HW2 Report

Link to visualization: https://info4310-hw2-disasters.netlify.app/

Data Description

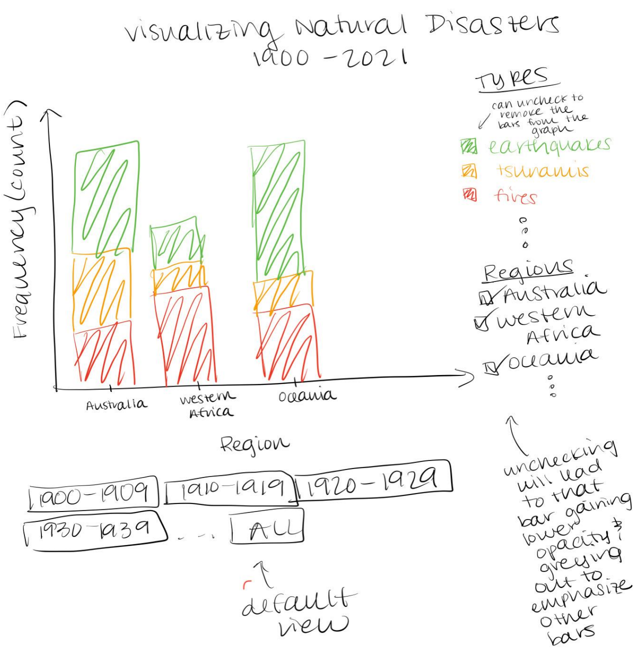
Our data is sourced from Kaggle (link to dataset) and provides information about the occurrences of natural disasters worldwide from 1900 to 2021. The original dataset was preprocessed to better fit our purposes.

Preprocessing included selecting relevant columns from the original dataset and recategorizing certain regions - for instance, “Polynesia”, “Melanesia”, and “Micronesia” were changed to “Oceania Islands”, and “Russian Federation” was combined into “Western Asia”. This was done to reduce the number of regions, as some had noticeably fewer occurrences, also making the information easier to digest. We knew early on that we wanted to present time as decades instead of individual years, so for ease of use in our visualization code, the “year” values were converted to the corresponding “decade” values beforehand.

While the dataset included many columns, many of them were missing specific data (e.g., most values for latitude and longitude were empty). We decided to explore the year, region, and disaster type fields, which were largely intact and covered a considerable range in all three aspects. A good use case is using it to analyze general trends at a larger scale.

Our interactions provide a certain amount of granularity for users to manipulate, but at the same time stay at an “overview” level of the data. This fits with our overall direction of showing trends instead of going too deep (for instance, individual disaster events are not presented), and also appeals to the strengths of the dataset itself.

Storyboards

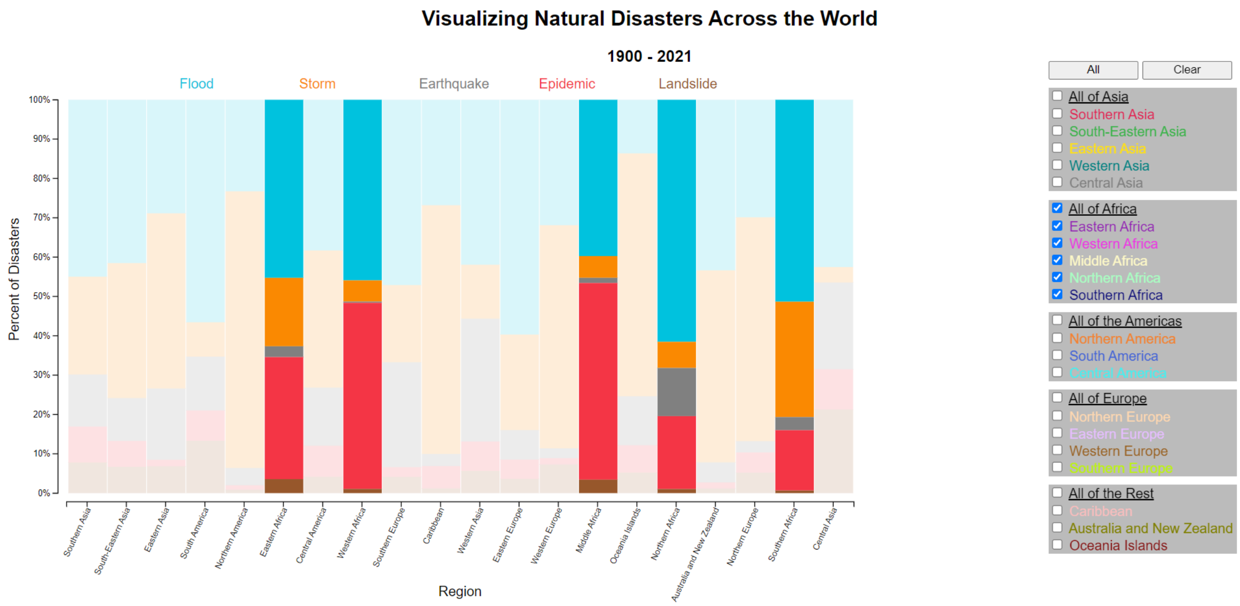


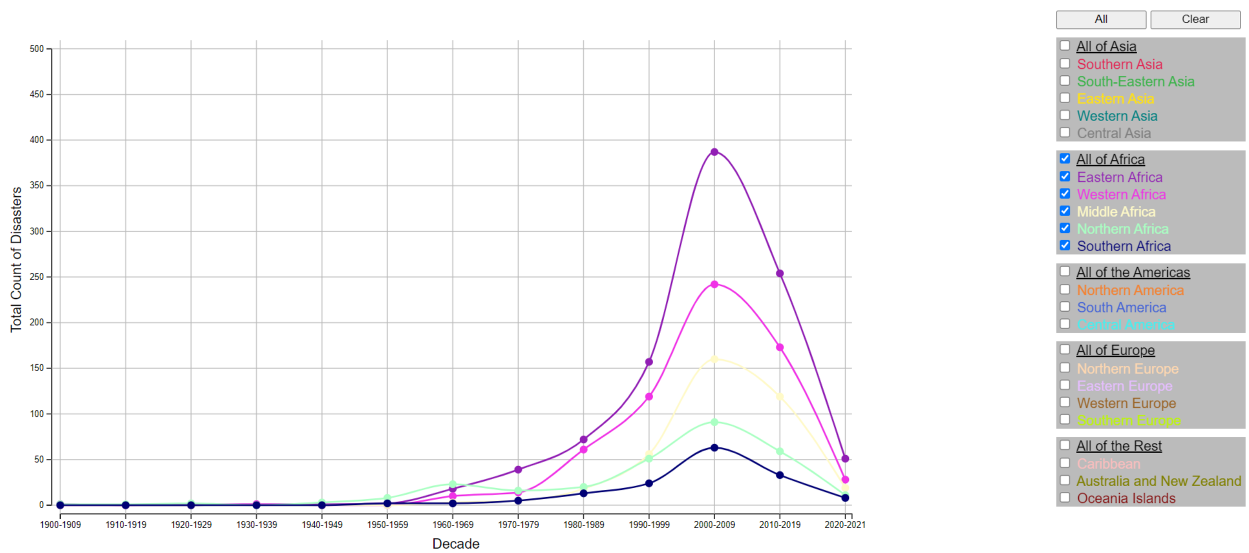
Our two main interactions provide filtering by disaster type and by region.

The distribution of disaster type is the aspect of this dataset that we wanted to focus on. The initial state of the bar graph shows all types, but users are able to hover over the provided labels to see specific types. This function makes it simple to compare the occurrence rates of each individual disaster type globally by “zooming in” on the categories one at a time.

Another important feature that we wanted to explore was regions in which natural disasters occurred. With 20 categories, a plain list of checkboxes would be inconvenient to click through, so we included options to select all, clear all, and choose subsets, which improves the interaction’s usability and speeds up the selection of desired regions.

Visualization





Our visualization consists of two graphs. The first displays the percentage of each type of disaster for every region (in total from the years 1900-2021), while the second shows the number of total natural disaster occurrences for each region per decade. In the screenshots above, “All of Africa” is selected in the region filter, which highlights the corresponding bars and lines - the two graphs are in this sense linked together. Additionally, the bar graph can be manipulated to show only one type of disaster by hovering over the labels.

Some of our design considerations/trade-offs include:

1. The y-axis of the bar graph is scaled by percentage instead of total count. Since we are

differentiating between types of disasters in this graph, it would be useful to be able to compare percentages; as an example, we can easily observe that western Asia has a higher percentage of earthquakes than southern Asia, even though the occurrence of natural disasters there overall is far lower than in southern Asia. Potential misleading representations are somewhat mitigated by the line graph below, which uses the actual number of disasters (aggregated) for each region. This way, users can benefit from information presented using both approaches.

2. The five most common disaster types were selected to include in the visualization because there are too many categories in the dataset, some of which are fairly trivial (there is only one incident in the disaster type "fog"), to include everything in one graph. We found five to be an acceptable cutoff to avoid overwhelming users, especially since there are already 20 regions presented in the graph. However, a disadvantage is that the exclusion of data might skew the percentages.

3. The line graph uses a linear scale for incident counts since the numbers are more or less evenly distributed and not too large.

4. The most recent decade (2020s) doesn't have enough data yet and thus is displayed as 2020-2021 instead of a decade, making it appear like the number of disasters sharply decreased, which is misleading if users don't look at it carefully. However, for sake of consistency, we decided to keep this bin in the line graph, as the aggregate data in the bar chart above also takes the disasters from 2020 and 2021 respectively into its total count. Moreover, the label on the bottom of the axis should signal to the user that this bin only contains two years. It is interesting to note that these two years already have more disasters than earlier decades like the 1930s and 1940s in some regions, signally the most likely continuation of increasing trends in the future.

Development Process

We initially considered mapping all occurrences to observe many distributions. However, the available data was not precise enough. We then realized that the dataset is more suited for discovering general trends, which led to the idea of comparing which types of disasters are more prevalent in what regions.

In the first sketch of our visualization, we wanted to filter by year as well but ultimately decided on a line graph instead. With the entire span of time displayed at once, the general upward trend for all regions is easily noticed.

Due to the limited width of the screen, we arranged the two graphs vertically. As the disaster type filter only applies to the bar graph, the labels are located close to it. Meanwhile, the region filter, which is linked to both graphs, is fixed at the right of the screen for ease of access.

Work Breakdown

- Isabelle: development & creation of the line graph (~7 hours), styling of the final visualization (~1 hour) final report editing (~30 min)

- Yvette: data preprocessing (~1 hour), final report (4 hours)
- Daniel: development and creation of bar graph (7 hours) , development of linked filters (1 hour), styling (2 hours)