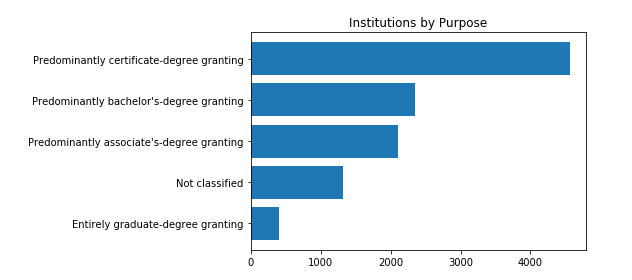
When I married my wife, we started our marriage off with over 50K in debt. Based on various statistics I’ve heard over the years, our debt levels were pretty typical for college graduates. We’ve been fortunate – steady employment and extended, tough sacrifices over the years enabled us to pay it off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

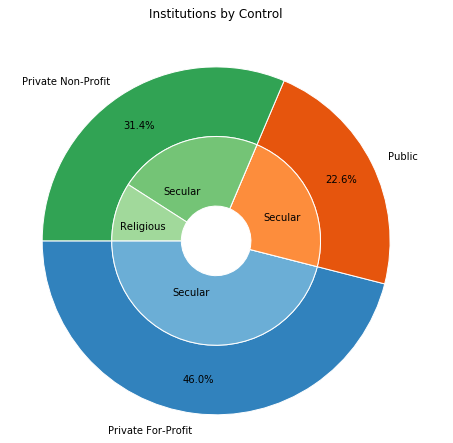
This formative experience motivated me to pick up the College Scorecard Dataset from the US Department of Education and see what I can learn about student debt – where it comes from and who’s most likely to graduate with a lot of it.

**Data Description:**

Each line on the college scorecard represents a *branch* of a higher education institution. Those branches that mutually are part of a single institution (for example, a main campus and its commuter branches) have a UNITID in common, but are rated separately on the scorecard.

Hundreds of columns then follow describing various performance metrics specific to higher education. These data come from a mix of sources: federal reporting form the institutions, federal financial aid data, and tax information. The data does not claim to comprehensively include private loans that may happen outside the federal financial aid process, although in the cases where data come from tax information, it may be that some of those loans are included.

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 (Public v. Private, Religious Affiliations)
* various student demographics for each branch: (Family Income Level, Sex, First Generation/Not, death rates after N years)

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**

#include a .describe table of debtAnalysis

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 over half 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.

In many columns, Nan values had been put in as strings: “Privacy Suppressed”. One easy way to reduce filesize was replacing all these with blanks, which I could later filter out entirely as needed in Jupyter.

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 make money, 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 or removed?.

My ultimate 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. 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.

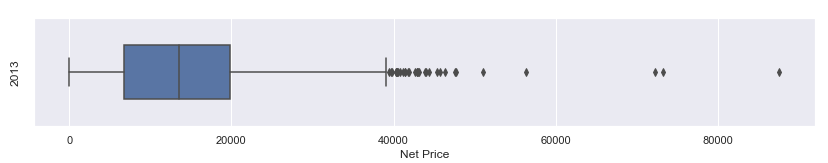
Dividing the student count by branch before multiplying it by the Debt median gave much clearer results for the branch’s **Total Debt**. Using the student demographic counts, I was able to “distribute” each branch’s Total Debt across demographics to get average debts for the following categories:

* 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



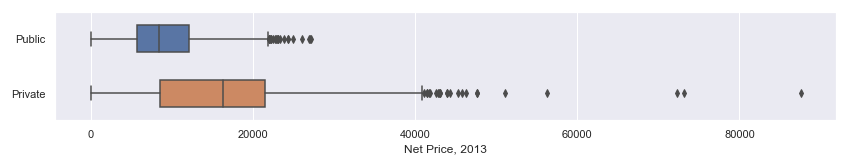
**Analysis**

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).

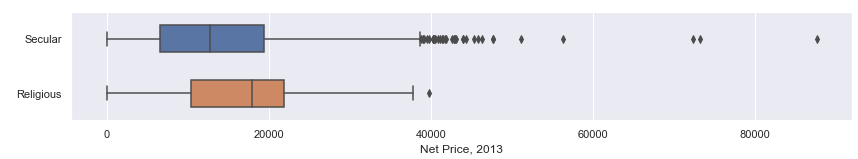


Interestingly, the 3 most extreme outliers are a Photography school and 2 Flight Schools.

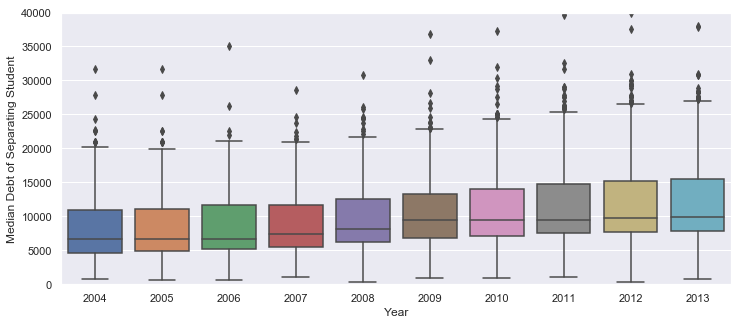
Private Universities are more expensive:



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



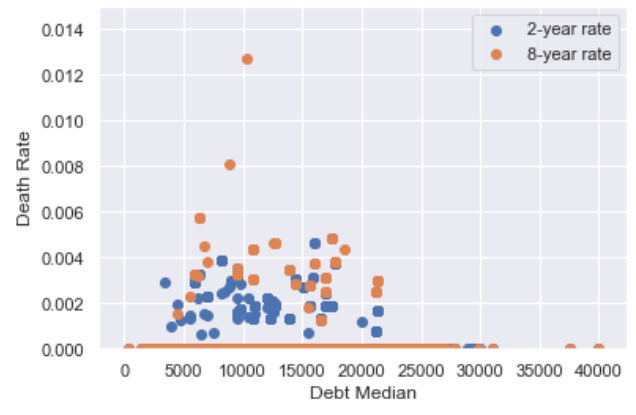
These price tags, in addition to various other factors, have contributed to a steadily increasing median debt for separating students for at least 10 years:



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 it’s 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 influencing mental health. At a minimum, I’m happy to report no meaningful correlation.

**Male & Female:**

Since each school reported median debts figures for their students, and also noted how many of their students were male or female, I was able to calculate how much of their debt for that year went to males and how much went to females. The two distributions of all the branches is seen below:

[]Gender image

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 men tend to have higher debt obligations than women upon separating from these institutions.

Further suggested research topics:

* This sample 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 problem of choice (men tend to choose more expensive schools), or are there other factors driving this difference?
  + Do men leaving these schools have higher earnings or better repayment rates to offset their higher debt balances?