Matthew Duffy
Dr. Zhushan Li
Multivariate Statistical Analysis
August 3rd, 2025

Socioeconomic Status and Education

A state government wants to know if socioeconomic and demographic factors generally affect a student's success. They believe if these factors affect student success during high school, it will also often set a trajectory for their life outcomes. Changing educational norms and implementing new strategies and programs to best educate people of different backgrounds could enable social mobility like never seen before. The government wants to know how these factors affect a student's academic success, but also more unorthodox measures of success like their sense of belonging and engagement in school. Academic achievement is a straightforward indicator of success today's society but feeling part of something bigger than oneself and engaging in one's community are some of the most important things one can do in their life. If there are substantial influences on unorthodox success metrics, corrective action should be taken.

The dataset, *hsls_17_student_pets_sr_v1_0.sav*, describes students that started high school in 2009 and studied them throughout their high school years and into college. The dataset contains information about the student (GPA, Math test scores, sense of belonging, sense of engagement, race, sex…), but also their socioeconomic status, family income, parent education, and many more. The dataset was built by The National Center for Educational Statistics.

The dataset has ~25,000 records and over 9,000 columns. For this analysis, the data set was cleaned and only a small subset of columns (15) was used. Our subset of columns contains: student ID, socioeconomic status composite, quintile coding of socioeconomic composite, total family income from all sources 2008, parents'/guardians' highest level of education, parent 1: employment status, mathematics theta score, GPA for all academic courses, ever dropout, student's sex, student's race/ethnicity, poverty indicator (relative to 100% of Census poverty threshold), school locale (urbanicity), scale of student's sense of school belonging, and scale of student's school engagement.

1. State your research question.

Do socioeconomic and demographic factors affect student success?

2. State the method(s) you used to study the relationships between socioeconomic and demographic factors and student success. What is your reasoning for your choices?

For this analysis, I will employ exploratory data analysis (EDA), cluster analysis, and logistic regression to address the research question. EDA will reveal important insights about the data and what relationships likely exist through visualization of relationships. Cluster Analysis will uncover the natural structures within the data. The cluster solution could confirm if socioeconomic metrics often coincide with student success measures, designating an association between them. Lastly, I will use logistic regression to explore the relationships between socioeconomic/demographic factors and student success more granularly. Logistic regression robustly examines relationships between both continuous and categorical variables and offers clear metric indicators of the extent and significance of those relationships (via coefficients and p-values).

3. Explain how you delt with missing values in the dataset. Did you process all missingness in the same fashion?

This dataset contains many missing values. As a survey-based dataset, respondents can skip questions, not answer full surveys, drop from the study, or really do anything they want; responses can also simply go missing during data collection. In this dataset, the types of missingness are differentiated as follows: -9 (missing), -8 (unit non-response), -7 (legitimate skip), -6 (component not applicable), -4 (Item not administered or abbreviated interview) and -1 (don't know). In our subset of columns, the only missing values were -1 (appeared 30 times), -8 (appeared 6,715 times), and -9 (appeared 1,000 times).

Since values of -8 indicate unit non-response (or some large grouping did not respond to the survey), that data could be missing not at random (MNAR). MNAR data can often have an explanation (people of lower income do not answer surveys about their socioeconomics for example), so simply removing it from the analysis may introduce bias. To combat this, I analyzed if the unit non-response has any difference in means between variables, which could indicate that the missing data represents some underrepresented group. The box plots (Figures 1-20) show that the missing values were either missing in other variables as well or do not add any unique perspective to the dataset, so I chose to remove them from the data for the sake of this analysis. I also elected to remove the small amount of -1 values present in the GPA variable since GPA is an important variable in the analysis and there were only uniquely missing records. The dataset is also quite large, so dropping those records allowed for a smaller sample to work as well.

For the values of -9, since they were designated as just *missing*, imputation seemed appropriate. Imputation allows us to avoid deleting any more of the data and should help maintain its representativeness. The only exceptions to imputation were for race and sex variables; if someone chose to not specify race or sex on a survey, it is unethical to take an educated guess based on other information they reported, so those values were simply dropped with the values of -8 and -1. For imputation, I used K nearest neighbors' (Knn) imputation to impute both categorical and continuous variables in the dataset. I first set all values of -9 in the dataset to null and then ran the Knn imputer from sklearn on the dataset. I set the number of neighbors to five, so the values are imputed as the average of the five closest points (of course barring the null values). This approach is more sophisticated than standard, group mean, or mode imputation, as it considers the non-null values in the observation when imputing missing data. Once imputed, the categorical values were then transformed into whole numbers of type category. Lastly, because GPA values are rounded to the nearest half I transformed the imputed Knn values to match that convention. The final cleaned and imputed dataset ended up with 15 columns and 15,480 rows.

4. Describe any relevant findings from exploratory data analysis (EDA). Were there insights that helped answer the research question?

In EDA, I analyzed descriptive statistics, distributions, correlations, group means and used visualizations to reveal patterns and relationships in the data. Many of the insights from this exploration guided my decisions during cluster analysis and logistic regression.

The descriptive statistics only had a few interesting aspects. For reference, the continuous variables in this analysis are GPA, math theta score, sense of belonging, school engagement, and socio-economic composite score. The categorical variables are socioeconomic quintile, family income, parent highest education, parent employment, ever dropout, sex, race, poverty, and locale. The average GPA in the dataset is 2.678876 and the most common quintile for socioeconomic status is 5 (the highest quintile). Each continuous variable except GPA spans negative and positive values (minimum school belonging is -4.35 and the maximum socioeconomic score is 2.8807 for reference), while GPA spans 0 to 4. These metrics are helpful to have in mind as analysis continued but did not describe any relevant information regarding the research question.

For distributions, socioeconomic score has some right skew (0.3), with more concentrated values on the left tail. It also has a negative kurtosis (-0.37), meaning the tails are lighter and the peak is flatter. The math theta score (a scaled metric that measures a student's ability in math by accounting for item difficulty) has an approximately normal distribution with small skew and kurtosis. GPA has a hefty left skew of -0.53, with the higher GPA's (2 and above) having more frequency. School belonging also has a high left skew of -0.48; there small amounts negative values spanning from about -4 to -1, but then the frequencies pick up drastically. That makes the right tail much heavier (which aligns with the kurtosis value of 0.52). School engagement is similar with a large left skew of -0.80 and kurtosis of 0.39.

Furthermore, I generated a correlation matrix for the continuous variables in the dataset (Figure 21). All continuous variables had significant and positive correlations with each other. The strongest correlation was between GPA and math score theta, with a correlation of 0.559. One's ability in math does not necessarily mean their GPA is high, but of course success in math often coincides with success in other subjects. It is worth noting that these are connected and significantly correlated, meaning they should not be used to predict each other. Socioeconomic score has the next two highest correlations with GPA and math theta score, 0.416 and 0.408 respectively. These relationships are another indication that socioeconomics affect academic success. Socioeconomic score has much lower correlations with school belonging and engagement, 0.164 and 0.146 respectively, indicating that socioeconomic does not affect those unorthodox success metrics in the same way as the academic ones, and may only affect them in minimal ways. Other notable correlations are GPA and school engagement, with a value of 0.327, and school belonging and school engagement, with a value of 0.289. Those correlations suggest that academic and extracurricular success have some association.

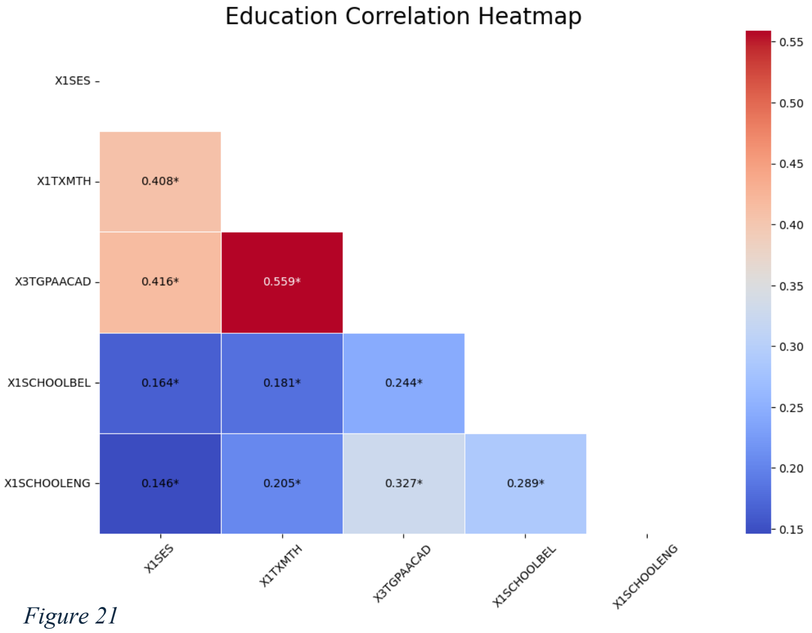


Figure 21

I then visualized the relationship between GPA and socioeconomic status using a simple line chart (Figure 22), which displays the mean socioeconomic composite score by GPA. The graph clearly shows that as GPA rises, so does the socioeconomic score, further demonstrating their positive relationship.

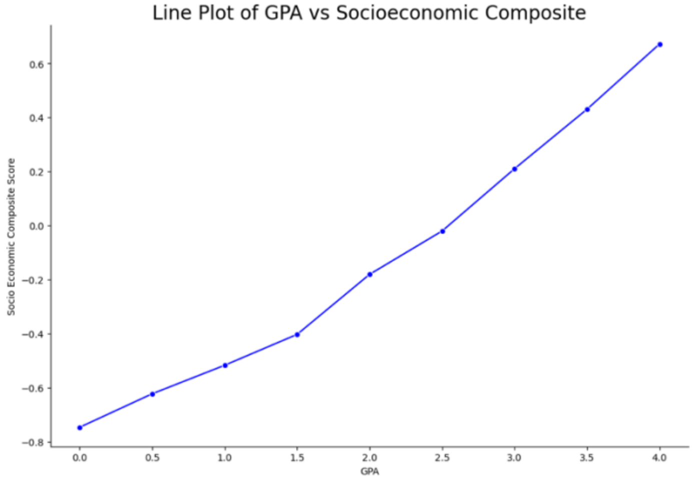
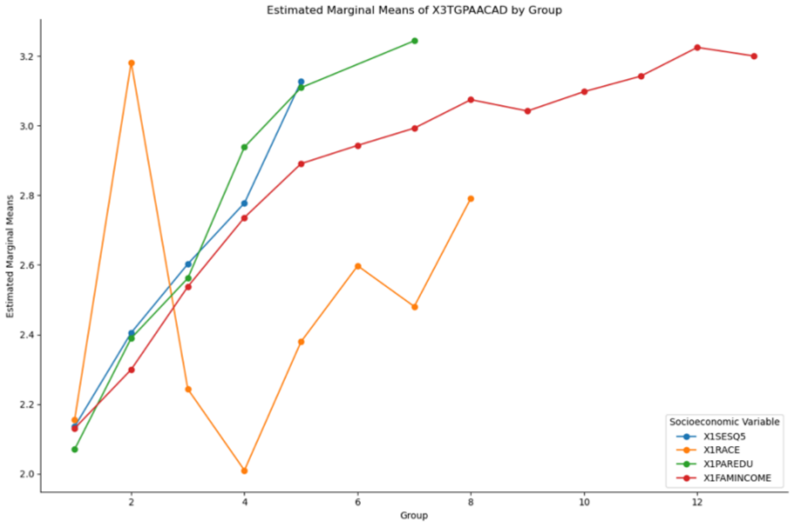


Figure 22

Lastly, I examined and visualized the group means for the continuous variables grouped by the main socioeconomic variables (Tables 1 and 2 in appendix show grouped means for socioeconomic quintile and race, Figures 23-26 are visualizations). The group means by socioeconomic quintile show that every student success metric has a positive relationship with socioeconomic status. The group means by race show that the race variables have very different means for success indicators. These mean differences suggests that socioeconomic and demographic factors may affect student success. Moreover, figures 23-26 continue to support that claim; they show that the group means move in a similar pattern across all continuous variables, just to slightly different degrees and each variable, except race, has a positive relationship with each success variable. Only student sense of belonging has a visually different graph with the same movements, the The similar mean movements across continuous variables indicate that the socioeconomic/ demographic factors may be affecting all the measures of student success in similar ways. The exploratory data analysis offers strong evidence that the success variables and socioeconomic/ demographic variables have some meaningful relationships.



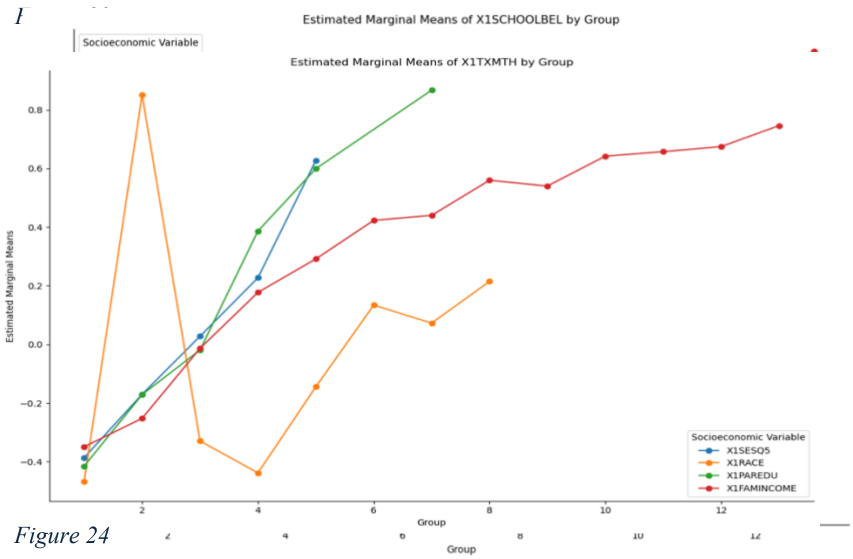


Figure 24

Figure 25

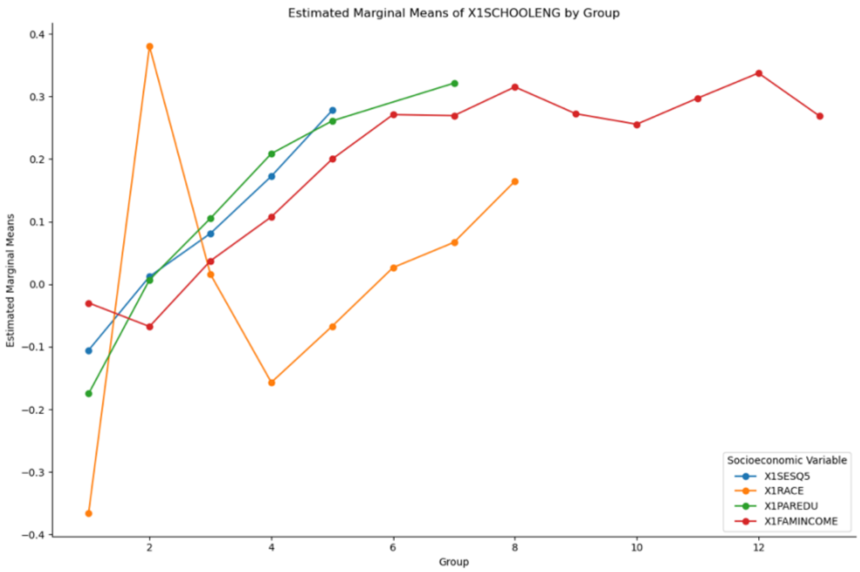


Figure 26

5. Are there underlying structures in the data? Describe how you discovered it and what insights they reveal about regarding the research question.

Cluster analysis revealed potential underlying structures in the data by identifying meaningful groupings in the data. I hypothesize that there should be at least two, if not more, distinct groupings influenced by socioeconomic factors. There should be one group representing students with higher socioeconomic status and another for those with lower socioeconomic status, and these groupings should align with student success metrics.

Before conducting the cluster analysis, I set up the data and addressed three assumptions of cluster analysis: existing structure, representativeness of the sample, and multicollinearity.

For preprocessing, I filtered the data for only continuous variables and removed 120 outliers from the dataset, leaving it with 15,360 records. Even though outliers make up less than 1% of the observations and represent valid observations, they were removed for cluster analysis since it is sensitive to outliers. I also transformed the data into z-scores, so the varying scales of the data did not distort the distance calculations used for cluster analysis.

Regarding the assumption of existing structure, cluster analysis will generate clusters, so it is important that a natural structure exists. Otherwise, the algorithm will group things with no meaning, leading to false conclusions. Exploratory data analysis and empirical knowledge suggest that higher socioeconomic status often coincides with higher student success, so there should be at least two distinct groupings.

Regarding sample representativeness, this dataset comes from the National Center for Educational Statistics (NCES). Data was collected from over 25,000 students that started high school in 2009. The cleaned version of the dataset that this analysis is using confirmed that none of the variables being dropped represented some unrepresented group in the population. Thus, we can extend the sample findings to the population.

Regarding multicollinearity, correlated variable can overly influence cluster solutions, causing the uncorrelated variables to have little impact. Therefore, it is vital to ensure no substantial multicollinearity exists in the data; I conducted a variance inflation factor (VIF) analysis on the potential cluster variables, and all variables had VIF scores less than 1.25. Since VIF scores have a general limit of 5, these scores indicate that there is little to no multicollinearity in the data.

Once I confirmed assumptions were met, I began the cluster analysis by performing hierarchical clustering to build a dendrogram, which visualized the clustering process (Figure 27). Because the dataset is so large, the visualization only displays the last 50 clusters formed. The dendrogram shows that there are two main groupings in the data, and then small variations within those main two groups. Based on the dendrogram, I will start with a k-means cluster analysis where k=2 to examine the main two groups and then look at one where k=4 to see any nuance.

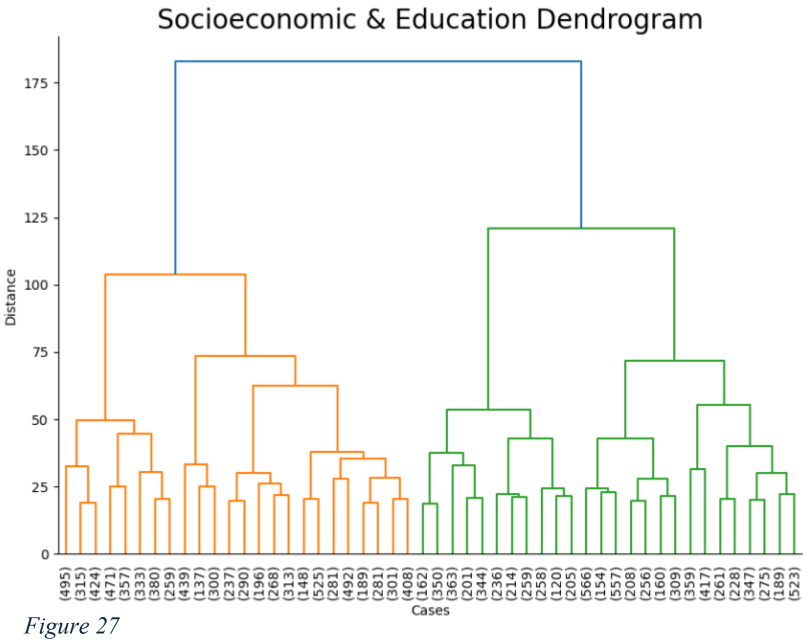


Figure 27

Moving into k-means clustering, I decided to use K-means++ instead of calculating initial centroids since it optimizes clusters through the reassignment of observations during the clustering process. After clustering the data, I verified that the differences in each variable mean across the clusters were statistically significant with ANOVA. The ANOVA results confirmed that the means are drastically different with F scores all over 5,000 and p-values of 0. Figure 28 visualizes the mean profile of the clusters revealing two distinct groups: cluster 0, those of higher socioeconomic status and student success, and cluster 1, those of lower socioeconomic status and student success. Cluster 0 scores above the mean while cluster 1 scores below. GPA has mean difference between clusters, with cluster 0 having a mean of about 0.6 and cluster 1 having a mean of about -0.75. Socioeconomic score and math theta score are similarly distant across the two clusters and school engagement and belonging close the gap a bit more than the rest (belonging and engagement, cluster 0 mean ≈ 0.4 and cluster 1 mean ≈ -0.4). This shows that while there are two distinct groups based on high and low socioeconomics and student success, the unorthodox success metrics have different relationships with socioeconomic factors than those for academic success.

Table 3 examines how well the clusters group students by family income, a categorical variable that was not included in the cluster analysis. The PercentOfGroup column displays how much of that family income category belongs that that cluster. From family income 1 – 3 (any income less than or equal to $55,000 a year), most people are in cluster 1. For family income less than or equal to $35,000 a year, cluster 1 completely dominates with over 70% majority. Once a family income gets above $75,000 a year, the people are predominately in cluster 0, with every category above 6 having over a 70% majority.

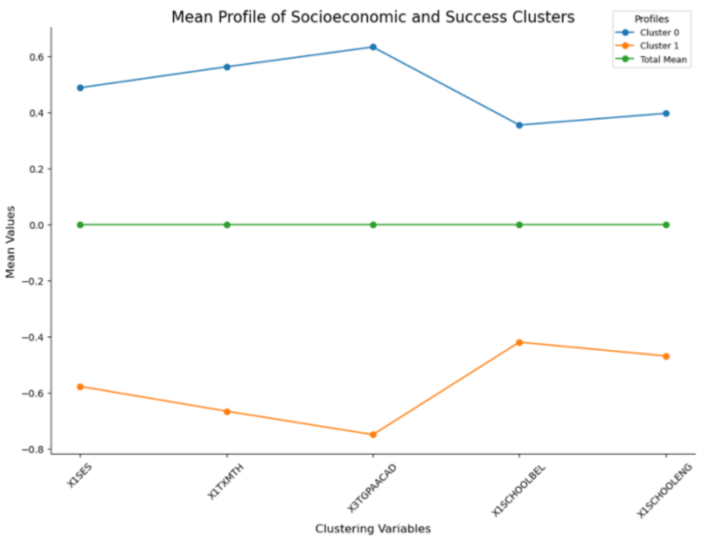


Figure 28

Income Level	Cluster	Count	PercentOfGroup
Family income less than or equal to $15,000	0	256	18.31%
Family income less than or equal to $15,000	1	1142	81.68%
Family income > $15,000 and <= $35,000	0	736	26.50%
Family income > $15,000 and <= $35,000	1	2041	73.50%
Family income > $35,000 and <= $55,000	0	1121	44.34%
Family income > $35,000 and <= $55,000	1	1407	55.65%
Family income > $55,000 and <= $75,000	0	1296	56.30%
Family income > $55,000 and <= $75,000	1	1006	43.70%
Family income > $75,000 and <= $95,000	0	1147	66.96%
Family income > $75,000 and <= $95,000	1	566	33.04%
Family income > $95,000 and <= $115,000	0	1013	73.94%
Family income > $95,000 and <= $115,000	1	357	26.058%
Family income > $115,000 and <= $135,000	0	696	77.85%
Family income > $115,000 and <= $135,000	1	198	22.147%
Family income > $135,000 and <= $155,000	0	586	84.68%
Family income > $135,000 and <= $155,000	1	106	15.32%
Family income > $155,000 and <=$175,000	0	281	82.16%
Family income > $155,000 and <=$175,000	1	61	17.83%
Family income > $175,000 and <= $195,000	0	191	86.82%
Family income > $175,000 and <= $195,000	1	29	13.18%
Family income > $195,000 and <= $215,000	0	261	89.38%
Family income > $195,000 and <= $215,000	1	31	10.62%
Family income > $215,000 and <= $235,000	0	94	86.24%
Family income > $215,000 and <= $235,000	1	15	13.76%
Family income > $235,000	0	653	90.32%
Family income > $235,000	1	70	9.68%

Table 3

Since socioeconomic composite score was included in that cluster analysis, it makes sense that there will be substantial deviation in family income, so I conducted another K-means cluster with only student success variables. I validated these means with ANOVA like before to confirm significantly different means across clusters (F statistics over 2,000 and p-values = 0). This mean profile (Figure 29) looks nearly identical to the cluster analysis with socioeconomic indicator showing that the success variables have this natural separation of successful and unsuccessful students which happens to closely align with their socioeconomic status. Looking at the breakdown of this cluster analysis by family income (Table 4), it follows the trend of cluster 1 dominating the lower income families and cluster 0 dominating the higher income ones. The groupings are less distinct since removing socioeconomic score, but the fact that the trend remains (increasing membership in cluster 0 as family income increases) demonstrates that a student success and socioeconomic status have a positive relationship.

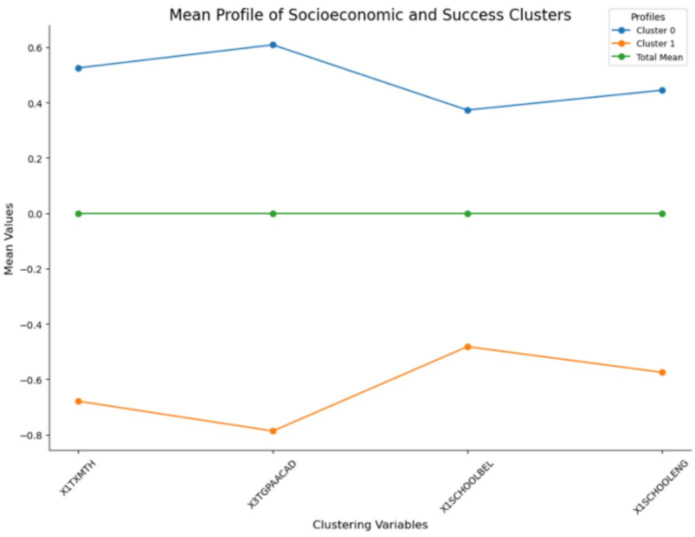


Figure 29

Lastly, I did a final cluster analysis with k=4, confirmed statistically significant mean differences across clusters and graphed the mean profile (Figure 30). The original bifurcation of the dataset is still present in cluster 0 and cluster 1, continuing to follow the trend of socioeconomic status coinciding with success. But cluster 2 and 3 show more nuanced structures than 0 and 1. Cluster 2 shows students that have above average socio-economic status and higher academic success metrics, but then lower score for school belonging and engagement. Cluster 3 is the contra to cluster 2, showing below-average socioeconomic status and academic performance, but higher levels of belonging and engagement. That major change suggests that unorthodox success metrics may not completely follow the same socioeconomic patterns as academic success .

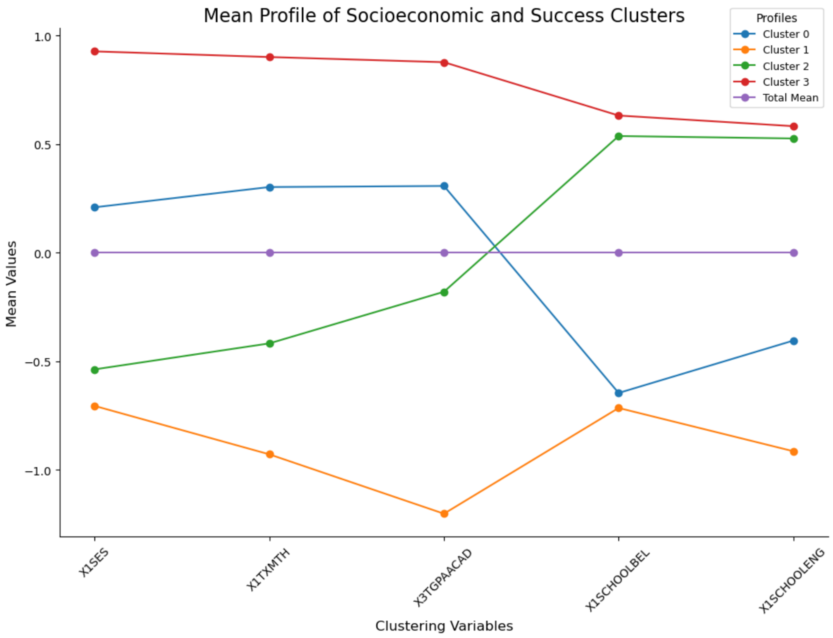


Figure 30

Income Level	Cluster Acad.	Count	PercentOfGroup
Family income less than or equal to $15,000	0	470	33.62%
Family income less than or equal to $15,000	1	928	66.38%
Family income > $15,000 and <= $35,000	0	1006	36.23%
Family income > $15,000 and <= $35,000	1	1771	63.77%
Family income > $35,000 and <= $55,000	0	1237	48.93%
Family income > $35,000 and <= $55,000	1	1291	51.07%
Family income > $55,000 and <= $75,000	0	1325	57.56%
Family income > $55,000 and <= $75,000	1	977	42.44%
Family income > $75,000 and <= $95,000	0	1120	65.38%
Family income > $75,000 and <= $95,000	1	593	34.62%
Family income > $95,000 and <= $115,000	0	964	70.36%
Family income > $95,000 and <= $115,000	1	406	29.64%
Family income > $115,000 and <= $135,000	0	641	71.70%
Family income > $115,000 and <= $135,000	1	253	28.30%
Family income > $135,000 and <= $155,000	0	539	77.89%
Family income > $135,000 and <= $155,000	1	153	22.11%
Family income > $155,000 and <=$175,000	0	260	76.02%
Family income > $155,000 and <=$175,000	1	82	23.98%
Family income > $175,000 and <= $195,000	0	170	77.27%
Family income > $175,000 and <= $195,000	1	50	22.73%
Family income > $195,000 and <= $215,000	0	235	80.48%
Family income > $195,000 and <= $215,000	1	57	19.52%
Family income > $215,000 and <= $235,000	0	91	83.49%
Family income > $215,000 and <= $235,000	1	18	16.51%
Family income > $235,000	0	604	83.54%
Family income > $235,000	1	119	16.46%

Table 4

6. Describe your model construction and state your hypothesis. Report the independent and dependent variables. Why did you choose them?

I fit a few logistic regression models to see if socioeconomic/demographic factors affect student success. The primary model regressed GPA on socioeconomic score, race, family income, and highest parent education level. My hypothesis is that socioeconomic/demographic factors affect student success. I then used the same model but changed the dependent variable to predict the other success metrics. My hypothesis for these remains the same: socioeconomic/ demographic factors affect student success. That would make the null hypothesis that those factors do not affect student success.

GPA was chosen as the primary dependent variable because it is a common metric used by colleges and businesses to evaluate people. GPA can partially determine what college someone gets into and what jobs they get offered in the future. The other eventual dependent variables are also strong metrics of success. Math prowess often leads to lucrative careers, and a sense of belonging and engagement in one's community is an important aspect of a life well-lived. Socioeconomic composite score was selected as a predictor because it quantifies socioeconomic status; it is the simplest way to understand where someone generally stands from a socioeconomic perspective. Race is our central demographic metric for the analysis, used to examine potential disparities in academic success across different racial and ethnic groups.

Given that systemic oppression may link race and socioeconomic indicators, potential multicollinearity will be monitored closely. Family income was chosen because it impacts a student's upbringing and educational opportunities (more income can allow students to have tutors and do more academics outside of school). While it is a significant component of socioeconomic status, having family income as its own variable allows for more insight into how it specifically affects a student's success. Similarly, parent education also allows for more direct insight into how it affects a student success. Having these big components of socioeconomic status in the model should allow the composite metric to explain how the interaction of many socioeconomic factors affect student success.

For data preprocessing, I made GPA binary by encoding GPAs of greater than or equal to 3 as successful (values of 1) and those less than 3 as unsuccessful (values of 0). Moreover, race, family income, and highest parent education level are all categorical variables, so all of them were one hot encoded. Family income and highest parent education level could be considered ordinal and left alone, but I wanted a more granular look at each component, so chose to one hot encode them as normal categorical variables. I also recoded the race value 8, white (non-Hispanic), to 0 so it would become the baseline variable, which works well because most of the survey respondents are white.

7. Present your model(s) equation and results. Were there any assumptions to confirm?

Logistic regression had four assumptions to address before modelling: binary outcome, sample size, independence of observations, and linearity between predictors and logits. The binary outcomes assumption was quite straightforward since I encoded in the preprocessing. There are 7,495 unsuccessful students and 7,985 successful students (based on GPA of 3.0). So, the distribution is very close to even, which is ideal for logistic regression. Moreover, the assumption of sample size was also met easily since the datasets contains over 15,000 observations. For independence of observations, I conducted a (variance inflation factor) VIF test so see if there was multicollinearity to be aware of. None of the variables have a VIF of more than 5, so the assumption is met! There are a few VIF scores that are close to 5, scores of 4.052 and 3.9718 for high school GED and bachelor's degree parent education levels respectively, so those are worth monitoring closely during modelling.

Lastly, I ran a model to check the interaction between socioeconomic status and its logit to check for linearity. This assumption only applies to continuous variables since categorical variables and inherently non-linear. For the test, I performed a logistic regression with the Broyden-Fletcher-Goldfarb-Shanno (bfgs) algorithm and a heteroscedasticity-consistent covariance matrix estimator (HC0). These specifications are best for checking for linearity. The test passed since the interaction term has a p-value greater than 0.05 (0.643), indicating indicates that there is no significant evidence to support non-linearity.

Once assumptions were confirmed, I fit the model. The model results are in Table 5 and coefficients are in Table 6.

Logit Regression Results - GPA			
Dep. Variable:	X3GPACAT	**No. Observations:**	15,480
Model:	Logit	**Df Residuals:**	15,454
Method:	MLE	**Df Model:**	25
Date:	Aug 2025	**Pseudo R^2:**	0.1281
Time:	12:00	**Log-Likelihood:**	-9,349
converged:	TRUE	**LL-Null:**	-10,722
Covariance Type:	nonrobust	**LLR p-value:**	0

Table 5

Variable	Coeff.	P-Value	Odds Ratio	Change in Odds
Intercept	-0.388512	0.001685	0.678065	-32.19%
Native American/Alaska Native, non-Hispanic	-0.981330	0.000016	0.374812	-62.52%
Asian, non-Hispanic	0.874024	0.000000	2.396535	139.65%
Black/African American, non-Hispanic	-0.925370	0.000000	0.396385	-60.36%
Hispanic, no race specified	-0.556806	0.022487	0.573036	-42.69%
Hispanic, race specified	-0.370326	0.000000	0.690509	-30.94%
More than one race, non-Hispanic	-0.246205	0.000077	0.781762	-21.82%
Native Hawaiian/Pacific Islander, non-Hispanic	-0.697198	0.005387	0.497979	-50.20%
Family income > $15,000 and <= $35,000	-0.093938	0.226962	0.910339	-8.97%
Family income > $35,000 and <= $55,000	0.098451	0.236587	1.103460	10.35%
Family income > $55,000 and <= $75,000	0.259390	0.003477	1.296139	29.61%
Family income > $75,000 and <= $95,000	0.355227	0.000255	1.426504	42.65%
Family income > $95,000 and <= $115,000	0.278736	0.007443	1.321458	32.15%
Family income > $115,000 and <= $135,000	0.358183	0.001980	1.430727	43.07%
Family income > $135,000 and <= $155,000	0.376441	0.002995	1.457090	45.71%
Family income > $155,000 and <=$175,000	0.314410	0.040093	1.369451	36.95%
Family income > $175,000 and <= $195,000	0.454434	0.014814	1.575282	57.53%
Family income > $195,000 and <= $215,000	0.455066	0.007932	1.576277	57.63%
Family income > $215,000 and <= $235,000	0.974023	0.000523	2.648578	164.86%
Family income > $235,000	0.597220	0.000016	1.817060	81.71%
High school diploma or GED	0.117394	0.195703	1.124562	12.46%
Associate's degree	0.140578	0.173042	1.150939	15.09%
Bachelor's degree	0.573725	0.000000	1.774866	77.49%
Master's degree	0.799899	0.000000	2.225316	122.53%
Ph.D./M.D/Law/other high lvl prof degree	0.670087	0.000038	1.954407	95.44%
X1 Socio-economic status composite	0.446987	0.000000	1.563594	56.36%

Table 6

The pseudo R^2 statistic of 0.1281 may seem modest, explaining only 13% of the variance in the dependent variable, but pseudo R^2 is generally lower than normal R^2, especially in social sciences like education research. Pseudo R^2 scores greater than 0.1 is generally thought of as moderate performance in social sciences. Moreover, the LLP p-value equal to 0 indicates that this model is significantly better than the null model which contains only an intercept. That result allows me to reject the null hypothesis and say that socioeconomic factors significantly affect a students' ability to succeed academically.

All race variables have statistically significant differences from their baseline (white students) in terms of GPA, indicated by the p-values < 0.05. Asian students generally perform better academically; their coefficient of 0.874024 indicates that the odds of having a GPA greater than 3.0 increases 140% compared to white students, holding all else constant. In contrast, students of all the other ethnic backgrounds have lower odds of having a GPA of 3.0 or higher: African American students are about 60% less likely to have a GPA of 3.0 of higher, Native American students' chances decrease about 63%, Hispanic (no race specified) students are about 43% less likely, and so on. These coefficients indicate that race has a significant effect on GPA.

Looking at family income, the first two categories (income between $15,000 and $55,000) have statistically insignificant variables, indicating that their odds having a GPA greater than 3.0 are not significantly different from baseline (income less than $15,000). According to the Office of the Assistant Secretary for Planning and Evaluation, the poverty line in the United States for families of 4 is $32,150 per year, so students living around the poverty line tend to have similar academic outcomes. However, every income range above $55,000 has statistically significant coefficients, indicating improved odds of achieving a GPA of 3.0 or higher. Income groups between $55,000 and $175,000, holding all else constant, have about a 30-45% better chance of achieving a higher GPA than the baseline less than $15,000. Income between $155,000 and $175,000 has a p-value of 0.04 (close to insignificant), so there seems to be some more nuance within that group, but it also one of the smaller groups making up about 2.22% of the sample, so there may just be unforeseen variance in that smaller category. Then there is a jump for students who families make between $175,000 and $215,000 a year, having a nearly 60% better chance to achieve a GPA of 3.0. Then there is a small group between $215,000 and $235,000 that have an increased chance by 164%, and anything over $235,000 a year coincides with an 80% better chance. Simply, higher socioeconomic status has a strong association with academic success, and it is hierarchical. The extremely wealthy students have better chances to achieve GPAs over 3.0 than moderately wealthy students.

Highest parent education level also has a significant impact on GPA. Like family income, the first two education levels do not differ significantly from baseline (less than high school education). But once a parent achieves a bachelor's degree, or greater, their student has a statistically significant better chance of achieving a GPA of 3.0 or higher. If a student's parent has a bachelor's degree, their odds of achieving a 3.0 increase by 77%, holding all else constant. If they have a master's degree, the chance increases by 122% and anything higher degree increases by 95%. These coefficients demonstrate that a parent's education level has significant impact on a student's chances of academic success.

Lastly, socioeconomic composite score (the only continuous indicator in the model) is also statistically significant. This composite score is a calculation taking in both parents' education, occupation, and income. So while we have some of that information in the model separately, socioeconomic score quantifies the interaction of all those elements and brings a bit more parental information as well. The other variables also do not consider the parent that handles more of the childcare; their education level can have a major impact on their child's academic abilities. The coefficient of 0.446987 indicates that for every one unit increase in socioeconomic composite score, the odds of achieving a GPA of 3.0 of higher increases by about 56%. The range of this score is from -1.7526 to 2.8807, so a one-unit increase of one is substantial, but this coefficient still exhibits the positive relationship between academics and socioeconomic status.

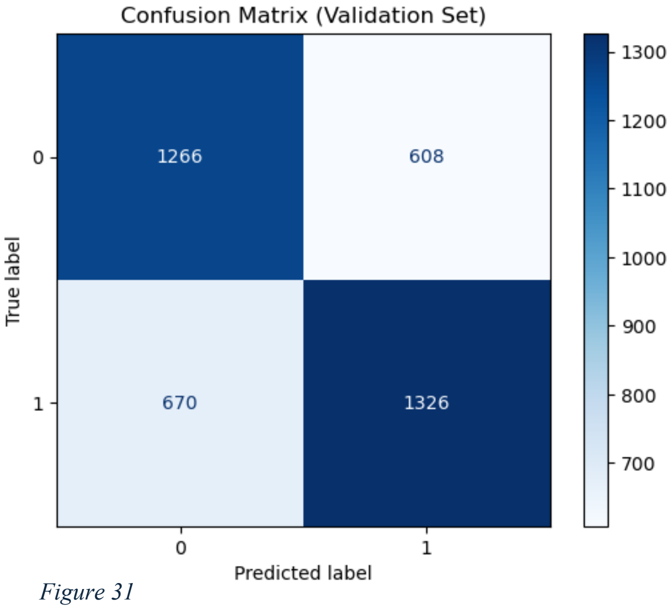


Figure 31

For validation, I performed a train-test split on the data, using the stratify argument to ensure a representativeness, and refit the model on the training set. This model had a similar pseudo R^2 of 0.1271, the same LLR p-value of 0, and very similar coefficients compared to the previous model (Table 7 and 8). The confusion matrix (Figure 31) shows that, while not highly accurate, the model predicts considerably better than random chance. The model scored an accuracy of 0.67, precision of 0.686, recall of 0.664, specificity of 0.676, and F1 of 0.675. Based off those scores, the model is balanced in predicting both positive and negative results, suggesting that the model captured genuine trends from the data.

Lastly, I regressed the rest of the success metrics (math theta score, school sense of belonging, and school engagement) on the same independent variables. Success was considered anything higher than the mean of the category.

The model for math theta score is very similar to the GPA model, which makes sense based on cluster analysis, their strong correlation, and domain knowledge (good math students are often good at other subjects). This models pseudo R^2 of 0.1139 and LLR p-value of 0 is nearly the same as the GPA model (allowing rejection of the null hypothesis). One main difference from this math model and the GPA model was that people who identified as biracial and pacific islander had statistically insignificant variation from the white baseline. People who identify as pacific islander are a very small group in the survey, but people who identify as biracial did not (making up nearly 9% of the survey). The results suggest that testing scores in math are a bit less affected by demographics than GPA. The math scores variable comes from a standardized test administered by the NCES, so perhaps that standard testing are more agnostic to socioeconomic differences than GPA. For family income, the increased odds of success are still present, but more regularized once an income threshold is hit. Income over $35,000 has a statistically significant p-value that indicates an increase in odds of a above mean math score of 20%. Then everything above $55,000, the increase in odds sways between 40%-75%, except incomes between $215,000 and $235,000, who has an increase in odds of 96%. For parent education all variables were significant, but the more education one has, the higher the increase in odds is for their child's math score. Ph.D.'s and other higher degrees holder's students had the largest increase in offs of 132.93%. Socioeconomic score is also statistically significant and for every one-unit increase, the odds of having an above average math increases by about 49%. These results support the idea that socioeconomic and demographic factors impact academic success.

The non-academic indicators of success tell a very different story than their academic counterparts. The first indication of that was a serious drop in pseudo R^2 to 0.016, explaining only about 1.6% of the variation in a student's sense of belonging. The LLR p-value is still close to 0 and significant but no longer displayed as simply 0; it is significantly better than the null model, but the coefficients tell more of the story. For sense of belonging, only those who identify as Asian and African American has statically significant statistics, but their p-values are 0.01269 and 0.007129 respectively, so they are much closer to insignificance than in previous models. Both of those groups also have increased chance of an above average sense of belonging compared to the white baseline (both around a 17% increased chance). Family income statistics are now mostly insignificant, except for those between $135,000 and $175,000 per year and those over $235,000 per year. Parent education also has mostly insignificant p-values except for parents with bachelor's and master's degrees. Socioeconomic score still has a statistically significant p-value, but it is not significant at the 0.01 level anymore (p-value equal to 0.0375) and the increase in odds is only about 10% now. The drastic change in results and coefficients from the academic models indicate that socioeconomics does not impact sense of belonging the same straight forward way that they affect academic success.

The model for school engagement, like the student sense of belonging model, has a very low R^2 score of 0.015. The LLR p-value is also still close to 0 and significant but no longer displayed as simply 0. Race variables are mostly insignificant, with only people who identify as Native American, Asian, African American, and Hispanic (race specified) having significant coefficients. Asian people have a p-value of 0 and a 50% increase in odds of above average school engagement. Native American students have a decrease in odds of 46%, but the rest of the demographic groups have relatively small odds differences, are insignificant, or both. For family income, the only significant statistic is students with family incomes between $15,000 and $35,000 per year which have a 20% less chance of being engaged in school than the baseline of family income under $15,000 per year. Parent education metrics are all significant at the 0.05 level, but they are also all insignificant at the 0.01 level. The odds of above average engagement increase around 20-30% once a parent has at least a high school education. The change in odds for this group are also small compared to the academic models. Socioeconomic status has an insignificant p-value and a 9% increase in odds. These results show that socioeconomic and demographic factors do not affect student engagement in the same way as academic success.

8. Make conclusions about the research question based on the analysis results. Do socioeconomic factors affect student success? Are all kinds of success impacted by socioeconomic and demographic factors in the same way (if at all)?

This analysis has yielded two main insights: one, socioeconomic/demographic factors significantly affect a student's academic success, and two, socioeconomic/demographic factors do not affect a student's unorthodox success in the same fashion as the former. The current education system allows for a student's circumstances to impact their academic achievement in school, which sets our society down a path to perpetuate cycles of poverty. That is something that the state government must address.

The evidence for socioeconomic and demographic factors affecting academic success is overwhelming. Through EDA, many visualizations and summary statistics showed the nature of these variables. Figure 22 shows very clearly that GPA and socioeconomic score generally have a positive relationship. Success means grouped by socioeconomic quintile, as well as the grouped mean charts (Figures 23-26) showed that relationship to be true across all success metrics. Then success means grouped by racial group showed that different ethnic backgrounds have generally different outcomes in terms of success. The correlation matrix also showed the strong association between socioeconomic status and academic success (correlations of 0.416 and 0.408 with math theta score and GPA respectively). Cluster analysis also showed the dominant structures in the dataset. The dataset distinctly groups into those of higher socioeconomic status and lower, and you did not even need to use a socioeconomic indicator to show that! Cluster analysis performed with only success metrics was able to group most the lower income families in the cluster with lower success metrics and most higher income families in the cluster with higher success metrics. The four- cluster analysis showed that students with above average socioeconomic status generally had above average academic success, while those with below average socioeconomic status tended to perform below average. The logistic regression also clearly show that socioeconomic and demographic factors affect academic success. Based on the coefficients and significant p-values, except for Asian people, every racial group had a worse chance to achieve a GPA of 3.0 or higher than white students (decreasing odds anywhere from 20-60%). Family income and parent income coefficients also saw students in wealthier and more families having increased odds of having a GPA of 3.0 or higher, far and away. The math theta score model offered very similar results but was generally less extreme with increased odds for better scores than the GPA (but they are still quite large).

The unorthodox success metrics were more complicated than their academic counterparts. EDA and cluster analysis was able to show that there is an association with higher socioeconomic status and one's sense of belonging and engagement at school, based on group means and the plotted group means for each continuous variable. But the four-cluster analysis did indicate the nuance with these metrics. Yes, there are overarching structures based on socioeconomic status (which is why in a large scale, data can be separated into two main clusters), but right underneath that are a collection of students that break the mold. There was a group student with moderately low socioeconomic status and academic performance that were above average in sense of belonging and engagement. Then there was another group that broke the mold in opposite way: above socioeconomic status and academic performance but below average in sense of belonging and engagement. The destruction of the trend was also reflected in the low model performance when predicting these unorthodox success metrics. That is why logistic regression was unable to model these relationships consistently. Money, status, and bias seem to have generally not corrupted these vitally important dispositions towards school and life, which is quite the relief.

So, yes, socioeconomic/demographic factors affect student success, but not all in the same ways and to the same extent. There seem to be some minor impacts on unorthodox success metrics, but overall, affects are not statistically significant (according to logistic regression). Students from lower income families may not have the luxury of engaging in after school activities due to parent work schedules and financial constraints. But many students that may not have access to the educational resources of wealthier students are still engaging with and feeling apart of their school communities. So, while there are things the government could do like offer more in school resources to students and other resources, socioeconomic/demographic factors do not seem to be concrete barriers to school belonging and engagement. In contrast, socioeconomic /demographic factors are barriers to academic success for students. Much of the disparity in GPA is likely connected to district funding. Students in higher income families are going to schools with more funding and better teachers, helping them get better grades and better life outcomes. Funding to schools also probably explains some of the demographic differences; it should not (and cannot) be that non-white students (excluding Asian students) have a significantly worse chance at academic success. That is strong evidence of systemic racism oppressing people through the guise of socioeconomics. It is also possible that students of color are not getting supported in the ways that they need academically. The math theta score model showed that when all is equal (like one a standardized test), the gap between students begins to shrink. The government should attempt to give schools more even funding and offer workshops on how to approach different students that come from different backgrounds. Equal education funding and care for all students may lead to a better tomorrow for all.

Appendix

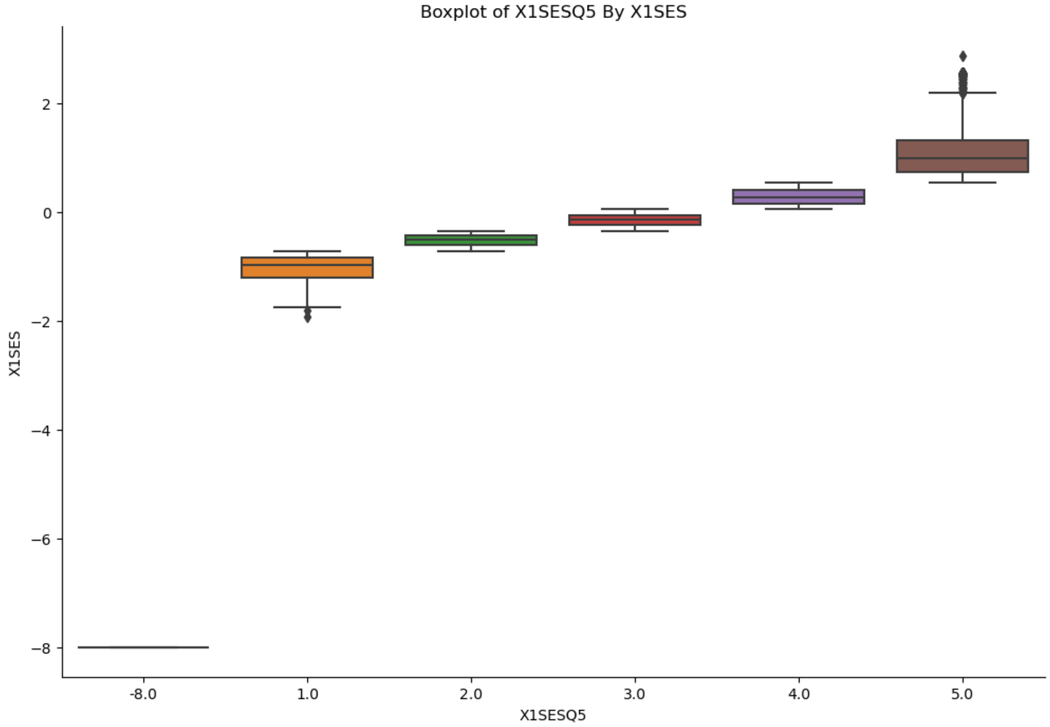


Figure 1

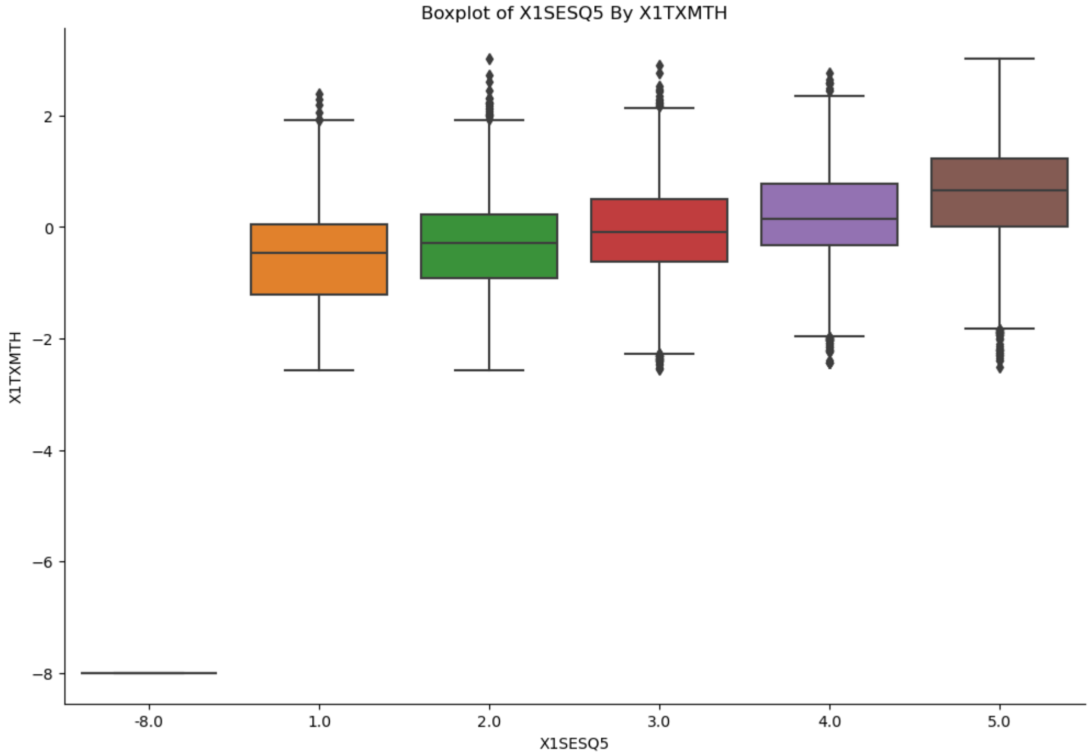


Figure 2

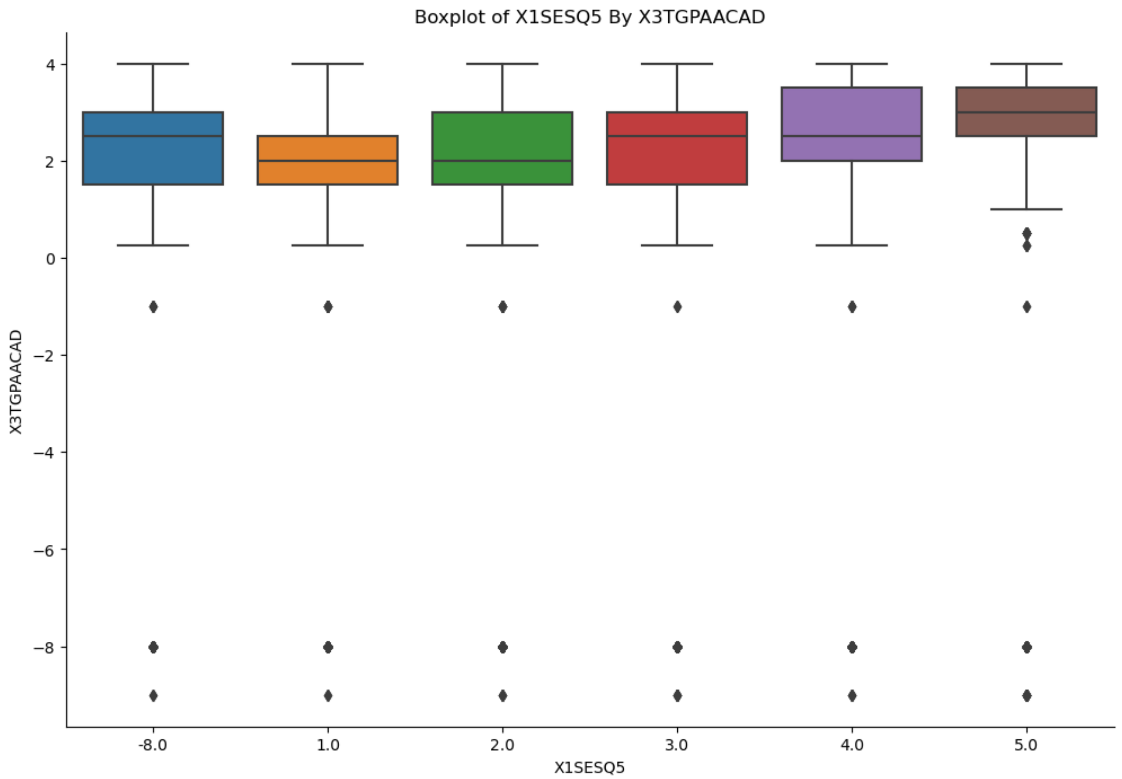


Figure 3

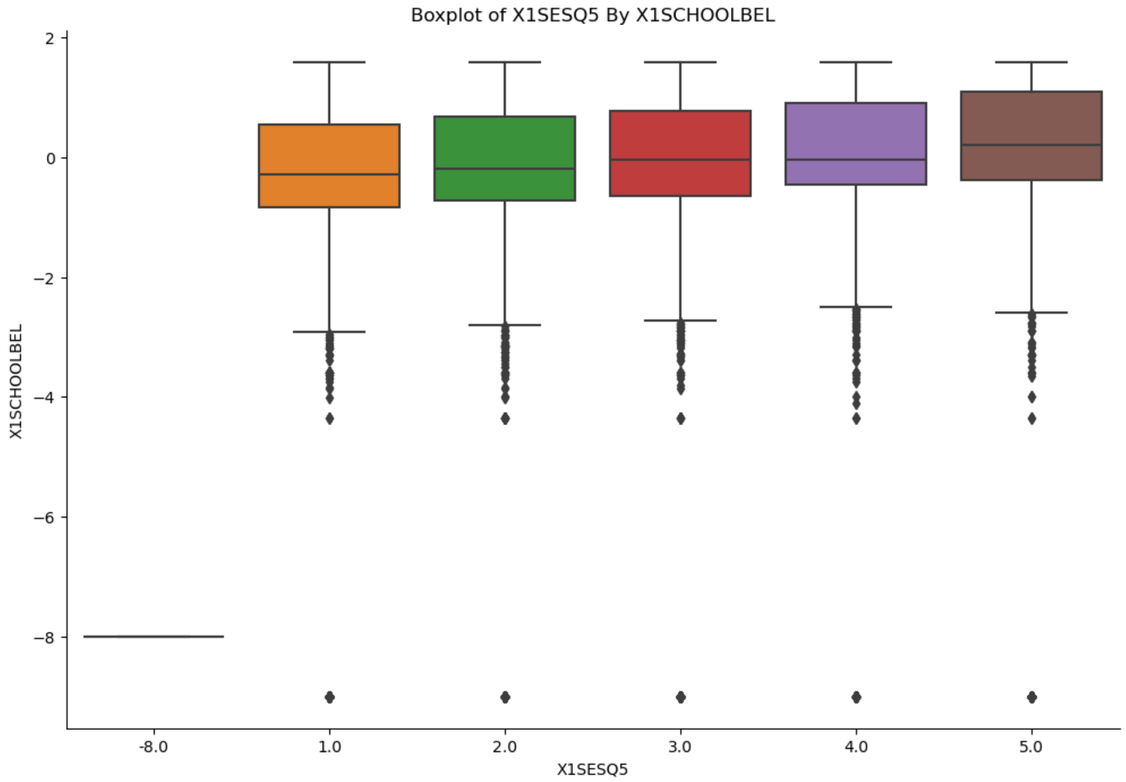


Figure 4

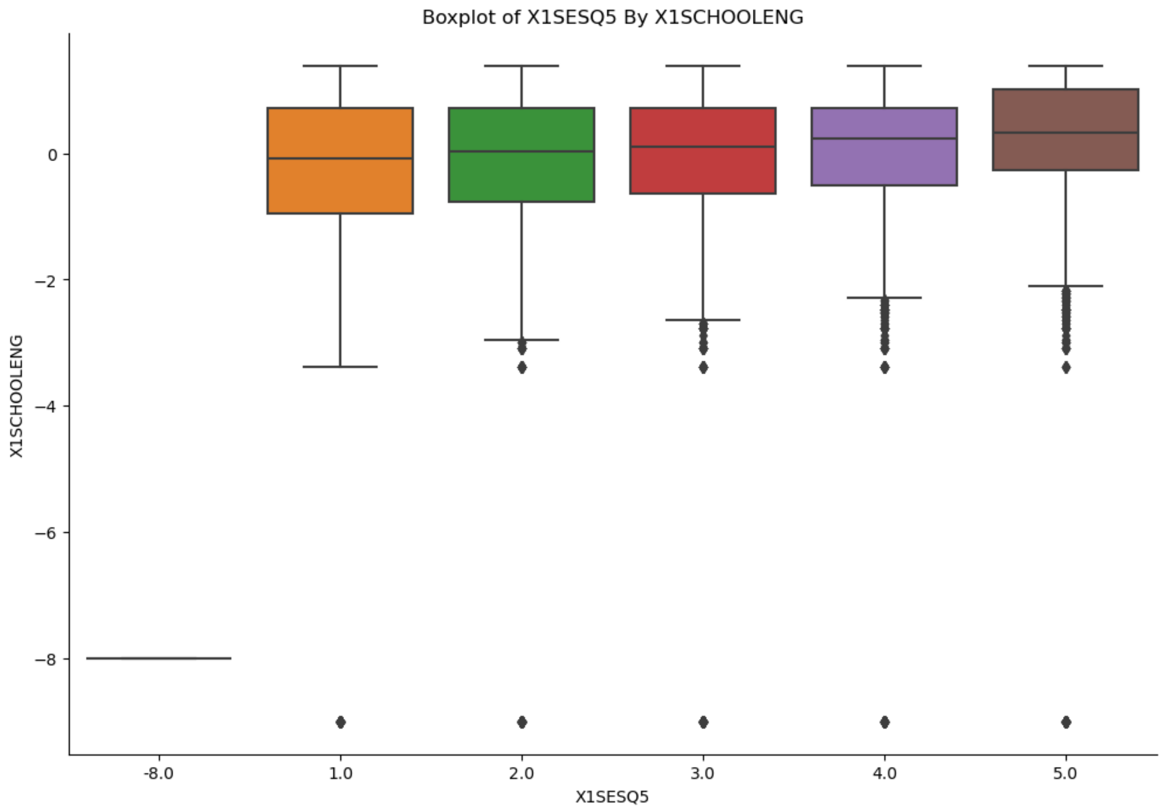


Figure 5

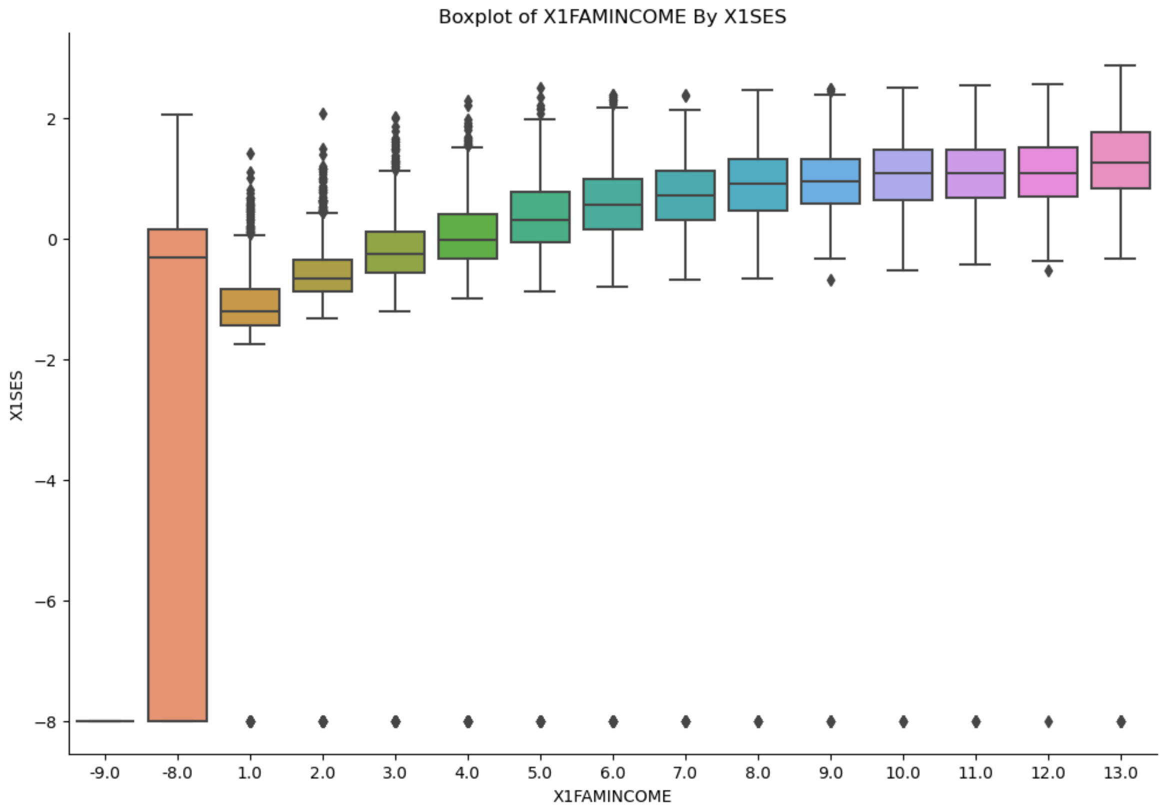


Figure 6

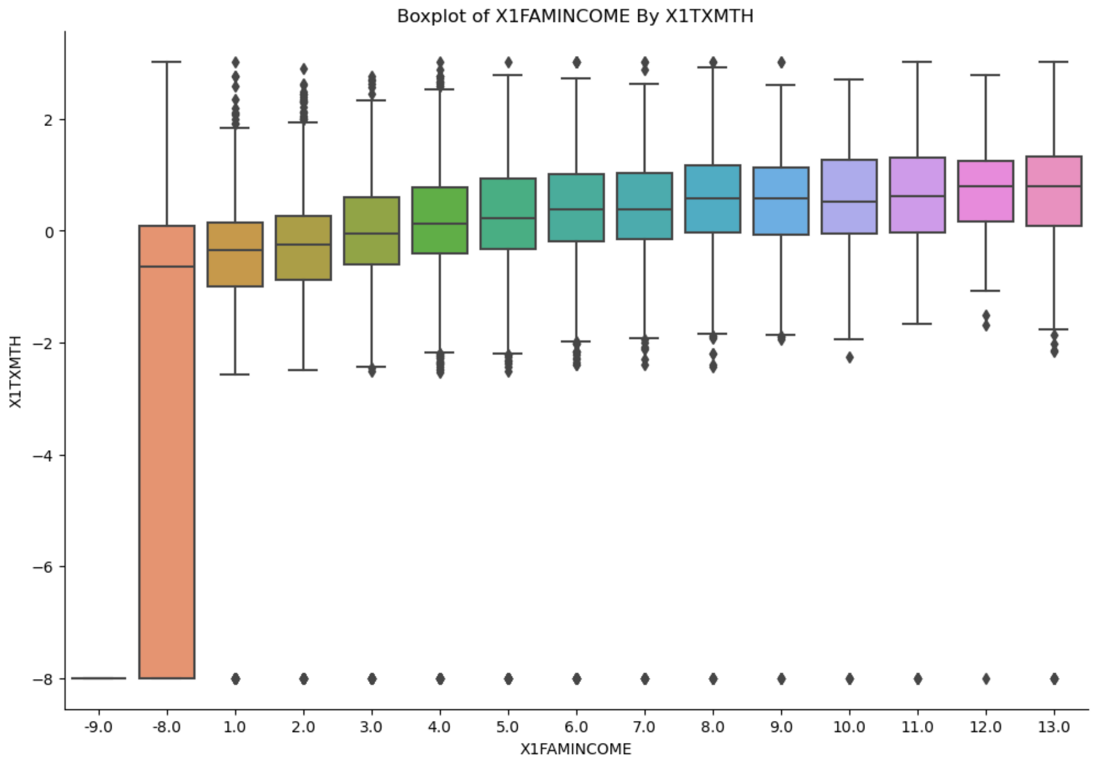


Figure 7

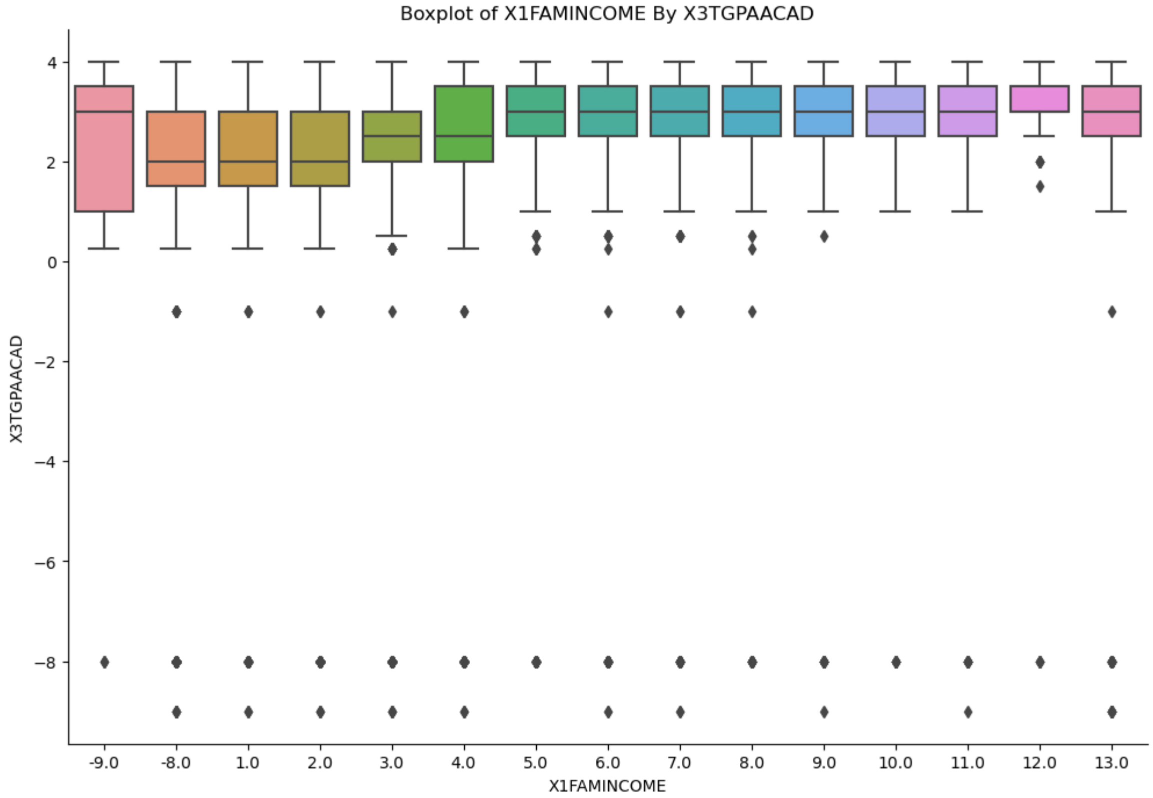


Figure 8

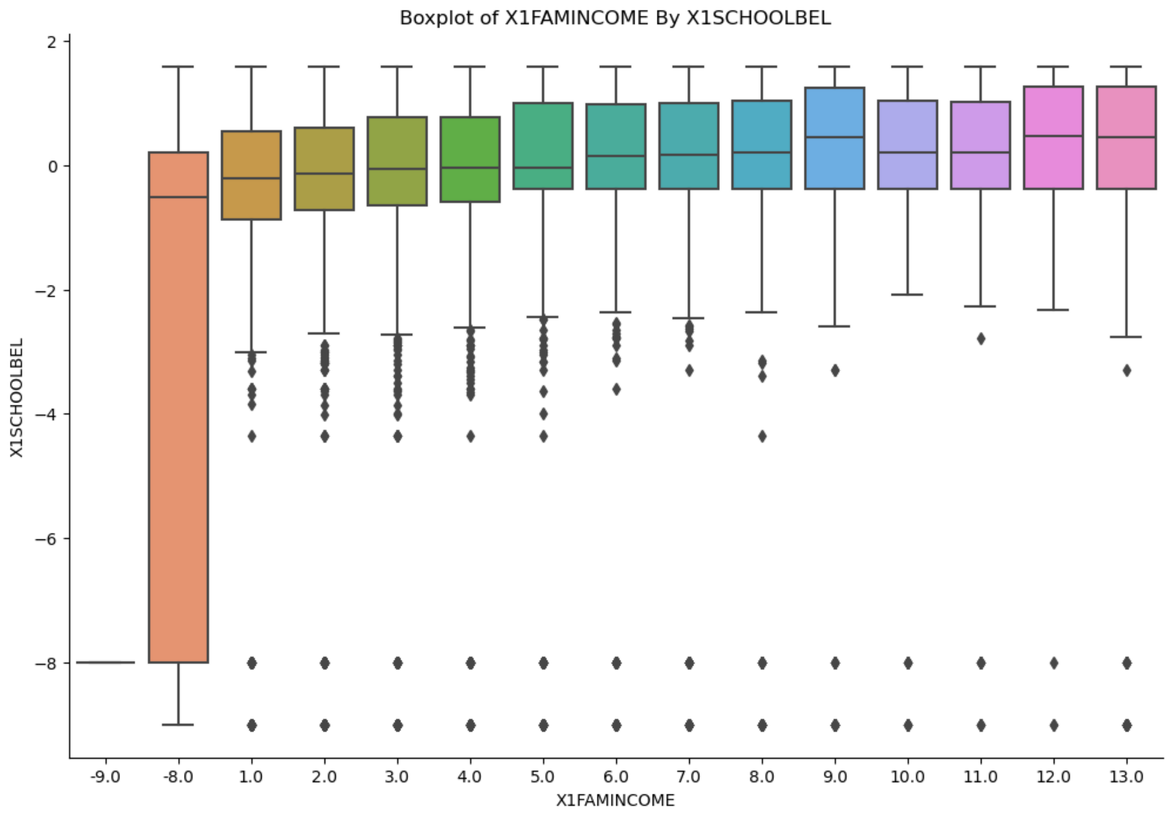


Figure 9

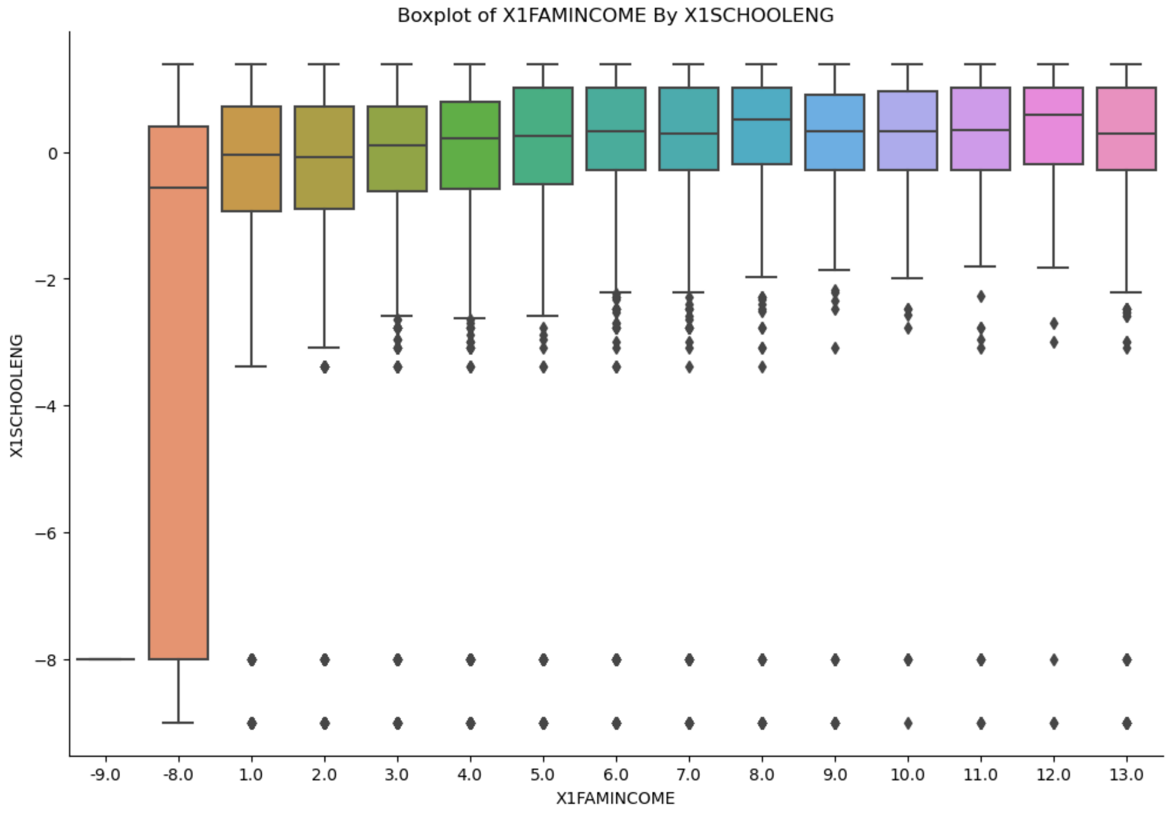


Figure 10

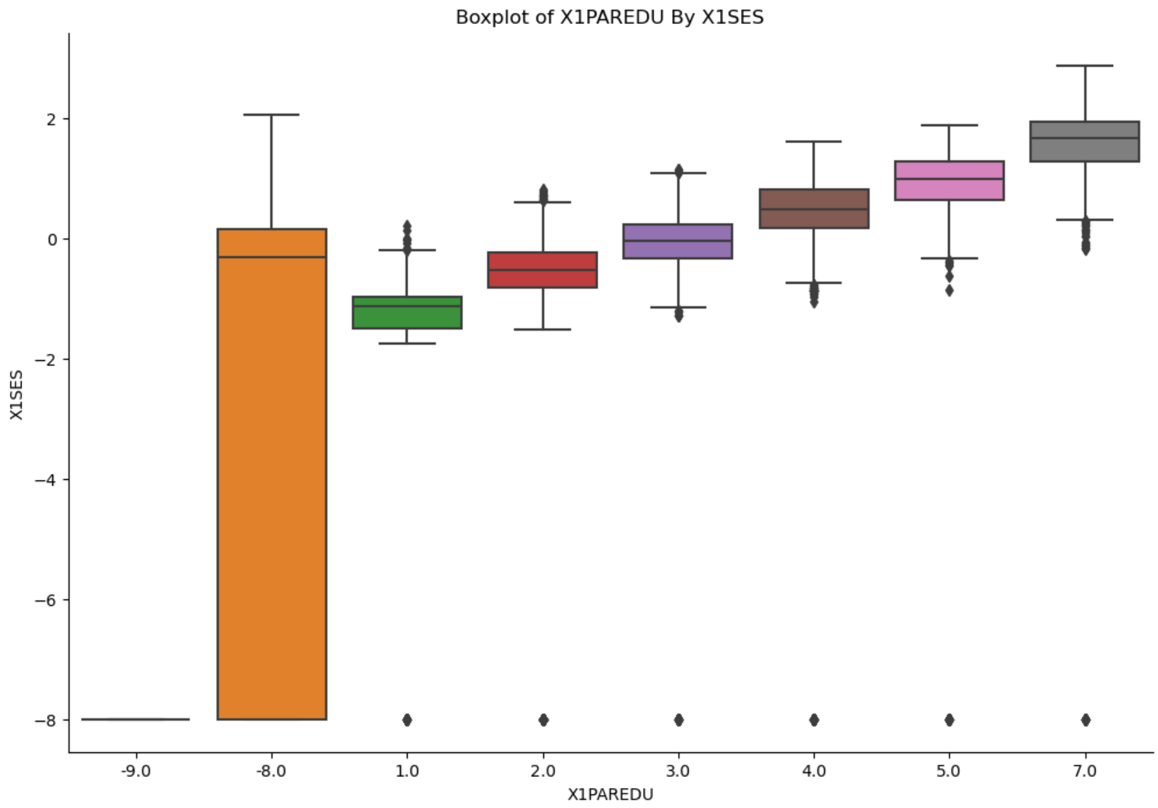


Figure 11

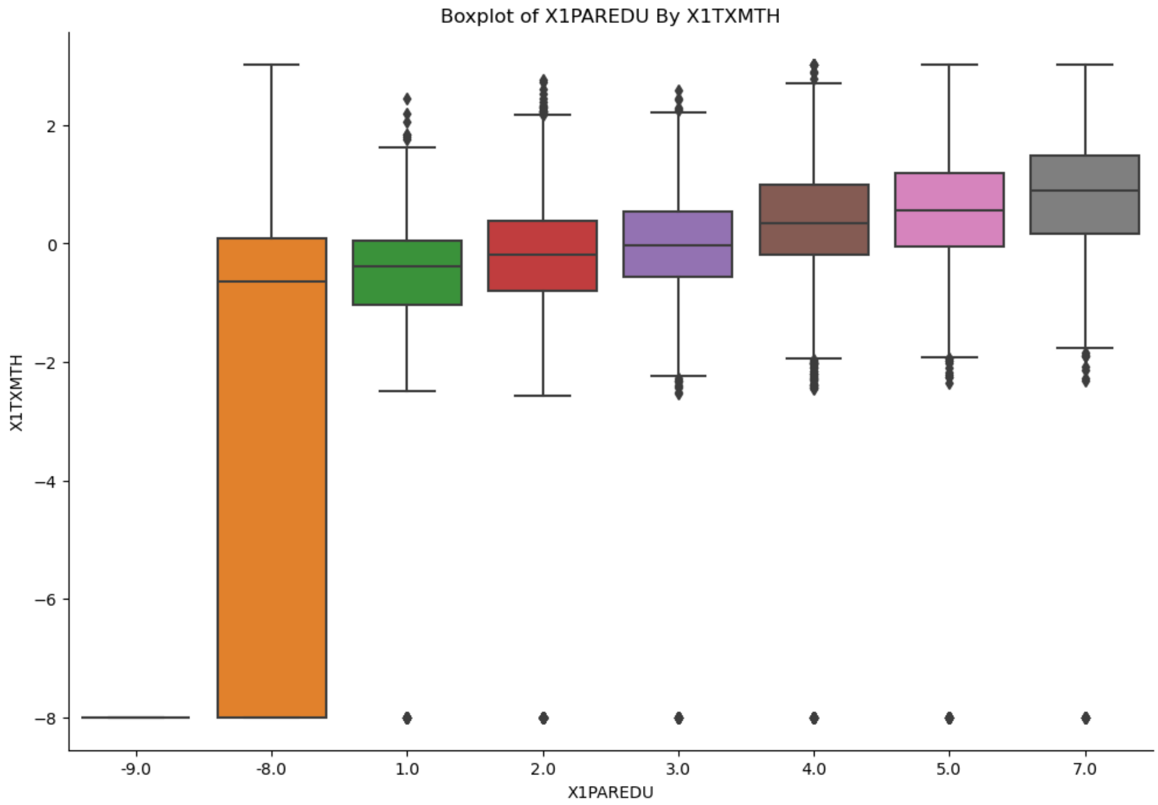


Figure 12

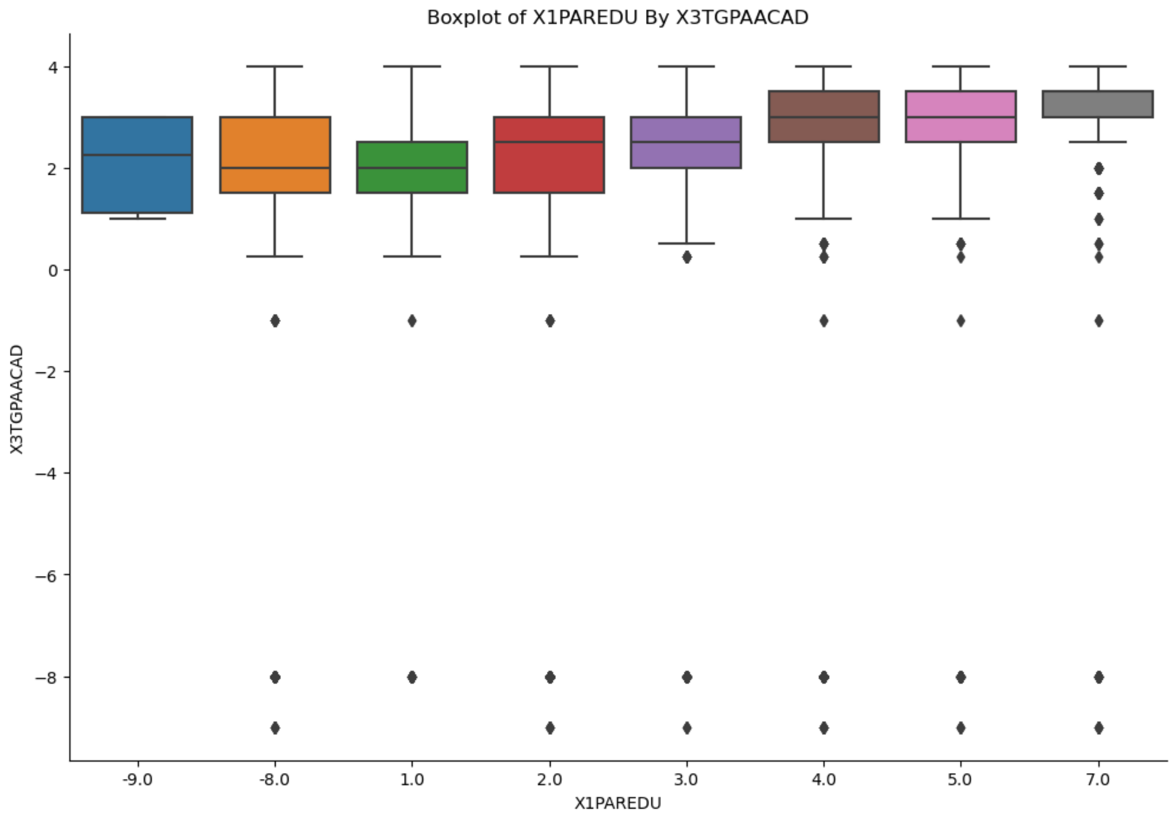


Figure 13

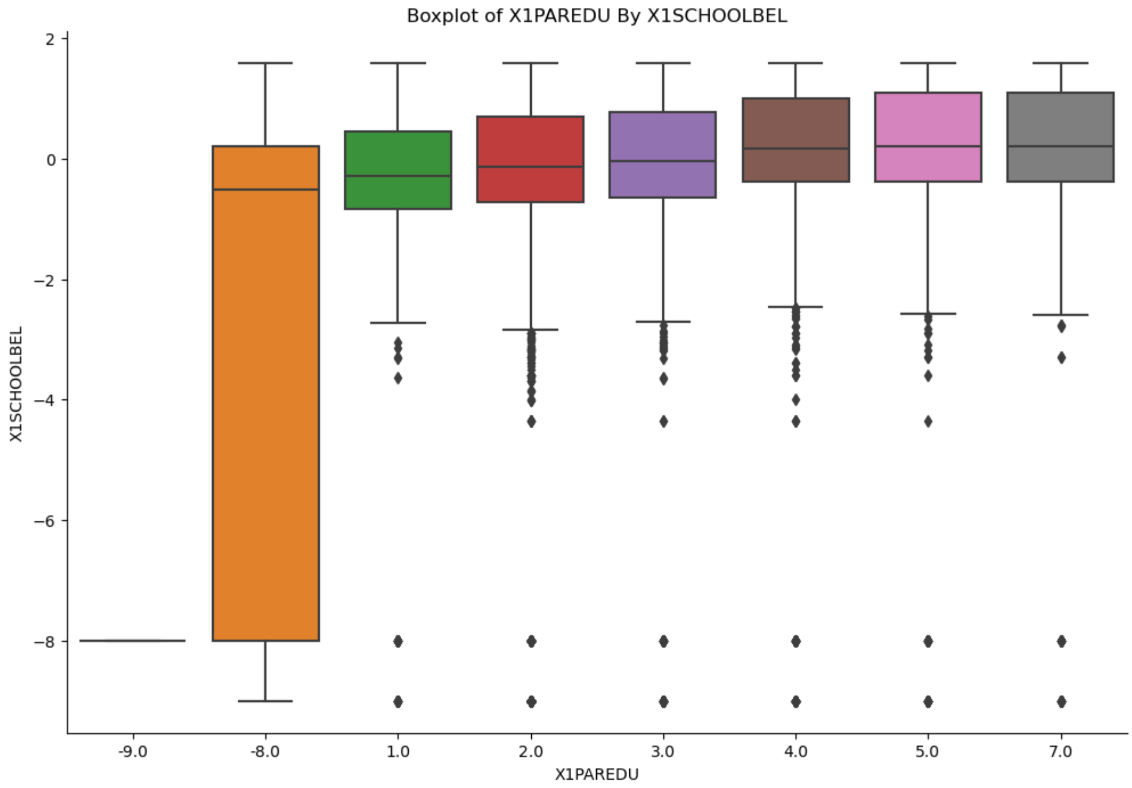


Figure 14

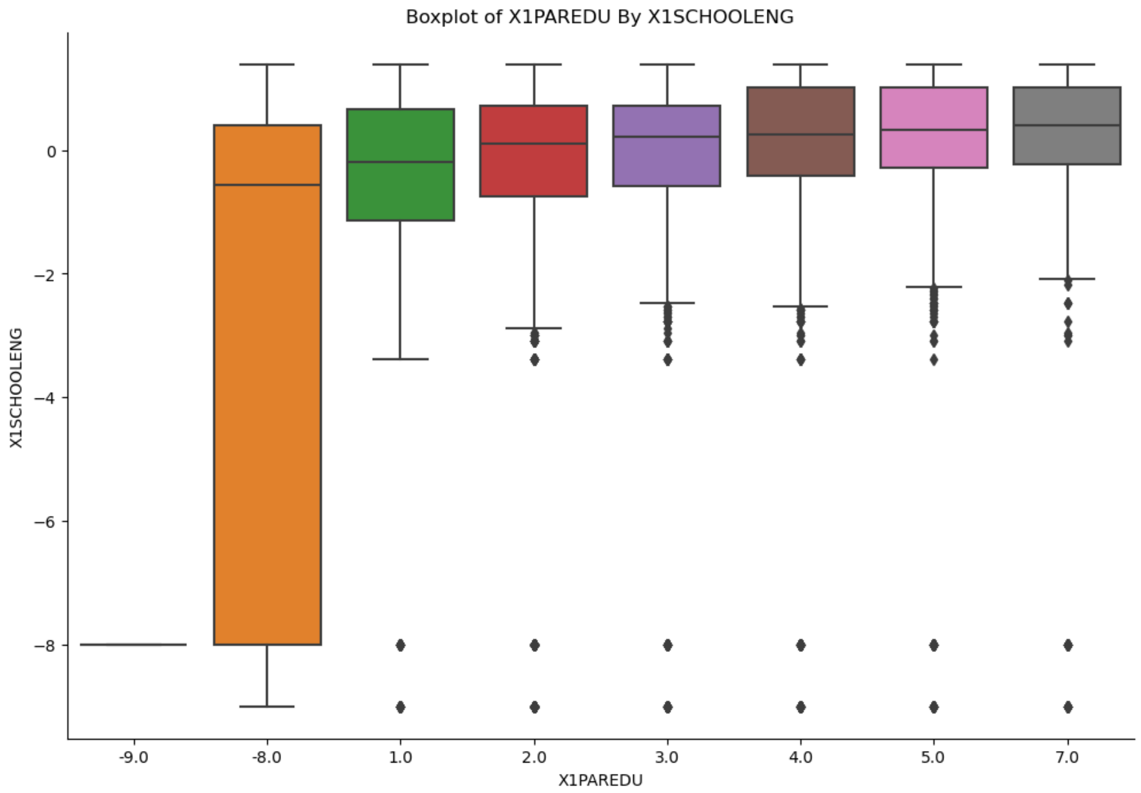


Figure 15

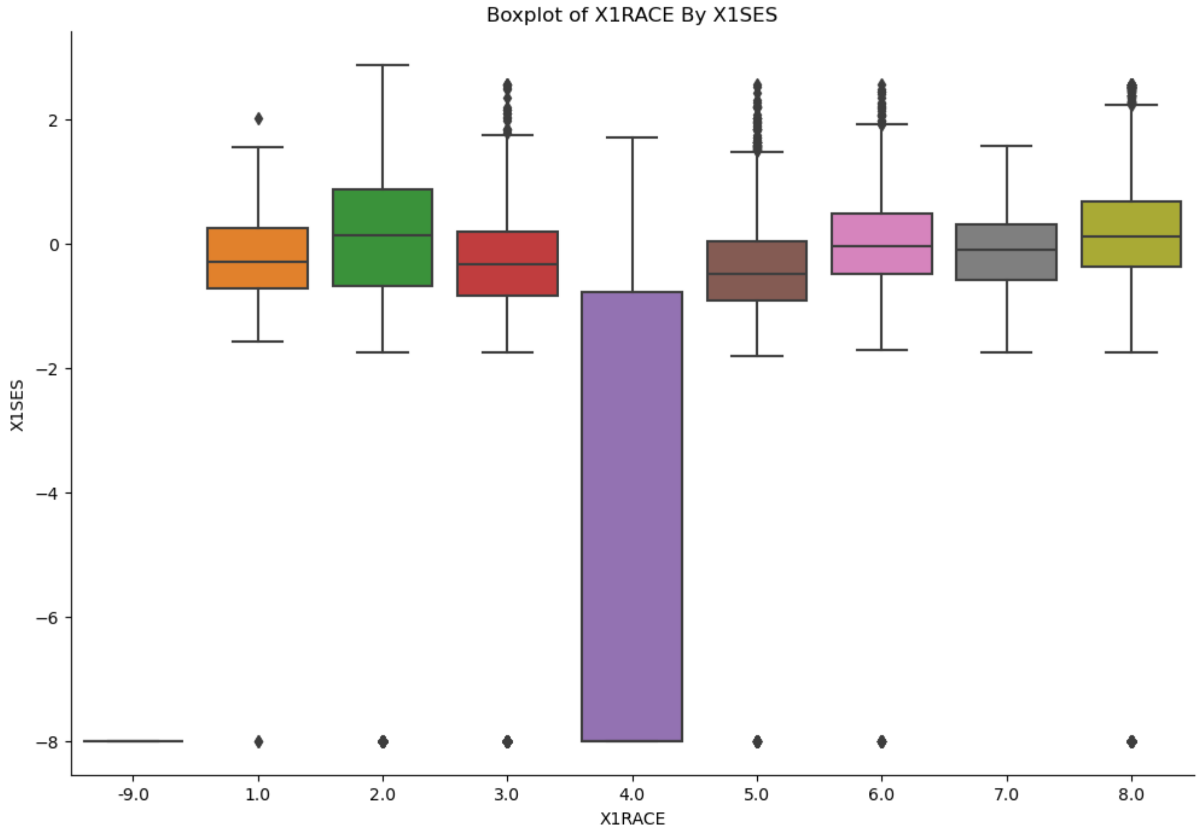


Figure 16

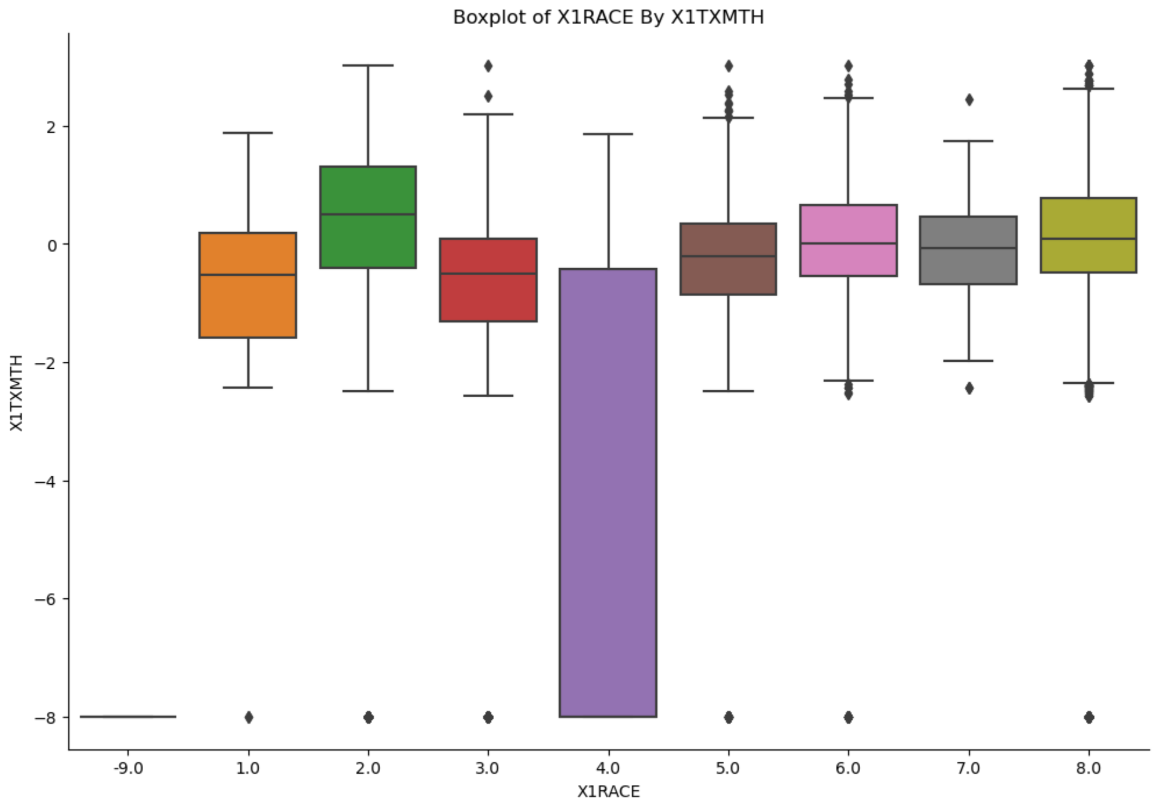


Figure 17

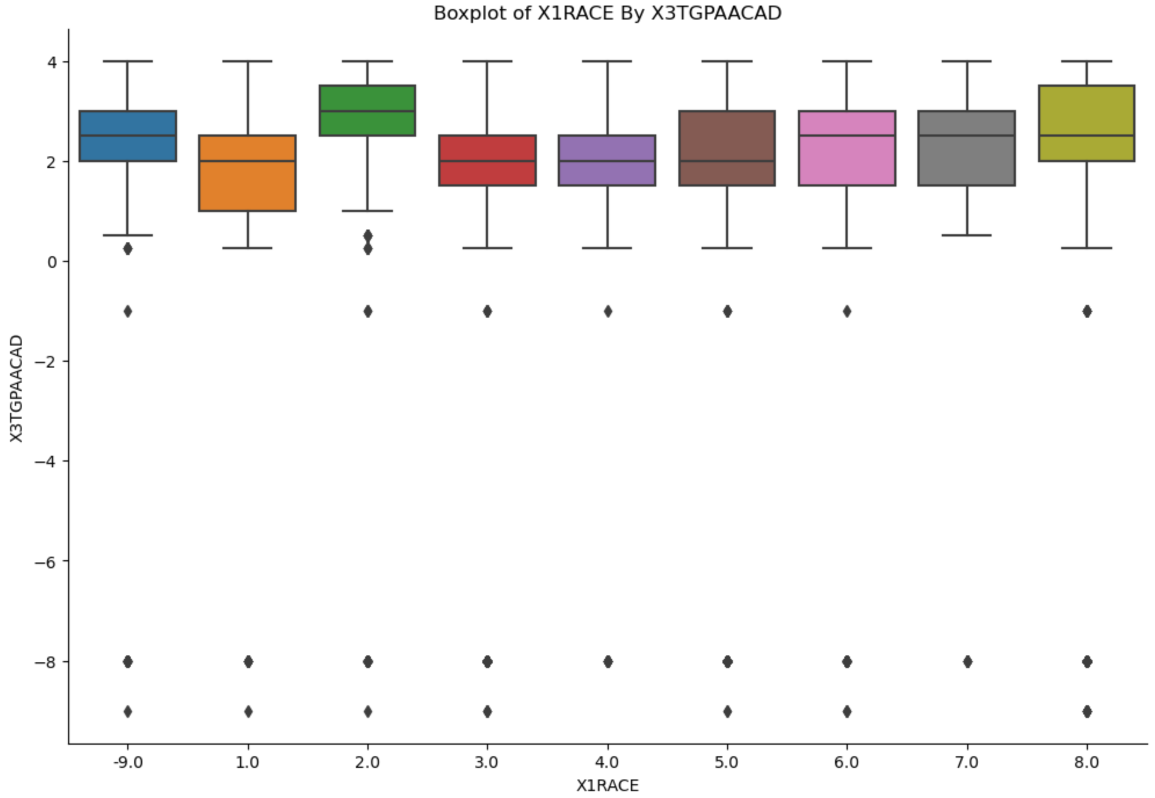


Figure 18

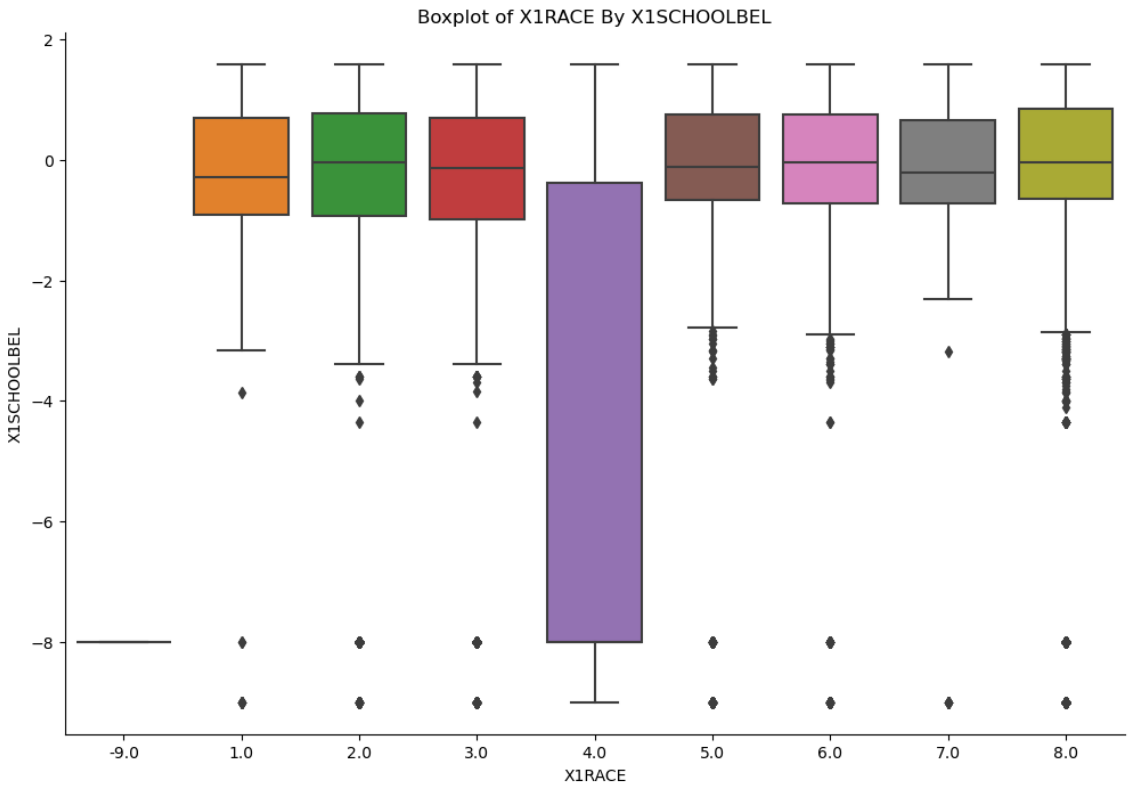


Figure 19

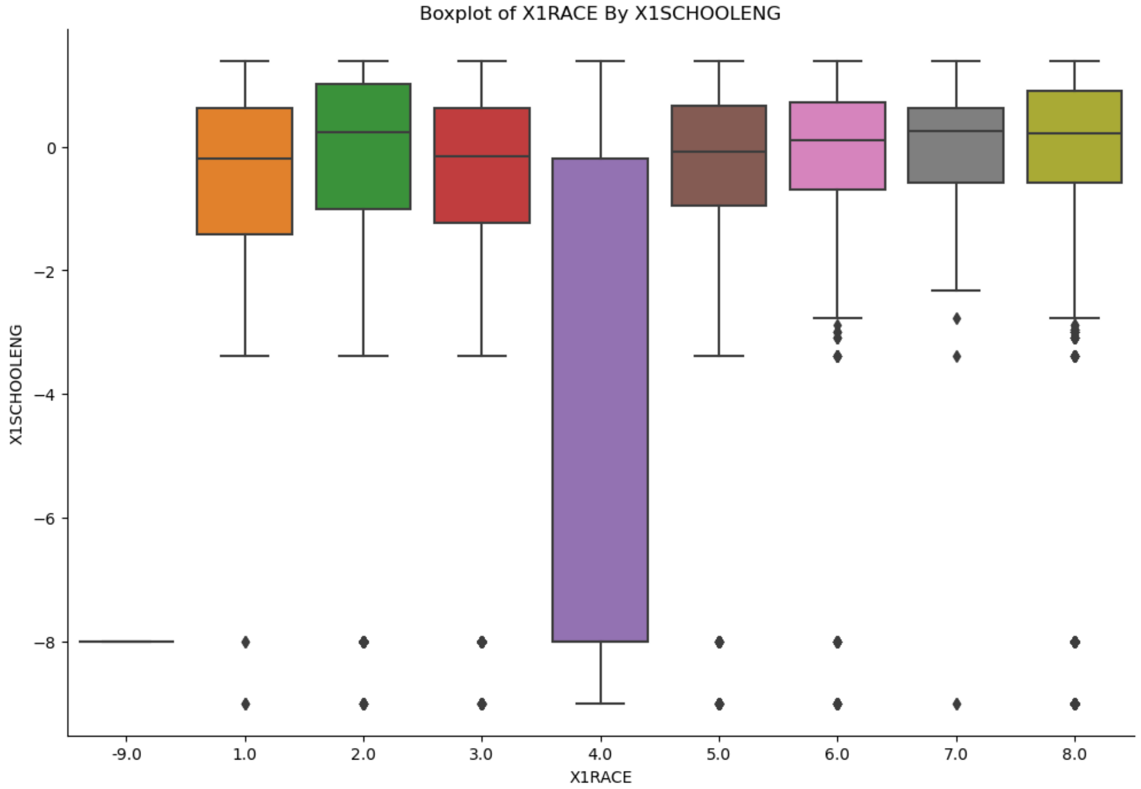


Figure 20

SES Quintile	Math theta	GPA	School Belonging	School Engage
1	-0.3858	2.1352	-0.0777	-0.1055
2	-0.1700	2.4050	-0.0204	0.0122
3	0.0279	2.6026	0.0798	0.0810
4	0.2271	2.7777	0.1783	0.1724
5	0.6270	3.1266	0.3271	0.2777

Table 1

Race	**Math Thea**	**GPA**	**Sense of Belonging**	**School Engagement**
Amer. Indian/Alaska Native	-0.4667	2.1545	-0.063	-0.3655
Asian, non-Hispanic	0.8511	3.1817	0.256	0.3798
Black/African-American, non-Hispanic	-0.3291	2.2432	0.1276	0.0159
Hispanic, no race specified	-0.438	2.0099	-0.073	-0.157
Hispanic, race specified	-0.1445	2.3793	0.0624	-0.0671
More than one race	0.1341	2.5975	0.0224	0.0266
Native Hawaiian/Pacific Islander, non-Hispanic	0.0727	2.4805	0.0401	0.0668
White, non-Hispanic	0.2147	2.7915	0.1506	0.1645

Table 2

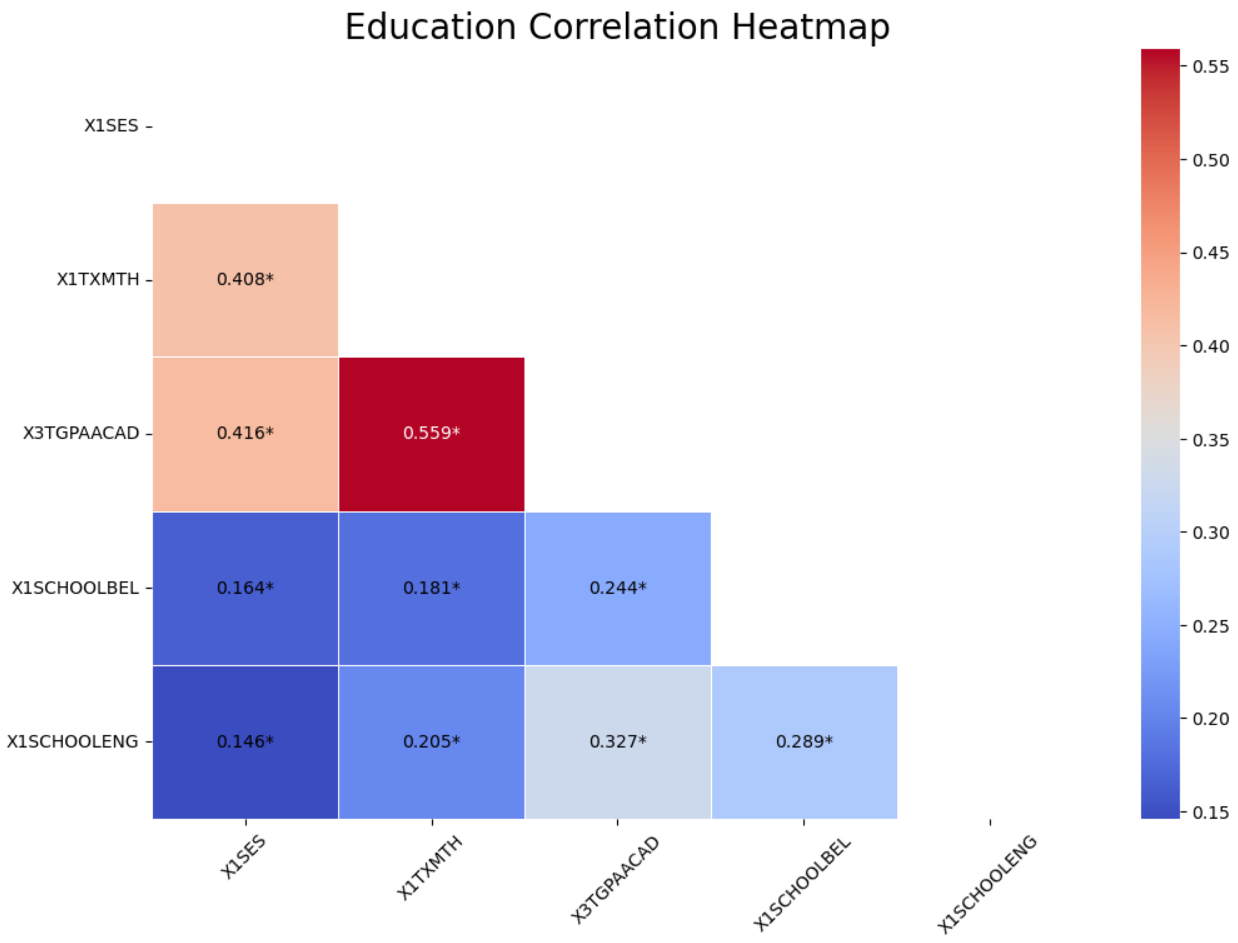


Figure 21

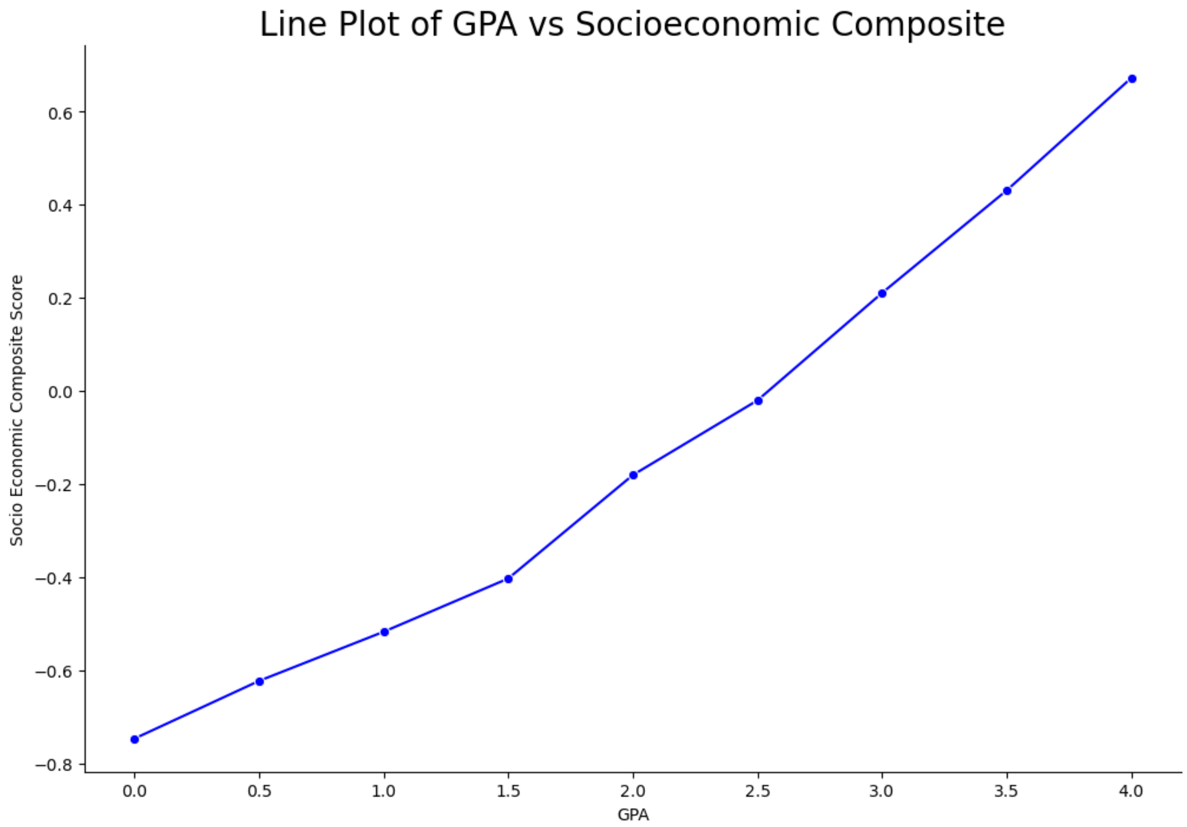


Figure 22

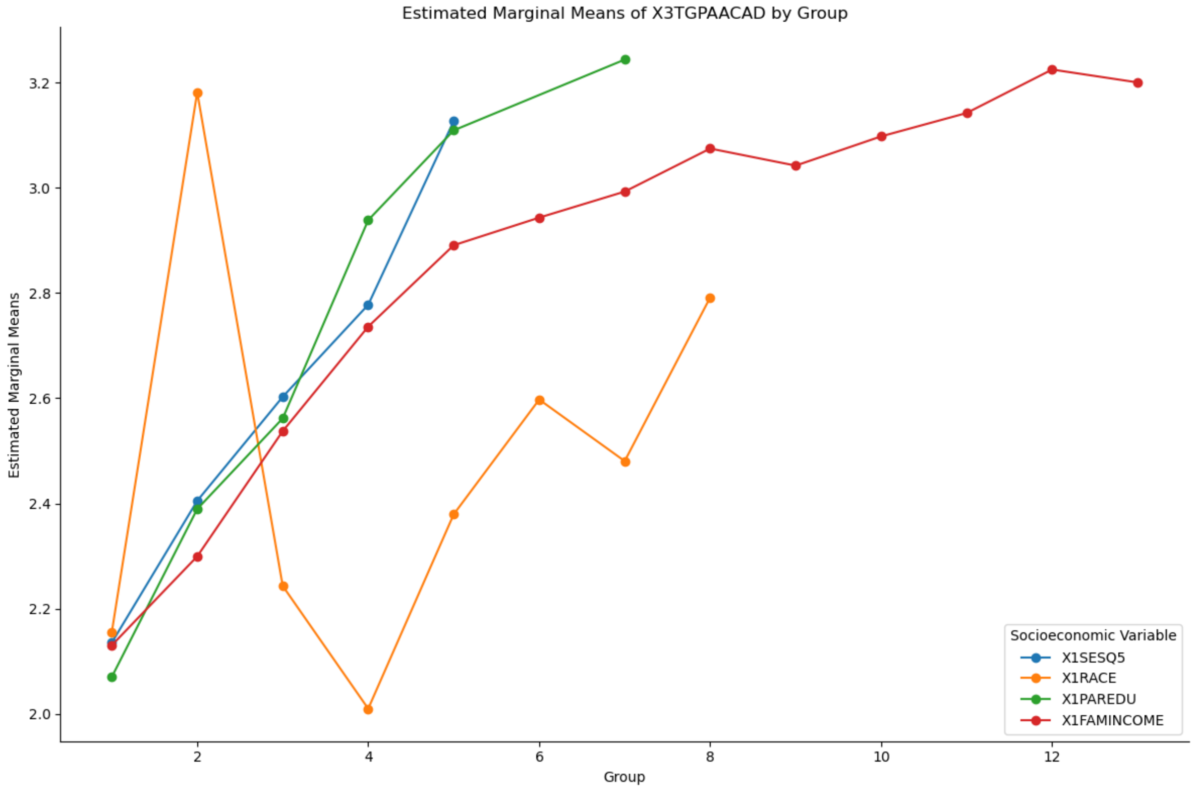


Figure 23

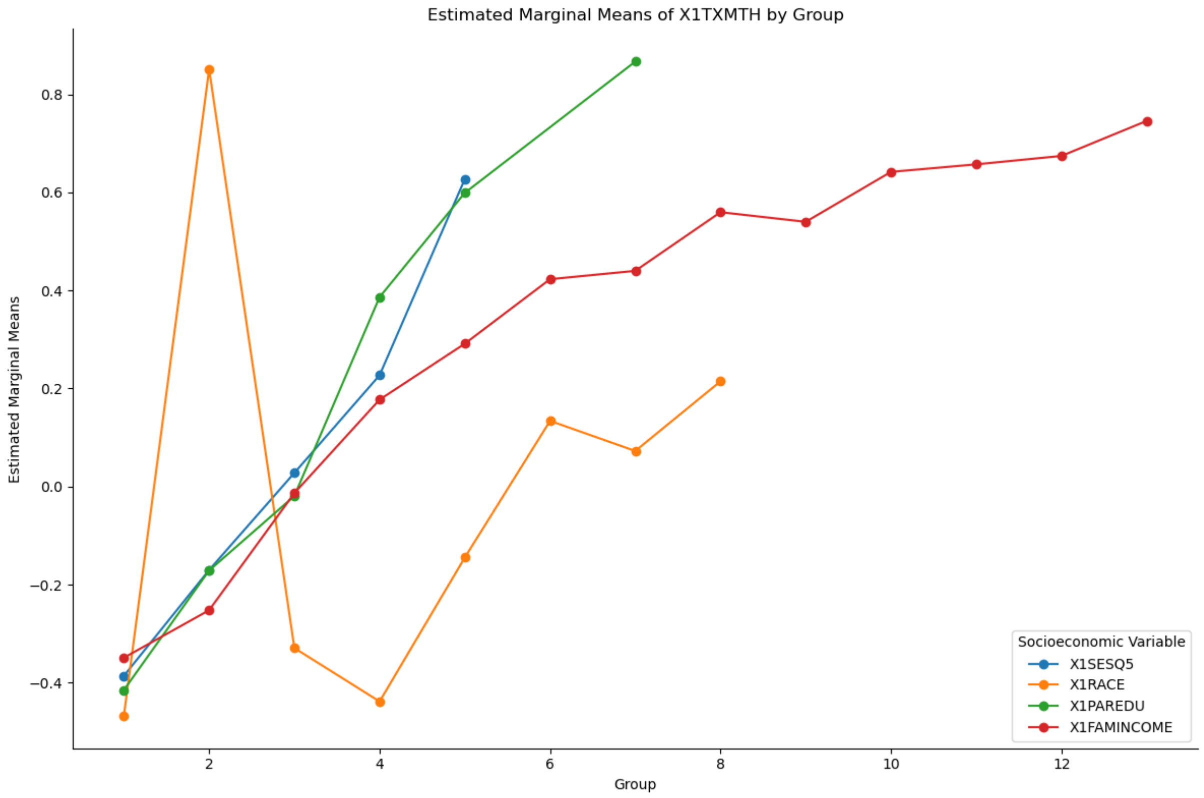


Figure 24

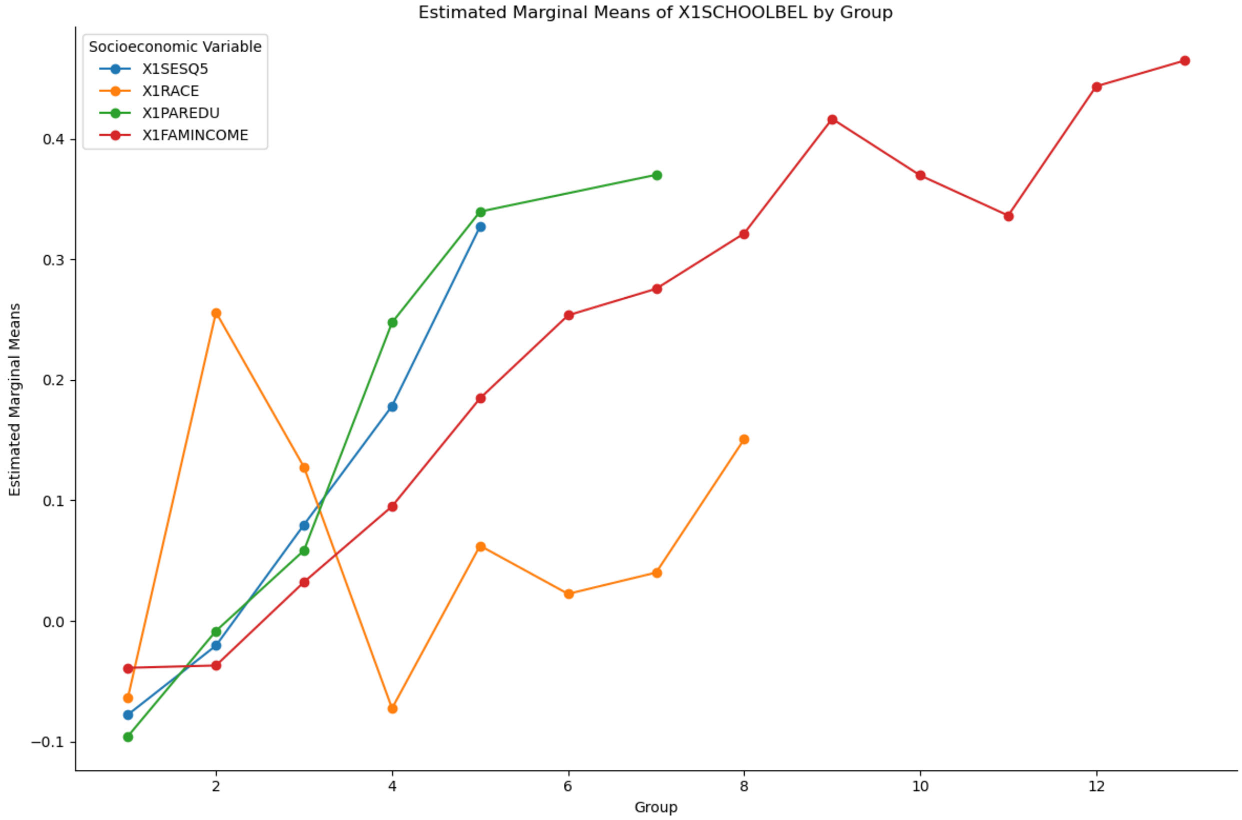


Figure 25

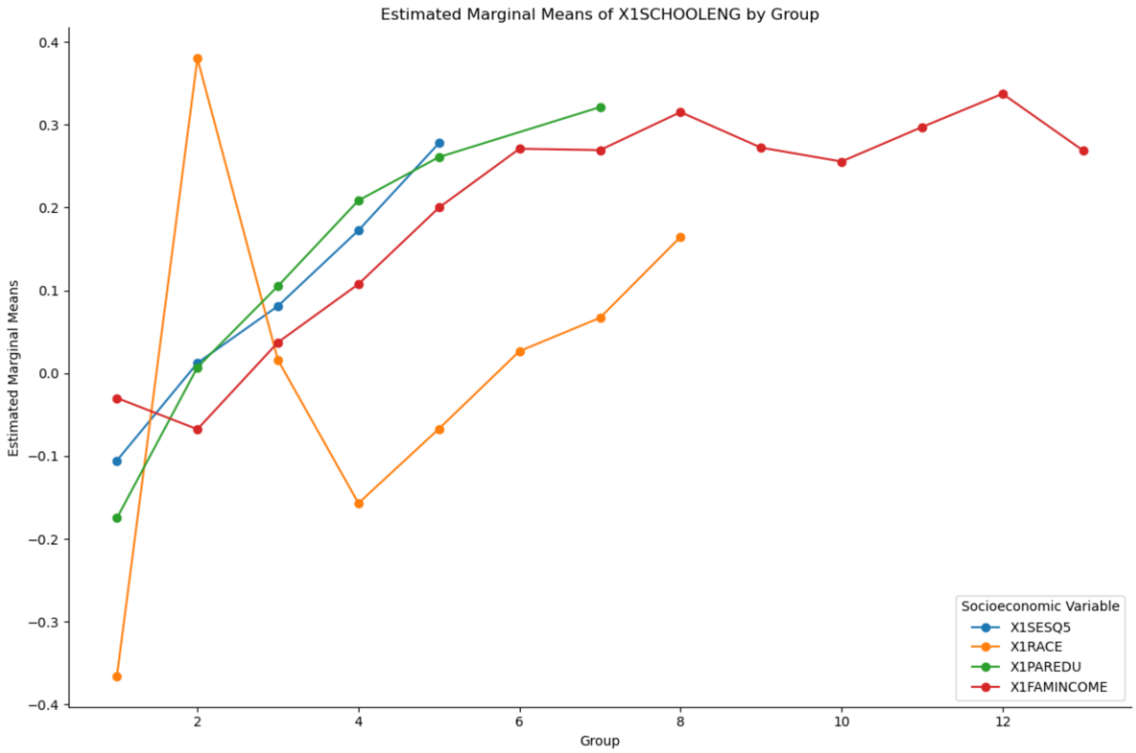


Figure 26

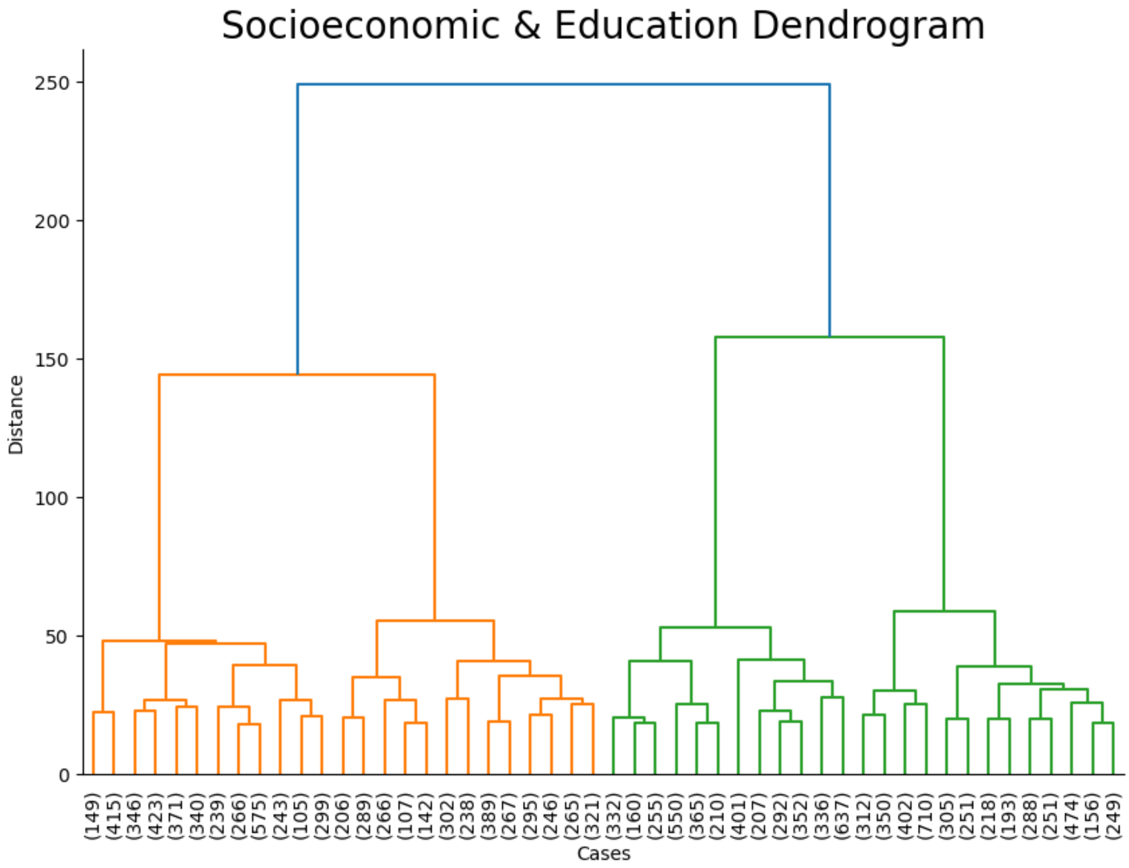


Figure 27

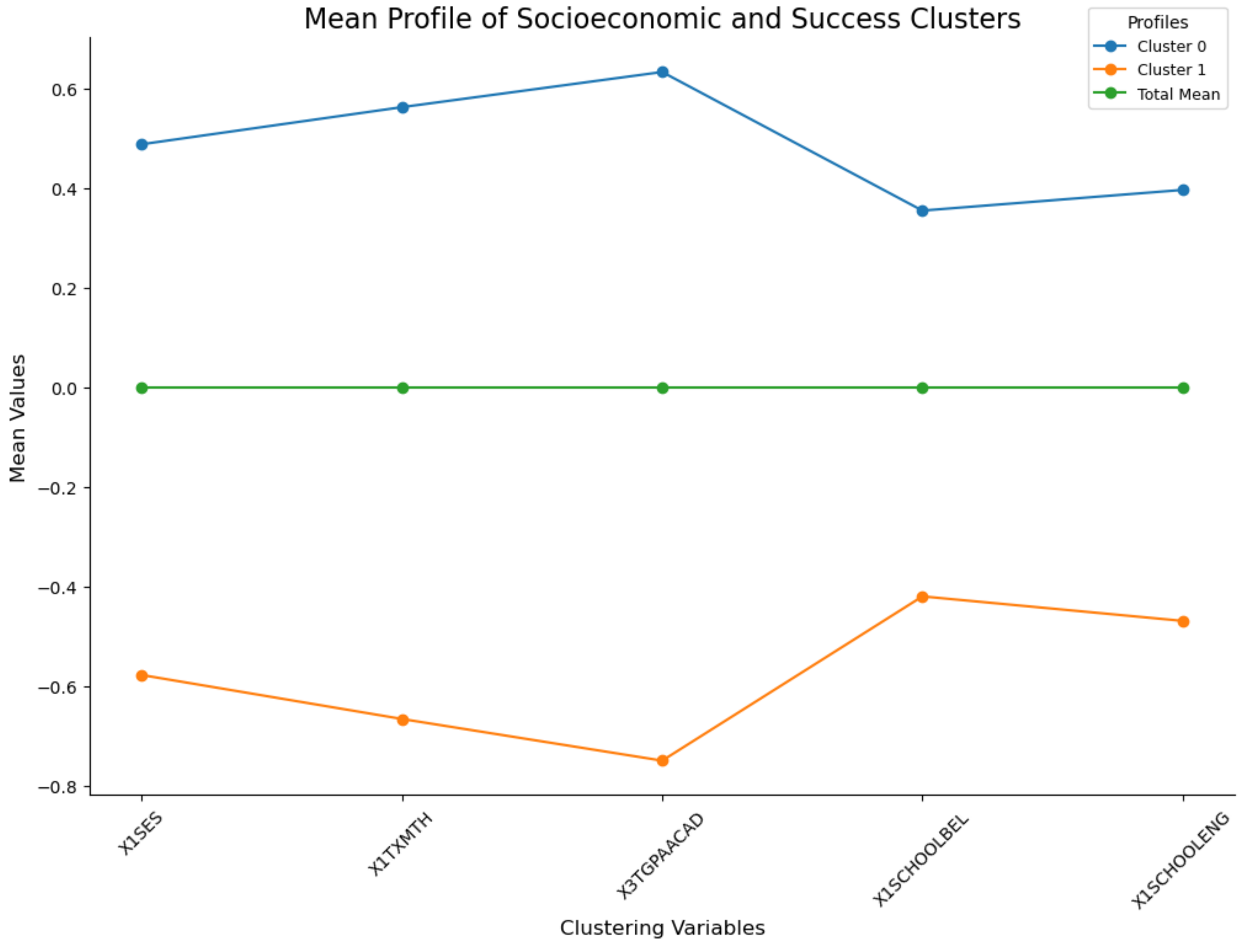

Mean Profile of Socioeconomic and Success Clusters
Profiles
Cluster 0
Cluster 1
Total Mean
Mean Values
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
X1SES
X1TXMTH
X3TGPAACAD
X1SCHOOLBEL
X1SCHOOLENG
Clustering Variables

Figure 28

Income Level	Cluster	Count	PercentOfGroup
Family income less than or equal to $15,000	0	256	18.31%
Family income less than or equal to $15,000	1	1142	81.68%
Family income > $15,000 and <= $35,000	0	736	26.50%
Family income > $15,000 and <= $35,000	1	2041	73.50%
Family income > $35,000 and <= $55,000	0	1121	44.34%
Family income > $35,000 and <= $55,000	1	1407	55.65%
Family income > $55,000 and <= $75,000	0	1296	56.30%
Family income > $55,000 and <= $75,000	1	1006	43.70%
Family income > $75,000 and <= $95,000	0	1147	66.96%
Family income > $75,000 and <= $95,000	1	566	33.04%
Family income > $95,000 and <= $115,000	0	1013	73.94%
Family income > $95,000 and <= $115,000	1	357	26.058%
Family income > $115,000 and <= $135,000	0	696	77.85%
Family income > $115,000 and <= $135,000	1	198	22.147%
Family income > $135,000 and <= $155,000	0	586	84.68%
Family income > $135,000 and <= $155,000	1	106	15.32%
Family income > $155,000 and <=$175,000	0	281	82.16%
Family income > $155,000 and <=$175,000	1	61	17.83%
Family income > $175,000 and <= $195,000	0	191	86.82%
Family income > $175,000 and <= $195,000	1	29	13.18%
Family income > $195,000 and <= $215,000	0	261	89.38%
Family income > $195,000 and <= $215,000	1	31	10.62%
Family income > $215,000 and <= $235,000	0	94	86.24%
Family income > $215,000 and <= $235,000	1	15	13.76%
Family income > $235,000	0	653	90.32%
Family income > $235,000	1	70	9.68%

Table 3

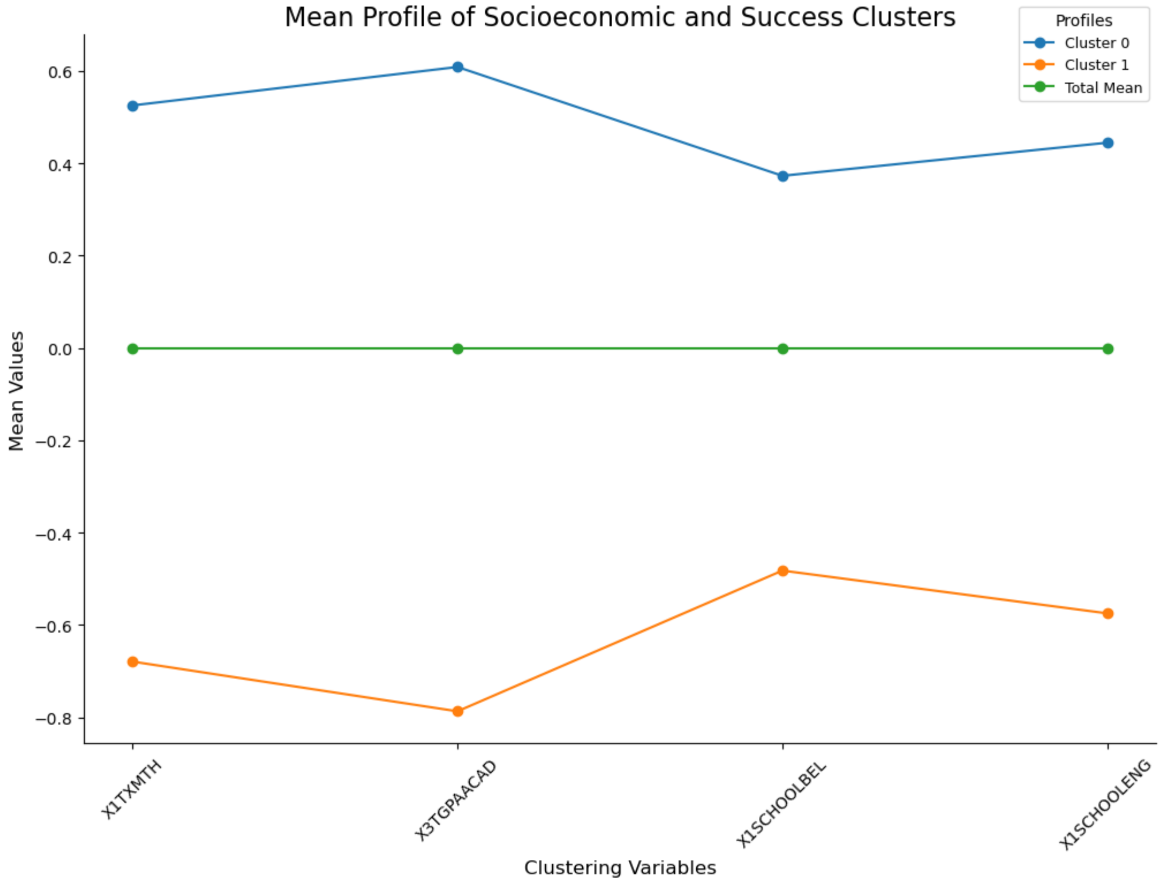

Mean Profile of Socioeconomic and Success Clusters
Profiles
Cluster 0
Cluster 1
Total Mean
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
Mean Values
X1TXMTH
X3TGPAACAD
X1SCHOOLBEL
X1SCHOOLENG
Clustering Variables

Figure 29

Income Level	Cluster Acad.	Count	PercentOfGroup
Family income less than or equal to $15,000	0	470	33.62%
Family income less than or equal to $15,000	1	928	66.38%
Family income > $15,000 and <= $35,000	0	1006	36.23%
Family income > $15,000 and <= $35,000	1	1771	63.77%
Family income > $35,000 and <= $55,000	0	1237	48.93%
Family income > $35,000 and <= $55,000	1	1291	51.07%
Family income > $55,000 and <= $75,000	0	1325	57.56%
Family income > $55,000 and <= $75,000	1	977	42.44%
Family income > $75,000 and <= $95,000	0	1120	65.38%
Family income > $75,000 and <= $95,000	1	593	34.62%
Family income > $95,000 and <= $115,000	0	964	70.36%
Family income > $95,000 and <= $115,000	1	406	29.64%
Family income > $115,000 and <= $135,000	0	641	71.70%
Family income > $115,000 and <= $135,000	1	253	28.30%
Family income > $135,000 and <= $155,000	0	539	77.89%
Family income > $135,000 and <= $155,000	1	153	22.11%
Family income > $155,000 and <=$175,000	0	260	76.02%
Family income > $155,000 and <=$175,000	1	82	23.98%
Family income > $175,000 and <= $195,000	0	170	77.27%
Family income > $175,000 and <= $195,000	1	50	22.73%
Family income > $195,000 and <= $215,000	0	235	80.48%
Family income > $195,000 and <= $215,000	1	57	19.52%
Family income > $215,000 and <= $235,000	0	91	83.49%
Family income > $215,000 and <= $235,000	1	18	16.51%
Family income > $235,000	0	604	83.54%
Family income > $235,000	1	119	16.46%

Table 4

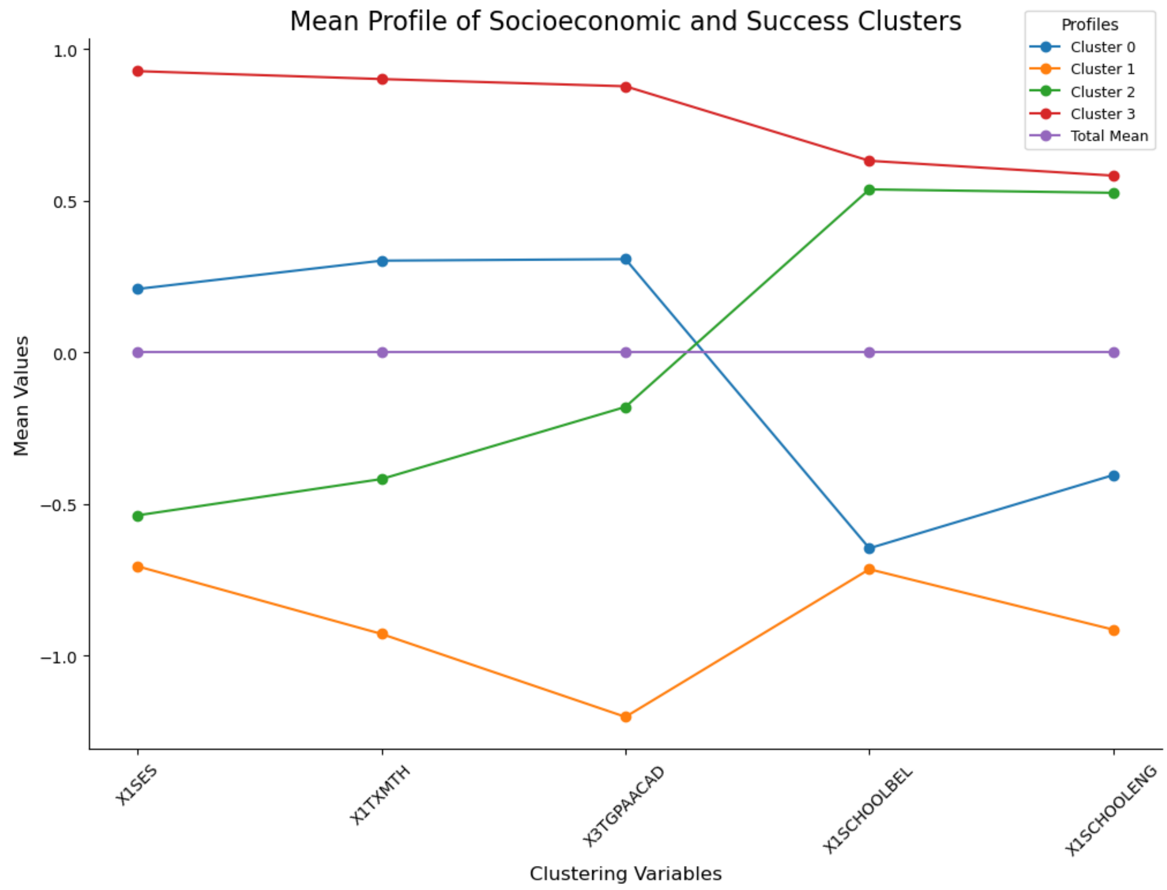


Figure 30

Logit Regression Results			
Dep. Variable:	X3GPACAT	**No. Observations:**	15,480
Model:	Logit	**Df Residuals:**	15,454
Method:	MLE	**Df Model:**	25
Date:	Aug 2025	**Pseudo R^2:**	0.1281
Time:	12:00	**Log-Likelihood:**	-9,349
converged:	TRUE	**LL-Null:**	-10,722
Covariance Type:	nonrobust	**LLR p-value:**	0

Table 5

Variable	Coeff.	P-Value	Odds Ratio	Change in Odds
Intercept	-0.388512	0.001685	0.678065	-32.19%
Native American/Alaska Native, non-Hispanic	-0.981330	0.000016	0.374812	-62.52%
Asian, non-Hispanic	0.874024	0.000000	2.396535	139.65%
Black/African American, non-Hispanic	-0.925370	0.000000	0.396385	-60.36%
Hispanic, no race specified	-0.556806	0.022487	0.573036	-42.69%
Hispanic, race specified	-0.370326	0.000000	0.690509	-30.94%
More than one race, non-Hispanic	-0.246205	0.000077	0.781762	-21.82%
Native Hawaiian/Pacific Islander, non-Hispanic	-0.697198	0.005387	0.497979	-50.20%
Family income > $15,000 and <= $35,000	-0.093938	0.226962	0.910339	-8.97%
Family income > $35,000 and <= $55,000	0.098451	0.236587	1.103460	10.35%
Family income > $55,000 and <= $75,000	0.259390	0.003477	1.296139	29.61%
Family income > $75,000 and <= $95,000	0.355227	0.000255	1.426504	42.65%
Family income > $95,000 and <= $115,000	0.278736	0.007443	1.321458	32.15%
Family income > $115,000 and <= $135,000	0.358183	0.001980	1.430727	43.07%
Family income > $135,000 and <= $155,000	0.376441	0.002995	1.457090	45.71%
Family income > $155,000 and <=$175,000	0.314410	0.040093	1.369451	36.95%
Family income > $175,000 and <= $195,000	0.454434	0.014814	1.575282	57.53%
Family income > $195,000 and <= $215,000	0.455066	0.007932	1.576277	57.63%
Family income > $215,000 and <= $235,000	0.974023	0.000523	2.648578	164.86%
Family income > $235,000	0.597220	0.000016	1.817060	81.71%
High school diploma or GED	0.117394	0.195703	1.124562	12.46%
Associate's degree	0.140578	0.173042	1.150939	15.09%
Bachelor's degree	0.573725	0.000000	1.774866	77.49%
Master's degree	0.799899	0.000000	2.225316	122.53%
Ph.D./M.D/Law/other high lvl prof degree	0.670087	0.000038	1.954407	95.44%
X1 Socio-economic status composite	0.446987	0.000000	1.563594	56.36%

Table 6

Logit Regression Results			
Dep. Variable:	X3GPACAT	**No. Observations:**	11610
Model:	Logit	**Df Residuals:**	11584
Method:	MLE	**Df Model:**	25
Date:	Aug 2025	**Pseudo R^2:**	0.1271
Time:	12:00	**Log-Likelihood:**	-7019.4
converged:	True	**LL-Null:**	-8041.6
Covariance Type:	nonrobust	**LLR p-value:**	0.000

Table 7

Variable	Coef.	P-value	Odds Ratio	Change in Odds
Intercept	-0.388512	0.001685	0.678065	-32.19%
Amer. Indian/Alaska Native, non-Hispanic	-0.98133	0.000016	0.374812	-62.52%
Asian, non-Hispanic	0.874024	0.000000	2.396535	139.65%
Black/African American, non-Hispanic	-0.92537	0.000000	0.396385	-60.36%
Hispanic, no race specified	-0.556806	0.022487	0.573036	-42.70%
Hispanic, race specified	-0.370326	0.000000	0.690509	-30.95%
More than one race, non-Hispanic	-0.246205	0.000077	0.781762	-21.82%
Native Hawaiian/Pacific Islander, non-Hispanic	-0.697198	0.005387	0.497979	-50.20%
Family income > $15,000 and <= $35,000	-0.093938	0.226962	0.910339	-8.97%
Family income > $35,000 and <= $55,000	0.098451	0.236587	1.10346	10.35%
Family income > $55,000 and <= $75,000	0.25939	0.003477	1.296139	29.61%
Family income > $75,000 and <= $95,000	0.355227	0.000255	1.426504	42.65%
Family income > $95,000 and <= $115,000	0.278736	0.007443	1.321458	32.15%
Family income > $115,000 and <= $135,000	0.358183	0.001980	1.430727	43.07%
Family income > $135,000 and <= $155,000	0.376441	0.002995	1.45709	45.71%
Family income > $155,000 and <=$175,000	0.31441	0.040093	1.369451	36.95%
Family income > $175,000 and <= $195,000	0.454434	0.014814	1.575282	57.53%
Family income > $195,000 and <= $215,000	0.455066	0.007932	1.576277	57.63%
Family income > $215,000 and <= $235,000	0.974023	0.000523	2.648578	164.86%
Family income > $235,000	0.59722	0.000016	1.81706	81.71%
High school diploma or GED	0.117394	0.195703	1.124562	12.46%
Associate's degree	0.140578	0.173042	1.150939	15.09%
Bachelor's degree	0.573725	0.000000	1.774866	77.49%
Master's degree	0.799899	0.000000	2.225316	122.53%
Ph.D/M.D/Law/other high lvl prof degree	0.670087	0.000038	1.954407	95.44%
X1 Socio-economic status composite	0.446987	0.000000	1.563594	56.36%

Table 8

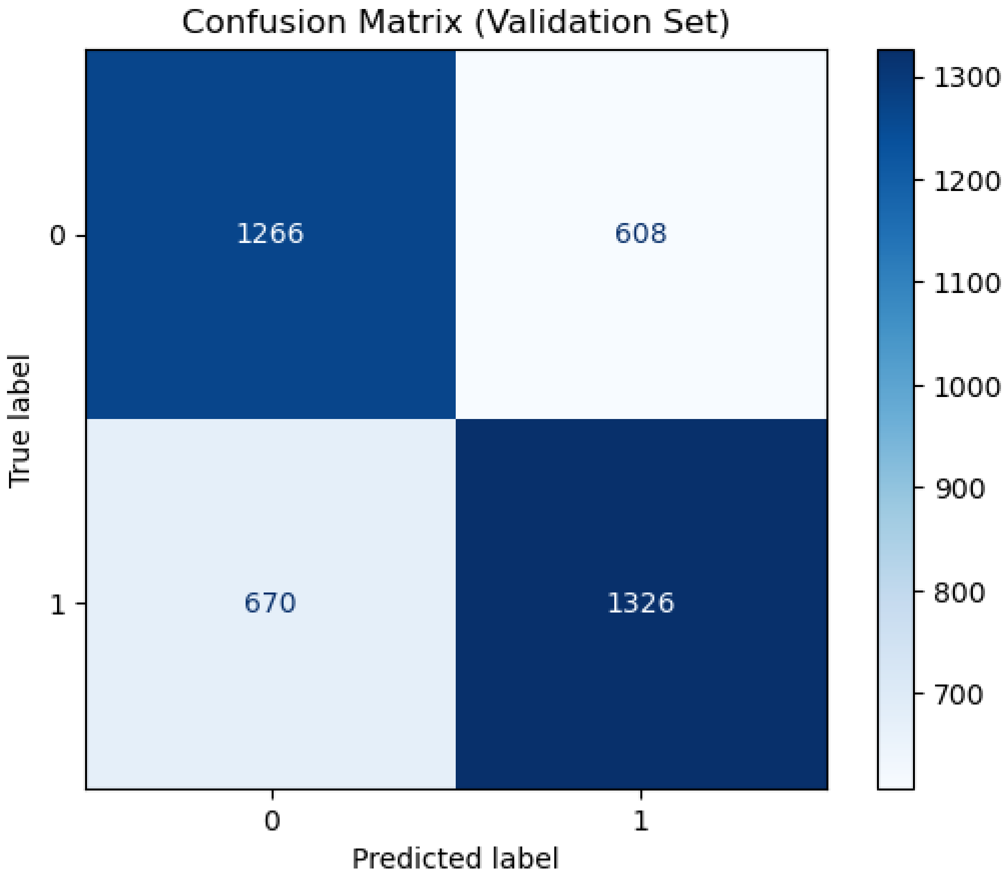

Confusion Matrix (Validation Set)
True label
Predicted label
0
1
0
1
1266
608
670
1326
1300
1200
1100
1000
900
800
700

Figure 31

Math Theta Score

Logit Regression Results			
Dep. Variable:	X1TXMTHCAT	**No. Observations:**	15480
Model:	Logit	**Df Residuals:**	15454
Method:	MLE	**Df Model:**	25
Date:	Aug 2025	**Pseudo R^2:**	0.1139
Time:	12:00	**Log-Likelihood:**	-9503.6
converged:	True	**LL-Null:**	-10725.
Covariance Type:	nonrobust	**LLR p-value:**	0.000

Table 9

Variable	Coef.	P-Value	Odds Ratio	Change in Odds
Intercept	-0.736333	0.00000	0.478867	-52.11%
Amer. Indian/Alaska Native, non-Hispanic	-0.583747	0.00746	0.557804	-44.22%
Asian, non-Hispanic	1.029402	0.00000	2.799391	179.94%
Black/African American, non-Hispanic	-0.756284	0.00000	0.469408	-53.06%
Hispanic, no race specified	-0.655555	0.01320	0.519154	-48.08%
Hispanic, race specified	-0.219526	0.00004	0.802899	-19.71%
More than one race, non-Hispanic	0.005301	0.93166	1.005315	0.53%
Native Hawaiian/Pacific Islander, non-Hispanic	-0.07955	0.74090	0.923532	-7.65%
Family income > $15,000 and <= $35,000	-0.058307	0.46013	0.94336	-5.66%
Family income > $35,000 and <= $55,000	0.184151	0.02808	1.202197	20.22%
Family income > $55,000 and <= $75,000	0.399276	0.00001	1.490745	49.07%
Family income > $75,000 and <= $95,000	0.344829	0.00038	1.411748	41.17%
Family income > $95,000 and <= $115,000	0.43026	0.00003	1.537657	53.77%
Family income > $115,000 and <= $135,000	0.428396	0.00019	1.534794	53.48%
Family income > $135,000 and <= $155,000	0.513568	0.00004	1.671244	67.12%
Family income > $155,000 and <=$175,000	0.508546	0.00078	1.662872	66.29%
Family income > $175,000 and <= $195,000	0.416626	0.01951	1.516835	51.68%
Family income > $195,000 and <= $215,000	0.398643	0.01496	1.489802	48.98%
Family income > $215,000 and <= $235,000	0.676377	0.00572	1.966739	96.67%
Family income > $235,000	0.575796	0.00002	1.778546	77.85%
High school diploma or GED	0.225316	0.01562	1.252719	25.27%
Associate's degree	0.238285	0.02354	1.269071	26.91%
Bachelor's degree	0.616816	0.00000	1.853019	85.30%
Master's degree	0.797395	0.00000	2.219751	121.98%
Ph.D/M.D/Law/other high lvl prof degree	0.845564	0.00000	2.329291	132.93%
X1 Socio-economic status composite	0.397381	0.00000	1.487923	48.79%

Table 10

School Belonging

Logit Regression Results			
Dep. Variable:	X1SCHOOLBELCAT	**No. Observations:**	15480
Model:	Logit	**Df Residuals:**	15454
Method:	MLE	**Df Model:**	25
Date:	Aug 2025	**Pseudo R^2:**	0.01641
Time:	12:00	**Log-Likelihood:**	-10542.
converged:	True	**LL-Null:**	-10717.
Covariance Type:	nonrobust	**LLR p-value:**	2.195e-59

Table 11

Variable	**Coef.**	**P-Value**	**Odds Ratio**	**Change in Odds**
Intercept	-0.340721	0.00214	0.711257	-28.87%
Amer. Indian/Alaska Native, non-Hispanic	-0.043077	0.82548	0.957838	-4.22%
Asian, non-Hispanic	0.157017	0.01436	1.170016	17.00%
Black/African American, non-Hispanic	0.159186	0.00609	1.172556	17.26%
Hispanic, no race specified	-0.049357	0.81389	0.951841	-4.82%
Hispanic, race specified	0.017954	0.71944	1.018116	1.81%
More than one race, non-Hispanic	-0.082362	0.16633	0.920939	-7.91%
Native Hawaiian/Pacific Islander, non-Hispanic	-0.286205	0.22620	0.751109	-24.89%
Family income > $15,000 and <= $35,000	-0.10558	0.13119	0.899802	-10.02%
Family income > $35,000 and <= $55,000	-0.042389	0.57894	0.958497	-4.15%
Family income > $55,000 and <= $75,000	-0.008998	0.91275	0.991042	-0.90%
Family income > $75,000 and <= $95,000	0.073458	0.41426	1.076223	7.62%
Family income > $95,000 and <= $115,000	0.169937	0.07786	1.18523	18.52%
Family income > $115,000 and <= $135,000	0.171245	0.10758	1.186781	18.68%
Family income > $135,000 and <= $155,000	0.240139	0.03697	1.271426	27.14%
Family income > $155,000 and <=$175,000	0.396281	0.00488	1.486287	48.63%
Family income > $175,000 and <= $195,000	0.153552	0.34574	1.165968	16.60%
Family income > $195,000 and <= $215,000	0.128965	0.38569	1.13765	13.77%
Family income > $215,000 and <= $235,000	0.376958	0.08107	1.457843	45.78%
Family income > $235,000	0.502002	0.00004	1.652025	65.20%
High school diploma or GED	0.086499	0.28202	1.09035	9.04%
Associate's degree	0.111544	0.23129	1.118003	11.80%
Bachelor's degree	0.29099	0.00373	1.337751	33.78%
Master's degree	0.337488	0.00396	1.401423	40.14%
Ph.D/M.D/Law/other high lvl prof degree	0.176676	0.22412	1.193244	19.32%
X1 Socio-economic status composite	0.106554	0.03261	1.112438	11.24%

Table 12

School Engagement

Logit Regression Results			
Dep. Variable:	X1SCHOOLENGCAT	**No. Observations:**	15480
Model:	Logit	**Df Residuals:**	15454
Method:	MLE	**Df Model:**	25
Date:	Aug 2025	**Pseudo R^2:**	0.01524
Time:	12:00	**Log-Likelihood:**	-10493.
converged:	True	**LL-Null:**	-10655.
Covariance Type:	nonrobust	**LLR p-value:**	6.130e-54

Table 13

Variable	Coef.	P-Value	Odds Ratio	Change in Odds
Intercept	0.021508	0.84565	1.021741	2.17%
Amer. Indian/Alaska Native, non-Hispanic	-0.628413	0.00157	0.533438	-46.66%
Asian, non-Hispanic	0.41148	0.00000	1.50905	50.90%
Black/African American, non-Hispanic	-0.146408	0.01135	0.863805	-13.62%
Hispanic, no race specified	-0.193672	0.34724	0.823928	-17.61%
Hispanic, race specified	-0.145433	0.00341	0.864648	-13.54%
More than one race, non-Hispanic	-0.114245	0.05339	0.892039	-10.80%
Native Hawaiian/Pacific Islander, non-Hispanic	0.058281	0.80170	1.060013	6.00%
Family income > $15,000 and <= $35,000	-0.209176	0.00255	0.811252	-18.87%
Family income > $35,000 and <= $55,000	-0.108949	0.15179	0.896776	-10.32%
Family income > $55,000 and <= $75,000	-0.049196	0.54832	0.951995	-4.80%
Family income > $75,000 and <= $95,000	0.048298	0.59232	1.049483	4.95%
Family income > $95,000 and <= $115,000	0.155298	0.10967	1.168006	16.80%
Family income > $115,000 and <= $135,000	0.165619	0.12351	1.180123	18.01%
Family income > $135,000 and <= $155,000	0.217316	0.06292	1.242737	24.27%
Family income > $155,000 and <=$175,000	0.228989	0.10767	1.257328	25.73%
Family income > $175,000 and <= $195,000	0.151154	0.36366	1.163176	16.32%
Family income > $195,000 and <= $215,000	0.179854	0.23788	1.197043	19.70%
Family income > $215,000 and <= $235,000	0.266523	0.22773	1.305418	30.54%
Family income > $235,000	0.009504	0.93790	1.009549	0.95%
High school diploma or GED	0.156652	0.04815	1.169589	16.96%
Associate's degree	0.260264	0.00482	1.297273	29.73%
Bachelor's degree	0.247987	0.01299	1.281443	28.14%
Master's degree	0.266594	0.02278	1.30551	30.55%
Ph.D/M.D/Law/other high lvl prof degree	0.227617	0.11917	1.255604	25.56%
X1 Socio-economic status composite	0.085316	0.08977	1.089061	8.91%

Table 14

Works Cited

Cirtautas, Justinas. *NBA Players*. Kaggle, n.d. Kaggle, https://www.kaggle.com/datasets/justinas/nba-players-data.

City of Los Angeles (LAPD OpenData). *Crime Data from 2020 to Present*. Data.lacity.org, 10 Feb. 2020. Updated 23 July 2025. https://data.lacity.org/d/2nrs-mtv8.

Food and Agriculture Organization of the United Nations. *FAOSTAT Statistical Database*. CC-BY-4.0 licence, https://www.fao.org/faostat/en/#data.

Hannah Ritchie, Pablo Rosado, and Max Roser (2023) - “CO_2 and Greenhouse Gas Emissions” Published online at OurWorldinData.org. Retrieved from: 'https://ourworldindata.org/co2-and-greenhouse-gas-emissions' [Online Resource]

U.S. Department of Education, National Center for Education Statistics. *High School Longitudinal Study of 2009 (HSLS:09) [Dataset]*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. Retrieved from https://nces.ed.gov/surveys/hsls09/

U.S. Department of Health and Human Services, *Poverty Guidelines*. Office of the Assistant Secretary for Planning and Evaluation, 17 Jan. 2025. ASPE, https://aspe.hhs.gov/topics/poverty-economic-mobility/poverty-guidelines.