University of Cape Town

STA5092Z

Exploratory Data Analysis

Assignment 2

Author:
Tinotenda Muzambi

Student Number:
MZMTIN002

March 13, 2024

Department of Statistical Sciences Plagiarism Declaration form

A copy of this form, completed and signed, to be attached to all coursework submissions to the Statistical Sciences Department. Submissions without this form will NOT be marked.

COURSE CODE: STA5092Z

COURSE NAME: Exploratory Data Analysis

STUDENT NAME: Tinotenda Muzambi

STUDENT NUMBER: MZMTIN002

PLAGIARISM DECLARATION

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.
2. I have used a generally accepted citation and referencing style. Each contribution to, and quotation in, this tutorial/report/project from the work(s) of other people has been attributed, and has been cited and referenced.
3. This tutorial/report/project is my own work.
4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.
5. I acknowledge that copying someone else's assignment or essay, or part of it, is wrong, and declare that this is my own work.

Note that agreement to this statement does not exonerate you from the University's plagiarism rules (http://www.uct.ac.za/uct/policies/plagiarism_students.pdf).

Signature: TM Date: 13/03/2023

Contents

Introduction

Earthquakes are a natural phenomenon with profound geological and societal impacts. This analysis delves into comprehensive datasets of recorded earthquakes spanning the period from January 2nd, 1965, to March 17th, 2023. The data come from two separate sources:

- Significant Earthquakes, 1965-2016 which was obtained from Kaggle[9]. It contains measurements of recorded earthquakes (or earthquake-like phenomena) ranging from 2 January 1965 until 30 December 2016.
- query.csv which was generated from the USGS website[8], and contains measurements from 1 December 2016 until 17 March 2023.

Data Wrangling

Because the data are from two distinct sources, there was no consistent naming scheme between the datasets. In order to standardise the variable names, all variables were converted to a common format (lowercase and with a period between words). Additionally variables that refer to the same property were renamed to be consistent between the two datasets.

Date-Time Parsing

Lubridate functions were used to parse the date/times from both datasets. To prevent redundant observations, entries were selectively retained in the "query.csv" dataset to dates strictly after December 30th, 2016. This filtering eliminated 243 observations, ensuring a temporally distinct subset for analysis alongside the Significant Earthquakes data.

Merging Datasets

At this point, the datasets were able to be merged. To ensure a cohesive and streamlined dataset suitable for further analysis, only variables present in both datasets were retained. This eliminated the inclusion of potentially irrelevant or incompatible variables that might have stemmed from differing data collection methodologies employed by the original sources.

Furthermore, to minimise inconsistencies within variable values, values of specific variables were transformed to either lowercase or uppercase. This standardisation process was particularly crucial for variables like location names or text descriptions, where minor inconsistencies in capitalisation could lead to erroneous interpretations during subsequent analysis.

Magnitude Ranges

Figure 1 illustrates the distribution of earthquake magnitudes in both datasets. The distribution exhibits a left-skewed pattern, indicating that earthquakes with lower magnitudes are more prevalent than those with higher magnitudes. Notably, the absence of earthquakes below a magnitude of five on the Richter scale suggests a threshold for reporting seismic events. The majority of reported earthquakes fall within the range of magnitudes between five and six on the Richter scale and all the earthquake magnitudes are within the normal range of Richter scale[11] magnitudes. As such, there is no need to exclude any measurements on the grounds of magnitude.

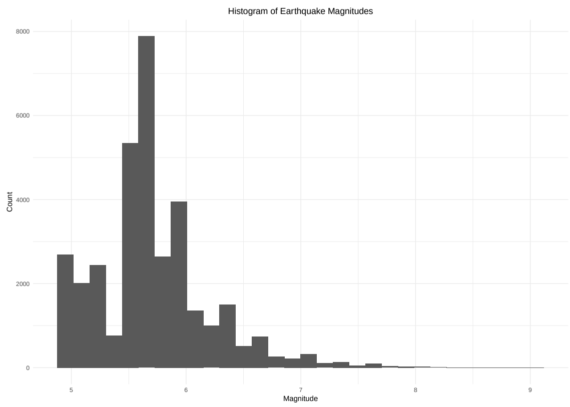

Figure 1: Histogram of Magnitude Distribution

Missing Data/Value Discrepancy Analysis

The majority of the variables have no missing data. Notable variables with missing data are magnitude.error (68.27%), distance.error (65.41%), distance (64.11%), depth.error (55.48%), azimuthal.gap (47.43%) and rms (17.74%).

Variables with a high proportion of missing data, such as magnitude.error, distance.error, and distance, are candidates for exclusion, as the missing data could lead to biased or unreliable results. Considering that this analysis does not make use of any of the variables with missing data, they can safely be excluded from the dataset. Status can also be excluded as it is not useful for this analysis.

The ranges of numeric variables were checked to ensure they are all within the acceptable range for that variable. All variables were found to be compliant.

Event Type Summary

Table 1 presents a summary of the various types of events recorded in the dataset, highlighting key metrics such as mean magnitude, mean depth, and count for each event category. It is evident that earthquakes constitute the predominant event type, comprising the vast majority of observations. However, a small number of other event categories, including explosions, volcanic eruptions, nuclear explosions, and a singular occurrence of a rock burst, are also documented within the dataset.

Event Type	Mean Magnitude	Mean Depth	Count
earthquake	5.71	65	33925
explosion	5.85	0	4
nuclear explosion	5.85	0	176
rock burst	6.2	1	1
volcanic eruption	5.29	0	55

Table 1: Table of Summary of Types of Events

Addition of Scale

A scale categorical variable was added which classifies the earthquakes according to their magnitude on this page[13]. This field is then converted into a factor to enforce the ordering of the various levels.

The final cleaned and merged dataset contained 33 925 observations of 12 variables.

Initial Exploration

Distribution of Features

Figure 2 shows the distribution of the scale variable that was added to the dataset. Similarly to figure 1, it is also left-skewed which aligns with expectations since scale is a categorical mapping of the magnitude. The majority of earthquakes are in the moderate category, which are 5.0–5.9 magnitude earthquakes which aligns with what was discovered with figure 1.

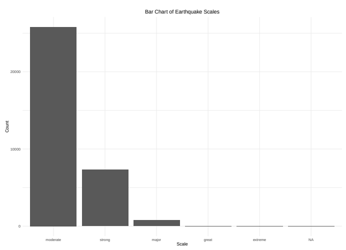

Figure 2: Bar Chart of Scale Distribution

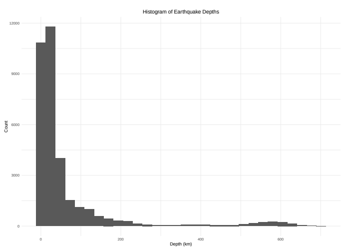

Figure 3: Histogram of Depth Distribution

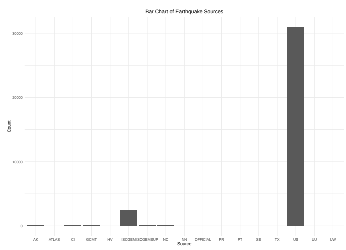

Figure 4: Bar Chart of Source Distribution

Figure 3 depicts the distribution of earthquake depths, representing the depth at which seismic ruptures originate[7]. The left-skewed bar chart indicates that the majority of earthquakes initiate at shallow depths, with a notable concentration between 0 and 25 kilometers.

Figure 4 presents the distribution of earthquake sources, identifying the primary data contributors or networks associated with seismic event reporting. The dominance of the US as the leading source, followed by ISCGEM and other minor contributors, underscores the significance of regional seismic monitoring networks in data collection and dissemination. Evaluating source distribution provides insights into data reliability, coverage, and potential biases.

Largest Earthquakes by Magnitude

Table 2 shows the three largest earthquakes by magnitude.

1. Indonesia Magnitude 9.1 Earthquake - This earthquake occurred along the Indian Ocean subduction zone triggering a massive tsunami that destroyed 800km of the coastal areas of Aceh Province. Post disaster damage and loss assessment revealed staggering numbers on the calamity that include over 220 000 human fatalities and the destruction of 139 000 houses.[12]
2. Japan Magnitude 9.1 Earthquake - This earthquake struck off the northeast coast of Honshu on the Japan Trench. A tsunami that was generated by the earthquake arrived at the coast within 30 minutes, overtopping seawalls and disabling three nuclear reactors within days. This resulted in over 18,000 dead, including several thousand victims who were never recovered.[5]
3. Chile Magnitude 8.8 Earthquake - This earthquake struck south-central Chile at 03:34 on Saturday, 27 February 2010, killing at least 799 people, triggering a tsunami and rattling buildings in the capital Santiago.[1]

Time	Latitude	Longitude	Depth	Magnitude	Scale
2004-12-26 00:58:53	3.295	95.982	30.0	9.1	extreme
2011-03-11 05:46:24	38.297	142.373	29.0	9.1	extreme
2010-02-27 06:34:12	-36.122	-72.898	22.9	8,8	great

Table 2: Table of Largest Earthquakes by Magnitude

Relationship between Depth and Magnitude

Figure 5 illustrates the scatter plot depicting the relationship between earthquake depth and magnitude. Upon visual inspection, the data points appear dispersed throughout the plot, indicating no clear linear correlation between depth and magnitude. However, the presence of an initial upward trajectory in the trendline suggests a potential association between shallower depths and higher magnitudes. This observation aligns with established seismological principles[10], as shallower earthquakes tend to produce stronger ground shaking due to their proximity to the Earth's surface.

Recognising the logarithmic nature of the Richter scale, figure 6 introduces adjustments for magnitude values to account for their exponential differences. By normalising magnitude values logarithmically, the scatter plot reveals a more compressed distribution of data points, reflecting the amplified magnitude differentials across the depth range. While this adjustment facilitates a more accurate representation of seismic energy release, it complicates the interpretation of patterns due to the expanded scale.

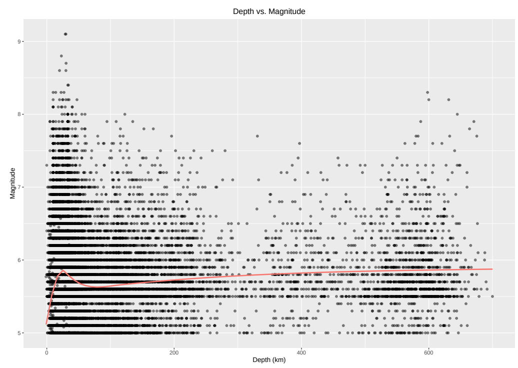

Figure 5: Scatter Plot of Depth vs Magnitude

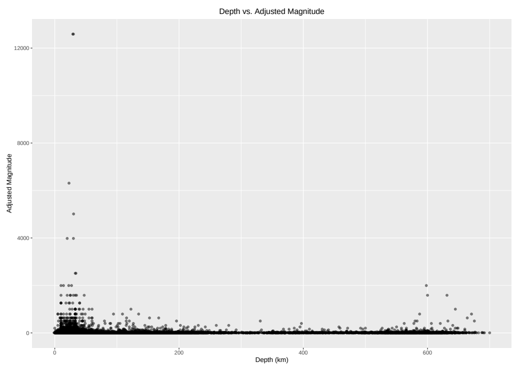

Figure 6: Scatter Plot of Depth vs Adjusted Magnitude

Nuclear Explosions

A nuclear explosion results from the detonation of a nuclear bomb, which induces a rapid release of energy in the form of intense heat, light, air pressure, and radiation[2]. These explosions, fueled by nuclear fission or fusion processes, produce devastating effects with far-reaching consequences for the surrounding environment and populations.

Nuclear explosion events have occurred at various locations across the globe. Figure 7 presents a map illustrating these sites, encompassing regions in Europe, Russia, North America, and the Pacific Ocean islands of Moruroa and Fangataufa. Notably, these events span a timeframe ranging from 1966 to 2017, reflecting historical and contemporary nuclear testing activities conducted by different nations[3].

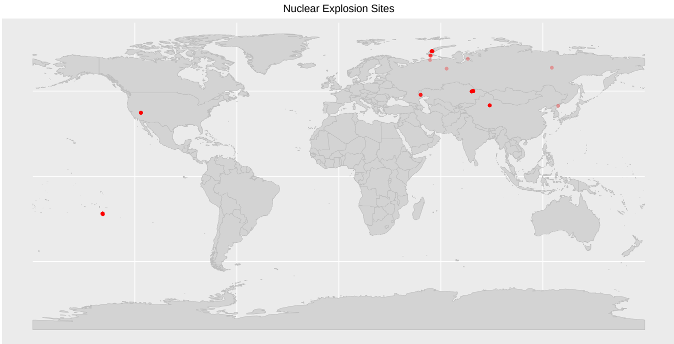

Figure 7: Map of Nuclear Explosion Sites

Temporal Investigation

Earthquake Frequency Comparison

Table 3 presents estimated earthquake frequencies globally[11], categorising seismic events into different magnitude classes ranging from micro to extreme. These estimates provide a benchmark for understanding the typical occurrence rates of earthquakes across varying magnitudes, serving as reference points for comparison with observed frequencies.

Description	**Average frequency of occurrence globally (estimated)**
Micro	Continual/several million per year
Minor	Over one million per year
Slight	Over 100,000 per year
Light	10,000 to 15,000 per year
Moderate	1,000 to 1,500 per year
Strong	100 to 150 per year
Major	10 to 20 per year
Great	One per year
Extreme	One to three per century

Table 3: Table of Estimated Earthquake Frequencies

Table 4 shows the observed earthquake frequencies. When comparing the two tables, we can see that the estimated frequencies match with what was observed in the dataset except for moderate earthquakes where fewer than the estimated number were observed. This discrepancy suggests potential under-representation or data limitations within the dataset for moderate seismic events.

Scale	Average per Year
moderate	437.37
strong	123.78
major	13.07
great	0.75
extreme	0.03

Table 4: Table of Observed Average Earthquake Frequencies

Patterns over Time

Figure 8 depicting earthquake magnitudes over individual years reveals temporal variations in seismic activity. Notable patterns may include fluctuations in the frequency and intensity of earthquakes over time, with certain periods exhibiting clusters of higher magnitude events. Observing lower magnitude earthquakes in recent decades may suggest shifts in seismic activity patterns or improvements in detection and reporting mechanisms.

Analysing earthquake magnitudes aggregated by decades in figure 9 provides a broader perspective on long-term seismic trends. Identifying patterns spanning multiple decades can offer insights into the cyclical nature of seismic activity, potential influences of geological processes, and the evolution of seismic monitoring capabilities over time.

While temporal patterns in earthquake magnitudes provide valuable insights into seismic activity dynamics, it is essential to acknowledge potential data artifacts that may influence observed trends. These artefacts could include changes in seismic monitoring networks, instrumentation upgrades, data recording practices, and reporting protocols. Additionally, factors such as population growth, urban development, and land use changes may impact the perceived frequency and intensity of seismic events over time.

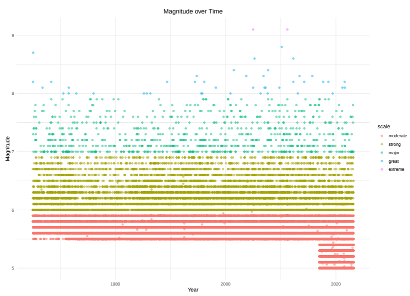

Figure 8: Scatter Plot of Magnitude vs Year

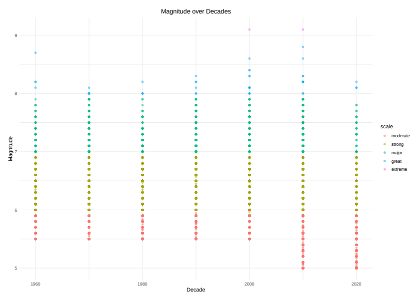

Figure 9: Scatter Plot of Magnitude vs Decade

Spatial Exploration

Figure 10 shows the earthquakes across the world with their magnitudes. Consistent with prior findings, we see that the majority of the earthquakes are low to moderate magnitude earthquakes. We also notice patterns of dense earthquake locations. This likely coincides with where tectonic plates connect and volcano location as these often are the cause of earthquakes.

Time	Latitude	Longitude	Depth	Magnitude	Scale
1967-07-22 16:57:00	40.751	30.8	30	7.4	major
1976-11-24 12:22:19	39.121	44.029	36	7.3	major
1999-08-17 00:01:39	40.748	29.864	17	7.6	major
1999-11-12 16:57:20	40.758	31.161	10	7.2	major
2011-10-23 10:41:23	38.721	43.508	18	7.1	major
2020-10-30 11:51:27	37.8973	26.7838	21	7	major
2023-02-06 01:17:34	37.2302	37.019	10	7.8	major
2023-02-06 10:24:49	38.008	37.2108	13.11	7.5	major

Table 5: Table of Major & Great Earthquakes in Turkey

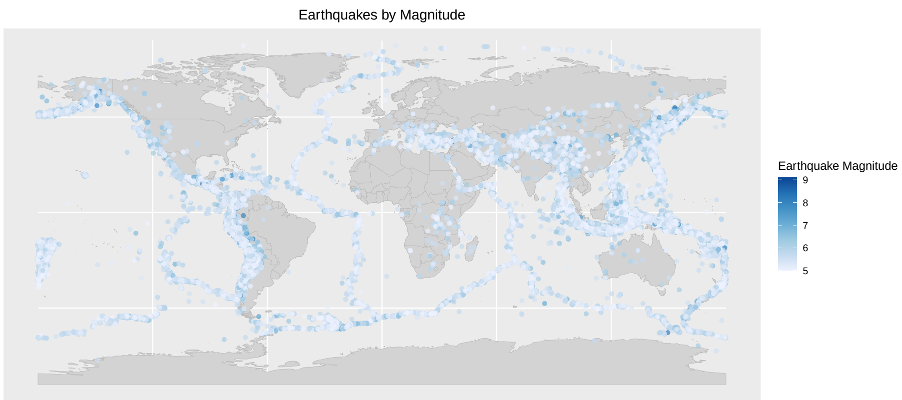

Figure 10: Map of Earthquakes by Magnitude

Turkey Earthquakes

The country of Turkey experienced a magnitude 7.8 earthquake on 6 February 2023 that devastated the southern region. Prior to this, Turkey experienced 198 earthquakes of which 152, 40 and 6 were moderate, strong and major respectively which can be seen in table 5 and visualised in figure 11.

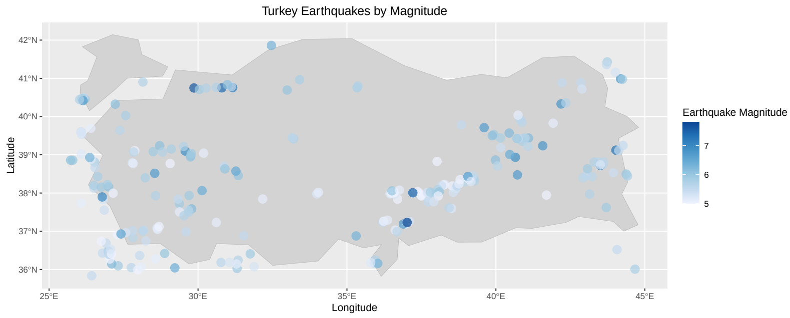

Figure 11: Map of Turkey Earthquakes by Magnitude

Table 5 shows the major earthquakes in Turkey and we can see that Turkey has only had eight of these previously spread out between 1967 and 2023. That equates to one major earthquake every seven years. The infrequent occurrence of major earthquakes in Turkey highlights the significance of each seismic event and underscores the need for comprehensive earthquake resilience and mitigation strategies.

South Africa Earthquakes

South Africa is characterised by relatively low seismicity compared to other tectonically active regions worldwide[4]. Despite this, sporadic seismic events have been recorded over the past decades, reflecting the presence of localised fault systems and geological complexities.

Table 6 provides a chronological overview of the earthquakes recorded in South Africa since 1965 (visualised in figure 12). With a total of 11 recorded events, these seismic occurrences are characterised by moderate magnitudes and shallow depths. The largest recorded earthquake in South Africa occurred on September 29, 1969, with a magnitude of 6.3. This seismic event, centered near the town of Tulbagh[6] in the Western Cape province, caused notable ground shaking and structural damage, highlighting the potential impact of moderate magnitude earthquakes on local communities and infrastructure.

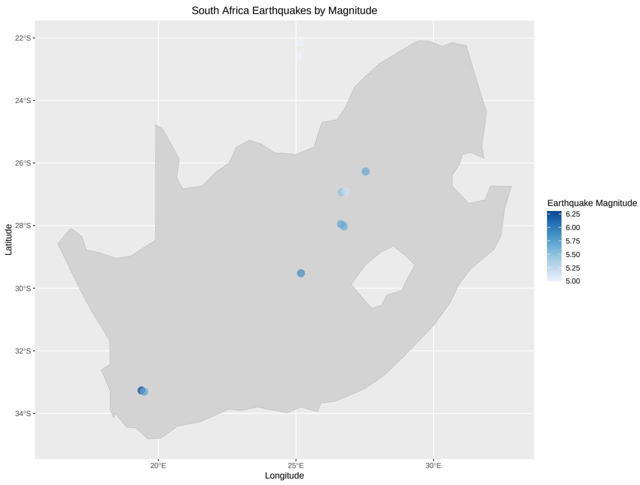

Figure 12: Map of South Africa Earthquakes by Magnitude

Time	Latitude	Longitude	Depth	Magnitude	Scale
1969-09-29 20:03:30	-33.268	19.386	15	6.3	strong
1970-04-14 19:08:19	-33.308	19.486	10	5.7	moderate
1974-05-14 06:51:15	-26.272	27.533	22	5.7	moderate
1976-07-01 11:24:05	-29.522	25.182	33	5.9	moderate
1977-04-07 11:54:37	-26.933	26.656	11	5.5	moderate
1989-01-25 10:14:33	-27.985	26.734	5	5.5	moderate
1994-10-30 06:06:27	-28.032	26.738	5	5.6	moderate
1999-04-22 22:19:37	-27.953	26.635	5	5.7	moderate
2017-04-05 00:55:50	-22.5646	25.0868	10	5	moderate
2017-04-03 17:50:16	-22.1421	25.1468	10	5	moderate
2017-04-03 03:08:51	-26.9092	26.8014	5	5.2	moderate

Table 6: Table of Earthquakes in South Africa

China Earthquakes

China, as one of the world's most populous countries and geographically diverse regions, experiences significant seismic activity. The analysis reveals approximately 2000 recorded earthquakes in China, primarily categorised as moderate and strong events, as shown in table 7, with a notable proportion of strong earthquakes accounting for approximately 23% of the total.

Figure 13 illustrates the spatial distribution of earthquakes in China, depicting both onshore and offshore seismic events. The presence of earthquakes occurring offshore underscores the maritime seismic hazards faced by coastal regions, highlighting the need for comprehensive earthquake monitoring and preparedness measures along China's coastline.

Scale	Average Magnitude	Count
moderate	5.481172	1365
strong	6.300706	425
major	7.269565	46

Table 7: Table of Summary China Earthquakes

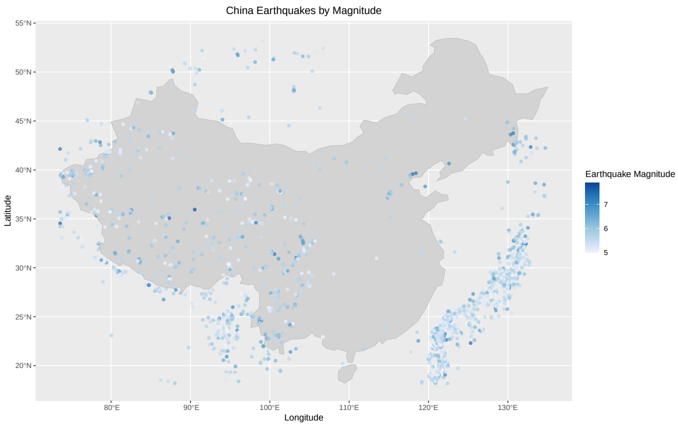

Figure 13: Map of China Earthquakes by Magnitude

Overlaying Population

Figure 14 extends the analysis by overlaying population density data onto the map of earthquake hotspots. This integration enables a comprehensive assessment of whether seismic activity coincides with areas of high human population density, thereby highlighting potential earthquake threats to densely populated regions.

By examining the map, it becomes evident that certain regions, notably China and India, exhibit both high seismic activity and dense human populations. This convergence of earthquake hotspots and large population centers underscores the heightened risk posed by seismic hazards to millions of people living in these areas.

While China and India stand out as prominent examples of regions where seismic hotspots coincide with large population centers, other countries such as the United States and Brazil also experience seismic activity near populated areas. Although these countries may have lower population densities compared to China and India, the potential impact of earthquakes on local communities remains significant.

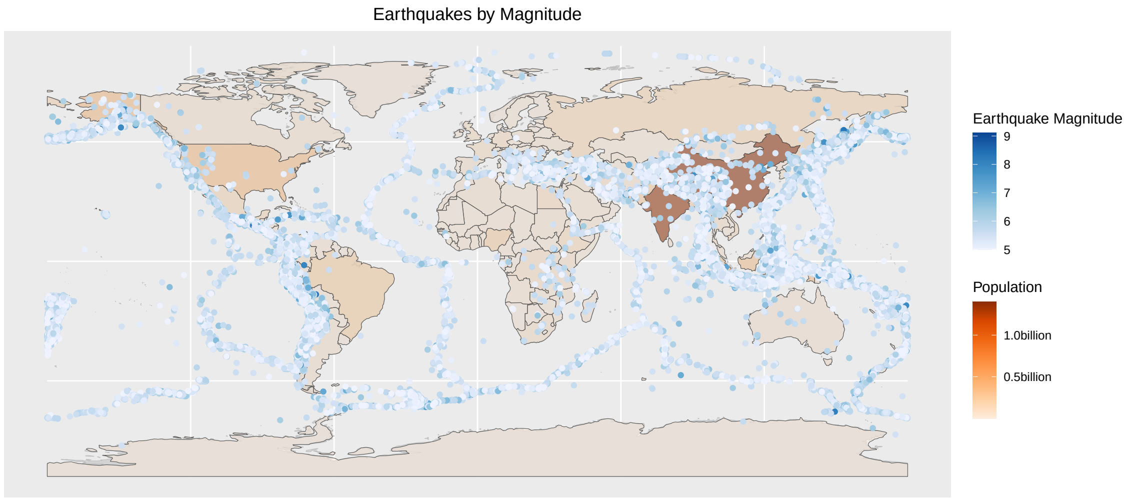

Figure 14: Map of Earthquakes by Magnitude with Population

Conclusion

The analysis of earthquake data spanning magnitude distribution, temporal patterns, depth-magnitude relationships, and regional seismic activity reveals insights into the dynamic nature of seismic hazards worldwide. The distribution of earthquake magnitudes, skewed towards moderate events, aligns with global seismicity patterns, with occasional extreme events.

Temporal trends indicate fluctuations in seismic activity over time, while the depth-magnitude relationship highlights the complex interplay between seismic depth and magnitude. Discrepancies between observed and estimated frequencies underscore potential data limitations or regional seismicity dynamics.

Overall, understanding seismic activity dynamics and their implications for human populations is crucial for prioritising resources and enhancing global resilience to seismic risks.

References

[1] ADRC. *Chile: Earthquake: 2010/02/27*. URL: https://www.adrc.asia/view_disaster_en.php?NationCode=152&Lang=en&Key=1369. Accessed: 2024-03-13.

[2] CDC. *Frequently Asked Questions About a Nuclear Blast*. URL: https://www.cdc.gov/nceh/radiation/emergencies/nuclearfaq.htm#:~:text=What%20is%20a%20nuclear%20blast,%2C%20air%20pressure%2C%20and%20radiation.. Accessed: 2024-03-13.

[3] Hanna Duggal and Mohammed Haddad. *Infographic: The impact of nuclear tests around the world*. URL: https://www.aljazeera.com/news/2022/8/29/infographic-what-is-the-impact-of-nuclear-tests-around-the-world-interactive. Accessed: 2024-03-13.

[4] Brassnavy Manzunzu et al. "Seismotectonics of South Africa". In: *Journal of African Earth Sciences* 149 (2019), pp. 271–279. ISSN: 1464-343X. DOI: https://doi.org/10.1016/j.jafrearsci.2018.08.012. URL: https://www.sciencedirect.com/science/article/pii/S1464343X18302620.

[5] NCEI. *On This Day: 2011 Tohoku Earthquake and Tsunami*. URL: https://www.ncei.noaa.gov/news/day-2011-japan-earthquake-and-tsunami. Accessed: 2024-03-13.

[6] South African History Online. *Earthquake shakes up Tulbagh*. URL: https://www.sahistory.org.za/dated-event/earthquake-shakes-tulbagh. Accessed: 2024-03-13.

[7] USGS. *ANSS Comprehensive Earthquake Catalog (ComCat) Documentation*. URL: https://earthquake.usgs.gov/data/comcat/#depth. Accessed: 2024-03-06.

[8] USGS. *Search Earthquake Catalog*. URL: https://earthquake.usgs.gov/earthquakes/search. Accessed: 2024-03-04.

[9] Abigal Larion USGS. *Significant Earthquakes, 1965-2016*. URL: https://www.kaggle.com/datasets/usgs/earthquake-database. Accessed: 2024-02-28.

[10] A.U. Vallina and E. Buforn. *Principles of Seismology*. Cambridge University Press, 2018. ISBN: 9781107138698. URL: https://books.google.co.za/books?id=x6o_DwAAQBAJ.

[11] SMS Tsunami Warning. *Richter Scale*. URL: https://www.sms-tsunami-warning.com/pages/richter-scale. Accessed: 2024-03-13.

[12] Prevention Web. *Recovery Collection: 2004 Indian Ocean Earthquake and Tsunami*. URL: https://recovery.preventionweb.net/collections/recovery-collection-2004-indian-ocean-earthquake-and-tsunami. Accessed: 2024-03-13.

[13] Wikipedia. *Richter scale*. URL: https://en.wikipedia.org/wiki/Richter_scale#cite_note-13. Accessed: 2024-03-03.