Project 1 – Final Report

Data Analysis of Four Decades of Global Earthquakes — Clean, Model, and Communicate

Creator: Reza Farzad

Mentor: Prof. Roya Ghiaseddin

ACMS 86700-02 Directed Readings

University of Notre Dame
FALL Semester 2025

Executive Summary

The current project analyzes ~20,000 earthquake records in the USGS ComCat catalog over the past forty years (1985.09.02 to 2025.09.02), restricted to events of magnitude larger than 5.5. The workflow builds a compact and reducible pipeline to inspect, clean, and transform the catalog data to explore three questions:

1. **Depth-magnitude relationship**: Distribution shape and if magnitude and depth correlate.
2. **Temporal behavior**: How annual frequencies and summary statistics evolve over time.
3. **Geospatial patterns**: Where large and deep events tend to occur across the globe.

Key findings:

- **Most events are shallow**: ~80% occur at depths < 70km, ~14% are 70-300km, and only 6% are > 300 km. The 90th percentile of depth is ~163 km.
- **Magnitudes are right-skewed**: values near 5.9 dominate; very few events exceed 8.
- There is **no strong depth-magnitude correlation**. The largest magnitudes (> 7.5) are mostly shallow.
- **Yearly counts has no persistent upward or downward trend.** Yearly counts of events (with magnitudes ≥ 5.5) are typically between ~450 and ~550 events. The strongest and most interpretable correlation between counts and statistical features of magnitude and depth is for **count vs max of magnitude = 0.42** (weak-to-moderate).
- Based on spatial analyses, **the strongest earthquakes occurred in South Asia, East Asia, South Australia, West of South America, and West of North America.**

Data & Scope

Source: USGS Earthquake Catalog (ComCat query), and the initial working table had 19,611 rows and 22 columns. They are shown in Figure 1.

Focus: A compact subset of nine columns was chosen for analysis: time, latitude, longitude, depth, mag, magType, depthError, magError, id.

Goal: Educational EDA rather than real-time hazard assessment; uncertainties are acknowledged and discussed.

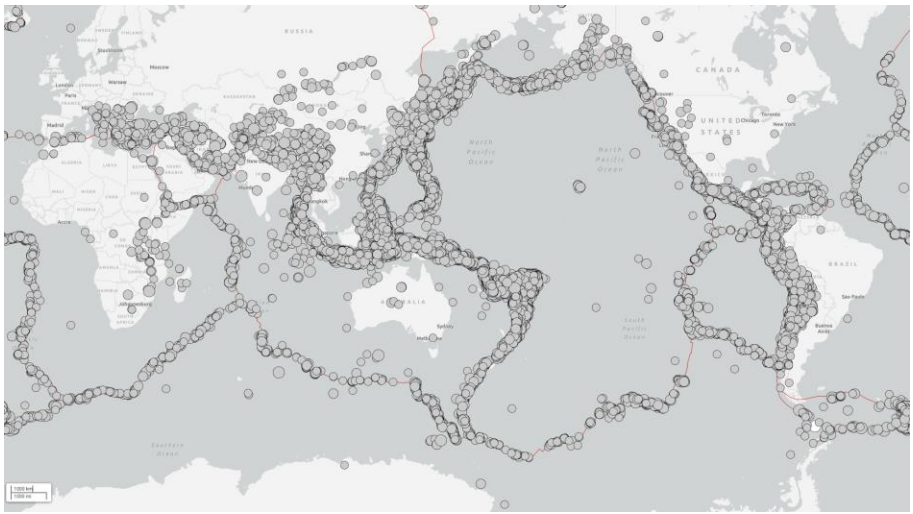

Figure 1. Location of earthquake dataset on the world map

Preprocessing

1) Understanding the table

- Basic inspection shows that there are 12 numeric and 10 object columns, and 8 out of 12 numeric features include missing values.
- A heatmap of missing values provides a quick look at where they are located.

2) Cleaning

Missing data (NaN):

- The 'depthError' is missing in ~60% of records; 'magError' is missing in ~79%.
- The correlation of missingness between these two variables is ~0.64, indicating temporal patterns (older events lack these uncertainties more often)

Sentinel/ out-of-range checks:

- Across inspected fields, values lie within plausible ranges for latitude (-90, 90), longitude (-180, 180), depth (0, 752km), and magnitude (5.5, 9.5).
- Five negative depths were found (small corrections); these were set to 0.0 km to denote very shallow events for robustness.

Outliers:

- Pair plots suggested outliers in depthError and magError.
- Using a z-score threshold $|z|>3$, the process indicates 38 depthError and 31 magError outliers, and 66 unique rows in total (assuming duplicates as one)
- Outliers are < 0.1% of the dataset and removed before visualization.

Errors & consistency:

- No duplicate rows were detected.
- All time values are clean (parse cleanly as datetimes with no invalid timestamp).
- magType shows 12 unique labels; counts confirm that the moment-based types are dominant. Therefore, for simplicity, we treat all magnitudes in the similar way.

3) Transformation & Feature Engineering

- Using categorical encoding, magType column was replaced by magTypeNum (via a frequency ranking 0 to 11).
- Temporal features, including year, month, and day were extracted from time.
- Final EDA table called 'df_all_numeric' has 19545 rows and 10 columns after cleaning and type conversion.

Results

A) Depth-Magnitude Structure

- **Depth**: KDE and quantiles highlight more concentration at shallow depths; P90=163km and P93.8=302km. One instance is the most recent earthquake occurred off the coast of Russia's Kamchatka Peninsula on July 29, 2025 with magnitude of 8.8 and 35 kilometers depth, which is considered to be shallow.
- **Magnitude**: Strong right skew; the majority are near 5.6-6.0, with a long tail at 9.1.
- **Joint view**: Scatter + KDE contours show no clear monotonic trend, and most events with magnitude greater than 7.5 occur at shallow depths

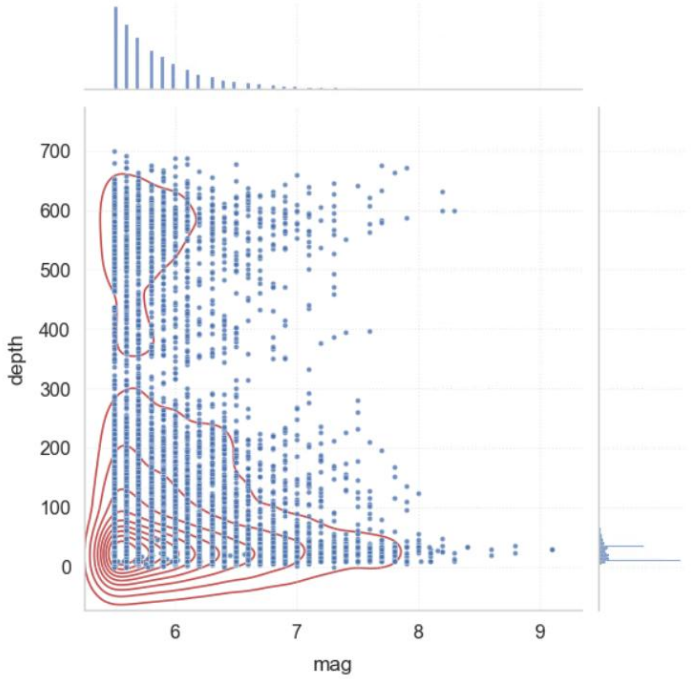

Figure 2. Scatter and kde plots of depth and magnitude relationships

B) Temporal Patterns (by Year)

- **Frequency**: After removing the partial initial year (1985), annual counts show no consistent trend; almost all years have ~450-550 events with magnitude higher than 5.5. The year with the lowest record has < 350 counts, and the year with the highest record shows > 700 event counts.
- **Year-level correlations** (from grouped summaries): The strongest association is count vs max magnitude, which is about 0.42, and other correlations are weaker, showing no relationship between statistics and event counts over years.

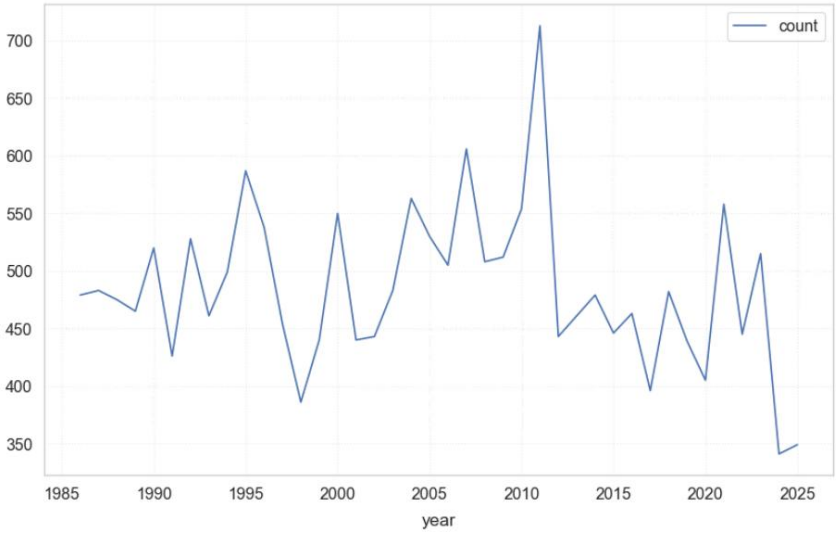

Figure 3. Trend of earthquake frequencies over last forty years

C) Interactive Global Maps

- **Depth & uncertainty**: Deeper events are clustered in western South America, East Asia, and East Australia with larger depth uncertainties appear in Pacific where fewer stations exist.
- **Magnitude**: The largest magnitude in the 40-year window focusses on South/East Asia, South Australia, western South America, and western North America.
- For better readability, a new mag_scaled feature linearly mapping magnitude is used for marker size.

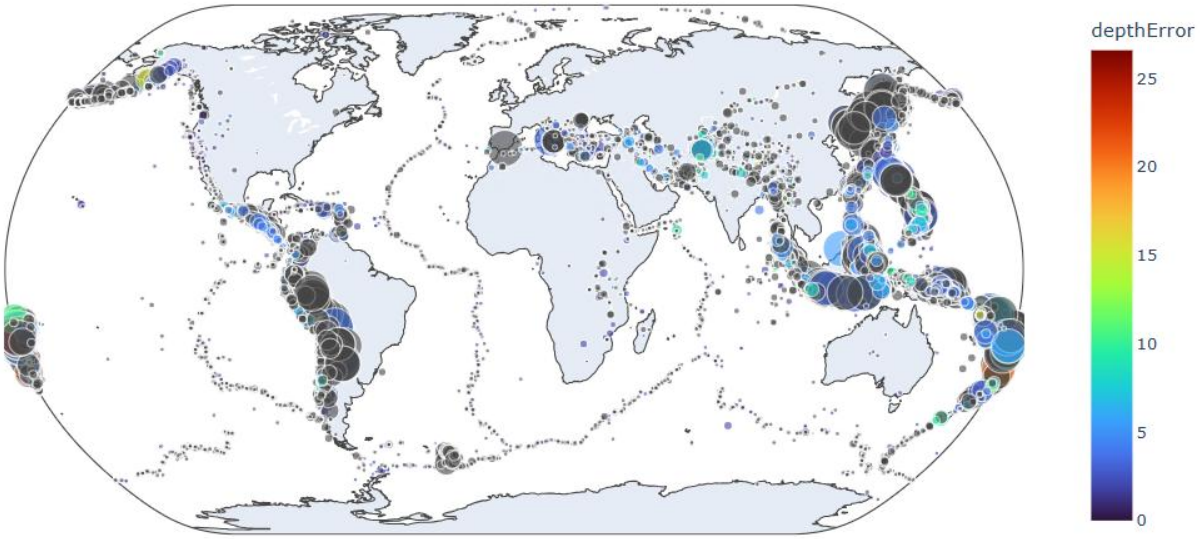

Figure 4. Spatial analysis of depth and its uncertainty

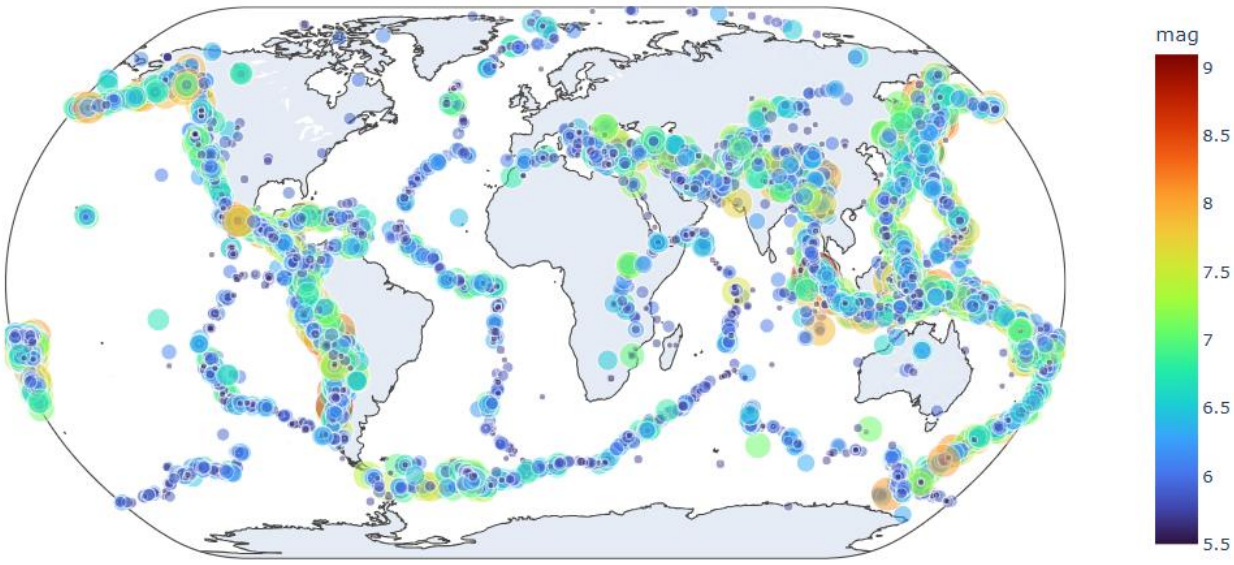

Figure 5. Spatial analysis of magnitude and its uncertainty

Discussion

Magnitude representation: The catalog mixes several magnitude types; nevertheless, ~92% start with mw, indicating that the method is based on moment. Keeping the type information in mind as well as encoding the types to numerics enables both interpretability and flexibility if necessary.

Missingness: The high and temporally structured missingness in depthError and magError (~60% and ~79%, with missingness correlation ~0.64) is important; uncertainty estimates are more frequent in older parts of the data. This helps with treating the uncertainty cautiously.

Outlier handling: A simple z-score rule ($|z|>3$) was chosen to determine extreme uncertainty values, leading to removing 66 rows (<0.1%) of the dataset.

Limitations

Magnitude threshold (≥ 5.5) intentionally biases the sample toward moderate-to-large events, and findings cannot generalize to smaller earthquakes.

Uncertainties coverage is uneven across time (especially for depthError/magError).

The work emphasizes **transparent EDA**, and there is no predictive models are trained.

Conclusions

This project demonstrates a reproducible EDA of global earthquakes focused on data cleaning and communicative visuals. After carefully cleaning, (addressing missing values, correcting rare negative depths, and removing the small set of outliers), the analysis shows:

- A shallow-event dominance with a thin deep tail to 700 km.
- Right-skewed magnitudes focused near the threshold (5.5) with few great earthquakes.
- Stable annual rates over four decades and a weak-to-moderate correlation between yearly activity and that year's maximum magnitude.
- Distinct global patterns for both depth and large magnitudes, available through interactive mapping.