Correlation Between Household Income and Student Academic Performance

INF2178

Zhuoying Li

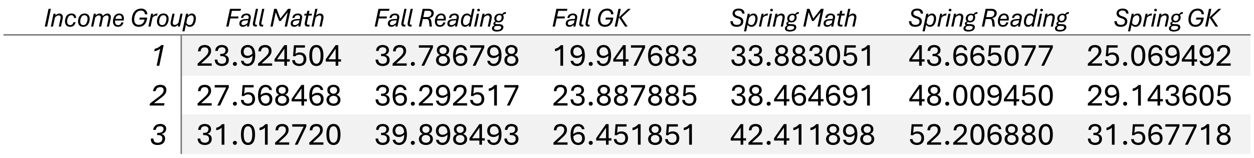
1004021202

Research Questions

1. **Research Question 1:** Is there a significant difference in student's academic performance (math and reading scores) between different income groups?
2. **Research Question 2:** Is there a significant difference in student's academic performance between fall and winter across different income groups? While using general knowledge as a baseline score (covariate).

Exploratory Data Analysis

We first do some basic summary statistics for various income groups, we are primarily interested in the mean, so following table includes mean scores for math, reading, and general knowledge across 3 different income groups.



Income Group	*Fall Math*	*Fall Reading*	*Fall GK*	*Spring Math*	*Spring Reading*	*Spring GK*
1	23.924504	32.786798	19.947683	33.883051	43.665077	25.069492
2	27.568468	36.292517	23.887885	38.464691	48.009450	29.143605
3	31.012720	39.898493	26.451851	42.411898	52.206880	31.567718

Figure 1: Mean scores for different income group

From the table, we can see a general trend for higher income groups to have higher scores overall. We further check this by drawing boxplot for different income groups, which confirms our observation that the overall score is higher when the income group is higher.

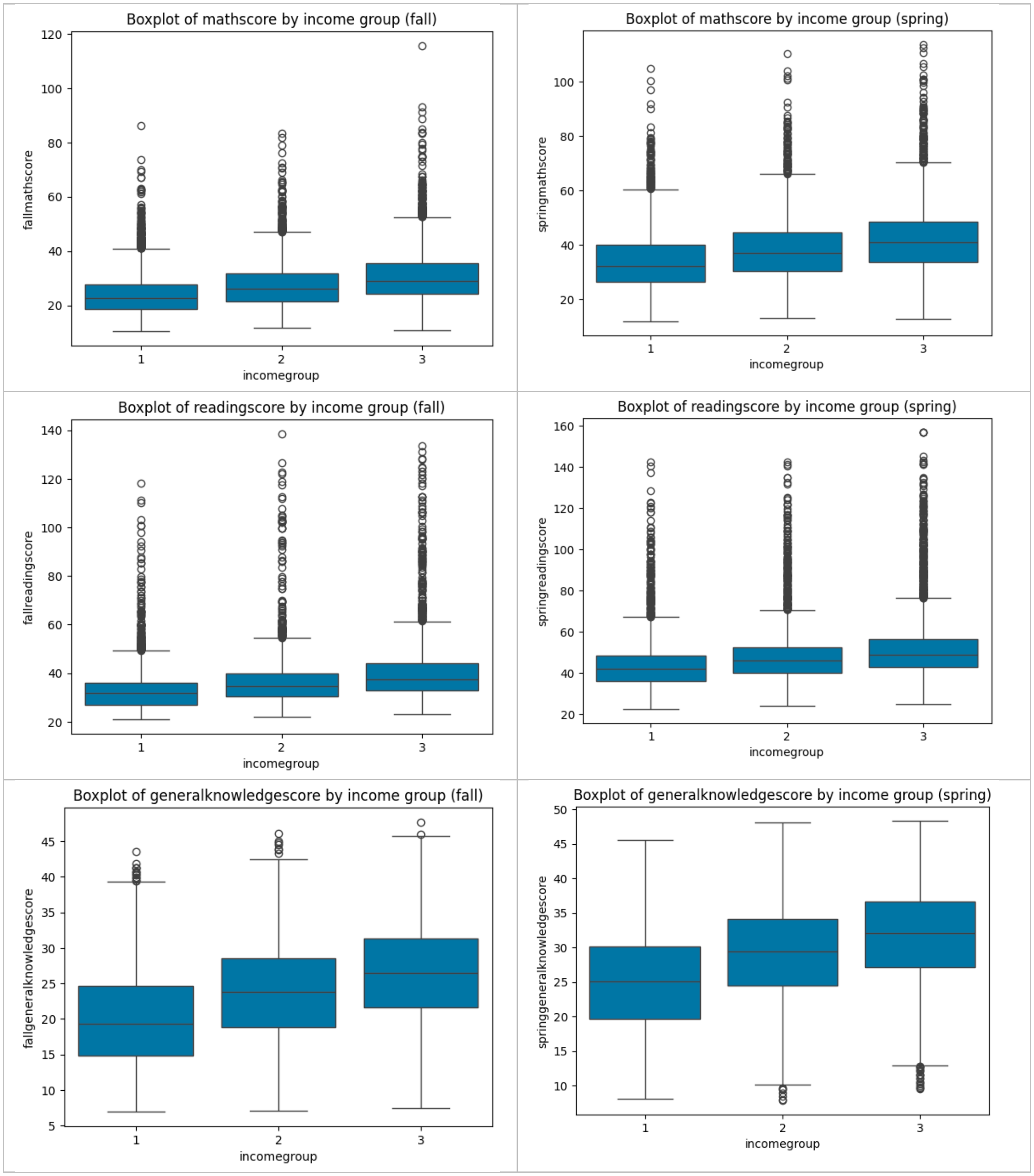


Figure 2: Boxplot for different scores by income group

One-Way ANCOVA

We utilize one-way ANCOVA as our analysis tool, this is because we have a covariate variable: general knowledge scores. Utilizing ANCOVA, we can get a better understanding of how exactly our independent variables affect our dependent variables without the effect of our covariate. This is important as we may not have a good, randomized selection of samples.

For the ANCOVA analysis, we have the following hypothesis.

1. H_0 – There is no difference in student's academic scores between different income groups.
2. H_1 – There is a significant difference in student's academic scores between different income groups.

We conducted many one-way ANCOVA tests to analyze how income groups may impact student's math and reading scores, with general knowledge score as the covariate. These models are constructed using the formula that's similar to: $fallmathscore \sim fallgeneralknowledgescore\ +\ incomegroup$. With the exact pairings listed below:

Dependent Variable	*Covariate*	*Independent Variable*
fallmathscore	fallgeneralknowledgescore	incomegroup
fallreadingscore	fallgeneralknowledgescore	incomegroup
springmathscore	springgeneralknowledgescore	incomegroup
springreadingscore	springgeneralknowledgescore	incomegroup

Figure 3: Dependent, covariate and independent variable tuples

After building all the ANCOVA models, we found that income group variable significantly affected reading and math scores for both fall and spring semesters. Giving us P-values <0.001 for all pairs. Therefore, **we reject our null hypothesis**, suggesting that there is a significant impact on academic scores given different income groups.

We will omit the ANCOVA table here as all P-values are <0.001, as is shown as 0.000 when using StatsModels library. Interested readers can check the attached Jupyter notebook for detailed results.

Checking Assumptions

For ANCOVA, there are four assumptions in total we must check.

Assumption 1 - Linearity of the relationship between the dependent variable and the covariant

The first assumption is verified by first plotting a scatter plot to visually inspect the relationships.

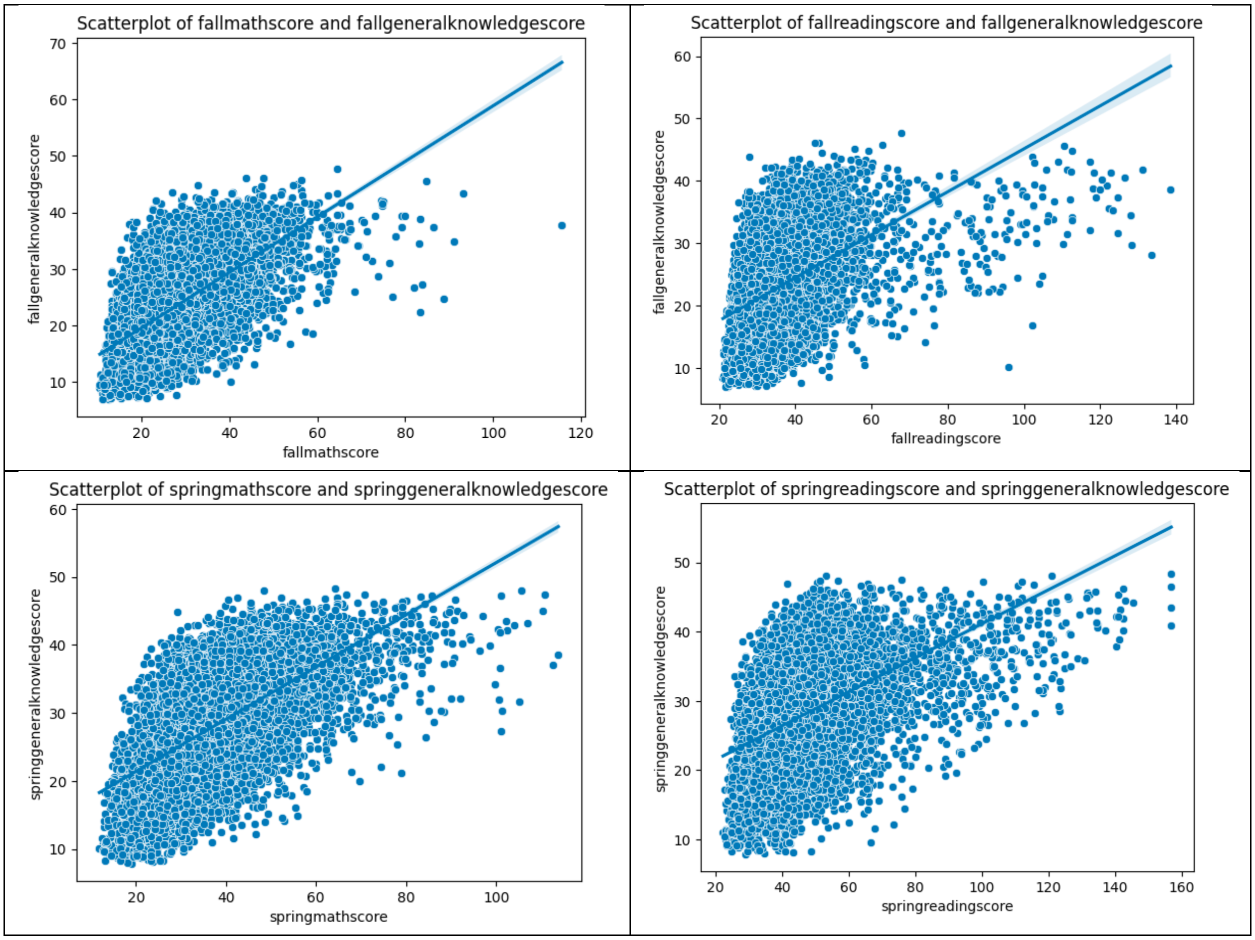


Figure 4: Scatter plot for dependent variable and covariate pairs

We can see the overall linear relationships between our variables, we also verified this by creating a linear regression model, which yields P < 0.001 for all pairs, suggesting strong linear relationships.

Assumption 2 - Homogeneity of regression slopes

This assumption can be checked by looking at the slopes from our assumption 1 testing, which we can see is mostly parallel. This can further be verified by doing a two-way ANOVA. We obtain the following ANOVA tables for each pair of dependent variable of our covariate

	P-Value	*df*
incomegroup	0.000000	2
Fall GK score	0.000000	1
Fall GK score: income	0.000028	2

Figure 5: ANOVA table for fall math score

	P-Value	*df*
incomegroup	4.014876e-257	2

Fall GK score	0.000000e+00	1
Fall GK score: income	6.178071e-05	2

Figure 6: ANOVA table for fall reading score

	P-Value	*df*
incomegroup	0.000000	2
Fall GK score	0.000000	1
Fall GK score: income	0.001266	2

Figure 7: ANOVA table for spring math score

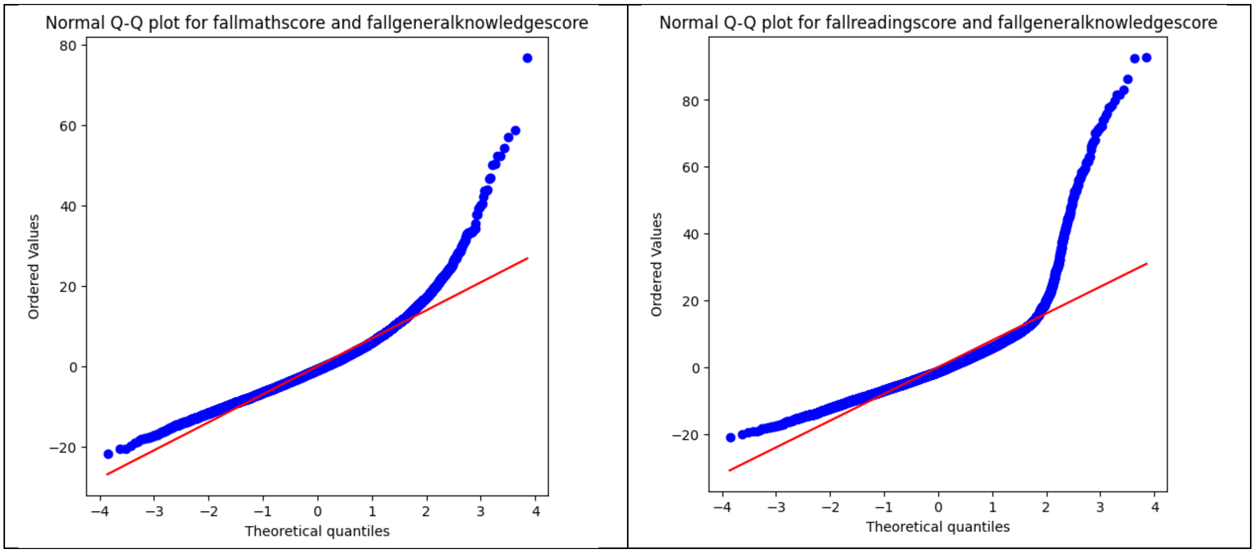
	P-Value	*df*
incomegroup	2.065001e-194	2
Fall GK score	0.000000e+00	1
Fall GK score: income	2.101536e-02	2

Figure 8: ANOVA table for spring reading score

All P-values are < 0.001, indicating the correctness of our assumption.

Assumption 3 – Normality of Residuals

Plotting the Q-Q plot for all our ANCOVA models, we found that none of the satisfy the normality requirement.



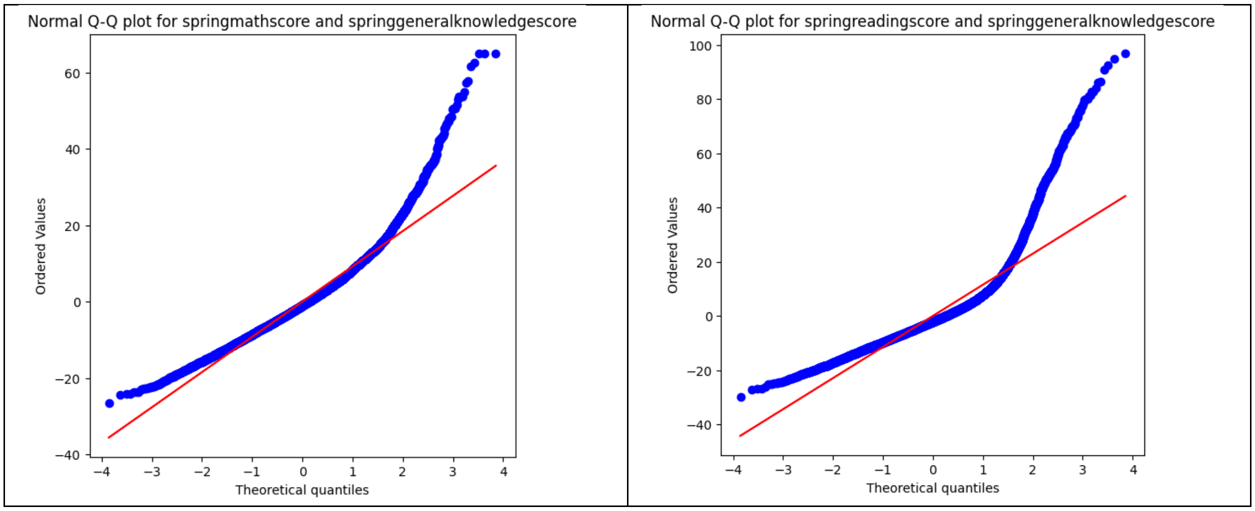


Figure 9: Q-Q plot for ANCOVA model residuals

A Shapiro-Wilk test is also conducted, that verifies our finding from the Q-Q plot, all tests have P-value < 0.001, indicating non-normal distributions.

Assumption 4 – Homogeneity of Variance

A Levene's test is conducted on all residuals from our ANCOVA models, which produced a statistically significant result of P-value < 0.001, indicating non-homogeneity of variances.

Conclusion

In conclusion, we found that there is a significant different between student's academic scores between fall and winter across various income groups, regardless of if we consider of our covariate or not. However, because some of our assumptions are not met for ANCOVA, further investigation is needed to verify our findings.