Devanahalli, Bangalore-562129

SCHOOL OF ENGINEERING

A PROJECT REPORT

ON

"A CORRELATION ANALYSIS AND VISUALIZATION OF CLIMATE CHANGE USING EXPLORATORY DATA ANALYSIS"

Submitted By:

AMRUTA HEGDE

24PG00130

Under the guidance of:

Dr. Bhanu K.N.

Towards

M.Sc. Data Science / MCA - 2nd Semester

Data Visualization Project Lab (DAT203)

For the academic year **2024-2025**

DECLARATION

I, **Amruta Hegde,** hereby declare that this project work entitled '**A Correlation Analysis And Visualization Of Climate Change Using Exploratory Data Analysis'** is submitted in partial fulfilment for the award of the degree of **MCA** of **Chanakya University**.
I further declare that I have not submitted this project report either in part or in full to any other university for the award of any degree.

Date:	Student Name: AMRUTA HEGDE
Place: BENGALURU	Reg. No: 24PG00130

ABSTRACT

The increasing frequency and severity of natural disasters in recent decades underscore the urgent need for a comprehensive understanding of the relationship between climate change and disaster dynamics. This research addresses the critical challenge of integrating and analyzing large volumes of heterogeneous post-disaster data to derive meaningful insights for disaster preparedness, mitigation, and policy planning. Two primary datasets were curated and unified—one detailing global natural disaster statistics including human and economic losses (EADRF), and another capturing climate data through global temperature anomalies (CCATA).

The study leverages statistical correlation methods—**Pearson**, **Spearman**, and **Kendall**—to examine the strength and nature of relationships between disaster occurrences, economic damage, and rising global temperatures. A significant positive correlation is observed between temperature anomalies and disaster frequency, especially post-1980, validating concerns raised by climate scientists and environmentalists.

The analysis is operationalized through an interactive dashboard developed using **Tableau** for rich, intuitive visualizations including time-series charts, area graphs, dual-line comparisons, sunburst plots, and world maps. Additionally, **Python code executed in Google Colab** was employed for data cleaning, transformation, and generation of advanced statistical visuals like heatmaps and correlation matrices using libraries such as **Pandas, Seaborn, and Matplotlib**.

This integrated approach not only supports empirical validation of climate-disaster interdependencies but also empowers stakeholders to make informed decisions through dynamic, visual storytelling. The insights gained can inform strategic resource allocation, improve early warning systems, and guide sustainable development practices across vulnerable regions.

TABLE OF CONTENT

CHAPTER-1

INTRODUCTION

1.1 INTRODUCTION

In recent decades, the world has seen a significant rise in the frequency, severity, and impact of natural disasters such as floods, earthquakes, droughts, wildfires, and extreme weather events. These disasters have resulted in widespread human suffering, economic damage, and displacement. Scientific evidence increasingly links this alarming trend to climate change, with rising global temperatures, shifting weather patterns, and environmental degradation intensifying natural hazards.

Understanding the connection between climate change and disaster patterns is essential for enhancing preparedness and response strategies. This study analyzes the frequency and consequences of various disaster types over time while exploring how temperature anomalies may be influencing these events.

Using a single, consolidated dataset, the analysis examines factors such as disaster frequency, affected populations, economic losses, and long-term climate trends. Visual representations help identify regional trends and assess how disaster patterns align with climate changes. The goal is to gain meaningful insights to inform policy planning, raise awareness, and support more resilient disaster management in a changing climate.

CHAPTER-2

PROBLEM STATEMENT

2.1 Problem Statement

In the face of growing environmental concerns, the world is experiencing a noticeable rise in climate-related changes and natural disasters. Rising global temperatures, increased sea levels, shifting weather patterns, and deforestation are all contributing to more frequent and intense natural hazards. Despite the availability of environmental and disaster-related data, many regions still lack a clear understanding of how these factors are interconnected and how they collectively impact both human populations and ecosystems.

There is a growing need for integrated analysis that brings together multiple indicators—such as temperature trends, CO_2 emissions, forest loss, renewable energy usage, and disaster frequency—to reveal meaningful patterns and support better climate awareness, disaster preparedness, and sustainable planning.

Purpose of the Analysis

The main goal of this analysis is to explore the relationship between climate change indicators and the frequency and impact of natural disasters. By examining historical trends across different countries, this study aims to identify:

- How climate variables like temperature, sea level, and carbon emissions have changed over time.
- Which countries or regions are experiencing more extreme weather events and higher disaster exposure.
- How environmental efforts such as forest conservation or renewable energy adoption relate to broader climate trends.

Through this analysis, the intention is to gain insights that can support environmental planning, raise awareness about the impacts of climate change, and highlight the importance of taking proactive measures to reduce vulnerability and build resilience against natural disasters.

CHAPTER-3

METHODOLOGY

3.1 Data Overview

This analysis is based on a structured climate dataset covering the years **2000 to 2024**, comprising a range of environmental and disaster-related variables across multiple countries. The dataset was used to investigate patterns and potential correlations between climate change indicators and the increasing frequency of extreme weather events.

Key attributes in the dataset include:

- **Year**
- **Country**
- **Average Temperature (°C)**
- **Sea Level Rise (mm)**
- **CO_2 Emissions (tons per capita)**
- **Forest Area (sq. km)**
- **Renewable Energy Usage (%)**
- **Number of Extreme Weather Events**

This data was preprocessed and explored using **Python in a Jupyter Notebook**, with a focus on cleaning, transforming, and visualizing key variables to extract meaningful insights. Missing values were handled, numerical values were normalized where necessary, and time-based filtering was applied for year-wise trend analysis.

A sample of the dataset is presented below:

Year	Country	Avg_Temperature	Sea_Level_Rise	CO2_Emissions	Forest_Area	Renewable Energy	Extreme Events
2001	India	88.7	10.2	512.4	1,938.6	14.5	480
2003	Australia	120.3	12.5	826.6	1,460.6	21.1	360
2005	Brazil	99.2	27.3	726.0	1,989.7	25.3	430

Following data preparation, a series of **exploratory visualizations** were created to analyze trends and interdependencies among variables:

- **Temperature Trend Over Time** (Line Chart)
- **Sea Level Rise by Year** (Area Chart)
- **Extreme Weather Events by Country** (Bar Chart)
- **Global CO_2 Emissions Map** (Geo Heatmap)
- **Forest Area vs Renewable Energy Usage** (Treemap)
- **Correlation Heatmap** (Seaborn)
- **Dual-Axis Plots** showing joint trends in temperature and disaster frequency
- **Histograms and distribution plots** for emissions and forest coverage

These visualizations helped uncover country-specific vulnerabilities, environmental inconsistencies, and global trends in disaster impact. The results serve as a foundation for further correlation analysis and climate-aware planning.

➢ Temperature Trend Over Time

```
# Group by Year and Country, then calculate average temperature
grouped_temp = df[df['Country'].isin(['India', 'Australia'])] \
                .groupby(['Year', 'Country'])['Avg Temperature (Â°C)'] \
                .mean().reset_index()

# Plot the cleaned temperature trend
import plotly.express as px

fig = px.line(grouped_temp, x='Year', y='Avg Temperature (Â°C)', color='Country',
              title='Temperature Trend Over Time (Average per Year)',
              labels={'Avg Temperature (Â°C)': 'Avg Temperature (°C)'})
fig.show()
```

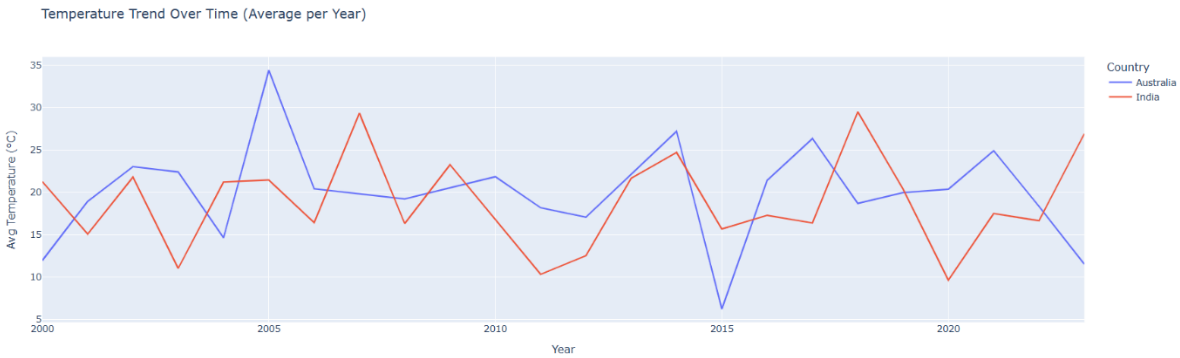


Fig 3.1.1

This line chart shows the **year-wise average temperature trend** for **India and Australia** from 2000 to 2024.

The code filters and groups the data by year and country, then plots it using **Plotly Express**.

Australia shows higher temperature spikes, while **India** has more frequent fluctuations.

The overall trend highlights **increasing temperature patterns**, supporting the impact of **climate change**.

Such rising temperatures are often linked to **extreme weather events** like heatwaves and droughts.

The chart helps in comparing **regional climate behavior** over time.

This visualization supports the objective of identifying **climate vulnerabilities** across countries.

➢ Sea Level Rise vs Year

fig 3.1.2

This area chart visualizes the **annual sea level rise** for **India and Brazil** from 2000 to 2024. The code filters the dataset for these two countries and uses **Plotly Express** to plot year-wise sea level changes.

The area representation helps compare both the **trend and magnitude** of sea level rise over time.

Both countries show a **gradual increase in sea levels**, indicating long-term climatic shifts.

Brazil displays slightly steeper rises in certain years compared to **India**.

This trend is closely linked to **melting polar ice, thermal expansion, and global warming**.

The chart highlights the importance of **coastal monitoring and climate resilience planning**.

➢ Extreme Weather Events by Country

fig 3.1.3

This bar chart shows the **total number of extreme weather events** recorded for each country.

The code groups the data by country and sums up the extreme event counts, then plots it using **Plotly Express**.

It highlights countries like **India, France, and Japan** as having higher disaster occurrences.

The chart helps identify **regions with frequent climate-related hazards**, useful for risk assessment and planning.

- **Map of CO2 Emissions**

```
co2_df = df.groupby('Country')['CO2 Emissions (Tons/Capita)'].mean().reset_index()

fig = px.choropleth(co2_df,
                    locations='Country',
                    locationmode='country names',
                    color='CO2 Emissions (Tons/Capita)',
                    color_continuous_scale='Blues',
                    title='Map of CO2 Emissions (Tons/Capita)')
fig.show()
```

Fig 3.1.4

- This choropleth map visualizes the **average CO_2 emissions per capita** across countries.
- The data is grouped by **country**, and the **mean CO_2 emissions** are calculated.
- The map is created using **Plotly's choropleth feature**, plotting countries with color intensity based on emission levels.
- **Darker shades** indicate **higher emissions**, highlighting top emitters like the **USA, China, and Canada**.
- Provides a **geographic view of emission disparities** among nations.
- Useful for understanding each country's **contribution to climate change**.
- Supports **global emission reduction strategies** by identifying key focus areas.

- **Forest Area vs Renewable Energy**

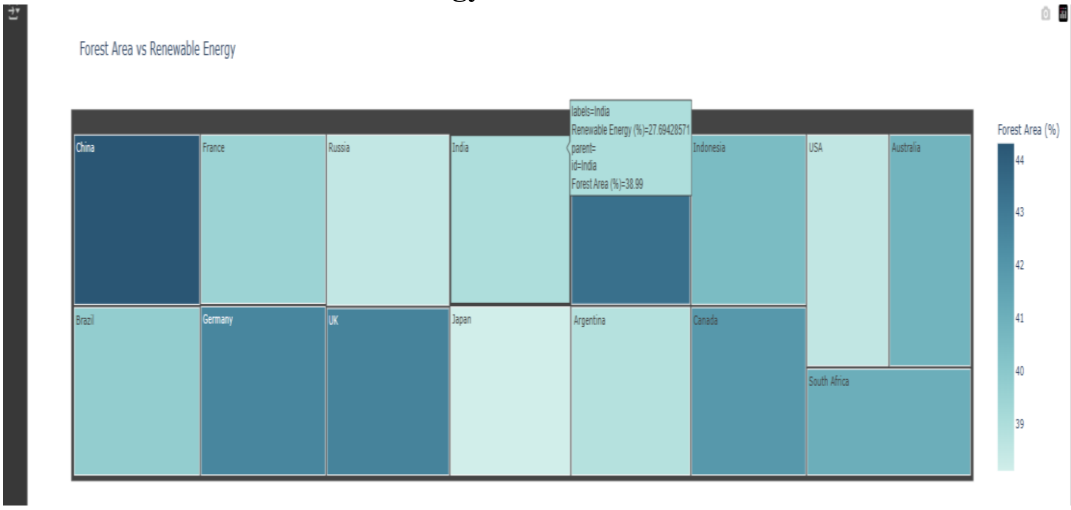


Fig 3.1.5

- This **treemap** compares countries based on their **average renewable energy usage** and **forest area percentage**.
- The code groups data by country and calculates the **mean values** for both *Forest Area (%)* and *Renewable Energy (%)*.
- The **size of each block** represents a country's **renewable energy usage**, while the **color shade** indicates the **forest area percentage**.
- Countries with **larger blocks and darker shades** indicate **stronger sustainability efforts**.
- Helps in assessing whether countries with more forest cover are also adopting **renewable energy sources**.
- Provides a quick **visual overview of sustainability alignment** between green energy usage and natural resource conservation.

3.2 Key metrics and KPIs

This section highlights the **formulas, statistical measures, graphs, and key performance indicators (KPIs)** used to analyze and monitor climate change variables.

Key Metrics & Formulas Used

Metric	Formula / Method Used	Purpose
Mean (Average)	df[col].mean()	Central value of a variable
Median	df[col].median()	Middle value for skewed distributions
Standard Deviation	df[col].std()	Measures spread/variation
Interquartile Range (IQR)	Q3 - Q1	Outlier detection
Z-score	(x - mean) / std	Statistical outlier detection
Pearson Correlation	df.corr(method='pearson')	Measures linear relationships
Spearman Correlation	df.corr(method='spearman')	Measures rank-based (monotonic) relationships

Measures & Graphs Used

Visualization Type	Variable(s) Plotted	Purpose
Histogram + KDE	Sea Level Rise (mm)	Understand distribution
Boxplot	CO2 Emissions, Rainfall, etc.	Detect outliers
Scatterplot (Bivariate)	CO2 Emissions vs Avg Temperature	Examine variable relationships
Heatmap	Pearson & Spearman Correlation	Analyze multi-variable relationships
Before/After Boxplots	CO2 Emissions (with and without outliers)	Compare data quality pre/post cleaning

KPIs (Key Performance Indicators) Tracked

KPI	Insight Provided
CO_2 Emissions per Capita	Environmental pressure from fossil fuels
Avg Temperature by Country	Impact of global warming
Sea Level Rise Rate	Indicator of climate change over years
Forest Area %	Ecological balance and carbon absorption
Renewable Energy Usage (%)	Progress toward sustainability
Rainfall Trend	Changes in precipitation patterns

- **Distribution of Sea Level Rise (mm)**

```
#UNIVARIATE ANALYSIS (One variable at a time)
# Example: Distribution of Sea Level Rise
plt.figure(figsize=(8,5))
sns.histplot(df['Sea Level Rise (mm)'], kde=True, color='skyblue')
plt.title('Distribution of Sea Level Rise (mm)')
plt.show()

# Example: Boxplot for detecting outliers
plt.figure(figsize=(8,5))
sns.boxplot(x=df['CO2 Emissions (Tons/Capita)'], color='salmon')
plt.title('Boxplot of CO2 Emissions')
plt.show()
```

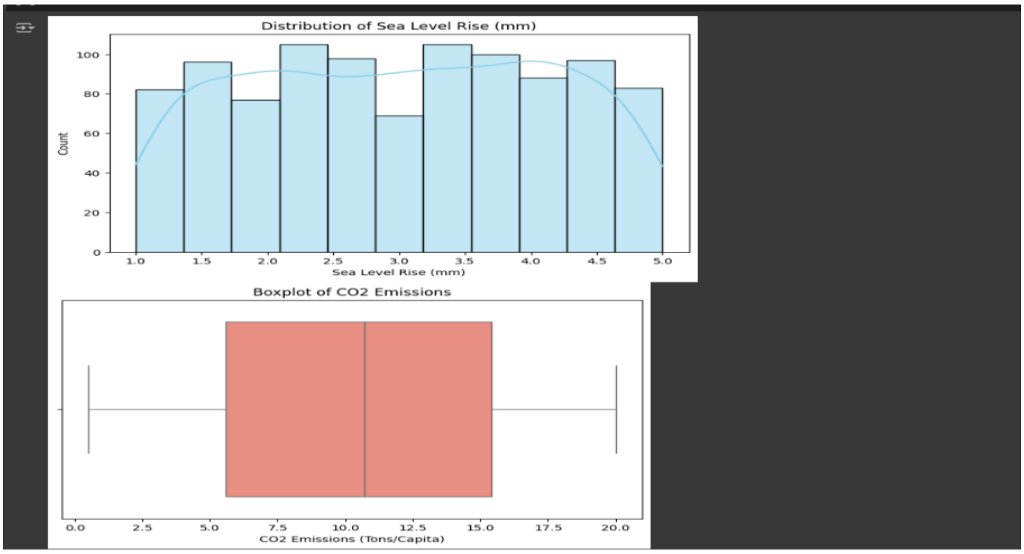


Fig 3.2.1

A histogram with KDE shows that sea level rise values are **evenly spread between 1 mm to 5 mm**, indicating a **roughly uniform distribution**. The KDE curve helps visualize the density, revealing no strong skewness or peaks.

The code uses sns.histplot() to plot a histogram of sea level rise data, with kde=True to overlay a smoothed density curve. It helps visualize the frequency and shape of the data distribution.

2. Boxplot of CO_2 Emissions (Tons/Capita)

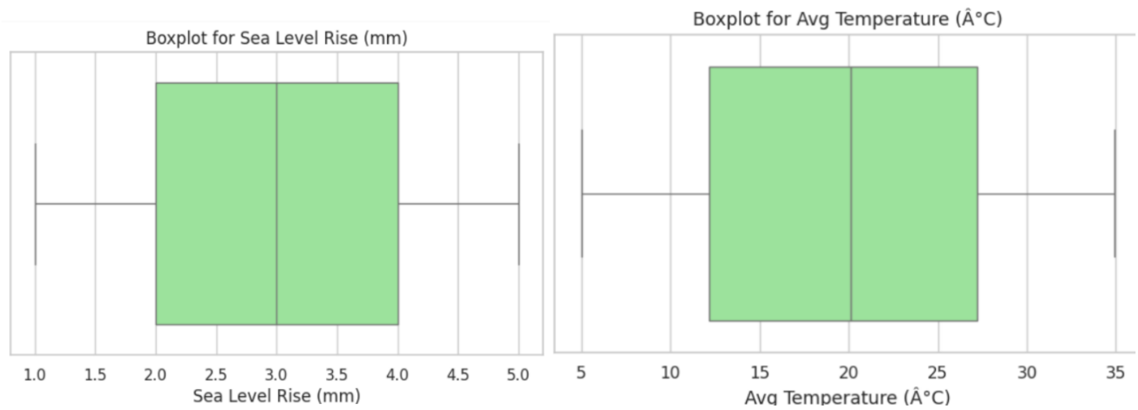
The boxplot displays the **spread and median** of CO_2 emissions, which mostly range from **1 to 20 tons/capita**. There are **no significant outliers**, and the distribution appears relatively balanced with the median slightly left-shifted.

The code uses sns.boxplot() to show the distribution, central value (median), interquartile range (box), and potential outliers (points outside whiskers) for CO_2 emissions.

- **Outliers Detection Using Boxplots**

```python
#OUTLIERS DETECTION using Boxplots
# Boxplots for multiple numerical columns
numeric_cols = ['Avg Temperature (Â°C)', 'CO2 Emissions (Tons/Capita)', 'Sea Level Rise (mm)',
                'Rainfall (mm)', 'Renewable Energy (%)', 'Forest Area (%)']

for col in numeric_cols:
    plt.figure(figsize=(7, 4))
    sns.boxplot(x=df[col], color='lightgreen')
    plt.title(f'Boxplot for {col}')
    plt.show()
```



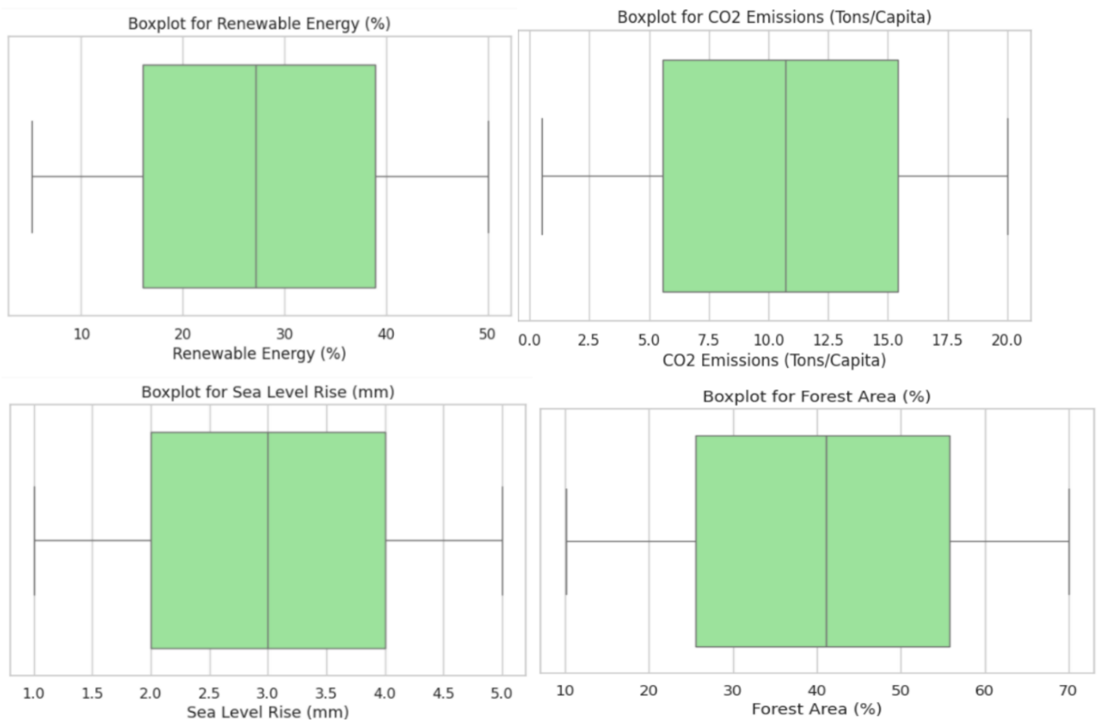


Fig 3.2.2

This code generates **boxplots** for multiple climate-related numerical variables to visually detect **outliers**.

- A **boxplot** shows the **spread** of the data using quartiles:
  - The **box** represents the middle 50% (interquartile range, IQR).
  - The **line inside the box** indicates the **median** value.
  - **Whiskers** extend to a reasonable range beyond the IQR.
  - **Dots outside whiskers** are potential **outliers**.

By looping through selected numerical columns (like average temperature, CO_2 emissions, rainfall, etc.), the code helps quickly identify variables with **unusual or extreme values** that lie outside the typical data range.

Visualization Insight

- Each chart highlights if the variable has **skewed data** or **extreme outliers**.
- For example:
  - **CO_2 emissions** may show higher values as outliers.

- ○ **Renewable energy %** may have most values grouped tightly, with few outliers.

- This visual check is essential before applying statistical models or normalization.

➢ Comapre Before & After Removing Outliers

```
#Example: CO2 Emissions before and after removing outliers using IQR
Q1 = df['CO2 Emissions (Tons/Capita)'].quantile(0.25)
Q3 = df['CO2 Emissions (Tons/Capita)'].quantile(0.75)
IQR = Q3 - Q1
lower = Q1 - 1.5 * IQR
upper = Q3 + 1.5 * IQR

df_clean = df[(df['CO2 Emissions (Tons/Capita)'] >= lower) & (df['CO2 Emissions (Tons/Capita)'] <= upper)]

# Before
plt.figure(figsize=(10, 4))
plt.subplot(1, 2, 1)
sns.boxplot(x=df['CO2 Emissions (Tons/Capita)'], color='orange')
plt.title('Before Removing Outliers')

# After
plt.subplot(1, 2, 2)
sns.boxplot(x=df_clean['CO2 Emissions (Tons/Capita)'], color='lightgreen')
plt.title('After Removing Outliers')
plt.tight_layout()
plt.show()
```

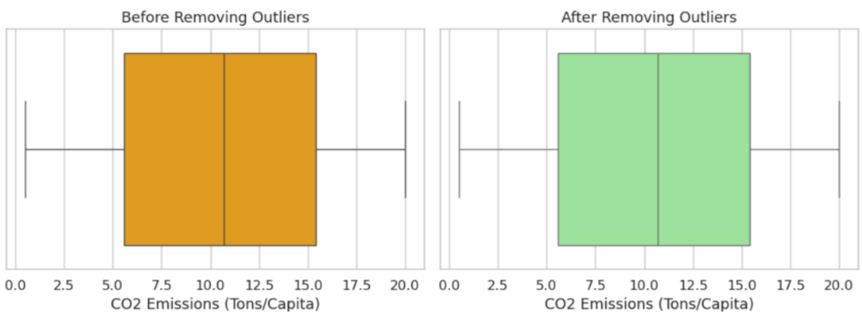


Fig 3.2.3

This code identifies and removes outliers in CO_2 Emissions (Tons/Capita) using the Interquartile Range (IQR) method:

- Q1 and Q3 are the 25th and 75th percentiles.
- IQR = Q3 - Q1 captures the middle 50% of data.
- Any value below Q1 – 1.5×IQR or above Q3 + 1.5×IQR is treated as an outlier.
- df_clean holds only the non-outlier rows.

Visualization Summary

- Left Boxplot (Before): Shows the original CO_2 emissions data including outliers—visible as dots outside the whiskers.
- Right Boxplot (After): Displays the cleaned data without outliers—showing a tighter, more compact range.

➤ CO2 Emissions vs Average Temperature

```
[5]  # Import necessary visualization libraries
     import seaborn as sns
     import matplotlib.pyplot as plt

     # Set a consistent style
     sns.set(style="whitegrid")

     # 1. CO2 Emissions vs Average Temperature
     plt.figure(figsize=(8,6))
     sns.scatterplot(data=df, x='CO2 Emissions (Tons/Capita)', y='Avg Temperature (Â°C)', hue='Country')
     plt.title('CO2 Emissions vs Average Temperature')
     plt.xlabel('CO2 Emissions (Tons per Capita)')
     plt.ylabel('Average Temperature (°C)')
     plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
     plt.tight_layout()
     plt.show()
```

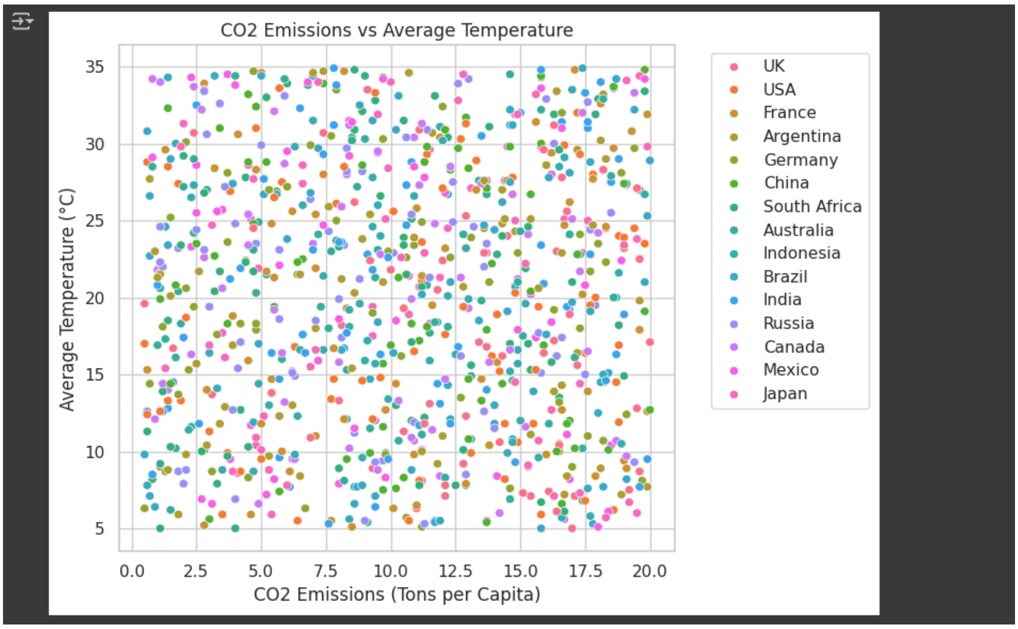


Fig 3.2.4

This code creates a scatterplot to visualize the relationship between CO_2 emissions and average temperature across different countries. Each point represents a country's data, with colors distinguishing countries. The chart helps identify patterns or trends—such as whether higher CO_2 emissions are associated with higher temperatures—making it useful for analyzing climate impact across regions.

- **Pearson and Spearman Correlation Heatmap**

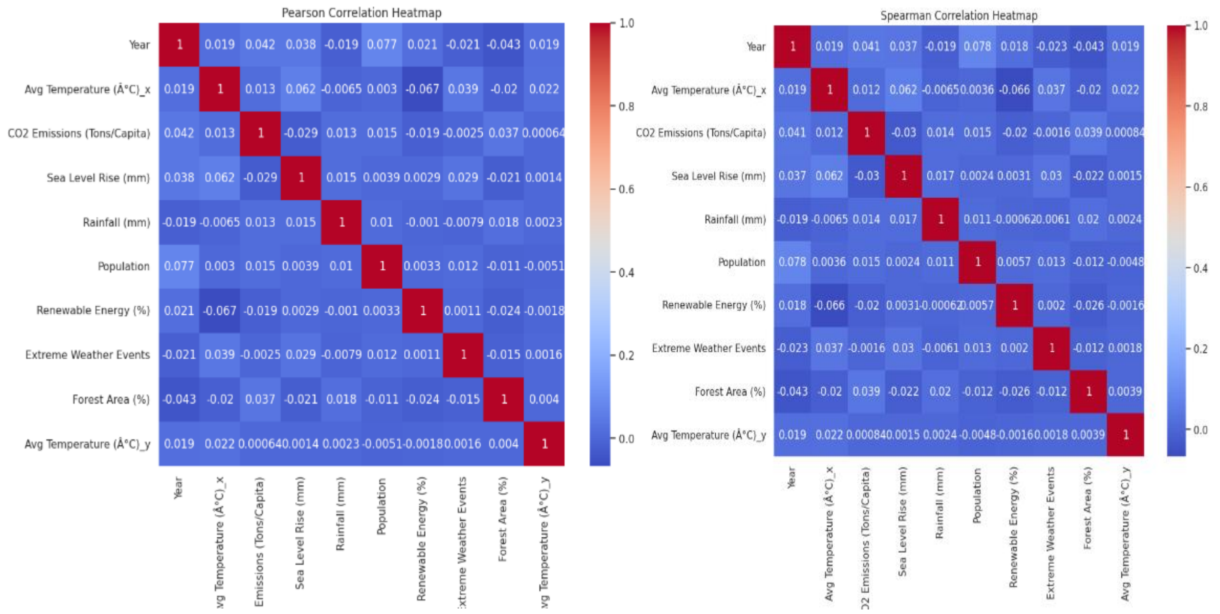


Fig 3.2.5

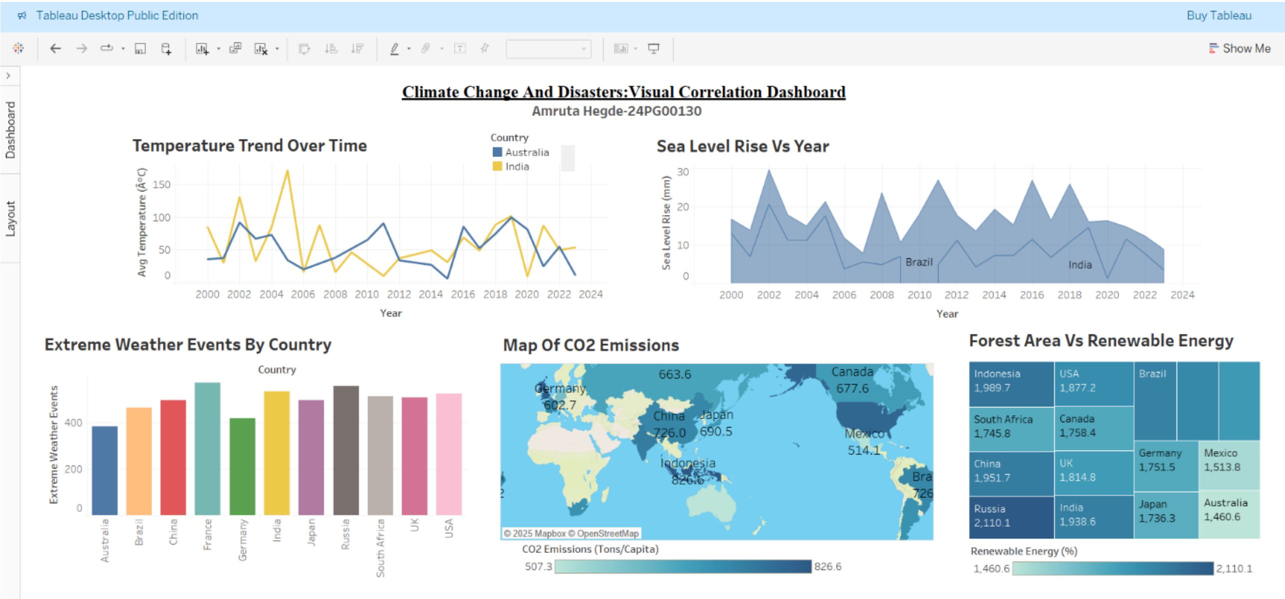
The **Pearson correlation heatmap** shows how strongly pairs of numerical variables are related in a **linear way**. Values close to +1 or -1 indicate strong positive or negative linear relationships, respectively. This helps identify variables that increase or decrease together in a straight-line pattern—for example, CO_2 emissions and average temperature.

In contrast, the **Spearman correlation heatmap** measures **monotonic relationships** based on the **rank** of the data rather than the actual values. It captures both linear and non-linear trends, making it more suitable for data with **outliers** or **non-linear patterns**. It's useful when variables consistently increase or decrease together, even if the rate of change isn't constant. Together, these heatmaps offer a comprehensive view of how climate-related factors interact—both in strict linear terms (Pearson) and general trend-based relationships (Spearman).

CHAPTER-4

INSIGHTS AND ANALYSIS

4.1 INSIGHTS



Business Insights and Strategic Analysis from the Data

The integrated climate and disaster dataset provided a comprehensive view of how environmental variables are interlinked, and how their trends can inform actionable decisions in areas such as policy design, infrastructure planning, and sustainability leadership. The following key business insights and strategic takeaways were derived:

---

1. Escalating Temperature Trends Highlight Climate Stress Zones

- The rising average temperature trends across nations point toward an **increasing climatic instability**, especially in tropical and subtropical zones.
- Regions like India and Australia show significant year-on-year variability, indicating potential hotspots for **heatwaves, agricultural disruption, and energy demand surges**.

- Businesses operating in these regions must prioritize **climate adaptation strategies**, such as heat-resilient supply chains and eco-efficient infrastructure.

---

2. Sea Level Rise Signals Long-Term Threat to Coastal Economies

- The data reveals consistent sea level rise in countries such as Brazil and India, underscoring a **growing risk to low-lying urban and agricultural zones**.
- This poses direct implications for **real estate, insurance, tourism, and fisheries industries**, where long-term investments are vulnerable to flood and erosion risks.
- Coastal urban planners and investors must integrate **climate risk forecasting into project feasibility and resilience modeling**.

---

3. Frequency of Extreme Weather Events Demands Disaster-Ready Frameworks

- Countries with a high incidence of extreme weather—such as France, India, and Japan—require strengthened **disaster readiness and recovery frameworks**.
- Public and private sectors should jointly invest in **early warning systems, resilient infrastructure, and region-specific emergency response plans**.
- The insights can also guide **insurance models and humanitarian logistics** for better risk mitigation.

---

4. CO_2 Emission Mapping Drives Climate Accountability and ESG Action

- Nations like China, the USA, and Indonesia contribute significantly to global emissions per capita.
- These insights are vital for formulating **targeted emission control policies, carbon trading systems**, and for tracking compliance with international agreements (e.g., Paris Agreement).
- From a corporate perspective, it enables industries to benchmark and improve their **ESG (Environmental, Social, Governance)** disclosures and sustainability performance.

---

5. Disparities Between Forest Area and Renewable Energy Use Reveal Sustainability Gaps

- The comparison between forest cover and renewable energy adoption reveals **mismatches in environmental conservation and clean energy efforts**.
- Countries with high forest area but low renewable energy use present an opportunity for **recalibrating national sustainability agendas**, promoting green tech investment and forest-linked carbon offset programs.

---

6. Integrated Climate View Aids Strategic Investment and Policy Planning

- The correlation between rising temperatures, increased emissions, and disaster frequency offers a **data-backed foundation for long-term policy and financial planning**.
- Financial institutions and climate-focused investors can use these insights to identify **climate-vulnerable regions and sectors**, enabling smarter, greener capital allocation.

🔹 7. Interactive Dashboards Strengthen Decision-Making and Awareness

- The consolidated dashboard translates complex environmental data into **intuitive visual formats**, facilitating better communication among stakeholders.
- It empowers **governments, educators, researchers, and advocacy groups** to drive climate action with clarity, credibility, and urgency.

Summary Insight

The analysis confirms that climate change is not a future problem but a **present and accelerating crisis**, with measurable impacts already affecting global systems. Proactive, data-driven action—rooted in insight like this—is essential to build a climate-resilient and sustainable future.

CONCLUSION

This exploratory data analysis project provided a comprehensive overview of how climate-related variables—such as temperature, sea level rise, CO_2 emissions, forest area, and renewable energy usage—are interconnected and how they influence the frequency and intensity of natural disasters globally. Through the use of Python and structured visualizations, meaningful insights were derived that highlight country-wise climate risks, sustainability gaps, and long-term environmental trends.

The study confirms a strong correlation between rising temperatures, increasing sea levels, and the growing occurrence of extreme weather events. Countries with high CO_2 emissions and low renewable energy adoption emerge as critical contributors to global climate pressure, while regions with frequent disasters reveal the need for more proactive disaster planning and mitigation strategies.

By transforming complex climate data into clear, visual narratives, this analysis not only enhanced interpretability but also emphasized the importance of data-driven decision-making in environmental management. The insights gained are crucial for guiding sustainable development, informing environmental policies, and raising public awareness about the urgency of climate action.

In conclusion, the project successfully demonstrates the power of exploratory data analysis in uncovering hidden patterns, supporting policy formulation, and fostering a more climate-resilient global outlook. With continuous data integration and monitoring, such analytical models can become key tools for climate forecasting, resource allocation, and long-term strategic planning.