Capstone 1 Mini-Project: Data Wrangling

James Catterall

Overall cleaning process

First, I imported the data dictionary to find which year's dataset would be most relevant to my project's goal, which is to predict a student's future earnings and debt repayment rate based on their choice of college. I determined that the 2014-2015 dataset had the most recent information for future earnings and imported that dataset. I then renamed the variables in the dataset to use the more developer-friendly names provided by the data dictionary, with some modifications for brevity.

The bulk of my time was spent narrowing down the over 1500 variables in my dataset to a smaller number (under 100) that I would include in the cleaned dataset, so I wouldn't have to deal with null values for so many variables. If a variable or set of variables didn't seem relevant to my project's goal or had more non-null values than a similar variable, then I removed it from consideration. After eliminating all but 78 of the variables, I checked their correlation and then chose 17 quantitative variables that seemed like the best fit, in addition to several categorical (mostly binary) variables which did not require cleaning, save for one.

I subset the dataset so that the rows only include colleges with non-null values for one of the two target variables, and the columns were limited to the 17 quantitative variables. I then dealt with null and excessive/erroneous zero values using a number of methods, which I will describe in the next section. There were no serious outliers that I could find in the dataset. After having trimmed and cleaned the dataset, I exported that dataset to a new csv file for later use.

Dealing with null and 0 values

If less than 1% of a column's values were null, then I left it alone, as they would have a negligible effect on analysis later in the capstone. For a few variables, I replaced the null values with the mean of the column. For the variables relating to tuition, I grouped the cases by the ownership variable, and then filled any null values with the mean of the variable of the group it was in. For null and 0 values in the 150% completion rate and retention rate columns, I filled them with values from the Title IV 6-year completion rate and 2-year continued enrollment rate columns, respectively. After that step, less than 1% of cases had null values in those columns, and only a single 0 value was in the retention rate column. For the share of full-time, first-time students, I used the numpy "mask" function to replace any null and 0 variables with 1 minus the share of part time students, provided that the latter value wasn't 0 or null. Finally, I filled all null values in the categorical variable "open_admissions_policy" with a newly-created "0" category, which indicates that no data is available.