When it comes to higher education, who really profits?:

An analysis of higher education's impact on socioeconomic mobility for low-income students by institutional control

Maggie Davidson, SI 618 WN 2020

I used the US Department of Education College Scorecard dataset to explore the relationship between the control of an institution (public, private nonprofit, or private for-profit) and outcomes related to socioeconomic mobility for low-income students. I am interested in this problem primarily because I volunteer with a college access organization that focuses on postsecondary educational success for low-income students. I am interested in learning more about outcomes to inform the educational programming that supports students in choosing an education that is right for them. Many people view higher education as a vehicle for upward socioeconomic mobility. However, some schools are more effective at facilitating this mobility for their students than others. Many of the students go on to for-profit postsecondary institutions, but it is my belief that students who attend for-profit institutions receive less benefit for their investment than students who attend public or private nonprofit institutions. However, I would like to explore this data and challenge my own beliefs. For this analysis, I will focus on 3 variables for low-income students:

- Earnings (how much students make 10 years after entering the institution) - a metric to assess upward socioeconomic mobility as a function of a student's education at this institution
- Completion (6-year completion rates for low-income students, either at the original institution or after transferring to a 2- or 4-year institution) - a metric to assess how frequently a student's investment in their education at this institution results a degree
- Cost (net price paid) - a metric to assess the financial investment a student undertakes in order to attend the institution

Institutions with high earnings, high completion, and low cost would indicate a greater likelihood of upward socioeconomic mobility for its students, whereas institutions with low earnings, low completion, and high cost would indicate a lower likelihood of upward socioeconomic mobility.

Questions

I chose to focus individually on each of the selected metrics, and then in the last question use unsupervised machine learning to cluster the institutions based on all 3 variables to identify a more holistic picture.

1. **Earnings**: How do postgraduate earnings for low-income students differ across institutions by control? Is there a statistically significant difference?
2. **Completion**: Are low-income students more likely to complete a degree (either at the original institution or after transferring to another institution) in 6 years at an institution of one control as compared to another?
3. **Cost**: What is the relationship between the control/ownership of the institution and net price for low-income students? Do low-income students pay more to go to one type of institution as compared to another?
4. **Trends**: When clustering institutions based on earnings, completion, and cost, what trends can be seen in institutions that are clustered together?

Data Source

I downloaded the raw data files directly from the US Department of Education (https://collegescorecard.ed.gov/data/). The data is in CSV file format, with one CSV file for each academic reporting year going back to 1996-97. The most recent data is from 2017-18. Each row in each CSV file represents one institution of postsecondary education. Each file has just under 2,000 columns representing information such as the demographics of the institution, the percentage of students who graduate with different types of degrees, the average price and average earnings of graduates, and more.

The particular data I was interested in were most recently complete in the 2014-2015 academic year dataset. Therefore, I opted to analyze this dataset, which contains 7,703 records (individual institutions) and 1,977 columns. I sliced the dataset down to some basic information about each institution, including the institution's name (str), control (int), average net cost of attendance for different populations (float), completion rate as a percentage for different populations in different timeframes (float), and earnings a certain number of years after graduation in dollars (float).

Methods

There were two primary ways that I cleaned the data and prepared it for analysis. The first is that the dataset contained values "PrivacySuppressed" to denote that data was collected, but was not included due to representing so few students that including it could be a violation of their privacy. For the purpose of this analysis, it is unimportant whether the value is "PrivacySuppressed" or null (unavailable), so I replaced all "PrivacySuppressed" with np.nan. The second is that I sliced out only those (groups of) columns that I intended to use to increase efficiency and usability using the data dictionary to select relevant data.

Question 1: Earnings

To answer my first question about earnings, I first identified the variables I wanted to use and cleaned and prepared the data further. I used the data dictionary to select 1 variable to define institutional control (CONTROL) and 3 variables for mean earnings of students who attended the institution and were working and not enrolled 10 years after entry. These 3 variables were split by the student's original entry income tertile--students whose income was under $30,000 (MN_EARN_WNE_INC1_P10), students whose income was between $30,000-$75,000 (MN_EARN_WNE_INC2_P10), and students whose income was over $75,000 (MN_EARN_WNE_INC3_P10). I then discovered that the earnings variables were stored as strings, so I converted them to numerical data.

Next, I created 3 side-by-side histogram plots, with each plot showing a different institutional control, to show mean earnings of students who attended institutions of that control. In each plot, I plotted 3 histograms--one for each income tertile--to do an easy visual comparison between the earnings of students from different income backgrounds.

To confirm my hypothesis from the histograms that there was a difference in how much money low-income (lowest income tertile) students make across institutional controls, I created summary tables and then performed an ANOVA test where I modeled mean earnings for the lowest income tertile students on control as a categorical variable.

My biggest challenge was creating the subplots in a way where they all looked the same. I kept trying things such as adding axis labels in the way I would with a regular single plot but they would only be applied to the last plot. I was able to use the subplot documentation to determine ways to format the plots as I needed.

Question 2: Completion

To answer my second question about completion, I first identified the variables I wanted to use and cleaned and prepared the data further. I used the data dictionary to select 3 variables for 6-year completion rates (as percents) of low-income students. These 3 variables were split by the way the student completed a degree--completed at the original institution (LO_INC_COMP_ORIG_YR6_RT), completed after transferring to a 4-year institution (LO_INC_COMP_4YR_TRANS_YR6_RT), and completed after transferring to a 2-year institution (LO_INC_COMP_2YR_TRANS_YR6_RT). I felt it was important to include the second two to avoid unfairly favoring nonprofit institutions; many students attend public community colleges with the plan to transfer to complete their degree, which should not count against the institutions. I then discovered that the earnings variables were stored as strings instead of numerical, so I converted them to numerical data.

A challenge I faced next was that I wanted to know the 6-year completion rate for low-income students at any institution, regardless of whether they completed at the original institution or transferred. This should be a simple sum of the 3 columns, but not all rows had values in all columns. In fact, only approximately 1,800 rows had values in all 3 columns. Summing rows that were missing data in at least one column could lead to the completion rate for that row (institution) being artificially low, because a null would be treated as a 0. The data in the first column were fairly complete, with over 5,000 observations out of the dataset of nearly 8,000, while the other two variables were much more limited, at between 3,000-4,000 observations. In inspecting the data, I found many observations were missing one of the second two observations, but rarely both. I decided to sum rows with at least 2 of the 3 values, since this would give me over 1,000 more observations and not skew the data as much as summing rows with only 1 value.

I proceeded to sum the columns and then round the values so that I could use them as categorical variables in a chi-square test and contingency table. I added a heatmap so that it was apparent where higher values lay. I normalized the contingency table across the index to demonstrate the percentage of institutions for each institutional control that fell into each completion rate range.

Question 3: Cost

To answer my third question about cost, I first identified the variables I wanted to use and cleaned and prepared the data further. I used the data dictionary to select 2 variables for average net price of attendance for low-income students. These 2 variables were split by institutional control--public (NPT41_PUB) and private for-profit and nonprofit (NPT41_PRIV). I then created a third variable with both costs combined, after confirming that there was no overlap between the two variables.

I was interested in approaching each of the first 3 questions in a similar way so that I could compare the results to determine whether for-profit institutions are worth the investment by earnings, completion, and cost. However, I needed to make sure to approach each question with some innovation and creativity. I decided that a violin plot would help to quickly visualize the difference or lack thereof in net price by control, and to then finish the analysis using an OLS regression to confirm or reject my hypothesis formed based on the violin plot.

Question 4: Trends

To answer my last question about trends, I used Agglomerative Clustering to cluster the institutions by the variables from the first 3 questions that were related to low-income students only; i.e. MN_EARN_WNE_INC1_P10 (mean earnings for low-income students), LO_INC_COMP_ORIG_OR_TRANS_YR6 (completion rate for low-income students either at any institution, not rounded), and NPT41_ALL (mean net price of attendance). I used the clusters to identify similarities and differences between the institutions. To

identify the correct number of clusters, I plotted a dendrogram cut off at only 2 levels. I wanted to use a small number to make it reasonable to describe the clusters. I then clustered the institutions into 8 clusters as identified by the dendrogram and assigned the cluster labels to the original dataset.

Next, I explored the relationship between the clusters and the institutional controls, which weren't used in clustering, using a violin plot. I then explored the relationship between the variables that were used as well as the clusters and institutional control using a pair plot. Last, I created violin plots to explore the relationship between clusters and the 3 variables used in clustering.

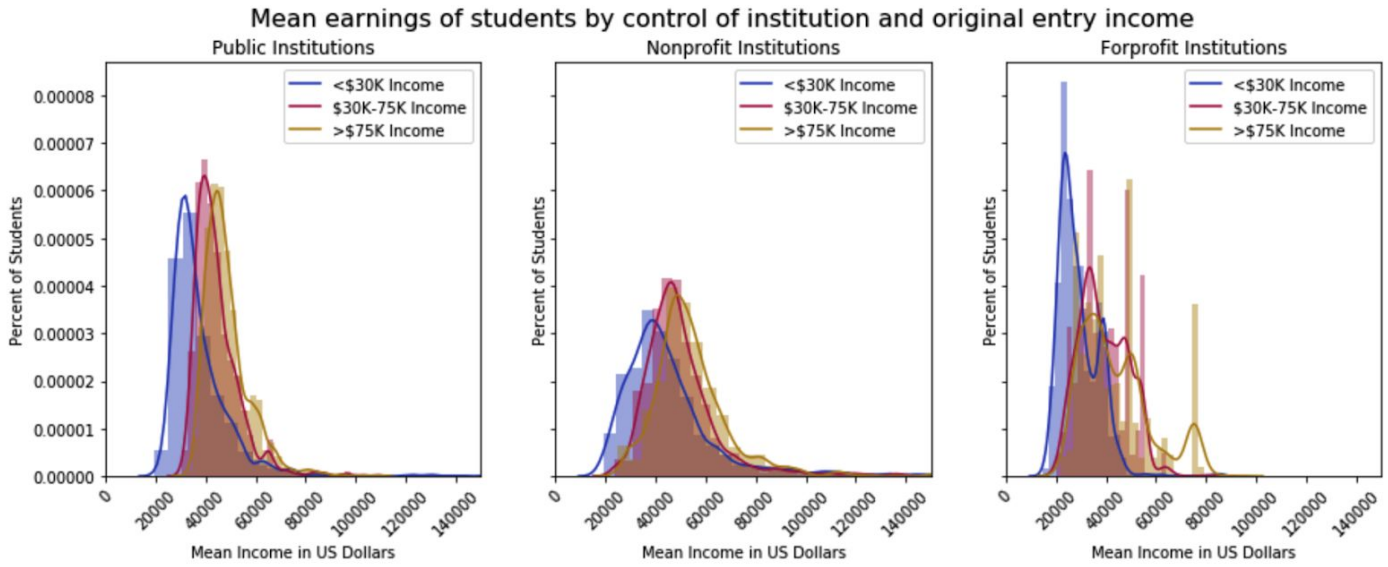
I then manually created a table to roughly demonstrate the attributes of each cluster in terms of control and each of the 3 variables. I made predictions about what type of institutions might fall into the 5 most interesting clusters (0, 3, 4, 6, and 7). Then, to confirm or reject my predictions, I plotted the 10 most commonly used words in names of institutions in each of those 5 clusters.

The part that I found most challenging was describing the clusters, since some of them had little obviously in common that I could glean without diving deep into the nearly 2,000 attributes available about each institution. I found it most impactful to generalize but be clear that I was doing so, as well as use a combination of control, the 3 outcome variables, most common words in institution name, and a sample of 10 institutions from the cluster to identify trends.

Analysis and Results

Question 1: Earnings

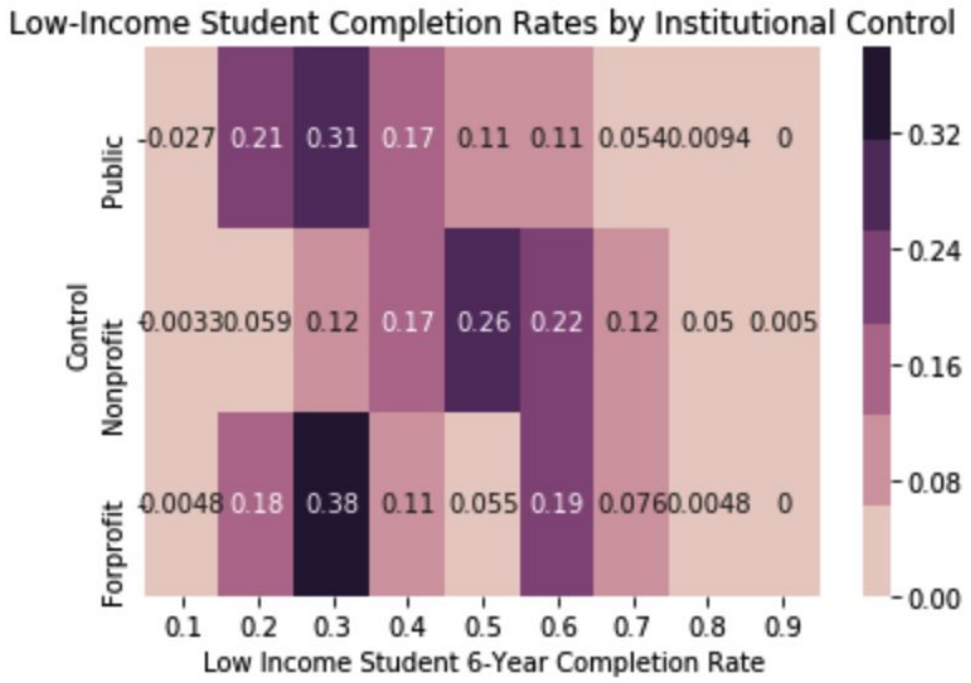
I found that there is a statistically significant difference in the earnings of low-income students between different institutional controls. Low-income students who attend nonprofit institutions on average earn the most ($45,014), while low-income students who attend for-profit institutions on average earn the least ($29,059). Low-income students who attend public institutions make on average $37,852. This difference is apparent in the following visualization. This visualization shows that the peak for low-income students at public institutions appears to be approximately $35,000, at nonprofit institutions about $40,000, and at for-profit institutions about $25,000. The difference in earnings between students of different income tertiles is very apparent in this visualization, a very clear indicator of the lasting impact of family income privilege.



Based on this analysis, I would conclude that for-profit institutions on average result in the lowest earnings for low-income students, while nonprofit institutions on average result in the highest earnings for low-income students.

Question 2: Completion

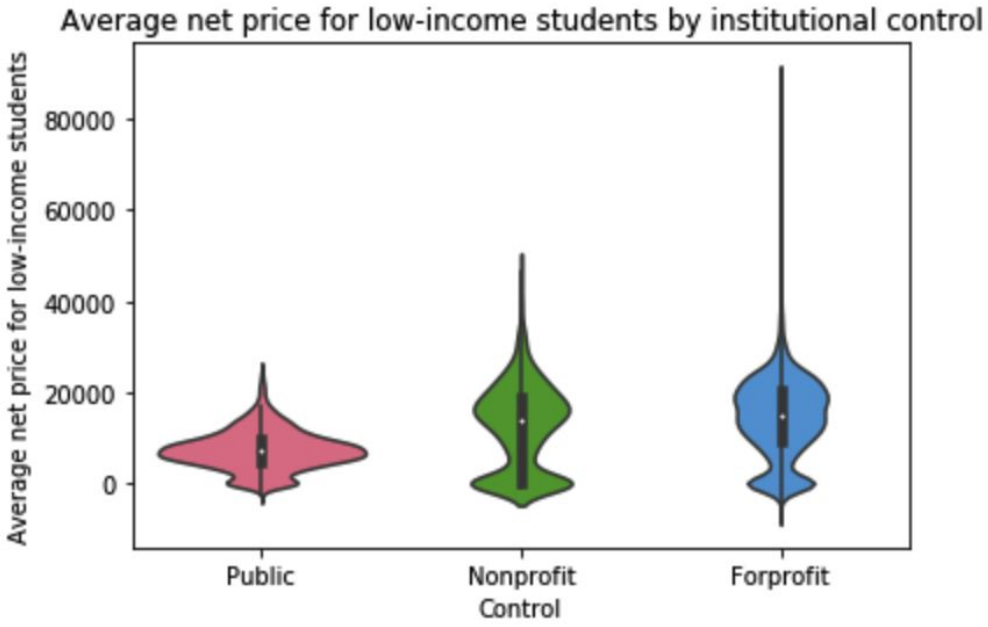
My null hypothesis was that there was no difference in completion rates based on institutional control. Using a chi-square test, I was able to reject my null hypothesis and determine that there was a difference in completion rates based on institutional control. Using a contingency table with a heatmap, it is apparent that 6-year completion rates for low-income students at nonprofit institutions are on average much higher than completion rates at either public or for-profit institutions.



Based on this analysis, I would conclude that for-profit and public institutions on average result in the lowest 6-year completion rates for low-income students, while nonprofit institutions on average result in the highest 6-year completion rates for low-income students.

Question 3: Cost

Using a combination of a violin plot, summary table, and OLS regression, I was able to reject the null hypothesis that there is no difference in net price across controls and determine that there is a statistically significant difference. It is apparent in the violin plot that the range of cost is narrowest with public institutions and widest with for-profit. It appears that, on average, the price is highest with for-profit institutions and lowest with public. I was able to confirm this suspicion with the summary table--the mean net price for public institutions was $7,467, approximately half of the mean net price for for-profit institutions--$14,396. Nonprofit institutions came out in the middle, as predicted, with a mean net price of

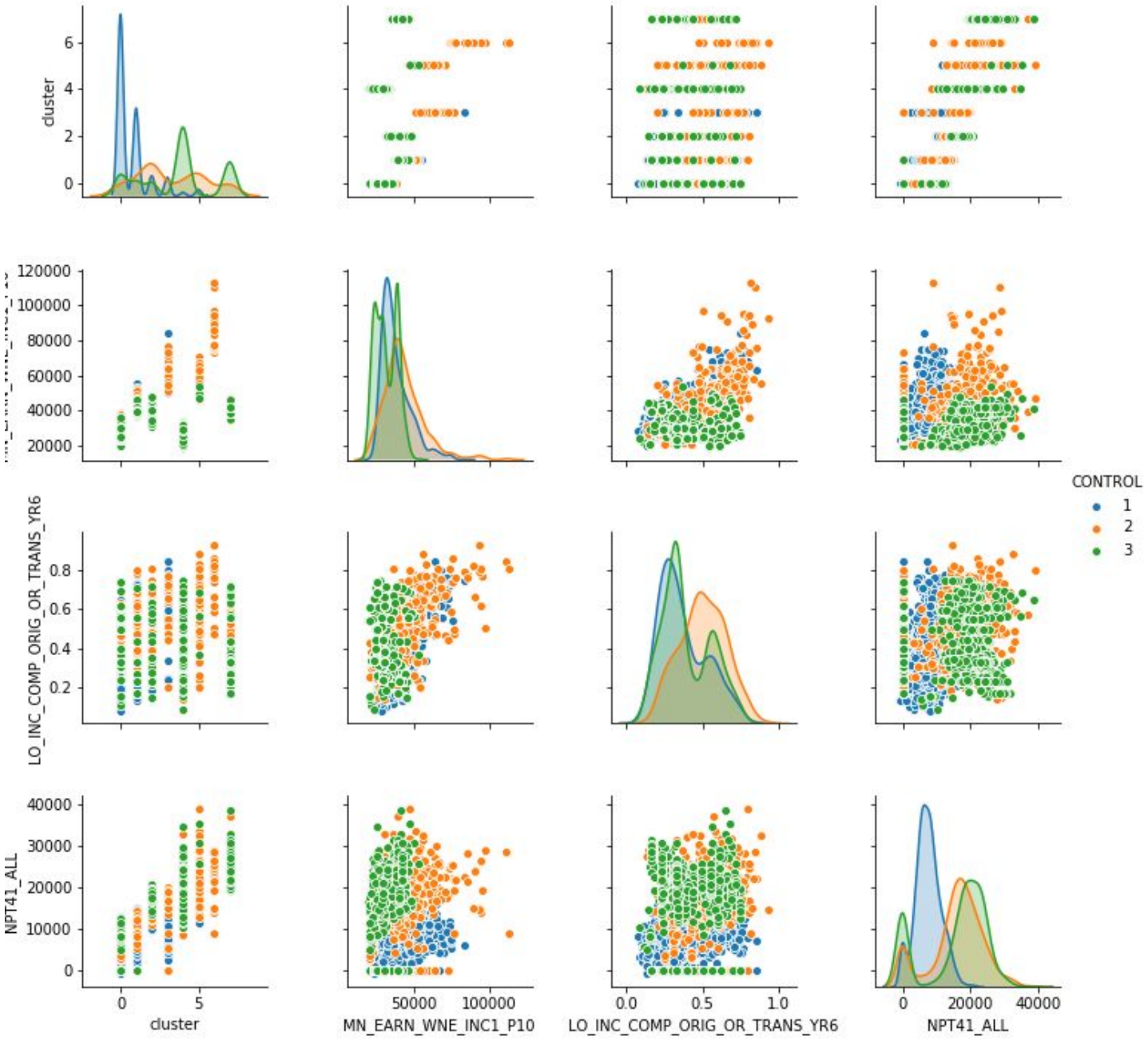


$12,414. I confirmed that these differences are statistically significant using an OLS regression. Therefore, low-income students generally pay the most to attend for-profit institutions, and the least to attend public institutions.

Based on this analysis, I would conclude that for-profit institutions on average result in the highest cost for low-income students, while public institutions on average result in the lowest cost for low-income students.

Question 4: Trends

After clustering institutions based on attributes related to low-income student socioeconomic mobility, I used a pairplot to quickly identify trends between clusters. As can be seen in the below plot, most public institutions can be found in the lower-numbered clusters, while for-profit institutions can be found in the higher-numbered clusters, with nonprofit institutions in the middle. There was a stronger relationship between the clusters and mean earnings and clusters and net price than cluster and completion rate.



I found that cluster 3, which consisted of mid-to-highly ranked public universities, tended to be the most reliable in facilitating upward socioeconomic mobility for low-income students, with high median earnings and completion rates but low net cost. However, cluster 4, for-profit trade schools for lower-earning trades, as well as cluster 7, "predatory" for-profit institutions, are the least reliable in facilitating upward socioeconomic mobility for low-income students, with low earnings and completion rates but high net cost. I was surprised to learn that community colleges may be inexpensive, but the likelihood of completing a degree and the earnings after completing that degree may not be worth the cost for low-income students. I created a table to roughly describe each cluster, with color-coding for the 3 continuous variables (red = worst, yellow = mid, green=best, gray=relationship between cluster and variable not meaningful enough to specify, as defined by a box (quartiles) that appear to span more than 20% of the available range and don't neatly fall into a bad/mid/good bucket) to quickly see the differences between clusters.

Cluster	**General Description**	**Control**	**Mean earnings (10 year) for low-income students** *Worst = low, median <$40K Best = high, median >$60K*	**6-year completion rate for low-income students** *Worst = low, median <30% Best = high, median >60%*	**Mean net price for low-income students** *Worst = high, median >$20K Best = low, median <$10K*
0	Community colleges	Primarily public, but includes institutions of all 3 controls	Low Median ~$31K	Low Median ~28%	Low Median ~$5K
1	Public non-community colleges and satellite public university campuses	Primarily public, but includes institutions of all 3 controls	Mid Median ~$41K	Minimal relationship between cluster and completion Median ~45%	Low Median ~$8K
2	Mid-ranked institutions, particularly nonprofit religious schools	Fairly evenly distributed across all 3 controls; however, the largest number of nonprofit institutions in any cluster is in this one	Mid Median ~$40K	Minimal relationship between cluster and completion Median ~45%	Mid Median ~$15K
3	Mid-to-highly-ranked public universities	Primarily public, but also includes nonprofit; no for-profit	High Median ~$65K	High Median ~65%	Low Median ~$9K
4	For-profit trade schools for trades that traditionally	Primarily for-profit, but includes	Low Median ~$28K	Minimal relationship between cluster	Mid Median ~$19K

	earn on the lower end, such as beauty and art	institutions of all 3 controls		and completion Median ~40%	
5	Mid-ranked nonprofit institutions	Primarily nonprofit, but includes institutions of all 3 controls	Mid Median ~$52K	High Median ~62%	High Median ~$20K
6	Highly ranked nonprofit undergraduate health sciences-focused universities	All nonprofit	High Median ~$85K	High Median ~75%	High Median ~$22K
7	For-profit institutions often described as "predatory" (such as ITT Tech)	Primarily for-profit, but also includes nonprofit; no public	Mid Median ~$40K	Mid Median ~30%	High Median ~$23K

Conclusion

I concluded that public universities are most successful at facilitating upward socioeconomic mobility for low-income students (high earnings, high completion rates), combined with being most accessible due to low net costs. Private nonprofit institutions, particularly those in the health sciences, lead to high earnings and have a high degree of success due to high completion rates, but tend to be very expensive, adding a large barrier to low-income students hoping to climb the socioeconomic ladder.

On the other hand, the institutions that demonstrated lower likelihoods of upward socioeconomic mobility for their students were for-profit institutions, either those that take predatory approaches, such as ITT Tech and DeVry, or those which train students for lower-earning trades, such as beauty and art. These institutions tend to have high costs for low-income students, but the chance of students graduating with a degree are significantly lower, as well as their earnings once they leave the institution.