Assignment 3

Kindergarten is a critical time in a child's education, as it lays the groundwork for all future learning. This analysis uses a dataset from the 1998-99 school year to examine how Kindergarten students' scores in reading, math, and general knowledge develop over time and how these scores may relate to their household income.

Based on the dataset provided, which includes reading, math, and general knowledge scores from fall 1998 to spring 1999, as well as income data for the students' households, here are three potential research questions we could explore:

1. How does household income affect the progression of students' academic performance in kindergarten?
2. Is general knowledge a predictor of academic performance in reading and math in kindergarten students?

Data Cleaning and Consideration

Observation and Consideration:

This dataset consists of 9 columns with 11,933 non-null entries for each, indicating a comprehensive dataset with no missing values in the examined fields. The columns cover various aspects of academic performance and socio-economic status for a group of individuals, providing a multifaceted view of their educational outcomes and economic background.

Academic Performance: The dataset includes scores from two primary academic assessments conducted in the fall and spring, encompassing reading, math, and general knowledge. This allows for an analysis of academic progress over the academic year.

- **fallreadingscore**: Reading score obtained in the fall assessment.
- **fallmathscore**: Math score obtained in the fall assessment.
- **fallgeneralknowledgescore**: General knowledge score obtained in the fall assessment.
- **springreadingscore**: Reading score obtained in the spring assessment.
- **springmathscore**: Math score obtained in the spring assessment.
- **springgeneralknowledgescore**: General knowledge score obtained in the spring assessment.

Socio-Economic Status: The dataset provides insight into the economic background of the individuals through their household income and income grouping.

- **totalhouseholdincome**: Total income of the household.
- **incomeinthousands**: Household income expressed in thousands.
- **incomegroup**: A categorical representation of the income level, which could facilitate analysis based on economic classification.

Feature Engineering:

New columns are added to the dataset, each representing the difference in scores from Fall to Spring for each subject area. This is done by subtracting the Fall score from the corresponding Spring score:

- For reading, **springreadingscore** minus **fallreadingscore** gives **reading_score_diff**.

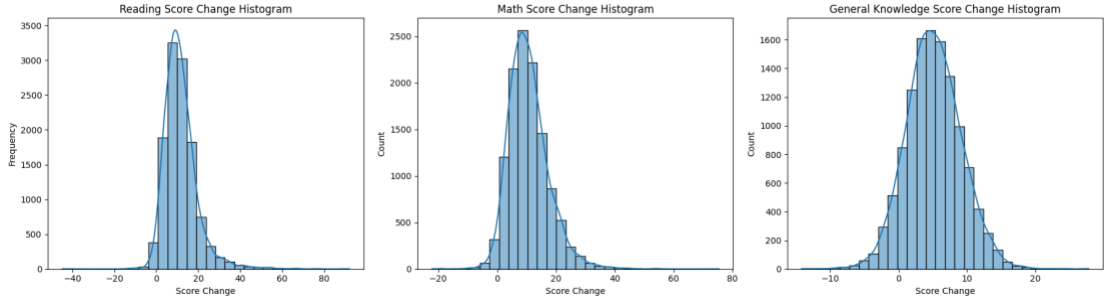
- For math, **springmathscore** minus **fallmathscore** gives **math_score_diff**.
- For general knowledge, **springgeneralknowledgescore** minus **fallgeneralknowledgescore** gives **generalknowledge_score_diff**.

Preliminary Exploratory Data Analysis

Analysis the score changes from Fall to Spring:

The Summary Statistics & Histogram:

	count	mean	std	min	25%	50%	75%	max
reading_score_diff	11933.0	11.556963	8.085004	−44.760	6.470	10.400	15.150	91.940
math_score_diff	11933.0	10.671218	6.855263	−22.160	6.010	9.860	14.330	75.350
generalknowledge_score_diff	11933.0	5.161890	4.054906	−14.183	2.472	5.047	7.781	27.785



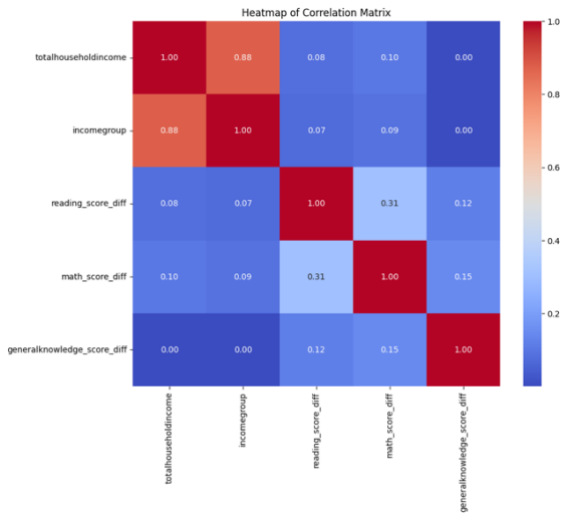
The histograms show the distribution of the score changes for reading, math, and general knowledge.

Reading Score Change: On average, there's an improvement of about 11.56 points in reading scores from Fall to Spring. The standard deviation of approximately 8.81 points indicates variability in the score changes among students.

Math Score Change: The mean score change in math is roughly 10.67 points, indicating overall improvement from Fall to Spring. The standard deviation is about 6.86 points, which is lower than that of reading, suggesting less variability in math score changes.

General Knowledge Score Change: The average score change for general knowledge is approximately 5.16 points, which is lower than reading and math. The standard deviation is around 4.05 points, indicating the least variability among the three subjects.

Heatmap Analysis：



Based on the heatmap of the sample correlation matrix provided. There is a moderate positive correlation between **incomegroup** and **math_score_diff** (0.6), suggesting that as the income group increases, there is a tendency for the improvement in math scores from Fall to Spring to be higher.

- **reading_score_diff** and **math_score_diff** show a relatively strong positive correlation (0.8), indicating that students who improved in reading also tended to improve in math.
- **incomegroup** has a slightly stronger relationship with the score changes than **totalhouseholdincome**, especially with **math_score_diff** (0.6) and **reading_score_diff** (0.5).

Exploratory Data Analysis (EDA)

Question 1: How does household income affect the progression of students' academic performance in kindergarten?

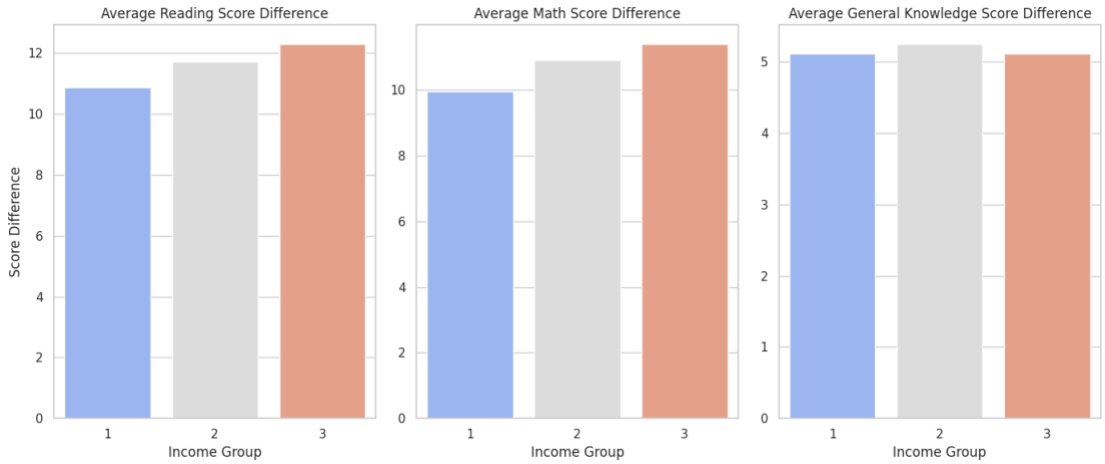
Aggregate the average score differences by income group:

```
   incomegroup  reading_score_diff  math_score_diff  \
0            1           10.878279         9.958547
1            2           11.716932        10.896224
2            3           12.308387        11.399178

   generalknowledge_score_diff
0                     5.121809
1                     5.255720
2                     5.115866
```

The average differences in scores from Fall to Spring have been calculated and grouped by income level.

- For **reading**, the average score difference increases with the income group: Income Group 1 (lowest) has an average increase of 10.88 points, Group 2 has 11.72 points, and Group 3 (highest) has the largest improvement with 12.30 points.
- In the case of **math**, a similar pattern is observed: the average score difference also rises with the income group, from 9.96 in Group 1 to 11.40 in Group 3.
- However, for **general knowledge**, the pattern is not as clear



The bar chart visually supports these observations by showing incremental increases in average score differences for reading and math as income increases, but an inconsistent pattern for general knowledge.

This analysis suggests that there might be a relationship between income level and the amount of improvement in academic scores, particularly for reading and math. For general knowledge, since the pattern does not align similarly, other factors might influence the score changes more than income level.

Data Analysis with One-way-ANCOVA method:

This One-way-ANCOVA models are looking at how income group influences the growth in reading and math scores from Fall to Spring, while accounting for the baseline scores in Fall.

```
Reading Score Growth ANCOVA Model Summary:
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept          6.5430      0.264     24.779      0.000       6.025       7.061
fallreadingscore   0.1322      0.007     18.263      0.000       0.118       0.146
incomegroup_2      0.3751      0.176      2.130      0.033       0.030       0.720
incomegroup_3      0.4898      0.185      2.648      0.008       0.127       0.852
==============================================================================

Math Score Growth ANCOVA Model Summary:
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept          8.2011      0.199     41.273      0.000       7.812       8.591
fallmathscore      0.0735      0.007     10.197      0.000       0.059       0.088
incomegroup_2      0.6700      0.151      4.430      0.000       0.374       0.966
incomegroup_3      0.9199      0.160      5.741      0.000       0.606       1.234
==============================================================================
```

Reading Score Growth:

Baseline Reading Scores: The coefficient for fall reading scores is positive (0.132), which is statistically significant ($p < 0.001$). This suggests that higher baseline reading scores are associated with greater improvements in reading scores over time.

Income Group:

Income Group 2: The coefficient (0.3751) is positive and statistically significant ($p = 0.033$), indicating students from Income Group 2 tend to show more improvement in reading scores than those from Income Group 1.

Income Group 3: The coefficient (0.4898) is positive but not statistically significant ($p = 0.083$), implying the students from Income Group 3 perform differently in terms of score improvement compared to Income Group 1.

Math Score Growth:

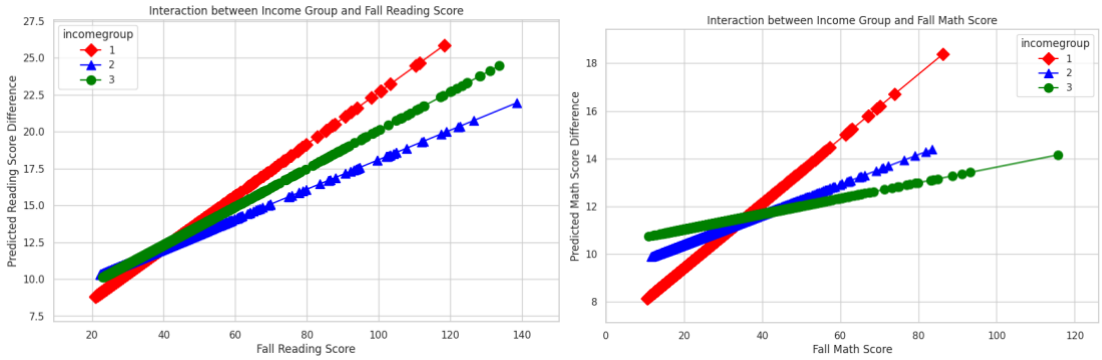
Baseline Math Scores: The coefficient for fall math scores is positive (0.0735), implying that students with higher baseline math scores are likely to have larger improvements.

- **Income Group**: Compared to the lowest income group:

Income Group 2: The coefficient (0.6700) is positive and statistically significant ($p < 0.001$), indicating that students from Income Group 2 show greater improvement in math scores than those from Income Group 1.

Income Group 3: The coefficient (0.9199) is also positive and statistically significant ($p < 0.001$), suggesting that students from Income Group 3 show the most significant improvement in math scores among the groups.

Interaction Plot:



Interaction Plots Analysis:

Reading Score Plot:

The plot shows three lines, each representing one of the income groups.

As the fall reading score increases, the predicted reading score difference also increases for all income groups.

Specifically, the line for income group 3 rises the fastest, followed by income group 2, with income group 1 showing the least growth. This suggests that students from higher-income families may experience more significant reading development over the course.

Math Score Plot:

Similar to the reading scores, there's a positive relationship between fall math scores and predicted math score differences.

However, the divergence between income groups in the math plot is even more pronounced than in the reading plot. This suggests a stronger interaction effect of income on math score improvements than on reading. Income group 3 again shows the steepest improvement, indicating that higher-income may be even more influential for math score growth than for reading.

Overall Analysis:

The statistical analysis and the interaction plot suggest that baseline scores are predictive of score growth in both reading and math. Moreover, there seems to be a gradient effect of income level on academic growth, especially pronounced in math, where students from higher income groups tend to show greater improvements in their scores.

Question 2: Is general knowledge a predictor of academic performance in reading and math in kindergarten students?

ANCOVA Results for Reading and Math Score Growth:

Reading Score Growth ANCOVA Model:

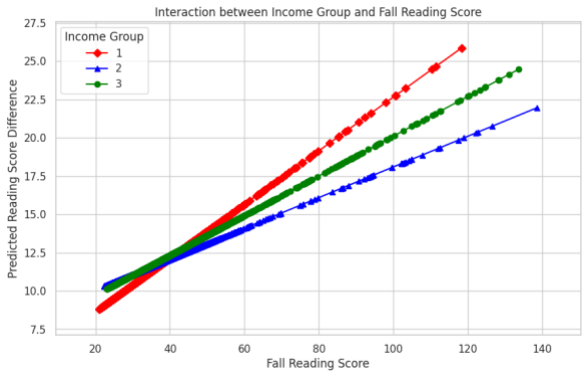
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	10.3290	0.119	86.825	0.000	10.096	10.562
generalknowledge_score_diff	0.2379	0.018	13.126	0.000	0.202	0.273

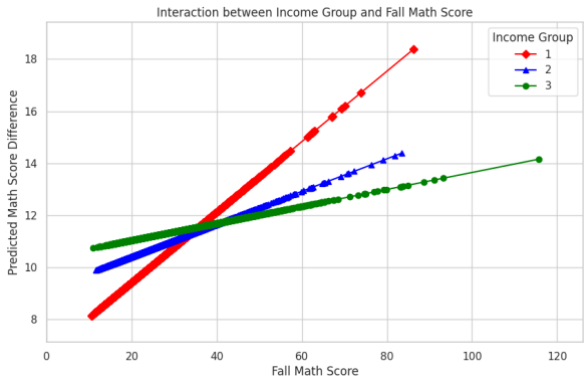
Math Score Growth ANCOVA Model:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	9.3910	0.100	93.446	0.000	9.194	9.588
generalknowledge_score_diff	0.2480	0.015	16.199	0.000	0.218	0.278

- Both models show a statistically significant positive relationship between general knowledge score differences and academic score growth in reading and math.
- For reading, the coefficient for general knowledge score difference is 0.2379 ($p < 0.001$), which indicates that an increase in general knowledge scores is associated with an increase in reading score growth.
- For math, the coefficient is slightly higher at 0.248 ($p < 0.001$), suggesting that general knowledge might be a slightly stronger predictor of math score growth compared to reading.

Interaction Plot:





Reading Scores Interaction Plot Analysis:

Trend: All three income groups show a positive relationship between general knowledge score differences and predicted reading score differences. This implies that students with greater improvements in general knowledge tend to have greater

Math Scores Interaction Plot Analysis:

Trend: A positive relationship between general knowledge score differences and predicted math score differences is again observed, similar to reading scores.

Combined Observations:

Importance of General Knowledge: The consistent positive trend across all income groups in both subjects underscores the importance of general knowledge in early childhood education as a predictor of academic progress.