INTRODUCTION TO DATA SCIENCE
MILESTONE - 1
REPORT

TABLE OF CONTENTS

S.NO.	STEPS INVOLVED	PAGE NO
1.	INTRODUCTION	3
	1.1 OBJECTIVE	3
	1.2 TYPE OF TOOL	3
	1.3 DATA SOURCES	3
	1.4 TECHNOLOGY STACK	3
2	PROJECT TIMELINE	4
3	DATA COLLECTION	4
4	DATA PREPROCESSING	7
	4.1 HANDLING MISSING DATA	7
	4.2 DATA TYPE CONVERSION	7
	4.3 OUTLIER DETECTION	7
	4.4 NORMALIZATION	7
5	EXPLORATORY DATA ANALYSIS	8
	5.1 CORRELATION MATRICES	8
	5.2 VISUALIZATIONS - GLOBAL RECORD TEMPERATURES	11
	5.3 VISUALIZATIONS - CO2 EMISSIONS	13
	5.4 VISUALIZATIONS - PRECIPITATION	15
	5.5 VISUALIZATIONS - GREEN COVER	17
	5.6 VISUALIZATIONS - NATURAL DISASTER AND EMERGENCY EVENTS	20
6	CONCLUSION	22

1. INTRODUCTION

1.1 OBJECTIVE

This project aims to develop an advanced recommender system that predicts climate temperature trends and assesses potential natural disasters based on the user-specified year and location (country). By leveraging historical data and predictive modeling, the system will provide valuable insights into environmental risks, enabling users to make informed decisions. This tool will be a crucial resource for climate analysis, disaster preparedness, and risk mitigation strategies.

1.2. TYPE OF TOOL

This project will develop a **Climate and Disaster Prediction Recommender System**, which will

- Recommends expected **temperature variations** according to time and location.
- Predicts the likelihood of **natural disasters** based on past trends.
- Deliver insights through an **interactive interface**.

1.3. DATA SOURCES

1. Global Temperature Records(1850-2013) - kaggle
2. Per capita Co2 emissions - Ourworldindata
3. Annual precipitation, 1940 to 2024 - Ourworldindata
4. Deforestation and Forest Loss - Ourworldindata
5. Natural Disasters Emergency Events Database & Country Profiles - Omdena

1.4. TECHNOLOGY STACK

Programming Language: Python

Libraries and Frameworks

- **Data Manipulation:** Pandas, NumPy
- **Visualization:** Matplotlib, Seaborn, Plotly
- **Machine Learning & Prediction:** Scikit-Learn, TensorFlow
- **Geospatial Analysis:** Geopandas, Folium
- **Report Generation:** Google docs

2. PROJECT TIMELINE

s.no	Task	Deadline
1.	Data collection and preprocessing	23rd February 2025
2.	Exploratory Data Analysis	23rd February 2025
3.	Feature Engineering and Feature selection	11th March 2025
4.	Data Modeling	21st March 2025
5.	Evaluation and testing	15th April 2025
6.	Interpreting and visualising results	23rd April 2025

3. DATA COLLECTION

For this climate prediction model, 5 datasets are collected that includes the global temperature records, per capita co2 emissions, Annual Precipitation, Deforestation and Forest Loss, and Natural disasters and Emergency events data. This section clearly explains the details of the datasets.

a. Global Temperature Records (1850 - 2013)

This dataset appears to be a historical climate dataset that records temperature trends across various cities worldwide. Based on the columns, this dataset provides detailed insights into global temperature changes over time. It has 239177 rows and 7 columns.

This dataset is collected from the kaggle.

Columns includes:

dt : This column indicates the date of the record temperature which is in YYYY-MM-DD format.

Average Temperature : The average temperature for the particular city to the corresponding date.

Average Temperature Uncertainty : The uncertainty in the recorded temperature.

City : The name of the city where the temperature is recorded.

Country : The country corresponding to the city name.

Latitude : The latitude coordinates of the city with the directions North and South

Longitude : The longitude coordinates of the city with the directions East and west

b. Per capita Co2 emissions

The dataset appears to contain historical CO_2 emissions data for different countries over multiple years. This dataset is useful for analyzing carbon emissions trends and their impact on climate change. This dataset is from OurworldinData which has 26182 rows and 3 columns.

Columns can be described as follows

Entity : Represents the country where the Co2 is recorded.

Year : Indicates the year of Co2 emission to that particular location.

Annual Co2 emission : Specifies the annual CO_2 emissions per person in metric tons.

c. Precipitation

The dataset appears to contain historical precipitation data for different countries over multiple years. This dataset is useful for analyzing rainfall trends and their impact on climate change. This dataset is from OurworldinData which has 16575 rows and 4 columns.

Columns data can be illustrated as follows:

Entity : Represents the country where the Co2 is recorded.

Code : This specifies the three letter ISO code for the country.

Year : Indicates the year of Co2 emission to that particular location.

Annual precipitation : This is the annual precipitation of that particular area represented in mm.

d. Deforestation and Forest Loss

The dataset keeps track of the annual net change in forest area for different countries over various years. It provides insights into deforestation and afforestation trends across different regions. It has 475 rows and 4 columns.

Columns consists of the data that includes:

Entity : Represents the country where the change in green area is recorded.

Code : This specifies the three letter ISO code for the country.

Year : Indicates the year when the change of the forest area is detected to that particular location.

Annual net change in forest area : Represents the average annual change in green cover where negative value indicates deforestation and positive value represents afforestation.

e. Natural Disasters Emergency Events Database & Country Profiles

This dataset maintains the data over the various Natural disasters over the period of time in various locations all over the world. It also gives detailed insights over the affected areas and population during the particular disaster. This dataset has 10432 rows and, 13 columns.

The columns of this dataset includes

Year : This is the year when the disaster occurred.

Country : Location where the disaster took place.

ISO : It is the three letter ISO code for the corresponding country mentioned.

Disaster Group : The broader category of disaster (e.g., Natural).

Disaster Subgroup : A more specific classification within the disaster group (e.g., Climatological, Hydrological, Geophysical).

Disaster Type : The specific type of disaster (e.g., Drought, Flood, Volcanic Activity).

Disaster Subtype : A further classification within the disaster type (e.g., Ash Fall under Volcanic Activity).

Total Events : The number of times this disaster event occurred in that year.

Total Affected : The total number of individuals affected by the disaster.

Total Deaths : The total number of people who died due to the disaster.

Total Damage (original) : The estimated financial damage caused by the disaster in original USD values.

Total Damage (adjusted) : The estimated financial damage adjusted for inflation.

CPI (Consumer Price Index) : The Consumer Price Index value, which is useful for adjusting economic damages for inflation.

4. DATA PREPROCESSING

Data Preprocessing is the process of transforming the raw data into clean and qualified data for further analysis. This process involves handling missing data, transforming the data and generating the new features or selecting the existing ones that help in better prediction for the model.

The basic statistics and information is known using df.shape, df.dtypes describe_data(), count_nulls() for all the datasets used. All the duplicate rows are removed using remove_duplicates() method in all the datasets.

4.1 HANDLING MISSING DATA

Climate & Disaster Data: linear interpolation (interpolate(method='linear')) is used to fill missing values (NaN) in the AverageTemperature and AverageTemperatureUncertainty columns.Since temperature data follows a continuous pattern, linear interpolation gives reasonable estimates.

Precipitation & Forest_area : Here the null values are present in the ISO codes for the location. These are handled by parsing the pycountry library and filling the missing ISO codes for the corresponding country names in both the datasets.

Natural Disasters : There are null values present in year, Disaster subtype, total events, total affected, total deaths, total damage in original and adjusted and CPI. Year missing value row is dropped using dropna(). Remaining columns are filled with the median values. Still there were some missing values that were identified due to incorrect country names. Those are fixed using the pycountry library and filled the disaster subtype with disaster type values corresponding to it.

4.2 DATA TYPE CONVERSION

- Dt object is converted to the proper date-time format.
- Latitude and Longitude are converted to float values in order to perform mathematical operations or to plot them in a map.

4.3 OUTLIER DETECTION

- Z-Score Method: Data points with Z-scores greater than 3 or less than -3 were considered extreme outliers.
- Interquartile Range (IQR) Method: Values falling outside 1.5 times the IQR range were flagged as potential outliers.
- Removed the outliers from the dataframes in all the datasets.

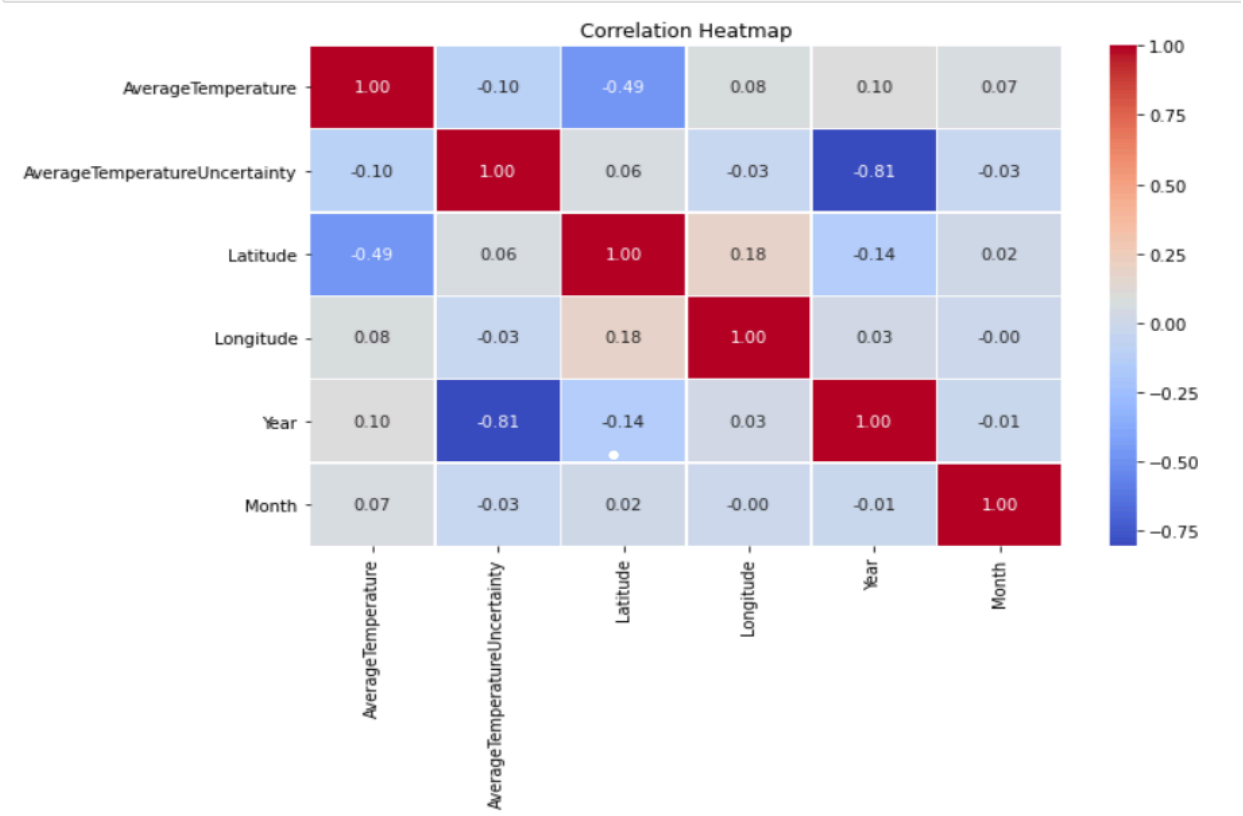
4.4 NORMALIZATION

- The features of all the datasets are normalized using z-score Normalization typically known as standardization using sklearn.preprocessing.StandardScaler.
- StandardScaler() is a preprocessing technique that transforms data to have a mean of 0 and standard deviation of 1 (Z-score normalization).
- It makes sure that numerical features have similar scales, improving model performance.

5. EXPLORATORY DATA ANALYSIS (EDA)

In this section, some of the visualizations are performed in order to understand and gain useful insights about the data. These visualizations are prepared using matplotlib and seaborn libraries.

5.1 CORRELATION MATRICES



These correlation matrices show the relationship between the features within the dataset.

Fig 1: Global Temperatures Dataset correlation matrix

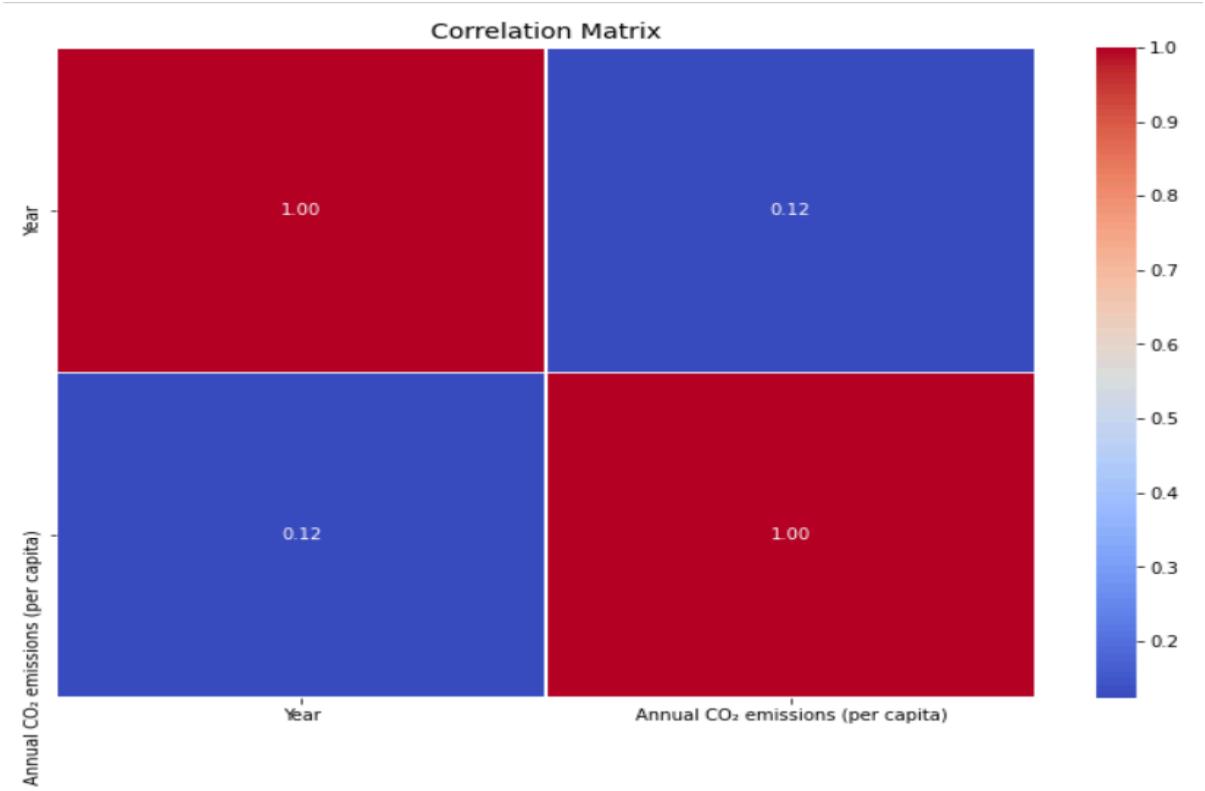


Fig.2 : Co2 emissions correlation matrix

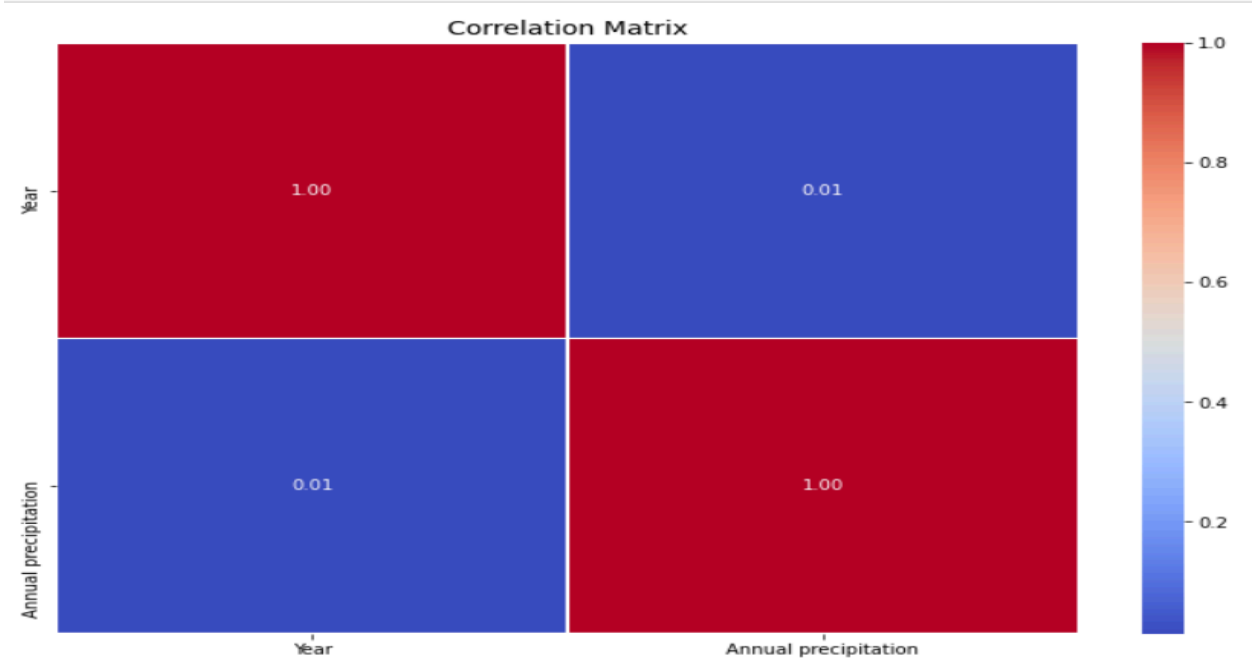


FIg 3: Annual Precipitation correlation matrix

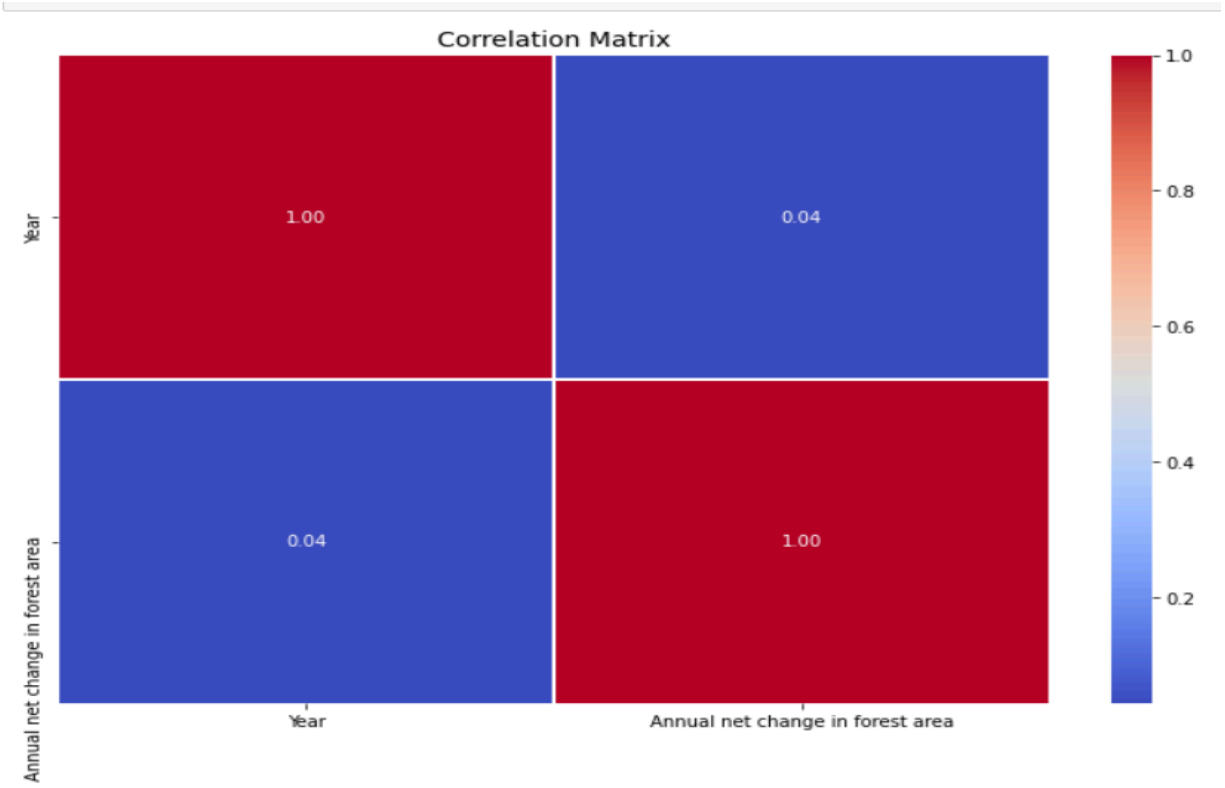


Fig. 4: Green cover correlation matrix

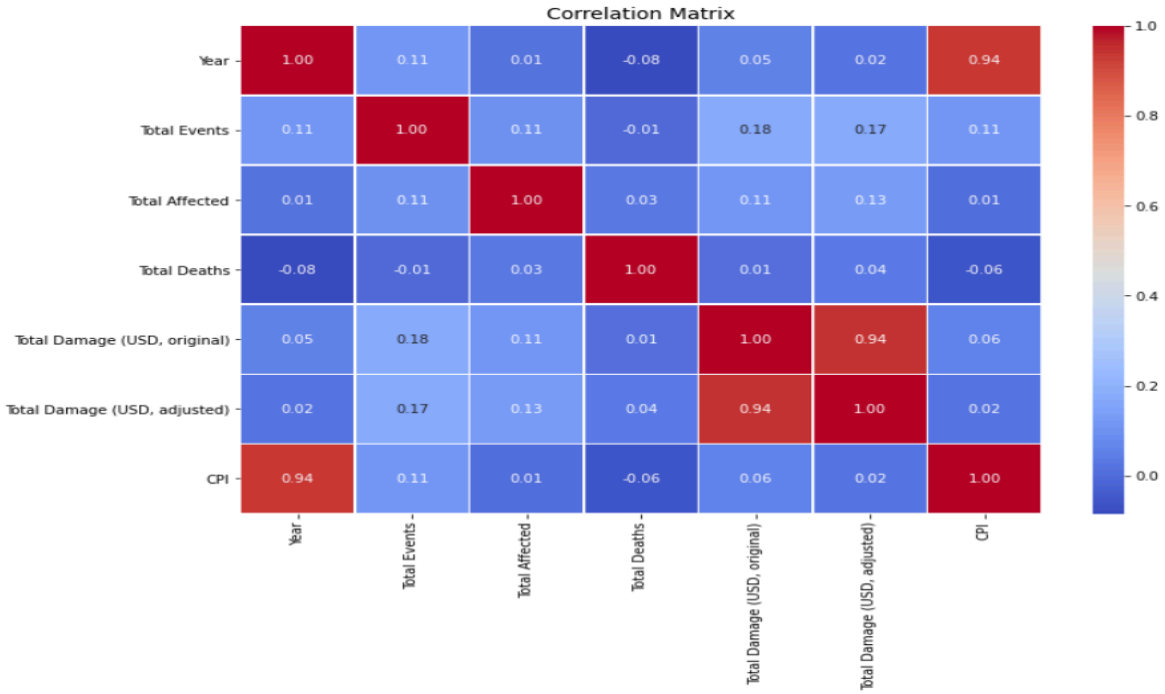


Fig. 5: Natural Disasters Correlation matrix

5.2 VISUALIZATIONS - GLOBAL RECORD TEMPERATURES

This visualization Displays the change in average temperature over time all over the various countries.

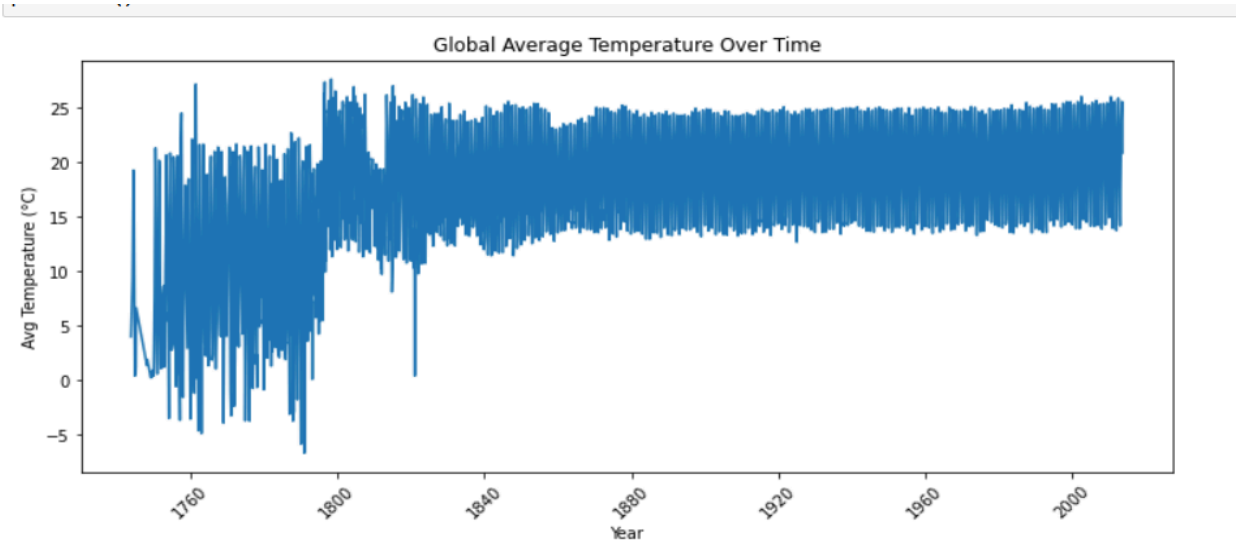


Fig.6 Global Average Temperature Over Time

This visualization provides key insights into the global distribution of average temperatures over time. The peak occurs between **20°C to 30°C**, indicating that most recorded temperatures fall in this range.

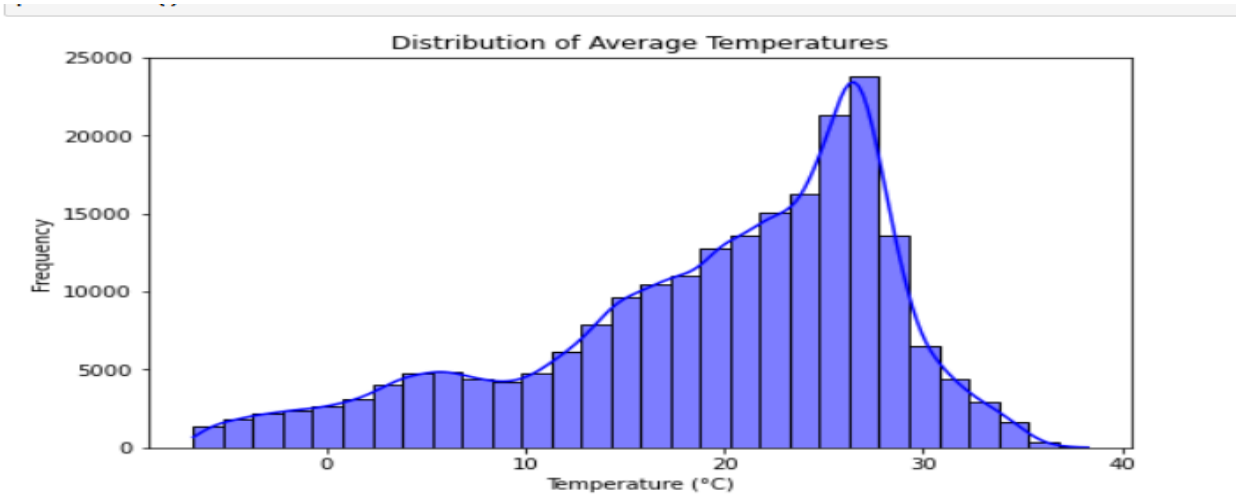


Fig.7. Histogram for distributions of temperatures

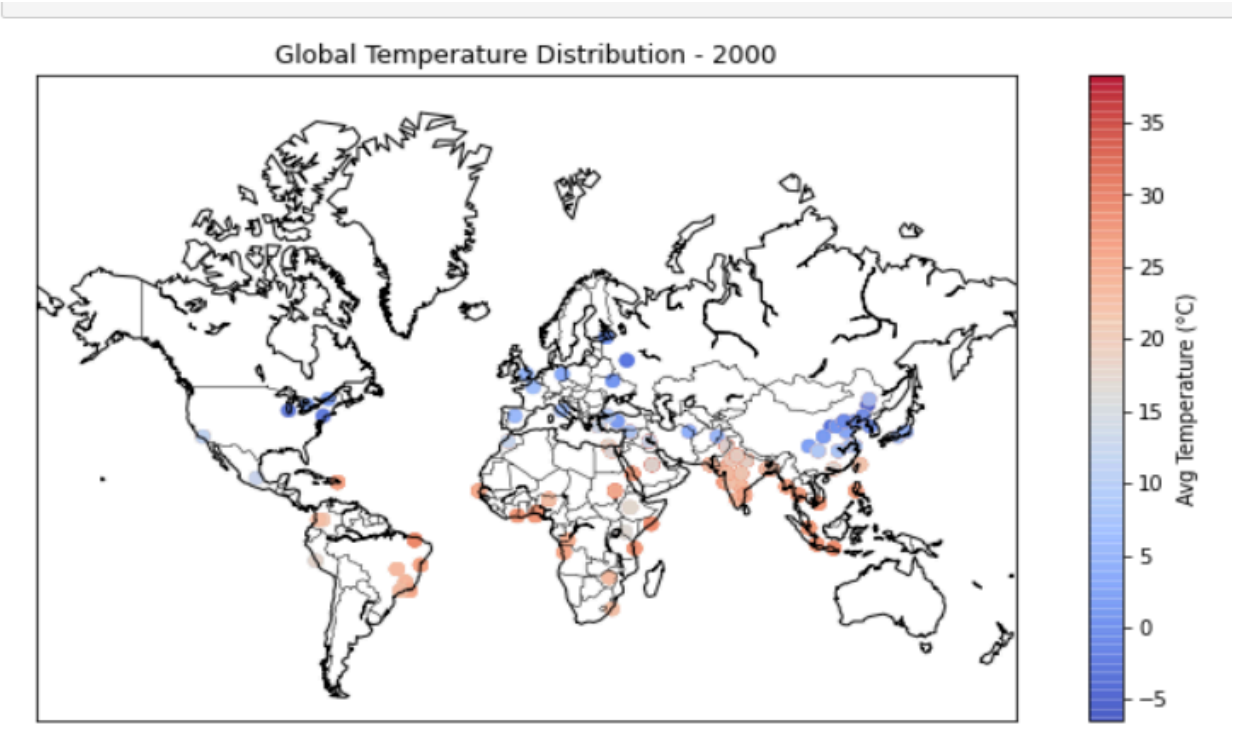


Fig.8. Global temperature distribution

By the following graphs, it can be considered that Iraq has had the highest temperatures and the US had the lowest temperatures over the time observed among the other countries.

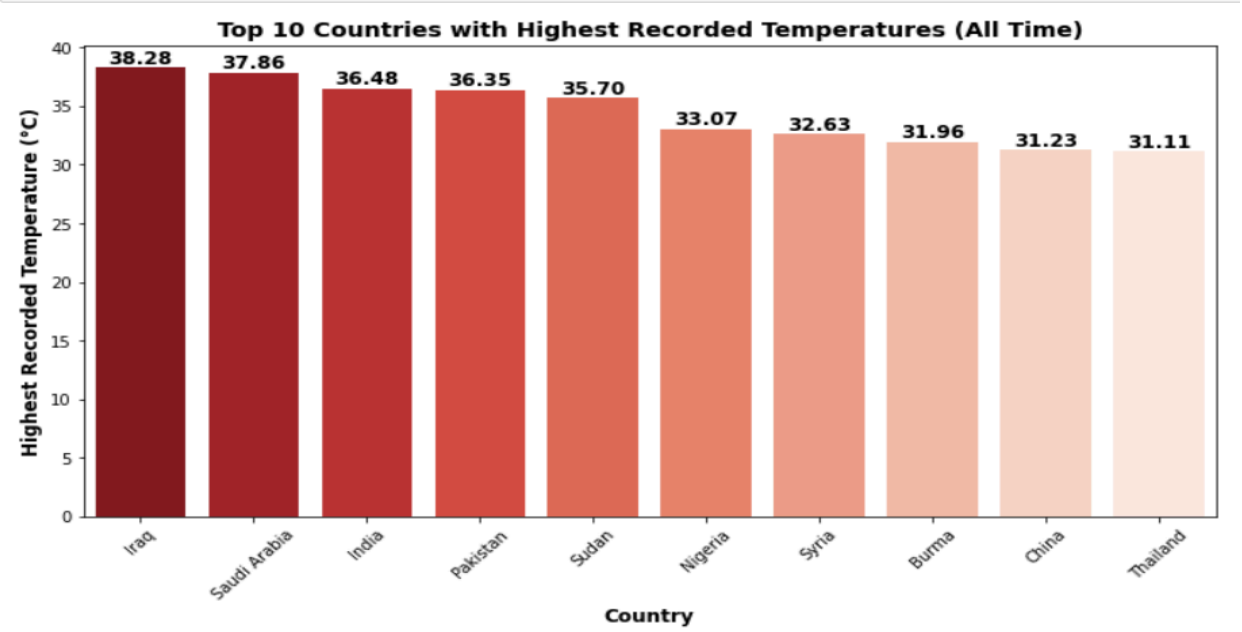


Fig.9. Barchart of countries with hottest temperatures

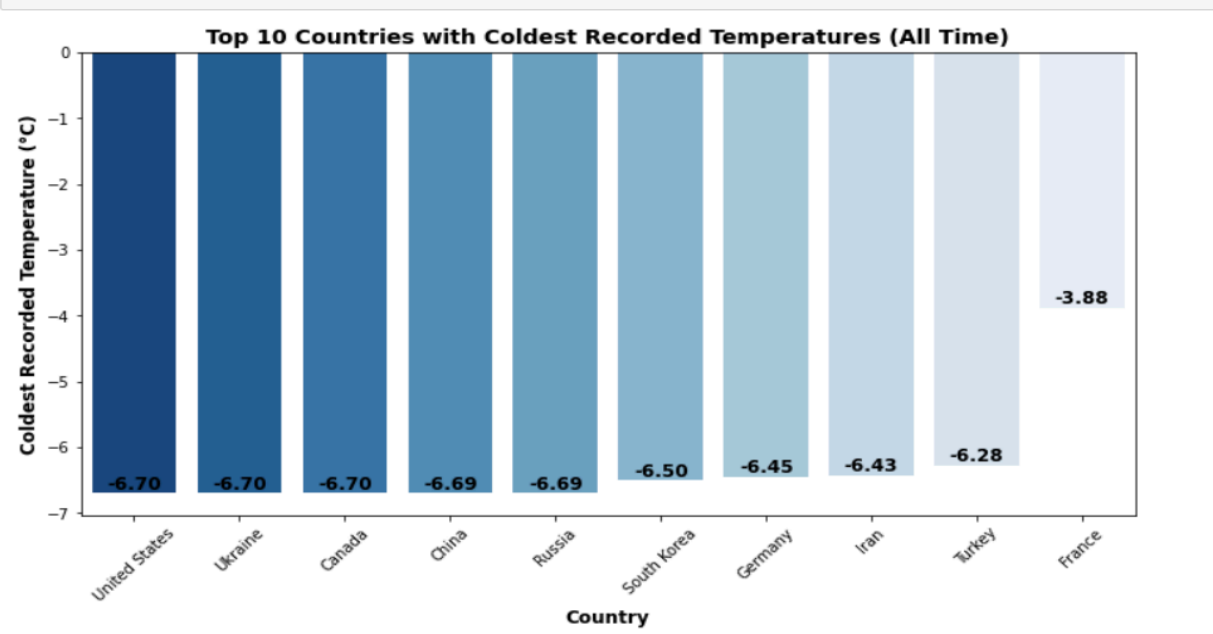


Fig.10. Barchart of countries with coldest temperatures

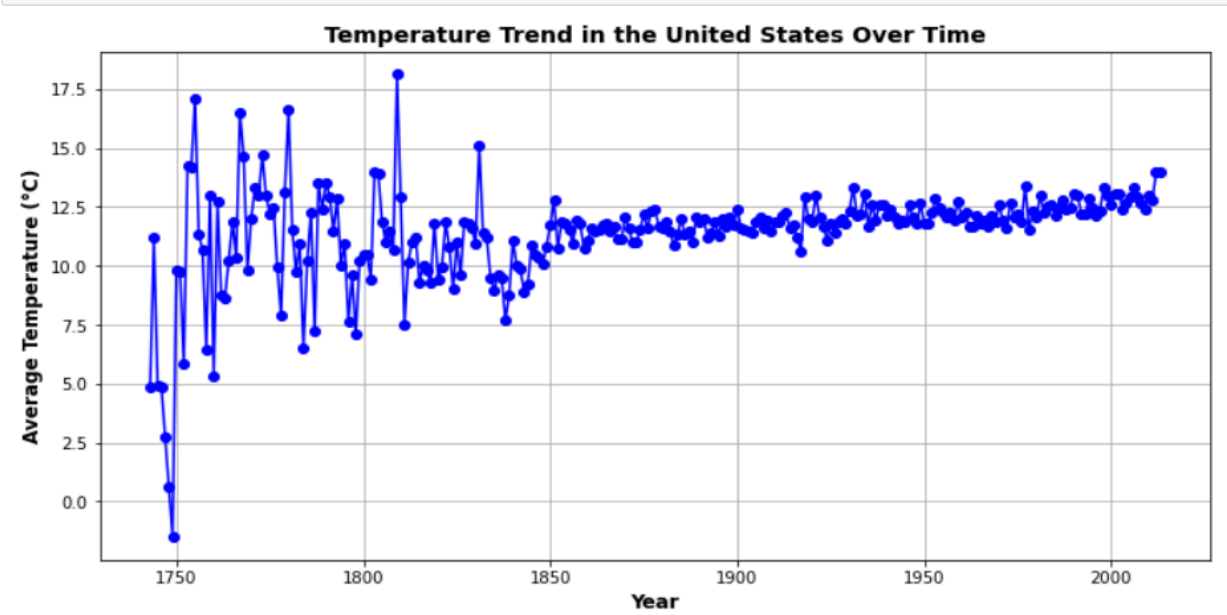


Fig.11. Time series trend of temperatures in US

5.3 VISUALIZATIONS - CO2 EMISSIONS

From the below diagram, it can be illustrated as co2 emissions were increasing over the time all around the world.

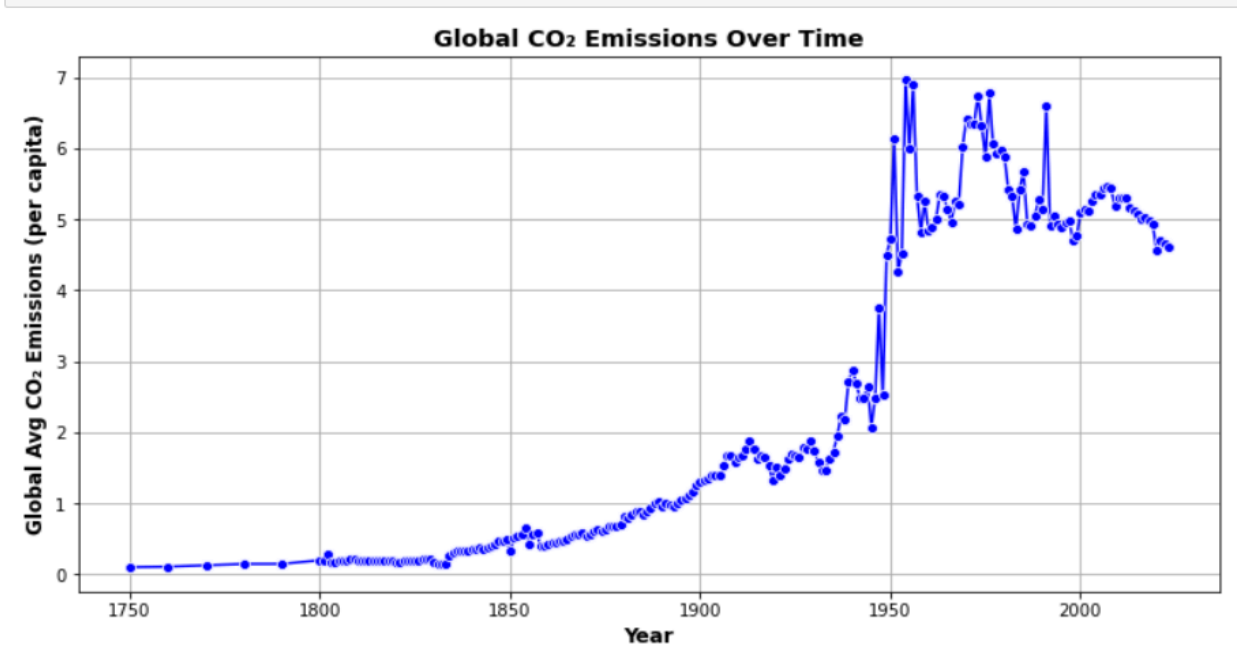


Fig.12. Time series trend of Co2 emission over the world

This diagram explains that dutch side countries are with highest co2 emissions and trinidad and tobago are with low co2 emissions and this level is increasing all over the world.

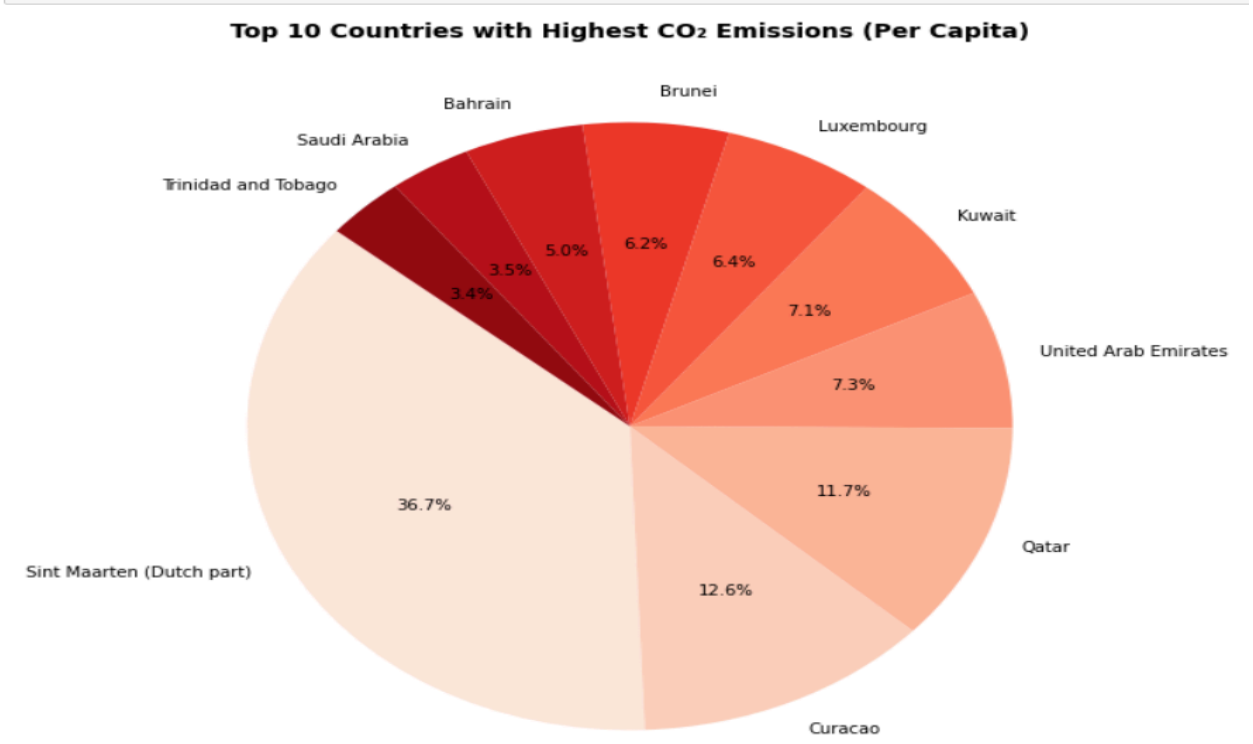


Fig.13. Pie chart of co2 emissions of top 10 countries

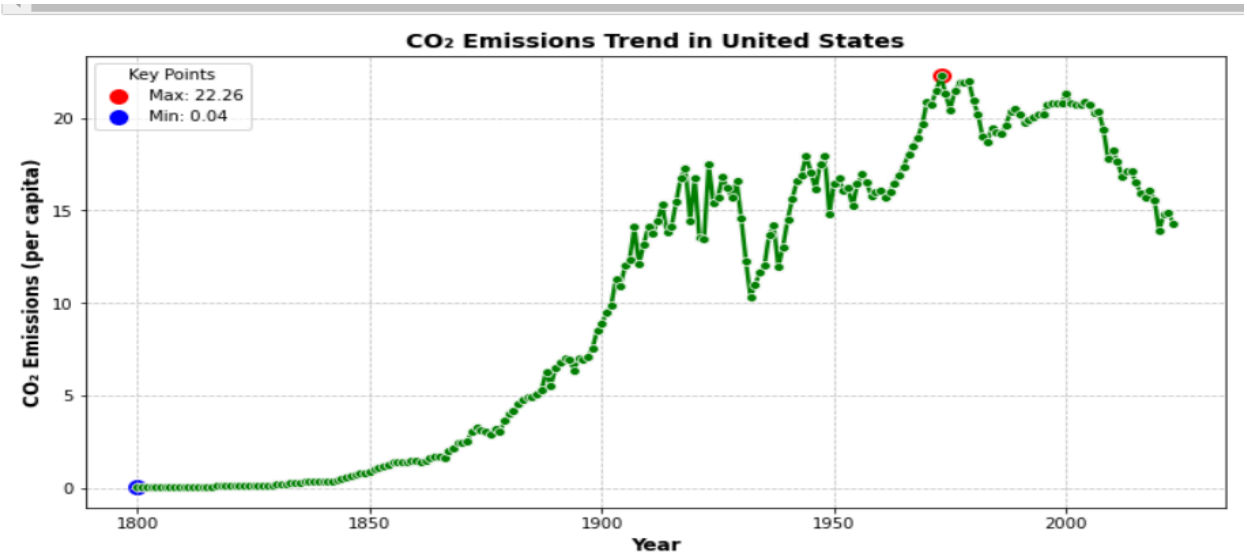


Fig.14. Time series trend of co2 emissions over the world

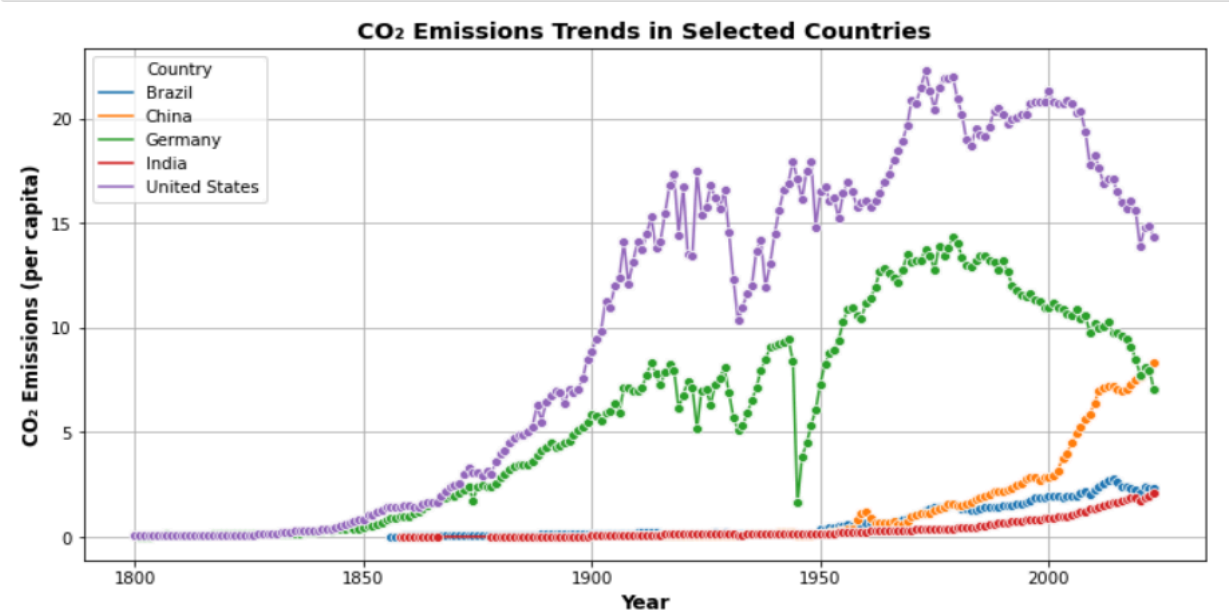


Fig.15. Time series trend of co2 emissions in selected countries

In the diagram below there are significant variations with multiple peaks and dips, indicating that precipitation levels are not constant but fluctuate significantly across decades.

5.4 VISUALIZATIONS - PRECIPITATION

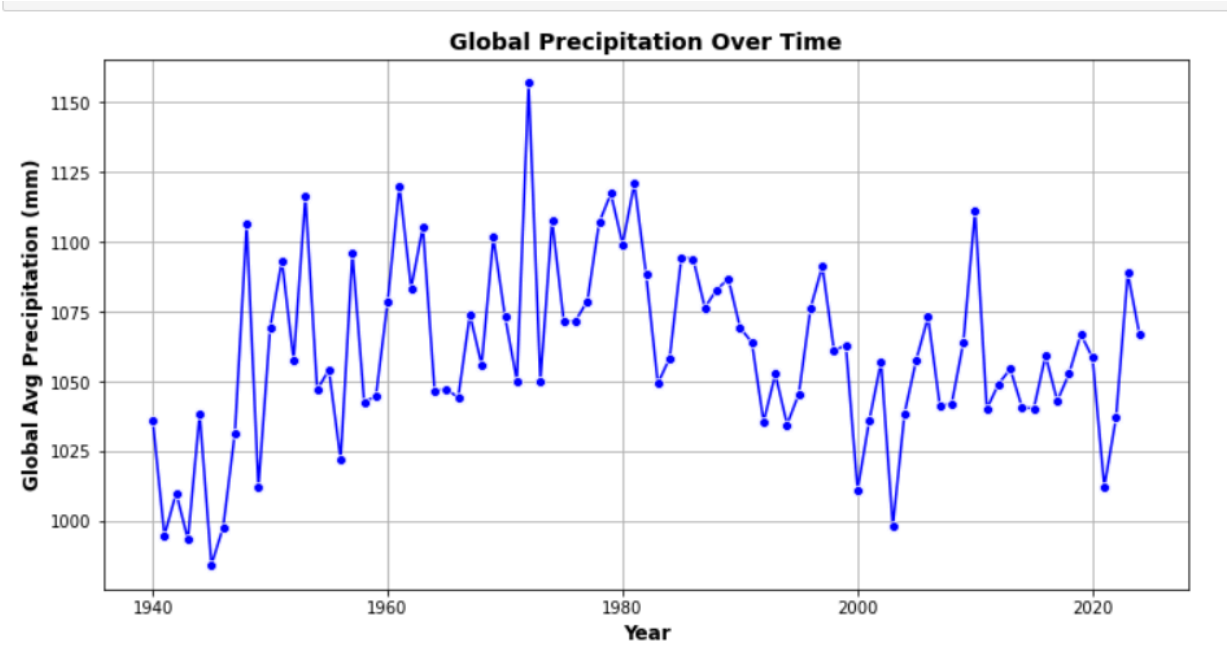


Fig.16. Time series trend of precipitation all over the world

Columbia has the highest precipitation rate and brunei has the lowest precipitation rate when compared to other countries

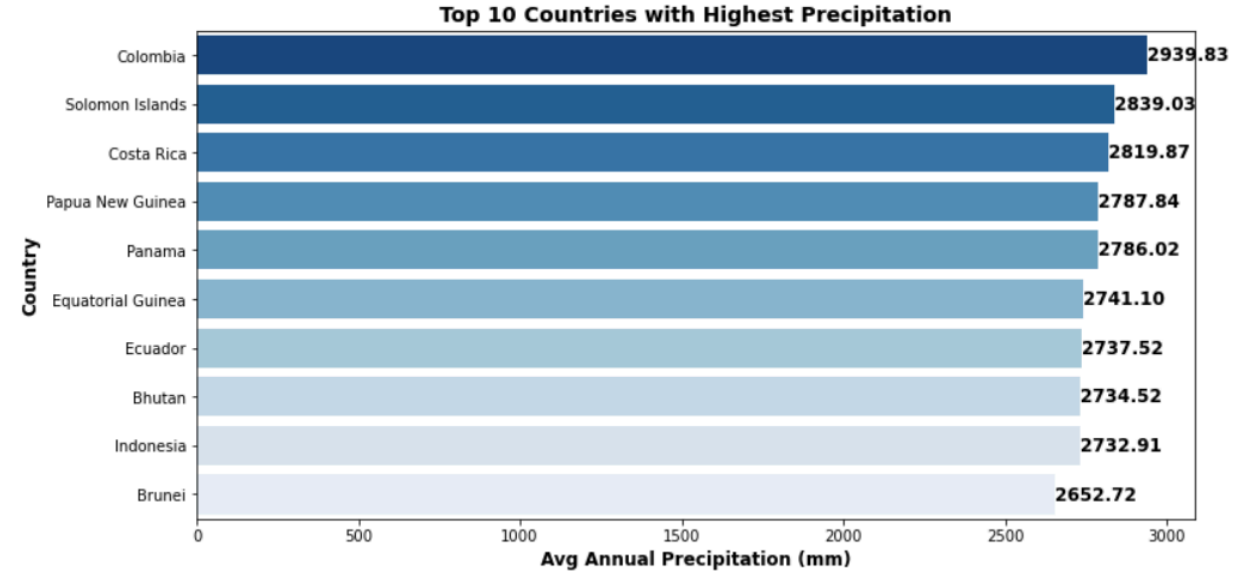


Fig.17. Barchart of highest precipitation of top 10 countries

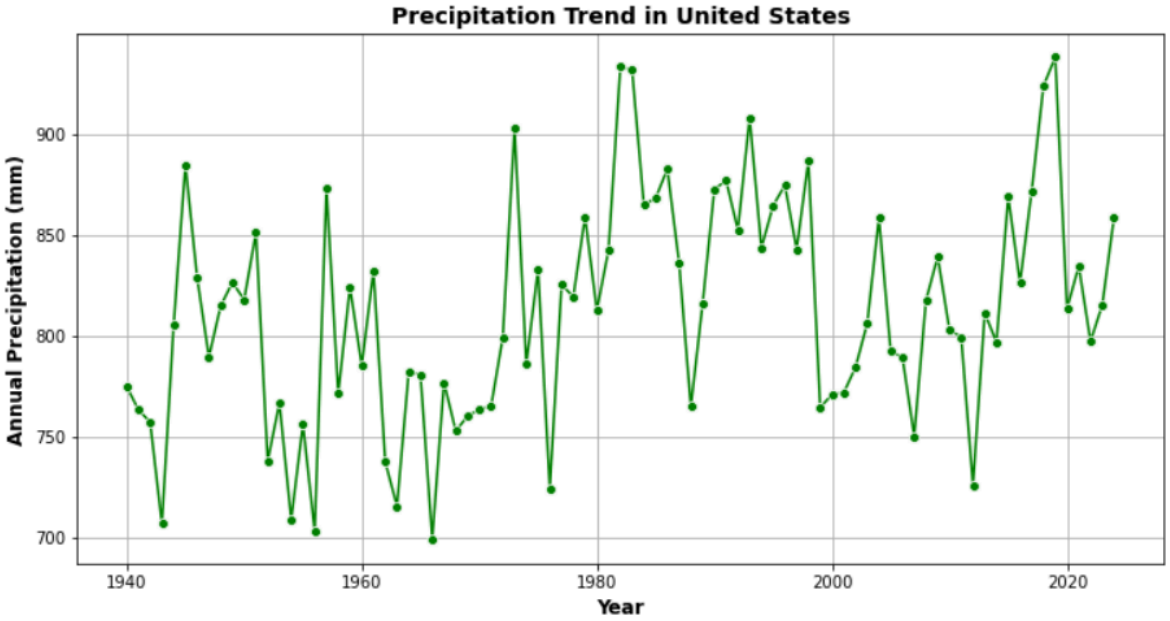


Fig.18. Time series trend of precipitation over US

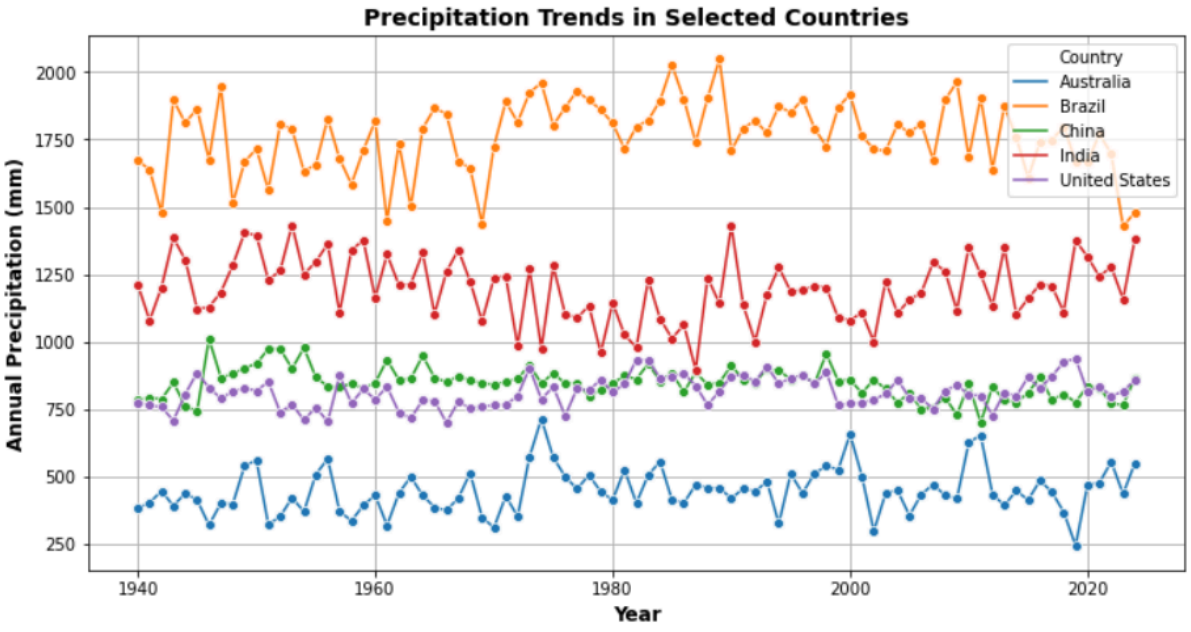


Fig.19. Time series trend of precipitation over selected countries

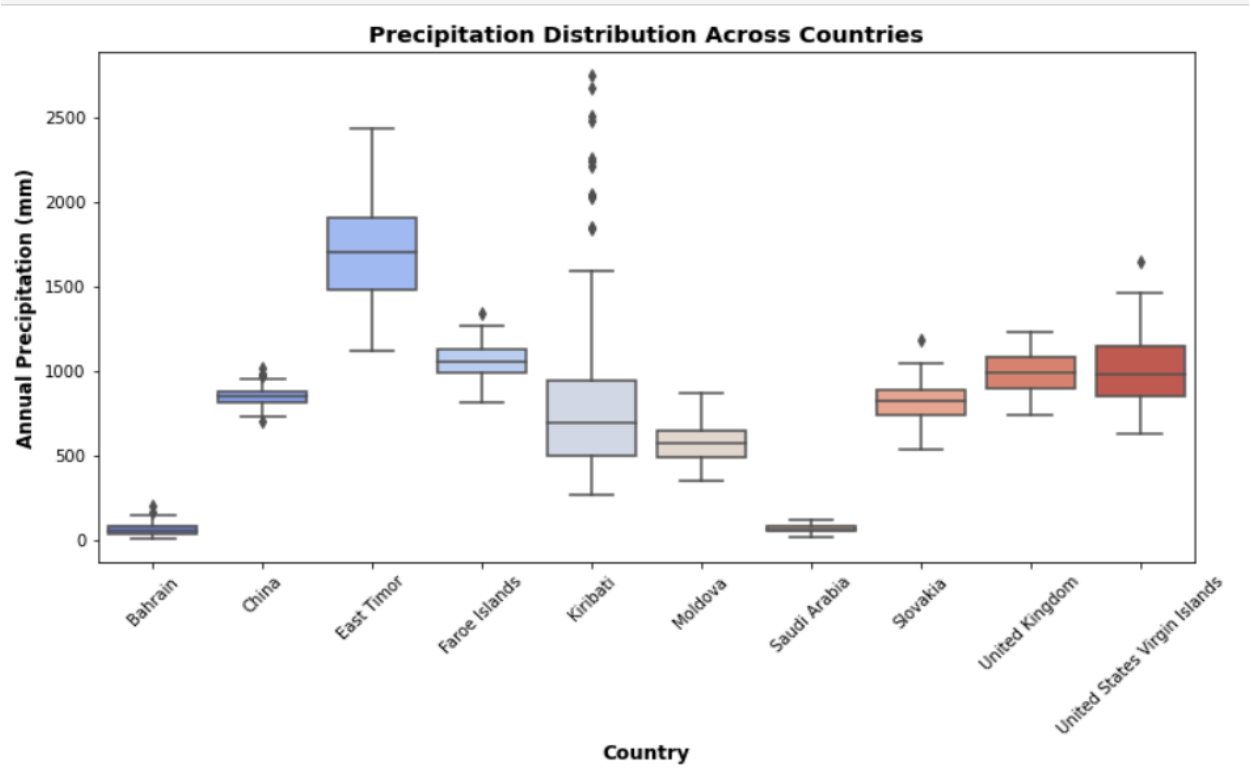


Fig.20. Time series trend of precipitation over some countries

5.5 VISUALIZATIONS - GREEN COVER

Here the forest area gradually increased and went to peak, then fell drastically.

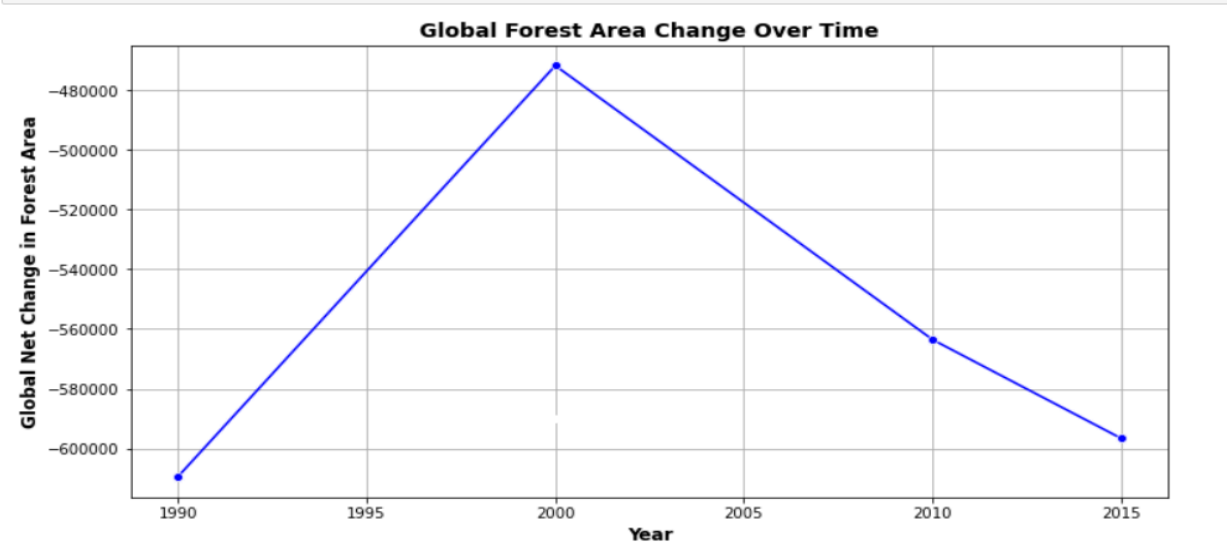


Fig.21. Time series trend of green cover over various countries in world

From below, somalia has the highest forest loss over the date

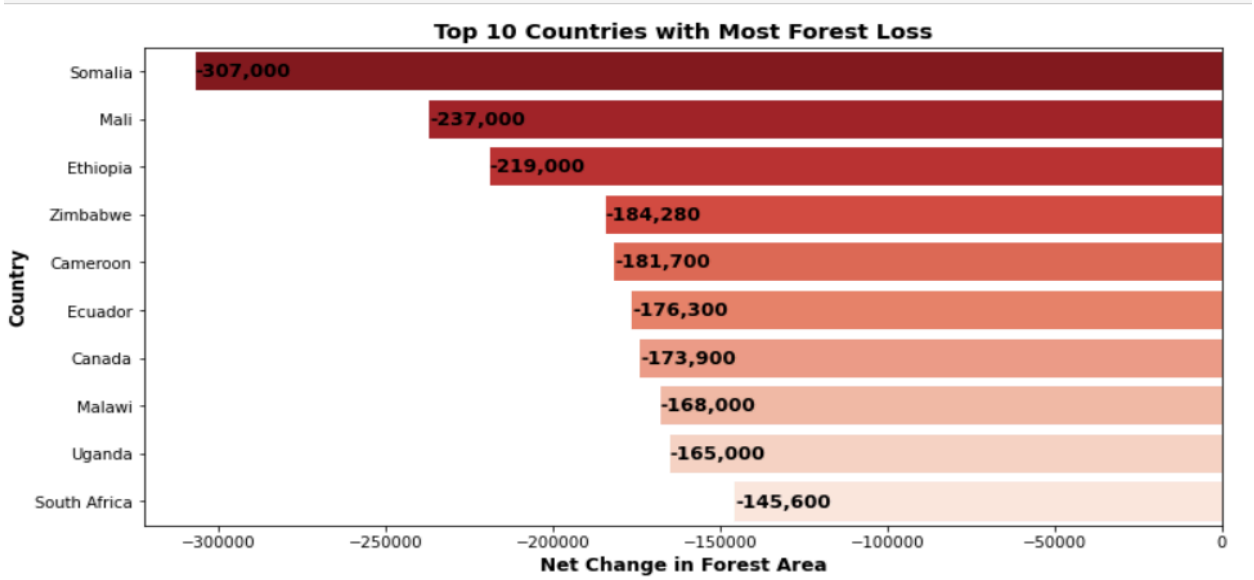


Fig.22. Time series trend of green cover for over top 10 countries

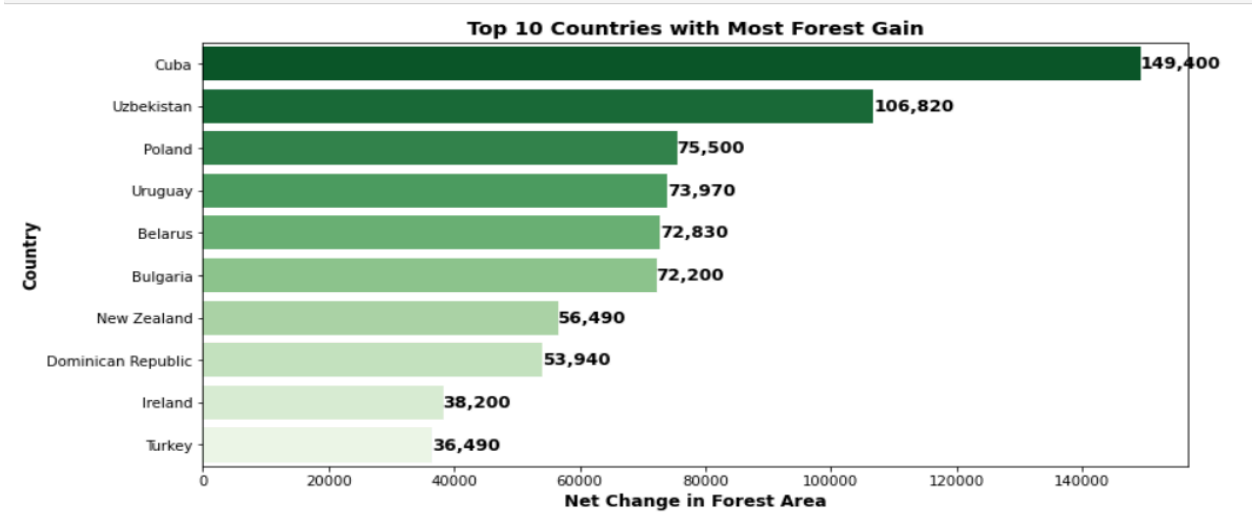


Fig.23. Barchart with less affected green cover area countries

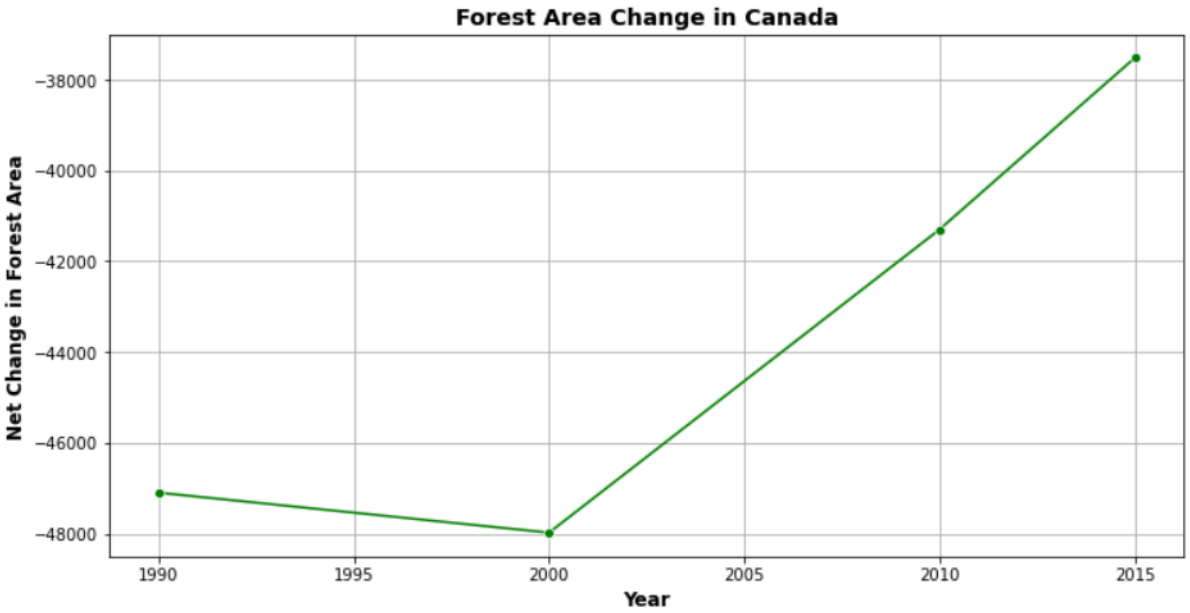


Fig.24. Time series trends for green cover in canada

5.6 VISUALIZATIONS - NATURAL DISASTER AND EMERGENCY EVENTS

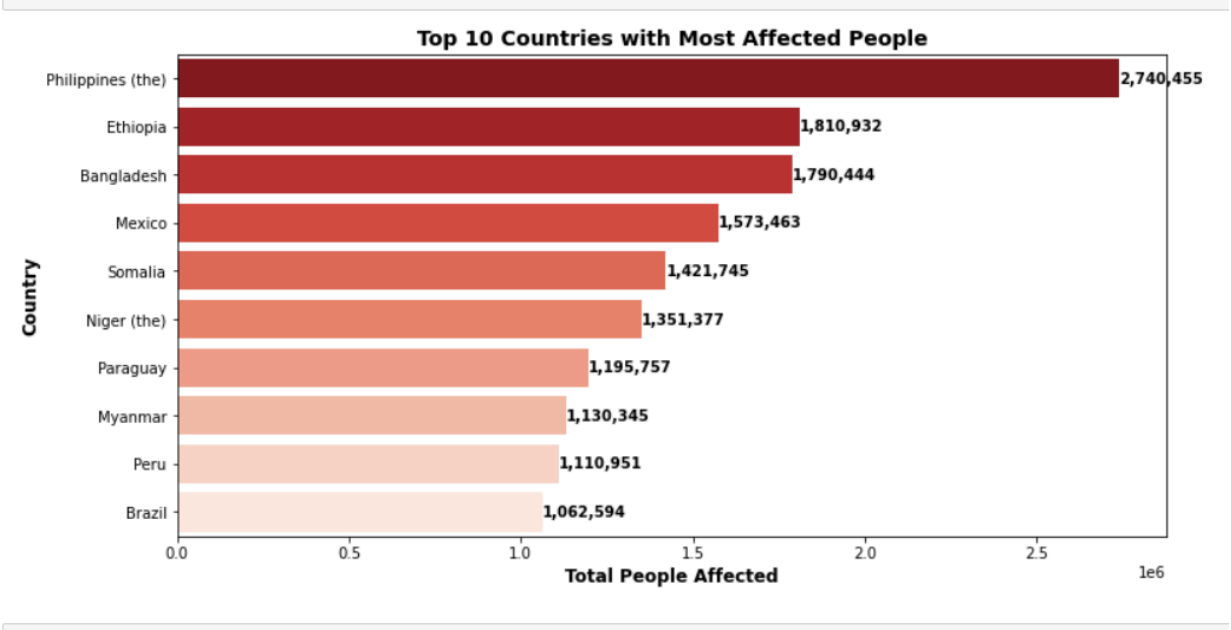


Fig.25. Barchart for top 10 countries with most affected people due to natural disaster.

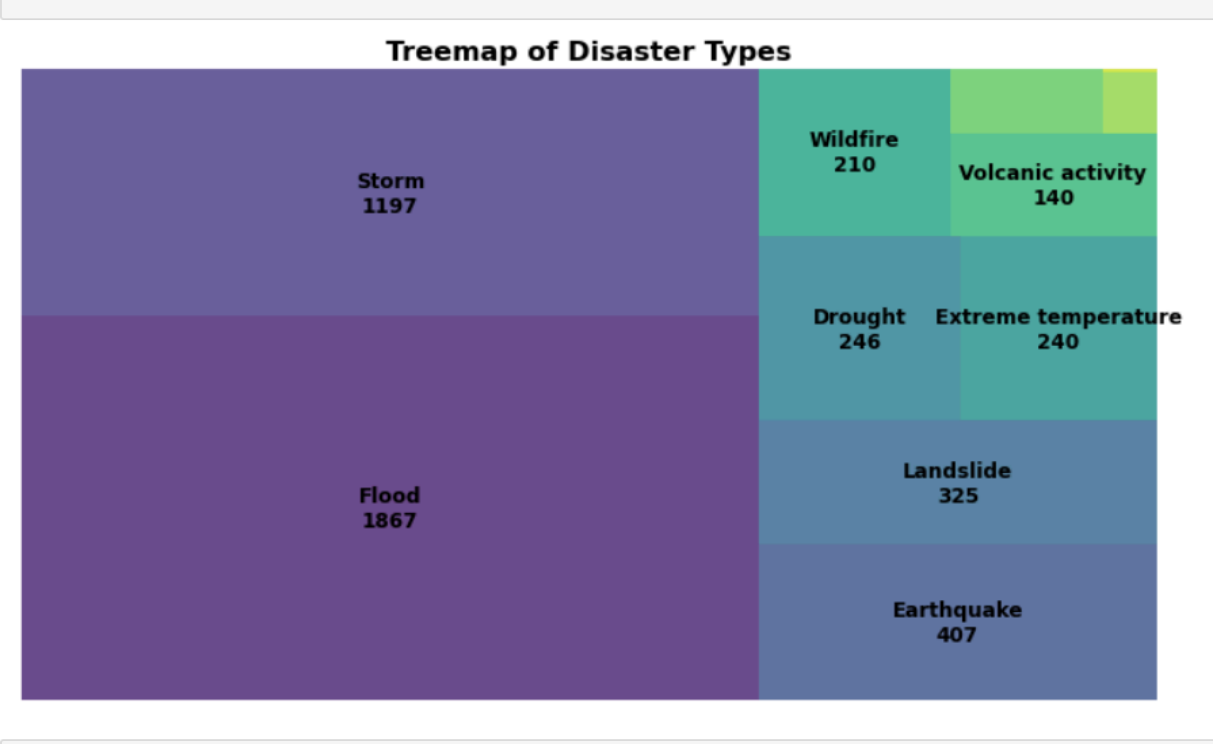


Fig.26. Barchart for top 10 countries with most affected people due to natural disasters.

This treemap visually represents the distribution of different types of disasters where each block corresponds to the frequency of that disaster type. From this, the most common disaster is strom.

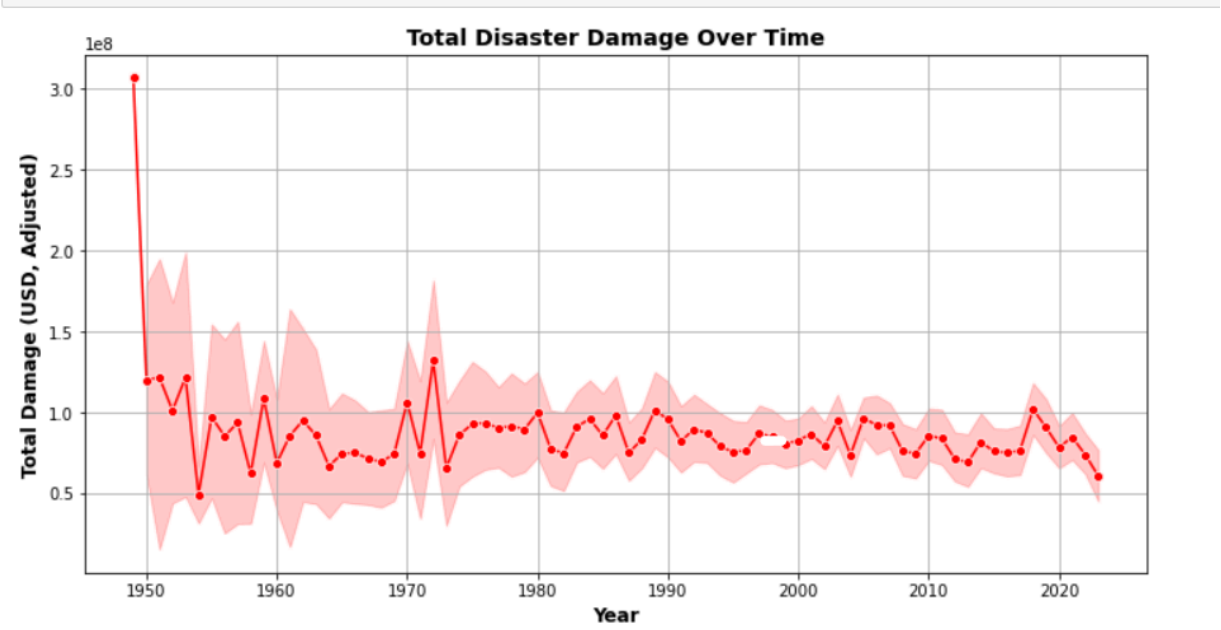


Fig.27. Total Disaster Damage in USD Over Time

A sharp spike in damage is visible around 1950, indicating a major disaster event that caused significant economic losses. After that, there are various fluctuations in the damage amount.

6. CONCLUSION

Data cleaning, preparation, and loading were carried out meticulously for sound analysis by checking for missing values, outliers, and consistency. EDA sets the successful stage for extraction of vital knowledge through descriptive statistics, visualizations, and trend analysis. Cumulatively, these steps are establishing a solid foundation for predictive modeling and informed decision-making in the next phases of the project. As per the EDA, temperature and occurrences of disasters are showing upward trends with large fluctuations through time. These extreme variations are reflective of increasing consequences of climate change and environmental instability.