EarthquakeSense
Machine Learning on Earthquake Impact and Damage

STEPHANIE NHI LE | EZANA SEYOUM | BEAMLAK ABDISA | AMANUEL DEMISSIE

Introduction

- This study looks at how earthquakes have severe damage in certain area.
- The objective is to use data on earthquake magnitude, location and depth to create a predictive models.
- The chance of major damage will be estimated using statistical models, which will serve as the basis for the analysis of different risk factors.
- To visualize data, detect trends, and guide more modeling, Exploratory Data Analysis (EDA), will be carried out.
- Evaluate whether Linear Regression or Logistic Regression implies a better efficiency on predicting and measuring the damage impact.

Research Question

How effectively can a predictive model, using earthquake characteristics such as magnitude, location, and depth, estimate the likelihood of severe damages in certain areas?

Objective: The study aims to explore the effects of earthquakes, with a particular focus on identifying the regions that are most vulnerable to severe damage during seismic events.

Context: Understanding the factors that contribute to earthquake damage is critical for improving disaster preparation, and possibly developing effective mitigation strategies.

Inspiration: Motivated by recent events and their devastating impact on communities, this research seeks to leverage historical data to predict and mitigate future damages.

Data Source: NOAA National Centers for Environmental Information (NCEI) from 1995 - 2024

Parameters: Key parameters include magnitude, location (latitude and longitude), and depth of the earthquake.

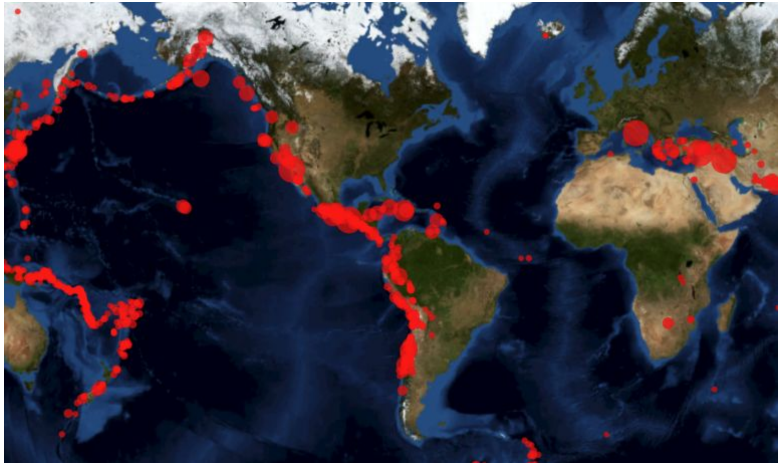
Preprocessing: Data cleaning involves handling missing values, normalizing location coordinates, and categorizing earthquake magnitudes into appropriate bins for analysis.

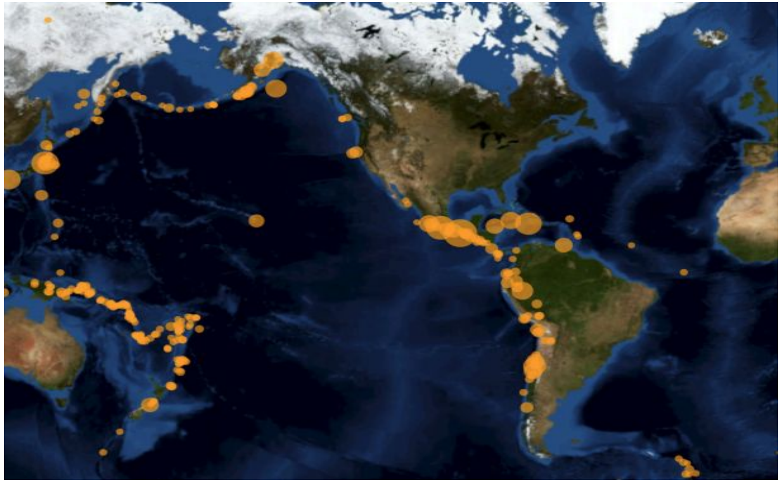
Analytical Approach: The study employs **Linear Regression** and **Logistic Regression** modeling to examine the relationship between earthquake characteristics.

Exploratory Data Analysis (EDA): To understand the distributions and interrelationships of the variables. Techniques used include creating histograms to visualize frequency distributions, scatter plots to explore correlations, and regression lines to model these relationships.

Visualizations: Histograms, Scatter Plots, Regression Lines, and Geo HeatMap.

Year	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
1995	9	14	14	4	31.4	2251		MEXICO: GUERRERO, OAXACA, PUEBLA, MEXICO CITY	16.779	-98.597	23	7.4		3	1			100	2		2	
1995	10	6	5	23	18.5			ALASKA: FAIRBANKS NORTH STAR COUNTY	65.17	-148.565	9	6									1	
1995	10	9	15	35	53.9	2252		MEXICO: JALISCO, MANZANILLO, SAN PATRICIO MELAQUE	19.055	-104.205	33	8		49	2			200	3		2	
1996	2	25	3	8	15.8	2262		MEXICO: OFF COAST OF GUERRERO	15.978	-98.07	21	7.1										
1996	6	10	4	3	35.4	2263		ALASKA: ANDREANOF ISLANDS	51.564	-177.632	33	7.9	6									
1996	6	10	15	24	56	2264		ALASKA: ANDREANOF ISLANDS	51.478	-176.847	24	7.3										
1997	1	11	20	28	26			MEXICO: MICHOACAN, ARTEAGA	18.219	-102.756	33	7.2		1	1						2	
1998	2	3	3	2	0.2			MEXICO: OAXACA, SAN AGUSTIN, SAN FRANCISCO	15.883	-96.298	33	6.3									2	
1999	6	15	20	42	5.9			MEXICO: PUEBLA, VERACRUZ, OAXACA, MORELOS, GUERRERO	18.386	-97.436	70	7		20	1			200	3	226.8	4	
1999	6	21	17	43	4.5			MEXICO: GUERRERO: COAHUAYUTLA; MICHOACAN: CUITZEO	18.324	-101.539	69	6.3									3	
1999	9	30	16	31	15.6			MEXICO: OAXACA	16.059	-96.931	61	7.5	8	35	1			215	3	164.8	4	
1999	10	16	9	46	44.1			CALIFORNIA: LUDLOW, LANDERS, TWENTYNINE PALMS	34.594	-116.271		7.2	7					4	1		1	
2000	9	3	8	36	30			CALIFORNIA: NAPA	38.379	-122.413	10	5	7					41	1	50	4	
2001	2	28	18	54	32.8			WASHINGTON: OLYMPIA, SEATTLE, TACOMA	47.149	-122.727	52	6.8	8	1	1			400	3	2000	4	
2001	9	9	23	59	18			CALIFORNIA: LOS ANGELES	34.059	-118.387	5	4.2	6								1	
2001	10	12	5	2	34	5609		CANADA: QUEEN CHARLOTTE ISLANDS	52.63	-132.2	20	6.1										
2002	1	30	8	42	3.4			MEXICO: VERACRUZ: SAN ANDRES TUXTLA, TUXTEPEC	18.194	-95.908	109	5.9									1	
2002	2	22	19	32	41.7			MEXICO: MEXICALI, BAJA CALIFORNIA	32.319	-115.322	7	5.5									1	
2002	4	20	10	50	47.5			NEW YORK: CLINTON, ESSEX, AU SABLE FORKS	44.513	-73.699	11	5.2	7								1	
2002	9	25	18	14	48.5			MEXICO: ACAPULCO	16.87	-100.113	6	5.3						2	1		1	
2002	10	23	11	27	19.4			ALASKA: CANTWELL, DENALI NATL PARK	63.514	-147.912	4	6.7	8								2	
2002	11	3	22	12	41			ALASKA: SLANA, MENTASTA LAKE, FAIRBANKS	63.517	-147.444	5	7.9	9					1	1	56	4	
2003	1	22	2	6	34.6	2402		MEXICO: VILLA DE ALVAREZ, COLIMA, TECOMAN, JALISCO	18.77	-104.104	24	7.5	8	29	1			300	3		3	2005
2003	2	22	12	19	10.5			CALIFORNIA: BIG BEAR CITY	34.31	-116.848	1	5.2	6								1	
2003	4	29	8	59	39			ALABAMA: FORT PAYNE,GAYLESVILLE,VALLEY HEAD	34.494	-85.629	20	4.6	6								1	
2003	6	6	12	29	34			KENTUCKY: BARDWELL	36.87	-88.98	3	4	6								1	
2003	11	17	6	43	6.8	2429		ALASKA: ALEUTIAN ISLANDS: RAT ISLANDS	51.146	178.65	33	7.8										
2003	12	22	19	15	56			CALIFORNIA: PASO ROBLES,TEMPLETON,ATASCADERO	35.706	-121.102	8	6.6	8	2	1			40	1	300	4	40
2004	1	1	23	31	50			MEXICO: GUERRERO, MEXICO CITY	17.488	-101.303	29	6.1									1	
2004	9	28	17	15	24.2			CALIFORNIA: CENTRAL: PARKFIELD, SAN MIGUEL	35.819	-120.364	9	6	6								1	
2004	11	2	10	2	12.8	3012		CANADA: VANCOUVER ISLAND	49.277	-128.772	10	6.6										
2005	6	15	2	50	53.1	2547		CALIFORNIA: OFF COAST NORTHERN	41.301	-125.97	10	7.2	4									
2005	7	26	4	8	37.1			MONTANA: DILLON, SILVER STAR, TWIN BRIDGES	45.365	-112.615	13	5.6	6								1	
2006	10	15	17	7	49.2	3017		HAWAIIAN ISLANDS	19.878	-155.935	39	6.7	8						3	73	4	
2007	4	13	5	42	23			MEXICO: GUERRERO, ATOYAC	17.302	-100.198	34	6	5								1	2
2007	5	8	15	46	49.1			MONTANA: SHERIDAN	45.394	-112.13	14	4.5	5								1	
2007	7	20	11	42	22.3			CALIFORNIA: MONTCLAIR	37.804	-122.193	5	4.2									1	
2007	8	2	3	21	42.8	3156		ALASKA: ALEUTIAN ISLANDS	51.307	-179.971	21	6.7										
2007	8	6	8	48	40			UTAH: HUNTINGTON	39.465	-111.237	2	4.2		9	1				1			
2007	8	17	0	38	56			UTAH	39.464	-111.207	0	1.6		3	1			6	1			
2007	10	31	3	4	54.8			CALIFORNIA: SAN JOSE	37.434	-121.774	10	5.6									1	
2008	2	9	7	12	5.8			MEXICO: BAJA CALIFORNIA	32.456	-115.315	3	5.1									1	

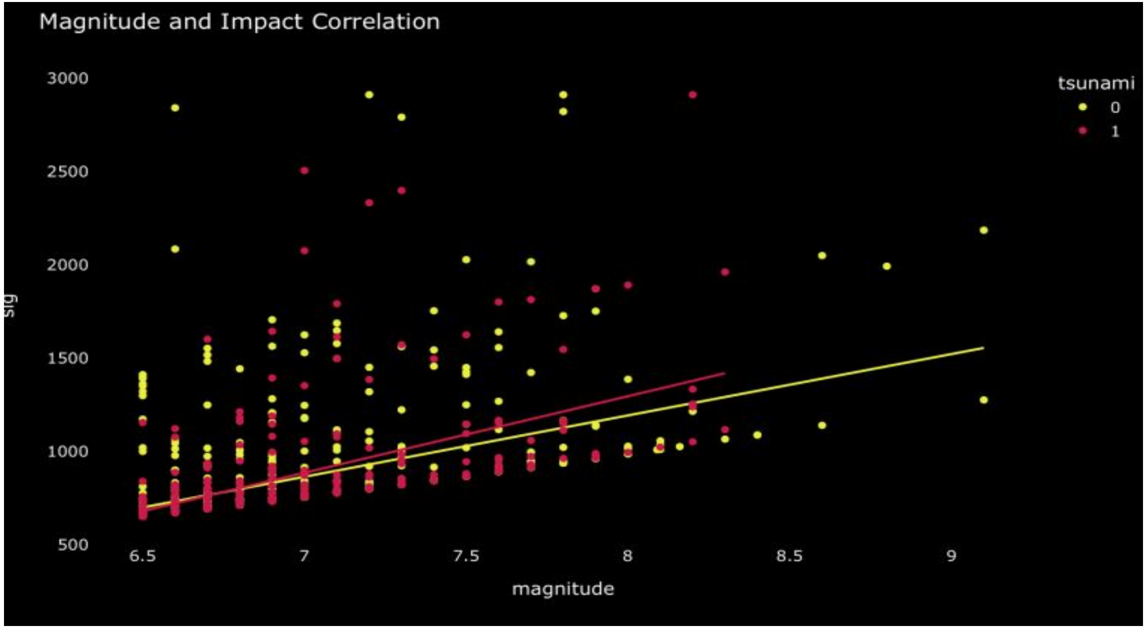




Analysis and results

We began by looking at a heat map to look at where earthquakes are located and the magnitude of the earthquake. Using the heatmap, we decided to focus on certain areas such as the west coast of USA to start small and work our way up. In correlation with earthquakes, we also saw that tsunamis are also significant signs of earthquakes.

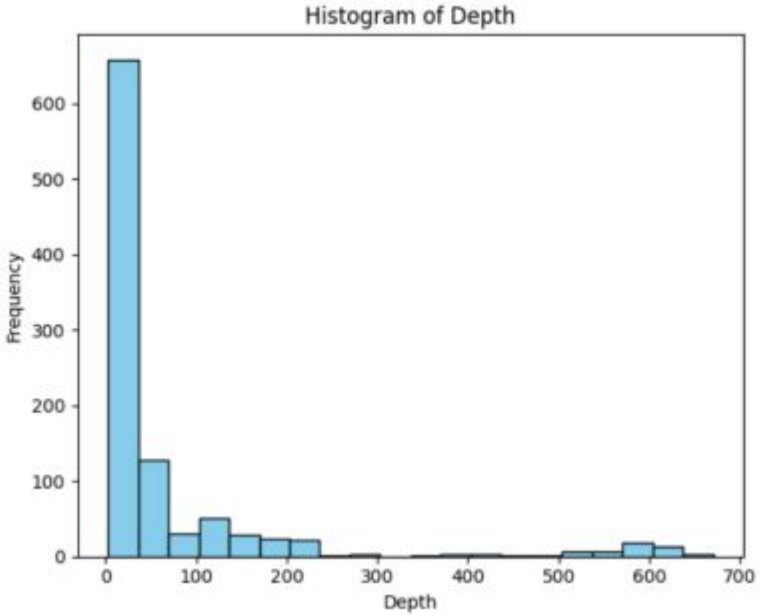
While comparing the two figures, we saw how they were very similar. Then using the data, we decided to do a scatter plot to see the exact correlation for earthquakes when there is a tsunami and when there isn't.



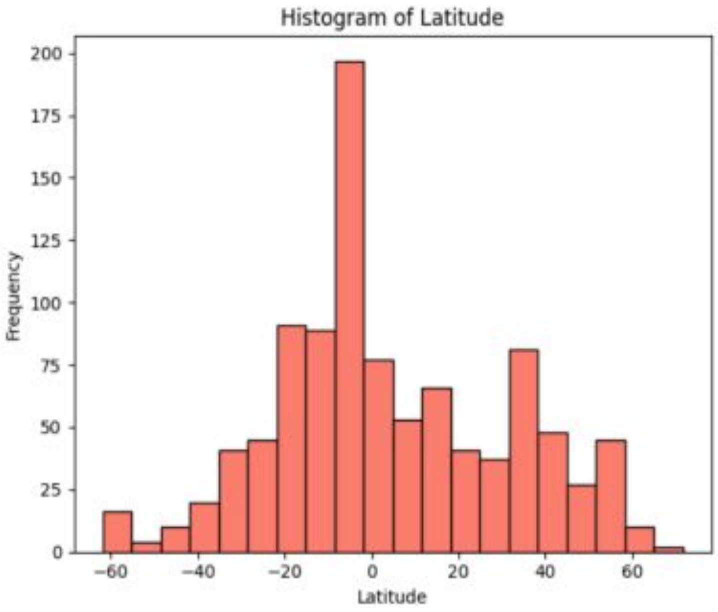
The graph illustrates the correlation between earthquake magnitude and impact severity, with an additional indication of whether a tsunami occurred. Here are the key observations:

1. **Positive Correlation**: There's a clear upward trend showing that as earthquake magnitude increases, so does the severity of its impact.
2. **Tsunami Indicator**: Earthquakes that resulted in tsunamis (red dots) are more frequent at higher magnitudes.
3. **Variability**: The spread of data points increases with magnitude, indicating greater variability in impact at higher magnitudes.
4. **Regression Line**: The line indicates a general trend but also highlights that magnitude alone does not account for all variations in impact.

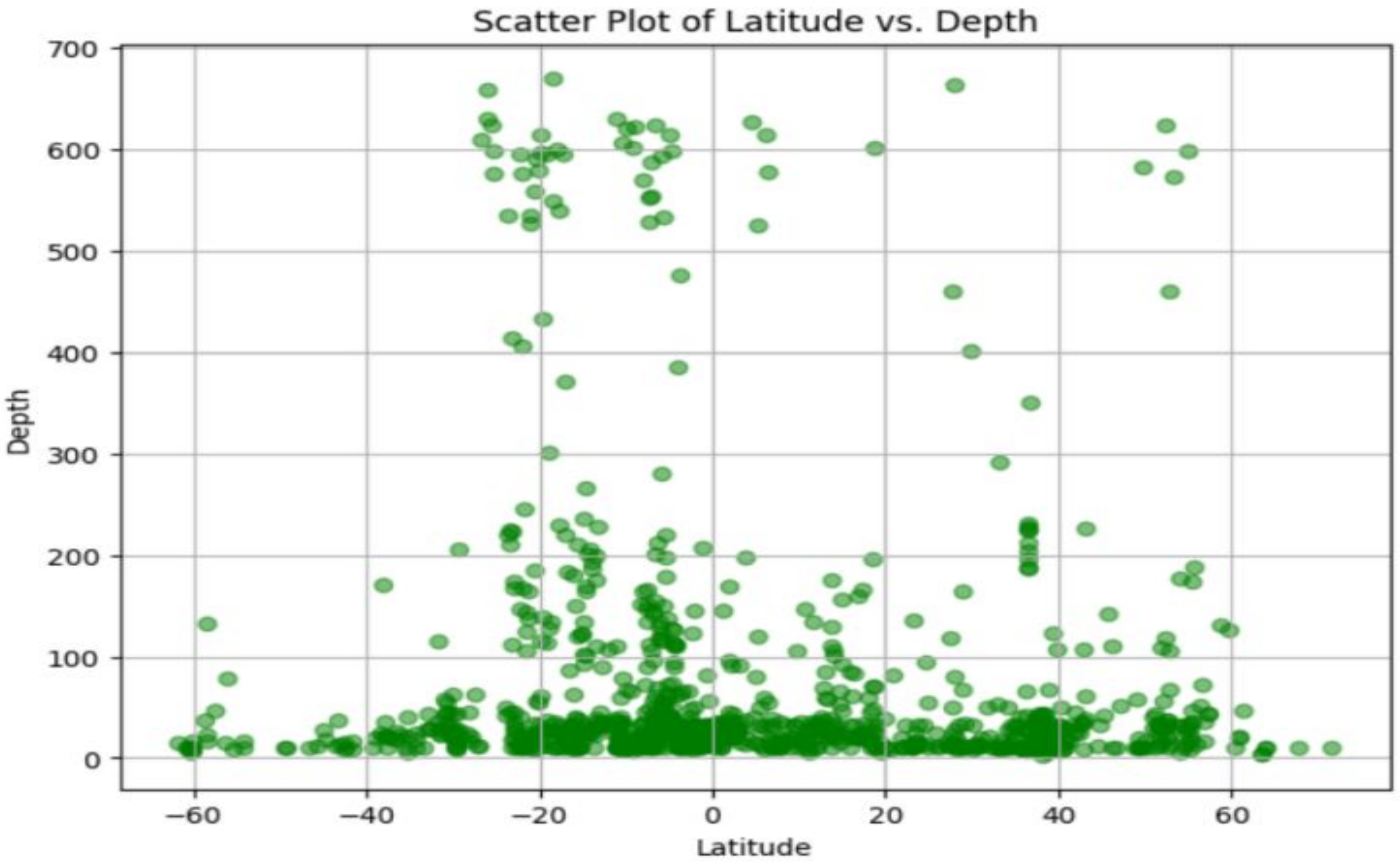
The graph underscores magnitude as a significant predictor of impact but also highlights the complex dynamics when tsunamis are involved, emphasizing the need for comprehensive risk assessments in earthquake-prone areas.



We used Histograms to provide insights into the distribution of earthquake depths and latitudes based on your dataset. The histogram of depth shows a strong skew towards shallower depths, with the majority of earthquakes occurring at depths less than 100 km. This is a common characteristic, as most seismic activity associated with tectonic plate interactions happens within this range.

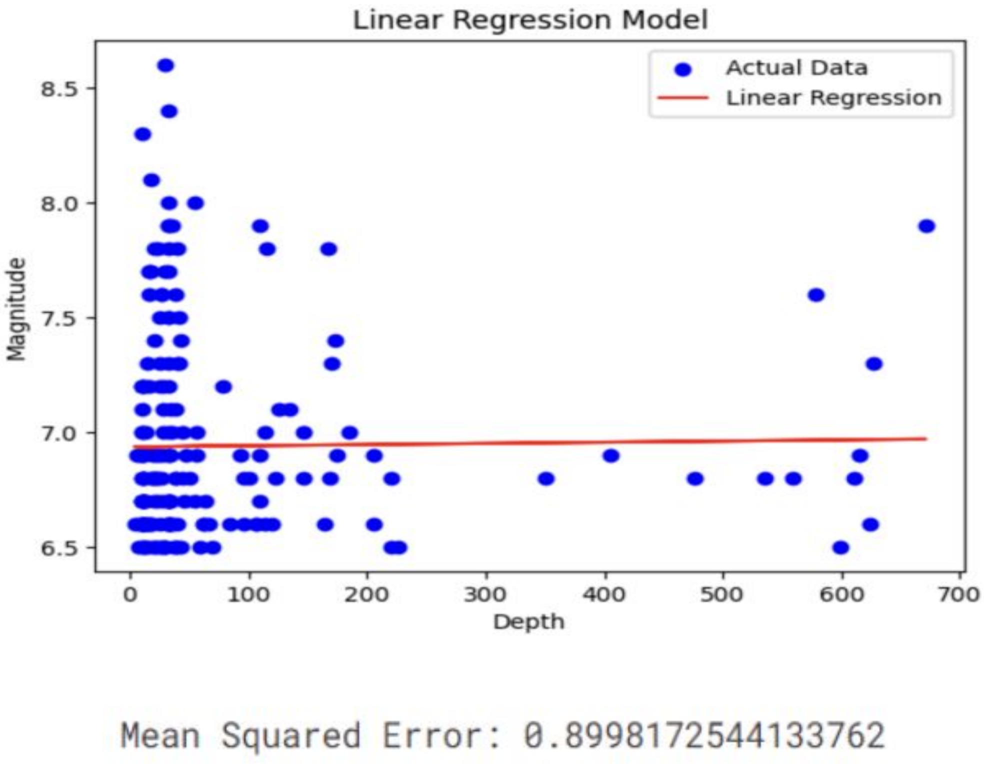


The histogram of latitude shows that the majority of earthquakes occur within a band around the equator between about -20 to 20 degrees latitude. This could be indicative of tectonic activity near the equatorial belt, which includes regions like the Pacific "Ring of Fire". There are also noticeable occurrences in both the northern and southern hemispheres, though these are less frequent than around the equatorial region. This spread might indicate global tectonic plate boundaries extending into higher latitudes. The concentration of earthquakes around the equatorial region could be due to the active tectonic boundaries in these areas, including subduction zones, transform faults, and rift zones that are geographically near the equator


Scatter Plot of Latitude vs. Depth
Depth
Latitude
700
600
500
400
300
200
100
0
−60
−40
−20
0
20
40
60

The scatter plot shows a wide range of earthquake depths across different latitudes. Most of the earthquakes are clustered at shallower depths (less than 100 km), which is typical for crustal earthquakes. However, there are significant occurrences of deeper earthquakes (over 300 km), particularly in specific latitude bands. Understanding the depth and location distribution of earthquakes can help in refining models for earthquake prediction and in improving preparedness and mitigation strategies in earthquake-prone regions

A graph depicting a linear regression analysis attempting to model the relationship between earthquake magnitude and depth.



Analysis of the Plot

- Data Points: The blue dots represent actual earthquake events, with the magnitude on the y-axis and depth on the x-axis.

- Linear Regression Line: The red line represents the linear regression fit to the data, which attempts to model the relationship between depth and magnitude.

Why Linear Regression May Not Be Suitable

1. Non-Linear Relationship: Earthquake data often involve complex relationships that are not linear.

2. High Variability: Earthquake magnitudes at similar depths can vary widely due to other factors not included in the model, such as the energy release, fault characteristics, and geological structures. This variability makes it difficult for a simple linear model to provide accurate predictions.

3. Residual Analysis: The suitability of a linear model can also be evaluated by analyzing the residuals (the differences between observed and predicted values).

Testing and training the data based on linear regression model

```
# Split the data into a training set and a test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25)

# Create the Linear Regression model
model = LinearRegression()

# Train the model
model.fit(X_train, y_train)

# Predict using the model
y_pred = model.predict(X_test)

# Calculate the mean squared error
mse = mean_squared_error(y_test, y_pred)

# Print the mean squared error
print("Mean Squared Error:", mse)
```

Outcome results comparing the actual and predicted magnitude

```
Actual Magnitude: 2.90  |  Predicted Magnitude: 1.76
Actual Magnitude: 1.60  |  Predicted Magnitude: 1.14
Actual Magnitude: 0.28  |  Predicted Magnitude: 1.48
Actual Magnitude: 1.60  |  Predicted Magnitude: 0.91
Actual Magnitude: 3.40  |  Predicted Magnitude: 2.61
Actual Magnitude: 0.60  |  Predicted Magnitude: 1.58
Actual Magnitude: 1.10  |  Predicted Magnitude: 1.57
Actual Magnitude: 2.70  |  Predicted Magnitude: 1.57
Actual Magnitude: 0.88  |  Predicted Magnitude: 1.48
```

An MSE of 0.89 indicates that the model's predictions deviate from the actual values by a mean squared error of 0.89. It's advisable to compare this MSE with benchmarks or alternative models, and consider refining the model based on a deeper analysis of its current performance and the underlying data characteristics.

Implications for Using Linear Regression

- **Poor Model Fit:** The poor fit of the linear regression model suggests that it may not be the best method for predicting earthquake magnitude based on depth alone.
- **Consideration of Other Variables:** It might be necessary to include additional variables such as geographical location, tectonic settings, or historical seismic activity to improve the model's accuracy.
- **Exploration of Non-Linear Models:** Given the apparent non-linear relationship, exploring other types of models could yield more accurate predictions.

Logistic Regression

Objective: Build a logistic regression model to predict the likelihood of an earthquake causing damage based on various features such as magnitude, depth, latitude, longitude, and the occurrence of a tsunami.

Data: NOAA Source - contains information on earthquakes that occurred between 1995 and 2024, including their characteristics and reported damage.

Design Choice: A statistical modeling technique used to analyze the relationship between a binary dependent variable (in this case, whether an earthquake caused damage or not) and one or more independent variables (earthquake features).

Logistic Regression Model

$$P(Damage = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot \mathrm{Mag} + \beta_2 \cdot \mathrm{Focal\ Depth} + \beta_3 \cdot \mathrm{Latitude} + \beta_4 \cdot \mathrm{Longitude} + \beta_5 \cdot \mathrm{Tsu})}}$$

where:

- $P(Damage = 1|X)$ is the probability of damage given the feature set X.
- β_0 is the intercept of the model.
- $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are the coefficients corresponding to each feature.

Training, Prediction & Evaluation Metrics

Prediction:

$$\hat{Y}_j = \begin{cases} 1 & \text{if } P(Damage = 1|X_j) \geq 0.5 \\ 0 & \text{if } P(Damage = 1|X_j) < 0.5 \end{cases}$$

where $\hat{Y}_j$ is the predicted label for the j-th test sample.

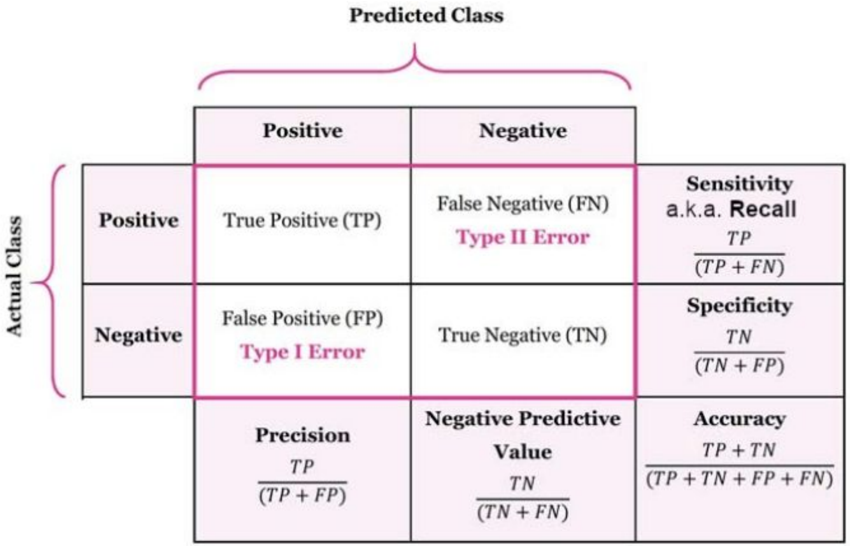
Accuracy:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Classification Report includes precision, recall, and F1-score for both classes (damage and no damage).

Classification Report

```
Accuracy: 0.7142857142857143
Classification Report:
              precision    recall  f1-score   support

           0       0.67      1.00      0.80         4
           1       1.00      0.33      0.50         3

    accuracy                           0.71         7
   macro avg       0.83      0.67      0.65         7
weighted avg       0.81      0.71      0.67         7
```



Train The Model

Classification Report

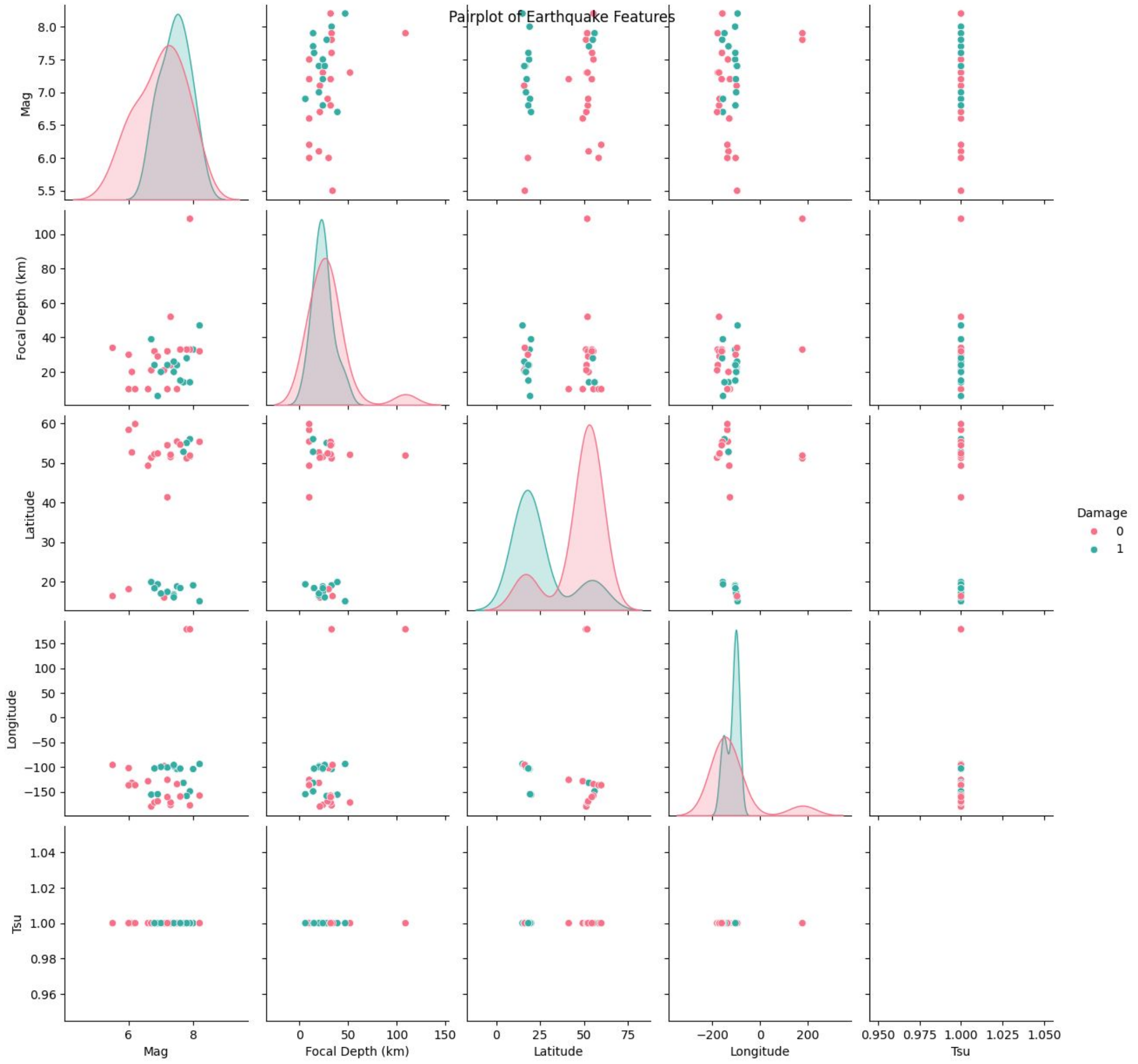
```
Accuracy: 0.7142857142857143
Classification Report:
              precision    recall  f1-score   support

           0       0.67      1.00      0.80         4
           1       1.00      0.33      0.50         3

    accuracy                           0.71         7
   macro avg       0.83      0.67      0.65         7
weighted avg       0.81      0.71      0.67         7
```

Data Processing

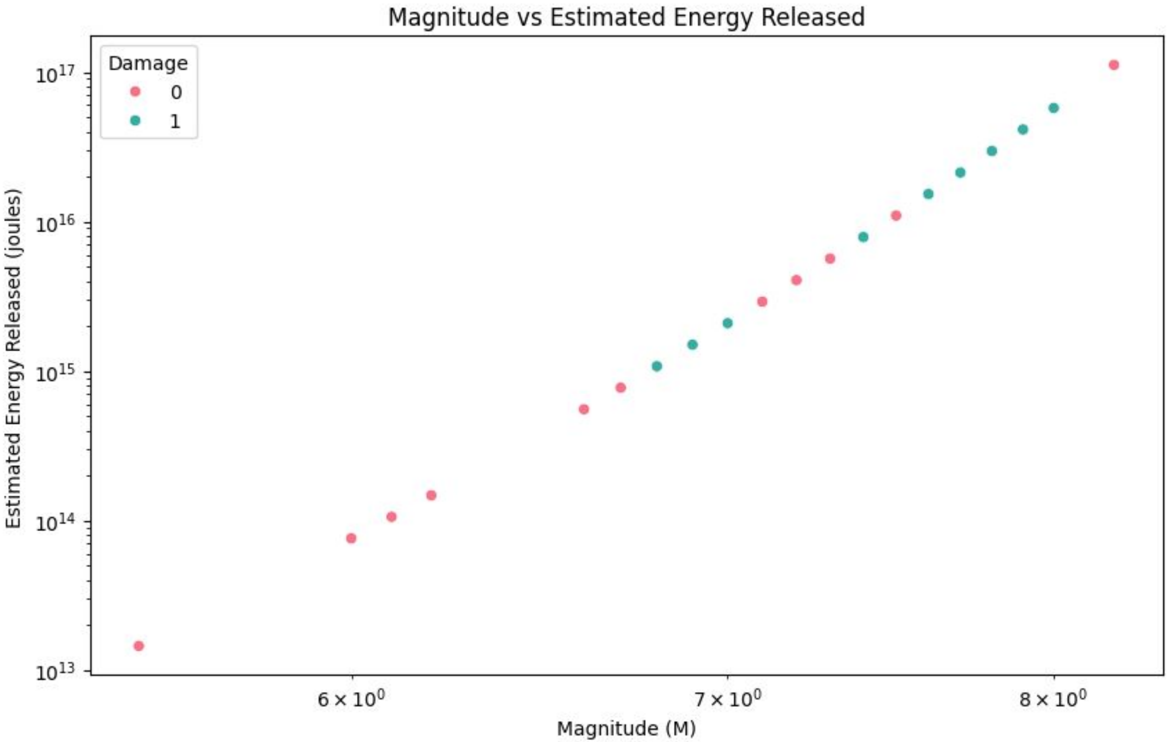
Design Choice: A Logistic Regression Model used to analyze the relationship between a binary dependent variable (in this case, whether an earthquake caused damage or not) and one or more independent variables (earthquake features).

a. Target Variable Creation: A binary target variable 'Damage' was created using a custom function damage_binary() that assigns a value of 1 if the 'Damage ($Mil)' column has a non-zero value or if the 'Damage Description' column is not null, and 0 otherwise.

b. Feature Selection: The relevant features ['Mag', 'Focal Depth (km)', 'Latitude', 'Longitude', 'Tsu'] were selected as independent variables.



Energy Release and Damage



Scientists can use seismogram data to estimate the energy released by an earthquake, including its focal depth. The energy released by an earthquake can be roughly estimated by using the equation $\log E = 5.24 + 1.44M$, where M is the magnitude. This relationship is only meant to work for earthquakes with a magnitude greater than 5.

Insights into the relationship between energy release and the likelihood of causing damage. If higher energy release is associated with a higher probability of damage, this could inform earthquake preparedness and mitigation strategies.

Conclusion

- The efficiency of a predictive model depends on which model shows the clearer predicted result and its impact.
- The use of Logistic Regression Model interprets more of the insights and intersection of multi variables which is in need of seeing a strong relationship of higher magnitude causing higher energy release and damage.
- However, there are some limitations and outliers that due to various factors, such as the depth of the event, local geological conditions, or measurement errors.

Sources

1. Khodaverdian, Leila, and Ehsan A. Safakish. "Machine Learning for Earthquake Prediction: A Review (2017–2021)." *Arabian Journal of Geosciences*, vol. 14, no. 78, 2021. SpringerLink.
2. Gono, Radomir, and Elżbieta Jasińska. "Analysis of Earthquake Forecasting in India Using Supervised Machine Learning Classifiers." *Sustainability*, vol. 13, no. 2, 2021, pp. 971. MDPI
3. Khosravikia, Farid, et al. "Earthquake Damage and Rehabilitation Intervention Prediction Using Machine Learning." *Journal of Building Engineering*, vol. 39, 2021. ScienceDirect
4. Varotsos, Panayiotis A., and Nicola Scordelis. "Towards Advancing Earthquake Forecasting by Machine Learning of Satellite Data." *Remote Sensing*, vol. 11, no. 12, 2019. MDPI
5. Liu, Zhengguo, et al. "Smart Earthquake Engineering: Machine Learning for Seismic Vulnerability Assessment." *Smart Cities*, vol. 4, 2021. Hindawi,