# Was That Swing You or Me?

## Analyzing the Effect of Pitches and Pitcher on Batter's Swing Lengths

Kori Thompson

Baseball is often called both the easiest and hardest game to master due to its simplicity of play, high skill requirements, and deep strategic underpinnings. These strategies range from pre-game decisions like lineup and batting order to in-game situational strategies such as which order of pitches to throw for a given batter. Every at bat and every pitch thrown is calculated to maximize the benefit to the team. All aspects of the game involve a strategy and none more so than the central interactions between the pitcher and the batter. Determining what pitches to throw to prevent a batter from successfully hitting the ball is a key part of a pitching strategy. Can pitchers or their pitches do more than outmaneuver a batter? The question arises, can the pitcher dictate the batter's swing? If so, to what extent is the swing due to the properties of pitch itself or to the pitcher individually? Such an advantage over a batter would greatly improve the odds of controlling a batter at the plate and lead to a successful defense. In baseball, even the slightest edge can have a dramatic impact on the game, making any potential for an advantage interesting.

This analysis seeks to answer the question of whether a pitcher can dictate a batter's swing and to what extent the effect is due to the pitch or the individual pitcher. The focus of the analysis was strictly on the effect of the pitch and the pitcher. It did not consider any of the strategic considerations that may have also influenced a batter's swing such as the proximity of the in-fielders or a runner on base. To examine the strength of the effects of the pitch and pitcher on a swing, a regression model was used to predict the batter's swing length. Several baseline regression models were constructed including a K-nearest neighbor regression, a decision tree regression, a random forest regression, and a gradient boosted tree regression. The models were then compared and the best-performing baseline model, the gradient boosted tree, was tuned to improve performance. The permuted feature importance scores were calculated and analyzed to examine whether the pitcher or the pitch had a greater impact on the batter's swing.

The properties of the pitch were found to be more important in predicting swing lengths than the pitchers themselves. While exploratory data analysis found that there were differences in the average swing lengths against some pitchers, the pitchers individually were not important in predicting the swing length. No feature representing a pitcher had an average feature importance above 0.0016. Instead, the properties of the pitch were more important in predicting the swing length against a pitch. In particular, the pitch's horizontal and vertical locations when it crossed the plate from the catcher's perspective. Similarly, the velocity and acceleration of the pitch were found to be important in predicting the swing length. The final model was found to explain only about 41.4% of the variance in the batter's swing lengths, suggesting that there are additional factors that likely help in explaining the differences in swing lengths. The analysis may be enhanced by including features that reflect the batter's strategic considerations, such as if a runner was on base or the number of strikes and balls before the pitch. Future research areas may include examining how a batter's stance, location in the batting box, or physical features could affect the batter's swing.

# Data

The data used in this analysis was provided by Statcast for the CSAS 2025 data challenge. In total the dataset consisted of 113 features and 701,557 observations from the entire 2024 MLB spring training, regular season, and post-season. 10 of the features were denoted in the dataset documentation to be deprecated fields from the legacy tracking system and were ignored when loading the dataset. The features in the dataset consisted of both high-level descriptive statistics, such as what was going on during the pitch, the runner on base, and the score before the pitch being thrown. It also contained low-level pitch and batting statistics including the speed of the pitch, the pitch's rotation, and batting speed. As the analysis was focused solely on the effects of the pitch and the pitcher, only the 20 features that are most related to the pitcher or the pitch were considered for the analysis. The target feature for the analysis was the length of the batter's swing.

The length of a swing was measured as the total distance traveled by the head of the bat in x, y, and z-dimensions from the tracking to contact with the ball. The features chosen for the pitcher were the pitcher's ID and the arm the pitch throws with. Regarding the pitch, the data points focused on the location of the pitch when it crossed the plate, the speed, acceleration, and velocity of the pitch, the movement of the pitch, the spin or rotation of the pitch, and the type of pitch thrown. The location of the pitch was measured horizontally and vertically from the catcher's perspective when it crossed the plate, corresponding to plate_x and plate_z respectively. The speed of the pitch was measured at the release of the pitch and the effective speed was derived from the release extension of the pitch. The acceleration of the pitch and the velocity of the pitch were measured in x, y, and z-dimensions, represented by ax, ay, az, vx0, vy0, vz0 respectively. The horizontal and vertical movement of the ball from the catcher's perspective was measured in feet denoted as pfx_x and pfx_z. The pitch's breaking location or the point of descent was measured with gravity in z-dimension as well as from the horizontal point of the pitcher's arm.

## Data Cleaning

Before beginning the analysis, the data was cleaned to remove missing values or unneeded features and to convert any mistyped data. One of the first things that was done to process the data was to check for duplicate records, which were not present in the data. We next limited the dataset to the subset of features that were of interest or use for further processing. This included columns related to the pitcher, the pitch, the swing length, or descriptions of what happened during play. The columns for the description of play, while not of interest for the analysis, were left in for use in processing the data and were later dropped from the dataset.

The data was also inspected for missing values. One trend in the missingness of the data is that many features related to the pitch have the same percentage of missing values, 0.000408. This is assumed to be due to a malfunction in the equipment resulting in no measurements being recorded. There appeared to be no further connection between the missingness of these features and the other features in the dataset. It was assumed that the data was missing completely at random. The swing_length feature had about 54% of its values missing. After examining this, it appeared to be due to a combination of no swing taking place and possible equipment malfunction.

Since more than 5% of the values were missing, the decision was made to impute the values. No values were imputed for instances where no swing was made. This was done to avoid changing the distribution of the swing lengths by imputing zeros for no swings. It was assumed that no swing took place for called balls, called strikes, a pitchout, a blocked ball, or when a batter was hit by a pitch. The median swing length for a bunt is 1.6 feet while the median swing length for a regular swing is 7.2 feet. As such, the imputation for the swings where the ball was hit into play was done by category of regular swing or bunt. It was assumed that the swings for swinging strikes, fouls, foul tips, and swinging strikes block were regular.

After imputation of values for swing lengths, it was noted that there were still 52% of the swing length values missing due to no swing taking place. Since the analysis is interested in the batter's swing rather than in the batter's judgement of whether to swing, it was decided to drop these instances. This allowed the analysis to focus exclusively on swing length rather than if a batter swung or not. Instances with one or more missing values for the other features were also dropped. After dropping the missing instances and non-swing instances, the dataset consisted of 334,905 unique instances.

## Exploratory Data Analysis

A combination of descriptive statistics, measures of skew, correlation coefficients, and bivariate charts were used to explore the data. The mean, standard deviation, minimum, maximum, and quartiles were examined for the numerical features while the frequency of counts was examined for the categorical variables. Of note from this analysis, was the range of values for swing lengths. The average length of a swing for the season is 7.21 feet with the shortest swing length being 0.3 feet and the longest swing being 12 feet. Since the swing length is the total distance traveled in the x, y, and z-dimensions from the start of the swing to connection with the ball, 0.3 feet seemed an improbable number. Similarly, the median swing length of a bunt is 1.6, implying that bunts generally have some range of movement. As such swings with abnormally short swing lengths, those less than one foot, were dropped from the dataset. While the longest swing was also noted to be quite far from the average swing length, those swings with abnormally long swing lengths were considered to be accurate observations and were left in the dataset.

Another interesting trend in the data was the apparent relationship between pitchers and average swing lengths. Based on the bar chart in Figure 1, it can be seen that some pitchers have notably shorter average swing lengths against them and others that have notably longer average swing lengths against them. This implies that there could be a relationship between the individual pitcher and the batters' swings. This does not take into consideration other factors that may influence the average swing length against a pitcher such as the types of pitches they most commonly

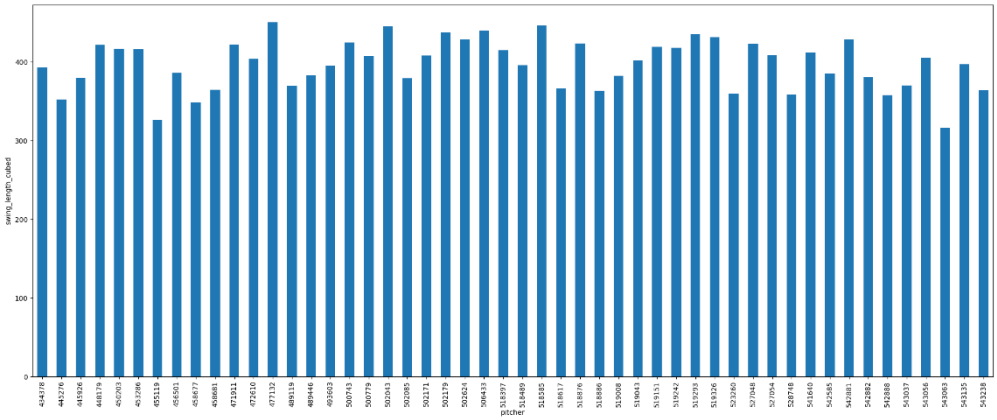


*Figure 1: Average swing length by pitcher*

throw or the average speed of their pitch. There is also an observable difference in the average swing length against certain pitches. Fastballs have a significantly shorter average swing length compared to other pitches while forkballs have a longer average swing length against them. This could imply that specific characteristics of each pitch may influence the swing length. For example, perhaps the speed of fastballs means they arrive at the plate sooner so the batter has less time and/or need to swing as long.

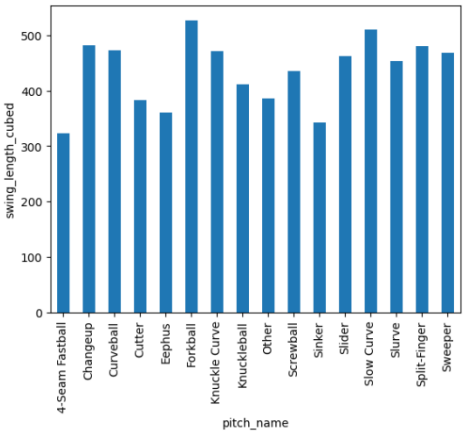



*Figure 2: Average Swing Lenght by Pitch Type*

A further trend of note is the relationship between the pitch's location and the swing lengths. The lower the swing was vertically from the catcher's perspective when it crossed the plate, the longer the swing length tended to be. As the height of the pitch increased, the swing lengths decreased. The relationship between the two, while present, is not particularly strong. There appears to be no strong linear relationship between swing length and the horizontal location of the pitch from the catcher's perspective when it crossed the plate. However, there does seem to be some relationship as shorter swings tend to happen closer to the center of the plate while shorter swings appear to occur closer to the edges of the catcher's perspective.

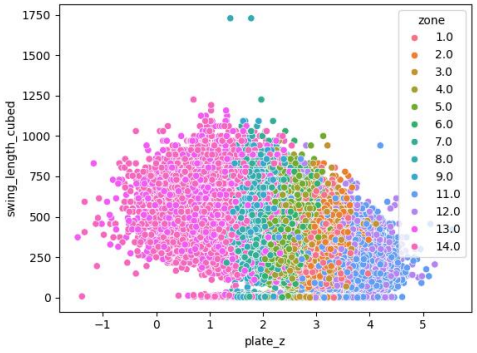



*Figure 3: Swing Length vs Vertical Position of the Pitch*

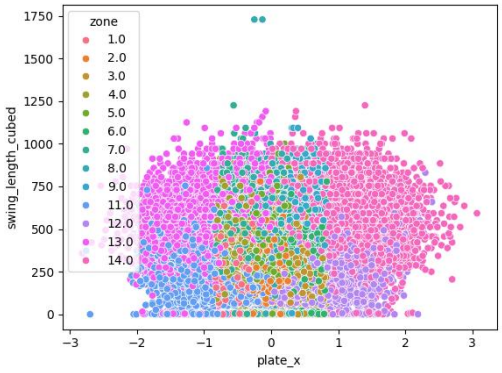



*Figure 4: Swing Length vs Horizontal Position of the Pitch*

One other trend of note is that the majority of input features did not have a linear relationship with the target feature. This can be seen in the seaborn pairplots, shown in *Appendix A*, which highlight the non-linear relationships between the data. The lack of linear relationships implies that the use of linear models would not be well suited to the data. As such, all models considered for the analysis were non-linear models.

The skew of the features was also checked to find features with highly non-normal distributions. Both swing length and release spin rate had skews below -1 implying that they were highly left skewed. A cubic transformation was used to make the swing length distribution closer to normal. For the release spin rate, a Yeo-Johnson transformation was used to correct skew. Since the data covered the entire season, this was a representative sample of the actual distributions of the data. This meant that transformations of the data would not be effected by the possibility of different distributions from the population. The decision to transform the features, particularly the target feature, was made to improve the accuracy of the modeling. Models are often sensitive to outliers and transforming the variables helps to tame some of the outliers, which improves the quality and stability of the model.

There was some multicollinearity present in the data. The release speed was highly correlated with effective speed and vy0 or velocity in the y-dimension. Similarly, vy0 was highly correlated with the

effective speed. The acceleration of the pitch in the x-dimension, ax, was highly correlated with the horizontal movement of the pitch, pfx_x. Likewise, the acceleration of the pitch in the z-dimension, az, was also highly correlated with the vertical movement of the pitch, pfx_z. The vertical movement of the pitch was also correlated with the location of the pitch broke in the z-dimension. To avoid multicollinearity obscuring the relationship of other features to the target feature, release speed, effective speed, and pfx_x, and pfx_z were dropped from the dataset.

### Feature Engineering

As previously noted, there appears to be a relationship between certain pitchers and the length of a batter's swings. There are 852 unique pitchers in the dataset and not all pitchers had a significant difference in the average swing length against them. To find the pitchers with the most significant impact on swing length, a linear regression was constructed. The linear regression used all dummy variables for the pitchers with the transformed swing length as the output. The ten most important pitchers were selected and added to the dataset as dummy variables. A pitcher's significance was determined by the absolute values of the coefficients from the regression.

## Method

### Data Preprocessing

The training and test datasets were created using a 70% training and 30% test split. The datasets were shuffled and randomly split using sklearn's train_test_split() method. Since there was a large difference in the scales of each numerical feature, they were scaled to avoid features of larger scales overpowering other features. All numerical features were scaled using sklearn's StandardScaler. The scaling of features was done after the train-test split to avoid information leakage from the test data. All categorical features were one-hot encoded using sklearn's OneHotEncoder.

### Modeling

To examine the role of the pitch and the pitcher on the batter's swing length, models were constructed to predict a batter's swing length. Four baseline models were built as candidates for the final model. These baseline models were assessed on the average root mean squared error (RMSE) and the average R-squared value from a 5-fold cross validation on the training data. The best performing model was the gradient boosted tree regression model which had an RMSE of 110.137 and an R-squared of 0.407. From there, the gradient boosted tree was tuned to improve the performance using a combination of random searches and a grid search. The tuned model achieved an RSME of 109.665 and an R-squared of 0.414 on the test data. The low R-squared value implies that there are features that could explain more of the variance in swing lengths in addition to the properties of the pitch and the pitcher.

#### Baseline Models

As previously noted, most of the input features do not appear to have a linear relationship with a batter's swing length. As such, the baseline models constructed for the analysis were limited to nonlinear models. The baseline models consisted of a K-nearest neighbor model, a regression decision tree, a

regression random forest model, and a regression gradient boosted tree (XGBoost) model. All of the baseline models used the default sklearn hyperparameter settings for each model. The baseline K-nearest neighbor model used a k of 5 and the decision tree and random forest models were allowed to grow without limits. The gradient boosted tree model used a learning rate of 0.3, a max tree depth of 6, lambda for the l2 regularization of 1, and an alpha for the l1 regularization of 0.

| Model | Average RMSE | Average $R^2$ |
|---|---|---|
| KNN Regressor | 119.80675605156709 | 0.29887912368847136 |
| Decision Tree Regressor | 159.4634424856843 | -0.24211732717962714 |
| Random Forest Regressor | 111.5255416368612 | 0.3924594458334558 |
| XGBoost Regressor | 110.13668048294919 | 0.4074942803729157 |

*Figure 5: Baseline Model Performance Metrics*

Of the baseline models, the XGBoost regressor performed the best in the 5-fold cross validation with the highest average R-squared of 0.407 and the smallest average RMSE of 110.137. The random forest model performed the second best with an average RMSE of 111.526 and an average R-squared of 0.392. The decision tree performed the worst of all the models with an RMSE of 159.463, which is well above the 143.221 standard deviation of the cubed swing lengths. This implies that the model is worse than chance at predicting swing lengths. Similarly, it also had a negative R-squared value implying that the model does a poor job of explaining the variance of swing lengths. As the best performing model from the cross-validation, the XGBoost regressor was selected as the model for hyperparameter tuning.

## Hyperparameter Tuning

To tune the XGBoost model, a combination of sklearn's RandomizedSearchCV and GridSearchCV was used. The purpose of using the RandomizedSearchCV and then the GridSearchCV was to go from more general ranges of values to very specific ranges of values. The results of each run of the RandomizedSearchCV were used to define more specific search areas for the next iteration. After the third iteration of the RandomizedSearchCV was completed, the ranges seemed to be specific enough to move to a GridSearchCV to find the exact values for the hyperparameters. The parameters tuned for the model were the number of estimators built by the model, the max depth for a tree, the max number of leaves for a tree, the learning rate, the alpha used for the l1 regularization, and lambda used for the l2 regularization. After the final round of tuning the best parameters were found to be 77 estimators, a max tree depth of 10, a maximum of 71 leaves, a learning rate of 0.1, an alpha of 1, and a lambda at 1.

## The Final Model

The final tuned model had a testing RMSE of 109.665, well below the standard deviation of the cubed swing lengths. This implies that the model does a better job than random chance at predicting swing lengths. The model had a final R-squared score of 0.414, which implies that only about 41% of variance in swing lengths was explained by the model. Using only features that relate to attributes of the pitch or who threw the pitch does a less than impressive job of predicting swing lengths. This suggests that other features were not considered in the model that may explain more of the variance in the length of a swing. These could include how far back a batter was standing in the batter's box or external factors such as if there was a runner on base or if the in-fielders were in closer. Since baseball is such a strategic

sport, it is unlikely that a model that excludes external factors that influence strategy will have much success at predicting swing length reliably.

## Analysis

To analyze whether the pitcher or the aspects of the pitcher were more important in predicting the swing length, the permuted feature importance scores were computed. The permuted feature importance was used over the feature importance from the model as it counteracts the bias towards high cardinality features of tree-based models. It also measures the decrease in RMSE over the decrease in purity. Overall, permuted feature importance is a more robust way of determining feature importance in the model and as such was the preferred method for this analysis.

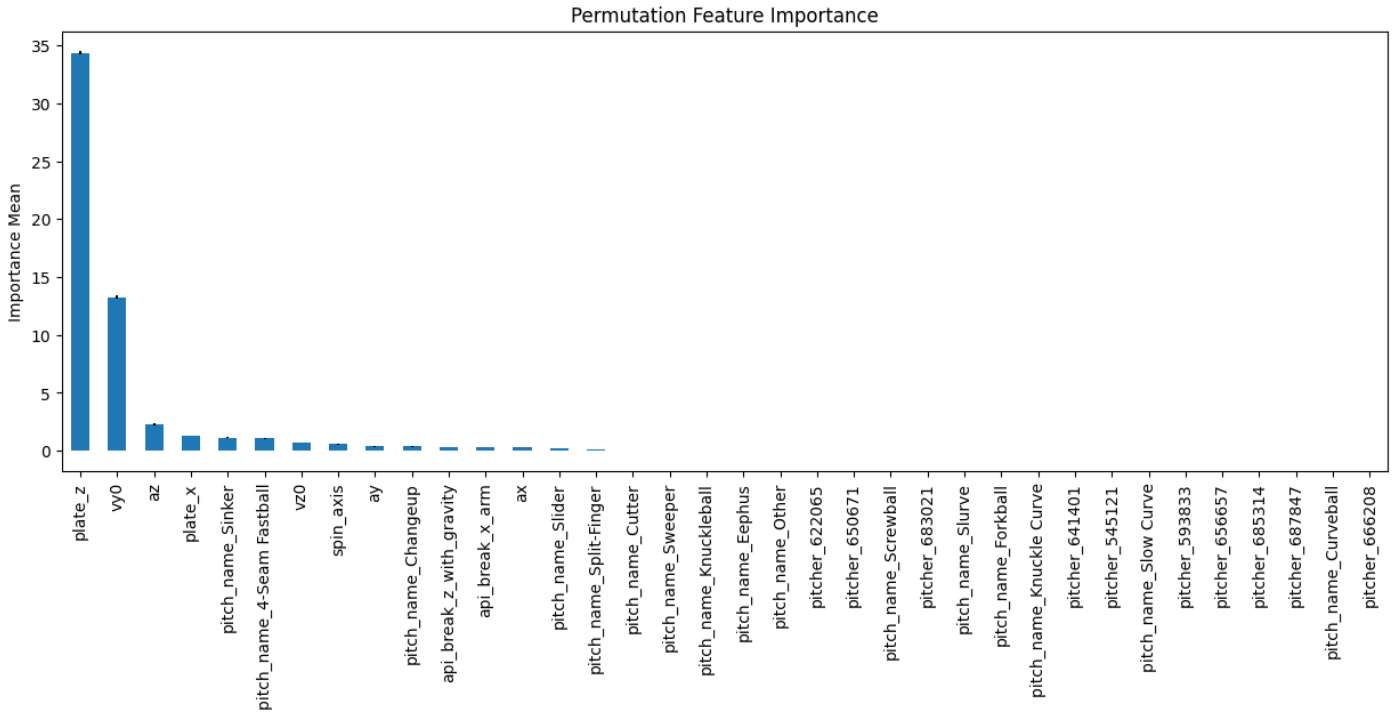



*Figure 3: Permuted Feature Importance*

The majority of features used in the model were unimportant with 21 input features having an average feature importance of less than 0.01. This included all the created pitcher dummy variables, implying that the pitchers themselves were not an important factor in predicting swing length for the model. This suggests that based on the model there is not a strong argument in favor of the pitcher dictating a batter's swing. This does not account for other factors such as the pitcher's stance, pitching form, or other factors that the pitch can control. This may be an area of future research into whether a pitcher's stance or pitching form rather than the pitcher individually can dictate a batter's swing.

Another interesting trend in the feature importance is the importance of the location of the pitch. The most important feature by far in predicting the length of a swing for the model was the plate_z feature or the vertical position of the ball when it crossed the plate from the catcher's perspective. As seen in the exploratory data analysis, there was a downward trend in the swing lengths as the vertical position of the ball increases. This makes sense as the swing length is the sum of all movement in the x, y, and z-

dimensions, and lower balls require more movement to hit than higher balls. This trend can also be seen in the fact that the sinker pitch also had high importance. Sinker pitches tend to sink low and thus influence the hit of the ball at the plate. The fourth most important feature was plate_x or the horizontal position of the ball when it crossed the plate from the catcher's perspective. This is interesting as from the exploratory data analysis it was clear that there was no strong linear relationship between plate_x and the swing length. However, it did appear that the longer swings tended to occur in the middle range of the horizontal position. The implication of this is that the location of the pitch is more influential in predicting the swing length than most other features.

A third interesting trend was the importance of the speed of a pitch. Whether a pitch was a fastball or a changeup, which effects the speed of the pitch, were both important features. Similarly, the acceleration of the pitch in all dimensions was important with the acceleration in the z-dimension being the third most important feature in the model. This implies that the speed of a pitch is important in predicting the length of a swing. With the fastball being one of the more important features, it is likely that the faster the pitch is, the greater the influence on the swing length. This could be due to the shorter amount of time the batter has to swing the bat before the pitch arrives at the plate. This is a place for further exploration in another analysis.

## Conclusions

This analysis sought to answer the question of whether a pitch dictates a batter's swing, and if so, to what extent is that the result of the pitcher or the pitch. Four candidate models were constructed to predict the length of a swing. These candidate models included a K-nearest neighbor model, a regression decision tree, a regression random forest model, and a gradient boosted tree model. The best performing model was the gradient boosted tree model, which was subsequently tuned to improve its performance. Overall, the model only achieved a test R-square score of 0.414, which was a sub-par performance. This implies that other features besides the aspects of the pitch and who threw the pitch will likely do a better job of explaining the variance in swing lengths.

Based on the model, the aspects of the pitch were found to be more important than the pitcher in predicting swing lengths. The most important features of the model tended to be those that involved the location of the pitch when it crossed the plate. The most important feature in the model was the vertical position of the ball when it crossed the plate from the catcher's perspective. The horizontal position of the pitch was also found to have high importance in predicting swing lengths. The suggestion is that the location of the pitch may influence the length of the swing. For example, lower pitches may require more vertical movement of the bat to connect with the ball, leading to longer swings. The speed of the pitch, particularly the velocity and the acceleration of the pitch, were also found to be important. This may be due to faster pitches leaving less time for the batter to swing, causing swing lengths to be shorter. The exact effects of this relationship may be a further area of research. Another interesting trend is that none of the features representing a pitcher were important for the model. This implies that the pitcher individually likely has little influence over the length of a swing. Rather, their pitch had a greater influence over the batter’s swing.

## Areas of Future Research

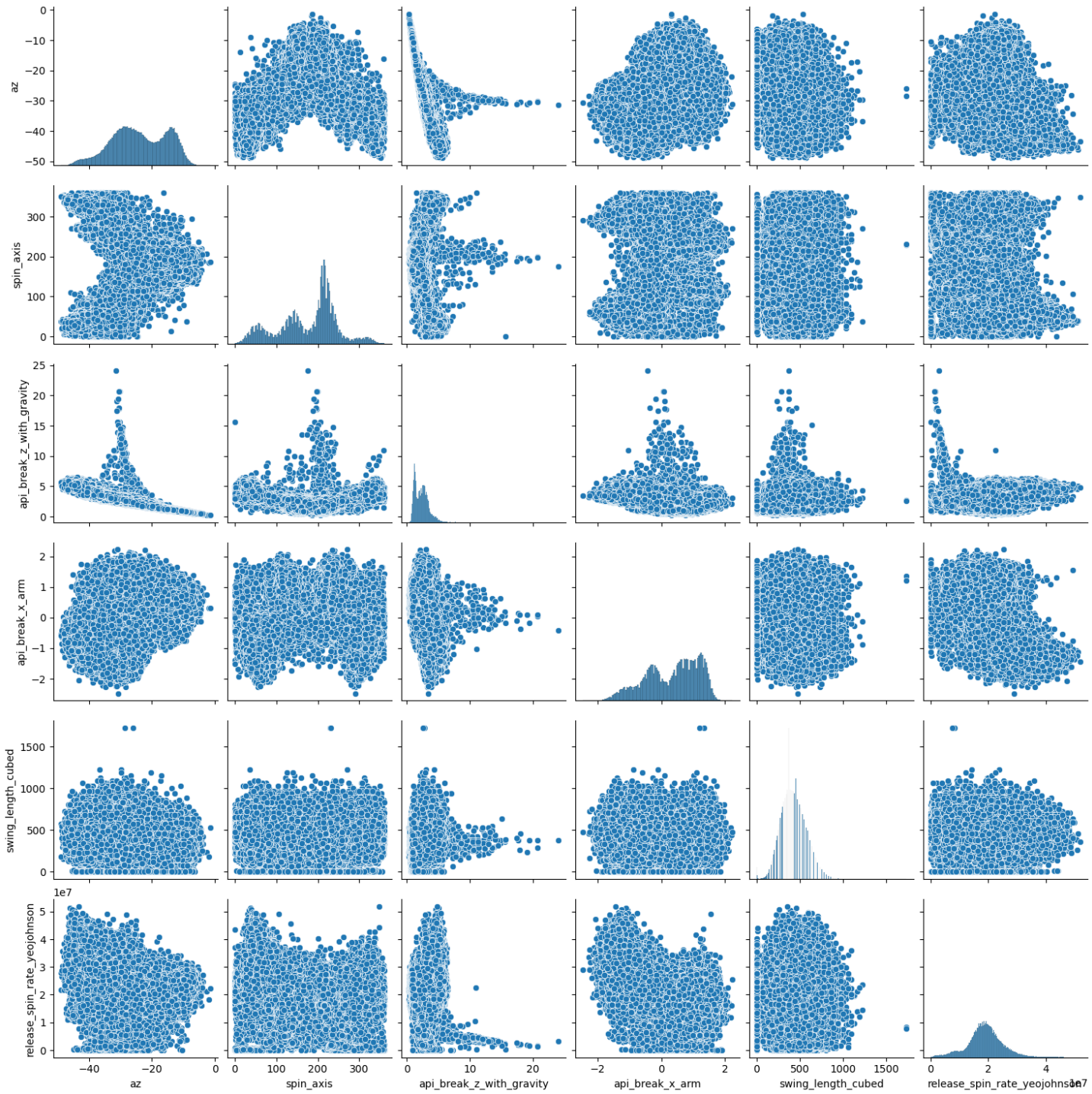
The scope of this analysis was limited to just the effects of the pitch and the pitcher on the swing length. It did not include other features that may influence the length of swing, which could provide interesting avenues for further research. This study could be enhanced by including strategic elements in the analysis to see if the pitch or the strategic considerations have more of an influence on the batter's swing. This may include data such as if a runner was on base, the score before the pitch, or even the number of strikes against the batter. Additionally, features such as the position of the in-fielders and out-fielders could be interesting to examine as well. One area for further research would be the inverse of this study, to see whether the pitch or the pitcher can influence whether or not a batter swings at the ball. Another area of future research could be how much control the batter has over their swing length. This could examine measurements such as where the batter was standing in the batting box, which side they bat, or even their physical characteristics such as height or wingspan.

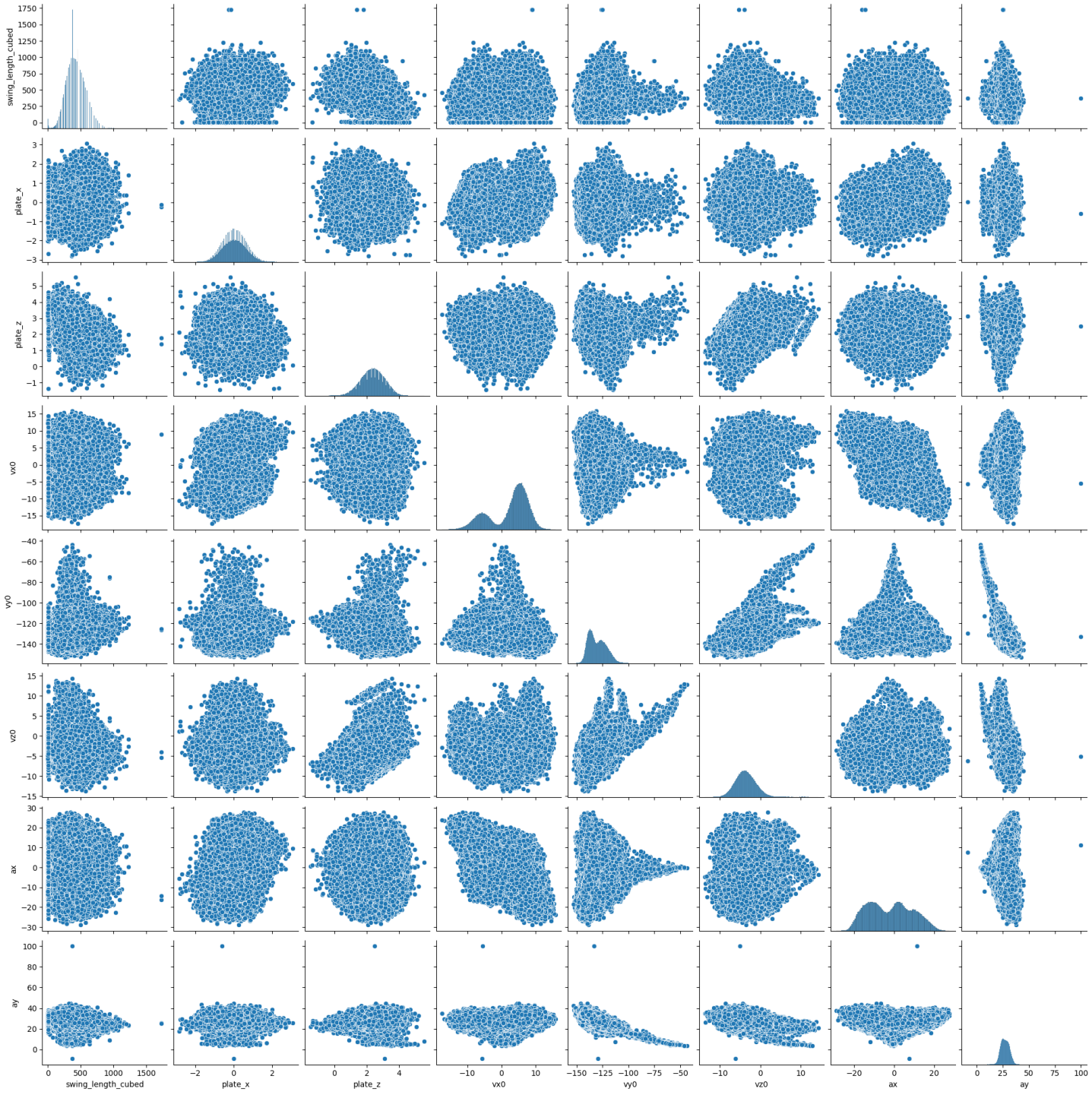
References

Baseball Savant. *Statcast Search CSV Documentation*. https://baseballsavant.mlb.com/csv-docs.

Statcast (2024). *statcast_pitch_swing_data_20240402_20241030_with_arm_angle2* [Data set]. CSAS. https://statds.org/events/csas2025/challenge.html

Appendix A


az
spin_axis
api_break_z_with_gravity
api_break_x_arm
swing_length_cubed
release_spin_rate_yeojohnson
1e7


swing_length_cubed
plate_x
plate_z
vx0
vy0
vz0
ax
ay