Investigating Baseball's Newest Scandal

Analyzing MLB's Pitch Spin Rate Data to Identify Potential Cheaters

DS 5100 Semester Project

MacKenzye Leroy, Jason Wang, & Stephen Whetzel

July 18, 2021

Abstract

On June 21, 2021, Major League Baseball began enforcing a ban on sticky substances used by pitchers to increase their spin rates. Over the past couple of years, before this ban was implemented, we have seen pitch rotations per minute (RPMs) skyrocket, leading to record lows of batting averages and hits. Since the rule was put in place, however, pitch RPMs have gone down, and more balls have been hit per game across the league. This paper investigates whether this trend can be attributed to this recent ban and investigates suspicious teams and individuals by looking at how their spin rate metrics have changed over time. By analyzing over 2.8 million individual pitches, we were able to analyze the drop in RPM across the league which has led to batters hitting the ball more. We were able to create plots that validated our initial beliefs of how various metrics, such as number of strikeouts or balls hit, were affected by the use of sticky substances. While there was no evidence of team driven cheating, there was significant evidence of league wide and player wide usage of the sticky substances to improve spin rates and therefore pitching performance.

Introduction

In 2015, Major League Baseball (MLB) introduced a barrage of cameras and radar guns in every single stadium (Cole 2014). The entire system is known as StatCast and captures every movement on a baseball field. If a player so much as takes a short half step off third base and sneezes, StatCast likely captures it, or at least could. StatCast also tracks the baseball itself in a previously unprecedented way. StatCast captures and logs the release point, velocity, spin rate, spin axis, and dozens of other variables for every single pitch thrown in competition. This new wealth of data has offered insight previous generations of players, executives, and fans could only dream of.

One of the key insights discovered was the importance of spin rate. For years, people have understood the importance of pitch velocity, with even your most novice baseball fan understanding that the higher the velocity of a pitch, the harder it is for a batter to hit it. What was not fully understood pre-2015 is that the spin rate of a pitch can often be even more predictive of whether a batter will swing and miss at it. In fact, per Berkeley Sports analytics, a 92-mph fastball with a spin rate of 2800+ RPM produced a slightly higher swinging strike rate than a 98-mph fastball with a spin rate of 2100 RPM (Duston 2020). In other words, a slower pitch with a higher spin rate is statistically more effective at getting batters to swing and miss.

Armed with this knowledge, pitchers across the league began trying to increase their own spin rates. The problem, however, is that this is nearly impossible to do naturally. According to Alan Nathan, a University of Illinois physics professor and MLB consultant, “It’s probably pretty hard to change that [fastball spin] ratio for an individual. I can see that you could do it for a curveball because a curveball involves some technique whereas a fastball is pure power. There is no finesse.” (Sawchik 2018)

Given this, pitchers started looking for alternative ways to increase their spin rate and therefore performance. They allegedly found a solution by applying foreign substances to their hands, whereby they could in fact increase their spin rates substantially without increasing their velocity (Sawchik 2018)). This has led to a surge in pitching performance over the last several years with offensive stats in MLB at all-time lows (Baseball-Reference.com). The controversy around spin rates and foreign substances hit a fever pitch this year when the 2021 season got off to one of the worst offensive starts

since the end of the dead ball era in 1919 (Janes 2021). As a result, in June MLB announced that it would officially be cracking down on the use of foreign substances in baseball. They did in fact start checking pitchers for the use of foreign substances multiple times a game starting on June 21 of this year (MLB 2021)

Given this background, we set out to investigate with two main goals. First, we wanted to confirm and quantify that over the last five years there has in fact been an increase in spin rate in MLB accompanied by a significant drop in June 2021. Second, we wanted to systematically analyze data leaguewide to find which teams and individual players have had the largest increases and subsequent decreases in spin rate during that time.

Data Description

Our data set comes from MLB itself which makes data from StatCast available through their Baseball Savant website. MLB does not make its data available through an API but instead releases limited amounts of data in CSV form which is made accessible through an on-site search function. We found this tool to be incredibly difficult to use. It was impossible to designate that you wanted the most granular level of pitch-by-pitch data and the CSVs that we downloaded were often grouped over some time frame or by team entity rather than on a pitch-by-pitch basis. This made our goal of getting individual pitch data for every regular season pitch from 2017 onward very difficult. We realized that by only requesting data from a single team over a single season we were able to get the data segmented how we wanted it, at the most granular level. However, for 30 teams over five seasons, this would mean that we would be manually downloading and concatenating 150 CSVs, something that we did not want to spend a significant amount of time doing.

Fortunately, we were able to isolate the relevant parameters for individual team and season in the URL released by Baseball Savant to download the CSVs. By looping through these parameters for each team and season in Python, we were able to request and add all 150 CSVs to a new master CSV. The result was a CSV containing all available data from every pitch thrown in MLB since 2017, about 2.8 million rows of data. This was an incredibly rich source of data to work with, allowing us to gain a complete and nuanced picture of everything that happened on the field in MLB for the last 5 seasons. The data included important metrics such as spin rate, velocity, pitcher ids, pitch type, and at bat results. There were also a lot of metrics that were not relevant to our study such as fielding statistics and information on batters. To cut down on the overall size of the data, we dropped columns outside of the scope of our study.

Missing data was rare in this dataset, but we did need to add a couple of columns to facilitate our research. The data included home and away teams but did not designate which team the pitcher in question played for. We created a new column with the pitcher's team by designating their team as the home team if they were pitching in the top of the inning and designating the team as the away team if they were pitching in the bottom of the inning. We also created a new column that converted game date strings to the Pandas Date Time format to facilitate our work with trends over time. The result of our efforts was a streamlined but incredibly rich data source formatted to suit the needs of our study.

League Wide Contextualization

Before diving into spin rate data, it is important to contextualize what all of this means for the game. Over the past five years, league wide offensive numbers have steadily decreased. On-base-percentage, batting average, and runs are all at historic lows (baseball-reference.com). One of the best illustrations of this precipitous drop-off in offensive production over the last five years is the ratio of strikeouts to hits. Throughout most of baseball's 100-plus-year history, there have been more hits than strikeouts every single month. This changed in April 2018, the first full month of play in which the league saw more strikeouts than hits (Blum 2018).

We decided to use our data set to explore this issue a bit more. By using the "events" column provided by StatCast we could determine how many strikeouts and hits (singles, doubles, triples, and homeruns) there were in the entirety of our dataset. By using our new column for month, we were able to group the count of strikeouts per month as well as hits per month and compare the ratio. Figure 1 below shows this ratio by month, with the red line being the baseline ratio of one strikeout per hit. As you can see, even though April 2018 was the first year to see more strikeouts than hits, it has become commonplace within the last six months for there to be more strikeouts than hits in any given month.

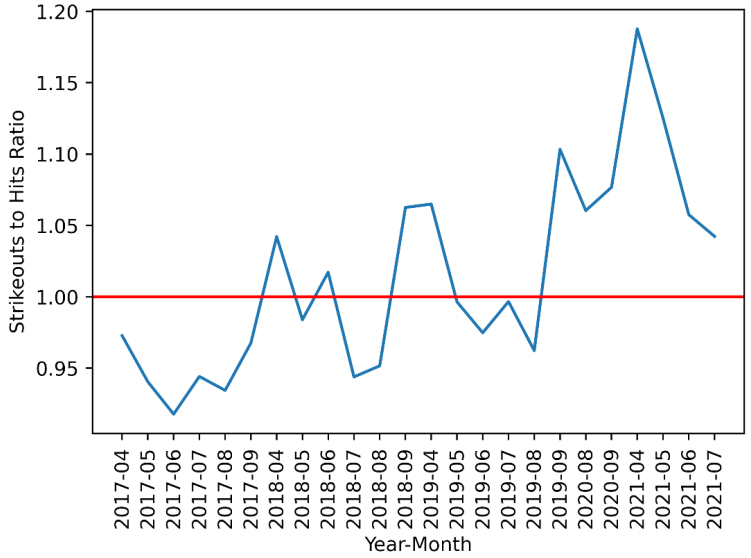

Figure 1- Ratio of strikeouts per hit by regular season month*. Post season or off-season months are not included here. Red line is the baseline 1 strikeout per hit, a ratio that was not exceeded in MLB's entire history until April 2018.*

Investigating League-Wide Trends

To explore the issue on a league-wide level, we first needed to group our data by month. By grouping our spin rate data by month and plotting the league-wide average over time, we can confirm our hypothesis. Since 2017, there has been a steady increase in the league-wide average spin rate followed by a steep drop-off in June and July of this year as displayed in Figure 2 below. This drop-off perfectly coincides with MLB's announcement that it would enforce its rules against foreign substances, providing strong evidence that the gains in RPM were made due to the use of foreign substances. We repeated the process with velocity and plotted the result (see Figure 3 below).

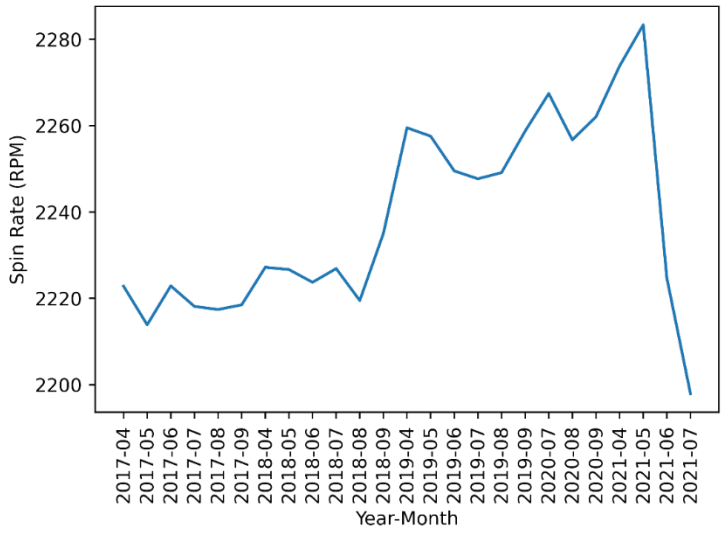

Figure 2 – Average spin rate league-wide from April 2017 to July 2021. *The upward trend since 2018, followed by the sharp decline in spin rates since the league's June 21, 2021, ban confirms our general hypothesis.*

Recall that it is extremely difficult (if not impossible) to increase a pitcher's spin rate without also increasing velocity. We would therefore expect to see velocity increase in parallel with spin rate if pitchers were in fact achieving these spin rate jumps without the aid of any foreign substances. Unfortunately, this is far from the case, as velocities have shown a wide range of variance in the period with seemingly no correlation with our spin rate data (Figure 3). In fact, July is on pace to have the lowest monthly spin rate since April 2017 as well as the highest velocity. We believe this may be due to pitchers throwing harder to compensate for drops in RPM as they forego the use of foreign substances.

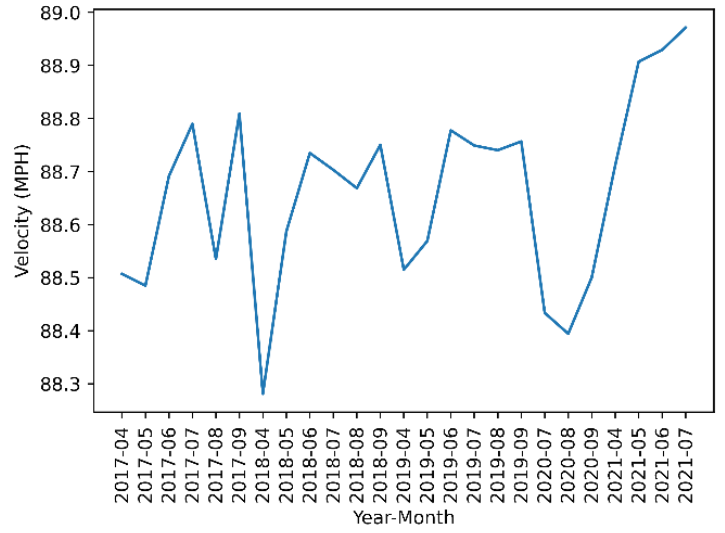

Figure 3 – Average MLB pitch velocity by month. *Data shows no correlation with spin rate trends and in fact may show a spike in the last two months as pitchers try and compensate for a lack of foreign substances by throwing the ball harder.*

Diving deeper into the data, we also investigated different pitches, where we grouped the average (mean) RPM of pitches by month and pitch type (see Figure 4 below). Looking at Figure 4, it is obvious that there is some variation in RPM over the last four years. Almost all these pitches, however, saw a

significant decrease starting in June of this season. This could be due to the increasing media pressure and MLB investigation into the issue before the official ban, leading pitchers to forego the use of foreign substances. The trend is clear, however, with RPM rates increasing steadily over the past four years and dropping substantially over the last two months.

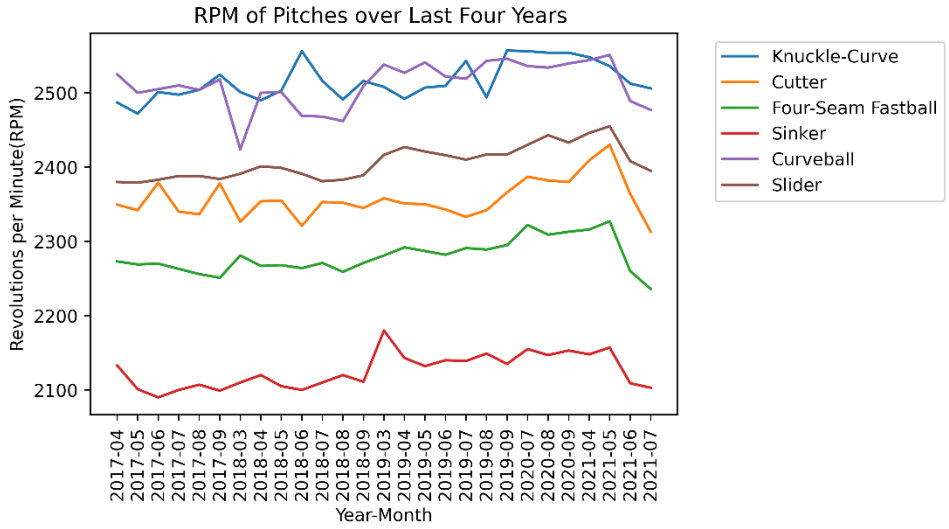

Figure 4 – Spin rate in RPM by pitch type, league-wide since April 2017. *There is a clear drop-off since the MLB instituted its ban on foreign substances on June 21, 2021.*

Investigating Individual Teams: Data Processing

When looking at whether teams were cheating as entire units—colluding together and cheating across an entire pitching staff or organization—we needed to ask three questions of each team: did the team see a large drop in spin rate after the June 21st ban, did players that joined the team usually see an increase in their spin rates once joining, and did players that left the team usually see a decrease in spin rates after leaving the team? To normalize for factors like team pitching strategy, where certain teams might be more likely to throw pitch types with higher spin rates than other teams, we looked only at Four-Seam Fastballs for the purpose of this study, and filtered our data frame to only use rows of data that had the pitch-type designation “FF”.

It was relatively simple to process the data to answer the first question, whether a team saw a decrease in spin rate after the ban on June 21st. To get team spin rates before the ban, we filtered the data for only games that occurred between April 1 and June 20, 2021 and then took the average fastball spin rate by taking the mean for each team. We did the same thing for all games on or after June 21. By adding this data to a new data frame we compiled a list of spin rates before the ban, after the ban, and the percent change for each team. We then plotted this data frame as shown below in Figure 5, with pre-ban rates for each team plotted on the x axis, post-ban rates plotted on the y axis, and the size of the data point corresponding to the magnitude in percentage drop.

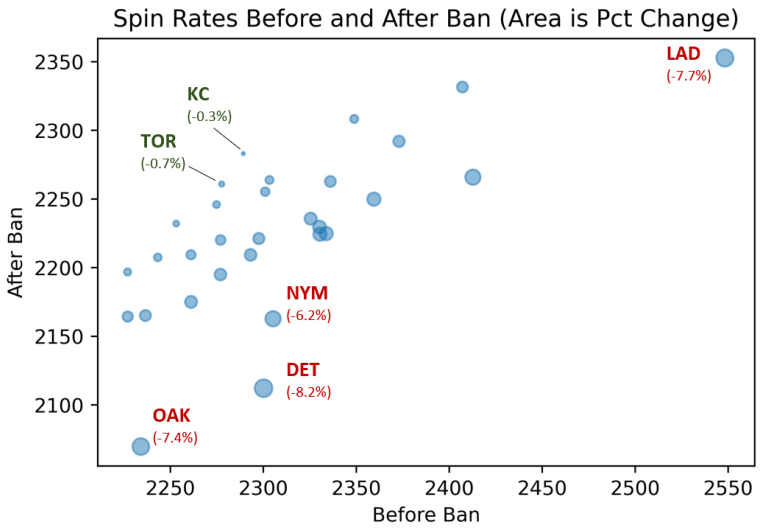

Figure 5 – Average spin rates by team before and after the foreign substance ban on June 21, 2021. *Size of plot point corresponds to magnitude of the percentage change before and after the ban. Red labels added for the teams with the four largest percent drops and green labels added for the two teams with the smallest drops.*

Processing the data to answer the second and third questions proved a bit more complicated as we first needed to look at player-by-player data and then summarize that data by which team a player either left or played for. First, we filtered the data frame by available pitcher ID, a unique id number corresponding to a single pitcher. By looping through all the rows available for each pitcher sorted by game date, we were able to group data by an individual stint that a player had with a team. We could not filter or group by each team that the player played for as they might have played for a single team multiple times in their career, so we had to move through the data row by row in chronological order to get the numbers we were looking for. By doing this we were able to create a data frame summarizing each player's spin rates for their entire time with a single team. We were also able to look at their average spin rates before they joined that specific team and after they left that team, if available.

Using this grouped and summarized data, we were then able to look at team-wide trends for how player spin rates changed once a player joined or left a team. For each player that had joined a team from another team we took their percentage change in average spin rates and added it to a list of changes in average spin rates. We then took the average of these percent changes to get an idea of how much a player's spin rate increased or decreased once they joined a specific team. We repeated the process for each player that left a team and dropped any players that did not throw at least 25 fastballs to make sure the averages were not skewed by small sample sizes. By using aggregated data by player and team instead of individual pitch data, we were able to ensure that the data was not skewed by a single pitcher playing for a much longer time than another pitcher on the team. We then plotted the results of this process (see Figure 6 below) with the average percentage change in spin rate after joining a team on the x axis and the average percentage change in spin rate after leaving a team on the y axis.

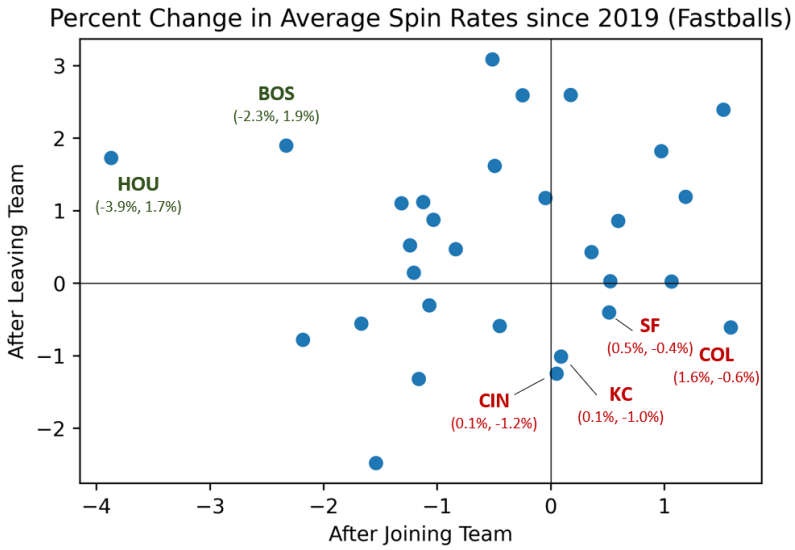

Figure 6 – Average percent change in average spin rate once a player joins or leaves a team (four seam fastballs only). *Red labels show the 4 teams who saw average increases in spin rates once players joined and decreases once players left. Green labels show two teams with significant drops in spin rates once players joined the team and significant increases in spin rates once players left the team.*

Investigating Individual Teams: Results

If a team were colluding together to cheat in a systematic fashion, we would expect that they would meet all three criteria outlined above: team-wide spin rates dropping drastically after the June 21 ban, an average rise in spin rates once a player joined the team in the last few years, and an average drop in spin rates once a player left the team in the last few years. For the latter two criteria there were four teams that met both requirements, San Francisco, Kansas City, Colorado, and Cincinnati (labeled in red in Figure 6). All four of these teams saw pitchers gain spin rate once joining the team while pitchers who left their team generally lost spin rate while pitching for their next team.

If these teams were systematically cheating, we would expect them to see some of the most drastic drops in spin rates after the June 21 ban. When looking at Figure 5, however, we see that none of these teams are in the top four of all MLB teams for spin rate loss since June 21. Instead, Cincinnati ranked 10th with a 3.8% drop, San Francisco 16th with a 3.1% drop, Colorado 21st with a 2.3% drop, and Kansas City ranked dead last, 30 out of 30, with a mere 0.3% drop in spin rate.

In the face of this mixed evidence, we must conclude one of two things. One, these teams are still cheating on a large scale even in the face of MLBs crackdown without getting caught, a conclusion we find unlikely given the urgency and diligence with which MLB is now pursuing this issue. Two, the apparent rise in spin rates after a player joins a team and then a corresponding drop once players leave the team may be indicative of some cheating but not a conspiracy to cheat. We favor this second conclusion and its implications, that some teams have higher concentrations of cheating players hence the large differences in team spin rate drops after the June 21 ban, but that it is unlikely that entire teams are colluding together to cheat systematically. Because there are not any teams that clearly fit all three of our criteria, we instead conclude that some teams have a higher rate of cheating pitchers, who may even work together on a limited basis to cheat, but that based on the evidence available it is unlikely that entire teams are cheating systematically at the level of team-wide policy.

Investigating Individual Players: Data Processing

While there may not be significant evidence to conclude that teams were cheating on a systematic basis, individual pitchers may have a different story. We took the data from the 2021 season and split up the data to pitches either before or after the ban on June 21st. Factoring out pitchers who did not have at least 1000 pitches throughout the entire season, the data is made up of established pitchers who have little variance in their pitch RPMs due to the large sample sizes.

Investigating Individual Players: Results

Looking at Figure 7 below, we see that Trever Bauer of the Los Angeles Dodgers has had the most change in RPM, dropping a little over 8% in RPM spin rate since MLB's June 21 ban. Not far behind him is Frankie Montas of the Oakland Athletics with a little over 7% drop in RPM after June 21. In fact, out of the 21 established pitchers that we measured, 17 of them saw a drop in RPM after the rule was enforced. Only four players had nearly the same or slightly higher spin rates after sticky substances were banned. The widespread decrease in spin rate RPM points to substantial usage of sticky substances by pitchers throughout the league, and helps explain the changes that we have seen over the past couple of months in batting performance.

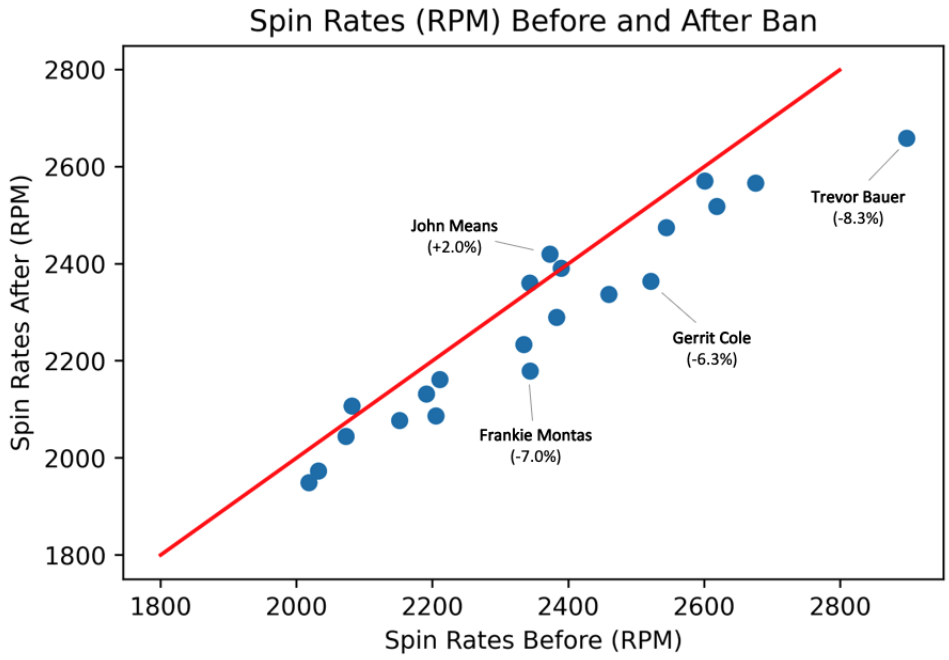

Figure 7 – Average spin rates in 2021 by individual pitcher before and after the June 21 ban, filtered for pitchers who have thrown 1000+ pitches in 2021. *Trevor Bauer, Frankie Montas, and Gerrit Cole have the most evidence for the use of foreign substances with the three largest drops in average spin rate since the ban. John Means of the Baltimore Orioles has seen an average rise in spin rate since June 21 indicating either an increased level of cheating or a lack of cheating altogether.*

Unit Testing: Initial Dataset

Although we did not write a thorough test to check that every single pitch for every single team made it into our dataset, we did check first that pitches from every team did make the dataset and check second

that the total pitch count for each team was within an expected range. First, we confirmed that the number of teams in our dataset was in fact 30, the number of teams in MLB. Moving on, we formulated an expected range of pitches thrown per team over the course of our study. A standard baseball season is 162 games long, but 2020 was shortened to 60 games due to the Covid-19 pandemic and we pulled our data when MLB was about 70 games into the 2021 season. In total, that means our data set spanned about 615 games. If we multiply this by the average number of pitches thrown in an MLB game (~150), we land at around 92,000 pitches thrown per team. We set our range from 85,000 pitches per team to 100,000 pitches per team expecting that there would be some natural fluctuation. We then checked that the max number of pitches thrown by a team was less than 100,000 and that the minimum number of pitches thrown by a team was less than 90,000. Any less and we had likely dropped a whole team's season (or seasons!) and any more would mean that we likely double counted a team's season or seasons.

```
#Testing that the number of rows we have per team match expected range

home_count = league.home_team.value_counts()
away_count = league.away_team.value_counts()

tc = TestCase()

tc.assertEqual(home_count.count(), 30)
tc.assertEqual(away_count.count(), 30)
tc.assertTrue(home_count.max() < 100000)
tc.assertTrue(away_count.max() < 100000)
tc.assertTrue(away_count.min() > 85000)
tc.assertTrue(home_count.max() > 85000)
```

Figure 9 – Unit tests written in Python to ensure data completeness when first aggregating our data source. *Tests were performed to ensure that the number of teams in the data was complete and then that the amount of total pitches for each team fell within an expected range.*

Unit Testing: Team Data Processing

To ensure that we had accurately grouped data together when getting spin rates for each stint a player had on a team, we ran a unit test to verify that at least one player had accurate data as cross referenced from Baseball Savant. It was beyond the scope of our study to verify the data from every single player, but if a random player in the middle portion of the data frame had correctly aggregated data, then we could be sure that the function was largely running as expected. We chose a random player who had played for more than one team since 2019, J.A. Happ on the Minnesota Twins, and tested his aggregated numbers against what we saw on the Baseball Savant site. This test ran after the data was compiled and ensured that minor changes that we made to the aggregating function were not fundamentally changing the way that the data was being processed. See Figure 10 below for the relevant code.

```
def test_player_stint_data_accuracy(player_stint_df):
    tc = TestCase()

    ja_happ_df = player_stint_df[player_stint_df['pitcher_id'] == 457918]
    spin_1 = int(round(ja_happ_df[ja_happ_df['team'] == 'NYY']['avg_spin_rate'],0))
    spin_2 = int(round(ja_happ_df[ja_happ_df['team'] == 'MIN']['avg_spin_rate'], 0))

    count_1 = int(ja_happ_df[ja_happ_df['team'] == 'NYY']['pitch_count'])
    count_2 = int(ja_happ_df[ja_happ_df['team'] == 'MIN']['pitch_count'])

    tc.assertEqual(spin_1, 2330)
    tc.assertEqual(spin_2, 2321)
    tc.assertEqual(count_1, 1654)
    tc.assertEqual(count_2, 853)
```

Figure 10 – Unit testing to ensure that data is being processed accurately. *This is done by cross checking the processed numbers against known numbers for a single player, J.A. Happ of the Minnesota Twins, from the Baseball Savant website.*

Conclusion

The landscape of Major League Baseball has changed over the last couple of months as officials started cracking down on pitchers' use of foreign substances, which had long been ignored throughout the league. Since the rule was first enforced on June 21st, there has been a noticeable impact on spin rates and batting statistics. After aggregating data for over 2.8 million pitches over the last four years, we were able to confirm the trend in higher RPMs in pitches which led to decreasing batting averages across the league. Looking closer at the data, we were able to see exactly how much more difficult hitting was for batters before the ban, as evidenced by rise in strikeout to hit ratios, with April 2018 representing the first month in MLB's 150-year history where there were more strikeouts than hits. The next step was investigating the effect of the ban on pitchers, where we compared pitchers before and after the ban to see if there were significant league-wide, team-wide, or individual changes.

We saw that league-wide spin rates had dropped as a whole and for every pitch that would benefit from increased spin rate. Individually, we were also able to see that the most established pitchers had their pitch RPMs drop significantly since the June 21 ban, one by as much as 8.3% (Trevor Bauer of the Los Angeles Dodgers), pointing to widespread usage of foreign substances amongst top pitchers. Analysis into teams as single units, however, yielded no significant or unified evidence of team collusion and did not suggest any team-wide conspiracies to use foreign substances to increase pitch effectiveness. Overall, we saw dramatic drops in spin rates and a corresponding rise in batting performance metrics at a league-wide, team, and individual basis since the enforcement of the ban.

Sources

Baccellieri, E. (2021, April 28). *How Worried Should MLB Be About Early Offensive Struggles?* Sports Illustrated. https://www.si.com/mlb/2021/04/28/offensive-struggles-low-batting-average-the-opener.

Blum, R. (2018, May 2). *There were more strikeouts than hits in a month for 1st time in MLB history*. Boston.com. https://www.boston.com/sports/mlb/2018/05/02/strikeouts-top-hits-in-month-for-1st-time/.

Cole, B. (2014, August 21). *Making sense of the sensors*. Beyond the Box Score. https://www.beyondtheboxscore.com/2014/8/21/6051679/statcast-pitchfx-trackman-biofx-saberseminar.

Duston, M. (2020, March 5). *What's in a Fastball: How a 4-Seamer Becomes Elite: Sports Analytics Group at Berkeley*. What's in a Fastball: How a 4-Seamer Becomes Elite | Sports Analytics Group at Berkeley. https://sportsanalytics.berkeley.edu/articles/whats-in-a-fastball.html.

Gonzalez, A., & Rogers, J. (2021, June 21). *Sticky stuff 101: Everything you need to know as MLB's foreign-substance crackdown begins*. ESPN. https://www.espn.com/mlb/story/_/id/31660574/sticky-stuff-101-everything-need-know-mlb-foreign-substance-crackdown-begins.

Janes, C. (2021, March 24). *MLB, seeking to crack down on doctored balls, turns to spin rate analysis*. The Washington Post. https://www.washingtonpost.com/sports/2021/03/24/mlb-foreign-substances-baseball-pitchers-crackdown/.

Janes, C. (2021, May 18). *Analysis | MLB's offensive woes are complicated, and they don't appear to be going away*. The Washington Post. https://www.washingtonpost.com/sports/2021/05/17/mlb-offense-complicated/.

MLB. (2021, June 15). *MLB announces new guidance to crack down against use of foreign substances, effective June 21*. MLB.com. https://www.mlb.com/press-release/press-release-mlb-new-guidance-against-use-of-foreign-substances.

MLB (2021) Statcast database. Retrieved from https://baseballsavant.mlb.com/statcast_search

Petriello, M. (2016, January 11). *The spectrum of Statcast: Spin vs. velocity*. MLB.com. https://www.mlb.com/news/statcast-spin-rate-compared-to-velocity-c160896926.

Rogers, J. (2021, March 24). *MLB memo warns teams about crackdown on use of foreign substances on baseballs*. ESPN. https://www.espn.com/mlb/story/_/id/31127060/mlb-memo-warns-teams-crackdown-use-foreign-substances-baseballs.

Sawchik, T. (2018, October 5). *Baseball's Top Staffs Have Come Around On The High-Spin Fastball*. FiveThirtyEight. https://fivethirtyeight.com/features/baseballs-top-staffs-have-come-around-on-the-high-spin-fastball/.