Deliverable #3: Initial Data Exploration

Tate Huffman

10/29/2021

Overview

The data collected for this thesis was obtained from Baseball Reference, a website that contains historical data for minor and major league baseball. The data was scraped using Python, the code for which can be found in my GitHub repository. It contains performance data in minor and major league baseball from 1998 through 2019 and is separated into batting and pitching data.

Format

The data was originally separated into thousands of different files by team, based on batting/pitching, year, level of baseball (e.g., Triple-A, MLB, etc.), and organization (i.e., the major league club of that baseball team). The data runs from 1998 through 2019, as 1998 was the first year in which MLB expanded to its current size of 30 organizations, and 2019 was the last full year of data available unaffected by the COVID-19 pandemic.

After scraping this raw data, I eliminated some extraneous columns from each file, as data from minor league teams did not match the format of major league teams, and then combined them into two long files, one for batting data and another for pitching. These files contain statistics describing playing time, player performance, and team played for, with a unique row for each player at each level for each team in each year. This means that if somebody plays at multiple levels in the same year, or changes organizations, they will have multiple entries in a given year.

Data

Disclaimer - this is significantly less formal than it will appear in the final paper.

Table 1: Means and Standard Deviations of Batter Statistics ($n = 23,210$, $n_{min} = 19,367$)

	No PA Minimum		PA >= 20	
Measure	M	SD	M	SD
Years Played	3.65	3.07	3.78	3.22
Organizations	1.61	1.31	1.62	1.37
Levels	2.64	1.64	2.85	1.69
PA per Year	254.92	190.32	293.92	175.17

Note - for PA minimum, player season only included if PA >= 20.

Above are the summary statistics for general player information, including the number of years played, organizations played for, and plate appearances per year. There are two different sets of statistics for each of these: one for the raw data, and one including only those player seasons where the hitter had at least twenty

plate appearances, in order to exclude years with injuries or instances where pitchers had plate appearances. We see that when we exclude those seasons, we have about 4,000 fewer players in the data.

From this table, we see that there is a wide range in key statistics across the player pool. Notably, the standard deviation for years played is nearly as large as the mean, indicating here that the distribution of years played is right-skewed, as would be expected in a population where many players have short careers but a select few play at the major league level for an extended period of time.

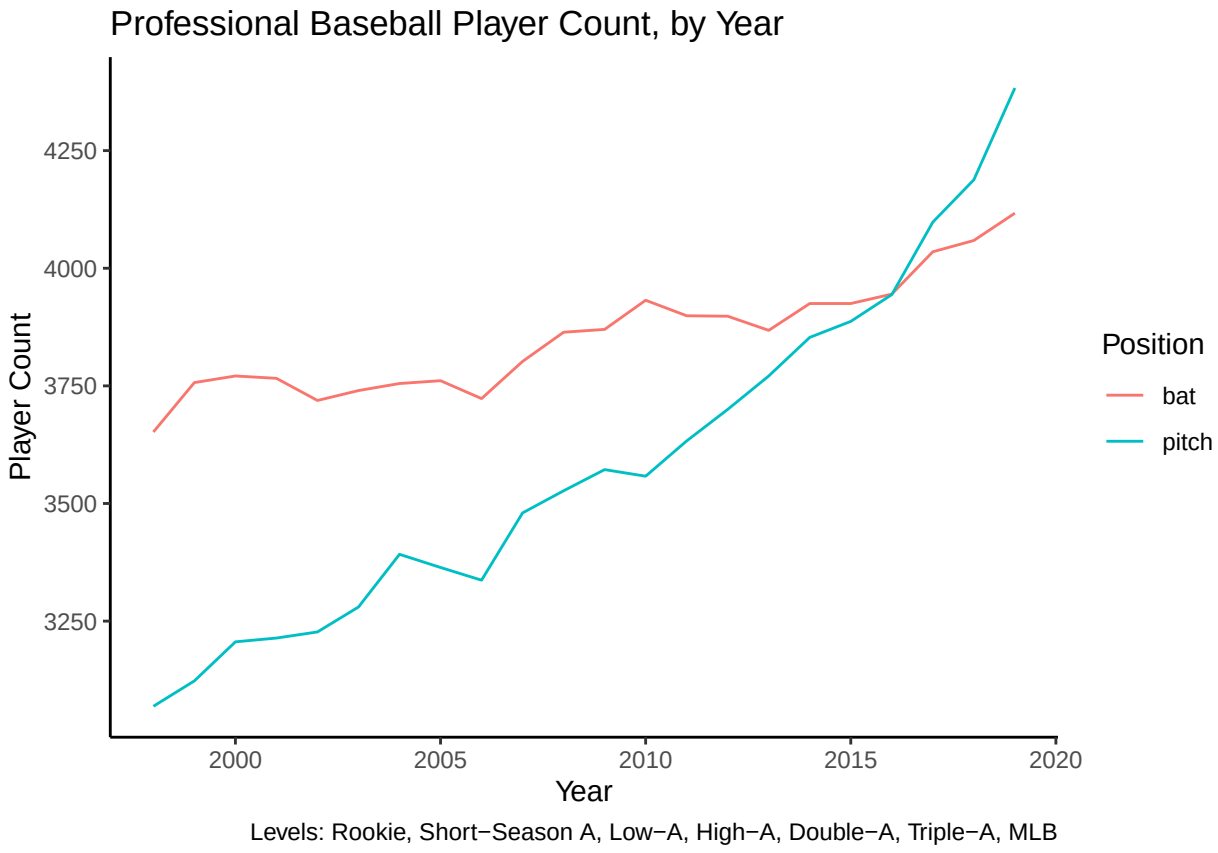
Table 2: Means and Standard Deviations of Pitcher Statistics ($n = 22,219$, $n_{min} = 19,677$)

	No BF Minimum		BF >= 20	
Measure	M	SD	M	SD
Years Played	3.55	3.02	3.76	3.07
Organizations	1.60	1.34	1.64	1.39
Levels	2.83	1.78	3.02	1.77
BF per Year	275.95	204.44	293.40	198.85

Note - for BF minimum, player season only included if BF >= 20.

The above table replicates what was done with the batting summary statistics, but for pitchers. The trends here are remarkably similar to those seen in the table for hitters, in both total number of players and their individual metrics. One small difference is that pitchers on average play at slightly more levels than hitters, but this is not an extreme.

There are many other summary statistics that will be included in the final paper - for example, the distribution of performance for players who are promoted, hazard rate, transition matrix, etc. - that were not included here. One variable of interest is how the number of players has increased year-over-year:



So we see that over the past two decades-plus, the number of pitchers in a given year has grown dramatically, a drastic increase that now outstrips that of the number of hitters per year.