Homework 3: Python and Web Scraper

David Fuentes (dmf4ns) and Dylan Howe (dsh7pd)
Due: September 30th, 2020

We decided to extend the Baseball Reference analysis we performed for the Module 4.7: NumPy and Pandas assignment, in which we manually downloaded batting data from the current 2020 season and performed some basic querying to determine the league leaders in hitting (as of September 15th), which batters played for the Yankees, and who was in the oldest and youngest 10th percentiles of batters.

We knew there was more that we could do with better access to the Baseball Reference database, and building a webscraper to access these data in an automated fashion seemed perfect for our needs. From more seasons' worth of data, we could perform time-series analyses, look at players by their comparable-age seasons, and extend our function to pull pitching data we could analyze in addition. So, we produced a webscraper that pulls data from the site for some n number of years until a specific year given by the user via a loop.

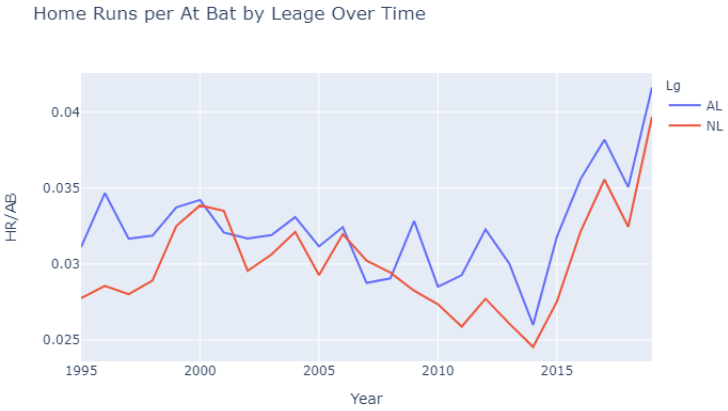
In order to do this, we implemented Selenium's WebDriver to run the crawler, BeautifulSoup to parse the data on the page and pull the data tables, Pandas to hold our data and perform any necessary cleansing/calculations, and finally, Plotly Express to produce our graphs. We imported some additional packages, like Time, to provide some useful output to the user.

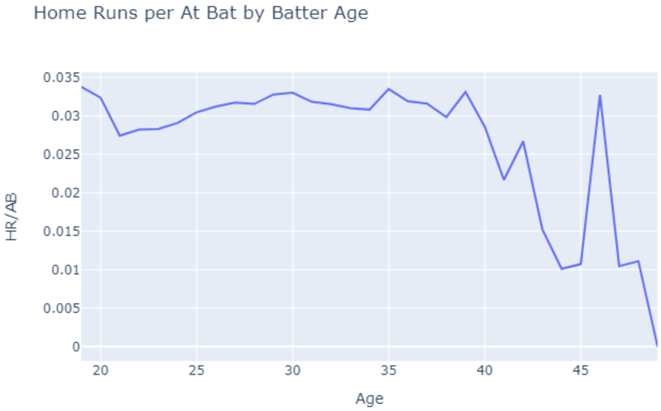
Our scaper function takes two parameters: the final year whose data we would like to pull, and the number of seasons' worth of data we are interested in pulling. These default to 2020 and 10, respectively, so running the function without parameters would return data from 2011-2020, for example. The function produces two .csv files – one containing year-over-year pitching statistics, and one containing year-over-year batting stats – and returns a tuple of the file names. We decided to scrape hitting and pitching data from 1995 through 2019 – the 2020 season is shortened due to COVID, so we decided not to include these stats – in order to run our analyses. The program took about 11 minutes to pull 55k rows of data for 25 seasons

The past 25 years are rather important as far as Major League Baseball is concerned. In the late 1990s, the MLB was at the height of the "Steroid Era", which was a period of time during which regulations on Performance Enhancing Drugs (PEDs) was very lax. Players like Mark McGwire and Barry Bonds were breaking long-standing homerun records, and homeruns were up all over the league. Eventually, the MLB cracked down, aided by The Mitchell Report and Congressional support, and homeruns tapered off for a few seasons.

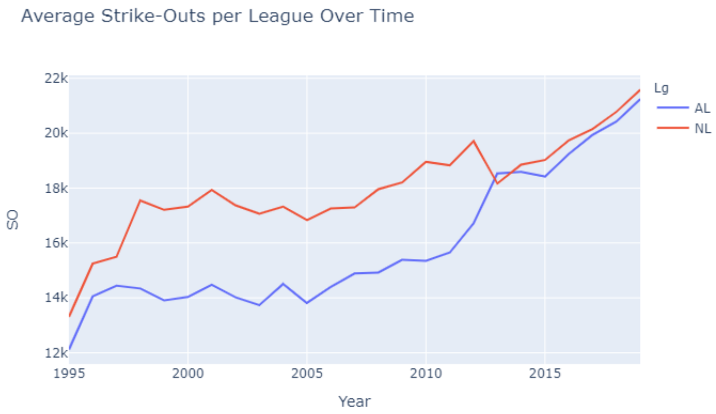
With the end of this era came a new one, but this time, instead of teams employing so many burly sluggers, they began to employ more statisticians and general data experts. Over time, teams realized that walks and homeruns were highly valuable; homeruns in particular. This sounds obvious, but teams, coaches, and players had valued events like singles (worth 1 base versus 4 for a homerun) and their corresponding high batting averages for more than a century at this point. Players over the past several years have completely revamped their swings to prioritize power (homeruns) over contact (singles and a high batting average).

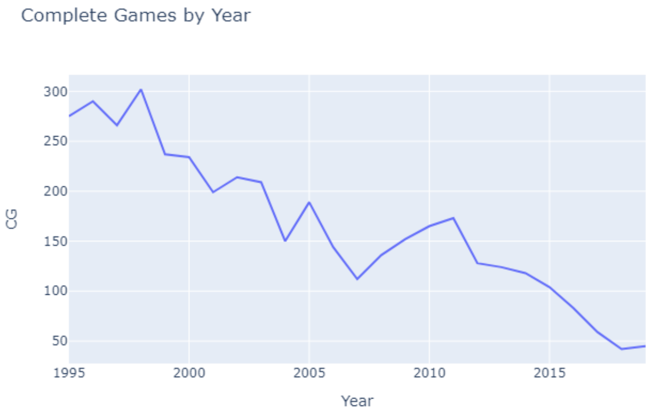
To see if the data supported this, we produced a Homeruns per At Bat statistic in Pandas – think of this as a homerun-frequency ratio that can be used to control for things like differences in total at bats across players. This ratio hit a peak in the late 1990s before dropping due to stricter steroid regulation before once again spiking to new all-time highs. We also noticed a pretty stark gap between the two leagues (AL and NL). This would make sense for a baseball fan – pitchers hit in the NL whereas they do not in the AL, and pitchers are not very good at hitting. We also see a small drop in 2018 – there were rumors that the MLB changed baseballs to try to curb the increase in homers, but the league vehemently denied this. We also show HR/AB by player age just to round things out – the spikes for older players likely exist due to one or two older outliers with very few at bats.





In order to combat this power surge, teams countered by producing pitchers who threw harder and had better movement on their pitches. Since batters were prioritizing power over contact and pitchers were prioritizing speed and movement, strikeouts surged. However, with this uptick in pitch speed and movement came more stress on pitchers' bodies, which led to more injuries. To protect their pitchers and their investments in them, MLB teams made them pitch less, which can be seen in the second graph of complete games (one pitcher pitching an entire game).



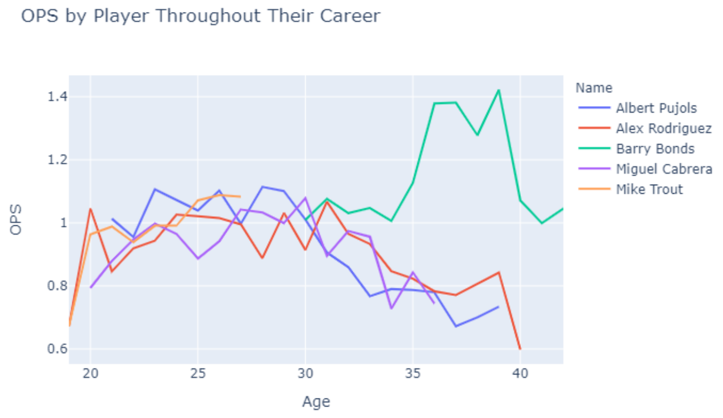


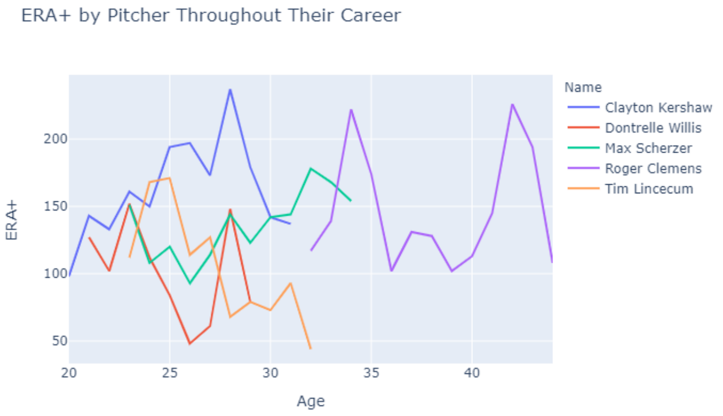
In addition to league-wide trends, we wanted to compare historically-good players against each other. Since players' production typically reaches a peak around their late 20s and trends down for the rest of their career, we wanted to compare these players by their age rather than any particular season – most of these players didn't play at the same time anyway.

The first graph shows the batter comparison using a statistic OPS, which is a holistic measure of power and contact. You can see these players hit their peak sometime in their mid-to-late 20s before their production drops off – all except for one. Barry Bonds played during the Steroid Era and has long been accused of using PEDs. Mike Trout is still playing, which is why his line ends before the rest.

We performed a similar analysis for pitchers using ERA+, which is an analogous all-encompassing statistic for pitchers. We see that these elite pitchers peaked sooner than their batter counterparts and didn't have the same career longevity – though Max Scherzer is still playing and is still enjoying some success relatively late into his career.

Notice again one player bucking this trend – Roger Clemens is another player long accused of using PEDs. His career started before 1995, which is why his line starts in his 30s.


OPS by Player Throughout Their Career


ERA+ by Pitcher Throughout Their Career

Someone interested in baseball data would find our program useful, and we can extend it to pull more Baseball Reference advanced statistics. We could also point the program to Baseball Reference's siter sites, which house NFL, NBA, and NHL data.

We think much of what we did is extra functionality beyond the given requirements, particularly the loop to pull time series data and the code to graph and compare those data.

Sources

Data:
Batting Table Example (for 2019):
https://www.baseball-reference.com/leagues/MLB/2019-standard-batting.shtml
Pitching Table Example (for
2019):https://www.baseball-reference.com/leagues/MLB/2019-standard-pitching.shtml

Python Documentation:
Selenium: https://selenium-python.readthedocs.io/
Beautiful Soup: https://www.crummy.com/software/BeautifulSoup/bs4/doc/
Pandas: https://pandas.pydata.org/docs/ plus the McKinney Text
Plotly Express: https://plotly.com/python/plotly-express/