By: Thejas P. Vidyasagar, Conner Gillette, Stephen Hung, Kshitij Verma

Project Proposal -- College Statistics

Education is a cornerstone to the modern society. One of the steps involved in the process of education is college. Nowadays, a degree in a college is almost a necessity for many jobs. To succeed in future career, students need to pick the right college that matches their preferences to spend about 4 years of their life in. Hence, it is important to look at the nationally accumulated data on Colleges to ultimately help students pick their preferred colleges. We will be observing the data on such colleges and present it in a helpful way for the users to understand. The source that we will be using comes from the source Data.gov, the link to which has been provided below. The data is collected, managed and hosted by the U.S. General Services Administration, Technology Transformation Service. We aim to make use of specific subsets of the data sets such as “most recent cohorts scorecard,” and “college scorecard raw data.” These data sets will help us relate the different requirements of different schools throughout the nation and also provide us with information like top 75 percentile admit SAT score, acceptance rate, city of the respective university, amount of earnings by graduates etc. Along with data sets we will also be using a data dictionary, the link to that is also provided below.
Link to the data dictionary: https://collegescorecard.ed.gov/assets/CollegeScorecardDataDictionary
Link to the source: https://catalog.data.gov/dataset/college-scorecard

The major audience that we hope to aim are the graduating high school seniors who will be looking at different colleges to apply to. There dire need to compare the 75/25 percentile of each college to their own scores, tuition fee of each university and suitable city will be fulfilled by this data representation. For example, John Wick, an average american high school senior who has bagged a decent score in SAT, limited financial standing and now wants to apply in different colleges. Since he cannot apply to all of the colleges (due to high application fee), John has to finalize few colleges where he thinks he can get in and where his other requirements are met. He turns to the information provided by us and he can easily list down universities according to his needs. Another audience that we can target will be the high school counselors who need to keep a track of each university, their requirements, tuition fee etc., so that they can be helpful towards their students seeking guidance. For example, Josh Spencer, a counselor at a public school, would need to consider the academic, financial and vocational situations of each of the student to suggest appropriate schools. Our presentation can assist people like Josh to consider preferable options for their students based on our easily understandable data and their experience.

We hope that our audience will be able to learn more about schools that they are interested in applying to, or to discover schools that fit their needs. One of the major criteria have that people have when applying for college is where it is, and how far it is from their home. Showing filtered results plotted on a map is a great way to visually show where schools of interest are, and their proximity to home. Another question that our presentation can answer is the size of the school’s student body, because some people have a specific idea about how large their school should be to make them the most comfortable. One last question that our presentation should answer is how good the school is, and where it lies on a ranked list in different categories. Our dataset should allow us to show which schools are ranked highest out of the filtered results. In summary, our presentation

should answer the following questions: *Where are these schools located? How large or small is the school? How good is this school?*

Our presentation will mostly be comprised of a map and a ranked table that displays results based on a series of user-submitted criteria such as location, size, demographic type, most prevalent majors, etc. The map will display hoverable points to show where schools are, and general information about each one. The accompanying table should show the statistics and relations by the colleges from user-submitted prospective list, in terms of ranking or amount of financial aid provided, etc. This will allow students easier access to a finite selection of schools that would fit their needs based on their criteria.

The data set we are working with comes as multiple large CSV files thus there are minimal reshaping or reformatting needed to be done to import it to RStudio. However, when analyzing the data we will have to reshape the data set. This is because the federal government gathered a large amount of specific information about each college from the degree of urbanization to the percent of students who died within 2 years at the institution. In order to reshape the data set, we will filter the information we will use. We do not plan to use any special new R libraries besides the ones we used in class (dplyr, tidyr, ggplot2, etc.). Our presentation will also for the user to keep track of a working list of prospective schools. Using this list, we generate a summary statistics table as we compare average size, majority type, majority major type and average national ranking collectively against the national average. This would help the users better understand the collective standing for the schools in the list against the national average.

We believe that the hardest part of the project will be sifting through the enormous data set in order to find information that will be useful to our project and visualization. Other challenges we anticipate include figuring out how to use Shiny and other libraries in the way we want. Some other challenges we anticipate include the general programming issues such as debugging and version control. In order to get help on these issues, we will search in Google for our issues and talk to the TAs for tips and advice on how to proceed.