Caiden Puma
DATA0200 – Group 2

Scaffold Activity 3 White Paper

Google Colab Link: scaffold_3.ipynb

The main decision we discussed within the group was the normalization of the data. It was agreed that the normalization was necessary because we identified in Scaffold 2 that the differences in district size were inherently skewing the results, particularly noted in the strong correlations between student cohort and metrics like diplomas, college enrollment, etc., and skewed data due to the few large districts. We discussed potentially only focusing on the data that was already provided as rates or percentages, but ultimately decided to normalize the raw data based on the total student cohort or total graduates (depending on the metric) to not jeopardize any of the variables and maintain robust analysis. Ultimately, this helped in having meaningful comparability and reducing instability in the OLS models.

The main question for the modeling is how school bus availability correlates with high school student success. Particularly, we are focused on examining whether access to school transportation (measured by the number of buses per 1000 students) is associated with differences in education outcomes across the California school districts. As such, we decided to use both a correlation analysis with the newly normalized data to assess the strength of the relationships between the variables as well as OLS regression to see if school bus availability significantly predicts the success metrics. These success metrics (continuous response variables) included high school completion rate, regular diploma rate, adult education diploma rate, GED rate, dropout rate, still enrolled rate, college-going rate, enrolled college rate, in-state rate, out-of-state rate, not enrolled rate, and chronic absenteeism rate. The explanatory variable (continuous) is the buses per 1000 students.

By using the correlation analysis on these variables, we were able to identify at a high level how the student success metrics were associated with bus accesses, with the heatmap allowing us to visualize the strength of these relationships to prepare expectations for the modeling. The OLS allowed us to more formally test these relationships, providing us with the magnitude of change per increase in 1 additional bus per 1000 students and significance of the associated (given R-square, P-values, confidence intervals, etc.) This is helpful because it tells us the explanatory power that access to buses has in the data!

The use of OLS assumes linearity, independence, and normality of residual values and variance. It also assumes the is no multicolinearity, but we are testing only one variable at a time, so the assumption does not apply in the context of our analysis. To test these assumptions, we used a Shapiro-Wilk test to assess the normality of the residuals and the Breusch-Pagan test to assess the variability of the residuals. We specifically used to p-values, so if the p-value for both the Shapiro-Wilk and Breusch-Pagan test is greater than 0.05, then the assumptions are satisfied.

For tests of independence, the Durbin-Watson test from the OLS output is used (in which values from 1.5-2.5 suggest independence). The output of the tests was:

```
 Assumption Check: HS Completion Rate ~ Buses per 1000 Students
Shapiro-Wilk Normality Test: t=0.711, p=0.000
Breusch-Pagan Residual Varianve Test: p=0.943

 Assumption Check: Regular Diploma Rate ~ Buses per 1000 Students
Shapiro-Wilk Normality Test: t=0.689, p=0.000
Breusch-Pagan Residual Varianve Test: p=0.235

 Assumption Check: Adult Ed Diploma Rate ~ Buses per 1000 Students
Shapiro-Wilk Normality Test: t=0.188, p=0.000
Breusch-Pagan Residual Varianve Test: p=0.429

 Assumption Check: GED Rate ~ Buses per 1000 Students
Shapiro-Wilk Normality Test: t=0.311, p=0.000
Breusch-Pagan Residual Varianve Test: p=0.451

 Assumption Check: Dropout Rate ~ Buses per 1000 Students
Shapiro-Wilk Normality Test: t=0.703, p=0.000
Breusch-Pagan Residual Varianve Test: p=0.973

 Assumption Check: Still Enrolled Rate ~ Buses per 1000 Students
Shapiro-Wilk Normality Test: t=0.399, p=0.000
Breusch-Pagan Residual Varianve Test: p=0.090

 Assumption Check: College Going Rate ~ Buses per 1000 Students
Shapiro-Wilk Normality Test: t=0.994, p=0.360
Breusch-Pagan Residual Varianve Test: p=0.167

 Assumption Check: Enrolled College Rate (Completers) ~ Buses per 1000 Students
Shapiro-Wilk Normality Test: t=0.994, p=0.355
Breusch-Pagan Residual Varianve Test: p=0.169

 Assumption Check: In-State Rate ~ Buses per 1000 Students
Shapiro-Wilk Normality Test: t=0.987, p=0.016
Breusch-Pagan Residual Varianve Test: p=0.001

 Assumption Check: Out-of-State Rate ~ Buses per 1000 Students
Shapiro-Wilk Normality Test: t=0.794, p=0.000
Breusch-Pagan Residual Varianve Test: p=0.201

 Assumption Check: Not Enrolled Rate ~ Buses per 1000 Students
Shapiro-Wilk Normality Test: t=0.994, p=0.355
Breusch-Pagan Residual Varianve Test: p=0.169

 Assumption Check: (1) Chronic Absenteeism Rate ~ Buses per 1000 Students
Shapiro-Wilk Normality Test: t=0.989, p=0.038
Breusch-Pagan Residual Varianve Test: p=0.102
```

so we can see that only the college-going rate, enrolled college rate, and not enrolled rate satisfy the assumptions for the Shapiro-Wilk test, and all but the in-state rate satisfy the assumptions for the Breusch-Pagan test. Thus, the p-values from the OLS may be a bit unreliable, seeing as many did not pass the normality test, though the sample size is decently large, so this is not an end-all. This is particularly true for the in-state rate as it is the only variable that failed both assumptions!

From the results of the assumption checking, we are going to make a note as to be cautious of the output, rather than completely changing the model, seeing as the sample size is relatively large, as previously discussed. Compiling the outputs of the results from the OLS and assumption notes in table gives:

Regression (Rate)	R-Squared	Adjusted R-Squared	F-Statistic	P-Value of F-Statistic	Coefficient (Buses per 1000 Students)	P-Value (Buses per 1000 Students)	Residual Normality Satisfied	Residual Variance Satisfied	Independence Satisfied
HS Completion	0.000	-0.003	0.1341	0.715	3.15e-5	0.715	No	Yes	Yes
Regular Diploma	0.005	0.001	1.398	0.283	0.0001	0.238	No	Yes	Yes
Adult ED Diploma	0.003	-0.001	0.7717	0.380	-1.848e-6	0.380	No	Yes	Yes
GED	0.004	0.001	1.220	0.270	-2.859e-6	0.270	No	Yes	Yes
Dropout	0.000	-0.004	0.01544	0.901	7.345e-6	0.901	No	Yes	Yes
Still Enrolled	0.027	0.023	7.541	0.00643	-0.0001	0.006	No	Yes	Yes
College Going	0.051	0.047	14.63	0.000162	-0.0441	0.000	Yes	Yes	Yes
Enrolled College	0.051	0.047	14.63	0.000162	-0.0004	0.000	Yes	Yes	Yes
In-State	0.064	0.061	18.88	1.97e-5	-0.0005	0.000	No	No	Yes
Out-of-State	0.002	-0.001	0.6458	0.422	3.793e-5	0.422	No	Yes	Yes
Not Enrolled	0.051	0.047	14.63	0.000162	0.0004	0.000	Yes	Yes	Yes
Chronic Absenteeism	0.033	0.030	9.380	0.00241	0.0294	0.000	No	Yes	Yes

For the sake of conciseness, I want to look here at only the significant results (as many of the results were not). The results here are interesting because the significant results seem to be associated with negative outcomes rather than positive ones! The significant relationships were found in the variables for still enrolled, college-going, enrolled college, in-state, not enrolled, and chronic absenteeism with coefficients of -0.0001, -0.0441, -0.0004, -0.0005, 0.0004, and 0.0294, respectively. While these are quite small, we can interpret that for each additional bus per 1000 students, we expect a decrease of 0.01% in rate of students still enrolled, a decrease of 4.41% in college-going rate, decrease of 0.04% in the enrolled college rate, decrease of 0.05% in in-state rate, increase of 0.04% in the rate of students not enrolled, and 2.94% increase in chronic absenteeism. We wouldn't expect this, as we thought positive outcomes would increase with more access to buses while negative ones would decrease. Thinking outside of the data, we might be able to explain this by the efficiency of resource allocation in districts, inefficiency of the transportation system in certain areas which requires more buses, the need for more busses in areas of lower socioeconomic status, or other confounding variables (larger geographic areas, longer communtes, greater poverty, rural-urban divide, etc.) that arent accounted for in the data. In all, it seems that more buses might instead be correlated with more logistical challenges in transportation access, longer commutes (increasing absenteeism), or other confounding factors (socioeconomic status of districts, rural vs. urban populations, etc.) that we need to look into the broader context to better understand.