Preliminary report

Research question: Which university characteristics correspond to higher future income?

Introduction

“The SAT is not perfect. We all know smart, knowledgeable people who do badly on standardized tests. But neither is it useless. SAT scores do measure both specific knowledge and valuable thinking skills.” (Virginia Postrel)

According to Virginia Postrel, since SAT measures those aspects, itself and ACT in general are the standard tools for universities to evaluate applicants. Students will be able to apply for the university that they think will provide for them a well base of knowledge in the fields that they might be involved in after graduation. The process of choosing the right college is difficult; colleges always advertise themselves as the right choice, thus presenting those interested in seeking higher education from high-school with too much information. Hence, we hope to discover college characteristics that greatly correspond to the early career earnings such as SAT/ACT scores, estimated price without aid, estimated price with avg. grant, percentage of students who get any grants, average student debt. We hope our project will save senior high school students time choosing their college. Since our data contain the location of the colleges, the audience may choose to filter colleges by location should they wish to attend colleges near their homes.

High school students have all been informed about the importance of SAT or ACT as they are the primary keys to get to a good college. These exams require an adequate amount of knowledge that a student has gained through their study journeys. Therefore, the SAT/ACT does represent their learning capability. Based on that reasoning, we expect the ACT/SAT score will correspond to higher income. It’s also necessary to note that some universities take SAT/ACT as an optional test, this detail will be addressed in the project. Beside SAT/ACT, we also investigate other variables such as estimated price, location to see if they affect the future income of a graduated college student. We predict that location is not a promising category since it depends on many other variables such as the economy and policy of that location, which are not included in the dataset. The estimated price without aid, estimated price with avg. grant, percentage of students who get any grants, average student debt could be good candidates to investigate in this research question. We expect to see a relationship between those features of each university and the future income. Overall, this project will provide another aspect of SAT/ACT and estimated price whether it directly influences the future income or not. Moreover, the result will help students to decide whether they should focus on practicing for ACT/SAT tests or should look at other features

like estimated price and location to pick a university that can ensure a better earning potential.

Method

At first, we found a dataset from the Data And Story Library which compiled information from Money's college ranking website. The original link to the website was no longer available, suggesting that the dataset is out-dated, so we decided to create a scraper to retrieve the dataset from Money's ranking of colleges in the United States in 2019-2020, this website also provides the latest update from all universities about SAT/ACT scores, estimated price and other aspects. We used Beautiful Soup in Python to scrape the table into a string data matrix, perform necessary string parsing (such as separating the college name and its location) before exporting to a .csv file. We then matched some colleges presented in the dataset from Data and Story Library regarding whether each college is public or private before manually matching the newer entries with their respective category.

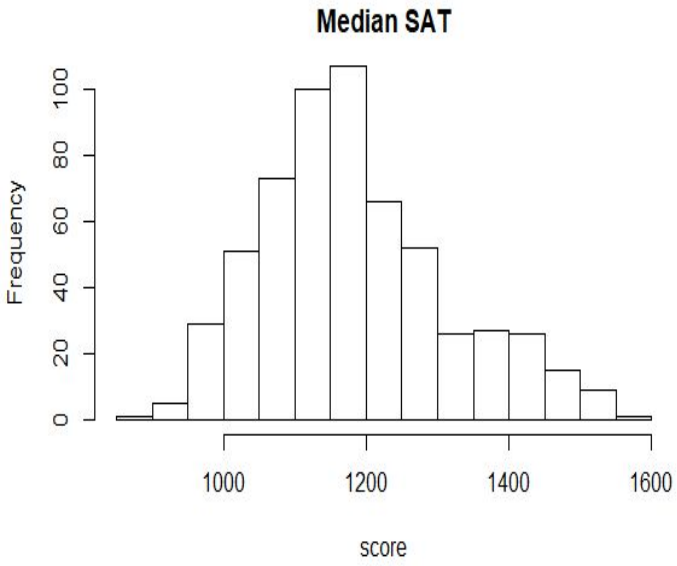
According to Money, the data collection and analysis process was led by the American Institute for Research. Each college in the dataset is guaranteed to have at least 500 students, have sufficient and reliable data for analysis, not be in financial distress, and have at least the median graduation rate of its category. The data on the estimated amount of money is retrieved from reported data from selected colleges to the US Department of Education. The early career earnings are retrieved from Payscale.com of alumni's earning at most three years after their bachelor's degree. Other reported data were gathered from Peterson's. While Money based their ranking on the dataset of many more features, we decided to use the published subset of their dataset for our research.

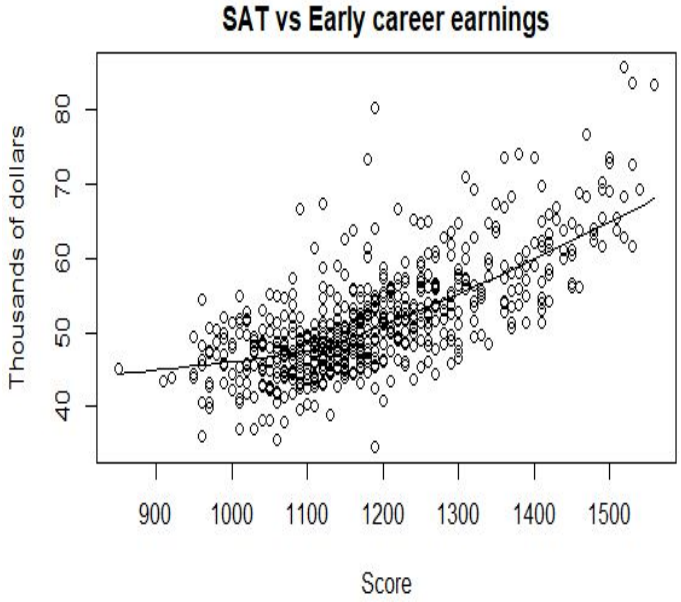
The explanatory variable in our project will be the characteristics of all universities in the United States and the response variable will be the early career earning. There are 9 variables in our dataset, three of which are categorical variables: school name, location, and rank. We expect to use 6 quantitative data as our explanatory variables in our project: median SAT score, median ACT score, estimated price without aid, estimated price with avg. grant, percentage of students who get any grants, average student debt. The response variable is early career earnings, which is quantitative data. We expect to see at least 2 explanatory variables will correspond either positive or negative to the early career earnings.

In recent years, colleges have become test-optional: including standardized test scores is optional. This may introduce bias in the median SAT and median ACT as students may choose to not include their low scores to make the application more appealing.

The obvious confounding variable is the location of each university as it may affect the college tuition fee and early career earnings due to local economy scaling and policies (taxes).

Results



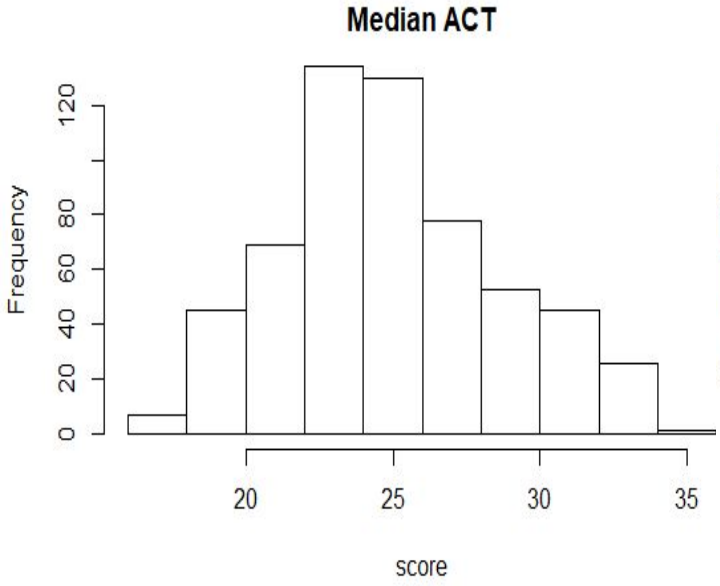


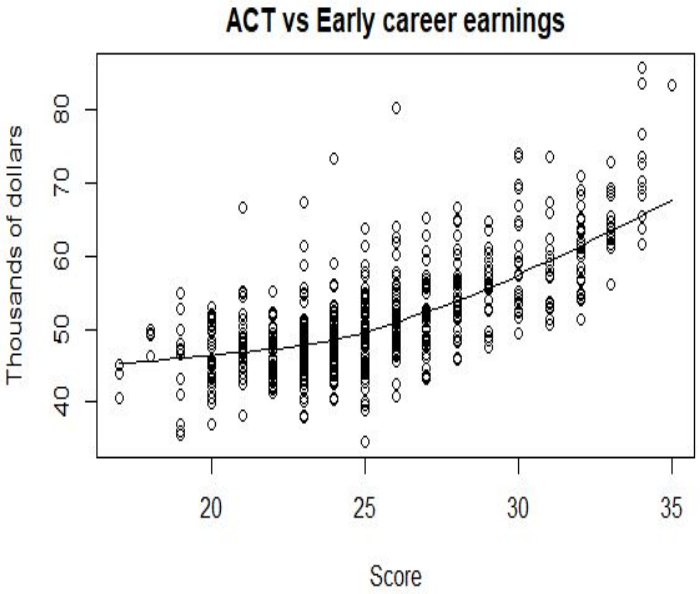
Summary of “Median SAT” variable:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
850	1098	1170	1187	1260	1560

Correlation between Median SAT and early career earning: 0.7091892

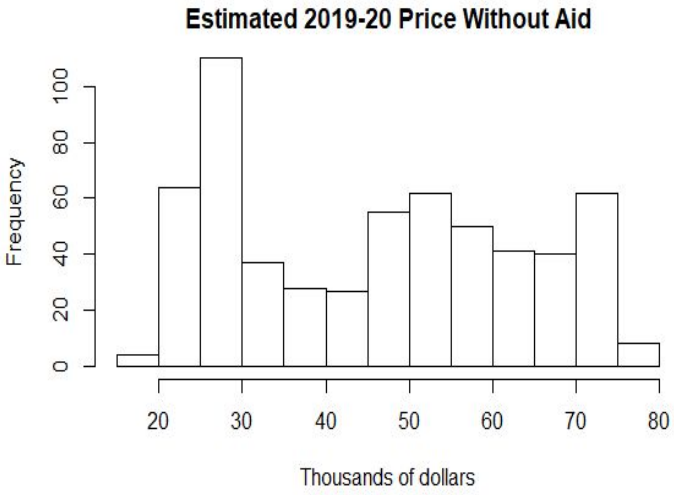
The shape of the histogram in the median SAT is slightly right-skewed. The mean (1187) surpasses the median (1170). The histogram of the frequency suggests that colleges tend to accept students with SAT scores slightly below the mean score (mean > median). In the plot between “Median SAT” and “Early career earnings”, we can see that there is a positive correlation between them. The data cluster around 1100, and the overall trend (the line shown in “SAT vs Early career earnings”) is positive. Moreover, the correlation between those 2 variables is 0.709, suggesting a strong relationship between the two variables. Therefore, we conclude that as the median of SAT increases, the earnings in the future also increases.

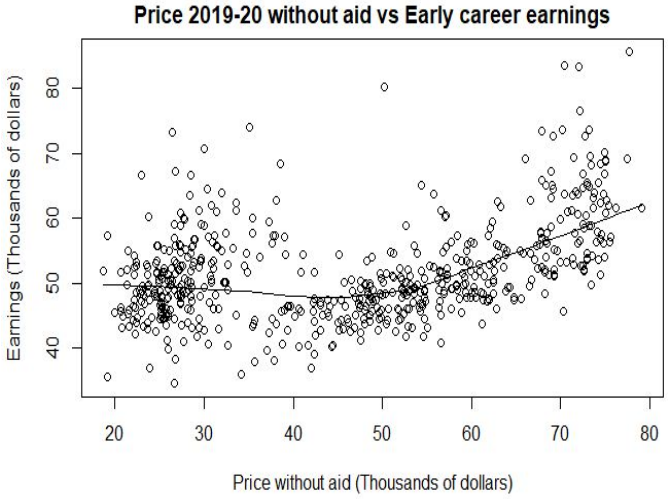




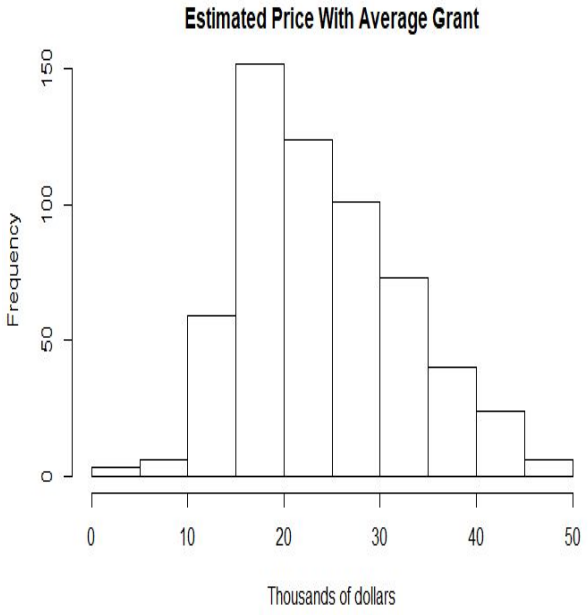
Summary of "Median ACT": Min. 1st Qu. Median Mean 3rd Qu. Max.
17.00 23.00 25.00 25.48 28.00 35.00
Correlation between Median ACT and early career earning: 0.6835765

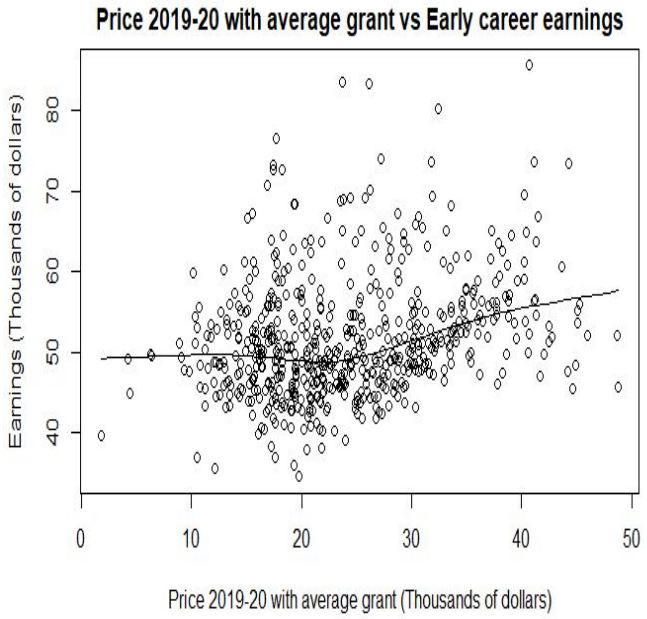
The median ACT score for the colleges in the database has the mean at around 25. The histogram of the frequency suggests that colleges tend to accept students with ACT scores slightly below the mean. The right skewness of the histogram suggests that colleges with median ACT score acceptance tend to be more specific on the requirements as the frequency spreads out as the score goes above the mean (25.5). In the second graph, we see that there is a positive correlation between the 2 variables. The correlation between those 2 variables also shows that they have a pretty strong correlation, approximately 0.7. This means as the median of ACT increases, the earnings in the future also increases.





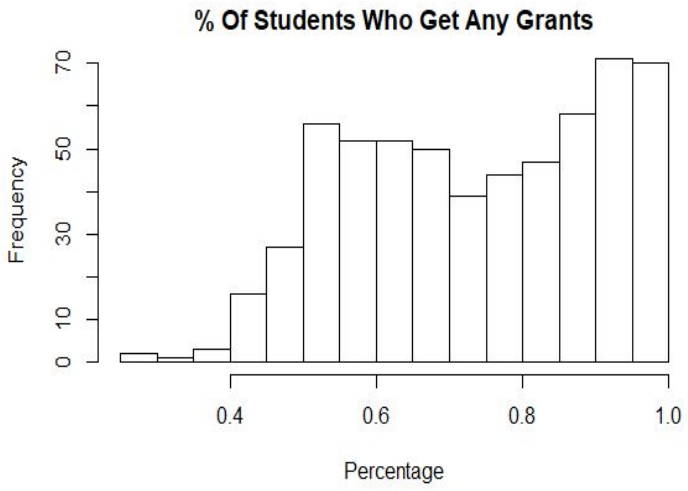
Summary: Min. 1st Qu. Median Mean 3rd Qu. Max.
18.70 28.27 47.95 46.15 60.42 79.10
corr : [1] 0.3952125
The histogram shows that the “Estimated 2019-20 Price Without Aid” does not follow any specific distributions. The scatter plot between the estimated price and the early career earnings shows that there is a positive relationship between those 2 variables since the line in the plot goes up. However, we can probably predict that this correlation will not be strong since the data all spread out. This prediction has also been proved by the correlation between the 2 variables. The correlation between those 2 variables is just approximate to 0.4, which is a weak correlation. Therefore, we can conclude that “Estimated 2019-20 Price Without Aid” does not correspond to the high future income.

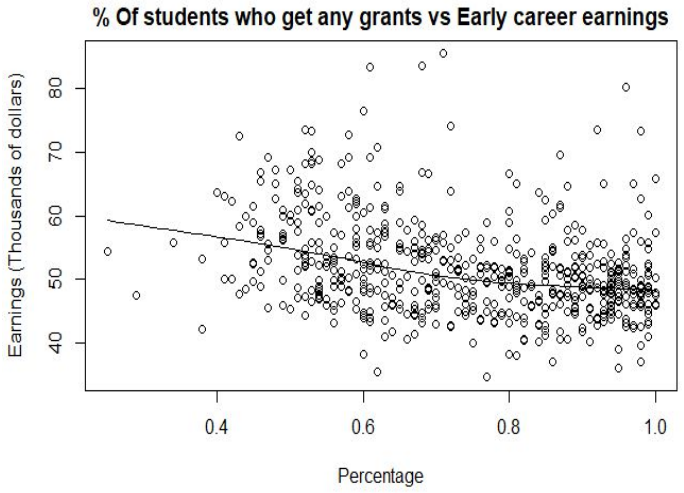




Summary: Min. 1st Qu. Median Mean 3rd Qu. Max.
1.80 17.68 22.50 24.08 29.73 48.80
Corr [1] 0.2645993

The histogram is right-skewed, suggesting that most selected colleges in the dataset require less fee than the mean price of alfdfsdfa selected colleges.We found that the correlation between the two features is too low for analysis. To put in context, this shows that Higher money investment in colleges does not necessarily correlate to the early career income. It does not, however, imply that investment in college tuition does not benefit the earning in the long run.

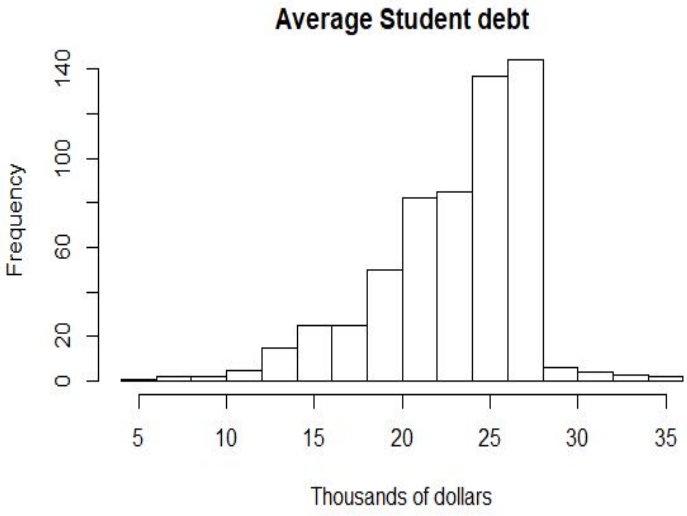


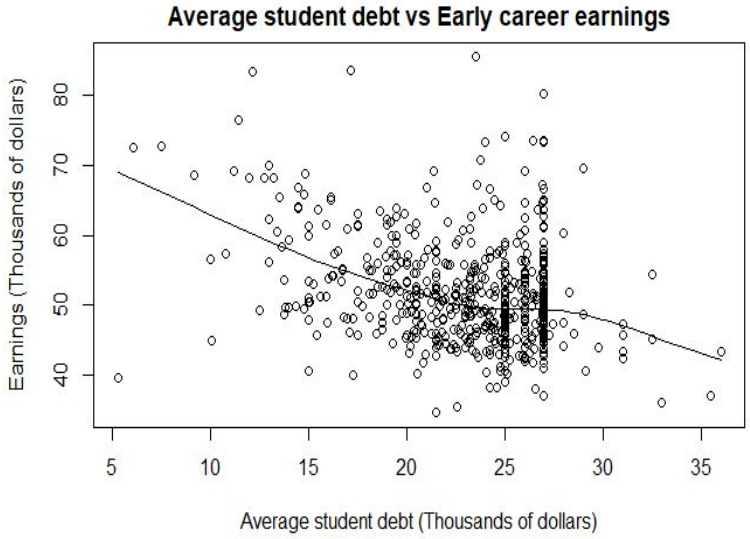


Summary of "% of students who get any grants": Min. 1st Qu. Median Mean 3rd Qu. Max.
0.2500 0.6000 0.7500 0.7419 0.9000 1.0000

Corr: -0.3194429

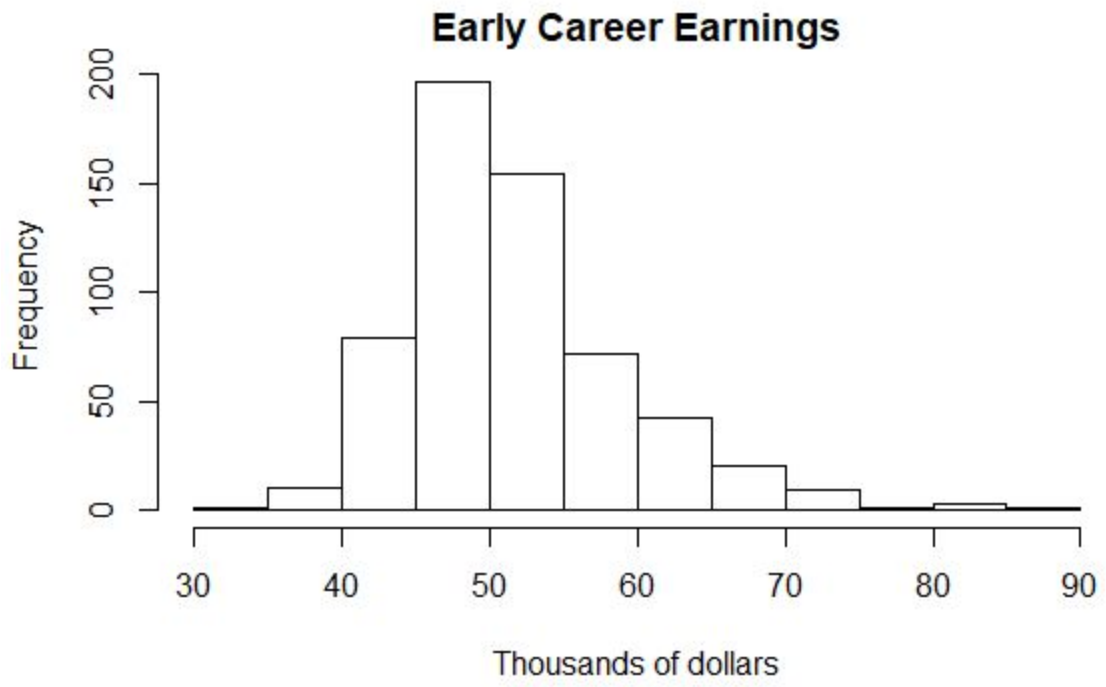
The histogram suggests that most colleges provide some kind of grant for around 75% of the students to cover the tuition fee. The minimum of 25% suggests that at least 25% of the students receive some kind of grant in the chosen colleges. There are also colleges that give grants to all of their students. The fact that the median and the mean surpasses 50% suggests that selected colleges support the majority of their students through grants. The correlation, however, suggests that the amount of support undermines the earning in their students' early career. This statement is not confidently asserted as the absolute value of the correlation suggests that this inverse relationship is poorly seen from the data set.





Summary : Min. 1st Qu. Median Mean 3rd Qua Max.
5.30 20.50 24.11 23.03 26.44 36.00
Corr (towards early career earning): -0.3170854

The histogram is left-skewed, suggesting that most colleges have the average debt above the mean average debt of the selected colleges.The average student debt is most frequently seen in the range of $26 000 to $29 000. The scatter plot introduces a vertical line at $25 000 and $27 000 average student debt, which suggests a low correlation between the two variables. The correlation states the dataset suggests the less the average student debt a college has, the more earning the students have in their early career. The low absolute value of correlation shows the aforementioned relationship is not easily seen.

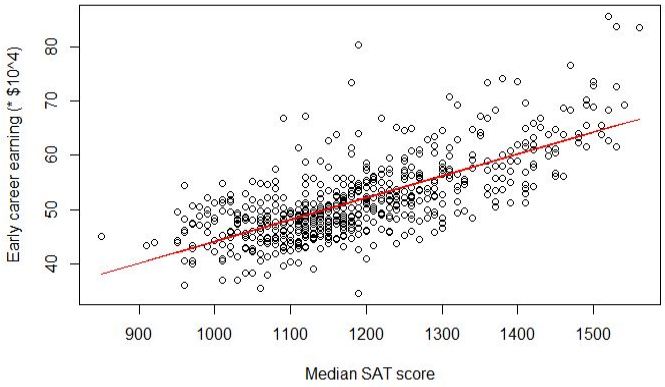


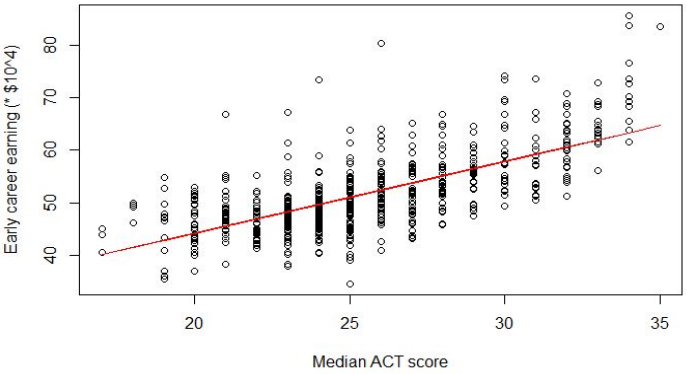
Summary of the “Early Career Earnings”:

Min	1st Qu.	Median	Mean	3rd Qu.	Max.
34.60	46.58	50.20	51.67	55.10	85.60

The shape of the histogram of early-career earnings also resembles the two above histograms. The mean earning ($51,670) also surpasses the median ($50,200). Overall, this histogram spreads out a lot less than the two above histograms, but it still remains right-skewed. This suggests that most people graduating from college will tend to have a specific amount of earnings in their early careers. Beyond this, there are cases where people may earn just less than twice the common amount of earning.

Linear regression between Median ACT/SAT to early career earning:





SAT:
Intercept: 3.85; Slope: 0.04; P-value: 2*10^(-16)
ACT:
Intercept: 16.87; Slope: 1.366; P-value: 2*10^(-16)

Since one main goal of the project is to measure the weight (the importance) of each characteristic listed in the dataset to determine the most valuable characteristic(s) in maximizing early career earnings, we decided to create a linear regression model of median SAT/ACT to predict the early career earnings to investigate its coefficient. If the model fits well, we can infer that colleges with higher SAT/ACT requirements correspond to the early career earnings in some ways. We chose median SAT/ACT because their histogram shapes resemble the shape of early career earnings, and each has strong correlation value to early career earnings. Each predicted value given the linear regression model reports to have very low p-value (less than 2*10^(-16)), so we are confident that our regression model fits the data very well.

The slope of each regression suggests that:

- Each 10 point in median SAT requirement corresponds to 0.04 * 10 * $1000 = $400 increment in early career earning.
- Each 1 point in median ACT requirement corresponds to 1.366 * $1000 = $1366 increment in early career earning.

- T-test:

1. T-test for Medians of ACT of private and public universities.
   H0: mu_median_ACT_public >= mu_median_ACT_private
   H1: mu_median_ACT_public < mu_median_ACT_private
   > p-value = 3.686e-06

   The p-value from the t-test between public college Median ACT and private college Median ACT is too small (p = 3.686e-06 < 5%), which indicates that we will reject the null hypothesis. In other words, public universities take lower median ACT than private universities.
2. T-test for Medians of SAT of private and public universities.
   H0: mu_median_SAT_public >= mu_median_SAT_private
   H1: mu_median_SAT_public < mu_median_SAT_private
   > p-value = 0.0002718

   The p-value from the t-test between public college Median.SAT and private college Median SAT is too small (p = 0.0002718 < 5%), which indicates that we will reject the null hypothesis. In other words, public universities take a lower median SAT than private universities.
3. T-test for Early career earning of private and public universities.
   H0: mu_career_earning of public >= mu_career_earning of private
   H1: mu_career_earning of public < mu_career_earning of private
   > p-value = 0.1494

   The p-value for early career earning between private and public college is significantly larger than 5% (it is just below 15%). This indicates that we do not have the evidence to reject the null hypothesis. The 2013 College Scorecard, which attempts to calculate the value of degrees from private and public colleges, states that public colleges lead to better returns than private colleges in the 10-year horizon. This supports the null hypothesis.

General conclusion: We conclude that public universities take lower median SAT and private universities take higher median SAT. In spite of that fact, our hypothesis test on the mean of early career earning from private and public universities concludes that public colleges lead to better returns than private colleges in the early years.

References

College Information - Peterson's - The Real Guide to Colleges and Universities. (2020). Retrieved 29 June 2020, from https://www.petersons.com/

Earnings | DASL . (2020). Retrieved 29 June 2020, from https://dasl.datadescription.com/datafile/earnings/?fbclid=IwAR2gj4eE0P8SGabCoVYpOx5PC9AJRZkjY3XLCHtv_xFd-KKaAFWiR_sP754

Money's 2019-20 Best Colleges Ranking. (2020). Retrieved 29 June 2020, from https://money.com/best-colleges/

Mulhere, K. (2019, August 12). How MONEY Ranked the 2019 Best Colleges. Retrieved June 29, 2020, from https://money.com/how-money-ranks-best-colleges-2019/

PayScale - Salary Comparison, Salary Survey, Search Wages. (2020). Retrieved 29 June 2020, from https://www.payscale.com/

Postrel, V. (2001, February 25). Dropping the SATs Is an Excuse to Drop Standards. Retrieved June 29, 2020, from https://www.latimes.com/archives/la-xpm-2001-feb-25-op-29979-story.html