The code for this project can be viewed here

When I married my wife, we started our marriage off with over 50K in debt. Based on various statistics I've heard over the years; our debt levels were typical for college graduates. We've been fortunate – steady employment and extended, tough sacrifices over the years enabled us to pay it off and enter a debt-free life, but we know that isn't typical of people in their 20s.

This formative experience motivated me to pick up the College Scorecard Dataset from the US Department of Education and see what I can learn about student debt – what schools it comes from and who's most likely to graduate with a lot of it.

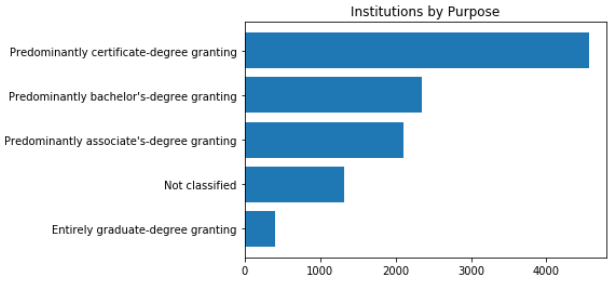
Data Description:

Each line on the college scorecard represents a *branch* of a higher education institution. I use the word **branch** when I report them separately**,** but **institution** when I group affiliated branches together.

Hundreds of columns then follow describing various performance metrics specific to higher education. These data come federal reporting form the institutions, federal financial aid data, and tax information. The data does not claim to include private loans outside the federal financial aid process.

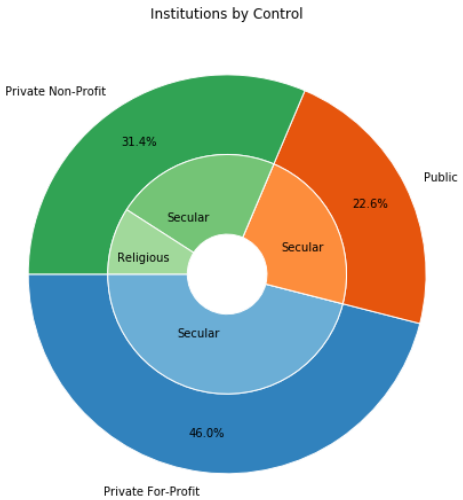
I isolated the columns I wanted to study – specifically:

- Average net price per year, and the median debt for students who separate from the branch, by year
- The count of students included in those measurements
- Various institution demographics
- various student demographics for each branch



Notable observations, some as seen in the figures to the right:

- Almost half of the 9,091 institutions analyzed are predominantly certificate granting
- Despite the performance metric sources being primarily federal in nature, less than 23% of the branches are publicly controlled. 31% are Private Non-profit, a small portion of which are Religious. The largest category is secular for-profits – 46% of the institutions graded.



Pre-Processing

Initially, the college scorecard was too large for my machine to import to a notebook. Fortunately, I was able to open the file in excel to reduce the size. I removed all data prior to 2004, leaving 10 years for analysis. I also removed the columns relating to things I wasn't interested in studying for this project:

Completion, Dropout, and Transfer rates, demographics besides the ones you'll find below, figures related to specific types of loans or grants, repayment rates, and other redundant metrics.

Once the file was importable, additional issues surfaced. Some branches had a negative net price; while it may be reasonable to think that occasionally students get scholarships that exceed their tuition and costs and actually earn money by attending school, it's not reasonable to assume that a branch's average price would be negative for the entire student body. These values were raised to zero.

My goal was to be able to answer some demographic questions about the debt students are leaving school with. Branches did not report average debt, but did report **median debt**. They reported the **count of students** included in the median measurement, as well as the same counts split by various demographics. I found that many of the demographic counts did not sum to the total as you'd expect (for example, *"Male_count plus Female_count"* would be expected to sum to the total count, but did not), so I converted the counts to percentages and used those. By multiplying the median debt by the count of students, I could theoretically achieve a **Total Debt** amount for all students leaving the branch in a given year. In practice, the results I was getting were unreasonably large. Particularly problematic were online-only schools, which in some cases unashamedly report a difference branch for each state. The codebook was unclear on this, but after further investigation I concluded that institutions must have reported student counts in aggregate, and each branch was mistakenly given the total. For example, the University of Phoenix, a prominent online school, reported 71 branches and each listed exactly 279,901 students – clearly incorrect.

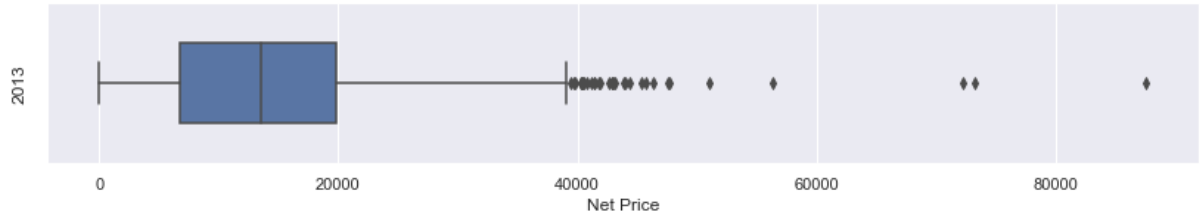
Dividing the student count by branch before multiplying it by the Debt median gave much clearer results for the branch's **Total Debt**. Using the student demographic percentages, I was able to "distribute" each branch's Total Debt across demographics to get average debts for the following categories:

- Male/Female
- First-Generation Student / Not First-Generation Student
- Family Income bracket in annual nominal dollars:
    - Low: Under $30,000
    - Medium: between $30,000 - $75,000
    - High: over $75,000

	studNum	totalDebtForBranch	malePCT	femalePCT	firstGenPCT	notFirstGenPCT	hiIncPCT	medIncPCT	lowIncPCT
count	58,917	58,917	37,522	37,522	55,446	55,446	32,761	32,761	32,761
mean	1,248	14,029,823	35.85%	64.15%	47.43%	52.57%	15.22%	29.19%	55.59%
std	2,066	25,115,605	17.05%	17.05%	12.32%	12.32%	11.97%	8.69%	18.12%
min	3	12,710	0.00%	0.00%	0.00%	4.99%	0.00%	0.00%	7.69%
25%	216	1,578,654	25.55%	56.01%	40.18%	44.46%	6.52%	23.08%	41.98%
50%	595	5,623,408	35.65%	64.35%	49.15%	50.85%	11.37%	29.86%	57.14%
75%	1,351	14,606,250	43.99%	74.45%	55.54%	59.82%	21.14%	35.34%	69.96%
max	101,600	965,200,000	100.00%	100.00%	95.01%	100.00%	79.20%	64.73%	100.00%

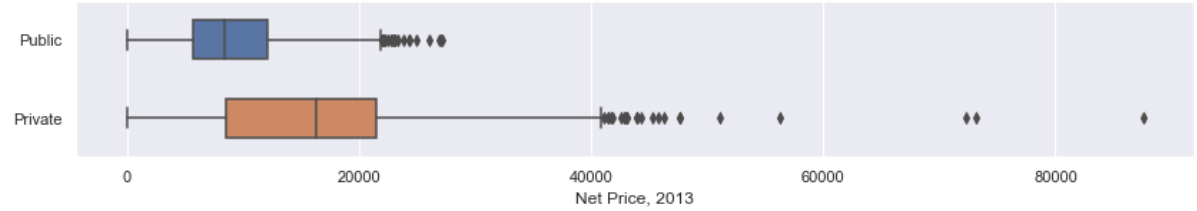
Analysis

One of the many factors contributing to the student debt crisis is the rising price of higher education. Below is the range of net prices (annual tuition and expenses, less average scholarship/grants) for the most recent year of study (2013).

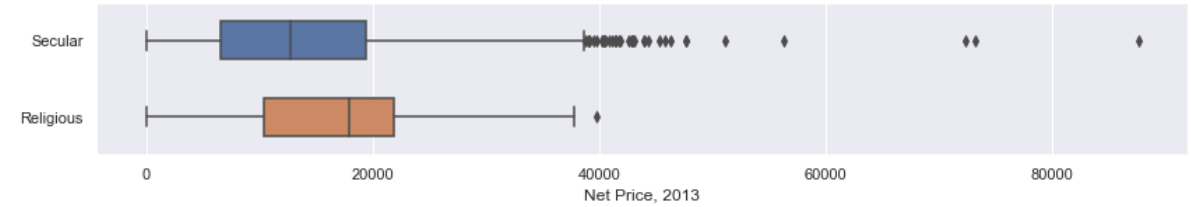


Interestingly, the 3 most extreme outliers are a Photography school and 2 Flight Schools.

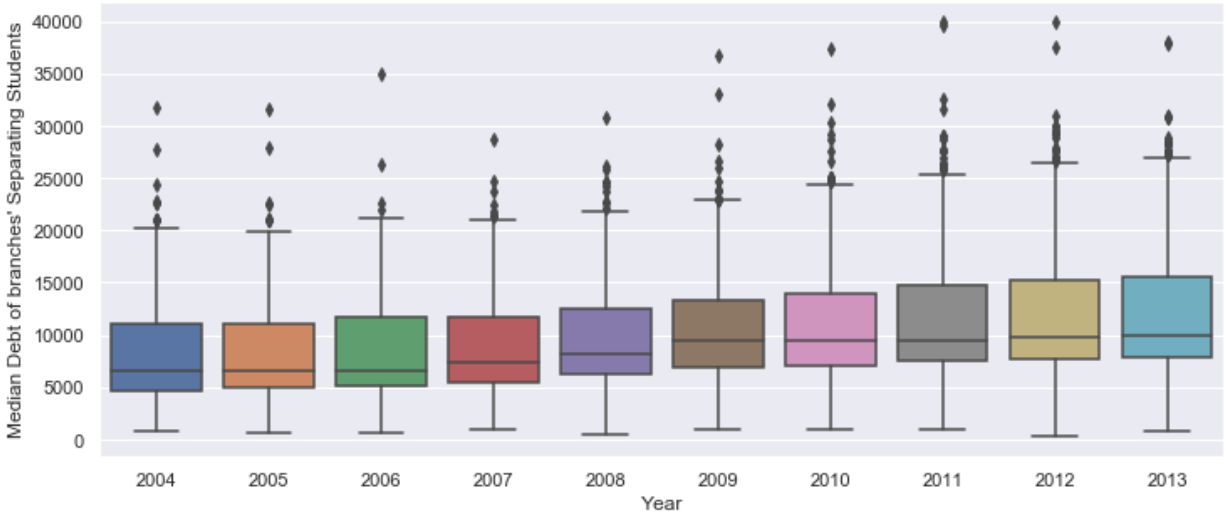
Private Universities are more expensive:



Though it's less decisive, Religious Universities tend to be more expensive than Secular ones:



These price tags, in addition to various other factors, have contributed to a steadily increasing median debt for separating students for at least 10 years:

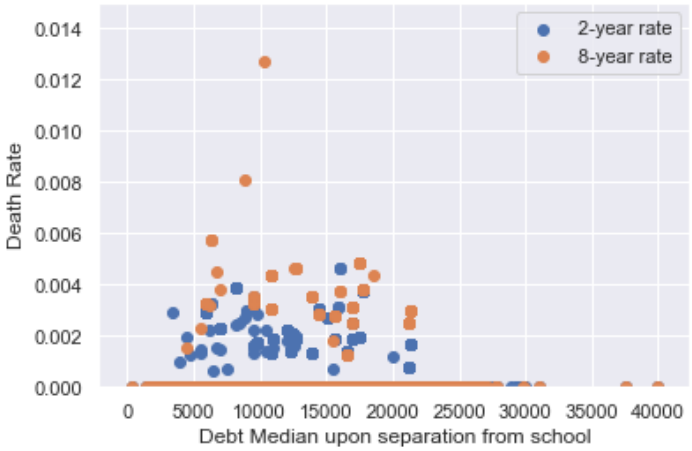


With the debt associated with each branch characterized with my chosen demographics, I set out to answer the following questions:

- Is the median debt of a branch correlated with the death rate of its students?
- Is there a significant difference between the debt balances of males and females?
- Is there a significant difference between the debt balances of first-generation college attendees and those born to college graduates?
- Is there a significant difference in debt balances based on your family's income?

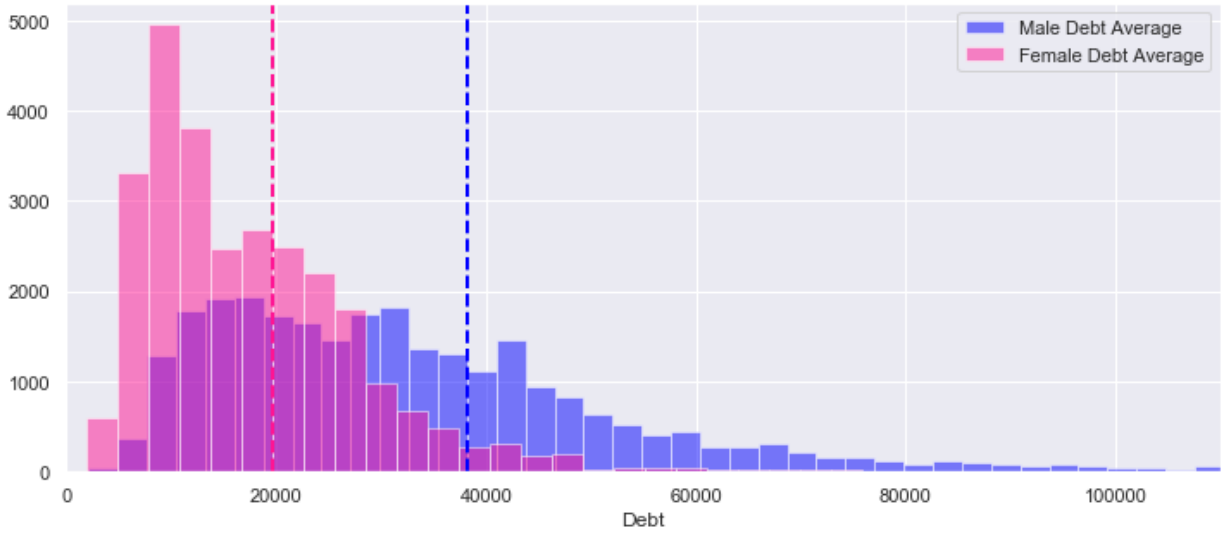
Death:

Death rates were recorded for the students from each branch at the 2-year and 8-year mark. As seen in the plot to the right, there was no found correlation between the median debt for a branch and the death rate of its past students, which would've been a very concerning suggestion that the increasing debt amounts were negatively influencing mental health or some other cause of death. At a minimum, I'm happy to report no meaningful correlation.



Male & Female:

Since each school reported median debts figures for their students and noted how many of their students were male or female, I was able to calculate how much of their *total* debt for that year went to males and how much went to females. By dividing those by the number of respective males/female students, I was able to calculate an average male/female debt. The two distributions of all the branches is seen below:

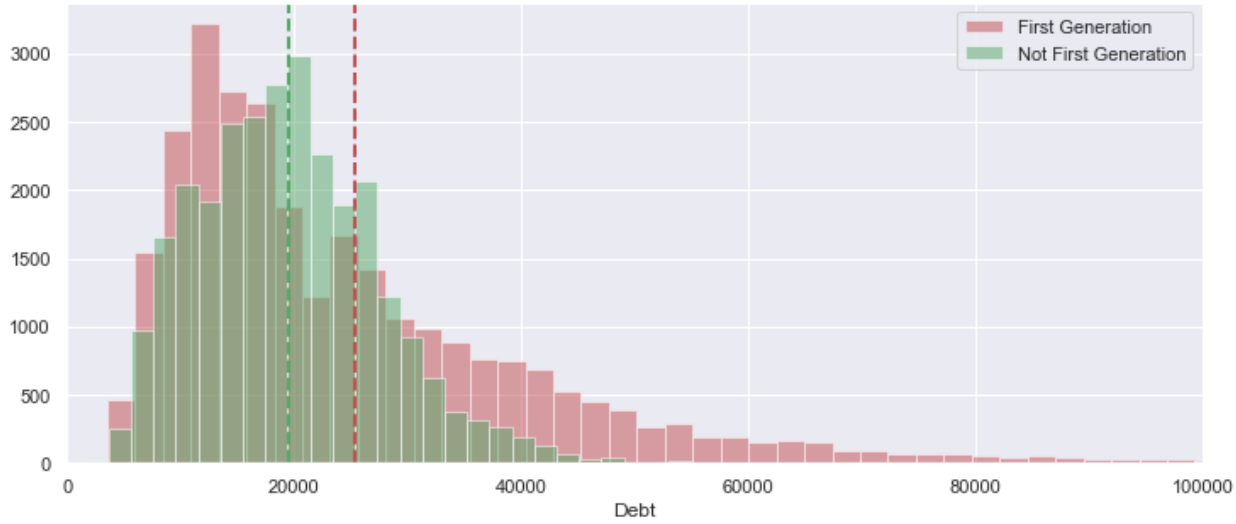


As a reminder, each row in this dataset is a branch of an institution. Therefore, this is a distribution of the average individual debt obligations for the males/females at the branch level. This chart shows that at a majority of branches, men have higher debt obligations than women upon separation (graduation or withdrawal).

First-Generation Students

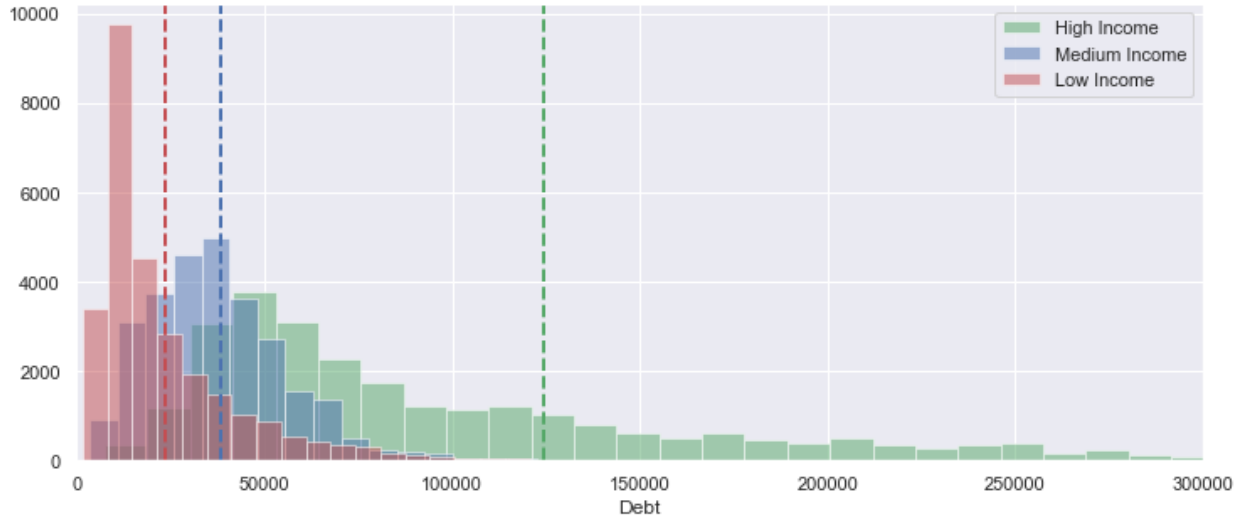
Average debt for first-generation students was calculated using the same methodology as male/female students. The means of each distribution are marked below.

Despite having a lesser mode and median, the mean debt for First Generation students has a long right tail and indicates that First Generation students tend to graduate with more debt.



Family Income

Average debt by family income bracket was calculated using the same methodology.



Perhaps surprisingly, debt appears to be correlated positively with your family's income bracket, suggesting that instead of using wealth to graduate without debt, families opt to use their wealth to get access to even more capital and pursue even more expensive educations.

Further suggested research topics:

- This suggests that men have higher debt obligations upon leaving school than women do.
    - Is this an income problem (men tend to have less resources to pay up front), or a selection problem (men tend to choose more expensive schools), or are there other factors driving this difference? By finding a source of *application* data (as opposed to ultimate enrollment), and by analyzing school prices against sex proportions, this could be hypothesized.
    - Do men leaving these schools have higher earnings or better repayment rates to offset their higher debt balances? The initial dataset had repayment and earnings data that could be aggregated to answer this question.
- This suggests that first-generation students have higher debt obligations upon leaving school than others.
    - Is this an income problem (first generation students come from less wealthy families) or a selection problem (they tend to choose more expensive schools), or are there other factors? By finding a source of *application* data (as opposed to ultimate enrollment), and by analyzing school prices against income bracket demographics, this could be hypothesized.
    - Is there a parental experiential advantage? If your parents had a college experience, and can help you prioritize what is worth taking on debt for and what isn't, is that advice enough to make a meaningful difference? If there is a way to hold constant other factors such as school price and parental income and times spent in school, and still have a meaningful sample, any difference found between debt values might suggest such an advantage.