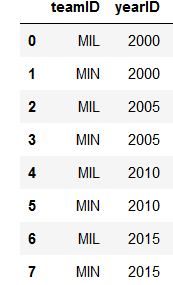
**Baseball Lab**

This lab will walk you through various uses of the SQL programming language to manage data and query information. The data for the lab is a collection of tables that contain batting and pitching statistics from 1871 to 2018, plus fielding statistics, standings, team stats, etc. The database is already uploaded into the Github site for this workshop. You may also want to reference the data documentation (<http://www.seanlahman.com/files/database/readme2017.txt>) to help aid in variable definitions and deciding which tables to utilize to solve the question at hand.

1. One measure of how well a player (or team) is doing is to calculate the Earned Run Average (ERA) which can be calculated using this formula:  . According to [Lootmeister Sports](http://www.lootmeister.com/mlb/good-bad-era.php), an ERA below 4.0 is considered OK.  
   1. Using the **Pitching Table**, verify the ERA for players that play for the Minnesota Twins (TeamID = MIN) and the Milwaukee Brewers (TeamID = MIL) in the year 2000. Additionally, include only the players whose ERA is considered OK.
   2. Using the **Team Table**, verify the ERA for teams in the years 2000, 2005, 2010, and 2015 for the Minnesota Twins and the Milwaukee Brewers. Your output should include the yearID, teamID, and Calculated ERA values. Additionally, arrange the values in your output to match the ordering given in the output below.



1. Another summary of interest might be salaries of the players.  
   1. Using the appropriate tables, find the average salary by team for 2006 and 2012. Your output should contain the following fields: Year, Franchise Name, Team, and Average Salary. (Hint: You will need to use three tables to accomplish this task.)

* 1. Find the team with the highest average salary overall, regardless of year.

1. Suppose I am interested in getting a list of all the players’ names (first and last) who play for the Detroit Tigers that have been born since 1980. Your output should only include TeamID and Player Name. Additionally, the names in your output should be organized alphabetically by last name with those at the end of the alphabet at the top of your output.

Comments:

* If you are getting duplicate rows, you can use the SELECT DISTINCT function to eliminate the duplicate rows.
* An INNER JOIN will only show the rows that are contained in BOTH tables being joined. You’ll want to use this rather than a LEFT JOIN in this case.
* It may be helpful to do any filtering before the tables are joined.
* You’ll have to figure out what the concatenate function is in PROC SQL to obtain the Player Name field.

1. Another statistic used to assess how well a player is performing is to use On-base plus slugging (OPS) (values closer to 1 are better), which is calculated using the following formula: OPS = OBP + SLG where

OBP =  and

SLG = 

Using the **Batting table**, compute the OPS for each player since 2010. Your output should include playerID, OBP, SLG, and OPS. Additionally, your output should NOT include missing values (i.e., you only want numbers in your output) or OPS values of 0.

1. The baseball database contains information about all the colleges and universities that the players came from. Suppose it is of interest to determine how many players there are from each college or university in the year 2009. Your output should include the full name of the college or university and the number of players and have the school with the most players at the top.

Comments:

* If you are getting duplicate rows, you can use the SELECT DISTINCT function to eliminate the duplicate rows.
* An INNER JOIN will only show the rows that are contained in BOTH tables being joined. You’ll want to use this rather than a LEFT JOIN in this case.
* It may be helpful to do any filtering before the tables are joined.