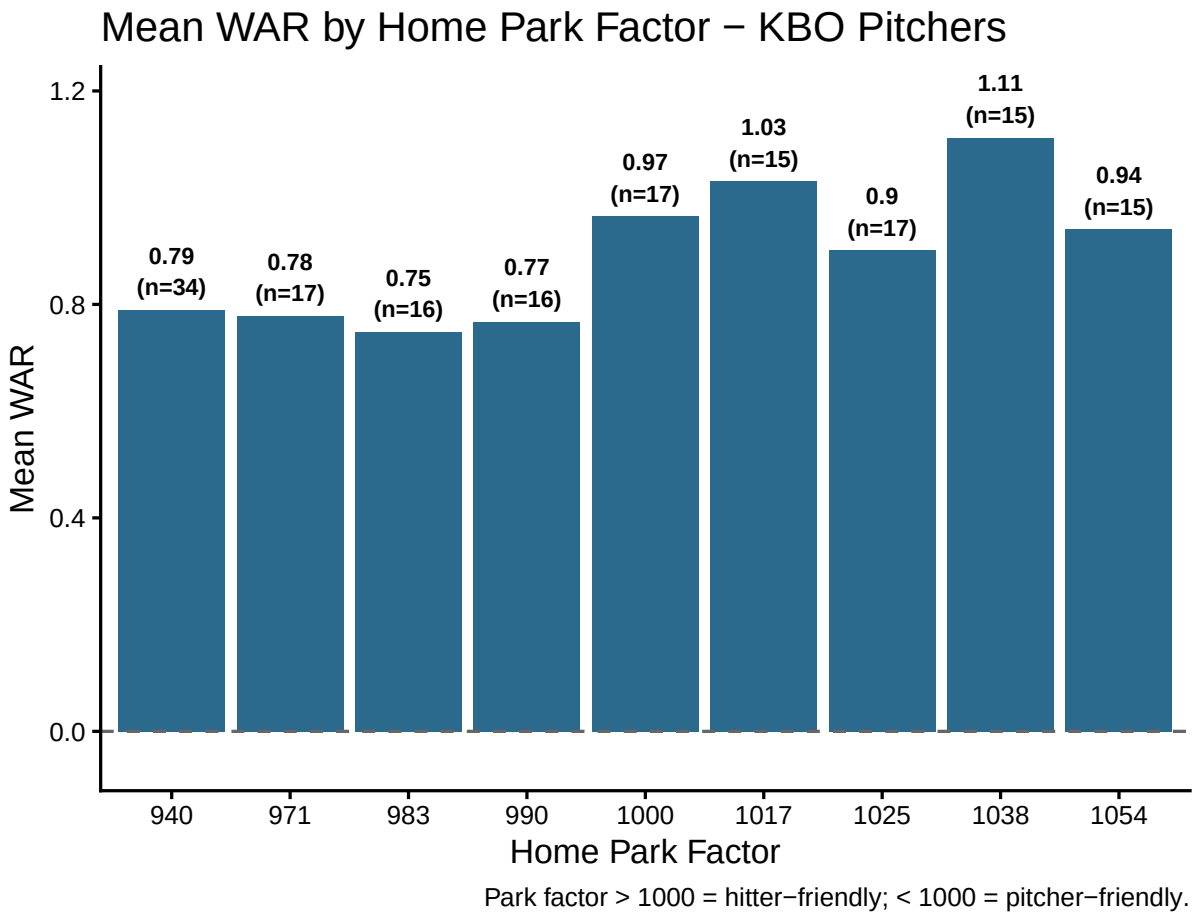


Figure 7: Mean WAR by home park factor. A park factor above 1000 favors hitters; below 1000 favors pitchers.

Top Pitchers by WAR

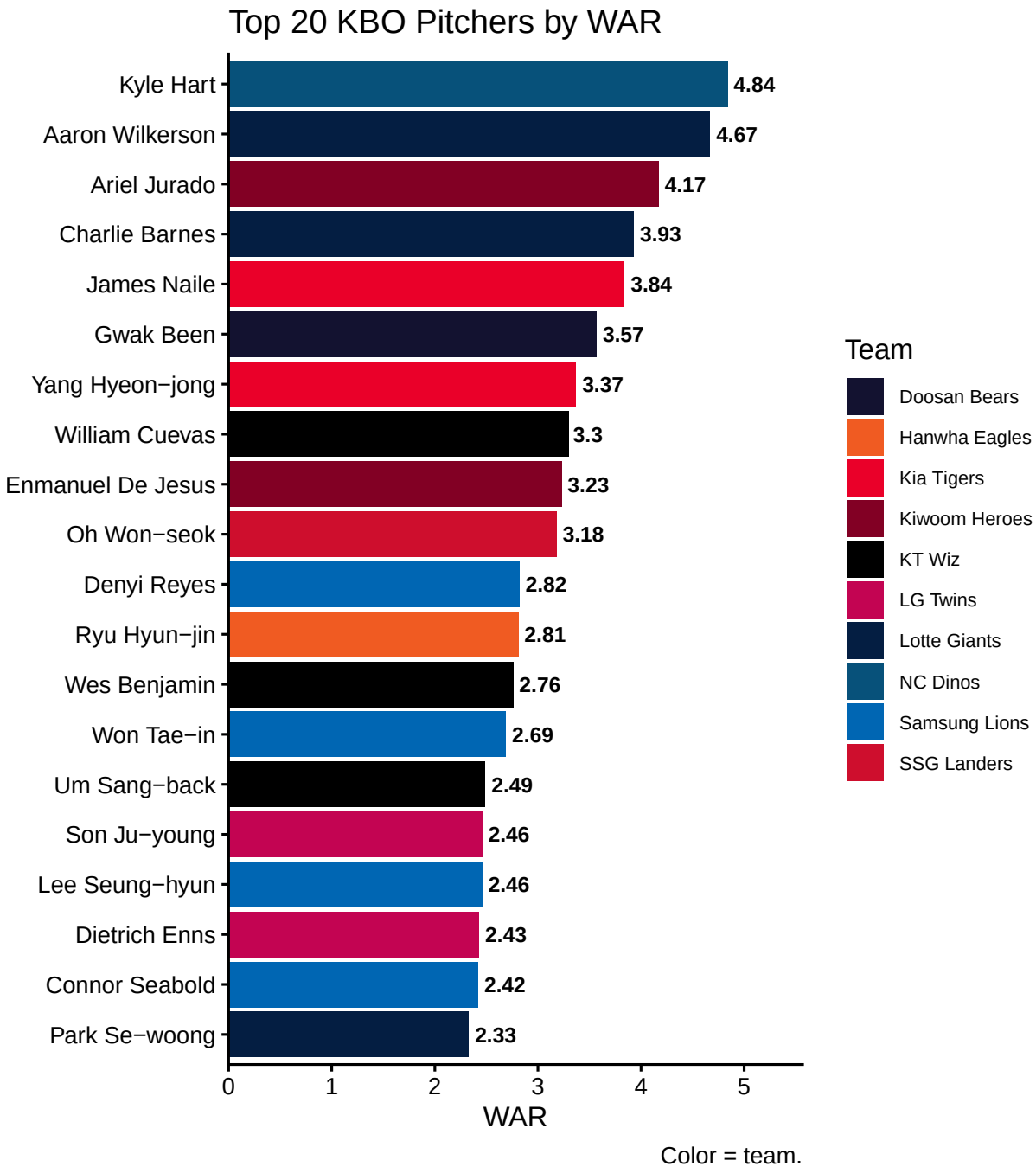


Figure 8: Top 20 KBO pitchers by WAR. Color = team.

Table 6: Top 20 KBO pitchers by WAR with key pitching metrics.

Name	Korean Name	Team	Pos	Hand	Age	WAR	ERA	FIP	WHIP	K%	BB%	SIERA
Kyle Hart	하트	NC Dinos	SP	RHP	31	4.84	2.44	2.96	1.07	16.2	12.1	2.78
Aaron Wilkerson	윌커슨	Lotte Giants	SP	RHP	35	4.67	3.61	3.86	1.18	22.8	5.1	3.21
Ariel Jurado	후라도	Kiwoom Heroes	SP	RHP	28	4.17	3.40	3.78	1.21	15.4	10.6	3.44
Charlie Barnes	반즈	Lotte Giants	SP	LHP	28	3.93	3.04	3.15	1.16	17.5	12.8	2.31
James Naile	네일	Kia Tigers	SP	RHP	31	3.84	2.88	3.72	1.21	12.1	8.5	3.26
Gwak Been	곽빈	Doosan Bears	SP	RHP	25	3.57	3.83	3.85	1.24	22.4	8.2	3.96
Yang Hyeon-jong	양현종	Kia Tigers	SP	LHP	36	3.37	4.06	4.62	1.26	17.1	8.1	4.11
William Cuevas	쿠에바스	KT Wiz	SP	RHP	33	3.30	4.09	4.21	1.22	25.1	6.4	3.68
Enmanuel De Jesus	헤이수스	Kiwoom Heroes	SP	LHP	27	3.23	3.34	3.88	1.18	22.4	8.9	3.10
Oh Won-seok	오원석	SSG Landers	SP/RP	LHP	23	3.18	4.00	4.71	1.30	13.6	5.5	4.05
Denyi Reyes	엄상백	Samsung Lions	SP	RHP	27	2.82	3.88	4.27	1.38	21.5	9.7	4.00
Ryu Hyun-jin	류현진	Hanwha Eagles	SP	LHP	37	2.81	3.76	3.36	1.30	16.1	14.0	3.61
Wes Benjamin	벤자민	KT Wiz	SP	LHP	30	2.76	4.18	4.32	1.12	17.1	8.9	3.11
Won Tae-in	원태인	Samsung Lions	SP	RHP	24	2.69	3.69	4.84	1.29	20.2	5.0	4.38
Um Sang-back	레예스	KT Wiz	SP	RHP	27	2.49	4.76	4.24	1.25	10.8	16.9	2.91
Son Ju-young	손주영	LG Twins	SP	LHP	25	2.46	3.48	4.78	1.46	19.0	4.9	4.42
Lee Seung-hyun	이승현	Samsung Lions	SP	LHP	22	2.46	3.58	4.85	1.40	16.4	12.3	4.51
Dietrich Enns	엔스	LG Twins	SP	LHP	33	2.43	4.15	3.66	1.31	23.0	7.2	3.46
Connor Seabold	코너	Samsung Lions	SP	RHP	28	2.42	4.14	4.88	1.21	13.3	13.3	3.21
Park Se-woong	박세웅	Lotte Giants	SP	RHP	28	2.33	5.15	4.34	1.44	26.6	6.5	4.39

Appendix C: Correlation and Multicollinearity

Sections: Correlations with WAR | Key Scatter Plots | Highly Correlated Pairs

Correlations with WAR

Table 7: Pearson correlations between continuous predictors and WAR, sorted by absolute value.

Variable	Correlation
Strikeouts	0.8839
InningsPitched	0.8612
PlateApperances	0.8238
Wins	0.8014
GamesStarted	0.7313
Hits	0.7145
WHIP	-0.6686
ERA	-0.6039
Doubles	0.5770
SIERA	-0.5748
FIP	-0.5512
Homeruns	0.4992
EarnedRuns	0.4797
Walks	0.4724
Loses	0.4608
GamesRelieved	-0.3606
Triples	0.2807
OBPAgainst	-0.1628
BABIP	-0.1428
Age	0.1295
GamesFinished	-0.1292
WalksPerNineInnings	-0.1277
WalkRate	-0.1191
StrikeoutRate	0.1139
Holds	-0.1111
YearsOfService	-0.1099
OPSAgainst	-0.1078
PopUpRate	0.0905
Games	-0.0847
StrikeoutsPerNineInnings	0.0772
StrikeoutToWalkRatio	0.0674
AVGAgainst	-0.0666
ThreeTrueOutcomesRate	-0.0618
SLGAgainst	-0.0595
HomerunsPerNineInnings	-0.0540
Saves	0.0287
FlyBallRate	0.0229
LineDriveRate	-0.0226
GroundBallRate	-0.0179

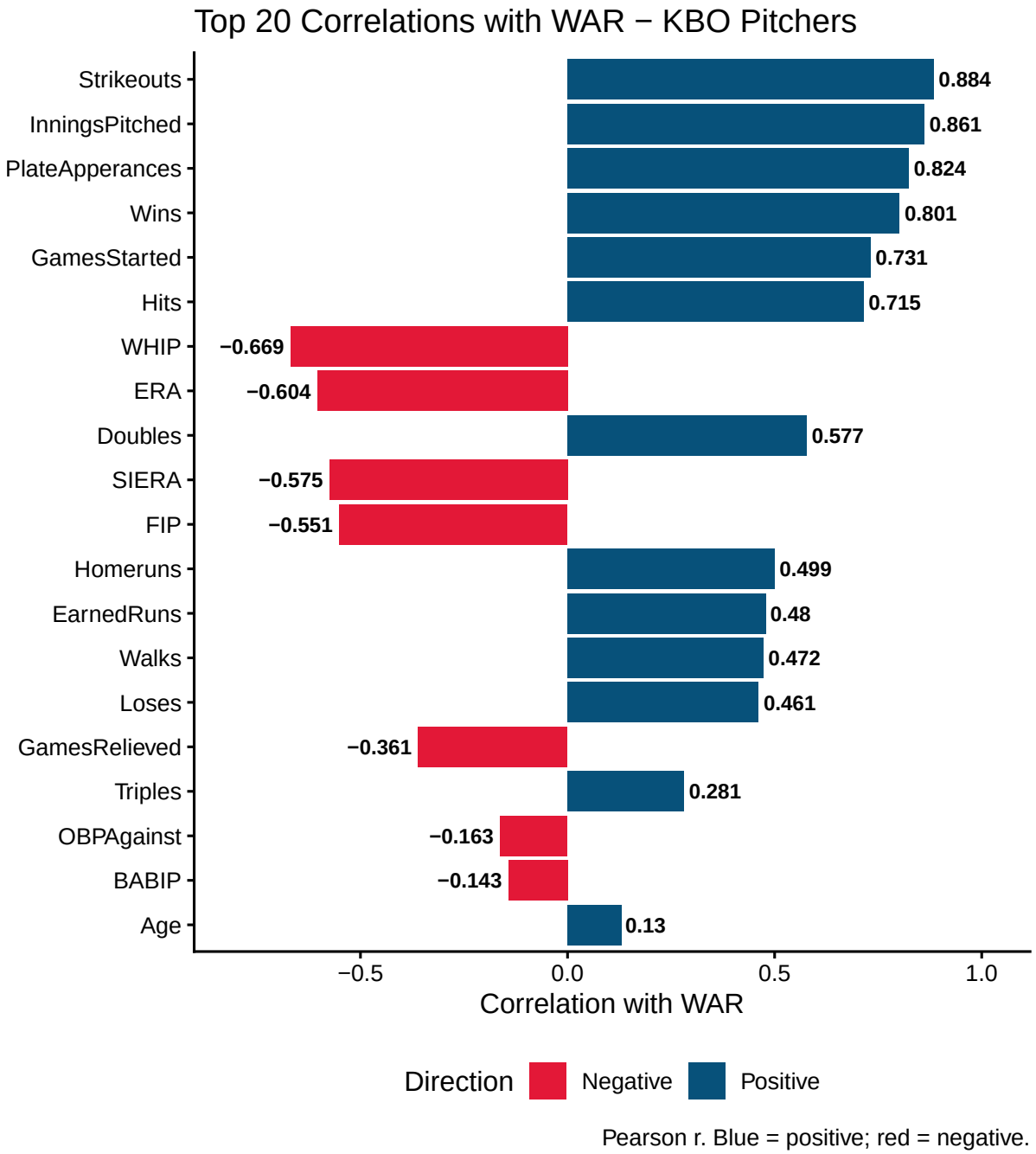


Figure 9: Top 20 variables by absolute correlation with WAR. Blue = positive; red = negative association.

Key Scatter Plots

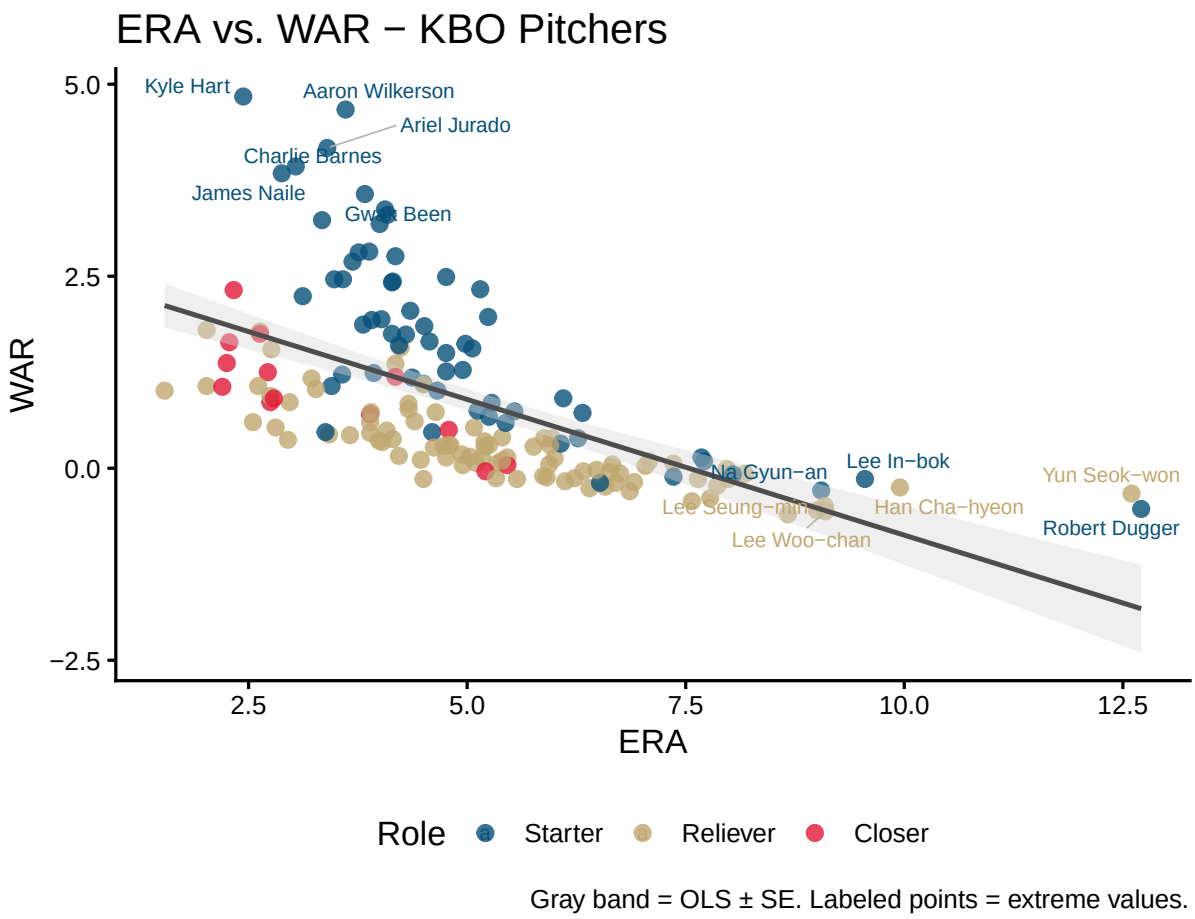


Figure 10: ERA vs. WAR for all KBO pitchers, colored by role. Lower ERA is expected to correlate with higher WAR.

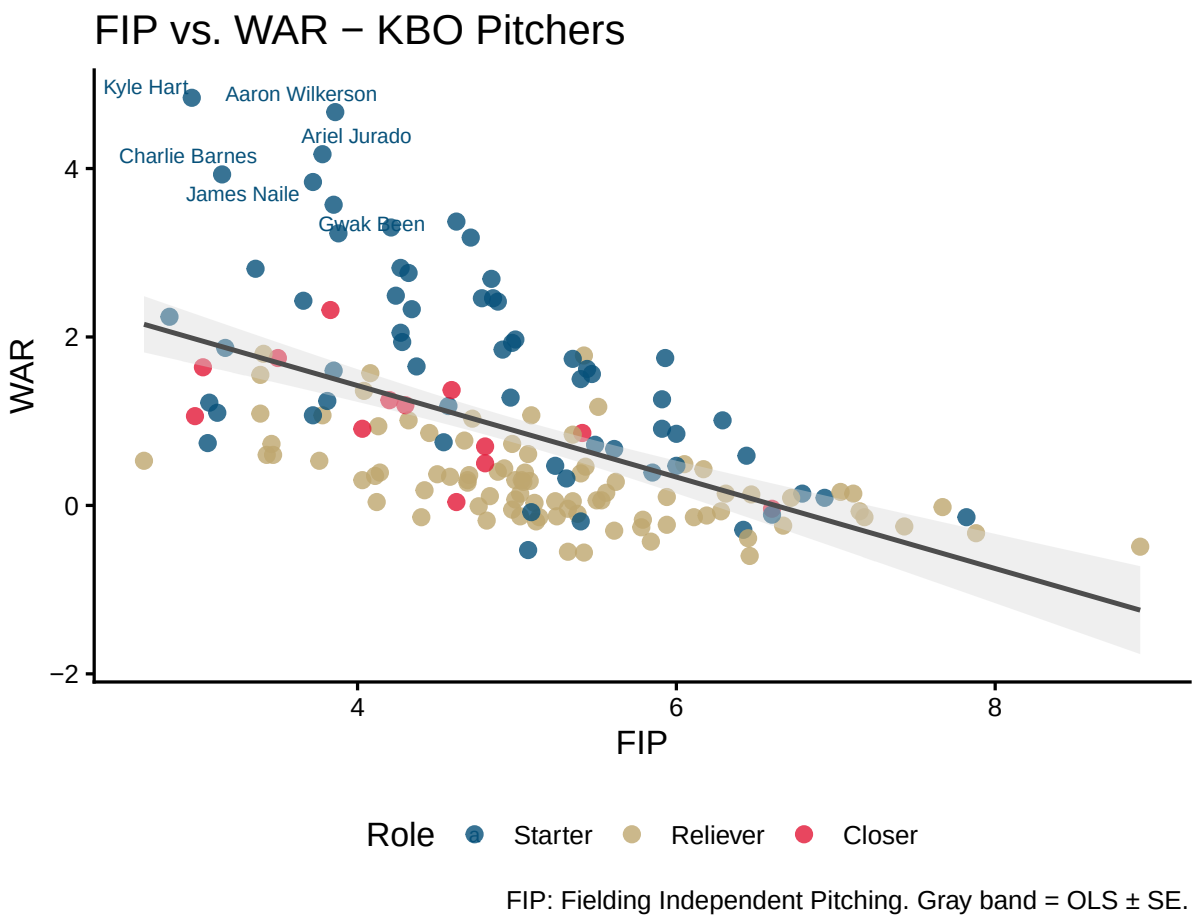


Figure 11: FIP (Fielding Independent Pitching) vs. WAR. FIP isolates outcomes the pitcher controls directly.

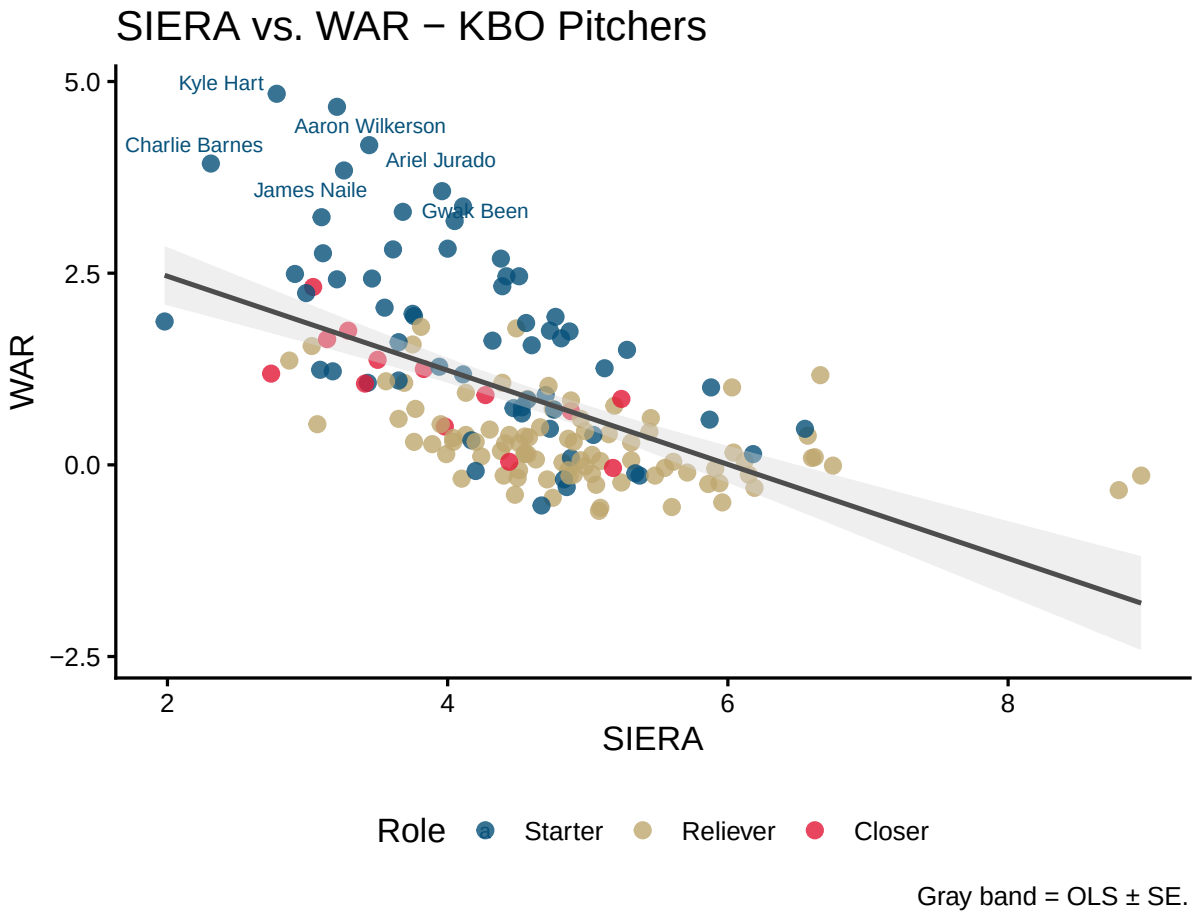


Figure 12: SIERA (Skill-Interactive ERA) vs. WAR. SIERA accounts for batted ball profiles and strikeout/walk rates.

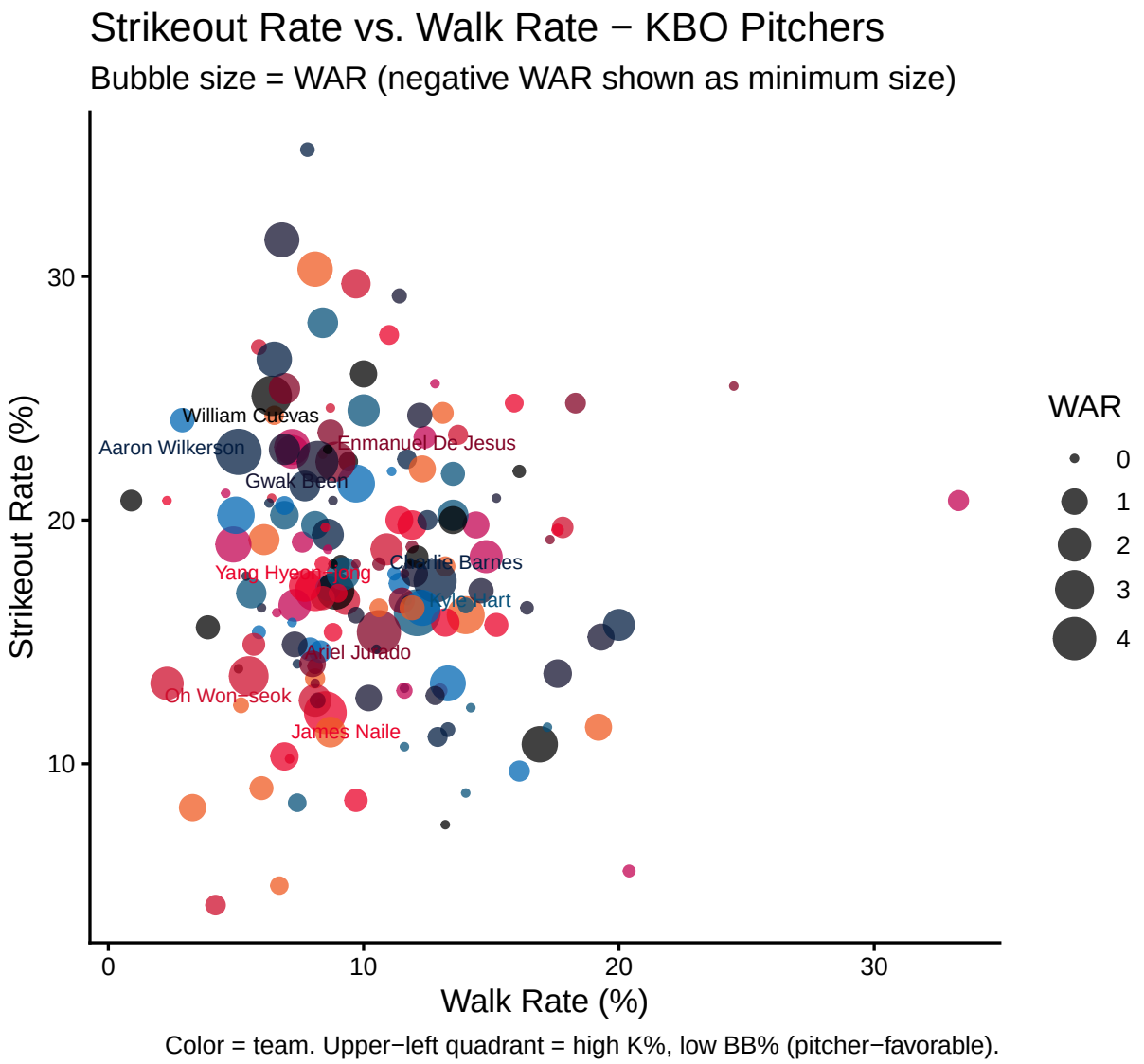


Figure 13: Strikeout rate vs. walk rate for all KBO pitchers, bubble size = WAR. The upper-left quadrant (high K%, low BB%) represents the most dominant profiles.

Highly Correlated Pairs

Table 8: Variable pairs with |r| > 0.8 (n = 29 pairs). High multicollinearity motivates the variable-selection procedure in the regression analysis.

Variable1	Variable2	Correlation
PlateApperances	InningsPitched	0.996
WalkRate	WalksPerNineInnings	0.988
FlyBallRate	GroundBallRate	-0.975
StrikeoutRate	StrikeoutsPerNineInnings	0.966
Hits	PlateApperances	0.965
Hits	InningsPitched	0.948
OPSAgainst	SLGAgainst	0.942
GamesRelieved	Games	0.938
Strikeouts	InningsPitched	0.933
Strikeouts	PlateApperances	0.918
Hits	EarnedRuns	0.906
Saves	GamesFinished	0.904
PlateApperances	GamesStarted	0.901
Hits	GamesStarted	0.890
InningsPitched	GamesStarted	0.889
OPSAgainst	AVGAgainst	0.876
Doubles	Hits	0.875
PlateApperances	EarnedRuns	0.861
InningsPitched	Wins	0.853
Strikeouts	Wins	0.847
EarnedRuns	GamesStarted	0.844
Doubles	EarnedRuns	0.843
Doubles	PlateApperances	0.839
PlateApperances	Wins	0.836
Strikeouts	Hits	0.834
InningsPitched	EarnedRuns	0.827
SLGAgainst	HomerunsPerNineInnings	0.822
Doubles	InningsPitched	0.820
SLGAgainst	AVGAgainst	0.806

Appendix D: Regression Analysis

Sections: Yeo-Johnson Transformations | Full Model | Stepwise Model Selection | Model Comparison | VIF and Multicollinearity Reduction | Final Model Diagnostics | Top Predicted Pitchers

Yeo-Johnson Transformations

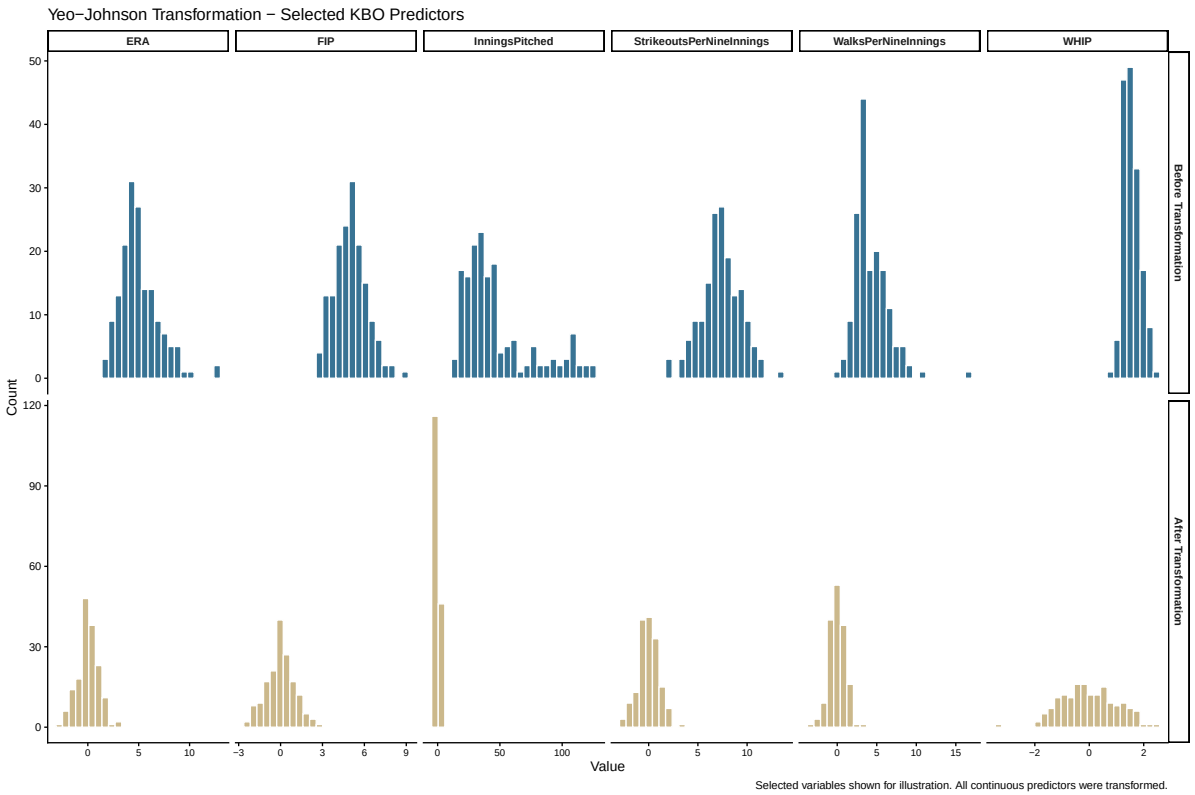


Figure 14: Before (left) and after (right) Yeo-Johnson transformation for selected continuous predictors. Transformation reduces skew and normalizes distributions.

Full Model

Table 9: Full model comparison (all continuous variables, with and without Position factor).

Model	Adj. R^2	Residual SE	BIC
All Continuous Variables	0.8635	0.4139	336.54
All + Position	0.8630	0.4146	352.06

Stepwise Model Selection

Table 10: Model comparison across all candidate models. Lower BIC = preferred model. Stepwise selection used AIC criterion in both directions.

Model	Adj. R^2	BIC
Stepwise (both)	0.8790	216.03
Stepwise (backward)	0.8790	216.03
Full (all continuous)	0.8635	336.54
Full + Position	0.8630	352.06
Stepwise (forward)	0.8630	352.06

Model Comparison

Table 11: Best model (lowest BIC) coefficient table — showing terms with p < 0.10. *** p < 0.001; ** p < 0.01; * p < 0.05; . p < 0.10.

Term	Estimate	CI Low	CI High	Std Err	t value	p-value	Sig
(Intercept)	0.8785	0.8180	0.9390	0.0306	28.6953	0.0000	***
GamesRelieved	-0.1684	-0.2590	-0.0778	0.0458	-3.6739	0.0003	***
EarnedRuns	1.4436	0.7774	2.1099	0.3371	4.2819	0.0000	***
InningsPitched	-3.5468	-5.0853	-2.0082	0.7786	-4.5554	0.0000	***
PlateApperances	1.3030	-0.1198	2.7259	0.7200	1.8097	0.0724	.
Hits	0.8025	0.3022	1.3027	0.2531	3.1700	0.0019	**
Homeruns	-0.0890	-0.1873	0.0093	0.0498	-1.7888	0.0757	.
Walks	0.2478	0.0431	0.4526	0.1036	2.3918	0.0180	*
Strikeouts	0.3222	0.1450	0.4994	0.0897	3.5926	0.0004	***
ERA	-1.3251	-1.7625	-0.8877	0.2213	-5.9868	0.0000	***
WHIP	-0.6990	-1.0147	-0.3834	0.1597	-4.3761	0.0000	***
HomerunsPerNineInnings	0.1410	0.0262	0.2558	0.0581	2.4272	0.0164	*
OBPAgainst	0.1338	-0.0085	0.2762	0.0720	1.8576	0.0652	.
OPSAgainst	-0.1931	-0.3712	-0.0150	0.0901	-2.1424	0.0338	*

VIF and Multicollinearity Reduction

Table 12: VIF values for the final reduced model. All values below threshold of 10 after iterative removal.

Variable	VIF
OPSAgainst	8.462
Strikeouts	7.593
OBPAgainst	5.384
Hits	4.943
WHIP	4.085
HomerunsPerNineInnings	3.487
ERA	2.988
Walks	2.640
Homeruns	2.546
GamesRelieved	1.699

Table 13: Final reduced model coefficients after VIF-based multicollinearity removal. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; . $p < 0.10$.

Term	Estimate	CI Low	CI High	Std Err	t value	p-value	Sig
(Intercept)	0.8785	0.8082	0.9487	0.0355	24.7187	0.0000	***
GamesRelieved	-0.3135	-0.4053	-0.2217	0.0465	-6.7483	0.0000	***
Hits	0.2295	0.0729	0.3861	0.0793	2.8949	0.0044	**
Homeruns	-0.0522	-0.1645	0.0602	0.0569	-0.9171	0.3605	
Walks	0.0540	-0.0605	0.1684	0.0579	0.9317	0.3530	
Strikeouts	0.3093	0.1152	0.5033	0.0982	3.1484	0.0020	**
ERA	-0.3583	-0.4800	-0.2365	0.0616	-5.8136	0.0000	***
WHIP	-0.2637	-0.4061	-0.1213	0.0721	-3.6599	0.0003	***
HomerunsPerNineInnings	0.1331	0.0016	0.2646	0.0666	1.9992	0.0474	*
OBPAgainst	0.1442	-0.0192	0.3077	0.0827	1.7435	0.0833	.
OPSAgainst	-0.2038	-0.4087	0.0011	0.1037	-1.9655	0.0512	.

Final Model Diagnostics

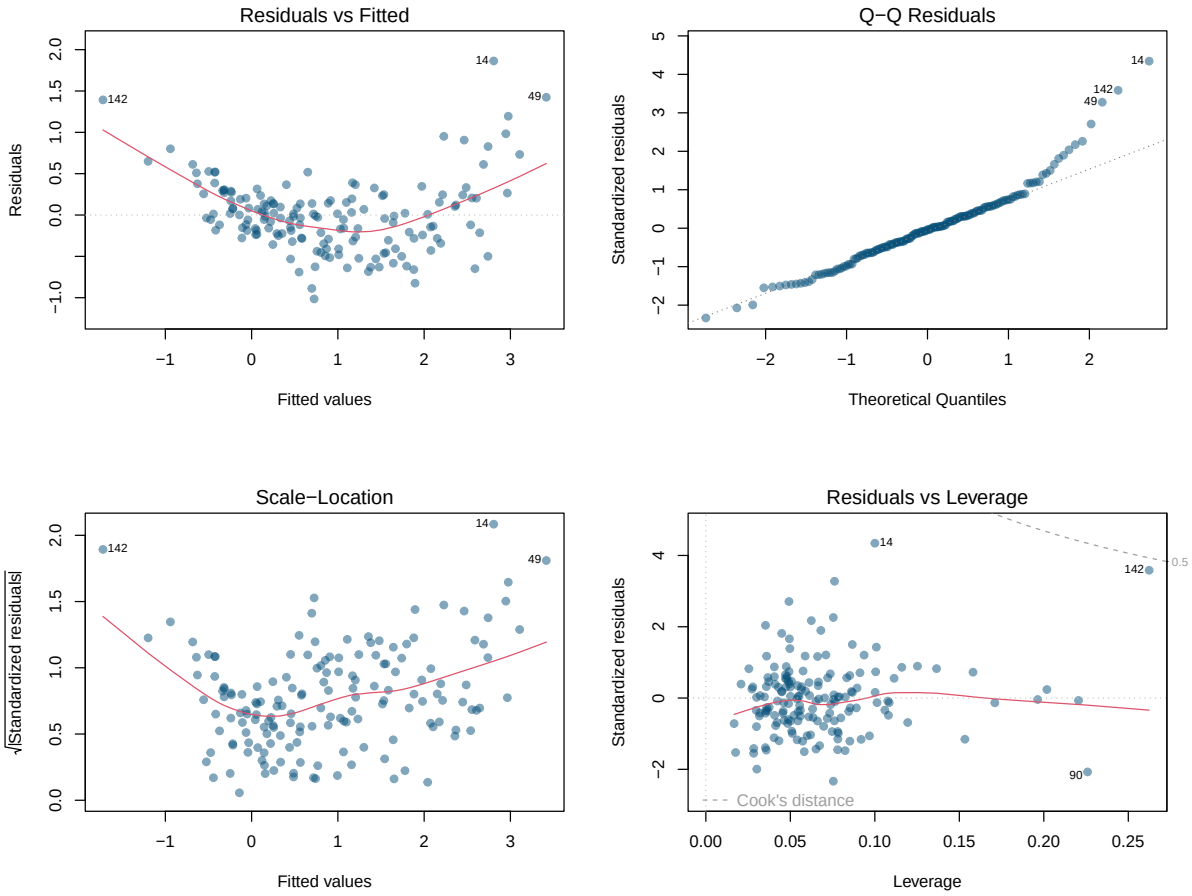


Figure 15: Diagnostic plots for the final reduced regression model. Top-left: residuals vs. fitted; top-right: Normal Q-Q; bottom-left: scale-location; bottom-right: residuals vs. leverage.

Table 14: Final reduced model goodness-of-fit statistics.

Metric	Value
R^2	0.8471
Adj. R^2	0.8370
Residual SE	0.4523
F stat	83.6616
p-value	0.0000
AIC	215.2988
BIC	252.3500
Resid. df	151.0000

Top Predicted Pitchers

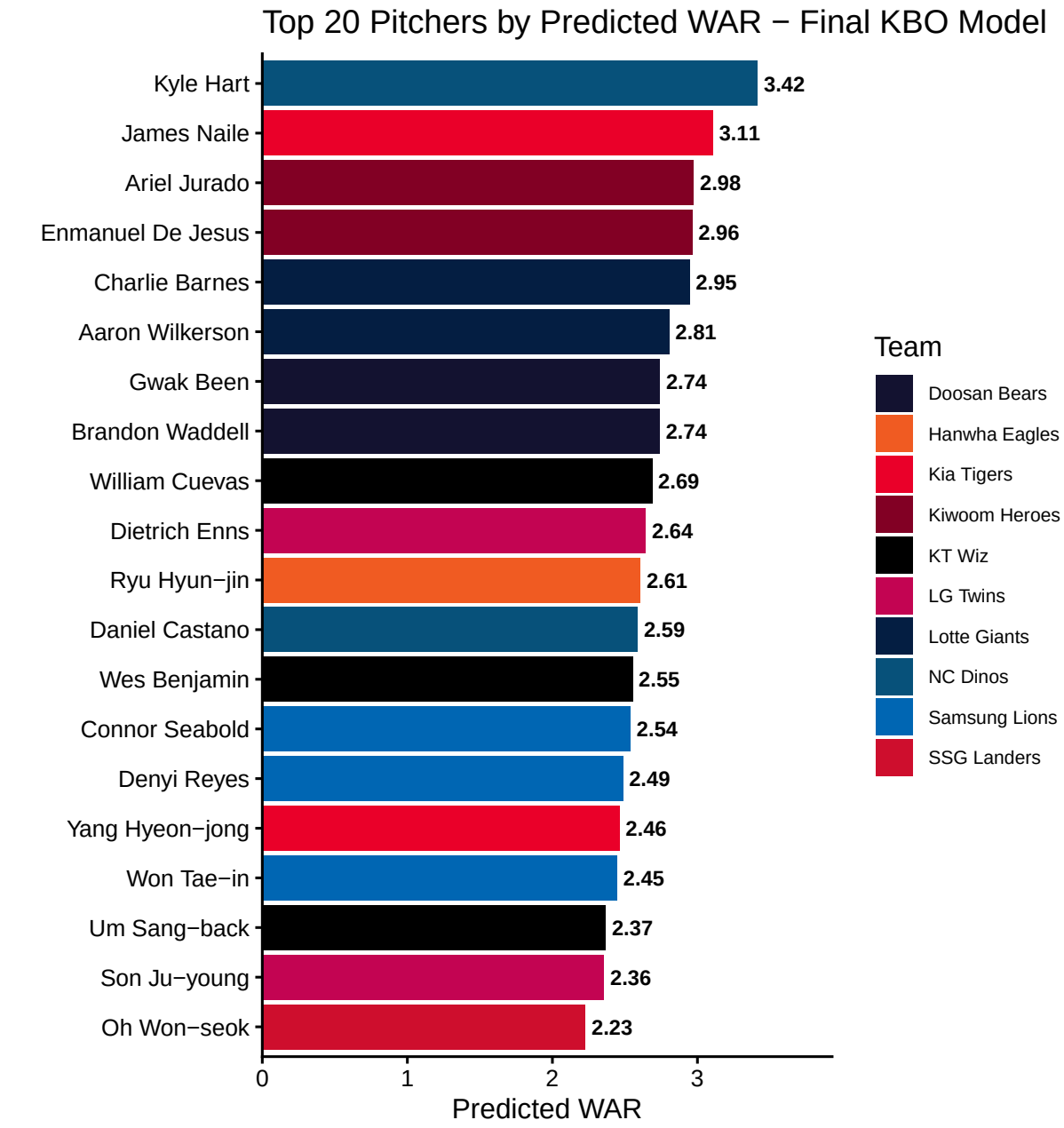


Figure 16: Top 20 KBO pitchers by model-predicted WAR from the final reduced regression model. Color = team.

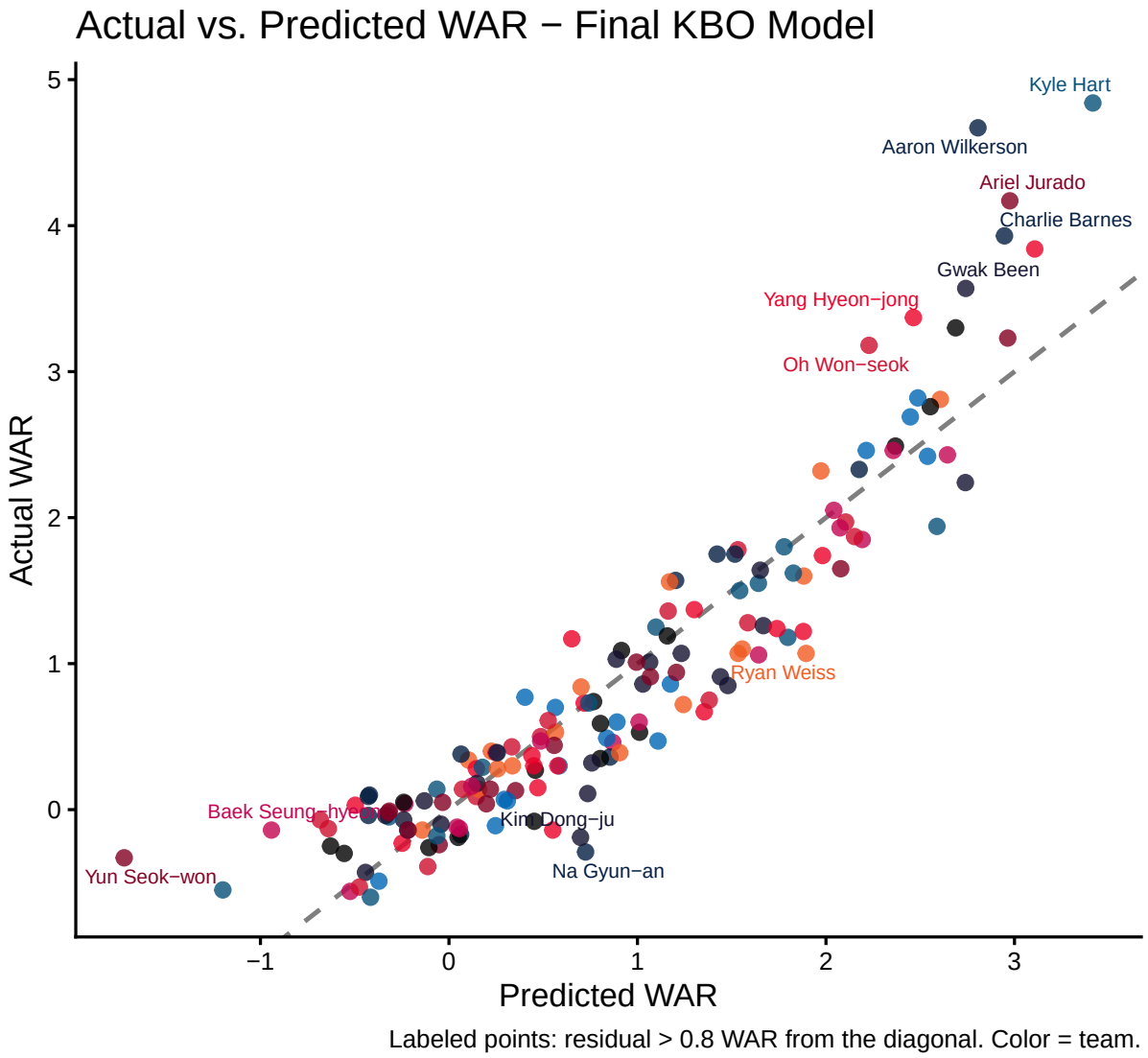


Figure 17: Actual vs. predicted WAR for all KBO pitchers from the final reduced model. Points near the diagonal indicate accurate predictions.

Table 15: Top 10 KBO pitchers by predicted WAR from the final reduced model.

Name	Team	Position	Actual WAR	Predicted WAR	Residual
Kyle Hart	NC Dinos	SP	4.84	3.416	1.424
James Naile	Kia Tigers	SP	3.84	3.107	0.733
Ariel Jurado	Kiwoom Heroes	SP	4.17	2.975	1.195
Enmanuel De Jesus	Kiwoom Heroes	SP	3.23	2.964	0.266
Charlie Barnes	Lotte Giants	SP	3.93	2.947	0.983
Aaron Wilkerson	Lotte Giants	SP	4.67	2.806	1.864
Gwak Been	Doosan Bears	SP	3.57	2.742	0.828
Brandon Waddell	Doosan Bears	SP	2.24	2.740	-0.500
William Cuevas	KT Wiz	SP	3.30	2.688	0.612
Dietrich Enns	LG Twins	SP	2.43	2.645	-0.215