Using Statcast Data to Calculate Expected Batting Average on Contact

By: Dieter Erben, Zach Gifford, Luc Maynor

Executive Summary

BACON, Batting Average on Contact, attempts to show a batter's quality of contact by measuring the number of batted balls that result in a hit or a home run. It differs from the traditional batting average since it subtracts strikeouts from at bats, measuring only the balls that were put in play. This statistic is similar to BABIP, Batting Average on Balls In Play, although BACON includes home runs while BABIP does not. In the light of these statistics, the main purpose of this project was to generate an expected BACON, or xBACON, to predict batted ball results based on contact quality. This statistic can be used to measure a player's batted ball skill, and to look at outliers to identify players expected to bounce back when their actual BACON is below their xBACON, or to regress when their actual BACON is above their xBACON. After looking at 62 different variables from Statcast data, 4 significant variables were chosen to generate a xBACON: launch angle, exit velocity, x coordinate and y coordinate. Based on these four variables, a classification tree was created to model xBACON, using approximately 230 nodes to increase accuracy without overfitting data. The model correctly predicted the outcome of 88.2% of the batted balls, with outs predicted more accurately (91.6%) than hits (82%).

The accuracy of the model can still be improved, as external factors such as park, weather and speed were not included. Additionally, this model does not take into account defensive positioning, and this variable could increase the model’s accuracy. Looking forward, this model could be used to create a hit outcome predictor, or to evaluate a pitcher’s contact quality allowed. Lastly, other sports could use this methodology to identify expected points or goals based on the quality of a player’s shot.

Project Objective

This project seeks to develop a model to quantify quality of contact for Major League Baseball (MLB) hitters. The project focuses on batting average on contact (BACON), measured as hits divided by batted balls. The necessary MLB data for training and testing the model is available publicly through Baseball Savant. The model uses binary tree classification to predict whether a batted ball with particular characteristics should be a hit. From the predictions for each player, an expected BACON (xBACON) is derived. Using xBACON, we can evaluate what a given player's BACON *should have been*, based solely on his contact quality. This method isolates player batted ball skill and is intended as a tool to measure hitter performance.

Traditionally, hitters have been measured by their batting average and home run totals, and more recently by more advanced metrics like batting average on balls in play (BABIP), isolated slugging percentage (ISO), home runs per fly ball (HR/FB), and weighted on-base average (wOBA). These statistics continue to be important in evaluating players. xBACON is intended to supplement other metrics to determine if players performed as expected, if they benefited from luck on batted balls, or were impeded by poor luck on batted balls.

Background Information

The primary motivation to pursue this project is the advent of Statcast batted ball data, made available to the public in full beginning with the 2015 season. MLB broadcasts began to

incorporate Statcast measures, like exit velocity and launch angle, when describing game action. There is a growing body of evidence that balls hit in the air are dramatically more productive than balls hit on the ground. This information appears to have entered the front office and trickled down to individual players, most notably Josh Donaldson and Justin Turner (Sawchik, 2017).

Research into batted ball data began with HITf/x, tracked by Sportvision. HITf/x debuted in 2009 and was the primary source for batted ball data until 2015. Full HITf/x data was never made available, even for teams, and the tracking system had numerous limitations. Thus, research and application of HITf/x era batted ball data is extremely limited.

The two most complete analyses were authored by Alan Nathan on Baseball Prospectus. In 2013, he determined that the initial velocity vector of a batted ball was not enough to accurately determine the distance which the ball traveled. He suggests that distance variation is not significantly impacted by wind, though the drag coefficient and spin on the baseball are major factors in the distance variance (Nathan, 2013).

In 2014, Nathan revisited his previous study. This time, he determined that the primary factor causing distance variance was the drag properties of a given baseball. The properties of the ball vary league-to-league, and even the balls for a given league displayed significant distance variance. This study determined that backspin was not a major factor in distance traveled, attributed to the increase in drag with increases in spin (Nathan, et. all, 2014).

In 2016, using Statcast batted ball data, Nathan analyzed batted ball flight variance attributable to park factors. His conclusion quantified what many had thought for years, that balls hit at similar or identical exit velocities and launch angles traveled different distances depending upon the stadium in which they were hit. These differences are attributable to factors such as temperature, humidity, and air pressure, among other possibilities (Nathan, 2016).

In May of 2016, Rob Arthur of FiveThirtyEight studied players with the largest changes in OPS from 2015 to 2016, which was compared to those players' changes in contact quality. This study set out to determine which players who had improved their OPS had done so by improving their batted ball profile, and which players whose improvement was due to luck. Additionally, the study analyzed players with the largest performance dropoffs to see which players might be unlucky. Arthur found that the players who improved their batted ball profile the most from 2015 to 2016 were younger players, which might suggest that these players are better able to consciously make changes to generate better contact quality. However, we note that approximately two months of batted ball data is likely too small a sample to determine whether these changes were significant. Additionally, with only two seasons of data, it is too early to tell whether the changes from 2015 to 2016 will continue to hold in future seasons (Arthur, 2016).

Lastly, Bill Petti of The Hardball Times published a study which sought to predict hits using Statcast batted ball data. The present study intends to build upon Petti's by using more batted ball data and applying it to the full 2016 season. Whereas Petti trained his model on

approximately 5,700 batted ball records from the 2016 season and tested against approximately 32,300 batted ball records also from the 2016 season, our model trains using about 110,000 records from 2015 and is tested against a similar number of records from 2016. Our method ensures that any potential self-fitting bias is eliminated from the data (Petti, 2016).

Methodology

Data Set

The primary data for this analysis was obtained from Baseball Savant Statcast Search. Baseball Savant has comprehensive data on batted balls and pitches, as well as aggregate data for players. Additionally, FanGraphs was used for players' season statistics. FanGraphs has comprehensive baseball data, including a wide array of both basic box score statistics and advanced statistics.

The 2015 season was chosen as our cutoff date because the full Statcast data, including batted ball exit velocity and launch angle, was made public beginning with the 2015 season. There were 62 variables in the initial data set, which we condensed to 4 significant variables, including exit velocity, launch angle, x hit location, and y hit location. Some of the variables that we tested, but were found to be insignificant included Speed Score, Handedness, Pitch Location, and Spin Rate. Over 100,000 records of batted ball data are included in each of the 2015 and 2016 seasons.

A few steps were taken to clean the data set to the specifications of the modeling method. Given that we classified hit outcomes, we developed a binary variable called “Hit” using an IF statement for the various outcomes spelled out in a categorical variable called “Events,” which described the outcome of the batted ball. Additionally, all entries with “null” or “0” values in any of our predictors were removed from the model, and we identify these as “untracked batted balls.” Then, we corrected our model with a constant derived from these deleted records. This constant, which we named “uBACON”, was derived as the percentage of hits resulting from records deleted from the model.

The primary data described above was gathered for the purpose of creating the classification tree model; however, more information was necessary to analyze the results. This data was taken from Fangraphs, and primarily included traditional statistics such as hits, strikeouts, and at-bats. Obtaining this data allowed us to examine specific players in order to identify those who overperformed and underperformed. This data required a slight amount of cleaning because of the differences in player names between Statcast and Fangraphs (Enrique vs. Kiké Hernandez).

Model

In order to create an effective model with thousands of entries, SPSS was the statistics software package that we utilized. From this, we derived a classification tree from the 2015 data set. After trying various methods such as logistic regression, k-Nearest Neighbors, and K-Means Clustering, the classification was determined to have the strongest outcome after assessing the information in the confusion matrix. Also, we ran multiple trials of the classification tree, so that

significant variables could be isolated within the model. This classification tree contained approximately 230 nodes, which we felt was a high enough number to effectively classify the data without overfitting the model. Once the model was created using the 2015 data, it was exported to a PMML file and imported to an SPSS file containing the 2016 data. Leveraging the power of the SPSS software allowed us to avoid partitioning the data, and attests to the strength of the model, given that it incorporates all records. Once we completed the model testing for the 2016 data, the data set was imported back into Excel for further analysis.

Model Analysis

Once the results of the model were imported into Excel, we created tables to analyze the results and step closer to deriving our xBACON statistic. Further data was also imported for the sake of analyzing the effectiveness of our xBACON stat. In Excel, a data table was created, giving each individual player their own record, rather than having the records be individual batted balls. This allowed us to aggregate the model's performance and helped to isolate specific players. Using the data taken from Fangraphs, the actual BACON was calculated for each individual player and then compared to our derived xBACON statistic. To obtain the xBACON statistic, we took the count of hits from our model's classifications, and added the expected count of hits using our uBACON constant. This provided an adjustment for untracked batted balls, which we called "xuHITS". Once we obtained our final xBACON measure, we calculated the difference between expected and actual BACON for each player, and made a separate column for its absolute value. Gathering all of this information in a data table made for an easy way to identify abnormal players, which will be discussed more in our Findings section.

Model Performance

Overall, the results of the model were strong from the beginning, with promising results from the initial model. During the process, we worried that the model was strong only because of overfitting to the data it was created with, but that notion was dispelled once the model was tested. Some statistics to highlight on the model include an error rate of 11.2%, sensitivity of 81.4%, and specificity of 93.0%. After testing the data on the 2016 data set, a confusion matrix was derived to show how the model actually performed once tested on a new set of data. The model was not biased by being trained to the 2016 data, so this is a stronger indicator of performance. Below is the confusion matrix for the test data:

Classification - 2016 Results			
	Predicted		
Observed	Out	Hit	% Correct
Out	65240	5976	91.6%
Hit	7061	32115	82.0%
% of Observations	65.5%	34.5%	88.2%

From this table, we can observe that the overall error rate was 11.8%. Looking further in depth at performance metrics, specificity was measured at 91.6% and sensitivity was 81.9%. Overall, this shows that the model was effective at achieving the goal it sets out to achieve.

Application of xBACON

Ultimately, the goal of this project was to quantify hitter performance, and to identify players due for a breakout and regression. By using a threshold of 100 tracked batting balls, we identified three players due to regress in 2017 based on 2016 performance, three that we expect

sustained success, and three that we expect improvement in 2017. Below is a chart outlining these players:

Notable Players			
Player	Tracked batted balls	xBACON	BACON
Jose Reyes	169	.383	.330
Jason Heyward	365	.321	.279
Juan Lagares	100	.321	.296
Bryce Harper	339	.315	.316
Carlos Correa	387	.362	.361
Mookie Betts	516	.356	.361
Jhonny Peralta	217	.245	.322
Jorge Soler	135	.257	.335
Willson Contreras	157	.306	.384

These players were not necessarily those with the highest error rates, as a common theme of speed emerged from both sides of the spectrum. Instead, we sifted through roughly the top-15 players with abnormally high error rates, and identified those with average speed. Given our lack of speed data, this was done with mostly reputation in mind, and help with statistics such as stolen bases. One specific player that is important to highlight is Jason Heyward, who just completed the first year of a massive 8 year/$184 million contract with the Cubs. After finishing the 2016 season with a .230 batting average, some may consider the contract to be a bust. However, our model is optimistic for 2017 based on the quality of contact Heyward displayed last season, and expects a bounceback season.

Potential for Improvement

The model might be improved assigning probabilities that batted balls with certain characteristics will be a hit. The model currently gives a batted ball a value of 0 or 1 to each record, which assumes a batted ball has a 0% or 100% chance of being a hit based on the measured variables. If

the model used probabilities, it could proportionately reward players for the quality of each batted ball contact. For example, if a ball hit at a 90 MPH exit velocity and 10 degree launch angle is a hit 50% of the time, the model could reward the hitter with 0.5 hit for that batted ball. These probabilities could be summed to determine an expected number of hits, and divided by batted balls to determine xBACON.

Further, the model in its current or improved state could be improved by incorporating handedness, defensive positioning, park factors, and batter speed. Shifts are a prominent part of the defensive game, and have proven effective at limiting BACON and BABIP by a small amount (Sarris 2017). Park factors are available at FanGraphs guts, and include adjustments for various hit statistics. Batter speed likely plays a significant factor in determining infield hits, since a faster runner is more likely to beat out a soft infield ground ball to earn a hit than a slower slugger. Overall, the model's greatest weakness was identifying Hits. We can attribute this difficulty to strong showings of defense, which some may refer to as "Web Gems." Errors are ignored in this model because they are not classified as hits, so only defensive plays that make a positive impact are highlighted. On the contrary, the model's difficulty in identifying outs can be attributed to speed. In the instance that a speedy player, such as Dee Gordon or Billy Hamilton, hits a groundball to the third baseman, he has a heightened likelihood of scoring a single because of his speed. In further iterations of this model, we would like to include significant variables that can quantify speed and defense.

Results, Findings, and Recommendations

Results and Findings

This project achieved the objective of quantifying the quality of batted balls. Overall, there was a promising error rate for the data, and the performance metrics all touted the strength of the classification tree. Players due for regression were also identified, and the 2017 season will show as evidence if these predictions can hold true. One important takeaway is that the data to achieve an even better model may not exist yet, but it is necessary to identify the flaws in the classification of batted balls. Furthermore, there is a promising foundation to go beyond the scope of BACON, in order to quantify more traditional baseball statistics such as AVG and OBP.

Further Research

There are additional applications for this model with slight variations. A similar model might be used to predict hit outcome (i.e., the probability that a batted ball goes for a single, double, triple, home run, or an out). The model could be expanded to include an expected strikeout rate and walk rate, from which one could derive an entire triple slash line (xAVG/xOBP/xSLG) and expected wOBA (xwOBA). Other expected statistics, such as expected ISO, expected home runs, and xBABIP could also be derived. Further, the model in any form could be applied on pitchers' batted ball against data to evaluate contact quality allowed. Also, there are various non-baseball applications for this study. Examples include: evaluating shot quality in hockey and lacrosse to calculate expected goals, or evaluating expected shots made in basketball.

References

Arthur, Rob. "Who's Hitting The Ball Harder This Year, And Who's Just Getting Lucky?" *FiveThirtyEight.* May 26, 2016. https://fivethirtyeight.com/features/whos-hitting-the-ball-harder-this-year-and-whos-just-getting-lucky/

Judge, Jonathan, Nick Wheatley-Schaller, and Sean O'Rourke. "The Need For Adjusted Exit Velocity." *Baseball Prospectus*. May 17, 2016. http://www.baseballprospectus.com/article.php?articleid=29210

Nathan, Alan. "Going Deep on Goin' Deep." *The Hardball Times*. April 6, 2016. http://www.hardballtimes.com/going-deep-on-goin-deep/

Nathan, Alan. "How Far Did That Fly Ball Travel?" *Baseball Prospectus*. January 8, 2013. http://www.baseballprospectus.com/article.php?articleid=19322

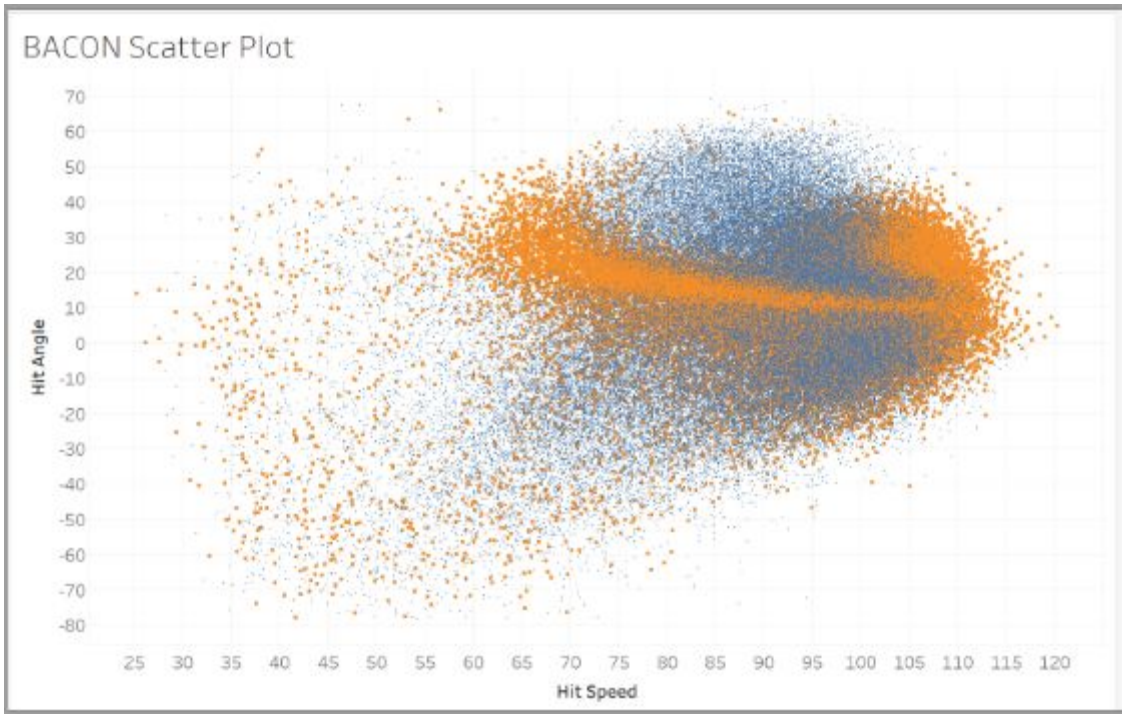
Nathan, Alan, Jeff Kensrud, Lloyd Smith, and Eric Lang. "How Far Did That Fly Ball Travel (Redux)?" *Baseball Prospectus*. December 9, 2014. http://www.baseballprospectus.com/article.php?articleid=25167

Petti, Bill. "Using Statcast Data to Predict Hits." *The Hardball Times.* June 14, 2016. http://www.hardballtimes.com/using-statcast-data-to-predict-hits/

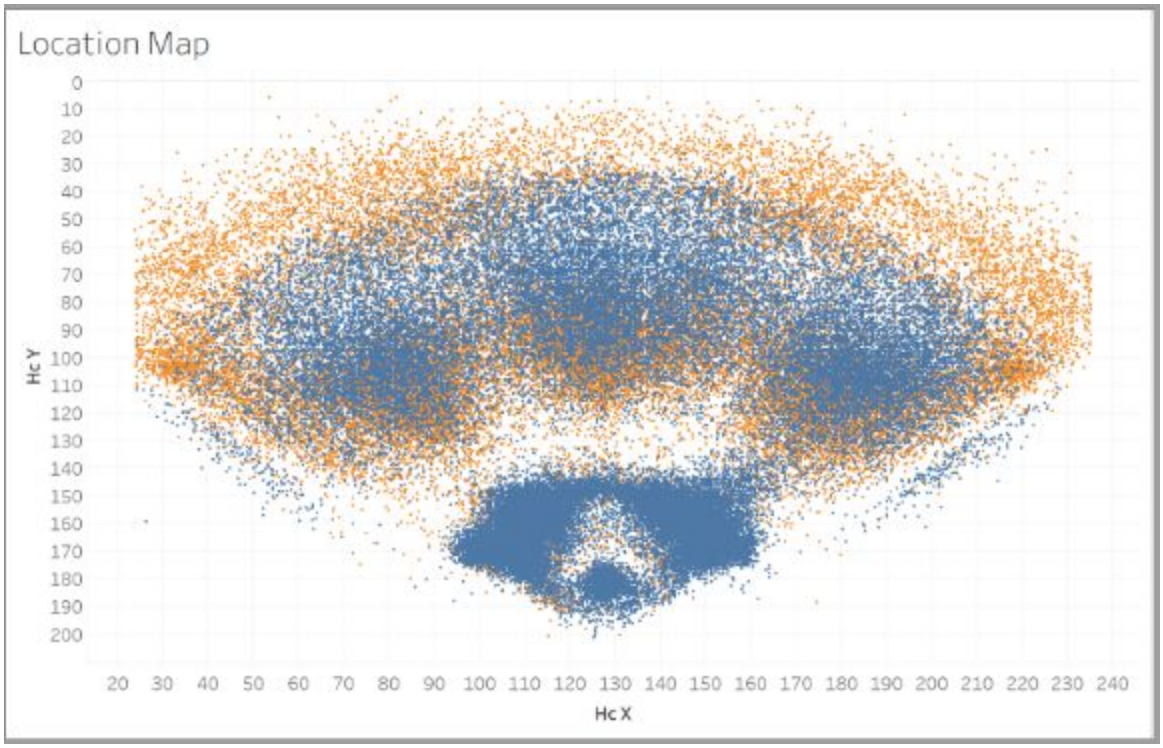
Sarris, Eno. "Are We at the High-Water Mark for Shifting in Baseball?" *FanGraphs*. February 27, 2017. http://www.fangraphs.com/blogs/are-we-at-the-high-water-mark-for-shifting-in-baseball/

Sawchik, Travis. "Has the Fly-Ball Revolution Begun?" *FanGraphs*. March 2, 2017. http://www.fangraphs.com/blogs/has-the-fly-ball-revolution-begun/

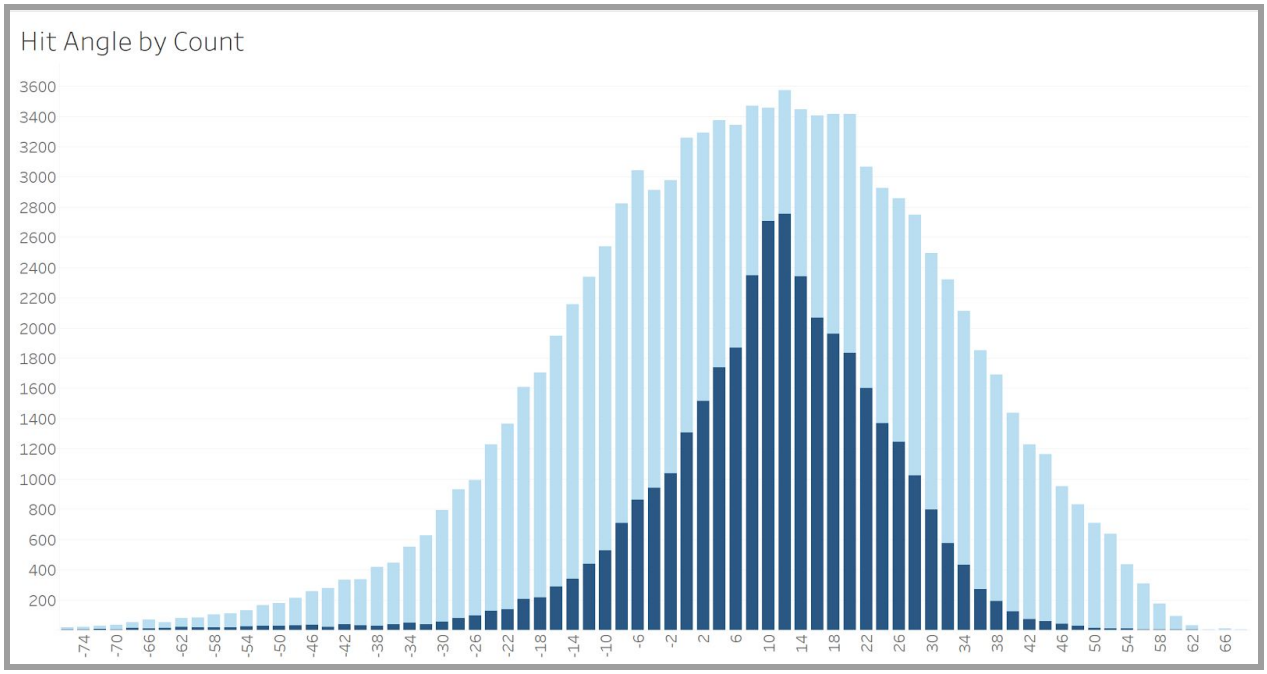
Appendices: Data Visualization



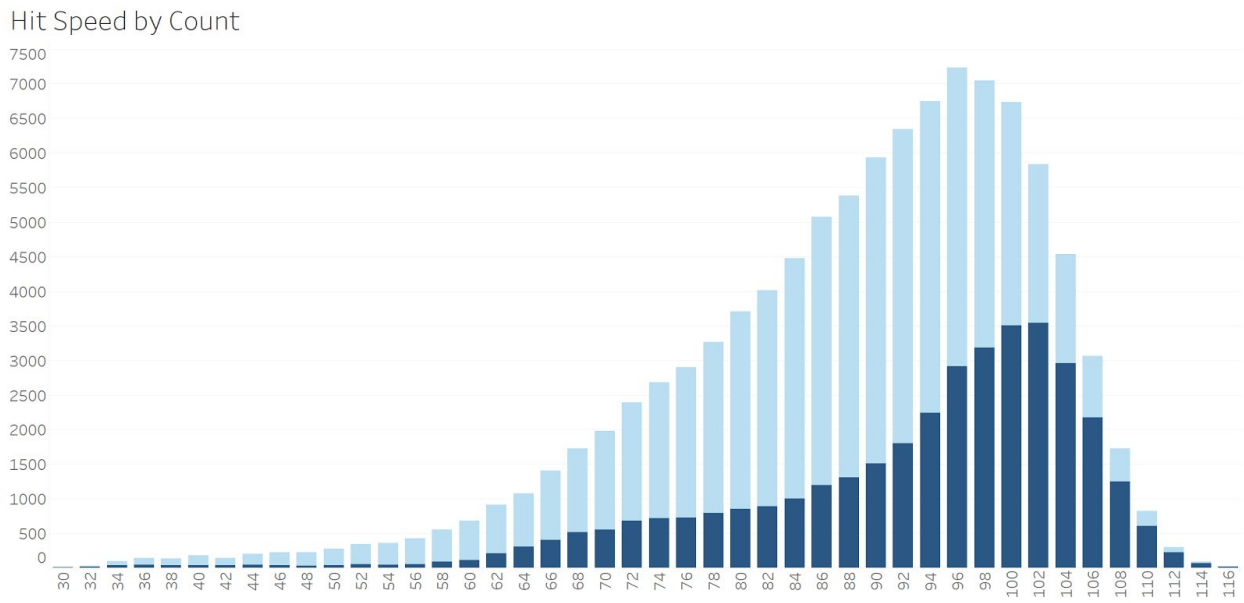
Appendix Figure 1: Batted balls plotted by exit velocity (x-axis) and launch angle (y-axis). Orange dots represent hits, and blue dots represent outs. The orange cluster at the top right includes many home runs, and the orange cluster on the left represents bloop hits.



Appendix Figure 2: Batted balls charted by the x-coordinates and y-coordinates provided by Statcast. Orange dots represent hits, and blue dots represent outs. The cluster of blue dots in short right field is evidence of the shift against pull-hitting left-handed hitters. Hits traditionally come on the outer edges of the defensive player's range, such as down the lines or in the gap.



Appendix Figure 3: Batted ball launch angle distribution. The light blue area is a count of batted balls at each two degree launch angle bin, and the dark blue area is a count of hits at each two degree launch angle bin. The larger the dark blue area is relative to the light blue area, the higher the batting average on contact for that launch angle.



Appendix Figure 4: Batted ball exit velocity distribution. The light blue area is a count of batted balls at each two degree launch angle bin, and the dark blue area is a count of hits at each two degree launch angle bin. The larger the dark blue area is relative to the light blue area, the higher the batting average on contact for that launch angle.