Project Report

STAT 184 Section 3

Gianna DeLorenzo & Sara Almansoori

1 Global Data - Learning and Skills

Education is the foundation of many countries worldwide. It is a fundamental right to have access to these tools. Over the past decade, global and regional efforts have been focusing on improving individuals' access to education, more specifically secondary education, addressing socio-economic disparities and integrating online learning tools; however, issues still arise within these efforts in providing equal education and access for all individuals. It is a social issue that affects everyone, whether indirectly or directly, and presents unequal treatment towards individuals' with less access to the same tools based on their regions and/or class levels.

Research Questions

In this report, we will be investigating the following research questions:

1. What are the global and regional trends in secondary schools enrollment rates over the past 10 years?

This research question explores patterns globally and regionally in terms of access to education. It focuses on the enrollment rates of adolescents of secondary school age.

2. How do socio-economic factors correlate with individuals' access to quality education and development?

This research question explores the patterns associated with socio-economic factors, and how they are connected with access to quality education and development. It goes into depth about which socio-economic factors impact this access the most and

3. How do gender disparities affect individuals' access to secondary education, and in what ways have these disparities changed over the course of 10 years ranging across different regions?

This research question explores the effects of gender disparities in secondary education. It is an important question to implement in our research because it can bring potential patterns in these disparities and can give a comparison between enrollment rates and completion rates in schooling.

We will first present the provenance of the data. Then, how it aligns with the FAIR and/or CARE principles. Then we will present the methodology of the data and the analysis attributes. After, we will explore the data visualizations and the results. Finally, we will close up the report with a discussion of our findings.

2 Dataset

2.1 Provenance

The dataset used for this analysis, the ‘Learning and Skills’ dataset from UNICEF, explores the global educational indicators ranging from 2016 to 2023. The indicators integrated include the percentage of children of lower secondary school age attending lower secondary school or higher, the percentage of children of upper secondary school age attending upper secondary school or higher

We obtained the data from UNICEF which is the United Nations Children’s Fund. It is a UN agency that is responsible for providing assistance to children worldwide pertaining humanitarian and developmental factors.

Source: UNICEF Data.

2.1.1 FAIR and CARE Principles

The dataset and analysis will adhere to the FAIR (Findable, Accessible, Interoperable, Reusable) principles to ensure transparency and reproducibility. Data sources and workflows will be documented and made available in open formats to facilitate future research. Additionally, the CARE (Collective benefit, Authority to control, Responsibility, Ethics) principles will guide the ethical use of data, particularly in ensuring respect for community rights and promoting equity in the dissemination of insights.

- **Findability**: The ‘Learning and Skills’ dataset can be accessed easily on the UNICEF Data explorer website under ‘Datasets’.

URL: UNICEF Data

- **Accessibility**: The ‘Learning and Skills’ dataset is publicly available as a CSV file and an Excel file.
- **Interoperability**: The ‘Learning and Skills’ dataset uses general formatting techniques and naming which makes it comparable using data analysis tools.
- **Reusability**: The ‘Learning and Skills’ dataset provides enough references and documented resources for similar research.
- **Collective benefit**: Insights from the ‘Learning and Skills’ dataset are shared for improvement purposes.
- **Authority to Control**: The ‘Learning and Skills’ dataset remains respectable for global and regional educational goals.
- **Responsibility**: The analysis of the ‘Learning and Skills’ dataset is conducted ethically and does not use the dataset and sensitive data for unethical reasons.
- **Ethics**: This research on the ‘Learning and Skills’ dataset focuses on equality to benefit communities who are at a disadvantage.

3 Methodology

3.1 Data Wrangling

As we mentioned prior, we filtered the ‘Learning and Skills’ dataset to focus on indicators including the percentage of children of lower secondary school age attending lower secondary school or higher, the percentage of children of upper secondary school age attending upper secondary school or higher.

This also provided us with a smaller sized dataset to work with. The original ‘Learning and Skills’ dataset was large and difficult to create visualizations with. It was also easier to analyze and discover readable trends from the sample dataset shown on the UNICEF website. We could come up with filters to integrate into the data visualizations and figure out which columns were the most essential to add in order to create a visualization that is best-suited to explain our research questions.

We found that it was necessary to filter **missing** and **null** values in the dataset. It was also found necessary to convert selected columns to numeric types which includes the Observation Values.

Overall, we put our focus on the following:

- **Global and Regional Enrollment Trends**
- **Completion Rates**
- **The Effects of Socio-Economic Factors and Gender Disparities**

3.2 Analysis Attributes

Temporal Trends: Changes in education metrics over time.

Regional Disparities: Variations across geographic regions.

Gender Disparities: Differences in metrics by gender.

Socio-Economic Correlations: Relationships between education metrics and socio-economic factors.

Indicator Trends: Examines the trends for indicators including attendance rates and completion rates regarding enrollment statistics.

4 Data Exploration

4.1 Exploratory Data Analysis

As we explore the structure of the UNICEF ‘Learning and Skills’ Dataset, the focus is on the patterns within the dataset, relationships between variables, and significant connections to our research questions. Our goals include examining the trends globally and regionally, identifying how these trends vary according to gender, and handling the null values to ensure the consistency of the data.

4.1.1 Summary Statistics

Here is a summary table that provides an overview of the observation values in the dataset according to the geographic region. The table includes the mean, minimum, median, maximum, and count for the observation values while removing null values.

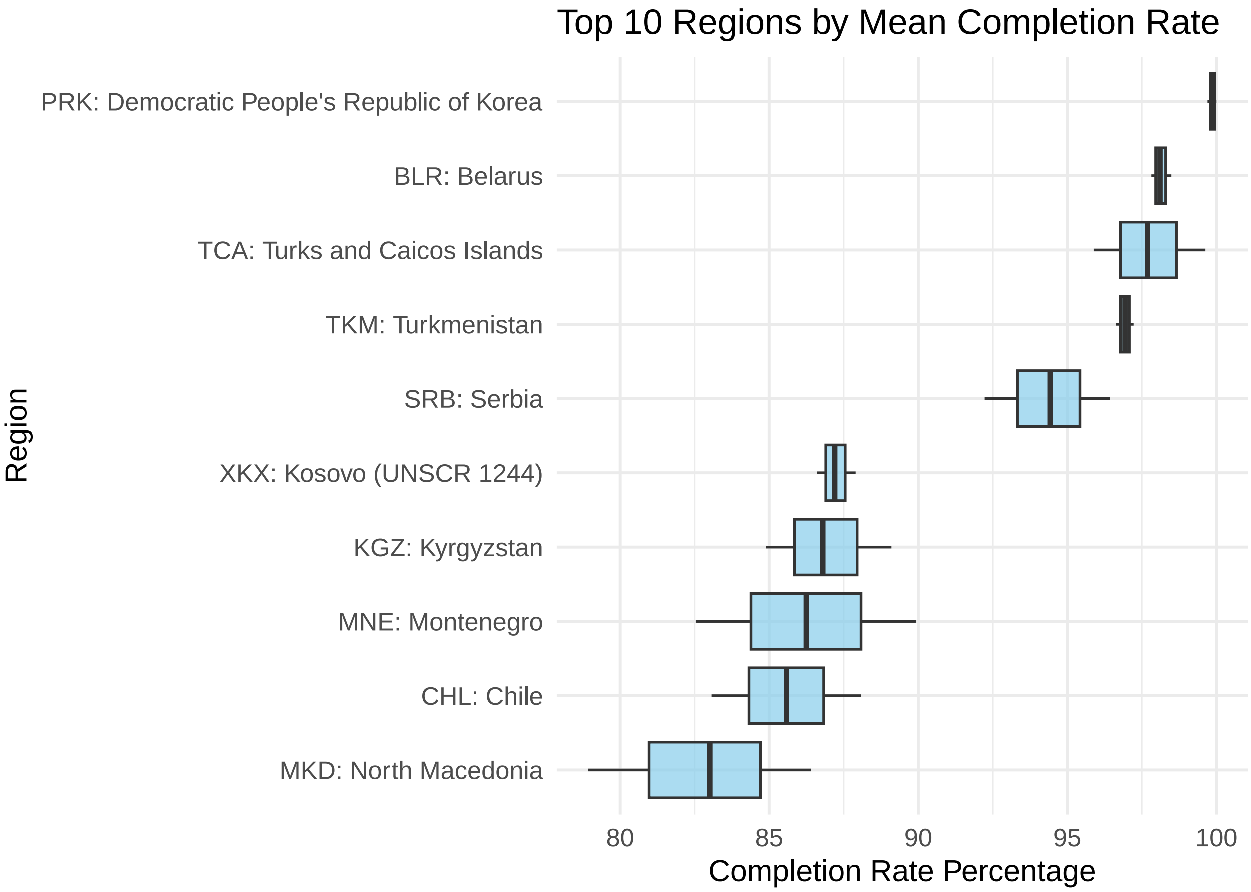
Summary Statistics

4.1.2 Enrollment Rates

4.1.2.1 Completion Rates

Furthermore, the completion rates of adolescents were one of the indicators that stood out to us. Here is a boxplot that displays the top 10 regions by the average completion rates ranging from the best to the worst regions.

Figure 1: Completion Rates



From this box-plot, it can be seen that the Democratic People's Republic of Korea is the top region with the highest average of completion rates in adolescents. The box-plot for this region is very close to 100% average completion rates. The small size of this box-plot also represents the minor variability in this average statistic value for this region, meaning that almost all individuals'

Table 1: Global Dataflow UNICEF

Geographic Area	Mean	Minimum	Median	Maximum	Count
AGO: Angola	26.84392	13.5266700	29.25522	42.05630	18
ALB: Albania	60.46288	2.0999999	77.79594	97.92905	18
ARG: Argentina	50.62327	1.0698500	64.36986	92.28950	18
ARM: Armenia	59.64951	0.7000000	80.38215	99.15201	18
BDI: Burundi	24.17530	3.4000001	23.19852	54.29890	18
BEN: Benin	28.24676	5.4404368	26.59325	66.10000	18
BGD: Bangladesh	40.76111	8.0000000	39.80000	70.50000	18
BLR: Belarus	63.02102	0.0000000	89.79915	98.81477	18
BLZ: Belize	37.80000	8.5000000	41.30000	65.50000	12
BOL: Bolivia (Plurinational State of)	51.44705	4.1562142	68.66000	84.25391	18
BRA: Brazil	52.22258	0.9666951	67.40969	88.60000	18
BWA: Botswana	47.22500	4.0000000	45.85000	91.60000	12
CAF: Central African Republic	18.34502	4.3049488	12.94070	63.46807	18
CHL: Chile	53.06566	0.4000000	66.60000	96.90509	18
CIV: Côte d'Ivoire	31.74076	13.5349200	30.41698	66.14774	18
CMR: Cameroon	35.20117	16.0521200	30.95905	53.22965	18
COD: Democratic Republic of the Congo	32.94103	15.7639900	32.20832	57.55380	18
CRI: Costa Rica	44.77120	2.3229780	54.28790	79.78559	18
CUB: Cuba	57.41271	1.5930210	66.46416	95.72341	18
DOM: Dominican Republic	50.55629	2.3194821	63.24412	85.76544	18
DZA: Algeria	46.45971	4.0125489	48.52436	88.48203	18
ECU: Ecuador	57.33333	5.4000001	75.15000	91.60000	18
ETH: Ethiopia	25.46559	6.6898718	23.76762	59.44491	18
GEO: Georgia	61.94678	0.7833119	82.90356	97.78119	18
GHA: Ghana	25.04929	6.4363341	20.61843	50.22569	18
GIN: Guinea	32.63889	11.2395800	30.44482	73.03444	18
GMB: Gambia	37.33333	29.0000000	35.20000	49.70000	6
GNB: Guinea-Bissau	16.64268	6.4768138	15.04884	35.93432	18
GUY: Guyana	55.21310	4.7565060	67.91807	91.86062	18
HND: Honduras	43.35637	25.8577310	43.06304	64.63129	18
HTI: Haiti	20.40539	5.2011690	18.14706	38.15783	18
IDN: Indonesia	55.56496	4.2785702	63.50632	90.76429	18
IND: India	49.81070	6.9056602	54.28550	82.44800	18
IRQ: Iraq	40.46111	14.7000000	43.80000	57.50000	18
JOR: Jordan	54.60252	5.0999999	65.10430	89.58794	18
KGZ: Kyrgyzstan	63.71667	1.2000000	86.75000	99.00000	18
KIR: Kiribati	46.52788	5.2956481	45.76121	88.42178	18
LAO: Lao People's Democratic Republic	40.24627	17.9000000	38.10000	61.50000	18
LKA: Sri Lanka	52.96170	0.7452087	55.46121	96.28514	18
LSO: Lesotho	33.75556	8.8000002	32.60000	66.90000	18
MDG: Madagascar	30.48333	11.4000000	26.45000	66.60000	18
MDV: Maldives	36.89503	3.8307080	27.88075	85.50000	12
MEX: Mexico	55.86111	5.4000001	63.05000	90.60000	18
MKD: North Macedonia	63.07876	1.0304180	88.06458	98.18111	18
MMR: Myanmar	43.11667	15.5000000	44.90000	71.00000	12
MNE: Montenegro	61.80625	0.9855446	86.80725	96.70629	18
MNG: Mongolia	48.62500	3.3000000	48.90000	94.30000	12
MWI: Malawi	19.85755	7.8872399	19.85000	36.39693	18
NGA: Nigeria	41.85000	24.6000000	39.25000	66.20000	18

complete upper secondary education. For all regions, the range stays between around 80% - 100% for completion rate percentages. Each individual box-plot for the region represents the quartile range going from 25th percentiles to 75th percentiles. The line in each boxplot represents the median and the whiskers (the lines connected to each side of the box) represent the minimum and maximum values. According to how this boxplot outputted, there are no outliers from this data.

On the other hand, regions such as North Macedonia and Chile have the lowest completion rates. Their large sizes indicate a major variability in the averages for these regions meaning that the completion rates are not consistent with the long whiskers for the boxes. The median for these boxes ranging between 83% to 86% are lower than all other regions with narrower boxes as well.

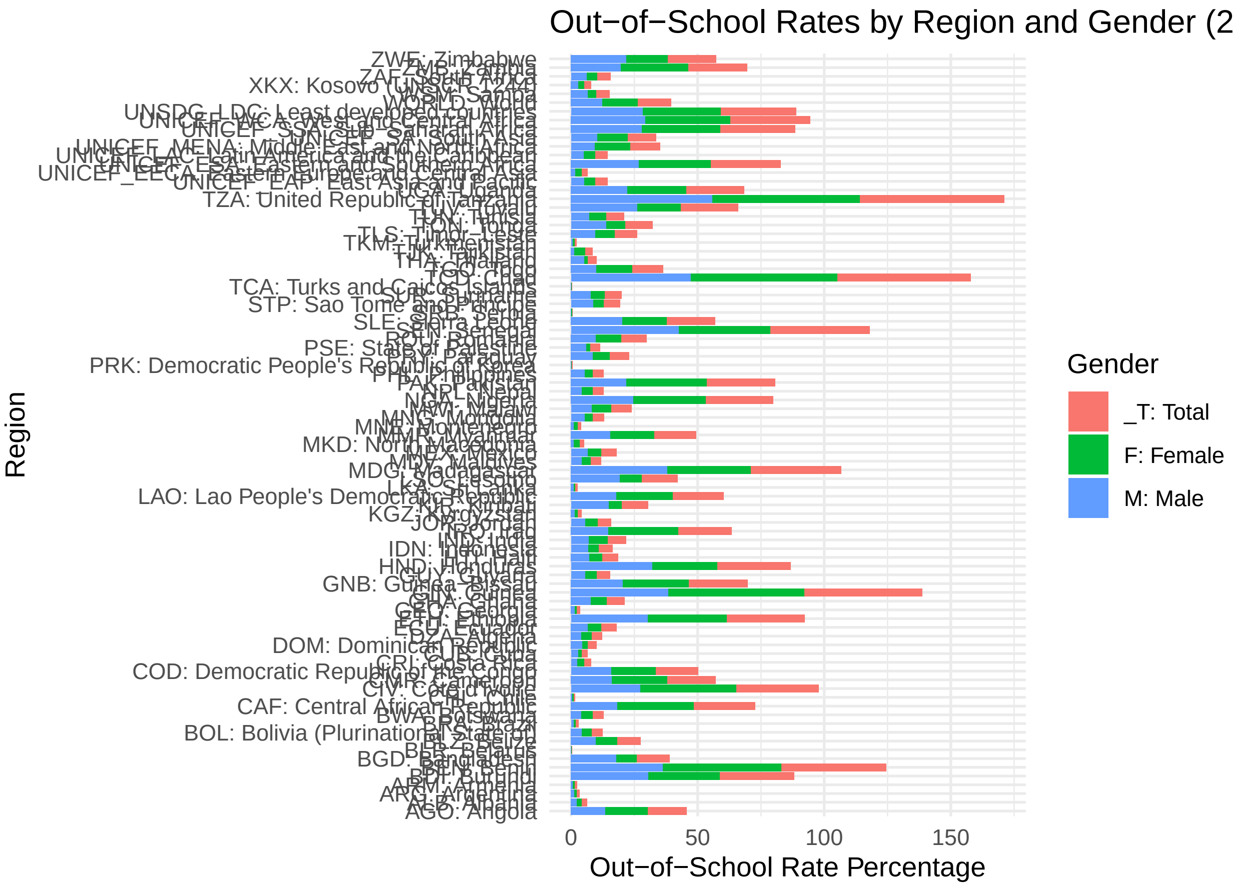
4.1.2.2 Net Attendance Rates

Figure 2: Net Attendance Rates

4.1.2.3 Out of School Rates

Additionally, out of school rates are another enrollment statistic that was significant to the ‘Learning and Skills’ dataset. It is layed out in the CSV file under the indicator column as “ED_ROFST_L2: Out-of-school rate for adolescents of lower secondary school age.” With our question in mind that explores the global and regional trends in secondary schools enrollment rates,

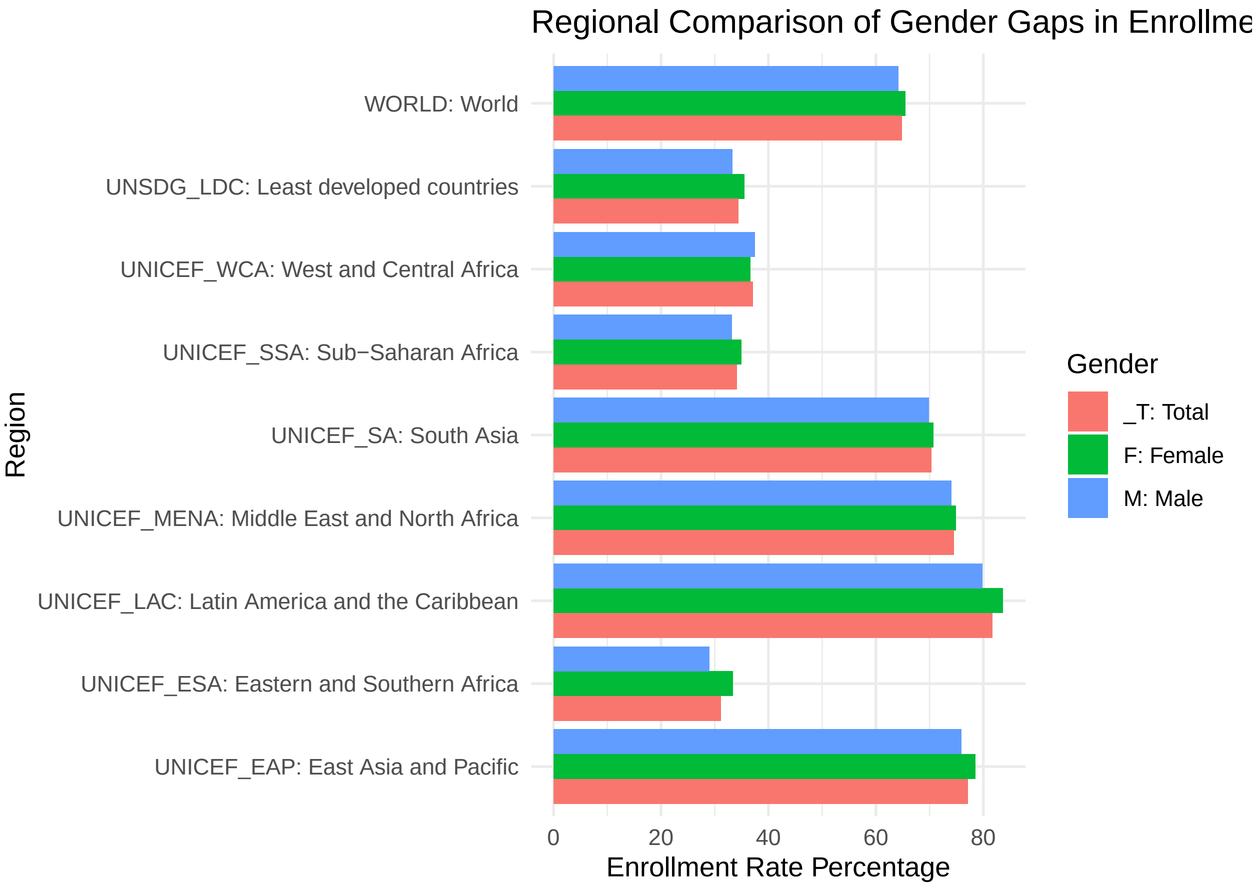
Figure 3: Out of School Rates



4.1.2.4 Gender Disparities

The following grouped bar plot highlights important trends in the data according to the net attendance rate which was another significant value under the indicator column in the ‘Learning and Skills’ dataset. This visualization provides a comparison between the gender disparities in lower secondary school enrollment rates ranging across multiple regions globally in 2022. The following plot also highlights areas where gender gaps would take place and where there has been progress towards achieving parity with both genders.

Figure 4: Gender Disparities



From the looks of this visualization output, the enrollment rates for females and males are close in range. This is suggesting to us that gender parity has been successfully achieved worldwide, although we cannot come to a clear conclusion when only looking at a visualization including a sample of 10 regions. The most significant differences between the enrollment rates according to the gender appears to be in regions including **East Asia and Pacific**, **Latin America and the Caribbean**, and **Eastern and Southern Africa** where the gender gaps are close displaying near-equal enrollment rates for females and males. This output indicate that the regions with closer gender gaps may be targeting interventions or implementing broader societal shifts towards gender equality.

5 Results

Here is summary of the statistics including the mean, standard deviation, minimum, and maximum values.

```
# A tibble: 85 x 6
# Groups:   REF_AREA.Geographic.area [85]
   REF_AREA.Geographic.area     TIME_PERIOD.Time.per~1  Mean      SD    Min    Max
   <chr>                                     <int> <dbl>   <dbl> <dbl> <dbl>
 1 AGO: Angola                                2016  31.5 0.702    30.8  32.2
 2 ALB: Albania                               2018  95.1 0.174    95.0  95.3
 3 ARG: Argentina                             2020  89.4 2.83     86.6  92.3
 4 ARM: Armenia                               2016  94.1 0.894    93.2  95.0
 5 BDI: Burundi                               2017  24.2 3.15     21.1  27.4
 6 BEN: Benin                                 2018  30.3 2.20     28.1  32.5
 7 BGD: Bangladesh                            2019  57.9 6.70     51.2  64.6
 8 BLR: Belarus                               2020  93.4 2.18     91.2  95.6
 9 BOL: Bolivia (Plurinational~               2016  71.7 0.00569  71.7  71.7
10 BRA: Brazil                                2019  85.1 2.03     83.1  87.1
# i 75 more rows
# i abbreviated name: 1: TIME_PERIOD.Time.period
```

Trends include …

6 Assumptions

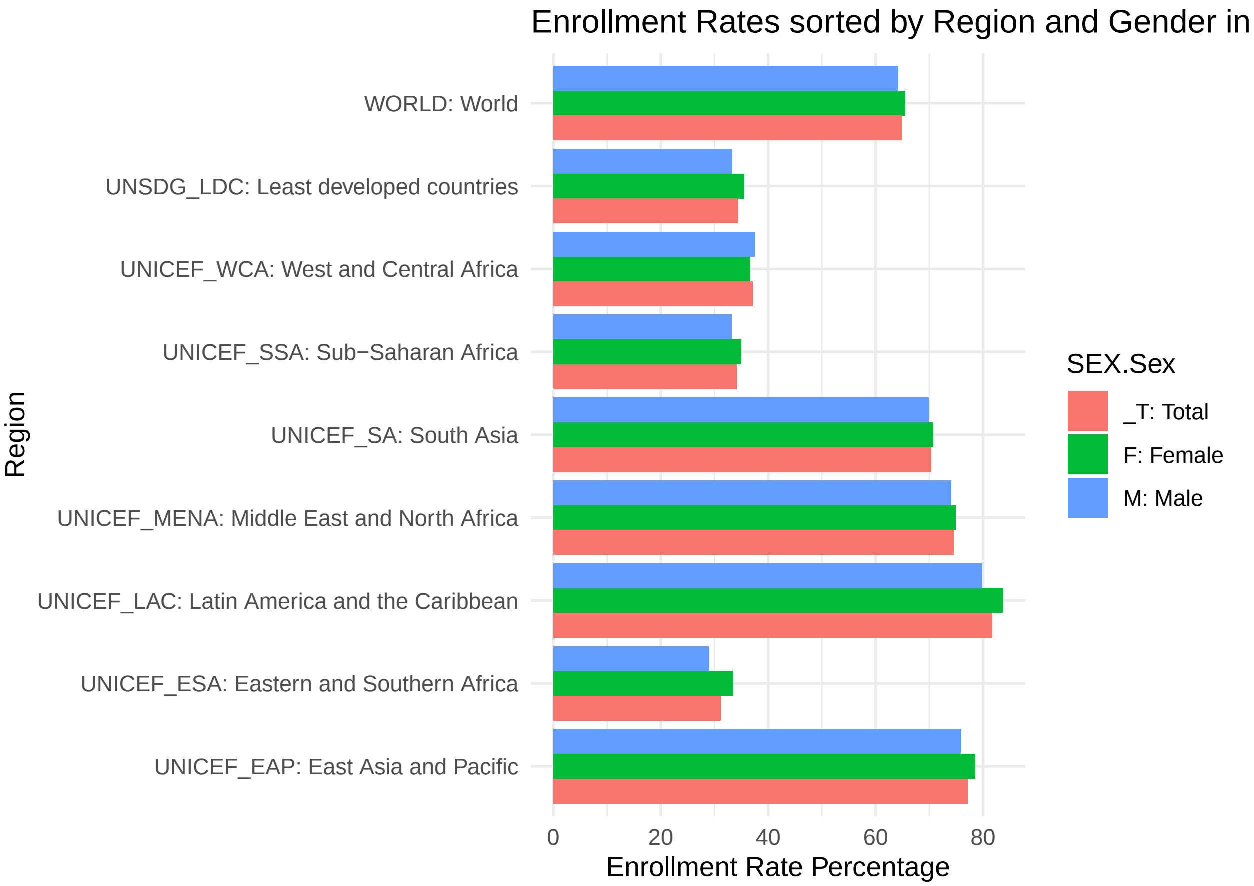
The ‘Learning and Skills’ dataset does not provide sufficient data to go through with hypothesis testing including a paired t-test for the first research question, Pearson’s Correlation for the second research question, and a Two-Sample t-test for the third research question, so we are implementing standard visualizations to effectively communicate the trends and disparities in the enrollment rates.

Research Question 1: What are the global and regional trends in secondary schools’ enrollment rates over the past 10 years?

Figure 5: Global and Regional Trends

Research Question 2: How do socio-economic factors correlate with individuals' access to quality education and development?

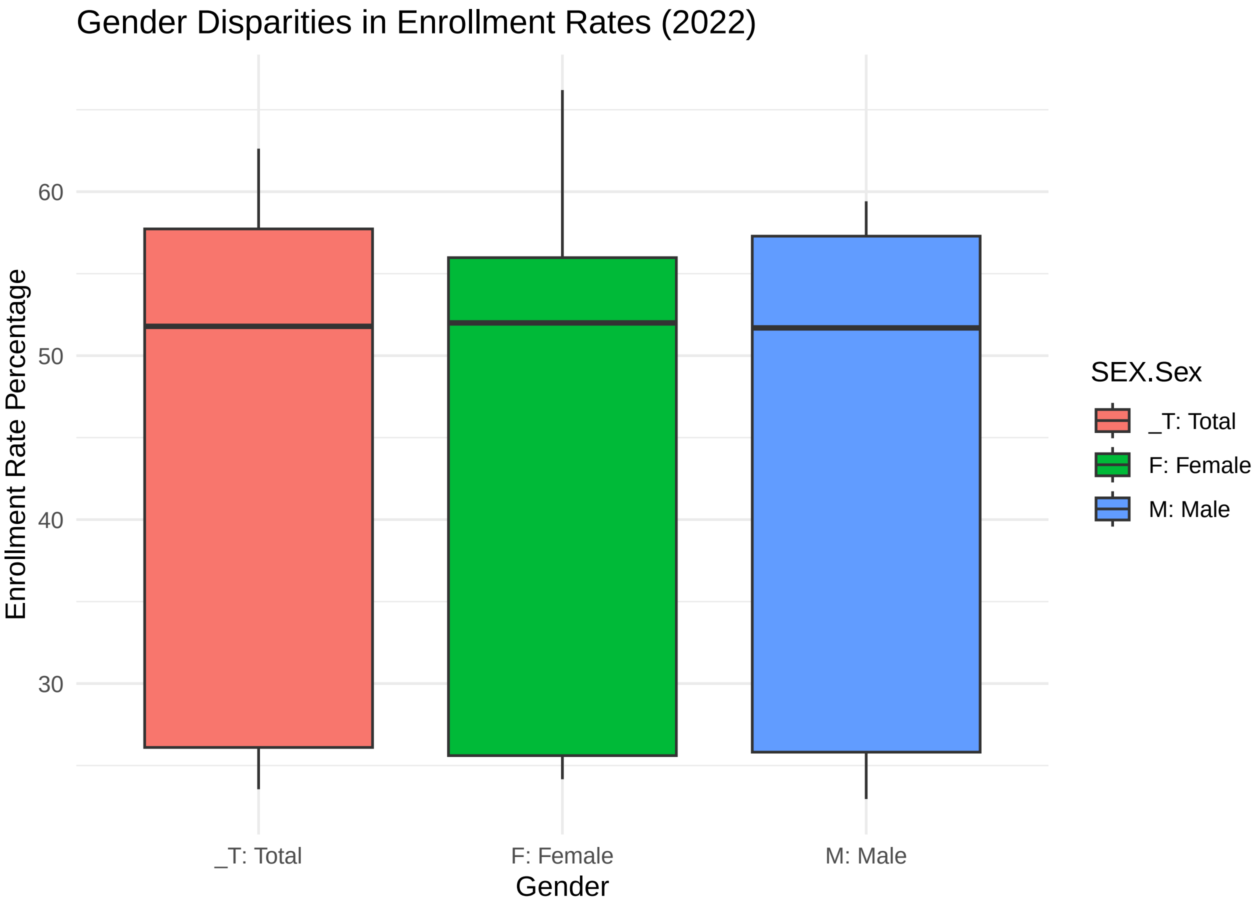
Figure 6: Region and Gender



This represents enrollment rates sorted by the Region and Gender in the year 2022. The enrollment rates appear balanced globally.

Research Question 3: How do gender disparities affect individuals' access to secondary education, and in what ways have these disparities changed over the course of 10 years?

Figure 7: Gender Disparities Enrollment



This represents the gender disparities in the 2022 enrollment rates. The average enrollment rates appear to be slightly higher for males than for females which can indicate some potential gender disparities, but not significantly.

7 Discussion

From the overall results of this report, we can fully analyze the global and regional trends in the **'Learning and Skills'** dataset in the secondary school enrollment rates. We can also fully analyze the gender disparities in the dataset in terms of access to education. While the global enrollment rates are showing improvement in the regions, there are significant gaps in regions such as **West and Central Africa**. Regions including **East Asia and the Pacific** and **Latin America** have shown high enrollment rates with boxplots having low variability and gender gaps being low, while regions such as **Sub-Saharian Afica** show the opposite. In certain regions, it was clear that male enrollment rates were found to be slightly higher in a global sense while having more variability in the female rates.

Based on our findings, it is recommended to …

This research aims to provide evidence-based insights into the factors influencing global education and skill acquisition. By highlighting trends, disparities, and socio-economic relationships, the study will contribute to the discourse on equitable access to education.

8 References

United Nations Children's Fund. (n.d.). Data Explorer: Global Dataflow. UNICEF. Retrieved December 16, 2024, from https://data.unicef.org/resources/data_explorer/unicef_f/?ag=UNICEF&df=GLOBAL_D

Best, D. J., & Roberts, D. E. (1975). Algorithm AS 89: The upper tail probabilities of Spearman's. Applied Statistics, 24(3), 377–379. https://doi.org/10.2307/2347111

Hollander, M., & Wolfe, D. A. (1973). Nonparametric statistical methods (pp. 185–194). John Wiley & Sons.

9 Code Appendix

```
#Load necessary libraries
library(dplyr)
library(kableExtra)


#Import UNICEF csv file
unicef_data <- read.csv(url("https://raw.githubusercontent.com/Stat184-Fall2024/Sec3_FP_Giannal

#Summary Statistics
unicef_summary_table <- unicef_data %>%
  group_by(REF_AREA.Geographic.area) %>%
  summarise(
    mean_value = mean(OBS_VALUE.Observation.Value, na.rm = TRUE),
    minimum_value = min(OBS_VALUE.Observation.Value, na.rm = TRUE),
    median_value = median(OBS_VALUE.Observation.Value, na.rm = TRUE),
    maximum_value = max(OBS_VALUE.Observation.Value, na.rm = TRUE),
    count = n()
  )

#Format the Summary Statistics in a Table using Kable Styling
kable(
  unicef_summary_table,
  format = "latex",
  caption = "Global Dataflow UNICEF",
  col.names = c('Geographic Area', 'Mean', 'Minimum', 'Median', 'Maximum', 'Count')
)

# Load necessary libraries
```

```r
library(dplyr)
library(ggplot2)

# Filtered data for Completion Rate Indicator
filtered_completion_data <- unicef_data %>%
  filter(
    INDICATOR.Indicator == "ED_CR_L3: Completion rate for youth of upper secondary education sc
    !is.na(OBS_VALUE.Observation.Value)) %>% # Remove missing values

  mutate(
    OBSERVATION_VALUE = as.numeric(OBS_VALUE.Observation.Value)
  )

# Filter for the Top 10 Regions by the average Completion Rate of Adolescents
top_ten_regions <- filtered_completion_data %>%
  group_by(REF_AREA.Geographic.area) %>%
  summarise(mean_value = mean(OBSERVATION_VALUE, na.rm = TRUE)) %>%
  top_n(10, mean_value) %>%
  pull(REF_AREA.Geographic.area)

# Filter for top regions
top_ten_filtered_data <- filtered_completion_data %>%
  filter(REF_AREA.Geographic.area %in% top_ten_regions)

# Create the boxplot
ggplot(top_ten_filtered_data,
       aes(x = reorder(REF_AREA.Geographic.area, OBSERVATION_VALUE, mean),
           y = OBSERVATION_VALUE)) +
  geom_boxplot(fill = "skyblue", alpha = 0.7) +
  coord_flip() +
  labs(
    title = "Top 10 Regions by Mean Completion Rate",
    x = "Region",
    y = "Completion Rate Percentage"
  ) +
  theme_minimal(base_size = 12)
# Load necessary libraries
library(dplyr)
library(ggplot2)
library(kableExtra)

# Filter for lower secondary school with adolescents
filtered_netattend_data <- unicef_data %>%
  filter(INDICATOR.Indicator == "ED_ANAR_L2: Adjusted net attendance rate for adolescents of lo
         !is.na(OBS_VALUE.Observation.Value)) %>% # Remove missing values

  mutate(
    Year = as.numeric(TIME_PERIOD.Time.period),
    Enrollment_Rate = as.numeric(OBS_VALUE.Observation.Value),
    Region = REF_AREA.Geographic.area)


# Create the Line Plot
ggplot(filtered_netattend_data, aes(x = Year,
                                    y = Enrollment_Rate,
                                    color = Region)) +
  geom_line(size = 1) +
  labs(
    title = "Global and Regional Trends in Secondary School Enrollment with Adolescents",
    x = "Year",
    y = "Enrollment Rate Percentage",
    color = "Region"
  ) +
  theme_minimal() +
  theme(
    legend.position = "bottom",
    legend.title = element_text(size = 10)
  )
# Load necessary libraries
library(dplyr)
library(ggplot2)

# Filter based on Out of School Rates
filtered_OOS_data <- unicef_data %>%
  filter(INDICATOR.Indicator == "ED_ROFST_L2: Out-of-school rate for adolescents of lower secon
         !is.na(OBS_VALUE.Observation.Value)) %>%
  mutate(
    OBS_VALUE = as.numeric(OBS_VALUE.Observation.Value)
  )

# Create a Stacked Bar Chart
ggplot(filtered_OOS_data, aes(x = REF_AREA.Geographic.area,
                              y = OBS_VALUE,
                              fill = SEX.Sex)) +
  geom_bar(stat = "identity", position = "stack") +
  coord_flip() +
  labs(
    title = "Out-of-School Rates by Region and Gender (2022)",
    x = "Region",
    y = "Out-of-School Rate Percentage",
    fill = "Gender"
  ) +
  theme_minimal()
# Load necessary libraries
```

```r
library(dplyr)
library(ggplot2)

filtered_disparities_data <- unicef_data %>%
  filter(
    INDICATOR.Indicator == "ED_ANAR_L2: Adjusted net attendance rate for adolescents of lower s
    TIME_PERIOD.Time.period == 2022,
    !is.na(OBS_VALUE.Observation.Value)) %>%

  mutate(
    OBS_VALUE = as.numeric(OBS_VALUE.Observation.Value))

# Plot gender gaps as grouped bar chart
ggplot(filtered_disparities_data, aes(x = REF_AREA.Geographic.area,
                       y = OBS_VALUE,
                       fill = SEX.Sex)) +

  geom_bar(stat = "identity", position = "dodge") +
  coord_flip() +
  labs(
    title = "Regional Comparison of Gender Gaps in Enrollment Rates in 2022",
    x = "Region",
    y = "Enrollment Rate Percentage",
    fill = "Gender"
  ) +
  theme_minimal()

# Load necessary libraries
library(dplyr)
library(ggplot2)

unicef_data %>%
  filter(INDICATOR.Indicator == "ED_ANAR_L2: Adjusted net attendance rate for adolescents of lo
  group_by(REF_AREA.Geographic.area,
           TIME_PERIOD.Time.period) %>%
  summarise(
    Mean = mean(OBS_VALUE.Observation.Value, na.rm = TRUE),
    SD = sd(OBS_VALUE.Observation.Value, na.rm = TRUE),
    Min = min(OBS_VALUE.Observation.Value, na.rm = TRUE),
    Max = max(OBS_VALUE.Observation.Value, na.rm = TRUE)
  )

# Load necessary libraries
library(dplyr)
library(ggplot2)

ggplot(unicef_data %>% filter(INDICATOR.Indicator == "ED_ANAR_L3: Adjusted net attendance rate
       aes(x = TIME_PERIOD.Time.period,
           y = OBS_VALUE.Observation.Value,
           color = REF_AREA.Geographic.area)) +
  geom_line() +
  labs(title = "Trends in Enrollment Rates Over Time",
       x = "Year",
       y = "Enrollment Rate Percentage") +
  theme_minimal()
# Load necessary libraries
library(dplyr)
library(ggplot2)

ggplot(unicef_data %>% filter(INDICATOR.Indicator == "ED_ANAR_L2: Adjusted net attendance rate
       aes(x = REF_AREA.Geographic.area,
           y = OBS_VALUE.Observation.Value,
           fill = SEX.Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  coord_flip() +
  labs(title = "Enrollment Rates sorted by Region and Gender in 2022",
       x = "Region",
       y = "Enrollment Rate Percentage") +
  theme_minimal()
# Load necessary libraries
library(dplyr)
library(ggplot2)

# Create Box Plot
ggplot(unicef_data %>% filter(INDICATOR.Indicator == "ED_ANAR_L3: Adjusted net attendance rate
       aes(x = SEX.Sex,
           y = OBS_VALUE.Observation.Value,
           fill = SEX.Sex)) +
  geom_boxplot() +
  labs(title = "Gender Disparities in Enrollment Rates (2022)",
       x = "Gender",
       y = "Enrollment Rate Percentage") +
  theme_minimal()
```