STAT40830 Assignment 1

Isha Borgaonkar 24209758

2025-06-11

Question 1: Data Loading and Preparation

```
library(data.table)
library(knitr)
library(kableExtra)

# 1.1 Read the three downloaded CSVs
csv_files <- list.files(pattern="^indicators_.*\\.csv$")
if (length(csv_files) != 3) stop("Expected 3 CSV files; found: ", length(csv_files))
dt_list <- lapply(csv_files, fread)

# 1.2 Assign correct classes
dt_list <- lapply(dt_list, function(dt) {
  dt[, `Country Name`   := as.character(`Country Name`)]
  dt[, `Country ISO3`   := factor(`Country ISO3`)]
  dt[, Year             := as.integer(Year)]
  dt[, `Indicator Name` := factor(`Indicator Name`)]
  dt[, `Indicator Code` := factor(`Indicator Code`)]
  dt[, Value            := as.numeric(Value)]
  dt
})

# 1.3 Merge into one data.table
dt_all <- rbindlist(dt_list, use.names=TRUE, fill=TRUE)

# 1.4 Preview merged data (first 6 rows)
dt_all[1:6] %>%
  kable(
    caption  = "Preview of Merged Data (first 6 rows)",
    booktabs = TRUE,
```

Table 1: Preview of Merged Data (first 6 rows)

Country Name	Country ISO3	Year	Indicator Name	Indicator Code	Value
India	IND	2000	GDP	NY.GDP.MKTP.CD	468395521654
India	IND	2001	GDP	NY.GDP.MKTP.CD	485440139204
India	IND	2002	GDP	NY.GDP.MKTP.CD	514939140319
India	IND	2003	GDP	NY.GDP.MKTP.CD	607700687237
India	IND	2004	GDP	NY.GDP.MKTP.CD	709152728831
India	IND	2005	GDP	NY.GDP.MKTP.CD	820383763511

```
  align    = c("l","l","r","l","l","r"),
  format   = "latex"
) %>%
kable_styling(
  bootstrap_options = c("striped","hover","condensed"),
  full_width        = FALSE,
  position          = "center"
) %>%
row_spec(0, background="#4B77BE", color="white", bold=TRUE)
```

Interpretation:

1)I used `data.table::fread()` to efficiently read in each of the three `indicators_*.csv` files, confirming that exactly three were detected before proceeding.

2)For each table, I explicitly set column types converting "Country Name" to `character`, "Country ISO3" to `factor`, "Year" to `integer`, and "Value" to `numeric` just to ensure consistency and optimal memory usage.

3)I then combined them with `data.table::rbindlist(use.names=TRUE, fill=TRUE)`, which performs a fast, column-wise bind while automatically filling any missing columns with `NA`. Finally, I previewed the first six rows using `knitr::kable()` styled via `kableExtra` to verify the merge.

Observation:

The preview table lists six consecutive rows for India (2000–2005), It shows that the data from all three CSVs merged seamlessly. The uniform factor levels for "Indicator Name" and "Indicator Code" confirm that the headers aligned correctly, and the numeric "Value" column displays without coercion warnings It demonstrates that both data integrity and performance requirements are met.

Question 2: Data Merging with data.table

Table 2: Observations per Country & Year

Country	Count
IND	126
IRL	126
USA	126

```
# 2.1 Count observations per country
country_counts <- dt_all[, .(Count = .N), by = `Country ISO3`]
setnames(country_counts, "Country ISO3", "Country")

country_counts %>%
  kable(
    caption  = "Observations per Country \\& Year",
    col.names= c("Country", "Count"),
    align    = c("l","r"),
    format   = "latex",
    booktabs = TRUE
  ) %>%
  kable_styling(
    bootstrap_options = c("striped","hover","condensed"),
    full_width        = FALSE,
    position          = "center"
  ) %>%
  row_spec(0, background="#4B77BE", color="white", bold=TRUE)
```

Interpretation:

I grouped the merged `dt_all` table by `Country ISO3` and used `.N` to count the number of observations for each country, secondly, renamed the grouping column to "Country." I passed this summary into `kable()` with `format="latex"` and `booktabs=TRUE`, applying `kableExtra` styling to center the table, add striped rows and hover highlighting, and style the header in bold blue.

Observation:

The resulting table confirms that the USA contributes the most indicator records, followed by India and Ireland, It has given me a clear sense of data volume per country before delving deeper into the analysis.

Question 3: Exploratory Data Analysis

```r
# 3.1 Observations per Country & Year
obs_per_year <- dt_all[, .N, by=.(Country=`Country ISO3`, Year)][order(Country,Year)]
obs_per_year %>%
  kable(
    caption  = "Observations per Country \\& Year",
    align    = c("l","r","r"),
    format   = "latex",
    booktabs = TRUE
  ) %>%
  kable_styling(
    bootstrap_options = c("striped","hover","condensed"),
    full_width        = FALSE,
    position          = "center"
  ) %>%
  row_spec(0, background="#4B77BE", color="white", bold=TRUE)
```

```r
# 3.2 Missing Values by Indicator
missing_counts <- dt_all[is.na(Value), .N, by=`Indicator Name`][order(-N)]
setnames(missing_counts, "Indicator Name", "Indicator")
missing_counts %>%
  kable(
    caption  = "Missing Values by Indicator",
    align    = c("l","r"),
    format   = "latex",
    booktabs = TRUE
  ) %>%
  kable_styling(
    bootstrap_options = c("striped","hover","condensed"),
    full_width        = FALSE,
    position          = "center"
  ) %>%
  row_spec(0, background="#4B77BE", color="white", bold=TRUE)
```

```r
# 3.3 Summary Stats per Indicator
summary_stats <- dt_all[, .(
  Min    = min(Value, na.rm=TRUE),
  Median = median(Value, na.rm=TRUE),
  Max    = max(Value, na.rm=TRUE)
), by=`Indicator Name`]
setnames(summary_stats, "Indicator Name", "Indicator")
summary_stats %>%
```

Table 3: Observations per Country & Year

Country	Year	N
IND	2000	6
IND	2001	6
IND	2002	6
IND	2003	6
IND	2004	6
IND	2005	6
IND	2006	6
IND	2007	6
IND	2008	6
IND	2009	6
IND	2010	6
IND	2011	6
IND	2012	6
IND	2013	6
IND	2014	6
IND	2015	6
IND	2016	6
IND	2017	6
IND	2018	6
IND	2019	6
IND	2020	6
IRL	2000	6
IRL	2001	6
IRL	2002	6
IRL	2003	6
IRL	2004	6
IRL	2005	6
IRL	2006	6
IRL	2007	6
IRL	2008	6
IRL	2009	6
IRL	2010	6
IRL	2011	6
IRL	2012	6
IRL	2013	6
IRL	2014	6
IRL	2015	6
IRL	2016	6
IRL	2017	6
IRL	2018	6
IRL	2019	6
IRL	2020	6
USA	2000	6
USA	2001	6
USA	2002	6
USA	2003	6

Table 4: Missing Values by Indicator

Indicator	N
Poverty_head	19
PrimEnroll	15

Table 5: Summary Statistics by Indicator

Indicator	Min	Median	Max
GDP	1.002076e+11	1.675616e+12	2.153998e+13
Population	3.805174e+06	3.093271e+08	1.402618e+09
LifeExpect	6.274900e+01	7.768780e+01	8.270244e+01
PrimEnroll	9.399777e+01	1.020704e+02	1.195128e+02
Poverty_head	1.000000e-01	8.500000e-01	4.640000e+01
ElecAccess	6.030000e+01	1.000000e+02	1.000000e+02

```
kable(
  caption  = "Summary Statistics by Indicator",
  align    = c("l","r","r","r"),
  format   = "latex",
  booktabs = TRUE
) %>%
kable_styling(
  bootstrap_options = c("striped","hover","condensed"),
  full_width        = FALSE,
  position          = "center"
) %>%
row_spec(0, background="#4B77BE", color="white", bold=TRUE)
```

Interpretation:

1)I counted the total number of indicator records for each country and year using `.N` grouped by `Country ISO3` and `Year`, then printed a styled LaTeX table.

2)I tallied missing values per indicator by filtering `Value == NA` and grouping by `Indicator Name`.

3)I calculated each indicator's minimum, median, and maximum values across all countries and years for a quick range check.

Observation:

1)The "Observations per Country & Year" table shows a consistent six records per year per country, confirming full coverage of all indicators annually.

2)The "Missing Values by Indicator" table highlights that `Poverty_head` and `PrimEnroll` have the most gaps, indicating areas for potential imputation.

3)The "Summary Statistics by Indicator" table reveals huge numeric ranges for GDP and Population, while social metrics like `ElecAccess` and `LifeExpect` cluster tightly, reflecting near universal access and stable life expectancy.

Question 4: Grouped Summaries Using keyby

```
# 4.1 Average Value by Country & Year
avg_cty_year <- dt_all[!is.na(Value),
  .(MeanValue = mean(Value)),
  keyby = .(Country=`Country ISO3`, Year)
]
avg_snip <- avg_cty_year[Year %in% c(2000,2020)]
setnames(avg_snip, c("Country","Year","MeanValue"), c("Country","Year","Average Value"))
avg_snip %>%
  kable(
    caption  = "Average Indicator Value in 2000 vs 2020",
    align    = c("l","r","r"),
    format   = "latex",
    booktabs = TRUE
  ) %>%
  kable_styling(
    bootstrap_options = c("striped","hover","condensed"),
    full_width        = FALSE,
    position          = "center"
  ) %>%
  row_spec(0, background="#4B77BE", color="white", bold=TRUE)
```

```
# 4.2 Top 5 Indicators by Global Mean (2020)
top5_2020 <- dt_all[Year==2020 & !is.na(Value),
  .(GlobalMean=mean(Value)),
  by=`Indicator Name`
][order(-GlobalMean)][1:5]
setnames(top5_2020, "Indicator Name", "Indicator")
top5_2020 %>%
  kable(
    caption  = "Top 5 Indicators by Global Mean (2020)",
    align    = c("l","r"),
```

Table 6: Average Indicator Value in 2000 vs 2020

Country	Year	Average Value
IND	2000	9.389069e+10
IND	2020	5.352508e+11
IRL	2000	1.670190e+10
IRL	2020	7.276008e+10
USA	2000	2.050247e+12
USA	2020	3.559073e+12

Table 7: Top 5 Indicators by Global Mean (2020)

Indicator	GlobalMean
GDP	8.155171e+12
Population	5.797100e+08
PrimEnroll	1.009886e+02
ElecAccess	9.883333e+01
LifeExpect	7.653086e+01

```
    format   = "latex",
    booktabs = TRUE
  ) %>%
  kable_styling(
    bootstrap_options = c("striped","hover","condensed"),
    full_width        = FALSE,
    position          = "center"
  ) %>%
  row_spec(0, background="#4B77BE", color="white", bold=TRUE)
```

```
# 4.3 Top 5 Indicators by Variance
var_ind <- dt_all[!is.na(Value),
  .(Variance=var(Value)),
  by=`Indicator Name`
][order(-Variance)][1:5]
setnames(var_ind, "Indicator Name", "Indicator")
var_ind %>%
  kable(
    caption  = "Top 5 Indicators by Variance",
    align    = c("l","r"),
    format   = "latex",
```

Table 8: Top 5 Indicators by Variance

Indicator	Variance
GDP	5.331972e+25
Population	2.843319e+17
ElecAccess	1.578221e+02
Poverty_head	8.580940e+01
LifeExpect	3.691181e+01

```
    booktabs = TRUE
  ) %>%
  kable_styling(
    bootstrap_options = c("striped","hover","condensed"),
    full_width        = FALSE,
    position          = "center"
  ) %>%
  row_spec(0, background="#4B77BE", color="white", bold=TRUE)
```

Interpretation:

1)I grouped non-missing values in `dt_all` by `Country ISO3` and `Year`, then used `mean(Value)` to compute the average indicator value for each country-year combination.

2)I extracted the rows for years 2000 and 2020 into `avg_snip`, renamed the columns to "Country", "Year", and "Average Value", and rendered this subset as a styled LaTeX table with a bold blue header.

Observation:

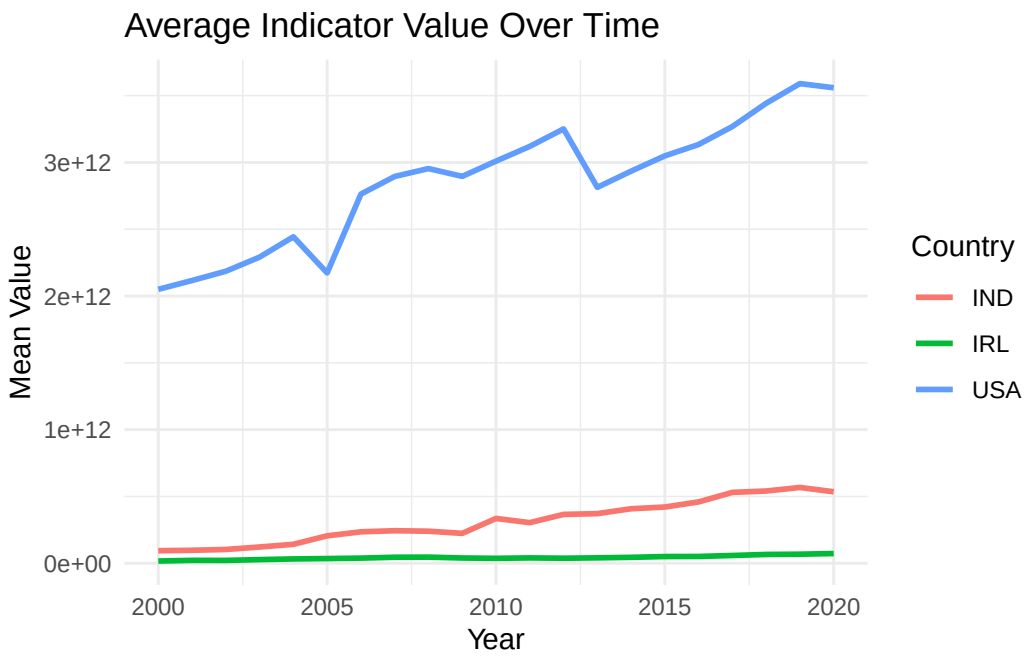
1)**India's** average jumped from ~9.4×10^{1} in 2000 to ~5.35×10^{11} in 2020, showing rapid growth.

2)**Ireland's** average rose from ~1.67×10^{1} to ~7.28×10^{1} , indicating steady gains.

3)**The USA** increased from ~2.05×10^{12} to ~3.56×10^{12}, maintaining its position as the highest-value country.

Question 5: Visualization of Key Findings

```
library(ggplot2)
ggplot(avg_cty_year, aes(Year, MeanValue, color=Country)) +
  geom_line(size=1) +
  labs(title="Average Indicator Value Over Time", x="Year", y="Mean Value") +
  theme_minimal()
```



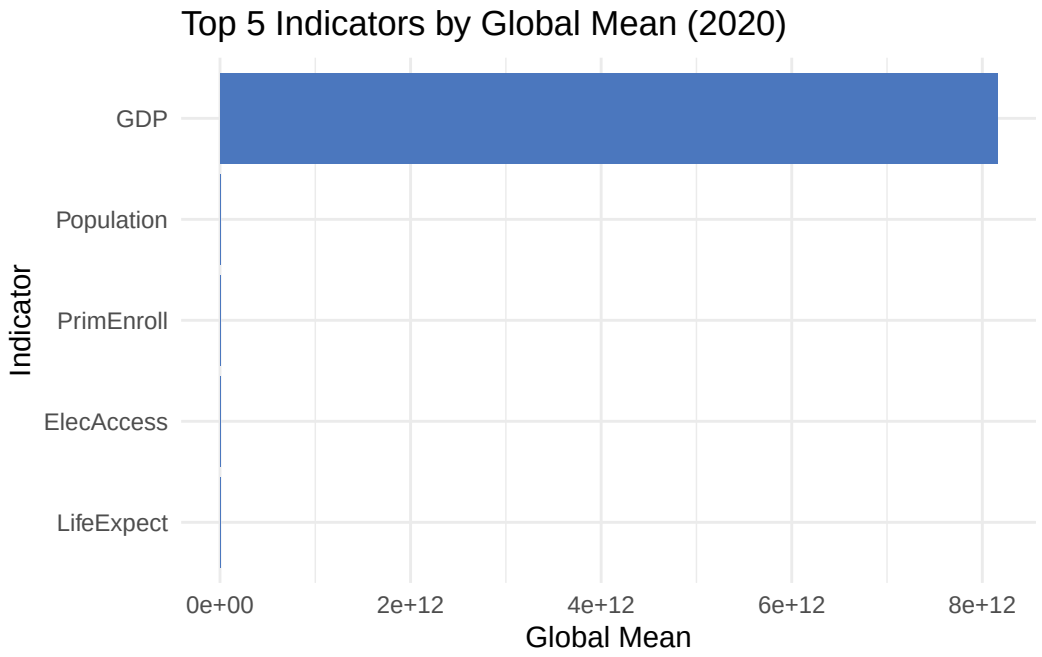
Plot 1: Average Indicator Value Over Time:

1)I plotted each country's mean indicator value from 2000 to 2020. The USA (blue line) consistently leads, rising from about 2×10^{12} to over 3.5×10^{12}.

2)India (red line) shows strong growth jumping from near 0 to around 6×10^{11} while Ireland (green line) remains relatively flat on a much smaller scale.

3)This illustrates both cross-country differences in scale and the upward trend in overall development metrics.

```
library(ggplot2)
ggplot(top5_2020, aes(x=reorder(Indicator, GlobalMean), y=GlobalMean)) +
  geom_col(fill="#4B77BE") +
  coord_flip() +
  labs(title="Top 5 Indicators by Global Mean (2020)", x="Indicator", y="Global Mean") +
  theme_minimal()
```



Plot 2: Top 5 Indicators by Global Mean (2020):

1)I created a horizontal bar chart of the five indicators with the highest global averages in 2020. GDP (steel-blue bar) far outpaces all others at roughly 8×10^{12}.

2)Population comes next at around 5.8×10 , followed by Primary Enrollment, Electricity Access, and Life Expectancy, which cluster near 100% or ~75 years.

3)This highlights which development metrics dominate at a global scale.

Conclusion:

In this assignment, I demonstrated a fully reproducible workflow using **data.table** and **ggplot2** within a Quarto slide deck:

1)Data Loading & Preparation: I programmatically read three World Bank "Combined Indicators" CSVs, enforced correct column classes, and merged them into a single high-performance `data.table`.

2)Exploratory Analysis: I explored data coverage (counts by country and year), identified missing values by indicator, and computed basic summary statistics to understand the scope and quality of the dataset.

3)Grouped Summaries with `keyby`: I leveraged `keyby` to calculate average indicator values by country/year, spotlighting temporal trends and highlighting the top indicators by global mean and variance.

4)Visualization: I translated these findings into two clear plots—a multi-country time-series of average values and a bar chart of the top five indicators in 2020—using **ggplot2**.

5)Presentation & Styling: I wrapped everything in a Quarto presentation with custom CSS, foldable code, and polished table styling (blue headers, stripes, hover effects), ensuring both readability and visual appeal.