## Lab 2.2 – Data Management for MLB baseball machine learning problems.

In this lab, we will construct data sets that will be used to build classifiers related to MLB baseball.

Link to my repo: <https://github.com/ShoSato-047/lab_2.2>

**Research Questions.**

1. **Classification problem.** How well fielding statistics can be used to predict whether a player wins a gold glove at their position.
2. **Regression problem.** How well batting, pitching, and fielding statistics help predict a player’s salaries.

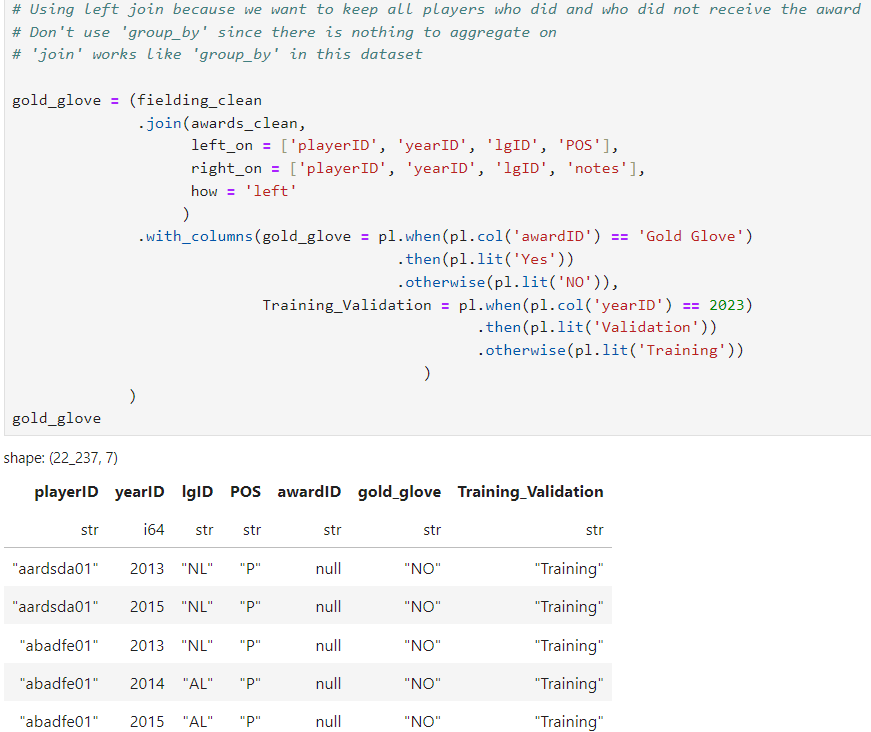
**Initial Tasks.** Perform each of the following tasks before constructing the two data sets.

1. Create a date repository,
2. Create a Python jupyter notebook in the root directory, and
3. Please the unzipped Lahman baseball database in a data folder.

**Construct the Gold Glove dataset.** First, be sure to read up on the [Gold Glove](https://en.wikipedia.org/wiki/Gold_Glove_Award) award. Our goal is to build a single data set covering years from 2013-present. The table will have one row per player-year and needs to contain the following

\*\*\*One row per player, year, position, league.

1. The target column should represent whether a player won a gold glove that year. This information needs to be extracted from the AwardsPlayers.csv file.
2. The features for this problem will be the fielding statistics found in the Fielding.csv file.
3. We will also want to create a Training-Validation column, which will be marked as **Validation** for the most recent complete season and **Training** otherwise.



**Construct the Salary dataset.** Our goal is to build a single data set covering years from 2013- present. The table will have one row per player-year and needs to contain the following

\*\*\* One row per player, year.

\*\*\* Final table should be around - 9657 rows, 63 columns

\*\*\* 2016 is a target year – 10 years before 2016 (include 11 years total)

\*\*\* stint, team, leagueIDs will be gone after group\_by & aggregation

1. The target column is the baseball players yearly salary, which can be found in the Salaries.csv file.
2. The features for this problem will be a combination the batting, pitching, and fielding statistics found in the respective files.
3. We will also want to create a Training-Validation column, which will be marked as **Validation** for the most recent complete season and **Training** otherwise.



