*The proposal should be a statement of your research question and a brief (1/2 page) description of your analysis plan. This will not be graded, but will need to be approved by me. If necessary, we may iterate on the proposal. After you submit it, I will either approve it, or send it back to you with comments for amending. If I send it back, please submit an updated version is a timely fashion - the longer it takes to finish your proposal, the less time you have for your final proje*ct.

My proposal consists of 2 interrelated research questions:

1. What is the relationship between PreK-8th grade attendance rate and freshman year risk of chronic absenteeism, controlling for student characteristics such as race, sex, and socioeconomic status?
2. Using logistic regressions with school random intercepts for predictive modeling (mimicking models closest to that used in district/state early warning systems), what is the incremental gain in predictive accuracy of each additional year of student data from PreK until 9th grade?

The sample for my project is the population of students who enrolled in the Boston Public Schools (BPS) PreK program for four-year-olds between the 2007-2008 and 2010-2011 school years, followed up until the 2022-2023 school year. I have merged student data with Massachusetts state administrative records covering those years for students’ PreK through high school grades. I have data on students’ demographic characteristics, schools attended, disciplinary information, and attendance rates for those years.

For my analysis, I will first conduct descriptive statistics of the time trends for attendance between each of the four cohorts in my sample (e.g., examine if attendance trends for 2007-2008 students differ from 2010-2011 students). Then, I will use both logistic regressions and linear probability models with school random intercepts, student characteristic covariates, and attendance rate from past school years to iteratively predict 9th grade chronic absenteeism. The logistic regressions will be used for predictive modeling while the linear probability models will be used moreso for greater interpretability for an education audience (it’s common in education papers to put odds ratios next to linear probability model coefficients since they are easier to interpret, and we assume that unless the outcome is close to 0/1 the average will be similar between both models–but happy to hear thoughts if you think this is not best practice with a binary outcome). Predictive accuracy will be assessed using common machine learning model accuracy metrics (accuracy, true positive rate, true negative rate, AUC, and BER).