Data Science Internship Assessment 2025

GitHub: github.com/Brian-Juma

Brian Juma

2025-04-29

Introduction

This report was prepared as part of the 2025 Data Science Task for my internship at the Center for Epidemiological Modelling and Analysis (CEMA), University of Nairobi. It presents an independent analysis of HIV burden, multidimensional poverty and child mortality in Africa using integrated datasets from the WHO Global Observatory (2000–2023), the World Bank Multidimensional Poverty Indicators and the UN Inter-agency Group for Child Mortality Estimation (UN IGME).

HIV continues to be a major public health challenge in Africa, particularly in sub-Saharan regions where prevalence remains high despite advances in prevention and treatment. Since HIV is a lifelong condition, understanding infection trends is essential for planning healthcare services, reducing stigma and improving equity in access to care. Beyond health, HIV also shapes broader socioeconomic outcomes by affecting labor markets, healthcare costs and poverty dynamics thus highlighting the importance of data driven targeted interventions.

This task required applying data science methods to address two key questions:

1. **HIV Trends and Poverty Linkages**
    - Identify countries contributing to 75% of the global HIV burden, and visualize their trends over time.
    - Perform a similar analysis at the WHO regional level.
    - Merge HIV prevalence data with World Bank multidimensional poverty indicators (income, education, infrastructure, water and sanitation).
    - Use a mixed-effects model to analyze the relationship between poverty and HIV, accounting for country,year level random effects.
2. **Child Mortality in East Africa**
    - Focus on the eight East African Community (EAC) countries: Burundi, Democratic Republic of the Congo, Kenya, Rwanda, South Sudan, Uganda, United Republic of Tanzania and Somalia.

- Visualize the latest estimates of under-five and neonatal mortality using spatial data (GADM shapefiles: www.gadm.org).

- Plot average trends over time, alongside country level data points, for both mortality indicators.

- Identify the countries with the highest under-five and neonatal mortality rates.

This document follows a research report structure: it motivates the problem, describes the data and methods used, presents statistical and visual results, and discusses policy relevant implications. **All code was written in R and is fully reproducible; source files are available on my GitHub: github.com/Brian-Juma.**

1.1 Data Preparation

1.1.1 Load Required Libraries

```
library(tidyverse)
library(lme4)
library(sf)
library(viridis)
library(patchwork)
library(scales)
library(ggrepel)
library(readxl)
library(kableExtra)
library(broom.mixed)
library(tinytex)
```

1.2 Setting Work Directory

1.2.1 Read and clean the data

```
hiv_data <- read_csv("HIV data 2000-2023.csv") %>%
  # Extract numeric value from the Value column (removing confidence intervals)
  mutate(
    Value_num = as.numeric(str_extract(Value, "^[0-9, ]+") %>%
                  str_remove_all("[ ,]")),
    Value_num = ifelse(is.na(Value_num), 0, Value_num)
  ) %>%
  filter(!is.na(Value_num), Value_num > 0) %>%
  # Clean up year and location names
  mutate(
    Year = as.numeric(Period),
```

```
    Country = Location
  )
```

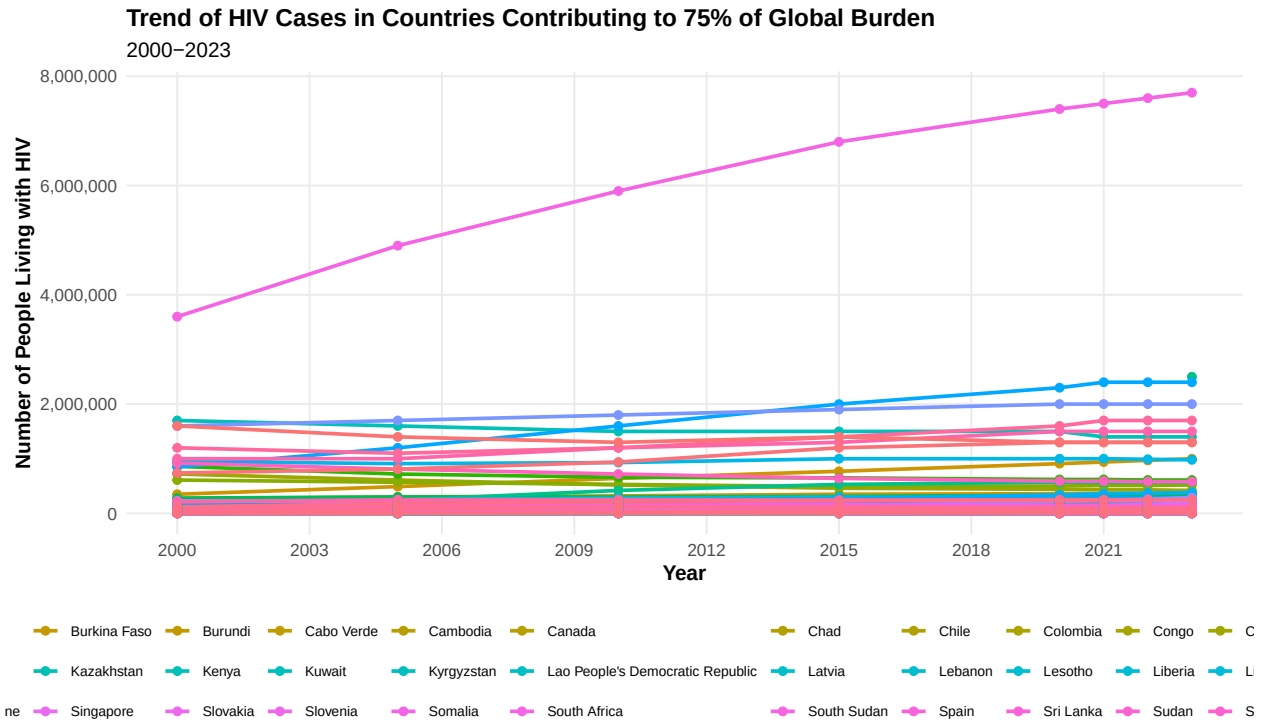
1.3 Global Trends

```
# Calculate global totals by year
global_totals <- hiv_data %>%
  group_by(Year) %>%
  summarise(Global_Total = sum(Value_num, na.rm = TRUE))

# Calculate country contributions
country_contributions <- hiv_data %>%
  group_by(Country, ParentLocationCode, ParentLocation, Year) %>%
  summarise(Country_Total = sum(Value_num, na.rm = TRUE)) %>%
  left_join(global_totals, by = "Year") %>%
  mutate(Contribution = Country_Total / Global_Total) %>%
  arrange(Year, desc(Contribution))

# Identify countries contributing to 75% of global burden
top_countries <- country_contributions %>%
  group_by(Country, ParentLocation) %>%
  summarise(Avg_Contribution = mean(Contribution, na.rm = TRUE)) %>%
  arrange(desc(Avg_Contribution)) %>%
  mutate(Cumulative_Contribution = cumsum(Avg_Contribution)) %>%
  filter(Cumulative_Contribution <= 0.75) %>%
  pull(Country)

# Filter data for top countries
top_countries_data <- hiv_data %>%
  filter(Country %in% top_countries) %>%
  group_by(Country, ParentLocation, Year) %>%
  summarise(Value_num = sum(Value_num, na.rm = TRUE))
```

1.4 Visualization

```
# Create visualization with improved legend layout
ggplot(top_countries_data, aes(x = Year, y = Value_num, color = Country)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  scale_y_continuous(labels = comma,
                     breaks = scales::pretty_breaks(n = 6),
                     limits = c(0, NA)) +
  scale_x_continuous(breaks = seq(2000, 2023, by = 3)) +
  labs(title = "Trend of HIV Cases in Countries Contributing to 75% of Global Burden",
```

```
    subtitle = "2000-2023",
    y = "Number of People Living with HIV",
    x = "Year",
    color = "Country") +
  theme_minimal(base_size = 12) +
  theme(
    legend.position = "bottom",
    plot.title = element_text(face = "bold", size = 14),
    axis.title = element_text(face = "bold"),
    panel.grid.minor = element_blank(),
    legend.text = element_text(size = 8),
    legend.box = "horizontal"
  ) +
  guides(color = guide_legend(nrow = 3, byrow = TRUE))
```

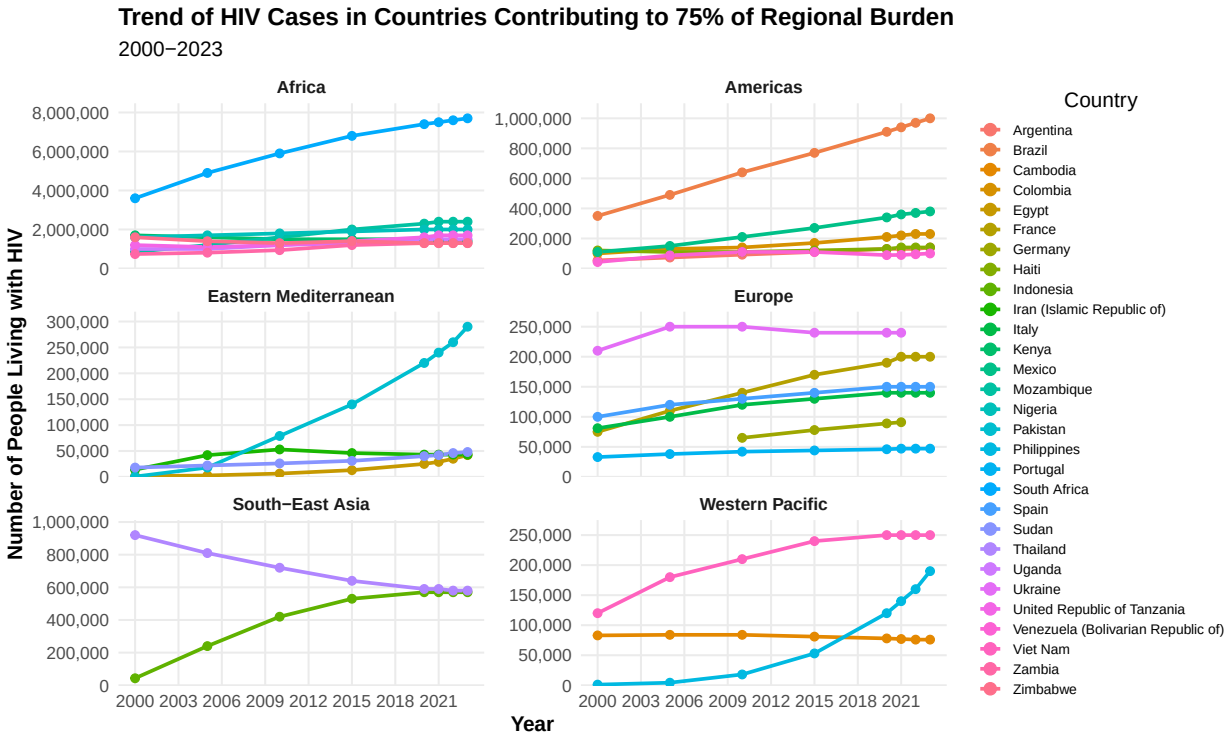


1.5 Regional Analysis

```
# Function to get top countries by region
get_regional_top_countries <- function(region_data) {
  region_data %>%
    group_by(Country) %>%
    summarise(Total_Value = sum(Value_num, na.rm = TRUE)) %>%
```

```
    arrange(desc(Total_Value)) %>%
    mutate(
      Region_Total = sum(Total_Value),
      Contribution = Total_Value / Region_Total,
      Cumulative_Contribution = cumsum(Contribution)
    ) %>%
    filter(Cumulative_Contribution <= 0.75) %>%
    pull(Country)
}

# Get top countries for each region
regional_top_countries <- hiv_data %>%
  group_by(ParentLocation) %>%
  group_modify(~ tibble(Country = get_regional_top_countries(.x)))

# Filter data for regional top countries
regional_top_data <- hiv_data %>%
  inner_join(regional_top_countries, by = c("ParentLocation", "Country")) %>%
  group_by(ParentLocation, Country, Year) %>%
  summarise(Value_num = sum(Value_num, na.rm = TRUE))

# Create visualization by region
ggplot(regional_top_data, aes(x = Year, y = Value_num, color = Country)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  scale_y_continuous(
    labels = comma,
    breaks = scales::pretty_breaks(n = 5),
    limits = c(0, NA),
    expand = expansion(mult = c(0, 0.1))) +
  scale_x_continuous(breaks = seq(2000, 2023, by = 3)) +
  facet_wrap(~ ParentLocation, scales = "free_y", ncol = 2) +
  labs(
    title = "Trend of HIV Cases in Countries Contributing to 75% of Regional Burden",
    subtitle = "2000-2023",
    y = "Number of People Living with HIV",
    x = "Year",
    color = "Country") +
  theme_minimal(base_size = 12) +
  theme(
    legend.position = "right",
    plot.title = element_text(face = "bold", size = 14),
    plot.subtitle = element_text(size = 12),
    axis.title = element_text(face = "bold"),
    panel.grid.minor = element_blank(),
    strip.text = element_text(face = "bold", size = 10),
    legend.text = element_text(size = 8),
```

```
    legend.key.height = unit(0.8, "lines"),
    legend.key.width = unit(1, "lines")
  ) +
  guides(color = guide_legend(
    ncol = 1,
    title.position = "top",
    title.hjust = 0.5,
    override.aes = list(size = 3)))
```



2. HIV and Poverty Analysis

```
library(lme4)
library(readxl)
library(tidyverse)



# Clean HIV data
hiv_data_clean <- read_csv("HIV data 2000-2023.csv") %>%
  filter(Value != "No data") %>%
  mutate(
    Value_clean = str_replace_all(Value, "\\[.*\\]", ""),
```

```
    Value_clean = str_trim(Value_clean),
    Value_clean = str_replace_all(Value_clean, "[^0-9.]", ""),
    Value = case_when(
      str_detect(Value_clean, "^\\d+\\.?\\d*$") ~ as.numeric(Value_clean),
      TRUE ~ NA_real_
    ),
    Year = as.numeric(Period),
    Country = Location
  ) %>%
  filter(!is.na(Value)) %>%
  select(Country, Year, Value) %>%
  rename(HIV_Population = Value)

# Read the multidimensional poverty dataset
poverty_data <- read_excel("multidimensional_poverty.xlsx", skip = 2)

# Assign correct column names
colnames(poverty_data) <- c(
  "Region", "Country code", "Economy", "Reporting year", "Survey name",
  "Survey year", "Survey coverage", "Welfare type", "Survey comparability",
  "Monetary (%)", "Educational attainment (%)", "Educational enrollment (%)",
  "Electricity (%)", "Sanitation (%)", "Drinking water (%)",
  "Multidimensional poverty headcount ratio (%)"
)

# Inspect raw poverty data for non-numeric values
poverty_data %>%
  select(
    `Multidimensional poverty headcount ratio (%)`, `Monetary (%)`,
    `Educational attainment (%)`, `Educational enrollment (%)`,
    `Electricity (%)`, `Sanitation (%)`, `Drinking water (%)`,
    `Reporting year`
  ) %>%
  summarise(across(everything(), ~paste(unique(.), collapse = ", "))) %>%
  print()
```

```
## # A tibble: 1 x 8
##   Multidimensional poverty headcount rat~1 `Monetary (%)` Educational attainme~2
##   <chr>                                  <chr>           <chr>
## 1 47.2036063671112, 0.293161417357623, 0.~ 31.1220049858~ 29.7534227371215, 0.1~
## # i abbreviated names: 1: `Multidimensional poverty headcount ratio (%)`,
## #   2: `Educational attainment (%)`
## # i 5 more variables: `Educational enrollment (%)` <chr>,
## #   `Electricity (%)` <chr>, `Sanitation (%)` <chr>,
## #   `Drinking water (%)` <chr>, `Reporting year` <chr>
```

```r
# Clean and preprocess poverty data
poverty_data_clean <- poverty_data %>%
  rename(
    Country = Economy,
    Year = `Reporting year`,
    Poverty_Ratio = `Multidimensional poverty headcount ratio (%)`,
    Monetary = `Monetary (%)`,
    Edu_Attainment = `Educational attainment (%)`,
    Edu_Enrollment = `Educational enrollment (%)`,
    Electricity = `Electricity (%)`,
    Sanitation = `Sanitation (%)`,
    Drinking_Water = `Drinking water (%)`
  ) %>%
  select(Country, Year, Poverty_Ratio, Monetary, Edu_Attainment, Edu_Enrollment, Electricity, S
  # Clean non-numeric values before coercion
  mutate(
    across(c(Poverty_Ratio, Monetary, Edu_Attainment, Edu_Enrollment, Electricity, Sanitation,
           ~as.numeric(str_replace_all(., "[^0-9.]", ""))),
    Year = as.numeric(str_replace_all(Year, "[^0-9]", ""))
  ) %>%
  # Print rows with NA in poverty indicators
  { print("Rows with NA in poverty indicators:");
    print(.[rowSums(is.na(select(., Poverty_Ratio, Monetary, Edu_Attainment, Edu_Enrollment, E
    . } %>%
  filter(!is.na(Poverty_Ratio) & !is.na(Year))  # Keep rows with valid Poverty_Ratio and Year
```

```
## [1] "Rows with NA in poverty indicators:"
## # A tibble: 39 x 9
##    Country         Year Poverty_Ratio Monetary Edu_Attainment Edu_Enrollment
##    <chr>          <dbl>         <dbl>    <dbl>          <dbl>          <dbl>
##  1 Albania         2012        0.293   0.0481           0.192          NA
##  2 Australia       2010        2.22    0.517            1.71           NA
##  3 Austria         2009        0.662   0.486            0.176          NA
##  4 Belgium         2009        0.680   0.0300           0.649          NA
##  5 Bulgaria        2009        1.33    0.699            0.629          NA
##  6 Belarus         2010        3.16    0                0              NA
##  7 Switzerland     2009        0.115   0.0372           7.80           NA
##  8 Cyprus          2009        0.917   0.00530          0.912          NA
##  9 Czech Republic  2009        0.0907  0.0599           3.08           NA
## 10 Germany         2010        0.322   0.209            2.11            2.60
## # i 29 more rows
## # i 3 more variables: Electricity <dbl>, Sanitation <dbl>, Drinking_Water <dbl>
```

```r
# Merge datasets
merged_data <- inner_join(hiv_data_clean, poverty_data_clean, by = c("Country", "Year"))
```

2.1 Fit mixed effects model

```
model <- lmer(
  log(HIV_Population + 1) ~ Poverty_Ratio + Monetary + Edu_Attainment + Edu_Enrollment +
    Electricity + Sanitation + Drinking_Water + (1 | Country) + (1 | Year),
  data = merged_data,
  control = lmerControl(check.nobs.vs.nlev = "ignore",
                        check.nobs.vs.rankZ = "ignore",
                        check.nobs.vs.nRE="ignore")
)

# Tidy model output
model_tidy <- tidy(model, effects = "fixed", conf.int = TRUE)

# Display in a nice table
kable(model_tidy, digits = 3, caption = "Mixed Effects Model: Impact of Poverty Indicators on H
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                full_width = FALSE,
                position = "center")
```

Table 1: Mixed Effects Model: Impact of Poverty Indicators on HIV Population

effect	term	estimate	std.error	statistic	conf.low	conf.high
fixed	(Intercept)	9.451	0.561	16.840	8.351	10.551
fixed	Poverty_Ratio	0.051	0.185	0.277	-0.311	0.414
fixed	Monetary	0.039	0.126	0.307	-0.209	0.286
fixed	Edu_Attainment	0.071	0.060	1.195	-0.046	0.188
fixed	Edu_Enrollment	0.012	0.074	0.168	-0.133	0.158
fixed	Electricity	-0.024	0.074	-0.327	-0.169	0.120
fixed	Sanitation	-0.062	0.034	-1.835	-0.129	0.004
fixed	Drinking_Water	0.053	0.076	0.705	-0.095	0.202

The mixed-effects model indicates that improved sanitation is modestly associated with lower HIV prevalence (= -0.062), while higher educational attainment shows a positive association (= 0.071), possibly due to increased HIV awareness and testing in more educated populations. Monetary poverty did not exhibit a direct relationship with HIV rates, and access to electricity showed a weak protective effect. High multicollinearity among predictors especially between poverty ratio and monetary poverty (correlation: -0.91)combined with a limited dataset (29 observations across 3 years), reduces the precision of the estimates. **Policy implication**: Integrating sanitation improvements into HIV prevention strategies may yield measurable benefits, while the observed education effect should be interpreted with caution.

3. Mortality Analysis in East Africa

3.1 Filtering Data for EAC Member States

```
library(sf)
library(viridis)

# Read the uploaded dataset
data <- read_csv("dataset_datascience.csv")

# Define EAC countries
eac_countries <- c("Burundi",
                   "Democratic Republic of the Congo",
                   "Kenya",
                   "Rwanda",
                   "South Sudan",
                   "Uganda",
                   "United Republic of Tanzania",
                   "Somalia")

# Filter the dataset for EAC countries
eac_data <- data %>%
  filter(`Geographic area` %in% eac_countries)

# Save the filtered data
write_csv(eac_data, "eac_mortality_data.csv")

head(eac_data)
```

```
## # A tibble: 6 x 23
##   REF_AREA `Geographic area` `Regional group` Indicator  Sex   `Wealth Quintile`
##   <chr>    <chr>             <chr>            <chr>      <chr> <chr>
## 1 BDI      Burundi           <NA>             Neonatal ~ Total Total
## 2 BDI      Burundi           <NA>             Neonatal ~ Total Total
## 3 BDI      Burundi           <NA>             Neonatal ~ Total Total
## 4 BDI      Burundi           <NA>             Neonatal ~ Total Total
## 5 BDI      Burundi           <NA>             Neonatal ~ Total Total
## 6 BDI      Burundi           <NA>             Neonatal ~ Total Total
## # i 17 more variables: `Series Name` <chr>, `Series Year` <chr>,
## #   `Reference Date` <dbl>, `Observation Value` <dbl>, `Lower Bound` <dbl>,
## #   `Upper Bound` <dbl>, `Standard Error` <dbl>, `Country notes` <chr>,
## #   `Observation Status` <chr>, `Unit of measure` <chr>, `Series Type` <chr>,
## #   `Series Category` <chr>, `Series Method` <chr>, `Age Group of Women` <chr>,
## #   `Time Since First Birth` <chr>, Definition <lgl>, Interval <dbl>
```

3.2 Visualizing Latest Indicator Estimates at the Country Level

```
# Load each country's shapefile
burundi <- st_read("C:/Users/BRIAN JUMA/Desktop/CEMA_INTERN/gadm41_BDI_shp/gadm41_BDI_0.shp")
```

```
## Reading layer `gadm41_BDI_0' from data source
##   `C:\Users\BRIAN JUMA\Desktop\CEMA_INTERN\gadm41_BDI_shp\gadm41_BDI_0.shp'
##   using driver `ESRI Shapefile'
## Simple feature collection with 1 feature and 2 fields
## Geometry type: POLYGON
## Dimension:     XY
## Bounding box:  xmin: 29.00035 ymin: -4.470001 xmax: 30.85023 ymax: -2.309823
## Geodetic CRS:  WGS 84
```

```
congo <- st_read("C:/Users/BRIAN JUMA/Desktop/CEMA_INTERN/gadm41_COD_shp/gadm41_COD_0.shp")
```

```
## Reading layer `gadm41_COD_0' from data source
##   `C:\Users\BRIAN JUMA\Desktop\CEMA_INTERN\gadm41_COD_shp\gadm41_COD_0.shp'
##   using driver `ESRI Shapefile'
## Simple feature collection with 1 feature and 2 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 12.20153 ymin: -13.45248 xmax: 31.30572 ymax: 5.386098
## Geodetic CRS:  WGS 84
```

```
kenya <- st_read("C:/Users/BRIAN JUMA/Desktop/CEMA_INTERN/gadm41_KEN_shp/gadm41_KEN_0.shp")
```

```
## Reading layer `gadm41_KEN_0' from data source
##   `C:\Users\BRIAN JUMA\Desktop\CEMA_INTERN\gadm41_KEN_shp\gadm41_KEN_0.shp'
##   using driver `ESRI Shapefile'
## Simple feature collection with 1 feature and 2 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 33.90959 ymin: -4.720417 xmax: 41.92622 ymax: 5.061166
## Geodetic CRS:  WGS 84
```

```
rwanda <- st_read("C:/Users/BRIAN JUMA/Desktop/CEMA_INTERN/gadm41_RWA_shp/gadm41_RWA_0.shp")
```

```
## Reading layer `gadm41_RWA_0' from data source
##   `C:\Users\BRIAN JUMA\Desktop\CEMA_INTERN\gadm41_RWA_shp\gadm41_RWA_0.shp'
##   using driver `ESRI Shapefile'
## Simple feature collection with 1 feature and 2 fields
## Geometry type: POLYGON
## Dimension:     XY
## Bounding box:  xmin: 28.86171 ymin: -2.839973 xmax: 30.89907 ymax: -1.04745
## Geodetic CRS:  WGS 84
```

```
south_sudan <- st_read("C:/Users/BRIAN JUMA/Desktop/CEMA_INTERN/gadm41_SSD_shp/gadm41_SSD_0.shp
```

```
## Reading layer `gadm41_SSD_0' from data source
##   `C:\Users\BRIAN JUMA\Desktop\CEMA_INTERN\gadm41_SSD_shp\gadm41_SSD_0.shp'
##   using driver `ESRI Shapefile'
## Simple feature collection with 1 feature and 2 fields
## Geometry type: POLYGON
## Dimension:     XY
## Bounding box:  xmin: 24.15193 ymin: 3.480999 xmax: 35.86995 ymax: 12.219
## Geodetic CRS:  WGS 84
```

```
uganda <- st_read("C:/Users/BRIAN JUMA/Desktop/CEMA_INTERN/gadm41_UGA_shp/gadm41_UGA_0.shp")
```

```
## Reading layer `gadm41_UGA_0' from data source
##   `C:\Users\BRIAN JUMA\Desktop\CEMA_INTERN\gadm41_UGA_shp\gadm41_UGA_0.shp'
##   using driver `ESRI Shapefile'
## Simple feature collection with 1 feature and 2 fields
## Geometry type: POLYGON
## Dimension:     XY
## Bounding box:  xmin: 29.5715 ymin: -1.48214 xmax: 35.00027 ymax: 4.234466
## Geodetic CRS:  WGS 84
```

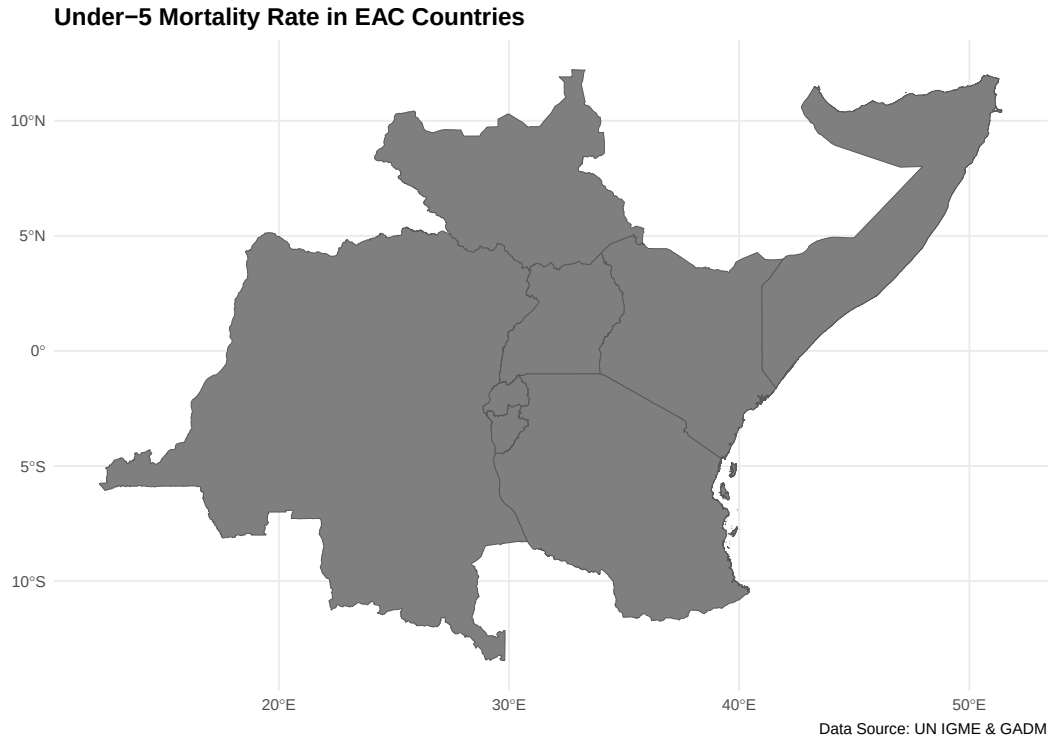
```
tanzania <- st_read("C:/Users/BRIAN JUMA/Desktop/CEMA_INTERN/gadm41_TZA_shp/gadm41_TZA_0.shp")
```

```
## Reading layer `gadm41_TZA_0' from data source
##   `C:\Users\BRIAN JUMA\Desktop\CEMA_INTERN\gadm41_TZA_shp\gadm41_TZA_0.shp'
##   using driver `ESRI Shapefile'
## Simple feature collection with 1 feature and 2 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 29.32717 ymin: -11.7457 xmax: 40.44514 ymax: -0.9857875
## Geodetic CRS:  WGS 84
```

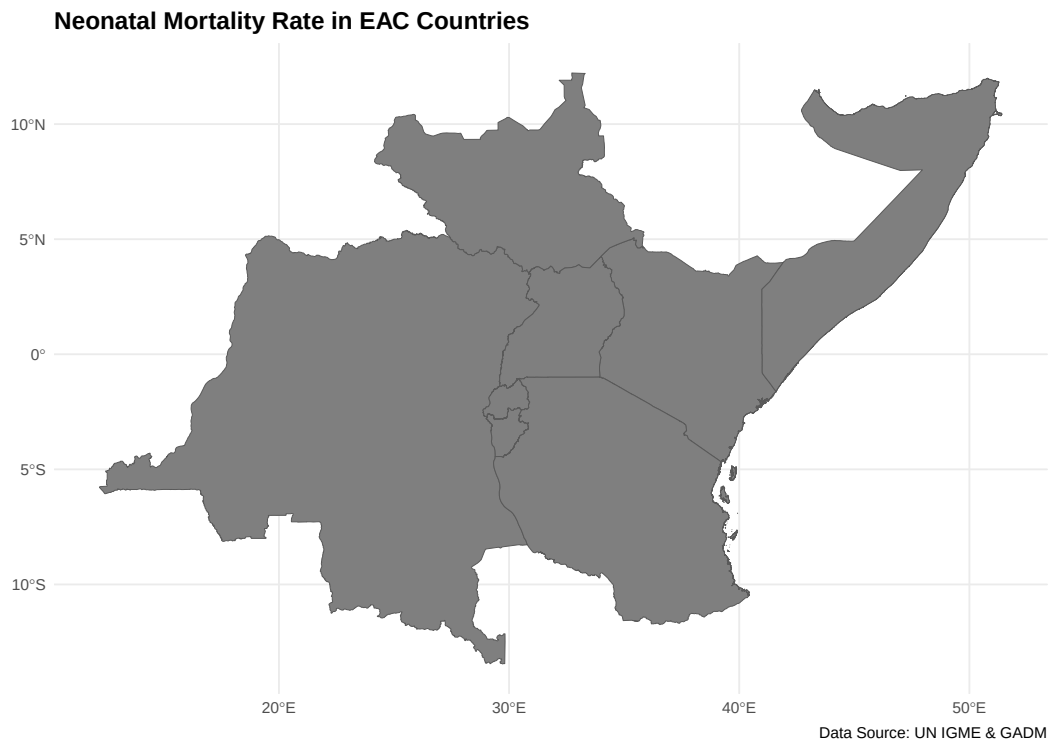
```
somalia <- st_read("C:/Users/BRIAN JUMA/Desktop/CEMA_INTERN/gadm41_SOM_shp/gadm41_SOM_0.shp")
```

```
## Reading layer `gadm41_SOM_0' from data source
##   `C:\Users\BRIAN JUMA\Desktop\CEMA_INTERN\gadm41_SOM_shp\gadm41_SOM_0.shp'
##   using driver `ESRI Shapefile'
## Simple feature collection with 1 feature and 2 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 40.9785 ymin: -1.647082 xmax: 51.4157 ymax: 11.98931
## Geodetic CRS:  WGS 84
```

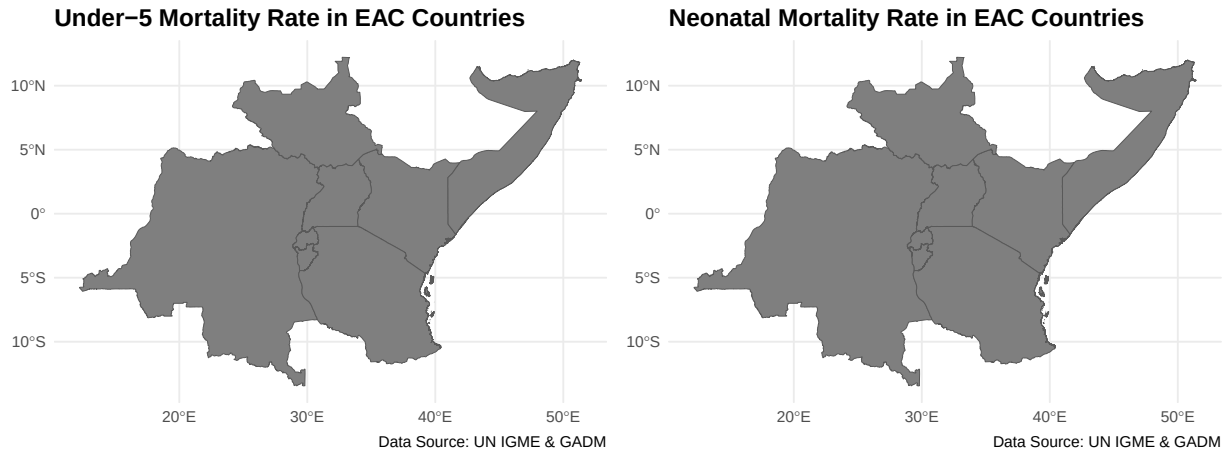
```r
# Combine all countries into one map
eac_map <- rbind(burundi, congo, kenya, rwanda,
                 south_sudan, uganda, tanzania, somalia)

# Filter only Under-5 Mortality, latest year
under5_data <- eac_data %>%
  filter(Indicator == "Under-five mortality rate",
         `Series Year` == max(`Series Year`, na.rm = TRUE)) %>%
  select(REF_AREA, `Observation Value`) %>%
  rename(Under5_Mortality = `Observation Value`)

# Merge Under-5 mortality data with shapefiles
eac_map_data <- eac_map %>%
  left_join(under5_data, by = c("COUNTRY" = "REF_AREA"))

# Plot Under-5 Mortality Map
under5_plot <- ggplot(data = eac_map_data) +
  geom_sf(aes(fill = Under5_Mortality)) +
  scale_fill_viridis_c(option = "plasma", name = "Under-5 Mortality") +
  theme_minimal() +
  labs(title = "Under-5 Mortality Rate in EAC Countries",
       caption = "Data Source: UN IGME & GADM") +
  theme(plot.title = element_text(size = 14, face = "bold"))
under5_plot
```



```r
# Filter only Neonatal Mortality, latest year
neonatal_data <- eac_data %>%
  filter(Indicator == "Neonatal mortality rate",
         `Series Year` == max(`Series Year`, na.rm = TRUE)) %>%
  select(REF_AREA, `Observation Value`) %>%
  rename(Neonatal_Mortality = `Observation Value`)

# Merge Neonatal Mortality data with shapefiles
eac_map_data <- eac_map_data %>%
  left_join(neonatal_data, by = c("COUNTRY" = "REF_AREA"))

# Plot Neonatal Mortality Map
neonatal_plot <- ggplot(data = eac_map_data) +
  geom_sf(aes(fill = Neonatal_Mortality)) +
  scale_fill_viridis_c(option = "magma", name = "Neonatal Mortality") +
  theme_minimal() +
  labs(title = "Neonatal Mortality Rate in EAC Countries",
       caption = "Data Source: UN IGME & GADM") +
  theme(plot.title = element_text(size = 14, face = "bold"))
neonatal_plot
```




```r
# Display side-by-side maps
under5_plot + neonatal_plot  # Using patchwork
```

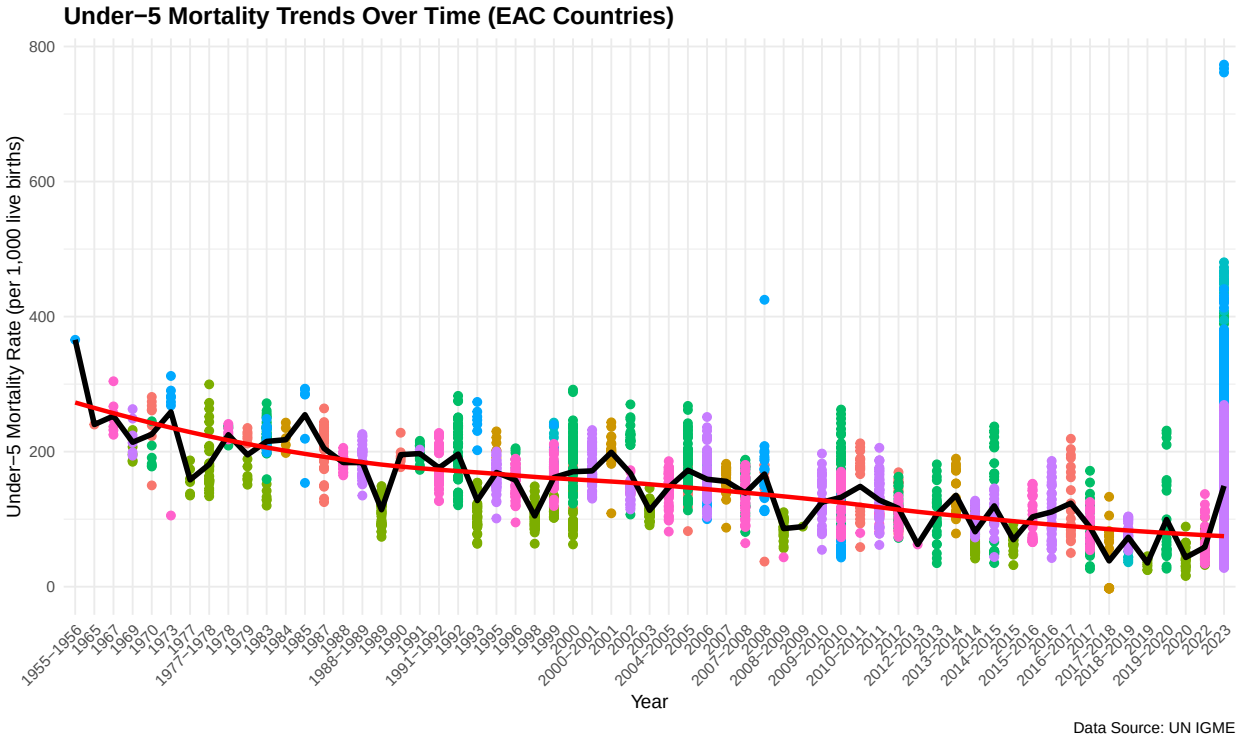


3.3 Analyzing Average Mortality Rate Trends Over Time

3.3.1 Under-5 Mortality Trends

```
# Prepare Under-5 Mortality data for trend plotting
under5_data_trend <- eac_data %>%
  filter(Indicator == "Under-five mortality rate") %>%
  select(`Geographic area`, `Series Year`, `Observation Value`) %>%
  rename(Country = `Geographic area`,
         Year = `Series Year`,
         Under5_Mortality = `Observation Value`)

# Calculate average Under-5 Mortality per year
under5_avg <- under5_data_trend %>%
  group_by(Year) %>%
  summarise(Average_Under5_Mortality = mean(Under5_Mortality, na.rm = TRUE))

# Plot Under-5 Mortality Trends
under5_trend_plot <- ggplot() +
  geom_point(data = under5_data_trend, aes(x = Year, y = Under5_Mortality, color = Country), si
  geom_line(data = under5_avg, aes(x = Year, y = Average_Under5_Mortality, group = 1), color =
  geom_smooth(data = under5_avg, aes(x = Year, y = Average_Under5_Mortality, group = 1),
              method = "loess", formula = 'y ~ x', color = "red", se = FALSE, linewidth = 1.2)
```

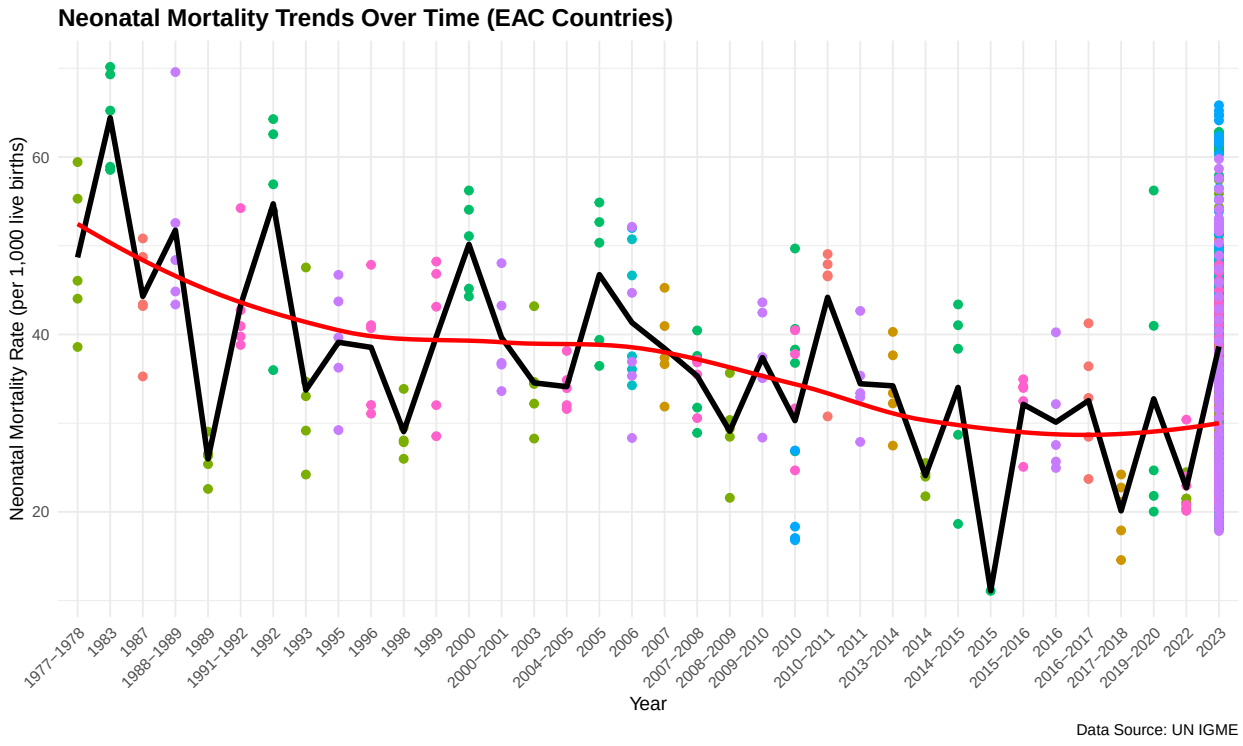
```
  theme_minimal() +
  labs(title = "Under-5 Mortality Trends Over Time (EAC Countries)",
       x = "Year", y = "Under-5 Mortality Rate (per 1,000 live births)",
       caption = "Data Source: UN IGME") +
  theme(plot.title = element_text(size = 14, face = "bold"),
        axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = "none")

# View Under-5 Trend Plot
under5_trend_plot
```



3.3.2 Neonatal Mortality Trends

```
# Prepare Neonatal Mortality data for trend plotting
neonatal_data_trend <- eac_data %>%
  filter(Indicator == "Neonatal mortality rate") %>%
  select(`Geographic area`, `Series Year`, `Observation Value`) %>%
  rename(Country = `Geographic area`,
         Year = `Series Year`,
         Neonatal_Mortality = `Observation Value`)

# Calculate average Neonatal Mortality per year
```

```r
neonatal_avg <- neonatal_data_trend %>%
  group_by(Year) %>%
  summarise(Average_Neonatal_Mortality = mean(Neonatal_Mortality, na.rm = TRUE))

# Plot Neonatal Mortality Trends
neonatal_trend_plot <- ggplot() +
  geom_point(data = neonatal_data_trend, aes(x = Year, y = Neonatal_Mortality, color = Country)
  geom_line(data = neonatal_avg, aes(x = Year, y = Average_Neonatal_Mortality, group = 1), colo
  geom_smooth(data = neonatal_avg, aes(x = Year, y = Average_Neonatal_Mortality, group = 1),
              method = "loess", formula = 'y ~ x', color = "red", se = FALSE, linewidth = 1.2)
  theme_minimal() +
  labs(title = "Neonatal Mortality Trends Over Time (EAC Countries)",
       x = "Year", y = "Neonatal Mortality Rate (per 1,000 live births)",
       caption = "Data Source: UN IGME") +
  theme(plot.title = element_text(size = 14, face = "bold"),
        axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = "none")

# View Neonatal Trend Plot
neonatal_trend_plot
```



3.4 Identifying Countries with Highest Mortality Rates

3.4.1 Country with the highest Under-five mortality rate (U5MR)

```
# Identify the country with the highest Under-five mortality rate (U5MR)
highest_u5mr <- eac_data %>%
  filter(Indicator == "Under-five mortality rate") %>%
  group_by(`Geographic area`) %>%
  summarise(max_u5mr = max(`Observation Value`, na.rm = TRUE)) %>%
  arrange(desc(max_u5mr)) %>%
  head(1)
print(highest_u5mr)
```

```
## # A tibble: 1 x 2
##   `Geographic area` max_u5mr
##   <chr>                <dbl>
## 1 South Sudan           773.
```

3.4.2 Country with the highest Neonatal mortality rate (NMR)

```
highest_nmr <- eac_data %>%
  filter(Indicator == "Neonatal mortality rate") %>%
  group_by(`Geographic area`) %>%
  summarise(max_nmr = max(`Observation Value`, na.rm = TRUE)) %>%
  arrange(desc(max_nmr)) %>%
  head(1)
print(highest_nmr)
```

```
## # A tibble: 1 x 2
##   `Geographic area` max_nmr
##   <chr>               <dbl>
## 1 Rwanda               70.2
```