Baseball Analysis - Tara Ghorpadkar

Motivation: I wanted to analyze the trends in baseball following the recent enforcement of the “sticky stuff” ball rule change in the MLB implemented on June 21, 2021. This rule prohibits the doctoring of balls by the pitching team in order to increase the spin rate by getting a better grip on the ball. By increasing the spin rate with the use of adhesives on the ball, this increases the ball’s time in the air, therefore making the ball’s vertical position higher than expected when it reaches the hitter. This rule was created prior to June 21, but it had been loosely enforced until recently. Suspicions of teams using sticky substances on baseballs arose back in 2018 when The Dodgers’ Trevor Bauer blew the whistle on the league, but enforcement continued to be loose until this past season saw the lowest batting averages and the highest no-hitters on record. The MLB came out with the news on June 5, 2021 to warn the teams of this upcoming change. This new rule will suspend a player for 10 days if they are caught using sticky substances to enhance their performance.For my project, I wanted to see how and when the rule change affected the league. Specifically, I looked to see which types of pitches were most impacted by the new rule, when exactly the change in spin rate occurred, which teams previously relied on the use of sticky substances the most, and whether batting averages were impacted by this enforcement.

```
library(devtools)
library(baseballr)
library(tidyverse)
library(lubridate)
library(ggplot2)
library(scales)
library(reshape2)
library(knitr)
```

First, I import the data from Baseball Savant Statcast Search from the beginning to July 30 of the 2021 season. In order to download the data from the website, you will need to uncomment the lines in this chunk.

```
#dates <- seq.Date(as.Date('2021-04-01'), as.Date('2021-07-30'), by = 7)

#date_grid <- tibble(start_date = dates, end_date = dates+6)

#savant_data <- purrr::map2_df(.x = date_grid$start_date,
#                              .y = date_grid$end_date,
#                              ~scrape_statcast_savant(start_date = .x,
#                                                      end_date = .y,
#                                                      player_type = 'pitcher'))
#write.csv(savant_data,"savant_data.csv", row.names = TRUE)

df = read.csv("savant_data.csv", header = TRUE)
paste('Number of pitches since July 30:', nrow(df))
```

```
## [1] "Number of pitches since July 30: 473481"
```

Then I clean the data frame by removing all of the columns that I am not using for my analysis as well as null values for the columns I am using.

```
df[,c("pitch_type", "game_date", "release_speed",
      "release_pos_x", "release_pos_z", "player_name",
      "zone", "des", "p_throws", "home_team", "away_team",
      "plate_x", "plate_z", "inning_topbot", "launch_speed",
      "launch_angle", "effective_speed", "release_spin_rate",
      "release_extension", "estimated_ba_using_speedangle",
      "estimated_woba_using_speedangle", "woba_value", "woba_denom",
      "babip_value", "iso_value", "launch_speed_angle",
      "at_bat_number", "pitch_number", "pitch_name", "spin_axis")]
```

pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name
<chr>	<chr>	<dbl>	<dbl>	<dbl>	<chr>
FF	2021-04-07	96.3	-1.43	5.72	Bauer, Trevor
FC	2021-04-07	87.8	-1.57	5.54	Bauer, Trevor
FC	2021-04-07	86.2	-1.55	5.66	Bauer, Trevor
FF	2021-04-07	95.2	-1.65	5.69	Bauer, Trevor
SL	2021-04-07	81.3	-1.77	5.52	Bauer, Trevor
SL	2021-04-07	82.5	-1.75	5.45	Bauer, Trevor
FF	2021-04-07	90.6	2.28	5.85	Luzardo, Jesús
FC	2021-04-07	85.4	-1.66	5.64	Bauer, Trevor
SL	2021-04-07	85.4	-1.23	5.40	Bieber, Shane
FF	2021-04-07	94.7	-1.51	5.67	Bauer, Trevor

1-10 of 10,000 rows | 1-7 of 30 columns Previous **1** 2 3 4 5 6 ... 1000 Next

```
# remove NAs
df <- df[!(is.na(df$release_spin_rate)|is.na(df$pitch_type)|is.na(df$pitch_name)|is.n
a(df$zone)), ]
```

To normalize the spin rates in order to compare spin rates with different pitch velocities, I convert the spin rates to Bauer Units in a Bauer Units Column. I also convert the date values to Date data types in R. I then designate which team is pitching. I also remove pitch types not commonly used in the league.

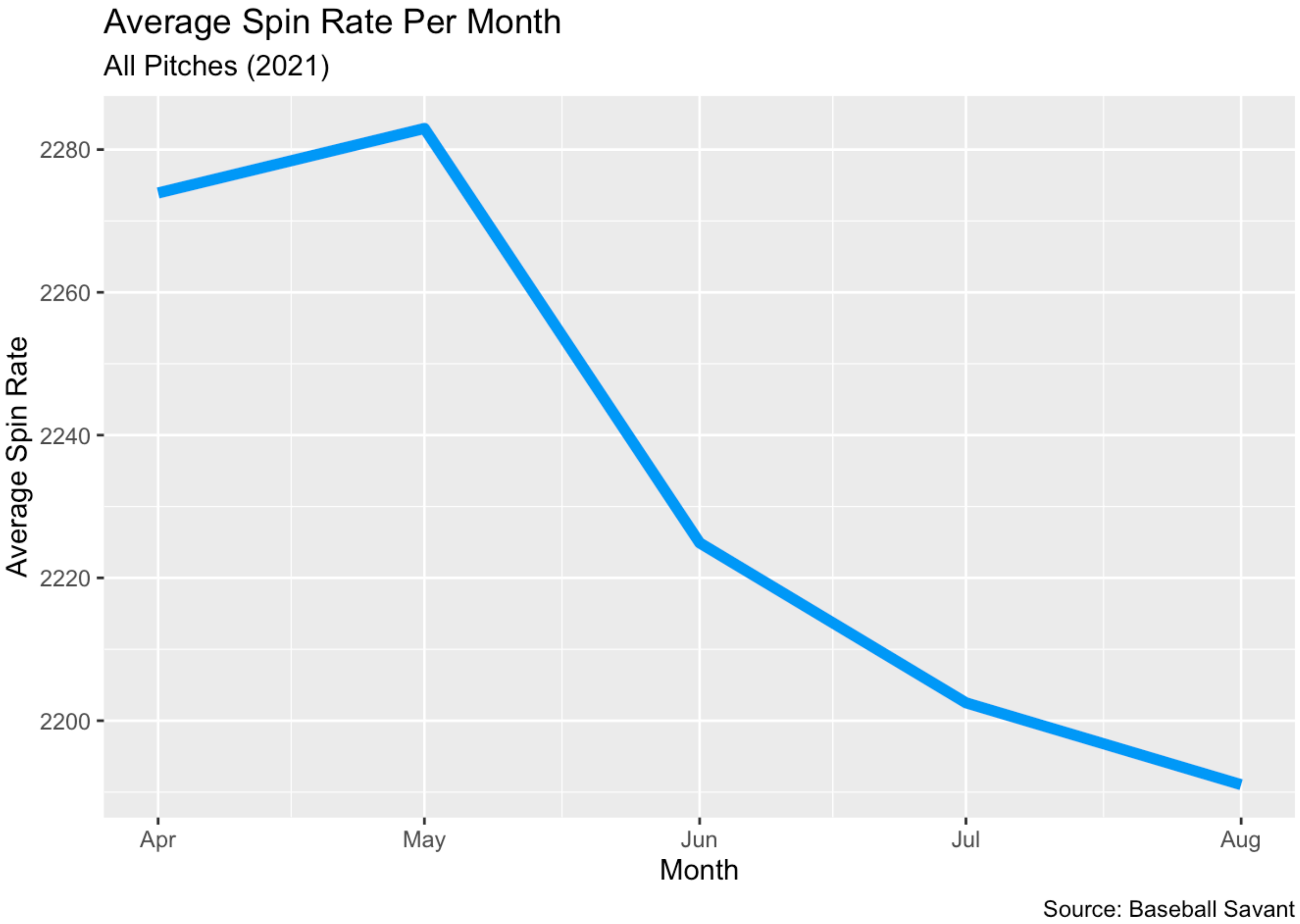
```r
#create Baur units column
df$b_units <- df$release_spin_rate/df$release_speed

#month and week of game date
df$month <- month(as_date(df$game_date))
df$week <- week(as_date(df$game_date))
df$week_of <- as.Date(sapply (as_date(df$game_date), function(d) { return (d + (-6 - as.POSIXlt(d)$wday %% -7 ))}), origin = "1970-01-01")
df$pitch_team <- ifelse(df$inning_topbot == "Bot", df$away_team,df$home_team)
df2 <- df[df$pitch_name != "Fastball" & df$pitch_name != "Knuckleball" & df$pitch_name != "Eephus" & df$pitch_name != "2-Seam Fastball" & df$pitch_name != "Screwball",]
```

Here, I have grouped the data by its average daily, weekly, and monthly spin rates and taken averages of the points of data to get one point for each day, week, and month. Then, I have put that data in a table.

```r
# start of analysis

daily_avg_spin <- aggregate(df2$release_spin_rate, list(df2$game_date), FUN=mean)
colnames(daily_avg_spin) <- c('date','spin_rate')
weekly_avg_spin <- aggregate(df2$release_spin_rate, list(df2$week_of), FUN=mean)
colnames(weekly_avg_spin) <- c('week_of','spin_rate')
monthly_avg_spin <- aggregate(df2$release_spin_rate, list(df2$month), FUN=mean)
colnames(monthly_avg_spin) <- c('month','spin_rate')
```

These are the graphs of the average daily, weekly, and monthly spin rates from April 1 to July 30.

```r
ggplot(data=monthly_avg_spin, aes(x=as.Date(paste0("2021-", month, "-1")), y=spin_rate, group=1)) +
  geom_line(linetype="solid", size=2, color="#0099f9")+
  labs(
    x = "Month",
    y = "Average Spin Rate",
    title = "Average Spin Rate Per Month",
    subtitle = "All Pitches (2021)",
    caption = "Source: Baseball Savant"
  )+
  scale_x_date(date_labels = "%b")
```

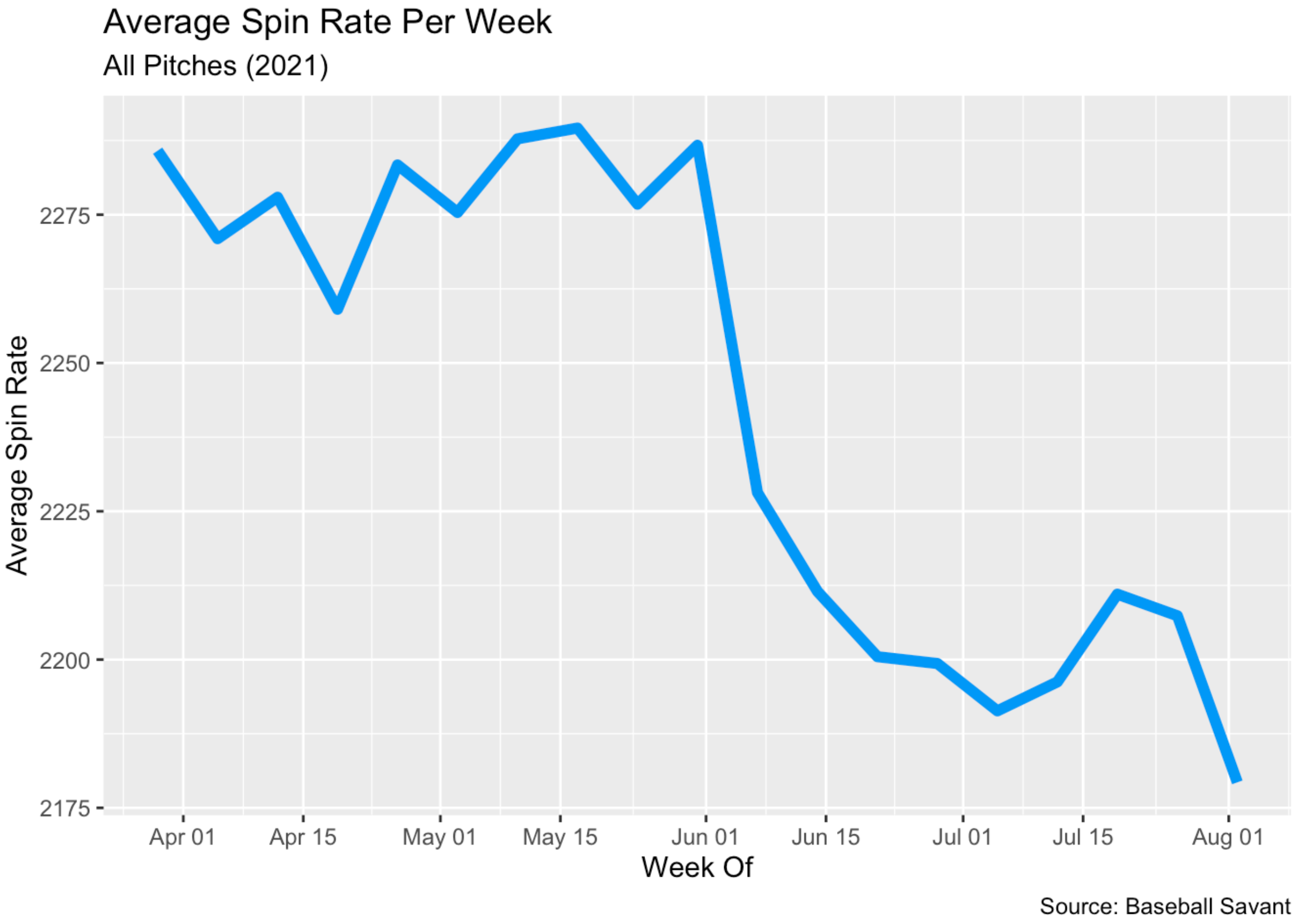


Average Spin Rate Per Month

All Pitches (2021)

Source: Baseball Savant

```
ggplot(data=weekly_avg_spin, aes(x=week_of, y=spin_rate, group=1)) +
  geom_line(linetype="solid", size=2, color="#0099f9")+
  labs(
    x = "Week Of",
    y = "Average Spin Rate",
    title = "Average Spin Rate Per Week",
    subtitle = "All Pitches (2021)",
    caption = "Source: Baseball Savant"
  ) +
  scale_x_date(breaks = scales::breaks_pretty(10))
```

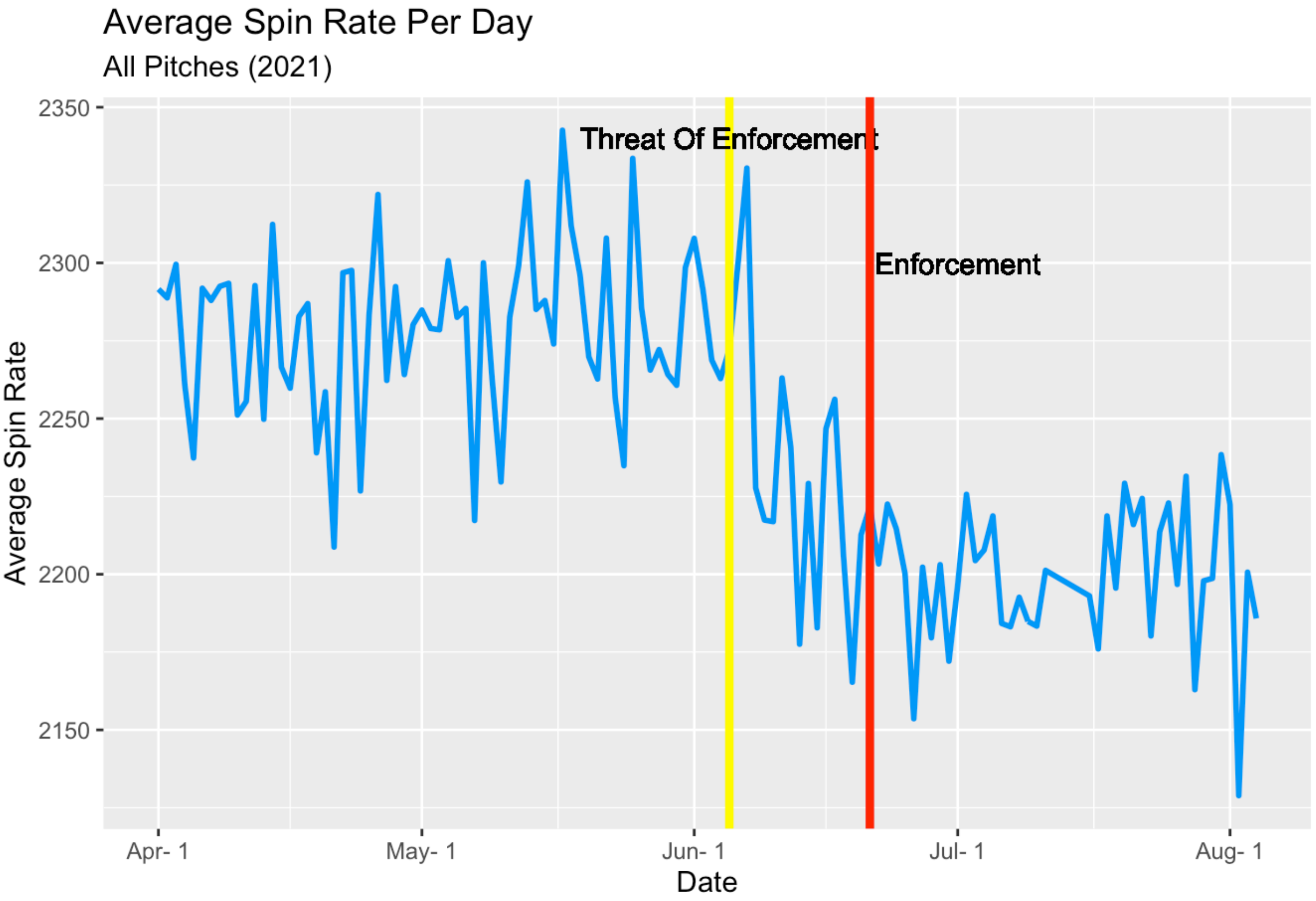


Average Spin Rate Per Week

All Pitches (2021)

Source: Baseball Savant

```
ggplot(data=daily_avg_spin, aes(x=as.Date(date), y=spin_rate, group=1)) +
  geom_line(linetype="solid", size=1, color="#0099f9")+
  labs(
    x = "Date",
    y = "Average Spin Rate",
    title = "Average Spin Rate Per Day",
    subtitle = "All Pitches (2021)",
    caption = "Source: Baseball Savant"
  ) +
  scale_x_date(date_labels = "%b-%e") +
  geom_vline(xintercept = as.numeric(ymd("2021-06-05")), linetype="solid", color = "yellow", size=1.5) +
  geom_text(aes(x=as.Date("2021-06-05"), label="Threat Of Enforcement", y=2340)) +
  geom_vline(xintercept = as.numeric(ymd("2021-06-21")), linetype="solid", color = "red", size=1.5) +
  geom_text(aes(x=as.Date("2021-06-21")+10, label="Enforcement", y=2300))
```



Source: Baseball Savant

For the graph of the Average Daily Spin Rate, I included a line to mark the date when the threat of enforcement was announced (June 5) as well as a line to mark the date of actual enforcement (June 21). It seems as if the change in spin rate took a dive shortly after threat of enforcement, so I created a graph and a table to get a specific date that accounts for the drop.

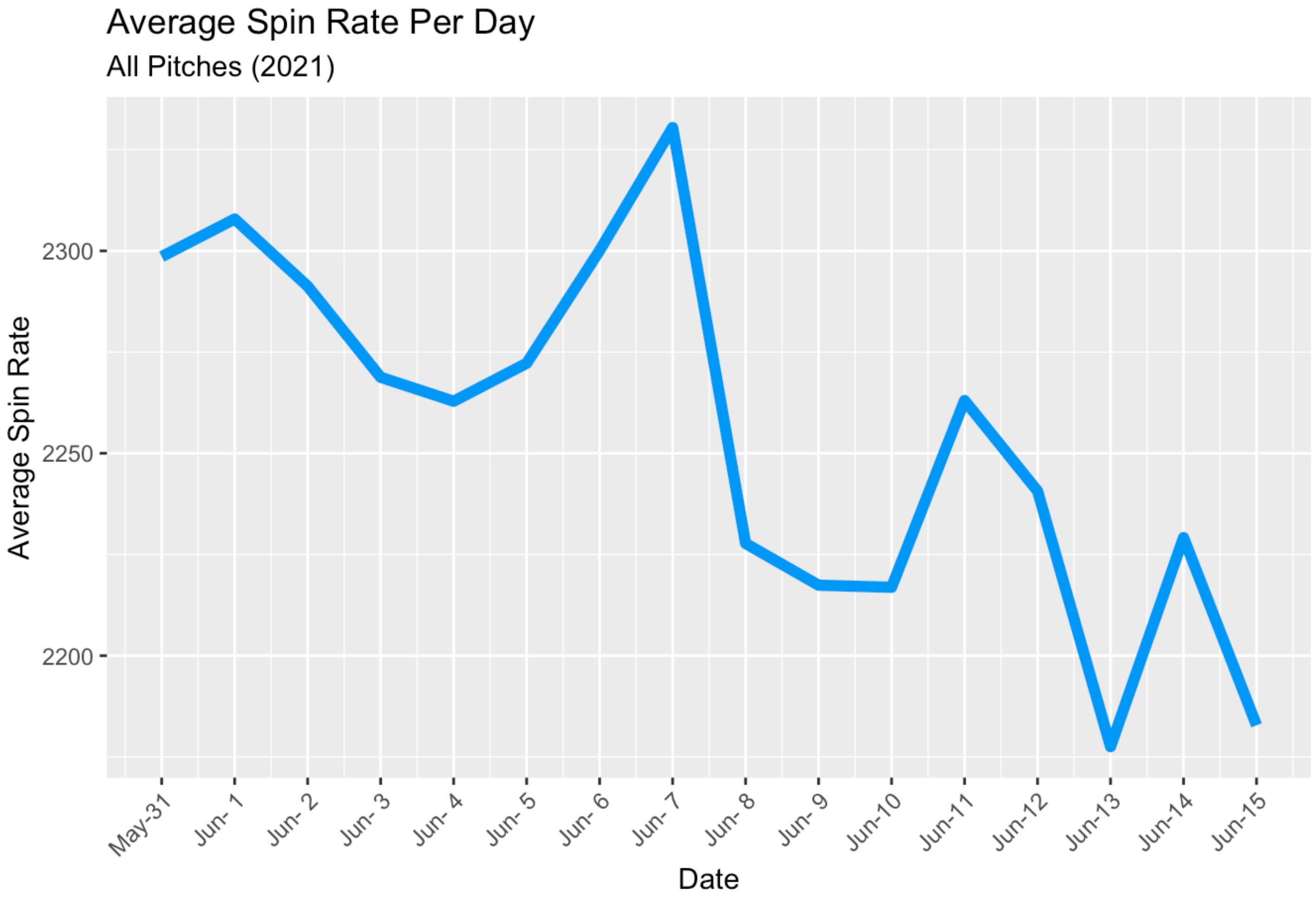
```
june_1_15_daily_spin <- daily_avg_spin[daily_avg_spin$date >= '2021-05-31' & daily_av
g_spin$date <= '2021-06-15',]

june_1_15_daily_spin
```

	date <chr>	spin_rate <dbl>
61	2021-05-31	2298.575
62	2021-06-01	2307.881
63	2021-06-02	2291.262
64	2021-06-03	2268.776

65	2021-06-04	2262.858
66	2021-06-05	2272.236
67	2021-06-06	2300.092
68	2021-06-07	2330.453
69	2021-06-08	2227.731
70	2021-06-09	2217.410

1-10 of 16 rows Previous **1** 2 Next

```
ggplot(data=june_1_15_daily_spin, aes(x=as.Date(date), y=spin_rate, group=1)) +
  geom_line(linetype="solid", size=2, color="#0099f9")+
  labs(
    x = "Date",
    y = "Average Spin Rate",
    title = "Average Spin Rate Per Day",
    subtitle = "All Pitches (2021)",
    caption = "Source: Baseball Savant"
  ) +
  scale_x_date(date_labels = "%b-%e",date_breaks  ="1 day") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Source: Baseball Savant

The day on which there is a significant drop is June 8th, so up till June 7th, the threat of enforcement did not seem to have an effect on spin rate. The drop, however, remained constant after the June 21 enforcement. This shows that spin rate significantly decreased just by threat of enforcement, which means that teams most likely stopped using sticky substances after they were threatened with a suspension and did not wait until actual enforcement to do so.

It is also clear that universally, pitch spin rates were decreased by this threat. Next, I want to check whether this change in spin rate was actually due to pitchers having a lower spin value regardless of pitch or was it due to the pitchers using pitches that naturally do not have as high of a spin rate. To check this, I am going to visualize the frequency of types of pitches being thrown before and after June 7th and the change in spin rate on each of those pitches.

Now, I am grouping the data by its pitch name and finding the frequency of each pitch being thrown before and after June 7th. I am also finding the change in frequency in each pitch.

```
pre_freq <- df2[df2$game_date <= '2021-06-07',]
pre_freq <- aggregate(pre_freq$release_spin_rate, list(pre_freq$pitch_name), FUN = le
ngth)
colnames(pre_freq) <- c('pitch_name','pre_count')

pre_freq$pre_frequency <- round(pre_freq$pre_count / sum(pre_freq$pre_count) * 100, d
igits = 2)

post_freq <- df2[df2$game_date > '2021-06-07',]
post_freq <- aggregate(post_freq$release_spin_rate, list(post_freq$pitch_name), FUN =
length)
colnames(post_freq) <- c('pitch_name', 'post_count')

post_freq$post_frequency <- round(post_freq$post_count / sum(post_freq$post_count) *
100, digits = 2)

pre_post_freq <- merge(pre_freq, post_freq, by = "pitch_name")

pre_post_freq$freq_diff <- pre_post_freq$post_frequency - pre_post_freq$pre_frequency

pre_post_freq$freq_diff_pct_freq <- round((pre_post_freq$freq_diff / pre_post_freq$pr
e_frequency) * 100, digits = 2)

pre_post_freq
```

pitch_name	pre_count	pre_frequency	post_count	post_frequency	freq_diff	fr
<chr>	<int>	<dbl>	<int>	<dbl>	<dbl>	
4-Seam Fastball	89784	35.06	75082	35.00	-0.06	
Changeup	29712	11.60	24034	11.20	-0.40	
Curveball	21013	8.21	17998	8.39	0.18	
Cutter	17867	6.98	14248	6.64	-0.34	
Knuckle Curve	5831	2.28	3978	1.85	-0.43	
Sinker	39184	15.30	34451	16.06	0.76	
Slider	49111	19.18	41164	19.19	0.01	
Split-Finger	3578	1.40	3573	1.67	0.27	

8 rows

As shown, the largest drop in frequency for a pitch was for the Knuckle-Curve pitch.

Next, I am going to look at the spin rates for each pitch before and after June 7th.

```
pre_spin <- df2[df2$game_date <= '2021-06-07',]
pre_spin <- aggregate(pre_spin$release_spin_rate, list(pre_spin$pitch_name), FUN = mean)
colnames(pre_spin) <- c('pitch_name','pre_spin')

post_spin <- df2[df2$game_date > '2021-06-07',]
post_spin <- aggregate(post_spin$release_spin_rate, list(post_spin$pitch_name), FUN = mean)
colnames(post_spin) <- c('pitch_name','post_spin')

pre_post_spin <- merge(pre_spin, post_spin, by = "pitch_name")

pre_post_spin$spin_diff <- pre_post_spin$post_spin - pre_post_spin$pre_spin

pre_post_spin$spin_diff_pct_spin <- (pre_post_spin$spin_diff / pre_post_spin$pre_spin) * 100

pre_post_spin
```

pitch_name	pre_spin	post_spin	spin_diff	spin_diff_pct_spin
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
4-Seam Fastball	2317.660	2241.492	-76.16811	-3.286422
Changeup	1784.398	1719.496	-64.90226	-3.637208
Curveball	2554.739	2474.215	-80.52378	-3.151938
Cutter	2416.577	2331.201	-85.37523	-3.532900
Knuckle Curve	2562.021	2500.527	-61.49412	-2.400219
Sinker	2158.137	2098.076	-60.06098	-2.783001
Slider	2462.543	2395.673	-66.87061	-2.715510
Split-Finger	1481.078	1300.416	-180.66124	-12.197958

8 rows

For each pitch, there is a 2.5 - 3.5 % decrease in the spin rate after June 7th, so there was no dramatic decrease in spin rate for the Knuckle-Curve Pitch. It appears as though no matter the pitch, the enforcement of the “sticky stuff” rule decreased spin rate around the same amount. The exception for this is the Split-Finger pitch, but because it has a much lower frequency than any other pitch on the table, it is susceptible to outliers in its data set because the sample size is so small, so this data may not be entirely reliable for that pitch.

Next, I wanted to see which teams were impacted the most by this rule change. Did this problem affect some teams in particular, or was it fairly even across the board?

Here, I am grouping the data by pitch team to find the differences in average spin rate per team before and after June 7th. I have ordered the data in terms of largest to smallest difference.

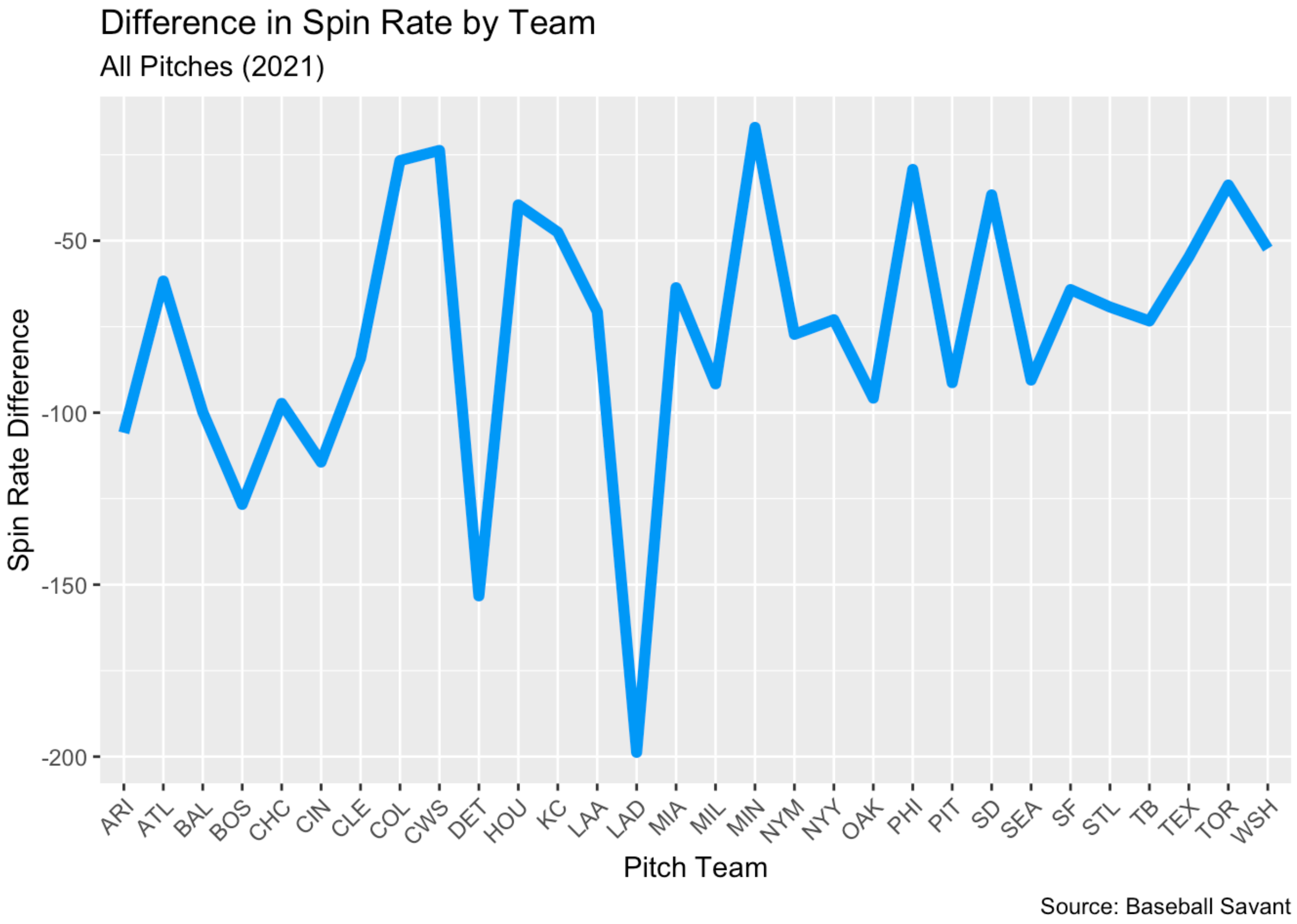
```
team_pre_spin <- df2[df2$game_date <= '2021-06-07',]
team_pre_spin <- aggregate(team_pre_spin$release_spin_rate, list(team_pre_spin$pitch_
team), FUN = mean)
colnames(team_pre_spin) <- c('pitch_team','pre_spin')

team_post_spin <- df2[df2$game_date > '2021-06-07',]
team_post_spin <- aggregate(team_post_spin$release_spin_rate, list(team_post_spin$pit
ch_team), FUN = mean)
colnames(team_post_spin) <- c('pitch_team','post_spin')

team_spin_comp <- merge(team_pre_spin, team_post_spin, by = "pitch_team")

team_spin_comp$difference <- team_spin_comp$post_spin - team_spin_comp$pre_spin

team_spin_comp <- team_spin_comp[order(team_spin_comp$difference),]
team_spin_comp
```

	pitch_team <chr>	pre_spin <dbl>	post_spin <dbl>	difference <dbl>
14	LAD	2587.406	2388.657	-198.74886
10	DET	2229.156	2075.900	-153.25564
4	BOS	2295.623	2168.964	-126.65864
6	CIN	2362.781	2248.410	-114.37181
1	ARI	2289.747	2183.881	-105.86539
3	BAL	2285.101	2185.288	-99.81346
5	CHC	2277.628	2180.323	-97.30454
20	OAK	2151.146	2055.381	-95.76406
16	MIL	2313.647	2222.030	-91.61710
22	PIT	2369.129	2277.853	-91.27569

1-10 of 30 rows Previous **1** 2 3 Next

As shown, the LA Dodgers and the Detroit Tigers have seen the largest drop in average spin rate, but the Chicago White Sox and the Minnesota Twins have not seen a noticable drop in spin rate. Therefore, this rule change does not affect the league evenly, as some teams may have relied more on sticky substances than others and with the new enforecement have seen drastic drops in spin rate as a result.

```
ggplot(team_spin_comp, aes(x = pitch_team, y = difference, group = 1)) +
  geom_line(linetype="solid", size=2, color="#0099f9")+
  labs(
    x = "Pitch Team",
    y = "Spin Rate Difference",
    title = "Difference in Spin Rate by Team",
    subtitle = "All Pitches (2021)",
    caption = "Source: Baseball Savant"
  ) + theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


Difference in Spin Rate by Team

All Pitches (2021)

Source: Baseball Savant

It is obvious that this rule has had an affect on teams in the league, but has this rule affected specific players? Next, I want to check which pitchers have seen the largest drops in spin rate as a result of the rule. I group the data by pitcher and average their spin rate before and after June 7th and calculate the difference in spin rate.

```
pitcher_pre_spin <- df2[df2$game_date <= '2021-06-07',]
pitcher_pre_spin <- pitcher_pre_spin <- aggregate(pitcher_pre_spin$release_spin_rate,
list(pitcher_pre_spin$player_name), FUN = mean)
colnames(pitcher_pre_spin) <- c('player_name','pre_spin')

pitcher_post_spin <- df2[df2$game_date > '2021-06-07',]
pitcher_post_spin <- pitcher_post_spin <- aggregate(pitcher_post_spin$release_spin_ra
te, list(pitcher_post_spin$player_name), FUN = mean)
colnames(pitcher_post_spin) <- c('player_name','post_spin')

pitcher_spin_comp <- merge(pitcher_pre_spin, pitcher_post_spin, by = "player_name")

pitcher_spin_comp$spin_diff <- pitcher_spin_comp$post_spin - pitcher_spin_comp$pre_sp
in

pitcher_spin_comp <- pitcher_spin_comp[order(pitcher_spin_comp$spin_diff),]
pitcher_spin_comp
```

	player_name	pre_spin	post_spin	spin_diff
	<chr>	<dbl>	<dbl>	<dbl>
60	Bumgarner, Madison	2532.237	2136.009	-396.2278658
381	Richards, Garrett	2783.719	2423.646	-360.0732331
282	Maples, Dillon	2879.107	2520.600	-358.5069519
383	Ríos, Yacksel	2200.500	1877.934	-322.5661578
36	Bender, Anthony	2565.160	2248.299	-316.8618699
363	Ponce, Cody	2500.230	2200.209	-300.0212520
274	Lynch, Daniel	2201.098	1921.609	-279.4887903
22	Antone, Tejay	2864.119	2594.324	-269.7950172
45	Borucki, Ryan	2514.561	2245.337	-269.2241458
107	de Geus, Brett	2297.339	2028.637	-268.7015310

1-10 of 524 rows Previous **1** 2 3 4 5 6 ... 53 Next

I also create a table to show the number of pitches thrown by each player.

```
pitches_thrown <- df2
pitches_thrown <- pitches_thrown <- aggregate(pitches_thrown$pitch_name, list(pitches
_thrown$player_name), FUN = length)
colnames(pitches_thrown) <- c('player_name', 'pitches_thrown')
```

Here, I create a table to show the number of appearances of each player in games before and after June 7th and the total appearances up till July 30th.

```
df_after_june_7 <- df2[df2$game_date >= '2021-06-07',]
apps_after_grouper <- aggregate(df_after_june_7$game_date, list(df_after_june_7$playe
r_name), FUN = function(x) length(unique(x)))
colnames(apps_after_grouper) <- c('player_name', 'appearances_june_7_later')

df_before_june_7 <- df2[df2$game_date < '2021-06-07',]
apps_before_grouper <- aggregate(df_before_june_7$game_date, list(df_before_june_7$pl
ayer_name), FUN = function(x) length(unique(x)))
colnames(apps_before_grouper) <- c('player_name', 'appearances_june_7_before')

apps <- merge(apps_after_grouper, apps_before_grouper, by = "player_name")
apps$total_apps <- apps$appearances_june_7_later + apps$appearances_june_7_before
```

This table shows the combined number of pitches and appearances made by each pitcher.

```
pitches_apps <- merge(apps, pitches_thrown, by = "player_name")
pitches_apps$pitches_per_app <- round((pitches_apps$pitches_thrown / pitches_apps$tot
al_apps),digits = 1)
```

This data frame combines the number of pitches, the number of appearances, and the spin rate of each player before and after June 7th.

```
apps_spin_diff <- merge(pitches_apps,pitcher_spin_comp,by = "player_name")
apps_spin_diff
```

player_name	appearances_june_7_later	appearances_june_7_before	total_ap
<chr>	<int>	<int>	<ir
Abbott, Cory	5	1	
Abreu, Albert	6	4	
Abreu, Bryan	9	21	
Adams, Austin	22	25	
Akin, Keegan	9	6	

Alcala, Jorge	19	23
Alcantara, Sandy	9	13
Alexander, Scott	5	13
Alexander, Tyler	14	16
Allard, Kolby	10	12

1-10 of 522 rows | 1-5 of 9 columns Previous **1** 2 3 4 5 6 ... 53 Next

Next, I am going to compare the xWOBA and the Bauer Units(pitch spin rate / pitch spin velocity) for each player before and after June 7th.

First, I am going to focus on xWOBA. This is the expected value of a batted ball for a player based on a launch angle and launch velocity when it leaves the bat. Here, I group the data based on player name and find the average xWOBA for each player before and after June 7th to find the difference in xWOBA based on pitcher appearances.

```
pitcher_pre_xwoba <- aggregate(df_before_june_7$estimated_woba_using_speedangle, list
(df_before_june_7$player_name), FUN = mean, na.rm=TRUE, na.action=na.pass)
colnames(pitcher_pre_xwoba) <- c('player_name', 'Pre_xwOBA')
pitcher_post_xwoba <- aggregate(df_after_june_7$estimated_woba_using_speedangle, list
(df_after_june_7$player_name), FUN = mean, na.rm=TRUE, na.action=na.pass)
colnames(pitcher_post_xwoba) <- c('player_name', 'Post_xwOBA')
pitcher_xwoba_comp <- merge(pitcher_pre_xwoba,pitcher_post_xwoba,by = "player_name")
pitcher_xwoba_comp$xwOBA_diff <- pitcher_xwoba_comp$Post_xwOBA - pitcher_xwoba_comp$P
re_xwOBA
apps_spin_xwoba <- merge(apps_spin_diff, pitcher_xwoba_comp, by = "player_name")
apps_spin_xwoba
```

player_name <chr>	appearances_june_7_later <int>	appearances_june_7_before <int>	total_ap <ir
Abbott, Cory	5	1	
Abreu, Albert	6	4	
Abreu, Bryan	9	21	
Adams, Austin	22	25	
Akin, Keegan	9	6	
Alcala, Jorge	19	23	
Alcantara, Sandy	9	13	
Alexander, Scott	5	13	

Alexander, Tyler	14	16
Allard, Kolby	10	12

1-10 of 522 rows | 1-5 of 12 columns Previous 1 2 3 4 5 6 ... 53 Next

Second, I am going to focus on Bauer Units. Here, I group the data based on player name and average the Bauer Units before and after June 7th to find the difference in Bauer Units due to the rule change.

```
pitcher_pre_bu <- aggregate(df_before_june_7$b_units, list(df_before_june_7$player_na
me), FUN = mean)
colnames(pitcher_pre_bu) <- c('player_name', 'Pre_b_units')
pitcher_post_bu <- aggregate(df_after_june_7$b_units, list(df_after_june_7$player_nam
e), FUN = mean)
colnames(pitcher_post_bu) <- c('player_name', 'Post_b_units')
pitcher_bu_comp <- merge(pitcher_pre_bu, pitcher_post_bu, by = "player_name")
pitcher_bu_comp$bu_diff <- pitcher_bu_comp$Post_b_units - pitcher_bu_comp$Pre_b_units
pitcher_bu_comp
```

player_name	**Pre_b_units**	**Post_b_units**	**bu_diff**
<chr>	<dbl>	<dbl>	<dbl>
Abbott, Cory	26.17766	25.80636	-0.37130591
Abreu, Albert	23.91939	22.69770	-1.22169255
Abreu, Bryan	26.80922	26.58625	-0.22296699
Adams, Austin	32.61611	31.96578	-0.65033681
Akin, Keegan	26.13866	25.52136	-0.61729623
Alcala, Jorge	25.22535	25.56480	0.33945766
Alcantara, Sandy	24.44300	23.45151	-0.99149232
Alexander, Scott	24.05888	24.58727	0.52838776
Alexander, Tyler	25.22101	24.25055	-0.97045927
Allard, Kolby	25.11576	22.70676	-2.40900109

1-10 of 522 rows Previous 1 2 3 4 5 6 ... 53 Next

Now I merge the xWOBA comparison data frame with the Bauer Units comparison data frame.

```
apps_spin_xwoba_bu <- merge(apps_spin_xwoba,pitcher_bu_comp, by = "player_name")
apps_spin_xwoba_bu$percent_spin_change <- (apps_spin_xwoba_bu$spin_diff / apps_spin_x
woba_bu$pre_spin)
apps_spin_xwoba_bu
```

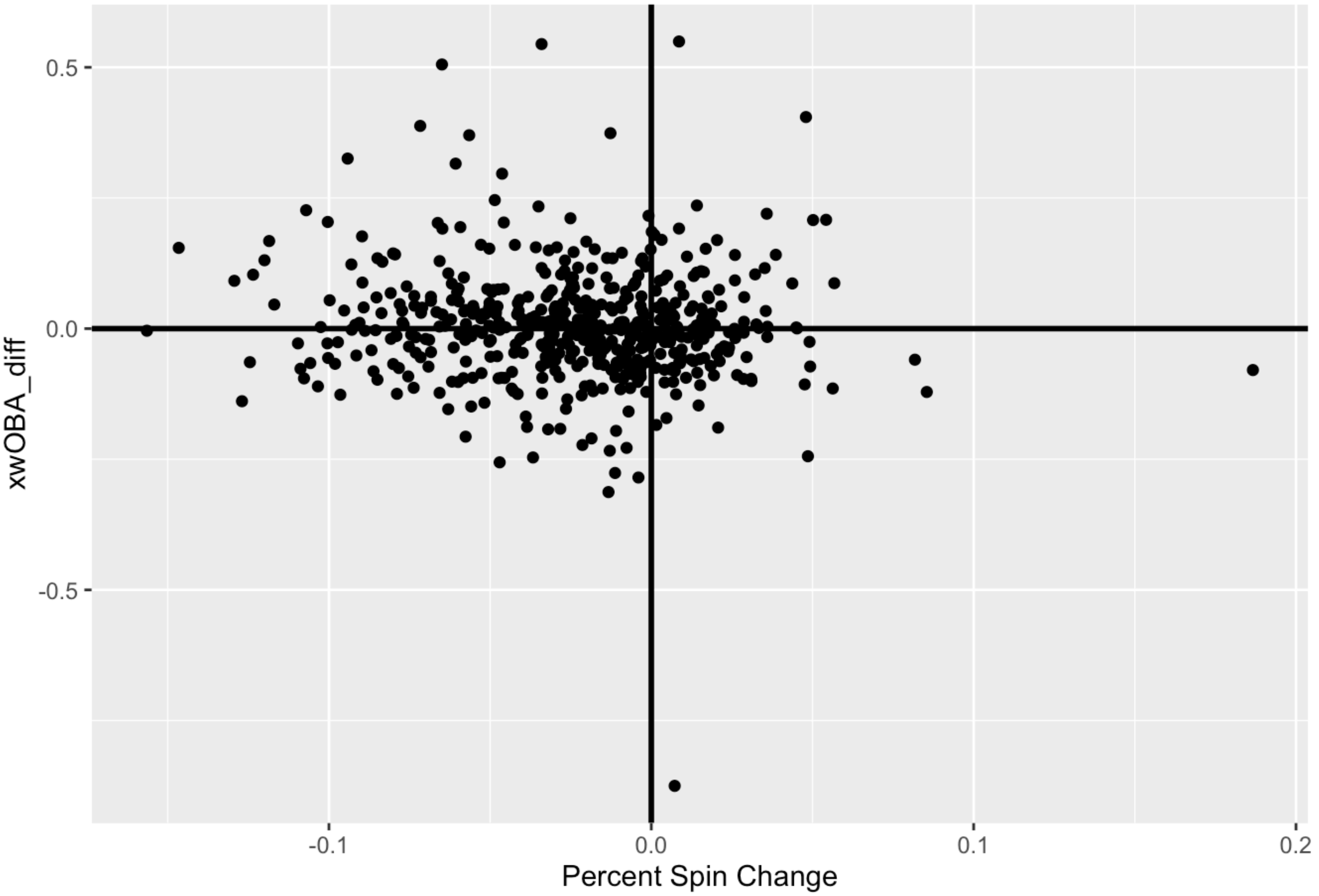
player_name	appearances_june_7_later	appearances_june_7_before	total_ap
<chr>	<int>	<int>	<ir
Abbott, Cory	5	1	
Abreu, Albert	6	4	
Abreu, Bryan	9	21	
Adams, Austin	22	25	
Akin, Keegan	9	6	
Alcala, Jorge	19	23	
Alcantara, Sandy	9	13	
Alexander, Scott	5	13	
Alexander, Tyler	14	16	
Allard, Kolby	10	12	

1-10 of 522 rows | 1-5 of 16 columns Previous **1** 2 3 4 5 6 ... 53 Next

And I plot the data, with the x axis indicating the percent change in spin rate for a pitcher and the y-axis indicating the xWOBA difference.

```
ggplot(apps_spin_xwoba_bu, aes(x = percent_spin_change , y = xwOBA_diff))+
  geom_point()+
  labs(
    x = "Percent Spin Change",
    y = "xwOBA_diff",
    caption = "Source: Baseball Savant"
  )+
  geom_vline(xintercept = 0.0, linetype="solid", color = "black", size=1)+
  geom_hline(yintercept = 0.0, linetype="solid", color = "black", size=1)
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

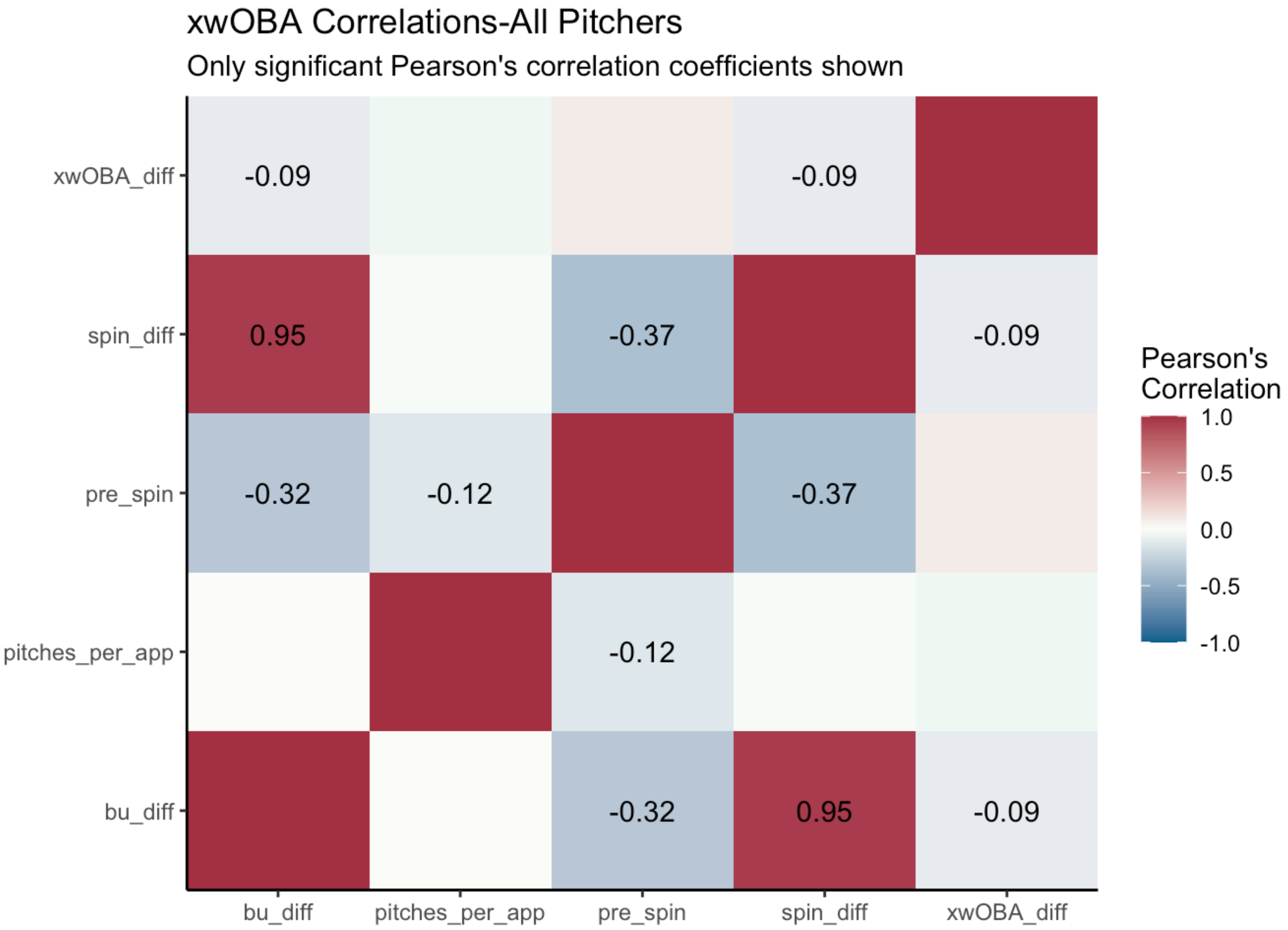


Source: Baseball Savant

It doesn't look like there is a correlation between spin change and xWOBA based on the graph. But, I am going to incorporate the difference in spin rate along with the spin rates before and after June 7th, the change in Bauer Units, and the difference in xWOBA into one graph to see if other variables cause an impact.

```
all_spin_xwoba_corr <- apps_spin_xwoba_bu[,c('xwOBA_diff', 'pitches_per_app', 'pre_sp
in', 'spin_diff', 'bu_diff')]

cors <- function(df) {
   # turn all three matrices (r, n, and P into a data frame)
   M <- Hmisc::rcorr(as.matrix(df))
   # return the three data frames in a list return(Mdf)
   Mdf <- map(M, ~data.frame(.x))
}
formatted_cors <- function(df){
 cors(df) %>%
 map(~rownames_to_column(.x, var="measure1")) %>%
 map(~pivot_longer(.x, -measure1, "measure2")) %>%
 bind_rows(.id = "id") %>%
 pivot_wider(names_from = id, values_from = value) %>%
 mutate(sig_p = ifelse(P < .05, T, F), p_if_sig = ifelse(P <.05, P, NA), r_if_sig = i
felse(P <.05, r, NA))
}

formatted_cors(all_spin_xwoba_corr) %>%
 ggplot(aes(measure1, measure2, fill=r, label=round(r_if_sig,2))) +
 geom_tile() +
 labs(x = NULL, y = NULL, fill = "Pearson's\nCorrelation", title="xwOBA Correlations-
All Pitchers", subtitle="Only significant Pearson's correlation coefficients shown")
+ scale_fill_gradient2(mid="#FBFEF9",low="#0C6291",high="#A63446", limits=c(-1,1)) +
 geom_text() +
 theme_classic() +
 scale_x_discrete(expand=c(0,0)) +
 scale_y_discrete(expand=c(0,0))
```

```
## Warning: Removed 13 rows containing missing values (geom_text).
```



However, after looking at the heat map, there doesn't seem to be a strong correlation between xWOBA and other variables as well, which means that spin rate did not have as much of an effect on a player's xWOBA.

Conclusion: The threat of enforcement of the "sticky stuff" ball rule change had a significant effect on the spin rate and most likely the usage of these substances caused this drop in spin rate by June 7th. However, after its actual enforcement on June 21, the spin rate remained fairly stagnant, indicating that by the time enforcement started, most of the players stopped the use of sticky substances in order to get a better grip. When looking at the distribution of spin rate drops, certain players and teams seemed to have been impacted by this more than others, indicating that some teams were relying on sticky substances more than others. However, the ban of sticky substances did not seem to have an effect on a player's performance in the game, as seen by the lack of correlation between spin rate and xWOBA.The pitchers' xWOBAs remain consistent even with the rule change amongst a variety of factors.

Potential Further Analysis: Since the data only covers up till the 2021 season, the lack of correlation may change in the upcoming seasons since the rule would then be implemented the entire season rather than just half of a season. As the sample size grows of spin rates and xWOBA beyond June 21, this may change the results. Additionally, I would look at the strikeout rate for these pitchers over time. Maybe the reason their xWOBAs are staying fairly constant is because their strikeouts are turning into weak contact outs or other types of non-hits.