High Speed and Spin: Hard to hit?

Walter Limm

5/7/2021

Summary

Baseball’s favorite buzzword “spin rate” burst on to the scene in 2015, when Statcast tracking was installed into each Major League stadium. Statcast tracks a massive amount of raw pitching data: pitch velocity, spin rate, spin direction, pitch movement, and more (as well as a similar amount of batted ball data and player tracking/defensive data). For a long time, velocity was the fan’s benchmark for a good fastball, but with this new information readily available, that perception has begun to shift - of course, TV broadcasts don’t display the spin rate and spin axis after each pitch in the same way velocity is shown, but “spin rate” is now in the casual sabermetric lexicon. But why do these numbers matter? I’m going to explore the connection between velocity, spin rate, and the “effectiveness” of a 4-seam fastball, the most commonly thrown pitch. Velocity is the speed the ball travels, measured in miles per hour, and spin rate is how fast the ball spins, measure in rotations per minute. “Effectiveness” will be defined as a swing and a miss, also called a “whiff”. I predict that 4-seam fastballs that are thrown faster and spin more will be more effecive. This data set includes all 4-seam fastballs thrown in the 2020 season by a pitcher who threw at least 200 pitches, to avoid including outlier appearances from a non-pitcher in a blowout game, or pitchers hampered by injuries.

```
fastballs <- read.csv("AllFSBalls.csv")
```

The tracked outcomes of each pitch are specific - they vary from “Ball” to “Called Strike” to “In play, out(s)”, among 19 total outcomes. I only care about swings and misses, a category 4 outcomes fall under: `swinging_strike` , `foul_tip` , `swinging_pitchout` , and `swinging_pitchout_blocked` . Of those 4, only the first two outcomes actually occurred, so they are the only ones I will define. This code adds the “result” observation to the dataset, defining a “foul_tip” and a “swinging_strike” as TRUE or 1 and all other outcomes as FALSE or 0.

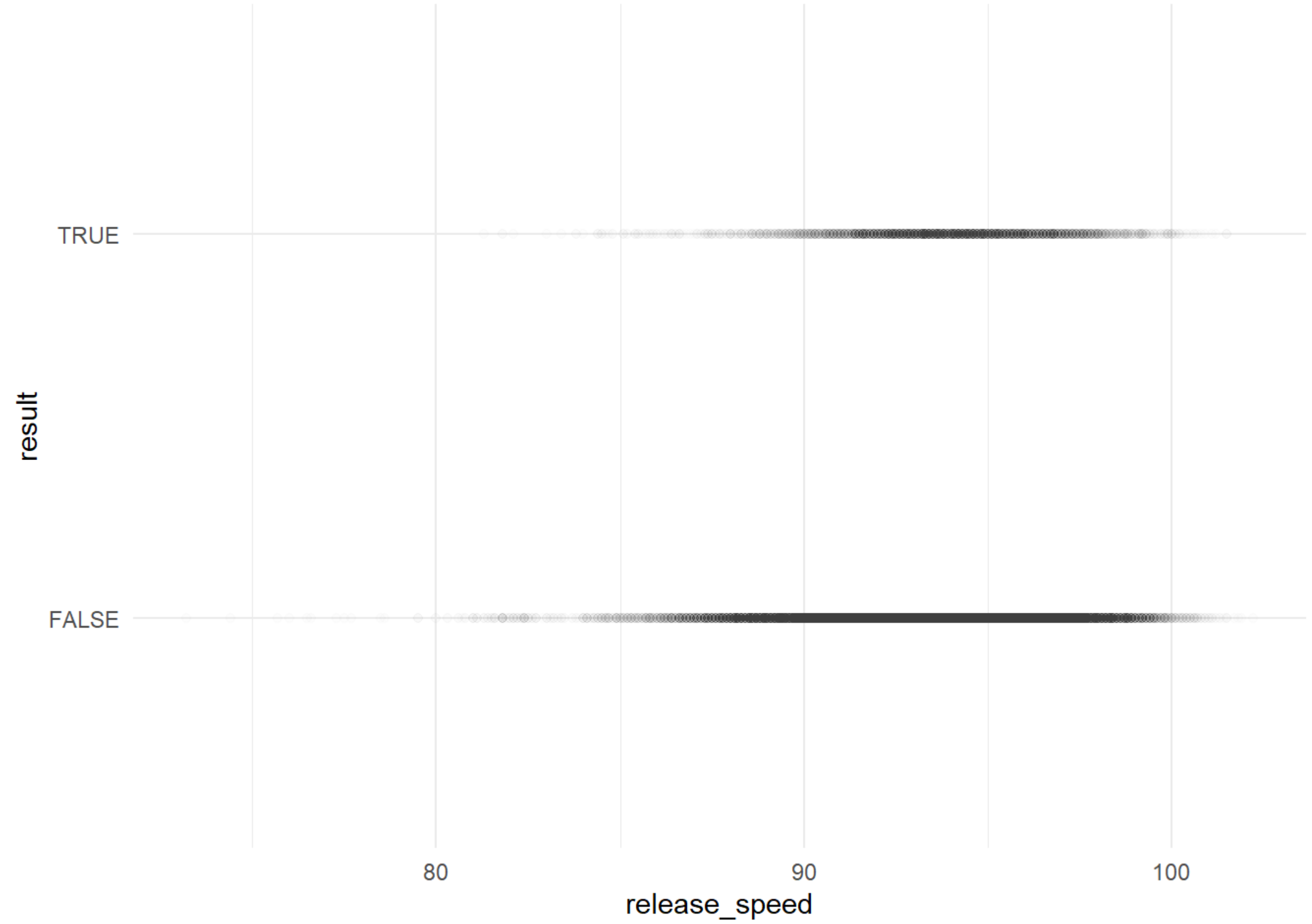
```
result <- fastballs$description == "foul_tip" | fastballs$description == "swinging_strike"

fastballs$result <- result
```

Let’s plot just velocity vs. result to get an idea of what we are looking at.

```
plot_velo <-
  fastballs %>%
  ggplot(aes(x = release_speed, y = result)) +
  geom_point(alpha = 0.0075) +
  theme_minimal()

plot_velo
```



Because of the sheer amount of data points (~40,000) it's not easy to get a sense of what's happening. There is actually a problem with how I have set up the data. If we are trying to judge whether a fastball is effective by its velocity and spin rate, we shouldn't look at all outcomes and then look how often the batter swung and missed. For example, a fastball thrown out of the strike zone will likely not be swung at because it is located poorly, not because of the spin rate or velocity. Instead, we should look at all outcomes where the batter swung, and of those see how many times they missed. This will give us a better idea of how difficult a fastball is to hit (because we only pay attention to the times batters try to hit it), and is a better definition of "effective". This new dataset includes all four-seam fastballs thrown in 2020 *that were swung at*, with the same 200 total pitch minimum.

```
fastball_swings <- read.csv("FSBallSwings.csv")
```

```
result2 <- fastball_swings$description == "foul_tip" | fastball_swings$description == "swinging_strike"

fastball_swings$result <- result2
```

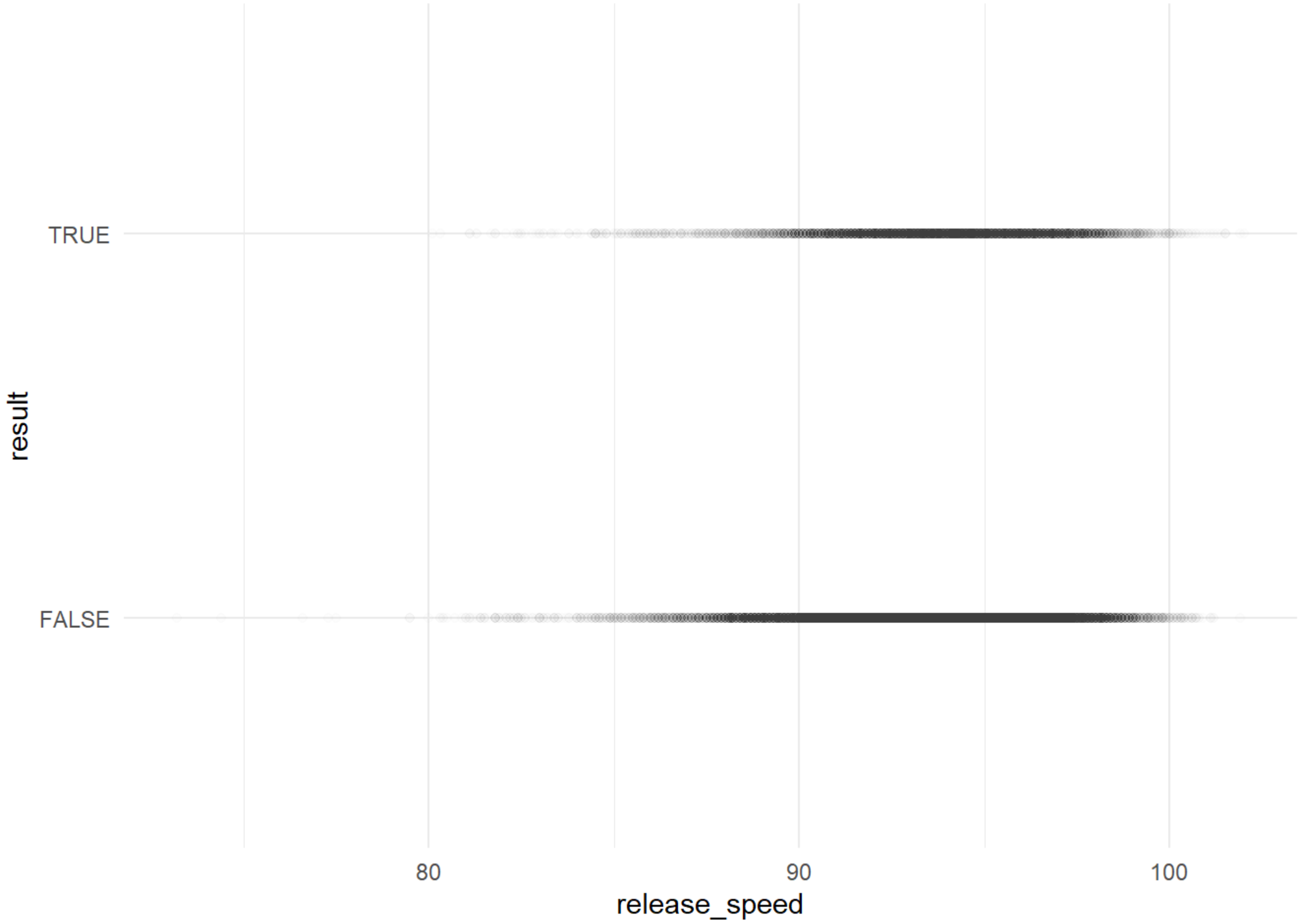
Let's also remove the observations we won't be paying attention to make looking at our data easier going forward; we don't need the 92 other variables this data has to offer

```
fastball_swings_clean <- fastball_swings[ ,c("result", "release_speed", "release_spin_rate")]

fastball_swings_clean <- fastball_swings_clean %>%
    mutate(response = ifelse(result == "TRUE", 1, 0))
```

```
plot_velo2 <-
  fastball_swings %>%
  ggplot(aes(x = release_speed, y = result)) +
  geom_point(alpha = 0.0075) +
  theme_minimal()

plot_velo2
```



I won't pretend that this is any easier to visualize, but this new data-set is better for our goals. Let's start with a logistic regression of each factor individually, before looking at the composite model. Velocity first!

```
velo_model <- glm(result ~ release_speed, family = "binomial", data = fastball_swings_clean)
```

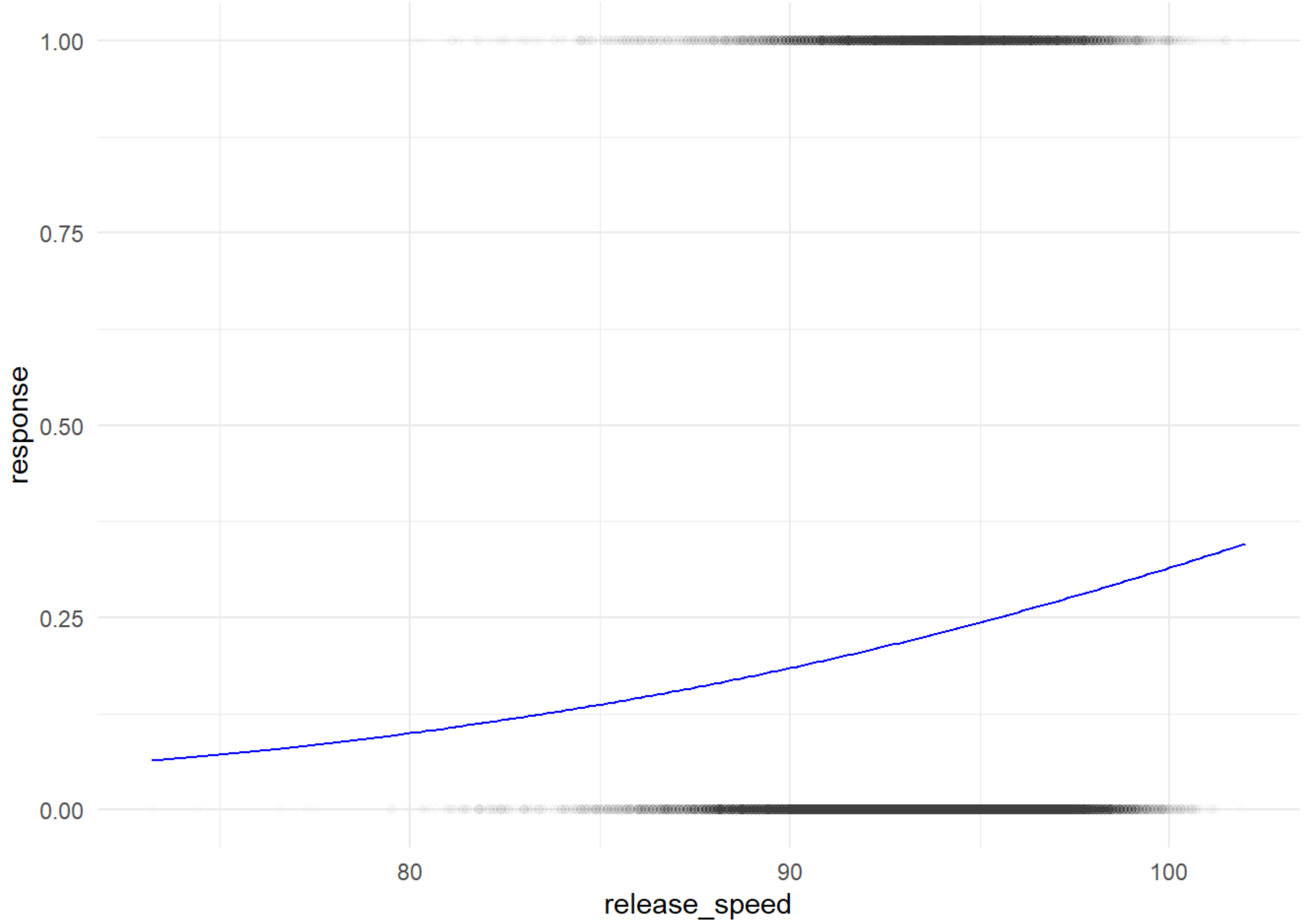
Now, I will add these predictions from the logistic regression where `release_speed` is the predictor, and graph the regression line.

```
velo_p_line <-
 fastball_swings_clean %>%
 add_predictions(velo_model) %>%
 mutate(velo_p_pred = exp(pred) / (1 + exp(pred)))
```

```
fastball_swings_clean %>%
  add_predictions(velo_model, var = "velo_p_pred", type = "response") %>% head(5) %>% kable() %
>% kable_styling()
```

result	release_speed	release_spin_rate	response	velo_p_pred
FALSE	98.2	2511	0	0.2880131
FALSE	97.0	2495	0	0.2708606
FALSE	88.6	2370	0	0.1698469
FALSE	95.8	2462	0	0.2543646
FALSE	95.4	2362	0	0.2490154

```
velo_p_line %>%
  ggplot() +
    geom_point(aes(x = release_speed, y = response), alpha = 0.0075) +
    geom_line(aes(x = release_speed, y = velo_p_pred), color = "blue") +
    theme_minimal()
```



Fast fastballs have long been considered good fastballs, and it seems like traditional baseball wisdom is fairly accurate, but there is not as much separation as one might expect. Interesting note: according this model, a fastball with a 50% whiff rate would have to travel

```
beta <- coefficients(velo_model)

log(1) - beta[1]/beta[2]
```

```
## (Intercept)
##    110.9461
```

at a blistering 110.9461 miles per hour.

Now, let’s look at spin rate

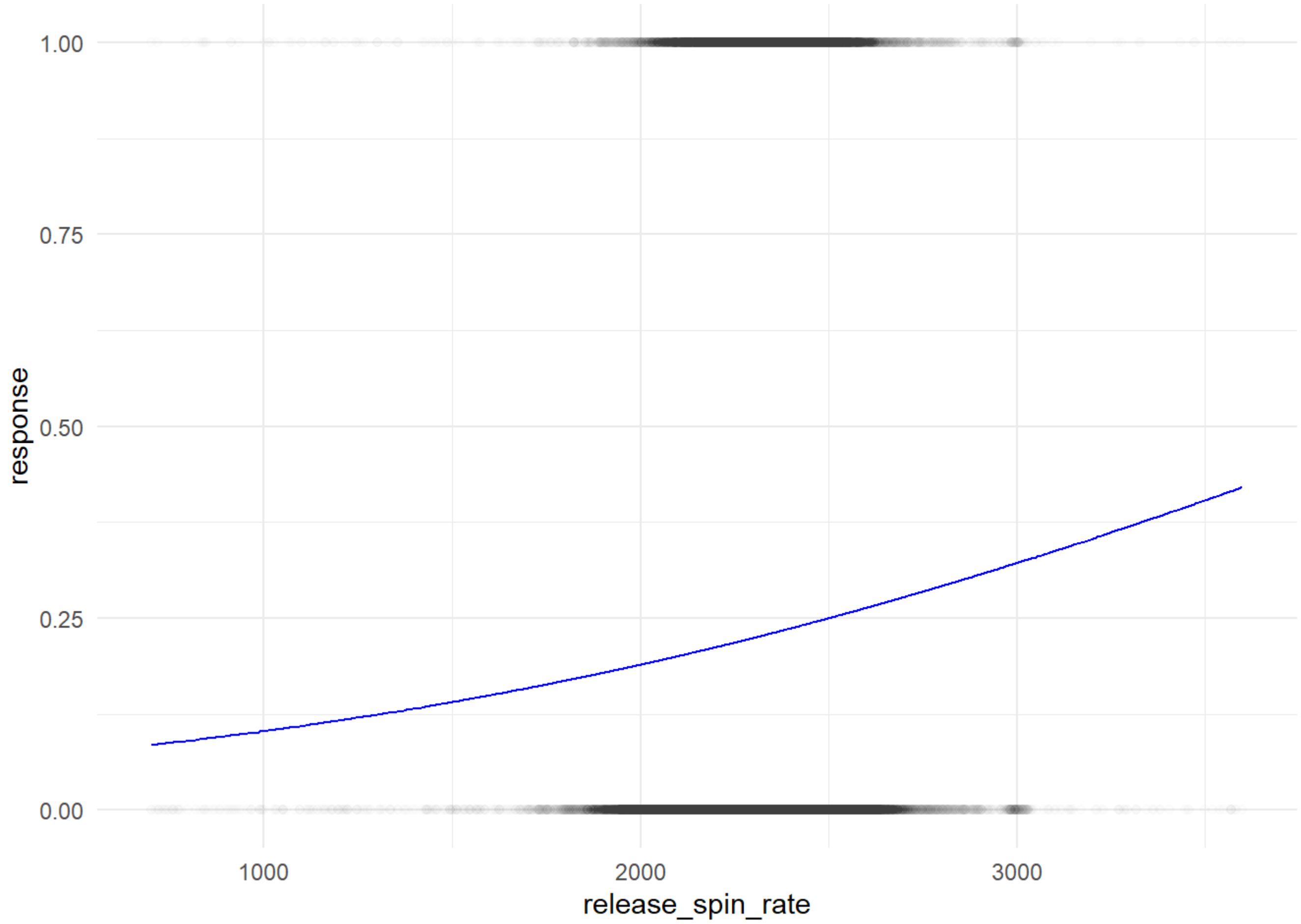
```
spin_model <- glm(result ~ release_spin_rate, family = "binomial", data = fastball_swings_clean)

spin_p_line <-
 fastball_swings_clean %>%
 add_predictions(spin_model) %>%
 mutate(spin_p_pred = exp(pred) / (1 + exp(pred)))
```

```
fastball_swings_clean %>%
  add_predictions(spin_model, var = "spin_p_pred", type = "response") %>% head(5) %>% kable() %
>% kable_styling()
```

result	release_speed	release_spin_rate	response	spin_p_pred
FALSE	98.2	2511	0	0.2512836
FALSE	97.0	2495	0	0.2491522
FALSE	88.6	2370	0	0.2329199
FALSE	95.8	2462	0	0.2447944
FALSE	95.4	2362	0	0.2319065

```
spin_p_line %>%
  ggplot() +
    geom_point(aes(x = release_spin_rate, y = response), alpha = 0.0075) +
    geom_line(aes(x = release_spin_rate, y = spin_p_pred), color = "blue") +
    theme_minimal()
```

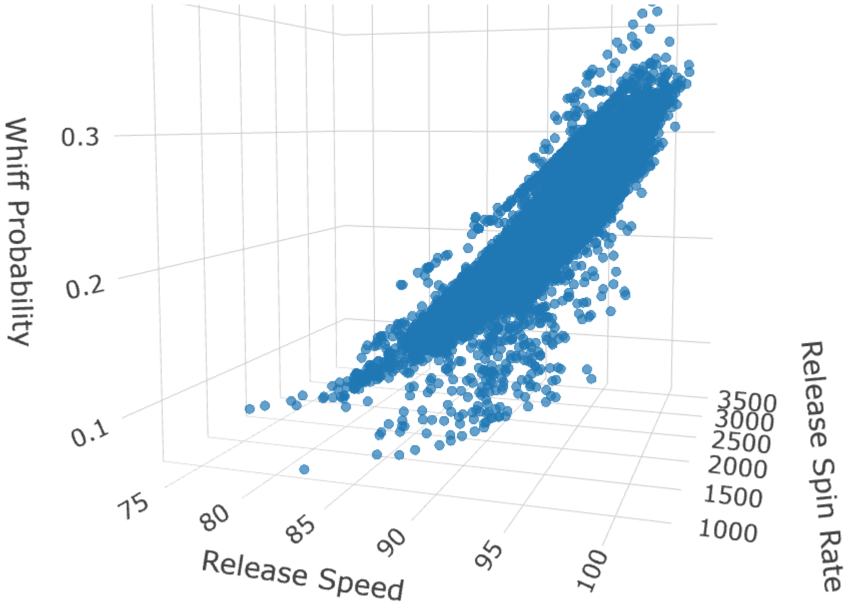


Similar with velocity, there is a slight increase in whiff probability with greater spin rate. Let’s look at them together

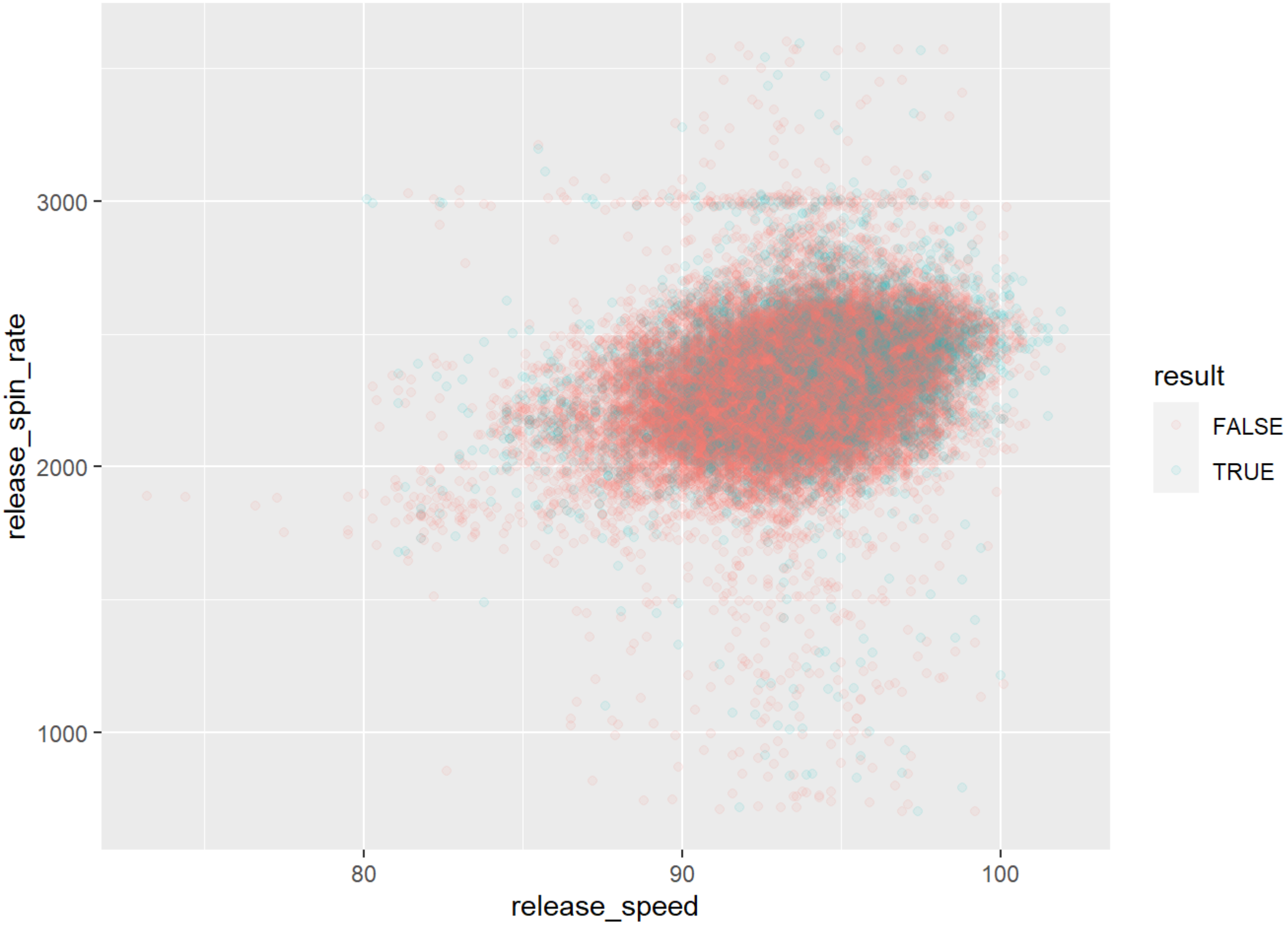
```
comp_model <- glm(result ~ release_speed + release_spin_rate, family = "binomial", data = fastball_swings_clean)
```

```
composite_data <- fastball_swings_clean %>%
  add_predictions(comp_model, var = "p_pred", type = "response")
```

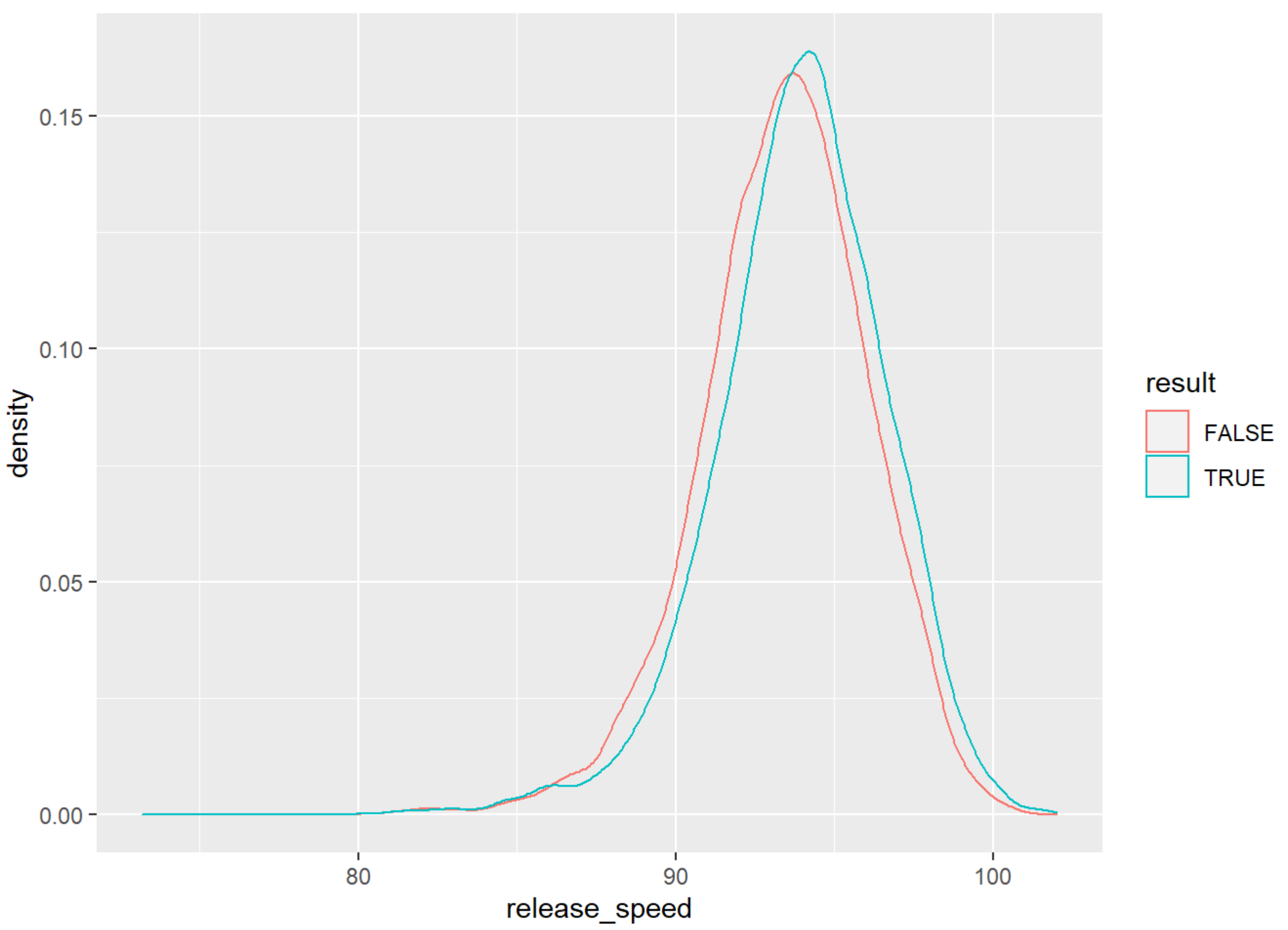
```
plot_ly(x = composite_data$release_speed, y = composite_data$release_spin_rate, z = composite_data$p_pred, type = "scatter3d", size = 10) %>%
  layout(
    scene = list(
      xaxis = list(title = "Release Speed"),
      yaxis = list(title = "Release Spin Rate"),
      zaxis = list(title = "Whiff Probability")
    )
  )
```

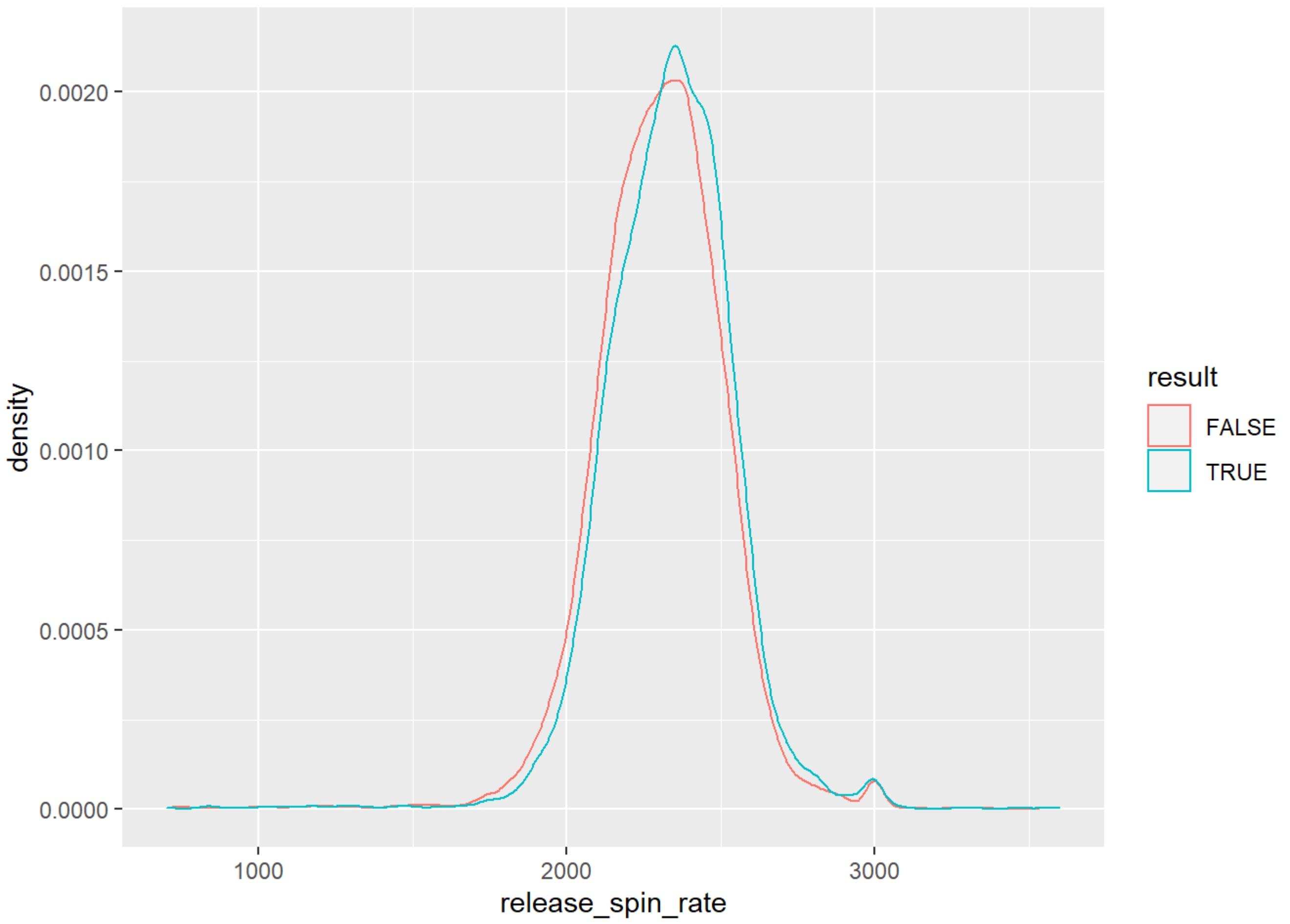


This is very interesting - the fact that no pitch has a whiff probability of over 50% could indicate that spin rate and velocity do not strongly correlate with whiff rate. We can get a better sense of this with a scatter plot where each point is colored depending on whether it resulted in a whiff or in contact

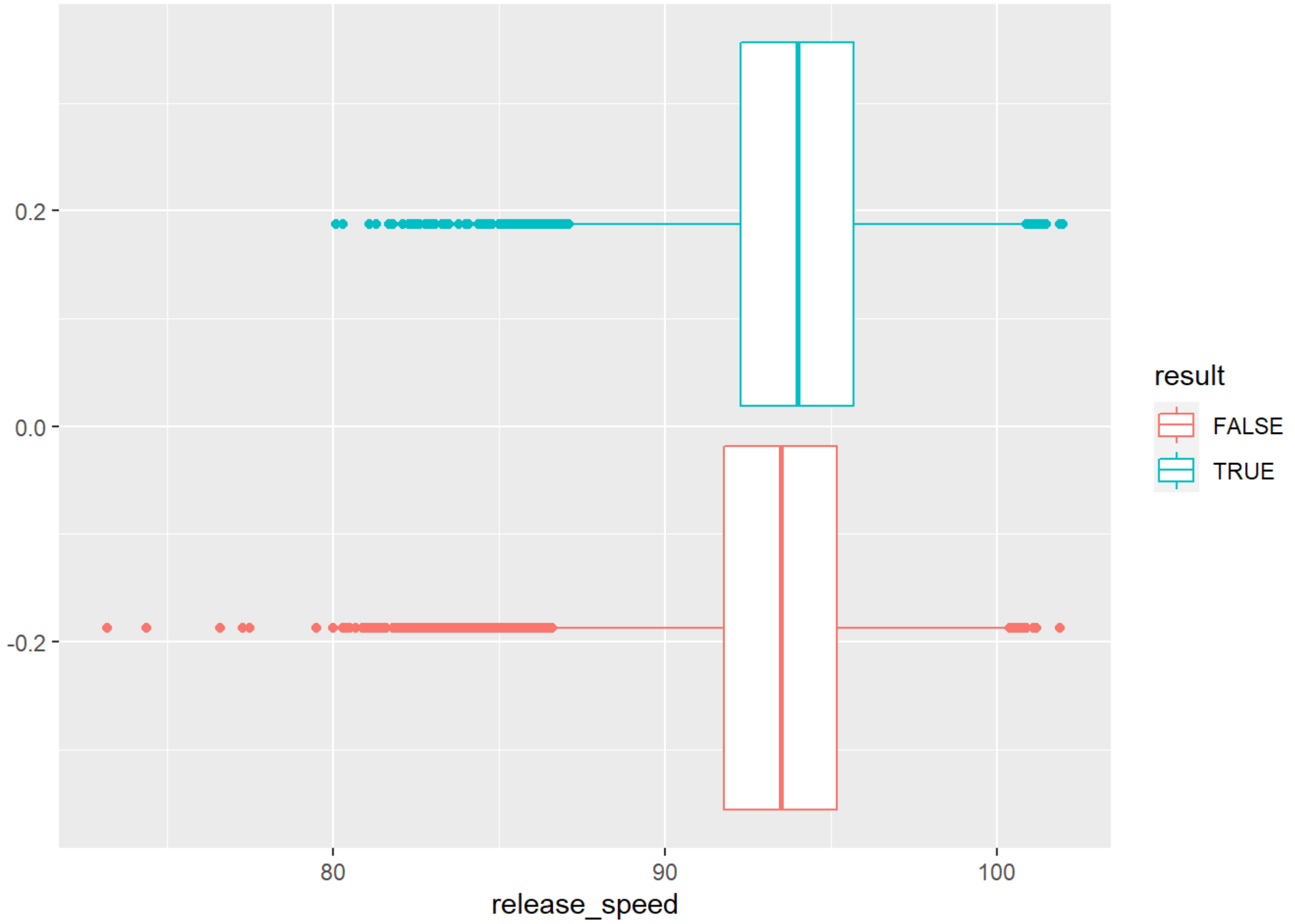


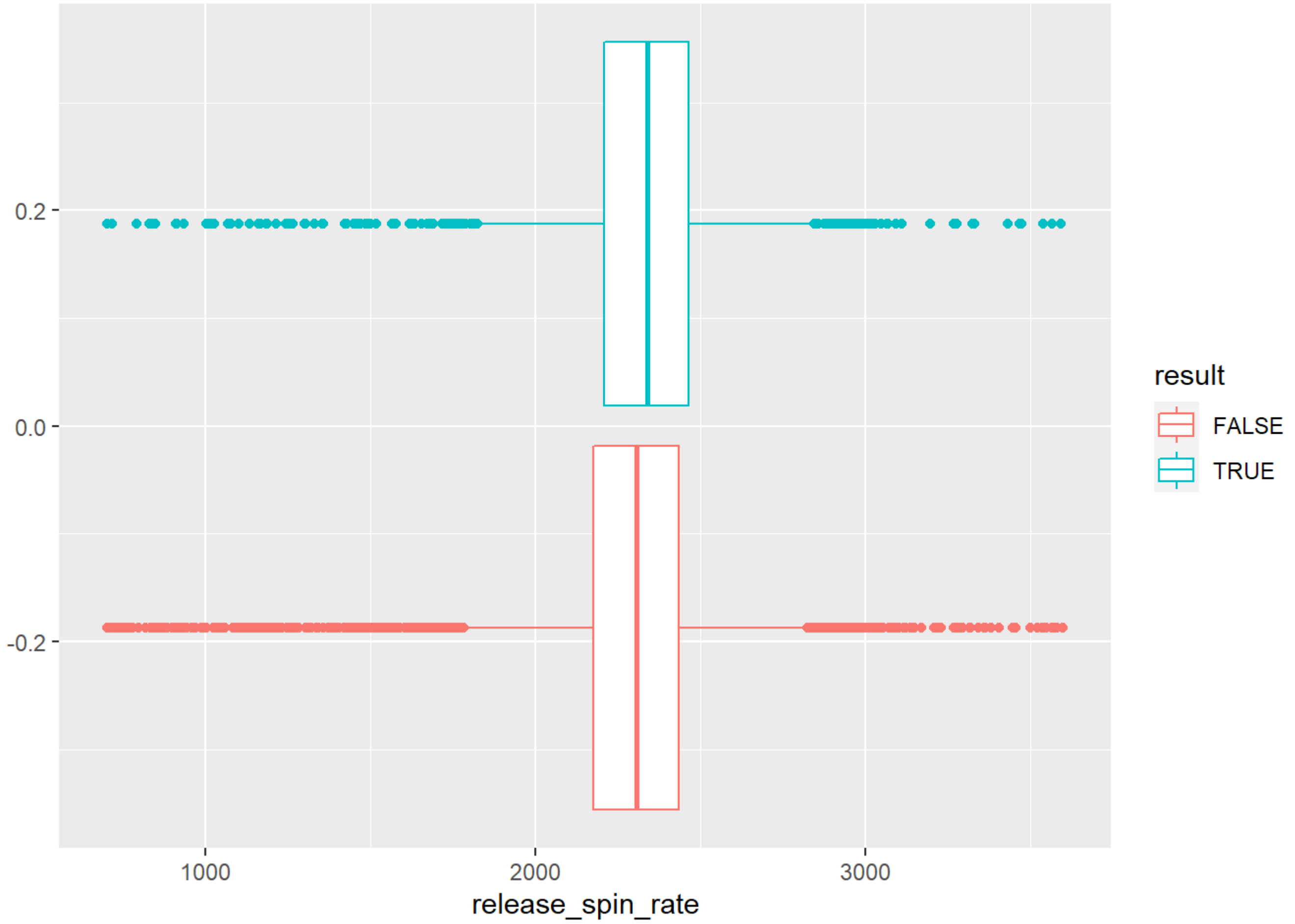
There aren't clear clusters anywhere, indicating that swinging and missing can happen at any speed or spin. This is confirmed by the kernel density plots of speed and spin rate divided by whiff vs. no contact.


0.15
0.10
0.05
0.00
density
80
90
100
release_speed
result
FALSE
TRUE


0.0020
0.0015
0.0010
0.0005
0.0000
density
1000
2000
3000
release_spin_rate
result
FALSE
TRUE

In each graph, the whiff density is shifted *slightly* to the right compared to the contact density, but they are nigh-identical. This similarity is confirmed again by boxplot graphs


0.2
0.0
-0.2
80
90
100
release_speed
result
FALSE
TRUE


0.2
0.0
-0.2
1000
2000
3000
release_spin_rate
result
FALSE
TRUE

Discussion

Why is this? This seems to go against conventional wisdom and the actions of baseball’s front offices - year after year, young pitchers who throw in the upper 90s are some of the highest draft picks, and pitcher use sticky substances to increase their spin rate (https://www.nytimes.com/2021/04/09/sports/baseball/trevor-bauer-dodgers.html# (https://www.nytimes.com/2021/04/09/sports/baseball/trevor-bauer-dodgers.html#)). Furthermore, the largest pitching contract ever, $324 million over 9 years for the New York Yankees, was signed in 2020 by Gerrit Cole, who was in the 94th percentile in both fastball velocity and fastball spin and the 87th percentile for whiff% - all elite numbers. Are baseball’s analytics departments misguided and focusing on the wrong metrics, disproved by an afternoon’s worth of work by an undergraduate student majoring in literature? I, for one, would bet on baseball, but that begs the question: what am I missing?

I predict that there are two main reasons front offices and pitchers care about velocity and spin.

The first is that, while it might not be significant, the work above does show a connection between whiffs and higher speed/spin. In a game where teams fight tooth and nail for any competitive advantage, this is entirely believable to me.

The second and more compelling reason is the idea that pitches do not exist in a vacuum. I don’t mean a literal vacuum where spin would have no bearing on movement - I mean that a 4-seam fastball is not effective if it’s the only pitch thrown. Most pitchers throw a variety of pitches, other common offerings being curveballs, sliders, changeups, and sinkers. What do all of these pitches have in common? Most often, they are slower than the fastball, and move down or laterally in some way. This motion is opposite to the fast, pseudo-rising trajectory of a 4-seamer (the pseudo-rise is because of the Magnus force - the more a 4-seam fastball spins, the more it fights the downward force of gravity). The most clear example is the fastball/changeup pairing - if the fastball is thrown at 85mph and the changeup at 83mph, the pitch lacks the “change” that its name claims to have. Similarly, most curveballs drop downward, so a high-spin fastball that drops less than expected would create more contrast. Essentially, a high-spin, high-speed pitch will often create greater separation in speed and movement compared to most secondary pitches. I hypothesize that greater contrast between pitches leads to better results.

Where do we go from here?

How can we test these predictions, specifically the idea that pitches with greater difference in speed and movement will be more effective? I think this would be difficult for a variety of reasons.

First, deciding what data to use presents an interesting challenge because pitchers don’t all throw the same pitches. Each pitcher has a unique arsenal of offerings, so deciding what pitches to compare is a puzzle - do we look at pitchers’ two most commonly used pitches? - should we only look at specific pitch pairings, like the fastball/changeup or fastball/curveball pairings discussed above (thus ignoring pitchers who don’t employ that pairing)? - do we only look pitch pairings, or do we find a way to compare the movement and velocity of a pitchers entire arsenal?

Second, our regression would likely follow a fundamentally different process (This of course depends on how we choose to compare pitches) - instead of looking at all fastballs, we would be looking at the contrast between pitches, meaning a logit model where effectiveness is defined as a whiff might not work well. We would need to define a new metric for effectiveness.

Finally, I would also predict that the contrast of velocity and movement would not be the only factors that contribute to effectiveness. Take an extreme example - let’s say a pitcher releases his fastball from one release point, and his curveball from another. No matter how extreme the contrast in movement and speed between the two pitches, a batter would likely be able to recognize which pitch is coming based on where the ball is released, mitigating the effectiveness of the contrast. Fortunately, Statcast tracks release point, but a batter can recognize/categorize a pitch at any point in it's trajectory from the pitcher's hand as it travels to the plate. This would imply, in the case of a curveball, it's not necessarily how *much* the curveball moves, it's also *when* the curveball moves. I'm describing an existing concept, caled "pitch tunneling", the idea being that different pitches should travel in the same "tunnel" as long as possible before it moves or breaks. As far as I know, Stacast does not include trajectory data. Essentially, there may be other variables that contribute to effectiveness that a study of contrast in velocity and movement of pitch arsenals would not capture.