San José State University

Department of Applied Data Science

DATA 230

Data Visualization

Fall 2023

Section 27

Instructor: Andrew H. Bond

Project Title: A Global Earthquake Analysis

Github Link:

Tableau Public Link:

https://public.tableau.com/app/profile/harshitha.lingala/viz/EQProject_17019240047540/EarthquakesDashboard?publish=yes)

Presented by:

Lingala Harshitha Reddy

SJSU ID – 017406545

Abstract

This report, "The Global Earthquake Analysis," presents a comprehensive analysis of earthquake data spanning over the years 1900 to 2023. The ‘Earthquakes Dataset 1900-2023’ is sourced from Kaggle, which provides information on seismic activities that occurred over the past century. It contains many key parameters like depth, magnitude, location, and time of occurrences. The key objective of this report is to find the locations where these earthquakes happen to be more concentrated, frequency of earthquakes in any given period, and the variances in the strength of these events. While this project is tiny proportion of what seismologists do with more comprehensive data in a more comprehensive manner, I believe could be a good starting point in crude analysis.

The project employs Python's pandas library for data exploration and cleaning, and Tableau for visualization, and analysis. It investigates various aspects such as frequency, magnitude distribution, depth, geographical distribution, and the relationship between depth and magnitude. The report also delves into historical trends, seasonal patterns, and the impact of man-made factors on seismic events.

The findings of this project offer valuable insights for real-time decision-making by various stakeholders, including governments, the general public, architects, and urban planners among other stakeholders. Governments can utilize the analysis and dashboard to identify high-risk areas and implement effective disaster management strategies, such as prompt emergency response and educate the residents of such areas on earthquake preparedness. Architects are encouraged to incorporate earthquake-resistant materials, such as steel or wood, while planning structures in these areas. While urban planning, care should be taken to prioritize low population density in these identified risk zones, and to accommodate emergency relief routes while planning.

TABLE OF CONTENTS

Introduction:

The 'Earthquakes Dataset 1900-2023.csv' provides insights into earthquakes that occurred over more than a century spanning almost 123 years. It contains information about each of these earthquakes that are represented by a unique ID. It has information about the depth, magnitude, place, latitude, longitude, and time of the occurrence apart from these it also has data regarding to magnitude type, nst, and dmin among others.

- ID – each event is represented with a unique ID.
- Depth - it is the distance between the hypocenter (earthquake's focus) and the earth's surface. It basically measures how the origin of the earthquake deep within the earth from the surface of the earth. It is also known as focal depth. It is a very important measure in analyzing and estimating the damage caused by an earthquake and while designing constructions, architects need to factor in the average depth of an earthquake in the region of construction.
- Magnitude – is the magnitude of the seismic waves that causes an earthquake. It basically denotes the earthquake's energy. It is measured across different scales but the most popular and widely followed is the Richter scale.
- Latitude and Longitude – gives the co-ordinates of the location of the occurrence of the earthquake event.
- Time – is the timestamp of the occurrence of earthquake.
- Mag Type – These indicate the magnitude scale that were used to measure the earthquakes.
- Place – has the details regarding the event occurrence near which country.
- Dmin – is the distance between the epicenter and its nearest seismic station, which gives us the accuracy about the location of the occurrence of the event.
- nst – represents 'Number of Stations' that recorded or detected a particular earthquake.
- Type – an earthquake is mainly a natural disaster caused by nature but in rare cases they are man-made disasters, like in the event of a nuclear explosion or dam explosions can trigger an earthquake.

Data Exploration:

For the initial understanding of the columns and data present in our dataset, used Pandas library of Python over Tableau Prep Builder for the usability and versatility.

Exploration of the Structure of the dataset:

```
import pandas as pd
```

```
df=pd.read_csv('Earthquakes Dataset 1900-2023.csv')
```

```
df.shape
```

```
(37331, 23)
```

```
df.columns
```

```
Index(['Time', 'Place', 'Latitude', 'Longitude', 'Depth', 'Mag', 'MagType',
       'nst', 'gap', 'dmin', 'rms', 'net', 'ID', 'Updated', 'Unnamed: 14',
       'Type', 'horizontalError', 'depthError', 'magError', 'magNst', 'status',
       'locationSource', 'magSource'],
      dtype='object')
```

```
df.isnull().sum()
```

```
Time                 0
Place              284
Latitude             0
Longitude            0
Depth              134
Mag                  0
MagType              0
nst              29858
gap              27244
dmin             32936
rms              17113
net                  0
ID                   0
Updated              0
Unnamed: 14      37331
Type                 0
horizontalError  33361
depthError       16504
magError         20780
magNst           31959
status               0
locationSource       0
magSource            0
dtype: int64
```

```
df.duplicated().sum()
```

```
0
```

Fig1.1. Data Exploration

Imported the dataset using ***pd.read_csv()*** method in pandas. Observed that there are in total 37331 records of earthquakes and 23 columns/parameters recorded about each earthquake event. We the see the all the column headers present in our dataset using ***df.columns*** , which lists all the names of our columns. The Earthquakes dataset has columns related to the occurrence of each earthquake event including depth, magnitude, place, latitude, longitude, and time of the occurrence apart from these it also has data regarding to magnitude type, nst, and dmin among other parameters. To arrive at cohesive analysis, we need to check for the presence of nulls. ***df.isnull().sum()***, checks for any missing values (NaN values) in all the columns present in the dataset and returns the sum of missing values in each column. We observed that columns, 'nst', 'dmin' , 'rms' among other have a huge number of missing values but since these are not our columns of interest, I chose to ignore and not impute these null values. Since the presence of duplicated records can hinder our analysis, I checked for their presence, if any using the ***df.duplicated().sum()*** method, which resulted 0, ie., none of our records are duplicated.

Data Cleaning:

```
: df['Unnamed: 14'].value_counts()

: Series([], Name: count, dtype: int64)

: df.drop(['Unnamed: 14','horizontalError', 'depthError', 'magError', 'magNst', 'status',
       'locationSource', 'magSource'],axis=1,inplace=True)

: df.columns

: Index(['Time', 'Place', 'Latitude', 'Longitude', 'Depth', 'Mag', 'MagType',
       'nst', 'gap', 'dmin', 'rms', 'net', 'ID', 'Updated', 'Type'],
      dtype='object')
```

Fig. 1.2. Data Cleaning

Observed a column named 'Unnamed: 14' which is empty, and other columns like 'horizontalError', 'depthError', 'magError', 'magNst', 'status', 'locationSource', 'magSource', which are of not much significance to our objective. Thus dropped these columns which were irrelevant.

```
df['Time']
```

```
0        2023-02-17T09:37:34.868Z
1        2023-02-16T05:37:05.138Z
2        2023-02-15T18:10:10.060Z
3        2023-02-15T06:38:09.034Z
4        2023-02-14T13:16:51.072Z
                  ...
37326    1902-01-01T05:20:30.000Z
37327    1901-12-31T09:02:30.000Z
37328    1901-12-30T22:34:00.000Z
37329    1901-03-03T07:45:00.000Z
37330    1900-10-09T12:25:00.000Z
Name: Time, Length: 37331, dtype: object
```

```
df["Time"] = pd.to_datetime(df["Time"]).dt.strftime('%Y-%m-%d %H:%M:%S')
```

```
df["Time"]
```

```
0        2023-02-17 09:37:34
1        2023-02-16 05:37:05
2        2023-02-15 18:10:10
3        2023-02-15 06:38:09
4        2023-02-14 13:16:51
                 ...
37326    1902-01-01 05:20:30
37327    1901-12-31 09:02:30
37328    1901-12-30 22:34:00
37329    1901-03-03 07:45:00
37330    1900-10-09 12:25:00
Name: Time, Length: 37331, dtype: object
```

```
df.dtypes
```

```
Time          object
Place         object
Latitude     float64
Longitude    float64
Depth        float64
Mag          float64
MagType       object
nst          float64
gap          float64
dmin         float64
rms          float64
net           object
ID            object
Updated       object
Type          object
dtype: object
```

```
df['Place'] = df['Place'].astype(str)
```

```
## Splitting the columns to form new columns with the desired data
```

```
for index, row in df.iterrows():
    df.loc[index, 'Date'] = row['Time'].split()[0]
    df.loc[index, 'time'] = row['Time'].split()[-1]
    df.loc[index, 'country'] = row['Place'].split()[-1]
```

Fig.1.3. Data Cleaning

When observed, Time column has the timestamp values of the occurrence of the event. For our analysis date of occurrence of the event is the most important parameter. So extracted the date from the Time column using ***pd.to_datetime()*** method and stored them in a new column, 'Date'. Also, when observed Place column has the location of the occurrence along with the distance in Kilometer from the nearest country for ex., '130 km SW of Tual, Indonesia' is one such entry. Country is also subject to our analysis so extracted the country by splitting the Place column and created a new column 'Country' to store all the countries.

Creating Parameters

Implemented parameters within Tableau named 'Earliest Date' and 'Latest Date' to facilitate date filtering through the creation of a field named 'Evaluation Period.' The earliest date corresponds to the initial recorded earthquake occurrence within the dataset, which is 10-09-1900. The latest date in the dataset is 02-31-2023; however, considering that the dataset does not encompass the entire year of 2023, limited occurrences to only those up to 12-31-2022 for better analysis.

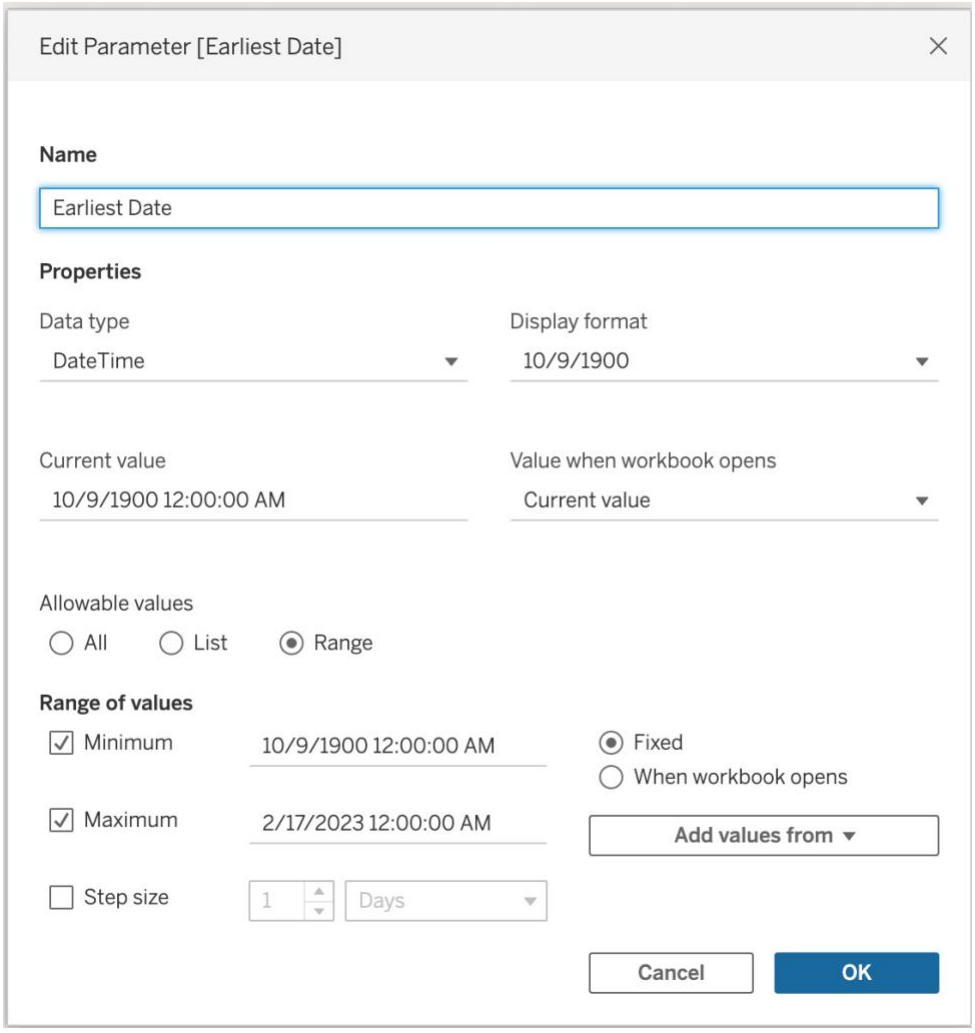

Fig 2.1. Earliest Date Parameter

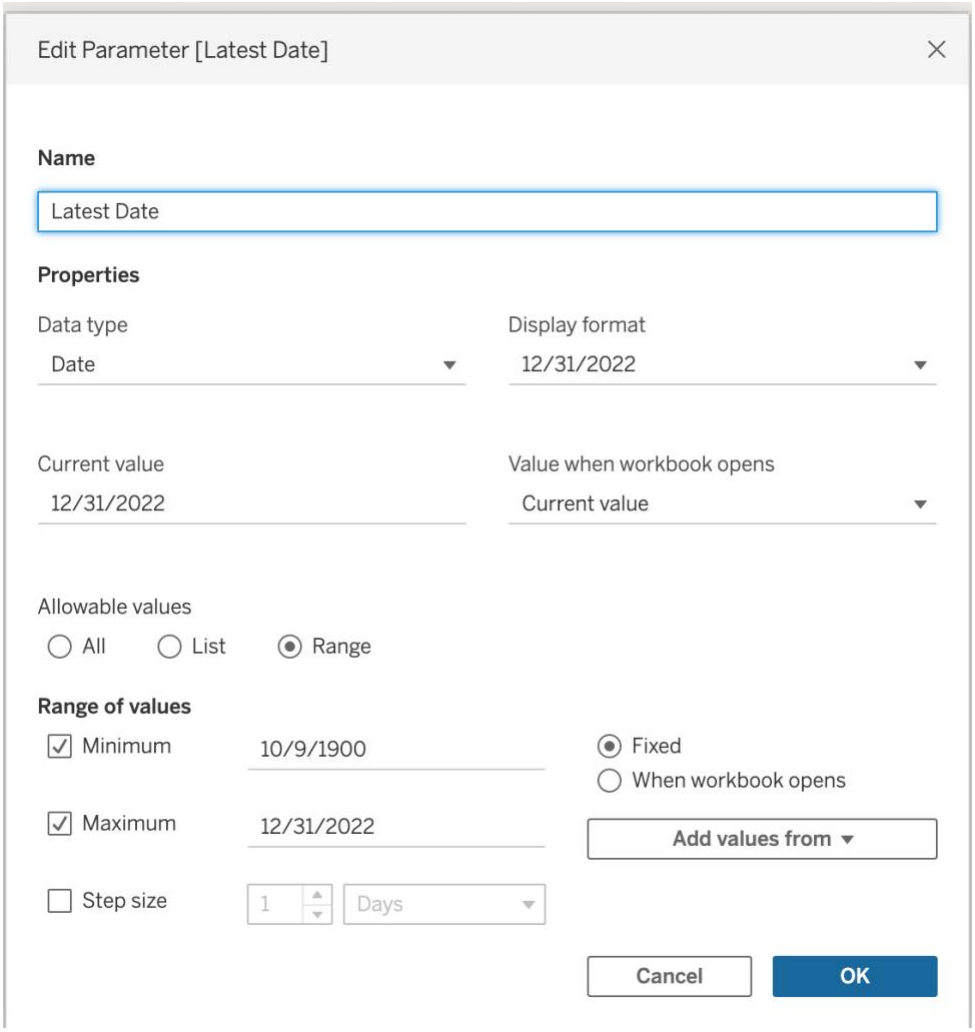

Fig 2.2. Latest Date Parameter

Creating Calculated Fields:

Implemented calculated fields within Tableau, named as 'Evaluation Period' and 'Magnitude Grouping,' leveraging the 'Earliest Date' and 'Latest Date' parameters for the former, and the earthquake magnitude attribute for the latter. The 'Magnitude Grouping' aligns with the categorization specified by the United States Geological Survey (USGS) website, wherein magnitudes are grouped into distinct categories such as 'Moderate' for 5-5.9, 'Strong' for 6-6.9, 'Major' for 7-7.9, and 'Great' for 8 and above.

Fig.3.1 Calculated Field 'Evaluation Period'

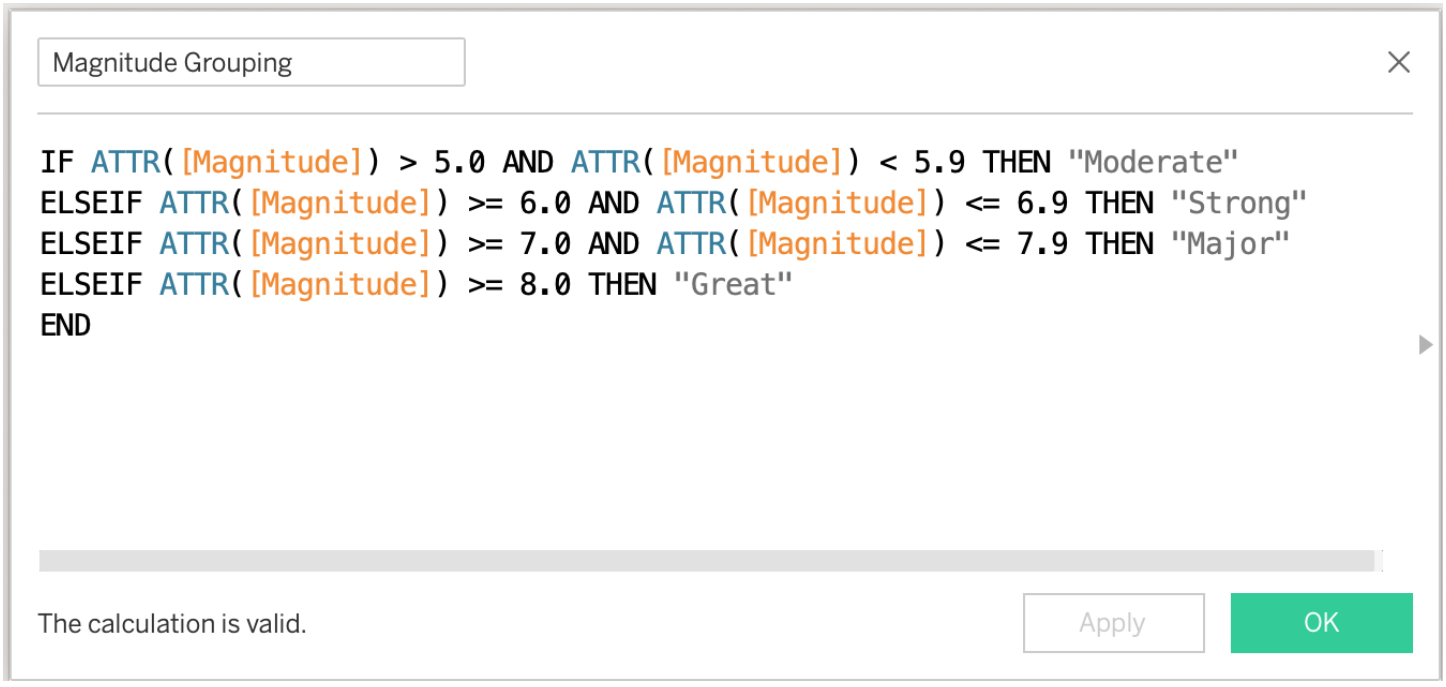

Fig.3.2 Calculated Field 'Magnitude Grouping'

Data Visualization and Analysis:

Leveraged Tableau, a powerful data visualization tool to get a deeper understanding on the occurrences of earthquakes by visual representations of the data. Exploited various features of Tableau like calculated fields, parameters among others to represent our data in a more intuitive and comprehensible manner.

Magnitude Counts over years

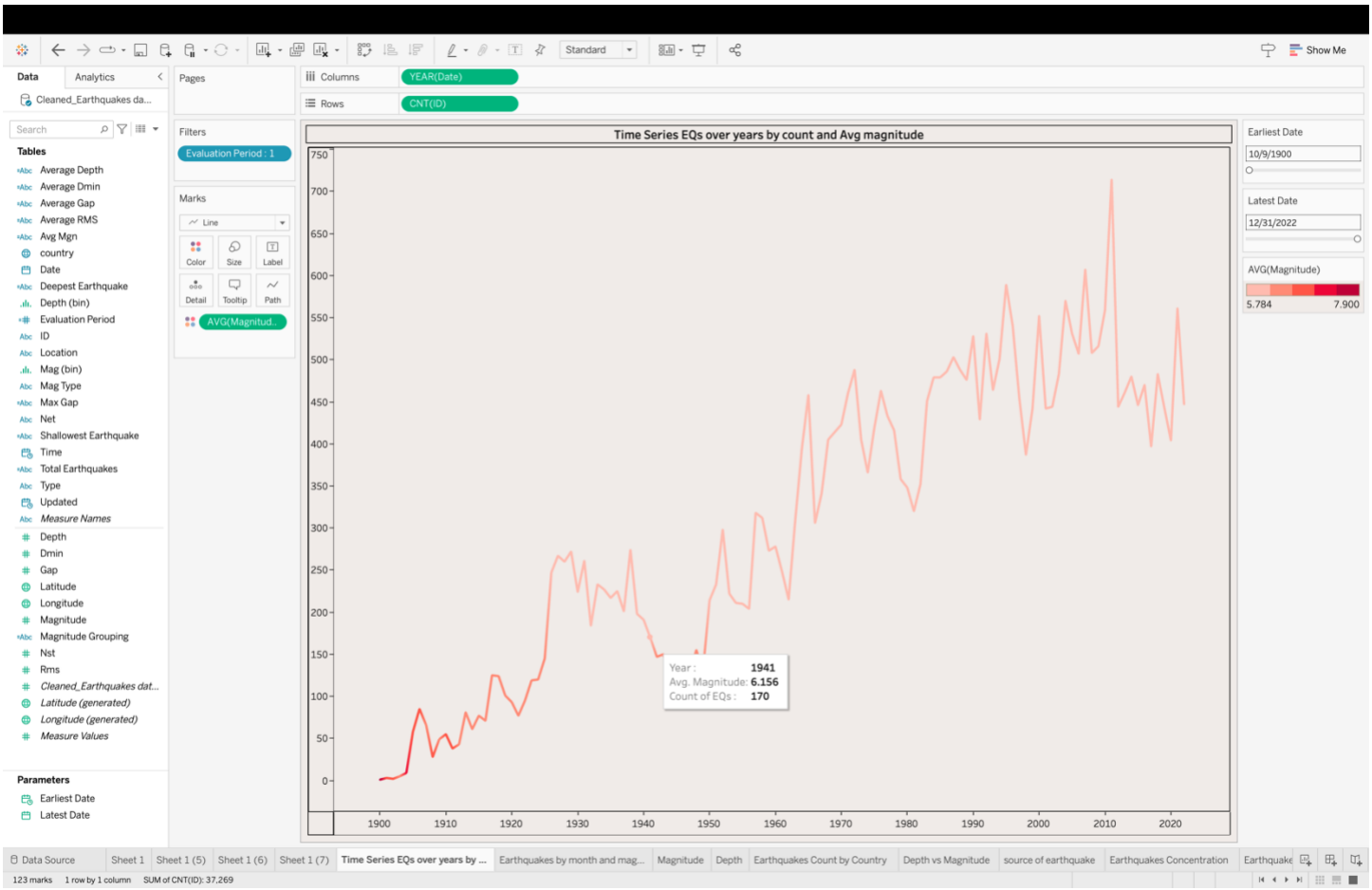

Fig.4.1. Time Series EQs over years by count and Avg magnitude

The above plot represents a time series graph with year on x-axis and the count of earthquakes on y-axis.
The height of the line typically represents the total number of occurrences of the event in a particular year and the color represents the average magnitude of the earthquake that occurred over that year. The plot is made interactive with respect to the dates by adding the Evaluation period filter. The color legend indicates the average magnitude ranging from 5.78 to 6.61. The darker the line, the higher the average magnitude for that year. When we hover on the line, we can see the year, average magnitude, and the count of occurrences.

From this time series plot it can be inferred that during the initial years of the 20th century though the occurrences are less compared to the later years, the average magnitudes were high. And from 1965 there's huge jump in the

number of occurrences and a decline in the magnitudes of the earthquakes. From this increasing tendency we can infer that either the frequency of earthquake occurrences is actually increasing or the technology of detecting the earthquakes have improved over time.

Seasonal Patterns:

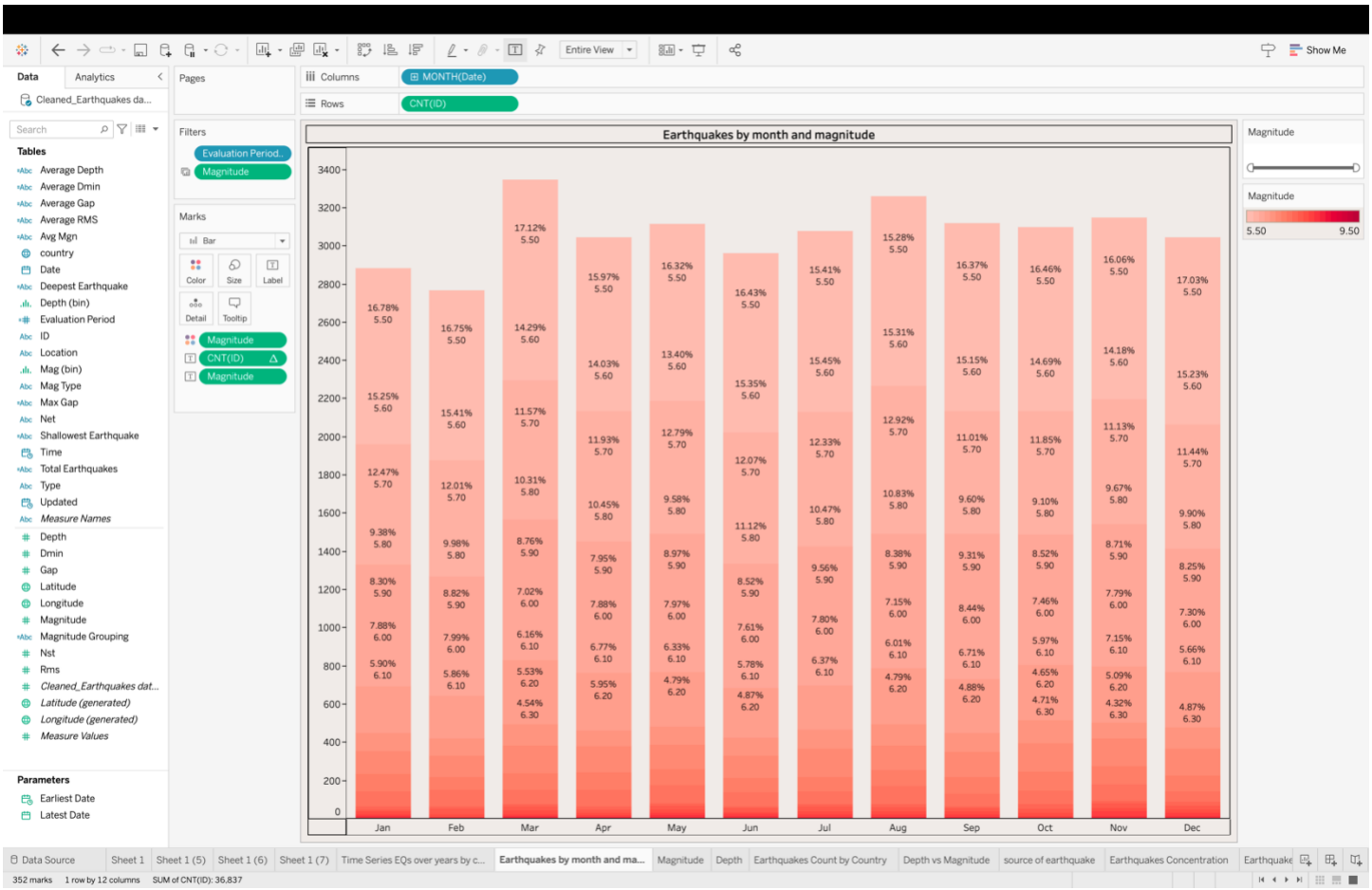

Fig.4.2. earthquakes by month and magnitude and percentages of counts

To understand any seasonal patterns in earthquake occurrences, created a stacked bar chart depicting months on the x-axis and counts on the y-axis. The magnitudes of the events are color coded, and the percentage of occurrences of each magnitude in a particular month has been displayed on the graph along with the corresponding magnitude. Used the 'Text Label' in the marks card to display the magnitude and the percentages.

From the above plot, it can be observed that the occurrences peaked in the month of March and as such that there is not much change in the bars of various months. Thus, can conclude that there is no such seasonal pattern observed in the occurrences of earthquakes.

Magnitude Histogram

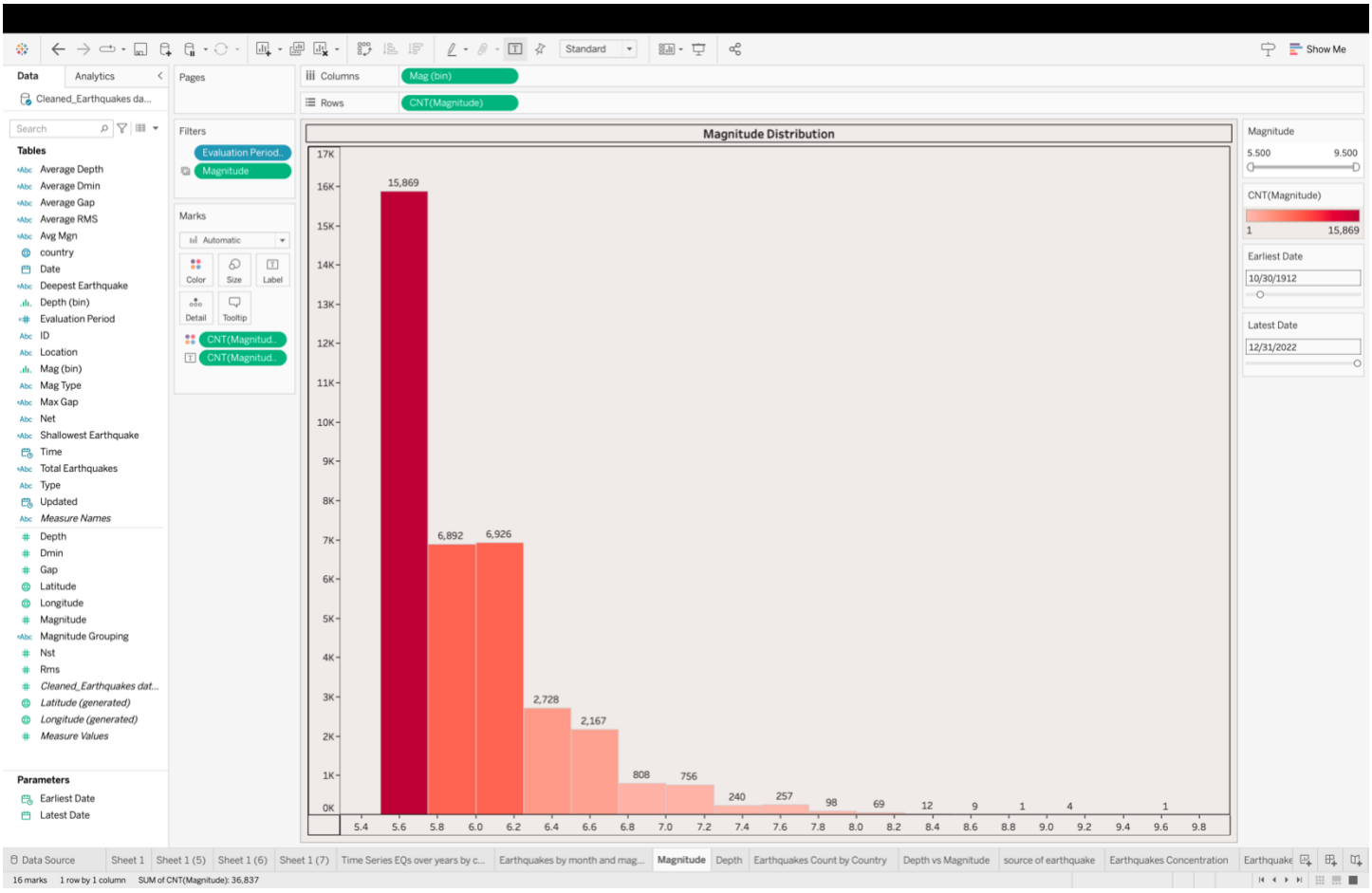

Fig.4.3. Magnitude Distribution

The above plot is a histogram of the magnitudes of all the earthquake occurrences, which shows the frequency of earthquakes across different magnitudes. The color legend indicates this frequency of occurrences. The count of ID (sum) is dragged onto the text label in marks card to display the total number of counts of each magnitude. The chart is made interactive by using various filters - magnitude, and dates. By changing the dates, we can see the frequency in that particular period.

From the above distribution, we can infer that as the magnitude increases the frequency of occurrences decrease, which typically proves the Gutenberg-Richter law, which states that the frequency of earthquakes decreases exponentially with increasing magnitudes.

Depth Histogram

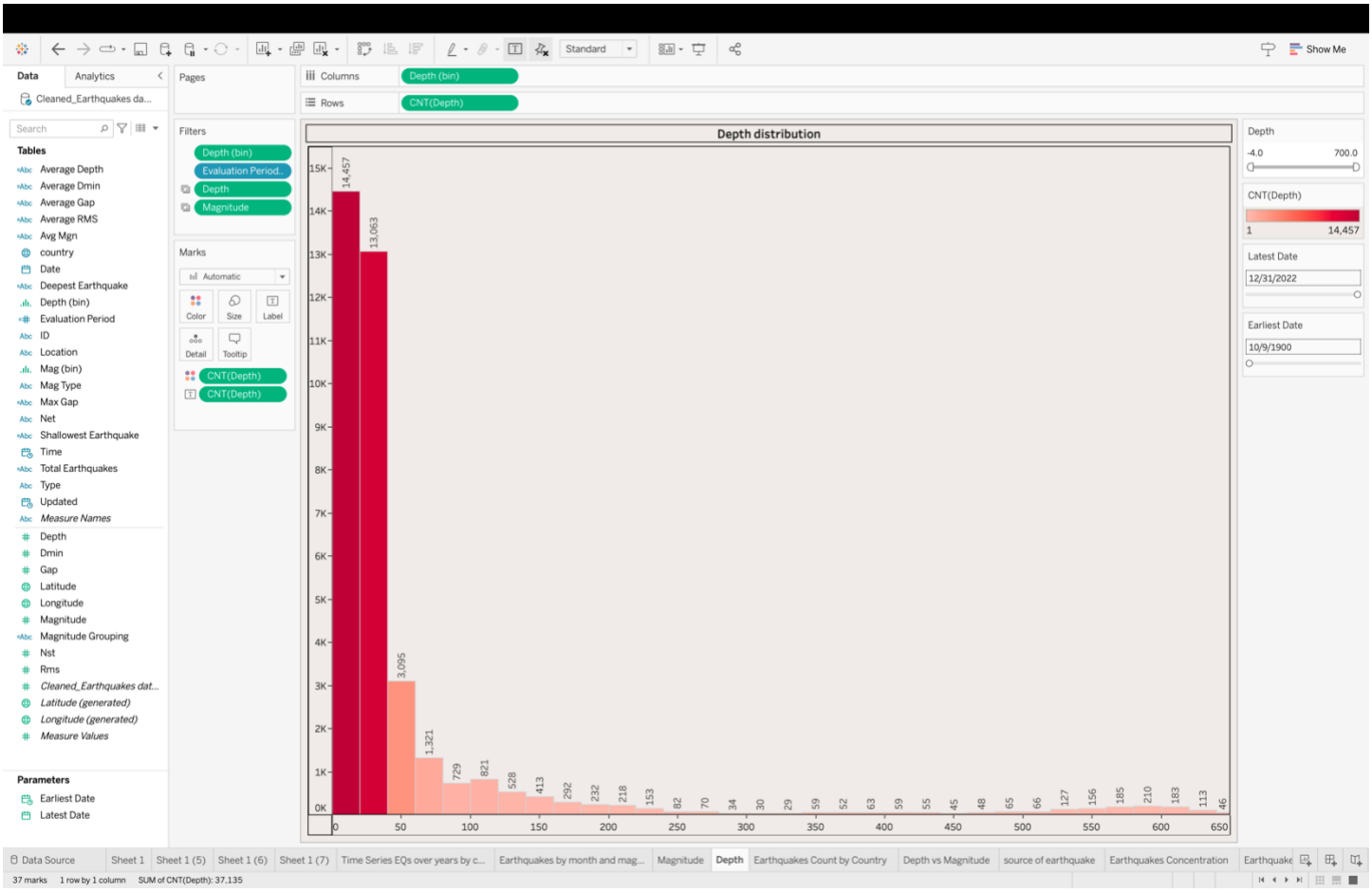

Fig.4.4. Distribution of Depth (Km)

The above plot is a histogram of the depths of all the occurrences, which shows the frequency of earthquakes across different depths. The color legend indicates this frequency of occurrences. The count of ID (sum) is dragged onto the text label in marks card to display the total number of counts of each depth. The chart is made interactive by using various filters - depth, and dates. By changing the dates, we can see the frequency in that particular period.

From the above distribution, we can observe that the from 0 to 50km depth there are around 27,520 earthquake occurrences and thereafter there's a steep decline in the occurrences, implying that shallow earthquakes are more common than the deeper ones.

Depth and Magnitude Relationship

Fig. 4.5. Depth vs Magnitude

To exploit the relationship between the depth and magnitude of an earthquake, I used scatter plot, with magnitude on the x-axis and depth on the y-axis and all these data points refer to the occurrences of earthquakes. The color legend depicts the magnitude variance, the darker the shade is the higher the magnitude. The plot is made interactive with the dates and the magnitude filters. We can observe that most of the data points are clustered at the shallow depths and in the lower range of magnitudes.

Heatmap of Earthquakes by Country

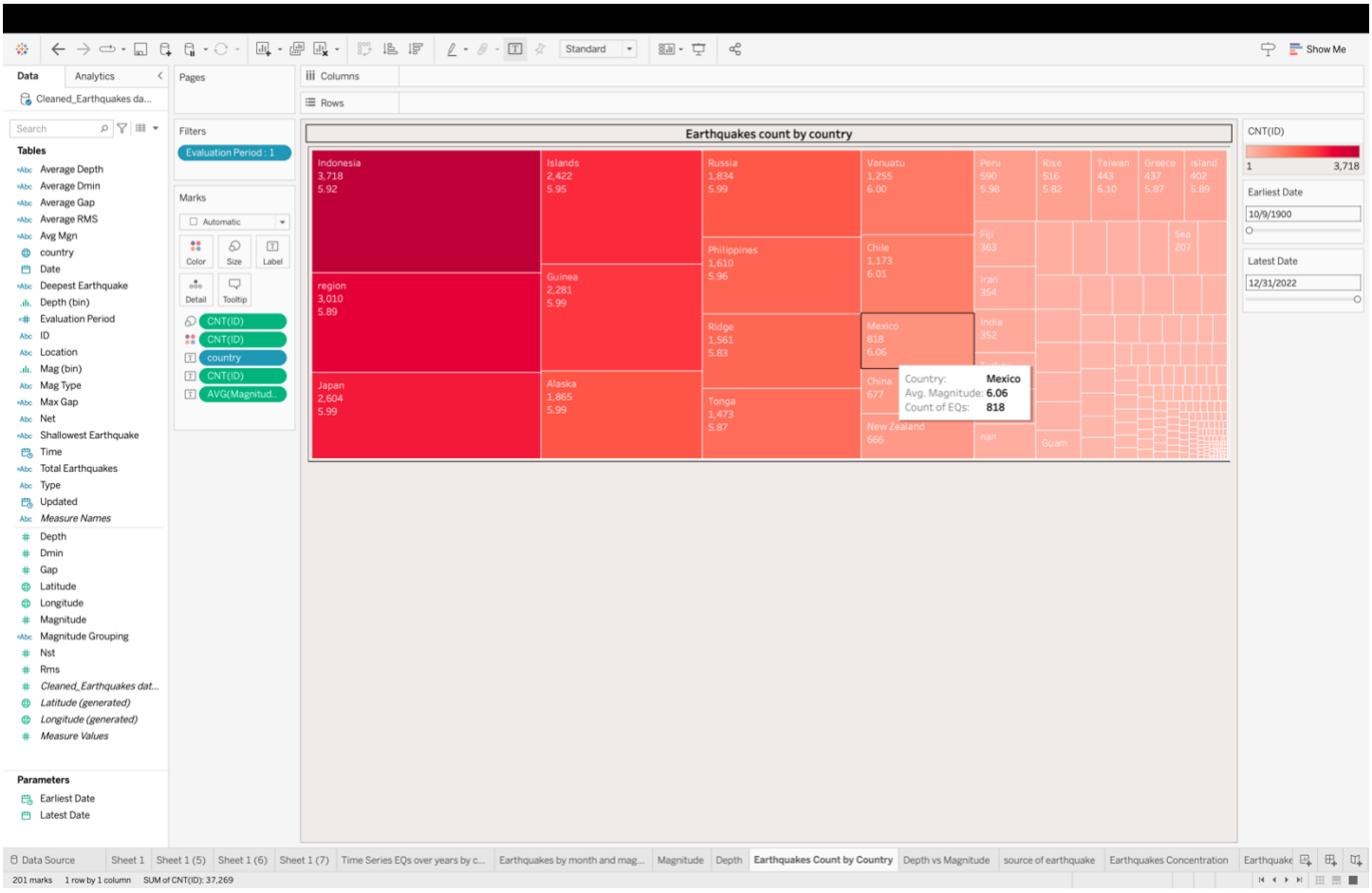

Fig.4.6. Heat map of earthquakes by country

The above plot is a heatmap that represents earthquake count grouped by each country, along with the average magnitude for each country. The color legend reflects the count of earthquakes, the darker the shade is the higher the count. By observing the heatmap we can tell which parts of the world are more prone to earthquakes and the strength of those earthquakes

From the above plot we can observe that Indonesia has around 3,718 occurrences over the last century with an average magnitude of 5.92, while India had around 352 earthquakes and the least is observed in Oman and Slovakia. This tells us that in regions of passive tectonic plates, the occurrences are rare which is obvious as there will be no tectonic movement.

Source of Earthquake

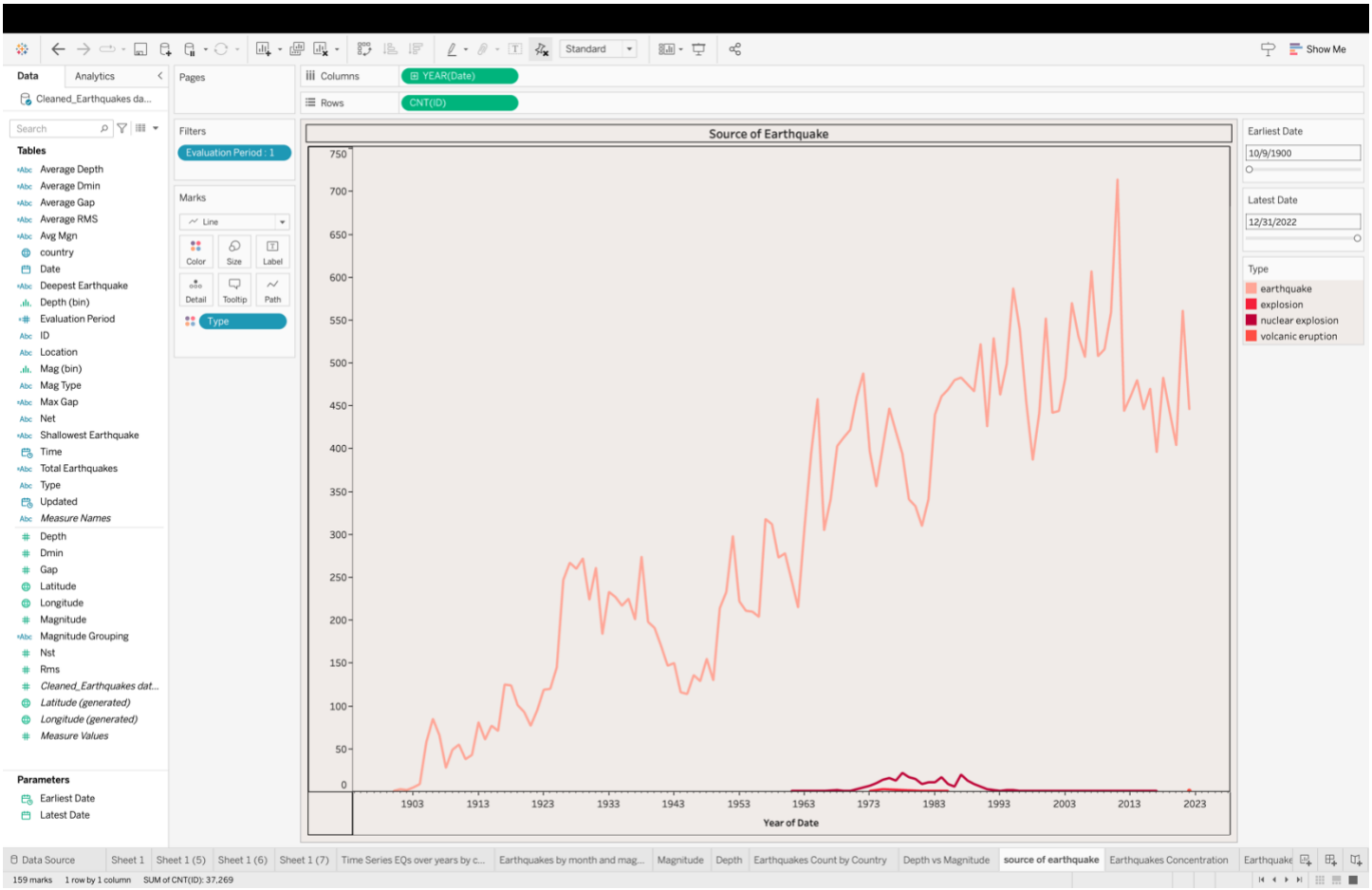

Fig.4.7 Source of Earthquake

This plots the various types of sources of earthquakes like seismic activity (earthquake), volcanic eruption, nuclear explosions, and other explosions. The plot is made interactive with the evaluation period filter.

We see that the primary cause of most earthquakes is natural seismic activity. However, between 1970 and 1992, there was a notable increase in the occurrence of earthquakes triggered by manmade factors, specifically nuclear explosions. Subsequently, there was a decline in such incidents as nations collaborated and implemented stringent nuclear regulations, contributing to a decrease in manmade seismic events.

Geographic Representation

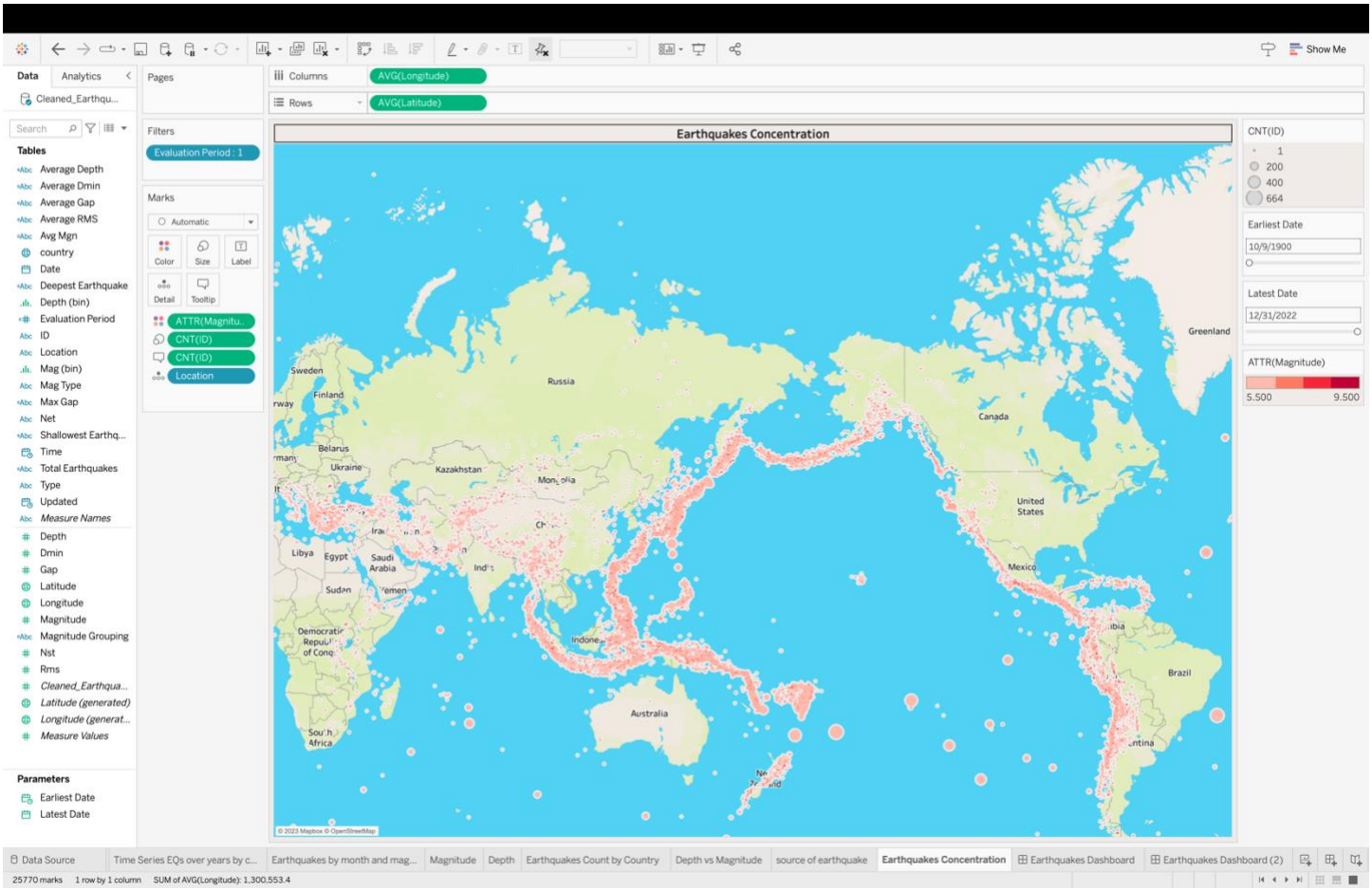

Fig.4.8. Earthquakes Concentration

The above map depicts the location of earthquake events with respect to their latitude and longitude of occurrence. The color legend depicts the magnitude variance, the darker the shade is the higher the magnitude and the size of the circle depicts the count of events, bigger the size higher the number of occurrences at a particular location. The map is made interactive with the evaluation period filter.

We can observe that almost 80% of the occurrences are concentrated along certain areas, like coastlines, island regions, and along the edges of the continents indicating that earthquakes occur where there is a constant tectonic plates movement.

The graph aptly points to the ‘Ring of Fire’, which is a tectonic belt that borders the Pacific Ocean which includes New Zealand, Japan and the costs of East Asia and the borders of North and South American continents, and various islands in the Pacific Ocean.

Dashboard: A Global Earthquake Analysis Dashboard

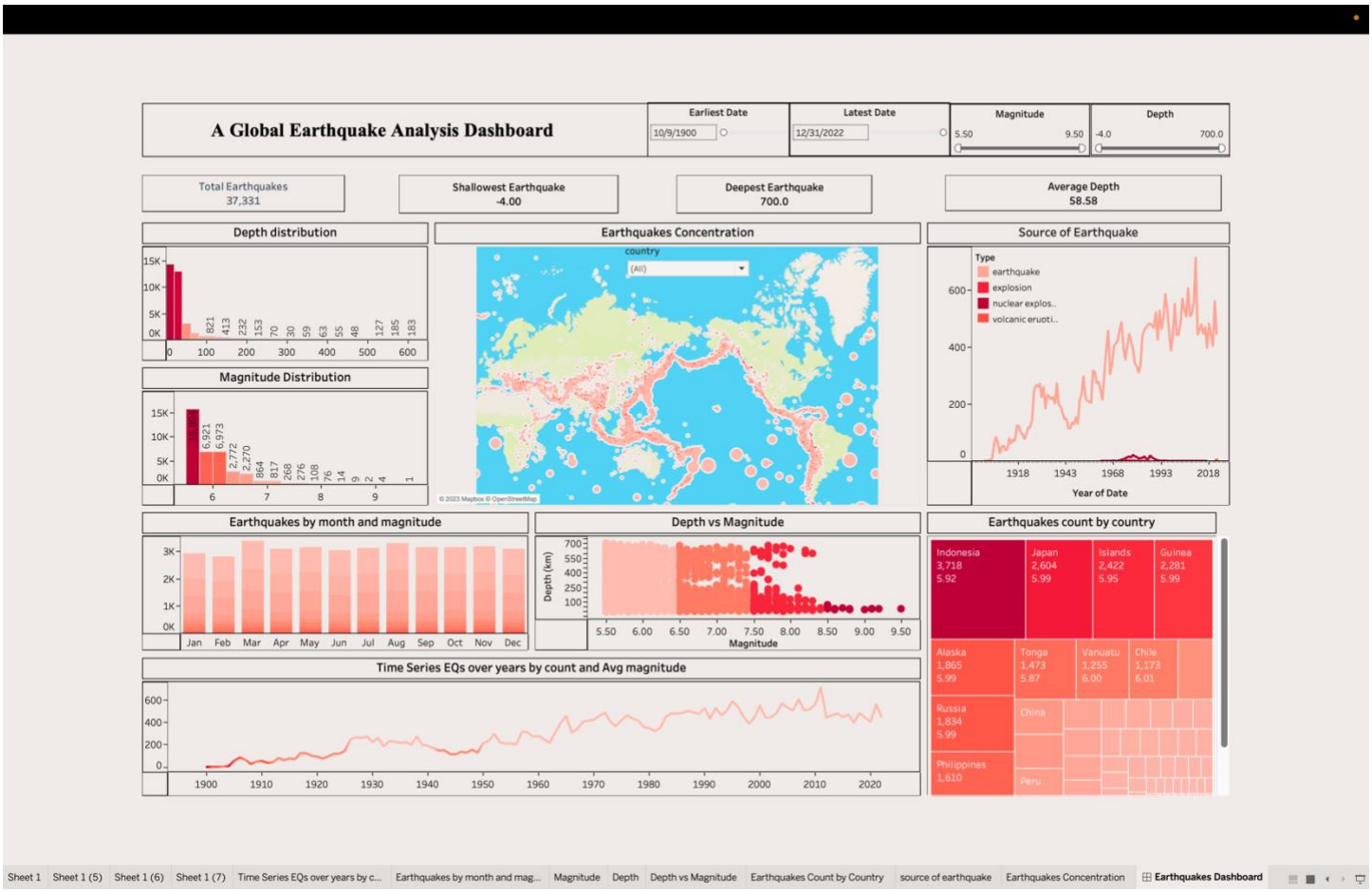

Fig 5.1. A Global Earthquake Analysis Dashboard

Conclusion:

The analysis of the 'Earthquakes Dataset 1900-2023' reveals significant findings about the nature and trends of earthquakes over the past century. The study confirms that shallow earthquakes are more common, and the frequency of earthquakes decreases with increasing magnitudes, aligning with the Gutenberg-Richter law. A notable increase in earthquake occurrences since 1965 may indicate improved detection technology or an actual increase in seismic activities. The data shows no apparent seasonal pattern in earthquake occurrences. Geographically, high concentrations of earthquakes are observed along active tectonic plate boundaries, particularly in the Pacific Ring of Fire. The study also highlights a period where man-made factors like nuclear explosions contributed to seismic events, though such occurrences have declined due to stringent regulations. This comprehensive analysis provides valuable insights into the patterns and impacts of earthquakes, aiding in better understanding and preparation for these natural phenomena.

References:

1. ‘Earthquake Dataset 1900-2023’, is from Kaggle, which is sourced from United States Geological Survey.
https://www.kaggle.com/datasets/jahaidulislam/significant-earthquake-dataset-1900-2023
2. Lecture Notes of course DATA230, by Andrew H Bond
3. Wikipedia on general understanding of earthquakes