Determining MVP Winners in Major League Baseball: A Statistical Baseline Analysis

By: Benjamin Michaels and Ethan Michaels[1]

August 25th, 2024

Github Repo

[1] Each author contributed equally to the design, coding & development, analysis, and writing of this project.

Abstract

This report investigates the predictive factors for determining the Most Valuable Player (MVP) in Major League Baseball (MLB). The central research question is “What statistical metrics are most indicative of a player being named MVP in the MLB?”. To answer this, the study employs statistical analysis and machine learning techniques, analyzing historical data on MVP winners and key performance indicators (KPIs) such as batting average, home runs, RBIs, and WAR (Wins Above Replacement).

This analysis reveals that while traditional metrics like batting average and home runs are significant, advanced metrics such as WAR and OPS (On- Base Plus Slugging) provide a more robust prediction of MVP outcomes. The study also explores the impact of team success and market size on MVP selection. The key finding is that a combination of traditional and advanced metrics, along with contextual factors like team performance, best predictsthe likelihood of a player winning the MVP Award. This approach not only offers a better understanding of MVP selection but also provides a framework for future predictions.

Background & Question

In the world of sports, players aim to be the best of the best, whether that is winning the end of the year championship (World Series, Super Bowl, etc.) or the individual awards. In most cases the most important individual award is the league’s Most Valuable Player.

The Most Valuable Player (MVP) award in Major League Baseball (MLB) is one of the most prestigious honors a player can receive, recognizing outstanding performance and contribution to their team. Despite its significance, the selection process for the MVP has often sparked debate among fans, analysts, and players alike. The subjective nature of the award, which takes into account various facets of a player's performance, raises the question: what combination of player performance metrics best predicts the winner of the Major League Baseball's MVP award.

Understanding the key metrics that most accurately predict the MVP winner not only offers insights into the criteria valued by voters but also enhances our comprehension of what constitutes "value" in a baseball context. Such an analysis can help to demystify the selection process, providing a more objective basis for predicting future MVPs and assessing player performance. It can also influence how players approach their development and how teams evaluate talent, ensuring that the most deserving players are recognized.

Moreover, this research has broader implications for statistical analysis in sports, as it contributes to the ongoing discourse on the quantification of player value—a topic of great importance in the era of advanced analytics. By identifying the optimal combination of performance metrics, this study aims to provide a more accurate and reliable model for predicting MVP outcomes, thereby enhancing both the analytical framework used in baseball and the overall understanding of player excellence.

We hypothesize that a player's offensive performance metrics (batting average, home runs, runs batted in, and on-base percentage) combined with advanced statistics

(such as WAR - Wins Above Replacement) are strong predictors of winning the American League and/or National League MVP Award.

Our prediction is that players with higher offensive performance metrics and superior advanced statistics are more likely to win the American League (AL) or National League (NL) Most Valuable Player award. By leveraging these metrics, it is possible to accurately forecast MVP award winners with a high degree of accuracy.

Data

The dataset used in our analysis is highly appropriate for addressing our research question for several key reasons. Firstly, it includes a comprehensive set of both basic and advanced player performance metrics, such as Wins Above Replacement (WAR), games played (G), runs (R), hits (H), home runs (HR), runs batted in (RBI), stolen bases (SB), batting average (BA), on-base percentage (OBP), slugging percentage (SLG), and on-base plus slugging (OPS). These metrics provide a robust foundation for analyzing the factors that most significantly contribute to a player's candidacy for the MVP award.

Second, the dataset's historical breadth, spanning several decades, encompasses the performance of players who were contenders for the MVP award in both the National League and American League. This historical depth allows for the identification of trends over time, thereby enhancing the predictive power of our analysis.

Additionally, the dataset includes variables directly related to MVP voting, such as "Vote.Pts," "1st.Place," and "Share," which enable a direct linkage between player

performance and their likelihood of winning the MVP award. The inclusion of the "Winners" variable, which indicates whether a player won the MVP, serves as a clear target variable for predictive modeling. Furthermore, the dataset distinguishes between the two leagues (American and National), which is essential given that the MLB awards two separate MVP honors. This distinction ensures that no player is incorrectly predicted for an award in a league in which they do not participate.

These characteristics collectively make the dataset a robust and relevant choice for exploring the relationship between player batting performance metrics and MVP outcomes in Major League Baseball.

Our report utilizes two distinct datasets. The first dataset comprises all MVP winners and contenders from 1956 to 2022[2], obtained through web scraping the Awards section from Baseball Reference and including both American League and National League MVPs. The second dataset contains the final batting performance data from 2023[3], which was used to evaluate our models' predictions for the most recent MVP awards.

The data cleaning process began by leveraging the high-quality data provided by Baseball Reference. Upon loading the dataset, we created a binary variable to identify the MVP winner for each year by ranking players based on MVP votes within their

[2] *2021 Awards Voting*. Baseball. (n.d.). https://www.baseball-reference.com/awards/awards_2021.shtml#AL_MVP_voting_link

[3] *2023 MLB Player Hitting Stat Leaders*. MLB.com. (n.d.). https://www.mlb.com/stats/batting-average/2023

respective year and league. Players with a rank of 1 were designated as MVPs, marked with a binary value of '1,' while all other contenders were assigned a value of '0.'

Next, we adjusted the 'Share' variable by converting it from a percentage to a decimal format, which involved removing the percentage sign and dividing the value by 100.

In terms of preprocessing, we focused specifically on data for batters post-1956, the year in which the Cy Young Award (exclusively for pitchers) was introduced. We included only players with over 150 at-bats, as those with fewer at-bats would likely not qualify as MVP candidates or were pitchers. Pitchers were excluded from our analysis for two primary reasons: first, they have a distinct set of statistics compared to batters, and second, pitchers accounted for only 11 of the 132 MVP winners (8.3% of the data). The sole exception was Shohei Ohtani, a two-way player (both pitching and batting); we excluded his pitching data rather than removing him from the dataset entirely. Finally, we ensured that the dataset of MLB MVPs and contenders was organized chronologically by year, facilitating easier navigation and analysis.

Methods and Models

In the modeling phase, we explored several predictive models, including Linear Regression, Lasso Regression, Ridge Regression, Neural Networks, and Random Forest. Following a thorough analysis, we excluded Lasso and Ridge Regression from this report due to their outcomes being consistently identical to those produced by the Linear Regression model. Additionally, challenges with the Neural Network model, coupled with time constraints, prevented its successful implementation. Consequently,

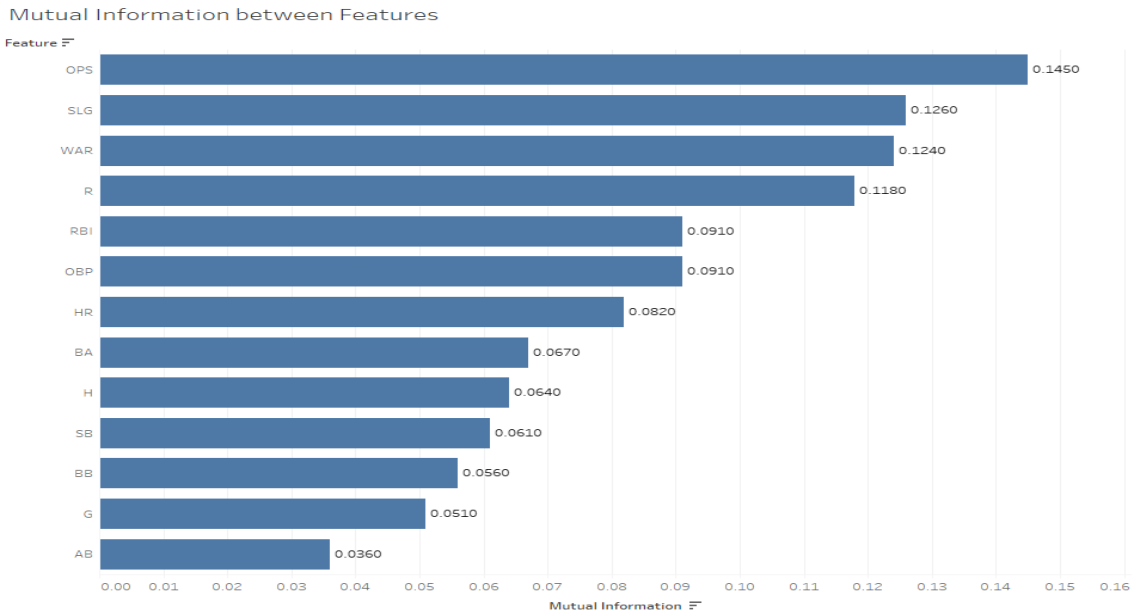
we focused on two primary models—Linear Regression and Random Forest—to identify the most effective method for predicting MLB MVPs.

Before implementing the predictive models, we evaluated three distinct methods to establish a foundation for feature selection: Correlation Matrix, Mutual Information, and Descriptive Statistics. The descriptive statistics, detailed in Appendix B, provided an initial overview of the data. For our correlation analysis, we examined the relationships between all independent variables and the dependent variable, 'Share,' visualized in Figure 1. The final method is examining Mutual Information between ‘Share’ and the ensuing independent variables. Mutual Information is a measure of dependency between two variables. In the context of machine learning, it helps quantify how much knowing one feature reduces uncertainty about another. Mutual Information values range from 0 to 1, with 1 indicating perfect dependency, where one feature completely determines another. Figure 2 illustrates the mutual information between various features and 'Share.' Notably, OPS has the highest score at 0.145, indicating it is the most informative feature, followed closely by SLG, WAR, and R. Despite OPS and SLG having the highest mutual information scores, these metrics were excluded from further analysis due to redundancy; OPS is derived by summing SLG and OBP, which compromises the independence of these features.The results gathered from Figure 1 and Figure 2, informed our decision to include the following variables in the models: 'WAR,' 'SB,' 'HR,' 'RBI,' 'BA,' and 'OBP.

Figure 1: Correlation Matrix Of All Relevant Statistics

	Share	WAR	G	AB	R	H	HR	RBI	SB	BB	BA	OBP	SLG	OPS
Share	1.00	0.39	0.05	0.04	0.33	0.15	0.28	0.27	0.06	0.15	0.22	0.24	0.37	0.36
WAR	0.39	1.00	0.26	0.19	0.58	0.35	0.22	0.13	0.29	0.40	0.36	0.48	0.35	0.44
G	0.05	0.26	1.00	0.89	0.52	0.69	0.20	0.43	0.07	0.22	-0.11	-0.14	-0.19	-0.19
AB	0.04	0.19	0.89	1.00	0.53	0.84	0.04	0.31	0.20	-0.12	0.00	-0.33	-0.29	-0.34
R	0.33	0.58	0.52	0.53	1.00	0.58	0.42	0.45	0.32	0.39	0.26	0.32	0.37	0.39
H	0.15	0.35	0.69	0.84	0.58	1.00	-0.08	0.26	0.23	-0.10	0.53	0.07	-0.07	-0.03
HR	0.28	0.22	0.20	0.04	0.42	-0.08	1.00	0.76	-0.29	0.42	-0.19	0.17	0.77	0.62
RBI	0.27	0.13	0.43	0.31	0.45	0.26	0.76	1.00	-0.32	0.31	0.01	0.14	0.57	0.46
SB	0.06	0.29	0.07	0.20	0.32	0.23	-0.29	-0.32	1.00	-0.09	0.12	-0.01	-0.23	-0.17
BB	0.15	0.40	0.22	-0.12	0.39	-0.10	0.42	0.31	-0.09	1.00	0.01	0.72	0.40	0.57
BA	0.22	0.36	-0.11	0.00	0.26	0.53	-0.19	0.01	0.12	0.01	1.00	0.64	0.34	0.50
OBP	0.24	0.48	-0.14	-0.33	0.32	0.07	0.17	0.14	-0.01	0.72	0.64	1.00	0.58	0.81
SLG	0.37	0.35	-0.19	-0.29	0.37	-0.07	0.77	0.57	-0.23	0.40	0.34	0.58	1.00	0.95
OPS	0.36	0.44	-0.19	-0.34	0.39	-0.03	0.62	0.46	-0.17	0.57	0.50	0.81	0.95	1.00

Figure 2: Mutual Information Between Features



To evaluate which model best addressed our research question, we assessed their performance using the root mean square error (RMSE), coefficient of determination (R^2), and an accuracy rating based on the percentage of accurately predicted MVP winners from 1956-2022 for both the American League and National

League. Below, we present the results and interpretations for both the Linear Regression and Random Forest models.

Linear Regression

The output of the Linear Regression analysis, with "Share" as the dependent variable and "WAR," "SB," "HR," "RBI," "BA," and "OBP" as independent variables, is displayed below.

Figure 3: Linear Regression Formula

```
Call:
lm(formula = Share ~ WAR + SB + HR + RBI + BA + OBP, data = train.data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.43302 -0.16316 -0.03381  0.13205  0.71346

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.4941957  0.0902164  -5.478 5.42e-08 ***
WAR          0.0308313  0.0043616   7.069 2.90e-12 ***
SB           0.0011292  0.0005075   2.225  0.02630 *
HR           0.0031521  0.0009865   3.195  0.00144 **
RBI          0.0012769  0.0004554   2.804  0.00514 **
BA           1.4846237  0.3668011   4.047 5.57e-05 ***
OBP          0.0905180  0.2583660   0.350  0.72615
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.211 on 1018 degrees of freedom
Multiple R-squared:  0.2232,    Adjusted R-squared:  0.2186
F-statistic: 48.75 on 6 and 1018 DF,  p-value: < 2.2e-16
```

Examining the distribution of the residuals reveals a range from -0.433 to 0.714, with a median of -0.03. The residuals appear to be reasonably symmetric around the median, suggesting a good fit for the model. Notably, the intercept, representing the expected value of "Share" when all independent variables are zero, is negative, indicating a baseline "Share" percentage of -49.41%. For instance, for each unit increase in WAR, "Share" increases by 3.08%, and for every 1% increase in batting average, "Share" increases by 1.48% . All variables, except for On-Base Percentage, are statistically significant.

The Linear Regression model yielded an R^2 value of 0.211 and an MSE of 0.507. To evaluate the model's predictive accuracy, we applied it to the dataset spanning 1956-2022, predicting MVP outcomes. The model accurately predicted 31 out of 67 American League MVPs and 35 out of 67 National League MVPs, resulting in an accuracy of 46% for the AL MVP and 53% for the NL MVP. These results suggest that the Linear Regression model may not be the most suitable for our hypothesis. However, the Random Forest model remains to be tested.

Random Forest

One of the key strengths of the Random Forest model is its ability to quantify the importance of each feature in making predictions, typically measured by the reduction in mean squared error (MSE) when a particular feature is used for splitting nodes in the trees. Below, we present our Random Forest model for predicting MLB MVP winners.

Figure 4: Random Forest Model

```
Call:
 randomForest(formula = Share ~ WAR + SB + HR + RBI + BA + OBP,      data = train.data, method =
"rf", trControl = control, tuneGrid = tuneGrid,      importance = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2

          Mean of squared residuals: 0.04493771
                    % Var explained: 21.31
```

The Random Forest model achieved an R^2 value of 0.2131 and an MSE of 0.025. To assess the model's accuracy, we generated confusion matrices for both American League and National League contenders, as shown below:

Figure 5: American League Confusion Matrix

	Actual

Predicted		0	1
	0	444	5
	1	5	62

Figure 6: National League Confusion Matrix

	Actual		
Predicted		0	1
	0	447	7
	1	7	60

In Figure 5, we present the confusion matrix for American League MVP contenders and winners. The table shows the actual and predicted outcomes, where '0' represents contenders (non-winners) and '1' represents winners. The model correctly identified 444 contenders and 62 winners, with an overall accuracy of 98.06%. Focusing specifically on predicting the winners, the model achieved an accuracy of 92.53%.

Figure 6, displays the confusion matrix for National League MVP contenders and winners, structured similarly to the American League matrix. The overall accuracy for NL MVP contenders and winners was 97.31%, with an accuracy of 89.55% in predicting the winners.

When combining the predictions for both the AL and NL MVPs, the model accurately predicted 122 out of 134 MVPs, resulting in an overall accuracy of 91.04%.

This accuracy is well-aligned with our research predictions, in contrast to the Linear Regression model.

Given the high accuracy of the Random Forest model, we conducted an additional test using the 2023 MVP data to see if the model could accurately predict the MVP winners from the 2023 season. We ran the model 100 times and documented the frequency with which each player was predicted to win MVP based on their statistics. The results are shown below in Figure 7 for the American League and Figure 8 for the National League.

Figure 7: Random Forest 2023 American League Predictions

Player	Occurrences
Shohei Ohtani	66%
Corey Seager	34%

Figure 8: Random Forest 2023 National League Predictions

Player	Occurrences
Ronald Acuña Jr.	76%
Matt Olson	24%

Our model predicted Shohei Ohtani as the 2023 AL MVP in 66% of simulations, with Corey Seager winning 34% of the time. Based on these results, we would predict Shohei Ohtani as the 2023 AL MVP, which indeed he was. For the National League, Ronald Acuña Jr. was predicted to win the NL MVP in 76% of simulations, with Matt

Olson winning 24% of the time. Again, our prediction was accurate, as Ronald Acuña Jr. did win the 2023 NL MVP. Thus, using our Random Forest model, we were able to accurately predict both the AL and NL MVP winners for the 2023 season.

Conclusion

In conclusion, the random forest model proved to be a powerful tool in predicting MLB MVP winners, offering a robust framework for understanding the metrics that potential voters value the most. This model could serve as a valuable resource for analysts, journalists, teams, and players alike, influencing future player development and evaluation strategies.

Discussion & Next Steps

In our analysis, we sought to determine the key player performance metrics that best predict the MLB MVP award, hypothesizing that a combination of traditional offensive metrics and advanced statistics like WAR would be strong predictors. The Random Forest model emerged as the most effective in addressing this question, achieving an overall accuracy of 91.04% in predicting MVP winners across both the American and National Leagues (Figures 4 and 5). This model's performance, demonstrated by its ability to accurately predict the 2023 MVP winners (Figures 6 and 7), confirms the importance of a multi-faceted approach that incorporates both traditional and advanced metrics. However, while the model performed well, it is important to acknowledge certain caveats. The exclusion of pitchers (except Shohei Ohtani's batting data) and the focus on post-1956 data may limit the generalizability of the findings. Additionally, the linear regression model's lower accuracy (46% for AL and 53% for NL) suggests that other factors not captured in our analysis might also

influence MVP voting. It should also be mentioned that it is not possible to predict injuries for any player in the MLB, thus if an MVP candidate gets injured then the chances of them winning the award is significantly reduced. Finally, the process of voting for the winner of the MVP award solely comes down to 30 individual journalists,which means that there is an introduction of human bias in that process.

Future analyses could explore the inclusion of more advanced metrics or even sentiment analysis from media coverage, which might provide further insights into the subjective aspects of MVP selection. From a journalistic perspective, these findings would help further the media in what to cover and when to cover certain stories based on who is ultimately predicted to win the award, depending on how often this model is run throughout the season. From a management perspective, these findings can guide teams in player development and talent evaluation, emphasizing the metrics that most strongly correlate with MVP outcomes.

Appendix A: Data Dictionary

Data Dictionary:

Rank

Description: Ranking of players in MVP Voting.
Options:

- 1: First
- 2: Second
- 3: Third
- 4: Fourth
- 5: Fifth
- 6: Sixth
- 7: Seventh
- 8: Eighth
- 9: Ninth

Name

Description: The name of the player.

Tm

Description: Team that the player played on that year of voting.

League

Description: Major League baseball is split equally into 2 leagues. League indicates which league the specific team was a part of.
Options:

- AL: American League
- NL: National League

Year

Description: Indicates the year in which the voting had occurred.

Vote.Pts

Description: Is the total number of points that is obtained from the amounts of placements per ballot. (Ex: 1st place is 14 pts, 2nd place is 9 pts, etc. from 30 ballots per league)

X1st.Place

Description: Number of 1st place votes

Share

Description: The vote points divided by most points possible. Unanimous choice is 100%.

WAR

Description: Wins Above Replacement. A single number that presents the number of wins the player added to the team above what a replacement player from the minor leagues would add.

- 8+ is MVP Quality
- 5+ is All-Star Quality
- 2+ is Starter
- 0-2 is Reserve
- < 0 is Replacement Level

G

Description: Games Played. The number of games played by the player in that specific season.

AB

Description: At Bats. The number of at bats the player had in the specific season.

R

Description: Runs Scored. The amount of times the player has scored during that season.

H

Description: Hits. The number of hits the player had during that season.

HR

Description: Home Runs. Number of home runs the player hit during that season.

RBI

Description: Runs Batted In. Is the number of times where the result of the player's plate appearance is a run being scored during that season.

SB

Description: Stolen Bases. THe number of stolen bases by the player during that season.

BB

Description: Bases on Balls/Walks. The number of times the player has been walked during their plate appearance over the course of the season.

BA

Description: Batting Average. H / AB

OBP

Description: On-base Percentage. (H + BB + HBP) / (AB + BB + HBP + SF)

- HBP: Hit By Pitch
- SF: Sacrifice Fly hit

SLG

Description: Slugging. Total Bases divided by At Bats **OR** (1B + 2*2B + 3*3B + 4*HR) / AB

- 1B: Single
- 2B: Double
- 3B: Triple

OPS

Description: On-base Plus Slugging. One-Base Percentage + Slugging Percentage.

Winners

Description: Indicates who won the MVP Award that season.
Options:

- 1: Won
- 0: Lost

Appendix B: Descriptive Statistics

Figure 9: Descriptive Statistics of All Major League Baseball MVP Winners

	Share	WAR	G	AB	R	H	HR	RBI	SB	BB	BA	OBP	SLG	OPS
Min	0.43	2.5	60.0	214.0	43.0	73.0	2.0	35.0	0	18.0	0.257	0.319	0.373	0.72
1Q	0.79	6.1	147.2	528.5	97.0	158.2	25.3	96.7	3.25	54.0	0.301	0.374	0.531	0.90
Mean	0.85	7.51	149.4	560.4	106.9	177.7	34.5	110.1	10.0	78.3	0.317	0.403	0.580	0.984
3Q	0.96	8.6	158.0	613.5	120.0	194.8	45.0	126.0	15.4	94.8	0.330	0.425	0.630	1.03
Max	1.00	12.5	165.0	716.0	143.0	242.0	73.0	158.0	104.0	232.0	0.390	0.609	0.863	1.42

Figure 10: Descriptive Statistics of All Major League Baseball MVP Contenders

	Share	WAR	G	AB	R	H	HR	RBI	SB	BB	BA	OBP	SLG	OPS
Min	0.08	0.3	47.0	154.0	33.0	52	0.0	21.0	0.0	10.0	0.230	0.259	0.279	0.538
1Q	0.21	4.6	146.0	532.0	86.0	155	21.0	86.0	3.0	50.0	0.287	0.358	0.490	0.853
Mean	0.35	5.8	148.0	560.3	96.9	170.4	28.7	99.4	12.2	69.4	0.304	0.382	0.532	0.914
3Q	0.48	7.0	158.0	605.0	109.0	188.0	37.0	115.0	17.0	85.0	0.320	0.404	0.578	0.970
Max	0.8	11	164.0	705.0	152.0	262.0	70.0	165.0	118.0	162.0	0.396	0.526	0.752	1.250

Appendix C: MVP and Contenders Occurrences by Team

Figure 11: Frequency of Winners and Contenders by Team

Abbreviation	Team Name	MVP Winners	Top 9 Contenders
ATL	Atlanta Braves	6	50
ARI	Arizona DiamondBacks	0	8
BAL	Baltimore Orioles	5	41
BOS	Boston Red Sox	7	70
CHC	Chicago Cubs	6	32
CHW	Chicago White Sox	5	35
CIN	Cincinnati Reds	9	52
CLE	Cleveland Guardians	0	34
COL	Colorado Rockies	1	21
DET	Detroit Tigers	2	38
HOU	Houston Astros	2	33
KCR	Kansas City Royals	1	22
LAA	Los Angeles Angels	1	31
LAD	Los Angeles Dodgers	4	56
MIA	Miami Marlins	1	7
MIL	Milwaukee Brewers	4	26
MIN	Minnesota Twins	5	36
NYM	New York Mets	0	28
NYY	New York Yankees	11	69
OAK	Oakland Athletics	5	29
PHI	Philadelphia Phillies	6	36
PIT	Pittsburgh Pirates	7	39
SD	San Diego Padres	1	21
SEA	Seattle Mariners	2	21
SF	San Francisco Giants	10	47
STL	St. Louis Cardinals	9	54
TBR	Tampa Bay Rays	0	7
TEX	Texas Rangers	6	23
TOR	Toronto Blue Jays	2	29
WSH	Washington Nationals	1	36