Suggestion of School Goodness: Average SAT admission score by institutional characteristics*

Zihan Jin

4/25/2022

Abstract

The SAT is a college entrance exam administered by the College Board that is taken by high school students across the United States and is widely used for college admissions. This paper is aim to assist students and families meet their educational goals by increasing transparency between school average SAT admission scores and institutional characteristics they are concerned about. By graphical analysis of gender and ethnicity enrollment share, institution controls, graduation rate, tuition, and student earnings, the results suggest that SAT is an effective indicator of the goodness of institution performance and Asian students have a higher share of outstanding groups in undergraduate education. Further discussion about Asian American-serving institutions indicates that they are well supported to students' success compare to other institutions. Besides, an multiple linear regression model was designed to estimate institutions' average SAT scores. These findings could help students select schools from the options that better meet their educational goals.

Introduction

The SAT is a college entrance exam administered by the College Board that is taken by high school students across the United States, typically in their junior or senior year. It is a standardized test widely used for college admissions. Some studies indicate that 90% of the institutions consider the SAT to be of "moderate" or "considerable" importance in the admission process (Clinedinst and Hawkins 2010).

Under the purpose of the established dataset, this paper is aim to assist students and families meet their educational goals by increasing transparency between school admission criteria, which is the average SAT admission score, and institutional characteristics they are most concerned about. The higher average reported scores are concluded to make an institution appear more selective and of higher quality (Robinson and Monks 2005). Students' earnings are a key component to explain the goodness of an institution. A past study has explained that earnings are strongly correlated with the selectivity level of the college attended (Tierney et al. 2000). Race, on the other hand, shows various performance of SAT scores in terms of distribution, timely improvement tec (Hedges et al. 1998; Grissmer et al. 1998, 1994). Many previous studies have been conducted investigating, but data from the most recent 2022 data and analysis of multiple considerations from students and families remain inadequate.

To fill the vacancy, this paper focuses on the factors that are often mentioned by candidates while selecting undergraduate institutions, including school gender and ethnicity ratio, institution controls, graduation rate, tuition, and student earnings, and conducts a comprehensive analysis through line charts. The result suggests that in the high SAT score segments, private schools account for a larger proportion. At the same time, both graduation rate and student earnings increase with the increase of SAT score, which shows that SAT is an effective indicator for measuring the goodness of a school. It is worth mentioning that Asian students have a higher share of outstanding groups: the enrollment share of Asian students increases with the increase of the institutions' average SAT admission score, and the graduation rate is the best among all ethnic groups. With a closer focus on the Asian student group, I found that Asian American-serving

*https://github.com/jin-zihan/sta304-final.git

institutions were well supportive of students' success. The specific performance is that the overall tuition level of Asian American-serving institutions is relatively low, which makes the loan burden of students lighter. Meanwhile, they help students obtain higher salaries than the institutions with equivalent admission SAT scores. Finally, after qualitatively considering the factors for school selection, this paper uses the multiple linear regression model to quantitatively analyze the factors. The results showed that, out of 20 variables, 8 variables were statistically significant. Among them, a 6-year graduation rate for Blacks, Hispanics, and Asians, Tuition of the out-of-state student, Enrollment share of Asians have a positive impact on average SAT admission score, while the Enrollment share of Blacks and Hispanics are negative.

The remaining part of the paper is organized into four sections, including Data, Results, Discussion, and Appendix. In the Data section, this paper explains the data source and processed methodology. One can also have quick browsing of data characteristics through graphical and statistical descriptions. In the result section, an analysis of characteristics of institutions with different average SAT score ranges (cf. Figure 3) and further discussion about Asian American-serving institutions (cf. Figure 4, 5, 6) are carried out by graphical analysis, and the future institutions average SAT admission score are estimated by multiple linear regression with regression output, the goodness of fit, and results explanation. The discussion section comment on the paper's findings, weaknesses, and future works. Finally, a data-sheet is included in the Appendix part to help users thoroughly understand the dataset in this paper.

Data

The dataset in this report is from the 2022 College Scorecard. College Scorecard collects data annually through surveys administered by the Department of Education's National Center for Education Statistics (NCES) and Integrated Postsecondary Education Data System (IPEDS) is the primary source of data on postsecondary education institutions in the United States. There are two data files provided by the College Scorecard, including data about institutions as a whole and data about specific fields of study within institutions. Since this paper is aim to analyze institutional-level attributes, instead of focusing on field study, we choose the former data file about institutions as a whole.

There are seven main aspects, including basic descriptive information about the institution, types of academic offerings available at each institution, admissions rate and SAT/ACT scores of students, students costs, demographic of the student body, completion and retention, students earnings and repayments. The dataset includes 2989 variables of 6694 institutions. No variables are newly created. These data are newly released and of high quality. They also better represent the whole US college population and therefore are less likely to suffer from the sampling bias issue present in many prior studies.

Data description

The key variable this paper focuses on to make analysis is the average SAT equivalent score of students admitted of institutions, which is hereinafter often referred to simply as "average SAT score". The SAT is a college entrance exam administered by the College Board that is taken by high school students across the United States, typically in their junior or senior year. The exam consists of math and critical reading sections scored between 200 and 800, so students can receive a combined score between 400 and 1600. Here is a quick view of its distribution.

According to figure 1, the average SAT score is around 1100.

Additionally, I pick 19 variables that are most frequently considered by students and their families from the remaining six aspects of the dataset. Here are the descriptive statistics for numerical variables.

Variable	Definition	Min	Max	Mean	SD
UGDS_WMEN	enrollment share of women	0.1265	1	0.5657	0.0890
UGDS_MEN	enrollment share of men	0	0.8735	0.4343	0.0890
UGDS_WHITE	enrollment share of whites	0.0105	0.8926	0.5516	0.2092
UGDS_BLACK	enrollment share of blacks	0.0027	0.9303	0.1074	0.1324

Variable	Definition	Min	Max	Mean	SD
UGDS_HISP	enrollment share of Hispanics	0.0043	0.9209	0.1502	0.1384
UGDS_ASIAN	enrollment share of Asians	0.0012	0.3841	0.0620	0.0720
UGDS_AIAN	enrollment share of American Indian/Alaska Native	0	0.6473	0.0076	0.0345
UGDS_NHPI	enrollment share of Native Hawaiian/Pacific Islander	0	0.0884	0.0030	0.0068
C150_4_WHITE	6-year graduation rate of whites	0.1429	0.984	0.6219	0.1595
C150_4_BLACK	6-year graduation rate of blacks	0	1	0.4827	0.2095
C150_4_HISP	6-year graduation rate of Hispanics	0	1	0.5479	0.1856
C150_4_ASIAN	6-year graduation rate of Asians	0	1	0.6364	0.2111
C150_4_AIAN	6-year graduation rate of American Indian/Alaska Native	0	1	0.4798	0.3318
C150_4_NHPI	6-year graduation rate of Hawaiian/Pacific Islander	0	1	0.5400	0.3901
TUITIONFEE_IN	in-state tuition	3260	61788	21097.61	15631.87
TUITIONFEE_OUT	out-of-state tuition	4208	61788	29095.23	11962.95
MD_EARN_WNE_P6	median earning 6 years after enrollment	25641	88873	44130.56	9346.41
MD_EARN_WNE_P8	median earning 8 years after enrollment	30193	102547	48895.15	10306.86
MD_EARN_WNE_P10	median earning 10 years after enrollment	28956	103246	53062.33	11285.38

Figure 2 are the box plots of variables to depict their distributions

From the above data descriptions, we know that

- the average SAT admission score is around 1100
- the enrollment share of American Indian and Native Hawaiian account for almost zero
- a completion rate of American Indian and Native Hawaiian have large fluctuations due to the small sample size, which should also be noted in the subsequent analysis
- the overall tuition level of out-of-state students is higher than that of in-state students
- the salary level of students will increase with the increase of employment years.

Methodology

Data visualization and linear regression model are adopted to analysis from both abstract and quantitative perspective. This paper first adopts line charts and box-plots to describe the possible indicators for average SAT score and the derived focus about Asian American-serving institutions. Then I design a multiple linear regression model to find what variables are statistically significant to the average SAT score. The initial model included all the 20 variables (19 numeric variables and 1 categorical variable) as indicators.

The dataset was processed and analyzed in R(R Core Team 2021) and I analyzed all these using R package including: tidyverse(Wickham et al. 2019), reshape2(Wickham 2007), patchwork(Pedersen 2020), car(Fox and Weisberg 2019).

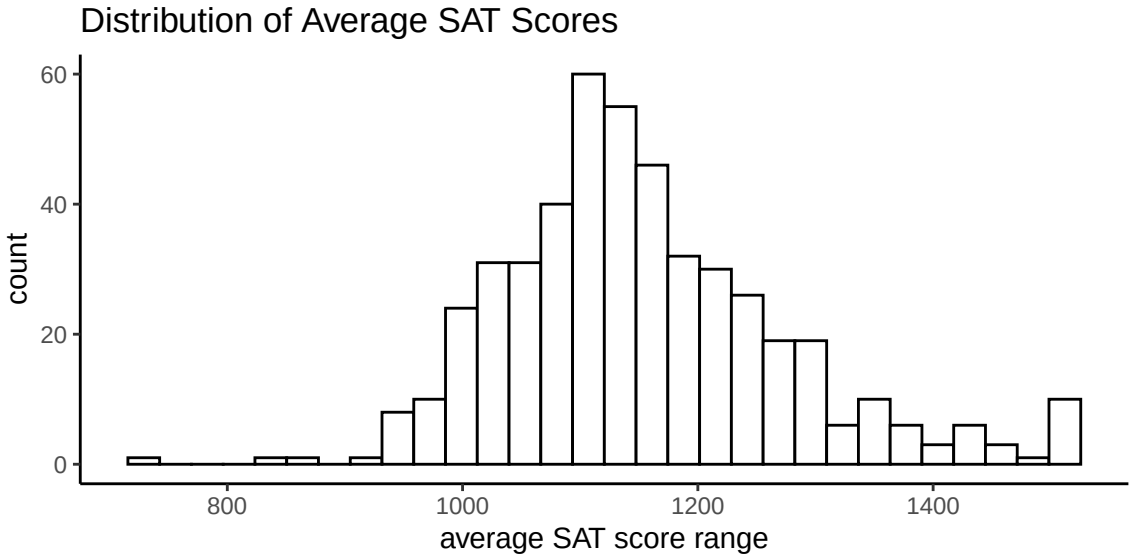

Figure 1: average SAT admission score of institutions

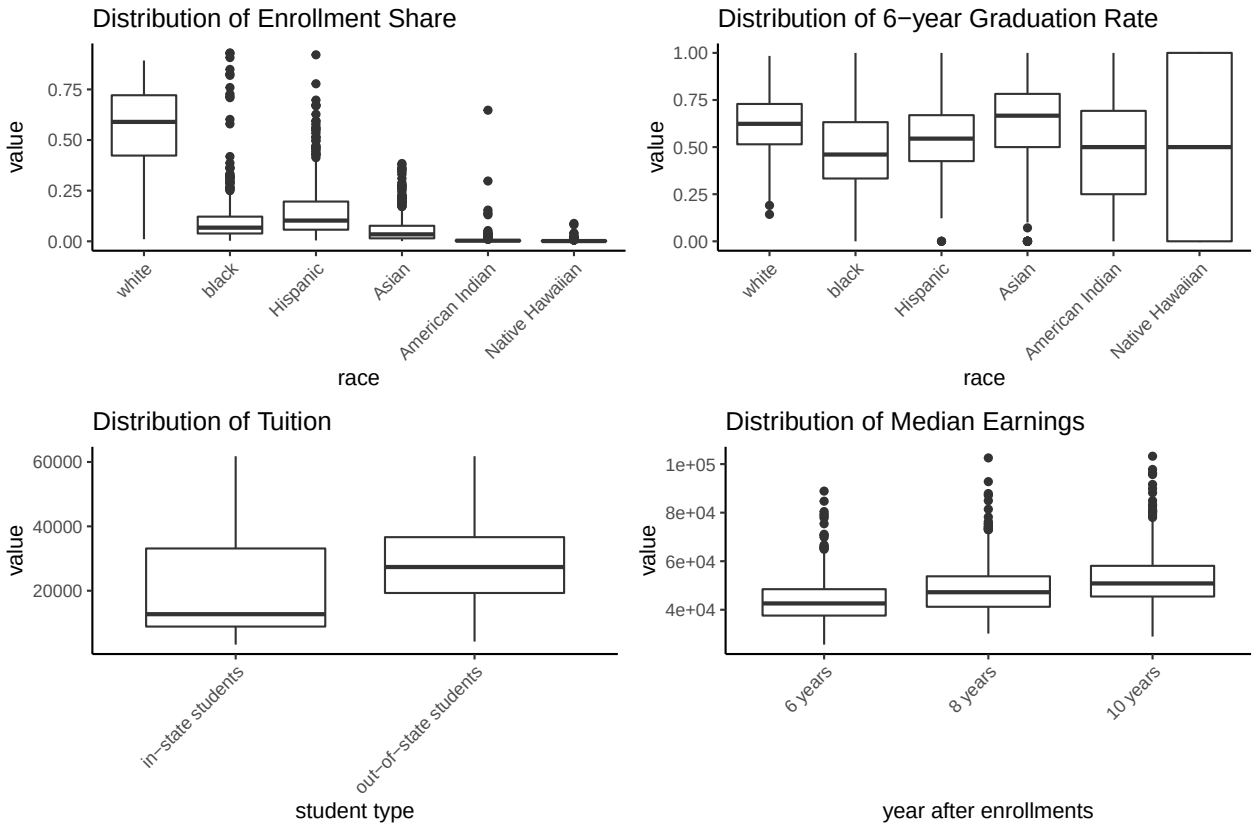

Figure 2: key variables description

Result

Characteristics of institutions with different average SAT score range

In the process of students' school selection, many questions are often raised, such as: What is the school's gender and race ratio? Is the school public or private? What is the school's graduation rate? Are tuition fees expensive? What is the salary level of graduates? By dividing the average SAT score into 9 segments with break of 100, the level of institutions and their different performances in these dimensions are more intuitively presented.

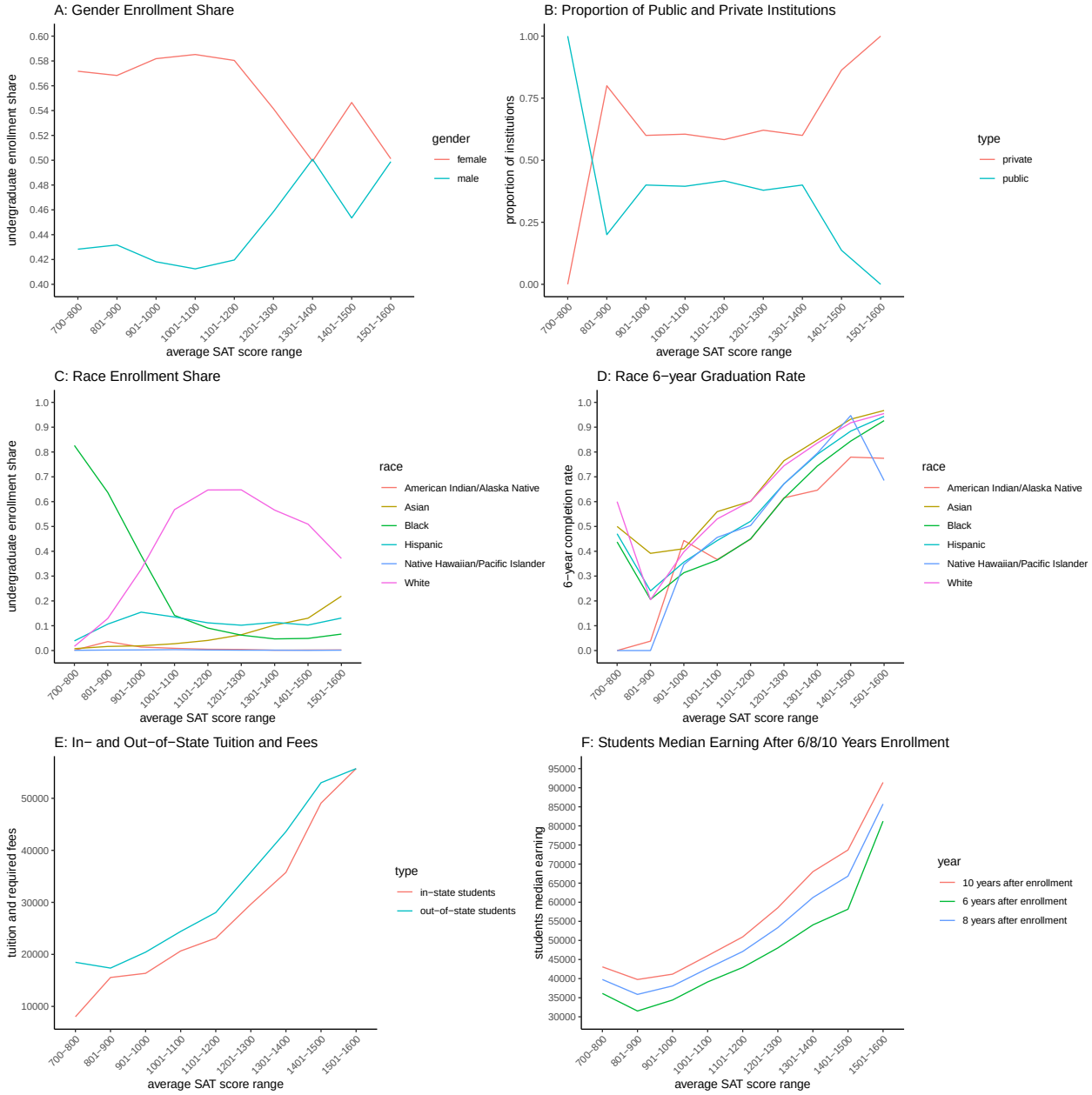

Figure 3: characteristics of institutions with different avg SAT score range

Panel A depicts the enrollment share of institutions with different average SAT score ranges by gender. It indicates that the enrollment rate of female is higher than that of male, but they are not as good as males in

terms of academic performance: the overall enrollment share of female declines with the increase of SAT admission score range. In addition, the gap in academic performance between is most pronounced in the low score segment, ie. 1200 and below, and the gap narrows overall after entering the mid-to-high score segment.

In panel B, it draws the average proportion of public and private institutions in different average SAT admission score range. It can be seen that the overall admission level of private schools is higher than that of public school. Such trends may indicate that the overall quality and popularity of private schools is higher than that of public schools because the average admission SAT score is an indicator of the admission intensity to the school. When the competition is fiercer, the admission score line is raised higher. In return, the high-quality students will further promote the academic and employment achievement of private institutions and form a virtuous circle.

The average enrollment rates of six main races in the institutions with different average SAT admission score ranges are shown in panel C.

- White has the best performance overall. Specifically, it has the highst overall proportion and distributes in a concentration of middle and high score segments. This trend is related to the population size and social advantage of White in American society, making them the main group on campus and have a relatively good educational background.
- The education situation of the Black is not optimistic, mainly concentrated in institutions with an average admission SAT score of 700-1000. The proportion of Blacks drops sharply with the increase of average SAT admission scores. Nonetheless, among the six races, Black is the race with the lowest enrollment rate except for the Native hawaiian/Pacific islander and American Indian/Alaska Native, which are about 0%. The extremely low proportion of these two races, however, is mainly resulted by the their sparse population.
- The enrollment rate of Asian tends to increase with the average SAT admission score. They have the lowest proportion in the 700-800 score segment. Then, an upward trend is shown and they reach the highest proportion at score range of 1501-1600. It reflects the significance that Asians place in education as well as the proficiency in SAT test. We cannot conclude that Asians perform better in admissions to all high-quality institutions, but the high-quality institutions they successfully attend are more likely to put higher percentage in SAT scores in admission. For example, a university with an average admission SAT score range of 1501-1600 may be as good as another university in the range of 1501-1600, but Asians are more likely to enter the institution with a higher SAT admission score.
- Hispanics show a relatively average performance in all ranges of average SAT admission score. We can conclude that the educational background of Hispanics students is more average and has no obvious preference.
- Native hawaiian/Pacific islander and American Indian/Alaska Native are two most extreme races in the performance of enrollment rate. Their proportions are close to zero in all score segemnt, which is related to their mere population size.

Panel D illustrates the 6-year graduation rate of six main races in 4-year institutions with different average SAT admission score range. In general, the 6-year graduation rate improves with institutions' average SAT admission score. It is reasonable to decide the degree of excellence of students by seeing that if they are able to graduate in 6 years, then this graph draws the conclusion that the SAT score is an effective criterion and better students can be selected. On the other hand, it shows that though high-quality institutions generally have more difficult curricula, students are not be hindered to complete their education.

- Asians performs an overall raising trend, with highest graduation rate in institution of every admission score segment. In the score segment of 1501-1600, its graduation rate reaches the peak, which is about 95%. This again reflects the importance Asians place on education, especially on the undergraduate degrees. Specifically, they are more prone to work hard to complete their study after entering the universities.
- Whites are close behind. Their graduation rates are quite close to Asians in the 901-1600 score range and also the reaches the highest graduation rate in the 1501-1600 range, about 94%.

- The trends of Hispanic and Blacks are relatively similar.
- At score range of 700-800, Native hawaiian/Pacific islander and American Indian/Alaska Native present a 0% graduation rate because there are no students of any of their races in this set of data. Since the sample size is small, the trend fluctuation is amplified and with large bias.

Panel E depicts the average tuition fees charged to in-state and out-of-state students by institutions with different average SAT admissions band. Generally speaking, schools with lower admission scores charge less, and schools with higher admission scores charge more. It may cause by the low attractiveness to the students, which fail to enable them collecting higher level of tuition fee. Moreover, according to the panel B, we know that there are more private schools with high admission score and they charge higher tuition fees, which also cause the phenomenon presented in panel E. In addition, tuition for in-state students is completely lower than tuition for out-of-state students.

Panel F reviews students' median earning of institutions with different average SAT admission scores after 6, 8, and 10 years after student enrolls. In general, average student earnings rises as the school admission score goes up and students at all institutions earn more as they works longer.

In summary, these variables all show some graphical association with the average SAT score.

Derived from Asian students: further discussion about Asian American-serving institutions

From the figure 3 panels C and D in the previous section, we can observe a prominent feature: compared with other races, Asian students have the most distinct trend of attending high-quality universities and have the highest graduation rate. It thus draws our attention to make further exploration of this population group. Since there is no direct race-level information, we select institutions with more concentration of Asian characteristics to make an analysis.

Among academic institutions, several have been awarded accreditation as Asian-American-/Native American-Pacific Islander-serving Institutions. These institutions meet the criteria that at the time of application, have an enrollment of undergraduate students that is not less than 10 percent students who are Asian American or Native American-Pacific Islander. Given my data description, the enrollment rate of Native American Pacific Islanders was found to be close to zero. So Asian American-/Native American-Pacific islander-serving (AANAPIS) institutions can be generalized as Asian American-serving institutions.

In general, the two main questions we raise to study are "does Asian American-serving institutions have higher average SAT admission score?" and "does Asian American-serving institutions have advantages in students' future performance?" To achieve this goal, I compare variables of average SAT score, student loan debt, and tuition fees between AANAPIS institutions and non-AANAPIS institutions.

Average SAT Score

Figure 4 depicts the difference in the distribution on average SAT Score of AANAPIS institution versus non-AANAPIS institution. It can be seen that the median average SAT score of AANAPIS institution is about 1150, which is slightly higher than that of other institutions (1100). Furthermore, the 25 percentile of AANAPIS institution is slightly higher than the 25 percentile of others, while the 75 percentile is almost the same. The outliers of non-AANAPIS institutions are higher than those of AANAPIS institutions, which is normal because of the larger number of institutions involved.

In general, there is no significant gap between the two types of institutions. However, according to the result in figure 3 panel C, institutions with average SAT score between 1501 and 1600 have an average enrollment rate of more than 20% of Asians. If all the eligible institutions apply to be AANAPIS institutions, the distribution of average SAT score should be more right-skewed.

Under this situation, any positive conclusions we draw from studying AANAPIS institutions should more depend on the institution itself, instead of advantages bring by larger proportion of Asian students.

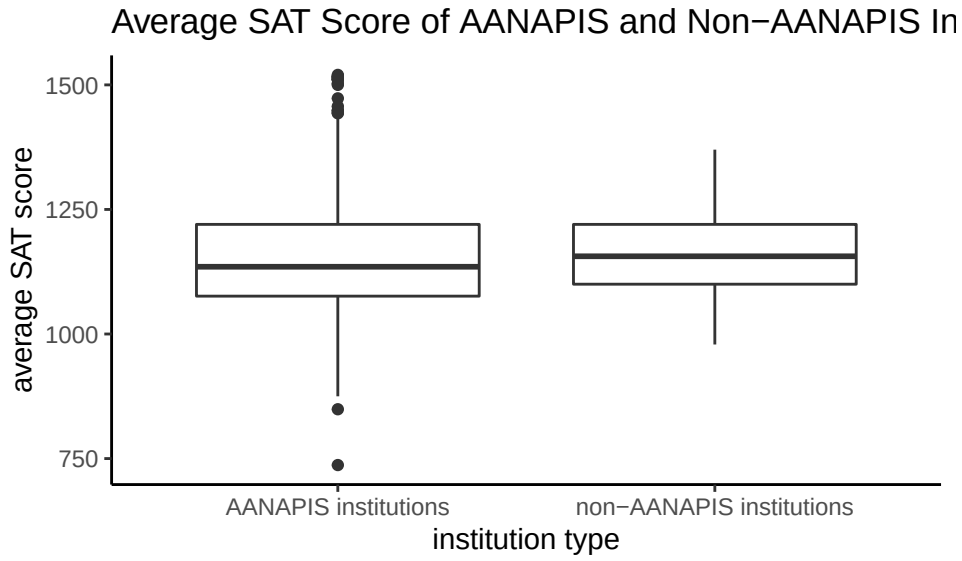

Figure 4: comparison between average SAT score

Students Earnings

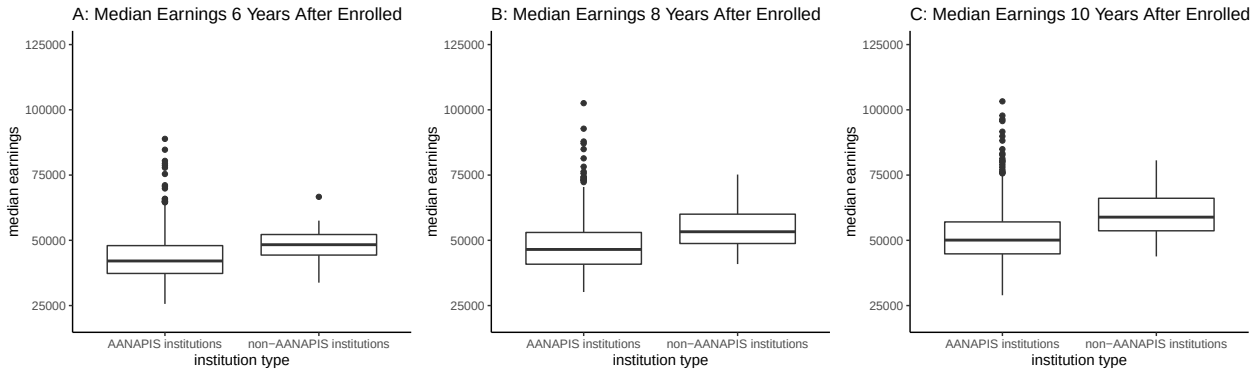

Figure 5: Median earnings after 6/8/10 years enrollments

Figure 5 depicts the median salary distribution of students from AANAPIS institutions and non-AANAPIS who enrolled for 6, 8, and 10 years. It can be seen that the overall distribution of median salary of AANAPIS institutions' students is higher than that of students from other institutions in all the panels. Specifically, after 6 years of enrollment, the median salary for non-AANAPIS institutions' students is around $38,000, while the median salary for AANAPIS institutions' students is around $47,000 US dollars, which is almost the same as the 75 percentile for non-AANAPIS institutions' students. The lower whisker of student earnings of AANAPIS institutions is even slightly higher than the median of that of non-AANAPIS institutions. This pattern remains unchanged in 8 years and 10 years of enrollment. The 50% percentile of median earnings of AANAPIS institutions students exceeded $50000 after 2 years and reaches $58000 after 4 years. It is worth noting that 10 years after enrollment, the distribution of student earnings of AANAPIS institutions changes from almost no skewed to slightly right-skewed, which means that their students with high salaries increased, and their 75 percentile reached about $82000.

As we discovered from figure 4, the average SAT score of AANAPIS institutions can be summarized as 1150. From figure 3 panel F, the average salary of schools with admission scores between 1100-1200 is around $40,000, $43,000, and $48,000 after 6, 8, and 10 years of enrollment, which are lower than the median salary for AANAPIS institutions students in this graph. Thus, it can be preliminarily judged that entering

AANAPIS institutions can help students achieve higher salaries in the future.

Student Loan Debt and Average Annual Cost

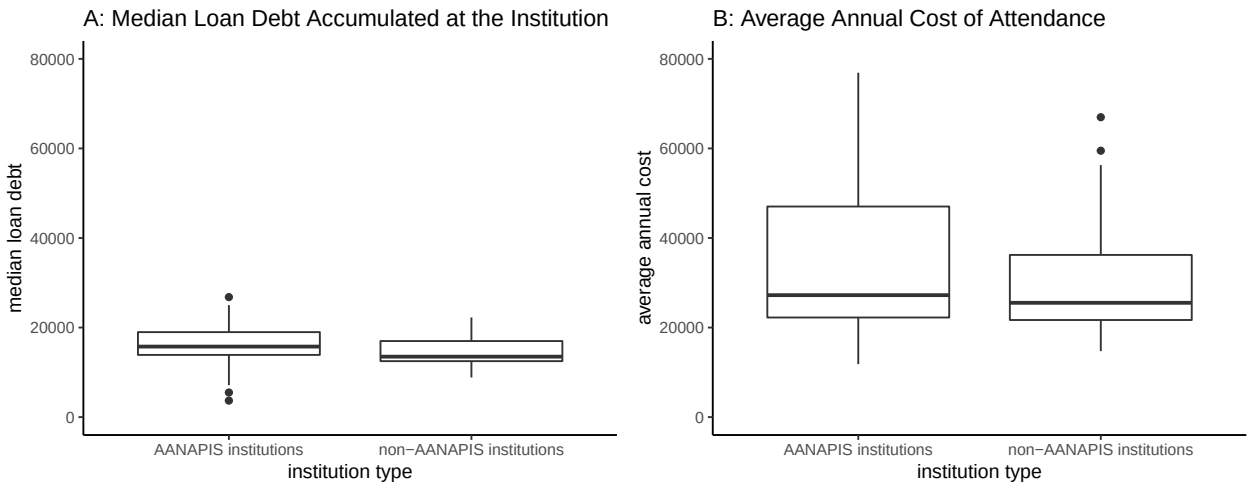

Figure 6: student loan debt and average anual cost

The median loan debt accumulated at the institution by all student borrowers of federal loans who separate (i.e., either graduate or withdraw) in a given fiscal year is measured at the point of separation. More specifically, the measure represents the sum of all undergraduate federal loans over students' college education at the institution for which the median debt is reported.

As can be seen from figure 6 panel A, the median loan debt borne by AANAPIS institutions' students is lower than students from other institutions. Specifically, the median loan debt of AANAPIS institutions is about $13500, roughly on par with the 25 percentile of non-AANAPIS institutions. The reason for the lower loan debt may because of the better family situation of students instead of the advantages provided by the universities. Therefore, I compare the annual cost of students from AANAPIS institutions and non-AANAPIS institutions (panel B). It shows that the overall tuition of AANAPIS institutions is lower than that of others, with a median difference of about $13,000, which is much higher than the difference in the median debt between the two types of institutions. This, in turn, suggests that AANAPIS institutions contribute to the low loan debt of their students.

In summary, the overall tuition level of AANAPIS institutions is relatively low, which makes the loan burden of students lighter. Meanwhile, they help students obtain higher salaries than the institutions with equivalent admission SAT scores. We can conclude that AANAPIS institutions is helpful to students' development prospects, especially in the financial aspect.

Estimating institutions average SAT admission score by linear regression model

In the above two sections, I draw conclusions based on data visualizations. Specifically, figure 3 describes the relationship between six institution attributes and the average SAT admission score. Now, I aim to associate these attributes with the average SAT admission score in a more accurate quantitative way, to achieve the purpose of estimating the future admission score of an institution through its characteristics and historical performance. As mentioned earlier, American Indians and Native Hawaiian have extremely small sample sizes and unreliable fluctuations, so I exclude these two races when considering enrollment share and graduation rate. Therefore, there are a total of 15 initial independent variables, which are shown in the following table. The final model is decided by auto-selection.

The regression model with real meaning is given below.

$$\begin{aligned}\widehat{\text{Average SAT scores}} = 1018 &+ 118.9 \cdot \text{six year graduation rate of Blacks} \\ &+ 135.5 \cdot \text{six year graduation rate of Hispanics} \\ &+ 54.34 \cdot \text{six year graduation rate of Asians} \\ &+ 0.58 \cdot \sqrt{\text{Tuition of out-of-state Student}} \\ &- 23.1 \cdot log(\text{Enrollment share of Blacks}) \\ &- 42.1 \cdot log(\text{Enrollment share of Hispanics}) \\ &+ 26.4 \cdot log(\text{Enrollment share of Asians}) \\ &- 40430 \cdot \frac{1}{\sqrt{\text{Median earnings after six year enrollment}}}\end{aligned}$$

Here is the regression output.

Variables	Estimate	Std. Error	t value	Pr(>
(intercept)	1.018e+03	6.756e+01	15.071	<2e-16
C150_4_BLACK	1.189e+02	2.469e+01	4.815	2.14e-06
C150_4_HISP	1.355e+02	3.024e+01	4.482	9.85e-06
C150_4_ASIAN	5.434e+01	1.995e+01	2.724	0.006745
TUITIONFEE_OUT^(0.5)	5.768e-01	1.272e-01	4.534	7.79e-06
log(UGDS_BLACK)	-2.314e+01	3.763e+00	-6.151	1.98e-09
log(UGDS_HISP)	-4.209e+01	4.681e+00	-8.991	<2e-16
log(UGDS_ASIAN)	2.641e+01	4.879e+00	5.413	1.11e-07
MD_EARN_WNE_P6^(-0.5)	-4.043e+04	1.184e+04	-3.415	0.000707

All the variables are statistically significant.

According to the model, we can conclude that

- the average SAT scores increases by 118.9 on average while the six-year graduation rate of Blacks increase by 1
- the average SAT scores increases by 135.5 on average while the six-year graduation rate of Hispanics increase by 1
- the average SAT scores increase by 54.34 on average while the six-year graduation rate of Asian increase by 1
- tuition of out-of-state students has a positive effect on average SAT scores and the marginal effect is decreasing
- the enrollment share of blacks has a negative effect on average SAT scores and the marginal effect is decreasing
- the enrollment share of Hispanics has a negative effect on average SAT score, which is stronger than that of blacks, and the marginal effect is decreasing
- the enrollment share of Asians has a positive effect on average SAT scores and the marginal effect is decreasing
- the median earnings after six-year enrollment has a positive effect on average SAT score and the marginal effect is decreasing

The following 2 graphs show the goodness of fit of the linear regression model. In residuals versus fitted graph, the points are scattered around residual of zero and there is no discernible pattern, which indicates good linearity. Independence and constant variance are also satisfied as there are no large clusters and fanning patterns. There are few leverage points in the normal Q-Q plot, so the normality may be slightly violated.

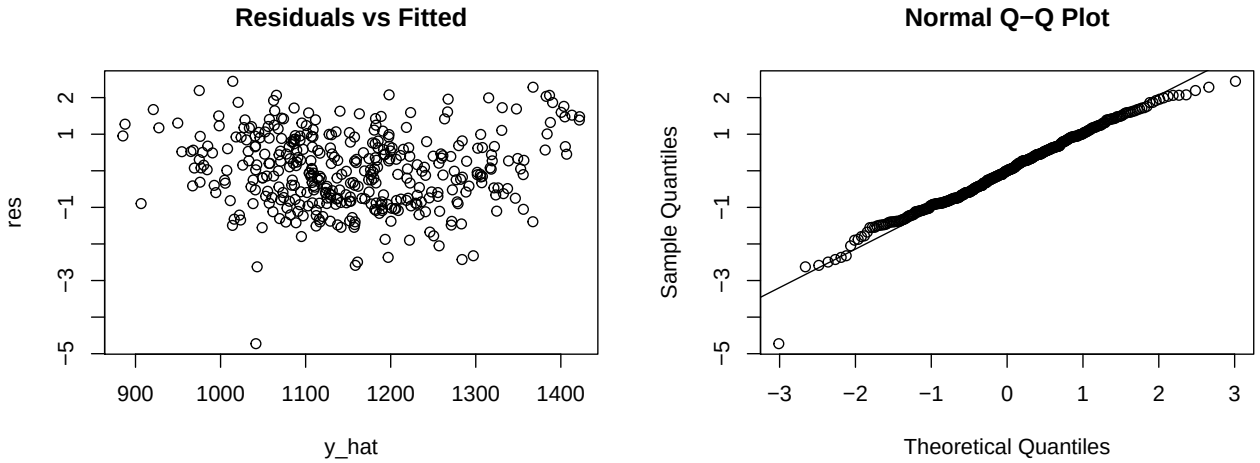

Figure 7: goodness of fit for the regression model

In summary, there are consistent conclusions that are drawn from the linear regression model and data visualization (c.f. graph 2). Specifically, Asian students indeed present intentions to the institutions with higher SAT admission scores, which in turn drive the estimated score higher when there are a large proportion of Asians. In contrast, the negative effect of black students' enrollment rate also has been reflected by the downward trend in graph 2 panel C.

Moreover, on the contrary to the negative effect of enrollment share produced by blacks and Hispanics, their six-year completion rates make a positive influence on the average SAT admission score. Meanwhile, they make a stronger effect than the completion rate of Asian as they have more extreme performance in different score segments as we find in graph 2 panel D.

At last, it is reasonable to see that the median earnings after a six-year enrollment rate have a positive effect on the average SAT admission score, which is also consistent with the visualization in graph 2 panel F.

Discussion

Findings

1. SAT is an effective indicator of the goodness of institution performance.

By categorizing undergraduate institutions by SAT score band with the break of 100, this paper plots the racial average 6-year graduation rates and students' earnings for institutions within each score band. Completion rates can reflect student performance in school, while earnings can reflect how competitive a student will be in society after enrollment. They are both significant factors in evaluating the goodness of an institution. Based on the line graphs, the results show that graduation rates and students earnings after 6/8/10 years for all races increased with the average SAT score of their institution. It indicates that students in schools with higher SAT admission scores perform better in school, which may be because they have a better quality of students or teaching; In addition, after receiving a university education, their students have higher social competitiveness and obtain monetary returns, which is also one of the main purposes of attending higher education.

2. Asian students have a higher share of outstanding groups in undergraduate education, while the opposite is true for blacks.

This conclusion was drawn by looking at the enrollment share and 6-year graduation rate of different races. Enrollment share represents whether the student can enter a school, and the graduation rate shows whether the student has sufficient academic ability to graduate. In terms of enrollment rate, Asians show higher enrollment rates as academic excellence rises. The admission rate in institutions with very low average SAT

scores is almost zero; blacks, on the other hand, have very high enrollment rates at institutions with very low admission scores but drop off a cliff when they reach the average admission score of institutions. Second, in terms of the 6-year graduation rate, the overall performance of Asians is better than that of all races, while blacks are ranked second from the bottom, only after American Indian/Alaska Native, which has few sample size. This also reflects that Asians place a higher emphasis on higher education, while blacks have the opposite.

3. Asian-American-/Native American- Pacific Islander-serving Institutions are well supportive of students' success.

In terms of average admissions SAT scores, Asian-American-/Native American- Pacific Islander-serving (AANAPS) Institutions are only slightly higher than other institutions. However, in terms of student earnings, AANAPS institutions have shown greater advantages. In particular, the average student earnings of AANAPS institutions are higher than the average of institutions with the same admission score, which indicates that the institution provides students with above-average educational services. In addition, students at AANAPS institutions have below-average loans, which may partly be due to their below-average tuition. In summary, the overall tuition level of AANAPIS institutions is relatively low, which makes the loan burden of students lighter. Meanwhile, they help students obtain higher salaries than the institutions with equivalent admission SAT scores. Therefore, one can conclude that the AANAPIS institution is above-average in helping students' future development, especially in terms of finances.

4. Multiple linear regression model to estimate institutions' average SAT admission score.

According to the model, the six-year completion rate of Blacks/Hispanics/Asians has a positive linear effect on the average SAT score. The enrollment share of Blacks and Hispanics negatively influences the average SAT score but in a decreasing marginal decrease. Meanwhile, the enrollment rate of Asians positively influences the score in a decreasing marginal effect as well. Specifically, Hispanics make a stronger effect than Asians as they have more extreme performance in different score segments. The median earnings after six-year enrollment is more statistically significant than that after eight and ten years of enrollment, which in a decreasing marginal increase. The above quantitative results are consistent with the graphical conclusion. In addition, tuition of out-of-state students has positive effect as well, but the marginal effect is decreasing.

Weakness

1. Inaccurate graduation rate

Data have several important limitations for measuring institutional performance. Specifically, graduation rates are only reported for cohorts of full-time, first-time students, so graduation rate information is not available for students who may have previous higher education experience or for part-time students. Moreover, outcomes are not recorded for students who transfer from the institutions. As a result, information on graduation rate outcomes is limited.

2. Under assumption of homogeneity

The number of schools in each fractional segment varies, even widely. For example, After the data is cleaned, there is only one school in the 700-800 range. This resulted in a large variance in the number of schools between each average SAT admission score segment. The results presented are reasonable based on assumption of homogeneity across institutions in each score band. But if there are actually differences between schools with the same score, then the sample is unbalanced and the result is biased.

Future works

1. Completion discussion on races

In the 2022 single-year dataset, the sample size of Native hawaiian/Pacific islander and American Indian/Alaska Native students is too small, resulting in undeniable bias. As a result, it has to be artificially excluded from the study. However, these are certainly groups that should not be ignored. Therefore, in future research, the sample size can be supplemented by adding datasets from the past years to conduct research. In addition,

the least squares method can be used to establish a linear regression equation on the average SAT score and races, and the significance of different races can be examined by t test.

2. Discover the field of study data

As mentioned above in Data section, there is another data file containing the data about specific fields of study within institutions. The major selection is also a key problem that most students have to face at the first or second year they get into universities. As a study that aim to help students do better selections, more sections to discover the fields of study within institutions should be added in. More specifically, we can again analysis the factors that have already covered, including gender and race enrollment share, graduation rate, students earnings, cost etc. As a result, one can not only make conclusion about fields study, but also understand the difference of fields performance in different level of institutions by average SAT score.

Appendix

Data sheet

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
   - The dataset was created to enable analysis of US postsecondary institutions. It can increase transparency and put the power in the hands of students and families to achieve their educational goals.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
   - The dataset was created by College Scorecard on behalf of U.S Department of Education.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
   - This part of information is missing
4. *Any other comments?*
   - None

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
   - There are two types of instances that comprised the dataset, including student performance information and institutional information.
2. *How many instances are there in total (of each type, if appropriate)?*
   - there are 327 student performance information data and 153 institutional information data.
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
   - The dataset is a sample of US postsecondary institutions and all students they enrolled. The dataset can be concluded as validated representative of the population. Because under the US Higher Education Act, all institutions that participate in Title IV federal student aid programs must complete the questionnaires. However, there is still some limitations. The most significant is that many outcomes are recorded for a limited subset of students.
4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*
   - There is no raw data.
5. *Is there a label or target associated with each instance? If so, please provide a description.*

- There is no label or target associated with each instance.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*
   - There is information missing in each instances because FSA data center does not list an active primary institutional accreditor.
7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
   - The studeny performance and institution characteristics may be mutually causal.
8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*
   - The data was splitted into training and testing segments while designing linear regression model. More specifically, 80% of the data are assigned to testing dataset and 20% of them are assigned to that of training.
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
   - There are no errors, sources of noise, or redundancies in the dataset.
10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*
    - The dataset is self-contained.
11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*
    - There is no data that might be considered confidential in the dataset.
12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
    - There is no data that might be offensive, insulting, threatening, or cause anxiety in the dataset.
13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
    - The dataset identifies population by gender or race. It concentrated on the stance of student performance.
14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*
    - It cannot identify individuals. All the data are institution-level.
15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*
    - The dataset does not contain such data.
16. *Any other comments?*
    - None.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*
   - The data is reported by subjects.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*
   - The data is collected through software APIs.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
   - No, the dataset is not a sample from a larger set.
4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*
   - This part of information is missing.
5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*
   - The data was collected in 2022 and last updated on March 14, 2022. Some data in this timeframe do not match the creation timeframe of the data associated with the instances, including information about student body, earning, and repayment, which were updated in 2018.
6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*
   - This part of information is missing.
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
   - Some data was collected through federal reporting from institutions, data on federal financial aid, and tax information. Others are collected through survey directly.
8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*
   - This part of information is missing.
9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*
   - Yes, because all institutions that participate in Title IV federal student aid programs must complete the IPEDS questionnaires.
10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*
    - This part of information is missing.
11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - This part of information is missing.
12. *Any other comments?*
    - None.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
   - There is no preprocessing/cleaning/labeling of the data.
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
   - There is no "raw" data.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

- This part of information is missing.

4. *Any other comments?*
   - None.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
   - Yes, this dataset is used for College Scorecard project, which is designed to increase transparency, putting the power in the hands of students and families to compare how well individual postsecondary institutions are preparing their students to be successful.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
   - There is no repository that links to any or all papers or systems that use the dataset.
3. *What (other) tasks could the dataset be used for?*
   - Tasks about US undergraduate institutions.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
   - Some data has stoped update in 2018, which may influence the judgement to some institutions that performed big fluctuations in recent years.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
   - This part of information is missing.
6. *Any other comments?*
   - None.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
   - This part of information is missing.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
   - The dataset is dictributed through U.S. Department of Education website and API.
3. *When will the dataset be distributed?*
   - The datset is distributed annually.
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
   - This part of information is missing.
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
   - This part of information is missing.
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
   - This part of information is missing.
7. *Any other comments?*
   - None.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

- Department of Education's National Center.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
   - Department of Education's National Center can be contacted through website and email address.
3. *Is there an erratum? If so, please provide a link or other access point.*
   - There is no erratum.
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
   - The dataset will be updated annually by Education's National Center in their official website.
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
   - This part of information is missing.
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*
   - The older versions of the dataset will continue to be supported and can be downloaded in a link in website with all data files.
7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*
   - This part of information is missing.
8. *Any other comments?*
   - None.

Reference

Clinedinst, M. E, and D. A. Hawkins. 2010. "NACAC State of College Admission 2010. Arlington." *National Association for College Admission Counseling*.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Grissmer, David Flanagan, Ann Williamson, and Stephanie. 1998. "Why Did the Black–White Score Gap Narrow in the 1970s and 1980s?" *American Psychological Association*.

Grissmer, D. W., Kirby, S. N. Berends, and M.& Williamson. 1994. "Student Performance and the Changing American Family." *PsycEXTRA Dataset*. https://doi.org/10.1037/e419112005-001.

Hedges, Larry V., Nowell, and Amy. 1998. "Black–White Test Score Convergence Since 1965." *American Psychological Association*.

Pedersen, Thomas Lin. 2020. *Patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, Michael, and James Monks. 2005. "Making Sat Scores Optional in Selective College Admissions: A Case Study." *Economics of Education Review*. https://doi.org/10.1016/j.econedurev.2004.06.006.

Tierney, William G., Jack K. Chung, William G. Bowen, and Derek Bok. 2000. "The Shape of the River: Long-Term Consequences of Considering Race in College and University Admissions." *The Journal of Higher Education*. https://doi.org/10.2307/2649250.

Wickham, Hadley. 2007. "Reshaping Data with the reshape Package." *Journal of Statistical Software* 21 (12): 1–20. http://www.jstatsoft.org/v21/i12/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.