Week 1 - Assignment Overview

In week 1 we used five seasons of data from Retrosheet to reproduce Table 1 from the Moneyball paper of Hakes and Sauer, to show that OBP is more significant in determining team winning percentage than is SLG.

The assignment for this week is to run a regression of team winning percentage on the oldest of all baseball statistics: batting average. Remember that batting average is defined simply as the number of hits divided by the number of at bats. Both hits and at bats are included in the Retrosheet data.

To complete the assignment you will need to repeat the same kinds of steps we took to generate Table 1. At each step you will be asked to answer a multiple choice question based on your data. After answering each question you will be shown code that can produce the correct answer, 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 wedo 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. Data Coding and Merging**

1. Load the “Game logs” data file, create a dummy variable for home wins (1 or 0) and visitor wins, and create the ‘year’ variable.
2. Create a dataframe that sums the statistics you need by season (year) and by home team. Rename the column for the home team as ‘Team’.
3. Create a dataframe that sums the statistics you need by season (year) and by visiting team. Rename the column for the visiting team as ‘Team’.
4. Merge the two dataframes on ‘year’ and ‘Team’ and create a variable for the total number of wins in the season for each team

# Assignment - Part 2

## **II. Calculating Performance Metrics**

1. Create batting average for and batting average against as the sum the relevant statistics for each team as home team and visiting team.
2. Create the win percentage variable for each team by generating a count of games played and dividing total wins by total games played.
3. Using a similar process to part 2), calculate each team’s home and away win percentage separately for each season.
4. Merge each team’s home and away win percentage for each season into one data frame. Calculate WPCT diff defined as home win percentage – away win percentage for every team in each season.

# Assignment - Part 3

## **III. Running Regressions**

1. Run an OLS (Ordinary Least Squares) regression of win percentage on batting average for and batting average against.
2. Run an OLS (Ordinary Least Squares) regression of win percentage on batting average for and batting average against,   constraining the coefficient on each variable to be equal and opposite.