A Novel Approach to Analyze and Predict Earthquake Patterns

Pranit Sawant[1]
Msc .Data Science
Vellore Institute of Technology,
Vellore-632014,Tamil Nadu,India
sawant.pranit2022@vitstudent.ac.in

Vashishth Pathak[2]
Msc .Data Science
Vellore Institute of Technology,
Vellore-632014,Tamil Nadu,India
pathak.vashishth2022@vitstudent.ac.in

Sounak Sarkar[3]
Msc .Data Science
Vellore Institute of Technology,
Vellore-632014,Tamil Nadu,India
sounak.sarkar2022@vitstudent.ac.in

ABSTRACT

This project focuses on conducting comprehensive data analysis of earthquake events to understand their characteristics and spatial distribution across different regions. Earthquakes are natural disasters that can have devastating consequences, and gaining insights into their attributes is vital for effective disaster management and preparedness. The study considers various attributes such as magnitude, date and time, intensity, alert level, tsunami occurrence, significance, seismic stations used, distance to the nearest station, azimuthal gap, magnitude calculation method, depth, geographic coordinates, location, continent, and affected country. This exploration will reveal the distribution of earthquake magnitudes, intensity levels, alert frequencies, and geographical patterns, thereby identifying notable trends and patterns within the data. Statistical techniques will be employed to explore relationships between attributes. This analysis will enable the identification of clustering or hotspot areas, contributing to a better understanding of seismic activity patterns within specific regions. In conclusion, this project aims to provide comprehensive insights into the characteristics and spatial distribution of earthquakes across different regions. By analyzing a diverse range of attributes, valuable information can be derived concerning earthquake magnitudes, intensities, geographical patterns, and influencing factors. The findings will contribute to improved disaster management strategies, enhancing preparedness and response measures in earthquake-prone areas.

Keywords-Earthquakes, earthquake-prone areas, tsunami, significance, signature

I.INTRODUCTION

Understanding the patterns and characteristics of earthquakes is crucial for effective disaster management and preparedness. Analyzing earthquake data provides valuable insights into seismic events' frequency, magnitude, and geographical distribution. This project aims to examine a comprehensive dataset encompassing various attributes related to earthquakes. These attributes include magnitude, date and time, intensity, alert level, tsunami occurrence, significance, seismic stations used, distance to the nearest station, azimuthal gap, magnitude calculation method, depth, geographic coordinates, location, continent, and affected country. By conducting a thorough analysis of this dataset, the study seeks to uncover trends, correlations, and spatial relationships, ultimately contributing to an improved understanding of earthquake occurrences.

Date and time information allows for temporal analysis, revealing patterns and seasonality in earthquake occurrences. The intensity and alert level indicate the impact and urgency of the response required. The occurrence of the tsunami is represented by a binary variable, signifies the potential for secondary hazards. The significance attribute quantifies the overall impact of an earthquake, considering factors such as magnitude, estimated intensity, felt reports, and estimated impact. The number of seismic stations used (not) and the azimuthal gap measure the reliability of earthquake location determination. The magnitude calculation method (magType) helps understand the technique employed to determine earthquake strength. Depth provides information about the subsurface rupture zone, while geographic coordinates enable precise location mapping. The location, continent and affected country attributes provide valuable contextual information regarding the geographical distribution of earthquakes.

By analyzing these attributes, this project aims to uncover meaningful insights regarding the characteristics and distribution of earthquakes. This knowledge can inform disaster management strategies, aid in preparedness measures, and facilitate targeted response efforts in earthquake-prone regions.

II.LITERATURE REVIEW

Earthquakes are natural phenomena that occur globally and have significant impacts on human lives, infrastructure, and the environment. Understanding earthquake patterns and their characteristics is crucial for risk assessment, disaster preparedness, and engineering design. This literature review aims to explore the existing research on analyzing earthquake patterns around the world, highlighting key findings and methodologies employed in the field.

Global Earthquake Databases: To investigate earthquake patterns on a global scale, researchers rely on comprehensive earthquake databases that provide information on seismic events worldwide. Databases such as the Global Centroid Moment Tensor (GCMT) catalog and the International Seismological Centre (ISC) catalog offer a wealth of data for analyzing earthquake occurrences. These databases include information on earthquake locations, magnitudes, focal mechanisms, and other relevant parameters.

Spatial and Temporal Patterns: Researchers have explored various spatial and temporal patterns of earthquakes to identify regions of high seismic activity and understand the underlying geodynamic processes. For instance, Gutenberg and Richter (1944) introduced the Gutenberg-Richter law, which describes the logarithmic relationship between earthquake magnitudes and their frequencies. This law has been widely used to analyze the frequency-magnitude distribution of earthquakes globally and regionally.

Furthermore, studies have examined seismicity rates along tectonic plate boundaries, subduction zones, and major fault systems. For example, Bird (2003) analyzed seismicity along subduction zones worldwide and identified regions with higher seismic activity and greater potential for large earthquakes. Other studies have focused on investigating the clustering of earthquakes and the occurrence of aftershocks following mainshocks, providing insights into the spatiotemporal behavior of seismic events.

Statistical and Machine Learning Approaches: To analyze earthquake patterns, researchers have employed statistical and machine learning methods to extract meaningful information from earthquake catalogs. Statistical techniques such as cluster analysis, regression analysis, and spectral analysis have been used to identify groups of earthquakes with similar characteristics, detect temporal variations, and investigate correlations between seismic events.

In recent years, machine learning algorithms have gained prominence in earthquake pattern analysis. These algorithms can uncover complex relationships and make predictions based on large volumes of data. For example, Deep Learning techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been applied to earthquake data to detect earthquake signals, classify seismic events, and predict future occurrences (Yoon et al., 2019; Ross et al., 2020). These approaches have shown promise in improving the accuracy of earthquake pattern analysis and prediction.

III.METHODOLOGY

A. Data Collection:

We collected earthquake data from Kaggle. Our data had 19 columns and 782 rows. We have data from all over the world having country names, latitudes, longitude, etc.

Out[3]:

	title	magnitude	date_time	cdi	mmi	alert	tsunami	sig	net	nst	dmin	gap	magType	depth	latitude	longitude	location	co
0	M 7.0 - 18 km SW of Malango, Solomon Islands	7.0	22-11-2022 2.03	8	7	green	1	768	us	117	0.509000	17.0	mww	14.000	-9.7963	159.5960	Malango, Solomon Islands	C
1	M 6.9 - 204 km SW of Bengkulu, Indonesia	6.9	18-11-2022 13.37	4	4	green	0	735	us	99	2.229000	34.0	mww	25.000	-4.9559	100.7380	Bengkulu, Indonesia	
2	M 7.0 -	7.0	12-11-2022 7.09	3	3	green	1	755	us	147	3.125000	18.0	mww	579.000	-20.0508	-178.3460	NaN	C
3	M 7.3 - 205 km ESE of Neiafu, Tonga	7.3	11-11-2022 10.48	5	5	green	1	833	us	149	1.865000	21.0	mww	37.000	-19.2918	-172.1290	Neiafu, Tonga	

B. Exploratory Data Analysis:

To explore the data and to find out patterns we performed exploratory data analysis. By doing EDA we got a clear understanding of the data.

We get information about data using the info () command in pandas.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 782 entries, 0 to 781
Data columns (total 19 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   title       782 non-null    object
 1   magnitude   782 non-null    float64
 2   date_time   782 non-null    object
 3   cdi         782 non-null    int64
 4   mmi         782 non-null    int64
 5   alert       415 non-null    object
 6   tsunami     782 non-null    int64
 7   sig         782 non-null    int64
 8   net         782 non-null    object
 9   nst         782 non-null    int64
 10  dmin        782 non-null    float64
 11  gap         782 non-null    float64
 12  magType     782 non-null    object
 13  depth       782 non-null    float64
 14  latitude    782 non-null    float64
 15  longitude   782 non-null    float64
 16  location    777 non-null    object
 17  continent   206 non-null    object
 18  country     484 non-null    object
dtypes: float64(6), int64(5), object(8)
memory usage: 116.2+ KB
```

a) Dealing with null values.
We had some null values in the columns like, location, continent, and country. We dealt with these missing values differently. For instance, In the column, if we have less than 5% of observations having missing values then we simply delete those rows. Continent columns we deleted as we already have location as well as latitude and longitude columns from those, we can easily determine continents. We replace null values in the alert column with the median as the values in the columns are of object types. Also, from the title column we created time and distance from the location column to check whether there is any relation between these. Finally, we left with the following data having zero null values.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 777 entries, 0 to 776
Data columns (total 20 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Unnamed: 0          777 non-null    int64
 1   locearthquake       777 non-null    object
 2   magnitude           777 non-null    float64
 3   date_time           777 non-null    object
 4   cdi                 777 non-null    int64
 5   mmi                 777 non-null    int64
 6   alert               777 non-null    int64
 7   tsunami             777 non-null    int64
 8   sig                 777 non-null    int64
 9   net                 777 non-null    object
 10  nst                 777 non-null    int64
 11  dmin                777 non-null    float64
 12  gap                 777 non-null    float64
 13  magType             777 non-null    object
 14  depth               777 non-null    float64
 15  latitude            777 non-null    float64
 16  longitude           777 non-null    float64
 17  location            777 non-null    object
 18  time                777 non-null    object
 19  dist_from_location  777 non-null    int64
dtypes: float64(6), int64(8), object(6)
memory usage: 121.5+ KB
```

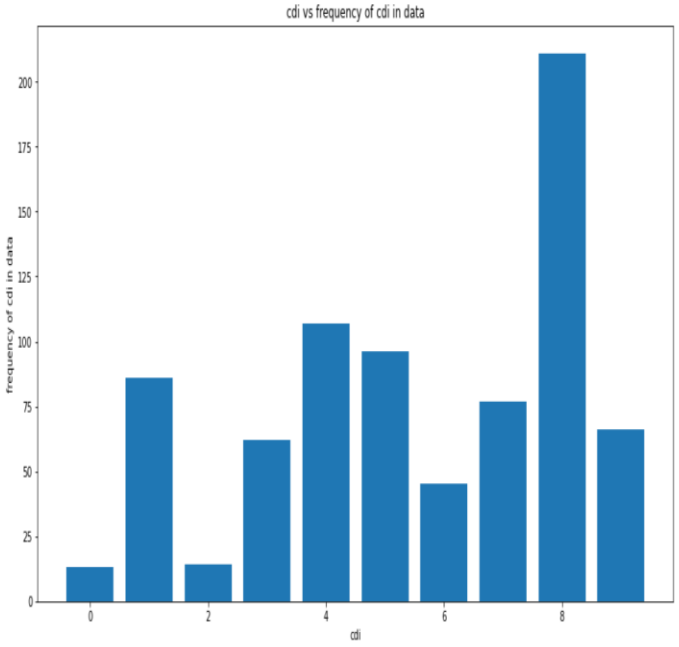
After dealing with missing values, we performed the following exploratory analysis to deal with the data.

b) Descriptive statistics:

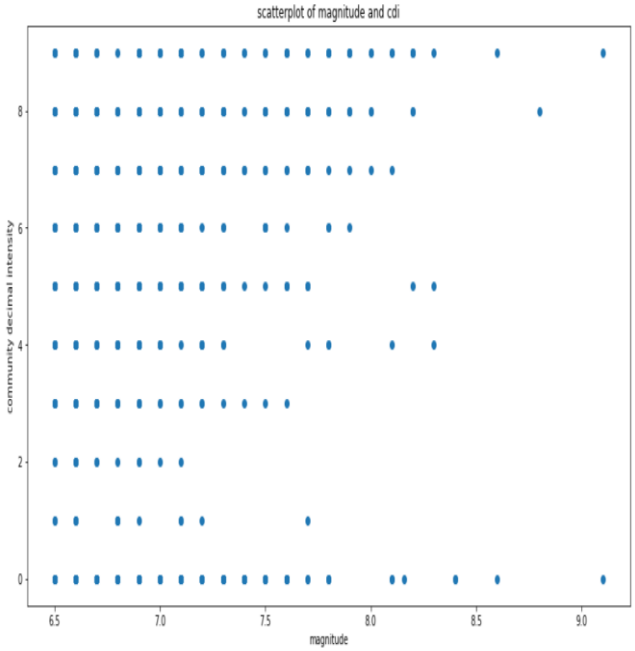
	count	mean	min	25%	50%	75%	max	std
magnitude	777.0	6.941647	6.5	6.6	6.8	7.1	9.1	0.446767
cdi	777.0	4.342342	0.0	0.0	5.0	7.0	9.0	3.173045
mmi	777.0	5.978121	1.0	5.0	6.0	7.0	9.0	1.452259
alert	777.0	0.275418	0.0	0.0	0.0	0.0	3.0	0.814701
tsunami	777.0	0.384813	0.0	0.0	0.0	1.0	1.0	0.486865
sig	777.0	870.804376	650.0	691.0	754.0	910.0	2910.0	323.315975
nst	777.0	231.211068	0.0	0.0	142.0	445.0	934.0	250.648184
dmin	777.0	1.312927	0.0	0.0	0.0	1.822	17.654	2.215283
gap	777.0	25.014788	0.0	14.2	20.0	30.0	239.0	24.287392
depth	777.0	74.232314	2.7	14.0	26.0	48.0	670.81	134.545639
latitude	777.0	3.668359	-61.8484	-14.344	-2.486	24.696	71.6312	27.235589
longitude	777.0	52.790185	-179.968	-71.649	108.174	148.887	179.662	117.503708
dist_from_location	777.0	94.407979	1.0	41.0	90.0	118.0	2011.0	114.509756
date	777	2012-09-23 03:14:35.67567568	2001-01-01 00:00:00	2007-09-26 00:00:00	2013-05-24 00:00:00	2017-08-08 00:00:00	2022-11-22 00:00:00	NaN

The earthquake data consists of 777 observations with various features such as magnitude, cdi, mmi, alert, tsunami, sig, not, dmin, gap, depth, latitude, longitude, and dist_from_location. The mean magnitude is approximately 6.94, indicating a significant level of seismic activity. The standard deviation of magnitude is 0.45, suggesting a relatively narrow range of variation. The target variable, sig, has a mean value of 0.38, indicating a moderate level of significance. The data exhibit a wide range of values, with some features showing high variability.

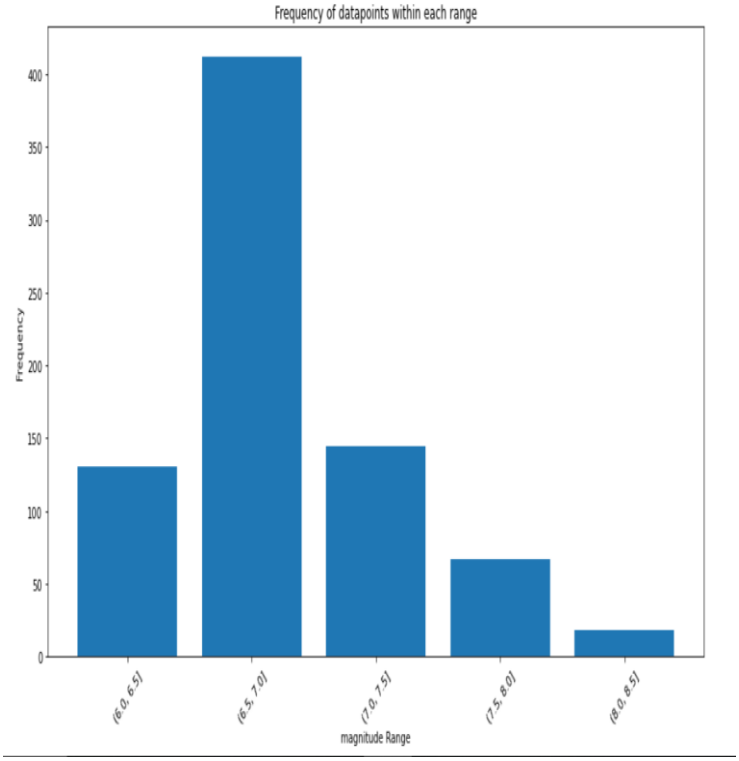
c) Data Visualization:



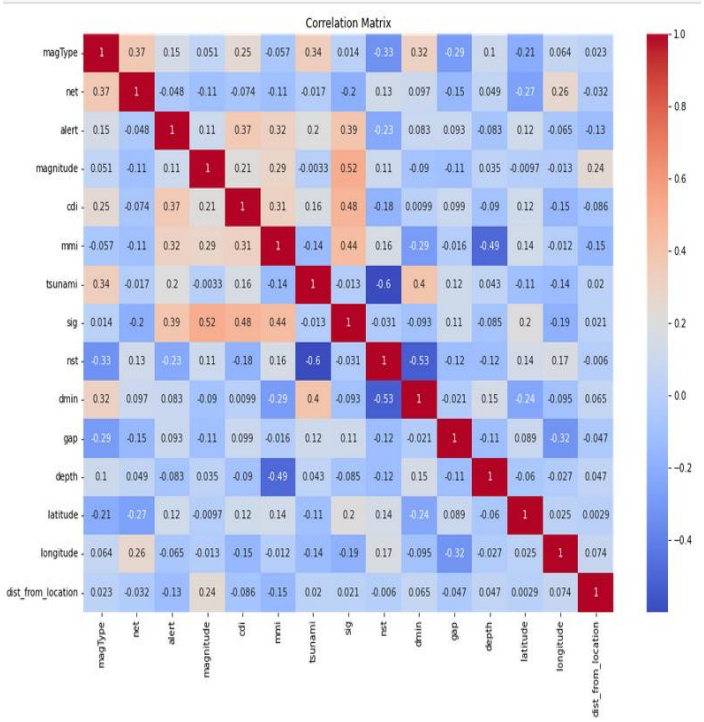
The cdi column in the earthquake data represents the Community Internet Intensity Map. From the descriptive statistics, we can observe that the cdi values range from 0 to 9, with a mean value of approximately 4.34. The frequency distribution of cdi in the data suggests that there is a diverse range of intensities reported by the community.



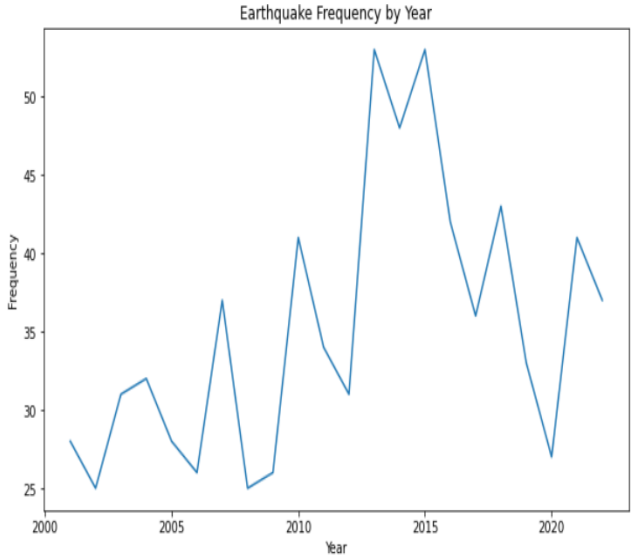
The scatter plot of magnitude and cdi in the earthquake data reveals a positive relationship between these two variables. As the magnitude increases, the cdi tends to be higher, indicating a stronger perceived intensity by the community. This suggests that there is a correlation between the seismic magnitude and the community's perception of the earthquake's intensity.



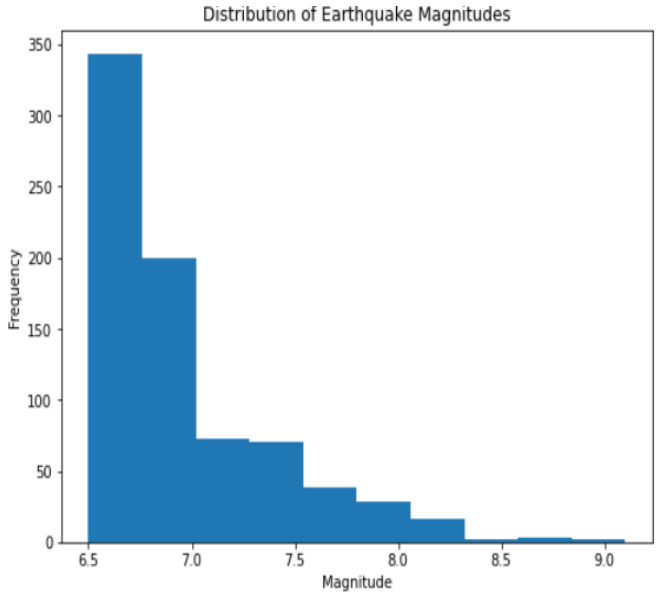
The frequency distribution analysis of the earthquake data reveals that the data points are distributed across various ranges. Most data points fall within the range of 6.5 to 7.1 in magnitude. There are relatively fewer data points with higher magnitudes. This distribution provides insights into the occurrence of earthquakes within different ranges and can help identify areas of higher seismic activity.



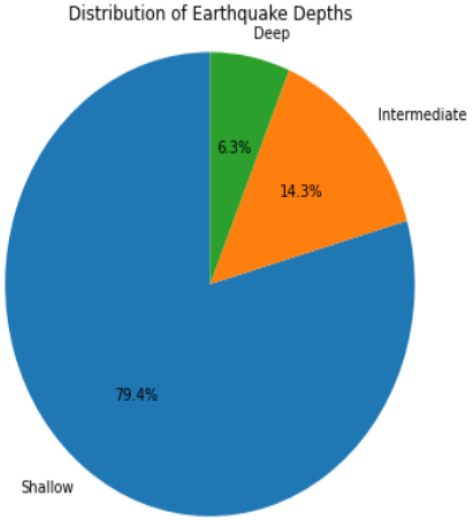
The correlation matrix of the earthquake data indicates the relationships between different variables. The magnitude shows a positive correlation with cdi and mmi, suggesting that higher magnitudes are associated with increased community and shaking intensity. There is a weak positive correlation between magnitude and longitude. Other variables show minimal correlations with each other. These findings provide insights into the interdependencies among the variables and can guide further analysis and modeling efforts in understanding earthquake characteristics and their effects.



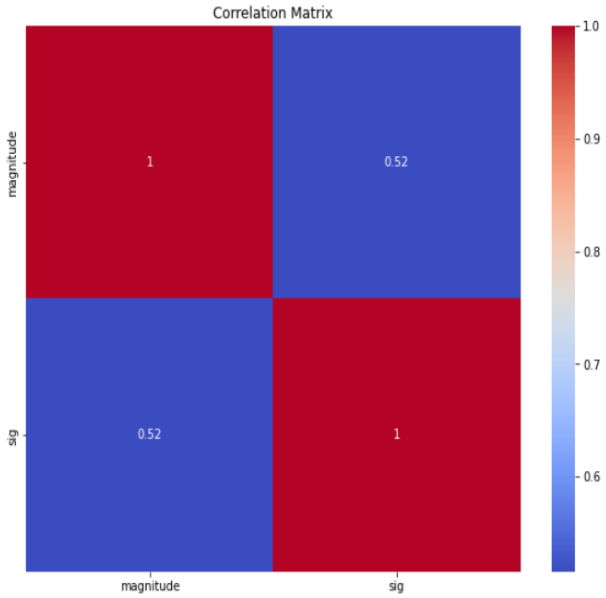
Analyzing the frequency of earthquakes by year in the data reveals that 2015 has the highest number of earthquakes. In contrast, the period between 2000 and 2005 exhibits the lowest number of earthquakes. This highlights the temporal variability in seismic activity, with certain years experiencing higher levels of seismicity compared to others. Further investigation can be conducted to understand the underlying factors contributing to the variations in earthquake occurrence during different periods.



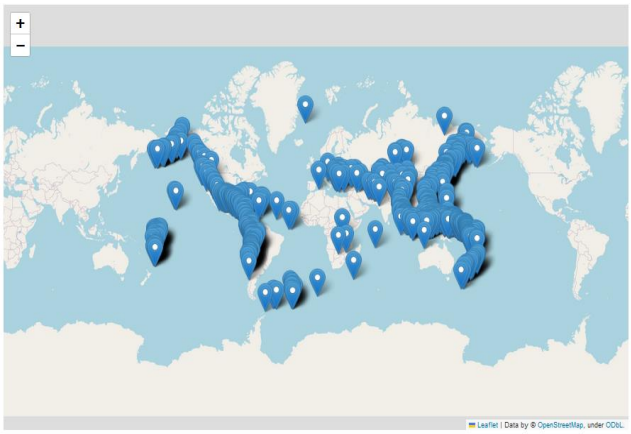
The distribution of earthquake magnitudes in the data indicates a range of seismic events. Most earthquakes fall within the range of 6.5 to 7.1 magnitudes, with a mean magnitude of approximately 6.94. The distribution is slightly right-skewed, with a higher concentration of moderate to high magnitudes. This suggests that the data contains a mix of moderate and significant seismic events. Understanding the distribution of earthquake magnitudes is crucial for assessing seismic hazards and implementing appropriate risk mitigation strategies.



The distribution of earthquake depths in the data reveals a wide range of values. The depths range from 0 to 239 units, with a mean depth of approximately 25.01 units. The distribution is positively skewed, with a higher concentration of earthquakes occurring at shallower depths. Understanding the distribution of earthquake depths is important for characterizing the subsurface processes associated with seismic activity and assessing the potential impacts on structures and infrastructure.



The correlation matrix between magnitude and sig reveals a positive correlation between these two variables. This suggests that as an earthquake's magnitude increases, the earthquake's significance or impact also tends to increase. This positive correlation implies that larger earthquakes are more likely to have a greater impact. Understanding the relationship between magnitude and sig is crucial for assessing the potential consequences of seismic events and informing disaster preparedness and response measures.



The above graph brings out the locations which have been affected by earthquakes from 2001 to 2022.

Seems like Most of the earthquakes have been happening in the faults under the Ocean floor and also along the shoreline of the continents.

IV.RESULTS

Some important findings in the data are as below:

The scatter plot of magnitude and cdi shows a positive relationship. Frequency distribution analysis reveals the distribution of data points across different magnitudes and intensities.

The correlation matrix indicates relationships between variables, with magnitude showing positive correlations with cdi and mmi. Analysis of earthquake frequency by year highlights temporal variability, with some years experiencing higher seismicity than others. The distribution of earthquake magnitudes shows a concentration in the range of 6.5 to 7.1, with a slightly right-skewed distribution. The distribution of earthquake depths indicates a wide range of values, with a higher concentration of shallow earthquakes. A positive correlation between magnitude and sig suggests that larger earthquakes tend to have a greater impact.

V.CONCLUSIONS

We have successfully performed visualizations on the data and find out some key observations. Key observations from the analysis have been identified. The next step is to build a model to determine earthquake significance.

VI.REFERENCES:

[1] "Introduction to Seismology" by Peter M. Shearer: This book offers a comprehensive introduction to seismology, covering topics such as earthquake sources, wave propagation, and seismic hazard analysis. It provides a solid foundation for understanding earthquake analysis techniques.

[2] "Earthquake Engineering: From Engineering Seismology to Performance-Based Engineering" by Yousef Bozorgnia and Vitelmo V. Bertero: This book focuses on the engineering aspects of earthquake analysis, including seismic hazard assessment, structural dynamics, and design considerations. It is a valuable resource for earthquake engineers and researchers.

[3] "Earthquake Ground Motion and Its Effects on Structures" edited by Pierre-Yves Bard: This book delves into the analysis of earthquake ground motion and its effects on structures. It covers topics such as ground motion prediction, soil-structure interaction, and seismic response analysis. It is particularly useful for those involved in structural engineering and design.

[4] "Seismic Design of Building Structures" by Michael R. Lindeburg and Kurt M. McMullin: This reference is specifically geared towards the seismic design of building structures. It covers seismic design principles, code requirements, and analysis methods for various types of structures. It is often used as a study guide for professional engineering exams' seismic principles and practices section.

[5] The publications and technical reports from reputable organizations such as the U.S. Geological Survey (USGS), the National Earthquake Hazards Reduction Program (NEHRP), and the International Building Code (IBC) are also excellent sources of information. These organizations provide valuable data, guidelines, and research findings related to earthquake analysis and seismic design.

[6] "Fundamentals of Earthquake Engineering" by Amr S. Elnashai and Luigi Di Sarno: This book provides a comprehensive introduction to earthquake engineering, covering topics such as seismic hazard analysis,

structural dynamics, and seismic design principles. It includes numerous examples and case studies.

[7] "Seismic Design and Assessment of Bridges: Inelastic Methods of Analysis and Case Studies" by M. J. N. Priestley, F. Seible, and G. M. Calvi: Focused specifically on bridge engineering, this book covers the analysis and design of bridges subjected to earthquakes. It explores inelastic methods of analysis and includes case studies for practical understanding.

[8] "Earthquake Resistant Design of Structures" by Shashikant K. Duggal: This book provides a practical approach to earthquake-resistant design, covering various aspects such as seismic hazard assessment, structural dynamics, and design principles for different types of structures.

[9] "Strong Motion Instrumentation for Civil Engineering Structures" edited by Miguel A. Pando: This reference book focuses on strong motion instrumentation and its application in earthquake engineering. It covers topics such as accelerometer selection, data acquisition, and analysis techniques for capturing and analyzing strong ground motion records.

[10] "Principles of Soil Dynamics" by Braja M. Das and G.V. Ramana: This book focuses on the dynamic behavior of soils during earthquakes. It covers topics such as seismic waves, soil response analysis, liquefaction, and soil-structure interaction.

[11] Technical reports and publications from organizations like the Pacific Earthquake Engineering Research Center (PEER), Earthquake Engineering Research Institute (EERI), and International Association of Earthquake Engineering (IAEE) are valuable resources for the latest research, case studies, and advancements in earthquake analysis.