Data Description
We chose Rotowire's datasets because they provide well-organized and extensive MLB statistics spanning multiple years. These stats are readily available across the internet, but Rotowire has them compiled nicely, and retrieving CSV files from them is made easy. Sports statistics are commonly collected to evaluate team and player performance, and to provide quantifiable data to determine awards winners, such as MVP, for example. Interestingly, baseball lends itself nicely to statistical analysis because of the high number of events (300 pitches a game, various velocities, etc.) Rotowire specifically seems to have ties to sports betting, which indicates another motive for collecting data. These datasets include 23 explanatory variables related to MLB batting statistics. Our unit of analysis is the team. We want to use player batting statistics to derive each team's offensive firepower and how likely they will be to succeed in the 2025 season.

Cleaning and Wrangling Steps
Both datasets were aligned with the same column structure. The dataset included various players—some only in 2023, some only in 2024, and others in both years. We retained only players who appeared in the 2024 season to simplify the analysis since we care about current and future offensive strength. If a player changed teams during the season or between 2023 and 2024, we assigned them to their most recent team. Additionally, players who switched teams during a season have more than one entry. We merged them and computed their aggregate stats. Because of missing values, we filled SF and SH with zeros as they were insignificant and added plate appearances (PA) to estimate the number of games played for those players with missing values. After merging, we performed some data wrangling. First, we analyzed the histogram for AVG and decided that there were outliers at both 0.0 and 0.4. We removed entries with these outliers. Then, we decided upon an arbitrary value of 20 games, slightly fewer than the 25th percentile, and removed players 20 or fewer games played. This is because they have a low number of plate appearances, and so a low sample size for their offensive strength. Therefore, we felt they were not a good representation of a team's offensive strength, and so they were removed.

Data Cleaning and Wrangling Steps:

1. Rectify Structural Misalignments – Dataset shape and column labels
2. Address Missing Values – Replace with zeros and calculated estimates
3. Merge Duplicate Entries – Combine data for players with multiple records
4. Standardize Data Formats – Numerical consistency and proper column identification
5. Merging - Remove unique season players and long merged by player name
6. Filter - Remove AVG = 0.0 and AVG >= 0.4. Remove G <= 20

Data Merging Issues
Before merging 2023 and 2024 statistics, we had to aggregate statistics within a year due to duplicate players. When merging 2023 and 2024 statistics, we added a column, *Year* to distinguish which year the player's data was collected in. Additionally, we wanted to ensure the team column reflected a player's current team, not any previous teams. We dropped

Team Contributions
David Sweasey: Added missing values to datasets, Data cleaning, Data merging
John Farrell: Found datasets, Data cleaning
Morgan Rivera: Documentation of README.md, Data cleaning

Timeline
1/27: Team meeting for contract and agenda
2/13: Meet with Dr. AM to get the datasets approved
2/23: Introduce missing values to data
2/24: Meet to clean and wrangle data and setup GitHub
2/25: Merge the datasets
2/26: Final cleaning and wrangling

Relevant Links
GitHub Link: https://github.com/csc442-team11/MLB-Data-Analysis.git
2023 MLB Dataset: https://www.rotowire.com/baseball/stats.php?season=2023
2024 MLB Dataset: https://www.rotowire.com/baseball/stats.php?season=2024