Stat159 Project 3: Add a title

Liang Hao, Bret Hart, Andrew Shibata, Gary Nguyen

December 3, 2016

1 Abstract

This project aims to design and build a website for an NGO in order to provide minority studetns a portfolio of colleges which would best serve their needs: a high-quality, low-cost education.

2 Introduction

For this project, we are assuming the role of an NGO which desires to provide at-risk, low-income, minority students a portfolio of colleges which would best serve their needs: a high-quality, low-cost education. This will ultimately done using the shiny tool: students will enter in various criteria (their GPA, location, desired field of study, etc) and it will output a list of colleges ranked according to a score which we arbitrarily, but intelligently create as an ad-hoc indicator of quality of school for these types of students. This college quality score will be our response variable, and will be compiled in multiple ways: one, by searching through the columns of data and doing basic research on what makes an education good, such as the average salary of each institutions graduates after college, graduation rate for at-risk students, or median debt of graduates; in addition, we can choose salient response variables by using various methods to cluster the data and examine what variables seem indicative or dependent upon a good education. We will then combine these responses into one college quality score for the school: if there turn out to be five really profound responses, each response would be 0.2 of the final quality score. In addition, we will be using the last 5 years of college data for our model, but each earlier year will be given less weight, in a polynomial manner, to still let the earlier data inform our model, but to not give it equivalent importance to the more recent years. To choose our predictor set, which will partially be entered in by students and partially inferred to predict how a student may do in respect to the score, we can use many different feature selection algorithms, such as lasso regression or pruning random forests or support vector machines for predictor selection. The interesting specificity of our model is that it is geared not toward creating a score for general school quality and a prediction for the average student, but

specifically for helping at-risk students who need assistance in deciding which schools are worth the application in their specific circumstance.

3 Data

3.1 Description

The data we used for this project come from the *College Scorecard* , which is developed by the U.S. Department of Education (under Obama's Administration) to provide "key indicators about the cost and value of institutions across the country to help students choose a school that is well-suited to meet their needs, priced affordably, and is consistent with their educational and career goals". The link to the data can be found [here](https://collegescorecard.ed.gov/data/). For now, the (website)[https://collegescorecard.ed.gov/] provides basic information about facts about educational institutions upon search.

3.2 Cleaning

For our purpose, we collect the raw data for the past five years, and clean up in the following way:

1: For dataset of each of the five years, we get the columns and rows with less than 10

2: Then we run through the data dictionary to extract useful columns that may help in building model and the shiny App for this project. The process decreases the number of columns to under 300.

3: We parse the five dataset with the selected columns and rows.

4: Believing that closer years have more value as information, we combine the five dataset with the following weight:

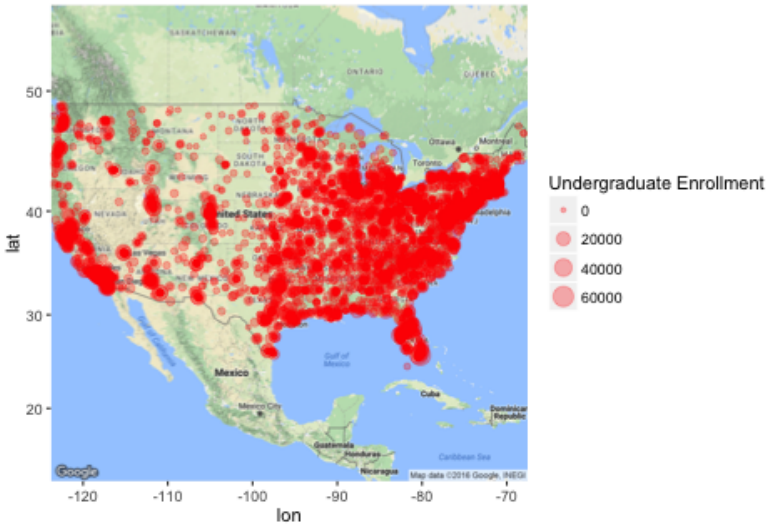
	Year 2014-2015	Year 2013-2014	Year 2012-2013	Year 2011-2012	Year 2010-2011
Weight	0.40	0.30	0.10	0.05	0.05

5: Lastly, we merge the combined dataset with the *Post-Graduation Salary* dataset to form the data frame to be used in this project.

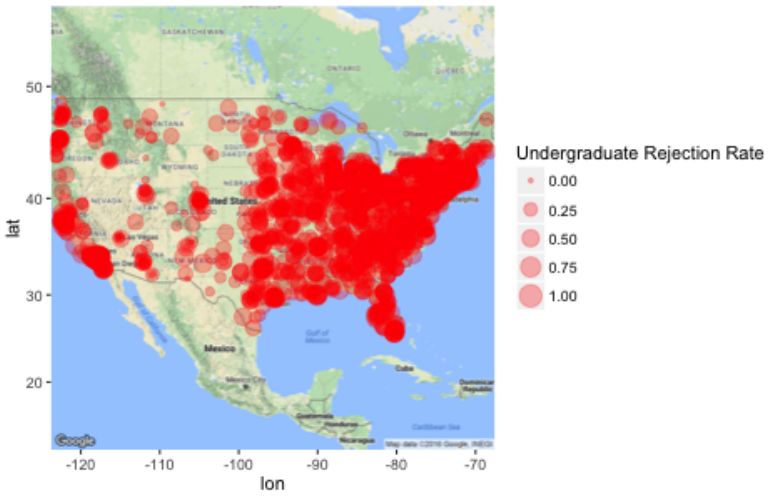
3.3 Exploratory Data Analysis

3.3.1 Geographical Distribution

We may first have a look at the geographical distribution of colleges in the dataset. For the purpose of graphing, the plot only shows colleges in mainland United States.



We may see that blablablablabla
Then, we may look at the



3.3.2 Ethnicity

We can also look at the historical statistics for educational conditions for minority groups. For such causes have long been drawing attention from the society and the public still complain about it.

	HBCU	Frequency	RelativeFrequency
1	No	6613	0.98
2	Yes	102	0.02

Table 1: Flag for Historically Black College and University

	PBI	Frequency	RelativeFrequency
1	No	6622	0.99
2	Yes	93	0.01

Table 2: Flag for predominantly Black Institution

From the tables above, we can see that the minority groups are underrepresented in most of the educational institutions in the U.S.. We may also

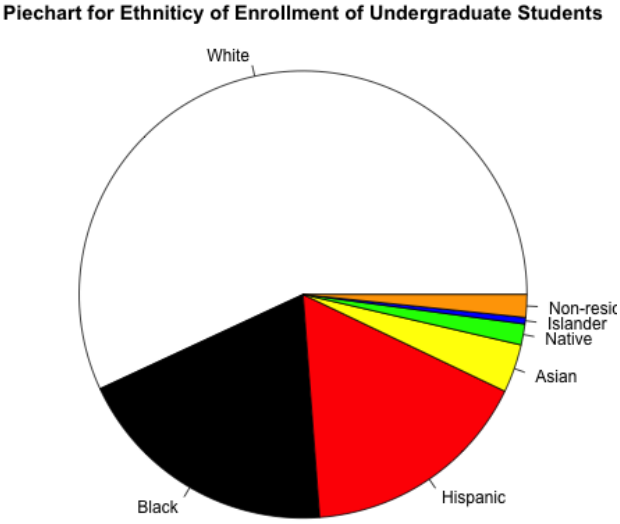
	ANNHI	Frequency	RelativeFrequency
1	No	6680	0.99
2	Yes	35	0.01

Table 3: Flag for Alaska Native, Native Hawaiian Serving Institution

	HSI	Frequency	RelativeFrequency
1	No	6368	0.95
2	Yes	347	0.05

Table 4: Flag for Hispanic-Serving Institution

look at the tthnitiy of total share of enrollment of undergraduate degree-seeking students in the dataset.



From the pie chart above, we may see that .....

3.3.3 Net Price

Net price is the .....

From the table above, we can see that .....

3.3.4 Standardized Test Scores

The Standardized Test Scores could be important indications for the quality of school and how likely a student may get admission. So we may look at the

	TRIBAL	Frequency	RelativeFrequency
1	No	6684	1.00
2	Yes	31	0.00

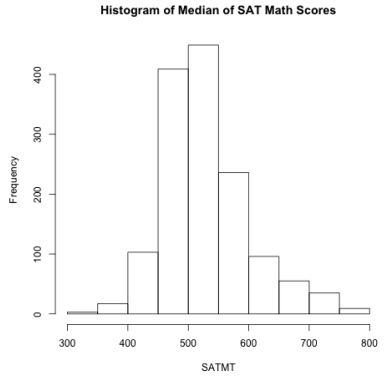
Table 5: Flag for Tribal College and University

	AANAPII	Frequency	RelativeFrequency
1	No	6586	0.98
2	Yes	129	0.02

Table 6: Flag for Asian American, Native American, Pacific Islander-serving Institution

general distribution of SAT and ACT scores.

From the histogram and the table above, we can see that the median of SAT Critical Reading scores basically follows a normal distribution, with most of the scores clustering between 450 and 550. We may consider colleges with such score higher than 556 to be outstanding as a college institution.



From the histogram and the table above, we can see that the median of SAT Mathematics scores basically follows a normal distribution, with most of

	Min.	Max.	Range	Median	25th	75th	IQR	Mean	SD
Public	126.32	27055.80	26929.48	8604.50	6217.02	12397.31	6180.28	9444.23	4466.77
Private	639.00	74245.42	73606.42	18337.28	13233.46	22543.12	9309.66	18143.31	7037.34

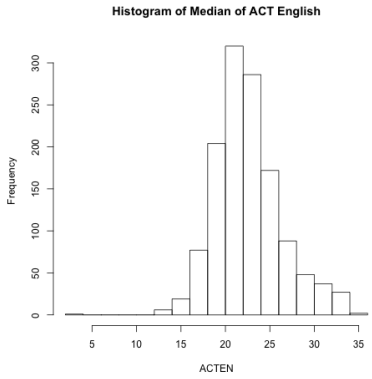
Table 7: Summary Statistics for Net Price of Public and Private Institutions

	Min.	Max.	Range	Median	25th	75th	IQR	Mean	SD
1	298.00	757.25	459.25	511.50	475.50	556.25	80.75	520.90	66.81

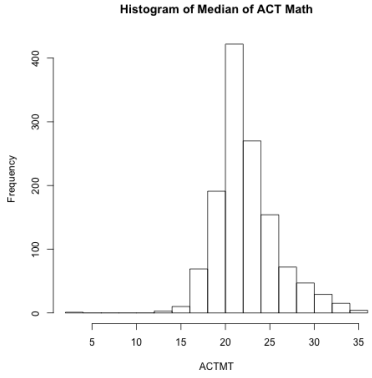
Table 8: Summary Statistics for SAT Critical Writing

the scores clustering between 450 and 550. We may consider colleges with such score higher than 563 to be outstanding as a college institution.

Because not all colleges recorded or submitted their SAT Writing scores to the original dataset, we may remove in the consideration. Also, we may conduct similar analysis on reported ACT scores:



From the histogram and the table above, we can see that the median of ACT English scores basically follows a normal distribution, with most of the scores clustering between 20 and 24. We may consider colleges with such score higher than 24.6 to be outstanding as a college institution.



	Min.	Max.	Range	Median	25th	75th	IQR	Mean	SD
1	305.00	783.75	478.75	516.00	482.31	563.17	80.86	528.66	70.38

Table 9: Summary Statistics for SAT Math

	Min.	Max.	Range	Median	25th	75th	IQR	Mean	SD
1	2.00	34.40	32.40	22.17	20.16	24.60	4.44	22.69	3.78

Table 10: Summary Statistics for ACT English

From the histogram and the table above, we can see that the median of ACT Math scores basically follows a normal distribution, with most of the scores clustering between 20 and 24. We may consider colleges with such score higher than 24 to be outstanding as a college institution.

4 Method

In this section, we will talk about how we decide on the criteria for high-quality, low-cost education for minority students.

One of the most common reasons students cite in choosing to go to college is the expansion of employment opportunities. To that end, data on the earnings and employment prospects of former students can provide key information. To measure the labor market outcomes of individual attending institutions of higher education, data on cohorts of federally aided students were linked with earnings data from de-identified tax records and reported back at the aggregate, institutional level.

	Min.	Max.	Range	Median	25th	75th	IQR	Mean	SD
1	2.00	35.40	33.40	21.95	20.40	24.00	3.60	22.47	3.41

Table 11: Summary Statistics for ACT Math