# Introduction

Researchers have been studying how family income affects early learning outcomes in order to inform policies and programs that aim to improve equitable educational opportunities for society more broadly. This study aims to explore the relationship between children's family income levels and their performance in kindergarten by examining test scores in reading, math, and general knowledge. We will use a dataset from the 1998-99 Early Childhood Longitudinal Study, which records these test scores for kindergarten students over several months. By applying ANCOVA (Analysis of Covariance), we plan to analyze how income influences learning achievements, adjusting for other factors that might affect these outcomes.

Our goal is to highlight the importance of addressing income disparities to support all children's educational success from an early age, based on the following research questions:

**Q:** How does household income impact the academic progression of students in kindergarten within a specified period of time?
**Q:** Does the relationship between household income and test scores differ by subject (reading, math, general knowledge)?
**Q:** Are there some potential differential income effects across various academic subjects?

# Data and Methodology

## Data Source

Our dataset was sourced from a dataset provided by the Faculty of Information at the University of Toronto for the INF2178 course in Winter 2024. It consists of 11,933 entries, including 9 columns and 11,933 rows correspondingly. We isolated our variables of concern from the dataset to focus primarily on the following:

- **Springgeneralknowledgescore:** The scores kindergarten students achieved on a general knowledge test taken in the spring.
- **Springmathscore**: Scores kindergarten students achieved on mathematics tests taken in the spring period.
- **Totalhouseholdincome:** The total amount of money earned by all members of a student's household within a year.
- **Income Group:** Categorizes families based on their total household income into different brackets. Helps to compare the economic backgrounds of students. 3 = high, 2 = medium, 1 = low.

We chose **Springgeneralknowledgescore** as a dependent variable to assess the knowledge that students have acquired by the end of the kindergarten year across various subjects, providing a comprehensive measure of their general cognitive development.

**Springmathscore** was chosen as another dependent variable reflecting the progress and proficiency in mathematical skills that students demonstrate at the conclusion of kindergarten, providing insight into the relationship between socioeconomic status and early numeracy development

**Totalhouseholdincome** was selected as a continuous covariate in our analysis to quantify the economic background of the students' families, offering a nuanced view of the socio-economic status that might correlate with academic performance in kindergarten.

**Income Group** was included as an independent categorical variable, dividing the sample into high, medium, and low tiers based on household income. This helps us check the presence of any threshold effects where income bands may differentially impact educational outcomes and to evaluate whether the effects of income on academic performance are consistent or vary across these predefined income categories.

### Data Preparation

Data preparation involved analyzing the basic contents of the dataset itself, before performing basic visual analysis to better understand the structure of our data. We did not perform extensive data cleaning or transformation since the dataset was presumed to be in a usable state for the intended analysis, and did not use any transformations such as melting or pivoting the data for this assignment.

### Statistical Analysis Methodology

Our project methodology uses ANCOVA analyses, including a thorough examination of the assumptions required for ANCOVA and subsequent post-hoc tests on the findings. ANCOVA helps assess the influence of both categorical independent variables and continuous covariates on a continuous dependent variable, allowing us to more deeply understand the relationships among variables. Specifically, ANCOVA allows adjustment for potential confounders and helps us examine the impact of categorical factors (such as income groups) on educational outcomes, controlling for continuous variables like total household income. This not only quantifies the effect of categorical variables on continuous outcomes, but also enriches our understanding of how these effects may be moderated or influenced by the presence of continuous covariates.

## Exploratory Data Analysis

This section will help examine, summarize, and visualize the main characteristics of our dataset. It will help provide a brief overview of some trends and patterns through descriptive and graphical visualizations.

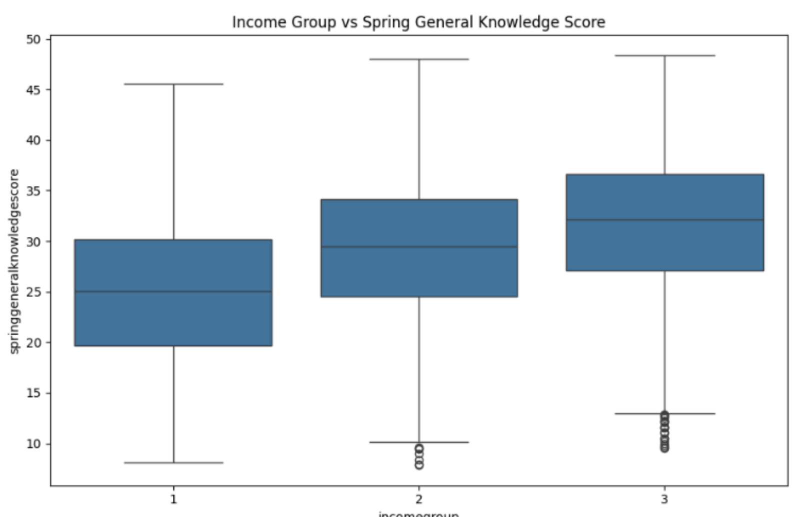


Figure 1: Income Group vs Spring General Knowledge Scores

The boxplot shows that there is a positive association between income group and general knowledge scores, with median scores increasing from the low to the high-income group. The lowest income group shows the greatest variability in scores, and the medium and high-income groups show less variability. The distribution within the middle income group is more symmetric and less varied than in the low-income group. This suggests that higher income may be associated with both higher general knowledge scores and more consistency in test score outcomes.

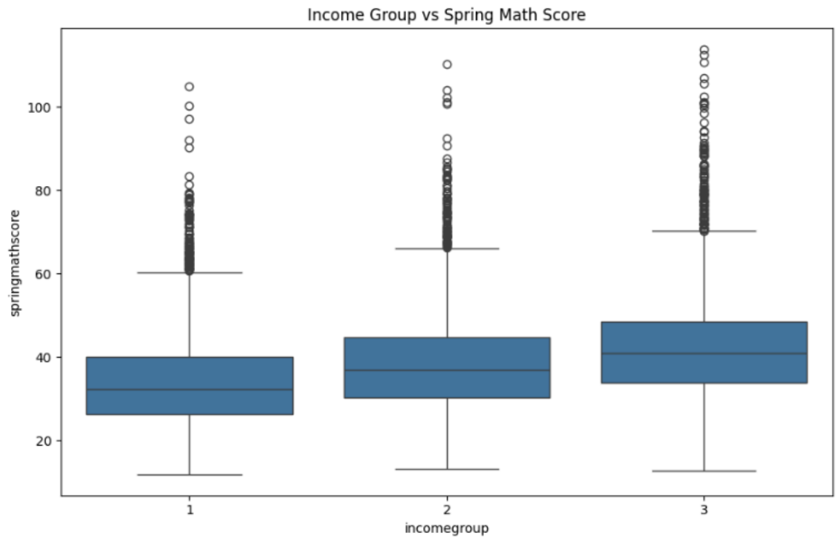


Figure 2: Income Group vs Spring Math Score

Figure 2 meanwhile shows that across three income groups,the central tendency, as indicated by median scores, is relatively consistent across groups. This means that the average math ability is comparable regardless of income group classification. The variability within each income group, as demonstrated by the interquartile range (IQR), is similar across the groups, suggesting a consistent spread of scores within low, medium, and high-income categories. A significant number of outliers are present for all groups, pointing to the existence of students with math scores much higher than typical within each income level. This could reflect differences in other factors outside those in the dataset, such as educational support, innate student ability, or other unknown variables. Despite the differing income brackets, the median scores do not show a stark gradient, implying that the central tendency for math achievement is akin across the income spectrum. The approximately equivalent IQRs across groups suggest that income level does not contribute majorly to variability in Spring Math Scores within this dataset.

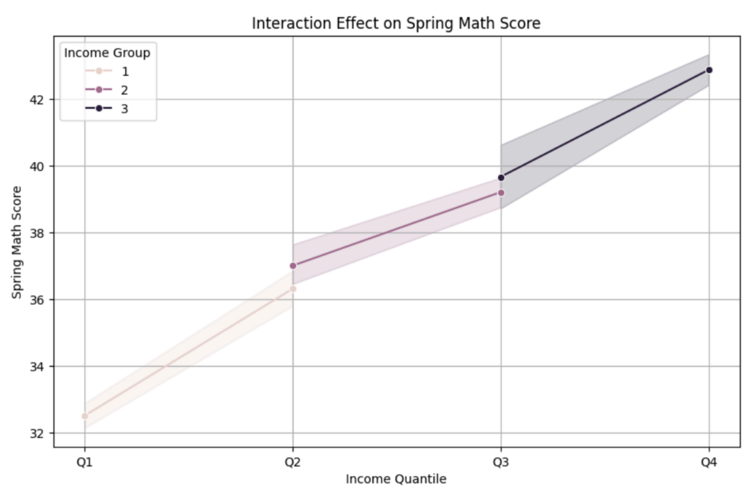


Figure 3: Interaction Plot for Spring Math Score

Figure 3 illustrates the interaction between Spring Math Scores and income quantiles across different income groups. The non-parallel lines suggest an interaction effect between income group and income quantile on math scores. Specifically, for all income groups, scores increase with higher income quantiles, but the rate of increase is more pronounced in the higher income groups.

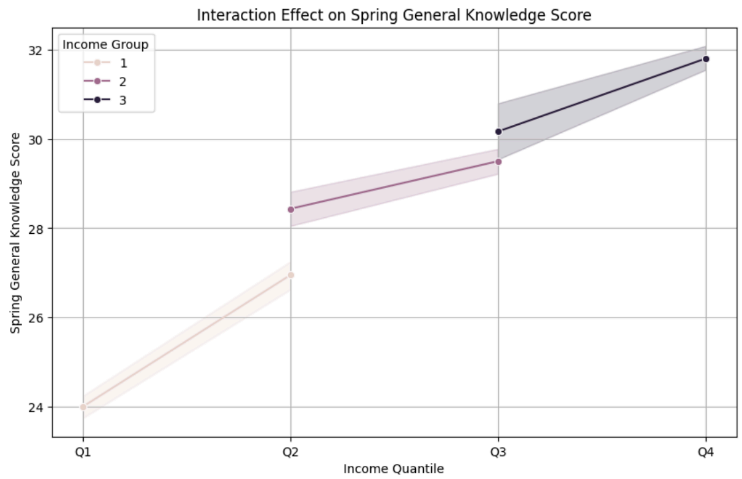


Figure 4: Interaction Plot: Spring General Knowledge Score

Figure 4 meanwhile shows the interaction between Spring General Knowledge Scores and income quantiles within different income groups. The ascending lines for all groups indicate that general knowledge scores increase with higher income quantiles. The slope of the line for Income Group 1 is steeper than those for Groups 2 and 3, which suggests a positive relationship between income and general knowledge scores for the lowest income category. This implies that increments in income within the lower bracket could be associated with substantial gains in general knowledge, possibly due to a higher marginal benefit of additional resources or opportunities. In addition, for Income Groups 2 and 3, while scores increase with income, the effect is less steep. This shows that higher income levels still benefit general knowledge scores, but the incremental advantage decreases as income rises.

# Results

### ANCOVA: Spring General Knowledge Scores

| Source | Sum of Squares | Degrees of Freedom | F-Statistic | P-Value | Partial ETA-Squared |
|---|---|---|---|---|---|
| incomegroup | 8159.5065 | 2 | 82.3213 | 1.159955 e-36 | 0.013777 |
| totalhousehold income | 11924.0708 | 1 | 243.5268 | 2.3049 e-54 | 0.0200 |
| Residual | 584092.8775 | 11929 | NaN | NaN | NaN |

The ANCOVA results for General Knowledge Scores indicate a definitive influence of economic factors on Spring General Knowledge Scores, as evidenced by significant F-values and near-zero p-values for both income group and total household income. The income group explains a notable portion of score variation, while total household income adds further differentiation within these groups. However, substantial variability remains unexplained by these factors, pointing to other influences beyond income on students' knowledge acquisition.

**ANCOVA: Spring Math Scores**

| Source | Sum of Squares | Degrees of Freedom | F-Statistic | P-Value | Partial ETA-Squared |
|---|---|---|---|---|---|
| incomegroup | 8159.5065 | 2 | 82.3213 | 1.159955 e-36 | 0.013777 |
| totalhousehold income | 11924.0708 | 1 | 243.5268 | 2.3049 e-54 | 0.0200 |
| Residual | 584092.8775 | 11929 | NaN | NaN | NaN |

Meanwhile, the results for Spring Math Scores show a significant influence for both the categorical income group and the continuous total household income on students' math scores. With significant F-values and p-values close to zero, it's clear that both income variables play a role in score variability. While the income group's contribution is present, the total household income shows a strong association with math score values. In summary, the ANCOVA overall shows that income, both at the grouped and individual household levels, has a statistically significant relationship with student performance in general knowledge and math, underscoring the importance of economic context in educational outcomes, yet also indicating the presence of other influential factors beyond income.

# Post-Hoc Analysis

This section will evaluate using statistical tests whether the conditions for ANCOVA were met based on the characteristics of our dataset.

## Assumption 1: Homogeneity of Variances

We will use Levene's test to check whether the variances of the scores are equal across the different levels of your independent variable (income group).

**Levene's Test Results**

| Dependent Variable | Test Statistic | P-Value |
|---|---|---|
| Spring General Knowledge Score | 18852.823934 | 0.0 |
| Spring Math Score | 14396.882936 | 0.0 |

As seen above, the Levene's results for both the springgeneralknowledgescore and springmathscore indicate significant p-values of 0.0, suggesting that the variances across income groups are significantly different for both dependent variables. This suggests the dataset violates the assumption of homogeneity of variances required for ANCOVA.

## Assumption 2: Normality of Residuals

We will use Shapiro-Wilk's test to check the normality of the residuals.

**Shapiro-Wilk's Test Results**

| Dependent Variable | Test Statistic | P-Value |
|---|---|---|
| Spring General Knowledge Score | 0.996955 | 5.78217 e-15 |
| Spring Math Score | 0.946478 | 0.0 |

Similar to above, the Shapiro-Wilk and Levene's tests for the two-way ANOVA reveal that two key assumptions, normality of residuals and homogeneity of variances, are not met. With p-values significantly below conventional significance levels, this implies that the assumption of normality is violated for the residuals. Overally, these violations imply that the standard ANOVA results may not be reliable.

# Discussion

Our findings indicate that the data does not satisfy the normality and homogeneity of variance assumptions, casting doubt on the validity of the ANCOVA results.

While our initial ANCOVA output indicated a strong association between income level and students' performance in both general knowledge and math, our post-hoc tests indicate certain risks with the reliability of our results, such as risk of Type I errors, potentially leading to incorrect conclusions about the statistical significance of the results. Non-homogeneity also means that the variances of scores within the income groups are not equal, and non-normality of residuals implies that the distribution of scores does not follow the assumed normal distribution. This implies a need for data transformation or the consideration of alternative statistical methods that do not assume normality of residuals. Furthermore, even while controlling for the impact of covariates, the possibility of other underlying factors explaining the relationship between test scores and income should not be understated. Other studies examining the relationship between academic achievement and family income associations have highlighted other confounding factors such as region, parental educational background and involvement, socio-psychological factors, and so on. Chevalier et al. (2013) for example highlights how parental union status also has certain impacts upon child educational attainment. Overall, we cannot generalize our findings without meeting criteria for our statistical techniques, as well as broadly considering other confounding factors as found in existing literature.

# Conclusion

The ANCOVA findings hint at a strong link between students' income levels and their performance on general knowledge and math tests. However, since the test's requirements were not fully satisfied (eg: groups having different variances and the test scores not spreading out in the expected pattern) we should exercise caution when relying on these results. To be more confident in our findings, it would be advisable to rework the data, try different statistical approaches that may not need these statistical requirements to be fully satisfied, or use other techniques such as bootstrapping, before we can be certain about the impact of economic factors on student performance.

# References

Chevalier, A., Harmon, C., O' Sullivan, V. *et al.* "The impact of parental income and education on the schooling of their children". *IZA J Labor Econ* **2**, 8 (2013). https://doi.org/10.1186/2193-8997-2-8

Phipps, Shelley, and Lynn Lethbridge. n.d. "Income and the Outcomes of Children." *Analytical Studies Branch Research Paper Series.* Statistics Canada. Accessed March 31, 2024. https://www150.statcan.gc.ca/n1/en/pub/11f0019m/11f0019m2006281-eng.pdf?st=KchUstx_.