# STAT40830 - Assignment 1

Suriya Raman S S

# Slide 1 – Loading the Datasets

```
fra = fread("indicators_fra.csv")
ind = fread("indicators_ind.csv")
irl = fread("indicators_irl.csv")
dataList = list(France = fra,
                India = ind, Ireland = irl)
```

Loading the datasets using fread()

# Slide 2 – Merge Datasets

```r
allData = rbindlist(dataList, idcol = "Country")
```

Merges all 3 datasets into one combined `data.table` named `allData`.

# Slide 3 – Assign Classes

```
allData[, Year := as.integer(Year)]
allData[, Value := as.numeric(Value)]
```

Converts Year and Value to numeric types for safe analysis.

# Slide 4 – Remove Missing Values

```
allData = allData[!is.na(Year) & !is.na(Value)]
```

Removed rows where `Year` or `Value` were missing — only 3 rows dropped.

# Slide 5 – Quick Data Exploration

```
cat("Rows:", nrow(allData), "Cols:", ncol(allData))
```

```
Rows: 236638 Cols: 7
```

```
colSums(is.na(allData))
```

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

```
str(allData)
```

```
Classes 'data.table' and 'data.frame':  236638 obs. of  7 variables:
 $ Country       : chr  "France" "France" "France" "France" ...
 $ Country Name  : chr  "France" "France" "France" "France" ...
 $ Country ISO3  : chr  "FRA" "FRA" "FRA" "FRA" ...
 $ Year          : int  2022 2021 2019 2018 2017 2016 2015 2014 2013 2012 ...
 $ Indicator Name: chr  "Fertilizer consumption (% of fertilizer production)" "
 $ Indicator Code: chr  "AG.CON.FERT.PT.ZS" "AG.CON.FERT.PT.ZS" "AG.CON.FERT.PT
 $ Value         : num  7181 1418 444 423 561 ...
 - attr(*, ".internal.selfref")=<externalptr>
```

# Slide 5 - Cont.

A basic check of the dataset shows it's well structured:

▶ 236,000+ rows across 7 columns covering 3 countries
▶ Only 3 missing values in Year and Value
▶ Year variable ranges from 1960 to 2024
▶ All variable types appear correctly assigned (e.g., character vs numeric)

# Slide 6 – Explore Value Distribution

```
allData[, .N, by = `Indicator Name`][order(-N)][1:10] |>
  ggplot(aes(x = reorder(`Indicator Name`, N), y = N)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(title = "Top 10 Most Common Indicators",
       x = "Indicator Name", y = "Count") +
  theme_minimal()
```
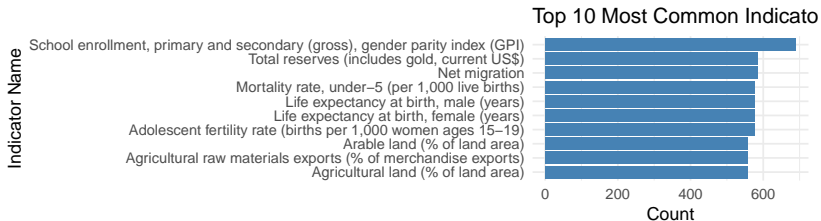


These are the most commonly reported indicators across the dataset.

# Slide 7 – Filter: GDP per Capita

```
gdp = allData[`Indicator Name` == "GDP per capita (current US$)"]
ggplot(gdp, aes(Year, Value, color = Country)) +
  geom_line(size = 0.7) +
  labs(title = "GDP per capita (in US Dollars $)", y = "USD")
```
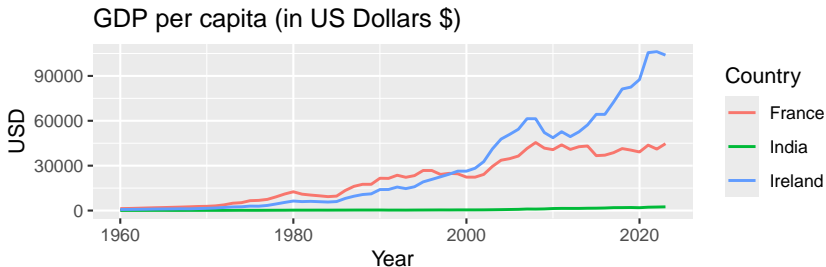
We extract and plot GDP per capita to compare how economic performance differs over time between the three countries.

Ireland's GDP per capita surged after 2000 and stayed highest. France grew slowly but steadily. India's growth is visible but still much lower.

# Slide 8 – Filter and Plot Population Growth

```
pop = allData[`Indicator Name` == "Population growth (annual %)"]

ggplot(pop, aes(Year, Value, color = Country)) +
  geom_line(size = 0.7) +
  labs(title = "Population Growth Rate", y = "% growth")
```
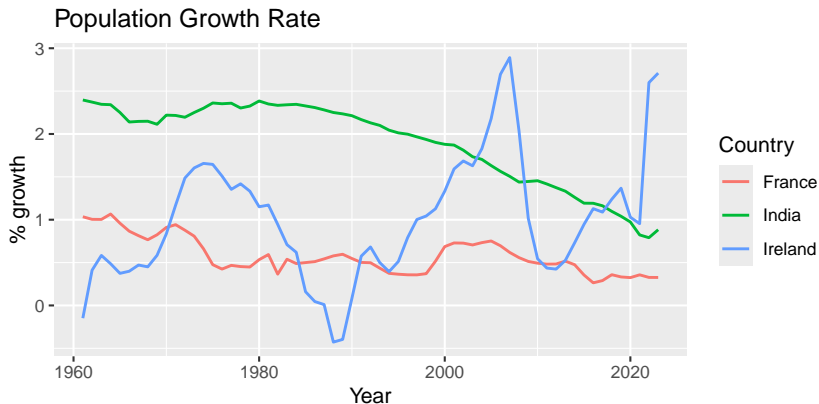
# Slide 8 - 2 – Filter and Plot Population Growth cont.

We filter and plot population growth over time.

India's rate is high but declining

France stays steady

Ireland fluctuates a lot in recent years

# Slide 9 – Filter and Plot: Education Spend

```
tmptxt = "Government expenditure on education, total (% of GDP)"
edu = allData[`Indicator Name` == tmptxt]

ggplot(edu, aes(Year, Value, color = Country)) +
  geom_line(size = 0.7) +
  labs(title = "Education Spend (% GDP)", y = "% of GDP")
```
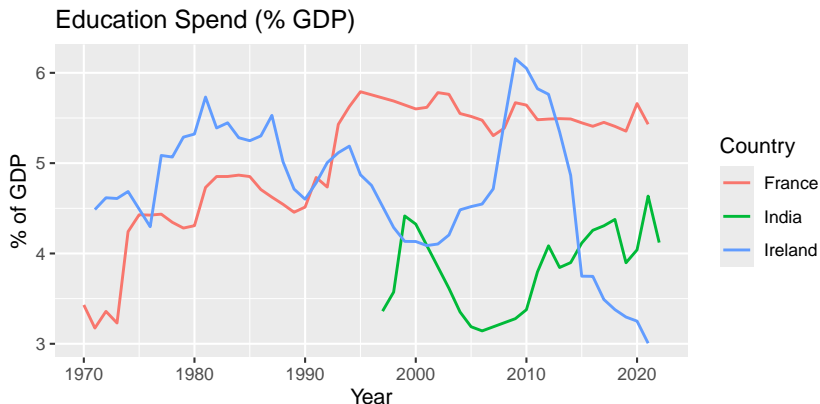


Education Spend (% GDP)

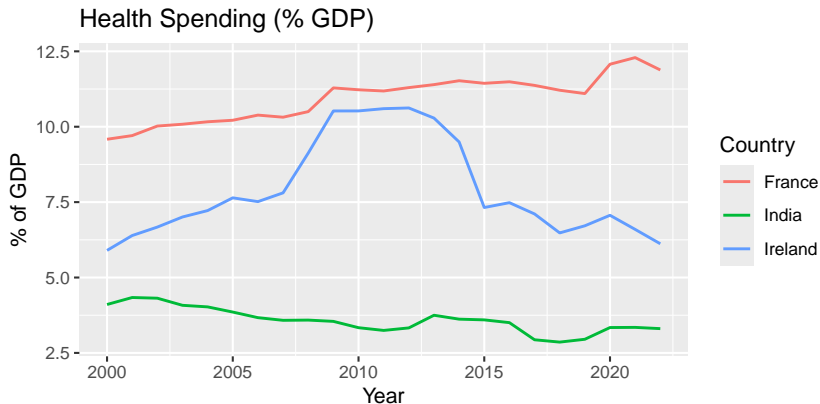We filter and plot education expenditure (% of GDP) over time.

France remains consistently high

Ireland shows fluctuation with spikes and recently its going down

India's spend is lower but gradually improving

# Slide 10 – Filter and Plot: Health Expenditure

```
tmptxt = "Current health expenditure (% of GDP)"
health = allData[`Indicator Name` == tmptxt]
ggplot(health, aes(Year, Value, color = Country)) +
  geom_line(size = 0.7) +
  labs(title = "Health Spending (% GDP)", y = "% of GDP")
```

We filter and plot Health expenditure (% of GDP) over time.

France leads with consistently high spend

Ireland rose until 2010 but declined after

India remains low and flat

# Slide 11 – Grouped Summary with keyby

```
allData[`Indicator Name` %in% c("GDP per capita (current US$)",
  "Population growth (annual %)",
  "Government expenditure on education, total (% of GDP)",
  "Current health expenditure (% of GDP)",
  "Carbon dioxide (CO2) emissions from
  Power Industry (Energy) (Mt CO2e)") ,
  .(avgValue = mean(Value, na.rm = TRUE)),
  keyby = .( `Indicator Name`, Country)]
```

```
Key: <Indicator Name, Country>
                                        Indicator Name Country     avgValue
                                                <char>  <char>        <num>
 1:          Current health expenditure (% of GDP)  France 1.094567e+01
 2:          Current health expenditure (% of GDP)   India 3.575042e+00
 3:          Current health expenditure (% of GDP) Ireland 7.922542e+00
 4:                    GDP per capita (current US$)  France 2.114789e+04
 5:                    GDP per capita (current US$)   India 6.560860e+02
 6:                    GDP per capita (current US$) Ireland 2.793495e+04
 7: Government expenditure on education, total (% of GDP)  France 4.998829e+00
 8: Government expenditure on education, total (% of GDP)   India 3.863645e+00
 9: Government expenditure on education, total (% of GDP) Ireland 4.735283e+00
10:                   Population growth (annual %)  France 5.790966e-01
11:                   Population growth (annual %)   India 1.894344e+00
12:                   Population growth (annual %) Ireland 9.989807e-01
```

# Slide 11 - 2 – Grouped Summary with keyby cont.

We group the data by Country and Indicator Name, then compute the average value using keyby. This helps us compare long-term trends between countries for select indicators.

# Slide 12 - Energy Expenditure

```
tmptxt = "Carbon dioxide (CO2) emissions from Power Industry (Energy) (Mt CO2e)
co2 = allData[`Indicator Name` == tmptxt]
ggplot(co2, aes(Year, Value, color = Country)) +
  geom_line(size = 0.7) +
  labs(title = "CO2 Emissions - Power Industry (Mt CO2e)", y = "Mt CO2e")
```



CO2 Emissions – Power Industry (Mt CO2e)
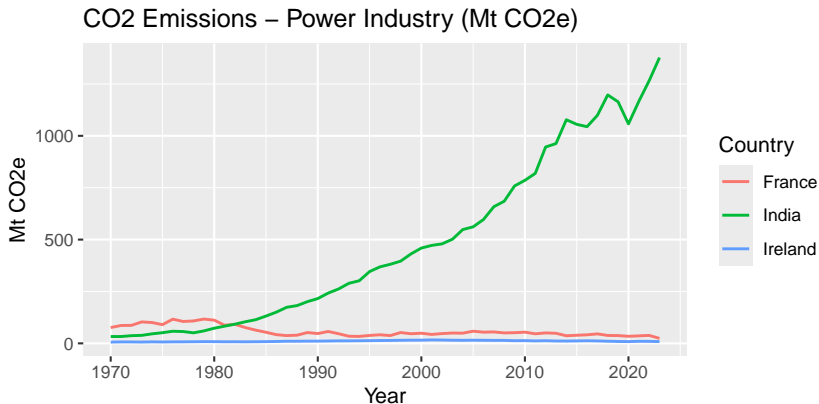
India's CO  emissions from power have grown massively since 1980.

France and Ireland remain flat with much lower levels.

Gap continues to widen due to India's rising energy demand.

# Slide 13 – Summary of Observations - Country Comparison

Ireland leads in GDP per capita, but fluctuates in social spending

France is consistent across indicators and invests the most in public services

India has rapid population growth and rising emissions, but low per-capita spending

GDP: Ireland highest, India lowest

Education & Health: France invests the most

Population: India growing fastest

Energy Use: France and Ireland higher per capita

CO Emissions: India's emissions rising fast from power sector