Caiden Puma
DATA0200 Spring 2025
Project Group 2

Scaffold Activity 1 White Paper

COLAB LINK: https://colab.research.google.com/drive/1YAmO83KNxgsz10BtEydAZrf2zMrk7P0l?usp=sharing

There were many considerations about the data that we made as a group. The most important was in switching the datasets that we wanted to use. Initially, we had decided to use datasets for the amount of buses in the fleet for each state and a dataset that held SAT scores as well as subject-specific GPAs for each state. After some discussion on how we felt about the data, we felt that it may be too general and (because of the large geographic area) may include more confounding factors than we would know to deal with. In the process of data collection, we came across data from the California Department of Education that was quite comprehensive. Because the bus data that we initially had also contained district-specific information, we felt that it would be more productive to focus on a singular state than all of them. Another factor in this decision was that the bus dataset did not have information for a few states including Hawaii and Colorado, among other states and territories, that would have made our analysis non-comprehensive and introduced more limitations. While we could have chosen any state here, as many state education departments provide data, we decided that California was most represented in the bus data and would likely allow for more usable data.

Some of the ethical considerations that may arise as a result of these analyses include a focus on transportation as a factor for these variables, rather than a more systematic approach (funding, resources, teacher quality, etc.). The most immediately use case for an analysis of this form would be in resource allocation which may have broader impacts. For example, if resources are rerouted from areas where there does not seem to be a significant relationship between transportation and student outcomes, there may be a future where the loss of that transportation negatively impacts the student outcomes. The analyses may also be a bit too broad, as differences in urban/suburban access may manifest in such a way that reallocation of resources would negatively impact students in suburban areas who have fewer transportation options and longer commutes.

Beyond these considerations, there may also be some limitations in the data itself. While the variables reflect requirements of the question, they do not necessarily fit our original approach to measure student outcomes in performance metrics (standardized test scores and GPA) and rather approach through attendance of college, graduation rates, and absenteeism which may be more broad. The focus on districts may also pose challenges, as they are not all represented in the bus data as a result of them being self-reported, though it is relatively comprehensive. Similarly, we are focused on the year 2022 which may not be representative of longitudinal trends in both buses (though fleet size should remain relatively consistent) or student outcomes (which may have more fluctuations from the pandemic, innovations in technology and online learning, etc.). Despite these, I think we have a solid foundation for the data as all of the variables are in workable forms (number of buses, categories of districts, etc.).

Data Dictionary (this is quite extensive, as we want to preserve as much data that may be potentially useful in the future though we did cut a lot of non-necessary variables out and used many datasets as seen in the colab):

Variable Name	**Number of Observations**	**Data Type**
State	290	Categorical
LEAID	290	Categorical
District Name	290	Categorical
Total number of buses	290	Discrete
Number of buses 2020-newer	290	Discrete
Percent of buses 2020-newer	290	Continuous
Number of buses 2010-2019	290	Discrete
Percent of buses 2010-2019	290	Continuous
Number of buses 2000-2009	290	Discrete
Percent of buses 2000-2009	290	Continuous
Number of buses 1999 and older	290	Discrete
Percent of buses 1999 and older	290	Continuous
Number of buses with age unknown	290	Discrete
Percent of buses with age unknown	290	Continuous
Academic Year	290	Discrete
District Code	290	Categorical
ChronicAbsenteeismRate	290	Continuous
High School Completers	290	Discrete
Enrolled In College - Total (12 Months)	290	Discrete
College Going Rate - Total (12 Months)	290	Continuous
Enrolled In-State (12 Months)	290	Discrete
Enrolled Out-of-State (12 Months)	290	Discrete
Not Enrolled In College (12 Months)	290	Discrete

CohortStudents	290	Discrete
Regular HS Diploma Graduates (Count)	290	Discrete
Regular HS Diploma Graduates (Rate)	290	Continuous
Adult Ed. HS Diploma (Count)	290	Discrete
Adult Ed. HS Diploma (Rate)	290	Continuous
GED Completer (Count)	290	Discrete
GED Completer (Rate)	290	Continuous
Dropout (Count)	290	Discrete
Dropout (Rate)	290	Continuous
Still Enrolled (Count)	290	Discrete
Still Enrolled (Rate)	290	Continuous