1 Exploratory Data Analysis

Our final dataset, after omitting null values, was around 2000 observations. We wanted to begin our informal analysis by plotting some data from schools with which are familiar. Therefore, we considered three cohorts - the California State University schools, the University of California schools, and the elite private schools, comprised of the Ivies and a few others. We actually considered a fourth cohort, the top-expenditure schools, but those mostly overlapped with the third cohort.

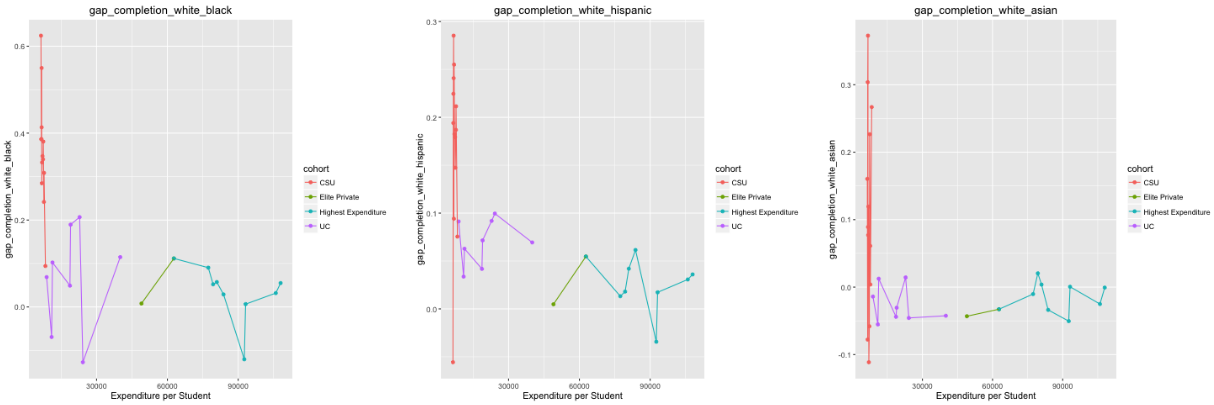

Figure 1: We see that while there is a clear segmentation of expenditures across cohorts, the difference in outcomes is not as clear-cut. There is a general negative relationship, but a lot of variance. The most obvious conclusion is that the CSUs suffer from much worse outcome gaps than the other cohorts.

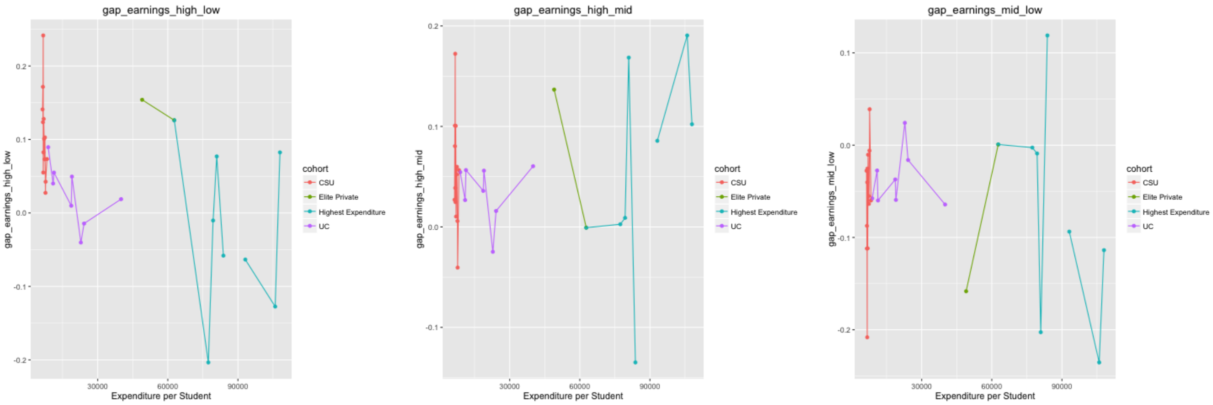

Figure 2: We see a similar pattern with earnings. Interestingly, the data actually suggests that middle-income students do worse in comparison to high-income students as expenditures increase.

We now look at the full dataset to get a sense of the relationship across all schools in the US. We begin with completion rates.

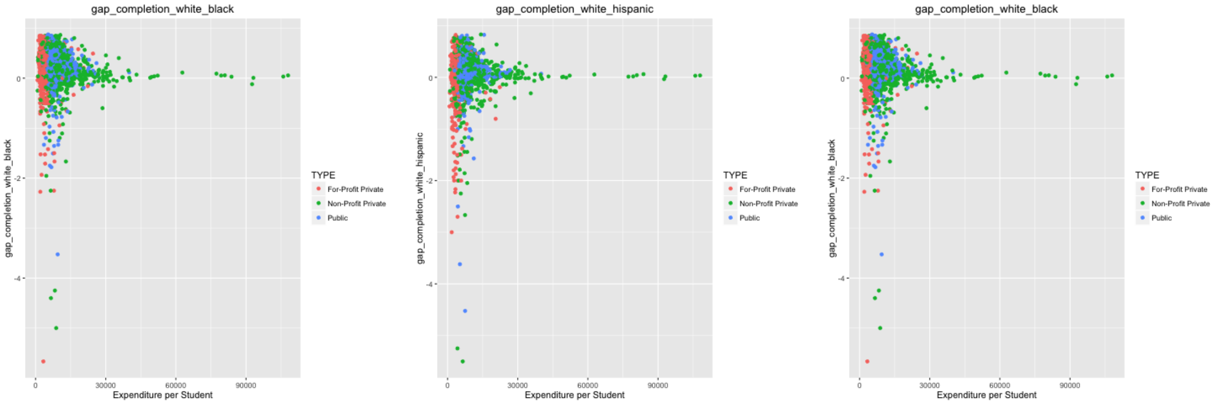

Figure 3: While there are definite differences in expenditure by type of school, it is not obvious that achievement gaps decrease with expenditures. Generally, achievement gaps are not centered too far from zero, and there are quite a few outliers in both dimensions. However, it does look like variance decreases and outcome gaps cluster more closely around zero as expenditures increase.

We restrict our bounds to exclude the outliers to see if we can visually detect any patterns.

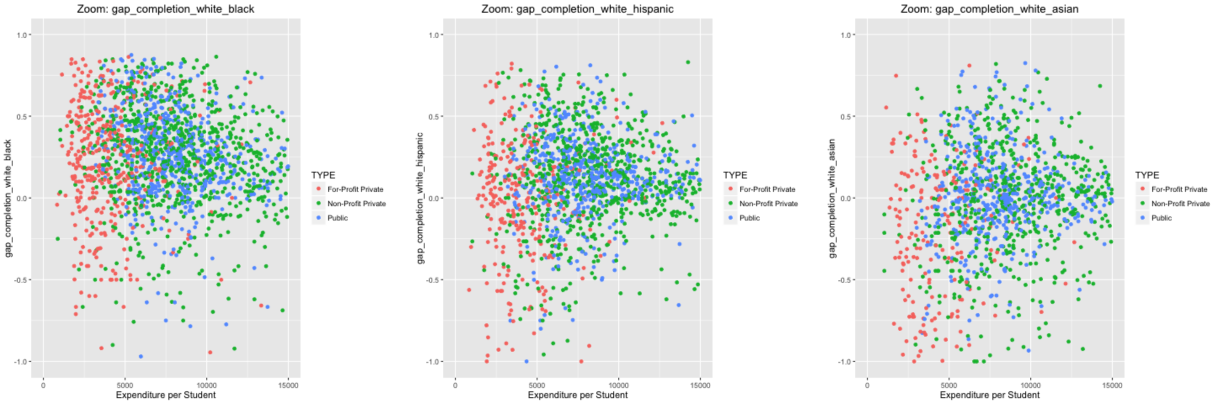

Figure 4: The data actually looks almost perfectly random.

Looking at our earnings metrics, we see a similar spread.

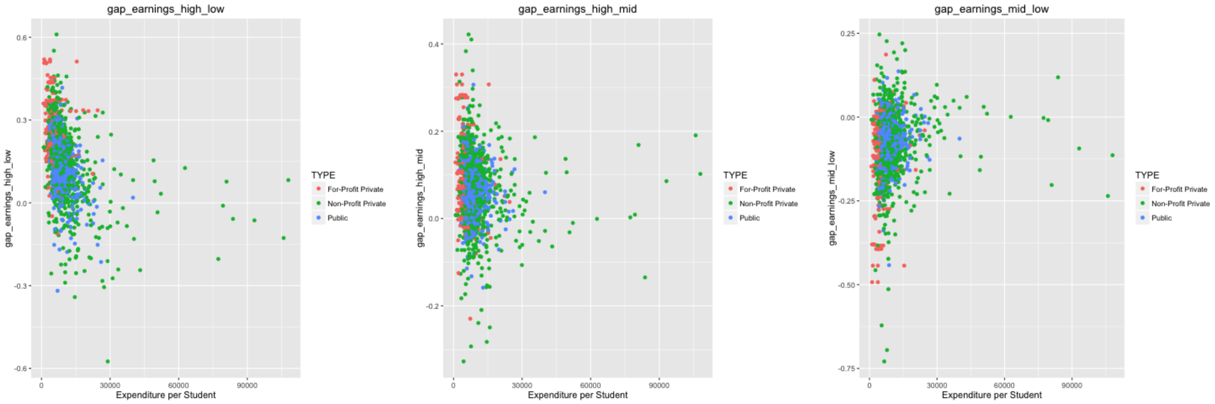

Figure 5: We see a similar pattern, where variance decreases and the gap clusters closer to zero as expenditure increases. An interesting observation is that the gap between medium and low income terciles actually seems centered below zero - low income students on average seem to perform a bit better than their middle income peers.

However, restricting bounds produces a more interesting visual insight.

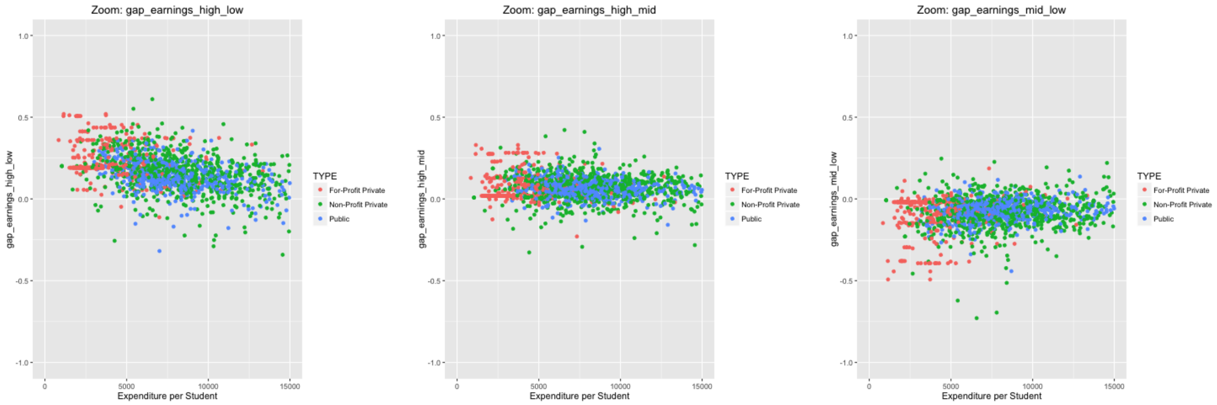

Figure 6: Fascinatingly, we actually see a very clear negative correlation between expenditure and outcome gap for High-Low and High-Medium. Medium-Low seems to have no relationship.

The relationship between expenditure and outcome gap is not obvious, but there definitely seem to be some non-random patterns worth investigating. In the next sections, we undertake formal statistical analysis to develop more concrete conclusions about these patterns.