**Baseball**

The premise in the movie *Moneyball* is searching for the secret of success in baseball. It follows a low-budget team, the Oakland Athletics, in the early 2000’s, who believed that underused statistics, such as a player’s ability to get on base, actually better predicts the ability to score runs than the typical, traditional statistics like homeruns, RBIs (runs batted in), and batting average. In fact, obtaining players who excelled in these underused statistics turned out to be much more affordable for the team.

In this activity, you are going to be looking at data from all 30 Major League Baseball teams from the 2009 season and examining the linear relationship between runs scored in a season and a number of other team statistics. The goal is to summarize these relationships both graphically and numerically in order to find which variable, if any, best helps us predict a team’s runs scored in a season.

**Research Question:** Which variable(s) in baseball helps predict a team’s runs scored in a season?

* Open up the data set: *Baseball.csv*.

The variables in the data set include the following:

1. runs: Total number of **runs scored**
2. at\_bats: Total number of **at-bats**
3. hits: Total number of **hits**
4. homeruns: Total number of **homeruns**
5. bat\_avg: **Batting average**
6. strikeouts: Total number of **strikeouts**
7. walks: Total number of **walks**
8. stolen\_bases: Total number of **stolen bases**
9. on\_base: **On-base percentage**
10. slugging: **Slugging percentage**
11. ob\_slg: **On-base plus slugging percentage**

**Discuss the Following Questions**

1. Which variable is the response variable? How do you know?
2. Which variables are the explanatory variables? How do you know?

The variables numbered 9-11 are the underused, newer variables in baseball and the variables numbered 2-8 are the traditionally-used variables.

* Use the statistical software of your choice (e.g., R, SAS, StatCrunch) to produce a scatterplot of *runs* verses *at\_bats*.

1. Sketch a scatterplot of the *runs* verses *at\_bats*.
2. Describe the relationship between the total number of at-bats and the total number of runs for the Major League Baseball teams. In particular, comment on:
3. The trend of the plot (linear, curved, no trend, etc.)
4. The direction of the trend
5. The clustering of the points around the trend (tightly clustered around the trend, more scattered around the trend)
6. Outliers

* Use the statistical software of your choice (e.g., R, SAS, StatCrunch) to compute a correlation between *runs* and *at\_bats*.

1. Report the correlation between *runs* and *at\_bats*.
2. Interpret the value of the correlation between *runs* and *at\_bats*.

* Choose another traditionally-used variable and use the statistical software of your choice (e.g., R, SAS, StatCrunch) to produce another scatterplot of *runs* verses *traditionally-used variable* of choice.

1. Sketch the scatterplot below.
2. Describe the relationship between your traditionally-used variable of choice and the total number of runs for the Major League Baseball teams.

* Use the statistical software of your choice (e.g., R, SAS, StatCrunch) to compute a correlation between *runs* and your *traditionally-used variable* of choice.

1. Report the correlation between runs and your *traditionally-used variable* of choice.
2. Interpret the value of the correlation between runs and your *traditionally-used variable* of choice.

* Choose an underused variable and use the statistical software of your choice (e.g., R, SAS, StatCrunch) to produce another scatterplot of *runs* verses the *underused variable* of choice.

1. Sketch the scatterplot below.
2. Describe the relationship between your underused variable of choice and the total number of runs for the Major League Baseball teams.

* Use the statistical software of your choice (e.g., R, SAS, StatCrunch) to compute a correlation between *runs* and your *underused variable* of choice.

1. Report the correlation between *runs* and your *underused variable* of choice.
2. Interpret the value of the correlation between *runs* and your *underused variable* of choice.

* Put the correlations and the name of the variables that you examined with *runs* on the table on the board.
* Discuss with others around you the variables that were examined with the variable *runs* and the correlations that were observed between those variables.

1. Which variable examined is the best predictor of the total number of runs for a team?

**EXTENSIONS**

The variables examined in this activity all had a positive relationship with the *runs* response variable.

1. What would a negative relationship look like? Sketch a scatterplot of a negative relationship between two variables below.
2. If a relationship is negative between two variables, what values could the correlation be?