IE6600 Computation and Visualization

Project 1:
Analysis and Visualization of Earthquake Data

Group 1:
Aishwarya Belakavadi Subrahmanya
Haodong Liu
Naveen Dhanasekaran
Yizhe Liu
Ziyue Yin

Date: January 24, 2024

Contents

1.INTRODUCTION

Earthquakes are natural phenomena that can have significant impacts on communities, infrastructure, and the environment. Understanding the patterns and dynamics of seismic activity is crucial for earthquake preparedness, risk assessment, and scientific research. In this context, the visualization of earthquake datasets becomes a powerful tool to uncover patterns, trends, and spatial distributions.

The dataset at hand comprises information about earthquakes, including their time of occurrence, geographic coordinates, depth, magnitude, and other relevant parameters. Visualization of this dataset provides an opportunity to gain insights into the temporal and spatial aspects of seismic activity.

Objectives of Visualization:

1. **Temporal Insights:**
    - Explore the temporal patterns of earthquake occurrences over time. Are there trends or cyclic patterns in seismic activity?
2. **Spatial Distribution:**
    - Analyze the geographic distribution of earthquakes to identify hotspots or regions with higher seismic activity. How are earthquakes distributed across different geographic locations?
3. **Magnitude Analysis:**
    - Visualize the distribution of earthquake magnitudes. Are there trends in the magnitude of seismic events over time or across regions?
4. **Cluster Analysis:**
    - Perform cluster analysis to identify groups of earthquake locations with similar characteristics. Are there distinct spatial clusters or patterns?
5. **Time Series Decomposition:**
    - Decompose the time series data to understand trends, seasonality, and anomalies in monthly earthquake occurrences.

Significance of Visualization:

- Visualization plays a crucial role in communicating complex seismic data to a wide audience, including scientists, policymakers, emergency response teams, and the general public.
- The insights gained from visualization can inform earthquake preparedness strategies, resource allocation, and decision-making related to infrastructure planning and disaster response.
- Identification of spatial clusters and hotspots aids in understanding the geological features and tectonic processes that contribute to seismic activity.
- Temporal patterns revealed through visualization contribute to scientific understanding and may guide further research into the factors influencing

earthquake occurrences.

Methods Used:

- The analysis involves data cleaning, normalization, and encoding of categorical variables.
- Exploratory Data Analysis (EDA) is conducted using static visualizations, leveraging libraries such as Matplotlib and Seaborn.
- Clustering analysis is performed to identify spatial groups of earthquake locations using KMeans.
- Time series decomposition is employed to extract trends, seasonality, and residuals from the monthly earthquake occurrences.

2.Data Inspection and Cleaning

Before starting to clean the raw data, we first inspect the data. Open the dataset and format it as a Pandas Dataframe. Implement the code below to display the first few rows to understand its general structure. There are 26642 rows and 22 columns.

	time	latitude	longitude	depth	mag	magType	nst	gap	dmin	rms	...	updated	place	type	horizontalError	depthError	magError	magNst	status	locationSource	magSource
0	2023-01-01T00:49:25.294Z	52.099900	178.521800	82.770	3.10	ml	14.0	139.0	0.87000	0.18	...	2023-03-11T22:51:52.040Z	Rat Islands, Aleutian Islands, Alaska	earthquake	8.46	21.213	0.097	14.0	reviewed	us	us
1	2023-01-01T01:41:43.755Z	7.139700	126.738000	79.194	4.50	mb	32.0	104.0	1.15200	0.47	...	2023-03-11T22:51:45.040Z	23 km ESE of Manay, Philippines	earthquake	5.51	7.445	0.083	43.0	reviewed	us	us
2	2023-01-01T03:29:31.070Z	19.163100	-66.525100	24.000	3.93	md	23.0	246.0	0.84790	0.22	...	2023-03-11T22:51:29.040Z	Puerto Rico region	earthquake	0.91	15.950	0.090	16.0	reviewed	pr	pr
3	2023-01-01T04:09:32.814Z	-4.780300	102.767500	63.787	4.30	mb	17.0	187.0	0.45700	0.51	...	2023-03-11T22:51:45.040Z	99 km SSW of Pagar Alam, Indonesia	earthquake	10.25	6.579	0.238	5.0	reviewed	us	us
4	2023-01-01T04:29:13.793Z	53.396500	-166.941700	10.000	3.00	ml	19.0	190.0	0.40000	0.31	...	2023-03-11T22:51:38.040Z	59 km SSW of Unalaska, Alaska	earthquake	1.41	1.999	0.085	18.0	reviewed	us	us
5	2023-01-01T04:50:17.639Z	19.281100	-155.428200	37.751	2.80	ml	19.0	127.0	0.06600	0.18	...	2023-03-11T22:51:29.040Z	10 km NNE of P?hala, Hawaii	earthquake	2.77	5.266	0.060	36.0	reviewed	us	us
6	2023-01-01T04:54:53.914Z	-19.041900	-177.542300	556.990	4.10	mb	15.0	87.0	3.05100	0.15	...	2023-03-11T22:51:45.040Z	Fiji region	earthquake	12.85	13.028	0.213	6.0	reviewed	us	us
7	2023-01-01T05:02:46.402Z	-15.321900	-174.875600	255.470	4.10	mb	40.0	81.0	3.41300	0.32	...	2023-03-11T22:51:45.040Z	Tonga	earthquake	9.84	6.047	0.095	34.0	reviewed	us	us

We used the ‘df.info()’ method in Pandas to obtain a concise summary of information about a DataFrame. This method can show us overall data frame information, names and non-null counts of each column, and memory usage. We also used the ‘df.describe(include = "all")’ method to gain an initial understanding of the distribution of data, detecting outliers, and understanding the central tendency and variability of the data.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26642 entries, 0 to 26641
Data columns (total 22 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   time             26642 non-null  object
 1   latitude         26642 non-null  float64
 2   longitude        26642 non-null  float64
 3   depth            26642 non-null  float64
 4   mag              26642 non-null  float64
 5   magType          26642 non-null  object
 6   nst              25227 non-null  float64
 7   gap              25225 non-null  float64
 8   dmin             24776 non-null  float64
 9   rms              26642 non-null  float64
 10  net              26642 non-null  object
 11  id               26642 non-null  object
 12  updated          26642 non-null  object
 13  place            25034 non-null  object
 14  type             26642 non-null  object
 15  horizontalError  25093 non-null  float64
 16  depthError       26642 non-null  float64
 17  magError         24970 non-null  float64
 18  magNst           25065 non-null  float64
 19  status           26642 non-null  object
 20  locationSource   26642 non-null  object
 21  magSource        26642 non-null  object
dtypes: float64(12), object(10)
memory usage: 4.5+ MB
```

	time	latitude	longitude	depth	mag	magType	nst	gap	dmin	rms	...
count	26642	26642.000000	26642.000000	26642.000000	26642.000000	26642	25227.000000	25225.000000	24776.000000	26642.000000	...
unique	24682	NaN	NaN	NaN	NaN	15	NaN	NaN	NaN	NaN	...
top	2023-04-06 01:08:20.773000+00:00	NaN	NaN	NaN	NaN	mb	NaN	NaN	NaN	NaN	...
freq	2	NaN	NaN	NaN	NaN	15906	NaN	NaN	NaN	NaN	...
first	2023-01-01 00:49:25.294000+00:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
last	2023-12-29 23:17:18.800000+00:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
mean	NaN	16.852798	-11.487497	67.491224	4.007395	NaN	42.571332	124.930971	2.692908	0.581575	...
std	NaN	30.389200	130.053399	116.762456	0.794423	NaN	37.662352	67.430145	4.043568	0.256276	...
min	NaN	-65.849700	-179.998700	-3.370000	2.600000	NaN	0.000000	8.000000	0.000000	0.010000	...
25%	NaN	-6.415275	-149.608650	10.000000	3.220000	NaN	19.000000	73.000000	0.612000	0.410000	...
50%	NaN	18.884167	-64.811833	21.998000	4.300000	NaN	30.000000	111.000000	1.579000	0.590000	...
75%	NaN	41.827950	126.965100	66.833000	4.500000	NaN	52.000000	165.000000	3.172000	0.750000	...
max	NaN	86.593900	179.999400	681.238000	7.800000	NaN	423.000000	350.000000	50.820000	1.880000	...

After that, we used’ df.isna().sum()’ method to count the number of missing values in each column of this DataFrame. There are 1415 missing values in the column nst, which we will fill with mean of the column. There are 1417 missing values in the column gap, which we will fill with mean of the column. There are 1866 missing values in the column dmin, which we will fill with mean of the column. There are 1866 missing values in the column place, which we will fill with “UNKNOWN”. There are 1549 missing values in the horizontal error, which we will fill with mean of the column. There are 1672 missing values in the mag error, which we will fill with mean of the column. There are 1577 missing values in the magnst, which we will fill with mean of the column. Then, we use “pd.to_datetime()” to change time and updated columns into time

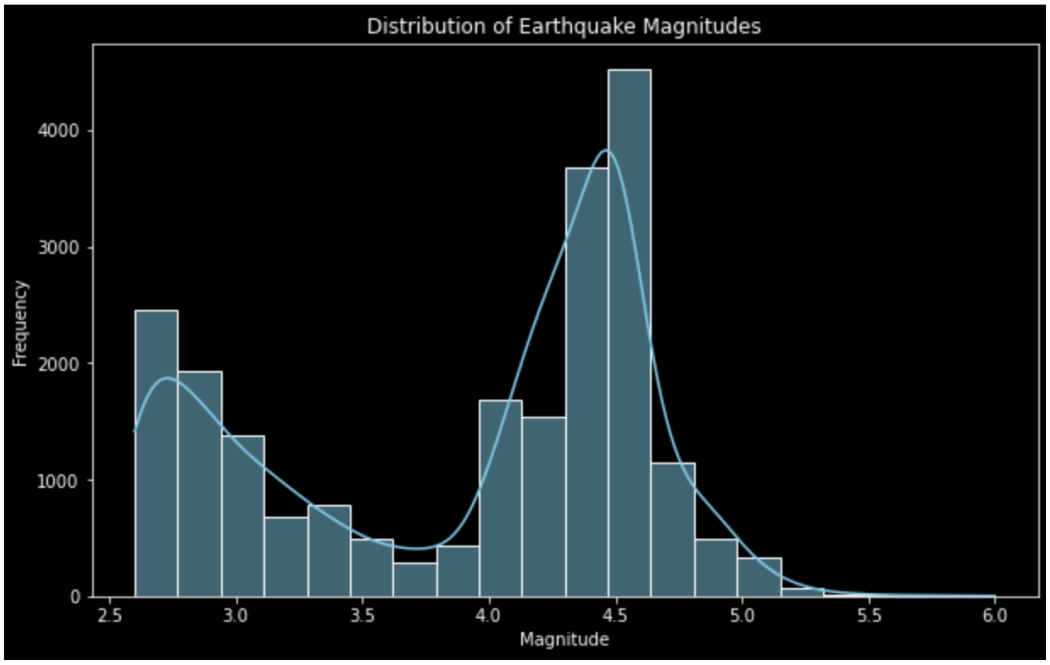
type. Finally, we use “drop_duplicates()” to drop 1960 columns and obtain a cleaned dataset with 24682 rows and 22 columns.

```
time                0
latitude            0
longitude           0
depth               0
mag                 0
magType             0
nst              1415
gap              1417
dmin             1866
rms                 0
net                 0
id                  0
updated             0
place            1608
type                0
horizontalError  1549
depthError          0
magError         1672
magNst           1577
status              0
locationSource      0
magSource           0
dtype: int64
```

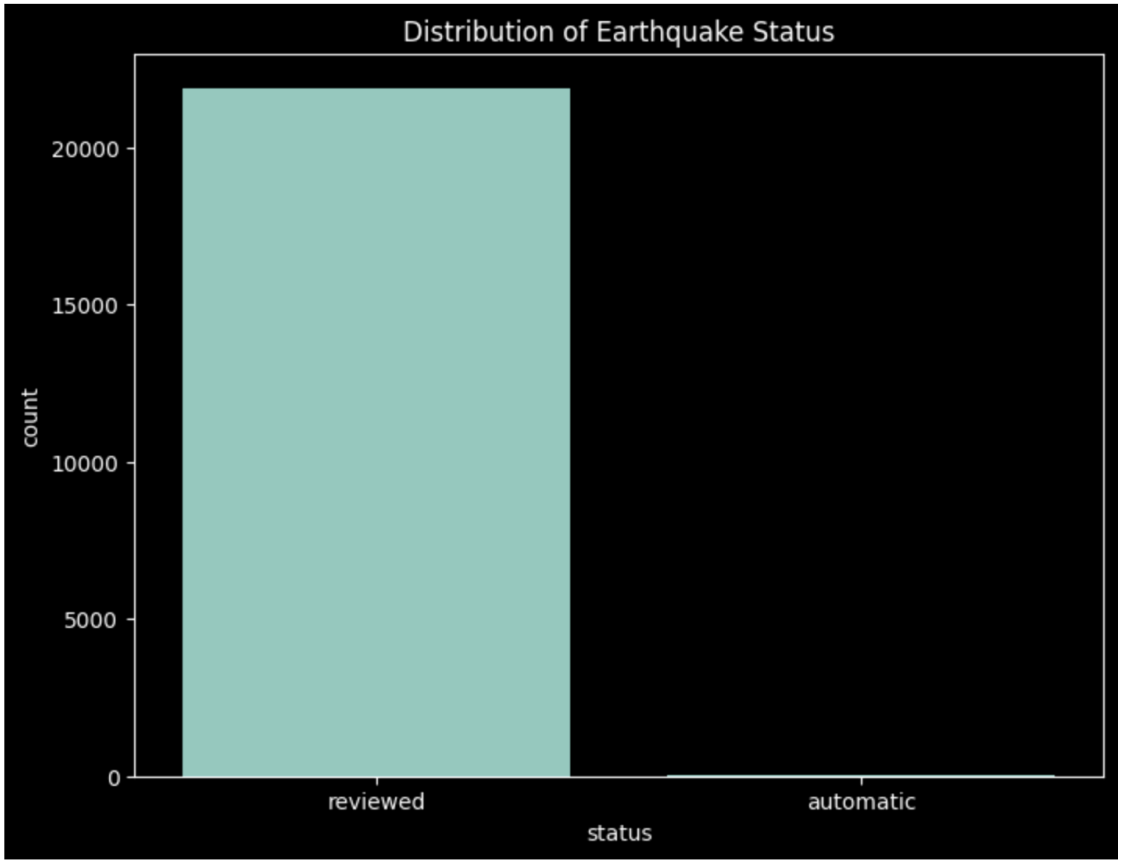
3. Exploratory Data Analysis

3.1: Earthquake Magnitude Distribution

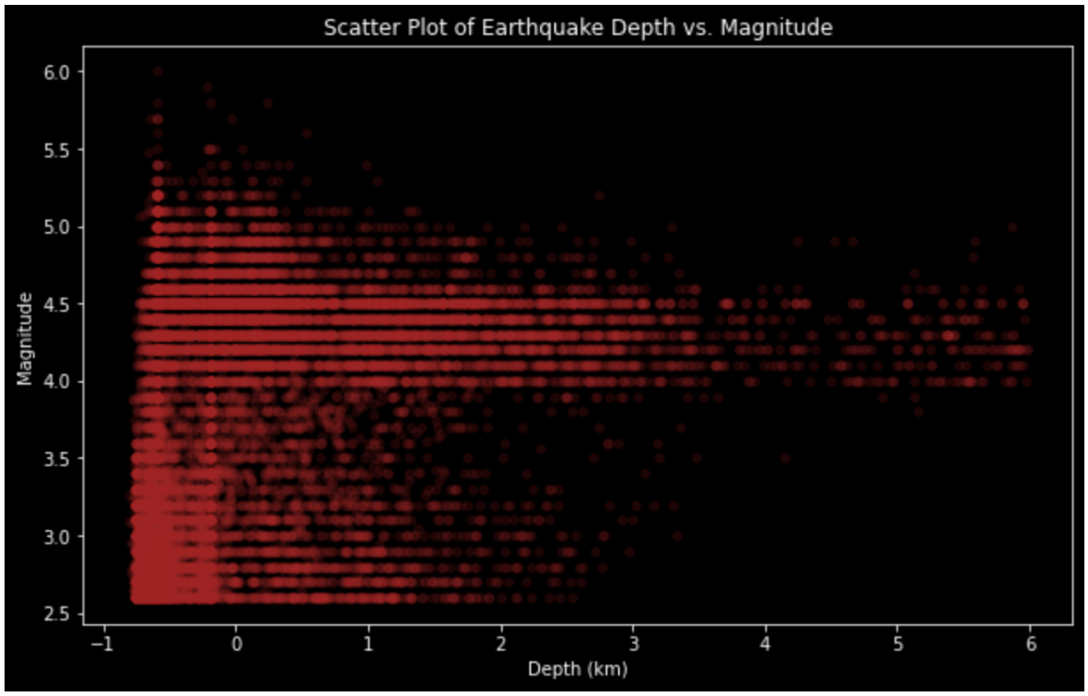
We used the histogram of the earthquake magnitude to create the visualization of the distribution of the earthquake magnitudes based on the x and y labels, which is corresponding to the frequency and the magnitude. By using the histogram plot, our group tries to get more insight and more direct visualization about the trend of the magnitude and frequency. And the conclusion is that the earthquake magnitude was around 4.5.The other thing is that the frequency of the magnitude around 2.5 and 4.0 is larger compared to the other values.



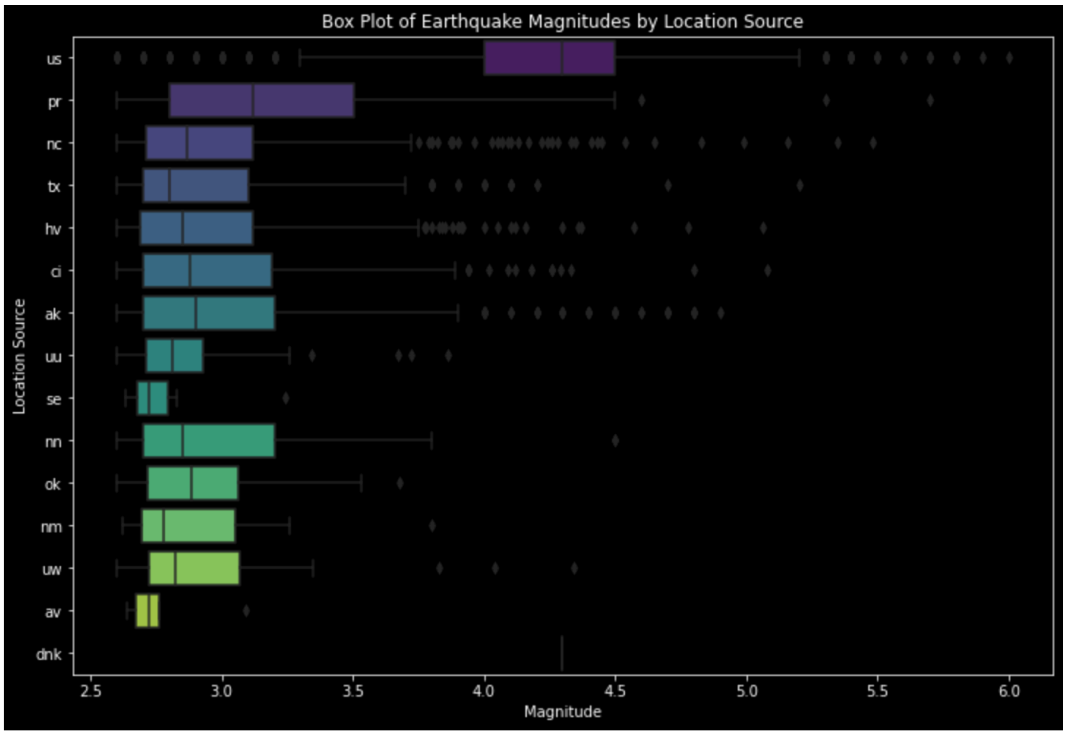
3.2:Earthquake Status Distribution:
In the code section our team chose to use the countplot, and in the following we show a bar graph to represent and compare the status of the earthquakes, and this graph provides us with more insights. From the data and the images we know that most of the earthquakes fall into the "reviewed" category, which indicates a high level of review and validation of the monitoring of seismic events. This indicates a high level of review and validation of seismic event monitoring.



3.3: Scatter plot of earthquake depth vs. magnitude:
In the code section our group chose to use scatterplo, which facilitates us to check the relationship between magnitude as well as depth in the subsequent analysis. Each point in the scatterplot is a seismic event, and the density of the points can be used to analyze the pattern and relationship between magnitude and depth. Using the dataset, we can see that the depths of lower magnitudes are more concentrated between 0 and -1, while the depths of larger magnitudes are spread out to greater depths and are evenly distributed between -1 and 2. But for Larger magnitudes, such as 4.5 and above, are distributed between 0 and -1.

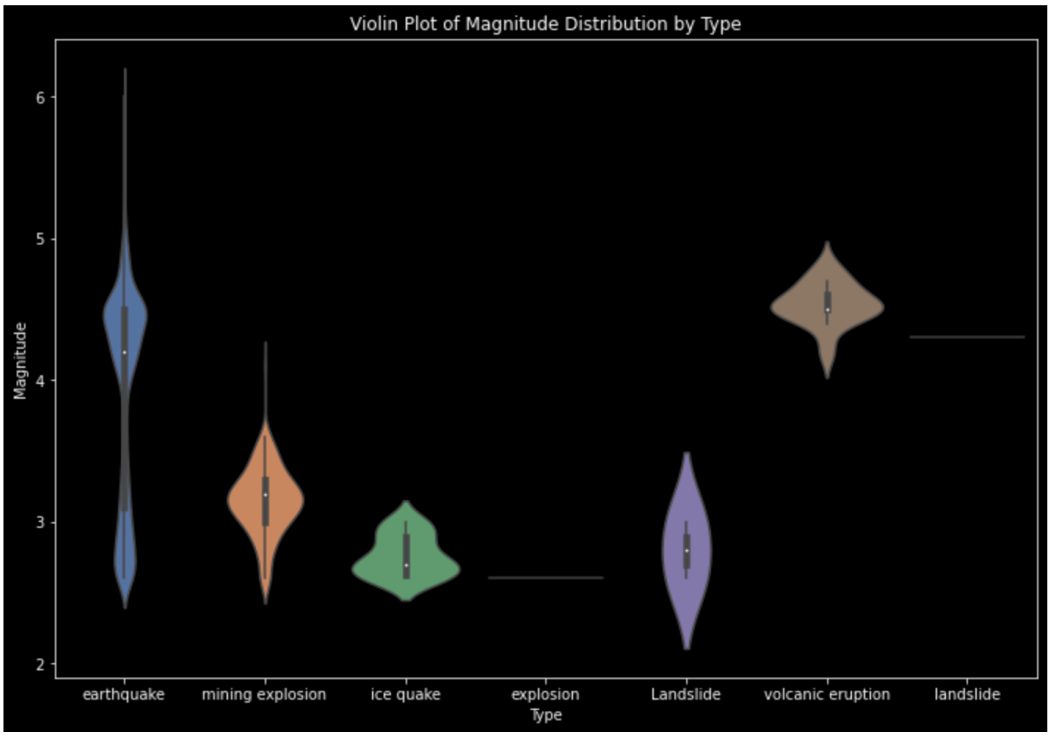


3.4: Box plot of earthquake magnitude by location sources:
In the code section our group chose to use boxplot to correlate the distribution of earthquake magnitudes in different regions. Our group chose to use boxplot in order to show the concentration trend of magnitude in different regions and to mark outlier anomalies. The image shows that the median earthquake magnitude is higher in the US region compared to other regions. The magnitude of the earthquakes in other regions is more concentrated in the range of 2.5-3.5. This is useful for future analysis and prediction in different regions.

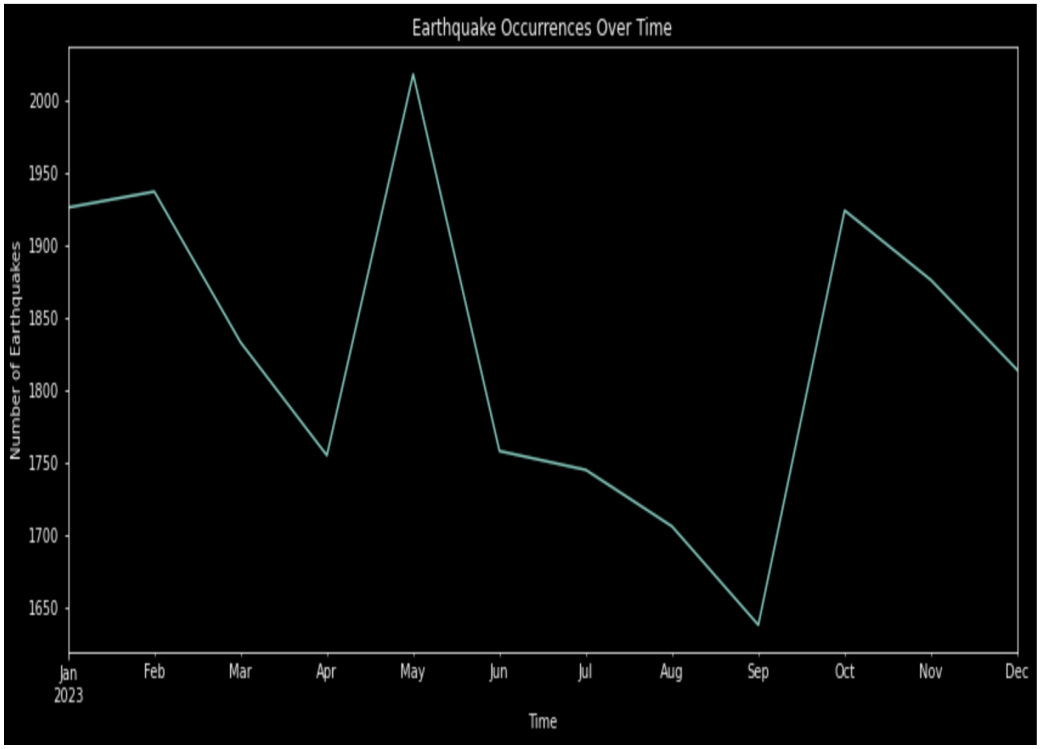


3.5:Violin Plot of Magnitude Distribution by Type

In the code section our group chose to use the violin plot in order to study the distribution of earthquakes in relation to the type and magnitude of the earthquakes. The widths and shapes provided in the violin plot allowed our group to understand the distribution of magnitudes and the concentration trends in the different types of earthquakes. The images obtained by our team show that the earthquake types are longer and have a larger magnitude range, with records from 2.5 to 6, but with a concentration in the 2.5-3 and 4.5. For mining explosions the magnitudes are wider and more concentrated around 3, but have some outlier events. Ice quakes are smaller and concentrated between 2.5 and 3, mainly around 2.5. landslides are similar to mining explosions in that the main magnitude is around 3, but the shapes are smoother. Volcanic eruption is similar to an ice quake, but the magnitude is concentrated around 4.5.

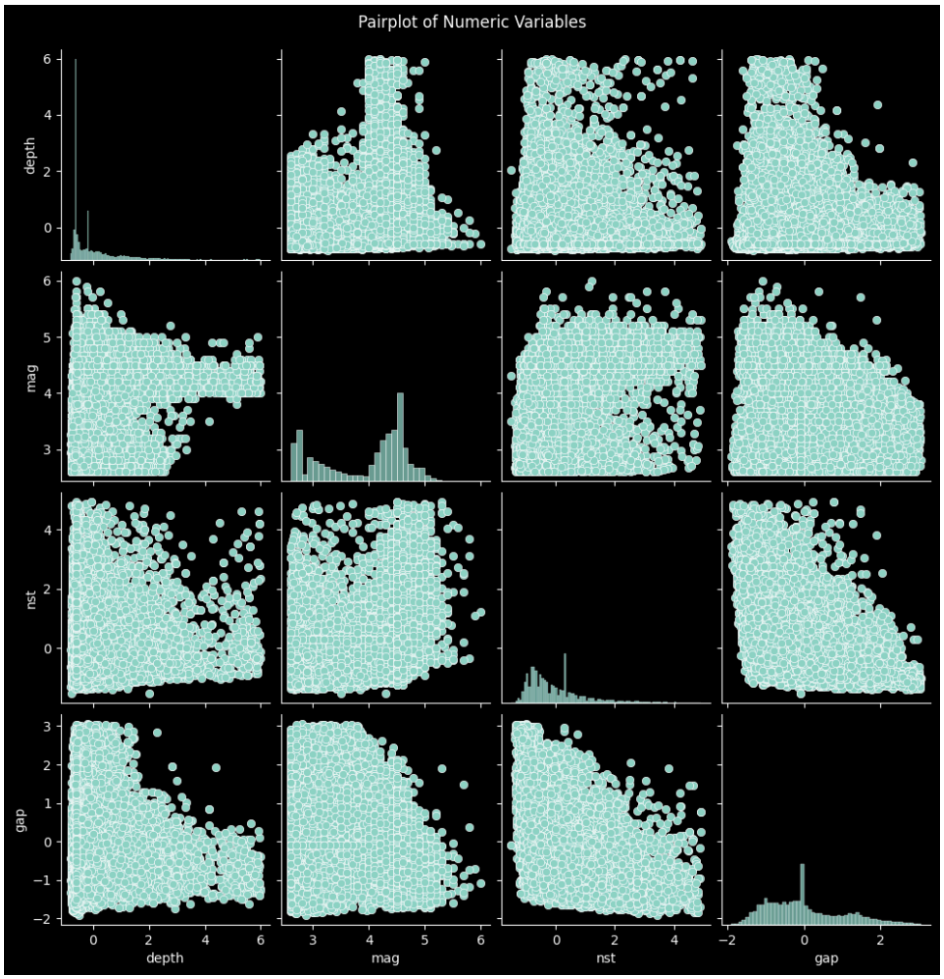


3.6: Time series plot of earthquake occurrence over time:
In the code section our team chose to use a line graph, which clearly expresses the change in the number of earthquakes in the whole time dimension. Based on the images obtained, our team has learned that the number of earthquakes peaks in May, but decreases rapidly before and after May, the line graphs are stable in January and February, and then increases but then decreases in October. The abnormal peak in May may be due to seasonal or geologic factors.



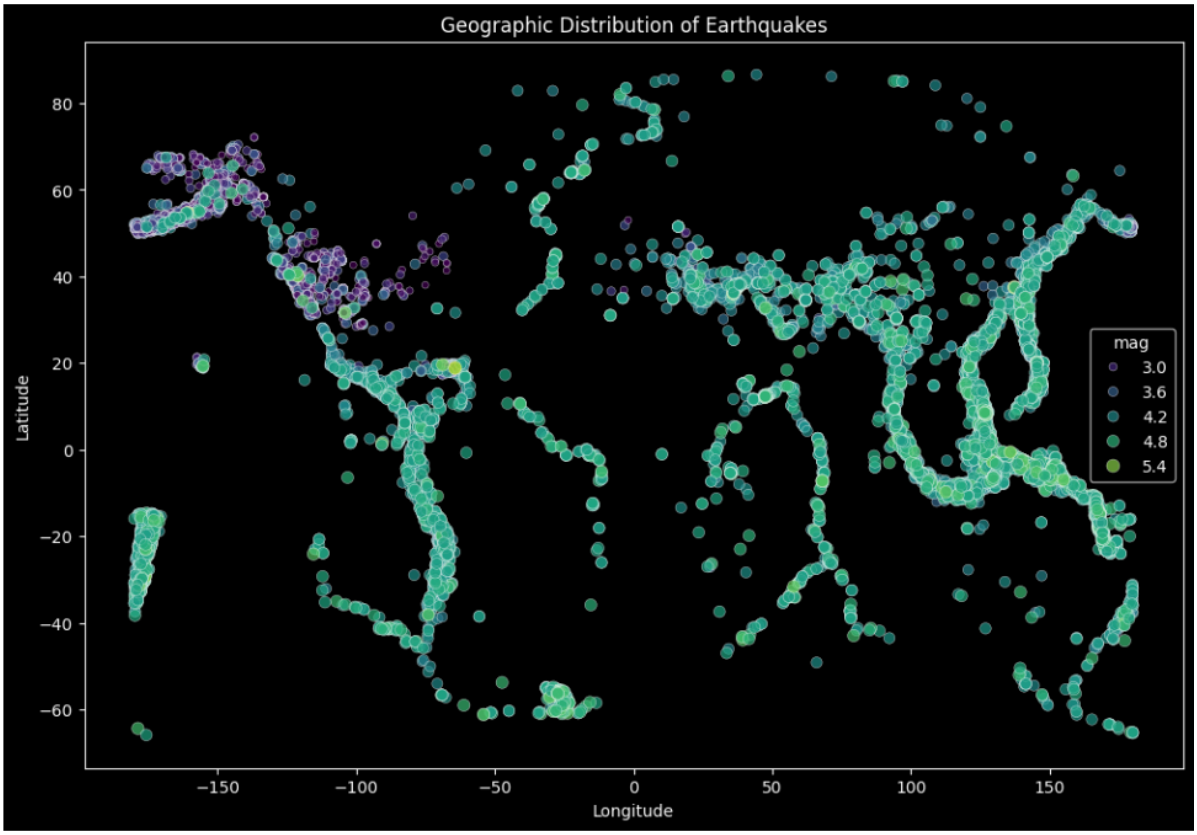
3.7: Pairplot

In the code section, our group chose to use pairplot, which can visually provide pairwise relationships between variables through scatterplots, and our group mainly chose the variables depth magnitude, nst, and gap in the dataset. Through the image representation, the overall distribution of points is rather chaotic and without any Our group initially thought that there might be no specific relationship between these pairs of variables.



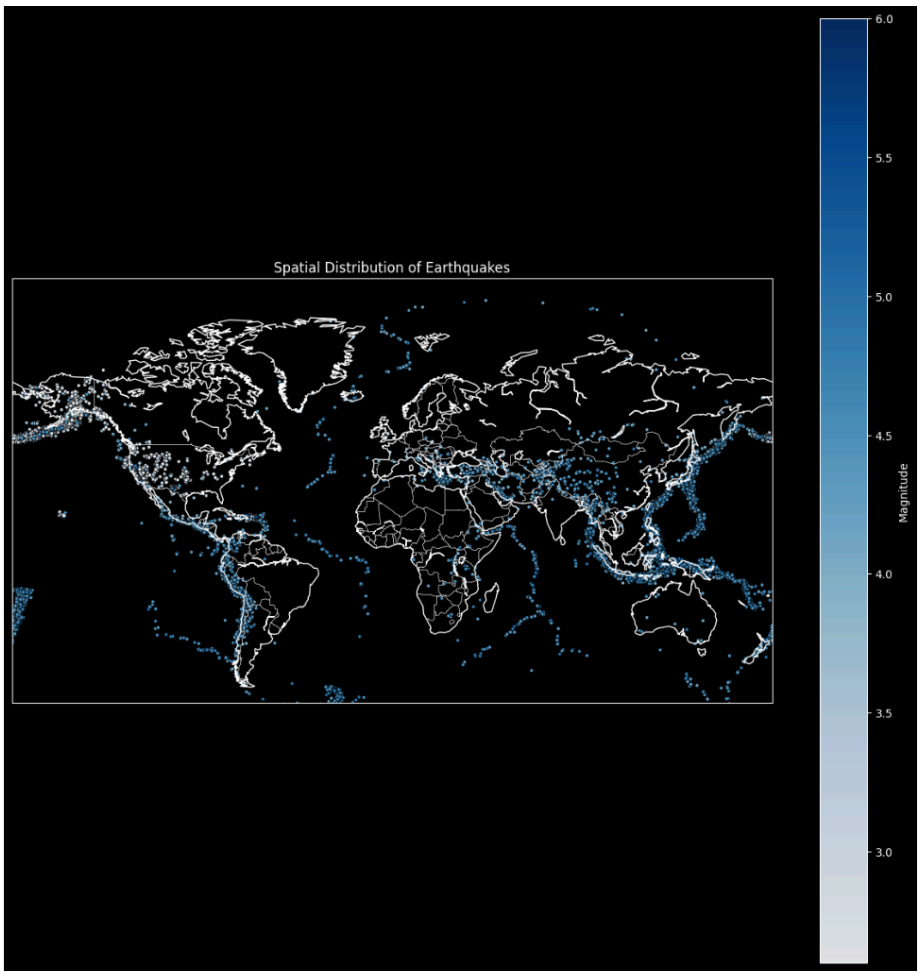
3.8: Geographic distribution with different magnitude

In the code section our group chose to use scatterplot, to obtain the magnitude and distribution of the earthquakes through scatterplot which mainly depends on the magnitude, longitude and latitude in the dataset. By observing the color and distribution of the points, our group obtained some patterns of magnitude and geographic location, in which the earthquakes in the -200 to -50 longtitude and the latitude from 20 to 60 are mainly 3.0, the rest of them are 4.2 and 4.8, and a small number of them are also 5.4 in magnitude. The image shows that 4.2 and 4.8 earthquakes are widely distributed, with a small number of 3.0 magnitude earthquakes concentrated in some areas.

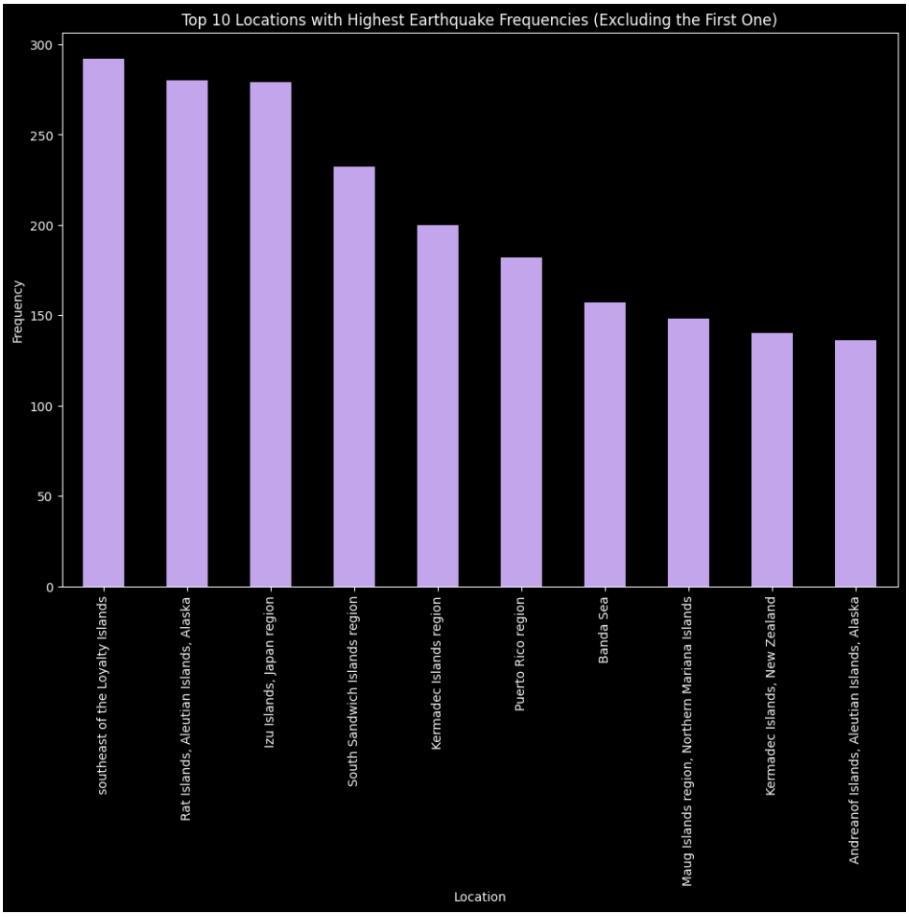


3.9: Spatial Distribution of Earthquakes

In the code section our group chose to use Basemap from mpl_toolkits to draw the relevant coastlines as well as the country boundaries to observe the relevant areas where earthquakes occur. The scatterplot allows for a better view of where earthquakes occur and is represented on the basemap. From our group's observations and the representation of the images, most of the earthquakes are centered on the country boundaries and many of them occur in the sea. Through the images our group analyzed the reason for the concentration of earthquakes in the areas in question, which may be mainly due to the geological plate boundaries and the collision of submarine plates.



3.10: Bar plot of the top N locations with the highest frequencies
In the code section our group chose to use a bar plot and show the top 10 regions in terms of the frequency of earthquakes, our group excluded the first point in order to compare the frequency of earthquakes in the other locations. By using the bar plot our group can get the relevant information about the locations. The top three locations do not vary much and the frequency of earthquakes in the bottom four locations also varies relatively smoothly. different groups of regions may have characteristics such as geographic location and climate that can be analyzed further.



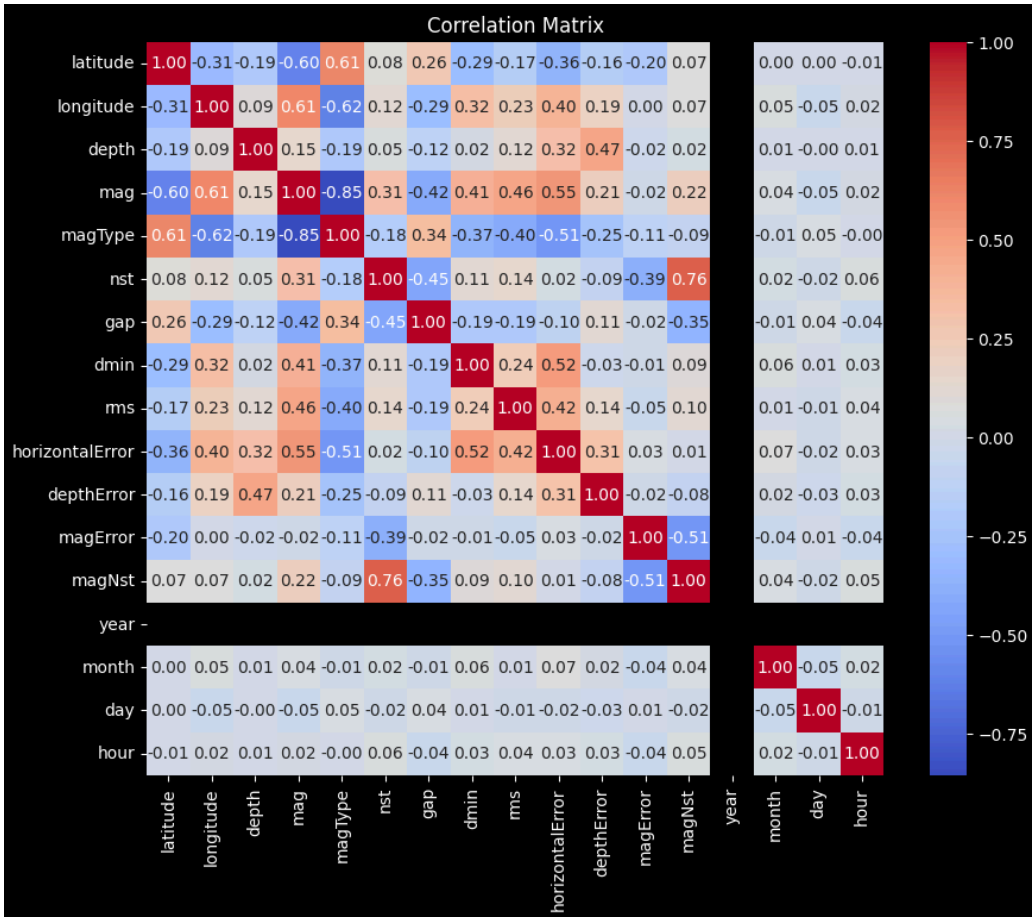
3.11: Statistic summary of earthquake magnitude

In the code section, our team chose to use info to show the statistics of the magnitude. From the graph, our team got the statistics of the number of magnitudes, the mean magnitude, the dispersion of the magnitude distribution, the minimum magnitude, the lower quartile, the median magnitude, the upper quartile, and the maximum magnitude. The mean value of the magnitude is 3.89 and the standard deviation value is 0.75 which is small indicating that the distribution of the magnitude is more concentrated. However, the maximum magnitude also shows that there are earthquakes with magnitude 6.0, but by analyzing the data, it can be concluded that similar earthquakes with larger magnitude are smaller. The main earthquakes are about 3.89 magnitude earthquakes.

```
Statistical Summary of Earthquake Magnitudes:
count    21930.000000
mean         3.891112
std          0.750497
min          2.600000
25%          3.100000
50%          4.200000
75%          4.500000
max          6.000000
Name: mag, dtype: float64
```

3.12: Correlation matrix among numerical variables

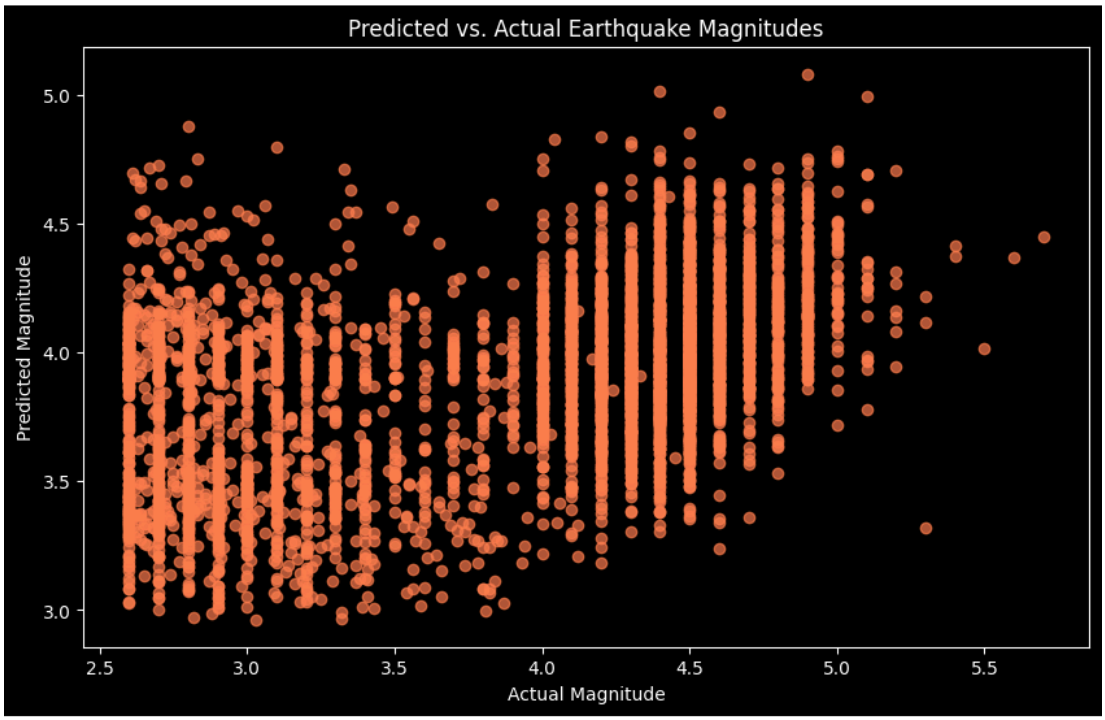
In the code section our group chose to use a heatmap to show the pairwise correlation of all the attributes, so that our group could more directly visualize the relationship between the variables through the numerical value results. In the image obtained by our group, most of the squares are in colors below 0, but there is important information such as the correlation between magtype and latitude, longitude and mag, MagNst (Number of seismic stations used to calculate the magnitude.) and MagNst (Number of seismic stations used to calculate the magnitude.). magnitude.) and Nst (Number of seismic stations that reported the earthquake). Where these paired variable relationships reflect a positive relationship, there is a correlation. This can inform subsequent advancement and analysis.



4. Advanced Analysis

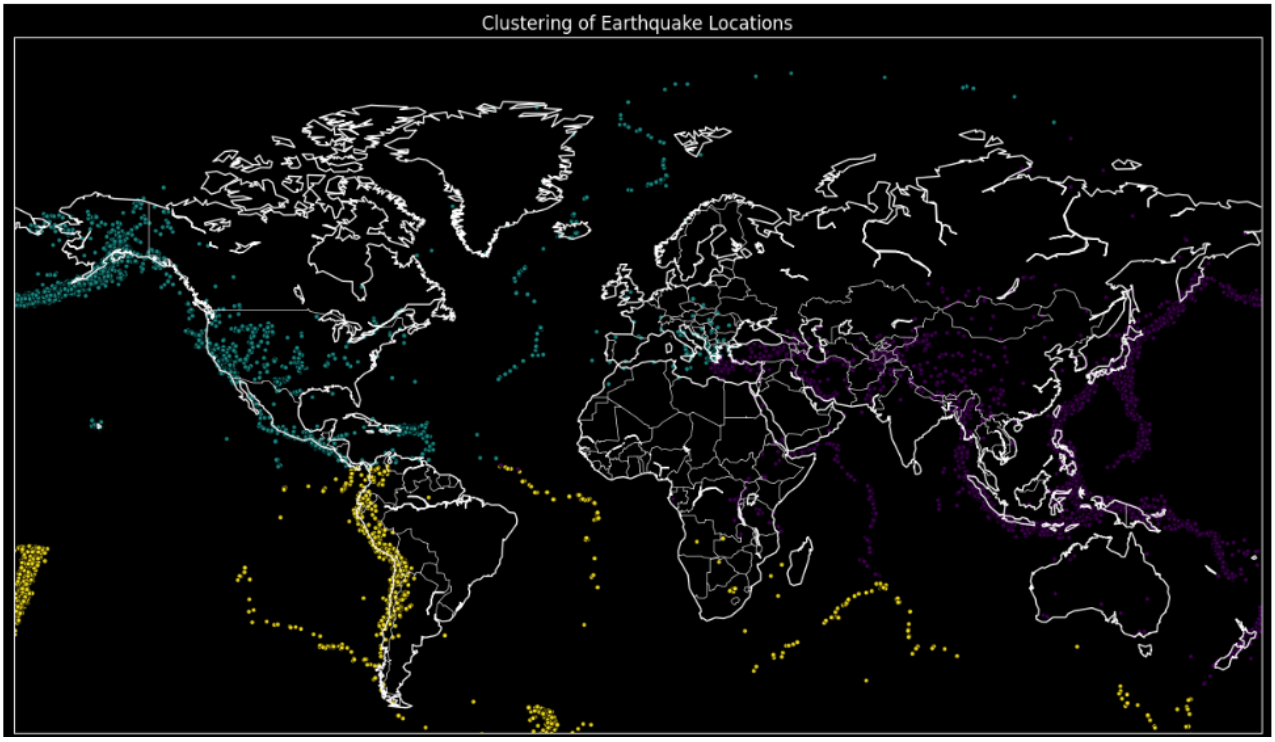
4.1: Predicting Earthquake magnitude

For Advanced analysis, our group first chose to use linear regression to try to make predictions about the magnitude of earthquakes. And we chose 'depth', 'nst', 'gap' as the features because our group thought that the magnitude is more related to the three. We first categorized the dataset into X_train, X_test, y_train, y_test and fitted the different datasets with linear regression. In this process the predicted results are compared with the actual results and at the end the Mean Squared Error is obtained as 0.45649. and the mean squared error is small and the prediction of the earthquake magnitude is relatively accurate. Linear regression provides relevant information and helps in aspects related to possible future prediction of earthquake magnitude.



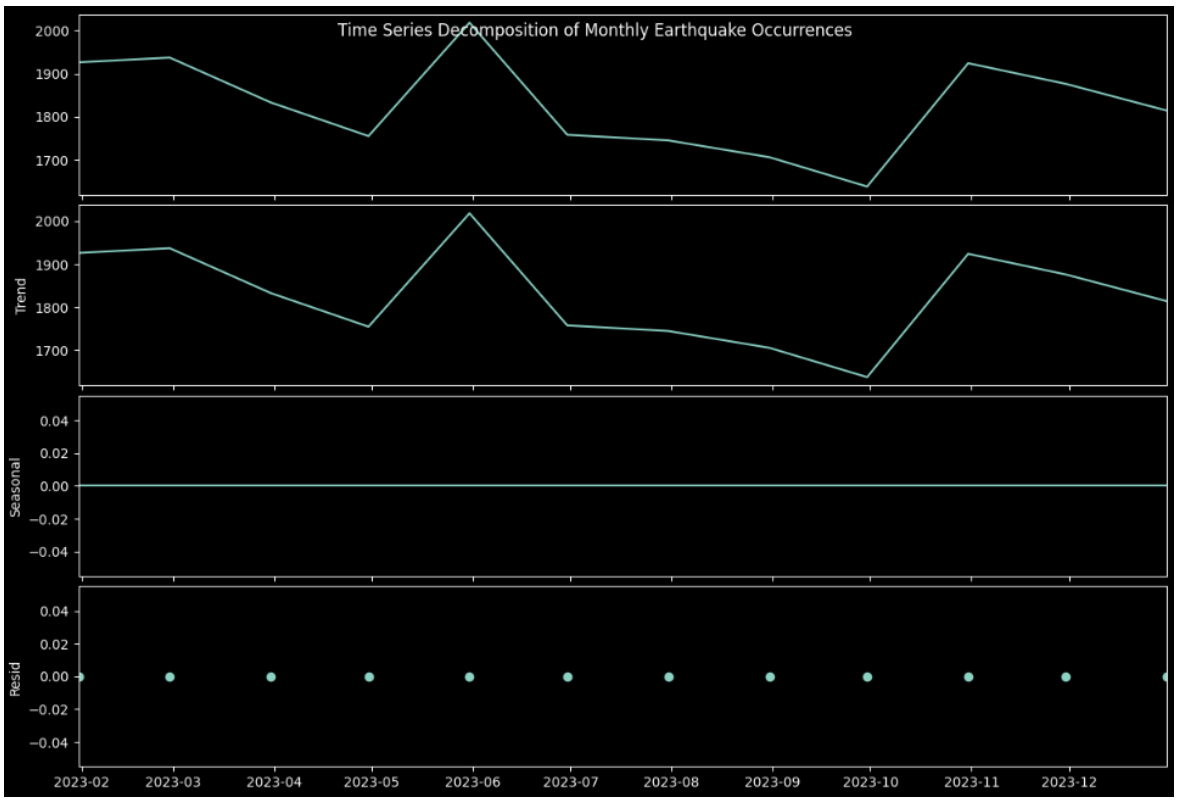
4.2: Clustering Earthquake locations

For the other advanced analysis, we decided to use clustering and used k-means clustering algorithm and basemap to further analyze the data and display it further on the map to make the data more visible and understandable. The latitude and longitude were first standardized using StandardScaler to ensure that the k-means clustering algorithm did not produce large differences in values. Based on the analysis of the data in the previous section and visualization of our parameters in k-means we decided to use n_clusters=3 and make predictions that are displayed on the basemap. On the image based clustering results, the clusters may be mainly based on the oceanic influence as well as the geological features. The oceanic and geoclimatic features are an important dividing line separating the western and northern parts. As well as possible geological features such as plate movements and tectonic relationships, the north and south sides are clustered separately. The clustering methods and predictions can help our group to have a deeper understanding of the distribution of seismicity and patterning in future applications.



4.3: Time Series Decomposition

For the advanced analysis of the last one, our group decided to use the seasonal_decompose function in statsmodels.api, using the initial idea of trying to count the number of earthquakes in all the months in the dataset and sum them. The seasonal_decompose function was then used to decompose the data into three attributes: trend, seasonal, and residual.The use of the additive model allowed us to observe the above trends, but as we got to know the function we realized that the number of datasets used was not large enough, which only contained data for the entire year 2023, and the seasonal_decompose function requires at least two cycles before it can be decomposed to comprehensive analysis. Therefore, our team used 1 as the value in the period of the time series only, resulting in an overfitting of the results to the original data without residuals, where the residuals are 0. In the future, if there is a larger dataset that can be used to improve the existing phenomena as well as the values.



5. Conclusion

In conclusion, the analysis and visualization of the earthquake dataset have provided valuable insights into the temporal and spatial patterns of seismic activity. Through the application of data cleaning, exploratory data analysis (EDA), clustering analysis, and time series decomposition, we have gained a comprehensive understanding of the dataset's characteristics.

Key Findings:

1. **Temporal Insights:**
    - The time series decomposition along with geographic scatter plot revealed trends and seasonal patterns in monthly earthquake occurrences. This understanding is crucial for predicting and preparing for future seismic activity.
2. **Spatial Distribution:**
    - Clustering analysis identified distinct spatial clusters, revealing regions with similar seismic characteristics. Hotspots and areas with higher earthquake frequencies were identified, contributing to risk assessment.
3. **Magnitude Analysis:**
    - Visualization of earthquake magnitudes provided insights into the distribution and intensity of seismic events. Understanding magnitude trends is vital for assessing potential impact.
4. **Cluster Analysis:**
    - The application of KMeans clustering helped identify spatial groups of earthquake locations, aiding in the identification of geological features and tectonic processes.
5. **Time Series Decomposition:**
    - Decomposing the time series data highlighted trends, seasonality, and residual components. This decomposition is foundational for modeling and forecasting future earthquake occurrences.

Thus, visualization and analysis is crucial when it comes to earthquake analysis and prediction.