**Project Description**

A csv file called ***baseball\_teams.csv*** is provided.

The data in the file describes historical baseball team performance and includes the following data fields

Descriptive Fields

* Year
* League
* Team
* Franchise
* Division

Team Performance

* Final Standing
* Games Played
* Games Won
* Games Lost
* League Win (y/n)
* World Series Win (y/n)

Offensive Measurements

* Runs Scored
* At Bats
* Hits
* Doubles
* Triples
* Home Runs
* Walks
* Strike Outs
* Stolen Bases
* Caught Stealing
* Hit By Pitch
* Sacrifice Fly

Defensive Measurements

* Runs Against
* Earned Runs
* Earned Run Average
* Complete Games
* Shutouts
* Saves
* Infield Put Outs
* Hits Allowed
* Home Run Allowed
* Walks Allowed
* Strikeouts Allowed
* Errors
* Double Plays
* Fielding Percentage

Miscellaneous

* Team Name
* Home Ball Park
* Attendance

**Project Objectives**

The project has the following objectives.

1. Determine if teams are more successful with high offenses, tight defenses or a balance of the two.
2. Determine if this conclusion is valid across multiple time periods.
   1. Period 1 – before 1920
   2. Period 2 - 1920 to 1960
   3. Period 3 – 1960 to 1990
   4. Period 4 – 1990 to 2010

**Approach**

1. Perform an exploratory data analysis to understand the statistical properties and distribution shapes of 8 key variables of your choice. Perform this analysis for a single dataset that spans 1960 to 2010
   1. Calculate the descriptive statistics for the 8 variables
   2. Show a histogram for each of the 8 variables
   3. Show a boxplot for each of the 8 variables
   4. Comment on the visual shape of the distribution of each variable
      1. Is it symmetric?
      2. Is there a single mode?
      3. Is it narrow or wide?
      4. Are there many outliers?
   5. Comment on the validity of the assumption that each of the 8 variables is normally distributed.
2. Do a correlation analysis for the following pairs of fields for the complete data spanning 1960 to 2010
   1. Games Won vs Runs Scored
   2. Games Won vs Runs Against
   3. Games Won vs (Runs Scored minus Runs Against)
   4. Games Lost vs Runs Scored
   5. Games Lost vs Runs Against
   6. Games Lost vs (Runs Scored minus Runs Against
   7. Provide an interpretation and commentary for your analysis.
3. Create a multiple linear regression model for each of the 4 time periods noted above and select the independent variables that are good predictors of games won.
   1. Keep the model as simple as possible
   2. Try combinations of offensive and defensive fields as independent variables.
   3. Justify your model in terms of the independent variables you selected
4. Evaluate each of your 4 regression models
   1. Statistical significance of the overall model
   2. Statistical significance of each coefficient, including the intercept
   3. Are your independent variables approximately normally distributed?
5. Use the 4th regression model from 1990 to 2010 and forecast the number of games won for the New York Yankees and the Toronto Blue Jays using values for the independent variables for 2012 and 2015.
   * 1. Discuss the error in your prediction by comparing the actual wins to predicted wins for the 2 teams in the 2 years.
     2. Quantify your comments about the model error by calculating the Root Mean Square Error for your set of predictions.. An example of this calculation is provided in the Week\_5\_Example\_Notebook.ipynb