EARTHQUAKE PREDICTION

A project report submitted to ICT Academy of Kerala

in partial fulfillment of the requirements

for the certification of

CERTIFIED SPECIALIST

IN

DATA SCIENCE & ANALYTICS

submitted by

Team members

Govind Raj
Gopika M
Shabna KP
Joyna Joy Perinchery
Teena Jose

ICT ACADEMY OF KERALA

THIRUVANANTHAPURAM, KERALA, INDIA

Oct 2024

List of Figures

List of Tables

Table of Contents

Abstract

The earthquake dataset provides records of major seismic events, including details such as magnitude, depth, location, date, time, and affected regions. This dataset helps in understanding the frequency and distribution of earthquakes across different geographic areas. This project aims to develop a machine learning model to predict earthquake magnitude based on key seismic factors such as latitude, longitude, depth, the number of seismic stations (nst), gap angle (gap), and root mean square error (rms), among others.

To improve model accuracy, data preprocessing techniques, including handling missing values and feature selection, are applied. Various machine learning algorithms, such as Decision Trees, Random Forest, Support Vector Machines (SVM), and Linear Regression, are implemented on a dataset containing 105,981 rows and 23 features to estimate earthquake magnitude based on seismic attributes.

Studying this dataset helps identify high-risk areas, contributing to better disaster preparedness, early warning systems, and improved earthquake risk assessment.

Problem Definition

Objective and Problem Statement

Earthquakes are one of the most devastating natural disasters, causing significant loss of life and infrastructure damage. Accurately predicting the magnitude of an earthquake can help mitigate risks by enabling early warning systems and disaster preparedness measures.

This project aims to build a machine learning model capable of predicting earthquake magnitude based on seismic factors such as location, depth, and geological characteristics. The dataset consists of 105,981 records with 23 features, providing a rich source of information for training predictive models.

Key challenges in earthquake magnitude prediction include:

- Develop predictive models that estimate earthquake magnitude.
- The presence of noise and missing data in earthquake records.
- The need for an accurate and generalized model that works across different regions.

By leveraging various machine learning techniques such as Decision Trees, Random Forest, SVM, and Linear Regression, this study attempts to analyze seismic patterns and improve earthquake magnitude prediction accuracy. The ultimate goal is to to analyze patterns, identify risk factors, and improve early warning systems.

Introduction

Earthquakes are sudden and often destructive natural disasters caused by the movement of tectonic plates beneath the Earth's surface. These movements release energy in the form of seismic waves, leading to ground shaking that can range from mild tremors to devastating quakes. Earthquakes can cause severe damage to infrastructure, loss of lives, and economic disruptions, making early detection and prediction essential for disaster preparedness.

Traditional earthquake prediction methods, such as historical data analysis and seismic monitoring, have limitations due to the unpredictable nature of earthquakes. While these methods provide useful insights, they often fail to detect subtle patterns that could indicate future seismic activity. This is where machine learning plays a crucial role. By leveraging large datasets and advanced algorithms, machine learning can recognize patterns and relationships within seismic data, improving the accuracy of earthquake predictions.

In this study, we apply predictive modelling using algorithms such as Decision Trees, Random Forest, Support Vector Machines (SVM), and Linear Regression to analyze a dataset containing 105,981 rows and 23 features. These models aim to estimate earthquake magnitude based on seismic attributes such as latitude, longitude, depth, number of seismic stations (nst), gap angle (gap), and root mean square error (rms). The goal is to enhance earthquake risk assessment, contribute to early warning systems, and improve disaster preparedness to minimize damage and save lives

Methodology

This section outlines the step-by-step approach used in this project, including data acquisition, analysis, feature engineering, model development, and evaluation. Additionally, different statistical analyses such as univariate, bivariate, and multivariate analysis are conducted to extract meaningful insights from the dataset

1. Data Acquisition

This dataset provides comprehensive information on significant earthquakes that have occurred around the world since 1900 with a magnitude of 5 or above. The data includes essential details such as location, date and time, magnitude, depth, and other relevant information about each earthquake.

Data Sources

The dataset used in this project was sourced from Kaggle, which is updated weekly and sourced from the United States Geological Survey (USGS), which maintains a global catalog of earthquake information. The dataset includes earthquakes from all regions of the world. This global coverage makes the data ideal for studying both high-risk and less active earthquake zones.

Earthquakes are among the most impactful natural disasters, capable of causing vast damage and loss of life. Hence, this dataset is valuable for research in:

- Earthquake pattern analysis
- Predictive modelling
- Infrastructure risk assessment
- Disaster Preparation

Each row in the dataset corresponds to a single seismic event and includes vital details such as:

- Time, location (latitude, longitude), and depth
- Magnitude and its type
- Seismic station data (e.g., nst, gap, rms)
- Descriptive location and event status

Column	Description
time	Timestamp of the earthquake
latitude , longitude	Geographic coordinates of the epicenter
depth	Depth of the earthquake in kilometers
mag	Magnitude of the event
magType	The scale used for measuring magnitude (e.g., mb, mw)
nst	Number of seismic stations used
gap	Largest gap between stations
dmin	Distance to the nearest station
rms	Root-mean-square of the location residuals
net, id	Identifiers for the seismic network and the event
updated	Time the event record was last updated
Place	Readable description of location
type	Type of seismic event (e.g., earthquake, explosion)
horizontalError , depthError , magError	Measurement errors
magNst	Stations used for magnitude calculation
status	Status in the catalog (e.g., reviewed)
locationSource, magSource	Networks responsible for location and magnitude reporting

Table 1.1 Columns of Dataset

Data Cleaning and Preprocessing

To ensure data quality and accuracy, the following preprocessing steps were applied:

- **Handling Missing Values**: When data is missing, we can fill it using different methods to keep the dataset useful:
    - Mean Imputation: Replace missing values with the average.
    - Median Imputation: Use the middle value instead (good for uneven data).
    - Mode Imputation: For categories, use the most common value.

  Missing entries in critical columns (e.g., magnitude, depth) were imputed using median values or others removed if necessary.
- **Removing Duplicates**: Any duplicate earthquake records were identified and dropped.
- **Outlier Detection**: Extreme values in magnitude and depth were analysed using boxplots and removed if they significantly deviated from realistic seismic activity.
- **Label Encoding**: Some features like magType, country, and net contain text values. We converted these text values into numbers using Label Encoding, which replaces each unique text with a unique number.
- **Standard Scaling**: Features like depth, gap, and rms have different ranges. To bring all features to the same scale, we used StandardScaler, which standardizes the data to have a mean of 0 and standard deviation of 1. This helps some models like Linear Regression and SVM work better.
- **Feature Selection**: the process of identifying the most relevant variables from a dataset that contribute to the predictive power of a model. In earthquake prediction, selecting the right features can significantly

enhance model performance.

- **Improves Accuracy**: Reduces overfitting by eliminating irrelevant features.
- **Enhances Interpretability**: Simplifies models, making them easier to understand and communicate.
- **Reduces Computational Cost**: Fewer features lead to lower processing time and resource usage.

2. Exploratory Data Analysis (EDA)

EDA is understanding the data sets by summarizing their main characteristics often plotting them visually. This step is very important especially when we arrive at modeling the data in order to apply Machine learning. Plotting in EDA consists of Histograms, Box plot, Scatter plot and many more. It often takes much time to explore the data. Through the process of EDA.

primary goals of EDA are to:

1. **Understand the data structure:** Gain insights into the shape, size, and composition of the dataset, including the types of variables, missing values, and data quality issues.
2. **Detect patterns and relationships:** Identify potentially interesting relationships, trends, or patterns within the data that may warrant further investigation.
3. **Identify outliers and anomalies:** Detect and analyse unusual or extreme values that could influence subsequent analyses or modelling efforts.
4. **Test underlying assumptions:** Assess whether the data meets the assumptions required for specific analytical techniques or models.
5. **Inform data transformations**: Determine if any data transformations, such as scaling, normalization, or encoding, are necessary to prepare the data for further analysis.

Types of Visualizations in EDA

1. Univariate Analysis

Univariate analysis looks at one variable at a time to understand its basic properties. It helps us see how data is spread, where most values lie (central tendency), and how much the data varies (dispersion). This is useful for spotting patterns or unusual values in each feature on its own.

- **Histogram**: Visualizes the frequency distribution of a single numerical variable.
- **Boxplot**: Summarizes the spread, central tendency, and outliers of a variable.
- **Violin Plot**: Combines a boxplot and density plot to show distribution and probability.
- **Bar Chart**: Displays the count or value of different categories.
- **Pie Chart**: Represents the proportion of each category as slices of a whole.
- **KDE Plot**: Displays the probability density function of two variables for smoother distribution comparison.

2. Bivariate Analysis

Bivariate analysis explores the relationship between two variables. This helps us understand how one variable might affect or relate to another. For example, we can check how earthquake magnitude varies with depth.

- **Scatter Plot**: Shows the relationship between two numerical variables.
- **Line Plot**: Depicts trends or changes over time between two variables.
- **Bar Plot**: Compares values across categories using bars for each group.

3. Multivariate Analysis

Multivariate analysis looks at three or more variables at once. It helps us see how multiple features work together and how they influence the

target variable (like earthquake magnitude). This is important for building better predictive models.

- **Pair Plot**: Displays scatterplots for all pairs of numerical features in one figure.
- **Correlation Heatmap**: Uses color to represent the correlation between multiple numerical variables.
- **3D Scatter Plot**: Represents the relationship between three numerical variables in 3D space.
- **Animated Timeline Plot**: Shows how data points or trends change over time through animation.

Data Visualization

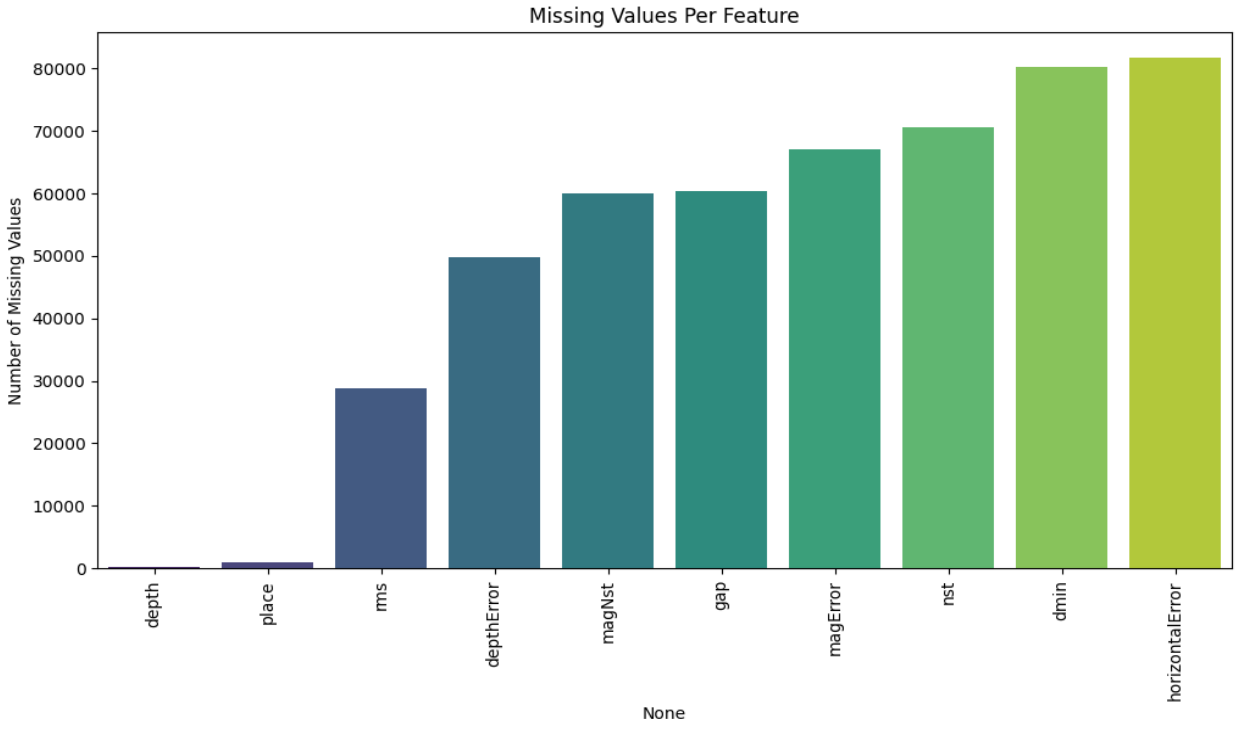

Fig 2.1.1: Bar Graph – Missing Values per Column

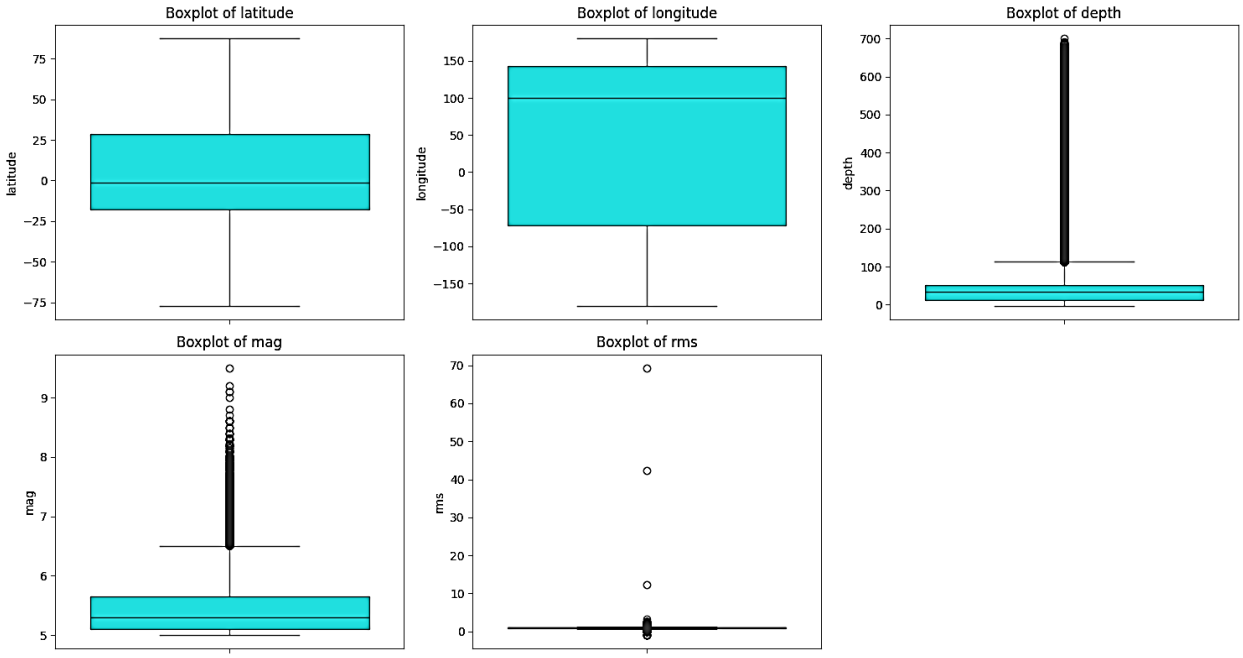

Fig 2.1.2: Box Plot – Identification of Outliers in Numerical Features

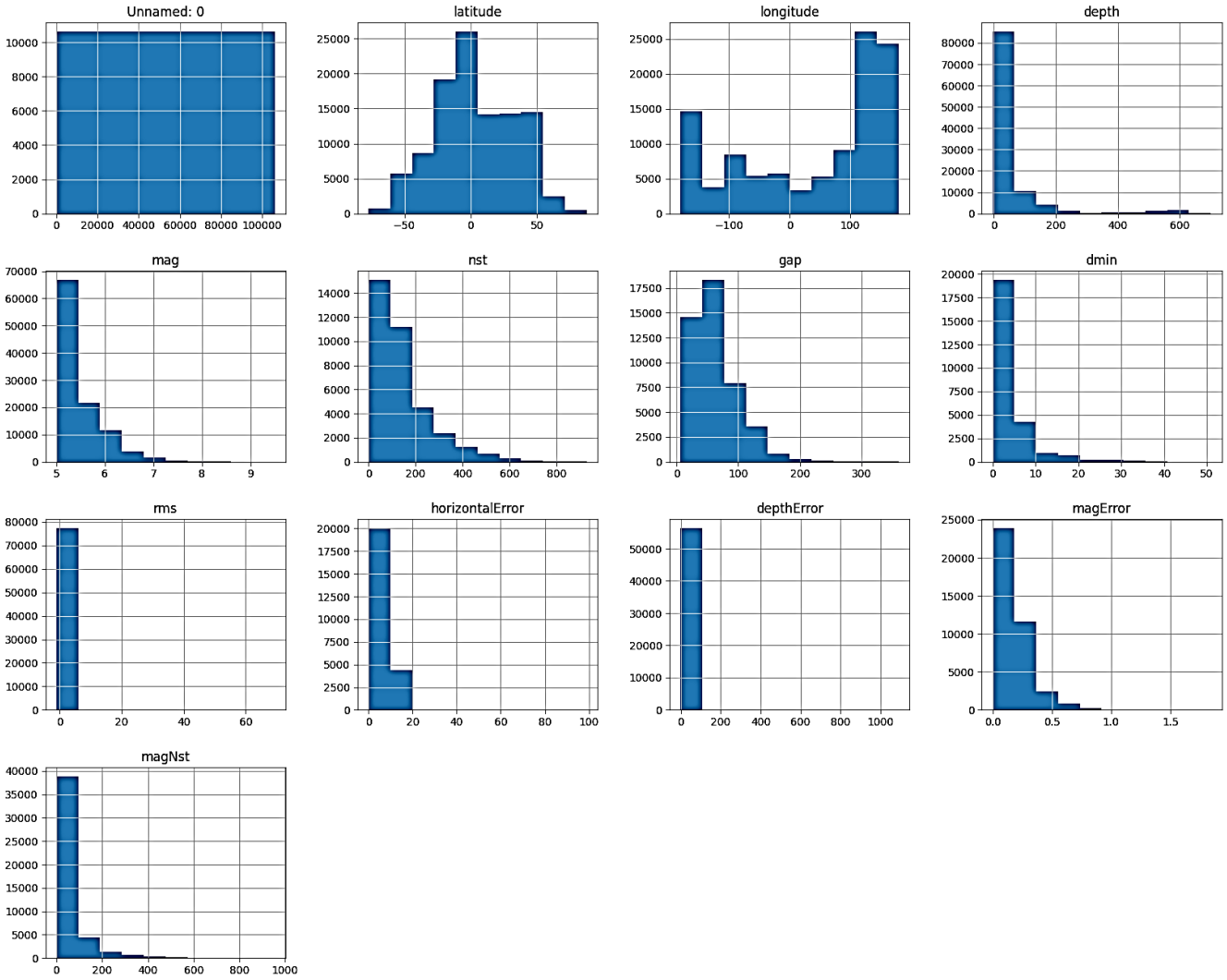

Fig 2.1.3: Histograms – Distribution of Numerical Columns

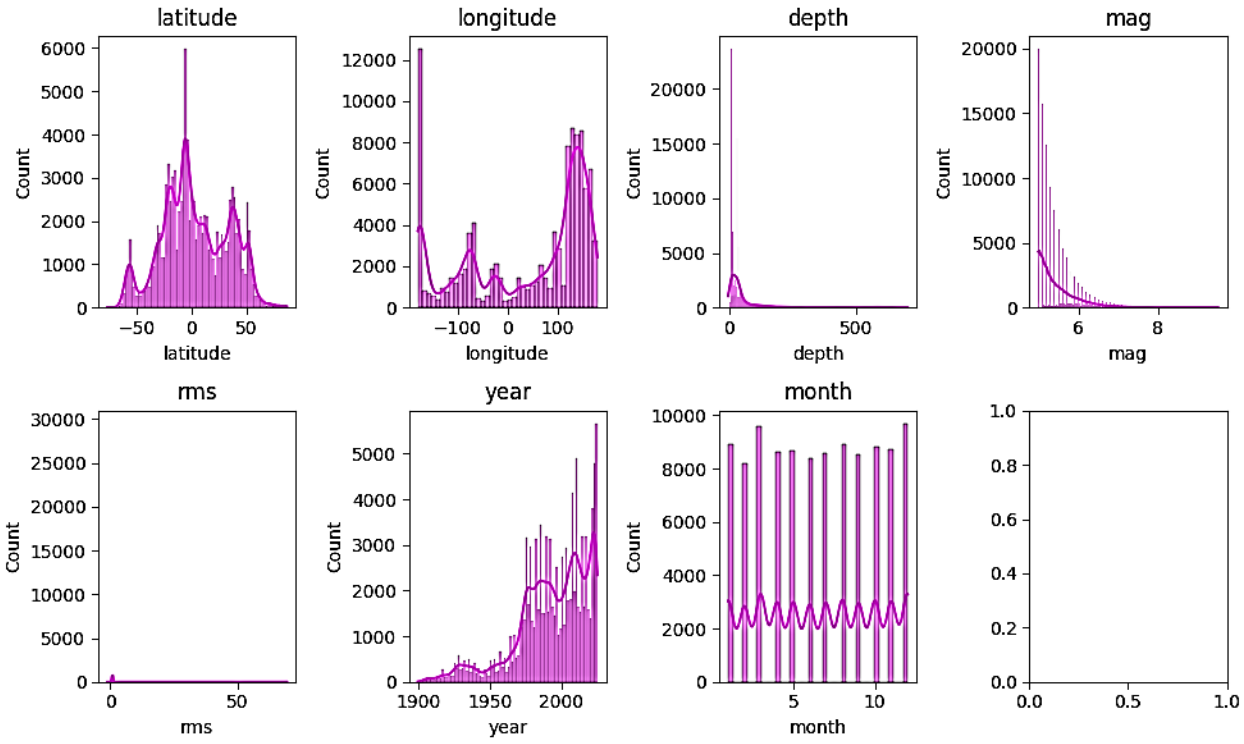

Fig 2.1.4: KDE - Distribution Plots of Key Features

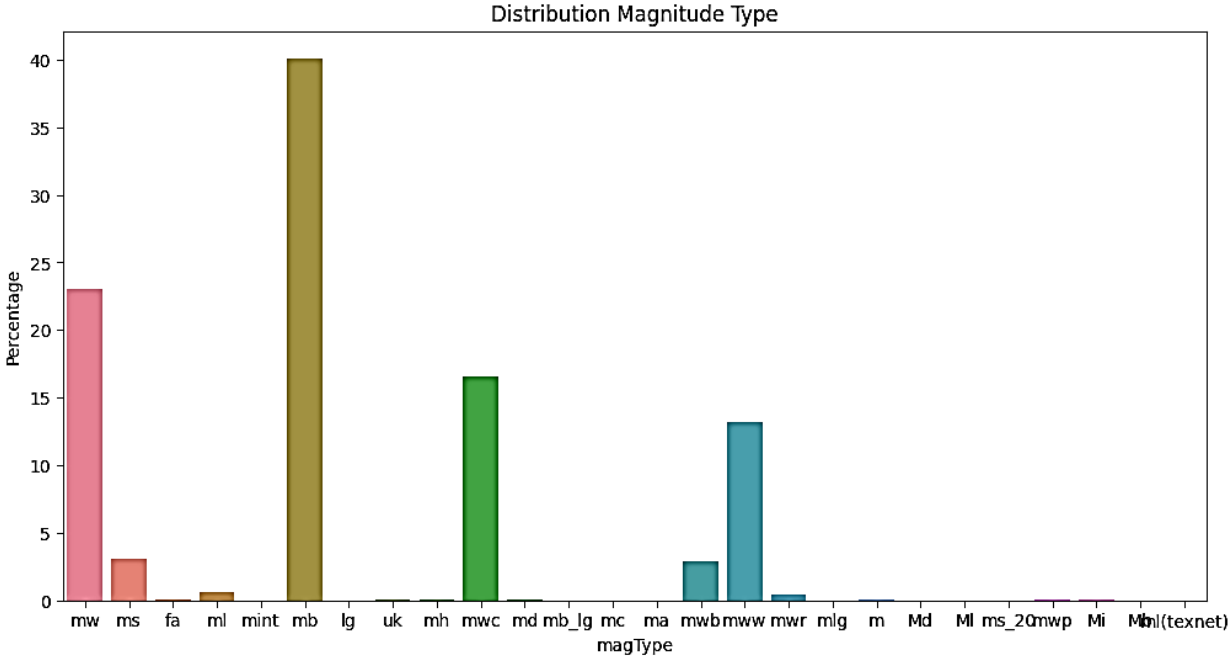

Fig 2.1.5: Bar Plot – Distribution of Magnitude Types

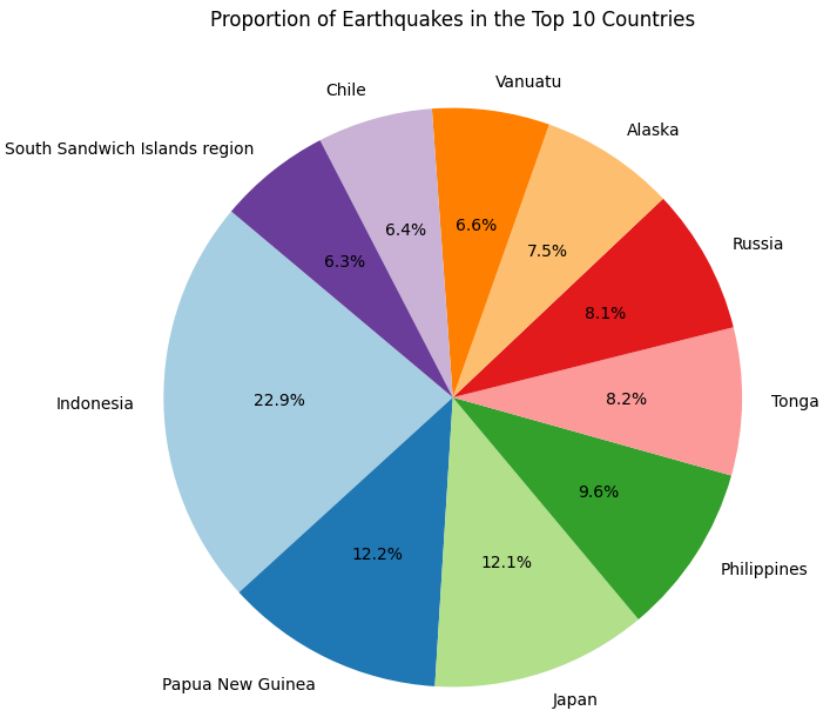

Fig 2.1.6: Pie Chart – Earthquakes by Top 10 Countries

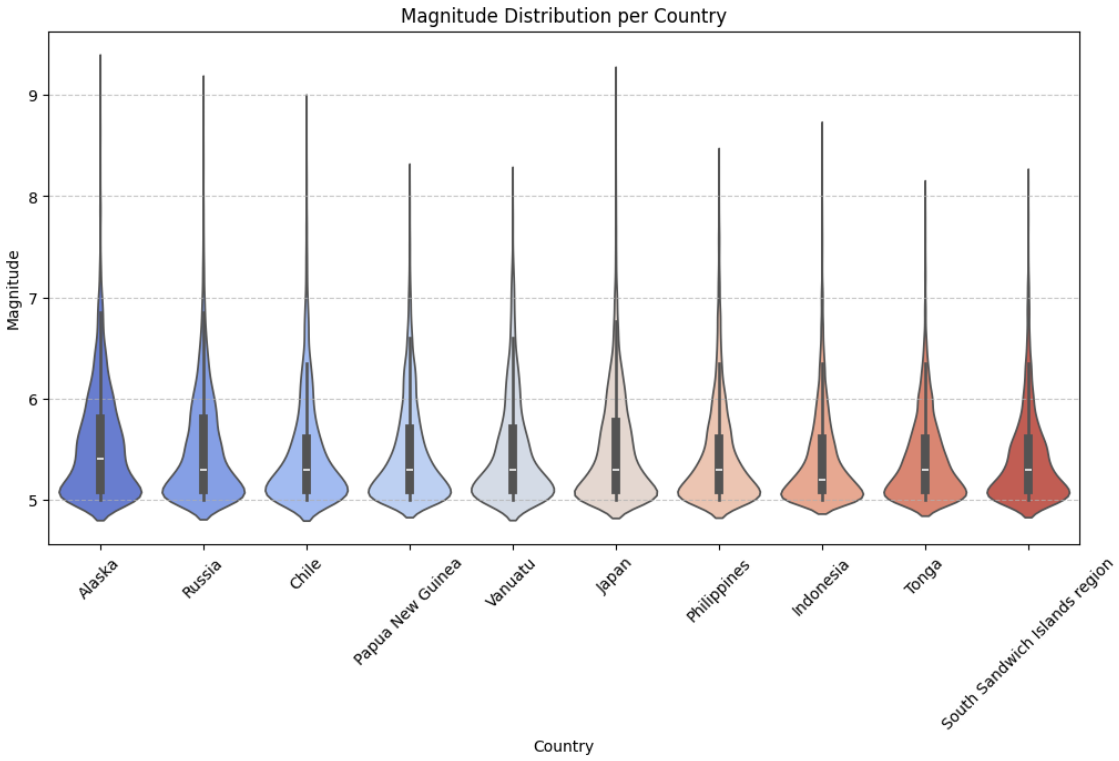

Fig 2.1.7: Violin Plot – Magnitude Distribution per Country

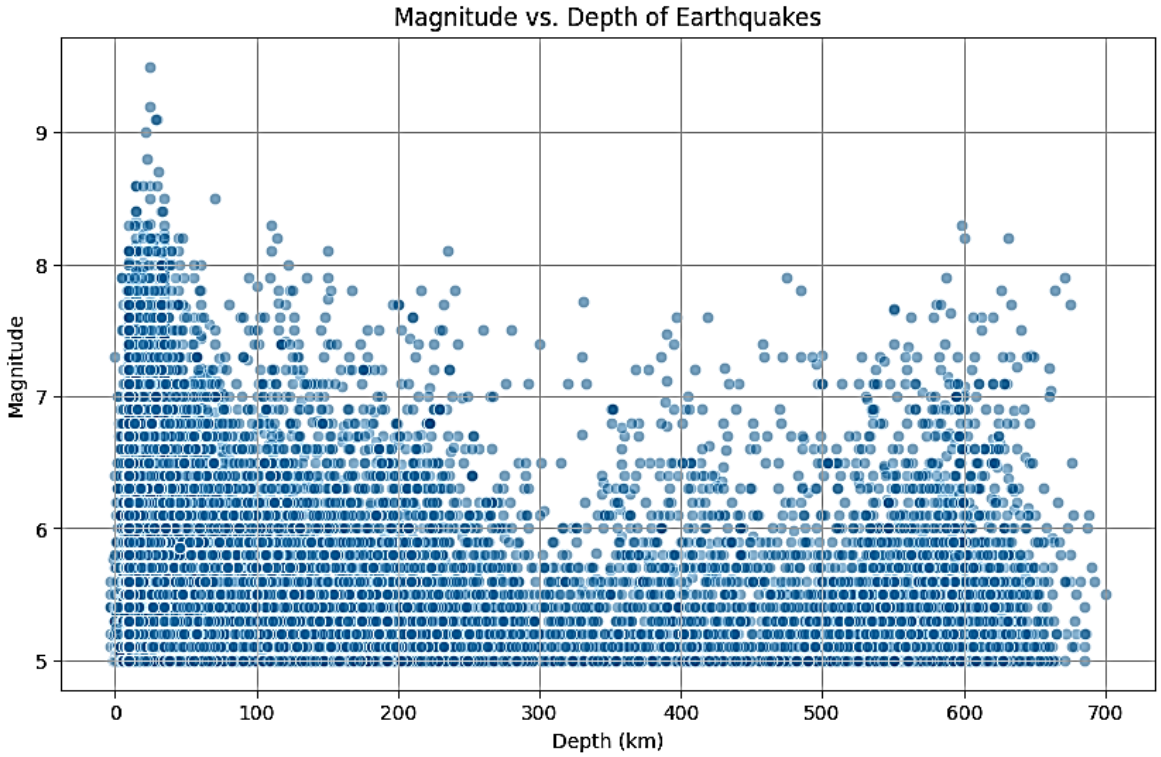

Fig 2.2.1: Scatter Plot – Earthquake Magnitude vs Depth

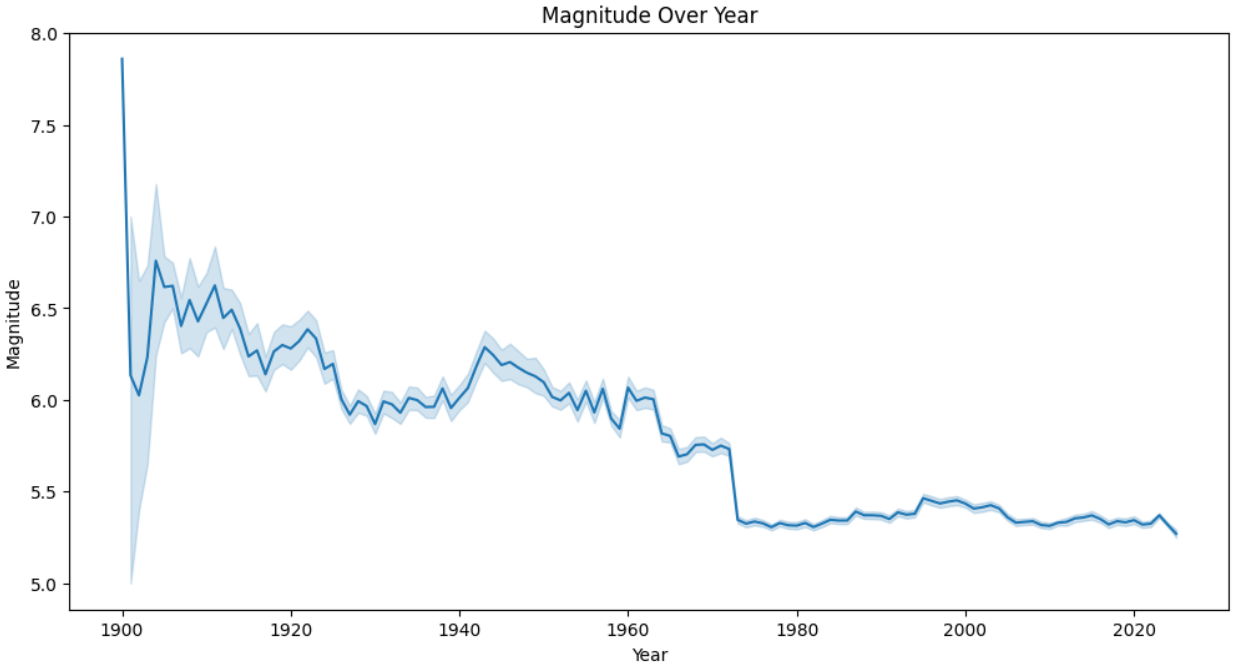

Fig 2.2.2: Line Plot – Earthquake Magnitude Trend Over Years

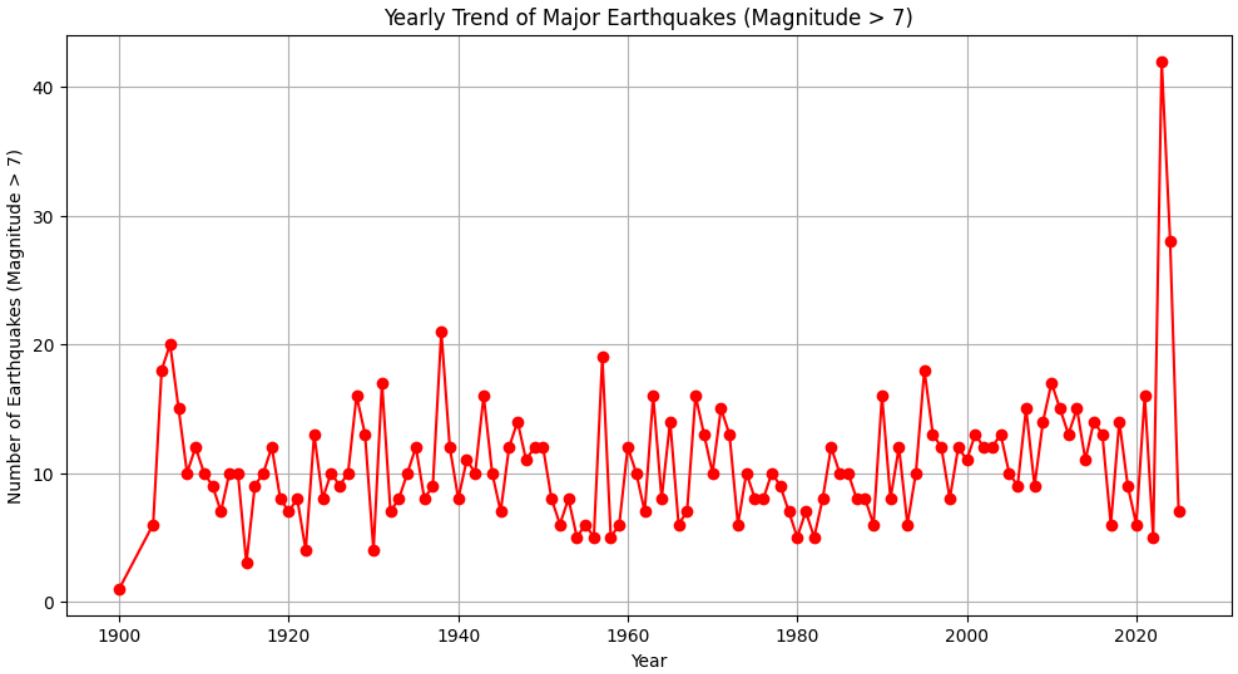

Fig 2.2.3: Line Plot – Yearly Trend of Major Earthquakes (Magnitude > 7)

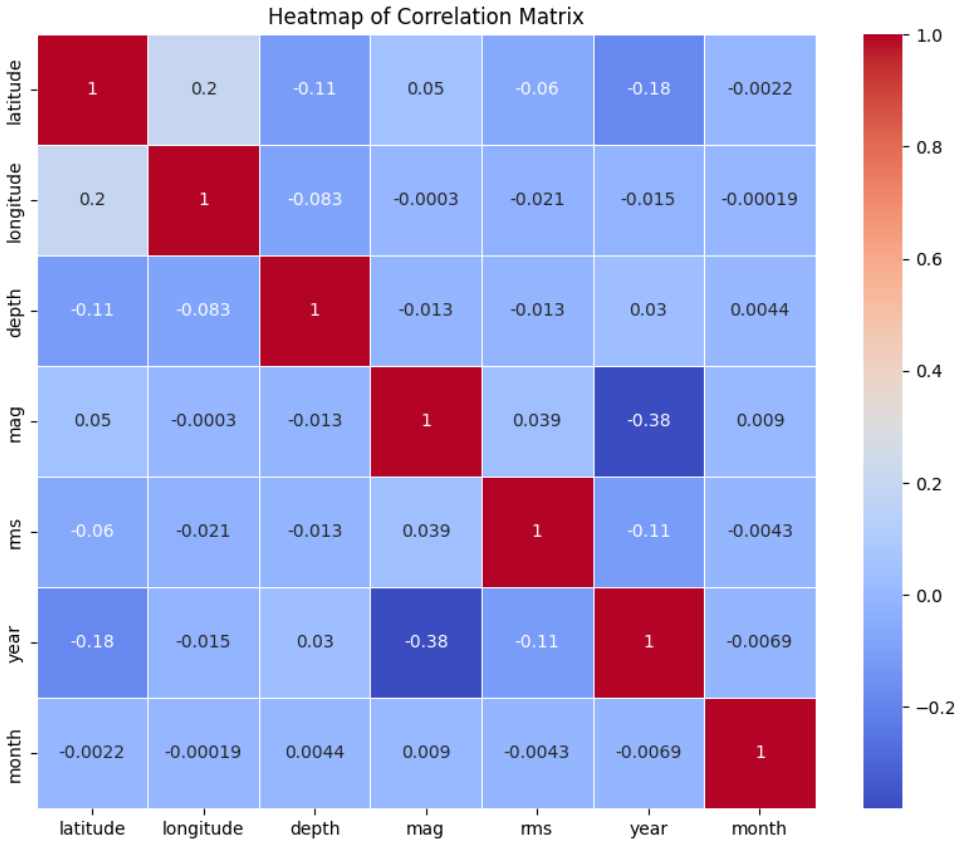

Fig 2.3.1: Heatmap – Correlation Among Earthquake Features

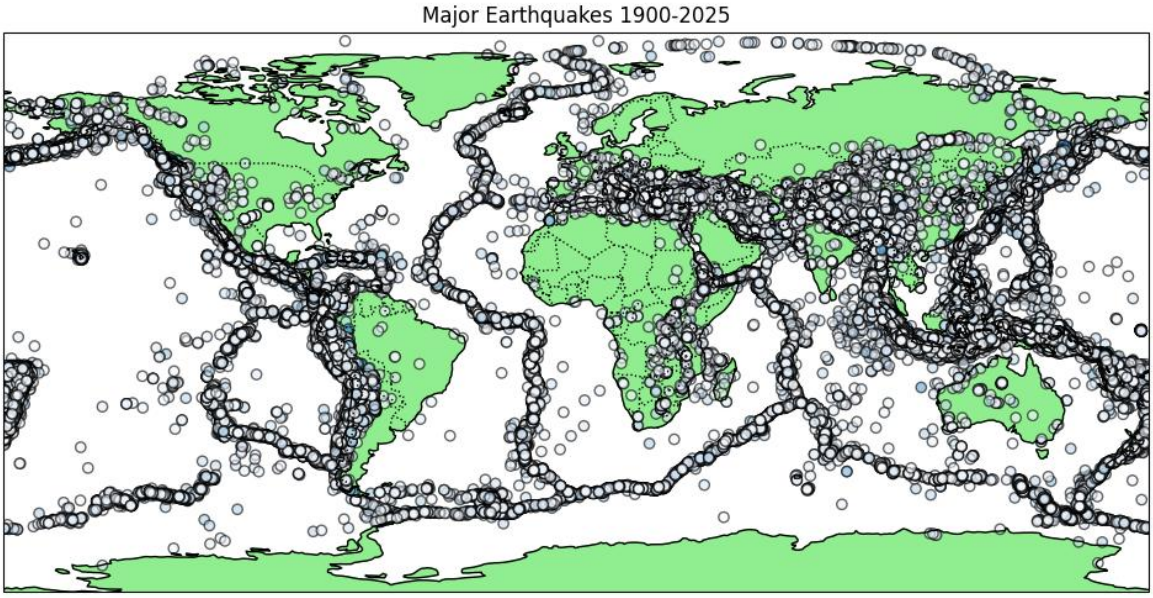

Fig 2.3.2: World Map – Global Distribution of Major Earthquakes (1900 to Present)

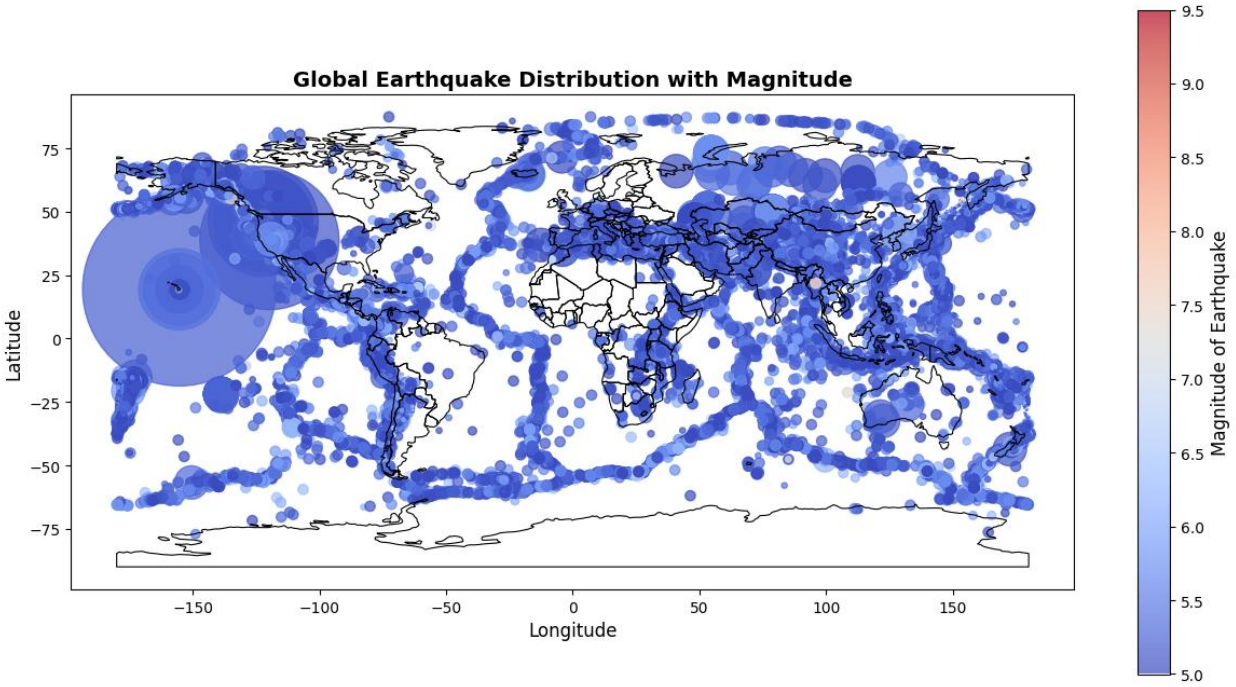

Fig 2.3.3: World Map – Earthquake Locations and Magnitude Representation

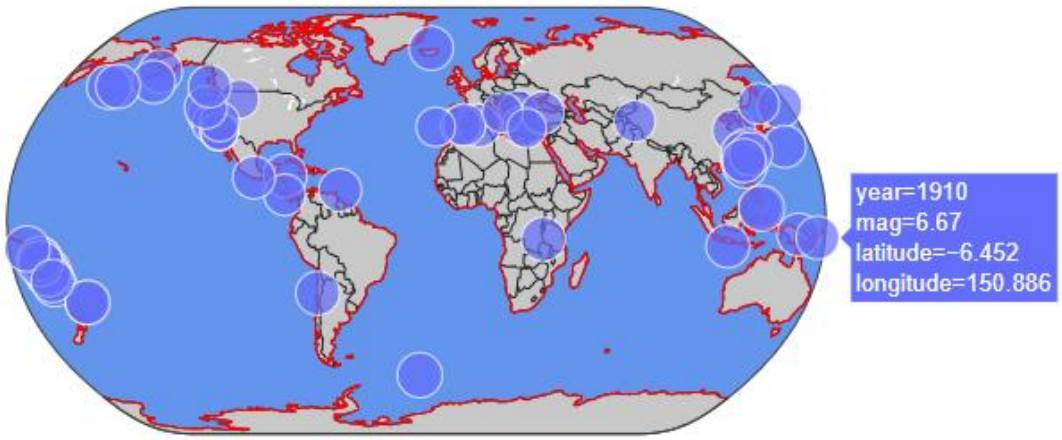

Fig 2.3.4: Animated World Map – Visualization of Earthquake Events

Insights from EDA

- **MissingValues:**
  Columns like depth have very few missing values, making them reliable for analysis. On the other hand, horizontalError has a lot of missing data, so it needs to be handled carefully. Columns such as nst, gap, and rms also have many missing values. These were either filled using imputation techniques or dropped if they were not useful. Visualizing the missing data was an important step, as it helped clean the dataset to make it ready for accurate modeling.
- **OutlierDetection:**
  Box plots showed some extreme values in depth, gap, and magnitude. These unusual values might be important because they could represent rare earthquakes. Instead of removing them, they were checked carefully to see how they affected the data. These outliers help show how different and unique some earthquakes can be.
- **Magnitude Distribution**:
  - **Magnitude Types**: The bar chart shows different types of magnitude used to measure earthquakes. The most common is mb, followed by mw, mwc, and mww. These are used most often by scientists. Less common types like Mi and Mh are

rarely used. This chart helps us understand how magnitudes are recorded.

- **Magnitude by Country**: This plot compares earthquake magnitudes in different countries. Countries like Japan and Indonesia have a wide range of magnitudes, showing they have both small and big earthquakes. Some areas, like South Sandwich Islands, show fewer changes in size. The thick part of the plot shows where most earthquakes happen in terms of magnitude.
- **Magnitude vs Depth**: This scatter plot shows the relationship between how deep the earthquake is and how strong it is. We can see that deeper earthquakes usually have smaller magnitudes, while shallow ones can be stronger. This helps us understand how depth affects an earthquake's impact.
- **MagnitudeOverTime:**The line graph shows how earthquake magnitudes have changed from 1900 to 2020. There is a drop in average magnitude around the 1960s, and then it stays low. This may be due to changes in how earthquakes were recorded or real changes in seismic activity.
- **Yearly Count of magnitude > 7: This** graph shows how many strong earthquakes (magnitude > 7) happened each year from 1900 to 2020. Some years have more earthquakes than others. A spike around 2020 may mean there was more activity or that earthquake detection got better.
- **Global Earthquake Map :**This world map shows where earthquakes happened. Bigger and darker circles mean stronger earthquakes. Most are seen along the Pacific Ring of Fire, which includes countries like Japan, Indonesia, and Chile. This helps us spot high-risk zones across the globe.

- **DepthAnalysis:**

  Most earthquakes are shallow (less than 100 km deep), but there are also deep-focus earthquakes that occur up to 700 km below the surface. These deep earthquakes are less frequent but can be

powerful.

A time-series analysis revealed an increasing trend in recorded earthquakes over the decades. This can be attributed to advancements in monitoring systems and wider data coverage, rather than an actual increase in earthquake frequency.

- **GeographicalPatterns:**

  This map shows where major earthquakes occurred around the world between 1900 and 2025. The continents are shown in green, oceans in white, and black dots represent earthquake locations. Many of these dots are grouped along tectonic plate boundaries, especially in active zones like the Pacific Ring of Fire, Mediterranean region, and the Himalayas. This map helps us clearly see which regions are more prone to frequent and strong earthquakes

- **CountryWiseDistribution:**

  The pie chart shows which countries have the most earthquakes. Indonesia has the highest number, making up 22.9% of all earthquakes. Next are Papua New Guinea (12.2%) and Japan (12.1%). Other countries like the Philippines, Tonga, Russia, Alaska, Vanuatu, Chile, and the South Sandwich Islands have smaller shares between 6.3% and 9.6%. This chart helps us see which countries face earthquakes most often, with Indonesia clearly being the most affected.

- **CorrelationMatrix:**

  The heatmap shows the correlation between different variables, such as latitude, longitude, depth, magnitude, root mean square (rms), year, and month. The colour scale ranges from blue (negative correlation) to red (positive correlation). It highlights that most variables have weak correlations with one another. For example, latitude and longitude have a slight positive correlation, while depth and year show a weak negative correlation. This indicates that no strong linear relationships exist among these variables.

- **AnimatedWorldMap:**

  This map shows where earthquakes have happened all over the world. Blue circles mark each earthquake location. When you hover

or click on a circle, it shows important details like the year, magnitude, latitude, and longitude of the event. This type of map helps us easily see which parts of the world experience the most earthquakes and spot any patterns in their location.

3.Feature Engineering

Feature engineering is a critical step in machine learning that involves transforming raw data into meaningful features that improve model performance. In this project, we applied several techniques to enrich our dataset and enhance predictive power.

1. Time-Based Feature Extraction

The time column in the dataset originally provided timestamps in ISO 8601 format (e.g., 2021-03-04T12:34:56.000Z). We converted this column into a datetime format using pandas.to_datetime(), allowing us to extract meaningful temporal features, such as:

- Year: To analyse earthquake trends over time.
- Month: For identifying seasonal or monthly earthquake patterns.
- is_night: A custom binary feature where 1 indicates the earthquake occurred during nighttime hours (before 6 AM or after 6 PM), and 0 indicates daytime. These time-derived features help the model better understand potential temporal patterns in earthquake occurrences.

2. Extracting Country Information from ‘place’ Column

The place column in the dataset contains human-readable location descriptions like:"10km NE of Santiago, Chile”. Since the dataset does not have a separate country column, we applied text processing techniques to extract the country name from the place string. This was typically done by splitting the string from the rightmost comma (,), assuming the last part represents the country

4. Modeling

We trained and tested five different machine learning models to compare their performance. which are the following:

1. **Linear Regression :**
   - A simple model that predicts the output based on a linear relationship between inputs and output.
   - Ideal for datasets with linear trends but struggles with complex, non-linear problems.
2. **Decision Tree Regressor :**
   - A tree-based model that splits data into smaller subsets based on conditions and makes predictions.
   - Easy to interpret and works well for non-linear relationships, but can overfit without proper pruning.
3. **Support Vector Regressor (SVR) :**
   - Uses support vectors to find the best fit line or curve for regression tasks.
   - The RBF kernel is commonly used to model non-linear relationships effectively.
4. **Random Forest Regressor :**
   - An ensemble model that combines multiple decision trees to improve accuracy and reduce overfitting.
   - Robust and versatile, handling complex datasets better than a single decision tree.
5. **Random Forest Regressor (Hyperparameter Tuned) :**
   - A refined version of Random Forest with optimized settings to enhance performance:
   - Number of trees: 1500
   - Max depth: 20
   - Minimum samples to split: 5
   - Minimum samples at leaf: 2

5. Model Evaluation

To check how well the models performed, we used the following metrics:

- **R^2 Score**: Tells how much of the variation in magnitude the model can explain higher is better

- **MSE (Mean Squared Error)**: The average of squared differences between actual and predicted values lower is better
- **RMSE (Root Mean Squared Error)**: The square root of MSE, gives error in the original unit (magnitude).

Model	**R^2 Score**	**MSE**	**RMSE**
Linear Regression	0.394946	0.143022	0.378182
Decision Tree	0.279359	0.170344	0.412727
Random Forest	0.617199	0.090486	0.300809
Support Vector Machine	0.533215	0.110338	0.332171
Random Forest (Tuned)	0.626290	0.088337	0.297215

Table 5.1 Model Evaluation

6. Web Appilcation

We developed a web application using Flask, a lightweight and powerful Python framework for backend development. Flask handles loading the trained earthquake magnitude prediction model, receiving user input, processing it, and returning the predicted result. On the frontend, we used HTML and CSS to design a clean and interactive interface. The form collects important earthquake details like latitude, longitude, depth, and magType. We also integrated a world map where users can click to select a location, making it more user-friendly. Country names are auto-filled based on the selected location to improve accuracy. This seamless integration of Flask backend with an HTML, CSS, and JavaScript frontend creates an engaging experience for predicting earthquake magnitudes easily

Result

After training the Random Forest Regressor model on the earthquake dataset, the model achieved strong performance. The model obtained a Mean Squared Error (MSE) of 0.0228, a Root Mean Squared Error (RMSE) of 0.1511, and an R^2 Score of 0.6261 on the test set.

The R^2 score indicates that approximately 62.6% of the variance in earthquake magnitudes can be explained by the model's input features. This demonstrates that the model is reasonably accurate and can effectively predict earthquake magnitudes based on inputs like latitude, longitude, depth, magType, and other seismic factors.

Overall, the Random Forest model performed well and provides reliable predictions for the web application.

Key Findings:

- The Random Forest Regressor model achieved a good performance with an R^2 score of 0.62, showing a decent ability to predict earthquake magnitudes.
- Important features influencing magnitude prediction included depth, latitude, longitude, gap, and magType.
- Proper handling of missing values (like filling medians, means, and modes) significantly improved the model's stability and accuracy.
- Feature scaling (using StandardScaler) helped the Random Forest model perform better and Encoding categorical features such as magType using Label Encoding made the data suitable for machine learning.
- Earthquake type filtering (removing explosions, volcanic eruptions, etc.) ensured that the model learned only from true earthquake data.
- Adding extracted features like year from the timestamp helped in capturing time-related patterns in the data.
- The final trained model, along with the scaler and label encoder, was successfully saved using pickle for deployment into the Flask web application.
- The web application visually enhances user experience by integrating interactive maps and country-level information

Discussion

Limitations

- The model is based only on historical data, so it may not generalize well to future or very rare seismic events.
- Some features had missing values, and dropping rows may reduce overall data size and affect model generalization.
- MagType and other categorical columns were label encoded, which does not consider relationships between categories (like ordinal importance).
- The dataset did not include real-time seismic waveforms or tectonic fault information, which could enhance predictions.

Future Work

- Include real-time seismic sensor data for early-warning systems.
- Add more geographic and geological features (e.g., tectonic plate data, soil type).
- Use deep learning models
- Try more advanced feature engineering like clustering or interaction features.
- Integrate the model into mobile or web-based alert systems with maps and real-time predictions.

Implications

- Earthquake magnitude prediction can support disaster planning and early warnings, potentially reducing loss of life and damage.
- This model can be embedded in local or national alert systems to provide risk estimations for decision-makers.
- Governments and researchers can use the model's insights to target infrastructure reinforcements in high-risk zone

Conclusion

Summary of Findings

- Multiple machine learning models were trained and tested to predict earthquake magnitudes.
- Among all, the Hyperparameter-Tuned Random Forest Regressor performed best, achieving a high R^2 score (~0.62).
- The most important features for prediction were: depth, latitude, longitude, magType, and gap.
- Preprocessing steps like label encoding and standard scaling significantly helped model performance.
- The project lays a solid foundation for future real-time earthquake prediction systems.

Impact

- This project shows how machine learning can be applied to natural disaster prediction. Accurate predictions of earthquake magnitude can help authorities:
- Save lives by issuing earlier warnings.
- Plan evacuation routes and medical support in high-risk zones.
- Reduce economic loss through better building codes and preparedness.

Recommendations

- Use the final Random Forest model in a web interface or dashboard for real time predictions.
- Regularly update and retrain the model as more recent data becomes available.
- Collaborate with geoscientists and disaster response agencies for practical deployment.

References

- **Kaggle**: https://www.kaggle.com/datasets/usamabuttar/significant-earthquakes/data
- https://www.trantorinc.com/blog/exploratory-data-analysis
- **JavaScript library used to create interactive maps** : https://leafletjs.com/
- https://www.openstreetmap.org :reverse geocoding (**converting coordinates to a country name**)
- https://nominatim.org/release-docs/develop/api/Search
- https://flask.palletsprojects.com/en/stable :**Flask**
- https://www.w3schools.com/html/ : **Html**