Assignment 1

Rakshitha Annaiah

2025-06-11

Introduction

This assignment focuses on using the data.table package in R to explore and analyze Economic, Environmental, and Agricultural indicators from the World Bank for three countries. I have chosen India, France, and the United States data.

It demonstrates: Efficient data handling using fread(), rbindlist(), keyby Data exploration and summarization Visualizations using ggplot2

Libraries Used

```
#Loading necessary Libraries
library(data.table)
library(ggplot2)
```

Step 1: Reading the Data

Each .csv file contains indicators for one country which are downloaded from data.humdata.org.

```
# Reading data
usa <- fread("indicators_usa.csv", skip = 1)
fra <- fread("indicators_fra.csv", skip = 1)
ind <- fread("indicators_ind.csv", skip = 1)
```

Step 2: Cleaning Column Names and Converting Types

We are standardizing column names across all three datasets to ensure consistency and convert the year and value columns to numeric types for analysis.

```
#Standardize column names
setnames(usa, old = c("#country+name", "#country+code", "#date+year"
                      "#indicator+name", "#indicator+code",
                      "#indicator+value+num"),
              new = c("Country Name", "Country ISO3", "Year",
                      "Indicator Name", "Indicator Code", "Value"))
setnames(fra, names(usa))
setnames(ind, names(usa))

# Convert year and value
cols <- c("Year", "Value")
for (df in list(usa, fra, ind)) {
  df[, (cols) := lapply(.SD, as.numeric), .SDcols = cols]
}
```

Step 3: Merging the Data

We are merging all three country datasets into one using rbindlist() and preview the result.

```
# Load required library
library(kableExtra)

# Combining the three cleaned datasets
combined <- rbindlist(list(usa, fra, ind))
```

```r
# Previewing first 10 rows
combined[1:5] |>
  kbl(booktabs = TRUE, caption = "Preview of Combined Dataset") |>
  kable_styling(latex_options = c("striped", "scale_down"))
```

Table 1: Preview of Combined Dataset

Country Name	Country ISO3	Year	Indicator Name	Indicator Code	Value
United States	USA	2022	Fertilizer consumption (% of fertilizer production)	AG.CON.FERT.PT.ZS	100.7912
United States	USA	2021	Fertilizer consumption (% of fertilizer production)	AG.CON.FERT.PT.ZS	110.0026
United States	USA	2020	Fertilizer consumption (% of fertilizer production)	AG.CON.FERT.PT.ZS	106.6126
United States	USA	2019	Fertilizer consumption (% of fertilizer production)	AG.CON.FERT.PT.ZS	104.6716
United States	USA	2018	Fertilizer consumption (% of fertilizer production)	AG.CON.FERT.PT.ZS	109.3041

Step 4: Selecting Key Indicators

We will focus on five important indicators to summarize and visualize:

```
# Key indicators
indicators <- c(
  "Fertilizer consumption (kilograms per hectare of arable land)",
  "Agricultural land (% of land area)",
  "Arable land (% of land area)",
  "Permanent cropland (% of land area)",
  "Land under cereal production (hectares)"
)
```

Step 5: Summary Table

We calculate the mean value of each indicator by country and format it into a neat wide-format table.

```
# Calculating mean indicator value by country
summary_dt <- combined[
  `Indicator Name` %in% indicators,
  .(Mean = mean(Value, na.rm = TRUE)),
  keyby = .(`Country Name`, `Indicator Name`)
]
# Converting to wide format
summary_wide <- dcast(summary_dt,
                      `Indicator Name` ~ `Country Name`,
                      value.var = "Mean")
```

```
#summary table
library(kableExtra)
summary_wide |>
  kbl(booktabs = TRUE, digits = 2,
      caption = "Average Indicator Values by Country") |>
  kable_styling(latex_options = c("striped", "scale_down"))
```

Table 2: Average Indicator Values by Country

Indicator Name	France	India	United States
Agricultural land (% of land area)	56.29	60.43	46.33
Arable land (% of land area)	33.03	53.77	19.20
Fertilizer consumption (kilograms per hectare of arable land)	236.17	84.12	103.27
Land under cereal production (hectares)	9323839.76	99966895.03	61706837.95
Permanent cropland (% of land area)	2.39	2.73	0.25

Interpretation: Average Indicator Values by Country

- India has the highest percentage of agricultural and arable land, as well as the largest land area under cereal production — highlighting its agrarian economy.
- France leads in fertilizer consumption, suggesting more intensive farming practices per hectare of arable land.
- The United States has significantly lower permanent cropland and arable land as a percentage of total land, but still maintains a large area of cereal production — indicating extensive, large-scale farming.

Step 6: Visualizations

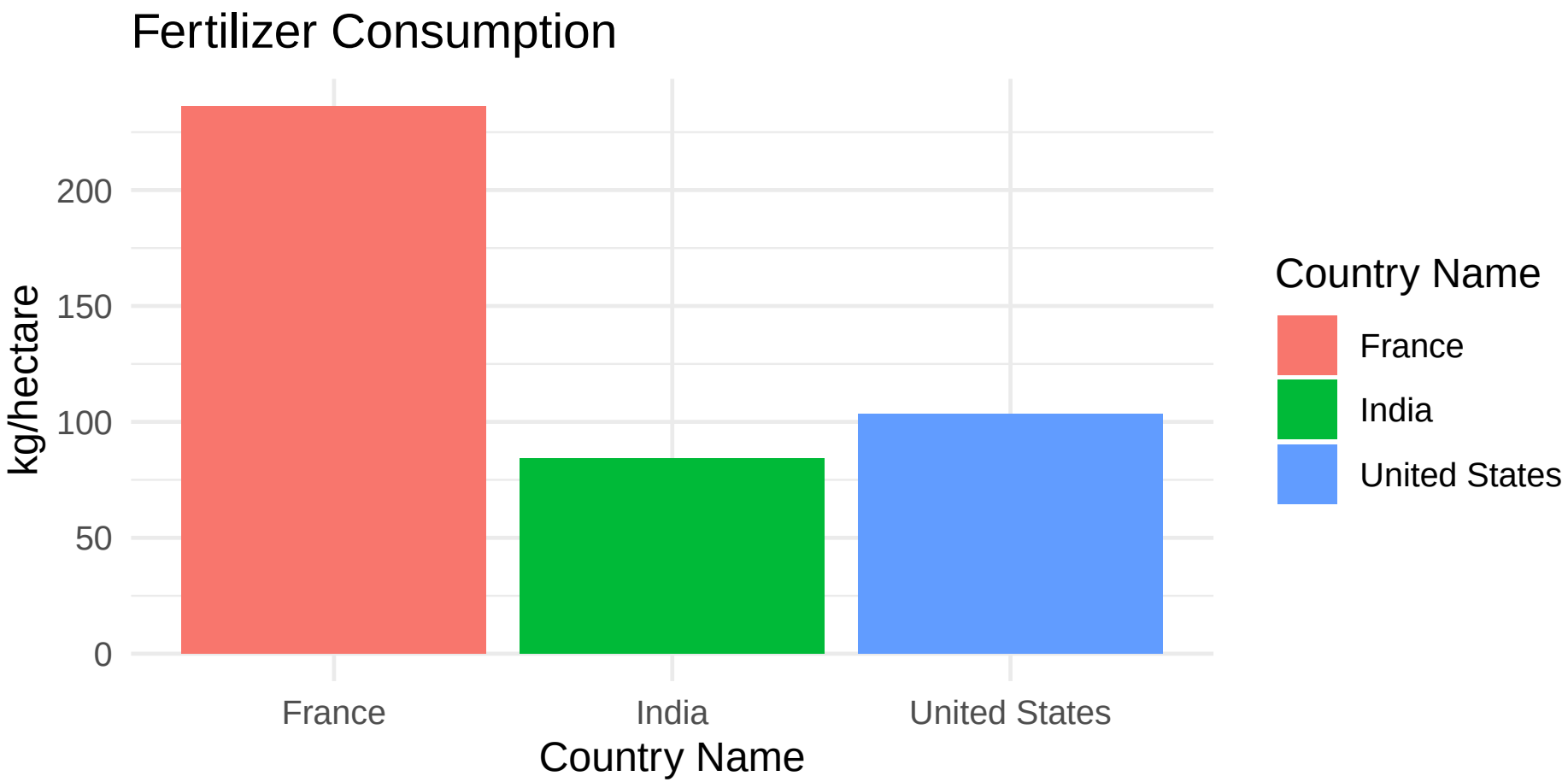
In this step, I created two visualizations to compare and analyze key agricultural indicators across countries:

- Plot 1 is a bar chart showing average fertilizer consumption (kg/hectare) for France, India, and the United States. This plot highlights the intensity of agricultural input usage.
- Plot 2 is a line chart comparing land use indicators (% of land area) such as agricultural land, arable land, and permanent cropland. This format provides a cleaner, more readable alternative to traditional bar plots.

Fertilizer Consumption by Country

```
# Plot 1: Fertilizer consumption
fert_data <- summary_dt[
  `Indicator Name` == indicators[1]
]
```

```
ggplot(fert_data, aes(x = `Country Name`, y = Mean, fill = `Country
  geom_bar(stat = "identity") +
  labs(title = "Fertilizer Consumption", y = "kg/hectare") +
  theme_minimal()
```



Interpretation: Fertilizer Consumption

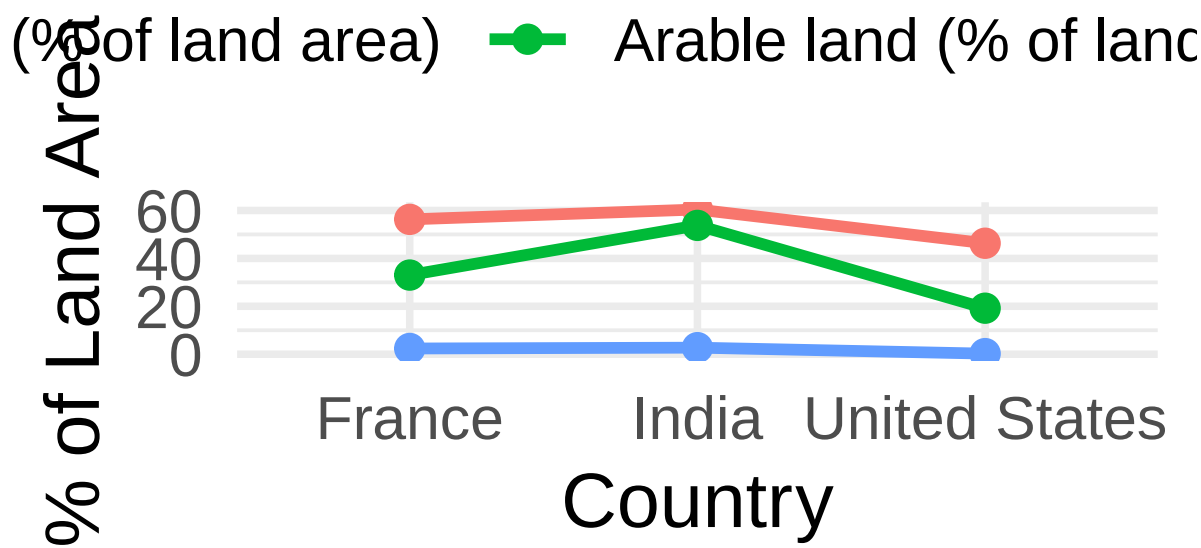
- France uses the highest amount of fertilizer per hectare, more than double that of India.
- India and the United States have moderate usage, but India's rate is slightly lower.
- The high usage in France suggests more intensive agricultural input practices, while India and the U.S. may rely more on extensive cultivation or different crop mixes.

Land Use Comparison

```
land_data <- summary_dt[
  `Indicator Name` %chin% indicators[2:4]
]

base_plot <- ggplot(land_data, aes(x = `Country Name`, y = Mean, col
  geom_line(linewidth = 1) +
  geom_point(size = 2)
```

```
base_plot +
  labs(title = "Land Use Indicators by Country",
       y = "% of Land Area", x = "Country") +
  theme_minimal(base_size = 14) +
  theme(
    legend.position = "top",
    plot.title = element_text(face = "bold", hjust = 0.5)
  )
```



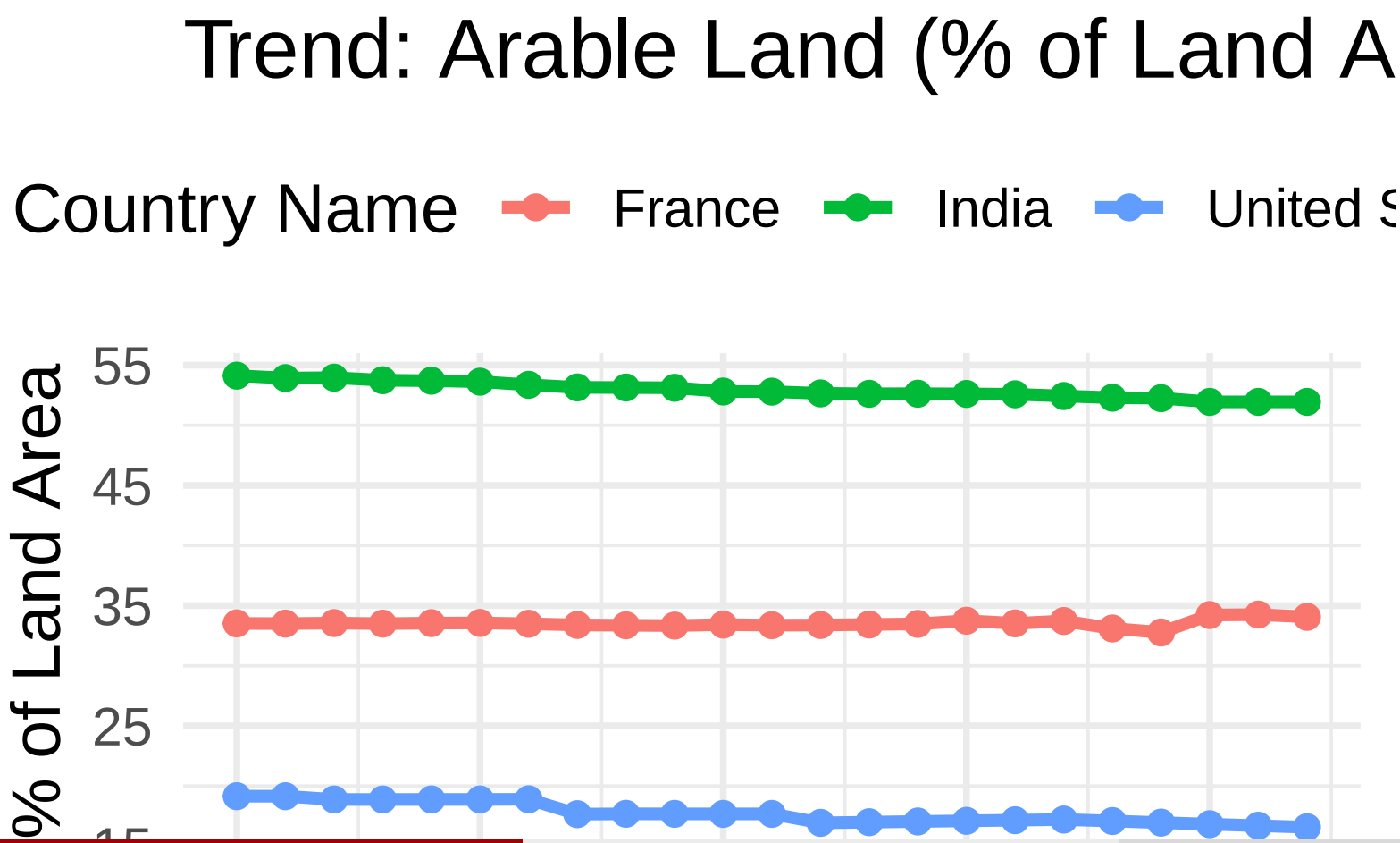
Interpretation: Land Use Indicators

- India has the highest percentages for both agricultural land and arable land, reflecting its large agrarian base.
- France follows, with moderate values in all three indicators, including a balanced share of permanent cropland.
- The United States has the lowest proportion of arable and permanent cropland, indicating more extensive and mechanized land use rather than intensive cultivation.

Step 7: Trend Analysis with keyby

```
trend <- combined[
  `Indicator Name` == indicators[3] & Year >= 2000,
  .(Mean = mean(Value, na.rm = TRUE)),
  keyby = .(Year, `Country Name`)
]
```

```r
ggplot(trend, aes(x = Year, y = Mean, color = `Country Name`)) +
  geom_line(linewidth = 1.2) +
  geom_point(size = 2) +
  labs(title = "Trend: Arable Land (% of Land Area)", y = "% of Land
  theme_minimal(base_size = 14) +
  theme(legend.position = "top")
```



Interpretation: Arable Land Trend (2000–2022)

- India consistently has the highest percentage of arable land (~53–54%), though it shows a slow, steady decline over time.
- France maintains a stable level of arable land (~34%), with minor yearly variations.
- The United States shows the lowest percentage, steadily decreasing from ~19% to around 17%, reflecting long-term land use shifts or reclassification.

  This trend suggests that India remains heavily dependent on arable land, while the U.S. may be using more land for other purposes or adopting more land-efficient agricultural practices.

Conclusion

In this assignment, I used data.table to efficiently analyze land and agriculture indicators for India, France, and the U.S.The analysis showed that India has the most arable land, France uses the most fertilizer, and the U.S. has the least cropland.