# Analysis of 2021 Major League Baseball Statcast Data

Austin Chen, Ambar Mishra, Koby Lieu, Su-Ting Tan, Sai Sriya Mudigonda

## Introduction

Major League Baseball (MLB) has always been an extraordinarily competitive business. Ever since the release of *Moneyball* by Michael Lewis, the fields of statistics and analytics have continuously become intertwined and synonymous with the baseball operations executives. Year after year, teams strive to find the most reliable ways to find and develop star talent.

In the past several years, there has been an unprecedented focus on the home run ball, the best way to score runs in baseball. In 2019, all 30 teams hit the most home runs in Major League history, which shattered the previous record (set in 2017) by 671.[1] The best hitters in the game generally hit lots of home runs while the best pitchers are able to limit home runs. Alongside this development has arisen the Statcast technology from MLB, which captures and publishes all the batted ball data from every Major League game.

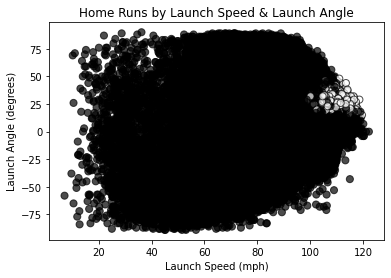
While there were not so many home run records shattered in 2021, avid baseball fans did observe similar trends in power numbers around the league as well. In this investigation, we will analyze the MLB Statcast data from the 2021 season[2] to determine the factors that most influence hitting a home run. Among them, we may discover the optimal conditions for home run hitting as well as the factors more in the control of the hitter. From these, we seek to find the changes hitters can make to optimize their swings for the ever-changing offensive baseball environment.

## 

## Exploratory Data Analysis

The Statcast data from all games in 2021 is on MLB’s Baseball Savant website, and we utilized a web scraping package in Python to extract the data directly to our program.[3]

As we investigate Major League Baseball batted ball data for the 2021 season, we are most interested in the elements of hitting home runs. We will start with looking at the several factors shown on most MLB broadcasts after a home run is hit: the pitch type, the pitch velocity, the pitch location, the launch angle, the launch speed (or exit velocity), and the projected distance.

First, we will plot the launch angle and launch speed of every batted ball in 2021 to see if there is an obvious correlation to start.

|  |
| --- |
| *In this scatterplot, a dark gradient indicates a lower proportion of batted balls are hit for home runs with those launch angle and launch speed while a lighter gradient indicates a higher proportion of batted balls are hit for home runs with those launch angle and launch speed.* |

We see right away that plotting all of the batted ball data is unhelpful, as only a small portion of the batted balls are going for any proportion of home runs. In order to get a better idea of how these factors correlate with home run hitting, we will first calculate the mean proportion of home runs hit (taking the total number of home runs hit and dividing by the number of batted ball events) and then filter only for launch angle and launch speed combinations where the proportion of home runs is greater than this mean proportion of home runs hit.

|  |
| --- |
| *In this scatterplot, a dark gradient indicates a lower proportion of batted balls are hit for home runs with those launch angle and launch speed while a lighter gradient indicates a higher proportion of batted balls are hit for home runs with those launch angle and launch speed.* |

In addition to launch angle and launch speed, we are interested in which types of pitches are hit for home runs compared to others. A pitcher seeking to limit home runs may try to throw pitches that are not hit for as many home runs, while hitters seeking to hit more home runs may seek out the pitches that are hit for a lot of home runs. Here are the home runs hit in 2021 separated by pitch type:

|  |
| --- |
| *In this bar plot, we observe the home runs hit by pitch type in 2021.* |

We see that the four-seam fastball (FF) is hit for the most home runs in 2021. While this information seems straightforward, we must recognize that the four-seam fastball is also the most common pitch in baseball, as illustrated below:

|  |
| --- |
| *In this bar plot, we observe the distribution of pitches thrown in 2021 by pitch type.* |

We thus may seek the proportions of each pitch hit for a home run.

|  |
| --- |
| *In this bar plot, we observe the proportion of each pitch hit for a home run in 2021. Note the category FA is a category of fastball that is neither classified as a four-seam fastball (FF) or a sinker (SI).* |

From this plot, we see that a much higher proportion of knuckle-curve balls (KN), unclassified fastballs (FA), and eephus pitches (EP) that are hit for home runs. However, it is worth noting that these three pitches are extraordinarily rare (see Pitch Distribution chart) - hence, the original finding that four-seam fastballs are hit for the most home runs turns out to be correct in this case.

Next, we investigate home runs by pitch velocity. This will help us to verify the findings of home runs by pitch type - traditionally, a fastball (four-seam or sinker) will be about 90 miles per hour or higher. Other possible pitches in the low 90 mph range include splitters and cutters. In the upper 80 mph speeds, we may see some cutters, splitters, and even sliders/changeups for those with much higher velocity fastballs. In the mid-80s and low-80s, we will see more sliders, curveballs, and changeups.

|  |
| --- |
| *In this bar plot, we observe the number of home runs hit off each pitch velocity (mph).* |

This data does support the premise that fastballs are the easiest pitches to hit for home runs. Worth noting is the tail off after 98 miles per hour - there are pitchers that throw higher than 98 miles per hour (upwards of 102 mph or 103 mph), so we notice that these pitches are much harder to hit for home runs than lower velocity pitches. Also, we see that the distribution is skewed to the left - with knowledge about the types of pitches, we can easily see that the movement on breaking pitches and offspeed pitches (curveballs, changeups, and sliders) makes them more difficult to hit for home runs than fastballs. All in all, we see that the lesser movement of fastballs compared to non-fastballs makes them the best pitches to hit home runs on. The trend that pure velocity means more difficult to hit for a home run only holds for the highest velocity fastballs (98 miles per hour or higher).

Next, we will focus specifically on balls hit for home runs, filtering out the batted balls that do not get hit for home runs. Below, we investigate home run balls separated by their launch angle and hit distance.

|  |
| --- |
| *In this scatterplot, we observe the home runs hit by launch angle (degrees) and hit distance (feet).* |

We observe that batted balls start to be hit for home runs when they reach roughly 340-350 feet in hit distance and hit at a launch angle of 17-18 degrees. We observe several points below the main cluster as well as to the left of it - these may be batted balls hit for inside-the-park home runs or batted balls hit for home runs at stadiums with much smaller dimensions (for example, Yankee Stadium’s “short porch” in right field or Fenway Park’s Pesky’s Pole). Again, we would classify most balls hit close to 25 degrees as line drives and most balls hit close to 50 degrees as fly balls.

Lastly, we will investigate the effect of the pitch zone on home runs.

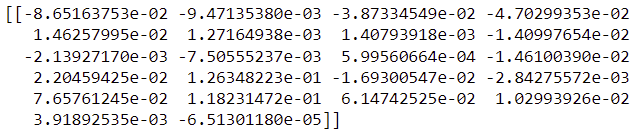
|  |
| --- |
| *In this scatterplot, we observe the home runs hit by pitch zone (the strike zone).* |
|  |
| *The above graphic depicts the numbered pitch zones from the catcher point of view.* |

We observe that the most home runs are hit on pitches of middle height (Zones 4, 5, 6). Additionally, for each height, pitches in the middle of the zone relative to home plate are hit for the most home runs (Zones 2, 5, 8). As such it stands to reason Zone 5 has the most home runs hit as pitches are “down the middle” when thrown there - pitches much easier to tell are strikes and pitches that are generally easier to hit. It follows that most hitters do not swing out of the strike zone as much as they swing in the strike zone, so therein lies the reason for decreased home run counts in Zones 11, 12, and 13.

## 

## Model Building

In addition to data analysis exploration, we wanted to build the best model to see which factors most influence hitting a home run. We included relevant quantitative variables for our model. We also split the data into 70% training data and 30% testing data when we build these models. The variables that we included are logistic model are: 'release\_speed', 'release\_pos\_x', 'release\_pos\_z', 'zone', 'balls', 'strikes', 'pfx\_x', 'pfx\_z', 'plate\_x', 'plate\_z', 'outs\_when\_up', 'inning', 'vx0', 'vy0', 'vz0', 'ax', 'ay' ,'az', 'hit\_distance\_sc', 'launch\_speed', 'launch\_angle', 'release\_spin\_rate'. The model coefficients are shown below.



Clearly, the largest coefficient in magnitude is ‘vy0’ and it is also positive, indicating that the velocity of the pitch, in feet per second, in y-dimension, will most likely influence hitting a successful home run. The second largest coefficient in magnitude is ‘az’, the acceleration of the pitch, in feet per second per second, in z-dimension, which means that the acceleration of the ball travels is also a big factor to see if we will get a home run. This matches the finding in our exploratory data analysis that fastballs are the pitches hit for the most home runs - the velocity of the pitch is largely correlated to the type of pitch, and with a positive coefficient value, we find that increased velocity correlates to a higher likelihood of being hit for a home run, with all other variables held equal. Similarly, our finding that the acceleration in the z-direction (which is upward and downward perpendicular to the ground) is the second largest coefficient in magnitude is not a surprise, as the aforementioned pitch types will have different accelerations upward and downward relative to the plate.

In addition to creating a logistic regression model, we included confusion matrices as well. Confusion matrices show how many true positive (predicted and true home runs), true negative (predicted and true non-home runs), false positive (predicted home run, true non-home run), and false negative (predicted non-home run, true home run) classifications we have. From this information, we compute the success rate from it. In a confusion matrix, the first value of the first row is true positive values; the second value is false positive; the first value of the second row is false negative; and the second value of the second row is true negative.



In the first confusion matrix, we created this using a logistic regression model. Using the logistic regression model, we have a misclassification error rate of 0.018380663541953866. The misclassification error tells us the percentage of times that a classifier is incorrect. We see that we have a pretty low percentage when using the logistic model; the misclassification rate is 0.0183, or 1.83% which means the model only predicts the home run or not incorrectly 1.83% of the time.



When we use the full logistic Generalized Additive Model, we see that the amount of true positive and true negative values has increased, while the false positive and false negative have decreased. The misclassification error is now 0.015807370646080322. This is a lower error rate as compared to what we found for the logistic regression model. This makes sense since we see that the true positive and true negative values have increased, which means that this model is more accurate than the logistic regression model that we created above. It is only a slight difference though, as it was 1.83% when we used the logistic model and 1.58% when we used the full logistic GAM model.



We also used an artificial neural network - specifically a multi-layer perceptron classifier to predict values and also created a confusion matrix with it. When we use a MLP classifier for the dataset, we see that the amount of true positive values has decreased, while the false positive, false negative and true negative values have increased, as compared to both the logistic regression and full logistic GAM model. The error rate for this confusion matrix is 0.01692859112213951. This is slightly higher than the full logistic GAM model, but still lower than the rate we get when we use the logistic regression model.

## Conclusion

As we conducted preliminary analysis on the MLB Statcast data set, we focused on the pitch type, the pitch velocity, the pitch location, the launch angle, the launch speed, and the projected distance. From the exploratory data analysis sections, we found (and later confirmed during logistic regression) which of these factors held the most importance in hitting a home run.

Looking at pitch type and velocity, we observe that fastballs are the easiest pitch to hit home runs, but usually below 98 miles since above this speed there is a tail off. Due to a left skewed distribution we see that the movement on breaking pitches and offspeed pitches makes it harder to hit a home run with than fastballs. We also see that the four-seam fastball (FF) is hit for the most home runs in 2021, and the only other types of pitches with higher proportion of home runs are extraordinarily rare. Therefore, we confirm the lesser movement of fastballs makes them the best to pitch home runs.

Observing launch angle and launch speed, we confirm that batted balls start to be hit for home runs when they are hit 340 to 350 feet or greater hit distance and a launch angle of 17 to 18 degrees up to 45 degrees. Most of the home runs are hit when the pitch is in the strike zone at middle height (Zones 4, 5, 6 of the strike zone). For each height, pitches in the middle of the zone relative to home plate are hit for the most home runs (Zones 2, 5, 8). We found Zone 5 has the most home runs hit as pitches thrown there are both middle in height and position relative to home plate.

In our model selection, we wanted to build the best model to predict whether a batted ball would be a home run or not based on the information provided in the Statcast data set. We see with the logistic model that we have a low percentage of the misclassification rate at 1.83%. Using the full logistic Generalized Additive Model, we see that the misclassification error fell to a lower rate of 1.58%, which means that this model is more accurate than the logistic regression model that we created above. With the neural network, specifically a Multi-layer Perceptron classifier, we see that, as compared to both the logistic regression and full logistic GAM model, the error rate for this model is 1.69%, which is slightly higher than the full logistic GAM model, but still lower than the rate we get when we use the logistic regression model. The logistic GAM model gives us the lowest misclassification rate and was the best model to use for home run prediction among the three options we explored.

Overall, our findings confirm that a lot of factors contribute to whether a pitch is hit for a home run. Some of which are determined by the pitcher’s pitch selection, velocity (which varies pitcher to pitcher), and the pitcher’s execution of the pitch itself, particularly the location of the pitch in the strike zone. The hitter must also execute a good swing to hit the ball hard with a high exit velocity and optimal launch angle to hit the ball far enough to be counted as a home run.

As baseball players continue to use technology to optimize their play, the importance of home runs in an increasingly offensive baseball environment will skyrocket. Especially with defensive shifts in baseball and highly researched and optimized pitching in the game, it will be ever important to score runs with the long ball when pitchers choose or execute poor pitches.

## References

[1] Stephen, Eric. “10 Reasons 2019 Was the Year of the Home Run.” *SBNation.com*, SBNation.com, 30 Sept. 2019, <https://www.sbnation.com/mlb/2019/9/30/20863062/mlb-home-run-records-2019>.

[2] “Statcast Search CSV Documentation.” *Baseballsavant.com*, Major League Baseball - Baseball Savant, <https://baseballsavant.mlb.com/csv-docs>.

[3] “Baseball-Scraper.” *PyPI*, <https://pypi.org/project/baseball-scraper/>.