# Statistical Analysis of Earthquake Detection


### Team Members
1. Devashi Ghoshal (NetID: dghoshal).
2. Kate Kang (NetID: kang232).
3. Maja Velichkovich (NetID: velichkovicm).
4. Navid Najmabadi (NetID: nnajmabadi).
5. Nevindu Batagoda (NetID: batagoda).


## Abstract


Earthquakes are one of the most impactful natural disasters which are difficult to prevent. As a result, detection technology has always been vital to gathering relevant data, knowing which regions are most susceptible to earthquakes, and making further improvements to technology. The following report is an analysis regarding earthquake detections with the improvement of such technologies over time. The question explored in this analysis was if there had been a significant increase in the frequency of earthquake detections over time, the detection of smaller magnitude earthquakes, and the nature of the relationship between the number of GSN stations and earthquake detections. The exploratory analysis was performed to learn the data and general trends and come up with the question. Then, the analysis began with testing, specifically with two-sample t-tests to test for a significant increase of earthquake detections before the establishment of GSN and while the number of stations is held the same. With this, there was a significant increase after the establishment but not a significant amount to conclude an increase between 2008 and 2020. To model and predict the relationship between GSN stations over time and earthquake detections, a linear model was created using the year, number of stations established by that year, and the interaction between those two as predictive terms. However, this model has room for improvement with more technological advances and other factors that provide more data. Afterward, to see if the GSN and year were the only driving factor of the increasing detection, an increase in the detection of smaller magnitude earthquakes over time was tested. The results demonstrated no significant evidence for an increase in smaller magnitude detections after the establishment of GSN stations as well.


## Motivation

While brainstorming we learned that we each are from different regions around the world, and thought something that all of us could relate to/have experienced are natural disasters. The two most common disasters that we all have some experience with are earthquakes and tsunamis. Since the two go hand in hand (tsunamis are caused by earthquakes that occur on the ocean floor) we thought that examining earthquake detection

would be an interesting topic. To do this, we found a dataset that contains information on earthquake occurrences and a dataset that contains information about the growth of GSN stations. This is an important topic because if it is found that an increase of GSN stations has a significant impact on the frequency of earthquakes detected, it could be beneficial to continue building GSN stations. If there is a sooner/more frequent detection of earthquakes this can help warn people earlier and give them time to prepare/go to a safe location.

# Data Set Description

## 1. All Natural Disasters 1900-2021

- The data set encompasses 16127 natural disasters that occurred worldwide in the years 1900 to 2021. All floods, storms, earthquakes, landslides, etc.
  Each natural disaster has its own entry with the corresponding characteristics.
- The dataset was last updated on 2021-10-10.
- URL: https://www.kaggle.com/brsdincer/all-natural-disasters-19002021-eosdis
- Source of Kaggle data: From the Earth Observing System Data and Information System website (EOSDIS)( https://earthdata.nasa.gov/eosdis)
- NASA's Earth Observing System Data and Information System (EOSDIS) provides end-to-end capabilities for managing NASA Earth science data from various sources--satellites, aircraft, field measurements, and various other programs.
- **Disaster Year**
  - The year in which the disaster occurred.
  - Data type: Numeric
- **Disaster Type**:
  - The category that the natural disaster falls under (we will be examining floods, droughts, storms, earthquakes, and landslides).
  - Data type: Factor.
- **ISO**:
  - Country code of the country in which the disaster occurred (e.g. ARG for Argentina).
  - Data type: Factor.
- **Region:**
  - The region where the disaster occurred: Southern Asia, South-Eastern Asia, etc.
  - Data type: Factor.
- **Continent**
  - Continent where the disaster occurred.
  - Data type: Factor.
- **Latitude**
  - The latitude of where the disaster occurred
  - Data type: Numeric
- **Longitude**
  - The longitude of where the disaster occurred
  - Data type: Numeric

## 2. Global Seismographic Network Dataset

- The Global Seismographic Network (GSN) is a 150 station, globally distributed, state-of-the-art digital seismic network that provides free, real-time, open-access data. GSN instrumentation measures and records with high fidelity all seismic vibrations possible from high-frequency, strong ground motions near an earthquake to the slowest global Earth oscillations excited by great earthquakes.
- We wrote a python script to scrape the GSN network website to create the GSN dataset.
- URL:https://earthquake.usgs.gov/monitoring/operations/network.php?virtual_net
- **Network Code**
  - Two-letter FDSN (International **F**ederation of **D**igital **S**eismographic **N**etworks) Network code that is the official designation of a particular GNS station's parent network.
- **Station Code**
  - 5 letter FDSN code of letters/numbers that are the official designation of a particular GNS station. Maximum 5 characters.
- **Name**
  - Name of the geographic location of the station.
- **Latitude**
  - The Latitude coordinates of the precise location of the station.
- **Longitude**
  - The Longitude coordinates the precise location of the station.
- **Operational From**
  - The year in which the station started operations.

**A block of code showing how to load the data into R**

```
disasters_1970 = read.csv("DISASTERS/1970-2021_DISASTERS.xlsx - emdat
data.csv")
disasters_1900 = read.csv("DISASTERS/1900_2021_DISASTERS.xlsx - emdat
data.csv")

disasters_1900  <- disasters_1900 %>%
                    select(Year, Disaster.Type, ISO, Region, Continent,
Latitude, Longitude)

disasters_1970  <- disasters_1970 %>%
                    select(Year, Disaster.Type, ISO, Region, Continent,
Latitude, Longitude)

gsn_data <- read_csv("gsn_data.csv")
```

## Statistical Questions

**Has there been a significant increase in the frequency of earthquake detections after the establishment of the Global Seismographic Network (GSN)?**

**Discussion of the question**

As described above the GSN is a global network of seismic stations that help in detecting earthquakes. Since its establishment in 1986, it has kept increasing the number of stations over the years up to the present day. Our hypothesis is that the increase in GSN stations worldwide is correlated to the increase in earthquake detections as well. That is the GSN has had a significant impact on accurately detecting earthquakes. We test this hypothesis to answer our statistical question if there has been a significant increase in earthquake detections after the establishment of the GSN.

## Exploratory Data Analysis

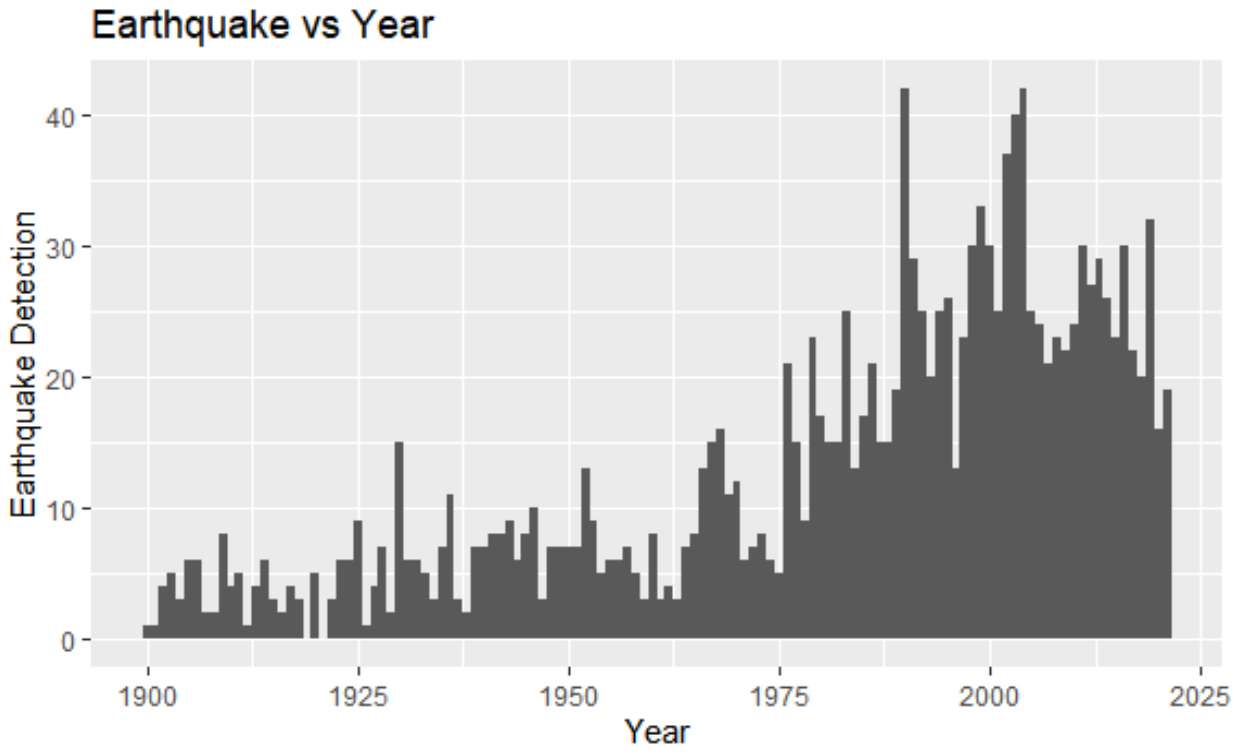



Fig 1: The above is the histogram of the number of earthquakes detected by year from 1900-2021. It seems like the number increased over the years, but the detection and reporting of earthquakes have improved dramatically over time. So to answer the question of whether the occurrence of earthquakes really increased over time, we decided to incorporate the GSN data.

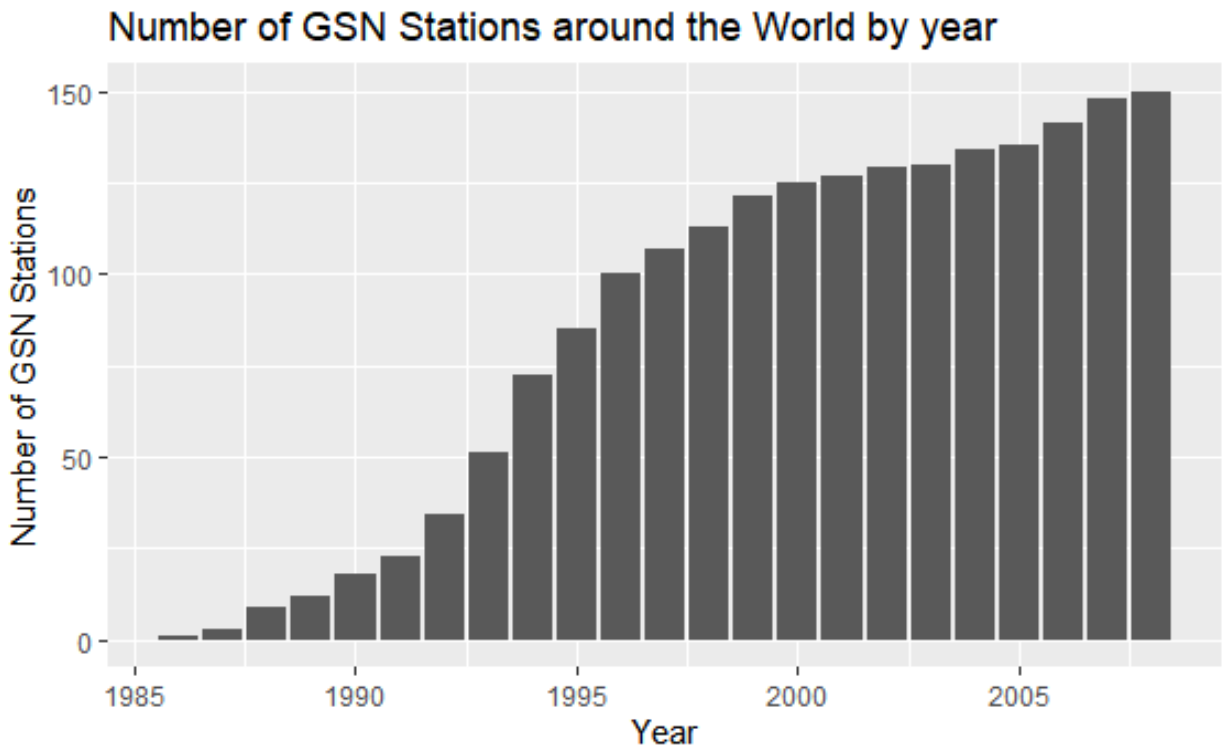



Fig 2: The above is the graph of the number of GSN stations around the world by year. As you can see, the number of stations increased over time and since 2008 is being operated by 150 stations.

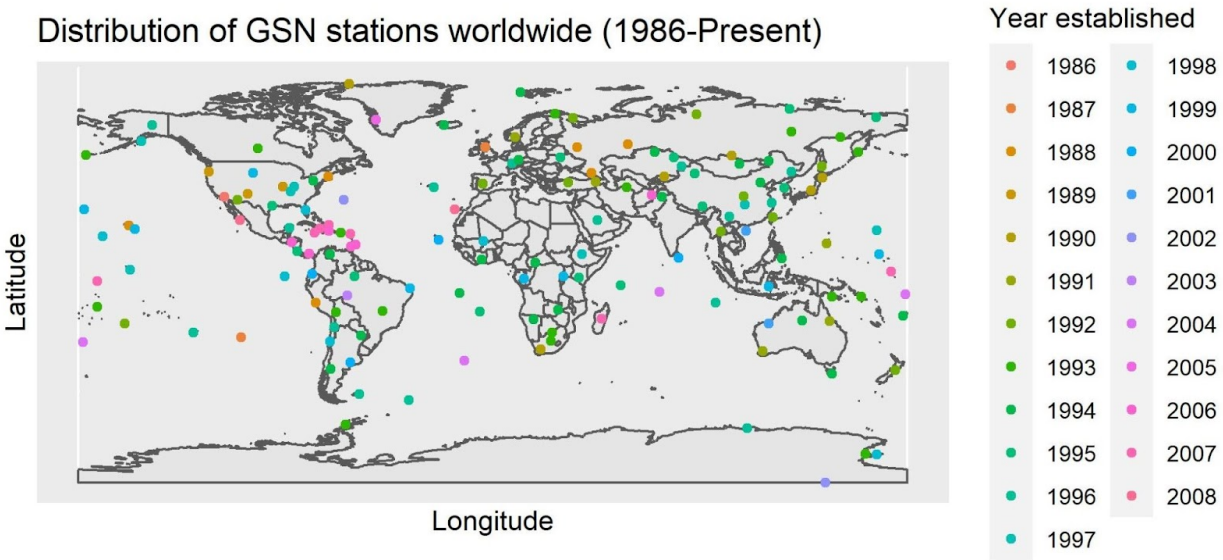



Fig 3:The above map shows the distribution of GSN stations worldwide, color-coded based on the year each station commenced operations.

Next, we visualize the relationship between GSN stations and the frequency of earthquakes detected.

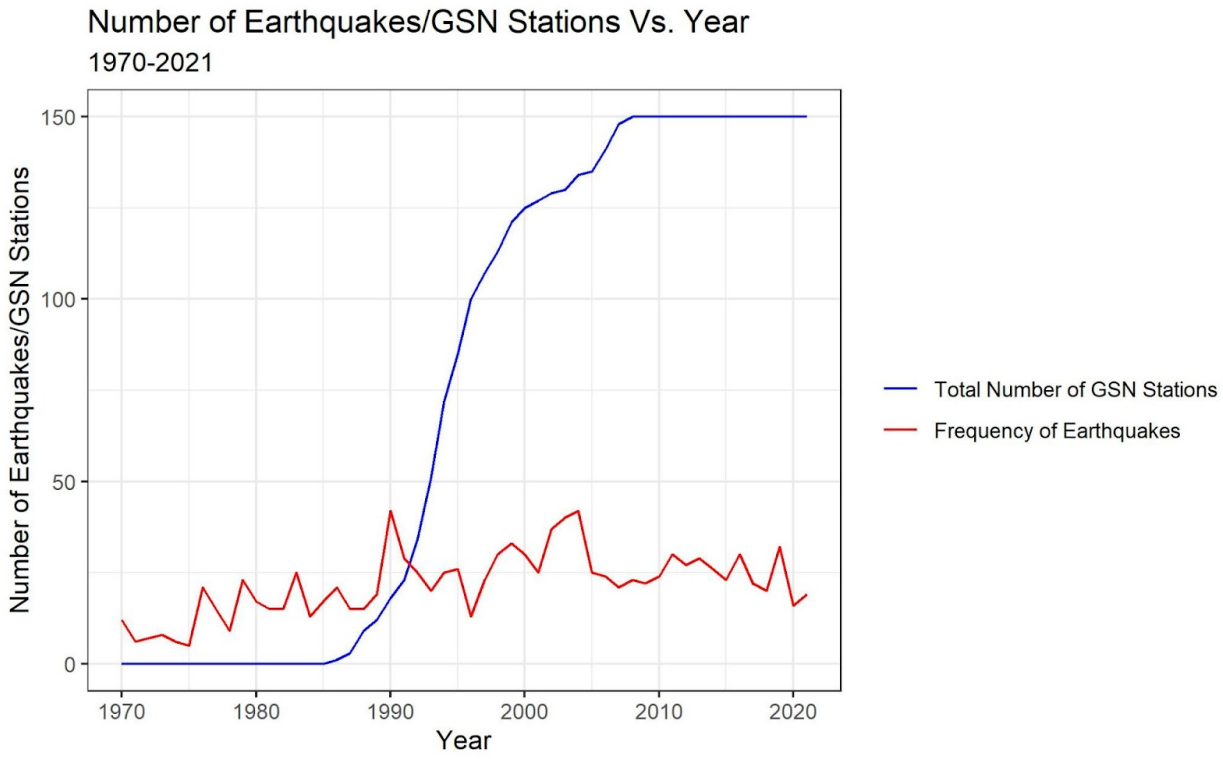



Fig 4: The above map shows the cumulative number of GSN stations (blue) by year and the number of earthquake detections (red) by year. We tried to see if there is a correlation between the detection of earthquakes with the number of observational stations around the world.

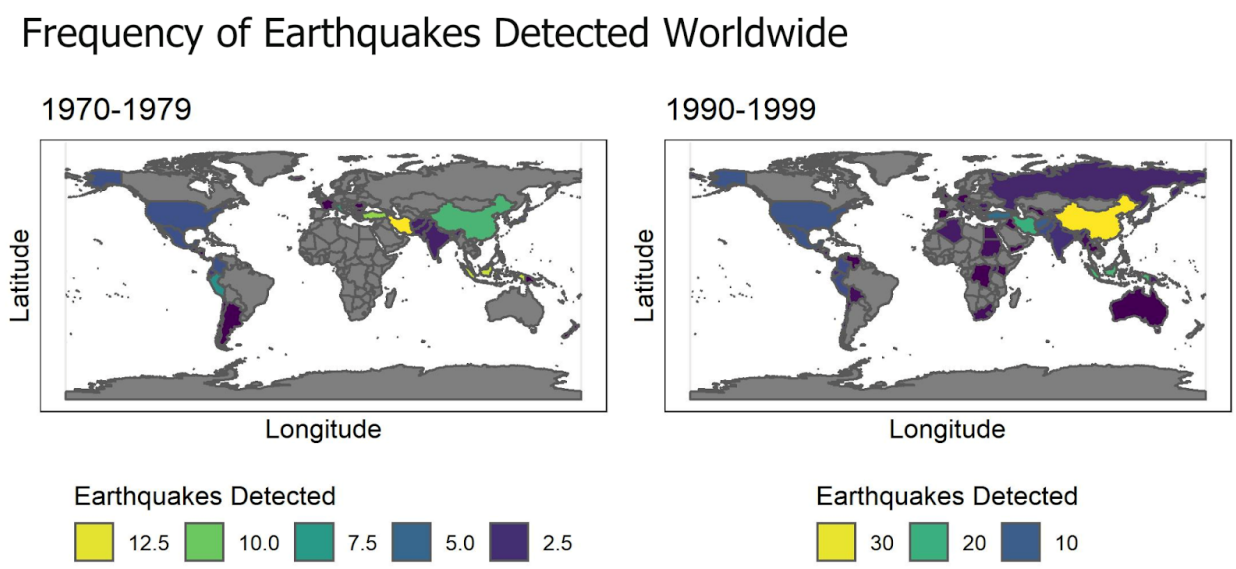



Fig 5:These two maps are a comparison of the number of earthquakes detected and their geographical spread in the ’70s (left) and the ’90s (right).

Comparing the two maps you can see that the frequency of detected earthquakes and the geographical spread of those detections have increased going from the ’70s to the ‘90s. One reason for the increase in detections and the geographical range could be the improvement in detection technologies and methods. We know that the Global Seismographic Network was first established in 1986, and at present, there are 152 stations active stations worldwide. Our hypothesis is that the establishment of the GSN network led to an increase in the detection of earthquakes.

To explore this hypothesis, we plotted a side-by-side comparison of the geographical spread of earthquakes detected in the 1990s and the 2000s and we overlaid each map with the locations of GSN stations active during those time periods. We chose to compare the 1990s and the 2000s because we observed a significant increase in the number of GSN stations active going from the 1990s to the 2000s. So, that time period would be ideal to check whether GSN stations had an impact on the detection rates of earthquakes.

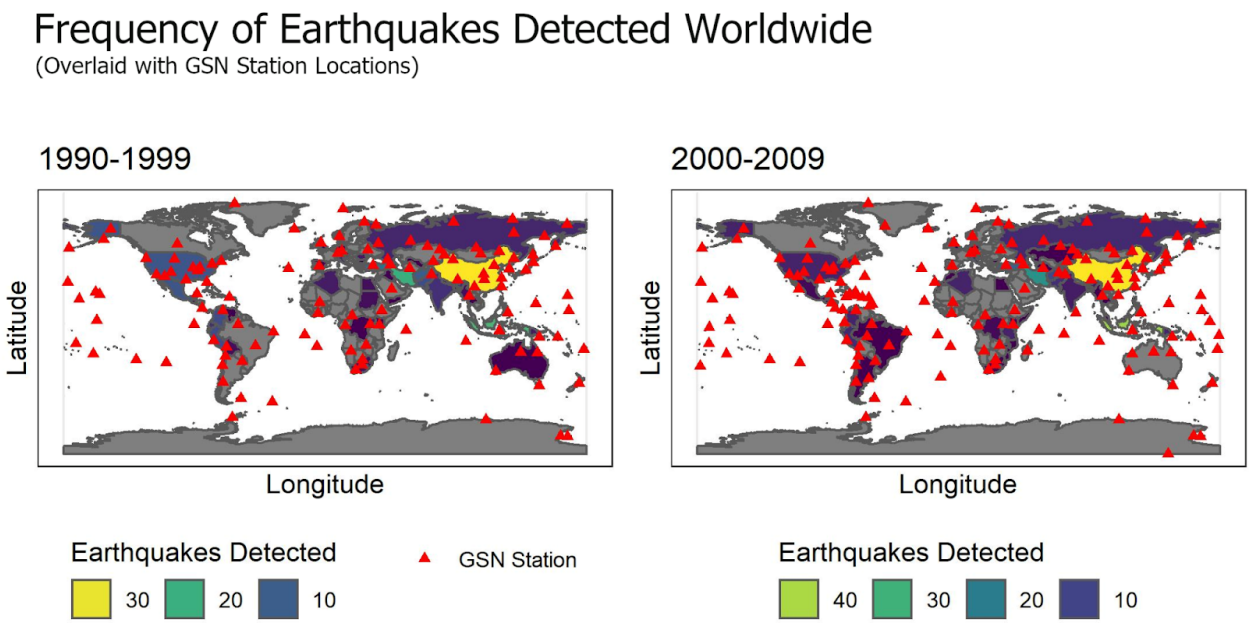



Fig 6: These two maps are a comparison of the number of earthquakes detected and their geographical spread in the 1999s (left) and the 2000s (right). And overlaid over each map is the location of GSN stations active during that time (red triangles).

Comparing the above two maps we can see that both the number of earthquakes detected and the amount of GSN stations active have increased in the span of a decade. This may suggest that there may be a correlation between the number of GSN stations active and the number of detections of earthquakes made. So we explore this further in the following sections.

# Analysis

### 1. Two-Sample t-tests

First, we tried to find whether there is a significant increase in earthquake detections after the establishment of the GSN in 1986. So to test this we conducted a one-sided independent two-sample t-test on the number of earthquakes detected before and after 1986. Our null hypothesis for this test is that the mean number of earthquakes detected before 1986 is equal to the mean number of earthquakes detected after 1986. The alternative is that the mean of earthquakes detected after 1986 is greater than that of before 1986.

```
        Welch Two Sample t-test

data:  after and before
t = 13.9, df = 50.301, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 16.04646      Inf
sample estimates:
mean of x mean of y
25.638889  7.392857
```

The test rejected the null hypothesis with a significant p-value < 2.2e-16 and concluded that there is a true difference in the mean frequency of earthquakes before 1986 and after 1986 and that the true difference is greater than 0. In other words, there is an increase.

To further verify the impact of the GSN on the frequency of detections of earthquakes. We decided to conduct another one-sided two-sample independent t-test on the number of earthquakes detected before 2008 and the number of earthquakes detected after 2008. We picked 2008 as a point in time because 2008 is when the last GSN station was added to the network and ever since the GSN has had 150 active stations worldwide. So analyzing if there's a significant increase in earthquake detections after 2008 when the network stopped increasing would suggest a true impact of the GSN on earthquake detection.

To test if there was an increase in the number of earthquake detections while the number of GSN stations is held constant, we took the earthquake data from 2008 to 2020 because, since 2008, the number of GSN stations around the world did not change. To conduct the test, we split our data into half, 2008 to 2014 and 2014 to 2020, and tested if there is an increase in the number of earthquake detections using the two-sample t-test with the testing hypothesis that there is no significant increase.

```
        Welch Two Sample t-test

data:  x and y
t = -0.71165, df = 9.2144, p-value = 0.7528
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -6.118372       Inf
sample estimates:
mean of x mean of y
 24.14286  25.85714
```

The test gave a very large p-value = 0.7528 and we failed to reject our null hypothesis. We concluded that there is no statistically significant evidence that there is an increase in earthquake detection while the number of GSN stations is held constant. This has the implication that increasing the number of GSN stations would lead to better earthquake detections.

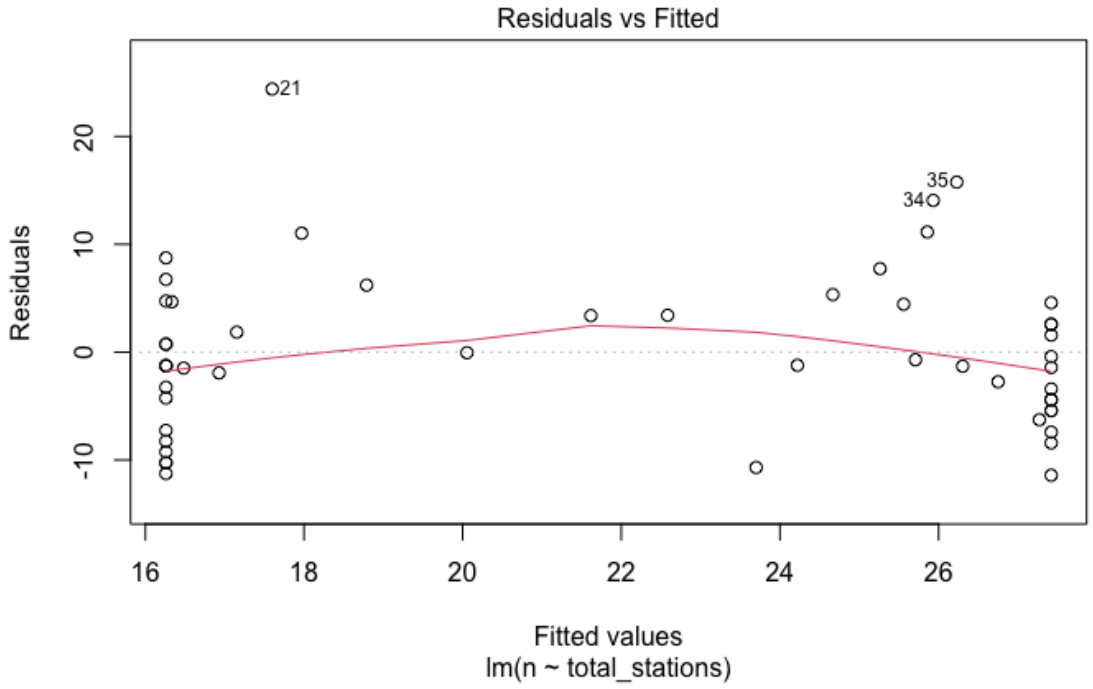
## 2. Fitting a linear regression model

Since the detection of earthquakes increased while the number of GSN stations was increasing, as mentioned above, we suspected that there would be some correlation between them. Running a correlation on the number of centers vs. number of earthquake detection, the correlation coefficient was 0.5596645. This represents that there is a moderately positive relationship between the number of earthquakes in a year and the number of stations in a year.

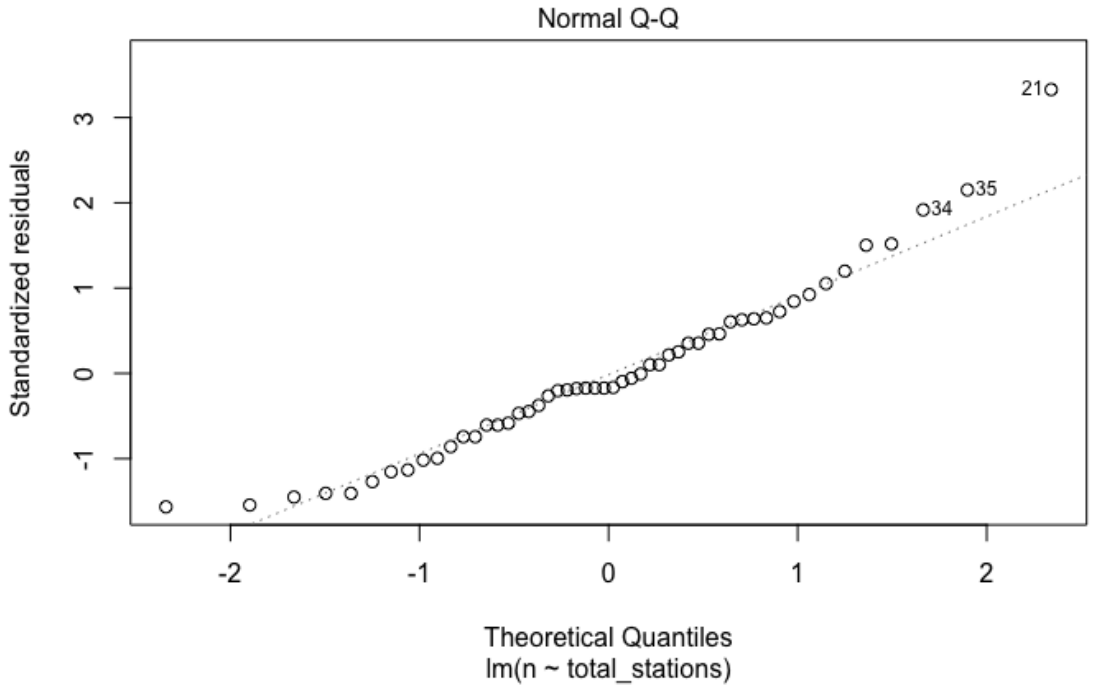
Knowing there is a moderate correlation between the number of earthquakes detected and the total number of stations, we created a linear regression model between them.

```
Call:
lm(formula = n ~ total_stations, data = by_year_new)

Residuals:
    Min      1Q  Median      3Q     Max
-11.420  -4.670  -1.240   4.475  24.401

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    16.25952    1.56457  10.392 4.33e-14 ***
total_stations  0.07440    0.01558   4.775 1.61e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.459 on 50 degrees of freedom
Multiple R-squared:  0.3132,    Adjusted R-squared:  0.2995
F-statistic:  22.8 on 1 and 50 DF,  p-value: 1.607e-05
```

The model above appears to be significant since the p-value is less than 0.05. However, in order to dive deep into the validity of the model, we can take a look at different plots like the residuals in order to determine if this is a valid model.




The residuals plot for the model above has a random scatter about the horizontal axis. However, there are some outliers for the residuals which may indicate that this model can be further improved.




The Q-Q plot is fairly linear with the exception of a few endpoints. The endpoints that skew the Q-Q plot seem to be the same outliers that exist in the residual plot. This confirms the

same thing as mentioned above that the model, although significant, can be improved. As a result, we decided to test different models to see if we could come up with a better model.

The correlation between Year and the detection of earthquakes was also moderately strong with a correlation coefficient of 0.5647147.

```
Call:
lm(formula = n ~ Year, data = by_year_new)

Residuals:
     Min       1Q   Median       3Q      Max
-14.0010  -4.7748  -0.8787   3.7987  21.9610

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -640.77418  136.95570  -4.679 2.23e-05 ***
Year           0.33207    0.06863   4.838 1.30e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.428 on 50 degrees of freedom
Multiple R-squared:  0.3189,    Adjusted R-squared:  0.3053
F-statistic: 23.41 on 1 and 50 DF,  p-value: 1.295e-05
```

The linear regression model above uses the year to predict the number of earthquakes detected. The p-value for this model is also less than 0.05 which means that in this model, the year is a valid predictive variable.

We tried to construct a model that takes into account both the number of stations and year to predict the detection of earthquakes. However, using the simple linear regression model, we could not create a reliable model.

```
Call:
lm(formula = n ~ total_stations + Year, data = by_year_new)

Residuals:
     Min       1Q   Median       3Q      Max
-13.1048  -5.0000  -0.6295   3.6318  23.0951

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -370.2104   401.0058  -0.923    0.360
total_stations    0.0329     0.0458   0.718    0.476
Year              0.1952     0.2026   0.964    0.340

Residual standard error: 7.464 on 49 degrees of freedom
Multiple R-squared:  0.326,     Adjusted R-squared:  0.2985
F-statistic: 11.85 on 2 and 49 DF,  p-value: 6.341e-05
```

We figured out that there is a really strong correlation between the number of stations and year (correlation coefficient was 0.9402685). Therefore, we included an interaction term in our model.

```
Call:
lm(formula = n ~ total_stations + Year + total_stations * Year,
    data = by_year_new)

Residuals:
    Min      1Q  Median      3Q     Max
-15.538  -4.022  -1.821   3.943  18.417

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1.478e+03  4.180e+02  -3.536 0.000911 ***
total_stations       1.322e+01  2.911e+00   4.543 3.74e-05 ***
Year                 7.541e-01  2.111e-01   3.573 0.000815 ***
total_stations:Year -6.619e-03  1.460e-03  -4.532 3.88e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.311 on 48 degrees of freedom
Multiple R-squared:  0.528,     Adjusted R-squared:  0.4985
F-statistic:  17.9 on 3 and 48 DF,  p-value: 6.203e-08
```

With this model, all the predictors are significant which suggests that this model is possibly valid to predict the number of earthquake detections in a year.

However, this model is not a perfect reflection of the real world. Our model only predicts the number of detections based on the GSN stations active at that time. It does not account for the changes in detection technology over time. For example, the improvement in the detection technology would allow stations to pick up earthquakes of smaller magnitude.

**Further Analysis**

So next as part of our analysis, we check if there is further technological development besides the number of GSN stations that impacts earthquake detection. For example, the development of more advanced detection technology like digital seismometers will allow us to detect earthquakes of smaller magnitude, even ones that humans cannot feel. Therefore, we checked if there is an increasing number of lower magnitude earthquakes detected while the number of GSN stations is held constant.

In this analysis, we classify earthquakes with “low magnitudes” as earthquakes that have a magnitude <= 5. We make this classification based on the following table:

Earthquake Magnitude Scale

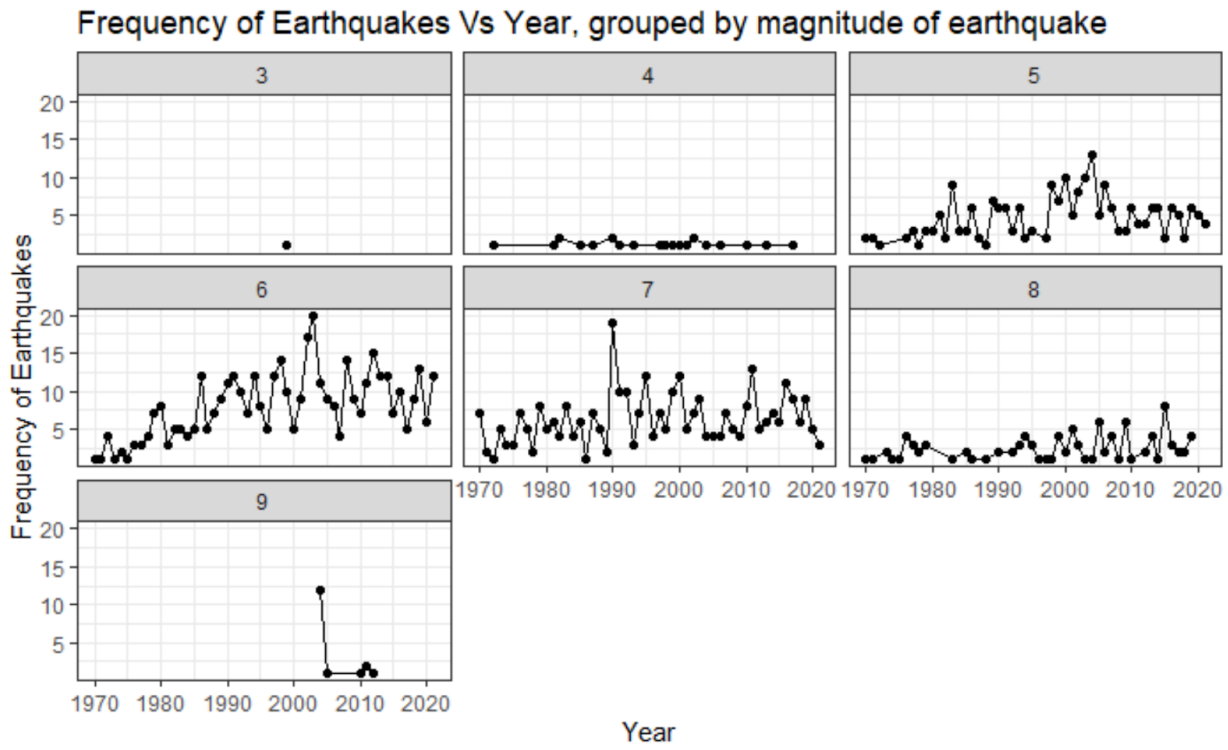
| Magnitude | Earthquake Effects | Estimated Number Each Year |
|---|---|---|
| 2.5 or less | Usually not felt, but can be recorded by seismograph. | Millions |
| 2.5 to 5.4 | Often felt, but only causes minor damage. | 500,000 |
| 5.5 to 6.0 | Slight damage to buildings and other structures. | 350 |
| 6.1 to 6.9 | May cause a lot of damage in very populated areas. | 100 |
| 7.0 to 7.9 | Major earthquake. Serious damage. | 10-15 |
| 8.0 or greater | Great earthquake. Can totally destroy communities near the epicenter. | One every year or two |

Source: https://www.mtu.edu/geo/community/seismology/learn/earthquake-measure/magnitude/

We then did a one-sided two-sample t-test to check whether there was an increase of low magnitude earthquakes after 2008 by comparing the sample means of low magnitude earthquakes before and after 2008. We picked 2008 because that is the year when the last GSN station was added to the current network. Our null hypothesis is that the population mean of low magnitude earthquakes before 2008 is equal to the population mean after 2008. The alternative is that the population mean after 2008 is greater than that of before 2008. The results of the test are as follows:

```
        Welch Two Sample t-test

data:  x and y
t = 0.30911, df = 45.093, p-value = 0.3793
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.8692011        Inf
sample estimates:
mean of x mean of y
 3.823529  3.627451
```

The test gave a very large p-value = 0.3793 and we failed to reject our null hypothesis. We concluded that there is no statistically significant evidence that there is an increase in detections of lower magnitude earthquakes after 2008 when the GSN stopped increasing. However, this result does not imply that the improvement of detection technology or the GSN does not have any effect on increasing detection rates of low magnitude earthquakes. Rather the case is that it is inconclusive. We believe that there could be a reporting bias in our dataset that includes only earthquakes that had enough impact on human lives to be reported. Hence, this dataset may not have recorded all the insignificant earthquakes, most of which are low magnitude earthquakes. This is evident from the graph below.




## Conclusion

Through the exploratory data analysis of our two datasets, we were able to see the general increasing trend in the number of earthquakes as well as the number of GSN stations over time. From there, we tested to see if there has been a significant increase in the frequency of earthquake detections after the establishment of the Global Seismographic Network (GSN). After testing for increase against multiple valuable dates (1986, 2014), we found that after 1986, there had been a significant increase in earthquake detection, which could be explained by the increase in the number of GSN stations whereas we cannot confirm that there is a significant increase in earthquake detection in years after 2008.

In order to further examine the relationship between the addition of GSN stations and earthquake detection, we found a positive, moderately strong correlation as well as created a linear model to predict the number of earthquakes based on the year, the total number of stations existing that year, and the interaction between those two terms.

(number of earthquake) = - 0.001478 + 0.1322*(number of stations) + 0.7541*(year)
- 0.006619*(number of stations)*(year)

This model definitely has room for improvement for more accuracy which is something that can be explored in the future after incorporating more data and factors like other improvements in technology. We also attempted to go further into the analysis of the impact of technology by testing a more specific aspect of earthquakes, testing if the addition of stations allowed for detecting smaller magnitudes (<= 5.0). We concluded that there was no significant increase in detections of smaller earthquakes after conducting a one-sided t-test.

Earthquake detection technology has improved immensely over the past few decades and continues to do so even now. Although it was difficult to find more data on the improvements of technology that could be factored directly into our data and analysis, further improvements in technology can be explored in the future and help expand our analysis to detect and predict a wider range of things such as the regions and future years that will experience earthquakes.

## Shortcomings and Potential Future Research

Throughout our project, we used a series of histograms, in order to determine the frequency of natural disasters, more specifically earthquakes since 1970 after GSN improvements and whether or not they have seen an increase over time. Our original plan was to see the general trend towards the frequency of earthquakes from 1900-2021 as the dataset given was titled. However, while scanning and plotting the data, one thing we noticed was that we didn't have enough data or recorded earthquakes from the years 1900-1970 and as a result, we decided to eliminate any data within those years due to the small sample size.

Additionally, one of our shortcomings stemmed from some missing data in our dataset. Our GSN dataset only included stations that are currently operating and not any discontinued station which could lead to possible errors and shortcomings in our analysis. Potential future research could be conducted to better reflect the status of earthquake detection facilities rather than limiting it to GSN related facilities.

There were multiple stations in our GSN dataset that were included in years before our original dataset. Since the GSN's help the frequency of detected earthquakes, the fact that we had stations added in years not included in our original dataset may not be as accurate of a representation of helping answer our statistical question of whether or not there has been a significant increase in the frequency of earthquake detections.

Some of our testing results failed to reject the null hypothesis such that we could not really draw a solid conclusion from it.

Another shortcoming, mentioned previously, was the need for more data such as more detections of earthquakes with smaller magnitudes. With future research, we can pull more data from multiple reliable sources and possibly even incorporate real-time data to have more information that would lead to more significant conclusions.