**Darin Young Term Project**

**Milestone 1**

**Project Subject Area:**

For my project, I will be looking at baseball related data from the MLB to hopefully find new insights (or confirm existing ones) regarding how different factors impact success in baseball.

**Data Sources:**

**Flat File:**

The flat file I will be using is a csv file that contains information about each of the ballparks in the MLB. It contains information like distance to the outfield walls, average temperature, elevation, and more. This will help me evaluate trends related to external factors that might have an impact on success. For the remainder of this project, I will refer to this dataset as ballpark\_data.

[MLB Ballparks (kaggle.com)](https://www.kaggle.com/datasets/paulrjohnson/mlb-ballparks)

**API:**

The API that I will be using is from the website “MLB Data API” which is owned and operated by the MLB/MLB Advanced Media. It contains player data, team data, injury data, roster data – pretty much all of the advanced data that I will need for my project. For the remainder of this project, I will refer to this dataset as player\_data.

[MLB Data API (appac.github.io)](https://appac.github.io/mlb-data-api-docs/#top)

**Website:**

The website I will be using comes from baseballreference.com, the leader in public access to baseball statistics. Specifically, I will be using the table of teams linked below to see how external factors (ballpark conditions from above) affect teams’ success in the aggregate. For the remainder of this project, I will refer to this dataset as team\_data.

[List of all the Major League Baseball Teams | Baseball-Reference.com](https://www.baseball-reference.com/teams/)

**Relationships:**

Ballpark\_data will be related to team\_data and player\_data via team\_id, which will be a 3 letter short name for the team (HOU for Houston Astros, NYY for New York Yankees, for example).

All 3 of the datasets will be related by team\_id. Ballpark\_data has a one to one relationship with team\_data (one team only has one ballpark) and a one to many relationship with player\_data (many players can play for the same team).

**Project approach/plan:**

My plan for this project is to analyze various factors that may impact team/player success, such as average temperature at ballparks, distance to outfield walls, and other external factors. My expectation is that teams that are located in places with favorable conditions (higher elevation, lower average temperatures, shorter distances to the outfield walls) will have generally more success over the long term. Specifically, I hope to find individual players and view their career longitudinally and see if their performance increases or decreases in relation to their home ballpark. Additionally, I will take a look at the teams as a whole over time to see if there is any overarching relationship between favorable ballpark conditions and success.

The biggest challenge I think will face is with narrowing down the data enough to make meaningful conclusions. The MLB is known for its massive amounts of data, so it will be challenging to ensure that I include all the data that is necessary while ensuring that I don’t include too much that may lead to noise. Additionally, I am concerned that there are so many factors that affect a players/team’s success, that much of the relationship (if any exists at all) will be masked by other factors.

I don’t believe there are any significant ethical implications for this project. Baseball is a sport meant for entertainment and any correlation with success/unsuccess only affects the entertainment value of the sport.