Activity: Beginning a Statistical Analysis Plan

The College Scorecard, compiled by the US Department of Education, is meant to be a tool for college applicants to compare different institutions. In this project, you will work with data from the scorecard to advise a high school senior on choosing a college.

The dataset provided for this project contains the following columns:

- `INSTNM`: school name
- `STABBR`: state in which the school is located
- `CONTROL`: whether the school is public or private
- `SATVRMID`: midpoint of SAT critical reading scores of students attending the school
- `SATMTMID`: midpoint of SAT math scores
- `SATWRMID`: midpoint of SAT writing scores
- `ACTCMMID`: midpoint of the ACT cumulative scores
- `ACTENMID`: midpoint of ACT English scores
- `ACTMTMID`: midpoint of ACT math scores
- `ACTWRMID`: midpoint of ACT writing scores
- `UGDS`: number of undergraduate students at the school
- `NPT4`: average cost to attend the school
- `PCTFLOAN`: fraction of undergraduates receiving a federal student loan
- `MD_EARN_WNE_P10`: median salary of students 10 years after graduation
- `GRAD_DEBT_MDN_SUPP`: median debt of graduated students
- `RPY_3YR_RT_SUPP`: proportion of students who are actively repaying their loans 3 years after graduation
- `RPY_3YR_70`: whether the proportion of students actively repaying their loans 3 years after graduation is $> 70\%$ (1 = yes, 0 = no)

Research questions

For this project, suppose you have been approached by a high school senior applying to college. They are looking through the college scorecard data, and they have some questions for you about the value of different schools.

- Do students who graduate from more expensive schools earn more money after graduation, after accounting for other variables?
- Is the relationship between school cost and graduate earnings the same for public and private schools, after accounting for other variables?

Questions

1. What variables should we consider to investigate these research questions?
2. What EDA would you do, using the variables from question 1?
3. Which statistical models will you fit to address the research questions? Provide the variables (explanatory and response) that will be included in the models.
4. How will you assess the suitability of these models? (Describe any diagnostics, metrics, etc. that you will use to check whether your chosen models are a good fit to the data).
5. How will you use the models to address the research question(s)? (hypothesis tests, confidence intervals, etc.)
6. What alternative strategies should be considered, and when? (i.e., what will you do if assumptions are violated? How would you modify the model or statistical analysis?)