UNIVERSITY OF TORONTO
FACULTY OF INFORMATION

INF 2178 - Experimental Design For Data Science

Technical Assignment 3

By

Saad Umar

Course Instructor: Shion Guha

Date: March 23, 2024

Examining the scores of Kindergarten children in Reading, Maths, & General Knowledge from Fall’1998 & Spring’1999

1. Introduction

This report presents an analysis of a subset of data derived from an early childhood longitudinal study, focusing on the academic performance of kindergarten students over the course of several months. The dataset comprises reading, math, and general knowledge scores obtained during fall 1998 and spring 1999 assessments. Additionally, income category, the sole categorical variable, is provided, enabling a comprehensive examination of socioeconomic influences on academic achievement. Through this analysis, we aim to gain insights into the factors influencing kindergarten students' performance and to inform strategies for fostering academic success in early childhood education settings.
This report offers a comprehensive data exploratory analysis with a goal to uncover the underlying trends. We delve into the dataset, named: ‘*INF2178_A3_data.csv*’ to discover the multi-dimensional aspects of socioeconomic status with scores of kindergarten students, shedding light to address three fundamental research questions, serving as guiding principles in discovering the patterns. By addressing these questions, we aim to contribute insights into dynamics of children's scores in reading, math, and general knowledge at the kindergarten level and provide a deeper understanding that can inform more effective interventions.

2. Data Cleaning and Data Wrangling

The raw dataset has a total of 9 columns with 10933 entities (rows). After initial review of the dataset, I was confident that not much data cleaning was necessary for the scope of my analysis. Below I have outlined my observations and the new feature which I added to the dataset:

a. Observations & Considerations

1. My analysis is quantitative and I’ve only used specific columns from the dataset to perform the analysis. However, I did not drop the other columns from the dataset as I feel those columns are needed for a more detailed analysis, which I plan to perform in the future. Below I have provided a short description of each columns that I have used in my analysis so far:
    - fallreadingscore: Scores of kindergarten students in reading for fall 1998
    - fallmathscore: Scores of kindergarten students in math for fall 1998
    - fallgeneralknowledgescore: Scores of kindergarten students in general knowledge for fall 1998
    - springreadingscore: Scores of kindergarten students in reading for spring 1999
    - springmathscore: Scores of kindergarten students in math for spring 1999
    - springgeneralknowledgescore: Scores of kindergarten students in general knowledge for spring 1999

b. Feature Engineering:

I created (3) new feature to add to the dataset to aid in my analysis. The feature is as follows:

1. **changeIn_Reading**: calculates the difference between a student's grade in fall 1998 and spring 1999 in reading: changeIn_Reading = spring reading score - fall reading score.
2. **changeIn_Maths**: calculates the difference between a student's grade in fall 1998 and spring 1999 in math: changeIn_Maths = spring math score - fall math score.
3. **changeIn_generalKnowledge**: calculates the difference between a student's grade in fall 1998 and spring 1999 in general knowledge (gk): changeIn_generalKnowledge = spring gk score - fall gk score.

3. Exploratory Data Analysis

I performed an extensive EDA to leverage insight that could potentially help me derive insightful research questions. I started by summarizing quantitative data and then using a bunch of bar plots and boxplots to see how different features varied and how the distributions differed across different levels.

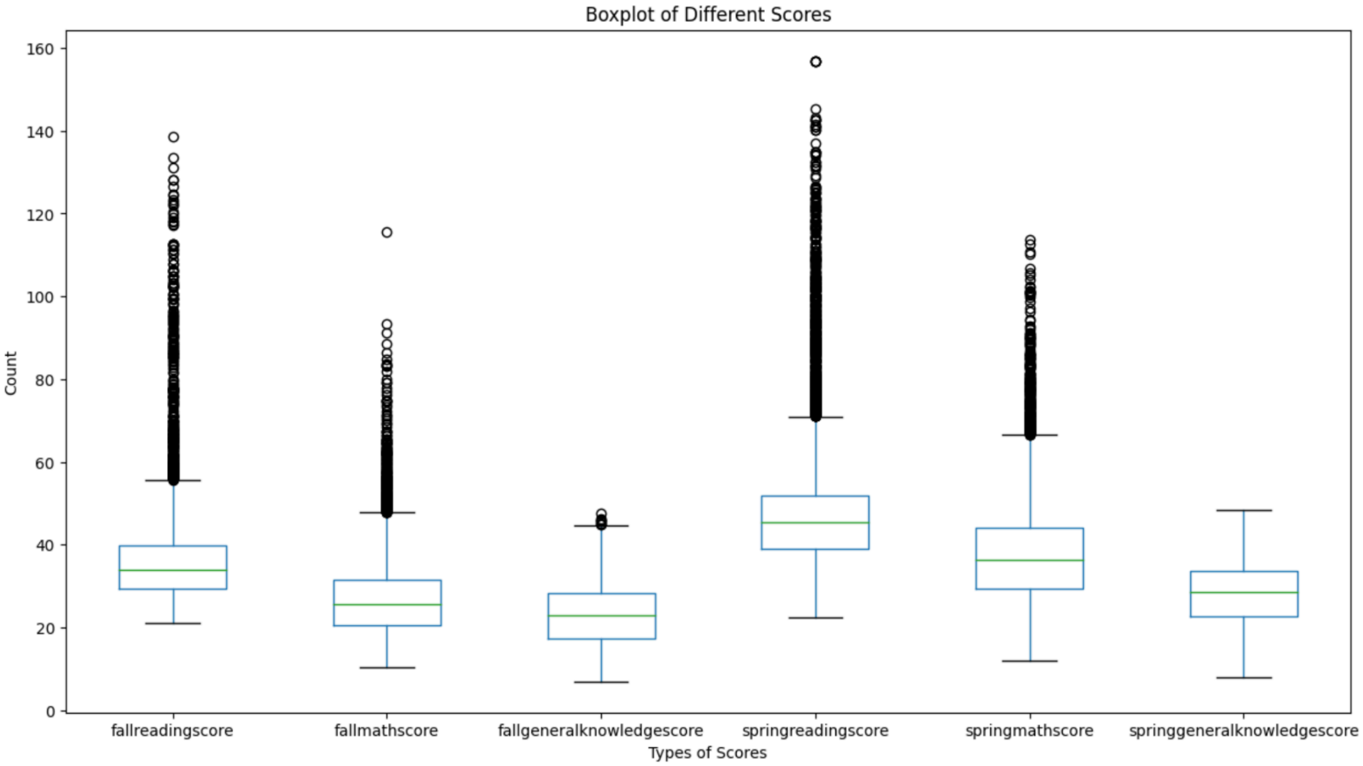

Fig 1 - Boxplot of the Different type of Scores (due to space constraint the titles are unclear but are in the order as in section 2a)

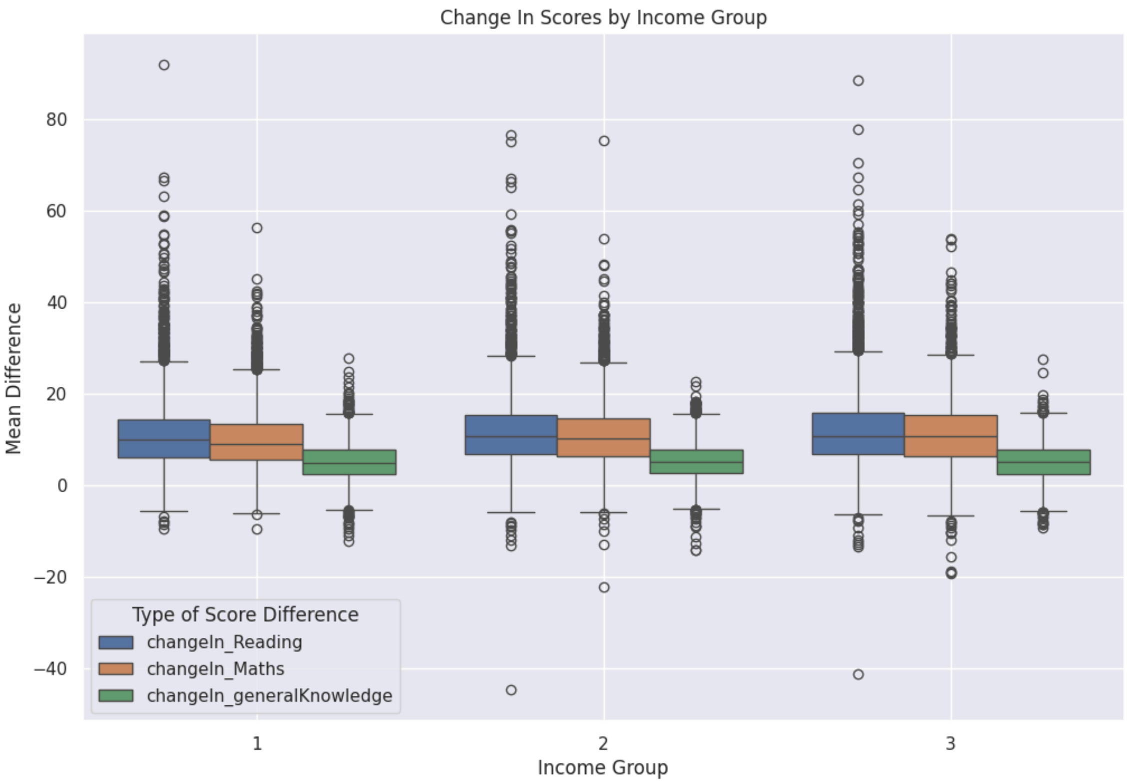

Fig 2 - Boxplot of Change in Scores by Income Groups

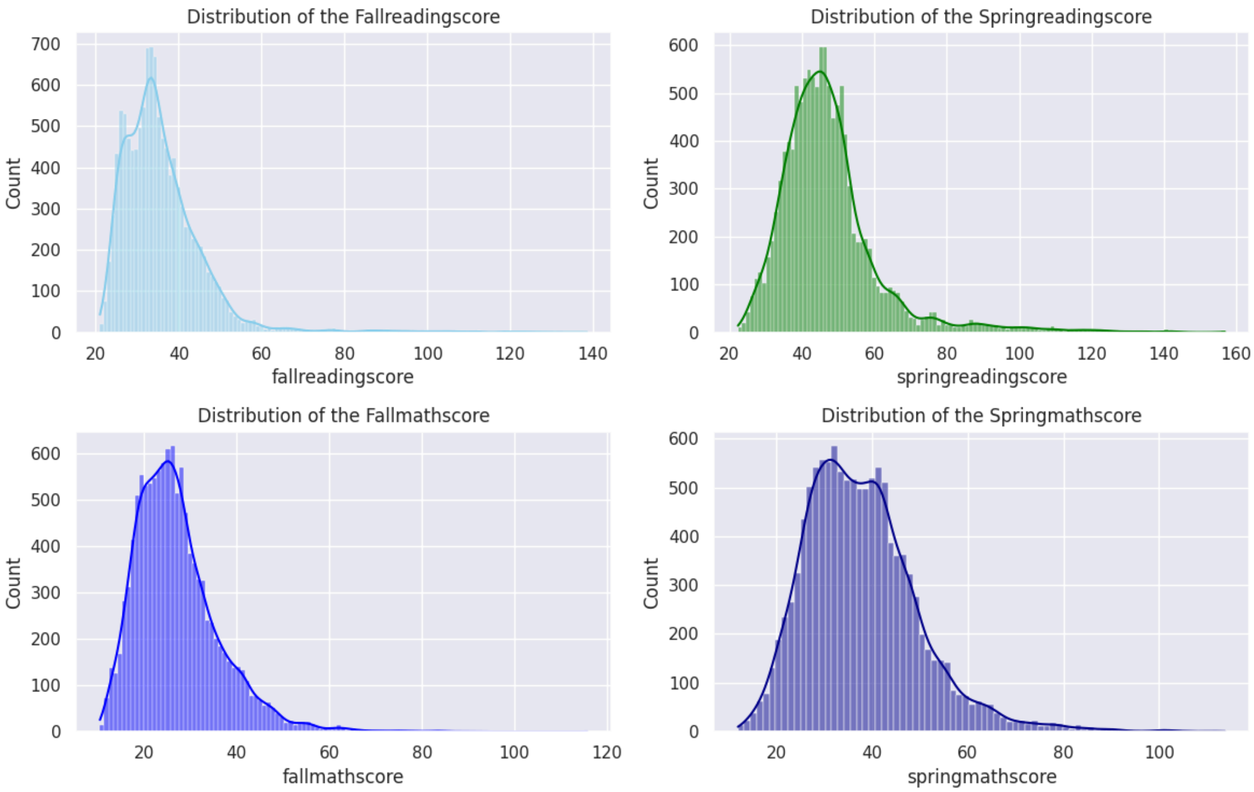

Fig 3 - Distribution of Different Types of Scores

4. Δ General Knowledge based on Income group with Δ Maths as a covariate

Research Question 1: ANCOVA1: Does the change in general knowledge vary at the end of the two semesters across different income groups, while controlling for the influence of the change in math scores?

I used an Ancova test to analyze the relationship between the dependent variable '*changeIn_generalKnowledge*' and two independent variables '*incomegroup*' and '*changeIn_Maths'.* Using table 1, the p-value for Income Group suggests that there may not be a statistically significant difference in the dependent variable across different income groups. The null hypothesis associated with this p-value test is that there is no significant difference in the dependent variable across different levels of the income group. A p-value of 0.179 suggests that there is no strong evidence to reject the null hypothesis at a conventional significance level (such as α = 0.05). In other words, the difference in the dependent variable across income groups may not be statistically significant.

	p-value
Income Group	0.179
changeIn_Maths	<0.001

Table 1: Ancova 1 Table

The p-value for changeIn_Maths suggests that there is a statistically significant relationship between the dependent variable and the change in math scores (changeIn_Maths). The null hypothesis associated with this p-value test is that there is no significant linear relationship between the dependent variable and changeIn_Maths. A p-value of <0.001 indicates strong evidence against the null hypothesis. It suggests that there is a statistically significant relationship between the dependent variable and changeIn_Maths.

The **second step** was to check for the assumptions: I used the **Shapiro Wilk** test to check for the normality. In table 2 below, we can see that the p-value is less than the significance level of 0.05, hence we can reject the null hypothesis of the data being normal. These results were also backed by the 'w test statistic' and the QQ plot in Figure 4, as we can see that the data does not seem to follow a normal distribution. The other assumption test I used was the **Levene's test** to check for variance homogeneity. From table 3 below, we can see that the p-value is greater than the significance level of 0.05, hence we do not have sufficient evidence to reject the null hypothesis that, "the variance across the three groups are equal" and thus, we can conclude that the variances across the different levels of income group are equal. The **third step** was to perform the post-hoc tests. These tests are used to explore specific pairwise comparisons between groups when the overall omnibus test indicates a significant difference. From Table 4 below, we can observe that p-value is greater than 0.05 (significance level) for all the different pairs, hence we **don't** have sufficient evidence to reject the null hypothesis that there is no significant difference in the different income groups.

Test statistic(w)	p-value
0.997	< 0.001

Table 2 - Shapiro Wilk Test Result

Parameter	Value
Test Statistic	0.182
P Value	0.834

Table 3 - Levene's Test Result

Conclusion: In conclusion, the analysis indicates no significant difference in the dependent variable across various income groups, as suggested by a p-value of 0.179. However, there exists a statistically significant relationship between the dependent variable and the change in math scores (changeIn_Maths), with a p-value of <0.001. Assumption tests reveal deviations from normality and variance homogeneity, yet post-hoc tests fail to provide sufficient evidence to reject the null hypothesis of no significant difference among income groups. Therefore, while individual variables may show significance, overall differences among income groups might not be statistically significant.

Group1	Group 2	meandiff	p-adj
1	2	0.134	0.287
1	3	-0.006	0.998
2	3	-0.140	0.309

Table 4 - Pairwise Comparison of Income Groups

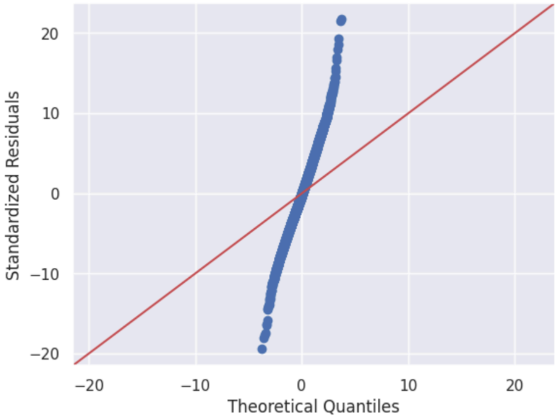

Fig 4 - QQ plot to check Normality

5. ΔReading score based on Income group with ΔGeneral knowledge as a covariate

Research Question 2: Ancova 2: Is there a change in the Reading score across the income groups if we set the 'ChangeIn_generalKnowledge' as the covariate.

	df	p-value
Income Group	2	<0.001
ΔGeneral Knowledge	1	<0.001
Residual	11929	NaN

Table 5 - Ancova 2 table

I used an Ancova test to analyze the relationship between the dependent variable '*changeIn_Reading'* and two independent variables '*incomegroup'* and '*changeIn_generalKnowledge'.* Using table 5, the p-value for Income Group suggests that there is a statistically significant difference in the dependent variable across different income groups. The null hypothesis associated with this p-value test is that there is no significant difference in the dependent variable across different levels of the income group. A p-value of <0.001 suggests that there is strong evidence to reject the null hypothesis at a conventional significance level (such as α = 0.05). In other words, the difference in the dependent variable across income groups is statistically significant.

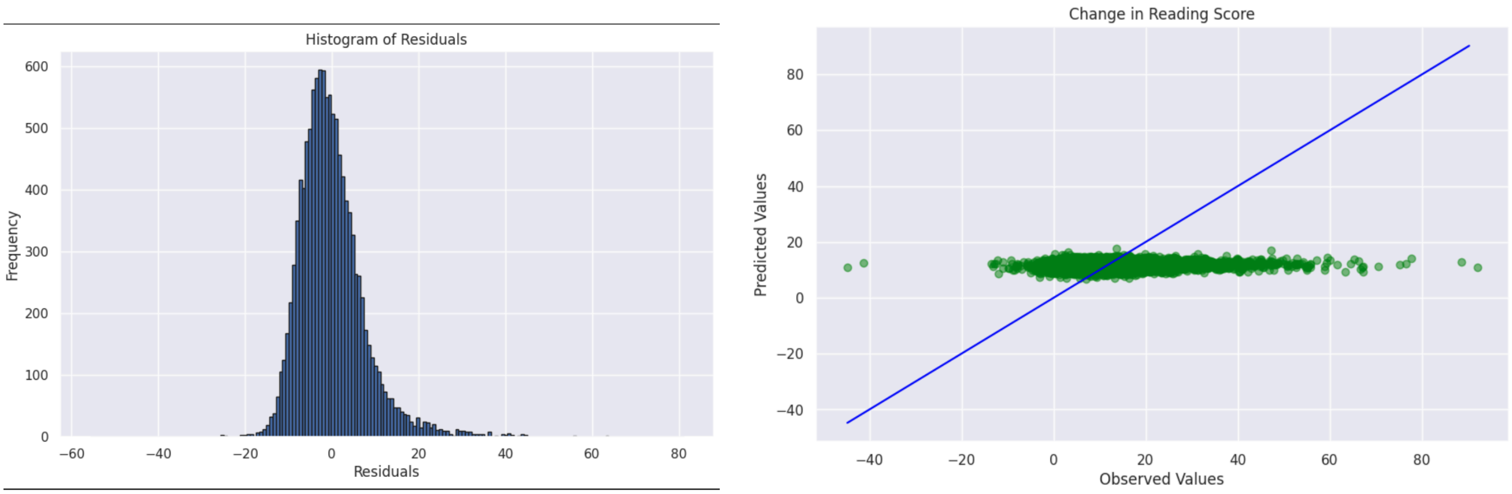

Fig 5 - Residual plot for Δ Reading Scores

Conclusion: In conclusion, the analysis indicates significant difference in the dependent variable across various income groups, as suggested by a p-value of <0.001. There exists a statistically significant relationship between the dependent variable and the income group with change in general knowledge score as the covariate. Additionally, the scatter plot shows a linear relation, thus meeting the assumptions of ancova.

6. ΔMaths score based on Income group with ΔGeneral knowledge as a covariate

Research Question 3: Ancova 3: Is there a significant change in the Maths score across the income groups if we set the 'ChangeIn_generakKnowledge' as the covariate.

I used an Ancova test to analyze the relationship between the dependent variable '*changeIn_Maths'* and two independent variables '*incomegroup'* and '*changeIn_generalKnowledge'*. Using table 6, the p-value for Income Group suggests that there is a statistically significant difference in the dependent variable across different income groups. The null hypothesis associated with this p-value test is that there is no significant difference in the dependent variable across different levels of the income group. A p-value of <0.001 suggests that there is strong evidence to reject the null hypothesis at a conventional significance level (such as α = 0.05). In other words, the difference in the dependent variable across income groups is statistically significant.

	df	**p-value**
Income Group	2	<0.001
ΔGeneral Knowledge	1	<0.001
Residual	11929	NaN

Table 6 - Ancova 3 table

Conclusion: In conclusion, the analysis indicates significant difference in the dependent variable across various income groups, as suggested by a p-value of <0.001. There exists a statistically significant relationship between the dependent variable and the income group with general knowledge score as the covariate. Additionally, the scatter plot shows a linear relation, thus meeting the assumptions of ancova.

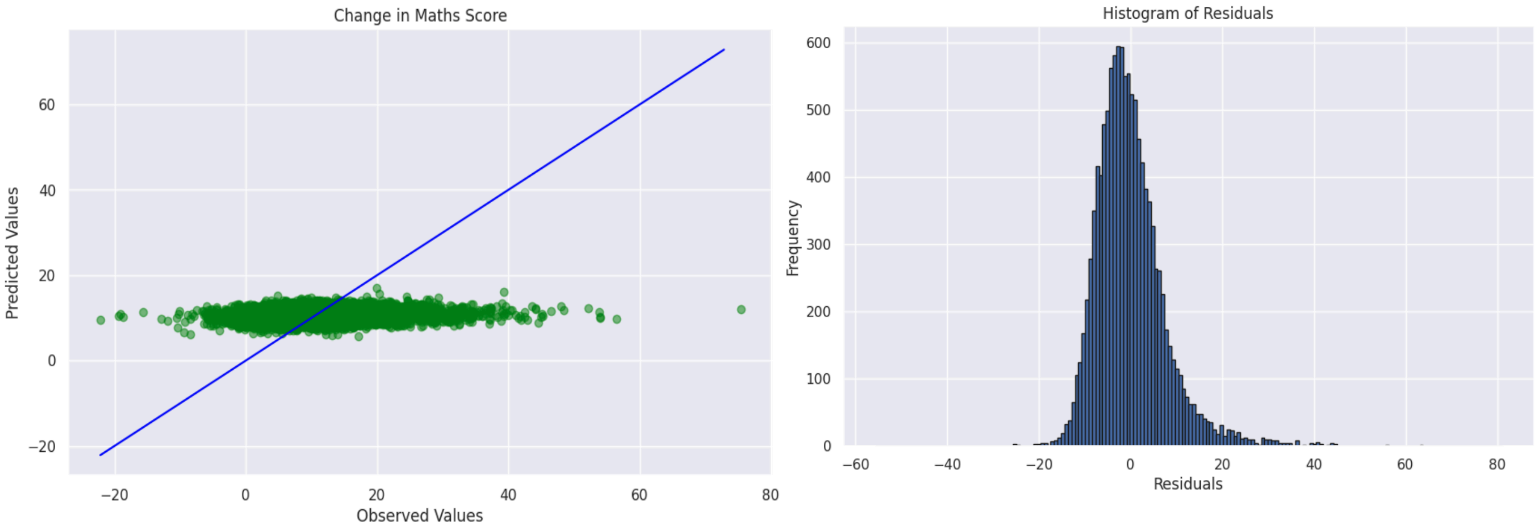

Fig 6 - Residual plot for Δ Maths score

7. Conclusion

In conclusion, this comprehensive analysis of kindergarten students' academic performance reveals intriguing insights into the interplay between socioeconomic status and educational outcomes. While there is no significant variance in general knowledge scores across income groups, a robust relationship exists between math scores and income levels, indicating socioeconomic disparities. Moreover, the reading scores exhibit statistically significant variations among income groups, emphasizing the role of socioeconomic factors in shaping early literacy skills. These findings underscore the importance of targeted interventions to mitigate disparities and promote equitable educational opportunities for all children, particularly in their formative years.

Note: Figures are small throughout the report because of space constraints, please refer to the .ipynb file for a clearer version of the figures.