concordance=TRUE

Stat159 Project 3: Add a title

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 (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 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, 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.

```
## Error:  'include_graphics' is not an exported object from 'namespace:knitr'
```

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 educational institutions.

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

```
## Error:  'include_graphics' is not an exported object from 'namespace:knitr'
```

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.

	HBCU	Frequency	RelativeFrequency
1	No	6613	0.98
2	Yes	102	0.02

Table 1: 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

	ANNHI	Frequency	RelativeFrequency
1	No	6680	0.99
2	Yes	35	0.01

Table 3: Flag for Alaska Native, Native Hawaiian Serving 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 look at the tthnitcy of total share of enrollment of undergraduate degree-seeking students in the dataset.

```
## Error:  'include_graphics' is not an exported object from 'namespace:knitr'
```

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

1.3.3 Net Price

Net price is the cost of attending an institution(public or private), which includes tuition and fees, books and supplies, and living expenses).

For this project, we are interested in the Average Net Price, which is derived from the full cost of attendance (including tuition and fees, books and supplies, and living expenses) minus federal, state, and institutional aid, for full-time, first-time undergraduate Title IV-receiving students.

Average net price (NPT4$_*$$forPUB[publiccolleges; forpublicinstitutions, thismetricislimitedtothoseunderg$
$statetuition]andPRIV[privatecolleges])includesaweightedaverageofallfull-time, first-$
$timeundergraduateTitleIV-receiving.$

Here, sumammary statistics is provided for both Private and Public Institutions

From the table above, we can see that the Net Price of Private Colleges are much higher than those of Public Colleges. All summary statistics of Private Colleges are much higher numbers and they are higher than Public Colleges by a factor of 3 or 4.

The following are the histograms for Net Price of both Public and Private colleges

```
## Error:  'include_graphics' is not an exported object from 'namespace:knitr'
## Error:  'include_graphics' is not an exported object from 'namespace:knitr'
```

From 2 histograms, we can see that the net price for public institution is skewed-left, while net price of private colleges tends to follow a normal distribution.

	HSI	Frequency	RelativeFrequency
1	No	6368	0.95
2	Yes	347	0.05

Table 4: Flag for Hispanic-Serving Institution

	TRIBAL	Frequency	RelativeFrequency
1	No	6684	1.00
2	Yes	31	0.00

Table 5: Flag for Tribal College and University

In addition, average net price is then categorized into different income quintiles for students.

(1)0-30,000;

(2)30,001-48,000;

(3)48,001-75,000;

(4)75,001-110,000

(5)110,000+.

FOr each quintile, average net price is divided into 'Public' or 'Private'.

1.3.4 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 general distribution of SAT and ACT scores.

```
## Error:  'include_graphics' is not an exported object from 'namespace:knitr'
```

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.

```
## Error:  'include_graphics' is not an exported object from 'namespace:knitr'
```

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 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:

	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

	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

```
## Error:  'include_graphics' is not an exported object from 'namespace:knitr'
```

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.

```
## Error:  'include_graphics' is not an exported object from 'namespace:knitr'
```

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.

	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

	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

	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