# Project 2 Proposal

## Research question

The research question I aim to answer in this analysis is the following: Can college readiness, student demographics, and school demographics predict graduation rates in California high schools? College readiness entails enrollment in AP/IB/Honors courses, taking the SAT/ACT, etc. Student demographics entails enrollment in free and/or reduced lunch programs, students living in poverty, absenteeism, number of suspensions, etc. School demographics entails number of students enrolled, teacher salaries, number of first-time teachers, etc.

## Data

The data I will use is from the Education Data Explorer from the Urban Institute's Center on Education Data and Policy (Education Data Portal (Version 0.23.0), Urban Institute 2023). It combines multiple national data sources on schools, districts, and colleges to better analyze trends. I focused on high schools in California and included data on school characteristics, absenteeism, college readiness, course offerings, student demographics, discipline, school safety, student outcomes, and teachers and staff. I've linked the query I used to generate my dataset. The observational unit is an individual high school in California. The key variables of interest are student enrollment, chronically absent students, students enrolled in the International Baccalaureate Diploma Program, students enrolled in the gifted and talented programs, students enrolled in at least one AP course, students who took one or more AP exams, number of students enrolled in a dual enrollment or dual credit program, students participating in the SAT and ACT tests, number of students eligible for free or reduced-price lunch, estimated percentage of students living in poverty, number of suspensions, 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.

## Model

The model I plan to use is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i + ... + \beta_p X_i + \varepsilon$$

where

- $Y_i$ represents the graduation rate at each school
- each $\beta_i$ for $i = 0, ..., p - 1$ represents the coefficient for each predictor selected
- $X_i$ is a predictor from college readiness, student demographics, and school demographics
- $\varepsilon$ represents the error

## Proposed extension

This model will go beyond simple linear regression in that it has multiple predictors and requires variable selection. I may also implement a transformation if necessary. Additionally, interaction terms may be needed depending on what predictors are selected.

## Analysis plan

Regression analysis will allow me to answer this research question because I will be able to quantify how much each predictor impacts predictions for graduation rates while holding other variables constant. I will evaluate my predictions by splitting the data into a train/test/validation split or use cross-validation. Then, I can make multiple models using the base model, interaction terms, transformations, and variable selection. I can compare the validation error between these different models.

## References

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