Data Science Project Report: Global EarthQuake Dataset

1. Introduction

The dataset contains information on 1,137 earthquakes, with 43 attributes, collected via the EveryEarthquake API from RapidAPI. The aim of this analysis is to understand the earthquake patterns globally through data cleaning, exploratory data analysis (EDA), correlation analysis, and predictive modeling. The ultimate goal is to predict the magnitude of an earthquake based on various features.

2. Data Cleaning

Initial inspection of the dataset revealed that several columns had missing values:

Alert: 373 missing values

Continent: 270 missing values

Country: 338 missing values

Subnational: 421 missing values

City: 463 missing values

Postcode: 940 missing values

To ensure a clean dataset, all rows with missing values were removed. The cleaned dataset had no null values, resulting in a reduced dataset with 68 entries.

3. Exploratory Data Analysis (EDA)

a. Descriptive Statistics

The dataset's key numerical attributes were summarized:

Magnitude: Ranges from 3.5 to 5.22, with an average magnitude of 4.47.

Felt Reports: High variance in public reports of earthquakes, with a mean of 5,996.

Tsunami Events: Only 14.7% of earthquakes were associated with tsunamis.

b. Visualization Insights

Magnitude Distribution:

The majority of recorded earthquake magnitudes ranged between 4.0 and 5.0, with a peak around 4.4.

Magnitude by Continent:

Boxplot analysis revealed that magnitudes vary slightly by continent, indicating regional differences in earthquake intensity.

Depth vs. Magnitude:

A scatter plot of earthquake depth against magnitude suggested no strong linear relationship, indicating that depth alone might not be a significant predictor of magnitude.

Earthquake Trend Over Time:

The monthly trend analysis showed fluctuations in the frequency of earthquakes over time, with no distinct increasing or decreasing pattern.

Tsunami Risk Distribution:

85.3% of earthquakes posed no tsunami risk, emphasizing that most seismic events are not directly linked to tsunami generation.

Magnitude Type Distribution:

A count plot of magnitude types showed diverse classifications, with some types more prevalent in the dataset.

4. Correlation Analysis

A heatmap of the correlation between numerical features indicated:

Magnitude had moderate positive correlations with features such as mmi (Modified Mercalli Intensity) and cdi (Community Determined Intensity).

There were weaker correlations between magnitude and depth or latitude/longitude, suggesting these geographical factors might play a lesser role in magnitude prediction.

5. Predictive Modeling

Linear Regression Model

A linear regression model was built to predict earthquake magnitudes using features like depth, latitude, longitude, mmi, cdi, and other numerical attributes. Key performance metrics were:

Mean Squared Error (MSE): 0.052

R² Score: 0.743

These metrics suggest a reasonably good fit, indicating that the model explains approximately 74% of the variability in earthquake magnitudes using the given features.

6. Conclusion

The analysis highlighted several key trends and factors influencing earthquake magnitudes globally. The linear regression model's performance suggests that certain seismological measurements are indeed predictive of magnitude, although additional features could potentially improve accuracy. Further studies could explore more sophisticated models to capture non-linear relationships or other influencing variables.