INF2178 A3 - Exploring Knowledge Scores of Children

in Different Times with Respect to Income Groups

Yuyang Liu, 1005965617

Introduction & Research Question

For the purpose of this project, we will be examining the dataset called “INF2178_A3_data.csv”. The dataset has 9 columns with 11933 observations excluding the header row. It collects information on children’s reading, math, and general knowledge scores across two different periods, fall 1998 and spring 1999. The dataset also collects information on respective household income and divide them into three income groups for further analysis.

This project is intended to explore some trends and conduct analyses on children’s knowledge score in different times with respect to income groups.

The research questions we are interested in are the followings:

1. *Does the change in children's reading scores from fall 1998 to spring 1999 differ across income groups, after controlling for their baseline knowledge?*
2. *Does the change in children's math scores from fall 1998 to spring 1999 differ across income groups, after controlling for their baseline knowledge?*

Data Cleaning and Preparation

To begin, we first import the dataset and relevant data analytical libraries into Python notebook.

Then we check to see if there are null values or empty cells in the dataset that need to be addressed. We find that all values are valid and can be used for analysis.

We are particularly interested in the knowledge score change of the children, so we compute the difference between fall 1998 and spring 1999 for the reading and the math scores.

Not all columns are useful, so we only select the columns that are relevant to our research questions, and analyses are conducted based on the selected columns only.

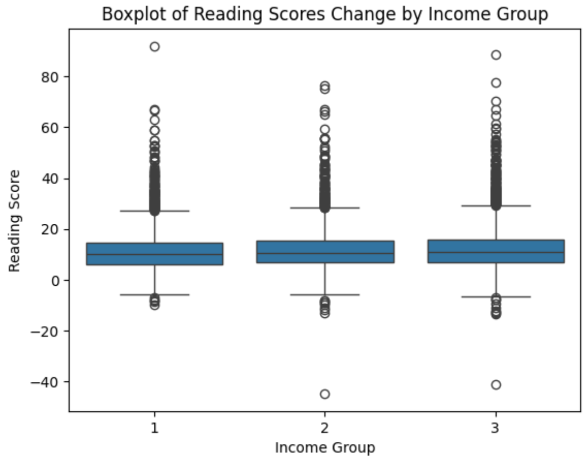
Exploratory Data Analysis (EDA)

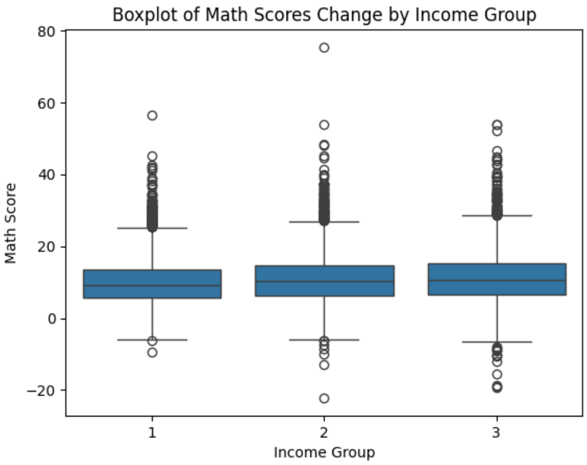
First of all, a summary statistics table is constructed to provide an overview of the selected columns we are interested in.

	fallgeneralknowledgescore	**fallreadingscore**	**fallmathscore**	**springreadingscore**	**springmathscore**	**readingscorediff**	**mathscorediff**
count	11933	11933	11933	11933	11933	11933	11933
mean	23.07	35.95	27.13	47.51	37.80	11.56	10.67
std	7.40	10.47	9.12	14.33	12.03	8.09	6.86
min	6.99	21.01	10.51	22.35	11.90	-44.76	-22.16
25%	17.39	29.34	20.68	38.95	29.27	6.47	6.01
50%	22.95	34.06	25.68	45.32	36.41	10.40	9.86
75%	28.31	39.89	31.59	51.77	44.22	15.15	14.33
max	47.69	138.51	115.65	156.85	113.80	91.94	75.35

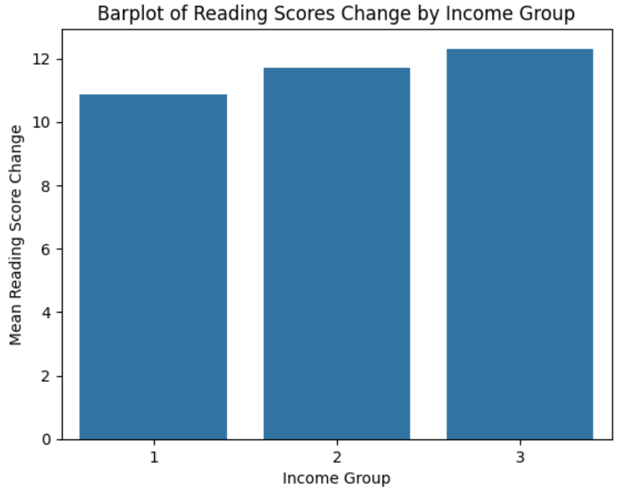
Given the summary statistics table output. Something interesting can be seen. We can clearly see that mean reading and math scores increase from fall to spring. However, there are exceptions. From the minimum value we can see that there are children actually experiencing a decrease in the knowledge scores. Therefore, we cannot be certain that the increase in the knowledge scores is a sure thing as time progresses. On the other hand, we are also interested in the potential household income effect on the knowledge score change. Hence, further analysis is needed.

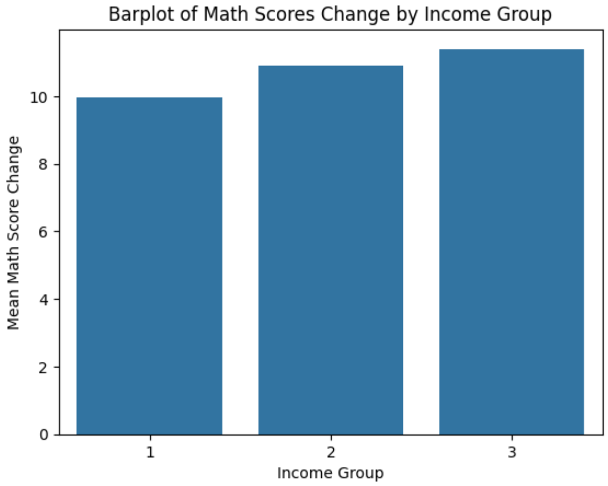
Then, we build two boxplots to see the distribution of data points. One for reading scores improvement by income group, and one for math scores improvement by income group.



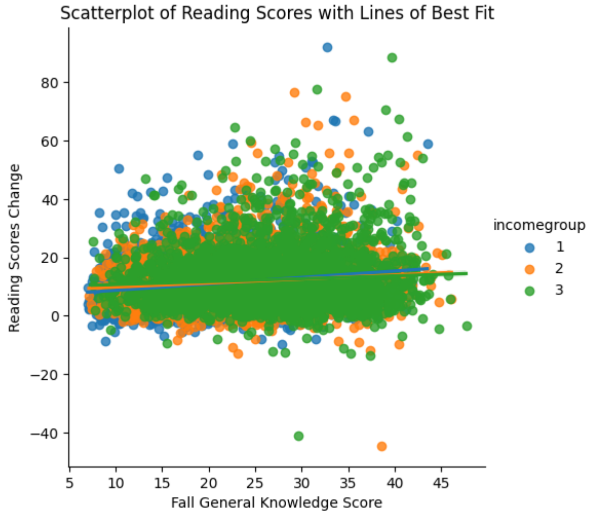


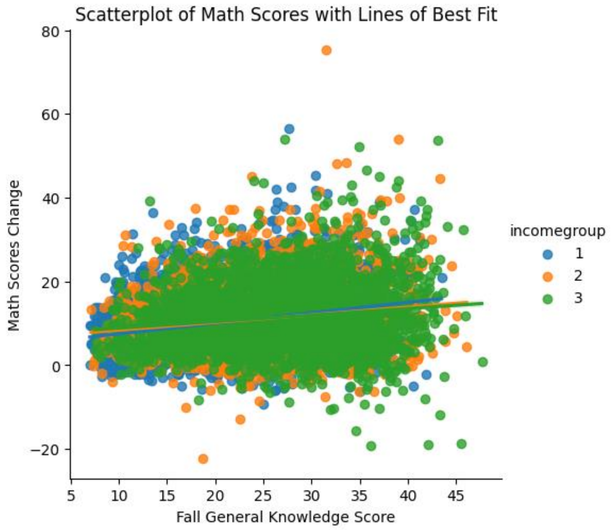
From the boxplots, there seems to be a small increase going from income group 1 to 3. But the trend is not very clear. So we use bar charts to capture the small differences that may be due to the income effect.





From the bar charts, we can see there is increase in both the reading and the math scores from fall to spring. But the effect is small, and we need to conduct statistical test to see whether the result is significant or not.





We also build scatterplots and line of best fit trying to support our findings. But there is no distinct differences between the three income groups.

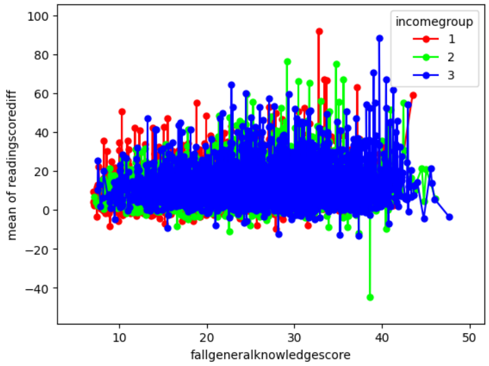
ANCOVA Test 1 – Reading Score Change by Income Group

Recall our first research question: Does the change in children's reading scores from fall 1998 to spring 1999 differ across income groups, after controlling for their baseline knowledge?

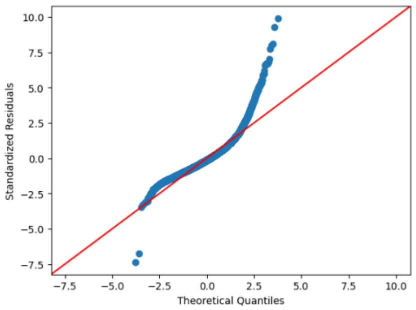
ANCOVA test is conducted to assess the significance of the result and the result is shown below:

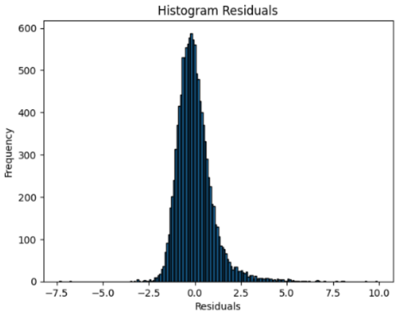
	Sum_sq	df	F	PR(>F)
C(incomegroup)	287.48	2.0	2.25	0.105
fallgeneralknowledgescore	14054.12	1.0	220.11	2.35e-49
Residual	761671.0	11929.0	NaN	NaN

We obtain a F statistic of 2.25 with p-value at approximately 0.105, the p-value is greater than any statistically significant level. Therefore, we cannot say that income differences have an impact on the reading score improvement.



An interaction plot is constructed, but there no distinct differences between the income groups with respect to reading scores change. Then, we are also interested in testing the assumptions of this ANCOVA test.





Shapiro Wilk Test w statistic	0.8996
p-value	< 0.001

Levene’s Test Parameter	Value
Test statistics (W)	19.728
Degrees of freedom (Df)	2
p-value	< 0.001

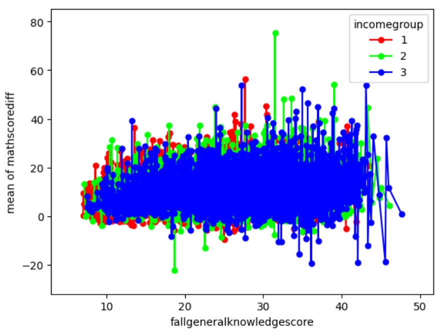
From the assumption check, we first see that residual normality assumption is violated. A test p-value < 0.001 and we can observe a clear trend in the residual plot. Then we use Levene’s test for the sample is not normally distributed, and we get a test statistic of 19.73 with a p-value < 0.001. This means that the homogeneity of variance assumption is also violated.

ANCOVA Test 2 – Math Score Change by Income Group

Recall our second research question: Does the change in children's math scores from fall 1998 to spring 1999 differ across income groups, after controlling for their baseline knowledge?

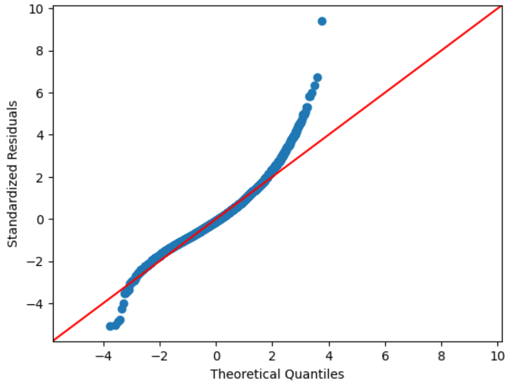
	Sum_sq	df	F	PR(>F)
C(incomegroup)	55.88	2.0	0.62	0.536
fallgeneralknowledgescore	22425.93	1.0	501.08	9.425e-109
Residual	533880.50	11929.0	NaN	NaN

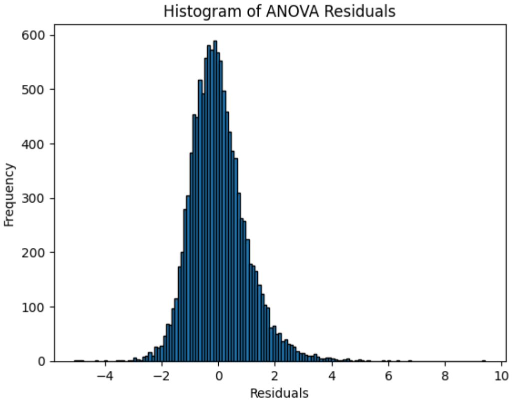
We obtain a F statistic of 0.62 with p-value greater than 0.5, the p-value is greater than any statistically significant level. Therefore, we cannot say that income differences have an impact on the reading score improvement.



An interaction plot is constructed, but there no distinct differences between the income groups with respect to reading scores change.

Then we check for test assumptions to understand the model fit and interpretation of the results.





Shapiro Wilk Test w statistic	0.966
p-value	< 0.001

Levene's Test Parameter	Value
Test statistics (W)	22.21
Degrees of freedom (Df)	2
p-value	< 0.001

Similar to our first ANCOVA test, we can observe a clear trend of the residuals. After conducting the relevant assumption tests and obtain the respective p-value, we can also see that both assumptions are violated.

Conclusion

In conclusion, we have conducted two ANCOVA tests in this study to determine whether there is an income effect on the children's reading and math score improvement. From the graphs analysis and statistic test results, we cannot say that income plays an important role in the score change.

However, both ANCOVA tests conducted in this experiment do not meet the assumptions of residuals normality and variances homogeneity. As a result, all the outcomes should be interpreted with caution, and we have to rethink and further explore the appropriateness of the model we utilized.