Analysis Of Global Earthquakes in 2023

Dhanya Maddi, Naga Sai Dhanya Veerepalli, Sriya Reddy Karegoud
Data Analytics Engineering
George Mason University

Abstract— Seismic activity describes the occurrence of waves or vibrations in the crust of the Earth, which are frequently brought on by tectonic plate movement, volcanic eruptions, or events that are triggered by humans. Earthquakes, volcanic eruptions, and other geological occurrences are examples of how these vibrations can appear. In this paper we explore the dataset “Earthquakes Global 2023” to certain patterns and relations between the different seismic activities using different data science steps.
The dataset contains 22 features related to earthquake events worldwide. After preprocessing, exploratory analysis revealed insights into depth, magnitude, and spatial distribution. Machine learning algorithms like decision trees, random forests, clustering, regression, and time series analysis were applied. Key findings indicate magnitude, depth, and location are crucial features for determining seismic activity. The Pacific Ocean region, Japan, Indonesia, and South America had the highest earthquake prevalence. A relationship exists between earthquake type, magnitude, and depth. Winter and spring months tend to have more earthquakes.The analysis provides descriptive insights into 2023's global earthquake patterns, informing disaster preparedness and mitigation strategies. Future work could integrate additional data sources and advanced predictive modeling.

Keywords: Decision Tree,Random forest,Clustering,Regression

I.Introduction

Earthquakes, one of nature's most powerful and unpredictable phenomena, have been a constant source of concern and intrigue throughout human history. These seismic events, caused by the sudden release of energy within the Earth's crust, can have devastating consequences, ranging from the loss of human life and widespread destruction of infrastructure to long-lasting environmental impacts. In 2023, the global seismic landscape continued to evolve, with

numerous earthquake occurrences recorded worldwide, each event serving as a stark reminder of the Earth's dynamic and ever-changing nature.
This comprehensive study aims to conduct an in-depth investigation into the global earthquake data from 2023, leveraging advanced data mining and machine learning techniques. By harnessing the power of these analytical approaches, the research endeavors to uncover meaningful insights, identify critical factors contributing to seismic activity, and explore potential relationships between various earthquake attributes.

The primary objective of this analysis is to shed light on the intricate patterns and underlying dynamics that govern the occurrence and behavior of earthquakes. Through a rigorous data preprocessing approach, exploratory data analysis, and the application of cutting-edge machine learning algorithms, this study seeks to unravel the complexities of seismic phenomena, providing a holistic understanding of the global earthquake landscape in 2023.

By delving into the rich dataset, encompassing a wide range of features such as time, location, magnitude, depth, and various seismic parameters, this research aims to uncover geographic hotspots, temporal distributions, and correlations between variables. These findings will not only contribute to the advancement of scientific knowledge but also hold significant implications for disaster preparedness, risk assessment, and the development of effective mitigation strategies.

Furthermore, the analysis will employ a diverse array of machine learning techniques, including decision trees, random forests, clustering algorithms, regression models, and time series analysis. These powerful methods will enable the identification of influential features, the detection of spatial and temporal patterns, the modeling of relationships between earthquake characteristics, and the forecasting of future seismic events.

By leveraging the insights gained from this study, authorities, organizations, and communities can prioritize resources, implement early warning systems, and develop robust disaster management plans. Ultimately, this research aims to contribute to the global effort in understanding and mitigating the devastating impacts of earthquakes, fostering a more resilient and prepared society in the face of these natural hazards.

II.Dataset Information

The dataset used in this analysis is sourced from Kaggle and contains comprehensive information on global earthquakes that occurred in 2023. The dataset, titled "Earthquakes 2023 - Global," is a compilation of seismic events recorded worldwide throughout the year.
The dataset consists of 22 features, providing a rich set of attributes for each earthquake event.

These features include:

Time: The date and time of the earthquake occurrence.
Latitude and Longitude: The geographic coordinates of the earthquake's epicenter.
Magnitude: A measure of the earthquake's size or strength, typically on the Richter scale.
Depth: The depth at which the earthquake originated, measured in kilometers.
RMS: The root mean square of the seismic wave amplitudes, a measure of the earthquake's intensity.
Horizontal Error and Depth Error: The estimated errors in the horizontal and depth measurements, respectively.
Additional features: Gap, Source, and various other parameters related to the seismic event.

The dataset covers a diverse range of earthquake magnitudes, depths, and locations, providing a comprehensive representation of the global seismic activity in 2023. It serves as a valuable resource for analyzing spatial and temporal patterns, investigating relationships between earthquake attributes, and developing predictive models for future seismic events.

III.Literature Review

Literature Review 1:

[7]This study investigates methods to estimate the magnitude of completeness (Mc) for earthquake catalogs, which is crucial for accurate seismic hazard assessment. The authors evaluate different techniques, including the maximum curvature method, the entire magnitude range method, and the Goodness-of-Fit method. They provide a comprehensive review of the strengths and limitations of each approach and offer recommendations for selecting the most appropriate method based on the characteristics of the earthquake catalog.

Literature Review 2:

[8]Scholz's book presents a comprehensive review of the relationship between earthquake occurrence and fault geometry. It explores the influence of fault orientation, stress fields, and tectonic processes on seismic activity patterns. The author discusses various fault models, their implications for earthquake behavior, and the role of fault interactions in controlling seismic sequences. This work provides valuable insights into understanding the spatial distribution and clustering of earthquakes.

Literature Review 3:

[9]This work examines recent earthquake occurrences in mainland China and their implications for seismic hazard assessment. The authors analyze the characteristics of major earthquake events, including their magnitudes, depths, and locations. They also discuss the potential influence of these events on seismic hazard maps and the need for updating risk assessments. The review highlights the importance of continuous monitoring and analysis of earthquake data for effective disaster preparedness and mitigation strategies.

Literature Review 4:

[10]This work examines techniques for observing and analyzing the earthquake rupture process, which is essential for understanding the physics of earthquakes and their potential impacts. It discusses integrating multiple data sources, such as seismic waveforms, geodetic measurements (e.g., GPS and InSAR), and tsunami observations, to characterize the spatial and temporal evolution of the rupture process. This includes studying the distribution of slip, rupture propagation, and coseismic deformation.

IV.Methodologies & Methods

Utilizing a variety of techniques and technology, the methodology for interpreting the seismic data from 2023 streamlines the analysis process leveraging Python, Tableau, and MongoDB for various stages .Establishing precise objectives that spell out the aims and queries the analysis is meant to answer is the first step in this procedure.First, the emphasis is on data preprocessing activities in a Python notebook. This stage is essential for creating prediction models and making sure the dataset is ready. This entails importing necessary libraries, including NumPy, Scikit-learn, and Pandas, and then loading the seismic dataset into a Pandas DataFrame. After the data is loaded, it is meticulously cleaned to remove numerous issues like duplicates, outliers, and missing values.This phase also encompasses standardizing data formats and converting data types to ensure consistency across the dataset. Additionally, feature engineering techniques are applied to either create new features or transform existing ones to enhance model performance. In this analysis pipeline, MongoDB acts as the repository for exploratory data analysis (EDA). After being cleaned and preprocessed, the seismic dataset is imported into MongoDB and sorted into collections according to pertinent characteristics such the depth, location, and magnitude of the earthquake. Then, using MongoDB's querying and aggregation features, subsets of data are extracted according to predetermined standards, and aggregation operations are carried out to retrieve insights for visualizations.Queries may involve retrieving earthquakes above a certain magnitude or aggregating data by location or time to get visualizations.

On the other hand,Tableau is used for transforming the seismic dataset housed in MongoDB into engaging and informative visualizations. Connecting Tableau and MongoDB together makes data easy to access and enables the development of dynamic dashboards and visuals.It made the locations, intensities, and depths of earthquakes easier to see and gave the seismic activity a spatial context. Furthermore, time series visualizations allow for the analysis of temporal trends and patterns, giving stakeholders a better understanding of seismic activity across time. Users can interactively examine the data by adding filters and parameters, which promotes a better comprehension of seismic trends and aids in well-informed decision-making.
Machine learning techniques are subsequently chosen and trained on the preprocessed data to forecast earthquake classes, depths, and magnitudes. These algorithms include Random Forest, Decision Tree, and Support Vector Machine (SVM),K Means Clustering,ARIMA. These models' accuracy and dependability are improved by rigorous evaluation of their performance, which is optimized.The methodology guarantees a comprehensive approach to earthquake analysis by integrating Tableau, Python, and MongoDB. This includes exploratory data analysis, machine learning, data visualization, and data preprocessing.

V.Exploratory Data Analysis:

[6]A number of visualizations used in the investigation of seismic activity offer important new perspectives on different facets of earthquake events.
The depth of earthquakes classified by type can be shown to exhibit clear patterns in the boxplot that shows depth versus seismic activity type. It provides important insights into the geological processes behind seismic occurrences by showcasing changes in seismic events such as tectonic plate shifts, volcanic activity, or artificial seismicity.

The boxplot that displays the magnitude in relation to the kind of seismic activity also provides insight into the distribution of earthquake magnitudes across various types, as well as the intensity and severity of seismic occurrences that fall into each category[1].

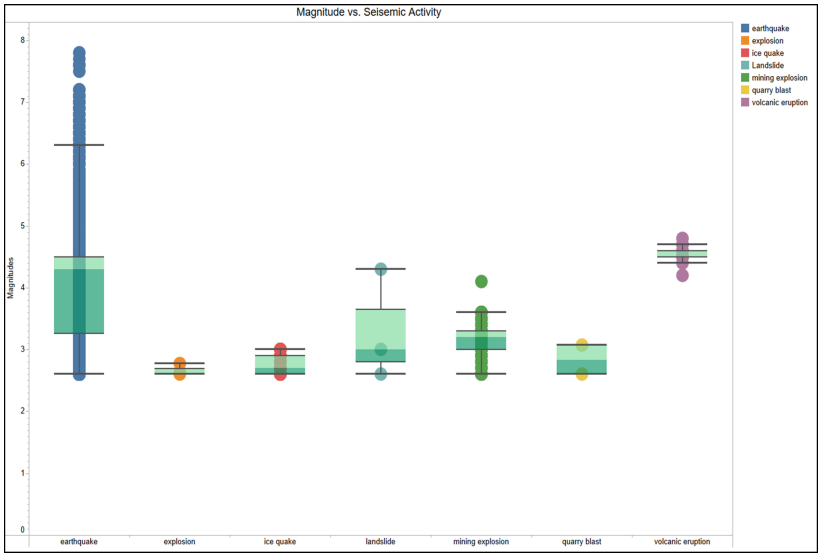

Fig 1: Magnitude of each seismic activity

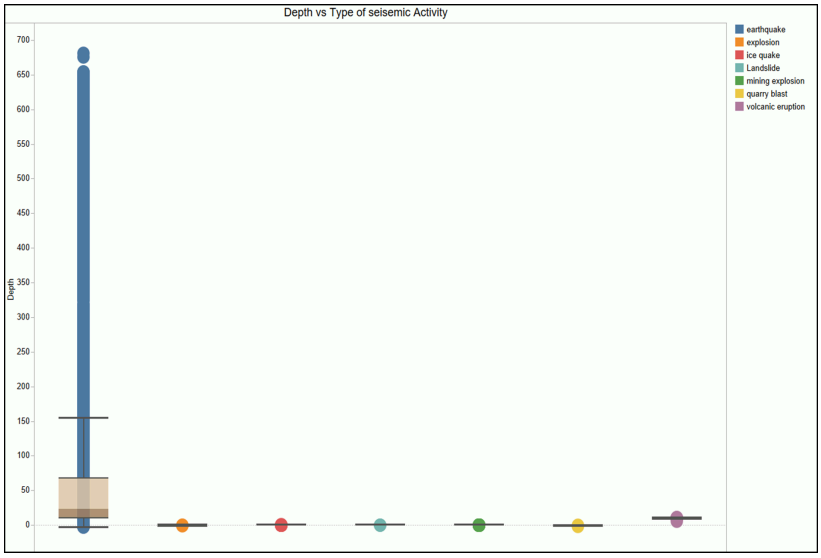

Fig2: Depth of each seismic activity

In the meantime, the scatterplot illustrating magnitude and depth error vs seismic activity type provides information about the quality and precision of seismic readings and highlights possible causes of uncertainty in seismic data analysis.

We can see more about the energy release linked to various kinds of seismic occurrences by navigating to the boxplot that compares seismic activity versus Root Mean Square (RMS). It is possible to compare RMS values across different forms of seismic activity with this graphic, which clarifies differences in the length and amplitude of seismic waves caused by various geological events.

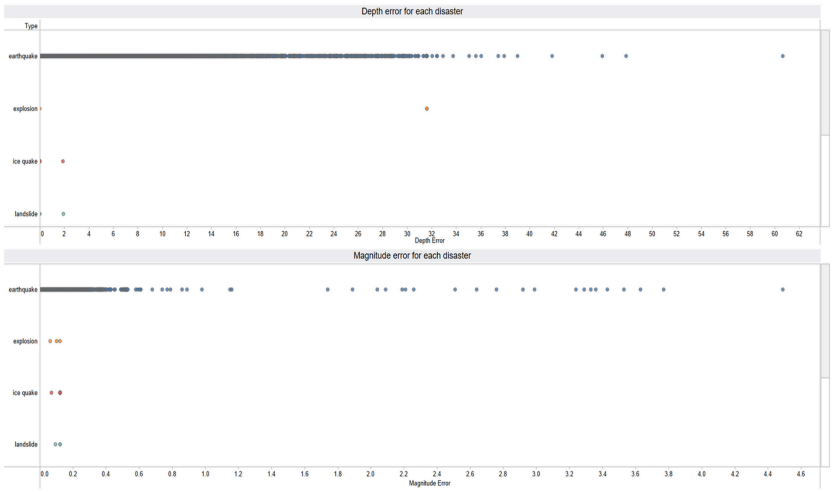

Fig 3: The magnitude and depth errors for every seismic activity

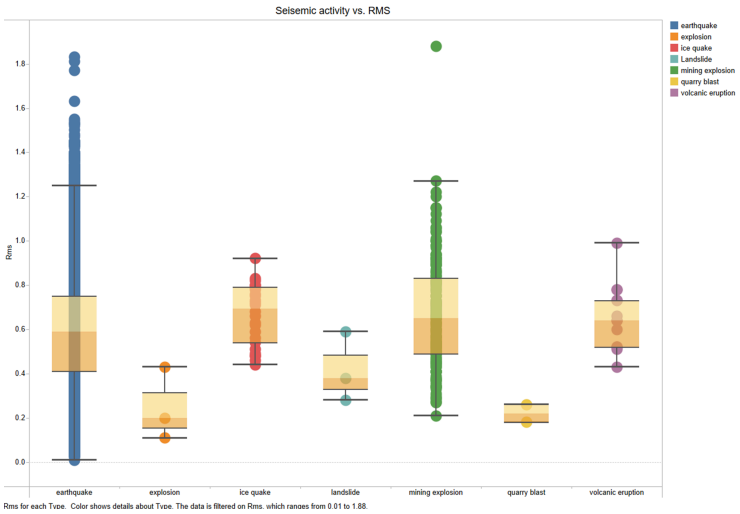

Fig 4: Rms vs. Seismic activity

Additionally, the seasonal trends in earthquake occurrences are revealed by the column bar plot that contrasts different seasons with different forms of seismic activity, which offers important context for comprehending the temporal patterns of seismic activity. Summer is the season mostly affected by the majority of the disasters.

Risk assessment and disaster preparedness activities are aided by knowing which regions are most likely to experience major seismic occurrences by identifying the top 10 places with the highest magnitude earthquakes.

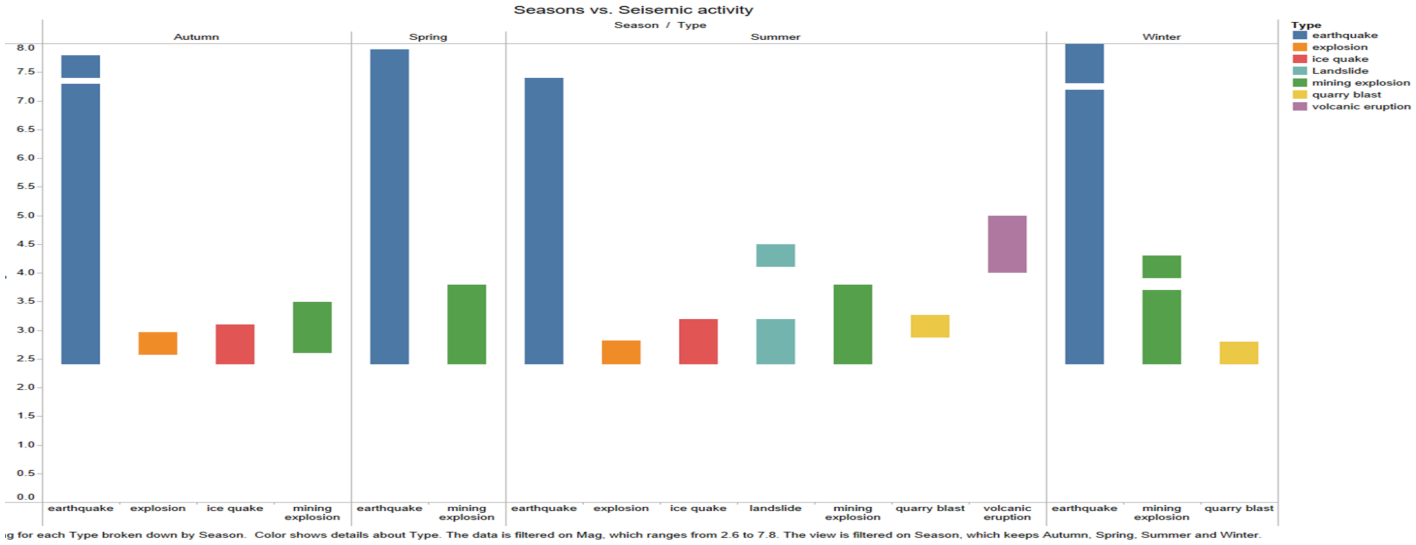

Fig 5: Seasons vs. Seismic activities

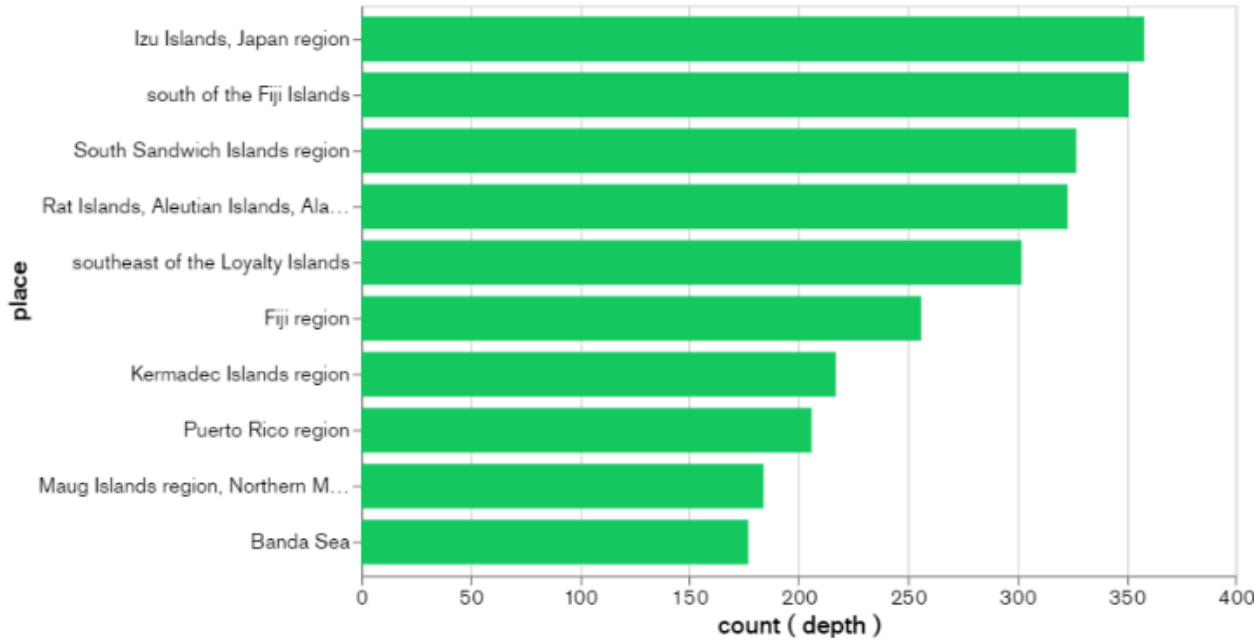

Fig 6: Top 10 places affected by earthquakes with respect to the magnitude

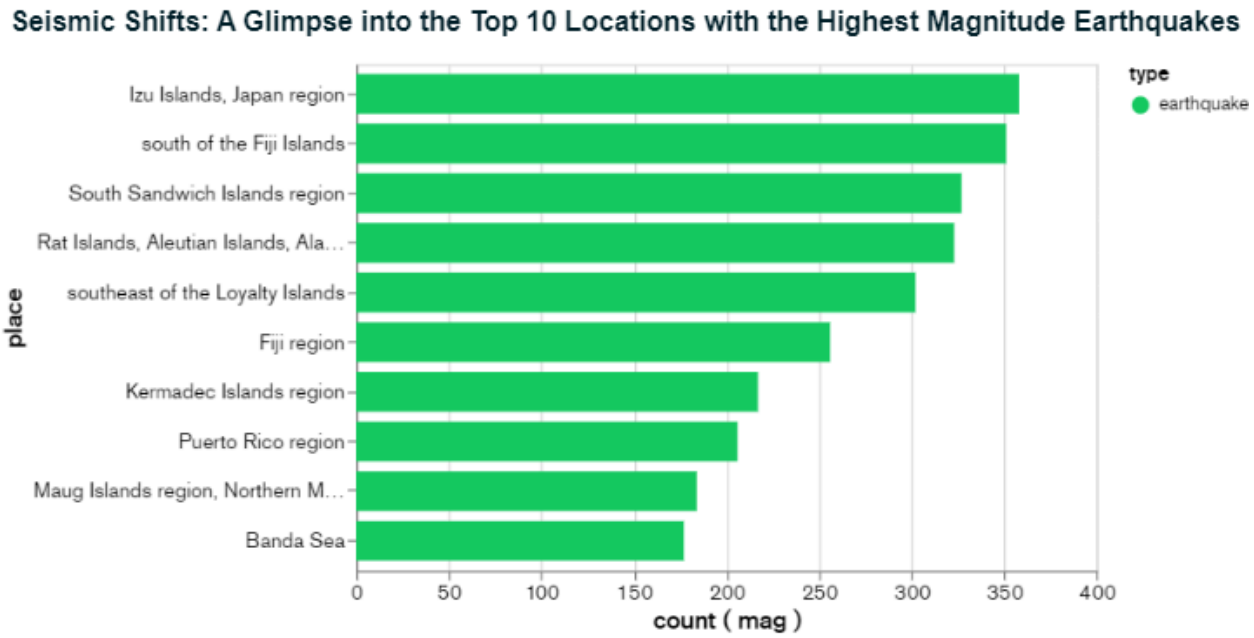

Fig 7: Top 10 places affected by earthquakes with respect to the depth

In the meantime, stakeholders may see spatial patterns and seismic event hotspots through interactive maps which display the type of seismic activity. This makes it easier to implement targeted mitigation and intervention efforts in high-risk locations.

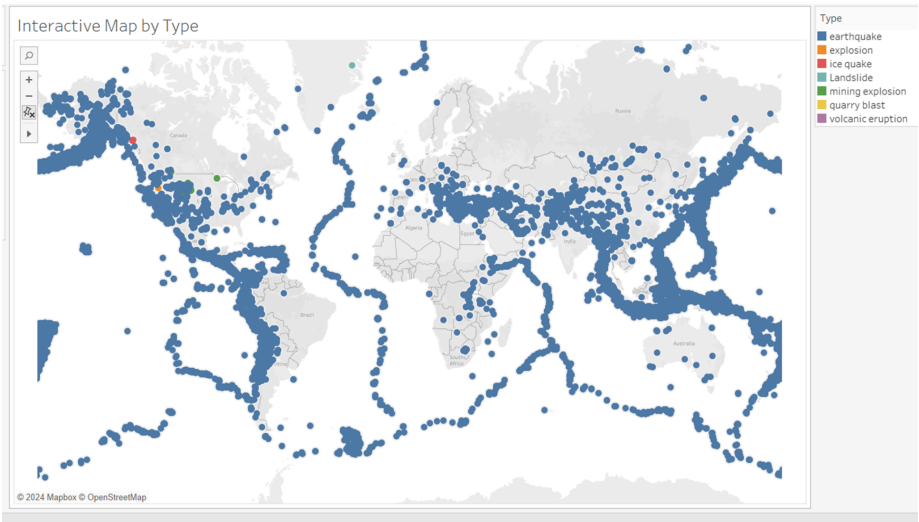

Fig 8: Spatial analysis of the all the seismic activities

In order to understand underlying links and dependencies within the dataset, it is necessary to investigate correlations between various seismic parameters, such as magnitude, depth, and seismic activity type. Analysts can identify factors influencing seismic activity and prioritize mitigation actions based on the degree and direction of these correlations.

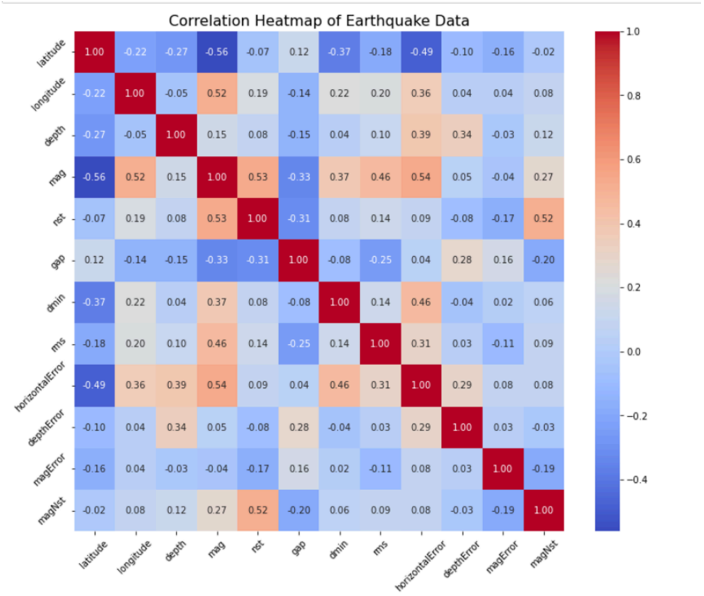

Fig 9: Correlation heat map

Overall, these visualizations collectively contribute to a comprehensive understanding of seismic activity, informing evidence-based decision-making and enhancing earthquake risk management strategies.

VI.Machine Learning Models:

A.Feature Importances:

Feature importances provide important insights into the dynamics of seismic activity prediction in the Random Forest,Decision Tree models research. Leading the pack, earthquake depth is the most significant predictor, contributing over 59.4% of the model's predictive ability. This emphasizes how earthquake depth has a significant influence on seismic activity, with deeper seismic events possibly suggesting different geological processes and offering different degrees of risk. Geographical variables come in second and third, accounting for 9.1% and 8.1% of the total, respectively, of longitude and latitude. These features draw attention to the spatial aspect of seismic activity and demonstrate how tectonic plate borders and local geological features influence patterns of earthquake occurrence.Furthermore, depth error and horizontal error, with contributions of 5.6% and 7.6%, respectively, are important components of the model. The uncertainty around estimates of earthquake depth and position is reflected in these measurements, highlighting how crucial precise data are to the prediction of seismic activity. Furthermore, the contribution of seismic station-related parameters such magnitude Nst and the

number of stations (Nst) is 2.7% and 3.4%, respectively, showing how important data coverage and quality are to improving forecast accuracy.Moreover, depth error and horizontal error play significant roles in the model, with contributions of 7.6% and 5.6%, respectively. These metrics reflect the uncertainty associated with earthquake location and depth estimates, underscoring the importance of accurate data in seismic activity prediction. Even though they are significant, magnitude and magnitude error have comparatively lower importances of 1.5% and 2.5%, respectively. This indicates that while earthquake magnitude plays a major role in predicting seismic activity, the Random Forest model places more weight on other characteristics like depth and location. Stakeholders can effectively prioritize resources and activities by concentrating on characteristics that have the greatest impact on seismic risk assessment and mitigation measures by knowing the relative relevance of these aspects.

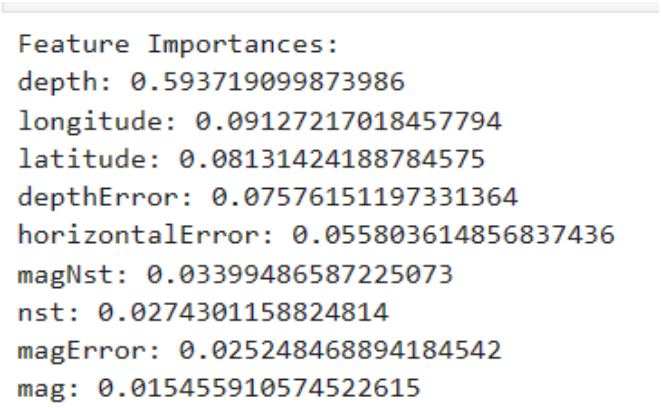

Fig 10: Feature importance from Random forest

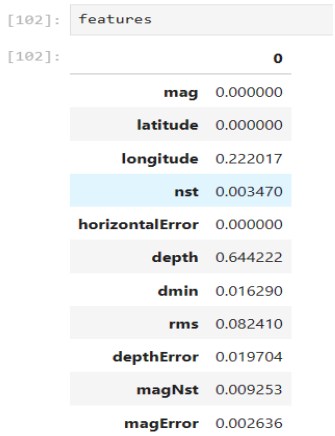

Fig 11: Feature importance from Decision tree

B.Classification Models:

Metrics like precision, recall, and F1-score offer valuable information about the Random Forest model's performance in varying seismic activity classes[3], and they are shown in the classification report. The model presents immaculate classification performance for the "earthquake" class, with perfect precision, recall, and an F1-score of 1.00. On the other hand, the precision, recall, and F1-score metrics for the "ice quake" and "volcanic eruption" classes are lower, indicating difficulties in correctly predicting these uncommon events. An F1-score of 0.91 indicates strong classification performance for the "mining explosion" class, which shows good precision and recall.

Support Vector Classifier's (SVC) classification report demonstrates that the model obtains precision, recall, and an F1-score of 1.00 for class "1," which probably corresponds to the majority class (such as "earthquake"). This indicates that the predictions for this category are

quite correct. The model appears to have failed to accurately classify any cases for classes "3", "5", and "7" as seen by the zero precision, recall, and F1-score metrics for these classes.
The model's overall accuracy of 99.36% appears high at first glance. The majority class's accurate classification, however, is what mostly determines this accuracy score; the model fails to accurately classify minority classes because of their unequal representation in the dataset.

The confusion matrix, which displays the number of true positive, false positive, true negative, and false negative predictions for each class, provides more insight into both model's performance. In this instance, the confusion matrix shows that the model has a high accuracy score since it accurately predicts every instance of the majority class (class "1"). For the minority classes, on the other hand, it is unable to predict any occurrences, which results in 0 precision, recall, and F1-score.
In summary, while the SVM,Random forest achieves high accuracy for the majority class, its performance is severely limited by its inability to accurately classify instances from minority classes. This highlights the importance of addressing class imbalance issues in the dataset and selecting appropriate evaluation metrics that account for class distribution when evaluating classifiers performance.

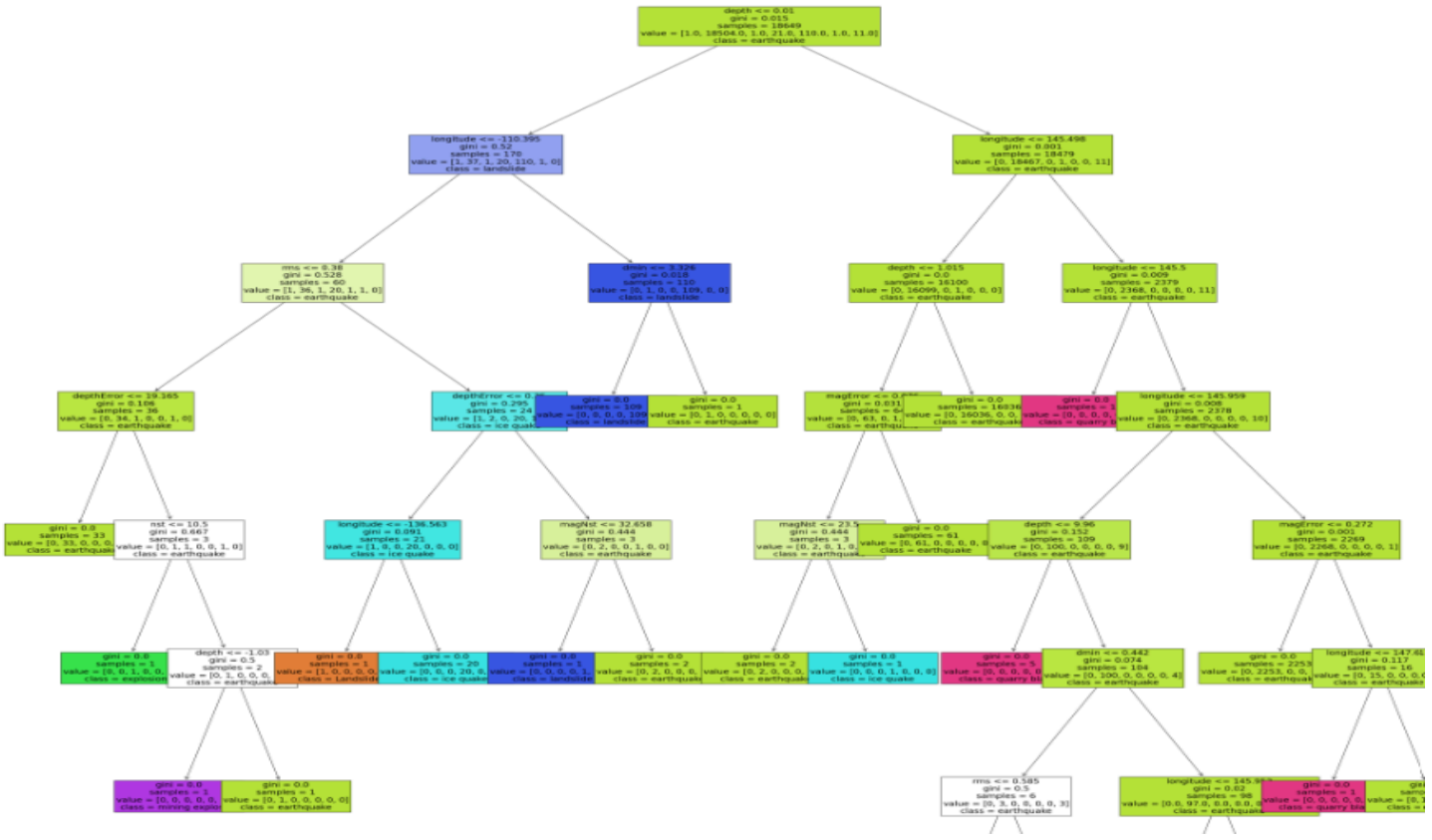

Fig 11: Decision tree

	precision	recall	f1-score	support
Landslide	0.00	0.00	0.00	1
earthquake	1.00	1.00	1.00	7924
explosion	0.00	0.00	0.00	2
ice quake	0.62	1.00	0.77	5
landslide	0.00	0.00	0.00	1
mining explosion	0.98	1.00	0.99	57
quarry blast	0.00	0.00	0.00	1
volcanic eruption	0.33	0.50	0.40	2
accuracy			1.00	7993
macro avg	0.37	0.44	0.40	7993
weighted avg	1.00	1.00	1.00	7993

Classification Report:

	precision	recall	f1-score	support
earthquake	1.00	1.00	1.00	5295
ice quake	0.20	0.50	0.29	2
mining explosion	0.91	0.94	0.92	31
volcanic eruption	0.00	0.00	0.00	1
accuracy			1.00	5329
macro avg	0.53	0.61	0.55	5329
weighted avg	1.00	1.00	1.00	5329

Confusion Matrix:

```
[[5289    2    2    2]
 [   0    1    1    0]
 [   0    2   29    0]
 [   1    0    0    0]]
```

Accuracy Score: 0.9981234753237005

Fig 12: Classification performance metrics

C.Regression:

Regression metrics, in addition to classification metrics, include information about how well the model predicts continuous variables like earthquake depth or magnitude. Lower numbers indicate better prediction accuracy. The mean absolute error and mean squared error quantify the average difference between expected and actual values[2]. Higher values indicate better predictive ability. The R-squared metric quantifies the percentage of variance in the target variable explained by the model.

With a mean absolute error of 0.011 and a mean squared error of 0.029 in our investigation, the Random Forest regression model shows comparatively modest prediction errors. With an R-squared value of 0.712, the model is able to capture the underlying associations between predictors and seismic features, accounting for about 71.2% of the variance in the target variable.

All things considered, the Random Forest model performs admirably in both classification and regression tasks, offering insightful information on the prediction and classification of seismic activity.

```
Regression Metrics:
Mean Absolute Error: 0.010701488403181528
Mean Squared Error: 0.02520981365898012
R-squared: 0.7496543824630768
```

Fig 13: Regression metrics

D. K-Means Clustering:

The cluster centers provide important information about the spatial distribution of seismic events that are derived using the K-Means clustering algorithm. Within each cluster, the average position of seismic occurrences is represented by the cluster center. Starting with Cluster Center 1, which is roughly positioned at (11.83, 125.02), this point most likely denotes an area with high seismic activity. The coordinates point to a region close to the Pacific Ring of Fire, which is well-known for being a seismically active zone with regular earthquakes and volcanic eruptions. Located approximately at 30.94 and -160.17, Cluster Center 2 is another unique geographic location with a unique seismic signature. The coordinates suggest a position in the Pacific Ocean or on the western coast of North America, where tectonic plate boundaries produce significant seismic activity. Lastly, Cluster Center 3, which is located close to (7.75, -65.54), indicates a distinct region with distinct seismic properties. These coordinates might indicate locations in South America that are well-known for experiencing seismic activity as a result of tectonic interactions, including the Caribbean Plate or the Andes mountain range. All things considered, the analysis of the cluster centers yields important geographical insights that enable stakeholders to pinpoint locations with comparable patterns of seismic activity and customize mitigation plans to address the particular risks associated with earthquakes in these places.

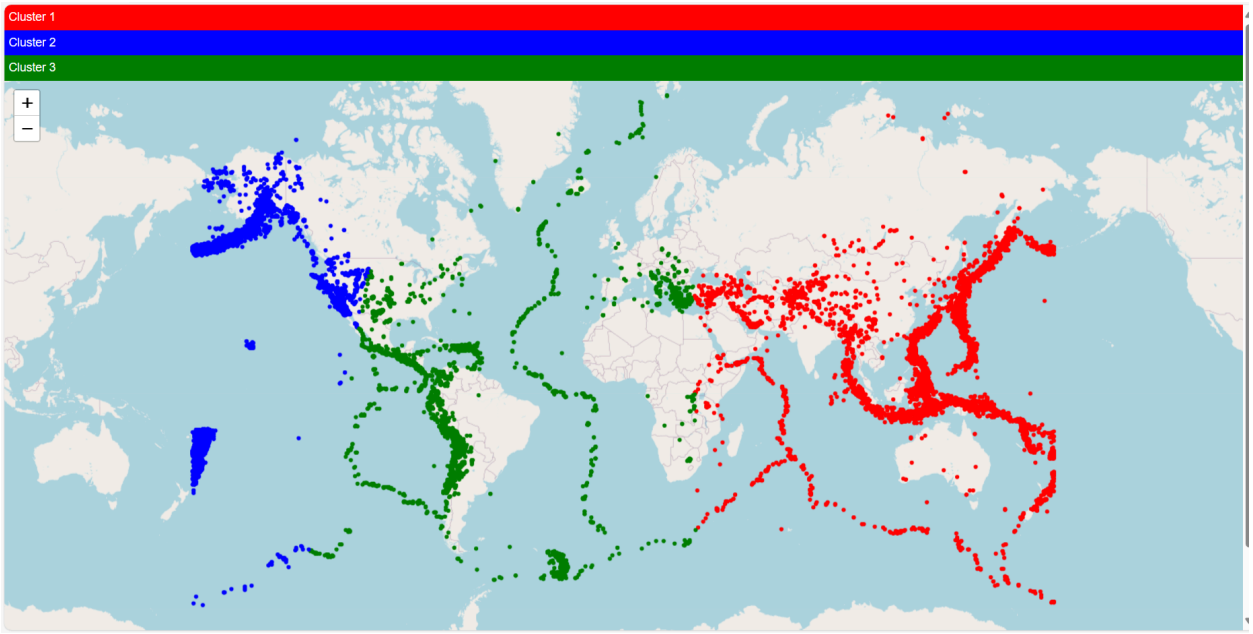

Fig 14:cluster analysis based on latitudes and longitudes

E.Time Series Analysis:

The data was fitted to an ARIMA(5, 1, 0) model in the time series analysis using the SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous components) model. The autoregressive link between the different forms of seismic activity over time is shown by the model's coefficients. In particular, the negative coefficients for the lag factors (ar.L1 to ar.L5) indicate a tendency toward a decline in the types of seismic activity, which, according to historical data, indicates a decreased chance of seeing a certain type of seismic event in the future.Each coefficient's standard errors and associated z-statistics shed light on how significant they are. At a high confidence level ($p < 0.05$), every coefficient is statistically significant, demonstrating the significance of each coefficient in the model. Furthermore, the model appears to have a decent fit to the data based on the log likelihood, AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), and HQIC (Hannan-Quinn Information Criterion) scores; lower numbers indicate a better match.

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   type   No. Observations:                26642
Model:                 ARIMA(5, 1, 0)   Log Likelihood              -11173.007
Date:                Wed, 10 Apr 2024   AIC                          22358.014
Time:                        14:07:56   BIC                          22407.156
Sample:                             0   HQIC                         22373.872
                                - 26642
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.7846      0.002   -501.338      0.000      -0.788      -0.782
ar.L2         -0.5986      0.002   -308.789      0.000      -0.602      -0.595
ar.L3         -0.4251      0.002   -193.140      0.000      -0.429      -0.421
ar.L4         -0.2872      0.002   -115.510      0.000      -0.292      -0.282
ar.L5         -0.1252      0.002    -52.076      0.000      -0.130      -0.120
sigma2         0.1354      0.000    818.777      0.000       0.135       0.136
===================================================================================
Ljung-Box (L1) (Q):                   4.84   Jarque-Bera (JB):          11408926.81
Prob(Q):                              0.03   Prob(JB):                         0.00
Heteroskedasticity (H):               0.11   Skew:                             7.95
Prob(H) (two-sided):                  0.00   Kurtosis:                       103.13
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
Forecasted earthquakes: 26642    1.0
26643    1.0
26644    1.0
26645    1.0
26646    1.0
26647    1.0
26648    1.0
26649    1.0
26650    1.0
26651    1.0
26652    1.0
26653    1.0
Name: predicted_mean, dtype: float64
```

Fig 15: Key findings from arima model

VII.Key Findings & Insights

Based on the analysis performed with Random Forest and Decision Trees classification models, location, depth, and magnitude are the most important features for determining seismic activity. These characteristics offer important new information about the magnitude, location, and intensity of seismic occurrences. An earthquake's strength is indicated by its magnitude, and its origin is located a certain distance below the surface of the earth. Location helps identify areas that are vulnerable to earthquakes by providing details on the geographic coordinates where seismic activity occurs.

K-means clustering analysis revealed that the Pacific Ocean, Japan, Indonesia, and South America are among the locations most prone to earthquakes in terms of seismic activity. Frequent seismic events are a result of geological activity and tectonic plate borders in certain places.

Relationships between the type, intensity, and depth of seismic activity were identified using the regression model analysis. The findings show that the magnitude and depth patterns of various seismic activity categories are distinct. Comprehending these correlations is essential for evaluating the likelihood of earthquakes and putting into practice efficient mitigation techniques.

The SARIMAX findings also shed light on the time series analysis of seismic activity. According to the ARIMA(5, 1, 0) model, there are autoregressive associations between different forms of seismic activity over time. Certain types of seismic occurrences appear to be declining, as shown by negative coefficients. While predicted earthquakes help stakeholders anticipate and get ready for possible seismic events, diagnostic tests confirm the model's goodness-of-fit.

VIII.Limitations & Future Works

Although the study of earthquakes offers important insights into patterns of seismic activity and related hazards, it is important to recognize a number of limitations and suggest possible directions for further research. Firstly, the results of the research are greatly influenced by the quality and availability of earthquake data, with conclusions that are skewed by incomplete or inconsistent data coverage. Future initiatives can concentrate on extending data collection activities to minority areas and enhancing data quality through standardized reporting standards. Furthermore, biases in the analysis may be introduced by imbalances in the distribution of earthquake types or occurrences across different locations, which calls for the investigation of methods to overcome concerns related to class imbalance. There may be issues with interpretability and processing resources due to the intricacy of the machine learning models and algorithms utilized in the investigation.Alternative methods or simplified models could be

investigated to lessen these difficulties without sacrificing performance. Incorporating other datasets, such as demographic data, climate data, or geological maps, may also contribute to a more thorough understanding of the temporal and spatial dynamics of earthquake activity. Improving the interpretability of models and measuring the uncertainty of predictions are essential for risk management and well-informed decision-making. Last but not least, integrating earthquake analysis with real-time disaster management platforms can help with prompt mitigation and reaction. Research projects that involve stakeholders and interdisciplinary teams must work together to overcome these obstacles and advance earthquake analysis in order to improve society's ability to withstand seismic risks.

Conclusion

Significant insights into seismic activity patterns gained by earthquake analysis, which advances our knowledge of related dangers and possible mitigation measures. The highest earthquake activity is found in locations like the Pacific Ocean, Japan, Indonesia, and South America. Other important findings include the identification of magnitude, depth, and position as significant features for determining seismic activity. Relationships between the kind, size, and depth of seismic activity are also found in the analysis, which offers important information for risk assessment. Furthermore, temporal trends in seismic activity are highlighted by time series analysis utilizing SARIMAX models, which supports forecasting and preparedness initiatives. Restrictions including poor data quality, complicated models, and unreliable predictions highlight the necessity of more study and cooperation in order to improve seismic analysis and catastrophe resistance. Overall, the study aids in risk management plans and well-informed decision-making for reducing the effects of earthquakes on infrastructure and populations.

References

[1] Patil, P. (2022, May 30). What is Exploratory Data Analysis? - Towards Data Science. ***Medium*****.** **https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15**

[2] Loh, W. (2011). Classification and regression trees. ***Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery/Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*****,** ***1*****(1), 14–23.** **https://doi.org/10.1002/widm.8**

[3] R, S. E. (2024, April 19). ***Understand random forest algorithms with examples (Updated 2024)*****. Analytics Vidhya.** **https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/**

[4] GeeksforGeeks. (2023, May 6). ***Data preprocessing in data mining*****. GeeksforGeeks.** **https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/**

[5] Algorithm AS 136: A K-Means clustering algorithm on JSTOR. (n.d.). ***www.jstor.org*****.** **https://www.jstor.org/stable/2346830**

[6]Visualize your data. Tableau. **https://www.tableau.com/trial/visualize-your-data?d=7013y0000020GqiAAE&msclkid=afdcbc7b65a111a71a64da4c9c2d0c8b&utm_content=7013y0000020GqiAAE**

[7] Arnaud Mignan, Jochen Woessner. Understanding Seismicity Catalogs and their Problems. **http://www.corssa.org/export/sites/corssa/.galleries/articles-pdf/Mignan-Woessner-2012-CORSSA-Magnitude-of-completeness.pdf**

[8] Scholz, C. H. (2019). ***The mechanics of earthquakes and faulting*****. Cambridge University Press.**

[9] Weijin xu, Jian Wu, Mengtan Gao. Seismic Hazard Analysis of China's Mainland Based on a New Seismicity Model. ResearchGate. **https://www.researchgate.net/publication/369988599_Seismic_Hazard_Analysis_of_China%27s_Mainland_Based_on_a_New_Seismicity_Model**

[10] Diego Melgar and Gavin P. Hayes. Systematic observations of the slip pulse properties of large earthquake ruptures. USGS. https://pubs.usgs.gov/publication/70192325