As a die-hard Cubs fan, I watch almost every game and have seen many different pitchers tear up the Cubs, but one specific game stood out to me during the 2021 season. In Joe Ryan's MLB debut on September 1 vs the Cubs, he was getting consistent swings and misses on center-cut 90-93 mph fastballs. My general thought would be that those pitches would result in hard-hit balls a majority of the time, but they weren't and so came the question: what makes a fastball so effective, and who throws the best fastball in the game?

When going about this question my first step was to create a ranking system and a set of criteria to find the best 4-seam fastballs for 2021. I wanted to focus on pitchers who throw their fastball as their primary pitch (greater than 40% usage rate) to dial in on how effective the fastball is on its own and not be distracted by pitchers who throw mostly offspeed pitches using their fastball as a "change of pace". I also focused on pitchers with a high sample size in 2021. I decided on 100 plate appearances as that minimum threshold as that would allow both starters and relievers to qualify while also getting rid of any "flukes" from limited innings (This excluded Joe Ryan from this specific ranking). Moving on to how to rank the pitchers, Baseball Savant has stats based on each pitcher's pitch, so I selected nine stats to evaluate each pitcher (BA, SLG, RV per 100 pitches, wOBA, Whiff %, xBA, xSLG, xwOBA, and Hard Hit %). I compared each pitcher to the league average for that stat and added a point for each incremental improvement over the league average and subtracted a point for each portion below average. For example, let's look at wOBA, league average was .314 for 2021 so if a pitcher allowed a wOBA of .300 they would receive 14 points for that category. This ranking is not exact and has a lot of problems, but it does provide a quick and fairly comprehensive comparison of each pitcher's fastball. The best fastball by almost 400 points was Josh Hader followed by Jordan Romano, Jacob deGrom, Blake Taylor, and Lucas Gilbreath.

A	B	C	E	G	H	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
ranking	last_name	first_name	pitch_name	pitches	pitch_usage	RV per 100 pitches	BA Dist	SLG Dist	wOBA Dist	Whiff Dist	xBA Dist	xSLG Dist	xwOBA Dist	Hard-Hit Dist	PTS		League Average	
1	Hader	Josh	4-Seamer	631	65.5	-2.5	-0.141	-0.24	-0.117	14.6	-0.103	-0.174	-0.085	-1.2	1043		BA	0.244
2	Romano	Jordan	4-Seamer	642	62.3	-1.8	-0.104	-0.122	-0.076	3	-0.079	-0.108	-0.065	-4.9	651		SLG	0.411
3	deGrom	Jacob	4-Seamer	704	57.4	-2.1	-0.086	-0.108	-0.095	4.1	-0.072	-0.093	-0.084	1.7	583		wOBA	0.314
4	Taylor	Blake	4-Seamer	542	72.5	-1.5	-0.028	-0.123	-0.012	-0.5	-0.059	-0.147	-0.053	-14.9	581		Whiff %	25.9
5	Gilbreath	Lucas	4-Seamer	461	63.1	-0.9	-0.057	-0.121	-0.023	0.1	-0.042	-0.098	-0.022	-17.7	550		xBA	0.24
6	Lynn	Lance	4-Seamer	1051	40.3	-1.5	-0.057	-0.114	-0.059	6.6	-0.067	-0.106	-0.075	7.2	487		xSLG	0.4
7	Peralta	Freddy	4-Seamer	1219	51.7	-1.7	-0.088	-0.103	-0.037	5	-0.064	-0.082	-0.035	3.2	444		xwOBA	0.315
8	Sewald	Paul	4-Seamer	644	58.3	-2.1	-0.064	-0.091	-0.052	7.1	-0.057	-0.052	-0.047	2	435		Hard-Hit%	38.7
9	Bednar	David	4-Seamer	521	55.8	-1.7	-0.049	-0.061	-0.045	2	-0.044	-0.081	-0.056	-6.2	435			

This top five is interesting because three types of pitchers and fastballs emerge. The first is Josh Hader in a group by himself. The next group is the two pitchers right behind Hader in the ranking (Romano, and deGrom). This group stands out for having high velocity, high spin, and high whiff rates. The final group is Taylor and Gilbreath who throw closer to league average and their effectiveness comes from limiting hard-hit rates. Hader is the most intriguing from this ranking as he scores so much higher in this metric as well as relies less on velocity than the

other members of this group. Hader's spin rate on his fastball is well below the league average, but because his spin is so efficient (98-99% active spin depending on the year) that his fastball (recently reclassified as a sinker that doesn't sink lol) has one of the best "rising" effects of any pitcher in the league. Most "flamethrowers" rely on high spin rates (or artificially create high spin rates) that allow them to pitch up in the zone, but Hader can go up there not because of the active spin that allows him to have more movement than most pitchers while having less spin. Combine that with a low three-quarters release point from the left side and you can see why Hader tops this ranking and has such an elite fastball.

Moving on to the second group, deGrom and Romano share the same characteristics with their fastball being in the top 3% of velocity (100th percentile for deGrom, 97th percentile for Romano) and top 20% of spin (87th and 83rd percentile respectively) in 2021. These characteristics lead to their fastball having less vertical break than average and missing bats by pitching up in the zone. These two also have similarities in release point and spin axis which could also factor into why their fastballs are so effective.

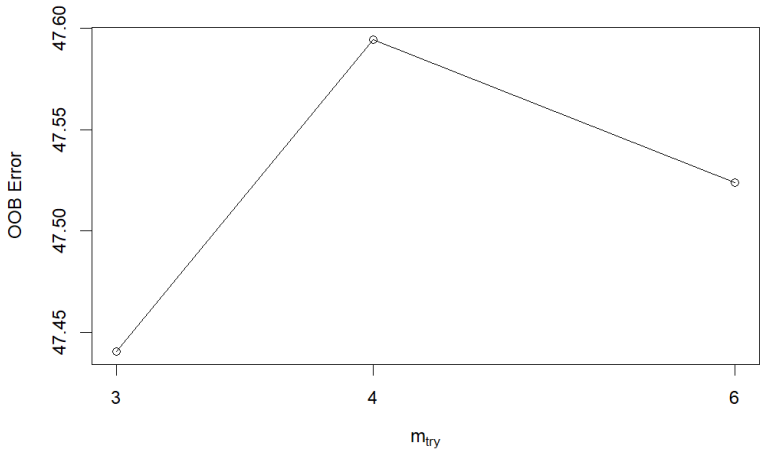
The final group, and the most confusing, is made up of two lefty relievers who are not household names in Gilbreath and Taylor. This grouping is puzzling as they are pitchers who throw league-average velocity and have average vertical movement so the indicators that we can see so far in our first three pitchers are not what is setting these pitchers apart. That makes sense on a certain level as we knew that the first three relied on whiffs to be effective while this group is more focused on limiting hard contact. The indicators that stand out with this group are horizontal movement (or lack thereof) and throwing the fastball anywhere in the zone. Gilbreath and Taylor both feature fastballs that are extremely straight and have almost no horizontal movement, league average is around 7-8 inches of arm side break. Gilbreath features a little more of a goofy motion whereas Taylor is more conventional, but both end up with similar results that stem from locating pitches and feature a different movement profile than the average pitcher.

These five pitchers have displayed different strengths when it comes to their fastball, but one overarching observation occurred to me during this analysis, pitch quality improves as it gets more unique. What I mean by that is when a pitch features something (movement, velocity, arm slot, spin axis, etc) that is far from the average and is not something that a batter sees every day it will be more effective. We saw this with Hader, deGrom, and Romano who featured velocity and vertical movement that are different than the rest of the league, and with Gilbreath and Taylor who throw abnormally straight fastballs. These differences cause deception and effective pitches. My background comes from statistics, so naturally I did not just want to leave it

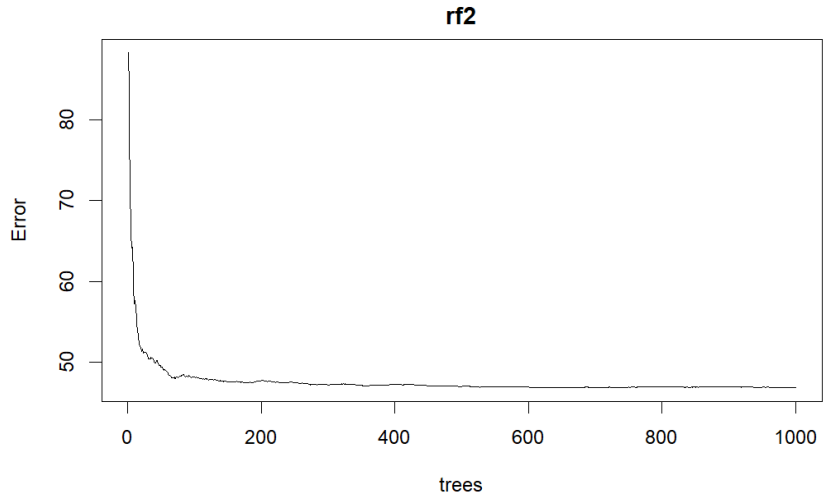
there, but I wanted to test my hypothesis and see statistically what leads to successful fastballs. I decided to look at both hard-hit rate and whiff% as response variables as those were the two big drivers of the leaders for my rankings. I also included data from 2020 and 2021 as that would give me a larger sample size and would also allow me to include active spin rate and other statistics that were introduced because of the Hawk-eye tracking system that MLB implemented starting in 2020. The independent variables that I used in the model are shown to the right, and I used two separate techniques to model each response variable, linear regression and random forest. The dataset was split into two smaller sets, 80% into the training set used to make the model and 20% into the testing set to evaluate the models. The linear regression model was tuned using an Avona type II only to use the significant variables (at the 5% level) given the other variables in the model. Beginning with the whiff% model, the final model included velocity, spin rate, vertical break (ride), release point (both x and y), and extension, this matches up almost exactly with what I observed when looking at the top 3 of the pitcher rankings. This model passed the Brown-Forsythe test, has a normal distribution of residuals, no influential points, no multicollinearity issues, and with a test RMSE of 6.87 (an improvement over the data's standard deviation of 7.718), I am confident in the effectiveness of the linear regression model. However, the R-squared value of only .251 leaves some room for improvement and reason to evaluate the random forest model.

```
ff_avg_speed
ff_avg_spin
ff_avg_break_x
ff_avg_break_z
ff_range_speed
release_pos_x
release_pos_z
effective_velocity
avg_extension
active_observed
seam_wake
lefty
```

Starting with the same independent variables, the random forest model has two main parameters that can be tuned to create a better model (mtry and ntree). The first parameter that I tuned was the mtry parameter which determines the number of variables to be randomly sampled at each split. The tuning function tests different amounts for the mtry parameter using their out-of-bag (OOB) error. The OOB error is calculated as the average error for the out-of-sample predictions for each sample grouping. The value for mtry that had the lowest OOB error was 3, shown to the left, so that will be the value used going forward.



The other parameter to be tested is the ntree value which is the number of trees to be used in the random forest model. Random forest models do not have the problem of overfitting by having too many trees because as the number of tree increase the sample mean values that are calculated within the forest will get closer to the true mean values because of the law of large numbers. That characteristic of random forest models means that to tune the ntree parameter you can increase the number of trees until the RMSE or the error plot levels out. This particular model has a lower RMSE using 1000 trees rather than the standard 500 trees and the error plot hits a plateau as the number of trees approaches 1000 indicating that increasing the number of trees further would not provide any further value to the model.



Comparing the results of the two models on the testing dataset leads to the clear conclusion that the random forest model is a better model than the linear regression model and will be used as the final model to estimate whiff percentage.

Model <chr>	R2 <dbl>	RMSE <dbl>	MAE <dbl>
Linear	0.2509991	6.870541	5.151320
RF	0.2758833	6.776230	5.077251

The hard-hit rate model follows a similar process starting with the same independent variables and using the Anova type II method to get to a reduced model that includes velocity, velocity range, x release point, effective velocity, extension, and seam shifted wake. This model failed the multicollinearity check which makes sense intuitively as the effective velocity is essentially a combination of the velocity and the extension of the pitcher. To address this issue I started building the model from scratch without effective velocity. This approach removed the multicollinearity issue but raised another issue, a model that simply was not good. The r-squared of the new model is only 0.03 and the RMSE of 10.58 is about the same as the

standard deviation of the hard-hit rate variable meaning that the model is not effective at modeling hard-hit rate whatsoever. This leads to the question of if limiting the hard-hit rate is actually a skill that is sticky year-to-year or if it is just plain luck. The hard-hit rate has been tracked since the introduction of statcast in 2015, so my first step was to collect as much data as I could for current pitchers. Baseball savant has a leaderboard that shows year-to-year changes in hard-hit rate for qualified pitchers between 2021 and 2022 (https://baseballsavant.mlb.com/leaderboard/statcast-year-to-year?group=Pitcher&type=hard_hit_percent). This covers current pitchers and is a perfect starting point for my analysis. Looking at the high-level data, the mean and median hard-hit rate over this group of pitchers has increased slightly year to year (naturally the sample size has also increased year to year as we work our way to 2022), but the trend is consistent enough that a pattern should emerge if there is a correlation. I chose to look at the years 2019-2022 as that would give me 4 seasons to compare and seek out a trend. The shortened season in 2020 leads to more outliers but the difference between the third and first quartile is similar to other years, so I am comfortable using that data. Looking at the correlation between these years you can clearly see that there is little to no correlation between the hard-hit rate a pitcher has in one year versus any other year. Comparing this trend to the whiff rate yearly correlations and you can see a clear positive correlation for each pitcher meaning that getting swings and misses is a skill that a pitcher will have whereas the hard hit rate will be more random. This validates my earlier thoughts on why the hard hit rate is difficult to model and is not fully dependent on the pitcher.

```
Hard Hit Rate Correlation
          2019      2020      2021      2022
2019 1.0000000 0.2295536 0.2437434 0.2270114
2020 0.2295536 1.0000000 0.2246250 0.2141892
2021 0.2437434 0.2246250 1.0000000 0.4120138
2022 0.2270114 0.2141892 0.4120138 1.0000000

 Whiff Rate Correlation
          2019      2020      2021      2022
2019 1.0000000 0.4940859 0.6265078 0.5237184
2020 0.4940859 1.0000000 0.6251555 0.5731421
2021 0.6265078 0.6251555 1.0000000 0.7711968
2022 0.5237184 0.5731421 0.7711968 1.0000000
```

This whole paper started by wondering why Joe Ryan was so good at getting swings and misses on average velocity. Going through this exercise did lead me to a few important conclusions, generating swings and misses will set the good pitchers apart from the great pitchers, and that a pitcher's deception (both in release and movement) joins velocity as being drivers of a great fastball.

All code and further investigation/validation provided at https://github.com/gschroeder2/MLB-Best-Fastball-2021/blob/main/R%20Code