STA141A Homework 1

Aleksandra Taranov
Statistics Department
University of California, Davis

October 2018

In 2015, the U.S. Department of Education launched an online tool called the College Scorecard with the purpose of providing information about cost and value of higher education to students and their families. Their website states that the tool is intended to increase transparency and put the power in the hands of the public. Their audience is not only students but also people who seek to improve college quality through public policy or other initiatives. According to the data document for The College Scorecard project, the data comes from federal reporting from institutions, data on federal financial aid, and tax information. This homework assignment looks at a subset of the dataset which has some years removed and feature names changed. The software used is R and Rstudio and whenever possible every action is done in both Rbase and GGplot. This document was written in texstudio on a laptop running a fedora distribution.

1 Exploratory Data Analysis

1.1 Rows

There are 38,068 rows in the dataset. Each row represents information about a university for a specific year. It is important to note that the information is by year because, otherwise, you would see the name of the college repeating many times. I used the R function titled unique to check to see if there were repeated college names and there were. Therefore, I looked at the columns or factors to see why this might be the case and concluded that it is just because there is information listed for multiple years for each college.

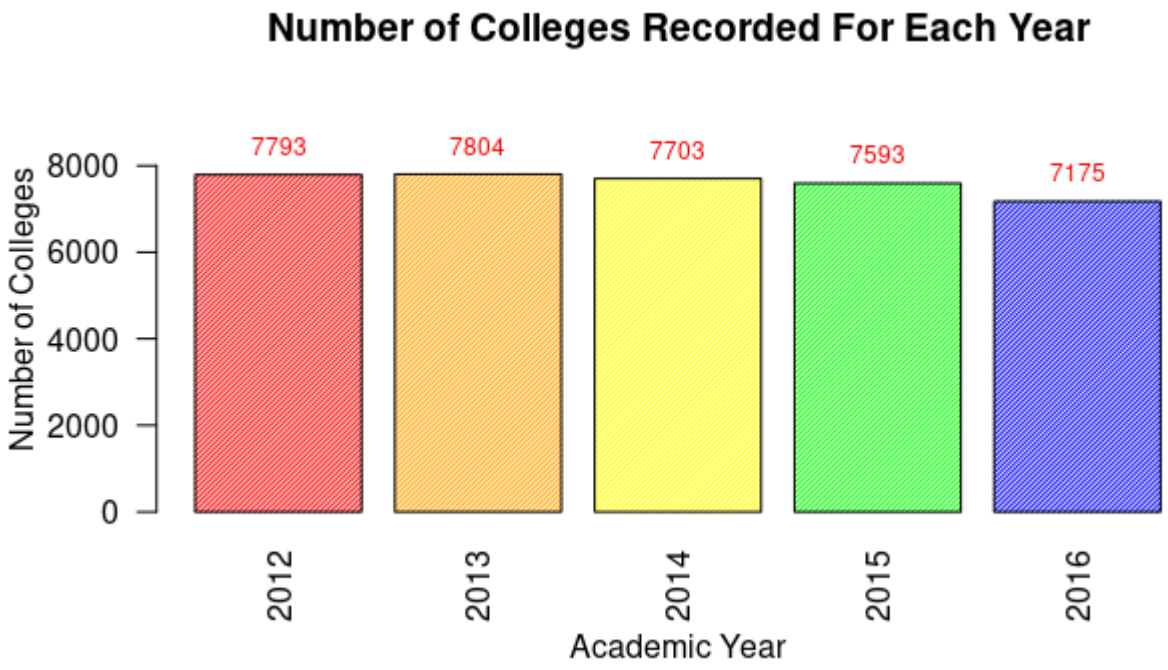
1.2 Columns

There are 142 columns in the dataset, which correspond to 142 different types of information gathered about each college. For instance, the documentation breaks down these factors into several categories: information about the institution, academics, admissions, costs, student body, financial

aid, completion, outcomes for Title IV students, Earnings, and Repayment. These categories represent groupings of the 142 columns, which then have more detailed information such as the 75th percentile of the math section of the SAT, which is part of the admissions group of columns.

1.3 Year Range and Number of Colleges Per Year

The range of our data set is 2012 to 2016 because it is a subset of the larger College Scorecard data set and has some years removed. The number of colleges recorded for each year is illustrated with the barplot below, with each value above each bar. Interestingly, earlier we found that there were 8,332 unique colleges, but none of the years have data for all 8,332. Therefore there is missing data for schools in every year.



2 Analysis by State

2.1 States with the Most and Least Colleges

In this problem I did not use the standard approach of aggregating colleges by state because I knew there would be duplicated college names as a result of having data for the same colleges for multiple years. Instead, I found how many colleges there were for each state per year and computed the mean over the 5 years from 2012-2016. My results were based on the ordering of the mean number of colleges for that state for those 5 years, as shown below.

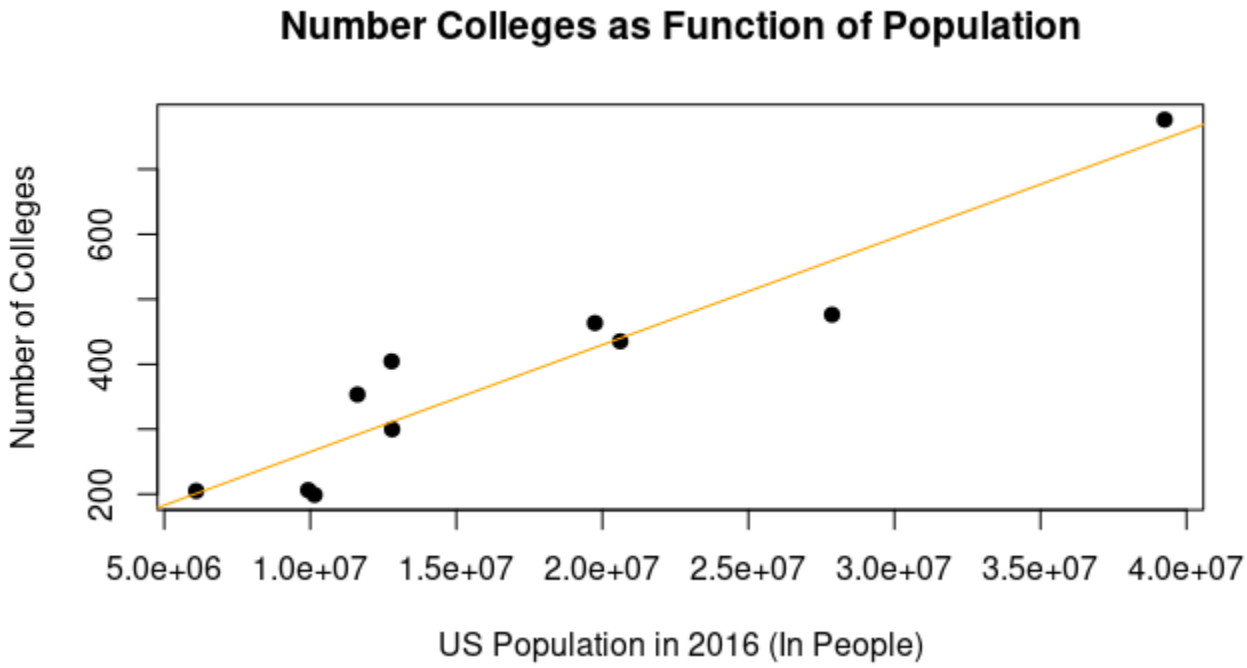
	year2012.state	mean
1	CA	776.20
2	TX	476.20
3	NY	463.40
4	FL	435.20
5	PA	404.40
6	OH	353.40
7	IL	299.80
8	MI	206.40
9	MO	205.00
10	NC	199.00

The five states with the most colleges were California, Texas, New York, Florida, and Pennsylvania. The five places with the fewest, including various territories, were American Samoa, Micronesia, Marshall Islands, Northern Mariana Islands, and Palau. I chose to only look at the 50 US states and not count Washington DC because I thought that territories are distinct enough from states that it would make it harder to see a pattern. When only the 50 US states were considered, the five with the lowest number of colleges were Alaska, Wyoming, Delaware, Hawaii, and Rhode Island.

2.2 Hypothesis

My hypothesis is that the number of colleges is greatly influenced by the population of each state and that states with many people will have many colleges. To test this hypothesis, I went on the US Census website and downloaded a dataset on populations by state for the year 2016. One major limitation is that I was comparing the mean of the number of colleges from 2012-2016 to just the populations in 2016. It would have been better to use the populations from all five years, in case the proportions changed year to year. I was also limited by the fact that I only used the 10 largest states to compare population and state and there could be a lot more information and possible contradictions to my hypothesis in the rest of the data.

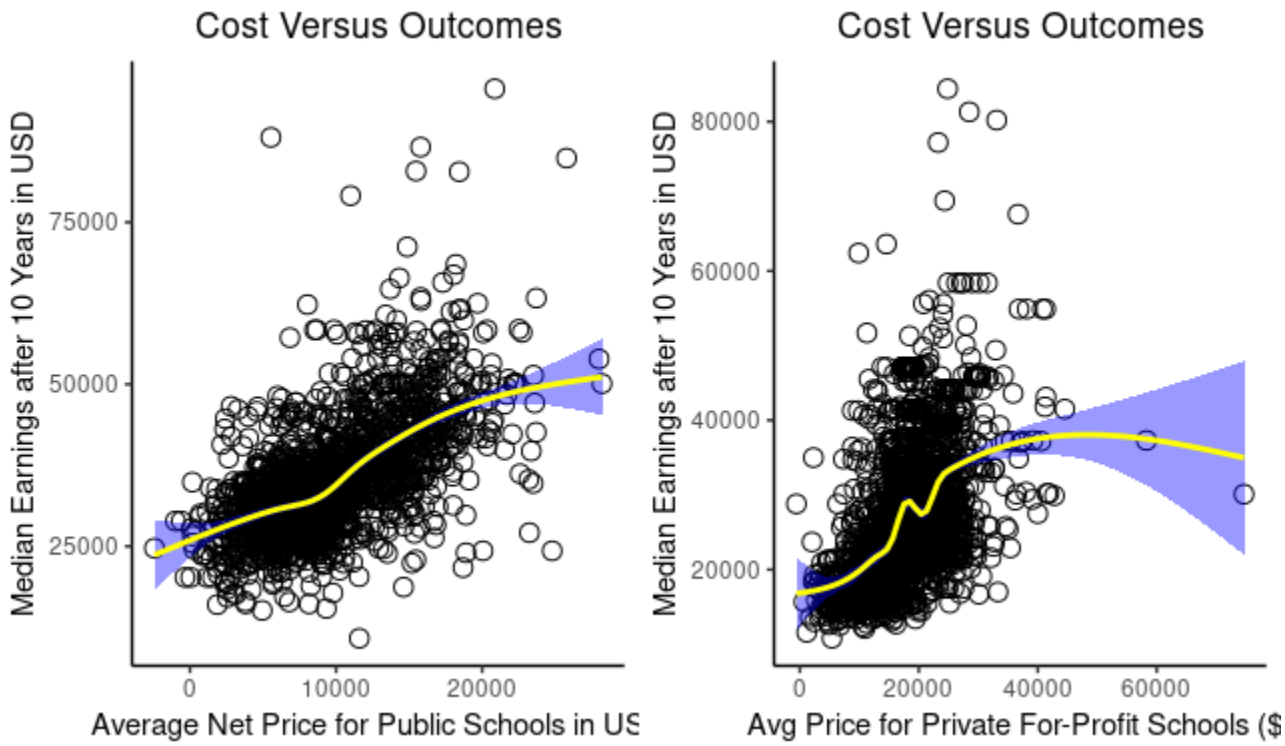
2.3 Testing of the Hypothesis



Having plotted the data and examined it, my hypothesis appears to be correct. Additionally, there is a regression line with a positive slope of 1.646e-05 and a regression coefficient of 0.945, which suggests a strong correlation. Indeed, the 6 states with the largest populations in order are California, Texas, New York, Florida, Illinois, and Pennsylvania and 5 of those were in the list of 5 colleges with the most states. While my hypothesis holds for the largest 4 states, it doesnt explain why Pennsylvania has more schools than Illinois even though Illinois has a larger population. Similarly for the states with the fewest schools, all of the five states with the lowest number of schools were in the lowest 10 states in terms of population. However, Montana, South Dakota, and North Dakota were once again the exceptions to the rule, since they have slightly more schools than Hawaii and Rhode Island despite having smaller populations. However, for the most part, population of state is a good predictor for the number of colleges.

3 Public Vs Private Schools

3.1 Comparing Average Net Price and Median Student Earnings for Public Schools



Above on the left you see the plot for the 2014 data on average net price versus median earnings after 10 years for public colleges. There is a positive slope of 1.204 that suggests that when the average net price is higher the median earnings are also higher. This is a correlation but I do not think it is a causation. Rather, I think there are lurking variables related to financial background of the student and their family. However, it certainly may result in students taking on more loans because they think that they will be able to earn more in the long run if they pay higher tuition now.

Furthermore, I think a major bias in the data is that because of the way it was collected it only shows the data for students who were on financial aid. This means that a huge number of US college students were not included in this data.

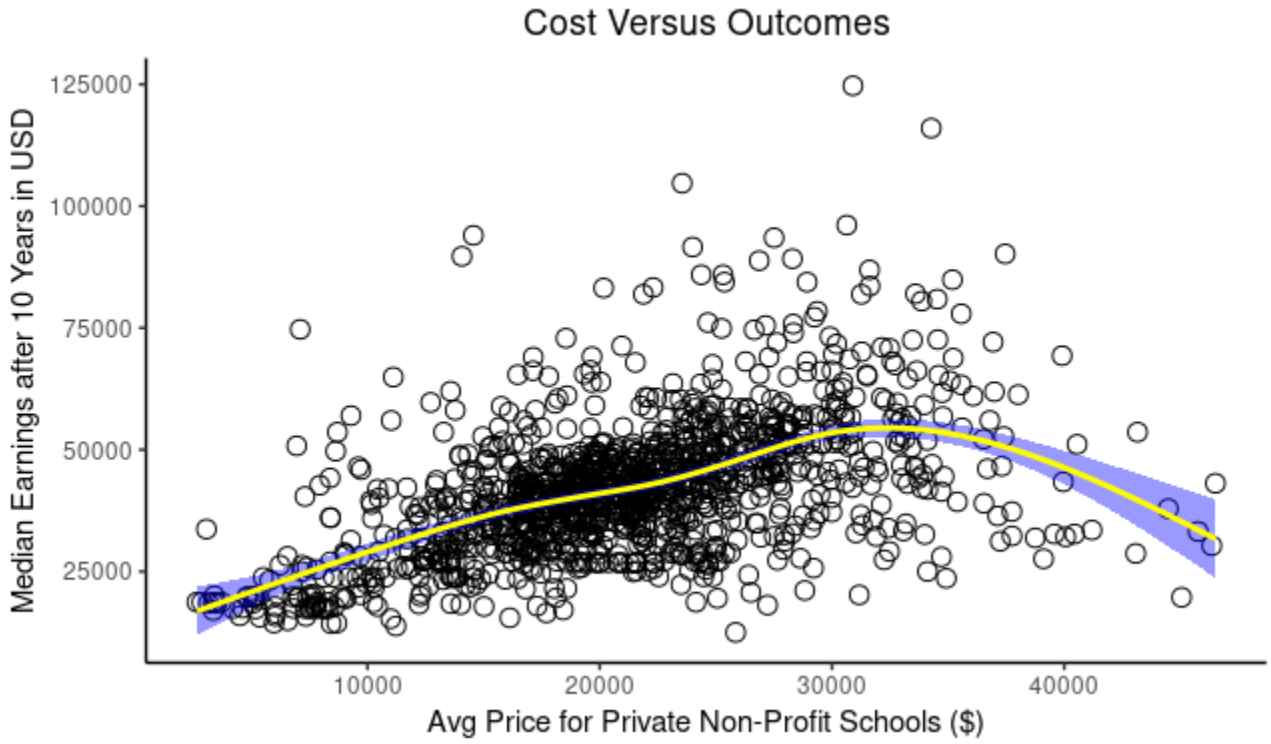
3.2 Comparing Average Net Price and Median Student Earnings for Private For Profit Schools

This plot on the right in the previous section is for private schools in year 2014. It is somewhat similar in the general trend to the previous plot, except that the slope is less steep. Whereas the public data had a slope of 1.204, the slope for the private data is .731. Therefore, although there is still a trend of higher median earnings when the college cost more, it is stronger for public than

for-profit private schools.

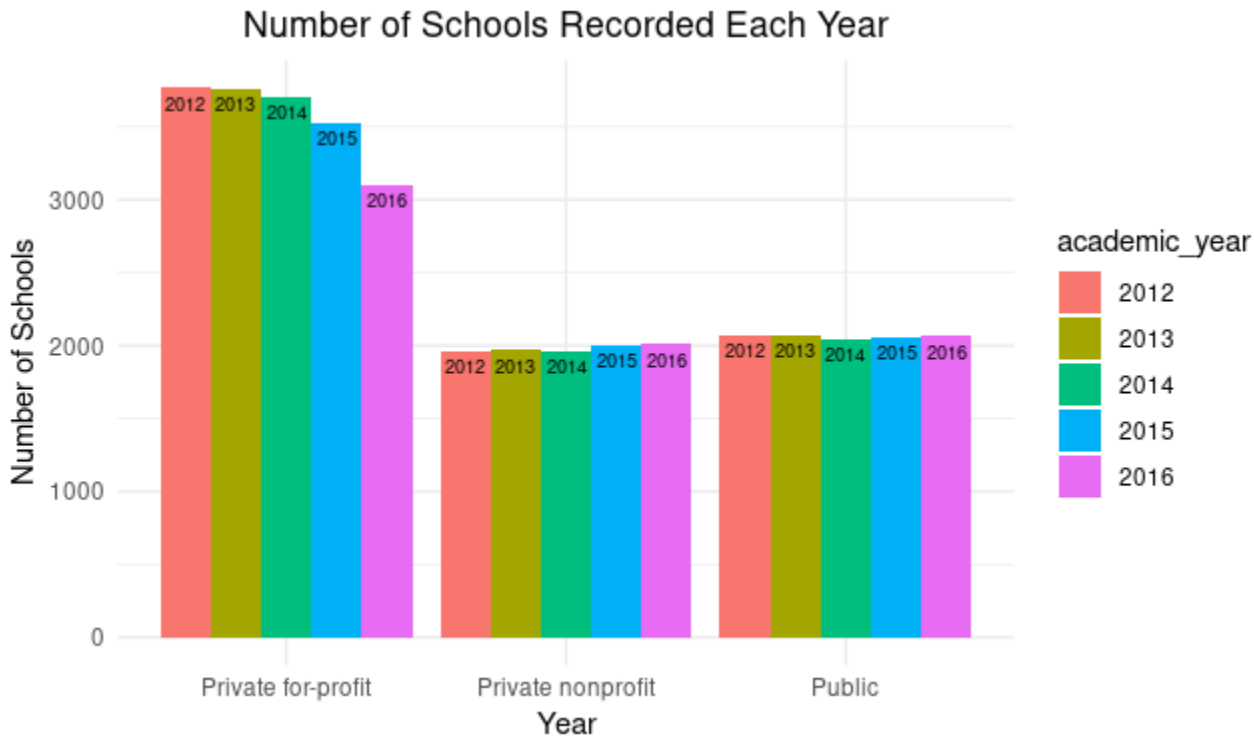
Another obvious comparison is that the range is very different for public schools and private for-profit schools. The range for public schools is $-2434 to 28201$, while the range for private for-profits is $-581 to 47994$. These values are sometimes negative due to the amount of financial aid that the student earns, which can exceed the cost of tuition. Private schools appear to be much more expensive. These two observations together imply that private schools are more expensive and not really giving as much upward mobility compared to the cost.

3.3 Comparing Average Net Price and Median Student Earnings for Profit Non-profit Schools



Continuing from the previous two questions, we can now see the three plots side-by-side for public schools, private for profits, and private non-profits. Public schools are the cheapest and have the higher return to cost ratio than private schools. Nonprofit private schools are more expensive than public schools but cheaper than private for-profit and seem to have a similar curve to the for-profit chart. As stated before, the ranges are quite different for the three groups and the slope is higher for public than for both private schools.

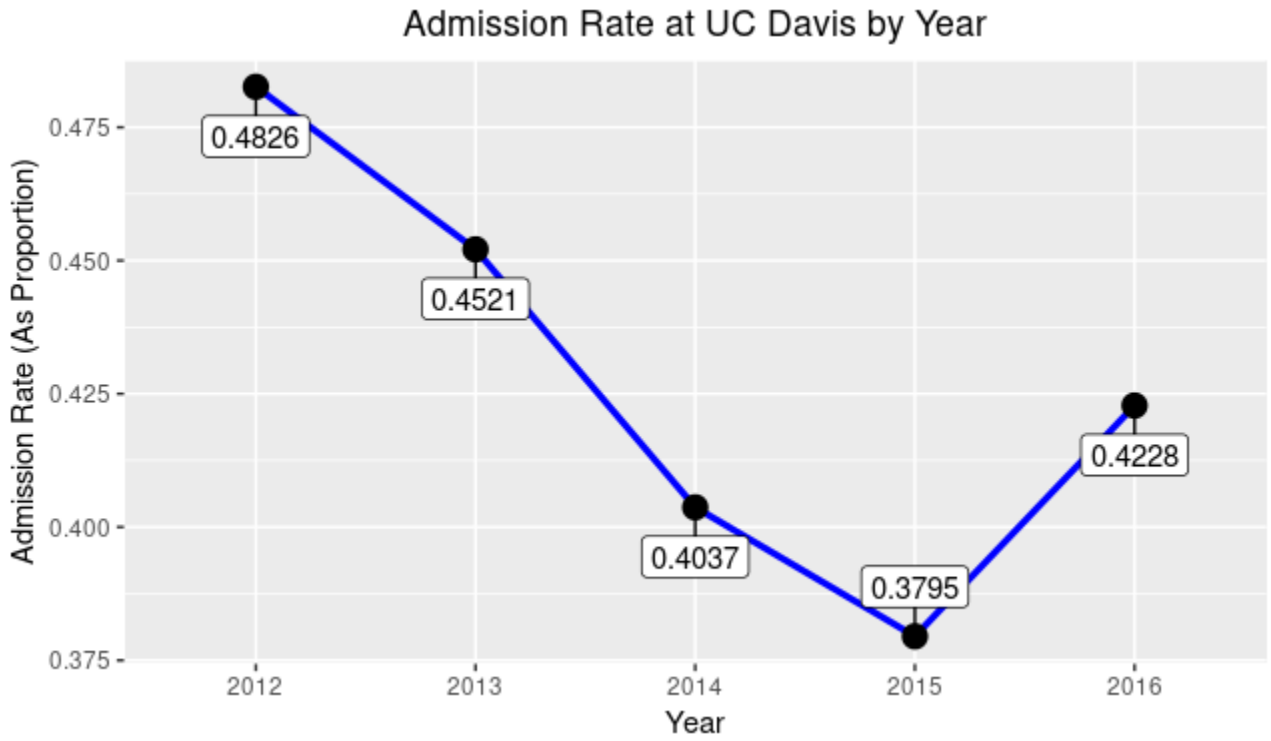
3.4 Ownership Trends



The number of schools recorded for each year seemed to be somewhat steady for public schools and private nonprofits but decreased for private for-profits. A dark hypothesis would be that some of these schools are diploma mills that get shut down. A more likely hypothesis might be that theres less data for private for-profits because the data is based on financial aid and it could be that no one at the for-profit private school gets financial aid in certain years.

4 Admissions

4.1 Comparing UC Davis Admissions Rate By Year



The admissions rates for UC Davis for the five years have a mean of 0.428, median of 0.4228, and the standard deviation is 0.040. This is a fairly low standard deviation and therefore I would say that it is not varying widely. However, there is an interesting trend in that the admissions rate was dropping from 2012-2015 and then increased by quite a bit from 2015 to 2016. I would attribute this increase to a push from UC Davis to enroll more international students to get more tuition. This policy is described in the following article: https://www.sacbee.com/news/local/education/article160029439.html

5 R Data Types

Running either str or lapply with class on the dataset allows me to see the data type for each column. In our dataset we have a wide variety represented: integer, character, logical, factor, numeric. Raw and complex were the only two R data types that we dont have in our dataset. R types do not match one-to-one with statistical data types and therefore sometimes need to be converted to the right data type to work with.

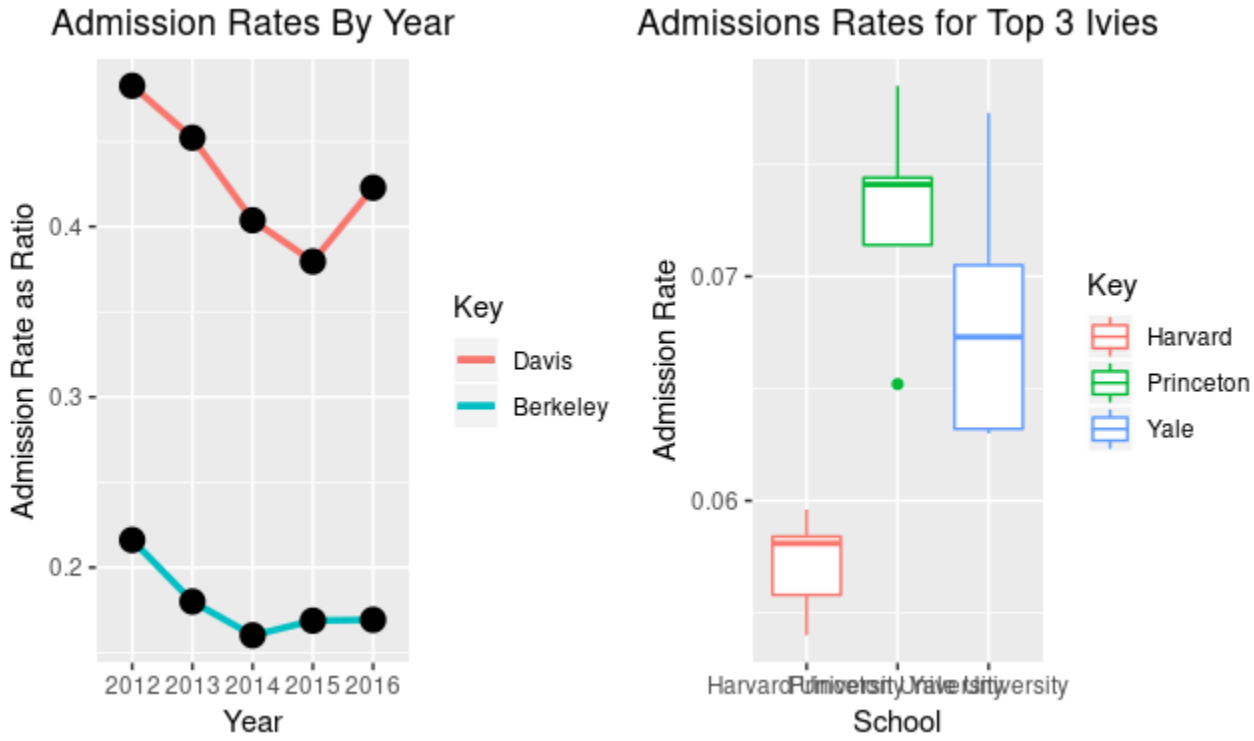
For instance, the column for state is of the type factor, while the column for city is of the type character. If I wanted to combine these two labels I would need them to both be strings and then I could use paste to combine the strings. However, first, to get the state to be a character, I would need to use as.character to coerce it into being a character rather than a factor.

Another example would be using as.numeric to convert character data to a numeric form so that you can take the mean or median of that vector. Here, the academic year is a character, so you would need to coerce it into a numeric form to be able to find the median year, which is 2014 (not that this is useful information.)

Another example involves trying to find how many values of TRUE there are in a logical vector such as the one called online only in our dataset. R sees logical values as logical, but if you use sum on them, true will be treated as the integer 1 and false as 0. However, you could also decide to coerce the values into numeric and then do the same addition to see how many TRUE labels there were in that vector.

Another useful conversion that came up in this project as using as.data.frame to convert vectors into dataframes before merging the two dataframes into one. I used this previously when attaching the 2016 population by state column to the rest of my dataframe about the number of colleges per state.

6 Further Questions



Question one: Which schools have more competitive admissions rates than other schools? This question is important because students use acceptance rates to decide where to apply. They may not want to feel the pain of rejection if it is too hard to get in or they might not want to spend the money on the application fee if their chances are very low. On the other hand, if the admission rate is high, it could calm their nerves and help them cope with the stress of waiting to hear back about their admissions. Additionally, this would be useful information for people who study admissions,

especially since schools like UC Davis have recently started allowing their admissions rates to increase by enrolling more international students who can pay higher tuition.

Here, I propose two types of plots that can answer this type of question, the line plot showing the different admission rates for UC Berkeley and UC Davis from 2012-2016 and the boxplot showing the different average admissions rates for Harvard, Yale, and Princeton for that same timeframe. Each type of plot has a strength and weakness here and they show different information. The line plot can show how the admission rate changes over time, while the boxplot does not capture this information at all. Without it, we would not know that UC Davis started accepting more students in 2016, raising the acceptance rate. However, it would be very difficult to compare schools if we started to plot 10 or 20 different lines on the same line plot. For those instances, boxplots are much easier to see side by side. They also make it easier to compare median values between schools. The variables that I used in the lineplot were the acceptance rate by year, while for the boxplot I used the admission rate by school and showed info for the combined 5 years in each boxplot.

Question two: I will not write the code for this one as it is not required, but I would be interested in seeing how the program percentages for engineering have changed from 2012 to 2016. This would be of interest to college administrators who want to see how many students are enrolling in each type of field so that they can allocate funding fairly between departments or potentially create more opportunities for engineering if it is indeed a growing field. The variables that I would use would be in the column titled program percentage.engineering and after grouping it by year, I would be able to identify trends in whether the number of students studying this subject was increasing or decreasing.

Question three: How do faculty salaries differ between public and private colleges? This question is of use to professors at public universities who are negotiating their salaries and could also inform whether someone decides to work for a public or private university. I would use the columns of ownership to sort the universities between public, private for-profit, and private non-profit and compare the faculty salaries for each.

7 Conclusion

Although this is a rich dataset with a lot of information, there are some glaring biases and omissions. Because the data is collected based on financial aid, we may be missing lots of important information about income of students who do not get financial aid. Additionally, the earnings after 10 years could be highly correlated both to level of preparedness and to the financial background of the student's family. It would always be better to have more and cleaner data, but this data set gave some useful insights.