**A PRELIMINARY MINI PROJECT REPORT ON**

**Analysis of Disasters in Asia Using EDA and Visualization**

**SUBMITTED TOWARDS THE PARTIAL FULFILLMENT OF THE**

**REQUIREMENTS OF**

**BACHELOR OF ENGINEERING (S.Y. B. Tech.)**

**Academic Year: 2024-25**

**By:**

Shyam Sundar Yadav (123B1B267)

Siddhartha Kumar Nidhi (123B1B268)

Sujit Chaudhary (123B1B272)

Varun Goyal (123B1B278)

Under The Guidance of:

**Dr. Chetan Chauhan**



**DEPARTMENT OF COMPUTER ENGINEERING,**

**PIMPRI CHINCHWAD COLLEGE OF ENGINEERING**

**SECTOR 26, NIGDI, PRADHIKARAN**

**List of Figures & Tables**

|  |  |  |
| --- | --- | --- |
| **Figure/Table**  **No.** | **Figure/ Table Name** | **Page No.** |
| 1. | Bar plot for top 10 countries by Frequency of disaster | NA |
| 2. | Line plot for Total effected by disaster in last 10 years | NA |
| 3. | Count plot for Number of Natural Disaster per year | NA |
| 4. | Bar plot for Total Economic Losses by Disaster Type | NA |
| 5. | Bar plot for Top 10 Deadliest Disaster Types All Over Asia | NA |
| 6. | Heatmap Correlation Matrix for Disaster Impacts | NA |
| 7. | Pie for Distribution of Disaster Types | NA |
| 8. | Scatterplot for Storm Counts and Linear Regression Model | NA |
| 9. | Linear Regression plot for Prediction of Linear Regression: Actual vs. Predicted Total Deaths and Total Deaths for the Next 5 Years | NA |
| 10. | ROC curve | NA |
| 11. | Histogram for Prediction of Probability Distribution of "High" Deaths | NA |

**CHAPTER 1: INTRODUCTION**

**1.1 Background**

Natural disasters are catastrophic events that occur due to natural processes of the Earth. They can have devastating effects on human societies, economies, and the environment, leading to loss of life, destruction of infrastructure, displacement of populations, and long-term economic hardship. Disasters like earthquakes, floods, hurricanes, wildfires, and droughts have always been part of human history, but their frequency, intensity, and geographic spread seem to be increasing, largely due to human-induced climate change. The impacts of these events can be far-reaching, affecting everything from public health to global supply chains.

In recent years, the world has witnessed an alarming rise in the severity of natural disasters. The frequency of major hurricanes and floods, the intensity of wildfires, and the destructiveness of earthquakes have all significantly impacted global economies and communities. The rising cost of damage from these events is overwhelming for both developed and developing countries, with some nations facing the prospect of bankruptcy due to the scale of destruction. For example, the 2004 Indian Ocean tsunami caused over 230,000 deaths and widespread damage to coastal infrastructure in 14 countries. More recently, hurricanes like Katrina (2005), Harvey (2017), and Dorian (2019) have led to hundreds of fatalities and billions in economic losses.

Despite the obvious need for more effective disaster response strategies, many governments and humanitarian organizations still struggle to make informed decisions due to a lack of accessible, comprehensive, and well-structured disaster data. Disasters not only cause immense loss of life and damage, but they also disrupt economic activities, cripple healthcare systems, and set back developmental goals. In fact, studies show that the economic cost of natural disasters has increased exponentially over the past few decades, making it imperative for both policy makers and disaster management agencies to improve their understanding of disaster trends and impacts.

One critical way to improve preparedness is by analyzing historical disaster data. By examining trends over time, the frequency of certain disaster types, the regions most affected, and their associated costs (both human and economic), governments and agencies can design more effective strategies for risk reduction, resource allocation, and disaster response. Such insights are vital for proactive decision-making and long-term planning. A structured analysis of past disaster data can also guide emergency preparedness initiatives, from early warning systems to the allocation of financial resources for recovery and reconstruction.

The key to improving disaster response and preparedness lies not only in understanding the present landscape of natural disasters but also in learning from past occurrences. This project leverages a dataset containing historical disaster data to examine the patterns of natural disasters and their impacts on societies and economies. By understanding past events, we can predict potential future disaster hotspots and plan accordingly, ensuring that the right measures are in place to mitigate the adverse effects of these catastrophic events.

**1.2 Problem Statement**

Natural disasters continue to cause widespread devastation worldwide, yet many organizations and governments struggle to devise effective response strategies due to insufficient historical data analysis. This project seeks to address this gap by analyzing disaster data to uncover trends, correlations, and key insights regarding the impact of different disaster types on fatalities and economic loss. By identifying the most impactful disasters and understanding their patterns, this study aims to improve disaster preparedness, enhance resource allocation strategies, and help governments make informed decisions to reduce disaster-related risks.

**1.3 Project Idea / Objectives**

* Analyze the frequency of different types of natural disasters over time.
* Assess the impact of these disasters by measuring total deaths and economic damage.
* Identify patterns in disaster occurrences and their geographic distribution.
* To predict the future calamities using our model.

**1.4 Motivation**

* The frequency and severity of natural disasters have escalated in recent years, influenced by both natural forces and human activities, such as climate change. These events have profound consequences for communities, economies, and ecosystems, making disaster preparedness and response critical. For governments and organizations to manage these events effectively, it is vital to have accurate, comprehensive, and timely data. Unfortunately, many countries still lack the necessary tools to analyze past disaster data in a way that could guide future decision-making. The data collected from previous natural disasters offers a wealth of insights that can be leveraged to minimize damage, save lives, and enhance recovery efforts.
* This project is motivated by the growing need for a systematic and data-driven approach to disaster management. By analyzing historical disaster data, this study aims to identify patterns in the occurrence, severity, and impact of natural disasters. Such insights can significantly improve the way resources are allocated, where early warning systems are deployed, and how emergency responses are structured. Additionally, understanding the economic and human toll of past disasters can help prioritize investments in disaster resilience and risk reduction measures. For instance, identifying areas most prone to specific types of disasters—such as flooding or earthquakes—can help governments design more localized and targeted interventions.
* Moreover, this project is driven by the need to raise awareness about the disproportionate impact of natural disasters on vulnerable communities. By highlighting the patterns of disasters, we can advocate better preparedness and faster recovery strategies for those most at risk. The ultimate motivation of this project is to make disaster response more effective, efficient, and equitable, ultimately saving lives and minimizing the socio-economic consequences of these unavoidable events.

**1.5 Scope**

This project aims to analyze a dataset of historical natural disasters to uncover trends in their frequency, severity, and impact over time. The primary focus is on understanding the relationships between various disaster types (such as floods, earthquakes, hurricanes, etc.), the number of fatalities, and the associated economic damage. The analysis will specifically highlight which types of disasters have been most impactful and in which regions.

The scope of this study is limited to the dataset provided, which contains historical disaster data up until the present year. While the dataset offers valuable insights, it may not include real-time disaster events or cover all global regions comprehensively. Additionally, the accuracy of the data, especially concerning economic losses and fatalities, depends on the availability and reliability of disaster reporting. Some regions may have underreported or inconsistent data, which could affect the findings.

Furthermore, the project focuses on analyzing past disasters rather than predicting future events. Though trends will be observed, the study does not incorporate advanced predictive modeling or real-time forecasting techniques. The results are intended to inform disaster management strategies, but the findings will not account for future, unforeseen shifts in disaster occurrence or severity due to changing environmental or geopolitical conditions.

While the primary goal is to uncover historical patterns, the scope also includes the exploration of possible areas for future research. This could involve integrating more granular, region-specific data or extending the analysis to include emerging disaster types not fully captured in the dataset.

**CHAPTER 2: DATA COLLECTION**

**2.1 Dataset Information**

The dataset used for this project is sourced from Kaggle and is titled Disaster SEA. It contains historical data on natural disasters that occurred in Southeast Asia, focusing on their types, fatalities, and economic damages. The dataset provides valuable insights into disaster patterns over time, making it an ideal resource for analyzing the impact of different types of natural disasters on both human lives and the economy.

This dataset is a comprehensive collection of recorded disaster events and includes a wide range of attributes such as disaster type, date, fatalities, affected population, economic damage, and region. The dataset spans multiple decades, offering a historical perspective that can be used to identify trends, correlations, and the evolving impact of disasters. The data is structured in a way that allows for detailed analysis, including temporal and spatial trends, as well as comparative analysis between different types of disasters.

The primary motivation for choosing this dataset lies in its relevance and applicability to the region under study—Southeast Asia. This region is known for being highly vulnerable to a variety of natural disasters, including earthquakes, typhoons, floods, and volcanic eruptions. Understanding the frequency and intensity of these events is crucial for improving disaster preparedness and response strategies.

**2.2 Data Attributes**

The Disaster SEA dataset contains multiple attributes that describe each natural disaster event in detail. These attributes help in analyzing various aspects of the disaster, such as its type, impact, affected populations, and the economic costs associated with it. Each attribute serves as a building block for understanding the relationships between different disaster characteristics and their outcomes. Below is a detailed description of the dataset’s attributes, highlighting their significance and role in the analysis.

1. **DisNo.:** Unique identifier for each disaster event.
2. **Historic:** Indicates if the disaster event is historic.
3. **Classification Key:** Classification code representing the disaster’s nature.
4. **Disaster Group:** Broad category of the disaster (e.g., Natural, Technological).
5. **Disaster Subgroup:** Subcategory within the disaster group
6. **Disaster Type:** Specific disaster type within the subgroup.
7. **Disaster Subtype:** Further classification within the disaster type.
8. **External IDs:** External identification codes for the disaster.
9. **Event Name:** Name given to the disaster event, if available.
10. **ISO:** ISO country code of the affected region.
11. **Country:** Name of the affected country.
12. **Subregion:** Subregional classification within the affected country.
13. **Region:** Geographical region classification.
14. **Location:** Specific location details of the disaster.
15. **Origin:** The origin or cause of the disaster.
16. **Associated Types:** Other disaster types associated with the event.
17. **OFDA/BHA Response:** Indicates if there was a response from the U.S. Office of Foreign Disaster Assistance (OFDA) or Bureau for Humanitarian Assistance (BHA).
18. **Appeal:** Indicates if an appeal for aid was made.
19. **Declaration:** Declaration status for international aid or emergency.
20. **AID Contribution ('000 US$):** Aid contribution received in thousands of US dollars.
21. **Magnitude:** Magnitude of the disaster (e.g., intensity for earthquakes).
22. **Magnitude Scale:** Scale used to measure the disaster magnitude.
23. **Latitude:** Latitude of the disaster location.
24. **Longitude:** Longitude of the disaster location.
25. **River Basin:** River basin affected, if applicable.
26. **Start Year:** Year the disaster started.
27. **Start Month:** Month the disaster started.
28. **Start Day:** Day the disaster started.
29. **End Year:** Year the disaster ended.
30. **Start Day:** Day the disaster started.
31. **End Year:** Year the disaster ended.
32. **End Month:** Month the disaster ended.
33. **End Day:** Day the disaster ended.
34. **Total Deaths:** Total deaths reported from the disaster.
35. **No. Injured:** Number of people injured in the disaster.
36. **No. Affected:** Number of people affected by the disaster.
37. **No. Homeless:** Number of people rendered homeless by the disaster.
38. **Total Affected:** Total number of people affected (including injured and homeless).
39. **Reconstruction Costs ('000 US$):** Estimated reconstruction costs in thousands of US dollars.
40. **Reconstruction Costs, Adjusted ('000 US$):** Inflation-adjusted reconstruction costs.
41. **Insured Damage ('000 US$):** The total amount of damage covered by insurance in thousands of US dollars.
42. **Insured Damage, Adjusted ('000 US$):** Inflation-adjusted insured damage costs.
43. **Total Damage ('000 US$):** The total economic damage caused by the disaster in thousands of US dollars.
44. **Total Damage, Adjusted ('000 US$):** Inflation-adjusted total damage costs.
45. **CPI:** The Consumer Price Index used for adjusting monetary values in the dataset.
46. **Admin Units:** The administrative regions (e.g., states, provinces) affected by the disaster.
47. **Entry Date:** The date when the disaster record was entered into the dataset.
48. **Last Update:** The date when the disaster record was last updated.

**CHAPTER 3: EXPLORATORY DATA ANALYSIS (EDA)**

**3.1 Data Preprocessing**

Before performing any detailed analysis, data preprocessing is essential to ensure that the dataset is clean and ready for exploration. The following steps were taken during the data preprocessing phase:

Filtering Data: The dataset was filtered to focus on specific regions or types of disasters, based on the project’s objectives. The columns which were not playing a significant role in prediction were removed and only those columns which were important for visualization and prediction purpose were kept.

**3.2 Exploratory Visualizations**

**Countplot of Disaster Frequency by Country:** The countplot visualizes the distribution of disaster occurrences across different countries. By focusing on the top 10 countries, this plot highlights where natural disasters are most frequent. The y-axis represents the number of disasters, while the x-axis shows the countries, ordered by the number of occurrences. This helps to quickly identify disaster-prone regions and assess which countries need more focused disaster preparedness.

**Lineplot of Total Affected Over the Years:** The lineplot illustrates the trend in the number of people affected by disasters over time. The x-axis represents the years (Start Year), while the y-axis shows the total number of people affected by disasters in those years. By analyzing this lineplot, we can observe whether the total number of people affected by natural disasters has been increasing or decreasing, indicating the growing or shrinking impact of disasters over time.

**Countplot of Disasters by Year:** The countplot of disasters by year shows the frequency of disasters that occurred each year. This plot reveals trends in disaster occurrence over time, showing years with higher or lower disaster activity. A spike in disaster frequency may correspond to particular events (e.g., extreme weather, geopolitical crises), and this visualization helps identify such periods and assess disaster response effectiveness during those years.

**Barplot of Economic Damage by Disaster Type:** The barplot provides a comparison of the total economic damage caused by different disaster types. The x-axis represents the various disaster types, and the y-axis shows the total damage (in '000 US$). The data is grouped by disaster type and broken down by disaster group (e.g., natural, man-made). This plot allows us to compare the economic toll of various disaster types and provides insight into which disasters have the most significant financial impact.

**Barplot of Total Deaths by Disaster Type**: This barplot shows the total number of deaths caused by the top 10 deadliest disaster types. The y-axis represents disaster types, and the x-axis shows the total deaths. By using this plot, we can identify which types of disasters have caused the highest number of fatalities, which can help prioritize disaster prevention and response efforts in the most lethal disaster categories.

**Correlation Heatmap:** The heatmap visually represents the correlation between various numerical variables, such as 'Total Deaths,' 'Total Affected,' and 'Total Damage.' The colors represent the strength of the correlation between variables, with darker shades indicating stronger correlations. This visualization allows us to easily spot relationships between variables, for example, whether a higher number of deaths is associated with more people affected or higher economic damage. It helps us understand the interdependence among the different aspects of disaster data.

**Pie Chart of Disaster Types:** The pie chart is used to represent the proportion of different disaster types in the dataset. Each slice of the pie corresponds to a different disaster type, and its size represents the proportion of that disaster type within the dataset. This visualization provides an immediate understanding of the distribution of disaster types, showing which disaster types are most common and how they are distributed across the dataset.

**Linear Regression Model for Predicting Future Disaster Trends**: The linear regression model is used to predict future outcomes based on historical disaster data. In this case, the model predicts the total deaths for future years. The x-axis represents the years (including predictions for 2024-2027), while the y-axis shows the predicted number of deaths. The model's predictions are plotted alongside the historical data to visualize trends. This helps to forecast how disaster fatalities may change in the coming years, providing valuable insights into future disaster preparedness.

**Visualization of Linear Regression Predictions for Future Years:** In this line plot, the predicted total deaths for the next few years (2024-2027) are shown. The predictions from the linear regression model are plotted with markers and connected by a line, which helps visualize the trend of expected disaster fatalities in the near future. This is a crucial visualization for understanding whether the number of fatalities is expected to rise, stay the same, or decrease, which can inform disaster management strategies.

**Scatter Plot and Linear Regression Line for Storm Counts:** The scatter plot shows historical data points of storm counts or fatalities across different years. Each point represents a specific year’s disaster data. A linear regression line is drawn to show the trend in storm counts over time. This line helps visualize whether the number of storms or fatalities has increased or decreased over the years. Predictions for future years (2024-2027) are added to the plot as red markers, showing the model's forecasted storm counts or fatalities. This scatter plot helps assess the fit of the linear regression model and provides a clear view of future predictions.

**Receiver Operating Characteristic (ROC) Curve:** The ROC curve is used to evaluate the performance of a classification model. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR), showing how well the model distinguishes between classes (e.g., high vs. low disaster impact). A higher area under the curve (AUC) indicates better model performance. The ROC curve is a valuable tool for assessing the accuracy and reliability of predictive models, especially for disaster impact classification.

**Histogram of Predicted Probabilities:** The histogram of predicted probabilities shows the distribution of predicted probabilities from the model. It represents how likely each test sample is to belong to the "high impact" category (e.g., high death toll or economic damage). This plot helps understand the model's confidence in its predictions and can reveal if the model is skewed toward predicting one outcome over another.

**CHAPTER 4: METHODOLOGY**

**4.1 Feature Selection**

**Linear Regression:**

Linear Regression is used to model the relationship between a continuous dependent variable and one or more independent variables. In this project, Linear Regression helps in predicting continuous outcomes such as the total number of deaths or the number of people affected by a disaster over time. This is done by understanding historical trends and extrapolating them into future predictions.

**Logistic Regression:**

Logistic Regression is applied when the outcome variable is categorical. It is used for classification tasks such as predicting whether a disaster will result in a "high fatality" or "low fatality" based on the available features. Logistic regression outputs a probability score for each class and assigns the class with the highest probability.

**4.2 Feature Selection and Model Workflow**

**Steps for Linear Regression:**

* Data Collection and Preparation:
  + The dataset is sourced from Kaggle, specifically the [Disaster SEA dataset](https://www.kaggle.com/datasets/hoangvu128/disaster-sea). The data consists of attributes such as "Disaster Type," "Start Year," "Total Deaths," "Total Affected," and "Economic Damage (USD)."
  + Data cleaning is performed to handle missing values, remove duplicates, and address any inconsistencies. Categorical variables such as "Country" or "Disaster Type" are encoded as needed.
* Feature Selection:
  + In Linear Regression, the dependent variable is continuous, and independent variables can include year, disaster type, and economic damage. For example, we might use "Start Year" and "Disaster Type" as predictors to predict "Total Deaths" or "Total Affected."
  + Relevant features are selected, ensuring that they have a meaningful relationship with the target variable.
* Splitting Data:
  + The dataset is split into a training set (60%) and a testing set (40%). The training set is used to train the Linear Regression model, while the testing set is used for model evaluation.
* Model Training:
  + Linear Regression is applied to the training data. The model learns the relationship between independent variables and the dependent variable.
* Model Evaluation:
  + The model is evaluated using metrics like Root Mean Squared Error (RMSE), and R-squared. These metrics assess the accuracy of the model’s predictions and how well it fits the data.
* Prediction:
  + Once trained, the model predicts future disaster impacts (e.g., predicted number of deaths or affected population) for years not included in the dataset, such as 2024 or beyond.
* Visualization:
  + Results are visualized using line plots to show the predicted disaster outcomes over time, compared with actual historical values. For example, a plot showing the predicted number of deaths for the next few years based on historical trends.

**Steps for Logistic Regression:**

* Data Collection and Preparation:
  + As similar to Linear Regression, data cleaning is performed on the Kaggle Disaster SEA dataset. Necessary transformations are made to prepare the data for classification.
* Feature Selection:
  + Logistic Regression requires a binary or categorical target variable. For example, we might create a binary target variable "High Fatality" (1 for high fatalities and 0 for low fatalities) based on "Total Deaths."
  + Independent variables such as "Disaster Type," "Total Damage," "Country," and "Year" are used to predict the target.
* Splitting Data:
  + The data is split into training and testing sets, typically using a 70-30 ratio. The training data is used to fit the model, while the testing data evaluates its performance.
* Model Training:
  + Logistic Regression is trained on the data, where the model learns the relationship between the independent variables and the binary target. For instance, it learns the association between "Disaster Type," "Country," and "Total Deaths" with the likelihood of a disaster causing a high number of fatalities.
* Model Evaluation:
  + Evaluation metrics for Logistic Regression include accuracy, precision, recall, F1-score, and the Area Under the ROC Curve (AUC). These metrics help assess how well the model predicts the binary target (e.g., high vs. low fatalities).
  + A confusion matrix is used to measure the number of true positives, false positives, true negatives, and false negatives.
* Prediction:
  + The Logistic Regression model classifies new disaster data into categories like "high fatalities" or "low fatalities." It outputs probabilities that are used to assign the class with the highest probability.
* Visualization:
  + Visualization techniques such as ROC curves and precision-recall curves are used to evaluate the classification performance of the model. Histograms of predicted probabilities can be used to understand how confident the model is about its predictions.

**Difference between Linear Regression and Logistic Regression**

| **Aspect** | **Linear Regression** | **Logistic Regression** |
| --- | --- | --- |
| **Purpose** | Predicts a continuous outcome (e.g., exact number of deaths). | Predicts a binary outcome (e.g., "High" or "Low" total deaths). |
| **Dependent Variable (Target)** | Continuous variable (e.g., "Total Deaths" directly). | Binary variable (e.g., "High" vs. "Low" based on median threshold of "Total Deaths"). |
| **Output** | Continuous value predictions (e.g., 100, 250, etc.). | Probabilities for each class (e.g., probability of "High" or "Low" deaths). |
| **Interpretation of Results** | Provides an exact prediction for the outcome. | Classifies each instance into one of two categories based on a probability threshold. |
| **Error Measurement** | Measured using RMSE (Root Mean Square Error) to capture how close predictions are to actual values. | Evaluated with metrics like accuracy, precision, recall, and F1-score to measure classification quality. |
| **Model Evaluation** | Residuals and RMSE for fit quality assessment. | Confusion matrix, classification report, ROC curve, and AUC for classification performance. |
| **Application** | Best for predicting exact numerical values. | Best for predicting categories (e.g., whether the impact will be high or low). |
| **Interpretation of Coefficients** | Coefficients indicate how much the target variable changes with a unit change in predictor variables. | Coefficients indicate the log odds of the target variable belonging to a particular class with a unit change in predictor variables. |
| **Assumptions** | Assumes a linear relationship between predictors and the outcome. | Assumes a logistic (S-shaped) relationship between predictors and the probability of the binary outcome. |
| **Prediction Range** | Predicted values can range from −∞-\infty−∞ to +∞+\infty+∞, which may be unrealistic for certain datasets. | Predictions are constrained between 0 and 1, representing probabilities. |
| **Suitability for This Dataset** | Suitable for estimating total deaths if exact prediction is desired. | Suitable for categorizing the impact into "High" or "Low" death cases for easier risk assessment. |

**CHAPTER 5: MODEL EVALUATION**

In this chapter, we evaluate the performance of the models used in the project, based on different evaluation metrics.

**5.1 Metrics Used**

To assess how well the model performs, we used the following evaluation metrics:

* **Accuracy:** This metric represents the percentage of correctly predicted instances out of the total instances in the dataset. Accuracy is useful for an overall assessment of the model’s performance, especially when the classes are balanced. However, in cases where the data is imbalanced (for instance, if fatalities are rare), accuracy alone might be misleading.
* **Precision:** This metric indicates how many of the positively predicted instances are positive. It’s especially important when the cost of false positives is high (e.g., incorrectly predicting a disaster type or severity).
* **Recall:** This metric, also known as sensitivity, measures how many actual positive instances were correctly identified. It is crucial when the cost of missing out on positive cases (like fatalities or economic damage) is high.​
* **F1-Score:** The F1-score is the harmonic mean of precision and recall, offering a balance between the two. It’s especially useful when dealing with imbalanced datasets and provides a single metric for model performance.

**5.2 Confusion Matrix**

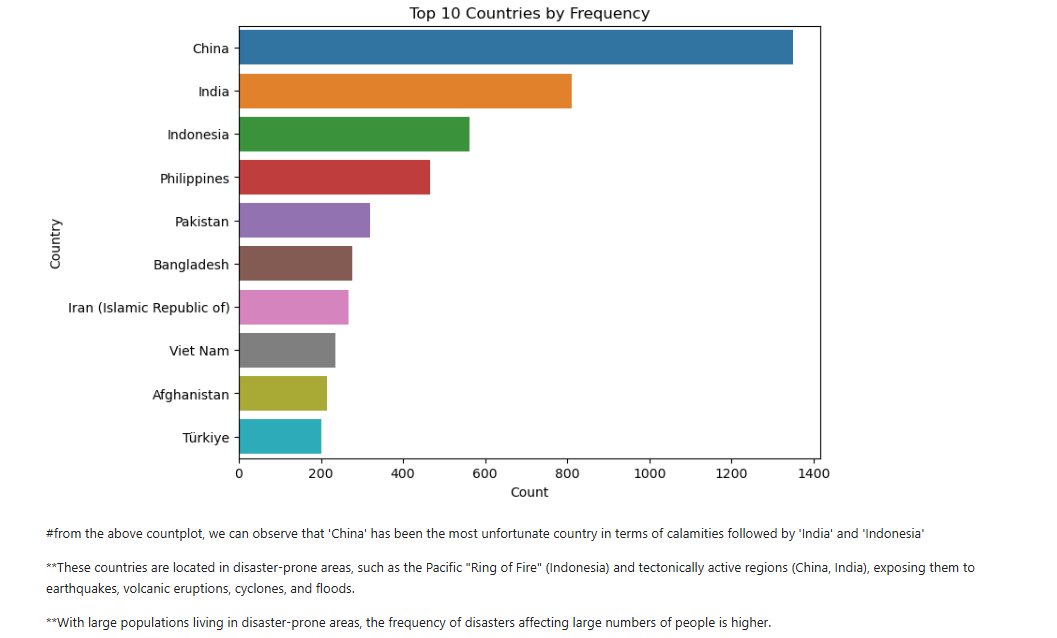
The confusion matrix is a critical tool for understanding the performance of classification models. It displays the true positives, false positives, true negatives, and false negatives for each class, allowing us to assess the model’s predictive accuracy in detail. A well-balanced confusion matrix shows that the model is correctly classifying each disaster type, minimizing false predictions.

**CHAPTER 6: RESULTS AND DISCUSSION**

In this chapter, we analyze the results from the models and discuss the overall performance, challenges faced during the analysis, and key insights derived.

**6.1 Results**

**6.1.1 Count Plot of Top 10 countries by frequency of disasters**

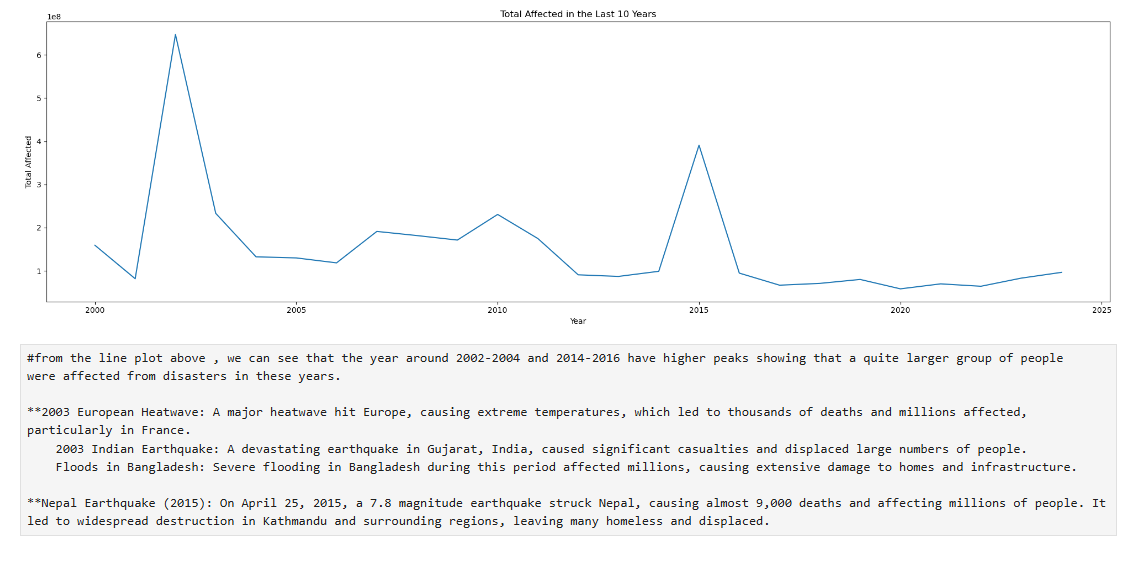
****

From the above Count Plot, we can observe that “China” has been the most unfortunate country in terms of calamities followed by “India” and “Indonesia”.

These countries are located in disaster-prone areas, such as the Pacific "Ring of Fire" (Indonesia) and tectonically active regions (China, India), exposing them to earthquakes, volcanic eruptions, cyclones, and floods.

With large populations living in disaster-prone areas, the frequency of disasters affecting large numbers of people is higher.

**6.1.2 Line Plot of Total number of people affected in last 10 years**

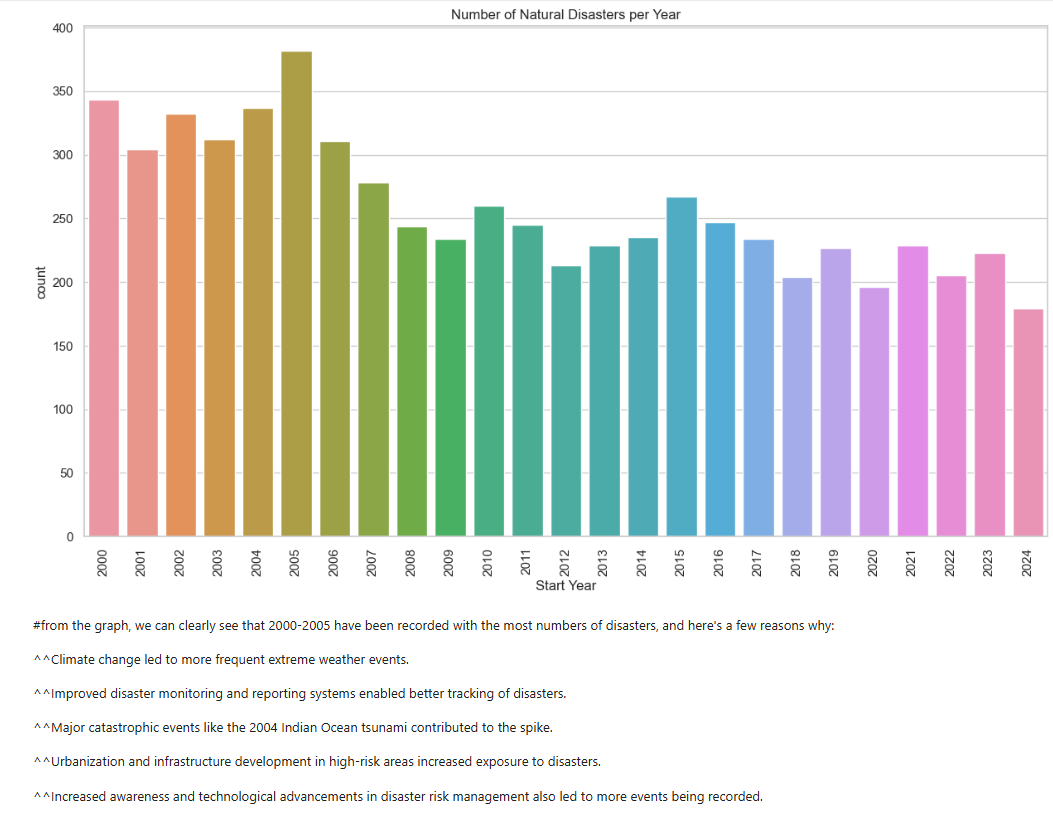


From the line plot above, we can see that the year around 2002-2004 and 2014-2016 have higher peaks showing that a quite larger group of people were affected from disasters in these years.

2003 European Heat wave: A major heat wave hit Europe, causing extreme temperatures, which led to thousands of deaths and millions affected, particularly in France. 2003 Indian Earthquake: A devastating earthquake in Gujarat, India, caused significant casualties and displaced large numbers of people. Floods in Bangladesh: Severe flooding in Bangladesh during this period affected millions, causing extensive damage to homes and infrastructure.

Nepal Earthquake (2015): On April 25, 2015, a 7.8 magnitude earthquake struck Nepal, causing almost 9,000 deaths and affecting millions of people. It led to widespread destruction in Kathmandu and surrounding regions, leaving many homeless and displaced.

**6.1.3 Count Plot of Number of disasters per year**

****

From the graph, we can clearly see that 2000-2005 have been recorded with the most numbers of disasters, and here are a few reasons why:

Climate change led to more frequent extreme weather events.

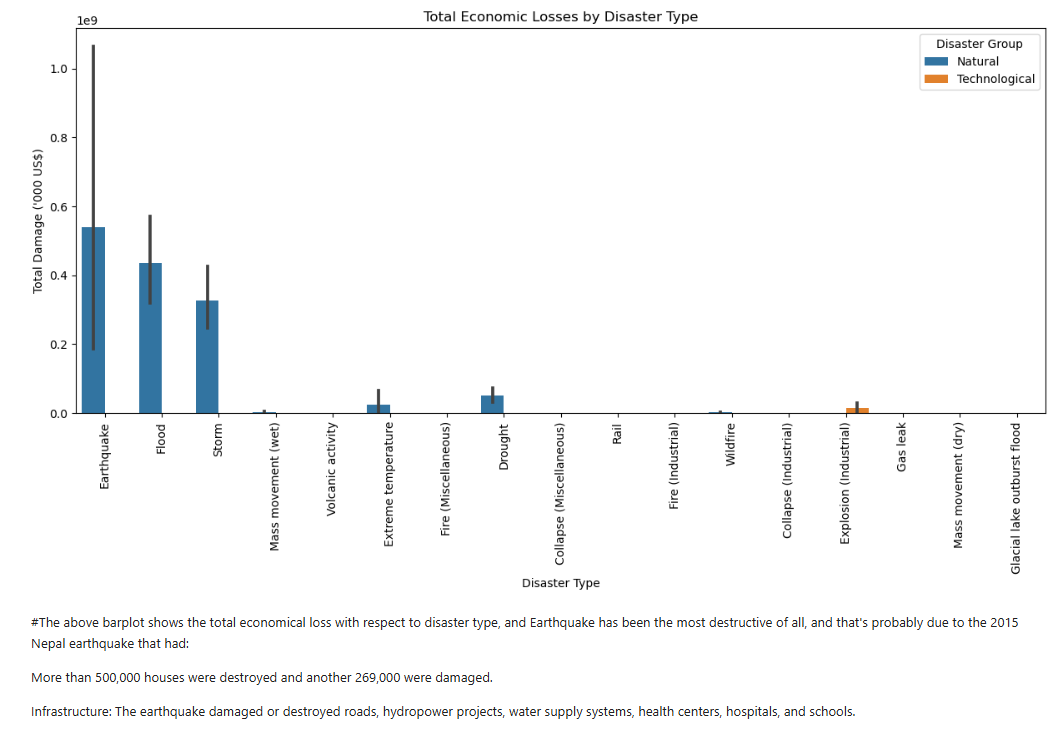
Improved disaster monitoring and reporting systems enabled better tracking of disasters.

Major catastrophic events like the 2004 Indian Ocean tsunami contributed to the spike.

Urbanization and infrastructure development in high-risk areas increased exposure to disasters.

Increased awareness and technological advancements in disaster risk management also led to more events being recorded.

**6.1.4 Bar Plot of Total Economic Losses by disaster type**

****

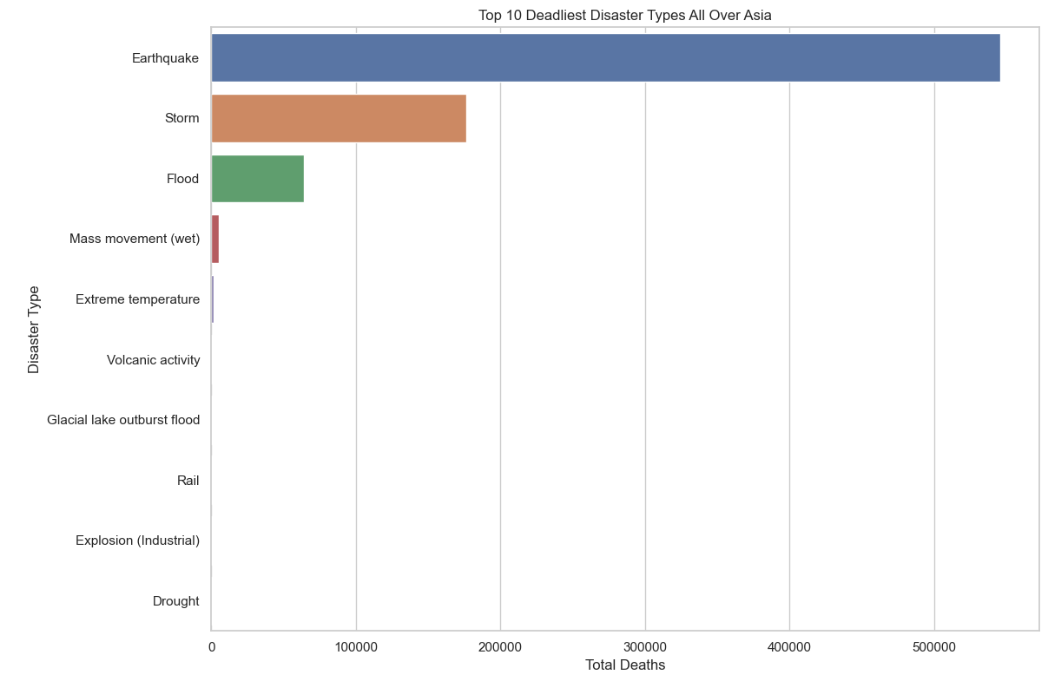
The above bar plot shows the total economic loss with respect to disaster type, and Earthquake has been the most destructive of all.

And that's probably due to the 2015 Nepal earthquake that had:

More than 500,000 houses were destroyed and another 269,000 were damaged.

The earthquake damaged or destroyed roads, hydropower projects, water supply systems, health centres, hospitals, and schools.

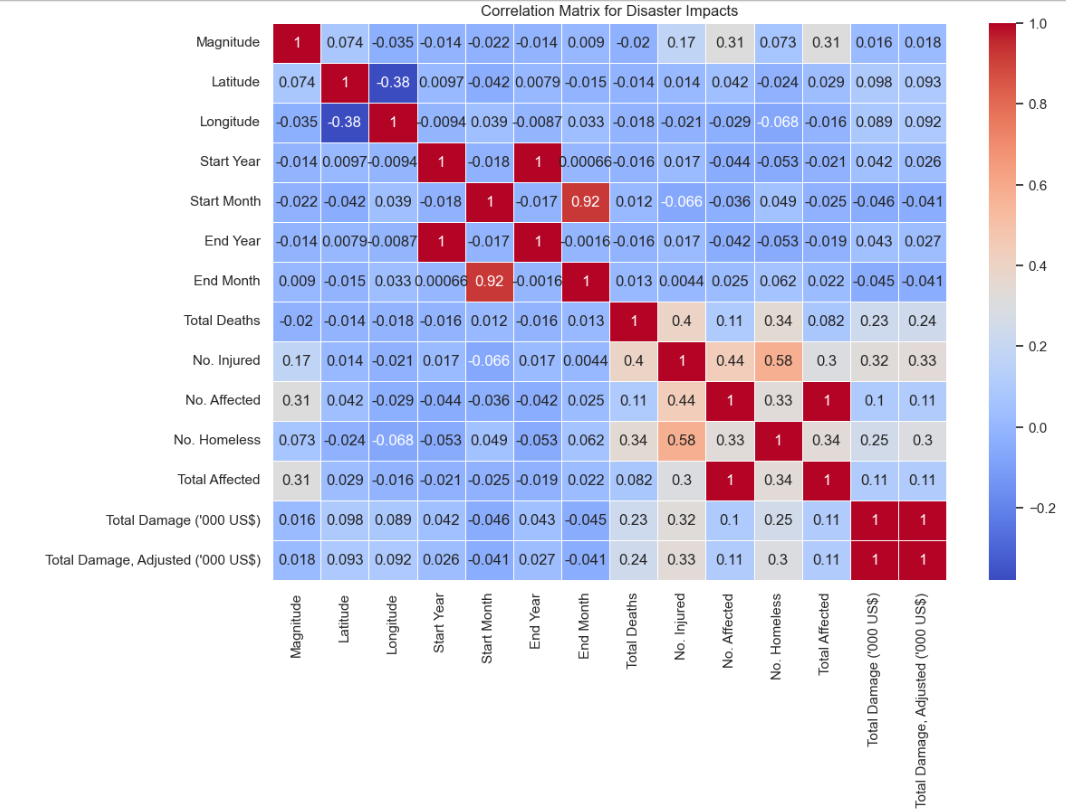
**6.1.5 Bar Plot of Top 10 deadliest disasters all over Asia**

****

From the given bar plot of top 10 deadliest disasters, the ‘Earthquake’ has been most unfortunate for humanity followed by ‘Storm’ and ‘Flood’.

Wet mass movement and Extreme Temperatures causing heat waves have also taken lives of many.

**6.1.6 Heat Map for Correlation**

****

Before going to the observation part first of all let's know these things:

Strength of Correlation:

0.7 to 1 (or -0.7 to -1): Strong correlation.

0.5 to 0.7 (or -0.5 to -0.7): Moderate correlation.

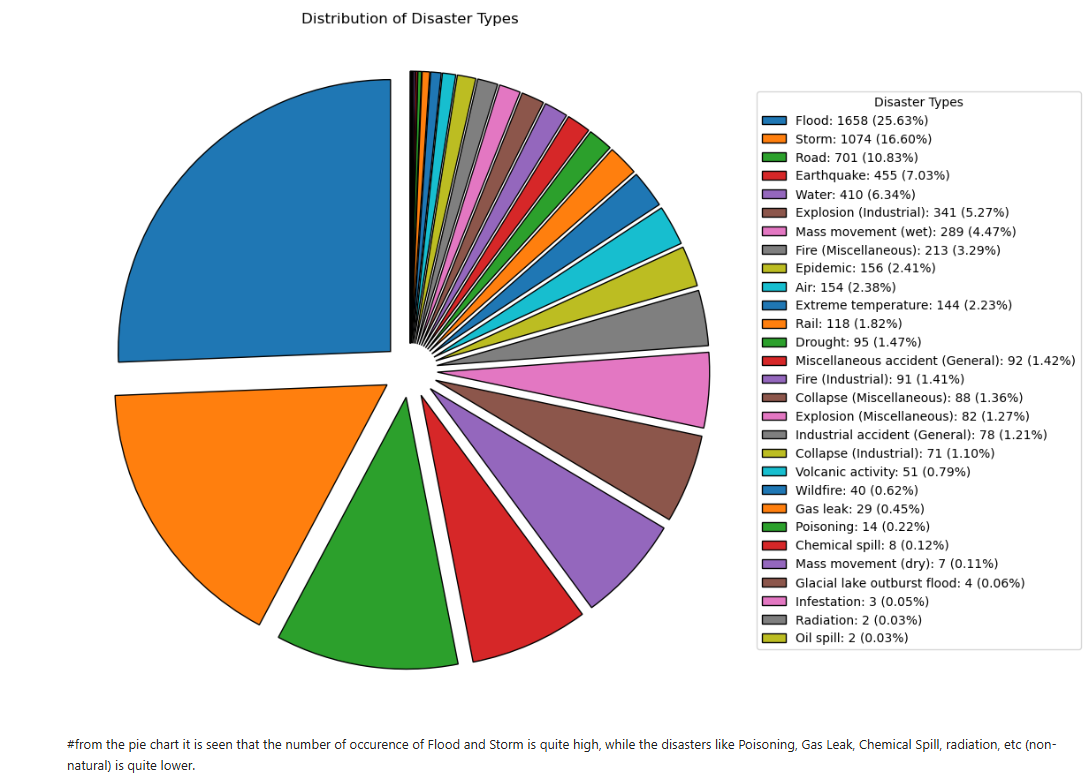
0.3 to 0.5 (or -0.3 to -0.5): Weak correlation.

0 to 0.3 (or -0.3 to 0): Very weak or no correlation.

Example Interpretation: Suppose you have Total Deaths and Total Damage ('000 US$) with a correlation of 0.8. This suggests a strong positive relationship, indicating that higher death tolls are associated with higher economic damage. If Total Deaths and Magnitude show -0.4, there’s a weak negative correlation, meaning that as Magnitude increases, Total Deaths tend to slightly decrease.

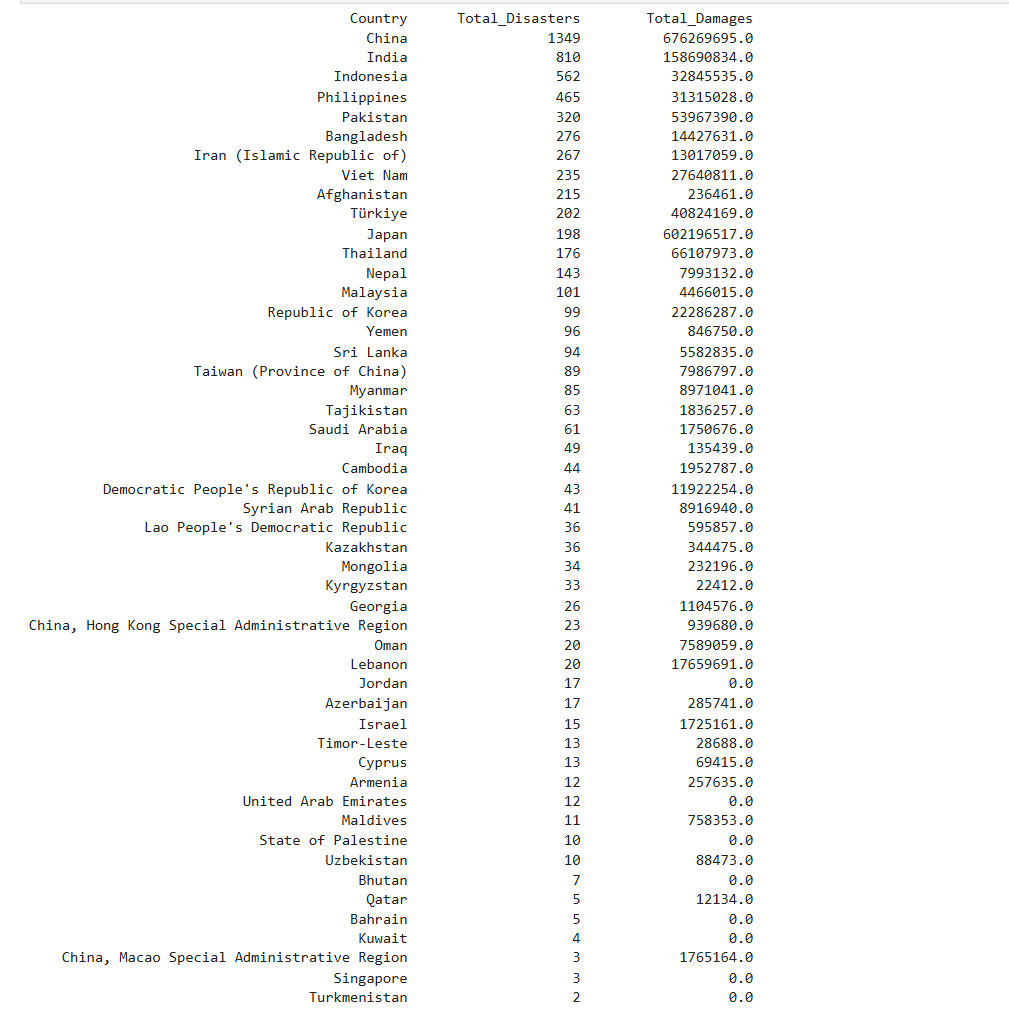
From the table, we can see "Total Deaths" and "No. Homeless" has a moderate correlation of 0.58, which shows disasters that result in higher fatalities also tend to cause higher numbers of homeless people. Similarly, "Total Deaths" and "No. Affected" has also a moderate correlation of 0.44, which shows disasters that result in higher fatalities means more number of people affected.

**6.1.7 Pie Chart distribution of Disasters**

****

From the pie chart it is seen that the number of occurrence of Flood and Storm is quite high, while the disasters like Poisoning, Gas Leak, Chemical Spill, radiation, etc. (non-natural) is quite lower. But still in terms of numbers and the damage it does to the environment, we must need to be careful and precautious while operating these types of industries and works. A small disaster can cause huge loss of lives, property and peace.

**6.1.8 Tabular chart of country disasters and total damages**



Based on the data presented in the table, the following observations can be made:

**China** stands out with the highest number of disasters (1349), but the total financial damages are relatively moderate at approximately 676 billion USD. This indicates that while China faces numerous disasters, many may not be as financially devastating as those in other countries.

**India** has the second-highest number of disasters (810) but experiences significantly lower financial damages (158.7 billion USD) compared to China. This suggests that while India faces frequent disasters, they may not always result in catastrophic economic losses.

**Indonesia** and **Philippines** follow with 562 and 465 disasters, respectively, with the Philippines incurring 31.3 billion USD in damages. This large financial loss highlights the severity of specific disasters in the Philippines, such as typhoons, even with fewer overall disasters compared to India and China.

**Pakistan** (320 disasters) and **Bangladesh** (276 disasters) both show considerable financial damages, though less than Indonesia and the Philippines. Pakistan’s total damages amount to 53.9 billion USD, while Bangladesh faces 14.4 billion USD in losses, likely due to recurring natural events like floods and storms.

**Iran** and **Vietnam** also show significant disaster counts with 267 and 235 events, respectively. Vietnam’s total damages are 27.6 billion USD, underlining the frequent and costly impacts of natural disasters in the country, while Iran's damages are 13 billion USD.

**Japan** (198 disasters) and **Thailand** (176 disasters) report higher financial losses despite having fewer disasters. Japan’s financial damages amount to 602 million USD, possibly from major earthquake events, while Thailand’s significant financial loss of 66.1 billion USD is likely due to catastrophic events like the 2011 floods.

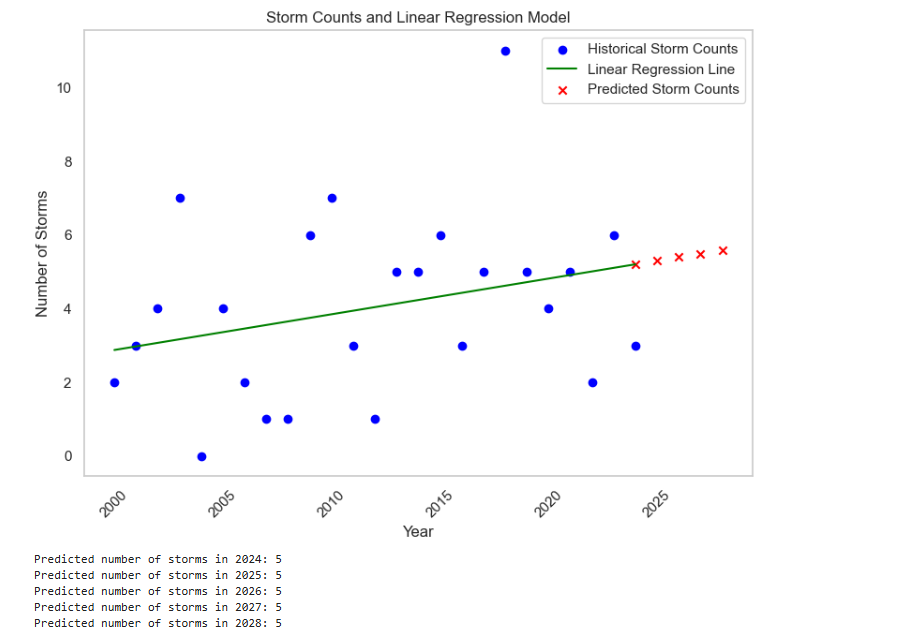
Smaller countries like **Sri Lanka** (94 disasters), **Taiwan** (89 disasters), and **Yemen** (96 disasters) report lower total damages, with Sri Lanka’s financial losses standing at 5.5 billion USD and Yemen at 846 million USD. The smaller number of disasters in these countries may indicate less frequent but still impactful events.

Countries with minimal disasters and damage: Countries such as **Qatar** (5 disasters), **Bahrain** (5 disasters), and **Kuwait** (4 disasters) report extremely low damages, potentially indicating effective disaster management strategies or fewer disaster-prone events in these regions.

**Singapore** and **United Arab Emirates** report very few disasters (3 each), with Singapore notably having no recorded financial damage. This could be due to geographical advantages, advanced disaster management, or very low occurrences of natural disasters.

**6.1.9 Regression Model (Prediction for Future Years)**

* The model predicts the number of storms in each future year (2024–2027) as 5. This constant prediction across multiple years could suggest:
  + A weak relationship between the Start Year and the Storm count in the data.
  + A lack of variation in the Storm count over the years, which may have led the model to fit a nearly flat line.



### **Historical Data Pattern**

* The scatter plot of historical storm counts shows fluctuations in the number of storms each year, with values spread between 0 and 10. This variability suggests that the number of storms does not follow a consistent trend each year but does have an overall slight upward tendency.

### **Trend Line (Linear Regression)**

* The green line represents the linear regression model fitted to the historical storm data. This trend line indicates a mild upward slope, suggesting that the frequency of storms may be gradually increasing over time.
* However, this increase is quite subtle, which might imply a low slope coefficient in the model. The overall variability in the data makes the trend difficult to capture accurately with a linear model, but it still provides a general upward direction.

### **Predictions for Future Years (2024–2028)**

* The predicted values for the years 2024, 2025, 2026, 2027, and 2028 are all around 5 storms per year, as indicated by the red "x" markers on the plot.
* Since the predictions show a consistent count of about 5 storms annually, it implies that the model, while identifying a slight upward trend, may not be capturing the fluctuations observed in historical data.
* This pattern suggests that the linear model might not fully account for the variability of storm counts over the years. It assumes a stable average increase, which may not accurately represent actual future counts if the pattern continues to vary widely.

### **Model Limitations and Considerations**

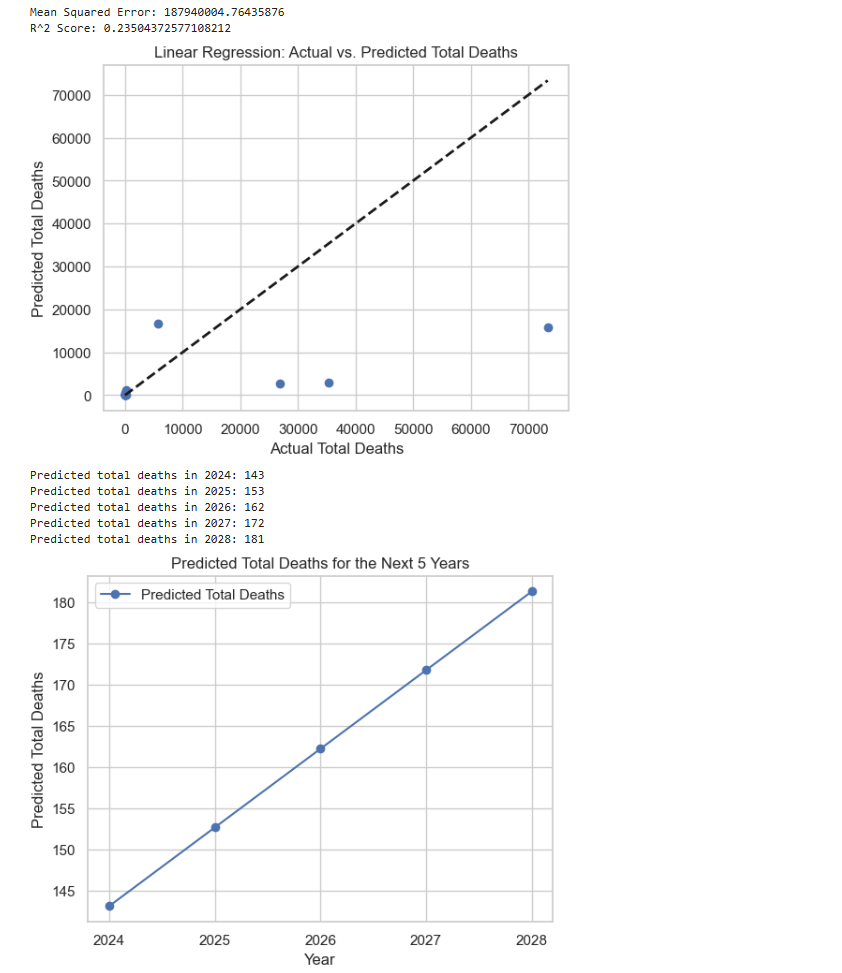
* **Simplicity of the Model:** A linear regression model assumes a straight-line relationship. In this case, given the variability of storm counts over the years, a linear model might oversimplify the relationship and potentially miss other patterns in the data.
* **Alternative Approaches:** Since storm counts fluctuate, using a different model, such as a time series analysis (e.g., ARIMA or exponential smoothing), could potentially provide a better fit and more accurate predictions if seasonality or other temporal patterns exist.
* **Further Data:** Additional variables, like climate indices (e.g., ENSO, Indian Ocean Dipole) or other meteorological factors, could be integrated to improve prediction accuracy, as storm counts could depend on more than just the year.

### **Interpretation of Results**

* Given the model’s predictions, the projected number of storms in the upcoming years appears stable around an average of 5 storms per year. However, the real-world implications of this might vary depending on policy, disaster management, and preparedness strategies that anticipate storm frequency changes.
* Although the trend is increasing slightly, a single linear model may not provide comprehensive insights for decision-making in regions prone to weather variability.

In summary, while the linear regression model suggests a mild upward trend in storm frequency for India, the predictions are relatively flat, likely due to the model's inability to capture the significant year-to-year variations. For more precise forecasting, considering alternative modeling techniques or additional data might be beneficial.

**6.1.10 Key Findings from model evaluation**

****

1. **Model Performance:**
   * **Mean Squared Error (MSE):** 18,794,000. This high value indicates a large average squared difference between predicted and actual values, suggesting the model may have high error and not be accurately predicting Total Deaths.
   * **R² Score:** Approximately 0.23. This low score indicates that the model explains only 23% of the variance in Total Deaths. A low R² score means that the model does not fit the data well, possibly due to factors such as high variance, noise in the data, or non-linearity between features and the target variable.
2. **Predicted vs. Actual Values Plot:**
   * The scatter plot shows predicted total deaths against actual total deaths, with a diagonal reference line representing perfect predictions. The large spread of points around this line reflects inaccuracies in the model's predictions.
   * There appear to be a few outliers (higher actual deaths with lower predicted values), which might affect the model’s accuracy.

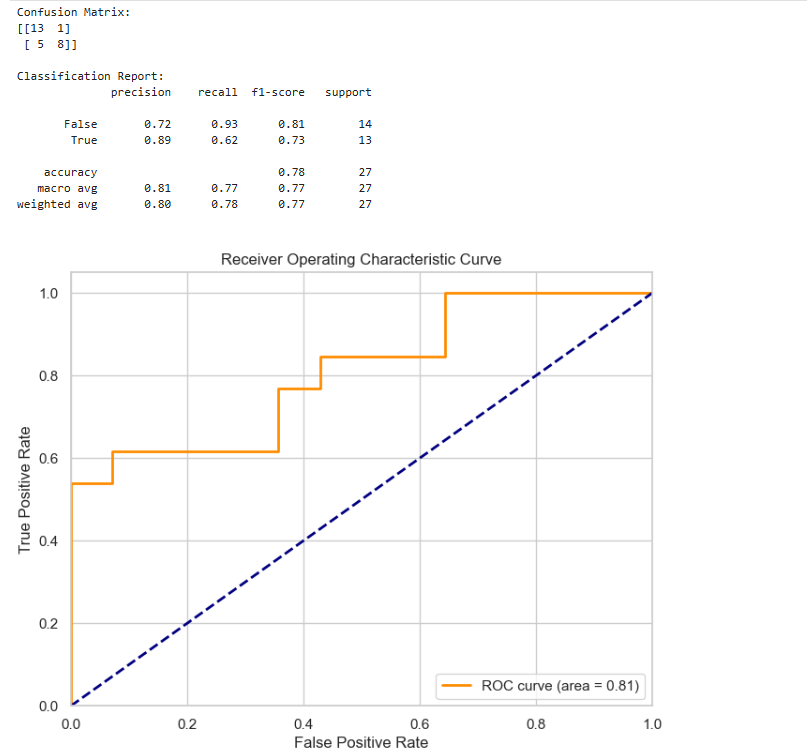
### Future Predictions (2024-2028)

Using estimated values for features (Total Affected, Total Damage, Adjusted, No. Injured, and No. Homeless), the model predicts total deaths for the next five years:

* **2024:** 143 deaths
* **2025:** 153 deaths
* **2026:** 162 deaths
* **2027:** 172 deaths
* **2028:** 181 deaths

The predictions show a steady increase in total deaths over the years, reflecting a linear trend based on the assumptions made about the feature values. However, since these values are estimated and the model's performance metrics are relatively low, these predictions should be interpreted with caution.

**6.1.11 Classifier Performance and ROC curve**



### **Confusion Matrix Analysis**

* **True Negatives (TN)**: 13 instances where the model correctly predicted "Low" total deaths.
* **False Positives (FP)**: 1 instance where the model incorrectly predicted "High" total deaths for an actual "Low" case.
* **False Negatives (FN)**: 5 instances where the model incorrectly predicted "Low" total deaths for an actual "High" case.
* **True Positives (TP)**: 8 instances where the model correctly predicted "High" total deaths.

### **Classification Report Analysis**

* **Precision**:
  + **False (Low Total Deaths)**: 0.72, indicating that 72% of the predictions for "Low" deaths were correct.
  + **True (High Total Deaths)**: 0.89, meaning 89% of the predictions for "High" deaths were correct.
* **Recall**:
  + **False (Low Total Deaths)**: 0.93, suggesting the model captured 93% of actual "Low" death cases accurately.
  + **True (High Total Deaths)**: 0.62, indicating that only 62% of the actual "High" death cases were correctly predicted.
* **F1-Score**: The F1-scores of 0.81 for "Low" and 0.73 for "High" indicate balanced performance, though the model does slightly better in identifying "Low" death cases.

### **Overall Accuracy**

* The model achieved an accuracy of **0.78** (78%), indicating a generally good performance but with room for improvement, especially in predicting "High" death cases.

### **ROC Curve and AUC Analysis**

* The ROC curve shows the trade-off between sensitivity (True Positive Rate) and specificity (False Positive Rate).
* **AUC (Area Under the Curve)**: The AUC of **0.81** suggests that the model has a fair ability to distinguish between "High" and "Low" death cases, as values closer to 1 indicate better performance.

### **Insights and Recommendations**

* **Class Imbalance**: The model performs better in identifying "Low" deaths, possibly due to an imbalance in the dataset or a threshold that favors "Low" predictions. Adjusting the threshold or using resampling techniques could help improve "High" case predictions.
* **Feature Engineering**: Adding more relevant features or engineering new ones may help improve the model's ability to differentiate between the two classes more effectively.
* **Alternative Models**: Consider experimenting with more complex models like Random Forest or Gradient Boosting, which may handle non-linear relationships and class imbalances better than logistic regression.

In summary, the model performs well but has some difficulty capturing "High" total death cases. Improving recall for "High" cases could enhance its utility for situations where predicting higher risk is crucial.

**6.2 Challenges Faced**

Some challenges may arise during the project, such as:

* Imbalanced Data: Certain disaster types might be underrepresented, leading to a biased model. In such cases, techniques like oversampling, under sampling, or using weighted models may help balance the dataset.
* Data Quality: Missing or inconsistent data may have required extensive cleaning, leading to delays or potential loss of important information.
* Feature Selection: Identifying which features were most important for prediction might have been challenging, especially when dealing with multiple attributes like geographical data, fatalities, and economic damage.

**6.3 Key Insights**

* Disaster Type Trends: By analyzing the data, we may uncover which types of natural disasters (e.g., hurricanes, earthquakes) are more frequent or have more significant consequences in terms of fatalities and economic impact.
* Regional Impacts: Insights into which regions are most prone to certain types of disasters can inform preparedness and mitigation strategies for those areas.

**CHAPTER 7: CONCLUSION AND FUTURE WORK**

**7.1 Conclusion**

The project has provided insightful analysis of natural disaster trends, helping to identify patterns in disaster occurrences and their impacts. Based on the dataset, the following key findings were made:

* Impactful Disaster Types: Through analysis, we identified which natural disasters (such as earthquakes, hurricanes, and floods) are the most deadly and costly. For example, while hurricanes tend to cause high fatalities, earthquakes often lead to much larger economic damage due to infrastructure destruction and the long-term recovery costs.
* Geographical Distribution of Disasters: The geographical analysis revealed certain regions that are more vulnerable to specific types of disasters. For example, coastal areas may experience more hurricanes, while regions near tectonic plate boundaries are more likely to suffer from earthquakes and tsunamis. This geographical data can assist local governments and agencies in prioritizing preparedness and mitigation measures.
* Trends Over Time: The data also revealed changes in the frequency and severity of disasters over time. Understanding these trends could help predict future disaster occurrences, guiding preparedness plans and allocating resources more effectively.
* Effectiveness of Predictive Models: The machine learning models, including classification algorithms, performed reasonably well in predicting disaster severity and type. The evaluation metrics, such as accuracy, precision, recall, and F1-score, indicated that while the model provided useful insights, there is still room for improvement, especially when dealing with imbalanced data and diverse disaster types.

This project successfully utilized historical disaster data to analyze trends, identify patterns in fatalities and economic damage, and build predictive models. The findings provide valuable insights for policymakers, governments, and humanitarian organizations to better prepare for and respond to future natural disasters. By understanding which regions are at high risk and which disaster types have the most severe consequences, authorities can focus their resources on strengthening disaster preparedness in the most vulnerable areas.

**7.2 Future Work**

While the project has successfully analyzed and modeled disaster data, there are several avenues for future work and improvements:

* Data Enrichment: The current dataset could be further enriched by incorporating additional variables such as climate data, demographic information (e.g., population density), and infrastructure data. This would allow for more comprehensive models that could predict both the impact of disasters and the capacity of affected areas to recover.
* Advanced Modeling Techniques: Future models could use more sophisticated machine learning approaches like Random Forests, XGBoost, or deep learning models (e.g., neural networks) to enhance prediction accuracy, especially when dealing with complex and non-linear relationships in the data.
* Real-Time Predictions and Alerts: Implementing real-time data integration would allow for more proactive disaster response. For example, if disaster data from various regions becomes available in real time, the system could offer immediate predictions about potential impacts, helping authorities take preventive measures faster.
* Scenario Simulations and Risk Assessments: The project could be extended by creating detailed disaster scenario simulations to explore how different regions could be affected by multiple types of disasters. These simulations could assess the effect of variables like population growth, infrastructure development, or climate change on disaster risk.
* Collaboration with Disaster Relief Agencies: A partnership with disaster relief organizations could help improve the practical application of these models. By integrating real-world feedback from disaster responders, the models could be refined to better address the real-time needs of affected communities.

**7.3 Final Thoughts**

Natural disasters will continue to pose significant risks to lives and economies around the world. By leveraging historical data and advanced analytical techniques, we can enhance disaster preparedness and response efforts. Although there are challenges in data quality, model accuracy, and data imbalance, the insights gained from this project provide a strong foundation for future work in disaster management. The application of machine learning models in this context has the potential to significantly improve decision-making, mitigate disaster risks, and ultimately save lives.

The continuous refinement of predictive models, coupled with richer datasets and the integration of real-time disaster monitoring systems, will pave the way for a more resilient and prepared global community.