1. Introduction

This analysis leverages a subset of data from an early childhood longitudinal study conducted in 1998-99, focusing on kindergarten students' academic progress over several months. Specifically, it examines changes in reading, math, and general knowledge scores from fall 1998 to spring 1999. Additionally, the dataset categorizes students by income group, the sole categorical variable used to assess the potential impact of socioeconomic status on academic improvement. This investigation aims to elucidate the relationship between income levels and educational outcomes in early childhood, providing insights into the broader implications of socioeconomic factors on learning and development.

2. Data Cleaning and Data Wrangling

The dataset comprises 11,933 entries across 9 columns, detailed as follows:

fallreadingscore: Reading scores recorded in the fall. The scores range from 21.01 to 138.51 with a mean of approximately 35.95.

fallmathscore: Math scores recorded in the fall. These scores range from 10.51 to 115.65 with a mean of about 27.13.

fallgeneralknowledgescore: General knowledge scores recorded in the fall, ranging from 6.985 to 47.691, with an average score of around 23.07.

springreadingscore: Reading scores recorded in the spring. Scores range from 22.35 to 156.85, with a mean score of approximately 47.51.

springmathscore: Math scores recorded in the spring, ranging from 11.9 to 113.8, with an average of about 37.80.

springgeneralknowledgescore: General knowledge scores recorded in the spring, ranging from 7.858 to 48.345, with an average of roughly 28.24.

totalhouseholdincome: Total household income, which ranges from 1 to 150,000, with a mean of approximately 54,317.20.

incomeinthousands: This appears to be a scaled version of the total household income, with the same range but expressed in thousands.

incomegroup: Categorized into three income groups (1 to 3), this column seems to categorize the total household income into distinct groups.

After inspecting the dataset, there is no null values in the dataset, indicating the dataset requires no further processing.

3. Exploratory Data Analysis

The exploratory data analysis reveals the following insights:

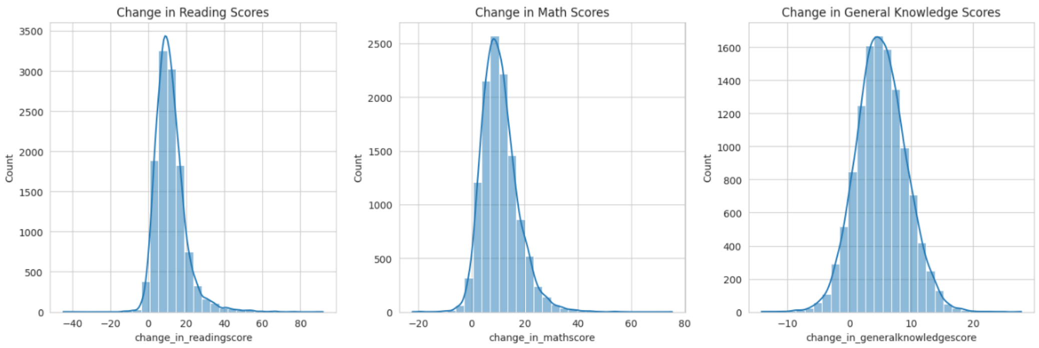


Figure 1: Score changes in reading, math and general knowledge

Change in Scores from Fall to Spring (As shown in Figure 1**):**

Reading Scores: The change in reading scores appears to be normally distributed with a mean increase suggesting that, on average, students improve their reading scores from fall to spring.

Math Scores: Similar to reading scores, the change in math scores also seems to be normally distributed with a mean increase, indicating improvement from fall to spring.

General Knowledge Scores: The distribution of changes in general knowledge scores follows the pattern of reading and math scores, with a mean increase that suggests general improvement.

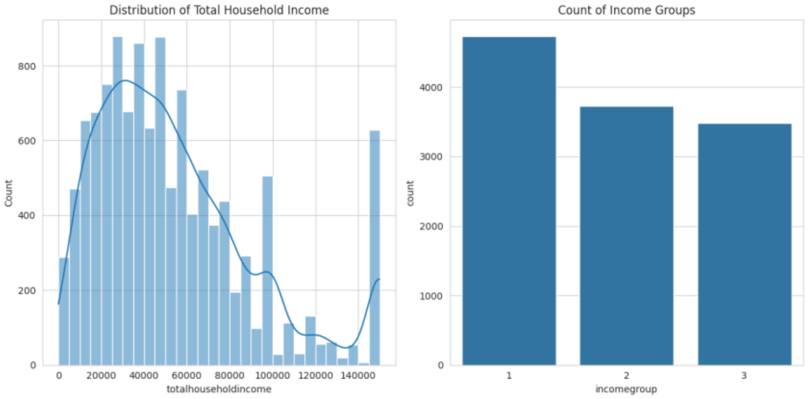


Figure 2: Income distribution

Income Distribution (As shown in Figure 2**):**

Total Household Income: The distribution of total household income is right-skewed, indicating that a larger number of households have incomes on the lower end of the scale, with fewer households having higher incomes.

Income Groups: The distribution across income groups shows a higher count in the lower income groups (1 and 2) compared to the highest income group (3).

Given these observations, our next step will involve conducting one-way ANCOVAs to explore how income group or total household income might influence the change in academic performance scores from fall to spring, controlling for other factors. This analysis will help us understand if there's a statistically significant difference in academic improvement across different income levels, controlling for initial performance scores.

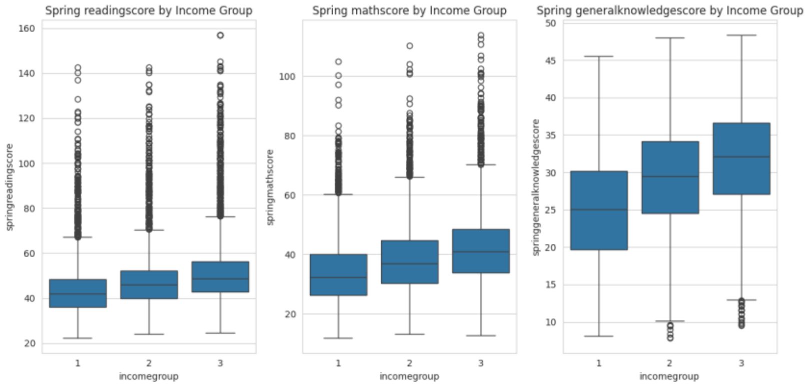


Figure 3: Boxplot for scores by income group

Research Questions:

1. Does the income group influence the degree of improvement in reading scores from fall to spring, after controlling for initial fall reading scores?
2. Does the income group influence the degree of improvement in math scores from fall to spring, after controlling for initial fall math scores?

Question 1

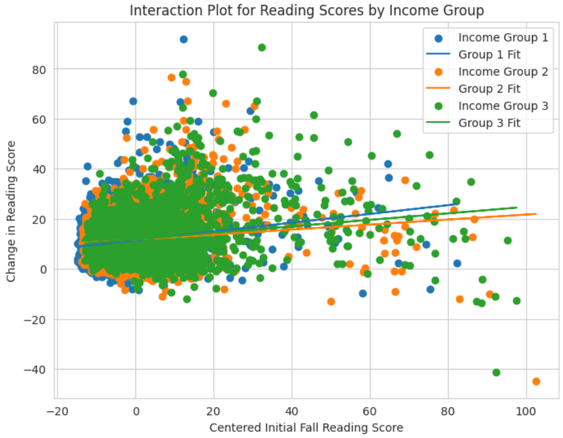


Figure 4: plot for reading scores

The distribution of the data points suggests variability in score changes within each income group. The slopes of the fitted lines for each group look quite similar, implying that the relationship between initial math ability and improvement does not vary drastically by income group. However, it's worth noting that the concentration of data points and the fitted line for Income Group 3 suggests this group has a broader range of initial scores and changes compared to the other groups. This plot is useful for visually assessing potential interaction effects between income group and initial math ability on students' score improvement.

Assumption Check:

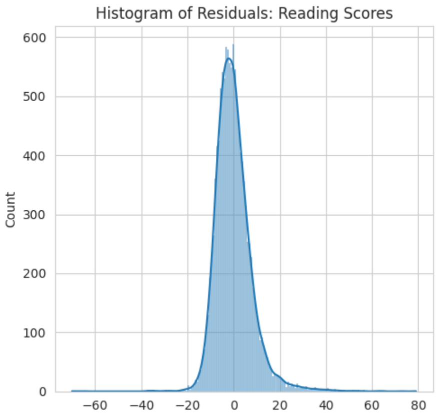


Figure 5: histogram of residuals

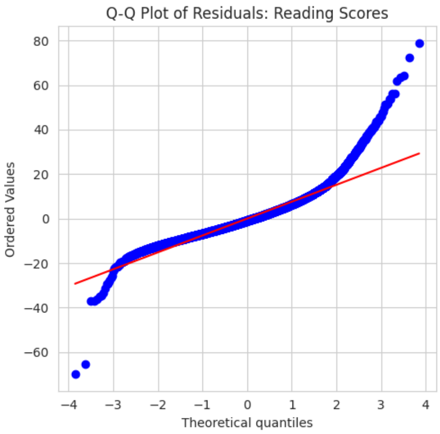


Figure 6: Q-Q plot of residuals

The histogram of residuals for reading scores presents a distribution that closely approximates a normal distribution, peaking at the center and tapering off symmetrically, which is indicative of the residuals aligning well with one of the key ANCOVA assumptions – the normality of residuals.

In the Q-Q plot of residuals for reading scores, the ordered values are plotted against the theoretical quantiles of a normal distribution. The blue points represent the residuals. The alignment along the red line in the center indicates normality, but there is notable deviation at the tails, suggesting that the residuals have heavier tails than expected under a normal distribution. This could impact the robustness of the ANCOVA results.

The Levene's test result, with a p-value significantly lower than the standard alpha level of 0.05, suggests that the assumption of equal variances (homoscedasticity) across the groups is not met.

Post-hoc Analysis:

Tukey's HSD test results show significant mean differences between all income groups for reading scores, with Group 1 differing from Group 2 by an average of 0.8387, Group 1 from Group 3 by 1.4301, and Group 2 from Group 3 by 0.5915. All comparisons are statistically significant, indicating that income group is a factor in reading score improvement.

group1	**group2**	**meandiff**	**p-adj**	**lower**	**upper**	**reject**
1	2	0.8387	0.0	0.4246	1.2527	True
1	3	1.4301	0.0	1.0079	1.8523	True
2	3	0.5915	0.0053	0.1458	1.0371	True

	sum_sq	**df**	**F**	**PR(&gt;F)**
incomegroup	513.1201	2.0	4.0557	0.0173
fallreadingscore	21099.6089	1.0	333.5392	1.6292e-73
Residual	754625.5522	11929.0	NaN	NaN

Table 1: Post-hoc test

Table 2: ANCOVA Reading Results

Change in Reading Scores:

Income Group Effect: The analysis indicated a statistically significant effect of the income group on the change in reading scores (F=4.06, p=0.017). This result suggests that students from different income groups experience varied degrees of improvement in reading scores from fall to spring, even when their initial reading scores are accounted for.

Fall Reading Score: The covariate, fall reading score, also significantly affects the change in reading scores (F=333.54, p<0.001), indicating that initial reading performance is a strong predictor of improvement.

Question 2

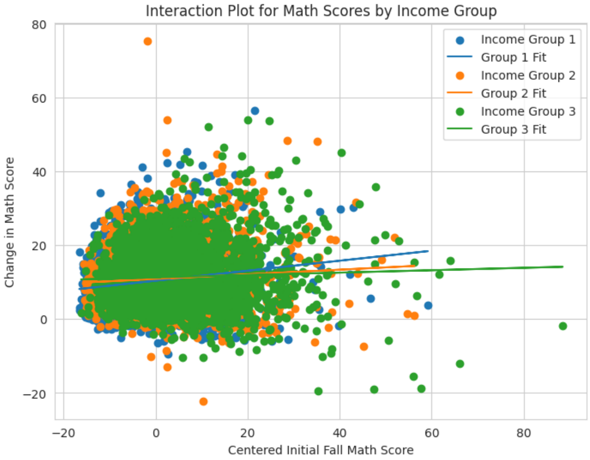
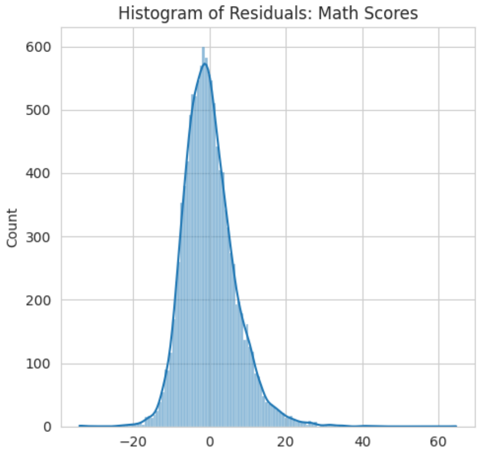


Figure 7: Interaction plot for math score

The interaction plot for math scores shows the change in scores by income group against centered initial fall scores. Data for Income Groups 1 (blue) and 2 (orange) display a modest upward trend, while Group 3 (green) indicates a more distinct positive correlation, suggesting higher initial scores may correspond to greater improvement in this group. The overlapping data points across the groups suggest similar variability in score changes, with no pronounced differences in the trends of score changes relative to initial performance across the different income groups.

Assumption Check:



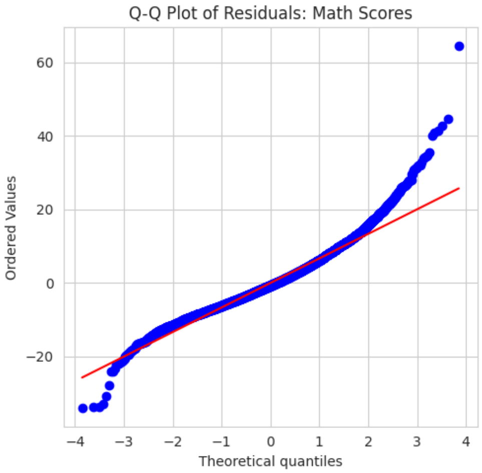


Figure 8: histogram of residuals

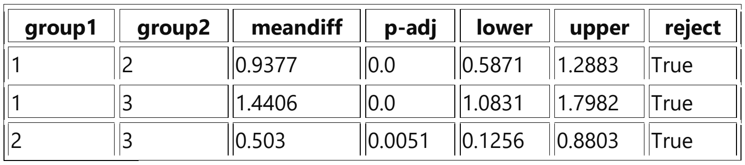
Figure 9: Q-Q plot of residuals

The histogram displays the residuals from the math score ANCOVA model. A perfect normal distribution would resemble a symmetrical bell curve centered around zero. Here, the residuals exhibit a right skew, with a tail extending towards the higher values. While not perfectly normal, the bulk of the data clusters around the center, which could be acceptable in large samples.

The Q-Q plot contrasts the observed residuals against those expected under a normal distribution, with points straying from the red line indicating departures from normality. The curvature and the marked deviation at the extreme ends suggest that the residuals have heavier tails than a normal distribution, which may affect the robustness of the ANCOVA results.

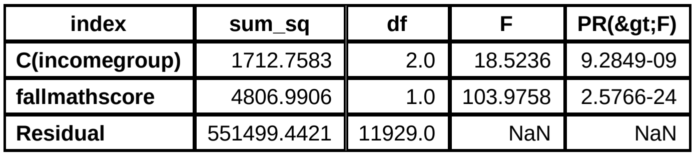
The Levene's test result, with a p-value significantly lower than the standard alpha level of 0.05, suggests that the assumption of equal variances (homoscedasticity) across the groups is not met.

Post-hoc Analysis:



group1	group2	meandiff	p-adj	lower	upper	reject
1	2	0.9377	0.0	0.5871	1.2883	True
1	3	1.4406	0.0	1.0831	1.7982	True
2	3	0.503	0.0051	0.1256	0.8803	True

Table 3: Post-hoc test



index	sum_sq	df	F	PR(&gt;F)
C(incomegroup)	1712.7583	2.0	18.5236	9.2849-09
fallmathscore	4806.9906	1.0	103.9758	2.5766-24
Residual	551499.4421	11929.0	NaN	NaN

Table 4: ANCOVA Reading Results

Change in Math Scores:

Income Group Effect: The results showed a significant effect of the income group on the change in math scores (F=18.52, p<0.001). This demonstrates that the income group plays a substantial role in the improvement of math scores across the academic year, beyond the initial performance level.

Fall Math Score: The fall math score is a significant covariate (F=103.98, p<0.001), showing that initial math performance strongly predicts the degree of improvement.

Conclusion:

The ANCOVA results collectively underscore the influence of socioeconomic status on academic improvement across different subject areas. Even after controlling for initial performance levels, the income group remains a significant predictor of academic progress, indicating that factors associated with socioeconomic status—such as access to resources, learning environments, or extracurricular support—may affect students' ability to improve academically. This underscores the need for educational policies and interventions that specifically address the disparities associated with socioeconomic status to ensure equitable academic improvement opportunities for all students.