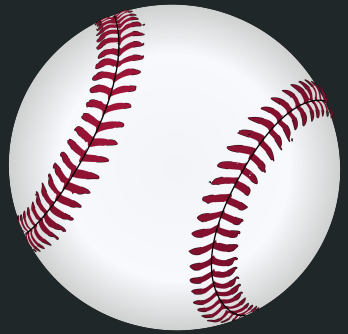
Baseball: player performance and its factors

Vail Dorchester and Jackson Mediavilla

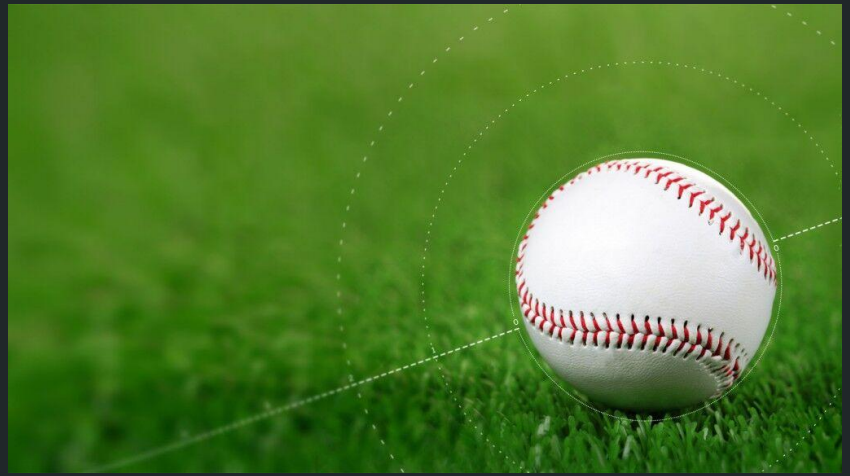
Project Description

We wish to examine the correlations between various factors and player performance in high stake situations. Data mining game events may reveal correlations that are not expected or easily identifiable. We will investigate which factors contribute most to player performance.



Prior Work

- There is an entire field of study based around the statistics of baseball called sabermetrics.
- There have been many studies focused on evaluating individual players, but it appears that less research has been done on evaluating overall player performance and how that correlates to outside factors such as weather, crowd presence, and closeness of the game.

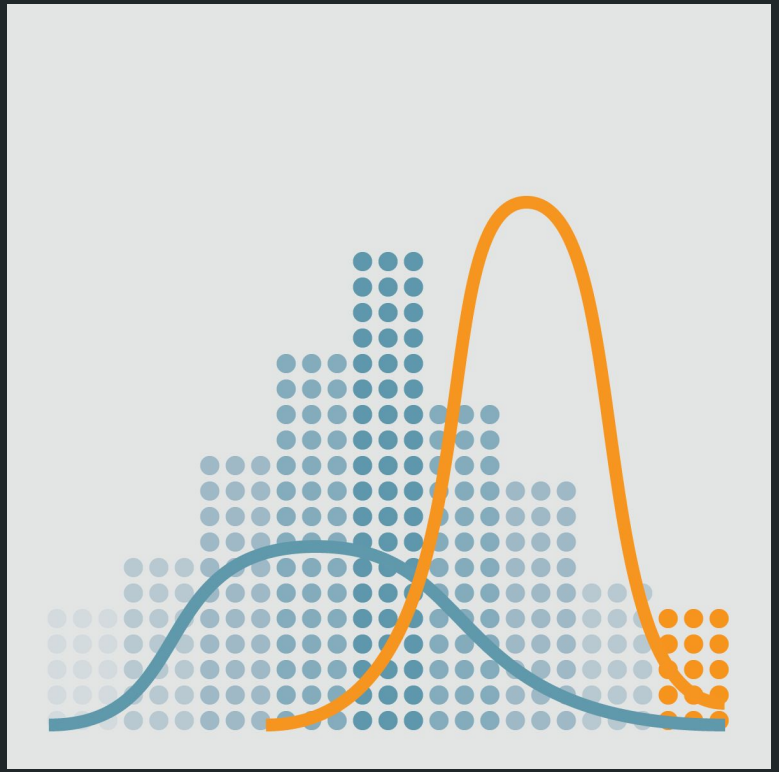


Datasets

- Retrosheet downloaded from retrosheet.org
  - Currently downloaded on Jackson’s computer
- Event files contain every play in the included games
  - Every event that changed the game state has its own entry
  - Information includes strikes, balls, batters on base, umpires and data about their assignments, bases stolen, etc.

Proposed Work

- Data cleaning and preprocessing
  - Some games are missing or incomplete
- Identify and define the ‘outside factors’ we are interested in
- Correlation analysis between performance and outside factors
- Linear regression analysis to predict performance as a function of outside factors
- Analyze this data to find the factors or situation that contribute to the best player performance



Proposed Tools

- Python
- Numpy
- Pandas - python data analysis library
- Git, GitHub
  - https://github.com/jackson-mediavilla/csci4502-project/
- Excel
- Jupyter Notebook

Evaluation and desired results

- We can do basic statistical analysis to discover relationships between various factors and player performance
  - This will require defining which situations qualify as “high stake” and a method to evaluate player performance. A possible method for evaluating player performance is to create a function that assigns each team a probability of winning given the current game state, and then analyze how player performance (i.e. specific events) change the probabilities.
- Pandas can be used to visualize the results.