Introduction

Educational achievement gaps based on socioeconomic status are a well-documented concern. This analysis seeks to explore these gaps in the context of kindergarten education by examining how improvements in reading and math scores from fall to spring are associated with income groups, controlling for students' baseline general knowledge scores.

Research Questions

1. How does improvement in reading scores from fall to spring vary by income group, after controlling for baseline general knowledge scores?
2. How does improvement in math scores from fall to spring vary by income group, after controlling for baseline general knowledge scores?

Hypotheses

Reading Scores

- H0: There is no difference in the improvement of reading scores from fall to spring between income groups, controlling for baseline general knowledge scores.
- HA: There is a difference in the improvement of reading scores from fall to spring between income groups, controlling for baseline general knowledge scores.

Math Scores

- H0: There is no difference in the improvement of math scores from fall to spring between income groups, controlling for baseline general knowledge scores.
- HA: There is a difference in the improvement of math scores from fall to spring between income groups, controlling for baseline general knowledge scores.

Data Cleaning

Step 1: Null Value Assessment

An initial examination was conducted to identify any missing or null values across the dataset's columns. This step is crucial for determining the completeness of the data.

Findings: No null values were detected, indicating that the dataset was complete and all entries across the variables of interest were filled.

Step 2: Removal of Redundant Columns

The dataset was scrutinized for any redundant information that could clutter the analysis. Redundancy in data can lead to confusion and misinterpretation of results.

Action Taken: The column incomeinthousands was identified as redundant since it provided the same information as totalhouseholdincome but in a different scale. It was removed to streamline the dataset.

Step 3: Derivation of New Variables

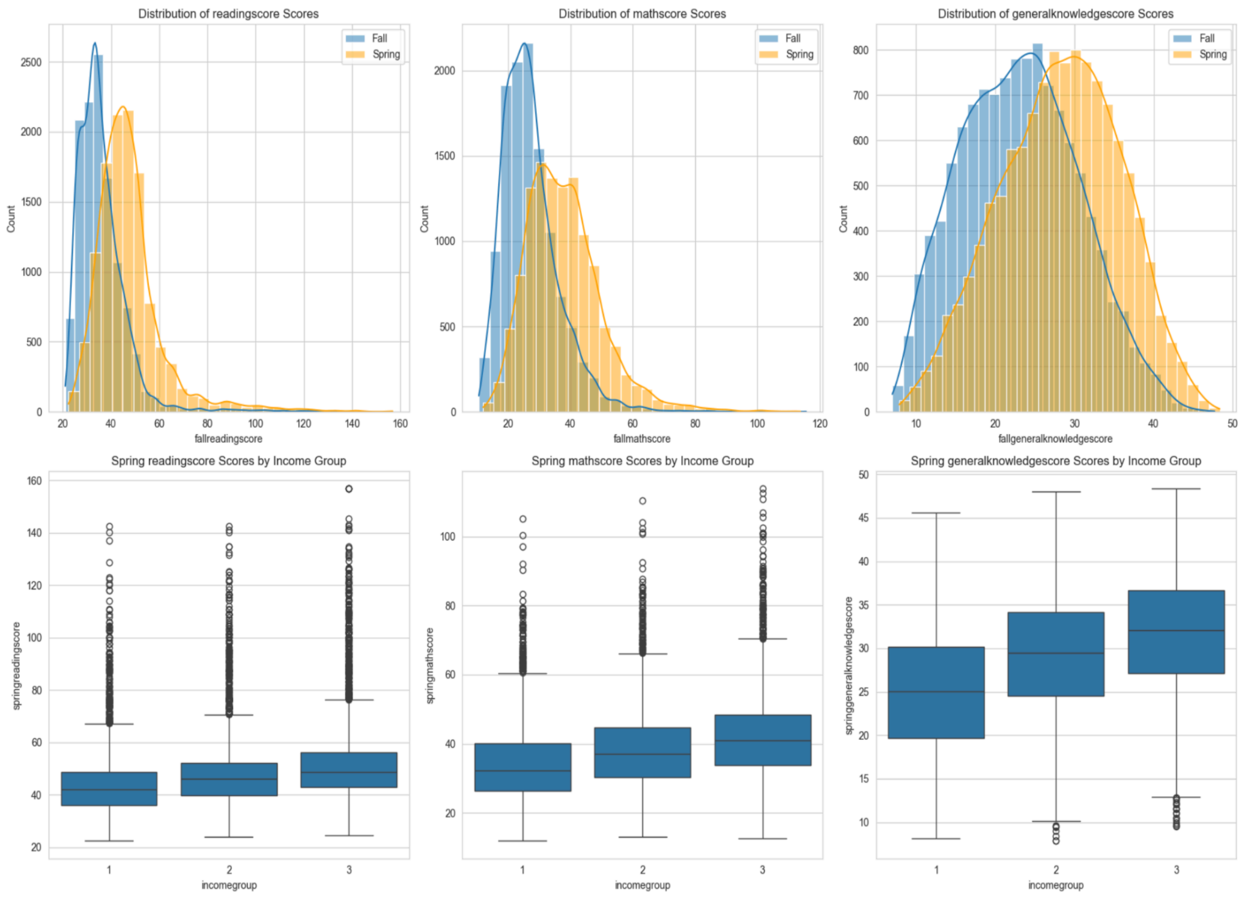
For the purpose of this analysis, it was necessary to compute the improvement in scores from fall to spring semesters for both reading and math.

New Variables Created: reading_improvement: Calculated as the difference between springreadingscore and fallreadingscore.

math_improvement: Calculated as the difference between springmathscore and fallmathscore. These variables serve as the dependent variables in our ANCOVA analysis, directly measuring academic improvement over the school year.

EDA

A preliminary exploratory data analysis (EDA) was conducted to examine the distributions of scores and income groups and ensure the dataset did not contain anomalies or outliers that could skew the analysis. The dataset comprises fall and spring scores for reading, math, and general knowledge for 11,933 kindergarten students, along with their household income information. Income groups were derived based on household income levels.



Descriptive Statistics

Scores: The analysis of fall and spring scores for reading, math, and general knowledge revealed improvements across all areas from fall to spring, indicating general academic growth over the school year.

Income Data: The totalhouseholdincome column was explored to understand the economic background of the students' families. The income data, categorized into three groups in the incomegroup column, showed a distribution that suggests a diverse socioeconomic composition of the student population.

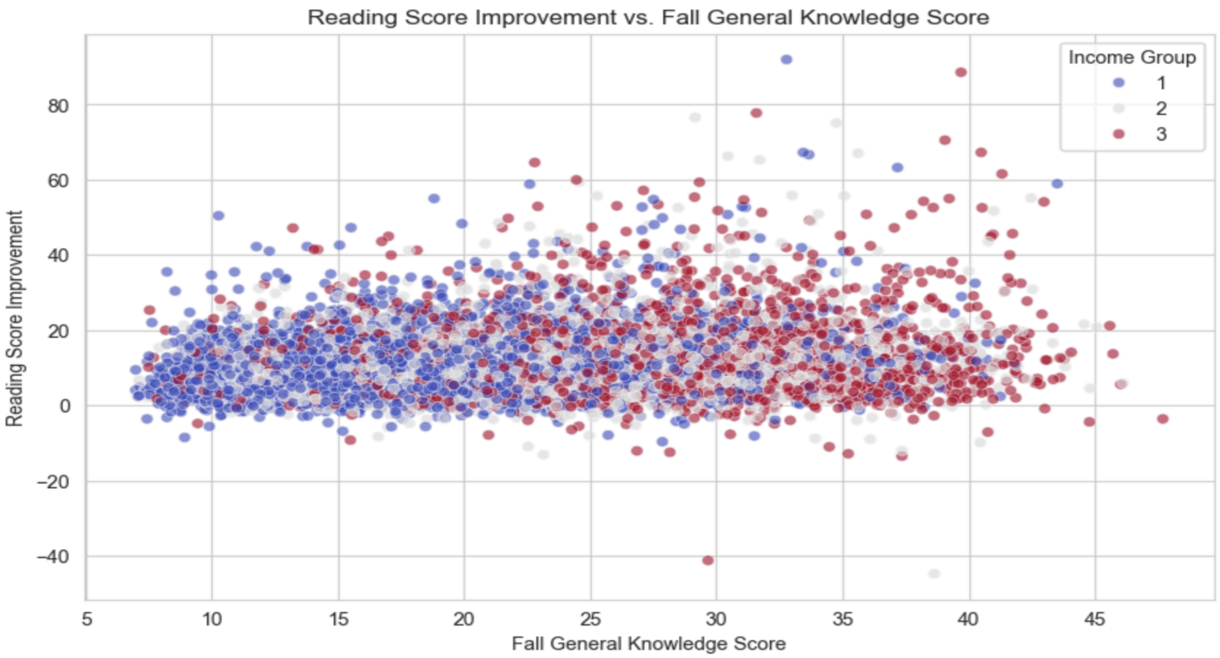
Distribution Analysis

Score Distributions: Histograms of fall and spring scores for reading, math, and general knowledge depicted a generally right-skewed distribution, with most students showing moderate scores but with a long tail of higher achievers.

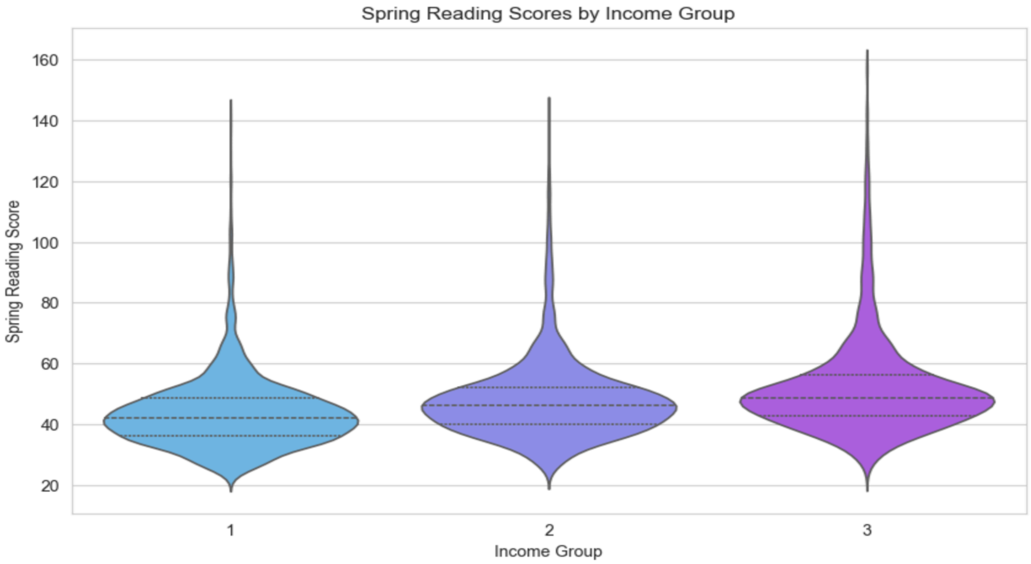
Income Group Distribution: Analysis of the incomegroup column indicated a relatively even distribution across the three income groups, with a slight overrepresentation of the middle-income group. This distribution is critical for ensuring our analysis covers a broad spectrum of socioeconomic statuses.

Correlation Analysis

Scores and Income Group: A preliminary correlation analysis was conducted to explore the relationship between income group and score improvements. Scatter plots and violin plots hinted



at a positive correlation, suggesting that students from higher-income groups tend to show greater improvements in scores. However, this observation required further statistical testing to confirm.



Preliminary Insights

Socioeconomic Influence: The EDA suggested a potential influence of socioeconomic status on academic improvement, aligning with expectations based on existing educational research.

Importance of Baseline Knowledge: The comparison of fall and spring scores highlighted the significant role of baseline knowledge in academic development, suggesting that students with higher baseline scores tend to show more substantial improvements.

Result

Reading Scores Improvement Analysis

Income Group Effect: There is a statistically significant association between income group and reading score improvement (p=0.034). Specifically, each increase in income group is associated with an additional 0.20-point increase in reading score improvement, indicating that students from higher-income families tend to show more significant improvements.

Baseline General Knowledge: Baseline general knowledge scores also significantly affect reading score improvement (p<.0001), with each point increase in baseline score associated with a 0.16-point increase in reading score improvement. This result underscores the importance of early general knowledge in reading development.

Model Insights: The ANCOVA model for reading score improvement showed that both the income group and the baseline general knowledge scores significantly influenced the reading score improvements. The positive coefficient for the income group suggests that being in a higher income group is associated with greater improvement in reading scores over the school year. This finding highlights the potential impact of socioeconomic factors on educational achievement, even at the kindergarten level.

Effectiveness of the Model: The model effectively demonstrated the relationship between income group and reading score improvement while controlling for the influence of baseline knowledge. However, the R-squared value indicates that a relatively small portion of the variance in reading score improvement is explained by these variables, suggesting that other factors not included in the model may also play significant roles.

Math Scores Improvement Analysis

Income Group Effect: The income group's effect on math score improvement was significant (p<.0001), but with a smaller coefficient (0.0752) than observed in reading scores. This indicates a positive but less pronounced relationship between income group and math score improvement.

Baseline General Knowledge: As with reading scores, baseline general knowledge significantly predicted math score improvement (p<.0001), with a 0.1996-point increase in math score improvement for each additional point in baseline general knowledge score.

Model Insights: Similar to the findings for reading scores, the ANCOVA model for math score improvement indicated that income group and baseline general knowledge scores significantly affected math score improvements. The significance of baseline general knowledge scores in both models underscores the foundational role of early general knowledge in academic development across different subject areas.

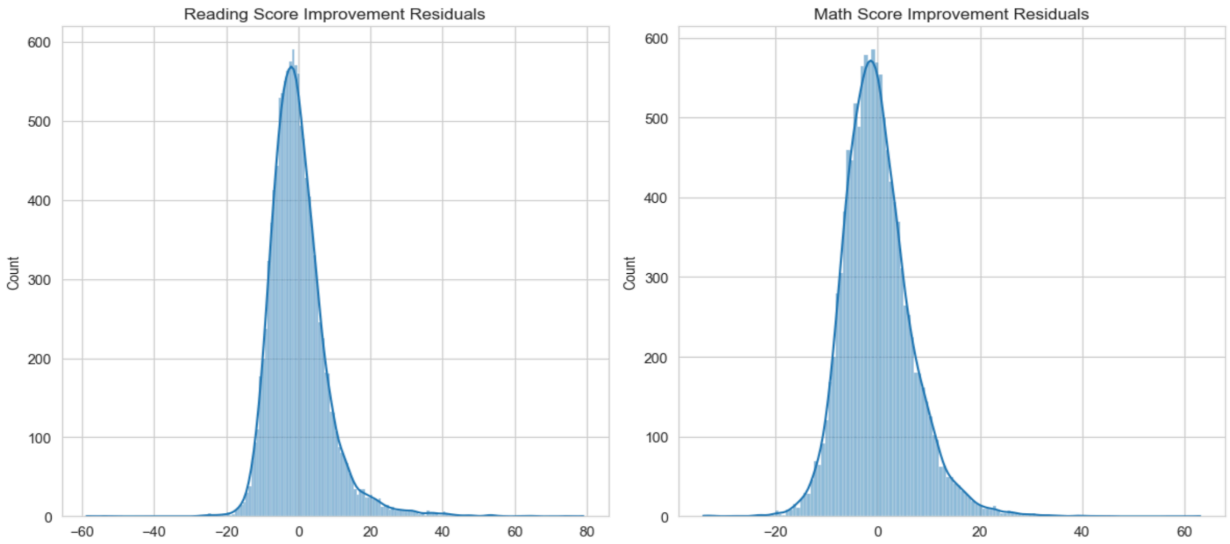
Effectiveness of the Model: The math score improvement model, like its reading counterpart, highlights the significance of socioeconomic status and early knowledge. The differences in the strength of association between income group and score improvements in reading versus math may indicate subject-specific influences of socioeconomic factors, an area ripe for further investigation.

Discussion

The results from both ANCOVA analyses highlight the critical role of socioeconomic status, as approximated by income group, in the academic development of kindergarten students. The findings corroborate existing literature that suggests children from higher-income families often have access to resources and environments conducive to more substantial academic growth.

Moreover, the significant effect of baseline general knowledge on both reading and math score improvements underscores the importance of early childhood education and the development of general knowledge before formal schooling begins. This could imply that interventions aimed at enriching children's early learning environments might yield significant improvements in academic outcomes.

Interestingly, while both reading and math improvements were positively associated with income group and baseline knowledge, the magnitude of these relationships differed between the subjects. This variance suggests subject-specific dynamics in how socioeconomic factors and early knowledge influence learning outcomes, warranting further investigation.



The analysis, while insightful, is not without limitations. The assumptions of ANCOVA, particularly concerning the normality of residuals and homogeneity of variances, were challenged in this dataset. Future research could benefit from employing robust statistical methods or non-parametric alternatives to address these concerns. Additionally, exploring other potential covariates and their interactions could provide a more nuanced understanding of the factors influencing academic improvement in early childhood.

Summary

This analysis emphasizes the impact of socioeconomic status on educational outcomes in early childhood, particularly in reading achievement. It underscores the importance of addressing educational disparities from a young age and suggests that interventions aimed at enhancing general knowledge could be beneficial across income groups.

This study acknowledges limitations, including the assumption violations in ANCOVA and the inability to account for all potential confounding variables. Future research could explore longitudinal data to assess long-term impacts and employ models that can accommodate non-normal distributions and heteroscedasticity.