Achievement 6 Project Documentation

Analysis of global tsunami trends 1985-2024

Data Summary

Personal Motivation

For my final achievement I have chosen to work with a data set containing 40 years of data related to runups (i.e. the waves formed as part of a tsunami). This time frame includes 20 years prior-to, and following, the devastating 2004 Indian Ocean earthquake and tsunami, which is considered to be the deadliest natural disaster of the 21st century.

The earthquake and tsunami dominated the news in Australia and the time and I have strong memories of this time. However, I don't think I truly understood the magnitude of the devastation until recently watching a 20-year anniversary special detailing the full extent of the deaths and destruction.

My goal with this analysis is to gain a better understanding for myself of how unique this event was in the context of other tsunamis/earthquakes during the same timeframe. I'm also hoping that I can make some visualisations that make the huge numbers of fatalities, as well as the wave size, easier for people to comprehend.

Source

This data set comes from an external source.

It was generated using the National Centers For Environmental Information (NCEI) / World Data Service (WDS) Global Historical Tsunami Database, which contains runup[1] information on locations where tsunami effects were observed.

The NCEI is a line office of the National Oceanic and Atmospheric Administration (NOAA), which is an agency that belongs to the US Department of Commerce.

The NCEI/WDS is the owner of this data. As these bodies belong to a US government agency, I believe it to be a trustworthy source.

This dataset was generated in four sections because of processing limitations of the database. I will join these together in python.

Below are the details of when each dataset was first generated and accessed:

[1] The maximum height above sea level a tsunami reaches on shore is called the *runup*.

- Runups_1985-1994 was first accessed on 2025-02-08 at 13:33:30
- Runups_1995-2004 was first accessed on 2025-02-08 at 13:36:29
- Runups_2005-2014 was first accessed on 2025-02-08 at 13:38:39
- Runups_2015-2024 was first accessed on 2025-02-08 at 13:40:57
- All files were accessed from the tsunami database here: Please cite DOI:10.7289/V5PN93H7

The above constitute the main dataset for my analysis. I suspect that I will want to find some secondary datasets for comparison of wave heights to more recognizable objects (i.e. buildings, animals, human heights etc).

Collection method

This data is administrative data that NCEI/WDS has gathered from various sources including eyewitnesses, field reconnaissance surveys, tide gauges or deep-ocean sensors.

Contents:

The data contains the date, time and location of tsunami runups between 1 January 1985 and 31 December 2024. In addition to these core variables, it also contains incomplete information on the max water height, max inundation distance, the time it took for the runup to travel from its source to the shore as well as the source of the tsunami.

Possible Limitations:

As some of the max water height can be reported by eye-witnesses there is opportunity for human error. As this data has been collected by a US agency, it is possible that the data set is skewed to the regions where the NOAA has their own measuring devices, and therefore, is missing key data for other regions.

Additionally, measurement abilities may have improved overtime. So, we might see an increase in tsunami's over time, but this might simply be that we are better at measuring these events. Patterns to look out for to confirm this would be an increase in the number of low impact tsunami events that may have gone undetected by humans in the past.

Data Cleaning Steps

Please view the attached Jupyter notebook 6.1 for details of the data cleaning steps performed.

Data Profile and Dictionaries

I have created a data profile as well as a few dictionaries for various categorical variables. These can be viewed in the attached excel sheet.

Questions for analysis

After performing the data cleaning and consistency checks, some questions that began to jump out at me from the data frame include:

- Is there a relationship between location and earthquake magnitude?
- Is there a relationship between location and runup wave height?
- What is the average number of runups per tsunami event?
- What is the average runup wave height?
- Does the distance of the tsunami source (i.e. earthquake / volcano) impact the wave height? Or wave inundation distance?
- Was there anything unique about the 2004 Indian Ocean earthquake and tsunami?