Project 2*

Megan Joseph

December 3, 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.

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).

[talk more about NA values, and transformations]

The dataset has 2125 observations while the original has 6111 observations.

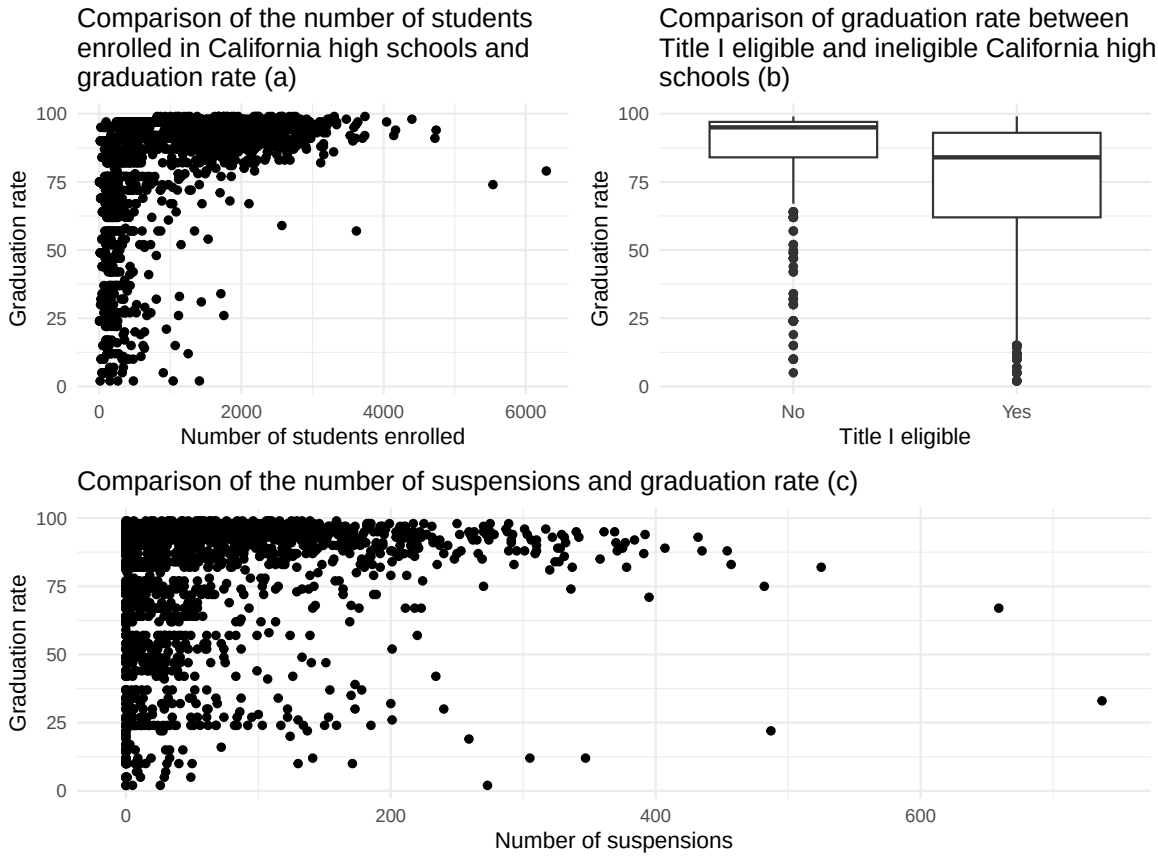

Figure 1: Plot (a) shows a scatter plot of the graduation rate over the number of students enrolled. Plot (b) shows a boxplot of graduation rate at Title I eligible and ineligible schools. Plot (c) shows the relationship between the number of suspensions in a school (x-axis) and graduation rate (y-axis).

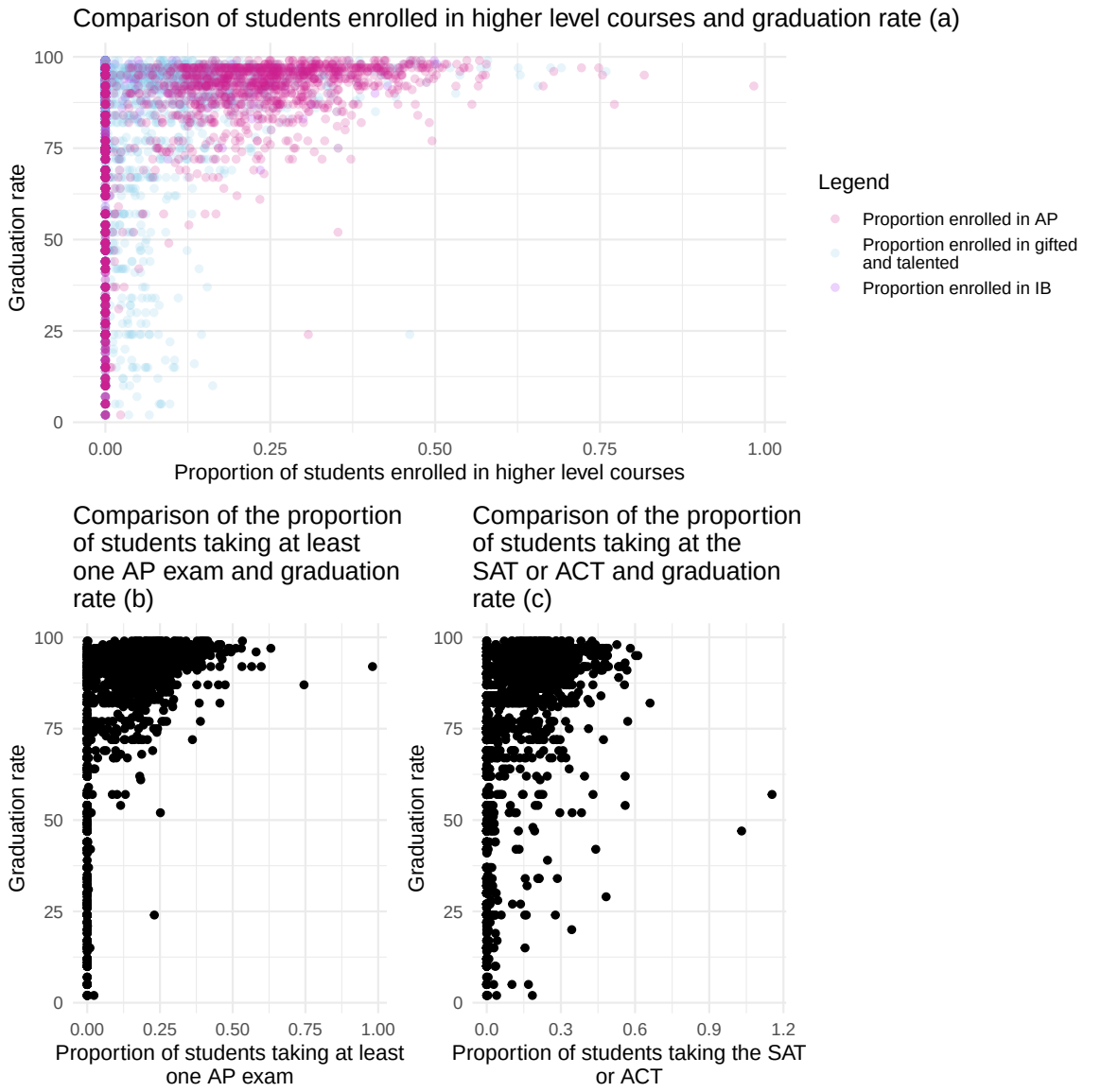

Figure 2: Plot (a) shows a scatter plot showing the relationship between the proportion of students enrolled in high level courses including the IB program (purple), AP program (pink), and gifted and talented programs (blue) (x-axis) and graduation rate (y-axis). Plot (b) shows a scatter plot of the graduation rate over the proportion of students taking at least one AP exam. Plot (c) shows a scatter plot of the graduation rate over the proportion of students taking either the SAT or ACT.

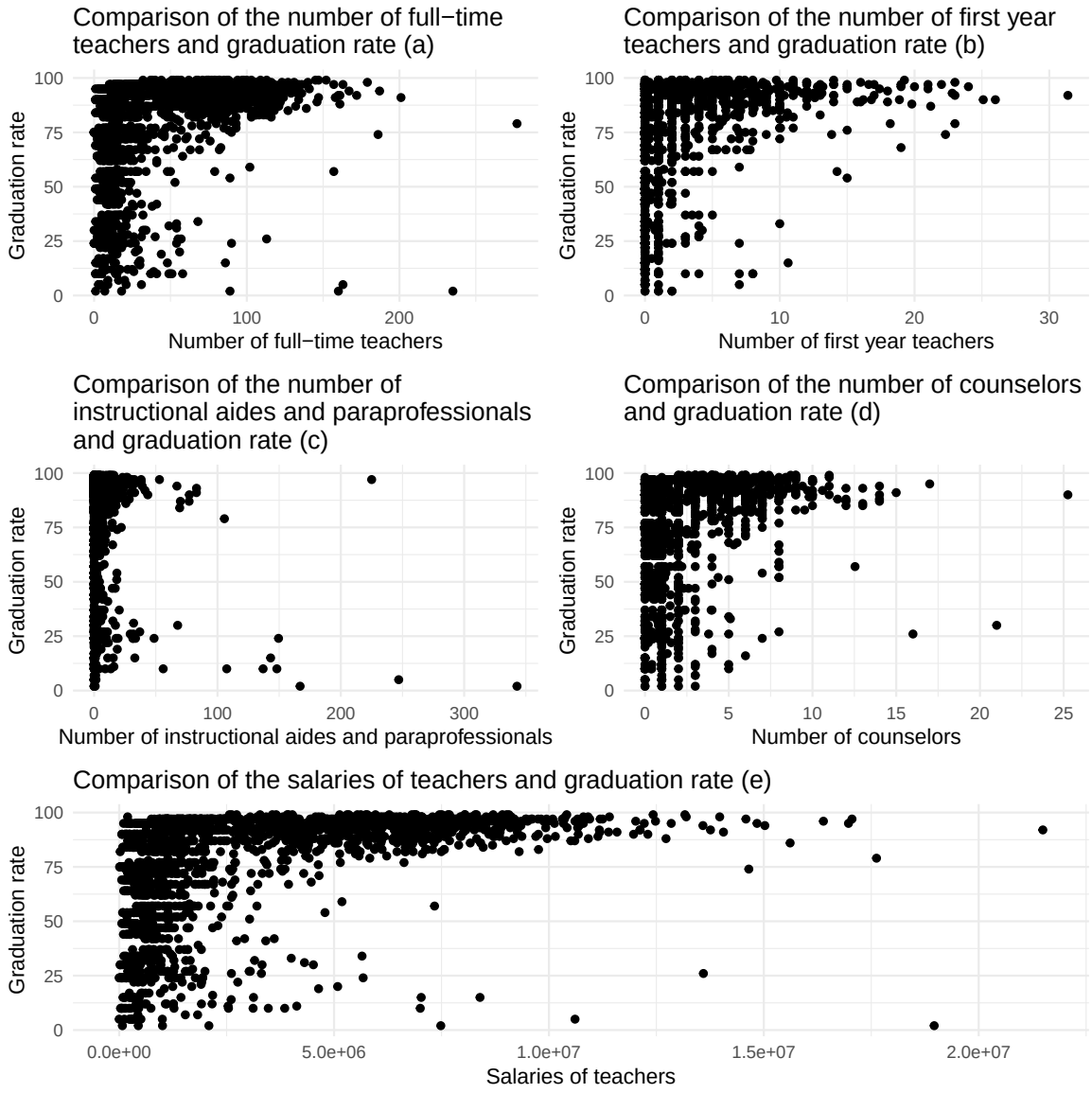

Figure 3: Plot (a) shows the relationship between the number of full-time teachers (x-axis) and graduation rate (y-axis). Plot (b) shows the relationship between the number of first year teachers (x-axis) and graduation rate (y-axis). Plot (c) shows the relationship between the number of instructional 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).

2.1 Common Core of Data

The Common Core of Data (CCD) is the United States Department of Education's database on all public elementary and secondary schools and school districts (2025b). The variables of interest from this dataset are the number of students enrolled and 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.

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, but I will use a square root transformation because of the large number of zeros. 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.).

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 key variables of interest 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 (a) shows a logarithmic relationship between these programs and graduation rate. Figure 2 (b) and (c), Figure 1 (c), and all plots in Figure 3 also show logarithmic relationships between the selected variables and graduation rate, so a square root 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 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*

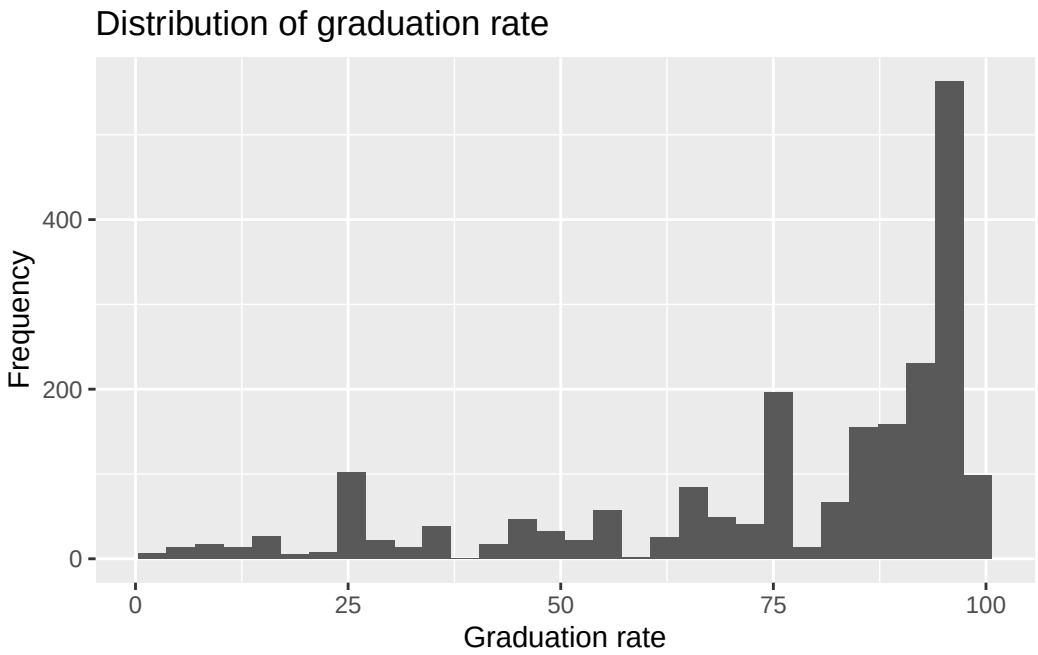

Figure 4: Histogram of graduation rate showing a left skewed distribution.

High school graduation rate comes from the ED*Facts* dataset. Figure 4 shows a left-skewed distribution in graduation rate.

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.

3 Methods

3.1 Base Model

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 be the 2125 x 1 vector where each entry represents the graduation rate at each school. Let X be the 2125 x 13 matrix where each column is a predictor. Let X_1 be the vector representing the number of students enrolled, X_2 be the vector equaling 1 if a school is Title I eligible and 0 if it's not, X_3 be the vector representing the proportion of students enrolled in the International Baccalaureate Diploma Program, X_4 be the vector representing the proportion of students enrolled in the AP program, X_5 be the vector representing the proportion of students enrolled in the gifted and talented program, X_6 be the vector representing the number of students that took one or more AP exams, X_7 be the vector representing the number of students participating in the SAT or ACT, X_8 be the vector representing the number of suspensions, X_9 be the vector representing the number of full-time teachers, X_{10} be the vector representing the number of first-year teachers, X_{11} be the vector representing the number of full-time equivalent instructional aides or paraprofessionals, X_{12} be the vector representing the number of full-time equivalent school counselors, and X_{13} be the vector representing teacher salary.

The model can we written as

$$Y = X\beta + \varepsilon$$

where β is the vector of coefficients for each predictor: β_0 represents the mean of the probability distribution of the graduation rate when all predictors are 0, β_1 represents the average change in graduation rate for an additional student enrolled, holding all other variables constant, β_2 represents the average change in graduation rate for a Title I school, holding all other predictors constant, β_3 represents the average change in graduation rate for an additional student enrolled in the IB program, holding all other variables constant, β_4 represents the average change in graduation rate for an additional student enrolled in the AP program, holding all other variables constant, β_5 represents the average change in graduation rate for an additional student enrolled in the gifted and talented program, holding all other variables constant, β_6 represents the average change in graduation rate for an additional student that took one or more AP exams, holding all other variables constant, β_7 represents the average change in graduation rate for an additional student participating in the SAT or ACT, holding all other variables constant, β_8 represents the average change in graduation rate for an additional suspension instance, holding all other variables constant, β_9 represents the average change in graduation rate for an additional full-time teacher, holding all other variables constant, β_{10} represents the average change in graduation rate for an additional first-year teacher, holding all other variables constant, β_{11} represents the average change in graduation rate for an additional full-time equivalent instructional aide or paraprofessional, holding all other variables constant, β_{12} represents the average change in graduation rate for an additional full-time equivalent school counselor, holding all other variables constant, β_{13} represents the average

change in graduation rate for a one USD increase in teacher salary, holding all other variables constant. ε is the vector containing the error.

3.2 Transformed Model

The next model used is the model where each predictor except Title I eligibility is transformed using the square root. While a log transformation would have been ideal, there are many zeros in the dataset which is not an acceptable input in a logarithmic function. A square root transformation is similar to a logarithmic one and allows zeros in its domain which makes it a viable substitution.

The model can be written as:

$$Y = \sqrt{X}\beta + \varepsilon$$

where each variable is equivalent to those in the base model.

3.3 Least Absolute Shrinkage and Selection Operator (LASSO) Regression Model

The last model created is one using LASSO regression to perform variable selection. LASSO is a shrinkage or regularization regression method that has the ability to "shrink" estimated coefficients to zero, carrying out variable selection. Alongside this, it reduces the variance of parameter estimates. LASSO aims to minimize the equation:

$$\sum_{i=1}^{n}(Y_i - Z_i^T\beta)^2 + \lambda\sum_{j=1}^{p-1}|\beta_j|$$

where $Z_i = \sqrt{X_i}$, $\lambda \geq 0$ is a tuning parameter that controls the level of regularization, n represents the number of observations, and p is the number of predictors. Essentially, we are minimizing the error sum of squares (SSE) with a penalty term added to it.

When picking the best lambda value, `cv.glmnet()` uses cross-validation to pick the best lambda that produces a mean squared error within one standard error of the mean squared error of the true best lambda (Friedman, Hastie, and Tibshirani 2010a). This is because the true best lambda may overfit the model by setting many estimated coefficients to zero. The user is able to pick any lambda used by the function based on their own analysis. For ease, I will use the default best lambda.

3.4 Model Selection

To determine which model is the best, I will first split the data into a 70% training, 15% testing, and 15% validation set. Then, I will train the models on the training set and find the mean squared prediction error using the validation set. This is defined as

$$E[(Y^* - \hat{Y}^*)^2]$$

where Y^* is the true graduation rate and $\hat{Y}^*$ is the predicted graduation rate using the model. This is the expected squared difference between the prediction and the true value over the validation data population. The one that gives the best mean squared error will be the best model. Finally, I will use the testing set to estimate the performance, such as error rates, for the final model.

3.5 Assumptions of Linear Regression

The assumptions of using linear regression for inference are:

1. The variables used accurately reflect the quantities of interest, the model should include all predictors, and the model should be applicable for the research question.
2. The sample data is representative of the population of interest.
3. The mean function must be correct.
4. Errors must be independent.
5. The variance of the errors must be equal.
6. Errors must be normal.

Since the goal of this analysis is point prediction, the most important assumption that should be met is linearity. If my goal was to create prediction intervals, I would need to meet the normality of errors assumption. After the transformations, there appears to be a linear relationship between the predictors and graduation rate.

3.6 Limitations of Analysis

One limitation of this analysis is that it is at the school level rather than the student level. While there are many reasons as to why this is the case, it does limit the analysis, and predictions can only be made for schools. An analysis at the student level, though, would change the problem from linear regression to logistic regression since the response would be whether the student graduated or not. Another limitation is that interaction effects are not included in any models. This is because it reduces the interpretability even though it could improve performance. In a future analysis, researching which variables have a theroretical interaction with each other would build upon this. Additionally, there may be models better suited for this research question that I have not learned yet.

3.7 Software

I use the `lm()` function from the R programming language (R Core Team 2025) to create the base model and the transformed model. I use the `cv.glmnet()` function from the `glmnet` package (Friedman, Hastie, and Tibshirani 2010b) to use LASSO regression to select the best model.

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. should i be having this many graphs?? does figure 2 look ok only 4-5, maybe if 0 remove it as a proportion
2. do i need to explain lasso more, change x to z and explain how lambda is calculated
3. do i need to add more info for describing the dataset

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/.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010b. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22. https://doi.org/10.18637/jss.v033.i01.

———. 2010a. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22. http://www.jstatsoft.org/.

"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.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.