Do Black students have a significantly higher graduation rate at HBCUs as opposed to other colleges/universities?

Introduction

As college students, we were in a position of researching colleges not too long ago. For many students, important considerations might include cost of attendance, acceptance rates, and job prospects. We wanted to explore a portion of these considerations; our analysis touches on the importance of considerations like institution type: HBCUs (historically Black colleges and universities) or Non HBCUs. The relevant population this data is geared towards would be students applying to academic institutions. We aim to answer the question: Do Black students have a significantly higher graduation rate at HBCUs than at other institutions of higher learning? We could use our results to see if institutions that are specifically created with Black students in mind are better at supporting Black students compared to other institutions.

Dataset and Exploratory Data Analysis

Our dataset is from the College Scorecard Project, which was collected by the US government from individual institutions, tax information, and records of federal financial aid. We accessed this data from https://collegescorecard.ed.gov/data/ on November 9, 2022. According to the website at the time, the data was last updated on September 14, 2022. The cases include 6,681 colleges and universities in the United States. The dataset as a whole discusses 2,989 variables.

We concluded that six variables were the most relevant to our research question. Our key outcome of interest is the four year completion rate for Black students (C150_4_BLACK). Our key predictor of interest is HBCU. We have two confounding variables: Average SAT (or equivalent) score of students admitted (SAT_AVG_ALL) and the percentage of first-generation students (PAR_ED_PCT_1STGEN). Finally, we have two precision variables: The average cost of attendance in dollars (COSTT4_A), and whether an institution is fully online (DISTANCEONLY). We fit models with interaction variables to check for effect modifiers and found none.

In order to work with our dataset to answer our research question, we first needed to clean our dataset. Since we are focusing on 4 year institutions, we filtered for ICLEVEL == 1, representing where level of institution is 4 years. We also filtered out NULL values for graduation rate and HBCU as an added measure and also because if there were NULL values, they won't make sense. For our variables, we have HBCU, a categorical variable, represented by 0 for "Not HBCU" and 1 for "HBCU". This is the same for our DISTANCEONLY variable, which is also categorical. Our dataset has 85 HBCUs and 1879 non-HBCUs. It also has 19 online-only institutions, and 1945 non-online-only. For our quantitative variables SAT_AVG_ALL, PAR_ED_PCT_1STGEN, and COSTT4_A, we turned them into numerics rather than floats and integers. Table 1 includes a summary of all of our quantitative variables.

Methods

In order to account for the other variables, we fit a linear model, which is as follows:

$$E[C150\_4\_BLACK \mid HBCU,\ COSTT4\_A,\ SAT\_AVG\_ALL,\ PAR\_ED\_PCT\_1STGEN] = \beta_0 + \beta_1 * HBCU + \beta_2 * COSTT4\_A + \beta_3 * SAT\_AVG\_ALL + B_4 * PAR\_ED\_PCT\_1STGEN$$

Ultimately, we dropped DISTANCEONLY from our model. There are no HBCUs that are also online only, which made fitting a model with both pretty much impossible.

Our null hypothesis (H_0: $\beta_1 = 0$) is that whether an institution is an HBCU is not associated with the graduation rate of Black students. Our alternative hypothesis (H_A: $\beta_1 \neq 0$) is that whether an institution is an HBCU is, in fact, associated with the graduation rate of Black students. We set a significance threshold of 0.05. A Type 1 error would be saying that HBCU status of a college significantly impacts the graduation rate of Black students when it doesn't, and a Type 2 error would be saying that

there is insufficient evidence of a relationship between HBCU status and the graduation rate of Black students when there is a relationship. Based on the implications of the two types of errors we determined that a threshold of 0.05 is reasonable.

Results

Figure 1, as well as Table 2, suggest that the graduation rate for Black students, without accounting for other impactful variables, is lower in HBCUs compared to non-HBCUs. We determine that this relationship isn't incredibly accurate as graduation rates can be affected by a variety of factors.

Figure 2 is a scatterplot visualization that shows the relationship between average cost of attendance (COSTT4_A) and Black graduation rates (C150_4_BLACK), the dark blue illustrating HBCUs compared to the gray non-HBCUs. We see here that holding the cost of attendance constant, we can generally observe that HBCUs have a higher percentage of Black graduation rates compared to Non HBCUs.

Figure 3 is a scatterplot visualization showing the relationship between average SAT scores (SAT_AVG_ALL) and Black graduation rates (C150_4_BLACK). Similar to cost of attendance, when an HBCU and Non HBCU have the same average SAT score, the Black graduation rate at the HBCU will be comparatively higher.

Figure 4 is a scatterplot visualization illustrating the relationship between percentage 1st generation students (PAR_ED_PCT_1STGEN) and Black graduation rates (C150_4_BLACK). Interestingly, when holding the percentage of first generation students constant, HBCUs have a lower Black graduation rate compared to Non HBCUs, when the percentage of first generation students is high.

Table 3 presents our results from fitting a multiple linear regression model, predicting our outcome (C150_4_BLACK) with (HBCU), (COSTT4_A), (SAT_AVG_ALL), and (PAR_ED_PCT_1STGEN). This table summarizes the estimated coefficients, standard error, t-value, and p-values of each variable and their relationships to (C150_4_BLACK).

The primary statistic of interest is β_1, the regression coefficient from our model that estimates the difference in graduation rate of Black students at US colleges between HBCUs and non-HBCUs in the broader population. β_1 is the expected change in graduation rate of Black students associated with a school being an HBCU, holding all other variables constant. Shown by table 3, β_1 is estimated to be 0.1576, meaning that holding cost, SAT scores, and 1st generation status constant, attending an HBCU is associated with a 15% higher graduation rate for Black students than attending a non-HBCU.

If HBCU status had no association with Black graduation rates, then we could expect to get an estimate at or more extreme than our estimate < 0.1% of the time ($p < 0.001$). This means that it would be surprising, given our data, if there were no relationship between HBCU status and the graduation rate of Black students, holding all other variables constant. So, we reject our null hypothesis, in favor of the alternative hypothesis: there is an association between HBCUs and the graduation rate of Black students.

Based on classical sampling, we have a confidence interval of 0.10748 to 0.20772. This means that we are 95% confident that the true difference in graduation rates of Black students between HBCUs and non-HBCUs, holding all other variables constant, is between 10.748% and 20.772%.

Discussion

When other variables are accounted for- likely costs of attendance and SAT (or equivalent) scores- we can see that Black students at HBCUs have a significantly higher graduation rate as opposed to other institutions. Our p-value (< 0.001) is very low and our confidence interval only contains what we think are significant values. Based on our findings, we believe that increased funding for HBCUs could

have an overall positive impact on academic and future successes of Black students in the US. Supporting HBCUs is treating a symptom of racial inequality, but there are also root causes that are perhaps more important to address.

We would say that our dataset is relatively representative of our target population. It includes information on every American college that receives federal funding, which is the vast majority of colleges in the United States. While that means that this data, and therefore analysis, isn't very useful outside of the US, it's still potentially useful to the hundreds of millions of people living within. It is important to note that the US Department of Education recognizes 107 HBCUs while our dataset is limited to 85. Even so, we determine that with the large number of cases, there's unlikely to be significant sampling bias. There could, however, be measurement errors in our analysis. Because the data is collected at the institutional level rather than the individual, we're making conclusions based on percentages of student bodies rather than the circumstances of actual students. For example, perhaps an institution could have very few Black students who are all first generation, but a student body which is overwhelmingly *not* first generation. Our analysis would essentially consider that case to be one of Black students who are unlikely to be first generation. Some ethical considerations we might have can be with data collection as there is no data on institutions that opt out of federal funding. This data is used to help prospective students make decisions about where they'll go for education, and these institutions have less of a chance of being in their radar.

The limitations of our analysis are mainly related to our predictors. We found that even though they serve to answer our research question, they themselves are not root causes. For example, we can see that HBCUs tend to have lower SAT scores, and that lower SAT scores are associated with lower graduation rates (see Figure 3). However, our analysis offers no insight into why either of those are the case, or what should be done about it. Not much can be done with our analysis on its own, but it points to things that are worth looking into. It doesn't appear that any causal claims can be made with our analysis. Another limitation is that the amount of funding a school receives could very well be an effect modifier, but isn't included in our analysis. While it would make sense that higher funding leads to higher graduation rates, it could be worth inspecting whether or not this impact is felt differently for Black students. With further analysis we might also look into the relationship of PAR_ED_PCT_1STGEN with our other variables as it resulted in a P-value above our significance threshold despite obviously corresponding with graduation rates, which might be worth looking into.

Tables and Figures

	Mean at HBCU	Mean at Non HBCU	Median	Min / Max	SD	Missing values
COSTT4_A	$25646.63	$38959.18	$35087	$9565 / $81531	17247.88	77
SAT_AVG_ALL	957.5882	1151.7400	1120	760 / 1566	130.9574	851
PAR_ED_PCT_1STGEN	36.82268%	30.72412%	31.84783%	8.866995% / 61.53394%	0.09628353 (9.628%)	89
C150_4_BLACK	33.43%	41.27%	42.86%	0% / 100%	0.2315627 (23.156%)	0 (removed null values)

Table 1. Summaries of key variables

HBCU Status	**Average Black Grad Rate**
No	41.27%
Yes	33.43%

Table 2. Graduation rates for Black students in HBCUs vs non-HBCUs. See Figure 1 for a visualization.

Variable	**Estimate**	**Standard error**	**t-statistic**	**p-value**
Intercept	-0.8624	-0.08073	-10.683	< 0.001
HBCU1	0.1576	0.02506	6.288	< 0.001
COSTT4_A	2.537e-06	3.546e-07	7.154	< 0.001
SAT_AVG_ALL	1.067e-03	5.706e-05	18.702	< 0.001
PAR_ED_PCT_1STGEN	-0.01177	0.07409	-0.159	0.874

Table 3. Results from multiple linear regression predicting C150_4_BLACK (Black graduation rates) with HBCU, COSTT4_A (average cost of attendance), SAT_AVG_ALL (average SAT or equivalent score), and PAR_ED_PCT_1STGEN (percentage of students at an institution who are first-generation).

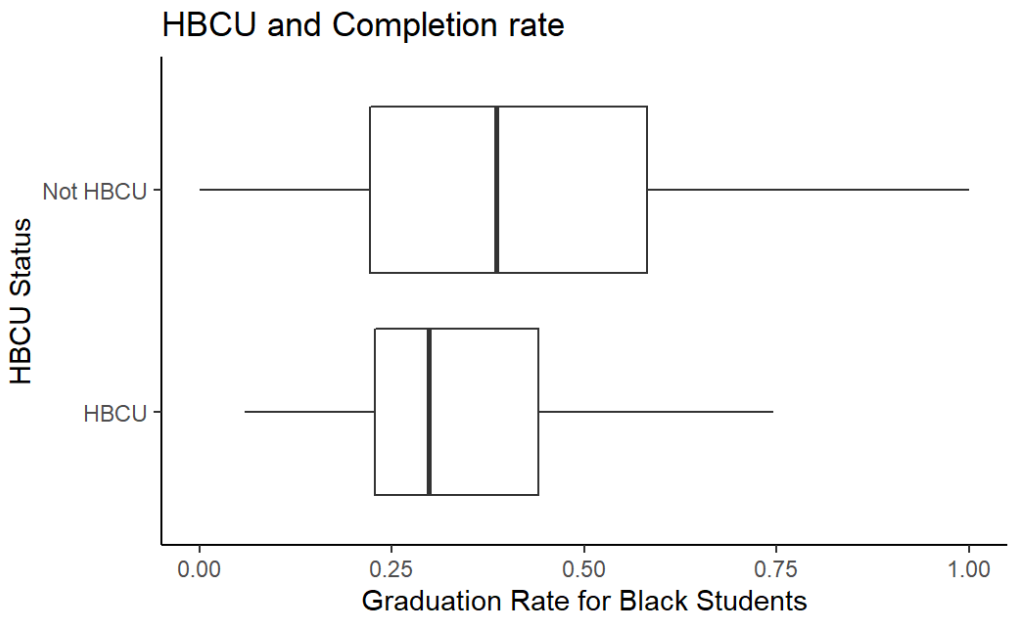

Figure 1. Boxplots showing graduation rates in HBCUs as opposed to other institutions before accounting for confounders.

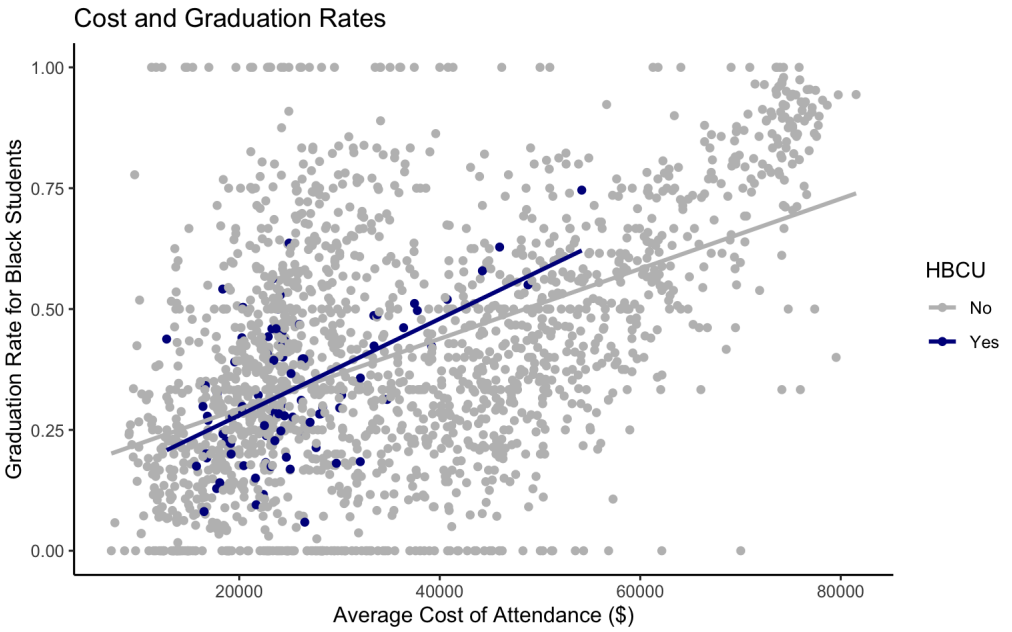

Figure 2. Scatterplot showing the relationship between cost of attendance and Black graduation rates at HBCUs and Non HBCUs.

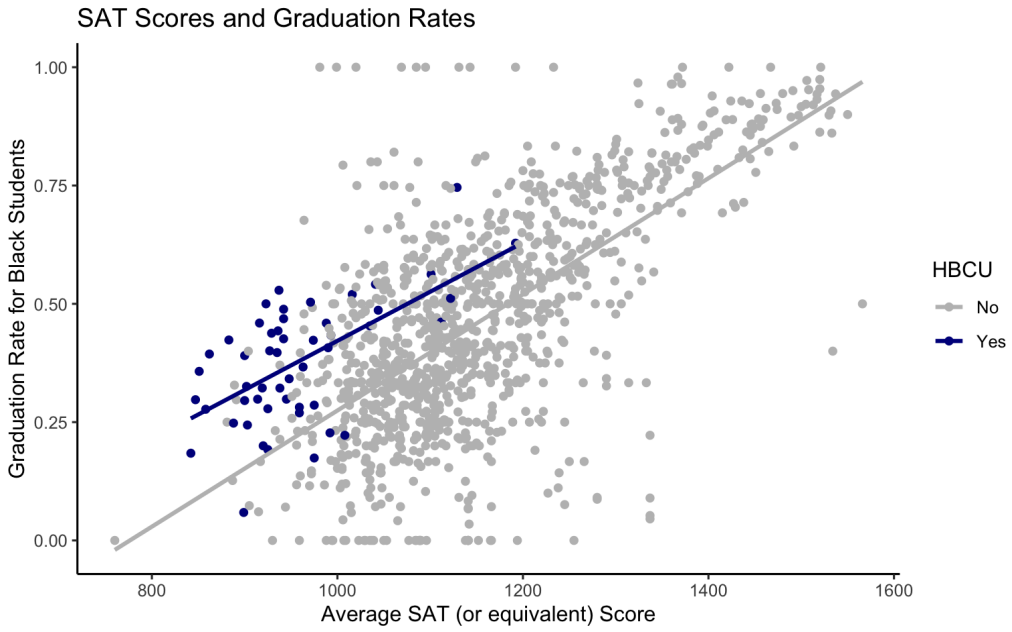

Figure 3. Scatterplot showing the relationship between SAT (or equivalent) scores and Black graduation rates at HBCUs and Non HBCUs.

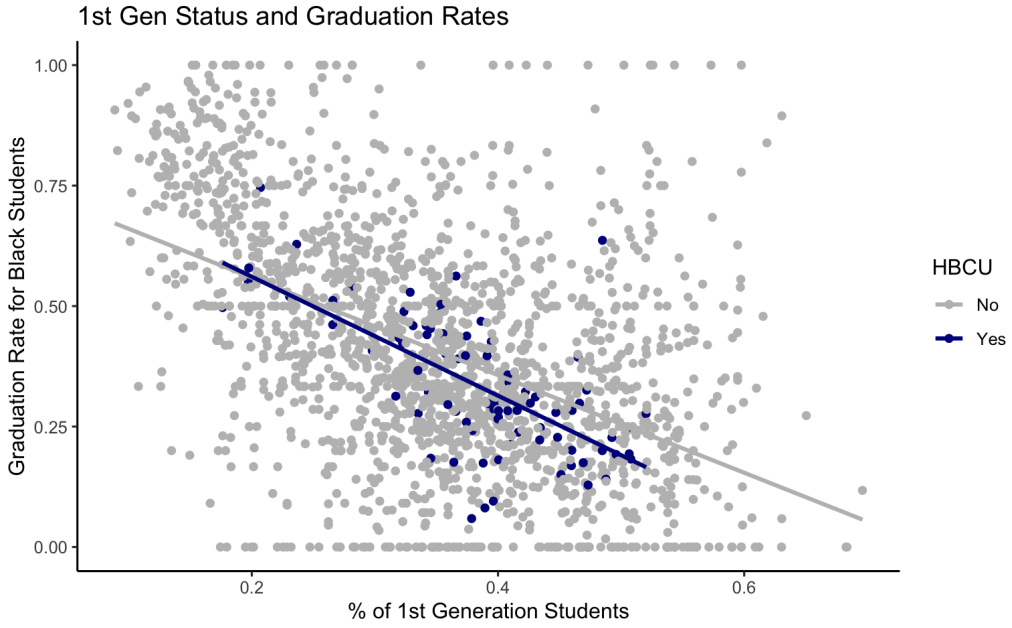

Figure 4. Scatterplot showing the relationship between percentage of 1st generation students and Black graduation rates at HBCUs and Non HBCUs.