Introduction

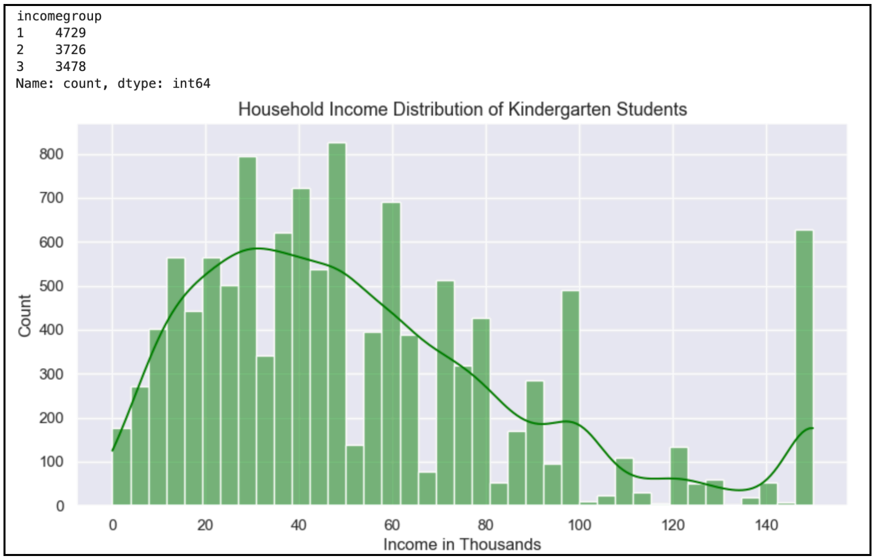
The dataset used in this assignment contains evaluation scores of Kindergarten students for Fall 1998 and Spring 1999. These scores include general knowledge, reading and math scores across the two semesters. The dataset also contains information on the household income of the students, which is then grouped into three categories based on their values. The aim of the assignment is to study the changes to the spring semester scores and study for any differences based on the income groups, controlling for the respective fall semester scores.

Data Pre-Processing

To access the variables more efficiently, I converted the 'incomegroup' variable into a object variable to maintain their structure and renamed the column names to a more concise version for readability. Further, I created three new variables ('Diff_GK_Score', 'Diff_Reading_Score', 'Diff_Math_Score') to assign the difference in values of the scores from Fall 1998 to Spring 1999, across all three subjects - general knowledge, reading and math.

Exploratory Data Analysis

To study the overall distribution of the scores for both semesters, I constructed a faceted summary statistics output based on the relevant subject and semester. Following this, I created a faceted output for the respective combinations of subject scores and semesters to visualize these distributions. The results show an overall larger spread of score values with higher means across the spring semester compared to the previous year's fall semester, confirmed by the summary statistics as well. Next, to understand the distribution of the household income of the kindergarten students, the results showed that group 1 of 'incomegroup' had the highest count of records, followed by groups 2 and 3; with the highest frequency between 25,000 to 40,000 units as seen in Figure 1.



```
incomegroup
1    4729
2    3726
3    3478
Name: count, dtype: int64
```

Figure 1. Kindergarten Students' Household Income Distribution

Additionally, to study for any trends between household income and the scores across both semesters, I created a series of trend plots, which confirmed a linear relationship across all combinations of subject scores and semesters. Lastly, I wanted to study the magnitude of differences between the subject scores between the two semesters.

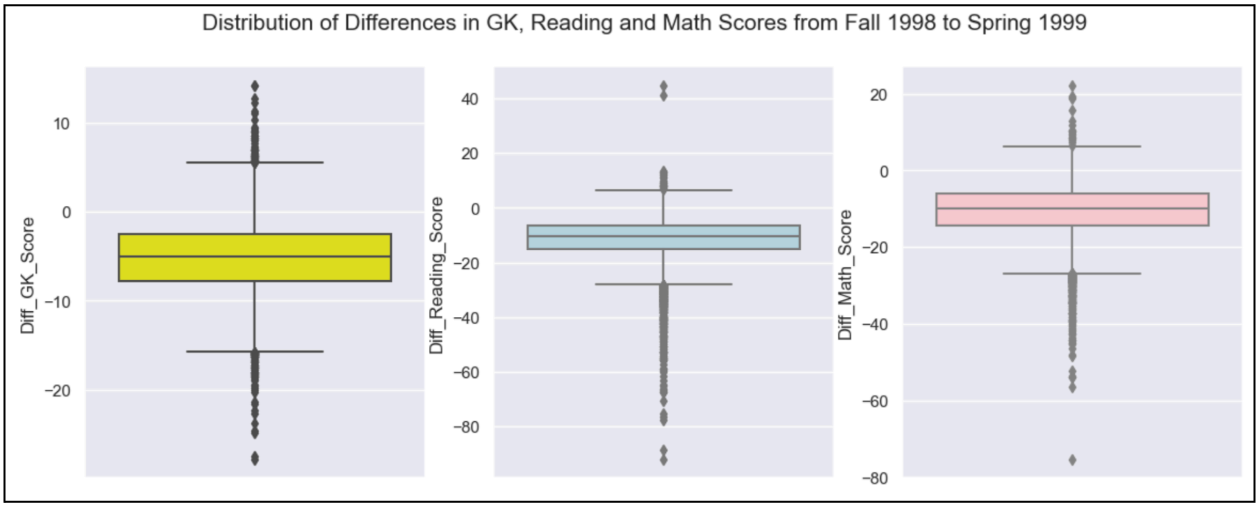


Figure 2. Distribution of Differences in Subject Scores from Fall 1998 to Spring 1999

As seen in Figure 2, the data shows a huge majority of negative differences among all the subject scores, which implies that the kindergarten students scored higher for these subjects in the following Spring semester than they did in their Fall semester.

For reference:

Hypotheses of Statistical Tests used for testing the assumptions -

- **Shapiro Wilk Test (For testing normality)**
    - **Null Hypothesis -** Data is drawn from normal distribution
    - **Alternative Hypothesis -** Data is not drawn from normal distribution
- **Levene's Test (For testing homogeneity of variances)**
    - **Null Hypothesis -** Samples from populations have equal variances
    - **Alternative Hypothesis -** Samples from populations do not have equal variances

One-Way ANCOVA - I

RQ: Does the income group have an impact on the kindergartner's spring general knowledge scores, having controlled for the student's fall general knowledge scores?

Null hypothesis: Means of all kindergartners' spring general knowledge scores based on each income group are equal after controlling the effect of kindergartners' fall general knowledge score i.e. adjusted means are equal

Alternative hypothesis: At least, one kindergartner's spring general knowledge score for an income group is different from other income groups after controlling the effect of kindergartners' fall general knowledge score i.e. adjusted means are not equal

Interaction Plot -

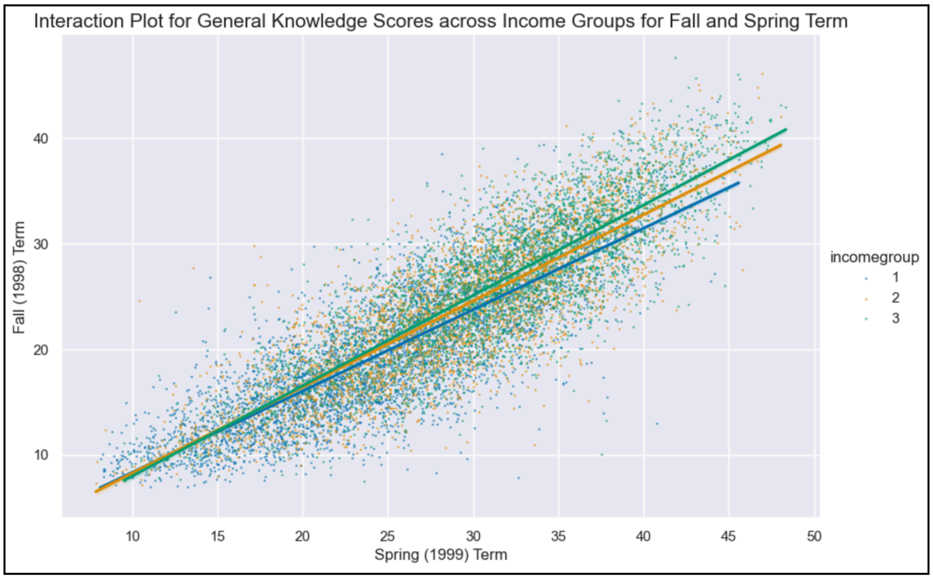


Figure 3. Interaction Plot

As seen in Figure 3, the plot lines are roughly parallel, implying that a no-interaction model is appropriate for studying differences in the general knowledge scores between the income groups for the Spring semester, controlling for the scores in the Fall.

Results from ANOVA: p-value < 0.001, F-statistic : 59.91
Given that the p-value < 0.05, the conclusion is that there are significant differences in the kindergartners' mean spring general knowledge scores among the income groups while adjusting the effect of their fall general knowledge scores.
Post-hoc Test - Based on the results from the Tukey-HSD test, for all pairwise combinations of income groups, since p-values < 0.05, we conclude there are statistically significant differences in kindergartners' mean spring general knowledge scores among all the three income groups.

Testing Assumptions

- Assumption 1: Residuals are normally distributed
  - The Shapiro-Wilk Test concludes that the residuals are not normally distributed since the p-value < 0.001, and therefore, we reject the null hypothesis.
- Assumption 2: Homogeneity of Variances
  - Given the data did not pass the normality assumption check, I used Levene's Test, which resulted in a p-value < 0.001, and since p-value < 0.05, we reject the null hypothesis and conclude that the variances are not homogeneous.

One-Way ANCOVA - II

RQ: Does the income group have an impact on the kindergartner's spring reading scores, having controlled for the student's fall reading scores?

Null hypothesis: Means of all kindergartners' spring reading scores based on each income group are equal after controlling the effect of kindergartners' fall reading score i.e. adjusted means are equal
Alternative hypothesis: At least, one kindergartner's spring reading score for an income group is different from other income groups after controlling the effect of kindergartners' fall reading score i.e. adjusted means are not equal

Interaction Plot -

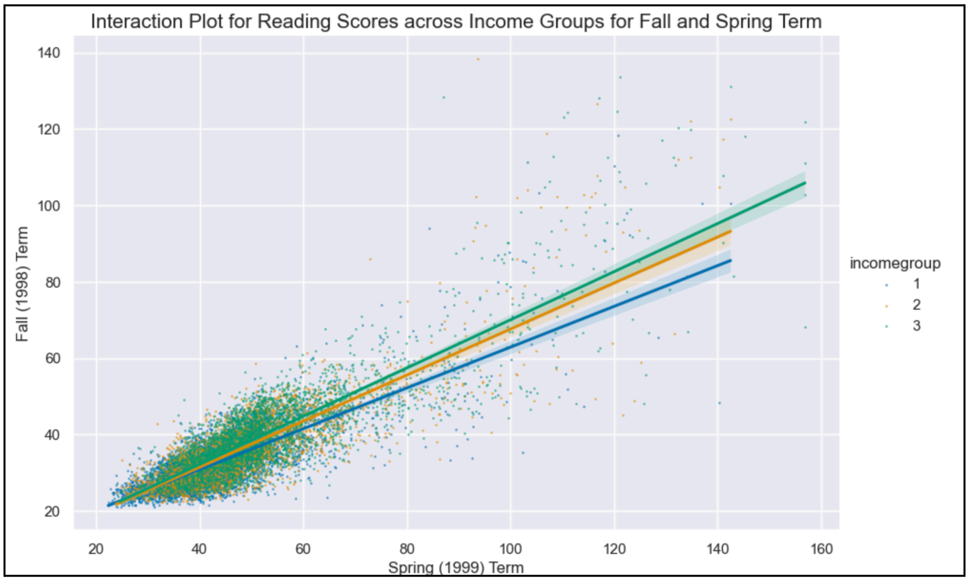


Figure 4. Interaction Plot

As seen in Figure 4, the plot lines are roughly parallel, implying that a no-interaction model is appropriate for studying differences in the reading scores between the income groups for the Spring semester, controlling for the scores in the Fall.

Results from ANOVA: p-value = 0.017, F-statistic : 4.05
Given that the p-value < 0.05, the conclusion is that there are significant differences in the kindergartners' mean spring reading scores among the income groups while adjusting the effect of their fall reading scores.
Post-hoc Test - Based on the results from the Tukey-HSD test, for all pairwise combinations of income groups, since p-values < 0.05, we conclude there are statistically significant differences in kindergartners' mean spring reading scores among all the three income groups.

Testing Assumptions

- Assumption 1: Residuals are normally distributed

  - The Shapiro-Wilk Test concludes that the residuals are not normally distributed since the p-value < 0.001, and therefore, we reject the null hypothesis.
- Assumption 2: Homogeneity of Variances
  - Given the data did not pass the normality assumption check, I used Levene's Test, which resulted in a p-value < 0.001, and since p-value < 0.05, we reject the null hypothesis and conclude the variances are not homogeneous.

One-Way ANCOVA - III

RQ: Does the income group have an impact on the kindergartner's spring math scores, having controlled for the student's fall math scores?

Null hypothesis: Means of all kindergartners' spring math scores based on each income group are equal after controlling the effect of kindergartners' fall math scores i.e. adjusted means are equal
Alternative hypothesis: At least, one kindergartner's spring math score for an income group is different from other income groups after controlling the effect of kindergartners' fall math score i.e. adjusted means are not equal

Interaction Plot -

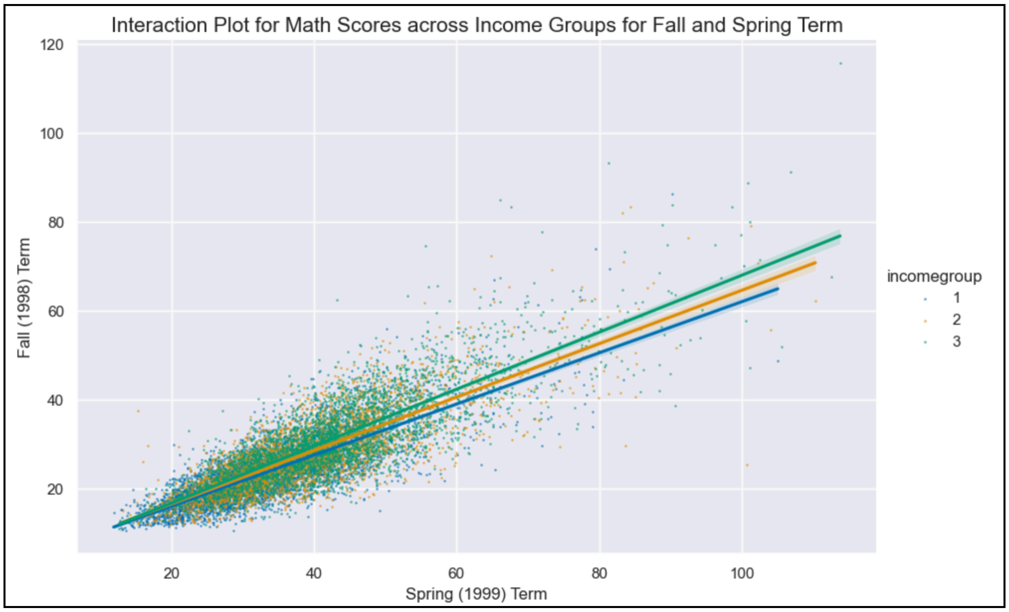


Figure 5. Interaction Plot

As seen in Figure 5, the plot lines are roughly parallel, implying that a no-interaction model is appropriate for studying differences in the math scores between the income groups for the Spring semester, controlling for the scores in the Fall.

Results from ANOVA: p-value < 0.001, F-statistic : 18.52

Given that the p-value < 0.05, the conclusion is that there are significant differences in the kindergartners' mean spring math scores among the income groups while adjusting the effect of their fall math scores.
Post-hoc Test - Based on the results from the Tukey-HSD test, or all pairwise combinations of income groups, since p-values < 0.05, we conclude there are statistically significant differences in kindergartners' mean spring math scores among all the three income groups.

Testing Assumptions

- Assumption 1: Residuals are normally distributed
  - The Shapiro-Wilk Test concludes that the residuals are not normally distributed since the p-value < 0.001, and therefore, we reject the null hypothesis.
- Assumption 2: Homogeneity of Variances
  - Given the data did not pass the normality assumption check, I used Levene's Test, which resulted in a p-value < 0.001, and since p-value < 0.05, we reject the null hypothesis and conclude the variances are not homogeneous.

Conclusion -

The results from this analysis confirm significant differences in the spring semester scores of kindergarten students of different income groups, across all three subjects - general knowledge, reading and math, while controlling for their respective fall semester scores. This suggests that their student's economic background seems to be a significant predictor role in the evaluation scores of each of the subjects. It would be interesting to extend this study to include other socio-economic factors to understand their impact on these scores, and compare these results to the latest scores to study for prominent historical trends in the data.

References -

- Python Graph Gallery - https://python-graph-gallery.com
- Stack Overflow - https://stackoverflow.com
- Seaborn Documentation - https://seaborn.pydata.org/archive/0.11/tutorial/categorical.html
- Pandas DF Documentation - https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html
- How to Perform ANOVA in Python, Renesh Bendre, https://www.reneshbedre.com/blog/ancova.html
- Displaying Multiple DataFrames Side By Side in Jupyter Lab/Notebook, Liu Zuo Lin, https://python.plainenglish.io/displaying-multiple-dataframes-side-by-side-in-jupyter-lab-notebook-9a4649a4940
- Seltman, H. J. (2018). Experimental Design and Analysis, Department of Statistics at Carnegie Mellon