CS586—group project, deliverable 1

Navya Sri Ambati & John O Brickley

This database is broadly about the relationship between college financing, academics, and athletics —but mainly about the relationship between financing and athletics. Depending on the kinds of relationships we are able to tease out of our datasets, we may or may not pursue the question as to the relationship between academics and athletics in colleges. We suspect that it will be somewhat more straightforward to investigate the relationship between colleges' finances overall and their spending on athletics. For the purposes of this project, we expect that any statistical analysis we carry out will be minimal. Rather, we plan to arrange our data so as to allow researchers to easily frame queries about any relationships between college financing, academics, and athletics that might be present.

Example questions

Some examples of the kinds of questions we would like to be able to ask follow.

1. What is the connection between college sports and academics?
2. Does spending on college sports affect student's academic performances?
3. How is cost of attendance affected by funding for college sports?
4. How does increased funding for college sports affect universities, colleges, and the areas surrounding them?
5. Does sports gambling correlate with subsidies to college sports?
6. Is there a connection between subsidies to college sports teams and Title-IX investigations?
7. How many students are in high school, under grad, grad etc programs?
8. How many students are involved in both academics and sports?
9. Which colleges have the highest fees?
10. What are the top 15 schools and colleges that have good ratings?
11. Which schools and colleges support financing for academically oriented students?
12. Which schools and colleges support financing for students involved in sports?
13. How many students are involved in college sports?
14. What kinds of funding are offered to students at schools and colleges?
15. How much funding is offered to students when comparing between states?
16. Which schools and colleges are highly supportive of sports?
17. What is the variation in funding for education versus sports?
18. Which schools and colleges participate and encourage students in more than 1 sport?
19. What is the average of students who are good at both sports and academics?
20. How many consecutive wins have various college sports team had?
21. How many students, by school, have received funding based on athletics participation?
22. What is the average amount or scholarship money offered per year in each category?

23. Is there a difference in the average funding for male and female college sports teams?
24. How does any difference in the previous question compare to differences in funding based on academic merit?

Sources and their ingestion

We have collected the following datasets:

- ☒ The "College scorecard data"[1], which, while off by an average of 10% in reported graduation rates among Pell-grant recipients[2], gives a comprehensive view of universities in the US as a whole[3];
- ☒ *The Huffington post* and *Chronicle of higher education*'s data on how colleges finance their athletics[4];
- ☒ The *Department of education*'s data on foreign gifts to and contracts with US colleges[5];
- ☒ *The census bureau*'s "Quarterly summary of state and local government tax revenue[6];
- ☒ NCAA data on student athletes' academic progress and graduation rates[7];
- ☒ The *Department of education*'s data on student loan default rates[8]; and
- ☒ The *Department of education*'s annual school- and team-level datasets on college sports' finances[9].

In aggregate, the above datasets amount to several tens of thousands of lines. We intend to convert each of these datasets to CSVs where they are not already distributed as such. For preprocessing and initial sanitization, we plan to use Unix tools like Sed and Awk—that is to say, we plan on carrying out our preprocessing in the usual Unix-y way. We plan to track down more subtle problems by enforcing column restrictions on our schemas[10] While we plan to import the data organized around individual colleges, we have yet to develop our database schema beyond that.

[1] https://collegescorecard.ed.gov/data

[2] https://hechingerreport.org/theres-finally-federal-data-on-low-income-college-graduation-rates-but-its-wrong/

[3] downloaded from https://ed-public-download.scorecard.network/downloads/College_Scorecard_Raw_Data_01162025.zip

[4] Described at http://projects.huffingtonpost.com/ncaa/reporters-note and directly downloadable from http://hpin.s3.amazonaws.com/ncaa-financials/ncaa-financials-data.zip.

[5] Described at https://studentaid.ed.gov/sa/about/data-center/school/foreign-gifts and downloadable from https://studentaid.gov/sites/default/files/ForeignGifts.xls.

[6] Indexed at https://www.census.gov/programs-surveys/qtax.html and downloadable from https://www2.census.gov/programs-surveys/qtax/tables/historical/2009Q1-2024Q3-QTAX-Table1.xlsx.

[7] Available from https://www.icpsr.umich.edu/web/ICPSR/studies/30022# with an institutional login via PSU.

[8] Downloadable from https://studentaid.gov/data-center/student/default.

[9] Bandied at https://ope.ed.gov/athletics/ and downloadable for the years 2003–2023 at https://ope.ed.gov/athletics/#/datafile/list.

[10] Or *schemata*, if you really must.