STA130 FINAL PROJECT

EXPLORING THE PROGRESS TOWARDS ACHIEVING QUALITY EDUCATION BETWEEN COUNTRIES BY DIFFERENT INDICATORS

Anh Dang Phuong, Emma Shen, Li Chen, Mariana Garcia, Richard Lin

April, 2024

PROJECT INTRODUCTION

Guiding Question: Are there any significant correlations between countries' progress towards achieving quality education and other indicators such as development level, life expectancy, and gender?

Goal and Motivation:

- Education is essential for a society to progress, as it allows people to break the cycle of poverty and have a better life quality (United Nations). We acknowledge the crucial impact of education on the development of not only individuals but also communities, which contributes to the country's sustainability. Unfortunately, despite the efforts to improve education standards across all regions, disparities persist, as many countries are still finding challenges to improve education.
- We aim to analyze how education relates to other factors such as child labor, life expectancy, and inequality to gain a further understanding and to be able to propose insights, supported by data and statistical analysis, on how to increase progress towards the SDGs.

Population and Sample:

- Our population is all countries around the world, while our research sample will be gained by utilizing the most out of the available datasets provided by UNICEF, depending on each question.

DATA SUMMARY

International Organization for Standardization (ISO) Code (Categorical Variable)

- Represents different countries, territories, or special geographical areas.

Development Level (Binary Categorical Variable)

- Categorizes a country to be either “Developed” or “Developing”.

UNDP's Rate of Inequality in Education (Continuous Quantitative Variable)

- Represents inequality rates from 2010 to 2021 with values ranging from 0.5501101 to 50.12411.

Life Expectancy at Birth (Continuous Quantitative Variable)

- Stores life expectancy at birth for different countries in 2021, number of years a person born in 2021 is expected to live if current trends continue.

Sustainable Development Goals’ Goal 4 Status (Nominal Categorical Variable)

- Classifier for level of achievement of SDG 4.

Sustainable Development Goals’ Goal 4 Score (Continuous Quantitative Variable)

- Score given based on progress towards achieving SDG4.

UNICEF's Child Labor Rate (Continuous Quantitative Variable)

- Includes three variables representing the proportion of child labor in each country in total, in females, and males, with values ranging from 0.12 to 50.6.

DATA WRANGLING

- Select all variables that are particularly necessary for this study and rename the columns by using **select()** and **rename()**.
- Combine the rows of ‘hdr_ineq_edu_2010’, ‘hdr_ineq_edu_2011’, … up to ‘hdr_ineq_edu_2021’ along with the corresponding development levels to make a tibble of two variables: ‘development_level’ and ‘hdr_ineq_edu’.
- **Merge** a tibble storing quality education scores (goal_4_score) with life expectancy information by country code iso3 and create data sets for each group of level of achievement using **filter()**.
- Create a child_labor_information data frame that stores child labor rates in total, female, and male in each country, then merge it with development level information by country code iso3, and finally create a new column for gender rate ratio (male_child_labor/female_child_labor) by the function **mutate()**.
- Remove all observations with missing value in any of the selected variables by **na.omit()**.

RESEARCH QUESTION 1

Was the median educational inequality rate in developing countries larger than the median rate of educational inequality in developed countries, in the period from 2010 to 2021?

- As highlighted by SDG4, "quality education must be accessible for all, leaving no one behind" (un.org, n.d.). However, the rates of inequality are not the same for all countries. Thus, we want to investigate if there is an educational inequality gap in different countries. If developing countries tend to have higher rates in educational inequality, it would mean that children in developing countries may have fewer chances to comprehensively grow because of less available assistance.

HYPOTHESES TEST STATEMENT

Null Hypothesis: There is no difference between the median educational inequality rates in developing and developed countries from 2010 to 2021.

$$H_0 : \text{median}_{developing} - \text{median}_{developed} = 0$$

Alternative Hypothesis: The median rate of educational inequality in developing countries is larger than the median rate of educational inequality in developed countries.

$$H_1 : \text{median}_{developing} - \text{median}_{developed} > 0$$

RESEARCH QUESTION 1: STATISTICAL METHOD

Objective:

- Estimate the difference of medians of educational inequality between developed and developing countries

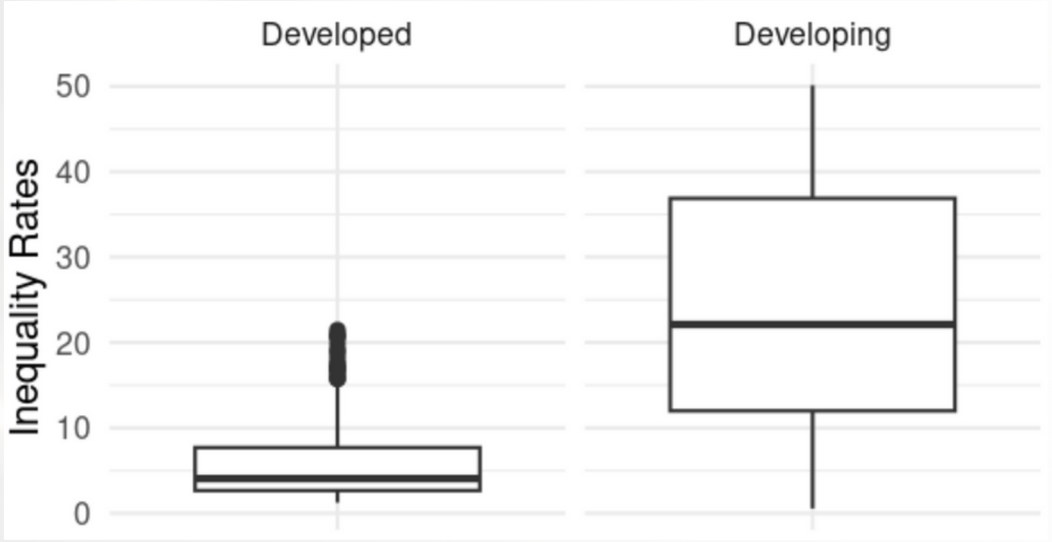
Variables:

- We used the hdr_ineq_edu variable to get all the available inequality rates from 2010 to 2021 and used the development_level variable to group the sample population based on development level (developed or developing).

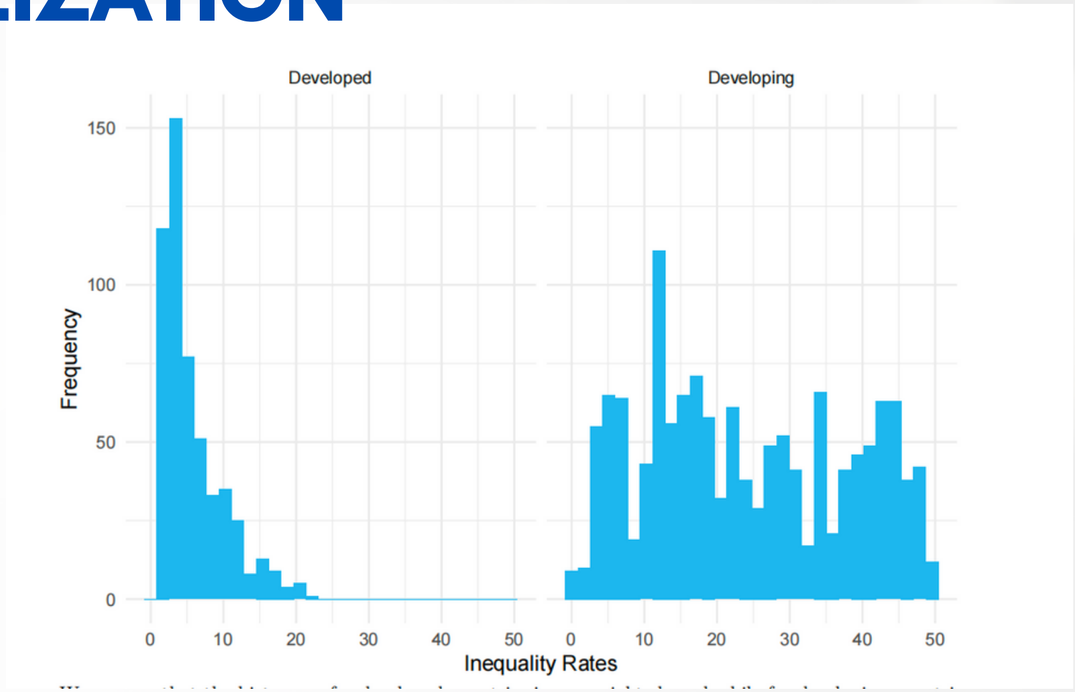
Hypothesis Testing:

- We used a two-sample hypothesis test to analyze the data. Under the assumption that the null hypothesis is true (i.e. there is no difference in inequality rates between developed and developing countries), we used R to randomly assign each inequality rate to either the developed group or the developing group.
- We decided to do a one-sided test.
- We set the alpha-level to be 0.05 to compare with the p-value after testing.

RESEARCH QUESTION 1: VISUALIZATION

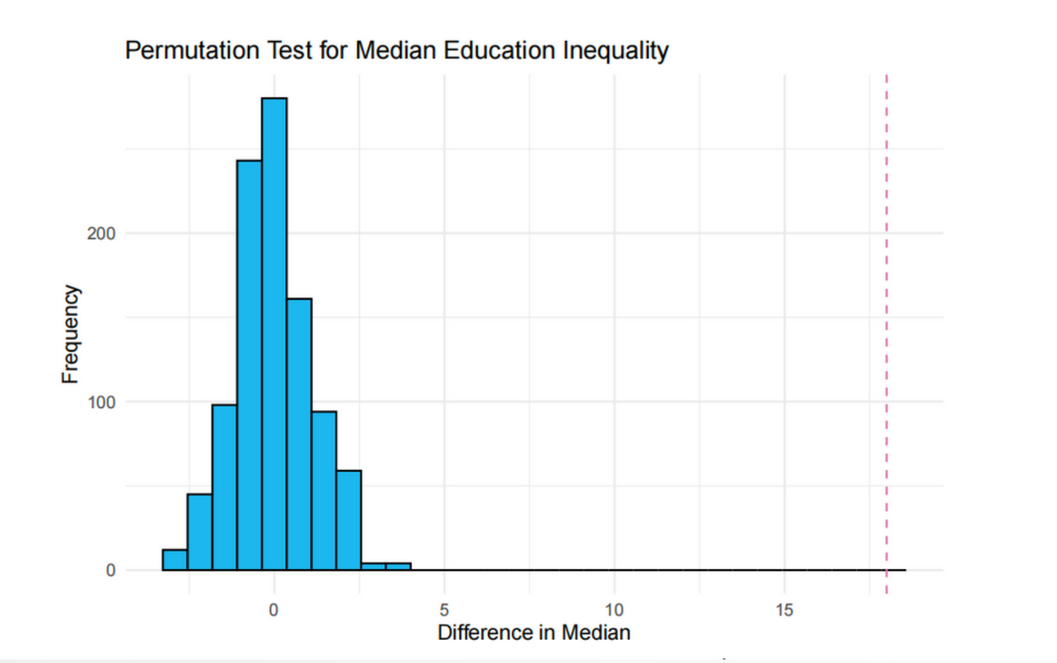


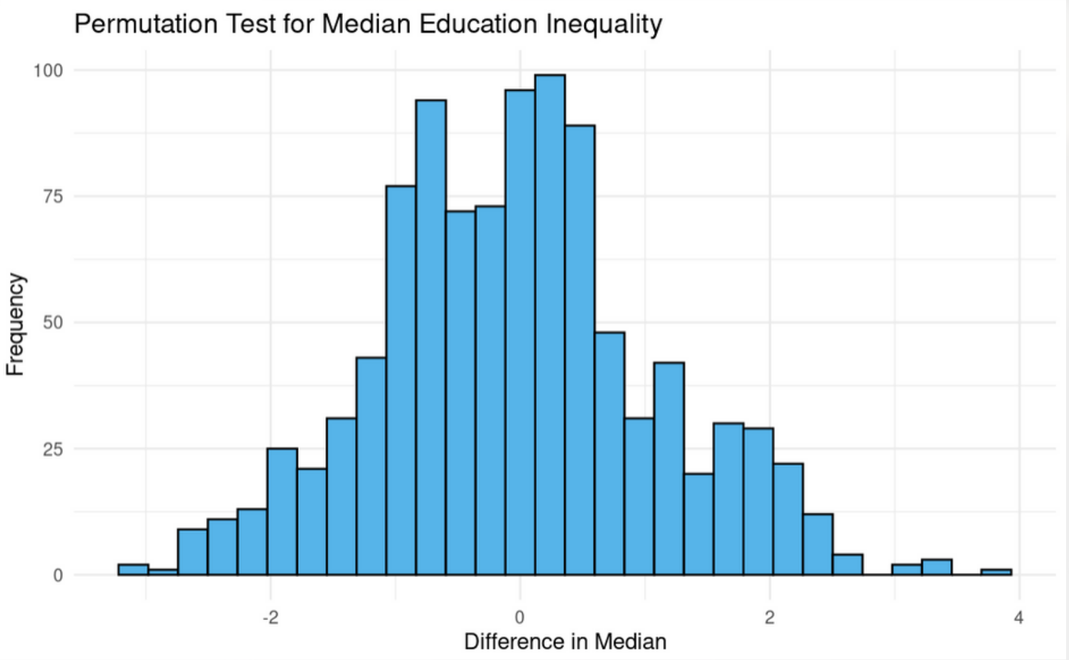
The boxplot shows that the median inequality rates in developing countries are roughly four times as higher as inequality rates in developed countries.



The distribution of developed nations is right-skewed, indicating a lower prevalence of extreme inequality but still being high enough to be concerning. The multimodal distribution seen in developing nations reflects the complexity and diversity of their educational institutions.

RESEARCH QUESTION 1: RESULTS AND INTERPRETATION





The observed test statistic is about 18.04 and the p-value equals 0 in this case. This suggests there is a very strong evidence to against the null hypothesis that there is no difference in the inequality rates between developing and developed countries.

RESEARCH QUESTION 1: RESULT AND INTERPRETATION

- Since the calculated p-value is less than 0.05, it provides strong evidence against the null hypothesis, meaning that the null hypothesis greatly reduces the likelihood that our test statistic would occur. Furthermore, we can accept the alternative hypothesis and reject the null hypothesis because the p-value is smaller than the statistical significance level of 0.05.
- The percentage of events with a decreased likelihood of happening is known as the p-value. In order to determine if a developing country has an impact on educational disparity, our alternative hypothesis asserts that rates of educational inequality are higher in rising nations.
- The percentage of events with a decreased likelihood of happening is known as the p-value. We have very little likelihood of making a Type 1 error because our p-value is so low.

RESEARCH QUESTION 2: INTRODUCTION

Does life expectancy at birth vary depending on the education level of different countries?

Why is this question important?

- **Previous studies:**
  - “People who are more highly educated live longer lives.” (Bull, 2015).
  - “The past century’s progress in medicine and public health has lengthened life expectancy, but the pace of progress has been modest” (Woolf et al., 2011).

- **In achieving SDGs:**
  - Observe how level of achievement in SDG4 might affect other SDGs.
  - Emphasize on the importance of achieving SDG4.
  - Closer look at groups of countries that need to work further on this goal.

RESEARCH QUESTION 2: STATISTICAL METHOD

Objective:

- Estimating the **median** life expectancy at birth for groups of countries with different SDG4 achievement levels.

Variables:

- life_expectancy_at_birth
- goal_4_status

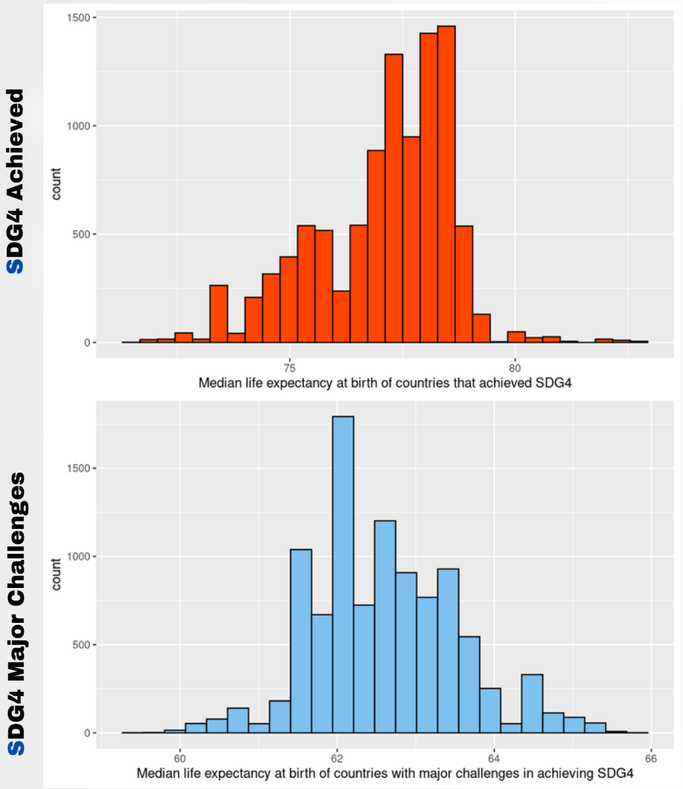
Bootstrapping:

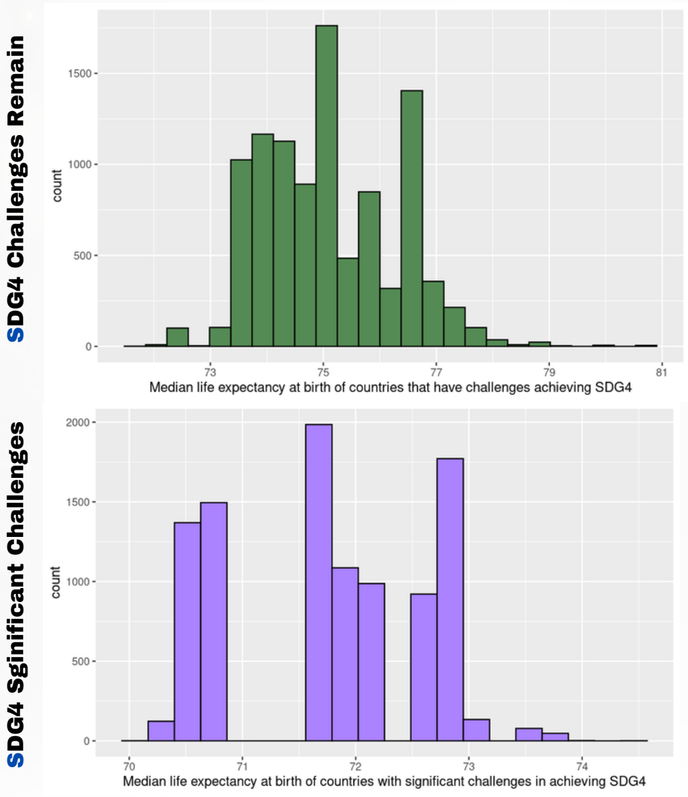
- Then, the bootstrapping method was implemented to create a **sampling distribution** for each group by **resampling** the observed sample **with replacement**.

Confidence interval:

- The **90%** confidence interval (middle 90% of values of the bootstrap statistics).
- **Range of plausible values** for the **median life expectancy for each group**.

RESEARCH QUESTION 2: DATA VISUALIZATION



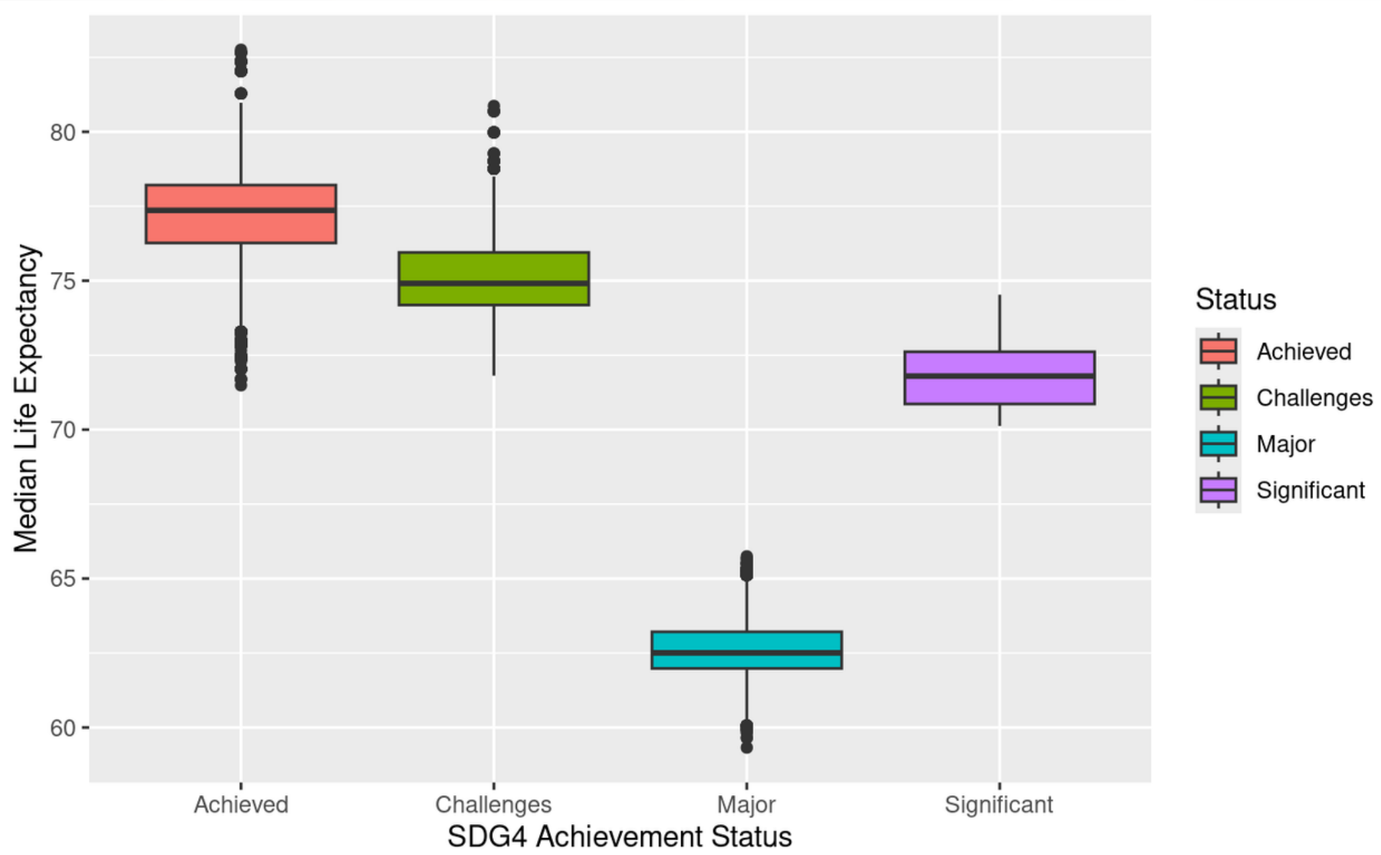


The distribution for SDG 4 achieved appears to be **left skewed** indicating a greater number of countries in this category has a median life expectancy on the higher end of the range of values.

Challenges remain and Major challenges remain categories both follow a **normal distribution.**

Significant challenges remain category follows a **multimodal** distribution.

RESEARCH QUESTION 2: VISUALIZATION



From observing the box plots of each category, it is **noteworthy** that countries which have achieved SDG4 **has a higher median life expectancy compared to the other categories.**

90% Confidence Intervals for each of the SDG4 categories

Achieved <dbl>	Challenges_remain <dbl>
74.13025	73.67645
78.71290	77.06710

Major_challenges <dbl>	Significant_challenges <dbl>
61.20495	70.4697
64.36360	72.8140

Similarly, from the confidence interval of each category, we notice the same pattern, the **estimated potential range of values** for the median life expectancy of the SDG4 achieved group is **higher than the others.**

RESEARCH QUESTION 2: ANALYSIS

As a general trend, we found that **countries with higher level of education**, in our case countries with higher progress towards SDG #4, **tend to have a higher life expectancy than those with less progress.**

However, it is important to consider that **these results have some limitations**, as they do not consider other possible confounding factors like health care or income.

Although further studies would be needed to assess the strength of this association, the results provided by the study can serve as support to state that **a lower progress towards achieving SDG 4 might be associated with a lower life expectancy** for the population of a country, emphasizing on the **importance of achieving this goal for the purpose of increasing life expectancy.**

RESEARCH QUESTION 3: INTRODUCTION

Are there any relationships between quality education (SDG4) and child labour (SDG8)?

Why is this question important?

- **Previous studies:**
  - **“**To achieve universal primary education, governments must address the link between child labor and education, dramatically escalating efforts to remove children from the worst forms of child labor” (HRW, n.d.).
  - “In many cases, employers actively prohibit children from attending school, while in others, the long hours demanded by employers make schooling practically impossible.” (HRW, n.d.).
- **In achieving SDGs:**
  - Possible relationship between achieving SDG4 and achieving SDG8 could help unveil a way of working towards SDG4 by reducing child labor through policy.

RESEARCH QUESTION 3: STATISTICAL METHOD

Objective:

- Understand the **relationship** between child labor and SDG4 achievement.

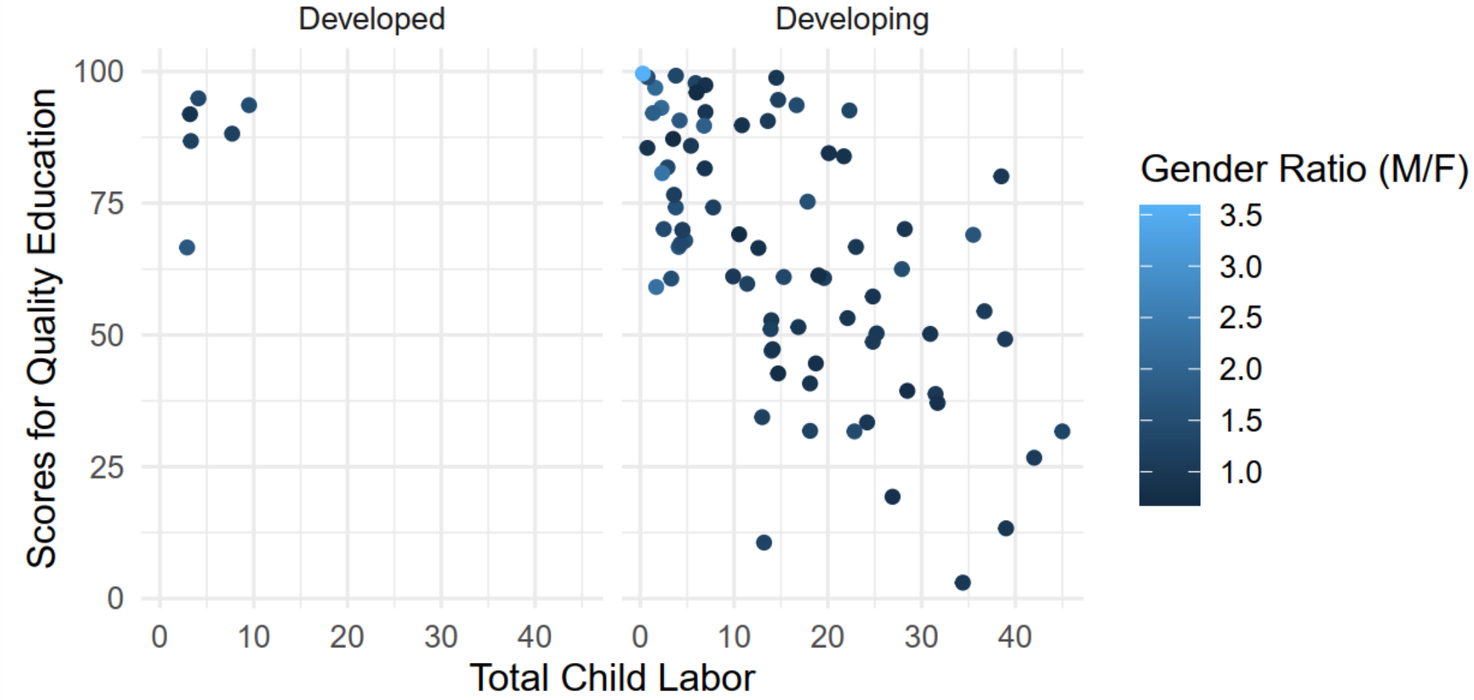
Variables:

- total_child_labor (predictor) & goal_4_score (response).

Linear regression:

- We used the lm() function to obtain the **regression coefficients** for the fitted linear regression equation.
- Then, we calculated the **correlation coefficient** and $\mathbf{R^2}$ in order to assess the **direction** and **strength** of the relationship, as well as the **variability** captured by our model.
- Finally, we performed a **hypothesis test** for the slope of the regression line to confirm that there's a correlation between the Child labor and Goal 4 score.

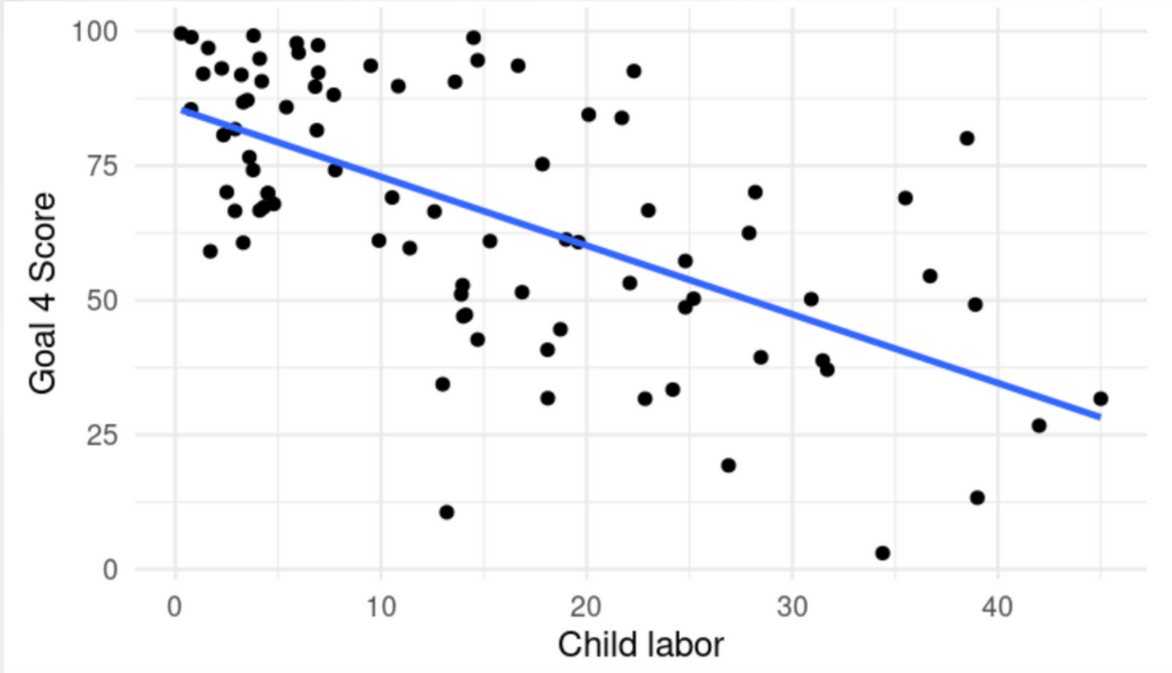
RESEARCH QUESTION 3: DATA VISUALIZATION



The scatterplot illustrates the total child labor percentage to its corresponding quality education score, divided into developed and developing categories, and colored by the gender ratio of child labor.

Noticeably, developed countries have much fewer points displayed in the scatterplot because of many missing entries. For developing countries, it is clearer that the data has a linear form with a negative association between 'total child labor' and 'quality education score'.

RESEARCH QUESTION 3: RESULTS AND INTERPRETATION




```
##                     Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)        85.332570  3.4073272 25.043844 1.846406e-38
## total_child_labor  -1.279891  0.1818508 -7.038137 7.441589e-10
```

$$\hat{y} = 85.33 - 1.28x$$

↓ quality education score

↓ child labor percentage

Hypothesis test to confirm the real correlation

$H_0 : \beta_1 = 0$ ⟶ no correlation

$H_1 : \beta_1 \neq 0$ ⟶ there's a correlation

Set alpha level = 0.05.

We have p-value = 7.44e-10

p-value ≤ alpha level, implying that there is a correlation between two variables (we **reject the null hypothesis**).

R = -0.6222196

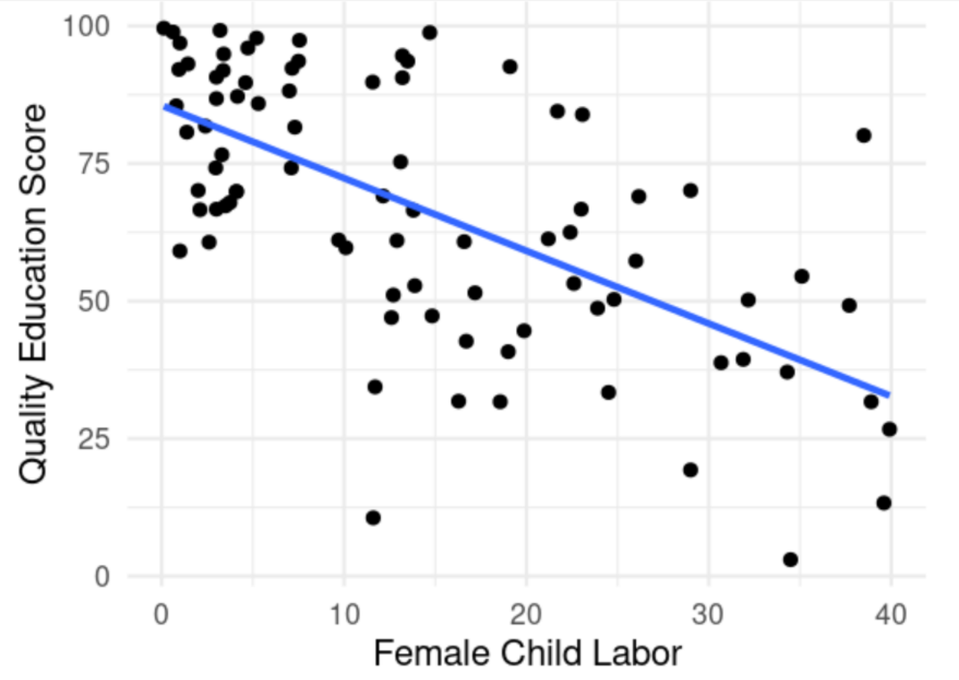
- negative correlation
- the variables have a somewhat strong association since |-0.6222196| is neither very big nor very small.

R^2 = 0.3871573

- the regression model explains very little of the variability of goal 4 score with respect to child labor.

RESEARCH QUESTION 3: ONE STEP FURTHER

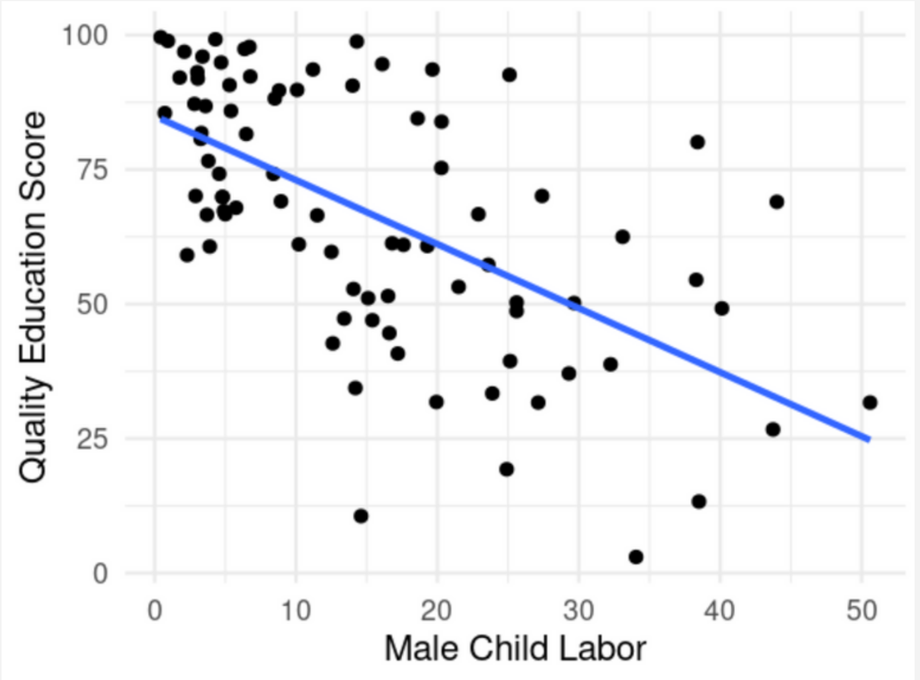
FEMALE



$$\hat{y} = 85.16 - 1.33x$$

R = -0.64

MALE



$$\hat{y} = 84.55 - 1.18x$$

R = -0.60

LIMITATION AND ASSUMPTIONS (FOR ALL QUESTIONS)

- **For research question 1...**
  - A lot of missing values in the file.
  - Data for developing countries almost 3 times more than for developed countries.
- **For research question 2...**
  - The observed sample for some education groups was smaller than for others, due to a different number of observations for each group.
- **For research question 3...**
  - We are assuming our data follows a linear form.
  - Survivorship bias, as some countries did not count with the data for 'child labor', and so we cannot be certain that the sample is entirely representative of the population.

PROJECT CONCLUSION

Research Question 1: The results of research question 1 suggest that we should support the alternative hypothesis and acknowledge a difference between the inequality rates. It is reasonable, as our hypothesis was children in developed countries may have more opportunities to go to school and can pursue a quality education than children in poorer countries.

Research Question 2: The results of research question 2 suggest that countries with higher levels of education, in our case those with higher progress towards SDG #4, tend to have a higher life expectancy than those with less progress. However, it is quite surprising to see a reversed pattern of countries with significant challenges and major challenges.

Research Question 3: The results of research question 3 suggest that there exists a correlation between 'goal 4 score' and 'child labor'. Insights were limited when splitting by development level because of missing data, yet we found that girls that engage in child labor are a little more likely to not have a good quality education.

REFERENCES AND ACKNOWLEDGEMENTS

Hummer, R. A., & Hernandez, E. M. (2013). The Effect of Educational Attainment on Adult Mortality in the United States. Population bulletin, 68(1), 1–16. Retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4435622/#R31

Woolf, S. H., Johnson, R. E., Phillips Jr, R. L., & Philipsen, M. (2007). Giving everyone the health of the educated: an examination of whether social change would save more lives than medical advances. American journal of public health, 97(4), 679-683. ISO 690. Retrieved from: https://ajph.aphapublications.org/doi/full/10.2105/AJPH.2005.084848

Report IV Child Labor (2005). Human Rights Watch. Retrieved from: https://www.hrw.org/reports/2005/education0905/8.htm

Goal 4: Quality Education. “The 17 Goals.” Sustainable Development Goals, United Nations, https://www.un.org/sustainabledevelopment/education/

Our group would like to thank Evan, Prof Speagle, Prof Moon, and the TAs for providing us this wonderful experience and for their advice to improve this project.

THANK YOU