Student ID: 1001420223
Student Name: Aaron Chen
Instructor: Shion Guha
Course: INF2178 Experimental Design for Data Science
Assignment: Technical Assignment 3

Exploring Kindergarten grades over time between different income group

1. Introduction

Early childhood development has always been the focus of many parents, the dataset under comprises data from an early child longitudinal study from Fall 1998 to Spring 1999, evaluating Kindergarten students over the span of several months.

This report entails a comprehensive data analysis of the above dataset with the objective of under-covering whether income groups affect the grade of child over time.

Research Questions:

Research Question 1: How does income group influence the change in students' math performance from Fall 1998 to Spring 1999, after controlling for their general knowledge as covariate?
Research Question 2: How does income group influence the change in students' reading performance from Fall 1998 to Spring 1999, after controlling for their general knowledge as covariate?

2. Data Cleaning and Data Wrangling

The dataset comprises 9 columns and 11933 rows. The dataset is pretty clean since initial inspection reveals no null data/missing data in the dataset.

Interested Columns:
fallreadingscore: reading mark of children during the fall 1998 term
fallmathscore: math mark of children during the fall 1998 term
fallgeneralknowledgescore: general knowledge mark of children during the fall 1998 term
springreadingscore: reading mark of children during the spring 1999 term
springmathscore: math mark of children during the spring 1999 term
incomegroup: representing the income group family the children is within

3. Exploratory Data Analysis (EDA)

Categorical Variable:

There is only one categorical variable which is the income group that has three types 1,2,3. The type of the income group is calculated by continuous variable totalhouseholdincome, by examining the data we see that the income group is represented in ascending order where 1 is the lowest income group.

Continuous Variables:

Table 1:

	fallreadingscore	fallmathscore	fallgeneralknowledgescore	springreadingscore	springmathscore	springgeneralknowledgescore
mean	35.95	27.13	23.07	47.51	37.80	28.23
std	10.47	9.12	7.40	14.33	12.03	7.58
min	21.01	10.51	6.98	22.35	11.9	7.86
25%	29.34	20.68	17.38	38.95	29.27	22.80
50%	34.06	25.68	22.95	45.32	36.41	28.58
75%	39.89	31.59	28.30	51.77	44.22	33.78
max	138.51	115.65	47.69	156.85	113.8	48.34

Referencing Table 1, just by looking at the stats for these data, we can see that on average reading score > math score > general knowledge score, this might be because they don't share the same scale when marked so we have to be careful if we try to compare across subjects. Another interesting point is that overall every subject sees an increase in scores across the span of time from fall 1998 to spring 1999. This is related to our research question, we now know that the score definitely increased but we want to delve deeper into investigating whether or not the increase is caused by the family income group.

Figure 1: income group count

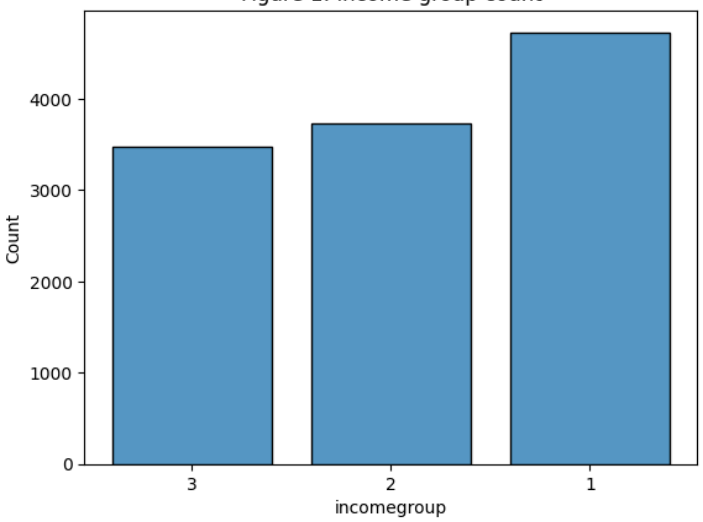


Figure 2: Math Scores by Income Group and Season

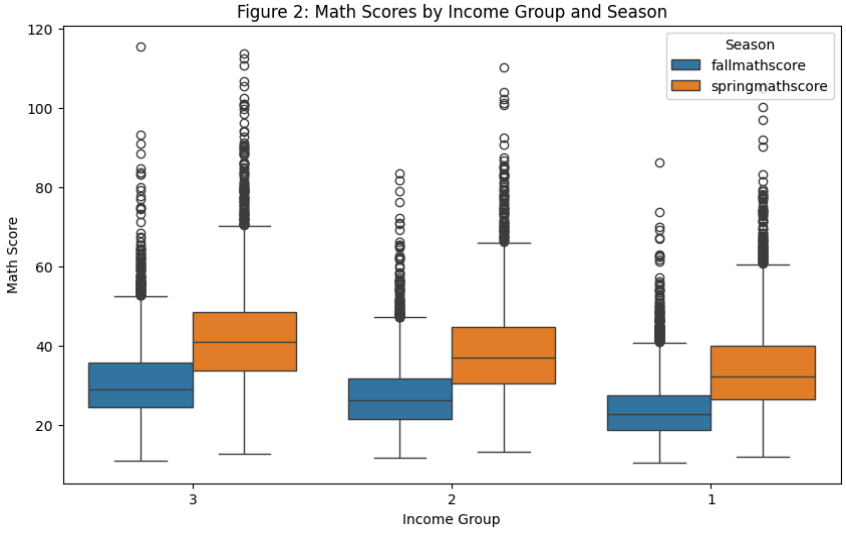


Figure 3: General Knowledge Scores by Income Group and Season

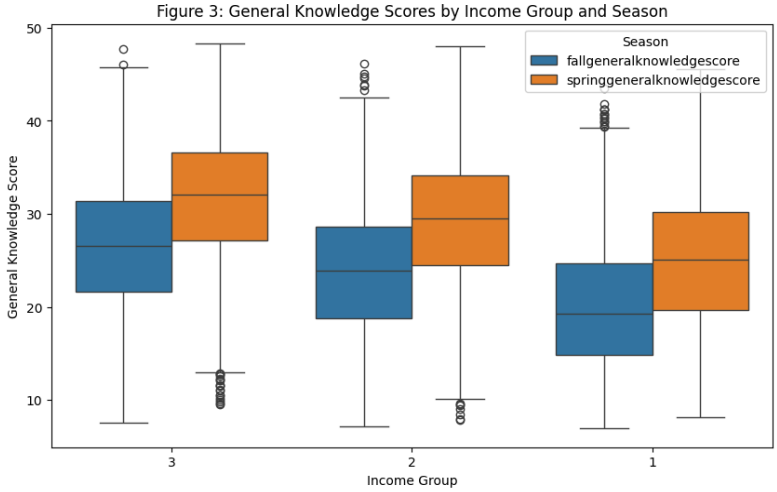
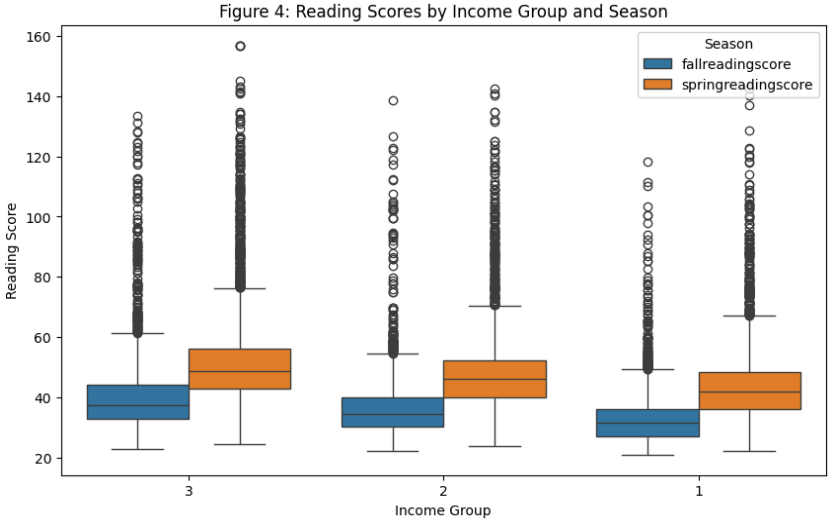


Figure 4: Reading Scores by Income Group and Season



Looking at figure 1 we see that the data is spread pretty equally across the different income groups, we can say that our dataset is balanced. Looking at figure 2, we see that the box shows a decrementing step from high to low (3 to 1) for both the fall score and the spring score; this suggests that high income family children do have an edge in testing math scores compared to low income families. Looking at figure 3 and 4 we can see the same trend happening for both general knowledge and reading scores. These box plots also support the finding we mentioned earlier on improvement in scores over time, by examining all the box plots it seems that the increase is the same for all income families. So by looking at the box plot it suggests that a higher income family child has a higher starting point in marks but the improvement overtime seems to be not affected by the income family.

Figure 5: Interaction Plot Math Score

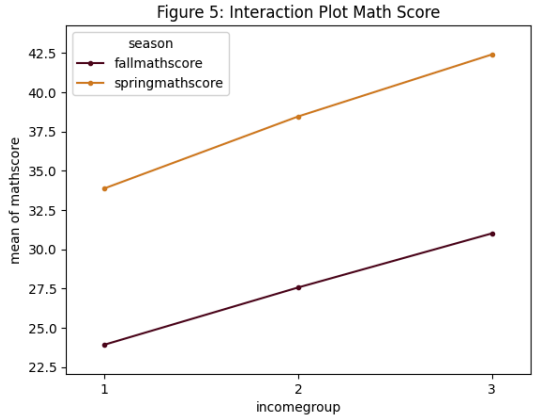


Figure 6: Interaction Plot General Knowledge Score

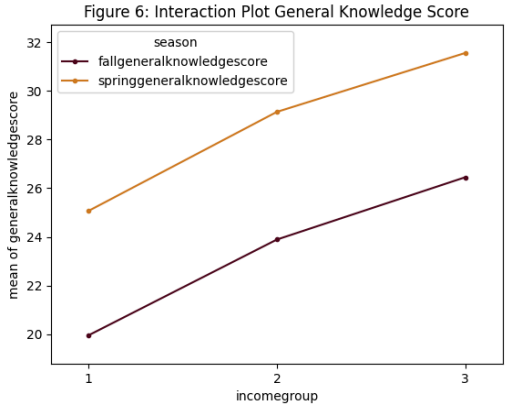
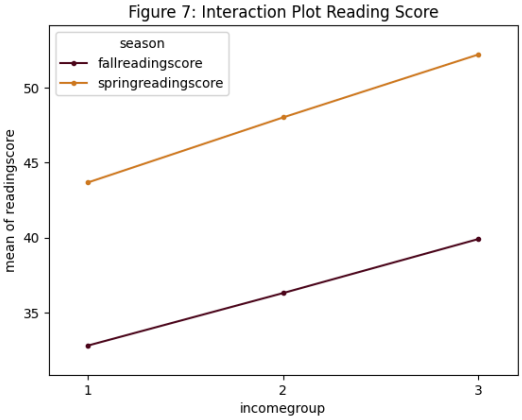


Figure 7: Interaction Plot Reading Score



From the above interaction plots we can see that the line runs parallel for all subjects, this suggests that there are no interaction effects. In other words, the effect of income groups on test scores is consistent across different income groups. But we have to be cautious of our results as there might be other factors that might influence the relationship between income group and test score.

Figure 8: Math Score change between fall 1998 - spring 1999 by income group

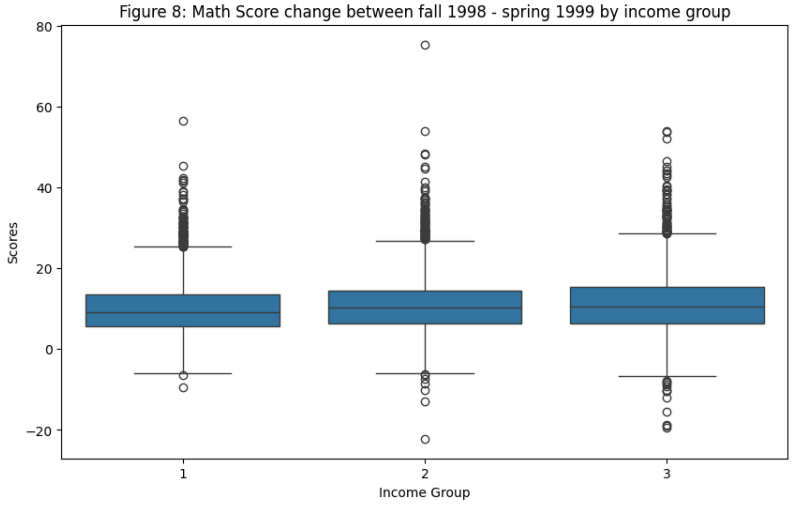


Figure 9: Read Score change between fall 1998 - spring 1999 by income group

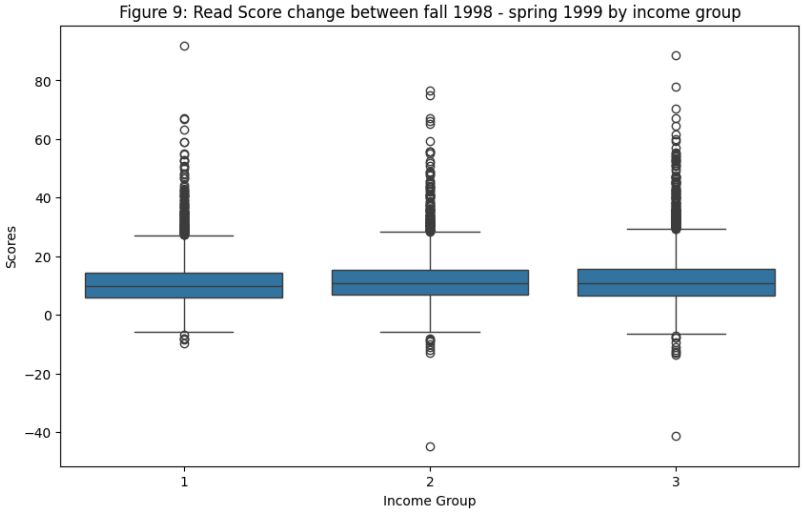


Figure 8 and 9 shows the box plot of the change in math and reading scores between spring and fall, as we can see above this furthermore suggests that income groups have no impact on scores over time.

4. ANCOVA

Below we will be conducting ANCOVA tests on both subjects to see if income groups have an effect on students' test marks over time controlling general knowledge.

Table 2: Change in Math score ANCOVA (mathdiffscore ~ incomegroup + fallgeneralknowledgescore)

Source	SS	DF	F	P-unc	np2
incomegroup	55.88	2	0.62	0.54	0.0001
fallgeneralknowledgescore	22425.93	1	501.08	<0.001	0.04
residual	551499.44	11929	Nan	Nan	Nan

From table 2 we can see that the ANCOVA result indicates that the income group does not have a significant effect on the change in math score ($p>0.05$). This result suggests that our null hypothesis is true and is also following what we suspected in our EDA.

Assumption Checks:

Residual Normality:

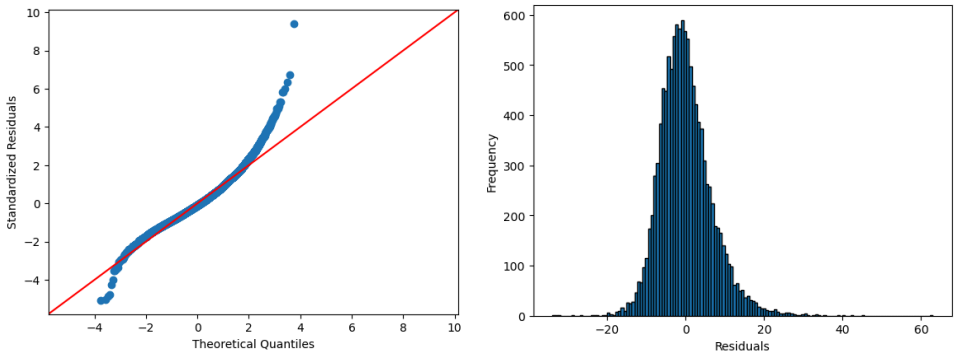


Table 3: Math score difference Shapiro Wilk test result

Statistic	P Value
0.966	<0.001

The Shapiro Wilk test reveals a significant departure from normality ($p<0.01$).

Table 4: Math score difference Homogeneity of Variances:
Since the sample is not normally distributed, we will use the Levene's test to assess homogeneity of variances. The result is below

	Parameter	Value
0	Test Statistics(W)	22.22
1	Degrees of freedom	2.0
2	P value	<0.001

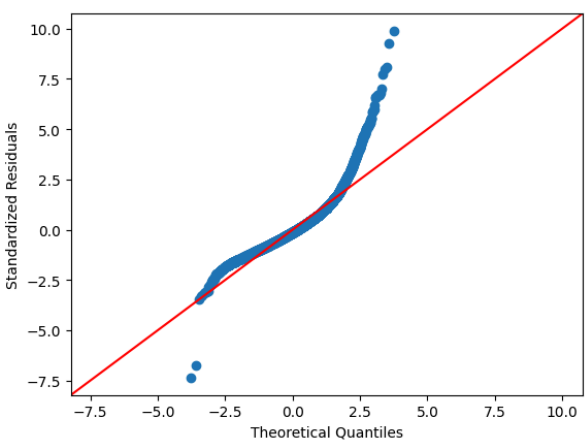
The test yielded a p-value less than 0.05, indicating violation of the assumption of homogeneity of variances. Given the violation of both assumptions, caution is warranted in interpreting the ANCOVA results.

Table 5: Change in Reading score ANCOVA (readdiffscore ~ incomegroup + fallgeneralknowledgescore)

Source	SS	DF	F	P-unc	np2
incomegroup	287.49	2	2.25	0.11	0.0004
fallgeneralknowledgescore	14054.12	1	220.11	<0.001	0.018
residual	761671.04	11929	Nan	Nan	Nan

From table 5 we can see that the ANCOVA result indicates that the income group does not have a significant effect on the change in reading score ($p>0.05$). This result suggests that our null hypothesis is true and is also following what we suspected in our EDA.

Assumption Checks:
Residual Normality:



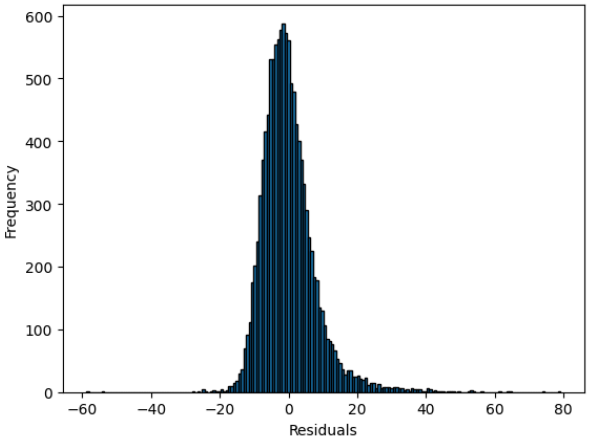


Table 6: Read score difference Shapiro Wilk test result

Statistic	P Value
0.899	<0.001

The Shapiro Wilk test reveals a significant departure from normality (p<0.01).

Table 7: Read score difference Homogeneity of Variances:
Since the sample is not normally distributed, we will use the Levene's test to assess homogeneity of variances. The result is below

	Parameter	Value
0	Test Statistics(W)	19.73
1	Degrees of freedom	2.0
2	P value	<0.001

The test yielded a p-value less than 0.05, indicating violation of the assumption of homogeneity of variances. Given the violation of both assumptions, caution is warranted in interpreting the ANCOVA results.

5. Conclusion

In this report, we analyzed the early child longitudinal study dataset to explore whether there is difference in score change for different income groups controlling general knowledge. We addressed two research question:

1. Math score difference: A one-way ANCOVA revealed no significant effect of income group on math score controlling general knowledge (p>0.05)
2. Read score difference: A one-way ANCOVA revealed no significant effect of income group on read score controlling general knowledge (p>0.05)

However, violated assumptions regarding normality and homogeneity of variances urge caution in interpreting the results for both ANCOVA tests.