Understanding Early Childhood Longitudinal study

1. Project overview

The dataset comes from an early childhood longitudinal study and includes reading, math, and general knowledge scores, as well as students' income group. We want to compare students' reading and math scores change over time by income group.

Our exploration will address three research questions, which is,

1. What is the impact of income group on spring exam scores while accounting for students' fall grades?
2. Is there any linear relationship between fall score and spring score?
3. Is there any difference between different groups when we consider the progress the students made?

2. Data cleaning, preprocessing

A. Dataset Overview

The dataset contains 9 columns, the first 6 columns is the score of math, reading, general knowledge in fall and spring, the last 3 columns is the total income and income group for each student.

B. Feature Engineering

To ensure the quality and availability of data, we do data cleaning and preprocessing including the following parts:

- Delete unnecessary columns such as 'totalhouseholdincome'.
- Create column progress_rate for each subject, we want to see how much progress a student makes in each subject, the equation is shown below,

$$progressrate = \frac{springscore - fallscore}{fallscore} * 100$$

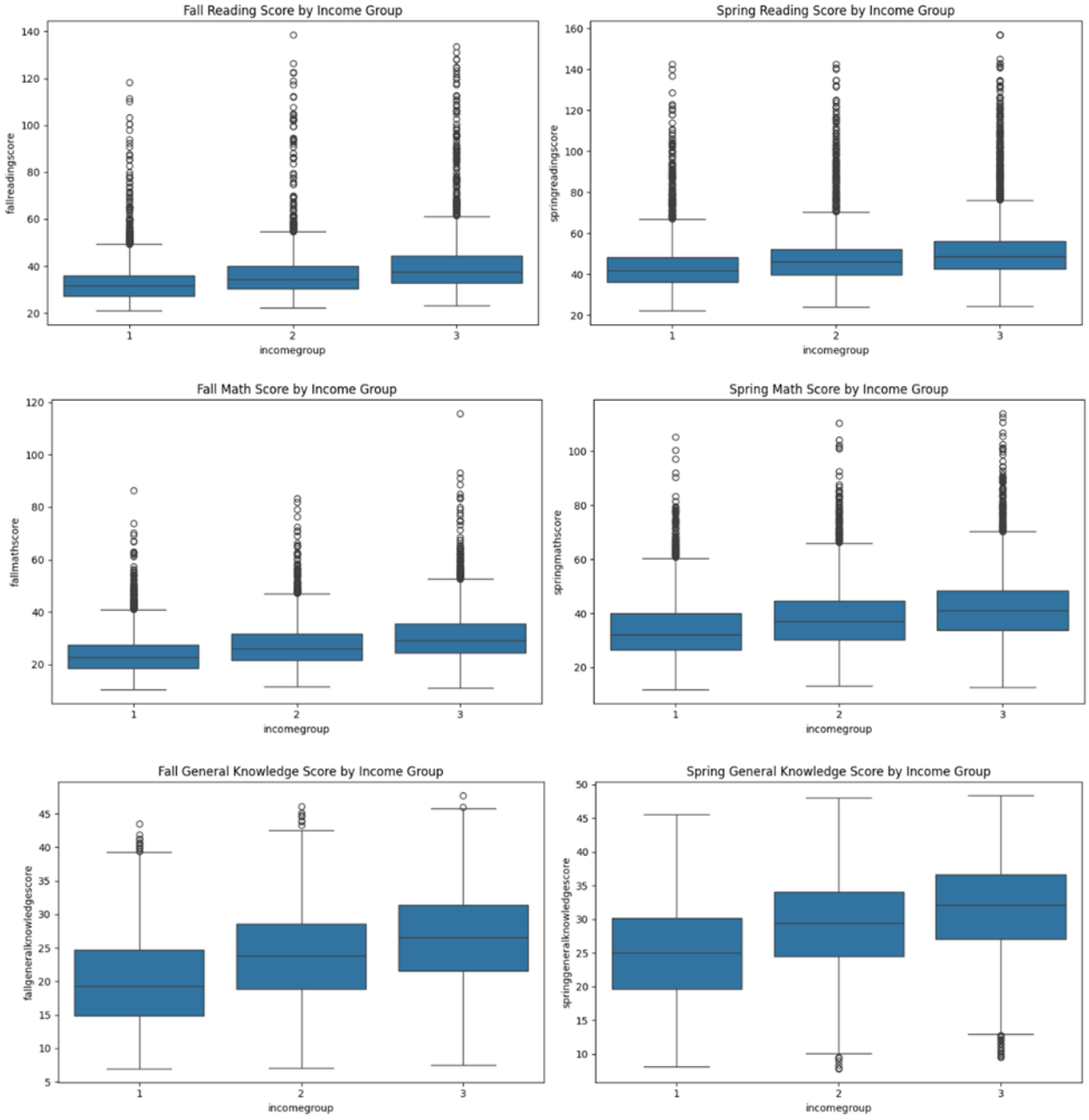
3. Exploratory Data Analysis

First, we start by describing our quantitative data analysis to have a visualization on these three subjects in fall and spring. This descriptive analysis, including mean, standard deviation, minimum, quartiles, median, and maximum values, provide a structured overview of the distribution and variability of scores within each subject and across the academic terms.

	fallreading	fallmath	fallknowledge	springreading	springmath	springknowledge
mean	35.95	27.13	23.07	47.51	37.79	28.23
std	10.47	9.12	7.39	14.33	12.03	7.58
min	21.01	10.51	6.99	22.35	11.9	7.86
25%	29.34	20.68	17.39	38.95	29.27	22.8
50%	34.06	25.68	22.95	45.32	36.41	28.58
75%	39.89	31.59	28.31	51.77	44.22	33.78
max	138.51	115.65	47.69	156.85	113.8	48.35

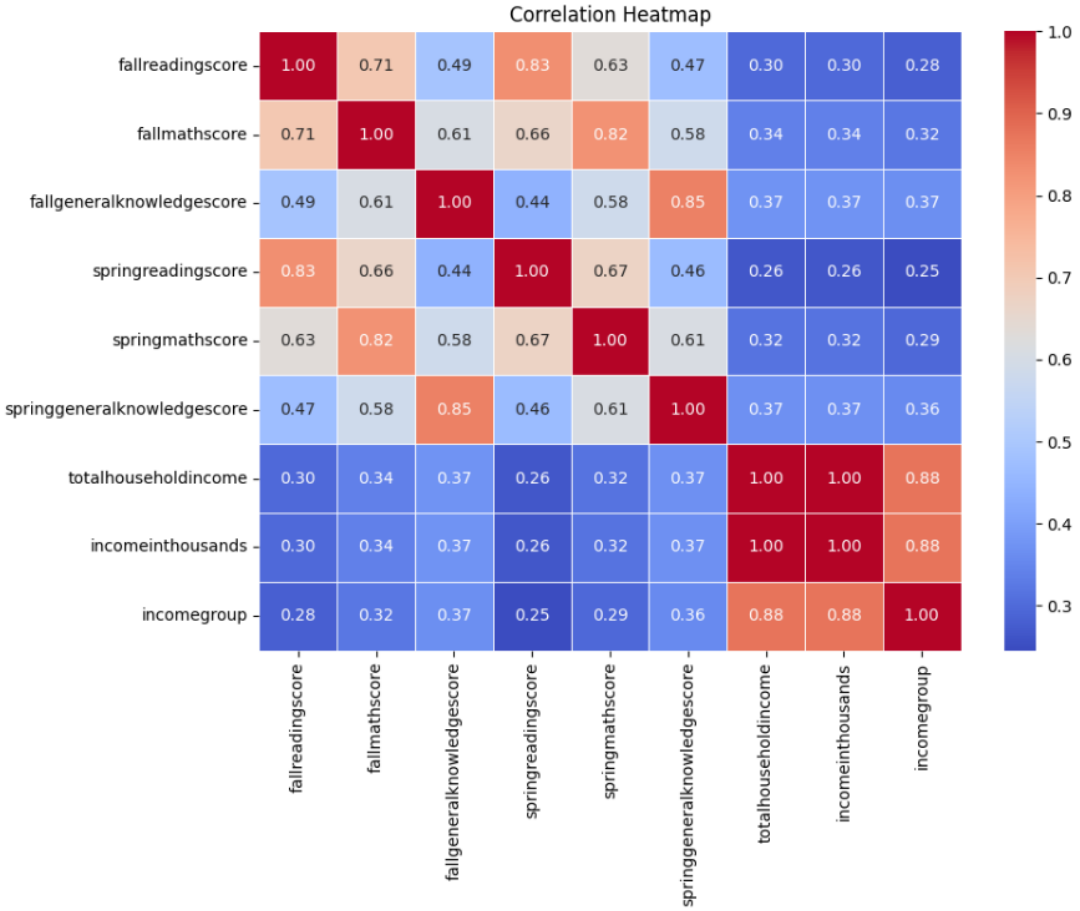
From statistical value above, we can see that reading score is the highest among all the subjects. And there is a general improvement from fall to spring across all subjects. However, the standard deviations are higher for spring scores, suggesting greater variability in spring performance.

To gain a comprehensive understanding of these data, we employ boxplots as a visualization tool. Boxplots provide a clear depiction of the distribution, central tendency, and variability of the data across different subjects and seasons. By examining boxplots for each subject in both fall and spring, we can identify trends, outliers, and overall patterns that can inform our analysis.



From the boxplots above, we can see that students with higher incomes tend to achieve higher scores. Moreover, the presence of noticeable outliers in math and reading scores suggests that these subjects may vary significantly in terms of learning difficulty or student performance. This implies that income level may play a crucial role in students' education.

Additionally, we want to explore the correlation between different variables, so we plot the correlation map below.



The correlation heatmap indicates a strong relationship between fall and spring scores, suggesting consistent performance over time. Additionally, there appears to be a potential relationship between math and reading scores, indicating that performance in one subject may influence the other. However, the income group shows a low correlation with all subjects, implying that income level may not directly impact academic performance in this dataset.

4. Impact of income group

Research Question 1: What is the impact of income group on spring exam scores while accounting for students' fall grades. For this research question, we perform One-Way ANCOVA to explore the data.

> ANCOVA (Analysis of Covariance) is a statistical analysis method that combines the features of Analysis of Variance (ANOVA) and Regression Analysis. It is used to evaluate the effects of one or more categorical

independent variables (factors) on a continuous dependent variable while controlling for the influence of one or more continuous covariates. In this experiment, we study the impact of income group on spring exam scores while accounting for students' fall grades.

The math score's result is shown below.

	coef	std err	t	P>|t|
Intercept	8.201	0.199	41.273	0
incomegroup[T.2]	0.67	0.151	4.430	0
incomegroup[T.3]	0.919	0.160	5.741	0
fallmathscore	1.073	0.007	149.007	0

	F	p-unc	np2
incomegroup	18.523	9.285e-09	0.003
fallmathscore	22203.081	0	0.651

The reading score's result is shown below.

	coef	std err	t	P>|t|
Intercept	6.543	0.264	24.779	0
incomegroup[T.2]	0.375	0.176	2.130	0.033
incomegroup[T.3]	0.49	0.160	2.648	0.008
fallreadingscore	1.132	0.007	156.382	0

	F	p-unc	np2
incomegroup	4.056	0.0173	0.00068
fallreadingscore	24455.397	0	0.672

The knowledge score's result is shown below.

	coef	std err	t	P>|t|
Intercept	8.03	0.119	67.519	0
incomegroup[T.2]	0.708	0.088	8.005	0
incomegroup[T.3]	0.942	0.094	10.013	0
fallgeneralknowledgescore	0.854	0.005	163.347	0

	F	p-unc	np2
incomegroup	56.908	2.525e-25	0.009
fallgeneralknowledgescore	26682.269	0	0.691

Conclusion: Income group has a significant impact on math scores, with incomegroup[T.2] and incomegroup[T.3] having p-values of 0, indicating a significant effect. Income group 3 has the largest impact, followed by income group 2.

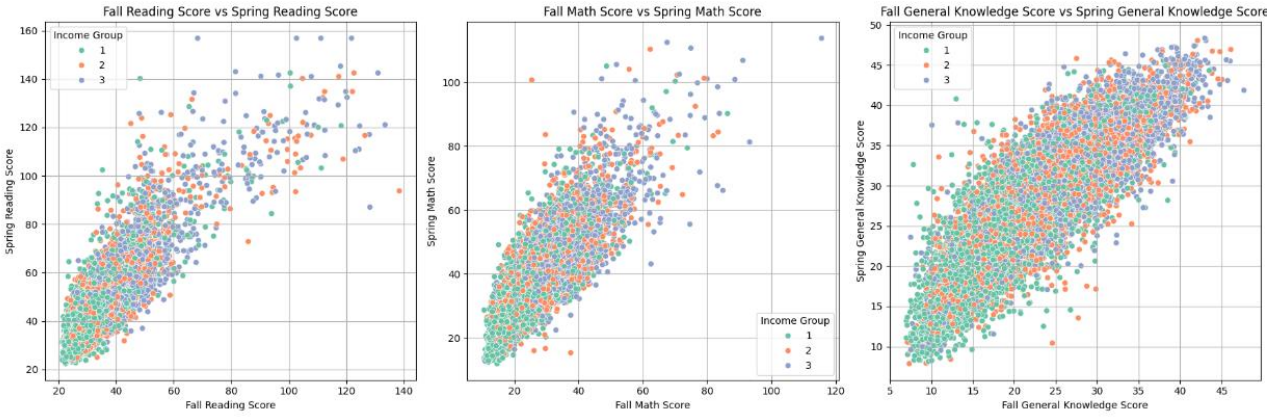
Fall score also significantly affects spring math score. In the ANCOVA table, both incomegroup and fallmathscore have F-values significantly below 0.05, also indicating a significant impact of income group and

fall math score on spring math score. So it can be inferred that family income level has a significant impact on students' spring general knowledge scores, especially for students from higher-income families who are more likely to perform well.

5. Relationship between spring and fall score

Research question2: Is there any linear relationship between fall score and spring score?

To observe the relationship between spring scores and fall scores and help us determine whether there is correlation, a linear relationship, or other trends between two variables, we also plot a scatter plot for spring and gall scores, which is shown below,

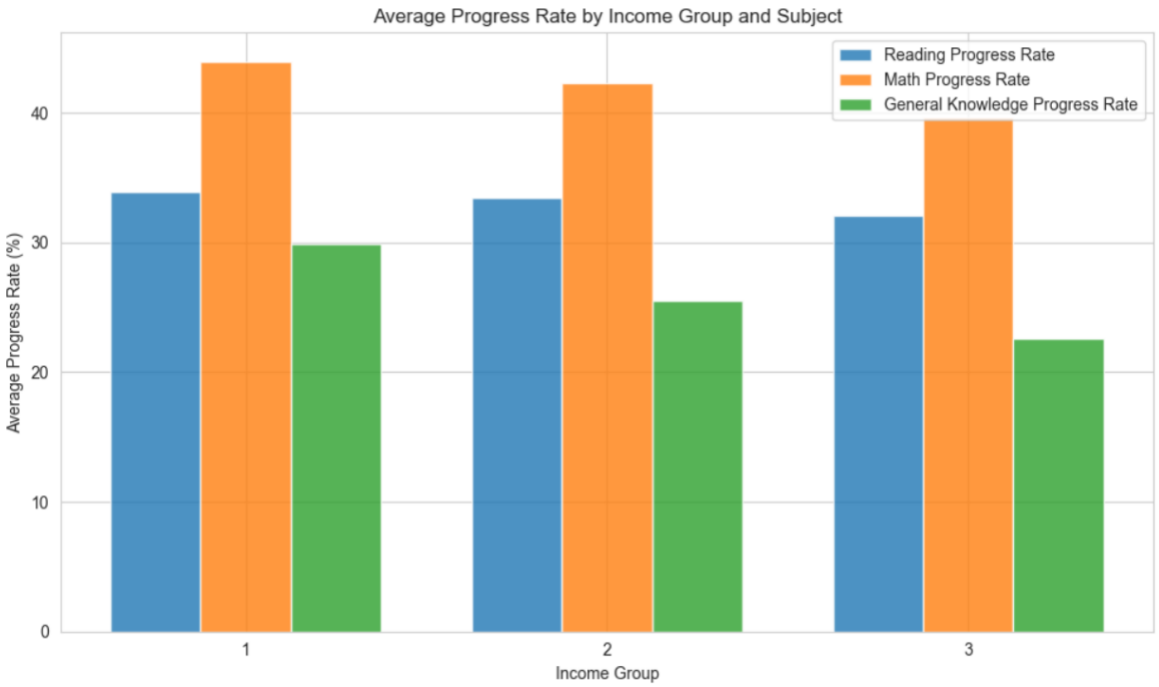


The scatter plot would show a pattern where as the fall scores increase, the spring scores also tend to increase or decrease in a linear manner. it is evident that there exists a linear relationship between spring scores and fall scores.

The results obtained from this visual analysis would suggest that students who perform well in the fall exams also tend to perform well in the spring exams, indicating a positive correlation between fall grades and spring scores. The strength of this correlation would be reflected in the slope of the trend line in the scatter plot.

6. Progress evaluation

Research Question 3: Is there any progress from fall to spring and what is the difference between different groups for each subject. To explore this question, we visualize the progress rate for each subject and each income group.



The plot illustrates the progress rates in reading, math, and general knowledge across various income groups. It is evident that students generally show more progress in math compared to reading and general knowledge. Specifically, income group 1 exhibits the highest progress rate across all subjects. This can be attributed to their lower initial scores in the fall, indicating a significant improvement over time.

From these observations, we can infer that lower initial scores may motivate students to strive for greater improvement, leading to higher progress rates.

7. Conclusion

In summary, the analysis reveals several key findings regarding students' academic performance, income group impact, and previous score impact. Based on all analysis results, we can conclude that both income group and fall score significantly affect spring score, with income group 3 having the greatest impact. Therefore, education policymakers may consider implementing differential education policies tailored to different income groups to better meet the learning needs of diverse student populations.

8. Further analysis

Currently we are exploring One-Way ANCOVA for the impact of income group on spring exam scores while accounting for students' fall grades. In future analysis, we can analyze the trend of specific income values rather than income groups on scores, we can use **regression analysis** or other statistical models. Regression analysis allows us to quantify the relationship between income values and scores and understand the impact of income value on scores over time.