# Project Report

Gianna DeLorenzo & Sara Almansoori

## 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.

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 does individuals' access to online learning tools differentiate according to the region and class levels, and what impact does this have on learning?

This research question explores the analysis of individuals' access to online learning tools as it pertains to equity in learning and outcomes.

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.

# Dataset

## 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 Explorer.

### 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 is https://data.unicef.org/resources/data_explorer/unicef_f/?ag=UNICEF&df=GLOBAL_DATAFLOW

**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.

## Methodology

### 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 including **

Overall, we put our focus on the following:

- **Global and Regional Enrollment Trends**
- **Completion Rates**
- **The Effects of Socio-Economic Factors and Digital Tools**

### 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.

**Digital Access**: Availability and impact of digital learning tools.

## Data Exploration

### Exploratory Data Analysis

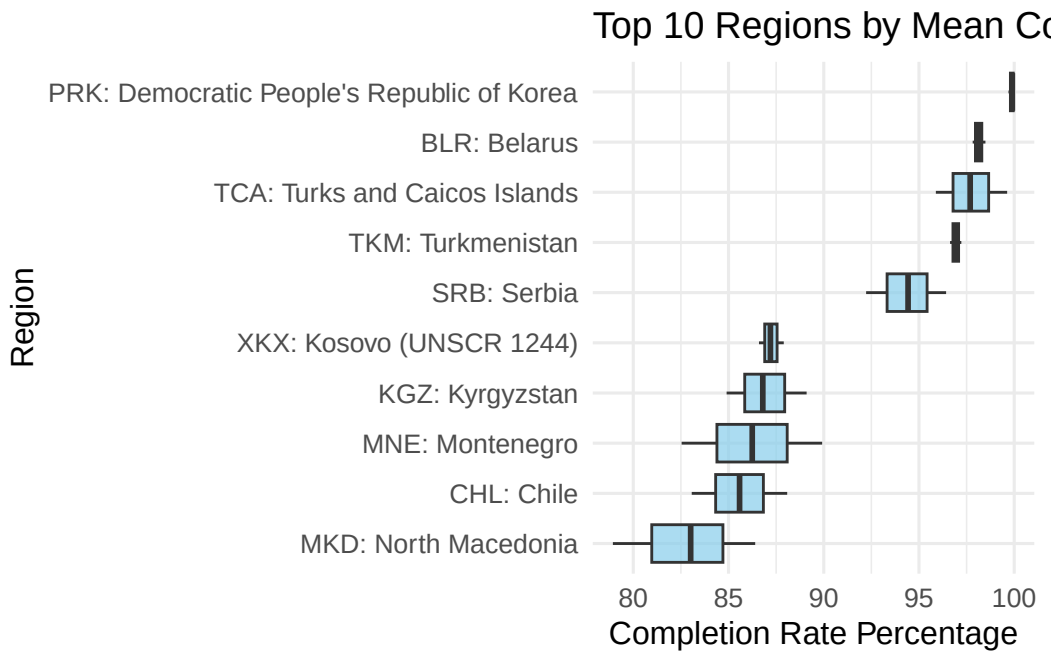
As we explore the structure of the UNICEF 'Learning and Skills' Dataset, the focus is on the patterns within the dataset, relationships between variables, and significant connections to our research questions. Our goals include examining the trends globally and regionally, identifying how these trends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

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.

Furthermore, the completion rates of adolescents were one of the indicators that stood out to us. Here is a generalized boxplot that displays the completion rate according to ...

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

# Import UNICEF csv file
unicef_data <- read.csv(url("https://raw.githubusercontent.com/Stat184-Fall2024/Sec3_FP_Gian

# Filtered data for Completion Rate Indicator
filtered_unicef_data <- unicef_data %>%
  filter(
    INDICATOR.Indicator == "ED_CR_L3: Completion rate for youth of upper secondary education
    !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_unicef_data %>%
  group_by(REF_AREA.Geographic.area) %>%
  summarise(mean_value = mean(OBSERVATION_VALUE, na.rm = TRUE)) %>%
```

```r
  top_n(10, mean_value) %>%
  pull(REF_AREA.Geographic.area)

# Filter for top regions
top_ten_filtered_data <- filtered_unicef_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)
```




From this boxplot, 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 boxplot for this region is very close to 100% average completion rates. The small size of this boxplot also represents

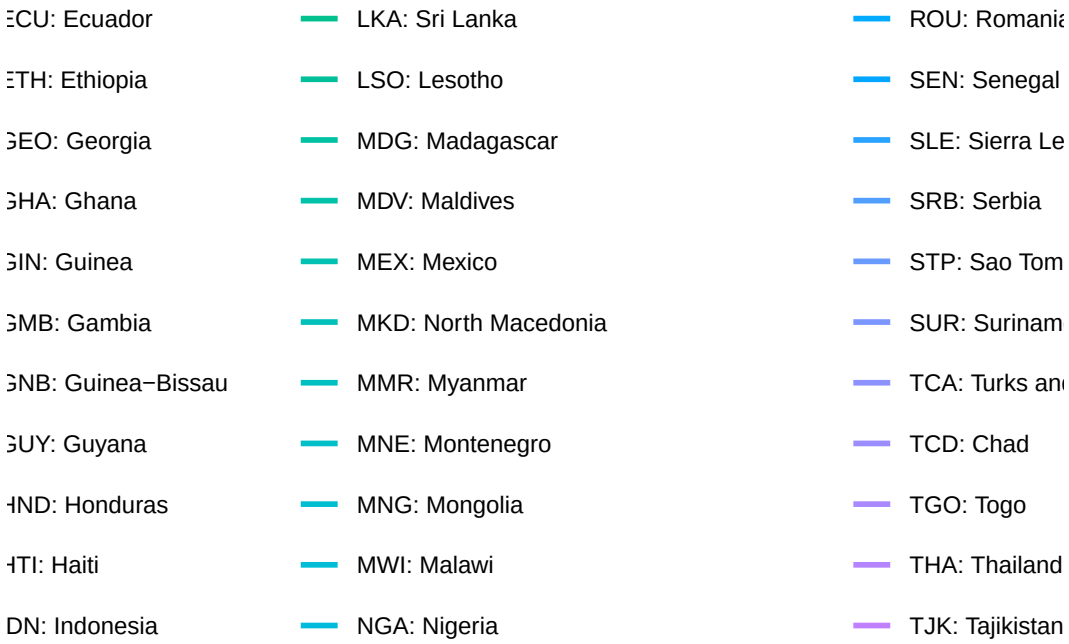
the minor variability in this average statistic value for this region, meaning that almost all individuals' complete upper secondary education.

On the other hand, regions such as North Macedonia and Chile have the lowest completion rates.

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

# Import UNICEF csv file
unicef_data <- read.csv(url("https://raw.githubusercontent.com/Stat184-Fall2024/Sec3_FP_Gian

# Filter for lower secondary school with adolescents
filtered_data <- unicef_data %>%
  filter(INDICATOR.Indicator == "ED_ANAR_L2: Adjusted net attendance rate for adolescents of
         !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)



ggplot(filtered_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)
  )
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
```

```
i Please use `linewidth` instead.
```




### Visualization 2

### Visualization 3

## Results

## Conclusion

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.

## 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=GLOBA

## Code Appendix

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

#Import UNICEF csv file
unicef_data <- read.csv(url("https://raw.githubusercontent.com/Stat184-Fall2024/Sec3_FP_Gian

#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
knitr::kable(
  unicef_summary_table,
  format = "latex",
  caption = "Global Dataflow UNICEF",
  col.names = c('Geographic Area', 'Mean', 'Minimum', 'Median', 'Maximum', 'Count')
)


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

# Import UNICEF csv file
unicef_data <- read.csv(url("https://raw.githubusercontent.com/Stat184-Fall2024/Sec3_FP_Gian

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

```r
  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_unicef_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_unicef_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)

# Import UNICEF csv file
unicef_data <- read.csv(url("https://raw.githubusercontent.com/Stat184-Fall2024/Sec3_FP_Gian

# Filter for lower secondary school with adolescents
filtered_data <- unicef_data %>%
  filter(INDICATOR.Indicator == "ED_ANAR_L2: Adjusted net attendance rate for adolescents of
         !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)



ggplot(filtered_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)
  )
```