Final Project

Three True Outcomes in MLB: Strikeout. Walk. Homerun.

Steven Latta, Matt Laken

IST 652 Scripting for Data Analysis
Professor Deborah Landowski
Syracuse University // June 2021

Table of Contents

Introduction

Our project provides an in-depth analysis of the sport of baseball utilizing data from Major League Baseball spanning as early as 1950 to the present day. We intend on showing how both the pitching and batting style has evolved over time as well as how this has impacted the way the sport is played. Our analysis consists of 5 individual parts beginning with a high-level view of the trends. Then, we dig deeper into the way the game's currently being played to investigate how the fastball velocity has changed within the past 10 years. Followed by how the specific pitch location has changed over the past decade, with regards to the batter's box. Then once we visualize the change in pitching zones, we see if this has elicited a response from the batters and altered their overall plate discipline. Lastly, we analyze the lure of launch angle play to determine if it plays a role in the rise in strikeouts. This essentially measures the angle at which the ball leaves the bat after making contact for a hit. Our goal is to see how the change in pitching, if any, has impacted the hitting across the league over the past decade.

Data Sources

Pybaseball, which is an open-source package of data from Statcast, FanGraphs and Baseball Reference. For purposes of this assignment, we utilize various modules from StatCast with data on pitching and hitting tracing back to 1950, with an emphasis on the last 30 years. We installed the PyBaseball package via pip "pip install pybaseball", however, it can also be installed from this git repo:

```
git clone https://github.com/jldbc/pybaseball
cd pybaseball
pip install -e .
```

Pre-processing

Each module is brought in from Pybaseball and called for a specific year range as we do not want to bring in too much data at one time. For the high-level analysis, we were able to bring in 70 years of data starting in the year 1950 spanning to 2020. Due to a large amount of pitching data – every pitch thrown in a season - we brought in a sample month from each year where the data was available. Since the 2020 season only contained 60 games, we chose to use the month of August from the years 2010 through 2020 so that we could have a representative sample from each year. We then viewed the columns for each module to determine what we need for the analysis. We had to create subsets of the data such as only displaying fastballs or batted balls. We then utilized the Pandas groupby function to aggerate our data and assign the specific columns/rows to a new variable.

PyBaseball Module	**Description**	**Questions Used For**
Team_Pitching	Contains pitcher's data such as K, ERA, W, L, and Season for each Team over the course of the MLB history.	Q1
Team_Batting	Contains batter's data such as K, BA, W, L, and Season For each Team over the course of the MLB history.	Q1
Statcast	Contains data at the game level for each pitch thrown In a single game for the history of the MLB.	Q2, Q3, Q5
Batting_Stats	Contains data pertaining to each player batting stats Throughout the history of MLB at the pitch level.	Q4

Method of Analysis

Analysis method 1	Data input	Output
Obtain the avg. strike outs per game per season, team batting average by season, homerun average by season, and team runs per season.	Season aggregation, to find mean of K/9, AVG (batting avg) and HR Used these columns for 1950-2020.	Plotted averages in form of a lin on the scatter plot of data for the duration of time – using matplotlib.pyplot

Analysis method 2	Data input	Output
Obtain avg release speed per year for specific pitch type.	Year aggregation for release type Avg of pitches that are "FF" from 2010-2020.	Histogram showing the average Velocity and the trend (line plot) across the past 10 years.

Analysis method 3	Data input	Output
Obtain pitch location avg for specific pitch type.	Year aggregation for pitch locations X and Z for fastballs thrown in past 10 years	Heat map showing the "Batters Box" with the warmer colors Showing the hotter zones.

Analysis method 4	Data input	Output
Obtain averages of metrics both inside and outside zone for swings.	Year aggregation for swings and Contact recorded inside and outside the zone, as well as walk%.	Multiple line graph showing the plate discipline, as a % of batters swinging the bat and the contact with the ball.

Analysis method 5	Data input	Output
Visualize the results of batted balls by the result, launch angle and launch speed. This analysis will show the launch angle and launch speed that produce the best batted ball results.	Subset of data that only contains batted ball data. Created a new Column to show the result based on the result in scatter plot using the Seaborn plot package.	Scatter plot that shows 3 dimensions: batted ball result, launch angle and launch speed.

Analysis & Interpretation

Our goal is to see the trends in pitching and hitting across the league over the past decade. With the rise in strikeouts and the three true outcomes in hitting – a walk, strikeout, or homerun – we wanted to view the data and provide insight into what factors are contributing to these outcomes. We will examine strikeout, homeruns, and batting averages over the past 70 years. From the results of this high analysis, we wanted to dig into specific pitching stats such as the rise in fastball velocity and pitch location trends using the Statcast package that returns advanced metrics for each pitch thrown and the batted ball result. From this, we hope to be able to tell the story as to why the MLB is in the current state that it is in.

Question 1: How have strikeouts, batting average, runs and homeruns trended in the past 70 years?

The first question set the stage for the project. Strikeouts are at an all-time high and batting average is at an all-time low, so we wanted to view how this has trended over an extended period, from 1950-2019 to be exact. Since 2020 was an abbreviated season in the MLB, we did not include this. We aggregated the data to find the mean strikeouts per game by season. As *Figure1* depicts, there is a clear, rising trend of strikeouts per game. The strikeouts per game doubled in the span of 70 years, going from 4 per game to over 8.

We performed similar analysis on batting average, runs, and homeruns. Batting average has varied jumped around quite a bit throughout the years, but we noticed that from about 2010 onward, there has been a downward trend that has lasted longer than any other dip in the 70-year period. What was interesting to see is that during this same period there was a very sharp increase in the number of homeruns being hit, even more than the steroid era that happened during the late 1990's and early 2000's. And lastly, when looking at the number of runs that are being scored, teams are not scoring many more runs than they have in the past. There was actually a slight dip during the early part of the 2010's and only in the past few years has there been a rise in runs scored across the league.

Interpretation

From this analysis, we think it is safe to say that hitters are willing to sacrifice their batting average going for the homerun. This makes sense as a homerun yields runs, whereas a hit might lead to runs but, also means a hitter must rely on the hitters behind him to get a hit or advance the run through some other means. We think this also explains the rise in strikeouts since hitters are swinging harder to generate the bat speed to hit a homerun and doing this leads to more strikeouts. Does this lead to ultimately scoring more runs? While there has been a rise in runs scored over the past few years, the data shows that teams are not scoring that many more runs than they have historically. Homeruns are a fun part of baseball but the lack of action that comes with this is also hurting the game. There is less action in the game more now than ever and the result is a more boring game to watch as your team now strikes out close to 10 times per game.

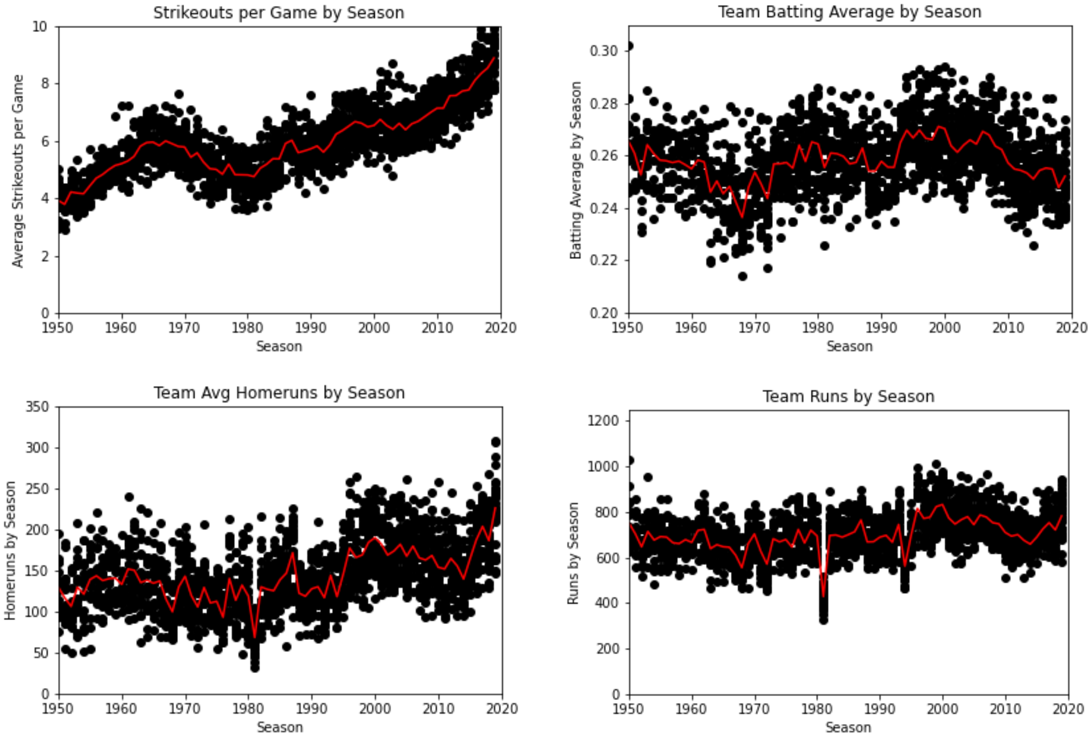

Figure1

Question 2: Has fastball velocity risen as dramatically as it seems?

If you turn on a baseball game and watch to see the radar gun's results from each pitch, it seems as if most pitchers these days are throwing 95 mph or higher, with some pitchers being able to pitch deep into games still hitting close to 100 mph. There is the saying that hitting a 90 mph is one of the hardest things to do in sports. After seeing how dramatically strikeouts have risen in the past 70 years, we wanted to look at the numbers and see how much fastball velocity has risen over the years. We were a bit limited in the number of years we could examine due to technology and data availability, so we had to settle with a 11-year window from 2010 to 2020. Because the Statcast package we used for this returns every pitch thrown in a game, we also chose to pick a sample month from each year to work with. We chose to look at the month of August due to 2020 being a shortened season and August was the only full month in 2020.

After we pulled the month of August from every year, we created a subset of data that only contained fastballs. Using this subset of the data we created a histogram plot showing the distribution of the release speed for every pitch in the data. We then created an aggregated data set on average release speed by season to chart over the top of the histogram in *Figure 2*. We can see the fastball velocity has been slowly rising over the past decade, from 92.3 to 93.4. The difference was not as drastic as we thought it might be but even a 1 mph across the league can make an impact. We then wanted to see if the number of high velocity pitches, pitchers greater than 95 mph or higher has increased. We created another subset of data and flagged each pitch that had a release speed of 95 mph or higher and then summed the total and aggregated this by year. The second chart in *Figure 2* clearly shows that the number of high velocity pitches has nearly doubled in the 10-year span.

Interpretation

While the league average velocity has seen some slight increase over the past decade, its noticeably clear in *Figure 2* that the number of high velocity pitches has dramatically increased. Hitters are at a disadvantage seeing more pitchers throwing harder now more so than ever. This also helps explain the rise in strikeouts and homeruns because when a hitter can guess correctly and make contact with higher velocity pitches, the result will be a harder hit ball, more likely a homerun.

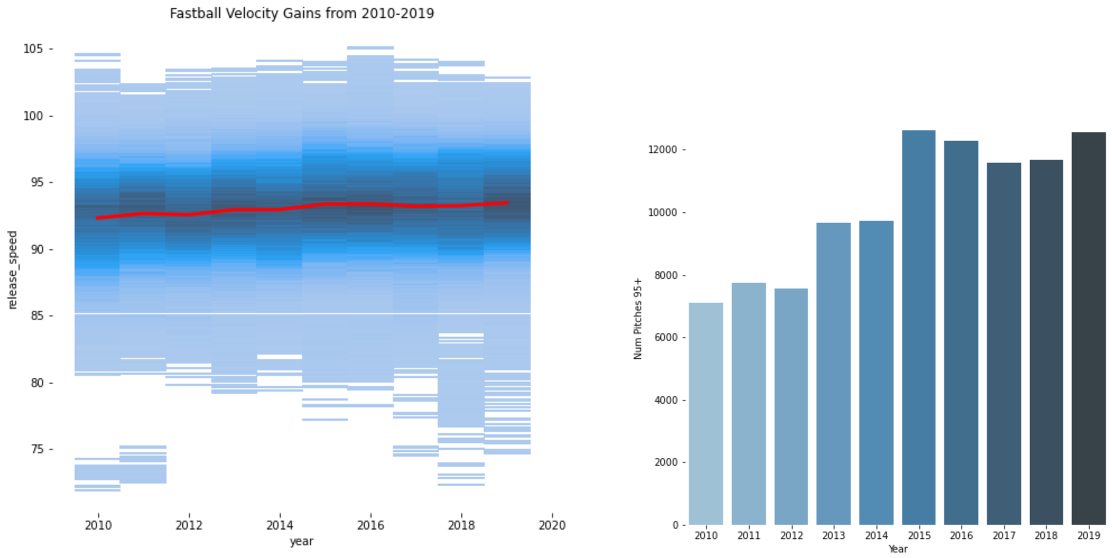

Figure 2

Question 3: Are pitchers adjusting the location of their pitches?

The next piece of analysis we conducted was to determine if pitchers are locating their fastballs differently now than in previous years. To accomplish this, we used the same set of data used for fastball velocity which contained the x and y location where the pitch crossed the plate. We then created three subsets of data for the years 2010, 2015 and 2019 followed by a heat map like chart using the matplotlib hexbin chart type, plotting the location of each pitch in data sets. Since the strike zone is different for every hitter, given their size, we found the average strike zone and created a static strike zone to get a sense of where the pitches are crossing the plate for the hitter.

Interpretation

Upon careful examination of each year in *Figure 3*, you can see that pitchers are throwing more high fastballs now than they did a decade ago. I think pitchers are starting to pitch higher in the strike zone because hitters have a tough time hitting high fastballs. Now that the league has more pitchers with high velocity fastballs it makes sense that pitchers would start to locate these fastballs higher in the zone to challenge the hitter.

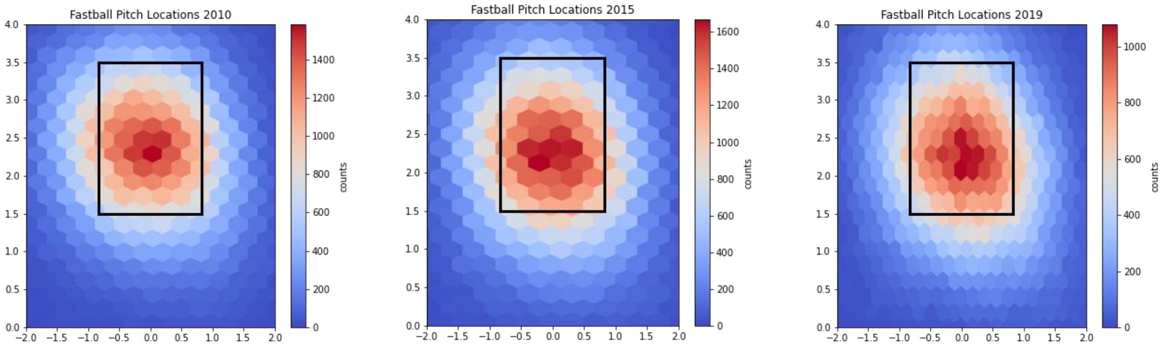

Figure 3

Question 4: Are hitters sacrificing plate discipline for the homerun?

After seeing how high the strikeout numbers and how low the batting averages are, we wanted to analyze the plate discipline metrics to determine if hitters are becoming more aggressive or more patient at the plate. To accomplish this, we needed to import batting statistics from the Pybaseball package. This dataset had more years available for the data that we needed so we were able to expand the scope and bring in data from 2002 to 2020. The fields we chose to examine were the percent of swings at pitches that were outside the zone, inside the zone and the percent of those in which the batter made contact. And lastly, we analyzed the percent of pitchers that were in the zone.

The data preparation was straight-forward for this analysis. The data was already aggregated at the player level so it was a matter of aggregating averages of each measure in question by year. After this was done we created a line chart that plotted the mean values for each year – see *Figure 4*. We found that the in the zone contact percentages remained relatively stable but had a slight decline start in 2016. The in the zone swing percentages was stable as well with values hovering at ~70%. The interesting finding was that the percentage of pitches in the zone has been dropping consistently since 2002. Not surprisingly, the percentage of outside the zone swings and contact has been increasing. The outside the zone contact shows a bit more volatility with a sizeable jump of 20% from 2002 to about 2011. The number of swings outside the zone pitches also rose ~15%.

Interpretation

As pitchers are throwing more pitches outside the zone it makes sense that hitters are responding by swinging at more pitches out of the zone. It is interesting to note that at first the hitters were making more contact with pitches outside of the zone but in 2011, a downward trend begins. Meanwhile, pitchers continued to throw a lower percentage of pitches in the zone. These findings complement the initial analysis of rising strikeouts and decreasing batting averages. Pitchers are forcing hitters to swing at more pitches outside of the strike zone thus increasing the number of strikeouts and lower opposing hitters batting averages.

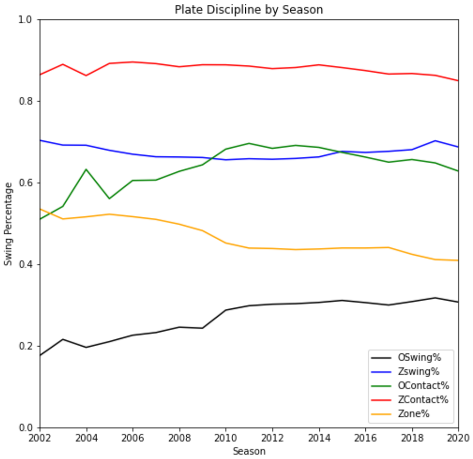

Figure 4

Question 5: How was launch angle trended and what are the batted ball results of launch angle and launch speed?

Having read many articles around MLB, the main topic of hitting revolves around the goal of increasing the launch angle of batted balls. This is an attempt by hitters to get "under" the ball and increase the number of fly balls they hit to achieve extra base hits. An extra base hit is any hit that results in a double, triple or homerun. A fly ball has a greater chance of becoming an extra base than a ground ball does so it makes sense that a hitter would want to strive to do as much damage as they can with their at bat. The last piece of analysis focuses on the launch angle and launch speed of a batted ball and the results that come the combination of the two.

The preprocessing steps to get the data for this analysis involved creating a subset of data that included only pitches that resulted in the hitter making contact with the pitch. After creating this subset, we had to change some data types and create a new column that flagged pitches that resulted in hits so we could plot the results against the launch angle and launch speed of the ball.

The results showed that there is a clear sweet spot for launch angle, launch speed and the desired outcome for the hitter, there is a correlation between the launch angle, launch speed and the outcome of the hit. Nearly all the hits that were hit with a negative launch angle ended up as singles or an out and those that became extra base hits were hit with a launch speed 80 mph or higher. All the homeruns were clustered in one area where the launch angle was between 18 and 50 degrees and the launch speed was greater than 80 mph. It was also interesting to see a distinct band of batted balls that ended up as singles or doubles. The data shows that if your launch angle is high enough, you do not have to hit the ball that hard to get a hit but if the launch angle is too high the result will be an out via a fly ball. If the hitter can square up the baseball more the result ends up being an extra base hit. If the hitter hits the ball at a lower launch angle, they can still be successful, but requires hitting the ball at a higher velocity.

Interpretation

The results in *Figure 5* of the analysis make sense, if you hit the ball hard with enough launch angle you are more likely to get an extra base hit. There are many factors involved that are hard to capture in a single chart, such as the direction the ball was hit, the dimensions of the stadium and defensive shifts designed to take away hits. A hitter can only take what the pitcher gives him but it's clear that trying to achieve launch angle and squaring up the baseball provide the hitter with the best chance of doing as much damage as possible in their at bat.

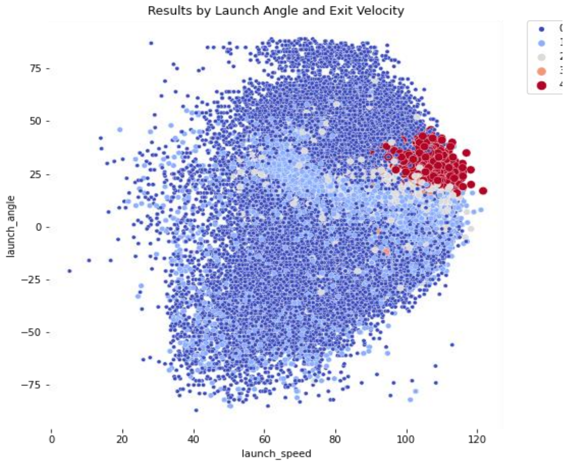

Figure 5

Conclusions

From our analysis, it appears that pitchers have the upper hand. Technology and improvements in physical conditioning have given pitchers the ability to throw a baseball harder than ever before. Hitters have had to adapt to seeing more high-velocity pitches by doing as much damage as possible when they hit a baseball. Players who hit more homeruns are more likely to get paid more than those who do not. It makes sense that hitters would work to retool their swings to achieve more launch angles to hit more extra-base hits and hopefully increase the number of homeruns they hit. But this comes at the cost of striking out more. If pitchers have responded by throwing more high fastballs, then hitters will struggle to make contact and hit these pitches because getting under a high fastball is much harder to do than a pitch that is in the middle or lower part of the strike zone. This, coupled with the uptick in velocity means that hitters are at a clear disadvantage.

The question remains, is this good for baseball? Pitchers are dominating which means that the action in the game has slowed down even more in an already slower-paced game. Will MLB make the adjustments necessary to even out the playing field? There is already talk of lowering the mound again and possibly shrinking the strike zone to create more action. One thing is for certain, pitchers and hitters will continue the cat and mouse game by continually adapting to find ways to swing the advantage in their favor.

Limitations of the Study

A major factor in determining what data to work with from Pybaseball was the need to focus in on a certain period as the datasets are so large and contain very in-depth metrics down to a specific pitch level. We were also limited in using more advanced metrics as the technology that tracks pitch data is relatively new so analysis using these fields was limited to 2010 and on. In future iterations we want to obtain a more in-depth dataset of baseball statistics reflecting the trends in modern baseball such as the use of illegal substances. Also, we want to be able to analyze sets of specific player data and see how they have adjusted their game throughout the duration of their career.

Appendix A: References

Intro to Pybaseball:
https://jamesrledoux.com/projects/open-source/introducing-pybaseball/

Pybaseball Repository:
https://github.com/jldbc/pybaseball

MLB Glossary:
https://www.mlb.com/glossary