INF2178 Assignment 3 - Mingli Wang

Introduction

Derived from a focused segment of an early childhood longitudinal study spanning the 1998-1999 academic year, this dataset seeks to explore the impact of family income on kindergarten students' academic achievements. It follows the evolution of reading, math, and general knowledge scores across two pivotal moments: fall 1998 and spring 1999, thereby mapping the students' developmental journey. Central to our analysis is the 'income category' variable, stratifying families by income level, which serves as a critical tool in examining the potential correlations and influences of socioeconomic factors on the educational outcomes of these young learners.

As for our research questions, in a manner akin to how the SAT evaluates students through aggregated scores in reading, math, and writing, we propose a composite score for each child by summing their reading, math, and general knowledge scores. Our first research question is to assess **[RQ1] how family income levels affect children's composite scores in the spring term, with an adjustment for their composite scores from the fall term.** Given the pivotal role of reading skills in facilitating new learning, we hypothesize that reading scores will serve as a significant covariate for general knowledge scores. Thus, our second research inquiry focuses on **[RQ2] the effects of income levels on children's general knowledge scores in the fall, controlling for the influence of fall reading scores.**

This dataset does more than just track the academic progress of kindergarteners; it invites a deeper conversation on educational equity and the influence of socioeconomic status on early educational experiences. Through this lens, we aim to uncover nuanced insights into how financial backgrounds shape or intersect with the educational pathways of children at a formative stage of their learning journey.

Data Wrangling & Feature Engineering

The dataset named "INF2178_A3_data.csv" has 9 columns and 11933 observations. Based on an initial examination, we determined that for the purpose of our analysis, not all columns are required and data cleaning is needed. Since the "totalhouseholdincome" and "incomeinthousands" are dependent features used to determine the income group, the two columns are not included in this analysis. In addition, the given dataset is relatively tidy with no missing values, therefore no action is required in this regard. The chart below provides the features used in this analysis and their corresponding description.

> **Feature Engineering:** To analyze our first research question, two new features are added by calculating a composite score of reading, math, and general knowledge scores of each student for the fall and spring terms respectively. They are named "fallsumscore" and "springsumscore".

Feature Name	Description	Type
fallreadingscore	Fall term reading score	Numerical
fallmathscore	Fall term math score	Numerical
fallgeneralknowledgescore	Fall term general knowledge score	Numerical
fallsumscore *(New feature)*	Summation of fall term math, reading, and general knowledge scores	Numerical
springreadingscore	Spring term reading score	Numerical
springmathscore	Spring term math score	Numerical
springgeneralknowledgescore	Spring term general knowledge score	Numerical
springsumscore *(New feature)*	Summation of spring term math, reading, and general knowledge scores	Numerical
incomegroup	The income group of the children's family is determined by total household incomes, split into 3 levels. "3" means a higher income, while "1" means a lower income.	Categorical

Exploratory Data Analysis

Our initial step in the analysis involved an exploratory data analysis (EDA) to delve into the dataset and unearth intriguing insights. We began by generating a summary of statistics for the numerical data, providing a comprehensive snapshot of children's scores across subjects from the fall to the spring terms, alongside an overview of household incomes.

	fall reading score	fall math score	fall general knowledge score	fall sum score	spring reading score	spring math score	spring general knowledge score	spring sum score	total household income
count	11933	11933	11933	11933	11933	11933	11933	11933	11933
mean	36	27.1	23.1	86.2	47.5	37.8	28.2	113.5	54317.2
std	10.5	9.1	7.4	23.2	14.3	12	7.6	29.2	36639.1
min	21	10.5	7	39.7	22.4	11.9	7.9	46	1
25%	29.3	20.7	17.4	69.8	39	29.3	22.8	93.6	27000
50%	34.1	25.7	23	83.3	45.3	36.4	28.6	111.2	47000
75%	39.9	31.6	28.3	98.5	51.8	44.2	33.8	128.7	72000
max	138.5	115.6	47.7	248.5	156.8	113.8	48.3	292.1	150000

To further our understanding of the data's distribution, histograms for each subject's scores in both the fall and spring terms, as well as for household incomes, were created (Figure 1). These visualizations aid in grasping the spread and central tendencies of the data.

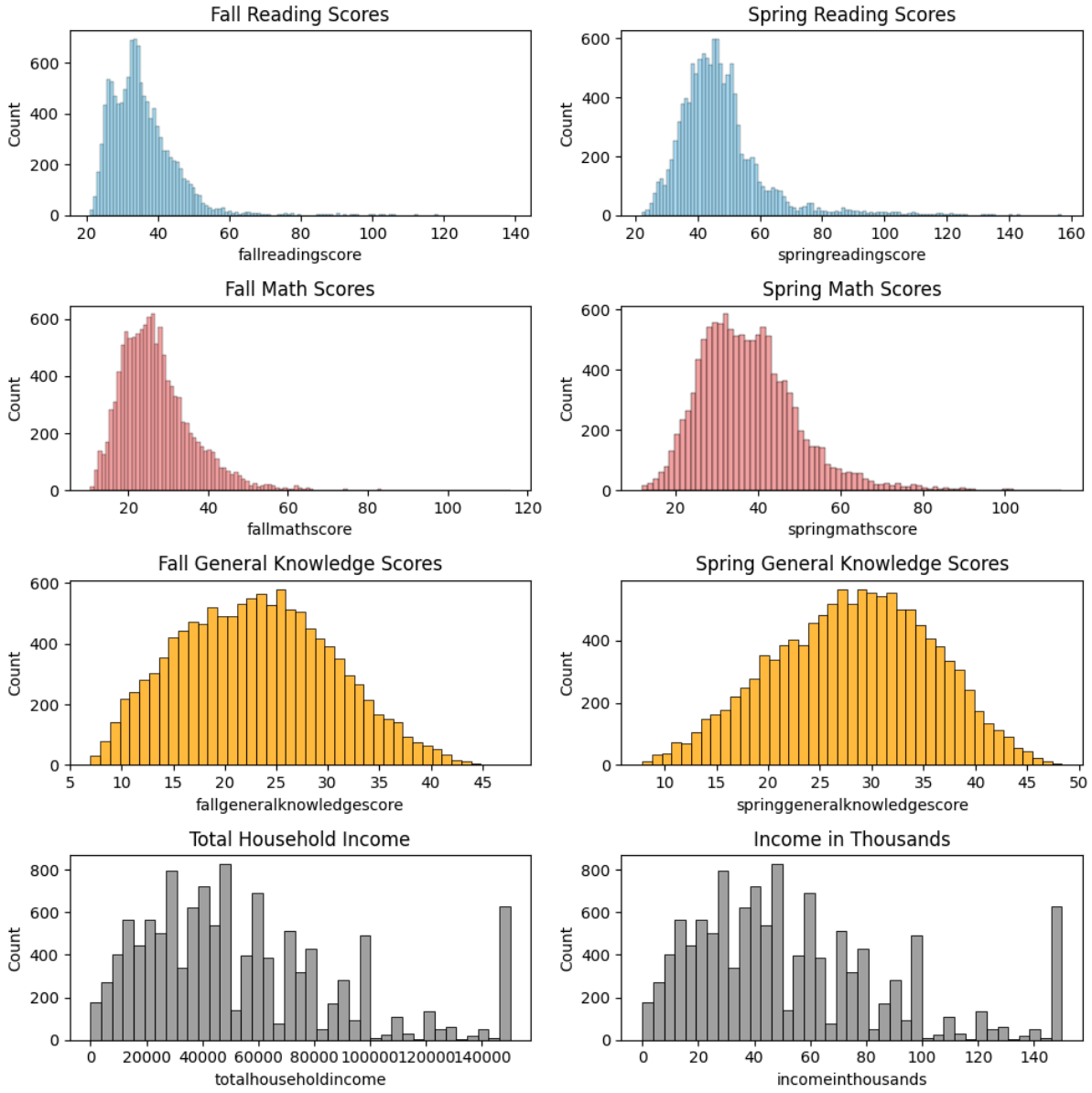

Figure 1

Key Findings:

1. The distributions of reading and math scores for both fall and spring terms exhibit a rightward skew, indicating a concentration of lower scores with fewer high-scoring outliers.
2. In contrast, the distributions of general knowledge scores in both terms approximate a normal distribution, showcasing a more balanced spread of scores around the mean.
3. Household income data reveals significant disparities and an uneven distribution, reflecting a broad spectrum of socioeconomic statuses among the children's families.
4. An overall trend observed is an improvement in children's scores from the fall term to the spring term, suggesting academic growth over the school year.

Family Income Group and Spring Composite Scores

We employed ANCOVA to investigate the influence of family income levels on children's composite scores in the spring term, adjusting for fall term scores. An initial scatter plot analysis revealed a stable linear relationship between fall and spring scores, with higher scores more prevalent among students from wealthier families (Figure 2). This trend was further supported by a boxplot analysis (Figure 3), showing that students in the highest income group outperformed those in lower income groups across various statistical measures.

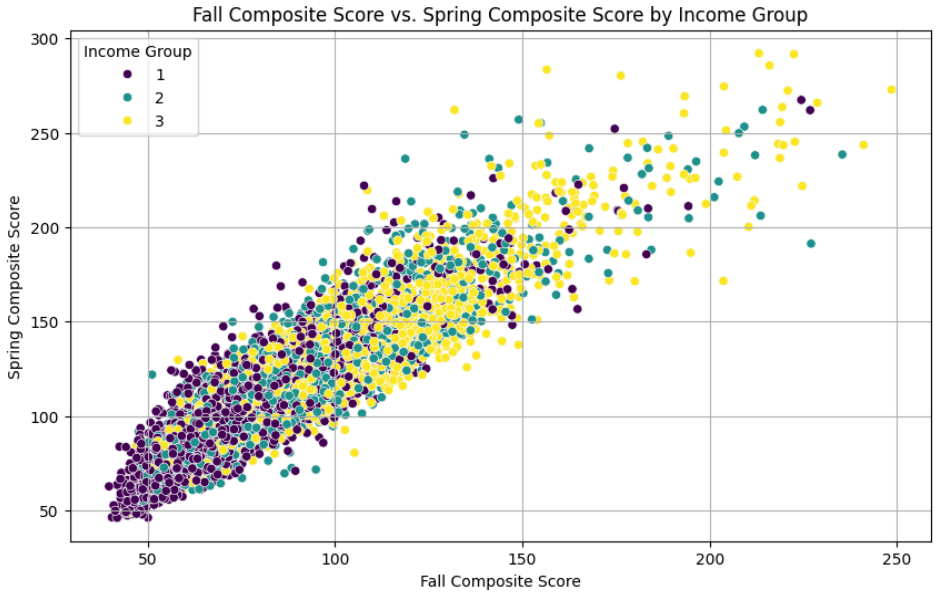

Figure 2

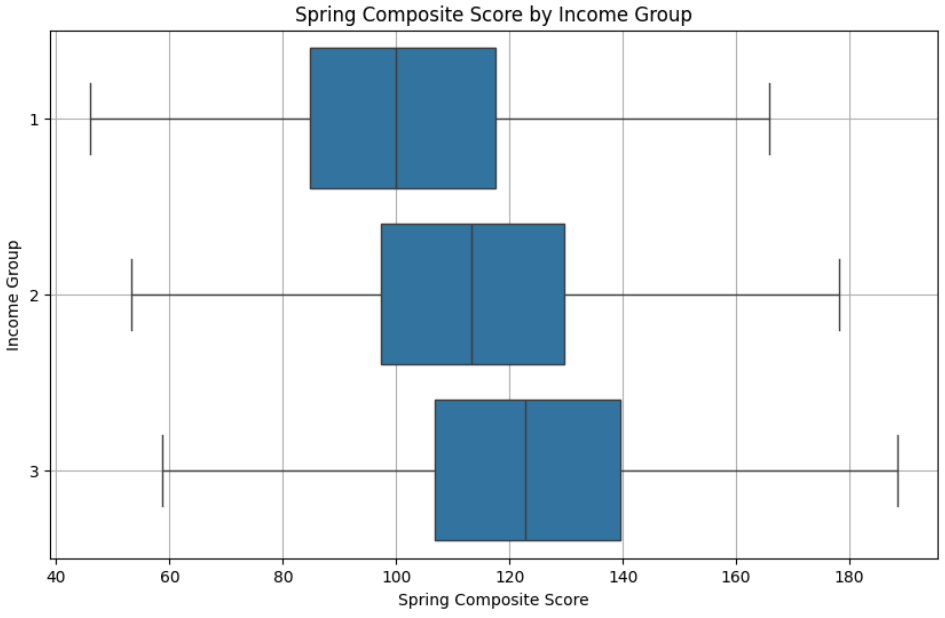

Figure 3

A one-way ANOVA confirmed significant differences in spring scores across income groups (p-value < 0.001). However, an ANCOVA analysis, controlling for fall scores, indicated that income group differences were not statistically significant (p-value = 0.174), highlighting fall scores as the dominant predictor of spring outcomes.

This suggests that while income may influence overall performance, academic progress over time is primarily dictated by previous scores. The necessity for targeted support to lower-income students is evident to bridge this performance gap. Nonetheless, it is critical to acknowledge that both the ANOVA and ANCOVA analyses did not satisfy the assumptions of normality and homogeneity of variances, necessitating a cautious interpretation of these results.

ANCOVA on Spring Comp Score and Income Group				
Source	**Sum of Squared**	**DF**	**F**	**P-value**
Income Group	7.012567e+02	2	2.055	0.128
Fall comp score	6.979486e+06	1	40898.997	<0.001
Residual	2.035875e+06	11930		
Shapiro Result	stat = 0.972		p-value < 0.001	
ANOVA on Spring Comp Score and Income Group				
Income Group	1.136497e+06	2	752.275	<0.001
Residual	9.011602e+06	11930		

Income Group and Fall-Term General Knowledge Scores

In our examination of how income groups affect children's general knowledge scores in the fall, after adjusting for fall reading scores, we first analyzed the relationship between reading and general knowledge scores through a scatter plot (Figure 4). This plot indicated a positive linear relationship, suggesting students with higher reading scores generally achieve higher general knowledge scores. Additionally, a boxplot categorized by income group (Figure 5) highlighted that students from higher-income families tend to score better, aligning with findings from previous analyses on composite scores.

To quantitatively assess the impact of income groups on general knowledge scores, controlling for reading proficiency, we conducted an ANCOVA. The results revealed statistically significant effects of both income groups (p-value < 0.001) and reading scores (p-value < 0.01) on general knowledge scores. This signifies that socioeconomic factors, alongside academic abilities, crucially influence children's educational outcomes.

ANCOVA on Fall General Knowledge Score and Income Group				
Source	**Sum of Squared**	**DF**	**F**	**P-value**
Income Group	38477.157	2	499.852	<0.001
Fall comp score	105360.481	1	2737.499	<0.001
Residual	459130.042	11929		
Shapiro Result	stat = 0.998		p-value < 0.001	

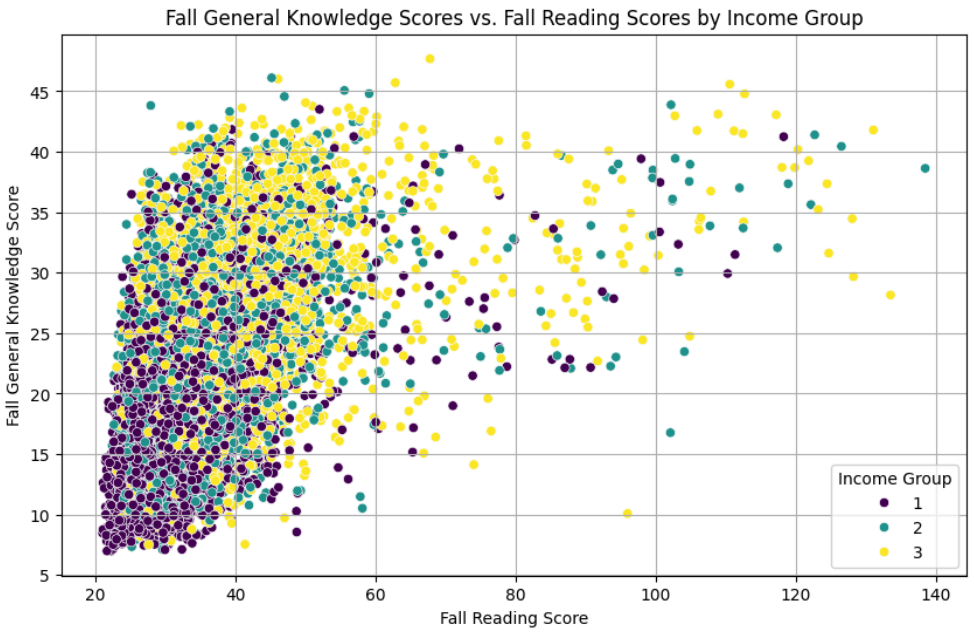

Figure 4

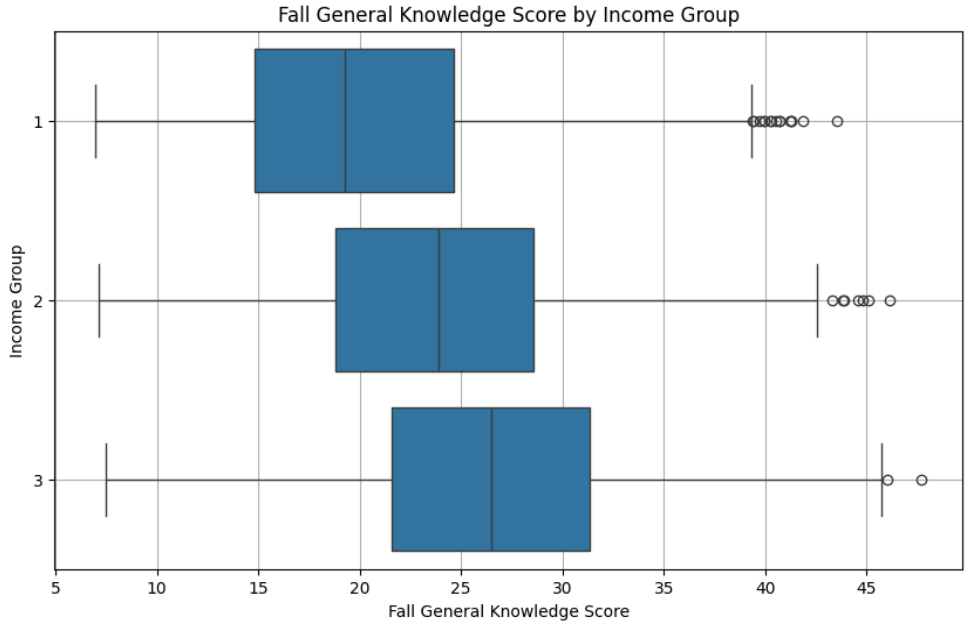

Figure 5

Conclusion

Our analysis addressed two key questions about family income's impact on children's education. First, we found that initial performance predicts spring outcomes, with income's direct effect being non-significant when controlling for fall scores. This indicates the potential of early academic interventions. Secondly, income levels and reading skills significantly affect fall general knowledge scores, highlighting the importance of addressing both educational and socioeconomic disparities to improve educational outcomes. If resource permits, future analysis can break down the income groups into smaller divisions for more intricate results.