



STTHK2113 DATA ANALYTICS

ASSIGNMENT 1

PREPARED FOR:

Assoc. Prof. Dr. Yuhanis Yusof

NAME	MATRIC NO.
MUHAMAD EZZ ZIKRY BIN EZRY	294672

Task 1 - Dataset Selection and Description

1. Dataset Title

- Global Earthquake and Tsunami Risk Assessment Dataset

2. Source and Citation

- Uzaki, A. (2023). *Global Earthquake and Tsunami Risk Assessment Dataset*. Kaggle. Retrieved October 2025 from <https://www.kaggle.com/datasets/ahmeduzaki/global-earthquake-tsunami-risk-assessment-dataset>

3. Domain/Context

- Environmental Analytics / Disaster Risk Assessment

The Global Earthquake and tsunami Risk Assessment Dataset is a comprehensive, research-ready dataset containing global seismic characteristics and tsunami potential indicators for 782 significant earthquakes recorded between January 1, 2001 and December 21, 2022. It was designed for earthquake-impact assessment, tsunami-risk prediction, and spatial coverage (latitude - 61.85° to 71.63° , longitude - 179.97° to 179.66°). It provides the foundation for understanding the relationship between earthquake magnitude, depth, and tsunami occurrence worldwide.

4. Dataset Overview

Attribute	Description	type	Range/ Values	Tsunami Relevance
magnitude	Earthquake magnitude (Richter Scale)	Float	6.5 - 9.1	High - Primary Actor
cdi	Community Decimal intensity (felt Intensity)	Integer	0 - 9	Medium - Population Impact
mmi	Modified Mercalli Intensity (Instrumental)	Integer	1 - 9	Medium - Structural Damage
sig	Event significance score	Integer	650 - 2910	High - Hazard Assessment

nst	Number of reporting stations	Integer	0 - 934	Low - Data Quality
dmin	Distance to nearest station (°)	Float	0 - 17.7	Low - Location precision
gap	Azimuthal gap (°)	Float	0 - 239	Low - Location reliability
depth	Earthquake focal depth (km)	Float	2.7 - 670.8	High - Shallow = higher tsunami risk
latitude	Epicenter latitude (WGS84)	Float	61.85° - 71.63°	High - Ocean Proximity
longitude	Epicenter longitude (WGS84)	Float	179.97° - 179.66°	High - Ocean Proximity
Year	Year of occurrence	Integer	2001 - 2022	Medium - Temporal patterns
Month	Month of occurrence	Integer	1 - 12	Low - Seasonal pattern
tsunami	Tsunami potential (1 = Yes, 0 = No)	Binary	0 / 1	Target variable

- Total Records : 782
- Attributes : 13 (numerical + binary/categorical)

5. Dataset Source Citation

- WarCoder. (2022). *Earthquake Dataset*. Kaggle. Retrieved October 2025, from <https://www.kaggle.com/datasets/warcoder/earthquake-dataset>

6. Justification for Dataset Selection

- It provides real-world relevance, addressing the global issue of earthquake and tsunami risk.
- It demonstrates sufficient breadth and depth, including geographic, temporal, and seismological factors.
- It is complete and well structured, and aligned with the requirements (>200 records, 5-10 attributes).
- It supports various analytical techniques, including descriptive, correlation, and classification analysis.
- It aligns with data analytics learning objectives, showcasing end-to-end preprocessing from cleaning to formatting.

Task 2 - Data Preprocessing

Step 1 : Data Cleaning (Checking Duplicates & Missing Values)

```
# Remove duplicates
df = df.drop_duplicates()

# Confirm dataset shape
df.shape
```

(782, 13)

```
print("\nChecking for missing values\n")
df.isnull().sum()
```

Checking for missing values

	0
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

Figure 1: Duplicate and missing-value check showing a complete and unique dataset.

- The dataset was first checked for duplicate and missing entries.
- Using `drop_duplicates()`, no duplicate records were found, and the dataset retained 782 rows and 13 attributes.
- The `info()` and `isnull().sum()` functions confirmed that all columns were complete, with zero missing values.

Step 2 : Outlier Detection and Removal

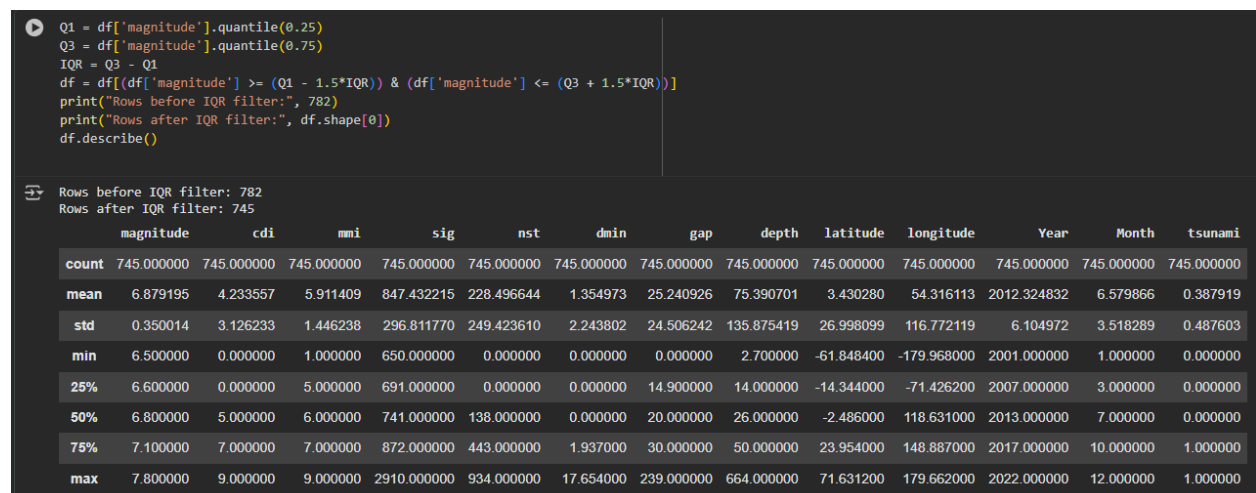
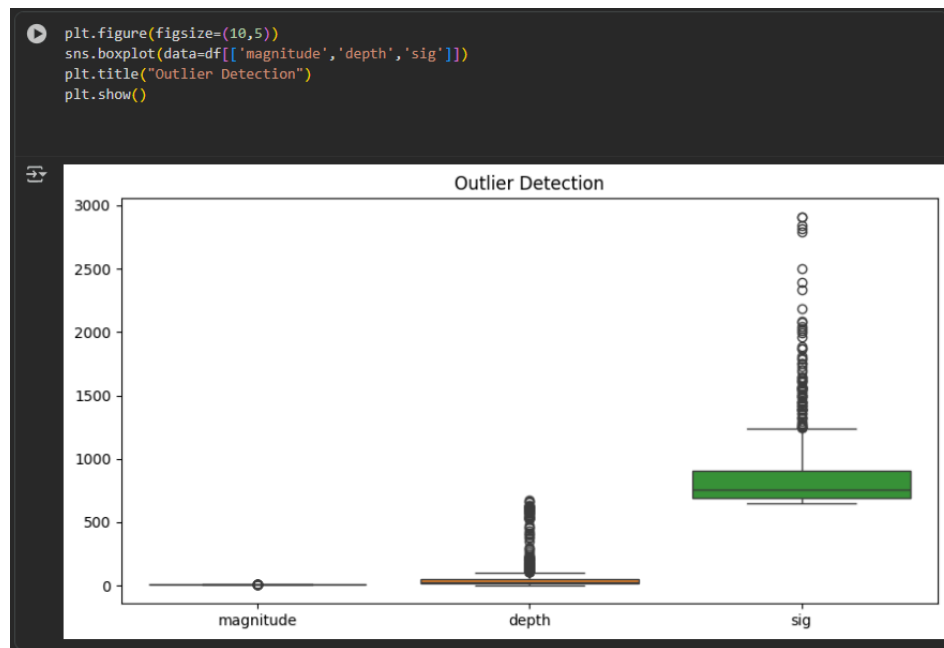


Figure 2: Outlier detection (boxplot) and IQR filtering reducing dataset to 745 consistent records.

- Outliers were identified using boxplots for the numerical variables magnitude, depth, and sig.
- The Interquartile Range (IQR) method was applied to remove extreme values in the magnitude column.
- After filtering, 37 outliers were removed, reducing the dataset from 782 to 745 records.
- The descriptive statistics confirmed that the maximum magnitude decreased from 9.1 to 7.8, indicating successful cleaning.

Step 3 : Data Transformation

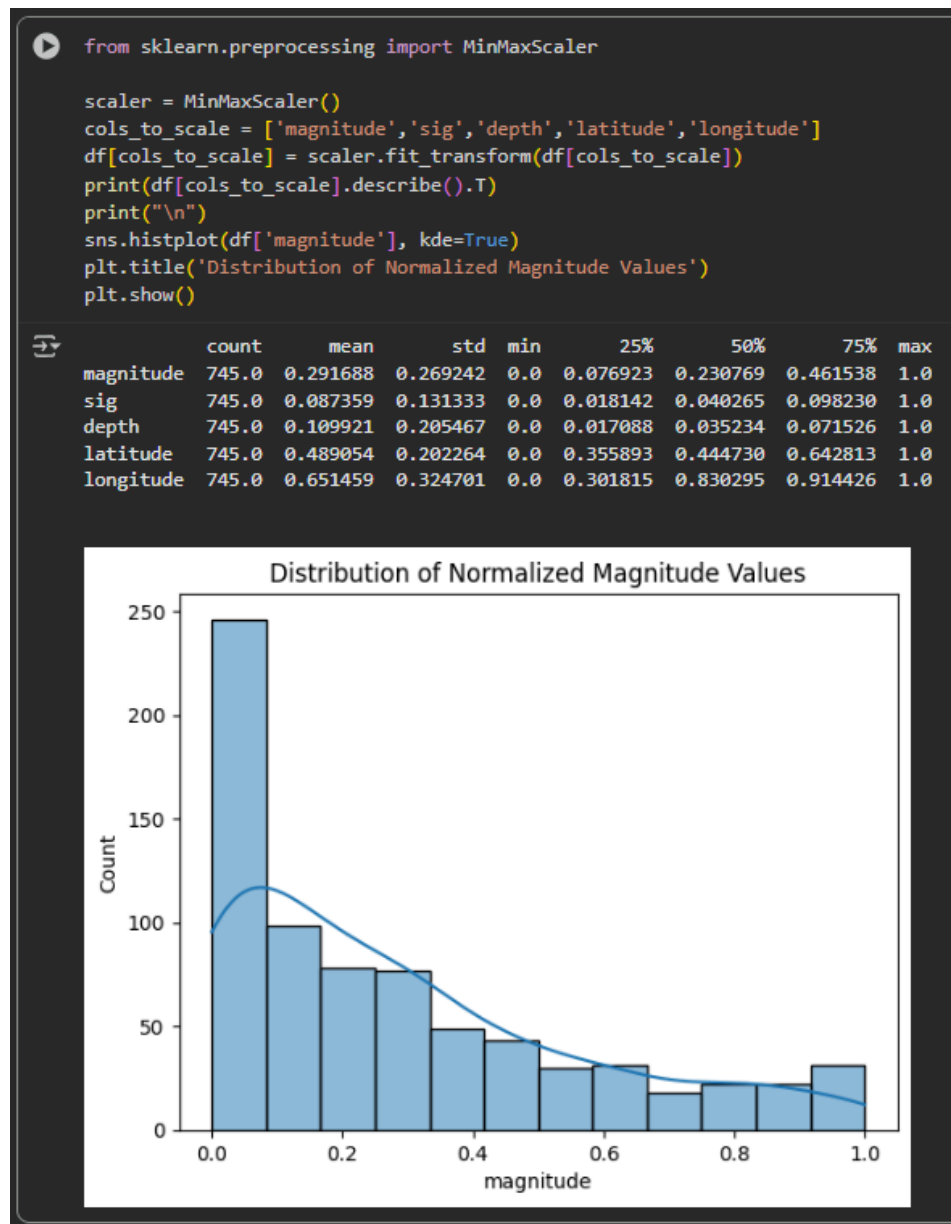


Figure 3: Normalization results confirming all numerical attributes rescaled between 0 and 1.

- To ensure consistent feature scaling, the Min-Max normalization technique was applied to numerical columns (magnitude, sig, depth, latitude, longitude).
- This transformation rescales data between 0 and 1, eliminating unit disparities.
- The descriptive statistics and histogram visualization confirmed successful normalization.

Step 4 : Categorical Encoding

```
df['tsunami'] = df['tsunami'].astype('category')
df['tsunami'] = df['tsunami'].cat.codes
print(df['tsunami'].unique())

[1 0]
```

Figure 4: Binary encoding applied to the tsunami column.

- The tsunami variable was encoded into numeric form for analytical compatibility.
- Using Pandas' category codes, values were converted to 1 = Tsunami occurred and 0 = No tsunami.
- The unique() function verified the correct binary encoding.

Step 5 : Final Validation and Summary

```
print("Rows:", df.shape[0])
print("Columns:", df.shape[1])
print("Missing values:", df.isnull().sum().sum())
df.describe()
```

Rows: 745
Columns: 13
Missing values: 0

	magnitude	cdi	mmi	sig	nst	dmin	gap	depth	latitude	longitude	Year	Month	tsunami
count	745.000000	745.000000	745.000000	745.000000	745.000000	745.000000	745.000000	745.000000	745.000000	745.000000	745.000000	745.000000	745.000000
mean	0.291688	4.233557	5.911409	0.087359	228.496644	1.354973	25.240926	0.109921	0.489054	0.651459	2012.324832	6.579866	0.387919
std	0.269242	3.126233	1.446238	0.131333	249.423610	2.243802	24.506242	0.205467	0.202264	0.324701	6.104972	3.518289	0.487603
min	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	2001.000000	1.000000	0.000000
25%	0.076923	0.000000	5.000000	0.018142	0.000000	0.000000	14.900000	0.017088	0.355893	0.301815	2007.000000	3.000000	0.000000
50%	0.230769	5.000000	6.000000	0.040265	138.000000	0.000000	20.000000	0.035234	0.444730	0.830295	2013.000000	7.000000	0.000000
75%	0.461538	7.000000	7.000000	0.098230	443.000000	1.937000	30.000000	0.071526	0.642813	0.914426	2017.000000	10.000000	1.000000
max	1.000000	9.000000	9.000000	1.000000	934.000000	17.654000	239.000000	1.000000	1.000000	1.000000	2022.000000	12.000000	1.000000

Figure 5: Final dataset summary confirming completeness and consistency after preprocessing.

- After preprocessing, the dataset contained 745 rows and 13 columns, with no missing values.
- The final structure confirms that all transformations were applied correctly, and the data is now clean, normalized, and ready for analysis.