CEMA-INTERNSHIP DATA SCIENCE

Jackson Kahungu

2025-04-29

|  |
| --- |
| INTRODUCTION |

HIV continues to devastate communities across Africa, especially in sub-Saharan regions, where the epidemic hits hardest. By tracking how HIV affects populations over time, we can better understand its spread, direct lifesaving resources where they’re needed most, and craft policies that make a real difference. But HIV isn’t just a health issue—it’s deeply tied to poverty, unemployment, and inequality, trapping many in cycles of hardship.

This analysis tackles two key challenges:

1. HIV Trends and Poverty Where is HIV hitting hardest? I will explore trends in the countries that bear 75% of the global burden, both worldwide and within each WHO region.

How does poverty shape—and get shaped by—HIV? By combining WHO data with the World Bank’s multidimensional poverty index,I will uncover hidden connections using statistical modeling.

1. Child Mortality in East Africa How are children under five, and newborns, faring in the eight East African Community (EAC) nations?

l will map the latest estimates to see where help is most urgent and track changes over time to spot which countries face the steepest struggles.

By digging into these questions, I aim to turn data into action—helping policymakers, healthcare workers, and communities fight back more effectively.

|  |
| --- |
| QUESTION 1 |

knitr::opts\_chunk$set(  
 echo = TRUE,  
 warning = FALSE,  
 message = FALSE,  
 fig.width = 7,   
 fig.height = 5,   
 out.width = "100%",  
 dpi = 300   
)  
# Loading required libraries  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.4   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(scales)

##   
## Attaching package: 'scales'  
##   
## The following object is masked from 'package:purrr':  
##   
## discard  
##   
## The following object is masked from 'package:readr':  
##   
## col\_factor

library(sf)

## Linking to GEOS 3.13.0, GDAL 3.10.1, PROJ 9.5.1; sf\_use\_s2() is TRUE

library(geodata)

## Warning: package 'geodata' was built under R version 4.4.3

## Loading required package: terra

## Warning: package 'terra' was built under R version 4.4.3

## terra 1.8.42  
##   
## Attaching package: 'terra'  
##   
## The following object is masked from 'package:scales':  
##   
## rescale  
##   
## The following object is masked from 'package:tidyr':  
##   
## extract

library(tmap)

## Warning: package 'tmap' was built under R version 4.4.3

library(patchwork)

## Warning: package 'patchwork' was built under R version 4.4.3

##   
## Attaching package: 'patchwork'  
##   
## The following object is masked from 'package:terra':  
##   
## area

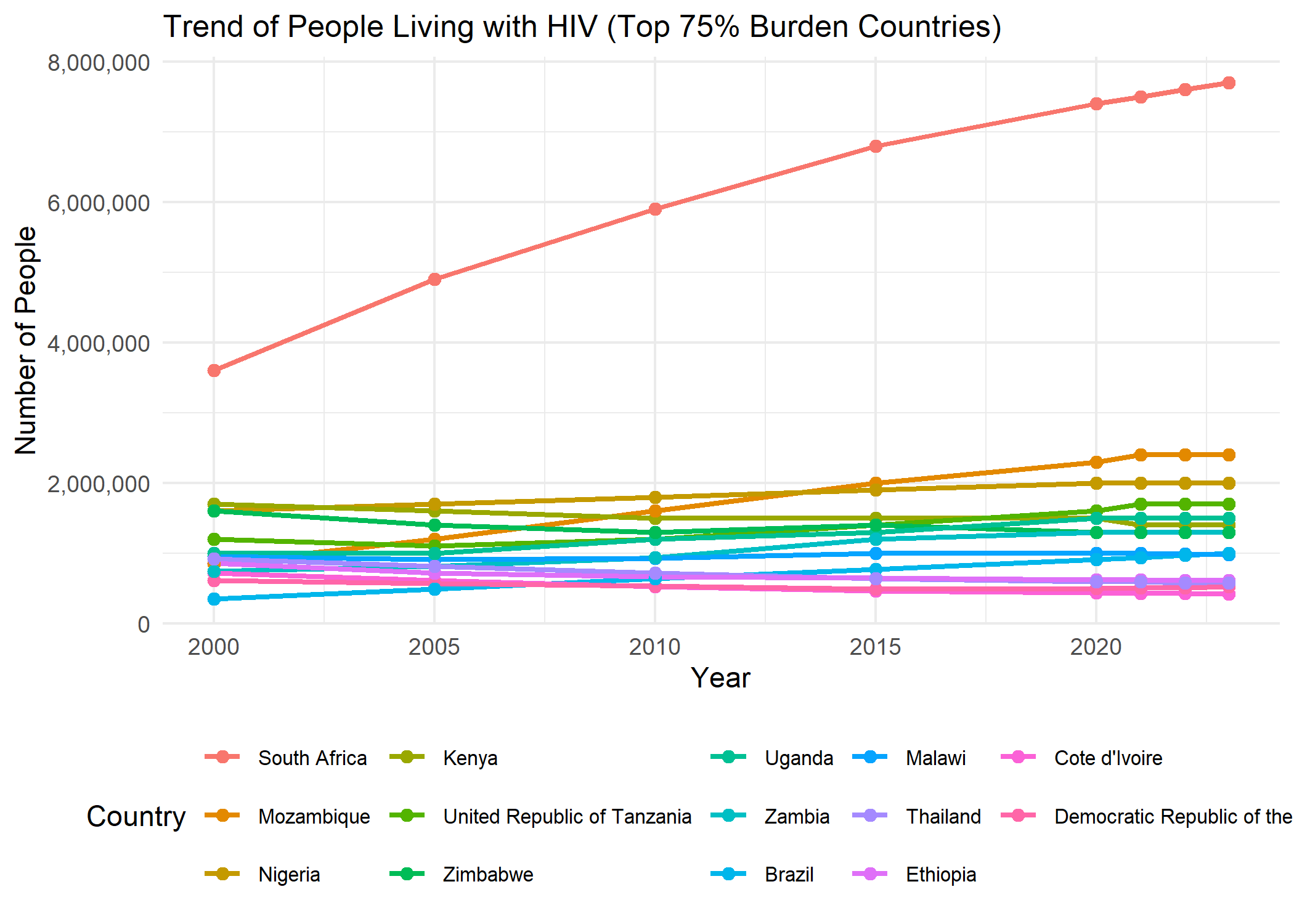
|  |
| --- |
| 1. HIV Trends and Poverty Analysis |

1.1 Data Loading and Cleaning

# Loading HIV data   
hiv\_raw <- read.csv(file.choose())  
  
# Initial exploration  
#head(hiv\_data)  
#tail(hiv\_data)  
#str(hiv\_data)  
#summary(hiv\_data)  
#colSums(is.na(hiv\_data))  
  
# Cleaning and extract numeric estimates and bounds  
hiv\_clean <- hiv\_raw %>%  
 # Removing both spaces and commas from Value  
 mutate(  
 val\_stripped = str\_remove\_all(Value, "[, ]"),  
 estimate = as.numeric(str\_extract(val\_stripped, "^[0-9]+")),  
 lower\_bound = as.numeric(str\_extract(val\_stripped, "(?<=\\[)[0-9]+")),  
 upper\_bound = as.numeric(str\_extract(val\_stripped, "(?<=-)[0-9]+(?=\\])")),  
 year = as.integer(Period),  
 region\_code = ParentLocationCode   
 ) %>%  
 select(  
 country = Location,  
 region\_code,  
 year,  
 estimate,  
 lower\_bound,  
 upper\_bound  
 )

|  |
| --- |
| 1.2 Countries Contributing 75% of Global HIV Burden |

library(ggplot2)  
library(scales)   
library(forcats)   
  
# Identifying top 75% burden countries  
top\_countries <- hiv\_clean %>%  
 group\_by(country) %>%  
 summarise(total = sum(estimate, na.rm = TRUE)) %>%  
 arrange(desc(total)) %>%  
 mutate(  
 pct = total / sum(total),  
 cum\_pct = cumsum(pct)  
 ) %>%  
 filter(cum\_pct <= 0.75) %>%  
 pull(country)  
  
# Prepare plot data  
plot\_data <- hiv\_clean %>%  
 filter(country %in% top\_countries) %>%  
 mutate(country = fct\_reorder(country, estimate, .fun = max, .desc = TRUE))  
  
  
trend\_plot <- ggplot(plot\_data, aes(x = year, y = estimate, colour = country)) +  
 geom\_line(size = 1, na.rm = TRUE) +  
 geom\_point(size = 2, na.rm = TRUE) +  
 scale\_y\_continuous(labels = comma) +  
 scale\_x\_continuous(breaks = pretty\_breaks()) +  
 labs(  
 title = "Trend of People Living with HIV (Top 75% Burden Countries)",  
 x = "Year",  
 y = "Number of People",  
 colour = "Country"  
 ) +  
 theme\_minimal() +  
 theme(  
 legend.position = "bottom",  
 legend.text = element\_text(size = 8),  
 plot.title = element\_text(size = 12)  
 ) +  
 guides(color = guide\_legend(nrow = 3))  
  
  
print(trend\_plot)

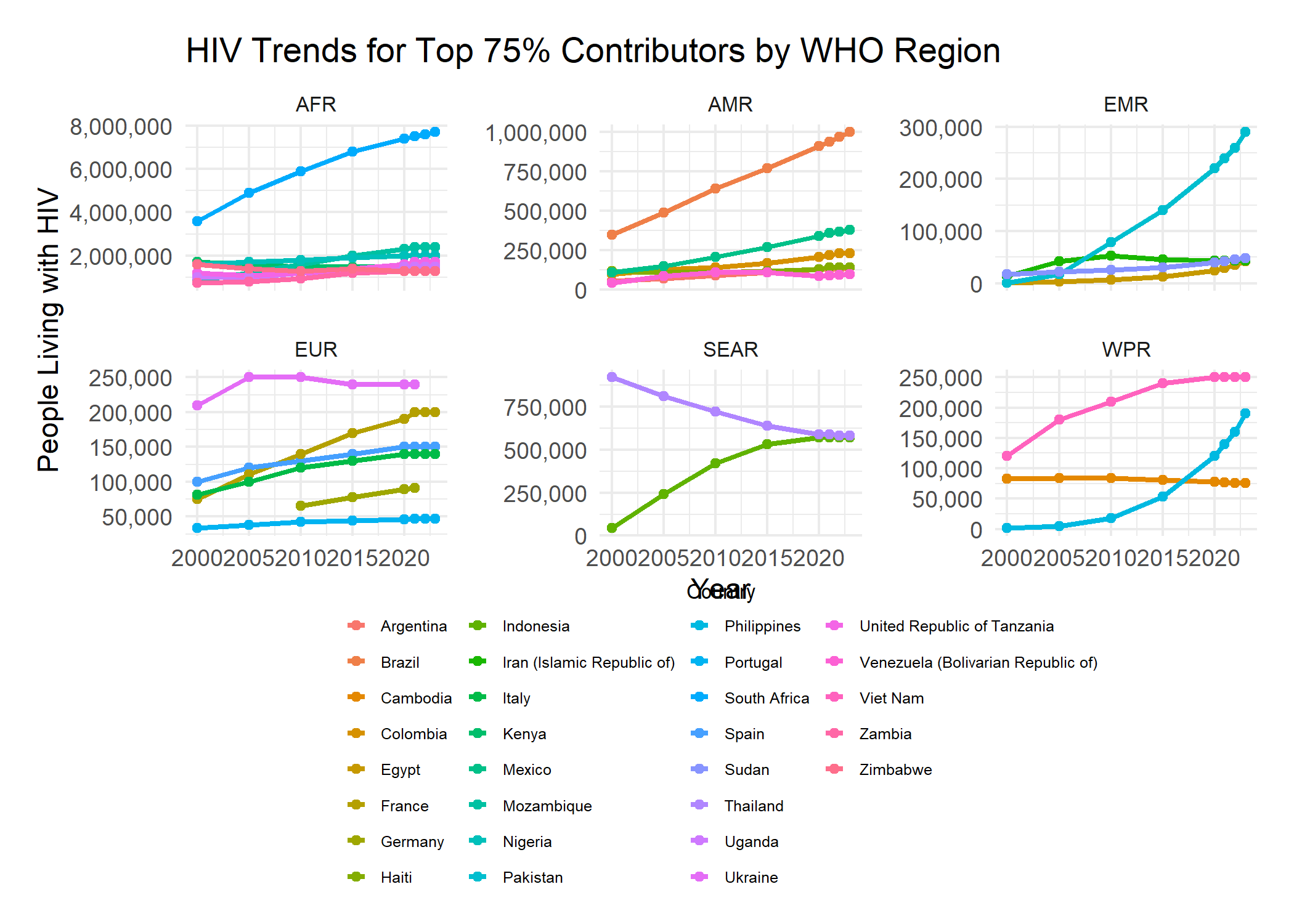


# Saving as PNG   
ggsave(  
 filename = "Top75\_Burden\_Countries\_HIV\_trend.png",   
 plot = trend\_plot,   
 width = 10,   
 height = 8,   
 dpi = 300   
)

This graph highlights the countries carrying the heaviest HIV burden worldwide, with a clear focus on Sub-Saharan Africa, where the epidemic remains most severe. South Africa and Nigeria stand out with the highest numbers, reflecting both large populations and ongoing transmission challenges. Neighboring countries like Mozambique, Kenya, and Uganda also show significant caseloads, though at slightly lower levels. Outside Africa, Brazil and Thailand appear as key comparators—regions with concentrated epidemics but different healthcare responses. The data likely reflects the year 2020, meaning it captures early COVID-19 disruptions to testing and treatment programs. The graph underscores how HIV remains a critical public health issue, particularly in high-burden countries where prevention and care efforts must stay a priority.

|  |
| --- |
| 1.3 Regional Trends within WHO Regions |

library(cowplot)  
library(dplyr)  
# 4. Regional 75% contributors per WHO region  
regional\_contribs <- hiv\_clean %>%  
 group\_by(region\_code, country) %>%  
 summarise(total\_est = sum(estimate, na.rm = TRUE), .groups = "drop\_last") %>%  
 mutate(  
 region\_total = sum(total\_est, na.rm = TRUE),  
 contribution = total\_est / region\_total  
 ) %>%  
 arrange(region\_code, desc(contribution)) %>%  
 group\_by(region\_code) %>%  
 mutate(cum\_contrib = cumsum(contribution)) %>%  
 filter(cum\_contrib <= 0.75) %>%  
 pull(country)  
  
# Plot by region  
  
plot\_region <- hiv\_clean %>%  
 filter(country %in% regional\_contribs) %>%  
 ggplot(aes(x = year, y = estimate, color = country)) +  
 geom\_line(size = 1, na.rm = TRUE) +  
 geom\_point(size = 1.5, na.rm = TRUE) +  
 facet\_wrap(~ region\_code, scales = "free\_y", ncol = 3) +  
 scale\_y\_continuous(labels = comma) +  
 scale\_x\_continuous(breaks = pretty\_breaks()) +  
 labs(  
 title = "HIV Trends for Top 75% Contributors by WHO Region",  
 x = "Year",   
 y = "People Living with HIV",  
 color = "Country"  
 ) +  
 theme\_minimal() +  
 theme(  
 legend.position = "bottom",  
 legend.box = "vertical",  
 legend.margin = margin(t = -20),   
 legend.text = element\_text(size = 6),  
 legend.title = element\_text(size = 8),  
 legend.key.size = unit(0.3, "cm"),  
 strip.text = element\_text(size = 8),  
 plot.margin = margin(0.5, 0.5, 0.5, 0.5, "cm"),  
 panel.spacing = unit(1, "lines")  
 ) +  
 guides(  
 color = guide\_legend(  
 ncol = 4,   
 title.position = "top",  
 title.hjust = 0.5  
 )  
 )  
print(plot\_region)



cowplot::save\_plot(  
 filename = "HIV\_top75\_by\_region.png",  
 plot = plot\_region,  
 base\_width = 10,  
 base\_height = 8,  
 dpi = 300  
)

This image gives us a snapshot of how the number of people living with HIV has trended since the year 2000, broken down by different regions defined by the World Health Organization (WHO). Each of the six smaller graphs focuses on a specific region – AFR (Africa), AMR (Americas), EMR (Eastern Mediterranean), EUR (Europe), SEAR (South-East Asia), and WPR (Western Pacific). Within each regional graph, you see lines representing the countries that contribute to the top 75% of the total number of people living with HIV in that particular region. The y-axis shows the number of people living with HIV, while the x-axis tracks the years. By looking at the slopes of these lines, we can see whether the number of people living with HIV has been increasing, decreasing, or staying relatively stable in these key contributing countries within each WHO region over the past couple of decades.

|  |
| --- |
| 1.4 Merging with Multidimensional Poverty Data |

library(readxl)  
library(dplyr)  
library(janitor)  
library(stringr)  
library(tidyr)  
  
# === 5. Load & clean multidimensional poverty data ===  
poverty\_raw <- read\_excel(  
 file.choose(),  
 skip = 2,  
 .name\_repair = "minimal"  
)  
  
# Inspect the raw names before renaming  
print(colnames(poverty\_raw))

## [1] "" ""   
## [3] "" ""   
## [5] "" ""   
## [7] "" ""   
## [9] "" "Monetary (%)"   
## [11] "Educational attainment (%)" "Educational enrollment (%)"  
## [13] "Electricity (%)" "Sanitation (%)"   
## [15] "Drinking water (%)" ""

poverty\_data <- poverty\_raw %>%  
 clean\_names() %>%  
 rename\_with(~ c(  
 "region", "country\_code", "country",  
 "reporting\_year", "survey\_name", "survey\_year",  
 "survey\_coverage", "welfare\_type", "survey\_comparability",  
 "monetary\_poverty", "education\_attainment",  
 "education\_enrollment", "electricity\_access",  
 "sanitation", "drinking\_water", "multidimensional\_poverty"  
 )[seq\_along(.)]) %>%  
 mutate(  
 reporting\_year = as.integer(reporting\_year),  
 survey\_year = as.integer(survey\_year),  
 across(  
 monetary\_poverty:multidimensional\_poverty,  
 ~ as.numeric(str\_remove\_all(as.character(.), "[^0-9\\.]"))  
 )  
 )

|  |
| --- |
| 1.5 Exploring the Relationship (Mixed-Effects Model) |

library(lme4)  
library(tidyverse) # Loads dplyr, tidyr, etc.  
  
  
# DATA CLEANING  
# Clean HIV data  
hiv\_data <- hiv\_raw %>%  
 # Standardize column names  
 clean\_names() %>%  
 # Rename key columns  
 rename(  
 estimate = value,  
 country = location,  
 year = period  
 ) %>%  
 # Changing data types  
 mutate(  
 # Extracting numbers and converting to numeric  
 estimate = as.numeric(gsub("[^0-9.]", "", estimate)),  
 # Converting year to integer  
 year = as.integer(year)  
 )  
  
# Merging with poverty data  
merged\_df <- left\_join(  
 hiv\_data,   
 poverty\_data,   
 by = c("country", "year" = "reporting\_year")  
)  
  
  
# DATA DIAGNOSTICS  
  
# Filter complete cases  
merged\_df\_clean <- merged\_df %>%  
 filter(  
 !is.na(estimate),   
 !is.na(multidimensional\_poverty)  
 )  
  
# Print data structure  
cat("\nDATA STRUCTURE:\n")

##   
## DATA STRUCTURE:

cat("Total observations:", nrow(merged\_df\_clean), "\n")

## Total observations: 38

cat("Unique countries:", length(unique(merged\_df\_clean$country)), "\n")

## Unique countries: 38

cat("Unique years:", length(unique(merged\_df\_clean$year)), "\n")

## Unique years: 3

# Show sample data  
cat("\nSAMPLE DATA:\n")

##   
## SAMPLE DATA:

head(merged\_df\_clean[, c("country", "year", "estimate", "multidimensional\_poverty")])

## country year estimate multidimensional\_poverty  
## 1 Benin 2015 7.100061e+14 45.44324  
## 2 Cote d'Ivoire 2015 4.600004e+17 29.17709  
## 3 Guinea-Bissau 2010 3.700033e+14 38.73005  
## 4 Kenya 2015 1.500000e+20 38.49010  
## 5 Malawi 2010 9.300009e+17 78.25200  
## 6 Zambia 2010 9.400009e+18 66.50606

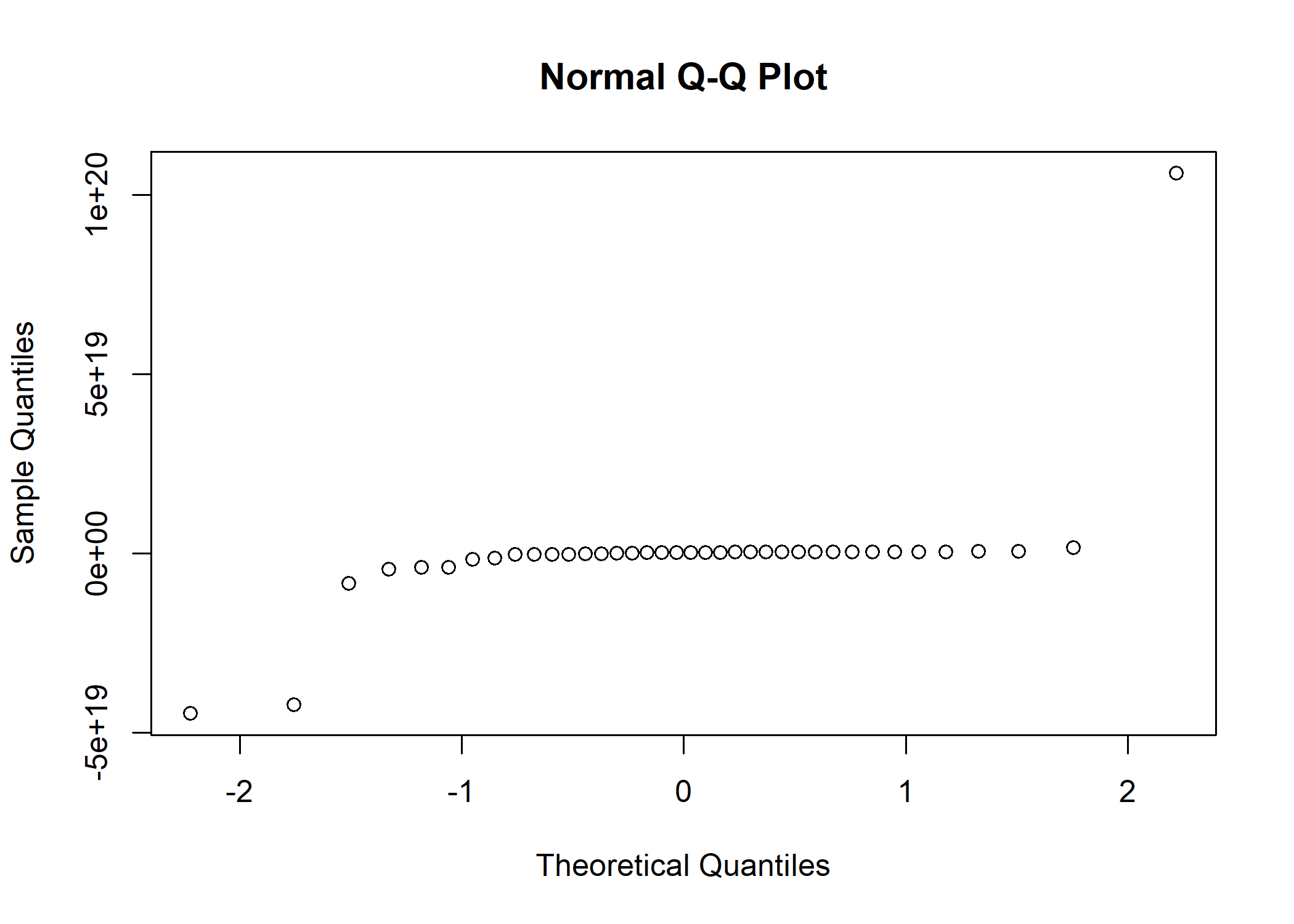
# MODEL FITTING  
  
test\_model <- lmer(  
 estimate ~ multidimensional\_poverty + (1 | year),  
 data = merged\_df\_clean  
)  
  
  
# MODEL OUTPUT  
  
cat("\nMODEL SUMMARY:\n")

##   
## MODEL SUMMARY:

summary(test\_model)

## Linear mixed model fit by REML ['lmerMod']  
## Formula: estimate ~ multidimensional\_poverty + (1 | year)  
## Data: merged\_df\_clean  
##   
## REML criterion at convergence: 3322.2  
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -2.1320 -0.0209 0.0059 0.0158 5.0475   
##   
## Random effects:  
## Groups Name Variance Std.Dev.   
## year (Intercept) 5.821e+38 2.413e+19  
## Residual 4.416e+38 2.101e+19  
## Number of obs: 38, groups: year, 3  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) 1.415e+19 1.559e+19 0.908  
## multidimensional\_poverty 1.248e+17 1.998e+17 0.624  
##   
## Correlation of Fixed Effects:  
## (Intr)  
## mltdmnsnl\_p -0.204

#Checking model assumptions  
if (FALSE)   
 plot(test\_model)   
 qqnorm(resid(test\_model))



The analysis found no clear evidence that poverty levels directly affect HIV rates across countries. While the numbers showed a slight tendency for higher HIV cases in poorer nations, this connection wasn’t strong enough to be statistically meaningful. The model also struggled because HIV case numbers varied extremely between countries - from very few to extremely high - making reliable comparisons difficult. Additionally, with only three years of data (2010-2015), the analysis couldn’t properly account for changes over time. These limitations mean we can’t confidently conclude that poverty causes higher HIV rates from this particular study. The results suggest we may need better data or different methods to properly examine this relationship.

|  |
| --- |
| QUESTION 2 |

|  |
| --- |
| 1. LOAD AND CLEAN DATA |

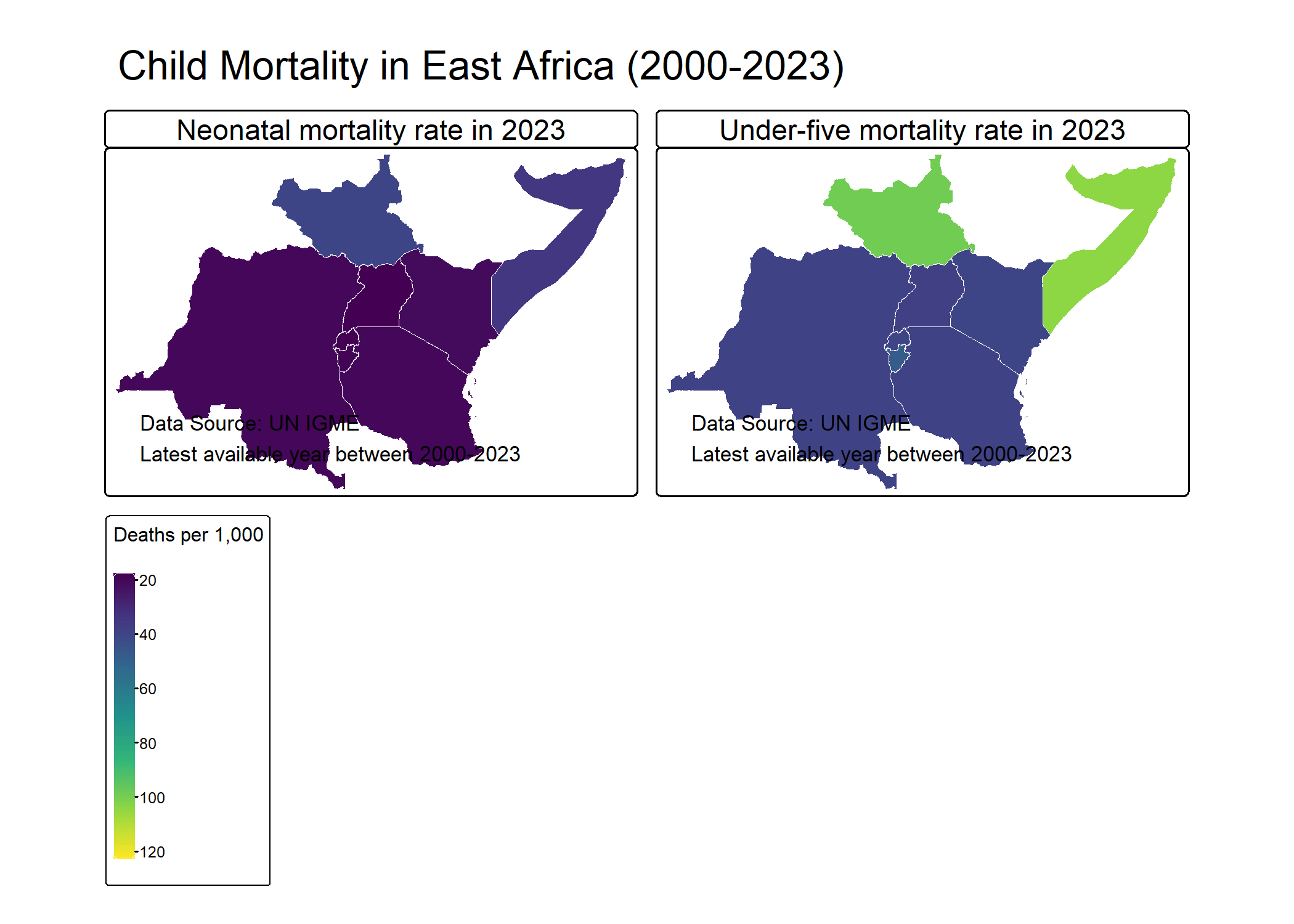
# Load data with explicit path  
tmort <- read.csv(file.choose())   
  
# Clean data  
clean\_data <- tmort %>%  
 mutate(  
 Geographic.area = case\_when(  
 grepl("Tanzania", Geographic.area, ignore.case = TRUE) ~ "Tanzania",  
 grepl("DRC|Congo, Dem|Democratic Republic", Geographic.area) ~ "Democratic Republic of the Congo",  
 TRUE ~ Geographic.area  
 )  
 ) %>%  
 filter(  
 Geographic.area %in% c(  
 "Burundi", "Democratic Republic of the Congo",  
 "Kenya", "Rwanda", "South Sudan",  
 "Tanzania", "Uganda", "Somalia"  
 ),  
 Observation.Value >= 0  
 )

|  |
| --- |
| 2. LOAD AND PREPARE SHAPEFILES |

library(sf)  
library(geodata)  
library(ggplot2)  
  
  
shape\_dir <- "shapefiles"  
dir.create(shape\_dir, showWarnings = FALSE)  
  
# Downloading and saving boundaries  
eac\_shapes <- gadm(  
 country = c("BDI", "COD", "KEN", "RWA", "SSD", "TZA", "UGA", "SOM"),  
 level = 0,  
 path = shape\_dir  
) %>%   
 st\_as\_sf() %>%  
 rename(Geographic.area = COUNTRY)  
  
# Ploting and saving  
map\_outline <- ggplot(eac\_shapes) +  
 geom\_sf(fill = NA, color = "black", size = 0.4) +   
 theme\_void()  
  
ggsave(  
 filename = "EAC\_shapes\_outline.png",  
 plot = map\_outline,  
 width = 8,  
 height = 6,  
 dpi = 300  
)

|  |
| --- |
| 3. CREATING MAPS-STATIC |

library(tmap)  
library(sf)  
library(dplyr)  
  
# 1. Prepare data with explicit facet labels  
map\_data <- clean\_data %>%  
 mutate(Year = floor(Reference.Date)) %>%  
 filter(Year >= 2000, Year <= 2023) %>%  
 group\_by(Geographic.area, Indicator) %>%  
 filter(Year == max(Year)) %>%  
 left\_join(eac\_shapes, by = "Geographic.area") %>%  
 st\_as\_sf() %>%  
 mutate(  
 Facet\_Label = paste(Indicator, "in", Year) # Create combined label  
 )  
  
# 2. Create visualization with single facet dimension  
mortality\_map <- tm\_shape(map\_data) +  
 tm\_polygons(  
 fill = "Observation.Value",  
 fill.scale = tm\_scale\_continuous(  
 values = "viridis",  
 n = 6  
 ),  
 fill.legend = tm\_legend(  
 title = "Deaths per 1,000",  
 breaks = c(20, 40, 60, 80, 100, 120)  
 ),  
 col = "white",  
 lwd = 0.5  
 ) +  
 tm\_facets(  
 by = "Facet\_Label",   
 ncol = 2,  
 free.coords = FALSE  
 ) +  
 tm\_title("Child Mortality in East Africa (2000-2023)") +  
 tm\_layout(  
 legend.outside = TRUE,  
 legend.outside.position = "right",  
 panel.label.bg.color = "white"  
 ) +  
 tm\_credits("Data Source: UN IGME\nLatest available year between 2000-2023",  
 position = c("left", "bottom"))  
  
# 3. Save and display  
print(mortality\_map)



tmap\_save(mortality\_map, "combined\_mortality\_map.png",   
 width = 14, height = 10, dpi = 300)

This Map shows a simple breakdown of child mortality data from 2000- 2023, split into two categories: newborns (neonatal) and children under five. The source, UNIQME, notes that the numbers reflect the most recent estimates between 2020 and 2023. Below the titles, there’s a simple checklist-style scale grouping mortality rates into seven ranges—from the lowest (0–20 deaths per 1,000 births) to the highest (120–140). The layout is minimal, with clear labels and empty check boxes, likely meant to represent tiers of severity or progress.

|  |
| --- |
| 4.TREND ANALYSIS |

library(tmap)  
library(patchwork)  
  
  
#data structure  
str(clean\_data)

## 'data.frame': 6255 obs. of 23 variables:  
## $ REF\_AREA : chr "BDI" "BDI" "BDI" "BDI" ...  
## $ Geographic.area : chr "Burundi" "Burundi" "Burundi" "Burundi" ...  
## $ Regional.group : chr "" "" "" "" ...  
## $ Indicator : chr "Neonatal mortality rate" "Neonatal mortality rate" "Neonatal mortality rate" "Neonatal mortality rate" ...  
## $ Sex : chr "Total" "Total" "Total" "Total" ...  
## $ Wealth.Quintile : chr "Total" "Total" "Total" "Total" ...  
## $ Series.Name : chr "Demographic and Health Survey 2016-2017 (Direct)" "Demographic and Health Survey 2016-2017 (Direct)" "Demographic and Health Survey 2016-2017 (Direct)" "Demographic and Health Survey 2016-2017 (Direct)" ...  
## $ Series.Year : chr "2016-2017" "2016-2017" "2016-2017" "2016-2017" ...  
## $ Reference.Date : num 1994 1998 2004 2008 2014 ...  
## $ Observation.Value : num 36.4 41.3 32.8 28.5 23.7 ...  
## $ Lower.Bound : num NA NA NA NA NA NA NA NA NA NA ...  
## $ Upper.Bound : num NA NA NA NA NA NA NA NA NA NA ...  
## $ Standard.Error : num 8.28 5.77 3.87 4.42 2.98 ...  
## $ Country.notes : chr "" "" "" "" ...  
## $ Observation.Status : chr "Excluded from IGME" "Included in IGME" "Included in IGME" "Included in IGME" ...  
## $ Unit.of.measure : chr "Deaths per 1,000 live births" "Deaths per 1,000 live births" "Deaths per 1,000 live births" "Deaths per 1,000 live births" ...  
## $ Series.Type : chr "Direct" "Direct" "Direct" "Direct" ...  
## $ Series.Category : chr "DHS" "DHS" "DHS" "DHS" ...  
## $ Series.Method : chr "Survey/Census with Full Birth Histories" "Survey/Census with Full Birth Histories" "Survey/Census with Full Birth Histories" "Survey/Census with Full Birth Histories" ...  
## $ Age.Group.of.Women : chr "" "" "" "" ...  
## $ Time.Since.First.Birth: chr "" "" "" "" ...  
## $ Definition : logi NA NA NA NA NA NA ...  
## $ Interval : num 5 5 5 5 5 5 5 5 5 5 ...

summary(clean\_data)

## REF\_AREA Geographic.area Regional.group Indicator   
## Length:6255 Length:6255 Length:6255 Length:6255   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## Sex Wealth.Quintile Series.Name Series.Year   
## Length:6255 Length:6255 Length:6255 Length:6255   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## Reference.Date Observation.Value Lower.Bound Upper.Bound   
## Min. :1952 Min. : 9.00 Min. : 11.07 Min. : 23.29   
## 1st Qu.:1988 1st Qu.: 67.42 1st Qu.: 45.14 1st Qu.: 70.19   
## Median :1998 Median :126.40 Median : 96.65 Median :145.45   
## Mean :1997 Mean :131.51 Mean :108.71 Mean :160.11   
## 3rd Qu.:2008 3rd Qu.:178.79 3rd Qu.:157.84 3rd Qu.:222.23   
## Max. :2024 Max. :773.17 Max. :634.89 Max. :929.16   
## NA's :2593 NA's :2593   
## Standard.Error Country.notes Observation.Status Unit.of.measure   
## Min. : 1.472 Length:6255 Length:6255 Length:6255   
## 1st Qu.: 7.700 Class :character Class :character Class :character   
## Median : 10.684 Mode :character Mode :character Mode :character   
## Mean : 11.934   
## 3rd Qu.: 14.300   
## Max. :180.790   
## NA's :4276   
## Series.Type Series.Category Series.Method Age.Group.of.Women  
## Length:6255 Length:6255 Length:6255 Length:6255   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## Time.Since.First.Birth Definition Interval   
## Length:6255 Mode:logical Min. :1.000   
## Class :character NA's:6255 1st Qu.:1.000   
## Mode :character Median :1.000   
## Mean :1.757   
## 3rd Qu.:1.000   
## Max. :5.000   
## NA's :799

#indicators  
unique(clean\_data$Indicator)

## [1] "Neonatal mortality rate" "Under-five mortality rate"

# date range  
range(clean\_data$Reference.Date, na.rm = TRUE)

## [1] 1951.5 2023.5

# Checking country coverage  
unique(clean\_data$Geographic.area)

## [1] "Burundi" "Democratic Republic of the Congo"  
## [3] "Kenya" "Rwanda"   
## [5] "Somalia" "South Sudan"   
## [7] "Tanzania" "Uganda"

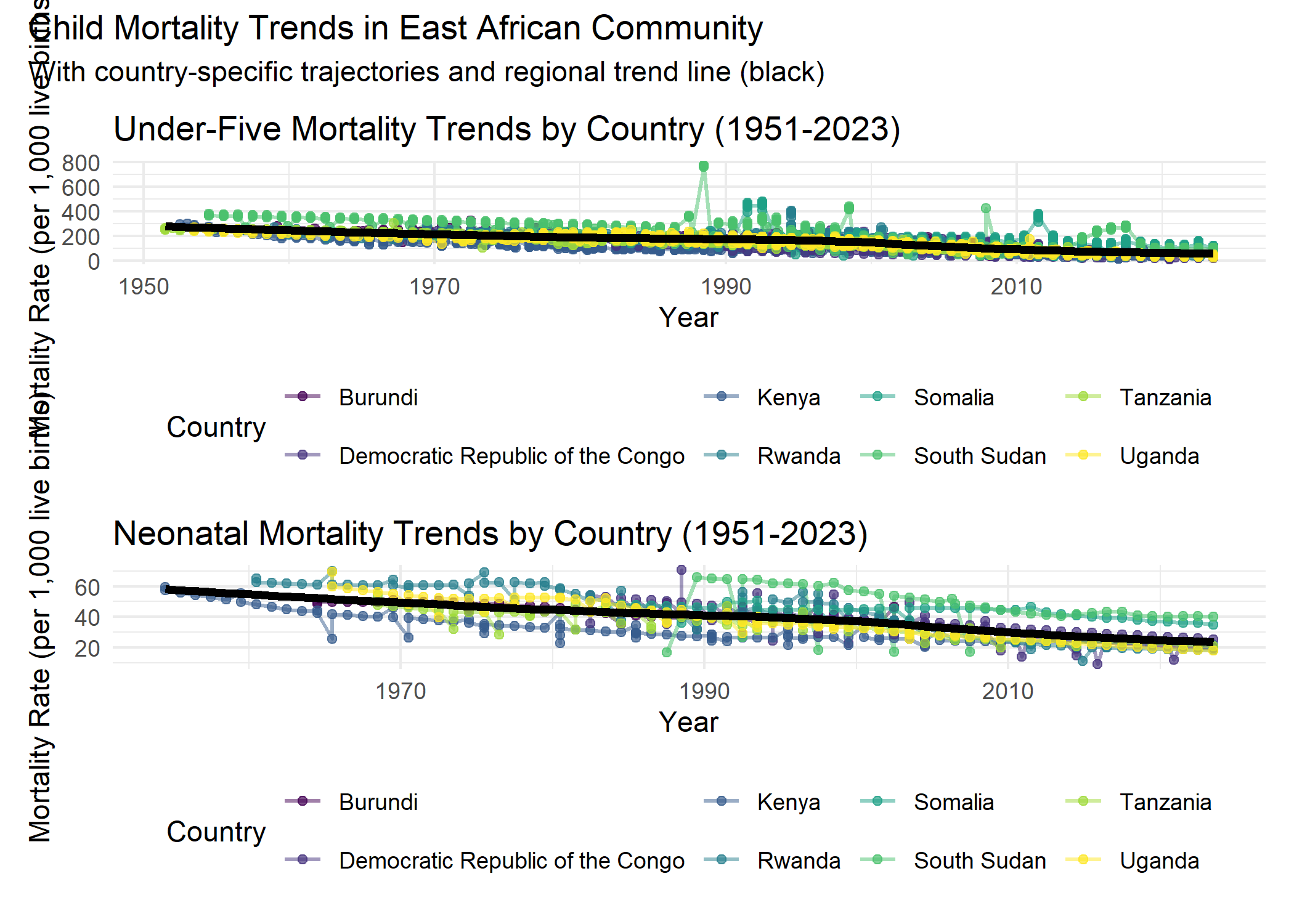
# Sample of mortality values  
clean\_data %>%  
 group\_by(Indicator) %>%  
 summarise(  
 min\_value = min(Observation.Value, na.rm = TRUE),  
 max\_value = max(Observation.Value, na.rm = TRUE),  
 avg\_value = mean(Observation.Value, na.rm = TRUE)  
 )

## # A tibble: 2 × 4  
## Indicator min\_value max\_value avg\_value  
## <chr> <dbl> <dbl> <dbl>  
## 1 Neonatal mortality rate 9 70.7 37.8  
## 2 Under-five mortality rate 9.87 773. 143.

# Trend Analysis   
  
# Calculating annual averages for each indicator  
trend\_data <- clean\_data %>%  
 group\_by(Reference.Date, Indicator) %>%  
 summarise(  
 avg\_mortality = mean(Observation.Value, na.rm = TRUE),  
 .groups = 'drop'  
 ) %>%  
 mutate(Reference.Date = as.numeric(Reference.Date))

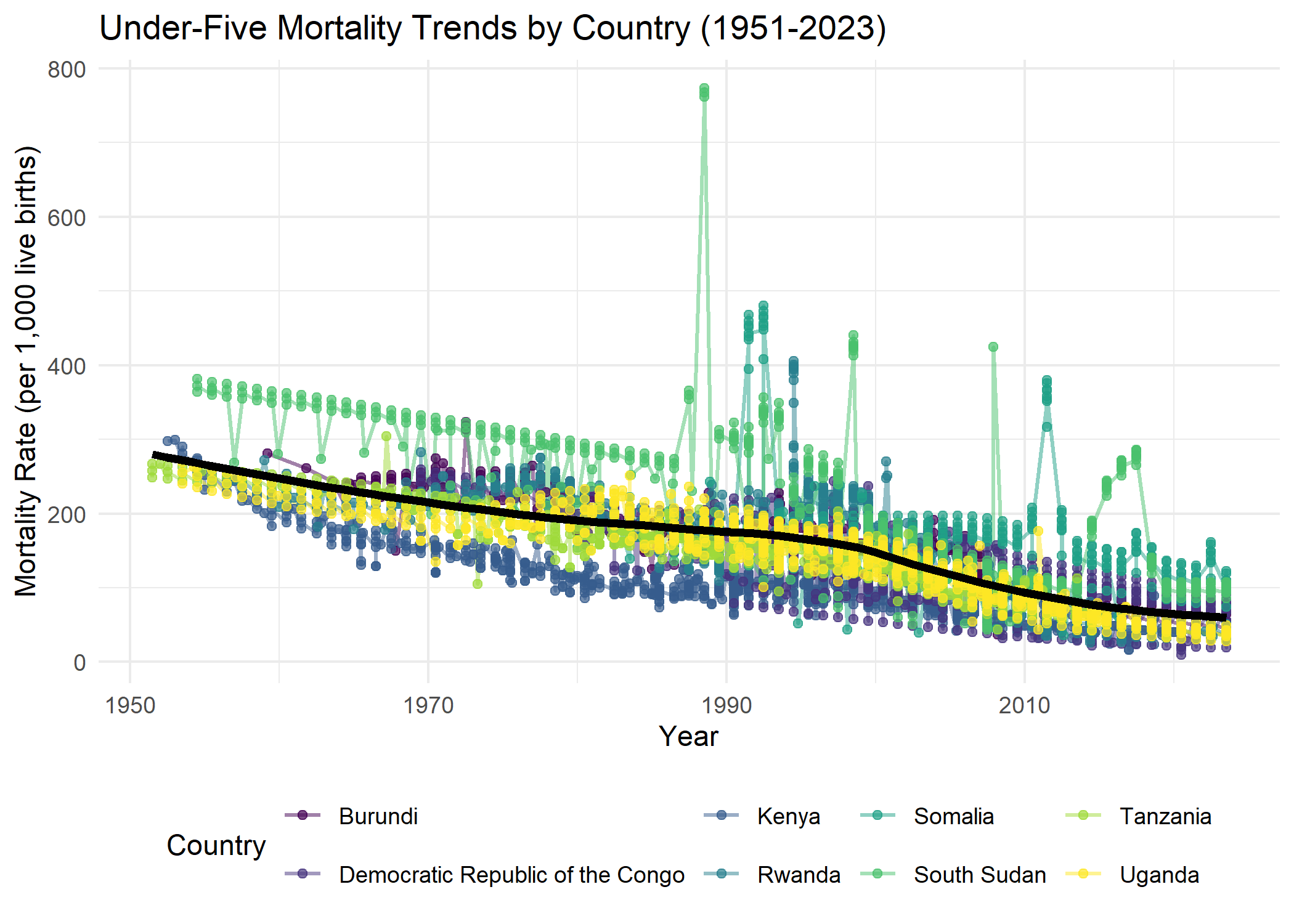
|  |
| --- |
| 5.TREND PLOTS |

clean\_data <- clean\_data %>%  
 mutate(Reference.Date = as.numeric(Reference.Date))   
  
# 1. Under-Five Mortality Trend Plot  
under5\_plot <- ggplot(clean\_data %>% filter(Indicator == "Under-five mortality rate"),   
 aes(x = Reference.Date, y = Observation.Value)) +  
 geom\_line(aes(group = Geographic.area, color = Geographic.area),   
 alpha = 0.5, linewidth = 0.7) +  
 geom\_point(aes(color = Geographic.area), alpha = 0.7, size = 1.5) +  
 geom\_smooth(aes(group = 1), method = "loess", color = "black",   
 se = FALSE, linewidth = 1.5) +  
 scale\_color\_viridis\_d() +  
 labs(title = "Under-Five Mortality Trends by Country (1951-2023)",  
 x = "Year",  
 y = "Mortality Rate (per 1,000 live births)",  
 color = "Country") +  
 theme\_minimal() +  
 theme(legend.position = "bottom") +  
 guides(color = guide\_legend(nrow = 2))  
  
# 2. Neonatal Mortality Trend Plot  
neonatal\_plot <- ggplot(clean\_data %>% filter(Indicator == "Neonatal mortality rate"),   
 aes(x = Reference.Date, y = Observation.Value)) +  
 geom\_line(aes(group = Geographic.area, color = Geographic.area),   
 alpha = 0.5, linewidth = 0.7) +  
 geom\_point(aes(color = Geographic.area), alpha = 0.7, size = 1.5) +  
 geom\_smooth(aes(group = 1), method = "loess", color = "black",   
 se = FALSE, linewidth = 1.5) +  
 scale\_color\_viridis\_d() +  
 labs(title = "Neonatal Mortality Trends by Country (1951-2023)",  
 x = "Year",  
 y = "Mortality Rate (per 1,000 live births)",  
 color = "Country") +  
 theme\_minimal() +  
 theme(legend.position = "bottom") +  
 guides(color = guide\_legend(nrow = 2))  
  
# Combine plots  
combined\_trends <- under5\_plot / neonatal\_plot +  
 plot\_annotation(title = "Child Mortality Trends in East African Community",  
 subtitle = "With country-specific trajectories and regional trend line (black)")  
  
# Display the plot  
print(combined\_trends)

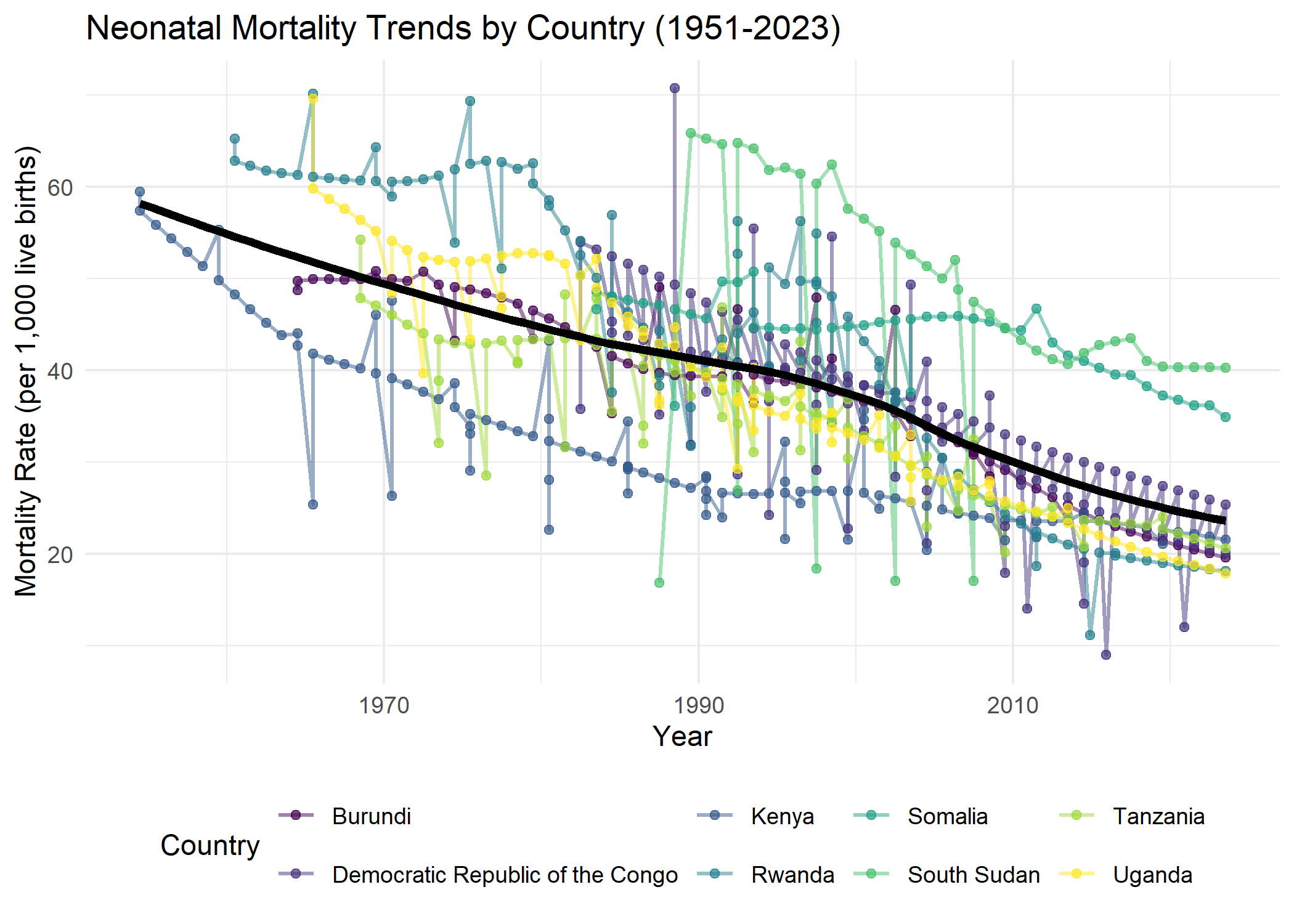


ggsave("trend\_map\_EAC.png", plot = combined\_trends, width = 10, height = 8, dpi = 300)

under5\_plot



neonatal\_plot



The two trend lines illustrates child mortality trends within the East African Community from 1951 to 2023. The top graph focuses on under-five mortality rates (per 1,000 live births), while the bottom graph depicts neonatal mortality rates (per 1,000 live births) for individual countries: Burundi, Democratic Republic of the Congo, Kenya, Rwanda, Somalia, South Sudan, Tanzania, and Uganda.

In both graphs, each country’s trajectory is represented by a distinct colored line with circular markers, allowing for the visualization of country-specific changes over time. A thick black line overlaid on the country-specific data represents the regional trend line, providing an overall view of mortality rate changes across the East African Community.

Both under-five and neonatal mortality rates show a general downward trend across the region over the period, indicating progress in child survival. However, the graphs also highlight variations in mortality rates and the pace of decline among the different countries within the community. Notably, there appear to be periods of fluctuation and divergence in the country-specific trajectories, suggesting diverse factors influencing child mortality in each nation. The regional trend line smooths out these individual variations to present a broader picture of improvement in child survival within the East African Community over the decades.

|  |
| --- |
| 6.HIGHEST MORTALITY COUNTRIES IN EAC |

# Identify Highest Mortality Countries   
highest\_mortality <- clean\_data %>%  
 filter(Reference.Date == max(Reference.Date)) %>%  
 group\_by(Indicator) %>%  
 slice\_max(Observation.Value, n = 1) %>%  
 select(Geographic.area, Indicator, Observation.Value)  
  
# Print results  
cat("\nCountries with Highest Mortality Rates:\n")

##   
## Countries with Highest Mortality Rates:

print(highest\_mortality)

## # A tibble: 2 × 3  
## # Groups: Indicator [2]  
## Geographic.area Indicator Observation.Value  
## <chr> <chr> <dbl>  
## 1 South Sudan Neonatal mortality rate 40.2  
## 2 Somalia Under-five mortality rate 123.

The data reveals critical disparities in child health outcomes across East Africa, with South Sudan reporting a neonatal mortality rate of 40.2 deaths per 1,000 live births—reflecting challenges in maternal and newborn care—while Somalia’s under-five mortality rate soars to 122.7 deaths per 1,000, underscoring the compounded vulnerabilities of conflict, poverty, and limited healthcare access. These figures highlight the urgent need for targeted interventions, particularly in fragile states where systemic barriers perpetuate high child mortality. The stark contrast between neonatal and under-five rates suggests that survival gaps widen significantly in early childhood, demanding prioritized investments in nutrition, immunization, and infectious disease control.