Visualizing the Increase in MLB Strikeouts Since 2007

Michael Calabro

8/3/2020

The goal of this document is to present, in a clean fashion, the most notable tables, graphs, and plots created in my "journey to insight". Graphs and tables which need explanations will be accompanied with explanations. And for now, I think we can dive right in!

```sql
SELECT
  yearID AS Year,
  SUM(SO) AS Strikeouts
FROM batting
WHERE yearID > 2006
GROUP BY yearID

-- Assigned to variable "total_so"
```

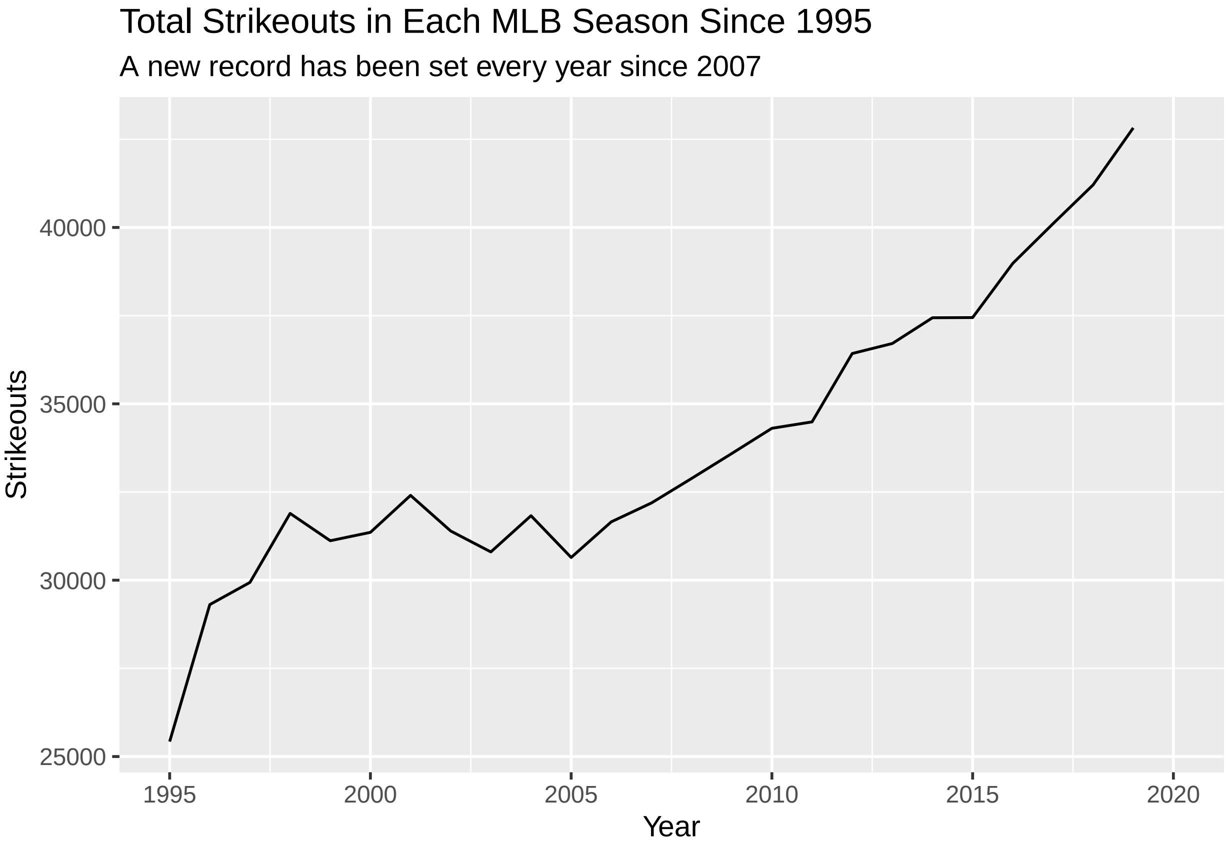
Table 1: Total Strikeouts in the MLB Per Season

Year	Strikeouts
2007	32189
2008	32884
2009	33591
2010	34306
2011	34488
2012	36426
2013	36710
2014	37441
2015	37446
2016	38982
2017	40104
2018	41207
2019	42823

This table provides a clean representation of what I wish to investigate throughout this document:

The continued increase in strikeouts in Major League Baseball since 2007.

```r
total_so %>%
  ggplot(aes(x = Year, y = Strikeouts)) +
  geom_line() +
  xlim(1995, 2020) +
  ggtitle("Total Strikeouts in Each MLB Season Since 1995",
          subtitle = "A new record has been set every year since 2007")
```



Of course, it is possible that this trend is due to an increase in at bats in this time frame. What I really want to know is if a higher *percentage* of at bats are resulting in strikeouts year over year.

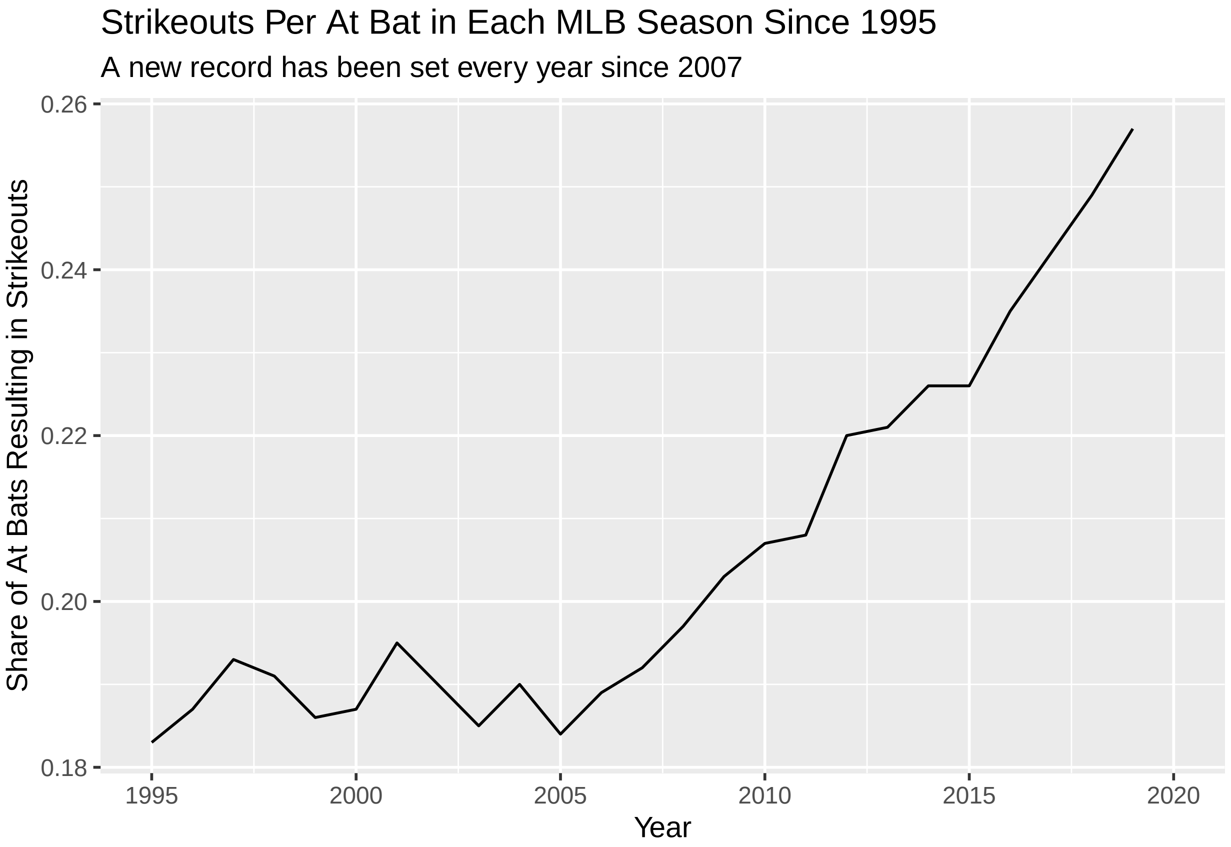
```sql
SELECT
  yearID AS Year,
  ROUND(CAST(SUM(SO) AS FLOAT) / SUM(AB), 3) AS "Strikeouts Per AB"
FROM batting
WHERE yearID > 2006
GROUP BY yearID

-- Assigned to variable strikeout_percent
```

Table 2: Strikeout Percentage in the MLB By Season

Year	Strikeouts Per AB
2007	0.192
2008	0.197
2009	0.203
2010	0.207
2011	0.208
2012	0.220
2013	0.221
2014	0.226
2015	0.226
2016	0.235
2017	0.242
2018	0.249
2019	0.257

```
strikeout_percent %>%
  ggplot(aes(x = Year, y = strikeouts_per_ab)) +
  geom_line() +
  xlim(1995, 2020) +
  ylab("Share of At Bats Resulting in Strikeouts") +
  ggtitle("Strikeouts Per At Bat in Each MLB Season Since 1995",
          subtitle = "A new record has been set every year since 2007")
```



I find it very notable that this trend began around 2005 - shortly after the release of "Moneyball", a book which emphasized a more data-driven approach to winning baseball games. Since the release of "Moneyball", by Michael Lewis, strategy in the all aspects of the MLB has slowly begun to evolve. For this reason, it is my belief that this increase in strikeouts is likely due to an evolution in both hitting and pitching strategy for MLB teams. To make it more specific, I have two theories:

1. Batters/ team batting strategies have changed to incentivize an increase in some metric(s) (i.e walks, homeruns, extra base hits, exit velocity) at the expense of more strikeouts, for the sake of more runs.

Simply: Batters changing strategy = more strikeouts

2. Pitchers/ team pitching strategies have outpaced hitting strategies in their ability to learn from the new advanced metrics.

Simply: Pitchers changing strategy = more strikeouts

So to what extent is each theory accurate? To find out, I need to come up with some investigatable hypotheses, and then utilize the data available to judge each theory's respective merit.

BATTING HYPOTHESIS #1: Some Hitting Statistic is Increasing Alongside Strikeouts

```sql
SELECT
  yearID AS Year,
  ROUND(CAST(SUM(SO) AS FLOAT) / SUM(AB), 3) AS "SO%",
  ROUND(CAST(SUM(H) AS FLOAT) / SUM(AB), 3) AS "Hit%",
  ROUND(CAST(SUM(HR) AS FLOAT) / SUM(AB), 3) AS "HR%",
  ROUND(CAST(SUM('2B') AS FLOAT) / SUM(AB), 3) AS "2B%",
  ROUND(CAST(SUM('3B') AS FLOAT) / SUM(AB), 3) AS "3B%",
  ROUND(CAST(SUM('3B') + SUM('2B') + SUM(HR) AS FLOAT) / SUM(AB), 3) AS "XBH%",
  ROUND(CAST(SUM(BB) AS FLOAT) / SUM(AB), 3) AS "Walk%"
FROM batting
WHERE yearID > 2006
GROUP BY yearID
```

Table 3: Hitting Percentages in the MLB By Season

Year	SO%	Hit%	HR%	2B%	3B%	XBH%	Walk%
2007	0.192	0.268	0.030	0.017	0.025	0.071	0.096
2008	0.197	0.264	0.029	0.017	0.025	0.071	0.098
2009	0.203	0.262	0.030	0.017	0.025	0.072	0.100
2010	0.207	0.257	0.028	0.016	0.025	0.069	0.095
2011	0.208	0.255	0.027	0.017	0.025	0.069	0.091
2012	0.220	0.255	0.030	0.017	0.026	0.072	0.089
2013	0.221	0.253	0.028	0.017	0.025	0.070	0.088
2014	0.226	0.251	0.025	0.017	0.026	0.069	0.085
2015	0.226	0.254	0.030	0.018	0.027	0.075	0.085
2016	0.235	0.255	0.034	0.018	0.027	0.079	0.091
2017	0.242	0.255	0.037	0.018	0.027	0.082	0.096
2018	0.249	0.248	0.034	0.019	0.028	0.080	0.095
2019	0.257	0.252	0.041	0.019	0.028	0.088	0.095

While an emphasis is often placed on Homeruns, it seems to me like the most notable trend is an increase in the percentage of all extra base hits (2B, 3B, and HR are all XBHs).

Furthermore, I want to see extra base hits as a share of all hits. The logic goes that players may be willing to strike out more, if it means that more of their hits are extra base hits.

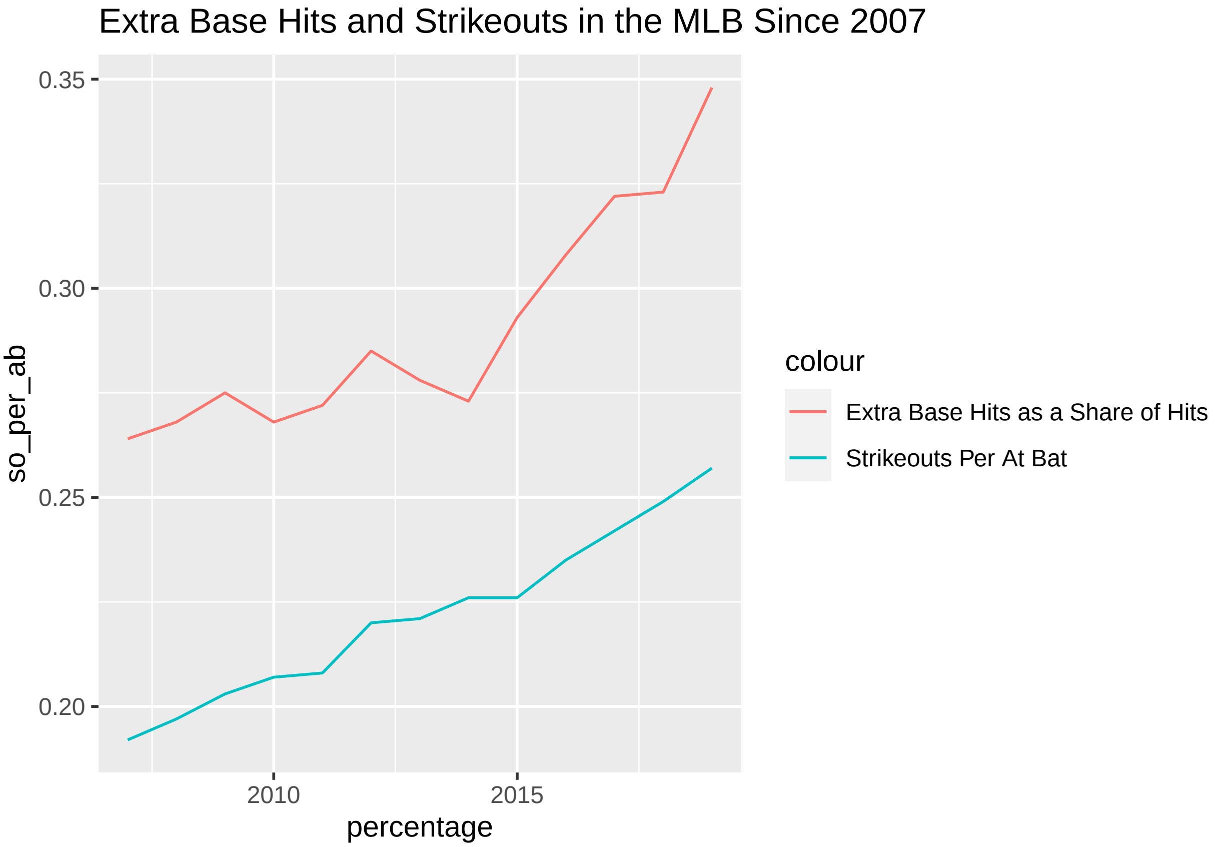
```sql
SELECT
  yearID AS Year,
  ROUND(CAST(SUM(SO) AS FLOAT) / SUM(AB), 3) AS "SO per AB",
  ROUND((SUM('3B') + SUM('2B') + SUM(HR)) / CAST(SUM(H) AS FLOAT), 3) AS "XBH per Hit"
FROM batting
WHERE yearID > 2006
GROUP BY yearID

-- Assigned to variable so_xbh
```

Table 4: Select Hitting Percentages in the MLB By Season

Year	SO per AB	XBH per Hit
2007	0.192	0.264
2008	0.197	0.268
2009	0.203	0.275
2010	0.207	0.268
2011	0.208	0.272
2012	0.220	0.285
2013	0.221	0.278
2014	0.226	0.273
2015	0.226	0.293
2016	0.235	0.308
2017	0.242	0.322
2018	0.249	0.323
2019	0.257	0.348

```
so_xbh %>%
  ggplot(aes(x = Year)) +
  geom_line(aes(y = so_per_ab, color = "Strikeouts Per At Bat")) +
  geom_line(aes(y = xbh_per_hit, color = "Extra Base Hits as a Share of Hits")) +
  xlab("percentage") +
  ggtitle("Extra Base Hits and Strikeouts in the MLB Since 2007")
```



I plan to do more statistical analysis in another document, but I think I see a pretty clear pattern here.

The next question is: why would someone want to sacrifice strikeouts for extra base hits?

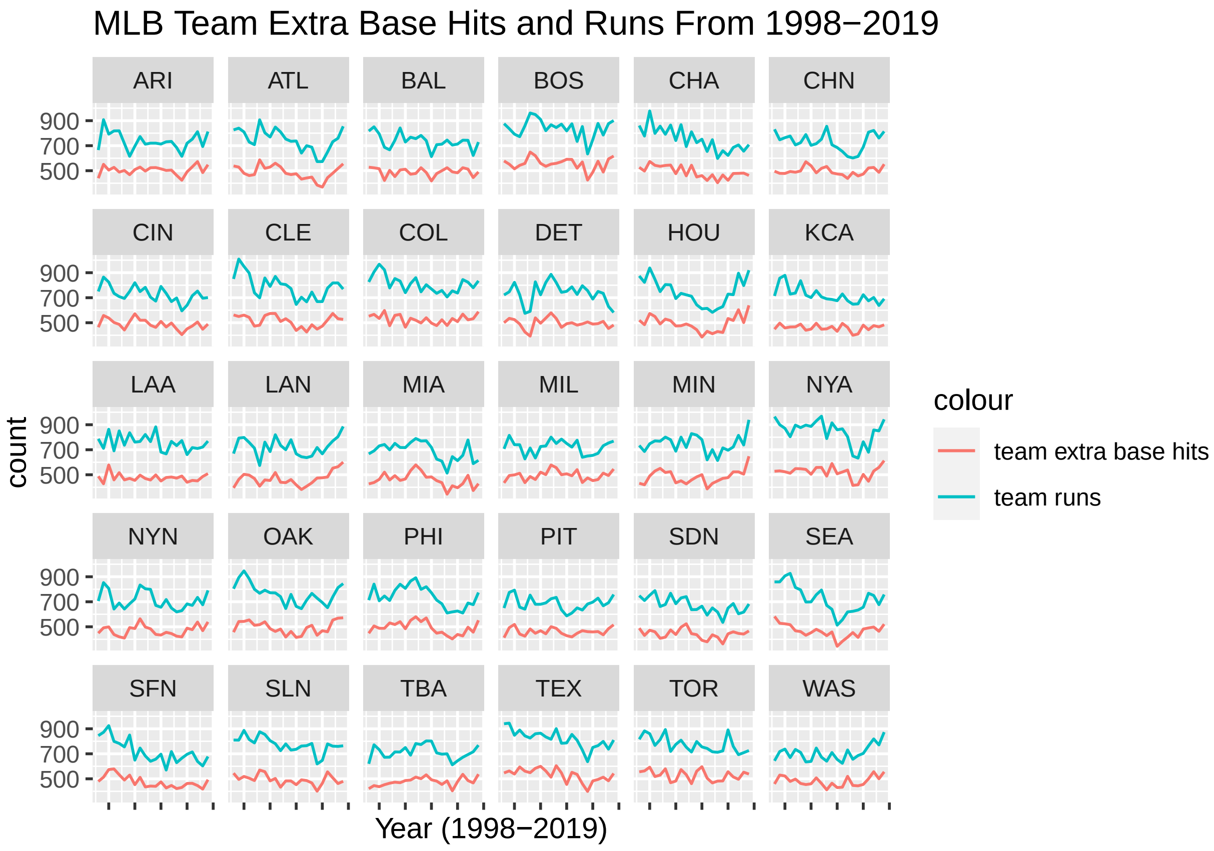
Answer: The offensive objective in baseball is to score as many runs as possible. Therefore, I would guess that team extra base hits play a bigger role in runs scored than strikeouts do. If that were the case, then teams would be incentivized to risk strikeouts for the sake of extra base hits, becasue then they would be likely to score more runs.

BATTING HYPOTHESIS #2: XBH Has a Greater Effect on Runs Scored Than Strikeouts

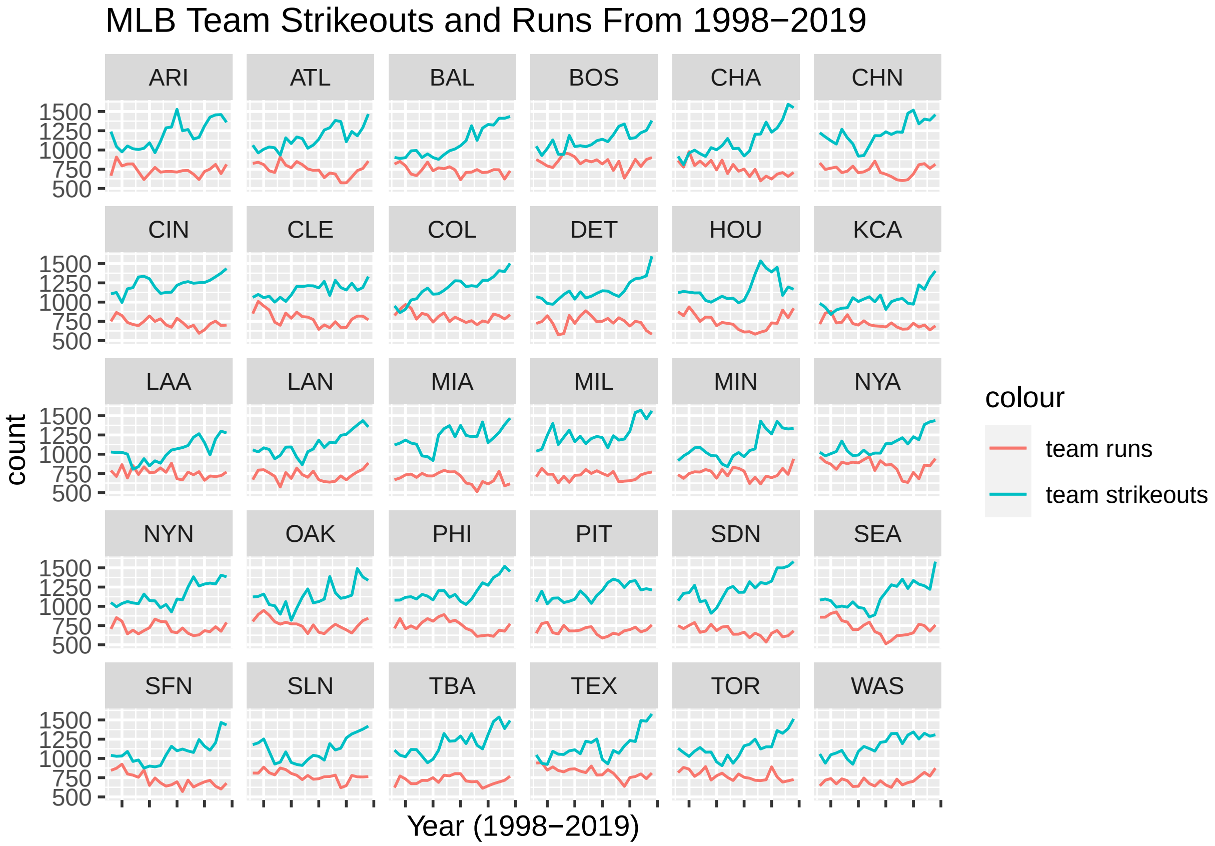
```
SELECT
  yearID,
  teamID,
  SUM(SO) AS team_strikeouts,
  SUM(R) AS team_runs,
  SUM(b.'2B') + SUM(b.'3B') + SUM(HR) AS team_xbh
FROM batting AS b
WHERE yearID > 1997
GROUP BY yearID, teamID

-- Assigned to variable "team_xbh_so_runs"
```

```
team_xbh_so_runs %>%
  # Mutate combines teams that moved cities in this time frame into the same plot
  mutate(teamID = ifelse(teamID == "ANA", "LAA",
                       ifelse(teamID == "MON", "WAS",
                              ifelse(teamID =="FLO", "MIA", teamID)))) %>%
  ggplot(aes(x = yearID)) +
  geom_line(aes(y = team_runs, color = "team runs")) +
  geom_line(aes(y = team_xbh, color = "team extra base hits")) +
  facet_wrap(~ teamID) +
  theme(axis.text.x=element_blank()) +
  xlab("Year (1998-2019)") +
  ylab("count") +
  ggtitle("MLB Team Extra Base Hits and Runs From 1998-2019")
```



```
team_xbh_so_runs %>%
  # Mutate combines teams that moved cities in this time frame into the same plot
  mutate(teamID = ifelse(teamID == "ANA", "LAA",
                         ifelse(teamID == "MON", "WAS",
                                ifelse(teamID =="FLO", "MIA", teamID)))) %>%
  ggplot(aes(x = yearID)) +
  geom_line(aes(y = team_runs, color = "team runs")) +
  geom_line(aes(y = team_strikeouts, color = "team strikeouts")) +
  facet_wrap(~ teamID) +
  theme(axis.text.x=element_blank()) +
  xlab("Year (1998-2019)") +
  ylab("count") +
  ggtitle("MLB Team Strikeouts and Runs From 1998-2019")
```



To my eye, it seems fairly clear that runs follow trend with extra base hits much more closely than they do with strikeouts. For that reason, it seems to make sense that coaches would want their hitters attempting to hit more extra base hits, even if that means striking out more often.

That being said, it is also fairly clear that nearly every team is striking out MUCH more in recent years, even if they aren't seein the accompanying rise in xbh and runs. This tells me that improved pitching across the MLB may also be playing a role in these crazy strikeout numbers.

Table 5: The Effect of Extra Base Hits and Strikeouts on Team Runs

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	215.8179296	20.3744664	10.59257	0
xbh	1.4031362	0.0333075	42.12676	0
so	-0.1375576	0.0104045	-13.22103	0

```
runs = team_xbh_so_runs$team_runs
xbh = team_xbh_so_runs$team_xbh
so = team_xbh_so_runs$team_strikeouts

summary(lm(runs ~ xbh + so))$coefficients %>%
  kableExtra::kable(caption = "The Effect of Extra Base Hits and Strikeouts on Team Runs") %>%
  kableExtra::kable_styling(full_width = FALSE)
```

```
summary(lm(runs ~ xbh + so))$r.squared
```

```
## [1] 0.7481196
```

As you can see from the regression, an increase in 1 extra base hit leads to an expected 1.4 runs added in a season, while an increase in 1 strikeout leads to an expected .137 runs lost. With this in mind, it would make sense that, if a team/player could get some more extra base hits at the expense of striking out more, they would choose to do so.

Now I am going to download some new data that was not in my "journey to insight", but which will provide clearer information for the advanced data that I attempted to visualize in that document. In my "journey to insight", I downloaded data from a "Leaderboard" page on Fangraphs.com, which compiled season stats for the best ~150 players in each season.

Instead, I will now be using pitch-by-pitch data scraped from the mlb's "Baseball Savant" website using the R package "baseballr". There are many data points(~4,000 per day), So I am going to select a week of the year and gather data from that week in every year since 2007. Not only will this data be more specific than the data from Fangraphs, but I will also get true exit veloicty data for batted balls since 2015 (The year they started keeping that stat).

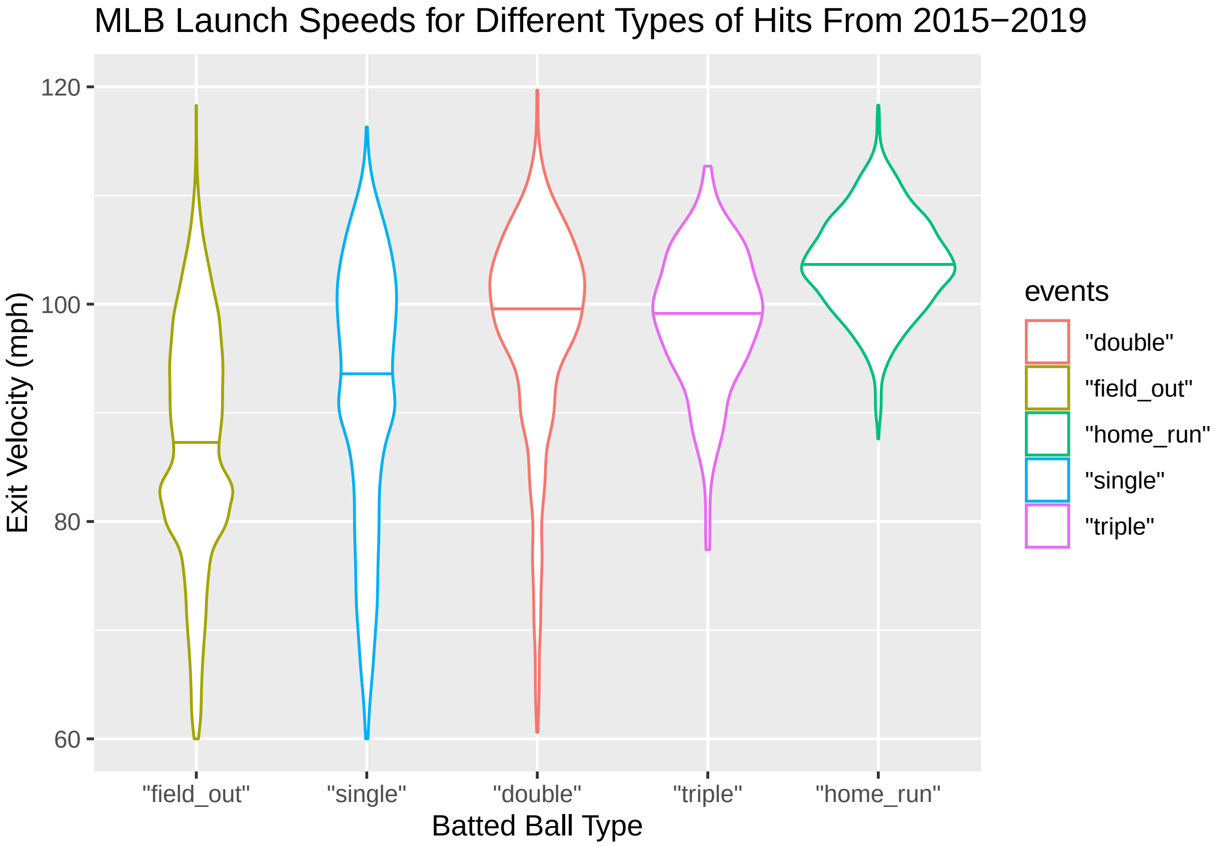
My process of gathering this data is shown in my "data playground", which is where I (messily) experiment with certain code. I turned the data into a csv so that it wouldn't take the rmarkdown forever to perform the functions every time I knit it.

BATTING HYPOTHESIS #3: Higher Exit Velocity = More Extra Base Hits

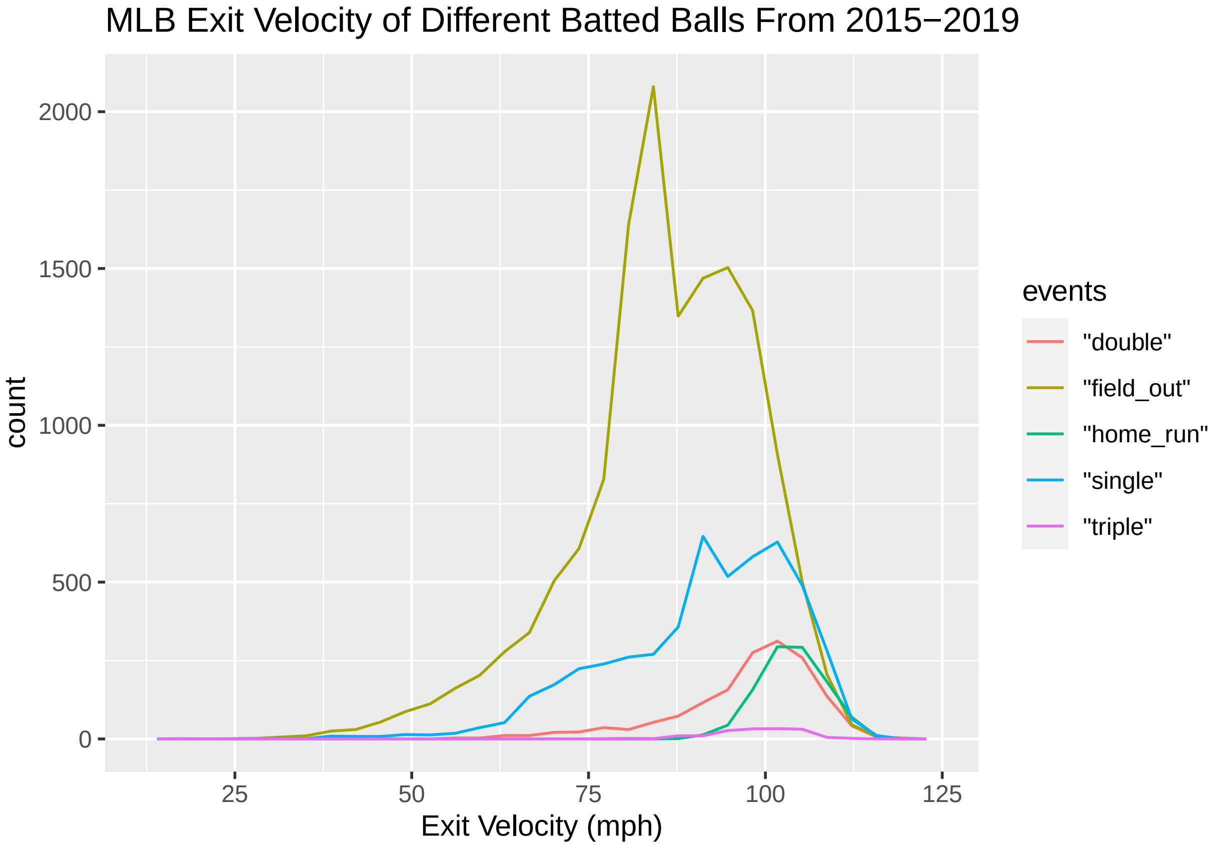
```
# Reading in the new csv data
# Clean names function from janitor fixed a glitch with the column names
library(janitor)

full_savant_data <- read_csv("full_savant_data.csv", col_names = TRUE) %>%
  clean_names()

event_list <- c('"field_out"', '"single"', '"double"', '"triple"', '"home_run"')
full_savant_data %>%
  filter(launch_speed > 0) %>%
  filter(events %in% event_list) %>%
  ggplot(aes(x = factor(events, level = event_list))) +
  geom_violin(aes(y = launch_speed, color = events),
              draw_quantiles = 0.5) +
  ylim(60, 120) +
  ylab("Exit Velocity (mph)") +
  xlab("Batted Ball Type") +
  ggtitle("MLB Launch Speeds for Different Types of Hits From 2015-2019")
```



```r
full_savant_data %>%
  filter(launch_speed > 0) %>%
  filter(events %in% event_list) %>%
  ggplot(aes(x = launch_speed)) +
  geom_freqpoly(aes(color = events)) +
  xlab("Exit Velocity (mph)") +
  ggtitle("MLB Exit Velocity of Different Batted Balls From 2015-2019")
```



It seems very clear that the slower the exit velocity of a batted ball, the more likely it is to be a single, or an out. As you hit the ball harder, your odds of getting a double, triple, or homerun increase.

As a player seeing this data, the message is clear: if you can hit the ball harder, you're more likely to get a hit, and you're more likely to help your team score runs and win. How can you hit the ball harder? By swinging harder, of course.

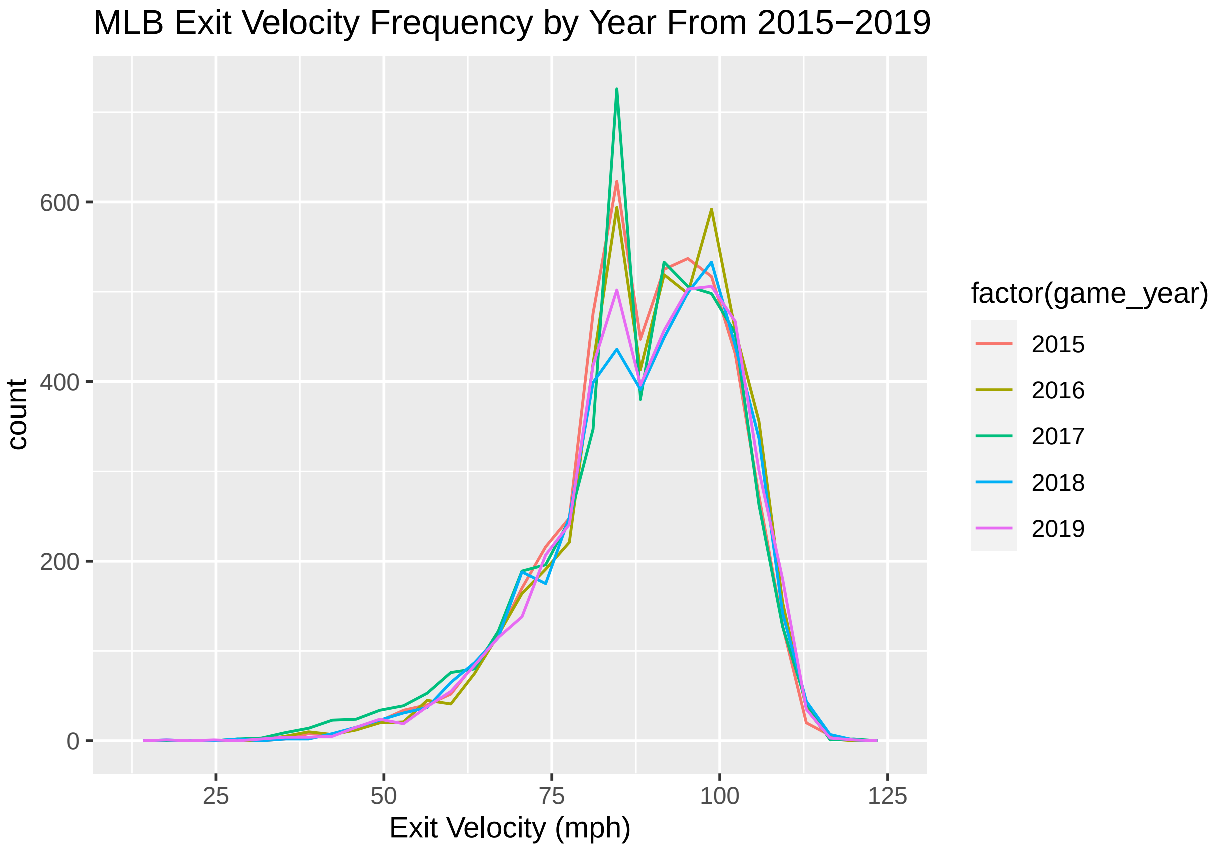
What typically hinders a player from swinging harder is the knowledge that they'll be less likely to make contact with the ball. However, with the knowledge that an extra base hit is more valuable than a whiff is detrimental, it would make sense to start putting more effort into every swing.

BATTING HYPOTHESIS #4: The Frequency of Low Exit Velo Batted Balls is Decreasing

I propose that, if my thought is correct, batters are moving towards replacing the balls hit ~80-90 mph with either balls hit harder, or strikeouts. Therefore, we should expect to see less of the soft hit outs/singles.

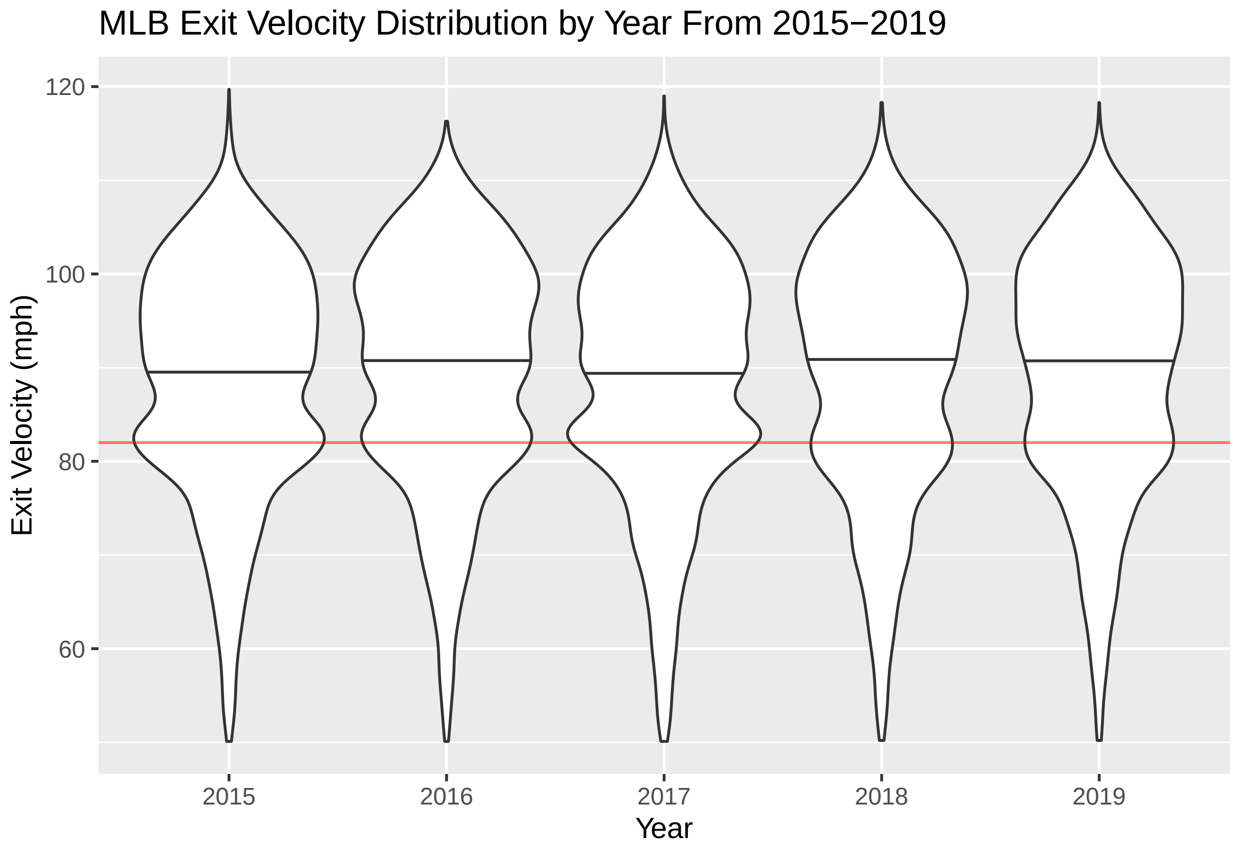
.

```r
full_savant_data %>%
  filter(launch_speed > 0) %>%
  ggplot(aes(x = launch_speed)) +
  geom_freqpoly(aes(color = factor(game_year))) +
  xlab("Exit Velocity (mph)") +
  ggtitle("MLB Exit Velocity Frequency by Year From 2015-2019")
```



It would be helpful to have data for this dating back to 2008, but instead we only have 2015-2019. With the data that we do have, it is quite clear that the soft-hit ball frequency has fallen pretty dramatically in 2018/2019. However, the peak for 2017 was the highest of the 5 years, which negates the idea that there is a constant trend.

```
full_savant_data %>%
  filter(launch_speed > 50) %>%
  ggplot(aes(x = factor(game_year))) +
  geom_violin(aes(y = launch_speed), draw_quantiles = 0.5, scale = "count") +
  geom_hline(yintercept = 82, color = "red", alpha = 0.5) +
  xlab("Year") +
  ylab("Exit Velocity (mph)") +
  ggtitle("MLB Exit Velocity Distribution by Year From 2015-2019")
```



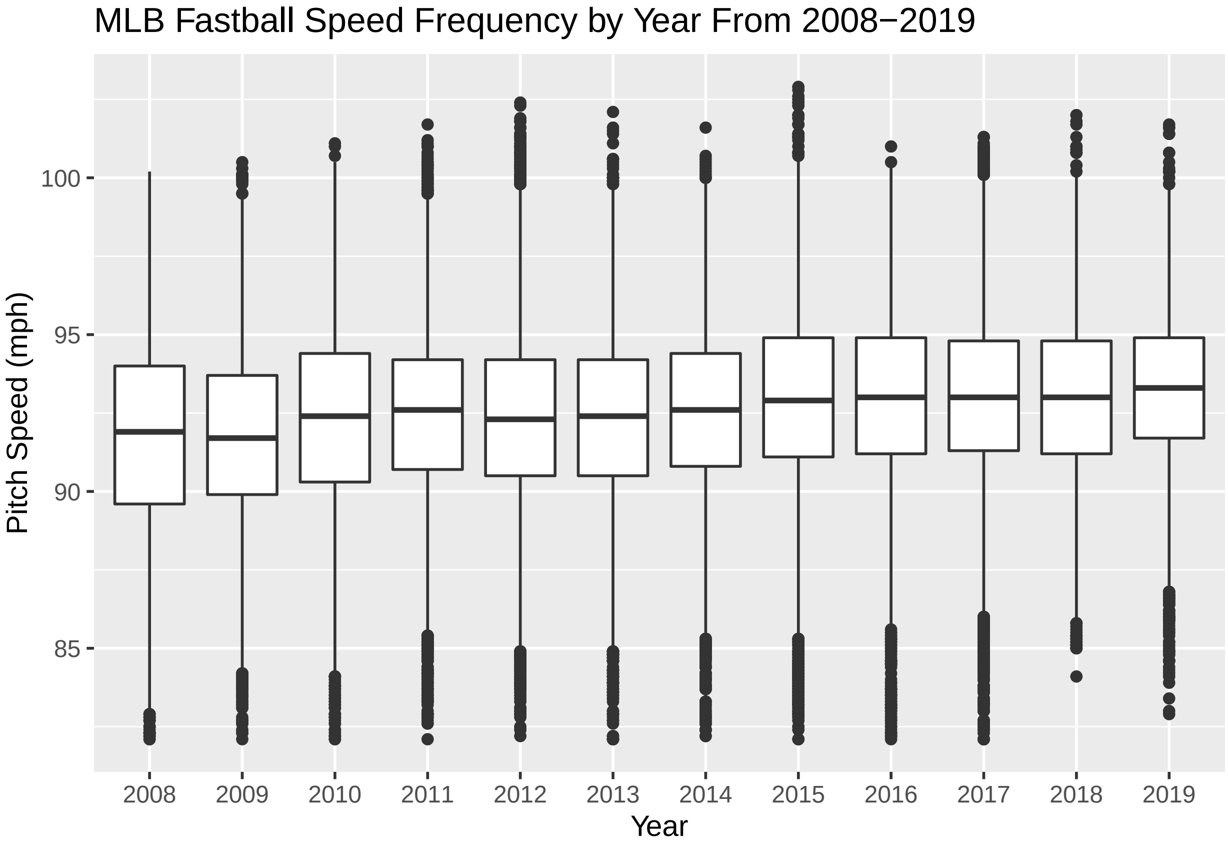
I predict that, in the coming years, the bump which occurs around the red line for the exit velocity violin plots will get thinner and thinner as batters and managers continue to realize that making soft contact is a suboptimal strategy, and that striking out isn't bad if it comes with a higher chance of hitting the ball harder.

One more thing I want to mention is that it makes sense why an individual would rather hit a soft flyout or groundout than get struck out. Quite often, a strikeout SEEMS and FEELS like a much worse outcome than a weak groundout, even though, most of the time, they both are simply equivalent to 1 out. As players, coaches, and managers begin to trust the data more, I believe the shame that comes with striking out will continue to go away.

OKAY! I think I have exhausted the batting hypotheses. Let's look at the roll pitching has played in the increase in strikeouts in the MLB

PITCHING HYPOTHESIS 1: Pitchers Are Throwing Harder Over Time

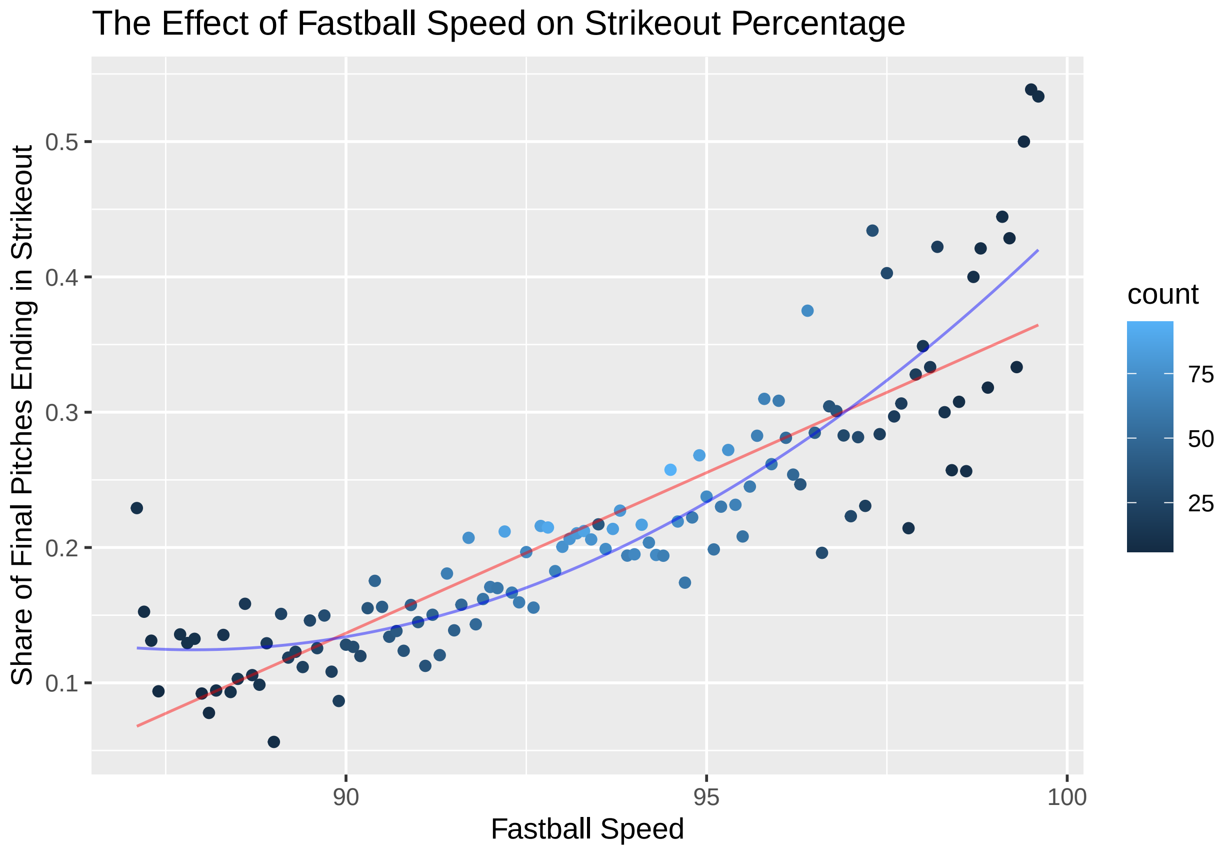
```
full_savant_data %>%
  filter(pitch_type == '"FF"') %>% #FF = Four seam Fastball
  filter(release_speed > 82) %>%
  ggplot(aes(x = factor(game_year))) +
  geom_boxplot(aes(y = release_speed)) +
  ylab("Pitch Speed (mph)") +
  xlab("Year") +
  ggtitle("MLB Fastball Speed Frequency by Year From 2008-2019")
```



Pitch Velocity has been generally rising (slowly) over the past decade, but not by much.

```
fastball_outcomes <- full_savant_data %>%
  filter(pitch_type == '"FF"') %>%
  filter(release_speed > 87) %>%
  filter(events %in% c('"field_out"', '"single"', '"double"', '"triple"',
                       '"home_run"', '"walk"', '"strikeout"')) %>%
  group_by(release_speed, events) %>%
  summarise(count = n()) %>%
  group_by(release_speed) %>%
  mutate(share = count / sum(count)) %>%
  filter(events == '"strikeout"') %>%
  filter(count > 5)

fastball_outcomes %>%
  ggplot(aes(x = release_speed, y = share)) +
  geom_point(aes(color = count)) +
  stat_smooth(geom = "line", method = "lm", se = FALSE, color = "red", alpha = 0.45) +
  stat_smooth(geom = "line", method = "lm", formula = "y ~ poly(x, 2)", se = FALSE, color = "blue", alph
  xlab("Fastball Speed") +
  ylab("Share of Final Pitches Ending in Strikeout") +
  ggtitle("The Effect of Fastball Speed on Strikeout Percentage")
```



The polynomial regression (blue line) seems to fit the data slightly better than the linear regression (red line). However, for the purposes of interpretaion and clarity, I am going to create a linear model on the next page.

Table 6: The Effect of Fastball Speed on Strikeout Percentage

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.998618	0.1176577	-16.98672	0
x	0.023726	0.0012588	18.84776	0

```
y = fastball_outcomes$share
x = fastball_outcomes$release_speed

summary(lm(y ~ x))$coefficients %>%
  kableExtra::kable(caption = "The Effect of Fastball Speed on Strikeout Percentage") %>%
  kableExtra::kable_styling(full_width = FALSE)
```

```
summary(lm(y ~ x))$r.squared
```

```
## [1] 0.7459254
```

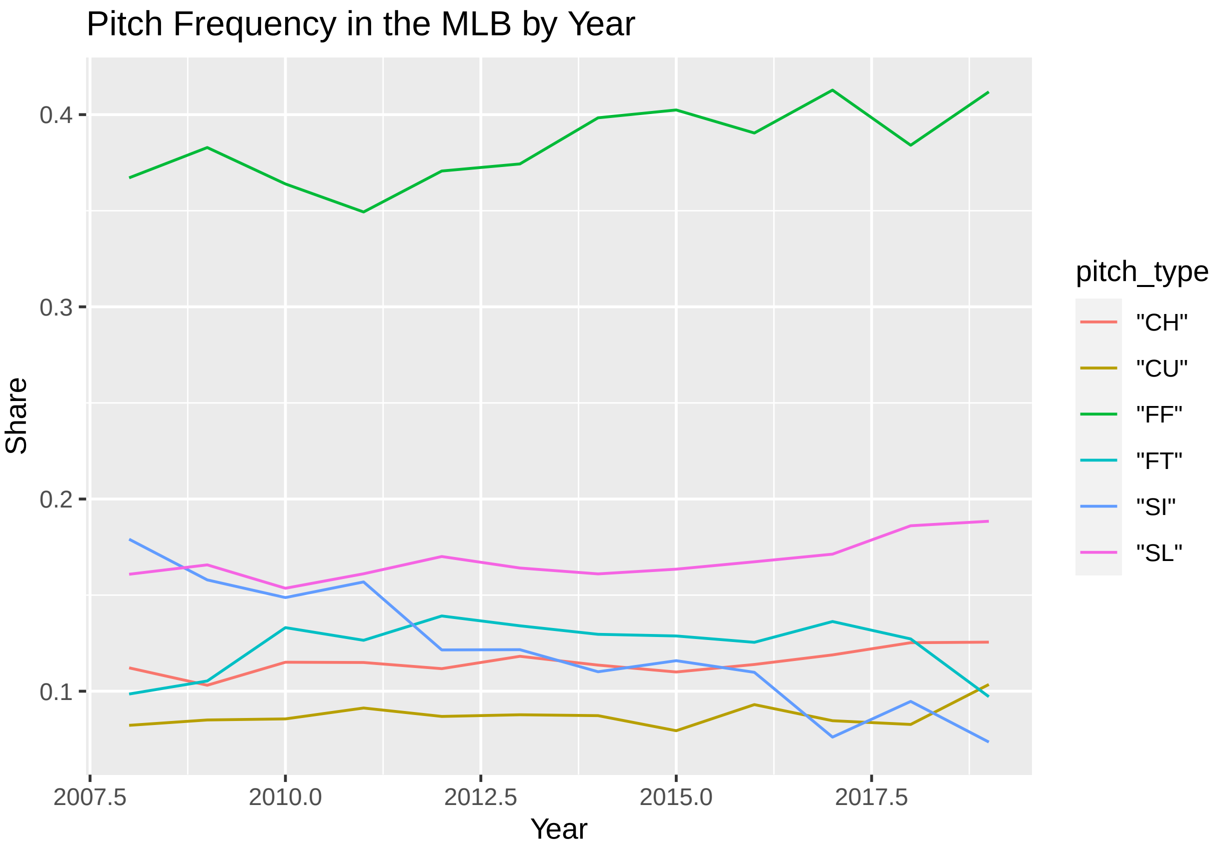
According to the model, every increase in 1 mph of a fastball should result in a 2.3% higher chance of getting a strikeout for the final pitch of an at bat. Therefore, 1 mph increase we have seen over the past decade could very well be contributing to the rise in strikeouts.

On the next page, I will be visualizing the usage of different pitches over the past decade.

```
key_pitches <- c('"SL"', '"CU"', '"CH"', '"FF"', '"SI"', '"FT"')
# SL - Slider
# CU - Curveball
# CH - Changeup
# FF - Four-Seam Fastball
# SI - Sinker
# FT - Two-Seam Fastball

full_savant_data %>%
  filter(pitch_type %in% key_pitches) %>%
  group_by(game_year, pitch_type) %>%
  summarise(count = n()) %>%
  group_by(game_year) %>%
  mutate(Share = count / sum(count)) %>%
  ggplot(aes(x = game_year)) +
  geom_line(aes(y = Share, color = pitch_type)) +
  xlim(2008, 2019) +
  xlab("Year") +
  ggtitle("Pitch Frequency in the MLB by Year")
```

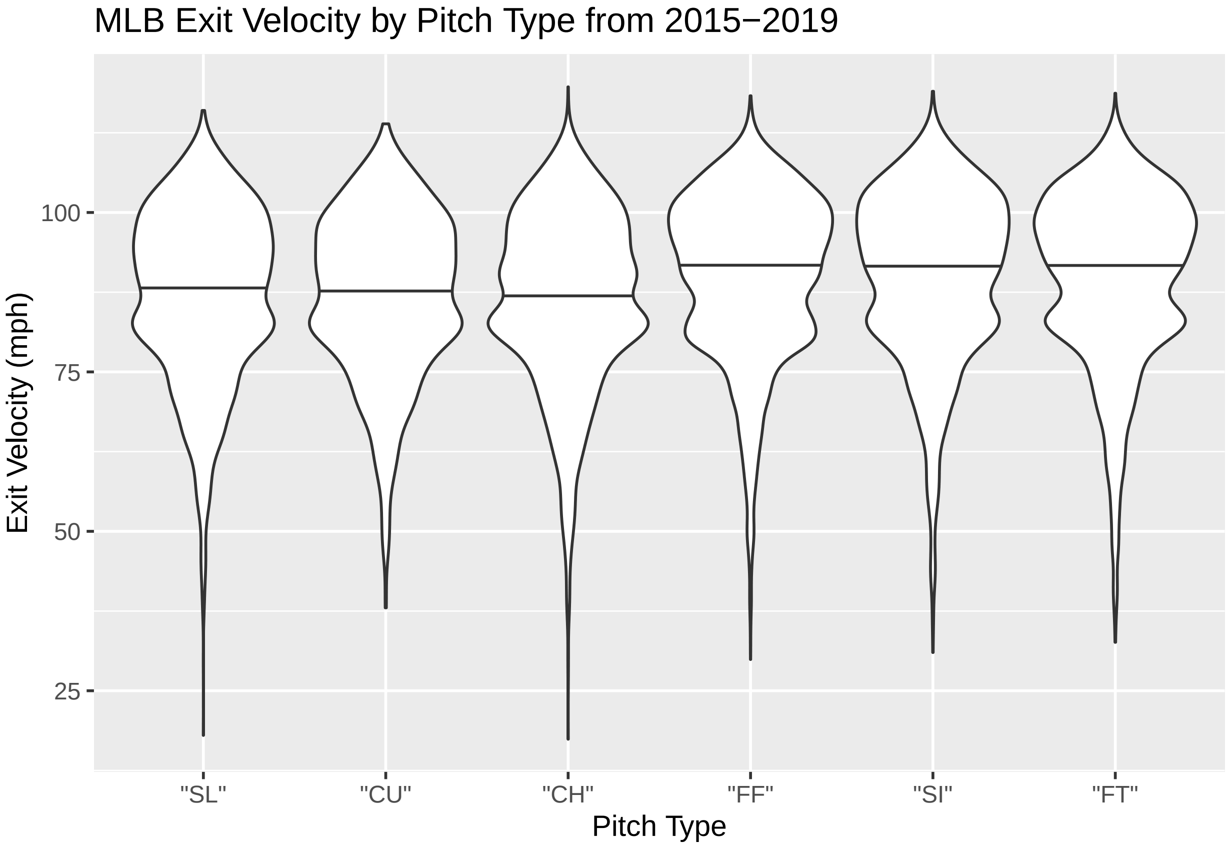
```
## `summarise()` regrouping output by 'game_year' (override with `.groups` argument)
```



Sliders have slowly been making their way as the go-to choice for offspeed pitch. If pitchers are adapting to get more strikeouts, it would make sense that sliders as a share of pitches are increasing because they are hard to hit, and/or hard to hit well.

PITCHING HYPOTHESIS #2: Sliders are a Hard Pitch to Hit/Hit Hard

```
full_savant_data %>%
  filter(launch_speed > 0) %>%
  filter(pitch_type %in% key_pitches) %>%
  ggplot(aes(x = factor(pitch_type, level = key_pitches))) +
  geom_violin(aes(y = launch_speed), draw_quantiles = 0.5) +
  xlab("Pitch Type") +
  ylab("Exit Velocity (mph)") +
  ggtitle("MLB Exit Velocity by Pitch Type from 2015-2019")
```



Sliders are actually hit slightly harder than changeups or curveballs on average.

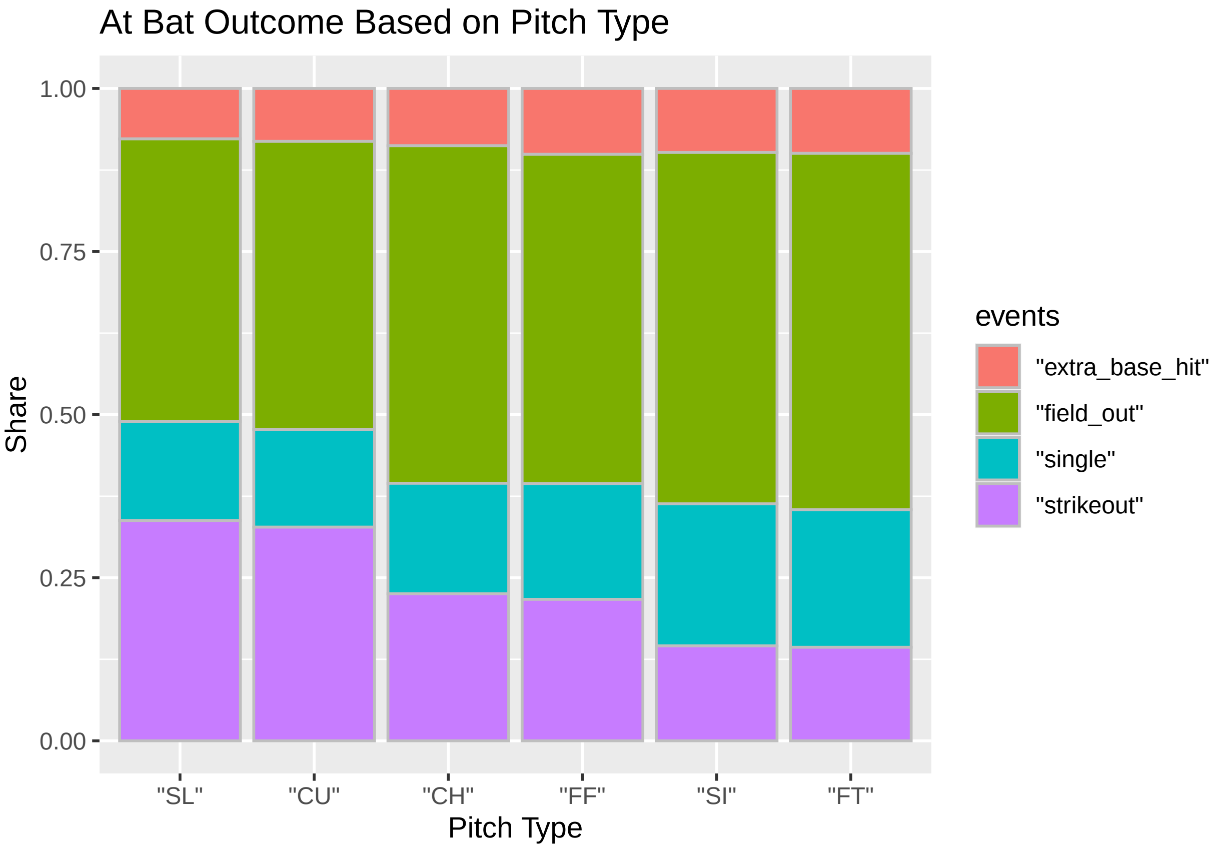
Table 7: Results of Different Pitch Types, 2008-2019

pitch_type	events	count	share
"CH"	"extra_base_hit"	741	0.0878066
"CU"	"extra_base_hit"	443	0.0811207
"FF"	"extra_base_hit"	2366	0.1009558
"FT"	"extra_base_hit"	784	0.0994419
"SI"	"extra_base_hit"	776	0.0980665
"SL"	"extra_base_hit"	912	0.0772096
"CH"	"field_out"	4366	0.5173599
"CU"	"field_out"	2411	0.4414942
"FF"	"field_out"	11831	0.5048216
"FT"	"field_out"	4308	0.5464231
"SI"	"field_out"	4263	0.5387337
"SL"	"field_out"	5119	0.4333728
"CH"	"single"	1430	0.1694514
"CU"	"single"	818	0.1497894
"FF"	"single"	4156	0.1773340
"FT"	"single"	1661	0.2106799
"SI"	"single"	1723	0.2177430
"SL"	"single"	1793	0.1517948
"CH"	"strikeout"	1902	0.2253822
"CU"	"strikeout"	1789	0.3275957
"FF"	"strikeout"	5083	0.2168885
"FT"	"strikeout"	1131	0.1434551
"SI"	"strikeout"	1151	0.1454568
"SL"	"strikeout"	3988	0.3376228

```r
full_savant_data %>%
  filter(events %in% c('"field_out"', '"single"', '"double"',
                       '"triple"', '"home_run"', '"strikeout"')) %>%
  mutate(events = ifelse(events == '"double"', '"extra_base_hit"',
                         ifelse(events == '"triple"', '"extra_base_hit"',
                                ifelse(events == '"home_run"', '"extra_base_hit"', events)))) %>%
  filter(pitch_type %in% key_pitches) %>%
  group_by(pitch_type, events) %>%
  summarise(count = n()) %>%
  group_by(pitch_type) %>%
  mutate(share = count / sum(count)) %>%
  arrange(events) %>%
  kableExtra::kable(caption = "Results of Different Pitch Types, 2008-2019")
```

```
## `summarise()` regrouping output by 'pitch_type' (override with `.groups` argument)
```

```
full_savant_data %>%
  filter(events %in% c('"field_out"', '"single"', '"double"', '"triple"', '"home_run"', '"strikeout"'))
  mutate(events = ifelse(events == '"double"', '"extra_base_hit"',
                         ifelse(events == '"triple"', '"extra_base_hit"',
                                ifelse(events == '"home_run"', '"extra_base_hit"', events)))) %>%
  filter(pitch_type %in% key_pitches) %>%
  mutate(pitch_type = factor(pitch_type, levels = key_pitches)) %>%
  ggplot() +
  geom_bar(aes(x = pitch_type, fill = events), position = "fill", color = "gray") +
  xlab("Pitch Type") +
  ylab("Share") +
  ggtitle("At Bat Outcome Based on Pitch Type")
```



When thrown as the last pitch of an at bat, a slider ("SL") has the greatest chance to strike out a batter, and offers the lowest chance to get an extra-base hit. It would make sense that pitchers would want to throw this pitch more often, since their goal is minimizing runs, which as we saw earlier goes hand in had with limiting extra base hits.

Similarly, sinkers ("SI") were a much more popular pitch in 2008 (the 2nd most popular in fact!), but their frequency has been on a sharp decline since then. This also seems to fit into the logic that pitchers want to limit runs, and sinkers do not help that cause.

So why would Sliders and Curveballs ("CU") not be thrown as often as fastballs? And why are Curveballs thrown less often than Sliders? My guess would be Accuracy, but I currently do not have the data necessary to investigate whether that may be true.