Group Project I - The College Scorecard Dataset

As part of an effort to counter rising tuition costs and conflicting incentives in higher education, the U.S. Department of Education has been collecting information for public and private colleges and universities, and assembling this into public 'scorecards'. The data link measures of affordability and ease of access to student outcomes post graduation, and can be summarized to highlight schools providing the best value in education. The current dataset and reports for September 2020 is available below:

Website: https://collegescorecard.ed.gov/
Dataset: https://collegescorecard.ed.gov/data/
Documentation: https://collegescorecard.ed.gov/data/documentation/ (Also see Canvas entries.)

Data are available in yearly reports from 1996 through 2020, and each includes roughly 7,000 observations (colleges) for 2000 variables. For privacy reasons, some data are masked as "PrivacySuppressed", and some values are collected as two-year moving averages to reduce variability. For encoded variables, values are translated in the CollegeScorecardDataDictionary .csv file..

Figure 3-1: Distribution of Median Total Debt for Graduates Across Institutions

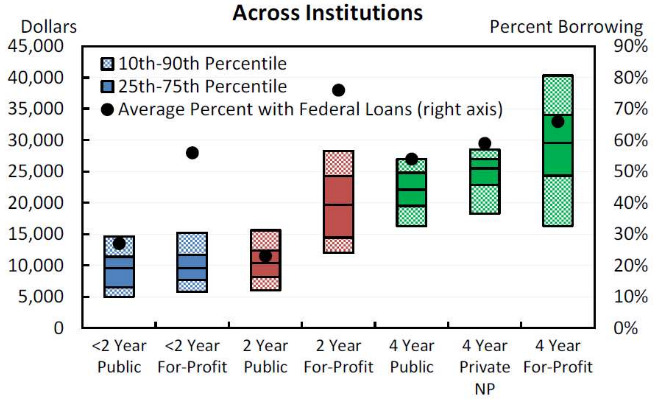
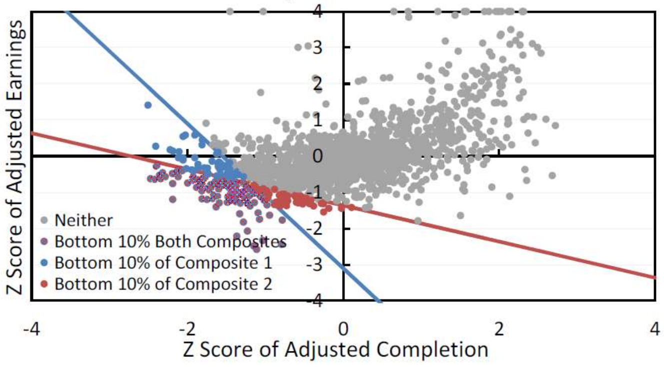


Figure 5-13: The Importance of Weights in Constructing Composite Indices



From Sept. 2015 College Scorecard report

Goals:

1. **Read the data into a language-native format**. This may involve:
   a. Either aggregating yearly files into a total file in R, or
   b. Using an interface between R and a native database format, or
   c. Accessing the API described in Documentation above, and at https://github.com/RTICWDT/college-scorecard
   d. Properly treating NAs and other encoded variables
2. **Assess the quality of the data**. This may include:
   a. Measuring fraction of missing values, or relationships among missing values and other columns
   b. Calculating numeric quality scores for several metrics
3. **Recode any variables of interest.** For example:
   a. Create combined categories from several columns
   b. Aggregate summary statistics from factor levels in several columns
   c. Normalize data between categories, where appropriate
4. **Identify and display any relationships of interest.**
   a. Ideally, these will use some of the recoded variables above
   b. Outcomes might cover access, affordability and student outcome categories