S&DS 425/625 Report: MLB Swing Length and Bat Speed Data

Julia Stiller and Alan Phlips

12/10/2024

Abstract

This report examines the relationship between swing mechanics (bat speed and swing length), game-state variables (balls, strikes, outs, and pitch type), and offensive production metrics such as weighted On-Base Average (wOBA) using data from the Baseball Savant Statcast system. The dataset comprises more than 300,000 Major League Baseball plate appearances from April to June 2024, provided for the CSAS 2025 Data Challenge. The dataset was cleaned and explored to identify patterns and trends in swing-related metrics and contextual variables. Multiple Linear Regression models were developed to analyze how swing mechanics are influenced by game-state factors and to predict wOBA based on these variables. Our findings indicate that game-state variables significantly affect swing length and bat speed, with longer swings correlating negatively with wOBA, despite a positive relationship between bat speed and offensive productivity. These results provide actionable insights for optimizing batter performance under varying game conditions. Further research will enhance model complexity by incorporating nonlinear relationships and batter-pitcher interactions to improve predictive accuracy and practical applicability.

Section 1: Introduction

The Connecticut Sports Analytics Symposium (CSAS) 2025 Data Challenge provides pitch-level data from Baseball Savant for 346,250 Major League Baseball plate appearances from 4/2/2024 to 6/30/2024, including relevant Statcast data along with bat speed and swing length on pitches with a swing tracked. The challenge is to use new baseball data on bat speed and swing length to analyze some aspect of the pitcher/batter interaction.

This study leverages that dataset to analyze how bat speed, swing length, and contextual factors such as the count, outs, and information on the batter-pitcher combinations influence outcomes like weighted On-Base Average (wOBA) and run expectancy. By modeling these relationships, we aim to provide actionable insights to help batters determine the optimal conditions under which to swing or not swing, ultimately improving offensive decision-making.

The dataset for this analysis was obtained as part of the CSAS 2025 Data Challenge and was provided via a SharePoint link. The dataset was subsequently downloaded as a csv file and then saved to an RDS file. To prepare the data, we removed columns that deprecated from the dataset and those that were found to be all empty values. We then also removed a small fraction of the rows that had no pitch type or release speed value due to a data entry error. Following data cleaning and preprocessing, we conducted exploratory data analysis to identify trends and validate the feasibility of modeling offensive production metrics (wOBA/run expectancy) based on contextual and swing-related factors in the dataset.

The dataset contains lots of Baseball specific terms and metrics, meaning a glossary is necessary to understand the data and a significant amount of time is needed to interpret the meaning of each column before analysis can begin. For example, wOBA, short for Weighted on-base average is a metric that quantifies offensive value of a play or player based on the relative value of each type of offensive event, normalized by season to be on a scale that can be compared for more apples-to-apples comparisons.

This paper is organized as follows: **Section 2** presents an exploratory analysis of the dataset, highlighting

patterns and trends in relevant predictors. **Section 3** outlines the development of predictive models, especially regression models, with an emphasis on evaluating their accuracy and interpretability in prediction. **Section 4** discusses the implications of findings up to this point, including interpretations and potential applications for hitters and coaching staff. Finally, **Section 5** concludes with a summary of the key findings and recommendations for further work.

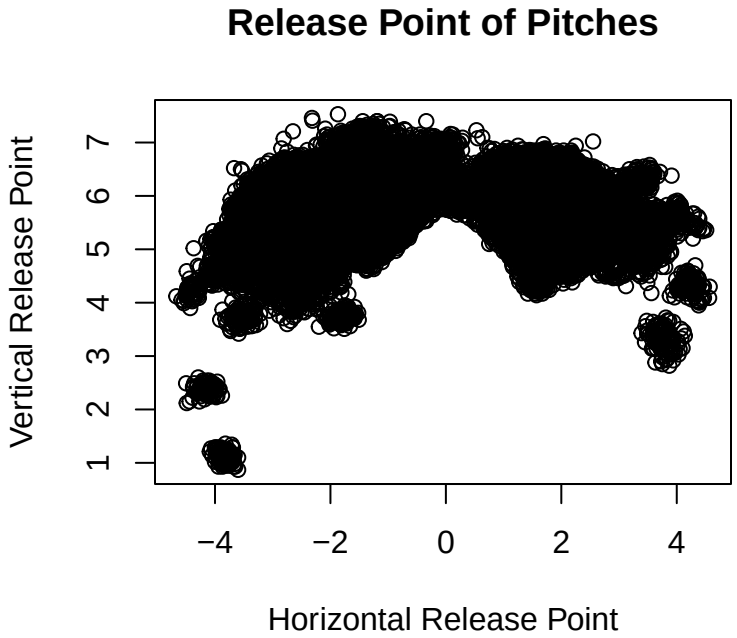
Ultimately, this project is the beginning of further research to be conducted on this dataset as part of the CSAS 2025 Data Challenge. The insights gained from this analysis will be used to inform future work and potentially contribute to the development of new strategies for optimizing offensive performance in Major League Baseball.

Section 2: Data exploration and Visualization

Given the context and rich dataset provided for this challenge, extensive data exploration and visualization was required for comprehension and to identify patterns and trends in the data and put into context. There are more than 85 unique variables for each of the more than 300,000 pitches in the cleaned dataset, each one giving unique information on the pitch, the batter, the pitcher, and the game situation. A significant part of the work for this study, and the essence of the data challenge as a whole, is to focus on a specific aspect of the game through this data to provide insights and recommendations for the batter-pitcher interaction. Going from the whole dataset to a specific question required a lot of data exploration and visualization to understand the data and guide what analysis will be completed.

The visualizations and commentary below provide a glimpse into the exploratory analysis conducted on the dataset.

As mentioned before, the dataset contains information on numerous aspects of the pitcher-batter interaction. One such aspect is the release point of each pitch as it moves towards the batter. The scatter plot below shows the release points of all pitches in the dataset. The x-axis represents the horizontal release point of the pitch, while the y-axis represents the vertical release point. This is information can be useful for batters to understand the trajectory of the pitch and make better decisions on whether to swing or not. While inherently very insightful, it was not used in this study beyond exploration due to the complexity of the data and the focus on swing length and bat speed.



Another example of useful information is the events column, which is summarized below. Each of these events represents a different outcome of the pitch not described by other quantitative columns. This information is crucial for understanding the context of the pitch and the outcome of the at-bat. This information was

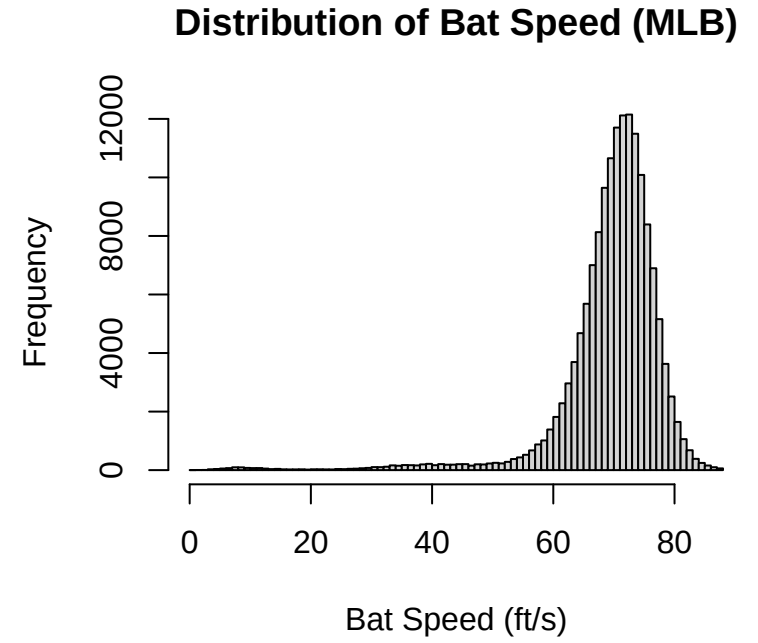
used in the study to understand the context of the pitch and the outcome of the at-bat, but was not used in the modeling process. Our analysis focuses on quantitative measures of offensive production to guide plate discipline and decision-making for batters.

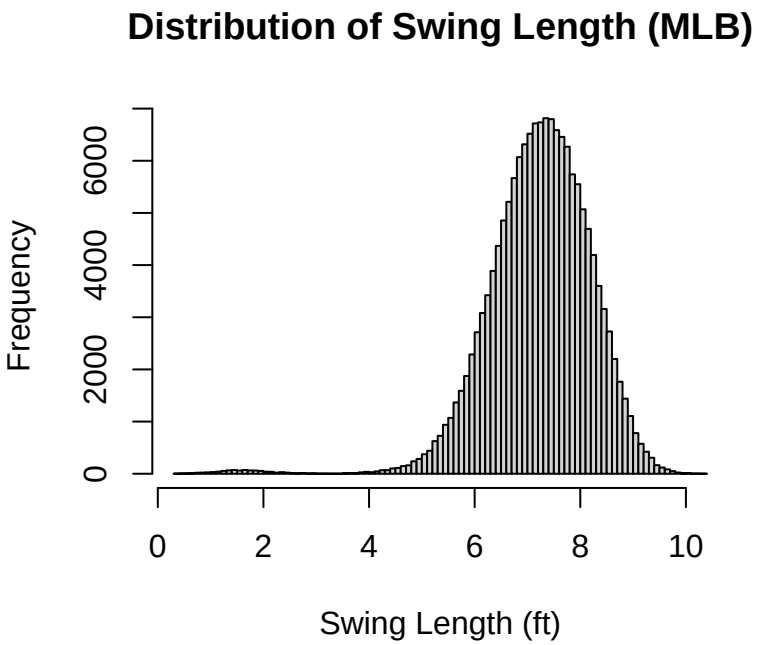
Event	Frequency
catcher_interf	55
caught_stealing_2b	131
caught_stealing_3b	6
caught_stealing_home	6
double	3812
double_play	161
field_error	570
field_out	36486
fielders_choice	176
fielders_choice_out	163
force_out	1716
grounded_into_double_play	1612
hit_by_pitch	1012
home_run	2529
other_out	9
pickoff_1b	4
pickoff_3b	1
pickoff_caught_stealing_3b	1
pickoff_caught_stealing_home	1
sac_bunt	225
sac_fly	620
sac_fly_double_play	2
single	12633
stolen_base_2b	2
strikeout	19696
strikeout_double_play	57
triple	351
triple_play	1
walk	7038

One aspect of the data that was used in the study is the number of balls and strikes present in the count as it is called before each pitch. The table below shows the number of balls and strikes in the dataset. This information is crucial for understanding the context of the pitcher-batter interaction as there are well known advantages and strategies for both sides of the interaction depending on the count.

	0	1	2	3
0	89083	33307	11063	3303
1	45413	35031	17869	7325
2	23649	33762	28727	17464

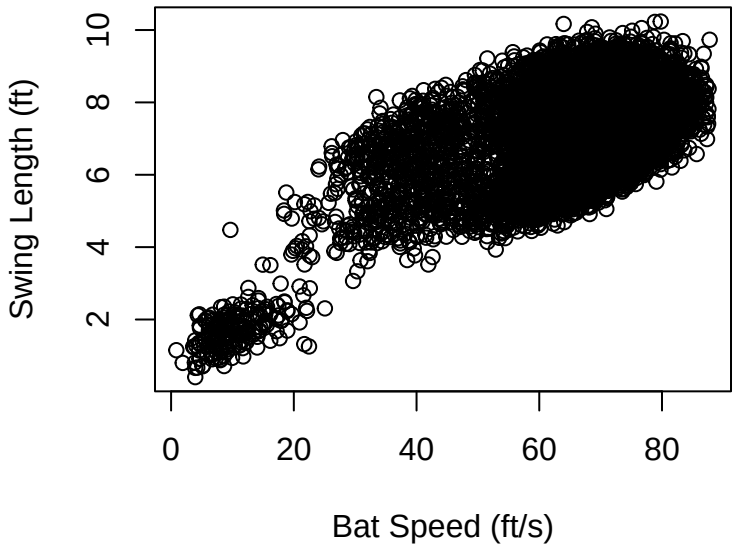
Given that our study focuses on the relationship between swing length, bat speed, and offensive production metrics, we explored the distribution of swing length and bat speed in the dataset. The histograms below show the distribution of bat speed and swing length in the dataset. Both individual plots revealed a bimodal distribution with one large peak at higher values (a full power swing) and a smaller and less frequent peak at lower values (these were investigated and found to mostly be bunts).





After the individual investigations, the two metrics were considered together in a scatterplot to understand the relationship between bat speed and swing length. The scatterplot below shows the relationship between bat speed and swing length for all batters in the dataset. While the correlation between the two metrics was calculated to only be around .5, the relationship is intuitive as a longer swing length allows for a greater range of motion and the potential for higher bat speed. Conversely, a shorter swing length at slower speeds is most often used for bunts or attempting to make contact with a pitch that is difficult to hit.

```
[1] 0.5344654
```

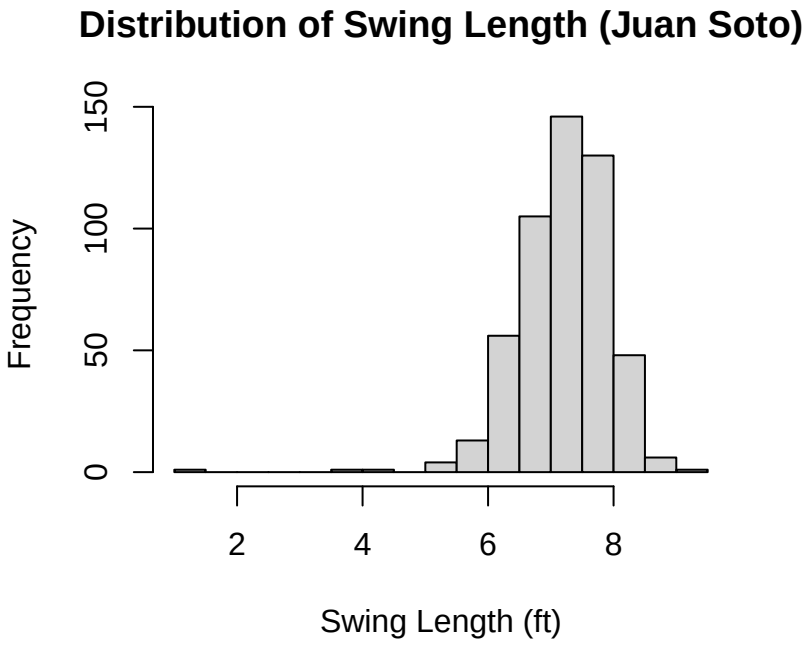


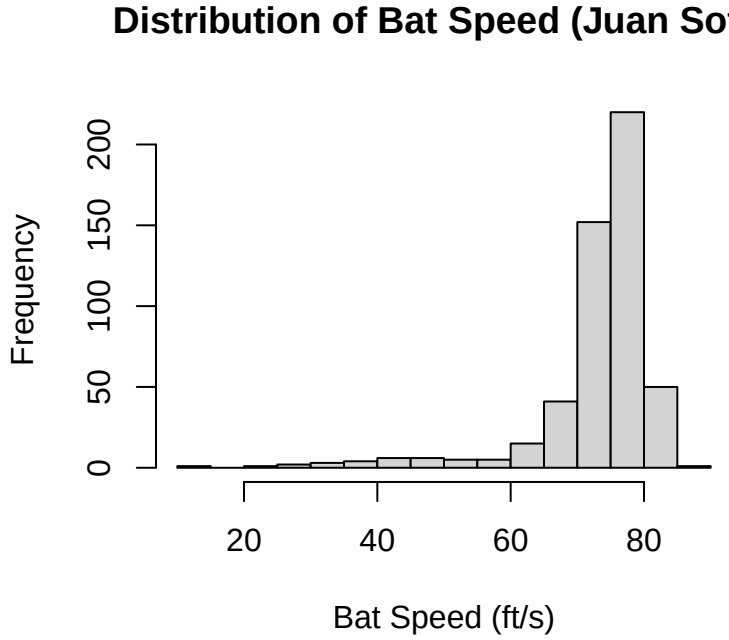
Up to this point, data exploration was conducted on the entire dataset. This made it simple to make overarching assumptions that could be verified with external sources or domain knowledge. However, it failed to consider the large number of combinations of batters, pitchers, pitch types, and game situations that cna impact the bat speed and swing length.

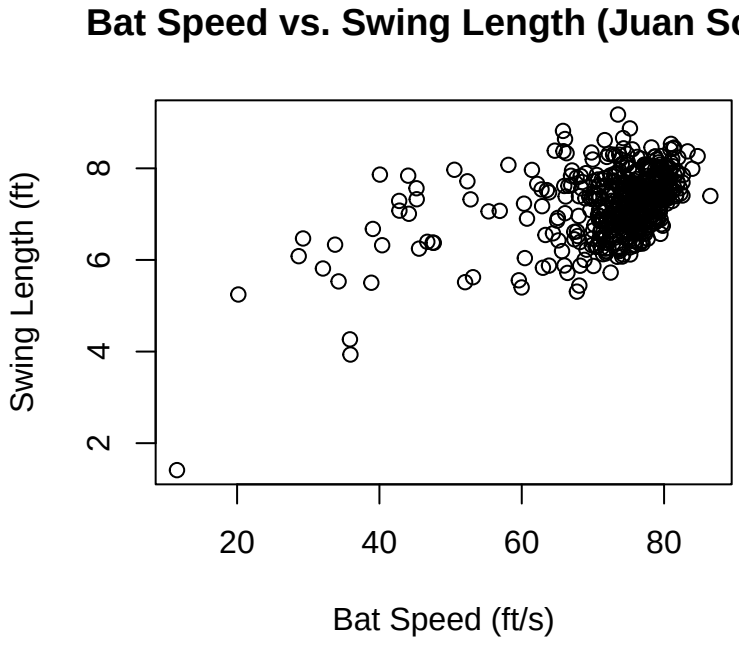
The next step was to see if the relationships observed for the whole dataset between swing length and bat speed held at a more granular level. The code below shows the first five rows of the dataset for Juan Soto, a well-known power hitter in Major League Baseball.

From the correlation coefficient and single variate plots a similar relationship between bat speed and swing length was observed for Juan Soto. The scatterplot below shows a similar association but not identical shaped cloud between bat speed and swing length. There are almost no bunt type swings and the majority of his swings are concentrated at high rates of bat speed. This suggests that the relationship between bat speed and swing length can vary across different batters due to their inherent approaches, abilities, and other characteristics. This is important to consider when ultimately giving recommendations to batters on how to optimize their swing length and bat speed for the best outcomes. What may work for one batter or even the league at large may not work for another specific batter.

```
[1] 0.4620462
```


Distribution of Swing Length (Juan Soto)
Frequency
0
50
100
150
2
4
6
8
Swing Length (ft)


Distribution of Bat Speed (Juan Soto)
Frequency
0
50
100
150
200
20
40
60
80
Bat Speed (ft/s)



Section 3: Modeling/Analysis

We developed linear regression models to examine two primary objectives: (1) predicting bat speed and swing length using game-state and pitch-specific variables and (2) predicting estimated weighted on-base average (wOBA) using swing mechanics and game context variables. Both analyses are grounded in linear regression, allowing us to quantify the relationships between predictors and outcomes and for a straightforward interpretation of results to ultimately help batters optimize their offensive performance.

Predicting Bat Speed and Swing Length

The first two models aimed to predict bat speed and swing, a key factor in the offensive production. Of the two new metrics in the dataset, bat speed is less dependent on pitcher controlled factors like pitch type, speed, and location, allowing for a greater focus on batter controlled inputs. We considered game-state variables such as the number of balls, strikes, and outs when up, as well as the type of pitch thrown (categorized by pitch category). These variables served as predictors, while the outcome variable was either bat speed or swing length. Observations were limited to rows where both `bat_speed` and `swing_length` were not missing, ensuring a focus solely on swings.

In our initial examination, we experimented with different transformations of the predictors to better capture relationships and improve model performance. Specifically, we considered converting the game-state variables (balls, strikes, and outs) into factors rather than keeping them as continuous numerical variables. This transformation allows us to account for the possibility that the relationship between these variables and the outcome (bat speed or swing length) might not be linear and might vary based on discrete states (e.g., 1 ball vs. 2 balls, 0 strikes vs. 1 strike).

By converting balls, strikes, and outs_when_up into factors, we hypothesize that each count represents a distinct game-state situation, which may influence the outcome in a way that is not strictly linear. We also included pitch_category as a categorical predictor to examine how the type of pitch impacts swing mechanics.

The next step was to evaluate the performance of the models by inspecting their summaries. These provide valuable information on the significance of each predictor, the goodness of fit, and how much variability in the outcome can be explained by the model.

```
Call:
```

```
lm(formula = swing_length ~ factor(balls) + factor(strikes) +
    factor(outs_when_up) + pitch_category, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-7.1430 -0.4692  0.0720  0.5685  3.0651

Coefficients:
                         Estimate Std. Error  t value Pr(>|t|)
(Intercept)              7.541241   0.007897  954.906  < 2e-16 ***
factor(balls)1           0.067118   0.006582   10.197  < 2e-16 ***
factor(balls)2           0.131167   0.007771   16.879  < 2e-16 ***
factor(balls)3           0.248414   0.010229   24.286  < 2e-16 ***
factor(strikes)1        -0.111094   0.006949  -15.987  < 2e-16 ***
factor(strikes)2        -0.224453   0.007294  -30.772  < 2e-16 ***
factor(outs_when_up)1    0.021957   0.006529    3.363 0.000772 ***
factor(outs_when_up)2    0.061886   0.006581    9.403  < 2e-16 ***
pitch_categoryCurveball  0.216222   0.011222   19.268  < 2e-16 ***
pitch_categoryFastball  -0.721622   0.006230 -115.822  < 2e-16 ***
pitch_categoryOffspeed   0.252413   0.008567   29.463  < 2e-16 ***
pitch_categoryOther     -0.231140   0.061771   -3.742 0.000183 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8859 on 109054 degrees of freedom
Multiple R-squared:  0.1819,    Adjusted R-squared:  0.1819
F-statistic:  2205 on 11 and 109054 DF,  p-value: < 2.2e-16


Call:
lm(formula = bat_speed ~ factor(balls) + factor(strikes) + factor(outs_when_up) +
    pitch_category, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-69.442  -2.444   1.301   4.550  20.758

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             69.34814    0.07570 916.066  < 2e-16 ***
factor(balls)1           1.06832    0.06309  16.933  < 2e-16 ***
factor(balls)2           1.84985    0.07449  24.834  < 2e-16 ***
factor(balls)3           3.21704    0.09805  32.810  < 2e-16 ***
factor(strikes)1        -0.79493    0.06661 -11.934  < 2e-16 ***
factor(strikes)2        -2.74942    0.06992 -39.323  < 2e-16 ***
factor(outs_when_up)1    0.30685    0.06259   4.903 9.47e-07 ***
factor(outs_when_up)2    0.34676    0.06309   5.497 3.88e-08 ***
pitch_categoryCurveball -0.06922    0.10757  -0.643  0.51994
pitch_categoryFastball   0.18948    0.05972   3.173  0.00151 **
pitch_categoryOffspeed   0.99744    0.08212  12.146  < 2e-16 ***
pitch_categoryOther      0.84777    0.59212   1.432  0.15222
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 8.492 on 109054 degrees of freedom
Multiple R-squared:  0.02127,   Adjusted R-squared:  0.02117
F-statistic: 215.4 on 11 and 109054 DF,  p-value: < 2.2e-16
```

For the swing length model, we observed that all of the game-state variables (balls, strikes, and outs_when_up) were statistically significant in comparison to their reference categories (e.g., 0 balls, 0 strikes, 0 outs). This suggests that changes in the count have meaningful impacts on swing length. Additionally, the pitch categories were significant when compared to the reference category of breaking balls, indicating that different pitch types lead to variations in swing length. The model explains 18.42% of the variability in swing length, as indicated by the R-squared value ($R^2 = 0.1842$). This suggests that while the game-state variables and pitch type account for some of the variance, much of the variability in swing length remains unexplained by the model. This will be a point for further analysis when developed for the CSAS 2025 Data Challenge.

Similarly, for the bat speed model, we found that all of the game-state variables were statistically significant when compared to their reference categories. This indicates that changes in the count (balls, strikes, outs) play a significant role in influencing bat speed. However, when examining the pitch categories, only certain types of pitches were statistically significant compared to the reference category of breaking balls. In particular, curve balls did not show a significant relationship with bat speed, suggesting that the batter's response to curve balls might differ from a pitch that is a breaking ball. The R-squared value for the bat speed model was very low at 2.16% ($R^2 = 0.02162$), indicating that the model explains only a small portion of the variability in bat speed. This suggests that while the model captures some relationships, there are likely other factors influencing bat speed that are not accounted for in this initial model. This, too, will be further examined in the future.

The coefficients from these models provide further insights into the relationship between predictors and the outcome variables (swing length and bat speed). For example, in the swing length model, the coefficient for balls (when factorized) suggests that batters adjust their swing length depending on how many balls they have in the count. A positive coefficient means that the swing length tends to increase as the count progresses in favor of the batter (e.g., 2 balls vs. 0 balls). Similarly, the coefficients for outs_when_up show the positive relationship between these game-state variables and swing mechanics, with different coefficients indicating the magnitude of these effects. For strikes, we see a negative relationship, where more strikes are associated with shorter swing lengths. The pitch category coefficients provide insights into how different pitch types influence swing length, with significant coefficients indicating that certain pitch types lead to longer or shorter swings compared to breaking balls.

For the bat speed model, the coefficients for the count variables (balls, strikes, outs_when_up) echoed the same relationships we saw in the swing length model. The pitch category coefficients showed that different pitch types have varying effects on bat speed, with some pitches leading to higher bat speeds than others. As previously mentioned, curve balls were not statistically significant in this model, suggesting that they do not have a significant impact on bat speed compared to breaking balls. However, Off speed balls are associated with higher bat speeds than fast balls, for example, when both are compared to breaking balls.

The results of this model indicated that all included game-state variables, such as the number of strikes and balls, play a significant role in shaping swing mechanics, including swing length and bat speed. While the models provide insights into how these factors affect performance, their relatively low R-squared values suggest that other unexamined variables— such as batter-pitcher interactions, location of the pitch, or batter-specific factors— likely contribute to swing dynamics in ways that are not fully captured by these initial models.

This model is appropriate given the focus on understanding linear relationships, but its simplicity may limit its predictive power. The results are easy to interpret, as they align with baseball intuition and can be explained to non-technical audiences in terms of game context and pitch type. In future work, it may be beneficial to explore more complex models, including interaction terms and additional predictors such as batter and pitcher handedness, pitch location, and pitch velocity. Additionally, nonlinear models or mixed effects models could provide a more nuanced understanding of how game-state variables and pitch type influence swing mechanics. By improving model accuracy, we can better predict batter performance and gain deeper insights into the relationship between swing mechanics and game context.

Predicting wOBA

The second set of models sought to predict estimated wOBA, a widely used measure of offensive productivity that accounts for various outcomes like singles, doubles, home runs, and walks, weighted by their run value. We used game-state variables as predictors, building on the models from the previous analysis. Specifically, we examined how bat speed and swing length— which reflect batter mechanics— contribute to offensive production, and how additional contextual factors like balls, strikes, outs when up, and pitch category affect this outcome.

Several models were considered:

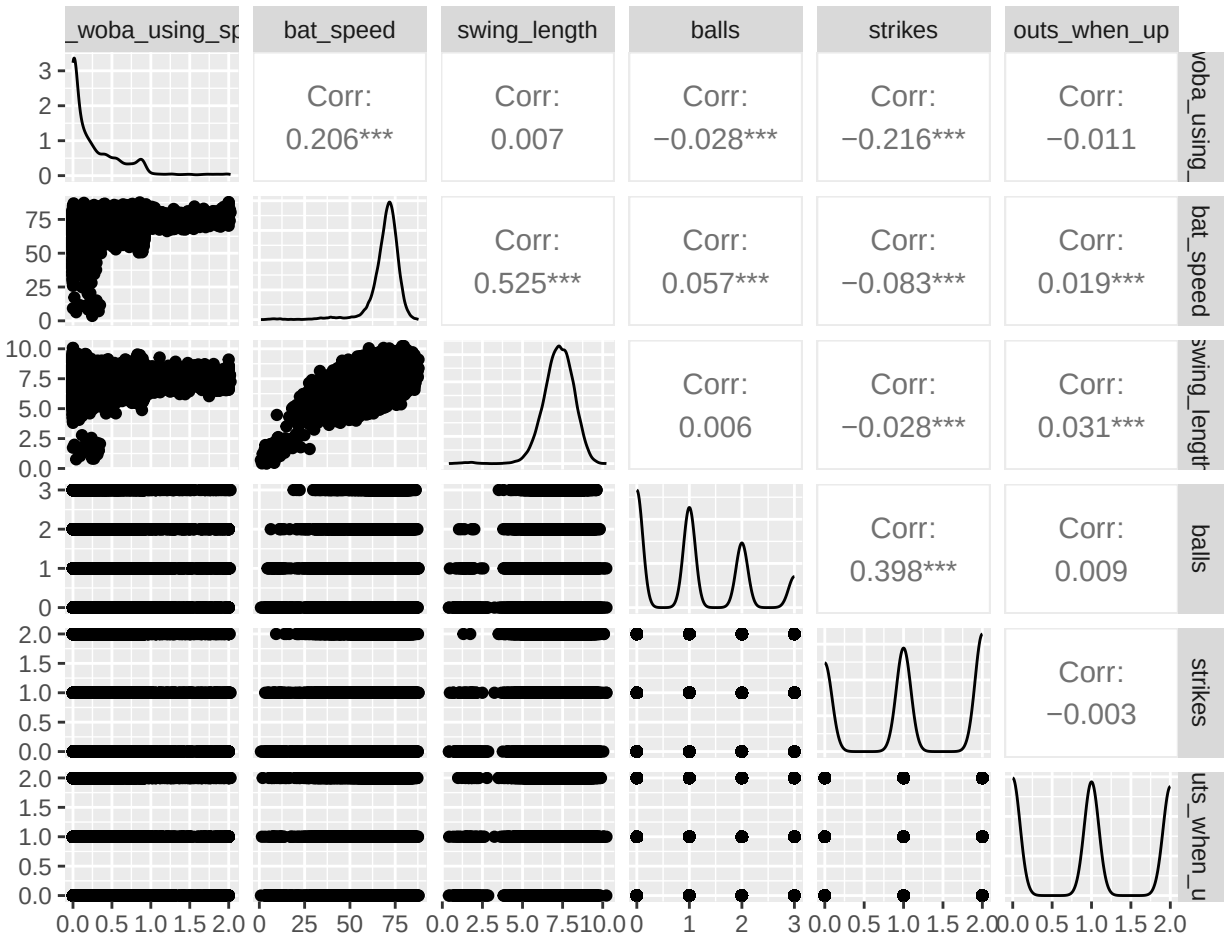
- A simple model using only bat_speed and swing_length as predictors.
- A more comprehensive model incorporating balls, strikes, outs_when_up, and pitch_category.
- A model with potential noise removed to see which combination of predictors best explains the variability in estimated wOBA.

In these models, we again focused on transforming the game-state variables to account for their categorical nature. As with the previous models predicting bat speed and swing length, we chose to treat balls, strikes, and outs_when_up as factors rather than numeric variables. This decision stemmed from the understanding that these counts represent discrete stages in an at-bat, and their influence on the outcome may not be linear. For example, the impact of being at 3 balls versus 0 balls might be more complex than a simple numerical difference.

In addition to the two batter mechanics variables (bat_speed and swing_length), the expanded model added the following predictors:

- balls and strikes: These game-state variables capture the batter's position in the count, influencing the likelihood of certain outcomes.
- outs_when_up: This variable reflects the pressure of the game situation. Batters might adjust their approach when there are more outs, and this is an important contextual factor.
- pitch_category: This factor variable categorizes the type of pitch faced (e.g., fastball, curveball, breaking ball), allowing us to see how pitch type influences wOBA.

We start by examining the scatterplot matrix to visualize the relationships between the predictors and the outcome variable (estimated wOBA). This visualization can help identify potential nonlinear relationships or interactions that may not be captured by linear models.



Based on the scatterplot matrix above, we can see the relationships between wOBA, bat speed, swing length, balls, strikes, and outs when up, providing a visual representation of how these variables interact with each other and with the outcome variable. This visualization can help identify potential nonlinear relationships or interactions that may not be captured by linear models.

```
Call:
lm(formula = estimated_woba_using_speedangle ~ bat_speed + swing_length,
    data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4940 -0.2584 -0.1259  0.1566  1.7560

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2643945  0.0181656  -14.55   <2e-16 ***
bat_speed     0.0125617  0.0002509   50.06   <2e-16 ***
swing_length -0.0447831  0.0021510  -20.82   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3669 on 49361 degrees of freedom
  (59702 observations deleted due to missingness)
Multiple R-squared:  0.04832,   Adjusted R-squared:  0.04828
```

```
F-statistic:  1253 on 2 and 49361 DF,  p-value: < 2.2e-16


Call:
lm(formula = estimated_woba_using_speedangle ~ bat_speed + swing_length +
    factor(balls) + factor(strikes) + factor(outs_when_up) +
    pitch_category, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5154 -0.2387 -0.1209  0.1462  1.8131

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -0.1072547  0.0195832  -5.477 4.35e-08 ***
bat_speed                 0.0097622  0.0002618  37.296  < 2e-16 ***
swing_length             -0.0296831  0.0023926 -12.406  < 2e-16 ***
factor(balls)1            0.0060015  0.0040924   1.466  0.14252
factor(balls)2            0.0239471  0.0046639   5.135 2.84e-07 ***
factor(balls)3            0.0385908  0.0058612   6.584 4.62e-11 ***
factor(strikes)1         -0.0142751  0.0047085  -3.032  0.00243 **
factor(strikes)2         -0.1527924  0.0045708 -33.428  < 2e-16 ***
factor(outs_when_up)1    -0.0047845  0.0039447  -1.213  0.22518
factor(outs_when_up)2    -0.0111045  0.0039752  -2.793  0.00522 **
pitch_categoryCurveball  -0.0037992  0.0066259  -0.573  0.56638
pitch_categoryFastball    0.0109155  0.0041161   2.652  0.00801 **
pitch_categoryOffspeed   -0.0071401  0.0050893  -1.403  0.16064
pitch_categoryOther      -0.0467305  0.0377088  -1.239  0.21526
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3605 on 49350 degrees of freedom
  (59702 observations deleted due to missingness)
Multiple R-squared:  0.08122,   Adjusted R-squared:  0.08098
F-statistic: 335.6 on 13 and 49350 DF,  p-value: < 2.2e-16
```

After training the models, we reviewed the model summaries to understand the contribution of each predictor. The results indicated that in model 1 (bat_speed and swing_length only), the R-squared value of 4.83% was quite low, suggesting that these two predictors alone do not explain much of the variance in wOBA, though they are both statistically significant predictors. This model provides a basic understanding but is insufficient for accurately predicting offensive performance.

Model 2, which included game-state variables like balls, strikes, outs_when_up, and pitch_category, showed a slight increase in R-squared to 8.12%. The addition of these predictors provided more explanatory power, with pitch_category contributing to the model— though it was only significant for certain pitch types, particularly fastballs. Some of the game-state variables (balls, strikes, outs_when_up) were significant in predicting estimated wOBA. However, it is noteworthy that 1 ball was not a significant predictor of wOBA when compared to 0 balls, and the same for 1 out compared to no outs when up. Pitch category also showed statistical significance for fastballs compared to breaking balls, suggesting that batters' responses to fastballs are more predictive of offensive productivity than their responses to breaking pitches. The model accounts for 8.12% of the variability in wOBA, indicating that contextual factors (such as pitch type and count) add some valuable explanatory power beyond the basic batter mechanics.

```
Call:
lm(formula = estimated_woba_using_speedangle ~ bat_speed + swing_length +
```

```
    factor(balls) + factor(strikes) + factor(outs_when_up), data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5092 -0.2390 -0.1212  0.1458  1.8243

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           -0.0873260  0.0186832  -4.674 2.96e-06 ***
bat_speed              0.0099681  0.0002552  39.054  < 2e-16 ***
swing_length          -0.0337306  0.0021319 -15.822  < 2e-16 ***
factor(balls)1         0.0059988  0.0040915   1.466 0.142605
factor(balls)2         0.0248792  0.0046562   5.343 9.17e-08 ***
factor(balls)3         0.0412331  0.0058199   7.085 1.41e-12 ***
factor(strikes)1      -0.0163738  0.0046779  -3.500 0.000465 ***
factor(strikes)2      -0.1553314  0.0045232 -34.341  < 2e-16 ***
factor(outs_when_up)1 -0.0051380  0.0039436  -1.303 0.192624
factor(outs_when_up)2 -0.0112057  0.0039752  -2.819 0.004821 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3606 on 49354 degrees of freedom
  (59702 observations deleted due to missingness)
Multiple R-squared:  0.08092,   Adjusted R-squared:  0.08075
F-statistic: 482.8 on 9 and 49354 DF,  p-value: < 2.2e-16
```

In Model 3, we further refined the model by removing the pitch_category variable and focusing on the same set of predictors (bat speed, swing length, and game-state variables). This model still performed reasonably well ($R^2 = 0.8092$) but showed that pitch category was an important variable in explaining wOBA, especially for fastballs, reinforcing the idea that pitch type plays a substantial role in offensive productivity.

In terms of model performance, the increase in R-squared values from Model 1 to Model 2 shows that adding contextual factors such as pitch type and game situation improves the model's explanatory power. The R-squared value for Model 2 was the highest of the three models, suggesting that the factors we included (bat speed, swing length, balls, strikes, outs when up, and pitch category) contribute meaningful predictive value to wOBA.

Model diagnostics indicated a reasonable fit for the data, though there are some areas for improvement. While the inclusion of game-state variables enhanced the model's performance, nonlinear relationships or the potential need for mixed-effects models (to account for batter and pitcher variability) could further refine the predictions. This could involve introducing interaction terms between game-state variables or exploring how batter-pitcher matchups affect wOBA in a more granular way.

As far as coefficients, interestingly, wOBA appears to have a negative relationship with swing length in all three models, suggesting that longer swings are associated with lower wOBA. This seems slightly counter intuitive as we know that swing length and bat speed have a positive correlation, but higher bat speeds are associated with higher wOBA, whereas longer swing lengths are not. This could be due to the fact that longer swings are more likely to be bunts or other types of swings that are not intended to produce hits.

The coefficients for the game-state variables provided insights into how balls, strikes, and outs_when_up influence wOBA. A higher count of balls tends to increase wOBA, whereas a higher count of strikes tend to decrease it. The number of outs when up also has a negative relationship with wOBA, suggesting that batters perform worse when there are more outs. The pitch category coefficients showed that fastballs are associated with higher wOBA compared to breaking balls, while other pitch categories did not differ significantly from breaking balls. This aligns with baseball intuition, as fastballs are generally easier to hit and more likely to produce hits than breaking balls.

Overall, this model provides valuable insights into the predictors of wOBA, but it is more complex and challenging to interpret for non-technical audiences due to the interplay of batter mechanics and game context. The most notable finding is how longer swing lengths, more strikes, and more outs when up are associated with lower wOBA, while fastballs are associated with higher wOBA. These insights can be used to inform batters on how to optimize their offensive performance based on the game situation and pitch type. While the model has some explanatory power, future improvements could focus on exploring nonlinear relationships, incorporating mixed-effects models to account for batter-pitcher interactions, and validating the model with additional features, such as pitch location or batter-pitcher handedness, to further enhance its predictive accuracy.

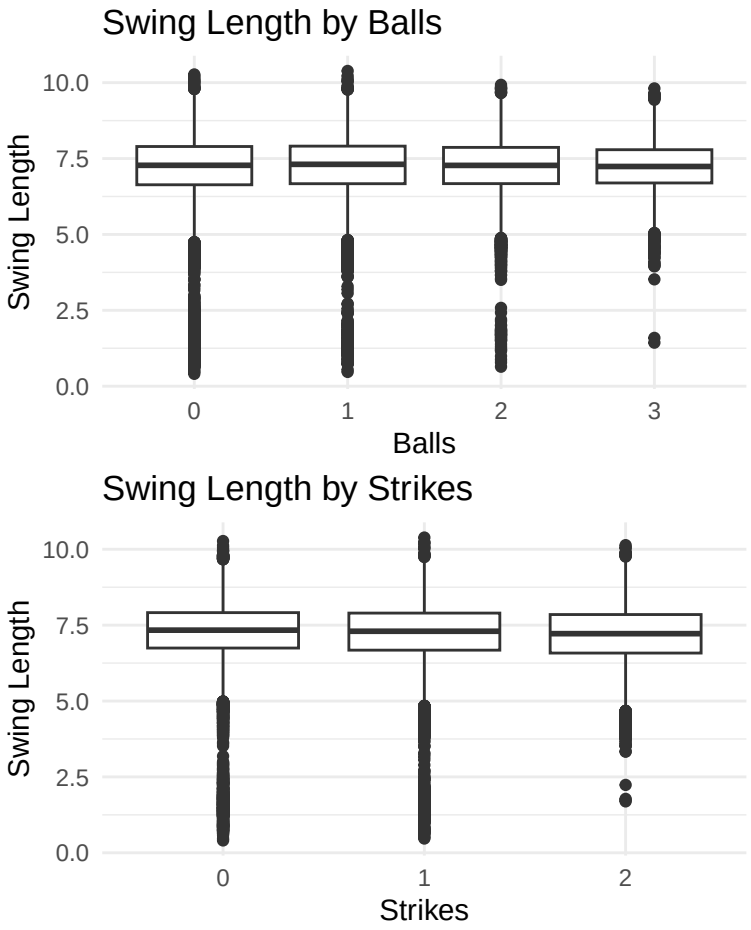
General Analysis Considerations

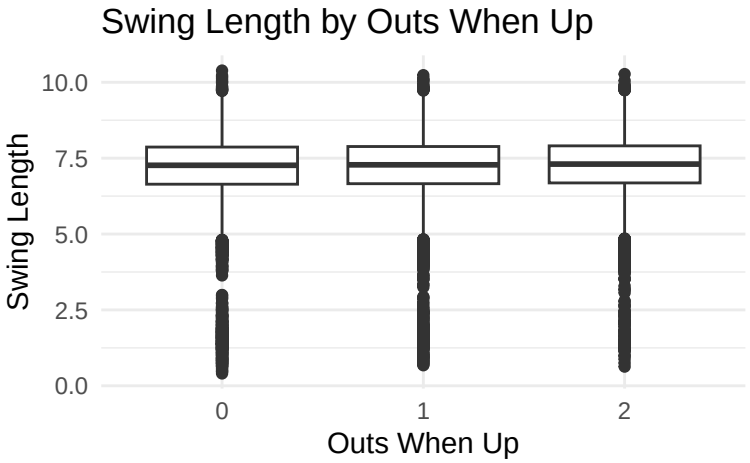
Both models rest on several assumptions standard in linear regression analysis, including linearity, normality of residuals, and homoscedasticity. Linear regression was chosen for its simplicity and interpretability, which align with the goals of understanding relationships and generating actionable insights. Future work may explore nonlinear methods, interaction terms, or advanced approaches like mixed-effects models to account for batter- and pitcher-specific effects.

Section 4: Visualization and interpretation of the results

Visualizing the Effect of Game-State Variables on Swing Length and Bat Speed

One of the primary insights from our models was that game-state variables such as balls, strikes, and outs when up, significantly influence both swing length and bat speed. To illustrate these relationships, we can create bar plots showing the average swing length and bat speed for different categories of these game-state variables.

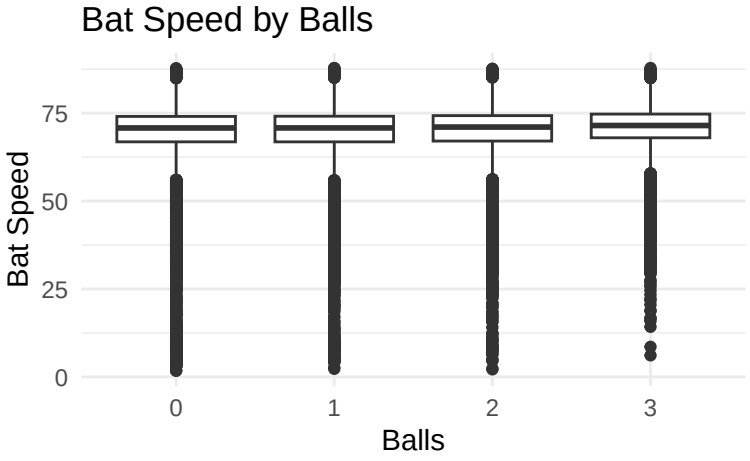


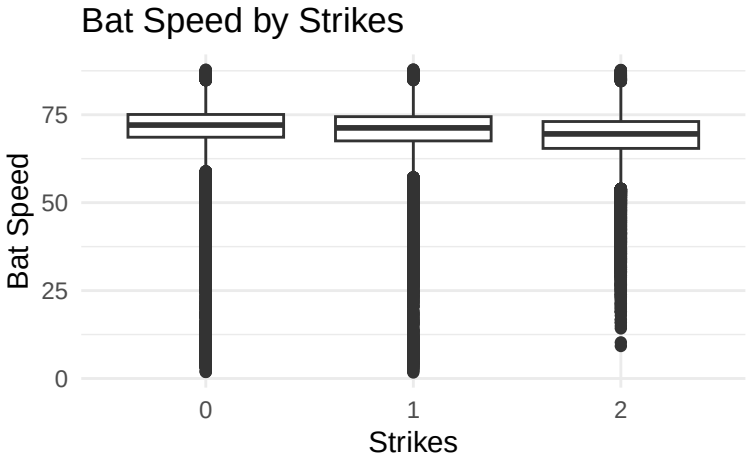


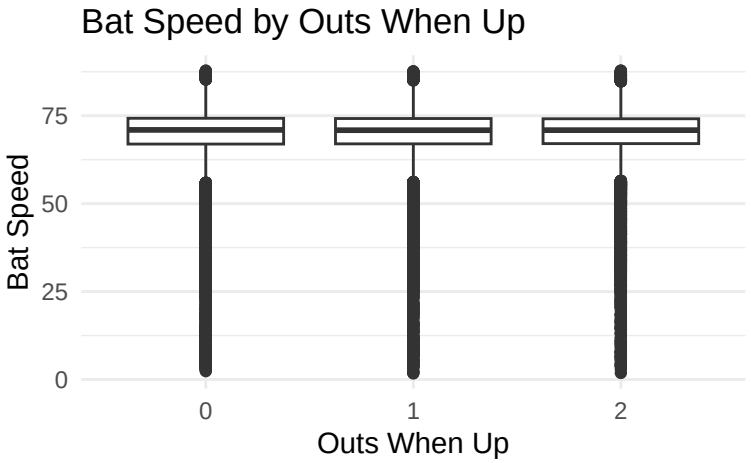
These visualizations help highlight how the swing length changes depending on the count and the number of outs. For example, we might observe that batters tend to take longer swings with a higher number of balls in the count, likely in anticipation of a more favorable pitch. On the other hand, batters with more strikes or more outs might shorten their swings to increase contact probability.

Bat Speed by Balls, Strikes, and Outs When Up

Similarly, we can visualize how bat speed varies with these game-state variables. By creating bar plots of average bat speed for different counts and outs when up, we can gain insights into how these factors influence bat speed.



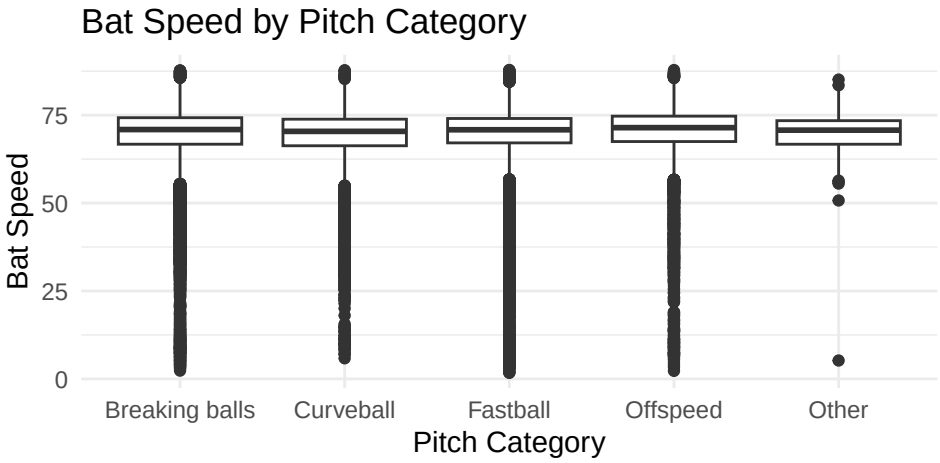


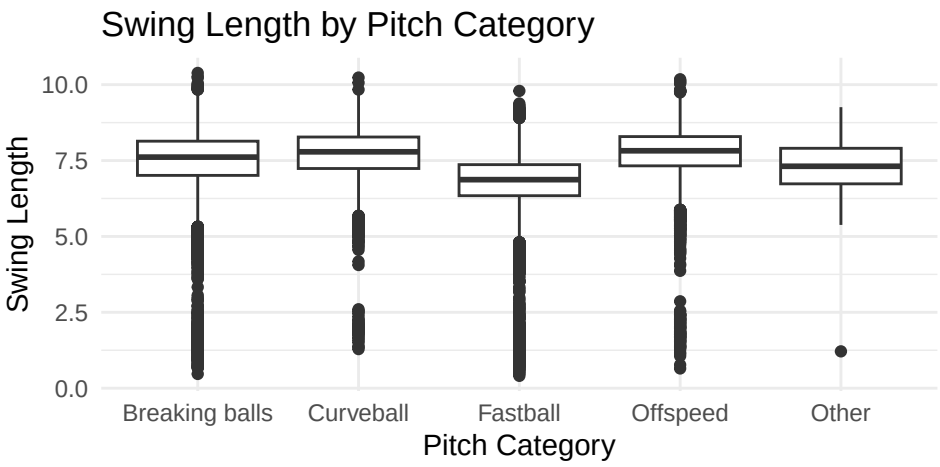


These box plots can illustrate how bat speed varies under different game conditions. We might find that bat speed tends to be higher when there are fewer outs, as batters may be more aggressive in those situations. The effect of the count (balls and strikes) may be more subtle, with batters perhaps swinging harder when ahead in the count.

Pitch Category Impact on Swing Length and Bat Speed

The pitch type also plays a significant role in influencing swing mechanics and performance. To illustrate how different pitch categories affect bat speed and swing length, we can use box plots to show the distribution of bat speed and swing length by pitch category.

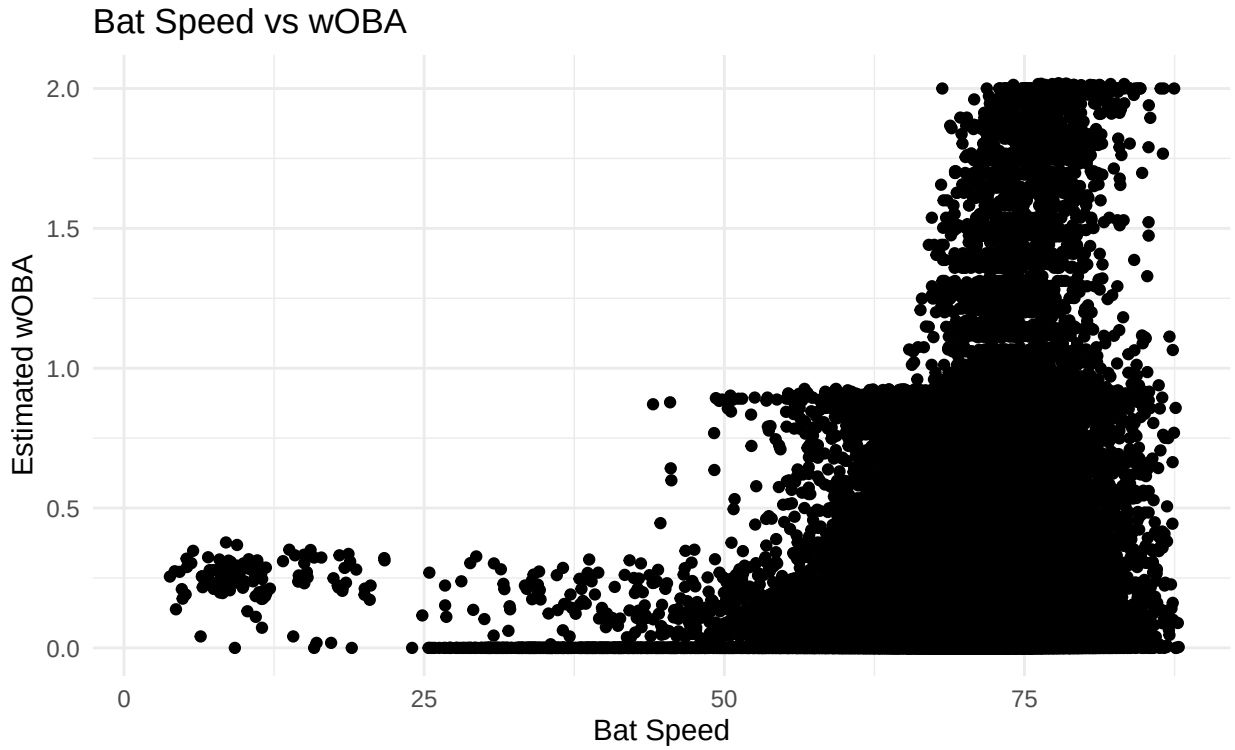


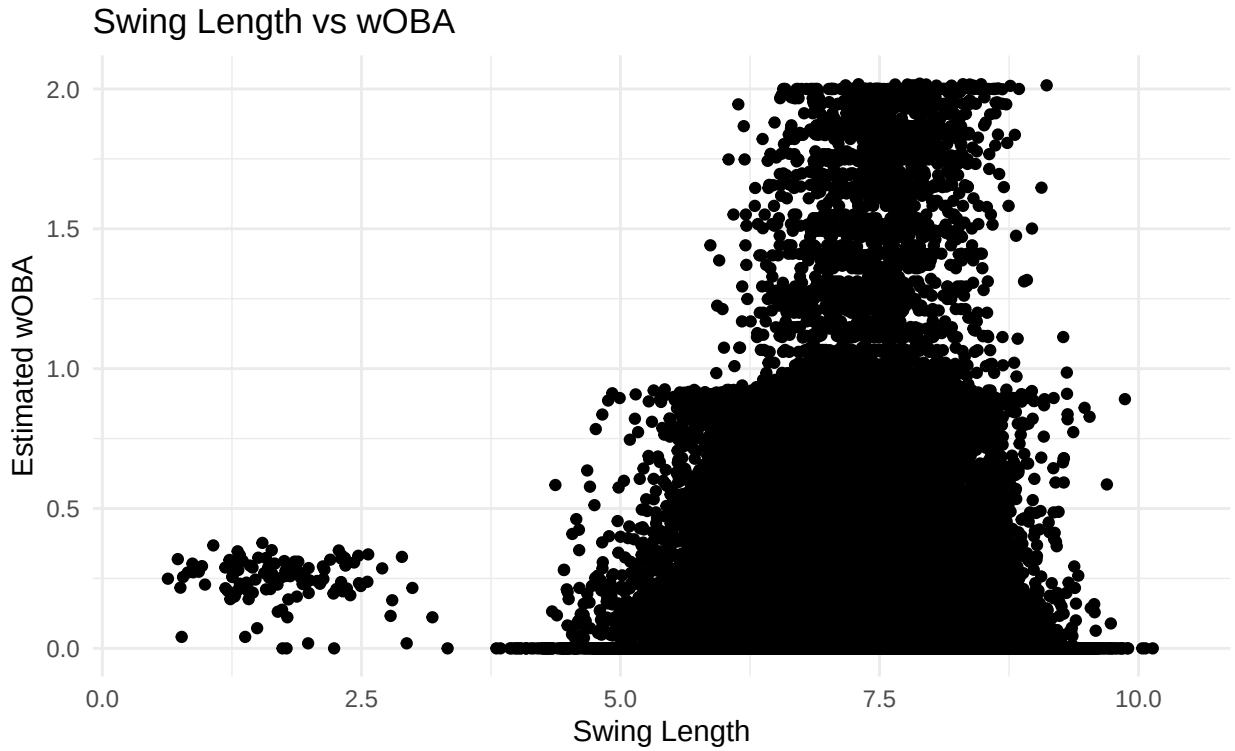


These plots help visually assess the impact of different pitch types. For example, we may find that bat speed is higher for fastballs compared to breaking balls, while swing length may vary more for certain types of pitches, like off-speed pitches.

Interpreting the Model Results: Swing Length and Bat Speed Impact on wOBA

In the wOBA prediction models, we found that bat speed and swing length had significant relationships with estimated wOBA. However, the effect of swing length was somewhat counterintuitive— longer swings were associated with lower wOBA, even though bat speed was positively correlated with wOBA. To further investigate this, we can create scatter plots to examine the relationship between bat speed, swing length, and wOBA.

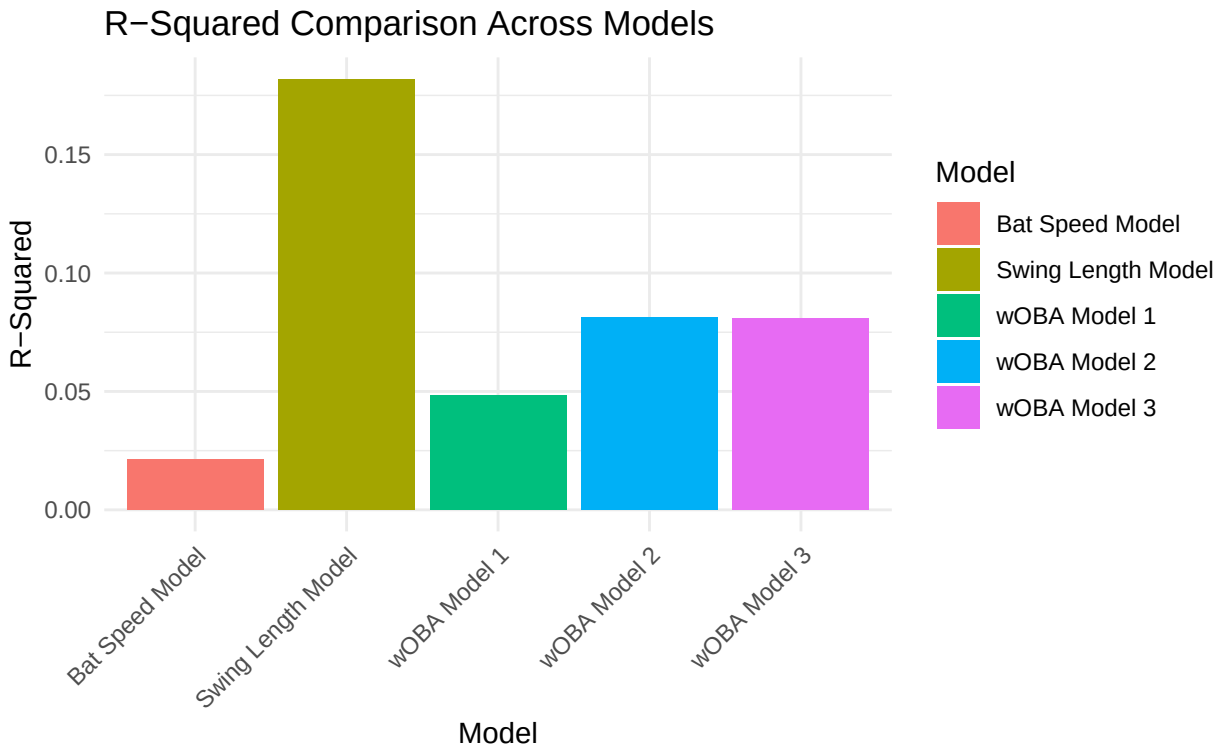




These scatter plots reveal the relationships between swing mechanics and wOBA. We might see a clearer positive correlation between bat speed and wOBA, while the relationship with swing length may appear weaker or even negative, providing further insight into how these two factors play different roles in offensive productivity.

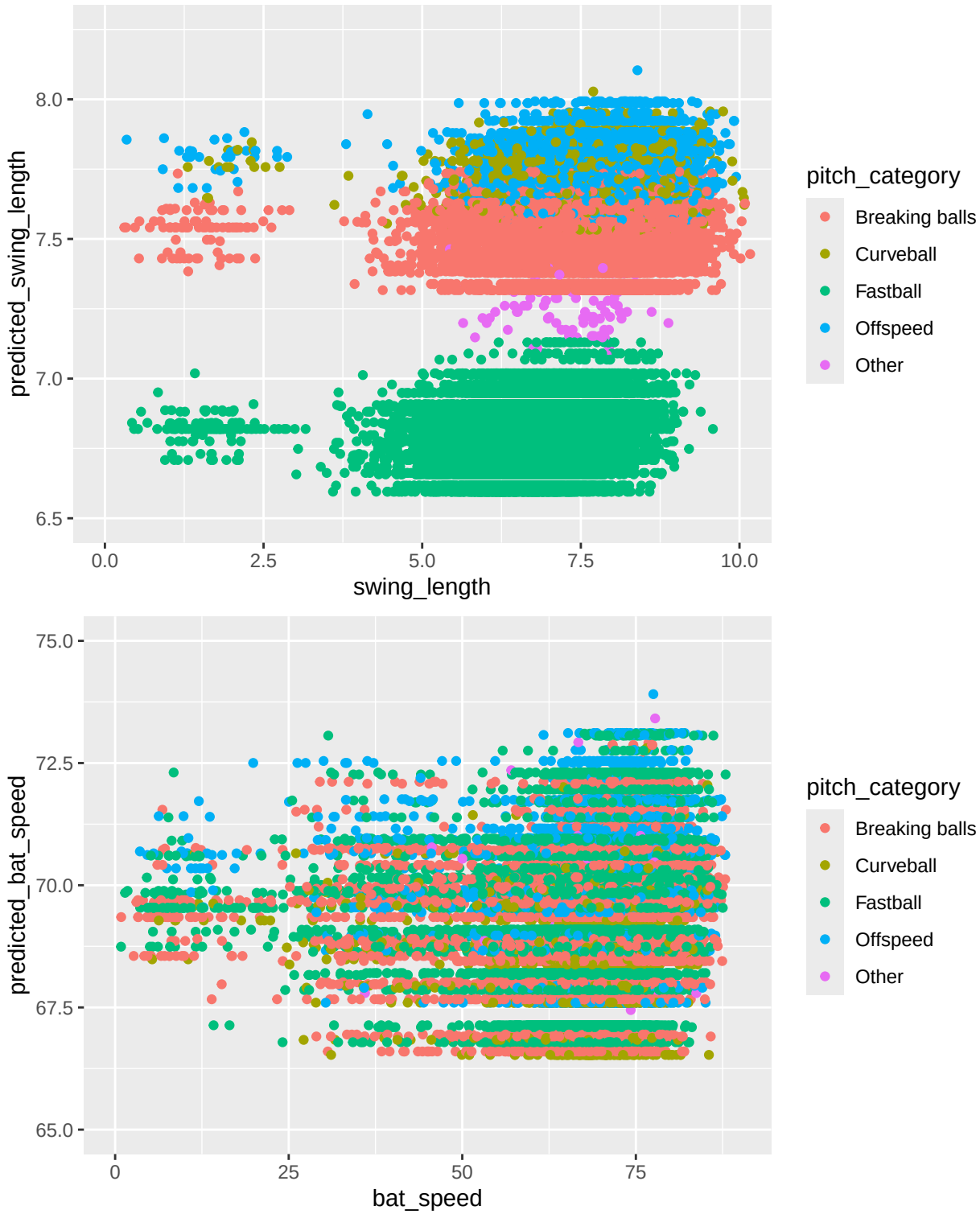
Comparing Model Performance: R-Squared and Predictive Power

To compare the performance of our models, we can use a series of plots that show the R-squared values and residuals of each model. This helps identify which model provides the best fit and predictive accuracy.



This bar plot visually demonstrates the relative explanatory power of each model. While the wOBA models (particularly Model 2) show the highest R-squared values, it's important to note that this measure alone does not capture the complete picture— model interpretability and practical application are also key factors in selecting the best model.

To assess how well the first two models predict swing length and bat speed, respectively, we plot the predicted values against the actual values in the test set. This helps visualize how closely the model's predictions align with the true data and provides an opportunity to observe any systematic errors or biases.



The plots show the actual swing length or bat speed on the x-axis and the predicted swing length or predicted bat speed on the y-axis, with data points colored by pitch category. A perfect model would result in points lying on the 45-degree diagonal line, meaning the predicted values would exactly match the actual values. However, in practice, the points tend to cluster around the diagonal line with some spread. This spread reflects the model's error— particularly, there is some degree of regression to the mean, where extreme values of swing length (either very high or low) tend to be pulled toward the average value in the predictions.

Regression to the mean indicates that the model has difficulty in accurately predicting extreme values of swing length, likely because the linear regression approach assumes a more uniform, less volatile relationship between predictors and the outcome. For pitch categories like fastballs or breaking balls, this behavior may be more pronounced as these types of pitches lead to a wider range of swing lengths, but the model may generalize too much, resulting in predictions that don't fully capture the variability in swing lengths for those pitches.

This visualization helps to highlight the overall model fit and any systematic bias, while also pointing out where improvements (such as incorporating more complex models or adding additional predictors) may be needed to better capture the variation in swing mechanics. We can clearly see that swing length is associated with certain pitch category, whereas bat speed is more uniformly distributed across pitch categories.

Section 5: Conclusions and recommendations

Our analysis of predicting bat speed, swing length, and estimated wOBA using game-state variables and swing mechanics reveals several key insights. Game-state factors like balls, strikes, and outs when up play a significant role in shaping swing mechanics, with batters adjusting their swing length and bat speed based on the count and the number of outs. Pitch type also impacts swing mechanics, with fastballs generally associated with higher bat speeds but shorter swing lengths.

For predicting wOBA, we found that bat speed and swing length are important predictors, but the relationship between swing length and wOBA was counterintuitive— longer swings tend to be associated with lower wOBA. The addition of game-state variables such as balls, strikes, outs when up, and pitch category improved the explanatory power of the model, though the overall R-squared values remained modest, indicating that there are other factors influencing offensive performance that are not captured by these models.

To advance in the CSAS 2025 Data Challenge, the next steps involve refining the existing models by adding interaction terms, polynomial features, and exploring non-linear techniques like Random Forest or XGBoost for better predictive accuracy. It's important to evaluate model performance to avoid overfitting and improve generalization. Additionally, incorporating features like pitch location, batter-pitcher interactions, and pitch speed can enhance model insights. By following these steps, we can refine our models to provide deeper insights into batting performance, ensuring that our predictions are both accurate and actionable, ultimately helping batters optimize their approach in the game.

References

Baseball Savant Data Dictionary

Petriello, M. 2024: Everything to know about Statcast's new bat-tracking data

Scott Powers and Ron Yurko: Swinging, Fast and Slow

Fangraphs Website

Goldbeck, G. 2023 (in MLB Technology Blog): Introducing Statcast 2023: High Frame Rate Bat and Biomechanics Tracking