STA130 - Final Project - Exploring The Progress Towards Achieving Quality Education Between Countries By Different Indicators

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

2024-03-10

```
# Set up the seed for later uses
SEED <- 13013020
```

I. Data Wrangling and Cleaning:

```
# load country indicators data
country_indicators <-
  read_csv("country_indicators.csv") %>%
  select(-...1) %>%  # remove first column
  select(iso3, everything()) %>%  # reorder the columns to put iso3 as column 1
  rename(country_code_iso3 = iso3)  # rename first column to country_code_iso3
```

```
## New names:
## Rows: 218 Columns: 1332
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (8): iso3, hdr_hdicode, hdr_region, wbi_income_group, wbi_lending_cat... dbl
## (1324): ...1, sowc_demographics__population-thousands-2021_total, sowc_d...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
# preview data
country_indicators
```

```
## # A tibble: 218 x 1,331
##    country_code_iso3 sowc_demographics__population-thou~1 sowc_demographics__p~2
##    <chr>                                           <dbl>                  <dbl>
##  1 AFG                                             40099.                 20298.
##  2 ALB                                              2855.                   574.
##  3 DZA                                             44178.                 15526.
##  4 AND                                                79.0                  12.8
##  5 AGO                                             34504.                 17833.
##  6 AIA                                                15.8                   3.29
##  7 ATG                                                93.2                  21.3
##  8 ARG                                             45277.                 12669.
##  9 ARM                                              2791.                   669.
## 10 AUS                                             25921.                  5667.
## # i 208 more rows
## # i abbreviated names: 1: `sowc_demographics__population-thousands-2021_total`,
```

```
## #   2: `sowc_demographics__population-thousands-2021_under-18`
## # i 1,328 more variables:
## #   `sowc_demographics__population-thousands-2021_under-5` <dbl>,
## #   `sowc_demographics__annual-population-growth-rate_2000-2020` <dbl>,
## #   `sowc_demographics__annual-population-growth-rate_2020-2030-a` <dbl>, ...
```

```
# load SDG data
sdg <-
  read_csv("sdr_fd5e4b5a.csv") %>%
  select(-...1)  # remove first column
```

```
## New names:
## Rows: 206 Columns: 59
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (36): Goal 1 Dash, Goal 1 Trend, Goal 2 Dash, Goal 2 Trend, Goal 3 Dash,... dbl
## (23): ...1, Goal 1 Score, Goal 2 Score, Goal 3 Score, Goal 4 Score, Goal...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
# rename columns
names(sdg)[1:(2*17)] <-
  paste(c(rep(paste("goal_", 1:17, sep=""), each=2)),
        rep(c("_status", "_trend"), times=17), sep="")
names(sdg)[(2*17 + 1):(3*17)] <-
  paste("goal_", 1:17, "_score", sep="")
names(sdg)[names(sdg)=="2023 SDG Index Score"] <-
  "SDG_index_score_2023"
names(sdg)[names(sdg)=="2023 SDG Index Rank"] <-
  "SDG_index_rank_2023"
names(sdg)[names(sdg)=="Percentage missing values"] <-
  "percentage_missing_values"
names(sdg)[names(sdg)=="International Spillovers Score (0-100)"] <-
  "international_spillover_score"
names(sdg)[names(sdg)=="International Spillovers Rank"] <-
  "international_spillover_rank"
names(sdg)[names(sdg)=="Country Code ISO3"] <-
  "country_code_iso3"

# preview data
sdg
```

```
## # A tibble: 206 x 58
##    goal_1_status     goal_1_trend       goal_2_status goal_2_trend goal_3_status
##    <chr>             <chr>              <chr>         <chr>        <chr>
##  1 SDG achieved      On track or maint~ Major challe~ Score stagn~ Challenges r~
##  2 SDG achieved      On track or maint~ Major challe~ Score stagn~ Challenges r~
##  3 SDG achieved      Score moderately ~ Significant ~ Score stagn~ Challenges r~
##  4 Challenges remain Decreasing         Significant ~ Score stagn~ Significant ~
##  5 SDG achieved      Score moderately ~ Significant ~ Score stagn~ Challenges r~
##  6 SDG achieved      Score moderately ~ Significant ~ Score stagn~ Significant ~
##  7 SDG achieved      Score stagnating ~ Major challe~ Score stagn~ SDG achieved
##  8 SDG achieved      On track or maint~ Major challe~ Score stagn~ Significant ~
##  9 SDG achieved      On track or maint~ Major challe~ Score stagn~ Significant ~
## 10 Challenges remain Score moderately ~ Major challe~ Score stagn~ Significant ~
```

```
## # i 196 more rows
## # i 53 more variables: goal_3_trend <chr>, goal_4_status <chr>,
## #   goal_4_trend <chr>, goal_5_status <chr>, goal_5_trend <chr>,
## #   goal_6_status <chr>, goal_6_trend <chr>, goal_7_status <chr>,
## #   goal_7_trend <chr>, goal_8_status <chr>, goal_8_trend <chr>,
## #   goal_9_status <chr>, goal_9_trend <chr>, goal_10_status <chr>,
## #   goal_10_trend <chr>, goal_11_status <chr>, goal_11_trend <chr>, ...
```

```
# load country codes data, and select only English language
country_codes <- read_csv("country_codes.csv") %>%
  select('ISO-alpha3 Code (M49)', 'Region Name_en (M49)',
         'Country or Area_en (M49)',
         'Developed / Developing Countries (M49)') %>%
  rename(country_code_iso3 = 'ISO-alpha3 Code (M49)',
           region = 'Region Name_en (M49)',
           country_label = 'Country or Area_en (M49)', development_level =
             'Developed / Developing Countries (M49)')
```

```
## New names:
## Rows: 298 Columns: 125
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (99): Global Name_en (M49), Region Name_en (M49), Sub-region Name_en (M4... dbl
## (22): ...1, Global Code (M49), Region Code (M49), Intermediate Region Co... lgl
## (4): Sub-region Code (M49), Least Developed Countries (LDC) (M49), Land...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
country_codes
```

```
## # A tibble: 298 x 4
##    country_code_iso3 region country_label                 development_level
##    <chr>             <chr>  <chr>                         <chr>
##  1 DZA               Africa Algeria                       Developing
##  2 EGY               Africa Egypt                         Developing
##  3 LBY               Africa Libya                         Developing
##  4 MAR               Africa Morocco                       Developing
##  5 SDN               Africa Sudan                         Developing
##  6 TUN               Africa Tunisia                       Developing
##  7 ESH               Africa Western Sahara                Developing
##  8 IOT               Africa British Indian Ocean Territory Developing
##  9 BDI               Africa Burundi                       Developing
## 10 COM               Africa Comoros                       Developing
## # i 288 more rows
```

```
# Join tables
indicators_n_sdg <- inner_join(x=country_indicators, y=sdg, by="country_code_iso3")

indicators_n_sdg
```

```
## # A tibble: 193 x 1,388
##    country_code_iso3 sowc_demographics__population-thou~1 sowc_demographics__p~2
##    <chr>                                           <dbl>                  <dbl>
##  1 AFG                                            40099.                 20298.
##  2 ALB                                             2855.                   574.
```

```
##  3 DZA                                                44178.                 15526.
##  4 AND                                                   79.0                   12.8
##  5 AGO                                                34504.                 17833.
##  6 ATG                                                   93.2                   21.3
##  7 ARG                                                45277.                 12669.
##  8 ARM                                                 2791.                   669.
##  9 AUS                                                25921.                  5667.
## 10 AUT                                                 8922.                  1542.
## # i 183 more rows
## # i abbreviated names: 1: `sowc_demographics__population-thousands-2021_total`,
## #   2: `sowc_demographics__population-thousands-2021_under-18`
## # i 1,385 more variables:
## #   `sowc_demographics__population-thousands-2021_under-5` <dbl>,
## #   `sowc_demographics__annual-population-growth-rate_2000-2020` <dbl>,
## #   `sowc_demographics__annual-population-growth-rate_2020-2030-a` <dbl>, ...
```

With all the fundamental datasets and potentially useful variables being selected, we are ready to go through each research question and draw any noticeable insights for the Quality of Education between developed and developing countries.

II. 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.

III. Research Question 1:

a) Introduction and motivations.

Are there any significant differences between developed and developing countries in learning literacy rates?

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.

b) Data exploration.

For Question 1 we will use data from `country_indicators` and `sdg` (development_level and some ineq_edu rates)

```
# Load essential data for Research Question 1:
rq1_general <- country_indicators %>% select(country_code_iso3,
```

```
                                          paste(c(rep(paste("hdr_ineq_edu_",
                                                            2010:2021,
                                                            sep=""), each=1)
                                                  )))

rq1_general <- inner_join(x=rq1_general, y=country_codes,
                          by='country_code_iso3') %>% select(development_level,
                                                             everything()) %>%
  select(-region, -country_label)

rq1_general
```

```
## # A tibble: 220 x 14
##    development_level country_code_iso3 hdr_ineq_edu_2010 hdr_ineq_edu_2011
##    <chr>             <chr>                         <dbl>             <dbl>
##  1 Developing        AFG                            42.8              44.8
##  2 Developed         ALB                            11.9              11.9
##  3 Developing        DZA                            NA                NA
##  4 Developed         AND                            15.2              15.2
##  5 Developing        AGO                            NA                NA
##  6 Developing        AIA                            NA                NA
##  7 Developing        ATG                            NA                NA
##  8 Developing        ARG                             6.91              6.83
##  9 Developing        ARM                             3.68              3.68
## 10 Developed         AUS                             2.75              2.48
## # i 210 more rows
## # i 10 more variables: hdr_ineq_edu_2012 <dbl>, hdr_ineq_edu_2013 <dbl>,
## #   hdr_ineq_edu_2014 <dbl>, hdr_ineq_edu_2015 <dbl>, hdr_ineq_edu_2016 <dbl>,
## #   hdr_ineq_edu_2017 <dbl>, hdr_ineq_edu_2018 <dbl>, hdr_ineq_edu_2019 <dbl>,
## #   hdr_ineq_edu_2020 <dbl>, hdr_ineq_edu_2021 <dbl>
```

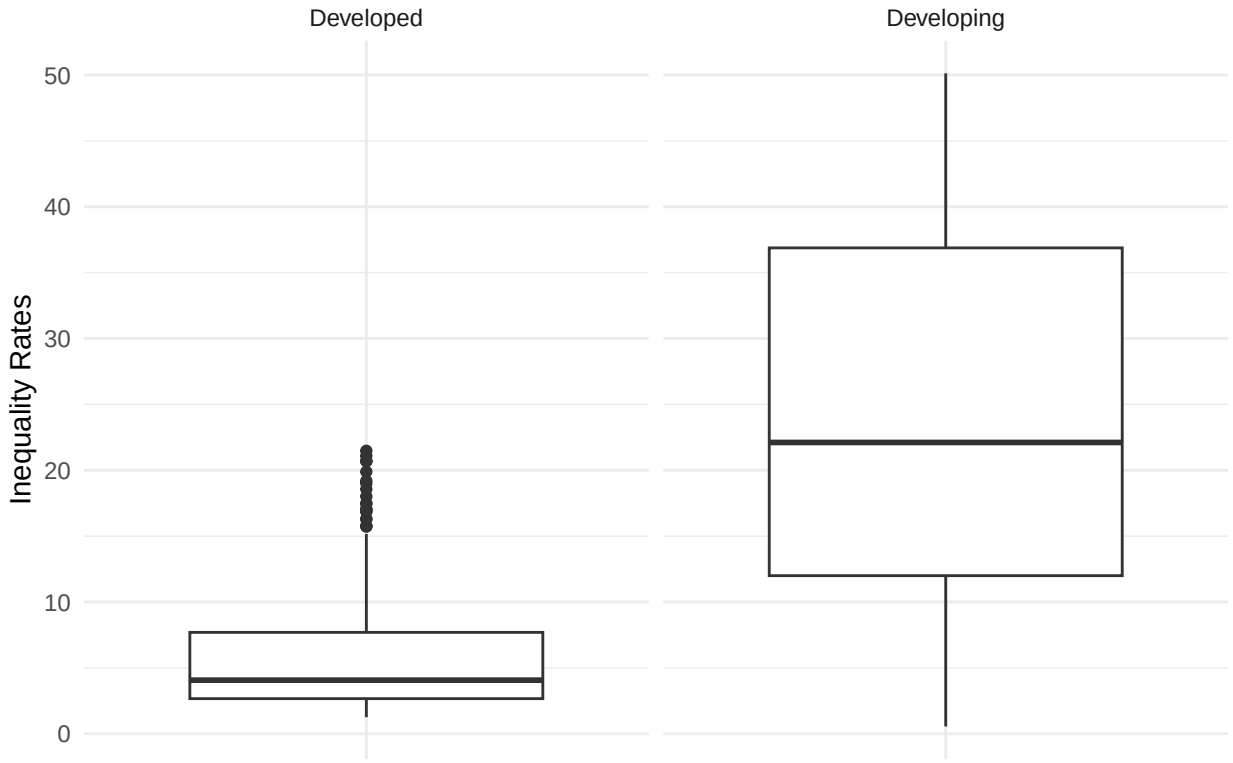
i) Deciding between the mean and median of inequality rates in education.

```
data_years <- list()

for (year in 2010:2021) {
  # Select columns for the current year, rename, and store in the list
  data_years[[as.character(year)]] <- rq1_general %>%
    select(development_level, !!sym(paste0("hdr_ineq_edu_", year))) %>%
    rename(hdr_ineq_edu = !!sym(paste0("hdr_ineq_edu_", year)))
}

# Combine row from those data years together
combine_data <- do.call(rbind, data_years) %>% na.omit()

combine_data
```

```
## # A tibble: 1,918 x 2
##    development_level hdr_ineq_edu
##    <chr>                   <dbl>
##  1 Developing               42.8
##  2 Developed                11.9
##  3 Developed                15.2
##  4 Developing                6.91
```

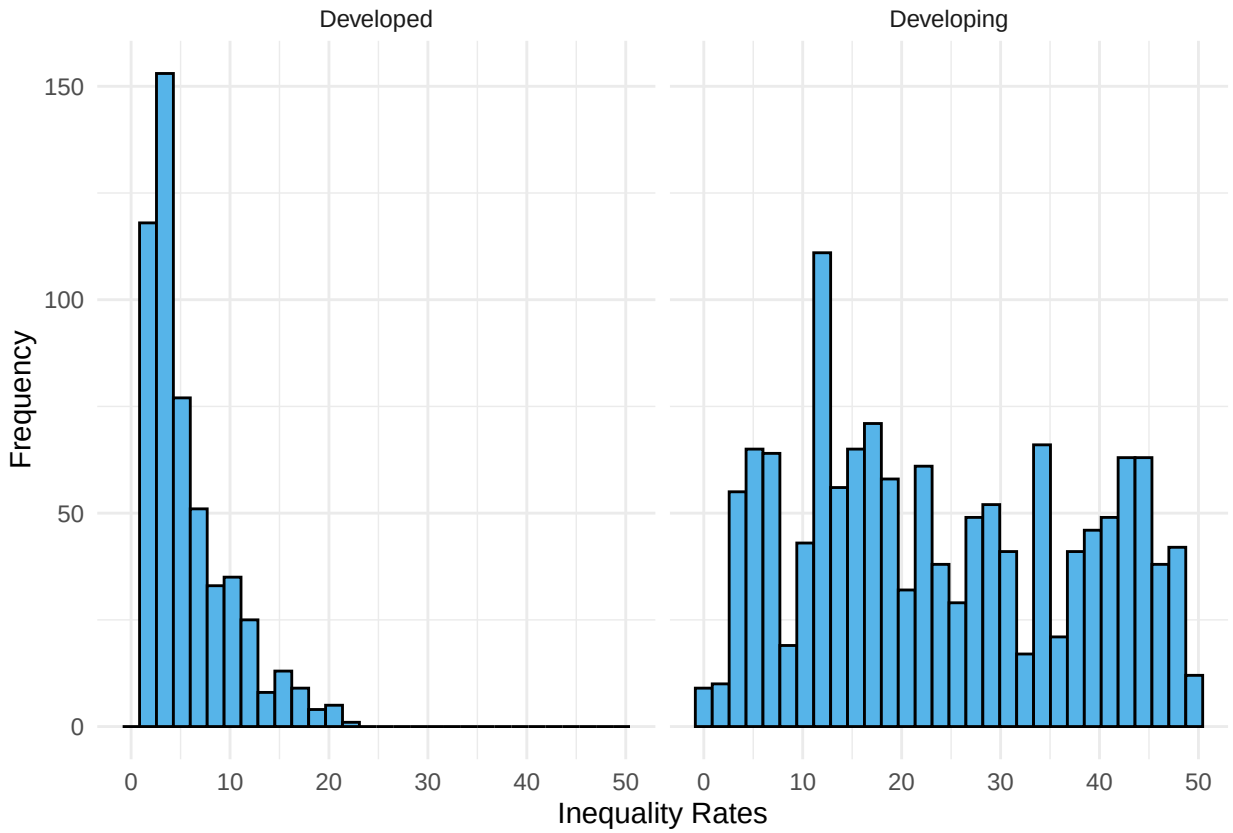
```
##  5 Developing                3.68
##  6 Developed                 2.75
##  7 Developed                 2.49
##  8 Developing                3.73
##  9 Developing                6.92
## 10 Developing               22.2
## # i 1,908 more rows
```

```r
# Plot to observe Mean or Median
# Use boxplots to look for outliers
combine_data %>% ggplot(aes(x='', y = hdr_ineq_edu)) +
  geom_boxplot() + facet_wrap(~development_level) +
  labs(x= '', y = 'Inequality Rates') +
  theme_minimal()
```




```r
# Use histogram to look for skewness
combine_data %>% ggplot(aes(x = hdr_ineq_edu)) +
  geom_histogram(fill = '#56B4E9', color = 'black') + facet_wrap(~development_level) +
  labs(x='Inequality Rates', y='Frequency') +
  theme_minimal()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We can see that the histogram for developed countries is more right skewed while for developing countries, the histogram is more of a multimodal and relatively symmetrical pattern. From the boxplots, we can see that there are many outliers for the developed category, though not very significant from the rest of the boxplot.

With those data collected, we decided to work on the median rates of educational inequality between developed and developing countries, since mean is more sensitive to outliers.

Let $M_{developed}$ be the median rate of educational inequality in developed countries and $M_{developing}$ be the median rate of educational inequality in developing countries. Therefore, our hypotheses are:

$$H_0 : M_{developing} - M_{developed} = 0$$

$$H_1 : M_{developing} - M_{developed} > 0$$

Calculate $\Delta\hat{M} = \hat{M}_{developing} - \hat{M}_{developed}$ (DEFINITION)

```
# Find median for each group
delta_median_hat <- combine_data %>% group_by(development_level) %>%
  summarise(medians = median(hdr_ineq_edu)) %>% summarise(value = diff(medians))

# Calculate f_hat by the observed median from our dataset
delta_median_hat
```

```
## # A tibble: 1 x 1
##   value
##   <dbl>
## 1  18.0
```

Therefore, the observed test statistic is $\Delta\hat{M} = 18.04$

c) Hypothesis Testing

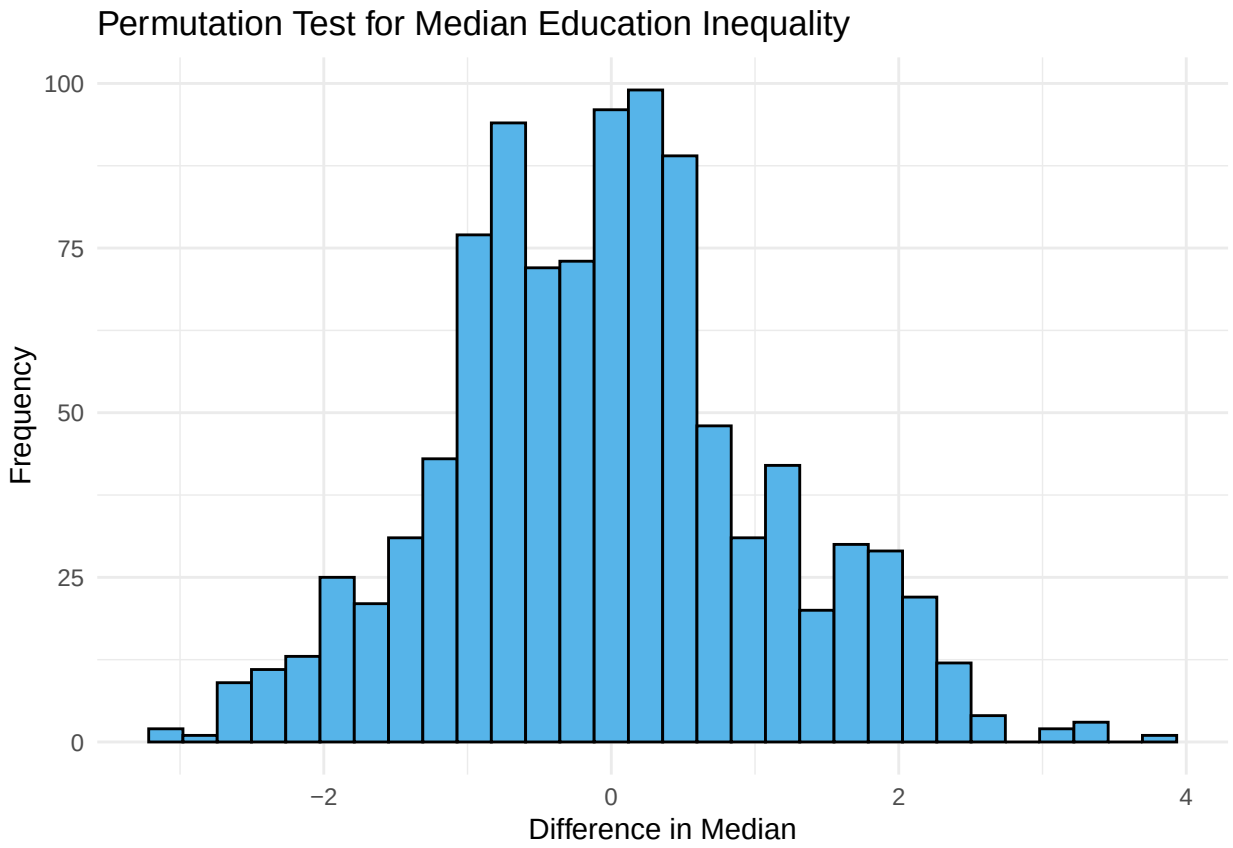
Below is R code that simulates $N = 1000$ values of the test statistic $\Delta\hat{M}_{\text{sim}}$ **under the null hypothesis** using a permutation test. In this test, we assume that our groups are identical under our null hypothesis. Mixing the two groups together, randomly generating new groups with the same sizes, and then recomputing our test statistic each time therefore should allow us to simulate values from the sampling distribution provided our sample size is large enough.

```
# Required for the reproducibility
set.seed(SEED)

# Set up
num_trials <- 1000
delta_median_simulations <- numeric(num_trials)

for(i in 1:num_trials){
  # Perform a random permutation
  permuted_data <- combine_data %>%
    mutate(development_level = sample(development_level, replace=FALSE))

  # Compute the simulated test statistic
  delta_median_sim <- permuted_data %>%
    group_by(development_level) %>%
    summarise(medians = median(hdr_ineq_edu), .groups="drop") %>%
    summarise(value = diff(medians)) %>% as.numeric()

  # Store the simulated value
  delta_median_simulations[i] <- delta_median_sim
}
```

Plot the distribution of the simulated test statistics (i.e. the sampling distribution) for the permutation test computed above.

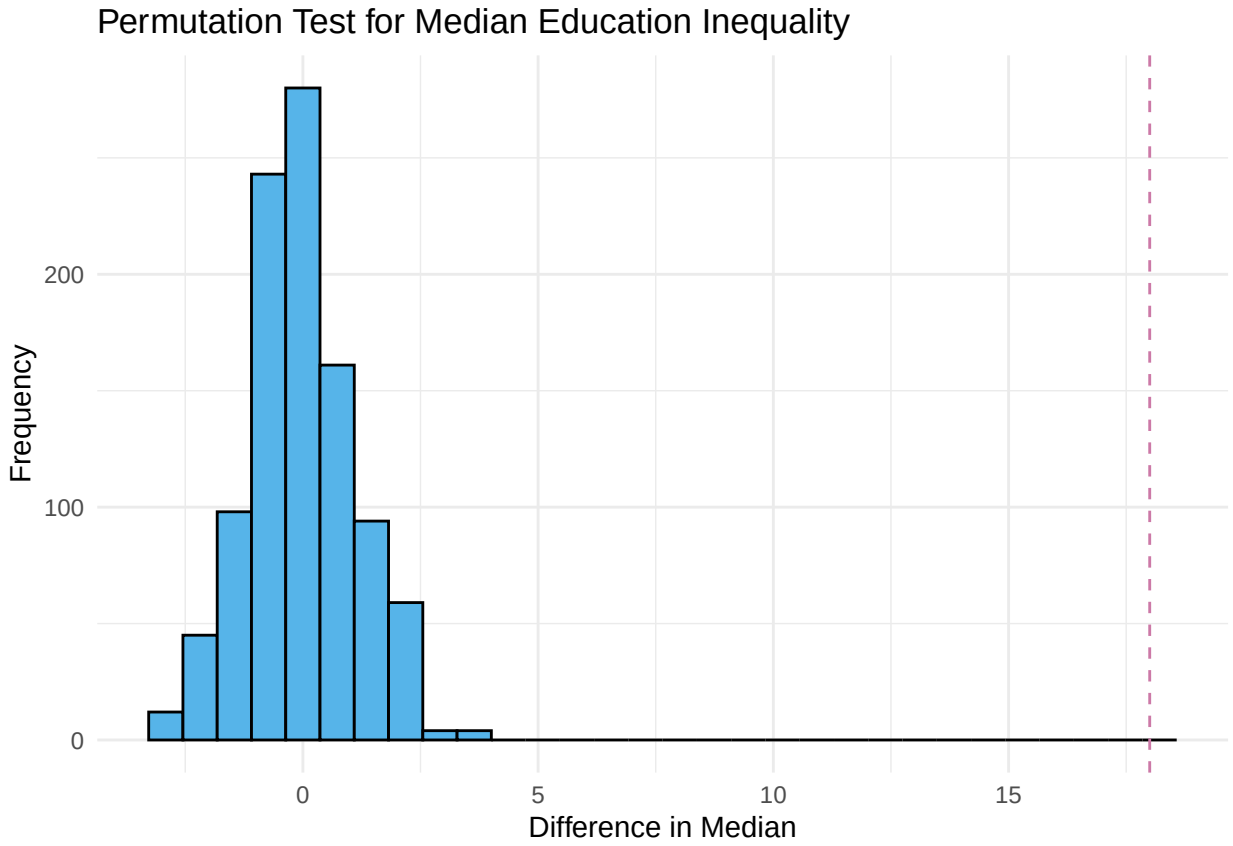
```
# The sampling distribution
sampling_distribution <- tibble(vhat_sim = delta_median_simulations)
ggplot(sampling_distribution, aes(x = vhat_sim), bins = 30) +
  geom_histogram(color = "black", fill = "#56B4E9") +
  labs(title = "Permutation Test for Median Education Inequality",
       x = "Difference in Median",
       y = "Frequency") +
  theme_minimal()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# The sampling distribution with the line for observed test statistic
sampling_distribution <- tibble(vhat_sim = delta_median_simulations)
ggplot(sampling_distribution, aes(x = vhat_sim), bins = 30) +
  geom_histogram(color = "black", fill = "#56B4E9") +
  geom_vline(xintercept = 18, color = "#CC79A7", linetype = "dashed",
             linewidth = 0.5) +
  labs(title = "Permutation Test for Median Education Inequality",
       x = "Difference in Median",
       y = "Frequency") +
  theme_minimal()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The histogram for the permutation test is relatively symmetrical around 0 (the null hypothesis). It also has two modes (roughly -0.7 and 0.3).

Compute the p-value based on the null hypothesis H_0 using the values of $\Delta\hat{M}_{\text{sim}}$ you computed from the sampling distribution, the observed test statistic $\Delta\hat{M}$, and the assumed true value under the null hypothesis $\Delta M = M_{developing} - M_{developed}$.

```
# 1-sided test (right)
rq1_p_value <- sum(delta_median_simulations >= delta_median_hat) / num_trials

# Print out the calculated p-value
rq1_p_value
```

```
## [1] 0
```

Thus, since the p-value is 0, we have a very strong evidence to against the null hypothesis. Therefore, there is a strong difference in the median of educational inequality rate between developing and developed countries, which means that in overall, developing countries face more inequalities in education than developed countries.

d) Analysis and Conclusion.

In conclusion, the hypothesis test above tells us to strongly reject H0, and we can conclude that the median inequality rate of developed countries is larger than that of developing countries. Although seeing the p-value equals zero is rare and often times it might be questionable due to a potential mistake in the dataset, it seems reasonable for this case, with the verified statistics from Unicef and with our expectations, to see that tremendous difference between the two groups. As the p-value is 0, we can also say that the probability of Type-1 error is very small.

IV. Research Question 2

a) Introduction and Motivations.

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

Previous studies have linked higher education with longer lifespans, and they've also demonstrated that despite progress in medicine and technology, the progress in lengthening life expectancy has been modest. Through our research question we aim to explore whether quality education might affect life expectancy in order to emphasize on the importance of achieving SDG4, as it might be affecting the progress towards other SDGs, like SDG3 (good health and well-being) through it's effect on factors such as life expectancy, a big indicator of health in a country.

b) Data Exploration

Before performing the test the data was wrangled to have a more manageable and organized data set. This section describes the process of data wrangling the data went through previous to the testing.

To perform the test we will use two numerical variables: 'goal_4_status' to measure quality education, and 'sowc_demographics___life-expectancy-at-birth-years_2021-0' to measure child labor.

Because of this, we'll create a data set 'goal_4_information' containing only the 'goal_4_status' variable and the ISO 3 country codes:

```
goal_4_information <- sdg %>%
  select(goal_4_status, country_code_iso3, country_label)

glimpse(goal_4_information)
```

```
## Rows: 206
## Columns: 3
## $ goal_4_status     <chr> "SDG achieved", "Challenges remain", "Challenges rem~
## $ country_code_iso3 <chr> "FIN", "SWE", "DNK", "DEU", "AUT", "FRA", "NOR", "CZ~
## $ country_label     <chr> "Finland", "Sweden", "Denmark", "Germany", "Austria"~
```

Next, we will extract the variable 'sowc_demographics___life-expectancy-at-birth-years_2021-0', as well as the ISO3 country codes.

```
# Extract the variables
life_expectancy_information <- country_indicators %>%
  select("sowc_demographics__life-expectancy-at-birth-years_2021-0",
         country_code_iso3) %>%
  rename(life_expectancy_at_birth =
           "sowc_demographics__life-expectancy-at-birth-years_2021-0")

glimpse(life_expectancy_information)
```

```
## Rows: 218
## Columns: 2
## $ life_expectancy_at_birth <dbl> 61.9824, 76.4626, 76.3767, 80.3684, 61.6434, ~
## $ country_code_iso3        <chr> "AFG", "ALB", "DZA", "AND", "AGO", "AIA", "AT~
```

Now we're going to merge the two data sets into one using the variable that they have in common: 'country_code_iso3'. Since the observations are matched to each country code, merging using this variable will allow the data from both data sets to be compatible once merged.

```
data_research_question_2 <- merge(
  goal_4_information, life_expectancy_information, by="country_code_iso3")
```

```r
glimpse(data_research_question_2)
```

```
## Rows: 193
## Columns: 4
## $ country_code_iso3        <chr> "AFG", "AGO", "ALB", "AND", "ARE", "ARG", "AR~
## $ goal_4_status            <chr> "Major challenges", "Major challenges", "Chal~
## $ country_label            <chr> "Afghanistan", "Angola", "Albania", "Andorra"~
## $ life_expectancy_at_birth <dbl> 61.9824, 61.6434, 76.4626, 80.3684, 78.7104, ~
```

Finally, we will split the countries into 4 groups each representing a different level of SDG Goal 4 achievement. These are, "SDG Achieved", "Challenges remain", "Major challenges remain" and "Significant challenges remain".

```r
achieved <- data_research_question_2 %>% filter(
  goal_4_status == "SDG achieved")

challenges_remain <- data_research_question_2 %>% filter(
  goal_4_status == "Challenges remain")

major_challenges_remain <- data_research_question_2 %>% filter(
  goal_4_status == "Major challenges")

significant_challenges_remain <- data_research_question_2 %>% filter(
  goal_4_status == "Significant challenges")
```

We can now use these data sets to perform the bootstrapping process. In the next section, we will construct a bootstrapping sample distribution for each education level group. This will enable us to find an estimate of the median 'life_expectancy_at_birth' for each group to therefore make inferences ad hopefully answer the research question.

c) Bootstrapping.

Assuming that our random sample is representative of the full population, we will generate random samples by randomly sampling with replacement from the observed sample (the data under the 'life_expectancy_at_birth' variable).

We decided to find an estimate for the median, as the mean might be very sensitive to possible outliers.

To generate resamples, we used a for loop and the sample() function, and we sampled with replacement, so replace=TRUE. For each of the four groups created, SDG4 achieved, challenges remain, major challenges remain and significant challenges we performed a simulation of 10000 repetitions, with sample size 26, 56, 48 and 57 respectively.

Then, the test statistics from the sample were stored in tibbles, which enabled us to plot easily the sample distribution for the estimated median 'life_expectancy_at_birth' for each of the groups in separate histograms.

Then, we estimated a potential range of values for the true median 'life_expectancy_at_birth' using a 90% confidence interval. This allowed us to be 90% sure that the true median 'life_expectancy_at_birth' for a given group is in that range.

We will proceed with the bootstrapping:

Education level group: **SDG Achieved**

```r
# Bootstrapping for SDG Achieved
n1 <- 26
repetitions <- 10000
```

```
sim1 <- rep(NA, repetitions)
set.seed(SEED %% 200)

for (i in 1:repetitions)
{
  new_sim1 <- sample(na.omit(achieved$life_expectancy_at_birth),
                     size = n1, replace=TRUE)
  sim_median1 <- median(new_sim1)
  sim1[i] <- sim_median1
}
sim1 <- tibble(median = sim1)
```

Next, we calculated a 90% confidence interval:

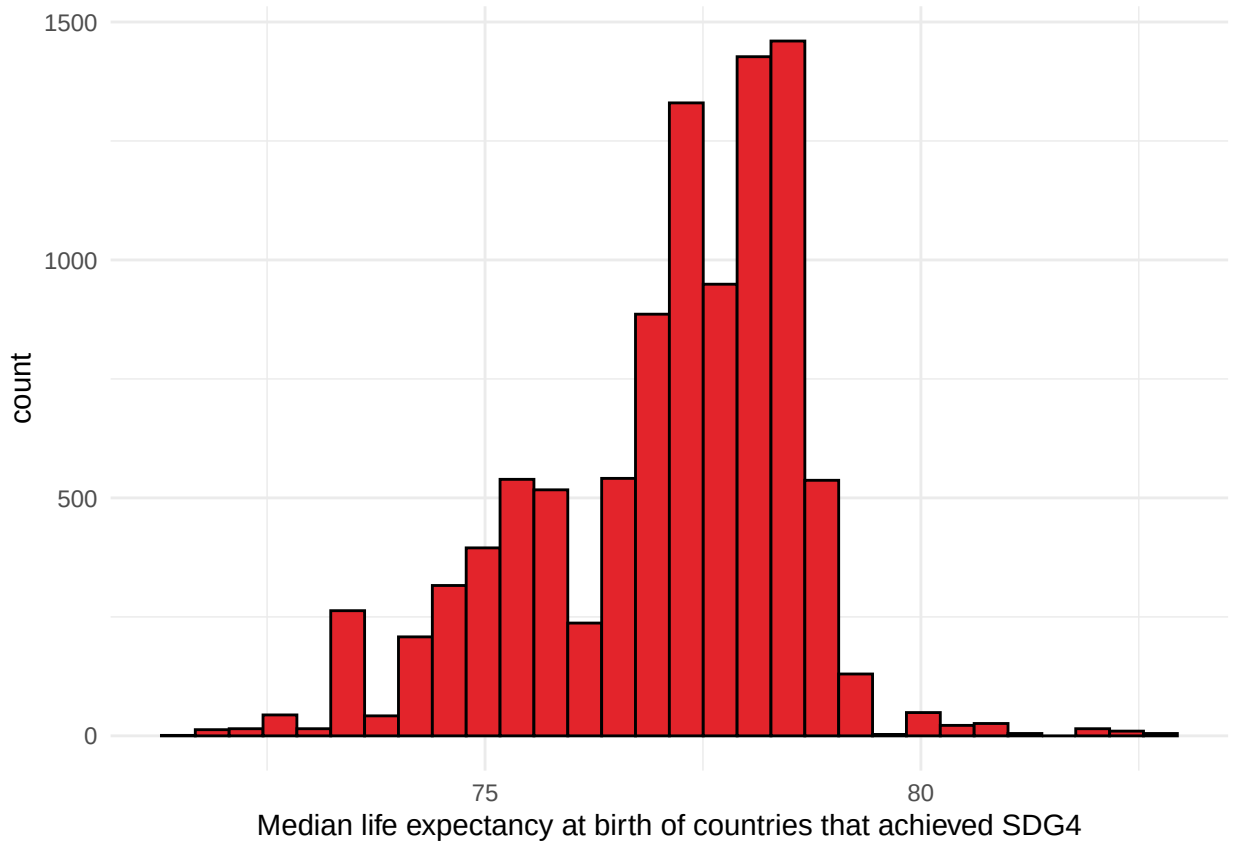
```
quantile(sim1$median, c(0.05, 0.95))
```

```
##       5%      95%
## 74.13025 78.71290
```

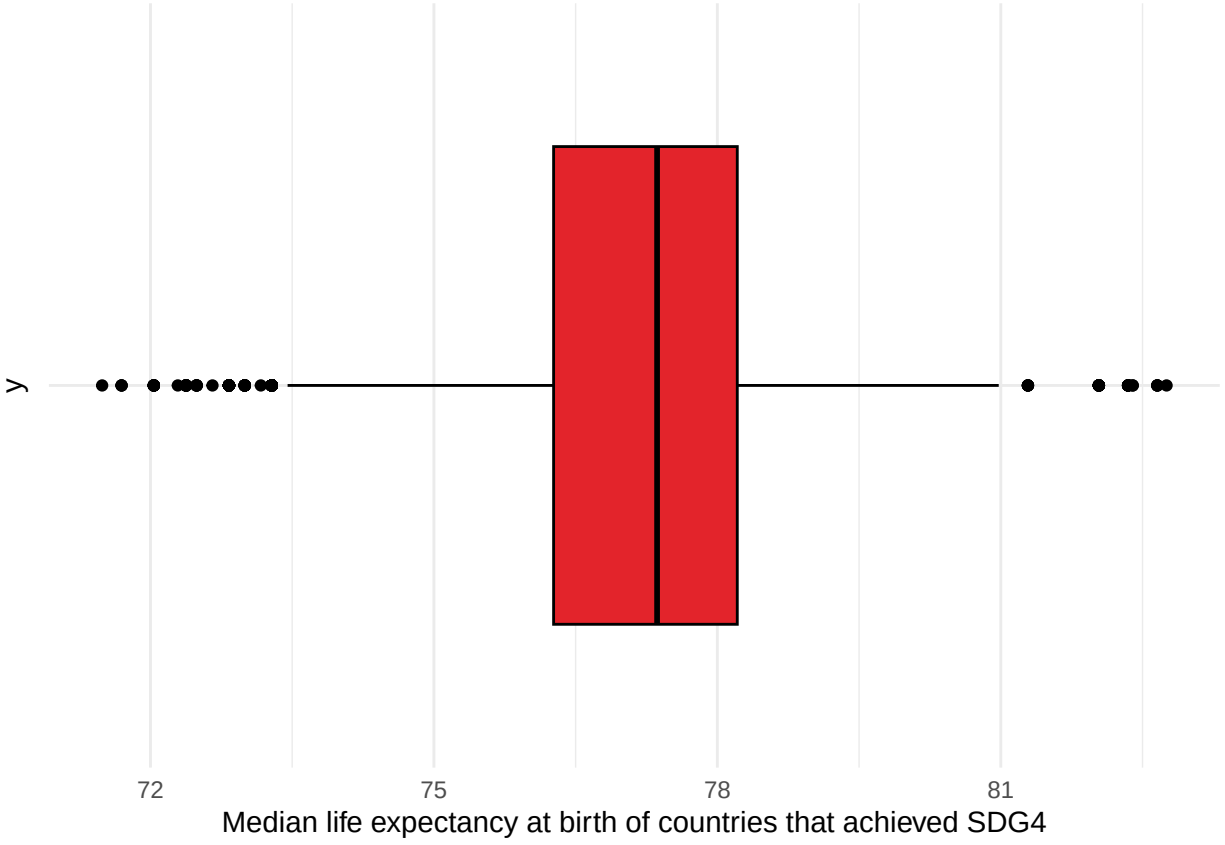
From this we know that the true median life expectancy at birth for countries who have already achieved SDG 4 is between 74.13 and 78.71.

Visualizations:

```
hist1 <- ggplot(data = sim1, aes(x = median))+
  geom_histogram(colour = "black", fill = "#e3242b", bins = 30) +
  labs(x = "Median life expectancy at birth of countries that achieved SDG4") +
  theme_minimal()


boxplot1 <- ggplot(data = sim1, aes(x = median, y = "")) +
  geom_boxplot(colour="black", fill="#e3242b") +
  labs(x = "Median life expectancy at birth of countries that achieved SDG4") +
  theme_minimal()

hist1
```


1500
1000
500
0
count
75
80
Median life expectancy at birth of countries that achieved SDG4

```
boxplot1
```



Education level group: **SDG Challenges remain**

```
# Bootstrapping for SDG Challenges remain
n2 <- 56
repetitions <- 10000
sim2 <- rep(NA, repetitions)
set.seed(SEED %% 200)

for (i in 1:repetitions)
{
  new_sim2 <- sample(na.omit(challenges_remain$life_expectancy_at_birth),
                     size = n2, replace=TRUE)
  sim_median2 <- median(new_sim2)
  sim2[i] <- sim_median2
}
sim2 <- tibble(median = sim2)
```

Next, we calculated a 90% confidence interval:

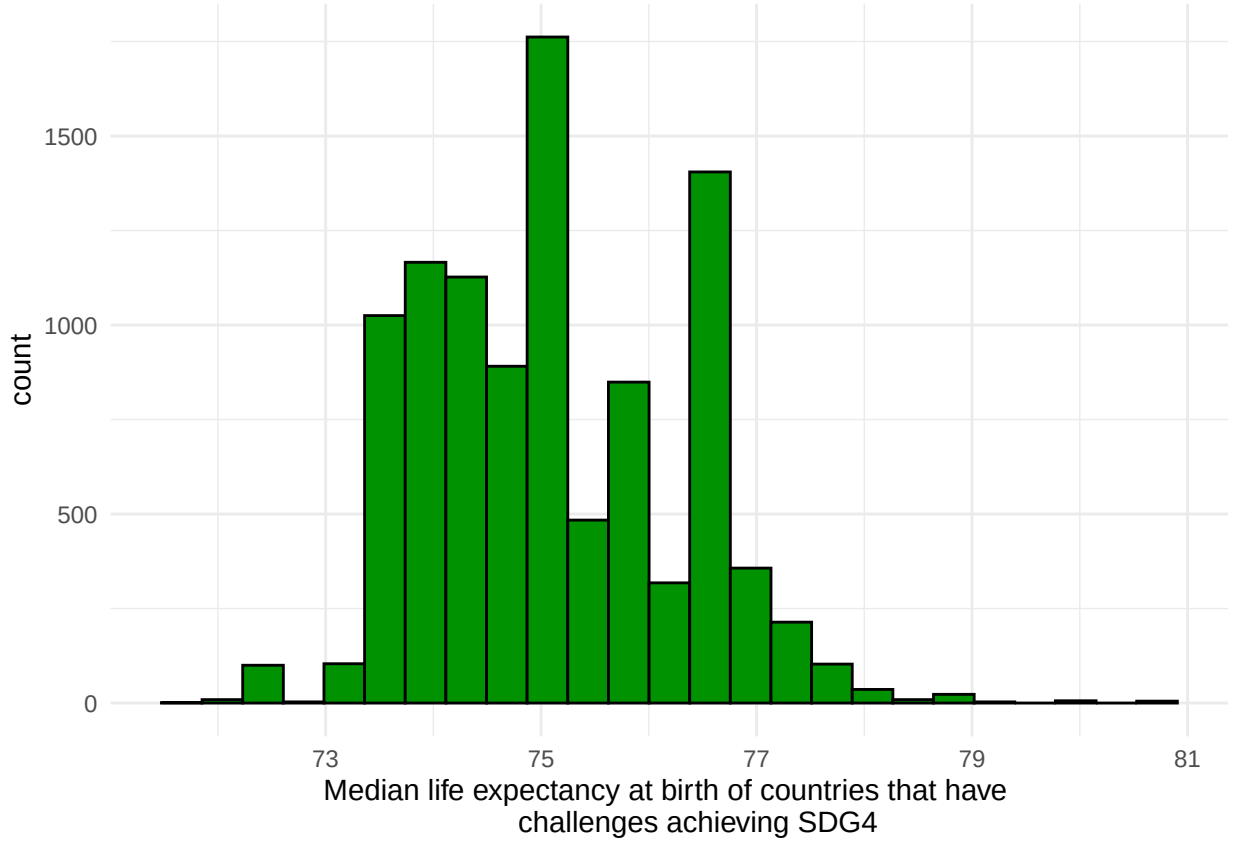
```
quantile(sim2$median, c(0.05, 0.95))
```

```
##       5%      95%
## 73.67645 77.06710
```

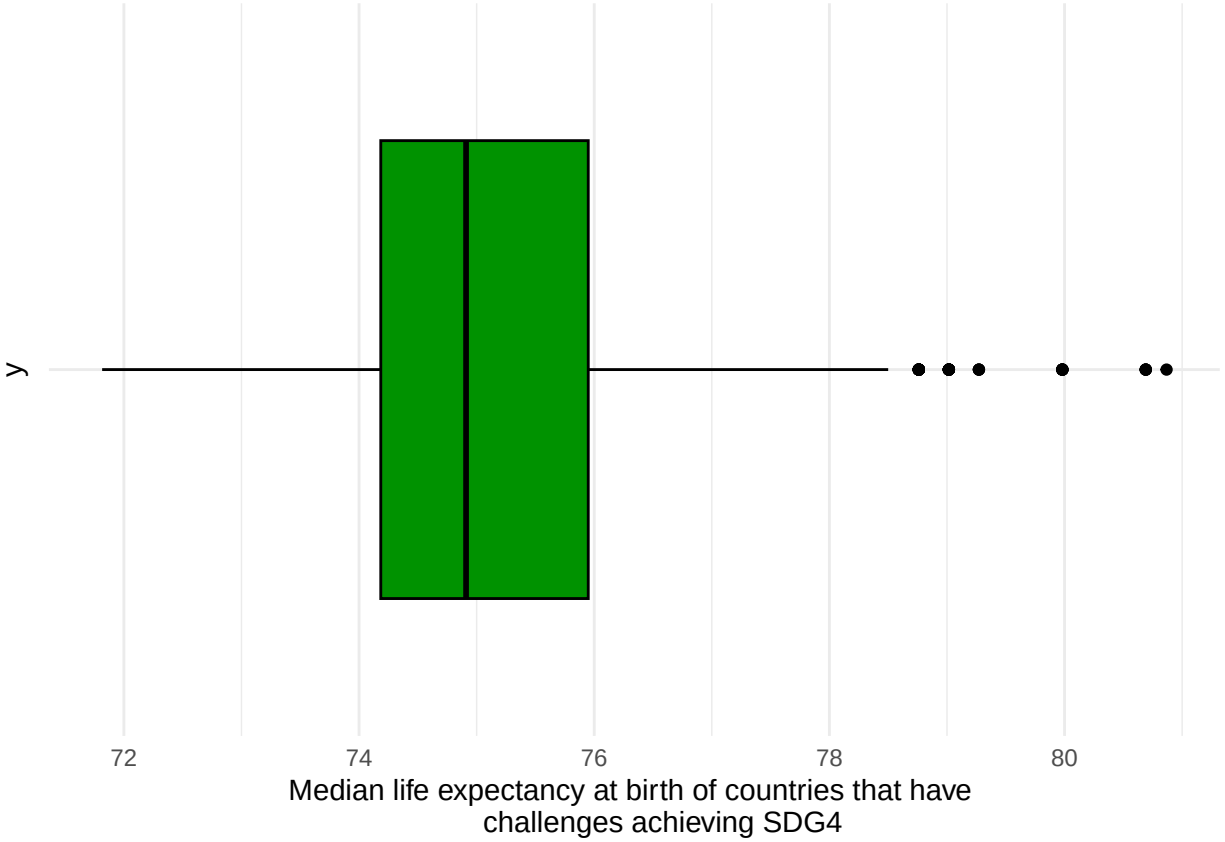
From this we know that the true median life expectancy at birth for countries who still have challenges regarding SDG 4 is between 73.67 and 77.07.

Visualizations

```r
hist2 <- ggplot(data = sim2, aes(x = median))+
  geom_histogram(colour = "black", fill = "#009200", bins = 25) +
  labs(x = "Median life expectancy at birth of countries that have
       challenges achieving SDG4") + theme_minimal()

boxplot2 <- ggplot(data = sim2, aes(x = median, y = "")) +
  geom_boxplot(colour="black", fill="#009200") +
  labs(x = "Median life expectancy at birth of countries that have
       challenges achieving SDG4") + theme_minimal()

hist2
```




```r
boxplot2
```



Education level group: **SDG Major Challenges remain**

```
# Bootstrapping for SDG Major Challenges remain
n3 <- 48
repetitions <- 10000
sim3 <- rep(NA, repetitions)
set.seed(SEED %% 200)

for (i in 1:repetitions)
{
  new_sim3 <- sample(na.omit(major_challenges_remain$life_expectancy_at_birth),
                     size = n3, replace=TRUE)
  sim_median3 <- median(new_sim3)
  sim3[i] <- sim_median3
}
sim3 <- tibble(median = sim3)
```

Next, we calculated a 90% confidence interval:

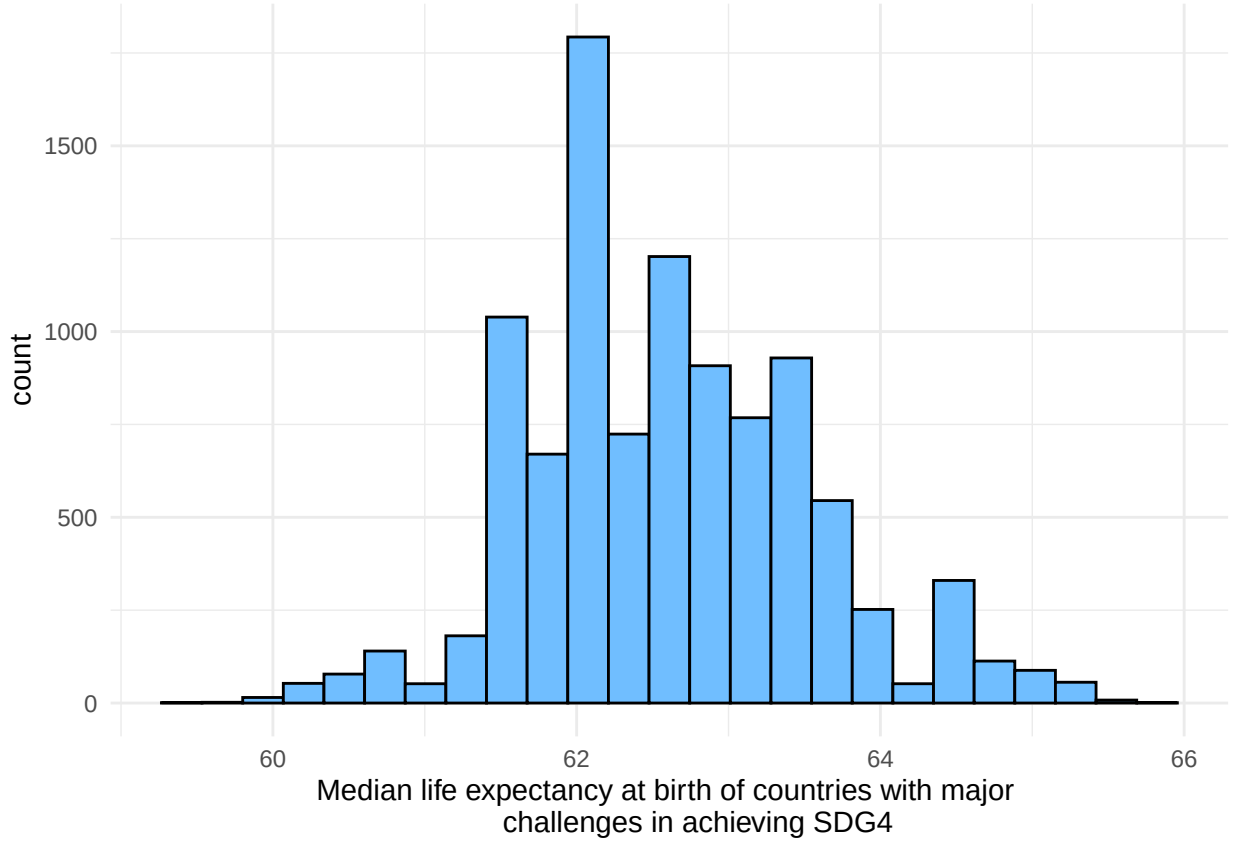
```
quantile(sim3$median, c(0.05, 0.95))
```

```
##       5%      95%
## 61.20495 64.36360
```

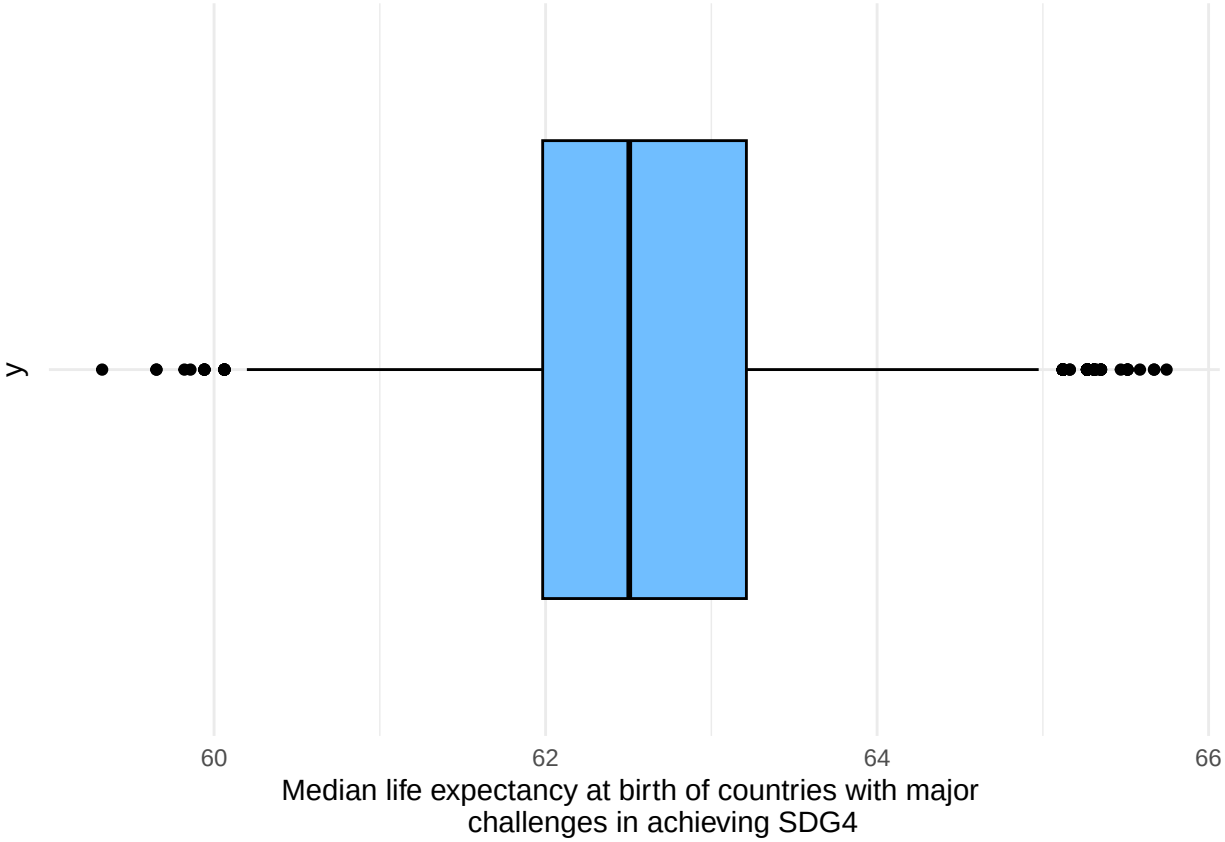
From this we know that the true median life expectancy at birth for countries who still have major challenges regarding SDG 4 is between 61.20 and 64.36

Visualizations

```r
hist3 <- ggplot(data = sim3, aes(x = median))+
  geom_histogram(colour = "black", fill = "#70beff", bins = 25) +
  labs(x = "Median life expectancy at birth of countries with major
       challenges in achieving SDG4") + theme_minimal()

boxplot3 <- ggplot(data = sim3, aes(x = median, y = "")) +
  geom_boxplot(colour="black", fill="#70beff") +
  labs(x = "Median life expectancy at birth of countries with major
       challenges in achieving SDG4") + theme_minimal()

hist3
```




```r
boxplot3
```



Education level group: **SDG Significant Challenges remain**

```
# Bootstrapping for SDG Significant Challenges remain
n4  <- 57
repetitions <- 10000
sim4 <- rep(NA, repetitions)
set.seed(SEED %% 200)

for (i in 1:repetitions)
{
  new_sim4 <- sample(na.omit(significant_challenges_remain$life_expectancy_at_birth),
                     size = n4, replace=TRUE)
  sim_median4 <- median(new_sim4)
  sim4[i] <- sim_median4
}
sim4 <- tibble(median = sim4)
```

Next, we calculated a 90% confidence interval:

```
quantile(sim4$median, c(0.05, 0.95))
```

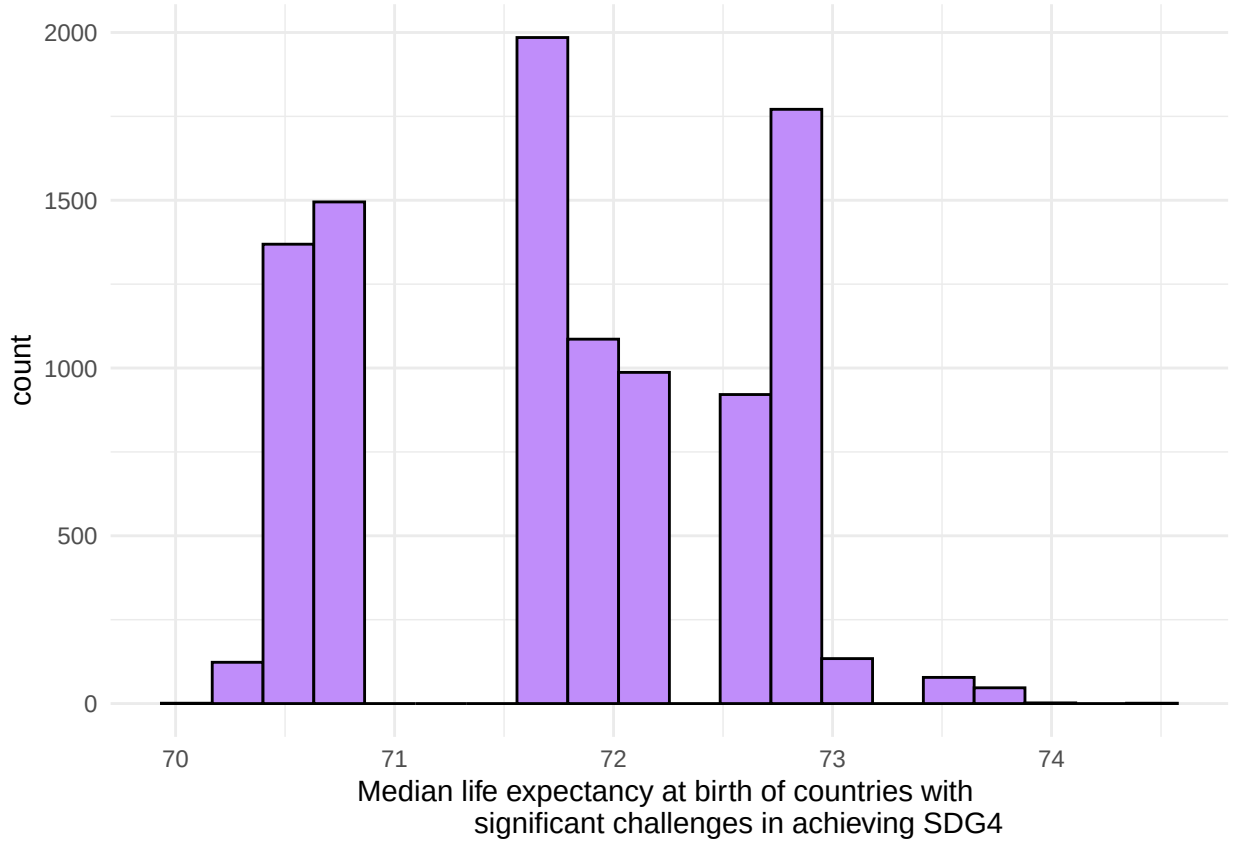
```
##      5%     95%
## 70.4697 72.8140
```

From this we know that the true median life expectancy at birth for countries who still have significant challenges regarding SDG 4 is between 70.47 and 72.81.

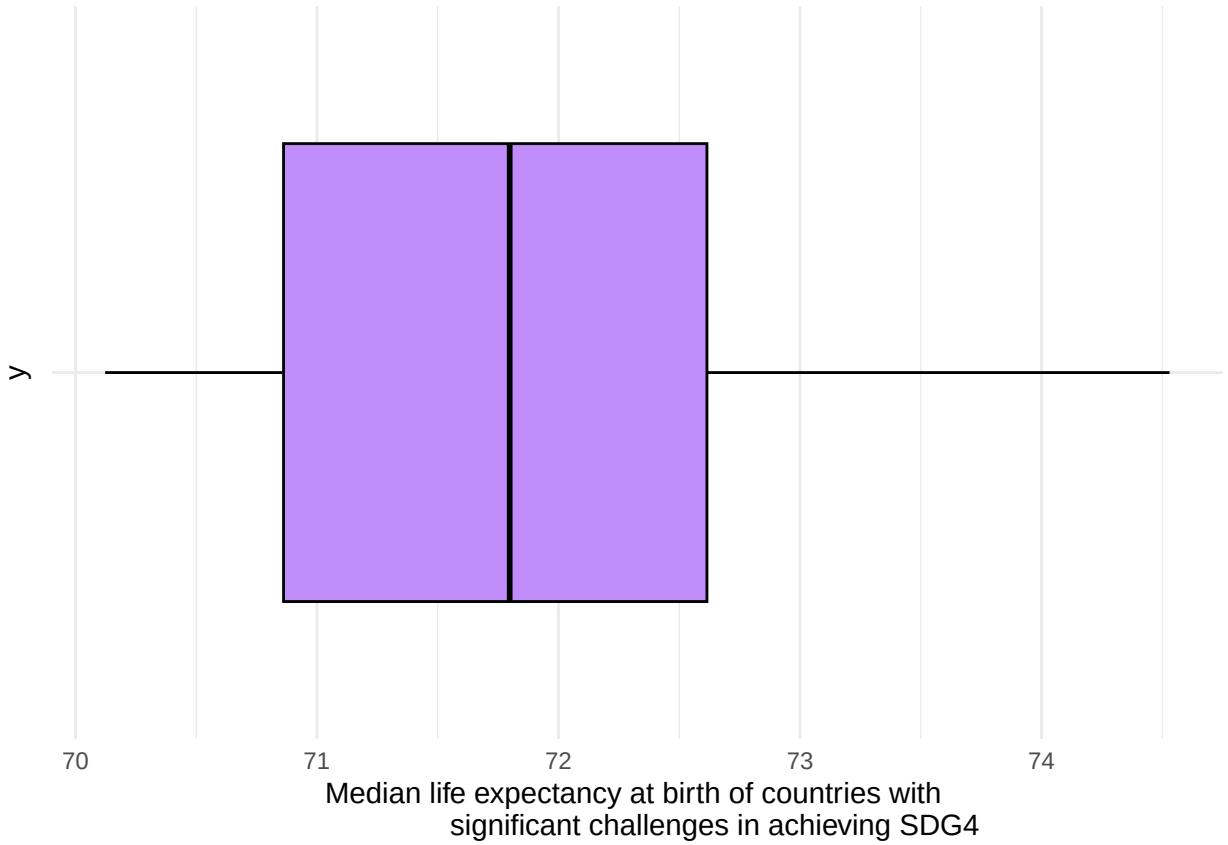
Visualizations

```r
hist4 <- ggplot(data = sim4, aes(x = median))+
            geom_histogram(colour = "black",
                 fill = "#c08dfa",
                 bins = 20) +
            labs(x = "Median life expectancy at birth of countries with
                 significant challenges in achieving SDG4") +
  theme_minimal()

boxplot4 <- ggplot(data = sim4, aes(x = median, y = "")) +
              geom_boxplot(colour="black", fill="#c08dfa") +
              labs(x = "Median life expectancy at birth of countries with
                   significant challenges in achieving SDG4") +
  theme_minimal()
```

```r
hist4
```

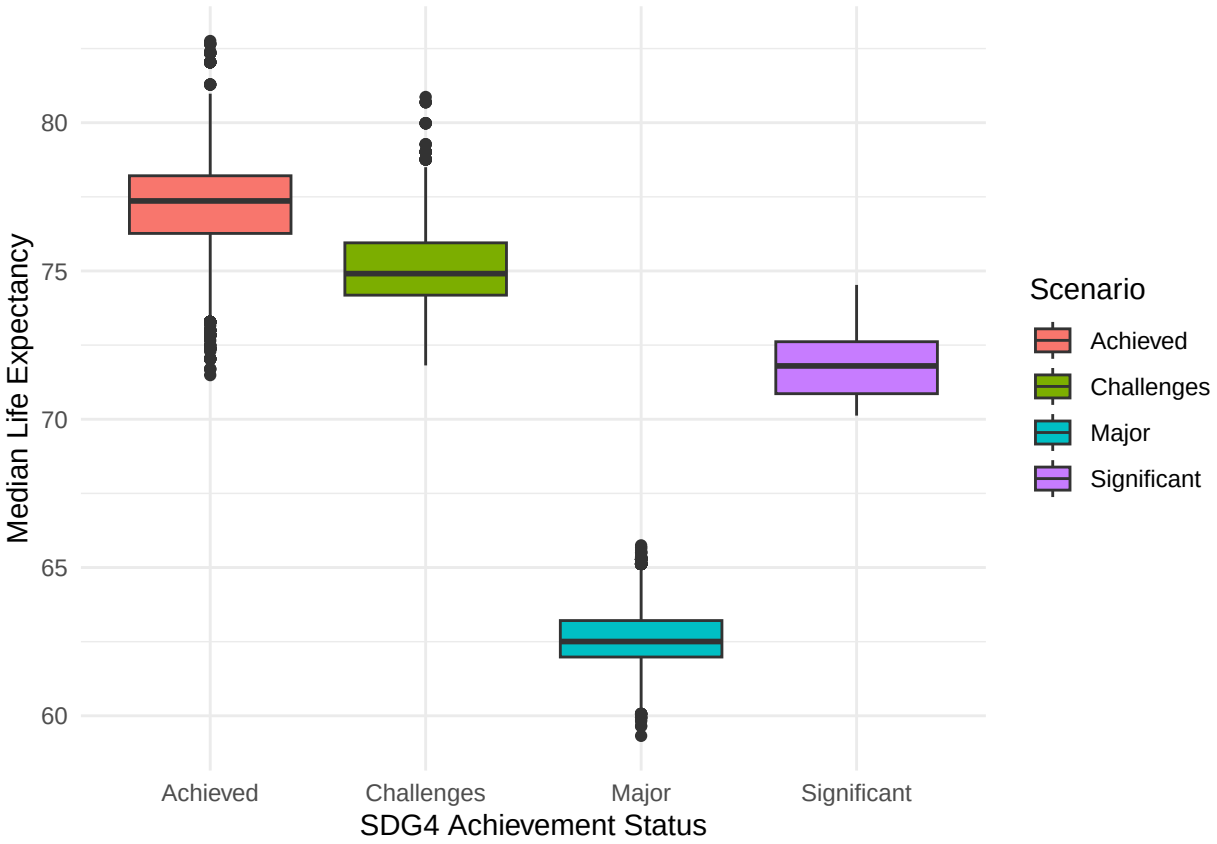



```r
boxplot4
```



To finalize, we produced a graph with all 4 box plots together, which allows us to compare the results of the bootstrapping for each group easily:

```
combined_data <- rbind(mutate(sim1, Scenario = "Achieved"),
                       mutate(sim2, Scenario = "Challenges"),
                       mutate(sim3, Scenario = "Major"),
                       mutate(sim4, Scenario = "Significant"))

ggplot(combined_data, aes(x = Scenario, y = median, fill = Scenario)) +
  geom_boxplot() +
  labs(x = "SDG4 Achievement Status",
       y = "Median Life Expectancy") +
  theme_minimal()
```



d) Analysis.

In relation to the histograms for each of the SDG4 status categories, the shape of the bootstrapping distributions appears to be uniform for countries which have achieved, have challenges remaining and have major challenges remaining towards SDG4, on the contrary, the shape of the bootstrapping distribution for countries which have significant challenges in achieving SDG4 appears to be multimodal. The values for SDG4 achieved, challenges remain and major challenges categories appears to be concentrated around the range 73-78, 73.5-76, 61.5-63.5 respectively, whereas, in the case of significant challenges remain category there appears to be 3 separate clusters of values. Additionally, in terms of centre they are, around 78, 75, 62 and 71.5 respectively.

Regarding the box-plots of each SDG4 status category, the median life expectancy of countries that have achieved SDG goal 4 is around 77, the next is by countries that have challenges remaining with a median life expectancy of around 75, which is followed by countries that have significant challenges at around 72 and lastly countries with major challenges at 62.5. The countries that have achieved SDG goal 4 appear to have the greatest number of extreme outliers when compared to the other 3 categories, conversely, the countries with significant challenges have no extreme outliers. Additionally, the interquartile range of countries that have achieved, have challenges remaining and have significant challenges remaining towards SDG4 are about the same, whereas, countries with major challenges in achieving SDG4 has the smallest IQR indicating that the spread of values is the lowest compared to the aforementioned categories.

From our confidence intervals we obtained the following ranges for true life expectancy at birth median:

SDG Achieved = 74.1 - 78.7

SDG Challenges Remain = 73.7 - 77.1

SDG Major Challenges Remain = 61.2 - 64.4

SDG Significant Challenges Remain = 70.5 - 72.8

As we can observe, the estimated median values for all for groups lie between 61.2 and 78.7, which means we have quite a wide range of values. We found that the groups where SDG 4 was already achieved and where there were still challenges had very similar ranges, with the group for SDG Achieved having a slightly higher end of the confidence interval.

The lowest range value is found in the SDG Major Challenges Remain group (61.2), on the other hand, countries in the group SDG Achieved had the highest possible value for the true life expectancy at birth (78.7). Furthermore, we see the largest range of possible for values for SDG Achieved (4.6), which means that the estimates from the sampling distribution were more spread apart, so there was a larger variability for the test statistics in this group. However, most of the ranges for the other groups also lie around 3.5, except for the group where significant challenges remain, where the range is only of 2.3 points, indicating the lowest variability in values.

We also observe a big difference when going from the two groups with higher advance in SDG 4 to the group where major challenges remain, as the lower end of the range of values drops from 73.6 to 60.7, more than 10 points. This was quite surprising for us, however, the most surprising finding was that the group where significant challenges remain, had a higher range of values for the median life expectancy at birth than the group where major challenges remain. The difference between the groups is almost 10 points on the lower end and 6 points on the higher end. This introduces the possibility of confounding factors which can skew the results of this study. These include, but are not limited to, the availability of healthcare to the general public, lifestyle choices and income disparity. One limitation of this study is the way SDG4 status is classified, this due to the fact that some countries that have significant challenges in achieving SDG4 have a higher goal 4 score than countries that have achieved this goal.

e) Conclusion.

In conclusion, the bootstrapping procedure allowed us to observe that life expectancy at birth does vary depending on the education level of different countries, which indicates a possible association between these two variables. 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.

V. Research Question 3

a) Introduction and Motivations.

Are there any relationships between quality education (goal 4) and child labour?

It's been widely observed that engaging in child labor could possibly be affecting quality education for children as "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.). As advised by Humans Right Watch, governments must address the link between child labor and education, in order to achieve SDG4, quality education. Our research question aims to explore this link further, considering that finding a possible relationship between achieving SDG4 and child labour could help unveil a way of working towards SDG4 by reducing child labor through policy.

b) Data Exploration

Before performing the test the data was wrangled to have a more manageable and organized data set. This section describes the process of data wrangling the data went through previous to the testing.

To perform the test we will use two numerical variables: 'goal_4_score' to measure quality education, and 'sowc_child-protection___child-labour-h-2013-2021-r_total' to measure child labor.

Because of this, we'll extract the variable 'goal_4_score' and create a data set containing only this information and the country codes. We also renamed variables `Goal 4 Score` and `Country Code ISO3`:

```
# Extract variables from sdg
goal_4_scores <- sdg %>%
  select(goal_4_score, country_code_iso3, country_label)

glimpse(goal_4_scores)
```

```
## Rows: 206
## Columns: 3
## $ goal_4_score      <dbl> 97.2, 99.8, 99.3, 97.2, 97.9, 99.6, 98.0, 93.9, 97.5~
## $ country_code_iso3 <chr> "FIN", "SWE", "DNK", "DEU", "AUT", "FRA", "NOR", "CZ~
## $ country_label     <chr> "Finland", "Sweden", "Denmark", "Germany", "Austria"~
```

Next, we'll extract the 'sowc_child-protection___child-labour-h-2013-2021-r_total' variable, as well as this same indicator for male and female from the country_indicators data set. We also renamed the variables to make them easier to use in the future:

```
# Extract variables from country_indicators
child_labor_information <- country_indicators %>%
  select("sowc_child-protection__child-labour-h-2013-2021-r_total",
         "sowc_child-protection__child-labour-h-2013-2021-r_male",
         "sowc_child-protection__child-labour-h-2013-2021-r_female",
         country_code_iso3) %>%
  rename(total_child_labor = "sowc_child-protection__child-labour-h-2013-2021-r_total",
         male_child_labor = "sowc_child-protection__child-labour-h-2013-2021-r_male",
         female_child_labor = "sowc_child-protection__child-labour-h-2013-2021-r_female")

glimpse(child_labor_information)
```

```
## Rows: 218
## Columns: 4
## $ total_child_labor  <dbl> 13.000, 3.300, 2.500, NA, 18.724, NA, NA, NA, 4.100~
## $ male_child_labor   <dbl> 14.200, 3.600, 2.900, NA, 16.600, NA, NA, NA, 5.000~
## $ female_child_labor <dbl> 11.700, 3.000, 2.000, NA, 19.870, NA, NA, NA, 3.000~
```

```
## $ country_code_iso3  <chr> "AFG", "ALB", "DZA", "AND", "AGO", "AIA", "ATG", "A~
```

```
child_labor_information %>% select(female_child_labor) %>% na.omit() %>% summarize(min = min(female_chi
```

```
## # A tibble: 1 x 2
##     min   max
##   <dbl> <dbl>
## 1  0.12  39.9
```

Next, we will load the country names and their development levels from the country_codes.csv:

```
# Create a copy of country_codes to use for this research question
countries_and_dev <- country_codes
glimpse(countries_and_dev)
```

```
## Rows: 298
## Columns: 4
## $ country_code_iso3 <chr> "DZA", "EGY", "LBY", "MAR", "SDN", "TUN", "ESH", "IO~
## $ region            <chr> "Africa", "Africa", "Africa", "Africa", "Africa", "A~
## $ country_label     <chr> "Algeria", "Egypt", "Libya", "Morocco", "Sudan", "Tu~
## $ development_level <chr> "Developing", "Developing", "Developing", "Developin~
```

Now we're going to merge the two data sets into one using the country_number variable that they have in common.

```
# Merge goal_4_scores and child_labor_information
data_research_question_3 <- inner_join(goal_4_scores, child_labor_information,
                                       by="country_code_iso3")
data_research_question_3 <- inner_join(x=data_research_question_3,
                                       countries_and_dev, by='country_code_iso3')
glimpse(data_research_question_3)
```

```
## Rows: 195
## Columns: 9
## $ goal_4_score       <dbl> 97.2, 99.8, 99.3, 97.2, 97.9, 99.6, 98.0, 93.9, 97.~
## $ country_code_iso3  <chr> "FIN", "SWE", "DNK", "DEU", "AUT", "FRA", "NOR", "C~
## $ country_label.x    <chr> "Finland", "Sweden", "Denmark", "Germany", "Austria~
## $ total_child_labor  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ male_child_labor   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ female_child_labor <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ region             <chr> "Europe", "Europe", "Europe", "Europe", "Europe", "~
## $ country_label.y    <chr> "Finland", "Sweden", "Denmark", "Germany", "Austria~
## $ development_level  <chr> "Developed", "Developed", "Developed", "Developed",~
```

Finally, we will remove the missing entries (entries with NA) and reorganize the variables a little bit:

```
# Remove missing entries
data_research_question_3 <- na.omit(data_research_question_3)

# Select necessary variables and create new variable: gender_ratio
data_research_question_3 <- data_research_question_3 %>% select(
  development_level, goal_4_score, total_child_labor, male_child_labor,
  female_child_labor, country_code_iso3) %>%
  mutate(gender_ratio = male_child_labor / female_child_labor)

glimpse(data_research_question_3)
```

```
## Rows: 84
## Columns: 7
```

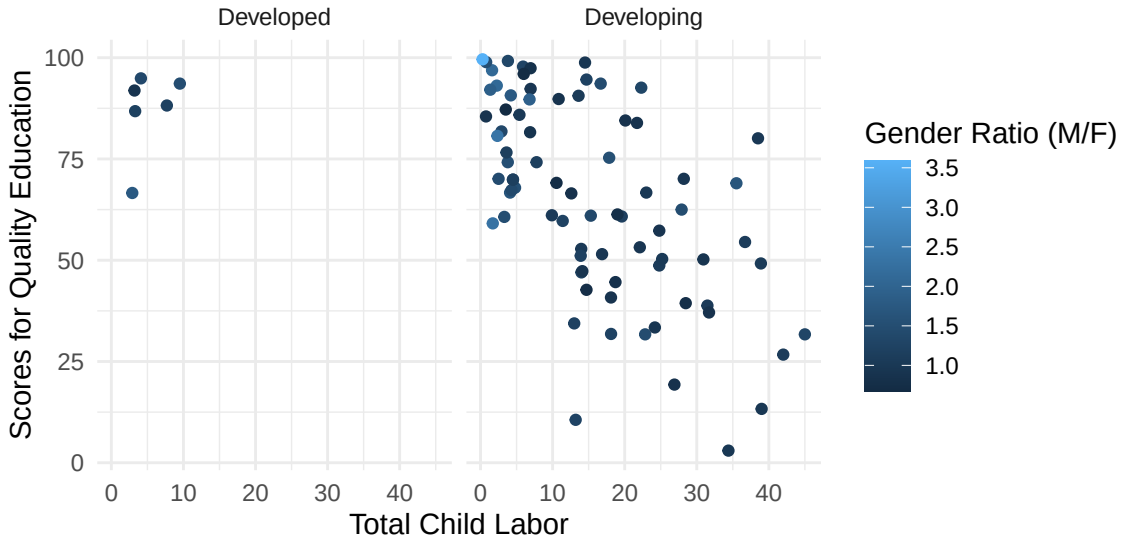
```
## $ development_level  <chr> "Developing", "Developing", "Developed", "Developed~
## $ goal_4_score       <dbl> 97.8, 90.7, 94.9, 93.6, 91.9, 96.9, 92.6, 85.9, 99.~
## $ total_child_labor  <dbl> 5.900, 4.200, 4.100, 9.500, 3.208, 1.600, 22.300, 5~
## $ male_child_labor   <dbl> 6.700, 5.300, 4.700, 11.200, 3.052, 2.100, 25.100, ~
## $ female_child_labor <dbl> 5.200, 3.000, 3.400, 7.500, 3.371, 1.000, 19.100, 5~
## $ country_code_iso3  <chr> "CHL", "URY", "BLR", "SRB", "UKR", "GEO", "KGZ", "B~
## $ gender_ratio       <dbl> 1.2884615, 1.7666667, 1.3823529, 1.4933333, 0.90536~
```

c) Linear Regression.

i) Setting Up

First, we briefly explored the correlation between 'goal_4_score' and 'total_child_labor' for developed and developing countries by producing a pair of scatterplots:

```
data_research_question_3 %>% ggplot(aes(x = total_child_labor,
                                        y = goal_4_score, col = gender_ratio)) +
  geom_point() +
  labs(x = 'Total Child Labor', y = 'Scores for Quality Education',
       col = 'Gender Ratio (M/F)') +
  facet_wrap(~development_level) +
  theme_minimal()
```



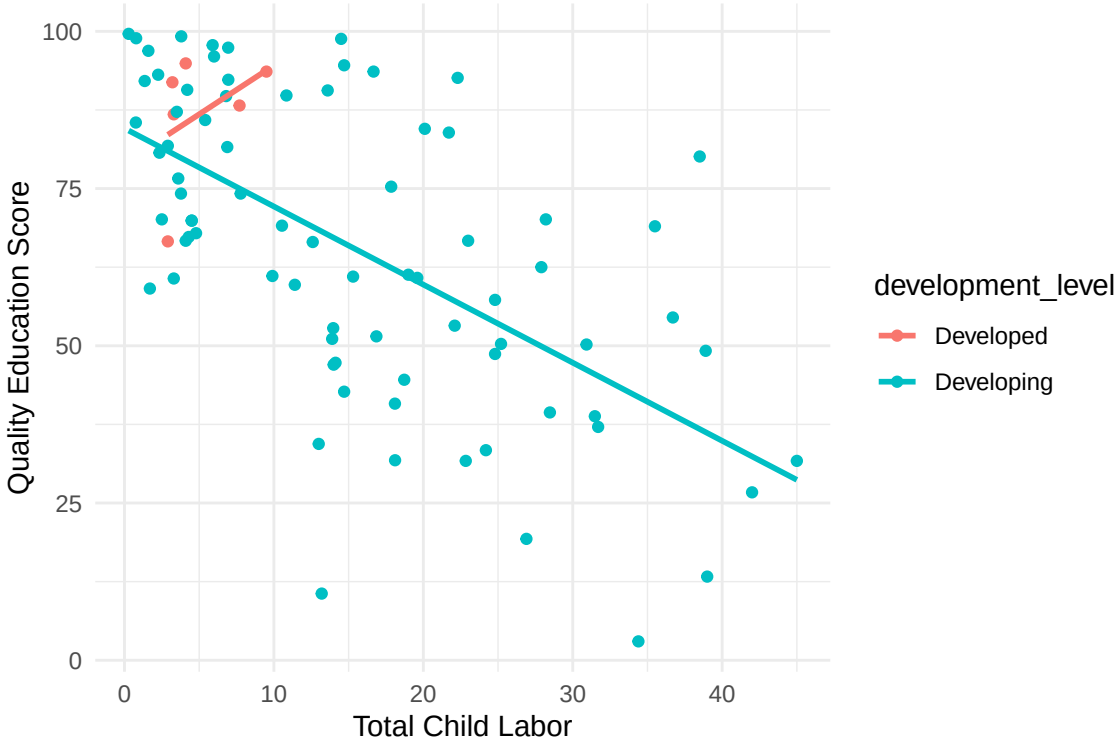
It is noticeable that developed countries have much fewer points represented in the scatterplot than developing countries. It is because the data set didn't have information about child labor in the original dataset for most of these, and thus were omitted during the data wrangling process. From the limited data of developed countries, we can see that all of them have relatively low rates of child labor (less than 10%), and most of them (5/6 points) have good scores on the quality of education. However, it is quite unclear to spot the association of those two goals in this group.

For the developing countries, it is clearer that the data has a linear form with a negative association between 'total child labor' and 'quality education score'. It seems like the values are not so concentrated but rather spreading out. We will investigate the association and strength further by using the linear regression method in the following steps.

We will try to fit predicted lines onto the scatterplot for each of the types of development level in order to have a clearer view of the correlation between the variables:

```r
# Fit a line onto the scatterplot
data_research_question_3 %>% ggplot(aes(x = total_child_labor,
                                        y = goal_4_score,
                                        col = development_level)) +
  geom_point() +
  labs(x = 'Total Child Labor', y = 'Quality Education Score') +
  geom_smooth(method = 'lm', se = FALSE) +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

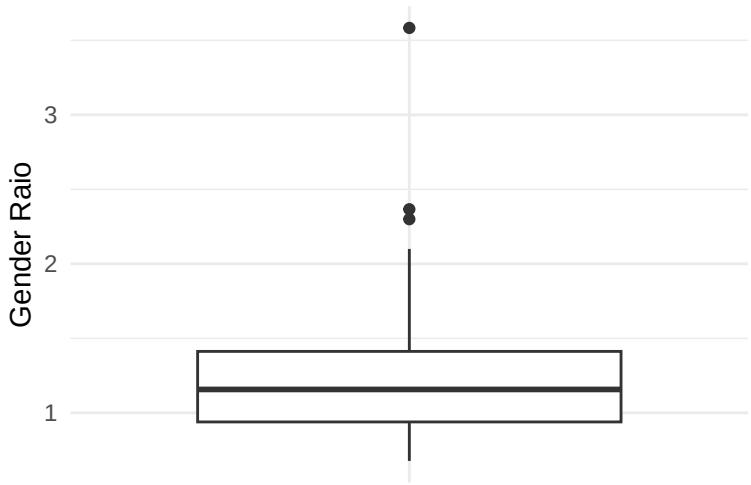


As expected, there is a negative association, due to a downward slope, for the total child labor and quality education score in developing countries. On the other hand, it is surprising that the predicted line of best fit for developed countries has an upward slope, implying that there might be an increasing trend: the more child labor, the higher the quality of education scores. However, with that being said, it is reasonable for the linear regression model to predict such an outcome, since due to the limited data, it's harder for to make better predictions, and we cannot trust it entirely.

Therefore, with an incredibly unequal number in developed and developing countries this wrangled data, it is not very beneficial to compare the correlation between those two goals by development level.

Let's take a closer look into another factor we mentioned earlier in the first graph, the gender ratio of children labor:

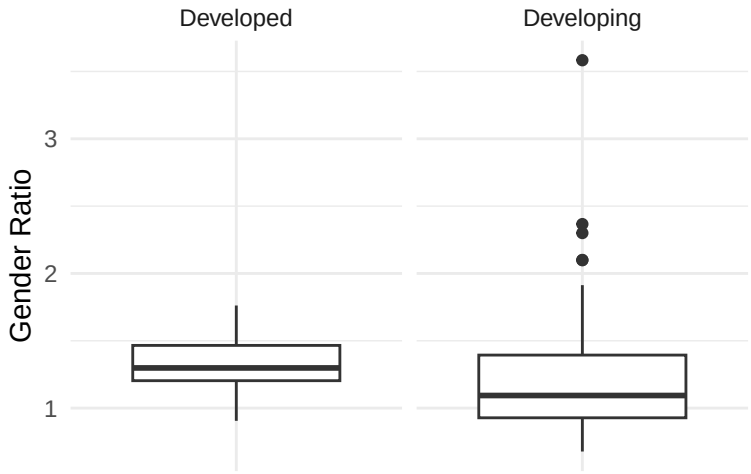
```r
data_research_question_3 %>% ggplot(aes(x = '', y = gender_ratio)) +
  geom_boxplot() +
  labs(x='', y='Gender Raio') +
  theme_minimal()
```



We can see that the gender ratio, the number of male children labour over the number of female children labour, spreads from around 0.7 to 2.3, with the most frequent lying on the IQR of about 0.9 - 1.4. There are noticeable outliers which have more than twice the number of male to female.There is even an outlier with the gender ratio being 3:1 Moreover, since the boxplot is right-skewed, it seems that male children are more likely to be involved in child labor than female children.

Next, we want to see if there are differences in the child labor gender ratio between developed and developing countries:

```
data_research_question_3 %>% ggplot(aes(x = '', y = gender_ratio)) +
  geom_boxplot() + facet_wrap(~development_level) +
  labs(x = '', y= 'Gender Ratio') +
  theme_minimal()
```



It is clear that the developing category has a larger spread and significant outliers of double, or even triple the percentage of male child labor to female child labor. It is also right-skewed, while the developed countries' boxplot is more symmetrical. However, it seems like the median gender ratio of developing countries is more toward the balance (the point where the ratio is 1:1) than developed ones.

ii) Linear Regression Models

To further explore difference in the trend due to genders, we carried out a **univariate (single-variable) linear regression analysis** to predict the score for quality education 'goal_4_score' variable based on the 'total_child_labor' variable

```
model4_1 <- lm(goal_4_score ~ total_child_labor, data = data_research_question_3)
summary(model4_1)$coefficients
```

```
##                     Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)       85.728371  3.3392707 25.672783 5.890438e-41
## total_child_labor -1.277849  0.1775429 -7.197408 2.649947e-10
```

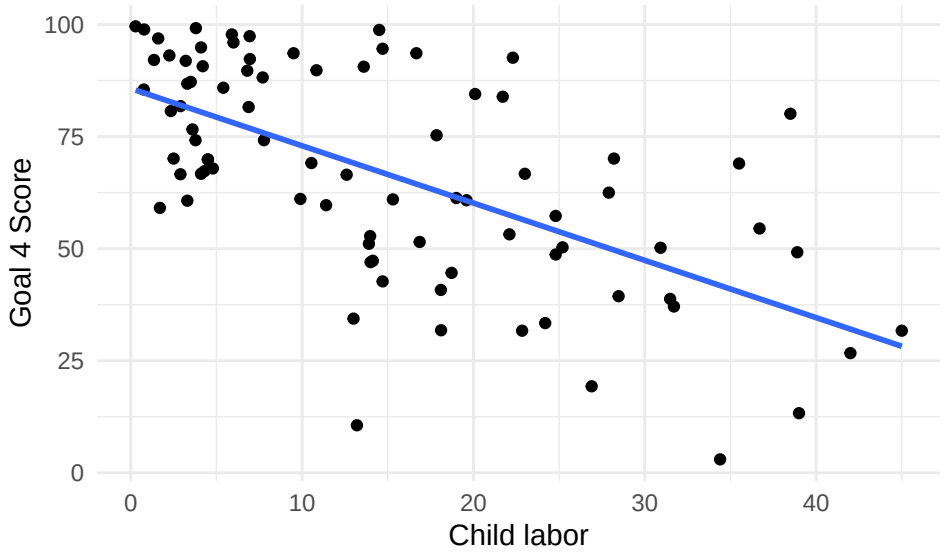
By extracting the coefficients, we obtain the linear regression equation: $\hat{y} = 85.33 - 1.28x$

Where $\hat{y}$ is the quality education and x is child labor. The y-intercept, 85.33 represents the predicted quality education when child labor = 0, so according to our model, countries with a proportion of 0 children engaged in child labor should have a Goal 4 Score of about 85.33. The slope represents the change in Goal 4 Score, for every additional change in unit of child labor, so essentially, for every additional point of child labor, Goal 4 Score, or level of education, goes down by 1.28. This means that child labor and Goal 4 Score are inversely related, because as child labor increases, Goal 4 Score decreases.

We can observe the inverse relationship between child labor and quality education in the graph below:

```
data_research_question_3 %>% ggplot(aes(x=total_child_labor, y=goal_4_score)) +
  geom_point() + geom_smooth(se=FALSE, method='lm') +
  labs(x = "Child labor", y = "Goal 4 Score") + theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Due to the assumptions of regression coefficient, we are assuming our data follows a linear form, i.e. has a linear relationship. The regression line coincides with our equation, and confirms that the variables are inversely related, as the slope is negative. Analyzing the points on the scatterplot we observe that the points have a negative association as the values of Goal 4 Score decrease as the values of Child labor increase. As for the strength, we can observe that the values aren't very concentrated, but we'll get a better sense of this by calculating the correlation coefficient:

```
cor(x = data_research_question_3$total_child_labor,
    y = data_research_question_3$goal_4_score)
```

```
## [1] -0.6222196
```

The correlation coefficient tells us the correlation is negative, because r<1, and the strength of the correlation is |-0.63|, which is 0.63. Although this value is not very big, it's not very small either, so we can say the variables have a somewhat strong association.

Next we're going to calculate R^2 to get some insights on the variability in Goal 4 Score, quality education, that our model captures:

```
summary(model4_1)$r.squared
```

```
## [1] 0.3871573
```

Here, R^2 is closer to 0 than to 1, so we can say that the regression model explains very little of the variability of Goal 4 Score with respect to Child labor.

Finally, we performed a hypothesis test for the slope of the regression line β_1 to further confirm that there's a correlation between the Child labor and Goal 4 score. The null hypothesis is that $\beta_1 = 0$, so there's no correlation between the variables, as the slope is 0. And the alternative hypothesis is that $\beta_1 \neq 0$, so there is a correlation between the variables.

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

We choose an alpha level of 0.05, and since the p-value = 7.44e-10, as retrieved from the table, we know that this p-value is way below 0.05, which leads us to reject the null hypothesis. Rejecting the null hypothesis implies that there is a correlation between the variables, Goal 4 Score and Child labor, answering our research question.

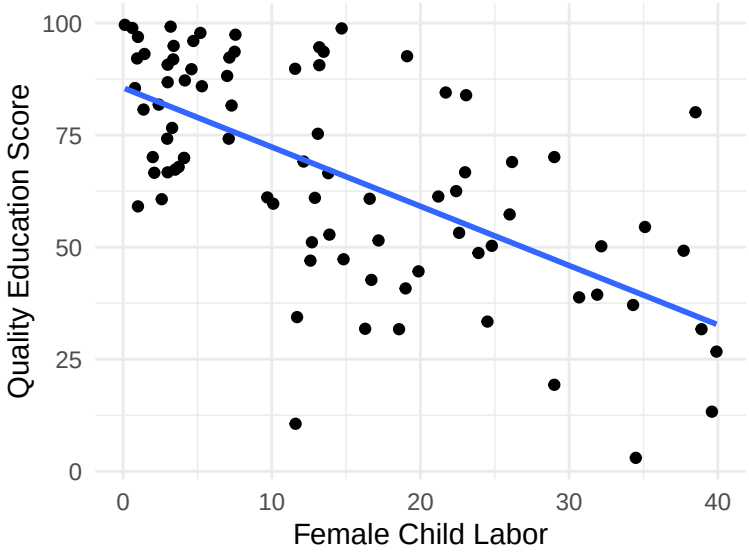
iii) One Step Further

Considering that we were also given child labor for male and female separately, we wanted to investigate whether there'd be any differences in the correlation between Goal 4 Score and Child labor for these two groups. Since this might be a confounding variable, we decided to investigate further.

First we graphed the data:

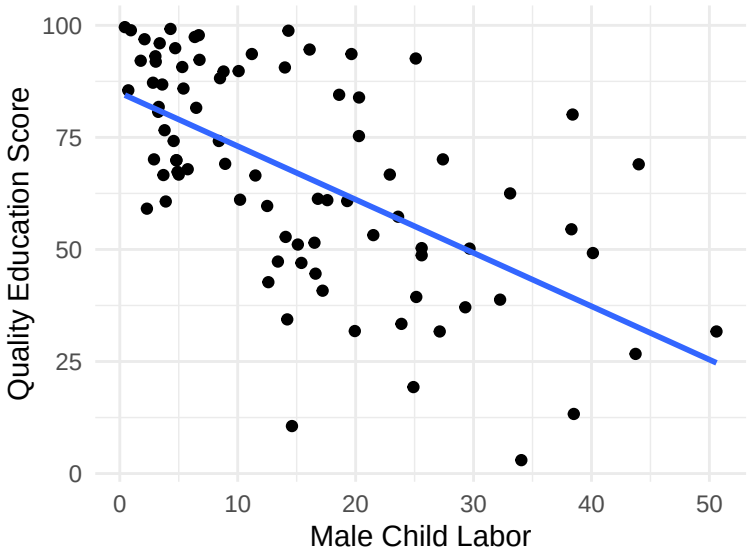
```
data_research_question_3 %>% ggplot(aes(x=female_child_labor, y=goal_4_score)) +
  geom_point() + geom_smooth(se=FALSE, method='lm') +
  labs(x = 'Female Child Labor', y = 'Quality Education Score') +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```
data_research_question_3 %>% ggplot(aes(x=male_child_labor, y=goal_4_score)) +
  geom_point() + geom_smooth(se=FALSE, method='lm') +
  labs(x = 'Male Child Labor', y = 'Quality Education Score') +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Just from the graphs, it's very hard to see potential differences in the slopes which is what essentially determines whether there's a correlation so we'll find the equations for each of these.

```
model_male <- lm(goal_4_score ~ male_child_labor, data = data_research_question_3)
summary(model_male)$coefficients
```

```
##                  Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)      84.94784  3.4039593 24.955597 4.634640e-40
## male_child_labor -1.19058  0.1755438 -6.782235 1.690808e-09
```

```
model_female <- lm(goal_4_score ~ female_child_labor, data = data_research_question_3)
summary(model_female)$coefficients
```

```
##                    Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)        85.556170  3.2306613 26.482556 6.051266e-42
## female_child_labor -1.321711  0.1771812 -7.459655 8.129367e-11
```

Extracting coefficients we get the following equations for male and female respectively:

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

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

We also calculated the correlation coefficients:

```
cor(x=data_research_question_3$male_child_labor, y=data_research_question_3$goal_4_score)
```

```
## [1] -0.5994736
```

```
cor(x=data_research_question_3$female_child_labor, y=data_research_question_3$goal_4_score)
```

```
## [1] -0.635823
```

Looking at the equations, we can observe that there is greater change in 'Goal 4 Score' as a result of a change in 'Child labour' for the female group than for the male group, as the slope is greater for female than for male ($1.33 > 1.18$), making the fitted linear regression line a little bit steeper.

From the correlation coefficients, we can say that there's a stronger correlation for the female group between 'Goal 4 Score' and 'Child labour', as $0.64 > 0.6$. This concludes that although the difference is very small, the correlation between 'Goal 4 Score' and 'Child labour' is stronger for female than for male.

d) Discussion and Conclusion.

From our first test, we concluded that there exists a correlation between 'Goal 4 Score' and 'Child labour', as we rejected the null hypothesis.

We also concluded that using the data provided, it's hard to find differences in the trends for developing vs. developed countries in these two variables, as there's not enough data on the 'Child labour' variable for developed countries.

The male vs. female tests results could mean that girls that engage in child labour are more likely to not have a good quality education, for example, not be able to attend school. However, we would need an expert's opinion to evaluate this conclusion. It's important to mention that our tests only allow us to find association, not causation, meaning that engaging in child labor does not necessarily cause a worse education.

It's also important to mention the limitations of this study, which include 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.

V. Summary, Discussion, and 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.

In summary, our research findings shed light on the intricate relationship between education, development, and different indicators. The insights above emphasize the importance of equitable education access and inform policy decisions for enhancing global well-being and achieving other SDGs, not just SDG4.

VI. References.

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.