International Disasters

Ela Apetrei

15/04/2023

The Why & How

- **Source:** EM-DAT, the International Disaster Database
  https://public.emdat.be/

- **Why?** Climate change

- **Timeline:** 1903 - 2022
  The dataset contains data from 1903 to 2022, but the amount of data for the earlier years is restricted likely due to limited data collection.

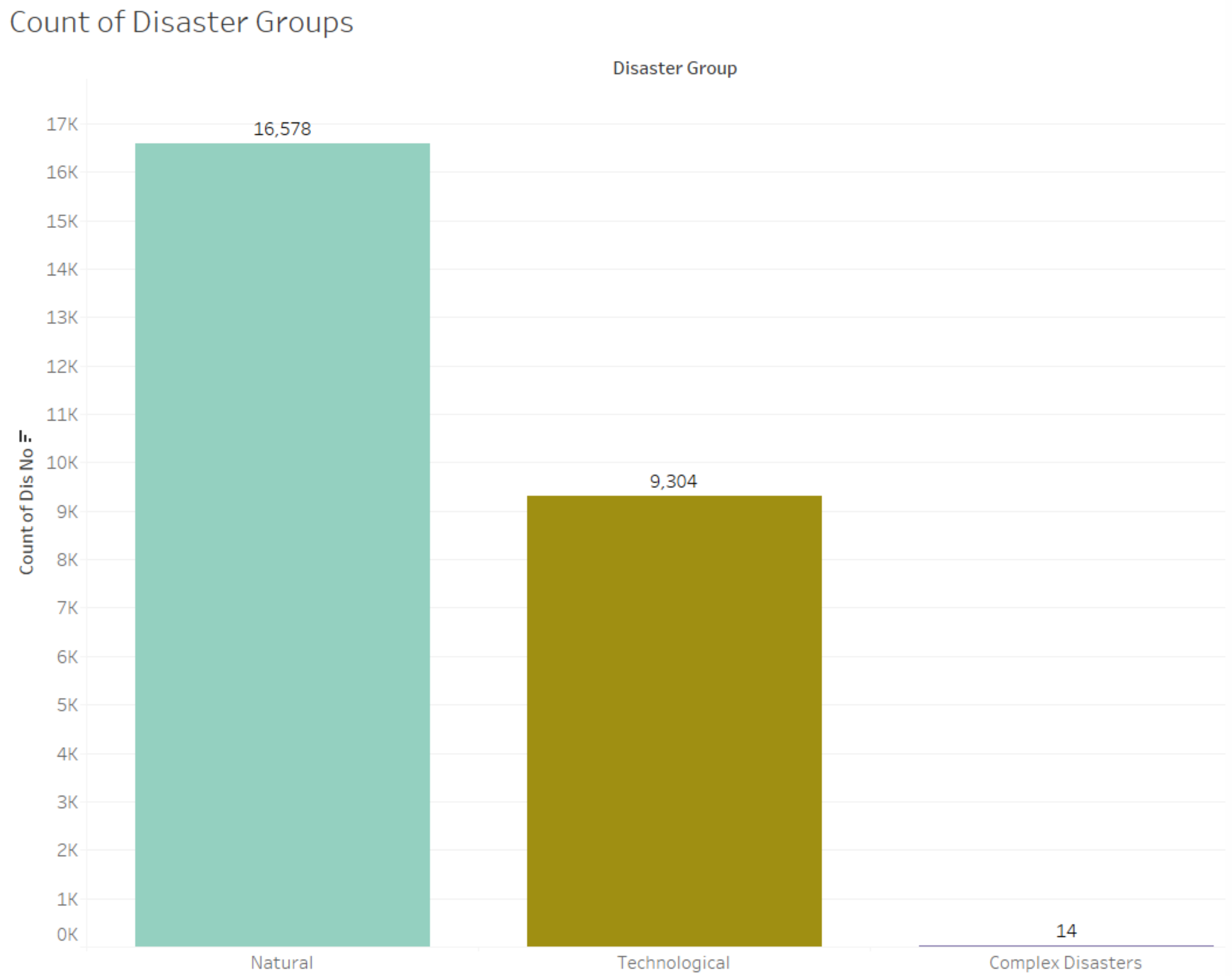
- **Challenges:** Real Data
  Cleaning the data was a labor-intensive, large number of null values.

- **Content:** Findings & Outcome of the ML algorithm
  First, we will showcase our findings, then we will discuss the outcomes of the Machine Learning algorithm.

Data Analysis Summary

1. Count of Disaster Groups
2. Split by Group & Type
3. Impact Over Time
4. Geographic Heatmap
5. Disaster Duration by Type
6. Seasonality of Disasters

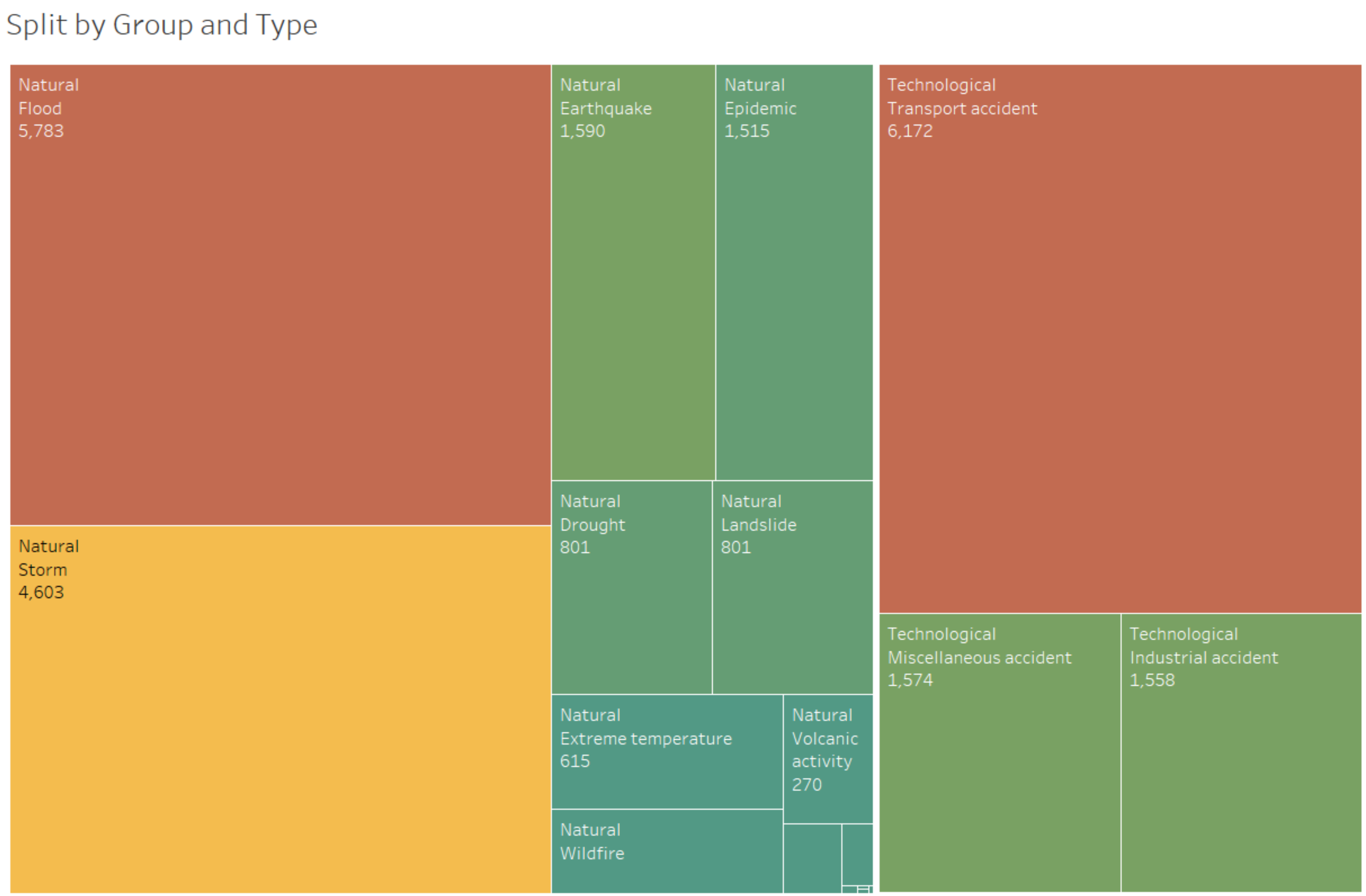
1. Count of Disaster Groups



According to our data, the number of natural disasters is almost twice that of technological disasters.

- **Natural Disaster:** Events caused by nature
- **Technological Disaster:** An event caused by a malfunction of a technological structure and/or some human error in controlling or handling the technology
- **Complex Disasters:** Major famine situation for which the drought were not the main casual factor

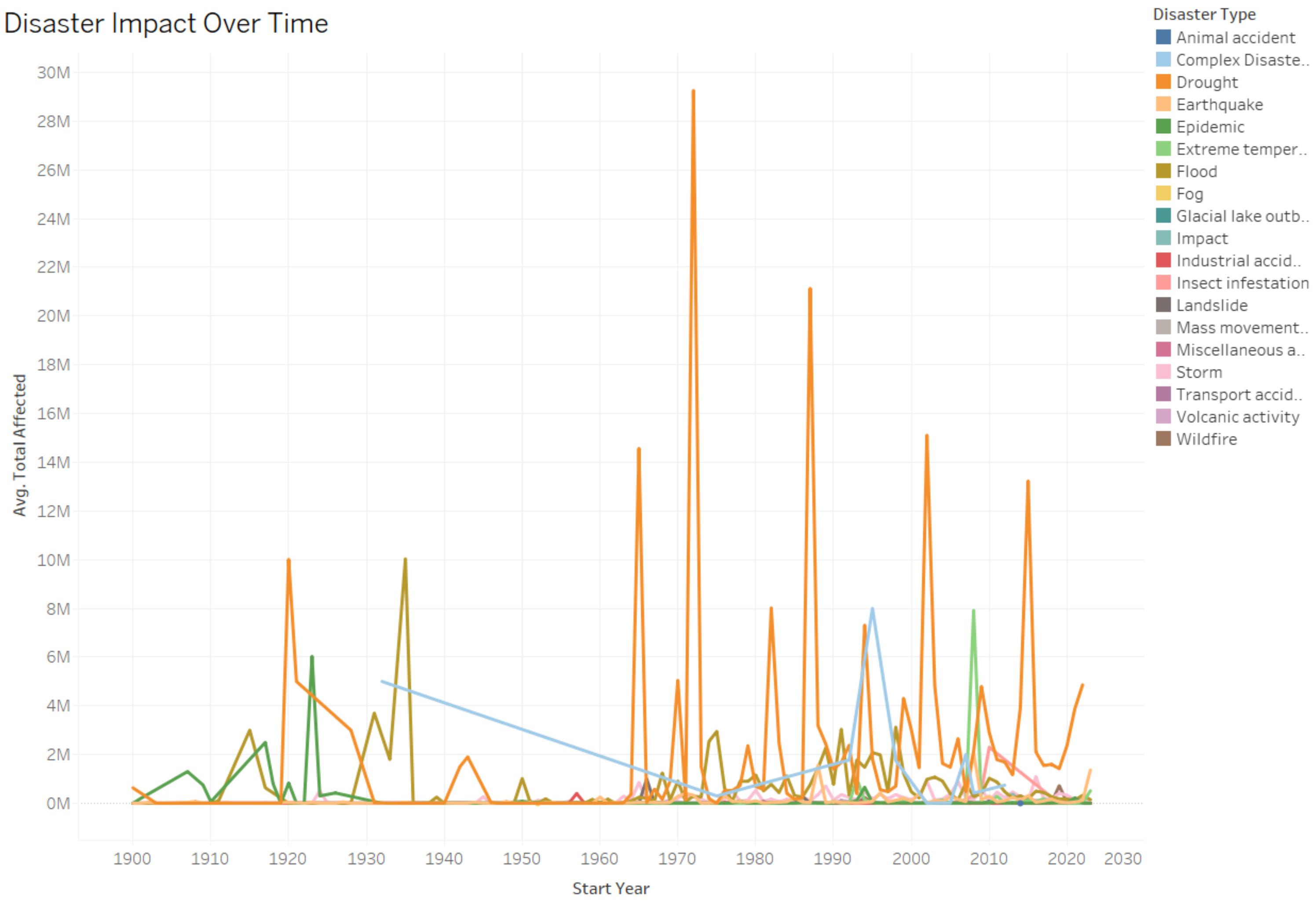
2. Split by Group & Type



In this treemap visualization, we have categorized disasters by their respective groups and types.

- **Natural floods** occur with striking frequency, nearly matching the incidence of **transport accidents**, which fall under the Technological Disasters group.

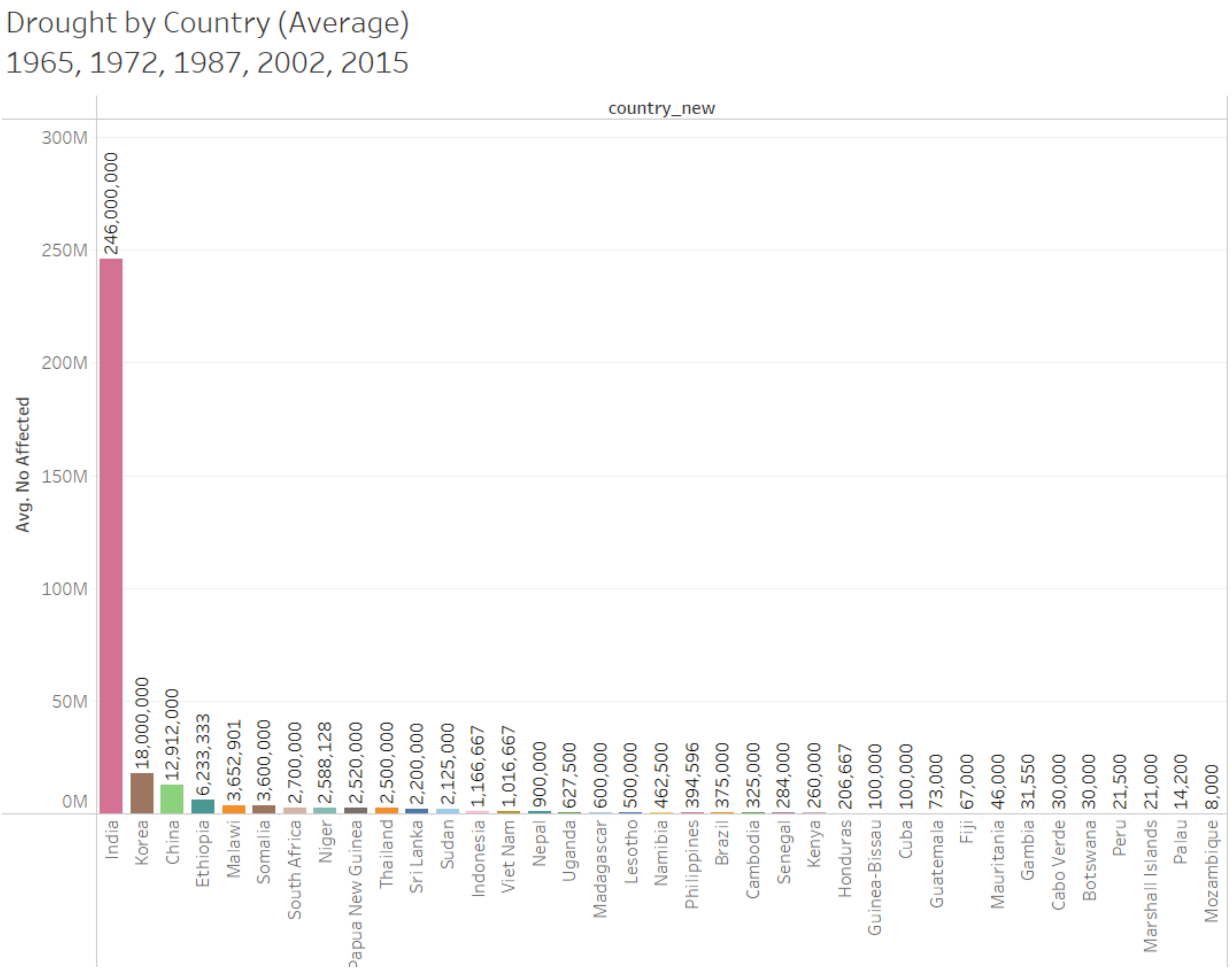
3. Impact Over Time



The area or line chart demonstrates the impact of disasters over time, measured by the number of affected individuals.

- We can observe the particularly devastating effects of drought on affected populations during specific years
- Notably, the years 1965, 1972, 1987, 2002, and 2015 stand out as having the highest impact on people due to drought.

4. Most impacted countries by Drought



The bar chart illustrates the countries that have experienced the most severe consequences of drought in the peak years.

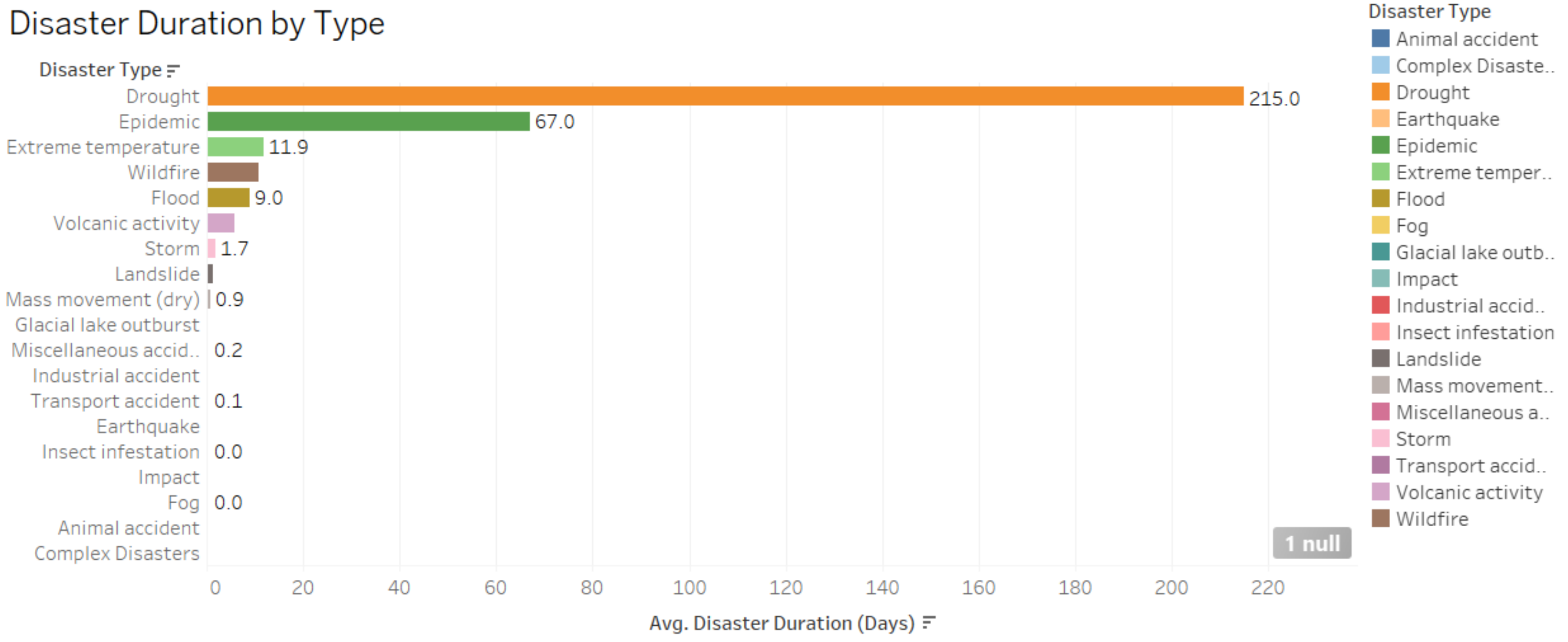
- It has been reported that India, Korea, and China are the three nations affected the most by this calamitous situation.

5. Geographic Heatmap

The map visualization displays the geographic distribution of disasters worldwide. We can better understand the regions that are most vulnerable

- As illustrated in the map visualization, China and the United States have experienced the highest number of disasters in our dataset.
- The reasons for this can be attributed to their vast geographical size, diverse climates, and population densities.

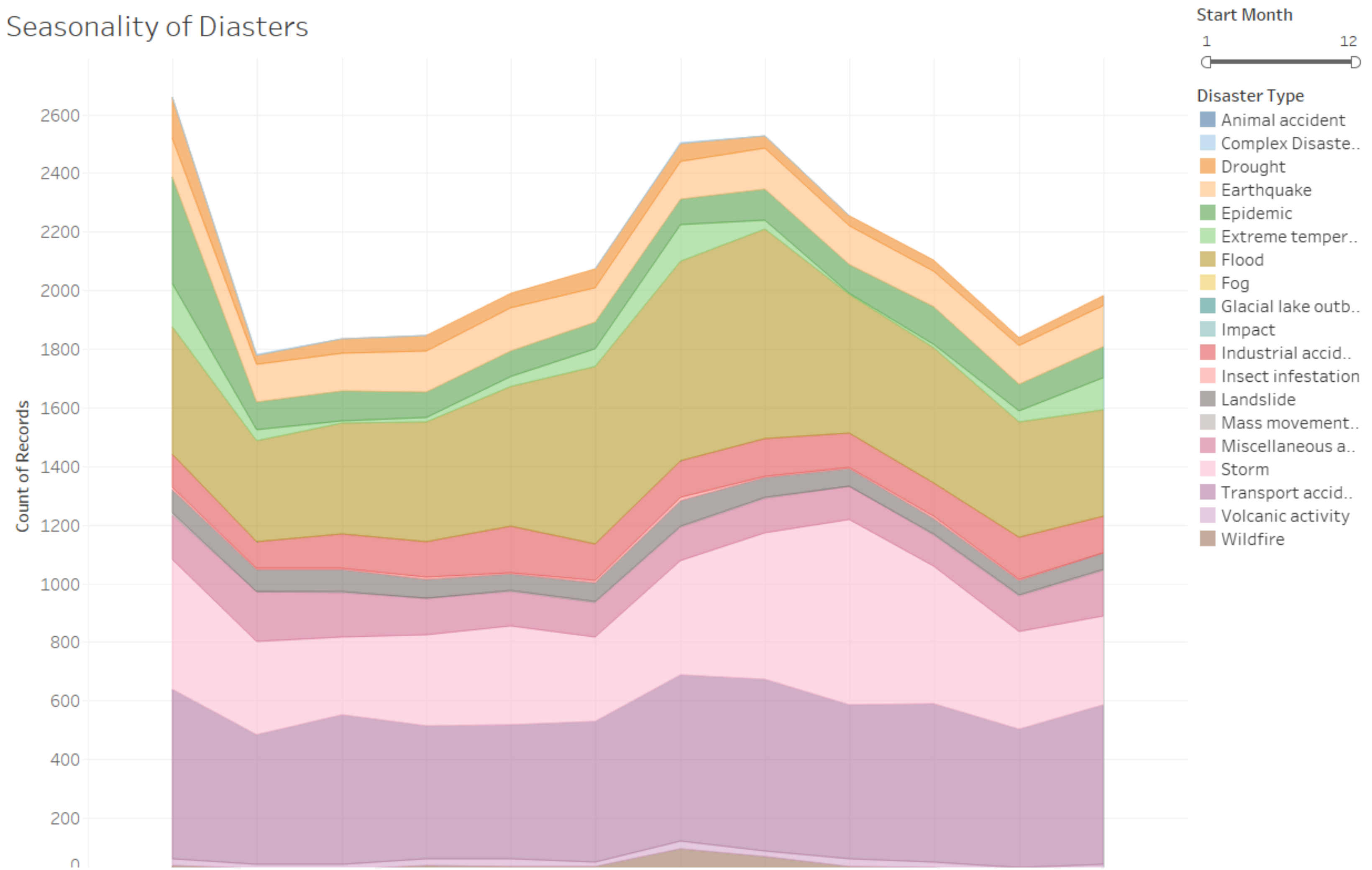
6. Disaster Duration by Type



This box plot illustrates the duration of different disaster types. By comparing the average duration of various disasters, we can better comprehend the potential long-term consequences

- Drought, epidemic, and extreme temperature disasters tend to have the longest durations among all disaster types.

7. Seasonality of Disasters



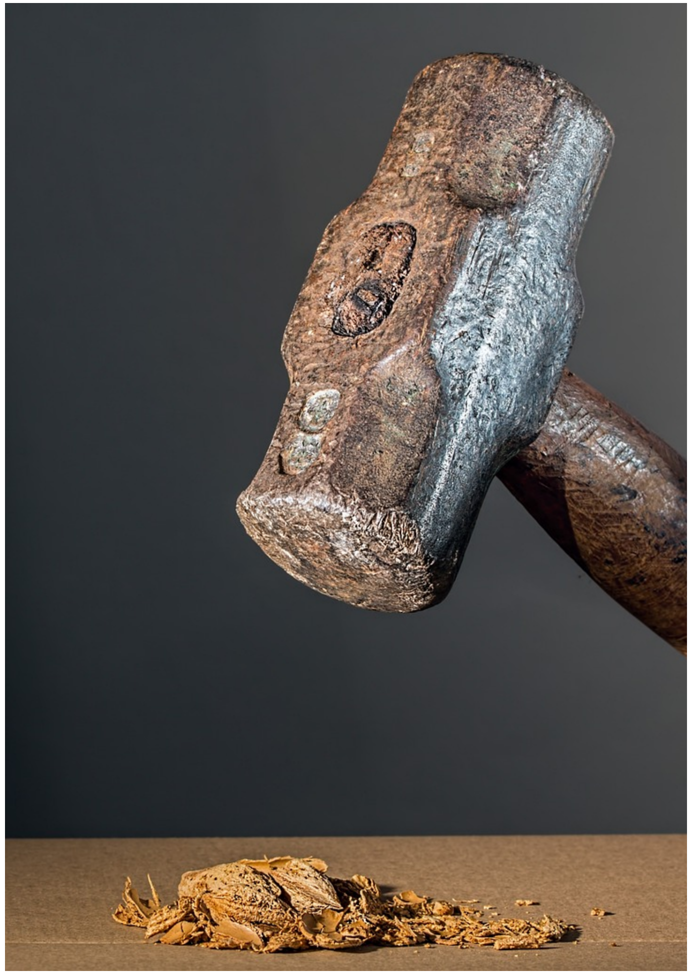
The number of disasters occurring in each month, highlighting the seasonality of various disaster types.

- **Floods** tend to have a peak season in **August**, which can be attributed to factors such as increased rainfall during the summer monsoon season in many regions, as well as melting snow and ice in the northern hemisphere
- Peak season for **storms,** particularly hurricanes and typhoons, occurs in **September**. This coincides with the period when ocean temperatures are at their warmest, providing the necessary energy for the development and intensification of tropical cyclones

Machine Learning Algorithm

1. Data Cleaning
2. Approach
3. Results
4. Suggestions

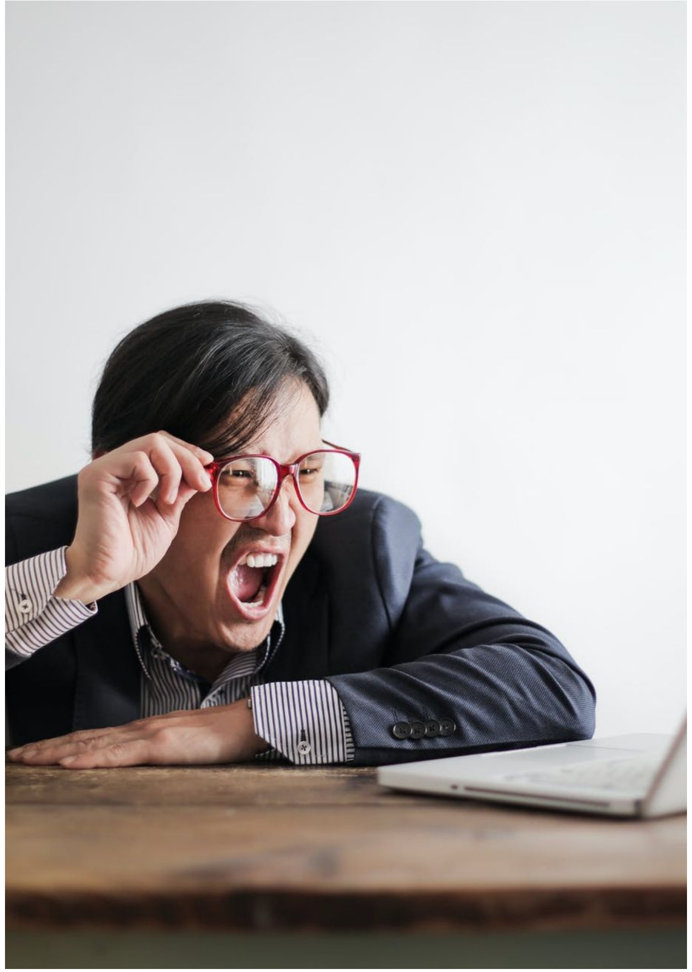
1. Data Cleaning



- **Dropped a few unnecessary columns**
- **Filling gaps with 'None' for some columns (associated_dis)**
- **Filling gaps with logic**

  total_deaths + no_injured + no_affected + no_homeless = total_affected

2. Approach



- **Model I - Dropping Columns + Encoding**
  - Removing the columns which have a high collinearity with our target but also has a lot of null values. Did not use any scaling methods

- **Model II - Simple Imputer + Standard Scaler**
  - Simple Imputer (mean) for "aid_contribution", "reconstruction_costs_adj", "insured_damages_adj"
  - Feature Scaling - Standard Scaler. Some columns have a wide range of values

- **Model III - Iterative Imputer + Normalization**
  - Iterative Imputer (mean) for "aid_contribution", "reconstruction_costs_adj", "insured_damages_adj"
  - Feature Scaling - Normalization

3. Results

Our Model III (Iterative Imputer and Normalization) is performing the best but still inadequately, evidenced by the low **R-squared** score, large **Mean Absolute Error** (MAE), and **Mean Squared Error** (MSE) values. We need to improve the model in order to reduce the MAE and MSE, as predicting the mean alone performs better than our current model.

Model	R2 Score	Mean Absolute Error	Mean Squared Error
Model I	-85,656	77,859,149	1,936,950,617
Model II	-24,138	45,049,101	1,028,237,083
Model III	-3,330	15,569,532	381,939,064

4. Suggestions

- Use the XGBoost technique since it is suitable with...

  1. Sparsity-aware learning. The tree learning algorithm is designed to handle sparse data and missing values effectively

  2. Missing value imputation. It doesn't explicitly impute missing values, it automatically learns the best direction to send a sample with missing values during the tree construction process

  3. Default direction for missing values: XGBoost allows users to specify a default direction when building trees. Useful in cases where the optimal direction for missing values is known or can be guessed beforehand

- Improve the quality of disaster-related data through

  1. Better coordination among organizations;

  2. Improved data collection and sharing technologies;

  3. Implementation of standardized methodologies for collecting and reporting data.

Thank you

Thank you!