# 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.

**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];

[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.

- ☒ 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.

[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.