Github: https://github.com/rrkapurberkeley/Project2_Kapur

Rohan Kapur
Data Sci 200
12/11/2024

Final Report: Baseball Strategies

Background

The 2024 World Series between the Los Angeles Dodgers and New York Yankees saw a boost in viewership and ratings compared to the previous year. This year's games averaged 15.8 million viewers, marking a 74% increase over 2023's World Series. Watching the games sparked an interest in me to understand strategies of both batters and pitchers. I was curious to see if there were any strategies that could be pulled out of data to inform players moving forward. The dataset I used was from Statcast, MLB's advanced analytics system, implemented in all major league stadiums since 2015. The dataset tracks a wide range of baseball data in real-time, capturing player performance metrics from totals of homeruns, strikes, whiffs to technical metrics such as pitch velocity, exit velocity, launch angle. The platform allows you to select hundreds of dataset points and to download it to a csv.

Questions

I used Batters and Pitchers to categorize my research.

The three questions I researched for Batters were:

1. What is the best indicator to give batters confidence to steal a base?
2. Is there any correlation between speed and success of stealing a base?
3. Which Base requires the highest speed to steal?

The three questions I researched for Pitchers were:

1. What is the best pitch for a Whiff, Groundball, Strike, Homerun?
2. Are there any correlations between pitch types and resulting hits?
3. Do any of these insights contradict with hitting home runs?

Data Cleaning / Sanity Checks

The initial batter dataset in its CSV format had a shape of (129,182) and included key variables such as Hit Outcome, First Strike %, and Hit Types, among others. With over a hundred columns, I had to carefully review and understand each one to answer my research questions effectively. At first, I struggled with some of the more technical terms like xBA, xSLg, xOBA, and xwOBA, which required some additional research to grasp. The pitchers dataset, with a shape of (107,279), included key variables such as pitch

types, pitch outcomes, and velocity. Each variable offered potential insights into my questions, so it was crucial to filter down to the specific columns that would provide the most relevant information.

Correlation Test Used:

The correlation test used in this analysis calculates the Pearson correlation coefficient to measure the linear relationship between the proportions of different pitch types—such as fastball, offspeed, and breaking—and the target variable, whiff_percent. The Pearson correlation coefficient ranges from -1 to 1, where a value of 1 indicates a perfect positive correlation (as one variable increases, so does the other), -1 indicates a perfect negative correlation (as one variable increases, the other decreases), and 0 indicates no linear relationship between the variables. This statistical test assumes a linear relationship, that both variables are approximately normally distributed, and that there are no extreme outliers that could skew the results. A high positive correlation (e.g., 0.5 or higher) suggests a strong association between a specific pitch type and a higher whiff_percent, while a high negative correlation (e.g., -0.5 or lower) indicates the opposite. A coefficient close to 0 suggests little to no relationship between the pitch type and the whiff_percent.

Compelling Text and Data Stories(Batters)
Compelling Figures

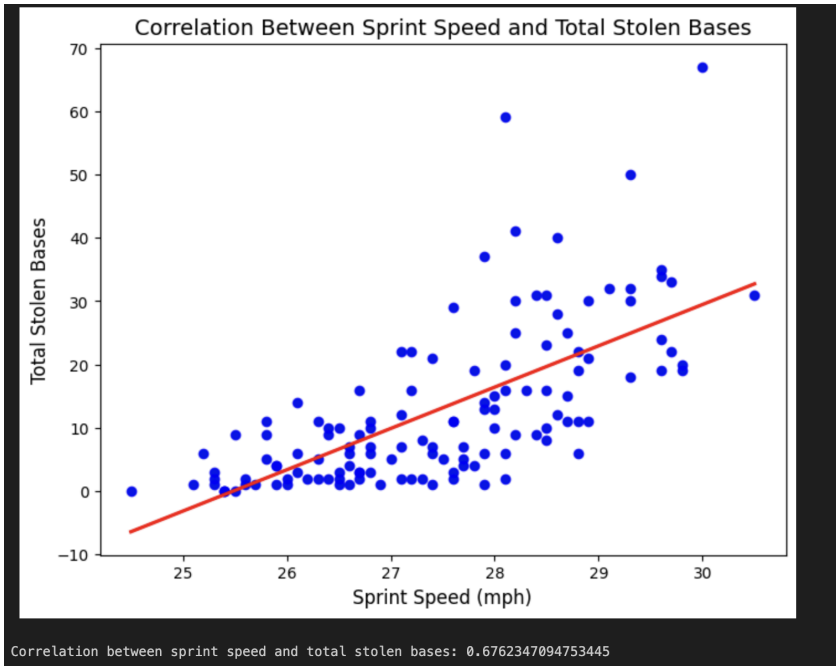
Top 10 fastest players by sprint speed:

	last_name, first_name	sprint_speed
2	Witt Jr., Bobby	30.5
26	De La Cruz, Elly	30.0
128	Peña, Jeremy	29.8
92	Langford, Wyatt	29.8
102	Chourio, Jackson	29.7
87	Young, Jacob	29.7
55	Carroll, Corbin	29.6
70	Duran, Jarren	29.6
32	Rodríguez, Julio	29.6
101	Turner, Trea	29.6

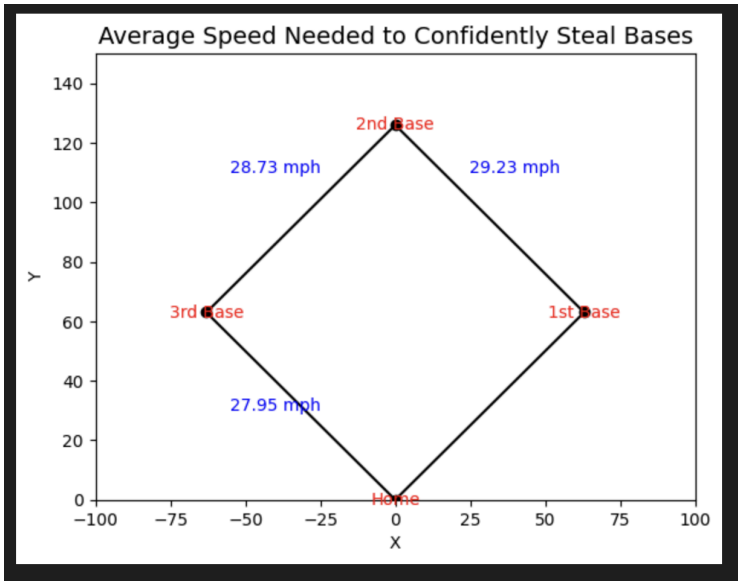
Top 10 players to steal the most bases:

	last_name, first_name	r_total_stolen_base
26	De La Cruz, Elly	67
61	Ohtani, Shohei	59
120	Turang, Brice	50
3	Ramírez, José	41
31	Chisholm Jr., Jazz	40
126	Garcia, Maikel	37
55	Carroll, Corbin	35
70	Duran, Jarren	34
87	Young, Jacob	33
77	Thomas, Lane	32

The graphs illustrate the top 10 fastest players by sprint speed and the players with the most stolen bases. When comparing these side by side, certain names overlap, but there are surprising exceptions, such as Shohei Ohtani, who ranked second for most bases stolen but did not appear in the top 10 fastest players. This finding intrigued me, prompting a deeper dive into whether there is a genuine correlation between sprint speed and stolen bases or if players often rely on other factors when deciding to steal.



A scatter plot comparing all players' sprint speed and stolen bases reveals a strong correlation (approximately 0.68), indicating that faster players are generally more successful in stealing bases. However, outliers stand out—players who are not the fastest but still achieve exceptional numbers of stolen bases. This suggests that stealing bases involves more than just raw speed; elements like timing, situational awareness, and strategy play a critical role.



Another graph represents a baseball diamond with the average sprint speeds required to successfully steal each base (home, first, second, and third). Interestingly, second base has the highest average sprint speed, challenging the common perception that stealing home is the most difficult. This insight provides batters with actionable benchmarks to gauge their readiness in high-pressure situations. While these speed averages are not strict thresholds, they offer guidance and boost confidence for players considering an attempt.

Final Recommendation for Batters:

Based on the 2024 data, here are my key recommendations for batters:

1. Use Speed Benchmarks Wisely: Be mindful of the average sprint speeds required to steal each base. These benchmarks can serve as a confidence booster when deciding to take a chance.
2. Focus on Second Base Speed: Second base demands the highest speed to successfully steal, so players should prioritize training for this sprint.
3. Avoid Imitating Outliers: Exceptional players, like Shohei Ohtani, succeed through a combination of skills beyond just speed. Do not rely solely on their patterns when making decisions, as they represent outliers in the data.

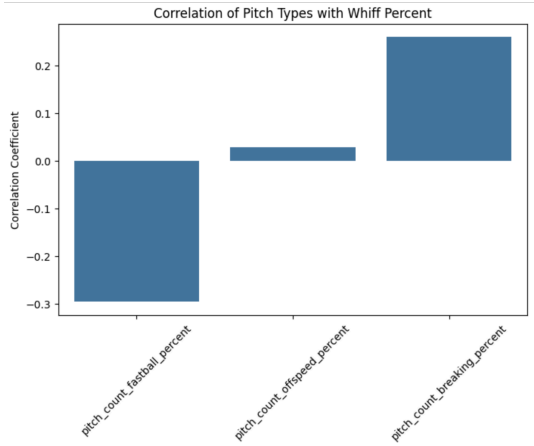
Background information on Pitching:

Pitching in baseball is a strategic art, with pitchers relying on different types of pitches to outwit batters. These pitches fall into three main categories:

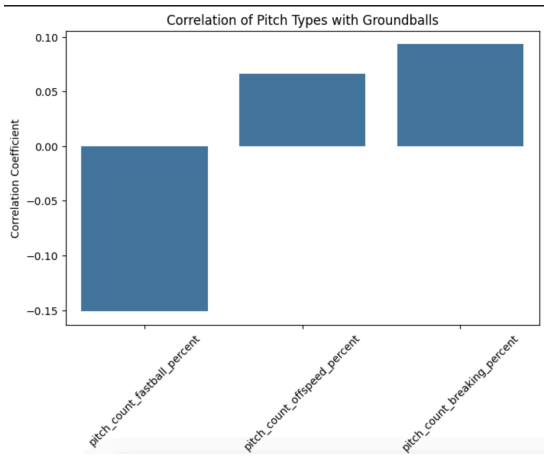
1. Fastball
    a. A fastball is the most basic and widely used pitch in baseball, characterized by its high velocity and relatively straight trajectory.
        i. Ex: Four Seam, Two Seam, Cut
2. Offspeed
    a. Offspeed pitches are thrown slower than fastballs and designed to disrupt the hitter's timing.
        i. Ex: Changeup, Splitter
3. Breaking
    a. Breaking pitches feature high spin and movement, often breaking downward or sideways as they approach the plate.
        i. Ex: Curveball, slider, sweeper

Each pitch type is used strategically depending on the game scenario, batter tendencies, and the pitcher's strengths. My goal was to analyze data to determine if specific pitches correlate well with certain outcomes in various situations.

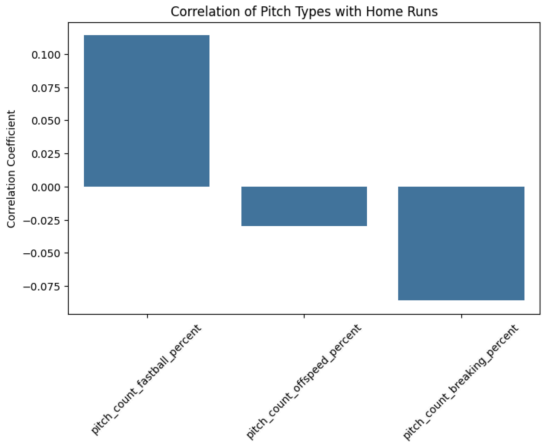
Compelling Text and Data Stories(Pitchers)
Compelling Figures



- The highest correlation with whiffs was observed with Breaking Pitches, which showed a correlation value of 0.259.
- Analysis: This suggests that breaking pitches are the most effective for inducing swings and misses. The spin and unpredictable movement of breaking pitches likely make them difficult for batters to track and hit cleanly.



- The highest correlation with ground balls was also found with Breaking Pitches, with a correlation value of 0.093.
- Analysis: While the correlation is not as strong, breaking pitches seem to effectively induce ground balls. This makes them a strategic choice in double-play scenarios or when aiming to minimize high-risk fly balls.



- The highest correlation with home runs was associated with Fastballs, showing a correlation value of 0.114.
- Analysis: Fastballs, while powerful, are relatively predictable and easier to time if the batter anticipates them correctly. This emphasizes the importance of mixing pitch types to keep batters off balance.
- This graph also served to validate strategies, ensuring no contradictions—such as breaking pitches excelling at inducing whiffs and ground balls while also being highly correlated with very successful hits

Final Recommendation for Pitchers:

1. When to Leverage Whiffs:
    a. Strategic Value: Whiffs disrupt the batter's rhythm, demoralize the offense, and give the pitcher confidence and momentum.
    b. Situational Use: Breaking pitches should be prioritized in high-pressure situations where minimizing batted-ball events is crucial, such as with two strikes or with runners in scoring position.
2. When to Target Ground Balls:
    a. Strategic Value: Ground balls are valuable for inducing double plays, particularly with a runner on first and less than two outs.
    b. Situational Use: Breaking pitches are effective in these scenarios but should be carefully located to avoid leaving the ball over the middle of the plate, which could lead to hard contact.
3. Avoiding Home Runs:
    a. Strategic Consideration: While fastballs are essential in a pitcher's repertoire, they are more likely to result in home runs if overused or poorly located. Mixing in offspeed and breaking pitches can mitigate this risk and keep batters guessing.

Limitations:

While this analysis provides valuable insights into batter and pitcher strategies, it has several limitations. The data is based solely on the 2024 season, which may not fully capture year-to-year variations in player performance or team strategies. Additionally, the correlation analysis assumes a linear relationship, which might oversimplify complex interactions between variables. Factors like player psychology, in-game decision-making, and defensive alignments were not accounted for, potentially limiting the applicability of the recommendations. Finally, outliers such as Shohei Ohtani highlight that some players excel due to unique skill sets, which may not align with the broader trends identified.

Conclusions:

This report highlights actionable insights for both batters and pitchers to optimize their strategies using data-driven approaches. For batters, understanding speed benchmarks and focusing on second-base readiness can improve base-stealing decisions, while pitchers can leverage breaking pitches strategically to maximize whiffs and ground balls while mitigating the risk of home runs. Despite its limitations, this analysis underscores the power of Statcast data in uncovering meaningful trends and provides a foundation for further exploration of baseball strategies in future seasons.

Works Cited:

1. "Custom Batter Leaderboard." Baseball Savant, MLB Advanced Media, 2024, https://baseballsavant.mlb.com/leaderboard/custom?year=2024&type=batter&filter=&min=q&selections=pa%2Ck_percent%2Cbb_percent%2Cwoba%2Cxwoba%2Csweet_spot_percent%2Cbarrel_batted_rate%2Chard_hit_percent%2Cavg_best_speed%2Cavg_hyper_speed%2Cwhiff_percent%2Cswing_percent&chart=false&x=pa&y=pa&r=no&chartType=beeswarm&sort=xwoba&sortDir=desc.
2. "Custom Pitcher Leaderboard." Baseball Savant, MLB Advanced Media, 2024, https://baseballsavant.mlb.com/leaderboard/custom?year=2024&type=pitcher&filter=&min=q&selections=pa%2Ck_percent%2Cbb_percent%2Cwoba%2Cxwoba%2Csweet_spot_percent%2Cbarrel_batted_rate%2Chard_hit_percent%2Cavg_best_speed%2Cavg_hyper_speed%2Cwhiff_percent%2Cswing_percent&chart=false&x=pa&y=pa&r=no&chartType=beeswarm&sort=xwoba&sortDir=desc.