Early Wins: Navigating Kindergarten Scores

1. Introduction

In the journey of a child's development, early period holds paramount significance as they gained the foundations for lifelong learning. Kindergarten scores serve as vital signals, providing insights into social adaptability, and overall readiness for the educational path that lies ahead. Moreover, the family's income level emerges as a pivotal role influencing early childhood development and their kindergarten scores. Economic circumstances can shape access to educational resources, enriching experiences, thereby impacting a child's social development. Understanding this connection becomes crucial in crafting educational interventions to ensure equitable opportunities for all early children.

This report provides a comprehensive exploratory data analysis of early child kindergarten scores focused on fall 1998 and spring 1999 with a goal to uncover any relationship between the individual scores and the family incomes. To unravel these patterns, we delve into the dataset named'INF2178_A3_data.csv' (accessible in the GitHub repository), tracking the early child's test scores of readings, math, and general knowledge to evaluate their performance over several months.

Our exploration will address two fundamental research questions, serving as guiding principles in examining whether the families from different income groups has effect on kindergarten students' grade performance patterns:

- **Research Question 1**: How do fall term general knowledge scores vary across different income groups, controlling the fall term reading scores as a covariate? Does the income group is an important indicator to affect fall term general knowledge scores?
- **Research Question 2**: Does the income group of kindergarten students impact their general knowledge scores in spring after controlling for their general knowledge scores in fall term?

By addressing these research questions, we aim to provide valuable insights into the dynamics of improving the kindergarten students' scores with a comprehensive understanding that can inform more effective interventions/solutions.

2. Data Cleaning and Data Processing

The raw dataset has a total of **9 columns** with **11,933** entities **(rows)**. After the initial review of the dataset, we thought a basic data cleaning was necessarily for the scope of our analysis. Below we showed the observations of our dataset.

A. Observations and Considerations:

Since our analysis is quantitative, we must work on the specific columns from the raw dataset. Below we provided a short description of 8 Numeric Variables and 1 Categorical Variable:

8 Numeric Variables:

- *fallreadingscore*: Reading scores of early children in the fall semester
- *fallmathscore*: Math scores of early children in the fall semester
- *fallgeneralknowledgescore*: General knowledge scores of early children in the fall semester
- *springreadingscore*: Reading scores of early children in the spring semester
- *springmathscore*: Math scores of early children in the spring semester
- *springgeneralknowledgescore*: General knowledge scores of early children in the spring semester
- *totalhouseholdincome*: Total household income from their families
- *incomeinthousands*: Total household income from their families (in thousands)

1 Categorical Variable:

- *incomegroup*: Derived from the income variable

B. The raw dataset does not have any missing values (NaN): We could say our data is carefully curated. The absence of missing values served as a testament to the robustness of the dataset's quality, ensuring that every data point was accounted for and accurately represented.

3. Exploratory Data Analysis (EDA)

We conducted with a comprehensive EDA to gain insight that could potentially lead to our research questions. As we started by describing the quantitative data as shown in *Figure 1* below. Additionally, we used histograms and bar plots seen in *Figure 2* and *Figure 3* to visualize the overall distribution of 8 numerical fields and 1 categorical field. The descriptive analysis provided **a clearer picture of kindergarten student's scores in different semesters and the general trends with income groups**.

	fallreadingscore	fallmathscore	fallgeneralknowledgesco	springreadingscore	springmathscore	springgeneralknowledge	totalhouseholdincome	incomegroup
count	11933	11933	11933	11933	11933	11933	11933	11933
mean	35.954215	27.128244	23.073694	47.511178	37.799461	28.235584	54317.19993	1.859165
std	10.47313	9.120505	7.396978	14.327111	12.027753	7.577457	36639.06115	0.822692
min	21.01	10.51	6.985	22.35	11.9	7.858	1	1
25%	29.34	20.68	17.385	38.95	29.27	22.802	27000	1
50%	34.06	25.68	22.954	45.32	36.41	28.583	47000	2
75%	39.89	31.59	28.305	51.77	44.22	33.782	72000	3
max	138.51	115.65	47.691	156.85	113.8	48.345	150000	3

Figure 1: Dataset Quantitative Data Statistics

Based on statistical data table above, we can observe that the average scores of three subjects in the spring term have an upward trend compared to the scores in previous fall semester. It might be proven that knowledge learned at fall term has a relationship with whether students can get a good grade in spring term.

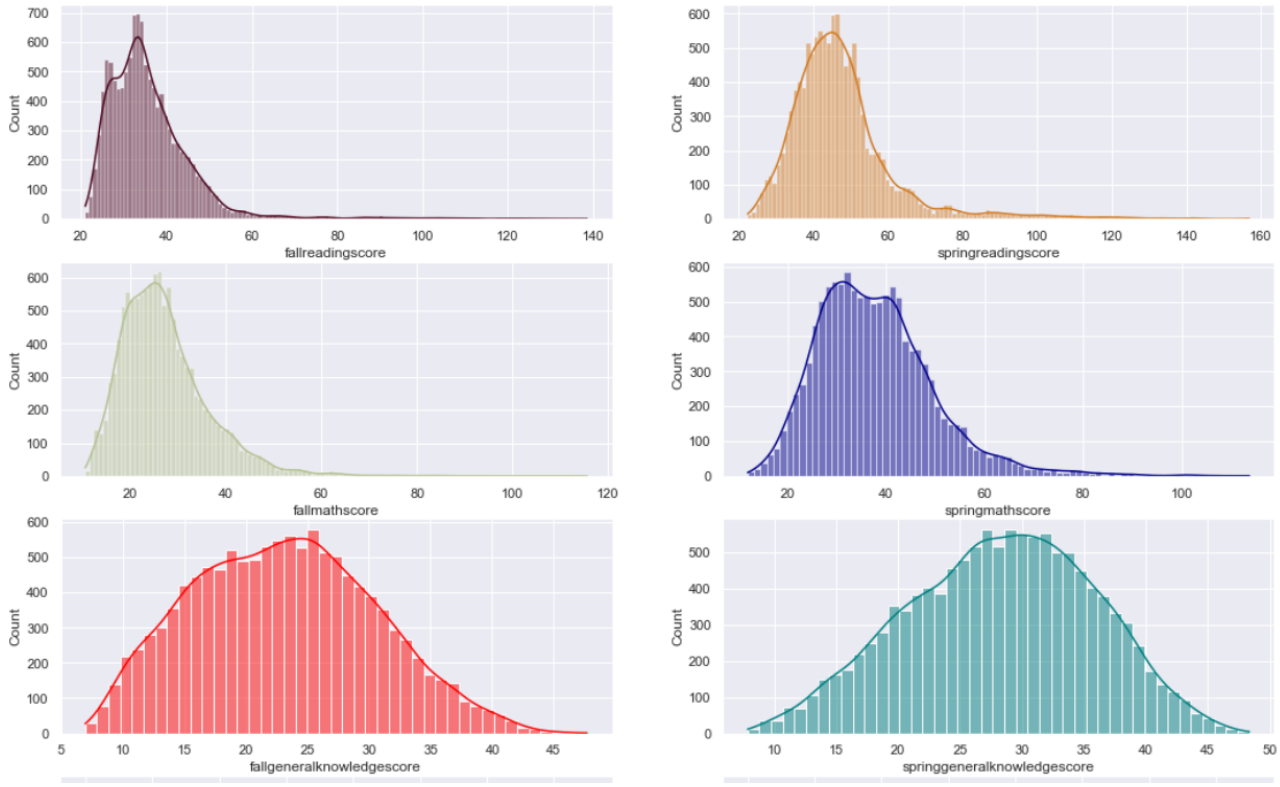

Figure 2: Histograms of Six Scores Data of Different Semesters

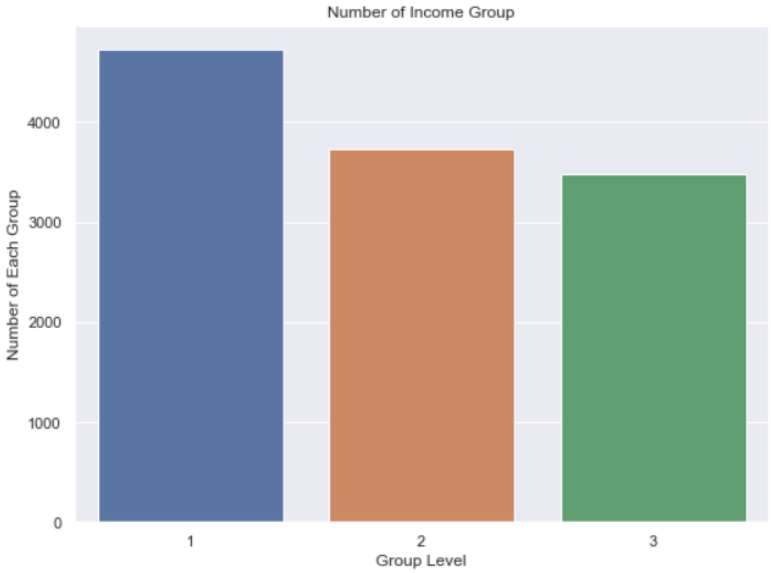

Figure 3: Bar plot of the Three Income Groups

Based on the histograms in *Figure 2*, we can see the overall distribution for **general knowledge scores** in both fall and spring term are **approximately normal**. As we can observe from first four plots, the **knowledge that students learned from fall term provide a foundation for the spring term.** In this way, we could say the students who did good at fall term are more likely to did good job again in the spring term.

4. Examine Fall General Knowledge Scores by Income Groups

Research Question #1: How do fall term general knowledge scores vary across different income groups, controlling the fall term reading scores as a covariate? Does the income group is an important indicator to affect fall term general knowledge scores?

For this analysis, we used ANCOVA to investigate whether the Fall General Knowledge Score is statistically significant under different Income Groups.

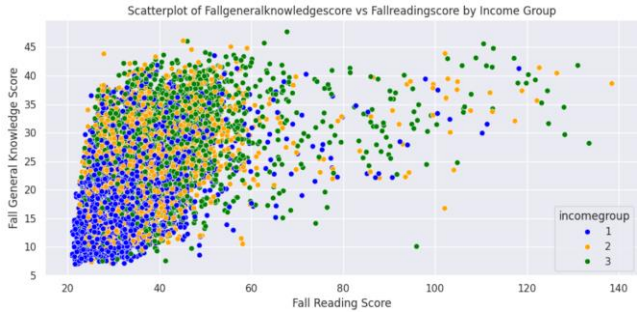

Figure 4: Scatterplot of Fall general knowledge score vs. Fall reading score by Income Group

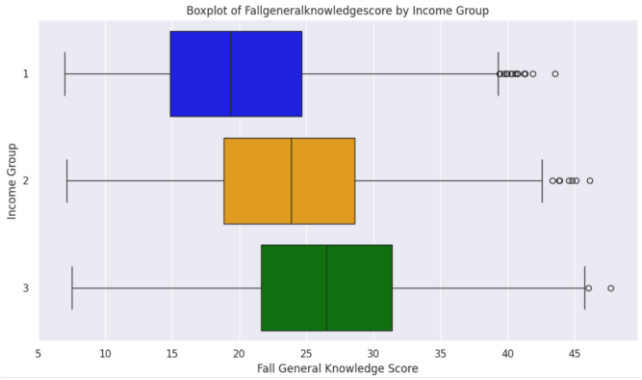

Figure 5: Boxplot of Fall general knowledge score by Income Group

At *Figure 4*, the scatterplot roughly showed a **positive linear relationship** that the higher reading score will have higher general knowledge score. We can also observe at the boxplot as *Figure 5*, **Fall General Knowledge Score is significant with different Income Groups**. Each group has a few data points are considered outliers in statistical area. In fact, these outliers are the highest scores from individual who is performing better than the most students. Keeping the outliers will improves the generalization of the model (positive noise) and provides more comprehensive data information.

	Source	Sum of Squared (SS)	DF	F	P-value	np2
0	Income group	38477.15676	2	499.851851	0.001	0.077324
1	fallreadingscore	105360.4815	1	2737.449236	0.001	0.186647
2	Residual	459130.0422	11929	NaN	NaN	NaN

Figure 6: ANCOVA Table for Fall General Knowledge Score

Observing from the ANCOVA table, both the **Income Group and Fall Reading Score showed statistically significant** to our dependent variable, as the P-value is very small

(P<0.001). At the beginning, we consider that student tends to get higher scores if their families have higher income. However, income was not the most important factor for Fall General Knowledge Score. As you can see from the table above, income groups only can explain about 7.7% of the variance in Fall General Knowledge Score and Fall Reading Score can explain about 18.7% of the variance. Therefore, **Fall Reading Score have a greater impact on Fall General Knowledge Score**. We should focus on how to improve student's reading score, such as build public libraries, to enhance their general knowledge scores.

5. Different Semesters of General Knowledge Score by Income Groups

Research Question #2: Does the income group of kindergarten students impact their general knowledge scores in spring after controlling for their general knowledge scores in fall term?

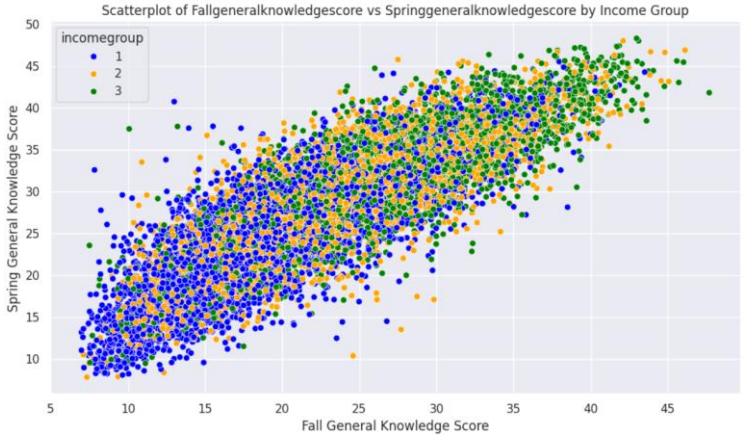

Figure 7: Scatterplot of Fall General Knowledge Score vs. Spring General Knowledge Score by Income Group

	Source	Sum of Squared (SS)	DF	F	P-value	np2
0	Income group	1756.905	2	56.908	0.001	0.009451
1	fallgeneralknowledgescore	411876.768	1	26682.269841	0.001	0.691049
2	Residual	184140.189	11929	NaN	NaN	NaN

Figure 8: ANCOVA Table for Spring General Knowledge Score

Similar to the method used in first research question, when tries to examine the factors that would influence Spring General Knowledge Score. We know there have mean difference of Spring General Knowledge Score by Income Group, which means higher income level will have higher scores. At *Figure 7*, there has a **strong positive linear relationship** between the fall scores and spring scores of general knowledge course. In other word, student who did well in fall term will have a good performance in spring term as well. Fall General Knowledge Score can explain about 69.11% of variance in Spring General Knowledge Score. Also, the **R-squared value is 0.731** means that

approximately 73.1% of the variability in the Spring General Knowledge Score can be explained by the Fall General Knowledge Score in our model. Therefore, the impact of income groups on Spring General Knowledge Score is not that much, students need to focus more on Fall General Knowledge course which can improve their grade performance in the right way.

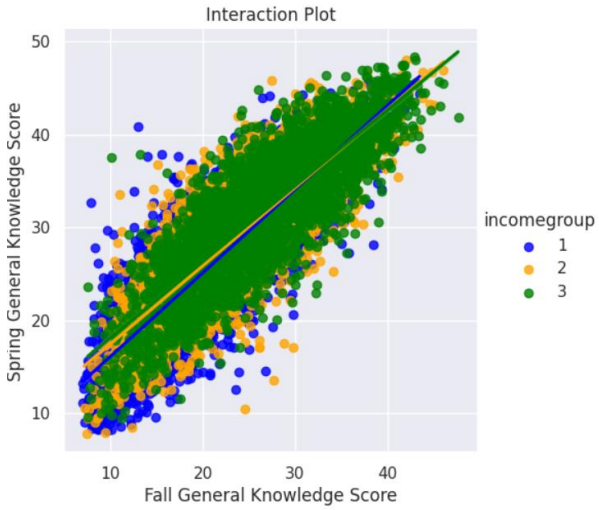

Figure 9: Interaction Plot for Spring General Knowledge Score

The interaction plot shows that all income groups exhibited a positive interaction between fall and spring general knowledge scores, the slopes indicating the strength of the relationship may vary across different income groups. However, the lower-income group may experience a higher rate of increase in scores, whereas the higher-income group may have a slower rate of increase in scores (displayed by flatter slopes in the plot).

Test Assumptions: Normality and Homogeneity of Variance

✧ **Levene's test P-value**: 8.286000485876772e-05 (P-value < 0.001)

After we tested the assumptions, we found the **normality assumption is satisfied**. However, the p-value is less than 0.001 (reject H0), it is **violated the assumption of homogeneity of variance**. It means that the model may incorrectly inform significant differences between certain groups or fail to detect the differences that may exist.

6. Conclusion

Through the data visualization and quantitative analysis, we gained valuable insights about key factors that may have impact on students' score performance. Our findings showed a substantial impact from three different income groups, which means the higher income families could access to more educational resources/opportunities. Perhaps, this knowledge can guide policymakers in creating more effective ways to improve students' grade. There could be some **educational interventions**. For example, the income groups have a significant impact on students' general knowledge scores in the Spring term. The government should consider **offering additional support for families in the lower-income group**, such as after-school tutoring, to help those students improve their grades.