Assignment1-AdvRprogramming

Vishweshwar Chowdhury; 24205804

Introduction

This project analyzes key development indicators for **France, Kazakhstan, and Ukraine** using World Bank data. We explore trends across areas like health, education, economy, and environment using the data.table package in R.

1) Read Data using data.table and assign correct class to variables

```r
# Reading each country's data
fra <- fread("indicators_fra.csv")
kaz <- fread("indicators_kaz.csv")
ukr <- fread("indicators_ukr.csv")
options(datatable.print.topn = 3,datatable.print.nrows = 5,width = 80)
str(fra)
```

```
Classes 'data.table' and 'data.frame':	78971 obs. of  6 variables:
 $ Country Name  : chr  "#country+name" "France" "France" "France" ...
 $ Country ISO3  : chr  "#country+code" "FRA" "FRA" "FRA" ...
 $ Year          : chr  "#date+year" "2022" "2021" "2019" ...
 $ Indicator Name: chr  "#indicator+name" "Fertilizer consumption (% of fertilizer production)" "Fertilizer consumption (% of fertilizer production)" "Fertilizer
consumption (% of fertilizer production)" ...
 $ Indicator Code: chr  "#indicator+code" "AG.CON.FERT.PT.ZS" "AG.CON.FERT.PT.ZS" "AG.CON.FERT.PT.ZS" ...
 $ Value         : chr  "#indicator+value+num" "7180.73874470283" "1418.35276478348" "444.042634876329" ...
 - attr(*, ".internal.selfref")=<externalptr>
```

```r
head(fra)
```

```
    Country Name  Country ISO3       Year
          <char>        <char>     <char>
1: #country+name #country+code #date+year
2:        France           FRA       2022
3:        France           FRA       2021
4:        France           FRA       2019
5:        France           FRA       2018
6:        France           FRA       2017
                                        Indicator Name    Indicator Code
                                                <char>            <char>
1:                                     #indicator+name   #indicator+code
2: Fertilizer consumption (% of fertilizer production) AG.CON.FERT.PT.ZS
3: Fertilizer consumption (% of fertilizer production) AG.CON.FERT.PT.ZS
4: Fertilizer consumption (% of fertilizer production) AG.CON.FERT.PT.ZS
5: Fertilizer consumption (% of fertilizer production) AG.CON.FERT.PT.ZS
6: Fertilizer consumption (% of fertilizer production) AG.CON.FERT.PT.ZS
                  Value
                 <char>
1: #indicator+value+num
2:     7180.73874470283
3:     1418.35276478348
4:     444.042634876329
5:     423.053706007171
6:     560.596036485363
```

```r
options(datatable.print.topn = 3,datatable.print.nrows = 5,width = 80)
str(kaz)
```

```
Classes 'data.table' and 'data.frame':	71716 obs. of  6 variables:
 $ Country Name  : chr  "#country+name" "Kazakhstan" "Kazakhstan" "Kazakhstan" ...
 $ Country ISO3  : chr  "#country+code" "KAZ" "KAZ" "KAZ" ...
 $ Year          : chr  "#date+year" "2022" "2021" "2020" ...
 $ Indicator Name: chr  "#indicator+name" "Fertilizer consumption (% of fertilizer production)" "Fertilizer consumption (% of fertilizer production)" "Fertilizer
consumption (% of fertilizer production)" ...
 $ Indicator Code: chr  "#indicator+code" "AG.CON.FERT.PT.ZS" "AG.CON.FERT.PT.ZS" "AG.CON.FERT.PT.ZS" ...
 $ Value         : chr  "#indicator+value+num" "32.0859019740788" "36.8331633627589" "46.0065478697083" ...
 - attr(*, ".internal.selfref")=<externalptr>
```

```r
head(kaz)
```

```
    Country Name  Country ISO3       Year
          <char>        <char>     <char>
1: #country+name #country+code #date+year
2:    Kazakhstan           KAZ       2022
3:    Kazakhstan           KAZ       2021
4:    Kazakhstan           KAZ       2020
5:    Kazakhstan           KAZ       2019
6:    Kazakhstan           KAZ       2018
                                        Indicator Name    Indicator Code
                                                <char>            <char>
1:                                     #indicator+name   #indicator+code
2: Fertilizer consumption (% of fertilizer production) AG.CON.FERT.PT.ZS
3: Fertilizer consumption (% of fertilizer production) AG.CON.FERT.PT.ZS
4: Fertilizer consumption (% of fertilizer production) AG.CON.FERT.PT.ZS
5: Fertilizer consumption (% of fertilizer production) AG.CON.FERT.PT.ZS
6: Fertilizer consumption (% of fertilizer production) AG.CON.FERT.PT.ZS
                  Value
                 <char>
1: #indicator+value+num
2:     32.0859019740788
3:     36.8331633627589
4:     46.0065478697083
5:     24.1064217210881
6:     40.4971037436908
```

```r
options(datatable.print.topn = 3,datatable.print.nrows = 5,width = 80)
str(ukr)
```

```
Classes 'data.table' and 'data.frame':	71557 obs. of  6 variables:
 $ Country Name  : chr  "#country+name" "Ukraine" "Ukraine" "Ukraine" ...
 $ Country ISO3  : chr  "#country+code" "UKR" "UKR" "UKR" ...
 $ Year          : chr  "#date+year" "2021" "2020" "2019" ...
 $ Indicator Name: chr  "#indicator+name" "Fertilizer consumption (% of fertilizer production)" "Fertilizer consumption (% of fertilizer production)" "Fertilizer 
consumption (% of fertilizer production)" ...
 $ Indicator Code: chr  "#indicator+code" "AG.CON.FERT.PT.ZS" "AG.CON.FERT.PT.ZS" "AG.CON.FERT.PT.ZS" ...
 $ Value         : chr  "#indicator+value+num" "262.754436707114" "253.050691645338" "208.823539445296" ...
 - attr(*, ".internal.selfref")=<externalptr>
```

```r
head(ukr)
```

```
    Country Name  Country ISO3       Year
          <char>        <char>     <char>
1: #country+name #country+code #date+year
2:       Ukraine           UKR       2021
3:       Ukraine           UKR       2020
4:       Ukraine           UKR       2019
5:       Ukraine           UKR       2018
6:       Ukraine           UKR       2017
                                        Indicator Name    Indicator Code
                                                <char>            <char>
1:                                     #indicator+name   #indicator+code
2: Fertilizer consumption (% of fertilizer production) AG.CON.FERT.PT.ZS
3: Fertilizer consumption (% of fertilizer production) AG.CON.FERT.PT.ZS
4: Fertilizer consumption (% of fertilizer production) AG.CON.FERT.PT.ZS
5: Fertilizer consumption (% of fertilizer production) AG.CON.FERT.PT.ZS
6: Fertilizer consumption (% of fertilizer production) AG.CON.FERT.PT.ZS
                  Value
                 <char>
1: #indicator+value+num
2:     262.754436707114
3:     253.050691645338
4:     208.823539445296
5:      209.61966292436
6:     188.314558332498
```

In the above output we can observe the following things:-

- Each dataset had 6 columns: Country Name, Country ISO3, Year, Indicator Name, Indicator Code, Value.
- The first row was incorrectly read as data (e.g., #country+name), so it was removed during cleaning. All columns were initially read as character; Year and Value were later converted to integer and numeric respectively.

2) Merging Datasets using data.table

```r
# Add a country label to each dataset
fra[, Country := "France"]
kaz[, Country := "Kazakhstan"]
ukr[, Country := "Ukraine"]

# Combine the three datasets into one
all_data <- rbindlist(list(fra, kaz, ukr),
use.names = TRUE, fill = TRUE)

# Check structure and preview
str(all_data)
```

```
Classes 'data.table' and 'data.frame':	222244 obs. of  7 variables:
 $ Country Name  : chr  "#country+name" "France" "France" "France" ...
 $ Country ISO3  : chr  "#country+code" "FRA" "FRA" "FRA" ...
 $ Year          : chr  "#date+year" "2022" "2021" "2019" ...
 $ Indicator Name: chr  "#indicator+name" "Fertilizer consumption (% of fertilizer production)" "Fertilizer consumption (% of fertilizer production)" "Fertilizer consumption (% of fertilizer production)" ...
 $ Indicator Code: chr  "#indicator+code" "AG.CON.FERT.PT.ZS" "AG.CON.FERT.PT.ZS" "AG.CON.FERT.PT.ZS" ...
 $ Value         : chr  "#indicator+value+num" "7180.73874470283" "1418.35276478348" "444.042634876329" ...
 $ Country       : chr  "France" "France" "France" "France" ...
 - attr(*, ".internal.selfref")=<externalptr>
```

```r
head(all_data)
```

```
    Country Name  Country ISO3       Year
          <char>        <char>     <char>
1: #country+name #country+code #date+year
2:        France           FRA       2022
3:        France           FRA       2021
4:        France           FRA       2019
5:        France           FRA       2018
6:        France           FRA       2017
                                        Indicator Name    Indicator Code
                                                <char>            <char>
1:                                     #indicator+name   #indicator+code
2: Fertilizer consumption (% of fertilizer production) AG.CON.FERT.PT.ZS
3: Fertilizer consumption (% of fertilizer production) AG.CON.FERT.PT.ZS
4: Fertilizer consumption (% of fertilizer production) AG.CON.FERT.PT.ZS
5: Fertilizer consumption (% of fertilizer production) AG.CON.FERT.PT.ZS
6: Fertilizer consumption (% of fertilizer production) AG.CON.FERT.PT.ZS
                  Value Country
                 <char>  <char>
1: #indicator+value+num  France
2:     7180.73874470283  France
3:     1418.35276478348  France
4:     444.042634876329  France
```

The merged dataset contains the following things:-

- The merged dataset contains **222,244 rows** and 7 variables.
- Column names and structure are consistent across all countries.
- The resulting dataset allows for cross-country comparisons across years and indicators.

3) Data Exploration

```r
# Basic summary
summary(all_data)
```

```
 Country Name       Country ISO3           Year           Indicator Name    
 Length:222244      Length:222244      Length:222244      Length:222244     
 Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Indicator Code        Value             Country         
 Length:222244      Length:222244      Length:222244     
 Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character  
```

```r
# Number of unique indicators
length(unique(all_data$`Indicator Name`))
```

```
[1] 3912
```

```r
# Time coverage
range(all_data$Year, na.rm = TRUE)
```

```
[1] "#date+year" "2024"
```

```r
# Count of rows per country
all_data[, .N, by = Country]
```

```
      Country     N
       <char> <int>
1:     France 78971
2: Kazakhstan 71716
3:    Ukraine 71557
```

```r
# Top 5 most common indicators
all_data[, .N, by = `Indicator Name`][order(-N)][1:5]
```

```
                                              Indicator Name     N
                                                      <char> <int>
1:                                             Net migration   585
2: Adolescent fertility rate (births per 1,000 women ages 15-19)   576
3:                      Life expectancy at birth, female (years)   576
4:                        Life expectancy at birth, male (years)   576
5:                 Mortality rate, under-5 (per 1,000 live births)   510
```

```r
# Missing values check
colSums(is.na(all_data))
```

```
  Country Name   Country ISO3           Year Indicator Name Indicator Code 
             0              0              0              0              0 
         Value        Country 
             0              0 
```

```r
# Number of observations per year per country
all_data[, .N, by = .(Country, Year)][order(Country, Year)]
```

```
     Country       Year     N
      <char>     <char> <int>
  1:  France #date+year     1
  2:  France       1960   634
  3:  France       1961   396
 ---                         
196: Ukraine       2022  1046
197: Ukraine       2023   845
198: Ukraine       2024    58
```

The findings of the data analysis task are as follows:-

- The merged dataset has **222,244 rows with 7 character columns**, indicating raw data still needed cleaning.
- Indicators are diverse, with the most common including:
  - **Net migration**
  - **Life expectancy**
  - **Adolescent fertility rate**
  - **Under 5-mortality rate**
- The dataset spans a wide time range — **from 1960 to 2023** - with varying data density per year.
- France had data available for nearly all years between **1960 and 2023**, with some variation in the number of observations per year.

4) Data Analysis task using keyby argument

```r
# Viewing the most common indicators
all_data[, .N, by = `Indicator Name`][order(-N)][1:20]
```

```
                                                  Indicator Name     N
                                                          <char> <int>
 1:                                                Net migration   585
 2: Adolescent fertility rate (births per 1,000 women ages 15-19)   576
 3:                     Life expectancy at birth, female (years)   576
---                                                                   
18:                                             Urban population   384
19:                    Urban population (% of total population)   384
20:       Mortality rate, adult, female (per 1,000 female adults)   382
```

```r
# Filter for selected indicators
focus_indicators <- c("Mortality rate, under-5 (per 1,000 live births)",
                      "Net migration",
                      "Population ages 15-64 (% of total population)")

selected_data <- all_data[`Indicator Name` %in% focus_indicators]
selected_data[, Value := as.numeric(Value)]
# Mean values per indicator per country over time
summary_data <- selected_data[, .(Average = mean(Value, na.rm = TRUE)),
                              keyby = .(Country, `Indicator Name`, Year)]

head(summary_data)
```

```
Key: <Country, Indicator Name, Year>
   Country                                  Indicator Name   Year Average
    <char>                                          <char> <char>   <num>
1:  France Mortality rate, under-5 (per 1,000 live births)   1960    28.5
2:  France Mortality rate, under-5 (per 1,000 live births)   1961    27.0
3:  France Mortality rate, under-5 (per 1,000 live births)   1962    25.7
4:  France Mortality rate, under-5 (per 1,000 live births)   1963    24.5
5:  France Mortality rate, under-5 (per 1,000 live births)   1964    23.4
6:  France Mortality rate, under-5 (per 1,000 live births)   1965    22.4
```

The following observations were made from above output:-

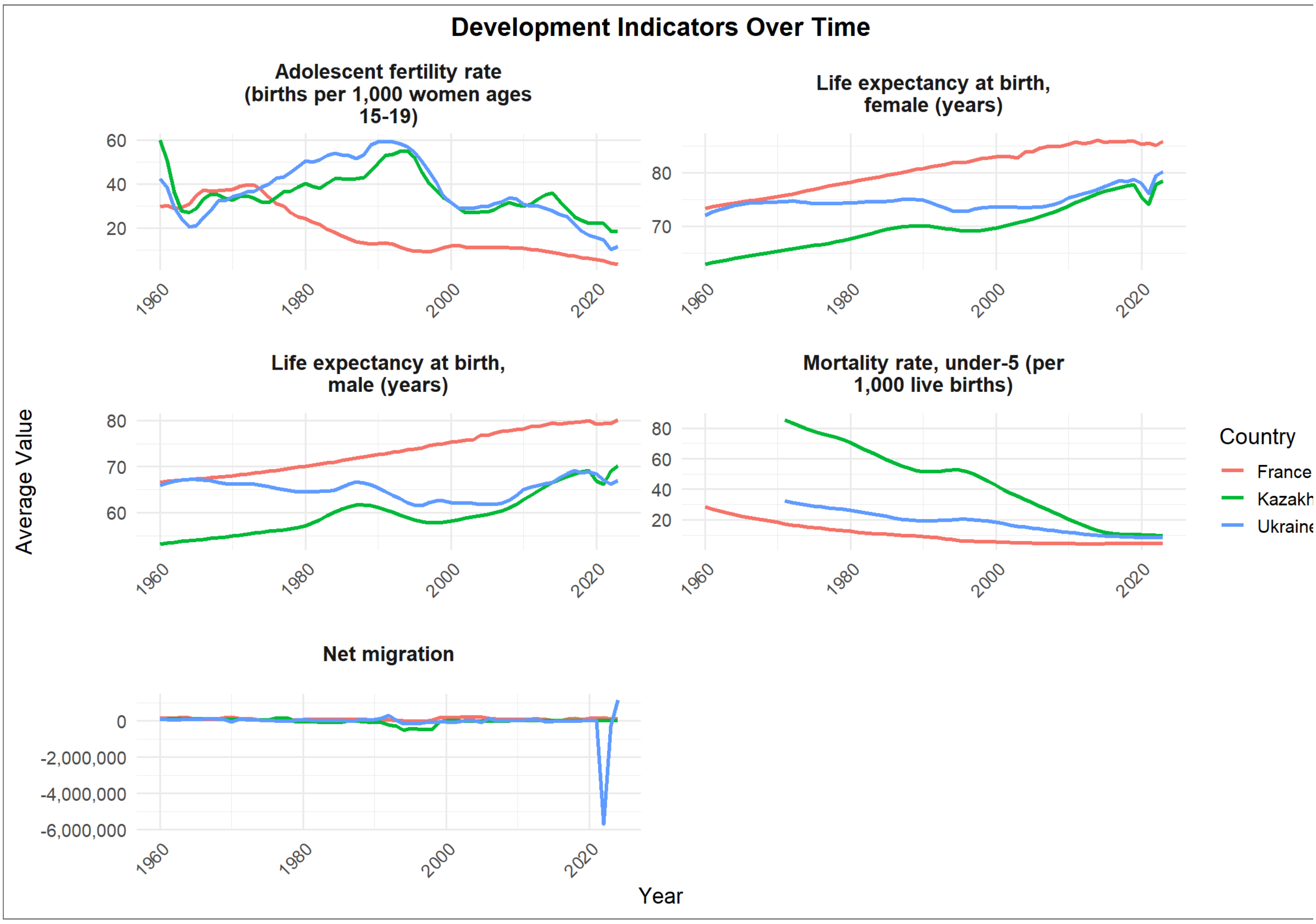
1. Using data.table 5 relevant development indicators were found which are net migration,Adolescent fertility rate , Life expectancy at birth and mortality rate under 5.
2. The data was grouped and summarized by Country, Indicator Name, and Year using **keyby** to calculate yearly averages.

5) Creation of Development Indicators over time plot and Under 5 Mortality Rate over time plot

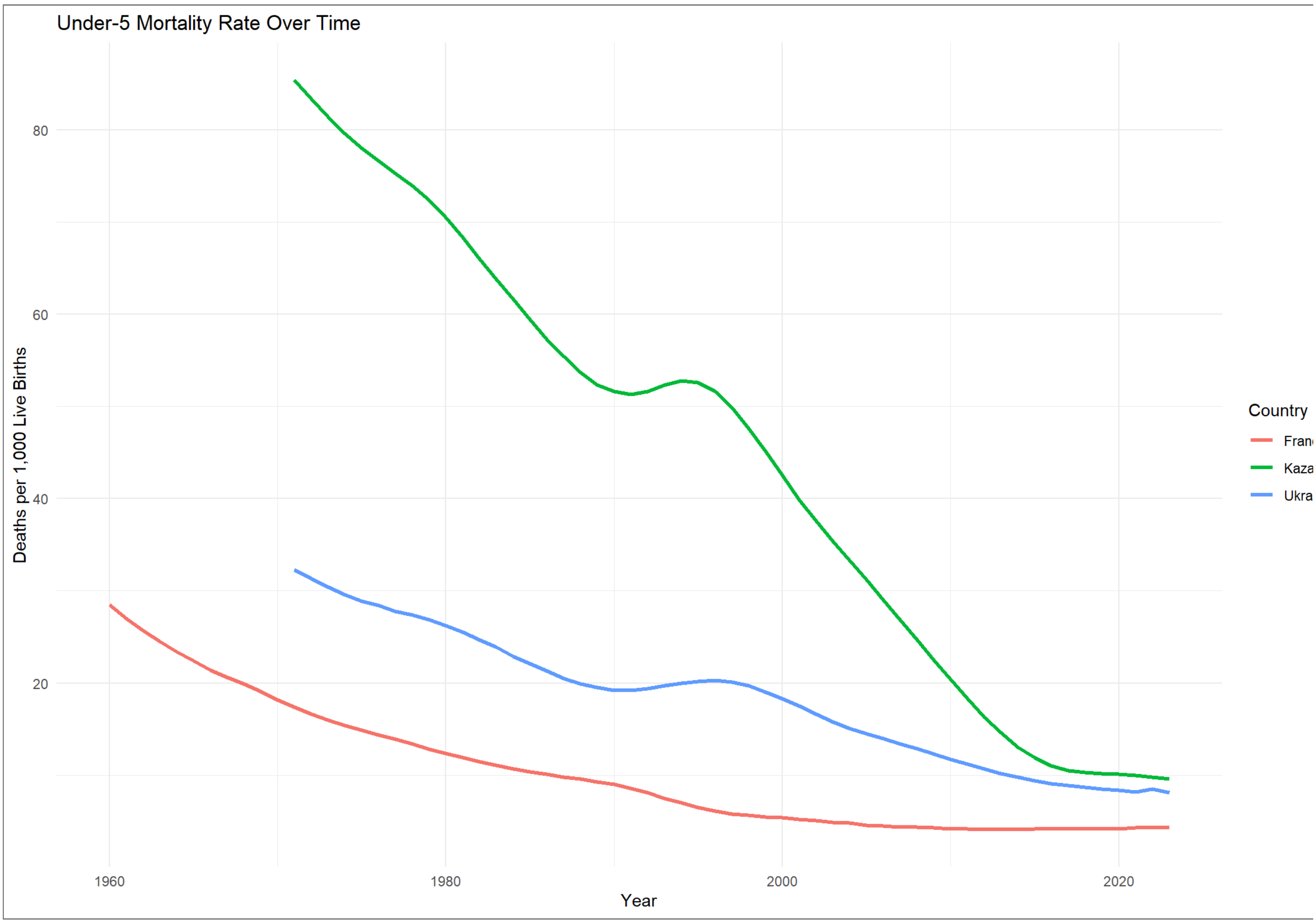
Code

```r
# Clean Year column
all_data <- all_data[!is.na(Year)]
all_data <- all_data[Year != "#date+year"]

# Convert Year to numeric
all_data[, Year := as.integer(Year)]

# Filter indicators
indicators_to_plot <- c(
  "Net migration",
  "Adolescent fertility rate (births per 1,000 women ages 15-19)",
  "Life expectancy at birth, female (years)",
  "Life expectancy at birth, male (years)",
  "Mortality rate, under-5 (per 1,000 live births)"
)

plot_data <- all_data[`Indicator Name` %in% indicators_to_plot]
plot_data[, Value := as.numeric(Value)]

# Recalculate summaries
summary_data <- plot_data[
  , .(Average = mean(Value, na.rm = TRUE)),
  keyby = .(Country, `Indicator Name`, Year)
]
```

Creation of Development Indicators over time- Plot



Under 5 mortality rate over time- Plot



Plot interpretations

1) Development Indicators over time

- **Life expectancy (male and female)** shows consistent growth across all countries, with **France** leading, followed by Ukraine and Kazakhstan.
- **Under-5 mortality rates** have significantly declined in all three countries — reflecting healthcare improvements — **though Kazakhstan started from a much higher rate**.
- **Adolescent fertility rates** have steadily **declined**, **especially in France**, indicating improved reproductive health awareness.
- **Net migration** shows **extreme reduction in Ukraine,** likely reflecting the recent **war** between Ukraine and palestine.

2) Under-5 Mortality Rate Over Time

- **Kazakhstan** had the **highest child mortality in the 1960s**, but saw major declines over time.
- **France** maintained the **lowest under-5 mortality rates throughout the timeline,** reflecting a consistently **strong public health system**.
- **Ukraine** showed moderate improvement, with a **steady decline from the 1970s to present.**
- Overall, the plot demonstrates strong downward trends for all three countries, **emphasizing global progress in reducing child mortality.**

Speaker notes