MLB Pitching Analysis

Adrean Lacay

Introduction

Among baseball fans, players, and everyone in between, there is a debate that is occasionally brought up: what is the most valuable part of the sport, pitching or hitting? Whether you like it or not, pitching is probably the most important aspect in the game of baseball. Yes, you need to score runs to win a game, but good pitching neutralizes good hitting most of the time. Even before the first ball is thrown, the team will work around their pitcher's abilities by setting up the most suitable defense behind them, as well as the opposition constructing a lineup to try and optimize their own offense. In an ideal world, pitchers can just follow a step-by-step guide to maximize not only their team's success, but their own success. However, that is not nearly the case since every pitcher's own personal make-up and upbringing is different. They throw pitches their body allows them to do and thus creates a several different pitching archetypes in the league today.

For this analysis, we will be looking at some of the more common standard and advanced statistics in baseball to look at pitching. More specifically, we will be analyzing the effects of a pitcher's arsenal, their most basic pitch, the 4-Seam fastball, and their most used breaking ball during the 2023 MLB season. We will examine the following questions:

1. *Does velocity and the number of pitch types play a factor in a pitcher's success?*
2. *Can we use a pitcher's speed difference from their fastball and secondary pitch to predict their whiff rate?*
3. *Is a pitcher's role reliant on their pitch usage and repertoire count?*

Terms & Definition

Fastballs & Breaking Balls

Although I never played or trained in organized baseball, the *4-Seam fastball* is the most basic and probably the first pitch a player learns. Once they get a grasp for it and develop their throwing mechanics, they may begin to throw pitches known as *breaking balls*. As the name suggests, these pitches break from the traditional straight path of their fastball and move across horizontally or down vertically. In this dataset, there are nine types of breaking balls: Changeup, Curveball, Cutter, Knuckleball, Sinker, Slider, Splitter, Slurve, and Sweeper.

Pitch Usage

Pitchers may work to include multiple breaking balls into their pitch arsenal along with their fastball and, depending on the scenario or their ability to throw each pitch, will have a different *pitch usage rate* from other pitchers. In most cases, being the simplest and fastest pitch, a pitcher will throw the fastball more often than their breaking ball. However, some will use a breaking ball more frequently if they feel that their fastball is not as effective. For the purpose of this analysis, regardless of its usage compared to the fastball, a pitcher's most used breaking ball will be referred to as their *secondary pitch*.

Starter vs. Reliever

Once they establish their pitch repertoire, pitchers may work towards becoming either a *starter* where they begin a game, or a *reliever* where they come in after the starter or another reliever. A pitcher's role comes down to not only their skillset, but also the team's need. While a starter will usually start throughout the season, a pitcher who may have been assigned as a reliever may also start if there is an injury or extra rest is needed. While the MLB has their own metrics when qualifying a pitcher as a starter or reliever, for this analysis, a starter must have started at least 50% of their games played.

FIP & Whiff Rate

In this day and age, there are a countless number of advance statistics used to quantify any aspect of the game. For this analysis, along with some categorical terminology, we will explore two of the more commonly used (and understandable) statistics: *FIP* and *whiff rate*. Fielding Independent Pitching, or FIP, only looks at results that are within the pitcher's control: strikeouts, walks, hit-by-pitches, and home runs, so it becomes a valuable statistic when looking at a pitcher's individual ability. For an in-depth look at its calculation, you can read this article by Piper Slowinski. The next statistic is whiff rate which is the percentage of swings-and-misses from the total number of swings. For example, if a pitcher drew 50 swings, 15 of which were missed by the batters, then their whiff rate would be 30%. Whiff rate becomes valuable when looking at a pitcher's pitch effectiveness as you want to minimize any balls in play.

Data Overview

As mentioned previously, the data consists of statistics from the 2023 MLB season. The site *Baseball Savant* was used to collect each pitcher's pitch arsenal and their statistics, and their overall season statistics was gathered from *Baseball Reference*. The data was initially filtered to include pitchers who faced at least 100 batters in the season. Additionally, as the pitcher's name were not identically formatted in Baseball Savant and Baseball Reference, changes were made to the Baseball Reference data through Excel and SQL so that each pitcher was assigned to an ID from Baseball Savant.

After removing duplicated pitchers and other rows that could affect our analysis, we are left with a dataset consisting of 442 pitchers.

```
## 'data.frame':    442 obs. of  10 variables:
##  $ pitcher_id     : num  425794 425844 434378 448179 450203 ...
##  $ pitcher_name   : chr  "Wainwright, Adam" "Greinke, Zack" "Verlander, Justin" "Hill, Rich" ...
##  $ fastball_usage : num  9.9 26.5 50 34.2 31.6 13.7 46.3 34.5 34.4 43.4 ...
##  $ fastball_speed : num  85.7 89.5 94.3 88.4 94.8 94.4 93.7 95.7 92.2 92.4 ...
##  $ secondary_pitch: Factor w/ 9 levels "Changeup","Curveball",..: 5 6 6 2 2 6 6 3 6 3 ...
##  $ secondary_usage: num  31.4 20 25.3 36.1 43.2 45.3 16.8 34.5 23.5 23.2 ...
##  $ secondary_speed: num  86.9 79.1 86.9 72 82.3 87.2 84 92.3 86.1 88.6 ...
##  $ whiff_pct      : num  13 17.3 22.5 19.3 32.2 21.7 28.6 23.3 19.3 28.7 ...
##  $ pitcher_role   : Factor w/ 2 levels "Reliever","Starter": 2 2 2 2 2 1 2 1 2 2 ...
##  $ fip            : num  5.99 4.74 3.85 4.87 3.87 6.13 4.32 2.44 7.02 5.53 ...
```

```
##   pitcher_id      pitcher_name fastball_usage fastball_speed secondary_pitch
## 1     425794  Wainwright, Adam            9.9           85.7          Sinker
## 2     425844     Greinke, Zack           26.5           89.5          Slider
## 3     434378 Verlander, Justin           50.0           94.3          Slider
## 4     448179        Hill, Rich           34.2           88.4       Curveball
## 5     450203   Morton, Charlie           31.6           94.8       Curveball
## 6     453268      Bard, Daniel           13.7           94.4          Slider
##   secondary_usage secondary_speed whiff_pct pitcher_role  fip
## 1            31.4            86.9      13.0      Starter 5.99
## 2            20.0            79.1      17.3      Starter 4.74
## 3            25.3            86.9      22.5      Starter 3.85
```

```
## 4          36.1          72.0      19.3      Starter 4.87
## 5          43.2          82.3      32.2      Starter 3.87
## 6          45.3          87.2      21.7     Reliever 6.13
```

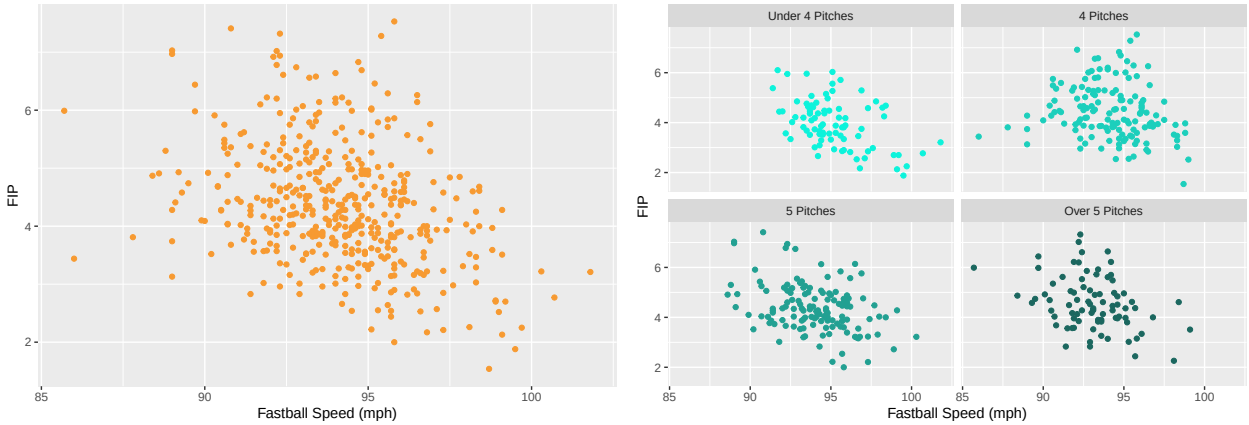
Analysis

1. Does velocity and number of pitches impact a pitcher's success?

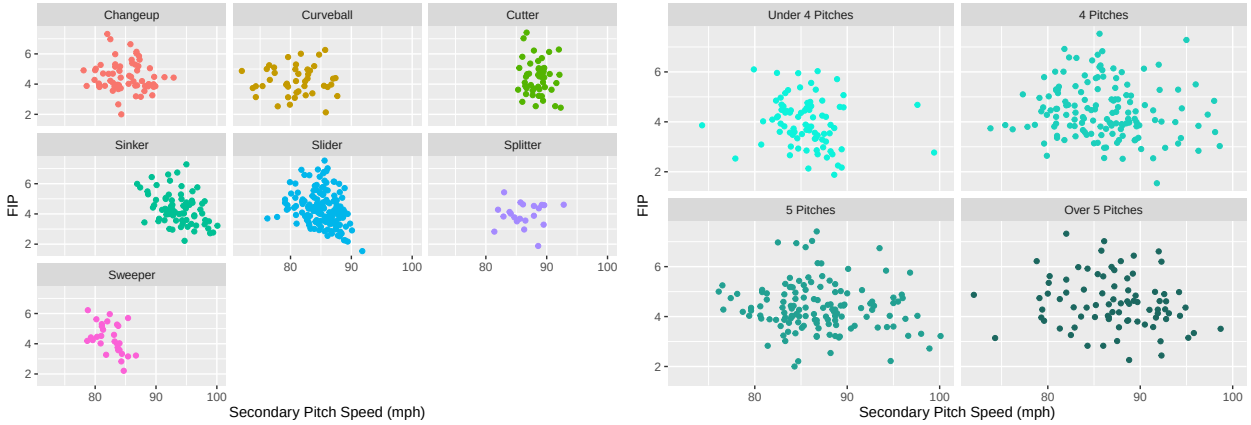
In recent years, velocity has become heavily emphasized as batters develop more power and become accustomed to using higher velocity pitching machines. As a result, pitchers have had to train to throw the ball harder to match the offense. Additionally, pitchers can couple their velocity with a larger repertoire to create some confusion in the batter's mind as to what pitch is coming next. Since FIP only accounts for outcomes that are in a pitcher's control, we can use it to measure their success against their fastball and secondary pitch velocity, along with their arsenal total. But first, we need to categorize each pitcher based on their number of pitches.

Arsenal Totals	Number of Pitchers
2	8
3	74
4	144
5	134
6	61
7	19
8	2

Based on these totals, most pitchers throw four or five different types of pitches with eight throwing two types and two throwing eight types (interestingly enough). We can use this information to create four categories: *Under 4 Pitches*, *4 Pitches*, *5 Pitches*, *Over 5 Pitches*.



Plotting fastball velocity and FIP shows a slight linear relationship that FIP decreases as fastballs increase in speed. Dividing the pitchers by their arsenal count category shows a similar result, except for pitchers throwing four pitches as it appears to be more scattered.



On the other hand, there seems to be no relationship between a pitcher's FIP and their secondary pitch speed, either as a whole or by their arsenal total. We can test our initial findings using linear regression to fit a model of fastball and secondary pitch speeds, separately, with a pitcher's arsenal count.

Coefficient Estimates of the Model: *FIP ~ Fastball Speed + Pitch Arsenal Category*

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	17.6676983	1.9853530	8.899021	0.0000000
Fastball Speed	-0.1439661	0.0208209	-6.914499	0.0000000
4 Pitches	0.3376659	0.1372135	2.460880	0.0142453
5 Pitches	0.2449158	0.1395968	1.754452	0.0800539
Over 5 Pitches	0.2917943	0.1593075	1.831642	0.0676853

Confidence Interval Test of the Estimates

	2.5 %	97.5 %
Intercept	13.7656710	21.5697256
Fastball Speed	-0.1848877	-0.1030446
4 Pitches	0.0679855	0.6073463
5 Pitches	-0.0294487	0.5192803
Over 5 Pitches	-0.0213098	0.6048984

With regards to their fastball velocity, using an arsenal of under four pitches as reference, the summary shows that a pitcher's fastball velocity is significant to determine their FIP. For every MPH, FIP decreases by 0.14. As pitchers develop a fourth pitch, their FIP increases by 0.34. If they decide to add a fifth pitch or more, their FIP increases by 0.24 and 0.29, respectively. However, based on their p-values and their confidence intervals containing 0, we cannot reject the null hypothesis. Therefore, these values are insignificant. This is somewhat surprising the plots suggested the opposite.

Coefficient Estimates of the Model: *FIP ~ (Secondary Pitch Speed x Secondary Pitch) + Arsenal Count Category*

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	6.3555447	3.3102033	1.9199862	0.0555287
Secondary Pitch Speed	-0.0261237	0.0388530	-0.6723733	0.5017115
4 Pitches	0.4477557	0.1419806	3.1536397	0.0017269

	Estimate	Std. Error	t value	Pr(>|t|)
5 Pitches	0.4343330	0.1471057	2.9525230	0.0033266
Over 5 Pitches	0.5430333	0.1677119	3.2378943	0.0012986
Secondary Pitch Speed * Curveball	0.0451215	0.0551846	0.8176472	0.4140168
Secondary Pitch Speed * Cutter	-0.0487239	0.0915491	-0.5322162	0.5948544
Secondary Pitch Speed * Sinker	-0.0954073	0.0555339	-1.7180028	0.0865245
Secondary Pitch Speed * Slider	-0.0760356	0.0487857	-1.5585621	0.1198444
Secondary Pitch Speed * Splitter	0.0526956	0.0872732	0.6038008	0.5462982
Secondary Pitch Speed * Sweeper	-0.1876530	0.0978417	-1.9179240	0.0557905

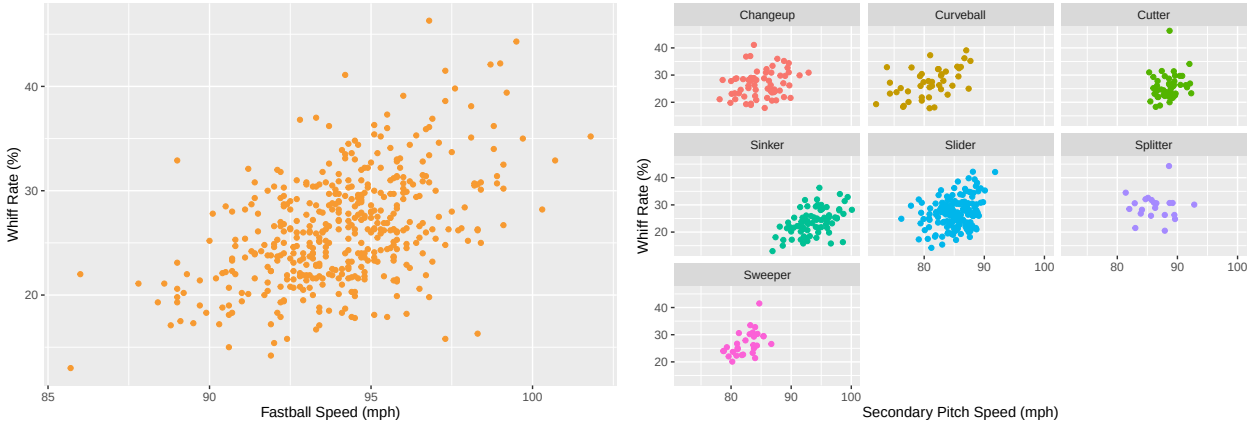
Confidence Interval Test of the Estimates

	2.5 %	97.5 %
Intercept	-0.1508633	12.8619526
Secondary Pitch Speed	-0.1024916	0.0502442
4 Pitches	0.1686841	0.7268273
5 Pitches	0.1451877	0.7234783
Over 5 Pitches	0.2133853	0.8726813
Secondary Pitch Speed * Curveball	-0.0633471	0.1535901
Secondary Pitch Speed * Cutter	-0.2286693	0.1312215
Secondary Pitch Speed * Sinker	-0.2045625	0.0137479
Secondary Pitch Speed * Slider	-0.1719270	0.0198558
Secondary Pitch Speed * Splitter	-0.1188452	0.2242365
Secondary Pitch Speed * Sweeper	-0.3799669	0.0046609

Meanwhile, having a Cutter, Sinker, Slider, and Sweeper appears to decrease FIP as you increase each pitch's velocity while a Curveball and Splitter increases it, as compared to having a Changeup with less than three pitches. However, these secondary pitches and their velocities are insignificant based on their p-values and confidence intervals. The only significant variable in this model is the pitcher's arsenal count. Using a Changeup and having a repertoire of under four pitches as reference, having four, five and above five increases a pitcher's FIP by 0.45, 0.43, and 0.54, respectively.

2. Can we use a pitcher's velocity difference from their fastball and secondary pitch to predict their whiff rate?

One of the most important skills for a pitcher is to minimize any contact. This will reduce the possibility of a hit or misplay by the defense, but batters have become trained to quickly identify and hit higher velocity pitches. Therefore, pitchers with slower pitches become more susceptible to contact. On the other hand, if they were to have a secondary pitch that is significantly slower than their fastball, then the difference in timing can force more swings-and-misses. We can plot a pitcher's whiff rate against their fastball and secondary pitch speed to gather some background information.



Doing so shows us that there is some linear relationship between whiff rate and fastball speed, but it is not as apparent with secondary pitches. Let's now take the difference in velocity from a pitcher's fastball and their secondary pitch.

Velocity Difference Summary

Minimum	Median	Maximum
-2.1	8.35	20.1

After taking the velocity difference from each pitcher, we see that the median difference is 8.35 mph. Given its name, it is safe to assume that the fastball is usually a pitcher's fastest pitch, but it appears that the lowest difference is -2.1 mph.

Pitcher Name	Fastball Speed	Secondary Pitch Speed	Velocity Difference	Secondary Pitch Type
Suter, Brent	86.0	88.1	-2.1	Sinker
Wainwright, Adam	85.7	86.9	-1.2	Sinker
Teheran, Julio	89.2	90.0	-0.8	Sinker
Priester, Quinn	92.8	93.5	-0.7	Sinker
Hudson, Dakota	91.0	91.6	-0.6	Sinker
Domínguez, Seranthony	97.5	98.0	-0.5	Sinker
Dunning, Dane	90.5	91.0	-0.5	Sinker
Floro, Dylan	92.3	92.7	-0.4	Sinker
Davies, Zach	89.3	89.7	-0.4	Sinker
Cuas, Jose	92.3	92.7	-0.4	Sinker
García, Luis	97.0	97.3	-0.3	Sinker
Corbin, Patrick	91.8	92.1	-0.3	Sinker
Graterol, Brusdar	98.3	98.6	-0.3	Sinker
Hoeing, Bryan	93.8	94.1	-0.3	Sinker
Winckowski, Josh	96.0	96.3	-0.3	Sinker
Quintana, José	90.2	90.4	-0.2	Sinker
Brito, Jhony	95.8	96.0	-0.2	Sinker
Syndergaard, Noah	92.2	92.3	-0.1	Sinker
Chafin, Andrew	92.2	92.3	-0.1	Sinker
Erceg, Lucas	97.9	98.0	-0.1	Sinker

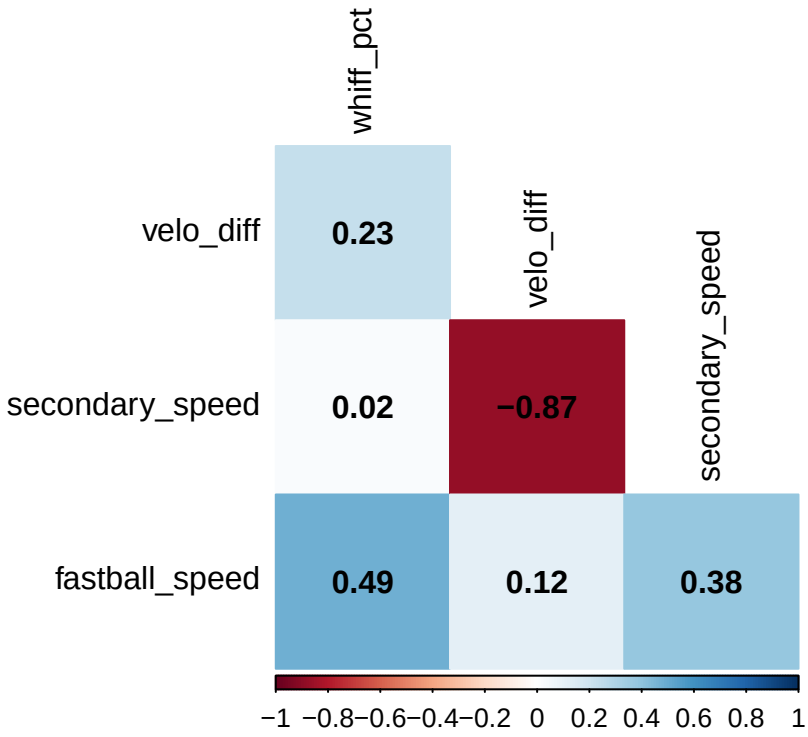
In total, there are 20 pitchers who have a negative velocity difference and, interestingly enough, all throw

Sinkers. Though pitching and physics may not be my specialty, this could be due to the higher amount of spin they put on this pitch causing it to create less air resistance. Regardless of the reason, we can keep this information for our model.

Coefficient Estimates of the Model: *Whiff Rate ~ Velocity Difference*

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	24.8972862	0.6235137	39.930617	0.0000000
Velocity Difference	0.2229088	0.0705633	3.158989	0.0017372

Once we split the data into a training and testing set and fit the model, we see that a pitcher's velocity difference has significant relationship to their whiff rate. However, according to the R^2 value, this model only explains about 3% of the variance of the expected whiff rates. This result carries over with an R^2 of 0.11 when using this model in trying to predict the whiff rates of the pitchers in our test data. Based on these findings, velocity difference alone may not be suitable in predicting whiff rate. We can go back and explore other possible predictors to add to this model to possibly improve it.



Looking at the correlation between fastball and secondary pitch speeds and whiff, there is very little relationship among them. Although, there is a strong negative correlation between velocity difference and secondary pitch speeds. We can interact these two terms to see if it improves the variability of our original model.

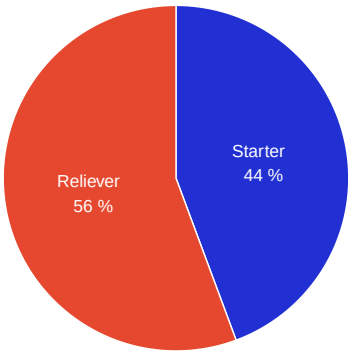
Coefficient Estimates of the New Model: *Whiff Rate ~ (Velocity Difference x Secondary Pitch Speed)*

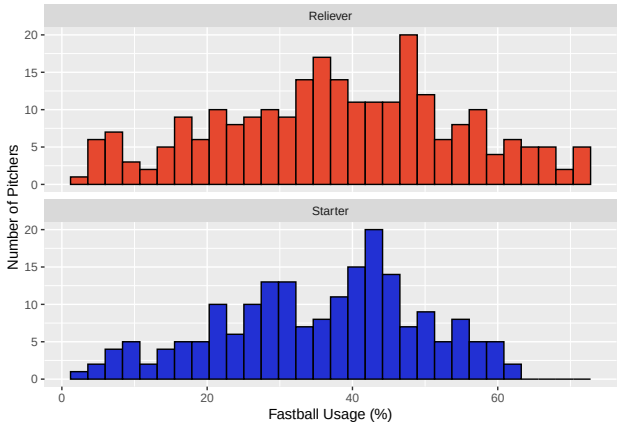
	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-61.6212441	12.6364146	-4.876482	0.0000017
Velocity Difference	-1.9310873	0.8868607	-2.177441	0.0301967
Secondary Pitch Speed	0.9083608	0.1357809	6.689900	0.0000000
Velocity Difference * Secondary Pitch Speed	0.0377301	0.0101106	3.731720	0.0002259

Introducing an interaction term with the secondary pitch speeds does improve the model. However, there is still a high level of variability with this predictor as the R^2 value is only 0.29 when trying to predict with our testing data. So, while this predictor is significant, there may be other factors outside of this data set that better predict a pitcher's whiff rate, such as pitch movement.

3. Is a pitcher's role reliant on their pitch usage and repertoire count?

In theory, since starters must pitch more innings, they would want to throw less fastballs and more breaking balls as they do not want the opposing batters to adjust to their fastball's speed and movement. In turn, these pitchers would likely have to have more pitches in their arsenal to keep the offense thinking about the next pitch.

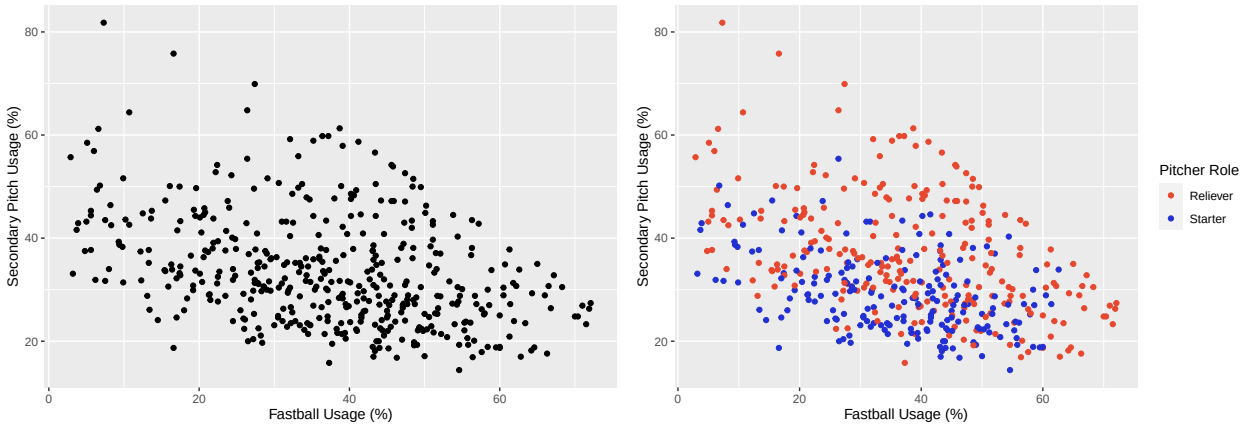




Fastball Usage Over 50%

Pitcher Role	# of Pitchers	% of Role
Reliever	59	24.0
Starter	31	15.8

Using the definition of a starter in this analysis, we see that there are less starters (196, 44%) than relievers (246, 56%). There are also more relievers who throw a fastball more than 50% of the time. So far, this follows our hypothesis regarding the fastball usage of starters vs. relievers.



When we plot the secondary pitch usage against the fastball usage, it appears as though secondary pitch usage decreases as the number of fastballs increases, which is not very surprising. However, if we colour each data by their pitching role, relievers take up the upper portion of the plot. This suggests that relievers not only throw more fastballs, but they also throw more of their secondary pitch than starters, which goes against our theory. Due to the variability in the data and lack of definitive clusters, we can use a random forest to try and predict a pitcher's role with their pitch usage.

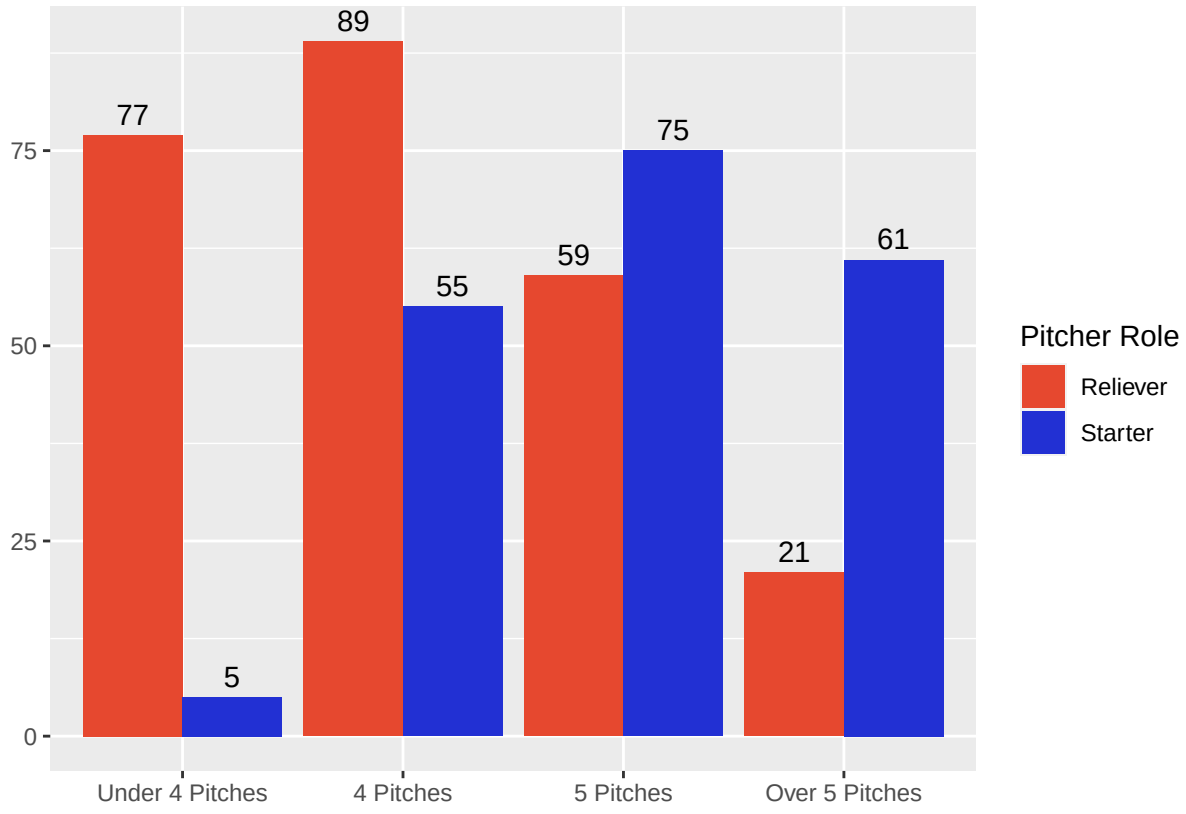
Results of Training Model: *Pitcher Role ~ Fastball Usage + Secondary Pitch Usage*

	Reliever	Starter	Class. Error	OOB Estimate
Reliever	113	57	33.53	37.97
Starter	63	83	43.15	

Prediction Results from the Test Data

	Reliever	Starter	Class. Error
Reliever	49	18	26.87
Starter	27	32	45.76

With fastball and secondary pitch usage as our predictors, a random forest does a fair job with an out-of-box (OOB) estimate error rate under 40%. Although, this model misclassifies over 40% of starters. This is reflected when attempting to predict pitching roles with the testing data. It appears that usage alone is not sufficient in determining a pitcher's role, so we can try including another predictor: arsenal count.



Five or More Pitches in Arsenal

Pitcher Role	# of Pitchers	% of Role
Reliever	80	32.5
Starter	136	69.4

There are 136 starters who throw five or more pitches as compared to 80 relievers. This difference in arsenal count can explain the disparity in pitch usage between roles. With a larger repertoire, one's pitch distribution is spread out more and the usage of each pitch becomes smaller. Since there are more starters who throw five or more pitches, their fastball and secondary pitch usage is lower than relievers that throw less pitches.

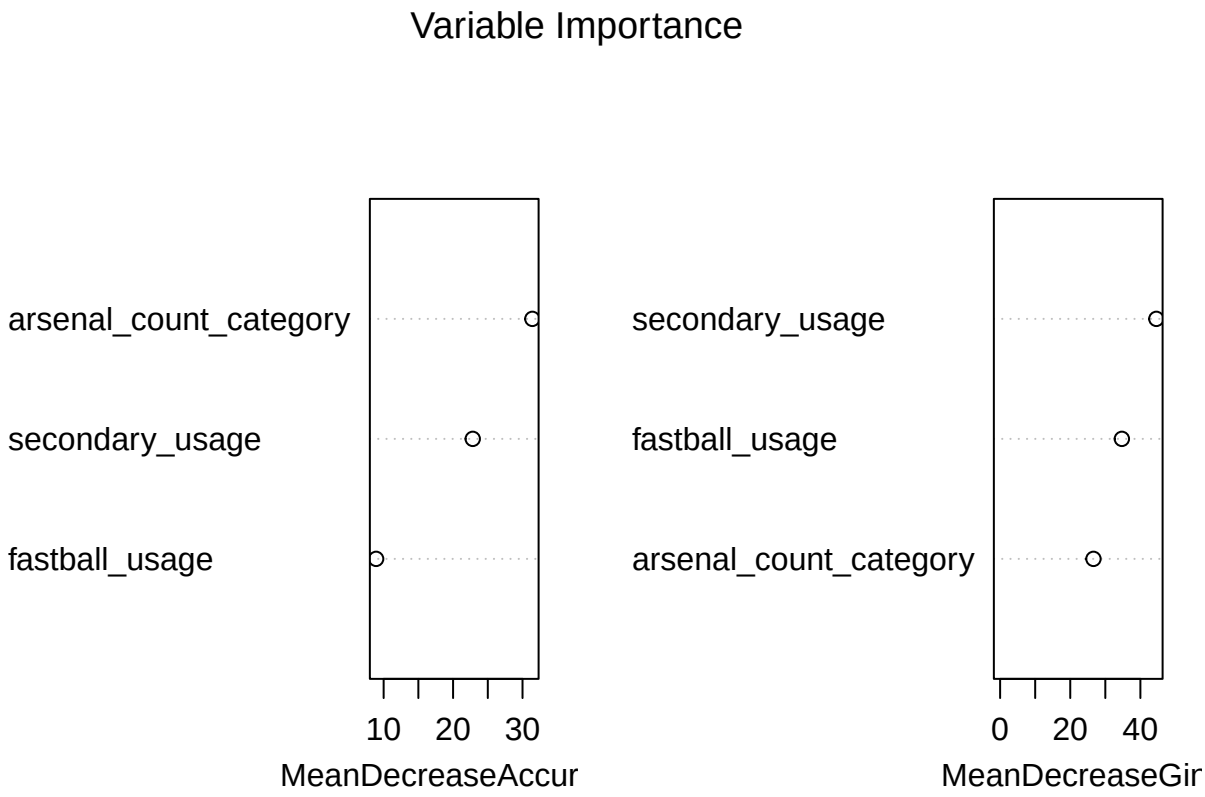
Results of Training Model: *Pitcher Role ~ Fastball Usage + Secondary Pitch Usage + Arsenal Count Category*

	Reliever	Starter	Class. Error	OOB Estimate
Reliever	111	59	34.71	31.96
Starter	42	104	28.77	

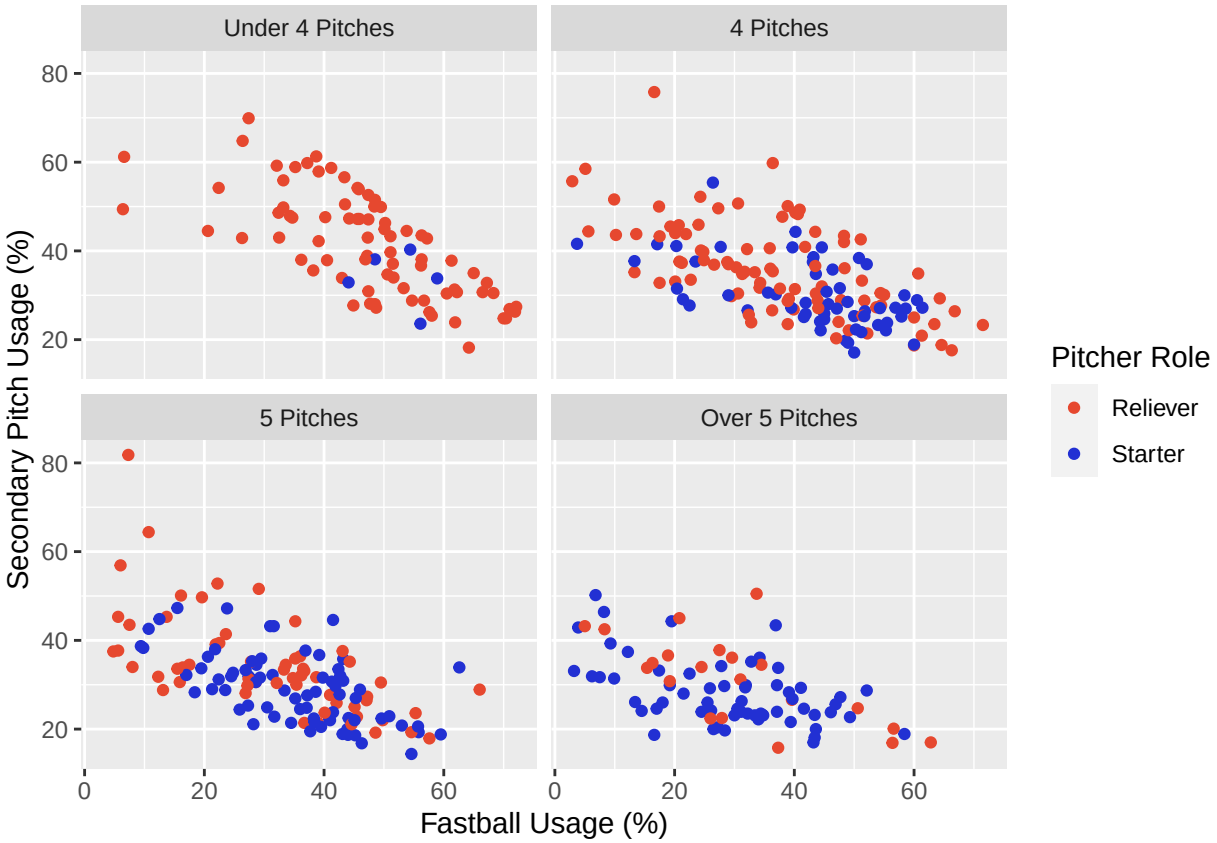
Prediction Results from the Test Data

	Reliever	Starter	Class. Error
Reliever	48	11	18.64
Starter	28	39	41.79

By introducing the arsenal count with the pitch usages, both OOB estimate and classification error rate for starters decrease. On the other hand, the error rate for classifying starters is around 40% once we try to predict the roles using our testing data. Perhaps with a larger sample size, these rates would decrease.



Including the arsenal count into the model seemed to have the most impact. As shown in the importance plot, a pitcher's arsenal total has the largest mean decrease in accuracy of the three predictors, so the model's accuracy becomes worse without it. On the contrary, its mean decrease in Gini is the lowest in the model. This suggests that a pitcher's repertoire count is not significant in distinguishing between starters and relievers, as compared to fastball and secondary pitch usage. This disparity in accuracy improvement but insignificance in classification could be caused by having both continuous and categorical predictors in this model. Due to the nature of the algorithm, random forests tend to favour continuous predictors because of the large number of possible points to split into a correct outcome or category. The same can be said for categorical variables with many factors. Since this predictor only has four categories, the probability of one's arsenal count predicting their role is lower than their pitch usage.



Moreover, while a pitcher's arsenal count alone may be considered when assessing their role (as seen with the bar graph and plot above), there are no distinguishing features when classifying roles with their pitch usage included.

Conclusion

Pitching, at the end of the day, is a human activity. It is a fact that there will be variability among pitchers, even if they throw the exact same pitches and velocities. Each pitcher has their own mechanics and limitations that ultimately affect their skillsets and overall success. So, while players in training look at their idols and other professionals, it is imperative that they understand what they can control and work on improving those aspects. One of the main reasons for this analysis was to not only examine the effects of pitches and pitch numbers on a player's success, but to also look at the increase in injury. More specifically, I wanted to see if velocity is as important as many teams believe.

For the past couple of seasons and, more noticeably, the beginning of this season, there has been an uptick in injuries among pitchers. This included All-Star and Cy Young-caliber pitchers like Shohei Ohtani, Spencer Strider and Shane Bieber. One possible culprit that players and officials look at is the introduction of the pitch clock. This has reduced the amount of rest in-between pitches and forcing pitchers to throw more pitches in a shorter amount of time, leading to more stress on the arm and causing injuries. However. older and former players are pointing to the increased focus on higher velocity. Most of these injured pitchers are known to throw nearly 100 MPH fastballs on a consistent basis as teams feel it's the best way to counteract the increase in power hitters. As this becomes an emphasis for being successful in the major leagues, pitchers have been taught at a young age to throw hard. As a result, their arm has experienced years of wear by the time they get called up.

Since the pitch clock was only implemented last season, there is not enough data to compare it to the past hundred or so years of baseball. Pitch speeds and arsenal data, however, is available and could be used to examine its impact on the game. With this information, I could try to get a sense of velocity's importance to a pitcher's success. If the results showed an insignificant relationship, this could have helped lower the emphasis on velocity, and save and extend the careers of many pitchers. This analysis also aimed to define pitching roles so that it could help pitchers understand what they need to focus on when training. While it appears that fastball velocity may be more important than their secondary pitch and factors other than usage may determine a pitcher's role, there is an abundance of statistics and tracking software that can be used to hopefully find some sort of solution to this dilemma. Keeping the best pitchers healthy should be prioritized to ensure the best possible game for both players and fans.