CSAS Data Challenge 2025: A Pitch-Level Analysis of MLB Swing Length and Bat Speed

Julia Stiller

1/15/2025

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 700,000 Major League Baseball plate appearances from April to October 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 701,557 Major League Baseball plate appearances from 4/2/2024 to 10/30/2024, 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 influence outcomes like weighted On-Base Average (wOBA). By modeling these relationships, we aim to provide actionable insights to help batters determine the optimal conditions under which to swing and what their swing mechanics should be, 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 were deprecated or 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) based on contextual and swing-related factors in the dataset.

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. Finally, **Section 4** concludes with a summary of the key findings and implications for hitters, as well as recommendations for further work. 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 revealed key patterns in a dataset containing over 700,000 pitches and 100 variables. 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. The visualizations and commentary below provide a glimpse into the exploratory analysis conducted on the dataset.

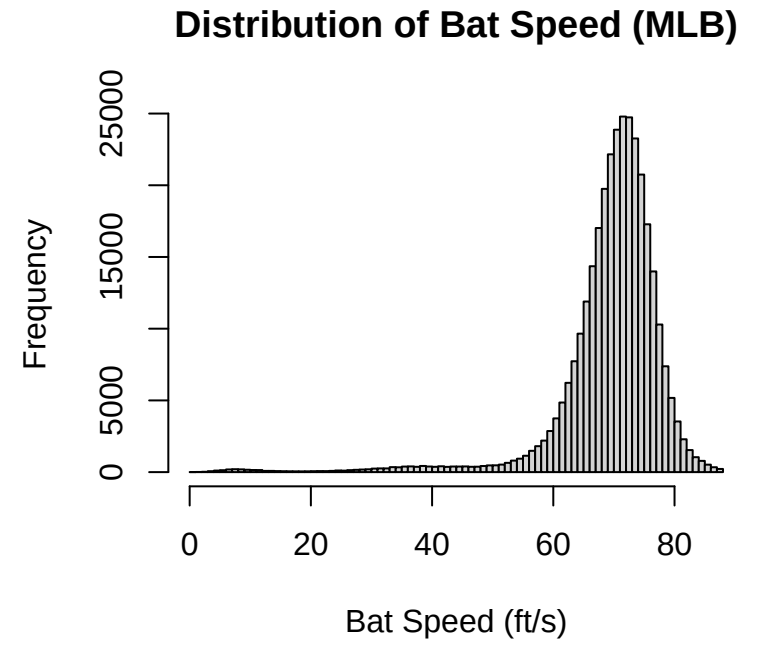
One aspect of the data that was used in the study is the number of balls and strikes present in the count as it is called before each pitch. The table below shows the number of balls and strikes in the dataset. This information is crucial for understanding the context of the pitcher-batter interaction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

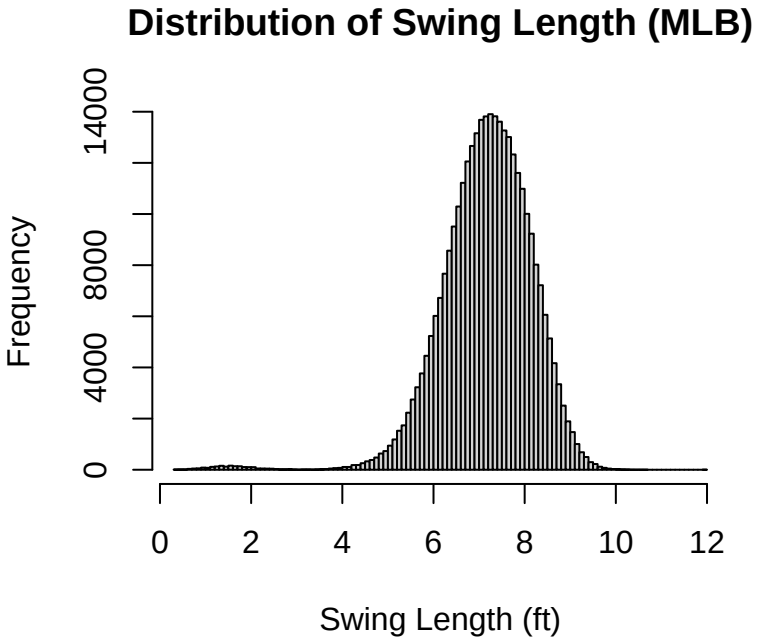
	0 Balls	1 Ball	2 Balls	3 Balls
0	180086	67165	22296	6762
1	91879	70663	35965	14762
2	48248	69333	58672	35438

The type of pitch thrown by the pitcher is another important aspect of the pitcher-batter interaction. Pitch types were categorized into 5 major groups and the table below shows the number of pitches in each of these categories. This information is crucial for batters to understand the type of pitch they are facing and adjust their swing mechanics accordingly.

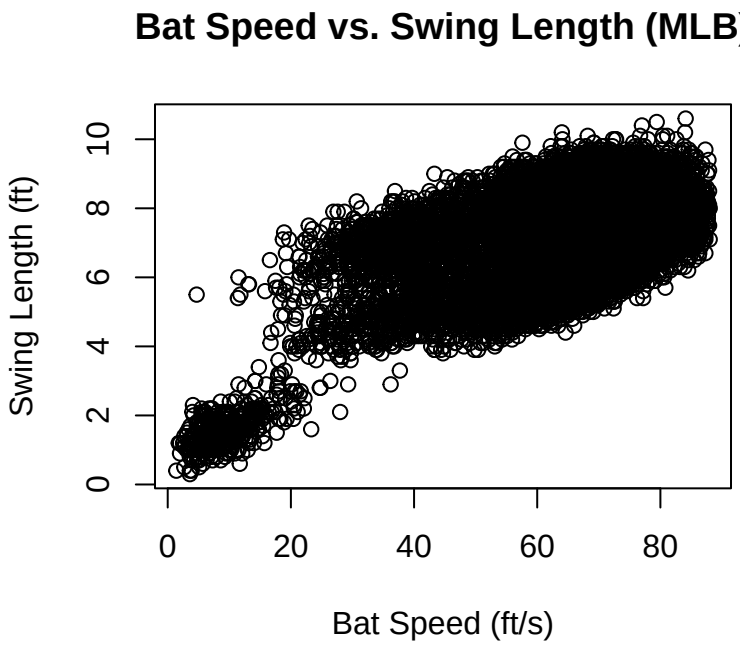
Pitch Category	Frequency
Breaking balls	213517
Curveball	57718
Fastball	335645
Offspeed	92651
Other	1738

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. Both histograms below revealed a bimodal distribution with one large peak at higher values (a full power swing) and a much smaller and less frequent peak at lower values (these were investigated and found to mostly be bunts).

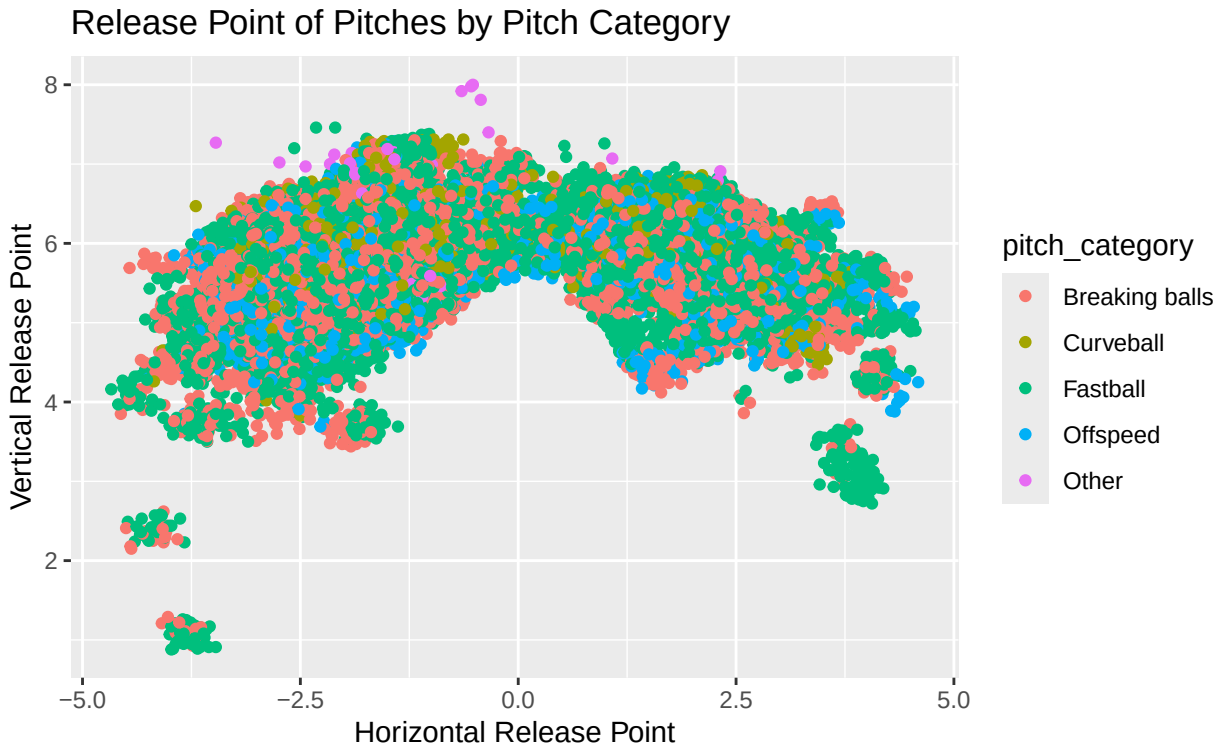




After the individual investigations, the two metrics were considered together in a scatterplot to understand the relationship between bat speed and swing length. The scatterplot below shows the relationship between bat speed and swing length for all batters in the dataset. While the correlation between the two metrics was calculated to be only 52.8%, the relationship is intuitive as a higher bat speed allows for a potentially greater range of motion. 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.

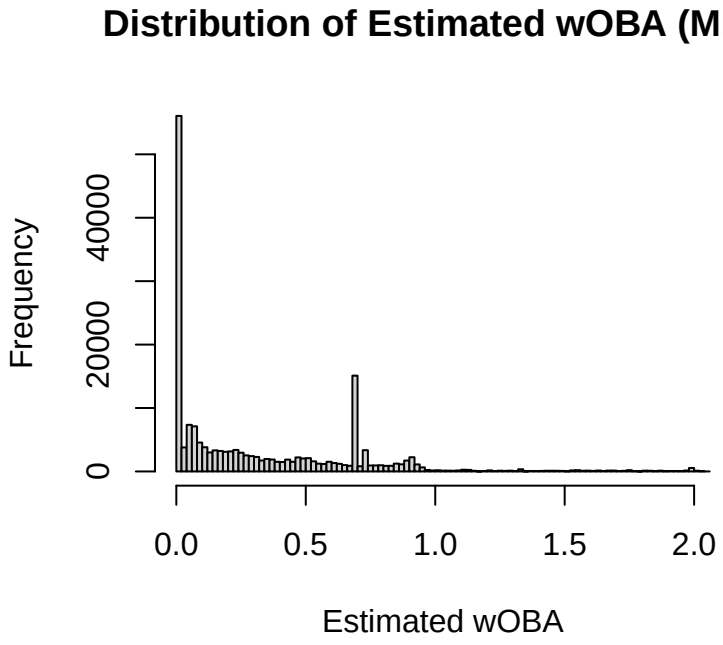


Another aspect of the pitcher-batter interaction is the release point of each pitch as it moves towards the batter. The scatter plot below shows the release points of all pitches in the dataset, color-coded by pitch category. The x-axis represents the horizontal release point of the pitch, while the 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.



One of the metrics that results from the pitcher-batter interaction is weighted On-Base Average (wOBA), which 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 comparable scales. The histogram below shows the distribution of wOBA in the dataset. This metric is important for understanding the overall offensive value of a player and can be

used to evaluate the effectiveness of a batter in a given situation.



Section 3: Models and Interpretations

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. Linear regression was chosen for its simplicity and interpretability, aligning with the goal of providing actionable insights to optimize offensive performance.

Predicting Bat Speed and Swing Length

We built models to predict bat speed and swing length using game-state variables (balls, strikes, outs) and pitch categories as predictors. To account for the categorical nature of these variables (i.e., the relationship with the response might not be linear), we transformed them into factors. Observations were limited to rows where both `bat_speed` and `swing_length` were not missing, ensuring a focus solely on swings.

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

Residuals:
    Min      1Q  Median      3Q     Max
-7.5629 -0.4743  0.0779  0.5750  4.2535

Coefficients:
                       Estimate Std. Error  t value Pr(>|t|)
(Intercept)            7.540122   0.005563 1355.411  < 2e-16 ***
factor(balls)1         0.065405   0.004640   14.096  < 2e-16 ***
factor(balls)2         0.141512   0.005485   25.799  < 2e-16 ***
factor(balls)3         0.256826   0.007184   35.748  < 2e-16 ***
factor(strikes)1      -0.107324   0.004901  -21.896  < 2e-16 ***
factor(strikes)2      -0.227295   0.005137  -44.247  < 2e-16 ***
factor(outs_when_up)1  0.027855   0.004599    6.057 1.39e-09 ***
factor(outs_when_up)2  0.058405   0.004629   12.617  < 2e-16 ***
```

```
pitch_categoryCurveball  0.214378   0.007841   27.340  < 2e-16 ***
pitch_categoryFastball  -0.749092   0.004385 -170.829  < 2e-16 ***
pitch_categoryOffspeed   0.264329   0.006101   43.327  < 2e-16 ***
pitch_categoryOther     -0.250174   0.041755   -5.992 2.08e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8974 on 225774 degrees of freedom
Multiple R-squared:  0.189, Adjusted R-squared:  0.1889
F-statistic:  4782 on 11 and 225774 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
-69.477  -2.443   1.361   4.648  21.089

Coefficients:
                        Estimate Std. Error  t value Pr(>|t|)
(Intercept)             69.26883    0.05451 1270.800  < 2e-16 ***
factor(balls)1           1.06985    0.04546   23.532  < 2e-16 ***
factor(balls)2           1.99788    0.05375   37.173  < 2e-16 ***
factor(balls)3           3.38605    0.07039   48.101  < 2e-16 ***
factor(strikes)1        -0.77456    0.04803  -16.128  < 2e-16 ***
factor(strikes)2        -2.74743    0.05033  -54.584  < 2e-16 ***
factor(outs_when_up)1    0.28919    0.04506    6.417 1.39e-10 ***
factor(outs_when_up)2    0.32143    0.04536    7.087 1.38e-12 ***
pitch_categoryCurveball -0.23707    0.07683   -3.086 0.002031 **
pitch_categoryFastball   0.15739    0.04297    3.663 0.000249 ***
pitch_categoryOffspeed   1.14927    0.05978   19.226  < 2e-16 ***
pitch_categoryOther      0.40064    0.40913    0.979 0.327452
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.793 on 225774 degrees of freedom
Multiple R-squared:  0.02122,   Adjusted R-squared:  0.02118
F-statistic: 445.1 on 11 and 225774 DF,  p-value: < 2.2e-16
```

Both models demonstrated that game-state variables significantly impact swing mechanics 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 bat speed and swing length.

In the swing length model, 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. This model explains 18.8% of the variability in swing length, as indicated by the R-squared value ($R^2 = 0.188$). 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.

For the bat speed model, only certain types of pitches were statistically significant compared to the reference category of breaking balls. In particular, the Other category did not show a significant relationship with bat speed, suggesting that the batter's response to unique pitches might differ from a pitch that is a breaking ball. The R-squared value for the bat speed model was very low at 2.08% ($R^2 = 0.02078$), indicating that the model explains only a small portion of the variability in bat speed and there are other factors influencing bat

speed that are not accounted for in this initial model.

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 (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 coefficients for the count variables (balls, strikes, outs_when_up) in the bat speed model echoed these same relationships.

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. Although the results are easy to interpret, it may be beneficial to explore more complex models (like nonlinear or mixed effects models) with interaction terms and additional predictors in future work.

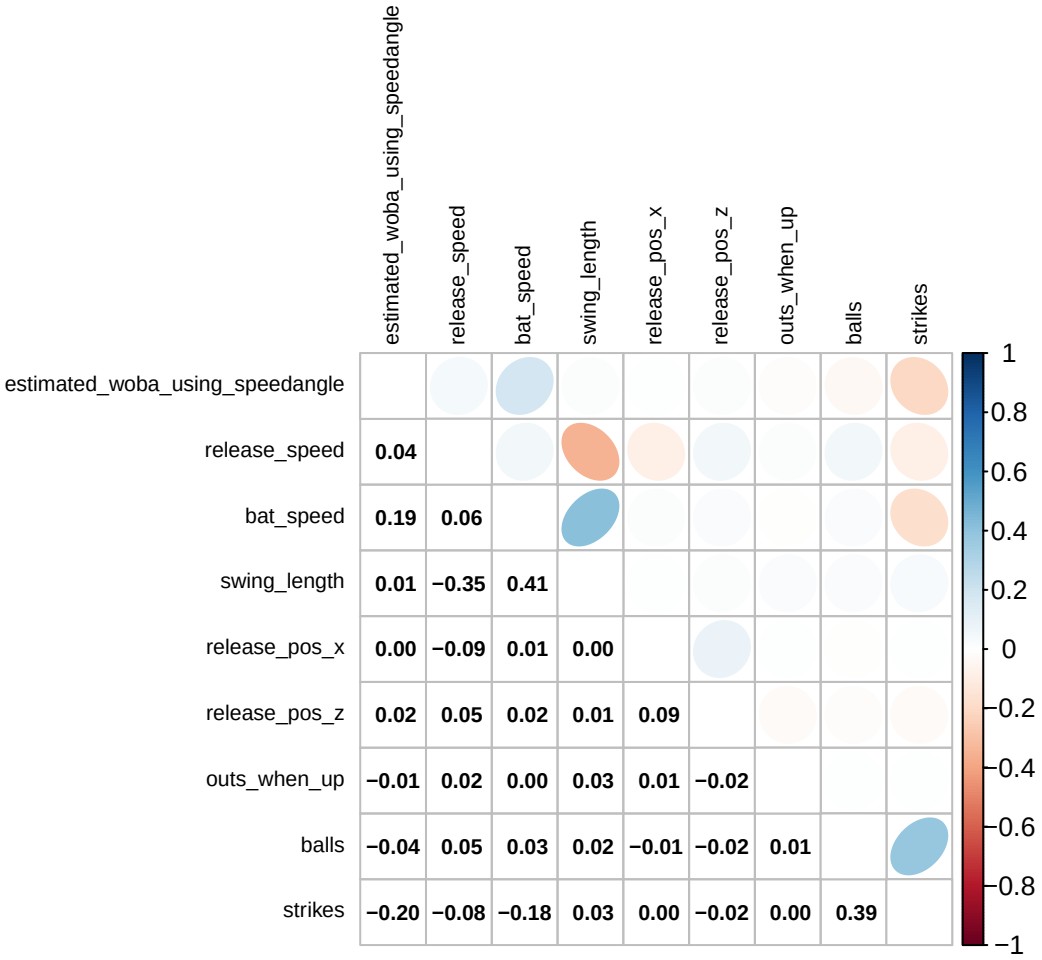
Predicting wOBA

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 now used bat speed and swing length as predictors, in addition to the game-state variables from the previous models plus additional contextual factors. Several models were initially considered and can be further explored in the appendix (A.2). Here, we examine our most comprehensive model which uses batter mechanics in addition to 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.
- release_pos_x and release_pos_z: These variables represent the horizontal and vertical release points of the pitch, respectively. They provide additional information on the pitch trajectory and location, which can influence the batter's decision-making and swing mechanics.
- release_speed: The speed at which the pitch is released can impact the batter's reaction time and swing mechanics, potentially affecting offensive productivity.

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 linear relationship.

We start by examining a correlation matrix to visualize how the numerical predictors relate to the outcome variable (estimated wOBA). We can see that the numerical variables are weakly correlated with each other, with the strongest correlation being between swing_length and bat_speed, which we examined in the prior exploration stages. This suggests that the predictors are not highly collinear, which is important for the stability of the regression models.




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

Residuals:
    Min      1Q  Median      3Q     Max
-0.5469 -0.2343 -0.1208  0.1434  1.8299

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.0430031  0.0324046   1.327  0.18449
bat_speed                0.0093575  0.0001741  53.747  < 2e-16 ***
swing_length            -0.0264462  0.0016357 -16.168  < 2e-16 ***
factor(balls)1           0.0072618  0.0028397   2.557  0.01055 *
factor(balls)2           0.0196639  0.0032405   6.068 1.30e-09 ***
factor(balls)3           0.0395401  0.0040493   9.765  < 2e-16 ***
factor(strikes)1        -0.0138870  0.0032793  -4.235 2.29e-05 ***
factor(strikes)2        -0.1570530  0.0031739 -49.483  < 2e-16 ***
factor(outs_when_up)1   -0.0070743  0.0027316  -2.590  0.00960 **
factor(outs_when_up)2   -0.0116111  0.0027544  -4.215 2.49e-05 ***
```

```
pitch_categoryCurveball  0.0002013  0.0049693   0.041  0.96769
pitch_categoryFastball   0.0265326  0.0038280   6.931 4.20e-12 ***
pitch_categoryOffspeed  -0.0057660  0.0035562  -1.621  0.10494
pitch_categoryOther     -0.0379719  0.0237401  -1.599  0.10972
release_pos_x           -0.0016005  0.0006140  -2.607  0.00915 **
release_pos_z            0.0066864  0.0021513   3.108  0.00188 **
release_speed           -0.0021576  0.0003354  -6.433 1.25e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3576 on 101455 degrees of freedom
  (124314 observations deleted due to missingness)
Multiple R-squared:  0.08483,   Adjusted R-squared:  0.08469
F-statistic: 587.8 on 16 and 101455 DF,  p-value: < 2.2e-16
```

After training and optimizing multiple models, we reviewed the summary of our most comprehensive model to understand the contribution of each predictor. This model accounts for 8.33% of the variability in estimated wOBA, as indicated by the R-squared value ($R^2 = 0.0833$). The results indicated several notable patterns and relationships that provide insights into the factors influencing weighted On-Base Average (wOBA).

Interestingly, the model shows a negative relationship between wOBA and swing length. This suggests that longer swings are associated with lower wOBA, which may initially seem counterintuitive given the positive correlation between swing length and bat speed. While higher bat speeds are typically linked to higher wOBA, longer swing lengths may not yield the same benefit. One possible explanation for this discrepancy is that longer swings are more likely to be bunts or other defensive swings that are not intended to produce strong contact or hits.

The coefficients for the game-state variables also provided valuable insights. The model indicates that a higher count of balls tends to increase wOBA, whereas a higher count of strikes tend to decrease it. This aligns with baseball strategy, as batters in favorable counts (more balls than strikes) have a higher likelihood of achieving positive outcomes. Additionally, the number of outs when up shows a negative relationship with wOBA, suggesting that batters perform worse when there are more outs. This could be attributed to increased pressure or changes in pitcher behavior as the number of outs increases.

Regarding pitch category, the model shows that fastballs and offspeed pitches are associated with higher wOBA compared to breaking balls. Curveballs and Other pitch types do not differ significantly from breaking balls. This aligns with baseball intuition, as fastballs are generally more likely to produce hits due to their predictability when compared to breaking balls, which tend to induce weaker contact or swings and misses.

The analysis of pitcher release variables provides further insights. The horizontal release point (release_pos_x) was not significant in the model, indicating that the lateral positioning of the pitch release has little impact on wOBA. In contrast, the vertical release point (release_pos_z) was somewhat significant, suggesting that the height at which the ball is released may influence batter performance. Specifically, higher release points are associated with higher wOBA. This finding could reflect batters' ability to adjust better to higher release points, potentially improving their ability to make contact. Finally, the release speed coefficient was significant and negative, indicating that faster pitches are associated with lower wOBA. This result aligns with the common understanding that faster pitches are harder to hit effectively, often resulting in more strikeouts or weak contact.

Overall, this model provides valuable insights into the predictors of wOBA, highlighting both batter mechanics and game-context variables. While the model's complexity may pose challenges for non-technical audiences, its key findings can inform batting strategies. For instance, batters might focus on optimizing their swing mechanics to avoid excessively long swings, aim to perform better in favorable counts, and adjust their approach based on the pitch type and release characteristics. The most notable takeaways from the model are the negative relationships between wOBA and longer swing lengths, more strikes, more outs when up, and faster release speeds, as well as the positive association between fastballs and higher wOBA compared to breaking balls. These insights can be used to inform batters on how to optimize their offensive performance

based on the game situation and pitch type.

General Analysis Considerations

Both sets of 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. While the model has some explanatory power, future improvements could focus on exploring nonlinear methods, interaction terms, additional features, or advanced approaches like mixed-effects models to account for batter- and pitcher-specific effects.

Section 4: 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. Notably, higher counts of strikes and outs had the most negative impact on wOBA, whereas favorable counts (more balls) positively influenced it. Fastballs were linked to higher wOBA compared to breaking balls, underscoring the importance of pitch type in offensive outcomes.

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. These findings highlight the need for batters to focus on shortening their swings in two-strike counts and recognizing pitch types to optimize outcomes.

Future improvements would 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 batter-pitcher interactions 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

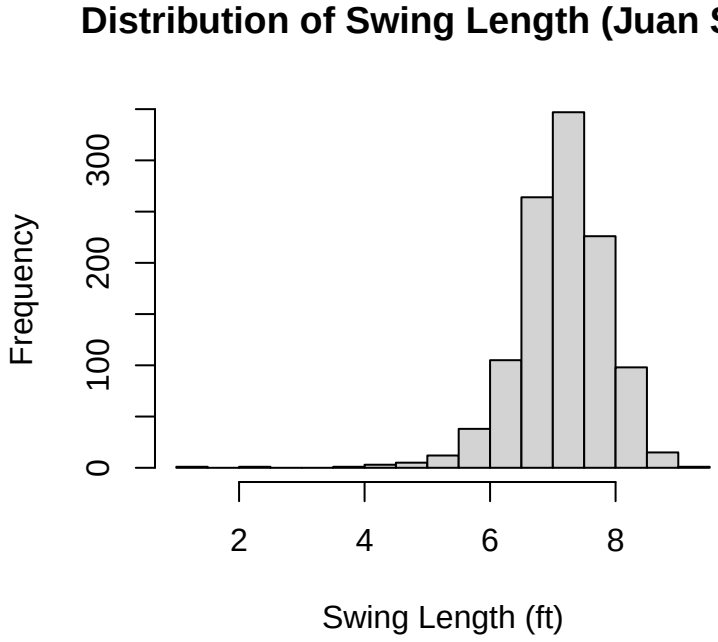
Appendix

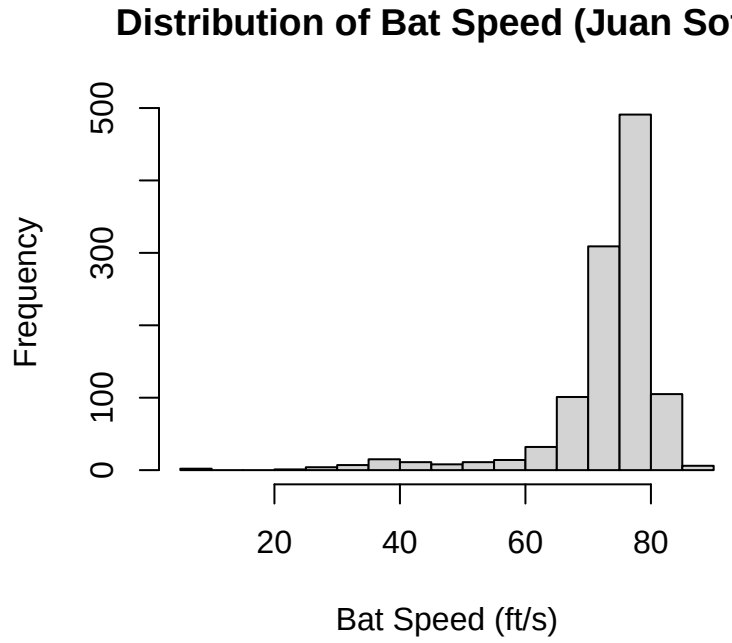
A.1 Sample Player Exploration: Juan Soto

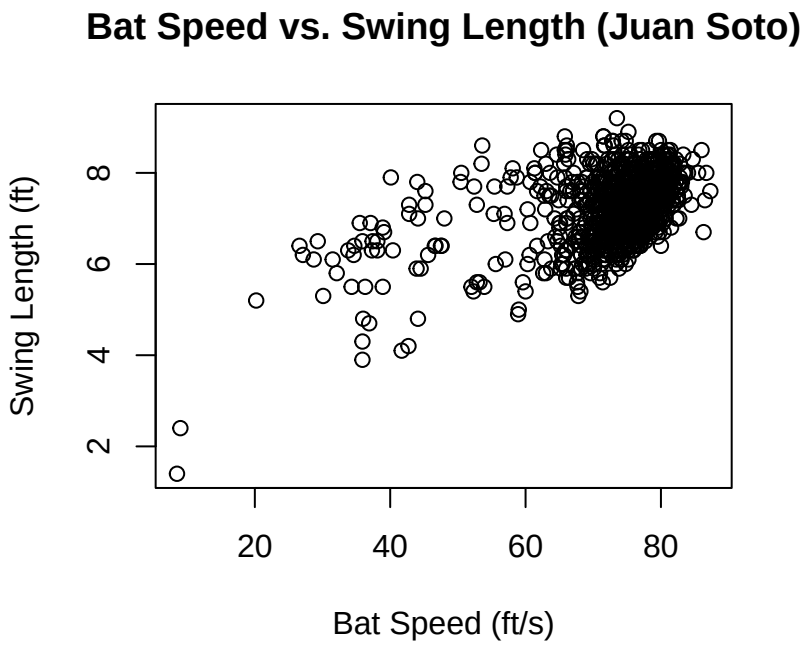
Data exploration was conducted on the entire dataset, which made it simple to make overarching assumptions that could be verified with external sources or domain knowledge. However, it failed to consider the large number of combinations of batters, pitchers, pitch types, and game situations that can impact the bat speed and swing length. Therefore, to see if the relationships observed for the whole dataset between swing length and bat speed held at a more granular level, we examined 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 an 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] 0.4796892
```







A.2 wOBA Model Refinement

Initially, simpler models were considered to predict wOBA using swing mechanics and game context variables. These models were built to explore the impact of different predictors on wOBA and to identify the most influential factors in offensive productivity. The models were evaluated based on their R-squared values and the significance of the predictors.

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

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-0.4926 -0.2578 -0.1270  0.1583  1.7565

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2704857  0.0121758  -22.21   <2e-16 ***
bat_speed     0.0119796  0.0001665   71.96   <2e-16 ***
swing_length -0.0388468  0.0014591  -26.62   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3645 on 101469 degrees of freedom
  (124314 observations deleted due to missingness)
Multiple R-squared:  0.04865,   Adjusted R-squared:  0.04863
F-statistic:  2594 on 2 and 101469 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
-0.5196 -0.2345 -0.1207  0.1429  1.8178

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -0.1059901  0.0131458  -8.063 7.54e-16 ***
bat_speed                0.0092784  0.0001737  53.429  < 2e-16 ***
swing_length            -0.0250548  0.0016213 -15.454  < 2e-16 ***
factor(balls)1           0.0071839  0.0028402   2.529 0.011428 *
factor(balls)2           0.0195176  0.0032411   6.022 1.73e-09 ***
factor(balls)3           0.0393384  0.0040497   9.714  < 2e-16 ***
factor(strikes)1        -0.0141549  0.0032796  -4.316 1.59e-05 ***
factor(strikes)2        -0.1583996  0.0031684 -49.993  < 2e-16 ***
factor(outs_when_up)1   -0.0078518  0.0027297  -2.876 0.004023 **
factor(outs_when_up)2   -0.0127032  0.0027503  -4.619 3.86e-06 ***
pitch_categoryCurveball  0.0134406  0.0045423   2.959 0.003087 **
pitch_categoryFastball   0.0098466  0.0028365   3.471 0.000518 ***
pitch_categoryOffspeed  -0.0058789  0.0035519  -1.655 0.097898 .
pitch_categoryOther      0.0086938  0.0226729   0.383 0.701389
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3576 on 101458 degrees of freedom
  (124314 observations deleted due to missingness)
Multiple R-squared:  0.08441,   Adjusted R-squared:  0.08429
F-statistic: 719.5 on 13 and 101458 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), data = train)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.5173 -0.2346 -0.1211  0.1430  1.8254

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           -0.0895613  0.0125236  -7.151 8.65e-13 ***
bat_speed              0.0093995  0.0001692  55.552  < 2e-16 ***
swing_length          -0.0277190  0.0014437 -19.200  < 2e-16 ***
factor(balls)1         0.0070394  0.0028399   2.479  0.01319 *
factor(balls)2         0.0198678  0.0032369   6.138 8.39e-10 ***
factor(balls)3         0.0406386  0.0040235  10.100  < 2e-16 ***
factor(strikes)1      -0.0155578  0.0032623  -4.769 1.85e-06 ***
factor(strikes)2      -0.1596796  0.0031396 -50.860  < 2e-16 ***
factor(outs_when_up)1 -0.0080600  0.0027297  -2.953  0.00315 **
factor(outs_when_up)2 -0.0127540  0.0027504  -4.637 3.54e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3577 on 101462 degrees of freedom
  (124314 observations deleted due to missingness)
Multiple R-squared:  0.08414,   Adjusted R-squared:  0.08406
F-statistic:  1036 on 9 and 101462 DF,  p-value: < 2.2e-16
```

The results indicated that in model 1 (bat_speed and swing_length only), the R-squared value of 4.86% 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 8.4%. 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.

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 had a slight decrease in R-squared ($R^2 = 0.08414$) which 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 initial 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.

Lastly, we also examined `p_throws` and `stand`, which represent the pitcher's throwing hand and the batter's stance, respectively. Batter-pitcher matchups can influence the outcome of the at-bat, and these variables could provide insights into batter-pitcher interactions. However, in model 6, we found that neither of these variables were significant predictors of wOBA. Despite this model having a comparable R-squared value to model 5, the lack of significance for these variables suggests that other factors may be more influential in determining offensive productivity, and we decided to examine model 5 as our most comprehensive model in the main section of this report.

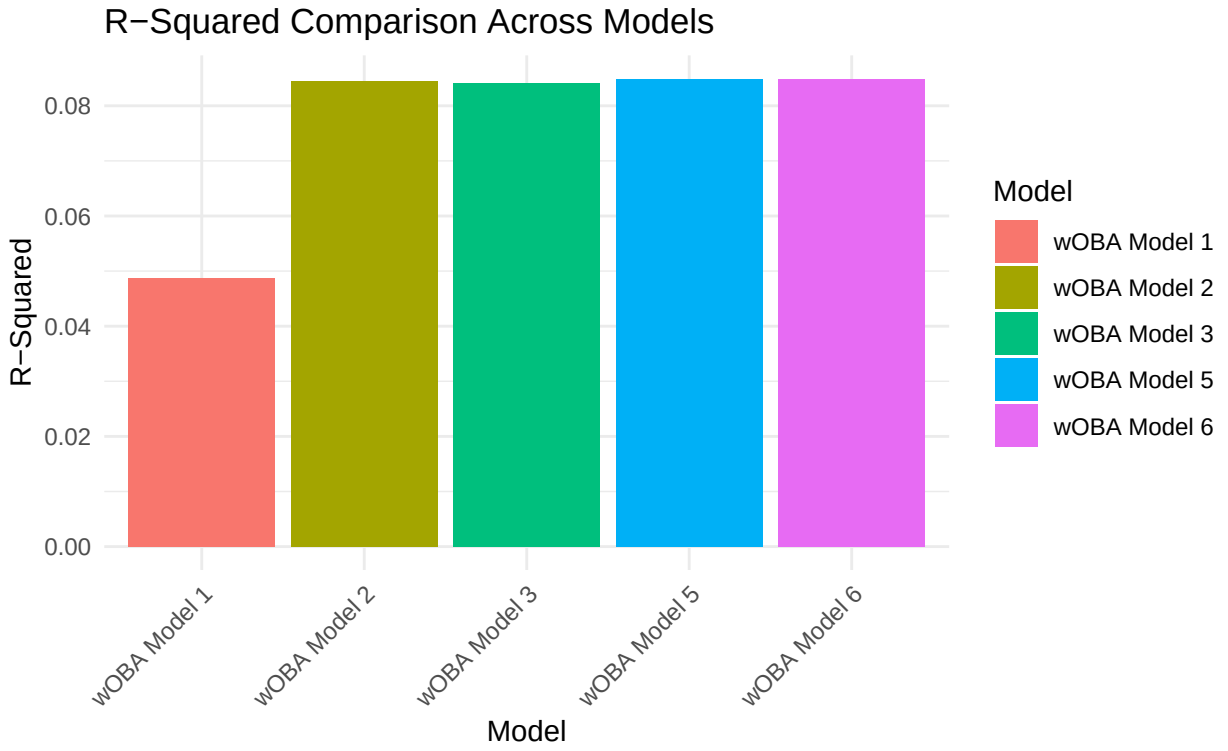
```
Call:
lm(formula = estimated_woba_using_speedangle ~ bat_speed + swing_length +
    factor(balls) + factor(strikes) + factor(outs_when_up) +
    pitch_category + release_pos_x + release_pos_z + release_speed +
    p_throws + stand, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5477 -0.2341 -0.1208  0.1434  1.8290

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.0455944  0.0324309   1.406  0.15976
bat_speed                0.0093560  0.0001742  53.723  < 2e-16 ***
swing_length            -0.0262733  0.0016481 -15.942  < 2e-16 ***
factor(balls)1           0.0072724  0.0028397   2.561  0.01044 *
factor(balls)2           0.0196181  0.0032411   6.053 1.43e-09 ***
factor(balls)3           0.0394408  0.0040508   9.736  < 2e-16 ***
factor(strikes)1        -0.0138729  0.0032794  -4.230 2.34e-05 ***
factor(strikes)2        -0.1569905  0.0031743 -49.457  < 2e-16 ***
factor(outs_when_up)1   -0.0070460  0.0027317  -2.579  0.00990 **
factor(outs_when_up)2   -0.0115765  0.0027544  -4.203 2.64e-05 ***
pitch_categoryCurveball -0.0005144  0.0049837  -0.103  0.91779
pitch_categoryFastball   0.0274786  0.0038435   7.149 8.77e-13 ***
pitch_categoryOffspeed  -0.0059430  0.0035797  -1.660  0.09687 .
pitch_categoryOther     -0.0422819  0.0237919  -1.777  0.07555 .
release_pos_x            0.0020305  0.0015518   1.308  0.19073
release_pos_z            0.0062172  0.0021613   2.877  0.00402 **
release_speed           -0.0022581  0.0003374  -6.693 2.20e-11 ***
p_throwsR                0.0162830  0.0065790   2.475  0.01333 *
standR                  -0.0023962  0.0023434  -1.023  0.30652
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3575 on 101453 degrees of freedom
  (124314 observations deleted due to missingness)
Multiple R-squared:  0.0849,    Adjusted R-squared:  0.08474
F-statistic: 522.9 on 18 and 101453 DF,  p-value: < 2.2e-16
```

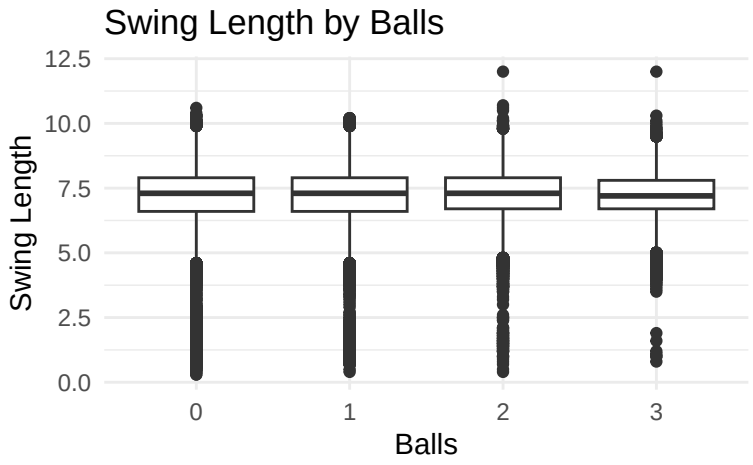
To compare the performance of our wOBA models, we used a series of plots that show the R-squared values of each model. This helps identify which model provides the best fit and predictive accuracy.

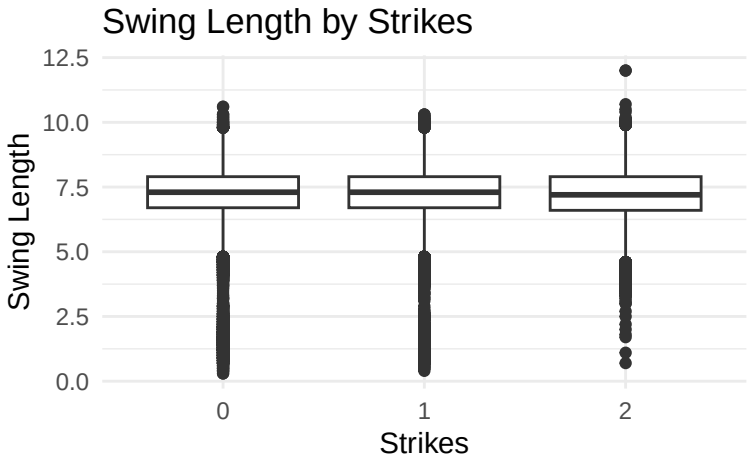


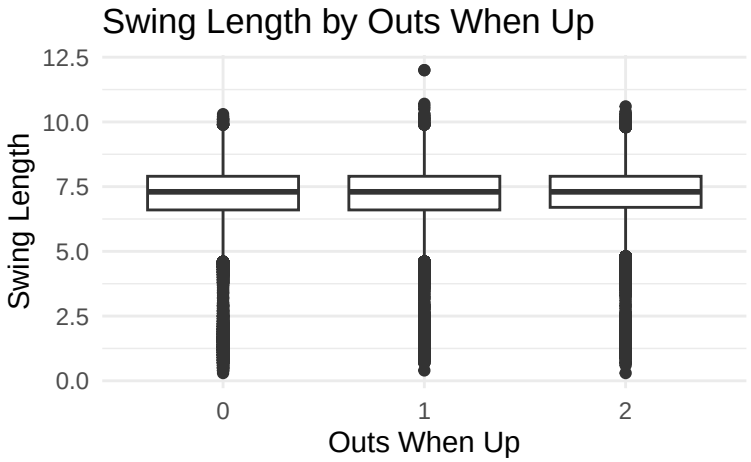
This bar plot visually demonstrates the relative explanatory power of each model. While model 5 shows the highest R-squared values, it's important to note that this measure alone does not capture the complete picture— model interpretability and practical application are also key factors in selecting the best model.

A.3 Additional Visualizations

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 created bar plots showing the average swing length and bat speed for different categories of these game-state variables.

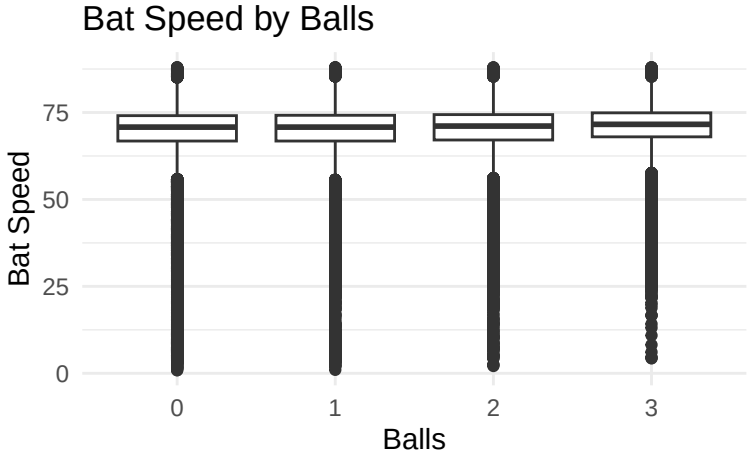


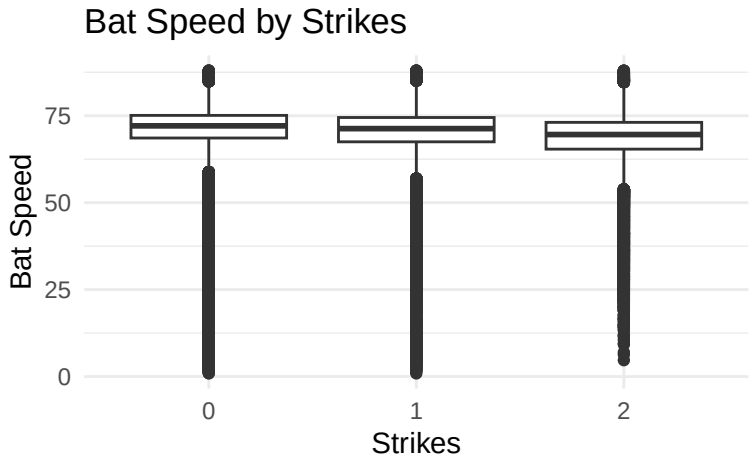


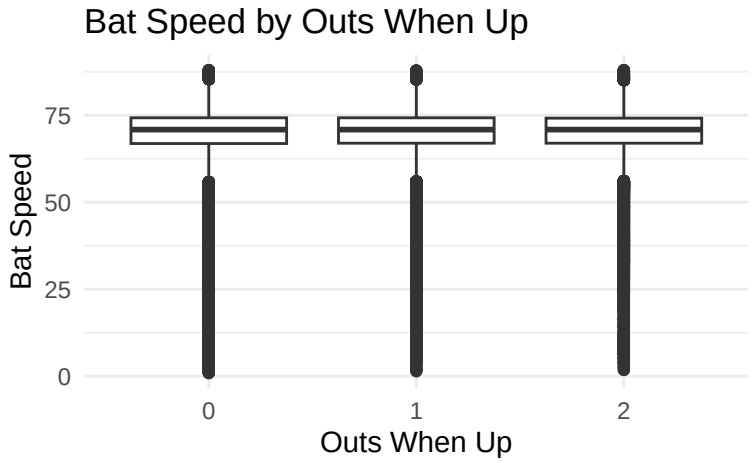


These visualizations help highlight how the swing length changes depending on the count and the number of outs. For example, we observed that batters tend to take longer swings with a higher number of balls in the count, likely in anticipation of a more favorable pitch. On the other hand, batters with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

Similarly, we visualized 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.

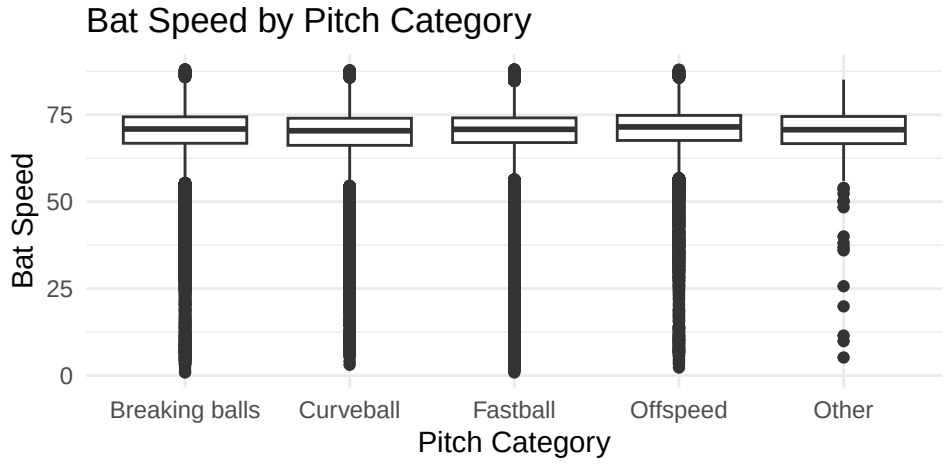


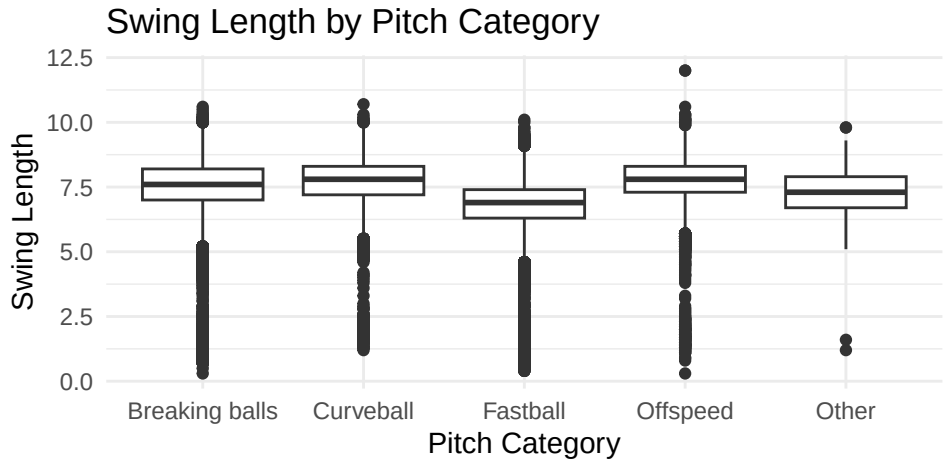




These box plots illustrate how bat speed varies under different game conditions. We found that bat speed tends to be higher when there are fewer outs, as batters may be more aggressive in those situations. The effect of the count (balls and strikes) may be more subtle, with batters perhaps swinging harder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

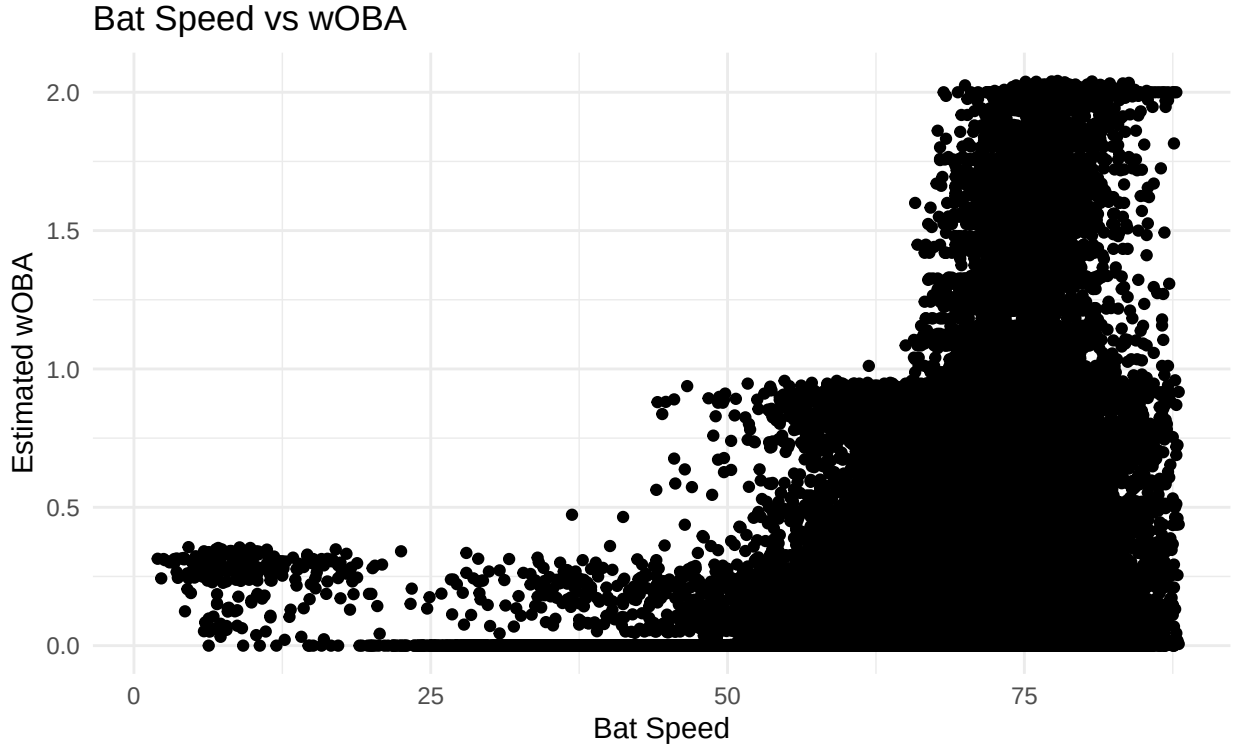
Pitch Category Impact on Swing Length and Bat Speed The pitch type also plays a significant role in influencing swing mechanics and performance. To illustrate how different pitch categories affect bat speed and swing length, we used box plots to show the distribution of bat speed and swing length by pitch category.

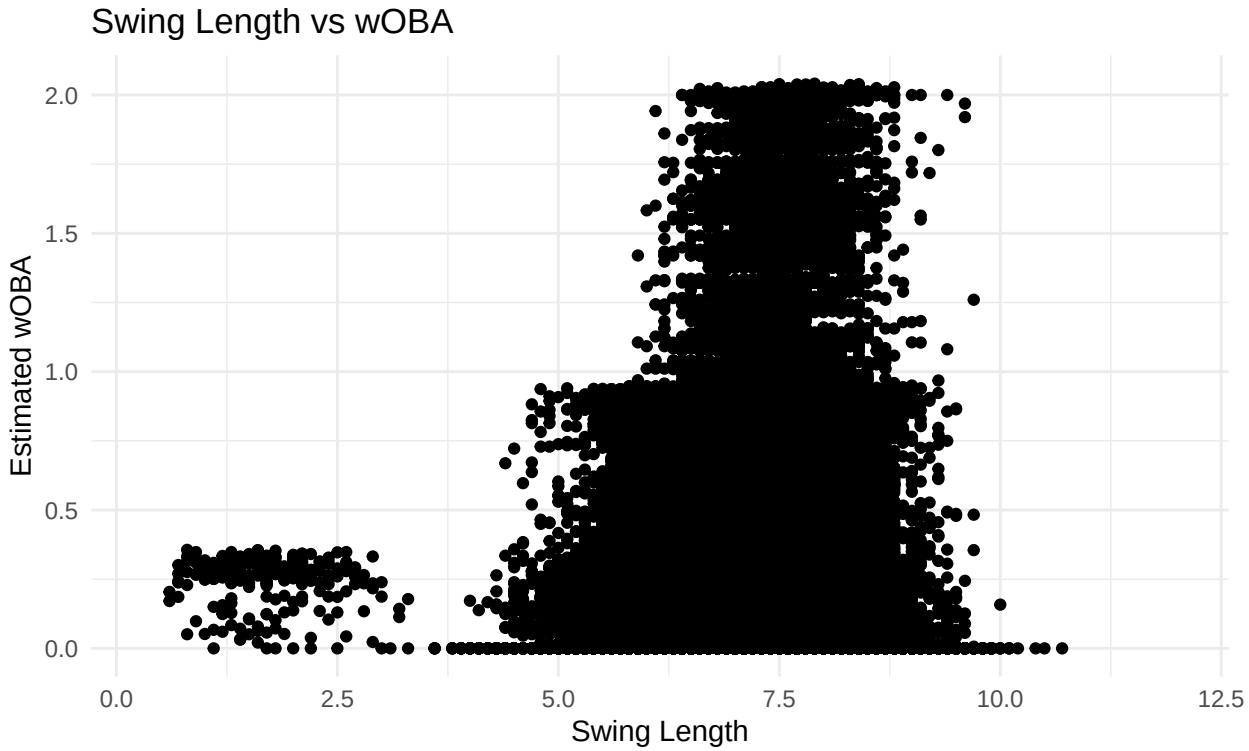




These plots help visually assess the impact of different pitch types. For example, we found that bat speed is higher for fastballs compared to breaking balls, while swing length may vary more for certain types of pitches, like off-speed pitches.

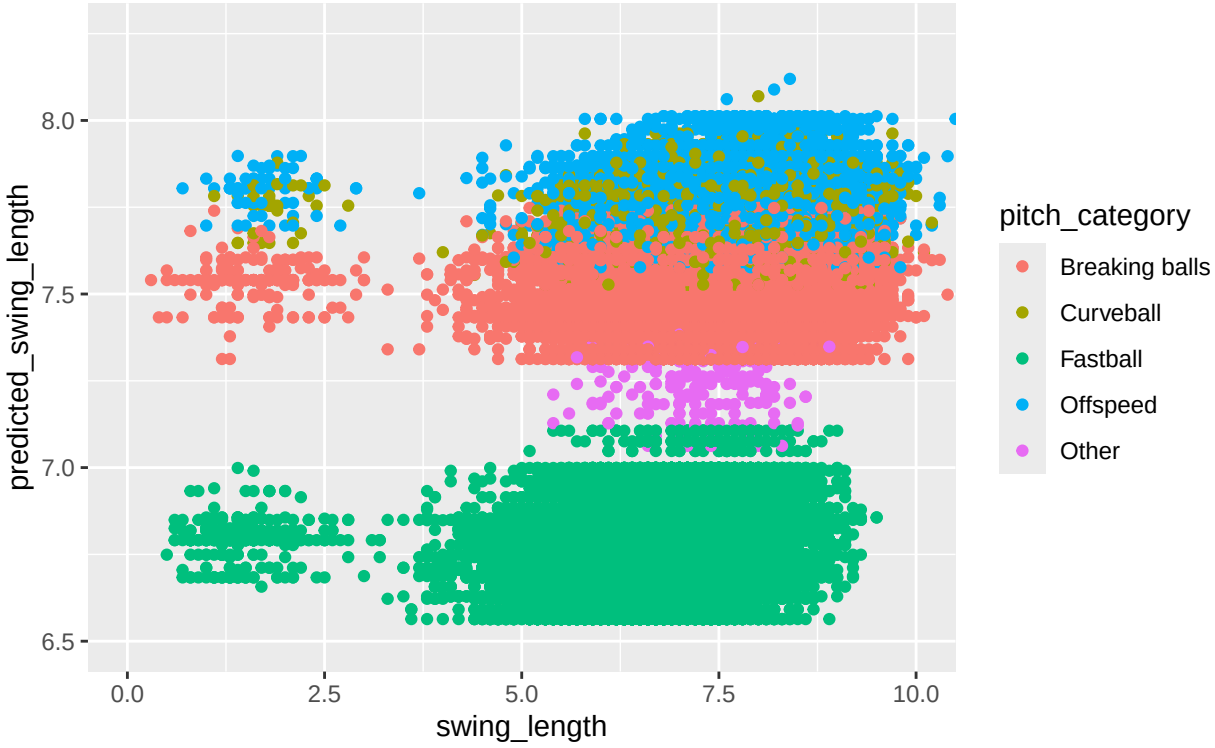
Interpreting the Model Results: Swing Length and Bat Speed Impact on wOBA In the wOBA prediction models, we found that bat speed and swing length had significant relationships with estimated wOBA. However, the effect of swing length was somewhat counterintuitive— longer swings were associated with lower 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.

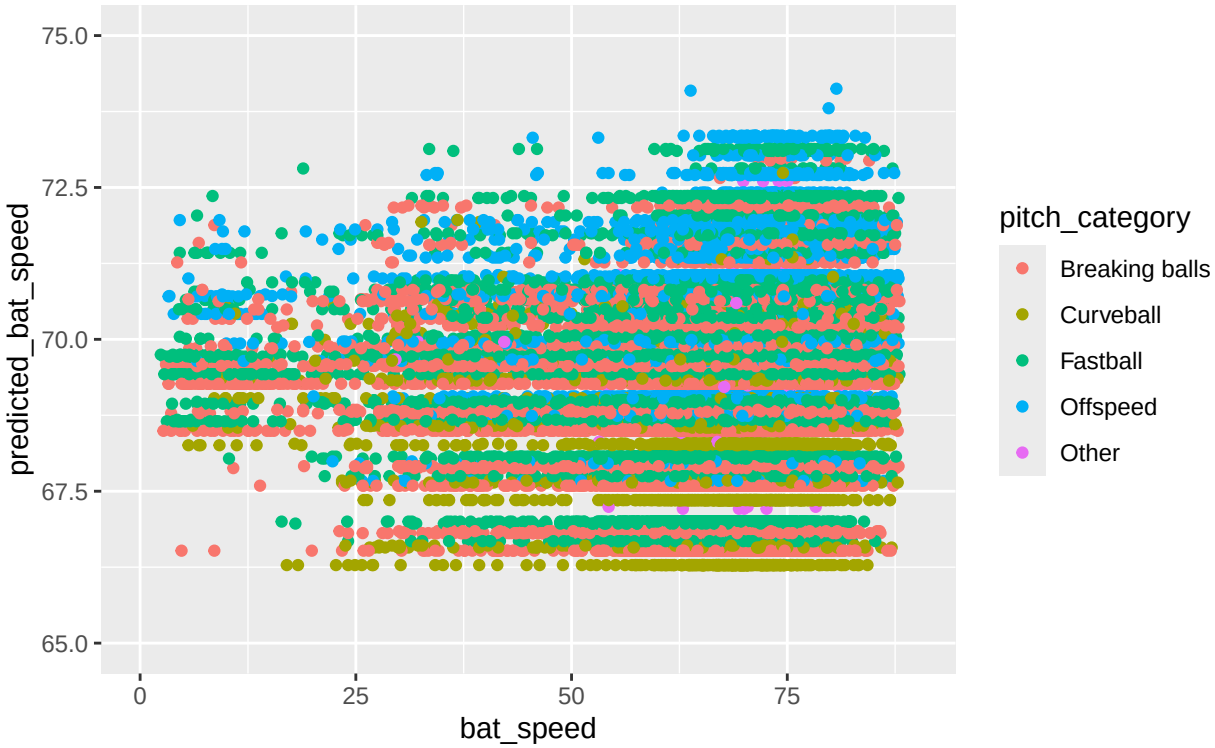




These scatter plots reveal the relationships between swing mechanics and wOBA. We might see a clearer positive correlation between bat speed and wOBA, while the relationship with swing length may appear weaker or even negative, providing further insight into how these two factors play different roles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

Comparing Model Performance: Predictive Power To assess how well the first two models predict swing length and bat speed, respectively, we plotted 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.