

Caiden Puma

Data 0200 – Group 2

### Scaffold 4 White Paper


Google Co-Lab:  DATA 0200 Scaffold Activity 4.ipynb

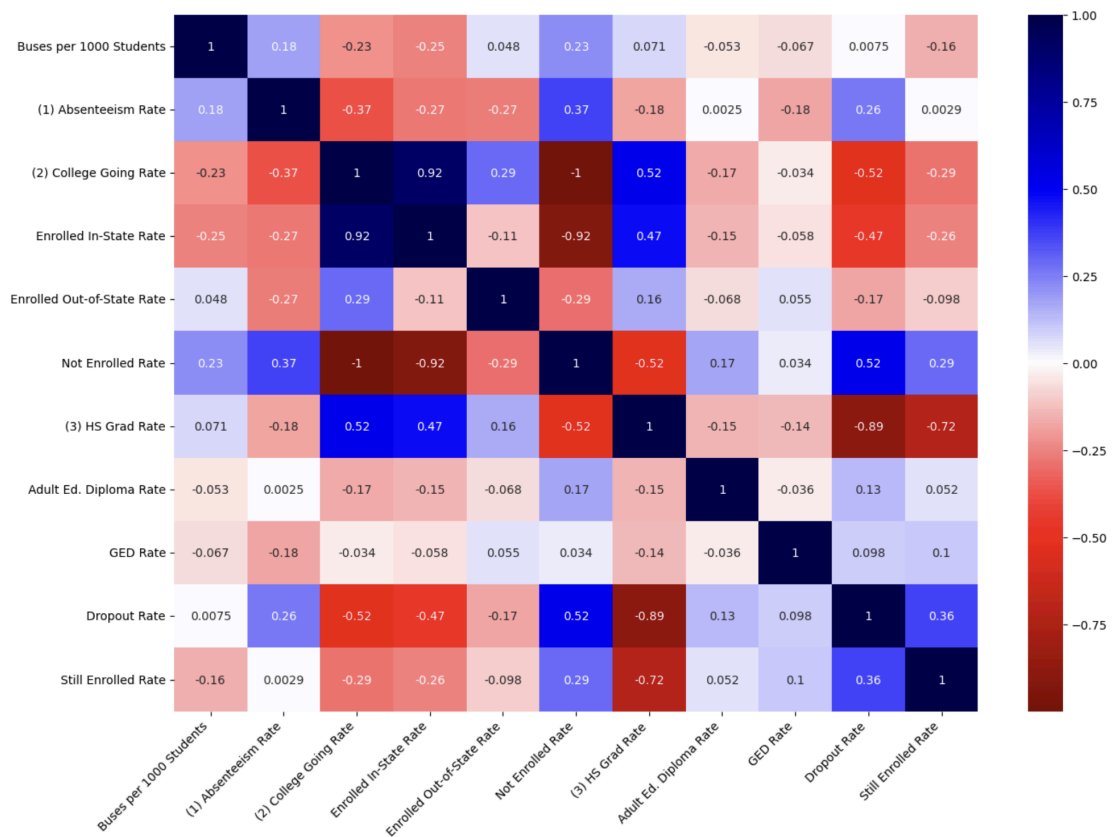
Tableau:

<https://public.tableau.com/app/profile/caiden.puma/viz/ScaffoldFour/Story1?publish=yes>

Because our group was assigned with the parent influence on education or school access and demographics topic for the project, we initially decided to test how national access to transportation affects student success outcomes via bus fleet size and standardized test (SAT) scores, but pivoted to focus on California-specific data due to the comprehensive data provided by the California Department of Education and consistency issues we ran into across states. Thus, our specific research question became: is there a statistically significant relationship between buses (per 1000 students) and metrics such as graduation rate, college enrollment, and chronic absenteeism?

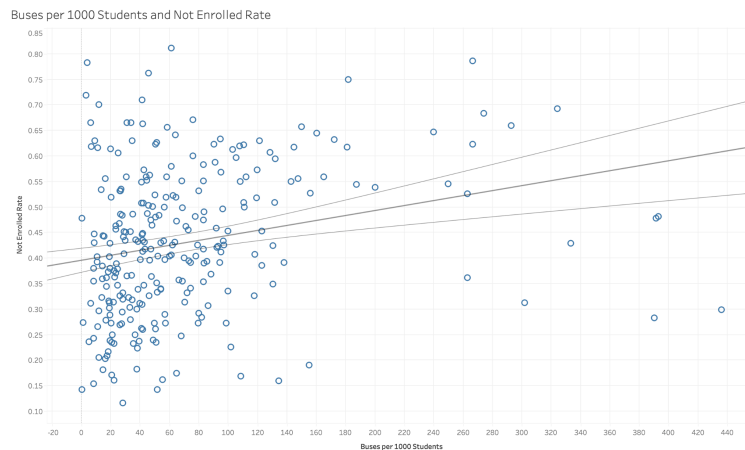
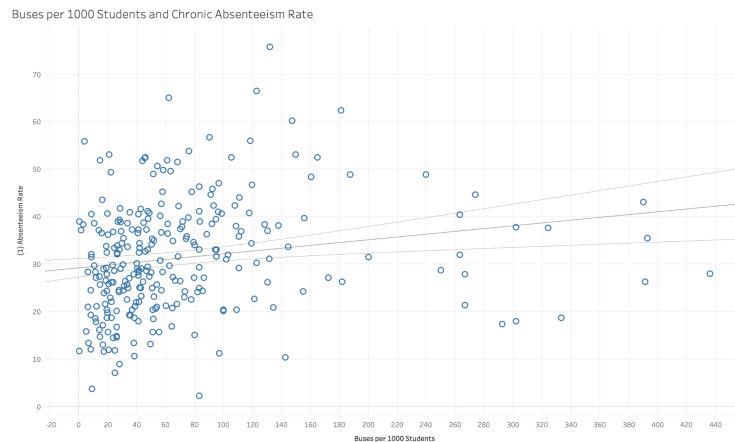
To explore these relationships, we performed both correlation analysis and OLS modeling with our explanatory variable of fleet size (normalized to buses per 1000 students) chosen to represent the transportation access with the success metrics. We found that larger districts were skewing the data, so we prepped the response variables by normalizing them to the student cohort size, effectively turning them into rates and allowing for comparability between districts. Additionally, we filtered any variables that were not provided to ensure consistency in the sample size amongst these variables. Finally, we ran Shapiro-Wilk, Breusch-Pagan, and Durbin-Watson analyses to ensure assumptions of normality of residuals, constant variance, and

independence for our OLS modeling, respectively. While not all of these assumptions were perfectly met, we had a fairly large sample size and proceeded with caution.



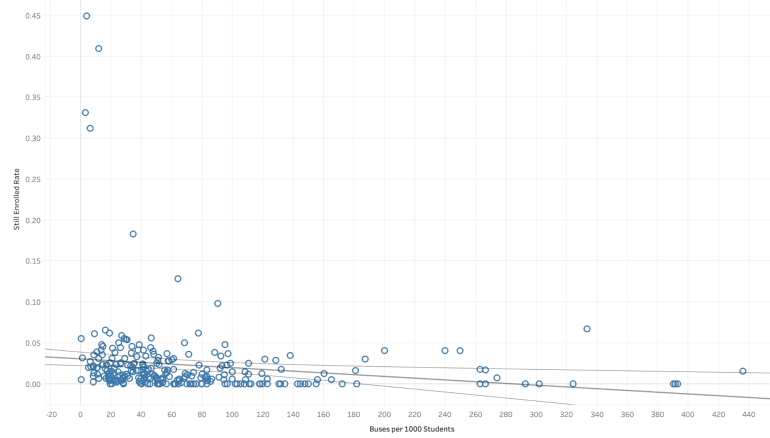
From the correlation heat map above, the key insights showed that bus access has moderately negative correlation with college going rate (-0.23) and enrolled in-state rate (-0.25), as well as moderately positive correlations with the not enrolled rate (0.23) and chronic absenteeism rate (0.18). This initially suggested that increased bus fleet size may be associated with lower college enrollment and higher drop-out rates, which was not what we expected to see as a result of increased access to transportation. This set the stage for some of the observations seen in the OLS regression as well, with success metrics having worse outcomes for those districts that had more buses per 1000 students which will be further discussed. The main take away from the heatmap was that while transportation access is correlated with these metrics, it is

likely not the defining factor and rather an indicator of more structural issues in those districts that have higher fleet sizes when considering the size of their cohorts, seeing as we would expect an increased access to transportation to motivate positive success metrics.

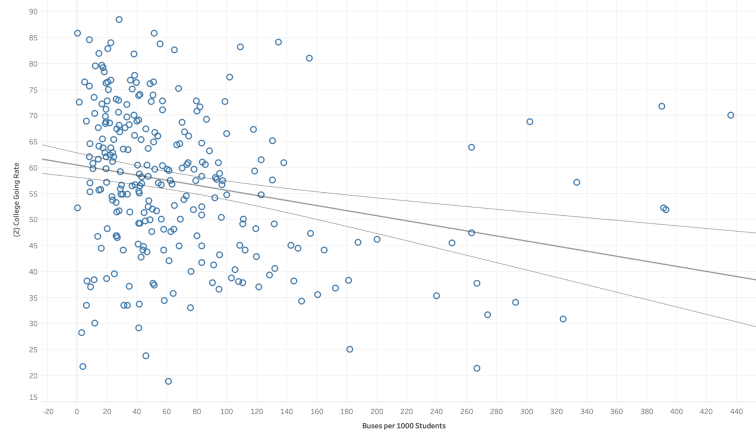


Those variables that did not have a statistically significant relationship with fleet size included high school completion rate, regular high school diploma rate, adult ED diploma rate, GED completion rate, dropout rate, and out-of-state enrollment rate. More interestingly, those with positive significant relationships included chronic absenteeism rate ( $p=0.002$ ) and not enrolled in rate ( $p<0.001$ ), with OLS results suggesting that for each additional bus per 1000 students, about 0.04% more students did not enroll and a rise of about 2.95 per 1000 student rise in chronic absenteeism.

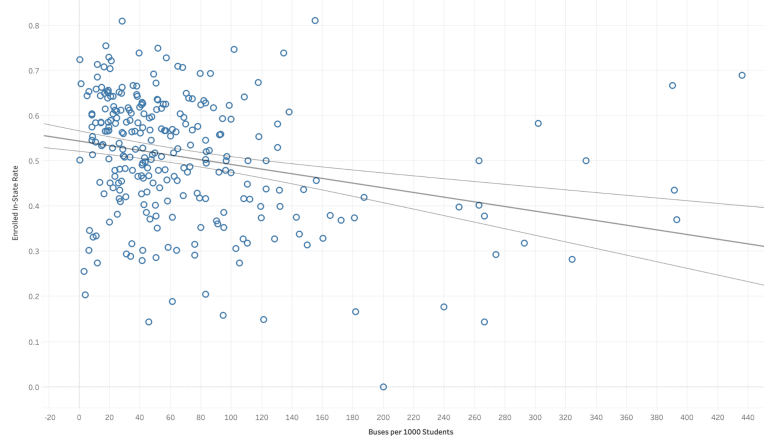
Buses per 1000 Students and Still Enrolled Rate



Buses per 1000 Students and College Going Rate



Buses per 1000 Students and Enrolled In-State Rate



Those with negative relationships included still enrolled rate, college-going/enrollment rate, and in-state enrollment rate. Results of the OLS modeling suggest that for each additional

buses per 1000 students, the rate of students enrolled and drops by about 0.01%, a 4.41% drop in students who attend college, and a slight drop in the percent of students currently enrolled in college and in-state college attendance (about 0.05% decline).

At first glance, these seem counterintuitive to what we would expect. As discussed, we hypothesized that increased access to transportation would lead to high success metrics, but our analysis found the opposite – districts with more buses per 1000 students tended to have slightly worse success outcomes. However, we can't definitively say that these are solely caused by access to buses—particularly because the R-squared values for our modeling were quite low, indicating that the bus variable explains little of the relationship in the grand scheme of things. Thinking critically about this relationship, we might find that those districts that are more rural and spread out require more buses and may also be more predisposed to economic or resource hardship than their urban counterparts. These are especially true as rural areas will have generally longer commute times, less public transportation, and less walkability. The limitations of this analysis focus more on correlation, which we cannot claim causation from, so the access to buses is likely not the cause of these outcomes but rather a symptom of more systematic issues that also affect these outcomes. We did not obtain variables relating to income levels, population demographics, location, funding, etc. and only worked with data for 2022, so these results are very limited. Thus, we should conclude that the analysis suggests that transportation access may be closer tied to deeper, systemic challenges in the education system. Those districts that require more buses likely face greater obstacles (longer distances, poverty, under-resourced schools, etc.) that influence whether students show up, graduate, and pursue higher education. These would need to be further explored in future iterations of the project, in order to assess any confounding variables that may come into play in the data.

Caiden Puma (me) 3 – Worked often on the code for the work alongside Sidney, contributing the normalization, heat maps, OLS regression, tests for normality/variance/etc.

Thomas Cronin 2 – Never missed a meeting, but did not contribute a ton to the actual work on the project or vocalize ideas regarding how we should approach the project. However, created tableau tables and helped in the process of visualizations.

Sidney Lin 3 – Did a lot of the coding, particularly in regard to cleaning the initial data and getting the data ready for analysis. Was consistent in working well with everyone, was communicative, and overall contributed most to the project.

Vicki Pu 3 – Never missed a meeting and was very proactive about setting up meetings and deadlines, had a good grasp of what needed to get done, and helped ensure things ran smoothly but did not contribute much to the actual coding or analysis – however, was vocal about ideas on how to approach our issues and helped in the approach.