

The College Scorecard

Varun Nadgir

September 27, 2017

Capstone Project

Introduction

The **College Scorecard** is a service meant to help prospective students make their college decision. Whether by comparing size, popular majors, or comparing costs to the national average, the site's goal is to help the user find a good fit. For my Springboard Capstone Project, I used the **dataset** made available by the College Scorecard to try and find additional ways to help users in their decision.

The Data

Available in a .zip file from the link above, the data is split into 19 .csv files - one for each academic year from 1996-'97 to 2014-'15. Each file contains 1,744 recorded data points (columns) and about 7,500 schools (rows). My first step was to add a DATAYEAR column to indicate the academic year and then merging the 19 files into one large .csv, which I called **fulldata**. From this 2 GB file, I would create subsets for plotting and studying trends. Next, I had to refer to **the data dictionary** to understand the column names and some of the placeholder values used. Once I merged the files and had a basic understanding of what data was available, I began to create some plots and documented my initial findings in my **data story**.

Deliverables

My first item will be using linear regression models to determine what variables are the most influential on the cost of a school, and to potentially predict what the cost of a school may be in the future. This would be useful in two ways. It could help students who are on the fence about going to college immediately after high school by suggesting a decline in cost. If a student sees that their ideal school is likely to be cheaper in two years, they may make the decision to find entry-level work or go traveling before going to college. It could also help the schools by indicating unknown areas of their budget that are influencing the cost of attendance. Of course, the goals of each school are different and they may not be interested in reducing cost, but if a school is experiencing a decline in applications, cost of attendance is likely to be something they look at.

The second item is a recommendation tool that works similarly to clustering methods used by media services such as Netflix, YouTube, and iTunes. Clustering based on things like location, cost, and SAT scores, a student can find options that are close to their top choice. Mentality is a very important part of finding success at school, and feeling out of place in freshman year can be quite discouraging. If their top choice is a far reach school, or it is too expensive/too far, then finding alternatives would hopefully help them to be satisfied in their decision.

My final item will be a basic UI that allows the user to explore the dataset on their own. Although the data will need to be curated and shaved down to a size that a standard internet browser can handle, my hope is that it will provide some transparency between students and universities. As an example, in my **data story**, I explored SAT averages and admission rates. Even though one could reasonably guess how they are related (higher SAT scores ~ lower admission rates), being able to plot the data and draw a conclusion from a graph is much more convincing. By putting this power in the hands of students and their families, they should be able to make much more educated decisions.