Sheng Zhang
eily.zhang@mail.utoronto.ca

An Investigation of an Early Child Longitudinal Study

1. Introduction

An early child longitudinal study that cross-evaluated the skills of reading, maths, and general knowledge for Kindergarten students from different families with different total household income was taken for fall 1998 and spring 1999. Based on the investigation of the dataset "INF2178_A3_data.xlsx," the report employs exploratory data analysis (EDA) and one-way ANCOVA to address the following research questions:

1. Does student score differ by the semester (fall vs spring) after controlling for the total household income?
2. Does student score differ by the subject (reading vs maths vs general knowledge) after controlling for the total household income?

With these research questions in mind, we can start our data analysis.

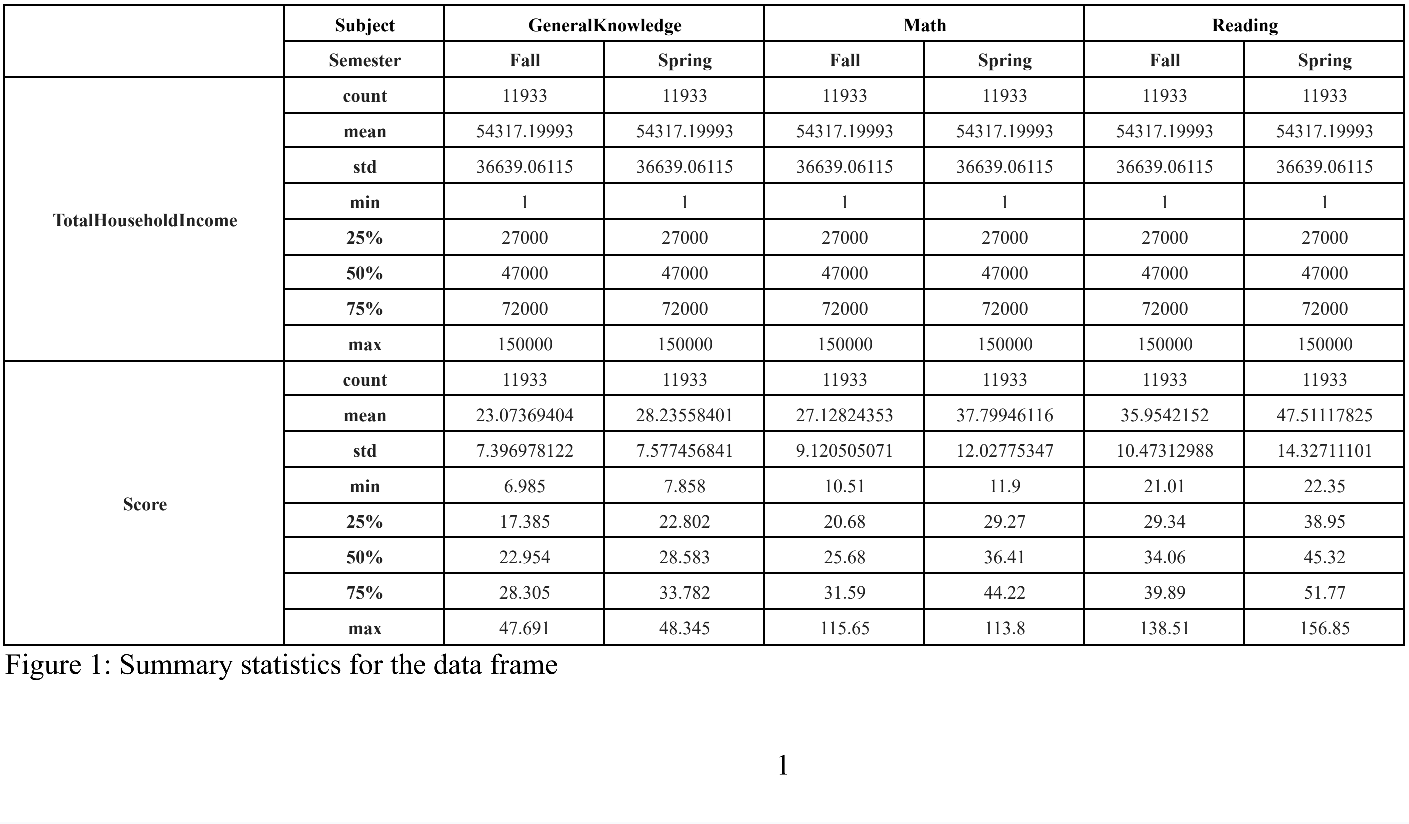
2. Data Cleaning and Data Wrangling

The initial dataset has 9 columns and 11933 rows. By checking the data and the data dictionary, we want to discard the variable “income in thousands” as it is basically the same as total household income. Additionally, since the Score is the dependent continuous variable that we are observing, we want to keep the scores in only one column. Then for the independent categorical variables IncomeGroup, Subject, Semester, we want to keep them in separate columns for the convenience of further data analysis. So we separate every single column that contained the scores originally from the template “[Semester][Subject][Score]” into three columns Semester, Subject, and Score respectively. For example, for the original column “fallreadingscore”, all the scores will be in the column Score, and the values in the columns Semester and Subject will be Fall and Reading respectively. Notice that originally, each row in the data frame represents information about a student, but after our conversion, each row presents a score, and we will notify what kind of scores it is, and the total household income of the student’s family that score belongs to. After conversion, we have the new data frame with the following key columns to start the data analysis:

- TotalHouseholdIncome: the total household income of the student’s family that the owner of that score corresponds to
- IncomeGroup: the group of the income after the classification of the total household income
- Subject: the subjects of the evaluation (reading, maths, general knowledge)
- Score: the score of the subject
- Semester: the semester of the school year (fall, spring)

3. Exploratory Data Analysis(EDA)

First we construct a summary table to check the statistics of the dataset [Figure 1].



	Subject	**GeneralKnowledge**		**Math**		**Reading**	
	Semester	**Fall**	**Spring**	**Fall**	**Spring**	**Fall**	**Spring**
TotalHouseholdIncome	**count**	11933	11933	11933	11933	11933	11933
	mean	54317.19993	54317.19993	54317.19993	54317.19993	54317.19993	54317.19993
	std	36639.06115	36639.06115	36639.06115	36639.06115	36639.06115	36639.06115
	min	1	1	1	1	1	1
	25%	27000	27000	27000	27000	27000	27000
	50%	47000	47000	47000	47000	47000	47000
	75%	72000	72000	72000	72000	72000	72000
	max	150000	150000	150000	150000	150000	150000
Score	**count**	11933	11933	11933	11933	11933	11933
	mean	23.07369404	28.23558401	27.12824353	37.79946116	35.9542152	47.51117825
	std	7.396978122	7.577456841	9.120505071	12.02775347	10.47312988	14.32711101
	min	6.985	7.858	10.51	11.9	21.01	22.35
	25%	17.385	22.802	20.68	29.27	29.34	38.95
	50%	22.954	28.583	25.68	36.41	34.06	45.32
	75%	28.305	33.782	31.59	44.22	39.89	51.77
	max	47.691	48.345	115.65	113.8	138.51	156.85

Figure 1: Summary statistics for the data frame

Please note that the columns of TotalHouseholdIncome are the same for each subject and each semester because no matter how to distribute the types of scores, we are still analysing the same group of students.

Before we consider the covariate, we can use boxplots to analyse the possible relationships with the dependent variable(score) and the independent variable(subject, semesters, and IncomeGroup). We have the boxplot of scores for different subjects [Figure 2], scores for different semesters [Figure 3], scores for different income groups [Figure 4].

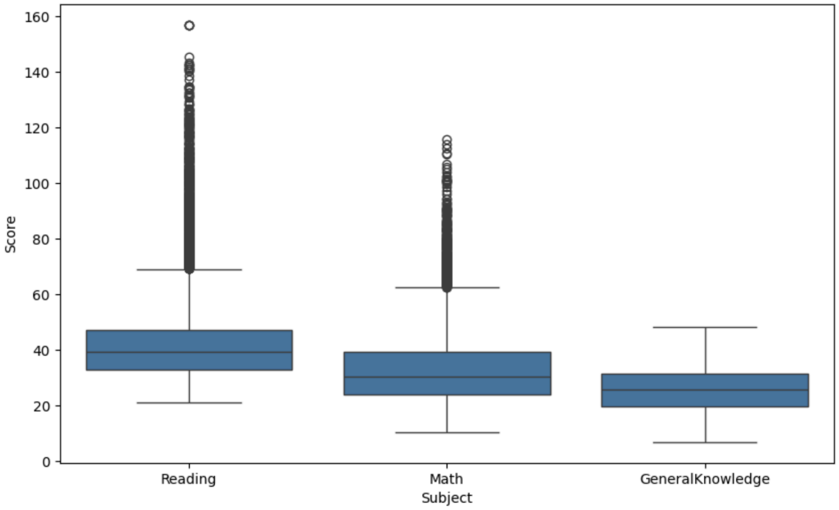
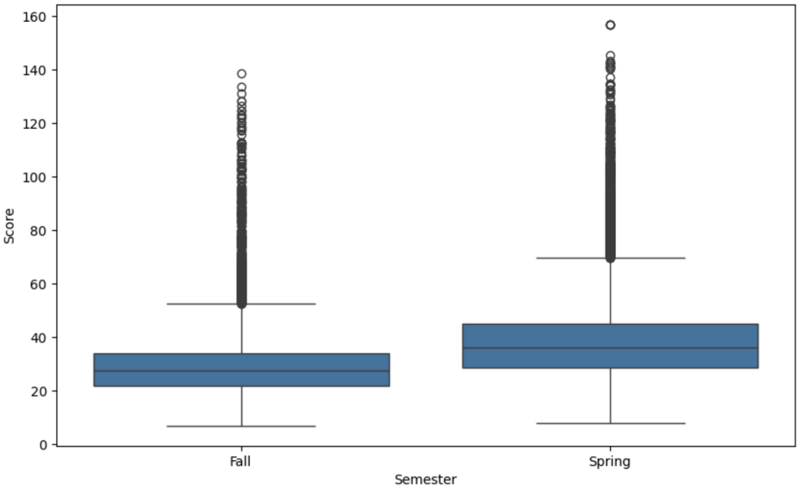
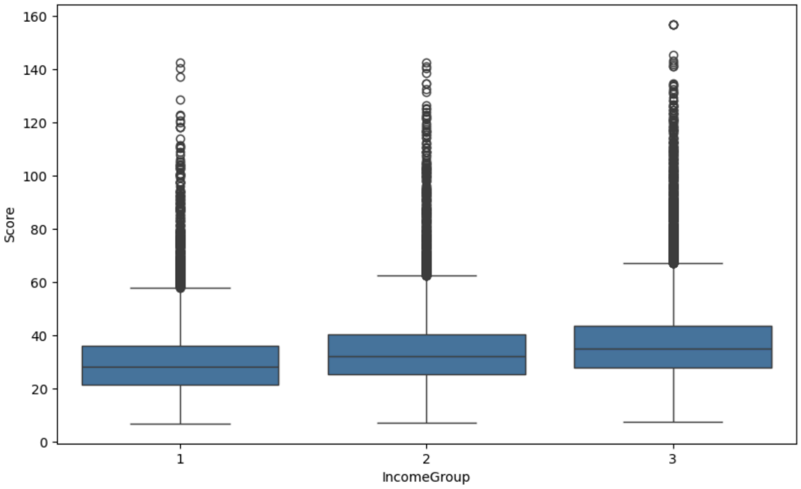


Figure 2[Left]: Box plot for scores across subjects

Figure 3 [Below Left]: Box plot for scores across semesters

Figure 4 [Below Right]: Box plot for scores across income groups





At the first glance of these boxplots above, we observe that students tend to score higher marks in reading and lower marks in general knowledge, also, they tend to score higher marks in spring semester compared to fall semester. Additionally, we can suspect that the scores might have a positive relationship with the total household income.

Now, we can check the cross-variable boxplots for us to understand the data in terms of considering multiple variables at the same time. We have the box plot for scores across different subjects and semesters [Figure 5], the box plot for scores across different subjects and income groups [Figure 6], and the box plot for scores across different semesters and income groups [Figure 7].

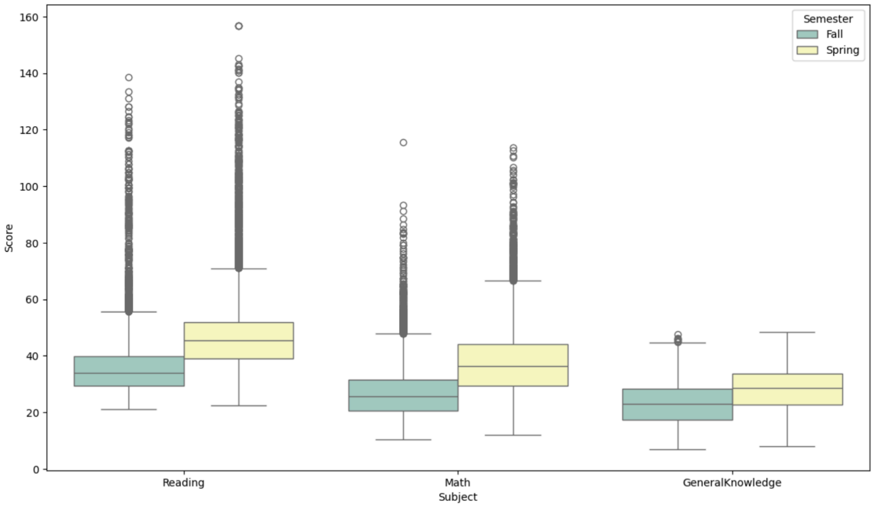


Figure 5: Box plot for scores across different subjects and semesters

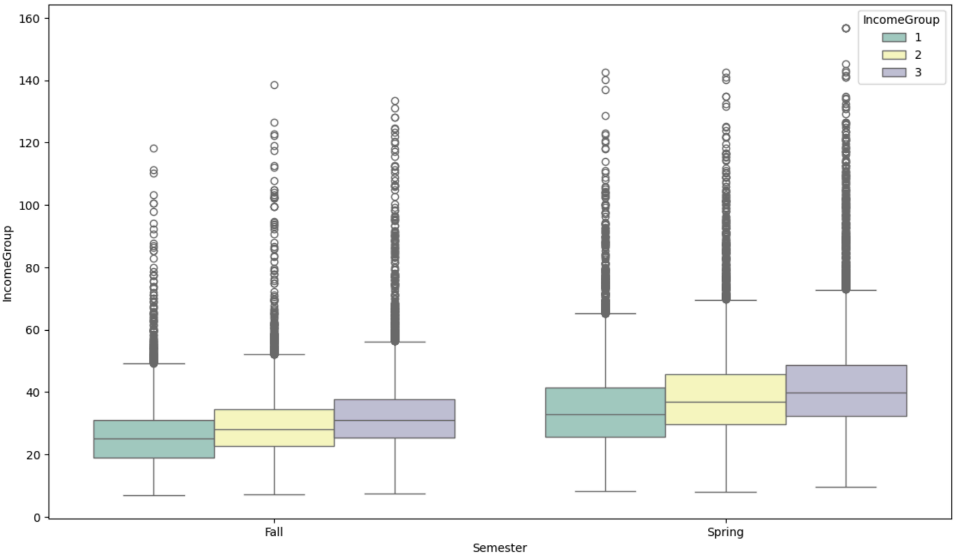


Figure 6: Box plot for scores across different subjects and income groups

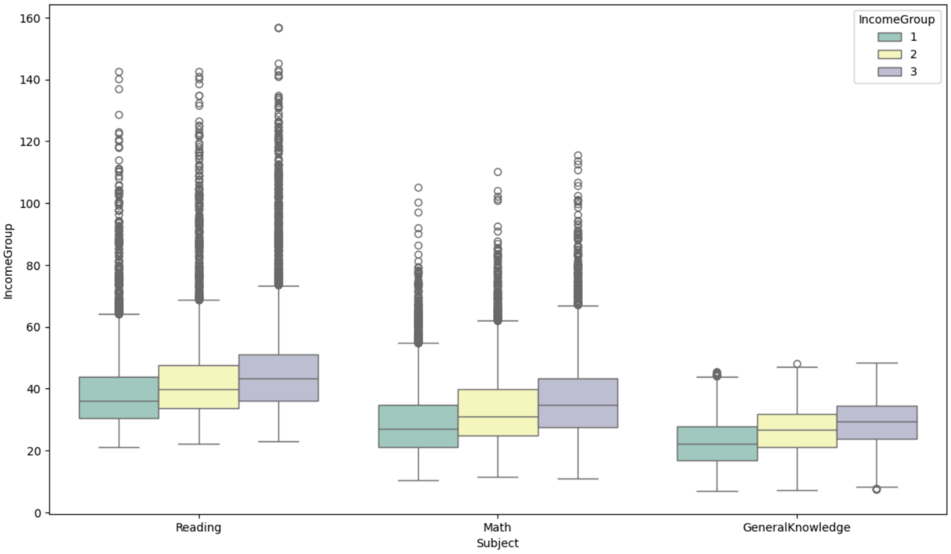


Figure 7: Box plot for scores across different semesters and income groups

The above multi-variable boxplots coincide with our assumptions from the previous boxplots, that is, students tend to score higher marks in reading subject, spring semester, and higher income group.

Now, we're prepared for deeper analysis. With each variable having at least three groups, and we want to consider the total household income as a potential covariate, we'll employ one-way ANCOVA to investigate how different independent variables(semesters and subjects) influence the continuous variable(scores) by controlling the covariate(total household income). All data are randomly drawn and independent, meeting the assumption of independence. Our significance level, or alpha-level, for all analyses will be set at 0.05. Detailed ANCOVA analyses for each research question and their assumptions will be covered in subsequent sections.

4. Scores across Semesters after Controlling the Total Household Income

Research Question 1: Does student score differ by the semester (fall vs spring) after controlling for the total household income?

The null hypothesis is that after controlling the total household income, there is no significant difference in the scores of students in fall semester versus spring semester, and the chosen alpha-level is 0.05. Thus, we'll present the table of the one-way ANCOVA in Figure 8.

Dep. Variable:	Score		**R-squared:**	0.178		**Df Model:**	3
Model:	OLS		**Adj. R-squared:**	0.178		**Covariance Type:**	nonrobust
Method:	Least Squares		**F-statistic:**	5177			
Date:	Wed, 13 Mar 2024		**Prob (F-statistic):**	<0.05			
Time:	17:12:53		**Log-Likelihood:**	-2.80E+05			
No. Observations:	71598		**AIC:**	5.59E+05			
Df Residuals:	71594		**BIC:**	5.59E+05			

	coef		std err	t	P>|t|	[0.025	0.975]
Intercept	24.2863		0.113	214.021	<0.05	24.064	24.509
Semester[T.Spring]	8.4438		0.16	52.616	<0.05	8.129	8.758
TotalHouseholdIncome	8.16E-05		1.73E-06	47.116	<0.05	7.82E-05	8.50E-05
Semester[T.Spring]:TotalHouseholdIncome	1.26E-05		2.45E-06	5.158	<0.05	7.83E-06	1.74E-05
Omnibus:	27260.147		Durbin-Watson:	1.206			
Prob(Omnibus):	<0.05		Jarque-Bera (JB):	202655.834			
Skew:	1.647		Prob(JB):	<0.05			
Kurtosis:	10.555		Cond. No.	3.07E+05			

Figure 8: One-way ANCOVA table for scores across semesters when controlling total household income

The ANCOVA p-value is below our chosen significance level of 0.05, leading us to reject the null hypothesis. This means that there is a significant relationship between the scores and the semesters(fall and spring), with total household income also affecting scores. The interaction terms suggest that the effect of total household income on scores varies by semesters. Next, we can perform a post hoc test as shown in Figure 9 below.

group1	group2	meandiff	p-adj	lower	upper	reject
Fall	Spring	9.13	<0.05	8.9479	9.3122	TRUE

Figure 9: Post-hoc table for scores across semesters when controlling total household income

From the table, we conclude that there is a statistically significant difference between the "Fall" and "Spring" groups.

Next, we'll examine the assumptions for one-way ANCOVA. The independence assumption was met during data collection. However, we need to check the assumptions for normality and homogeneity of variance:

1. Assumption 1: Normality of Residuals
   We perform the Shapiro-Wilk test and obtain a p-value of <0.05, indicating there is no deviation from normality. Thus, the assumption of normality is met
2. Assumption 2: Homogeneity of Variance
   We perform Levene's test and obtain a p-value of <0.05, suggesting equal variance. Therefore, the assumption of homogeneity of variance is also met.
3. Assumption 3: Linearity
   We can check the scatter plot for the residuals in the Figure 10 below.

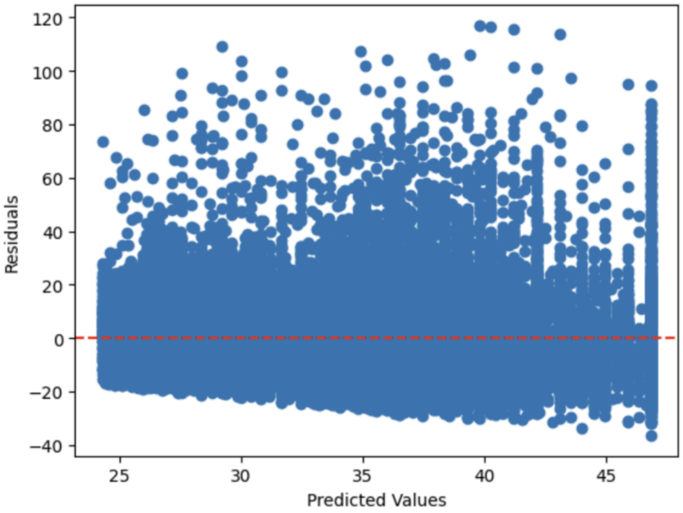


Figure 10: Scatter Plot of the residuals for scores across semesters when controlling total household income

   We observe that the residuals are randomly distributed around the horizontal line at zero without any clear pattern, which suggests the linearity. Thus, the assumption of linearity is also met.
4. Assumption 4: Homogeneity of Regression Slopes
   This can be verified by looking at the interaction plot as shown in Figure 11 below.

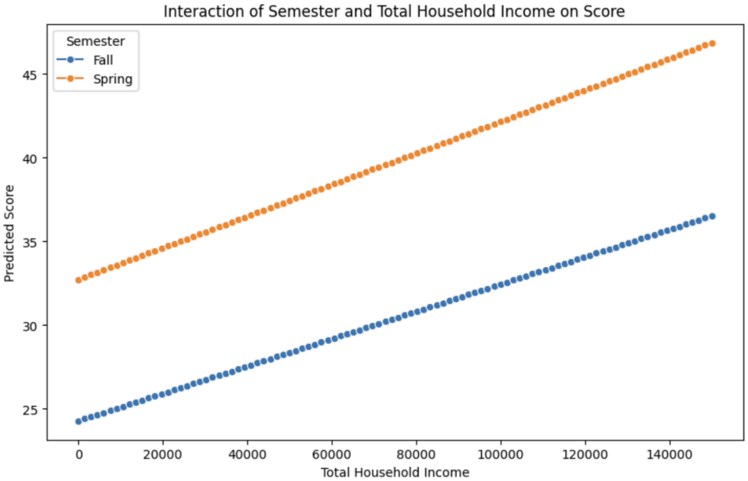


Figure 11: Interaction Plot of semesters and total household income in scores.

We observe that the two lines are nearly parallel to each other, which indicates no interaction and homogeneity of regression slopes. Thus the assumption for homogeneity of regression slopes is also met.

As all assumptions of one-way ANCOVA are met, which indicates that our conclusion is valid. Thus, there is a significant relationship between the scores and the semesters(fall and spring), with total household income also affecting scores.

5. Scores across Subjects after Controlling the Total Household Income

Research Question 2: Does student score differ by the subject (reading vs maths vs general knowledge) after controlling for the total household income?

The null hypothesis is that after controlling the total household income, there is no significant difference in the scores of students in reading versus maths versus general knowledge, and the chosen alpha-level is 0.05. Thus, we'll present the table of the one-way ANCOVA in Figure 12.

Dep. Variable:	Score	**R-squared:**	0.307		**Df Model:**	5
Model:	OLS	**Adj. R-squared:**	0.307		**Covariance Type:**	nonrobust
Method:	Least Squares	**F-statistic:**	6350			
Date:	Wed, 13 Mar 2024	**Prob (F-statistic):**	<0.05			
Time:	17:46:49	**Log-Likelihood:**	-2.73E+05			
No. Observations:	71598	**AIC:**	5.47E+05			
Df Residuals:	71592	**BIC:**	5.47E+05			
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	21.5327	0.128	168.74	<0.05	21.283	21.783
Subject[T.Math]	5.8299	0.18	32.305	<0.05	5.476	6.184
Subject[T.Reading]	15.0965	0.18	83.653	<0.05	14.743	15.45
TotalHouseholdIncome	7.59E-05	1.95E-06	38.963	<0.05	7.21E-05	7.97E-05
Subject[T.Math]:TotalHouseholdIncome	1.80E-05	2.75E-06	6.545	<0.05	1.26E-05	2.34E-05
Subject[T.Reading]:TotalHouseholdIncome	1.81E-05	2.75E-06	6.561	<0.05	1.27E-05	2.35E-05
Omnibus:	29253.73	**Durbin-Watson:**	1.43			
Prob(Omnibus):	<0.05	**Jarque-Bera (JB):**	232449.262			

Figure 12: One-way ANCOVA table for scores across subjects when controlling total household income

The ANCOVA p-value is below our chosen significance level of 0.05, leading us to reject the null hypothesis. This means that there is a significant relationship between the scores and the subjects(reading, maths, and general knowledge), with total household income also affecting scores. The interaction terms suggest that the effect of total household income on scores varies by subjects. Next, we can perform a post hoc test as shown in Figure 13 below.

group1	group2	meandiff	p-adj	lower	upper	reject
GeneralKnowledge	Math	6.8092	<0.05	6.5627	7.0557	TRUE
GeneralKnowledge	Reading	16.0781	<0.05	15.8316	16.3246	TRUE
Math	Reading	9.2688	<0.05	9.0223	9.5153	TRUE

Figure 13: Post-hoc table for scores across subjects when controlling total household income

Thus, we conclude that there is a statistically significant difference between any two groups among the three subjects.

Next, we'll examine the assumptions for one-way ANCOVA. The independence assumption was met during data collection. However, we need to check the assumptions for normality and homogeneity of variance:

5. Assumption 1: Normality of Residuals
   We perform the Shapiro-Wilk test and obtain a p-value of <0.05, indicating there is no deviation from normality. Thus, the assumption of normality is met
6. Assumption 2: Homogeneity of Variance
   We perform Levene's test and obtain a p-value of <0.05, suggesting equal variance. Therefore, the assumption of homogeneity of variance is also met.
7. Assumption 3: Linearity
   We can check the scatter plot for the residuals in Figure 14 below.

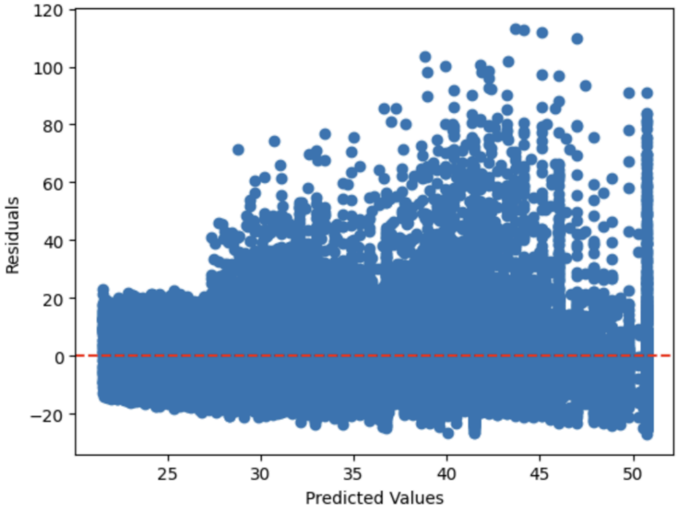


Figure 14: Scatter Plot of the residuals for scores across subjects when controlling total household income

We observe that the residuals are randomly distributed around the horizontal line at zero without any clear pattern, which suggests the linearity. Thus, the assumption of linearity is also met.

8. Assumption 4: Homogeneity of Regression Slopes
   This can be verified by looking at the interaction plot as shown in Figure 15 below.

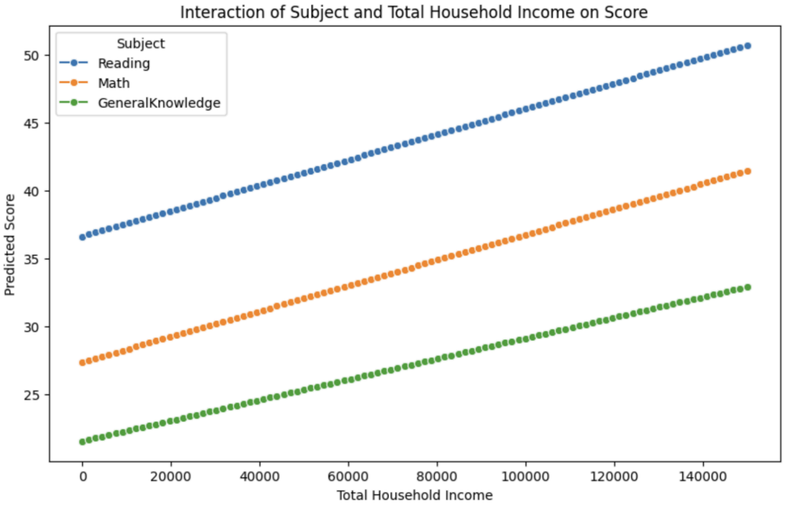


Figure 15: Interaction Plot of subjects and total household income in scores.

We observe that the two lines are nearly parallel to each other, which indicates no interaction and homogeneity of regression slopes. Thus the assumption for homogeneity of regression slopes is also met.

As all assumptions of one-way ANCOVA are met, which indicates that our conclusion is valid. Thus, there is a significant relationship between the scores and the subjects(reading, maths, and general knowledge), and total household income affects scores.

6. Conclusion

After the one two-way ANOVA analyses, we observe significant differences between scores and subjects&semesters after controlling the total household income, which means that the semesters and subjects of studying, and financial background can all influence the academic performance for Kindergarten students.