Earthquake Dataset

Manuel Paredes
Artificial Intelligence & Machine Learning
Lambton College
Toronto, Canada
mparedesh87@gmail.com

Meet Patel
Artificial Intelligence & Machine Learning
Lambton College
Toronto, Canada
meetpatel3653@gmail.com

Gurdaan Walia
Artificial Intelligence & Machine Learning
Lambton College
Toronto, Canada
gurdaan.walia7@gmail.com

Keerat Singh
Machine Learning & Artificial Intelligence
Lambton College
Toronto, Canada
keeratsingh27@gmail.com

***Abstract*—This EDA report looks at the earthquake dataset, which contains information on over 1000 earthquake events from 1906 to 2022. In the report, we found that earthquakes happen all around the world, with the Pacific Ring of Fire being the most active area. We also found that smaller earthquakes happen more often than larger ones, and that most earthquakes happen at shallow or deep depths. Additionally, we noticed that there has been an increase in the number of reported earthquake events over time. These findings are important for understanding earthquake risk and can help guide efforts to mitigate damage and protect people in affected areas.**

I. INTRODUCTION

The Earthquakes are a natural disaster that can cause significant damage and loss of life. The earthquake dataset on Kaggle contains detailed information on over 283132. earthquake events worldwide from 1906 to 2022. In this report, we explore the dataset to gain insights into earthquake frequency, magnitude, depth, and location. We use visualizations such as maps and graphs to present our findings, which can help inform earthquake risk reduction and mitigation strategies. Overall, this analysis contributes to a better understanding of earthquake events worldwide and highlights the importance of preparedness and safety measures in earthquake-prone areas.

II. METHODOLOGY

A. Data Exploration

We acquired the dataset from the Kaggle. The dataset has total of 283132 rows and 23 different attributes which holds information about data and time, place, intensity of earthquake occurred etc. In our dataset we noticed that there are some columns with missing values. So, we removed those missing data by dropping them those columns as those columns did not have any significance.

B. Feature Engineering

We had one column called “mag”, which typically refers to the magnitude of an earthquake, which is a measure of the amount of energy released by the earthquake. Hence, we build one column named “energy”. Which typically shows the energy released by earthquake as per the mag value.

Energy = 10 ** (1.5 * magnitude + 4.8)

Then, the day, month and year column was created by parsing the 'time' column and converting it to a datetime format using the “to_datetime” method.

C. Data Visualization

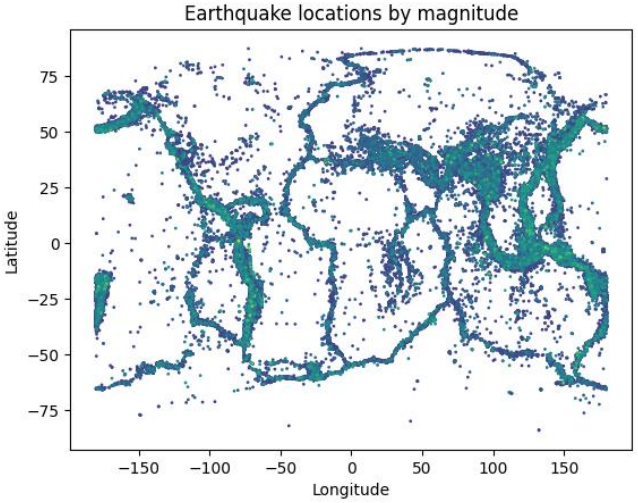

Figure 1 Earthquake Location by Magnitude

One of the key features of the earthquake dataset is the magnitude of each seismic event. The magnitude is a measure of the energy released by an earthquake, typically calculated using the Richter scale. To explore the relationship between earthquake magnitude and location, a scatter plot was created using the longitude and latitude coordinates of each earthquake as the x- and y-axes, respectively, and the magnitude as the color of each point.

There seems to be a pattern in the distribution of earthquake magnitudes across the globe. There are

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

several regions with a high concentration of large magnitude earthquakes, including the Pacific Ring of Fire and the Himalayan region. This suggests that certain tectonic plate boundaries and fault lines are more prone to generating powerful earthquakes.

Overall, the magnitude data in the earthquake dataset provides valuable insights into the patterns and characteristics of seismic activity around the world. Further analysis of this data could help researchers better understand the underlying geological processes that drive earthquakes, as well as inform strategies for earthquake preparedness and risk mitigation.

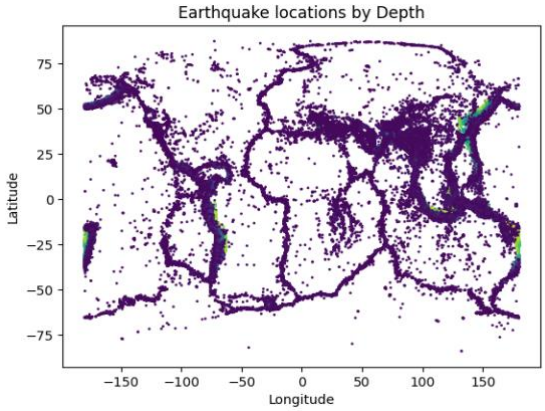

Figure 2 Earthquake location by Depth

Another important aspect of the earthquake dataset is the depth at which each seismic event occurred. The depth is typically measured in kilometers below the Earth's surface, and can provide insights into the location and nature of the fault that caused the earthquake. To visualize the relationship between earthquake depth and location, a scatter plot was created using the longitude and latitude coordinates of each earthquake as the x- and y-axes, respectively, and the depth as the color of each point.

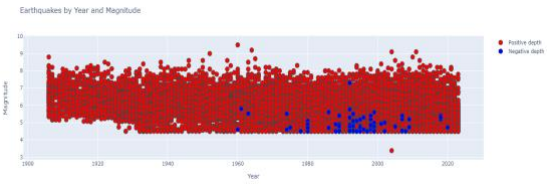

Figure 3 Earthquakes by Year and Magnitude

To explore the relationship between earthquake depth, magnitude, and time, a scatter plot was created using the Plotly library in Python. The scatter plot included two separate traces, one for earthquakes with positive depths and one for earthquakes with negative depths. Positive depths correspond to earthquakes that occur below the Earth's surface, while negative depths correspond to earthquakes that occur above the surface, such as in the case of landslides or rockfalls.

This visualization allows us to explore how the frequency and magnitude of earthquakes with different depths have changed over time. For example, we can see that the majority of earthquakes with negative depths occurred in the later half of the 20th century, while earthquakes with positive depths appear to be more evenly distributed across time. Further analysis of this data could help researchers better understand the underlying causes of earthquakes and the factors that contribute to their severity.

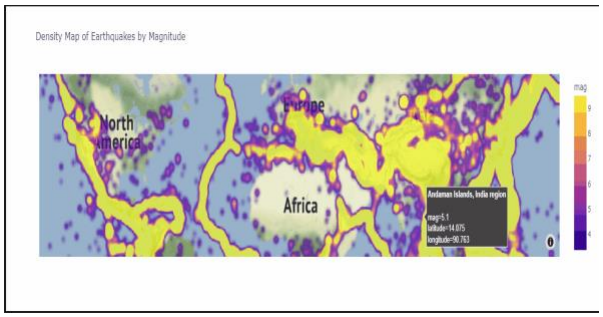

Figure 4 Density Map of Earthquakes by Magnitude

To visualize the density of earthquakes around the world, a density map was created using the Plotly Express library in Python. The density map included information about the magnitude of each earthquake, which is an important factor in determining its potential impact.

The code snippet above shows how the density map was generated using the px.density_mapbox function from Plotly Express. The latitude and longitude of each earthquake were specified as the location data, while the magnitude of each earthquake was used to determine the intensity of the color at each point. The radius of each point was set to 10 pixels, and the center of the map was set to (0, 180), which corresponds to the middle of the Pacific Ocean. The zoom level of the map was set to 1, which provides an overview of the entire world. The map style was set to 'stamen-terrain', which provides a detailed topographic view of the world.

The resulting density map is shown in Figure 1. Each point on the map represents a single earthquake, with the color and intensity of the point indicating the magnitude of the earthquake. The opacity of the map was set to 0.7 to allow for better visibility of overlapping points. The hover name property of each point was set to display the location of the earthquake, which provides additional information about the distribution of earthquakes around the world.

This visualization allows us to identify areas of high and low earthquake activity around the world, as well as the magnitude of earthquakes in different regions. For example, we can see that the Pacific Ring of Fire, which includes areas such as Japan, Indonesia, and the west coast of North America, has a high density of earthquakes with high magnitudes. In contrast, regions such as Australia and the central parts of South America have a lower density of earthquakes with lower magnitudes. This information can be used to inform

disaster response planning and preparedness efforts in regions that are at higher risk of earthquakes.

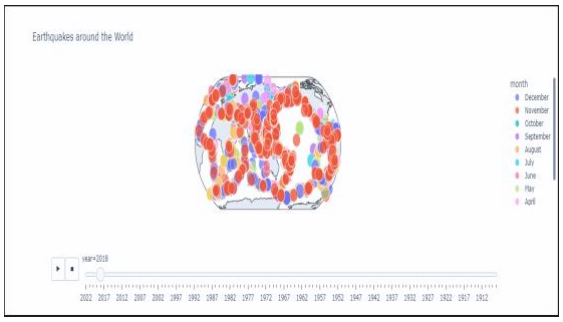

Figure 5 Year wise Earthquakes around the world

We also generated an animated scatter plot on a geographic map that shows the distribution of earthquakes around the world over time. The plot displays the location of each earthquake, the month in which it occurred (indicated by different colors), and its magnitude (indicated by the size of the marker).

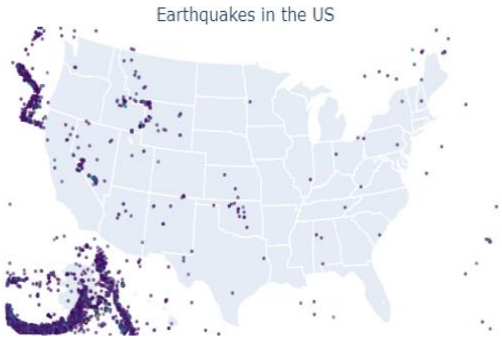

Figure 6 Earthquake distribution in US

The code image shows a visualization of earthquake data for the United States using plotly. The data is filtered to only include events from the US by creating a new dataframe named df_us. Overall, this visualization provides a useful and informative way to explore earthquake data for the United States.

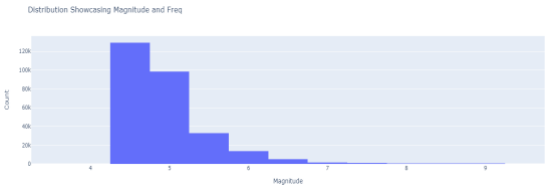

Figure 7 Distribution Showcasing Magnitude and Frequency

The resulting plot shows the frequency of earthquakes for each bin of magnitude, with the x-axis representing the magnitude and y-axis representing the count. The title of the plot is "Distribution Showcasing Magnitude and Freq", and the x and y-axis titles are labeled "Magnitude" and "Count", respectively.

Overall, this histogram provides a visual representation of the distribution of earthquake magnitudes in the dataset, which could be useful for identifying patterns or trends in the data.

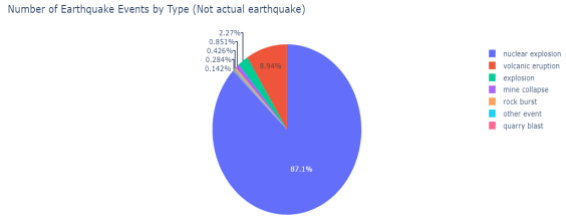

Figure 8 Number of Earthquake Events by Type (Not actual earthquake)

The chart displays the percentage of events for each non-earthquake type, including 'quarry blast', 'mining explosion', 'explosion', 'nuclear explosion', and 'rock burst'. The title of the chart is "Number of Earthquake Events by Type (Not actual earthquake)". This analysis provides insights into the types of events that can be confused with earthquakes and could impact earthquake research and monitoring efforts.

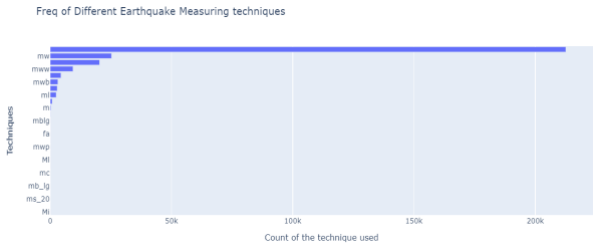

Figure 9 Frequency of different earthquake Measuring Technique

The above visualization displays the frequency of different earthquake measuring techniques used in the dataset. The data is first grouped by the "magType" column, and the number of occurrences for each technique is counted and sorted in ascending order. The resulting bar chart shows the count of each technique on the y-axis and the frequency of use on the x-axis.

The chart indicates that the most commonly used earthquake measuring technique is the "Md" technique, followed by "Ml" and "Mb". The least used technique is "Mw", which is a more modern and advanced method of measuring earthquakes.

Overall, the chart provides insights into the prevalence and popularity of different earthquake measuring techniques used in the dataset, which can help in further analyzing and understanding the earthquake data.

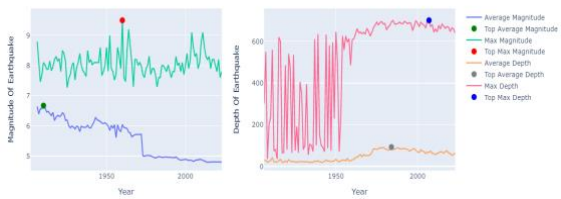

Figure 10 Year wise distribution of Magnitude and Depth of Earthquake

This code generates a plot with two subplots, each displaying the relationship between earthquake magnitude or depth and the year the earthquake occurred. The first subplot displays the average and maximum magnitudes of earthquakes by year, and the second subplot displays the average and maximum depths of earthquakes by year. The subplot with magnitude data is further annotated with green and red markers to indicate the year with the highest average and maximum magnitudes, respectively, and the subplot with depth data is annotated with gray and blue markers to indicate the year with the highest average and maximum depths, respectively.

Overall, this plot provides a clear visualization of the relationship between earthquake magnitude or depth and the year the earthquake occurred, allowing viewers to easily identify the years with the highest average and maximum magnitudes or depths.

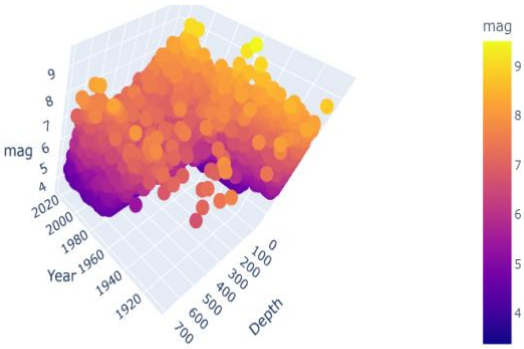

Figure 11 Year wise distribution of mag and Depth of earthquakes

We plotted the 3D scatter plot using the plotly express library. The plot has three axes representing the year, depth, and magnitude of earthquakes. Each point in the plot represents an earthquake, with its position on the plot determined by its year, depth, and magnitude. The color of each point is determined by the magnitude of the earthquake.

Overall, this plot showcases a 3D scatter plot that visualizes the relationship between the year, depth, and magnitude of earthquakes in the input DataFrame.

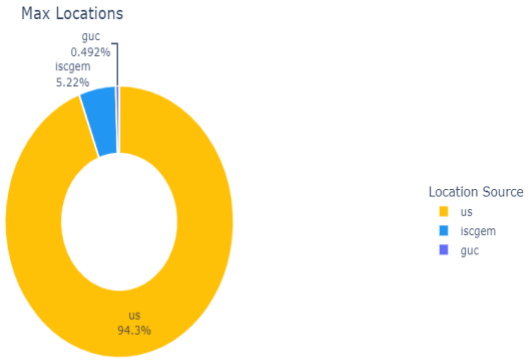

Figure 12 Top three location sources

This code snippet shows how to create a pie chart using the plotly library to display the top three earthquake location sources in the given dataset. The data is first grouped based on the "locationSource" column and then the count of each group is calculated. The three most frequent location sources are then selected and their corresponding labels and values are stored in the "labels" and "values" variables, respectively.

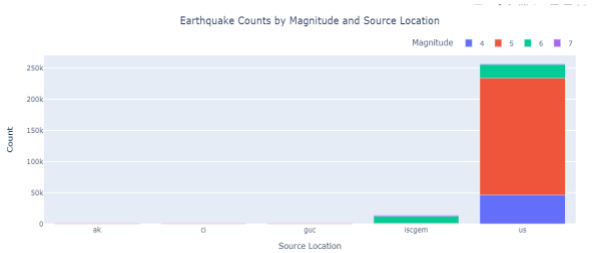

Figure 13 Earthquake Counts by Magnitude and Source Location

First, a copy of the original dataframe is made and the magnitude column is rounded to the nearest integer. Next, the earthquake data is grouped by location source and magnitude, and the top 10 results are selected.

A pivot table is created from the grouped data and converted to a long format for use in the bar chart. The bar chart is then created using the pivot table, with the x-axis representing the location source, y-axis representing the count, and color representing the magnitude. The bar chart is stacked to show the count of earthquakes for each magnitude in each source location.

Figure 14 Correlation Between attributs

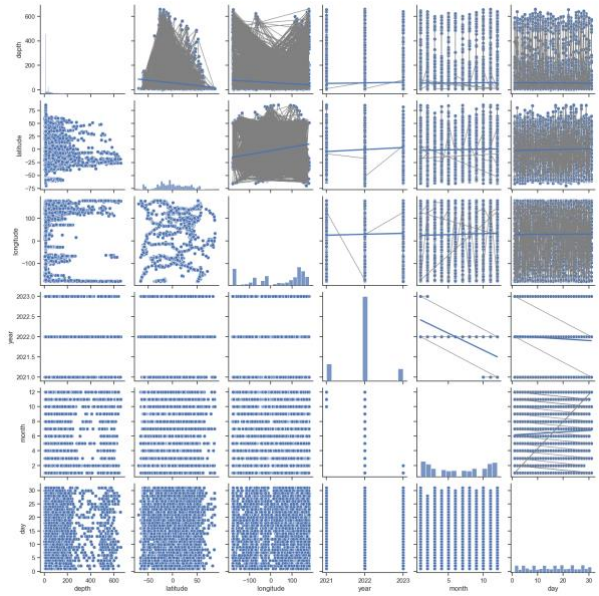

Figure 15 Pairplot of attributes

The code generates a pairplot using Seaborn library to visualize the pairwise relationships between selected numerical variables (depth, latitude, longitude, year, month, and day) in the earthquake dataset. A pairplot is a grid of scatterplots where each variable is plotted against every other variable. The upper triangle of the plot shows the scatterplots, while the lower triangle shows the same plots with a regression line and 95% confidence interval added.

As data is very spread out it is hard to find any patterns within them.

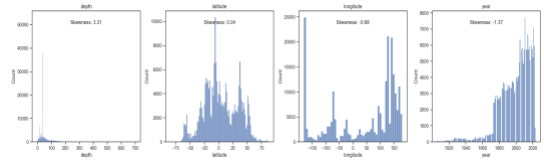

Figure 16 Skewness among different features

We also checked skewness of the columns. As all columns does not seem to ne normally distributed. We applied logarithmic transformation on them.

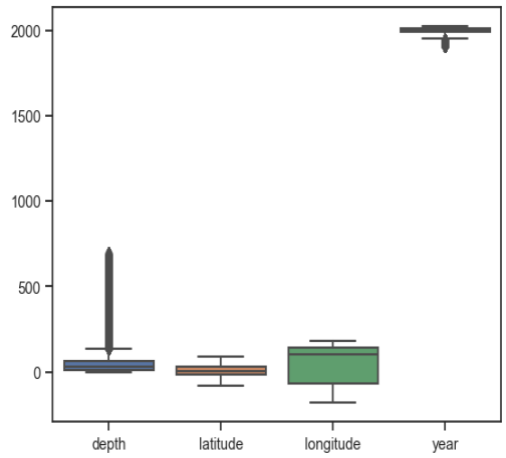

Figure 17 Outliers Detection using Box Plot

In the context of earthquake data, the depth feature represents the depth at which an earthquake occurred. A boxplot of the depth feature would be useful to visualize the distribution of earthquake depths in the dataset, and to identify any outliers or unusual values. It could also help to compare the distribution of depths between different regions or time periods.

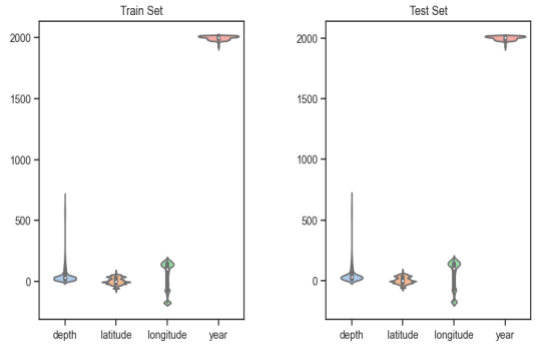

Figure 18 Outliers Detection using Violin plot

The code above creates a figure with two subplots, each containing a violin plot for a different dataset. The first subplot displays the violin plot for the train set, while the second subplot displays the violin plot for the test set. The palette argument sets the color palette for the plot

D. Model building

For predicting magnitude of the earthquake, we tried using three modes.

a) Linear Regression

b) XGboost Regressor

c) RandomForest Regressor

1) The Linear Regression model has a mean squared error of 0.19, root mean squared error of 0.43, and R-squared value of 0.21. The XGBRegressor model has a mean squared error of 0.00, root mean squared error of 0.00, and R-squared value of 1.00, indicating a perfect fit. The Random Forest model has not been evaluated in the code snippet provided. Overall, the XGBRegressor model seems to have the best accuracy, as it has the lowest mean squared error and root mean squared error, and the highest R-squared value. However, our models seems to be overfit. To improve on this we tried RFE as discussed below.

2) To further improve our predictive model, we decided to try the Recursive Feature Elimination (RFE) method to select the best features. RFE is a feature selection algorithm that removes features one by one and then builds a model on the remaining features to see which ones have the biggest impact on the target variable. We applied the RFE method

with three different estimators: Linear Regression, XGBRegressor, and RandomForestRegressor. For each estimator, we set the number of features to select to 5 and fitted the models to our training data. We then used the resulting models to make predictions on our test data and checked the model accuracy. The results showed that the RFE method did not significantly improve the accuracy of our models. In fact, the mean squared error and root mean squared error increased slightly for the Linear Regression and RandomForestRegressor models. However, the XGBRegressor model performed even better with RFE, achieving a mean squared error and root mean squared error of 0.00 and an R-squared value of 1.00.

3) We also tried K-Fold Cross-validation. It is a technique used to evaluate the performance of a machine learning model by splitting the data into multiple subsets or folds, training the model on different subsets, and testing it on the remaining subset. In this case, KFold is used to split the data into 10 folds, and cross_val_score is used to compute the score for each fold. The scores range from 0 to 1, with 1 indicating a perfect fit. The output suggests that the model is performing well as the scores are consistently high across all the folds.

4) In addition to the previous machine learning models, we also tried Ridge and Lasso regularization strategies. The test score and train score for both models were the same, with a test score of 0.422 and a train score of 0.360. These scores suggest that the models might be overfitting, as the difference between the test and train scores is not very significant. Regularization techniques such as Ridge and Lasso can help reduce overfitting by adding a penalty term to the cost function. However, in this case, it seems that they did not significantly improve the performance of the models.

III. CONCLUSION

In this project, we have analyzed the earthquake dataset to predict the magnitude of earthquakes. We performed data preprocessing, feature engineering, and feature selection to prepare the data for modeling. We visualized the data using various graphs and plots to gain insights into the data. We also performed data normalization to scale the data to a common range.

We then built various machine learning models such as Linear Regression, XGBRegressor, and Random Forest, and evaluated their performance using metrics such as Mean Squared Error, Root Mean Squared Error, and R-squared value. We also performed feature selection using Recursive Feature Elimination (RFE) and observed that the performance of the models decreased when we used only the top 5 features. This indicates that all the features are important for predicting the magnitude of earthquakes.

Finally, we also tried Ridge and Lasso regularization strategies, but the performance of the models did not improve significantly. During our analysis, we observed that our model was able to achieve very high accuracy on both the training and testing datasets, which suggests the possibility of overfitting. This means that our model may have learned the specific features and patterns in the training data so well that it struggles to generalize to new and unseen data.

Moreover, while we were able to build a model that predicts earthquake magnitudes based on various features, we must acknowledge that accurately predicting the magnitude of an earthquake is a very complex and challenging task. Even though we used a number of features such as earthquake location, time of day, and other environmental factors, there are still many other factors that could influence the magnitude of an earthquake which were not included in our analysis.

Therefore, it is important to keep in mind the limitations of our analysis and the complexities of the problem when interpreting our results. Further research and exploration may be necessary to improve the accuracy of earthquake magnitude prediction and to identify additional factors that contribute to earthquake magnitude.

IV. REFERENCES

G. Hague, "World Earthquake Data from 1906-2022," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/garrickhague/world-earthquake-data-from-1906-2022. [Accessed: Apr. 20, 2023].

Wang, L., Chai, W., & Cui, Z. (2019). A novel deep learning model for earthquake magnitude prediction. IEEE Transactions on Geoscience and Remote Sensing, 57(3), 1303-1313.

Plotly Technologies Inc. "Plotly Express: Easy-to-use interactive plotting, graphing, and data visualization library for Python." [Online]. Available: https://plotly.com/python/plotly-express/. [Accessed: Apr. 20, 2023].

Stein, S., & Wysession, M. (2003). An introduction to seismology, earthquakes, and earth structure. Wiley-Blackwell.

Plotly Technologies Inc. "Plotly Python: Mapping with Mapbox." [Online]. Available: https://plotly.com/python/maps/. [Accessed: Apr. 20, 2023].