DSC 465 – DATA VISUALIZATION

JOURNEY TO THE CENTER OF THE EARTH

1. Introduction

Natural disasters, such as earthquakes, have significant implications for human life, infrastructure, and the environment. Understanding the patterns and characteristics of earthquakes is crucial for developing effective mitigation strategies, designing resilient structures, and ensuring the safety of communities at risk. The Earthquake Dataset, available on Kaggle, provides a comprehensive collection of earthquake records, offering a wealth of information for analysis and exploration.

Dataset Overview
The Earthquake Dataset encompasses a vast repository of earthquake data spanning different geographical regions and time periods. It includes various attributes, such as the earthquake's location, magnitude, depth, and date. The dataset also provides supplementary information, such as the type of seismic activity, focal mechanism, and associated tsunami events. With thousands of records, this dataset allows researchers, scientists, and analysts to delve into the patterns and underlying factors of seismic activities.

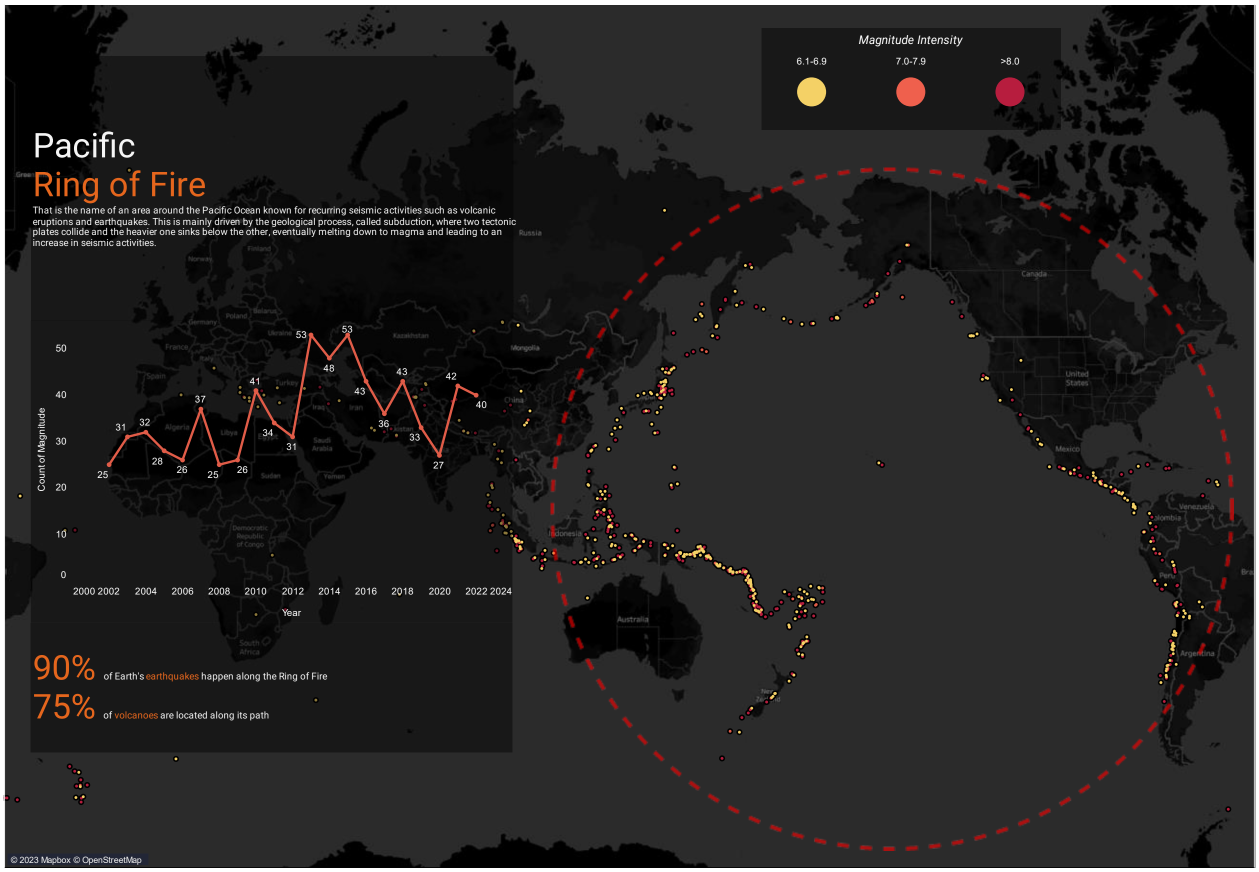
2. Data Preprocessing

- **Checking for Nulls and Replacing Alert Column**: The first step is to check for null values in the dataset and address them. The 'alert' column was found to have null values and was subsequently replaced with a new column called 'Intensity'. This new column categorizes the severity of the earthquake based on its magnitude. This classification follows the scale provided by the United States Geological Survey. This is achieved using nested 'ifelse' statements in R.
- **Adding a Year Column**: To further facilitate analysis, a new column called 'Year' is extracted from the existing 'date and time' column.
- **Addressing the Country Column**: The 'country' column also had null values. These null values were filled with appropriate country names using Python's delimiter and split function. This procedure was applied on the 'location' column to retrieve country names.
- **Dropping the Continent Column**: The 'continent' column was found to have insignificant data and was therefore dropped from the dataset.
- **Reading Economic Damage Data**: A separate dataset containing information about economic damage from natural disasters was read. This data contained information about various types of natural disasters.
- **Extracting Earthquake Data**: Since the focus of the analysis is on earthquakes, a new data frame was created that only contained information about earthquakes from the economic damage data.
- **Merging the Earthquake Data**: The earthquake data from the economic damage dataset was then merged with the original earthquake dataset based on the 'Year' column. The 'alert' column, which was previously replaced, was dropped during this step.
- **Adding Tsunami Data**: After merging, the 'Earthquake Damage' column was renamed for clarity.

- **Adding Overall Economic Damage Data**: Another column from the economic damage dataset was added to the merged table. This column contained information about economic damage from all types of disasters, providing a point of comparison for the economic damage caused specifically by earthquakes.

3. Visualizations

Overall Understanding of Earthquake Occurance



The visualization represents the occurrence of earthquakes in the world. The Pacific Ring of Fire is specifically highlighted by being encircled with a dotted circle, indicating its significance in terms of seismic activity (Earthquake activities).

The earthquakes are categorized based on their intensity using a scale with three levels: low (6.1 - 6.9), moderate (7.0 - 7.9), and severe (above 8.0). Each intensity level is distinguished by a different color, allowing for easy differentiation of earthquake severity.

Additionally, the plot includes the count of earthquake occurrences over a period, ranging from 2001 to 2022. This count provides valuable information about the frequency and distribution of earthquakes both within the Pacific Ring of Fire and in other regions across the globe.

By encircling the Pacific Ring of Fire with a dotted circle, the visualization highlights this region as a focal point of earthquake activity. This is because the Pacific Ring of Fire is known for its elevated seismicity due to the tectonic plate boundaries surrounding the Pacific Ocean.

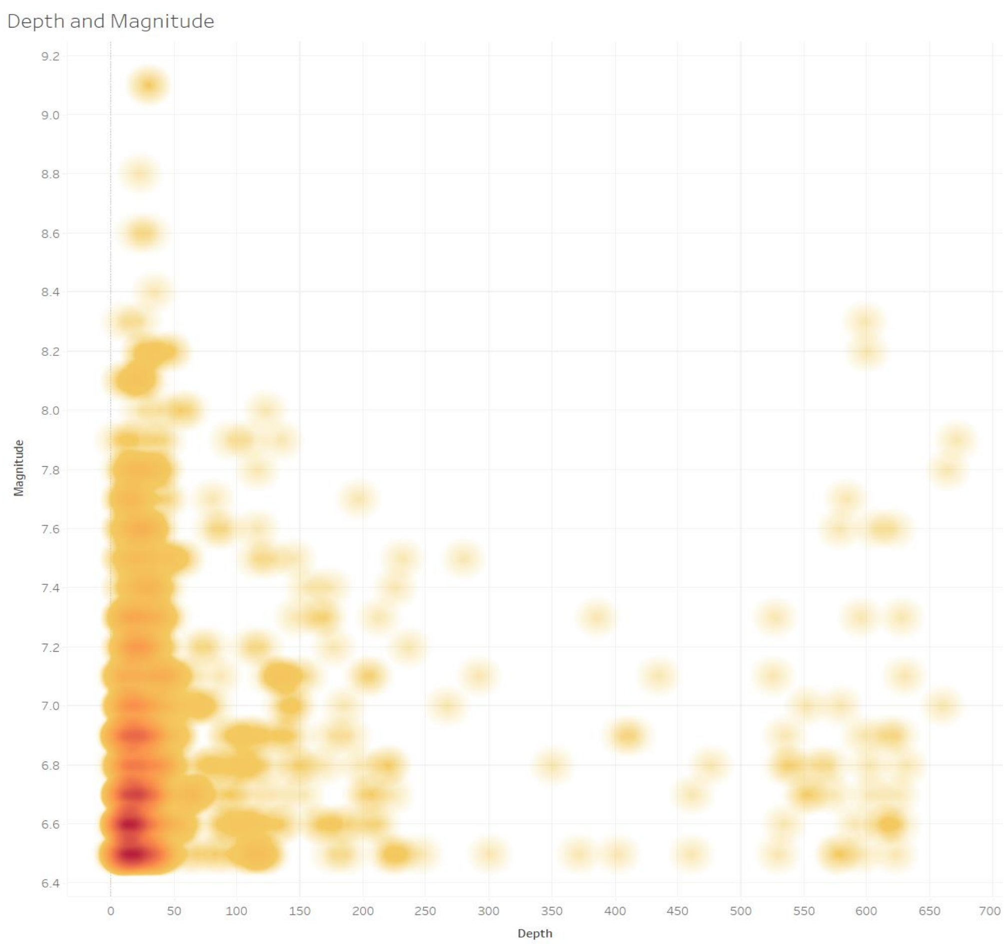
The intensity of earthquake clearly depicts the severity of earthquake. In some parts of the Ring of fire like Indonesia, Japan, Mexico, and Southern part of America has the clusters of earthquakes which shows the frequency of seismic activities. When comparing all these Indonesian regions and Japanese regions has the higher severity earthquakes.

In summary, the visualization provides a comprehensive understanding of earthquake occurrence by depicting both the intensity and count of earthquakes within the Pacific Ring of Fire and the rest of the world. But most of the earthquakes happen in or around the Pacific Ring of fire.

3.1. Depth Analysis with the correlation of magnitude.

Earthquakes damage is dependent on the magnitude and depth of the seismic activity. Two crucial parameters provide valuable insights into the behavior and impact of seismic events.

Shallow earthquakes, which occur near the Earth's surface, tend to have a higher potential for causing significant damage. This is because the seismic energy is released closer to populated areas, resulting in stronger shaking at the surface. Shallow earthquakes can have a wide range of magnitudes, from small tremors to very large events.



Depth vs. Magnitude.

The magnitudes range from 6.5 to 9.1, indicating a variety of seismic activity.

The depths of the earthquakes vary widely, ranging from 4.2 to 675 kilometers.

The highest magnitude in the list is 9.1, which is a significant earthquake and it happened in the shallow part of the Earth's surface.

Most earthquakes in the list have magnitudes between 6.5 and 7.9, suggesting a range of moderate to strong seismic events.

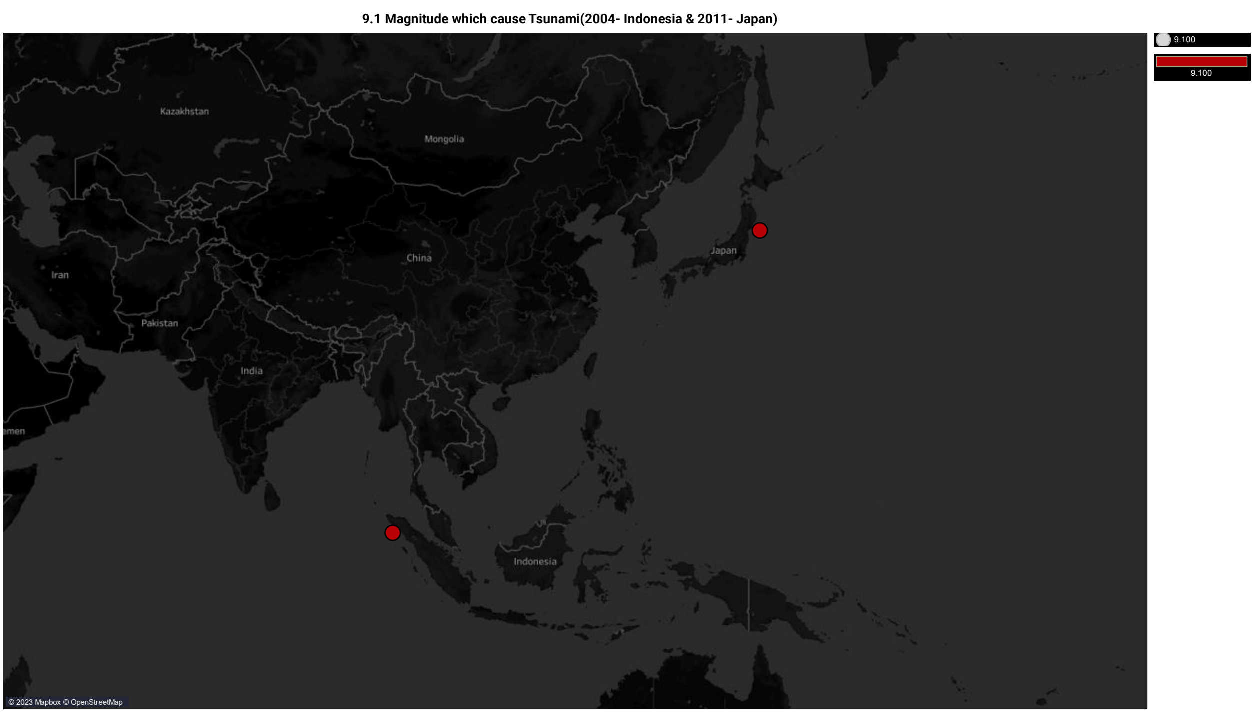
There are instances of earthquakes with relatively high magnitudes occurring at shallow depths, which could result in stronger ground shaking and potentially more significant damage.

The list includes earthquakes with varying depths, indicating seismic activity at different levels within the Earth's crust and mantle.

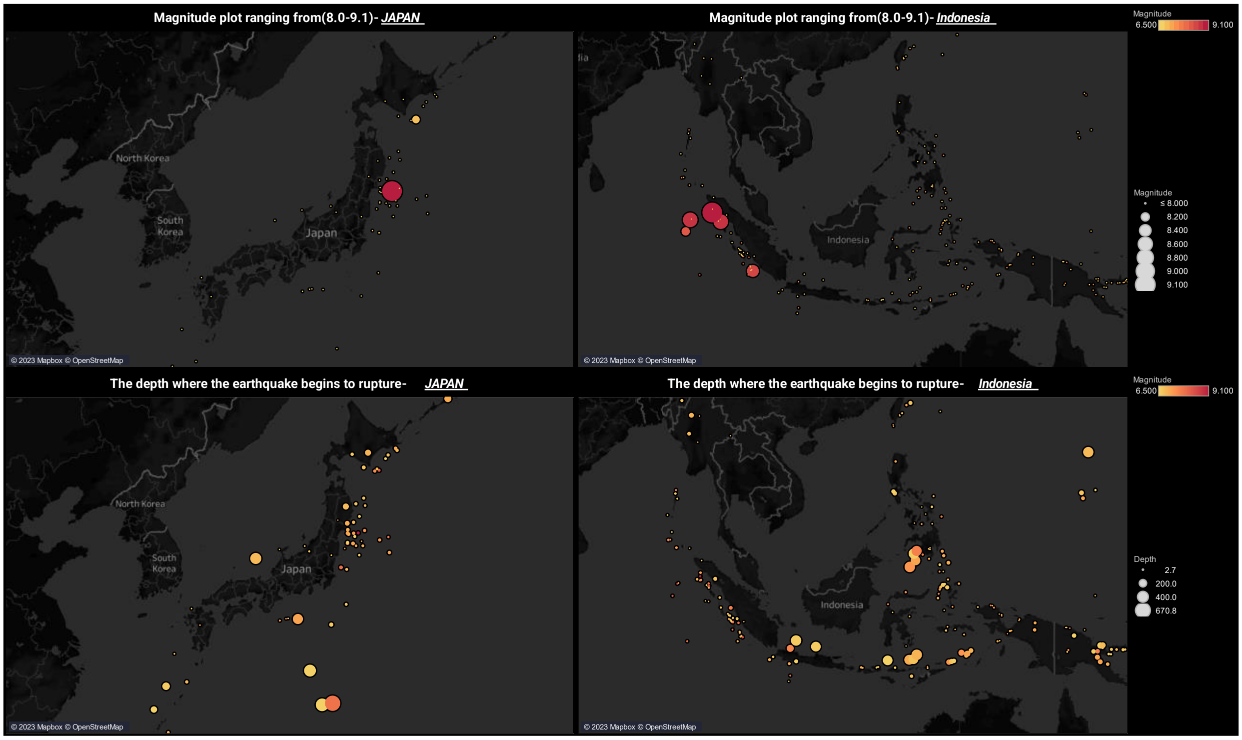
Overall, the list showcases a diverse range of earthquakes in terms of both magnitude and depth, reflecting the dynamic nature of tectonic activity around the world. Further analysis for the 9 Richter magnitude scale and most count of the earthquake regions and depth and magnitude proportions will give a clear picture of depth analysis.

3.2. 9-Richter earthquake and depth analysis

The plot is filtered to magnitude 9 and above. To plot where the deadliest earthquake occurred in world.



The above graph illustrates the magnitude which caused the tsunami in 2004 Indonesia and Indian regions, and 2011 Japan region.

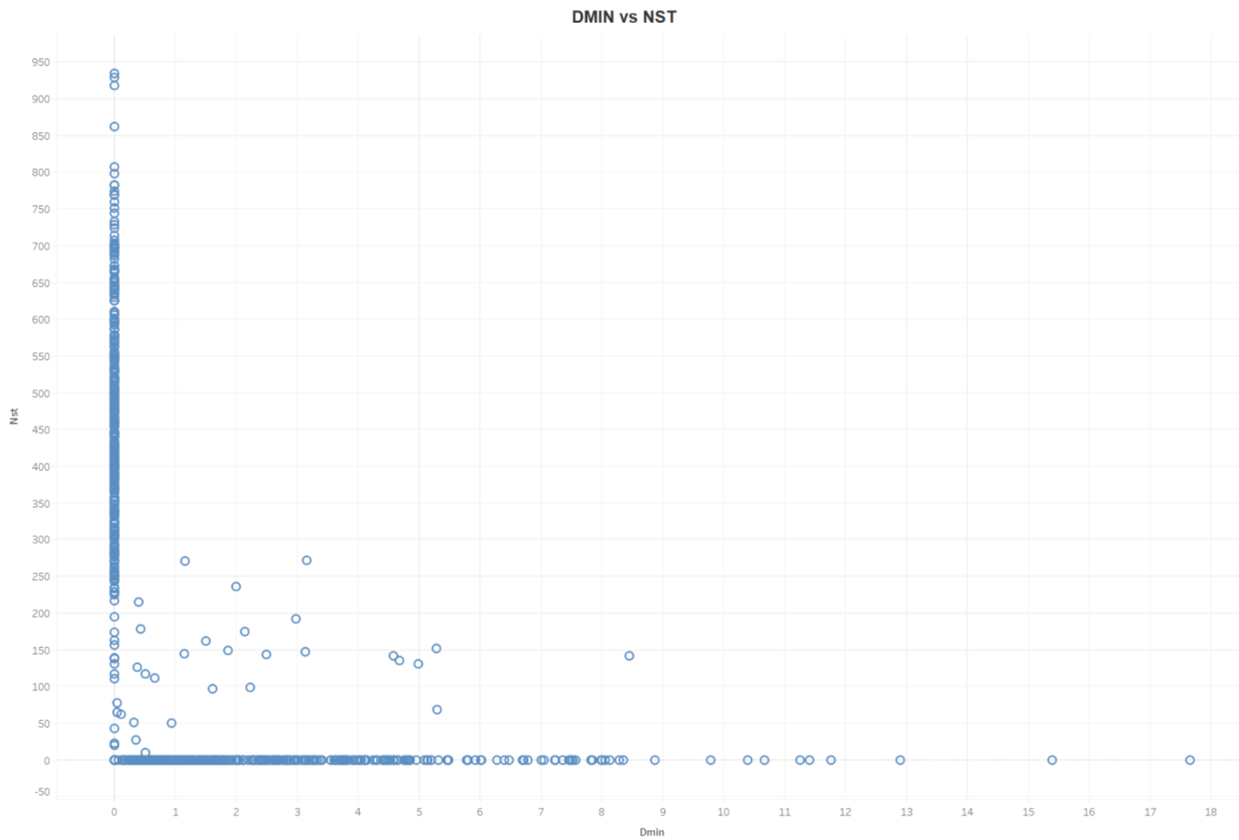


The plot has 4 choropleth graphs which are in dashboard. The two plots above depict the Magnitude intensity with color and size of the magnitude, and below two plots have size which says the depth and color which says magnitude. A choropleth visualization can be produced to show how earthquakes are distributed over various locations. This visualization incorporates earthquake information such as magnitude, depth, latitude, and longitude. The spatial distribution and intensity of earthquakes over the specified latitude and longitude coordinates would be the message delivered by the visualization. A choropleth map often uses shading or coloring to depict data for regions or nations based on the variable of interest. While an earthquake's magnitude is a measurement of the energy produced at the source of the earthquake, the depth of an earthquake is the distance from the Earth's surface to the earthquake's hypocenter. From the graph We can clearly infer that Although shallow earthquakes are frequently more violent at the surface than deep earthquakes of the same magnitude, deep earthquakes are less potent. Even though deep earthquakes can have large magnitudes, their impacts might not be as obvious at the surface, particularly if they happen in isolated, sparsely populated places. In the Upcoming plots we can understand more about how the seismic activity are measured and recorded.

3.3. DMIN and NST

DMIN (Minimum Distance): DMIN is the shortest distance between an earthquake event and the nearest seismic recording station (**NST**). It aids in determining the spatial coverage and dispersion of seismic data.

NST (Number of Stations): The number of seismic stations that recorded a certain earthquake is represented as NST. It indicates the dependability and quality of the earthquake information.



Dmin vs. Nst.

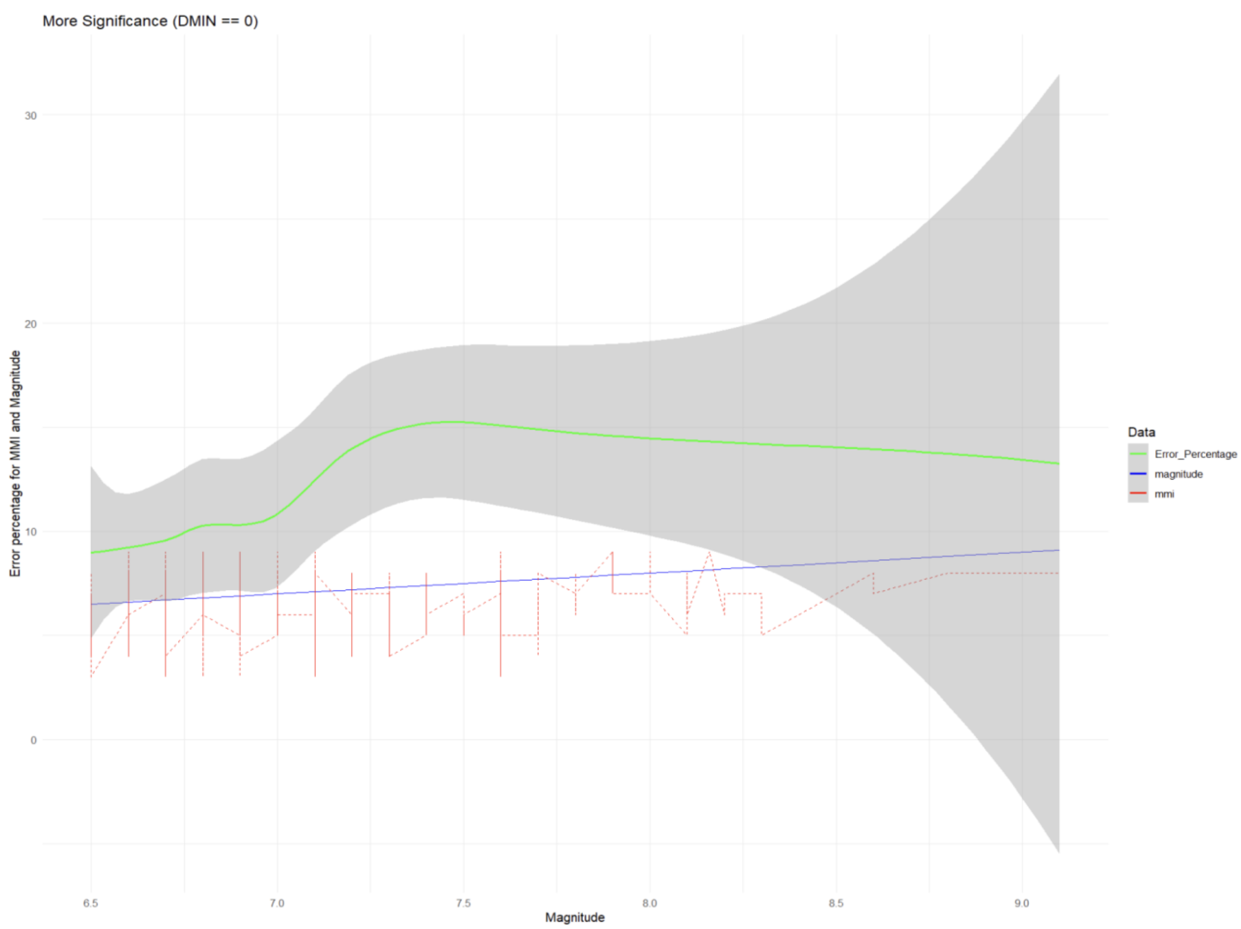
From the above plot DMIN and NST have been distributed across the 0^{th} point of X and Y axis were uniformly distributed. On the other hand, a high DMIN and a low NST value can point to coverage gaps and probable earthquake parameter uncertainties, and low DMIN and a high NST value shows that the recording of earthquake event is significant.

3.4. Analyzing Error percentage.

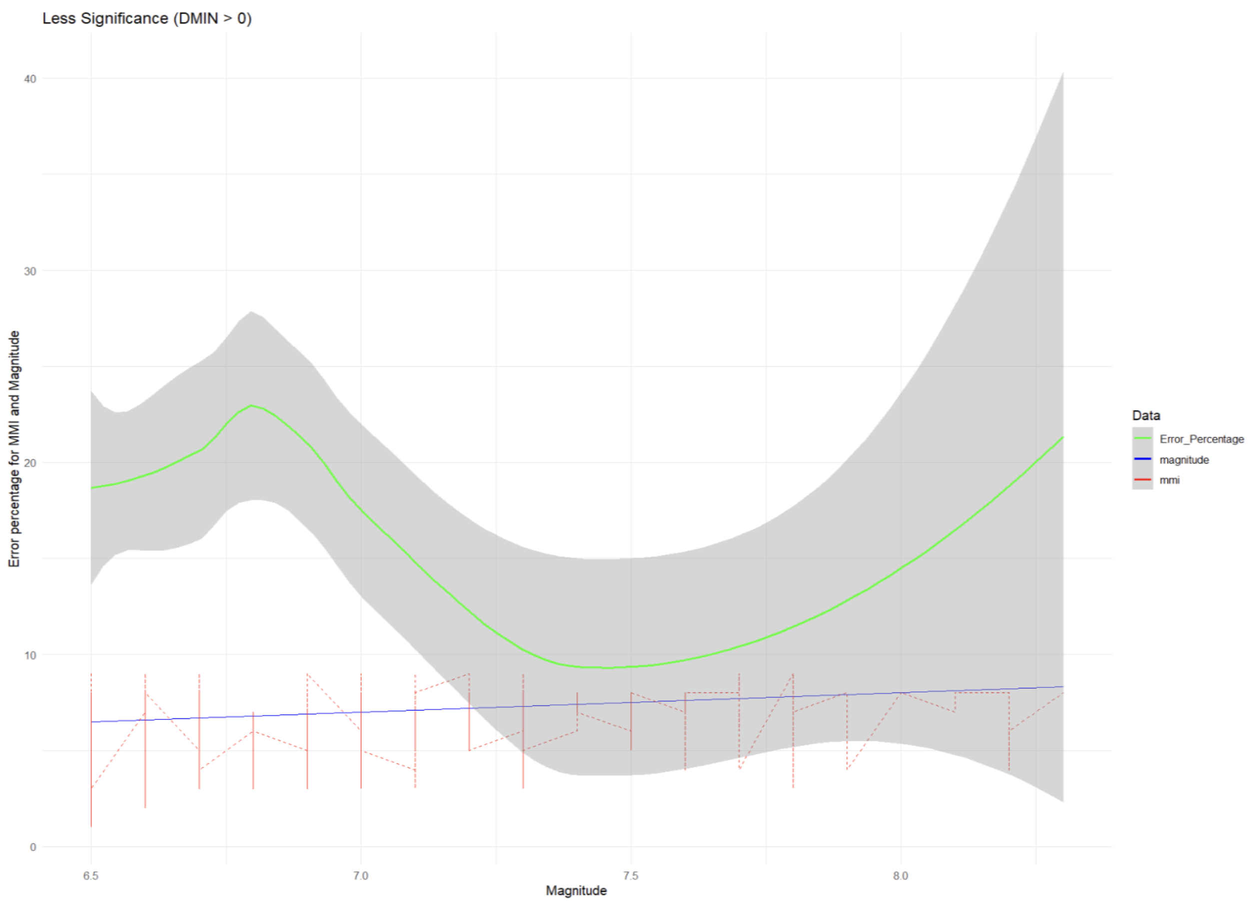
The data has been transformed into two classifications and grouped by the (DMIN == 0) and (DMIN > 0). So, it will create an earthquake reading that has more significance and less significance.

The first reporting values from the NST have been recorded in the column of **MMI (***Modified Mercalli Intensity, which is a measure of the intensity of shaking experienced during an earthquake***),** where it will have the readings of which has been recorded as first information but in the **MAGNITUDE** table values are official's declared values which has been calculated by the average of all **MMI**.

So, the difference between the **MAGNITUDE** and **MMI** will give the error percentage of **MAGNITUDE** and **MMI**. With this we can find how much the value of the actual reading and official declared magnitude is diverged.

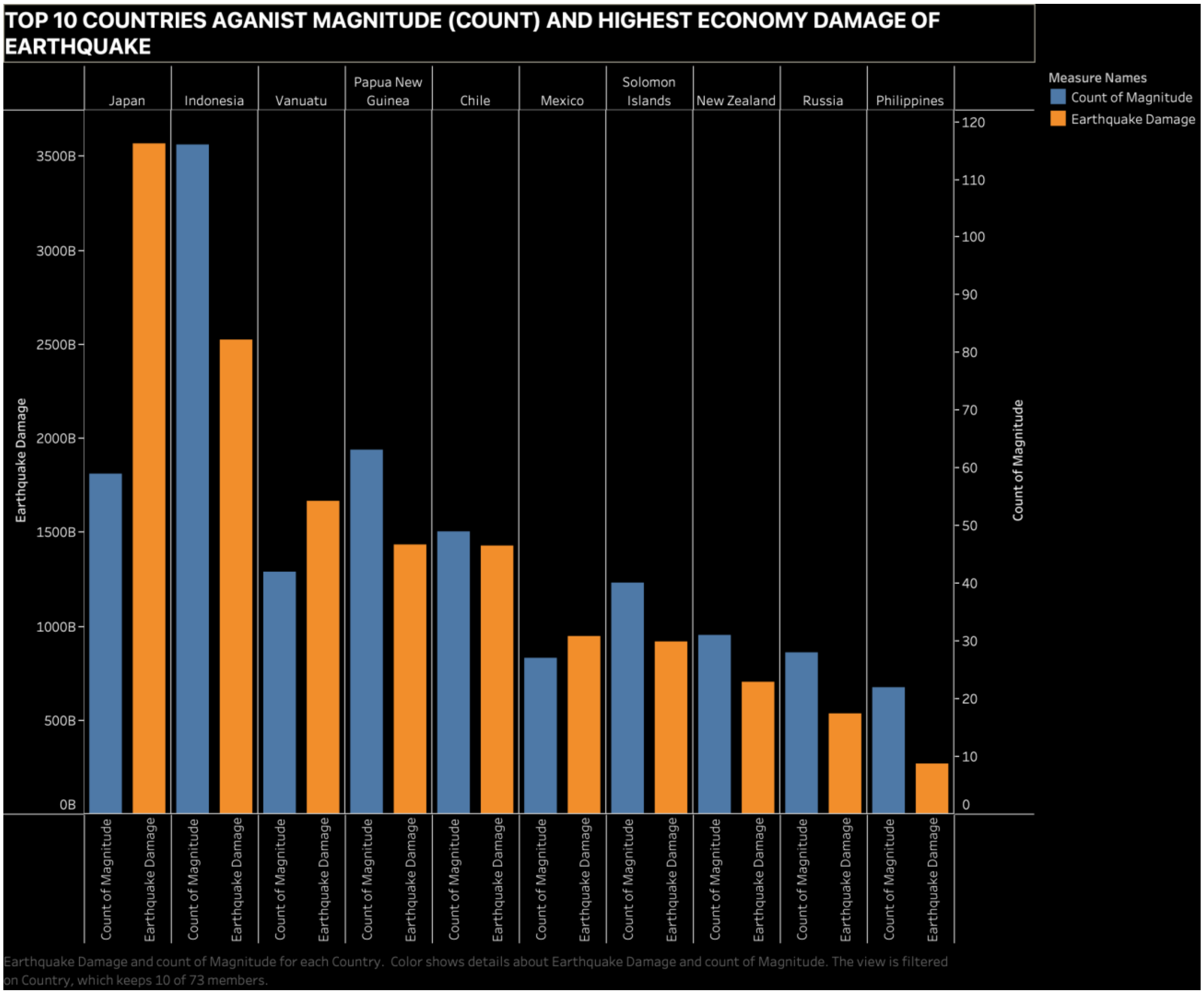


The plot above has two-line plot which is **MMI** and **MAGNITUDE** and geom_smooth line plot to represent the error percentage between the **MMI** and **MAGNTIUDE**. From the graph we can clearly say that the error percentage lies between 8% to 15%. The **MMI** reading count for each magnitude level (which is NST station's). The plot has a low percentage of error in the readings.



From the graph we can clearly say that the error percentage lies between 8% to 23%. The **MMI** reading count for each magnitude level (which is NST station's). The plot has a high percentage of error in the readings. When the DMIN is greater than zero. So, there will be less seismic station near to the earthquake happening, that's why the percentage error is more.

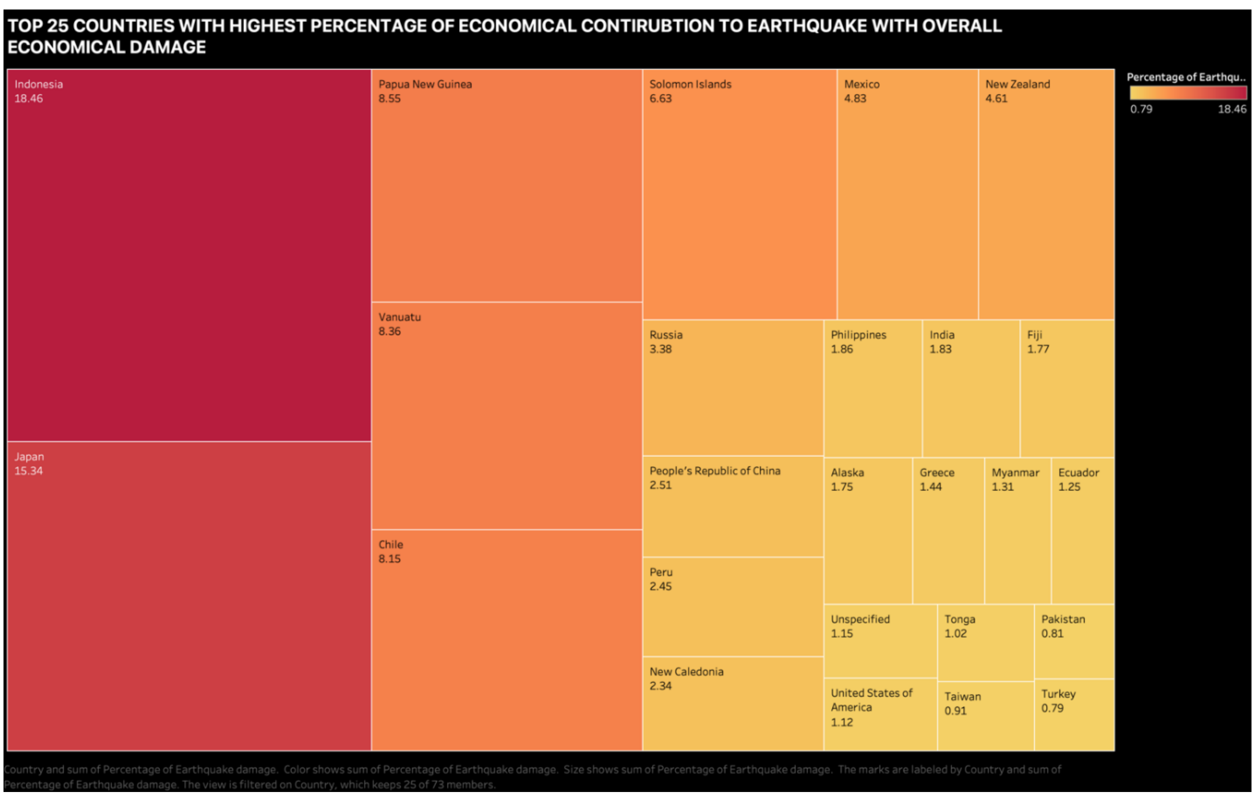
3.5. Economical Damage



The dataset includes information about earthquake magnitude, earthquake damage, and the corresponding countries. To represent this data, a bar chart visualization was chosen. The bar chart is an appropriate choice for displaying earthquake magnitude, country, and earthquake damage. The x-axis represents the countries, while the y-axis represents both the magnitude and earthquake damage.

By using different colors or shades within each bar, we successfully differentiated between earthquake magnitude and earthquake damage. This color variation helps the audience easily distinguish between the two variables and comprehend their relative proportions within each country. The stacked bar chart effectively presents the distribution of earthquakes and their associated damage across different countries, allowing for clear comparisons of magnitude and damage levels.

From the above bar chart we can clearly infer that the magnitude of high impact earthquakes have a clear influence of the Economical Damage contributed because of it. For example, we can take Indonesia the magnitude of earth quake damage is close to 115 and they have impact close to 2500 Billion $ and lets also take Philippines they have the least damage of close to 23 magnitude and the economic damage caused by earthquake is 300 billion $. So we can conclude that higher the magnitude higher the Economical damage as they have a clear correlation and we can also see Japan as a special case because we can see average magnitude is close to 60 which is the third highest but the economic damage is close to 3500 Billion $ this is because the Pacific Ring of Fire, where Japan is located, is a region with tremendous tectonic activity and frequent earthquakes. Japan is a strong technical and economic powerhouse with many manufacturing facilities, power plants, and important infrastructure. The nation also endures a substantial number of seismic occurrences. Earthquake damage to these sites can lead to severe financial losses, including interruptions in supply chains, income loss, and output. Tsunamis, which are frequently brought on by massive underwater earthquakes, are another threat to Japan. These tsunamis have the potential to wreak havoc along the coasts, severely damaging nearby companies, homes, and infrastructure. Considering these elements, Japan suffers greater economic loss from earthquakes than other nations.



The statistics of the nations and the proportion of earthquake damage may be shown using tree maps. The tree map uses stacked rectangles to display hierarchical data, with each rectangle's area representing the percentage of earthquake damage in a particular nation. It is suitable to use a tree map to compare the relative proportions of earthquake damage percentages across various nations. The country, the proportion of earthquake damage, and the overall economic loss from natural catastrophes are all included in the apparent statistics. This will make it clear how much damage is caused by earthquakes alone in various nations and how it relates to overall economic loss. The message of the data is the relative impact of earthquakes in different countries in comparison to the overall economic damage they cause. To accomplish this, we also added a new column to the dataset called **"Percentage of Earthquake Damage"** by using the calculated field option in Tableau. We then implemented a formula to combine this specific damage share of the earthquake damage alone with the overall damage and the Formula used to achieve this field is **[1-([Total economic damage from natural disasters]-[Earthquake Damage])/[Total economic damage from natural disasters]].**

eq.csv Magnitude	eq.csv Country	eq.csv Earthquake Damage	eq.csv Total economic damage ...	Calculation Percentage of Earthqua...
6.70000	Chile	2,067,714,000	51,617,152,000	0.040059
6.80000	Mexico	2,067,714,000	51,617,152,000	0.040059
7.10000	Taiwan	2,067,714,000	51,617,152,000	0.040059
7.50000	Philippines	2,067,714,000	51,617,152,000	0.040059
7.40000	Afghanistan	2,067,714,000	51,617,152,000	0.040059
6.60000	Papua New Guinea	2,067,714,000	51,617,152,000	0.040059
6.50000	Chile	2,067,714,000	51,617,152,000	0.040059
6.70000	Papua New Guinea	2,067,714,000	51,617,152,000	0.040059
6.60000	Vanuatu	2,067,714,000	51,617,152,000	0.040059
7.20000	Vanuatu	2,067,714,000	51,617,152,000	0.040059

Also, from the above tree maps we can clearly infer that those countries with the highest earthquake magnitude, such as Indonesia, Japan, and Papua New Guinea, have the highest percentage share of the economic impact caused by earthquakes from their total economic share, which is 18.46, 15.34, and 8.55 respectively, while the nations with the lowest magnitude, such as Pakistan, Taiwan, and Turkey, have the lowest percentage shares, 0.81, 0.91, and 0.79, respectively.

4. Conclusion

According to the statistics, the Pacific Ring of Fire has a substantial amount of seismic activity and is frequently shaken by earthquakes, especially in places like Mexico, Indonesia, Japan, and the southern United States. An earthquake's depth and magnitude are inversely proportional. Regardless of their magnitude, shallow earthquakes have the potential to cause more surface damage because of how close they are to populated areas. The veracity of the data can be determined by an earthquake's DMIN and NST. Low

DMIN and high NST values signify significant and reliable recording of an earthquake occurrence, but high DMIN and low NST values could indicate potential uncertainty in the earthquake parameters. The error study revealed that, when the DMIN is equal to zero, the discrepancy between the original (MMI) and officially stated (Magnitude) values ranges from 8% to 15%. When the DMIN is more than 0, the error proportion rises to 8% to 23%. This can be explained by the dearth of adjacent seismic stations that would allow for precise measurements. The bar chart shows how the size of the earthquake and the economic damage it causes are directly related. High-magnitude seismic activity nations, like Indonesia, have sustained significant economic harm. The tree map also corroborates the fact that countries with higher earthquake magnitudes, such as Indonesia, Japan, and Papua New Guinea, also bear the most significant economic impact from these earthquakes.

In conclusion, the study of this earthquake data offers essential insights into understanding the severity, frequency, and impact of earthquakes worldwide. Understanding these patterns can aid in the development of better earthquake preparedness and response strategies and provided the significance of depth and magnitude correlation.

6. Appendix

Data Preprocessing:

```
{r}
library(readr)
library(dplyr)
library(ggplot2)


{r}
earthquake <- read_csv("E:/Masters/Data Viz/Meta Data Project/Dataset/archive (4)/earthquake_data.csv")


{r}
head(earthquake)
df <- earthquake
```

I'm planning to merge the economy send on natural disaster. So, I need years seperately from the "date_time" column to by using lubridate I'm extracting the years of the earthquake and made into an new col named as "Years".

```
{r}
head(df)


{r}
df$Year <- format(as.Date(df$date_time, format = "%d-%m-%Y "), format = "%Y")
head(df)
#"22-11-2022 02:03"


{r}


#format(as.Date(df$date_time, format =  "%d-%m-%Y %H:%M:%S"), format = "")
```

There is col named alert level in dataset which has lot of null values, so I'm writing an condition which has been referenced from "https://www.norfolk.gov/DocumentCenter/View/2556/Earthquake-Information?bidId=" it say's alert is respective to the magnitude level.

Reference from : United States Geological Survey.

{r}

#write.table(df)

df$Intensity <- ifelse(df$magnitude >= 0 & df$magnitude <= 6.8, "Moderate",

ifelse(df$magnitude >= 6.9 & df$magnitude <= 7.8, "Strong",

ifelse(df$magnitude >= 7.9 & df$magnitude <= 10, "Severe", NA)))

head(df)

 # Print the updated dataframe

Reading a Economical Damage from natural Disaster.

{r}

economy <- read_csv("E:/Masters/Data Viz/Meta Data Project/Dataset/archive (4)/economic-damage-from-natural-disasters.csv")

head(economy)

It has all Kind of Disasters, So the target values from the data set are Earthquake(2001 - 2022) and Tsunami(2001 - 2022)

First, extracting the earthquake.

{r}

df_eco <- economy %>% filter(economy$Entity == 'Earthquake')

df_eco

df_eco <- df_eco %>% filter(Year > 2001) %>% select(-Code)

head(df_eco)

Transforming the data as it requires and merging with the meta data set(Earthquake) where the year has been extracted from the date_time col., So, the Year col is common in the both of the data set.

{r}

merged_table <- merge(df, df_eco, by = "Year") %>% select(-alert) %>% rename()

head(merged_table)

Now, Tsunami

{r}

merged_table <- merged_table %>% rename("Earthquake Damage" = "Total economic damage from natural disasters")

head(merged_table)

Now. adding an another col from economy data set. for comparing the Earthquake damage to the overall economy damage across the world.

```{r}
df_eco <- economy %>% filter(economy$Entity == "All disasters")

df_eco <- df_eco %>% filter(df_eco$Year > 2001) %>% select(-Code, -Entity)

head(df_eco)

merged_table <- merge(merged_table, df_eco, by = "Year")

head(merged_table)
```

Visualization Plots

```{r}
df <- cleaneddata

trans_table <- df %>% filter(dmin == 0) %>% group_by(dmin)

trans_table$Error_Percentage <- (((trans_table$magnitude)-(trans_table$mmi))/(trans_table$magnitude)*100)

#trans_table
```

```{r}
ggplot(trans_table, aes(magnitude)) +

  geom_line(aes(y = magnitude, color = "magnitude"), linetype = "solid") +
```

````
  geom_line(aes(y = mmi, color = "mmi"), linetype = "dashed") +
  geom_smooth(aes(y = Error_Percentage, color = "Error_Percentage")) +
  scale_color_manual(values = c("magnitude" = "blue", "mmi" = "red", "Error_Percentage" = "green")) +ggtitle("More Significance (DMIN == 0)") +
  xlab("Magnitude") +
  ylab("Error percentage for MMI and Magnitude") +
  labs(color = "Data") +
  theme_minimal()
```
```{r}
df <- cleaneddata
trans_table <- df %>% filter(dmin > 0) %>% group_by(dmin)
trans_table$Error_Percentage <- (((trans_table$magnitude)-(trans_table$mmi))/(trans_table$magnitude)*100)
trans_table
ggplot(trans_table, aes(magnitude)) +
  geom_line(aes(y = magnitude, color = "magnitude"), linetype = "solid") +
  geom_line(aes(y = mmi, color = "mmi"), linetype = "dashed") +
   geom_smooth(aes(y = Error_Percentage, color = "Error_Percentage")) +
  scale_color_manual(values = c("magnitude" = "blue", "mmi" = "red", "Error_Percentage" = "green")) + ggtitle("Less Significance (DMIN > 0)") +
  xlab("Magnitude") +
  ylab("Error percentage for MMI and Magnitude") +
  labs(color = "Data") +
  theme_minimal()
```
````
````