Dataset: Academic Scores for NCAA Athletic Programs

**Where is it from?**

This data set was compiled by the NCAA (National Collegiate Athletic Association), and we found it on Kaggle (https://www.kaggle.com/ncaa/academic-scores). It’s an attempt to compare academic information about college student-athletes across schools and sports. The NCAA created the Academic Progress Rate, a composite score based on eligibility, retention, and graduation rates (higher score means better academics), to ensure athletic teams are held accountable for their academics. The NCAA penalizes teams with low a APR (e.g. make them ineligible for postseason), as well reward teams with high scores.

**How many rows/columns are in this dataset?**

There are 6511 rows and 57 columns. We plan to merge this dataset with the college scorecard~~, as well as information about each team and college’s overall athletics’ performance[ie revenue, total number of championships, home game attendance~~].

**What information does it contain?**

SCHOOL\_ID – unique school identifier

SCHOOL\_NAME – school name

SCHOOL\_TYPE – 0 = public, 1 = private

SPORT\_CODE – sport identifier

SPORT\_NAME – sport name

NCAA\_DIVISION – whether sport is Division 1, 2, or 3 (based on school size)

NCAA\_CONFERENCE – school’s conference affiliation

FOURYEAR\_ATHLETES – team’s total number of athletes from 2011-2014

FOURYEAR\_SCORE – team’s average APR score from 2011-2014

FOURYEAR\_ELIGIBILITY – team’s average

FOURYEAR\_RETENTION – team’s average

20XX\_ATHLETES – team’s number of athletes in a certain year from 2004-2014

20XX\_SCORE – team’s APR score in a certain year from 2004-2014

20XX\_ELIGIBILITY – team’s eligibility in a certain year from 2004-2014

20XX\_RETENTION – team’s retention in a certain year from 2004-2014

**Questions we hope to answer:**

We are particularly interested in retention, and we hope to answer the question, “Which factors are most important in determining retention rates for NCAA athletic teams?” To answer this, we will look at…

* + - * Which sports, schools, conferences, geographic regions have the highest and lowest retention rates?
      * Is there a difference between public, private institutions? Men’s, women’s sports?
      * Which schools retention rates have increased/decreased the most in the last 10 years?
      * How does the overall student body compare to the athletic student body in regards to graduation rates?
      * How does school tuition affect retention?

We will also try to predict retention rates with multi linear regression (using stepwise regression to determine which variables are significant). Finally, we will determine which team has the best APR (but we know it’s going to be Yale Women’s Tennis!)