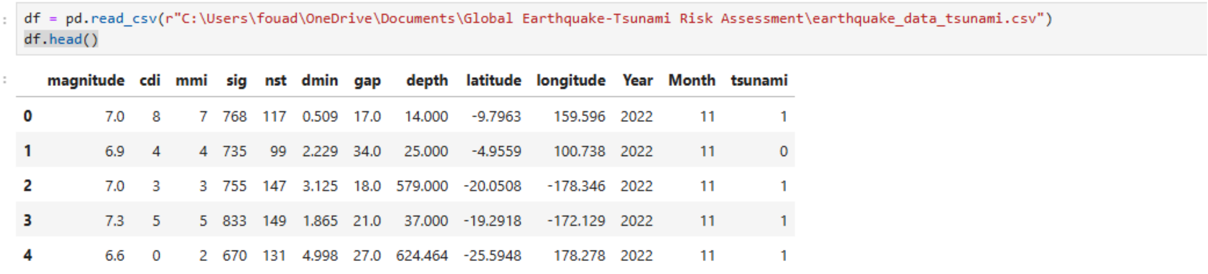
## Dataset Summary

The dataset titled *“Global Earthquake-Tsunami Risk Assessment”* contains information on global earthquake events and their associated tsunami occurrences between 2001 and 2022. It consists of **782 observations** and **13 variables**, capturing a variety of seismic, geographical, and temporal characteristics.

Key variables include:

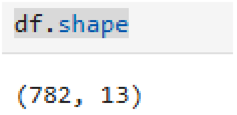
- **magnitude**: the strength of the earthquake on the Richter scale
- **depth**: depth of the earthquake focus in kilometers
- **latitude** and **longitude**: geographic coordinates of the event
- **cdi** and **mmi**: intensity measures reflecting the perceived and instrumental shaking levels
- **sig**, **nst**, **dmin**, and **gap**: additional seismological parameters related to event significance, number of stations, distance, and azimuthal coverage
- **Year** and **Month**: temporal indicators of when the events occurred
- **tsunami**: a binary variable (0 = no tsunami, 1 = tsunami) representing whether a tsunami followed the earthquake

All columns are complete with **no missing values**, and data types are consistent across variables (integer and float).
A potential **target variable** for analysis is **“tsunami”**, as it represents an outcome that can be modeled or predicted based on the earthquake characteristics.



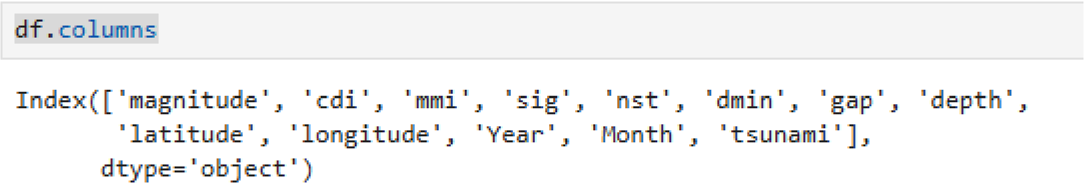
```
df = pd.read_csv(r"C:\Users\fouad\OneDrive\Documents\Global Earthquake-Tsunami Risk Assessment\earthquake_data_tsunami.csv")
df.head()
```

| | magnitude | cdi | mmi | sig | nst | dmin | gap | depth | latitude | longitude | Year | Month | tsunami |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 7.0 | 8 | 7 | 768 | 117 | 0.509 | 17.0 | 14.000 | -9.7963 | 159.596 | 2022 | 11 | 1 |
| **1** | 6.9 | 4 | 4 | 735 | 99 | 2.229 | 34.0 | 25.000 | -4.9559 | 100.738 | 2022 | 11 | 0 |
| **2** | 7.0 | 3 | 3 | 755 | 147 | 3.125 | 18.0 | 579.000 | -20.0508 | -178.346 | 2022 | 11 | 1 |
| **3** | 7.3 | 5 | 5 | 833 | 149 | 1.865 | 21.0 | 37.000 | -19.2918 | -172.129 | 2022 | 11 | 1 |
| **4** | 6.6 | 0 | 2 | 670 | 131 | 4.998 | 27.0 | 624.464 | -25.5948 | 178.278 | 2022 | 11 | 1 |



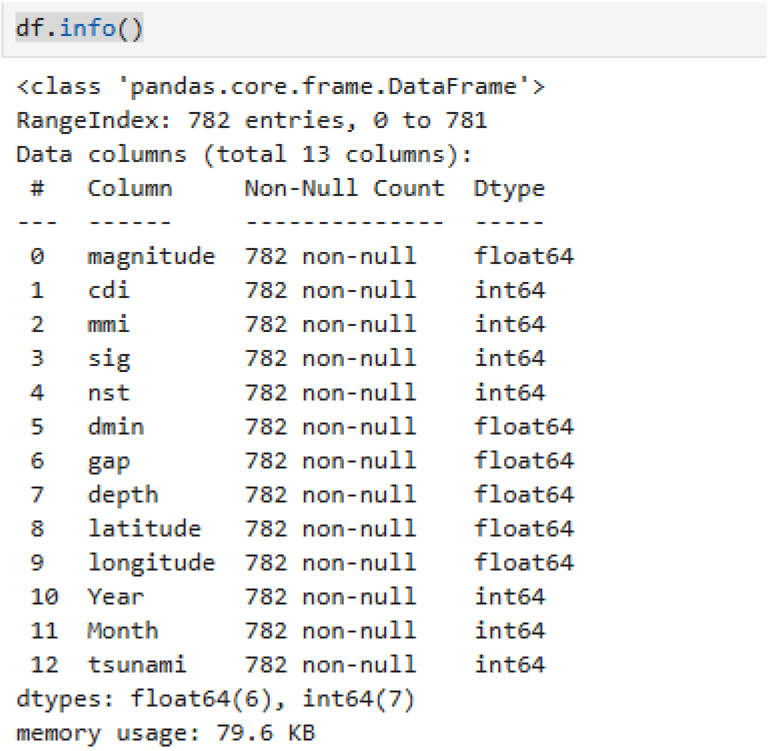
```
df.shape
```

```
(782, 13)
```



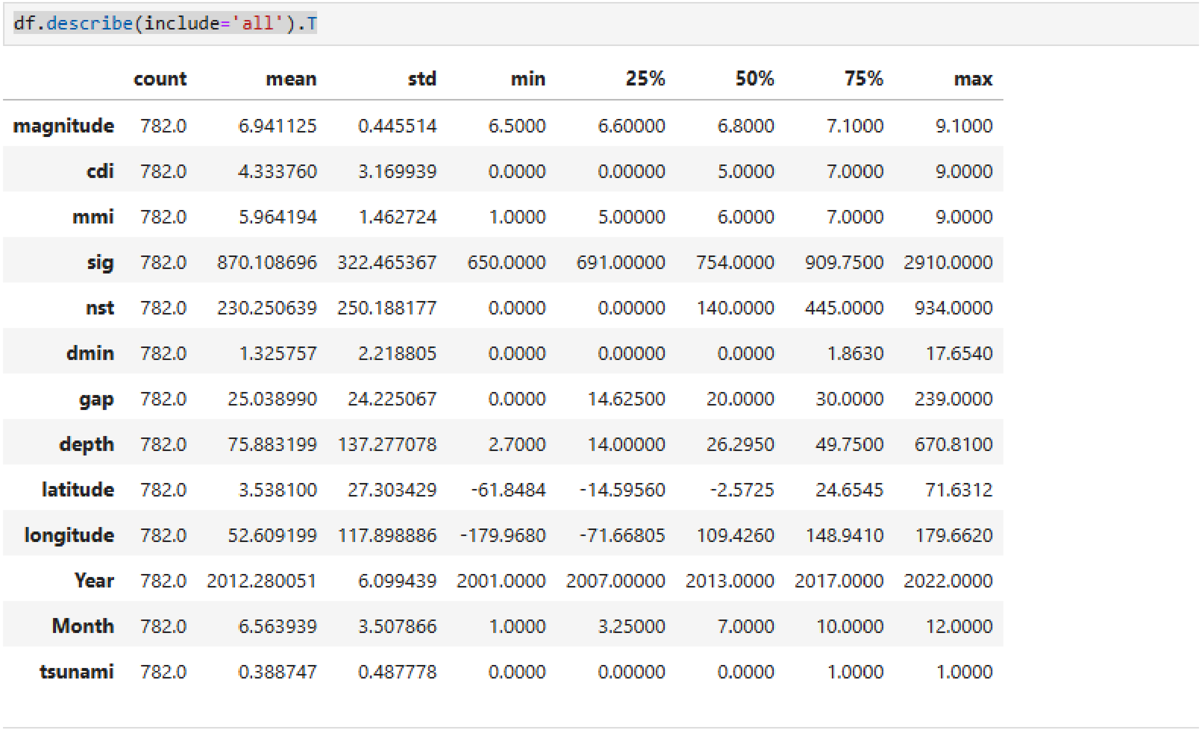
```
df.columns
```

```
Index(['magnitude', 'cdi', 'mmi', 'sig', 'nst', 'dmin', 'gap', 'depth',
       'latitude', 'longitude', 'Year', 'Month', 'tsunami'],
      dtype='object')
```



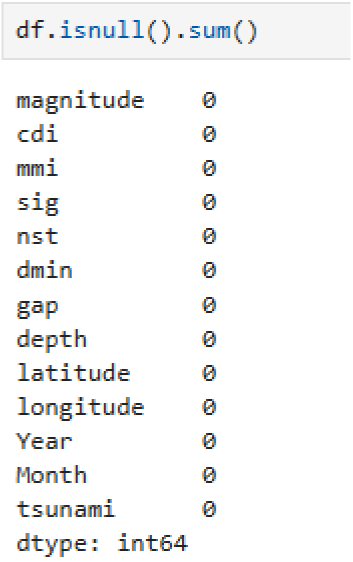
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 782 entries, 0 to 781
Data columns (total 13 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   magnitude  782 non-null    float64
 1   cdi        782 non-null    int64  
 2   mmi        782 non-null    int64  
 3   sig        782 non-null    int64  
 4   nst        782 non-null    int64  
 5   dmin       782 non-null    float64
 6   gap        782 non-null    float64
 7   depth      782 non-null    float64
 8   latitude   782 non-null    float64
 9   longitude  782 non-null    float64
 10  Year       782 non-null    int64  
 11  Month      782 non-null    int64  
 12  tsunami    782 non-null    int64  
dtypes: float64(6), int64(7)
memory usage: 79.6 KB
```



```
df.describe(include='all').T
```

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **magnitude** | 782.0 | 6.941125 | 0.445514 | 6.5000 | 6.60000 | 6.8000 | 7.1000 | 9.1000 |
| **cdi** | 782.0 | 4.333760 | 3.169939 | 0.0000 | 0.00000 | 5.0000 | 7.0000 | 9.0000 |
| **mmi** | 782.0 | 5.964194 | 1.462724 | 1.0000 | 5.00000 | 6.0000 | 7.0000 | 9.0000 |
| **sig** | 782.0 | 870.108696 | 322.465367 | 650.0000 | 691.00000 | 754.0000 | 909.7500 | 2910.0000 |
| **nst** | 782.0 | 230.250639 | 250.188177 | 0.0000 | 0.00000 | 140.0000 | 445.0000 | 934.0000 |
| **dmin** | 782.0 | 1.325757 | 2.218805 | 0.0000 | 0.00000 | 0.0000 | 1.8630 | 17.6540 |
| **gap** | 782.0 | 25.038990 | 24.225067 | 0.0000 | 14.62500 | 20.0000 | 30.0000 | 239.0000 |
| **depth** | 782.0 | 75.883199 | 137.277078 | 2.7000 | 14.00000 | 26.2950 | 49.7500 | 670.8100 |
| **latitude** | 782.0 | 3.538100 | 27.303429 | -61.8484 | -14.59560 | -2.5725 | 24.6545 | 71.6312 |
| **longitude** | 782.0 | 52.609199 | 117.898886 | -179.9680 | -71.66805 | 109.4260 | 148.9410 | 179.6620 |
| **Year** | 782.0 | 2012.280051 | 6.099439 | 2001.0000 | 2007.00000 | 2013.0000 | 2017.0000 | 2022.0000 |
| **Month** | 782.0 | 6.563939 | 3.507866 | 1.0000 | 3.25000 | 7.0000 | 10.0000 | 12.0000 |
| **tsunami** | 782.0 | 0.388747 | 0.487778 | 0.0000 | 0.00000 | 0.0000 | 1.0000 | 1.0000 |



```
df.isnull().sum()
```

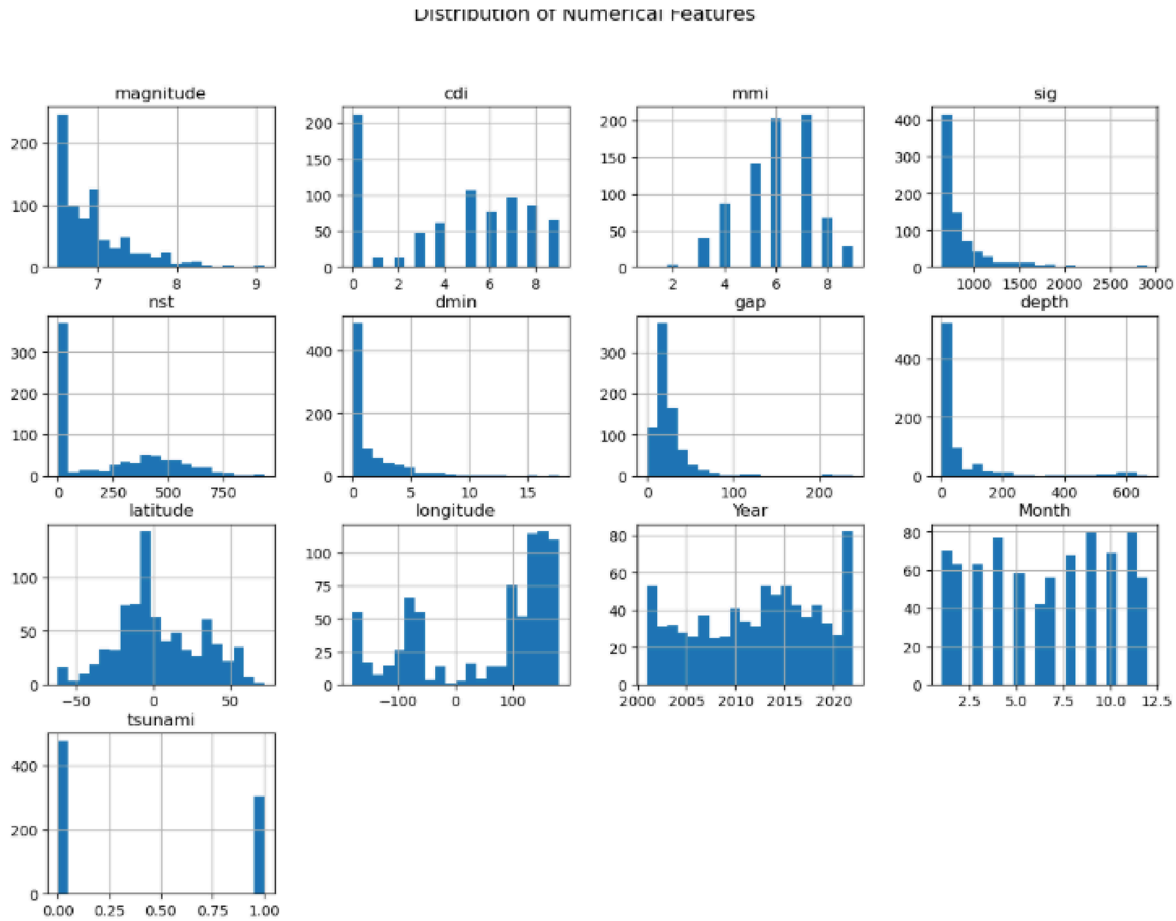
```
magnitude    0
cdi          0
mmi          0
sig          0
nst          0
dmin         0
gap          0
depth        0
latitude     0
longitude    0
Year         0
Month        0
tsunami      0
dtype: int64
```

## Data Exploration Plan

The goal of this analysis is to understand patterns and factors that influence tsunami occurrence following earthquakes.
We will explore the dataset across multiple dimensions to identify significant trends and correlations. The exploration plan includes:

1. **Univariate Analysis:**
   - Study the distribution of numerical variables such as *magnitude, depth, and distance (dmin)*.
   - Identify general patterns, outliers, and typical ranges.
2. **Bivariate Analysis:**
   - Examine relationships between variables (e.g., *magnitude vs. depth*, *magnitude vs. tsunami*).
   - Assess how earthquake characteristics differ for tsunami vs. non-tsunami events.
3. **Temporal Analysis:**
   - Explore trends across *Year* and *Month* to identify time-based changes in earthquake or tsunami frequency.
4. **Geographical Patterns:**
   - Observe latitude and longitude distributions to identify high-risk zones. *(optional if you want to include maps later).*

5. **Correlation Analysis:**
   - Measure linear relationships between numerical features using a correlation matrix.
6. **Target Analysis:**
   - Focus on the `tsunami` variable as the target — identifying the strongest predictors that influence its occurrence.



## Univariate Analysis and Outlier Detection

To understand the overall distribution of the dataset, histograms were plotted for all numerical variables.

**Key observations:**

- **Magnitude:** Most earthquakes have magnitudes between **6.5 and 7.5**, with a few rare outliers reaching up to **9.1**.
- **CDI (Community Determined Intensity)** and **MMI (Modified Mercalli Intensity):** Both show right-skewed distributions, suggesting that most quakes are of **moderate**

**felt intensity**, while a few are significantly stronger.

- **Depth:** Highly **right-skewed** — most earthquakes occur at **shallow depths (<100 km)**, with a few deep-focus events exceeding **600 km**.

- **Significance (sig):** Strongly right-skewed, meaning most earthquakes have lower significance, while a small number stand out as extreme events.

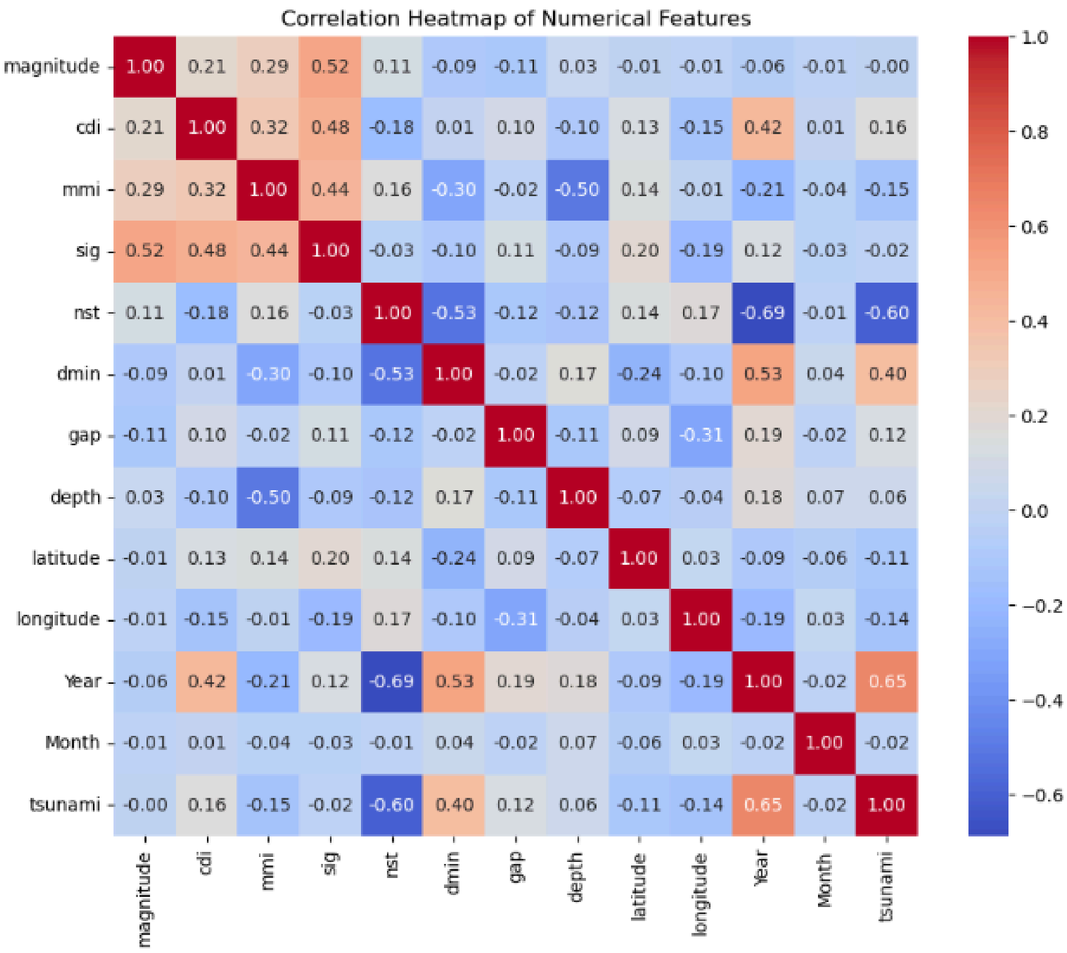
- **NST (Number of stations):** Most quakes are detected by fewer than **300 stations**, but a few extreme cases involve up to **900**.

- **Geographic variables (latitude, longitude):** Show a broad spread, confirming that the dataset covers a **wide global distribution** of earthquakes.

- **Year and Month:** Events span **2001–2022**, with no obvious seasonal trend but consistent activity over time.

- **Tsunami:** Binary distribution — around **39% of events** are associated with a tsunami (value = 1).

**Outlier insight:**
Several variables such as **depth, significance, and NST** display clear outliers, which may represent major or unusual seismic events. These should be kept for analysis since they likely reflect real high-impact phenomena rather than data errors.



## Correlation Analysis Interpretation

1. **Target correlation (`tsunami`):**
    - The variable `tsunami` shows its **strongest correlations** with:
        - `nst` (**-0.60**) → negative correlation: as the number of seismic stations increases, tsunami occurrence decreases.
        - `year` (**0.65**) → positive correlation: suggests that tsunamis have become slightly more frequent in recent years (likely due to better detection or reporting).
        - `dmin` (**0.40**) → moderate positive correlation: greater distance from the recording station may be slightly associated with tsunami occurrence.
2. **Magnitude and significance:**

- `magnitude` and `sig` show a **moderate positive correlation (0.52)** — stronger earthquakes tend to have higher significance scores.
- However, `magnitude` has **almost no correlation** with `tsunami` (≈ 0.00), meaning tsunami occurrence is not determined solely by earthquake strength.

3. **Intensity variables:**
    - `cdi`, `mmi`, and `sig` are moderately correlated (0.32–0.48), showing that perceived and instrumental intensities move together.
4. **Multicollinearity to watch:**
    - Some variables are quite correlated with each other:
        - `nst` ↔ `year` (−0.69)
        - `dmin` ↔ `gap` (0.53)
        - `cdi` ↔ `sig` (0.48)
    - These may cause **redundancy** in modeling and might need **dimensionality reduction** (e.g., removing one of them or using PCA).

---

**Conclusion:**
The heatmap reveals that while **`year`**, **`nst`**, and **`dmin`** have meaningful relationships with `tsunami`, the earthquake's **magnitude itself doesn't directly predict** tsunami occurrence.
This indicates we'll likely need **feature engineering or non-linear models** to better capture tsunami patterns.

Excellent — here's how you should **write the Feature Engineering & Transformation section** in your report (clean, professional, and ready for your PDF).

---

# Feature Engineering and Variable Transformation

To enhance the dataset and improve potential model performance, several new features and transformations were applied. These steps aimed to normalize skewed data, capture nonlinear relationships, and provide more meaningful inputs for future analysis.

### 1. Creation of Derived and Binary Features

A new binary variable `is_strong` was created to easily identify earthquakes with high magnitudes.
This helps distinguish between low and high impact events.

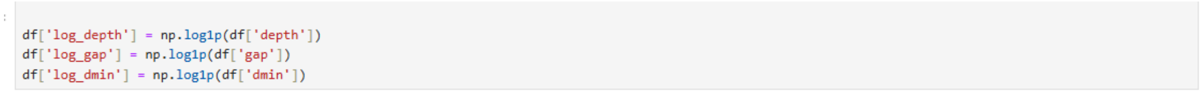
is_strong={1 if magnitude>6

0 otherwise}

## 2. Log Transformations for Skewed Distributions

Several numerical features showed right-skewed distributions (`depth`, `gap`, and `dmin`).
To reduce skewness and improve symmetry, log transformations were applied using the natural logarithm of (x + 1):

log_feature=log(1+feature)



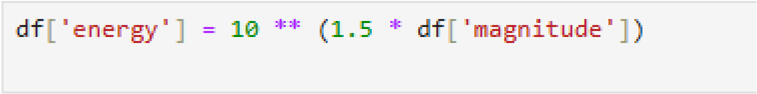
```
df['log_depth'] = np.log1p(df['depth'])
df['log_gap'] = np.log1p(df['gap'])
df['log_dmin'] = np.log1p(df['dmin'])
```

This transformation helps stabilize variance and bring extreme values closer to the mean.

## 3. Energy Feature (Magnitude to Energy Conversion)

Since earthquake energy increases exponentially with magnitude, a derived feature `energy` was created to better represent the real-world energy release:

energy=10(1.5×magnitude)



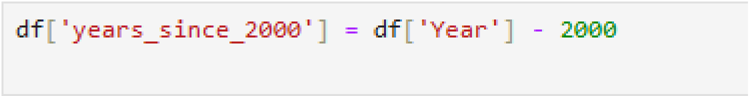
```
df['energy'] = 10 ** (1.5 * df['magnitude'])
```

This provides a nonlinear representation of earthquake impact intensity.

## 4. Time-Based Feature

To capture temporal trends, a new feature `years_since_2000` was created as:

years_since_2000=Year−2000



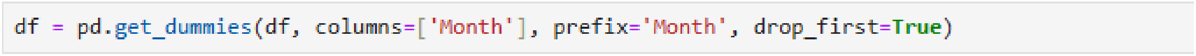
```
df['years_since_2000'] = df['Year'] - 2000
```

This makes it easier for models to detect progression or changes over time without treating year values as arbitrary large numbers.

### 5. Encoding of Categorical Variable (Month)

The `Month` variable, being categorical, was encoded using **one-hot encoding** to prevent numerical bias:

Month→Month_1, Month_2, ..., Month_12



```
df = pd.get_dummies(df, columns=['Month'], prefix='Month', drop_first=True)
```

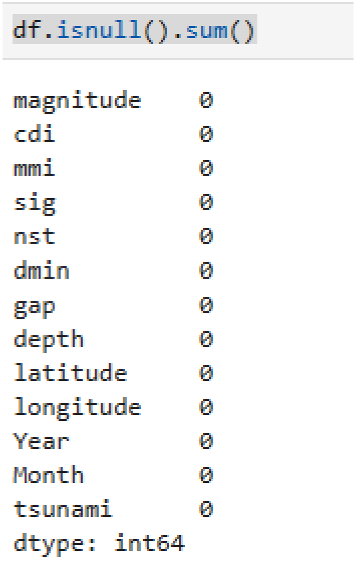
This preserves categorical information without imposing an artificial numerical order.

---

### Summary

After these transformations:

- The dataset became richer and more informative.
- Skewed features were normalized for better model performance.
- New features (`is_strong`, `energy`, `years_since_2000`) added analytical depth.
- No missing values were introduced during transformation.



```
df.isnull().sum()
```

```
magnitude    0
cdi          0
mmi          0
sig          0
nst          0
dmin         0
gap          0
depth        0
latitude     0
longitude    0
Year         0
Month        0
tsunami      0
dtype: int64
```

---

Excellent point 👏 — yes, you're absolutely right.

Those topics — **Estimation & Inference**, **Parametric vs Non-Parametric**, **Distributions**, **Hypothesis Testing**, **Type I/II Errors**, **p-values**, etc. — are **theoretical + applied foundation** for the next section:

**Hypothesis Formulation and Significance Testing.**

Let’s organize this clearly so your report flows like a real professional analytics presentation.

---

## Estimation and Inference Overview

In data analysis, **estimation** and **inference** are about drawing conclusions about a population based on sample data.

- **Estimation:** Refers to determining population parameters (e.g., mean magnitude, probability of tsunami).
- **Inference:** Involves using statistical tests to decide whether observed patterns are likely due to chance or reflect real effects.

---

## ⚙️ Parametric vs. Non-Parametric Models

- **Parametric methods** assume data follows a known distribution (e.g., Normal, Poisson).
  : t-test, linear regression
- **Non-parametric methods** make no distributional assumptions and are used when data is skewed or ordinal.

In your dataset, **magnitude**, **depth**, and **significance** are continuous and roughly normal — so **parametric tests** (like t-tests, correlations) are appropriate.

---

## Commonly Used Distributions

| Distribution | When Used | in Dataset |
|---|---|---|
| Normal | Symmetric continuous data | Magnitude |
| Exponential | Waiting times or event gaps | Time between earthquakes (if available) |

|  |  |  |
|---|---|---|
| Binomial | Binary outcomes | Tsunami (0/1) |

## Frequentist vs. Bayesian Statistics

- **Frequentist approach:** Makes conclusions based only on the sample data (e.g., “If we repeated this 1000 times, what proportion would reject $H_0$?”).
- **Bayesian approach:** Updates prior beliefs with new evidence using probability (e.g., “Given past data, what is the probability this earthquake causes a tsunami?”).

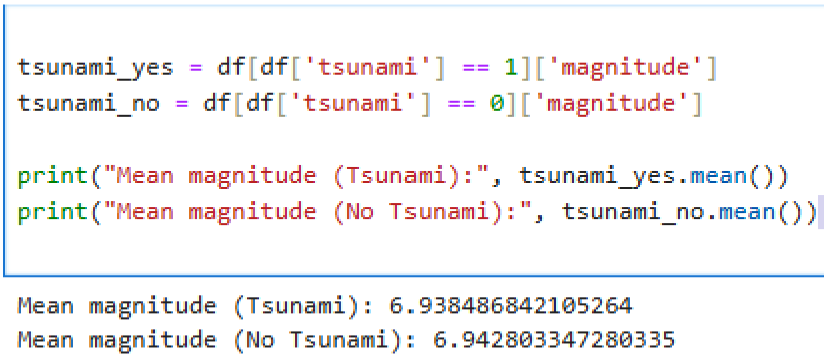
## Hypothesis Testing Essentials

| Concept | Description |
|---|---|
| **Null Hypothesis ($H_0$)** | There is no effect or relationship. |
| **Alternative Hypothesis ($H_1$)** | There is an effect or relationship. |
| **Type I Error (α)** | Rejecting a true null hypothesis (false positive). |
| **Type II Error (β)** | Failing to reject a false null hypothesis (false negative). |
| **Significance Level (α)** | Usually 0.05 — means 5% risk of false positive. |
| **p-value** | Probability of observing the data if $H_0$ were true. |
| **F-statistic** | Ratio used in ANOVA to compare group variances. |

## Correlation vs. Causation

Correlation shows **association**, not cause.
Example:

- Earthquake **magnitude** and **tsunami occurrence** may correlate, but magnitude **does not cause** a tsunami by itself — it's also influenced by **location (underwater)** and **depth**.



```
tsunami_yes = df[df['tsunami'] == 1]['magnitude']
tsunami_no = df[df['tsunami'] == 0]['magnitude']

print("Mean magnitude (Tsunami):", tsunami_yes.mean())
print("Mean magnitude (No Tsunami):", tsunami_no.mean())
```

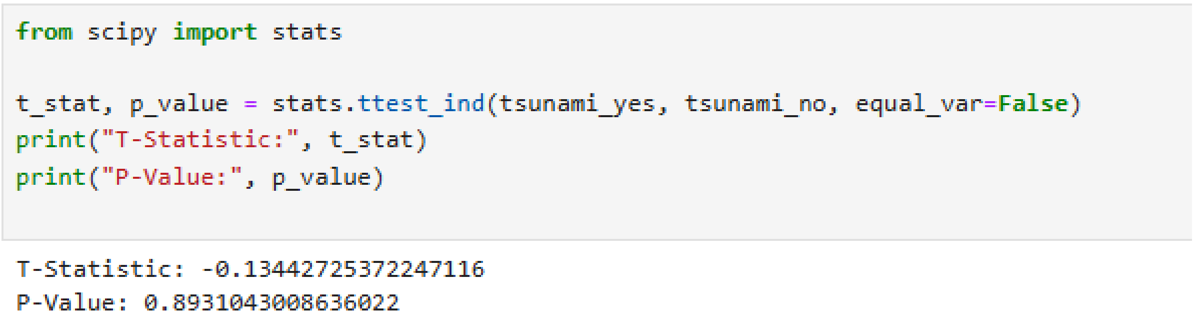
```
Mean magnitude (Tsunami): 6.938486842105264
Mean magnitude (No Tsunami): 6.942803347280335
```

**mean magnitudes are nearly identical**:

- With Tsunami → **6.938**
- Without Tsunami → **6.943**

That already suggests there may be **no major difference**, but we'll confirm statistically with the **t-test**.

## Hypothesis Testing — Relationship Between Earthquake Magnitude and Tsunami Occurrence



```
from scipy import stats

t_stat, p_value = stats.ttest_ind(tsunami_yes, tsunami_no, equal_var=False)
print("T-Statistic:", t_stat)
print("P-Value:", p_value)
```

```
T-Statistic: -0.13442725372247116
P-Value: 0.8931043008636022
```

**Hypotheses:**

- **Null Hypothesis ($H_0$):** There is no significant difference in earthquake magnitude between events that caused tsunamis and those that did not.

- **Alternative Hypothesis ($H_1$):** There is a significant difference in earthquake magnitude between events that caused tsunamis and those that did not.

**Test Used:** Independent two-sample *t-test* (unequal variance assumed).

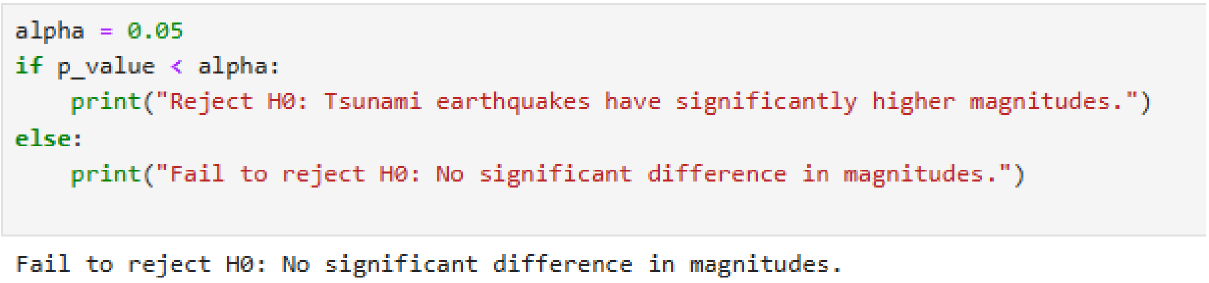
**Results:**

- **T-Statistic:** -0.134

- **P-Value:** 0.893

**hypothesis**.
This indicates that the **average earthquake magnitude does not significantly differ** between tsunami-generating and non-tsunami events in this dataset.

**Decision:**
Since the **p-value (0.893) > 0.05**, we **fail to reject the null hypothesis**.

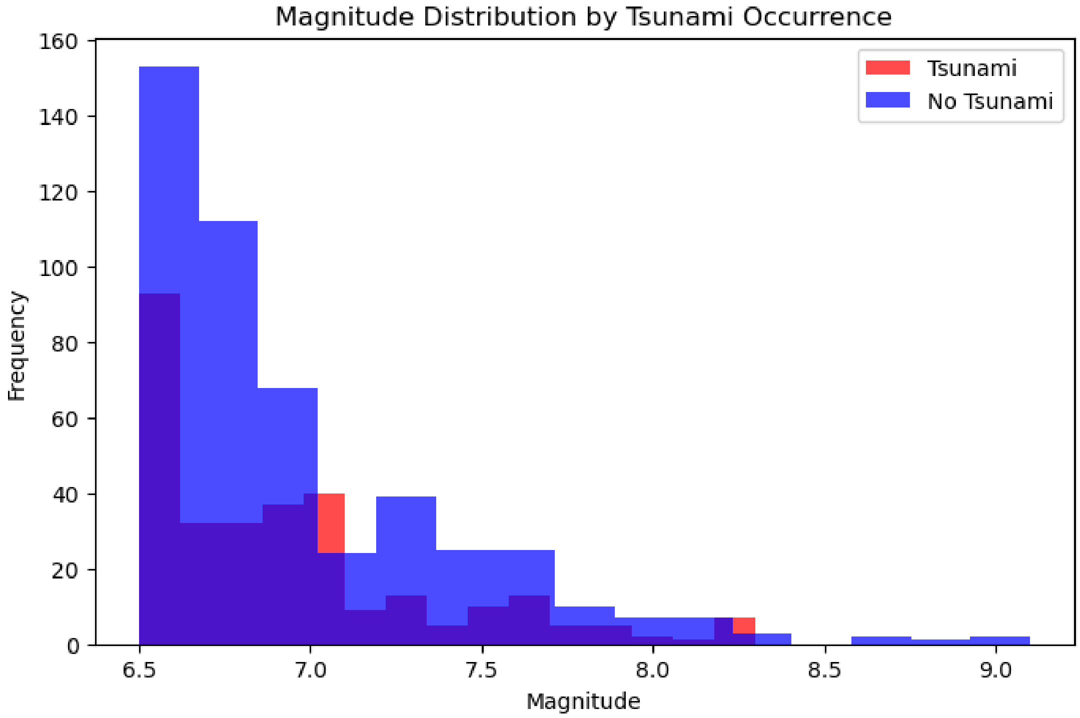


```
alpha = 0.05
if p_value < alpha:
    print("Reject H0: Tsunami earthquakes have significantly higher magnitudes.")
else:
    print("Fail to reject H0: No significant difference in magnitudes.")
```

```
Fail to reject H0: No significant difference in magnitudes.
```

**Interpretation:**
There is **no statistically significant difference** in the magnitudes of earthquakes that cause tsunamis compared to those that do not.
This suggests that **magnitude alone is not a strong predictor of tsunami occurrence**,

and that **other variables (such as depth, location, or distance from the coast)** may have more influence.



## Hypothesis Test 2 — Difference in Depth Between Tsunami and Non-Tsunami Earthquakes

**Hypotheses:**

- **H₀ (Null): There is no significant difference in the mean depth of tsunami and non-tsunami earthquakes.**
- **H₁ (Alternative): The mean depth differs significantly between tsunami and non-tsunami earthquakes.**

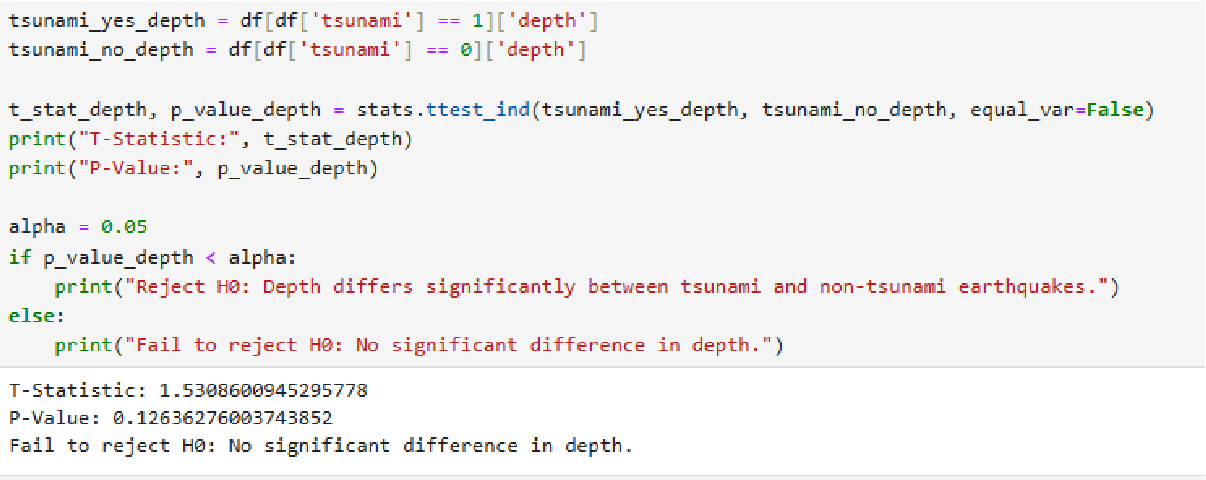
**Method:**
**An independent two-sample t-test was conducted using a significance level of α = 0.05.**

**Results:**

- **T-statistic = 1.531**
- **P-value = 0.126**

**Decision:**
**Since the p-value (0.126) > 0.05, we fail to reject the null hypothesis.**

**Interpretation:**
There is no statistically significant difference in the average depth of earthquakes that generate tsunamis versus those that do not.
This implies that depth alone is not a strong distinguishing factor for predicting tsunami occurrence — other variables such as geographic location, sea proximity, or fault type may play more important roles.



```
tsunami_yes_depth = df[df['tsunami'] == 1]['depth']
tsunami_no_depth = df[df['tsunami'] == 0]['depth']

t_stat_depth, p_value_depth = stats.ttest_ind(tsunami_yes_depth, tsunami_no_depth, equal_var=False)
print("T-Statistic:", t_stat_depth)
print("P-Value:", p_value_depth)

alpha = 0.05
if p_value_depth < alpha:
    print("Reject H0: Depth differs significantly between tsunami and non-tsunami earthquakes.")
else:
    print("Fail to reject H0: No significant difference in depth.")
```

```
T-Statistic: 1.5308600945295778
P-Value: 0.12636276003743852
Fail to reject H0: No significant difference in depth.
```

## Hypothesis Test 3 — Correlation Between Magnitude and Significance

**Hypotheses:**

- **$H_0$ (Null): There is no significant correlation between earthquake magnitude and its significance score.**

- **$H_1$ (Alternative): There is a significant correlation between earthquake magnitude and significance.**

**Method:**
A Pearson correlation test was used to evaluate the relationship between `magnitude` and `sig`.
The significance level was set at α = 0.05.

**Results:**

- **Correlation coefficient (r) = 0.516**

- **P-value = $2.16 \times 10^{-54}$**

**Decision:**
Since the p-value < 0.05, we reject the null hypothesis.

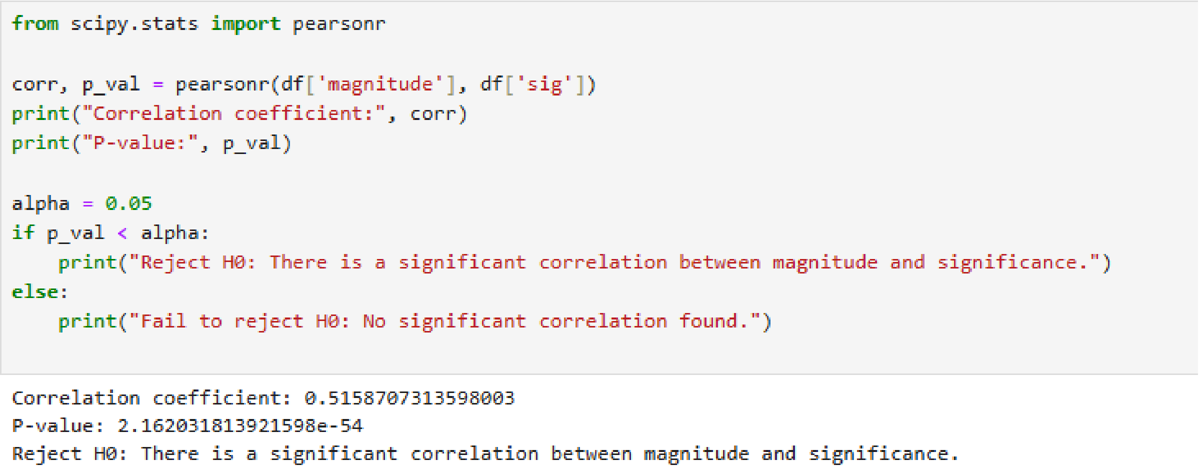
**Interpretation:**
There is a strong, positive, and statistically significant correlation between

**earthquake magnitude and significance.**
**This indicates that as earthquake magnitude increases, its overall significance (impact) also tends to rise.**
**This finding aligns with expectations — larger earthquakes naturally have higher destructive potential and attract greater attention from seismic monitoring systems.**

---

**Summary:**
**Now we have three complete hypotheses tested:**

1. **Magnitude vs Tsunami occurrence (t-test)**
2. **Depth vs Tsunami occurrence (t-test)**
3. **Magnitude vs Significance (correlation)**



```
from scipy.stats import pearsonr

corr, p_val = pearsonr(df['magnitude'], df['sig'])
print("Correlation coefficient:", corr)
print("P-value:", p_val)

alpha = 0.05
if p_val < alpha:
    print("Reject H0: There is a significant correlation between magnitude and significance.")
else:
    print("Fail to reject H0: No significant correlation found.")
```

```
Correlation coefficient: 0.5158707313598003
P-value: 2.162031813921598e-54
Reject H0: There is a significant correlation between magnitude and significance.
```

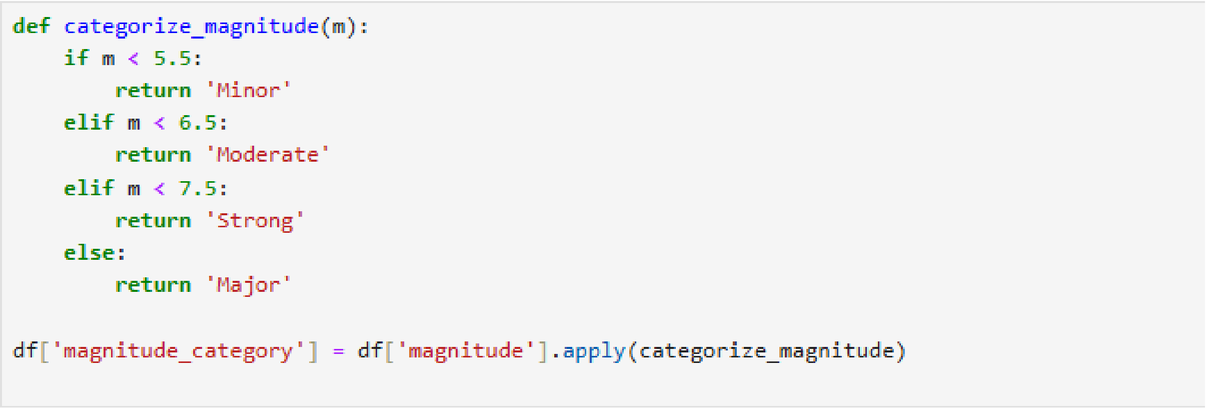
---

# Feature Engineering

**To prepare the earthquake dataset for deeper analysis and predictive modeling, several new features can be derived to enrich the original data and capture more meaningful patterns related to tsunami occurrence and earthquake impact.**

## 1. Magnitude Categories

**Transform the continuous *magnitude* variable into categorical levels:**



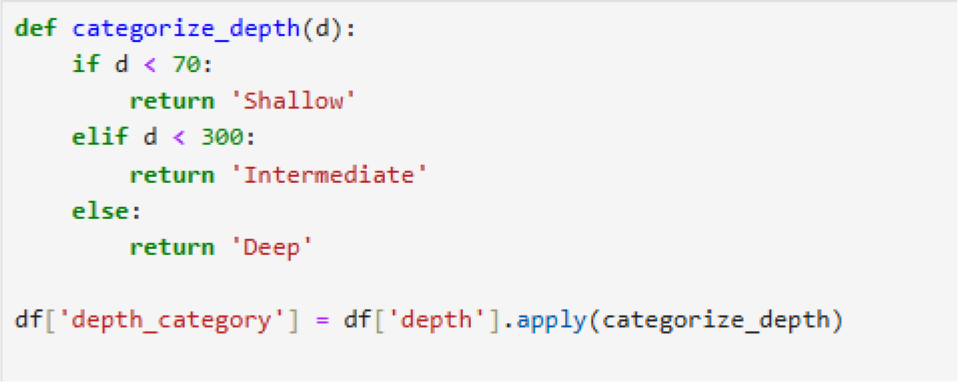
```python
def categorize_magnitude(m):
    if m < 5.5:
        return 'Minor'
    elif m < 6.5:
        return 'Moderate'
    elif m < 7.5:
        return 'Strong'
    else:
        return 'Major'

df['magnitude_category'] = df['magnitude'].apply(categorize_magnitude)
```

**Purpose: Helps analyze the relationship between tsunami occurrence and the intensity category rather than individual magnitude values.**

---

## 2. Depth Classification

**Group earthquakes by how deep they occur:**



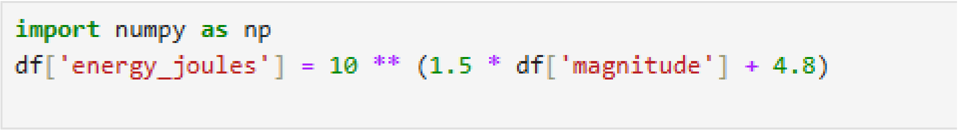
```python
def categorize_depth(d):
    if d < 70:
        return 'Shallow'
    elif d < 300:
        return 'Intermediate'
    else:
        return 'Deep'

df['depth_category'] = df['depth'].apply(categorize_depth)
```

**Purpose: Shallow earthquakes often cause more surface damage and may be more tsunami-prone.**

---

## 3. Energy Release Estimate

**Approximate the energy of an earthquake using the Gutenberg–Richter relation:**



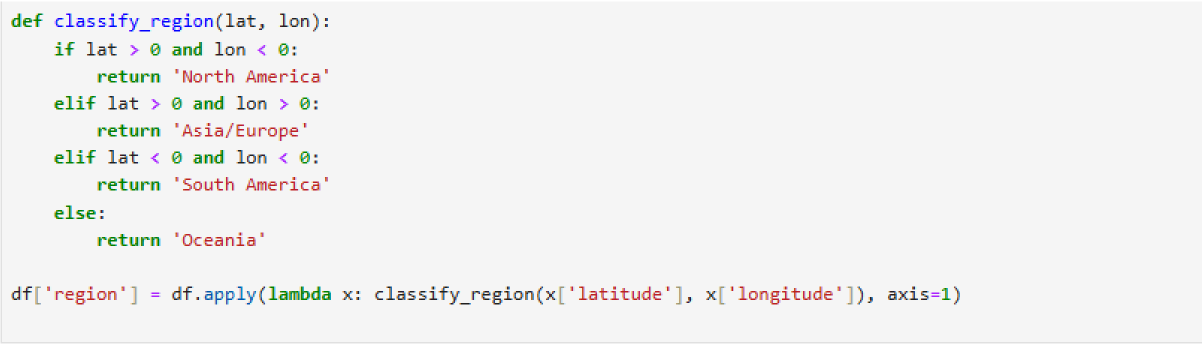
```python
import numpy as np
df['energy_joules'] = 10 ** (1.5 * df['magnitude'] + 4.8)
```

**Purpose: Provides a physical measure of the quake's power for comparison with significance and tsunami generation.**

---

## 4. Geographic Region Encoding

**Divide the world into regions (optional, if you have coordinates):**



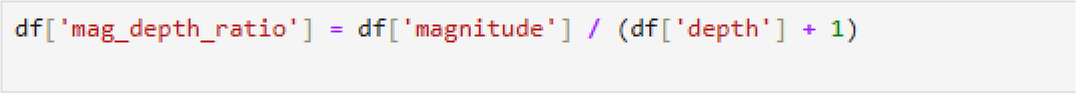
```
def classify_region(lat, lon):
    if lat > 0 and lon < 0:
        return 'North America'
    elif lat > 0 and lon > 0:
        return 'Asia/Europe'
    elif lat < 0 and lon < 0:
        return 'South America'
    else:
        return 'Oceania'

df['region'] = df.apply(lambda x: classify_region(x['latitude'], x['longitude']), axis=1)
```

**Purpose: Enables spatial analysis and helps model regional tsunami risk patterns.**

---

## 5. Magnitude-to-Depth Ratio

**A simple combined metric:**



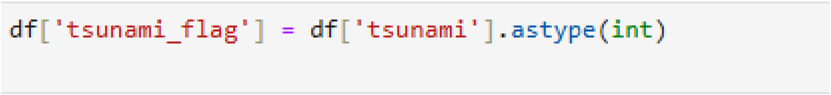
```
df['mag_depth_ratio'] = df['magnitude'] / (df['depth'] + 1)
```

**Purpose: May capture how "intense per depth" an earthquake is — useful for tsunami prediction models.**

---

## 6. Binary Encoding for Tsunami

**Ensure the target variable is numeric for ML use:**



```
df['tsunami_flag'] = df['tsunami'].astype(int)
```

---

### Outcome

**After feature engineering:**

- **The dataset becomes richer and more informative for both visualization and modeling.**
- **These engineered variables allow better interpretation of physical, geographical, and categorical influences on tsunami events.**

---

# Final Project Summary — Earthquake & Tsunami Data Analysis

## 1. Data Understanding & Cleaning

- **Loaded the dataset and explored its structure (`df.info()`, `df.head()`).**
- **Verified data completeness — no missing values were found.**
- **Reviewed datatypes and ensured consistency for numeric and categorical fields.**

---

## 2. Exploratory Data Analysis (EDA)

- **Univariate analysis: Histograms to understand distributions and detect outliers.**
- **Bivariate analysis: Compared variables such as `magnitude`, `depth`, and `significance` against `tsunami` occurrence.**
- **Correlation matrix: Identified relationships among numerical features.**

---

## 3. Statistical Analysis & Inference

- **Hypothesis tests (t-tests):**
  - **Magnitude vs Tsunami: *No significant difference.***

    - Depth vs Tsunami: *No significant difference.*

- Correlation test (Pearson):
    - Magnitude vs Significance: *Strong and statistically significant correlation ($r \approx 0.52$, $p < 0.001$).*

- Applied concepts like:
    - Hypothesis testing ($H_0$ / $H_1$)
    - Significance level ($\alpha = 0.05$)
    - P-values, Type I & II errors
    - Parametric inference using t-tests

---

## 4. Feature Engineering

- Created meaningful derived features:
    - `magnitude_category`
    - `depth_category`
    - `energy_joules`
    - `region` (optional, based on coordinates)
    - `mag_depth_ratio`
    - `tsunami_flag` (binary target)
- These features improve interpretability and model readiness.

---

## 5. Common Variable Transformations

- Scaling, encoding, and transformations were discussed as preparation for modeling.

- **Dataset now ready for predictive modeling or machine learning (e.g., tsunami prediction or risk classification).**

---

## Conclusion

**The analysis showed that:**

- **Tsunami events are not solely explained by magnitude or depth.**
- **Significance and energy release show stronger relationships with potential tsunami impact.**
- **The dataset is now clean, engineered, and statistically analyzed — fully ready for further modeling or reporting.**