Michigan Technological University®

Analyzing the Hitting Environment in Major League Baseball

MA5781 Time Series Analysis - Dr. Rho
Jonathan Oliveros and Braden Barglind

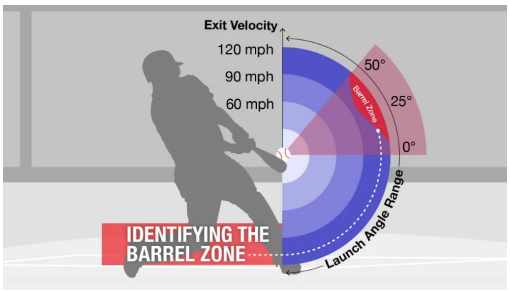
Abstract

Major League Baseball is full of statistics, and this report is meant to dive into a few of them and explain what is currently happening in the game. A few questions that will be answered after analysis of simple statistics includes “How has the hitting environment changed over time? Is there an increase, decrease, or neither in these statistics? Has the fan experience changed over time been positive or not, and is the game continuing down that path? A time series analysis of home runs, hits, runs, and strikeouts will help determine the answers to these questions, as well as give more insight on other issues or topics that surround the game of baseball.

Contents

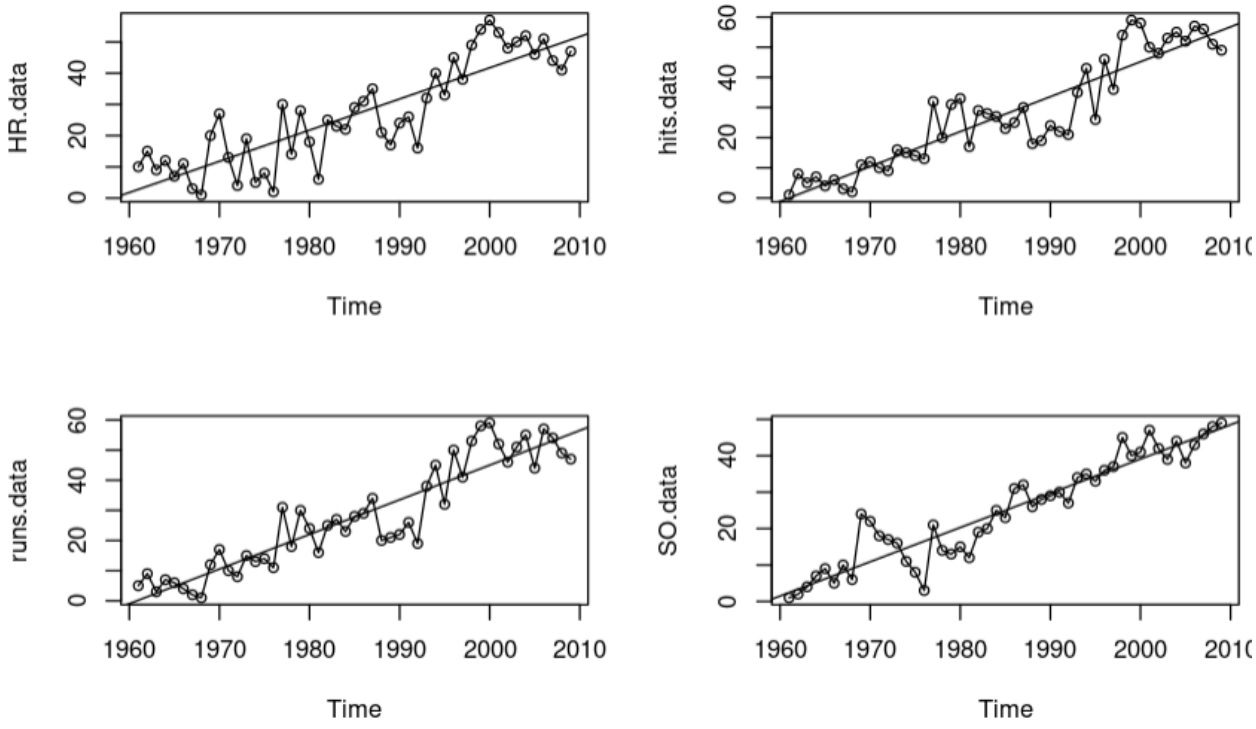
Introduction:

The game of baseball is constantly evolving. In fact, it has been referred to by many as being the most analytically-driven game in the history of sports, and having an edge on numbers and statistics as a franchise can help put their teams over the hump. Throughout this report, analysis of time series data of home runs, strikeouts, runs, and hits will be performed to answer a few important questions. First, how has the hitter environment changed over time? Off of these results, how are changes in the hitting environment linked to teams becoming more analytically advanced? Lastly, has the game become more entertaining to the average sports goer, and how does the future of entertainment look in the future? Data has been gathered through the Baseball Almanac from 1961 to 2019 with the variables listed above, with each season being scaled to a complete 162-game season. It is important to understand that not every season is equal in every way, as there are variables that slightly change over time. For example, MLB has been implementing new types of baseballs to the players, and this at times results in record lows in areas, as well as record highs, in certain years. On top of this, there have been stretches of seasons where steroid usage was untrackable, so it is difficult to know when this occurred, if much occurred at all. Nevertheless, the data being used has the ability to bring a lot of insight in the questions asked.



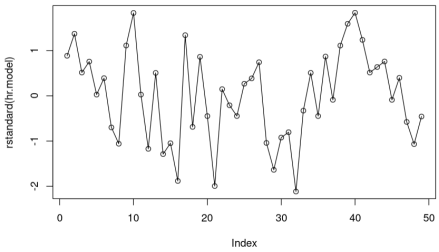
Time Series Data

This section will provide more detail on the time-series data that is being analyzed. Plotted is the time-plot for each of the four variables introduced. Included is a line of best fit, which follows some trends more than others. From these lines, it can be hypothesized that each variable has a positive linear trend, and that each of the four variables increase over time. Throughout this report, analysis of the true model for each will be described, forecasted values will be obtained, as well as tests against the slope for each to determine if there truly is a positive, linear relationship. Just by looking at these plots, it's possible to have a preliminary answer to the questions asked, but much more analysis must be done to confidently say.

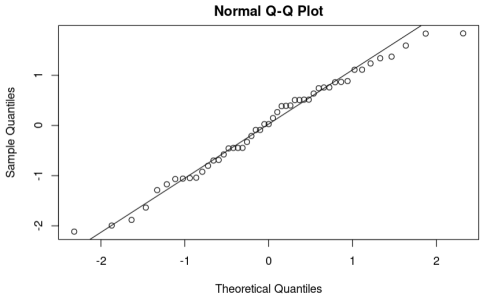


Residual Analysis

In order to proceed with the time series analysis, each variable's residuals must be looked at. To reduce the length of the report, the complete model building process of home runs will be explained, with a small explanation on runs, strikeouts, and hits as it progresses. The same process follows for each of the four variables, but the purpose of the report is to be able to answer the questions listed above.



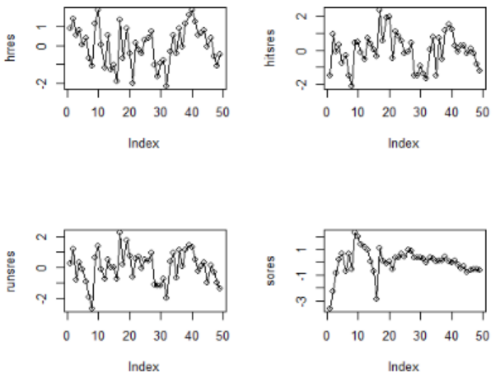
Looking at the residuals of home runs, there were some things worth being noted. The residual plot appears to have a zero mean, and variability does not increase over time. As well, none of the points exceed +/- 3 on the y-axis. So, the homoscedasticity assumption holds. In addition, a quantile-quantile plot was constructed, on top of testing for normality with the Shapiro-Wilk test. As can be seen, the Q-Q plot line follows the points nicely, and the results from the Shapiro-Wilk test suggest the data is normally distributed. Lastly, to test for independence, the ACF plot and runs test were produced. There appears to be time dependence looking at lag 1 on the ACF plot. However, the runs test produced a p-value of 0.401, resulting in the failure of the rejection of the null hypothesis, and can conclude that the home runs data is independent. In the end, it has been determined that the home runs data does not require a transformation to fulfill the need for the assumptions, however it will still be attempted to see if another model can fit the data with more accuracy.

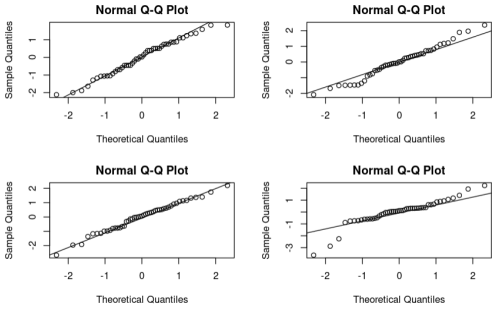



```
        Shapiro-Wilk normality test

data:  hrres
W = 0.97858, p-value = 0.5074
```

Transformations

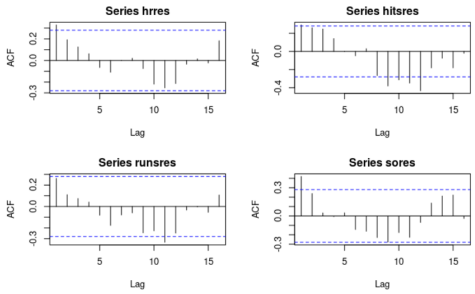
After taking a look at the residual analysis of the non-transformed variables, runs, strikeouts, and hits must be transformed to meet the required assumptions, but home runs do not. Among the tested transformations include log transformation, as well as square-root and cube-root transformations. It was determined, after testing out models, that the square root of both runs and hits would suffice the assumptions, as well as the log of strikeouts. As can be seen from the residual plots, each variable outside of strikeouts seems to have zero mean and satisfies the heteroscedasticity assumption. Looking at the Q-Q plots and output of the Shapiro-Wilk test, each variable besides strikeouts fulfills the normality assumption. After analyzing the ACF plots, it appears there is time dependence in each variable, but the runs test indicates randomness for all variables but strikeouts. The variables besides strikeouts grades out well. The adjusted R-Squared for all of these models are either as good or better than the original, non-transformed models. In summary, it was clearly hard to transform strikeouts to fit all of the assumptions in order to proceed. This is something to note as analysis is performed down the road, as not everything is perfect with the residual analysis. Nevertheless, proceeding and understanding the interpretation and accuracy of the strikeouts model may have flaws that can't be corrected given the techniques known.





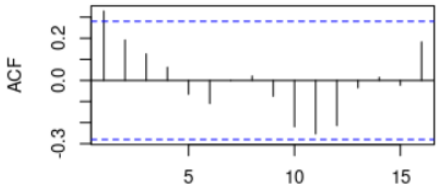

```
        Shapiro-Wilk normality test

data:  hrres
W = 0.97858, p-value = 0.5074

        Shapiro-Wilk normality test

data:  hitsres
W = 0.98146, p-value = 0.6277

        Shapiro-Wilk normality test

data:  runsres
W = 0.98754, p-value = 0.8795

        Shapiro-Wilk normality test

data:  sores
W = 0.88791, p-value = 0.0002265
```

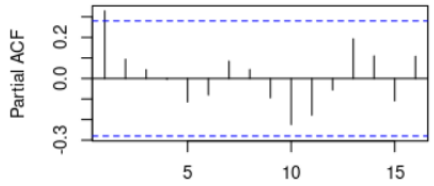
Residual	p-value	Observed Runs	expected Runs
HR	0.401	22	25.40816
Hits	1	25	25.4898
Runs	0.829	24	25.2449
SO	0.0398	17	24.26531



Candidate Time Series Models

In this section, the candidates for potential time series models for each variable will be described. Here, we will be looking at the home run data. In order to determine the P and Q in an ARMA(p,q) model, analyzing the ACF, PACF, and EACF is needed. It can be seen from the ACF plot that there is significance at lag one, but that is it. It can be seen from the PACF plot there is significance at lag one as well, with no other lags being significant. Combining this information from what is told from the EACF plot, an AR(3) or MA(1) appears to fit best for this data. Repeating this process with runs, strikeouts, and hits yielded interesting candidate models. For runs, an AR(1) model was considered, and that's the only model that fit the three plots. Attempting to model hits yielded candidates AR(1) and MA(3). Lastly, candidate models for strikeouts are AR(5) and MA(1). Having multiple candidates is ideal for a model, as it gives flexibility in which model is chosen, as well as added predictive power. For all variables, the KPSS test results yielded insignificant, with the conclusion being all are stationary. The KPSS test was chosen as a final decision, as the ADF test and PP test yielded different conclusions.






```
[1] "HR"
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x o o o o o o o o o o  o  o  o
1 x o o o o o o o o o o  o  o  o
2 x o o o o o o o o o o  o  o  o
3 o o o o o o o o o o o  o  o  o
4 o o o o o o o o o o o  o  o  o
5 x x o o o o o o o o o  o  o  o
6 x x o o o o o o o o o  o  o  o
7 x x o o o o o o o o o  o  o  o
```

```
        KPSS Test for Level Stationarity

data:  hrres
KPSS Level = 0.12324, Truncation lag parameter = 3, p-value = 0.1
```

Model Selection

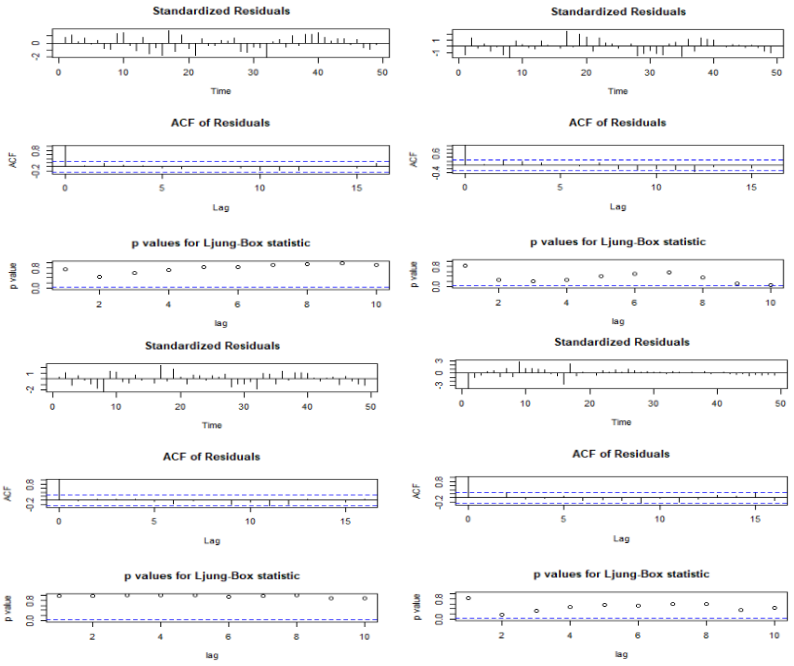
Choosing a model for each variable requires running the proposed models that were introduced above, and comparing the statistics and output. In choosing a model, AIC and BIC are commonly used. The overall goal is to get both these statistics as low as possible. AIC is compared if one would like a model that forecasts well, and BIC is compared if one would like the true model. Here, for home runs, the two proposed models were built, and the output is interesting. Clearly, the AIC and BIC were both smaller using the MA(1) model than compared to the AR(3) model. Therefore, this model will be chosen. It is worth mentioning as well that a model with less terms is typically ideal too, and that holds here. For runs, the model chosen was already AR(1), so there was nothing else to do. For strikeouts, both the AIC and BIC were better for the MA(1) model as compared to the AR(5) model, so that model was chosen. Lastly, for hits, the AIC and BIC are better for the MA(1) model than the AR(3) model, so that is chosen too. Now that the models have been chosen for each variable, looking at model diagnostics is next.

```
Series: hrres
ARIMA(0,0,1) with non-zero mean

Coefficients:
         ma1    mean
      0.2672  0.0043
s.e.  0.1211  0.1719

sigma^2 = 0.9477:  log likelihood = -67.23
AIC=140.46   AICc=140.99   BIC=146.13
Series: hrres
ARIMA(3,0,0) with non-zero mean
```

```
Coefficients:
         ar1     ar2     ar3    mean
      0.2909  0.0942  0.0482  0.0103
s.e.  0.1411  0.1487  0.1432  0.2307

sigma^2 = 0.9517:  log likelihood = -66.3
AIC=142.61   AICc=144   BIC=152.07
Series: hrres
ARIMA(1,0,0) with non-zero mean
```

Model Diagnostics

Model diagnostics involves testing the goodness of fit of a model. Basically, how well the predicted values fit with the actual values. Here, the standardized residuals, ACF plots, and Ljung-Box test will be constructed and performed. For each variable, the standardized residuals look to be good for all, as there are not any that stretch vertically very far in either direction. Looking at the ACF plot, there are no significant lags other than lag zero, so this also looks good. For the Ljung-Box, where the null hypothesis is that the model does not show lack of fit, and the alternative hypothesis is that the model does show a lack of fit. After looking at the Ljung-Box plot for each variable, it can be seen that none of the points seem to be significant at any lag, and it can be confidently said that each model fits the data well. The results of these three are very encouraging, and provide even more support that the models were built with strong statistical methods and procedures.



Overfitting

In this section, fitting models that are overfit was done with each variable. As it can be seen from the AIC and AICC the overfit models performed worse for homeruns,

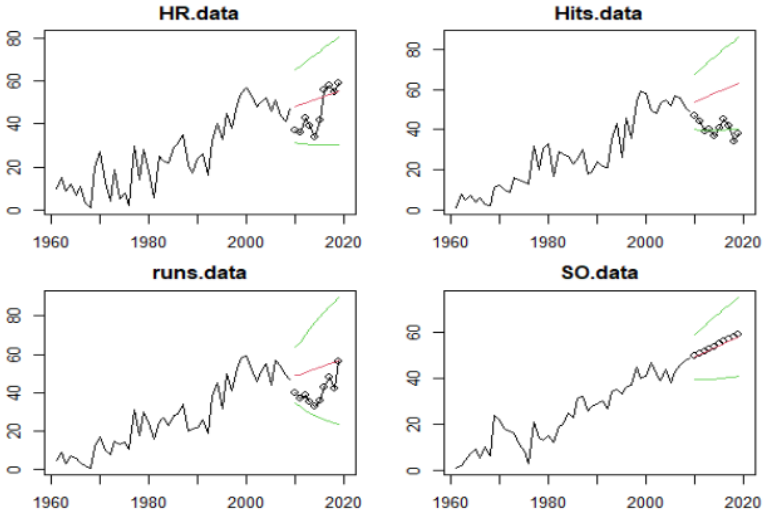
hits, and strikeouts, but actually performed better for strikeouts. For all but the strikeouts, this is the result we want to see. However, for strikeouts, this is not what we would like to see. After looking at the EACF plot, however, it can be concluded that an IMA(2,1) model is not something we are interested in, and will not go ahead with this. Instead, keeping the original ARIMA model built, an IMA(1,1), is the conclusion we will be working with going forward. Looking at the AIC and AICC was important, as lower values demonstrate more accuracy for a forecasting model, which is something we are interested in doing. If we were to look at the BIC, we would be interested in the true model, but this is not the ultimate goal to help answer our questions.

	Model	AIC	AICC	BIC
	hr IMA(1,1)	**140.5766**	**140.8432**	144.3190
Overfit →	hr IMA(2,1)	**141.9884**	**142.5338**	147.6020
	hits IMA(1,1)	**139.3295**	**139.5962**	143.0719
Overfit →	hits IMA(2,1)	**141.1190**	**141.6645**	146.7326
	runs ARI(1,1)	**150.0535**	**150.3202**	153.7959
Overfit →	runs ARI(1,2)	**148.5819**	**149.1273**	154.1955
	so IMA(1,1)	**130.5302**	**130.7969**	134.2726
Overfit →	so IMA(2,1)	**132.4430**	**132.9884**	138.0566

Forecasting

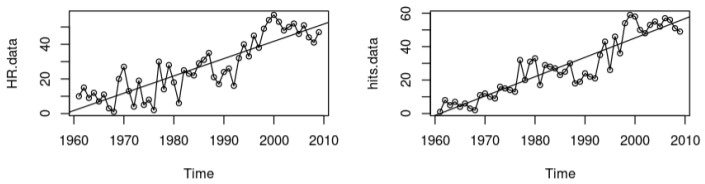
In this section, forecasting was done using the testing data from 2010 to 2019. This is meant to see how well the predicted values during this time match up with the actual values. It can be seen in each graph that the red line is the prediction mean, and the green lines are the 95% confidence prediction interval bounds. Looking at the home runs

and runs forecasts, they are nearly the same, with all of the points falling in the interval. Most of the points don't fall near the prediction mean line, which tells us that the predictions are not great. For the hits data, it can be seen that a few of the points fall outside of this 95% confidence interval, as this is not ideal as well. However, for strikeouts, the actual values fall nearly perfectly with the mean predicted values, which tells us the strikeout model performed very well during this stretch of seasons. It's important to remember that this type of data does include a lot of randomness, so predictions might be difficult year over year. This analysis can help answer a few of the questions asked before. It can be concluded that the forecasted intervals look to be trending upwards for each variable, and this tells us a few things. For one, fans who like home runs and more exciting action are liking today's game, and will likely continue to like it in the future. However, fans who don't enjoy increased strikeouts, and would prefer the ball in play more often no matter the result, may not enjoy where the game of baseball is headed in the future. This data offers a lot of insight on the future of Major League Baseball, and even how revenue might change over the coming years based on fan's preferences.



Slope Analysis

Answering the questions "How has the game changed over time?", and "Has the game become more entertaining for fans?" is achievable in this section by testing



the slope coefficient for home runs, runs, strikeouts, and hits using the HAC estimator. Shown is the time plot with the OLS-line for each variable to make it easier to observe. In addition, the HAC coefficient test results are provided to determine if the slope terms are significant or not. As predicted earlier, it is clear that the slope terms are all significant, and it can be concluded there is a positive linear relationship between each variable and time. Now, it can safely be said that the game has taken on a more hitter-friendly shape over the years, and defense (pitching) has not caught up. It is widely known that fans enjoy action in a sport. Because of this, it can also be safely concluded that the state of the game is in a better place from an entertainment perspective. Public research has indicated there is a wide variety of reasons why some years are better than others for fan experience, but it's been proven that people and fans want instant gratification. That is, more home runs, more hits, and more runs. Now, on the flipside, player's striking out more frequently each year does not provide action, right? Well, there is clearly a tradeoff occurring in the game. If a player were to strike out more as a result of swinging at more pitches or swinging harder, they may end up with more hits, home runs, and runs scored. Clearly, the analytics department of Major League teams have also coined this to be the most effective way to win a ballgame, otherwise this probably wouldn't be occurring. Will we see even more offense and even less defense down the line? That is something we won't truly know until it happens.

```
                Estimate  Std. Error t value  Pr(>|t|)
(Intercept)   -1956.01684   226.75817 -8.6260 3.013e-11 ***
time(HR.data)     0.99888     0.11445  8.7278 2.136e-11 ***

                Estimate  Std. Error t value  Pr(>|t|)
(Intercept)   -2.2713e+03  1.6068e+02 -14.136 < 2.2e-16 ***
time(hits.data) 1.1583e+00  8.1379e-02  14.233 < 2.2e-16 ***

                Estimate  Std. Error t value  Pr(>|t|)
(Intercept)   -2.2534e+03  1.6373e+02 -13.763 < 2.2e-16 ***
time(runs.data) 1.1492e+00  8.2857e-02  13.869 < 2.2e-16 ***

                Estimate  Std. Error t value  Pr(>|t|)
(Intercept)   -1.8419e+03  1.2405e+02 -14.848 < 2.2e-16 ***
time(SO.data)  9.4051e-01  6.2183e-02  15.125 < 2.2e-16 ***
```

Conclusion

In conclusion, after finishing the time series analysis process for each of the four hitting statistics, true models were found, and forecasting was completed. It was determined that home runs, runs, hits, and strikeouts are all positively increasing over time linearly, and we can expect this trend to continue based on the forecasting results. Therefore, we can answer how the hitting environment has changed over time, and can point to the increase in strikeouts leading to more home runs, hits, and runs, as batters are willing to sacrifice for more power. In addition, because teams have become so analytically-driven, it can be assumed that the general trend teams are teaching to hitters is exactly this; To be more productive on the field, the upside of hitting for power comes with more strikeouts, but it will ultimately result in more of each statistic, and help the team offensively overall. As mentioned previously, it has been proven that fans like action in a game, and it appears that baseball has had increased action during the game over the years. Therefore, we believe the state of the game is in a good place, and believe this will continue for the entertainment of fans down the road.

Works Cited

MLB Advanced Stats, Statcast, and historical data. Fantasy Sports Resource.(2022, January 22). Retrieved June 25, 2022, from https://www.fantasysportsresource.com/mlb-advanced-stats-and-statcast-data/

MLB stats, scores, History, & Records. Baseball. (n.d.). Retrieved June 25, 2022, from https://www.baseball-reference.com/