NICAR Scorecard User Guide

You must (please) read the **Data Documentation**, as well as the **NICAR_Data_Dictionary** before working with the data. You can also consult the **Technical Paper and Policy Paper**, which will answer a lot of questions as you go and provide context into some national figures derived from the data. Those papers also provide key insights and several examples for the types of analyses best suited to the data.

Earlier this year, for the first time ever, the Obama Administration released a comprehensive intersection of student population, college performance and “outcome” data, measuring with precise detail who gets into what school and what they do after graduation. But the Department of Education’s raw “College Scorecard” is a labyrinth of information covering some 7,800 campuses all over the country. The original data was really wide, more than 1,700 columns from three main sources: an annual survey collected by the Department of Education (IPEDS), the national student loan database (NSLDS), and administrative earnings data from tax records.

Using the data, newsrooms can track and compare schools’ accessibility across different income levels alongside performance metrics and ultimate outcomes. Many schools all over the country are under budget constraints, so understanding how appropriations have impacted students can paint a vivid picture in your city or state. The Scorecard also provides insight into the world of private for-profit colleges, which have sprung up in cities -- and online -- across the country. Most importantly, the data covers all federal grant and loan recipients, so reporters can now measure the effectiveness of government aid across different types of students and schools.

The original purpose of the dataset, its creators say in the documentation, is to allow students to make school-to-school comparisons so they can decide what institution best fits them and their individual goals. It’s also meant to help regulators assess accessibility and outcomes. The idea being that you could make some conclusions about a school’s performance based what types of students come in and what they do after they leave. But those conclusions come with caveats and disclaimers (read, warnings), which need to be considered before during your analysis. Please see the **NICAR Caveat Guide** for more information.

It helps to think of the NICAR College Scorecard data in three general groups, which get progressively granular in their data.

1. **Colleges**
2. **Students**
3. **Outcomes**

I. Colleges

Keep in mind these basic -- but important -- categorical distinctions in this data slice. If this was data on football, these columns would show you things like teams, leagues, conferences, and divisions. They describe and define each school.

1. There's an important distinction between **UNITID** and **OPEID,** the data's two unique identifiers for each college. The **Data Documentation** and the **NICAR Caveat Guide** outline exactly what you need to know. But basically **UNITID** is specific to each campus, while **OPEID** bundles branch campuses together under the umbrella of its flagship/main campus.[1]
2. **Type:** Public, Private Non-Profit and Private For-Profit.
3. **Degree:** Associate, Certificate and Bachelors. (Note: This is not the highest degree awarded, but the most predominant one. **See page 3 of the Data Documentation**.) The associate degrees are usually for two year programs, many community colleges, which has an effect on outcome measures.
4. There's the typical location stuff: **city, state and region** (check the **Data Dictionary** to see what states belong in what region). You can use that to easily localize your story.
5. **Campus:** Is it the main campus or a branch? There's an important caveat with branches in this data, explained in the **NICAR Caveat Guide and page 28 of the Technical Paper.**

II. Students

Then there's the more specific demographic data for the schools' student population. A lot of this is based on survey responses collected from the Department of Education. (Note: these columns are different and separate from **outcome** data, which we'll do next.) If you'll buy the football metaphor, think of these columns as individual teams' roster information -- not their performance measures. However, these columns are crucial when looking at outcomes and building ratios.

1. Enrollment size (**undergrad_enrollment**), which is then broken out by some demographic information:
   a. Race (**percentage_white, percentage_black, percentage_asian, etc.**)
   b. Test scores (**ACT_median, SAT_average)**
2. How many are they accepting? **Admission_rate**
3. Economic status of the entire student body:
   a. **median_family_income,**

[1] As noted in the **NICAR Caveat Guide** and **data documentation**: While some schools report these data at the campus level (8-digit OPE ID), data produced for this site are rolled up to the institution level (6-digit OPE ID). In these cases, IPEDS institutions sharing a common 6-digit OPEID are all assigned the (student-weighted) average outcome or median outcome for students across all branches of the institution for NSLDS or tax-data derived measures

   b. **pell_recipients, fed_loan_recipients,** which percentages of the student body are receiving which type of Title IV Aid. (Likewise, we also give **population_recieving_loan, population_recieving_pell,** for these and a few other measures, showing the number count of students in each category.)

4. Affordability and price. Here we give tuition (**instate_tuition, outofstate_tuition**), **attendance_cost**, and **avg_net_price**, which is the most accurate of the three. Average net price includes the full cost of attendance (including tuition and fees, books and supplies, and living expenses) minus federal, state, and institutional aid cumulative for all students.
   a. Importantly, price is also broken out by income bracket quartiles -- five different columns that show how the real price of attending school differs depending on family income. (See **page 11 of the Technical Paper** for more on net price calculations.)
5. **Debt**: The median debt accrued by students, broken out by income bracket.

III. Outcomes

This data is like individual player stats' on each football team. Gathered largely by tax records and the student loan database, the outcome measures aim to quantify students' performance, and thus, the schools'.

1. **Completion_Rate** measures what percentage of students graduate within six years (three years, for two-year institutions) of starting at a school. This data is broken out by race **(for four year institutions only**) but not income brackets because of reporting errors with the NLSDS completion data.
   a. **Important:** See the caveat guides for the limitations in completion data for non-four-year institutions.
2. **Repayment_Rate** is reported six years after students began attending schools. Repayment is broken out across students who graduated, did not graduate, Pell Recipients, not Pell recipients (which is a proxy for federal loan recipients), family income level, and whether or not they are first-generation college students.
3. **Median_earnings** measured 10 years after entering school and broken out by income bracket terciles. (**Important**: note these income bracket terciles are not the same quartiles as used in some other measurements, so direct ratios will be less exact.)
   a. **Earning_above_threshold**: is the rate of students earning above the $25,000 threshold, which was chosen since it approximately corresponds to the median wage of workers age 25 to 34 with a high-school degree only.

Steps for Analysis

1. First, read what other stories have been done using the scorecard data. We'd recommend ProPublica's and the Wall Street Journal's coverage.
2. Before you start tackling the data, make sure -- again -- that you've read over the documentation. For good points of comparison and for context and possible trends within the data, refer to the **technical** and **policy papers**. They're full of useful information for how your city, state or reporting area might figure in nationally.

   Some examples from the documentation:

   > "Among the 10 percent of four-year schools with the lowest earnings, more than two-thirds of students are from families with incomes below $30,000, whereas, in the top 10 percent of four-year institutions, nearly the opposite is true with roughly one-third of low-income students. This gives the impression that factors associated with family income differences may be partially responsible for the differences in student outcomes across these institutions." (Technical Paper 46)

   > "At 53 percent of institutions, more than half of alumni are not even earning more than a typical high school graduate within six years after starting at the school." (Policy Paper 15)

3. The first thing you should do when starting with the data is narrow down which columns you're definitely not interested in, and make a copy of the data without including those columns. At 78 variables wide, the NICAR Scorecard Data is substantially more wieldy than the raw form from the Department of Education. Still, you'll have an easier time if you analyze step-by-step. For instance, it may make sense to pick just one outcome measure at first, and see how different students or types of schools stack up against one another.

4. Annie Waldman, an education reporter at ProPublica, said she starts developing story ideas at this early stage by running a summary analysis on the columns that most interest her. In knowing the mean, median, highs, lows and quartiles of each numeric field, you can start developing some empirical statements about some of the demographics represented in the columns.

   In Excel, that's:

   ```
   =average(x:y)
   =median(x:y)
   =max(x:y)
   =min(x:y)
   =quartile(x:y)
   =quartile(x:y)
   =quartile(x:y)
   =quartile(x:y)
   ```

5. After running each field you’re interested in, write a sentence or two about what you might be able to conclude from further analysis. How do the median test scores stack up in the for-profit private schools versus the non-profit? How do students’ repayment rates[2] compare at small colleges versus the big university in your city? And those of the Pell recipients versus those who didn’t receive one? You’ll probably come up with more questions than you’ll answer, but simply understanding the summary facts about each field will help you start.

6. Next you should parse what general observations you just made. As noted in the **NICAR Caveat Guide**, any sort of statements made in broad strokes has to be taken with a grain of salt. So a good way to take your analysis to the next level is to filter the outcome data for specific types of schools, sizes, geographic locations, degrees awarded, or any of the other categorical data you might be able to use to group certain schools together. You’ll need a database manager such as MySQL, Access or SQLite.

   Here’s an example of the type of script you might run to see how students at the medium-to-large public colleges in Texas repay their loans, how their debt levels compare, as well as the initial price it cost to attend broken out by different income brackets:

   ```
   SELECT  UNITID, INSTNM, undergrad_enrollment, state, region, type, degree,
   campus, branches, median_family_income, first_bracket_net_price,
   second_bracket_net_price, third_bracket_net_price, fourth_bracket_net_price,
   fifth_bracket_net_price, repayment_rate_nonpell_recipients,
   repayment_rate_pell_recipients, repayment_rate_lowincome,
   repayment_rate_midincome, repayment_rate_highincome,
   FROM NICARscorecard2016
   WHERE state= "TX", undergrad_enrollment >= 2000, type=”Public”
   ORDER BY undergrad_enrollment
   DESC;
   ```

   An analysis like this not only lets you compare different students at the same school, but also different student types (say, low income) across schools of similar size in the same state.

[2] From ProPublica’s Debt by Degrees, “Behind the Data”: “The cohort default rate has long been used as an indicator of how well students are able to pay off their loans after graduation. But schools can manipulate this rate by pushing their students into deferment or forbearance. The repayment rate ... is a better indicator of how many students are struggling to pay off their student debt, as it includes default, deferment, forbearance, as well as students who are unable to start paying off the interest on their debt.”

5. From there, you can start reporting out any trends or surprises you find in the data. ProPublica offers a comprehensive guide to reporting out from the scorecard. Look at how schools in your state are appropriating their (often constraining) budgets. For example:

> Florida reduced its per-student appropriations by 32 percent from fiscal year (FY) 2008 to FY 2014, and tuition rose 53 percent over that time period.16 At the same time, fewer public institutions are helping make up the difference in costs for low-income students. Many public colleges and universities—including well-resourced ones—are reacting to budget constraints, contracting enrollment, and college rankings that emphasize spending over outcomes by diverting their institutional aid to attract high-performing students, which can drive up costs without improving quality. (Policy Paper 11)

What sort of programs do they prioritize? Many, you might learn, place a premium on performance instead of accessibility, or completion instead of earnings. Understanding the mission of individual schools is important context for this data.

6. In addition to the Policy Paper, there are also several studies that discuss the data's implications, impetus and possible explanations for some of your questions. See **page 81 of the Technical Paper** for a list of those studies.