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.

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.