R Assignment 1 - Summer

Sanmesh Sanjay Shintre

Table of contents I

Section 1

Use data.table to read in the data

Use data.table to read in the data

```
library(data.table)
library(ggplot2)
library(knitr)
library(dplyr)

#Loading the datasets
dt_india <- fread("indicators_ind.csv")[-1,]
dt_are <- fread("indicators_are.csv")[-1,]
dt_usa <- fread("indicators_usa.csv")[-1,]
```

- Loading the packages required.
- Loaded the dataset using fread for 3 countries.

Section 2

Assign the correct class to the variables

Assign the correct class to the variables

```
options(width = 70)
dt_india[, `:=`(
  `Country Name` = as.factor(`Country Name`),
  `Country ISO3` = as.factor(`Country ISO3`),
  `Indicator Name` = as.factor(`Indicator Name`),
  `Indicator Code` = as.factor(`Indicator Code`),
  Year = as.integer(Year),
  Value = as.numeric(Value)
)]
dt_are[, `:=`(
  `Country Name` = as.factor(`Country Name`),
  `Country ISO3` = as.factor(`Country ISO3`),
  `Indicator Name` = as.factor(`Indicator Name`),
  `Indicator Code` = as.factor(`Indicator Code`),
  Year = as.integer(Year),
  Value = as.numeric(Value)
)]
dt_usa[, `:=`(
  `Country Name` = as.factor(`Country Name`),
  `Country ISO3` = as.factor(`Country ISO3`),
  `Indicator Name` = as.factor(`Indicator Name`),
  `Indicator Code` = as.factor(`Indicator Code`),
  Year = as.integer(Year),
  Value = as.numeric(Value)
)]
```

Combining 3 datasets

```
dt_all <- rbindlist(list(dt_india, dt_are, dt_usa)
                    , use.names = TRUE, fill = TRUE)
```

- Assigned the correct class for each of the column of dataset for respective countries.
- Such as integer and numeric for Year and Value respectively.
- Combined 3 datasets into one single dataset using rbindlist().

Section 3

Data Exploration

Data Exploration

```
dt_all[, .N, by = .(`Country Name`,
                    `Indicator Name`)][, .N, by = `Country Name`]
```

```
           Country Name     N
                 <fctr> <int>
1:                India  3633
2: United Arab Emirates  1921
3:        United States  2000
```

```
common_indicators <- Reduce(intersect, list(
  unique(dt_india$`Indicator Name`),
  unique(dt_are$`Indicator Name`),
  unique(dt_usa$`Indicator Name`)
))
length(common_indicators)
```

```
[1] 1739
```

- Taking the common indicators from dt_all to common_indicators.
- Printing the number of common_indicators in the dataset.

Section 4

Data Exploration and plots

Data Exploration and plots

```
indicator_counts <- dt_all[, .N, by = .(`Country Name`, `Indicator Name`)][, .N, by

top_indicators_overall <- dt_all %>%
  group_by(`Indicator Name`) %>%
  summarise(Count = n(), .groups = "drop") %>%
  slice_max(order_by = Count, n = 15)
```

- Count **how many indicators** each country has: two-step grouping with .N.
- Prepare two exploratory outputs:
  1. **indicator_counts** (per country)
  2. **top_indicators_overall** (top 15 by frequency)

Section 5

Data Analysis using data.table (with keyby)

Data Analysis using data.table (with keyby)

```
dt_filtered <- dt_all[`Indicator Name` %in% common_indicators]
dt_migration_summary <- dt_filtered[
  `Indicator Name` == "Net migration",
  .(Average_Migration = mean(Value, na.rm = TRUE)),
  keyby = .(Year, `Country Name`)
]
dt_fuel_exports <- dt_filtered[`Indicator Name` ==
                                 "Fuel exports (% of merchandise exports)"]
dt_fuel_summary <- dt_fuel_exports[,.(
    Avg_Export = mean(Value, na.rm = TRUE),
    SD_Export = sd(Value, na.rm = TRUE),
    Min = min(Value, na.rm = TRUE),
    Max = max(Value, na.rm = TRUE),
    Observations = .N
  ),keyby = `Country Name`
]
```

Data Analysis using data.table (with keyby)

```
dt_fuel_summary
```

```
Key: <Country Name>
           Country Name Avg_Export SD_Export       Min      Max
                 <fctr>      <num>     <num>     <num>    <num>
1:               India   6.845227  7.041200 0.2473565 21.75247
2: United Arab Emirates  56.501599 24.217227 7.2975111 94.91210
3:        United States   5.571262  4.617027 1.5251860 21.41017
   Observations
          <int>
1:          186
2:           93
3:          186
```

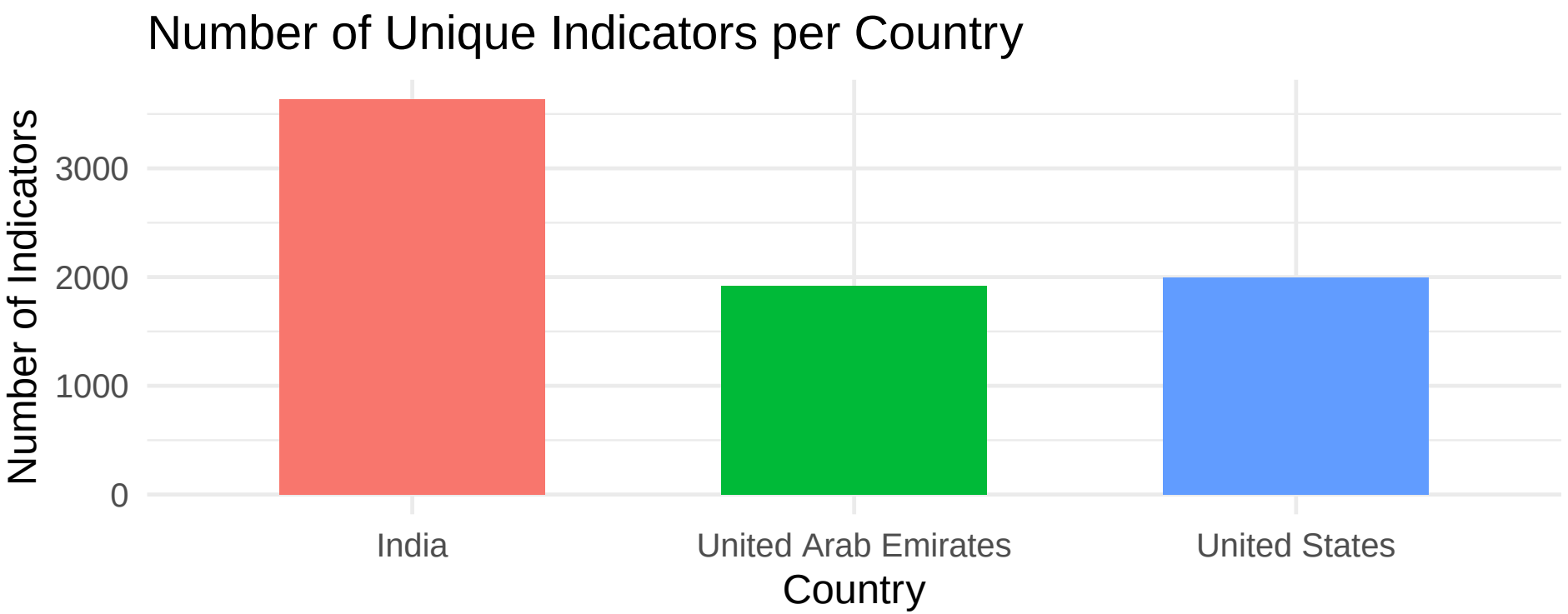
- **Filter** to only the common_indicators.
- Use **keyby** to efficiently compute:
  - **Average Net Migration** by Year & Country
  - **Fuel Exports** summary stats by Country

Section 6

Plots

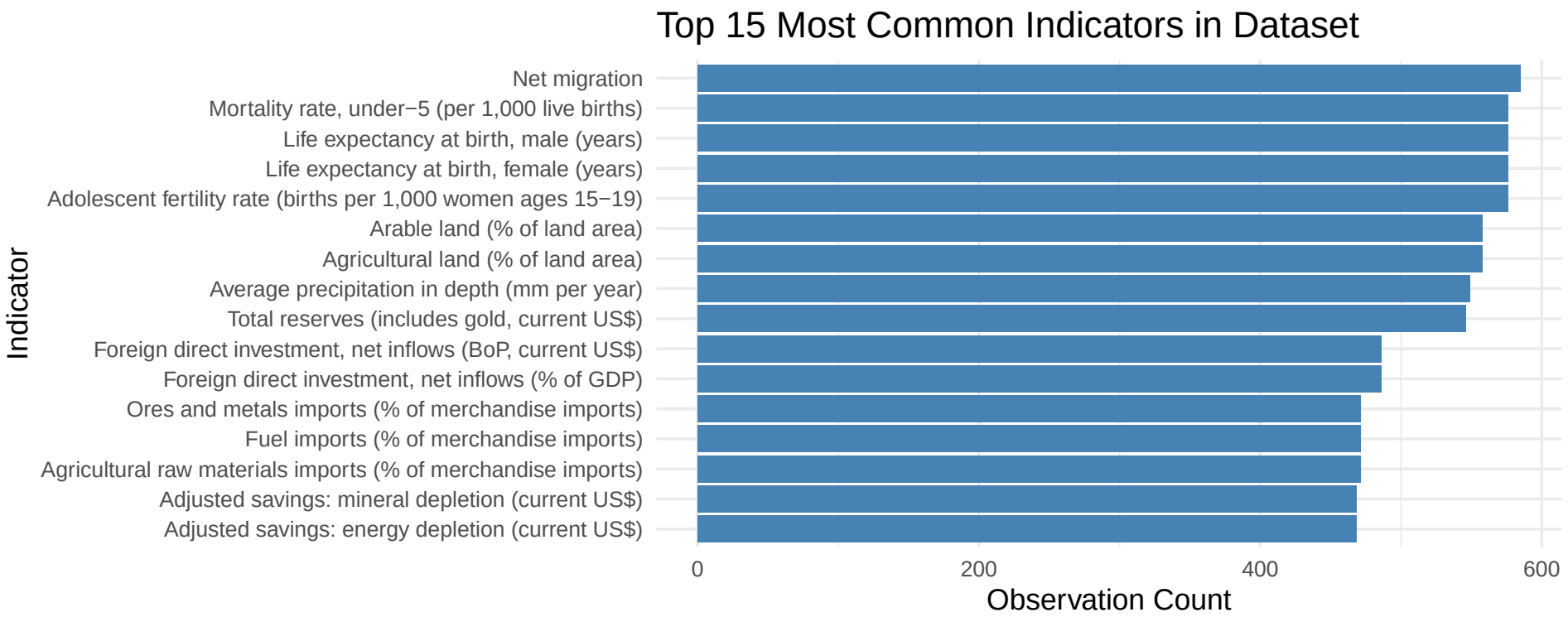
Plot 1 of Analysis

```r
ggplot(indicator_counts, aes(x = `Country Name`, y = N, fill = `Country Name`)) +
  geom_col(width = 0.6) +
  theme_minimal() +
  labs(
    title = "Number of Unique Indicators per Country",
    x = "Country",
    y = "Number of Indicators"
  ) +
  theme(legend.position = "none")
```



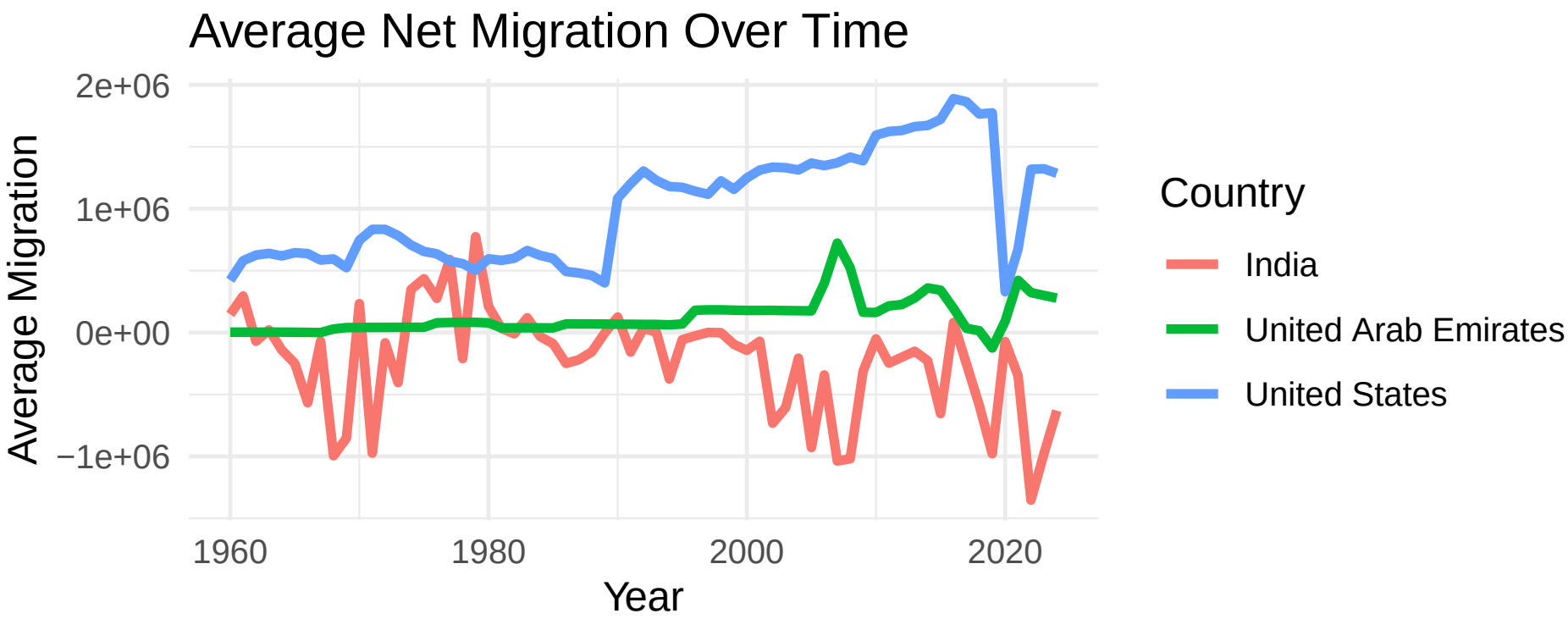
Plot 2 of Analysis

```
ggplot(top_indicators_overall, aes(x = reorder(`Indicator Name`, Count), y = Count))
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(
    title = "Top 15 Most Common Indicators in Dataset",
    x = "Indicator",
    y = "Observation Count"
  ) +
  theme_minimal(base_size = 8)
```



Plot 3 of Analysis

```
ggplot(dt_migration_summary, aes(x = Year, y = Average_Migration,
                                 color = `Country Name`)) +
  geom_line(linewidth = 1.2) +
  theme_minimal() +
  labs(
    title = "Average Net Migration Over Time",
    x = "Year",
    y = "Average Migration",
    color = "Country"
  )
```



Plot 4 of Analysis

```r
ggplot(dt_fuel_summary, aes(x = `Country Name`, y = Avg_Export,
                            fill = `Country Name`)) +
  geom_col(width = 0.6) +
  theme_minimal() +
  labs(
    title = "Average Fuel Exports (% of Merchandise Exports)",
    x = "Country", y = "Average % of Fuel Exports"
  ) +
  theme(legend.position = "none")
```