Exploring the Impacts on Early Children's Academic Progress

1. Introduction:

Since the early 2000s, the academic community has increasingly focused on early childhood longitudinal studies to understand the myriad factors influencing young learners' educational outcomes. These factors range from their initial educational preparedness to various socioeconomic conditions, notably family income, which collectively shape their learning trajectories and ability to assimilate new knowledge.

This report uses a dataset "INF2178_A3_data.csv", which contains the data from the 1998-99 Early Childhood Longitudinal Study. This report focuses on kindergarten students' reading, math, and general knowledge scores from fall 1998 to spring 1999. Employing Analysis of Covariance (ANCOVA) seeks to identify underlying factors affecting kindergarten scores, especially insights into the educational disparities in economic conditions. The report focuses on the following research questions:

- How does income group (**incomegroup**) affect students' reading scores in the spring(**springreadingscore**), after controlling for their reading scores in the fall (**fallreadingscore**)?
- How does income group (i**ncomegroup**) affect students' math scores in the spring (**springmathscore**), after controlling for their math scores in the fall(**fallmathscore**)?
- How does income group (**incomegroup**) affect students' general knowledge scores (**springgeneralknowledgescore**) in the spring, after controlling for their general knowledge scores in the fall(**springgeneralknowledgescore**)?

2. Data Cleaning:

This dataset contains a total of 9 columns and 11933 rows, since the income group represents the household income level, we can drop the two columns **totalhouseholdincome** and **incomeinthousands** which represent the exact amount of the household income. To appropriately represent its unique characteristic as the only non-continuous variable in the dataset, the data structure for the 'income group' , which is Initially formatted as an integer, has been changed into a categorical variable to better align with its distinct nature. Below are the descriptions for each column:

- **fallreadingcore**: Early children's reading score in fall 1998;
- **fallmathscore**: Early children's math score in fall 1998;
- **fallgeneralknowledgesoce**: Early children's general knowledge score in fall 1998;
- **springreadingscore**: Early children's reading score in spring 1999;
- **springmathscore**: Early children's math score in spring 1999;
- **springgeneralknowledgesoce**: Early children's general knowledge score in spring 1999;
- **incomegroup**: Early children's household income level.

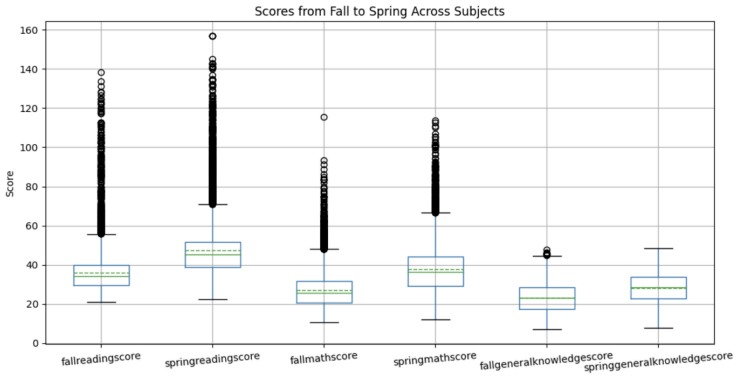
Upon checking, there are no missing values in the dataset so we can start our analysis.

3. Exploratory Data Analysis (EDA):

In our exploratory data analysis, an initial review of the summary statistics and boxplot for different subjects provides an overarching view of the kindergarten students' academic performance.

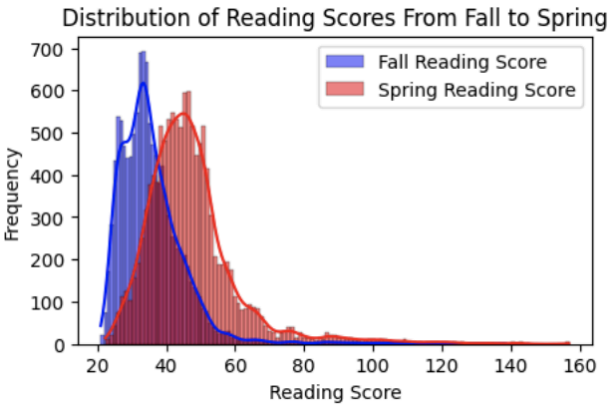
index	fallreadingscore	fallmathscore	fallgeneralknowledgescore	springreadingscore	springmathscore	springgeneralknowledgescore
count	11933.0	11933.0	11933.0	11933.0	11933.0	11933.0
mean	35.954	27.128	23.074	47.511	37.799	28.236
std	10.473	9.121	7.400	14.327	12.028	7.577
min	21.01	10.51	6.985	22.35	11.9	7.858
25%	29.34	20.68	17.385	38.95	29.27	22.802
50%	34.06	25.68	22.954	45.32	36.41	28.583
75%	39.89	31.59	28.305	51.77	44.22	33.782
max	138.51	115.65	47.691	156.85	113.8	48.345

- The summary statistics of that dataset indicate that kindergarten students show a clear progression in their academic abilities from fall to spring. The mean scores for reading and math have increased, with reading scores showing a particularly strong improvement. The standard deviations suggest a considerable spread in students' scores, implying varied individual performance levels. The range between the minimum and maximum values further underscores the wide disparity in students' academic achievements. This preliminary analysis sets the stage for a more detailed exploration of the factors influencing such educational outcomes.

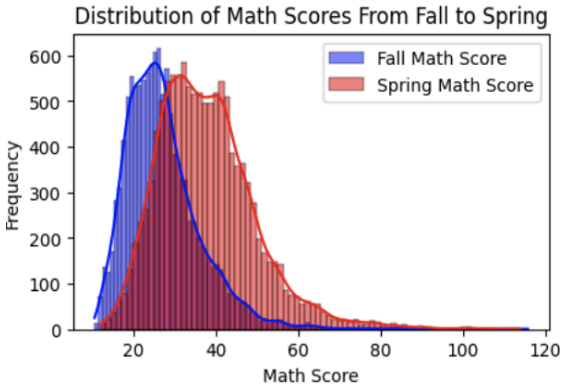


- The boxplot displays the scores distribution across subjects from fall to spring. There's a noticeable increase in median scores from fall to spring, with a considerable number of outliers suggesting variability among students. The spread of scores, as indicated by the interquartile range, is consistent across seasons, yet it widens slightly in spring, hinting at a greater disparity in student performance as the year progresses.

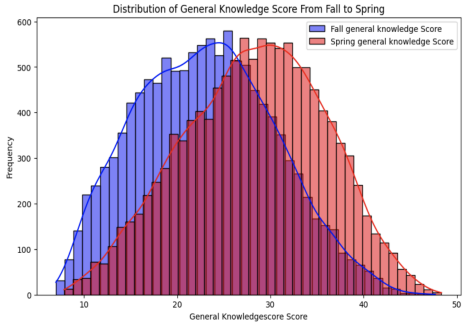
Next, let's delve into the score distributions for each subject from fall to spring, which will illuminate the nuances of students' academic growth.



- The reading scores histogram shows students' scores have shifted higher from fall to spring, indicating improved reading skills across the board. The spread of scores in spring is wider, suggesting not only general improvement but also greater variability in students' performance gains.



- Math scores also trend upward from fall to spring, with a wider distribution in the spring, which points to varied improvements among the students. The increase in spread indicates that while all students generally improved, the extent of their improvement is not uniform.



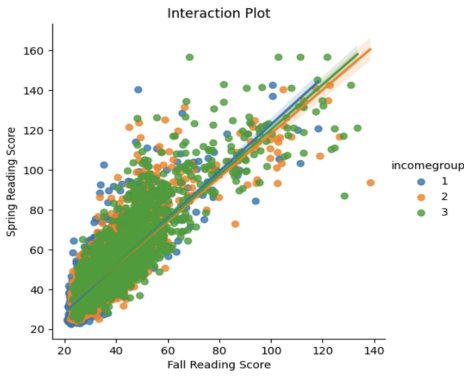
- For general knowledge, the spring scores exhibit a small shift toward higher scores, but the distribution remains quite similar to the fall. This suggests that while there is some improvement in general knowledge, it is less pronounced compared to the gains seen in reading and math.

4. Analysis of Covariance (ANCOVA)

In the ANCOVA section of our analysis, we initiate the process by examining interaction plots that correspond to our research questions, offering a preliminary glimpse into the potential relationships within our dataset. This visual appraisal is a crucial step, as it helps hypothesize about the interactions between variables. We then delve into a detailed interpretation of the ANCOVA results, which will quantify the impact of independent variables while controlling for covariates. The final phase of the analysis involves checking the assumptions of the ANCOVA. By ensuring that these assumptions hold, such as normality and homogeneity of variances, we can substantiate the statistical validity of our study's outcomes.

Research Question 1: Income group with Reading Score

- Interaction plot



The interaction plot indicates a positive correlation between fall and spring reading scores for all income groups, with the strength of this relationship varying by income level. Group 1 shows a gentle slope, suggesting modest gains from fall to spring. Group 2's steeper slope hints at a more significant improvement in reading scores over time. Group 3, the highest

income bracket, also demonstrates a steep increase, implying substantial score gains. The diverging slopes of the lines for each income group suggest that socioeconomic status may influence how students' reading abilities develop over the school year.

- ANCOVA results

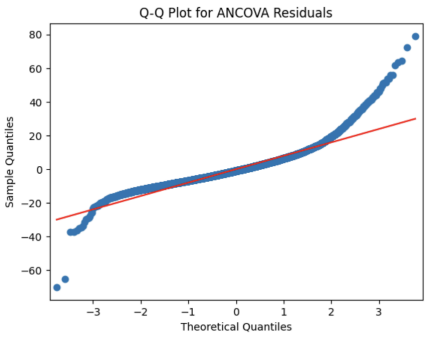
index	Source	F	p-unc	np2
0	incomegroup	4.055660265335827	0.017347930946815	0.0006795044724698
1	fallreadingscore	24455.397576420415	< 0.001	0.6721396863876946

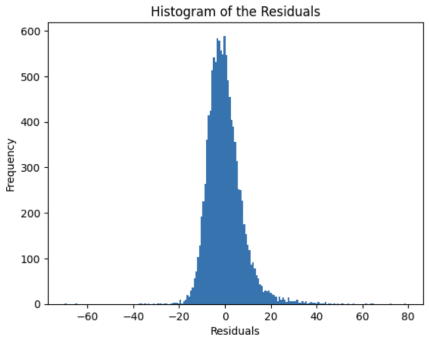
IncomeGroup: The F-value of approximately 4.056 and p-value of around 0.017 suggest that there is a statistically significant difference in reading scores across different income groups, but with a partial eta squared of about 0.00067, the actual impact of income on reading scores is very small.
Fall Reading Score: An extremely high F-value of around 24455.4 and a p-value of 0.0 indicate a very strong statistical association between fall reading scores and the dependent variable. The partial eta squared of approximately 0.672 indicates that fall reading scores are a major factor, explaining a significant portion of the variance in the reading scores.

	coef	P>|t|
Intercept	6.5430	< 0.001
C(incomegroup)[T.2]	0.3751	0.033
C(incomegroup)[T.3]	0.4898	0.008
fallreadingscore	1.1322	< 0.001

The coefficient for intercept is 6.5430 with a highly significant p-value ($p < 0.001$). This indicates that students from the second and third income groups score significantly higher on spring reading tests than the reference group, likely the first income group, with increases of 0.375 and 0.490 points respectively and both showing statistical significance (p-values < 0.05). Additionally, fall reading scores are strongly and positively correlated with spring reading scores, with each point increasing in the fall translating to a 1.132-point rise in the spring, a relationship that is highly statistically significant ($p < 0.001$). These results underscore the importance of both income group and fall reading performance as influential predictors of spring reading success.

- Ancova Assumptions:



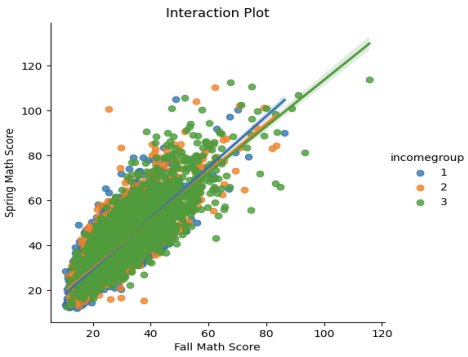


	statistics	p value
Shapiro	0.912	< 0.001
Levene	39.553	< 0.001

Normality of residuals: The Q-Q plot and histogram for the ANCOVA residuals, along with the Shapiro-Wilk test, suggest that the assumption of normality may not hold, as indicated by the deviation from the line in the Q-Q plot and the p-value <0.01 in the test.
Homogeneity of variances: Levene's test result with a p-value less than 0.01 suggests the variances are not equal across groups, violating the homogeneity of variance assumption for ANCOVA.

Research Question 2: Income group with Math Score

- Interaction plot



The interaction plot shows a positive correlation between fall and spring math scores across all income groups, with parallel trend lines indicating a consistent relationship regardless of income. This suggests no significant interaction effect; students' math score improvements from fall to spring don't vary by income group.

- ANCOVA results

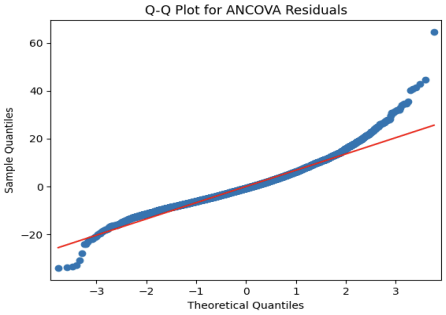
index	Source	F	p-unc	np2
0	incomegroup	18.5235850007212	9.284861e-09	0.00309602406501
1	fallmathscore	22203.0812384309	< 0.001	0.65050475777701

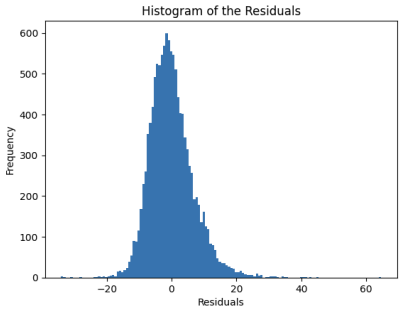
The ANCOVA results indicate that the income group significantly affects the dependent variable ($F = 18.52$, $p < 0.001$, $np^2 = 0.003$), though with a small effect size. The fall math score is a strong predictor, with an overwhelming effect size ($F = 22203.08$, $p < 0.001$, $np^2 = 0.650$).

	coef	P>|t|
Intercept	8.2011	< 0.001
C(incomegroup)[T.2]	0.6700	< 0.001
C(incomegroup)[T.3]	0.9199	< 0.001
fallmathscore	1.0735	< 0.001

The ANCOVA results indicate that after controlling for fall math scores, both income group 2 and income group 3 have significantly higher spring math scores compared to the reference income group (likely income group 1), with increases of 0.6700 and 0.9199 units, respectively. Additionally, fall math scores are a significant predictor of spring math scores; for every one-unit increase in the fall score, the spring score increases by 1.0735 units. All these findings are statistically significant with p-values less than 0.001.

- Ancova Assumptions:



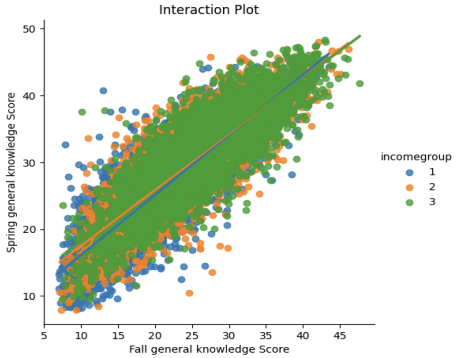


	statistics	p value
Shapiro	0.9649	< 0.001
Levene	18.8999	< 0.001

Similar to the first question, the Q-Q plot, histogram, and the Shapiro-Wilk test with $P < 0.001$ suggest that the assumption of normality may not hold. Levene's test result with a p-value < 0.01 suggests that the homogeneity of variance assumption is not met either.

Research Question 3: Income group with General Knowledge Score

- Interaction plot



Like the other two plots, this interaction plot indicates that higher income groups have a stronger positive correlation between fall and spring general knowledge scores, with income group 3 showing the greatest

increase, suggesting income may influence the growth of general knowledge scores over time.

- ANCOVA results

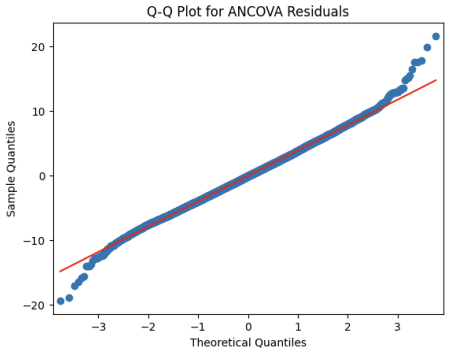
index	Source	F	p-unc	np2
0	incomegroup	56.9080398966	2.525291e-25	0.00945095225
1	fallgeneralknowledgescore	26682.2698405	<0.001	0.69104875210

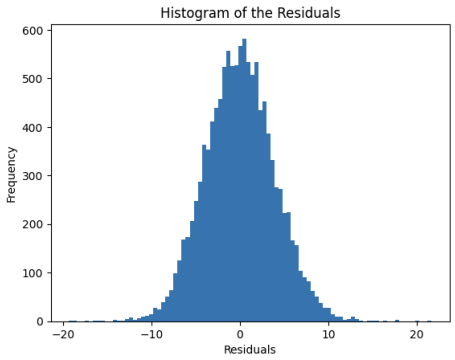
The ANCOVA results indicate that the income group has a significant but small effect on spring general knowledge scores ($F = 56.98$, $p < .001$, $np^2 = 0.00945$), while fall scores are a highly significant and strong predictor of spring scores ($F = 26682.27$, $p < .001$, $np^2 = 0.69105$).

	coef	P>|t|
Intercept	8.0303	< 0.001
C(incomegroup)[T.2]	0.7084	< 0.001
C(incomegroup)[T.3]	0.9424	< 0.001
fallgeneralknowledge	0.8542	< 0.001

Similar from above,the results demonstrate that after controlling for fall general knowledge scores, both income group 2 and income group 3 have significantly higher spring general knowledge scores compared to the reference group, with increases of 0.7084 and 0.9424 units, respectively. Moreover, fall general knowledge scores are a significant predictor of spring scores; for every one-unit increase in the fall score, the spring score increases by 0.8542 units. All these findings are statistically significant with p-values less than 0.001.

- Ancova Assumptions:





	statistics	p value
Shapiro	0.9980	3.155e-11
Levene	9.4058	0.0001

Lastly, the normality assumption and homogeneity of variance assumption both failed from the graphs the extremely low p-value for both Shapiro and Levene tests.

5. Conclusion:

Our analysis revealed that fall academic performance is a robust predictor of spring scores in reading, math, and general knowledge for kindergarten students. Income group also emerged as a significant factor, albeit with a smaller effect size, indicating that socioeconomic status does play a role in academic progress, but to a lesser extent than the continuity of individual performance.

However, the study encountered limitations. The assumptions of homogeneity of variance and normality were not met, as indicated by the Shapiro-Wilk and Levene's tests. These violations necessitate a careful approach to interpreting the ANCOVA results and suggest the potential benefit of employing alternative analytical methods or data transformations in future research.