Earthquake Analysis 1900-2020 – Jason Hunt

Exploratory Analysis

Data sources		
Type	Name	Records
MS Excel	earthquakes raw data	6215

Tables	
Table	earthquakes raw data
Field name	Description
Year	*To be completed*
Mo	
Dy	
Hr	
Mn	
Sec	
Tsu	
Vol	
Location Name	
Latitude	
Longitude	
Focal Depth (km)	
Mag	
MMI Int	
Deaths	
Death Description	
Missing	
Missing Description	
Injuries	
Injuries Description	
Damage ($Mil)	
Damage Description	
Houses Destroyed	
Houses Destroyed Description	
Houses Damaged	
Houses Damaged Description	
Total Deaths	
Total Death Description	
Total Missing	
Total Missing Description	
Total Injuries	
Total Injuries Description	
Total Damage ($Mil)	
Total Damage Description	
Total Houses Destroyed	
Total Houses Destroyed Description	
Total Houses Damaged	
Total Houses Damaged Description	

For this project only one data source with one table/sheet was provided for analysis and no data dictionary was provided. If further analysis was to be completed then obtaining this definition would be very helpful.

Source sites provided:	
Site name	Address
NOAA National Centers for Environmental Information (NCEI)	https://www.ngdc.noaa.gov/hazel/view/hazards/earthquake/search

Initial Analysis

- It appears to be a results table rather than an original source table.
- A unique earthquake (record) identifier is missing from this dataset.
- Each result in the table appears to hold descriptive data such as:
    Year, Month, Day, Hour, Minutes, Latitude, Longitude, Location name
  As well as measurements such as: Magnitude, MMI, Death description etc.
- Latitude, Longitude: provide us with an accurate method to map our data set.
    - 48 records are missing both fields.
- Time columns: Year, Mon, Dy, Hr, Mn, Sec   These provide useful information for the earthquakes, however for many earthquakes this level of detail is not available in the data and as such I have chosen to apply my analysis to Year only.
- There are possible references to other datasets in the fields Tsu and Vol however after a short exploration of a related source site, I could not determine any relevant data that would prove fruitful.
- The table appears to contain data that may have gone through a style of cleaning or manipulation process. Many of the fields that would be expected to contain values are either empty or hold very few values. This could be due to the method of collection data or that these fields have been considered obsolete.
- Several of the fields have an associated "description" field which holds a classification or grouping for that result type rather than the individual entries. This can give the impression that the data entries have been groups/categorised, however I then wonder why the original entries where not left for more detailed analysis.   e.g.   Deaths / Death Description
- The possible source site for this dataset supplies a definition of where the data originated and if this is the same dataset then it has possibly been constructed from more than one original data source. In fact, five sources are detailed on the site. This would explain the large gaps in the dataset as different sources collect their data in a different manner, storage, and formatting. (Note: I say possible as we did not take this data directly from the stated source site.)
- The dataset did display a higher level of completeness as you move through to the younger records, and certain fields previously mentioned did display a possible pattern of incompleteness. However as also previously mentioned, as the dataset could be the result of joining and unioning different data sources, which would explain some of the inconsistency.
- Another reason for the inconsistency is the failure of the global community to agree upon one method of recorded earthquake activity and events. The shift from magnitude to Modified Mercalli intensity could also explain why some variables are not fully populated
- I decided after getting a better understanding of the data that I would use the period 1900-2020. For two reasons:
    - This period meant that several of the fields were more populated with entries than the period before 1900
    - Researching the topic of earthquakes, I discovered that "*Events prior to 1900 were not instrumentally located, therefore, the location given is based on the latitude and longitude of the city where the maximum effects occurred.*" After 1900 data collection improved as did the methods of recording earthquake measurements and effects. This mean that several of the issues seen in the absence of data.
- Working on the period 1900 to 2020 I have found that only one of my Longitude entries was NULL for Costa Rica-Panama. I will filter this entry out from my GIS Analysis, yet still use it in my overall counts.
- Several of the categorisation descriptive fields hold Nulls but rather than exclude their other important information from the analysis I am choosing to set a value for them. Given the data that I am reporting for those fields especially Death Toll or Damage Level I have chosen not to be incentive to the data and simply set those values to 0, instead where I will set to 'Unreported'.

Calculations and definitions

During my investigation I found the following that helped add more depth the dataset.

Definition: Modified Mercalli Intensity
A seismic intensity scale used for measuring the intensity of shaking produced by an earthquake. It measures the effects of an earthquake at a given location, distinguished from the earthquake's inherent force or strength as measured by seismic magnitude scales (such as the "Mw" magnitude usually reported for an earthquake)

Definition: A significant earthquake
Classified as one that meets at least one of the following criteria: caused deaths, caused moderate damage (approximately $1 million or more), magnitude 7.5 or greater, Modified Mercalli Intensity (MMI) X or greater, or the earthquake generated a tsunami. **Used to calculate: Significant earthquake**

Definition: Damage Description
0 - None 1 - Limited (<$1 million) 2 - Moderate (~1$ to $5 million) 3 - Severe (~>$5 to $24 million) 4 - Extreme **Used to calculate: Damage class**

Definition: Total Death Description
0 - None 1 - Few (~1 to 50 people) 2 - Some (~51 to 100 people) 3 - Many (~101 to 1000 people) 4 - Very Many (~1001 or more people) **Used to calculate: Total death toll**

Definition: Latitude versus Longitude
Latitude: 0 to 90 (Northern Hemisphere) -90 to 0 (Southern Hemisphere) Longitude: 0 to 180 (Eastern Hemisphere) -180 to 0 (Western Hemisphere) **Used to calculate: Which hemisphere**

Definition: Earthquake Magnitude Scale / Earthquake Magnitude Classes		
Magnitude	*Earthquake Effects*	*Estimated Number Each Year*
2.5 or less	Usually not felt, but can be recorded by seismograph.	900,000
2.5 to 5.4	Often felt, but only causes minor damage.	30,000
5.5 to 6.0	Slight damage to buildings and other structures.	500
6.1 to 6.9	May cause a lot of damage in very populated areas.	100
7.0 to 7.9	Major earthquake. Serious damage.	20
8.0 or greater	Great earthquake. Can totally destroy communities near the epicenter.	One every 5 to 10 years

Earthquake Magnitude Classes
Earthquakes are also classified in categories ranging from minor to great, depending on their magnitude.

Class	*Magnitude*
Great	8 or more
Major	7 - 7.9
Strong	6 - 6.9
Moderate	5 - 5.9
Light	4 - 4.9

Minor	3 -3.9	
		Used to calculate: Earthquake class

Summary and Presentation

Outcomes

- The number of unrecorded entries in or records or NULL entries meant that there were several large sections of our records that we could not analyse effectively. To resolve this is would be important to confirm that these fields were not collected or that their value is zero (especially as several of our categories did not have any 0 entries.
- Overall, I was able to gain a better understanding of earthquakes, where they occur, methods of measuring them and their devastating effects. I have also come to learn that I need to change the way that I talk about earthquakes by using the MMI measurement.
- My initial planned analysis was traded in for an analysis into the differing measurements of magnitude and MMI and I would have liked more time and more accurate or complete data to see how MMI is linked to actual events.

Further analysis

- Looking into further comparisons on the MMI measurement and the events that occurred in unison with an earthquake to visualize/interpret if the MMI does hold as a more accurate measurement of earthquakes than the magnitude. To perform this task, it would be more important to have more complete data or to focus on an area such as the US where the one method of collection it used. Instead of the world where multiple differing methods may be in play. Perhaps a little closer to home, it would be interesting to analyse the data available to us from New Zealand.
- Looking into completed time collections and being able to compare earthquakes with the different time of the day or month to determine if that effects the death toll. e.g. Are people being alerted while they slumber.
- Looking further into the "Ring of Fire" and being able to find linked Volcano and Tsunami information so we could perhaps create a reconstructive image of the earthquake and resulting effects and movements.
- Correct separation of the location name to provide a more precise detail or ability to analyse the data by different locations in a bar chart or other visualisation than a GIS (i.e. using latitude and longitude)

Calculated fields

- Several fields were introduced to allow me an extra level of overlay or shading of my data.
- A combination field that used previously created fields was developed to answer the question of Significant earthquake.
- A future analysis could involve creating some extra calculated fields to provide percentage results or average values.

Custom Split

- Using the custom split an extra field was created from the location name. Attempts were also made to perform a further split on one of the resulting fields. However, the structure of the location columns does not lead to being able to split using this method as all rows do not follow the same format. If I were to guess, based on my earlier statements, it could be that several countries data results have been put together, so different data entry format issues may exist.
- Some of the entries in the location name field look like the headers of a country set. They are also missing any entry for subcategory or lower levels. Whilst other fields that appear to be from the same country hold multiple levels. So creating an automatic system to split this column is difficult.
- Two suggested approaches: Perform a search for a country and separated each country into its own datafile. Then strip/separate the data in the location name column to separate details. Followed by reunioning the countries together in one larger set.
- The second suggestion is to use the latitude or longitude to reference another dataset for definitions of the location, or to check the location. Similar to those displayed in region on the linked source site:
  - No Selection
  - 20 - Antarctica
  - 70 - Atlantic Ocean
  - 80 - Bering Sea

○ 90 - Caribbean
○ 100 - Central America
■ Where I guess but did not explore that these numbers or connected to the latitude or longitude values.
Ask a classmate to review your visualisations
Completed. A useful process as gaining a different perspective can help you to develop and also highlight issues that the author might have failed to notice by themselves.