DSCD 611 — Programming for Data Scientists I

Topic:

Few Events, Many Lives: Exploratory Data Analysis of Global Disaster Impacts

Group C4

Members:

- Julius Aban Odai (team lead) 22424770
- Frederick Tettey-Lowor 22424676
- Michael Kusi-Appiah 22424580

1. Topic Studied

This project examined **global natural disasters and their human impacts** using data from the Emergency Events Database (EM-DAT). The study focused on identifying patterns in disaster frequency, disaster types, geographic distribution, and human consequences such as fatalities and affected populations.

A central focus of the project was the concentration of disaster impacts. Rather than treating all disasters as equally consequential, the analysis explored whether a small number of extreme events accounted for a large share of total human losses worldwide.

To guide the exploratory analysis, the project addressed the following research questions:

- How has the frequency of natural disasters changed over time?
- Which disaster types occur most frequently worldwide?
- Which disaster types cause the greatest total deaths?
- Which disaster types affect the largest populations?
- How does disaster impact vary by geographic region?
- Are disaster impacts concentrated in a small number of catastrophic events?
- How have disaster impacts, measured by deaths and affected populations, changed over time?

These questions structured the data preparation, analysis, and interpretation of results presented in the subsequent sections.

2. What Is Known About the Topic

Existing research shows that natural disasters are reported more frequently over time, due to a combination of climatic factors, population growth in vulnerable areas, and improvements in data collection and reporting systems. Prior studies also show that disaster types differ significantly in both frequency and severity. Floods and storms tend to occur most often, while earthquakes and droughts, though less frequent, often result in higher mortality or long-term disruption.

It is also well established that disaster impacts are unevenly distributed across regions. Countries with limited infrastructure and disaster preparedness capacity often experience higher death tolls and greater long-term effects. Research further indicates that a small number of extreme disasters typically drive aggregate statistics, meaning that averages may obscure the disproportionate influence of rare but catastrophic events.

3. Why the Topic Is Interesting, Relevant, or Important

Understanding global disaster impacts is important for disaster risk reduction, humanitarian planning, and public policy. Governments and international organizations rely on such information to allocate limited resources for preparedness, response, and recovery.

This topic is particularly interesting because it challenges common assumptions. Disasters that occur most frequently are not always those that cause the greatest harm. In addition, the concentration of impacts in a small number of events suggests that focusing on extreme risks may be more effective than planning based on average outcomes. Exploratory data analysis provides a useful approach for uncovering these patterns and highlighting inequalities in global disaster vulnerability.

4. Description of the Data Used

The dataset used in this project was obtained from the Emergency Events Database (EM-DAT), a globally recognized disaster database maintained by the Centre for Research on the Epidemiology of Disasters (CRED). The full EM-DAT database contains records of more than 22,000 mass disaster events worldwide from 1900 to the present. However, the analysis focused on disaster events occurring from the year 2000 onward, comprising over 16,000 disaster records. This restriction was applied because disaster data prior to 2000 is more susceptible to reporting bias due to limited data collection and documentation practices in earlier periods.

Each observation in the dataset represents a single disaster event and includes variables related to the following dimensions:

- **Shape:** 16,657 rows × 48 columns
- **Time:** start year, month, and day
- **Classification:** disaster group, type, and subtype
- **Geography:** country, region, and subregion
- **Human impact:** total deaths, total affected, injured, and homeless populations

Although the dataset includes variables related to economic damage and reconstruction costs, these fields contain substantial missing values as a result of uneven and incomplete reporting across countries and time periods. As a result, the analysis primarily focused on disaster frequency and human impact metrics, which were more consistently reported and therefore more appropriate for exploratory analysis.

5. How the Project Was Done (Tools and Methods)

The project was conducted using **Python** within a Jupyter Notebook environment. Key libraries included:

- Python for data analysis
- Jupyter Notebook for development and documentation
- Pandas and numpy for data cleaning and aggregation
- Matplotlib for visualization
- Streamlit for an interactive dashboard

Data preprocessing involved inspecting missing values, checking for placeholder or sentinel values, and constructing usable time variables from partial date fields. No imputation was performed for highly sparse variables; instead, the analysis was restricted to reliable fields.

Exploratory data analysis techniques were applied, including:

- Time-series analysis of disaster frequency
- Grouped aggregations by disaster type and region
- Ranking analyses for deaths and affected populations
- Concentration analysis to assess the contribution of top events

An interactive dashboard was developed to allow dynamic filtering by year, region, disaster type, and impact thresholds.

6. Results and Societal Impact

The analysis identified clear patterns in global disaster occurrence and impact. Recorded disaster events peaked in the early 2000s, followed by a modest decline, likely reflecting changes in reporting practices rather than a sustained reduction in risk.

Floods were the most frequent disaster type, followed by storms and transport-related events. However, the most frequent disasters were not the most severe. Earthquakes caused the highest number of deaths, with extreme temperature events ranking second, while floods and droughts affected the largest populations worldwide.

Substantial regional inequality was observed, with Asia bearing the highest burden in both total deaths and affected populations. The analysis also showed strong concentration of impacts, as approximately half of all reported disaster-related deaths were attributable to the ten deadliest events. No clear long-term decline in disaster impacts was observed, as annual totals were driven by rare but catastrophic events.

From a societal perspective, these findings emphasize the importance of prioritizing preparedness for high-impact, low-frequency disasters. Targeted investments in disaster preparedness and early warning systems, particularly in highly affected regions such as Asia, could significantly reduce future loss of life.

7. Team Contributions

The project was completed collaboratively, with responsibilities divided as follows:

- **Julius Aban Odai (22424770):** Data analysis, coding, and dashboard development.
- **Frederick Tettey-Lowor (22424676):** Data visualization, code review, and preparation of the presentation slides.
- **Michael Kusi-Appiah (22424580):** Data analysis review, report writing, and organization of findings

8. Reflections on the Project

This project demonstrated the value of exploratory data analysis in understanding complex, real-world phenomena. Working with disaster data highlighted challenges such as missing values, reporting biases, and the need for careful interpretation of aggregated statistics.

A key takeaway is that impactful insights can be generated without predictive modeling by thoughtfully exploring structure, distributions, and inequalities within the data. The project also reinforced the importance of aligning analytical questions with data quality constraints.

Future extensions could include predictive modeling, integration of socioeconomic indicators, or deeper geographic analysis to further explore drivers of disaster vulnerability.

Acknowledgements

Generative AI tools were used as a support resource for code debugging, syntax clarification, and improving the clarity and organization of the report. All analysis, code, and interpretations were independently developed and verified by the authors.

Appendix:

Data source: https://public.emdat.be — requires login

Github repository: https://github.com/jlsodai/c4_dcsd611

Hosted dashboard: https://c4dcsd611.streamlit.app/

Charts and figures

Figure 1 — Number of Recorded Disaster Events per Year (Q1)

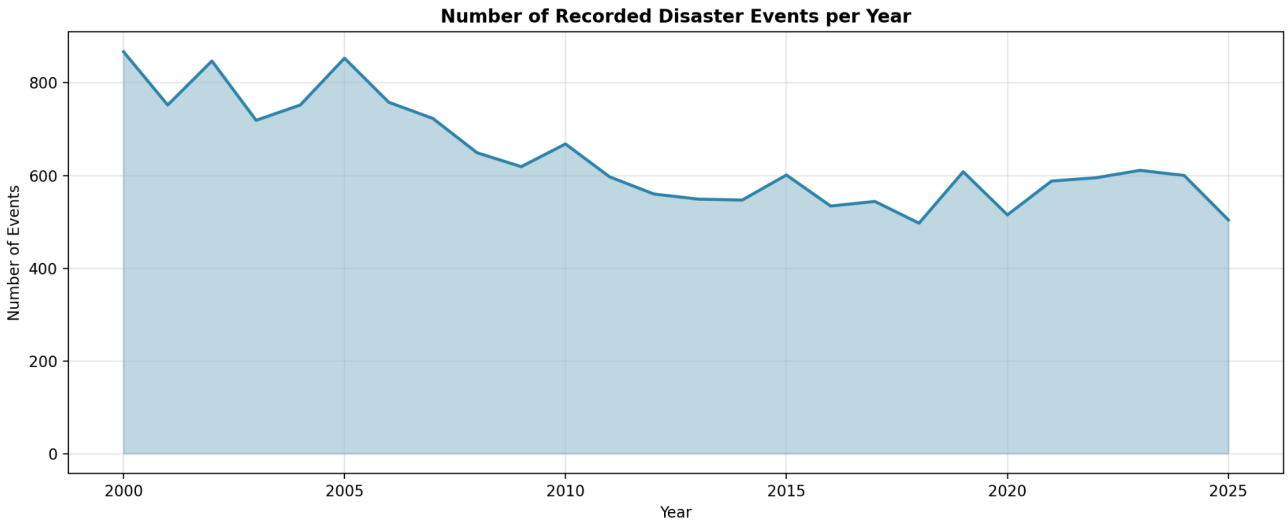

Figure 2 — Top 10 Disaster Types by Frequency (Q2)

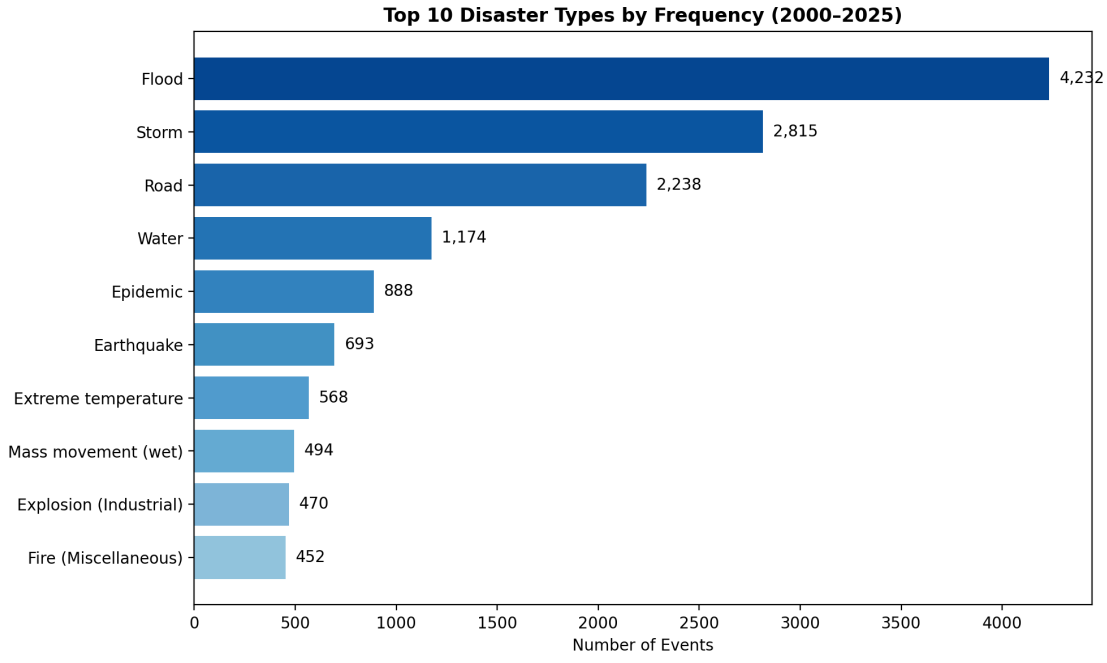

Figure 3 — Top 10 Disaster Types by Total Deaths (Q3)

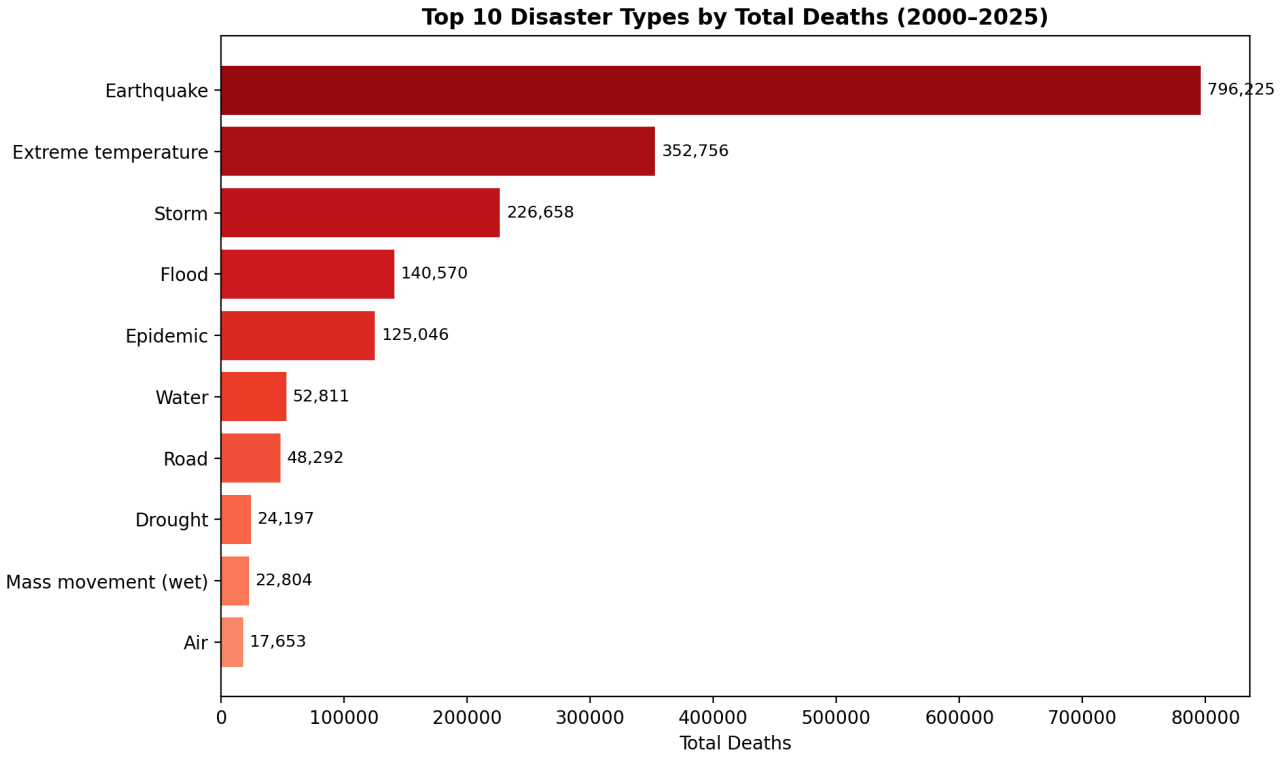

Figure 4 — Top 10 Disaster Types by Total Affected Population (Q4)

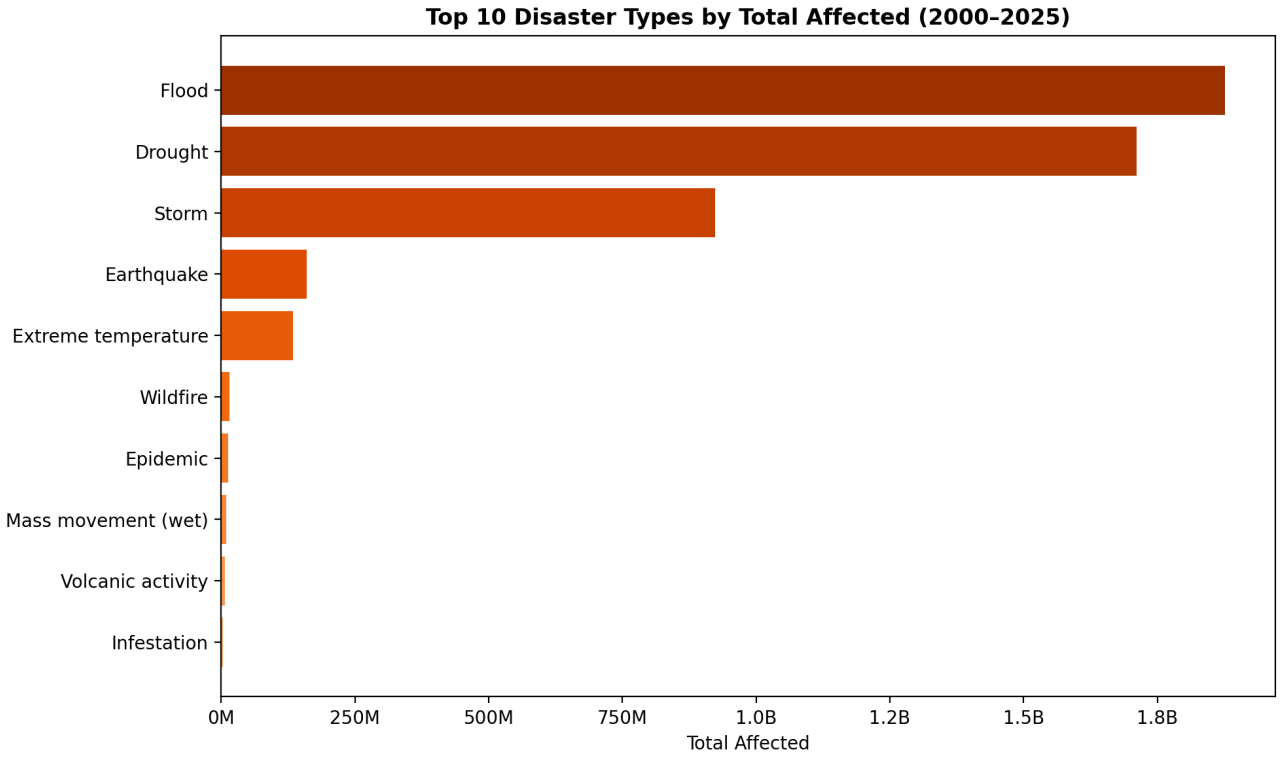

Figure 5 — Total Disaster Impact by Region (Q5)

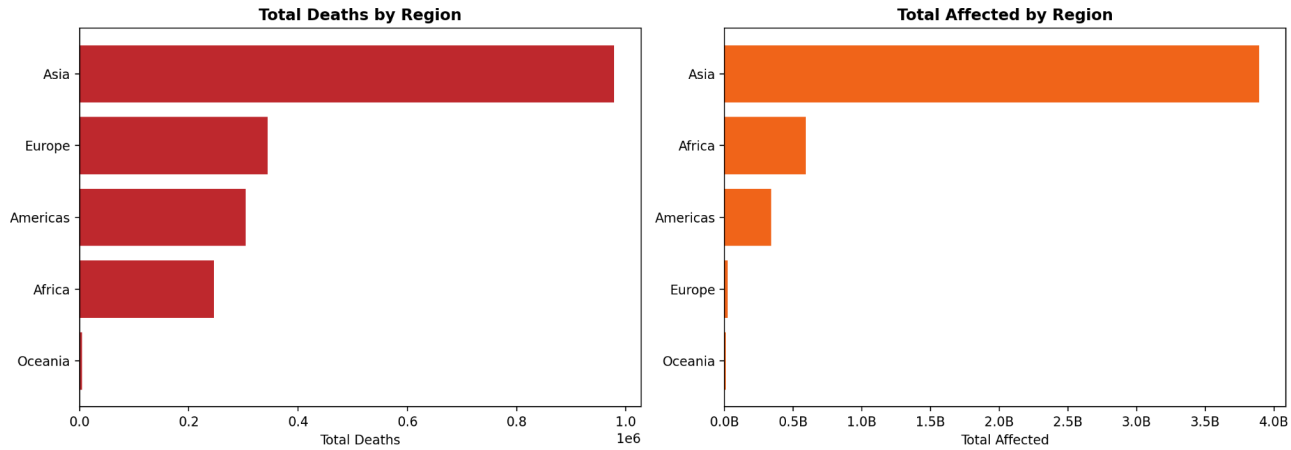

Figure 6 — Concentration of Disaster Mortality (Q6)

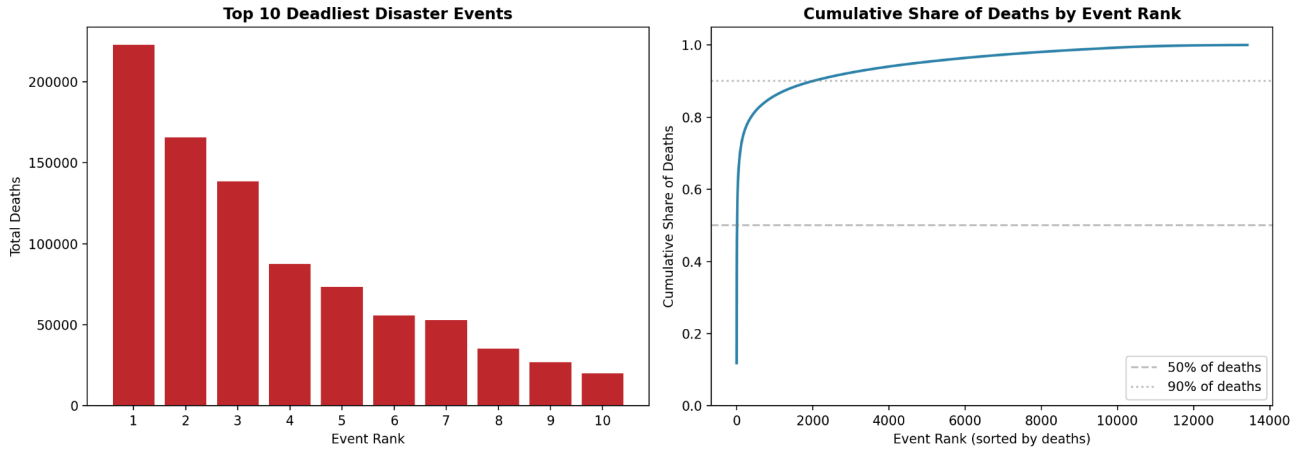

Figure 7 — Trends in Disaster Impacts Over Time (Q7)