Kevin Nitièma | 1002673298
INF2178H: Experimental Design for Data Science – Assignment 3

Dissecting Disparities in Early Childhood Education

1. Introduction

Access to quality education in early childhood is a crucial factor in shaping future opportunities for children. Nonetheless, disparities in accessing equitable education persist with socioeconomic factors. To explore this topic, this study investigates the relationship between household income and academic performance among Kindergarten students using data from a longitudinal study conducted between 1998-99.

Our analysis focuses on understanding how income levels and prior academic performance may influence reading, math, and general knowledge scores among students over the course of a school year. We aim to address any variations in academic performance across income groups and explore whether income plays a role in academic growth and success over time.

Our exploration will address three (3) fundamental research questions, serving as guiding principles in discovering the nature of educational disparities. The research questions are as follows:

1. **Research Question 1:** Does the change in general knowledge scores from fall to spring evaluations vary based on income groups?
2. **Research Question 2:** How does income group predict spring math scores among kindergarten students, after controlling for their fall math scores?
3. **Research Question 3:** Is there a significant difference in general knowledge scores between fall and spring evaluations, when controlling for household income?

By addressing these questions, we hope to contribute to the discussion on educational equity and inform evidence-based interventions to promote equal opportunities for all children. Through a comprehensive analysis, our findings will hopefully help advocate for policies that prioritize the educational needs by ensuring that every child has access to quality education regardless of socioeconomic background.

2. Data Cleaning and Data Wrangling

The raw dataset includes **11,933 entries (rows)** and **9 features (columns)**. Minimal data cleaning was required initially, but we identified discrepancies and introduced new features for analysis. Below, we summarize our observations and the new feature(s):

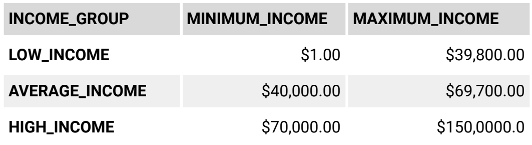
A. **Observations and Considerations:**

1. Recognizing the quantitative nature of our analysis, we'll be working with most columns from the raw dataset. Below we provide a brief description of each column of the raw dataset:
   - `FALL_READING_SCORE:` Students' reading score in fall.
   - `FALL_MATH_SCORE:` Students' math score in fall.
   - `FALL_GK_SCORE:` Students' general knowledge score in fall.
   - `SPRING_READING_SCORE:` Students' reading score in spring.
   - `SPRING_MATH_SCORE:` Students' math score in spring.

    - `SPRING_GK_SCORE:` Students' general knowledge score in spring.
    - `TOTAL_HOUSEHOLD_INCOME:` Total household income of students' families.
    - `INCOME_THOUSANDS:` Total household income, in thousands of dollars.
    - `INCOME_GROUP:` Categorical variable indicating the income group of each household, based on total household income.

2. The `INCOME_THOUSANDS` column exhibited mixed formatting with both integer and float values. To standardize the data, we converted the column to solely contain integer values.

3. The `INCOME_GROUP` feature was reworked by replacing numeric labels (1, 2, and 3) with more descriptive categories: **`'LOW_INCOME'`**, **`'AVERAGE_INCOME'`**, and **`'HIGH_INCOME'`**, respectively. As documented throughout our study, we acknowledge that the income range provided may not precisely correspond to conventional socioeconomic classifications of low, average, and high-income brackets. Nonetheless, for the sake of simplicity in this study and for the convenience of our readers, we have adopted these labels. (*Figure 1*)



INCOME_GROUP	MINIMUM_INCOME	MAXIMUM_INCOME
LOW_INCOME	$1.00	$39,800.00
AVERAGE_INCOME	$40,000.00	$69,700.00
HIGH_INCOME	$70,000.00	$150,0000.0

Figure 1: Table of income range by income group.

B. Feature Engineering:

Here, we created **two (3) new features** to add to our dataset to aid in later analysis (we also later worked with additional features which we address in our report). The features are as follows:

1. `DELTA_GK_SCORE:` this feature represents the difference in the General Knowledge (GK) scores between two evaluation periods: spring and fall. This difference is calculated by subtracting the General Knowledge scores obtained during the fall evaluation (`FALL_GK_SCORE)` from those obtained during the spring evaluation (`SPRING_GK_SCORE)`. The same principle applies to the following features: `DELTA_READING_SCORE, DELTA_MATH_SCORE`.
    - $DELTA\_GK\_SCORE = SPRING\_GK\_SCORE - FALL\_GK\_SCORE$

3. Exploratory Data Analysis (EDA)

After cleaning and enhancing our dataset with additional information, we dove into a thorough EDA to uncover valuable insights that could lead to **interesting research questions**. We began by exploring the numerical aspects of our data and visually representing it through plots, including boxplots (as depicted in *Figure 2*), histograms (detailed in the accompanying notebook/code), and heatmap correlation matrices (showcased in *Figure 3*). These data visualization strategies allowed us to identify hidden patterns and trends within the dataset. For a more detailed account of our exploration process, please refer to the provided code.

During our EDA, we noticed a consistent normal distribution among most continuous variables, more precisely, the scores of students across various disciplines and evaluation periods. However, there was a slight right skewness observed in the distribution of income, which aligns with typical socioeconomic trends.

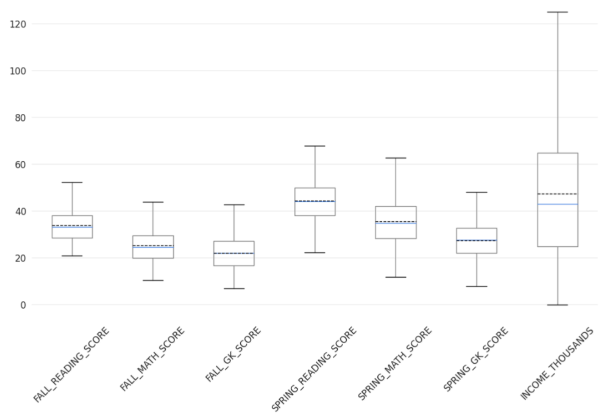


Figure 2: Boxplot of Continuous Variables

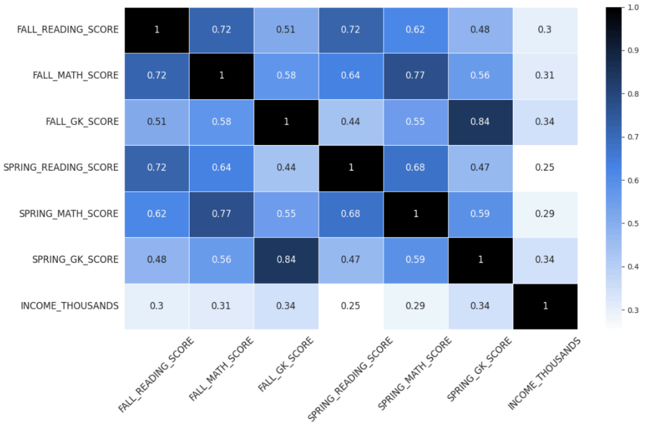


Figure 3: Correlation Matrix Heatmap

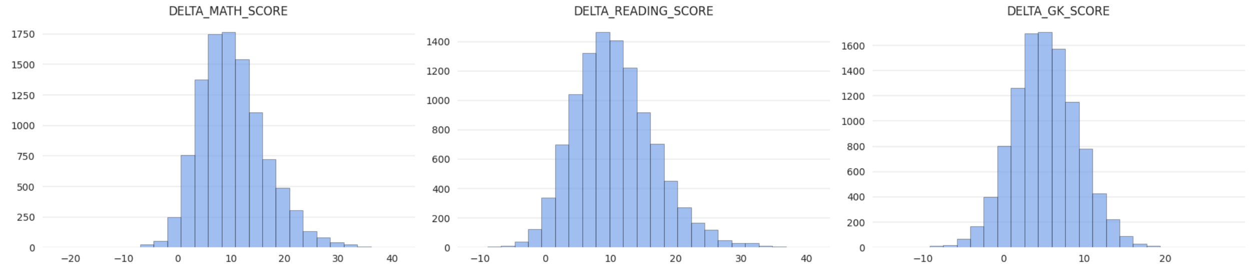


Figure 4: Distribution Net Performance (spring-fall) across Disciplines

Moreover, upon examining the correlation matrix heatmap, we identified **strong positive correlations** ($r > 0.7$) among all disciplines when comparing fall and spring evaluation periods *(Figure 2)*. This finding led us to formulate a hypothesis suggesting the existence of a predictable relationship between fall scores and spring scores.

Research Questions #1 (EDA): Does the change in general knowledge scores from fall to spring evaluations vary based on income groups?

In addition to analyzing the normality of the dataset, we examined the distribution of score differences – i.e., `DELTA_MATH_SCORE, DELTA_READING_SCORE, DELTA_GK_SCORE` – across disciplines *(Figure 4)*. Interestingly, our investigation revealed nearly perfect normal distributions for these differences. However, a notable observation was that none of these distributions were centered at 0; rather, most were **significantly centered above 0.** This finding implies that students exhibited **improvements in their scores across all disciplines** between the fall and spring evaluation periods.

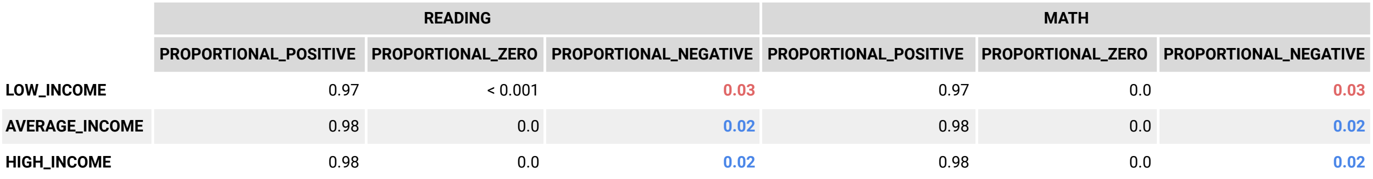
To further explore this hypothesis, we investigated the number of students who achieved net positive, net zero, and net negative scores across all disciplines (reading, math, and general knowledge) segmented by income group. *Figure 5* illustrates these counts specifically for general knowledge.

INCOME_GROUP	TOTAL_STUDENTS	NET_POSITIVE	NET_ZERO	NET_NEGATIVE	PROPORTIONAL_POSITIVE	PROPORTIONAL_ZERO	PROPORTIONAL_NEGATIVE
LOW_INCOME	4540	4085	0	455	0.90	0.0	0.10
AVERAGE_INCOME	3460	3164	0	296	0.91	0.0	0.09
HIGH_INCOME	2446	2213	1	232	0.90	< 0.001	0.09

Figure 5: Table of Net Performance and Net Proportional Performance for General Knowledge Score

The table illustrates significant net positive scores across all income groups. However, a noteworthy observation is the larger count of net negative scores for lower-income students $count = 455$, especially when contrasted with the 296 and 232 negative scores recorded for average and high-income students, respectively.

Given the varying sample sizes across income groups, we standardized net performance scores by analyzing their proportional frequencies, as shown in *Figure 5* and *Figure 6*. In reading and math assessments, around 3% of lower-income students scored negatively, while about 2% did so in average and higher-income groups. This pattern extends to general knowledge, with roughly 1% of lower-income students scoring negatively, compared to 0.09% in average and higher-income groups. Although small at first glance, these trends indicate disparities in student achievement based on income level and evaluations period.



	READING			MATH		
	PROPORTIONAL_POSITIVE	PROPORTIONAL_ZERO	PROPORTIONAL_NEGATIVE	PROPORTIONAL_POSITIVE	PROPORTIONAL_ZERO	PROPORTIONAL_NEGATIVE
LOW_INCOME	0.97	< 0.001	0.03	0.97	0.0	0.03
AVERAGE_INCOME	0.98	0.0	0.02	0.98	0.0	0.02
HIGH_INCOME	0.98	0.0	0.02	0.98	0.0	0.02

Figure 6: Table of Net Proportional Performance for Reading and Math Scores

Next, we plotted the mean score of each income level across evaluations to further investigate how income could potentially play into student academic success rate *(Figure 7)*. As expected, a consistent hierarchy emerged in score rankings. Consistently, lower-income students attained lower scores compared to average-income students, who in turn scored lower than higher-income students across all assessments.

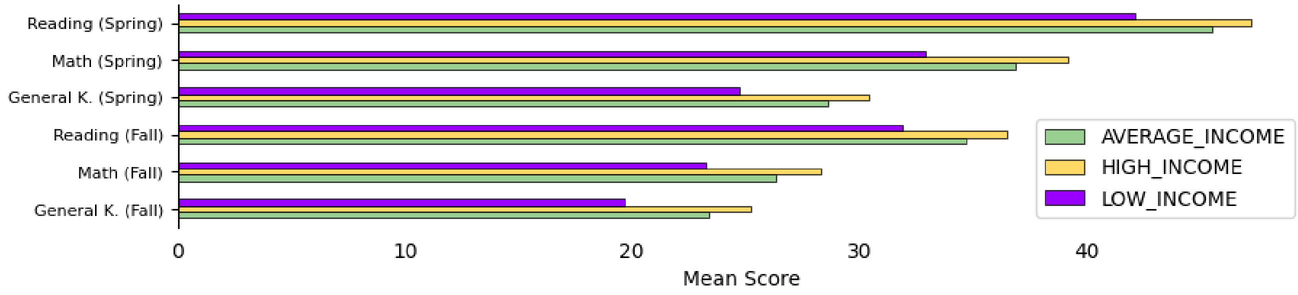


Figure 7: Bar chart of Income-based Mean Test Scores across Disciplines and Evaluation Period

4. Math Education: The Role of Income and Prior Achievement

Research Question #2: How does income group predict ,pring math scores among kindergarten students, after controlling for their Fall math scores?

By exploring this question, we aim to understand how **wealth levels** (`INCOME_GROUP`) and **prior knowledge** (`FALL_MATH_SCORE`) contribute to the variation in spring math scores (`SPRING_MATH_SCORE`) among kindergarten students. To explore this question, we conducted a **one-way ANCOVA model**, controlling for `FALL_MATH_SCORE` (covariate) to isolate the effect of income group on spring math scores.

Note: Interaction Plots included in Notebook/code.

When outlining the assumptions and conditions of our model, we took into account the following:

1. ✔ **Homogeneity of Variance:** we conducted Levene's test on our dataset which yielded $p - value = 0.99$ with $\alpha = 0.05$. We conclude that our model meets this assumption.

2. ✔ **Normal Distribution of Residual:** we plotted the residuals' distribution (*Figure 9*) which suggested normality. As a disclaimer, Shapiro-Wilk tests showed significant departure from normality. However, this discrepancy likely stems from the test's limitations with very large sample sizes ($n > 5000$).
3. ✖ **Homogeneity of the Regression Slopes:** our model suggests an interaction effect between the fall math score and income group on the spring math score ($p - value < 0.001$), which implies a violation of the assumption of homogeneity of regression slopes (*Figure 8*). Nonetheless, we discuss below how despite this violation this model provides valuable insights and helps us dive into further analysis.

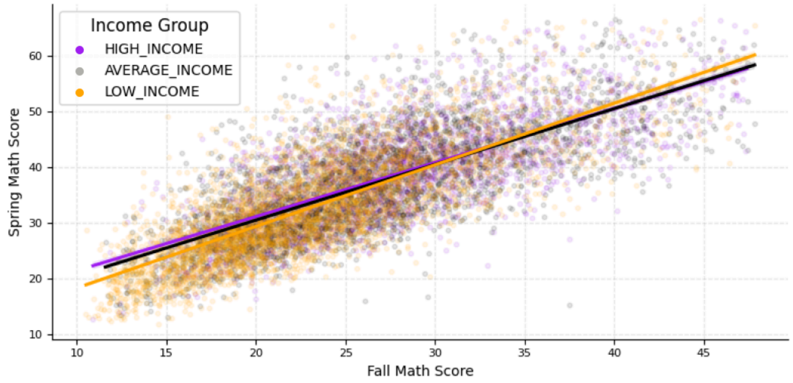


Figure 8: Fall Math Score vs. Spring Math Score by Income group (Regression Slopes)

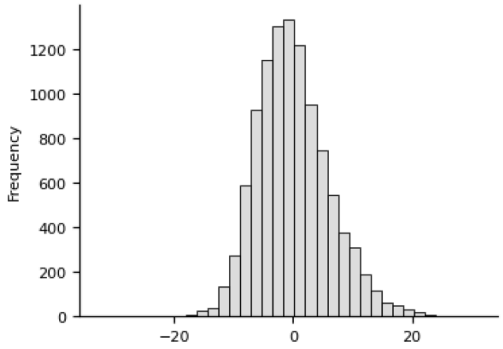


Figure 9: Residuals Histogram Plot

	coef	p-value
Intercept (Low Income)	8.9119	< 0.001
C(INCOME_GROUP_NUM)[T.2] (Average Income)	0.6787	< 0.001
C(INCOME_GROUP_NUM)[T.3] (High Income)	0.9864	< 0.001
FALL_MATH_SCORE	1.0346	< 0.001

R-squared	Adj. R-squared:
0.601	0.600

Figure 9: One-Way ANCOVA Results

Results:
The model's overall fit is strong, with $R^2 = 0.601$, indicating that **approximately 60.1%** of the variability in spring math scores is explained by the model. Both independent variables, i.e., `INCOME_GROUP` and `FALL_MATH_SCORE`, show statistically significant coefficients ($p - value < 0.001$) at a significance level of $\alpha = 0.05$. This suggests that both **socioeconomic status** (`INCOME_GROUP`) and **prior academic performance** (`FALL_MATH_SCORE`) are significantly associated with spring math scores.

Despite the violation of the Homogeneity of the Regression Slopes assumption, the main effects of income group and fall math score remain significant predictors of spring math scores. However, the interpretation of the **interaction term becomes more complex** and may require additional analysis to gauge the impact of socioeconomic factors, like income bracket, on students' general academic success.

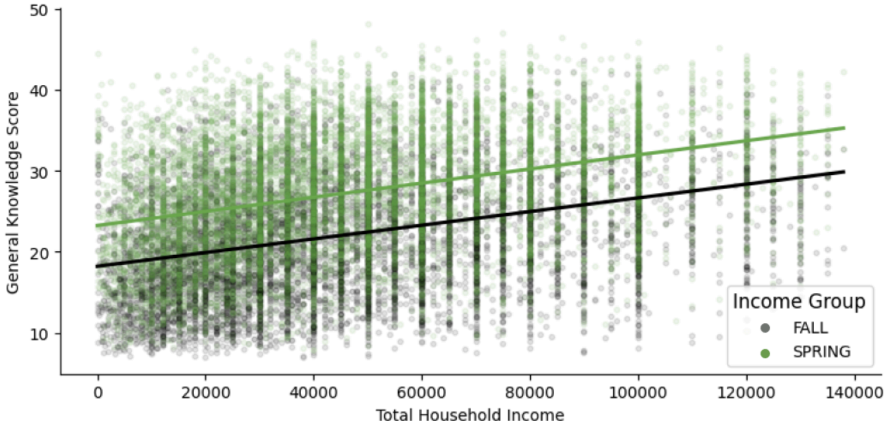
5. Academic Success: Combining Household Income & Sessional Evaluation

Research Question #3: Is there a significant difference in general knowledge scores between fall and spring evaluations, when controlling for household income?

To explore this questions, we conducted another **one-way ANCOVA** model to understand how **wealth** (`TOTAL_HOUSEHOLD_INCOME` {covariate}) and **assessment timing** (`EVALUATION_PERIOD` {categorical variable}) – i.e., Fall and Spring– impact general knowledge score (`GENERAL_KNOWLEDGE_SCORE` {dependent variable).

We use a similar approach to evaluate our assumptions and conditions as detailed in Section 4:

1. ✔ **Homogeneity of Variance:** we conducted Levene's test which yielded a $p - value < 0.001$ only due to its sensitivity to large sample sizes. Our residuals confirm that our **model meets this assumption**.
2. ✔ **Normal Distribution of Residual:** we plotted the residuals' distribution (*Figure 11*), which points to normality. Again, Shapiro-Wilk test fails only due the very large sample sizes ($n > 5000$).
3. ✔ **Homogeneity of the Regression Slopes:** our analysis indicated no significant interaction effect between total household income and evaluation period on the general knowledge score ($p - value = 0.4$). Additionally, the **parallel regression** slopes (*Figure 10)* suggests homogeneity.



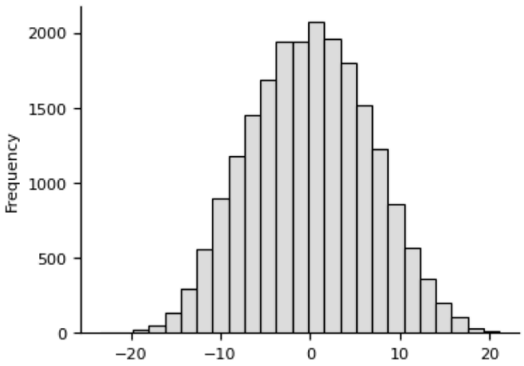
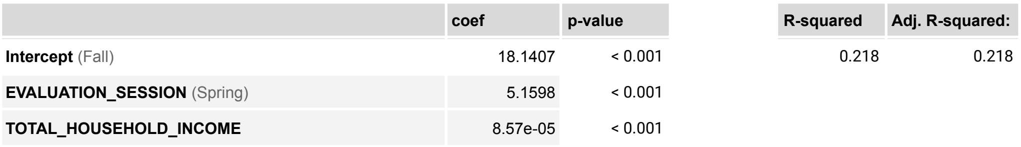


Figure 10: Household Income vs. General Knowledge Score per Season (Regression Slopes)

Figure 11: Residuals Histogram Plot



	coef	p-value
Intercept (Fall)	18.1407	< 0.001
EVALUATION_SESSION (Spring)	5.1598	< 0.001
TOTAL_HOUSEHOLD_INCOME	8.57e-05	< 0.001

R-squared	Adj. R-squared:
0.218	0.218

Figure 12: One-Way ANCOVA Results

Results:

Although our model returns an $R^2 = 0.218$, it holds strong in identifying the **'EVALUATION_SESSION[T.SPRING]'** and **'TOTAL_HOUSEHOLD_INCOME'** coefficients ($p - value < 0.001$) as having a statistically significant effect on general knowledge scores. Firstly, our analysis reveals a significant difference in general knowledge scores between fall and spring evaluations, with students achieving approximately 5.16 points higher on average during the spring session, after adjusting for household income. This seasonal effect suggests potential influences from curriculum and/or instructional strategies. Furthermore, household income also serves as a significant predictor of general knowledge scores, with higher income households showing slightly higher scores independent of evaluation sessions.

6. Conclusion

Our study revealed the significant influence of household income and prior academic knowledge on Kindergarten students' performance. We found strong correlations between income levels and seasonal variations in math, reading, and general knowledge scores, emphasizing the role of socioeconomic status in shaping student outcomes. Despite minor challenges, such as assumptions violations, our analysis remained robust, highlighting the persistent impact of socioeconomic factors on education. These findings could inform interventions for promoting educational equity, though further analysis is needed to account for external socioeconomic influences on early childhood education like parental involvement or teacher qualifications.