# Project 2*

Megan Joseph

November 30, 2025

## 1 Introduction

the motivation for the analysis

Can college readiness, staff demographics, and school demographics predict graduation rates in California high schools?

a brief overview of the broader context

the knowledge gap that is addressed by the paper

what was done in the analysis

the key findings and why they are important

the structure of the paper

This project draws from the data sources Common Core of Data, Small Area Income and Poverty Estimates, The Civil Rights Data Collection, ED*Facts*, Integrated Postsecondary Education Data System, College Scorecard, National Historical Geographic Information System, Federal Student Aid, National Association of College and University Business Officers, National Center for Charitable Statistics, Model Estimates of Poverty in Schools (MEPS), Equity in Athletics Data (EADA), and Campus Safety and Security (2025a).

---

*Project repository available at: https://github.com/meganajoseph/261a_project2.

1

## 2 Data

The data are from the Education Data Explorer from the Urban Institute's Center on Education Data and Policy. This tool compiles data from national data sources on schools, districts, and colleges in order to assist in creating insights to improve student outcomes (2025a).

I used this query to produce a dataset of high schools in California from 2016-2017 with information on school characteristics, absenteeism, college readiness, course offerings, student demographics, discipline, school safety, student outcomes, and teachers and staff (Education Data Portal (Version 0.23.0), Urban Institute 2023). The observational unit is an individual high school in California. The key variables of interest are student enrollment, whether or not a school is eligible for participation in either Target Assistance program or school-wide program authorized by Title I of Public Law 103-382, the number of students enrolled in the International Baccalaureate Diploma Program, the number of students enrolled in the gifted and talented programs, the number of students enrolled in at least one AP course, the number of students who took one or more AP exams, the number of students participating in the SAT and ACT tests, the number of suspensions, the number of full-time teachers, the number of first-year teachers, the number of full-time equivalent instructional aides or paraprofessionals, the number of full-time equivalent school counselors, and teacher salary.

Removing all NA values would result in an empty dataset, so I removed a few of them and left the for variables like number of students enrolled in the IB program because that would most likely mean it is not offered at that school while another program like AP is offered. After removing these, the dataset has 2125 observations while the original has 6111 observations.

### 2.1 Common Core of Data

The variables student enrollment and Title I eligibility are sourced from Common Core of Data (CCD) (Education Data Portal (Version 0.23.0), Urban Institute 2023). The CCD is the United States Department of Education's database on all public elementary and secondary schools and school districts (2025b).

Student enrollment reports the number of students enrolled in the school. Figure 1 (a) shows a logarithmic relationship, so a log transformation should be applied. A school is Title I eligible if it is in a school district for children from low-income families (2025c). Figure 1 (b) shows that schools that are not Title I eligible have a higher median graduation rate and is less varied, whereas schools that are Title I eligible are much more varied with a larger number of outliers. I picked these variables because small schools have been linked to higher graduation rates, and students from low-income families tend to have higher dropout rates ("High School Graduation," n.d.).
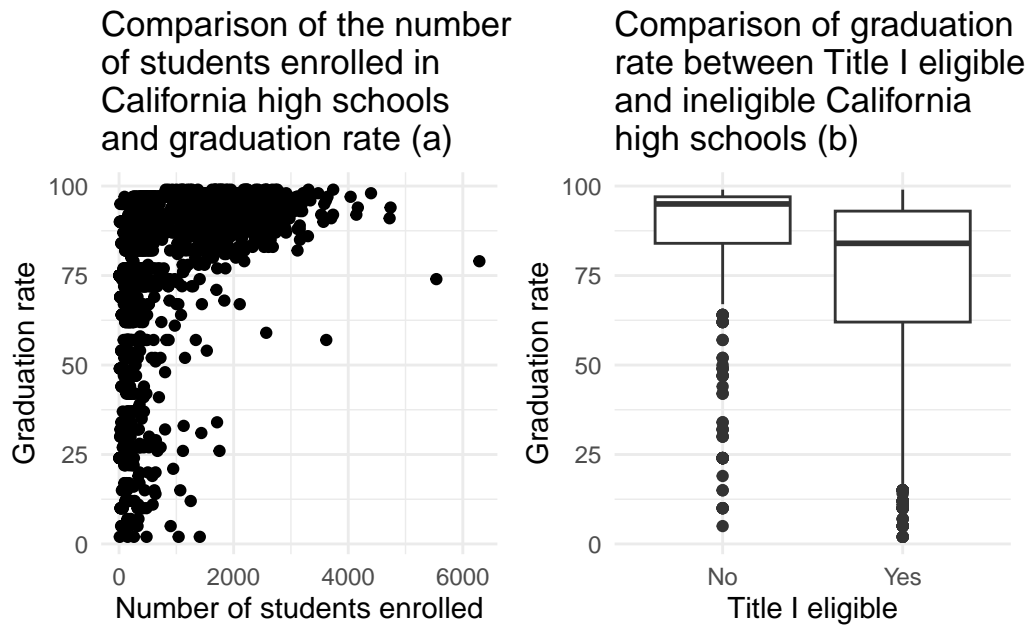
Figure 1: Plot (a) shows a scatter plot of the graduation rate over the number of students enrolled. Plot (b) shows a boxplot of graduation rate at Title I eligible and ineligible schools.
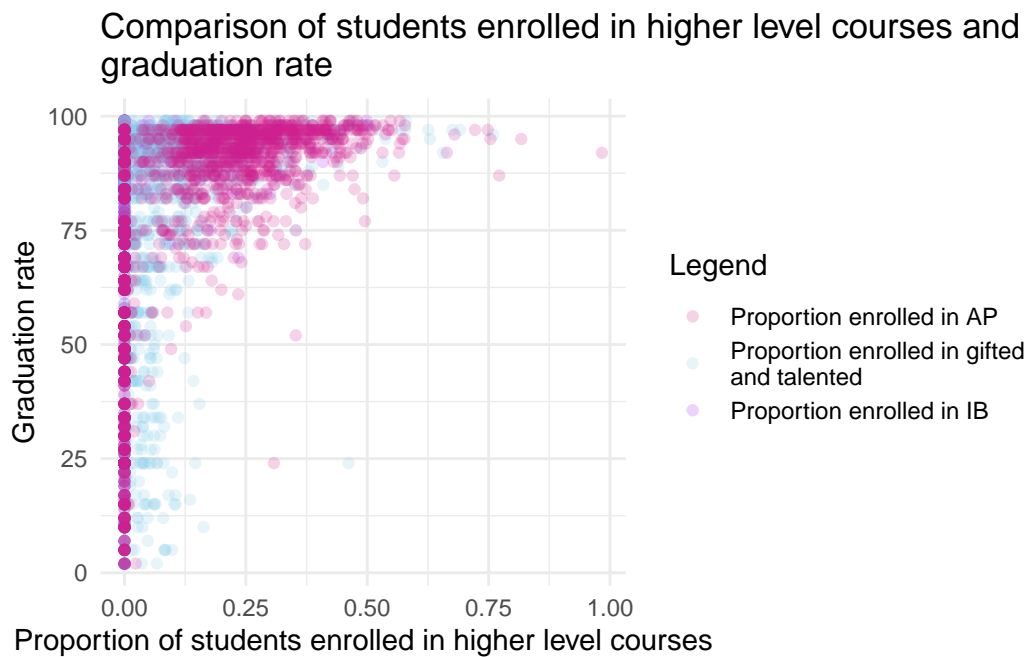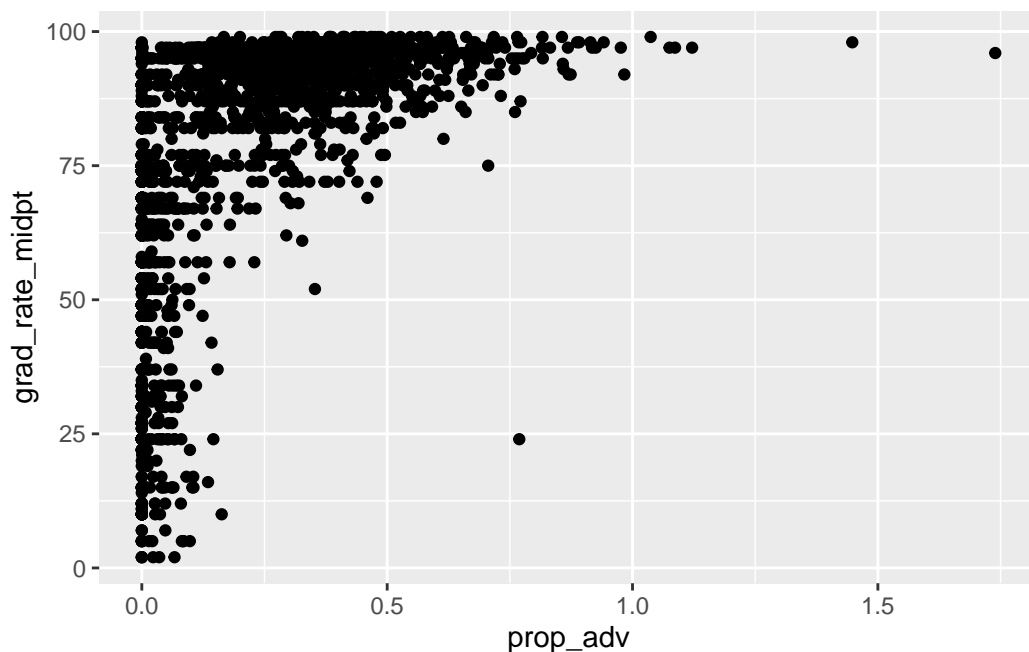
Figure 2: A scatter plot showing the relationship between the proportion of students enrolled in high level courses including the IB program (purple), AP program (pink), and gifted and talented programs (blue) (x-axis) and graduation rate (y-axis).

## 2.2 The Civil Rights Data Collection



The Civil Rights Data Collection (CRDC) comes from the United States Department of Education's Office for Civil Rights. This data is collection in order to ensure schools give students equal access to educational opportunities (2025e). Data are collected from public schools and districts, justice facilities, charter schools, alternative schools, and special education schools (2025e).

The variables from this dataset are the number of students enrolled in the International Baccalaureate Diploma Program, the number of students enrolled in the gifted and talented programs, the number of students enrolled in at least one AP course, the number of students that took one or more AP exams, the number of students participating in the SAT or ACT, the number of suspensions, the number of full-time teachers, number of first-year teachers, number of full-time equivalent instructional aides or paraprofessionals, number of full-time equivalent school counselors, and teacher salary. The number of students enrolled in the International Baccalaureate Diploma Program, the number of students enrolled in the gifted and talented programs, and the number of students enrolled in at least one AP course were transformed to the proportion of students enrolled in each respective program because that is more informative. Figure 2 shows a logarithmic relationship between these programs and graduation rate. Figure 3, Figure 4, Figure 5, and **?@fig-crdc-5** also show logarithmic relationships between the selected variables and graduation rate, so a log transformation is appropriate.

I picked these variables because students participating in higher level courses and/or participating in college-preparatory exams would be most interested in graduating and going to
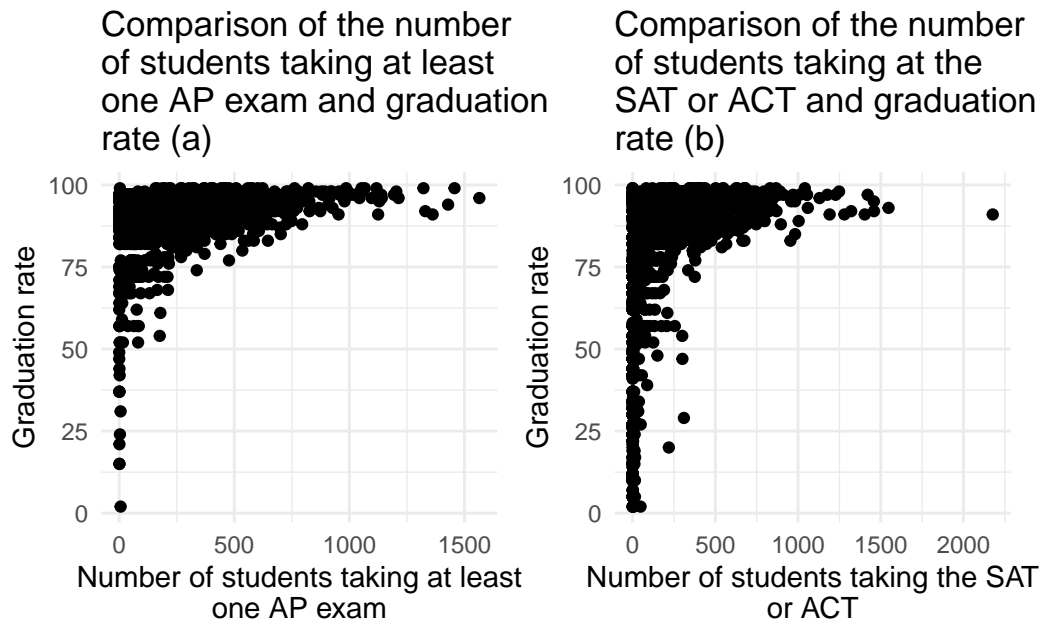
Figure 3: Plot (a) shows a scatter plot of the graduation rate over the number of students taking at least one AP exam. Plot (b) shows a scatter plot of the graduation rate over the number of students taking either the SAT or ACT.
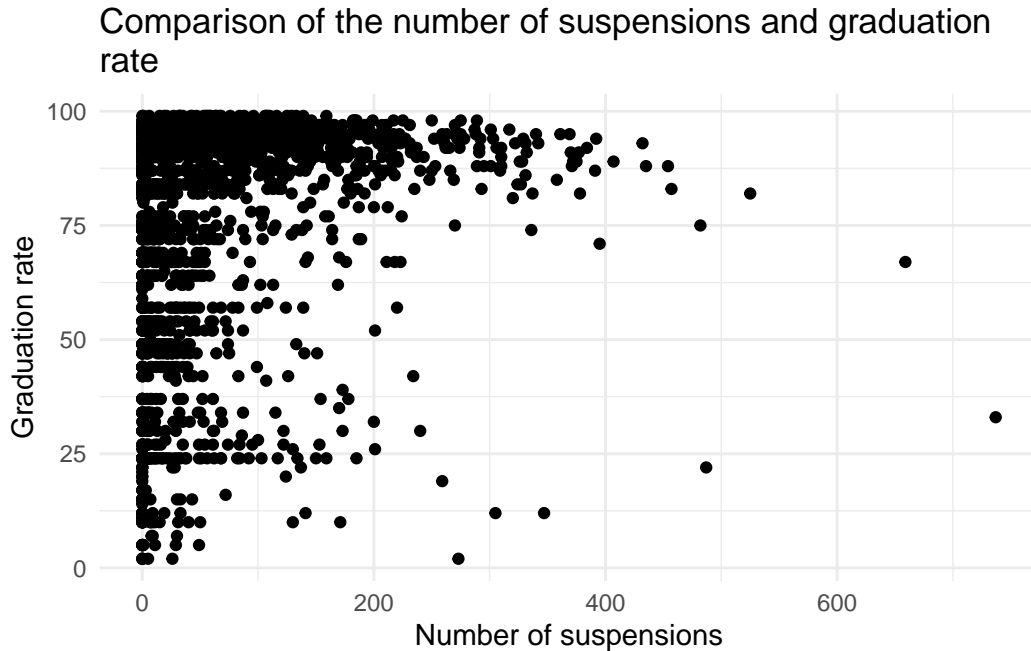
Figure 4: This shows the relationship between the number of suspensions in a school (x-axis) and graduation rate (y-axis).

college. Suspensions would make it more difficult for students to graduate. Increased support from experienced teachers and other aides can motivate students to do well in school. Higher teacher salary will motivate teachers to put more effort in helping their students.

## 2.3 ED*Facts*

High school graduation rate comes from the ED*Facts* dataset. ED*Facts* collects and analyzes data from pre-kindergarten through twelfth grade supplied by state education agencies with other data within the U.S. Department of Education (2025d).

## 2.4 Alternative Data Sources

A good alternate data source would be at the student level rather than the school level. This can be difficult to find, though, because the students are not legal adults and due to Family Educational Rights and Privacy Act (FERPA) which protects the distribution of educational records without the consent of parents or the student if they are over 18 years old. Information on absenteeism, family income level, extracurriculars of the students, involvement of parents and teachers, and reading level would also help in predicting graduation rates.
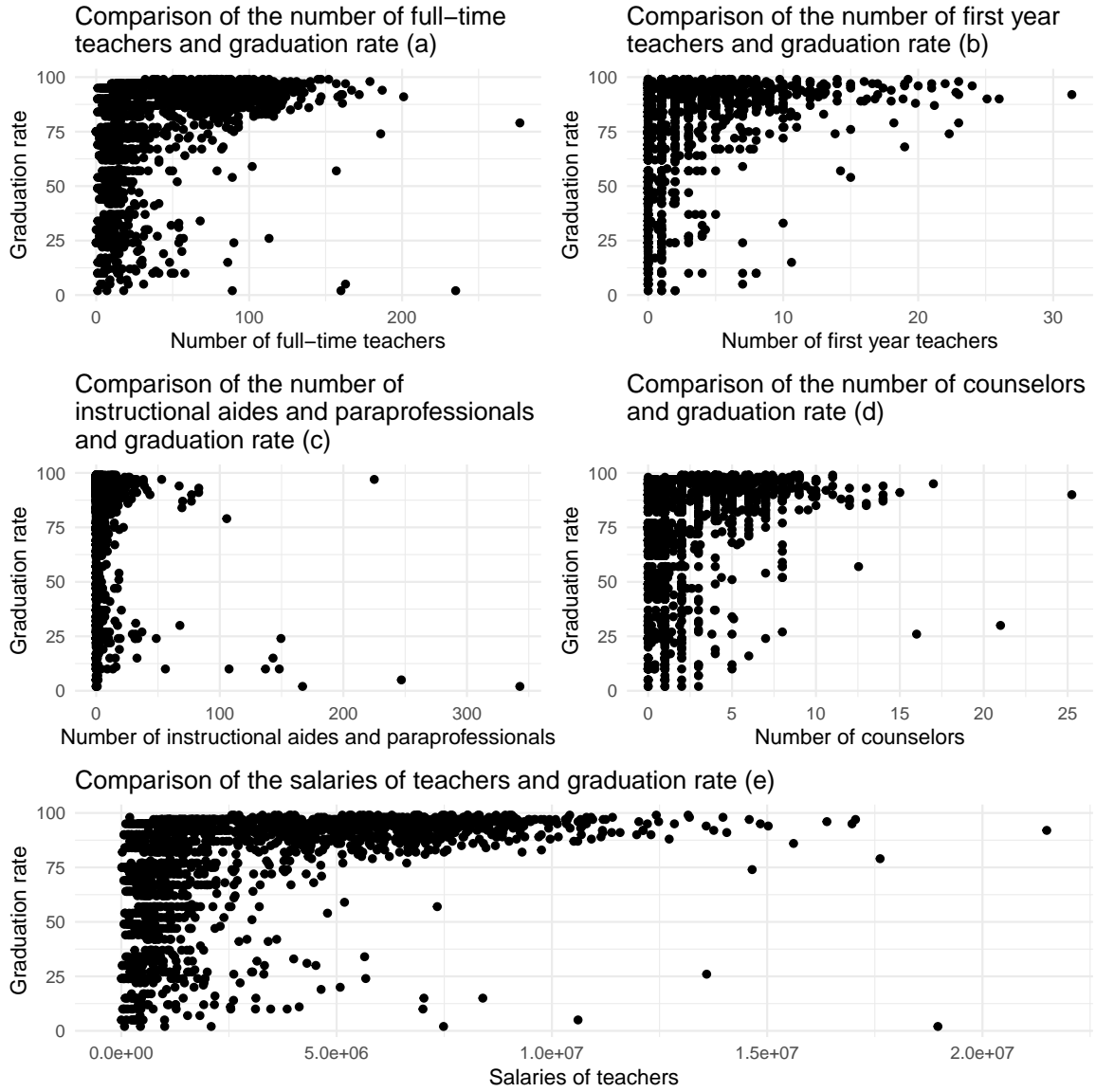
Figure 5: Plot (a) shows the relationship between the number of full-time teachers (x-axis) and graduation rate (y-axis). Plot (b) shows the relationship between the number of first year teachers (x-axis) and graduation rate (y-axis). Plot (c) shows the relationship between the number of instructional aides and paraprofessionals (x-axis) and graduation rate (y-axis). Plot (d) shows the relationship between the number of counselors (x-axis) and graduation rate (y-axis). Plot (e) shows the relationship between teacher salary (x-axis) and graduation rate (y-axis).

## 3 Methods

In an attempt to answer this research question, I will first fit a multiple linear regression model using all variables and no transformations. Let $Y_i$ represent the graduation rate at an individual school. Let $X_{i1}$ represent the number of students enrolled at that school, $X_{i2}$ equal 1 if a school is Title I eligible and 0 if it's not, $X_{i3}$ represent the number of students enrolled in the International Baccalaureate Diploma Program, $X_{i4}$ represent the number of students enrolled in the AP program, $X_{i5}$ represent the number of students enrolled in the gifted and talented program, $X_{i6}$ represent the number of students that took one or more AP exams, $X_{i7}$ represent the number of students participating in the SAT or ACT, $X_{i8}$ represent the number of suspensions, $X_{i9}$ represent the number of full-time teachers, $X_{i,10}$ represent the number of first-year teachers, $X_{i,11}$ represent the number of full-time equivalent instructional aides or paraprofessionals, $X_{i,12}$ represent the number of full-time equivalent school counselors, and $X_{i,13}$ represent teacher salary.

The model can we written as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \beta_7 X_{i7} + \beta_8 X_{i8} + \beta_9 X_{i9} + \beta_{10} X_{i,10} + \beta_{11} X_{i,11} + \beta_{12} X_{i,12} + \beta_{13}$$

where $\beta_0$ represents the mean of the probability distribution of the graduation rate when all predictors are 0, $\beta_1$ represents the average change in graduation rate for an additional student enrolled, holding all other variables constant, $beta_2$ represents the average change in graduation rate for a Title I school, holding all other predictors constant, $\beta_3$ represents the average change in graduation rate for an additional student enrolled in the IB program, holding all other variables constant, $\beta_4$ represents the average change in graduation rate for an additional student enrolled in the AP program, holding all other variables constant, $\beta_5$ represents the average change in graduation rate for an additional student enrolled in the gifted and talented program, holding all other variables constant, $\beta_6$ represents the average change in graduation rate for an additional student that took one or more AP exams, holding all other variables constant, $\beta_7$ represents the average change in graduation rate for an additional student participating in the SAT or ACT, holding all other variables constant, $\beta_8$ represents the average change in graduation rate for an additional suspension instance, holding all other variables constant, $\beta_9$ represents the average change in graduation rate for an additional full-time teacher, holding all other variables constant, $\beta_{10}$ represents the average change in graduation rate for an additional first-year teacher, holding all other variables constant, $\beta_{11}$ represents the average change in graduation rate for an additional full-time equivalent instructional aide or paraprofessional, holding all other variables constant, $\beta_{12}$ represents the average change in graduation rate for an additional full-time equivalent school counselor, holding all other variables constant, $\beta_{13}$ represents the average change in graduation rate for a one USD increase in teacher salary, holding all other variables constant, $n$ is the number of schools in the dataset, and $\varepsilon_i$ represents the error.

present any model or statistical methods clearly using mathematical notation and plain English explanations.

Every symbol and component of the model should be introduced and described.

This section must clearly describe any relevant assumptions and whether they are met.

Describe possible limitations of your analysis.

describe the software used for implementation

include any necessary steps taken for model selection and validation.

# 4 Results

describe the fitted model and its implications.
This section should include relevant tables, graphs, and quantitative results, as well as plain
English explanations of all of these components.

Clearly explain what the model results show and connect your results to the central research
question(s) and their broader context

# 5 Discussion

A brief summary of the paper

Discussion of your key findings and what their implications are for your research question

Discussion of some potential weaknesses of your study

Identification of potential improvements and extensions or future areas of related research.

# 6 Questions

1. can't remove all NAs without getting an empty dataset. what to do?
2. should i be having this many graphs?? does figure 2 look ok
3. still confused about interaction, esp how to check
4. should i put a graph of the log transformations?
5. do i need to make a graph or anything for graduation rate?
6. instead of seeing if there is enrollment in either ap, ib, or other program, should i do any
   program at all? (get issue with students being in multiple programs)
7. can i write the model in matrix form or does it have to be all written out

# References

2025a. *Education Data Explorer.* https://educationdata.urban.org/data-explorer/about.

———. 2025b. *Common Core of Data (CCD) - Common Core of Data (CCD).* National Center for Education Statistics. https://nces.ed.gov/ccd/.

———. 2025c. *National Center for Education Statistics (NCES).* U.S. Department of Education. https://nces.ed.gov/fastfacts/display.asp?id=158.

———. 2025d. *U.S. Department of Education.* https://www.ed.gov/data/edfacts-initiative.

———. 2025e. *Ocrdata.ed.gov.* https://ocrdata.ed.gov/.

Education Data Portal (Version 0.23.0), Urban Institute. 2023. "Common Core of Data; the Civil Rights Data Collection; Model Estimates of Poverty in Schools (MEPS); EDFacts." https://educationdata.urban.org/documentation/.

"High School Graduation." n.d. *Office of Disease Prevention and Health Promotion.* https://odphp.health.gov/healthypeople/priority-areas/social-determinants-health/literature-summaries/high-school-graduation.