Exploring Early Child Data in Toronto

Introduction

This study examines a subset of data from an early childhood longitudinal study conducted between 1998 and 1999. It seeks to analyze the connection between socioeconomic factors and educational outcomes, specifically focusing on reading, math, and general knowledge scores of Kindergarten students over several months. This data, gathered from fall 1998 and spring 1999, offers a unique opportunity to investigate the educational trajectories of young learners as they progress through an academic year, with an emphasis on how these trajectories differ across income groups.

The dataset, named "INF2178_A3_data.csv," contains continuous variables that measure academic performance in reading, math, and general knowledge, as well as income category, the sole categorical variable. By examining Kindergarten scores, this study aims to use general knowledge scores as a baseline for evaluating how students' reading and math scores evolve over time in relation to their income group.

Research Question: How does household income influence the development of reading and math skills in young learners, and what does this imply for our understanding and intervention strategies regarding educational disparities?

The findings of this research will contribute to a broader discourse on educational equity and may inform policy decisions that strive to offer equal learning opportunities to students from all socioeconomic backgrounds.

Data Exploration

To begin our analysis, I first examined the summary statistics on the dataset, which include information on income groups, fall standardized testing, and spring standardized testing.

Variable	Min	Mean	Max	25th	Median	75th	IQR
All Reading Score	21.01	35.95	138.51	29.34	34.06	39.89	10.55
Fall Math Score	10.51	27.13	115.65	20.68	25.68	31.59	10.91
Fall General Knowledge Score	6.98	23.07	47.69	17.39	22.95	28.3	10.91
Spring Reading Score	22.35	47.51	156.85	38.95	45.32	51.77	12.82
Spring Math Score	11.9	37.8	113.8	29.27	36.41	44.22	14.95
Spring General Knowledge Score	7.86	28.24	48.34	22.8	28.58	33.78	10.98
Total Household Income	1.0	54317.2	150000	27000	47000	72000	45000
Income in Thousands	0.0	54.32	150.0	27.0	47.0	72.0	45.0
Income Group	1	1.9	3	1.0	2.0	3.0	2.0

Data Cleaning and Wrangling

The raw dataset comprises 9 columns with 11,933 entries. The dataset encapsulates student performance across various assessments over two semesters, along with socioeconomic indicators. My preliminary examination suggests that the dataset is well-structured and requires minimal cleaning for the analysis I intended to perform.

```
Data Overview
-------------------------------------------|--------------------------------|------------------
 - The dataset has a RangeIndex of 11,933 entries, ranging from 0 to 11,932.
 - There are 9 columns, consisting of a mix of float64, int64, and object data types.

Column Descriptions
-------------------------------------------|--------------------------------|------------------
 - **fallreadingscore**: Reading score for the fall semester.
 - **fallmathscore**: Math score for the fall semester.
 - **fallgeneralknowledgescore**: General knowledge score for the fall semester.
 - **springreadingscore**: Reading score for the spring semester.
 - **springmathscore**: Math score for the spring semester.
 - **springgeneralknowledgescore**: General knowledge score for the spring semester.
 - **totalhouseholdincome**: Total household income for the student's family.
 - **incomeinthousands**: Household income in thousands.
 - **incomegroup**: Categorization of income into groups for analysis.
 - **generalknowledge**: Combined general knowledge score, aggregating performances across semesters.

Missing Values
-------------------------------------------|--------------------------------|------------------
 - The dataset contains no missing values, indicating a complete dataset ready for analysis.
```

Feature Engineering:
In my endeavor to extract more nuanced insights from the dataset, I engaged in feature engineering to enrich my understanding of educational outcomes across different socioeconomic backgrounds. Specifically, I developed new features to examine how various factors influence students' performance and to identify potential areas for targeted interventions.

General Knowledge:
A feature calculating the average between springgeneralknowledgescore and fallgeneralknowledgescore, aimed at quantifying individual student progress over the whole academic year.

Educational Improvement:
A metric derived from the difference between spring and fall scores across reading, math, and general knowledge, providing a direct measure of academic progress.

ANCOVA Resuls

The ANCOVA sought to assess the relative academic gains in reading, math, and general knowledge by considering the interaction of income groups with changes in scores from fall to spring. The goal was to quantify not only the educational progress but also to gauge how income levels may affect this progress.

1. Sample sizes (n) for each income group:
   a. Income Group 1: n = 4729
   b. Income Group 2: n = 3726
   c. Income Group 3: n = 3478
2. Mean general knowledge scores for each income group in the spring:

   a. Income Group 1: Mean = 22.51
   b. Income Group 2: Mean = 26.52
   c. Income Group 3: Mean = 29.01

3. Standard deviations (std) of general knowledge scores for each income group in the spring:
   a. Income Group 1: Std = 6.68
   b. Income Group 2: Std = 6.61
   c. Income Group 3: Std = 6.72
4. Regression analysis results (from OLS regression summary) showing the effect of income group and fall general knowledge score on spring general knowledge score:
   a. Intercept: 8.0303
   b. Effect of being in Income Group 2: +0.7084
   c. Effect of being in Income Group 3: +0.9424
   d. Effect of fall general knowledge score: +0.8542 per point increase
5. Significance levels:
   a. All p-values for the income group effects and the fall score effect are less than 0.001, indicating that these are statistically significant predictors of spring general knowledge score.
6. Model fit:
   a. R-squared: 0.731, indicating that approximately 73.1% of the variability in spring general knowledge score is explained by the model.
7. Additional interaction effects
   a. Interaction between Income Group 2 and fall general knowledge score: -0.0585
   b. Interaction between Income Group 3 and fall general knowledge score: -0.0829

```
                                OLS Regression Results
========================================================================================
Dep. Variable:     springgeneralknowledgescore   R-squared:                        0.731
Model:                                     OLS   Adj. R-squared:                   0.731
Method:                          Least Squares   F-statistic:                  1.082e+04
Date:                         Sat, 23 Mar 2024   Prob (F-statistic):                0.00
Time:                                 18:40:16   Log-Likelihood:                 -33259.
No. Observations:                        11933   AIC:                          6.653e+04
Df Residuals:                            11929   BIC:                          6.656e+04
Df Model:                                    3
Covariance Type:                     nonrobust
==============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Intercept                      8.0303      0.119     67.519      0.000       7.797       8.263
C(incomegroup)[T.2]            0.7084      0.088      8.005      0.000       0.535       0.882
C(incomegroup)[T.3]            0.9424      0.094     10.013      0.000       0.758       1.127
fallgeneralknowledgescore      0.8542      0.005    163.347      0.000       0.844       0.864
===============================================================================
Omnibus:                        75.905  Durbin-Watson:                    1.867
Prob(Omnibus):                   0.000  Jarque-Bera (JB):               101.391
Skew:                            0.090  Prob(JB):                      9.62e-23
Kurtosis:                        3.414  Cond. No.                          80.9
===============================================================================
```

These results suggest that there are significant differences in spring general knowledge scores across different income groups and that these differences are further influenced by the scores from the fall. Higher income groups tend to have higher scores in general, and the positive coefficients indicate that students with higher fall scores are likely to have higher spring scores. The negative interaction terms suggest that the effect of the fall score on the spring score decreases as the income group increases.

Results

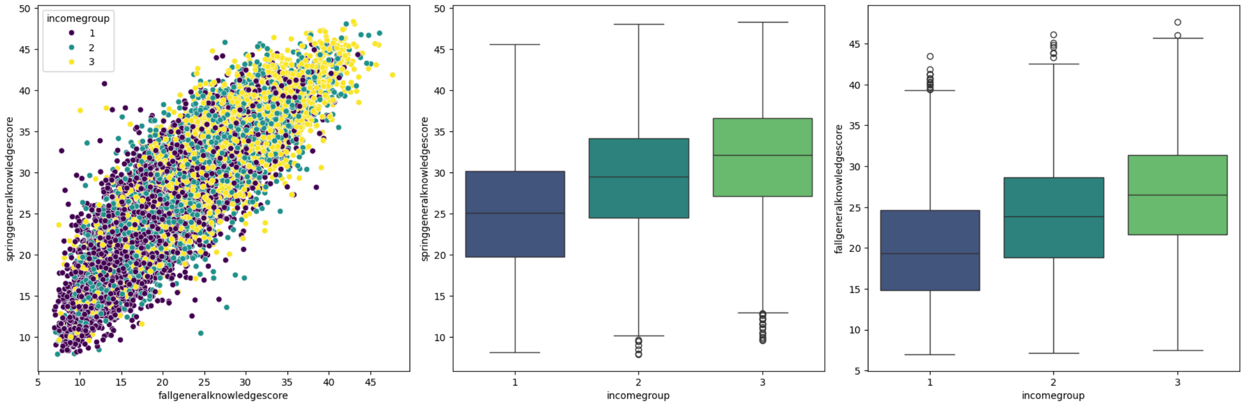
After reviewing the dataset and examining the visualizations of educational scores in connection to economic groups, substantial differences were discovered that warranted additional investigation. The observed variations in educational achievements between income groups can be attributed to the unique influences that socioeconomic status has on learning and development.

It is possible that households with diverse financial levels would apply different approaches to education, impacting the resources available for learning and potentially resulting to disparities in academic accomplishment. Higher-income households may have access to a broader selection of educational resources, which can enhance their children's learning opportunities. Conversely, lower-income families may face obstacles that limit their children's educational chances.

Additionally, systemic issues such as educational policy, financial support programs, and demographic features may have a substantial impact on these educational outcomes. For example, regions with greater investment in educational infrastructure may have higher student performance, demonstrating the impact of community resources on educational outcomes. On the other side, places with limited access to such services may see a decrease in academic advancement among pupils who live there.

The findings of this investigation, as well as the graphical representations, suggest that addressing educational gaps requires a multifaceted strategy. Understanding the impact of income on educational achievements enables stakeholders to develop targeted interventions that build an educational environment in which all students, regardless of socioeconomic status, have the ability to thrive.

Data Visualization



Visualization and Diagnostic Tools:

The report contains a number of visualizations that show the association between students' general knowledge scores and their economic level. Scatter plots and boxplots are the key diagnostic tools, illustrating the distribution and variance of scores across income levels, as well as their change from fall to spring.

The scatterplot shows the association between fall and spring general knowledge scores, color-coded by income category. This image clearly demonstrates how autumn semester scores compare to spring semester scores, with the distribution of data points demonstrating variability within each income level. The gradient of points indicates a positive association, meaning that greater fall scores are often related to higher spring scores.

Boxplots provide a summary of the distribution of general knowledge scores, with separate representations for fall and spring. The boxplots show median scores and interquartile ranges for each income category, providing information on central tendencies and dispersions. The boxplots reveal that higher income groups had higher median scores in both semesters, and the score distribution, as illustrated by the interquartile ranges, varies between income groups.

Conclusion

This analysis reveals differences that necessitate a more nuanced approach to educational policy and resource allocation. Interventions and support mechanisms customized to lower-income populations may be helpful in closing the educational achievement gap. Future research should consider longitudinal data and additional factors to provide a comprehensive picture of educational equity and guide impactful educational reforms.