Pitching Data Analysis Project

Baseball has been and always will be a game centered around statistics. From the mid-19th century creation of the basic stats found within a box score up to the modern age of sabermetrics, people have continued to create new ways to determine the value of each player. As I begin my career, I aim to manipulate these existing statistics, and create entirely new ones, in order to help a team build a successful future, both on and off the field. With this project, I am laying the foundation towards these goals by examining some pitching statistics that have intrigued me and looking at the impact that they have on one another.

I was a pitcher from youth through college, which has given me a very strong understanding of what makes each pitch successful. While your average baseball viewer may think velocity and control is the end all be all of what makes a pitcher "good," my years of experience have definitively proven that this view is inaccurate. Factors such as pitch sequencing, spin rate, extension, and arm slot all play major roles in determining how well a pitcher does. However, anecdotal evidence needs to be quantified in order to choose which specific players are picked when building a team or to fix flaws in pitchers that are already with a team.

In order to begin this analysis, I needed a set of data to work with. I initially tried to download the pitch-by-pitch Statcast data for the 2020 season, however it was too large of a selection to process. As a result, I chose to use a sample of the Statcast readings of 100k pitches from the 2017 season (link). Although this does not cover the entire season, it is a large enough sample to be able to get accurate estimates for the statistics that I will be examining. I will also be using the Lahman database to compare these Statcast measurements to the year-end

statistics for each pitcher (link). I acknowledge that these year-end statistics are based off of more data than the sample that I am using, and therefore my results may not be completely precise if applied in the real world. However the methods that I am using would give completely precise results when applied to the entire season's worth of data, if my computer had the capability to process that amount of information. Lastly, I will use R to conduct my analysis, which I have used extensively during my time at UC Davis.

I first wanted to look at the impact of pitch sequencing on the outcome of each at bat, however the data that I used was a random selection of 100,000 pitches from the 2017 season, rather than a random selection of at bats from that season. As a result, I was not able to analyze individual at bats, and therefore was not able to determine the sequences that pitchers used within each at bat. I then pivoted towards analyzing the impact that first pitch strike percentage had on a pitcher's effectiveness. In order to do this, I first separated the first pitches from the overall Statcast data. I then created a function which calculates the first pitch strike percentage for a given pitcher and ran this function across all pitchers for which Statcast data was recorded to create a new data frame composed of each pitcher's MLB ID and their corresponding first pitch strike percentage. I then moved over to the Lahman database, which contains the year end statistics for each pitcher. I first combined the statistics for the pitchers who changed teams midseason in order to create a unique ID for each pitcher, as well as adding a few statistics that I was interested in examining, K/9 and BB/9. I could now add my calculated first pitch strike percentages to this data frame and look at the relationship between first pitch strike percentage and any of the statistics contained within this data set.

The most basic statistic which measures a pitcher's success is ERA, so although it is dependent on more than a pitcher's raw ability, I thought it was a good place to start. I first narrowed down the data to pitchers who have thrown at least 25 innings, as pitchers with very small sample sizes may skew the data. I then regressed ERA on first pitch strike percentage and created a plot to show the relationship between the two variables. When looking at the graph, it is apparent that the relationship between ERA and first pitch strike percentage will not fit a linear model well (Figure 1). However, when analyzing the regression output, there is very clearly a negative relationship between the two variables, as indicated by the very low p-values. The low R-squared confirms the inability to forecast ERA using this model, as the variability is very high. Although the coefficients should not be used to predict a pitcher's ERA solely based off of his first pitch strike percentage, this model confirms that there is a relationship between the two variables, and if a large regression model is created to accurately predict ERA, including first pitch strike percentage would be warranted.

I then decided to analyze the impact of first pitch strike percentage on a couple statistics that are more easily controlled solely by the pitcher, K/9 and BB/9. I followed the same steps that I used when analyzing ERA, however the results were much different. The K/9 vs first pitch strike percentage plot showed zero relationship between the two variables, and the linear model confirmed this (Figure 2). This result was surprising and is something that I would like to analyze further, as there are numerous possible reasons for this lack of a relationship. On the other hand, the BB/9 vs first pitch strike percentage plot showed a clear negative relationship between the two variables (Figure 3). As with the ERA model, the R-squared was very low, so the model should not be used to predict ERA, but it does show that pitchers who throw more

first pitch strikes tend to walk less batters, which lines up with my expectations. In conclusion, the ability to consistently throw first pitch strikes does not seem to have an impact on the ability to strike batters out, but does correlate with a pitchers ability to control walks.

I next wanted to look into some more advanced metrics, rather than statistics that purely measure ends results. For example, statistics such as ERA, K/9, and first pitch strike percentage all measure outcomes of pitches, whereas statistics such as spin rate and exit velocity measure the pitches or batted balls themselves without looking at whether or not those pitches were strikes or those batted balls fell in for hits. These raw statistics can be used to more accurately predict future performance for a player, as they eliminate luck. For example, a batter may have elite exit velocity, but if he gets unlucky and hits the ball right at the fielder every time, the more common statistics such as batting average and slugging percentage will indicate that he is not a good hitter. As I did with the first pitch strike percentage, I first created a function to calculate the average exit velocity against for a given pitcher in the data set, then added the exit velocities for each pitcher to the existing data frame. I thought it would be interesting to look at the impact that first pitch strike percentage had on exit velocity. However after regressing exit velocity on first pitch strike percentage and comparing the two variables in a plot, I found no significant relationship.

I next wanted to examine the relationship between spin rate and opponent's exit velocity. On one hand, higher spin rates tend to correlate with higher velocity, and if one barrels up a pitch with higher velocity the exit velocity tends to be higher. On the other hand, higher spin rates also indicate more movement, which makes it much harder to barrel up a pitch. As a result, I am very intrigued to see the true impact of spin rate on exit velocity. It was

important to first split the data into separate sets for each pitch. Spin rates will vary greatly depending on which pitch is thrown, and for some pitches, such as a sinker, a lower spin rate is actually more desirable. I then displayed the relationship between spin rate and opponent's exit velocity for each pitch (Figure 4), which appeared to show little or no relationship between the two variables, regardless of pitch. To prove this, I ran regressions for both fastballs and sliders and found insignificant p-values, which indicates a lack of linear relationship. I then looked at the relationship between velocity and spin rate to help explain this. I again looked at every individual pitch in a graph, then ran regressions for sliders and fastballs. Looking at the graph (Figure 5), it appeared that fastballs with higher velocities tended to also have higher spin rates, while off speed pitches such as changeups and sliders seemed to have no relationship between the two variables. This was confirmed when looking at the results of the regressions, as the fastball regression showed a statistically significant positive relationship, while the slider regression showed no evidence of any relationship. The higher velocities for higher spin rate fastballs may explain why spin rate does not directly affect exit velocity when looking at the data in whole. As a result, in order to see the true impact spin rate has on exit velocity, it would be important to further separate the pitches by velocity, then repeat this process for each velocity group.

Lastly, I wanted to take a look at extension and the effects that a greater extension has on a pitcher's ability to miss barrels. If a pitcher has a greater extension, that means that they release the ball closer to home plate, which should increase the perceived velocity from the batter's perspective. I first wanted to confirm this relationship between extension and effective velocity, so I displayed the two variables in a plot (Figure 6), separated by each pitch type, and

ran a regression on the fastball data. Both the plot and the regression output indicated that there was a significant relationship between the two variables. I then compared extension and effective velocity to exit velocity. Both the plot (Figure 7) and the regression output indicated zero relationship between the two variables. This seems to show the conflicting effects of velocity. On one hand, when a faster pitch is barreled the resulting exit velocity will also be higher. However, it tends to be much harder to barrel a faster pitch, which creates a larger amount of batted balls with very low exit velocities. The data that I used, which looks at the entire range of batted balls, shows that these two effects essentially even each other out in aggregate.

In conclusion, although it is simple enough to compare pitching variables with one another, it is impossible to be able to predict a pitcher's effectiveness using any single statistic. In order to accurately gauge a pitcher's overall abilities, a combination of many different statistics must be looked at and weighed against one another. For example, a pitcher may have elite velocity and spin rate, but if he has zero control of the strike zone he will not be effective. I aim to find a system that will determine which individual characteristics of pitching truly matter the most. In an age where masters of control like Kyle Hendricks are achieving similar end results to flame throwers such as Luis Castillo, it can be difficult to determine how to weigh control versus velocity. Is it better to have a pitcher that can evenly balance the two values, or is it better to have a pitcher that has mastered one of them? I would like to be able to find a value that can accurately compare the two and determine how much an added unit of control is worth in relation to an added unit of power.

Appendix

Figure 1

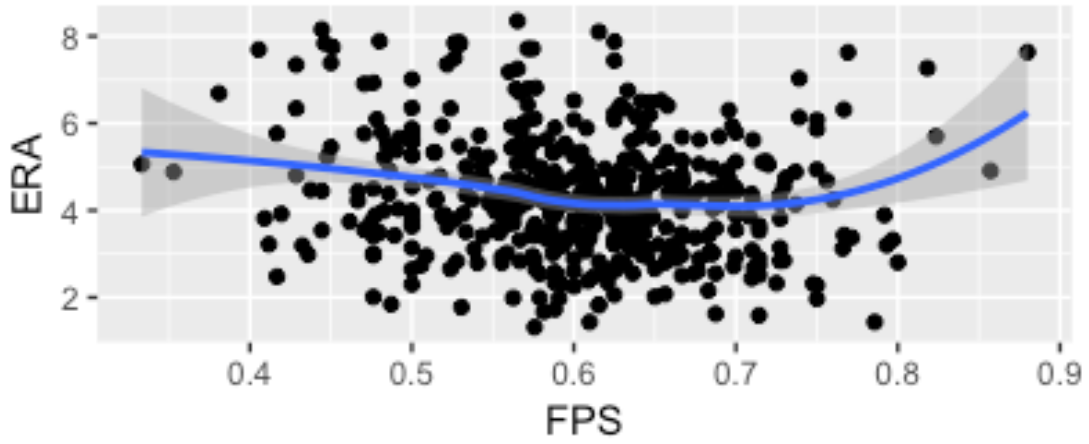

Figure 2

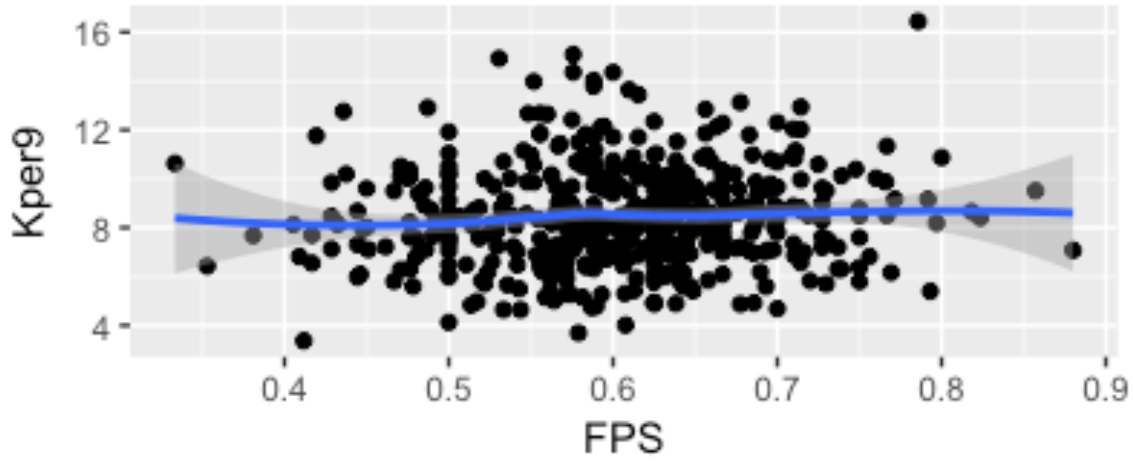

Figure 3

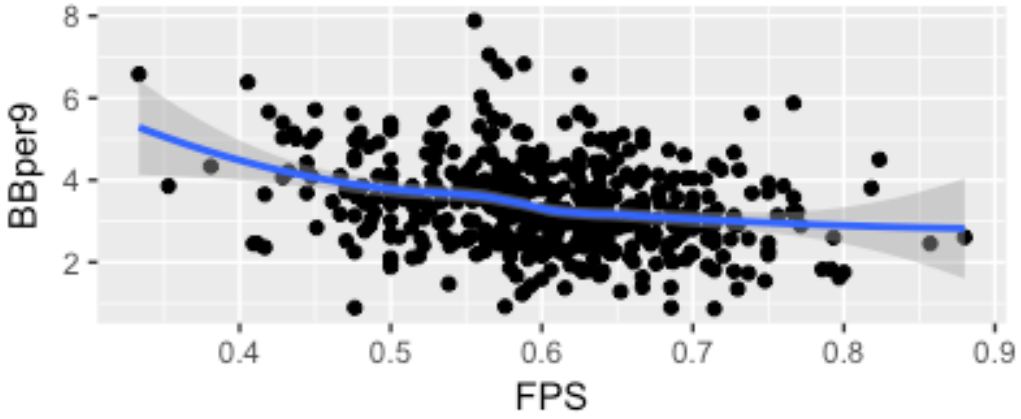

Figure 4

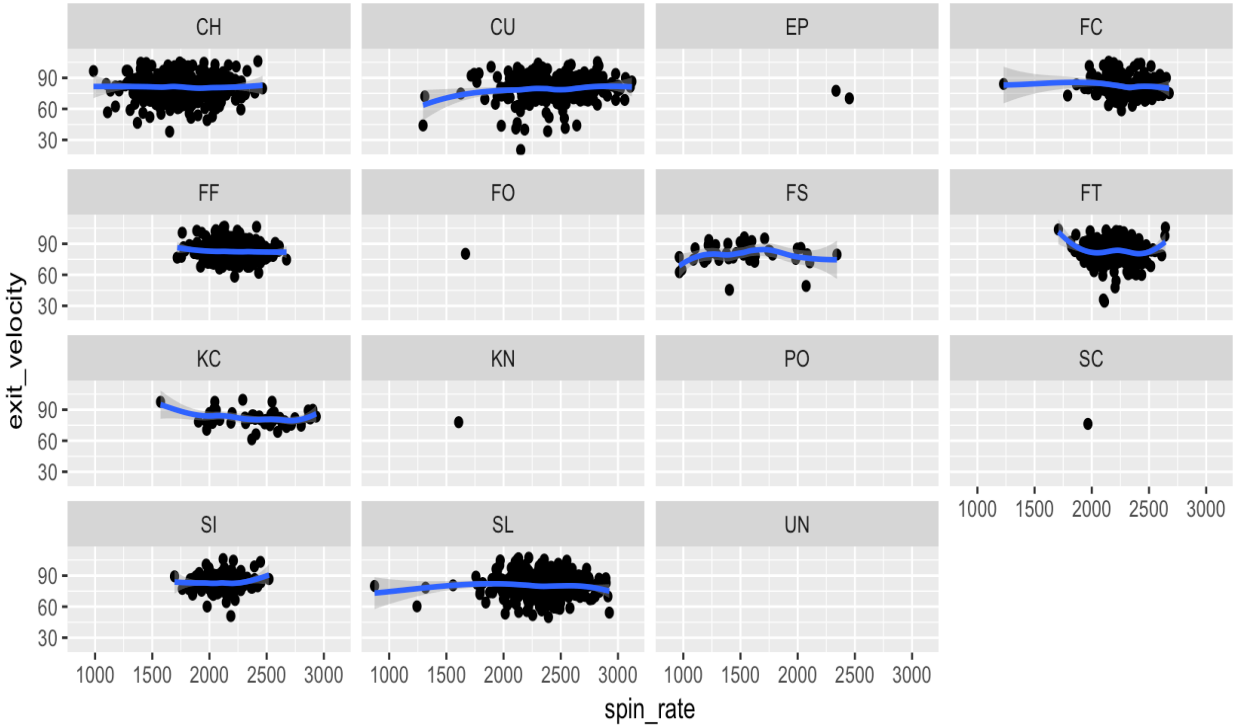

Figure 5

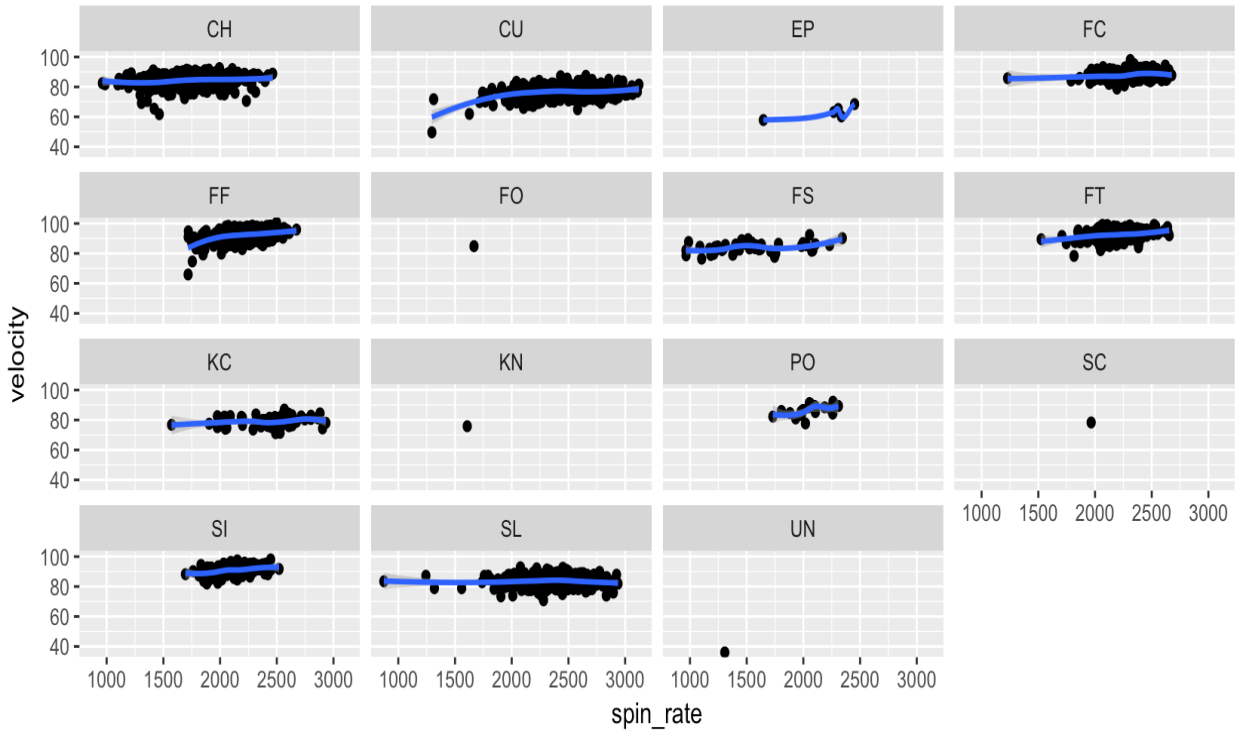

Figure 6

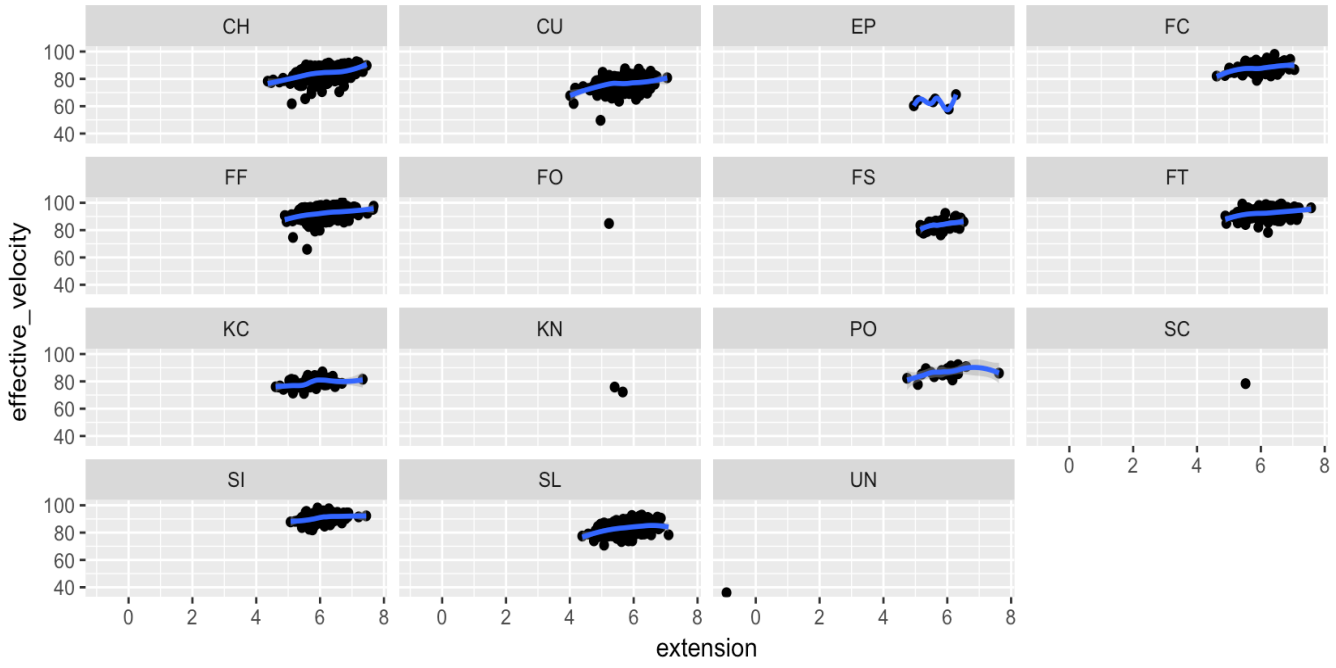

Figure 7