California Lutheran University

Project 1

ECON-562: Advanced Analytics

Andrew Frohner
3-24-2025

Exploratory Data Analysis: College Scorecard

Introduction:

The United States Department of Education collects information on universities around the country to better help students make decisions about their higher education. The areas in which they gather information ideally try to get at what a student cares about when it comes to their education. To name some of the primary items they cover Tuition costs, Student retention rates, Student body diversity, Faculty demographics, Student outcomes (employment, earnings, loan defaults), Student borrowing.

By gathering this data students have objective and measurable criteria to help them make a more informed decision on whether higher education is right for them. This data also helps increase accountability of the institutions providing education. The US dept of education aggregates this data from multiple sources including the Natl Student Loan Data System (NSLDS), US Department of Treasury, and the Integrated Postsecondary Education Data System (IPEDS).

Data used for analysis came from the most recent collection period dating 2022-2023 and it contained over 3200 variables across 6,484 educational institutions in the United States. This would be considered a large data set.

A dataset with dimensions such as these allows a researcher to potentially answer several questions about a college, however some of those questions are off limits. Not because they are necessarily a secret. This is more attributable to the nature of the collection process, organizations involved, and even timing of when research is conducted. For example, earnings and debt data included in this effort are only shared for students who received financial aid (Chingos & Whitehurst, 2015). This inherently biases and we must be conscious of that when we are asking questions and drawing conclusions about the relationships we discover. The limitation that is a bigger issue with regards to this particular analysis is timing of collection vs publication. In our data set, we have over forty variables related to student earnings, and across all 6448 institutions those observations are missing.

Immediately, the relationships we will find and the research questions we will ask must go in a “non-earnings” direction.

Data Description & Data Quality

This research was conducted using the latest college score card data file. (MERGED2022_23_PP.csv)

This data was last updated in January 2025 – but most there were many data items that were missing.
The focus of the research was initially focused on a select subset of 31 data elements (see full detail in the appendix).

Of those 31 data elements I narrowed the research further to only 4-year institutions under the Carnegie classification (CCSIZSET)

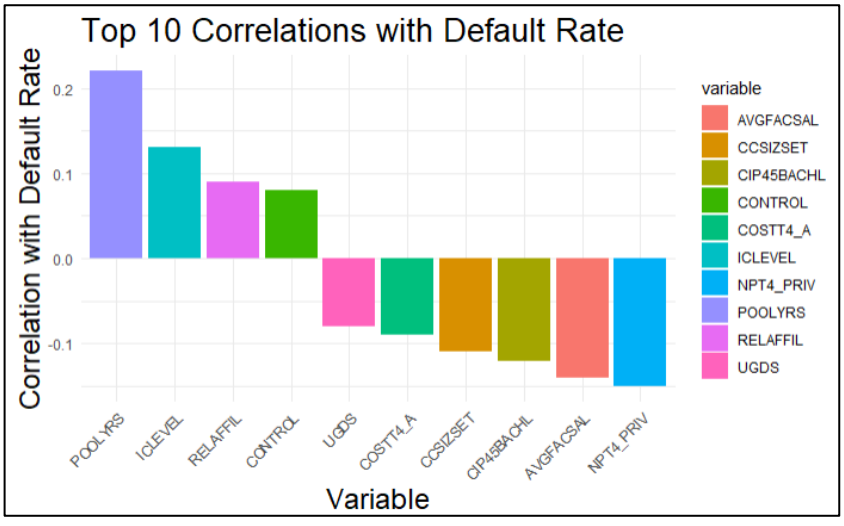
variable	*Percent missing*	*Count missing*
SCH_DEG	100	2585
FAMINC	100	2585
LNFAMINC	100	2585
MD_EARN_WNE_P10	100	2585
GRAD_DEBT_MDN	100	2585
UG25ABV	100	2585
RELAFFIL	67.81	1753

This left 2,585 observations to study, and among those schools, 7 variables had over 50% of their observations missing:

Research Question 1

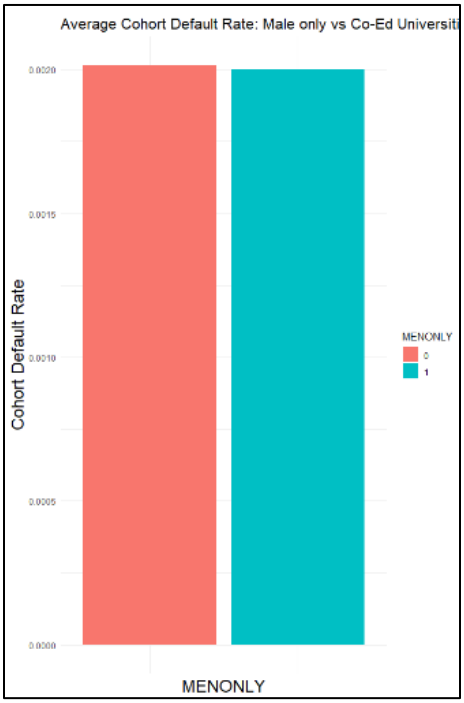
This provides us with the opportunity to ask: What factors impact the student default rate at large schools?

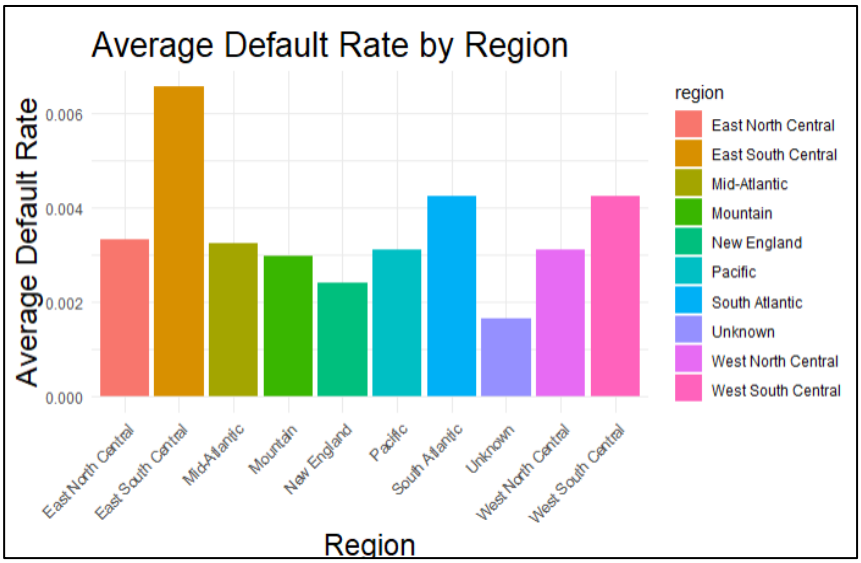
In the subset I selected the Default Rate of the 3-year cohort (CDR3) only had 7% missing values and a high number of outliers 3 times below the 25% quartile.



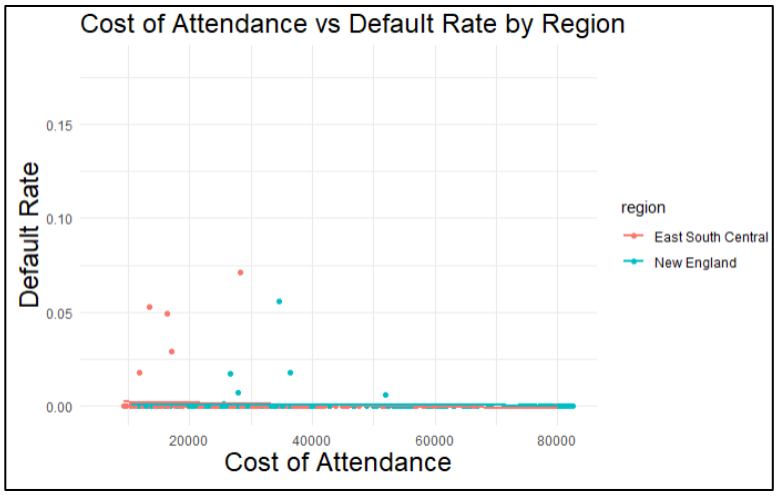
Interestingly, price variables (COSTT4_A, NPT$_PRIV) showed negative relationships with default rates. Carnegie Classification (CCSIZSET) is an ordinal variable by institution size was also negative. Suggestive that larger and higher cost institutions do not experience higher default rates among the 3-year cohort. On the other end – Private schools (CONTROL) were shown to have positive correlation.

Next, I tried to locate where higher default rate might be concentrated, and whether it is distinctly different between Co-Ed vs Male universities.





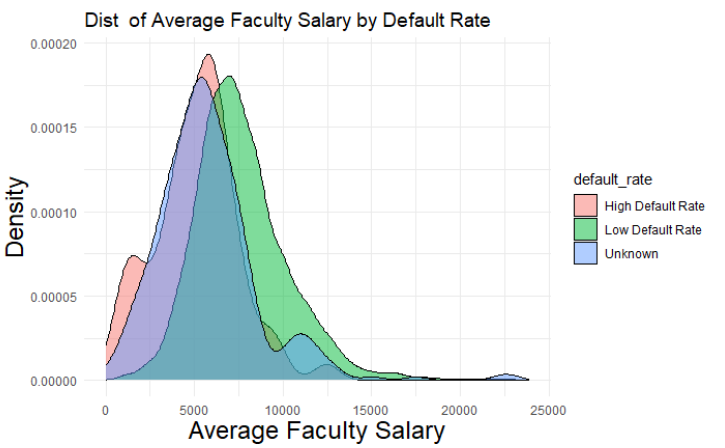
There is no difference in this default rate at men only universities, and aside from the East South-Central region of the US, default rates are stable.

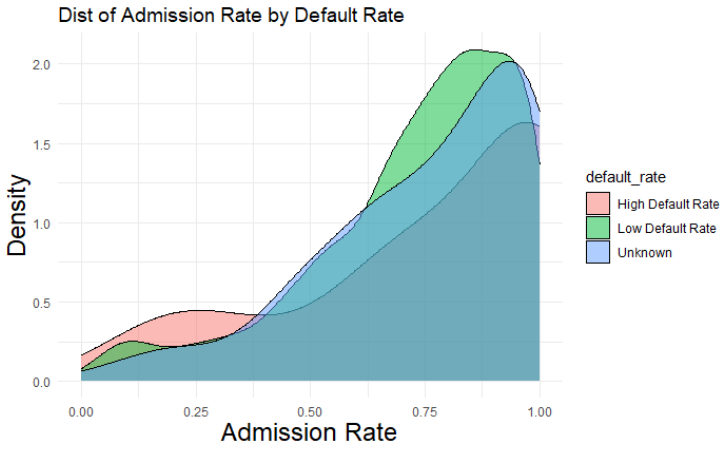


Comparing a high default and low default region, we find that default rates spike at the 4-year schools where yearly cost of attendance is below $40K a year.

This is a good insight to have, the data suggests that default rate is not necessarily linked to a region. Both regions show higher levels of default at a similar average cost of attendance.

Another insight that could be gained is looking at what schools with "low default rates" have in common. **Low default rate** was defined as any school with less than 5% of their 3-year cohort in default.





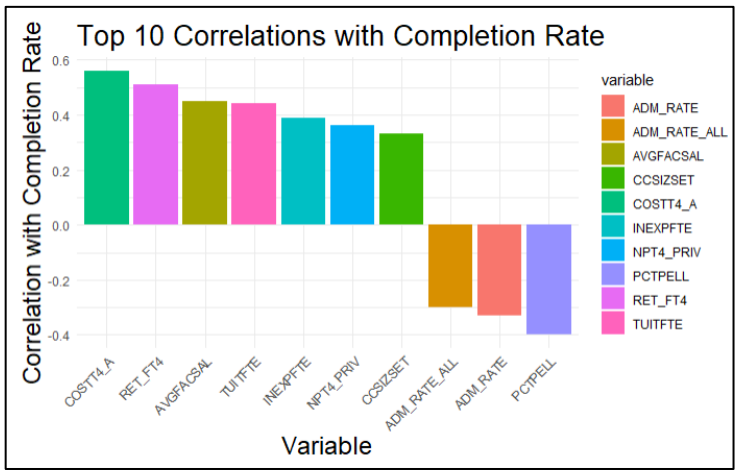
A Breakthrough insight does not jump out here—but the wider distributions for these variables among the high default rate schools could be useful for classification.

The EDA performed here is attempting to classify schools as ones where their student body has a high low default rate.

Research Question 2

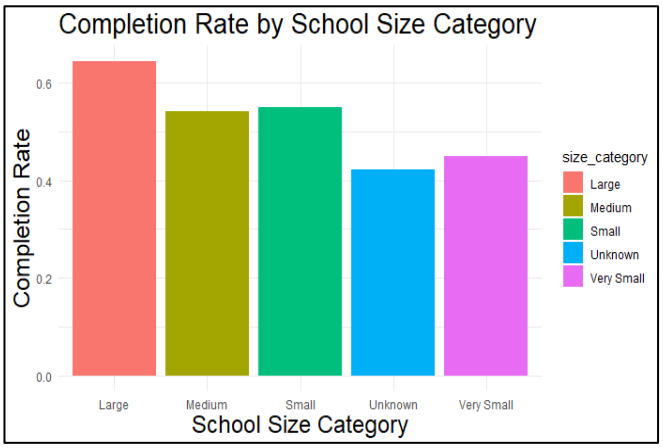
What can be used to categorize schools where a high percentage of students complete their degree vs schools with low percentages?

My sample is exclusively concerned with 4-year institutions. The variable C150_4 works nicely for this analysis. Note that the college scorecard uses 6 years as the expected completion time for a 4-year degree.



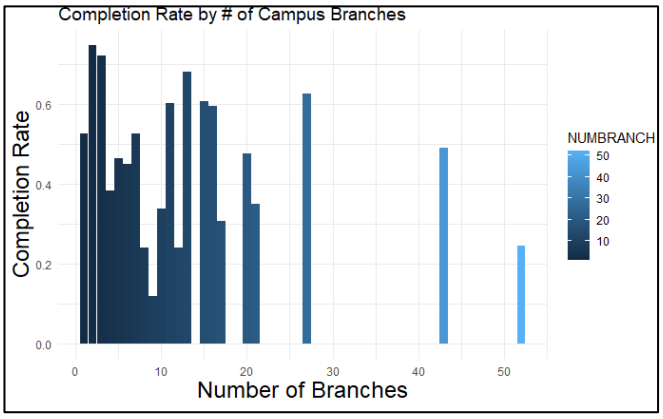
Completion rate exhibits stronger relationships with other predictors in my subset. Most of these variables have an intuition reason to be here. Variables like Average Cost to attend (COSTT4_A) and Retention Rate (RET_FT4) have strong positive correlations as they should. Most people see education as an investment in the future, an expensive school might provide a great return, and the student is motivated to complete their

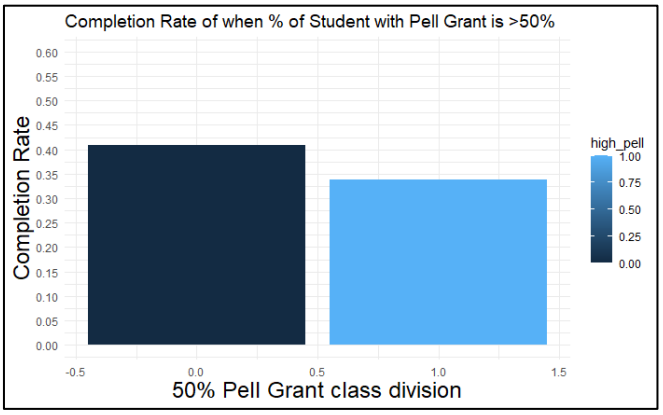
degree. A universities Retention Rate measures the percentage of students that stay enrolled – students cannot unenroll and complete their degree.



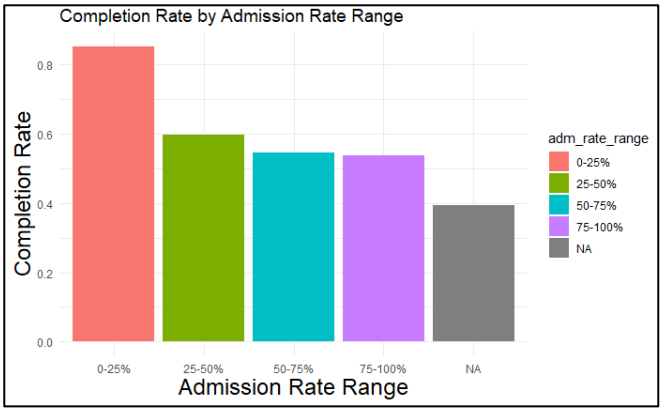
This visual shows the completion stepping lower as the school size gets lower. Interesting because categorial relationships such as this lend itself useful in a classification setting.

This categorization can be explored with other variables that exhibited a strong relationship with the completion rate.





There is a noticeable difference between completion rate and % of student body with Pell grant. Universities with lower Pell grant student proportions experience almost a 25% higher student completion rate.
Similar to the school size visual – admission rates have a step-down relationship at 4 consecutive levels.



Variables of school size, Admissions rate, Branch campuses, and Pell grant percentage have some ability to separate the completion rate variable into distinct levels (Higher or lower). This information is what we hope to find in an exploratory data exercise because it makes us more equipped to model.

We could take this insight and characterize universities as a High/Low completion rate institution.

Appendix

R Markdown file containing EDA Project 1

562_PROJ1_AF.html