Hailey Yabroudy
Project 3 Part 4
December 15, 2020

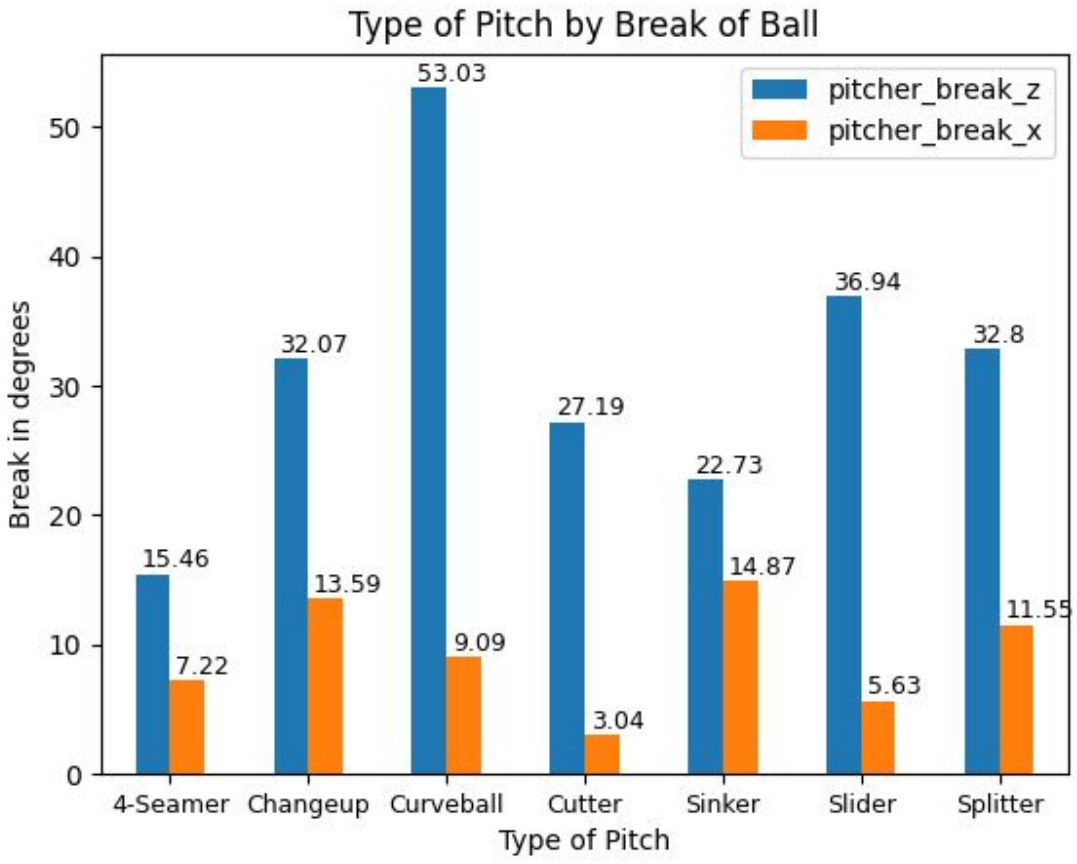
Statcast Project for Data Interoperability

For my project, I wanted to be able to work with the data that I would potentially be working with in future jobs. My dream career is to be a baseball statistics analyst working for a team or the MLB itself. This past summer, I was supposed to have an internship in New York City with the MLB and from what my contacts there have told me, I would have spent a majority of the internship working with a website called baseball savant. This has basically every piece of data that you could think of for baseball. They have data available on specific players and teams as a whole. I got the data that I chose to work with directly from the website after doing a little searching and was able to just download it as a csv file. The specific data is about all of the pitchers in the MLB and which of their pitches moves the most. It also covers how much the pitch moves vertically (z break) and horizontally (x break). The data also compares the individual pitchers numbers to the league average and the differences, their average speed, what team they're on, how many pitches they've thrown in total, etc. For this project, I only focused on the pitch type and it's break. I wanted to work with pitchers data rather than hitting data because I myself am a pitcher and I personally find it more interesting information. I also focused on break rather than velocity because velocity can only get you so far as a pitcher, having spin and break on the ball can make a pitch unhittable.

Part one of my project was fairly simple. My goal was to convert the csv file into an html file and it was pretty straightforward. I opened the csv, read it in, and then used pandas to convert it into an html file. Having practice doing this throughout the year was very helpful and while in the past, I have had issues with getting the data to show up in the new html file, this one showed up on my first try and looks good.

For part two of my project, I wanted to make a bar chart showing the z break and x break for each pitch type. To do this, I would need to average z and x break for each pitch. I started by following what was done in the activities from week 9 but it became quickly evident to me that this would be somewhat hard to make work with the type of data I was using and I wanted to find another solution to do it. I then tried to put all of the values into separate lists that I could then put together into the chart, but this way felt very tedious and I knew that there had to be a better way to do what I wanted. I then remembered our work with pandas and the groupby function. This made it very easy to get each pitch type data and be able to find the average for

each pitch. The new problem I thought would be getting that into the form of a list or dictionary and be able to create the graph from that. After a lot of research and trial and error, I found that I could simply just add ".apply(list)" to the end of my groupby line of code and then just use that to create the bar graph. I wouldn't need to convert it from there or anything like that. My groupby statement was called "movement" and I was able to create the bar chart with "movement.plot.bar()". This made my code look a lot cleaner and streamline the data. From there, I just added the y label, x label, and a title. I liked how the rest of the chart looked, so I didn't need to change too much. I was able to learn an easier and simplistic way to create a bar chart with 3 variables in a way that I liked more than the ways we learned in class. Below is the chart that I made:



Hailey Yabroudy
Project 3 Part 4
December 15, 2020

After finishing part two, I added some more questions that I wanted to answer for part three. I initially just wanted to find the average break for each pitch type and how much break left and right handed pitchers get. Since I had to find the averages in order to make the chart, half of my analysis was done. I wanted to challenge myself and be able to find out more information from this data. So, I decided to work on more of my questions to answer:

1. What is the average z and x break for each pitch type? (Here I wanted to just print it in written form so there are numbers to go with the chart)
2. Which pitch moves the most vertically and which moves the most horizontally?
3. Do right handed or left handed pitchers get more movement on their pitches?
4. Which pitch is most popular for left and right handed pitchers?
5. Who has had the most movement vertically and who has had the most movement horizontally?

I started by quickly getting the first question out of the way and just printing the results I've already found. Questions two and three were also fairly simple because I was able to just tweak the line of code I used for question one. The results for question one can be seen in the chart above. For question two, the curveball moves the most vertically with 53 degrees of break and the sinker moves the most horizontally with 14 degrees of break. For question 3, I found that left handed pitchers on average get more movement vertically and horizontally.

Question four was the hardest one for me to figure out. After messing around with the .groupby() part for a while, I was able to get it to display all of the information that I was looking for but I wanted there to be two separate lists for the right handed pitchers and the left handed pitchers. This is where the trouble started. I tried .sort_values(), if/else statements, for loops, and so many other ways that I can't even remember them all. For some reason, it wouldn't identify the difference between "L" and "R" within the column. After many, many, many, MANY google searches, I was able to find a solution that used a filter method. Using this solution, I created two new data frames, one with only the left handed pitchers and the other with only the right handed pitchers. From there, I would just apply that new data frame to my .groupby() statement, and I was able to successfully make the two lists I wanted in the format I wanted. I had never used a filtering method before (at least not a time I can remember), and I'm excited to use it again in the future because of how simple and easy it is to use. The results I found were that most left handed and right handed pitchers get the most movement on their 4-seamers. This means that for most pitchers in the MLB, this is the pitch that they can get the most movement on, but that doesn't necessarily mean it is the most effective. The list for left handed pitchers for most popular to least is:

Hailey Yabroudy
Project 3 Part 4
December 15, 2020

Left Handed:		Right Handed:	
4-seamer	82	4-seamer	241
Sinker	44	Slider	151
Slider	41	Sinker	111
Changeup	40	Curveball	84
Curveball	28	Changeup	63
Cutter	21	Cutter	48
Splitter	1	Splitter	16

For question five, I took a slightly different approach. So instead of using .groupby(), I sorted the csv based on the z and x break in descending order, so it would show the highest values at the top. I then had to figure out how to show the specific details I wanted to know rather than the entire row. I decided to show the first name, last name, the team they play for, and what the break value was. This part helped me be able to better understand how to show specific values in a table and a little more of the mechanics around using .iloc(). I found that Alec Mills with the Cubs got the most vertical movement with his curveball and he got 72.5 degrees of break. For the most horizontal movement, Aaron Loup with the Rays got 19.7 degrees of break on his sinker.

Please note, I could only upload a single file. I would have loved to have been able to upload a zip file with all of my python files showing my work and the csv file that I was working with. If you would like to see any of this, I would be happy to send it.