Faster and Longer: The Swing That Makes You Stronger

Charlotte Imbert

January 13, 2025

Introduction

The baseball swing is a thing of beauty, and has been an object of biomechanical research for decades. Coaches often spend years experimenting in search of the 'perfect' swing. A batter can't control the pitch he receives, but he can, to some extent, control elements his swing. This makes bat speed and swing length ideal targets for players to modify and optimize. But what makes a swing good? Hitting home run after home run would be a near-impossible best-case scenario, but batters can certainly use their swing to bring value to their team at every pitch in other ways.

A longer swing allows a batter to generate more momentum and kinetic energy, which should in turn send the ball further, but it also increases the chance of hitting the bat late. Swinging at high speeds seems like a good idea, but it can increase the risk of injury, and there are limits to the speed that the human body can generate. In this analysis, data from the 2024 Major League Baseball season will be analyzed with the aim of determining how bat speed and swing length affect pitch outcomes. The goal is to identify the values of bat speed and swing length that are associated with a batter's success. The interactions between these two metrics and how they impact a batter's success will also be examined. As a quick (and fun) disclaimer, my baseball experience is limited to playing Wii Sports and watching the most recent World Series, so the insights in this report will be entirely data-driven.

Data Wrangling

The data were wrangled with three objectives in mind:

1. Handling missing values,
2. Reducing dimensionality and multicollinearity by selectively removing features,
3. Creating new, potentially valuable features based on existing ones.

Each step in the data wrangling process is summarized in the flowchart below, with key visualizations used to make cleaning decisions included and described in subsequent figures.

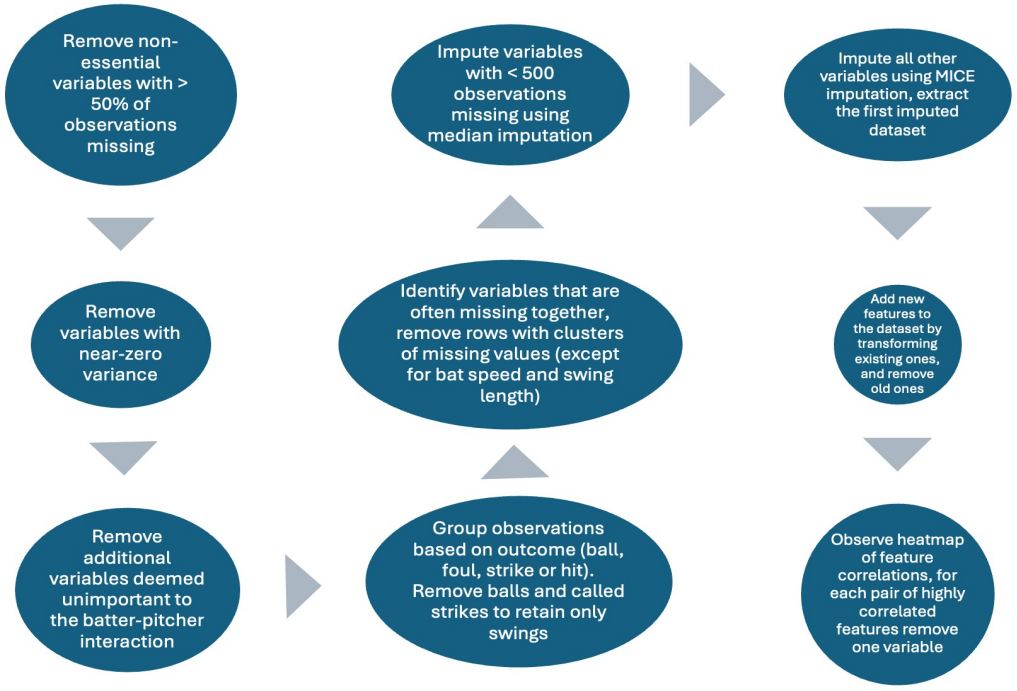


Figure 1: The data wrangling process, step-by-step.

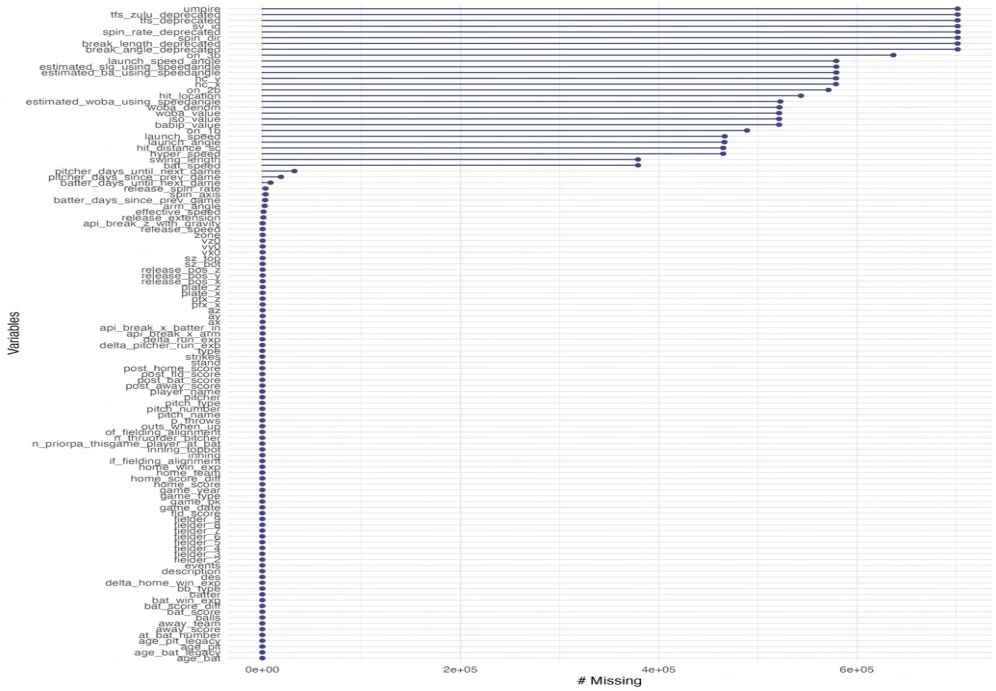


Figure 2: Number of missing observations per variable in the original dataset. Variables between launch speed and bat speed have greater than 50% of observations missing, but were kept due to their high relevance to bat speed and swing length. Patterns of missingness were observed using upset plots. Clusters of variables frequently missing together suggest that the missingness is not random for certain observations. Rows where variables were missing in clusters were consequently removed.

Feature Engineering - New Variables

New Variable	Description
swing_combined	The product of bat speed and swing length
swing_efficiency	Bat speed divided by the swing length
pitch_combined	Product of pitch release speed and release spin rate
batter_pitcher_opposite	Binary variable indicating whether or not the batter and pitcher have the same handedness
bat_team_lead	The batting team's lead pre-pitch, measured as the difference between the batting and fielding teams' scores
sz_height	The height of the strike zone, calculated as the difference between the top and bottom of the strike zone

Table 1: New variables created based on existing ones during the feature engineering stage. Simple transformations were applied to bat speed and swing length to analyze how these metrics interact. A composite pitching variable (pitch_combined) was created to quantify the interaction between the speed and spin delivered by the pitcher.

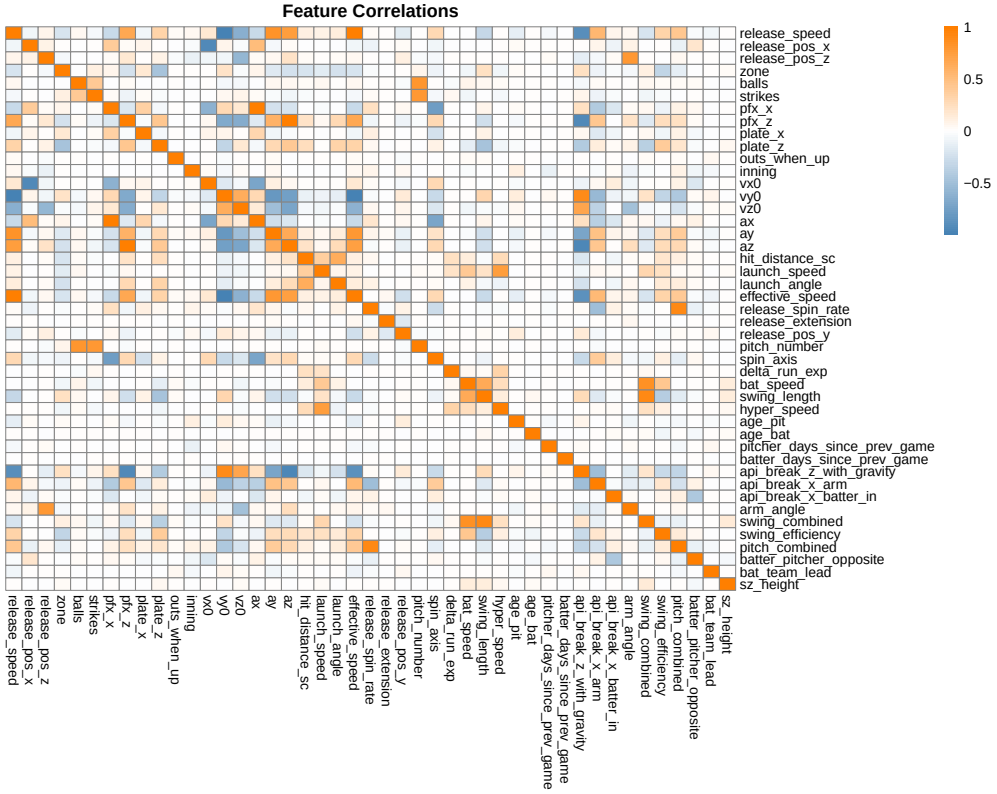


Figure 3: Heatmap of feature correlations in the wrangled dataset. Orange regions indicate positive correlations, while blue regions indicate negative correlations. With the large number of features in the dataset, feature selection and dimensionality reduction are important steps to ensure that subsequent models remain interpretable and do not overfit the data. Multicollinearity introduces noise and increases the likelihood of overfitting. To address this, one feature from each pair of highly correlated features was removed.

Data Analysis

Random Forest Classifier

A random forest classifier was trained and evaluated on the cleaned dataset. The response variable was the pitch outcome, taking on the labels hit, strike, or foul. The predictors used in the model were all of the remaining variables in the dataset except for those containing information about what happened after the ball was hit, as these would reveal information about the outcome. The random forest classifier was utilized to gauge which aspects of the swing and pitch were most important for determining whether that swing will result in a hit, strike or foul. Feature importances were calculated using Gini impurity.

The important takeaway from this model is that the composite metrics of the swing (swing efficiency and swing combined) are more important features than bat speed and swing length. This suggests that the interaction between bat speed and swing length is more predictive of the pitch outcome than either of these metrics individually.

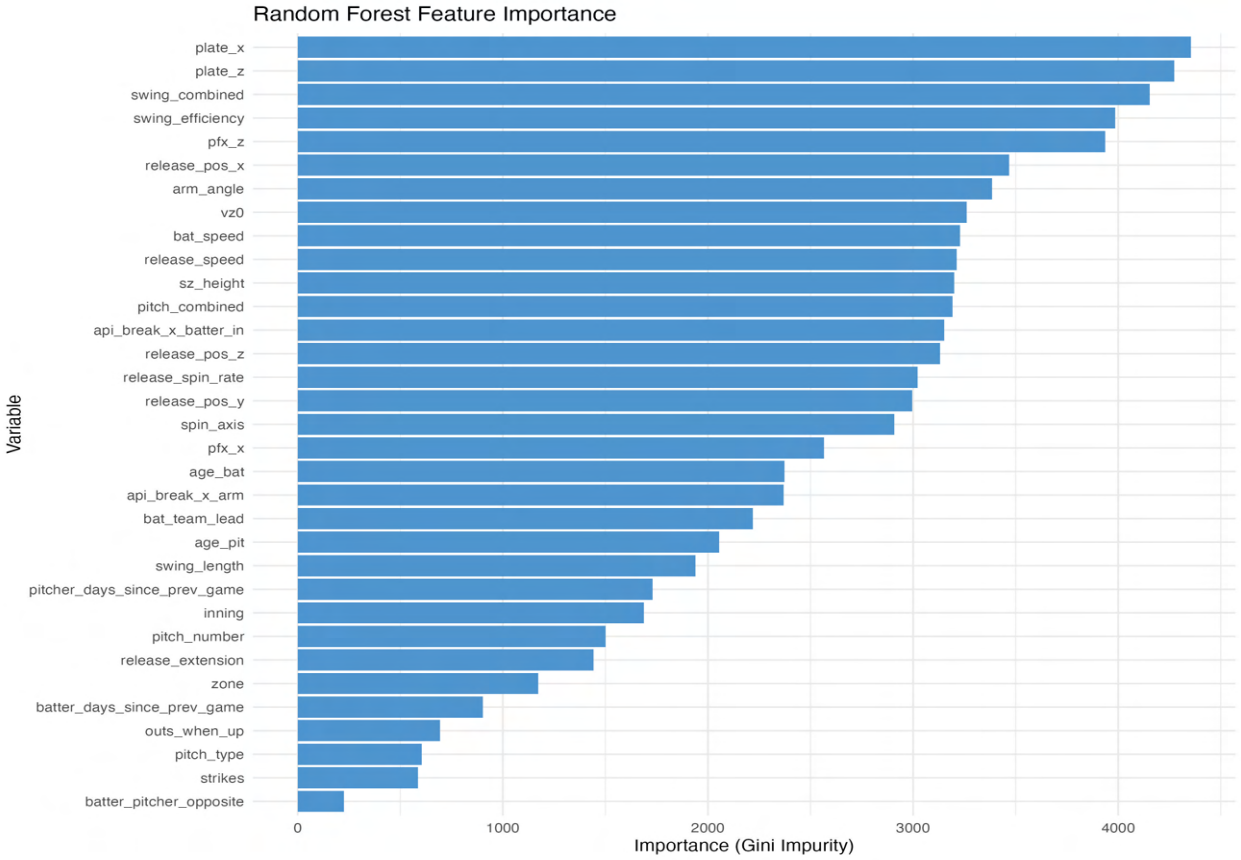


Figure 4: Feature importances for the random forest classifier predicting pitch outcomes. Model accuracy was 0.6157. The horizontal and vertical position of the ball when it crosses the home plate were the most important features, followed by the composite metrics swing combined and swing efficiency.

Batter Score Metric

While this initial model indicates which features are important for predicting the outcome of a pitch, they lack specificity about how those outcomes impact the game. We don't just want to know if a pitch was hit; we want to know if that hit was valuable! For pitches that resulted in hits into play, we can analyze the impact of bat speed, swing length and other swing and pitch metrics on the value of those hits. For this analysis, the value of a hit is defined in terms of its offensive contribution, quantified using a 'batter score' metric. Values were assigned to each pitch event to quantify its impact on the game. Positive values indicate an offensive contribution to the batting team, values close to 0 indicate neutral events, and negative values indicate detrimental events, such as those leading to outs. Values assigned for each event are displayed in the table below:

Event	**Value**
Single	1.0
Double	2.0
Triple	3.0
Home Run	4.0
Sac Fly	0.8
Sac Bunt	0.5
Field Error	0.2
Catcher Interference	0.5
Field Out	-1.0
Force Out	-1.0
Double Play	-2.0
Grounded Into Double Play	-2.0
Fielders Choice	-1.0
Fielders Choice Out	-1.0
Triple Play	-3.0
Sac Fly Double Play	-2.5

Table 2: Event values for batter contributions.

To incorporate game-specific context, the change in run expectancy before and after the pitch is considered. The composite batter score is therefore calculated as:

$$\text{Batter Score} = \text{Event Value} \times (1 + \text{Change in Run Expectancy})$$

Thus, the event is weighted by its actual effect on run expectancy during the game. As an example, singles are valued at 1 (thank you Wii Sports). Without weighting, our new scoring metric would completely ignore the single's effect in a specific game. The event value alone cannot tell us if the single resulted in a run and ignores the number of batters on the bases. This composite batter score therefore reflects specific game situations more accurately than an unweighted score. So, how do bat speed and swing length impact this new batter score?

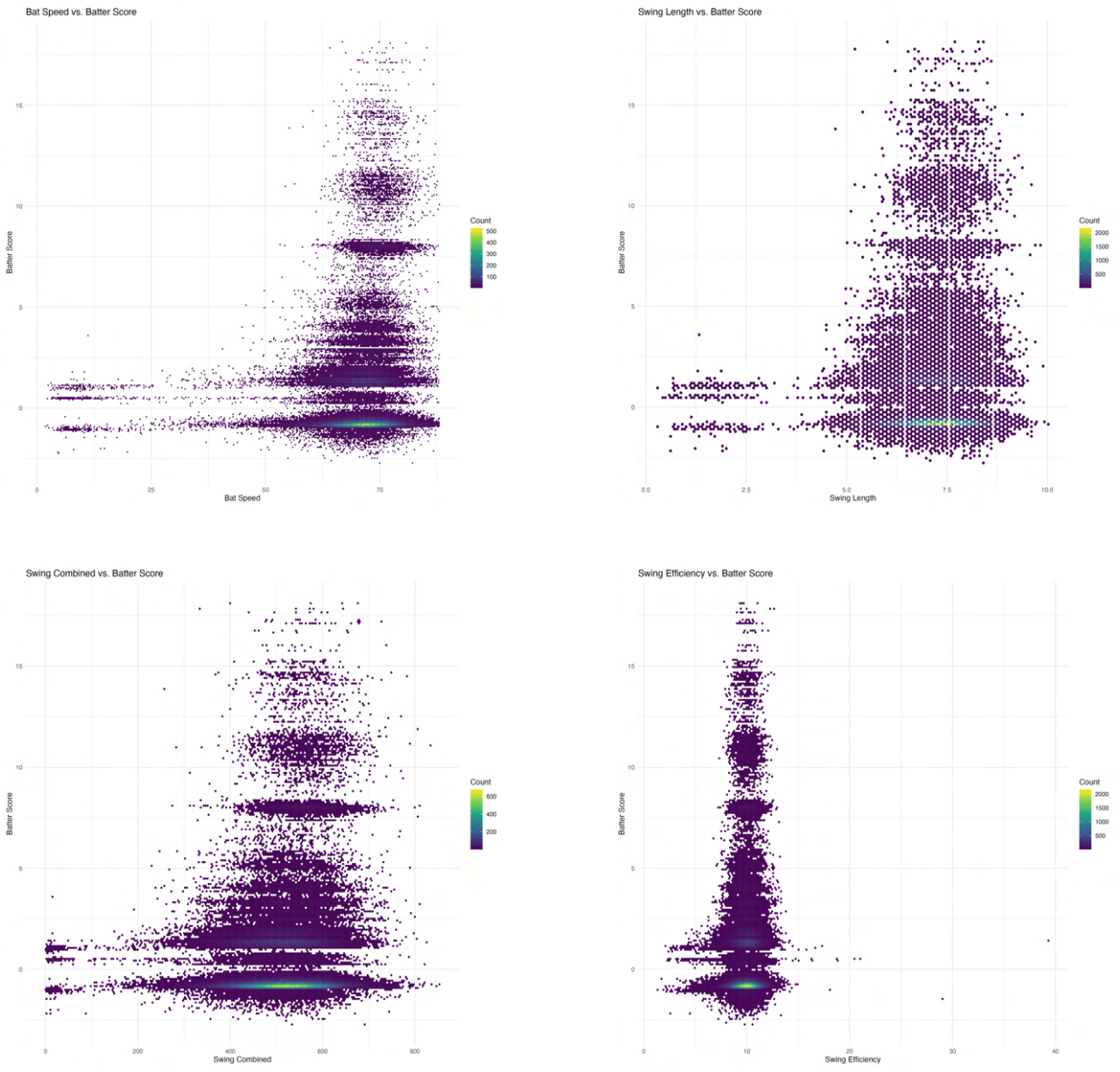
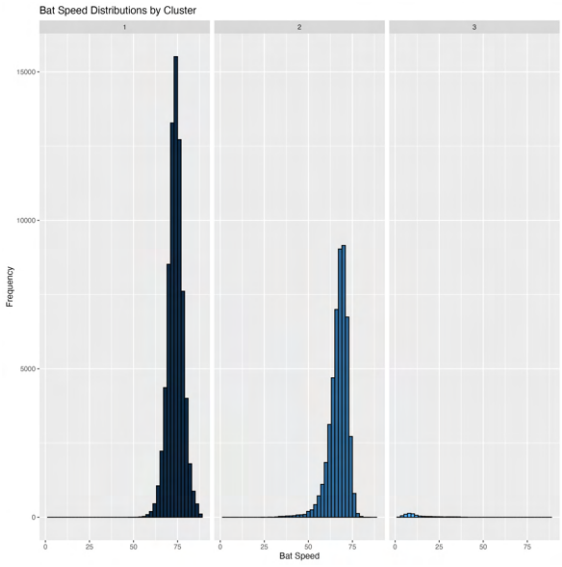


Figure 5: Distributions of batter score by bat speed, swing length, and the composite metrics swing combined and swing efficiency.

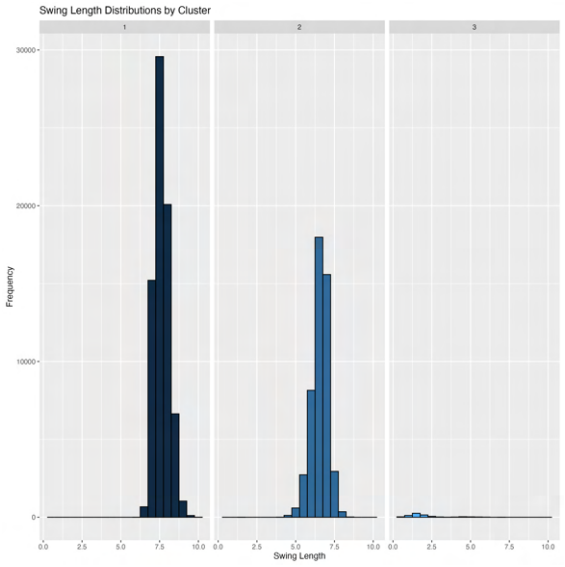
K-Means Clustering

Though none of these 4 swing metrics display a clear relationship with batter score, our random forest classifier hinted at the importance of bat speed and swing length metrics in terms of predicting outcomes. To identify more obscure relationships, K-means clustering can be used to group observations that display similar characteristics of bat speed and swing length. We can then compare the characteristics of each cluster.

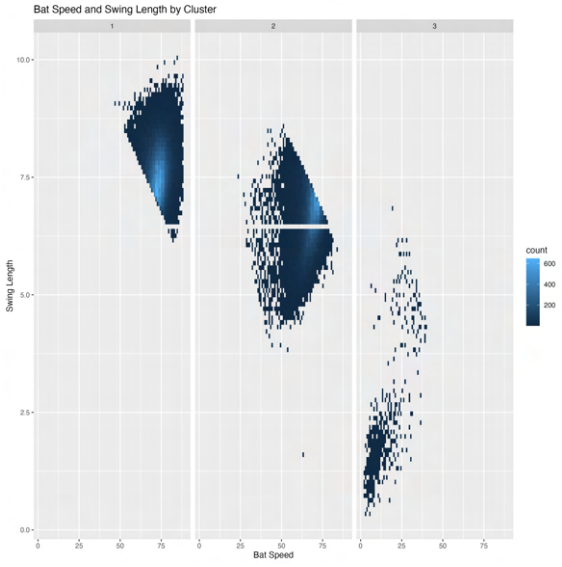
Swing Characteristics by Cluster



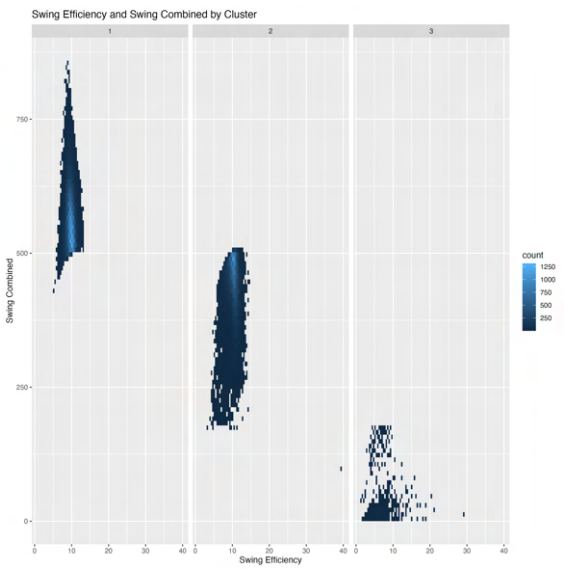
(a) Distribution of bat speed for observations in each cluster.



(b) Distribution of swing length for observations in each cluster



(c) Heatmap of bat speed and swing length for observations in each cluster, making the characteristics of each cluster more clear.



(d) Heatmap of swing efficiency and swing combined for observations in each cluster.

Figure 6: Visualizations of bat speed and swing length characteristics for each cluster.

Cluster	Count	Median Bat Speed	Median Swing Length
1	73334	73.9	7.6
2	48466	68.1	6.6
3	718	10.3	1.7

Table 3: Number of observations, median bat speed and median swing length for each cluster.

By visual inspection, several swing variables in the dataset exhibited noticeably different distributions by cluster. These were bat speed, swing length, swing combined, and swing efficiency. For each of these variables, Kruskal-Wallis and Dunn's tests of multiple comparison (using Bonferroni to control the experiment-wise error rate) were applied to confirm that there were indeed statistically significant differences between clusters for these metrics. These tests were chosen over a standard ANOVA and Tukey's HSD comparisons because they do not require the normality of residuals assumption to be satisfied.

The visualizations show the following swing characteristics for each cluster:

- **Cluster 1:**
  - Highest bat speeds (median 73.9)
  - Largest swing lengths (median 7.6)
  - Swing efficiency similar to other clusters
  - Largest swing combined
- **Cluster 2:**
  - Lower bat speeds than cluster 1, larger than cluster 3 (median 68.1)
  - Shorter swing lengths than cluster 1, longer than cluster 3 (median 6.6)
  - Swing efficiency similar to other clusters
  - Swing combined values between clusters 1 and 3
- **Cluster 3:**
  - Lowest bat speeds (median 10.3)
  - Shortest swing lengths (median 1.7)
  - Swing efficiency similar to other clusters
  - Lowest swing combined

Batter Performance by Cluster

Now that we have determined the differences in swing characteristics between clusters, we can examine how these swing characteristics impact batter performance (measured by our new batter score).

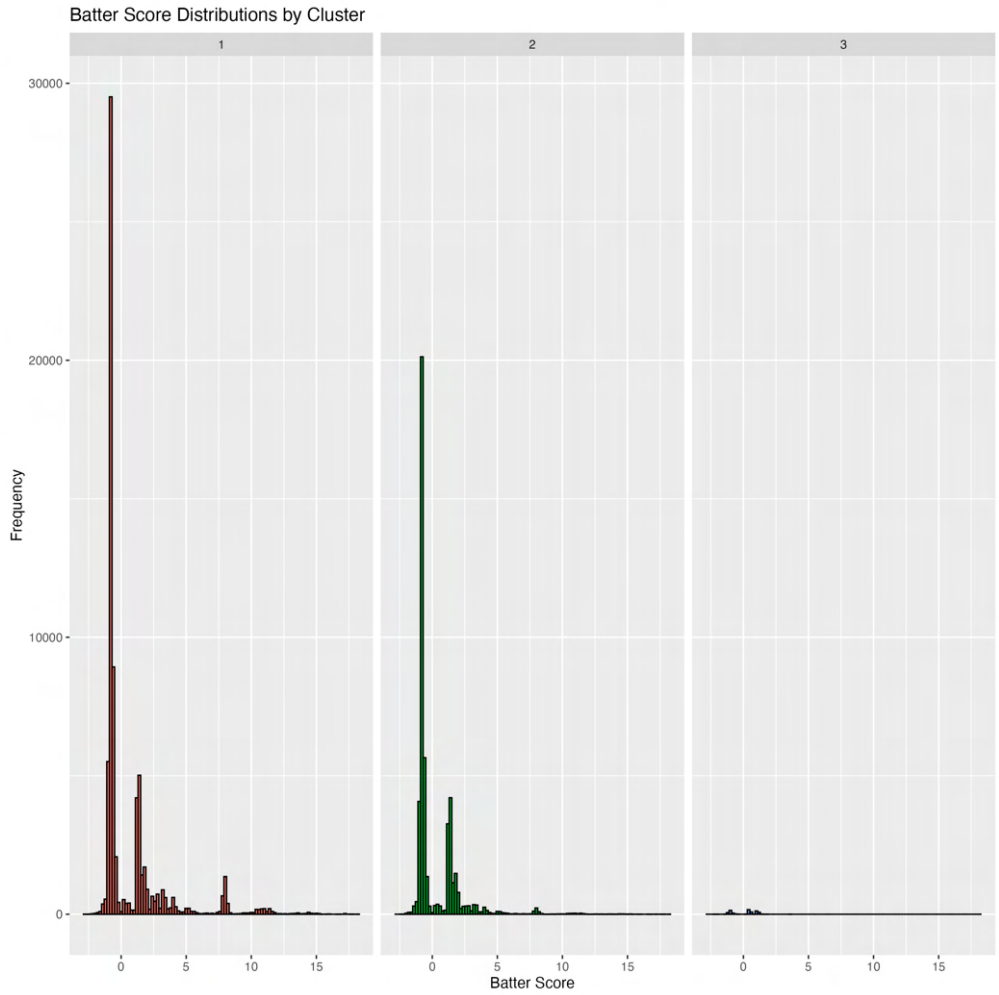


Figure 7: Distributions of batter score for each cluster, labeled at the top of the graph.

Upon visual inspection, it seems that cluster 1 has slightly higher batter scores than cluster 2, which in turn has higher batter scores than cluster 3. We can use a permutation test to determine the statistical significance of the cluster differences in batter score. Permutation tests involve:

1. Calculating a test statistic based on the observed data,
2. Creating a null distribution by resampling from the observed data, where labels (clusters) have been shuffled,
3. Calculating the same test statistic based on the shuffled data,
4. Visually comparing the null and observed distributions,
5. Returning a p-value for the differences between the distributions, and either rejecting or failing to reject the null hypothesis of no difference between the observed and shuffled data.

The chosen test statistic was the variance between cluster group means in order to gauge the variability between the mean batter score for each cluster.

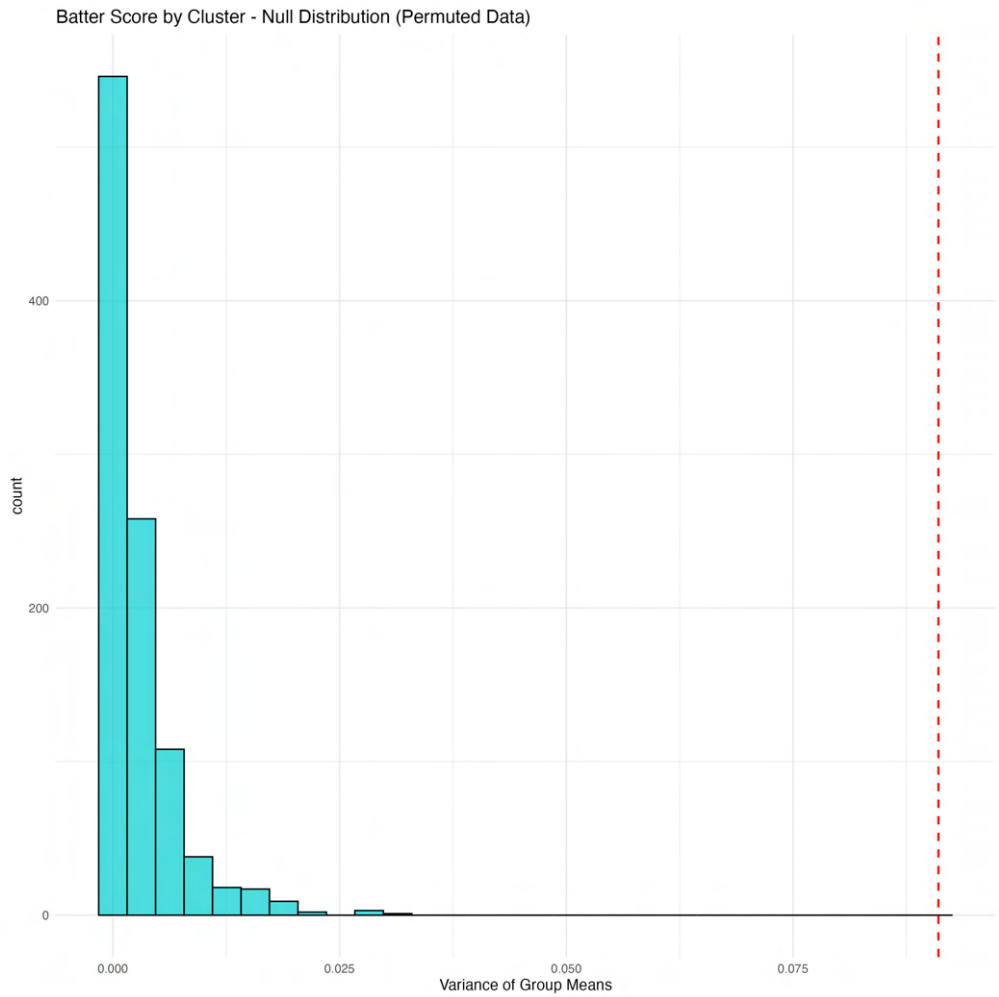


Figure 8: Results of the permutation test. The red dashed line indicates the observed value of the test statistic (variance of group means), while the histogram displays the values of the test statistic calculated using the shuffled data. The p-value was 0, allowing us to reject the null hypothesis. It is extremely unlikely, at the 5% significance level, that batter score does not vary across clusters.

Additionally, Kruskal-Wallis and Dunn's tests of multiple comparisons, followed by mean rank comparisons, were used to quantify the differences between the clusters. This revealed that cluster 1, on average, has larger values for batter score than cluster 2, which in turn has larger values than cluster 3.

Cluster Comparison	**Mean Rank Difference**
1 - 2	3475.504
1 - 3	5607.748
2 - 3	2132.244

Table 4: Batter score mean rank differences. Cluster 1 has larger batter score values, on average, than clusters 2 and 3. Cluster 2 has larger values than cluster 3.

The Effect of Strike Count

The clusters also differ in their distributions of strike counts, as can be seen in the bar chart below:

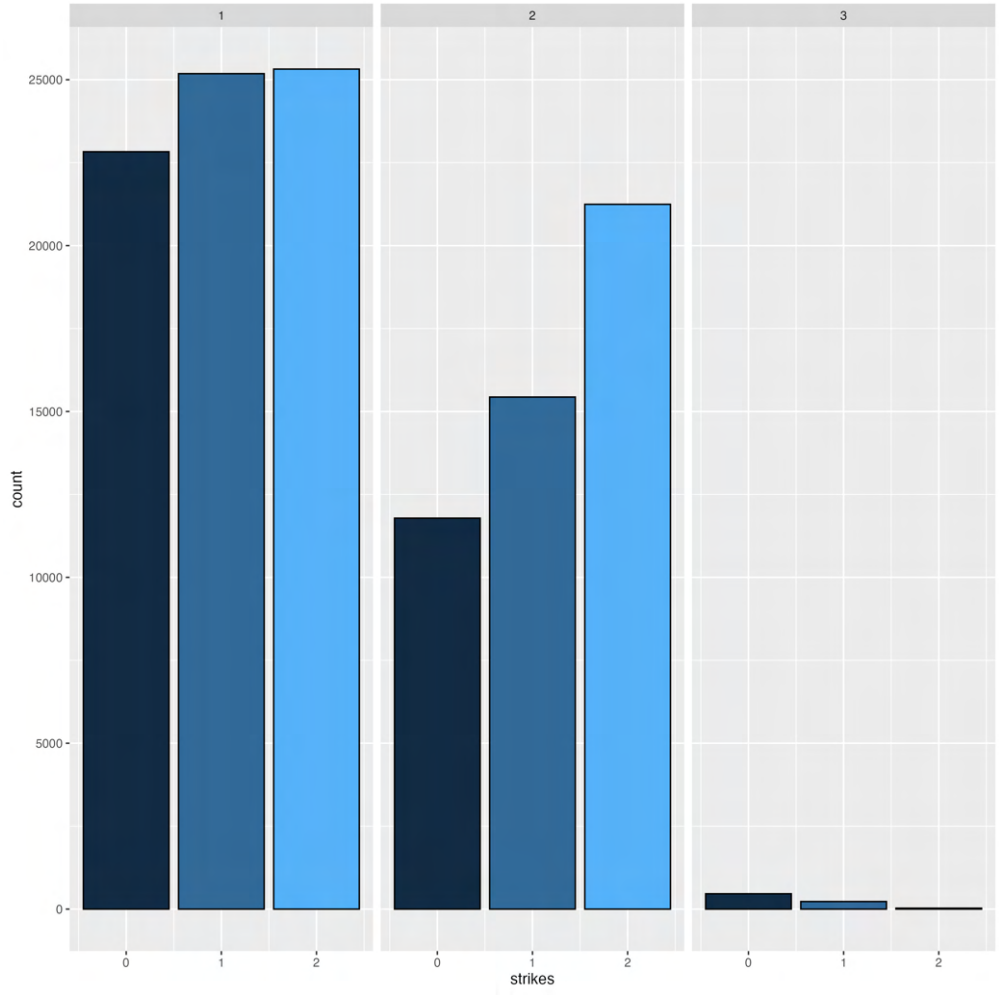


Figure 9: Distribution of pitch strike counts across clusters. Clusters are labeled at the top of the graph.

Metric	Cluster 1	Cluster 2	Cluster 3
% of all pitches in data	59.86%	39.56%	0.58%
% of 0-strike pitches in data	65.08%	33.60%	1.32%
% of 1-strike pitches in data	61.65%	37.79%	0.56%
% of 2-strike pitches in data	54.34%	45.60%	0.06%

Table 5: Strike count compositions and overall pitch proportions for each cluster. Cluster 1 has a higher proportion of observations at 0- and 1-strike counts, while cluster 2 has a greater share at 2-strike counts. Cluster 3 contains relatively few observations across all strike counts.

The observations belonging to cluster 3 have the highest proportion of 0-strike pitches, followed by 1-strike pitches and lastly 2-strike pitches. On the other hand, 2-strike pitches occupy the

highest proportion of observations in clusters 1 and 2. Interestingly, cluster 1 has very similar proportions for both 1-strike and 2-strike pitches, while cluster 2 has a much larger proportion of 2-strike pitches. In fact, despite consisting of approximately 25,000 fewer observations than cluster 1, cluster 2 makes up 46.6% of all 2-strike pitches resulting in hits into play.

These observations have implications since we have already established that the clusters differ in terms of their offensive success. Cluster 1 observations, on average, have higher batter scores than those in cluster 2. 2-strike pitches represent high-pressure situations, since a strike here results in an out (again, thank you Wii Sports - I learned the hard way). The observed strike count proportions imply that the performance of a batter at 2-strikes experiences a slight drop: Almost half of all 2-strike pitches are in cluster 2, which has lower bat speeds, shorter swing lengths, and lower batter scores on average, relative to cluster 1.

Conversely, the extremely small proportion of 2-strike pitches and relatively high proportion of 0-strike pitches in cluster 3 suggests that the observations in cluster 3 are swings where batters were willing to make riskier decisions due to the lower stakes that accompany a 0-strike pitch, since there is a reduced threat of an imminent out.

Conclusions and Takeaways

The first finding produced by this analysis is that the horizontal and vertical position of the ball as it crosses the plate, swing combined (the product of bat speed and swing length), and swing efficiency (bat speed divided by swing length), were the most important variables for predicting whether a pitch resulted in a hit, strike or foul. This pointed to the relatively higher importance of the interactions between bat speed and swing length, compared to bat speed and swing length individually, in predicting pitch outcomes.

The focus of this analysis then shifted to pitches resulting in hits into play in order to determine how bat speed and swing length affect hit outcomes. To do this, a new metric was defined, called batter score, to provide an indication of the offensive contribution made by a batter at each pitch. This metric incorporated both the change in run expectancy and the standard value of pitch events (singles, doubles, etc.) in order to quantify the context-specific influence of a swing.

Since no clear relationship could initially be visualized between swing metrics (bat speed, swing length, swing efficiency, and swing combined) and batter score, a k-means clustering algorithm was implemented. The clustering analysis found that pitches resulting in hits into play can be grouped into 3 clusters based on bat speed and swing length. Pitches with higher bat speeds and longest swing lengths on average made up cluster 1. Cluster 2 contained pitches with slower bat speeds and shorter swing lengths on average than cluster 1, and cluster 3 was composed of anomalous pitches with the slowest bat speeds and shortest swing lengths.

These findings are important because there were significant differences in batting performance between clusters. Cluster 1 had higher batter scores on average, followed by cluster 2, and lastly, cluster 3. This indicates that both a high bat speed and a long swing length are associated with greater offensive contributions by the batter. While we might expect swing efficiency (the bat speed divided by the swing length) to be decisive in a batter's performance, our analysis does not suggest this as swing efficiency was similar across the clusters. A batter does not appear to gain

any meaningful offensive advantage from compacting a high bat speed into a shorter swing.

Importantly, the clusters displayed interesting differences in their strike count proportions. 2-strike pitches were almost exclusively found in clusters 1 and 2. Close to half of all 2-strikes pitches hit into play were in cluster 2, despite this cluster making up only approximately 40% of all hits. This disproportionately higher allocation of 2-strike pitches to cluster 2 might reflect the drop in a batter's performance during a high-pressure situation, such as a 2-strike pitch. This drop in bat speed and swing length could alternatively demonstrate a batter's desire to take fewer risks during a high-stakes situation. Cluster 2 also contains a disproportionately low percentage of 0-strike observations, indicating a batter's willingness to swing more riskily when the stakes are low (i.e. much slower and shorter like in cluster 3, or much faster and longer like in cluster 1).

Overall, this analysis provides strong evidence that higher bat speed and longer swing length lead to greater offensive contributions. It also demonstrates a batter's strategic adjustments according to strike counts, such as reducing risk by reducing bat speed and swing length during 2-strike pitches. It seems that Daft Punk, and later, Kanye West, were right: Harder, Better, Faster, Stronger!