Stat159 Project 3: College Recommendation

Liang Hao, Bret Hart, Andrew Shibata, Gary Nguyen

December 5, 2016

1 Data

1.1 Description

The data used in the construction of this app originates from the *College Scorecard* database, 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 dataset can be accessed here: (https://collegescorecard.ed.gov/data/).

1.2 Cleaning

For our purpose, we interpret and intercalate the data from the past five years, cleaning and merging the sets by implementing the following algorithm:

1: For each dataset, we extract the columns and rows with less than 10% *PrivacySuppressed* or *NaN* to obtain usable datasets with each column posessing enough information to be meaningful. We then take the unions and overlaps of the columns and rows across the years. This leaves us with around 600 columns and 6700 colleges.

2: We examine the data dictionary and relevant literature, noting useful columns that may be relevant to our response variable. This arbitrarion process decreases the number of columns to under 300. We then parse the five datasets with the logically collected columns and rows.

4: While we include five years of data to ease out short-term variances, we still believe more recent data has value in this circumstance, so we combine the five datasets with the following weights:

	Year 14-15	Year 13-14	Year 12-13	Year 11-12	Year 10-11
Weight	0.40	0.30	0.10	0.05	0.05

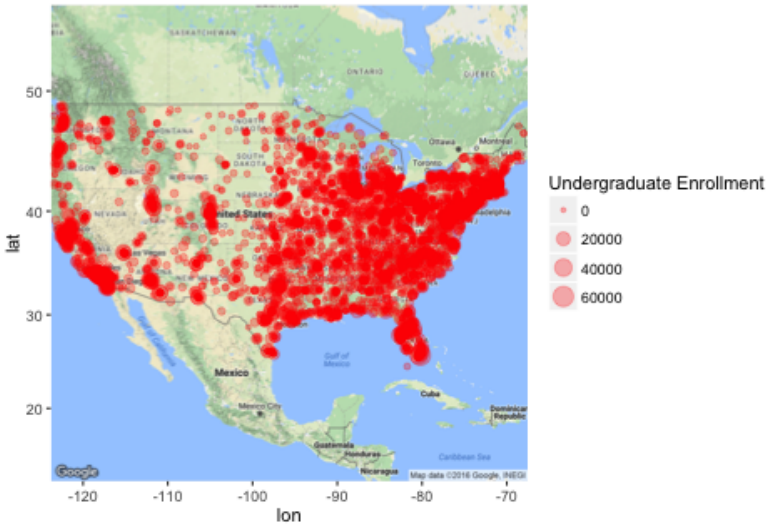
5: Lastly, we merge the combined dataset with the *Post-Graduation Salary* dataset, which contains various variables not included in the original database (despite its insane size) to form the data frame used in this project.

1.3 Exploratory Data Analysis

1.3.1 Geographical Distribution

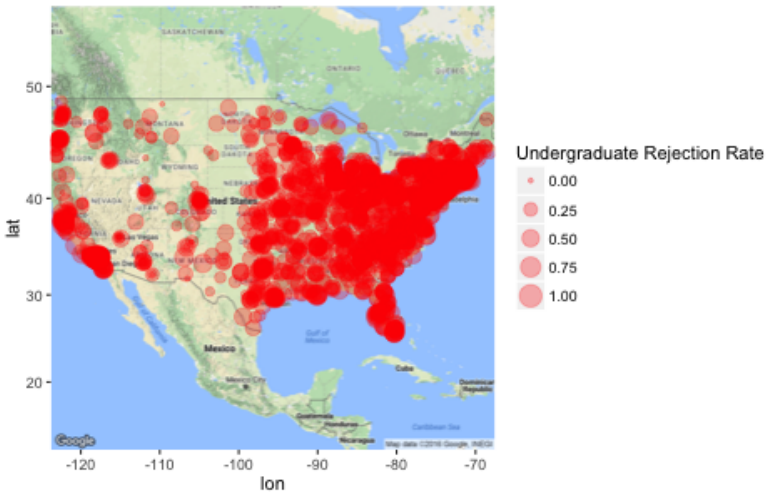
To gain a better sense of what we want students to input or what we should include as an important piece in the response metric, We carry out substantial exploratory data analysis. We begin by looking at colleges around the country as a whole.

First, let's look at the geographical distribution of colleges in the dataset. (For the purpose of efficient and realistic graphing, the plot unfortunately only shows colleges present in the mainland United States.



We see that there are a good number of schools of substantial size in most states, with the mid-western, mountain hemisphere perhaps a little low on number of educational institutions.

Then, we may look at the "difficulty" of schools around the country, that is, the rejection rates, from 0% to 100%.



It's hard to tell as the points are smaller for lower acceptance rates, but there honestly doesn't seem to be a deficiency anywhere in schools that are, perhaps, easier to get in to, but which will undoubtedly provide a good education - this is exactly what we wanted to know. Is it even worth it to work on this recommendation platform at all? Yes, because students may have schools in their own backyards that they just need to be told about to explore and succeed.

1.3.2 Ethnicity

It is no secret that those of minority racial/ethnic status in the United States may not be given the same institutional affordances in the college process, from admissions to graduation to post-graduation achievement. Even beyond negative and arbitrary admissions processes, it has been theorized that the existence of a community which is composed of students of your background, etc. can have a profound effect on success. Thus, we can examine the various ways race has been historicized in the college process, as it is and always will be of great public concern. Additionally, due to our NGO affiliation, we want to approach these institutional inequalities with a constructive open mind and a tool which suggests schools of better, or at least matching, diversity.

From these tables, we gathered that minority groups are underrepresented in the vast majority of institutions, and thus, these students may feel out of

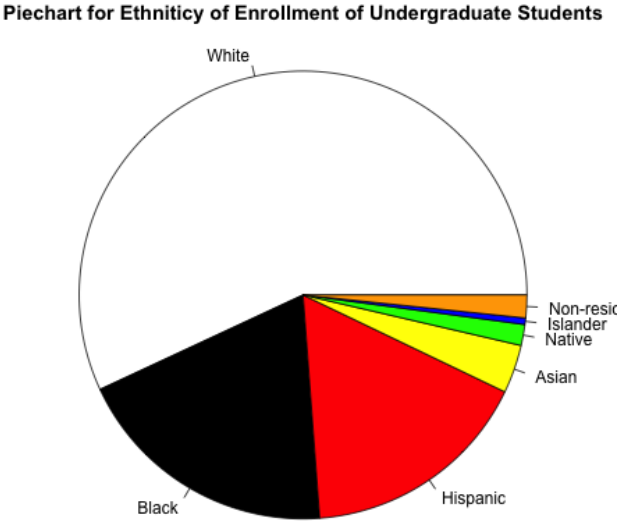
	HBCU	Frequency	RelativeFrequency
1	No	6613	0.98
2	Yes	102	0.02

Table 1: Percentages of Historically Black Colleges and Universities

	PBI	Frequency	RelativeFrequency
1	No	6622	0.99
2	Yes	93	0.01

Table 2: Percentages of predominantly Black Institutions

place in an environment that is mostly made up of people different than them. As a contrast to these damning tables, we offer a simple pie chart which shows racial/ethnic demographics across the schools.



From the pie chart above, we may see that, while it is of course a problematic and contentious postulation, that white students make up the largest majority of the student body - thus, our metric of predominance in an institution may be unrealistic or not the greatest actual marker of diversity.

1.3.3 Net Price

Net price is the final cost of attending an institution, which includes the amount required for tuition and fees, books and supplies, and living expenses after

	ANNHI	Frequency	RelativeFrequency
1	No	6680	0.99
2	Yes	35	0.01

Table 3: Percentages of predominantly Pacific Islander Institutions

	HSI	Frequency	RelativeFrequency
1	No	6368	0.95
2	Yes	347	0.05

Table 4: Percentages of predominantly Hispanic Institutions

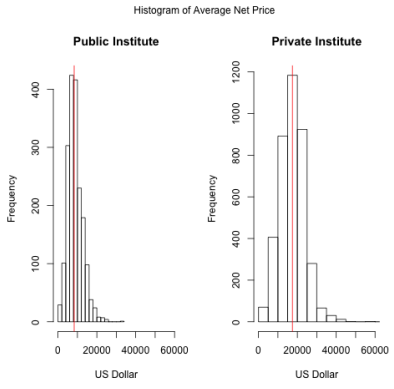
financial aid which would be on-default given to certain income levels. Thus, the Net Price is different for each income bracket - and our model reflects this, by pulling from different columns for net price in response to the student's input. For this project, we are obviously interested in the Average Net Price for full-time, first-time undergraduate Title IV-receiving students.

Average net price ($NPT4*$ for PUB [public colleges; for public institutions, this metric is limited to those undergraduates who pay in-state tuition] and $PRIV$ [private colleges]) contains a weighted average of the net cost of attendance for all full-time, first-time undergraduate Title IV-receiving students.

Here, sumammary statistics are provided for both Private and Public Institutions.

From the table above, we (logically) see that the Net Price of Private Colleges is much higher than that of Public Colleges. As you can see, every summary statistic for Private Colleges is larger than the equivalent value for Public Colleges by a factor of 3 or 4.

The following are the histograms for Net Price of both Public and Private colleges by income bracket.



By comparing the 2 histograms, we can see that the simplified distribution of net price for public institutions is skewed-left, while net price of private colleges tends to follow a more normal-appearing distribution.

The data reflects these differences by having stored many different net prices

	TRIBAL	Frequency	RelativeFrequency
1	No	6684	1.00
2	Yes	31	0.00

Table 5: Percentages of predominantly Native American Institutions

	AANAPII	Frequency	RelativeFrequency
1	No	6586	0.98
2	Yes	129	0.02

Table 6: Percentage of predominantly Asian American, Native American, Pacific Islander-serving Institutions

- average net price is categorized into different income quintiles for students.

	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