Assignment #2: Signal v. Noise in Pitcher Outcomes

Your task is to predict the LW allowed by each pitcher in 2024, using only data from 2023 and ignoring age.

Why are you being asked to do this?

Predicting player performance is of great interest to teams, who need to make decisions on trades and free agent signings. Usually, when creating projections, you have multiple years of data, and age is an important consideration. We are simplifying the task to practice the fundamental skill of extracting signal from noise in a single season of data.

What (exactly) are you being asked to do?

This assignment is very open-ended. You are to create a set of predictions that covers every pitcher who appears in the provided data (mlb_event_2023.csv). Your goal is to predict as accurately as possible the average linear weight (per plate appearance) allowed by each pitcher in 2024. More important that the accuracy of your predictions is how well you explain your methodology and reasoning. We will not know the accuracy of your predictions until well after the assignment is graded, so you will be evaluated on your methodology and reasoning.

The expectation is that you will use some of the tools we have covered in the class (e.g. regression to the mean) in the context of batter evaluation. We have learned how to separate signal from noise for batters, and now you must think critically about how to apply that framework to pitchers. Although you are expected to use some of the tools we have learned in class, there is no right answer. You are also encouraged to think on your own about how to predict pitcher performance. Importantly, you are expected to do some sanity-checking of your results. Just as you would in a real job, make sure your results make sense before you submit them! Do you believe in the pitchers at the top of your list?

Note that you are discouraged from incorporating outside data in your predictions. For example, pitch trajectory information is very useful for pitcher projection, but we will cover that on the next assignment. For now, focus on extracting the signal from the event data.

Submission Requirements

- A PDF report (max 4 pages) summarizing your findings, including at minimum the following:
  - A description of (and reasoning for) your methodology for predicting pitcher performance
  - At least one data visualization that tells a story about your results
- A CSV of predictions, with columns `pitcher_id` and `pred_lw` (average LW per plate appearance)
- A R script (.R file extension) with all of the code you used to generate your report and the CSV

Reminders

- Prepare your report as if your audience is a baseball executive who has not seen the assignment prompt. Write clearly and concisely, and format your report in a way that makes it easy to read.
- In this class we value **critical thinking**! Don't just parrot what you've been taught—bring your own experiences to bear on the assignment. If you disagree with something we've covered in class, that's strongly encouraged! Be sure to explain your reasoning.
- Please anonymize your submission by removing any personally identifiable information.

How will your grade be determined?

You will get feedback on your work product based on several criteria. Within each of those criteria, the feedback will be: Missing (0%), Needs Improvement (70%), Good (85%) or Exceeds Expectations (100%). Your grade on the assignment will be the average of the grades across criteria. The criteria are:

1. **Description of prediction methodology.** Did you explain your prediction methodology with sufficient detail that another baseball data scientist could replicate it?
2. **Implementation of prediction methodology.** Did you correctly implement the prediction methodology (or did you make a mistake)?
3. **Data visualization.** Did you include a data visualization that tells a compelling story?
4. **Critical thinking.** Does your analysis exhibit a depth of thinking about the problem, or are you just applying the methods we've covered in class?
5. **Written communication.** Did you write clearly and concisely? Did you organize your key ideas with the evidence supporting them? Did you format your report in a way that makes it easy to read?