Moneyball Week 4 - Assignment Overview

In week 4 we used play by play data from 2018 to estimate the run value of players and events using the run expectancy framework. The assignment for this week is to repeat this process for the seasons 2016 and 2017.

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. Coding Run Expectancy Dataset (2017)**

1. Import pandas and numpy packages and read in “MLBAM17.csv” data
2. Keep only the relevant columns (see exercise notebook)
3. Create three indicator variables (one for each base) to denote whether or not each base is occupied prior to the plate appearance
4. Create a “start state” variable to denote the base out state prior to each plate appearance
5. Create three indicator variables (one for each base) to denote whether or not each base is occupied after the plate appearance
6. Create an “end state” variable to denote the base out state after each plate appearance
7. Restrict data to plays where either there is a change in state/runs scored and innings in which there were exactly 3 outs recorded
8. Calculate run expectancy by starting state and merge this into play by play data
9. Using the run expectancy data by starting state, create a variable to denote run expectancy by ending state (in addition, make sure to include the base out states for which there are 3 outs with run expectancies of 0 to the run expectancy by ending state data)
10. Merge the ending state run expectancy into the play by play data
11. Calculate the run value of each event

# Assignment - Part 2

## **II. Coding Run Expectancy Dataset (2016)**

Repeat part I for 2016 using the MLBAM16.csv data (Instructions are included again here for convenience)

1. Read in “MLBAM16.csv” data
2. Keep only the relevant columns (see exercise notebook)
3. Create three indicator variables (one for each base) to denote whether or not each base is occupied prior to the plate appearance
4. Create a “start state” variable to denote the base out state prior to each plate appearance
5. Create three indicator variables (one for each base) to denote whether or not each base is occupied after the plate appearance
6. Create an “end state” variable to denote the base out state after each plate appearance
7. Restrict data to plays where either there is a change in state/runs scored and innings in which there were exactly 3 outs recorded
8. Calculate run expectancy by starting state and merge this into play by play data
9. Using the run expectancy data by starting state, create a variable to denote run expectancy by ending state (in addition, make sure to include the base out states for which there are 3 outs with run expectancies of 0 to the run expectancy by ending state data)
10. Merge the ending state run expectancy into the play by play data
11. Calculate the run value of each event

# Assignment - Part 3

## **III. Comparing 2016 vs. 2017**

1. Calculate the average run value for each event for both 2016 and 2017.
2. Merge the 2016 and 2017 run values by event into one dataframe (before merging, it may be helpful to rename the run value variables Run\_Value16 and Run\_Value17 respectively)
3. Calculate the aggregate run value (sum) for all players in both 2016 and 2017
4. Merge the 2016 and 2017 run values by player into one dataframe so that only players appearing in both datasets appear in the merged data (before merging, it may be helpful to rename the run value variables Run\_Value16 and Run\_Value17 respectively)