College of Computer and Information Sciences
Department of Information Technology

IT 362 Course Project
Semester-2, 1446H

IT362: Principles of Data Science

< PREDINA >

Phase 2: Data Collection, Processing, Cleaning, and Exploratory Data Analysis (EDA)

(note: everything mentioned in the report was also explained in the jupyter notebook)

Prepared by

Student name	Student ID
Nouf AlMansour	444200525
Sara AlOqiel	444203016
Shahad AlMutairi	444200935

Supervised by:
Dr. Mashael AlDayel

1 Primary Data

Overview

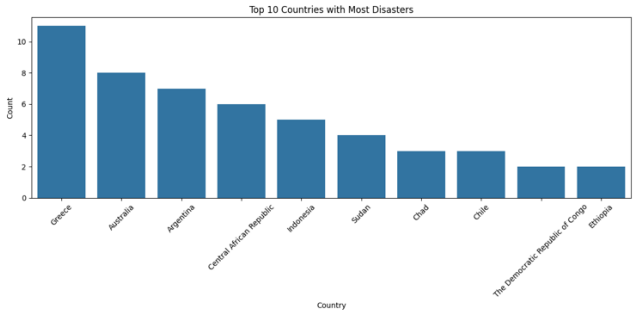
The primary dataset contains less than 100 entries with 15 columns, including categorical and numerical variables. Key columns include:

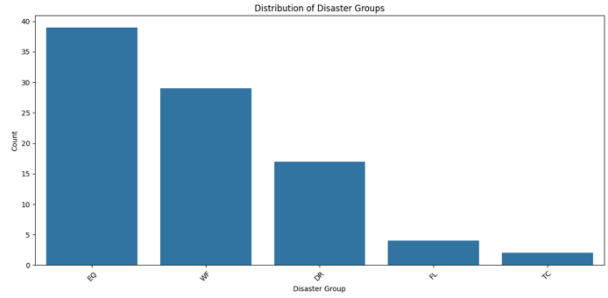
- **Event Name**: 20 non-null values (high missingness)
- **Country**: 91 non-null values
- **ISO Code**: 90 non-null values
- **Disaster Group**: Fully populated
- **Geographical Coordinates**: Latitude and Longitude available for all entries
- **Temporal Data**: Start and End Year, Month, and Day
- **Magnitude & Unit**: Present in most cases
- **Losses**: Completely missing

Key Observations

- There is substantial missing data in Event Name and Losses, potentially requiring imputation or removal.
- The dataset covers a diverse range of disaster events across multiple countries.
- Temporal data is complete, enabling trend analysis over time.
- Geographical data allows for spatial visualization of disasters.

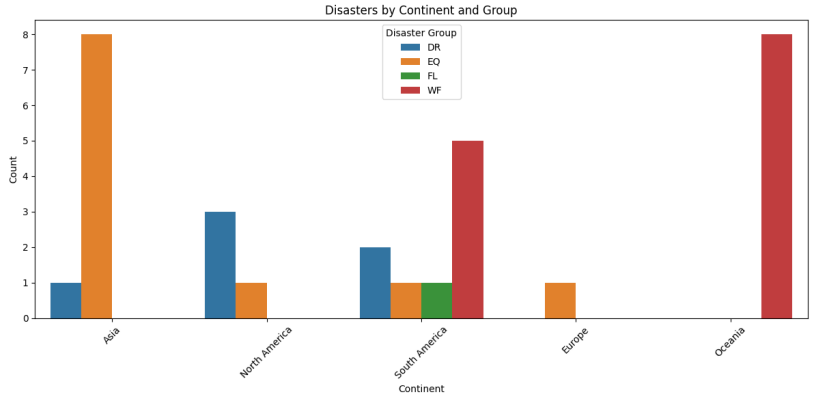
Visualizations



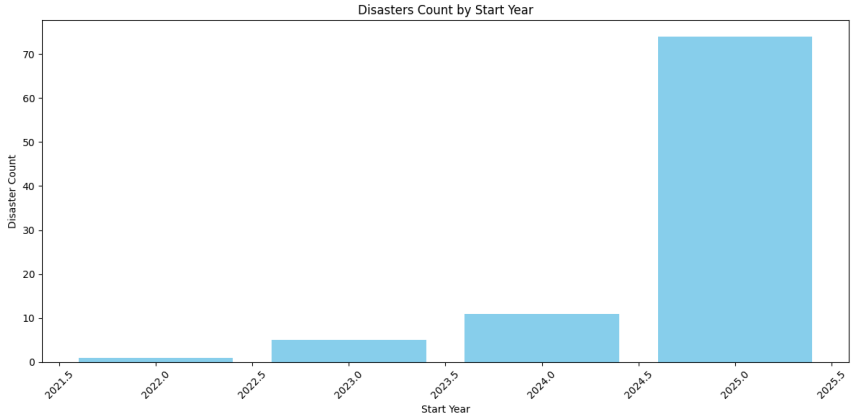


Countries with the most occurance in diasters are Greece followed by Australia and Argentina

From the dataset exploration, we observed that **EQ (Earthquake)** is the most frequent disaster type, followed by **WF (Wildfire)**.



- **Wildfires (WF) occur most frequently in Oceania**, followed by South America indicating a significant climate-driven disaster pattern in the region.
- **Earthquakes (EQ) are most prevalent in Asia**, highlighting the region's vulnerability to seismic activity.
- Other disaster groups show less occurrences across continents, but these two stand out as the most frequent occurrences.



2 Secondary Data

Overview

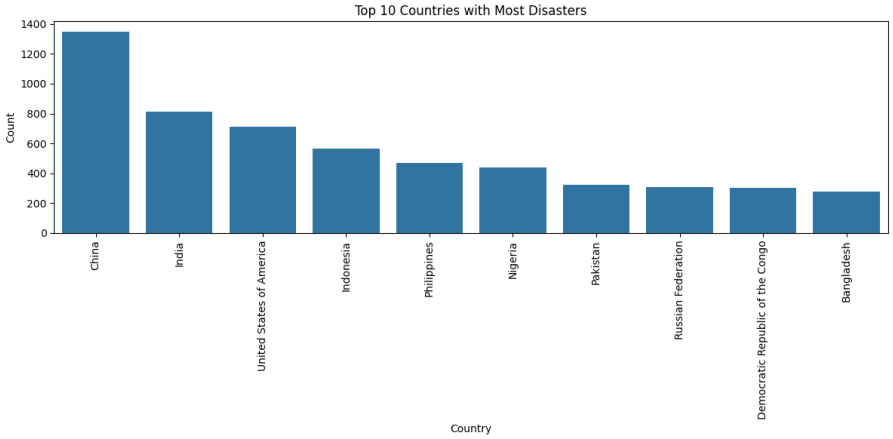
The secondary dataset provides pre-existing information that complements the primary dataset. This includes:

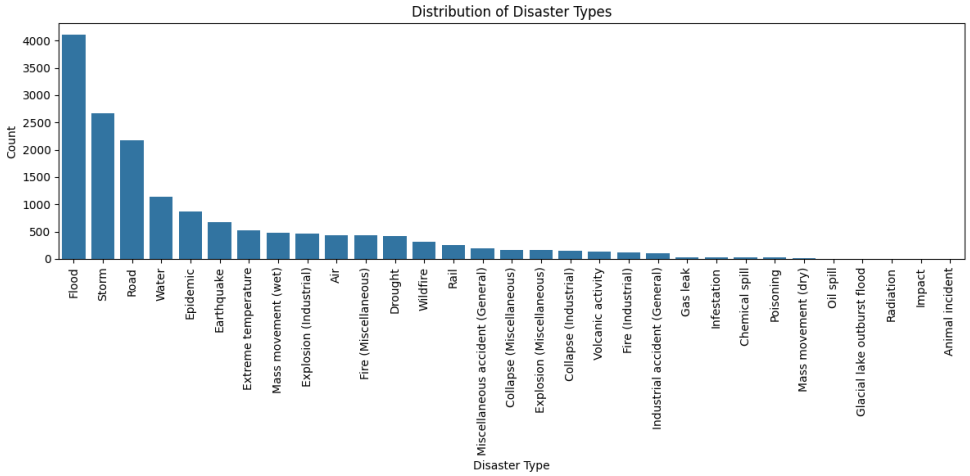
- **Historical Disaster Records**: Potentially overlapping with primary data
- **Economic Impact Data**: Used to infer losses where missing in the primary dataset
- **Geographical Information**: Enhancing location-based insights

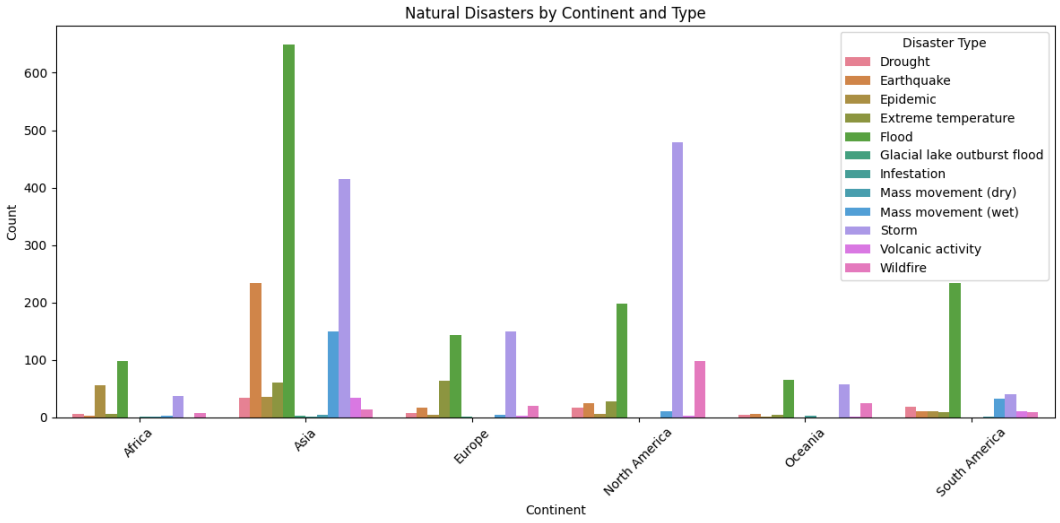
Key Observations

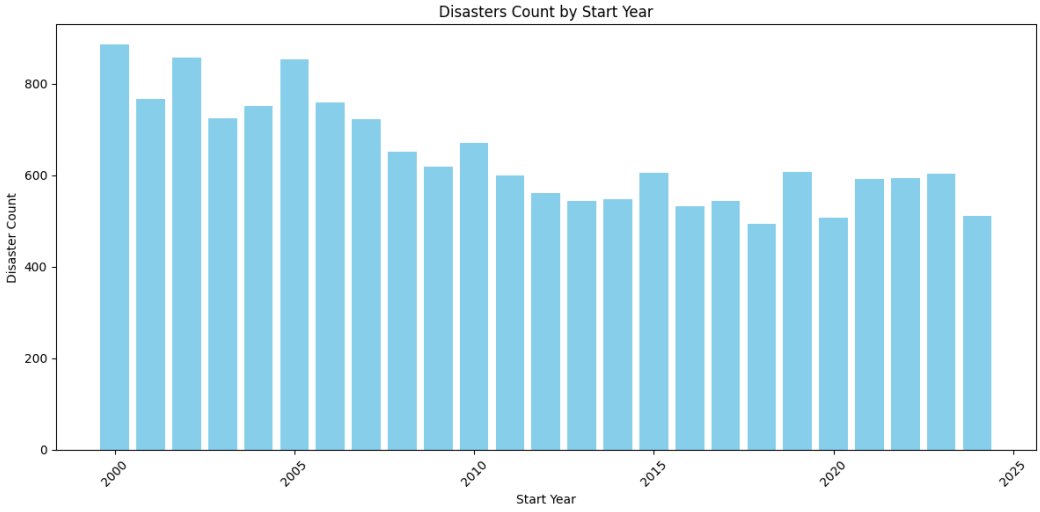
- Some discrepancies exist between recorded disasters in primary and secondary datasets.
- Losses information is more complete, allowing potential imputation for primary data.
- Differences in categorical classifications may require standardization.

Visualizations









Upon comparing the **primary dataset** with the **secondary dataset**, we observed a significant difference in the **number of rows, attributes, and overall structure**. To ensure a **comprehensive analysis**, we proceeded with **EDA on both datasets** to extract key insights and compare them effectively.

However, for a more **unified and enriched analysis**, we integrated both datasets into a single **combined dataset**. This allowed us to capture a broader perspective while maintaining consistency in our exploration. The EDA was then primarily focused on the **integrated dataset**, ensuring that all relevant patterns and trends were thoroughly examined.

3 Comparison: Primary vs. Secondary Data

Alignments

- Both datasets capture similar disaster events and geographical data.
- Temporal trends largely align, reinforcing the reliability of time-based analyses.

Discrepancies

- Missing values in the primary dataset contrast with more complete data in the secondary source.
- Differences in classification schemes for disaster types.
- Some inconsistencies in recorded magnitudes.

Comparison

Aspect	Primary Data (API)	Secondary Data (Historical)
Row Count	<100	16,000+
Disaster Count	5 types	Over 10 types
Time Period	2020 and later	1900s and later
Most Affected Country	Greece (10 disasters)	China (1,200+ disasters)
Most Frequent Disaster	Varies by country; Asia → Earthquakes Oceania → Wildfires	Asia → **Floods** (most frequent overall)
Disaster Type Mapping	Limited to **news-reported disasters**	Covers a **wider variety** of disasters

4 Summary of New Insights and Hypotheses

Insights

- The primary dataset requires imputation, especially for financial losses.
- The secondary dataset can act as a validation tool for historical disaster trends.
- Discrepancies highlight the importance of standardization in disaster reporting.

Hypotheses

- The missing Losses column in the primary dataset can be estimated using secondary data trends.
- Different disaster classification schemes affect analytical conclusions and should be harmonized.
- Temporal and geographical patterns can help predict disaster occurrences and economic impacts.