College Scorecard Analysis of the Institutions of America

KYLE HARTLAND BROWN, VICTORIA VALESQUEZ, and RAMI ALQUNAIBIT

The goal of our project is to analyze a series of documents created by the Department of Education called "College Scorecards". These consist of various categories concerning the institutions of the United States, including but not limited to: Graduating Debt, Graduating Income, Transfer status during schooling, and acceptance rate(among many others). Our hope is to analyze this data and discover trends concerning the average debt of different institutions and other categories listed. We hope to find answers to questions various questions, foremost among these being "Do institutions with graduating debt tend to have higher graduating income?" and "Do institutions with higher acceptance rates tend to have higher levels of debt on graduation?". We also hope to discover many more interesting correlations in the data sets as we explore the hundreds of columns of information collected about every institution.

Previous Works

The previous work done on this data set is extensive, and should prove incredibly useful for our project. These are listed as Kernels on Kaggle.com. The most interesting of the kernels at first glance have been: "Admission"(https://www.kaggle.com/stuffypuppy/admission), "For Whom the Pell Tolls" (https://www.kaggle.com/wrudebusch/for-whom-the-pell-tolls), "Are Affordable Schools a Good Deal?" (https://www.kaggle.com/michaelpawlus/rmarkdown-default-script), "Escape From Poverty" (https://www.kaggle.com/wrudebusch/escape-from-poverty), and "College Earnings Premium & Value Proposition"(https://www.kaggle.com/apollostar/college-earnings-premium-value-proposition).

"Admission" is a paper concerning the admission rates of schools in regards to the overall SAT scores of that school. They also analyze the different institutions with higher earning rates, whether STEM majors make admission more competetive at each institution, and the general price of each institution compared to it's admission rate.

"For Whom the Pell Tolls" gives an excellent analysis of where students with pell grants are most likely to attend, here finding that online for profit colleges such as DeVry University get many of the students, as well as that the pell grant students still graduate with high debt.

"Are Affordable Schools a Good Deal?" Was able to show a strong correlation between a high cost of attendance and a high repay rate of those students loans, as well as graphs showing a general lower repay rate based with attendance cost below $40,000.

"Escape Fron Poverty" Sorts all colleges by their median earnings six years after graduation and compares that to whether or not they are first generation college students to determine if college was a positive decision for those graduates, their findings were that online for-profit schools keep the best track of their students, are high in first generation student percentage, and generally lead to a rather low sallary, with none exceeding $70,000 a year on average.

"College Earnings Premium & Value Proposition" was undoubtably the most interesting concerning ourt project, as they went through and compared SAT scores, income, and various other attributes to create an Earnings Premium chart in which they predict the expected earnings of a group based on their degree and university.

Proposed Work

We propose to sort through the data and determine how much can be learned about the average debt of each institution over the decades of study, and check to see if graduating income increases with level of debt. This will require cleaning the data of the many null values that appear, sorting through schools that have not included this information, and attempting to compensate for any lack of information by using updated college scorecards provided by the department of education. A big challenge will be to overcome the amount of "Privacy Supressed" data that has been ommited by institutions that do not wish to have certain statistics made public. This is where we can set ourselves appart from previous projects, as no previous projects used the most up to date scorecards from the department of educations website, we will have access to four years more data, as well as an opportunity to gain insight into what institutions supress what information and correlate that with other information they may have provided. The privacy supressed data seems to become more available in more recent years and could have some very interesting stories to tell about what institutions were charging high rates, but not giving students as large of an advantage upon graduation.

Data Set https://collegescorecard.ed.gov/data/

Although the dataset is available on Kaggle, we have chosen to set ourselves apart from the other kernels on this website by acquiring the data directly from the department of education, this has a number of benefits, but mostly it has ensured that we are in posession of the most recent data possible.
The categories that the data has available are extremely numerous, but are organized into: School, Academics, Admissions, Student, Cost, Aid, Repayment, Completion, and Earnings. All of which have dozens of sub-categories, for example Student can be expanded into Number of Undergrad Students, Race of Undergrads, Undergrad part-time percentage, Age, Income Brackets, First-Generation percentage, and FAFSA submissions. This results in close to a hundred individual categories that we can use over the hundreds of thousands of school entries to find as much meaningful data that can reasonably be acquired by the collegiate institutions of the United States. We also fully intend to start exploring other data sets as we answer our questions with the scorecards. For example, it may be interesting to search for data concering the average income of residents in cities surrounding colleges to see if there is any affect on the loan rates and tuition costs due to poverty or wealth of certain areas.

Evaluation Methods

To evaluate our data the larges challenge will be actually sorting through it to find the percentages that could be correlated. Otherwise the entire data set is composed of percentages which we assume to be taken out of the total population of students for each university. So at that point we can reference how other papers evaluated things like the predicted income or debt, but most of our work in evaluating the data will be in checking the Confidence and Support of the relationships we hope to draw. At that point we will determine a minimum support that would make the data relavent and draw conclusions based on what these metrics tell us.

Tools

Our projected needs for tools do not currently exceed the functions that numpy, pandas, and other python libraries are able to provide for us concerning the calculations we need to make and how we want to sort through as well as manage our data.

Milestones

3/9 - have completely explored the data and determined all questions we would like to answer and whether null values or ommited data will be a problem
3/18 - be close to or finished with cleaning all data of null values, and have decided how to handle privacy ommited data.
3/24 - determine if we wish to use other data to draw more information about the colleges.
4/1 - be close to if not done with preliminary findings, start creating graphs and graphics, incorperate new methods learned in class if desired.
4/10 - take care of any unforseen changes and create project final report when ready.

Summary of peer review

The peer review was encouraging, we feel as though we have a grasp on what we want to do and how to do it, however, after listening to some groups talk about large numbers of null values we did become concerned about the fact that we cannot immediately determine the true number of nulls and Privacy Omitions, only that they are incredibly numerous, our first goal is to see exactly how extensive the omissions are, and whether we should adjust the questions we would like to answer about our set according to the data actually available in the data set. We will do this by simple omission of schools with null values in our relavent categories such as debt and income, and find the percentage of schools that provided the relevant information.