INF2178 Experimental Design For Data Science

Technical Assignment #3

Student Name: Liguangxuan He | Student ID: 1006141809

Background Information About This Data

This assignments' data-set is derived from an early childhood longitudinal study conducted in 1998-1999, focusing on kindergarten students. It includes their scores in reading, math, and general knowledge both at the beginning and end of the school year (fall 1998 and spring 1999, respectively) and categorizes students by income level (1, 2 or 3) of their parents. Statistical summaries of these 11933 data points reveal average scores in reading, math, and general knowledge increase from fall to spring, suggesting academic growth over several months. Additionally, the data outlines household income, showing a broad range from very low to high, with a mean annual household income of approximately $54,320. This comprehensive data-set not only allows for the assessment of academic progress over a school year but also offers insights into how socioeconomic factors may influence educational outcomes.

---

1. Explanatory Data Analysis (EDA)

In this part of my analysis, I delve into how students' academic performance are influenced by time and socioeconomic status. By presenting the data through a series of graphical representations, aims to better understand and explain why some students get better grades over time and if their parents make more money at home has anything affecting it. So, following figures from 1 to 3 serves as visual anchors for this exploration, guiding me through the variances in score improvements.

Research Question #1: How do students' scores improve from fall to spring?

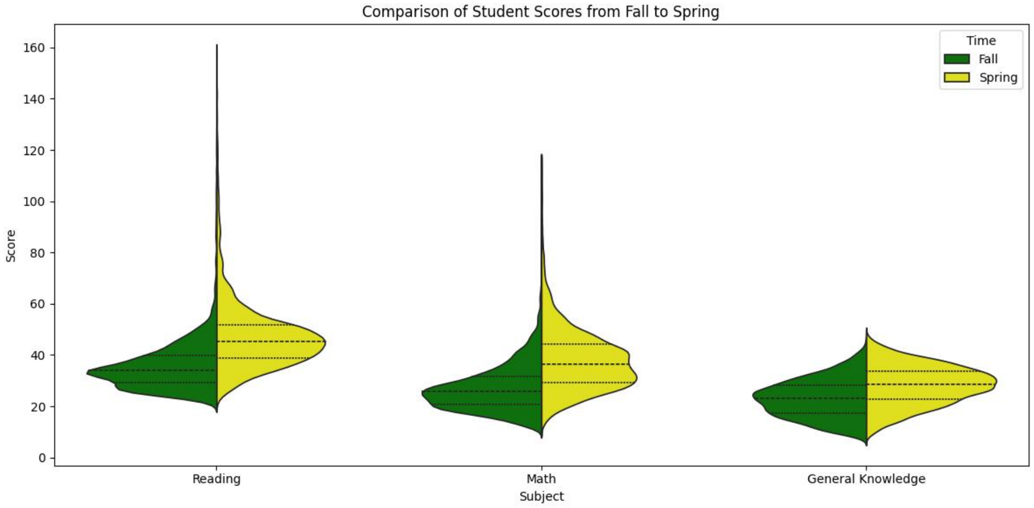

Figure 1. Comparison of Student Scores from Fall to Spring (Violin plot)

Referring to Figure 1, the violin plot reveals a noticeable advancement in student performance across reading, math, and general knowledge from the fall to the spring semester. In reading and math, the spread and median of the scores rise, with the spring semester showing a wider distribution and higher peaks, suggesting significant improvement. General knowledge scores exhibit a more modest elevation, with the spring scores slightly higher, indicating a less pronounced but consistent gain. Overall, the broader spring score distributions in all subjects imply not only enhanced performance but also increased variation in student achievements as the year progresses.

Research Question #2: Is there a relationship between household income and students' scores?

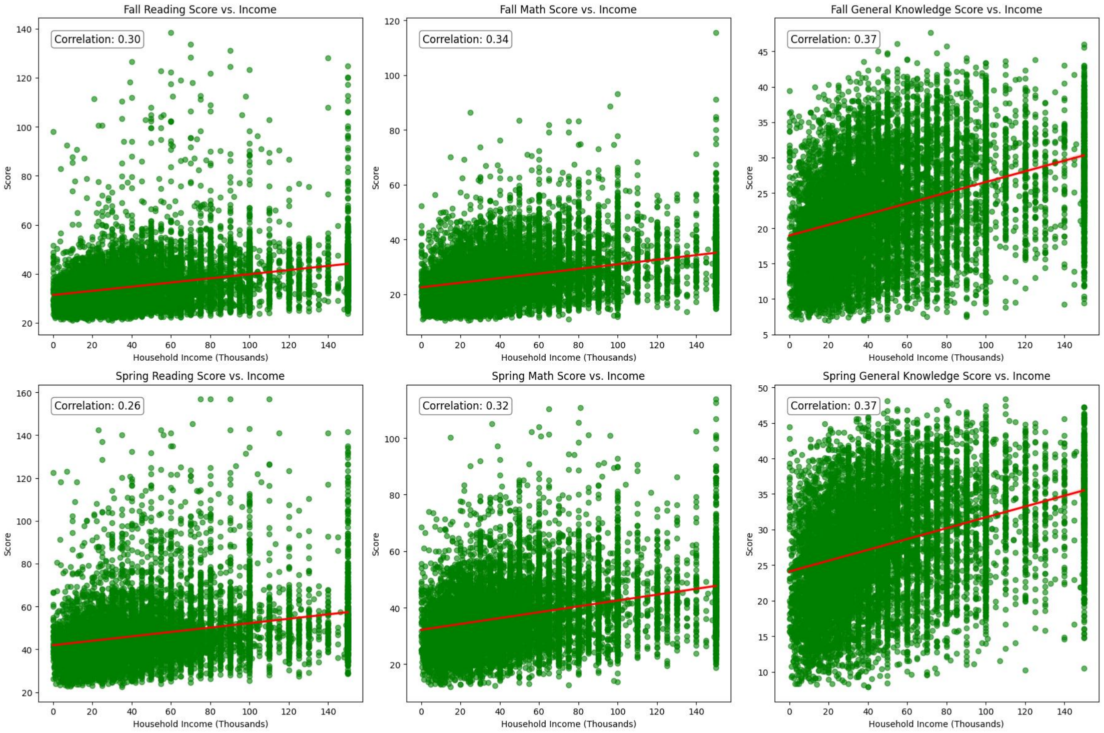

Figure 2. Regression Plot Graphs with Correlation Coefficient Labeled

Referring to Figure 2, the scatter plots present positive correlation between household income and student scores, with the strength of the relationship denoted by correlation coefficients from 0.26 to 0.37. In the 1998 fall data, the correlation for reading is 0.30, for math 0.34, and general knowledge shows the strongest relationship at 0.37. Moving into the spring of 1999, the reading score's correlation with income slightly weakens to 0.26, whereas math sees a marginal decrease to 0.32, and general knowledge maintains a consistent correlation of 0.37. The spread of scores is notably wider at higher income levels, indicating variability in outcomes within wealthier households. The data from fall 1998 and spring 1999 combined suggests that while household income generally aligns with better academic performance, its impact varies across subjects and over the course of the school year.

Research Question #3: Does the income group affect the scores' improvement from fall to spring?

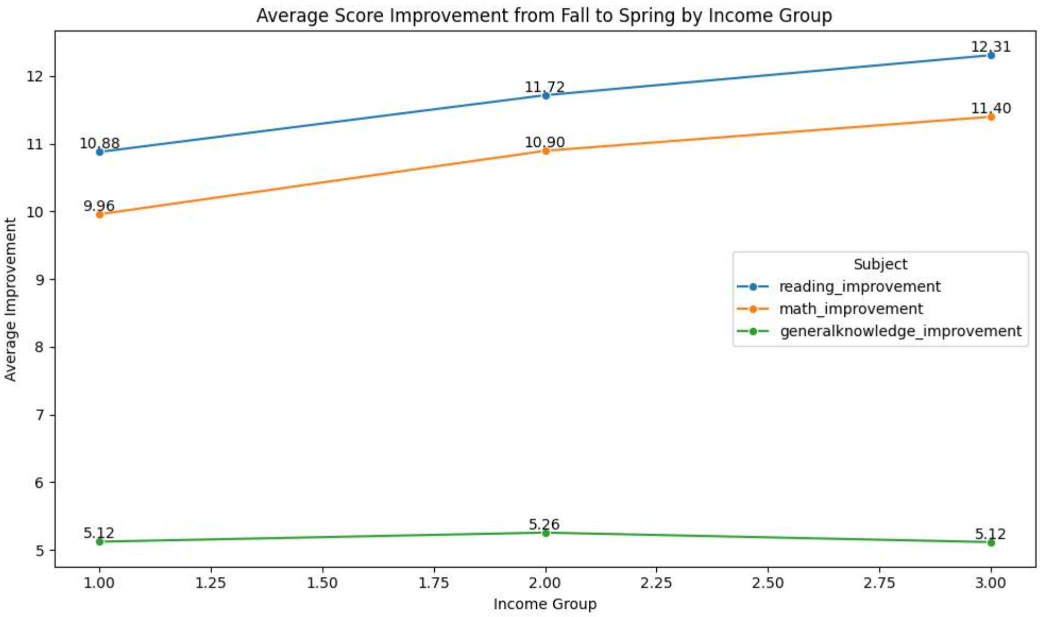

Figure 3. Average Score Improvement from Fall to Spring by Income Group

Referring to Figure 3, the line graph illustrates that students from higher-earning households tend to exhibit greater average academic gains from fall to spring. In reading, for instance, the progression is evident with a notable increase from an improvement score of 10.88 in the lowest income group to 12.31 in the highest. Math scores follow a similar upward trend, with the lowest income group showing an average improvement of 9.96, which rises to 11.40 in the wealthiest group. Interestingly, general knowledge scores peak with the most substantial improvement of 5.26 in the middle income group, then slightly retreat to 5.12 in the highest income bracket, suggesting a more complex interplay between income and gains in general knowledge.

---

2. Prior Visualization

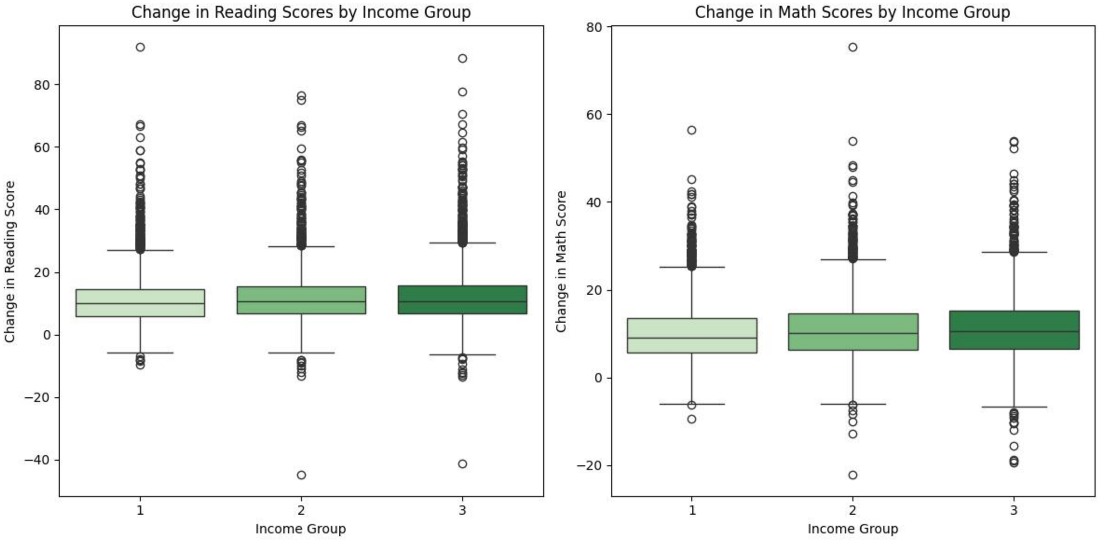

Figure 4. Change in Reading & Math Scores by Income Group

The box plots visualize the changes in reading and math scores by income group, providing a comparison of educational progress across different socioeconomic statuses. The median improvement across all income groups is above zero, suggesting a general increase in scores from fall to spring. However, the range of improvements is quite broad, especially in the highest income group, which shows both higher median improvement and greater variability, as indicated by a larger inter-quartile range and more extreme outliers. For math, the median improvements are consistent across income groups, with a slight uptick in the higher income category, but less variability compared to reading, as evidenced by tighter inter-quartile ranges and fewer outliers. These trends suggest that while students from all backgrounds are improving, those from higher-income families might have access to resources or opportunities that contribute to a greater range of score improvement, particularly in reading.

3.1 One-way ANCOVA analysis

In the ANCOVA model, I will examine the change in reading scores across different income groups, while adjusting for students' initial general knowledge scores from the fall. This analytical approach seeks to determine the extent to which students' reading proficiency, as it develops from fall to spring, can be attributed to their income group when baseline academic knowledge is held constant. The hypothesis driving this analysis is that income group has a significant effect on the improvement of reading scores over time, with the expectation that this relationship will persist even when accounting for the students’ starting level of general knowledge.

	Coefficient	Standard Error	t-value	p-value	95% Confidence Interval
Intercept	7.73	0.24	31.96	<0.001	(7.257, 8.205)
Income Group 2 vs 1	0.22	0.18	1.21	0.23	(-0.136, 0.570)
Income Group 3 vs 1	0.40	0.19	2.11	0.04	(0.029, 0.779)
Fall General Knowledge Score	0.16	0.01	14.84	<0.001	(0.137, 0.179)

Parameter	Value
R-squared	0.023
Adjusted R-squared	0.023
F-statistic	95.49

Table 1. One-way ANCOVA summary results

Examining the ANCOVA model outcomes presented in Table 1, the data reveal a complex picture regarding how students' socioeconomic backgrounds impact their academic progress in reading. The estimated coefficient for students in Income Group 2, when compared with the baseline Income Group 1, is noted to be 0.22, but this difference fails to reach statistical significance, with a p-value of 0.23. Essentially, this means that the growth in reading scores for Income Group 2 over the academic year is not statistically different from that of Income Group 1. On the other hand, students from Income Group 3 show a distinct pattern; they have a coefficient of 0.40

compared to Income Group 1, and with a p-value of 0.04, this result is statistically significant. Therefore, there's an approximate 0.4-point greater improvement in reading scores for Income Group 3, which aligns with the prediction that students from higher-income backgrounds would demonstrate a noticeable increase in reading score improvements.

Further, the role of students' initial general knowledge is quite pronounced; with a coefficient of 0.16 and a p-value close to zero, the baseline general knowledge strongly correlates with reading improvement, asserting that regardless of income group, those with a more robust foundation in general knowledge at the start are likely to see more considerable gains in their reading scores. The R-squared value of 0.023 in the model suggests that only a minor fraction of the variability in reading score improvements is captured by these factors, hinting at the presence of additional variables beyond income group and initial general knowledge that could be contributing to changes in reading scores throughout the school year.

3.2 One-way ANCOVA: post hoc test using Tukey's HSD

group 1	group 2	MeanDiff	p-adj	Lower	Upper	reject
1	2	0.84	0.00	0.42	1.25	True
1	3	1.43	0.00	1.01	1.85	True
2	3	0.59	0.0053	0.15	1.04	True

Table 2. One-way ANCOVA post hoc results

The post hoc analysis using Tukey's HSD as part of the one-way ANCOVA reveals statistically significant differences in reading score improvements between all income groups. Specifically, the mean difference in improvement between Group 1 and Group 2 is 0.84 units, with a p-value effectively at zero, indicating a significant increase for Group 2. Even more substantial is the difference between Group 1 and Group 3, which stands at 1.43 units, again with a p-value of nearly zero, denoting a significant leap in reading scores for Group 3 over Group 1. Additionally, when comparing Group 2 and Group 3, there is a mean difference of 0.59 units with a p-value of 0.0053, suggesting a smaller yet still significant difference in score improvements favoring Group 3. These results, summarized in Table 2, consistently affirm that reading score improvements are significantly influenced by income group, with higher income groups exhibiting greater gains.

3.3 One-way ANCOVA Analysis Assumption Check

Assumption 1: Normality of residual

Referring to Figure 5. The histogram and Q-Q plot are employed to assess the normality of residuals. The histogram shows the distribution of residuals is roughly bell-shaped, suggesting an approximation to a normal distribution. However, there is a slight left skewness as evidenced by a longer tail on the left side. The Q-Q plot further aids this analysis; most data points fall along the reference line, which implies that the residuals are normally distributed for the most part. But there are deviations, particularly at the ends of the distribution, indicating potential outliers or extreme values. These outliers are more pronounced in the tails and may affect the normality assumption. Despite these slight deviations, the overall pattern does not suggest a severe departure from normality, although the exact adherence to the normality assumption might require a more rigorous statistical test to confirm.

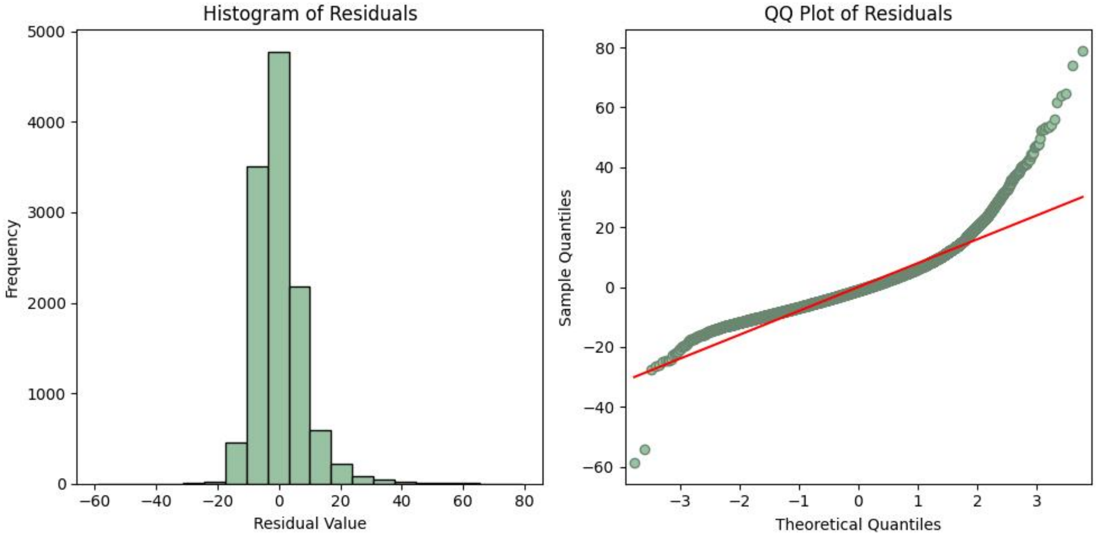

Figure 5. Histogram and QQ plot for residuals

Assumption 2: Homogenity of variances (Levene's test)

Parameter	Value
Test Statistic (W)	26.96
df1 (Between Groups)	2.0000
df2 (Within Groups)	1193
p-value	0.0000

Table 3. Levene's test results

The results of Levene's test, as detailed in Table 3, are crucial for evaluating the assumption of homogeneity of variances across groups in the ANCOVA analysis. A significant Levene's test indicates that the assumption of equal variances has been violated. The test statistic (W) is 26.96 with a p-value of almost zero, which is highly significant. This suggests that the variances in reading improvement scores are not consistent across the three income groups. With the degrees of freedom between groups (df1) at 2 and within groups (df2) at 11930, the test has ample power to detect any inequality of variances. Given these results, can infer that the variances of reading score improvements differ significantly between the income groups.

Conclusion

This study's analysis, rooted in a 1998-1999 longitudinal dataset, has illuminated the impact of socioeconomic factors on the academic growth of Kindergarten students. The ANCOVA results substantiate a clear socioeconomic gradient in reading score improvements, maintained even after adjusting for baseline knowledge. However, challenges in assumptions, like homogeneity of variances, suggest the need for nuanced approaches in future analyses. The forthcoming ANCOVA on math score changes will further explore these educational dynamics, leveraging prior insights to deepen our understanding of the interplay between household income and learning progress.