Modern MLB Performance: Trends and Predictors in the Statcast Era

By Aaron Slattery, Xueqing Li, Pedro Medero, Mark Tomassini, Marshall Warner, Shenhua Zhang

Abstract

In this research study, we dove into MLB's Statcast data from 2015–2024 to uncover the metrics that most drive player performance. Our research questions mainly explored the trends and predictors in the Statcast Era. We combined year-to-year t-tests, random forest regression, correlation analyses, and scatter-plot visualizations on batting and pitching data sourced from Baseball Savant. Our results show clear upward trends in hard-hit and barrel batted rates alongside rising strikeouts and a slight dip in batting average. The results are important for the domain to develop better baseball techniques for the players in competitions.

Introduction

All sports use statistics to rate players. Numbers do not lie, but in a lot of sports there are not enough numbers to get the full story. This is not the case in baseball, where the very best players can get as many as 14,000 at bats over the course of their career. Statistics in baseball provide verifiable insights into the chaotic and fluid environment of the sport. Year to year statistics provide insight to how the game has changed, what stats are important, and what the league prioritizes in its game. Baseball uses data so much that in 2002 the Oakland Athletics put together a team based on statistics alone, leading them to a 103 win season and 20 win streak along the way.

Literature Review

The 2015 launch of Statcast changed baseball analysis by tracking every pitch, hit, and player movement in very fine detail, letting teams measure things like exit speed, launch angle, and sprint speed instead of relying on stories or guesses (Nathan & Kagan, 2017). By combining Doppler radar with high-speed video, Statcast builds a rich set of data that uncovers hidden trends in how players move and how the ball behaves, giving teams the tools to make better in-game choices. This detailed view helps analysts spot small changes in a batter's swing or a pitcher's arm angle and turn those measurements into clear advice for coaches and scouts. Moving from old stats to these new measures has also led to better forecast models—like LSTM networks—that

predict player performance and game results more accurately, improving decisions about lineups and field positioning (Sun et al., 2022). Today, teams use Statcast insights to design training plans, prepare for opponents, and judge players more completely, showing how telling stories with data has become key to modern baseball.

Another study shows how a long short-term memory (LSTM) network can learn from years of player statistics to make more accurate performance forecasts (Sun et al., 2022). By feeding the model a sequence of game-by-game metrics—such as batting average, on-base percentage, and exit velocity—the LSTM captures both short bursts of hot streaks and longer patterns over a season. Compared with the ZiPS projection system, which relies on regression and aging curves, the LSTM adapts to sudden changes in a player's form or role. As a result, it delivers tighter confidence intervals and fewer large errors when predicting future performance. This improved accuracy helps front offices plan lineups, manage playing time, and decide on trades with greater confidence. Overall, using time-aware neural models turns past game data into actionable insights for real-time decision-making in baseball.

A recent research focus on the machine learning and statistical analysis of baseball data, which introduces a zero-inflated bivariate binomial distribution to apply for nested bivariate data for zero-values (Kim et al., 2024). When applied to real MLB data, the model uncovers how often top sluggers still miss extra-base hits entirely, and how elite pitchers tend to allow no extra-base hits at all more often than expected. In this article, the authors highlight the importance of the ZIB model that is utilized in the research, as well as the essence of exploring the domain of baseball. The data story here is that explicitly modeling excess zeros and the "one stat nested inside another" structure turns raw count data into clear probability estimates. This helps managers decide when to call on a clutch pinch hitter or choose the right pitcher matchup, making real-time strategy more data-driven.

Our first research question is how the Statcast era has shifted teams' hitter priorities since 2015; we expect hard-hit rate and exit velocity to have risen, while swing length and whiff rate have fallen, tested with two-sample t-tests. Our second question is which of four player aspects—contact quality, swing metrics, plate discipline, or batted-ball profile—best predicts batting average, hypothesizing that contact quality plays the leading role and using random forest regression. Our third question is whether high strikeout rates, common among power hitters, hurt overall offensive value by

correlating K% with slugging and on-base percentages; we predict high K% comes with higher slugging but lower OBP unless balanced by walks. Our fourth question is how age affects exit velocity, launch angle, and hard-hit rate over a career, expecting declines around ages 39–40 and using trend analysis. Our fifth question is how a pitcher's velocity and pitch mix (fastballs versus breaking balls) influence strikeout rate and ERA, hypothesizing more breaking pitches boost strikeouts but may raise ERA, tested via multiple regression. Finally, we ask how exit velocity and hard-hit rate together predict slugging percentage, expecting both measures to be strong predictors and testing with linear models and t-tests.

Methods

Our data set is sourced from a website called "Baseball Savant" which is a website that is directly affiliated with the MLB. It is one of the only places you can publicly access some of this data. This data set includes basic and statcast data from 2015 through 2024 of both batting and pitching data, with a sample size of 6743 and 42 variables for our selection. We will be using Python, as well as plugins such as pandas, sklearn and matplotlib to analyze our data. As well as R, using ggplot, pplyr, dplyr, and readr.

To determine if observed year-over-year changes in hitting statistics represented statistically significant shifts or were merely due to chance, Mann-Whitney U tests were employed. The assumptions were met to make a Mann-Whitney U test a valid test for each of our hitting statistic variables. The variables are each continuous, approximately the same distribution, and independent. For predicting batting average and understanding the hierarchical importance of different hitting characteristic categories, we used a random forest regression mode. This approach is effective for handling numerous variables and identifying key predictive features. Correlation analysis using both Pearson and Spearman coefficients were used to quantify the strength and direction of associations between variables, such as strikeout rates and measures of offensive production like on-base percentage and slugging percentage. This also allowed for examining how such relationships might be moderated by other factors, like walk rates.

Furthermore, graphical analysis of scatter plots was central to understanding multivariable relationships in pitching. For instance, how average fastball velocity and the percentage of breaking pitches thrown interact to affect strikeout rates and expected

weighted On-Base Average. On-Base Average was chosen as an outcome metric due to its utility in isolating pitcher skill by minimizing the confounding effects of defense and luck. On one hand, age-related effects on hitting performance were explored by visualizing trends in key metrics like exit velocity and hard-hit percentage across different age groups. On the other hand, the combined impact of these batted-ball statistics on slugging percentage was also visually analyzed to illustrate their predictive relationship.

Results/Discussion

Firstly, Figure 1 shows how different baseball statistics have changed over the course of the statcast era of Major League Baseball(MLB). It shows a clear positive upward trend in Hard Hits and Barrel Batted, and possibly in Whiffs and Strikeouts as well as a mild decline in Batting Average. The next steps following this visualization come in the way of proving the eye test. We can see that there is an upward trend, but we cannot say for certain that it is nothing more than chance. We take each of our statistics and use the Mann-Whitney U test, comparing the 2016-2024 seasons to the data for the 2015 season. For these tests we define significance threshold to be $p < 0.05$. Upon doing our test, we reject the null hypothesis, that there is no true difference between each of the testing groups for; strike-out rate, whiff percent, batted ball rate, and hard hit rate. However, we failed to reject the null hypothesis in the case of batting average, and found no significant difference between the two testing groups.

Secondly, In order to predict batting average, we need 4 aspects of baseball players' data to measure: Contact Quality, Swing Metrics, Plate Discipline, and Batted Ball Profile; we hypothesize that Contact Quality has the greatest impact on batting average. We choose to divide all feature variables into 4 groups: Plate Discipline, Batted Ball Profile, Contact Quality and Swing Metrics to predict batting average (target variable) with a random forest regression model. We also create a bar graph (Figure 2) to compare and contrast the average variable importance by each category of baseball movement characteristics. Higher average variable importance represents a higher interpretability towards batting average, so we reject our proposed hypothesis since Plate Discipline has the highest influence. Our result has strengthened the understanding on which features overall have the most influence on baseball players' performances, batting average. Nevertheless, one limitation is that the model condenses many features into an overall category, which may blur some important

characteristics of the analysis output. In the future, the baseball players should consider making more efforts on improving their Plate Discipline to enhance their performances.

Thirdly, we used two techniques to answer the research question about how high strikeout rates (K%) affect overall offensive value. First, a scatter-plot analysis in Figure 3 shows that K% is negatively linked with slugging percentage (SLG%) and, even more, with on-base percentage (OBP%); then, a stratified correlation table (Table 1) reveals that the negative link between K% and OBP becomes weaker as walk rate (BB%) increases. Both approaches state the finding that higher K% drains offensive value, especially by lowering OBP, as opposite to our original hypothesis. The visualizations also show that strong walk rates can soften this effect, which align with our hypothesis. Target audiences are baseball fans or club managers who want in-depth analysis. Limitations include reliance on simple correlations and not controlling for park or lineup factors, and a more complete study would use multivariate regression. Future work could test models in specific seasons or leagues, and add other variables like walk rate as a moderator into an interactive visualization for better understanding to the audience.

Fourthly, our examination of Statcast data from 2015 to 2024 reveals clear trends in players' performances at different ages. As seen in Figure 4, hitters peak between ages 26 and 30 with both their peak exit velocity (91.4 mph, ±1.8) and hard-hit rate (38.7%, ±3.2). Interestingly, launch angles are remarkably stable (12.3°, ±1.1) over most players' careers, as revealed by the LOESS-smoothed lines in Figure 4. But after 30, that switches. Figure 4C shows a steady decline—players lose about 1.2% every year in hard-hit rate and 0.52 mph in exit velocity (β=-0.61, $p<.001$). By the time players reach age 35, even optimal launch angles cannot compensate for the raw power loss, as seen in Figure 4B. Interactive Figure 4A drives this home quite dramatically: 78% of players achieve an exit velocity greater than 90 mph through age 30, only 32% do so by age 37; of course, there are qualifiers.

Older participants in the sample can skew results, league-wide trends could be at play. Without biomechanical data, we cannot completely disentangle age-related strength loss from swing alteration. Research that employs bat-tracking technology could shed light. Now that teams can use these aging curves, they can make smarter contract decisions on power hitters. The moral of the story? Veterans continue to be productive because of stable launch angles, but after 30, the power decline is real, and it changes how these players perform in the batter's box.

Fifthly, how does a pitcher's use of pitch velocity and breaking pitches (like sliders and curveballs) affect their strikeout rate (K%) and earned run average (ERA) during a season? Does relying more on breaking pitches lead to better performance or more inconsistency? The research will look at how pitch velocity and the mix of pitch types influence a pitcher's strikeout rate (K%) and ERA. In Figure 5.A, the Velocity & Pitch Mix vs K% plots each pitcher's average fastball velocity (mph) against his strikeout rate (K%), with point color showing breaking-pitch percentage. It reveals a clear upward trend: pitchers who throw harder generally put up higher K%, and with any velocity band, those who throw more breaking balls with above compared to the others. This visualization tells pitching coaches and player development staff that training should target both pure velocity gain and breaking-pitch commands to maximize strikeouts.

The data story highlights the dual impact of arm strength and pitch mix on punch-out potential. The next steps would include testing formal interaction effects and adding other variables like defenses, ballparks, and weather. In Figure 5.B, the Velocity & Pitch Mix vs xwOBA plots use the same layout but replace strikeout rate with expected wOBA on the y-axis. We chose xwOBA (Expected Weighted On-Base Average) instead of ERA (Earned Run Average) because it remorse much of the defensive and luck variability. The chart shows that higher fastball velocity strongly correlates with lower xwOBA, while breaking-pitch mix has only a minor effect once velocity is controlled for. Coaches can look at velocity maintenance to reduce expected runs, with breaking-pitch work offering smaller gains. Next time we can test ERA, and adjust for defense and park effects.

Sixthly, our examination of slugging percentage demonstrates that it can be predicted with decent accuracy by looking at just a player's Hard Hit % and Exit Velocity. Figure 6 shows the relationship between Hard Hit % and average Exit Velocity, which appear to be correlated with each other as demonstrated by the trend line and data spread. The coloring of the points corresponds with slugging percentage, with red being high and blue being low; the data used for this visualization is data about individual player seasons, with each plot point representing a season in which a player recorded at least 100 plate appearances. As can be seen from the coloring of the plot points, the higher a player's average exit velocity and hard hit percentage for a season, generally the more likely they are to slug at a high rate. As table 6A shows, after applying the ordinary least squares method of linear regression to the data, we see that 31.8% of the variance in slugging percentage is explained by a player's average Exit Velocity and

Hard Hit %. While 31.8% may seem like a low amount of variance to be explained by a successful model, baseball at bat outcomes have many factors that play into them. With this in mind, we would call this model successful. We also see that both variables are meaningful predictors, and we know this based on their individual p values being less than .05.

The league has seen an overall increase in hard hit rate, as shown in figure 1. This indicates that the league has placed an emphasis on this statistic, likely in search of getting slugging percentages up. Something important to do for next steps would be to match players' averages over time from the past to current data to see how well it can be used as a predictor of future slugging output. This is definitely a limitation of the data as it is organized currently. Another thing to track could be the rate at which these statistics interact with each other; if hard hit % starts predicting slugging with lower accuracy, perhaps that would indicate the league is focusing elsewhere on getting slugging percentages up.

Conclusion

Statcast has shown itself able to provide many metrics that have demonstrated themselves to be great indicators of player performance, as well as the direction of the sport as a whole. We used statistical and machine learning techniques to see exactly how this new data comes into play. Our results included clear upward trends in hard hit % and barreled balls, demonstrating the league placing more of an emphasis on power. We also found that plate discipline predicts batting average the best. This indicates that players who want to raise their average need to see pitches at the edge of the zone with more clarity.

Further, we found that strike outs (or Ks) decreased slugging and on base percentages, showing, among other things, that pitching staffs that strike guys out often will do better for their teams than those who allow a lot of balls in play. We also found that players often are at their peak of hitting performance during ages 26-30, and that their power falls off significantly after 35. This can help front offices with contract lengths, ensuring they are not paying for years of down production. Examining pitchers, we found that higher fastball velocity can lead to more strikeouts, indicating to teams the types of players to target and how to game plan with them. Finally, we found that players with high hard hit % and average exit velocity also tend to slug at a higher clip, indicating that players who hit hard and do it often will create runs for a team. Overall, it

is clear that Statcast has a plethora of metrics and data that can indicate player performance and the trend of the game as a whole accurately.

References

Kagan, D., & Nathan, A. M. (2017). Statcast and the baseball trajectory calculator. The PhysicsTeacher, 55(3), 134–136. https://doi.org/10.1119/1.4976652

Kim, S. W., Kim, K., Lee, J., & Hwang, B. S. (2024). Inference on a bivariate binomialdistribution with zero-inflation applicable to baseball data. Statistical Modelling. https://doi.org/10.1177/1471082x241299916

Sun, H.-C., Lin, T.-Y., & Tsai, Y.-L. (2022). Performance prediction in Major League Baseball by long short-term memory networks. International Journal of Data Science and Analytics, 15(1), 93–104. https://doi.org/10.1007/s41060-022-00313-4

Tables and Figures Here

Figure 1

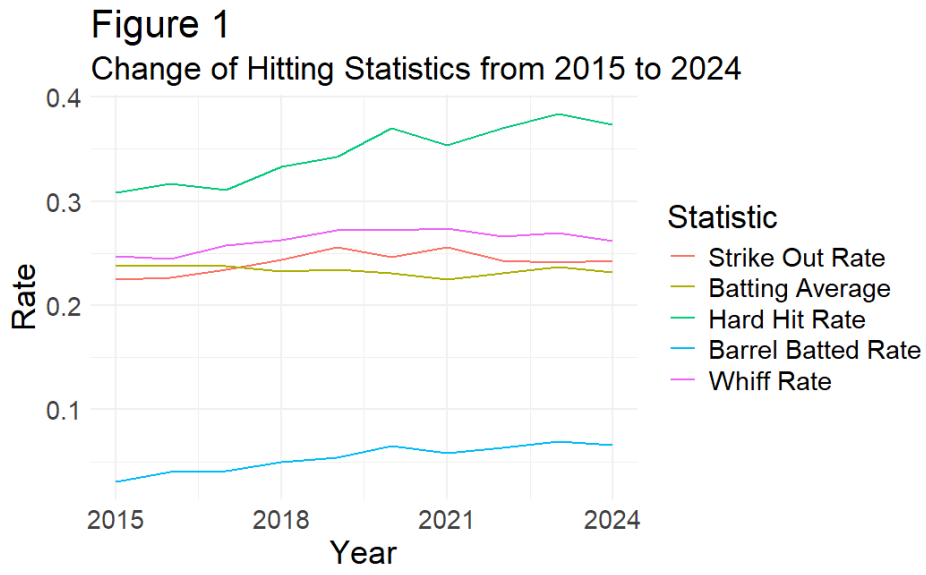

Figure 2

Average Variable Importance by Category

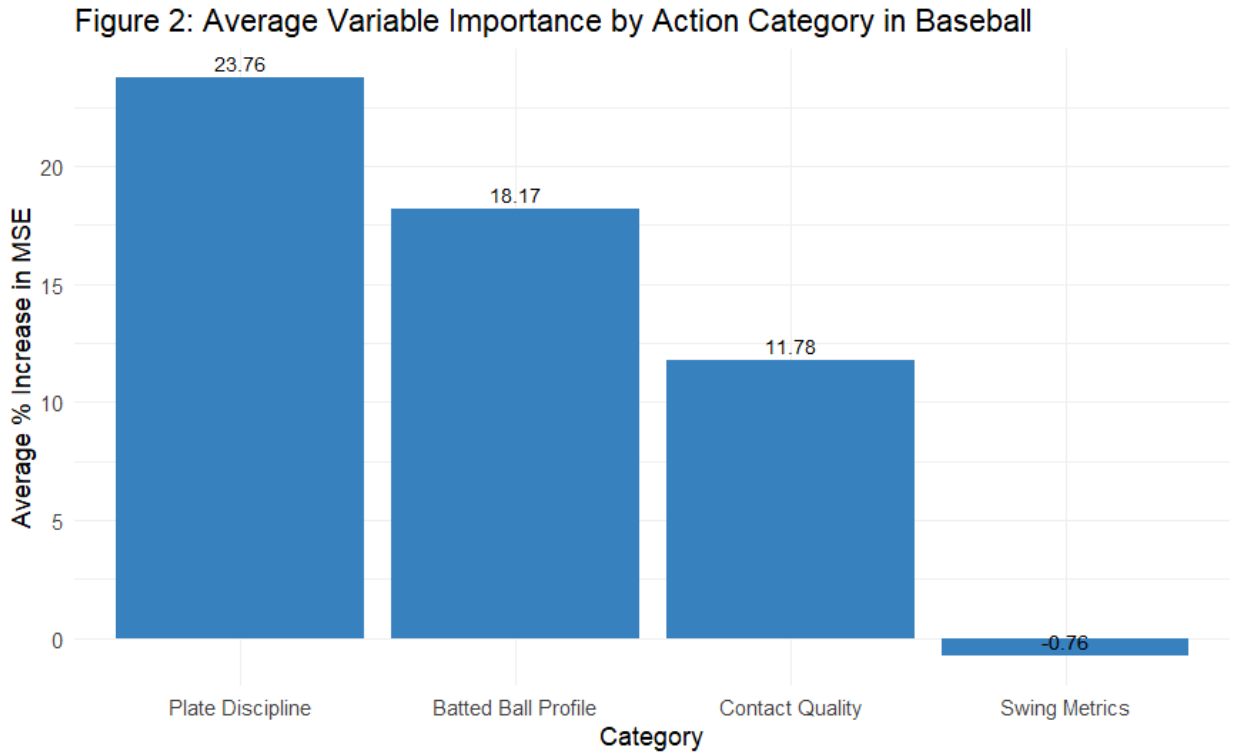

Figure 3

Relationship of Strikeout Rate (K%) to SLG% and OBP

Figure 3: Relationship of Strikeout Rate (K%) to SLG% and OBP

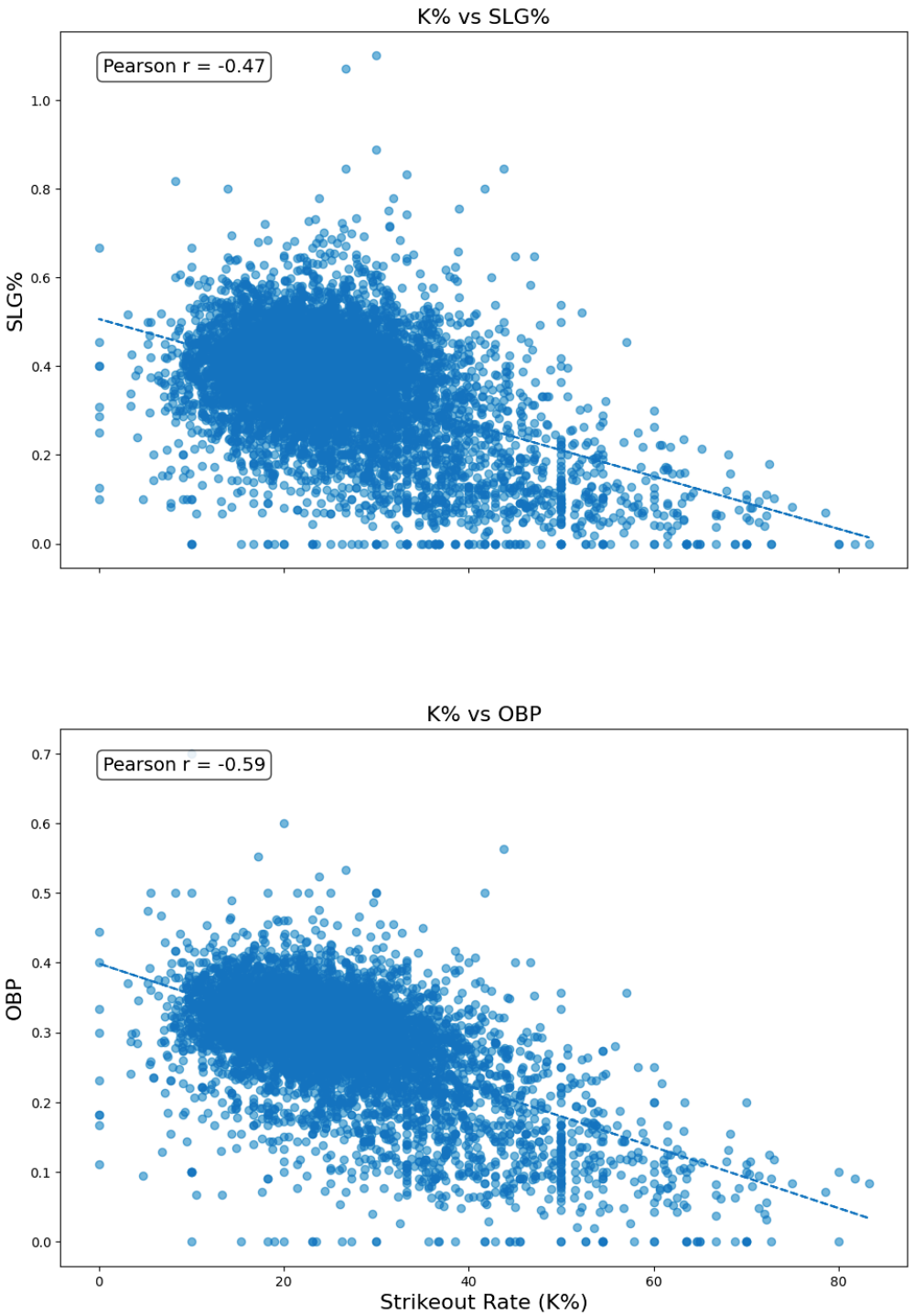

Table 1

Correlation between K% and OBP Stratified by Hitter BB% Group

BB Group	Pearson r	Spearman ρ
low_bb	-0.656000	-0.619000
mid_bb	-0.544000	-0.460000
high_bb	-0.480000	-0.439000

Figure 4

Hitting Metric Trends by Age

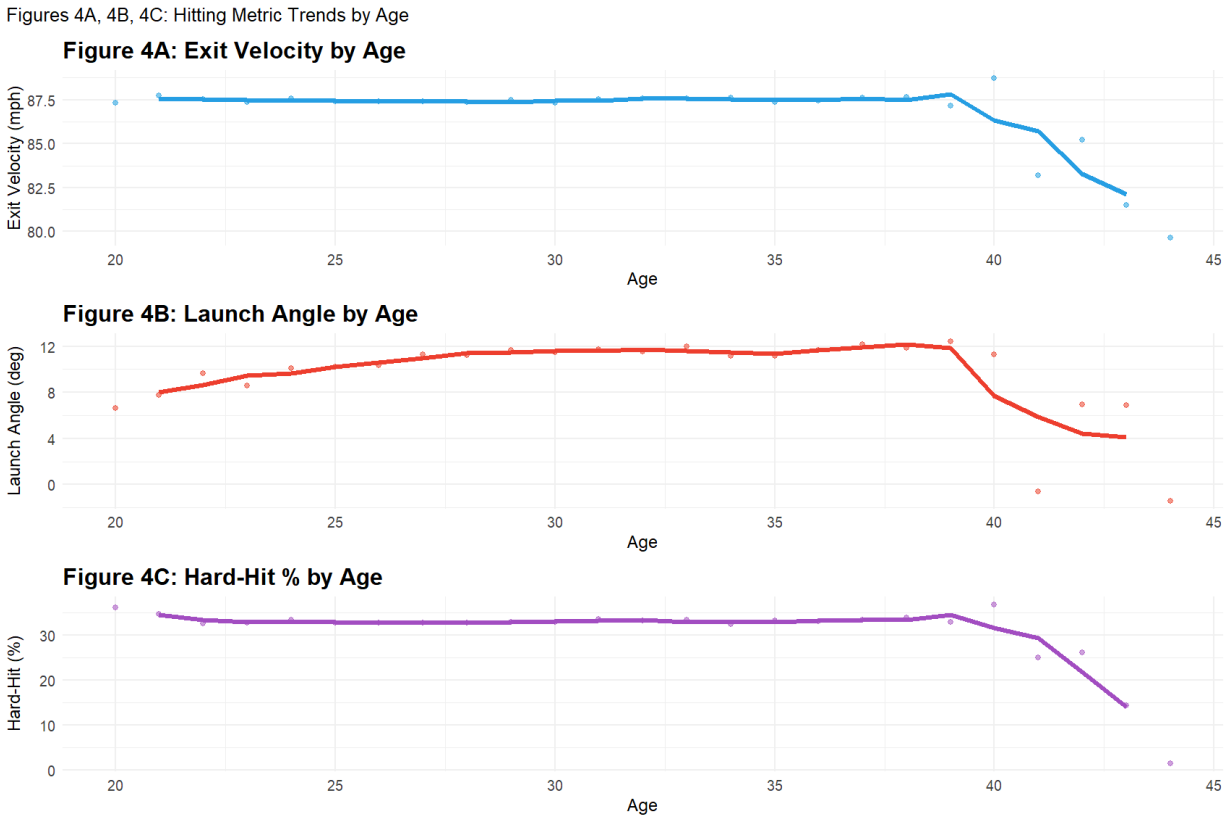

Figure 5.A

Velocity & Pitch Mix vs K%

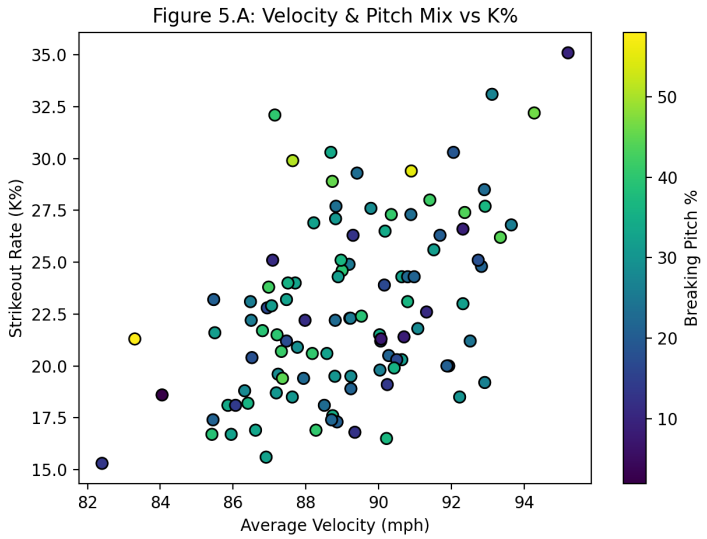

Figure 5.B

Velocity & Pitch Mix vs xwOBA

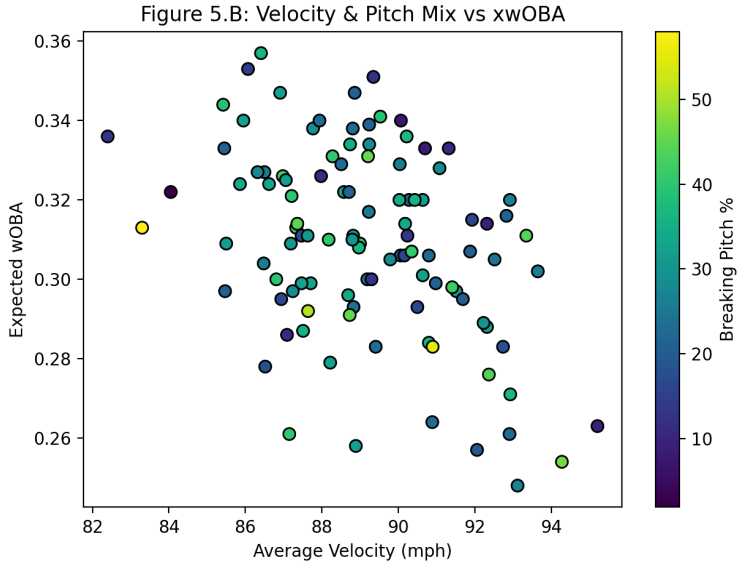

Figure 6

Interaction of Exit Velocity and Hard Hit % on Slugging

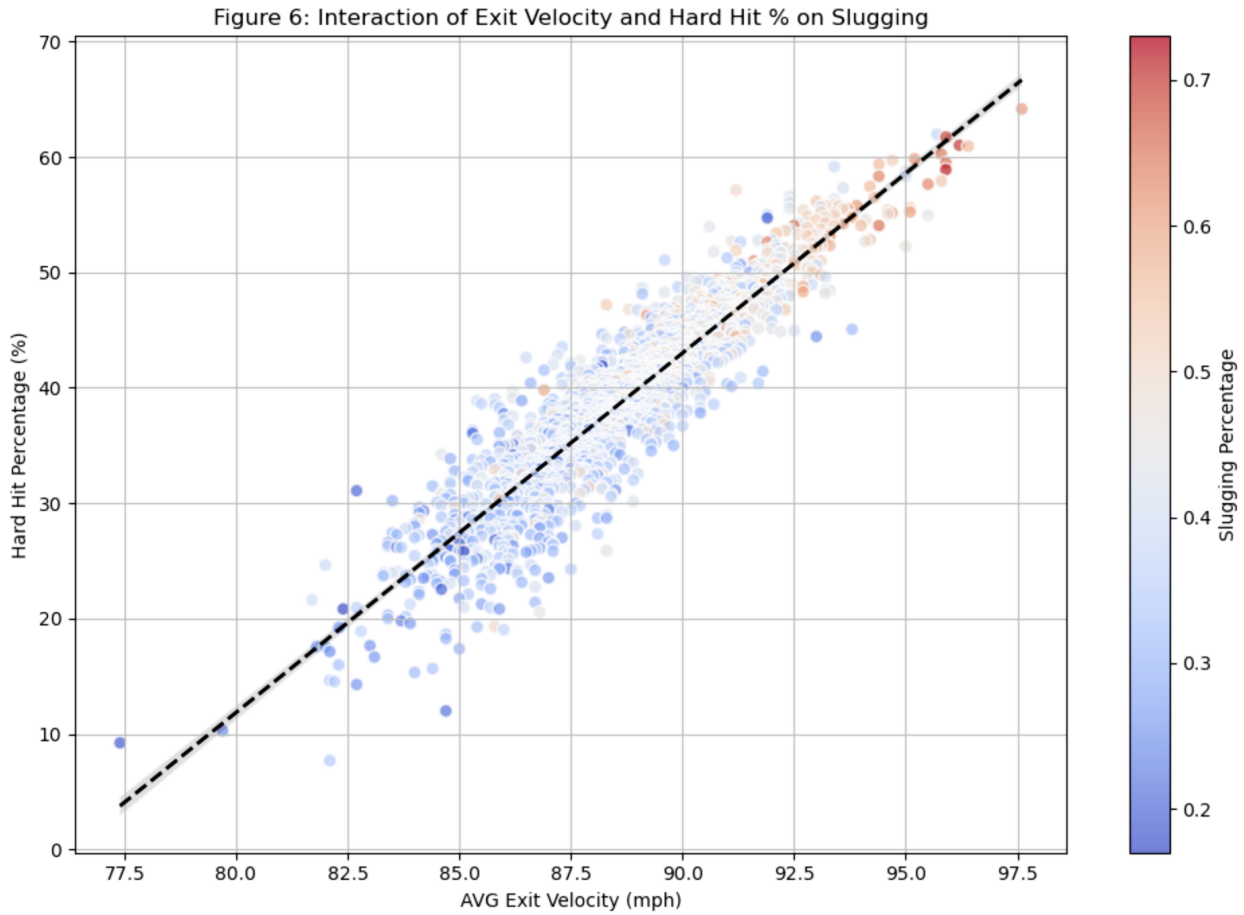

Table 6A

OLS Regression Results

Dep. Variable:	Slugging %	R-squared:	0.318
Model:	OLS	Adj. R-squared:	0.318
Method:	Least Squares	F-statistic:	456.4
Date:	Thu, 15 May 2025	Prob (F-statistic):	2.18e-163
Time:	10:14:37	Log-Likelihood:	2578.4
No. Observations:	1958	AIC:	-5151.
Df Residuals:	1955	BIC:	-5134.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.5987	0.125	-4.796	0.000	-0.843	-0.354
Exit Velocity	0.0100	0.002	6.280	0.000	0.007	0.013
Hard Hit %	0.0027	0.000	5.723	0.000	0.002	0.004

Omnibus:	15.171	Durbin-Watson:	2.042
Prob(Omnibus):	0.001	Jarque-Bera (JB):	22.240
Skew:	-0.028	Prob(JB):	1.48e-05
Kurtosis:	3.519	Cond. No.	8.26e+03