Student: Yixin Chang (1005991651)
INF2178 - Assignment 3

A Study on Kindergarten Academic Progress and Income

1. Introduction

The early years of education provide the crucial foundation for future learning and life success, marking a period of rapid growth and the development of essential skills in children. The dataset "INF2178_A3_data.csv" tracks incomes and Kindergarten students' reading, math, and general knowledge scores for fall 1998 and spring 1999. This report examines this dataset and focuses on the students' progress, attempting to provide insights into their academic development.
Our exploration will address two research questions:

1. **Research Question 1:** How does income group influence the change in reading scores from fall to spring among Kindergarten students, controlling for variations in baseline general knowledge scores?
2. **Research Question 2:** Does the impact of different income group on math score improvement differ among Kindergarten students, controlling for initial general knowledge scores?

2. Data Cleaning

For convenience, we drop unnecessary columns and reduce our dataset to the following columns:
‘fallreadingscore’, ‘fallmathscore’, ‘fallgeneralknowledgescore’, ‘springreadingscore’, ‘springmathscore’, ‘springgeneralknowledgescore’, ‘incomegroup’
The current dataset has 11933 rows and 7 columns. After critically reviewing the data, we found that there are no missing values (NaN) in the current dataset. Then, we can work on our current dataset for further analysis.

3. Exploratory Data Analysis (EDA)

We perform exploratory data analysis to visualize data distribution. The following graph presents a comparison of reading scores for Kindergarten students in fall and spring.

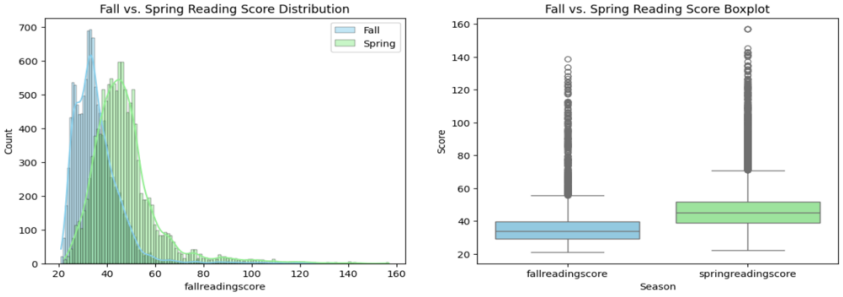

Figure 1: Histogram (left) and Boxplots (right) for reading scores by seasons (fall vs spring)

Based on histograms in Figure 1, both distributions are skewed to the right, indicating that a majority of the students have lower scores with fewer kids scoring extremely high. There is a noticeable shift to the right from fall to spring, suggesting an overall improvement in reading scores. Boxplots shows the median spring reading score is higher than the median fall reading

score, supporting the idea of an overall improvement again. Furthermore, the spring score has a larger IQR, which means spring scores vary more than fall scores.
Figure 2/3 presents the comparisons of math/general knowledge scores in fall and spring.

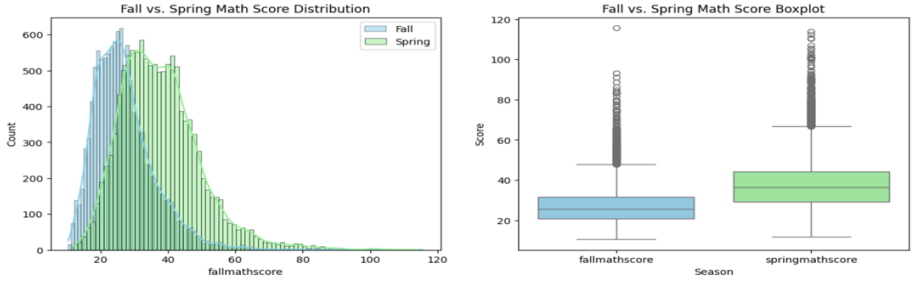

Figure 2: Histogram (left) and Boxplots (right) for math scores by seasons (fall vs spring)

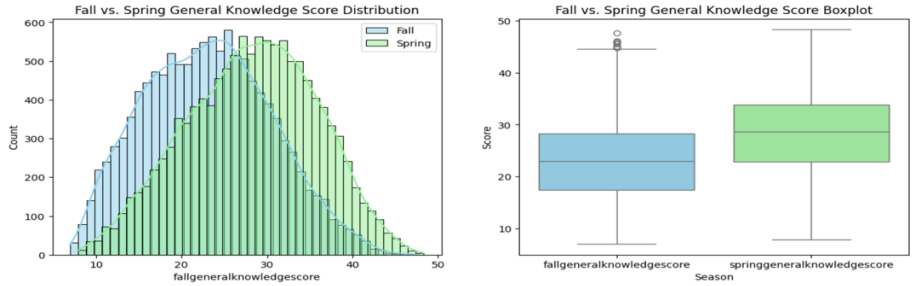

Figure 3: Histogram (left) and Boxplots (right) for general knowledge scores by seasons (fall vs spring)

All histograms display a right shift from fall to spring, and all boxplots show that spring has a higher median score than fall. Thus, based on these distributions, we can conclude that all types of scores improved from fall to spring.

Then, we use a correlation heatmap for examining relationships between multiple variables.

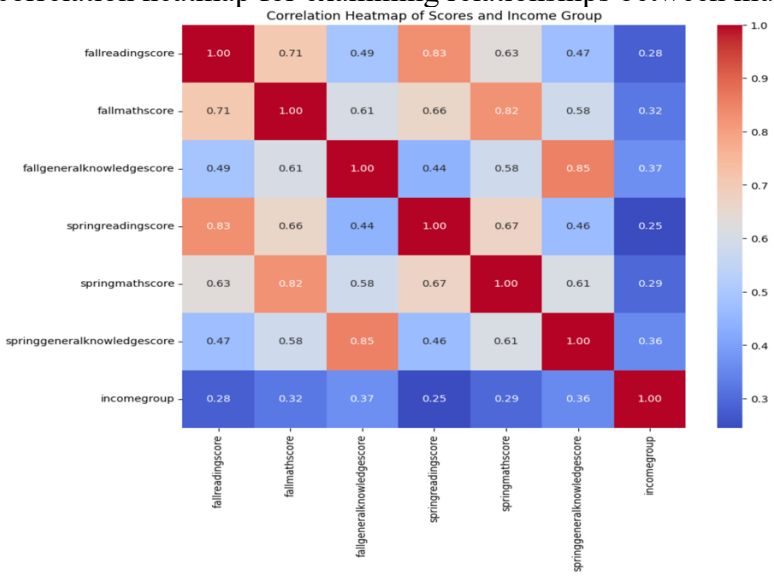

Figure 4: Correlation Heatmap for all variables in current dataset

Considering our research questions, we are interested in the relation between scores and income groups, or between reading/math scores and general knowledge scores. We can see positive correlation around 0.5-0.6 between reading/math scores and general knowledge scores. However, the correlation between scores and income groups are small around 0.3. Since we only contain scores but not the improvement of scores in current dataset, two features representing improvement are created and added to the dataset:

ReadingDiff: this feature refers to the difference between fall and spring reading scores.

$$ReadingDiff = springreadingscore - fallreadingscore$$

MathDiff: this feature refers to the difference between fall and spring math scores.

$$MathDiff = springmathscore - fallmathscore$$

Then, we want to explore what factors may influence the improvement of reading/math scores. An interaction plot is provided in Figure 5 to visualize the relationship between income group and changes in reading/math scores.

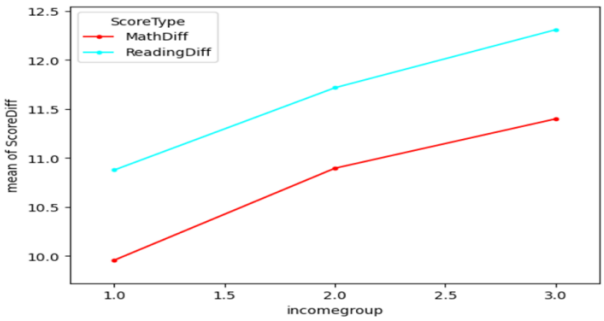

Figure 5: Interaction Plot ('ScoreType' and 'incomegroup')

Both lines are sloping upwards, which indicates a positive association between income level and score differences in both subjects. The plot suggests that students in higher income groups are showing greater improvement in both subjects. Furthermore, lines are separate and do not cross, so the relationship between income group and the change in scores is consistent across the different subjects.

4. One-way ANCOVA (ReadingDiff – for Research Question 1)

In order to explore our first research question, we will use plots and conduct one-way ANCOVA. Firstly, we use Figure 6 to perform some visualization.

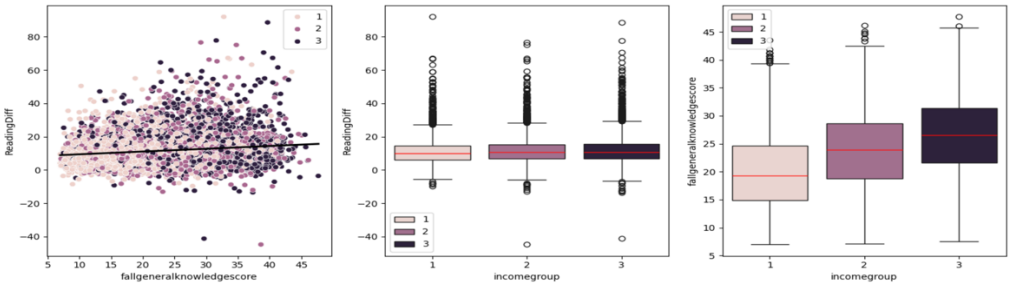

Figure 6: Scatterplots: 'ReadingDiff' vs 'fallgeneralknowledgescore' (left), Boxplots: 'ReadingDiff' across different income group (middle), Boxplots: 'fallgeneralknowledgescore' across different income group (right)

From left to right, the first scatterplot reveals a general trend where higher general knowledge scores correlate with higher ReadingDiff, although the data is quite spread out indicating a lot of variability. This scatterplot and the regression line also help us check the assumption of linearity between the covariates ('fallgeneralknowledgescore') and the dependent variable ('ReadingDiff') for ANCOVA that we will perform. However, the linearity is not apparent. We cannot see a significant difference between medians in the middle graph for ReadingDiff, while higher income group has higher median general knowledge score in the third one.

Figure 7 shows results of ANCOVA. The null hypothesis can be drawn as follow: there is no difference in the mean of reading score difference (ReadingDiff) across different income groups after controlling for the covariate (fall general knowledge score).

	coef	std err	t	P > |t|	Confidence interval

Intercept	7.731	0.242	31.960	< 0.001	(7.257, 8.205)
C (incomegroup) [T.2]	0.217	0.180	1.205	0.228	(-0.136, 0.570)
C (incomegroup) [T.3]	0.404	0.191	2.110	0.035	(0.029, 0.779)
fallgeneralknowledgescore	0.158	0.011	14.836	< 0.001	(0.137, 0.179)

R-squared:	0.023
Adj.R-squared	0.023
F-statistic:	95.49
Prob (F-statistic):	<0.001

Figure 7: Results after using OLS model to perform ANCOVA

Since p-value < 0.001 < 0.05 in the second table, we reject the null hypothesis and conclude the income group and/or general knowledge scores play a significant role in explaining the variability in reading score differences among Kindergarten students. The results in the first table indicate that, while controlling for general knowledge, there is no significant difference in reading score changes between the baseline income group (1) and income group 2 (p-value = 0.228 > 0.05). However, income group 3 differs significantly from the baseline (income group 1), implying an income-related discrepancy in reading score improvement (p-value = 0.035 < 0.05). The fall general knowledge score is a significant covariate with p-value < 0.001 < 0.05, and the positive coefficient (0.158) indicates that higher baseline knowledge is associated with greater reading score improvements. The model's low R-squared (0.023) means the model explains a small percentage of the variance in the change in reading scores. Hence, other variables not included in the model might also affect reading score changes.

Checking assumptions is an important step when performing one-way ANCOVA. The assumption check about linearity is mentioned above in Figure 6, and the linearity is not obvious. To check the assumption of homogeneity of regression slopes, we add an interaction term between the covariate and the independent variable to the model and then testing the significance of this interaction. Both p-values in Figure 8 are less than 0.05, so the assumption of homogeneity of regression slopes is violated.

	coef	**std err**	**t**	**P > |t|**	**Confidence interval**
C (incomegroup) [T.2] : fallgeneralknowledgescore	-0.081	0.026	-3.155	0.002	(-0.132, -0.031)
C (incomegroup) [T.3] : fallgeneralknowledgescore	-0.121	0.026	-4.693	< 0.001	(-0.171, -0.070)

Figure 8: Interaction term results

Then, QQplot in Figure 9 and Shapiro-Wilk test with the results in Figure 10 is used for normality assumption.

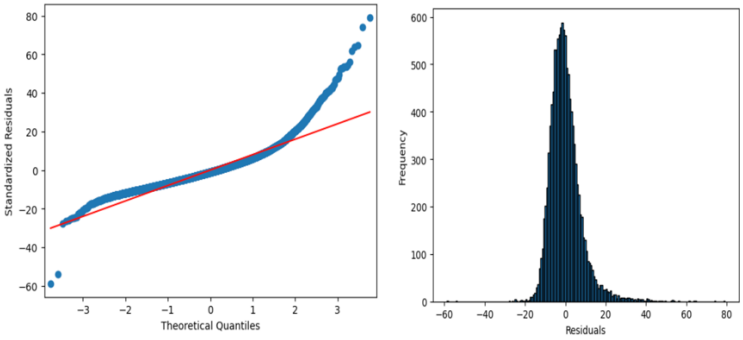

Figure 9: QQ-plot and Histogram from standardized residuals

Test Statistics: W	p-value
0.900	< 0.001

Figure 10: Results of Shapiro-Wilk test

Since the p-value < 0.001 < 0.05, we reject the null hypothesis and conclude that the residuals are not normally distributed. Because the data is not drawn from normal distribution, Levene's test is used to check the assumption, the Homogeneity of variances.

	Parameter	Value
0	Test statistics (W)	19.728
1	Degrees of freedom (Df)	2.000
2	P value	< 0.001

Figure 11: Results of Levene's test

As the p value (< 0.05) is significant in Figure 11, we reject null hypothesis and conclude the assumption of homogeneity of variances has been violated.

5. One-way ANCOVA (MathDiff – for Research Question 2)

We also use plots and conduct one-way ANCOVA to explore Research Question 2.

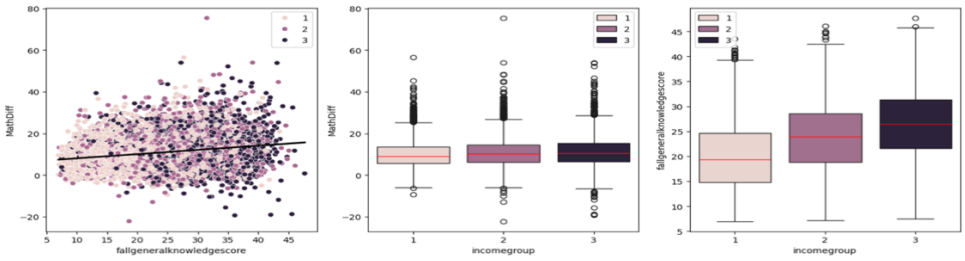

Figure 12: Scatterplots: 'MathDiff' vs 'fallgeneralknowledgescore' (left), Boxplots: 'MathDiff' across different income group (middle), Boxplots: 'fallgeneralknowledgescore' across different income group (right)

The first scatterplot shows a general trend where higher general knowledge scores correlate with higher MathDiff, and it help us check the assumption of linearity. We cannot see a significant difference between medians in the middle graph for MathDiff.

ANCOVA results are shown in Figure 13. The null hypothesis is: there is no difference in the mean of math score difference (MathDiff) across different income groups after controlling for the covariate (fall general knowledge score).

	coef	**std err**	**t**	**P > |t|**	**Confidence interval**
Intercept	5.9826	0.203	29.542	< 0.001	(5.586, 6.380)
C (incomegroup) [T.2]	0.1523	0.151	1.011	0.312	(-0.143, 0.448)
C (incomegroup) [T.3]	0.1442	0.160	0.900	0.368	(-0.170, 0.458)
fallgeneralknowledgescore	0.1993	0.009	22.385	< 0.001	(0.182, 0.217)

R-squared:	0.048
Adj.R-squared	0.048
F-statistic:	200
Prob (F-statistic):	<0.001
Log-Likelihood:	-39610
AIC:	79230
BIC:	79260

Figure 13: Results after using OLS model to perform ANCOVA

Since p-value < 0.001 < 0.05 in the second table, we reject the null hypothesis and conclude the income group and/or general knowledge scores play a significant role in explaining the variability in math score differences among Kindergarten students. The results in the first table indicate that, while controlling for general knowledge, there is no significant difference in math score changes between the income group (1) and income group 2/3 (p-value = 0.312/0.368 > 0.05. The model's low R-squared (0.048) means the model explains a small percentage of the variance in the change in math scores.

We checking assumptions using the same step in section 4.

- Assumption of linearity between 'MathDiff' and fall general knowledge score is violated, which is not apparent in Figure 12.
- Homogeneity of regression slopes is also violated with all p-values are less than 0.05 in Figure 14.
- Normally distributed (Figure 15/16) is violated (the p-value < 0.001 < 0.05 in Shapiro-Wilk test).
- Homogeneity of variances (Figure 17) is violated (the p-value < 0.001 < 0.05 in Levene's test).

	coef	**std err**	**t**	**P > |t|**	**Confidence interval**
C (incomegroup) [T.2] : fallgeneralknowledgescore	-0.068	0.022	-3.172	0.002	(-0.110, -0.026)
C (incomegroup) [T.3] : fallgeneralknowledgescore	-0.095	0.022	-4.386	< 0.001	(-0.137, -0.052)

Figure 14: Interaction term results

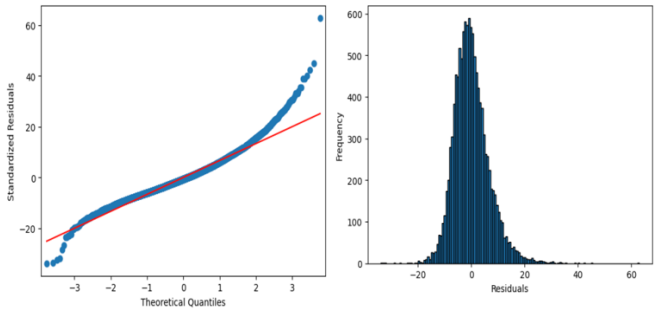

Figure 15: QQ-plot and Histogram from standardized residuals

Test Statistics: W	p-value
0.966	< 0.001

Figure 16: Results of Shapiro-Wilk test

	Parameter	Value
0	Test statistics (W)	22.215
1	Degrees of freedom (Df)	2.000
2	P value	< 0.001

Figure 17: Results of Levene's test

6. Conclusion

In conclusion, the impact of different income group on reading/math score improvement differ among Kindergarten students, controlling for initial general knowledge scores. However, in both one-way ANCOVA, many assumptions are violated, which may lead to inaccurate results. More models and further analysis are needed in this case.