Maya Barron

Tsunami Damage Predictions

With my final project, I would like to examine the factors that contribute towards damage caused by tsunamis around the world. The damage would include deaths, number of houses destroyed, and total cost of the event. Technology has come a long way in recent decades in regards to detecting incoming tsunamis and building infrastructure to mitigate the effects of the storm hitting, so it'd be interesting to know what effect this shift in technology has had for damage caused by tsunamis.

The dataset I will be using is an offshoot of The Global Historical Tsunami Database from NOAA. I would have used this full dataset but I could not figure out how to download the dataset. Therefore, I will be using a tsunami dataset from Kaggle which has all of the same main variables of interest as the NOAA dataset, but it has 2259 observations in contrast to NOAA's 2947. This dataset is a csv file with 21 columns, and each row represents one single recorded tsunami. This dataset includes information about tsunamis from 2100 BC to present day, so there is quite a bit of missing data. The description of the dataset also states that locations may not be entirely accurate for older tsunamis, as these events obviously don't have official records. Luckily, there is a column that indicates the trustworthiness of that particular entry in the dataset, which I plan to use to get rid of some of the more unreliable entries. Outside of this tsunami dataset, I would like to also make use of a dataset that maps country boundaries so I can map out where tsunamis have historically occurred the most by using the latitude and longitude variables from my chosen dataset.

To inspect the factors that contribute towards damage caused by tsunamis around the world, I will be using machine learning regressors. I will start with a simple linear regression by looking at how tsunami intensity impacts damage, as one would expect these two variables to be highly positively correlated. I'll then look to add on more regressors like latitude and longitude to look at other factors that may influence damage. I may look to try other regression models as well, but my first main focus will be a linear regression model. Since my data goes back so far, I may also look to do something relating to time series analysis/forecasting to see whether death tolls have been increasing or whether the number of tsunamis have been on the rise over the years. Finally, I'll visualize where tsunamis are hitting and with what intensity with a proportional dot map, as well as how the number of or intensity of tsunamis has changed over time with a time series graph.