1. Introduction

In recent years, the early childhood education journey has received much attention for its critical role in shaping foundational skills and knowledge. Within this context, a unique subset of data from the 1998-99 Early Childhood Longitudinal Study became the focus of our investigation. This dataset covers reading, math, and general knowledge scores from fall 1998 through spring 1999, providing a perspective for assessing the academic growth of kindergarten students over several months. The inclusion of income category, the only categorical variable, adds a dimension of socioeconomic analysis to our study.

A comprehensive exploratory data analysis was conducted for this report using a dataset called "INF2178 A3 data.csv". It scrutinizes the first six columns of continuous variables that represent individual students' fall and spring scores in reading, mathematics, and general knowledge. It closely examined the first six columns of continuous variables that represent individual student fall and spring scores in reading, mathematics, and general knowledge. It utilizes general knowledge scores as benchmarks to facilitate comparative studies of how reading and math scores change over time for students in different income groups.

Our exploration will address three fundamental research questions:

1. Research Question 1: Does the family income group of students influence the change in their math scores over time, and is this influence still significant after controlling for fall general knowledge scores?
2. Research Question 2: Does the family income group of students influence the change in their reading scores over time, and is this influence still significant after controlling for fall general knowledge scores?

Through this analysis, we aim to contribute a deeper understanding of educational disparities, thereby informing more targeted and effective educational interventions.

2. Data Cleaning and Data Wrangling

This original dataset has 9 columns and 11,933 entities (rows). To better analyze the research, we have adjusted the original dataset:

2.1 Since our analysis is quantitative in nature, I have considered deleting unimportant columns and keeping the following to make the data clearer and easier to understand:

‘fallreadingscore’, ‘fallmathscore’, ‘fallgeneralknowledgescore’, ‘springreadingscore’, ‘springmathscore’, ‘springgeneralknowledgescore’, ‘incomegroup’

2.2 I have conducted a preliminary exploration of the reduced data. based statistical summary:

The dataset has a total of 11933 entries with 7 features.

One feature is recognized as integer type (int64) and 8 features are recognized as float types

There are no significant missing values as each feature has 11933 non-null values.

3. Data Engineer

To explore changes in students' reading and math scores over time, two continuous variables ‘math_over_time’ and ‘reading_over_time’ were created using:

math_over_time = Math scores in the spring - Math scores in the fall;

reading_over_time= reading scores in the spring - reading scores in the fall.

4. ANCOVA

4.1 Math Over Time

4.1.1 EDA

The table 1 shows that the income group increases from 1 to 3, there is also an increase in the average improvement of students' math scores over time. The average increase for Group 1 is about 9.96 points,

for Group 2 it is 10.90 points, and for Group 3 it is 11.40 points. Accompanying this trend is an increase in standard deviation, from 6.41 in Group 1 to 7.34 in Group 3, which indicates a greater variability in the improvement of math scores among students in higher income groups.

incomegroup	n	mean	std
1	4729	9.958547	6.406540
2	3726	10.896224	6.853736
3	3478	11.399178	7.339003

Table 1

The figure shows that the Scatter Plot: Displays the relationship between students' general knowledge scores in the fall and the change in math scores over time, with different colors representing different income groups. The plot indicates a positive correlation between students' fall general knowledge scores and their subsequent improvement in math scores over time. Box Plot (Middle): Compares the income groups with the changes in math scores over time, showing that students from higher-income groups tend to have a larger median change in math scores. As income groups increase, variability also appears to increase. Box Plot (Right): Shows the distribution of fall general knowledge scores across different income groups. Students from higher-income groups have higher median general knowledge scores. This may suggest that students from higher-income families have access to better educational resources. Additionally, the charts contain a significant number of outliers, which I have decided to keep. Removing them could lead to a misunderstanding of the overall data structure and distribution, whereas retaining them ensures the integrity of the data.

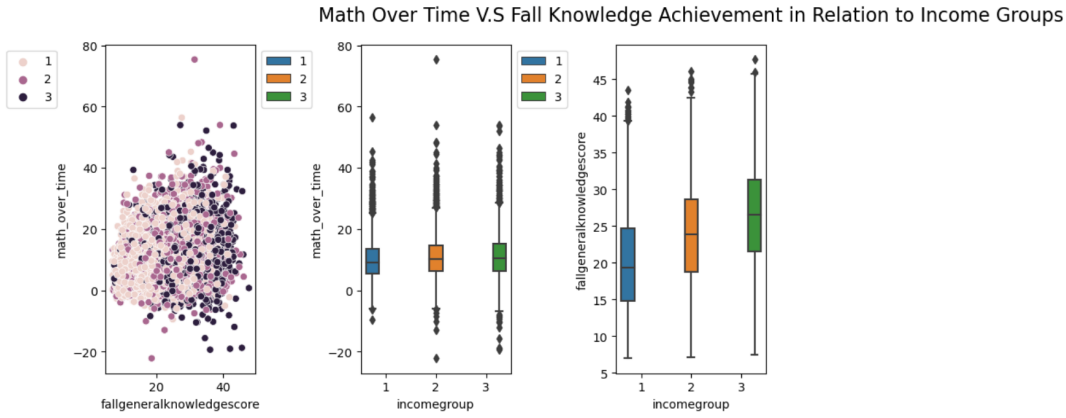

Figure1

4.1.2 ANCOVA

- The effect of income group on the change in math scores:H0: The average change in math scores is equal across all income groups, which means the income group does not have an effect on the change in students' math scores.H1: There is at least one income group whose average change in math scores is different from the others, suggesting that the income group does influence the change in students' math scores.
- The effect of fall general knowledge scores as a covariate:H0: There is no relationship between fall general knowledge scores and the change in math scores.H1: There is a relationship between fall general knowledge scores and the change in math scores.

From the Table 2, for the independent variable, income group, the impact on the change in math scores over time was found to be F = 0.62 with a p-value = 0.540 > 0.05, indicating that we cannot reject the null hypothesis, i.e.the average change in math scores is equal across all income groups, which means the income group does not have an effect on the change in students' math scores.For the covariate, fall general knowledge score , the impact on the change in math scores over time was significant with F = 501.08 and a p-value < 0.001, suggesting that there is sufficient evidence to reject the null hypothesis,i.e.there is no relationship between fall general knowledge scores and the change in math scores.

Source	SS	DF	F	p-unc	np2
incomegroup	55.88	2	0.62	0.540	0.00
fallgeneralknowledgescore	22425.93	1	501.08	<0.001	0.04
Residual	533880.50	11929	NaN	nan	NaN

Table 2

From the Table 3, the results indicate that the coefficient for the income group is 0.0752, with a p-value of 0.347>0.05, suggesting that we cannot reject the null hypothesis, i.e. the average change in math scores is equal across all income groups, which means the income group does not influence the change in students' math scores. The coefficient for fall general knowledge scores is 0.1996, with a p-value<0.001, indicating that there is sufficient evidence to reject the null hypothesis i.e. there is no relationship between fall general knowledge scores and the change in math scores, and there is a significant positive correlation between fall general knowledge scores and the change in math scores. The R-squared value is 0.048, suggesting that the variables in the model explain only about 4.8% of the variability in the change in math scores. This implies that there are other factors not included in the model that are affecting the change in math scores. The Omnibus test's p-value is close to 0, indicating that the residuals may not be normally distributed.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         math_over_time   R-squared:                       0.048
Model:                            OLS   Adj. R-squared:                  0.048
Method:                 Least Squares   F-statistic:                     299.9
Date:                Sat, 23 Mar 2024   Prob (F-statistic):          8.33e-128
Time:                        13:35:09   Log-Likelihood:                -39610.
No. Observations:               11933   AIC:                         7.923e+04
Df Residuals:                   11930   BIC:                         7.925e+04
Df Model:                           2
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                        5.9240      0.215     27.589      0.000       5.503       6.345
incomegroup                      0.0752      0.080      0.940      0.347      -0.082       0.232
fallgeneralknowledgescore        0.1996      0.009     22.437      0.000       0.182       0.217
==============================================================================
Omnibus:                     1731.033   Durbin-Watson:                   1.804
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4987.511
Skew:                           0.777   Prob(JB):                         0.00
Kurtosis:                       5.760   Cond. No.                         86.2
==============================================================================

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

Table 3

- H0: There is no interaction between income group and fall general knowledge scores, meaning their impacts on the change in math scores are independent of each other.
- H1: There is an interaction between income group and fall general knowledge scores, meaning they jointly influence the change in math scores.

The Table 4 shows that the coefficient for the income group is 1.1572, with a p-value < 0.001, indicating that we have sufficient evidence to reject the null hypothesis that the average change in math scores is the same across all income groups. The coefficient for the fall general knowledge score is 0.2889, with a p-value < 0.001, indicating that we have sufficient evidence to reject the null hypothesis that there is no relationship between fall general knowledge scores and the change in math scores. The coefficient for the interaction term incomegroup: fallgeneralknowledge score is -0.0464, with a p-value < 0.001, indicating that we have sufficient evidence to reject the null hypothesis that there is no interaction between income group and fall general knowledge scores.

OLS Regression Results

Dep. Variable:	math_over_time	R-squared:	0.049
Model:	OLS	Adj. R-squared:	0.049
Method:	Least Squares	F-statistic:	206.9
Date:	Sat, 23 Mar 2024	Prob (F-statistic):	7.97e-131
Time:	13:35:09	Log-Likelihood:	-39600.
No. Observations:	11933	AIC:	7.921e+04
Df Residuals:	11929	BIC:	7.924e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.9471	0.492	8.017	0.000	2.982	4.912
incomegroup	1.1572	0.255	4.531	0.000	0.657	1.658
fallgeneralknowledgescore	0.2889	0.022	13.189	0.000	0.246	0.332
incomegroup:fallgeneralknowledgescore	-0.0464	0.010	-4.461	0.000	-0.067	-0.026

Omnibus:	1747.450	Durbin-Watson:	1.804
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5012.870
Skew:	0.785	Prob(JB):	0.00
Kurtosis:	5.760	Cond. No.	524.

Table 4

4.2 Reading Over Time

4.1.1 EDA

The table 5 shows that as the income group increases from 1 to 3, there is also an increase in the average improvement of students' reading scores over time. The average increase for Group 1 is about 10.88 points, for Group 2 it is 11.72 points, and for Group 3 it is 12.30 points. Accompanying this trend is an increase in standard deviation, from 7.46 in Group 1 to 8.97 in Group 3, which indicates a greater variability in the improvement of reading scores among students in higher income groups.

incomegroup	n	mean	std
1	4729	10.878279	7.462176
2	3726	11.716932	7.906122
3	3478	12.308387	8.967606

Table 5

The Figure 2 shows that the Scatter Plot: Displays the relationship between students' general knowledge scores in the fall and the change in reading scores over time, with different colors representing different income groups. The plot indicates a positive correlation between students' fall general knowledge scores and their subsequent improvement in reading scores over time. Box Plot (Middle): Compares the income groups with the changes in reading scores over time, showing that students from higher-income groups tend to have a larger median change in reading scores. As income groups increase, variability also appears to increase. Box Plot (Right): Shows the distribution of fall general knowledge scores across different income groups. Students from higher-income groups have higher median general knowledge scores. This may suggest that students from higher-income families have access to better educational resources. Additionally, the charts contain a significant number of

outliers, which I have decided to keep. Removing them could lead to a misunderstanding of the overall data structure and distribution, whereas retaining them ensures the integrity of the data.

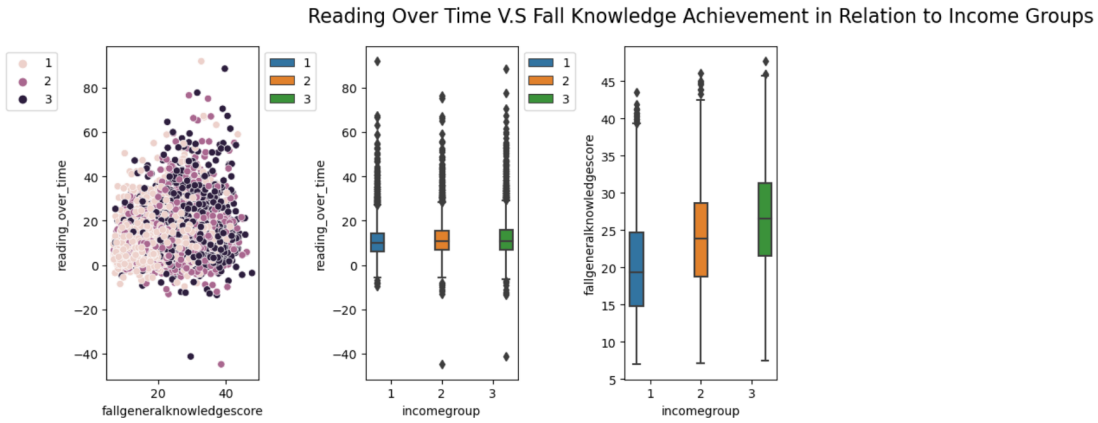

Figure 2

4.2.2 ANCOVA

- The effect of income group on the change in reading scores:H0: The average change in reading scores is equal across all income groups, which means the income group does not have an effect on the change in students' reading scores.H1: There is at least one income group whose average change in reading scores is different from the others, suggesting that the income group does influence the change in students' reading scores.
- The effect of fall general knowledge scores as a covariate:H0: There is no relationship between fall general knowledge scores and the change in reading scores.H1: There is a relationship between fall general knowledge scores and the change in reading scores.

From the Table 6, the independent variable, income group, the impact on the change in reading scores over time was found to be F = 2.25 with a p-value = 0.110 > 0.05, indicating that we cannot reject the null hypothesis, i.e.the average change in reading scores is equal across all income groups, which means the income group does not have an effect on the change in students' reading scores. For the covariate, fall general knowledge score , the impact on the change in reading scores over time was significant with F = 220.11 and a p-value < 0.001, suggesting that there is sufficient evidence to reject the null hypothesis,i.e.there is no relationship between fall general knowledge scores and the change in reading scores.

Source	SS	DF	F	p-unc	np2
incomegroup	287.49	2	2.25	0.110	0.00
fallgeneralknowledgescore	14054.12	1	220.11	<0.001	0.02
Residual	761671.04	11929	NaN	nan	NaN

Table 6

From the Table 7, for the income group incomegroup, the coefficient is 0.2025 with a p-value=0.034<0.05. This means we have sufficient evidence to reject the null hypothesis, i.e. the average change in reading scores is equal across all income groups, which means the income group does not influence the change in students' reading scores. For the fall general knowledge score, the coefficient is 0.1578 with a p-value<0.001, indicating that we have sufficient evidence to reject the null hypothesis i.e. there is no relationship between fall general knowledge scores and the change in reading scores, and it demonstrates that the fall general knowledge score has a significant positive impact on

the change in reading scores. The R-squared value is 0.023, indicating that these variables only explain approximately 2.3% of the variability in the change in reading scores. The Omnibus test's p-value is close to 0, suggesting that the residuals may not follow a normal distribution.

OLS Regression Results

Dep. Variable:	reading_over_time	R-squared:	0.023
Model:	OLS	Adj. R-squared:	0.023
Method:	Least Squares	F-statistic:	143.2
Date:	Sat, 23 Mar 2024	Prob (F-statistic):	3.35e-62
Time:	13:35:12	Log-Likelihood:	-41730.
No. Observations:	11933	AIC:	8.347e+04
Df Residuals:	11930	BIC:	8.349e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	7.5314	0.256	29.365	0.000	7.029	8.034
incomegroup	0.2025	0.096	2.120	0.034	0.015	0.390
fallgeneralknowledgescore	0.1578	0.011	14.857	0.000	0.137	0.179

Omnibus:	4517.094	Durbin-Watson:	1.715
Prob(Omnibus):	0.000	Jarque-Bera (JB):	33257.080
Skew:	1.632	Prob(JB):	0.00
Kurtosis:	10.499	Cond. No.	86.2

Table 7

- H0: There is no interaction between income group and fall general knowledge scores, meaning their impacts on the change in reading scores are independent of each other.
- H1: There is an interaction between income group and fall general knowledge scores, meaning they jointly influence the change in reading score.

Based on the Table 8, the coefficient for the income group is 1.5409, with a p-value < 0.001, indicating that we have sufficient evidence to reject the null hypothesis that the average change in reading scores is the same across all income groups. The coefficient for the fall general knowledge score is 0.2683, with a p-value < 0.001, indicating that we have sufficient evidence to reject the null hypothesis that there is no relationship between fall general knowledge scores and the change in reading scores. The coefficient for the interaction term incomegroup: fallgeneralknowledge score is -0.0575, with a p-value < 0.001, indicating that we have sufficient evidence to reject the null hypothesis that there is no interaction between income group and fall general knowledge scores.

OLS Regression Results

Dep. Variable:	reading_over_time	R-squared:	0.025
Model:	OLS	Adj. R-squared:	0.025
Method:	Least Squares	F-statistic:	102.8
Date:	Sat, 23 Mar 2024	Prob (F-statistic):	1.10e-65
Time:	13:35:12	Log-Likelihood:	-41720.
No. Observations:	11933	AIC:	8.345e+04
Df Residuals:	11929	BIC:	8.348e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.0859	0.588	8.649	0.000	3.933	6.239
incomegroup	1.5409	0.305	5.052	0.000	0.943	2.139
fallgeneralknowledgescore	0.2683	0.026	10.256	0.000	0.217	0.320
incomegroup:fallgeneralknowledgescore	-0.0575	0.012	-4.620	0.000	-0.082	-0.033

Omnibus:	4552.226	Durbin-Watson:	1.716
Prob(Omnibus):	0.000	Jarque-Bera (JB):	33764.091
Skew:	1.645	Prob(JB):	0.00
Kurtosis:	10.555	Cond. No.	524.

Table 8