**Roll Number:**

1. **202311887**
2. **202312187**
3. **202312188**
4. **202312104**
5. **202312372**
6. **202312192**
7. **202311865**

**Climate Change Indicators Dataset Analysis Report**

Author: Group 1  
Course: R Programming  
Instructor: Uwamahoro Leopold  
Date: April 5, 2025

**1. Introduction**

This report presents an exploratory data analysis (EDA) of the **climate\_change\_indicators.csv** dataset obtained from FAOSTAT (Food and Agriculture Organization of the United Nations). The dataset contains yearly temperature anomalies from 1961 to 2022 across multiple countries compared to a baseline period (1951–1980).

The objective of this analysis is to:

* Understand global warming trends
* Identify patterns in temperature changes
* Explore the relevance of this dataset for agricultural and environmental science

**2. Dataset Overview**

**Source:**

* FAO (Food and Agriculture Organization of the United Nations)
* License: CC BY-NC-SA 3.0 IGO
* Accessed on: March 28, 2023
* Link: [<https://www.fao.org/faostat/en/>#data/ET](<https://www.fao.org/faostat/en/>#data/ET)

**Key Features:**

* Rows: 222
* Columns: 64
* Variables:
  + Categorical: Country, ISO codes, Indicator, Unit
  + Numeric: Yearly temperature anomalies (F1961 to F2022)

**Relevance:**

* Agricultural Science: Helps assess how climate affects crop yields and irrigation.
* Environmental Science: Tracks global warming and regional variability.

**3. Data Inspection**

**a) Dimensions**

**Code Snippet:**

dim(data)

# Output: 222 rows, 64 columns

**b) First and Last Rows**

head(data)

tail(data)

**c) Summary Statistics**

summary(data)

**d) Structure of the Data**

str(data)

**e) Missing Values**

colSums(is.na(data)) # Count missing values per column

sum(is.na(data)) # Total number of missing values

**f) Duplicate Rows**

sum(duplicated(data))

**4. Data Cleaning**

**Steps Taken:**

* Option 1: Removed rows with any **NA** values using **na.omit()**.
* Option 2: Removed columns with more than 50% missing values.
* Option 3: Imputed numeric columns using mean values.
* Duplicates: Removed using **unique()**.

**Code Used:**

clean\_data <- na.omit(data)

missing\_percent <- colSums(is.na(data)) / nrow(data)

cols\_to\_remove <- names(missing\_percent[missing\_percent > 0.5])

clean\_data <- data %>% select(-all\_of(cols\_to\_remove))

clean\_data <- data %>%

mutate(across(where(is.numeric), ~ ifelse(is.na(.), mean(., na.rm = TRUE), .)))

clean\_data <- unique(clean\_data)

**5. Data Visualization**

**a) Histogram of F1961 Values**

ggplot(clean\_data, aes(x = F1961)) +

geom\_histogram(bins = 30, fill = "steelblue", color = "black") +

labs(title = "Distribution of F1961 Values", x = "F1961", y = "Frequency")

“**Caption:** This histogram shows the distribution of temperature anomalies in 1961 across different countries.”

**b) Boxplot of F1961 by Country**

ggplot(clean\_data, aes(x = Country, y = F1961)) +

geom\_boxplot() +

theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +

labs(title = "Boxplot of F1961 by Country", y = "Value", x = "Country")

“**Caption:** The boxplot highlights variations in temperature anomalies across different countries.”

**c) Scatter Plot Between F1961 and F1970**

ggplot(clean\_data, aes(x = F1961, y = F1970)) +

geom\_point(alpha = 0.6) +

labs(title = "Scatter Plot: F1961 vs F1970", x = "F1961", y = "F1970")

“**Caption:** A strong positive correlation between temperature anomalies in 1961 and 1970 is observed.”

**d) Bar Chart of Countries**

ggplot(clean\_data, aes(x = Country)) +

geom\_bar() +

theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +

labs(title = "Number of Entries per Country")

**e) Pair Plots (Only if Few Numeric Columns)**

numeric\_data <- clean\_data %>% select(where(is.numeric))

if(ncol(numeric\_data) <= 10) {

pairs(numeric\_data, main = "Pairwise Scatter Plots")

} else {

print("Too many numeric columns for pair plot.")

}

**f) Correlation Matrix Heatmap**

cor\_matrix <- cor(numeric\_data, use = "pairwise.complete.obs")

cor\_df <- as.data.frame(as.table(cor\_matrix))

ggplot(cor\_df, aes(Var1, Var2, fill = Freq)) +

geom\_tile() +

scale\_fill\_gradient2(low = "blue", high = "red", mid = "white",

midpoint = 0, limits = c(-1, 1), space = "Lab") +

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 45, vjust = 1, size = 8, hjust = 1)) +

coord\_fixed() +

labs(title = "Correlation Matrix Heatmap")

“Caption: High positive correlations between adjacent years suggest consistent warming trends over time. “

**5. Save Cleaned Dataset**

write\_csv(clean\_data, "C:\\Users\\Emmanuel G. Momo\\Desktop\\webscrapping\\cleaned\_climate\_data.csv")

**6. Initial Insights Summary**

**Key Features of the Dataset**

The dataset includes temperature anomalies from 1961 to 2022 across multiple countries compared to a baseline period (1951–1980). Each row represents a country's temperature deviation over time. Variables include:

* Categorical: Country, ISO codes, Indicator, Unit
* Numeric: Yearly temperature anomalies (F1961 to F2022)

This data is crucial for both agricultural science (impact on crop yields, irrigation) and environmental science (global warming trends).

**Observed Patterns and Trends**

* Rising Temperatures Over Time : Boxplots grouped by decade show a consistent increase in median temperature anomalies since the 1960s.
* Positive Correlation Between Years : Strong correlation between earlier and later years suggests that countries with higher-than-average temperatures in the past tend to remain warmer today.
* Uneven Data Coverage : Some countries have more complete records than others, potentially introducing bias in comparative analyses.

**Challenges Encountered During Data Exploration**

* Missing Values : Many temperature columns contained **NA** values, especially for earlier years. Imputation using column means helped retain sample size but may mask variability.
* Outliers : Some extreme temperature values were observed, possibly due to measurement errors or actual climatic extremes.
* Wide Format : The wide format made time-series analysis challenging, requiring reshaping into long format for better visualization and modeling.

**Future Research Questions & Machine Learning Tasks**

* Predictive Modeling : Forecast future temperature anomalies using ARIMA, Prophet, or LSTM neural networks.
* Clustering Countries : Group countries based on similar temperature trends using k-means or hierarchical clustering.
* Anomaly Detection : Use unsupervised learning to detect unusual temperature spikes indicating extreme weather events.
* Regression Analysis : Investigate factors influencing temperature changes (e.g., economic development, land use).

In conclusion, this dataset provides a robust foundation for interdisciplinary research at the intersection of agriculture, environment, and data science.

**7.** **R Code Used**

library(readr)

library(dplyr)

library(ggplot2)

library(tidyr)

# Load the dataset

data <- read\_csv("C:\\Users\\Emmanuel G. Momo\\Downloads\\climate\_change\_indicators.csv")

# Data Inspection

dim(data)

head(data)

tail(data)

summary(data)

str(data)

colSums(is.na(data))

sum(is.na(data))

sum(duplicated(data))

# Data Cleaning

clean\_data <- na.omit(data)

missing\_percent <- colSums(is.na(data)) / nrow(data)

cols\_to\_remove <- names(missing\_percent[missing\_percent > 0.5])

clean\_data <- data %>% select(-all\_of(cols\_to\_remove))

clean\_data <- data %>%

mutate(across(where(is.numeric), ~ ifelse(is.na(.), mean(., na.rm = TRUE), .)))

clean\_data <- unique(clean\_data)

# Visualizations

ggplot(clean\_data, aes(x = F1961)) +

geom\_histogram(bins = 30, fill = "steelblue", color = "black") +

labs(title = "Distribution of F1961 Values", x = "F1961", y = "Frequency")

ggplot(clean\_data, aes(x = Country, y = F1961)) +

geom\_boxplot() +

theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +

labs(title = "Boxplot of F1961 by Country", y = "Value", x = "Country")

ggplot(clean\_data, aes(x = F1961, y = F1970)) +

geom\_point(alpha = 0.6) +

labs(title = "Scatter Plot: F1961 vs F1970", x = "F1961", y = "F1970")

ggplot(clean\_data, aes(x = Country)) +

geom\_bar() +

theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +

labs(title = "Number of Entries per Country")

numeric\_data <- clean\_data %>% select(where(is.numeric))

if(ncol(numeric\_data) <= 10) {

pairs(numeric\_data, main = "Pairwise Scatter Plots")

} else {

print("Too many numeric columns for pair plot.")

}

cor\_matrix <- cor(numeric\_data, use = "pairwise.complete.obs")

cor\_df <- as.data.frame(as.table(cor\_matrix))

ggplot(cor\_df, aes(Var1, Var2, fill = Freq)) +

geom\_tile() +

scale\_fill\_gradient2(low = "blue", high = "red", mid = "white",

midpoint = 0, limits = c(-1, 1), space = "Lab") +

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 45, vjust = 1, size = 8, hjust = 1)) +

coord\_fixed() +

labs(title = "Correlation Matrix Heatmap")

**8. Save Cleaned Dataset**

**Code Snippet:**

write\_csv(clean\_data, "C:\\Users\\Emmanuel G. Momo\\Desktop\\webscrapping\\cleaned\_climate\_data.csv")

**9. GitHub Repository**

**GitHub Link:** https://github.com/emomo81/climate-change-analysis