INF2178_A3

Student Name: Nianchuer Liu
Student Number: 1010332454
Email:niancher.liu@mail.utoronto.ca

1. Research Questions

After loading the dataset, it can be known that the dataset contains scores for reading, math, and general knowledge taken in both fall and spring, along with information about household income. So, here is my research questions based on the data:

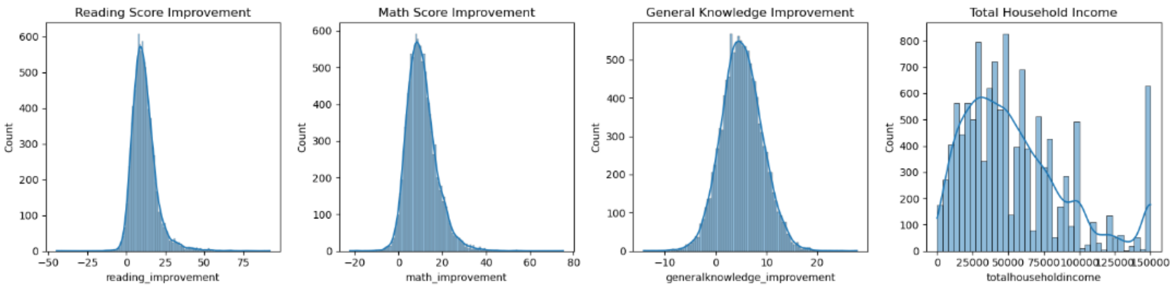
1. **Does household income affect the improvement in academic scores from fall to spring?** I can analyze the difference in scores from fall to spring for each subject and see if there's a significant difference across income groups.
2. **Is there a subject in which the influence of income on score improvement is more pronounced?** Comparing the effects of income on reading, math, and general knowledge score improvements could highlight disparities in academic areas.
3. **How does income group correlate with initial fall scores across subjects?** This can help people understand if income has an apparent impact on starting academic performance.

2. Exploratory Data Analysis (EDA)

For the EDA, I will:

- Check the distribution of scores and income through summary statistics and histograms.
- Examine any potential outliers.
- Calculate the score improvements from fall to spring for each subject.

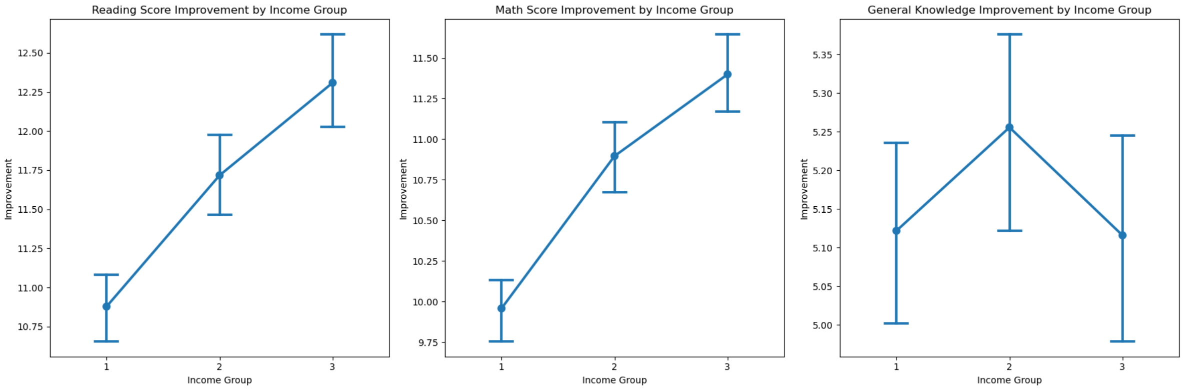
2.1 Interaction Plots and Summary Statistics



Based on the graph, the improvements in reading, math, and general knowledge scores from fall to spring show positive mean values, indicating an overall increase. Specifically, the average improvements are approximately 11.56 for reading, 10.67 for math, and 5.16 for general knowledge, with reading showing the highest mean improvement. Moreover, The total household income has a wide distribution, with a mean of approximately 54,317 and a standard deviation of 36,639, indicating significant variability among the households in the dataset. The income range extends from as low as 1 to as high as $150,000. These findings suggest potential areas of investigation, particularly regarding how these improvements correlate with income groups and if income has a different impact on the improvement of different academic subjects.

2.2. Interaction Plots Creation

I am going to make interaction plots to show how income group affects learning gains in reading, math, and general knowledge. These graphs will help us see if there is a clear pattern in how income levels affect how well students do in these topics.



Based on the plots, there seems to be a trend that reading scores might improve more significantly for students from higher-income groups. However, the differences between income groups are not very clear. Also, like with reading, there's a chance that people with better incomes may see bigger gains in math scores, though the trend seems a bit clearer than with reading. There is also a pattern in the growth in general knowledge scores across income groups. Improvements may be bigger in higher income groups, but the differences don't seem to be as clear as they are in reading and math.

3. Check Assumptions:

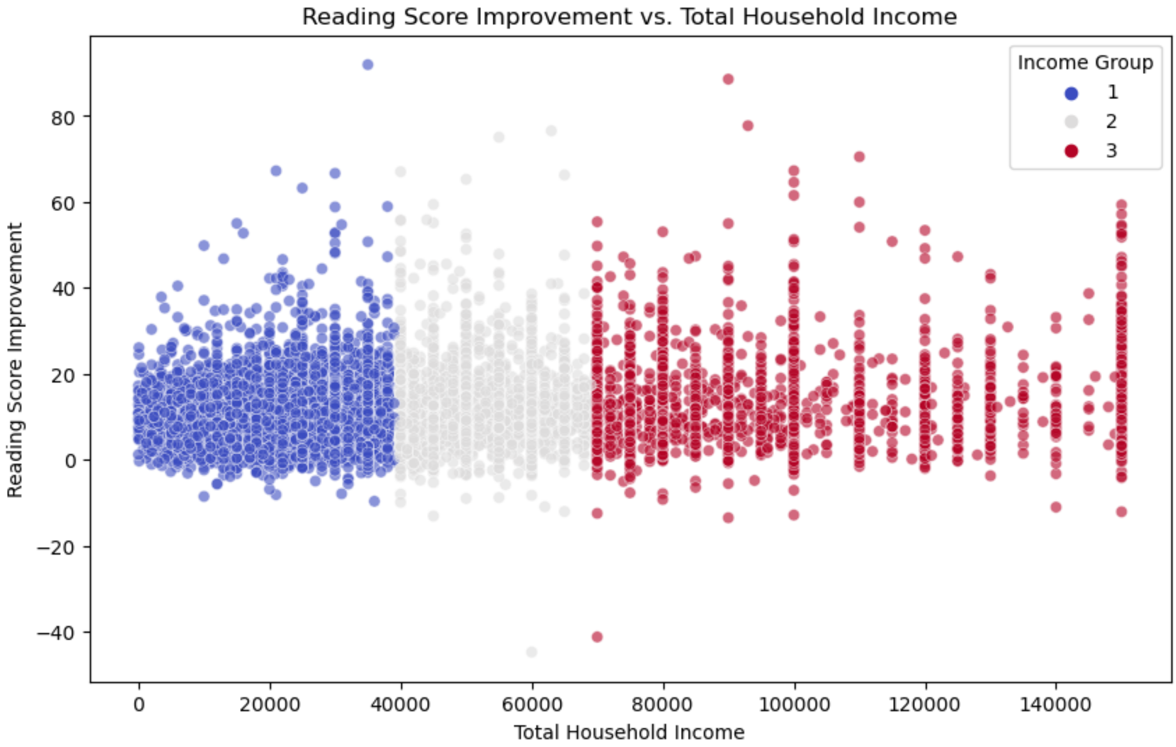
The results from checking the assumptions for conducting an ANCOVA on the reading score improvements are :

1. The p-values for the Shapiro-Wilk test of normality are very low (significantly less than 0.05) for each income group. This means that the null hypothesis of normality within each group is not true. This means that the reading improvement scores are not usually spread out among the income groups.
2. The p-value for Levene's test is also very low (2.794929503613517e-09), which means that the differences in reading score gains between income groups are not the same. This goes against the idea that variances should be homogeneous.
3. The p-value for the test of homogeneity of regression slopes (2.0561997106146038e-05) shows that there is a different relationship between total family income (as a covariate) and reading score improvement across income groups. This goes against the assumption that regression slopes should be the same for all income groups.

Based on these results, it's clear that the conditions needed to do an ANCOVA on reading score increases are not fully met. Specifically, ANCOVA data may not be valid if the normality and homogeneity of variances assumptions are broken. The important interaction effect shows that the covariate (total family income) does not have the same effect on the dependent variable (reading improvement) across groups. This means that a simple ANCOVA might not be the best method without making some changes.

4. Perform One-way ANCOVA

The ANCOVA results for reading score improvements show that total household income has a significant positive effect (F=17.413538, p=0.000030), while income group as a categorical variable does not have a significant effect on reading score improvements (F=1.775743, p=0.169402). This shows that the different income groups don't really explain the differences in reading score gains. Instead, the actual household income is what matters. The sum of squares for income groups and family income shows how much of the variation can be explained by these two factors. The continuous measure of income (1130.725900) explains a lot more of the variation than the variation between income groups (230.611199). The difference in reading skills across the collection that can't be explained is shown by the residual sum of squares (774594.435176). This study shows that family income levels are more important than income groups when it comes to improving academic performance.



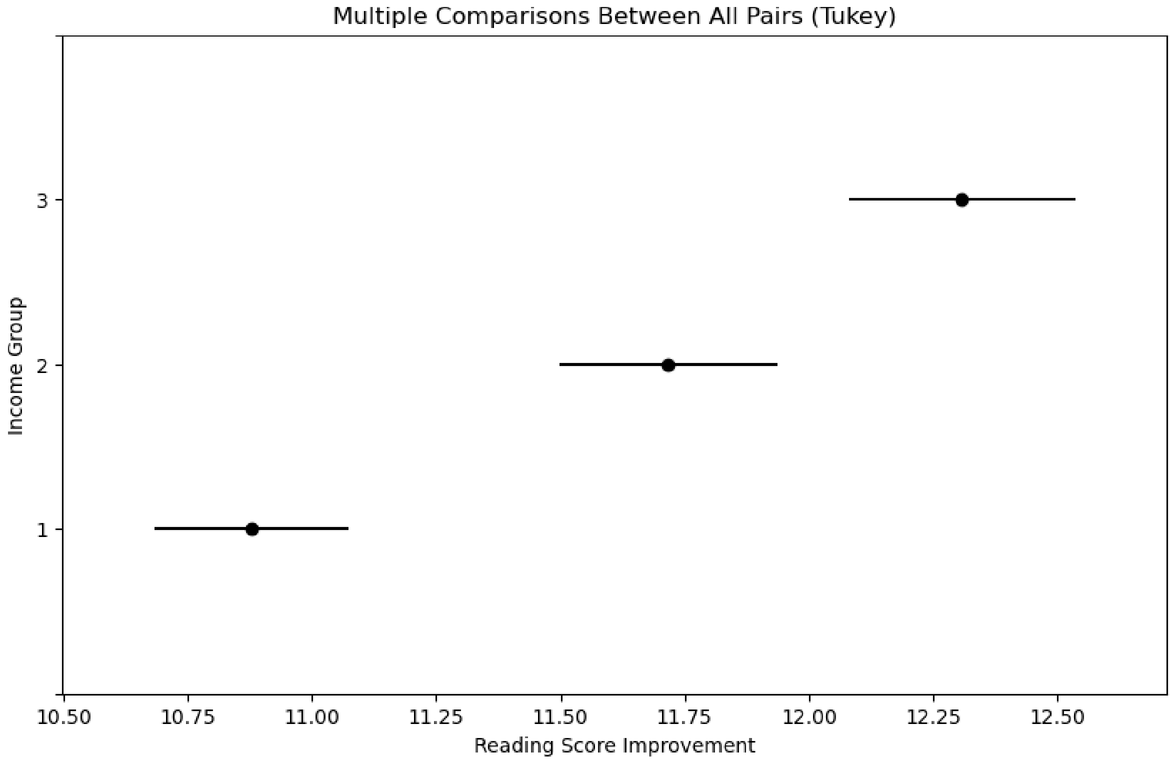
The graph backs up the ANCOVA finding that total household income has a big impact on reading score improvement. However, it also shows how complicated this relationship is and suggests that there are other factors at play that aren't just captured by income group labels.

5. post hoc tests

5.1 Tukey HSD Results:

- **Group 1 vs. Group 2**: With a p-value of 0.0, which is less than the alpha level of 0.05, the mean difference is 0.8387. The null hypothesis is not true because the 95% confidence range for the mean difference does not include zero (0.4246 to 1.2527). This shows that there was a statistically significant difference in the amount of progress made in reading between income groups 1 and 2, with group 2 making more progress overall.
- **Group 1 vs. Group 3**: There is a statistically significant difference between these groups, as shown by the larger mean difference of 1.4301 and the p-value of 0.0. There is no zero in the confidence range (1.0079 to 1.8523); therefore, the null hypothesis is not true. Reading scores have improved a lot more in Group 3 than in Group 1.
- **Group 2 vs. Group 3**: The mean difference is 0.5915, with a p-value of 0.0053, which is also statistically significant. The confidence interval (0.1458 to 1.0371) does not include zero. Therefore, group 3 shows significantly higher reading score improvements than group 2, but the difference is smaller than that between groups 1 and 3

The graph and table below show the outcomes of Tukey's Honestly Significant Difference (HSD) test, which compares the changes in reading scores between groups of people with different incomes, using a Family-Wise Error Rate (FWER) of 0.05 as the significance level.



There are confidence intervals for the mean differences in reading score increases between the income groups shown in the graph. We can see the overlaps because the intervals are plotted along the axis of reading score increase. The fact that the lines don't meet with the line of no difference (which would be at zero on the horizontal axis) shows that the differences between the groups are important. After using Tukey's HSD test for post hoc analysis, it was found that reading score increases are significantly different between all pairs of income groups. More specifically, it seems that reading score gains go up as the income group goes up. This shows that groups with higher incomes see bigger gains in reading scores than groups with lower incomes. This is in line with what you might expect given how socio-economic factors affect educational outcomes.