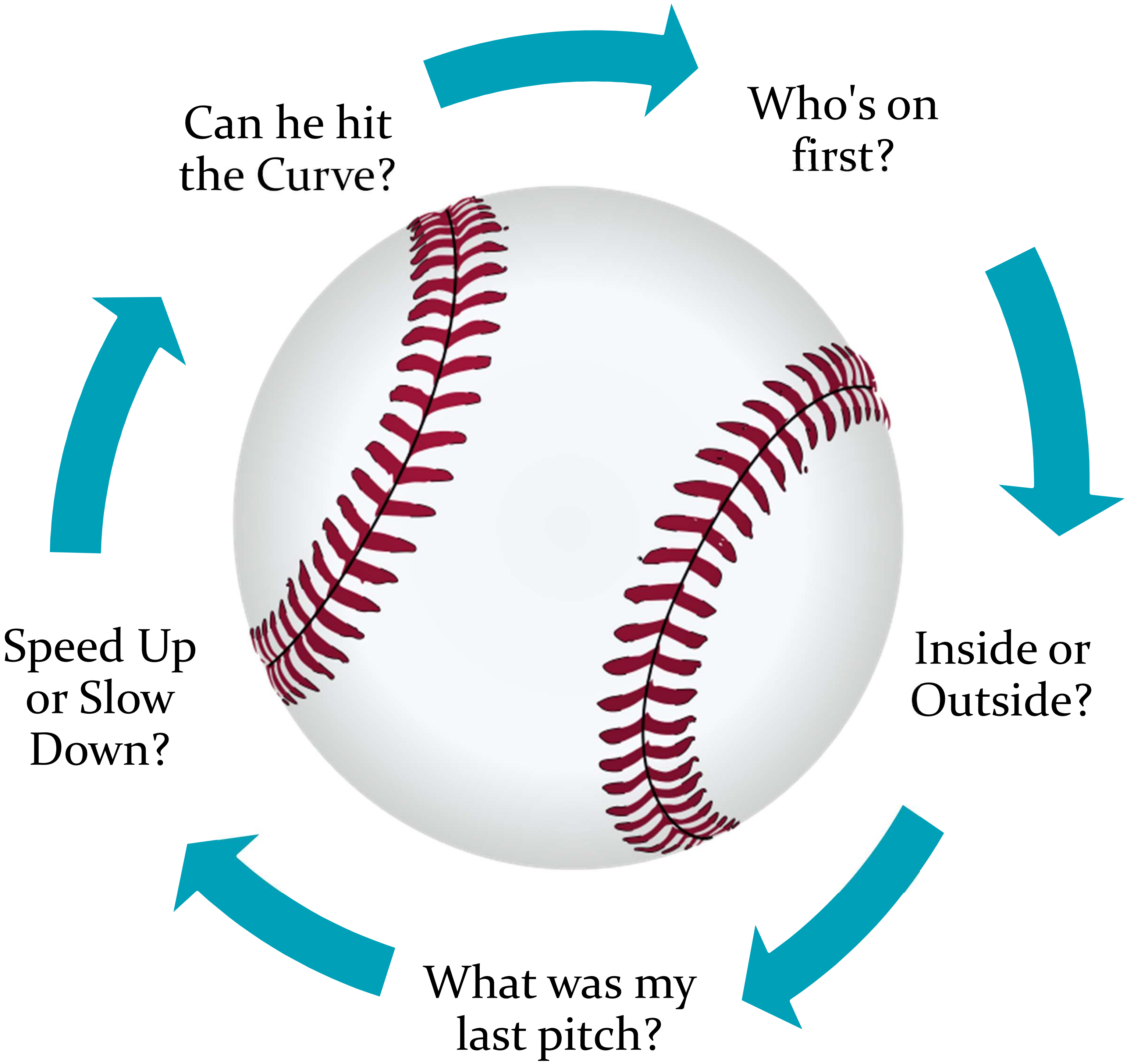



# The Evaluation of a Pitch

A CLOSER LOOK AT DATA-DRIVEN DECISIONS IN BASEBALL

J.C. Surber | TruMedia 2017 MLB Hackathon | February 6, 2017

# INTRODUCTION

Albert Einstein once said, "The definition of genius is taking the complex and making it simple." He was also credited with saying, "If you can't explain it to a six year old, you don't understand it yourself." As I began working on this project, these two quotes resonated strongly with me. With this report, my goals were simple:

1. Answer a group of questions that I've always been curious about utilizing real, live data from Major League Baseball (MLB) statistics.
2. Be able to answer these questions in a way that is simple and easy to understand (even for a six year old).

I am not a programmer and I don't work in the baseball industry. I'm just a novice fan who likes the movie "Moneyball" and enjoys data-driven exercises. Because of this, I stuck to a very simplistic method of analyzing the data provided in an easily digested format. There was no coding or programming as part of this exercise. There was no complicated math. There ARE a lot of data tables and charts, along with some interactive files that helped me to draw the conclusions you will find in the subsequent sections of this report. These files are all included as part of this report. All of the calculations were performed by hand, calculator, or using Microsoft Excel (so no wizardry there either).

I will say, **I had a blast** participating in this competition and I hope that is evident throughout the hard work that went into developing and refining this report.

So, about these questions I was hoping to answer…

# WHAT QUESTIONS ARE WE TRYING TO ANSWER?

As stated above, the goal of this report was to simplistically answer a cadre of questions related to baseball, specifically pitching, using data provided to enhance decisions for future pitching programs and plans. The questions I was hoping to decipher were:

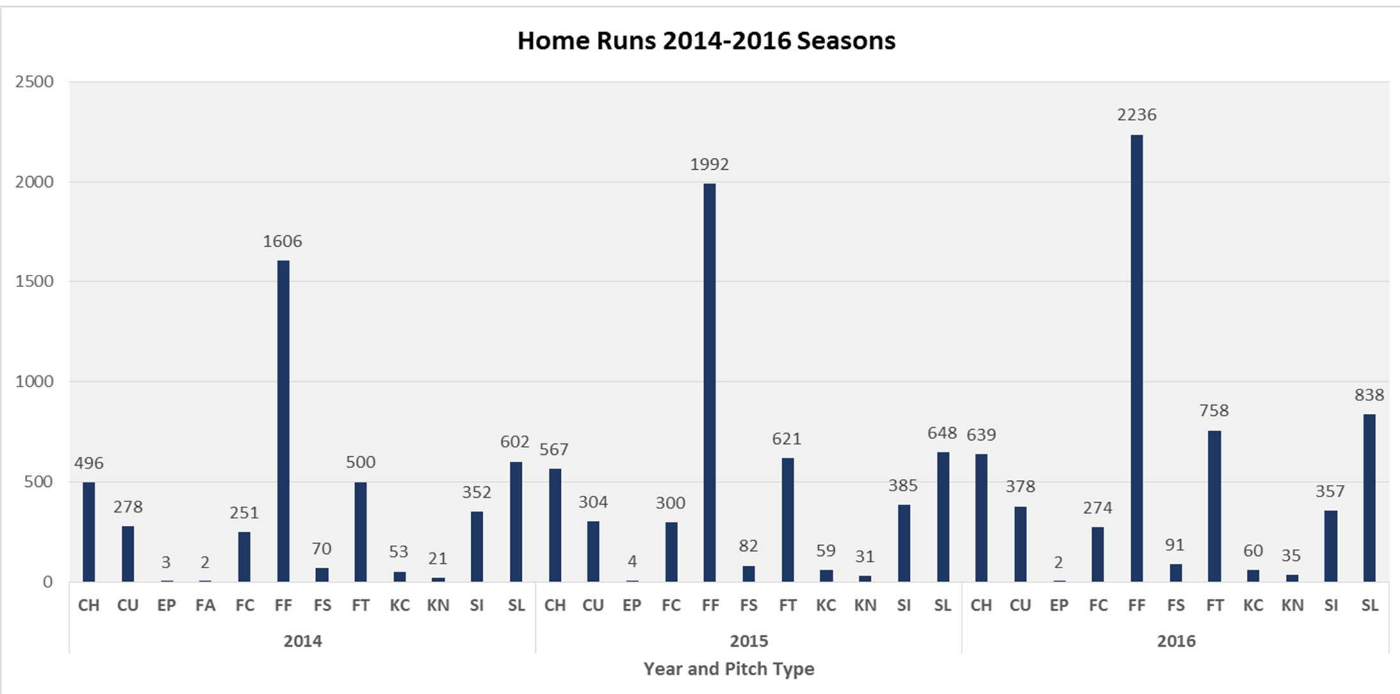
1. What type of pitches tend to yield **home runs** to a batter?
2. In cases where a pitcher faces **a full count** (3 balls, 2 strikes) who were the **most successful batters** and which pitches would a pitcher want to **avoid?**
3. Do umpires really have **biases against certain players or pitches**, and is there any advantage that these biases might produce?
4. Does it matter if a pitcher and/or batter are **right or left handed**?

All of these questions and the corresponding hours of research and conclusions are provided in sections below. There is a detailed explanation for every step that was taken to get to the conclusion with tables, charts, and calculations along the way. I have provided all of the support files for additional reference and to allow you to interact with the data that I filtered and sorted.

# SECTION 1: WHAT TYPE OF PITCHES TEND TO YIELD THE MOST HOME RUNS?

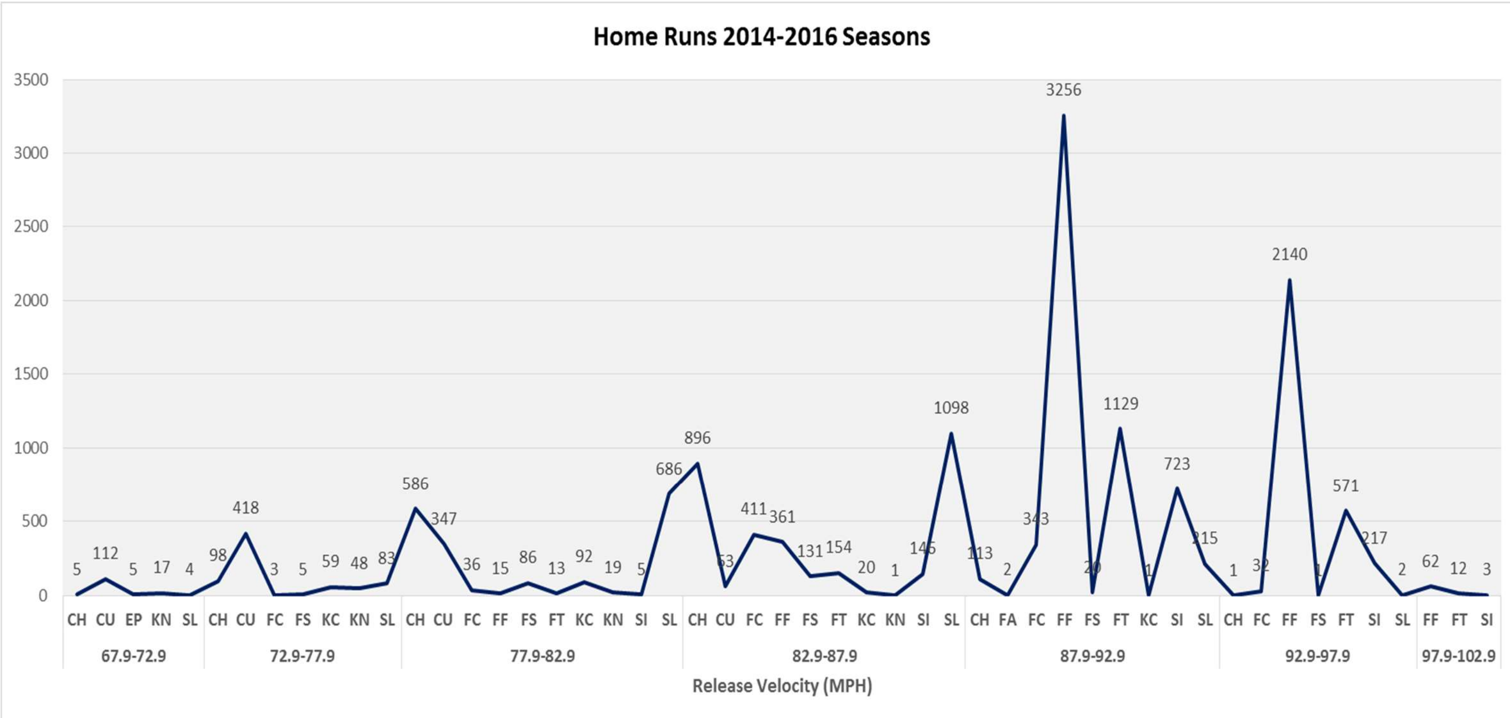
Home runs are a pitchers worst nightmare. They can swing the score of a game in a team's favor in one pitch. They can destroy the confidence of a pitcher almost instantly. They can bolster the confidence of a batter, again, instantly. No pitcher ever wants to give up a crucial home run in a game where the pressure is on.

So, looking at this assignment through the eyes of a pitcher, I wanted to discern which pitches over the course of the 2014-2016 MLB seasons yielded the most home runs. I filtered out a number of different criteria provided and assessed the outcomes based on these pitching inputs. Please reference the file titled "Home Runs" for an interactive excel file. The first step I took was to filter out what types of pitches equaled home runs (excluding unknowns and zero values). The chart for each of the last three seasons can be found below:



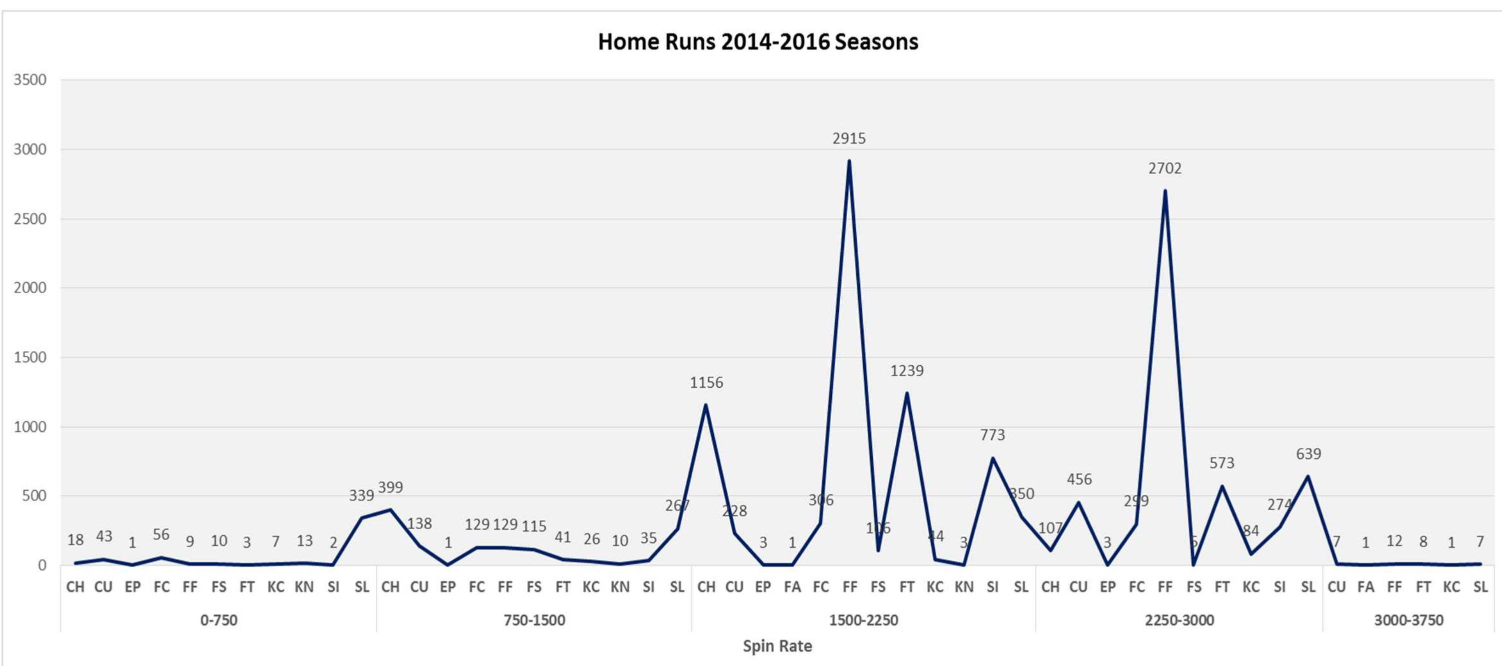

One of the biggest findings from this chart is the fact that total home runs have increased every season over the last 3 seasons. Many of the same pitches were the culprits of home run results with the **Four Seamer, Slider, Two Seamer, and Changeup** accounting for the overwhelming majority of home runs.

The next step was to filter and sort the data to uncover other criteria that would show a correlation to home runs. The first was a look at release velocity and its production of home runs for the past three seasons. That chart can be found on the next page:



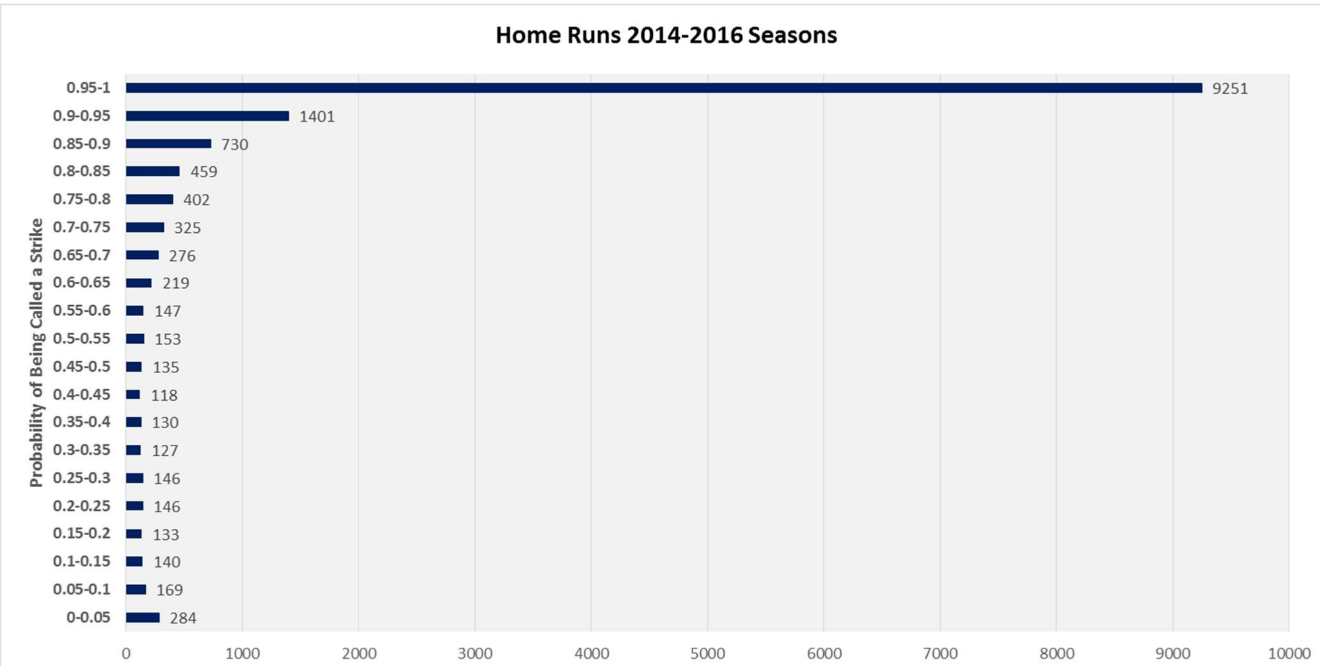

Immediately, the four seamer in the 87-97 MPH range jumps out. This isn't shocking as the four seamer is typically thrown in this range and has a much higher frequency of pitches thrown leading to a higher probability of home run potential. Interesting that the slider had the most home runs in the 82-89 range. If you can throw a changeup in the 87 MPH range, you have a good chance of avoiding a home run fate.

Next, I wanted to examine the spin rate of pitches vs. home runs produced to see if there was any correlation over the past three seasons. That chart can be found below:



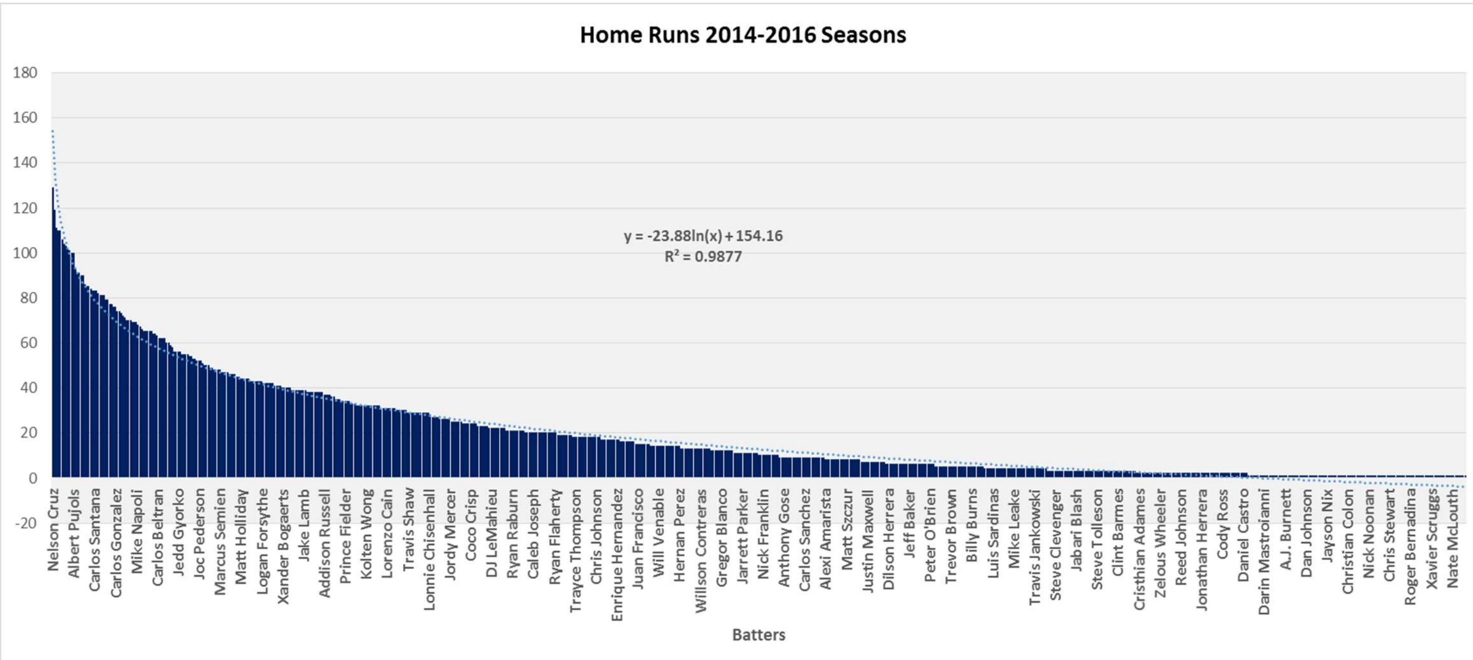
There wasn't a strong correlation from the spin rate other than the most home runs were hit between 1500-3000. Again, most pitches thrown fall in this range which certainly leads to this higher frequency.

I also wanted to examine the TruMedia strike probability model and see how the home runs stacked up against the probabilities of the model output. That chart can be found below:




As one would expect, the odds of the pitch being a strike have a direct relationship with home runs. However, what really surprised me was that over **62%** of the home runs were hit when the probability of a strike fell in the **0.95-1 range**. What this told me is that batters are not some inhuman species able to hit a multitude of pitches at various locations for home runs. They are waiting on pitches nearly smack in the middle of the strike zone and capitalizing on those. This begs the question as to whether or not a pitcher should just attempt to avoid direct middle pitches and aim for other "softer" areas of the strike zone. This would be an interesting hypothesis to test. One that is understandably difficult to try as the ability to control pitches that well is a special art.

I stumbled across an interesting trend when assessing batters and who were the top home run hitters over the past 3 years. When I let the data reflect every batter, the chart below was produced:




The chart does not include every batter's name over the past three years that hit a home run, but their data is captured. The interesting thing that came from this chart was that I was able to create a trendline for the logarithmic curve that had nearly a 99% fit with the data. It produced the equation:

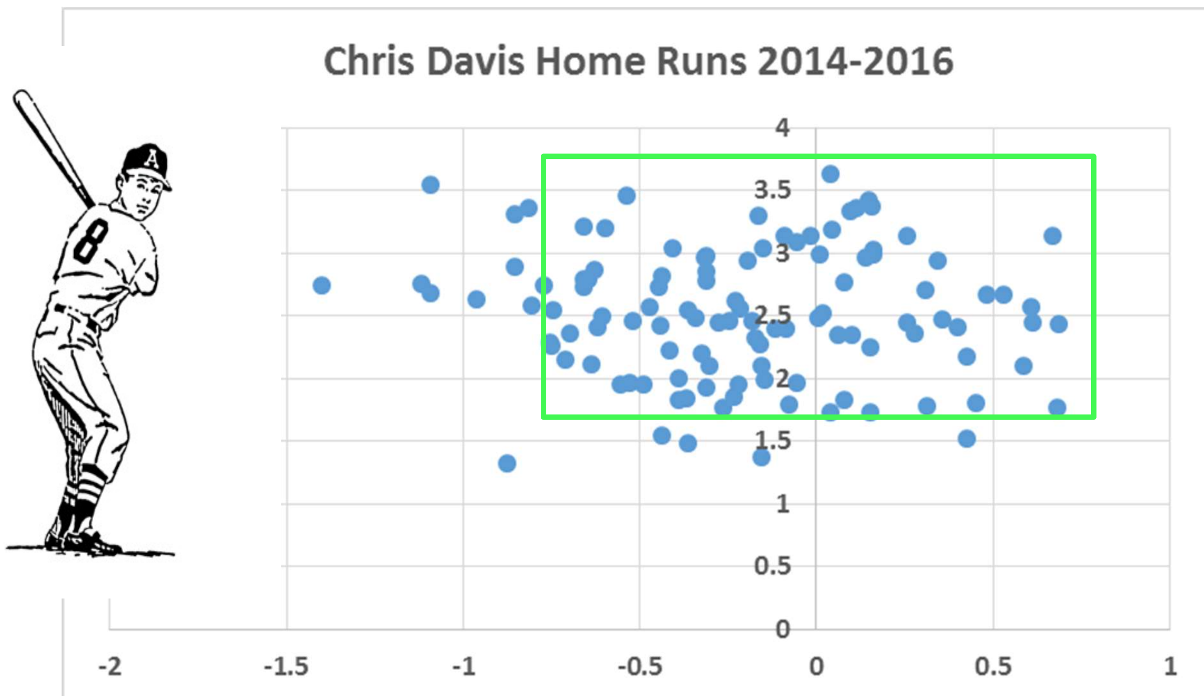
$$Y=-23.88\ln(x) + 154.16$$

This means, you could enter in values for x or y and determine how many batters would hit a certain threshold of home runs over a three season span. For instance, using this data, let's say I wanted to understand how many batters would hit **80 or more home runs** over a three season stretch. I would simply set y=80 and the equation would look like this:

$$80=-23.88\ln(x) + 154.16$$

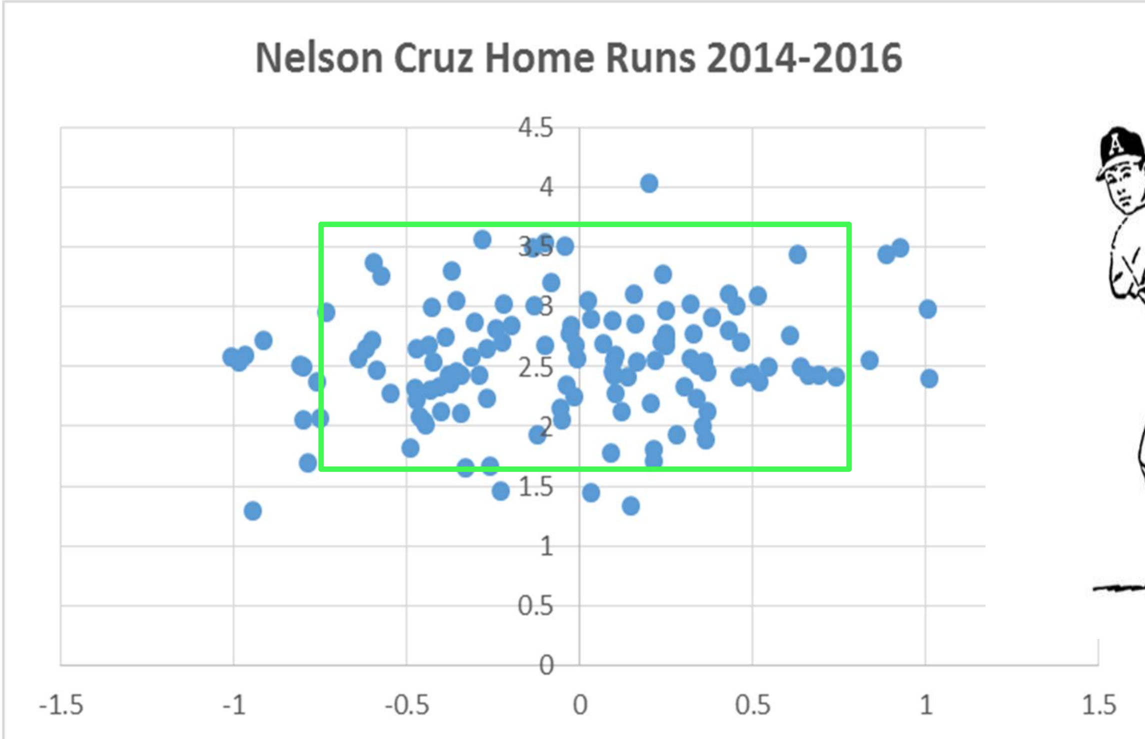
Solving for x gives you approximately 22. Therefore, you would expect around 22 players to hit over 80 home runs over the course of a three season period. Not too bad news for the pitchers out there, unless you're having to face one of those 22 batters on a regular basis. This type of information is useful for people who draft and trade players knowing the expectations of the players they select vs. the pool of talent out there. You could expand this exercise to focus on a specific team and perform the same exercise to help a general manager forecast expectations of batters and performance.

The final thing I wanted to look at was location of pitches and their correlation to home runs.  I sorted out specific batters to analyze their home run pitches and where they were located as they crossed the plate.  The first player I analyzed was Chris Davis.  Chris came in 3rd in home runs over the past three seasons with 111.  Also, I went to high school with Chris and know him personally as we both grew up in Longview, TX (everyone is extremely proud of the success he has had as a professional ball player).  The location of pitches chart for his home runs can be found below:




Since Chris is a left-handed hitter, you can immediately notice that he hits far more home runs **down and inside of his strike zone** (strike zone is outlined by the green box) than high or outside.  So, if I was a pitcher facing Chris, and I wasn't looking to give up a home run, I would test him **high and/or outside** and see what type of results I received.

The final batter I wanted to analyze was none other than the top home run hitter over the past three years, Nelson Cruz.  Cruz came in with the most home runs hit over the past three years with 129.  I don't know him personally, but I do know the impact he had on many Texas Rangers fans with his debacle in the 2011 World Series.  Nonetheless, he has an impressive swing and his location of pitches chart can be found on the next page:




This is a tough one.  Cruz seems to have the ability to hit home runs from most sides of his strike zone. As a right-handed batter, I would most likely test his abilities to hit **low and inside** as he doesn't have any data points that show he can hit home runs from that location**.**  **High and outside** would be the other approach, but with a guy like Cruz, you are truly hoping for the best.

## <u>Conclusion:</u>

Home runs are a tough event to analyze.  There are some obvious conclusions from this data study such as most of the home runs are hit from the strike zone, the four seamer, which is the most common pitch thrown in baseball, yields the most home runs, and fast pitches tend to yield more home runs than slow pitches.  From the batter performance curve, we quickly realize that not every player is going to hit record home runs every year (Nelson Cruz, the top performer, averaged just over **43 home runs** over the last three years).  These numbers are impressive, but they aren't a common event for the number of pitches thrown over the course of a season.  The real analysis to utilize the concepts and charts above for individual players, and develop a pitch plan against them based off of the findings, similar to the Nelson Cruz and Chris Davis analyses.  This would lead to an understanding of what the lowest risk pitches, speeds, and locations are for any given batter that one might face.  THAT is information worth having as a pitcher.

## SECTION 2: WITH A FULL COUNT (3 BALLS, 2 STRIKES), WHICH BATTERS WERE MOST EFFECTIVE AND WHICH PITCHES WERE THEY MOST EFFECTIVE AGAINST?

Sounds like a torturous tongue twister!  The premise of this section was to analyze pitches thrown with a 3-2 count over the past three seasons.  The idea was to look at the data through a pitchers eyes to determine what the optimal pitches would be given a particular batter and his previous decisions and performance in 3-2 situations.

The first step I took was to filter out the Top 20 performing batters against a full count for each of the three regular seasons.  I ranked them in a category called "Hits" which was comprised of Singles, Doubles, Triples, Home Runs, and Walks (excluding Intentional Walks, the batter doesn't get credit for those). The results from these rankings can be found below:

| 2016 Reg Season Batters | Hits |
|---|---|
| Mike Trout | 87 |
| Joey Votto | 85 |
| Brandon Belt | 81 |
| Josh Donaldson | 72 |
| Wil Myers | 69 |
| Paul Goldschmidt | 69 |
| Jose Bautista | 68 |
| Kris Bryant | 67 |
| Carlos Santana | 67 |
| Jayson Werth | 67 |
| Mike Napoli | 64 |
| Dexter Fowler | 64 |
| Jonathan Villar | 63 |
| Edwin Encarnacion | 63 |
| Chris Carter | 61 |
| George Springer | 61 |
| David Ortiz | 60 |
| Christian Yelich | 59 |
| Nelson Cruz | 59 |
| Cesar Hernandez | 58 |
| Bryce Harper | 57 |

| 2015 Reg Season Batters | Hits |
|---|---|
| Joey Votto | 108 |
| Jose Bautista | 83 |
| Bryce Harper | 70 |
| Kris Bryant | 70 |
| Mike Trout | 68 |
| Josh Donaldson | 68 |
| Carlos Santana | 67 |
| Chris Davis | 67 |
| Andrew McCutchen | 66 |
| Curtis Granderson | 65 |
| Manny Machado | 64 |
| Matt Carpenter | 64 |
| Paul Goldschmidt | 61 |
| Alex Rodriguez | 57 |
| Shin-Soo Choo | 57 |
| Dexter Fowler | 56 |
| Joc Pederson | 56 |
| Edwin Encarnacion | 55 |
| Miguel Sano | 54 |
| Brandon Belt | 53 |
| Justin Upton | 53 |

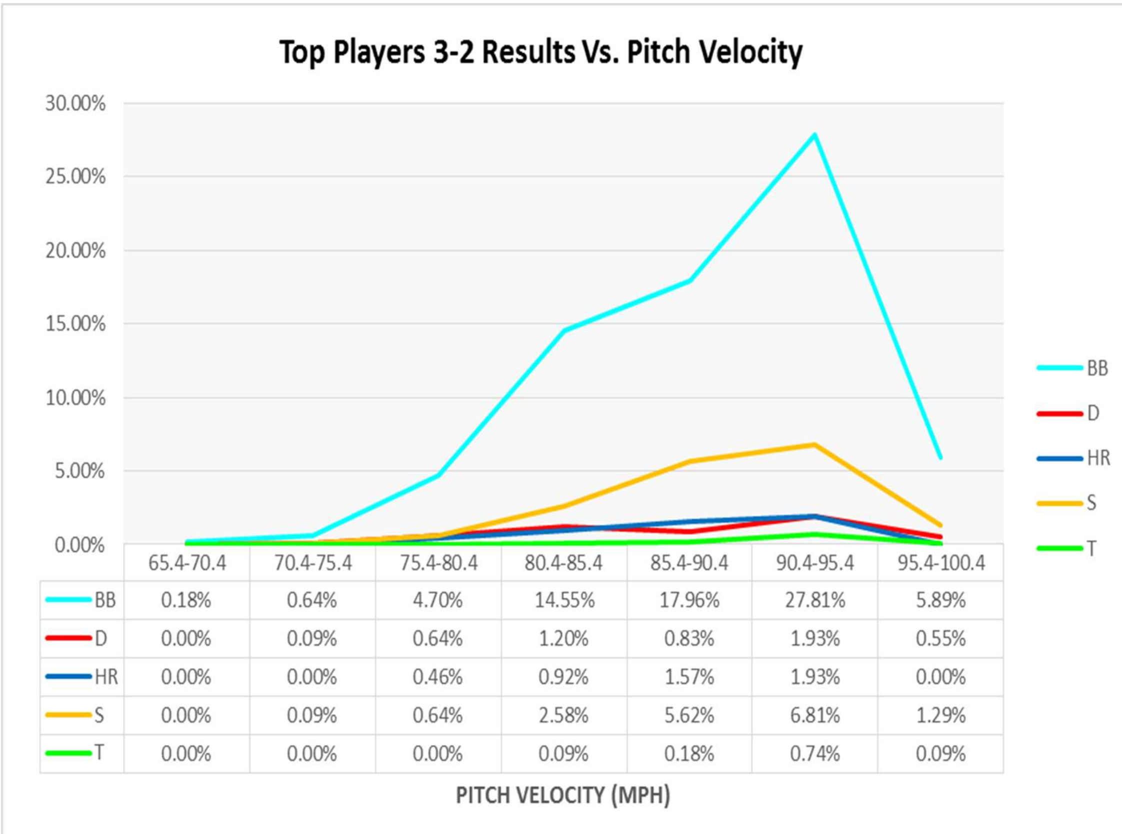
| 2014 Reg Season Batters | Hits |
|---|---|
| Freddie Freeman | 73 |
| Mike Trout | 73 |
| Matt Carpenter | 72 |
| Carlos Santana | 68 |
| Christian Yelich | 66 |
| Jayson Werth | 65 |
| Jose Bautista | 64 |
| Andrew McCutchen | 61 |
| Mike Napoli | 59 |
| Alex Gordon | 58 |
| Jason Heyward | 55 |
| Brian Dozier | 55 |
| Ryan Howard | 55 |
| Robbie Grossman | 54 |
| Adam LaRoche | 53 |
| Brett Gardner | 52 |
| Jhonny Peralta | 52 |
| Trevor Plouffe | 51 |
| Matt Holliday | 51 |
| Dexter Fowler | 51 |
| Anthony Rendon | 50 |

Right away, I noticed multiple players appearing across multiple seasons. Clearly, certain players have had consistent success against pitchers on full count pitches. The six names that jumped out were:

1. Mike Trout
2. Jose Bautista
3. Joey Votto
4. Andrew McCutchen
5. Carlos Santana
6. Mike Napoli

As a pitcher, these players are the individuals you never want to face with a full count. But, obviously, that isn't an option. Therefore, I wanted to know what type of pitches these individuals had success and struggles with in full count scenarios. All of the following data and charts are based off the **combined data performance of the 2014-2016 seasons when these six individuals finished in the top 20 in my metric "hits."**

The first criteria I wanted to look at was the breakdown of the results based on the opposing pitchers release velocity. That chart can be found below:

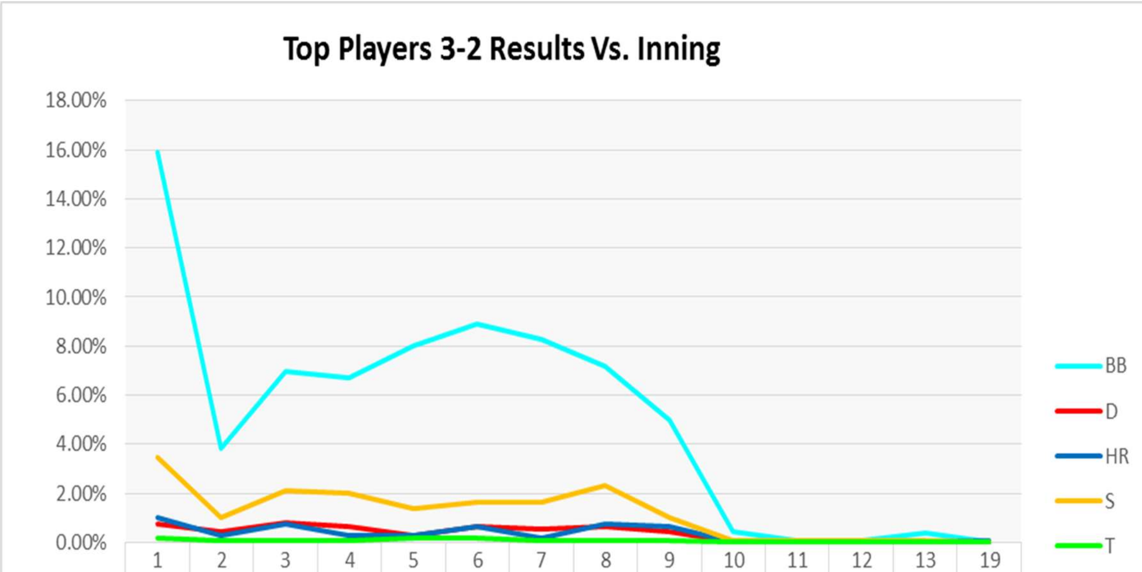



| | 65.4-70.4 | 70.4-75.4 | 75.4-80.4 | 80.4-85.4 | 85.4-90.4 | 90.4-95.4 | 95.4-100.4 |
|---|---|---|---|---|---|---|---|
| BB | 0.18% | 0.64% | 4.70% | 14.55% | 17.96% | 27.81% | 5.89% |
| D | 0.00% | 0.09% | 0.64% | 1.20% | 0.83% | 1.93% | 0.55% |
| HR | 0.00% | 0.00% | 0.46% | 0.92% | 1.57% | 1.93% | 0.00% |
| S | 0.00% | 0.09% | 0.64% | 2.58% | 5.62% | 6.81% | 1.29% |
| T | 0.00% | 0.00% | 0.00% | 0.09% | 0.18% | 0.74% | 0.09% |

There were a few interesting facts that can be seen from this chart:

1. These players hit **no home runs** once the pitch speed was over 95 mph. However, they hit the most homeruns in the 90-95 mph range.
2. The majority of the data can be captured in the walks between **80-95 mph**. In fact, those three columns of walks account for just over **60%** of the pitches. This can be for a number of reasons such as the players are very focused on these pitches and looking for bad pitches. Or, maybe pitchers are just afraid to throw strikes as all of these batters are very good hitters.

The second criteria I wanted to review was the results over the course of all innings. That chart can be found below:

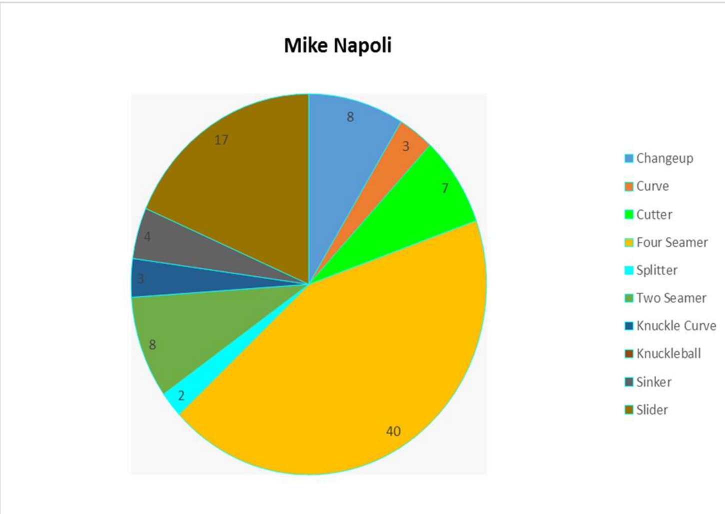


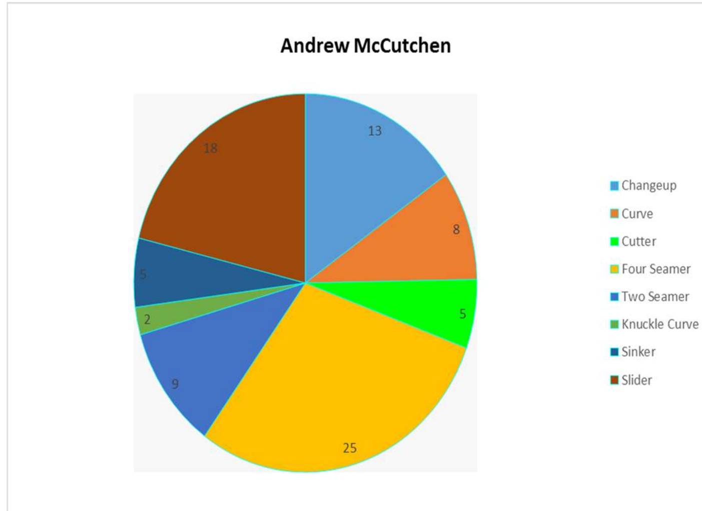

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BB | 15.90% | 3.86% | 6.99% | 6.71% | 8.00% | 8.92% | 8.27% | 7.17% | 4.96% | 0.46% | 0.09% | 0.09% | 0.37% | 0.00% |
| D | 0.74% | 0.46% | 0.83% | 0.64% | 0.28% | 0.64% | 0.55% | 0.64% | 0.46% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| HR | 1.01% | 0.28% | 0.74% | 0.28% | 0.28% | 0.64% | 0.18% | 0.74% | 0.64% | 0.00% | 0.00% | 0.00% | 0.00% | 0.09% |
| S | 3.49% | 1.01% | 2.11% | 2.02% | 1.38% | 1.65% | 1.65% | 2.30% | 1.01% | 0.09% | 0.09% | 0.09% | 0.09% | 0.00% |
| T | 0.18% | 0.09% | 0.09% | 0.09% | 0.18% | 0.18% | 0.09% | 0.09% | 0.09% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |

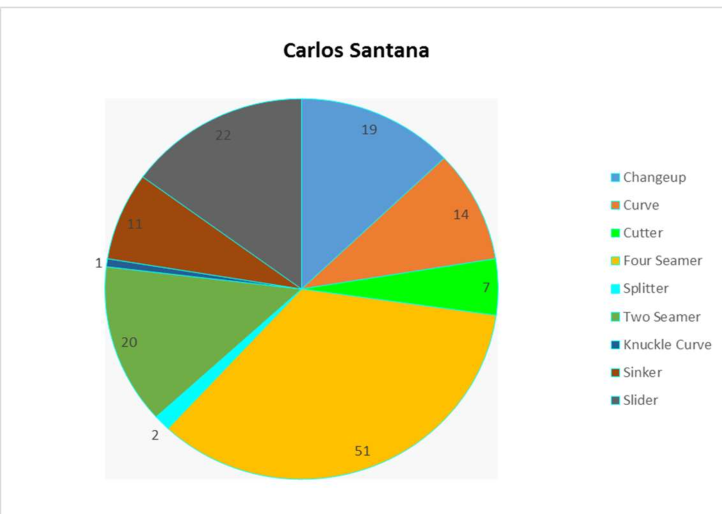
There were a few observations from this chart as well:

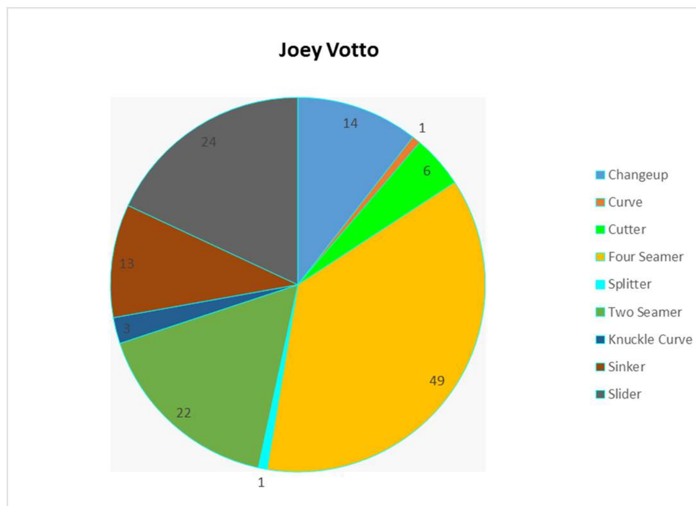
1. The overall trend across the board is that there are a lot of full counts in the early innings and **decrease as the game progresses**. It's ironic that the most singles, walks, and homeruns were all in the first inning. These batters must be laser focused in the first inning, trying to take the pitcher as deep as possible.
2. Pitchers clearly pitch away from these strong batters as the **home run numbers are minimal, while the walk numbers are higher**, even while decreasing as the game progresses.

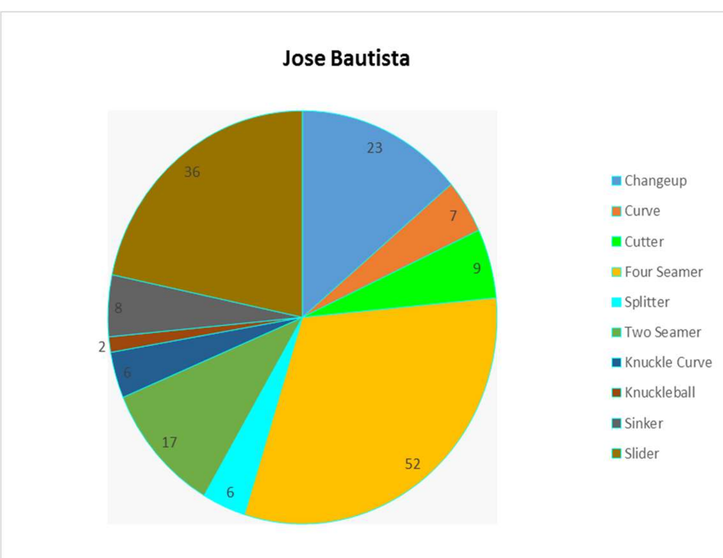
The third criteria to assess was the variety of pitches that resulted in **walks** for each of the batters in full count situations. This would help determine the pitches each batter had a good eye for and should be avoided if throwing junk. That dashboard can be found below:

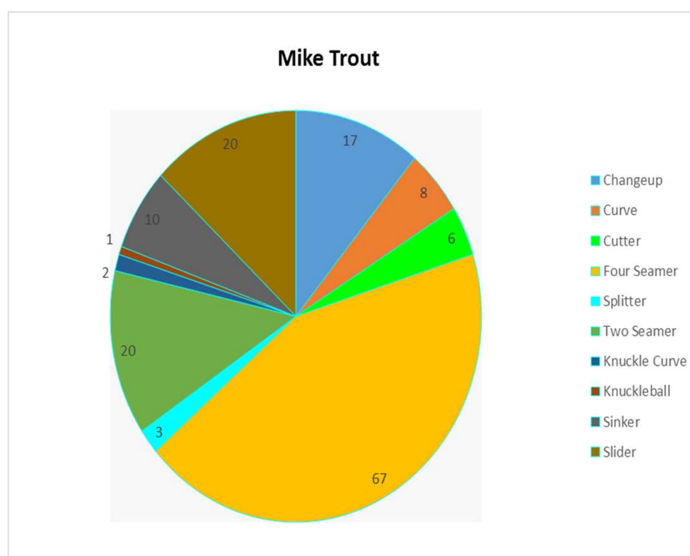


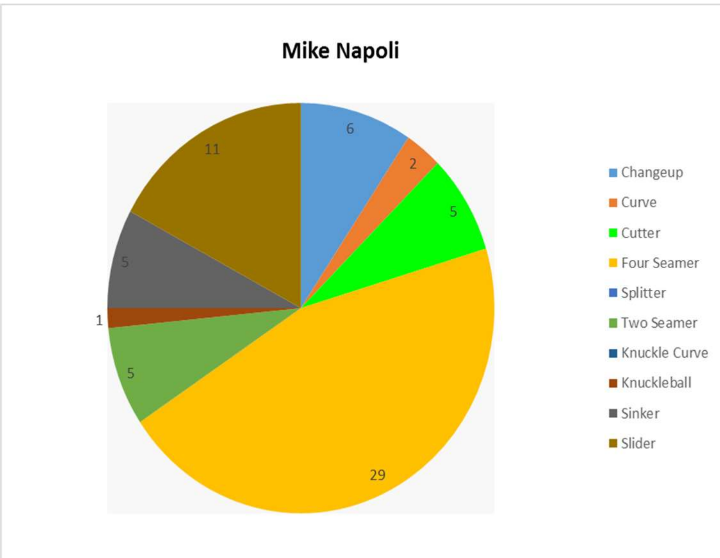



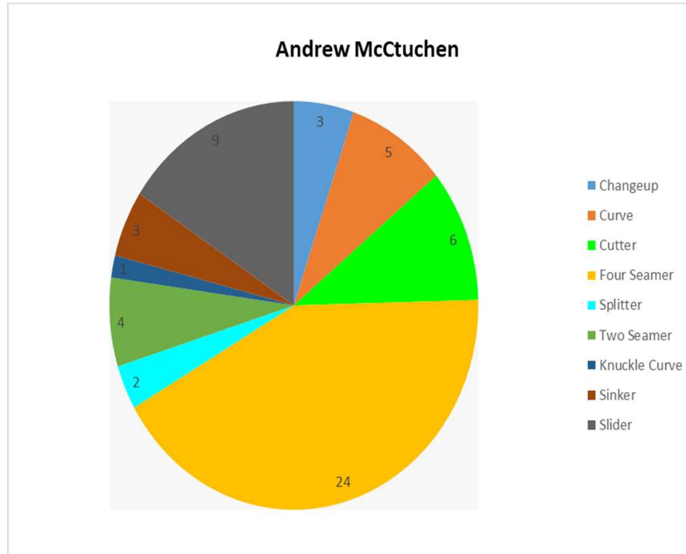



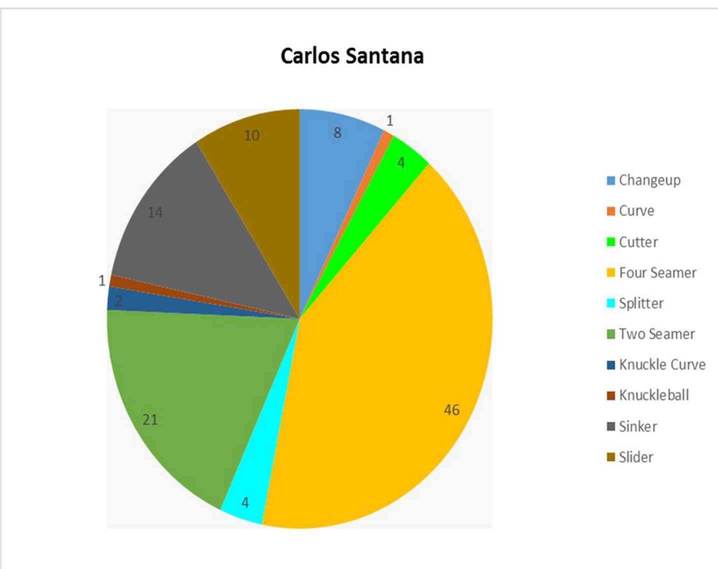



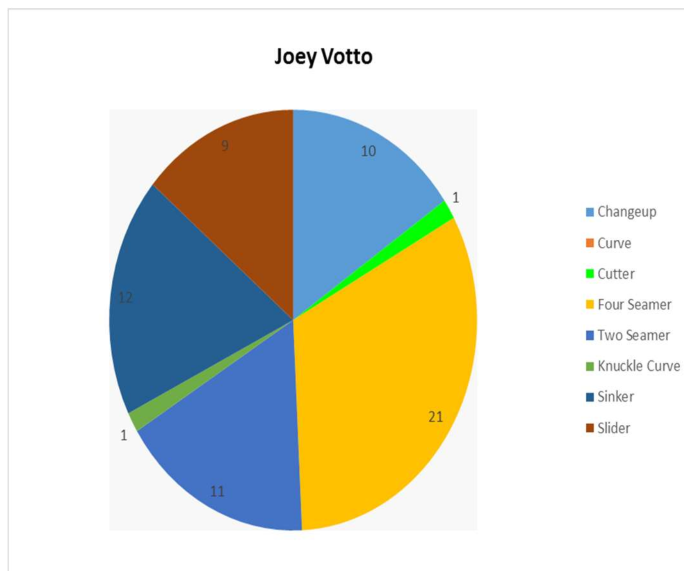



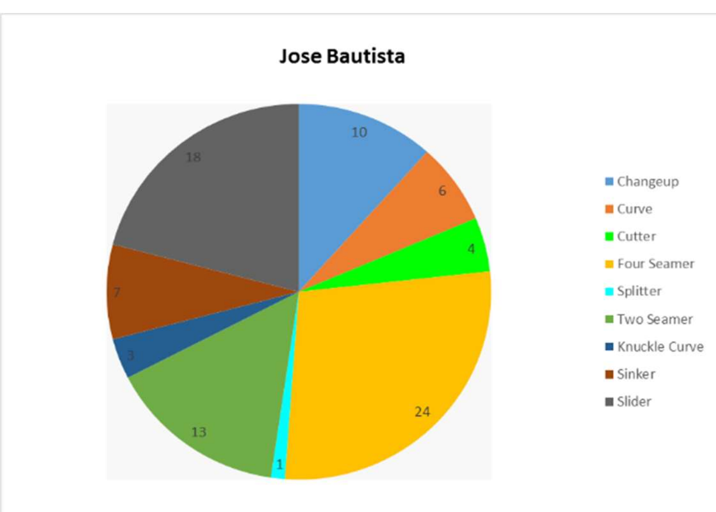
Now that I knew their strengths, I wanted to investigate the weaknesses of these star batters and determine if there were any pitches that resulted in strikeouts or in play outs that they struggled to have success with. That dashboard can be found below:

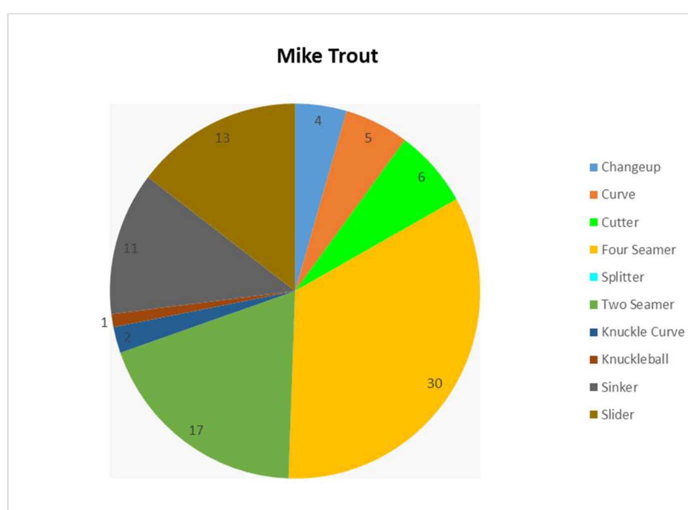


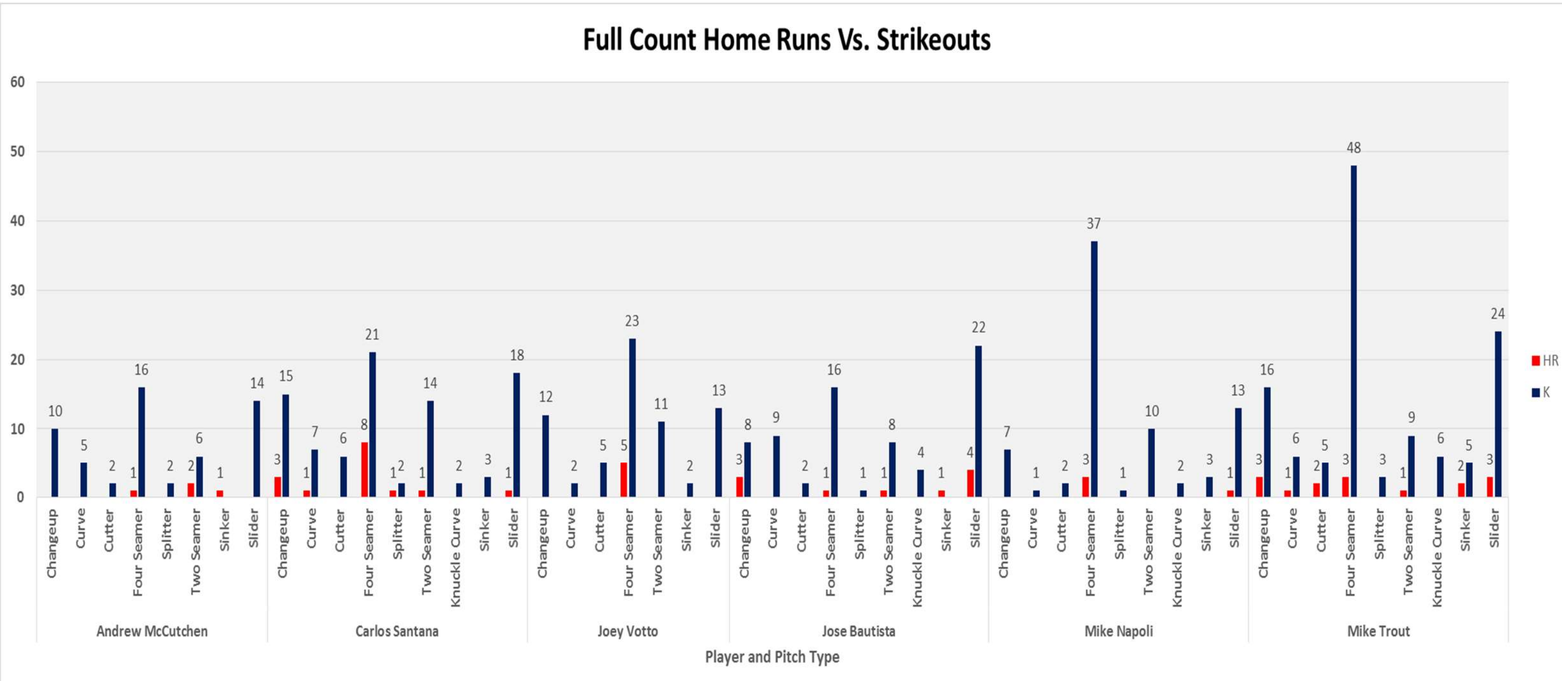



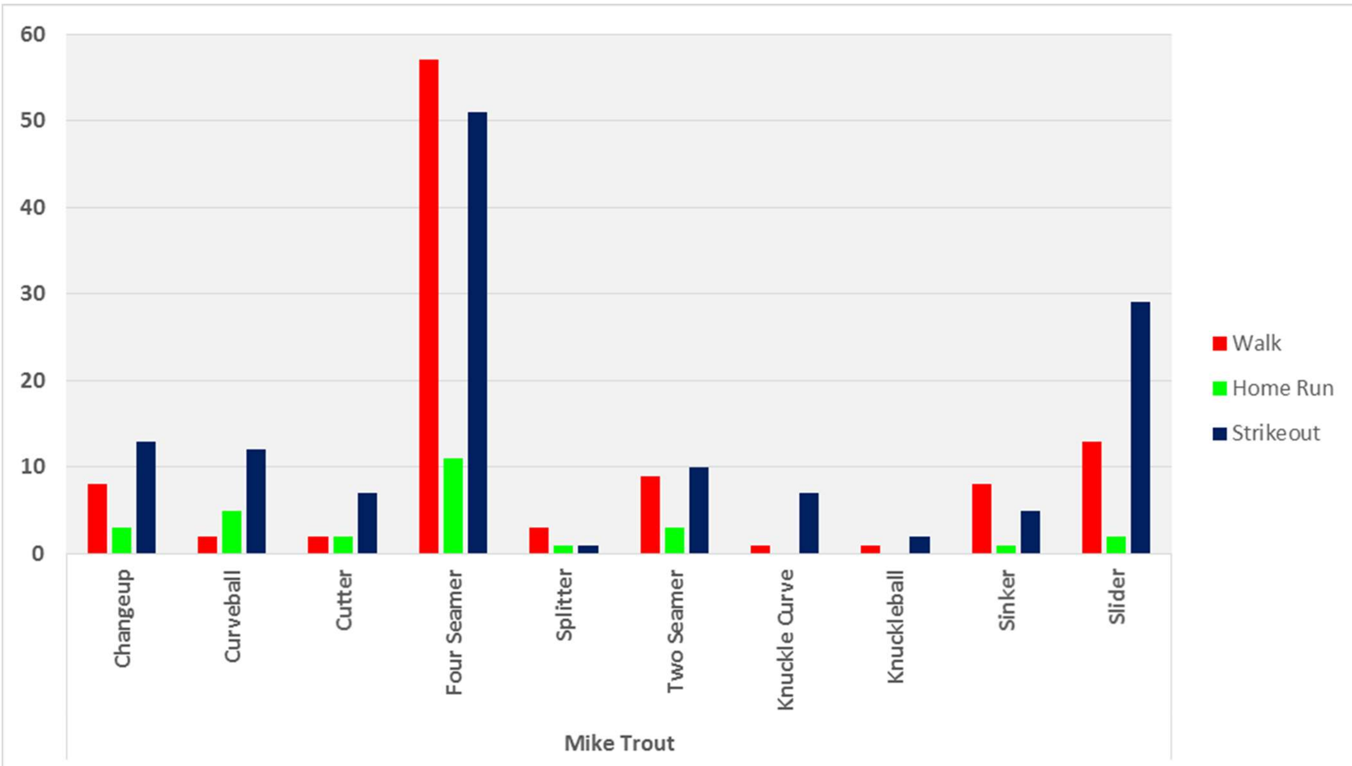
This dashboard didn't exactly tell me anything that I wanted to know. So, I utilized a different approach to see if I could produce a data set that could provide a valuable comparison of pitches for these players. I simply filtered down to home runs and strikeouts to see if that data told us anything. That chart is provided below:




Now we're talking. I really liked this chart. I could conclude a few things very quickly. On a 3-2 count, here are some quick observations:

1. I would look at throwing Andrew McCutchen, Joey Votto, and Mike Napoli a **changeup or a curve**.
2. I would throw a **four seamer as fast as possible** at Mike Trout. With more time, I'd like to see if location, spin rate, or a certain velocity would make the pitch even more effective against him. I'd also consider a knuckle curve.
3. I would **not throw a Four Seamer** at Carlos Santana. He's a tough player to pitch to, but if I had to pitch to him in this scenario, I would throw some type of junk like a sinker, knuckle curve, or a cutter.
4. I would **avoid a changeup** to Jose Bautista. I would throw a curve or a cutter at him.

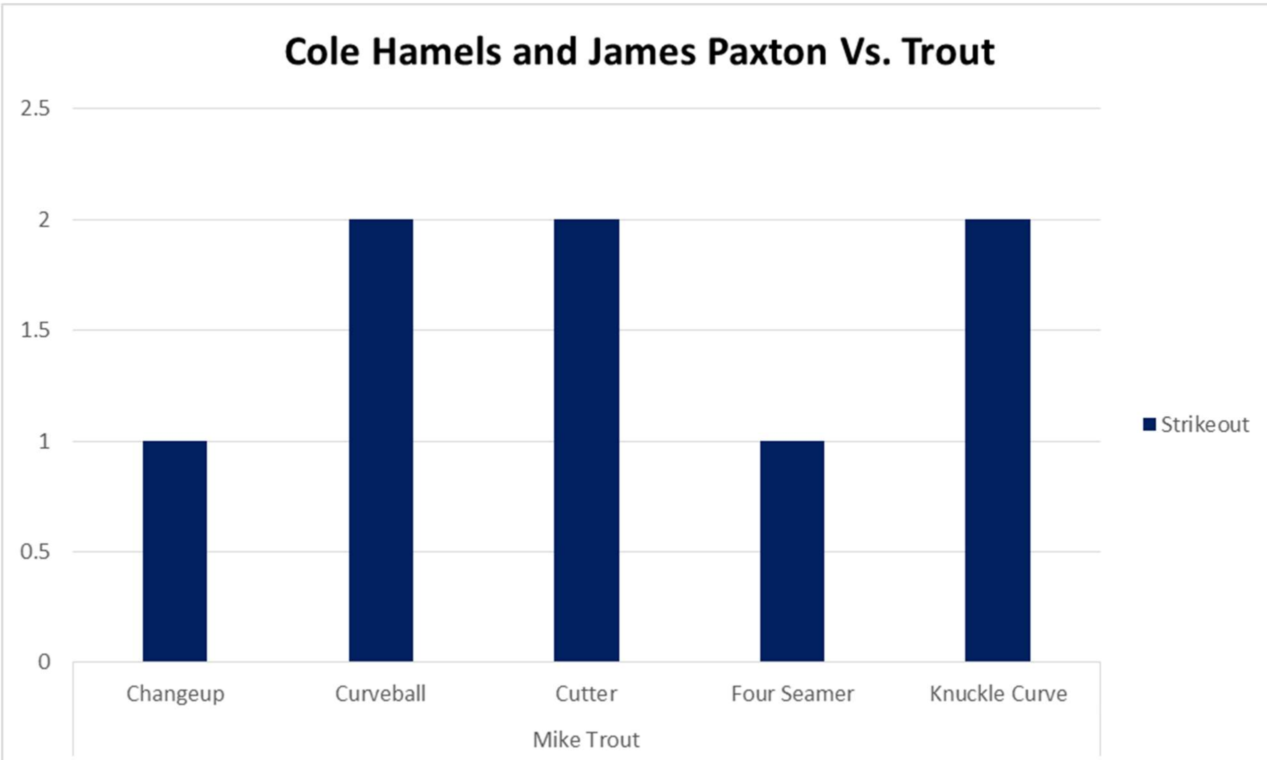
With more time, I would go back and see if the same type of trends are reflective in these players' bats when the count is different. If similar data was portrayed for all pitches, we may have a breakthrough. With the time I have, I wanted to see if the 2016 season data matched for Mike Trout. I filtered out just 2016 regular season data for **Mike Trout** to analyze pitches that yielded walks, home runs, or strikeouts. That can be found below:




A couple of things jumped out immediately:

1. Trout has **better success against the four seamer** overall than he has had over the past three seasons when the count was full.
2. In both charts, he has hit **no home runs against a knuckleball or knuckle curve**. In clutch pitching situations, these would be the types of pitches I'd be looking to throw at him. Save them for those crucial times when a strikeout is desperately needed.
3. The **slider** is an interesting pitch for Trout. He has been struck out a lot with it, hit relatively few home runs, but seems to have a good eye for the pitch as he has drawn a number of walks with it. This would be a pitch I'd like to study further to see if there is a certain type of slider that is more successful against him than others.

Then, finally, I wanted to look and see which pitchers were the most successful against a guy like Mike Trout, and what type of pitches they threw so that I would know what film to focus on and see if I could determine any tendencies that I could use to my advantage. A chart of two successful pitchers is shown below:




Cole Hamels and James Paxton were two of the more successful pitchers against Mike Trout, giving up no walks or homeruns to him. These particular pitches and pitchers would be good study if one had to face Mike Trout, especially in a full count scenario.

**CONCLUSION:**

Looking at pitch choice in a full count situation is very subjective depending on a number of factors. The research above shows that certain criteria such as velocity and what inning it is may show a small trend for the identified players, but have a relatively small impact. The type of pitch that these top batters have success/struggles against is wildly variable and has to be looked at on a specific player and specific outcome basis to have applicable meaning, much like the Mike Trout charts above. Utilizing player specific information in in certain scenarios has value and can be analyzed further to develop pitch plans against certain players. From the data above, it is very clear that you have to study individuals' historic performance against specific criteria to develop an approach that has the chance to be successful. Pitches like a four seam fastball have such mixed results that I wouldn't be able to say whether they are good or bad pitches in a full count scenario. With more time, I would want to look at a few other factors such as location of pitches for various types and analyze whether or not there was an advantage to be had.

# SECTION 3: ARE UMPIRES REALLY BIASED? AND WHAT SHOULD WE DO ABOUT IT?

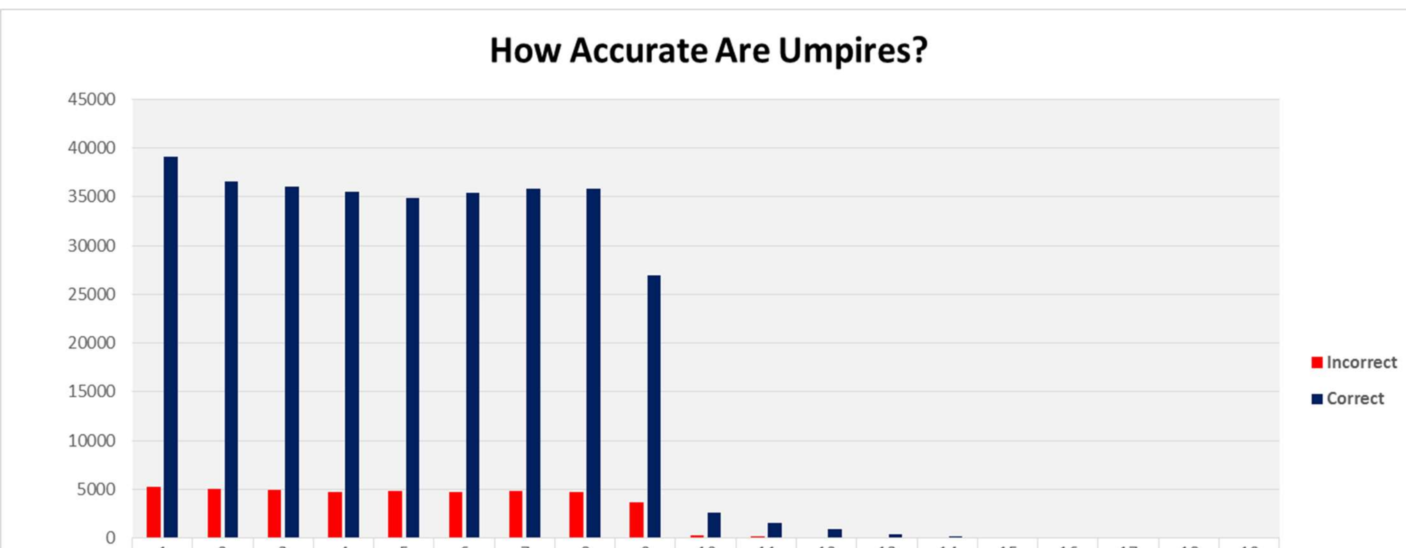
In all sports, people like to blame the outcomes of events on the officials. It is very easy to point to one particular play, pitch, missed call, etc. and credit that event as the sole reason for a win/loss. While there is proof through behavioral economics that officials tend to show biases at various moments of a sporting event, there is very little data to explain why.

My goal in this section was to let the data help us understand the true accuracy of umpires, and assess whether we should alter our behavior (pitches) based on the results. Because I wanted to delve into some pretty specific detail on this topic, I only used one data set (2016 regular season) with a lot of in depth analysis about umpire performance.

The way I analyzed the umpires in the data set was relatively simple. I established a strike zone given the data points: **px, pz, szt, szb**. I then wrote a couple of "OR" statements in excel to determine whether the call was a strike or ball, and whether it was correct or not. I then filtered the data to **only include strikes accepted looking and called balls.**

With some basic calculations of the data set I determined whether the umpire made the correct or incorrect call. The results produced good news. Per the calculations, the collective group of umpires made the **correct** call **88%** of the time. The results and chart can be seen below.

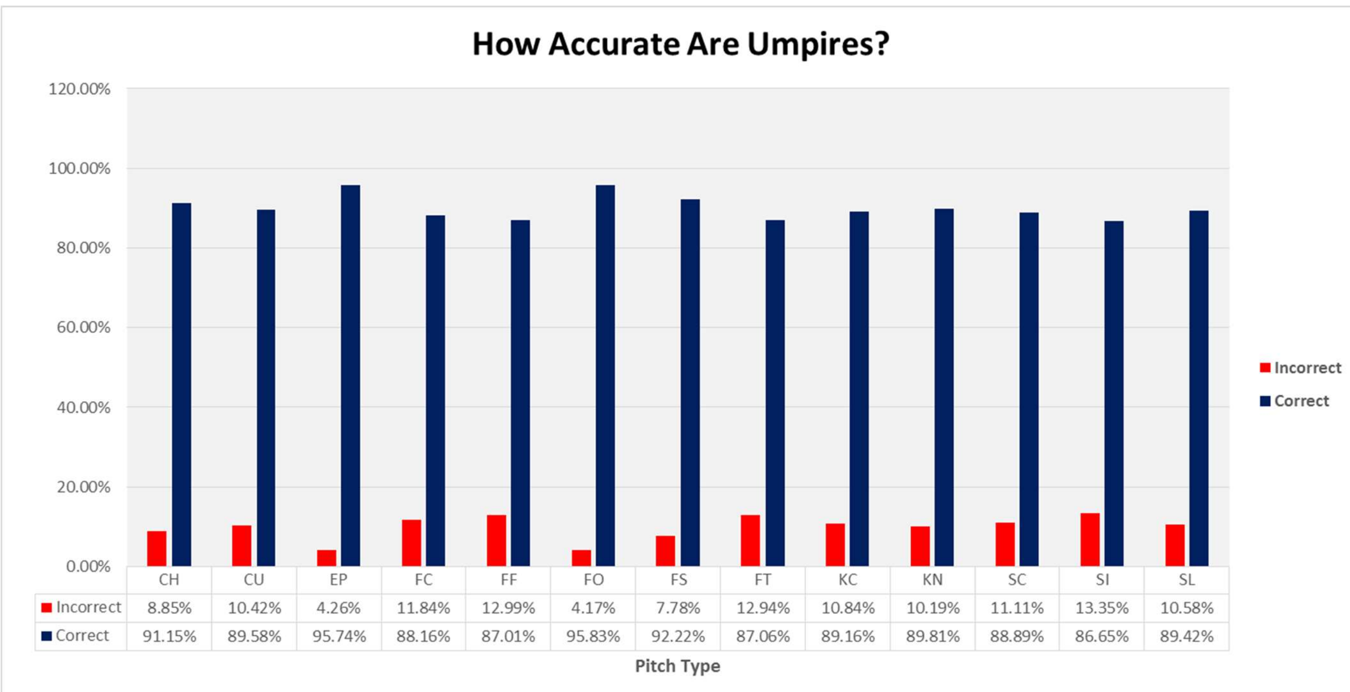
For an in-depth look and interactive chart for the data and analysis for this section, please reference the Excel file in the depository titled "2016 Umpire.




| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Incorrect | 5292 | 5013 | 4932 | 4719 | 4851 | 4738 | 4880 | 4710 | 3649 | 328 | 189 | 94 | 50 | 23 | 25 | 6 | 7 | 6 | 2 |
| Correct | 39059 | 36545 | 36027 | 35503 | 34864 | 35383 | 35849 | 35805 | 26903 | 2620 | 1550 | 896 | 438 | 167 | 108 | 51 | 55 | 30 | 10 |

Inning

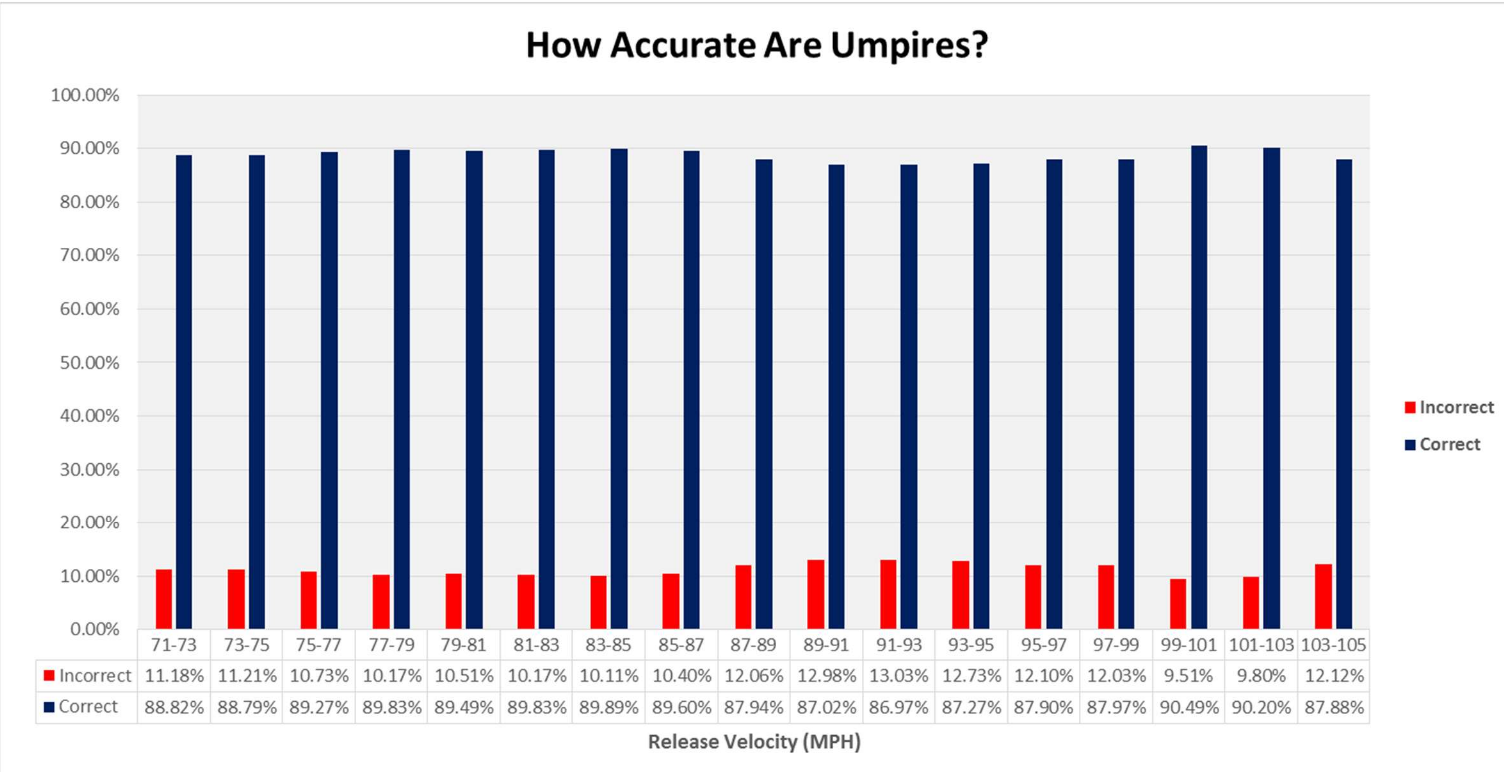
So yes, good news on the umpire accuracy front. However, there were still some incorrect calls and these are the ones I wanted to drill down on as a pitcher. I wanted to understand how I could use these missed calls and potential biases to my advantage. My next steps were to understand what particular attributes of a pitch were most commonly called incorrectly so that I could avoid those pitches at the opportune times. There was not a lot of variance between Right and Left-handed pitchers and the percentage of incorrect calls. Both Right and Left-handed pitchers were victims of incorrect calls 11-12% of the time. The chart that breaks down the pitch type vs. calls is below:



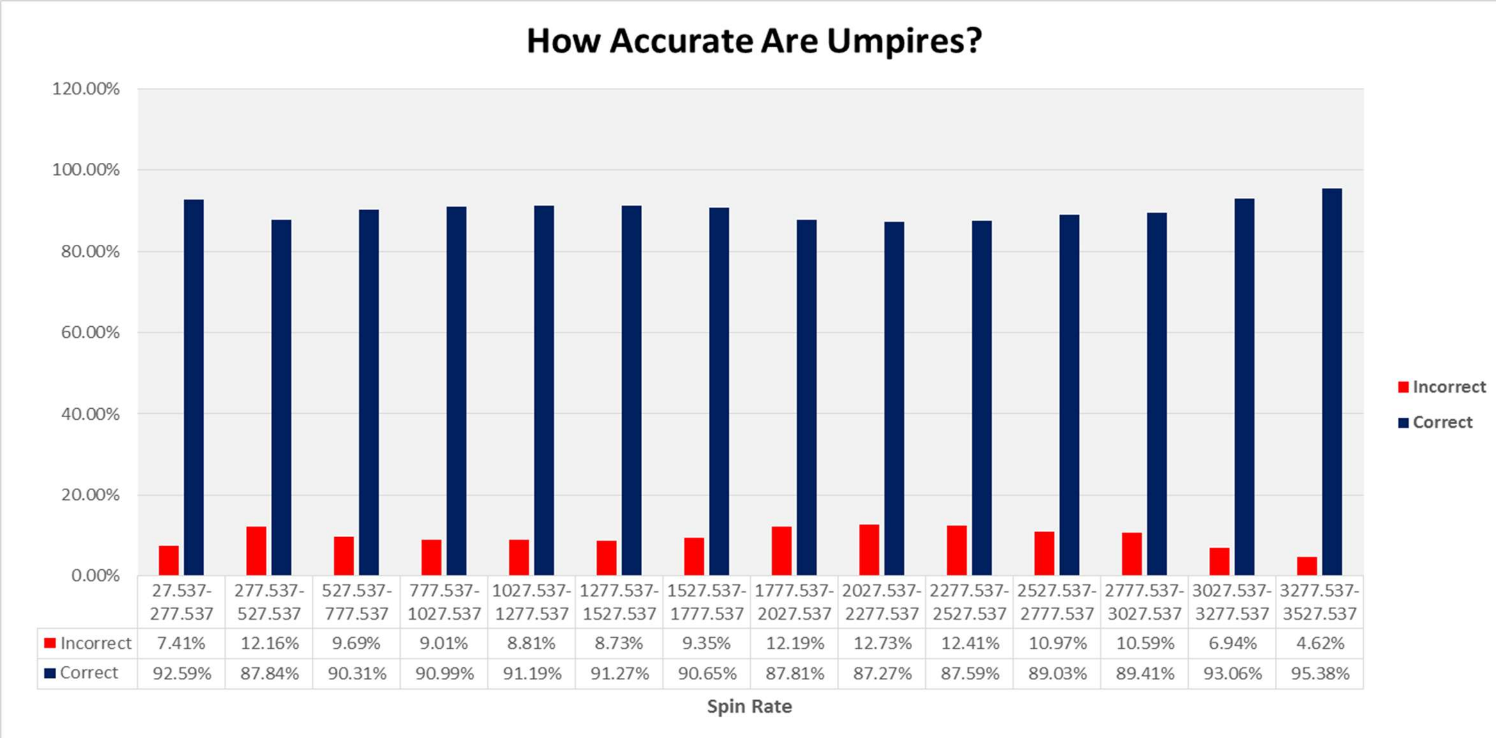

With this information, one could determine to alter their pitch count of certain pitches with the assumption that umpires would continue to follow their trends of incorrect vs. correct calls. Particularly when combined with risky vs. risk averse pitches, this data could generate a different pitch plan than what would traditionally be done. Perhaps one would want to add more splitters and two seamers to their game plan knowing that the umpire will call them more consistent compared to other pitches.

But I wanted to take this analysis a step further. What other criteria about umpire tendencies could provide a small edge in developing pitching plans?

First, I wanted to see if the release velocity showed any type of trend with umpire performance. I filtered out pitches that were less than 70 MPH as the data was minimal at that threshold. The correlating chart can be seen below:




This chart didn't provide any real correlation other than the fact that the umpires were only accurate over **90%** of the time on pitches in the **99-103 MPH range**. The calls were pretty consistent across the board. Next, I wanted to determine if the **spin rate** applied to the ball had any impact on umpire accuracy. That chart can be found on the next page:

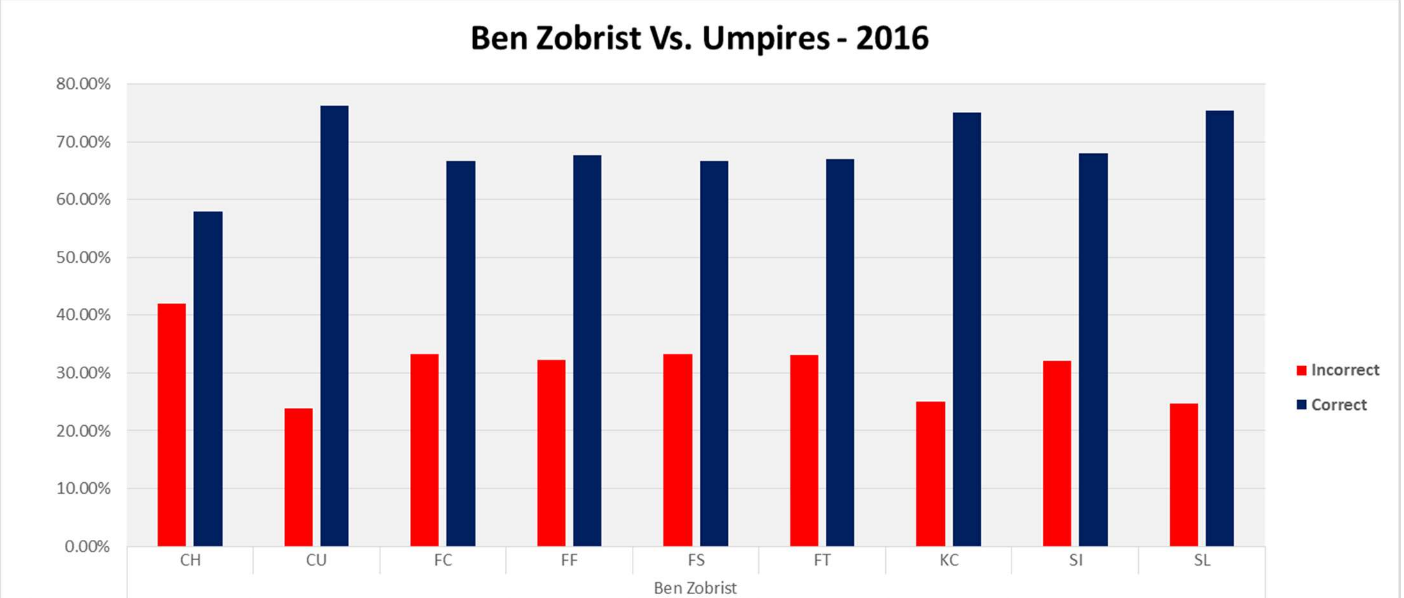



The only interesting observation here is that the incorrect percentage was above 12% for spin rates from around **1775 – 2500**. This range contains a lot of the data in the sample.

Then, I decided to get a little creative. If I was a pitcher, I would be most interested in the types of pitches that would yield strikes while looking that were questionable or incorrect calls. Filtering the data to only include strikes while looking, I ran a few different cases to determine which type of pitches would yield the best advantage with which umpire. First, I wanted to analyze which batters were most likely to suffer a **strike while looking** call that was incorrect. This would hopefully produce a list of umpire biases against certain players. The top 10 results can be seen below.

| Batters | Incorrect |
|---|---|
| Ben Zobrist | 212 |
| Curtis Granderson | 211 |
| Nick Markakis | 182 |
| Brian Dozier | 181 |
| Paul Goldschmidt | 181 |
| Jayson Werth | 174 |
| Jason Kipnis | 173 |
| Christian Yelich | 172 |
| Brett Gardner | 164 |
| Carlos Santana | 161 |
| Matt Carpenter | 161 |
| Gregory Polanco | 159 |

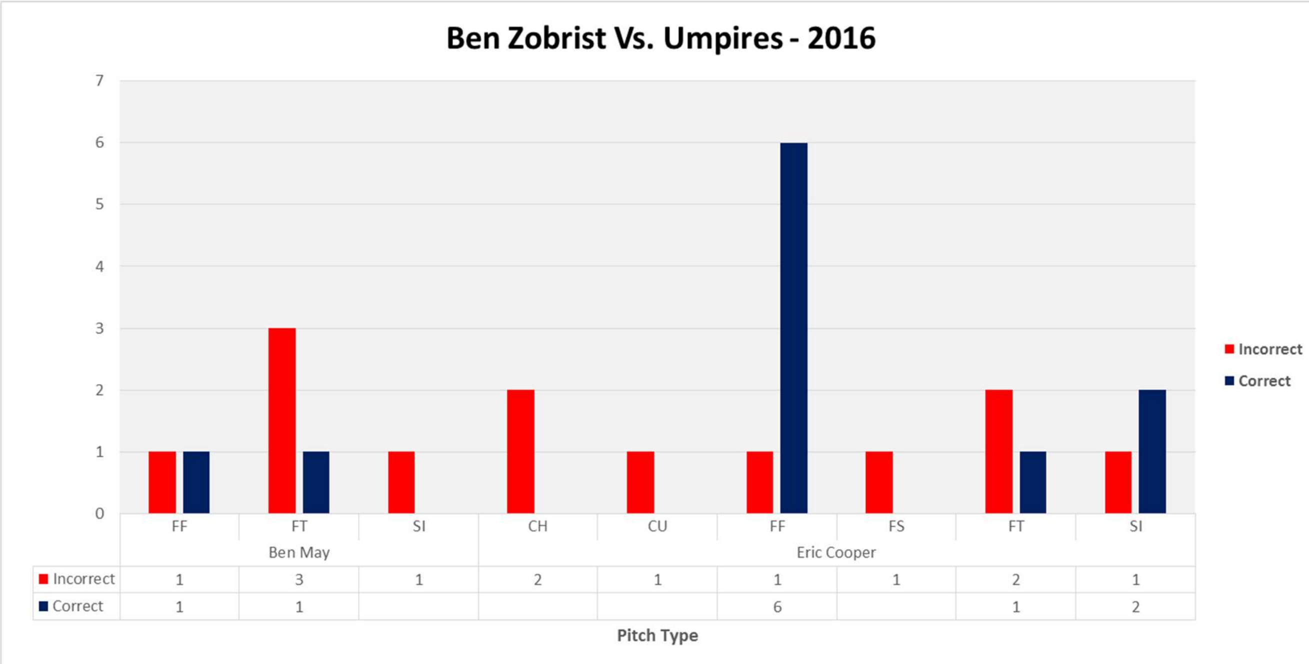
Having this information about potential batters I would be facing, the next step was to drill down to specifics about each batter to see where potential advantages would exists based on a variety of criteria. Take the top victim, **Ben Zobrist**. It's not a huge surprise to see Zobrist at the top of the chart as he has one of the lowest swing-rates in the MLB. Therefore, he gives the umpire more chances to make a mistake. If I wanted to analyze the various pitches that umpires **call strikes while looking** incorrectly, that chart would look like this:




| | CH | CU | FC | FF | FS | FT | KC | SI | SL |
|---|---|---|---|---|---|---|---|---|---|
| Incorrect | 42.03% | 23.81% | 33.33% | 32.27% | 33.33% | 33.05% | 25.00% | 32.00% | 24.64% |
| Correct | 57.97% | 76.19% | 66.67% | 67.73% | 66.67% | 66.95% | 75.00% | 68.00% | 75.36% |

The biggest thing that jumps out from this chart is that Ben Zobrist is a victim of a strike being called incorrectly almost half the time when the pitch is a **changeup**. The changeup is the pitch with the most bias from umpires and would be worth studying and researching whether there is an opportunity to exploit against him.

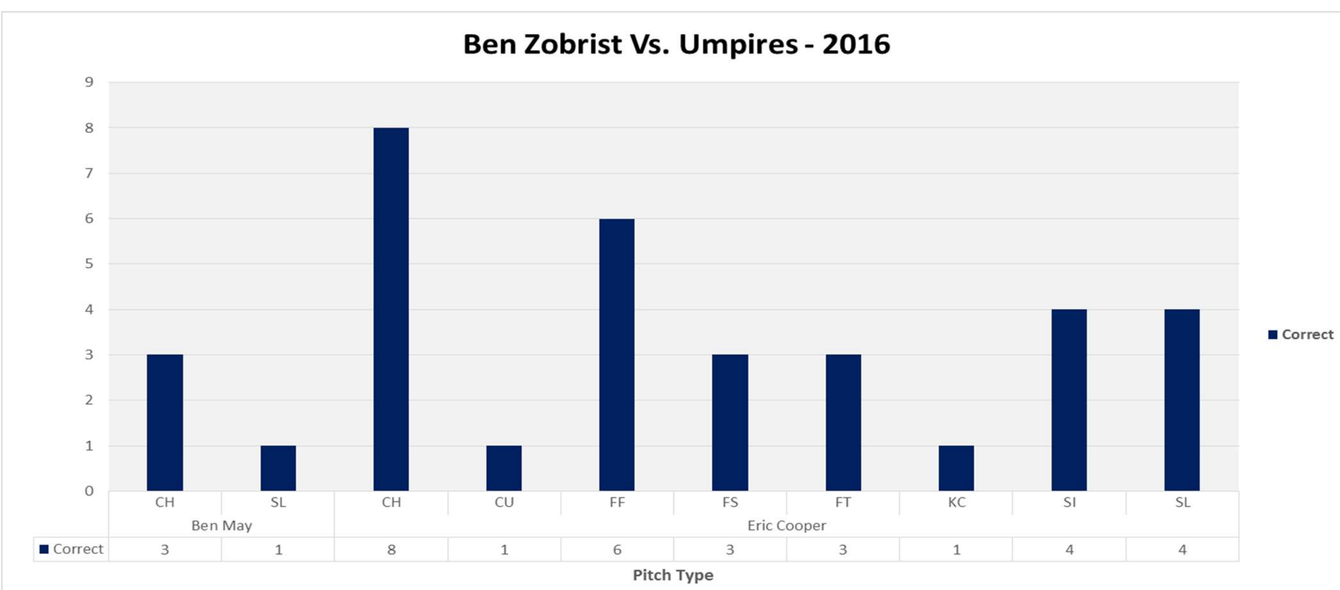
But what if we could take this data and look at specific umpires and their biases against Ben Zobrist? For instance, I stayed with the data set for **Strikes Accepted Looking** and filtered out certain umpires. The following chart shows the data from umpires **Ben May** and **Eric Cooper**:



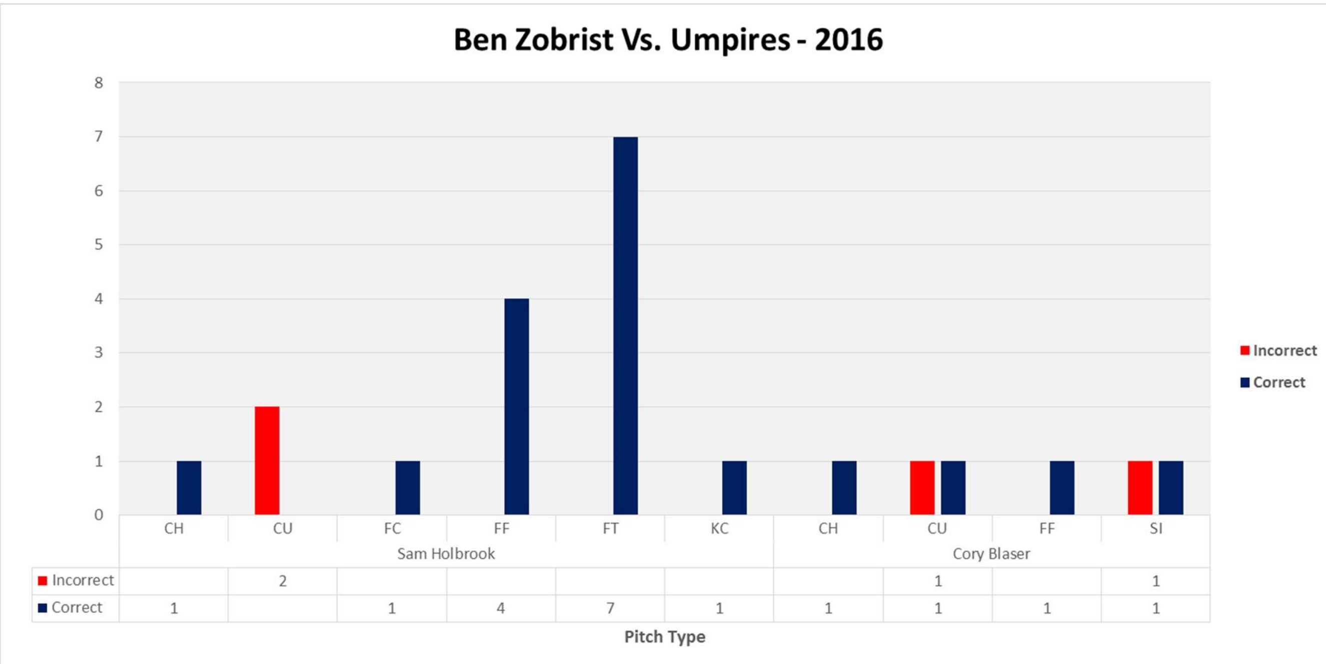

Immediately, a few things jump out:

1. Ben on Ben crime! Ben May must have a strong bias against Ben Zobrist as he makes more incorrect calls on a variety of pitches than he does correct calls, leaving Zobrist the victim of many strikes he shouldn't have been charged with.
2. If I was trying to slide a questionable one past Zobrist with Eric Cooper calling the game, I would NOT throw a four seamer. Cooper gets those right with strong consistency. Instead, I would opt for the **two seamer or splitter**.

So if we can analyze strikes called incorrectly down to this level, couldn't we perform the same analysis for **balls called incorrectly**? This would expose where I, as a pitcher, would be at risk from the umpire. Leveraging from our same umpires as above, Ben May and Eric Cooper, here is the chart of their called balls:



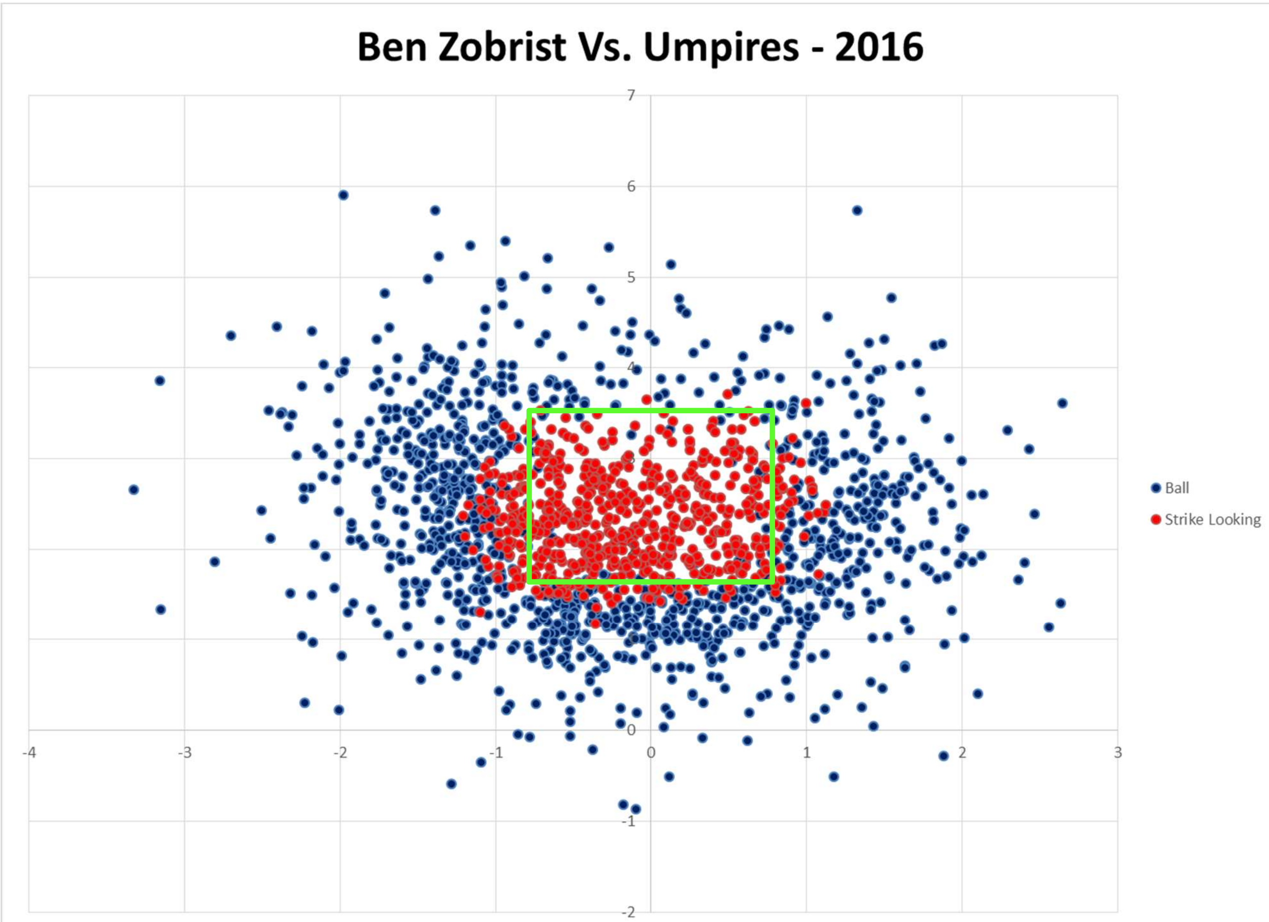
Wow! The story kept getting better for a pitcher. Ben May and Eric Cooper did not have a single incorrect ball call against Ben Zobrist. They were 100% accurate on balls, certainly to a pitcher's advantage. Balls are typically easier to call on the whole, but not every umpire was perfect against Ben Zobrist. See the chart for Sam Holbrook and Cory Blaser below:




A few observations from this chart:

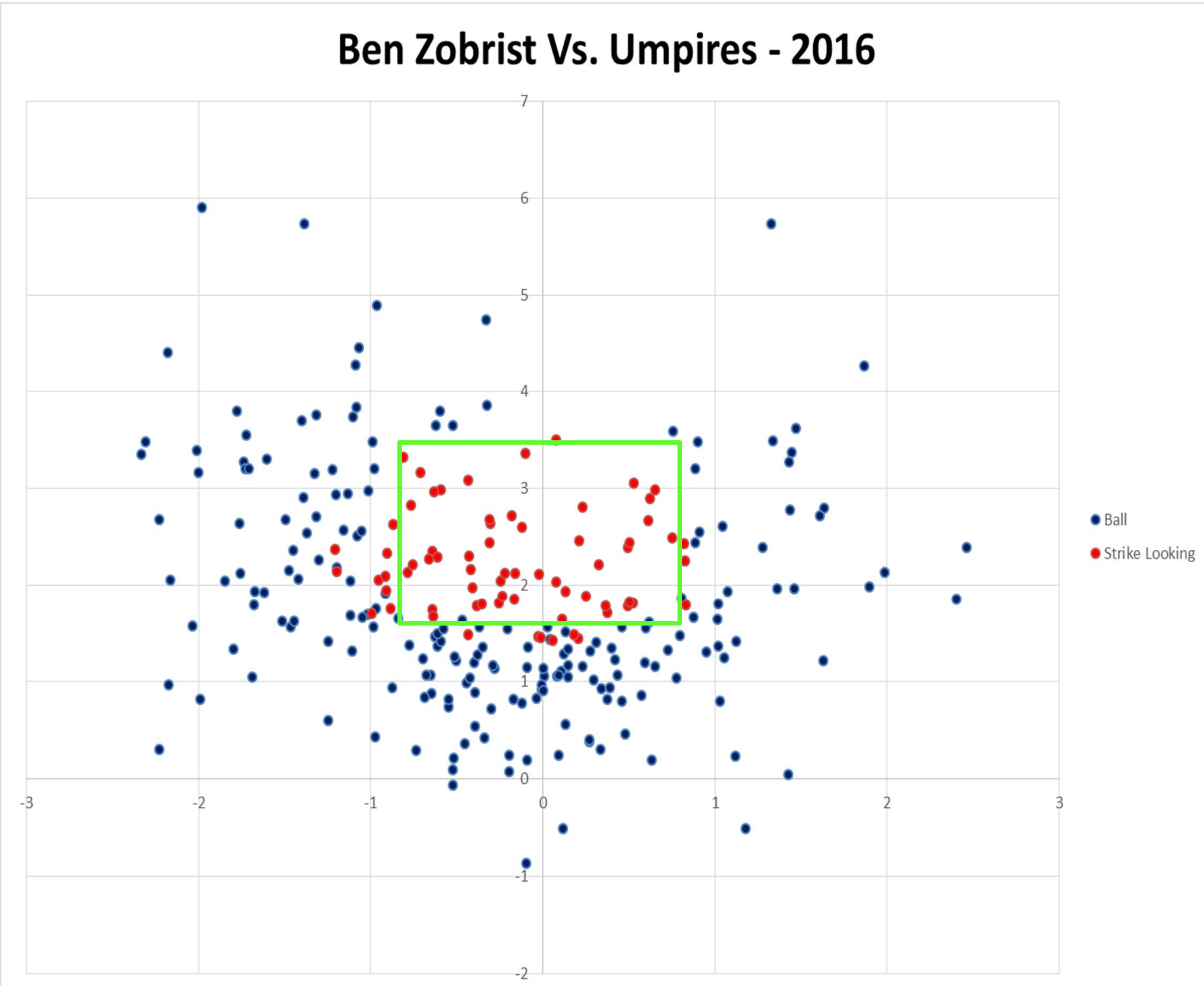
1. I would want to stay away from a **questionable curve** if Sam Holbrook was calling the pitches against Zobrist.
2. **No changeups** were called incorrectly. That's actually good news as we know Zobrist is victim to the changeups from the charts above.

And finally, I wanted to chart Ben Zobrist's pitch locations where an umpire could affect the outcome of the call.  I charted all of Ben's pitches, both balls and strikes called looking, with an average of his strike zone from the **szt and szb data**.  That chart can be found below:




Looking at this data, it confirmed that umpires are very good at calling balls when they are earned.  I found it interesting that there are more strikes called that are outside of the established strike zone on the **left side of the grid than the right side**.  Since Zobrist is a switch-hitter, we would have to review these specific pitches and see whether they were typically called as inside or outside pitches.  There are also a cluster of pitches that are just outside the **bottom of the strike zone all the way across**.  These would provide a target area in the course of facing Zobrist as umpires show a small bias in calling strikes outside of the zone.

**Bonus:** since Zobrist falls victim to umpires the most on **changeups**, I wanted to look at his pitch location grid for changeups only. That chart can be found below:




Very analogous to the chart above it. **The left side of the grid and just below the strike zone are good areas to target against Zobrist** as umpires tend to call the most strikes looking incorrectly in those two areas. The pitches in this chart would guide my pitch plan to Ben Zobrist, who is one of the toughest batters you'll ever face.

**<u>Conclusion</u>:**

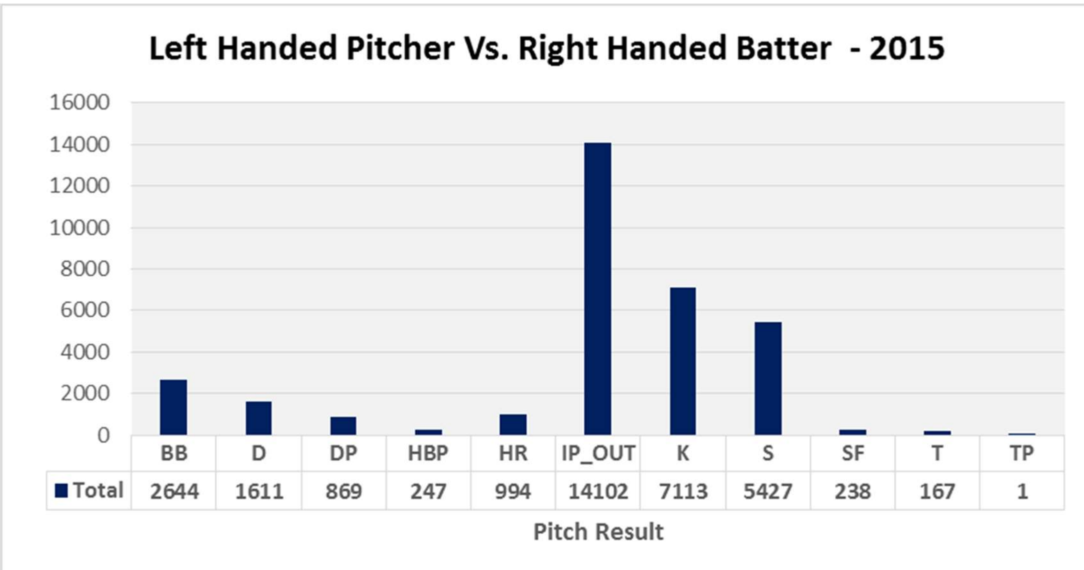
Obviously, with more time and data, one could look at a variety of criteria in relation to umpires to determine if there are any biases against certain pitchers, batters, pitch locations, etc. The data will be somewhat limited and tests would need to be trialed to determine if the umpires would continue to follow their biases and if advantages could be realized. According to the data provided, **umpires get the calls right the overwhelming majority of the time**. Finding the few cases where they don't, and utilizing them to your advantage, is the skill in data analytics that could mean a questionable strikeout vs. a questionable walk.

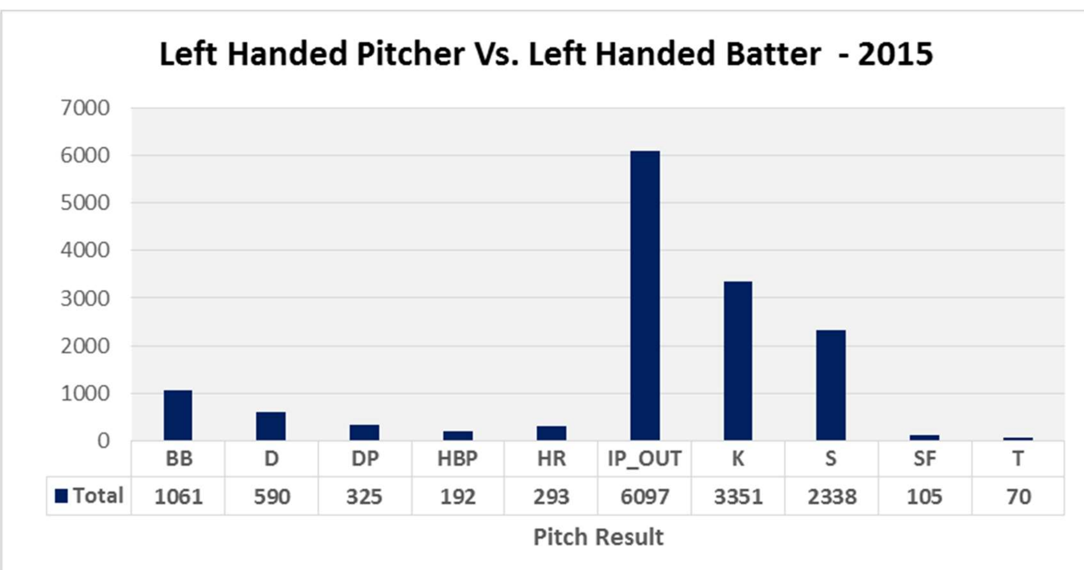
I have heard many people say that it is better for the batter to see more pitches and let the pitcher make the mistake. If the pitcher knows the tendencies and the biases of both the batter and the umpire calling the game, seeing more pitches could mean for a risky endeavor on the part of the batter. Question is, are they willing to roll the dice?

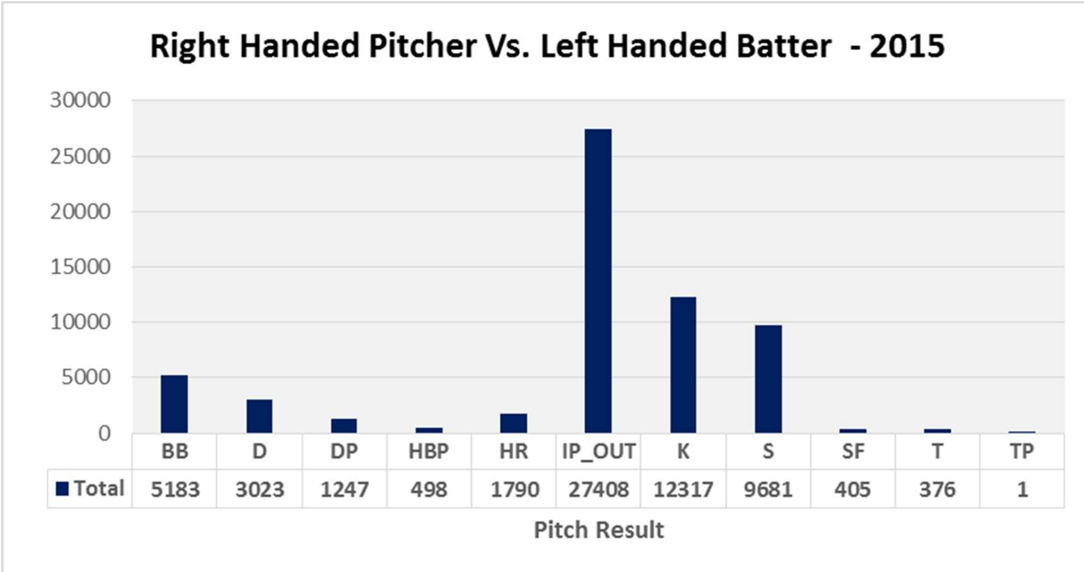
## SECTION 4: DOES IT MATTER IF A PITCHER AND/OR BATTER ARE RIGHT OR LEFT HANDED?

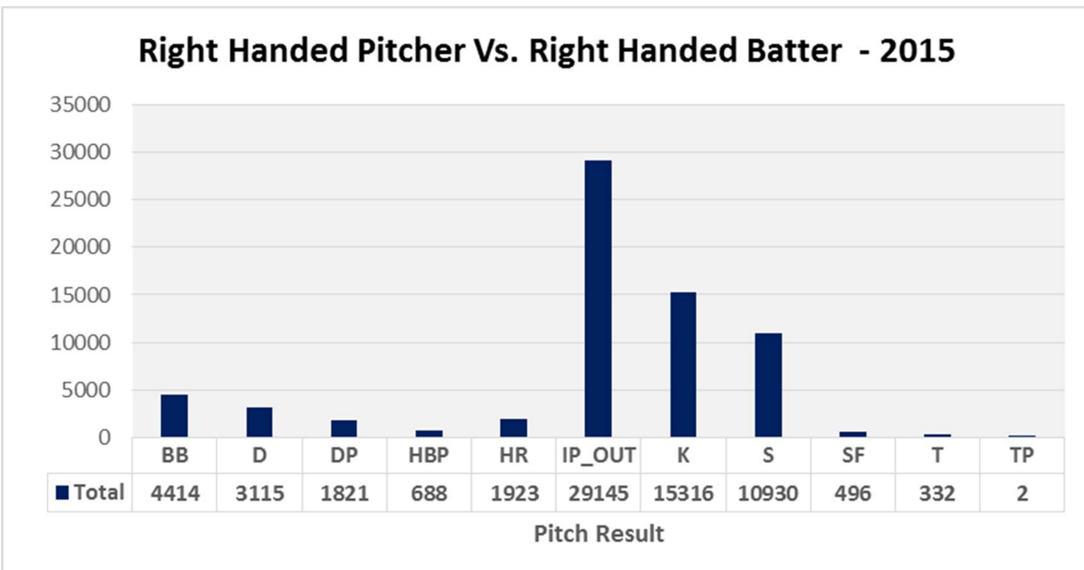
This was the shortest section of all. I had some extra time and wanted to give a cursory look at whether the myth of having opposite handed pitchers and batters really provided an advantage one way or the other. For an interactive excel file please reference the file entitled “Pitch Hand.” I kept this analysis brief and only utilized the **2015** season as a random sample.

Below, you will see a group of charts that reflect all cases of a certain dominant hand batter vs. a dominant pitcher. These were generated from some very quick slicer sorting in excel:






Conclusions can be found on the following page:

**Conclusions:**

1. The ratios of pitches was:
    a. Right-handed batter – Left-handed pitcher = 33,413
    b. Right-handed batter – Right-handed pitcher = 68,182
    c. Left-handed batter – Right-handed pitcher = 61,929
    d. Left-handed batter – Left-handed pitcher = 14,422

    So, when a right-handed batter was at the plate, the ratio of pitches was **2.04 to 1**. When a left-handed batter was at the plate, the ratio of pitches was **4.29 to 1**.

2. In analyzing **double plays**, when a batter was right handed the ratio of double plays was **2.09 to 1**, almost EXACTLY equal to the ratio of pitches thrown. When a batter was left-handed the ratio of double plays was **3.84**. This may lead to a small advantage in facing a left-handed pitcher against a left-handed batter, but it appears that there is not a strong correlation to pitcher/batter stance and double plays.

3. In both cases, more **walks** were given up when the pitcher and batter had **opposite stances**. The margin isn't large in either case and could be attributed to the frequency of pitches.