University of Cape Town

5092

EDA

STA5092Z-Exploratory Data Analysis Assignment 2

Author:
Dibanisa Fakude

Student Number:
FKDDIB001

March 15, 2024

Contents

Authorship Declaration

I, Dibanisa Fakude, declare that:

1. This research report and the work presented in it, is my own.
2. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.
3. These calculations/report/plans are 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.

Signature: D. Fakude

1

1.1 Introduction

The STA5092Z Exploratory Data Analysis Assignment 2 investigates earthquake data spanning from 1965 to March 2023, sourced from two distinct datasets. The assignment begins with data wrangling to merge the datasets, ensuring consistency and handling missing values. Initial exploration focuses on understanding feature distributions, identifying the largest earthquakes based on magnitude, and exploring the relationship between depth and magnitude. Subsequent temporal investigation compares observed earthquake frequencies with estimated frequencies and explores temporal patterns. Spatial exploration provides a summary of earthquake spatial distributions globally and in specific regions, including Turkey and Southern Africa. The assignment concludes with key findings summarized to enhance understanding of earthquake occurrences and patterns globally.

1.2 Data Wrangling

The data wrangling process involves several steps to prepare and combine earthquake data from two distinct sources. Initially, the earthquake data, obtained from ”Earthquakes 1965 - 2016.csv,” is read into R using ‘read.csv()‘ and converted into a tibble. The ”Date” and ”Time” columns are concatenated into a single ”Date_Time” column using ‘mutate()‘ and ‘mdy_hms()‘ functions from the ‘lubridate‘ package. Redundant columns are then removed using ‘select()‘and Column are renamed to match in both datasets. Similarly, the query data, sourced from ”query.csv,” is read and renamed to match earthquake data variables. The ”Type” column is capitalized using ‘str_to_title()‘, and the ”Date_Time” column is converted to date format using ‘ymd_hms()‘. Finally, the datasets are merged using ‘bind_rows()‘ to create a single dataframe named ”merged_data.” This comprehensive data wrangling ensures uniformity, consistency, and compatibility between the datasets, facilitating further analysis and exploration.

1.3 Question 1

Figure 1 shows the magnitude ranges for the datasets ,for the first dataset, magnitudes range from 5.5 to 9.1 on the Richter scale, indicating a broad spectrum of seismic events ranging from moderate to very strong earthquakes. Given this wide range, there seems to be no immediate justification for excluding measurements based solely on their magnitude range. Additionally, the magnitude ranges in both datasets are comparable. In the second dataset, magnitudes range from 5 to 8.2, starting from a slightly lower minimum magnitude compared to the first dataset. As with the first dataset, there's no apparent reason to exclude any magnitudes from analysis.

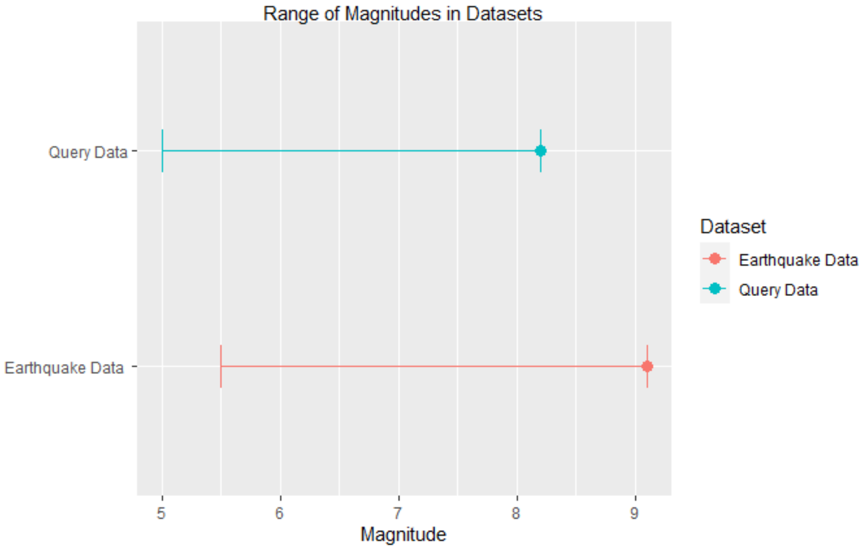

Figure 1: Range of Magnitudes for both datasets

The report provides an overview of missing data across common variables within the datasets. While some variables, such as Latitude, Longitude, Type, Depth, and Magnitude, demonstrate complete datasets with no missing values, others like Depth.Error, Depth.Seismic.Stations, Magnitude.Error, and Magnitude.Seismic.Stations exhibit substantial missing counts, suggesting potential data quality concerns. In addressing missing values, it's crucial to assess their impact on the analysis and the nature of missingness. In my analysis, I've determined that the missing values do not significantly affect the outcomes, so I've chosen to retain them with appropriate documentation. Additionally, discrepancies in values have been resolved through data wrangling efforts, ensuring consistency by aligning words in columns and capitalizing words in rows. This approach maintains data integrity and consistency, aligning with the specific requirements and objectives of the analysis.

Table 1: Missing Data Counts for Variables with Missing Values

Variable	**Missing Count**
Latitude	0
Longitude	0
Type	0
Depth	0
Depth.Error	18952
Depth.Seismic.Stations	27307
Magnitude	0
Magnitude.Type	0
Magnitude.Error	23382
Magnitude.Seismic.Stations	21060

The figure below shows the count of missing values in the datasets

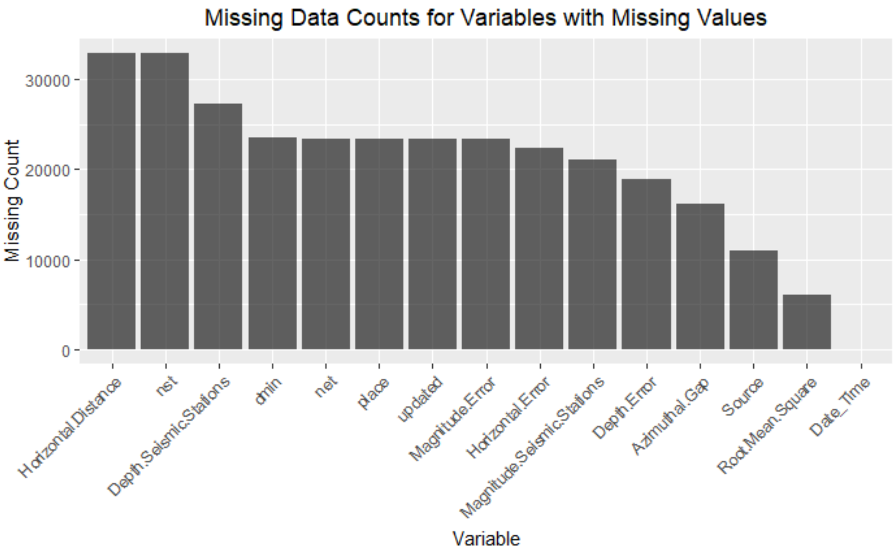

Figure 2: Count of zeros

The table below provides a summary of events categorized by type, including earthquakes, explosions, nuclear explosions, rock bursts, and volcanic eruptions. For each event type, the table displays the mean magnitude, maximum magnitude, minimum magnitude, and the number of events recorded. Earthquakes, the most common event type, exhibit a mean magnitude of 5.705 with a maximum magnitude of 9.1 and a minimum magnitude of 5.0, comprising the majority of events with 34,168 recorded occurrences. Explosions, nuclear explosions, and volcanic eruptions also feature in the dataset, each with varying magnitudes and frequencies. Notably, rock bursts have the least representation, with only one event recorded. Overall, the table offers insights into the distribution and characteristics of seismic events based on their type and magnitude.

Table 2: Summary of Events by Type

Type	**Mean**	**Max Magnitude**	**Min Magnitude**	**Number of Events**
Earthquake	5.705	9.1	5.0	34168
Explosion	5.850	6.4	5.6	4
Nuclear Explosion	5.853	6.9	5.5	176
Rock Burst	6.200	6.2	6.2	1
Volcanic Eruption	5.289	5.8	5.0	55

2 Initial Exploration

The provided code generates histograms for two selected features, ”Magnitude” and ”Depth,” from the datasets. Upon analysis, it is observed that both histograms exhibit right-skewness, indicating that the majority of the data points are clustered towards the lower end of the feature range, with a few extreme values towards the higher end. This skewness suggests that the distributions of both ”Magnitude” and ”Depth” are positively skewed, with a tail extending towards higher values.

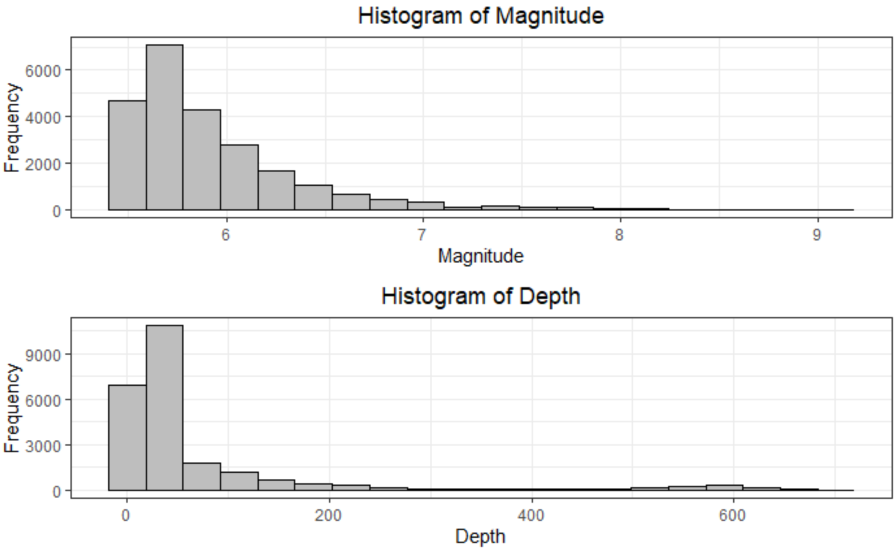

Figure 3: Distribution of Magnitudes

The largest earthquakes, as determined by their magnitude, represent seismic events of considerable significance due to their potential to cause widespread devastation and profound impacts on affected regions. By selecting a specific number of these significant earthquakes, we can gain insights into the characteristics and potential consequences of such seismic events. These earthquakes often result in extensive damage to infrastructure, loss of life, and environmental disturbances, underscoring the critical importance of understanding and monitoring seismic activity.

Table 3: Largest Earthquakes

Latitude	**Longitude**	**Magnitude**	**Scale**
3.295	95.982	9.1	Extreme
38.297	142.373	9.1	Extreme

The correlation coefficient of 0.0794 indicates a very weak positive correlation between depth and magnitude in seismic events. This implies that while there may be a slight tendency for deeper earthquakes to have larger magnitudes, the relationship is not strong. In other words, the depth of an earthquake alone does not provide a reliable predictor of its magnitude, and vice versa.

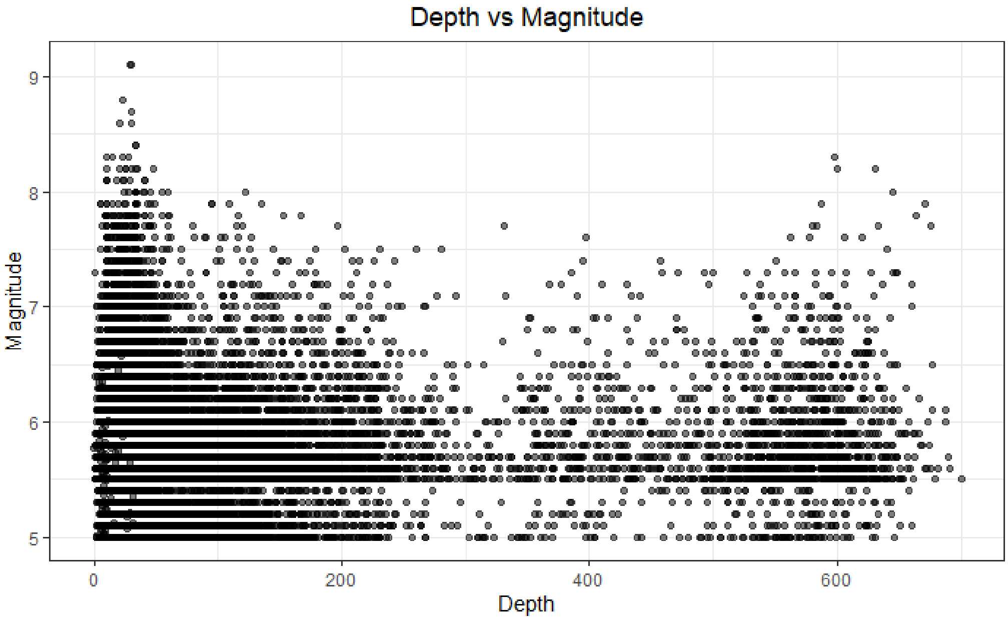

Figure 4: Depth Vs Magnitude

When adjusting the magnitude accordingly, it is essential to recognize that small changes in magnitude can have a substantial impact on the energy released by an earthquake. Therefore, when analyzing the relationship between depth and magnitude with adjusted magnitudes, the picture may change significantly. Adjusting for the logarithmic scale can highlight variations in the energy release associated with earthquakes of different depths

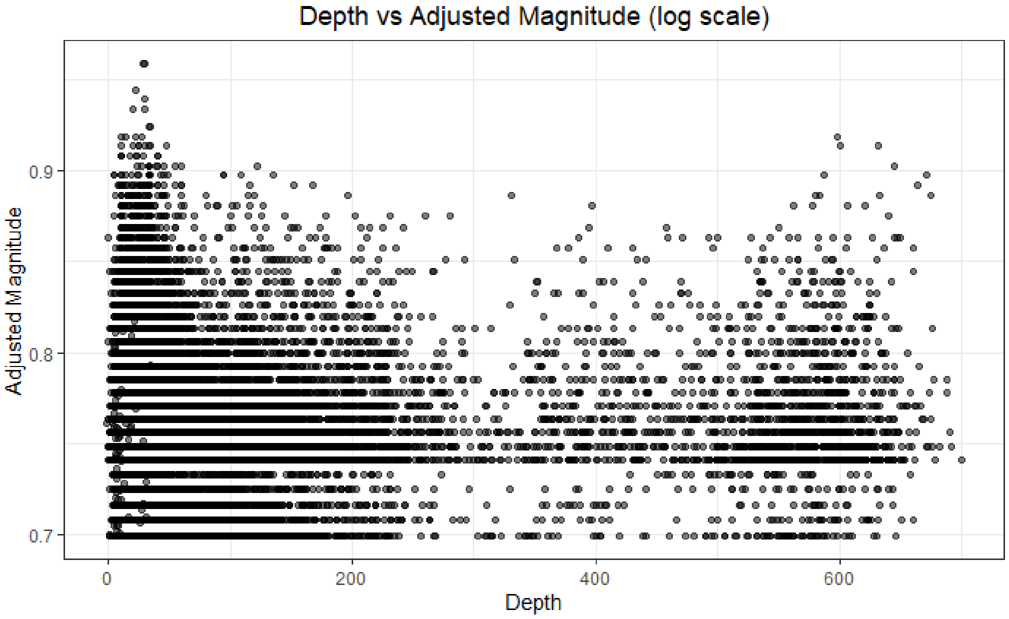

Figure 5: Depth Vs Adjusted Magnitude

In 2018, a series of volcanic eruptions occurred in Hawaii, contributing to significant geological activity in the region. These eruptions, marked by their occurrence in the Hawaiian archipelago, notably affected the island's landscape and local communities. To visualize the spatial distribution of these volcanic events, a map of Hawaii has been generated, pinpointing the locations of the eruptions

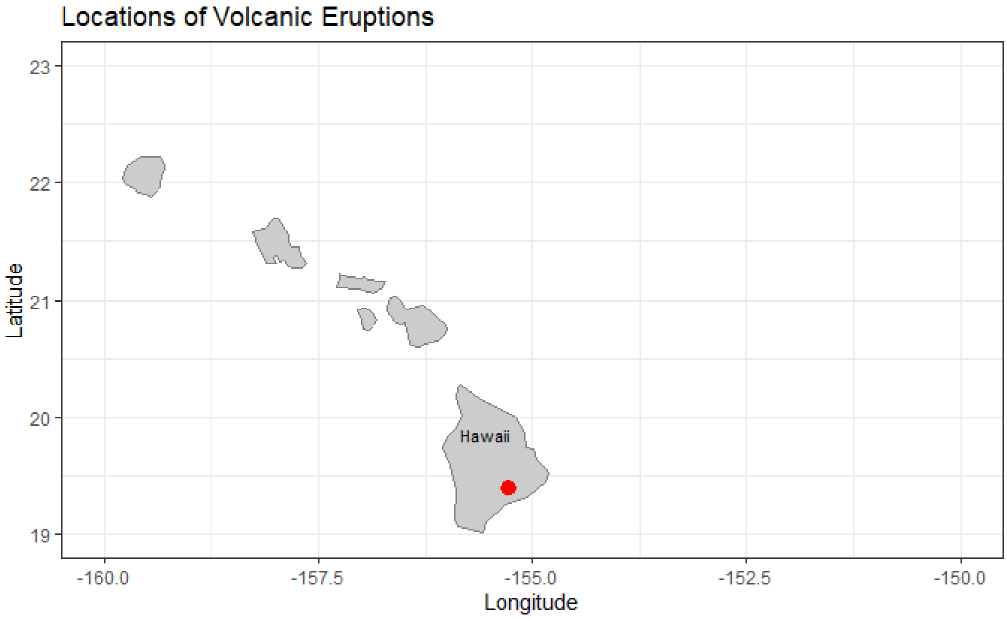

Figure 6: Location of volcanic eruption in Hawaii that took place in 2018

3 Temporal Investigation

The comparison between the frequency of earthquakes within specific magnitude ranges and their estimated frequencies offers insight into the relative occurrence rates of seismic events of varying intensities. According to data provided, earthquakes ranging from 1.0 to 1.9 on the Richter scale, categorized as ”Micro,” are estimated to occur continually or several million times per year. As the magnitude increases, the frequency decreases exponentially. For instance, earthquakes in the range of 2.0 to 2.9, classified as ”Minor,” are estimated to occur over one million times annually, whereas those in the 3.0 to 3.9 range, termed ”Slight,” occur over 100,000 times a year. The trend continues with diminishing frequency as magnitude increases, with ”Extreme” earthquakes (9.0–9.9) occurring only once to three times per century. These estimations provide a broad understanding of the relative occurrence rates of earthquakes across different magnitudes, highlighting the significantly reduced frequency of stronger seismic events compared to smaller ones.Table 4 shows the frequence(count) and Average magnitude of each catagory and Figure 6 visualizes the frequency of each catagory.

Table 4: Summary of Earthquake Magnitudes

Scale	Count	Average Magnitude
Extreme	2	9.100
Great	44	8.180
Major	774	7.285
Moderate	26,026	5.495
Strong	7,320	6.271
Unknown	2	5.955

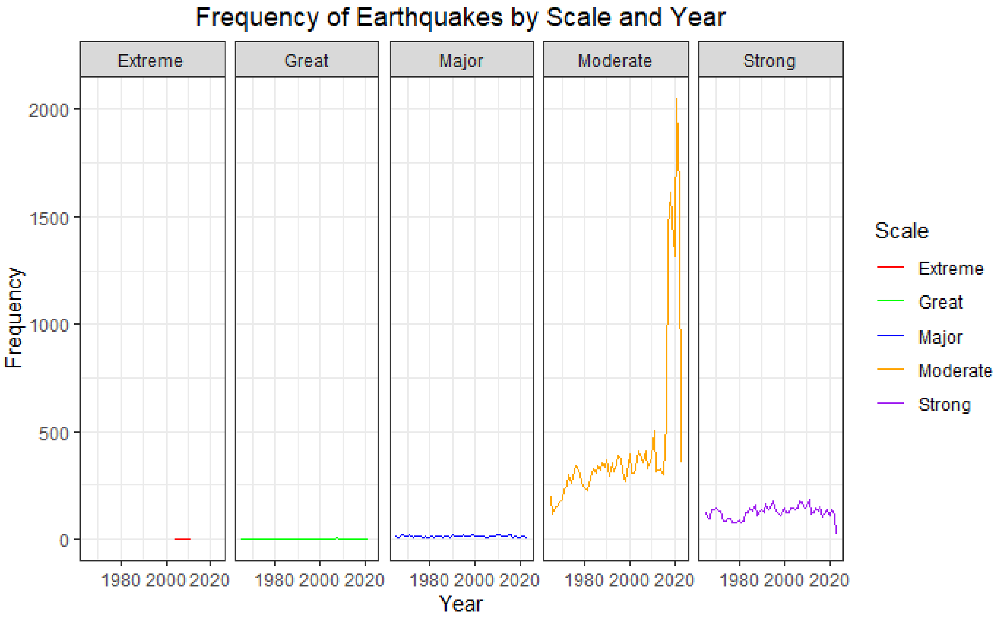

Figure 7: Frequency of Earthquakes over time by scale

4 Spatial Exploration

The devastating magnitude 7.8 earthquake that struck the southern region of Turkey on February 6, 2023, is a significant event in the context of seismic activity in the area. Comparing this earthquake with previous occurrences reveals that it falls within the ”Major” category based on its magnitude. Previous earthquakes in the region have varied in intensity, with magnitudes ranging from 5.5 for moderate earthquakes to 9.1 for extreme events. While the earthquake of February 6, 2023, ranks lower in magnitude compared to some historical events, its impact on the affected region is nonetheless substantial. They have frequently experienced 818 great and major earthquakes,

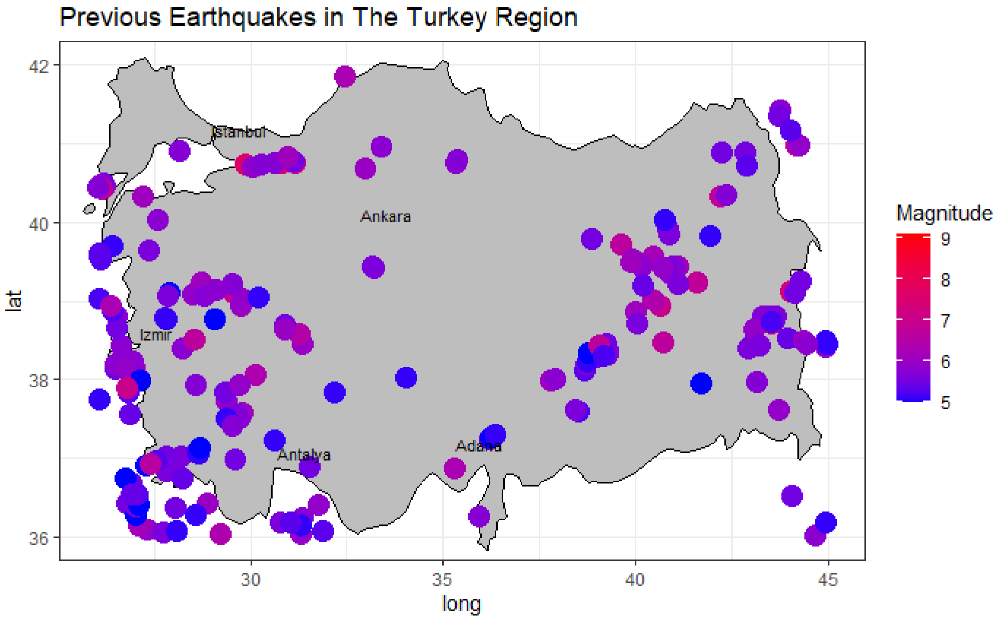

Figure 8: Previous Earthquakes in The Turkey Region

The analysis reveals that Southern Africa, including areas off the South African coast, has experienced a total of 34 earthquakes since 1965. The largest earthquake recorded in the region since then had a magnitude of 7, placing it in the "Major" category based on its intensity. This seismic event represents a significant occurrence in the region's seismic history. The plot visualizes the geographical distribution of these earthquakes, highlighting their magnitude with varying colors. Such data and analyses are crucial for understanding the seismic activity in Southern Africa and its potential impact on the region's communities and infrastructure.Figure 8 show the earthquakes distribution in the Southern Africa

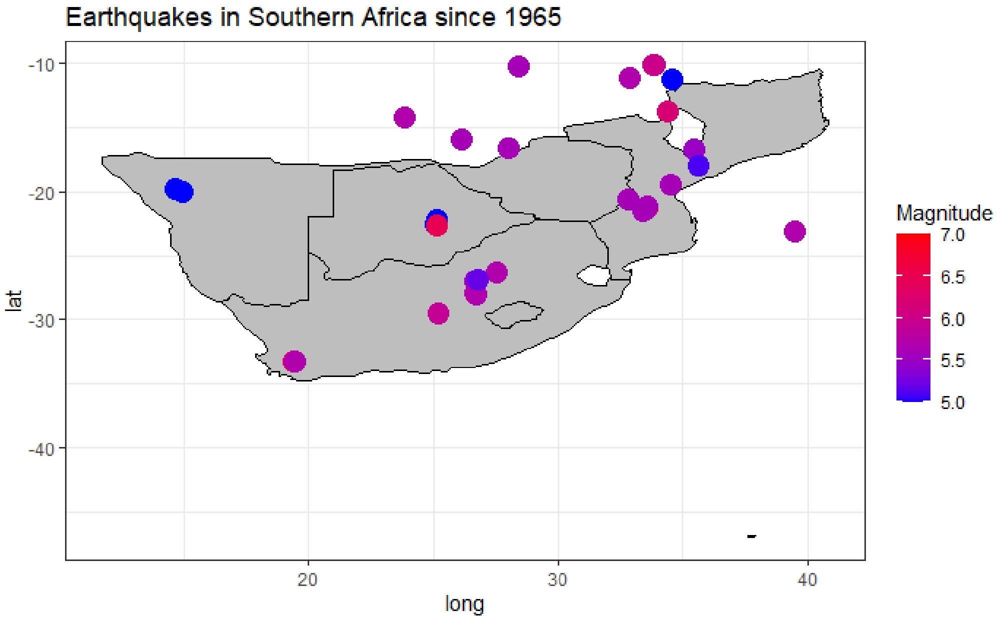

Figure 9: earthquakes distribution in the Southern Africa since 1965

4.1 Summary

Our examination provided valuable insights into the frequency, severity, and spatial spread of earthquakes. Through our analysis, we uncovered temporal trends and notable occurrences, enhancing our understanding of earthquake dynamics and aiding disaster preparedness initiatives. Additionally, we observed that moderate earthquakes occur more frequently than any other category. Notably, we discovered that the impact of earthquakes cannot be solely determined by magnitude. For instance, despite a magnitude 7.8 earthquake causing significant destruction, previous events with higher magnitudes in the same location had less severe effects.