Using IPEDS to document institution-level variation and trends in higher education

Memo for Axel Morales Sanchez

Written by Marshall Steinbaum

February 29, 2024

The purpose of this project is to use institution-level data from IPEDS (Integrated Postsecondary Education Data System) to document trends in the political economy of higher education. IPEDS covers the years 1984-2022, but it becomes more informative over time in general. That’s good news generally, but it also means that there are variables that come into existence part-way through, and others that disappear.

There is a tutorial called [Overview of IPEDS Data](https://nces.ed.gov/ipeds/use-the-data/overview-of-ipeds-data) on the website of the National Center for Education Statistics (the agency that creates IPEDS). It’s probably worth going through its five modules, although you may also want to return to them once you get your hands dirty with the data. The complete data files are [here](https://nces.ed.gov/ipeds/datacenter/DataFiles.aspx?goToReportId=7&sid=fea8ae80-53ed-4207-ae64-780911b658f8&rtid=7). For each year of the survey, the files are divided, basically by type of variable (institution characteristics, enrollment, completion, financial aid, faculty and staff, etc.) So for each of the analyses below, you will be downloading and analyzing different pieces of the overall IPEDS database. Each file/group of variables has an accompanying data dictionary.

In fact, what I want you to do is create a master IPEDS panel dataset on the University of Utah server, in which each row is an institution-year combination. By “institutions,” I mean colleges and universities, which are very diverse institution types ranging from enormous state universities to small undergraduate-only colleges to for-profit certificate-granting trade schools.

In addition to the data visualizations, a very important part of this project is to produce code that’s replicable. I’d like you to do this programming and data analysis in Stata and write the code so that it can easily be updated when new years of IPEDS data are published. My collaborator Sergio Pinto has code libraries for analyzing IPEDS that I will share with you as a starting point, and he is a very conscientious and careful coder so his model is a good one to follow. His Stata do-files are in the directory S:\steinbaum\_share\IPEDS-CS\IPEDS-Sergio.

**National IPEDS Sample** (i.e., in principle, all institutions in the US)

The object is to build a panel dataset at the institution-by-year level, from which all of the following charts can be created. Each variable should be a different column of the data. If there’s no information for a given variable for an institution in a given year, then the value for that variable corresponding to the institution-year should be missing. But in principle, the panel is balanced in the sense that there is an observation (row of the data) for every combination of institution and year. The only reason the panel would not be balanced is that the institution didn’t exist/didn’t report information to IPEDs in a given year. In those cases, you can “fill out” the panel by having institution-year observations where every variable except for institution, unitid, and year has a missing value.

Once you’ve created the panel data, I’d like you to write the code to create each of these charts in Stata. Using one Stata master do-file and individual sub-do-files for each section of the memo (similar to the setup Sergio is using), bring the entire panel dataset into memory at the beginning of each of the sub-do-files and then make the chart(s) within that section from there. Imagine that I will be adding other do-files to your master to create other visualizations, and it should be possible to run the master do-file by commenting out all but the sub-do-file(s) you want to run to create a given visualization.

For the actual visualizations to be created in Stata, use the “cleanplots” “scheme.” So you should have a line of code in the master do-file that says set scheme cleanplots. Below I make some comments about formatting the visualizations, which are based on the ones you’ve copied over to the Utah server that were generated in R or Python. Once you’re using Stata, those defaults (like scientific notation) may no longer be relevant. The point is: use whatever the cleanplots default is for each Stata plot type, then tweak formatting from there.

The individual institution identifier in IPEDS is unitid. Each variable in the panel dataset should be given a meaningful label in Stata, i.e. more than just the IPEDS variable name.

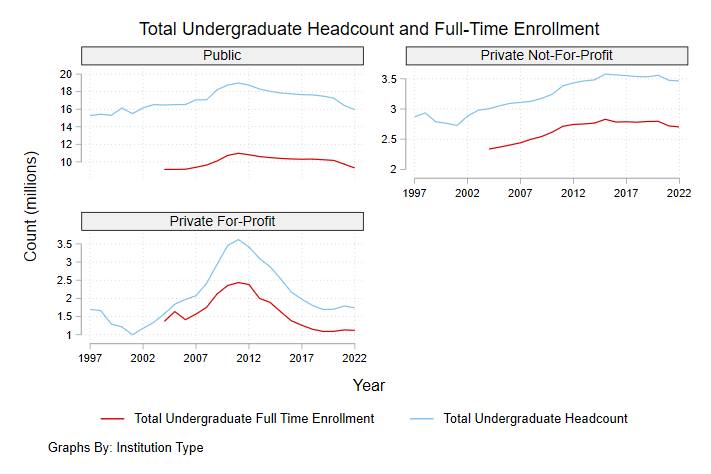
1. Time series of undergraduate/graduate student enrollment by institution type

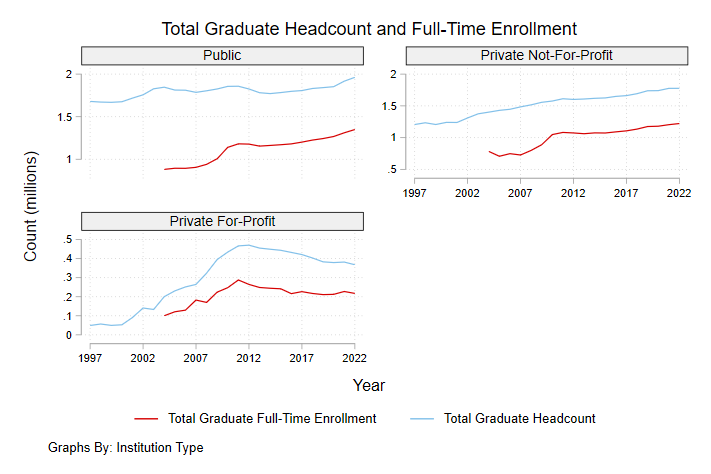
For each institution reported in IPEDS, count the number of enrolled students at the undergraduate and graduate level. Assign institution type (private nonprofit, private for-profit, public) to each institution. Sum across institutions within a given year and type, and plot that as a time series. This gives rise to six figures (graduate/undergraduate and three institution types), plus two more summing enrollment across types.

The institution type is reported with the variable control. If that variable is missing for some institution-years, you can assume that the same value can be inferred from future observations.

The enrollment count variables are fteug xfteug ftegd xftegd ftedpp xftedpp for the most recent years. (“FTE” means full-time equivalent). Before FTE enrollment counts are available, I think IPEDS reports head count enrollment (i.e. not taking into account credit hours). For each time series chart, I want you to include all the years for which there is data, but if it’s not actually the same variable being plotted, use a different color scheme. So, for example, if you have headcount enrollment from 1984 to 2000, and then FTE enrollment from 2001-2021, then you’d have two different color line plots all on the same set of axes.

IPEDS does both fall and full-year enrollment counts. For the purpose of this item and the long series, use full-year enrollment counts. For the California-specific analysis below, however, you’ll use both.



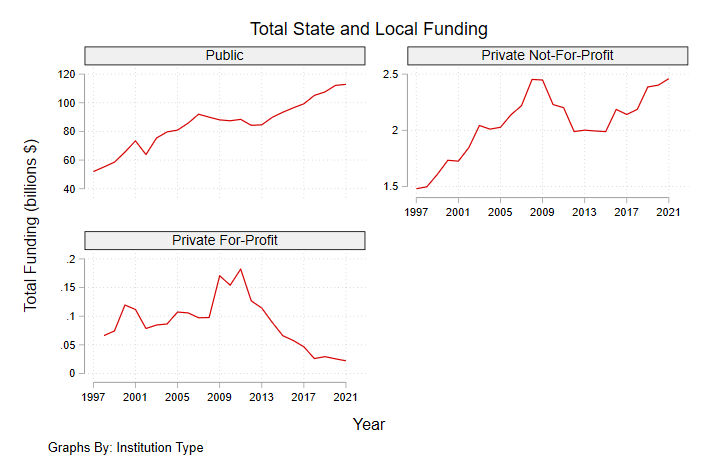


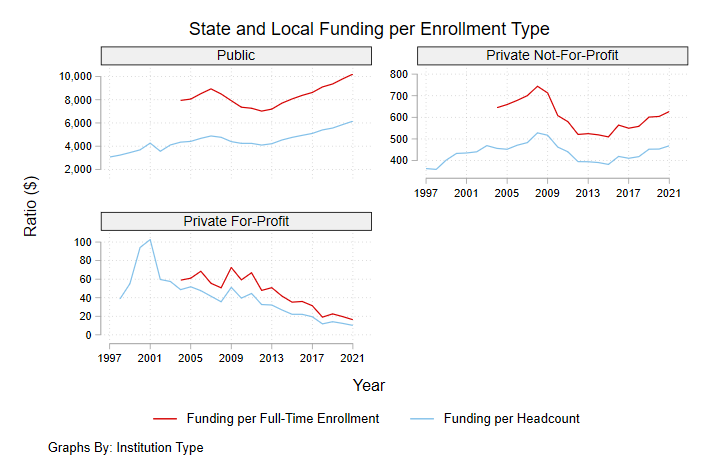
1. Time series of state and local funding and state and local funding per FTE.

For each institution-year, sum the financial variables that report institutional funding from state and local governments. There are a number of variables in the “Finance” category whose names begin with f1, f2, and f3. For example “state appropriations,” “state grants and contracts,” “state nonoperating grants,” and then equivalent ones for “local.” The state appropriations ones should be zero for the private institutions and positive for the public institutions. Private institutions might get state grants and contracts and nonoperating grants, i.e. other forms of state funding, but not appropriations.

Once you have the sum of the state and local funding variables at the institution-year level, divide that by the total FTE enrollment (undergraduate and graduate) computed in (1). Plot both the total state and local funding in dollars over time, and the ratio of funding to FTE enrollment ($$ per FTE). Don’t label the vertical axis with “G” since it’s not clear what that means.

Also make equivalent charts in which the denominator is headcount enrollment, not FTE (and which will thus go back farther in time).





1. Layer cake chart of institutional funding sources over time.

For each type of institution (public, private non-profit, private for-profit), create two layer cake charts of institutional funding sources by broad category over time by summing over institutions for each category. The two charts are in dollars (so the total will rise over time) and as percentages of the total (so by construction they sum to 100%)

The categories are

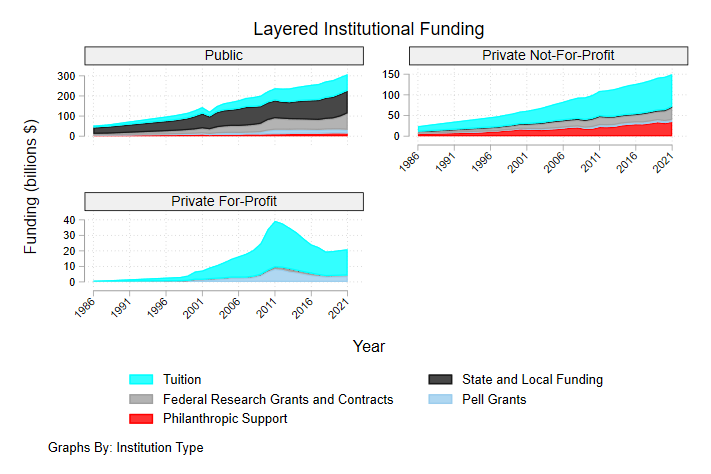
* State and local funding (as in 2 above)
* Tuition, net of financial aid
* Federal research grants and contracts (not including Pell)
* Pell grants
* Philanthropic support (“gifts, including contributions from affiliated organizations” and “capital grants and gifts”)

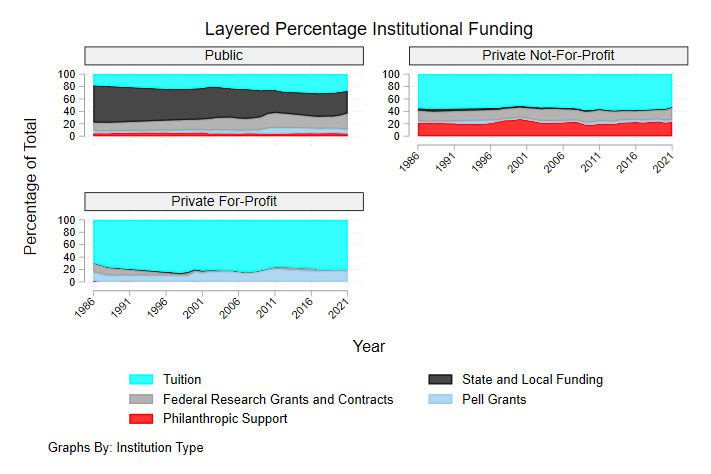
[These are non-exhaustive—they don’t have to add up to the total revenue field, since we’re ignoring some smaller categories. For the percentage version of the layer cake chart, the denominator should be the sum of the things we are counting, not the total revenue.]

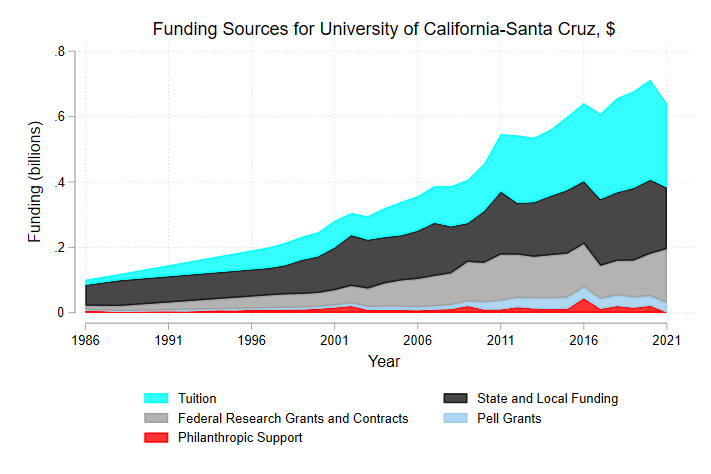
For these charts, please also create a functionality that plots the equivalent figures for a single institution over time. So there should be a line of code that says keep if unitid == XXXXX, where you can select the institution to be plotted. Then you create each of the two budget figures just for that institution, i.e. not summing over institutions (which is no longer possible since only that institution’s data will still be in memory).

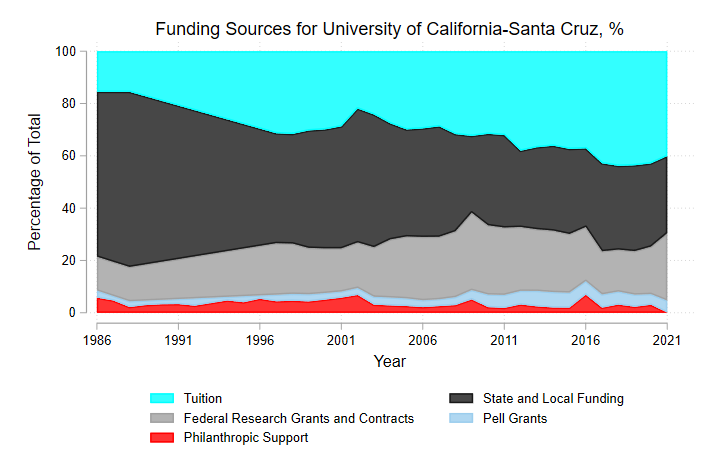
On formatting the charts:

* When the vertical axis is in dollars, please don’t use scientific notation. If it’s in billions of dollars, label the axis “$bil” or something like that and just use numbers. You can also use “M” and “B” on the axis tick labels, since it’s clear what those mean, but not “G.”









UPDATE 2/7/24: For the single-institution funding layer cake charts, change the chart titles to “Funding Sources for \_\_\_\_, $” and “Funding Sources for \_\_\_\_\_, %” where \_\_\_\_ populates with the name of the institution, given the UNITID you enter to create the chart.

In these examples, I’m worried about the temporary appearance of the Pell segment in only a few years. I would have thought that Pell grants were broadly constant over time as a source of funding (many institutions have reduced the number of Pell-eligible students they admit, but the total number of Pell-eligible students in the system is increasing, and more to the point, within any one institution there won’t be much variation over a short period of time). Hence, that makes me worried that somehow the code that looks at Pell funding to create these charts is doing so inconsistently, i.e. that variable is called one thing in some years and something slightly different in others.

1. Number of professional masters programs since uncapping of graduate loan limits.

This is going to be tricky to figure out in IPEDS data. To state upfront what we’re going for: in 2006, loan limits for grad-plus student loans were uncapped, which created a windfall for institutions to offer what I call “credentializing masters programs,” i.e. a graduate degree tied to a profession/career that previously wasn’t necessary to advance in said career. These are cash cows for universities, and a major reason for the increase in high student loan balances that increase over time (because borrowers who ‘start’ with high balances from such a credentializing masters program tend to enroll in Income-Driven Repayment programs since salaries are nowhere near high enough to repay the debt, so it keeps rising from there).

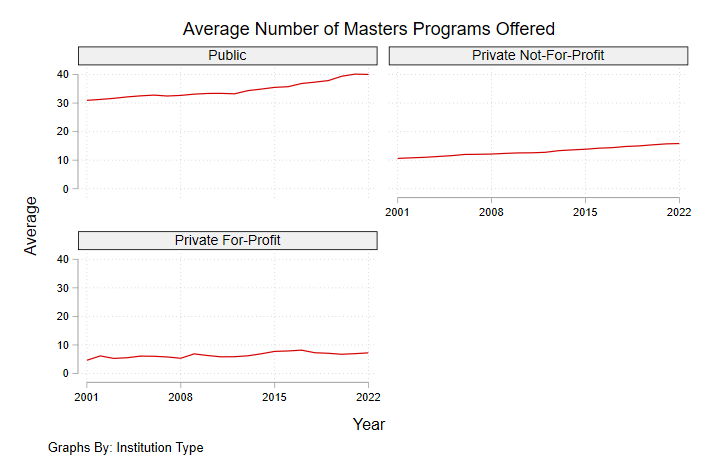
There’s something called a CIP code, which is a taxonomy that governs programs of study. I have to look into exactly which ones we’ll call credentializing masters programs. For whatever set of programs that is, we want to track two things:

* The number of them offered at the institution level. From my read of IPEDS documentation, this information is not directly available. Instead, we do get graduation and possibly enrollment counts by CIP, so we’ll want to measure the distinct number of credentializing masters programs by taking the number of such CIPs that have any graduates/any enrollment, by institution.

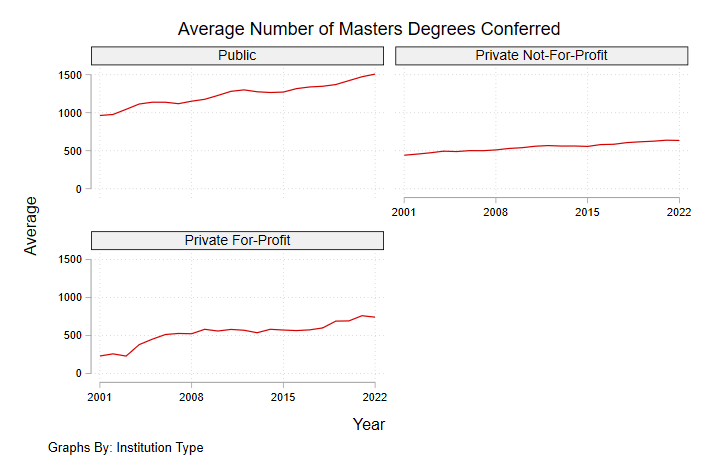
2/7/24: It seems like a “0” entered into the degrees-conferred variable indicates that a program is offered, but no degrees were conferred. Hence, number of masters programs offered is computed based on non-missing values for degrees conferred. – This behavior seems to only be well defined for years >1999. Although on the actual survey forms, the instructions given to institutions were to fill a 0 if no degrees were conferred at an award level for a program if it is still offered in that school year, for years 1984-1989, 1991, and 1999 there are no programs that conferred 0 degrees, which seems unlikely. Furthermore, for the year 1998, there are an abnormal amount of institutions (~600 more than the previous and next year) in the completions data. Hence, only including years >1999 in the graphs for now.

I will try to find programs offered data directly.

* Number of masters degrees conferred per institution.
* Total enrollment in all masters programs, average per institution .



For “Number of Masters Programs Offered,” compute the average per institution. So the numerator is the number you’re currently plotting, and the denominator is the count of UNITIDs in each control-type-by-year cell.



Likewise, you should plot the average number of masters degrees conferred per institution.

\*\* TO COME FROM MARSHALL: We will look into using PSEO to plot earnings for masters degree holders over time.

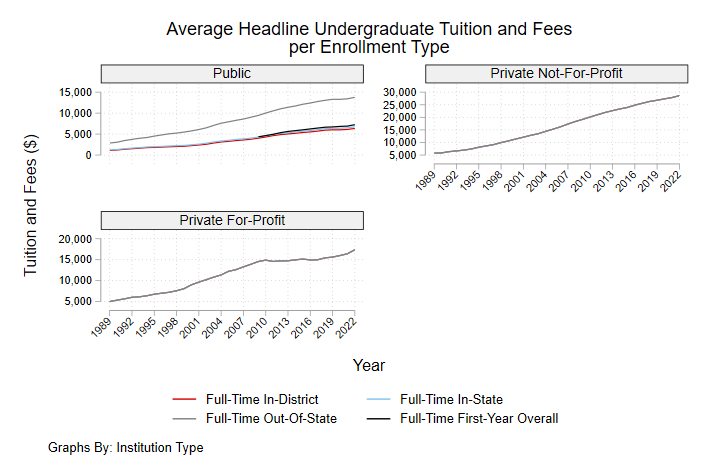
1. Tuition cost per FTE over time

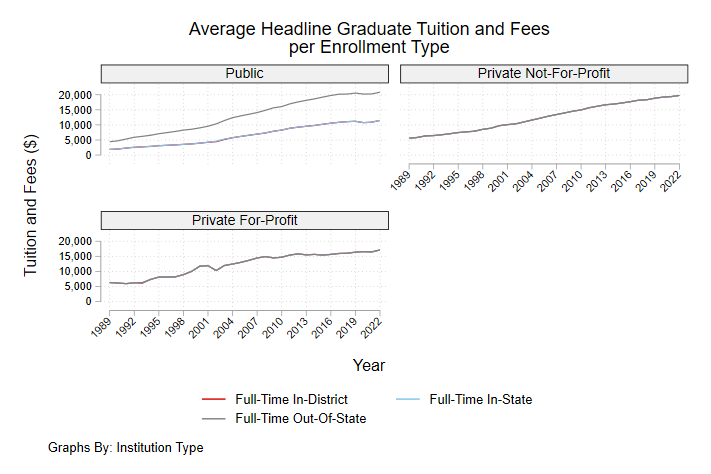
The total “student charges” (tuition as well as fees)—the numerator for this item—are contained in one of the variable blocks, for full time/part time undergrads and all manner of academic and professional programs. For this, we want a time series of the average student charges per FTE for each type of institution (public, private nonprofit, and private for-profit), over time. For publics, IPEDS differentiates in-state and out-of-state tuition. Most importantly, this item is about tracking what’s called the sticker price, as opposed to the net price (what students actually pay), which is covered in the next item.

The denominator for this item, total FTE enrollment by degree program, would come from item 1.

For this item, focus exclusively on what IPEDS calls academic programs, as opposed to vocational ones.

In addition, I’d like to be able to produce this time series chart for a single institution, given an inputted unitid. For those figures, the heading should be “Tuition per FTE at \_\_\_\_”





1. Charts of net tuition and enrollment by parent/household income by institution, over time.

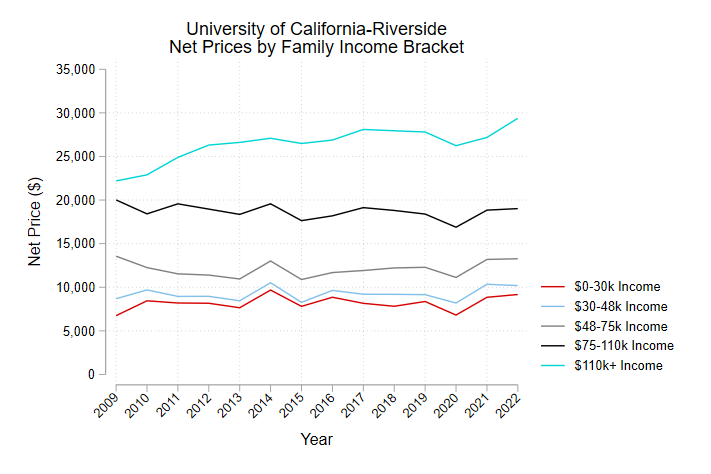
I want you to replicate figures that look like this, which pertain to institution financial aid and access policies. [These figures were created by a higher ed researcher named Brett Herda.] The left panel plots “average net price” by family income bucket over time, and the right panel plots the enrollment count by family income bucket. Both sets of variables are part of the “student financial aid and net price” set of IPEDS variables. In addition to these plots, create a time series of the overall net price for each institution (not conditional on family/household income bucket).

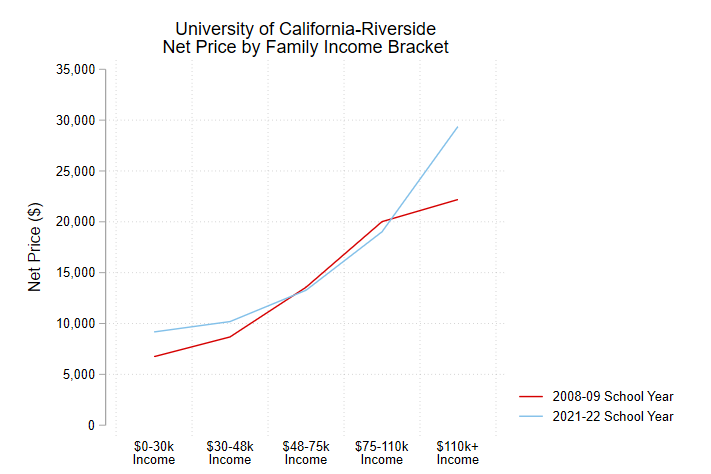
Chart

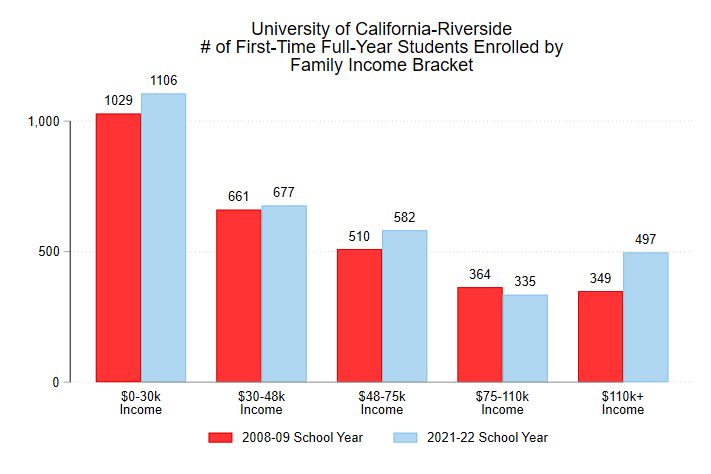
Description automatically generated

Southern New Hampshire University is an extreme case where they actually charge higher tuition to students with lower-income parents. I suspect for most institutions, what you would see is that net price has become more flat over time, but not actually downward-sloping (as in the left panel above for 2018-2019). In addition, public state flagship institutions will show declining enrollment of the children of low-income parents.

I want you to write code that does this institution-by-institution the way it’s done in the figure above for SNHU. For example, if I wanted to generate the same figures for UC-Irvine, I could simply enter UC-Irvine’s unitid, as well as the academic years I want to plot, and it would pop out the equivalent two panels. In addition, create a third plot type: the time series of net tuition for each parental income category.







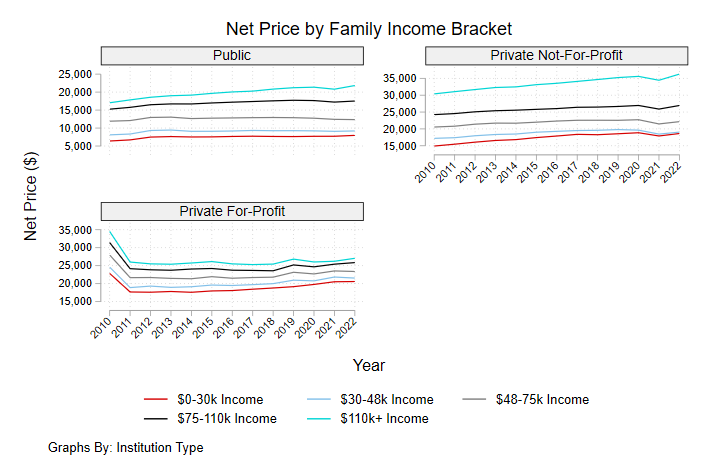
**UPDATE 2/29/24**: We now have three “financial aid panels”: the two above from the Herda chart, and the annual time series of net price by family income bucket. I would like you to create versions of all three panels for the following three breakdowns:

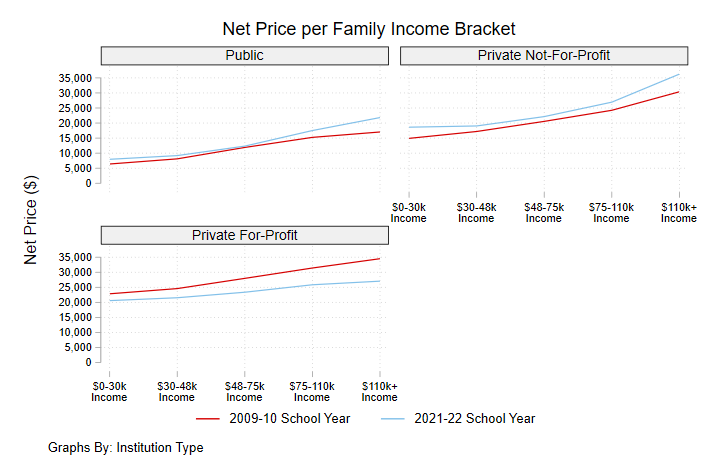
1. All publics/all private non-profits/all private for-profits.
2. Every state individually, public universities only (50 states x 3 charts per state, aggregating over all public universities within each state).
3. Same as (2), except only “R1” universities according to the Carnegie Classification. According to [this page](https://nces.ed.gov/ipeds/use-the-data/institutional-groupings-in-ipeds#fn2), that variable should be available in IPEDS.
4. HBCUs (likewise, I believe there is a variable in IPEDS that designates that)

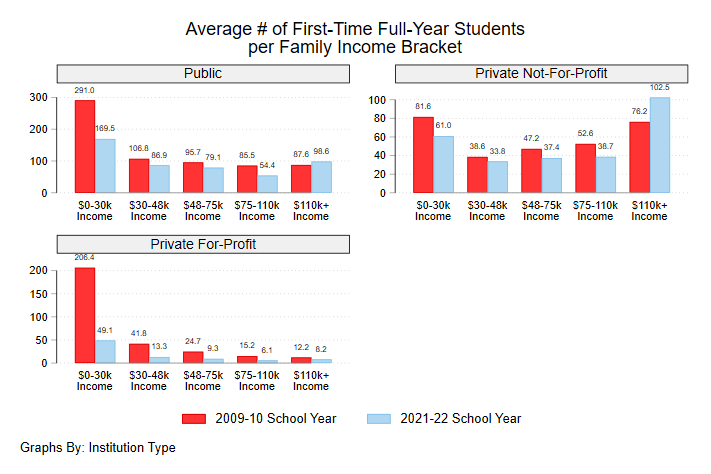
For the panel that shows the net price by family income category in two academic years, use the most recent available rather than 2018-2019, as in the Herda chart.

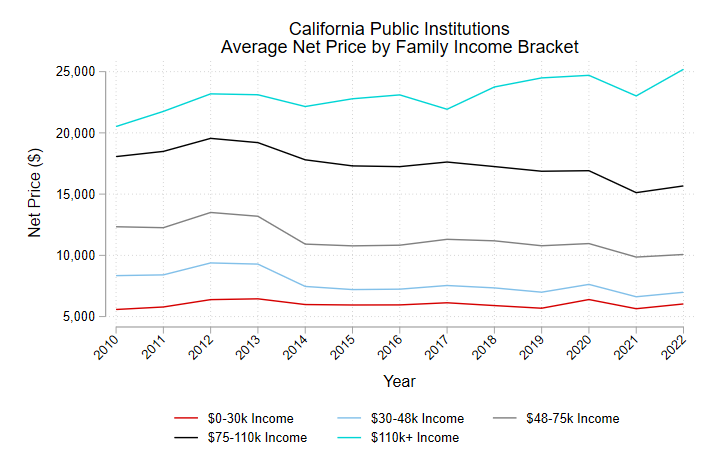
UPDATE: 3/15/24: For the year 2009, privates did not report data for 110k+ income net price, but they did report enrollment counts for 110k+ income. Overall this has the effect of artificially decreasing the weighted average net price for the year 2009 when aggregating by R1 and HBCU universities, as well as aggregating over a state. Due to this, I decided to plot starting from the next best year 2010, as the artificially low average for 2009 is especially deceptive in the two-year comparisons for those respective aggregation levels, as well as in the time series.

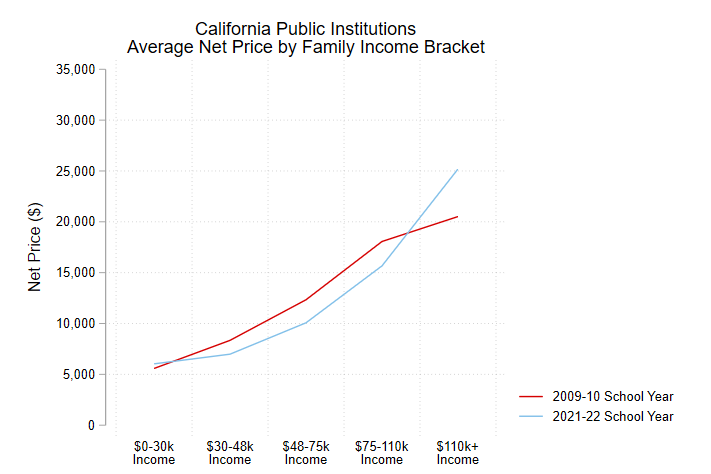
A potential fix would be to set private enrollment counts for that income bucket and year to missing when enrollment counts are used as the weights, then resetting back to the original value after the weighted mean computation. This of course will have the effect of reporting the average net price at that aggregation level for that year as just the weighted average of publics. Another fix is imputation.

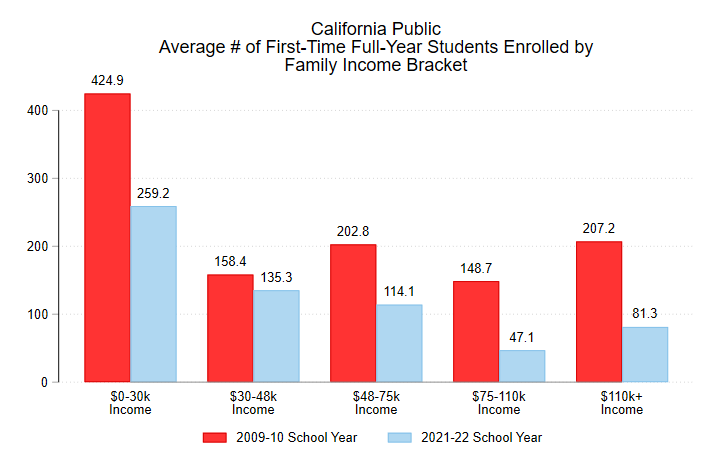


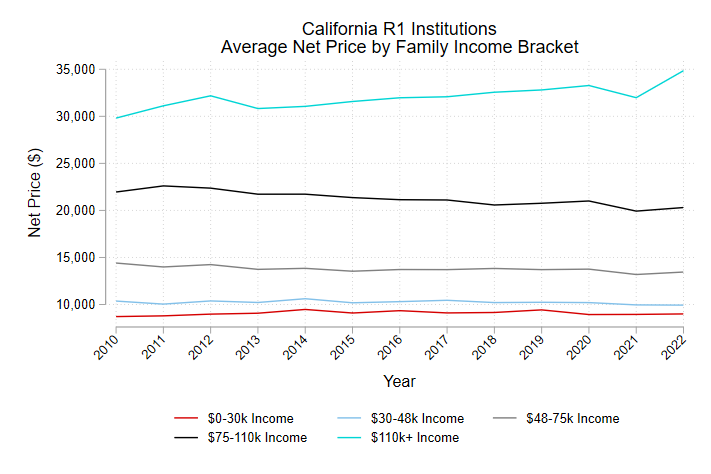


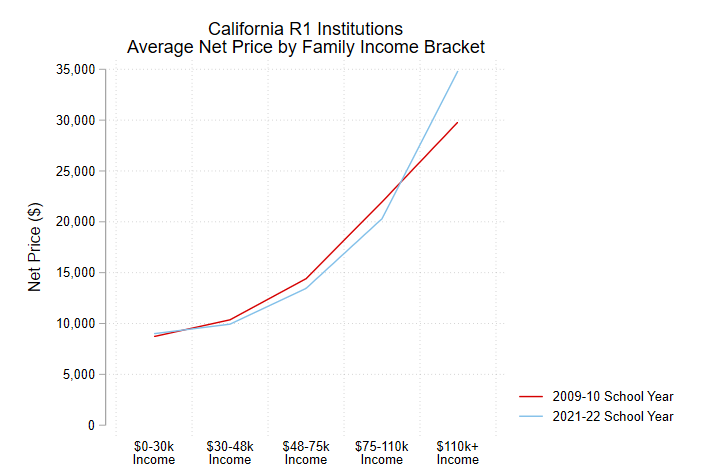


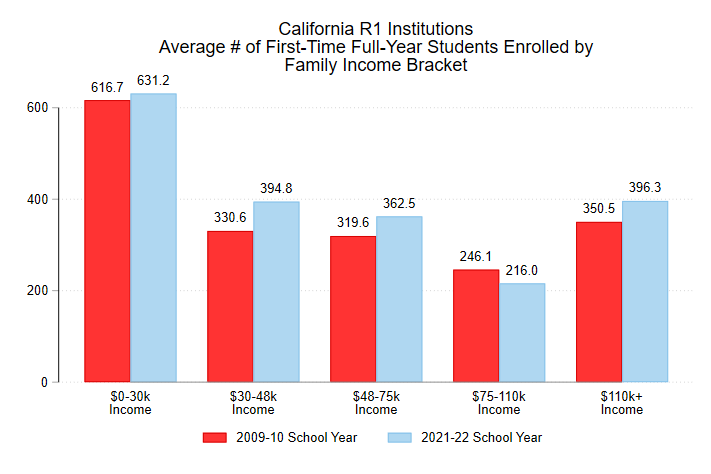


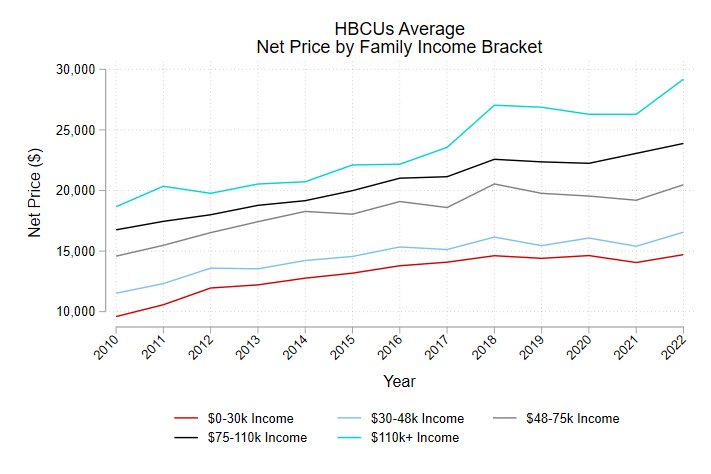


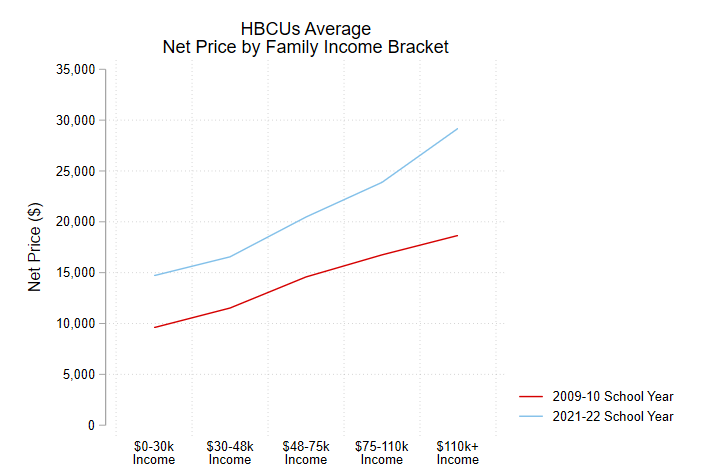


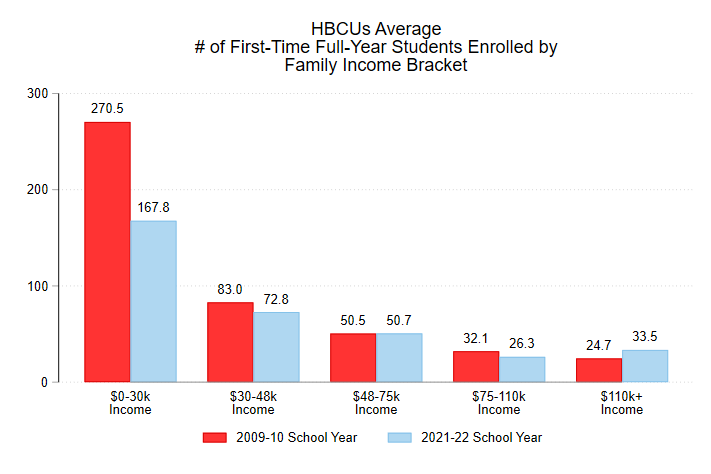


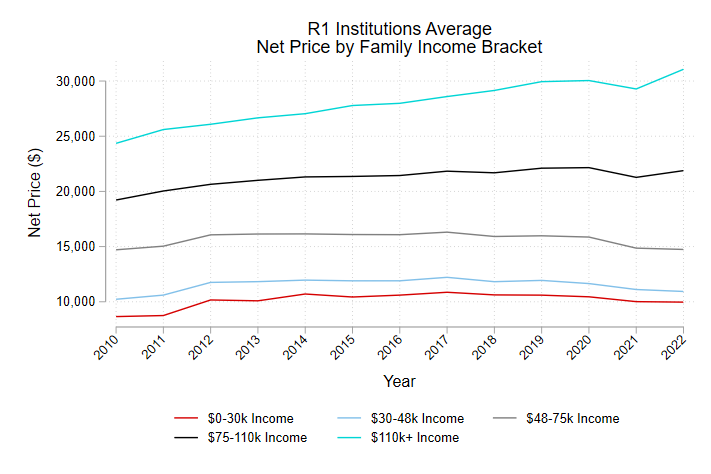


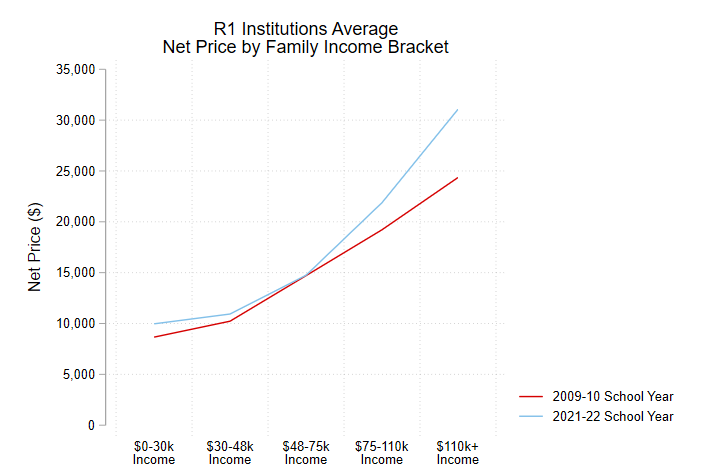


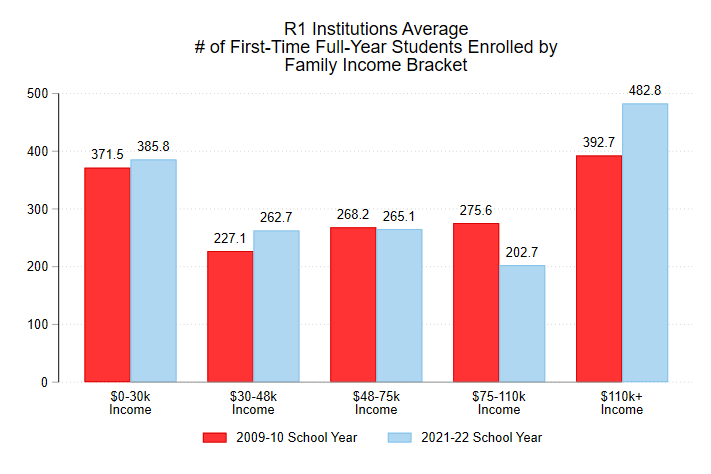












**California IPEDS Sample for use with the UC-CCP**

We also want to isolate institution-level variation in aid and other policies for California institutions only, for use with the California sample of UC-CCP. We also want the other 49 states as possible control groups. Hence, create a panel data extract from your overall IPEDS panel at the state-by-academic year level.

The net price for all in-state and out-of-state students at California institutions for 2016-2017, 2017-2018, 2018-2019, 2019-2020, 2020-2021, and 2021-2022 (the final year available), as well as every other state. Hence, you want a (short) panel dataset that has each of the 50 states as the panel ID variable and the time dimension is six academic years.

We’re particularly interested in the change between 2019-2020 and 2020-2021, for the purpose of evaluating the effect of pandemic-era financial aid policies. (This is already visible in the four UC plots you made.)

The variables are:

* Net price for each of the five family income buckets.
* Enrollment counts for each of the five family income buckets (use the average per institution, so you’re summing the enrollment count for a given family income bucket across institutions within a state-academic year cell, then dividing that count by the count of the number of institutions you’re summing over.

UPDATE 3/15/24: See 14 - IPEDS - Create State-Year Level Panel Data for details on how I made the dataset. In summary, I computed the weighted average net price at the state-year level for each income bracket, where the (frequency) weights are the enrollment counts in the corresponding income bracket. I also computed the simple average of the enrollment counts for each income bracket at the state-year level (i.e. sum of enrollment counts for a given bracket at the state-year level, divided by the corresponding frequency).

**To come**

1. Vocational programs:
   * Institution-level offering
   * Tuition
   * Federal funding for apprenticeships.
2. PSEO Earnings by masters degree holders.