Advanced Data Programming with R - Assignment 1

Vanessa Carvalho - 25267189

June 11, 2025

Introduction

For this first Assignment, I have selected Brazil, Angola and the United Kingdom to explore and analyse their *Economic, Social, Environmental, Health, Education, Development and Energy* data.

As countries with quite a different development throughout the decades, I have selected the following specific indicators to be used in **comparisons** between each of them:

1. Life expectancy at birth (total) years
2. Net migration
3. Unemployment, total (% of total labor force) (modeled ILO estimate)

Description of datasets and indicators

Life expectancy

Indicates the number of **years** a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.

Net migration

The net **total of migrants** during the period, that is, the number of immigrants minus the number of emigrants, including both citizens and non citizens.

Unemployment

The share of the labor force that is **without work** but available for and seeking employment.

Data Import and Overview

Here we import the datasets on Education, Health, and Energy using `fread()` from the `data.table` package.

```
library(data.table)

#read each dataset
bra <- fread("indicators_bra.csv")
ago <- fread("indicators_ago.csv")
gbr <- fread("indicators_gbr.csv")
```

Resumed Version of the Dataset (Table Preview)

Each table includes the first 6 rows of the datasets

- Brazil demographics

```
library(knitr)

# this code is for printing only the first 5 rows, minus the second one
bra_preview <- head(bra, 6)
bra_preview <- bra_preview[-2]
bra_preview[, `Indicator Code` := NULL]
kable(bra_preview)
```

Country Name	Country ISO3	Year	Indicator Name	Value
Brazil	BRA	2022	Fertilizer consumption (% of fertilizer production)	663.0297
Brazil	BRA	2020	Fertilizer consumption (% of fertilizer production)	790.8127
Brazil	BRA	2019	Fertilizer consumption (% of fertilizer production)	803.6594
Brazil	BRA	2018	Fertilizer consumption (% of fertilizer production)	590.8927

Country Name	Country ISO3	Year	Indicator Name	Value
Brazil	BRA	2017	Fertilizer consumption (% of fertilizer production)	635.8644

- **Angola demographics**

```
library(knitr)

# this code is for printing only the first 5 rows, minus the second one
ago_preview <- head(ago, 6)
ago_preview <- ago_preview[-2]
ago_preview[, `Indicator Code` := NULL]
kable(ago_preview)
```

Country Name	Country ISO3	Year	Indicator Name	Value
Angola	AGO	2022	Fertilizer consumption (kilograms per hectare of arable land)	13.258306
Angola	AGO	2020	Fertilizer consumption (kilograms per hectare of arable land)	7.849204
Angola	AGO	2019	Fertilizer consumption (kilograms per hectare of arable land)	6.652424
Angola	AGO	2018	Fertilizer consumption (kilograms per hectare of arable land)	7.103552
Angola	AGO	2017	Fertilizer consumption (kilograms per hectare of arable land)	9.773158

- **United Kingdom demographics**

```r
library(knitr)

# this code is for printing only the first 5 rows, minus the second one
gbr_preview <- head(gbr, 6)
gbr_preview <- gbr_preview[-2]
gbr_preview[, `Indicator Code` := NULL]
kable(gbr_preview)
```

Country Name	Country ISO3	Year	Indicator Name	Value
United Kingdom	GBR	2022	Fertilizer consumption (% of fertilizer production)	215.6584
United Kingdom	GBR	2020	Fertilizer consumption (% of fertilizer production)	341.8342
United Kingdom	GBR	2019	Fertilizer consumption (% of fertilizer production)	439.1753
United Kingdom	GBR	2018	Fertilizer consumption (% of fertilizer production)	436.8189
United Kingdom	GBR	2017	Fertilizer consumption (% of fertilizer production)	412.5512

Data Cleaning

In this step each dataset will be cleaned by *keeping only my target indicators* and **removing the duplicates**

```r
# Defining indicators to keep
indicators_keep <-
  c("Life expectancy at birth, total (years)",
    "Net migration", "Unemployment, total (% of total labor force) (modeled ILO estimate)")

# Filter the datasets
bra_clean <- bra[`Indicator Name`%in% indicators_keep]
ago_clean <- ago[`Indicator Name`%in% indicators_keep]
gbr_clean <- gbr[`Indicator Name`%in% indicators_keep]

# Remove the duplicates
bra_clean <- unique(bra_clean)
ago_clean <- unique(ago_clean)
gbr_clean <- unique(gbr_clean)
```

Merging Datasets

The following process is to merge both datasets for an easier look and comparative analysis of the indicators

```
all_data <- rbindlist(list(bra_clean, ago_clean, gbr_clean),
                      use.names = TRUE, fill = TRUE)
```

Table preview of merged data with basic summary statistics

```
# Average of each indicator value for each country
summary_stats <- all_data[
  , .(Average = mean(Value, na.rm = TRUE)),
  keyby = .(`Country Name`, `Indicator Name`)
]

# Renaming the indicators for a cleaner look
summary_stats[,`Indicator Name`:= fcase(
  `Indicator Name` == "Life expectancy at birth, total (years)",
  "Life Expectancy",  `Indicator Name` == "Net migration", "Net migration",
  `Indicator Name` ==
    "Unemployment, total (% of total labor force) (modeled ILO estimate)", "Unemployment"
)]

kable(summary_stats)
```

Country Name	Indicator Name	Average
Angola	Life Expectancy	47.304828
Angola	Net migration	43500.061538
Angola	Unemployment	16.378235
Brazil	Life Expectancy	65.917937
Brazil	Net migration	-78913.800000
Brazil	Unemployment	9.368853
United Kingdom	Life Expectancy	76.203563
United Kingdom	Net migration	116425.892308
United Kingdom	Unemployment	6.183088

Key Analysis and Plots

Two analysis were done to assess the following:

1. How each indicator has evolved year by year within and across countries
2. A comparison in between countries for each indicator in the last year **(2024)**

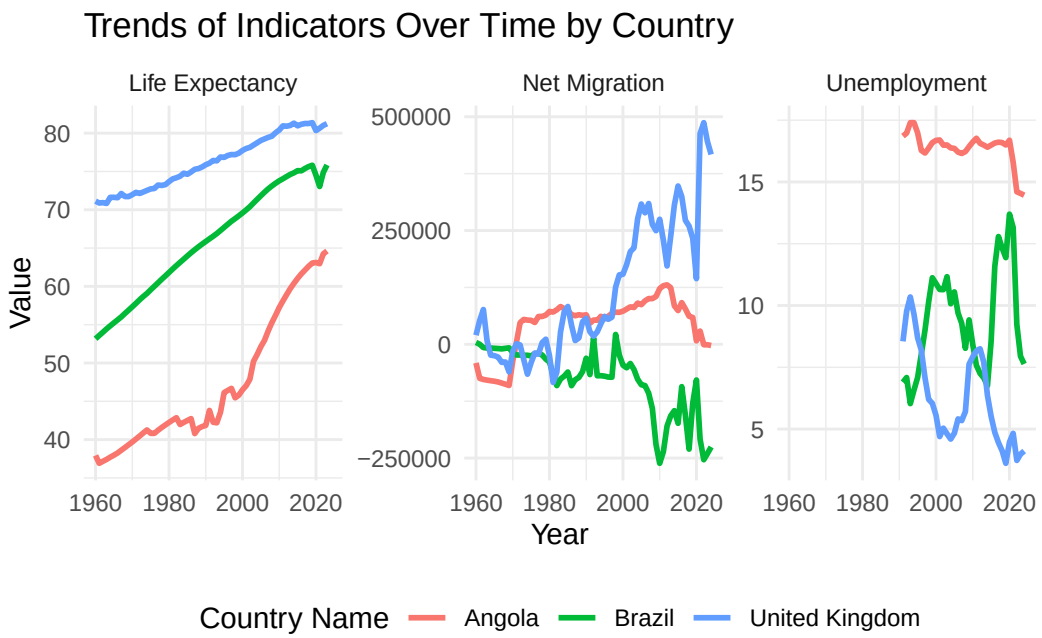
Analysis 1 - Trend over Time

```
# Renaming indicators for better look
all_data[,`Indicator Name`:= fcase(
  `Indicator Name` == "Life expectancy at birth, total (years)",
  "Life Expectancy",  `Indicator Name` == "Net migration", "Net Migration",
  `Indicator Name` ==
    "Unemployment, total (% of total labor force) (modeled ILO estimate)", "Unemployment"
)]

# Creating the trend analysis
trend <- all_data[
  `Indicator Name` %in% c("Life Expectancy", "Net Migration", "Unemployment"),
  .(Average = mean(Value, na.rm = TRUE)),
  by = .(`Year`, `Country Name`, `Indicator Name`)]

library(ggplot2)

ggplot(trend, aes(x = Year, y = Average, color = `Country Name`)) +
  geom_line(size = 1) +
  facet_wrap(~ `Indicator Name`, scales = "free_y") +
  labs(title = "Trends of Indicators Over Time by Country",
       y = "Value",
       x = "Year") +
  theme_minimal() +
  theme(legend.position = "bottom")
```

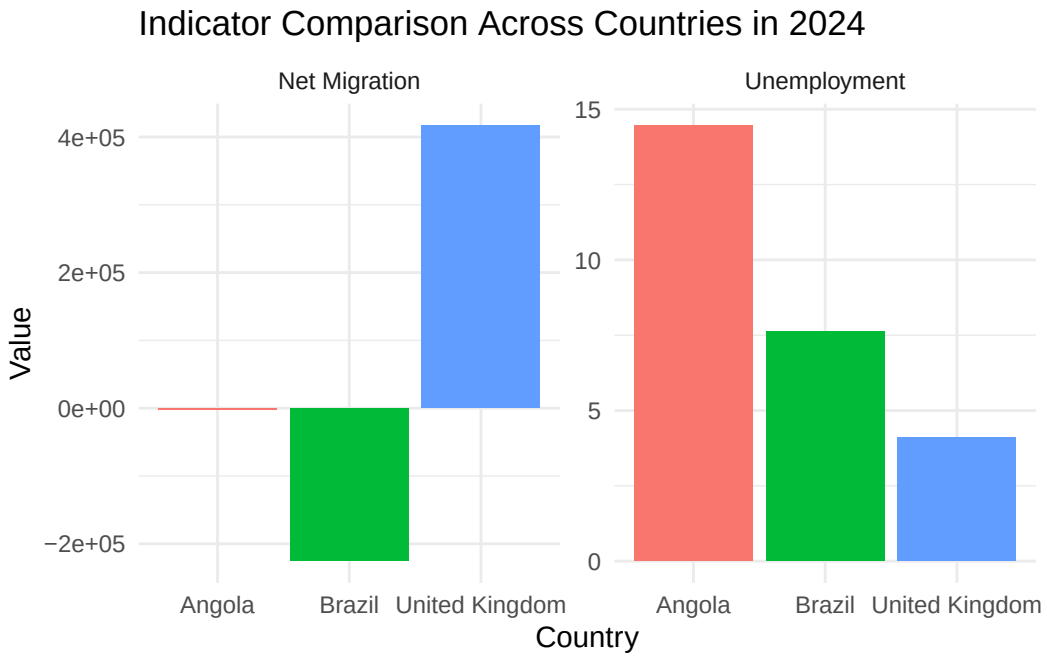


Analysis 2 - Cross-Country Comparison

For this analysis, only `Net Migration` and `Unemployment` are considered due to absence of data for 2024 on `Life Expectancy`

```
latest_year <- max(all_data$Year, na.rm = TRUE)
comparison <- all_data[Year == latest_year,
  .(Value = mean(Value, na.rm = TRUE)),
  by = .(`Country Name`,`Indicator Name`)
]

ggplot(comparison, aes(x = `Country Name`, y = Value, fill = `Country Name`)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ `Indicator Name`, scales = "free_y") +
  labs(title = paste("Indicator Comparison Across Countries in", latest_year),
       y = "Value",
       x = "Country") +
  theme_minimal()
```



Insights and Findings

Trend over Time

This trend analysis have shown:

- Across 6 decades, `Life Expectancy` considerably increased for all 3 countries. Whilst Brazil and the United Kingdom had a consistent exponential gain throughout, Angola had its major boost in the past 20 years
- Regarding the `Net Migration`, each country had their own different development, with very unstable peaks throughout the years. The United Kingdom presented the highest increase overall, whilst Brazil kept negative numbers for the vast majority of the years. Angola on the other hand had its numbers quite stable, until it started to decrease around 2010, reaching a 0 mark in most recent years
- `Unemployment` data was only captured from the year 1991, showing Angola with the highest incidence of unemployed people. The United Kingdom and Brazil numbers varied, with many high and low peaks in different years.

Cross-Country Comparison

Giving a closer look into the indicators in the past year, the database for these 3 countries have demonstrated:

- The United Kingdom have had the *highest* `Net Migration`, but the *lowest* `Unemployment` rate
- Brazil was the country with *lowest* `Net Migration`, however with an *average* `Unemployment` rate
- Angola barely presented `Net Migration` in 2024, with its rate *close to 0*, compensating in its `Unemployment` rate which was the *highest* among all countries.

Conclusion

- Compiling both analysis made for this assignment, an overall difference in the demographics of all 3 countries is noted. The results exemplify the development of countries like the United Kingdom, the slow progress of Brazil and the under development of an African nation like Angola
- A special mention to the reflection of the **Covid-19 pandemic break in 2020**, where `Net Migration` differed in the United Kingdom (with a drastic decrease) and a big peak in Brazil. Also, with a massive peak in Unemployment in all 3 countries.