Evan Calkins

Data Visualization Final Project

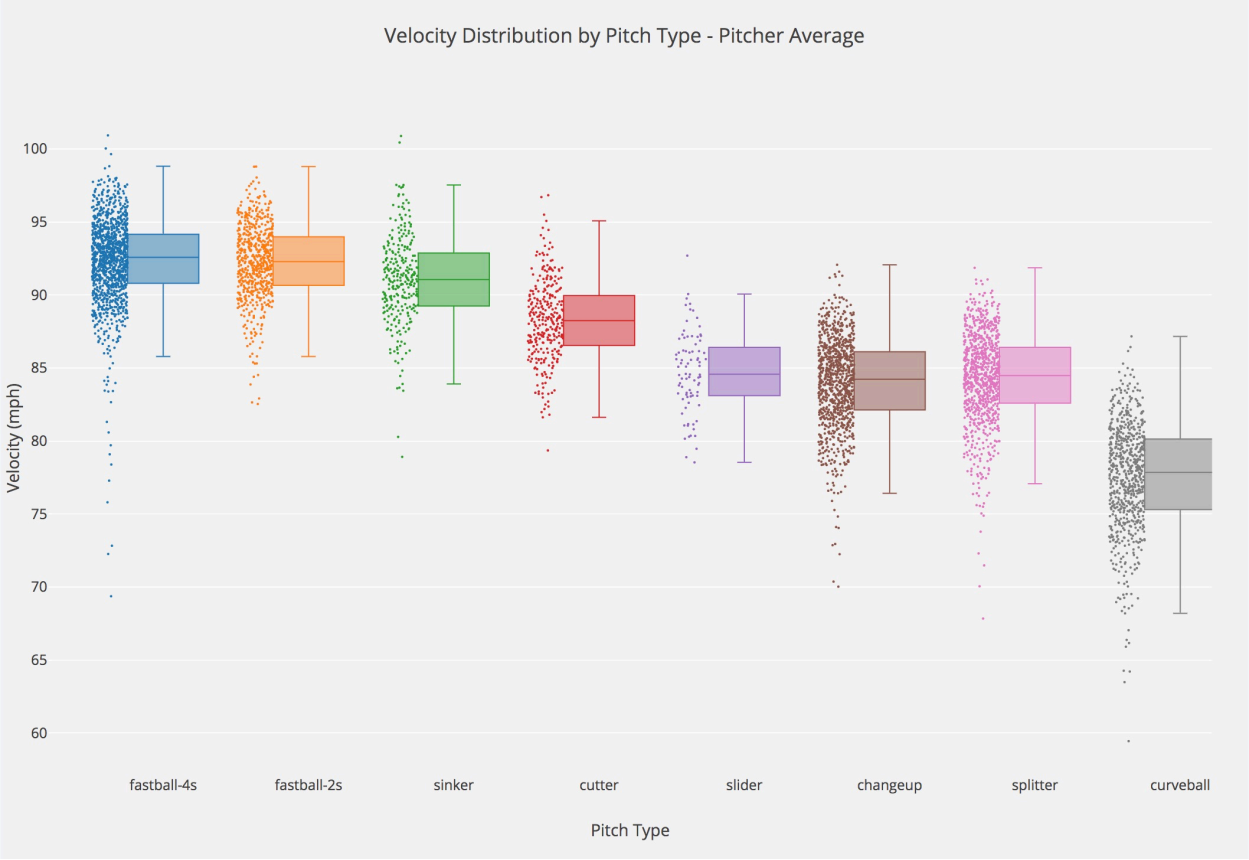
Introduction

The dataset I chose to analyze for my data visualization final project was a dataset of Major League Baseball (MLB) pitch data from the 2018 season that I joined with many different supplemental tables including details on the game situation, at bat information, stadium location, etc. The primary tables (pitches, at_bats, games) came from Kaggle and were originally scraped from the MLB.com website. I sourced the stadium details from Fusion Tables and the list of top MLB prospects from "The Baseball Cube".

I chose to analyze this data because it combines two of my passions – data science and baseball. Baseball is perhaps the most analytics-driven sport out there and I have not yet had the opportunity to apply my newly developed data science skills to a baseball-related question up until this project. I chose to add in the stadium location data because I thought it would be interesting to visualize certain things about team performance and metrics with respect to their location.

There are many questions that I am looking to explore with this data. My main goal is to better understand the types of pitches thrown in the MLB, how the different features of pitches vary (pitch type, velocity, location), and what features makes a pitch or pitcher particularly effective. In answering this last question, I plan to analyze the relationship between performance metrics for a pitcher and the features of their typical pitches. Since map-based plots don't necessarily apply to the main goal of my project, I will explore tangential questions using these plots. In particular, I want to look at how the number of home runs varies by location (due to ball park size) and how that correlates with the stadium's average attendance.

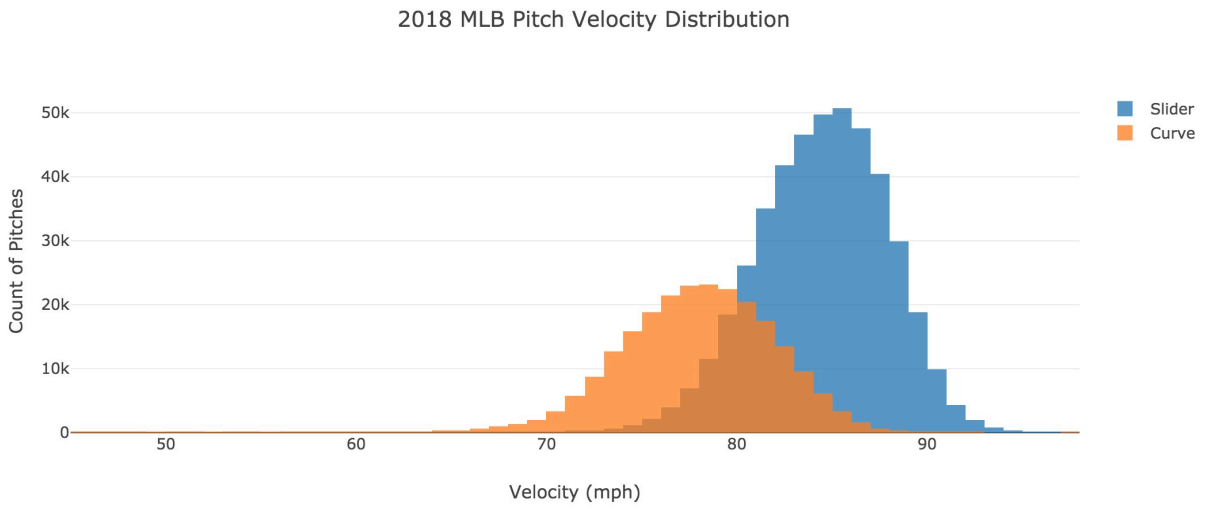
Boxplot (interactive)



https://plot.ly/~ecalkins/86/velocity-distribution-by-pitch-type-pitcher-average/#/

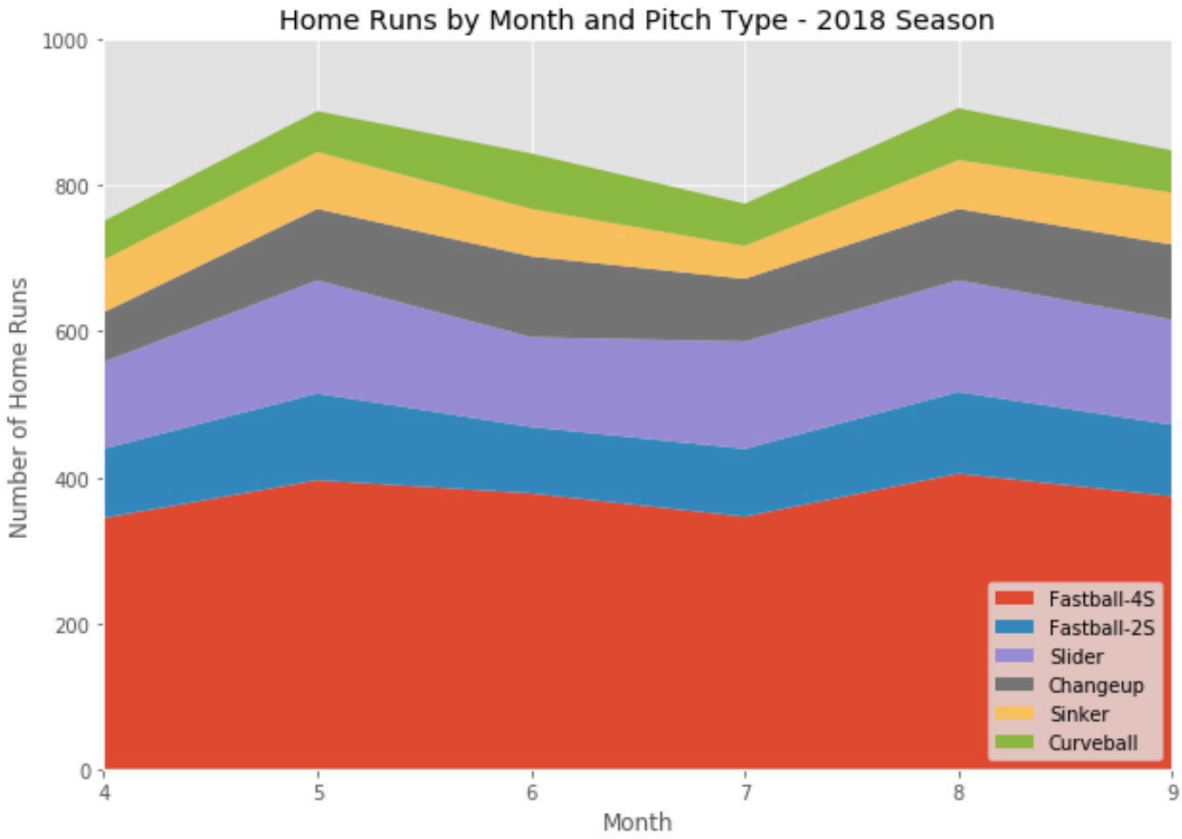
This plot shows the distribution of average pitch velocity by pitch type for MLB pitchers in the 2018 season. Each data point represents the average for a particular pitcher across all of the pitches of that type that they threw in the season. It is interactive in that you can see who the pitcher is for each data point. I chose this plot to get an initial idea of the velocity distribution for each pitch type before diving deeper into this feature.

Histogram



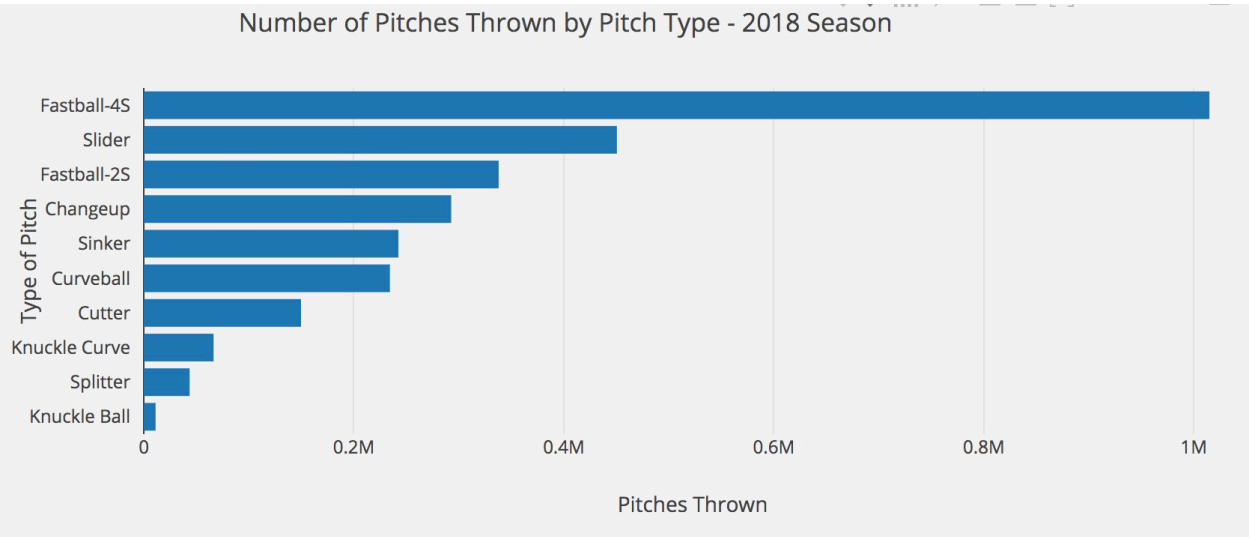
This histogram shows the count of pitches by velocity for pitch types of slider and curve for the 2018 season. These are two pitch types that are very similar and I had always wondered exactly what separates one from the other. I knew that a curve was generally slower than a slider, and this plot depicts by exactly how much, on average.

Stacked Area Graph



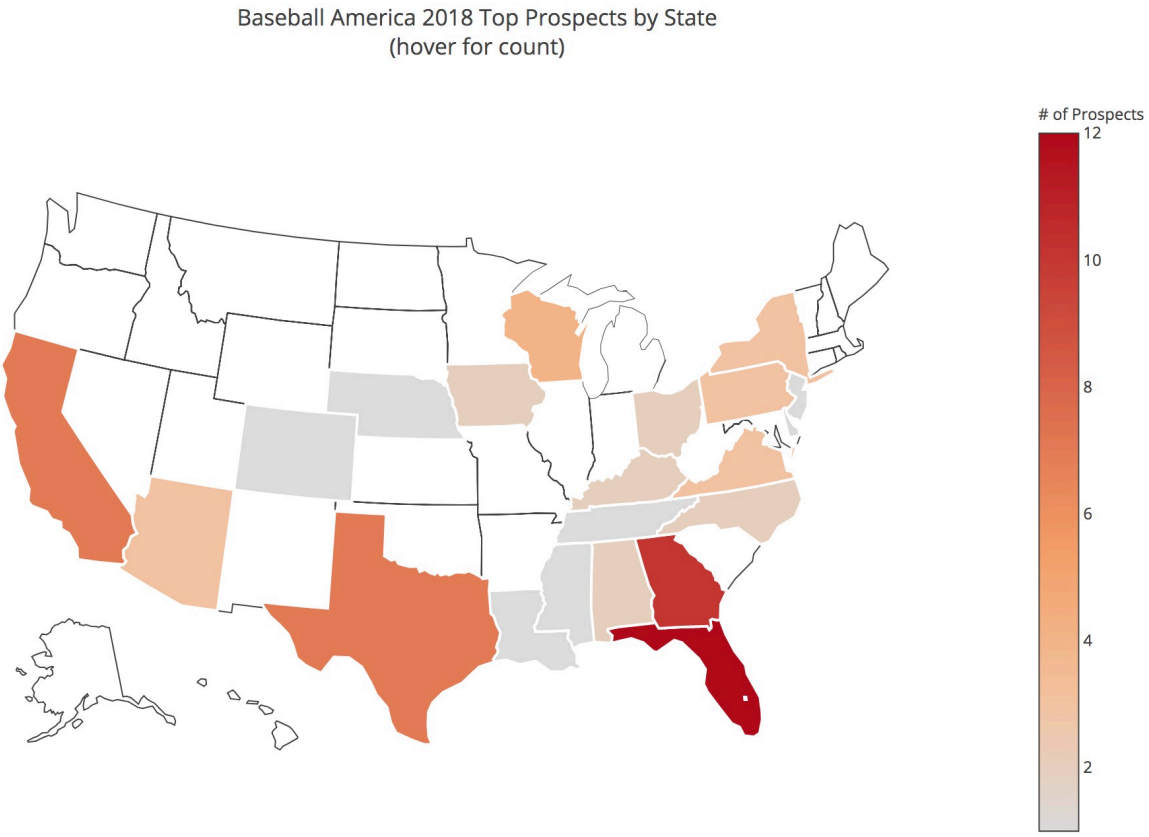
I'm also interested in exploring the relationship between pitch type and the number of home runs given up, so I decided to plot a stacked area graph of the count of home runs by pitch type for each month in the 2018 baseball season for the top six most commonly thrown pitches. Here we see that there is a cyclical trend for home runs throughout the season, and the four-seam fastball is the most common pitch to hit a home run off of.

Bar Plot



Based on the results from the previous plot, I was interested in exploring the number of pitches thrown by pitch type. An interesting finding is that changeups are thrown less than two seam fastballs, but they have very similar counts of home runs hit against them. It seems that changeups are slightly harder to hit home runs off of than two-seam fastballs.

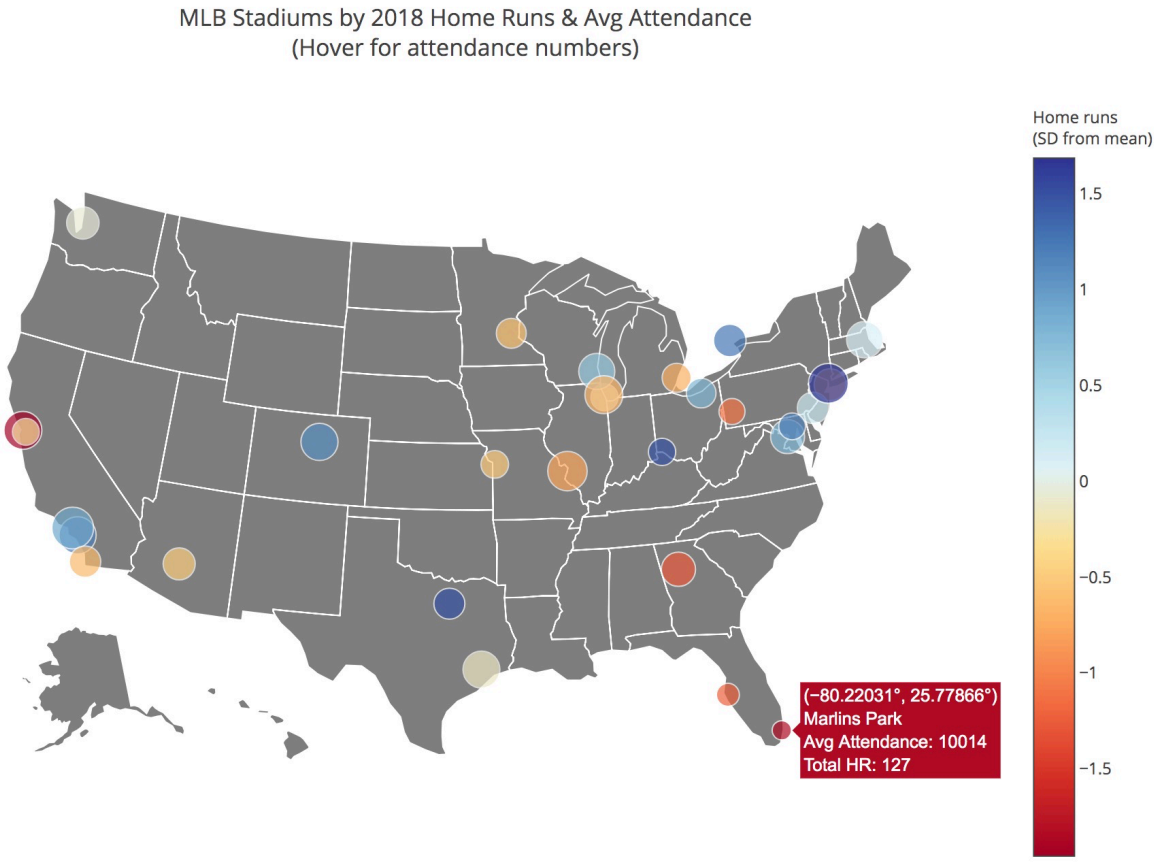
Chloropleth



https://plot.ly/~ecalkins/110

This plot incorporates the prospects data, since it was not interesting to show maps of pitch data. Here we see the number of prospects coming from each state in the U.S. (by hometown) for the top MLB prospects in 2018 as rated by Baseball America magazine. This plot is interactive in nature, so the number of prospects that helm from each state can be seen by hovering over the state.

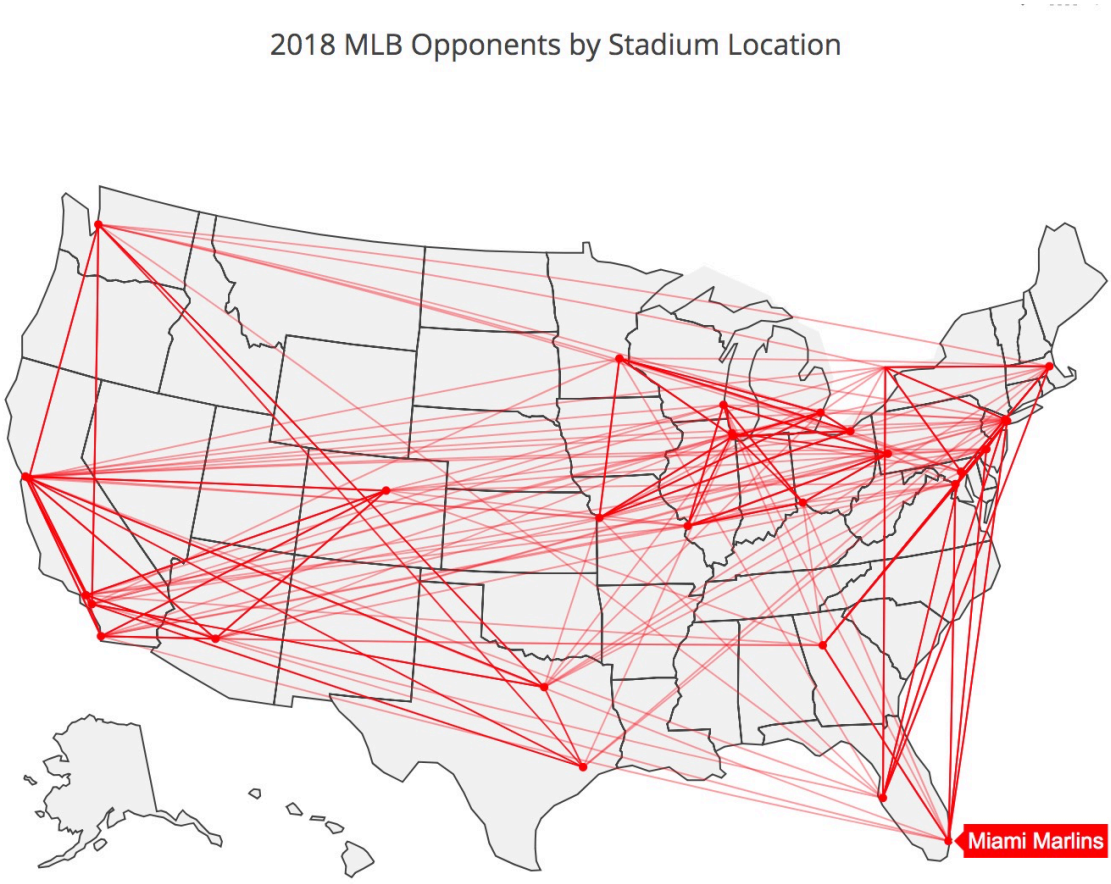
Bubble Map (interactive)



https://plot.ly/~ecalkins/106/mlb-stadiums-by-2018-home-runs-avg-attendance-hover-for-attendance-numbers/#/

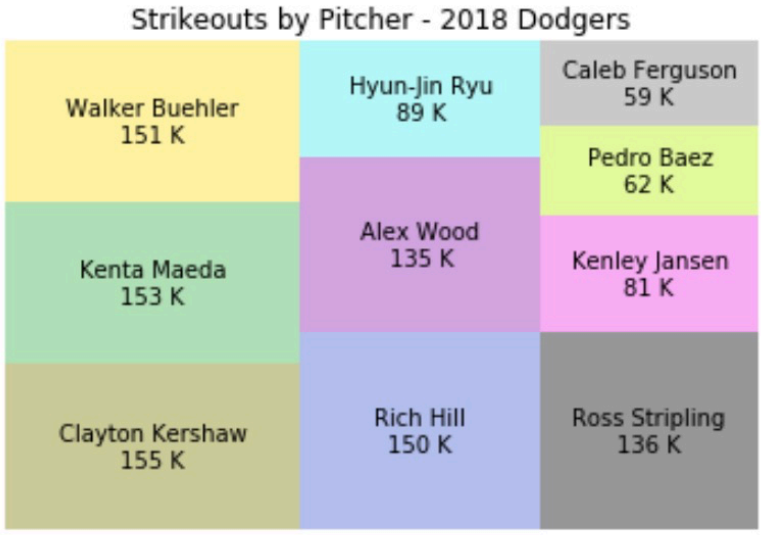
In this bubble map, the size of the bubble corresponds to the average attendance for games in that stadium in 2018, whereas the color of the bubble corresponds to the stadium's home run factor. By home run factor, I mean the number of standard deviations from the mean number of home runs hit in an average stadium. In baseball, some stadiums are at higher elevation, have shorter fences, or have less wind, and the likelihood of a home run is dependent on these factors. Fans also enjoy watching home runs, so it makes sense that parks with more home runs may see a higher average attendance.

Connection Map



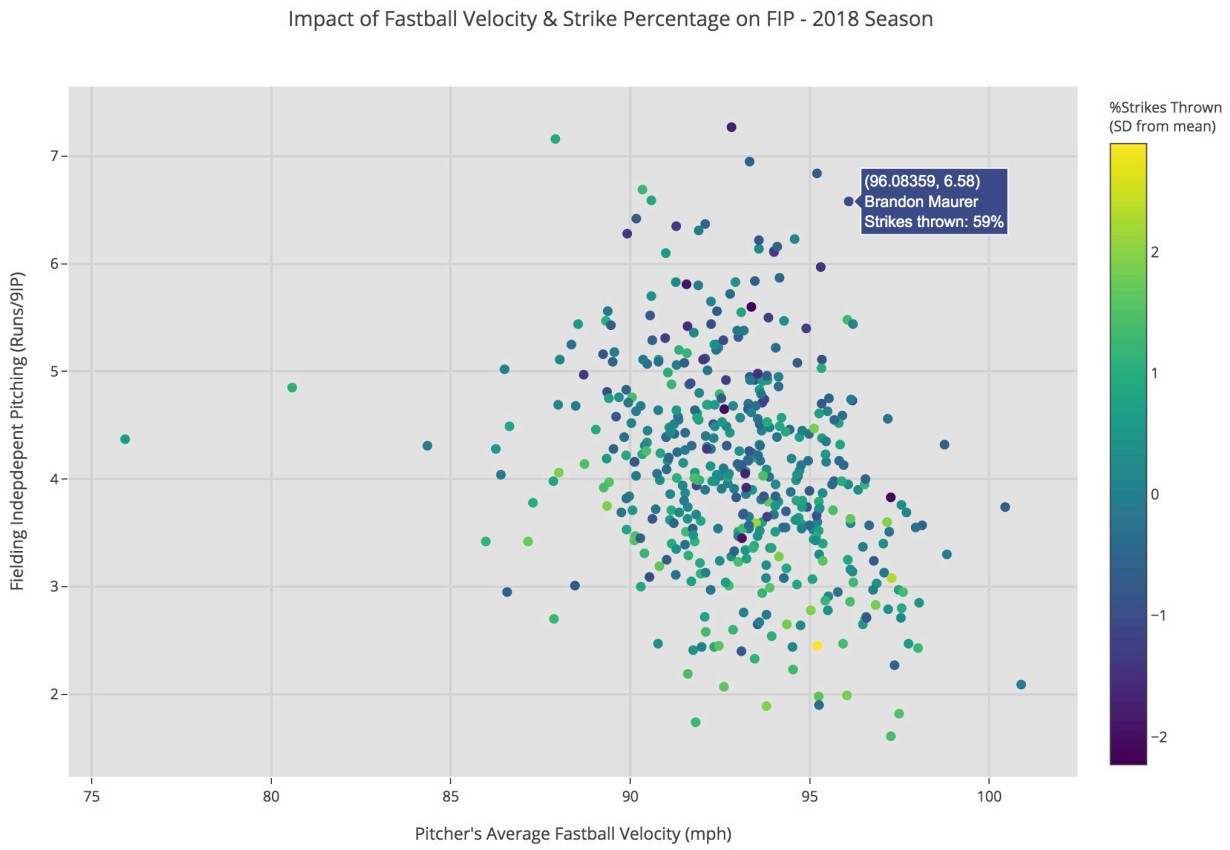
This interactive plot shows which teams played each other in the 2018 MLB season based on the location of their stadiums. It could also reflect the flight path for the players if they were flying direct to the game. Since the MLB has both a National League and an American League, not every team plays each other each year.

Treemap



This plot was not very applicable to my primary research questions, so I decided to look at the breakdown of number of strikeouts for the top 10 strikeout pitchers on my favorite team, the LA Dodgers. From this plot, it's clear that the five main starters on the team account for the majority of team's strikeout total.

Scatterplot (interactive)



https://plot.ly/~ecalkins/102/impact-of-fastball-velocity-strike-percentage-on-fip-2018-season/#/

Here I decided to plot the Fielding Independent Pitching (FIP) statistic with respect to both a pitcher's average fastball velocity and the percentage of strikes they throw. FIP is a somewhat complicated stat, but it can be thought of as a better version of the well-known earned run average (ERA) statistic, which is the average number of earned runs a pitcher gives up over a full game (9 innings pitched). Since small changes in the percentage of strikes a pitcher throws can be important, yet hard to visualize, I decided to standardize the percentage so that the plot can give us a better idea of how pitchers compare to each other.

There are two main takeaways from this plot. Firstly, there is a slight inversely proportional relationship between average fastball velocity and FIP. Additionally, we see that pitchers in the lower FIP ranges (1.5 – 3.5) generally have a lighter green shade indicating that they are above the mean on percent strikes thrown. Looking at the high FIP range (5-7), we see that the majority of pitchers have a below average strike percentage. We will discuss this further in the storyline and conclusion sections.

Heat Map - Storyline

We saw from the scatterplot above that the percentage of strikes thrown correlates highly with a pitcher's success (FIP). That said, the strike zone is nearly four square feet in area, so this only tells us so much in terms of what separates the good pitchers from the bad pitchers (studs from duds). So for my final storyline, I decided to dive deeper into what separates the studs from the duds in terms of their specific pitch locations.

The plot above shows the distribution of pitch locations for good and bad pitchers in four different scenarios: when a lefty pitcher is pitching to a lefty hitter, lefty pitching to a righty, righty vs. lefty, and righty vs. righty. The color denotes the number of pitches and the red box denotes the strike zone. The reason I decided to show eight plots instead of two is because baseball is highly contextual – pitchers pitch differently to righties vs. lefties and there are also natural differences in the way a righty and lefty pitcher throws the ball. Another important point is regarding how I selected the good and bad pitchers. This was done by taking the pitchers in the bottom and top decile, respectively, for the pitching statistic of Fielding Independent Pitching (FIP) for the 2018 season (minimum 50 innings pitched). FIP is an adjusted version of Earned Run Average (ERA), or the average number of earned runs given up in a typical game.

From the above plots, we can see clearly that good pitchers are more willing to challenge hitters over the middle of the plate. From all of the plots, but in particular the L vs. L and L vs. R, we see a higher concentration of pitches thrown over the middle of the plate for stud pitchers than we do for dud pitchers. This is somewhat surprising to me; I expected to see a higher concentration of pitches thrown in the strike zone, but I thought those pitches would be more

toward the corners of the strike zone than in the center. There are two possible explanations for why the stud pitchers are throwing more pitches over the plate. The first is that they have better control of their pitches are therefore missing less (throwing fewer balls when they meant to throw strikes). The second is that they have better "stuff" (baseball jargon for how difficult to hit they are), and can therefore challenge hitters over the middle of the plate more often and get away with it. This will be discussed further in the next section.

Conclusion

In combining our analysis from both the heatmap and scatter plot, the scatter plot results fit best with the second narrative: that good pitchers throw more over the plate because their "stuff" is better and they can get away with it. Pitch velocity is a component of a pitcher's stuff, and from the scatter plot we saw that velocity correlates with a pitcher's FIP. Based on my experience, I think it's a combination of the two factors: good pitchers throw more over the plate both because they are more accurate and because they can get away with it. At the end of the day, in order to be an effective pitcher, you need to throw strikes so that you don't walk hitters, but you also need good enough stuff such that you don't get hit hard when you do throw those strikes.

As a next step in this project, I would continue to explore the affect of pitch features on pitchers' performance. In particular, I would look more into spin rate as it relates to specific breaking pitches and their outcomes.

Github/code

https://github.com/ecalkins/mlb_pitch_data_visualization/

Citations

https://www.kaggle.com/pschale/mlb-pitch-data-20152018
https://fusiontables.google.com/DataSource?docid=1EXApOoxEgJUFlbMjUodfxBSWlRvgQNpABeddHqiN
http://www.thebaseballcube.com/prospects/byYear.asp