Project 4 Outline

By: Allen, Matthew, Michael, Cole

1. Problem worth solving, analyzing, or visualizing:

Which MLB team in the NL West is most likely to win the division in 2023, using 2015, 2016, 2017, 2018, 2019, 2021 as training data & 2022 statistics as the test data

Using: batting data, pitching data, wins

- Batting
  - Batting Average
  - Slugging Percentage
- Pitching
  - ERA
  - WHIP

2. We will be using (Select 2):

- **Python Pandas - Cleaning and creating dataframes + Spark (ML)**
- **Tableau- Displaying Data**

And Machine learning (ML) + Scikit-learn

Data Model Implementation (25 points)

- **A Python script initializes, trains, and evaluates a model (10 points)**
- **The data is cleaned, normalized, and standardized prior to modeling (5 points)**
- **The model utilizes data retrieved from Spark (5 points)**
- **The model demonstrates meaningful predictive power at least 75% classification accuracy or 0.80 R-squared. (5 points)**

Data Model Optimization (25 points)

- **The model optimization and evaluation process showing iterative changes made to the model and the resulting changes in model performance is documented in either a CSV/Excel table or in the Python script itself (15 points)**
- **Overall model performance is printed or displayed at the end of the script (10 points)**