Student Name: Zhibing Liu Student Number: 1010506515

Bridging the Gap: An Analysis of Socioeconomic Influences on Academic Achievement in Early Education

1. Introduction

The interplay between income levels and academic performance has been a subject of extensive research within the field of education. Studies have repeatedly shown that socioeconomic status significantly influences students' academic outcomes, access to educational resources, and overall learning environment. This research project seeks to contribute to this body of knowledge by examining the early childhood longitudinal study dataset, which provides a rich source of information on kindergarten students' reading, math, and general knowledge scores across two semesters.

In particular, this study will focus on the differences in academic performance among students from various income groups, aiming to understand how socioeconomic factors influence learning outcomes over the academic year. By employing quantitative analysis methods, including one-way ANCOVAs, this project will analyze the dataset to answer pivotal questions regarding the relationship between income levels and educational achievement. Through this examination, the research aims to highlight areas where policy interventions could be targeted to support students from lower-income families, thereby contributing to efforts to level the educational playing field and ensure equal opportunities for all students.

Our exploration will address two fundamental research questions, sets the stage for a detailed examination of the dataset, guiding the analysis towards understanding the nuanced ways in which income levels might affect academic performance among kindergarten students.
Research Question 1: How do changes in students' reading scores from fall to spring term vary across different income groups, after controlling for their reading scores at the start of the fall term?
Research Question 2: How do changes in students' math scores from fall to spring term vary across different income groups, after controlling for their reading scores at the start of the fall term?

2. Data Cleaning and Data Wrangling

The dataset contains 11,933 entries and 9 columns, comprising several columns relevant to our study on the impact of income levels on students' academic performance, with a focus on reading, math, and general knowledge scores across two terms: fall and spring. After a thorough review, we determined that the dataset was in good shape for our analysis, requiring minimal data cleaning. This review also allowed us to identify which columns were essential for our analysis and which could be excluded.

A. Observations and Considerations:

For our quantitative analysis, we streamlined the dataset by selecting only the columns that are crucial for understanding the changes in academic performance across income groups and between the fall and spring terms. The following is a brief description of each retained column:

- fallreadingscore: Reading scores of students in the fall term.
- fallmathscore: Math scores of students in the fall term.
- fallgeneralknowledgescore: General knowledge scores of students in the fall term.
- springreadingscore: Reading scores of students in the spring term.
- springmathscore: Math scores of students in the spring term.
- springgeneralknowledgescore: General knowledge scores of students in the spring term.
- incomegroup: Categorized income levels of students' households.

The dataset was found to be complete with no significant missing values within the selected columns, ensuring the integrity of our analysis.

B. Feature Engineering:

To enrich our dataset and facilitate more insightful analysis, we introduced three new features that reflect the changes in scores from the fall to the spring term. These features are as follows:

reading_score_change: This feature represents the change in reading scores from the fall to the spring term, calculated as the difference between springreadingscore and fallreadingscore.

math_score_change: Similarly, this feature denotes the change in math scores from the fall to the spring term, calculated as the difference between springmathscore and fallmathscore.

general_knowledge_score_change: This feature illustrates the change in general knowledge scores from the fall to the spring term.

3. Exploratory Data Analysis (EDA)

After adding our new features to our new working dataset, we proceeded with a comprehensive EDA to leverage insight that could potentially lead to interesting research questions. We started by describing our quantitative data as shown in Figure 1 below.

	fallreadingscore	fallmath score	fallgeneral knowledgescore	springreading score	springmath score	springgeneral knowledgescore	income group	reading_score_change	math_score_change	general_knowledge_score_change
count	11933	11933	11933	11933	11933	11933	11933	11933	11933	11933
mean	35.9542	27.1282	23.0737	47.5111	37.7994	28.236	1.8952	11.557	10.6712	5.1619
std	10.44731	9.1205	7.3970	14.3271	12.0277	7.5775	0.8227	8.085	6.8552	4.0550
min	21.01	10.51	6.985	22.35	11.9	7.858	1	-44.76	-22.16	-14.183
25%	29.34	20.68	17.385	38.95	29.27	22.802	1	6.47	6.01	2.472
50%	34.06	25.68	22.954	45.32	36.41	28.583	2	10.4	9.86	5.047
75%	39.89	31.5	28.305	51.77	44.22	33.782	3	15.15	14.33	7.781
max	138.51	115.65	47.691	156.85	113.8	48.345	3	91.94	75.35	27.785

Figure 1: *Dataset quantitative Data statistics*

Additionally, we employed boxplots seen in Figure 2,3,4 to visually represent the distribution of these features, after removing outliers. This descriptive analysis offered a clearer understanding of the general trends within each feature.

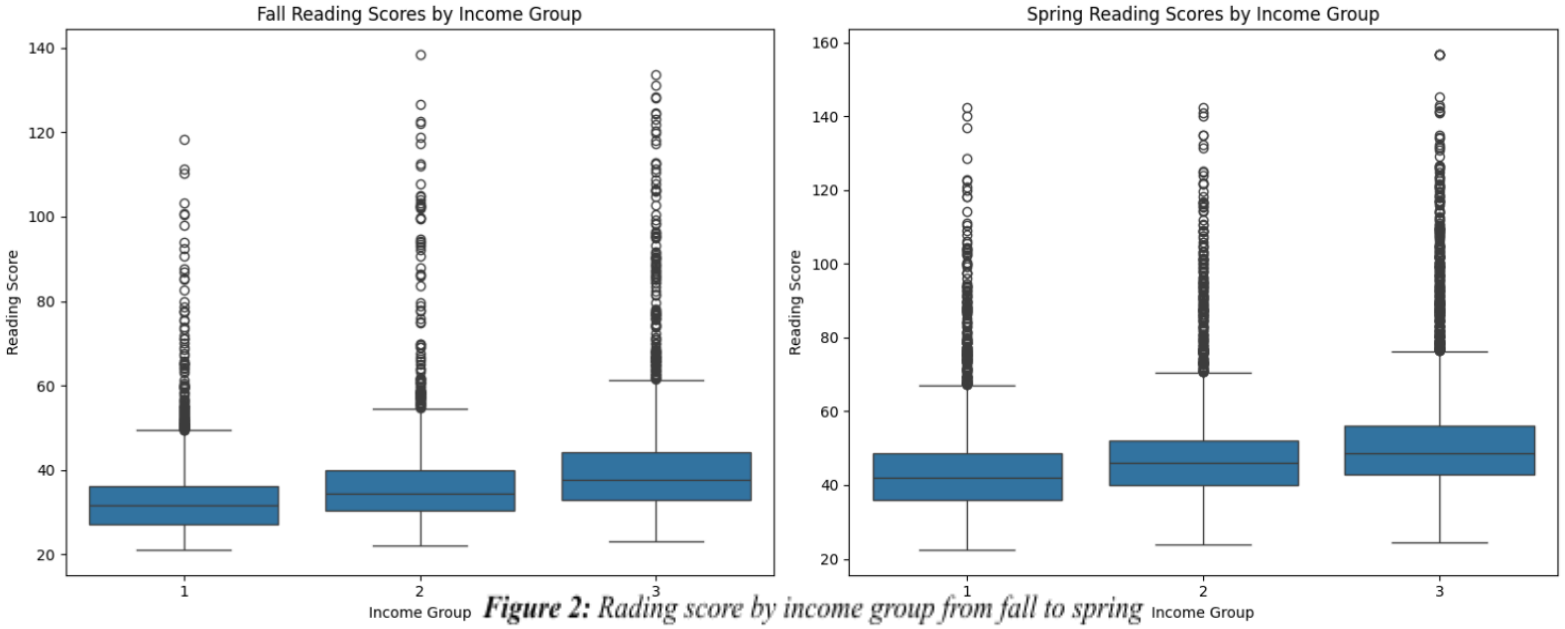

Figure 2: *Rading score by income group from fall to spring*

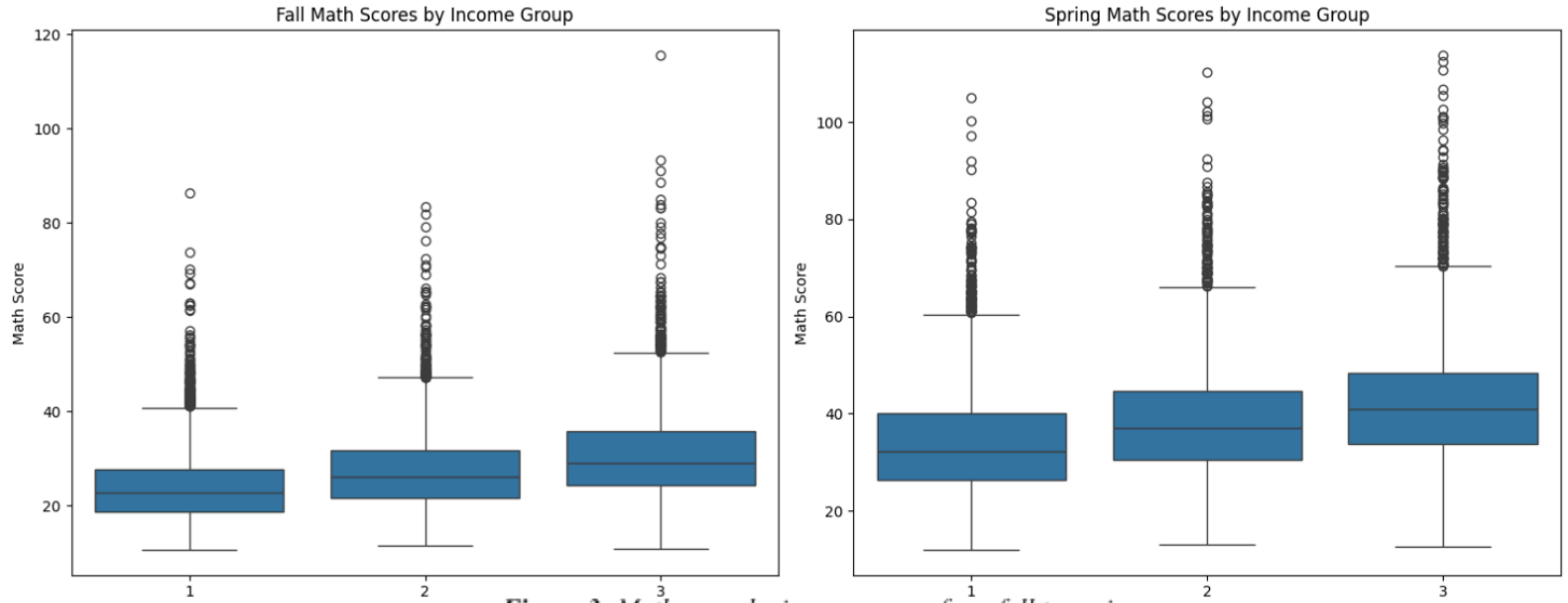

Figure 3: *Math score by income group from fall to spring*

Based on the box plots figure2, figure3 provided, we can observe some distinct distribution patterns in students' academic scores across income groups. The median score, denoted by the line within each box, tends to rise from income group 1 to income group 3 in reading, math scores. This trend suggests that students from higher-income families may have an advantage in academic performance compared to their peers from lower-income groups. Notably, the spread of the scores, as shown by the interquartile range (the height of the boxes), is also wider in the higher income groups, indicating greater variability in those students' performance..

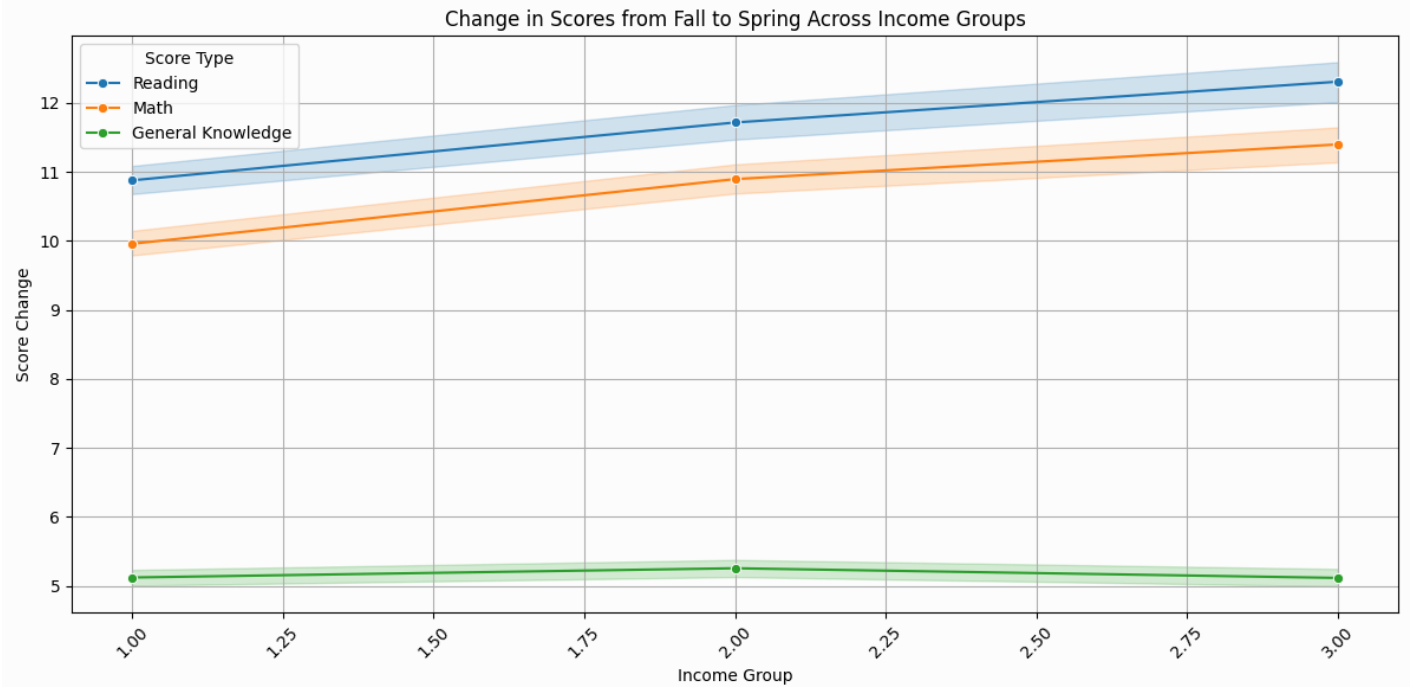

Figure 4 *change in Scores from fall to spring across income groups*

Figure 4 uses shaded areas around each line to depict confidence intervals, providing a visual representation of the uncertainty around the mean score change estimates. The fact that these intervals are relatively narrow for reading and math but wider for general knowledge suggests that there is more variation in the change in general knowledge scores across the population than in reading or math scores.

Overall, the graph provides evidence that may support the hypothesis that higher income is associated with greater improvements in academic performance, particularly in reading and math, over the course of the

academic year. However, general knowledge does not show the same level of disparity based on income group. This visual evidence would be a compelling part of a narrative on the influence of socioeconomic factors on educational progress.

4. Reading scores from fall to spring term vary across different income groups

Research Question #1: How do changes in students' reading scores from fall to spring term vary across different income groups, after controlling for their reading scores at the start of the fall term?". follow the requirement

For this question, We use ANCOVA to investigate whether the reading and math scores vary across different income groups over time in fall term

	Source	SS	DF	F	p-unc	np2
0	incomegroup	513.120105	2	4.055660	1.734793e-02	0.00068
1	fallreadingscore	21099.608888	1	333.539242	1.629240e-73	0.02720
2	Residual	754625.552188	11929	NaN	NaN	NaN

Figure5 *ANCOVA table for reading score in fall term*

The table figure 5 indicates that there is a statistically significant effect of the covariate fallreadingscore on reading_score_change with a p-value much less than 0.05. The incomegroup also shows a significant effect on the change in reading scores. This suggests that after controlling for the initial reading scores from the fall term, there are significant differences in how reading scores changed by the spring term across different income groups.

The F-statistic for income group is significant ($p < 0.05$), which indicates that the mean reading score changes are not the same across all income groups. The np^2 value represents the partial eta squared, which is a measure of effect size; here it suggests that income group has a small to moderate effect on the reading score change.

Figure 6: *Interaction plot for reading scores*

Assumption Checks:

Linearity of the Relationship:

The scatter plot shows the relationship between fallreadingscore and reading_score_change for different income groups. The fitted lines for each group suggest that there is a positive relationship between the fall the reading score and the reading score change, but the slope and position of these lines vary by income group. It looks like there's a reasonably linear relationship, but it does differ by income group, which is meaningful.

Homogeneity of Variance (Levene's Test):
Levene's test p-value for Homogeneity of Variances: 1.0904328845666816e-11
Levene's test is significant ($p < 0.05$), which suggests a violation of the assumption of homogeneity of variances. This means that the variances of reading score changes are not equal across all income groups, which is an important assumption for ANCOVA.
Normality of Residuals (Shapiro-Wilk Test):
Shapiro-Wilk test p-value for Normality of Residuals: <0.001
The Shapiro-Wilk test is also significant ($p < 0.05$), indicating that the residuals from the ANCOVA model are not normally distributed, which is another violation of ANCOVA assumptions.
The significant p-values for both the income group effect and the covariate indicate that both are important in explaining the change in reading scores. However, the violation of the homogeneity of variances and the normality of residual assumptions means that the ANCOVA results may not be fully reliable.

5. Math scores from fall to spring term vary across different income groups

Research Question #2: How do changes in students' math scores from fall to spring term vary across different income groups, after controlling for their reading scores at the start of the fall term?". follow the requirement

	Source	SS	DF	F	p-unc	np2
0	incomegroup	1712.758286	2	18.523585	9.284861e-09	0.003096
1	fallmathscore	4806.990621	1	103.975792	2.576573e-24	0.008641
2	Residual	551499.442116	11929	NaN	NaN	NaN

Figure7 *ANCOVA table for math score in fall term*
In our examination of the impact of income groups on the change in students' math scores from fall to spring, controlling for their fall math scores, a one-way ANCOVA was conducted. The analysis revealed that both the covariate, the fall math scores ($p < 0.001$, $np^2 = 0.00864$), and the factor of income groups ($p < 0.001$, $np^2 = 0.00396$) significantly affected the math score changes. This indicates a discernible difference in how math scores evolved during the academic year across income groups, even after adjusting for the students' initial math scores.
The F-statistic for the income group was significant, suggesting that the math score changes are not uniform across all income groups. The effect size, as measured by partial eta squared, was small but meaningful, indicating that income group accounts for some variation in math score change.

The interaction plot figure 8 further underscores these results, showing the relationship between fall math scores and math score change across income groups. The positive slopes suggest that students with higher initial math scores tended to have greater increases in their scores, irrespective of their income group. However, the steeper slope for the highest income group indicates a potential interaction effect where students from higher income backgrounds may experience more pronounced progress.

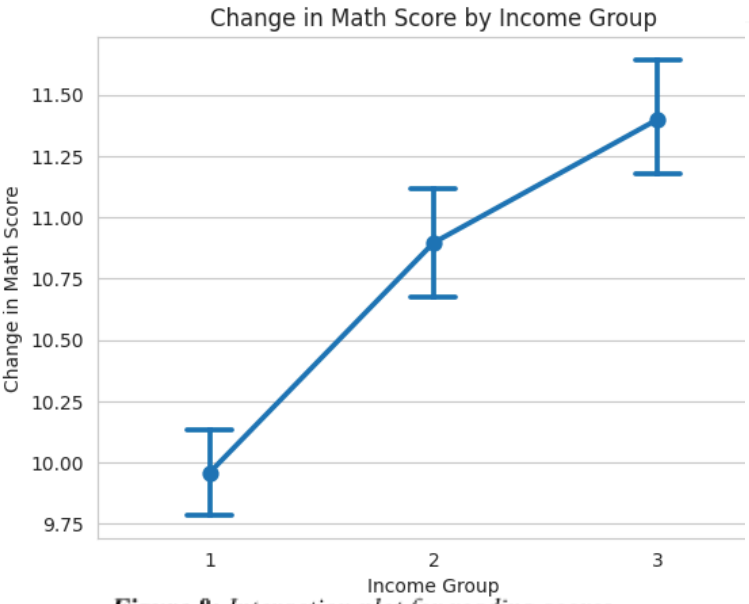

Figure 8: *Interaction plot for reading scores*

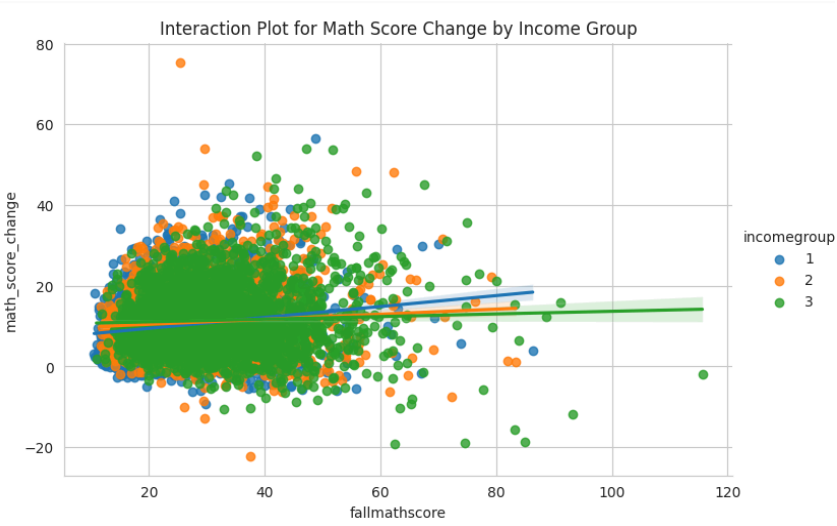

Figure 9: *Interaction plot for reading scores*

Assumption Checks:

Linearity of the Relationship: The provided scatter plot illustrates the relationship between fall math scores and math score change for the various income groups. The fitted lines across groups 1 (blue), 2 (orange), and 3 (green) suggest a positive relationship between initial fall math scores and subsequent math score change. Despite the scatter of points, the linear relationship appears consistent across income groups, suggesting that higher initial math scores are generally associated with larger increases in math scores, irrespective of income group. However, it's worth noting that the slopes of the lines differ among the income groups, hinting at the possibility that the extent to which initial math ability predicts the change in scores may vary depending on the students' socioeconomic status. This variation is significant as it suggests that income groups may interact with initial math ability to influence the change in math scores over the school year.

Homogeneity of Variance (Levene's Test): Levene's test for homogeneity of variances in math score changes across income groups returned a p-value of 1.6601388771886818e-10, which is significantly less than the conventional alpha level of 0.05. This significant result indicates that the assumption of homogeneity of variances has been violated. In practical terms, this suggests that the variances in math score changes are not consistent across all income groups, which is a vital consideration when interpreting the results of ANCOVA.

Normality of Residuals (Shapiro-Wilk Test): The Shapiro-Wilk test for normality of residuals in the math ANCOVA model reports a p-value of <0.001, indicating a significant deviation from normality. This suggests that the residuals of the model are not normally distributed, which violates another fundamental assumption of ANCOVA. This violation implies that the results of the ANCOVA, including the significance tests and confidence inter

6. Conclusion

In summary, while the analysis reveals significant effects of both income group and initial math ability on the change in math scores, the violation of key assumptions of ANCOVA—homogeneity of variances and normality of residuals—calls for caution in interpreting these results. Further analysis using alternative methods that are robust to these violations may be necessary to confirm these findings.