Report on Natural Disasters

Jonah Harmon
Dana Strong
Presentation:
https://docs.google.com/presentation/d/1Iyrx4GMLWqHg0IceIVnqw62l6MSgUDKJqFUJ_gMCIt4/edit?usp=sharing
Git Repo: https://github.com/JHamoni676/cs5830_project3

Introduction

Natural disasters have been plaguing humans for as long as we've been around. While humans have been able to mitigate many of these crises, like infectious diseases, there are many natural disasters that we have not been able to prevent and/or predict. In this analysis, we focus on the three big natural disasters: earthquakes, tsunamis, and volcanic eruptions from the last century. We aim to determine which of these three disasters is the most destructive, both in terms of lives lost and damage cost in millions. We also want to see if these measures of destructiveness have changed over time, AKA, determine whether our mitigation efforts have helped lessen the blow of these disasters. We want to do this both worldwide, and by the countries most affected by these disasters, so researchers and governments alike can allocate resources to help mitigate disasters in the future.

One thing to keep in mind, however, is that our data is from 1924-2024. While data today is quite accurate, the data in 1924 was spotty in most areas, especially in developing countries at the time. This may affect our time analysis, especially when we try to determine if these disasters have increased or decreased over time. We also need to keep in mind that the human population has increased since 1924, and the number of people who live in cities, especially along the coast, has also increased over time. While we can account for inflation in the damage per millions of dollars, we will not quantify urbanization of human populations since 1924.

Dataset

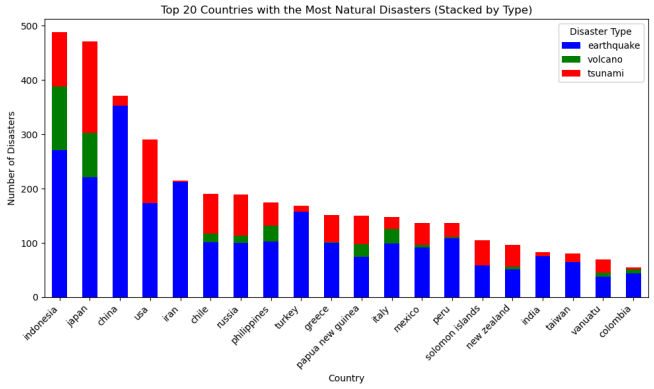
The dataset for this project comes from the National Centers for Environmental Information, part of the National Oceanic and Atmospheric Administration (NOAA). It includes data on all major earthquakes, tsunamis, and volcanic eruptions over the last 100 years, starting in 1924. The key details we focused on are the year the disaster happened, the country it affected, the number of deaths, and the damage caused (adjusted for inflation). These factors help us understand how severe, frequent, and costly these natural disasters are. The dataset gives us a clear look at the history of these events and helps us spot patterns, trends, and any unusual occurrences.

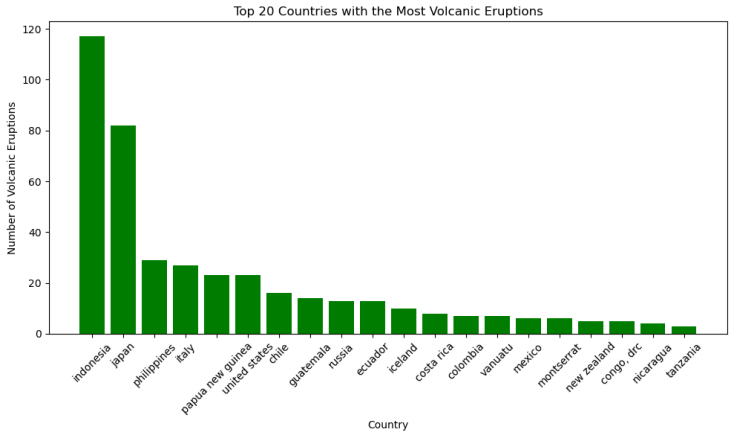
Analysis

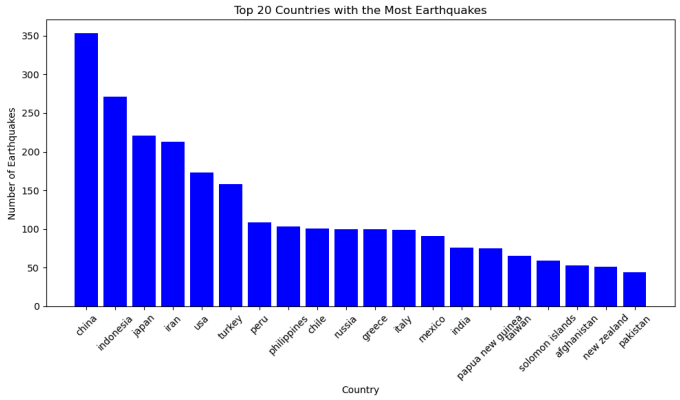
The first set of analyses we wanted to do was to see which countries are most affected by these natural disasters. We did this by filtering by country and disaster type and plotting the countries with the highest counts of each. We then wanted to determine which disaster was the most costly in terms of deaths and inflation-adjusted damage (in millions of dollars). We used an ANOVA test to determine if these differences were significant in both cases. Next we wanted to see if there was any correlation between the number of disasters and the total impact of

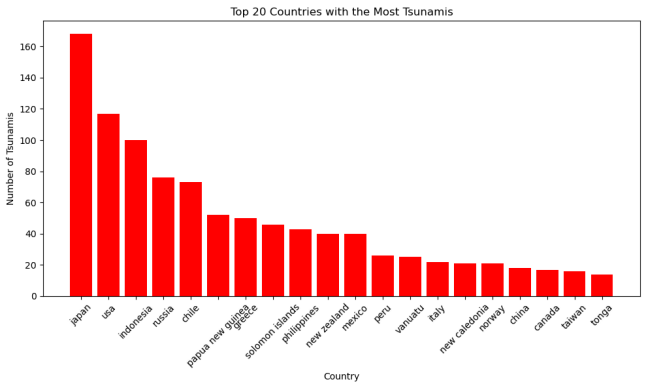
disasters, using the Pearson correlation coefficient, and if there was any country that was particularly burdened with any natural disasters. Finally, we want to see if the number of natural disasters, the number of deaths, and inflation-adjusted damage changes with time. We did this with three time-series plots, and calculated the Pearson correlation coefficient.

Results



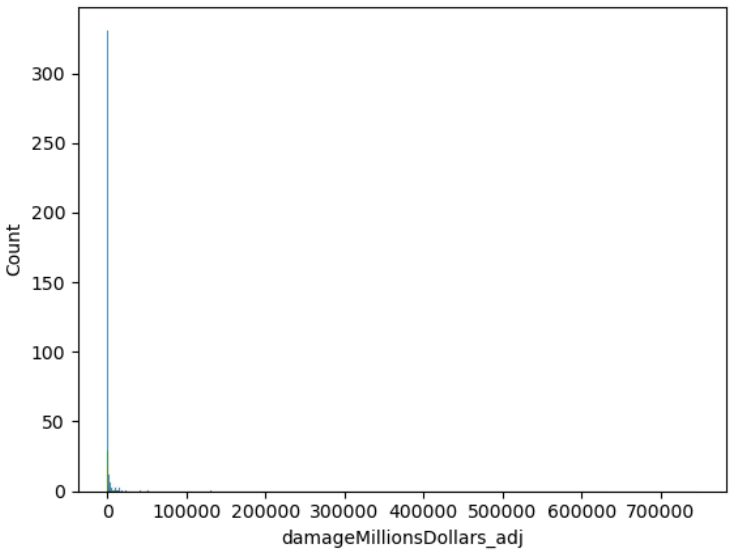




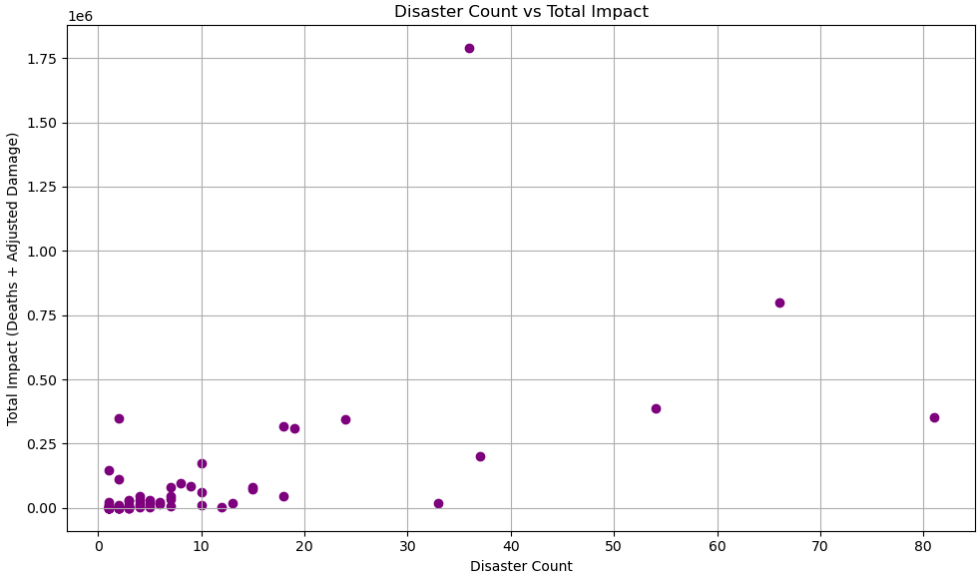


We can see the top 5 countries are Indonesia, Japan, China, the United States, and Iran. We also notice that Indonesia and Japan dominate the majority of the volcanic eruptions (green) while the other disasters are more balanced between countries.

We then tested whether there's a significant difference in how many deaths and for each disaster. We did this using the ANOVA three-way test of the number of deaths and the number of damages for each disaster. We did make the NAN values 0, because we figured that in this case, if there was no reporting of people dying/damages, there must have not been any significant number to. The ANOVA test for number of deaths gave us a non-significant

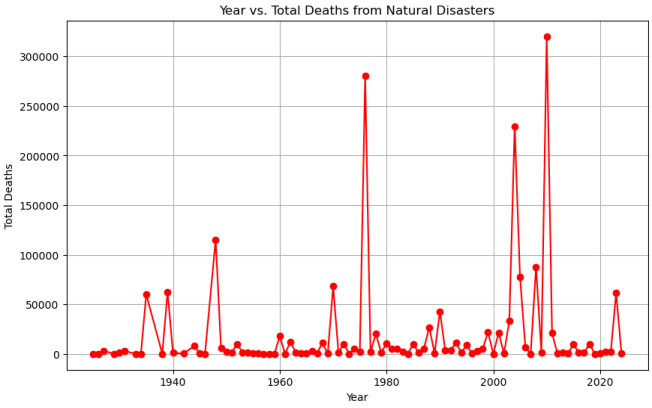


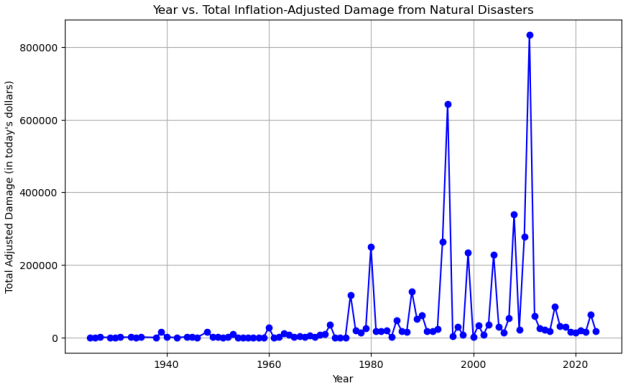
value for all data and when we filtered it for disasters with large amount of deaths, but was significant (4*10^-7) when we filtered for a small number of deaths (<1000), which we suspect is simply because the variation in deaths from volcanic eruptions is so large. The ANOVA to test if there's a significant difference between the damages between the natural disasters gave us a non-significant p-value for all data, and a marginally non-significant value (0.06) when we filtered for all values where the number of damages > 0. We do think, however, that that p-value 0.06 is truly non-significant because the data is so intensely right-skewed for all of the disasters, and the fact that there is so many more earthquakes (~2000 data points) than tsunamis (~200) and volcanic eruptions (~100) combined.

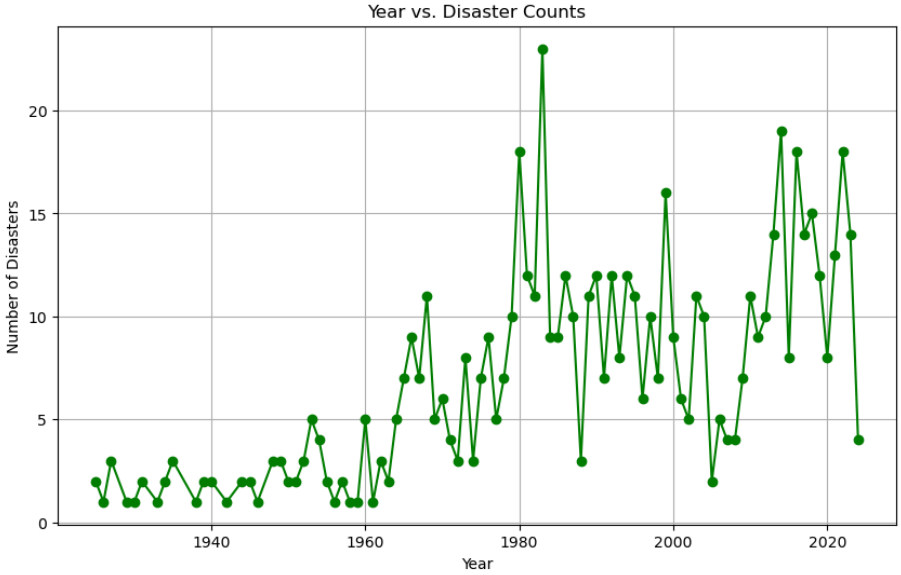


We then tried to determine whether there was any correlation between the countries with the most disasters and the total impact from these disasters (the addition of deaths and damages in millions), and whether there was any country that was particularly burdened. The Pearson correlation coefficient unsurprisingly gave us a significant p-value of 5*10^-9, meaning as the number of disasters a country has increases, so will the total impact of those disasters. More interesting, however, is the outlier: Japan. Japan has a much larger total impact than expected due to the enormous damage in millions of dollars it has endured.

The last set of analyses we did was to see if the number, deaths, and damages done by natural disasters has changed over time. The Pearson correlation coefficient gives us a value of 2*10^-14, 0.18, and 0.002, respectively. While these changes over time may be legitimate, it is important to remember that the data collection was much worse and much less accurate even 50 years ago than it is today.







Technical

The data preparation in this project was not near as bad as the last. NOAA keeps good data. The dataset offered by NOAA started in -2000 BCE with earthquakes, but the API only allowed to pull 200 rows per call. So we kept it to the last 100 years, also because reporting before isn't very good. We chose to fill null values of damages and deaths with 0, the thought being that if nothing was reported it's because there were no damages or no related deaths. One important part of our analysis was adjusting for inflation. We calculated the value of one dollar in 1924 today (~$18.36) then used that to calculate a linear approximation of inflation over time. It is the reason why we got such large numbers in recent damages.

We used basic statistical analysis methods including bar charts, histograms, a scatterplot, ANOVA test, and the Pearson test to find significance. We felt these were appropriate because they let us see the distribution of disasters over time. They allowed us to look at trends in the frequency and impact of disasters. Through the process we tried to use a KDE plot, but for some reason our numbers cause the distribution to center over 0 and go into the negatives, so we decided to leave it out. The histograms let us find interesting extreme outliers.