J. Sci. Inform. Soc. 2023, Vol. 1, No. 1
ISSN XXXX-XXXX

Revealing Insights Through Exploratory Data Analysis on Earthquake Dataset

Kiagus Muhammad Arsyad[1], Ariana Yunita[1*], Haniifah Mas'uudah Krismartopo[2], Aghnia Syahputri Dimar[2], Kartika Dewi[2], Iktri Madrinovella[2]

[1]Department of Computer Science, Universitas Pertamina, Jakarta 12220, Indonesia
[2]Department of Geophysics Engineering, Universitas Pertamina, Jakarta 12220, Indonesia
*Corresponding Author: ariana.yunita@universitaspertamina.ac.id

Received: 15 October 2022
Accepted: 13 December 2022

Abstracts

Exploratory Data Analysis (EDA) is a critical approach in developing machine learning models because the goal is to summarize the main characteristics of the data, often with visual methods, before modeling. It is frequently used as a prerequisite for more advanced data analytics techniques. Earthquakes are one of the natural disasters that commonly happen worldwide and lead to many victims. Research on machine learning for predicting earthquakes has been conducted a lot in recent years. This is a preliminary study for understanding an earthquake dataset to reveal several insights. This study aims to perform EDA using a dataset available on Kaggle, the Earthquake dataset from 1965 until 2016. Using several libraries in Python for data visualization and correlation analysis, this study results that depth does not correlate with magnitude, and the most frequent earthquake happened in 2011. Recommendations for further research are to cluster the dataset using clustering algorithms, such as K-means and hierarchical clustering, and then classify using several classifier algorithms.

Keywords: Correlation analysis, data visualization, earthquake dataset, exploratory data analysis

Introduction

Earthquakes are common natural disasters that happen in Indonesia. It is aligned with Indonesia's complex geological features where tectonic plates meet. This natural calamity usually causes human victims in terms of economy, environment, health, and other factors. Currently, data science and artificial intelligence have been utilized in the context of natural disasters [1]. Implementing machine learning to predict earthquakes is promising for helping to understand this phenomenon and preventing victims [2][3].

Machine learning is one of many artificial intelligence branches that allows a program to learn, adapt, and improve its prediction accuracy using a set of codes [4]. Data understanding is a prerequisite step before beginning to create a predictive model. Exploratory Data Analysis (EDA) is like "taking a peek at data" and is commonly used for data understanding. EDA aims to familiarize with a new dataset and to find something interesting. Understanding a new dataset means knowing the number of features, missing values, and some statistical numbers. To find something interesting refers to revealing outliers and finding a correlation between features. The term EDA was coined by John W. Tukey in 1977.

Several previous studies also conducted EDA, such as Purwoningsih et al. used an online learning dataset of Open University in Indonesia to detect online learners' behavior using EDA. They used data visualization, correlational, and clustering analysis [5]. Another study used the Covid-19 dataset to visualize and gain insights using EDA [6]. Yunita et al. used data from multiple sources to create a dataset and apply EDA in the higher education dataset in Indonesia

[7]. Xiangrong used the same earthquake dataset from Kaggle and analyzed it using data science and statistical methods [8]. Another study used the same earthquake dataset [9] and conducted a time series analysis using the dataset. This study complements the previous research by explaining how to preprocess data and conduct correlational analysis.

In general, this study attempts to conduct EDA using the earthquake dataset available on Kaggle. However, this study specifically aims to find the most frequent earthquake from 1965 until 2016. Second, visualize geospatial analysis from the earthquake dataset, and third, find a correlation between each feature. Section 2 discusses the theoretical foundations; Section 3 briefly explains the research method; Section 4 shows the results, and the last section concludes and gives recommendations.

Theoretical Foundations

Feature correlation

Examining the correlation between features is one quick method for performing feature selection. There are several methods to explore the correlation between independent and dependent variables, which depends on the data types. For example, if dependent variables are numerical data, and independent variables are also numerical data, the method to find the correlation is Pearson product-moment correlation or Spearman's correlation [10].

Scatterplots can be used in complement to the techniques explained above to analyze the correlations between the features. Positive, negative, no correlation, strong correlation, weak correlation, and non-linear correlation are the categories associated with feature relationships [10]. If the scatterplots are sloping to the upper right, it indicates a positive correlation. Otherwise, if the scatterplots fall to the left, it means a negative correlation. Furthermore, if there is no pattern, we can conclude zero correlation. Weak or strong correlations can also be viewed through scatterplots. If the data spreads out, it indicates that the correlation is weak. Otherwise, if the data are close to each other, it suggests that the correlation is strong [10].

Data Visualization

Data visualization deals with developing, designing, and applying graphical representations of data to make it easier to make sense of the data. Data visualization is also known as scientific or information visualization [11]. Using images, graphs, charts, and maps to understand data and information has been around for centuries. Computer advancements have made it possible to process vast amounts of data quickly. Today, data visualization is becoming a blend of art and science, and it will make a tangible difference in the years to come. Visualizing data can be complex, but it is much easier to understand data in a visual format than in large tables with text, numbers, and many rows and columns. Understanding data and its structure enable our data visualization techniques to be appropriately chosen.

All visualization techniques attempt to solve the same problem but in different ways. It has two data visualization categories for various purposes: description and exploration. Exploratory data visualization is useful when there is a large amount of data but little understanding of the data and a vague goal. Use descriptive data visualization when a lot of data is coming back, but you know what it is. Both categories are helpful for visually representing data [11].

There are a lot of visualization techniques based on the Python data analysis libraries, such as Matplotlib, Seaborn, ggplot, Plotly, PyQtGraph, VisPy, Bokeh, Altair, Pygal, Geoplotlib, Missingno, and Leather. Those libraries are currently popular and can be used for a wide range of dataset analyses. The visualization system also has the advantages of being simple to implement and having a high level of interactivity. GeoSeries and GeoDataFrame, which expand the capabilities of Series and DataFrames from pandas, are the primary data structures of geopandas. As the earthquake happened in specific locations, geopandas seems to be an appropriate library for visualizing earthquake.

Material and Methods

There are three steps to conduct this research: collecting, preprocessing, and performing exploratory data analysis through visualization and correlation analysis. The dataset in this study was retrieved from Kaggle [12], and then the data analysis used in this research is Exploratory Data Analysis. Google Colab used a platform to conduct EDA. The original dataset retrieved from Kaggle consists of 23412 rows and twenty-one columns.

Results and Discussion

Data Visualization

First, one of the aims of conducting EDA is to understand whether the dataset is imbalanced or balanced. In the dataset, there are four types of earthquakes: earthquake, nuclear explosion, explosion dan rock burst. Figure 1 shows a bar chart of earthquake types. The number of rows labeled as "Earthquake" was more than 20,000. Afterward, in the dataset, the types of an earthquake were replaced with earthquakes and not earthquakes. The earthquake was labeled as 1, and not earthquake was labeled as 0.

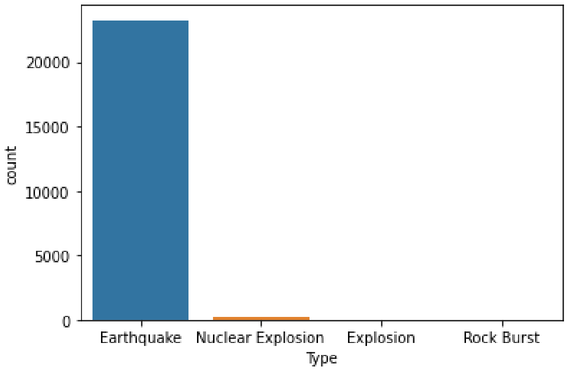


Figure 1. Type of Earthquake

Furthermore, a simple map was displayed to gain insights into the earthquake's location. Geopandas, a library in Python, was used to visualize the latitude and longitude to show the geospatial analysis. Using the 'naturalearth_lowres' dataset, which is coloring as blue, the earthquake is symbolized as red dots. It can be seen that earthquake commonly happens in the Pacific Ring of Fire.

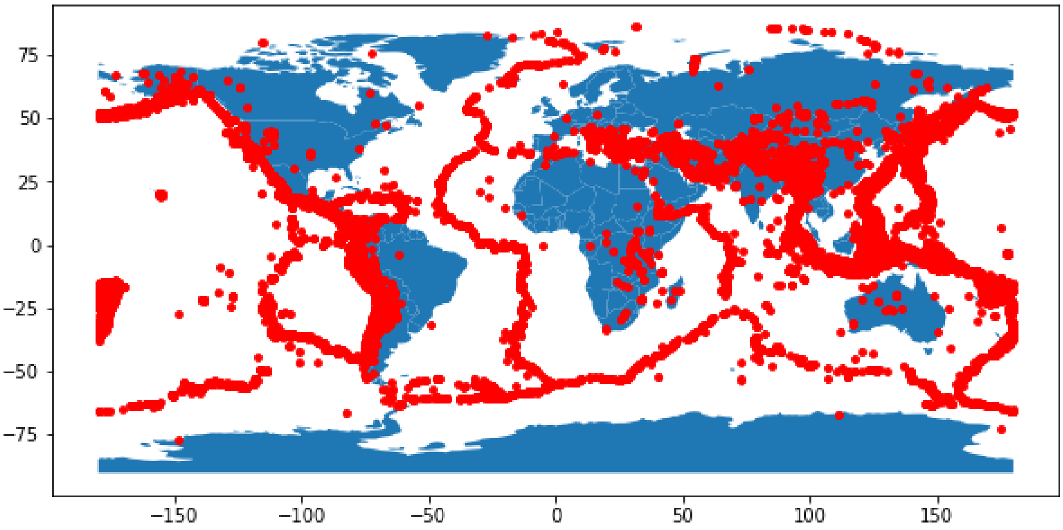
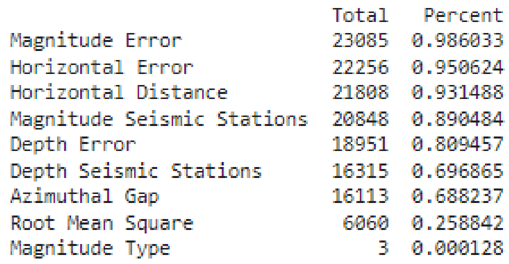


Figure 2. Two Dimension Analysis for Earthquake

Data Preprocessing

Missing values should be identified in the dataset. Below is the result of several missing values. Several missing values were found in several columns, such as Magnitude Error, Horizontal Error, Horizontal Distance, Magnitude Seismic Stations, Depth Error, Depth Seismic Stations, Azimuthal Gap, Root Mean

Square, and Magnitude Type. Of twenty-one columns, nine columns have missing values.



```
                            Total   Percent
Magnitude Error             23085  0.986033
Horizontal Error            22256  0.950624
Horizontal Distance         21808  0.931488
Magnitude Seismic Stations  20848  0.890484
Depth Error                 18951  0.809457
Depth Seismic Stations      16315  0.696865
Azimuthal Gap               16113  0.688237
Root Mean Square             6060  0.258842
Magnitude Type                  3  0.000128
```

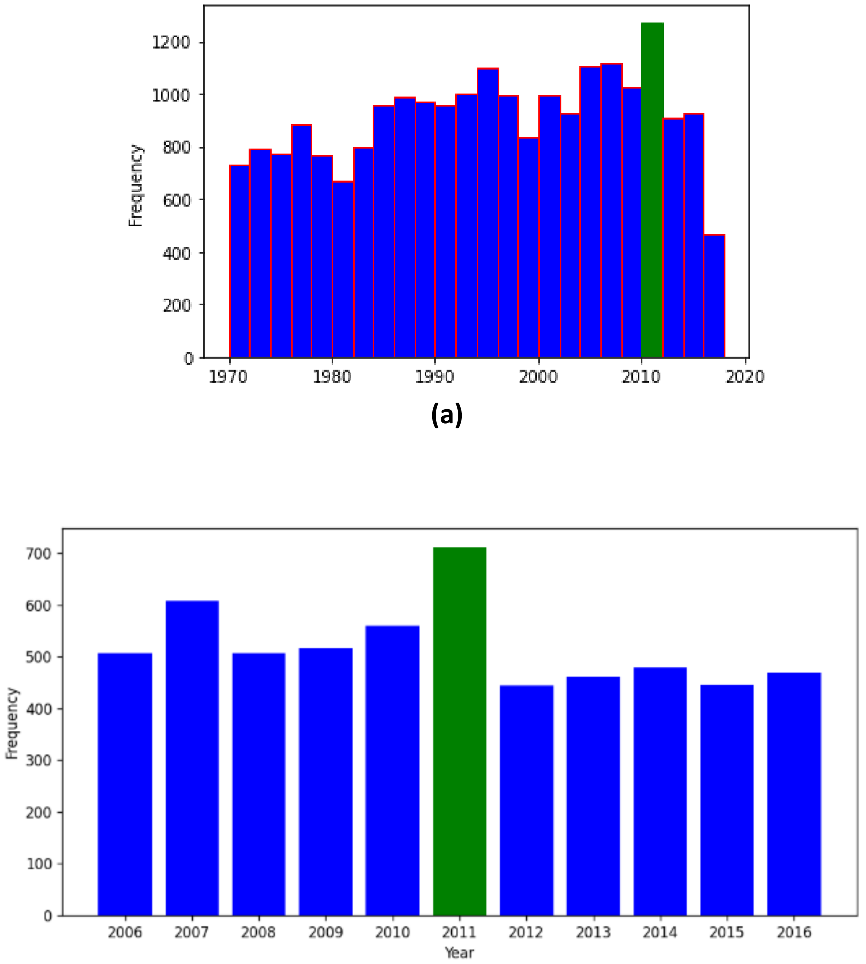
Figure 3. Percentage of Missing Values of Each Columns in the Earthquake Dataset

Another task in data preprocessing is to identify inconsistency of data. When the year of the data column was extracted, there was an error notification. It can be considered that there was inconsistency. The format should be '%m/%d/%Y', but another form was found. Therefore, the inconsistent data were replaced, as shown in Table 2.

Table 2. Inconsistency Data and The Replacement Value

Date	replaced with
1975-02-23T02:58:41.000Z	02/23/1975
1985-04-28T02:53:41.530Z	04/28/1985
2011-03-13T02:23:34.520Z	03/13/2011

After the values of the Year columns are consistent, the visualization can be built in Python. As seen, in Figure 4 a and b, the most earthquake occurrences in 2011 were 713. As reported by National Geographic [16], in March 2011, Japan experienced its most significant violent earthquake in recorded history. The earthquake beneath the North Pacific Ocean struck Sendai, the largest city in the Tohoku region, located north of the island of Honshu. In Tohoku, an earthquake caused a tsunami.



(a)

(b)

Figure 4. (a) Histogram of Earthquake Frequency from 1965-2016 (b) Bar Chart of Earthquake Frequency from 2006-2016

Feature Correlation Analysis

Knowledge of the correlations between these features is required to support further data analysis. We used Spearsman feature correlation because the Depth, Magnitude are numerical data. From Figure 5, the darker the color in the correlation matrix means no correlation. It can be seen that nuclear and the type of earthquake do not correlate. In addition, the variety of nuclear explosions and class also do not correlate.

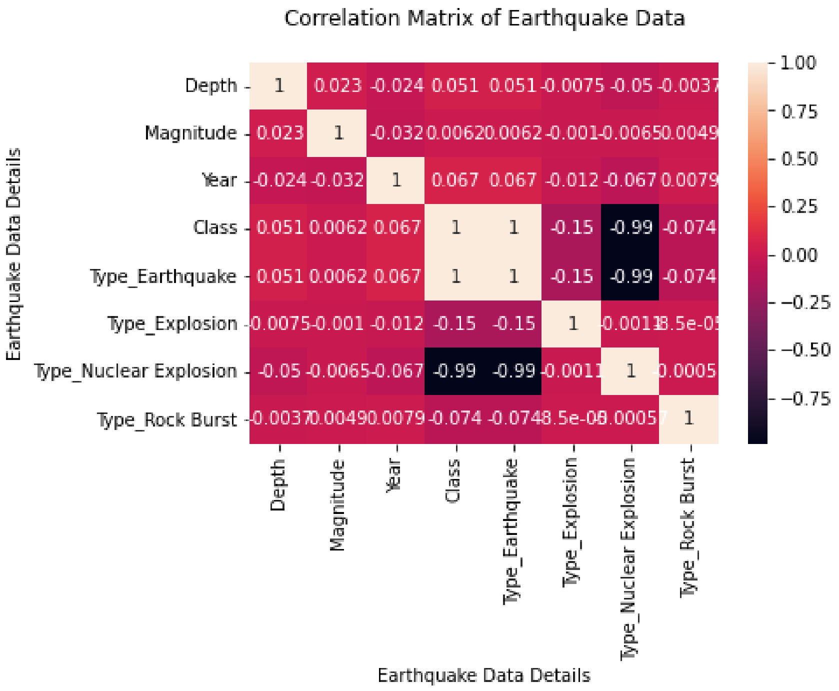


Figure 5. Correlation Matrix for Earthquake Dataset

Figure 5 is the result of the correlation matrix between features. In addition, a scatter plot is also made to show the correlation between magnitude and depth.

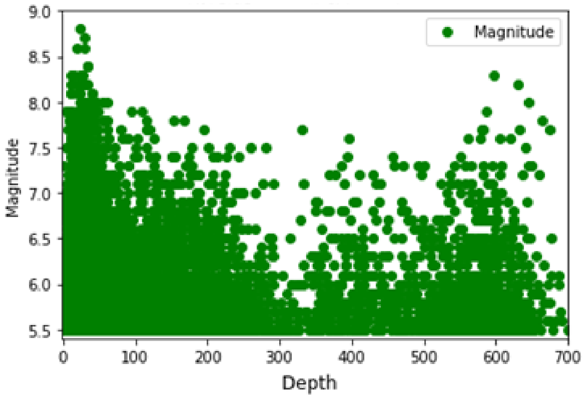


Figure 6. Scatter Plot of Depth and Magnitude

From the scatter plot, the two features (Magnitude and Depth) have a weak correlation. If an earthquake occurs in shallow depth with high magnitude, the affected area will be severely damaged. A formula that empirically explains how earthquakes range in size from small to large [14]:

$$log_{10}N = a - bM$$

N is the number of earthquakes with magnitudes greater than or equal to M, and M is the earthquake's magnitude. The distribution of earthquakes from minor to large is described by the parameters a and b; which physically a value indicates the level of seismicity and b value is related to the stress accumulation. Generally, b tends to be near 1. The smaller b-value means the energy of the tectonic plate motion is stored and has not been released. It is a logarithmic relationship, so plotting it with a logarithmic axis seems like a straight line. It shows the value of a = 0.9738 (medium level of seismicity) and b = 0.018 (high level of energy stored).

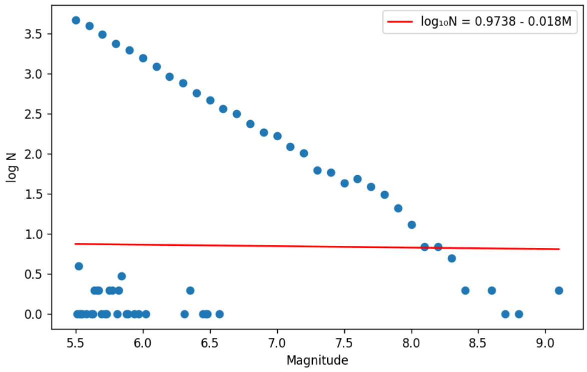


Figure 7. Scatter Plot of $\log N$ and Magnitude

Conclusion

This study demonstrates how to conduct EDA for earthquake dataset analysis. Earthquakes can be formulated as logarithmic relationship. Besides descriptive statistical analysis and data visualization, clustering algorithms can also be used to complete the Exploratory Data Analysis (EDA) process. Another recommendation is to use other earthquake datasets. Predictive analysis using several classifiers also can be conducted.

References

[1] D. A. Nurdeni, I. Budi, and A. Yunita, “Extracting Information from Twitter Data To Identify Types of Assistance for Victims of Natural Disasters: An Indonesian Case Study,” *J. Manag. Inf. \& Decis. Sci.*, vol. 25, 2022.

[2] Y. Xie, M. Ebad Sichani, J. E. Padgett, and R. DesRoches, “The promise of implementing machine learning in earthquake engineering: A state-of-the-art review,” *Earthq. Spectra*, vol. 36, no. 4, pp. 1769–1801, 2020.

[3] R. Tehseen, M. S. Farooq, and A. Abid, “Earthquake prediction using expert systems: a systematic mapping study,” *Sustainability*, vol. 12, no. 6, p. 2420, 2020.

[4] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. “ O’Reilly Media, Inc.,” 2022.

[5] T. Purwoningsih, H. B. Santoso, and Z. A. Hasibuan, “Online Learners’ Behaviors Detection Using Exploratory Data Analysis and Machine Learning Approach,” *Proc. 2019 4th Int. Conf. Informatics Comput. ICIC 2019*, no. Icic, pp. 1–8, 2019.

[6] J. DSouza, “Using Exploratory Data Analysis for Generating Inferences on the Correlation of COVID-19 cases,” in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2020, pp. 1–6.

[7] A. Yunita, H. B. Santoso, and Z. A. Hasibuan, “‘Everything is Data’: Towards one big data ecosystem using multiple sources of data on Higher Education in Indonesia,” *J. Big Data*, vol. 9, pp. 1–22, 2022.

[8] X. Xiangrong, “Visual Analysis of World Earthquakes based on Data Science and Statistical Methods,” *J. Phys. Conf. Ser.*, vol. 1684, no. 1, p. 12031, Nov. 2020.

[9] M. F. A. Azis, F. Darari, and M. R. Septyandy, “Time Series Analysis on Earthquakes Using EDA and Machine Learning,” in *2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2020, pp. 405–412.

[10] S. Kumar and I. Chong, “Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states,” *Int. J. Environ. Res. Public Health*, vol. 15, no. 12, 2018.

[11] P. Gandhi and J. Pruthi, “Data Visualization Techniques: Traditional Data to Big Data,” in *Data Visualization*, 2020, pp. 53–74.

[12] - US GEOLOGICAL SURVEY, “Significant Earthquakes, 1965-2016,” 2016. [Online]. Available: https://www.kaggle.com/datasets/usgs/earthquake-database. [Accessed: 12-Jun-2021].

[13] “Mar 11, 2011 CE: Tohoku Earthquake and Tsunami.” [Online]. Available: https://education.nationalgeographic.org/resource/tohoku-earthquake-and-tsunami.

[14] J. Russels, “Exploring Earthquake Magnitude and Depth,” 2020. [Online]. Available: https://jbrussell.github.io/eilive2020/part1a3_magnitudedepth/.