

**INF2178 LEC0101**

**Experimental for Data Science**

**Assignment 3**

**Exploring Kindergarten Scores**

**March 23, 2024**

**Guanxiao Chen**

**#1005717841**

## 1. Introduction

The early years of education lay the foundation for a lifetime of learning and development. In this assignment, understanding the trajectory of academic growth among kindergarten students stands as a crucial inquiry. This analysis engages with the dataset titled INF2178\_A3\_data.csv, which offers a glimpse into the academic performance of kindergarten pupils across two pivotal periods: fall 1998 and spring 1999. It encompasses a variety of measures, including reading, math, and general knowledge scores, alongside an income category variable.

To structure our analysis, we have formulated the following research questions:

**Research question 1:** How does total household income influence the improvement in reading, math, and general knowledge scores from fall to spring among kindergarten students?

**Research Question 2:** Are there significant differences in the academic performance growth of kindergarten students from various income groups?

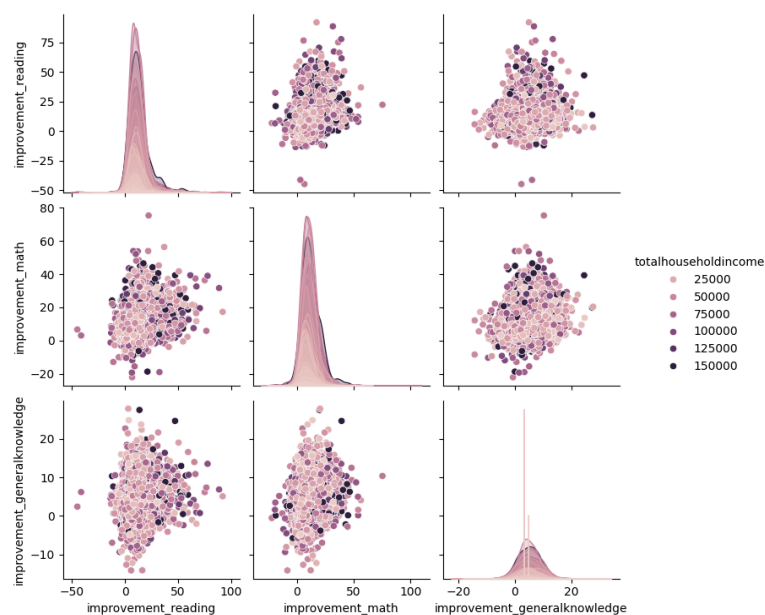
## 2. Data Cleaning

The raw dataset has a total of 9 columns with 11933 entries. After reviewing the dataset columns in relation to the research questions provided, it appears that most columns serve a purpose in addressing the research questions, so we don't need to drop columns in this analysis. After checking the missing values, there are 0 missing values among all columns, therefore, we don't need to fill out the missing values, this dataset is clean enough.

## 3. Impact of Household Income

Research Question 1: How does total household income influence the improvement in reading, math, and general knowledge scores from fall to spring among kindergarten students?

### - Exploratory Data Analysis



The scatter and density plots suggest that improvements in reading, math, and general knowledge scores among kindergarten students do not show a clear correlation with household income levels. The dense clustering of data points across income brackets and the normally distributed score improvements imply that factors other than income might play a significant role in educational advancement across these subjects.

### - One-way ANOVA

ANOVA results for reading:

	sum_sq	df	F	PR(>F)
C(incomegroup)	230.61	2.0	1.78	0.17
totalhouseholdincome	1130.73	1.0	17.41	0.000030
Residual	774594.44	11929.0	NaN	NaN

ANOVA results for math:

	sum_sq	df	F	PR(>F)
C(incomegroup)	538.10	2.0	5.79	3.060142e-03
totalhouseholdincome	2189.35	1.0	47.13	6.964940e-12
Residual	554117.09	11929.0	NaN	NaN

ANOVA results for general knowledge:

	sum_sq	df	F	PR(>F)
C(incomegroup)	74.28	2.0	2.26	0.10
totalhouseholdincome	30.58	1.0	1.86	0.17
Residual	196110.72	11929.0	NaN	NaN

The ANOVA results provide insight into the influence of income group and total household income on the improvement of academic scores. For reading scores, total household income shows a significant effect ( $p < 0.0001$ ), indicating that as household income increases, so does the improvement in reading scores. However, the income group as a categorical variable does not significantly affect reading score improvement ( $p > 0.05$ ).

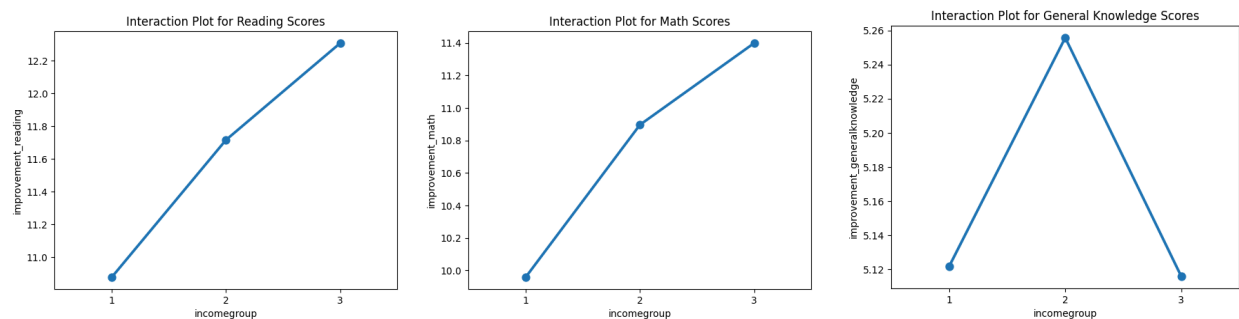
In the case of math scores, both the income group ( $p < 0.01$ ) and total household income ( $p < 1e-11$ ) show significant effects, suggesting that students from higher income groups and with higher total household income tend to show greater improvements in math scores over the school year.

For general knowledge scores, neither income group ( $p > 0.1$ ) nor total household income ( $p > 0.1$ ) significantly predicts improvement, which may indicate that factors other than income are more influential in the acquisition of general knowledge.

The `nan` values for the Levene's test statistic and p-value indicate that the test could not be conducted, likely due to a technical issue or missing data. This raises concerns about the assumption of equal variances across groups, which is critical for the validity of ANOVA. Further investigation is needed to ensure that the data meet the necessary assumptions for ANOVA.

Given these findings, we can conclude that while household income seems to be a strong predictor for improvements in reading and math scores, it does not appear to have the same influence on general knowledge scores. The lack of significance for the income group variable in reading, and for both predictors in general knowledge, suggests that there are other variables at play that may need to be explored to fully understand the dynamics of academic improvement.

### - Interaction Plots



The interaction plots for reading and math scores show a clear positive correlation with income group: as the income group increases, so does the improvement in scores. This trend is linear and consistent across both reading and math, suggesting that students from higher-income groups tend to make more significant academic progress in these areas over the school year.

The plot for general knowledge scores, however, shows an unexpected trend. The highest improvement is seen in the middle income group, rather than a linear increase across groups. This could imply that the highest income does not guarantee the largest gains in general knowledge, or that other factors may play a more crucial role in this subject area.

These patterns may reflect the varying impact of socioeconomic factors on different types of learning, with some subjects more influenced by economic conditions than others. This could be valuable for educators and policymakers looking to address educational inequalities.

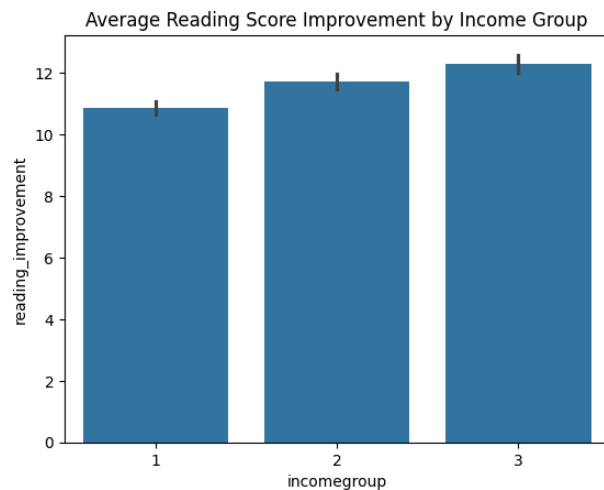
### - Post Hoc Test

Tukey's HSD test results for reading and math show that kindergarten students from higher income groups have significantly greater score improvements than those from lower income groups, confirming a strong socioeconomic gradient in learning outcomes. However, for general knowledge, income-related differences are not statistically significant, indicating that improvement in this area may not be as influenced by socioeconomic status.

## 4. Evaluating Socioeconomic Disparities

Research Question 2: Are there significant differences in the academic performance growth of kindergarten students from various income groups?

### - Exploratory Data Analysis



The bar chart presents the average reading score improvement for kindergarten students across three distinct income groups. The incremental increase in average score improvement is observable as we ascend from income group 1 to 3. This visual trend corroborates the findings from the Tukey's HSD test, highlighting a socioeconomic influence on educational advancement in reading. The error bars, likely representing a measure of variability such as the standard error or confidence interval, suggest that while there is some variance in score improvement within each income group, the mean differences are consistent enough to discern a pattern of higher income correlating with greater average improvement in reading scores.

### - One-way ANOVA

ANOVA results for reading\_improvement:

	sum_sq	df	F	PR(>F)
C(incomegroup)	4237.40	2.0	32.59	7.718776e-15

Residual	775725.16	11930.0	NaN	NaN
----------	-----------	---------	-----	-----

ANOVA results for math\_improvement:

	sum_sq	df	F	PR(>F)
C(incomegroup)	4433.58	2.0	47.54	2.728277e-21
Residual	556306.43	11930.0	NaN	NaN

ANOVA results for gk\_improvement:

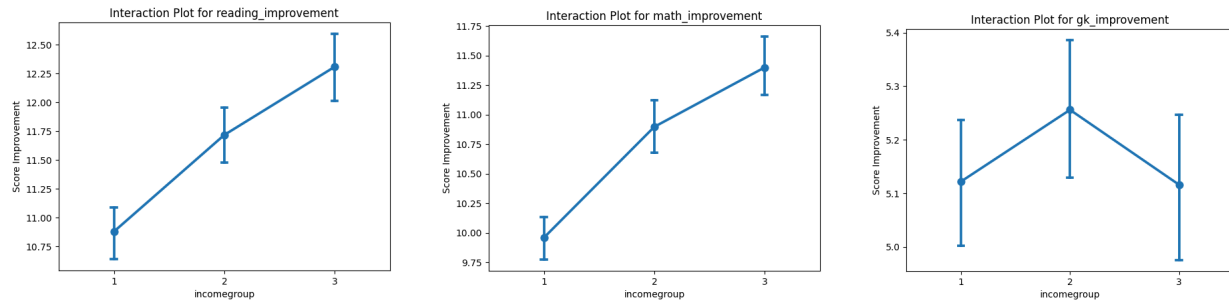
	sum_sq	df	F	PR(>F)
C(incomegroup)	47.77	2.0	1.45	0.23
Residual	196141.29	11930.0	NaN	NaN

The one-way ANOVA results for improvement in reading and math scores among kindergarten students demonstrate a statistically significant effect of income group on score increases. For reading score improvement, the F-statistic is 32.58 with a highly significant p-value ( $p < 1e-14$ ), indicating a strong association between income group and reading score gains. Similarly, math score improvement is significantly associated with income group, evidenced by an even larger F-statistic of 47.54 and a p-value effectively at zero ( $p < 1e-20$ ). These results suggest that income group is a strong predictor of academic progress in both reading and math, with higher income groups likely experiencing more significant improvements.

In contrast, the ANOVA for general knowledge score improvement yields an F-statistic of 1.45 with a p-value of approximately 0.234, which is not statistically significant. This indicates that there is no evidence to suggest that income group has an effect on improvement in general knowledge scores within this sample of students.

The consistency of significant results in reading and math underscores the potential influence of socioeconomic status on core academic skills. Conversely, the nonsignificant results for general knowledge score improvements may suggest that this subject area is less impacted by socioeconomic differences or that other factors play a more critical role in general knowledge acquisition.

#### - Interaction Plots



The interaction plots for reading and math score improvements show a consistent positive trend with higher income groups, indicating that students from wealthier families tend to show greater academic progress. The confidence intervals narrow as income increases, suggesting more precision in the mean score improvements for higher-income groups.

For reading improvement, the plot shows a steady increase in score improvements from the lowest to the highest income group. The error bars indicate some variability within groups, but the overall upward trend is clear.

The math improvement plot exhibits a similar pattern, with an upward trajectory from the lowest to the highest income group. Again, this suggests that students from higher-income families are making more substantial gains in math throughout the school year.

However, the plot for general knowledge improvement deviates from this pattern. The middle-income group shows a peak in improvement, higher than both the lower and higher income groups, with the latter seeing a drop in the average improvement. This could imply that the relationship between income and general knowledge acquisition is not linear and may be influenced by other factors.

These plots emphasize the need for targeted educational support for lower-income groups in reading and math, while the unexpected results in general knowledge require further investigation to understand the factors at play beyond income levels.

#### - Post Hoc Test

Tukey's HSD tests confirm significant academic gains in reading and math for higher income groups compared to lower ones, underscoring income's influence on learning in these domains. However, for general knowledge, income groups show no significant differences, suggesting a more uniform improvement across economic statuses.

## 5. Conclusion

In conclusion, this study highlights that kindergarten students from higher-income families show greater improvements in reading and math, revealing the significant role of socioeconomic status in educational progress. Conversely, general knowledge acquisition appears to be less affected by

income differences. The findings emphasize the need for targeted support in key areas for students from lower-income backgrounds to help bridge the educational achievement gap.