2021 Major League Baseball season
Final Project
Data Warehousing (ISM6208)

Executive Summary:

For this project, we assembled a data warehouse using Major League Baseball Statcast data to store every single pitch thrown so far in the 2021 MLB season. With every pitch in the fact table, we are able to see all sorts of details on it such as the pitch's velocity, spin rate, location, movement, and pitch type. Our pitcher, batter, team, and date dimension tables allow us to view this data with the contextual details of each pitch: when the pitch was thrown, the season stats of the pitcher/batter, the player's height, weight, handedness, team, and many more. With the data warehouse created, we analyzed our data to see who the most talented pitchers were for velocity and spin rate, how those metrics change for pitch type, and how those metrics have changed throughout the season.

Authors

Michael Zeolla
MS Business Analytics and Information Systems
Muma College of Business
University of South Florida

Meghla Sarkar
MS Business Analytics and Information Systems
Muma College of Business
University of South Florida

Samrat Korupolu
MS Business Analytics and Information Systems
Muma College of Business
University of South Florida

Samarpan Dutta
MS Business Analytics and Information Systems
Muma College of Business
University of South Florida

Problem Statement

Our data warehouse allows us to examine all sorts of information on pitching throughout the 2021 Major League Baseball season, providing all sorts of options to analyze the season. Often as a fan of the sport, it can be difficult to find quick answers on questions about baseball statistics, but now the option of database querying provides relief for that. A very specific issue comes from baseball's current "sticky-stuff controversy." As of June 21st, umpires are checking pitchers to ensure they are not using foreign substances to increase their spin rate (rotations per minute on their pitches) for performance enhancement. This data warehouse can help us track spin rates and determine how spin rates have changed throughout the season, as well as identify potential pitchers that were previously using foreign substances.

Literature Review

Travis Sawchik is a baseball writer who has written about the "sticky-stuff controversy" a couple times throughout this baseball season. It was actually Sawchik who first brought the concept of using foreign substances to enhance spin rate to my attention, when his book *The MVP Machine* came out in 2019. The book detailed pitcher Trevor Bauer's thoughts on spin rate, and how he believes the only way to increase RPM on a fastball is through using substances for increased grip of the baseball. The Houston Astros had a few pitchers, most notably Gerrit Cole, Justin Verlander, and Charlie Morton, who saw massive jumps in both their success and fastball spin rates when they switched teams to the Astros.

Bauer at the time was criticized for accusing other pitchers of using foreign substances, but it increasingly became apparent that the trend across baseball was more pitchers using foreign substances. Previously, pitchers using foreign substances was ignored because it was seen as only a way to help grip, but now we know many are doing it to increase spin rate and improve performance.

In 2021, Sawchik has written a couple articles that I think provide a lot of insight on this controversy in baseball. One article from June 3rd, titled *Baseball's dirty little secret is out. We decided to experiment*, goes into detail about the controversy and then describes an experiment that Sawchik himself underwent to determine which substances best enhance spin rate.

On June 3rd, the date of the first article's release, Major League Baseball announced that they would start enforcing rules to ban pitchers using foreign substances. This was prompted by offense being down throughout baseball at historically low rates, with most people understanding that pitching had just gotten too far ahead of offense. The official date that baseball's new rules would be enforced was announced as June 21st, but people began noticing many pitcher's spin rates already plummeting since the original June 3rd date. Sawchik detailed this in his next article, *Is MLB's threat to police sticky stuff an effective deterrent for pitchers?* Sawchik has since released one more article, How pitchers are adjusting to baseball's sticky-stuff crackdown. Simply put, a lot of pitchers are experiencing diminished spin rate.

This YouTube video does a good job of summarizing the recent history of foreign substance use in baseball, and particularly details Trevor Bauer. Bauer, as previously mentioned, was a voice against foreign substances but has since gained notoriety for ironically becoming one of the pitchers likely benefitting from foreign substance use himself.

Data Collection and Preparation

For data collection we have used primarily three resources:

1. **MLB Data API:**

This is an excellent open-source API collection that exposes endpoint for player-level data, season-level, career-level, and league-level hitting and pitching stats for all the players and also team data along with their 40-man rosters for each season. We created python scripts (Included in Appendix 1) to fetch data using the API and export them as a CSV file. The table to below mentions the API endpoints we used and their purpose.

API Endpoint	Purpose
GET http://lookup-service-prod.mlb.com/json/named.team_all_season.bam?sport_code='mlb'&all_star_sw='N'&sort_order=name_asc&season='2021'	Team data for all the teams for the season 2021. (TEAM_DIM table)
GET http://lookup-service-prod.mlb.com/json/named.player_info.bam?sport_code='mlb'&player_id='493316'	Player information (player_id, name, height, weight, primary position etc.) for all the MLB players who played in season 2021. (PLAYER_OUTRIGGER table) **Note:** (The request URL is for fetching the details of the player having player_id = 493316)

2. **Baseball Savant:**

Baseball Savant is the main website for data on Statcast, a data tool used by Major League Baseball to quantify the talent and physical skills of players using high-speed cameras and doppler radar. It has been fully implemented since 2015. It is through Statcast that we can get the bulk of our project's data—our main fact table and for our pitcher and batter dimension tables.

Our fact table contains every single pitch thrown during the 2021 season through June 21st, which totals up to a whopping 313,274 pitches. In order to get such a large amount of Statcast data, we had to scrape it from the Baseball Savant website using R. Luckily for us, the R package *baseballr*, created by Bill Petti, allows us to accomplish this. In fact, on Bill Petti's blog, he details very

specifically how to obtain all pitch-level data from Statcast, exactly what we needed for our fact table.

In our pitcher and batter dimension tables, we wanted season-summary statistics for results-based metrics as characteristics of talent for players. Originally, we planned on using Fangraphs, another great baseball data website, for our pitcher and batter dimension tables. Fangraphs offers a season stats page that can easily be exported to a CSV file, as well as a lengthy and diverse list of stats. Unfortunately, Fangraphs uses a different player ID system than Baseball Savant, a discrepancy that ultimately led us to look into other options. Instead, we used season-summary statistics from Baseball Savant.

This was done using the *scrape_statcast_savant_pitcher*, *scrape_statcast_savant_batter* and *statline_from_statcast* functions in *baseballr*, where a for loop ran the functions for 695 pitchers and 837 batters to create the pitcher and batter dimension tables. The resulting stats are not as rich as what we would have had with Fangraphs, but we still get the "slashline" statistics (batting average, on-base percentage, and slugging percentage) and can calculate rate stats like strikeout percentage and walk percentage. For a glossary of baseball stats, read here.

3. **Chadwick Baseball Bureau**

The other data source, Chadwick Baseball Bureau, was used for the player outrigger table. Their "people.csv" table contains information about players such as name, height, weight, birthday, debut date, country of origin, and handedness and gets updated frequently throughout the season. But we could see there were some noises in the data and so, we collected the MLB player ids from here and used them to fetch player data from the MLB Data API. The table can be accessed from a raw CSV file from their GitHub.

Database Design

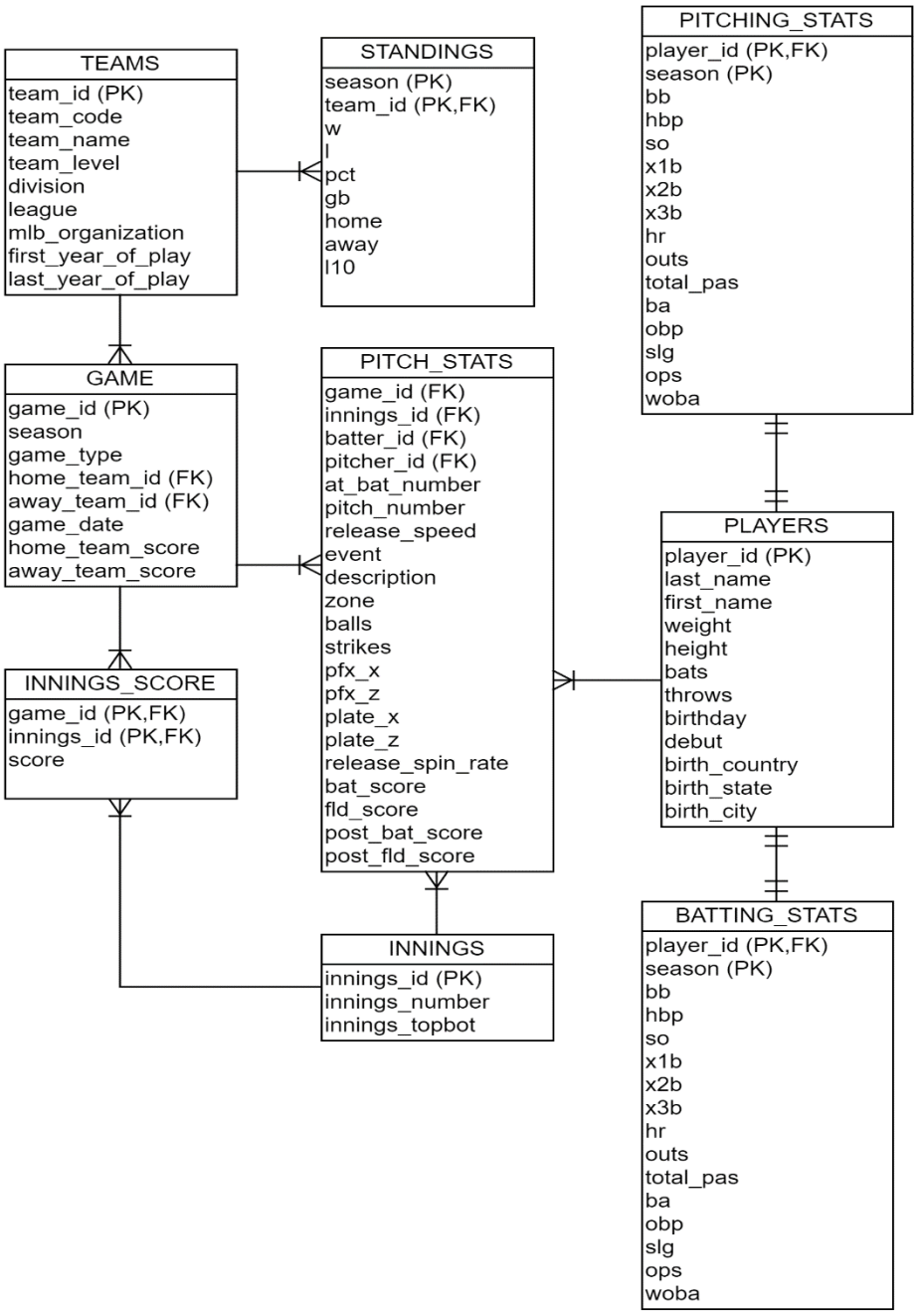
In this section we will look at both our transactional and dimensional schema. While the transactional schema involves normalized tables suitable for OLTP, the dimensional schema de-normalizes and flattens them based on the business process for which the warehouse being designed.

1. **Transactional Model:**

The table below mentions all the entities along with a short description.

Entity Name	Description
PLAYERS	Captures player information
BATTING_STATS	Captures batting stat of a player
PITCHING_STATS	Captures pitching stat of a player
INNINGS	18 entries consist of 9 innings and top/bottom with innings_id for each entry.
INNINGS_SCORE	Innings level score for every game
GAME	Captures list of matches played in the season 2021 until now.

TEAM	List of all the teams participating in the season.
STANDINGS	Standings table for a particular season based on league
PITCH_STATS	This table captures stats for every pitch in a particular game.



Dimensional Model:

To design dimensional model, we took help of the Kimball's four step dimensional design process.

Granularity: Every single pitch thrown during the 2021 season through June 21st

Dimensions:

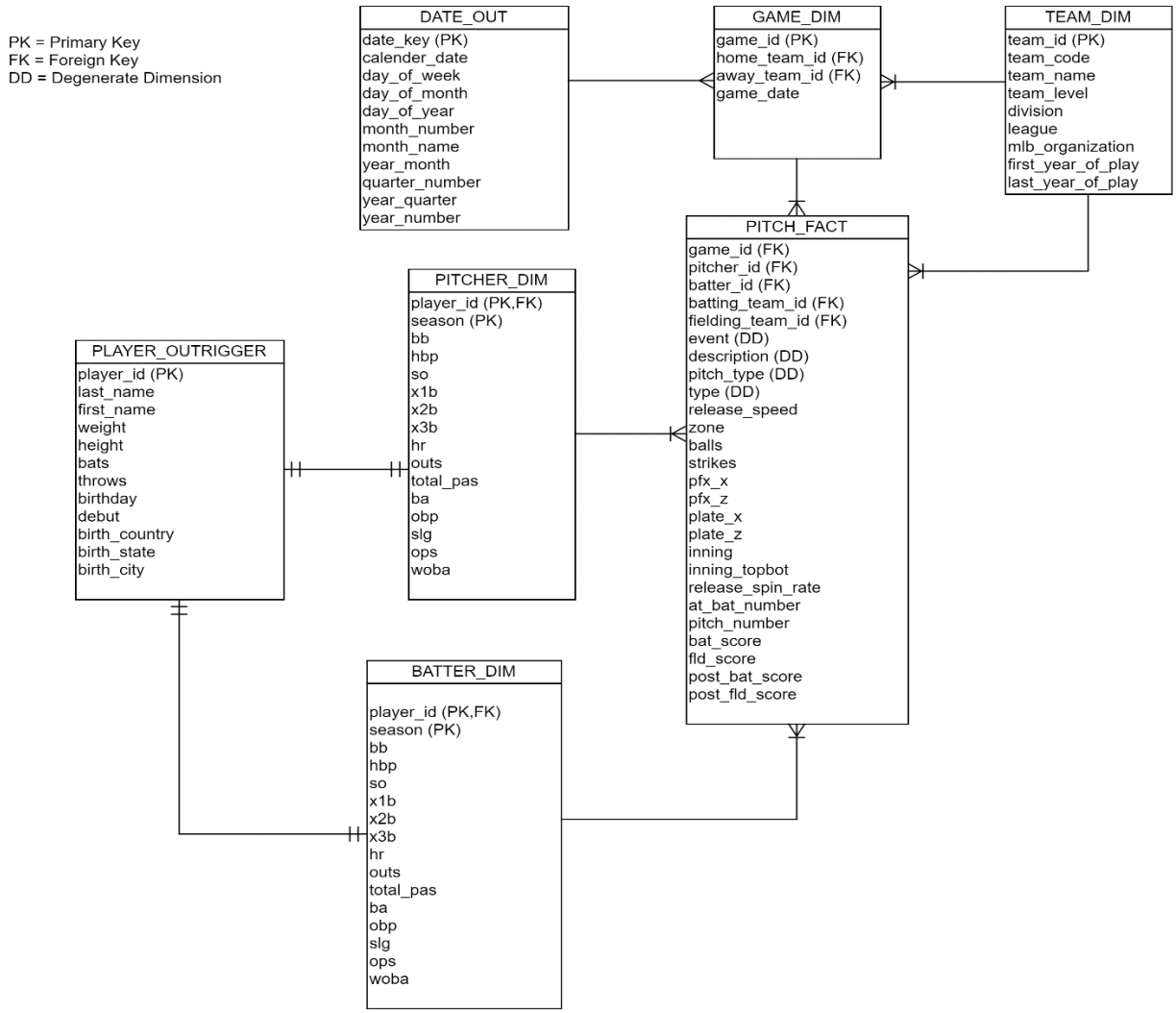
Dimension Table Name	Number of Records
PITCHER_DIM	570
BATTER_DIM	837
GAME_DIM	1076
TEAM_DIM	30

Facts:

Fact Table Name	Number of Records
PITCH_FACT	313275

Outrigger:

Table Name	Number of Records
DATE_OUT	438656
PLAYER_OUTRIGGER	1242



Reporting, Modeling and Storytelling

Query 1: Pitchers with the best average fastball velocity this season. Pitches that are categorized as a fastball include a four-seam fastball (FF), sinker (SI), two-seam fastball (FT), and cutter/cut fastball (FC).

```
SELECT
    pitcher_id, first_name, last_name,
    round(avg(release_speed),1) as fastball_velocity_mph,
    round(avg(release_speed) * 1.60934,1) as fastball_velocity_kph,
    count(*) as total_pitches
FROM
    pitch_fact p LEFT JOIN player_outrigger o
    ON p.pitcher_id = o.player_id
WHERE
    pitch_type IN ('FF','SI','FC','FT')
GROUP BY
```

```
    pitcher_id, first_name, last_name
ORDER BY
   fastball_velocity_mph DESC;
```

	PITCHER_ID	FIRST_NAME	LAST_NAME	FASTBALL_VELOCITY_MPH	FASTBALL_VELOCITY_KPH	TOTAL_PITCHES
1	661403	Emmanuel	Clase	100	161	375
2	594798	Jacob	deGrom	99.2	159.7	584
3	621242	Edwin	Diaz	99.1	159.5	286
4	547973	Aroldis	Chapman	99.1	159.5	276
5	660813	Brusdar	Graterol	99	159.4	22
6	656803	James	Norwood	98.7	158.8	15
7	621237	Jose	Alvarado	98.5	158.5	450
8	657240	Julian	Merryweather	98.2	158	30
9	612434	Miguel	Castro	98.2	158.1	222
10	666808	Camilo	Doval	98.2	158	85

Query 2: This table lets us know what the average spin rate and velocity for each pitch is. I think this is important because as a baseball fan, often you hear what a pitcher's spin rate is without knowing what a good spin rate actually is, with no benchmark to reference it to. This way we get a good idea of knowing what metrics a standard pitch type has. From the results, we see that breaking balls (knuckle-curve, curveball, slider) have higher spin rates, which makes sense because they move more. Bauer units, named after pitcher Trevor Bauer, is spin rate divided by velocity. This is a way to standardize spin rate because faster pitches are going to have a higher RPM. For a glossary on baseball pitch types, view here.

```
SELECT
    pitch_type,
    round(avg(release_spin_rate),0) as spin_rate,
    round(avg(release_speed),1) as velocity,
    round(avg(release_spin_rate)/avg(release_speed),1) as bauer_units
FROM
    pitch_fact
WHERE
    pitch_type IS NOT NULL
GROUP BY
    pitch_type
ORDER BY
    spin_rate DESC;
```

	PITCH_TYPE	SPIN_RATE	VELOCITY	BAUER_UNITS
1	KC	2557	81.1	31.6
2	CU	2540	78.9	32.2
3	SL	2452	84.7	28.9
4	FC	2408	88.6	27.2
5	FF	2307	93.7	24.6
6	FT	2303	89.8	25.7
7	CS	2276	71.9	31.7
8	SI	2151	92.9	23.1
9	SC	1945	82.3	23.6
10	KN	1845	61	30.3
11	FA	1809	74.3	24.3
12	CH	1777	84.8	20.9
13	FS	1465	86.1	17
14	EP	1161	47.2	24.6

Query 3: We can observe the pitchers with best fastball spin rates before June 3rd and compare that to their spin rates after the announcement of cracking down on foreign substances. Keep in mind that this data runs through June 21st. We can look at this through Bauer units, and ironically, Trevor Bauer is the pitcher with the best Bauer units pre-crackdown.

```
SELECT
    pitcher_id, first_name, last_name,
    round(avg(release_spin_rate) / avg(release_speed),1) as
bauer_units
FROM
    player_outrigger o RIGHT JOIN pitch_fact p
    ON o.player_id = p.pitcher_id
    LEFT JOIN game_dim g
    ON p.game_id = g.game_id
WHERE
    pitch_type IN ('FF','SI','FC','FT')
    AND g.game_date_key < 20210603
GROUP BY
    pitcher_id, first_name, last_name
HAVING
    count(*) > 50
ORDER BY
    bauer_units DESC;
```

	PITCHER_ID	FIRST_NAME	LAST_NAME	BAUER_UNITS
1	545333	Trevor	Bauer	31.4
2	506433	Yu	Darvish	30.3
3	476595	Lucas	Luetge	29.6
4	669203	Corbin	Burnes	29.5
5	542888	Shawn	Armstrong	29.3
6	445276	Kenley	Jansen	29
7	458708	Josh	Tomlin	28.9
8	518516	Madison	Bumgarner	28.6
9	458676	Josh	Lindblom	28.6
10	596027	Dillon	Maples	28.5

If we run the same query, but look at spin rate after June 3rd, I can calculate pitchers with the biggest difference.

Pitcher ID	First Name	Last Name	Post-Bauer Units	Pre-Bauer Units	Difference
448855	Junior	Guerra	22.6	26.1	-3.5
641927	Bailey	Ober	22.9	25.9	-3
680704	Nick	Sandlin	22.9	25.8	-2.9
607572	Sam	Howard	23.9	26.7	-2.8
656322	Sam	Coonrod	20.5	23.1	-2.6
445276	Kenley	Jansen	26.5	29	-2.5
675916	James	Karinchak	23.1	25.6	-2.5
669135	Mac	Sceroler	25.7	28.2	-2.5
545333	Trevor	Bauer	29	31.4	-2.4
656638	Alex	Lange	23.3	25.6	-2.3
453192	Andrew	Miller	21	23.3	-2.3
675921	Spencer	Howard	21.9	24.1	-2.2
621076	James	Kaprielian	21.3	23.4	-2.1

Here we can see every pitcher whose Bauer units on fastballs dropped by over 2.0. The most notable pitchers on here are Trevor Bauer, Kenley Jansen, and James Karinchak. Of the 379 pitchers who threw at least 50 fastballs before and after June 3rd, 79 saw a drop of over 1.0 units.

Query 4: It is important to know why exactly pitchers want an increased spin rate. Spin equals movement, so it is insightful to see how spin rates correlate with horizontal and vertical movements of pitches. For reference, sinkers (SI) typically drop more vertically, and from a right-handed pitcher's perspective, a cut fastball (FC) moves from right to left and a two-seam fastball from left to right. Spin rate seems to be most effective on four-seam fastball vertical movement.

```
SELECT
    pitch_type,
    ROUND(CORR((release_spin_rate/release_speed), pfx_z),3) AS
vertical_corr,
    ROUND(CORR((release_spin_rate/release_speed), pfx_x),3) AS
```

```
horizontal_corr
FROM
    pitch_fact
WHERE
    pitch_type IN ('FF','SI','FC','FT')
GROUP BY
    pitch_type;
```

	PITCH_TYPE	VERTICAL_CORR	HORIZONTAL_CORR
1	SI	0.206	-0.122
2	FF	0.248	0.013
3	FT	-0.135	-0.316
4	FC	-0.262	0.212

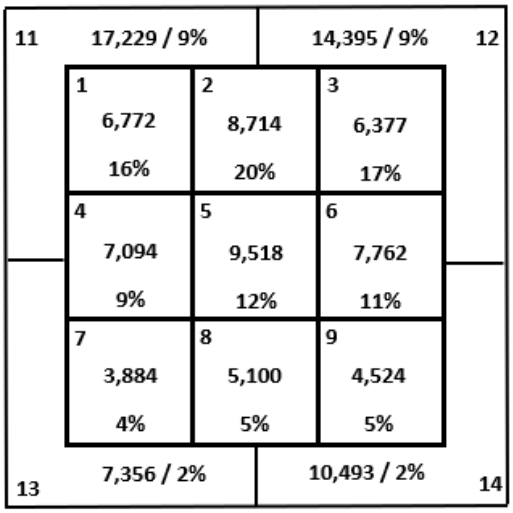
Query 5: Query 4 gives some understanding on a pitching style that has gotten more popular over the years, especially by the Houston Astros starting in 2017. The Astros famously targeted high velocity pitchers with good spin rates and popularized a new style of pitching high in the strike zone with four-seam fastballs, with the idea being that high fastballs are difficult for batters to get "on top of," generating lots of swings and misses. Spin rate (and foreign substances) contributed to this trend because the more movement a high fastball has, the better. I wanted to view the number of pitches in each section of the strike zone and see if this trend is as popular around baseball as it now seems to be, while also comparing how effective four-seam fastballs in each zone are at generating swings and misses. Swinging strike rate (swinging strikes divided by total pitches) is becoming a popular metric, and FanGraphs writer Ben Clemens does a great job of explaining why:

"*If I could see only one pitching statistic, I'd choose swinging-strike rate. That's not to say that nothing else matters; that's decidedly not the case, and there are easy examples of both pitchers who miss bats but aren't effective and pitchers who are effective without missing bats. But as a first pass, swinging strikes are great. Everything else is contextual. Called strike? That's because the batter didn't swing. Foul? It's not always worth a strike. Groundball? The batter could hit it through the defense or find a gap. A swing and miss is absolute*."

```
SELECT
    zone,
    /*count(description)  as whiffs*/
    count(description) as total_pitches
FROM
    pitch_fact
WHERE
    /*description = 'swinging_strike' AND*/
```

```
    pitch_type = 'FF'
GROUP BY
    zone;
```

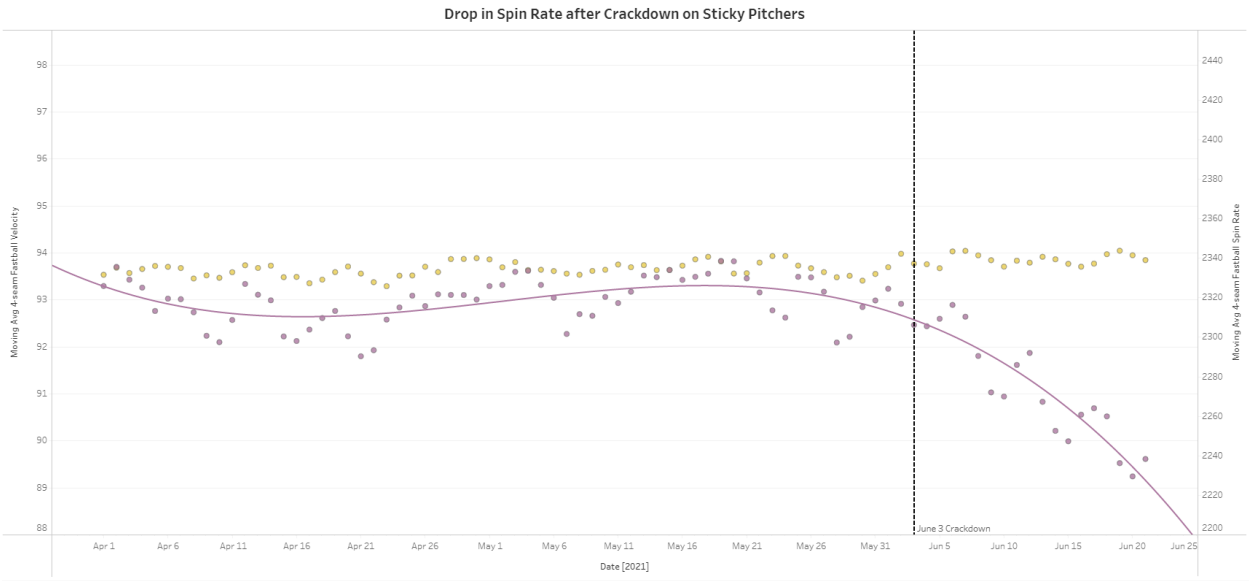
I formatted the results in a picture of a strike zone—the same strike zone format that Baseball Savant uses when it classifies zones for pitches—showing the total number of pitches thrown in that zone and its corresponding swinging strike rate. Zones 1-9 are the strike zone while zones 11-14 are for pitches outside of the strike zone. Keep in mind that this strike zone is from the catcher and umpire's perspective. From these results, we do see that high four-seam fastballs are thrown more than low ones, although the middle three zones of the strike zone are actually more popular than the upper three. It is also confirmed that pitches high in the zone generate more swings and misses. For zones 11 and 12, results are a little inaccurate because hitters are less likely to swing at balls outside the strike zone. In this instance, it may be helpful to have a supplemental strike zone graphic that uses whiff rate (swinging strikes divided pitches swung at) instead of swinging strike rate (swinging strikes divided by all pitches).



Storytelling

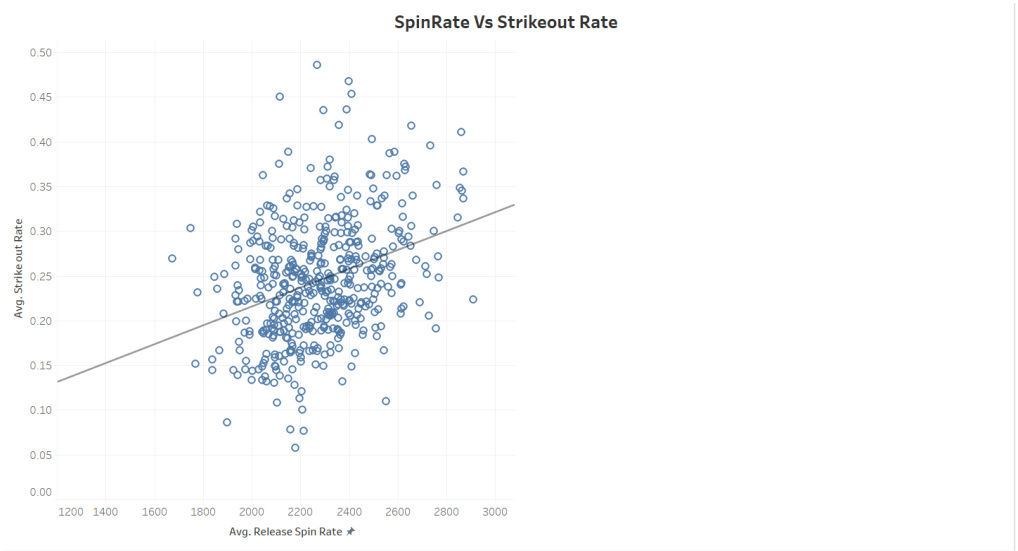
For an interactive view of these visualizations, view them on Tableau Public here.

1. Drop-in Spin Rate after Crackdown on Sticky Substances



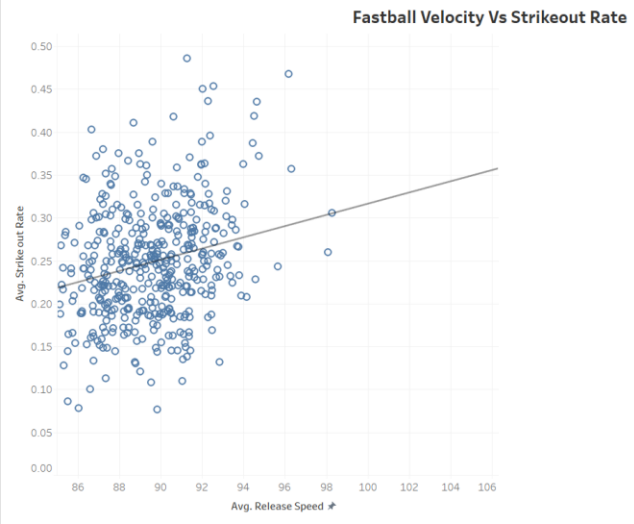
Moving averages are calculated for spin rate and velocity on four-seam fastballs throughout the season. We see that while four-seam fastball velocity (yellow) has stayed relatively the same, spin rate (purple) takes a dive after June 3rd. The data points for spin rate are fitted with a logarithmic curve.

2. Average Four-Seam Fastball Spin Rate Vs Pitcher Season Strikeout Rate



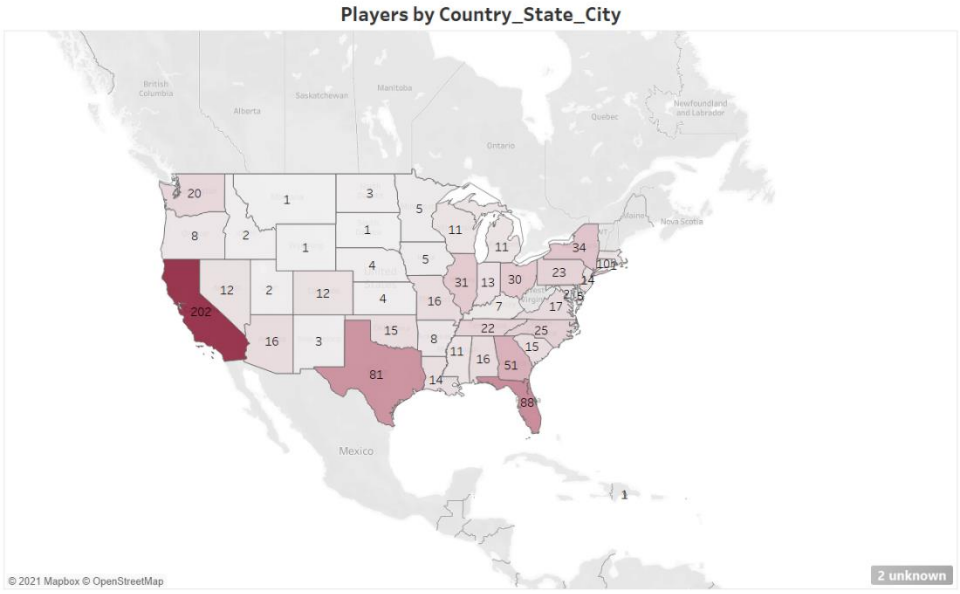
There appears to be some correlative power between a pitcher's fastball spin rate and how well they have performed so far during the season, using strikeout rate (total strikeouts divided by total batters faced).

3. Average Four-Seam Fastball Velocity Vs Pitcher Season Strikeout Rate

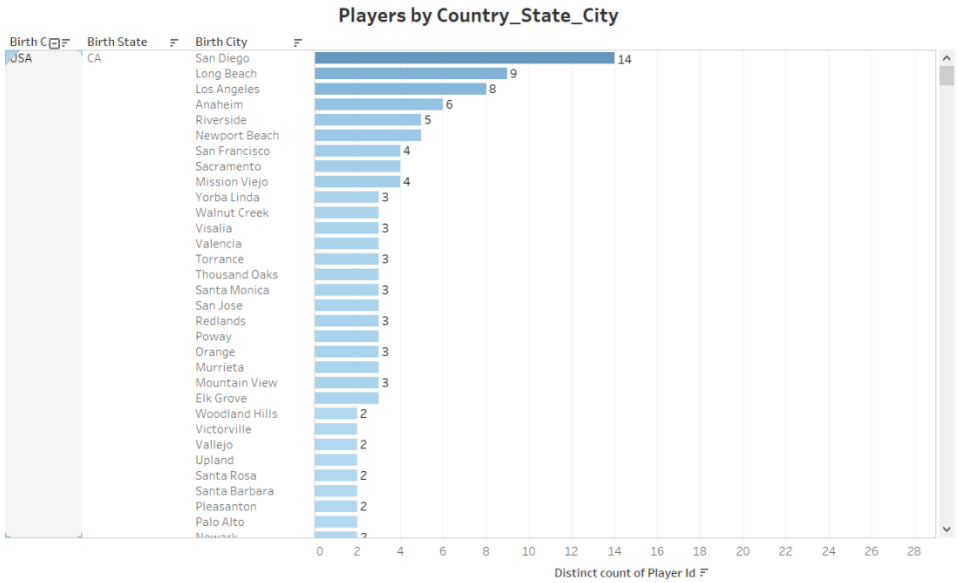


There also is some correlative power between fastball velocity and the pitcher's strikeout rate for the season. These correlation scatter plots are best viewed on the Tableau Public page, where you can select pitch type, and not just four-seam fastballs as seen above.

4. Players by Country State City



Of the Major League Baseball players from the United States, a large portion are from California, Texas, and Florida.



Appendix

1. **RStudio code for generating player's season stats:**

```
lsf.str("package:baseballr")
library(rio)
pitcher_IDs <- import("savant_pitchers.xlsx", sheet = 3)
pitcherID_list <- pitcher_IDs$pitcherID
savant_pitcher_stats <- NULL
for (i in seq_along(pitcherID_list)) {
temp <- NULL
temp <- scrape_statcast_savant_pitcher("2021-04-01", "2021-06-21",
pitcherID_list[i])
temp_stats <- statline_from_statcast(temp)
savant_pitcher_stats <- rbind(savant_pitcher_stats, temp_stats)
}
library(writexl)
write_xlsx(savant_pitcher_stats, "savant_pitcher_R.xlsx")
batter_IDs <- import("savant_pitchers.xlsx", sheet = 4)
batterID_list <- batter_IDs$batterID
savant_batter_stats <- NULL
for (i in seq_along(batterID_list)) {
temp <- NULL
temp <- scrape_statcast_savant_batter("2021-04-01", "2021-06-21",
batterID_list[i])
temp_stat <- statline_from_statcast(temp)
savant_batter_stats <- rbind(savant_batter_stats, temp_stat)
}
write_xlsx(savant_batter_stats, "savant_batter_R.xlsx")
```

2. **Python script to fetch Player Data:**

```
# -*- coding: utf-8 -*-
"""
Created on Sat Jun 26 17:19:27 2021

@author: samar
"""

import pandas as pd
import requests

def date_processor(d):
    d = d.split("T")[0]
    (year,month,day)=d.split("-")
    return str(year)+str(month)+str(day)


player_id_list = pd.read_excel("player_id.xlsx")['player_id'].tolist()

dict = {
            "player_id": [],
            "last_name": [],
            "first_name": [],
            "weight": [],
            "height": [],
            "bats": [],
            "throws": [],
            "birthday": [],
            "debut": [],
            "birth_country": [],
            "birth_state": [],
            "birth_city": []
        }

for id in player_id_list:

    parameters = {
        "sport_code": "\'mlb\'",
        "player_id": "\'"+str(id)+"\'"
    }

    response = requests.get("http://lookup-service-
prod.mlb.com/json/named.player_info.bam", params=parameters)
    response = response.json()
    body = response["player_info"]['queryResults']['row']
    dict['player_id'].append(id)
    dict['last_name'].append(body['name_last'])
```

```
    dict['first_name'].append(body['name_use'])
    dict['weight'].append(int(body['weight']))

dict['height'].append(int(body['height_feet'])*12+int(body['height_inc
hes']))
    dict['bats'].append(body['bats'])
    dict['throws'].append(body['throws'])
    dict['birthday'].append(date_processor(body['birth_date']))
    dict['debut'].append(date_processor(body['pro_debut_date']))
    dict['birth_country'].append(body['birth_country'])
    dict['birth_state'].append(body['birth_state'])
    dict['birth_city'].append(body['birth_city'])

player_outrigger = pd.DataFrame(dict)
player_outrigger.to_csv('player_outrigger.csv', index=False)
```

3. **Python script to fetch Player Data:**

```
# -*- coding: utf-8 -*-
"""
Created on Sat Jun 26 02:35:17 2021

@author: samar
"""

# Fetch mlb team data
import pandas as pd
import requests
import json

def jprint(obj):
    # create a formatted string of the Python JSON object
    text = json.dumps(obj, sort_keys=True, indent=4)
    print(text)


parameters = {
    "sport_code": "\'mlb\'",
    "all_star_sw": "\'N\'",
    "sort_order": 'name_asc',
    "season" :"\'2017\'"
}

response = requests.get("http://lookup-service-
prod.mlb.com/json/named.team_all_season.bam", params=parameters)
response = response.json()
```

```
dict = {
        'team_id':[],
        'team_code':[],
        'team_name':[],
        'team_level':[],
        'division':[],
        'league':[],
        'mlb_organization':[],
        'first_year_of_play':[],
        'last_year_of_play':[]
       }
for i in response["team_all_season"]["queryResults"]['row']:
    dict['team_id'].append(i['team_id'])
    dict['team_code'].append(i['team_code'].upper())
    dict['team_name'].append(i['name_display_full'])
    dict['team_level'].append(i['sport_code'])
    dict['division'].append(i['division'])
    dict['league'].append(i['league_abbrev'])
    dict['mlb_organization'].append(i['mlb_org_abbrev'])
    dict['first_year_of_play'].append(i['first_year_of_play'])
    dict['last_year_of_play'].append(i['last_year_of_play'])

team_dim = pd.DataFrame(dict)
team_dim.to_csv('team_dim.csv', index=False)
```