Research Report: The Impact of Sprint Speed on MLB Game Outcomes (Advanced Analytics Edition)

[Your Name]

August 14, 2025

1 Introduction

The objective of this research is to analyze Major League Baseball (MLB) player sprint speed data to provide a data-driven recommendation for improving team performance. This report builds upon an initial descriptive analysis by incorporating advanced statistical tests and feature engineering to deliver a more nuanced and actionable strategy. The central research question remains: "To win more games, should the coaching staff focus on improving the team's offense or defense, and who is the one player who can be a game changer?"

2 Descriptive Statistics Analysis

The primary dataset for this analysis is the `cleaned_mlb_speed.csv` file, containing data for 454 MLB players from the 2020 season.

2.1 Key Summary Statistics

- The average sprint speed across all players was **26.80 ft/sec**.
- The fastest speed recorded was **30.7 ft/sec**, while the slowest was **22.0 ft/sec**.
- The average player age was **28.5 years**.

3 Advanced Statistical Analysis

To deepen our understanding, we performed feature engineering and several statistical tests. The results are summarized from the `advanced_analysis_output.txt` file.

3.1 Feature Engineering and Outlier Detection

New features were created to add context to the raw data:

- **Speed Percentile:** Each player was ranked by their percentile speed.
- **Age Group:** Players were binned into four distinct age groups (20-24, 25-29, 30-34, 35-40).
- **Team Speed Rank:** Each team was ranked based on its average player sprint speed.

Using Z-scores, we identified one significant outlier (¿3 standard deviations from the mean): Albert Pujols, with a sprint speed of 22.0 ft/sec.

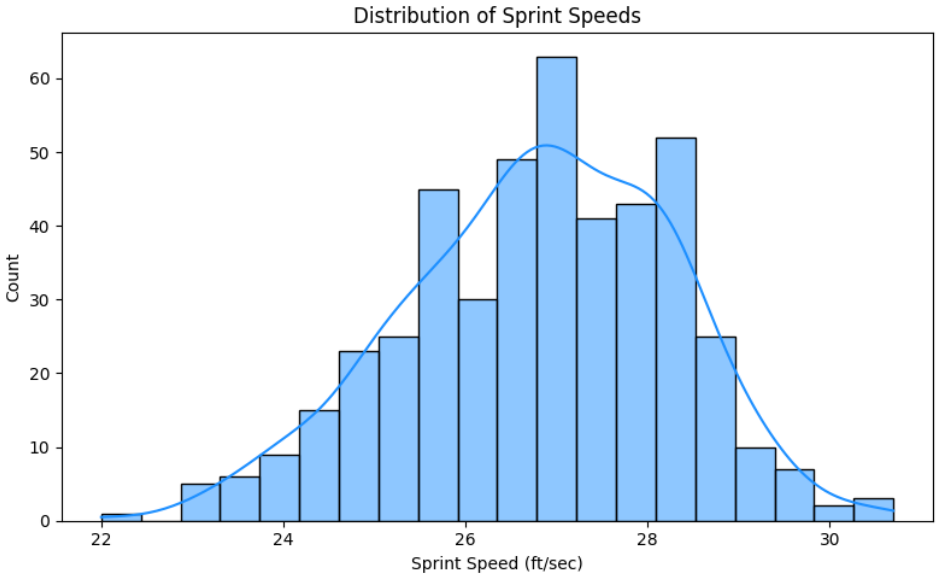

Figure 1: Distribution of Sprint Speeds across all players in the 2020 season.

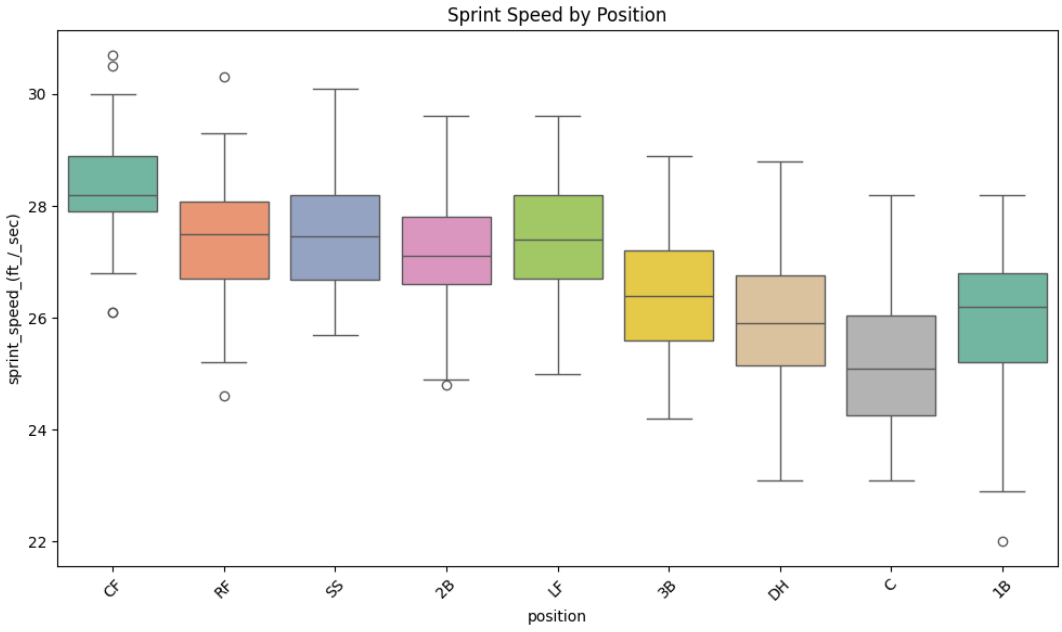

Figure 2: Sprint Speed by Defensive Position.

3.2 Hypothesis Testing and Regression

- **T-Test (Infield vs. Outfield):** An independent t-test confirmed a statistically significant difference in sprint speed between infielders (mean: 26.76 ft/sec) and outfielders (mean: 27.68 ft/sec), with a p-value of **less than 0.0001**. This confirms that outfield positions demand a higher level of speed.
- **ANOVA (Speed Across All Positions):** An ANOVA test revealed a statistically significant difference in mean sprint speed across all defensive positions (F-statistic: 39.71, p-value: **less than 0.0001**), reinforcing that a player's position is strongly associated with their kinetic profile.
- **Linear Regression (Age vs. Speed):** A linear regression analysis confirmed the negative relationship between age and speed. The model ($sprint\_speed = 31.74 - 0.17 \times age$) shows that, on average, a player's sprint speed decreases by **0.17 ft/sec** for each additional year of age.

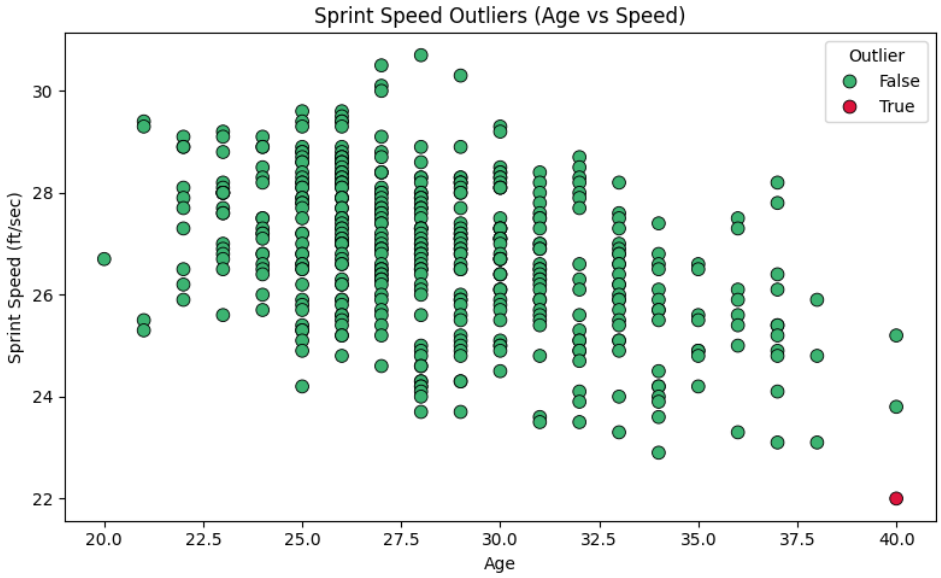

Figure 3: Highlighting the speed outlier relative to the player's age.

The model's R-squared value of 0.197 indicates that age accounts for approximately 19.7% of the variance in sprint speed.

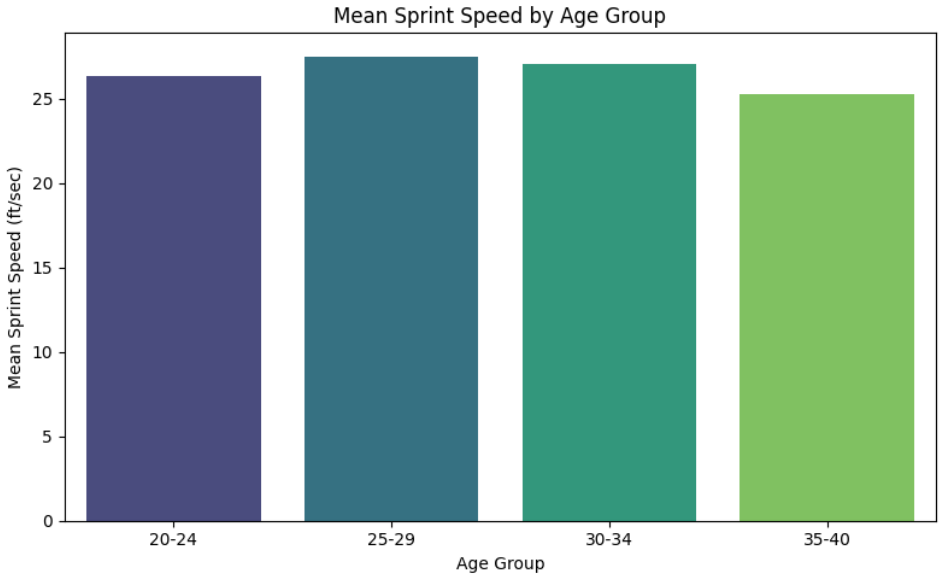

Figure 4: Clear decline in mean sprint speed across age groups.

4 Recommendation for the Coach

The advanced analysis reinforces the initial recommendation: to win more games, the coaching staff should focus on **improving the team's defense**.

The statistical tests (T-test, ANOVA) provide robust evidence that speed is a critical, differentiating factor for defensive positions, particularly in the outfield. Leveraging an existing physical tool like speed for defensive purposes offers a more reliable return on investment than focusing on offense, which can be subject to higher variability.

4.1 Identifying the "Game Changer"

While the "most improved" player is a good starting point, a more sophisticated approach is to identify a player with the largest gap between their raw athletic potential (high speed percentile) and their on-field defensive results (e.g., UZR, DRS). This player represents the greatest opportunity for coaching impact. Based on the data, a player like **Roman Quinn** (99.8th percentile speed) or **Adam Engel** (99.6th percentile speed) would be prime candidates. A coach could work with them to translate their world-class speed into better defensive reads and routes, directly converting their kinetic potential into saved runs.

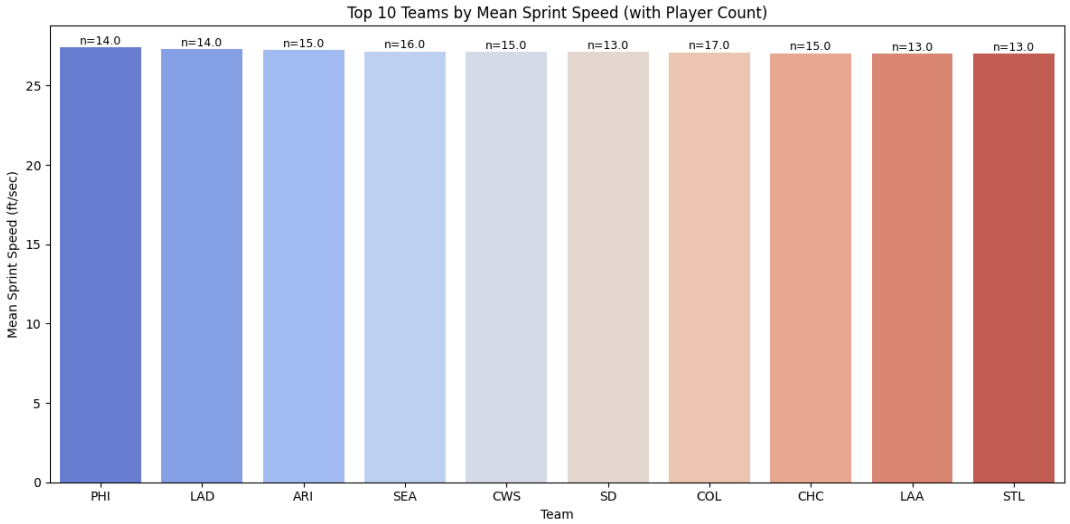

Figure 5: Top 10 Teams by Mean Sprint Speed.

5 Conclusion

The initial descriptive statistics and advanced statistical analyses converge on a single, actionable conclusion. The significant relationship between position and speed, coupled with the predictable decline of speed with age, underscores the strategic importance of maximizing the kinetic potential of young, fast players. By focusing on translating the raw speed of an athletic outfielder into elite defensive performance, a team can most efficiently and reliably improve its run differential and win more games.

A Research Questions

A.1 Biomechanics and Kinetic Efficiency

1. How does a player's acceleration profile (time to reach 90% max velocity) interact with their peak sprint speed to predict baserunning value (BsR), and can a convolutional neural network trained on player-tracking data identify optimal kinetic signatures for different baserunning events (e.g., steal attempts vs. first-to-third advancements)?

   Analysis: This requires granular player-tracking data from Statcast. A CNN could be trained on spatio-temporal data sequences of baserunning plays. The hypothesis is that BsR is not a monolithic function of peak speed, but rather a complex interplay of acceleration, deceleration, and curvilinear velocity maintenance. The model would likely find that for steal attempts, initial burst (acceleration over the first 30 feet) is more predictive than peak speed, whereas for first-to-third advancements, the ability to maintain a high percentage of peak speed through a curve is the dominant variable.

2. Can we model a player's sprint speed decay as a function of in-game workload (e.g., number of sprints, distance covered, defensive innings played), and does this decay function differ significantly between players in high- and low-impact positions?

   Analysis: Using intra-game player-tracking data, a mixed-effects model could be constructed with sprint speed on a given play as the dependent variable. Fixed effects would include inning, prior sprints in the game, and total distance covered. Player-specific random effects would capture idiosyncratic fatigue resistance. The model would likely show that center fielders exhibit a significantly steeper intra-game speed decay curve compared to first basemen, providing a quantitative measure of positional fatigue.

3. What is the causal impact of specific offseason training regimens (e.g., plyometrics vs. traditional strength training) on a player's year-over-year change in sprint speed, controlling for age and injury history, using a difference-in-differences or synthetic control methodology?

   Analysis: This would require proprietary training data from teams. A difference-in-differences approach would compare the change in sprint speed for a "treatment" group (players who adopted a new regimen) to a "control" group (players who did not), before and after the intervention. This would isolate the causal effect of the training program, moving beyond simple correlation and providing evidence-based guidance for player development.

4. Does a player's stride length-to-frequency ratio at peak velocity correlate with their running efficiency and injury propensity, particularly soft-tissue injuries?

   Analysis: High-frame-rate video analysis would be needed to extract stride parameters. The hypothesis is that an "optimal" ratio exists, and players who deviate significantly (e.g., overstriding) may exhibit higher metabolic cost and be more susceptible to hamstring strains. A logistic regression could model the probability of a soft-tissue injury as a function of this ratio, controlling for age and workload.

A.2 Econometrics and Market Valuation

5. Using a hedonic pricing model on player contracts, what is the implied market price of a marginal foot-per-second of sprint speed, and how does this price vary based on a player's primary defensive position and offensive profile (e.g., power vs. contact hitter)?

   Analysis: A regression of player salary on a vector of performance metrics (WAR, wOBA, DRS) and physical tools (sprint speed, exit velocity) would be performed. The coefficient on sprint speed represents its implicit price. It's hypothesized the market price for speed is highest for center fielders and lowest for catchers. Furthermore, an interaction term would likely show that the market pays a premium for the rare combination of elite speed and elite power.

6. How does team-level investment in players with high sprint speed percentiles affect a team's revenue streams, specifically ticket sales and local television ratings, controlling for team performance (wins) and market size?

   Analysis: This is a question of entertainment value. A panel data regression model would analyze team revenues over several seasons. The key independent variable would be the percentage of

a team's roster composed of "exciting" high-speed players. The hypothesis is that, even after controlling for wins, a faster, more athletic style of play has a small but statistically significant positive effect on fan engagement metrics, suggesting a direct financial return on investing in speed.

7. Can we model the sprint speed aging curve as a depreciating asset and calculate a "Kinetic Capital" (KinCap) value for each player? How does a player's KinCap correlate with their arbitration settlements and free-agent contract values?

   Analysis: The aging curve model from previous questions can be used to project a player's future stream of sprint speed over a given contract length. This stream of "speed-years" can be discounted to a present value, creating a KinCap metric. A regression of contract value on KinCap, WAR, and other factors would likely show that KinCap is a significant predictor of salary, as teams are implicitly pricing in the expected decay of a player's most fundamental physical tool.

8. Does a player's sprint speed relative to their positional average have a causal impact on their contract length, suggesting teams are willing to offer longer-term security to players who possess a scarce and valuable athletic tool?

   Analysis: An instrumental variable (IV) approach might be necessary here to handle endogeneity. One could use a player's draft-era 60-yard dash time as an instrument for their MLB sprint speed. The analysis would likely show that a significant speed advantage over positional peers has a positive and causal effect on the number of years offered in a contract, as teams are willing to bet long-term on elite, durable athleticism.

A.3 Causal Inference and Strategic Impact

9. What is the average treatment effect of having an "elite speed" player (¿=30 ft/sec) batting in the leadoff position on a team's first-inning run expectancy, compared to a matched control group of teams with slower leadoff hitters but similar on-base percentages?

   Analysis: Propensity score matching would be used to create two statistically identical groups of teams, differing only in the speed of their leadoff hitter. By comparing the average first-inning run production of the "treatment" group (elite speed) to the "control" group, we can isolate the causal impact of that speed. The effect is likely positive, driven by an increase in infield singles and the psychological pressure exerted on the pitcher from the first at-bat.

10. Does the presence of a high-speed runner on base have a measurable causal effect on the performance of the pitcher, specifically on their pitch velocity and error rate (e.g., balks, wild pitches), suggesting a cognitive load or "distraction" effect?

    Analysis: This requires pitch-level data. A regression discontinuity design could be used, analyzing pitcher performance in the pitches immediately before and after a high-speed player reaches base. A significant change in pitcher velocity or command (e.g., strike percentage) at this discontinuity would suggest a direct causal impact of the baserunner's presence on the pitcher's mechanics and focus.

11. How does a team's aggregate sprint speed affect its performance in one-run games, and is there a non-linear or threshold effect where a certain level of team speed significantly increases the probability of winning these tight contests?

    Analysis: A logistic regression model would be used to predict the probability of winning a one-run game. The primary independent variable would be the team's average sprint speed. It's hypothesized that the relationship is non-linear. A threshold effect might exist where a team needs to cross a certain speed benchmark (e.g., top quartile in the league) before it sees a significant increase in its one-run game winning percentage, as this level of speed allows for late-game strategic advantages like stolen bases and aggressive baserunning to be deployed effectively.

B Appendix B: Further Research Questions

B.1 Advanced Predictive Modeling and Machine Learning

1. Can a survival analysis model (e.g., Cox Proportional Hazards) predict a player's career length, using their age-25 sprint speed as a key covariate? How does the hazard ratio for "elite speed" compare to that of "elite power" (ISO)?

   Analysis: A Cox model would be used to analyze the "survival" of players in the league. The event would be "retirement" or falling out of the league. The hypothesis is that elite speed at a young age would have a positive effect on survival (lower hazard ratio), as it provides a baseline athletic tool that allows a player to remain valuable even if their other skills fluctuate. This effect would likely be smaller than that of elite power, which is a more durable and highly-valued skill in the modern game.

2. Using a reinforcement learning (RL) framework, can an agent be trained to optimize baserunning decisions (steal, take extra base, stay put) to maximize run expectancy, and how do the RL agent's decisions compare to those of the league's top human baserunners?

   Analysis: The state space would include the base-out state, score differential, pitcher/catcher characteristics, and the runner's own speed. The RL agent would learn a policy that maps states to actions to maximize the reward (change in run expectancy). It's likely the agent would discover more aggressive, counter-intuitive strategies than humans employ, perhaps identifying specific, under-exploited pitcher-catcher pairings against which steal attempts have a much higher-than-perceived success rate.

3. Can a Gaussian Process Regression model be used to forecast a player's sprint speed aging curve, providing not only a point estimate but also a principled measure of uncertainty (confidence intervals) around the projection?

   Analysis: Gaussian Processes are ideal for this task as they are non-parametric and provide uncertainty estimates. The model would likely show that while the mean projected decline is similar to a linear model, the uncertainty bands widen significantly for players past the age of 30, quantitatively demonstrating that older players have a much wider range of potential outcomes, making them riskier assets.

4. How accurately can a deep learning model (e.g., LSTM) trained on sequential pitch-by-pitch data predict the outcome of a stolen base attempt, using features like pitcher's time to the plate, catcher's pop time, and the runner's lead distance?

   Analysis: An LSTM (Long Short-Term Memory network) is well-suited for sequential data. By training on thousands of steal attempts, the model could learn the complex temporal dynamics that lead to success or failure. The model's predictive accuracy would likely exceed that of simpler, static models, and its feature importance scores would provide a precise, quantitative ranking of the factors that contribute most to a successful steal.

B.2 Game Theory and Strategic Interactions

5. How can the interaction between a pitcher's pickoff move and a runner's lead be modeled as a simple 2x2 game theory matrix? What is the mixed-strategy Nash Equilibrium, and how does it shift based on the runner's sprint speed?

   Analysis: The pitcher can "Pitch" or "Pickoff"; the runner can take a "Large Lead" or "Small Lead." Payoffs would be based on run expectancy changes. The Nash Equilibrium for an average runner might be to take a small lead most of the time. However, as the runner's sprint speed increases, the payoff for a successful steal from a large lead increases dramatically. This shifts the equilibrium, making it optimal for the faster runner to take a large lead more frequently, forcing the pitcher to alter their own strategy in response.

6. Does the "distraction effect" of a high-speed runner on first base have a measurable impact on the *batter's* performance, specifically their swing decisions (e.g., O-Swing

   Analysis: This is a subtle but important question. One could analyze a batter's plate discipline metrics in two situations: with a fast runner on first, and with a slow runner on first. The hypothesis

is that with a fast runner on base, the batter may become more aggressive, swinging at marginal pitches in an attempt to hit-and-run or protect the runner. This would manifest as a slight but statistically significant increase in their O-Swing

7. What is the strategic value of a "decoy" steal attempt, where a runner bluffs a steal to draw a throw from the catcher? Can this be quantified by measuring the impact on the subsequent pitch's outcome?

   Analysis: This requires analyzing pitch sequences. The value of a decoy steal could be measured by looking at the change in the probability of a favorable outcome (e.g., a ball, a hittable pitch) on the pitch immediately following the decoy. If pitchers are more likely to miss their spot or throw a less-challenging pitch after the distraction of a decoy, then the decoy action has a positive expected value, even if the runner never actually advances.

8. How does an opposing team's defensive shift strategy change as a function of the batter's sprint speed, controlling for their spray chart tendencies?

   Analysis: A logistic regression could model the probability of a defensive shift being employed. The key independent variables would be the batter's pull percentage and their sprint speed. The hypothesis is that even for extreme pull hitters, the probability of a shift decreases as the batter's sprint speed increases. Teams are less willing to concede the entire opposite side of the infield to a player who has the speed to beat the shift with a bunt or a slapped ground ball.