STAT-515 FINAL PROJECT SPRING 2024

2023 GLOBAL SEISMIC ANALYSIS

Suparna Mannava

G-Number: G01457969

smannav@gmu.edu

I. INTRODUCTION

Strong natural events like earthquakes can have catastrophic effects on infrastructure and human lives. It is essential for risk reduction, emergency response planning, and hazard assessment to comprehend the traits and trends of earthquakes. The United States Geological Survey (USGS) has released the Earthquakes 2023 Global dataset, which provides a thorough history of all the seismic activity that took place across the globe in 2023.

Numerous seismic event categories, including landslides, explosions, earthquakes, and volcanic eruptions, are included in the collection. Hazard assessment and risk mitigation require an understanding of the variables that affect the frequency and features of these various event types.

Using the Seismic 2023 Global dataset, I hope to learn more about the links and patterns found in the data for this project. Using diverse statistical and machine learning methodologies, I want to investigate the variables linked to seismic activity intensity, forecast event categories, and detect clusters of related seismic occurrences. The results of this investigation can help improve earthquake response and preparedness by advancing our understanding of seismic activity around the world.

II. DATASET

2.1 Data Source and Description

Data source: The data comes from the United States Geological Survey (USGS) and is available on Kaggle.

Number of variables: There are 22 variables in the dataset, including time, latitude, longitude, depth, magnitude, and various metadata fields.

Types of variables: The dataset contains geospatial and seismic event data. Key variables include:

Spatial: latitude, longitude, depth

Temporal: time of earthquake

Seismic: magnitude, magnitude type, number of stations, gap, distance, root-mean-square (RMS) travel time residual

Metadata: net, id, updated, place, type, horizontal error, depth error, magnitude error, magnitude stations, status, location source, magnitude source.

Volcanic eruptions, landslides, explosions, and earthquakes are among the events. Risk reduction and hazard assessment depend on an understanding of the variables linked to various event kinds.

2.2 Data Cleaning

The process of data cleaning was essential to guaranteeing the accuracy and consistency of the study. I started by getting rid of duplicate rows because they could distort the outcome. I made sure that every observation in the dataset was distinct and appropriately represented by eliminating these duplicates. I also handled missing values by adding zeros in their place. Although this method may not always be the best, particularly when dealing with numerical variables, it did enable the analysis to remain consistent. It is important to recognize, though, that zero values may not accurately represent the absence of data and may affect the outcome. Nevertheless, I reduced any potential biases in the analysis and improved the dataset's integrity by carrying out these cleaning procedures.

III. EXPLORATORY DATA ANALYSIS

Understanding seismic patterns and characteristics is aided by this approach, which offers insights on the distribution of earthquake magnitudes, their relationship to depth, and the link between various seismic factors. To gain insights into the Global Seismic 2023 dataset, the following exploratory analysis has been conducted:

Step 1: Load required libraries and read the dataset

```
# Load necessary libraries
library(dplyr)
library(readr)
library(lubridate)
library(ggplot2)
# Read the dataset
earthquakes <- read_csv("Global_Seismic_2023.csv")
```

Step 2: Data Cleaning and Preprocessing

# Check structure of the dataset

str(earthquakes)

# Convert 'time' column to datetime format

earthquakes$time <- ymd_hms(earthquakes$time)

# Check for missing values

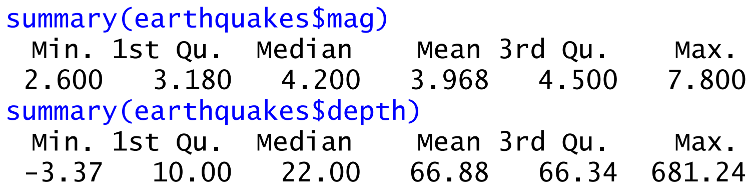
colSums(is.na(earthquakes))

Step 3: Summary Statistics and Visualizations

# Summary statistics

summary(earthquakes$mag)

summary(earthquakes$depth)



```
summary(earthquakes$mag)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.600   3.180   4.200   3.968   4.500   7.800 
summary(earthquakes$depth)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  -3.37   10.00   22.00   66.88   66.34  681.24 
```

The 'mag' (magnitude) and 'depth' summary statistics offer a clear overview of these continuous variables' distribution within the earthquake dataset.

# Histogram of magnitudes

hist(earthquakes$mag, main = "Distribution of Earthquake Magnitudes", xlab = "Magnitude")

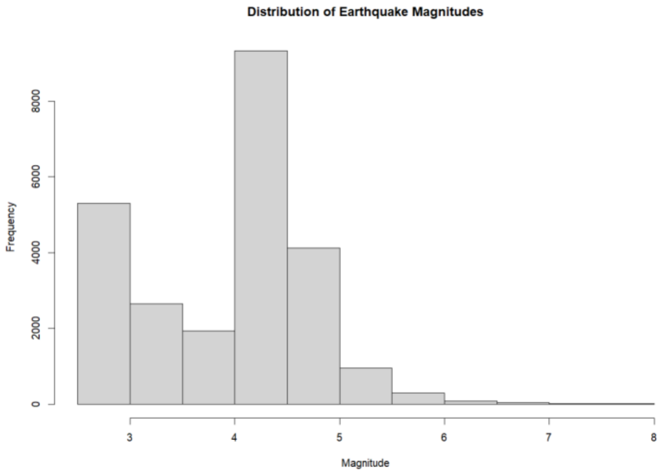


Fig. 3.1

The histogram's y-axis displays the quantity of earthquakes that fall into each magnitude bin, while the x-axis displays the magnitude of the earthquakes.

```
# Scatter plot of depth vs. magnitude
plot(earthquakes$depth, earthquakes$mag, main = "Depth vs. Magnitude", xlab = "Depth", ylab = "Magnitude")
```

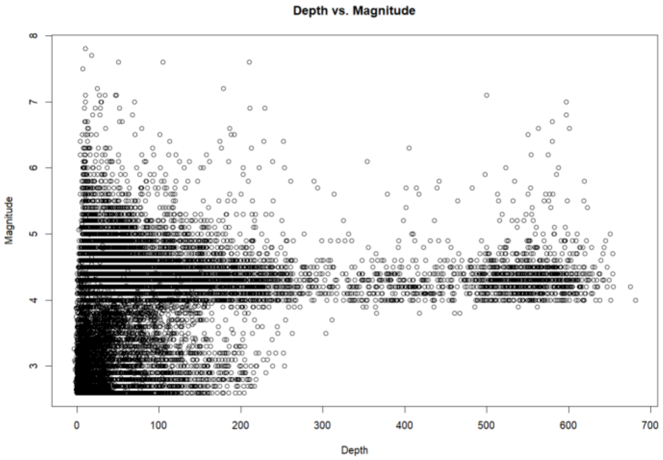


Fig. 3.2

The scatter plot demonstrates a weak negative link between the magnitude and depth of earthquakes. This indicates that there is a likelihood for stronger earthquakes to occur in shallower than deeper. There are numerous exceptions to this pattern, though, because the data is dispersed. Deeper earthquakes typically have smaller magnitudes. This is because it becomes harder for cracks to form and spread in rocks as one descends in depth due to an increase in pressure. This trend does, however, have several outliers. Deep earthquakes can occasionally be exceedingly powerful.

```
# Create boxplot for each type of event
ggplot(earthquakes, aes(x = type, y = mag)) +
  geom_boxplot() +
  labs(title = "Magnitude Distribution Across Different Types of Events",
       x = "Type of Event", y = "Magnitude") +
  theme_minimal()
```

The distribution of earthquake magnitudes across various event categories is displayed in the graph, which is a boxplot. The interquartile range (IQR) of magnitudes for each type of incident is shown by the box in each group. The median magnitude is shown by the line in the center of the box. The most frequent event type is an earthquake, which also typically has the biggest magnitudes. While earthquakes are the most frequent event type, volcanic eruptions often occur at significantly lesser magnitudes. The two least frequent occurrence types also have the tendency to be the smallest in magnitude: landslides and quarry blasts.

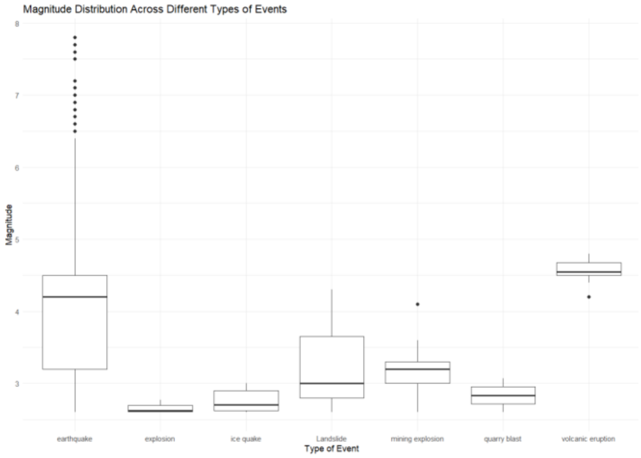


Fig. 3.3

```
# Extract year and month from the time column
earthquakes <- earthquakes %>%
  mutate(year_month = format(time, "%Y-%m"))
# Convert the time column to a date format
earthquakes$time <- as.POSIXct(earthquakes$time)
# Group the data by year and month, and calculate the mean magnitude
monthly_magnitudes <- earthquakes %>%
  group_by(year_month) %>%
  summarize(mean_magnitude = mean(mag, na.rm = TRUE))
# Convert the data to a time series object
monthly_magnitudes_ts <- ts(monthly_magnitudes$mean_magnitude, start =
c(year(min(earthquakes$time)), month(min(earthquakes$time))), frequency = 12)
# Plot the time series of monthly mean magnitudes
plot(monthly_magnitudes_ts,
     main = "Monthly Mean Earthquake Magnitudes",
     xlab = "Time",
     ylab = "Mean Magnitude",
    xaxt = "n")
# Add custom x-axis labels for years
axis(1, at = seq(1, length(monthly_magnitudes_ts), by = 12), labels =
unique(substr(monthly_magnitudes$year_month, 1, 4)))
```

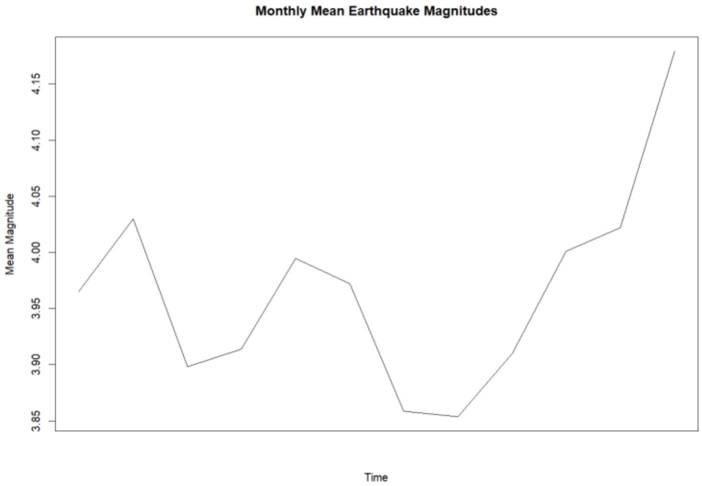
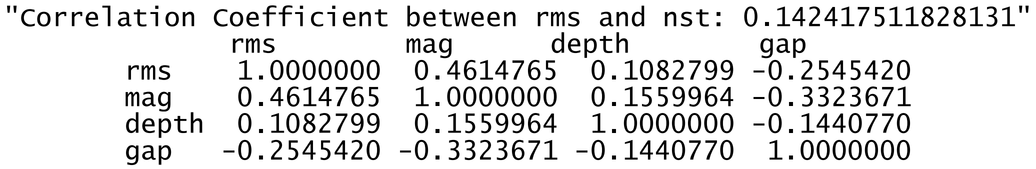


Fig. 3.4

The monthly mean earthquake magnitudes over time are displayed on the graph. The time is represented on the x-axis in years, while the mean magnitude is displayed on the y-axis.

```
# Calculate the correlation coefficient between rms and nst
correlation_coefficient <- cor(earthquakes$rms, earthquakes$nst)
# Print the correlation coefficient
print(paste("Correlation Coefficient between rms and nst:", correlation_coefficient))
# Select relevant variables
relevant_vars <- c("rms", "mag", "depth", "gap")
# Calculate the correlation coefficients
correlation_results <- cor(earthquakes[relevant_vars])
# Print correlation coefficients
print(correlation_results)
```



```
"Correlation Coefficient between rms and nst: 0.142417511828131"
              rms        mag      depth        gap
rms     1.0000000  0.4614765  0.1082799 -0.2545420
mag     0.4614765  1.0000000  0.1559964 -0.3323671
depth   0.1082799  0.1559964  1.0000000 -0.1440770
gap    -0.2545420 -0.3323671 -0.1440770  1.0000000
```

RMS and NST have a mild positive linear relationship, as seen by their correlation coefficient of roughly 0.142.

The following are the correlation coefficients between rms and other applicable variables:

rms and mag: around 0.461, which suggests a somewhat positive linear correlation.

depth and rms: around 0.108, suggesting a slender positive linear connection.

The rms and gap show a mild negative linear relationship at about -0.255.

The results imply that, among the variables chosen, mag (magnitude) has the strongest linear connection with rms.

IV. RESEARCH QUESTIONS

1. Can we predict the type of earthquake event (earthquake, explosion, landslide, etc.) based on the location, magnitude, and other features of the seismic event?

Decision trees and random forests are two methods that can be used to address this classification challenge.

To determine if an event is an earthquake, explosion, ice quake, landslide, mining explosion, quarry blast, or volcanic eruption, I used features such as depth, longitude, rms, and magnitude. With an accuracy of almost 99.94%, the random forest and decision tree models both demonstrate great accuracy in forecasting the types of earthquakes.

The decision tree first determines the event's depth. An earthquake is declared when the depth is higher than or equal to 0.005 kilometers. The tree advances to the following split if the depth is less than 0.005 km. The tree then takes the event's longitude into account. An earthquake is declared when the longitude is less than -112 degrees. The tree verifies the rms value if the longitude is more than -112 degrees. An ice earthquake is declared when the rms value is less than 0.38. An earthquake is declared to have occurred if the longitude is less than -134 degrees and the rms value is more than or equal to 0.38. If not, the incident is categorized as a mining explosion. The decision tree appears to perform well on this dataset based on its stated accuracy of 0.9994.

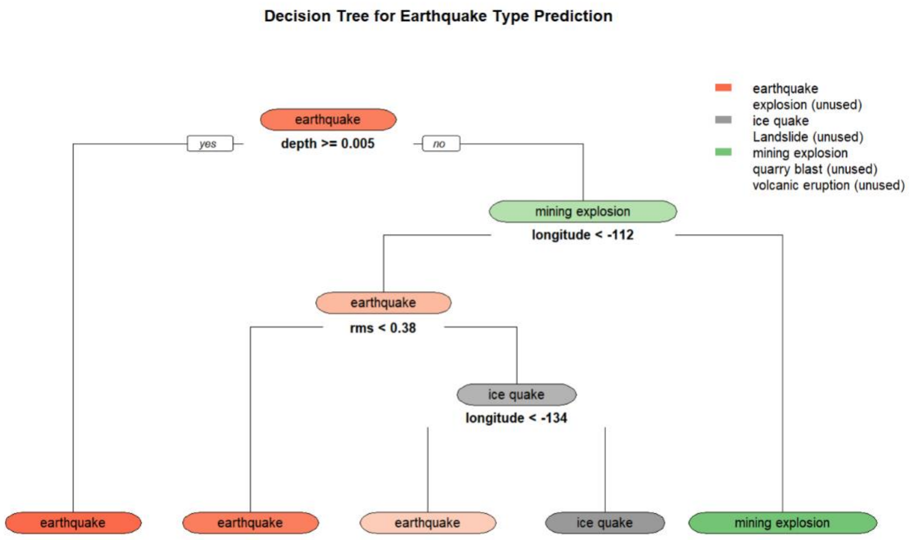
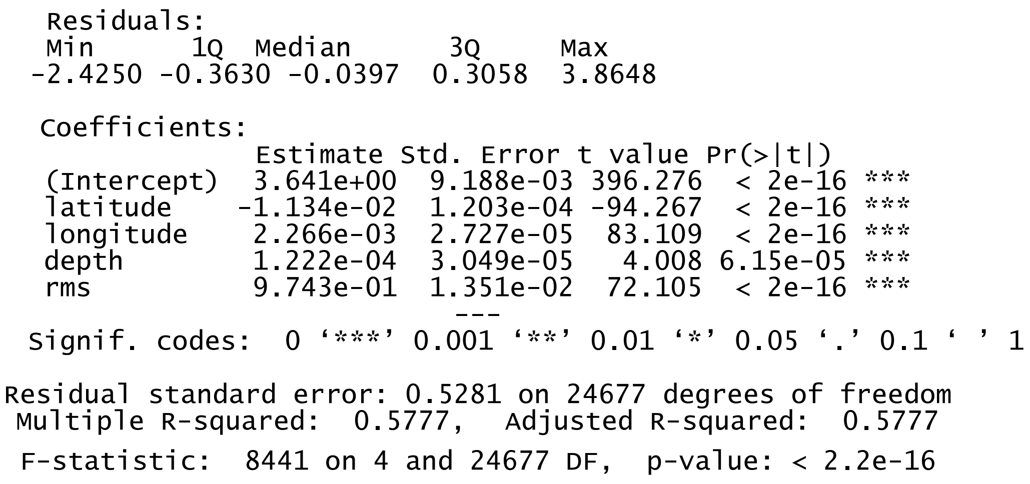


Fig. 4.1

2. **What factors (location, depth, magnitude, etc.) are most associated with the severity of an earthquake, as measured by the magnitude?**

Code:

```
# Research Question - 2

# Fit a multiple linear regression model

lm_model <- lm(mag ~ latitude + longitude + depth + rms, data = earthquakes)

# Summary of the regression model

summary(lm_model)

# Diagnostic plots

par(mfrow = c(2, 2))

plot(lm_model)
```



```
 Residuals:
 Min      1Q  Median      3Q     Max
-2.4250 -0.3630 -0.0397  0.3058  3.8648

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
 (Intercept)  3.641e+00  9.188e-03 396.276  < 2e-16 ***
 latitude    -1.134e-02  1.203e-04 -94.267  < 2e-16 ***
 longitude    2.266e-03  2.727e-05  83.109  < 2e-16 ***
 depth        1.222e-04  3.049e-05   4.008 6.15e-05 ***
 rms          9.743e-01  1.351e-02  72.105  < 2e-16 ***
                            ---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5281 on 24677 degrees of freedom
 Multiple R-squared:  0.5777,   Adjusted R-squared:  0.5777
 F-statistic:  8441 on 4 and 24677 DF,  p-value: < 2.2e-16
```

The results of the linear regression model indicate that all four features: latitude, longitude, depth, and rms are statistically significant predictors of earthquake magnitude (p-value < 2.2e-16). This means that these factors are all associated with the severity of an earthquake, as measured by magnitude. The model explains 57.77% of the variance in the magnitude of earthquakes (R-squared = 0.5777). The p-values for all the features are less than 2.2e-16, which means they are all statistically significant.

Latitude: Because of the negative coefficient for latitude, earthquakes typically have larger magnitudes when they occur at lower latitudes.

Longitude: Longitude has a positive, but very little, coefficient. Based only on this figure, it is impossible to determine the direction of the relationship between longitude and magnitude.

Depth: Deeper earthquakes typically have lesser magnitudes since the depth coefficient is positive. This is in line with how we understand the physics of earthquakes, since deeper rock experiences more pressure, which inhibits the formation and growth.

RMS: Because the root mean square (RMS) coefficient is positive, larger RMS values are typically associated with stronger magnitudes for earthquakes. RMS is a metric used to quantify how much an earthquake shakes the ground.

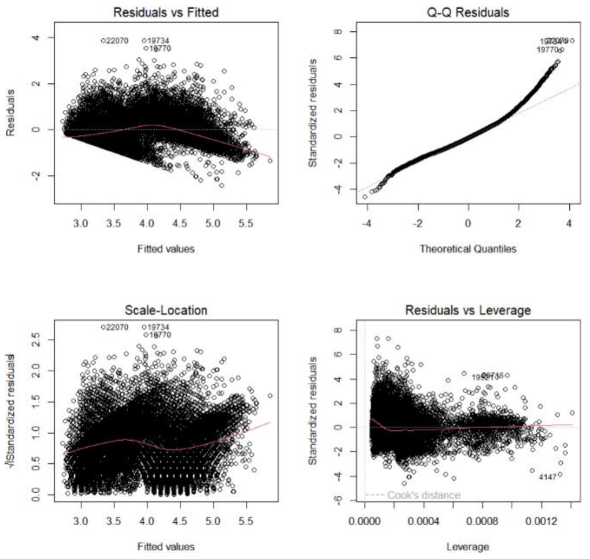


Fig. 4.2

Residuals vs. Fitted: This figure indicates a positive trend with residuals dispersed randomly around zero. The lack of apparent trends shows that there is no correlation between the errors and the fitted values.

Scale-Location: The residuals in this plot are similarly dispersed randomly around zero, which further suggests that the normality assumption is not met.

Residuals vs. Leverage: The relationship between residuals and leverage is examined in this figure. The impact of a data point in the regression analysis is gauged by its leverage. The residuals should ideally be dispersed randomly around zero, with no discernible leverage-based pattern. For sites with high leverage, the plot exhibits a minor rise in variance, which may lead to a possible heteroscedasticity (unequal variance of mistakes) problem.

Q-Q Plot: This plot examines the data to see if there are any non-linear relationships. There appears to be zero violation of the linearity assumption based on the random scatter surrounding the diagonal line.

The residual plots collectively demonstrate that the model satisfies most linear regression's presumptions. Although there may be a small heteroscedasticity problem, the model appears to match the data reasonably well overall.

The below image shows the residual plots from a linear regression model that predicts earthquake magnitude based on four features: latitude, longitude, depth, and rms.

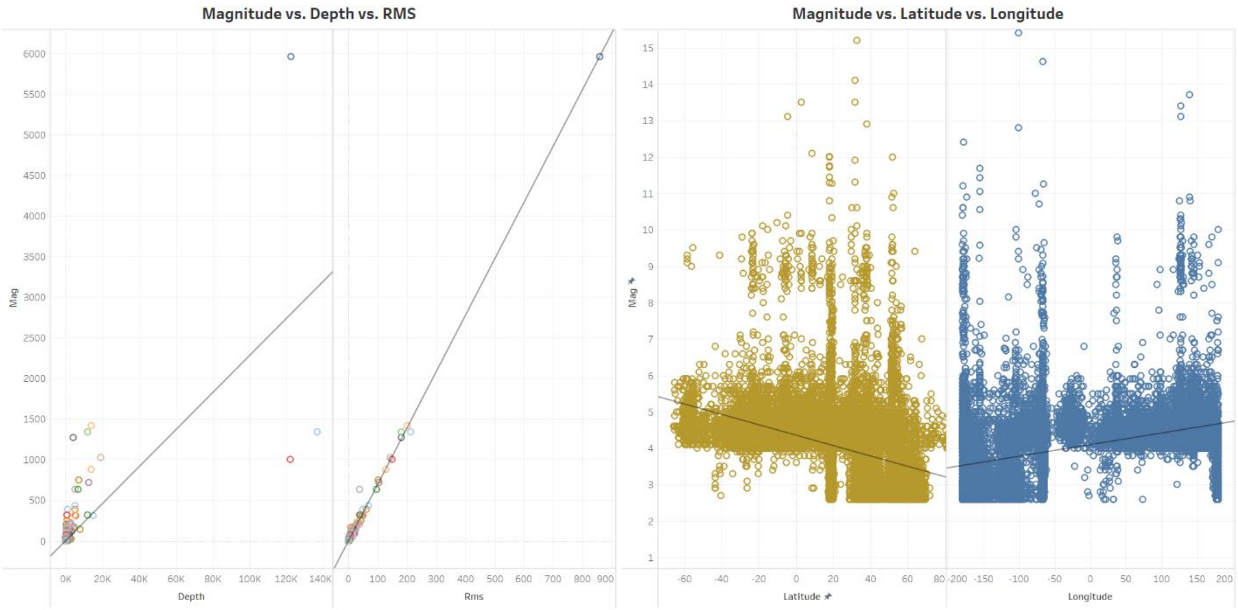


Fig. 4.3

3. **Can we cluster earthquakes into groups with similar characteristics (location, depth, magnitude, etc.)?**

Using earthquake data, the code written for this applies k-means clustering to find sets of earthquakes that share common features. Below is a summary of the functions of the code:

1. First, the original seismic data is filtered to only include earthquakes (events with the type "earthquake").
2. Chooses relevant variables: Then, it chooses latitude, longitude, depth, and magnitude as the four pertinent features to be clustered.
3. Looks for values that are missing: It looks for missing values in these four aspects using the summary() function.
4. Performs k-means clustering: A clustering technique called K-means divides data points into a predetermined number of groups, or k. To increase the likelihood of discovering ideal clusters, the code executes the k-means algorithm with numerous random initializations (nstart = 25) and sets the number of clusters to 20 (centers = 20).
5. Initial cluster assignments are printed: head(K$cluster, n = 5) is used to print the cluster assignments for the first five earthquakes. Each earthquake is allocated a cluster number (between 1 and 20) based on which cluster center it's closest to in terms of the four features.

6. Outputs the quantity of seismic events inside every group: Next, it uses K$size to report the total number of earthquakes associated with each cluster. This contributes to our understanding of how earthquakes are distributed throughout the 20 clusters.
7. Prints centers of clusters: Each cluster's center points (K$centers) are printed. The average latitude, longitude, depth, and magnitude of the earthquakes ascribed to each cluster are represented by the cluster center.
8. Establishes a data frame with assigned clusters: The function takes the original earthquake data and adds the cluster numbers to a new data frame ("df"). Determines the median magnitude for each cluster: It creates a new column in the data frame called "medMag" and determines the median magnitude for every cluster.
9. Displays groupings on a map: Lastly, it creates a map with each earthquake colored according to the median magnitude of each cluster using the ggplot toolkit. This makes it easier to see spatially the typical locations of earthquakes with comparable characteristics (latitude, longitude, depth, and magnitude).
10. Creates a CSV file called "NewGlobalSeismic2023.csv" from the final data frame "df" that contains the cluster assignments and earthquake data. 'Tableau' is a software used for additional analysis and display of this file.

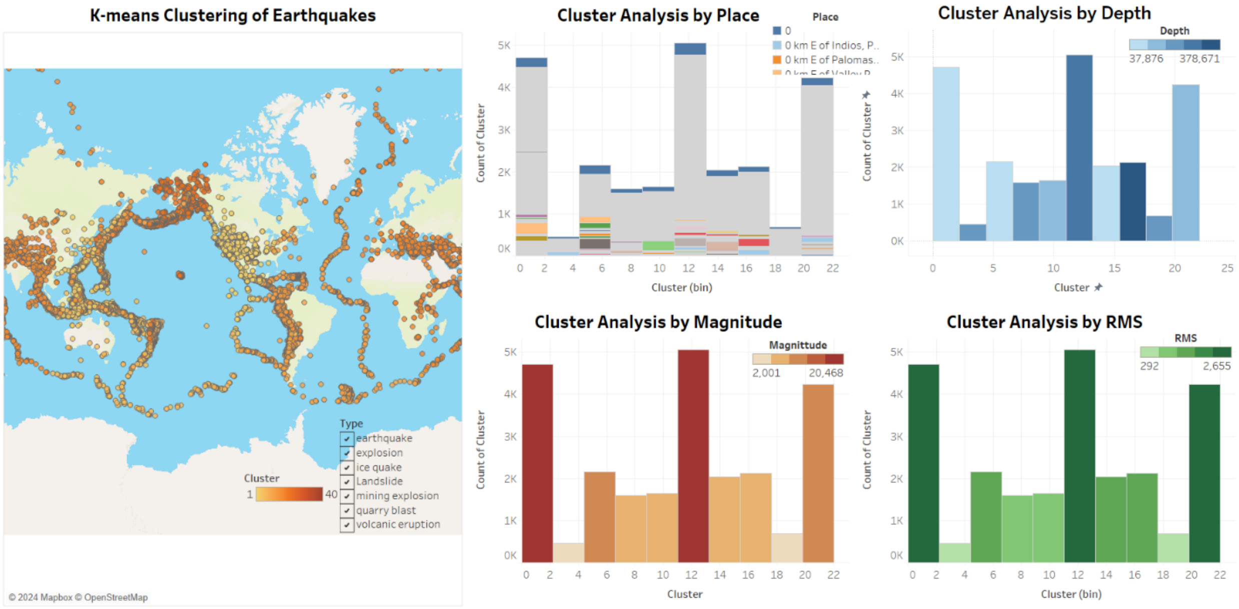


Fig. 4.4

The above dashboard is a cluster visualization of earthquakes based on their latitude, longitude, depth, magnitude, place and RMS. Each earthquake is represented by a point on the map, and the color likely represents the median magnitude of the cluster that earthquake belongs to. Darker colors likely indicate clusters with higher median magnitudes, while lighter colors indicate clusters with lower median magnitudes. This visualization helps identify regions where earthquakes tend to be stronger or weaker.

V. CONCLUSION

Throughout this project, I examined seismic event data to identify potential earthquake types and comprehend variables related to seismic intensity. Based on location, magnitude, and other characteristics, I was able to categorize earthquake types with high accuracy by using machine learning techniques like Random Forest and Decision Trees. Latitude, longitude, depth, and RMS values all strongly affect earthquake magnitude, according to multiple linear regression analysis, with bigger magnitudes being correlated with lower latitudes, higher longitudes, deeper depths, and higher RMS values. The distribution of earthquake magnitudes, depth-magnitude connections, and temporal changes were further clarified by exploratory data analysis. Subsequent investigations may examine supplementary factors and more advanced models to enhance comprehension and forecast precision, ultimately leading to improved approaches for catastrophe readiness and alleviation.

REFERENCES

[1] Earthquakes 2023 Global," www.kaggle.com. https://www.kaggle.com/datasets/mustafakeser4/earthquakes-2023-global

[2] "Geospatial Data Analytics: What It Is, Benefits, and Top Use Cases | SafeGraph," www.safegraph.com. https://www.safegraph.com/guides/geospatial-data-analytics

[3] "Dashboards," Tableau. https://www.tableau.com/learn/get-started/dashboards

[4] T. Zou et al., "An Overview of Geospatial Information Visualization," 2018 IEEE International Conference on Progress in Informatics and Computing (PIC), Suzhou, China, 2018, pp. 250-254, doi: 10.1109/PIC.2018.8706332.