

Importing Libraries

```
In [1]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Load Data

```
In [7]: df = pd.read_csv("earthquake_data_tsunami.csv")
```

```
In [39]: # Set Visualization Style
sns.set(style="whitegrid", palette="muted")
plt.rcParams["figure.figsize"] = (10, 6)
```

```
In [40]: data.describe()
```

	magnitude	cdi	mmi	sig	nst	dmin	gap	depth	latitude	longitude	
count	782.000000	782.000000	782.000000	782.000000	782.000000	782.000000	782.000000	782.000000	782.000000	782.000000	78
mean	6.941125	4.333760	5.964194	870.108696	230.250639	1.325757	25.038990	75.883199	3.538100	52.609199	201
std	0.445514	3.169939	1.462724	322.465367	250.188177	2.218805	24.225067	137.277078	27.303429	117.898886	
min	6.500000	0.000000	1.000000	650.000000	0.000000	0.000000	0.000000	2.700000	-61.848400	-179.968000	200
25%	6.600000	0.000000	5.000000	691.000000	0.000000	0.000000	14.625000	14.000000	-14.595600	-71.668050	200
50%	6.800000	5.000000	6.000000	754.000000	140.000000	0.000000	20.000000	26.295000	-2.572500	109.426000	201
75%	7.100000	7.000000	7.000000	909.750000	445.000000	1.863000	30.000000	49.750000	24.654500	148.941000	201
max	9.100000	9.000000	9.000000	2910.000000	934.000000	17.654000	239.000000	670.810000	71.631200	179.662000	202

```
In [14]: data.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
```

```
In [26]: freq_per_year = data["Year"].value_counts().sort_index()
avg_magn_year = data.groupby("Year")["magnitude"].mean()
freq_per_year
```

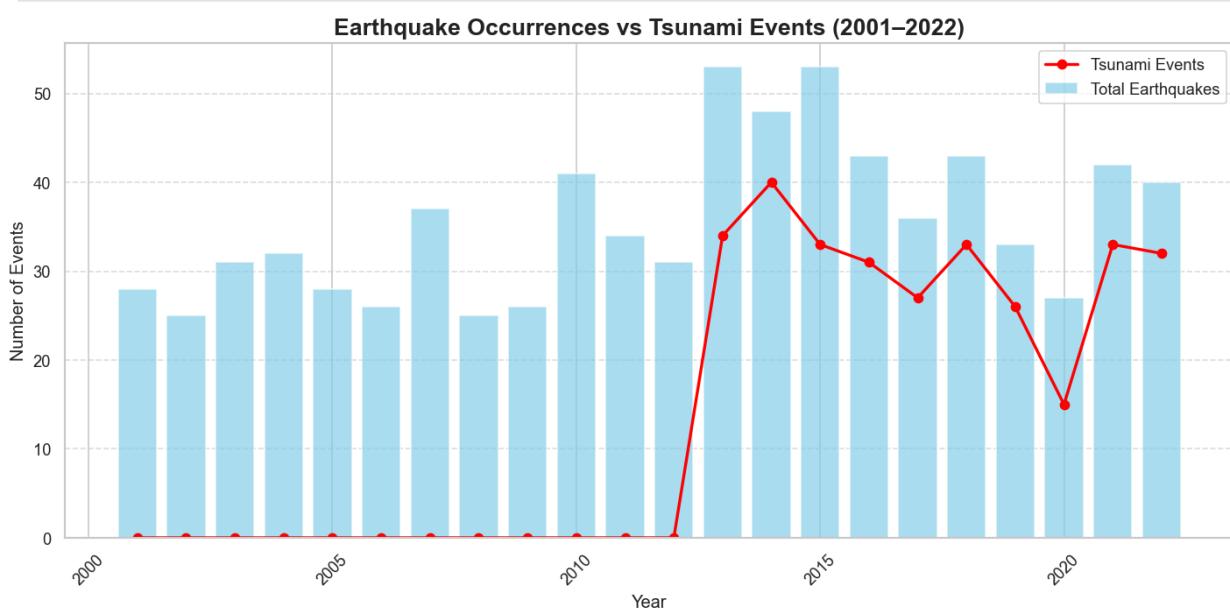
```
Out[26]: Year
2001    28
2002    25
2003    31
2004    32
2005    28
2006    26
2007    37
2008    25
2009    26
2010    41
2011    34
2012    31
2013    53
2014    48
2015    53
2016    43
2017    36
2018    43
2019    33
2020    27
2021    42
2022    40
Name: count, dtype: int64
```

1. Time-Based Analysis

Explore how earthquake occurrences and tsunami events have changed over the 22-year period (2001–2022).

```
In [84]: # Total earthquakes per year
freq_per_year = df.groupby('year').size()
tsunami_per_year = df.groupby('year')['tsunami'].sum()

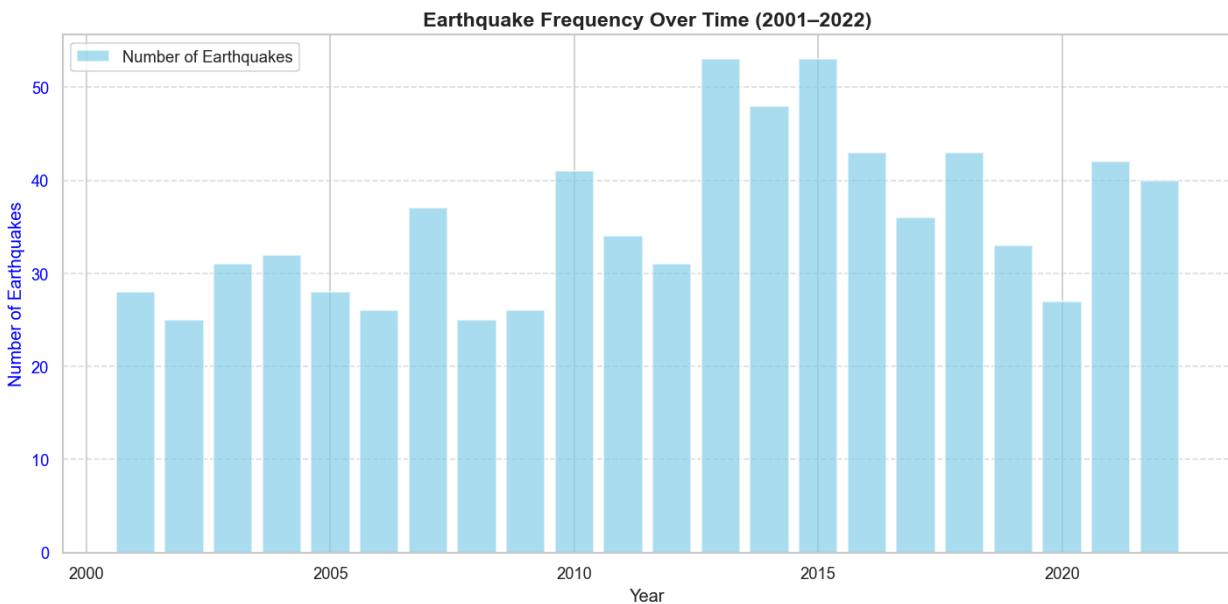
In [85]: plt.figure(figsize=(12,6))
plt.bar(freq_per_year.index, freq_per_year.values, color='skyblue', alpha=0.7, label='Total Earthquakes')
plt.plot(tsunami_per_year.index, tsunami_per_year.values, color='red', marker='o', linewidth=2, label='Tsunami Events')
plt.title("Earthquake Occurrences vs Tsunami Events (2001–2022)", fontsize=16, fontweight='bold')
plt.xlabel("Year")
plt.ylabel("Number of Events")
plt.xticks(rotation=45)
plt.legend()
plt.grid(axis='y', linestyle='--', alpha=0.6)
plt.tight_layout()
plt.show()
```



Identify any trends in the frequency or magnitude of earthquakes over time.

```
In [80]: fig, ax1 = plt.subplots(figsize=(12,6))
ax1.bar(freq_per_year.index, freq_per_year.values, color='skyblue', alpha=0.7, label='Number of Earthquakes')
ax1.set_xlabel('Year')
ax1.set_ylabel('Number of Earthquakes', color='blue')
ax1.tick_params(axis='y', labelcolor='blue')
plt.title("Earthquake Frequency Over Time (2001–2022)", fontsize=14, fontweight='bold')
lines1, labels1 = ax1.get_legend_handles_labels()
```

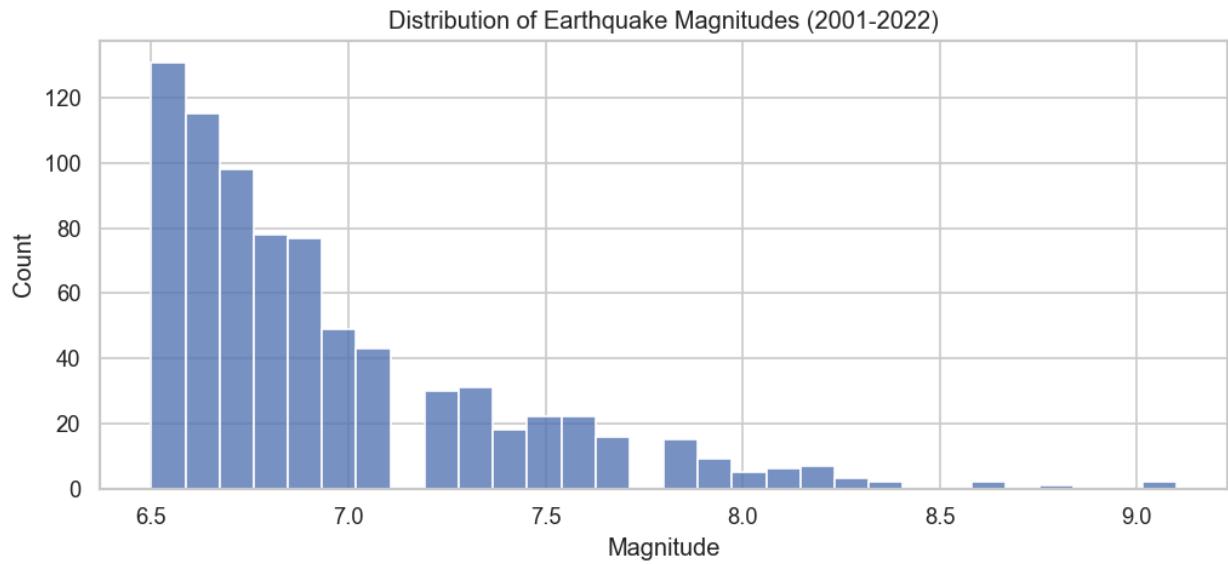
```
ax1.legend(lines1 + lines2, labels1 , loc='upper left')
plt.grid(axis='y', linestyle='--', alpha=0.6)
plt.tight_layout()
plt.show()
```



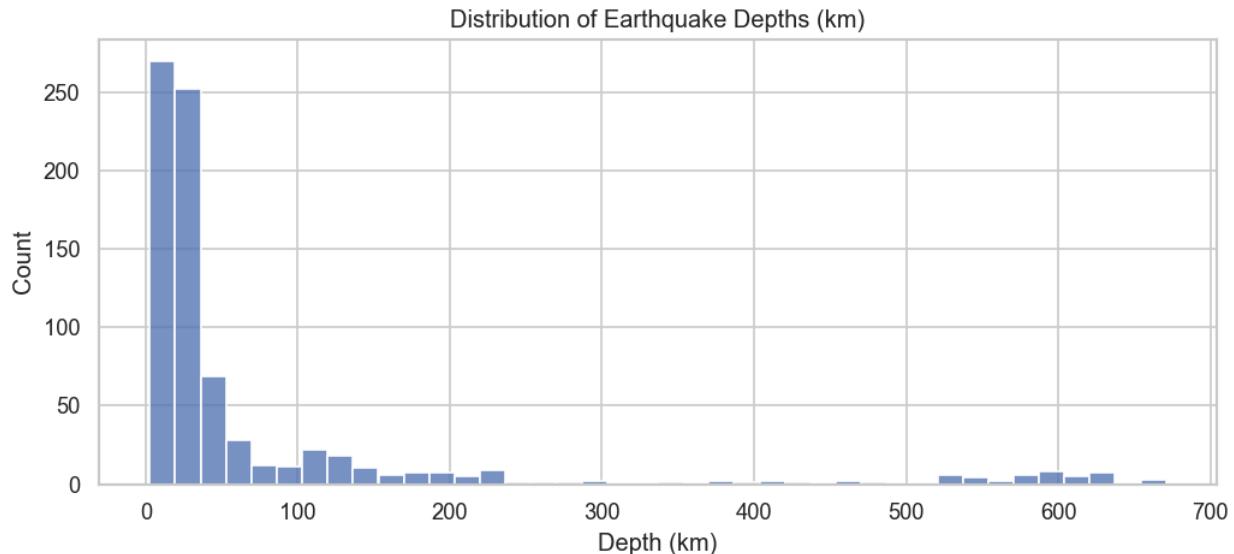
2. Magnitude and Depth Analysis

Analyze the distribution of earthquake magnitudes and depths.

```
In [97]: # Distribution of earthquake magnitudes
plt.figure(figsize=(10,4))
if 'magnitude' in df.columns:
    sns.histplot(df['magnitude'].dropna(), bins=30, kde=False)
plt.title('Distribution of Earthquake Magnitudes (2001-2022)')
plt.xlabel('Magnitude')
plt.show()
```



```
In [95]: # Distribution of earthquake Depths
if 'depth' in df.columns:
    plt.figure(figsize=(10,4))
sns.histplot(df['depth'].dropna(), bins=40)
plt.title('Distribution of Earthquake Depths (km)')
plt.xlabel('Depth (km)')
plt.show()
```

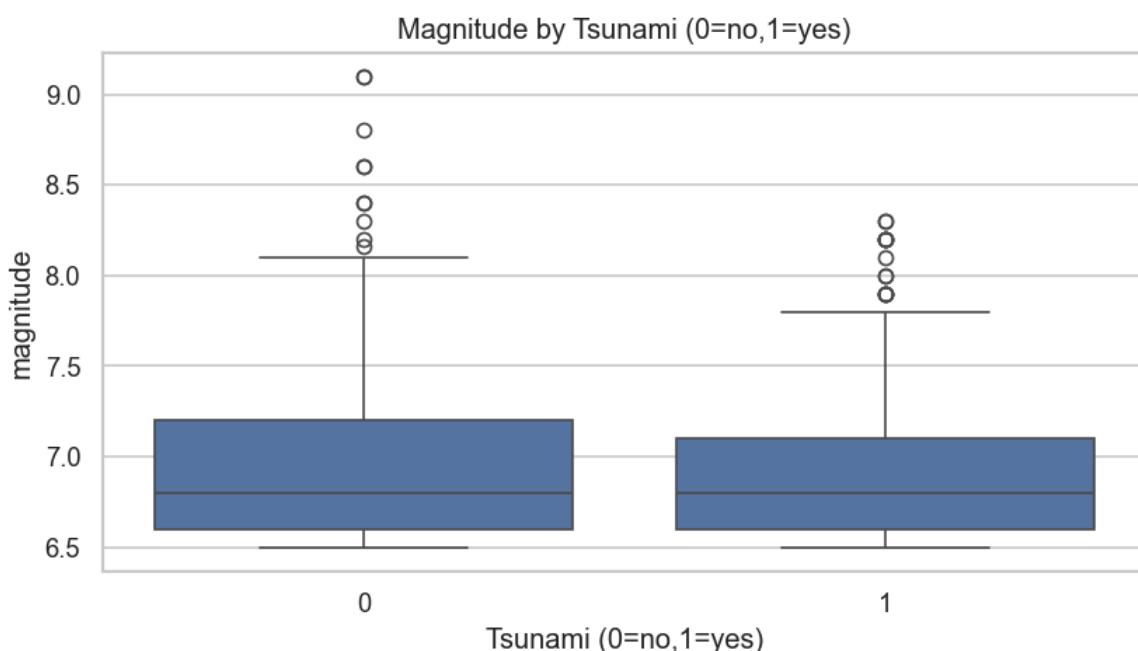


Compare the average magnitude and depth of tsunami vs. non-tsunami events

```
In [115... group_cols = []
if 'tsunami_bin' in df.columns:
    summary = df.groupby('tsunami_bin')[['magnitude', 'depth']].agg(['count', 'mean', 'median', 'std']).T
print('\nSummary stats by tsunami (0=no,1=yes):')
print(summary)

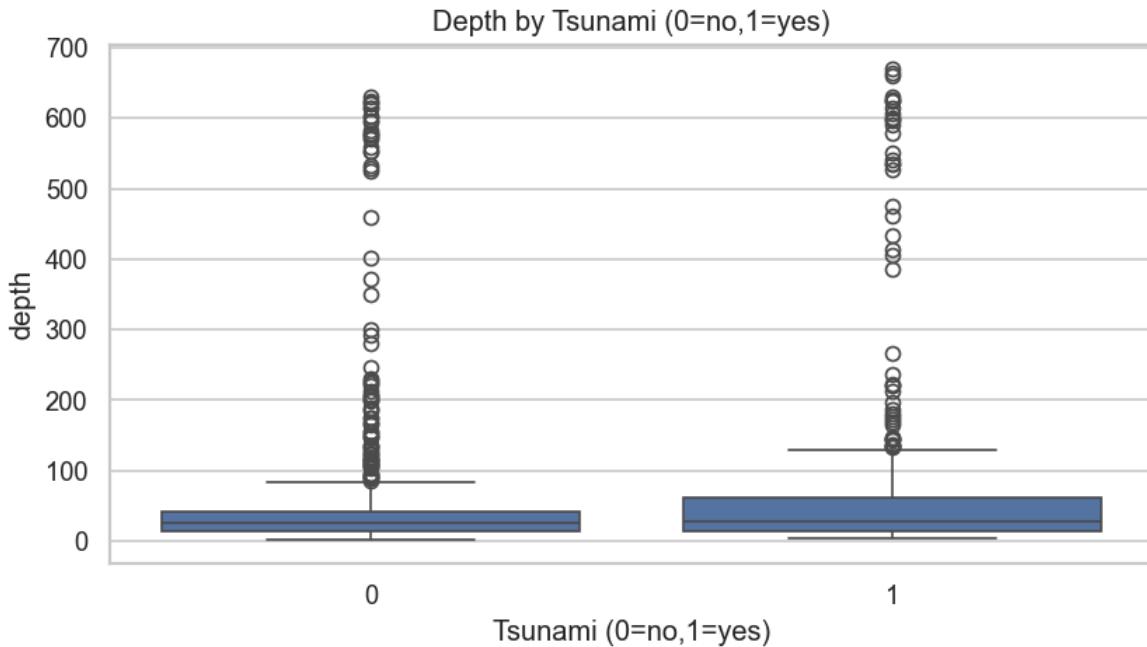
Summary stats by tsunami (0=no,1=yes):
tsunami_bin          0           1
magnitude count    478.000000  304.000000
               mean     6.942803   6.938487
               median    6.800000   6.800000
               std      0.459541   0.423250
depth       count    478.000000  304.000000
               mean    69.667356  85.656796
               median   26.000000  26.971500
               std     127.501206 151.080342
```

```
In [118... if 'magnitude' in df.columns:
    plt.figure(figsize=(8,4))
    sns.boxplot(x='tsunami_bin', y='magnitude', data=df)
    plt.title('Magnitude by Tsunami (0=no,1=yes)')
    plt.xlabel('Tsunami (0=no,1=yes)')
    plt.show()
```



```
In [119... if 'depth' in df.columns:
    plt.figure(figsize=(8,4))
    sns.boxplot(x='tsunami_bin', y='depth', data=df)
    plt.title('Depth by Tsunami (0=no,1=yes)')
```

```
plt.xlabel('Tsunami (0=no,1=yes)')
plt.show()
```



Highlight major earthquakes (≥ 8.0) and their characteristics

```
In [121...]:
if 'magnitude' in df.columns:
    majors = df[df['magnitude'] >= 8.0].sort_values('date')
print('\nMajor earthquakes (magnitude >= 8.0):')
if majors.empty:
    print('None found in dataset')
else:
    print(majors[['date', 'magnitude', 'depth', 'latitude', 'longitude', 'tsunami']])

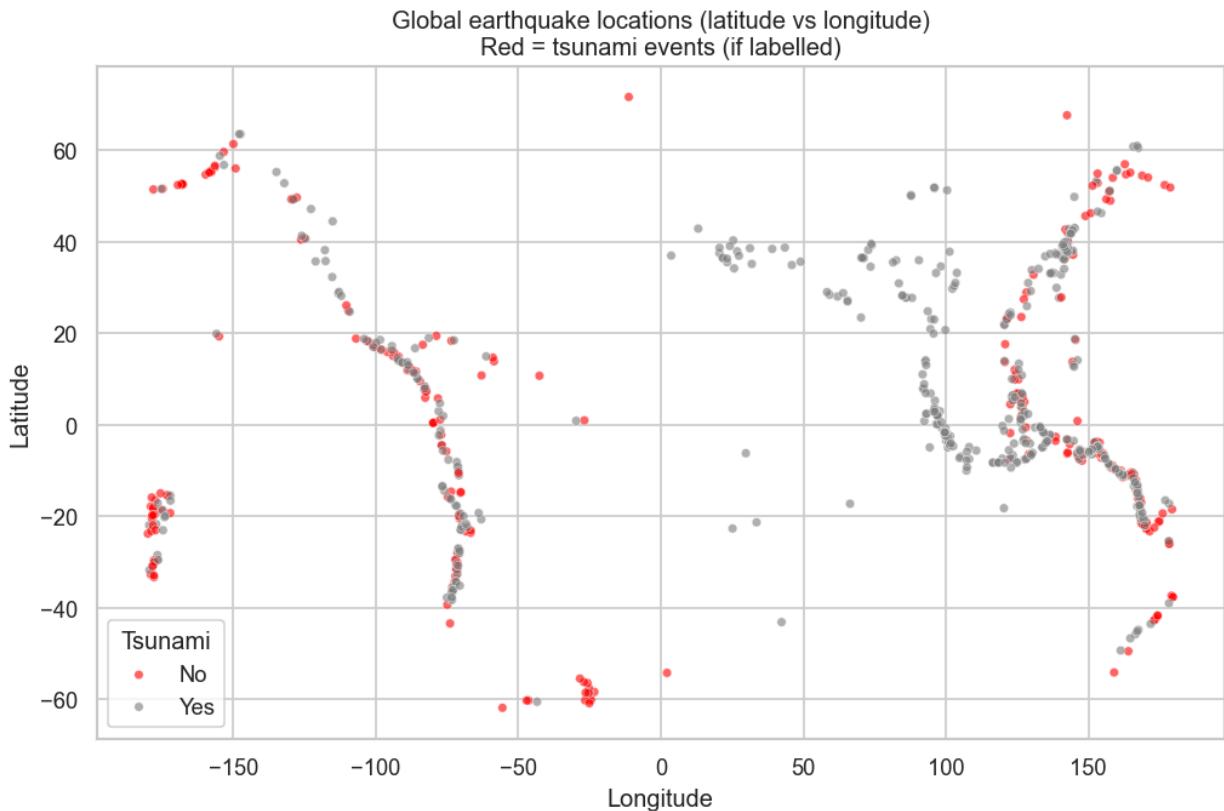
Major earthquakes (magnitude >= 8.0):
   date  magnitude  depth  latitude  longitude  tsunami
767  2001-06-01      8.40  33.00 -16.2650  -73.6410      0
712  2003-09-01      8.16  27.00  41.8150  143.9100      0
669  2004-12-01      8.10  10.00 -49.3120  161.3450      0
668  2004-12-01      9.10  30.00   3.2950   95.9820      0
657  2005-03-01      8.60  30.00   2.0850   97.1080      0
628  2006-05-01      8.00  60.50 -19.9900  -173.9070      0
627  2006-05-01      8.00  55.00 -20.1870  -174.1230      0
614  2006-11-01      8.30  10.00  46.5920  153.2660      0
611  2007-01-01      8.10  10.00  46.2430  154.5240      0
606  2007-04-01      8.10  24.00  -8.4660  157.0430      0
597  2007-08-01      8.00  39.00 -13.3860  -76.6030      0
593  2007-09-01      8.40  34.00  -4.4380  101.3670      0
535  2009-09-01      8.10  18.00 -15.4890  -172.0950      0
517  2010-02-01      8.80  22.90 -36.1220  -72.8980      0
476  2011-03-01      9.10  29.00  38.2970  142.3730      0
440  2012-04-01      8.20  25.10   0.8020   92.4630      0
441  2012-04-01      8.60  20.00   2.3270   93.0630      0
414  2013-02-01      8.00  24.00 -10.7990  165.1140      1
393  2013-05-01      8.30  598.10  54.8920  153.2210      1
356  2014-04-01      8.20  25.00 -19.6097  -70.7691      1
285  2015-09-01      8.30  22.44 -31.5729  -71.6744      1
198  2017-09-01      8.20  47.39  15.0222  -93.8993      1
170  2018-08-01      8.20  600.00 -18.1125  -178.1530      1
129  2019-05-01      8.00  122.57  -5.8119  -75.2697      1
74   2021-03-01      8.10  28.93  -29.7466  -177.2240      1
60   2021-07-01      8.20  35.00  55.3154  -157.8290      1
59   2021-07-01      8.20  46.66  55.4742  -157.9170      1
56   2021-08-01      8.10  22.79 -58.4157  -25.3206      0
```

3. Geographic Distribution Using 2D Plotting

Plot earthquake locations using latitude and longitude on a 2D scatter plot.

```
In [125...]:
if 'latitude' in df.columns and 'longitude' in df.columns:
    plt.figure(figsize=(10,6))
    sns.scatterplot(x='longitude', y='latitude', hue='tsunami_bin', data=df, palette=['gray','red'], alpha=0.6, s=20)
    plt.title('Global earthquake locations (latitude vs longitude)\nRed = tsunami events (if labelled)')
    plt.xlabel('Longitude')
```

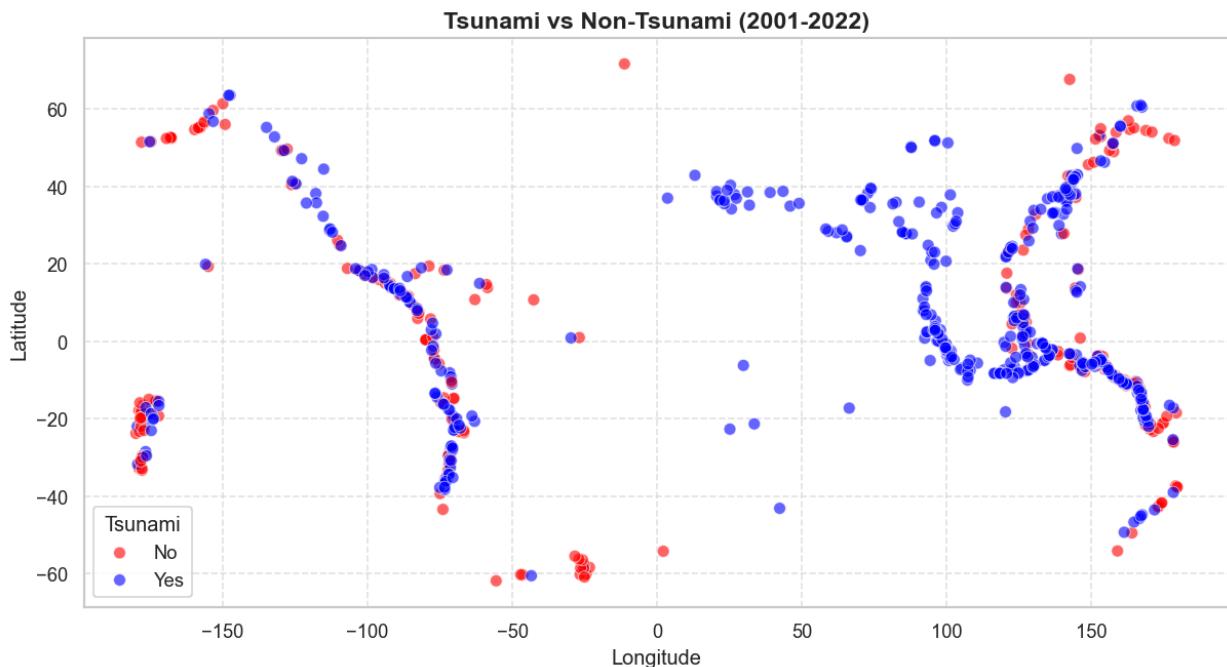
```
plt.ylabel('Latitude')
plt.legend(title='Tsunami', labels=['No', 'Yes'] if df['tsunami_bin'].nunique()>1 else ['Unknown'])
plt.show()
```



Visually distinguish between tsunami and non-tsunami events.

```
In [132...]:
if 'tsunami' in df.columns:
    df['tsunami_bin'] = df['tsunami'].apply(lambda x: 1 if x == 1 else 0)
plt.figure(figsize=(12,6))
sns.scatterplot(
    x='longitude',
    y='latitude',
    hue='tsunami_bin',
    data=df,
    palette={0: 'blue', 1: 'red'},
    alpha=0.6,
    s=50
)

plt.title("Tsunami vs Non-Tsunami (2001-2022)", fontsize=14, fontweight='bold')
plt.xlabel("Longitude")
plt.ylabel("Latitude")
plt.legend(title='Tsunami', labels=['No', 'Yes'])
plt.grid(True, linestyle='--', alpha=0.5)
plt.show()
```



Identify clusters or regions with higher concentration of tsunami events (without using map tiles or interactive maps)

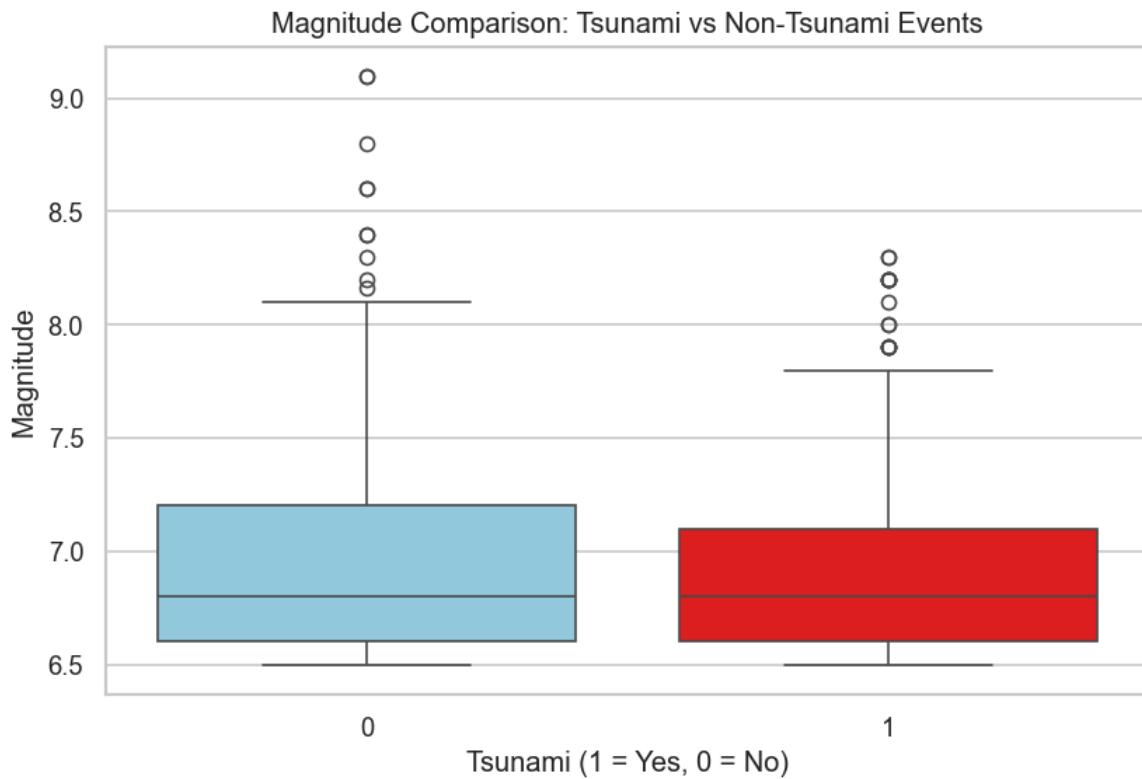
```
In [143... # Define number of bins for grid
nbins = 60
lon_idx = np.floor((df['longitude'] - df['longitude'].min()) / (df['longitude'].max() - df['longitude'].min()) * nbins)
lat_idx = np.floor((df['latitude'] - df['latitude'].min()) / (df['latitude'].max() - df['latitude'].min()) * nbins)
df['lon_bin'] = lon_idx.astype(int)
df['lat_bin'] = lat_idx.astype(int)
grid = df.groupby(['lon_bin', 'lat_bin'])['tsunami_bin'].sum().reset_index(name='tsunami_count')
top_cells = grid.sort_values('tsunami_count', ascending=False).head(10)
print("\nTop grid cells with most tsunami events (approx regions):")
print(top_cells)

Top grid cells with most tsunami events (approx regions):
   lon_bin  lat_bin  tsunami_count
216        55       25              11
235        57       22              11
215        55       24               9
234        57       21               9
224        56       23               8
242        58       17               7
2         0       14               7
243        58       18               7
14         3       52               6
42        14       34               6
```

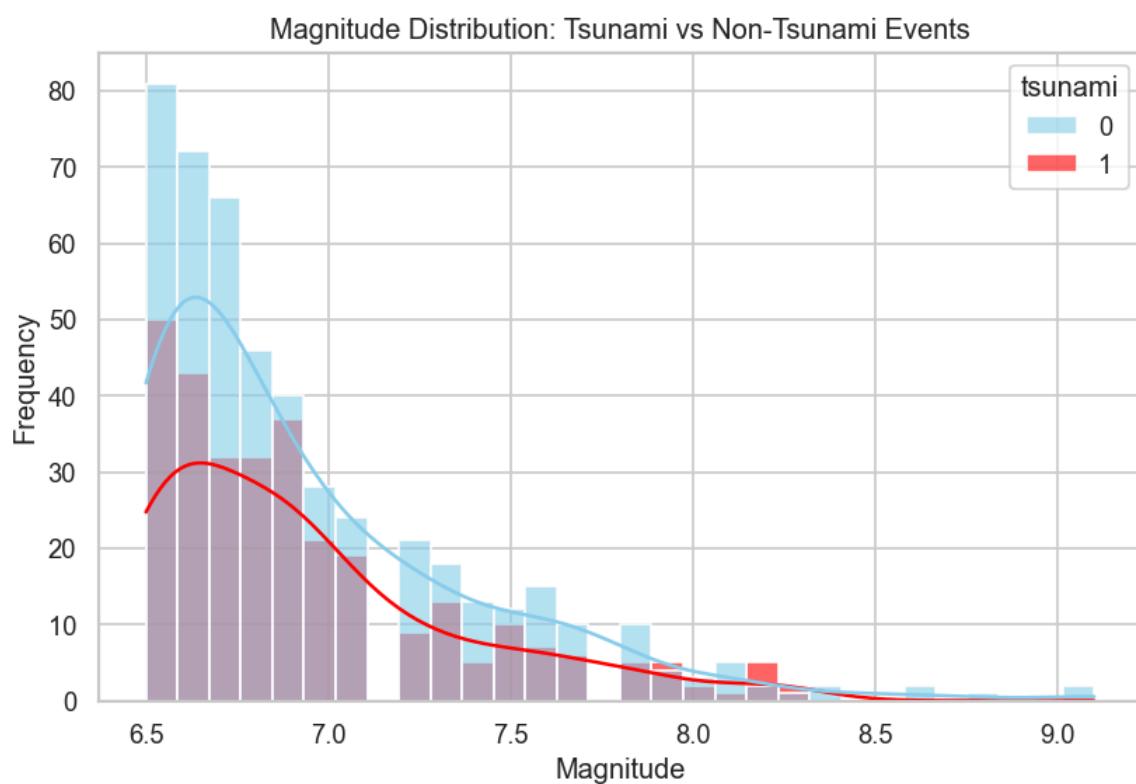
4. Statistical and Comparative Analysis:

Use box plots, histograms, and bar charts to compare seismic features between tsunami and non-tsunami events.

```
In [152... # Boxplot: Magnitude
plt.figure(figsize=(8, 5))
sns.boxplot(
    x='tsunami', y='magnitude',
    hue='tsunami', data=df,
    palette={0: 'skyblue', 1: 'red'},
    legend=False
)
plt.title("Magnitude Comparison: Tsunami vs Non-Tsunami Events")
plt.xlabel("Tsunami (1 = Yes, 0 = No)")
plt.ylabel("Magnitude")
plt.show()
sns.set(style="whitegrid", palette="coolwarm")
df['tsunami'] = df['tsunami'].astype(int)
```

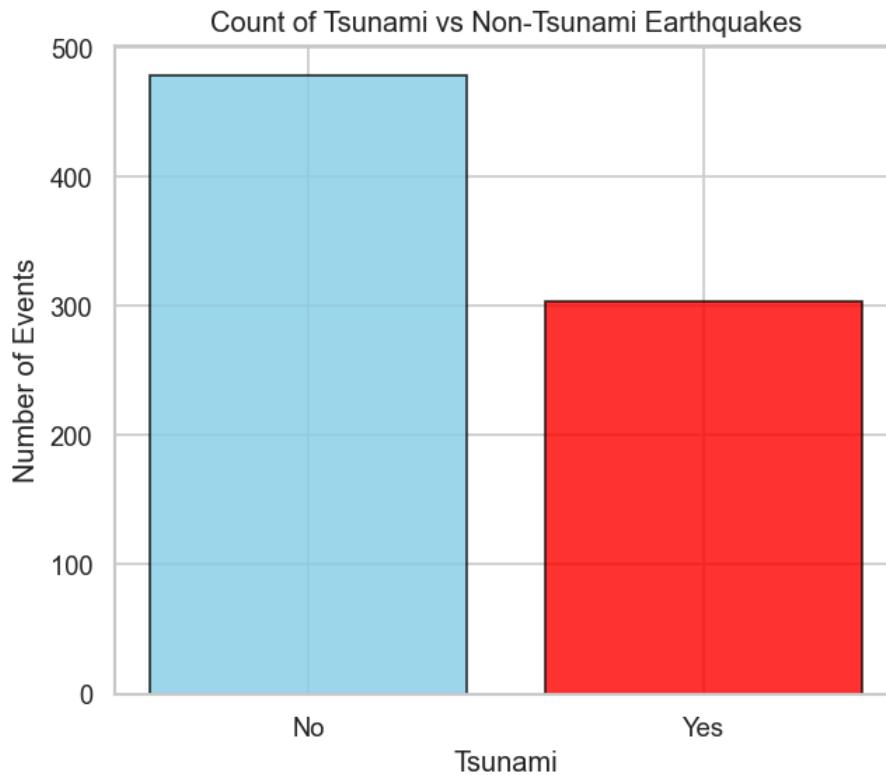


```
In [154...]
# Histogram: =Magnitude
plt.figure(figsize=(8, 5))
sns.histplot(
    data=df,
    x='magnitude',
    hue='tsunami',
    bins=30,
    kde=True,
    palette={0: 'skyblue', 1: 'red'},
    alpha=0.6
)
plt.title("Magnitude Distribution: Tsunami vs Non-Tsunami Events")
plt.xlabel("Magnitude")
plt.ylabel("Frequency")
plt.show()
```



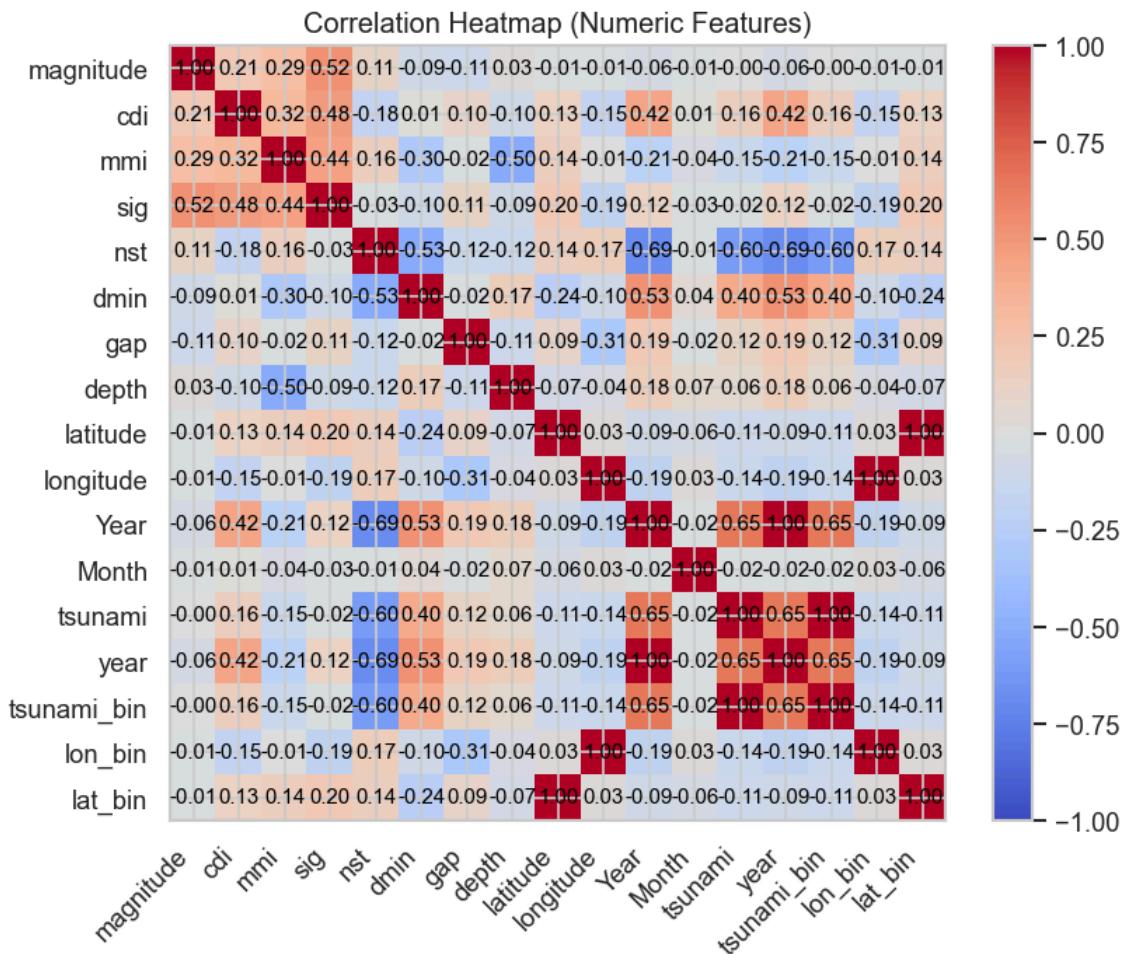
```
In [159]: # Bar plot
if 'tsunami' in df.columns:
    counts = df['tsunami'].value_counts().sort_index() # 0=no, 1=yes
    plt.figure(figsize=(6,5))
    plt.bar(['No', 'Yes'], counts.values, color=['skyblue','red'], edgecolor='black', alpha=0.8)

    plt.title("Count of Tsunami vs Non-Tsunami Earthquakes", fontsize=12)
    plt.xlabel("Tsunami")
    plt.ylabel("Number of Events")
    plt.show()
```



Analyze correlations between variables using heatmaps.

```
In [164]: numeric_cols = df.select_dtypes(include=[np.number]).columns.tolist()
if numeric_cols:
    corr = df[numeric_cols].corr()
    plt.figure(figsize=(8,6))
    im = plt.imshow(corr, cmap="coolwarm", vmin=-1, vmax=1)
    plt.colorbar(im, fraction=0.046, pad=0.04)
    for i in range(len(corr)):
        for j in range(len(corr)):
            plt.text(j, i, f"{corr.iloc[i, j]:.2f}", ha='center', va='center', color='black', fontsize=9)
    plt.xticks(range(len(numeric_cols)), numeric_cols, rotation=45, ha='right')
    plt.yticks(range(len(numeric_cols)), numeric_cols)
    plt.title("Correlation Heatmap (Numeric Features)")
    plt.tight_layout()
    plt.show()
    print("\nNumeric columns used in correlation:", numeric_cols)
```



Numeric columns used in correlation: ['magnitude', 'cdi', 'mmi', 'sig', 'nst', 'dmin', 'gap', 'depth', 'latitude', 'longitude', 'Year', 'Month', 'tsunami', 'year', 'tsunami_bin', 'lon_bin', 'lat_bin']

5. Insights and Observations

Summarize key differences in seismic behavior between tsunami and non-tsunami earthquakes.

Key Differences

Magnitude

Tsunami-generating earthquakes generally have higher magnitudes (often ≥ 7.0 – 8.0) compared to non-tsunami earthquakes.

Depth

Tsunami earthquakes are usually shallow, typically occurring at depths <70 km, which allows energy to displace water efficiently.

Non-tsunami earthquakes often occur at deeper levels.

Location / Geography

Tsunami events are clustered along subduction zones and coastal regions, whereas non-tsunami earthquakes can occur more broadly, including inland.

Frequency

Non-tsunami earthquakes are far more frequent, but only a small fraction generate tsunamis.

Correlation of Features

Magnitude and shallow depth are the strongest indicators for potential tsunamis.

Other features (e.g., latitude, longitude) are secondary but highlight high-risk zones.

Identify seismic thresholds or indicators associated with increased tsunami potential

Key Indicators for Tsunami Potential

Magnitude Threshold

Earthquakes with magnitude ≥ 7.0 are far more likely to generate tsunamis.

Major tsunamis often occur with $M \geq 8.0$ events.

Shallow Depth

Depth < 70 km is a critical factor.

Shallow earthquakes transfer more energy to the ocean floor, displacing water efficiently.

Location / Tectonic Setting

Subduction zones (convergent plate boundaries) are hotspots.

Coastal regions near trenches or faults are at higher risk.

Energy Release and Rupture Type

Vertical displacement of the seafloor (thrust faults) increases tsunami potential.

Horizontal strike-slip faults rarely produce significant tsunamis.

Historical Clusters

Areas with repeated shallow, high-magnitude events are more prone to tsunami recurrence.

In []: