STATS 100: Project Proposal

Team Members: Rosman R Cariño

Research Question:

- How can we maximize the number of expected runs a baseball team can produce based within an inning?

Why is it Important?

- I think this question is important because we want to maximize the number of runs given an opposing pitcher and the batting lineup that our team has. Each baseball team uses a subset of players every game to use in their batting order, but it would be best to find out which lineup combination for a given team will maximize their number of expected runs.
- By maximizing our expected number runs given then this will contribute to the team possibly having a higher likelihood to win the game because the definition of winning a baseball game is by having more runs than the opposing team.

What data will we use, and how you procure them?

- We will use the following datasets in no order:
    - Lahman Package: https://cran.r-project.org/web/packages/Lahman/index.html
        - Lahman Package provides pitching, hitting and fielding performance and other tables from 1871 through 2022, as recorded in the 2023 version of the database.
    - Retrosheet: https://www.retrosheet.org/
        - Retrosheet provides complete play-by-play data for all AL and NL seasons for 1912 and from 1917 – 2023 (To a certain extent)
    - PyBaseball: https://github.com/jldbc/pybaseball
        - PyBaseball scrapes Baseball Reference, Baseball Savant, and FanGraphs.
- We will use the Lahman Package/PyBaseball to possibly obtain a historical average over some statistics. We will then use the Retrosheet to go over some play-by-play data that might be useful when constructing this model.

What methods will you use to answer your research question?

- I will be using the idea of a Markov Chain like we discussed in class and if the project is going well, I will reach out to the teaching staff to incorporate more ideas to make the project more interesting.
- Some ideas that I have discussed with the teaching staff:
    - Using a Bayesian model
    - Incorporating more states in the Markov Chain
    - Incorporating more probability data within the Markov Chain