1. Introduction

This research aims to explore the relationship between income levels and educational outcomes among kindergarten students. Specifically, it investigates how students' performances in reading, math, and general knowledge evolve over the course of an academic year, from fall to spring.

This report using the data from “INF2178_A3_data.csv” to analyze the relationship between income level and student’s academic performance. The dataset is prepared, cleaned, and organized for further using.

This research focus on two issues and apply them as directions to organize and analyze the relationship:

Research Question 1: Does income level have significant impact on the math score improvement from fall to spring? Use fall math score as baseline.

Research Question 2: Does income level have significant impact on the reading score improvement from fall to spring? Use fall reading score as baseline.

2. Data preparation

Dataset contains 9 columns and 11933 rows. By analyzing the data structure, the first eight variables are continuing variable. “incomegroup” is only one categorical which is derived from continuous variable income. Below is the describe of each variable and their structure:

1. **fallreadingscore:** Continuous variable. Student’s reading score in Fall.
2. **fallmathscore:** Continuous variable. Student’s math score in Fall.
3. **fallgeneralknowledgescore:** Continuous variable. Student’s general knowledge score in Fall.
4. **springreadingscore:** Continuous variable. Student’s reading score in Spring.
5. **springmathscore:** Continuous variable. Student’s math score in Spring.
6. **springgeneralkowledgescore:** Continuous variable. Student’s general knowledge score in Spring.
7. **totalhouseholdincome:** Continuous variable. Student’s family’s total household income.
8. **incomeinthousands:** Continuous variable. Student’s family’s total household income in thousands.
9. **incomegroup:** Categorical variable. Categorizing the income into groups

Missing value examine:

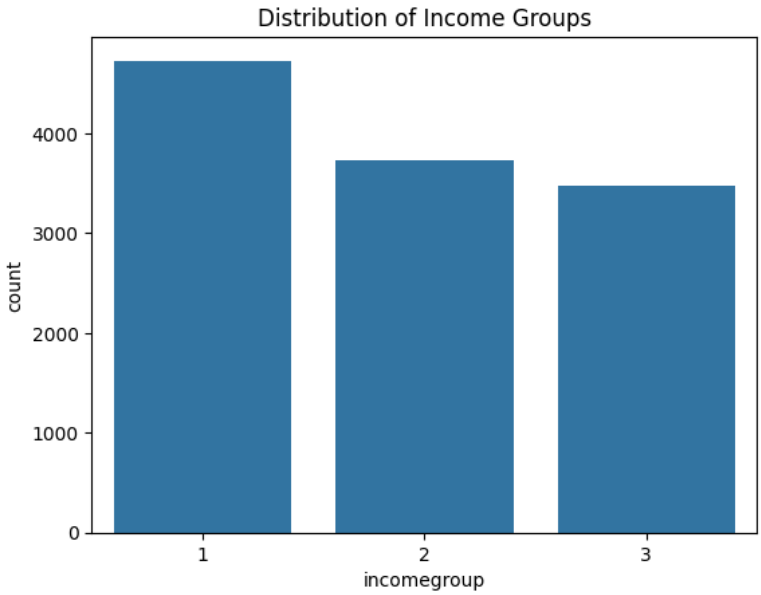
After examining the dataset, there are no missing value in each variable.

3. Exploratory Data Analysis (EDA)

	fallreadingscore	fallmathscore	fallgeneralknowledgescore	springreadingscore	springmathscore	springgeneralknowledgescore
count	11933.0	11933.0	11933.0	11933.0	11933.0	11933.0
mean	36.0	27.1	23.1	47.5	37.8	28.2
std	10.5	9.1	7.4	14.3	12.0	7.6
min	21.0	10.5	7.0	22.4	11.9	7.9
25%	29.3	20.7	17.4	39.0	29.3	22.8
50%	34.1	25.7	23.0	45.3	36.4	28.6
75%	39.9	31.6	28.3	51.8	44.2	33.8
max	138.5	115.7	47.7	156.9	113.8	48.3

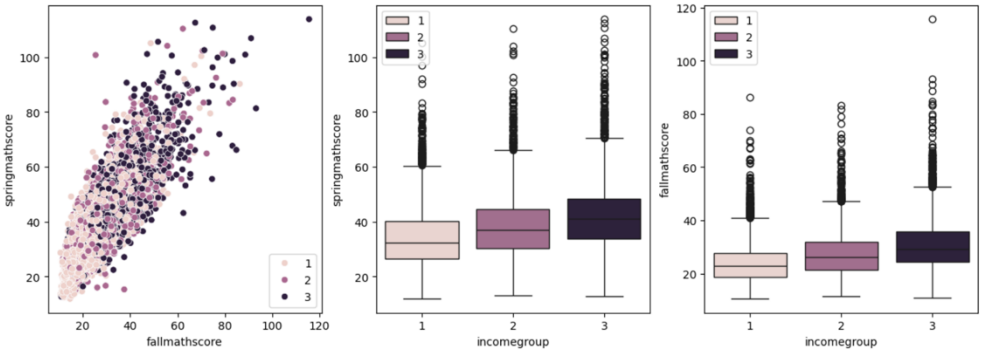
Table 1: Quantitative Data Summary

This table summarizes the kindergarten students' scores across fall and spring. The **mean scores** for reading, math, and general knowledge increase from fall to spring, indicating overall improvement. **Reading scores** show the most significant variability and range. **Math scores** have less variability but still exhibit a significant range. **General knowledge scores** are the most consistent, with the most minor variability and a narrower range in both terms.



This graph illustrates the distribution of each **income group**. Group 1 which is the lowest amount group of income, has the most counts. Group 3, which is the highest income group has the least count.

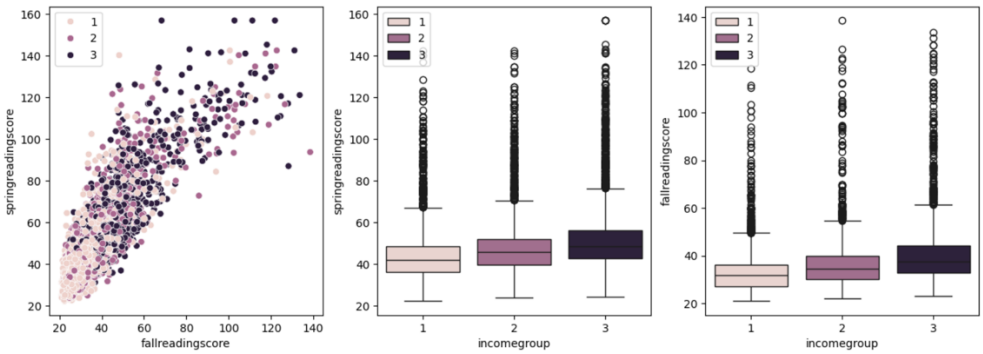
Graph 1: Distribution of Income Groups



Graph 2:

Scatter plot and Boxplot of math score across spring and fall

The scatter plot on the left illustrates a positive connection between fall math results and spring math scores, suggesting that there is a tendency for students with higher scores in the fall to also exhibit higher scores in the spring. The boxplots on the right compare spring math results and fall math scores across three income categories. Group 3, representing the highest income group, consistently exhibits higher median scores in both the spring and autumn math tests when compared to Groups 1 and 2.



Graph 3: Scatter plot and Boxplot of math score across spring and fall

The scatter plot indicates a positive correlation between fall and spring reading scores, meaning higher scores in the fall lead to higher scores in the spring. The boxplots compare spring and fall reading scores across three income groups. In both the spring and fall reading scores, Group 3 tends to have higher median scores than Groups 1 and 2.

4. One-Way ANCOVA for Math Score

In this part, One-Way ANCOVA is conducted by setting fall math score as baseline, spring math score as dependent variable, across different income groups. Below is the ANCOVA table:

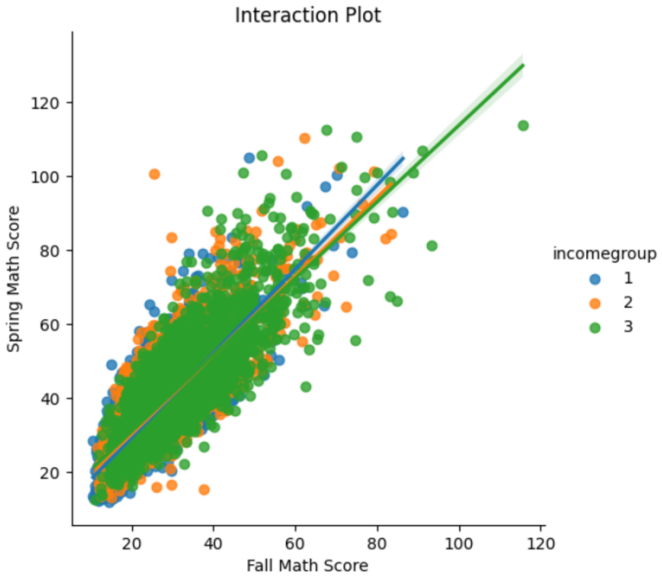
	Source	SS	DF	F	p-unc	np2
0	incomegroup	1.71E+03	2	18.523585	9.28E-09	0.003096
1	fallmathscore	1.03E+06	1	22203.081	0.00E+00	0.650505
2	Residual	5.51E+05	11929	NaN	NaN	NaN

Table 2: ANCOVA table for Math score across income group

The ANCOVA table shows that both income group and fall math scores significantly affect the spring math score.

R-squared	0.680	
F-statistic	1.27E+04	
Prob (F-statistic)	<0.01	
	coef	**P>|t|**
Intercept	7.7862	<0.01
fallmathscore	1.0735	<0.01
incomegroup	0.4699	<0.01

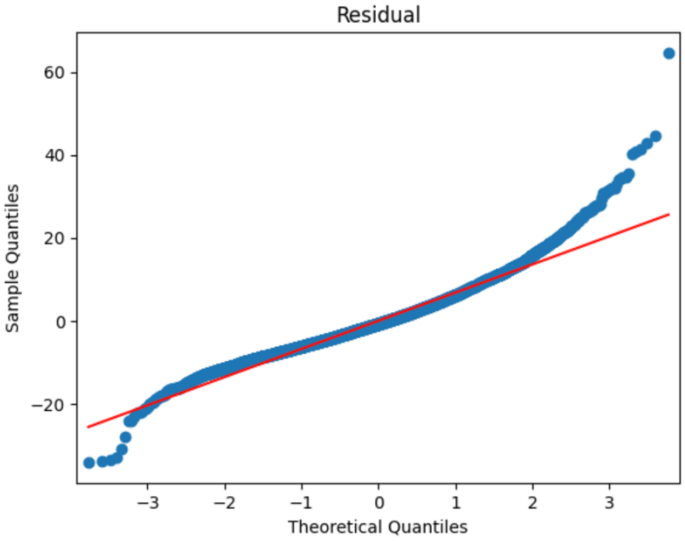
Table 3: ANCOVA table of Math Score



Graph 4: Interaction Plot of Math score in Spring and Fall across income groups

The ANCOVA table reveals that the model explains 68% of the dependent variable's variance with an R-squared value of 0.680. The overall model p-value is below 0.01. The dependent variable is positively influenced by fall math results (coefficient = 1.0735, $p < 0.01$) and income group (coefficient = 0.4699, $p < 0.01$). The intercept value of 7.7862 is statistically significant ($p < 0.01$), indicating a significant impact on the outcome variable. This table shows a strong income group impact on spring math scores.

In the Interaction plot, lines of best fit for each income group show a positive correlation between fall and spring scores, with income group 3 (green) having a steeper slope. This suggests a stronger positive relationship for higher-income families.



Graph 5: Q-Q Plot of Math ANCOVA

	Statistic	P-value
Shpiro Wilk Test	0.96	<0.01
Levene's Test	18.90	6.38e-09

Table 4: Two Tests Result

Assumption 1: Normality of Residuals

This suggests that the residuals have heavier tails than a normal distribution, indicating the presence of outliers or a non-normal distribution of the residuals. Since the p-value of Shipro-Wilk Test is below the common alpha level of 0.05, we reject the null hypothesis, indicating that the data does not follow a normal distribution. **Assumption not met.**

Assumption 2: Homogeneity of Variances

P-value less than 0.01 suggests a rejection of the null hypothesis that variances are equal across the groups, indicating that there is significant heterogeneity of variances. **Assumption not met.**

5. One-Way ANCOVA for Reading score

	Source	SS	DF	F	p-unc	np2
0	incomegroup	5.13E+02	2	4.05566	0.017348	0.00068
1	fallreadingscore	1.55E+06	1	24455.3976	0	0.67214
2	Residual	7.55E+05	11929	NaN	NaN	NaN

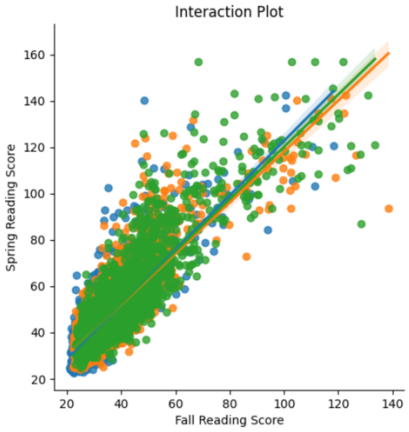
Table 5: ANCOVA table for Reading Score

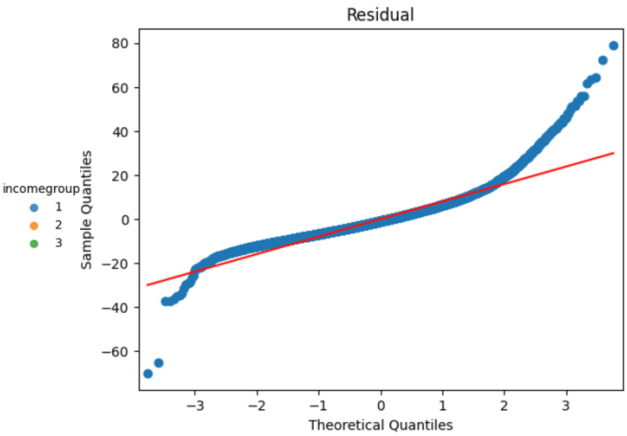
The ANCOVA table shows that both income group and fall reading scores significantly affect the spring reading score.

Table 6: ANCOVA table for Reading score factors

			coef	P > |t|
R-squared	0.692	**Intercept**	6.3272	<0.01
F-statistic	1.34E+04	**fallreadingscore**	1.1322	<0.01
Prob(F-statistic)	<0.01	**incomegroup**	0.2512	0.006

With an R-squared of 0.692, the ANCOVA table shows that the model accounts for 69.2% of the variance in the dependent variable. With a p-value of less than 0.01 and an F-statistic of 1.34E+04, the model is highly significant. Strong predictors in the model are indicated by the significant coefficients for the intercept and fall reading scores, which are 6.3272 and 1.1322, respectively (both $p < 0.01$), and the income group coefficient, which is 0.2512 ($p = 0.006$).





Graph 6: Interaction Plot of Reading score across fall and spring(Top Left)

Graph 7: Q-Q plot of Reading score ANCOVA(Top Right)

	Statistic	P-value
Shpiro Wilk Test	0.91	<0.01
Levene's Test	39.56	7.57E-18

Table 7: Two Tests Results of reading score

In the **Interaction plot**, the lines of best fit for each group show a positive correlation between scores from fall to spring, and the slopes of these lines suggest that this relationship may differ slightly by income group, with group 3 possibly having a steeper slope.

In the **Q-Q Plot**, the lower tail drops below, and the upper tail rises above the expected line, which indicates the residuals have heavier tails than a normal distribution. This indicates **Assumption 1 of Normality of Residuals is not hold.** In the Two tests table, P-value of Levene's Test less than 0.01 indicates that variances are not equal across the groups, indicating that there is significant heterogeneity of variances. **Assumption 2 of Homogeneity of Variances not met.**

6. Conclusion

Based on the findings and results from One-Way ANCOVA, it si clear that income group has impact on both math and reading score change from fall to spring. Both Assumptions are not hold suggests we may need to choose another appropriate to conduct the research. In conclusion, this research provide valuable analysis on how income group will influence the student's academic performance.