Name: Yun Zhou (1009709442)
Assignment 3: INF2178 Technical Assignment 3

Explore Kindergarten students scores

Introduction

The dataset is related to Kindergarten student academic performance and household income, which includes both academic performance and household income data. Based on preliminary analysis, the data tells us story of how academic progress over time in relation to household income. To be specific, in the academic performance dataset, it shows an academic growth or change across the school year from fall to spring semester based on academic scores in reading, math and general knowledge.

With the household income data as continuous variable, alongside a categorical income group, the data helps us understand how household income might have influence on academic performance and progress.

1. Data Cleaning and preparation

The dataset has a total of 9 columns with 11,933 rows, presenting information individual student academic performance in reading, math and general knowledge over fall and spring, alongside household income data.

From initial observations, the "incomegroup" dataset divides households income into 3 different income brackets, which can be used for data analysis in examining the relationship in income and educational performance. There is wide range income data which shows a diverse set of economic situation among students. Also, there is improvement in educational performance scores from fall to spring, it shows the either the effect of schooling over year or impact of other variables. There is two income columns ("incomecolumn" and "incomecolumn" by thousands) with the same data may require further clean-up during further analysis, as to reduce redundancy.

To delve deeper, below are the research questions that this dataset can help answer:

> **Research Question 1:**
> Are there any statistically differences in academic progress for reading and math scores between income groups, by controlling the general knowledge levels?
> **Research Question 2:** Does different income groups affect the change in reading and math scores, by controlling for general knowledge?
> **Research Question 3:** Hoes does student's initial general knowledge has impact on their improvements in specific areas of subjects (ex. Reading) over the academic year?
> **Research Question 4:** How does household income affect the academic progress of students over academic year, by controlling the initial knowledge levels?
> **Research Question 5:** How effective of education in improving student scores across different areas of subjects over an academic year from fall to spring?

The goal is to understand the relationships between household income, initial knowledge levels and academic progress over the academic year during fall and spring.

2. Exploratory data analysis

First is to examine the data, including checking for missing values and understanding the distribution of income groups, in order to identify immediate cleaning needs.

Per the **summary statistics** for all numerical columns (student scores and income data), we can refer that there is overall improvement in scores across the academic year as mean scores has increased from fall to spring. Ex. The mean reading score has increased from 35.95 (fall) to 47.51(spring). The math score has increased from 27.12 (fall) to 37.80 (spring). General knowledge has increased from 23.07 (fall) to 28.23 (spring).

The “incomegroup” data with 3 distinct categories from 1 to 3, the mean is 1.89 which is close to 2, suggesting a relatively **even distribution** of the data across these groups, but need the actual frequency distribution to analysis further.

We learnt that there are **not** any **missing values** or absent entries and dataset is complete. Per income distribution, group 1 accounts for 39.63% of the data, group 2 comprise for 31.22% and group 3 made up of 29.1% Therefore, Group 1 – lowest income has the largest proportion of dataset. It gives us a clear picture of the sample's economic diversity based on distribution.

To calculate the change in score, first is to create two new columns in df: “reading_change” and “math_change”, this to represent the **change in reading** and **math scores** from fall to spring, calculated by spring math/reading score subtract fall math/reading score for each student.

Two boxplots are created to visualize the dataset. In the 1st boxplot, it represents changes in reading scores across different income groups in academic year. The median change is above 0, therefore students in all income groups has improved their reading scores on an average. The box represents the interquartile range (IQR), which is the middle 50% of the change in scores which shows a range of change that are consistent across all income groups. Therefore, there is a similar range of improvement in reading scores across all 3 income groups. The whisker represents the lines extends both upwards and downwards, represent 1.5 * IQR (interquartile range) above and below box. From the boxplot, the length whisker in income group 3 are longer than group1 and 2, therefore a more variability in reading score change for Group 3. Lastly, there are outliners in all 3 groups.

In **2nd boxplot**, it represents change in math score across 3 different income group from fall to spring. Similar to reading score changes above, the median change in math scores for all income groups is positive, which indicates academic improvement in spring. The IQR in group 2&3 is slightly greater than for income group 1, this shows income Groups 2 and 3 had a broader spread of score changes than in Group 1. Also, Group 3 shows longer whiskers as well, it similarly indicates a greater variability in the math score changes within the group as compared to Groups 1 and 2. However, group 3 have a slightly lower range of outliers compared to groups 1 & 2.

Overall, the boxplot shows the scores have improved over the academic year for all 3 income groups for both reading and math, also, the change in reading and math scores from fall to spring indicate some differences among 3 income groups. It seems like students in higher income group (Group 3) tend to have greater positive changes in scores. Further data analysis (ANCOVA) would need to test if these differences are statistically significant.

3. **Data analysis**

Research Question #1: Are there any statistically differences in academic progress for reading and math scores between income groups, by controlling the general knowledge levels?

First is to **test the assumptions** necessary to conduct Analysis of Covariance (ANCOVA), the test assumptions are "reading change" and "math change" as dependent variables, controlling the "fall general knowledge score", the dataset is being divided by 3 "income group".

As a note, the p-value tested result might not be accurate as natural due to the small dataset, the p-value might change if testing on large dataset.

1) <u>For change in reading scores from fall to spring (reading change</u>): **Levene's test** p-value is extremely small (**2.79e-09**) and test statistic is **19.728**. The low p-value means the test result is statistic significant, therefore, the homogeneity of variance assumption is rejected, which means the variances in reading change scores aren't equal across the 3 income groups.
   Shaprio-Wilk - p-value is **0.0** (less than alpha 0.5) and test statistic is **0.899**, the result also is significant. Therefore, rejecting normality assumption for residuals and indicating the residuals are not normally distributed.
2) <u>For change in math scores over fall to spring (math change):</u>
   Levene Statistic is **22.215** and p-value is **2.34e-10**, which indicates statistic significant and therefore reject the null hypothesis of equal variances across groups. Therefore, non-homogeneity of variances for math change.
   The Shapiro-Wilk test statistic is **0.966** and p-value is **0.0**. similar to above, this suggests significant therefore residuals do not follow the normal distribution.

Overall, the test result indicates a statistic significant for change in reading and math scores, therefore there are challenges with the dataset's compliance with ANCOVA assumptions: variance homogeneity and normality of residuals are not met for both dependent variables. (This might affect the validity of the ANCOVA results.) We will perform ANCOVA though noting the violations.

Research Question 2: Does different income groups affect the change in reading and math scores, by controlling for general knowledge?

Then is to perform two separate AVNOCA analyses for reading and math score changes, by controlling for the general knowledge score for fall term.

<u>For reading change:</u>

The income group is not a significant predictor of the change in reading scores, as F statistic is **2.251** and p-value is around **0.105**. As p-value is greater than 0.05, the result is not significant and therefore the income group is not significantly affected reading change, while controlling for general knowledge score.

<u>For math change:</u>

The income group is not a significant predictor of the change in math scores, due to F statistic is **0.624** and p-value of **0.536** (greater than 0.05), therefore not significant by controlling general knowledge score.

Therefore, the level of income does not influence reading and math scores, by controlling general knowledge variable. Although there are visible differences in score changes across the three income groups, per ANCOVA result did not find these differences to be statistically significant once controlling for general knowledge scores.

To visualize the output, we create the **interaction plots** to display how the relationship between income groups and academic changes.

<u>Interaction plot for reading (score) change:</u>
It shows a positive relationship between income group and reading change. While the income group number increases so is income levels becomes greater, the reading change score also increases. The result shows student from higher income groups have a greater increase in reading skills than student from lower income groups.

<u>Interaction plot in math (score) change:</u>
Same as above plot, the plot shows upward trend and therefore also a positive relationship between income group and math score change. Student with higher income groups is having greater increase in math skills. The result again student from higher income groups have larger improvement in math score skills compared with those from lower income group.

Overall, the interaction plot shows income group level is positively associated with academic skills improvement in both reading score and math score change.

In conclusion, by combing both statistical analysis and plot trends, while the interaction plots suggest there may be a positive relationship between income group and changes in reading and math skills, the statistical analysis indicates that these differences are not significant when accounting for general knowledge.

Research Question 3: Hoes does student's initial general knowledge has impact on their improvements in specific areas of subjects (ex. Reading) over the academic year?

Per ANCOVA results:

For reading change:

The fall general knowledge score is a significant covariate as F statistic is **220.110** and p-value is **2.35e-49** (less than alpha significant 0.05), therefore strong effect on reading change.

For math change:
The fall general knowledge shows F statistic is **501.084** and p-value of 9.43e-109 ($p<0.01$), the smaller than alpha means test significant and so there is significant effect on math change.

Overall, the general knowledge score has a significant impact on both reading and math score changes income among groups, we can refer that student's original knowledge does strongly impact their academic progress.

Research Question 4: How does household income affect the academic progress of students over academic year, by controlling the initial knowledge levels?

Based on the above ANCOVA results, household income does not appear to have a significant effect on academic progress over the academic year when controlling for initial knowledge levels. Since there are greater p-values ($p>0.05$) associated with the income group variable for both reading and math changes:

According to change in reading score, the income group had a p-value of approximately 0.105, which is higher than the alpha 0.05.

For change in math score, the p-value was even higher, at 0.536.

Therefore, the p-values means that the change in scores (academic progress) measured over the academic year are not statistically significant with respect to household income when controlling initial knowledge levels.

Research Question 5: How effective of education in improving student scores across different areas of subjects over an academic year from fall to spring?

Based on prior ANCOVA results, we learnt general knowledge has a significant impact on both reading and math changes. This implies that education, as it enhances the general knowledge, can be quite effective in improving student scores.

Further, per reading change, the general knowledge score had an F-statistic of 220.110 For math change, the general knowledge score had an F-statistic of 501.084. The F statistic were large for general knowledge in both reading and math models indicates strong link between students' general knowledge and their academic progress in both reading and math.

In summary, general knowledge plays a crucial role in academic improvement in reading, it has significant impact on education's effectiveness in improving on student scores for reading and math.