Examination Of Early Child Longitudinal Study

1. Introduction

This report explores child longitudinal study data through the dataset INF2178_A3_data.csv that outlines data points on fall/spring reading, math, and general knowledge scores of individual students by income. The data shows the scores for fall 1998 and spring 1999 measurements, evaluating kindergarten students over the span of several months. The analysis is guided by the following questions:

Research question 1: Does income group significantly predict differences in spring general knowledge scores among kindergarten students after controlling baseline reading and math abilities?

- Null Hypothesis (H0): There is no significant difference in spring general knowledge scores among kindergarten students across different income groups after controlling baseline reading and math abilities.
- Alternative Hypothesis (H1): There is a significant difference in spring general knowledge scores among kindergarten students across different income groups after controlling baseline reading and math abilities.

Research question 2: Does income group predict spring math scores while controlling for baseline fall math scores?

- Null Hypothesis (H0): There is no significant difference in spring math scores among income groups after controlling for baseline fall math scores.
- Alternative Hypothesis (H1): There is a significant difference in spring math scores among income groups after controlling for baseline fall math scores.

2. Data Cleaning and Wrangling

The dataset INF2178_A2_data.xlsx consists of 11,933 rows and 9 columns. Among these columns, 8 are of the float64 data type, while one column, 'incomegroup', is represented as an integer. The column 'incomeinthousands' is derived from the 'totalhouseholdincome' column, providing the same data in a condensed format. Similarly, 'incomegroup' is also a derivative of 'totalhouseholdincome'. To facilitate analysis, the 'incomegroup' variable has been converted into a categorical variable. Furthermore, for visualization purposes, the values in the 'incomegroup' variable have been transformed as follows: 1 represents low income, 2 represents medium income, and 3 represents high income. It is worth noting that no missing values have been identified in the dataset

3. Exploratory Data Analysis

The examination of the dataset was thorough, as evidenced by the content of the Python notebook. Only essential findings from this Explorative Data Analysis are presented in this paper.

In Fall, average scores of 35.95 in Reading, 27.13 in Math, and 23.07 in General Knowledge underscore proficiency levels close to the middle of the range. With standard deviation of 10.47 in Reading, 9.12 in Math, and 7.40 in General Knowledge in Fall, it became apparent that while some students clustered closely around the mean, others exhibited greater divergence in performance. In Spring, with mean scores of 47.51 (STD=14.33) in Reading, 37.80(STD=12.03) in Math, and 28.24(STD=7.58) in General Knowledge, the data suggested that a significant proportion of students achieved proficiency levels close to these central

values, indicating overall progress within the cohort. The variability in student performance persisted, suggesting continued diversity in achievement levels. In all three subjects—Reading, Math, and General Knowledge—the mean scores increased from fall to spring(Table 1).

	Fall Reading	Fall Math	Fall General Knowledge	Spring Reading	Spring Math	Spring General Knowledge	Total Household Income	Income (in thsd.)
Count	11933	11933	11933	11933	11933	11933	11933	11933
Mean	35.95	27.13	23.07	47.51	37.80	28.24	54317.20	54.32
Std	10.47	9.12	7.40	14.33	12.03	7.58	36639.06	36.64
Max	138.51	115.65	47.691	156.85	113.8	48.345	150000	150
Min	21.01	10.51	6.985	22.35	11.9	7.858	1	0.001
Median	34.06	25.68	22.954	45.32	36.41	28.583	47000	47
IQR	10.55	10.91	10.92	12.82	14.95	10.98	45000	45

Table 1

The median scores in spring were consistently higher than those in fall for all three subjects (Figure 2, Figure 3). The interquartile ranges in spring generally exhibited similar trends to those observed in fall, albeit with slight variations (Table 1). The maximum scores observed in spring were generally higher than those recorded in fall, reflecting the substantial progress made by some students. Similarly, the minimum scores in spring tended to be higher than those in fall, indicating overall advancement in academic performance across the student population. Additionally, Reading and Math scores in Spring and Fall have significant outliers, while few outliers are observed for the general knowledge score (Figure 2, Figure 3).

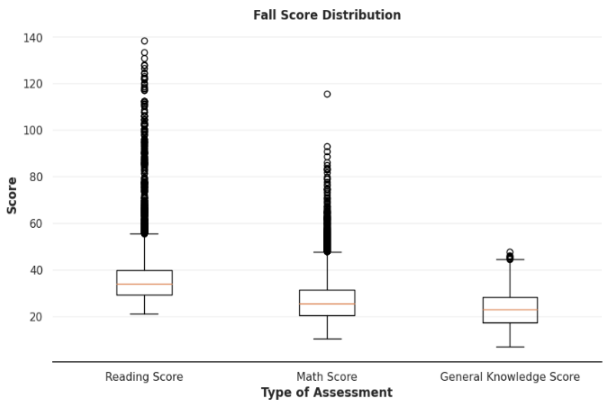


Figure 2

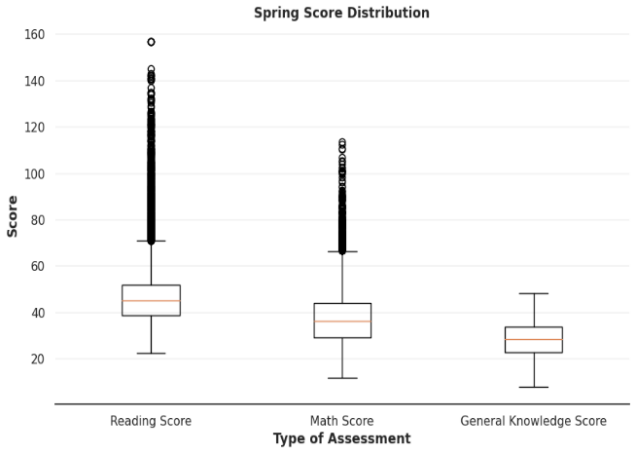


Figure 3

An analysis has been conducted to fall 1998 and spring 1999 scores in three subjects: reading, math, and general knowledge (Figure 4, Figure 5, Figure 6). Each subplot within the figure represents one of these subjects, with fall scores plotted on the x-axis and spring scores plotted on the y-axis. For all three subjects fall and spring scores mirror each other with linearity in the relations. The higher scores in reading are significantly scattered, and the same can be observed for the highest score in math (Figure 4, Figure 5, Figure 6).

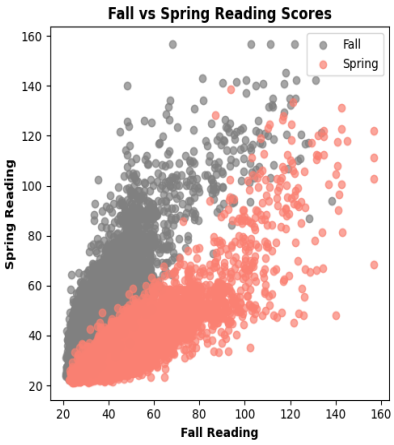


Figure 4

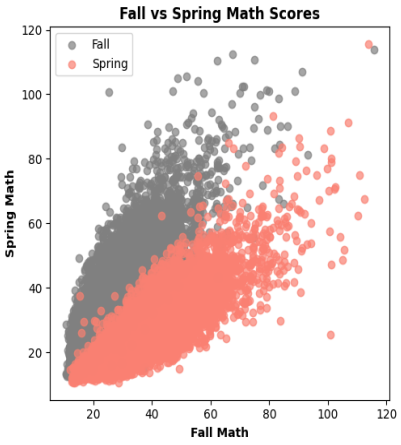


Figure 5

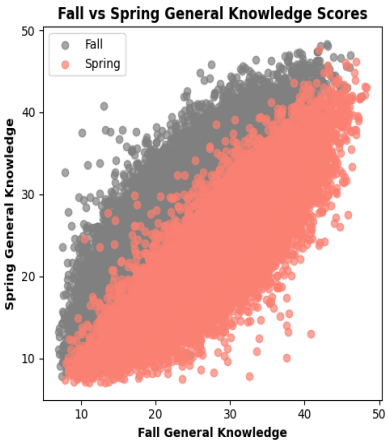


Figure 6

Figure 7 and Figure 8 compare the relationship between reading and math scores for fall and spring, while considering the income groups. The examination of Figure 7 shows that children belonging to income group 1 are predominantly clustered towards the lower end of the plot, indicating lower proficiency in both math and reading during fall. However, during spring, children from income group 1 are also observed in the middle of the scatter plot (see Figure 8), suggesting an improvement in their performance. Additionally, it's notable that the data points in spring display a more linear relationship compared to those in fall. Finally, the datapoint on y-axis in Figure 8 has higher values across all income groups suggesting that children across the groups have improved their results.

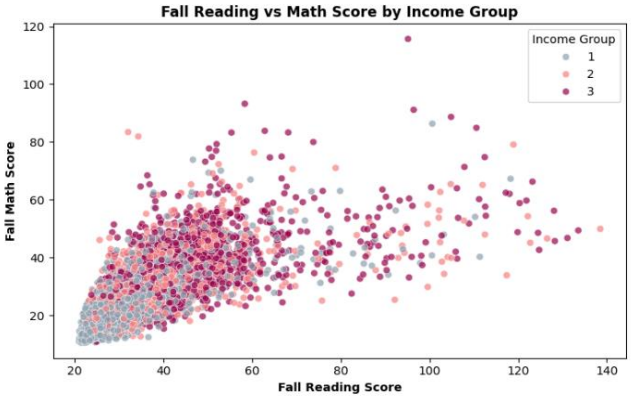


Figure 7

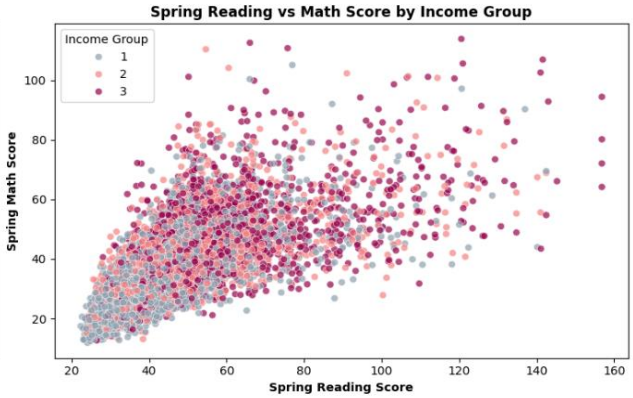
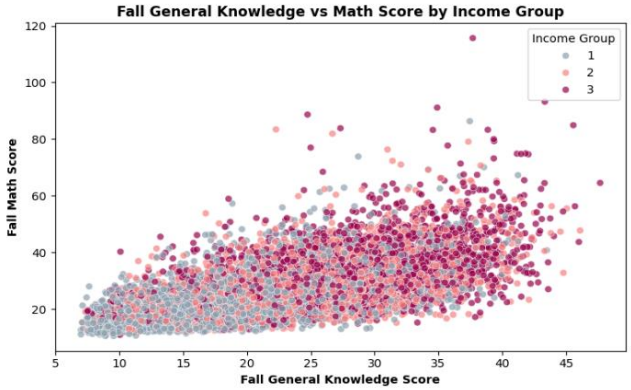


Figure 8

The scatter plots (Figure 9 and Figure 10) compare the relation between general knowledge on x-axis and math score on y-axis for fall and spring respectively considering the income group of the children. Based on the figures, the relation between math and general knowledge is mostly linear, with the children from the income group 3 primarily concentrating at the end of the pot. Based on Figure 10, the children from income group 1 improved their performance both in math and general knowledge in spring. Additionally, the datapoints are more scattered at the end showing the higher performance of children in math particularly for the children of income group 3. The figures above drive formulation of the research questions.



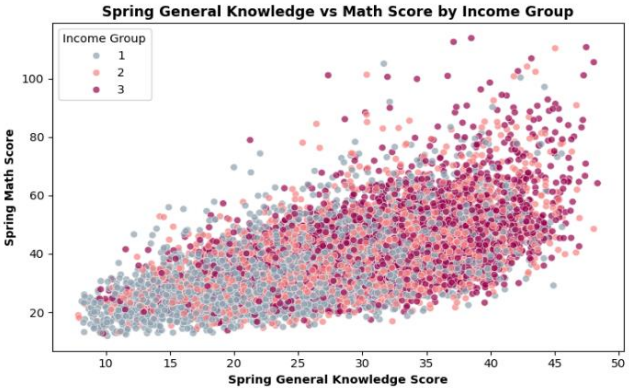
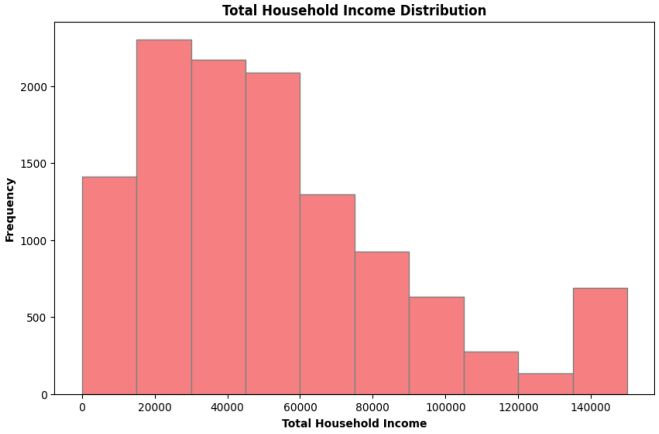


Figure 9 **Figure 10**

Based on the histogram (Figure 9), the total household income has been rightly skewed suggesting that most children in the dataset are in the lower income group. The bar chart in Figure 10 compares the average general knowledge scores between income groups both for fall and spring. Based on Figure 10, the children from high income group have overall higher average performance than other children. Additionally, Figure10 shows that the students have improved their scores in general knowledge in spring across all three income groups.



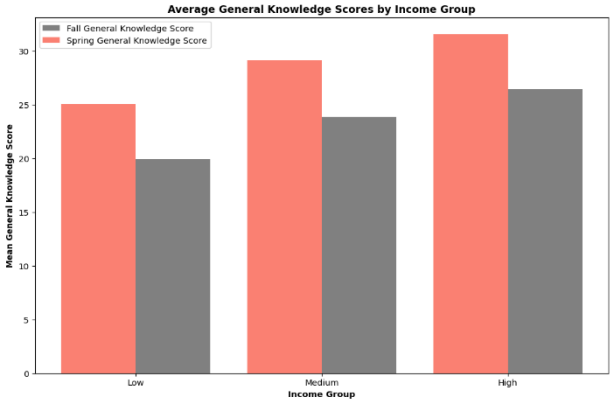


Figure 9 **Figure 10**

4. Discussion

4.1. Effect of Income Group on Kindergarten Students' General Knowledge Scores in Spring 1999

To answer the research question on whether the income group significantly predicts differences in spring general knowledge scores among kindergarten students after controlling baseline reading and math abilities the one-way ANCOVA has been employed.

	Source	SS	DF	F	p-unc	np2
0	incomegroup	24052.3	2	356.128	5.99e-151	0.05635
1	springreadingscore	2726.94	1	80.7523	2.94e-19	0.00672
2	springmathscore	89622.1	1	2653.96	0.00e+00	0.182
3	Residual	402799	11928	NaN	NaN	NaN

Table 2

Based on the Table 2, the p-value is smaller than significance level (5.99e-151<0.05), so the null hypothesis has been rejected. This suggests that income group does significantly predict differences in spring general knowledge scores among kindergarten students after controlling baseline reading and math abilities. The p-values for the baseline reading and math scores are also extremely small.

There is a positive interaction between spring reading score, spring math score, and income group, so there is combined effect of increasing spring reading score, spring math score, and income group on spring general knowledge score is augmented (Figure 11).

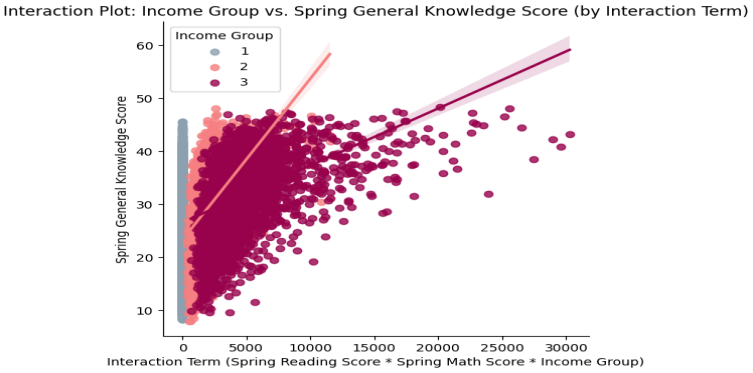


Figure 11

The assumption tests have been utilized to evaluate two critical assumptions: normality and homogeneity of variances. The assumption check suggests the following:

- **Assumption on normality**. Koglomorov Smirnov test has been conducted instead of Shapiro-Wilk test considering that the number of data points(n=11,933), which exceeds 5000. With the small p-value equals to 0.0 < 0.05), the null hypothesis has been rejected. Based on the results of Kolmogorov-Smirnov test, the residuals do not appear to be normally distributed. This suggests that the assumption of normality for the residuals in the ANCOVA model is violated.
- **Assumption on homogeneity of variances.** Based on the results of the Barlett test, p-value equals 0.3498, which is greater than significance level of 0.05. Thus, the null hypothesis has failed to be rejected. The variances of the residuals appear to be homogeneous, so the assumption of homogeneity of variances for the residuals in the ANCOVA model is met.

4.2. Effect of Income Group on Spring Math Scores Adjusted for Baseline Fall Math Performance

The one-way ANCOVA has been conducted to respond to the research question whether the income group predicts spring math scores while controlling for baseline fall math scores. Based on the results in Table 3, the p-values associated with each coefficient are all less than 0.05, indicating that all predictors, including intercept and fall math score, are statistically significant in predicting spring math scores. There is a significant difference in spring math scores among income groups after controlling for baseline fall math scores. Specifically, students from higher income groups such as income group 2 and income group 3 tend to have higher spring math scores compared to students from income group 1, even when controlling for their baseline fall math scores.

OLS Regression Results

Dep. Variable:	springmathscore	R-squared:	0.681
Model:	OLS	Adj. R-squared:	0.680
Method:	Least Squares	F-statistic:	8469.
Date:	Sat, 23 Mar 2024	Prob (F-statistic):	0.00
Time:	19:41:19	Log-Likelihood:	-39804.
No. Observations:	11933	AIC:	7.962e+04
Df Residuals:	11929	BIC:	7.965e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	8.2011	0.199	41.273	0.000	7.812	8.591
incomegroup[T.2]	0.6700	0.151	4.430	0.000	0.374	0.966
incomegroup[T.3]	0.9199	0.160	5.741	0.000	0.606	1.234
fallmathscore	1.0735	0.007	149.007	0.000	1.059	1.088

Omnibus:	1782.310	Durbin-Watson:	1.808
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5069.769
Skew:	0.802	Prob(JB):	0.00
Kurtosis:	5.761	Cond. No.	94.8

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified. **Table 3**

The absence of crossing lines indicates that the effect of fall math scores on spring math scores does not vary depending on income group. This suggests that the relationship between fall math scores and spring math scores remains consistent regardless of the income group of the students. Since there is no interaction effect observed, it implies that income group does not change the relationship between fall math scores and spring math scores. Thus, the influence of fall math scores on spring math scores is uniform across all income groups.

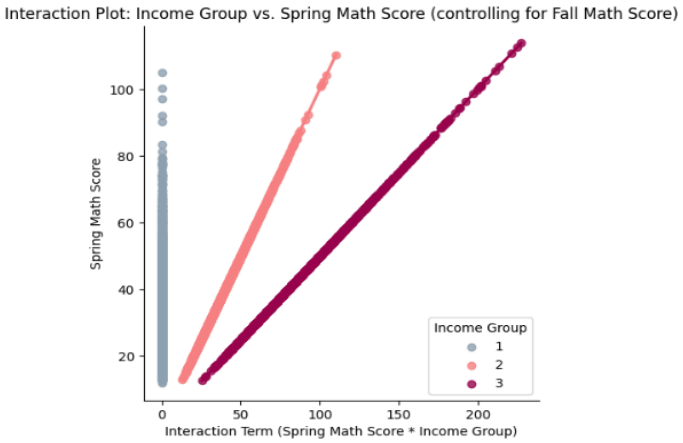


Figure 12

The assumption tests have been utilized to evaluate two critical assumptions: normality and homogeneity of variances. The assumption check suggests the following:

- **Assumption on normality**. The findings of the Kolmogorov-Smirnov test (Kolmogorov-Smirnov test statistic is 0.3892435359616416 with a p-value of 0.0.) suggest that the assumption of normality for the residuals of the model is violated.
- **Assumption on homogeneity of variances.** Based on the results of the Barlett test, the p-value equals to 9.226636634960654e-20 is less than the significance level 0.05, the null hypothesis is rejected. This indicates that there is sufficient evidence to suggest that the variances of the residuals are not equal across groups or levels of the independent variables, which the assumption of homogeneity of variances is violated.

For the one-way ANCOVA test on the first research question, the violation of normality is concerning, since it affects the validity of the test. The one-way ANCOVA on the second research question violates both normality and homogeneity of variance assumptions. This affects the accuracy of the parameter estimates and the validity of the statistics as well as leading to biases in estimates and an increased Type I error rate. This means that both tests are not reliable to draw conclusions, and the violations should be addressed before processing further.