Analysis and experimental forecasting of earthquakes in the US

BDA 600 Big Data Analytics Capstone Seminar

Submitted by

Parul Jain - pjain8319@sdsu.edu

Nishu Singh - nsingh2878@sdsu.edu

Umadevi Betageri - ubetageri5637@sdsu.edu

2nd May 2023

Big Data Analytics

San Diego State University

Website: https://arcg.is/nCj8b
Video: https://youtu.be/LAACGzyAksU
GitHub: https://github.com/nishusingh11/Captstone_Project

Team Contributions

Task	Member
Research on the topic	Umadevi, Nishu, Parul
Data collection 2000 – 2010 2010 – 2023 Alaska and Hawaii	 Umadevi Parul Nishu
Data cleaning and pre-processing	Umadevi, Nishu, Parul
EDA and Data Visualization Overall Analysis Year-wise Analysis Region-wise Analysis	 Umadevi Nishu Parul
Machine learning Models and Forecasting Clustering Timeseries forecasting (ARIMA, Exponential Smoothing, LSTM)	 Umadevi Nishu & Parul
Website development	Umadevi, Nishu, Parul
Video making	Umadevi, Nishu, Parul
Group report writing	Umadevi, Nishu, Parul
Individual Report writing	Umadevi, Nishu, Parul

Abstract : The project covers an in-depth analysis of earthquake patterns in the United States, spanning area, year, and the top five earthquakes recorded in the country from 2000 to Feb 2023. The study involves collecting and analysing historical data, as well as employing experimental approaches to forecast future earthquakes. For time series forecasting of earthquake data, this study makes use of ARIMA, Exponential Smoothing, and LSTM models. The goal is to get a better knowledge of the patterns and effects of seismic activity, as well as to develop forecasting tools to assist in mitigating the effects of earthquakes on infrastructure and communities. The study's findings will help to improve understanding of earthquake trends in the United States, as well as support the creation of effective disaster mitigation and management policies and methods.

Keywords - *Earthquakes, time series analysis, clustering, forecasting, Machine Learning.*

1. Introduction

The natural disasters that regularly occur on our planet include floods, tornadoes, tsunamis, volcanic eruptions, earthquakes, and storms, among others. It seriously harms both nature and all living things. It is irregular and unpredictable. The Earthquake, usually called a quake or a tremor, is one of the perilous calamities. If there is a dense population, there is tremendous harm.

An earthquake is a natural occurrence that happens when the earth's crust suddenly moves or shakes. The energy that has accumulated over time as a result of the movement of tectonic plates is released during this shaking. Large plates that are part of the earth's crust sit on top of the mantle below, which is molten. Over time, these plates move very slowly, and when they collide, pressure builds up along the fault lines where the plates meet. Other factors, such as landslides, volcanic activity, and human activities like chemical explosions, underground mining or the building of major dams, can also produce earthquakes. The likelihood and severity of induced earthquakes depend on a variety of factors, including the geological features of the area, the specific activities being undertaken, and the level of monitoring and prevention measures in place.

To better understand their causes and foretell when they might happen, seismologists research earthquakes. This information is crucial for enhancing earthquake preparedness and reducing the damage and casualties brought on by earthquakes. The US has suffered several earthquakes of varied magnitude since 2000, some resulting in severe damage and fatalities. In this capstone report, we will examine the frequency, magnitude, location, and other pertinent factors of the earthquakes that have occurred in the US throughout this time period. We intend to learn more about the trends and patterns of earthquakes in the US.

To help with a more in-depth analysis of the data and draw insightful conclusions, a variety of fundamental visual representations, including bar charts, line charts, and tables, among others, were created. In order to fully understand historical trends, studies of time series and geographic data were also performed. Furthermore, we also conducted experimental forecasting. Currently, earthquake prediction with a higher of accuracy and precision is not possible. and is an intricate problem, experimental earthquake forecasting is currently a subject of research in seismology.

Earthquakes occur due to complex and unpredictable geological processes, and there is no known method to predict when and where they will occur with certainty. However, scientists are continuing to conduct research and develop new methods to improve our understanding of earthquake occurrence and potentially predict them in the future. Additionally, it is important to focus on earthquake preparedness, such as building resilient infrastructure and emergency response plans, rather than relying solely on prediction.

2. Motivation and Literature Review

Natural disasters like earthquakes have the potential to seriously harm infrastructure, harm people, and severely damage communities. Because of the numerous active fault lines that are spread throughout the country, the United States is particularly vulnerable to earthquakes. The study aims to uncover patterns and trends in seismic activity by analysing earthquake data gathered from 2000 through 2023. By examining these patterns, scientists can better understand the causes of earthquakes and create forecasting models that are more precise in predicting future seismic events.

Also, the project seeks to raise earthquake safety knowledge among the general public. Researchers can identify regions that are more vulnerable to seismic activity and alert the public to possible dangers through the study of earthquake data. This is necessary because after effects of earthquakes can lead to landslide, tsunami, surface faulting and ground failure as given in USGS article. [1]

West M et al., 2018[2] emphasizes the significance of taking ground motion's non-uniformity into account when evaluating earthquake risks and hazards. Moreover, it highlights the requirement for improved modelling and monitoring of earthquake sources and their impacts on ground motion. The results of this study can assist Alaska and other areas with comparable geological circumstances estimate earthquake risk and take mitigation measures.

The source features and seismotectonic background of the Ridgecrest earthquake sequence are well-explained in USGS article[3]. The findings have important implication for estimating earthquake risks and reducing them in California and other places with comparable geological characteristics.

The research paper by Stover C et al., 1527[4] provides important information about several regions susceptible to earthquakes of the United States. The authors define the seismically active locations and discuss the different kinds of earthquakes that are usually observed in these places. They additionally explore how earthquakes affect various sorts of infrastructure, such as dams, bridges and buildings.

As far as predicting earthquakes is concerned, an article from USGS[5] clearly states that "Neither the USGS nor any other scientists have ever predicted a major earthquake." Only the probability of a major earthquake can be calculated in an area within a certain time frame. Also, a vague claim that an earthquake will occur now cannot be called a prediction because the movement of earth's crust is continuous and an earthquake is very likely to occur[6]. For instance, Geologists were aware that an earthquake would occur in southeast Turkey decades ago, but accurate forecasting is still the stuff of science fiction.[7]

The USGS distinguishes four terms: early warning, probabilities, forecasts, and prediction. Earthquake early warning systems alert people and devices when waves of shaking are expected in their area. Earthquake forecasts are used to predict the likelihood of aftershocks in the time period following an earthquake. Probabilities of earthquake indicate the long-term probability of an earthquake of a specific magnitude occurring during a given time span. Earthquake prediction necessitates identifying the date and time, location, and magnitude, but scientific data is currently insufficient to predict earthquakes. False earthquake predictions are based on non-scientific evidence and fail to define all three aspects of a prediction.[8]

A study done by Chittora P et al., 2022[9] uses the Machine learning classifiers to predict the magnitude of the future earthquake in Indian subcontinent. This paper mainly focuses on the relationship of depth and magnitude to make the predictions.

In another study by Yuan X et al., 2022[10], time series algorithms are used to analyse the irregularities of the origin times of the earthquake in Longmen Mountain fault zone in southwest china.

Another study by Amei A et al., 2012[11] uses the time series models to predict the large scale earthquakes worldwide having the magnitude greater than or equal to 8. Using the motivation and reference from these papers, we have implemented data science techniques to analyse and perform experimental forecasting on the earthquake data for the USA.

Lastly, the initiative is driven by the chance to further ongoing seismic research. Even though there has been a lot of progress in our knowledge of seismic activity, there are still a lot of open concerns regarding how earthquakes behave. The project aims to increase our knowledge of earthquakes and improve our ability to predict and mitigate their impacts.

3. Data

3.1. Dataset Description

This data is collected from the United States Geological Survey (USGS) website [12], one of the top experts on seismic activity in the country. The USGS keeps an extensive database of US earthquakes that is frequently updated with the most recent data. All US earthquakes between 2000 and 2023 (till February) having a magnitude of 2.5 or higher are included in our dataset. This covers both naturally occurring earthquakes and those brought on by human activities, like mining explosions and rock busts.

In order to better understand the patterns and trends of earthquakes in the US, including their frequency, magnitude, location, and other important characteristics, we are undertaking this investigation.

The CSV file for the dataset has 22 attributes and 100892 records. The following are relevant attribute[2] fields used for this project:

Attribute	Description
time	When event has occurred
longitude	Decimal degrees longitude.
latitude	Decimal degrees latitude.
depth	Depth of the event in kilometres.
mag	The magnitude of event
magType	The method or algorithm used to calculate the preferred magnitude for the event
nst	The total number of seismic stations used to determine earthquake location.
gap	The stations that have the greatest azimuthal separation (in degrees)
dmin	Horizontal distance between the closest station and the epicentre (in degrees)
rms	The RMS travel time residual with all weights, expressed in seconds.
net	a data contributor's ID
id	A unique identifier for the event
updated	Date of the most recent update to the event
place	area of geography close to the incident
type	seismic event type
horizontalError	Uncertainty regarding the event's reported position in kilometres
depthError	Uncertainty over the event's claimed depth in kilometres.
magError	Uncertainty over the event's claimed magnitude in kilometres.
magNst	the total number of seismic stations utilized to determine this earthquake's magnitude.
status	shows if a person has examined the incident.
locationSource	the station first reported the scene of this incident.
magSource	initially recorded magnitude for this event by the network.

The detail description of dataset attributes are available at ANSS Comprehensive Earthquake CatLog (ComCat) website[13]. The agencies, teams, and individuals whose information or materials were used in the creation of the report are included under contributors in USGS earthquake reports [14]. We created a tour map in our website to show the details about these contributors by creating table with the columns source code, description, images, latitude and longitude. The magnitude of an earthquake occurrence is often calculated using many approaches or algorithms that take into consideration various earthquake criteria such as seismic wave amplitude, distance from the epicentre to the recording station, and earthquake type [15].

3.2. Data Pre-processing

This is the vital initial phase in both data analysis and machine learning. The project's first stage is to clean the data with Pandas and NumPy. In this phase, we have the following data cleaning steps:

Data Merging: Downloaded three separate data files for contiguous US, Alaska, and Hawaii. Merged these files into one data file (complete data of all states in the US)

Website: https://arcg.is/nCj8b

***Converting and Extracting attributes**:* We extracted the region name from the entire address in the "place" attribute and converted the "datetime" format of the time attribute into "time" and "date" separately.

***Data filtration**:* For this project, we have taken a dataset that is reviewed by humans i.e., "status" value is set to "reviewed".

***Handling "NA" and "Null" values**:* Missing or null values in an earthquake dataset can be particularly troublesome since they could correspond to crucial seismic measures like the earthquake's magnitude or epicentre. The dataset is made full and correct by removing NA and null values, which can produce analytical and modelling outputs that are more trustworthy and insightful.

***Removing irrelevant attributes**:* Here irrelevant means the features that don't offer any meaningful information or have little bearing on how the analysis or model turns out are considered irrelevant. The dataset may be made simpler and the complexity of the analysis or model can be decreased by removing unnecessary features. Removed attributes are: id, status, rms, updated, net, horizontalError, gap, updated, magError, and depthError.

Renaming the attributes: At last, we modified the attribute names for easier understanding after completing the procedures above.

The figures 3.1 and 3.2 shows the raw data and data after cleaning and pre-processing respectively. With this our dataset is ready for exploratory data analysis and experimental forecasting with 12 significant attributes.

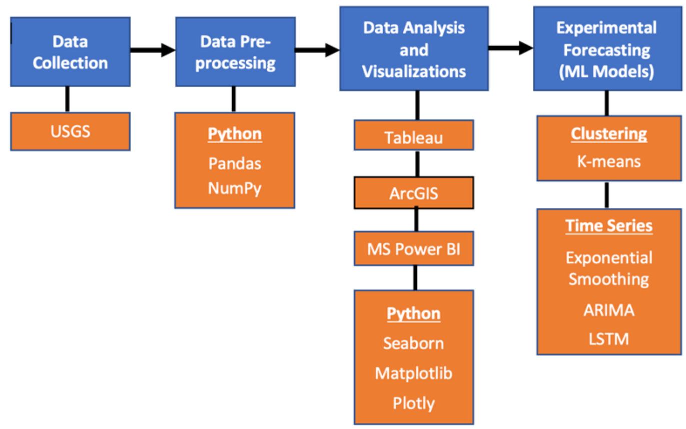
time	latitude	longitude	depth	mag	mag Type	nst	gap	dmin	rms	net	id	updated	place	type	horizont-alError	depth Error	mag Error	mag Nst	status	loc Source	mag Source
2003-12-31 T22:01:47.59	35.56133	-120.806	7.45	2.59	md	49.0	56.0	0.08649	0.06	nc	nc213 28106	2017-01-04 T23:39:23.08	9 km W of Templeton, California	earthquake	0.15	0.88	0.1	50.0	reviewed	nc	nc
2003-12-31 T20:17:47.38	35.698	-121.1415	5.94	2.92	md	50.0	118.0	0.1045	0.07	nc	nc213 28075	2017-01-04 T23:38:57.27	7 km NE of San Simeon, California	earthquake	0.15	0.58	0.07	71.0	reviewed	nc	nc
2003-12-31 T20:14:26.60	35.69983	-121.1393	5.25	2.89	md	50.0	116.0	0.1063	0.06	nc	nc305 00600	2017-01-04 T23:38:46.65	7 km NE of San Simeon, California	earthquake	0.13	0.51	0.09	71.0	reviewed	nc	nc
2003-12-31 T16:20:49.55	35.586	-120.9825	5.763	2.82	md	14.0	183.0	0.08108	0.13	nc	nc680 07512	2017-01-04 T23:37:40.39	9 km ENE of Cambria, California	earthquake	0.7	0.63	0.07	13.0	reviewed	nc	nc
2003-12-31 T15:08:06.59	33.66216	-91.7686	1.13	2.8	mlg	11.0	235.0	1.208	0.43	nm	nm60 5328	2016-04-11 T14:26:52.06	4 km NNE of Monticello, Arkansas	earthquake	3.8	7.3	0.351	2.0	reviewed	nm	nm
2003-12-31 T10:27:33.59	35.5695	-120.8078	3.685	2.99	md	15.0	201.0	0.08829	0.19	nc	nc680 07505	2017-01-04 T23:37:01.05	9 km WNW of Templeton, California	earthquake	0.7	1.52	0.09	9.0	reviewed	nc	nc
2003-12-31 T09:46:17.69	35.53683	-120.8535	6.45	3.33	ml	57.0	49.0	0.04324	0.06	nc	nc213 27818	2017-01-04 T23:36:47.34	11 km NNE of Cayucos, California	earthquake	0.15	0.48		2.0	reviewed	nc	nc
2003-12-31 T09:18:46.97	35.69483	-120.9996	6.364	2.75	md	46.0	45.0	0.02703	0.12	nc	nc213 27812	2017-01-04 T23:36:22.77	8 km SSW of Oak Shores, California	earthquake	0.17	0.63	0.09	53.0	reviewed	nc	nc
2003-12-31 T09:08:02.06	35.697	-121.1428	6.076	2.8	md	32.0	120.0	0.1045	0.06	nc	nc213 27808	2017-01-04 T23:36:13.12	7 km NE of San Simeon, California	earthquake	0.2	0.69	0.08	32.0	reviewed	nc	nc

Figure 3.1. Raw data in excel format.

Latitude	Longitude	Depth	Magnitude	Magnitude _Method	D_min	Type	Location_ Source	Magnitude_ Source	Region	Date	Time
25.6559	-120.9243	10.0	4.9	mb	7.229	earthquake	US	US	North Pacific Ocean	2015-12-27	19:24:56
26.1818	-66.6189	10.0	4.2	mb	6.219	earthquake	US	US	North Atlantic Ocean	2019-06-01	21:22:51
25.7194	-66.6056	2.08	5.2	mwb	5.896	earthquake	US	US	North Atlantic Ocean	2013-12-23	16:45:49
40.32	-124.733	12.543	4.5	mh	5.654	earthquake	CI	CI	California	2010-03-06	08:47:37
30.1795	-74.1579	19.04	4.4	mb	5.138	earthquake	US	US	East Coast	2013-10-08	01:58:09
26.1169	-92.1408	10.0	4.3	mb	4.707	earthquake	US	US	Gulf of Mexico	2018-02-26	04:44:54
40.0	-124.0	4.813	3.3	ml	4.665	earthquake	CI	CI	California	2001-08-10	20:19:55
42.2726	-115.7232	0.0	3.0	ml	3.869	earthquake	NN	NN	Nevada	2010-07-22	08:19:29
27.359	-111.2764	10.0	4.1	mb	3.85	earthquake	US	US	Mexico	2017-01-18	04:09:37
31.9883333	-124.1363333	9.835	3.51	ml	3.78	earthquake	CI	CI	California	2010-10-12	01:32:38
26.7675	-110.8671	10.0	5.3	mww	3.773	earthquake	US	US	Mexico	2022-01-21	21:26:35

Figure 3.2. Transformed data for analysis.

4. Methodology



Flowchart 4.1. Project phases.

4.1. Analysis and Visualizations

Understanding and predicting seismic events requires a rigorous analysis and visualization of earthquake data. Scientists can build models for earthquake forecasting by analysing seismic data to find trends and patterns in seismic activity. Scientists can better comprehend the intricate connections among earthquakes and geological causes by using visualization tools to convey the data in a comprehensible way.

4.1.1. Overall Trends

The analysis of disaster occurrences that occurred in the USA between 2000 and 2023 is shown in figure 4.1. We can see with clarity that the majority of our data has a high percentage of natural earthquakes roughly 94%. The remaining ones include rock bursts, explosions in mines, and so on. Our examination of the different types of explosives led us to the conclusion that Wyoming exhibits about 34% of explosions and 24% of mining explosion activity. The third graph displays the top 6 geographic areas with the largest explosions, with Wyoming having the largest explosions with a magnitude of 4.8. The highest explosion magnitude as a function of time is depicted in our last graph. 2003 saw a 4.8M explosion, and 2012 saw a 4.5M explosion.

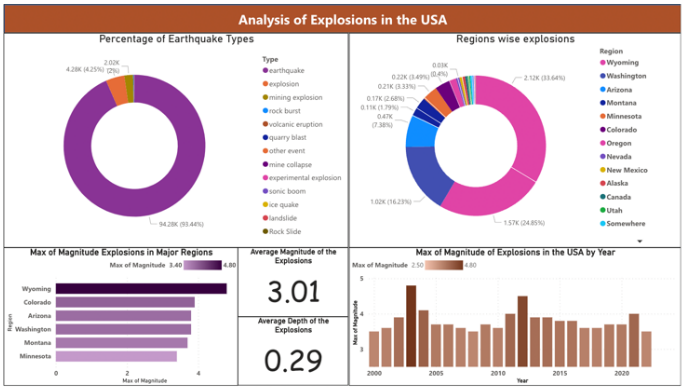


Figure 4.1. Analysis of earthquake types and explosions.

The highest magnitude of each region for each year between 2000 and 2023 is displayed in the figure 4.2. According to this data, Alaska, California, and Mexico are the countries that have experienced big earthquakes with magnitudes of 7.9, 7.1, 7.1, 7.2, and 7.0 in 2002, 2018, 2019, 2010, and 2016, respectively. In the following section, a detailed study of these earthquakes is shown.

Region	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Alabama	3.000	3.900	3.000	4.600	4.300	3.300	2.800	3.000	3.100	3.400	3.200	3.500	2.900
Alaska	5.800	5.500	7.900	5.400	5.500	5.400	5.400	5.200	5.200	5.900	5.400	5.300	5.800
Arizona	3.500	3.400	3.500	3.800	2.800	5.100	2.900	3.400	3.700	3.100	3.600	3.800	3.200
Arkansas	3.900	4.300	2.800	4.000	2.900	4.200	2.800	3.100	2.800	3.000	4.100	4.700	3.900
Arkansas-O..													
Bering Strait									3.200	3.300			
Bermuda												4.600	
California	4.900	5.200	5.200	6.500	5.970	5.200	5.000	5.450	5.440	5.190	6.500	4.730	6.300
California-..	3.130	3.170	3.330		3.610		3.800	3.300	3.730	3.320	4.400	3.500	
California-..										2.900			
Canada	4.700	3.900	4.500	4.200	3.100	4.900	4.200	4.100	3.800	3.600	5.350	3.870	4.500
Colorado	4.400	4.500	4.300	3.800	4.400	5.000	3.800	4.400	3.400	3.900	3.800	5.300	3.500
Connecticut													
Delaware													
East Coast			5.400					4.600					
Florida		3.300											
Georgia	3.500	3.200		3.600	2.500	2.600	2.900	2.600	3.100	3.200	2.600	2.700	2.700
Gulf of Alas..													
Gulf of Calif..			4.400	3.500				4.300		4.400		4.300	
Gulf of Mex..				3.200			5.900						
Hawaii	5.000	4.900	4.100	5.000	4.500	5.300	6.700	5.400	4.300	4.910	4.700	4.570	4.880
Hawaiian Is..		3.000										3.100	
Idaho	3.700	5.300	4.100	3.500	3.900	4.500	4.400	3.800	3.700	3.110	3.400	4.400	3.400
Idaho-Mont..	3.300		3.100	3.000						3.100	2.600		
Illinois	2.900			3.200	4.200	3.000	3.600	2.500	5.200		3.800	2.700	3.600
Illinois-Indi..													
Illinois-Ken..									2.500				

Figure 4.2. Region-wise maximum earthquakes in each year.

The map shown figure 4.3 displays earthquake points from 2000 to 2023 mapped on top of a USGS earthquake faults layer. The base layer comprises locations and details on faults and related folds that are thought to have caused large earthquakes (those with magnitudes of 6 or higher) over the past 1.6 million years in the United States. An earthquake is the consequence of a movement along a fault, which is a crack or region of cracks in the Earth's mantle along which the crustal blocks have shifted in relation to one another.

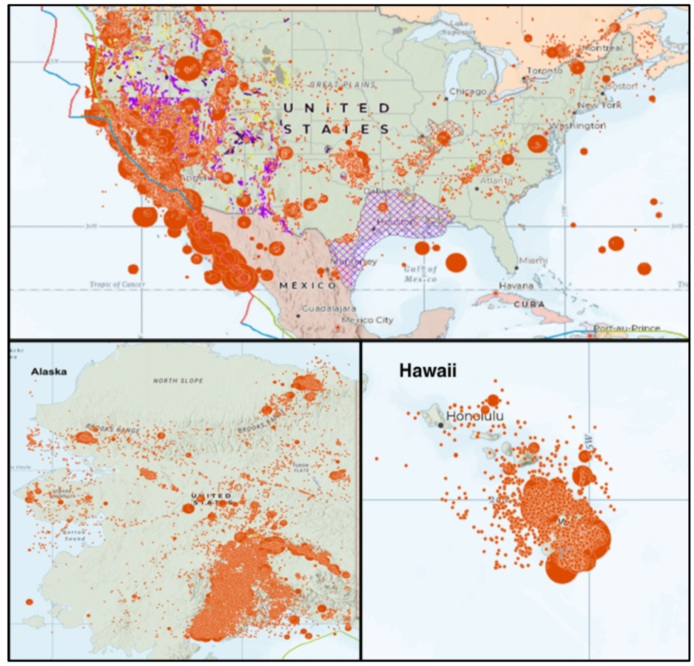


Figure 4.3. USA map showing details of all the earthquakes.

The map shown in the figure 4.4 displays earthquakes of a magnitude greater than 4 on top of a layer that is susceptible to landslides. Earthquakes are one of the major causes of the landslides. As a result, the base map highlights regions in the contiguous United States where numerous landslides have occurred, and it also includes regions that are prone to landslides. Nevertheless it is noted that earthquake sites are less visible in the red areas where landslides are more likely to occur.

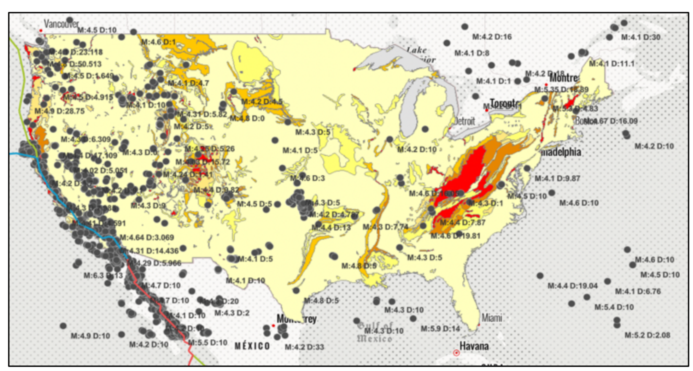


Figure 4.4. The Map of Earthquakes on the layer of Landslide Susceptibility (UCSG).

The number of earthquakes in the top 20 regions from 2000 to 2023 is depicted in the animated graph shown in figure 4.5. One of the states with the most number of earthquakes is California, which recorded 1119, 1170, and 3546 in the years 2000, 2001, and 2019, respectively. Alaska and Hawaii recorded 2985 and 3206 earthquakes, respectively, between the years 2002 and 2018. Moreover, Oklahoma reports 2760 earthquakes in 2015. Mexico, which is a country bordering the United States, recorded 3397 earthquakes in 2010. As a result, Alaska and California are covered in more detail in the following section.

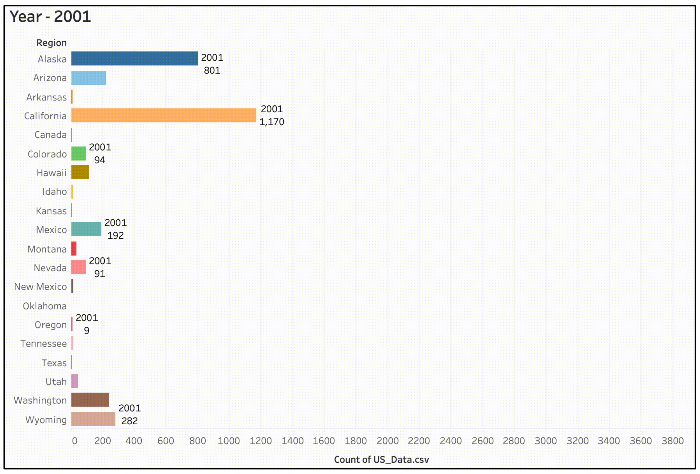


Figure 4.5. Total number of earthquakes in the top 20 regions of the USA.

Figure 4.6 shows a bar graph that displays the annual total number of earthquakes. According to this graph, the year 2018 had the most earthquakes (8,224), followed by the years 2010, 2019, and 2020 having 6679, 6501, 6149, and 6111 earthquakes, respectively. As a result, these years are covered in more detail in the following study.

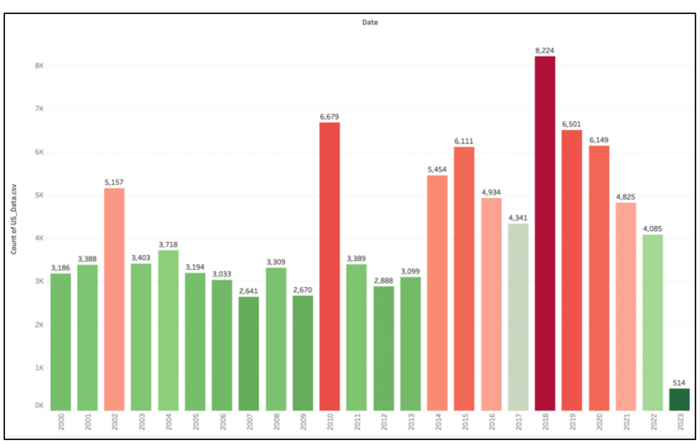


Figure 4.6. Number of earthquakes that occurred each year.

The graph in the figure 4.7 describes the magnitude method, which is used to determine the year's highest magnitude. The years 2002, 2010, 2015, 2018, 2019 and 2020 are taken into consideration for the analysis because they have notable earthquakes. It can be seen that large earthquakes of magnitudes of 7.9, 7.2, 6.4, 7.1, and 5.8 are reliably measured using the "mw" (Moment W-phase) method. Since 2002, the approaches "mb" and "ml" have been applied regularly.

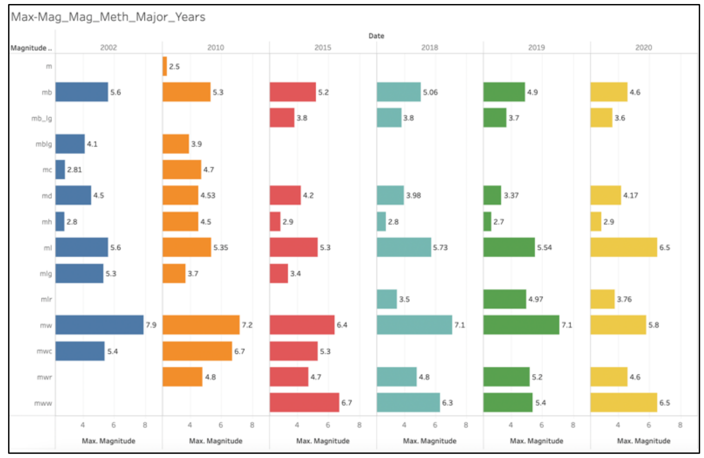


Figure 4.7. Magnitude method vs maximum magnitude graph.

The line graph in figure 4.8 display the highest magnitude and corresponding depth for each year, as well as the highest depth and corresponding magnitude. We can see from the graph of highest magnitude that the depth of the earthquake with 7.9M is only 4.2km. The earthquake at a depth of 300 km has 2.7M, as can be seen sequentially in the highest depth graph. The earthquake with a depth of 213.5 km also had 2.8M.

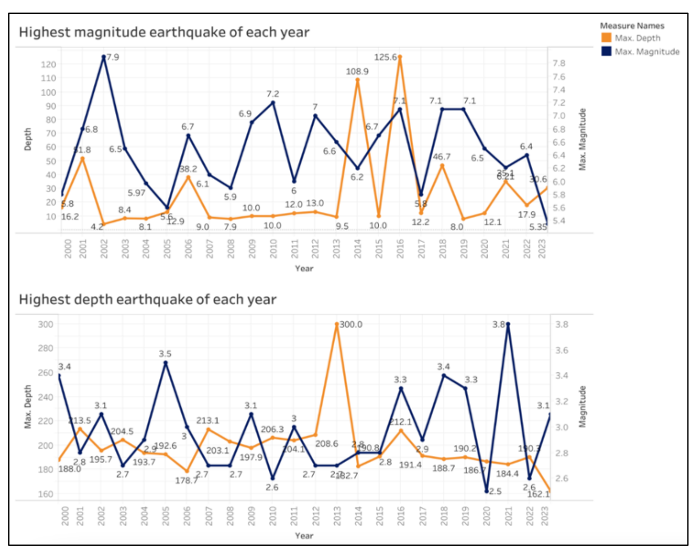


Figure 4.8. Highest magnitude and depth earthquake of each year.

Hence, we can draw the conclusion that the relationship between magnitude and depth is not always obvious. Slightly deeper earthquakes of the same size can cause greater damage than deeper earthquakes of the same magnitude because the energy produced by the earthquake is transmitted more directly to the surface. The local rock formations and soil type may have an impact on the earthquake's effects, among other factors.

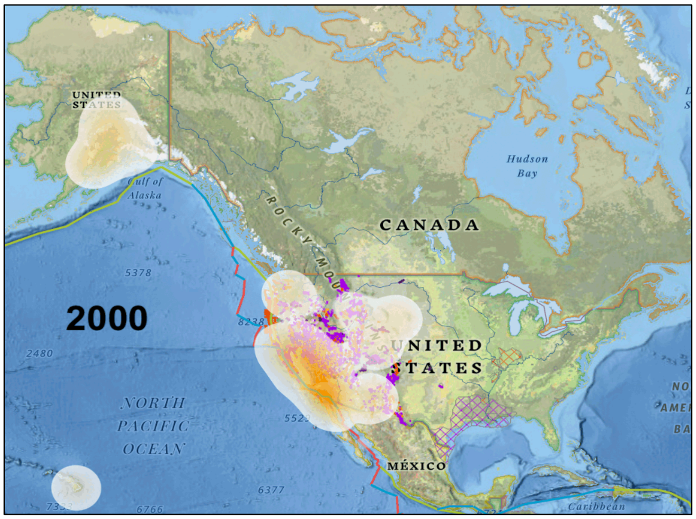


Figure 4.9. Magnitude method vs maximum magnitude graph.

The hear map in the figure 4.9 displays the earthquake intensity over time with respect to magnitude. The USGS Earthquake Fault base map has been added to this map. It can be seen that Hawaii, Alaska and California are in the darker shade i.e., the most affected regions.

4.1.2. Major Earthquakes

We have carried out a research on major 5 earthquakes in the US throughout the period. The details are as follows.

Denali earthquake – 2002

At its epicentre 66 kilometres (41 miles) east of Denali National Park in Alaska, the US, the Denali earthquake struck on Nov 3, 2002 at 10:12:41pm UTC (1:12 PM local time). A 7.9 Mw earthquake has not struck the US since 1973 (after the Rat Islands earthquake in 1965). The shock was the most intense ever recorded in Alaska's interior. Due to the remote location, there was only one injury and no fatalities. It might be felt all the way in Seattle because to the shallow depth.

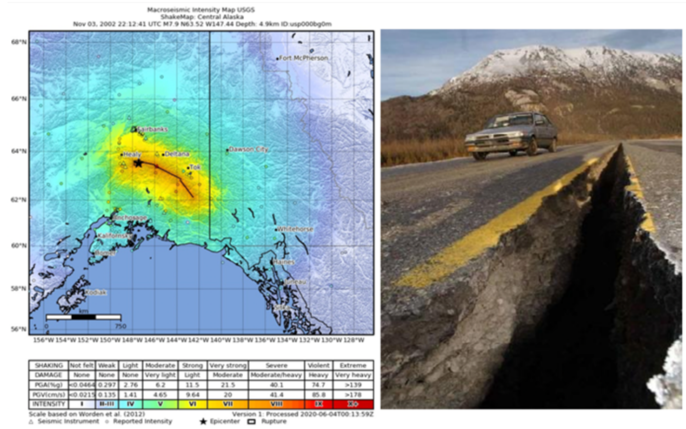


Figure 4.10. Pictures of Denali earthquake – 2002.

Old Iliamna earthquake - 2016

On January 24, 2016, at around 1:30 a.m. AKST, the earthquake rocked the Cook Inlet region of Alaska near Iliamna. About 65 miles (105 km) from Homer and 162 miles (261 km) from Anchorage were the epicentres of the earthquake. [6] The 7.1 on the Richter scale earthquake was felt throughout much of Southcentral Alaska and as far away as Juneau, which is around 700 mi (1,100 kms) southeast of the epicentre. Homes, roads, and commercial buildings sustained moderate to severe damage over a wide area.

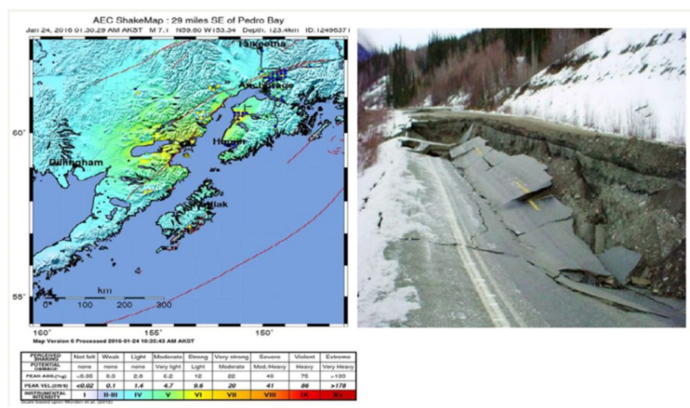


Figure 4.11. Pictures of Old Iliamna earthquake - 2016.

Anchorage earthquake – 2018

A 7.1 magnitude earthquake struck South Central Alaska (05:29 PM GMT) on November 30, 2018. The centre of the earthquake was situated at a depth of 29 miles in Point Mackenzie, about 10 miles (16 kms) north of Anchorage (47 kms). A magnitude 5.7 tremor with a centre 4.0 kms (2.5 miles) to the northwest of the municipality was felt six minutes after the original aftershock. The earthquake was felt all the way in Fairbanks. The National Tsunami Warning Centre in Palmer, Alaska, 42 miles (68 kms) northeast of Anchorage and within the earthquake zone, issued tsunami warnings for neighbouring coastal locations including Cook Inlet and the Kenai Peninsula. However, the warnings were quickly lifted.

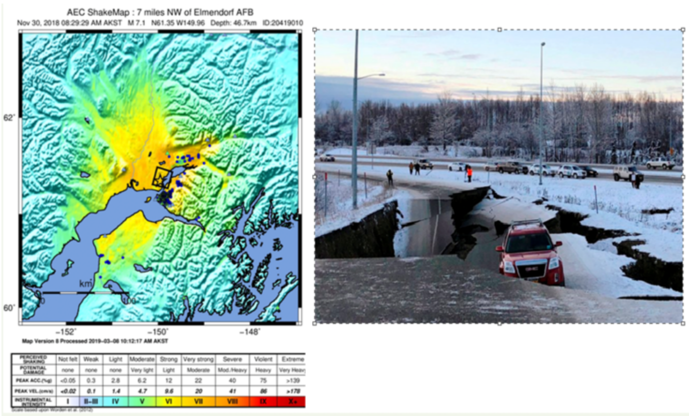


Figure 4.12. Pictures of Anchorage earthquake – 2018.

Ridgecrest earthquakes – 2019

The strongest earthquakes to hit California in more than 20 years were the Ridgecrest Earthquakes in July 2019. A magnitude 6.4 earthquake occurred on July 4 at 10:33 AM PST, roughly 12 km (10.5 miles) southwest of Searles Valley. Following several aftershocks, on July 5, 2019, an earthquake with a magnitude of 7.1 ripped the earth in the Mojave Desert, unleashing the equivalent of 45 nuclear bombs. The earthquake's strength was comparable to that of the atomic bomb that was dropped on Hiroshima.

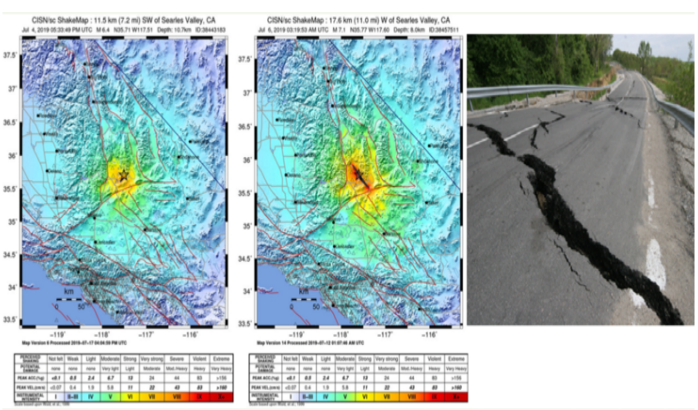


Figure 4.13. Pictures of Ridgecrest earthquakes – 2019.

Easter earthquake – 2010

On 4th of April (Easter Sunday), a 7.2 magnitude earthquake with a very severe Mercalli intensity (VII) rocked the Baja California region. It has several names, including 2010 Easter, 2010 Sierra El Mayor, and 2010 El Mayor - Cucapah. At 03:40:41 PM PDT, an earthquake hit the Mexican state of Baja California, just south of the town of Guadalupe Victoria. Aftershocks from the 8.9 magnitude quake were felt strongly in both southwestern Mexico and southern California. The last earthquake of the same magnitude was in 1952 in Kern County, California. The last time an earthquake of such size hit southern California was in 1992 when the 7.3 magnitude Landers earthquake occurred. All of these tremors had around the same magnitude and were felt all throughout the world.

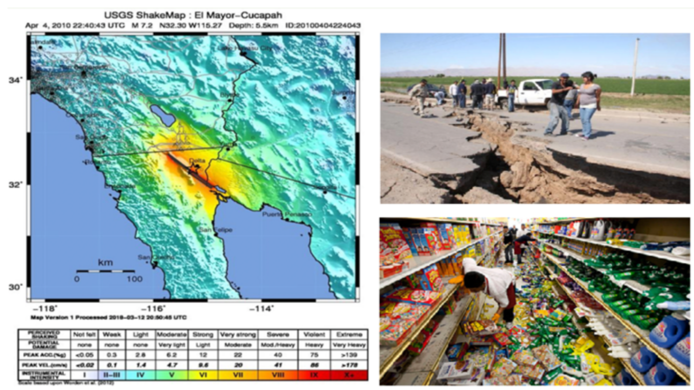


Figure 4.14. Pictures of Easter earthquake – 2010.

4.1.3. Year-wise Analysis

Year-wise analysis is a valuable technique for understanding seismic activity and observing trends and patterns over time. The significant earthquakes that happened across the US in the years 2002, 2010, 2015, 2018, and 2019 were the focus of this section. These years saw a number of significant earthquakes that resulted in significant destruction and fatalities, underlining the importance of continued research and readiness initiatives in the field of seismology.

Year - 2002

The map shown figure 4.15 shows the total earthquakes in 2002 with magnitude above 3. We can observe that density of earthquakes is more seen in Alaska and California regions.

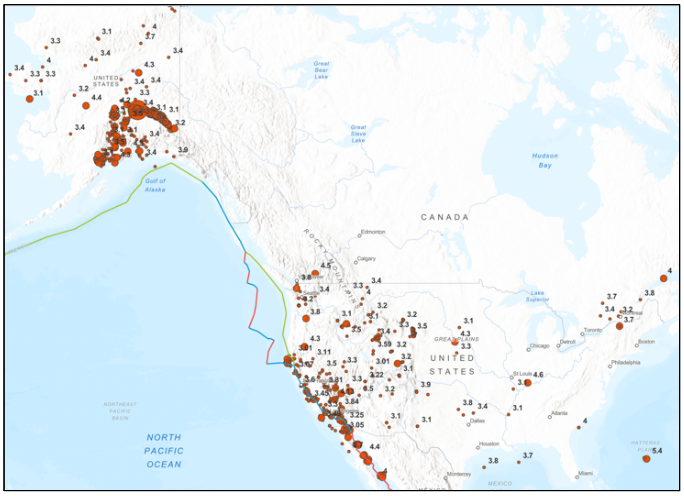


Figure 4.15. Map of the US showing all earthquakes in 2002 with magnitude above 3.

The graph shown in 4.16 depicts the earthquake with the highest magnitude and depth that occurred on each day of the year 2002. It also displays how many earthquakes occurred each day in 2002. It is clear that on November 4th, there were 535 earthquakes. As previously reported, an M7.9 earthquake was occurred in the same month. In addition, on July 8 there was a 195.7-kilometer-deep earthquake.

The Denali fault was the site of an earthquake of a magnitude of 6.7 on October 23, 2002. The Denali fault ruptured 45 km (28 mi) in length as a result of the earthquake, but aerial observation failed to find a surface rupture. The epicentre of the mainshock is 10 km (6.2 mi) west of the location of this rupture. As a result, there were numerous small rockfalls and snow avalanches in the area.

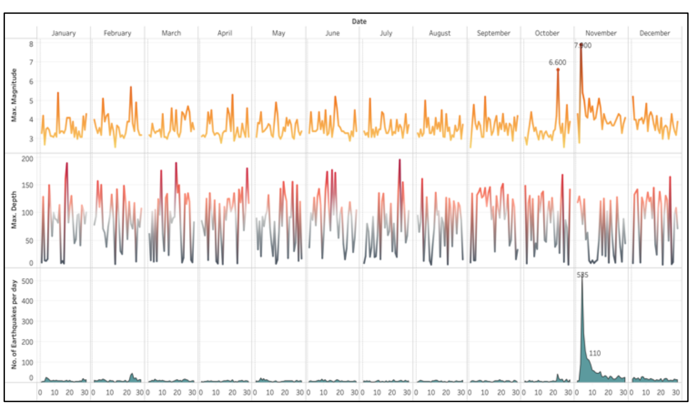


Figure 4.16. Time Series Analysis of the year 2002.

Figure 4.17 shows the earthquake patterns in November 2002, a month with a lot of quakes. On November 3, 2002, an earthquake with a maximum magnitude of 7.9 occurred in Alaska. Each earthquake that happened in Alaska during the month of November 2002 is shown on the map. Many of the aftershocks occurred on nearby faults that aren't believed to have ruptured and could only be allowing for stress changes. On the Denali fault itself, the largest aftershock was only a Mw 5.8 event, and there were fewer and smaller ones than expected.

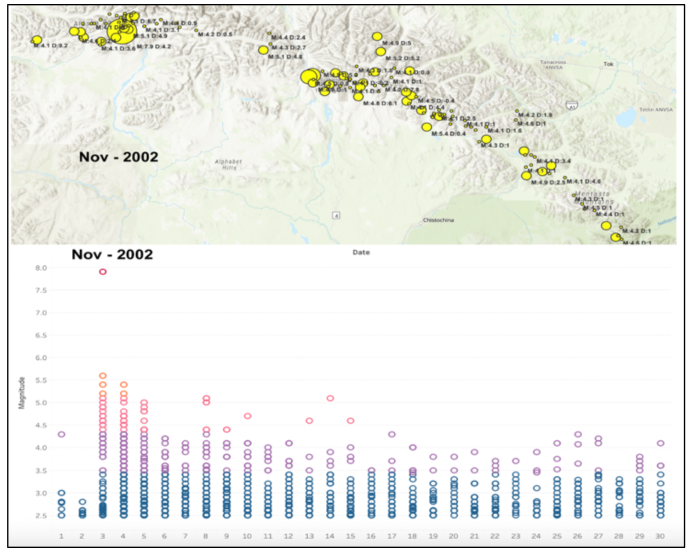


Figure 4.17. Earthquake trends in Nov-2002.

Year – 2010

The figure 4.18 shows the map of locations of all earthquakes with a magnitude greater than 3 in 2010. On tectonic plates, it can be seen that the California region experienced more earthquakes than other regions. In the vicinity of Mexico, a major earthquake measuring M7.2 occurred.

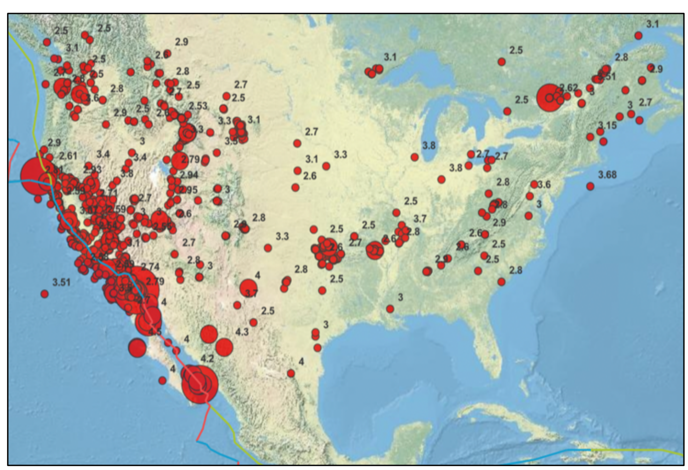


Figure 4.18. Map of the US showing all earthquakes in 2010 with magnitude above 3.

The graph in figure 4.19 displays the day's largest and deepest earthquakes for the entire year of 2010. It also displays how many earthquakes there were each day in 2010. It is evident that three significant earthquakes with magnitudes of 6.5, 7.2, and 6.7 occurred on January 10, April 4, and October 21, respectively. These earthquakes' depths weren't all that noteworthy, though. Although the strongest magnitude earthquake that day was just 3.3M, the 206 km deep earthquake was recorded as the deepest earthquake of the year. It should be mentioned that in April, a month with significant earthquake, there were 664 aftershocks.

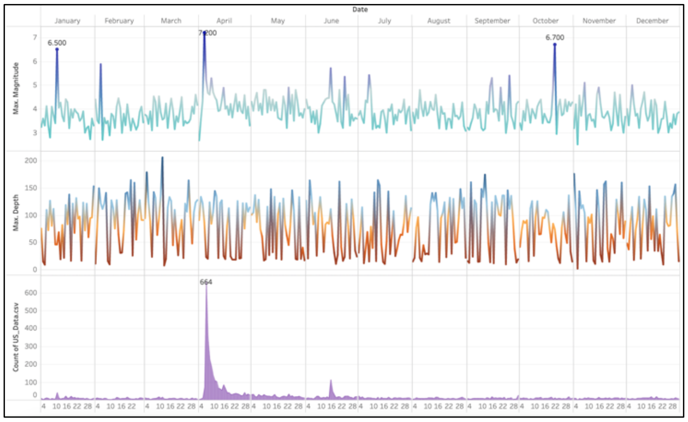


Figure 4.19. Time Series Analysis of the year 2010.

The figure 4.20 the earthquake patterns in April 2010, a month with a lot of earthquakes. The map displays all earthquakes that occurred in April of 2010 in Baja California region. Magnitude of 5.2, 5.4, 5.1, and magnitude of 5.7, all four within an hour, have been reported as aftershocks after 7.2 magnitude earthquake on April 4. In total, there were 9 major aftershocks. 6 hours after the first earthquake, nearly 90 aftershocks or generated earthquakes with magnitudes ranging from 3.0 to 5.1 were occurred in Southern California and northern Baja California.

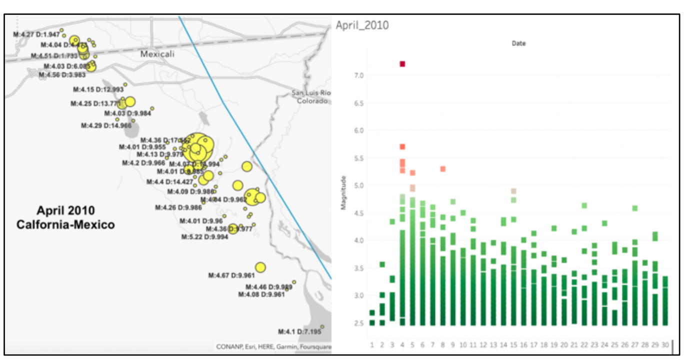


Figure 4.20. Earthquake trends in April-2010.

Year - 2015

All US earthquakes in 2010 with a magnitude greater than 3 are displayed on the map shown in figure 4.21. It was evident that there were many little earthquakes close to Oklahoma State. Significant amount earthquakes were also felt in Alaska and California.

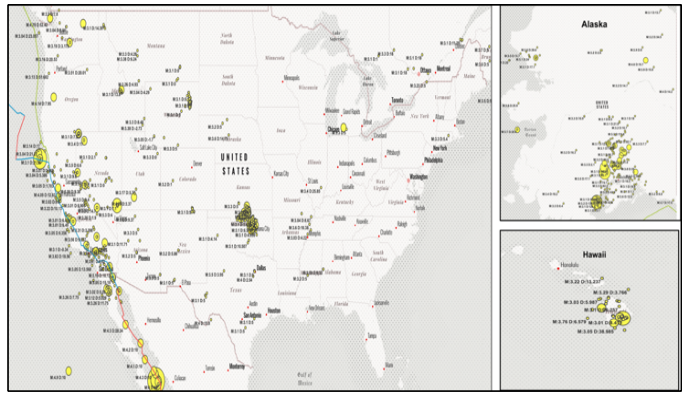


Figure 4.21. Map of the US showing all earthquakes in 2015 with magnitude above 3.

The figure 4.22 displays the day's largest and deepest earthquakes for the entire year of 2015. It also shows number of earthquakes in each day of year 2010. It was noted that there were 35 earthquakes on an average each day. On July 29, at 119.3 kms depth , a magnitude 6.4 earthquake occurred in Alaska, US, 52 km (32 mi) east northeast of Pedro Bay (74.1 mi) and on September 13, at 10.0 km depth, a magnitude 6.7 earthquake occurred off the coast of Mexico 95 kilometres (59 miles) southwest of Topolobampo (6.2 mi). Many earthquakes of depth over 150 kms were occurred in the month of July and around three in the month of October.

Figure 4.22. Time Series Analysis of the year 2015.

Year - 2018

All earthquakes of a magnitude greater than 3 that occurred in significant locations in 2018 are depicted on the map shown in figure 4.23. Significant amount of earthquakes can be seen in California, Alaska and Oklahoma regions.

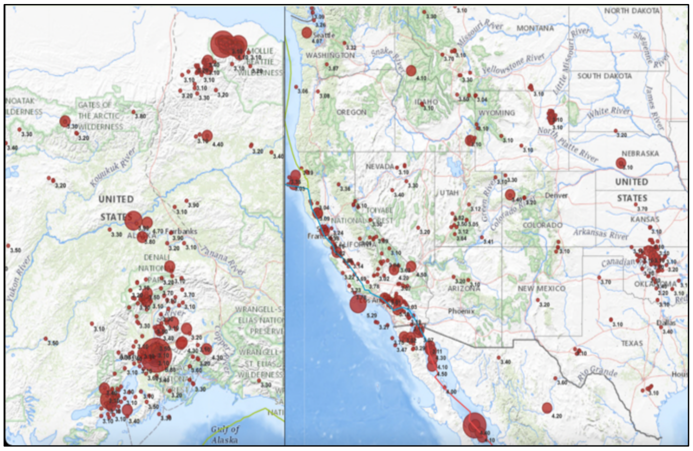


Figure 4.23. Map of the US showing all earthquakes in 2018 with magnitude above 3.

The figure 4.24 displays the day's largest and deepest earthquakes for the entire year of 2018. Moreover, it displays the daily earthquake total. It can be seen that there were more than 200 earthquakes on the 13th of August and the 30th of November. On May 4, 2018, a large magnitude-6.9 earthquake shook the island of Hawaii on the south slope of Kilauea Volcano. This was the biggest earthquake in last 43 years in Hawaii. Smaller-magnitude earthquakes are still happening in the same region more than five months later.

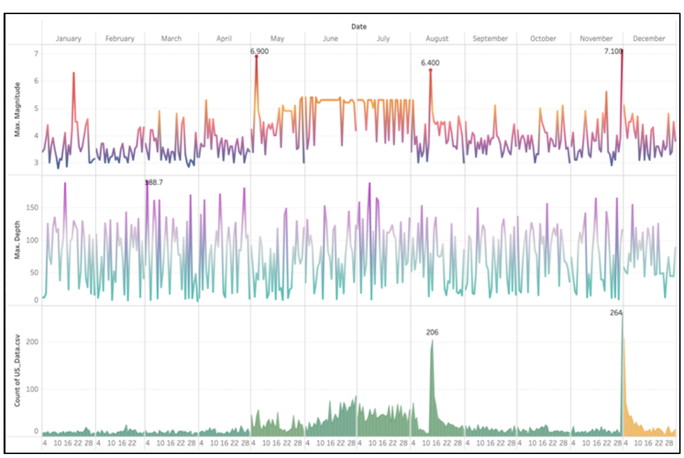


Figure 4.24. Time Series Analysis of the year 2018.

Year- 2019

Every earthquake in 2019 with a magnitude greater than 3 is depicted on the map in figure 4.25. Two significant earthquakes struck California this year, the state endured widespread power outages due to dry weather and wildfires, and homelessness in Los Angeles became a major problem.

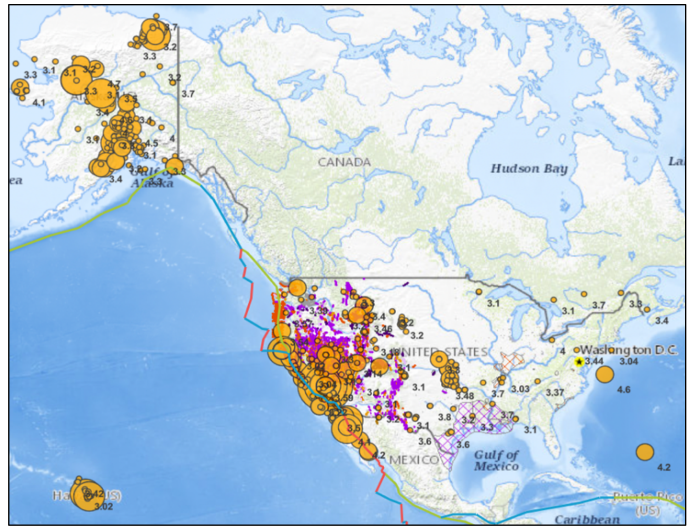


Figure 4.25. Map of the US showing all earthquakes in 2019 with magnitude above 3.

The figure 4.26 displays the day's largest and deepest earthquakes for the entire year of 2019. Additionally, it displays the daily earthquake total. It can be seen that when the M7.1 earthquakes happened on July 6th, there were approximately 1250 earthquakes overall. Moreover, there are about 15 earthquakes with a depth of more than 150 km. Learn more about major earthquakes in this year in region-wise analysis.

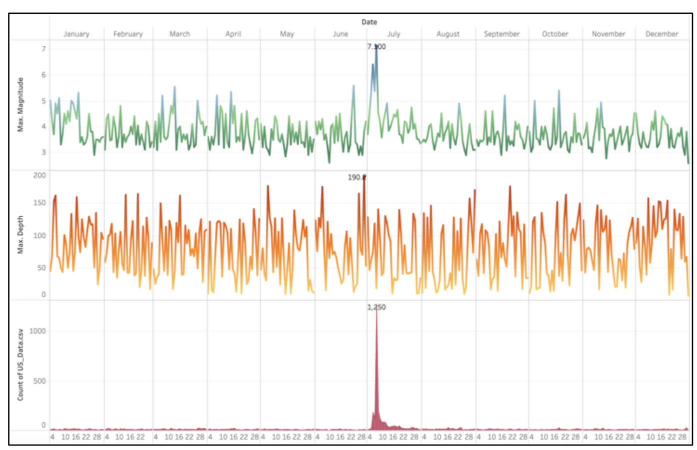


Figure 4.26. Time Series Analysis of the year 2019.

Due to the significant number of earthquakes that occurred in July 2019, the figure 4.26 reflect the patterns in earthquake activity during that month. A major earthquake of M6.4 and M7.1 were struck on 4th and 6th of July in the California. It is also noticed that number earthquakes during initial days of the month were high. However, USGS claimed that "*A temporary increase or decrease in seismicity is part of the normal fluctuation of earthquake rates. Neither an increase nor decrease worldwide is a positive indication that a large earthquake is imminent*" [16].

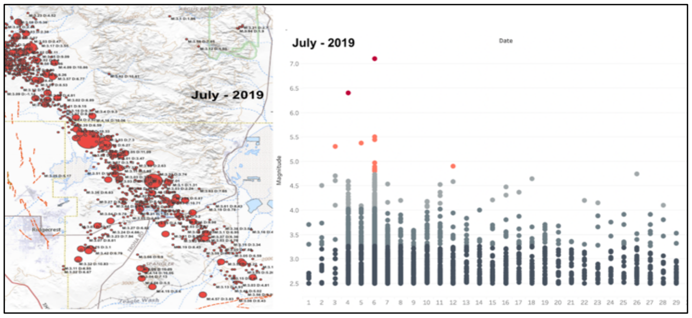


Figure 4.27. Earthquake trends in July-2019.

4.1.4. Region-wise Analysis

California Regional Analysis

In this analysis, we are going to analyse the areas which were impacted by earthquakes, shake intensity, and after effect of earthquakes, and lastly will analyse the highest magnitude earthquake in California in detail. The map in figure 4.28 shows all earthquakes with a magnitude greater than 4 in California.

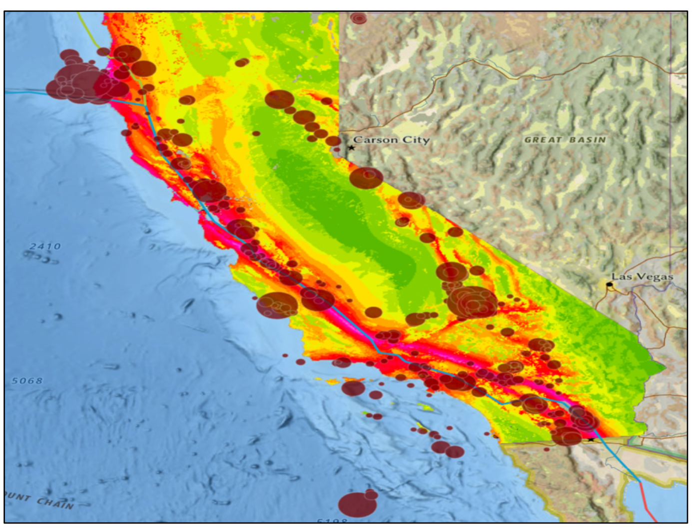


Figure 4.28. Map of the California showing all earthquakes from 2000 to 2023 with magnitude above 4.

The density of earthquakes of various magnitudes across California would be displayed on a density map of earthquake magnitudes in figure 4.29. California is one of the world's seismically active regions, with multiple faults running across it, notably the widely known San Andreas Fault. The density map would reveal that California has a lot of earthquakes with magnitudes under 2.0 that are rarely felt by people. The frequency of earthquakes declines as the magnitude rises. The map would demonstrate that earthquakes with a magnitude of 2.0 to 3.0 are more frequent than earthquakes with a magnitude of 3.0 to 4.0, and so on. The mean line in green color is represented as a straight line on the density map, with areas above the line showing a higher density of data points, while areas below the line indicate a lower density.

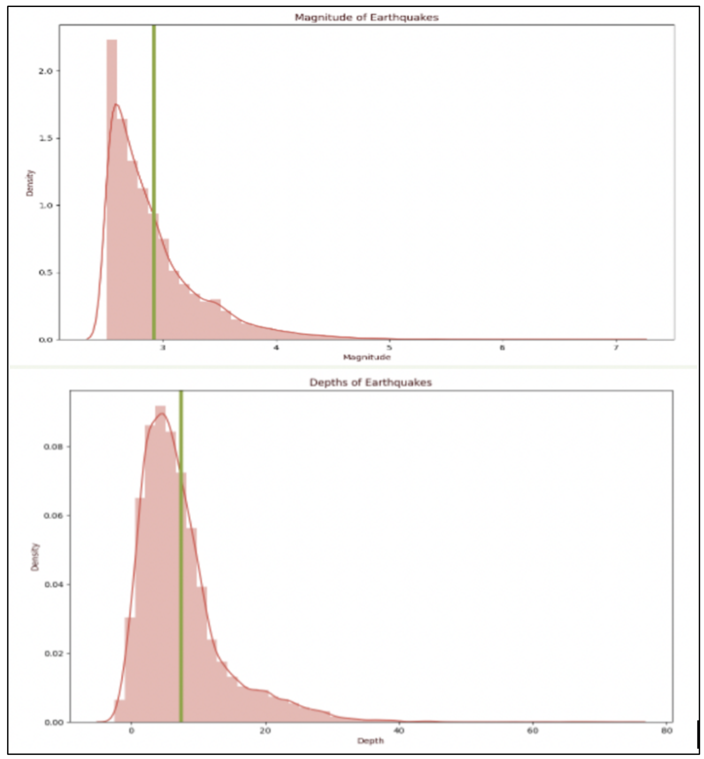


Figure 4.29. Density map of earthquake magnitude and depth.

The density map would demonstrate the range of depths at which earthquakes occur in California. The San Andreas Fault and other active faults in the state are frequently linked to the shallowest earthquakes (less than 10 km deep). Compared to earthquakes that happen at deeper depths, this depth's earthquake density is typically higher. The density map of earthquake depths in California would show the seismic activity in the state and the regions that are most vulnerable to earthquakes at various depths. It can be a helpful tool for figuring out how earthquakes occur in California and for locating places that might be more vulnerable to seismic hazards. In a density map of earthquake depths in California, the mean line in green color represents the average depth of earthquakes across the entire region.

The following layers are used from the ArcGIS to create map in figure 4.30.

Layer	Detail
CA Data	This layer shows the size of magnitude of earthquakes.
Surface_Rupture_Ridgecrest_Prov_rel_1	Documentation of Surface Fault Rupture and Ground-Deformation Features Produced by the 4 and 5 July 2019 Mw 6.4 and Mw 7.1 Ridgecrest Earthquake Sequence.
Shelly_stations	A High-Resolution Seismic Catalog for the Initial 2019 Ridgecrest Earthquake Sequence - Foreshocks, Aftershocks, and Faulting Complexity.
lidar_012020 & lidar_102019	Airborne Lidar and Electro-Optical Imagery along Surface Ruptures of the 2019 Ridgecrest Earthquake Sequence, Southern California.
USGS_ESC_Stations & USGS_ASL_Stations	The U.S. Geological Survey's Rapid Seismic Array Deployment for the 2019 Ridgecrest Earthquake Sequence.

The map in figure 4.30 shows earthquakes that occurred in July 2019 in California.

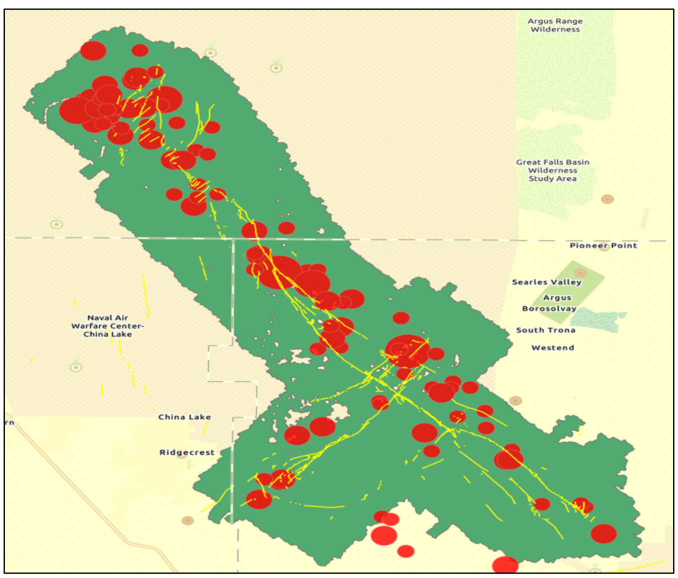


Figure 4.30. A map of California Ridgecrest earthquakes.

Figure 4.31 depicts the 6.4 magnitude earthquake with a centre located 18.2 kms (11.3 mi) west-southwest of Ridgecrest, California, which jolted more than 47k individual in southern California, as well as those in Northern-California and Arizona, Phoenix at 10:34am local time on July 4, 2019. A series of extremely minor earthquakes (foreshocks) lasting over an hour, including an M4.0 event that occurred approximately 30 minutes earlier, preceded the M6.4.

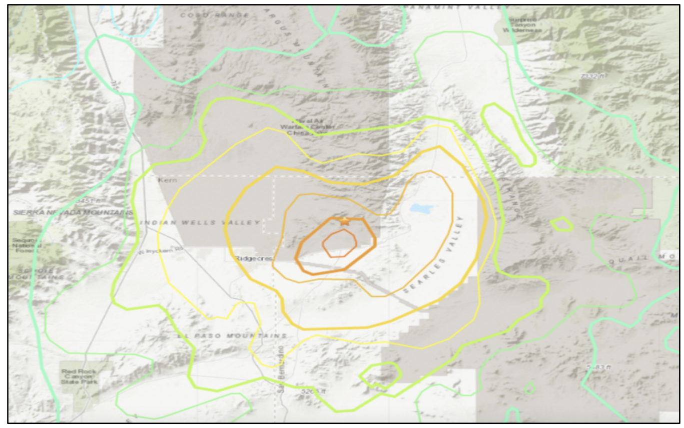


Figure 4.31. Intensity map of 6.4M.

According to figure 4.32, the early information suggested that the earthquake occurred on a shallow strike slip fault, and later aftershocks revealed a fault zone that extended from the southwest to the northeast. Following the 6.4M, there were around 250 2.5M or bigger aftershocks, but it wasn't until 34 hours later, at 08:19 pm local-time, and 11 kms (6.8 mi) to the northwest of the 6.4M event, that another, more significant earthquake struck. This 7.1M earthquake was a shallow strike-slip event that occurred on a fault that was orthogonal to the strike of the 6.4M (rotated 90 degrees).

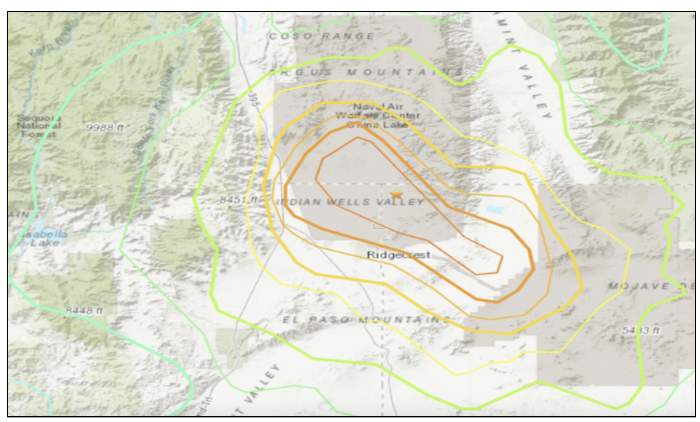


Figure 4.32. Intensity map of 7.1M.

The figure 4.33 describes earthquakes that were recorded during the first three days of Ridgecrest earthquake sequence on July 2019. A circle is used to depict every earthquake, with a blue circle used to indicate earthquakes larger than M4.0. The horizontal and vertical axes, respectively, show the magnitude and time of each earthquake. There is a gap at low magnitudes following the M6.4 and M7.1 earthquakes because there are too many aftershocks happening for them all to be detected.

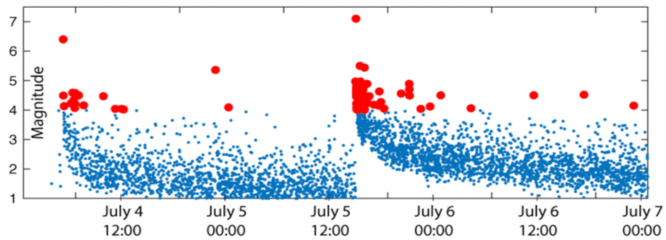


Figure 4.33. The earthquake sequence that occurred in Ridgecrest in July 2019.

Geologists quickly began gathering their tools to investigate any potential earth characteristics left over from the rupture and shaking as the second mainshock was starting. Seismologists were keen to place temporary equipment close to the epicentres of the 6.4M and 7.1M earthquakes in order to get information on the ensuing aftershock sequences due to the comparatively large distance of about 20km (12.4 mi) between permanent seismic stations in that area.

San Andreas Fault Line

A significant geological fault can be found in California, USA, called the San Andreas Fault. It is a transform fault, which means that the Pacific Plate and the North American Plate are separated by it. The fault stretches for approximately 800 miles (1,300 kms) through California, from the Salton Sea in its southern portion to Cape Mendocino in the north The city of San Francisco was completely destroyed by a 7.8M earthquake in 1906, which was the fault's most recent significant earthquake. Since then, there have been numerous smaller earthquakes along the fault, and scientists believe there is a substantial likelihood that a significant earthquake will happen soon.

Figure 4.34. Details about San Andreas fault line.

Alaska Regional Analysis

In this section, we are analysing the density maps of depth and magnitude, date-wise analysis of max. magnitude & number of earthquakes that occurred on a particular date, and lastly discuss the region and after-effects of the highest magnitude in Alaska.

The density of earthquakes at various magnitudes over Alaska is shown in figure 4.35. Because to its location at the meeting point of the Pacific and North American plates, Alaska is one of the seismically active areas in the globe. The purple mean line on an earthquake density map for Alaska reflects the average magnitude of earthquakes throughout the region.

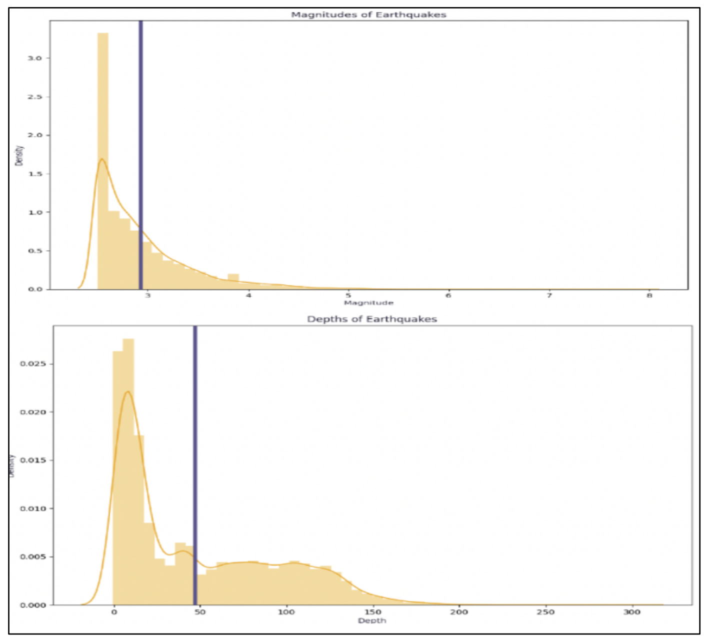


Figure 4.35. Density map of earthquake magnitude and depth of Alaska.

The density map of earthquake depths in Alaska would provide a visual representation of the state's seismic activity and the areas that are most prone to earthquakes at different depths. It can be a useful tool for understanding the distribution of earthquakes in Alaska and for identifying areas that may be at higher risk of seismic hazards. In a density map of earthquake depths shown in figure 4.35 in Alaska, the mean line in purple color represents the average depth of earthquakes across the entire region.

The line graph in figure 4.36 shows the maximum magnitude and total earthquakes that occurred on a specific date. Also, it has a filter by which we can filter out the specific range of earthquakes for better analysis.

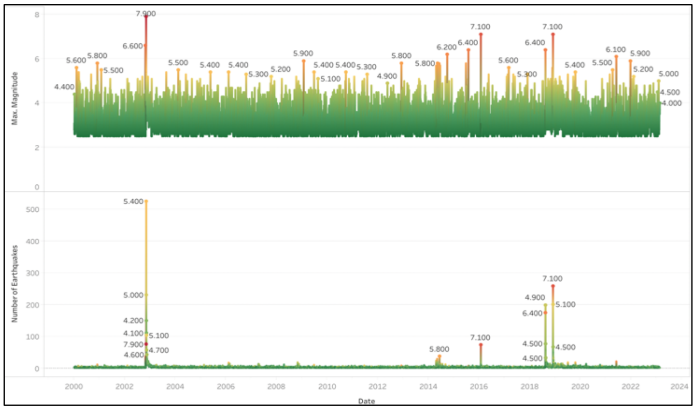


Figure 4.36. Magnitude Peaks of Earthquakes, by Date.

The table in figure 4.37 displays the highest magnitude earthquakes and the total number of earthquakes in Alaska by year. The darker color indicates the earthquakes with the highest magnitudes. If you wish to investigate the data for a specific year, such as 2002, which had the highest magnitude earthquake, you can click on it to access a month-wise analysis.

On November 3rd, 2002, a 7.9 magnitude earthquake occurred, which was the strongest in the past 23 years. If you would like to learn more about the impact and aftermath of the earthquake, you can click on the block corresponding to November 3rd. This will take you to a webpage with additional information about the Denali Fault earthquake.

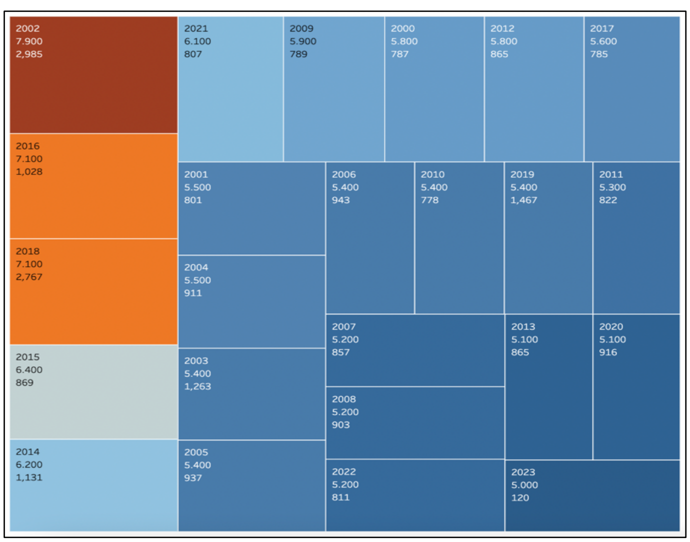


Figure 4.37. Alaska Earthquake Data by Years, Months, and Days.

4.2. Experimental Forecasting

An earthquake forecast is a prediction of the likelihood, timing, location, and magnitude of a future earthquake. Earthquake forecasting is a difficult topic because earthquakes are complicated occurrences that are difficult to anticipate completely. However, many tools and methodologies, such as statistical analysis of past earthquake data, monitoring of seismic activity, and modelling of the Earth's crust and faults, can be used to produce probabilistic estimations of earthquake occurrence.

Earthquake forecasting is significant because it can provide people living in earthquake-prone areas with early notice, allowing them time to prepare and potentially avoid harm or death. It can also aid in the planning and design of earthquake-resistant structures and infrastructure. Here are a few approaches that have been explored for experimental earthquake forecasting:

Seismic monitoring: Seismometers can be used to measure ground vibrations and detect patterns that may indicate seismic activity. This data can be analysed using various techniques, such as clustering, to identify potential precursors of earthquakes.

Satellite monitoring: Satellites can be used to detect changes in the Earth's surface that may be indicative of potential seismic activity, such as ground deformation or changes in the Earth's magnetic field.

Animal behaviour: Some animals have been observed to exhibit unusual behaviour prior to earthquakes, such as changes in their feeding patterns or movements. Monitoring animal behaviour in earthquake-prone areas may provide insight into potential seismic activity.

Machine learning: On the basis of several features, such as location, magnitude, and time, machine learning algorithms can be used to forecast the likelihood of future earthquakes.

From the resources available to us and based on our computer science knowledge, we have deployed some machine learning techniques to provide an experimental forecast of the quakes from historic data.

4.2.1. Clustering

Clustering can be helpful in earthquake forecasting by recognizing patterns and trends in data that may predict the risk of future earthquakes. Clustering techniques can be used to group earthquakes based on their location, magnitude, depth, and other properties. This can assist in identifying clusters of earthquakes that occur in a certain region or along a fault line, which may indicate increasing seismic activity in that area. It may be feasible to detect variations in seismic activity that could be an early warning indication of an upcoming earthquake by monitoring these clusters over time.

K-Means Clustering

K-means clustering is a unsupervised learning algorithm in machine learning and data science. In the context of earthquake forecasting, the algorithm works by grouping together data points that are similar to each other based on a set of features or characteristics. For example, earthquakes can be based on their location, magnitude, depth, and time. The algorithm will start by selecting k randomly, where k is the number of clusters desired. Then, it will assign each earthquake to the nearest centroid, based on a distance metric, and calculate the mean of each cluster. The algorithm will iterate until the centroids no longer move significantly, resulting in k distinct clusters of earthquakes.

By analysing the resulting clusters, researchers and seismologists can gain insights into the patterns and trends of seismic activity. This can help to identify areas or fault lines where earthquakes are more likely to occur, and potentially provide early warning of impending seismic activity. For this study, earthquakes are clustered using the parameters - Location, Magnitude and Location, Depth separately.

Fig.4.38 and 4.40 shows the elbow curve to select the optimal number of characters to get maximum model performance. Here, Silhouette analysis is done and the K with the maximum Silhouette score is chosen as the number of clusters for grouping the data.

To create clusters for Latitude, Longitude and Magnitude, K=4 gives the maximum performance as can be seen from the elbow curve in Fig 4.38.

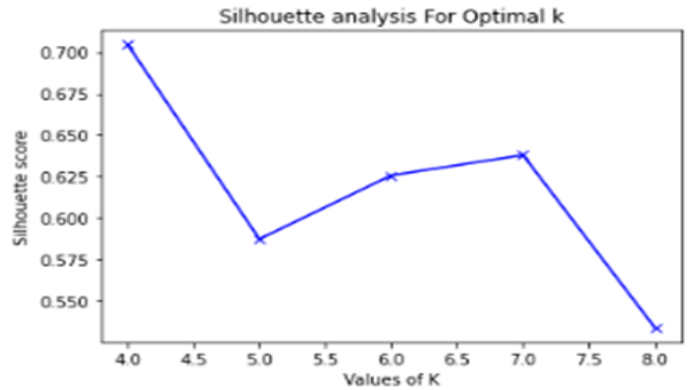


Figure 4.38. Elbow graph for Location and Magnitude.

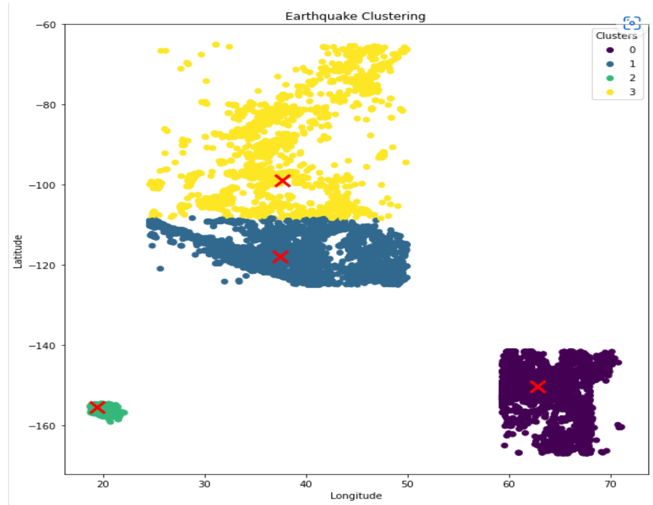


Figure 4.39. Clusters with centroid for Location and Magnitude

When applied on the data, a clear cluster formation with the cluster's centroid can be seen in the figure 4.39.

To create clusters for Latitude, Longitude and Depth, K=4 gives the maximum performance as can be seen from the elbow curve in Fig 4.40.

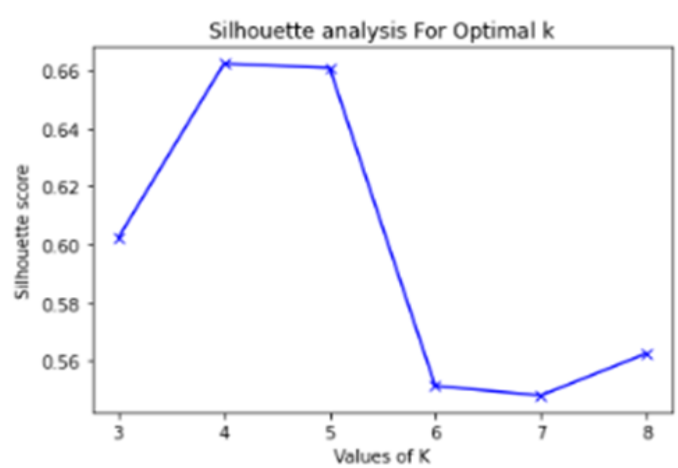


Figure 4.40. Elbow graph for Location and Depth.

The figure 4.41 shows the clusters formed with the data. These clusters are not as defined as the clusters in Figure 4.39. Some data points from yellow and purple clusters can be seen in blue cluster. This might be because the depth of an earthquake does not depend on the location where the earthquake occurs. Whereas, some locations have a similar trend of earthquake magnitudes.

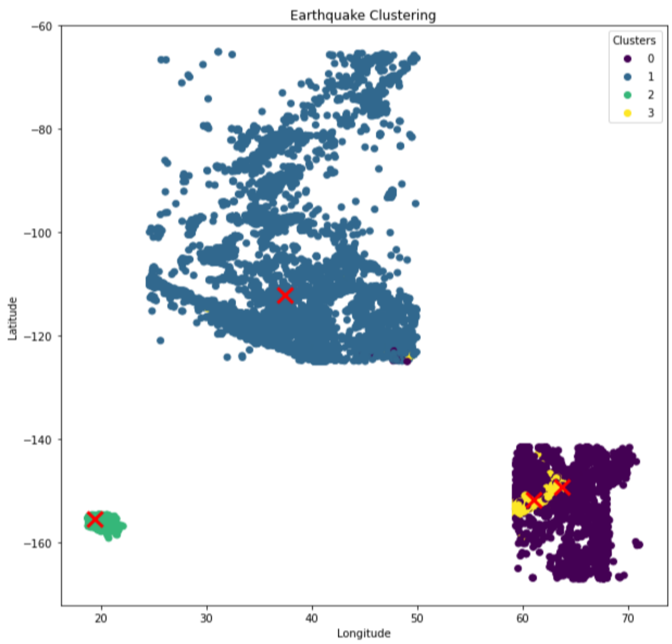


Figure 4.41. Clusters with centroid for Location and Depth.

The figure 4.42 captures the maximum and minimum depth and magnitude of the earthquakes in each cluster. Clusters 0 and 1 capture the events that have maximum magnitude whereas Cluster 3 in the second part of the algorithm captures the earthquakes with maximum depth. The locations in these clusters require more attention and measures to deal with earthquakes in future.

```
Cluster : 0
Minimum Magnitude : 2.5
Maximum Magnitude : 7.9
Cluster : 1
Minimum Magnitude : 2.5
Maximum Magnitude : 7.2
Cluster : 2
Minimum Magnitude : 2.5
Maximum Magnitude : 6.9
Cluster : 3
Minimum Magnitude : 2.12
Maximum Magnitude : 5.9
```

```
Cluster : 0
Minimum Depth : -0.5
Maximum Depth : 59.8
Cluster : 1
Minimum Depth : -3.48
Maximum Depth : 60.47
Cluster : 2
Minimum Depth : -3.74
Maximum Depth : 63.03
Cluster : 3
Minimum Depth : 59.8
Maximum Depth : 300.0
```

Figure 4.42 Clusters and their Max and Min values.

Three different types of performance scores are calculated for the algorithms as shown in figure 4.43. Both the implementations have good silhouette scores (ranges from [-1,1]) and are closer to the higher range of the matrix.

Cluster Features / Metrics	Lat, Long, Mag	Lat, Long, Depth
Silhouette Score	0.662219227	0.704608764
Calinski_harabasz_score	176613.946	529513.0904
Davies_bouldin_score	0.549742097	0.376954952

Figure 4.43. Clustering Performance Metrics.

4.2.2. Time Series Forecasting

Time series analysis can be used to detect patterns and trends that may be useful for earthquake forecasting. However, it is important to note that there is no single method that can reliably predict when and where earthquakes will occur.

One approach to using time series analysis for earthquake forecasting is examining the frequency of earthquakes over a period of time and then analysing these patterns to identify certain trends or changes in seismic activity that could indicate an increased likelihood of an earthquake occurring.

Another approach is to use machine learning techniques, such as neural networks, to analyse seismic data and identify patterns that may be useful for earthquake forecasting.

The data for time series algorithms is prepared from the original US dataset and consists of two columns: Date, Count where count is the number of earthquakes that occurred that day as shown in figure 4.44.

Date	Count
2000-01-01	9
2000-01-02	6
2000-01-03	12
2000-01-04	6
2000-01-05	9

Figure 4.44 Data for time-series forecasting models.

__ARIMA (Autoregressive integrated moving average)__

ARIMA is a popular time series forecasting model used to predict future values based on past data. In the context of earthquake forecasting, ARIMA can be used to analyse historical earthquake data and forecast the likelihood and magnitude of future earthquakes. The ARIMA model works by identifying and modelling the patterns and trends in the historical data. It consists of three components:

<u>Autoregression (AR)</u>: This component represents the relationship between a lagged value a from the past and time series' present value. The present value of the time series is related to the values from the two prior time points, for instance, if the AR component has a value of 2.

<u>Integration (I)</u>: The time series' trend, which may be linear or nonlinear, is modelled by this component. It is applied to make the time series stationary, which ensures that its statistical characteristics, such as the average and variance, remain constant over time.

<u>Moving Average (MA)</u>: This component models the relationship between the moving average of the past values and the current value of the time series. For example, if the MA component has a value of 2, it means that the current value of the time series is related to the moving average of the two previous time points.

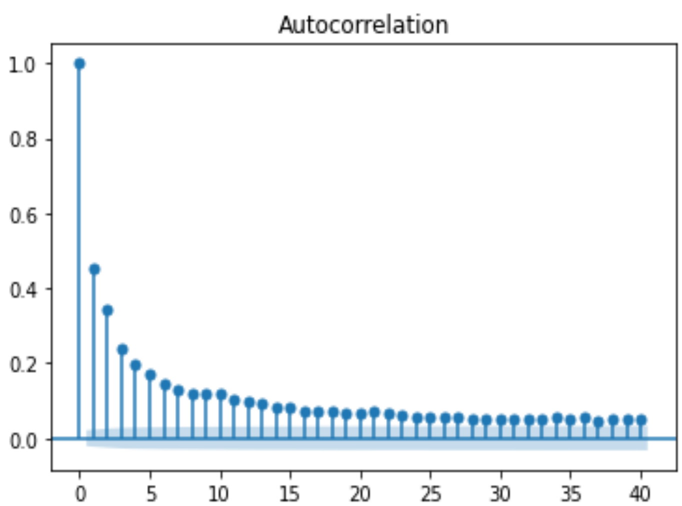


Figure 4.45. Autocorrelation graph for ARIMA.

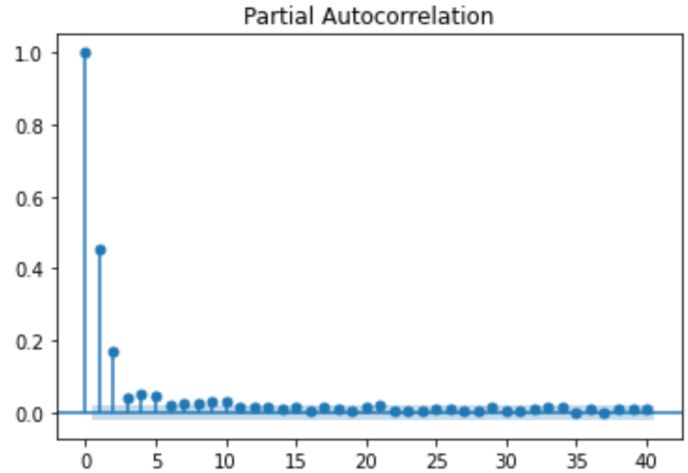


Figure 4.46. Partial Autocorrelation graph for ARIMA.

ARIMA models can be customized by adjusting the values of components, which can be determined through a process called parameter tuning. Once the model is trained on the historical data, it can be used to make forecasts for future time periods.
A nonseasonal ARIMA model is defined by the following parameters:

p	number of autoregressive terms
d	number of nonseasonal differences neededfor stationarity
q	number of lagged forecast errors in the prediction equation.

```
ARIMA(0, 0, 0) ,  RMSE=11.884
ARIMA(0, 0, 1) ,  RMSE=11.885
ARIMA(0, 0, 2) ,  RMSE=11.883
ARIMA(0, 1, 0) ,  RMSE=17.148
ARIMA(0, 1, 1) ,  RMSE=12.161
ARIMA(0, 1, 2) ,  RMSE=12.135
ARIMA(0, 2, 0) ,  RMSE=17343.762
ARIMA(0, 2, 1) ,  RMSE=17.888
ARIMA(0, 2, 2) ,  RMSE=12.145
ARIMA(1, 0, 0) ,  RMSE=11.884
ARIMA(1, 0, 1) ,  RMSE=11.883
ARIMA(1, 0, 2) ,  RMSE=11.883
ARIMA(1, 1, 0) ,  RMSE=12.780
ARIMA(1, 1, 1) ,  RMSE=12.068
ARIMA(1, 1, 2) ,  RMSE=12.476
ARIMA(1, 2, 0) ,  RMSE=6115.451
ARIMA(1, 2, 1) ,  RMSE=12.773
ARIMA(1, 2, 2) ,  RMSE=12.977
ARIMA(2, 0, 0) ,  RMSE=11.883
ARIMA(2, 0, 1) ,  RMSE=11.883
ARIMA(2, 0, 2) ,  RMSE=11.884
ARIMA(2, 1, 0) ,  RMSE=12.408
ARIMA(2, 1, 1) ,  RMSE=11.757
ARIMA(2, 1, 2) ,  RMSE=12.457
ARIMA(2, 2, 0) ,  RMSE=2732.268
ARIMA(2, 2, 1) ,  RMSE=12.392
ARIMA(2, 2, 2) ,  RMSE=17.515
```

Figure 4.47 RMSE scores for combinations of (p, d, q).

The values for the parameters (p, d, q) can be estimated from the Autocorrelation and Partial Autocorrelation graphs in Figure 4.45 and 4.46. However, the model is run on different combinations to get the best result.

The figure 4.47 shows the RMSE (Root Mean Squared Error) scores calculated for combinations for p, d, q values. This was done to find the values that give the minimum residual error while prediction. The best performing model gives the RMSE value of 11.757 for p = 2, d = 1 and q =1.

However, even though the algorithm made some predictions on the unseen data, it was unable to capture the trend and seasonality of the historical data. Which gives a straight line when plotted against the actual data as seen in Figure 4.48.

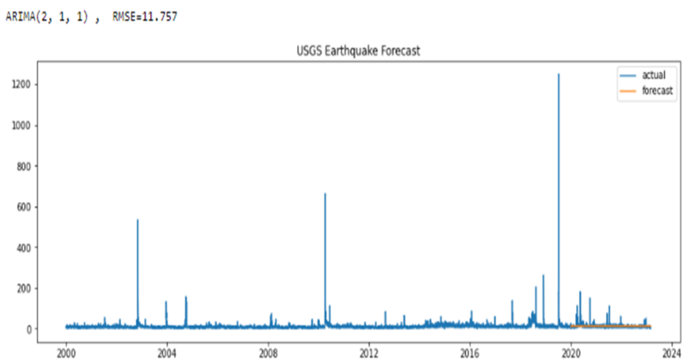


Figure 4.48 Actual number of earthquake vs forecasted.

Exponential Smoothing

Holt-Winters exponential smoothing is a forecasting approach that predicts future values by using weighted averages of prior data. It is a versatile strategy that can anticipate a wide range of time series, including trended, seasonal, or both. The Holt-Winters model is estimated using a recursive approach, which means that as new data becomes available, the estimates of the trend, seasonality, and irregularity components are updated. As a result, the model is well-suited for forecasting time series that change over time.

Holt-Winters exponential smoothing has been effectively applied to forecasting a wide range of time series, including economic data, weather data, and stock prices. It has also been used to predict seismic activity. It is a three-parameter model, which means that it requires three parameters to be estimated:

Trend: The trend component of the model represents the long-term direction of the time series.

Seasonality: The seasonality component of the model represents the regular fluctuations in the time series that occur over a fixed period of time, such as daily, weekly, or monthly.

Irregularity: The time series' random fluctuations that cannot be explained by the trend or seasonality factors are represented by the model's irregularity component.

In one study, Holt-Winters exponential smoothing was used to forecast the number of earthquakes in California. The model was able to forecast the number of earthquakes with a high degree of accuracy, even when the earthquakes were clustered in time.[12]

Taking this reference, we built a model on the US data by setting different values of hyper parameters: **trend, seasonal, seasonal_periods** and parameter of fit function: **remove_bias.** Figure 4.50 shows the RMSE scores for different combinations of parameters. More than one combination of hyperparameters gave the best RMSE score (12.16082). However, we found an interesting observation with the results.

```
> Model params : ['add', 'add', 30, True] , RMSE : 13.43272
> Model params : ['add', 'add', 30, False] , RMSE : 13.43201
> Model params : ['add', 'add', 60, True] , RMSE : 13.37217
> Model params : ['add', 'add', 60, False] , RMSE : 13.37173
> Model params : ['add', 'add', 180, True] , RMSE : 19.62049
> Model params : ['add', 'add', 180, False] , RMSE : 19.62116
> Model params : ['add', 'add', 365, True] , RMSE : 42.82408
> Model params : ['add', 'add', 365, False] , RMSE : 42.82809
> Model params : ['add', None, 30, True] , RMSE : 38.33356
> Model params : ['add', None, 30, False] , RMSE : 38.34498
> Model params : ['add', None, 60, True] , RMSE : 38.33356
> Model params : ['add', None, 60, False] , RMSE : 38.34498
> Model params : ['add', None, 180, True] , RMSE : 38.33356
> Model params : ['add', None, 180, False] , RMSE : 38.34498
> Model params : ['add', None, 365, True] , RMSE : 38.33356
> Model params : ['add', None, 365, False] , RMSE : 38.34498
> Model params : [None, 'add', 30, True] , RMSE : 13.10435
> Model params : [None, 'add', 30, False] , RMSE : 13.10449
> Model params : [None, 'add', 60, True] , RMSE : 13.23410
> Model params : [None, 'add', 60, False] , RMSE : 13.23420
> Model params : [None, 'add', 180, True] , RMSE : 14.63084
> Model params : [None, 'add', 180, False] , RMSE : 14.63047
> Model params : [None, 'add', 365, True] , RMSE : 13.21652
> Model params : [None, 'add', 365, False] , RMSE : 13.21672
> Model params : [None, None, 30, True] , RMSE : 12.16082
> Model params : [None, None, 30, False] , RMSE : 12.16107
> Model params : [None, None, 60, True] , RMSE : 12.16082
> Model params : [None, None, 60, False] , RMSE : 12.16107
> Model params : [None, None, 180, True] , RMSE : 12.16082
> Model params : [None, None, 180, False] , RMSE : 12.16107
> Model params : [None, None, 365, True] , RMSE : 12.16082
> Model params : [None, None, 365, False] , RMSE : 12.16107
```

Figure 4.49 RMSE scores for different values of parameters.

Figures 4.50 through 4.52 visualizes the actual number of earthquakes against the forecasted value for different values of RMSE. It is interesting to see that the best model (Figure 4.50) misses the trend and seasonality of the earthquakes and makes no reliable predictions. The model with maximum RMSE (Figure 4.51) successfully captures the trend in data to a certain extent but forecasts the negative value which is not possible. Moving further, the second best model with RMSE 13.21652 (Figure 4.52) is the best model in terms of forecasting and capturing the trend in data. Here, it is safe to compromise a little on RMSE metrics because the model forecasts the future values that can be used to mitigate the consequences of earthquakes.

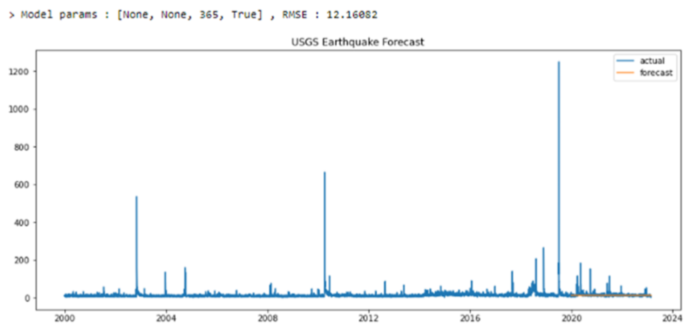


Figure 4.50 Actual number of earthquake vs forecasted for minimum RMSE.

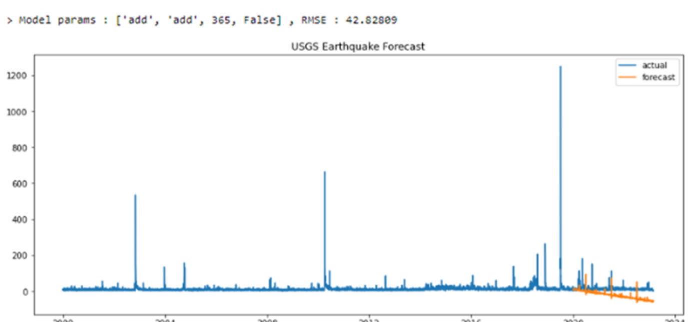


Figure 4.51 Actual number of earthquake vs forecasted for maximum RMSE.

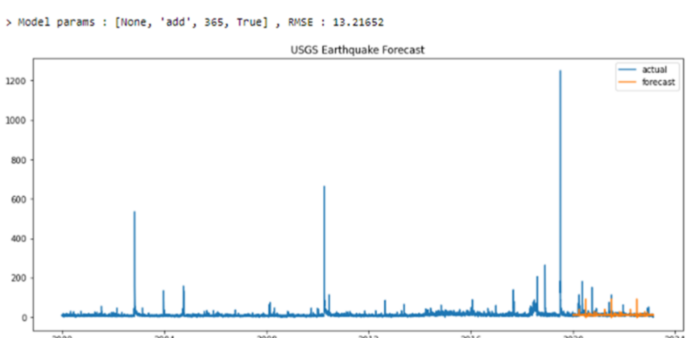


Figure 4.52 Actual number of earthquake vs forecasted for second best RMSE.

LSTM (Long Short-Term Memory)

Time series forecasting, such as earthquake forecasting, is a common use of LSTM, a form of recurrent neural network (RNN). The model analyses the input data sequence one time step at a time and produces an output for that time step. Particularly applicable to time series data, which may contain complicated links between past and future values, the LSTM model's capacity to learn long-term dependencies in the data is one of its important strengths.

The LSTM is able to accomplish this by employing a memory cell and gates to regulate the influx and egress of data into and out of the cell. At each time step, the outputs are influenced by the memory cell's stored knowledge about previous inputs. The model can learn to recall and forget data based on its relevance to the job at hand thanks to the gates' ability to regulate the flow of data into and out of the cell.

In the context of earthquake forecasting, the LSTM model can be trained on historical earthquake data to identify patterns and trends that are related to future earthquakes. LSTM models can be customized by adjusting the number of memory cells, the number of layers, and the training parameters. Once the model is trained, it can be evaluated based on its accuracy and used to make predictions for future earthquakes.

As can be seen in Figure 4.53, we begin the model implementation with very few parameters with default activation function and optimizer. By doing so, you may evaluate how well the model fits the data using the standard settings. The prior timesteps that were utilized to make predictions for the next timestep are represented by the lookback number 1. The network is made up of a single input layer, a hidden layer with 50 LSTM neurons, and a layer of output that predicts a single value. By default, the LSTM blocks employ the sigmoid activation function. The training of the network takes place across 20 epochs, and 64 samples are processed before the model is updated (the batch size).

```
model = Sequential()
model.add(LSTM(50, input_shape=(lookback, 1)))
model.add(Dense(1))
model.compile(loss='mean_squared_error', optimizer='adam')
```

Figure 4.53 LSTM Model-1 parameters.

The first iteration of the model gives the best result and forecasts the number of earthquakes almost with perfect accuracy as seen in Figure 4.54. The model underperforms for the higher earthquake counts but successfully learns noise in the data. This is the best performing model so far with **RMSE of 9.68**.

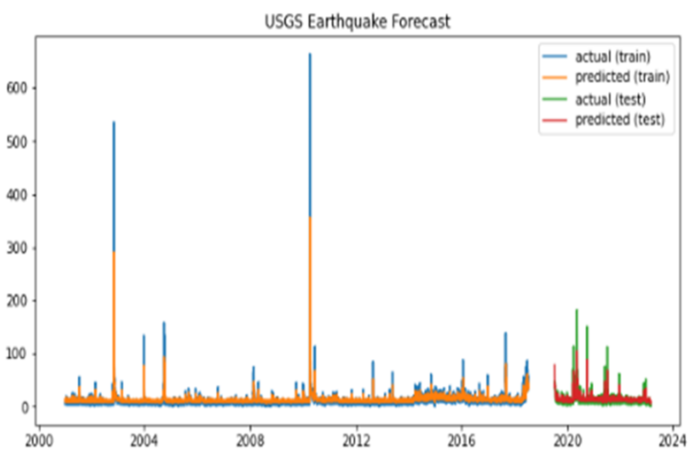


Figure 4.54 Actual number of earthquake vs forecasted for Model-1.

To further improve the performance and forecast, we tried to add more parameters for a well-defined model. A dropout value to regularize the model and a SoftMax activation function to transform the raw outputs into a vector of probabilities is added for better training. The additional parameters do not help and is unable to learn anything from data as shown in Figure 4.56.

```
model = Sequential()
model.add(LSTM(50, input_shape=(lookback, 1)))
model.add(Dense(1))
model.add(Dropout(0.2))
model.add(Activation('softmax'))
model.compile(loss='mean_squared_error', optimizer='adam')
```

Figure 4.55 LSTM Model-2 parameters.

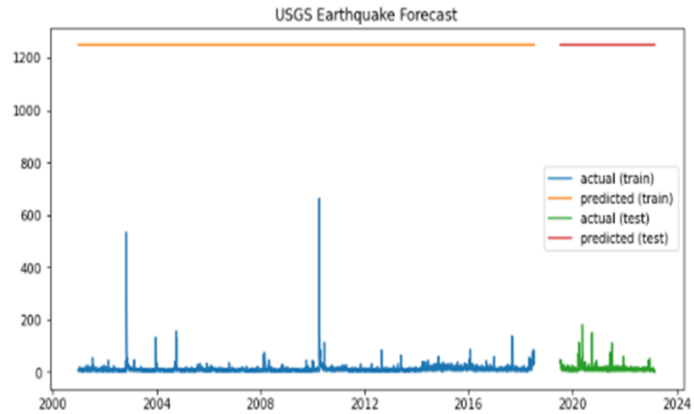


Figure 4.56 Actual number of earthquake vs forecasted for Model-2.

5. Discussion And Conclusion

This section shows the SWOT Analysis of the project and concludes with important learnings and analysis made throughout the project. It also covers how the project can be used and modified in future for better analysis and forecasting of earthquakes.

5.1. Conclusions

In conclusion, the analysis of earthquake data from the USGS has provided meaningful insights into the frequency, location, and magnitude of seismic activity in various parts of the US. We were able to identify the most seismically active places and comprehend the factors that lead to earthquake occurrence by studying the data.

The project also focuses on exploring the Machine Learning and Deep Learning Algorithms for forecasting the earthquakes. We use the word experimental because as we saw that predicting the earthquake is highly unlikely but understanding the trend and patterns of previous earthquakes can give guidance on how to prepare better for the future occurrences. Clustering identified the group of locations having similar magnitude earthquakes and some clusters contained the locations with highest magnitude. This information can be used to track the seismic activity of such regions more closely.

The time-series models like Exponential Smoothing and LSTM showed the capability of identifying the trend in data to some extent and even made predictions on unseen data. Using the high compute systems and more data can help improve the models' performance and the results can be used to proactively take actions.

5.2. SWOT Analysis

SWOT analysis is a strategy development tool used to evaluate a project's strengths, weaknesses, threats and opportunities, where strengths are intrinsic positive aspects of a project that give it a competitive advantage or value proposition. Weaknesses are the intrinsic, unfavourable aspects of a project that impede its ability to function effectively. Opportunities are externally derived positives that have the potential to be utilised to the advantage of the project. Threats are externally derived unfavourable factors that may provide difficulties or dangers for the undertaking. The strengths, weaknesses, opportunities, and threats facing our project are outlined in figure 5.1.

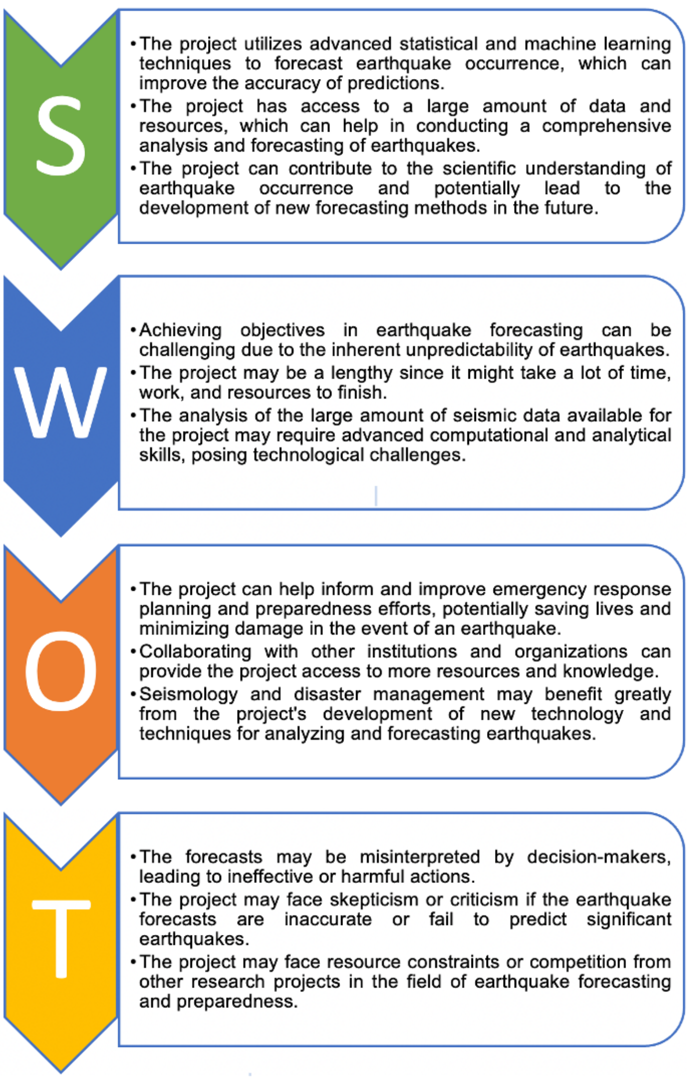


Figure 5.1. SWOT Analysis.

5.3. Future Work

In order to improve the accuracy of earthquake analysis and forecasting, it is important to consider the use of more data and better compute systems in future works. With access to more data, we can develop more comprehensive models that take into account a wide range of variables that may affect earthquake activity. Additionally, better compute systems can process and analyse larger datasets more efficiently, which can help us make more accurate and timely predictions about earthquakes.

The study uses the clustering and time-series models separately. to build a more robust model, these algorithms can be combined together for accurate predictions. The patterns identified from clustering can be used as an input to the time-series model for the locations that are prone to high intensity earthquakes. Understanding seismic activity patterns and trends will help us create communities that are more robust and less vulnerable to future earthquakes.

6. References

[1] *"What are the Effects of Earthquakes",* by Earthquake Hazards Program, USGS,
Link: https://www.usgs.gov/programs/earthquake-hazards/what-are-effects-earthquakes

[2] *"The 30 November 2018 Mw 7.1 Anchorage Earthquake",* by Michael E. West, Adrian Bender, Matthew Gardine and others, 2018
Link: https://par.nsf.gov/servlets/purl/10202260

[3] "*The 2019 Ridgecrest, California, Earthquake Sequence*", by Lisa Wald and Camille Collett, USGS, May 10, 2021
Link: https://earthquake.usgs.gov/storymap/index-ridgecrest.html

[4] *"Seismicity of the United States, 1568-1989 (Revised)*", by Carl W. Stover and jerry L. Coffman, USGS Professional Paper 1527
Link: https://pubs.usgs.gov/pp/1527/report.pdf

[5] *"Can you predict earthquakes?"*, by USGS
Link: https://www.usgs.gov/faqs/can-you-predict-earthquakes

[6] *"100% Chance of an Earthquake"*, by Earthquake Hazards Program, USGS,
Link: https://www.usgs.gov/programs/earthquake-hazards/100-chance-earthquake

[7] "*What Turkey's earthquake tells us about the science of seismic forecasting",* by nature article 06 March 2023
Link: https://www.nature.com/articles/d41586-023-00685-y

[8] "*What is the difference between earthquake early warning, earthquake forecasts, earthquake probabilities, and earthquake prediction?"*, by USGS article
Link: https://www.usgs.gov/faqs/what-difference-between-earthquake-early-warning-earthquake-forecasts-earthquake-probabilities

[9] "*Experimental analysis of earthquake prediction using machine learning classifiers, curve fitting, and neural modeling"*, by Pankaj Chittora, Tulika Chakrabarti, and others, September 1st, 2022
Link: https://assets.researchsquare.com/files/rs-1896823/v2/b190f239-8da8-4078-8cb5-0f18c2e17d0a.pdf?c=1663935282

[10] "*ARIMA Model Analysis of the Regularities of Earthquake Origin Times in the Longmen Mountain Fault Zone"*, by Xue Yuan, Hu Dan, and others, October 13th, 2022
Link: https://assets.researchsquare.com/files/rs-2118474/v1/0db5ecd9-5b81-4335-9a23-5ee13ccb95e5.pdf?c=1673963064

[11] "*Time Series Analysis for Predicting the Occurrences of Large Scale Earthquakes",* by Amei, Wandong Fu, Chih-Hsiang Ho, August 2012 Link: https://www.ijastnet.com/journals/Vol_2_No_7_August_2012/8.pdf

[12] *"Search Earthquake Catalog"*, by Earthquake Hazards Program, USGS,
Link: https://earthquake.usgs.gov/earthquakes/search/

[13] "*ANSS Comprehensive Earthquake Catalog (ComCat) Documentation*", by Earthquake Hazards Program, USGS
Link: https://earthquake.usgs.gov/data/comcat/

[14] "*USGS Contributors List*", by Earthquake Hazards Program, USGS
Link:
https://earthquake.usgs.gov/data/comcat/contributor

[15] *"Magnitude Types",* by Earthquake Hazards Program, USGS
Link: https://www.usgs.gov/programs/earthquake-hazards/magnitude-types

[16] "*Why are we having so many earthquakes? Has naturally occurring earthquake activity been increasing? Does this mean a big one is going to hit? OR We haven't had any earthquakes in a long time; does this mean that the pressure is building up for a big one?"*, by USGS
Link: https://www.usgs.gov/faqs/why-are-we-having-so-many-earthquakes-has-naturally-occurring-earthquake-activity-been

[17] "*Time series Forecasting using Holt-Winters Exponential Smoothing"*, by Prajakta S. Kalekar (04329008) Kanwal Rekhi, School of Information Technology, December 6, 2004
Link:
https://caohock24.files.wordpress.com/2012/11/04329008_exponentialsmoothing.pdf