FINAL PROJECT WRITEUP: PITCHF/X DATA

USING PITCHF/X DATA TO SORT AND ANALYZE PITCH BREAK AND CALLED STRIKES

PAUL VARJAN, STEVE PELKEY, KENT OWEN

DATA SET AND AREAS OF ANALYSIS

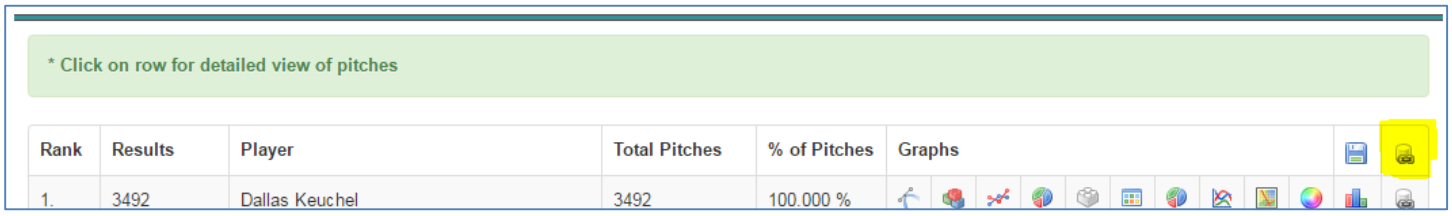
The primary dataset used in this analysis is the Pitchf/x dataset owned and maintained by Major League Baseball Advanced Media, which tracks the speeds and trajectories of every baseball pitched in every Major League Baseball game. All 2015 data were joined with weather data from a separate source for every game. We have two primary areas of interest:

1. **EXPLORATION OF TOTAL PITCH BREAK BASED ON SPIN RATE, VELOCITY, STADIUM ALTITUDE, WEATHER**
2. **EXPLORATION OF HOME FIELD IMPACT ON UMPIRE CALLED STRIKES**

Many of these factors had surprisingly clear relationships. We will provide key takeaways on each of these later on in our writeup.

CORE DATA SET OVERVIEW

Our core data set was downloaded from a public aggregator website called http://baseballsavant.com/. To download an example of the data set, click the following link 2015 Data Set Download Example. Once the data set loads, click the "Export to .csv" button, as highlighted in yellow below:



Additionally, you can reference a Glossary of Fields here: Glossary of Fields & Column Headers in Data. The data set contains information on every pitch thrown, with fields included for:

- Pitch velocity (in x, y, z directions)
- Break on pitch (in x, z directions)
- End location of pitch (in x, z directions)
- End location *Zone* (pre-set bins labeled 1-14 of where ball crossed plate)
- Spin rate on pitch
- Type of pitch (Fastball, Changeup, Slider, Knuckleball, etc.)
- Result of Pitch (Ball, Swinging Strike, Called Strike, Foul, Ball in Play, etc.)
- Pitcher ID and Name
- Time of Pitch
- Top or bottom of inning

Additionally, data can be downloaded and filtered based off stadium the pitch was thrown at.

However, this data set in and of itself was not sufficient to get to all of the different analysis that we wanted to get to. In order to do this, we also needed to separately merge *Stadium Altitude* data and *Weather* data from various other publically-available sources.

ADDITIONAL DATA SET OVERVIEW – DATA WRANGLING PROCESS AND CHALLENGES

Stadium elevation was copied into a csv file from http://baseballjudgments.tripod.com/id62.html. Data integrity was checked using elevation maps from the US Geological Survey http://nationalmap.gov/elevation.html. The elevation CSV was joined to the pitchf/x dataset using the stadium column.

The weather data work on this project conjured up all the horror stories of data wrangling at one point or another. For starters, we looked to a number of verified clean sources for the data we needed, but none offered the ability to pull hourly temperature, humidity and barometer data for all 30 relevant locations across a 6+ month period. We finally found such a resource in the mesowest.utah.edu site. However, the site had a query tool that required weather station ID. In order to find weather station, IDs we researched zip codes of each park on Google and used a tool within mesowest to find the nearest reliable stations for each park. At this point we also captured the altitudes and created a small location dataframe with team name, weather stationID, altitude and if the park had a roof.

In retrospect, we could have forced the weather station IDs into the Mesowest API and programmatically concatenated the resulting data, but we actually just manually pasted each of the 30 station IDs in and pulled down a .csv file, which was copy/pasted into Atom. Yet another problem arose when we realized the 30th weather station was located in Toronto - a city which was not included in the Mesowest data set. For Toronto we had to pull in monthly data, which had all the information we needed...but in metric measurements. Ultimately we did not convert and massage the Toronto data and wound up using data from the other 29 weather stations located in the US, but here again was another complication in accumulating, aggregating and harmonizing open source data.

Once all raw weather data was collated into one single file, we set about the task of reading the .csv into a jupyter notebook. Here we encountered our first need for manual intervention - one of the 29 technicians of the weather stations used multiple commas in the text of their remarks for about a 6 week period. The commas were interfering with pandas read_csv function, so we had to identify all rows with extraneous commas, go into the text editor, and manually remove the offending characters. The field in question contained important location information so we had to be careful not to wipe out useful data in the process.

With the unexpected commas dealt with, pandas was able to read in the data. We were able to eliminate a number of columns with irrelevant data such as visibility and dew point, as well as columns with virtually no data across the set. Within each relevant time series (temperature, humidity, barometric pressure) we finally caught somewhat of a break. There were few gaps or outlier values (there were humidity readings over 100% and barometer readings close to 300 but they were easily repaired). Once we removed outliers and identified that there were no gaps of over a few hours, we wrote cleaning functions that filled in NaN values with interpolated values.

Finally we created a single time stamp out of the various time fields including time zone. In the pitch data, we reasoned that the single pitch time ID field (inscrutably called 'tfs_zulu') was in GMT, so once we collated the time stamp data, we used a dictionary to adjust the discrete weather capture times back to GMT. Our basic clean weather set was complete.

The last piece of work related to cleaning and conditioning weather and location data series was to incorporate the location data into the weather. Some weather research led us to the fact that reported barometric readings are adjusted to sea level, so we used a first order approximation to calculate a local barometric reading for each site.

The most important, yet most difficult piece of the puzzle proved to be fuzzy matching each element of the pitch dataframe to the nearest location and time stamped weather data for that pitch. The pandas asof function turned out to be less useful than advertised, as it was not consistently matching times up correctly. We suspect small disparities in the date formats between the weather dataframe and pitch dataframe, but were never able to pinpoint the exact issue. Additionally, mapping data against one another by hour created unwanted duplicate entries for hours where more than one weather measurement had been taken. However, an “eleventh hour” stroke of insight caused us to instead take the weather data, groupby team and take mean measurements for each hour. Finally we were able to collate the pitch and weather data into a single data frame. From here we were able to use the strikezone heat maps to perform weather driven analysis.

1) EXPLORATION OF TOTAL PITCH BREAK BASED ON SPIN RATE, VELOCITY, STADIUM ALTITUDE

SPIN RATE AND TOTAL BREAK OVERVIEW AND IMPORTANCE

In professional baseball, pitchers spin the ball at delivery to alter the overall trajectory of the pitch. The change in trajectory is determined mainly by the Magnus effect – a fundamental principle of aerodynamics. This phenomenon states that the rotation of any cylinder or ball causes an angular deflection of the surrounding air depending on the type of spin (e.g. backspin, topspin, and sidespin). Therefore, pitchers vary the spin rate, angle, and overall velocity of the ball to produce over 10 unique pitch types (Fastball, Curveball, Changeup, etc.) that are categorized by their speed and movement along the x and z axes. The relative x/z movement of the pitch is called **pitch break**. **Figure A** depicts the relationship between ball spin and break for two pitch types and nicely demonstrates the Magnus effect. The slope differences between curveball and fastball are due differences in velocities of pitch type.

Figure A

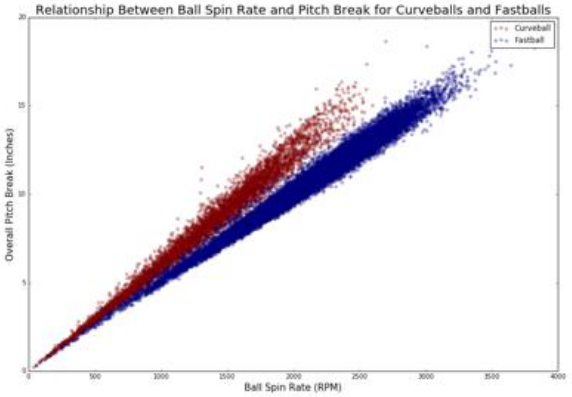
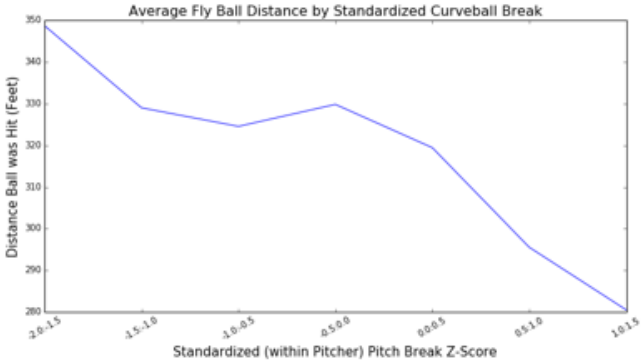


Figure B



Since the Magnus effect depends on the turbulent wake created in the airflow surrounding a ball, it logically follows that the deflective force of the air might significantly vary by meteorological factors such as air density, temperature, or humidity. We will be further exploring this relationship because breaking balls (which are typically slower pitches that do not travel straight as they approach the batter, such as a Curveball) with less break result in good bat contact, demonstrated by **Figure B**. As such, weather may have impact on batting outcomes purely by modifying total break of the pitch.

STADIUM ELEVATION

Of the 30 stadiums in major league baseball, the average elevation is 517 feet above sea level; however, Coors Field in Denver, Colorado is a notable outlier at 5,183 feet. Since higher elevations have considerably thinner air, it is hypothesized that the thin air will exert less force upon the ball causing less pitch break.

It is at first difficult to reasonably infer that the relative lack of pitch break at Coors Field is actually due to altitude rather than other factors unique to Colorado. For example, a potential confounding factor is the frequency at which the Colorado pitching rotation pitches at home. The Colorado Rockies are one of the worst teams in baseball, and have difficulty attracting pitching talent. However, overall correlation between average standardized pitch break for home and away pitching is a very tight .94 making this point moot.

In order to determine the relationship between stadium elevation and pitch break, z-scores were calculated for pitch movement according to pitch type for each unique pitcher. In other words, the break of every pitch is compared to that pitcher's average respective pitch break in an attempt to minimize the effects of between-pitcher variation.

After computing the mean pitch break z-score for each stadium, an inverse relationship with stadium altitude is easily apparent (see **Figure C**). In fact, the average break of a knuckle curve thrown at Coors Field is 1.5 standard deviations below the mean, which

corresponds to roughly 4.4 inches. The overall correlation between stadium elevation and mean pitch break is -.54, but the exact correlation coefficient varies by pitch type (**see figure D**).

Figure C

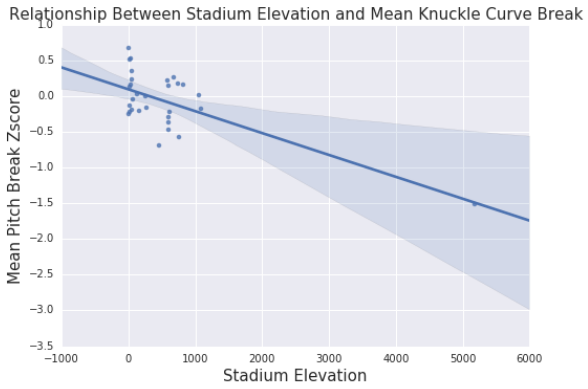
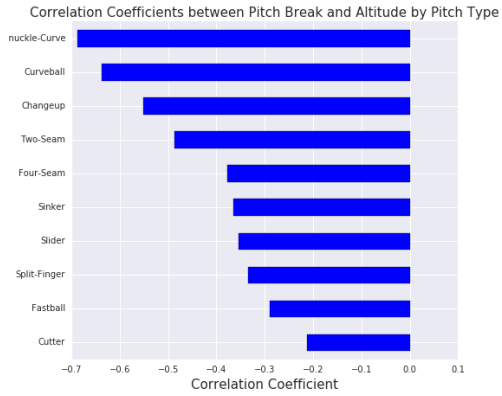


Figure D



WEATHER IMPACT ON PITCH BREAK

Several weather variables were brought into the Pitchf/x dataset, including temperature and barometric pressure at the time of the pitch. Each of the weather variables were appropriately binned in order clarify the relationship with pitch break in our graphs. Barometric pressure is of particular relevance because it is negatively related to elevation, so if our elevation/break hypothesis is correct, we should observe a positive relationship between barometric pressure and pitch break. To our relief, **Figure E** supports our hypothesis, which also suggests that our difficult weather/pitch dataset merge was successful.

Figure F shows a highly correlated negative relationship between temperature and mean pitch break for all the stadiums without roofs. For every ten degree increase in temperature, pitches will break less by about .1 standard deviations, which doesn't sound like much, but it adds up. This makes sense regarding the Magnus effect because the molecules in warm air are more spaced out resulting in less force enacted on the ball.

Figure E

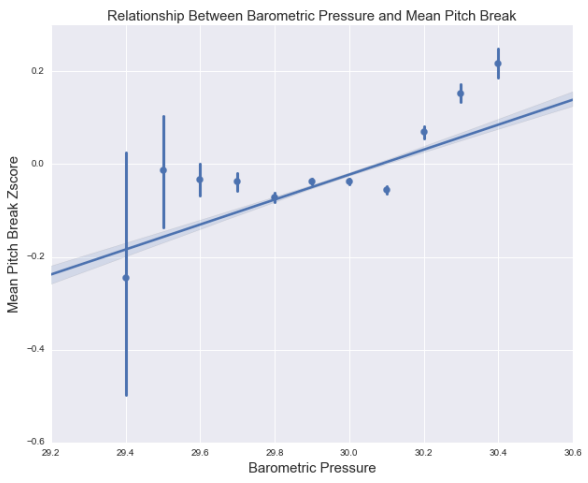
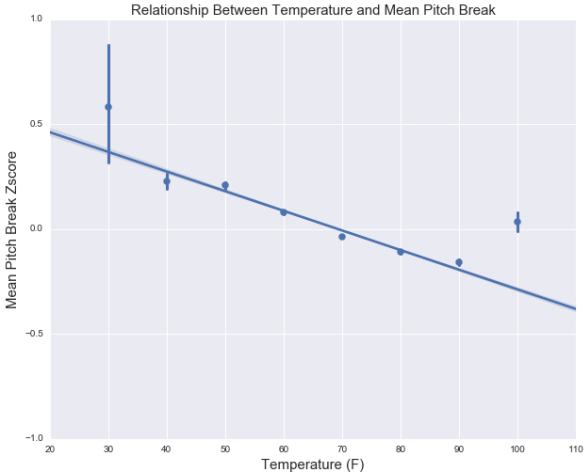


Figure F



2) EXPLORATION OF HOME FIELD IMPACT ON UMPIRE CALLED STRIKES

HOME-FIELD IMPACT ON UMPIRE STRIKE CALLING

In Major League Baseball, there is a distinct advantage to playing at a team's "Home" stadium. In fact, from 1920 onwards, the home team has won 54% of MLB games (http://www.baseball-reference.com/blog/archives/9916), despite having equal talent to the away team (when compared across large enough sample sizes). Though umpires are *trained* to call a consistent game, we have been exploring whether there could actually be a mental factor to calling a baseball game, and whether umpires are influenced to call more strikes by the home crowd. An umpire is meant to call a consistent "Strike Zone" of called balls v. called strikes, but if they are influenced by the crowd or other home factors in any way, we will see material differences in their Called Strike rates.

Figure G (Home Team Pitching)

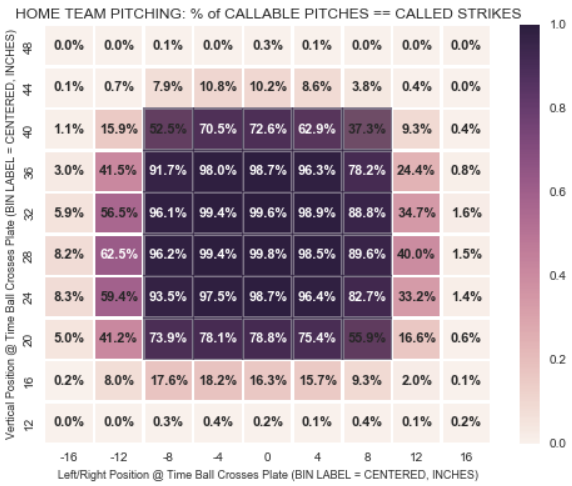


Figure H (Away Team Pitching)

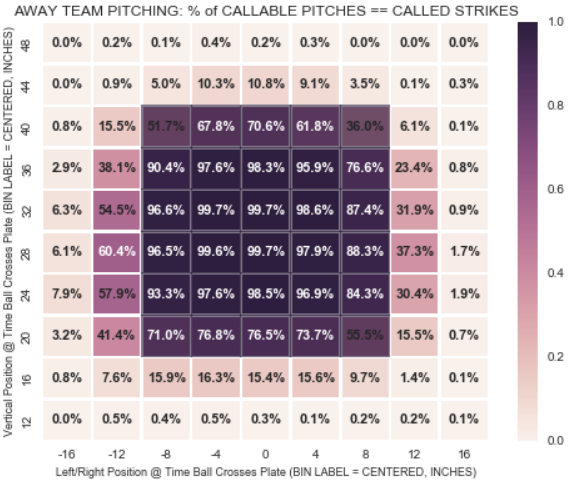
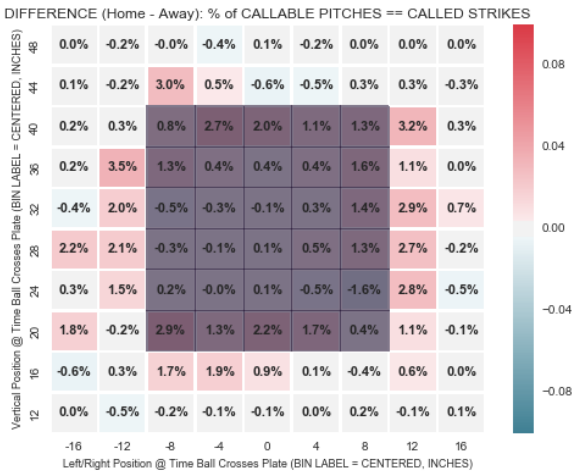


Figure I (Difference between Home & Away)



First, we filtered the data to be only on "Callable" pitches (called balls and called strikes), removing any data where the batter made a swing (appx. 50% of the data). However, we cannot just take the percentage of called strikes for when a home team is pitching v. when an away team is pitching on this subset to derive any meaningful conclusions, as there may be other factors at play. For example, if a pitcher is better at pitching into the strike zone at home compared to when he is away, fully unbiased umpiring would still show a higher percentage of called strikes for the home team. As such, we need to analyze the ***percentage of called strikes by exact location*** as the ball crosses the plate in order to remove for overlapping home/away pitcher performance impact. To do this, we "binned" the location of each pitch by 4x4 inch sections of when the ball crosses the plate, then brought in the rate of called strikes by each location bin.

Note that in each of our charts, we mapped the average strike zone size over top of the bins (dark grey box) for a visual representation of the strike zone. We are also only showing just two extra bins outside of the strike zone, as pitches more than 2 bins (8 inches) outside of the average strike zone are rarely called strikes for either the home or away team, and as such have minimal impact on analysis. See **Figure G (Home Team Pitching), Figure H (Away Team Pitching), and Figure I (Difference between Home and Away Teams)** for a comparison of these bins.

If umpires were uninfluenced by whether a home v. away team is pitching, we would see minimal differences by bin that average near 0.0% **in Figure G**. However, in Figure G, you can visually see more red than blue bins, indicating that the home team gets strikes called at a higher rate. In fact, as a % of total pitches per bin, the home pitcher gets:

- 0.1% more strikes called in the inner bins of the standard strike zone
- 0.6% more strikes called in the outermost bins *within* the standard strike zone
- 1.7% more strikes called *one bin outside* the standard strike zone
- 0.2% more strikes called *two bins outside* the standard strike zone
- 0.1% more strikes called *three or more bins outside* the standard strike zone

Overall, as a % of callable pitches by weighted bin, there are **0.54%** more strikes called for the home pitcher than the away pitcher! As 50.2% of pitches are callable pitches, this equates to 0.27% of pitches in the full data set that that the home team will receive an extra advantage of getting a called strike on. While this may not seem especially material at first, this has a significant impact over the course of a game, and especially over the course of the season. Separate research has shown the value of an additional strike called to be worth ~0.14 runs (http://www.baseballprospectus.com/article.php?articleid=22934). As such, over the course of a typical baseball game of ~290 pitches (http://www.baseball-reference.com/blog/archives/7533), the impact of an umpire is approximately 290 pitches/game*0.27% difference in called strikes per pitch*0.14 run value per additional called strike = **0.11** runs per game. As such, umpire impact in itself is enough to give a legitimate home field advantage! Extrapolating this further, over the course of a 162-game season, a team can expect to score **18** more runs playing at home than playing away purely due to intentional or unintentional umpire bias.

Furthermore, we wanted to explore when the pressure and importance increases on a callable pitch whether the umpire actually gives an even greater bias towards the home team. We did this by exploring the subset of when there were two strikes in the count; meaning one more strike and the batter strikes out.

When there is increased pressure and importance on the call, umpires potentially get *slightly* more biased (**Figures J-L**). As a % of total pitches per bin, with two strikes in the count the home pitcher gets:

- 1.2% more strikes called in the inner bins of the standard strike zone
- 1.4% more strikes called in the outermost bins *within* the standard strike zone
- -0.1% more strikes called *one bin outside* the standard strike zone
- 0.1% more strikes called *two bins outside* the standard strike zone
- 0.0% more strikes called *three or more bins outside* the standard strike zone

Overall, as a % of callable pitches, the home team gets a 0.61% advantage of called strikes when there are already two strikes in the count. This is slightly more than the 0.54% rate advantage on all strike counts. However, because filtering on two strikes only gives us a smaller sample size that the full data set, and because the difference between 0.61% and 0.54% is only +0.07%, the data suggests that there is *likely* an extra advantage received by home pitchers on a two strike count, but we want to hold off on concluding this *definitely*.

Nonetheless, we can definitively conclude from our analysis that there is an overall umpire bias, whether intentional or unintentional, towards the home team pitcher.

Figure J (Home Team Pitching, 2 STRIKES)

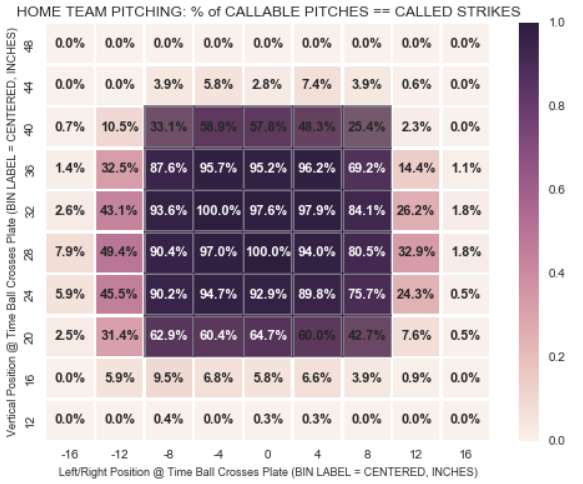


Figure K (Away Team Pitching, 2 STRIKES)

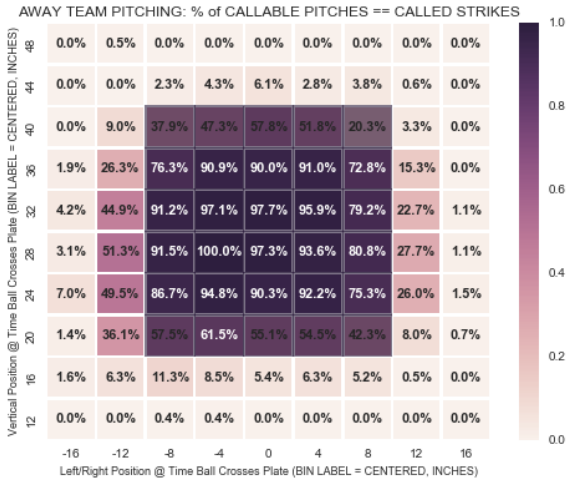


Figure L (Difference between Home and Away, 2 STRIKES)