The Pitching Evolution:

The Effect of Pitching Velocity on
Strikeout Rates in Major League Baseball

Authored By: Clay Hamilton and Max DeWit

Econ 57, Professor Zemel

December 14, 2018

Introduction:

A 100-mph fastball takes about 375-400 milliseconds to cross home plate. The normal blink of an eye takes 300-400 milliseconds.[1] In baseball, pitching is the first line of defense, and the better pitching a team has, the better their chance of winning. The most recent trend in developing pitchers is increasing the overall velocity of pitches, specifically the fastball. One way coaches are helping develop pitchers' fastballs is with new technology that helps identify mechanical issues in a pitcher's form. When a mechanical issue is identified, coaches can then train pitchers out of the inherent bad habits to help them gain more control or velocity in their pitches. As a result of this new technology and mode of teaching, we are currently witnessing the fastest ever individual pitchers in the game of baseball. Currently, Aroldis Chapman, a pitcher for the New York Yankees, pitches the fastest fastball ever thrown in game, clocking in at 105.1 MPH.[2] The second fastest ever recorded in game pitch was in the 2018 season, thrown by a rookie Jordan Hicks, who also threw 105 MPH.[3] With these two in mind, coaches have become so asphyxiated with increasing pitchers' fastball velocity because they, along with many players, believe that the increase in pitching velocity is correlated with the simultaneous decrease in batting percentages and on-base percentages. Coaches are changing the way their batters approach hitting in an attempt to deal with the new velocity pitchers are capable of. In fact, dominant aspects of hitting, such as bunting, are being left in the dust as the new strategy for hitters manifests itself. In our study ask if there is a significant relationship between faster fastballs and more strikeouts? And if we do find a relationship between faster fastballs and more strikeouts, how different is the relationship between the 2012 season and the 2018 season?

We expect to find that as fastball velocity goes up, the strikeout rate will increase significantly. We also think that there will be an increase in the slope comparing fastball velocity and strikeouts in the 2018 season, compared to the 2012 season.

---

[1] https://projects.seattletimes.com/2017/mariners-preview/science/

[2] https://www.mlb.com/news/aroldis-chapman-throws-a-pitch-1051-mph/c-190404910

[3] Cardinals' Jordan Hicks ties MLB record by throwing 105 mph -- twicehttps://www.usatoday.com/story/sports/mlb/.../05/...mlb...throwing.../627391002/

Data:

This "data revolution" has led to the MLB making available an enormous quantity of statistics about all facets of the game, for teams' data scientists and hardcore fans alike to pore over. The data for this experiment comes from the advanced pitching statistics of fangraphs.com, which aggregates data from the MLB and other private sources[4]. The site offers more than 50 different statistical variables for pitching alone; this project uses several of the most fundamental ones.

The main variables which the project considers are average fastball velocity (FBv) and strikeouts (shortened as K's) per 9 innings. Their definitions are as follows:

Fastball Velocity (FBv): The average velocity (in mph) of every fastball a pitcher threw over the course of a season.

Strikeouts per 9 Innings (K/9): The number of strikeouts a pitcher threw over the course of a season, divided by every 9 innings (a full game worth of pitching). This number represents the average number of batters a pitcher would have struck out over a 9 inning span.

$$\text{Calculation: } \frac{\text{Total \# of Strikeouts}}{\text{\# of Innings Pitched}} * 9$$

These measurements were selected because they are standard in the field for measuring pitching speed and strikeout rate. The baseball community uses fastball velocity specifically because every pitcher has a fastball in their arsenal, whereas types of off-speed pitches (pitches that aren't fastballs) vary from pitcher to pitcher. For example, one pitcher might use a fastball, curveball, and a changeup, while another might use a fastball, slider, and a cutter. For this reason, the fastball velocity of a pitcher is the most accurate and applicable measure of generally how hard they throw. K/9 is the measurement used for strikeout rate because it gives an easily understandable number for how many batters a pitcher strikes out. In 2018, for example, pitchers struck out between approximately 5 and 15 batters every 9 innings of pitching. Analyzed together, these variables demonstrate how a pitcher's pitch speed affects the amount of batters they strike out.

[4] https://www.fangraphs.com/library/pitching/complete-list-pitching/ (note: see Sources section for more)

The sample periods contain the entire MLB seasons of 2012 and 2018. We picked 2012 because it was the year that most MLB teams had fully embraced the use of advanced statistical analysis. Additionally, the movie “Moneyball” had just been released, chronicling the A’s early adoption of statistics and the advantage it gave them. All told, this was the first year “sabermetrics” and the “data revolution” really began to make their mark in baseball, and thus should capture the inception of this new approach to baseball.

For our sample, we used every qualified pitcher from the 2012 and 2018 seasons. In order to be qualified, a pitcher had to pitch 50 innings or more over the course of the season. We chose this cutoff because we wanted to avoid having the results be skewed by pitchers who had only pitched a few innings (for example, position players pitching for an inning, or pitchers who were only called up from the minors for a game or two). We chose 50 innings specifically because that is the lower bound for MLB relief pitchers, and thus any pitcher who pitched at least 50 innings was a regular pitcher for a team. This gave us $n=327$ observations (eligible pitchers) for the 2012 season and $n=337$ for the 2018 season.

There are five further metrics used in the calculations for this paper. Two of them, fastball percentage and innings pitched, are used as control variables. The other three, walks per 9 innings, strikeouts - walks per 9 innings, and home runs per 9 innings are dependent variables which provide further insight about the effects of the increased pitching velocity. These metrics are defined as:

Fastball Percentage (FB%): The percentage of times a pitcher threw a fastball vs. another type of pitch (curveball, slider, etc.).

Calculation: $\frac{\text{Total \# of Fastballs}}{\text{\# of Innings Pitched}} * 100$

Innings Pitched (IP): The number of total innings a pitcher completed in a season.

Walks per 9 Innings (BB/9): The number of walks a pitcher allowed over the course of a season, divided by every 9 innings. This number represents the average number of batters a pitcher would have walked over a 9 inning span.

Calculation: $\frac{\text{Total \# of Walks}}{\text{\# of Innings Pitched}} * 9$

Strikeouts per 9 - Walks per 9 (K/9 - BB/9): Per 9 innings, how many more strikeouts than walks a pitcher would have on average.

Calculation: $(\frac{\text{Total \# of Strikeouts}}{\text{\# of Innings Pitched}} * 9) - (\frac{\text{Total \# of Walks}}{\text{\# of Innings Pitched}} * 9)$

Home Runs per 9 Innings: The number of home runs a pitcher allowed over the course of a season, divided by every 9 innings. This number represents the average number of home runs a pitcher would have given up over a 9 inning span.

Calculation: $\frac{\text{Total \# of Home Runs}}{\text{\# of Innings Pitched}} * 9$

Table 1: Summary Statistics

		2012	2018
	Observations (n)	327	337
FBv	Mean	91.460	92.871
	Standard Deviation	2.6276	2.6106
FB%	Mean	0.5835	0.5546
	Standard Deviation	0.1246	0.1262
IP	Mean	111.33	103.04
	Standard Deviation	56.398	48.224
K/9	Mean	7.7520	8.7036
	Standard Deviation	2.0269	2.0885
BB/9	Mean	3.0423	3.1841
	Standard Deviation	1.0452	1.0127
K/9 - BB/9	Mean	4.7096	5.5195
	Standard Deviation	2.1123	2.2016
HR/9	Mean	0.9937	1.0883
	Standard Deviation	0.3781	0.4297

Chart 1: Histograms of K/9 per player, by year

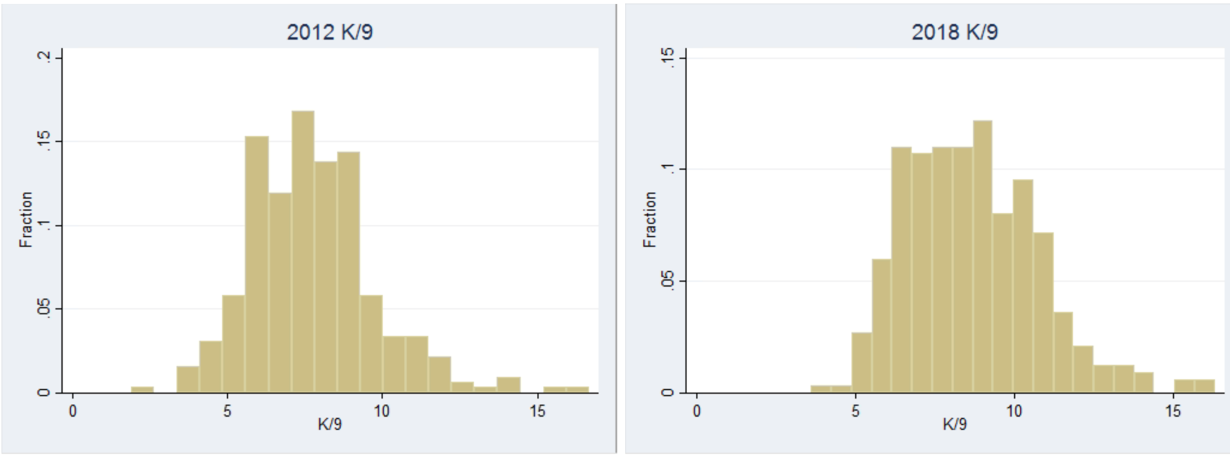


Chart 2: Histograms of FBv per player, by year

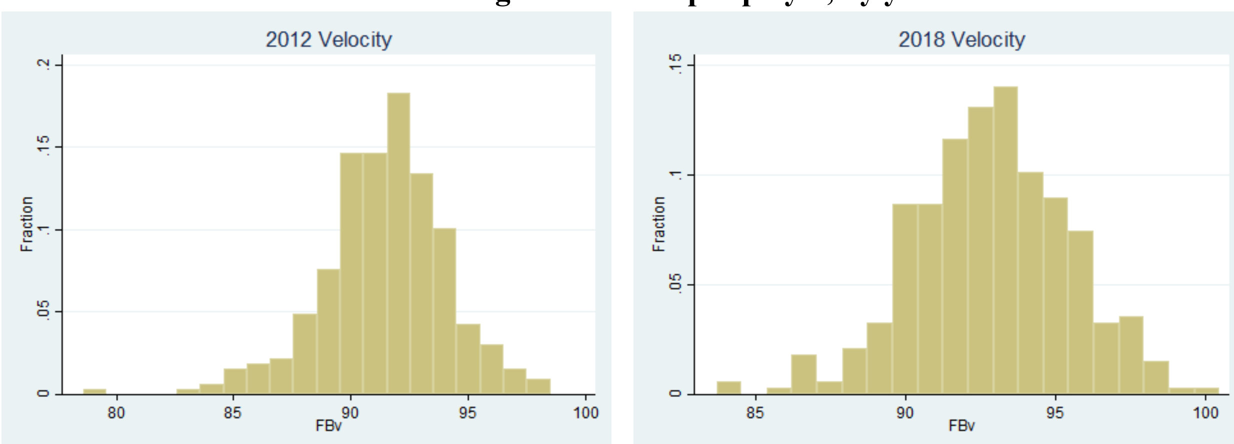
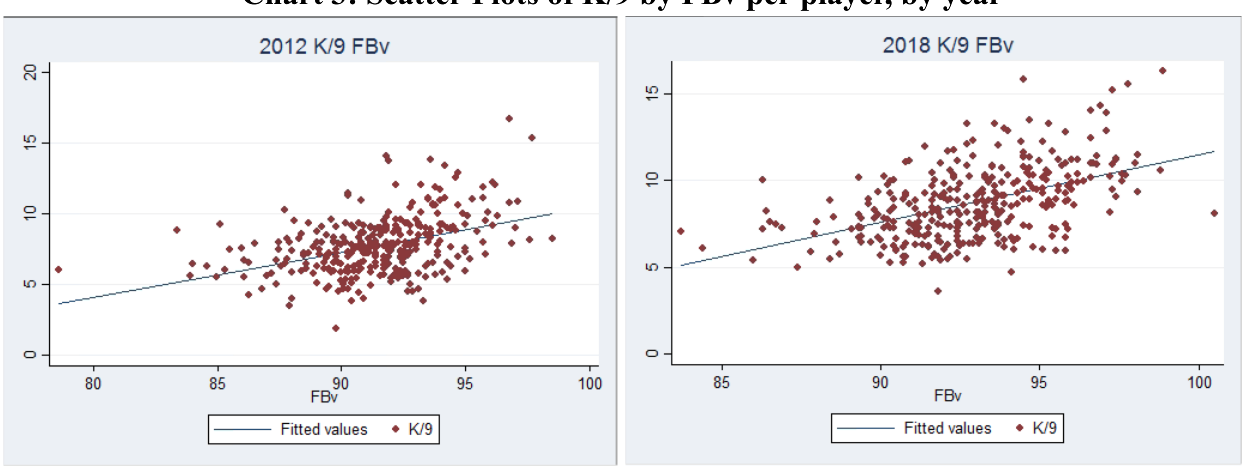


Chart 3: Scatter Plots of K/9 by FBv per player, by year



Methodology:

These summary statistics seem to show an increase in pitching velocity and strikeout rate, but in order to determine for sure the best method is to use a multivariable regression. Using multivariable regression is appropriate because it will determine: a) whether there exists a statistically significant difference in pitching speed and strikeout rate between the years 2012 and 2018, b) whether there exists a statistically significant relationship between pitching speed and strikeout rate for the years 2012 and 2018, and c) whether the relationship between pitching speed and strikeout rate has changed between the two years. The regressions control for percentage of fastballs thrown and for number of total innings pitched by a pitcher. Controlling for percentage of fastballs thrown ensures that the results are not skewed by the effect of pitchers changing up their pitch types (for example, if pitchers use other pitches more often to get strikeouts). Controlling for number of total innings pitched negates the possible effect that pitching fewer innings over the course of a season might have on a pitcher's average velocity and strikeout rate (for example, if a pitcher pitches slower and strikes out fewer batters because of seasonal fatigue). These two control variables will allow the effect of pitching speed to be isolated as best as possible.

In order to determine the difference in effect between the years of 2012 and 2018, this regression also uses a dummy variable and an interaction term. The dummy variable, Ind18, is an indicator which gives a 1 for data from 2018 and 0 for data from 2012. The interaction term, intFBV, is equal to FBv * Ind18, and will allow the data to show whether the relationship (the slope of the line) between velocity and strikeouts has changed from 2012 to 2018.

Additionally, there are several dependent variables which are calculated to determine some of the additional effects of the increase in fastball velocity. These are the walk rates per 9 innings, the difference between K/9 and BB/9, and the home runs allowed per 9 innings. The Stata abbreviations for every variable are:

K9: K/9 (strikeouts/9 innings)
FBv: FBv (fastball velocity)
FBp: FB% (fastball percentage)
IP: IP (# of innings pitched)
BB9: BB/9 (walks/9 innings)

K9BB9: K/9 - BB/9 (strikeout rate - walk rate)
HR9: HR/9 (home runs/9 innings)
Ind18: Dummy variable for 2018 (1 = 2018, 0 = 2012)
intFBV: Interaction term for FBV * Ind18

Each regression was designed to test the model that β_{FBV} is > 0 with a p-value of < .05, which would indicate that fastball velocity is positively correlated with strikeout rate. Figures 1, 3, 4, and 5 used the following equation:

$$y_i = \alpha + \beta_{1FBVi} + \beta_{2FBPi} + \beta_{3IPi} + \beta_4 M_{18i} + \varepsilon_i \qquad i = 1, 2 \ldots n$$

The independent variables in this case are FBv, FBp, IP, and the dummy variable Ind18.

Figure 2, which included the interaction term intFBv, used the following equation:

$$y_i = \alpha + \beta_{1FBVi} + \beta_{2FBPi} + \beta_{3IPi} + \beta_4 M_{18i} + \beta_{5FBVi} M_{18i} + \varepsilon_i \qquad i = 1, 2 \ldots n$$

These yielded the following hypotheses:

Null Hypothesis: There is no relationship between fastball velocity and strikeout rate.
$\beta_{FBV} = 0$

Alternative Hypothesis: There is a relationship between fastball velocity and strikeout rate.
$\beta_{FBV} \neq 0$

Results:

Figure 1: Multivariable Regression for K/9

```
. reg K9 FBv FBp IP Ind18

      Source |       SS           df       MS      Number of obs   =       663
-------------+----------------------------------   F(4, 658)       =     57.17
       Model |  760.985397         4  190.246349   Prob > F        =    0.0000
    Residual |  2189.76567       658  3.32791135   R-squared       =    0.2579
-------------+----------------------------------   Adj R-squared   =    0.2534
       Total |  2950.75106       662  4.45732789   Root MSE        =    1.8243

------------------------------------------------------------------------------
          K9 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         FBv |   .3659846   .0290307    12.61   0.000     .3089806    .4229886
         FBp |  -1.326731   .6074495    -2.18   0.029    -2.519504   -.1339581
          IP |  -.0037671   .0013828    -2.72   0.007    -.0064823   -.0010519
       Ind18 |   .3656531   .1503918     2.43   0.015     .0703474    .6609589
       _cons |  -24.52755   2.587874    -9.48   0.000    -29.60904   -19.44606
------------------------------------------------------------------------------
```

Figure 2: Multivariable Regression for K/9 (with interaction term)

```
. reg K9 FBv FBp IP Ind18 intFBV

      Source |       SS           df       MS      Number of obs   =       663
-------------+----------------------------------   F(5, 657)       =     46.01
       Model |  765.273227         5  153.054645   Prob > F        =    0.0000
    Residual |  2185.47784       657  3.32645028   R-squared       =    0.2593
-------------+----------------------------------   Adj R-squared   =    0.2537
       Total |  2950.75106       662  4.45732789   Root MSE        =    1.8239

------------------------------------------------------------------------------
          K9 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         FBv |    .333741   .0406074     8.22   0.000     .2540051    .4134769
         FBp |  -1.256881   .6104244    -2.06   0.040    -2.455499   -.0582628
          IP |  -.0037542   .0013825    -2.72   0.007    -.0064689   -.0010395
       Ind18 |  -5.328152   5.017293    -1.06   0.289    -15.18001     4.52371
      intFBV |   .0618211   .0544513     1.14   0.257    -.0450985    .1687406
       _cons |  -21.62072   3.639956    -5.94   0.000    -28.76807   -14.47338
------------------------------------------------------------------------------
```

Figure 3: Multivariable Regression for BB/9

. reg BB9 FBv FBp IP Ind18

Source	SS	df	MS
Model	75.9548503	4	18.9887126
Residual	627.131925	658	.953088031
Total	703.086775	662	1.06206461

Number of obs	=	663
F(4, 658)	=	19.92
Prob > F	=	0.0000
R-squared	=	0.1080
Adj R-squared	=	0.1026
Root MSE	=	.97626

BB9	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
FBv	.0737415	.015536	4.75	0.000	.0432354	.1042475
FBp	-.0154907	.3250804	-0.05	0.962	-.6538107	.6228293
IP	-.0046007	.00074	-6.22	0.000	-.0060537	-.0031476
Ind18	-.0008884	.0804831	-0.01	0.991	-.1589231	.1571464
_cons	-3.180802	1.384917	-2.30	0.022	-5.900192	-.4614128

Figure 4: Multivariable Regression for K/9 - BB/9

. reg K9BB9 FBv FBp IP Ind18

Source	SS	df	MS
Model	452.275204	4	113.068801
Residual	2734.88824	658	4.1563651
Total	3187.16344	662	4.81444628

Number of obs	=	663
F(4, 658)	=	27.20
Prob > F	=	0.0000
R-squared	=	0.1419
Adj R-squared	=	0.1367
Root MSE	=	2.0387

K9BB9	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
FBv	.2922431	.0324435	9.01	0.000	.2285378	.3559485
FBp	-1.31124	.6788614	-1.93	0.054	-2.644236	.0217553
IP	.0008336	.0015453	0.54	0.590	-.0022008	.003868
Ind18	.3665415	.168072	2.18	0.030	.0365195	.6965635
_cons	-21.34675	2.892105	-7.38	0.000	-27.02561	-15.66788

Figure 5: Multivariable Regression for HR/9

```
. reg HR9 FBv FBp IP Ind18
```

Source	SS	df	MS			
Model	11.4093758	4	2.85234394	Number of obs	=	663
Residual	98.5603983	658	.149787839	F(4, 658)	=	19.04
				Prob > F	=	0.0000
Total	109.969774	662	.166117483	R-squared	=	0.1038
				Adj R-squared	=	0.0983
				Root MSE	=	.38702

HR9	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
FBv	-.0463206	.006159	-7.52	0.000	-.0584143	-.034227
FBp	.0784091	.1288732	0.61	0.543	-.1746432	.3314613
IP	.0003622	.0002934	1.23	0.217	-.0002139	.0009382
Ind18	.1651712	.0319063	5.18	0.000	.1025208	.2278217
_cons	5.144225	.5490293	9.37	0.000	4.066164	6.222285

These regressions lead to several important takeaways. First, for K/9, the p-values for FBv, FBp, and IP were < .05, which means the null hypothesis is rejected. More specifically, FBv has a positive correlation with K/9, with the result showing that each 1 mph increase in fastball velocity, will result in .3659846 more strikeouts per nine innings; this supports an alternative hypothesis of **H1:** $\beta_{FBV} > 0$. FBp and IP are negatively correlated with K/9, meaning higher fastball percentages and more innings pitched lead to fewer strikeouts.

The dummy variable representing the year also had a p-value of < .05, meaning the intercept for 2018 FBv/K/9 is provably greater than that of 2012.

Including the interaction term intFBv changes the results, giving p-values of > .05 for the dummy variable and the interaction term. These results fail to reject the null hypothesis, and therefore we cannot show that the relationship between fastball velocity and strikeout rates has changed from 2012 to 2018. Put simply, the results show that pitching speed and strikeouts are correlated, and the number of both has increased from 2012 to 2018, but it can't be shown relationship between the two changed between the two years.

There are a few supplementary outcomes which the other regressions provide. There is a significant correlation between FBv and BB/9 (p-value < .05), suggesting that this increase in pitching speed leads to more walks as well as more strikeouts. However, looking at K/9-BB/9, the regression shows that there is a positive relationship between that and FBv, again with p-value < .05. This means that although walk rates increase with the increase in fastball velocity, strikeout rates increase more, and therefore the difference becomes more pronounced at higher velocity.

Finally, as for HR/9, the regression also gives a p-value of < .05, and therefore there is a significant relationship between HR/9 and FBv. Interestingly, this is a negative relationship: an increase in fastball velocity leads to a lower rate of home runs allowed.

Discussion:

In the last 10-20 years, baseball has used much more technology and generated many new statistics to evaluate players. All of this data is compiled into larger datasets that are generally accessible to anyone who is interested in the statistics of the game. Moreover, most of these statistics are super easy to record with these new technological advancements. The accessibility and precision of these data sets allows us to find data that accurately measures our variables and does so holistically. Because of this, our sample can include data for every pitcher in the MLB who pitched over 50 innings in a season; this makes the data representative, as the sample size is the population. Our threshold for pitchers is inclusive enough to be representative of the overall population of pitchers in the league because it is a majority of the pitchers to start with. Because the sample we used (almost every pitcher in the league) is basically equal to the population in this case, there’s not really any selection bias that could have affected the results. The only possible bias is that using pitchers with 50 or more innings most likely drove up the result for number of strikeouts and fastball velocity; however, this choice likely made the sample mean and standard deviation more accurate, as all of the pitchers with just a few innings and just a few strikeouts (who were excluded by our criteria) would drag the results down inaccurately.

Based on the results of our data, we can infer a correlation between our independent variable, velocity of pitch, and our dependent variable, total number of strikeouts. We evaluated this relationship in two different ways: 1) testing the relationship between pitch velocity and strikeouts in the 2018 season and 2) testing if the slope of the velocity and strikeouts line was steeper in the 2018 season or the 2012 season.

The regressions illuminate several important findings. For K/9, the p-values for FBv, FBp, and IP were < .05, which means the null hypothesis is rejected. We observed that FBv has a positive correlation with K/9, with the result showing that each 1 mph increase in fastball velocity, will result in .3659846 more strikeouts per nine innings. This finding makes sense because the fast the ball is thrown the less time a batter has to adjust their swing, making it ultimately harder to hit. Additionally, we find that higher fastball percentages and more innings pitched lead to fewer strikeouts, which also makes sense because pitchers arms get tired the more they pitch. Generally, the more fatigue the pitchers experience, the less accurate and effective their pitching gets.

The next part of our results show that pitching speed and strikeouts are correlated, and the number of both has increased from 2012 to 2018, but that the relationship between the two has not significantly changed between the two years. One possible explanation for this is that batters are beginning to adjust to the increase in velocity of pitches and changing their approach to how they hit. A strikeout is the worst possible outcome for a batter because it requires not defense from the otherside, so hitters must emphasize the importance of not striking out relative to their abilities with faster pitching.

We also find a significant correlation between FBv and BB/9 (p-value < .05), showing that as pitching velocity increases the likelihood of walks also increases. This intuitively makes sense, as in order to achieve this increased velocity pitchers might have to sacrifice some control over where their pitches go. However, even though walk rates have increased, strikeout rates have increased more.

Ultimately, our data gave us very interesting yet logical outputs that baseball minds have been grappling with as they try to give their team the slight edge. While it is easier said than done, pitching coaches are focusing on how to increase pitching velocity while also increasing pitching accuracy. Once this evolution in the teaching process is a little more developed, batters will face even larger challenges than they already do.

One of the confounding variables we weren't able to fully include in the regression we ran is the relationship that batters have with pitchers. Throughout the history of baseball the relationship of the batters mindset to the pitchers abilities has had ebbs and flows. The pitcher will change an aspect of their strategy as a means of gaining the edge on the batter and the batter will respond by changing part of their strategy to account for the pitcher's new development. This process continues over and over again, allowing for the game to develop as a whole. Currently, the pitchers have the upper hand as demonstrated by the increase in strikeouts. However, batters in the last couple seasons are changing their approaches to combat this new reality. The main new philosophy of hitting begins with accepting that pitchers are better at striking out batters than they used to be, meaning, it is harder to get guys on base in scoring position.

Additionally, it is harder to string together a series of hits that lead to a run being scored because on base percentages have gone down. So, what does one do if it is difficult to get someone on base in scoring position and even more difficult to get a hit that brings that runner in? The answer is to hit home runs. Coaches and players alike believe that batters should adjust their swings to be more conducive towards hitting home runs because if they are going to hit the ball, they might as well hit it harder and further. While the adjustments aren't too big, to someone who has been swinging for contact their whole life, it can feel pretty different to swing for the fences more often than not. Coaches are pushing players to be stronger in areas of their body that adds power to their swings, ie quads, glutes, core and arms. In addition to adding strength in key areas, coaches are tweaking batters swings to have a slightly more upward motion so that when hit, the ball is given a better chance of going over the wall, and these tactics seem to be working in the last couple seasons. In the last few years, the league witnessed two substantial homerun records: 1) the most home runs by a single team, done by the Yankees who hit 267 total home runs in the 2018 season, and 2) the most home runs hit in total hit by all teams, which is 6,105 done in the 2017 season.[5] Moreover, the last three seasons have 3 of the top five most homeruns in total hit by all teams.[6] With the shift in the mechanics and strategy of hitting, we are

[5] http://www.baseball-almanac.com/hitting/hihr6.shtml
[6] http://www.baseball-almanac.com/hitting/hihr6.shtml

seeing batters begin to catch up to the dominant pitchers in the league. Unfortunately, these ebbs and flows are very difficult to measure because they are more about the mental side of baseball, which can not be directly statistically represented. This aforementioned issue gets to the larger problem of coming to conclusions about baseball based solely based on statistics. In a perfect world, we could somehow control for the emotions of athletes and their respective thought processes, so that the statistics reflect abilities on any given day, rather than including these abnormalities that are caused by the mental side of the game.

One of the main elements we wish we had included in our test is the development of other types of pitches and any significant changes that they have experienced. This was very difficult to do because while every pitcher does have a fastball, they do not have the same combination of other pitches. For example, some pitcher have just a fastball, a change up, and a curveball, while others may have a fastball, a slider, and a cutter. Due to the variety of combinations that make up pitchers' repertoires and the lack of other pitches that are so universal throughout the league's pitchers, we chose to just look at the fastball as that is what most baseball fanatics pay most attention to. One of the main reasons looking at other the development of other pitches could have lead to a better answer is because it would show us the relationship between each of these pitches and the overall amount of strikeouts, as opposed to just the relationship between fastball velocity and strikeouts.

These results answer many important questions, but they raise some important questions as well. What pitch do pitchers use as their strikeout pitch? Is it a fastball? How do pitchers react to different hitters, and how do they select their pitches accordingly? Where in the strike zone are pitchers throwing, and how has that changed over the course of the data revolution? How exactly does this increase in velocity affect pitchers' control? These, and more questions, remain to be answered. Luckily, as more and more data comes out, we will soon be able to learn more about the game, and answer some of these unanswerable questions.

Data Sources:

2012 Data:
https://www.fangraphs.com/leaders.aspx?pos=all&stats=pit&lg=all&qual=50&type=c,76,75,24,13,36,37,38,40,120,121,217&season=2012&month=0&season1=2012&ind=0&team=0&rost=0&age=0&filter=&players=0

2018 Data:
https://www.fangraphs.com/leaders.aspx?pos=all&stats=pit&lg=all&qual=50&type=c,76,75,24,13,36,37,38,40,120,121,217&season=2018&month=0&season1=2018&ind=0&team=0&rost=0&age=0&filter=&players=0