Springboard - DSCT
Capstone Project 2 - Milestone Report 1

David Buehler

December 2019

At this point in my project, things have gone well. I have set out to get a good idea on the best way to predict pitches from MLB pitchers and what factors are the best in determining what pitch is going to come next. Some people that might be interested in this kind of analysis would be MLB coaches and managers. They're always trying to get a leg up on the competition, so if there can be a model built that could predict pitches with any form of accuracy, they will be very interested in getting all the information on the opposing pitcher they can get. Hitters also watch a lot of film on the pitchers they'll be facing, so they'd be interested in knowing the results the models I am building would predict.

I pulled the data from Kaggle, and it all came really clean already. There was very little cleaning I had to do to get it in a good shape for analysis and visualization. What I did do though was a couple things, first I noticed there were a few pitches that occured with 4 balls in the count, and since you can have no more than 3 balls in the count before a ball is thrown, those needed to be dropped. Next I wanted to know what pitch in the at-bat each and every pitch was, so I looped through every pitch and was able to do just that. I added that as a column to each data frame and then I needed to group pitches into 4 groups: fastballs, breaking balls,

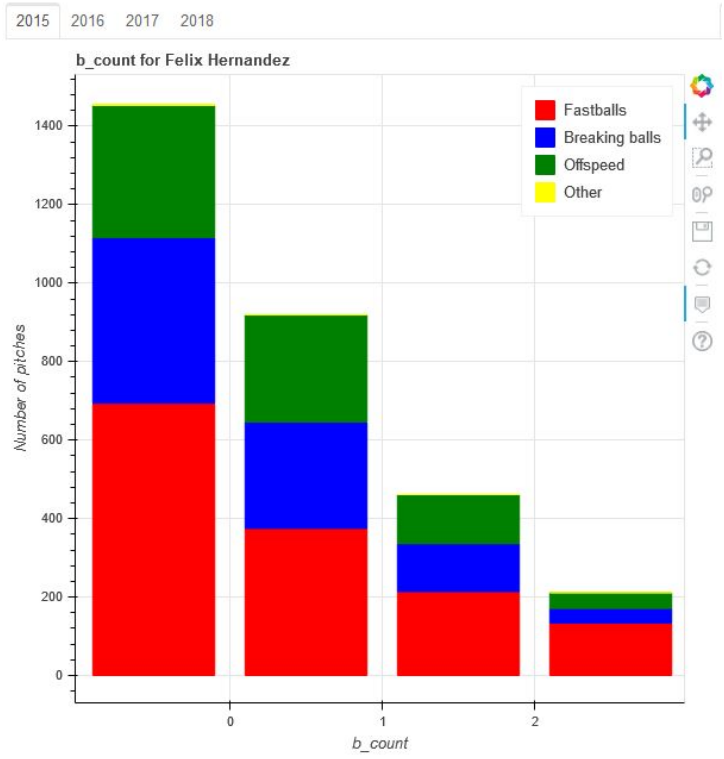
offspeed pitches, and other pitches. Fortunately Pandas has an easy function to do that, and was able to accomplish that with relative ease. Finally, I wanted to know what pitcher was throwing each pitch, and the Kaggle site I got my data from had another csv file with each player ID and their corresponding first and last name. With a simple merge call, I made two new columns with a pitcher's first and last name in separate columns and my data wrangling was done after putting the wrangled data frames into CSV files to use next.

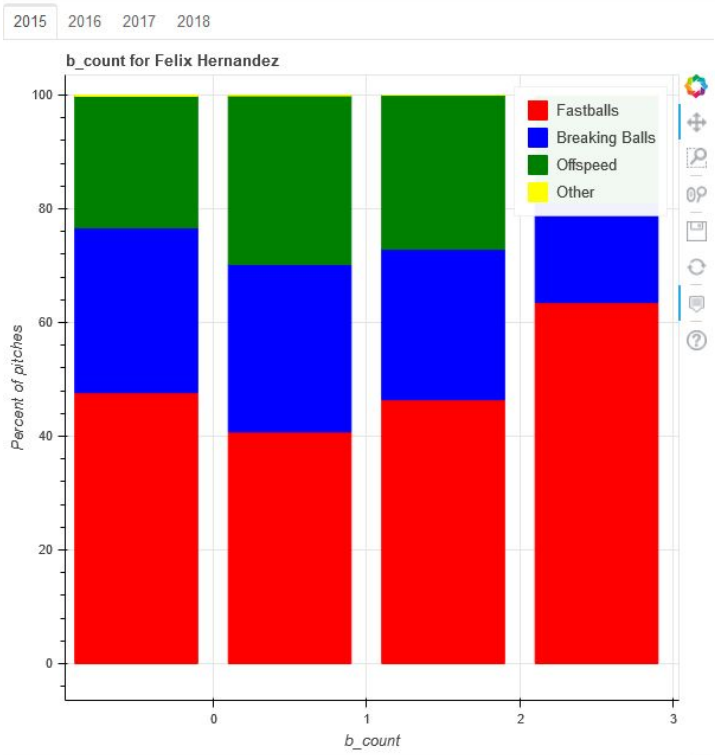
Now it was time for the data storytelling, and this was where I was able to get deep into the data. I wanted to use the Bokeh library because I knew that could be a powerful tool in data visualization, and I was able to do so after coming to grips on what data I wanted to show. First I needed to create a couple functions that would pull out the needed data. I was able to pull out how many fastballs, breaking balls, offspeed and other pitches that pitchers threw based on the variables in the data frames I'm working with. For example. I used a function that would return a data frame that tells me how many, and the percentage of fastballs/breaking balls/offspeed/other pitches a certain pitcher threw with so many balls in the count, broken up by year. An example of this would be this data frame for Felix Hernandez:

	count	year	number_of_FB	number_of_BB	number_of_OS	number_of_OT	percent_FB	percent_BB	percent_OS	percent_OT
0	0	2015	694	422	337	3	47.664835	28.983516	23.145604	0.206044
1	1	2015	375	271	273	1	40.760870	29.456522	29.673913	0.108696
2	2	2015	214	122	125	0	46.420824	26.464208	27.114967	0.000000
3	3	2015	134	37	40	0	63.507109	17.535545	18.957346	0.000000
4	0	2016	592	295	211	0	53.916211	26.867031	19.216758	0.000000
5	1	2016	313	201	211	0	43.172414	27.724138	29.103448	0.000000
6	2	2016	195	100	138	0	45.034642	23.094688	31.870670	0.000000
7	3	2016	124	33	40	0	62.944162	16.751269	20.304569	0.000000
8	0	2017	298	215	140	0	45.635528	32.924962	21.439510	0.000000
9	1	2017	165	115	121	0	41.147132	28.678304	30.174564	0.000000
10	2	2017	86	59	71	0	39.814815	27.314815	32.870370	0.000000
11	3	2017	62	23	26	0	55.855856	20.720721	23.423423	0.000000
12	0	2018	503	411	264	2	42.627119	34.830508	22.372881	0.169492
13	1	2018	272	232	213	1	37.883008	32.311978	29.665738	0.139276
14	2	2018	185	132	113	0	43.023256	30.697674	26.279070	0.000000
15	3	2018	138	67	28	0	59.227468	28.755365	12.017167	0.000000

This data frame would then get used to make a

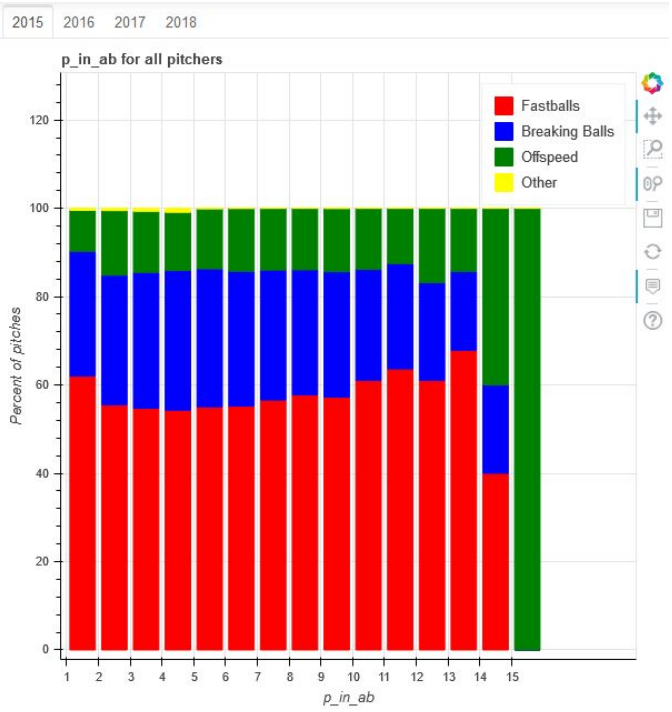
plot of this data using bokeh, and would turn out like this:





The top graph being the total number of types of pitches thrown with each number of balls in the count, and the bottom being the percentage of types of pitches thrown. The percentage graph usually gives more information, as trends are able to be seen better than looking at the total number of pitches.

Those graphs were made using the function that took in fixed number variables (ball count, strike count, batter side etc.), I was also able to put together a function that took in variable number features and plot them for me, like innings and what the pitch in the at-bat is. Unfortunately I wasn't able to look at this for specific pitchers, so it's looking at every pitcher in the data frames. The percentage pitch in at-bat graph looks like the graph below:



The graph shown is only for 2015, but some trends were noticed as the years went on. First pitch fastball occurs 62% of the time in 2015, and actually decreases over time. Down to 60% in 2018. It's not a huge variation but certainly notable. Breaking balls have seem to

become more prevalent as first pitches to batters as time has gone on, rising from 28% in 2015 to 32% in 2018. Pitchers seem to agree that the best time for an off-speed pitch, most commonly a changeup, is the 2nd pitch of an at bat, and has stayed at around 14% of all types of pitches. These were just a few things that were noticed in an otherwise information filled graph.

Next I wanted to take a look at some Cy Young award winners, and see how their pitches stacked up against the rest of the MLB. Starting in 2015, Jake Arrieta of the Chicago Cubs won the National League Cy Young award in 2015, and Dallas Keuchel of the Houston Astros who won the American League Cy Young award in 2015. These two pitchers were at the top of their game in 2015, and the data showed why. I also looked at the 2018 Cy Young award winners and not the 2016 or 2017 winners as I didn't want the document to get too long and drag on. Finally, I thought it would be interesting to look at the pitches thrown with the number of runners on base, so I was able to make a column in the data frame that had the total number of runners on base, but not where they were, and made my bokeh plots on those.

Now time for the inferential statistics portion. This data doesn't come with too many continuous variables, mainly batter score and pitcher score, almost all the rest is categorical. Looking at batter score and pitcher score, I thought it would be interesting to see how significant the score differential was on the type of pitch thrown. Creating the score differential column was easy enough, then I created a function that makes the pitch type I want to look at a 1, then all the other pitches a 0 in the pitch_type column of the data frame. Then created a bootstrap test function, that takes the target pitch type as a variable, and created histograms with it, and performed a p-test from scipy's ttest_ind function. I operated under the null

hypothesis of: “On average, pitchers will throw the same amount of fastballs/breaking balls/offspeed pitches no matter the score differential" and performed the bootstrap test at a 95% confidence interval. Testing all four pitch types, all of them came under the 0.05 threshold, so I knew these results were statistically significant and we reject our null hypothesis. While doing this analysis, something really interesting came up while looking at the “other” pitch type. A lot of other pitches, which include pitch outs that usually lead to intentional walks, were used where the score differential was around 1. Which means pitchers were intentionally walking hitters that could either tie the game, or put the hitting team up with one swing of the bat, and I found that to be really interesting.

Next I wanted to take a look at how different variables affected what pitch was thrown using statsmodels’ logit function. This would perform a very basic logistic regression formula, but would only look at the summary from the fit() call and not make any predictions. First up was ball count and strike count for each type of pitch. For fastballs, as ball count goes up, it’s more likely that a pitcher would throw a fastball, and as the strike count went up it was less likely that a fastball would be thrown. Next up was breaking balls, which came with some information that hasn’t matched any trends seen up to this point. As both the ball and strike count went up, pitchers were less likely to throw a breaking ball. However the graphs and the data have consistently showed that as strike count goes up, breaking balls are more likely to be thrown. An interesting discrepancy there for sure. Next up came runners on, as that was only looked at briefly in the data storytelling. Statsmodels confirmed what the graph showed. As more runners are on, fastball usage is likely to go up, and breaking balls and offspeed usage are likely to go down.