Moneyball Week 5 - Assignment Overview

In week 5 we used play by play data from 2016 and 2017 to explore the relationship for run values between seasons. The assignment for this week is to add the seasons 2014 and 2015 to this analysis.

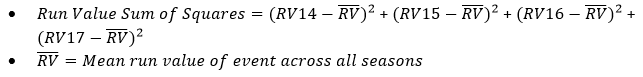
To complete the assignment you will need to repeat the same kinds of steps we took to generate our run value calculations. This assignment will include three checkpoints where you will be asked to answer multiple choice questions based on your results. After answering each quiz you will be shown code that can produce the correct answers, which you should use as the basis for moving onto the next question.

**Beware:** Even though your code might get you to the correct answer at a given point, it is sometimes possible that the way you write it might interfere with completing a further step. So even if you get the answer right, you should look at the code we supply to check if you are going the same way. In practice, there are often many ways to get to answer in Python, and we do not insist that you follow our approach exactly – but simply warn you to be aware that differences could turn out to be problematic later.

# Assignment - Part 1

## **I. Comparing Event Run Values**

1. Import pandas, numpy
2. Using a function similar to what was used in the week’s exercise notebook, calculate the run value for every event in 2014, 2015, 2016 and 2017.
3. Calculate the average run value for each type of event for every season
4. Merge the event level run values for each season into one data frame.  The data frame should include the event name and then four columns with the run values (one for each season).
5. Delete the row for event “Sacrifice Bunt DP”
6. Compute the correlation matrix for event level run values across all seasons
7. For each event, calculate the sum of squares between run values using the data from all four seasons and create a column for this sum of squares variable



# Assignment - Part 2

## **II. Comparing Player Run Values**

1. Compute the aggregate player level run values for each season
2. Merge player run values for each season into one data frame **so that only players with run values for all four seasons are included in the data frame.** The data frame should include each player’s name and then four columns with the run values (one for each season).
3. Compute the correlation matrix for event level run values across all seasons
4. Run a regression model by regressing player run values from 2017 (dependent variable) on player run values from 2014, 2015, and 2016 (independent variables)

# Assignment - Part 3

## **III. Comparing Team Run Values**

1. For each season’s run expectancy data frame, create a variable “team” to denote the batting team
2. Compute the aggregate team level run values for each season
3. Merge team run values for each season into one data frame. The data frame should include each team’s name and then four columns with the run values (one for each season).
4. Compute the correlation matrix for event level run values across all seasons
5. Run a regression model by regressing team run values from 2017 (dependent variable) on team run values from 2014, 2015, and 2016 (independent variables)