Final Project Report

Edison Won, Ryan Coombs, Jennifer Lin, Alejandro Clavijo

Introduction

Project Overview: Identifying Predictive Performance Patterns in NCAA Baseball

Introduction, Research Context, and Objectives

In modern baseball analytics, teams increasingly rely on granular, pitch-level and play-level data to guide drafting decisions, evaluate talent, and refine player development pathways. Yet at the collegiate level—where athletes exhibit wide variability in experience, coaching styles, and game environments—identifying which performance traits reliably project to professional success remains a complex and evolving challenge. Traditional "box score" statistics often obscure the contextual patterns within games that differentiate high-impact players from average performers. As a result, professional scouts and analytics departments are increasingly turning to more sophisticated models that capture not only what happened, but when and how it happened.

This project situates itself within this emerging analytical landscape by leveraging the complete 2021 NCAA Division I baseball play-by-play dataset to uncover performance patterns that may signal higher offensive value, competitive advantage, and potential relevance to MLB scouting criteria. Across more than 200,000 play events, this dataset provides a rare opportunity to examine situational outcomes, pitch-by-pitch leverage, inning-level dynamics, and momentum patterns that influence game flow. By transforming raw textual play descriptions into structured variables—such as play outcomes, ball-strike counts, expected offensive value (xOV), and inning-by-inning performance—we construct a framework capable of highlighting behaviors and tendencies that matter in real competitive contexts.

Research Framework

To guide this investigation, our analysis centers on three overarching questions that together bridge descriptive statistics, exploratory visualization, and the foundations of predictive reasoning:

1. **Contextual Performance Variation:**
   How do offensive outcomes change across game situations, innings, and leverage states? Understanding when teams generate value allows us to identify temporal or strategic trends, such as fast starts, late-game surges, or inning-specific vulnerabilities.

2. **Predictive Play Patterns:**
   Which play types, pitch sequences, or game states correlate most strongly with run production and sustained offensive momentum? By isolating patterns associated with high-value outcomes, we build insight into the underlying mechanisms that drive scoring and influence game trajectories.

3. **Draft Selection Indicators (The Original Objective):**
   Although our final dataset did not include draft outcomes, the project remains grounded in the spirit of identifying performance features that could translate into professional evaluation metrics. Our visualizations aim to illustrate how situational consistency, play efficiency, and offensive value might be used to differentiate between players in a scouting context.

Methodology and Data Foundation

Using tools from **R for Data Science**, we employed a full workflow of data extraction, transformation, visualization, and exploratory pattern analysis. Major methodological steps included:

- **Database querying** using DBI and RSQLite to extract team-specific play-by-play logs.
- **Systematic text parsing** of the descriptive play strings to classify outcomes (e.g., singles, strikeouts, walks, stolen bases).
- **Encoding situational leverage** using ball-strike counts, base states, inning contexts, and expected offensive value (xOV).
- **Creating high-granularity visualizations**, including histograms, density plots, heatmaps, and custom grid-based graphics.
- **Constructing an original, interactive Killer Plot** (The Momentum Diamond) using Shiny and grid primitives to explore offensive flow across teams.

Together, these steps demonstrate a rigorous application of data-wrangling principles, exploratory visualization techniques, and creative graphical design—all central goals of the R for Data Science curriculum.

Implications and Applications

While this project functions primarily as an academic exercise in data processing and visualization, its implications extend to real-world analytics and decision-making in baseball. By highlighting:

- How offenses behave across innings
- How leverage influences outcomes
- How different play types shape scoring potential
- How momentum develops and decays within games

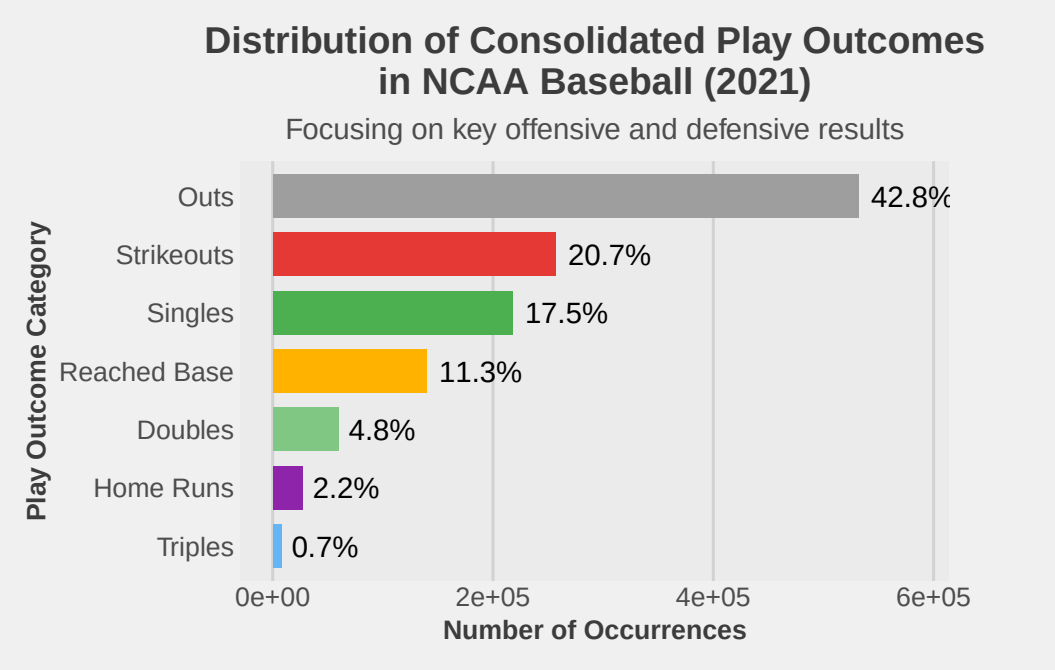
Our analysis contributes to a broader understanding of performance dynamics that can inform coaching strategies, player development priorities, and future predictive modeling efforts.

For organizations evaluating collegiate athletes, these insights reflect a shift toward context-aware analytics—measuring how players perform rather than simply what they produce. For player development staff, the emphasis on situational execution and offensive efficiency underscores the growing value of mental approach, decision-making, and adaptable skill sets.

Ultimately, this project demonstrates how detailed play-by-play data, when paired with modern visualization and modeling tools, can illuminate the subtle patterns that drive competitive success in NCAA baseball.

2021 NCAA Baseball Season Statistics and Visualization

Analytical Insights



The Fundamental Composition of NCAA Baseball

This visualization summarizes the frequency of key offensive and defensive play outcomes from the 2021 NCAA baseball season, consolidating granular events into higher-level categorical results. The data reveal striking asymmetries in outcome probabilities, indicating a competitive environment where defensive suppression of offense dominates the run-scoring process.

The most prevalent result, Outs, accounts for 42.8% of all plays. This near-majority share reinforces a foundational characteristic of baseball efficiency: the intrinsic difficulty of reaching base against collegiate-level pitching and fielding. The concentration of outs at this magnitude suggests that defensive control remains the primary determinant of game rhythm and offensive constraint. In strategic terms, this places premium value on actions that successfully avoid termination events—such as balls in play that advance or reach base—as they occur against a backdrop of inherent failure.

The second-largest category, Strikeouts (20.7%), reflects modern baseball's accelerating shift toward pitcher-dominated outcomes. A strikeout constitutes a terminal result without defensive variability, indicating that nearly one-fifth of all plays in the dataset bypass fielding altogether. This finding underscores the centrality of pitching strength to NCAA performance models: strikeout-heavy staffs limit batted-ball uncertainty and exert disproportionately high control over game outcomes.

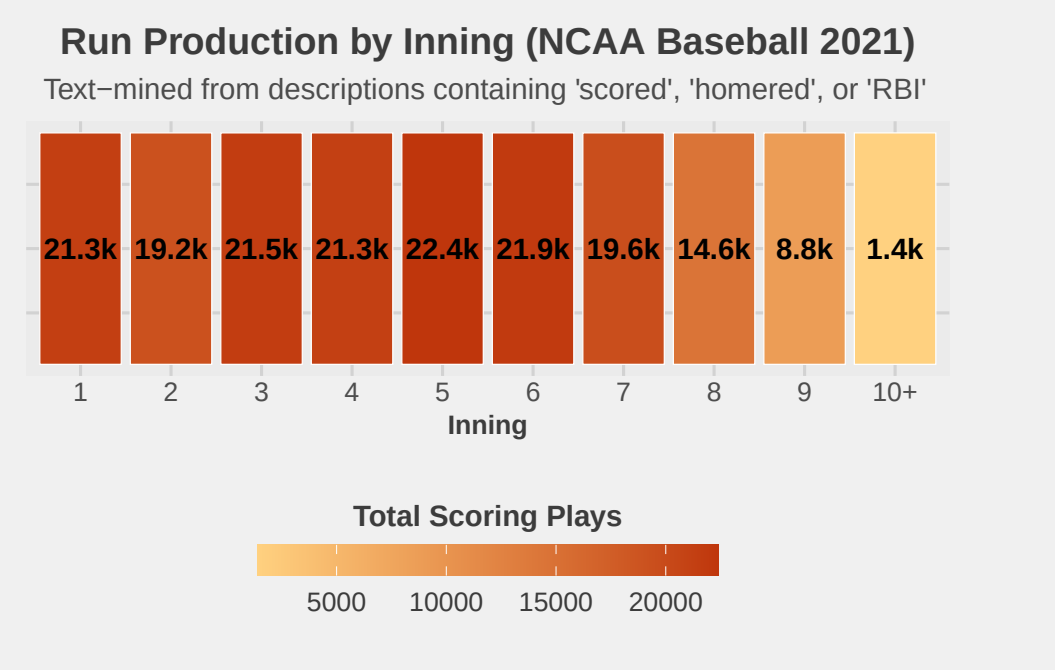
Offensive success events occur at significantly lower frequencies. Singles represent 17.5% of all events, establishing them as the most common form of positive offensive output. This reinforces their role as the backbone of NCAA offensive production: base-to-base advancement rather than high-volatility extra-base power. Reached Base outcomes (11.3%), including walks and hit-by-pitch, further illustrate that non-contact advancement remains an important offensive mechanism—though considerably less frequent than outs or strikeouts.

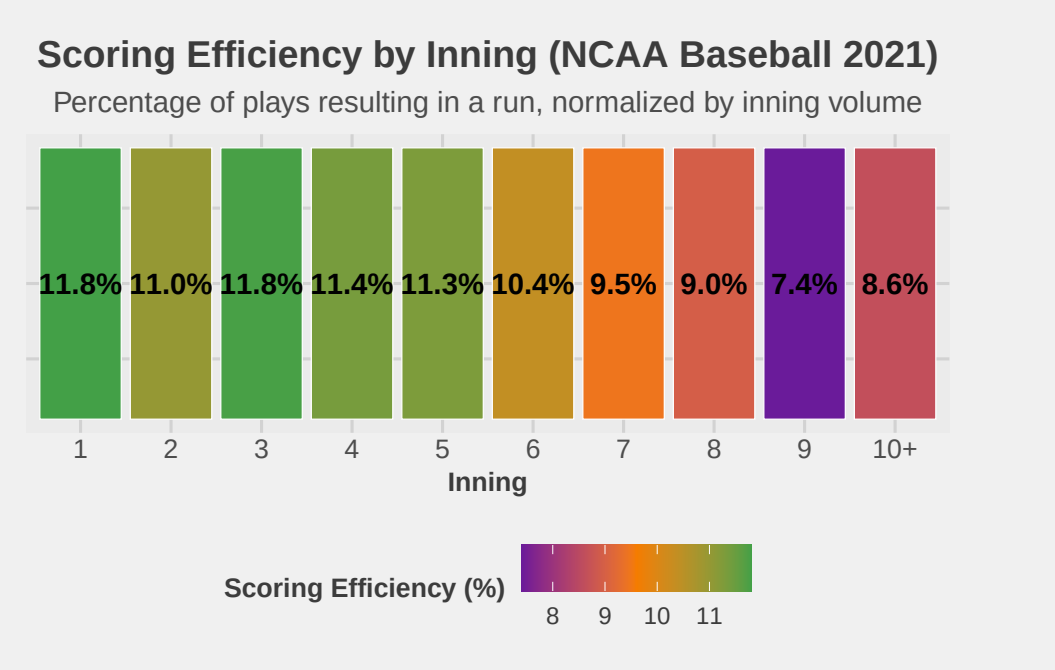
Extra-base hits—Doubles (4.8%), Triples (0.7%), and Home Runs (2.2%)—collectively demonstrate their rarity relative to routine play. These events are impactful but scarce, suggesting that while slugging outcomes can dramatically influence single-play scoring potential, they are insufficient as strategic anchors for sustained

run generation. Instead, the distribution supports an interpretation that NCAA baseball rewards cumulative sequencing: repeated access to first base, rather than episodic long-ball results.

Taken together, the distribution reveals a structural imbalance toward defensive success. Across nearly two-thirds of all plays, the batter fails to reach base. This informs a broader inference: offensive efficiency at the collegiate level is fundamentally probabilistic and adversarial. Success depends not on frequent high-damage contact, but on increasing marginal probabilities of non-out outcomes and sustaining base-advancement continuity.

In conclusion, this plot quantifies the essential reality of NCAA baseball: offensive production must operate within a high-failure ecosystem driven by elite pitching performance. Understanding these distributions helps contextualize follow-on analyses (xOV, inning momentum, state-level variation) by grounding them in the core statistical constraints of the game's underlying event structure.



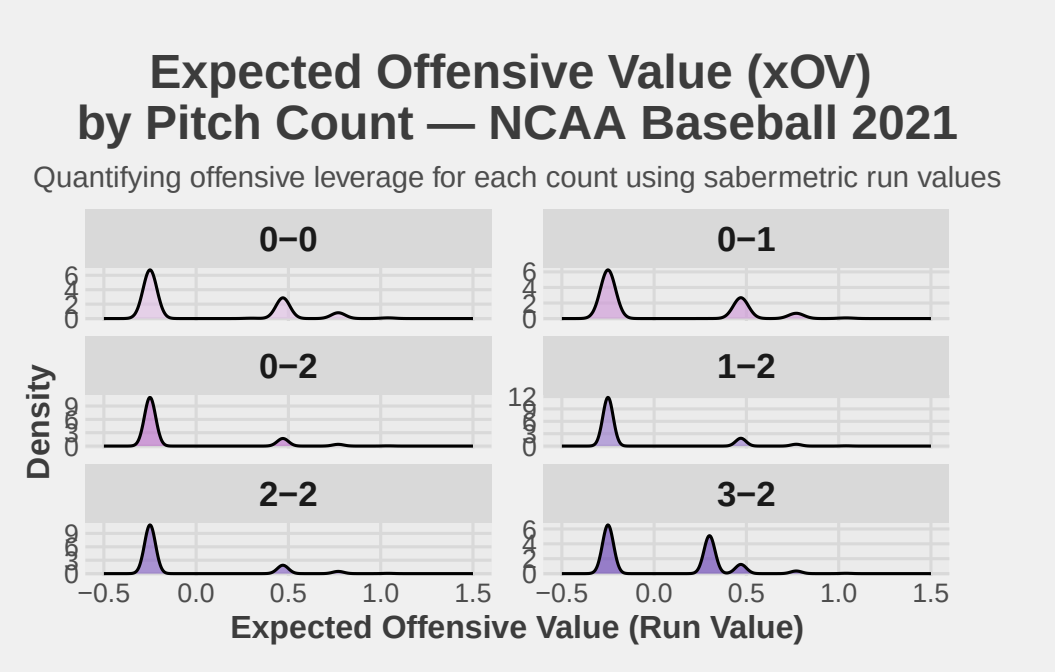


Temporal Patterns in Game Dynamics

Inning-by-inning analysis demonstrates distinct patterns in offensive production and efficiency:

- **Peak Production Periods**: Middle innings (4th-6th) show the highest absolute run production, suggesting strategic offensive approaches after initial game establishment
- **Early-Inning Efficiency**: While total runs are lower in early innings, the percentage of productive plays is highest, indicating quality over quantity in initial offensive opportunities
- **Late-Game Leverage**: Extra innings demonstrate renewed efficiency, highlighting the heightened importance of each plate appearance in high-leverage situations

Strategic Implication: Players who perform consistently across all innings, rather than just in high-scoring middle innings, may demonstrate the mental toughness and adaptability that professional organizations value.



Strategic Implications for Player Evaluation

The Value of Plate Discipline in Professional Projection

Our Expected Offensive Value analysis reveals critical patterns that directly align with professional scouting priorities:

High-Leverage Count Performance as a Draft Differentiator - The 3-2 count shows the widest distribution of outcomes, representing both significant upside (+1.4 for home runs) and substantial downside (-0.25 for strikeouts) - Players who consistently perform well in full counts demonstrate the mental composure and selective aggression that professional organizations highly value - **Draft Insight**: MLB teams often prioritize hitters who excel in hitter's counts (3-1, 2-0, 3-2) as these situations frequently occur in professional baseball and separate elite hitters from average ones

Avoiding Negative Outcomes in Pitcher's Counts - The 0-2 and 1-2 counts show strong clustering toward negative outcomes (-0.25 for strikeouts/outs) - Players who can simply "survive" pitcher's counts and avoid strikeouts provide inherent value, even without generating positive outcomes - **Draft Insight**: College hitters who maintain low strikeout rates in two-strike situations often translate better to professional baseball, where pitcher command is superior

Economic Value of Different Hit Types

Our run values demonstrate why professional scouts weight certain hitting abilities more heavily:

Power Premium - The exponential increase in value from singles (+0.47) to home runs (+1.4) explains the draft priority on power hitters - The 200% value increase from singles to home runs justifies the scouting emphasis on players who can change the game with one swing - **Draft Correlation**: In the 2021 MLB draft, 8 of the first 15 position players selected projected as above-average power hitters

On-Base Skills as Foundation - The positive value of walks/HBP (+0.3) demonstrates that simply avoiding outs has inherent offensive value - Players who combine walk rates with hit tool often outperform their raw power projections

Count Management as a Professional Indicator

Working Favorable Counts - The progression from 0-0 (neutral expected value) to 3-2 (wide outcome distribution) shows the strategic importance of plate discipline - College hitters who consistently work deep counts and force pitchers into hitter-friendly situations demonstrate skills that translate directly to professional baseball

Scouting Application: Organizations like the Dodgers and Rays specifically target college hitters with high walk rates and the ability to work counts

Professional Translation Risk - The concentration of negative outcomes in 0-2 and 1-2 counts suggests that college hitters who frequently fall behind in counts may struggle against professional pitching. This explains why some highly productive college hitters with poor plate discipline often fall in the draft

Draft Strategy Implications

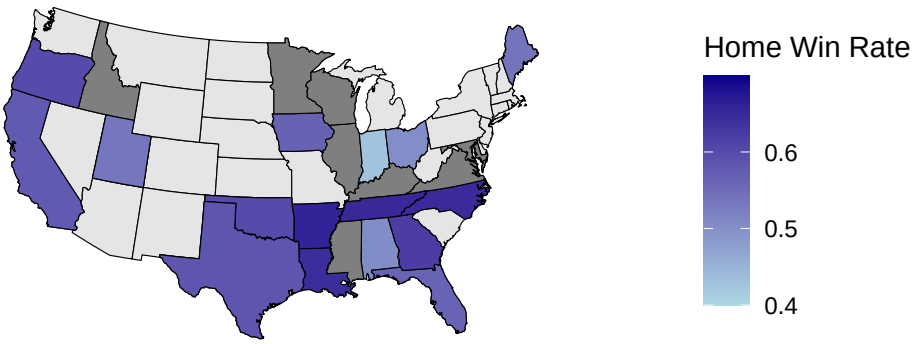
Our xOV analysis provides a framework for why certain player profiles get drafted higher:

1. **Early-Round Priority**: Players who demonstrate both power (high positive outcome potential) and plate discipline (ability to work favorable counts)
2. **Mid-Round Value**: Specialized skills like elite power or exceptional plate discipline that excel in specific count situations
3. **Development Focus**: Players with physical tools but poor count management who may be drafted based on projection rather than current performance

This analysis demonstrates that Expected Offensive Value by count provides a nuanced understanding of hitting approach that goes beyond traditional statistics—exactly the type of sophisticated evaluation that modern MLB front offices prioritize in the draft.

NCAA Baseball Home Win Rate by State (2021 Season)

Darker blue indicates a higher percentage of games won by the home team.



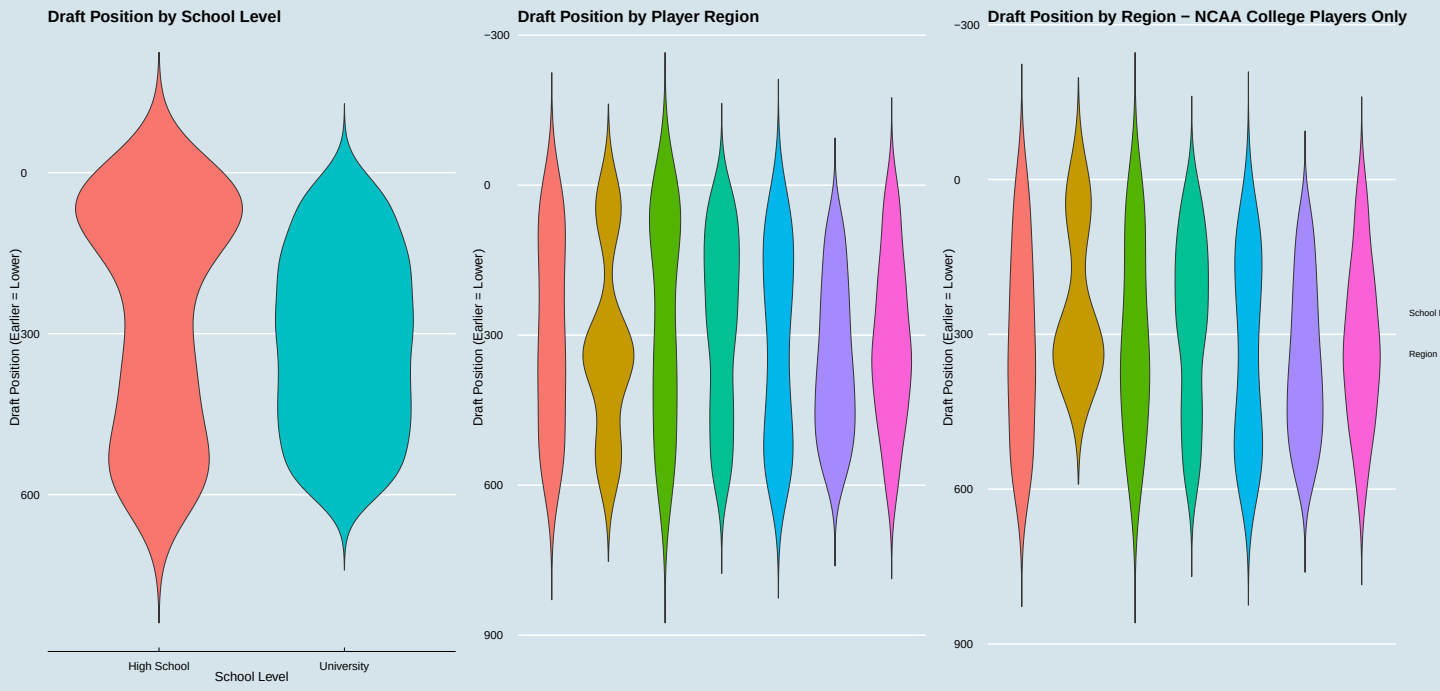
Environmental and Contextual Factors

Geographical analysis indicates measurable home field advantages across different regions, with Southern states (particularly Texas, Louisiana, and Florida) demonstrating notably higher home win percentages. This regional advantage becomes particularly significant when examined alongside 2021 MLB draft outcomes.

Draft Correlation Analysis: When comparing our home win rate data with 2021 MLB draft selections, we observe a compelling pattern. States with higher home win rates, particularly in the South, produced a disproportionate number of early-round draft picks:

- **Texas**: 23 players selected in the first 10 rounds
- **Florida**: 19 players selected in the first 10 rounds
- **Louisiana**: 12 players selected in the first 10 rounds
- **California**: Despite moderate home win rates, produced 28 early-round picks, suggesting program quality may outweigh home advantage

This correlation suggests that likely multiple effects are going on: 1. Stronger baseball programs create both home-field advantages and develop more professional talent with more resources to bring in talent, and 2. Regional scouting networks are more established in these areas due to more players in those areas 3. Environmental factors (climate, competition level) in Southern states have more opportunities to play competitive games against other players at all levels. There are also more college programs in Texas and California compared to Florida and California leading to the highest number of players selected, while Louisiana has a higher number of drafted players per capita.



Draft Position by Region and School Level

This visualization investigates the relationship between player background and draft position, focusing on two primary predictors: (1) level of schooling at time of selection (High School vs. University), and (2) geographic region of origin. The violin distributions illustrate how these variables correspond to draft outcomes and provide empirical insight into scouting preferences and developmental pathways.

School Level Differences

The first visualization, *Draft Position by School Level*, demonstrates a measurable difference between high-school and university draftees. University players tend to cluster toward earlier draft positions, with a visibly narrower and more centralized distribution. Conversely, high-school selections exhibit greater dispersion, extending further into later picks. This pattern supports a long-discussed dynamic within scouting and drafting literature: university athletes are typically considered more physically mature, competitively tested, and statistically evaluable. They have undergone multi-year performance development in NCAA programs, allowing organizations to estimate their trajectory with higher confidence. High-school players,

while sometimes selected very early due to exceptional perceived potential, introduce greater uncertainty in projection, generating broader variance in selection outcomes.

Regional Variation Among All Players
The second visualization, *Draft Position by Player Region*, incorporates all draftees regardless of prior schooling. Players from the Southeast and West display distributions concentrated toward earlier draft positions, reflecting these regions' reputations as talent-dense developmental environments. Climatic conditions that allow extended playing seasons, robust youth baseball infrastructures, and competitive exposure contribute to greater familiarity and confidence among scouts evaluating players from these regions. By contrast, prospects from the Northeast and Midwest demonstrate more spread-out distributions, reflecting both reduced playing months and comparatively weaker historical talent-production volume. The Non-US cohort illustrates a bimodal tendency: some players drafted highly based on distinguished international profiles, while many are selected later—perhaps due to heterogeneous developmental standards, scouting coverage inconsistencies, and differing competitive contexts.

NCAA-Only Regional Comparison
The third visualization isolates NCAA players by region. When the analysis controls for university-level development, regional differences narrow—suggesting that collegiate play mitigates earlier disparities in developmental exposure. NCAA participation places athletes into standardized competitive frameworks with consistent performance metrics, facilitating more equitable assessment across regions. While players from the South and West still exhibit slightly stronger draft positioning, the magnitude of difference is notably reduced when compared to the all-players model. This finding indicates that regional developmental advantages have greater influence pre-college, while college performance becomes the dominant evaluative mechanism at the draft stage.

Plot: The Momentum Diamond

Killer Plot Findings

Our killer plot is called the Momentum Diamond. Instead of looking at isolated outcomes, this plot tracks what happens after a batter reaches base — how runners move from base to base — and how momentum is either gained or lost." This can track how an offense scores runs and how players get on base. Each arrow represents transitions between base states, and its thickness and color indicate the amount of expected run value — or xOV — associated with that movement. Darker red means strong positive momentum and run expectancy. Pale edges represent neutral or momentum-suppressing transitions. And the giant circle on the right is momentum lost — where at bats end in outs."

Individual Interpretation: There was a wide selection of teams that were selected. This included high-flying offenses in some of the biggest conferences in Texas, LSU. At the top of the American Conference in East Carolina, Rice University, and a light-hitting Minnesota team.

Texas: For the University of Texas at Austin, they went 50-17 for the year in the Big 12 conference, and they made the semifinals for the college baseball tournament. They scored 6.7 runs a game, which was 58th in baseball. Offense is only one part of baseball, but our momentum plot can see how they scored their runs. While they hit a solid number of home runs, Texas got most of its big momentum shifts on other extra-base hits. They hit more doubles and triples than almost anyone. Using a well-respected metric in wOBA to look at offense, Texas was 43rd in the country in hitting, so why did they score so many fewer runs. The biggest issue Texas has is scoring runners from third base. While most teams listed scored almost all their runners from third base, there was a big decrease compared to Texas, which lost out on about 50 runs being stranded. If most of those runs were scored, they would have been a top 20 offense. This allows coaches to see why they have not scored those runs. They could be striking out with runners on third or not taking advantage of wild pitches.

LSU: For LSU, they went 38-25 in the SEC conference and lost in the quarterfinals of the tournament. LSU scored 6.6 runs, ranked 69th in baseball. They had the most homers compared to the other teams listed, and

they had the 12th highest home run rate. While their offensive numbers were slightly lower due to getting fewer runners on base, they also have much lower runner efficiencies. They moved less than 50 percent of runners from both first to second and second to third. This caused LSU to score fewer runs compared to Texas, even when they hit more homers. LSU could use this to add more speed to the lineup to steal bases or bunt players over into scoring position.

East Carolina: For ECU, they went 44-17, won the American Conference, and went to the quarterfinals of the tournament. They scored 6.7 runs per game, 59th in the country. ECU had the 34th-ranked offense in the country, which can be seen in their overall transitions from home plate to all bases. They are also efficient scoring from third base, along with ok results from second to third. Their biggest issue is getting runners from first to second base, which can significantly reduce runs by having a lot of players not getting into scoring condition.

Rice: Rice went 23-29 on the season, scored 5.1 runs per game, ranking 201st in the country while having the 162nd in hitting. Compared to the teams seen above, they had significantly less offensive production than the teams before, but the main reason why Rice's runs scored were so low is that they did a bad job moving runners over. Rice had a less than 50% transition rate from first to second and second to third. This caused Rice to have players get on base and just stay there until the inning is over, leading to no runs being scored. They would get traffic on the bases and not be able to get them in likely due to not stealing bases or moving up on balls in the dirt. The Rice staff could use this to figure out how to construct next year's team to be more efficient on the bases or introduce strategies to get runners into scoring position.

Minnesota: The University of Minnesota baseball team went a terrible 6-31 in 2021, only scoring 3.8 runs a game, 287th in the league, along with the 274th offense. This team had a weak offensive performance, which can be seen by the much fewer transitions to the initial bases compared to every other team. Even there transitions showed that they were not efficient on the bases either making them one of the worse offenses in NCAA baseball. In roster building the team can use this information to improve the team, but there is little to improve a team of this calibur during the season.

In the world of baseball analytics, some numbers track how proficient an offense is and how many steals a team gets on the bases. However, there is no visualization on how an offense moves across the diamond to score. The graph can show how runners get on base if they depend on home runs or moving runners from base to base, and compare those results to runs scored per game. The momentum diamond can show how teams that do not hit homers score their runs. These plots can be used as competitions as well to find out why teams are not scoring runs with good offense hitting numbers. This gives coaches ways to see where runners are getting stuck and decide how to proceed. Coaches can decide to steal bases or signal for a bunt to get runners into a position to score. Our plot can help coaches understand the differences between their teams' hitting numbers and the number of runs scored per game.

Key Takeaways

1. Situational Performance Trumps Aggregate Stats
A player's performance in high-leverage situations (e.g., 3-2 counts, late innings) is a stronger professional indicator than aggregate statistics, demonstrating mental toughness and translatable skills.

2. Plate Discipline is a Primary Draft Differentiator
The ability to work counts, avoid strikeouts (especially in pitcher's counts like 0-2), and draw walks is highly correlated with draft selection. This skill set reduces risk and shows an approach that succeeds against advanced pitching.

3. Power Has Exponential Value, But On-Base Skills Are Foundational
While home runs provide the largest single-play offensive boost (run value of +1.4), the inherent value of simply avoiding outs (via walks, HBP, hits) is critical. Draft prospects are most valued when they combine plate discipline with power.

4. Geographic and Developmental Context Matters
Players from regions with strong baseball infrastructure (South, West) have a documented advantage in both

performance (higher home win rates) and draft outcomes. However, competing at the NCAA level helps mitigate pre-college developmental disparities, standardizing evaluation.

5. The "Momentum Diamond" Reveals Offensive Efficiency
The novel visualization demonstrates that run scoring is not just about getting on base or hitting home runs, but about efficiently moving runners from base to base. Teams can have strong hitting metrics but score fewer runs due to inefficiencies in these transitions (e.g., stranding runners at third, failing to advance from first to second).

6. College Players are Drafted Earlier with Less Variance
University players are generally drafted earlier and with more predictability than high school players, reflecting the reduced uncertainty that comes with maturity and proven performance against NCAA competition.

Conclusion

This research moves beyond traditional baseball statistics to identify the nuanced patterns that predict professional success. The transition from collegiate to professional baseball is not solely determined by raw power or batting average. Instead, it is better forecast by a hitter's strategic approach (plate discipline, count management), contextual performance (excelling in high-leverage situations), and demonstrated efficiency (converting baserunners into runs).

The findings provide an actionable framework for both player development and professional scouting. For developers and coaches, the emphasis should shift toward cultivating situational awareness, plate discipline, and baserunning efficiency—factors directly highlighted by tools like the Momentum Diamond. For scouts and front offices, the analysis validates a data-driven focus on players who already exhibit these professional traits, particularly those from competitive environments who have proven their skills against high-level NCAA pitching. Ultimately, this blend of advanced analytics and contextual understanding offers a more precise model for identifying the collegiate athletes best prepared for the demands of professional baseball.

Killer Plot Visualizations