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. In addition, we plan to merge this with the college scorecard to compare athletes to the general student population, as well as data on each team and college’s overall athletic performance (ie: revenue, championships, etc…).

**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 – average % of the team academically eligible to play (satisfied GPA and credit requirements) from 2011-2014

FOURYEAR\_RETENTION – average % of the team that stays in school from 2011-2014

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 – % of team academically eligible to play (satisfied GPA and credit requirements) in a certain year from 2004-2014

20XX\_RETENTION – % of the team that stays in school in a certain year from 2004-2014

**Questions we hope to answer:**

We are particularly interested in APR, and will focus on the question, “Which factors are most important in determining APRs for NCAA athletic teams?” We will look at…

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

We will also try to predict APRs 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!)