Milica Cvrkota

Shuyi Qi

Madeleine Vessely

**Abstract:**

Making a decision about where to go to college is a big deal as more and more people are deciding to attend college. In the interest of helping individuals decide which college to attend, we carried out a series of analyses on College Scorecard data, aiming to visualize and compare the characteristics of different colleges and universities in the United States. The analyses include visualizing colleges on maps, clustering, predicting for the best fit college and historical comparison of colleges. Our end product is an R Shiny App, which can be applied as a tool to compare colleges and find the best fit by users.

**The Analysis of the College Characteristics in the United States**

**Introduction**

Making a decision about where to go to college is a big deal as more and more people are deciding to attend college. In 2016, almost 70% of high school graduates enrolled in college directly after high school graduation. Clearly, this is a decision that affects many people in the United States. College is a huge financial and temporal investment and therefore it is important to make informed choices. Having access to data about colleges in the United States, and being able to easily visualize it, is essential. The goal of this project is to conduct a retrospective observational study in college data by visualizing and comparing the characteristics of different colleges and universities in the United States.

**Methods**

Our dataset comes from the government website College Scorecard. The dataset was provided through federal reporting from all undergraduate degree-granting institutions of higher education, and it contains variables describing federal financial aid and tax information, which was reported to Integrated Postsecondary Education Data System (IPEDS). Each year (1996 through 2017) is contained in a separate dataset, and each case in the datasets represents an institution, with all its relevant information in 1899 variables.

Many of the 1899 variables are far too specific to be of general use. Consequently, we narrowed down the variables to 20 which we considered to be of general interest. These include 7 categorical variables: type of school (public, private nonprofit, private for-profit), region (nine regions overall), locale (twelve types of locations, ranging from rural to large city), religious affiliation, institution name, men only, and women only) and 13 numeric variables (admissions rate, verbal SAT score, math SAT score, writing SAT score, english ACT score, math ACT score, writing ACT score, number of undergraduates, annual cost of attendance, average faculty salary, percent full-time faculty and percentage of first-generation students). Next, we looked at the years present and selected those in which all of these variables were present. In the end, we decided to use data from the 2009-2010 through 2016-2017 school year.

To process the data, we first selected the 20 variables of interest in all eight datasets, then added a year column in each of them and merged the data from all years in one dataset. Additionally, we removed any schools that were not present in the dataset in all eight years, along with any schools that were present multiple times. We also created a new binary variable First\_gen, which equals 1 when over 30% of students at the school are first-generation students and 0 otherwise. We decided on 30% as a threshold because it was median and mean across all schools from our dataset. Moreover, to deal with “NULL” strings, we either deleted the entries in which some columns contain “NULL” strings or turned the “NULL” strings into another factor level or integer value depending on the situation. Finally, we altered the types of some variables from factor to integer/double and modified the levels of some factor variables where necessary.

As our data is a combination of categorical and quantitative variables, we decided to analyze it using visualization in Shiny App that could support both types. There are four parts to our App: visualizing colleges on maps, clustering colleges, prediction of similar schools, and historical comparison of schools. In the first part of the App, we created interactive maps with data filtered by 6 of the 7 categorical variables (excluding religious affiliation), the first generation variable, and some of the quantitative variables (SAT/ACT scores, number of undergraduates and cost). The interactive mapping gives a comprehensive view of admissions rates of colleges across the states and the filtering options allowed us to examine the relationships between various variables.

The second part of the App is applying a clustering algorithm to the 2016-2017  data to look for characteristics that naturally group colleges together. We chose seven numeric variables (admissions rate, SAT verbal, SAT Maths, ACT English, ACT Maths, cost of attendance and number of degree-seeking undergraduate students) to use for modeling. We then performed PCA on the data and applied the PAM algorithm to group colleges into clusters. The advantage of using PAM is that data points are used as cluster centers. The centers are representative of their cluster and thus allow us to infer common characteristics in each cluster. We also created a 2-D visualization of clustering with the top two dimensions calculated from PCA as y and x-axis. In addition, the App provides the functionality to control for four categorical variables (type of school, region, locale and religious affiliation) in order to eliminate their confounding effects. Besides using PCA, we also created 2-D visualizations with all combinations of two numeric variables in the clustering dataset being the axes to examine their relationships.

The third part of our App is the prediction of best fit schools given a set of user inputs. The user has the option of entering desirable characteristics, such as school size, type of school, region, location, annual cost and standardized tests’ scores. There are additional options for predicting women-only, men-only, and schools with over 30% first-generation students, and based on the inputs the program produces 10 closest college matches. Our prediction algorithm is based off of Gower distance, which can be applied to clustering data with a mix of categorical and numeric data. Once inputs are provided, a new row containing the inputs is added to the end of the 2016-2017 dataset and we obtain the dissimilarity value between the new row and every other row from the distance matrix. We computed Gower distance between the newly created point and every other point on the plane with the top two dimensions produced by PCA as x and y-axis to find the 10 closest points. We were able to compute the percentage of fitness based on the distance of 10 points from our new point based on the values of Gower distance, which gives us values between 0 and 1. Our new point would have distance 0, which means that it fits 100% to user input, while the furthest point has distance 1, which would give 0% fitness. Using this approach, we were able to compute the percentage of fitness for the best college matches, where for example, the point with distance 0.1 would have 90% fitness. This approach implies that the percentage is also relevant to other data points and not just the user’s input since they are used for computing the percentage of fitness. Although we learned about Gower distance in class, in our project we used it to develop our own prediction algorithm, which is a novel application of the method.

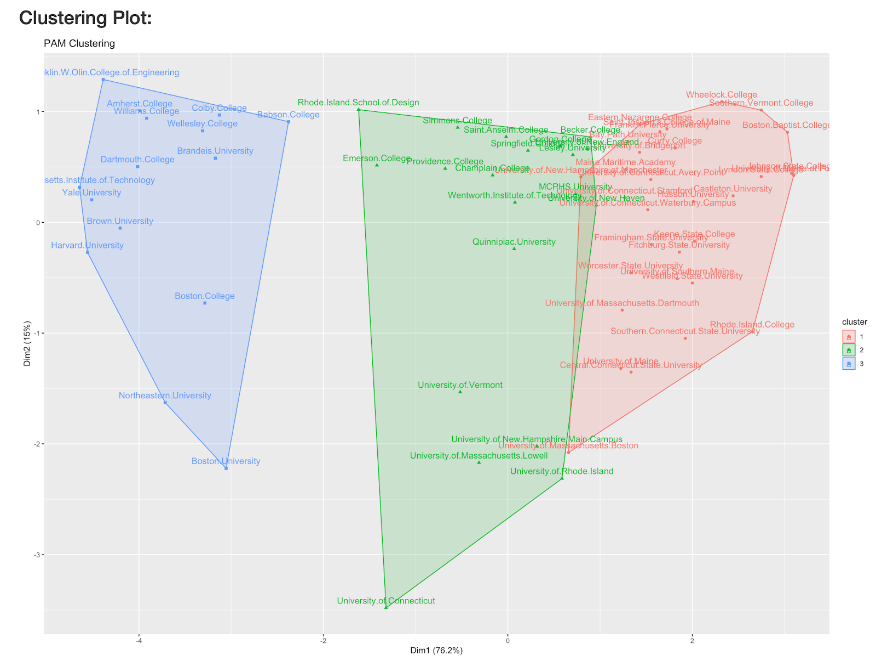
For the historical comparison part of the App, the user has the option of choosing which variable they want to see throughout the course of all years in the dataset. They are also able to select schools of interest by entering the name of the school. Multiple schools can be selected at once, enabling the user to compare the performance of multiple schools over time. There is also a slider range allowing the user to select which range of years they are interested in.

**Results**

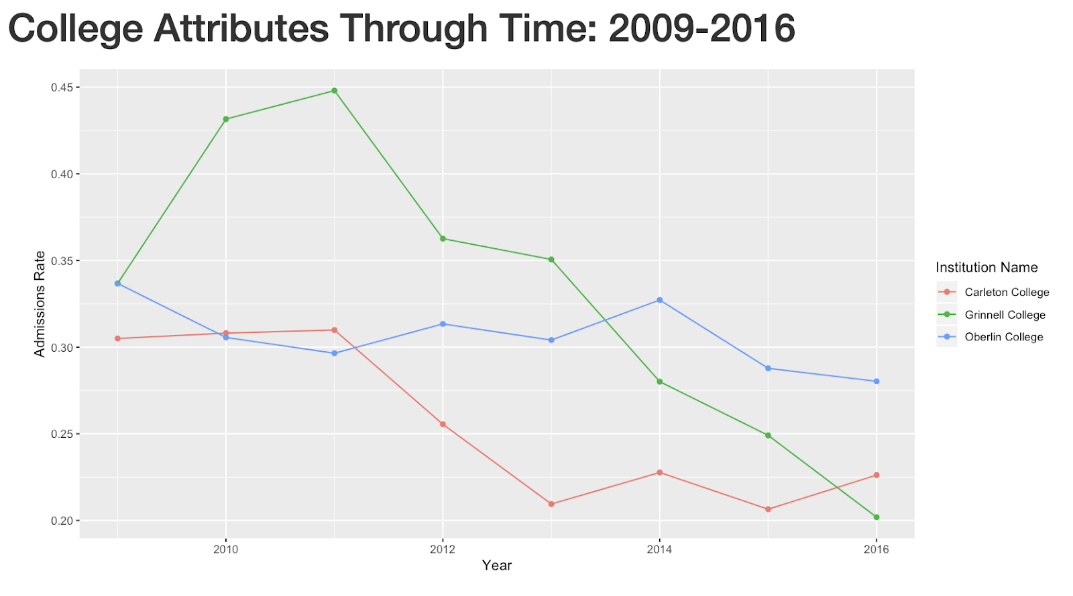
While we wanted to use the map for visualizing the whole dataset, we were still able to observe some trends while filtering. As we would expect, private schools tend to be smaller and more expensive comparing to public schools. Women and men-only schools are all private non-profit schools, located mainly on the East coast. Furthermore, there are slightly more private schools in the East, while public schools are evenly distributed.

Figure 1 shows a visualization of clustering based on the top two dimensions from PCA. The visualization shows that PAM generally groups colleges based on dimension 1, which describes their middle standardized scores, regardless of region, locale, religious affiliation and type of the school. Interestingly, PAM groups colleges that are commonly thought of as in the same tier and that have similar rankings in U.S. News.

The results from the historical comparison functionality of the App depend on the inputs. In Figure 2, an example of a possible comparison of admissions rates at three similar school is shown. The admissions rate at Grinnell College was high in 2010 and 2011, while it did not change much for Oberlin or Carleton during these years. Beginning in 2012, there has been a steady decrease in the admissions rate at both Grinnell and Carleton. Comparatively, the admissions rate at Oberlin remained relatively stable as compared to the other two schools.



*Figure 1. Clustering plot for public and private nonprofit colleges in New England*



*Figure 2. Line plot for historical comparison*

**Conclusion and Discussion**

Admissions rates of colleges distribute uniformly across fifty states and these rates have fluctuated for almost every college during the eight years. Colleges with higher standardized scores are those that are commonly considered to be in the first tier and that are ranked high by U.S. News, which, as we would expect, suggests that colleges that are nationally recognized as having a high education quality are more competitive and harder to get in.

One limitation of our data is that they do not cover every variable that can affect one’s college decision, such as majors. And our analysis is possibly not representative of all colleges in the United states because we eliminated many colleges with incomplete information for the variables we selected. We also made assumptions about the religious affiliation variable in the data that if a school’s religious affiliation is “not reported” or “not applicable”, then it does not have a religious affiliation. The PAM algorithm also introduces random errors because the clusters are sensitive to choice of initial cluster centers.

**References**

James, Gareth, et al. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2017.

**Appendix 1: PCA**

From the PCA output, we knew that the first dimension roughly describes the standardized test scores while the second dimension primarily describes the number of degree-seeking undergraduate students enrolled (UGDS).

