Atal Bihari Vajpayee-Indian Institute of Information Technology and Management Gwalior

विश्वजीवनामृतं ज्ञानम्

Natural Disasters Data Analysis Report

Submitted to Dr. Santosh Singh Rathore

Group Members :

Mekala Bhavana	2021BCS-041
Nelluri Pavithra Sai Lakshmi	2021BCS-049

Table of Contents

1. Introduction

Natural disasters, ranging from earthquakes to hurricanes, have had a tremendous impact on human societies, influencing the course of history. This study examines the patterns and implications of events that occurred from 1900 and 2021. The EM-DAT database, a vital resource, meticulously records a century of catastrophes, providing insights into their types, locations, and socioeconomic repercussions.

1.1 How natural disasters have occured ?

Natural disasters are events caused by natural processes of the Earth that result in significant and often catastrophic consequences for people, wildlife, and the environment. These events can be sudden or develop over time.

It's crucial to remember that human actions, such as urbanization, deforestation, and climate change, can cause or worsen certain natural disasters. It is essential to comprehend the mechanisms and causes underlying these occurrences in order to create plans to lessen their effects and boost readiness.

1.2 Types of Natural Disasters

Here are some common types of natural disasters and their causes:

Earthquakes : An abrupt release of energy from the Earth's crust results in seismic waves, which in turn cause earthquakes. Tectonic plate movement beneath the Earth's surface is typically the cause of this energy release.

Hurricanes, Typhoons and Cyclones : The same meteorological occurrence goes by all of these names: tropical cyclones. When there is adequate heat and moisture in the atmosphere, they form over warm ocean waters. These systems intensify into strong storms as a result of Earth's rotation.

Tornadoes : The combination of warm, humid air and chilly, dry air can cause a tornado to form. Rotating columns of air arise as a result of the instability this impact causes in the atmosphere.

Floods : Numerous things can cause floods, such as excessive rain, storm surges, snowmelt that happens quickly, or the failure of levees or dams. The overabundance of water overwhelms the natural drainage systems, resulting in extensive flooding.

Volcanic Eruptions : When magma beneath the Earth's crust is released, volcanic eruptions take place. This is possible because of the tectonic plate movement, which causes pressure to release and magma to erupt onto the surface of the Earth.

Wildfires : High temperatures, dry weather, and the presence of flammable elements like dry vegetation can all contribute to wildfires. Wildfires can also be started by human activity, lightning strikes, or volcanic activity.

Droughts : Extended stretches of unusually low rainfall are known as droughts, and they can cause water shortages. Climate trends, temperature swings, and human actions like deforestation and excessive water resource extraction can all have an impact on them.

Landslides : Large amounts of earth material fall quickly and violently down a slope, causing landslides. Landslides can be caused by strong winds, earthquakes, volcanic eruptions, and human activity such as deforestation.

Tsunamis : Generally speaking, underwater landslides, volcanic eruptions, or earthquakes cause tsunamis. Strong ocean waves that can travel great distances are produced by the displacement of water brought on by these processes.

2. Motivation

The motive for examining natural catastrophes from 1900 to 2021 is to better understand the changing nature of these events, their frequency, and the differing levels of devastation they impose on different places. Researchers, policymakers, and emergency responders can obtain useful insights into the elements that contribute to the occurrence and severity of natural disasters by analyzing historical data. This knowledge is required for the development of successful readiness, response, and recovery strategies.

3. Objectives

- Data Collection: Obtain and clean data from the EM-DAT database, focusing on the attributes supplied.

- Exploratory Data Analysis (EDA): Use descriptive statistics, data profiling, and visualizations to better understand the features of the dataset.

- Examine the temporal trends of natural disasters to detect any patterns or abnormalities.

- Geospatial Analysis: Use geographical information (latitude and longitude) to map the global distribution of disasters.

- Categorical Analysis: Sort and classify disasters into kinds, subtypes, and subsubtypes.

- Impact Assessment: Determine the severity of a disaster by looking at total deaths, injuries, affected populations, and economic damages.

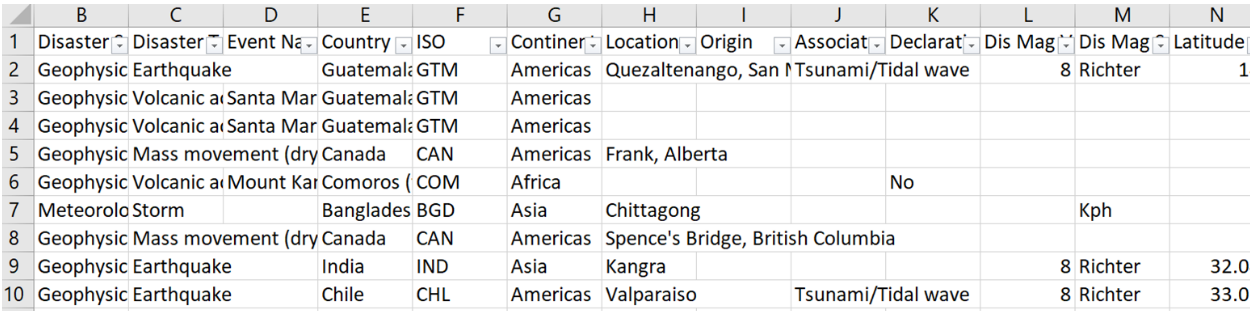
- Response and assistance Analysis: Examine the efficacy of response mechanisms such as OFDA responses, appeals, and assistance contributions.

4. Descriptive Analysis

4.1 Data Collection

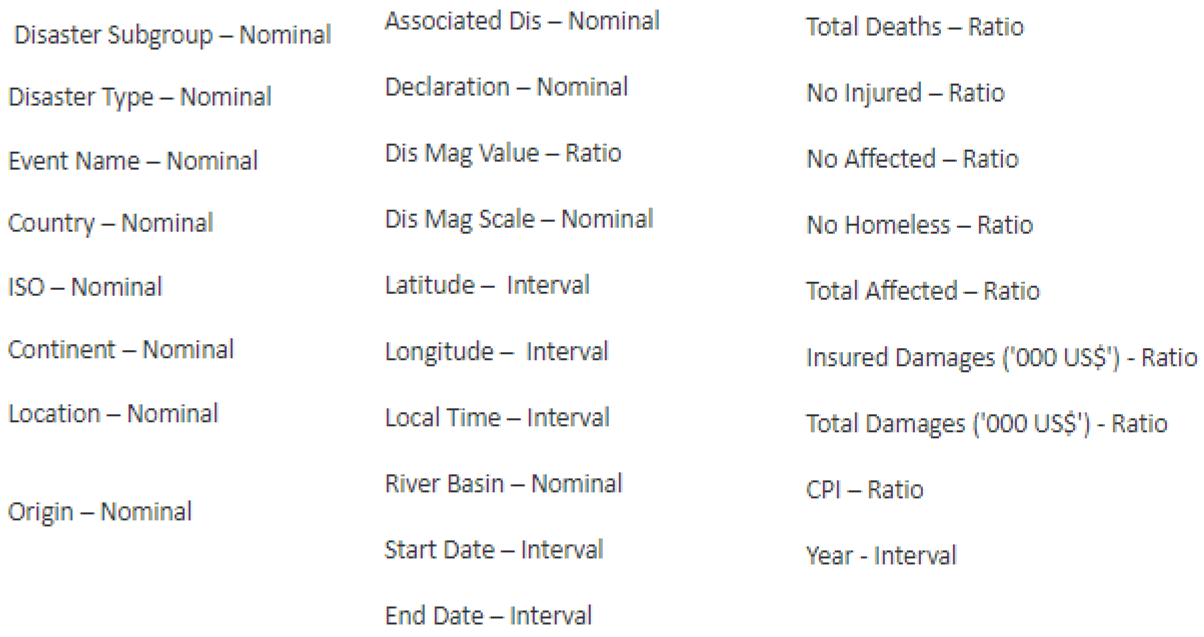
We have collected our dataset from Kaggle which is taken from the EMDAT database. EMDAT is an emergency events database.
EM-DAT contains essential core data on the occurrence and effects of over 22,000 mass disasters in the world from 1900 to the present day.

Dataset before Cleaning :



	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Disaster	Disaster	Event Na	Country	ISO	Continer	Location	Origin	Associat	Declarat	Dis Mag	Dis Mag	Latitude
2	Geophysic	Earthquake		Guatemala	GTM	Americas	Quezaltenango, San M		Tsunami/Tidal wave		8	Richter	1
3	Geophysic	Volcanic a	Santa Mar	Guatemala	GTM	Americas							
4	Geophysic	Volcanic a	Santa Mar	Guatemala	GTM	Americas							
5	Geophysic	Mass movement (dry		Canada	CAN	Americas	Frank, Alberta						
6	Geophysic	Volcanic a	Mount Kar	Comoros (	COM	Africa				No			
7	Meteorolo	Storm		Banglades	BGD	Asia	Chittagong					Kph	
8	Geophysic	Mass movement (dry		Canada	CAN	Americas	Spence's Bridge, British Columbia						
9	Geophysic	Earthquake		India	IND	Asia	Kangra				8	Richter	32.0
10	Geophysic	Earthquake		Chile	CHL	Americas	Valparaiso		Tsunami/Tidal wave		8	Richter	33.0

Data Types of Attributes



Disaster Subgroup – Nominal

Disaster Type – Nominal

Event Name – Nominal

Country – Nominal

ISO – Nominal

Continent – Nominal

Location – Nominal

Origin – Nominal

Associated Dis – Nominal

Declaration – Nominal

Dis Mag Value – Ratio

Dis Mag Scale – Nominal

Latitude – Interval

Longitude – Interval

Local Time – Interval

River Basin – Nominal

Start Date – Interval

End Date – Interval

Total Deaths – Ratio

No Injured – Ratio

No Affected – Ratio

No Homeless – Ratio

Total Affected – Ratio

Insured Damages ('000 US$') - Ratio

Total Damages ('000 US$') - Ratio

CPI – Ratio

Year - Interval

4.2 Data Preprocessing

An essential phase in the pipeline for data analysis and machine learning is data pretreatment. In order to prepare raw data for analysis or machine learning model training, it must be cleaned and transformed.

Data preprocessing is necessary in Handling Missing Values, Handling Noisy Data, Handling Inconsistent Data, Handling Duplicate Data, Scaling and Normalization, Encoding Categorical Variables, Feature Engineering,Handling Imbalanced Datasets, Text

Preprocessing, Reducing Dimensionality, Preparing Data for Modelling.

In our dataset,

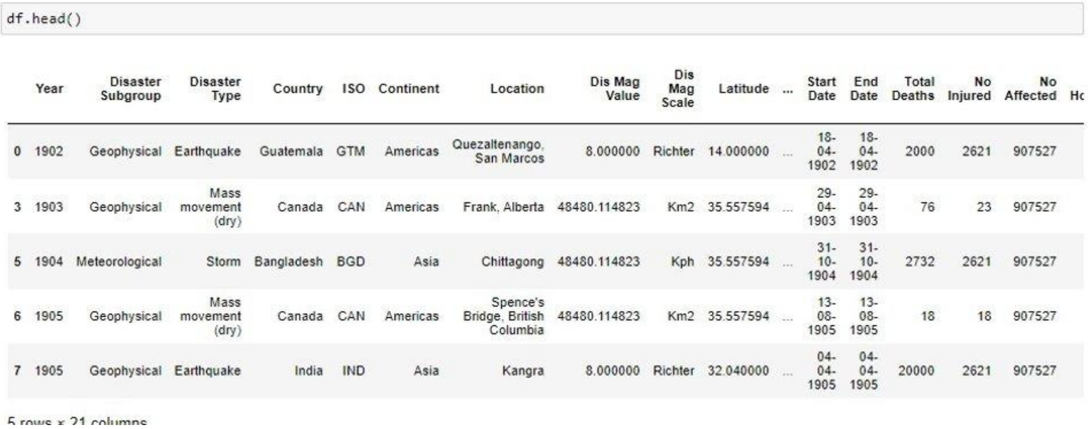
Start Day, Start Month, Start Year are combined to a single new attribute “Start Date”. Similarly for End Day, End Month, End Year combined to a single row attribute End Date.

Handling Missing Values :

For numerical values, we replaced them with mean and for categorical values, we replaced them with a common value or mode.

We Dropped redundant or uninformative columns like Seq no.

Dataset after Pre-processing :

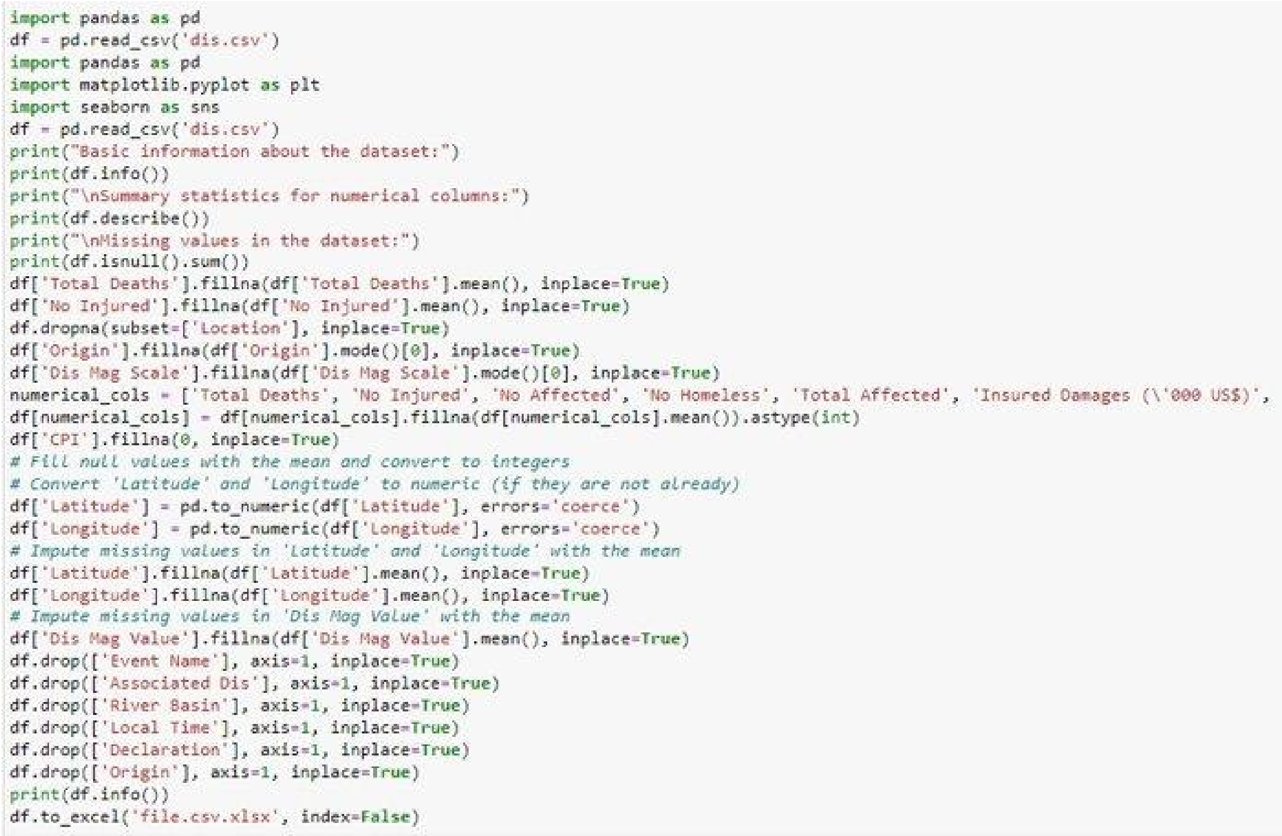


```
df.head()
```

	Year	Disaster Subgroup	Disaster Type	Country	ISO	Continent	Location	Dis Mag Value	Dis Mag Scale	Latitude	...	Start Date	End Date	Total Deaths	No Injured	No Affected	Hc
0	1902	Geophysical	Earthquake	Guatemala	GTM	Americas	Quezaltenango, San Marcos	8.000000	Richter	14.000000	...	18-04-1902	18-04-1902	2000	2621	907527	
3	1903	Geophysical	Mass movement (dry)	Canada	CAN	Americas	Frank, Alberta	48480.114823	Km2	35.557594	...	29-04-1903	29-04-1903	76	23	907527	
5	1904	Meteorological	Storm	Bangladesh	BGD	Asia	Chittagong	48480.114823	Kph	35.557594	...	31-10-1904	31-10-1904	2732	2621	907527	
6	1905	Geophysical	Mass movement (dry)	Canada	CAN	Americas	Spence's Bridge, British Columbia	48480.114823	Km2	35.557594	...	13-08-1905	13-08-1905	18	18	907527	
7	1905	Geophysical	Earthquake	India	IND	Asia	Kangra	8.000000	Richter	32.040000	...	04-04-1905	04-04-1905	20000	2621	907527	

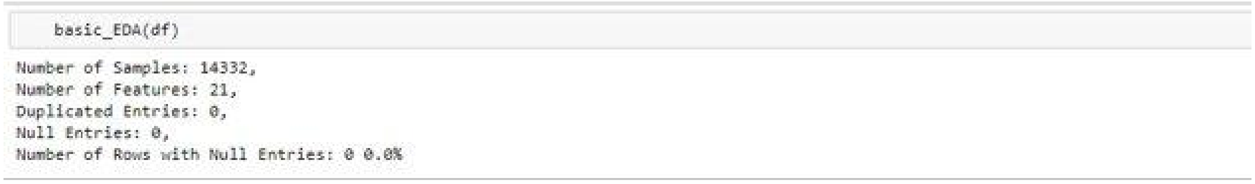
5 rows × 21 columns

We can observe that before the dataset had 31 columns and now reduced to 21 columns after performing data pre-processing.



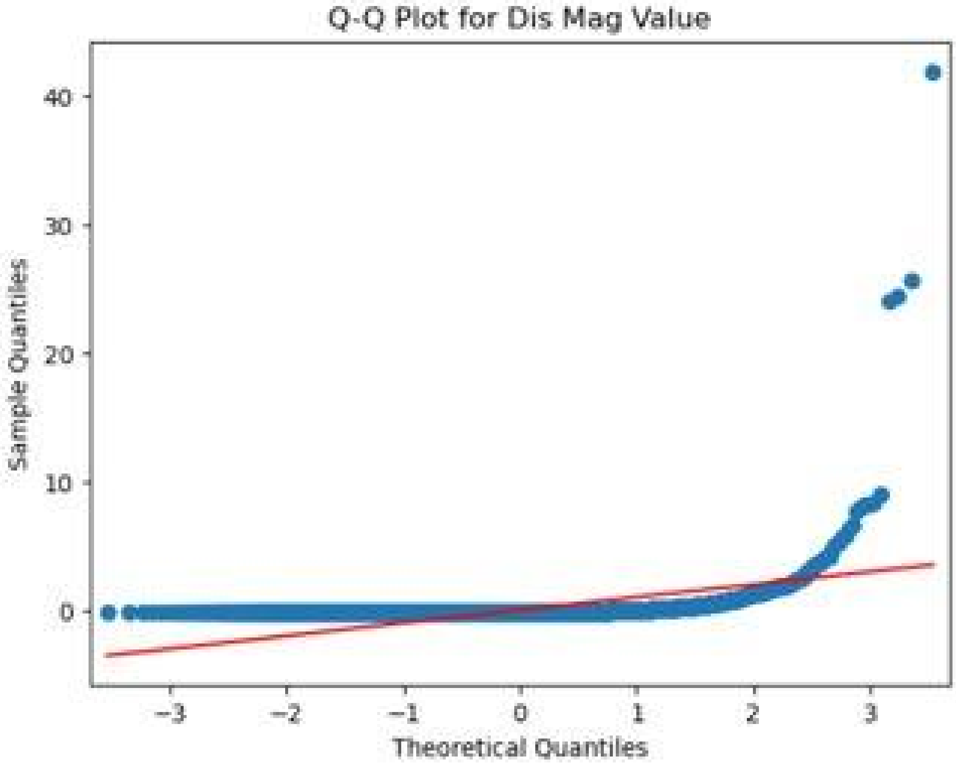
```
import pandas as pd
df = pd.read_csv('dis.csv')
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
df = pd.read_csv('dis.csv')
print("Basic information about the dataset:")
print(df.info())
print("\nSummary statistics for numerical columns:")
print(df.describe())
print("\nMissing values in the dataset:")
print(df.isnull().sum())
df['Total Deaths'].fillna(df['Total Deaths'].mean(), inplace=True)
df['No Injured'].fillna(df['No Injured'].mean(), inplace=True)
df.dropna(subset=['Location'], inplace=True)
df['Origin'].fillna(df['Origin'].mode()[0], inplace=True)
df['Dis Mag Scale'].fillna(df['Dis Mag Scale'].mode()[0], inplace=True)
numerical_cols = ['Total Deaths', 'No Injured', 'No Affected', 'No Homeless', 'Total Affected', 'Insured Damages (\'000 US$)',
df[numerical_cols] = df[numerical_cols].fillna(df[numerical_cols].mean()).astype(int)
df['CPI'].fillna(0, inplace=True)
# Fill null values with the mean and convert to integers
# Convert 'Latitude' and 'Longitude' to numeric (if they are not already)
df['Latitude'] = pd.to_numeric(df['Latitude'], errors='coerce')
df['Longitude'] = pd.to_numeric(df['Longitude'], errors='coerce')
# Impute missing values in 'Latitude' and 'Longitude' with the mean
df['Latitude'].fillna(df['Latitude'].mean(), inplace=True)
df['Longitude'].fillna(df['Longitude'].mean(), inplace=True)
# Impute missing values in 'Dis Mag Value' with the mean
df['Dis Mag Value'].fillna(df['Dis Mag Value'].mean(), inplace=True)
df.drop(['Event Name'], axis=1, inplace=True)
df.drop(['Associated Dis'], axis=1, inplace=True)
df.drop(['River Basin'], axis=1, inplace=True)
df.drop(['Local Time'], axis=1, inplace=True)
df.drop(['Declaration'], axis=1, inplace=True)
df.drop(['Origin'], axis=1, inplace=True)
print(df.info())
df.to_excel('file.csv.xlsx', index=False)
```

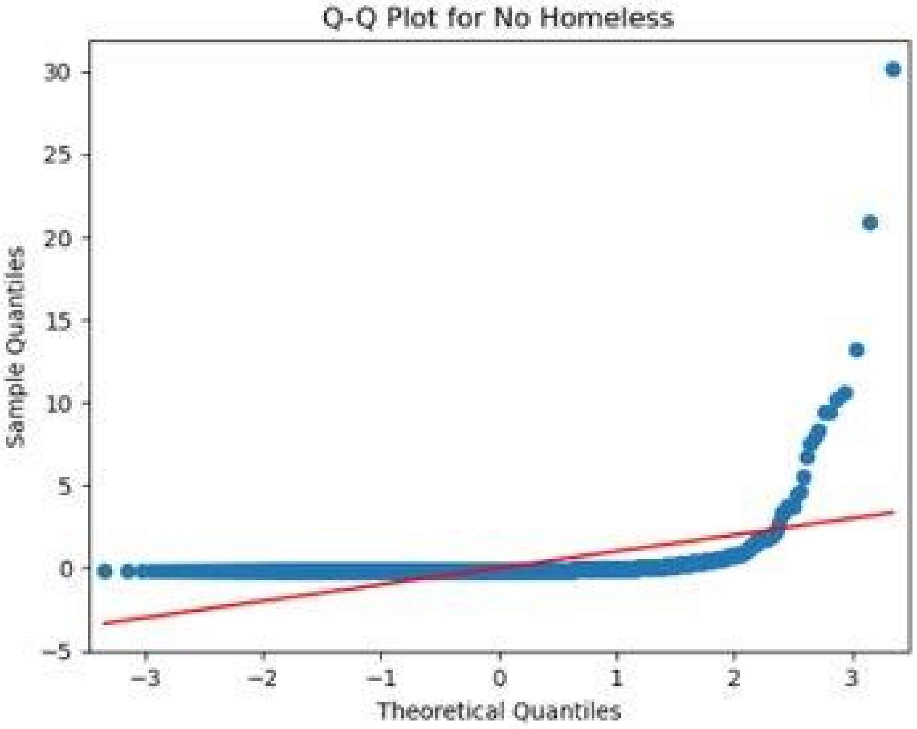
4.3 Exploratory Data Analysis

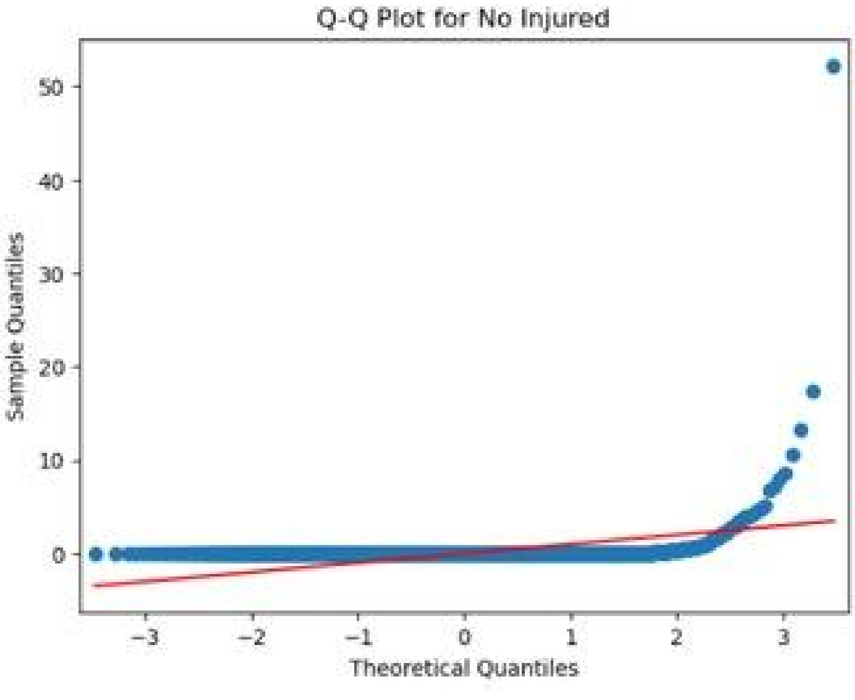


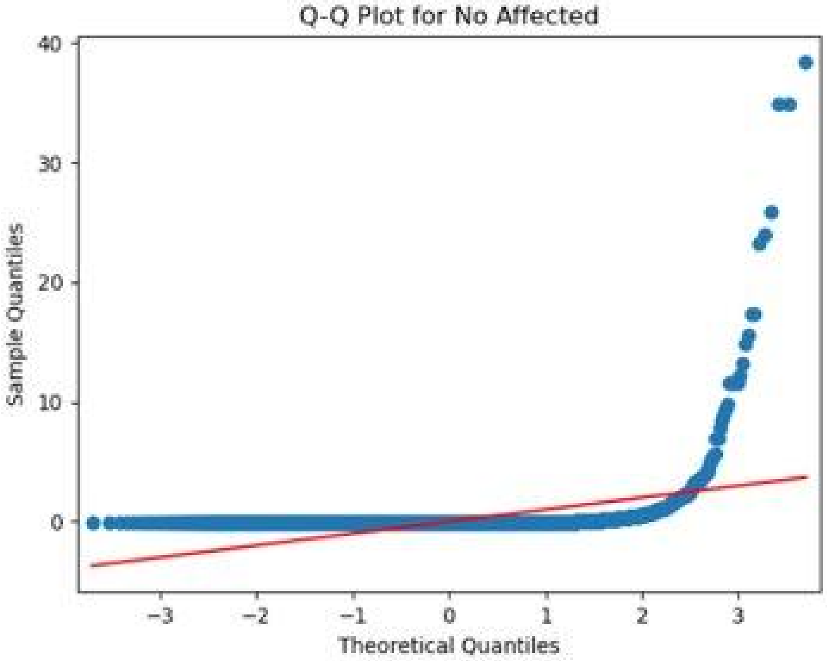
```
    basic_EDA(df)
```

```
Number of Samples: 14332,
Number of Features: 21,
Duplicated Entries: 0,
Null Entries: 0,
Number of Rows with Null Entries: 0 0.0%
```

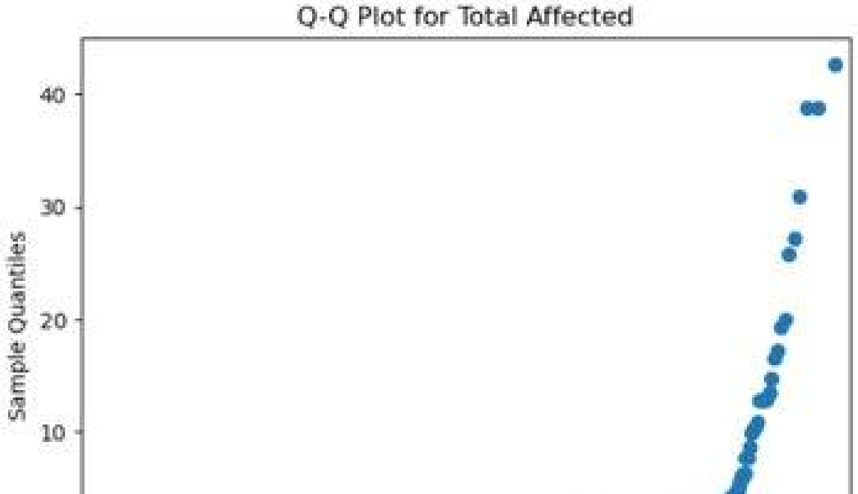



Q-Q Plot for No Homeless
Sample Quantiles
Theoretical Quantiles
30
25
20
15
10
5
0
−5
−3
−2
−1
0
1
2
3

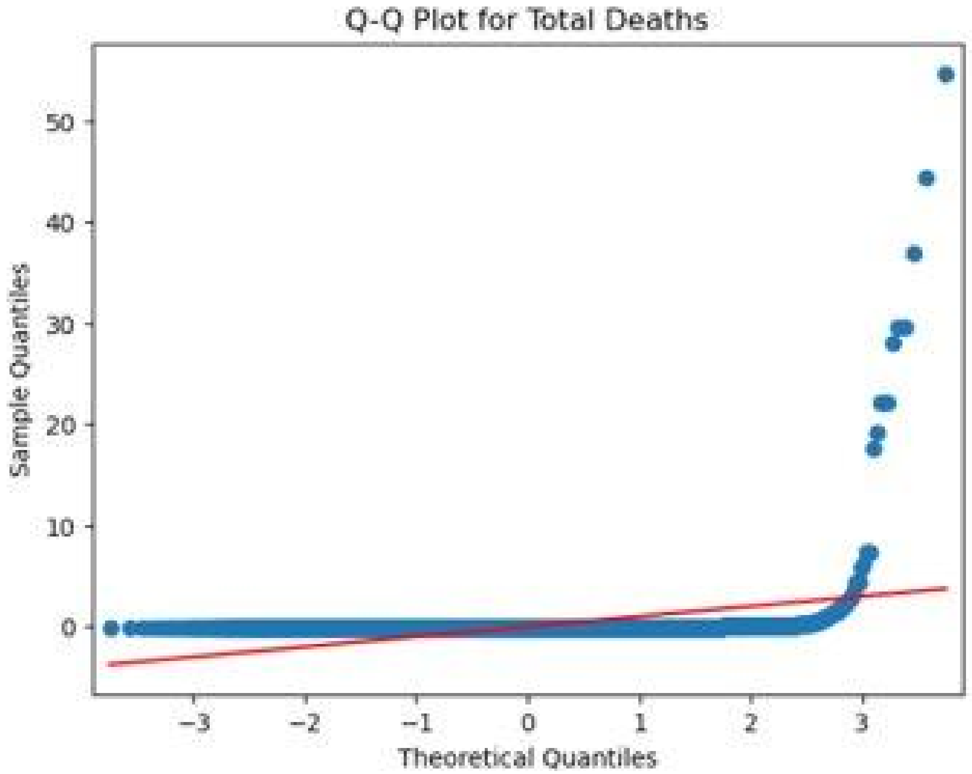

Q-Q Plot for No Injured
Sample Quantiles
Theoretical Quantiles
50
40
30
20
10
0
−3
−2
−1
0
1
2
3

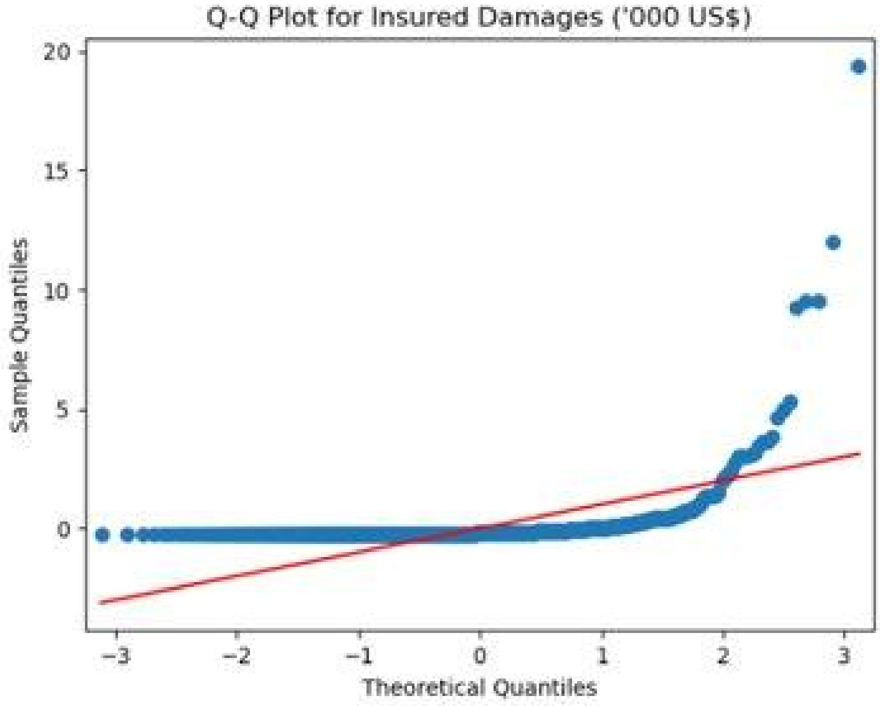


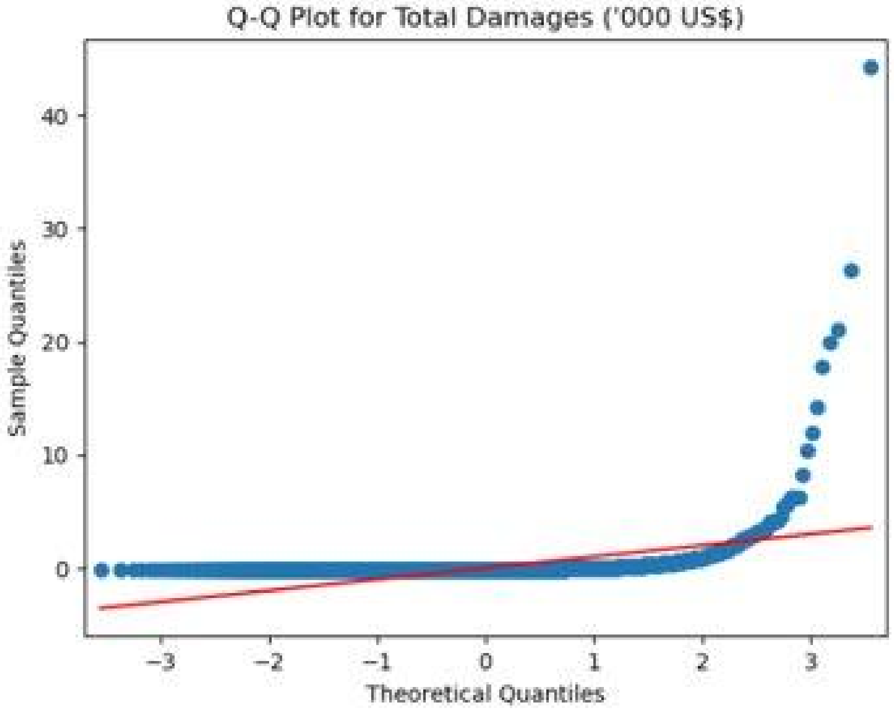
Q-Q Plot for No Affected

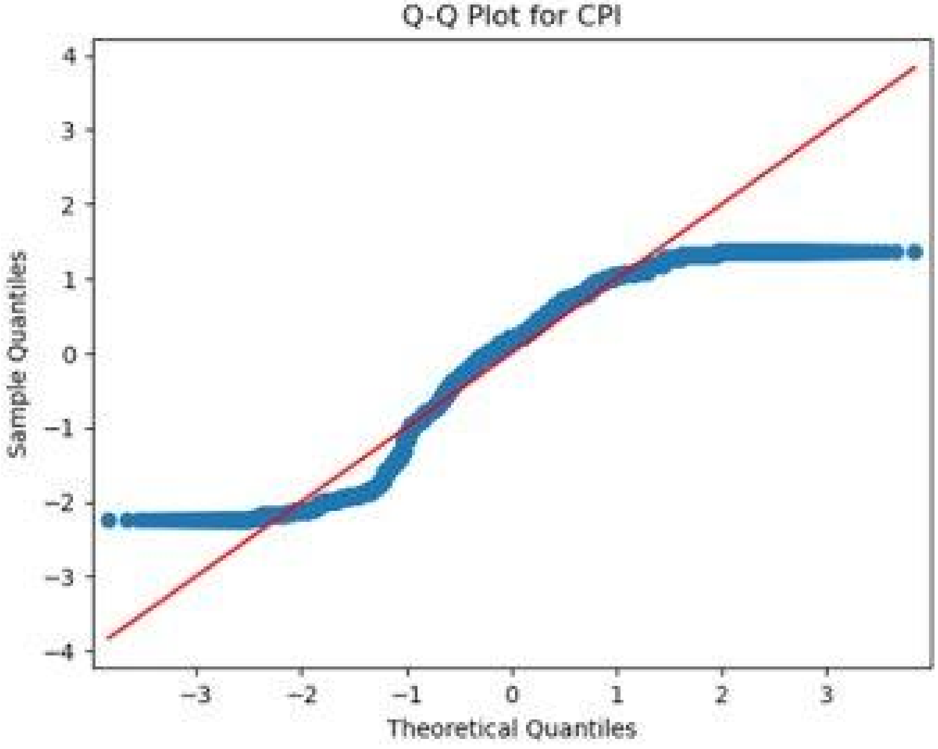


Q-Q Plot for Total Affected


Q-Q Plot for Total Deaths
Sample Quantiles
Theoretical Quantiles
0
10
20
30
40
50
−3
−2
−1
0
1
2
3


Q-Q Plot for Insured Damages ('000 US$)
Sample Quantiles
Theoretical Quantiles
0
5
10
15
20
−3
−2
−1
0
1
2
3

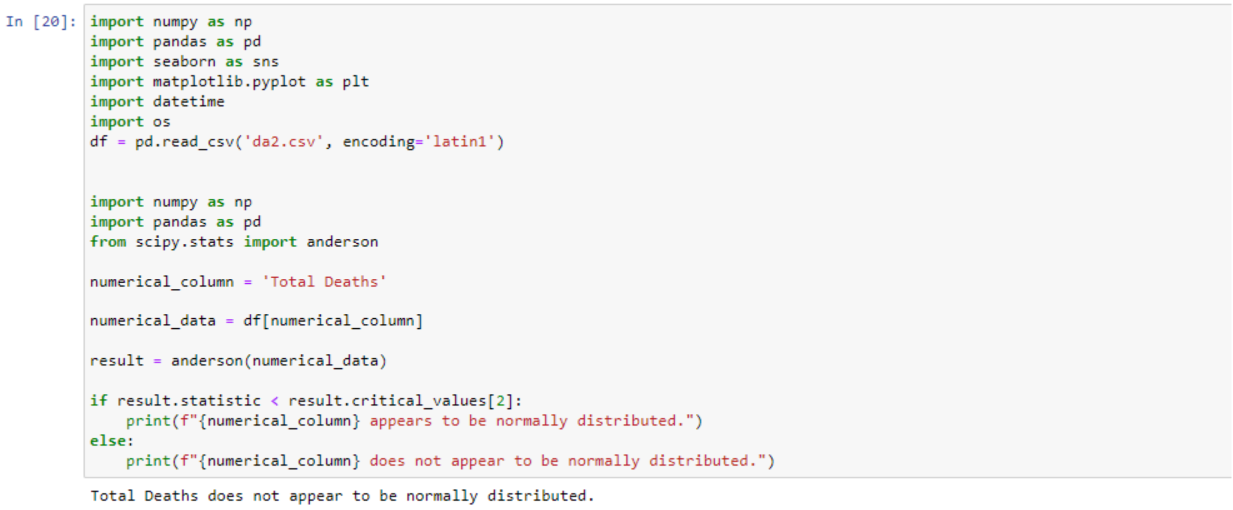

Q-Q Plot for Total Damages ('000 US$)
Sample Quantiles
Theoretical Quantiles
40
30
20
10
0
−3
−2
−1
0
1
2
3


Q-Q Plot for CPI
Sample Quantiles
Theoretical Quantiles
4
3
2
1
0
−1
−2
−3
−4
−3
−2
−1
0
1
2
3

From the above diagram, we can observe that the observed points differ from the predicted straight line proving that the distribution is not normal.

In light of these findings, non-parametric or strong statistical alternatives might be taken into account for analyses in which the normalcy assumptions are not satisfied.

We also used the **Anderson - Darling test** to check the normality of distribution.



```
In [20]: import numpy as np
         import pandas as pd
         import seaborn as sns
         import matplotlib.pyplot as plt
         import datetime
         import os
         df = pd.read_csv('da2.csv', encoding='latin1')


         import numpy as np
         import pandas as pd
         from scipy.stats import anderson

         numerical_column = 'Total Deaths'

         numerical_data = df[numerical_column]

         result = anderson(numerical_data)

         if result.statistic < result.critical_values[2]:
             print(f"{numerical_column} appears to be normally distributed.")
         else:
             print(f"{numerical_column} does not appear to be normally distributed.")
```

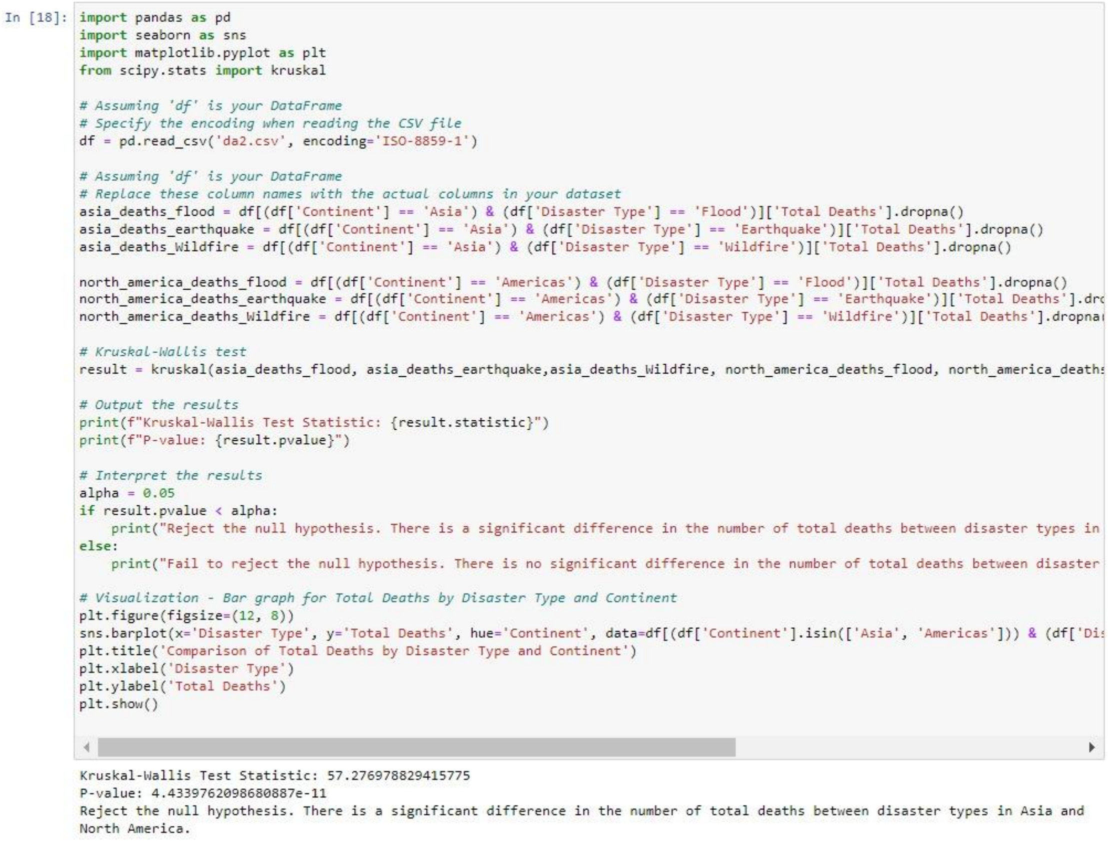
```
Total Deaths does not appear to be normally distributed.
```

Comparison of Total Deaths by Disaster Type :

A non-parametric test,Kruskal-Wallis test, is used to assess if two or more independent groups differ statistically significantly from one another.
It is used to determine whether there are statistically significant variations in the total deaths for each disaster subtype between the two continents in the context of your research comparing total deaths between disaster subtypes in Asia and North America.

Null Hypothesis (H0): The distributions of total deaths for each disaster subtype are the same across Asia and North America.

Alternative Hypothesis (H1): At least one group (disaster subtype) has a different distribution of total deaths between Asia and North America.

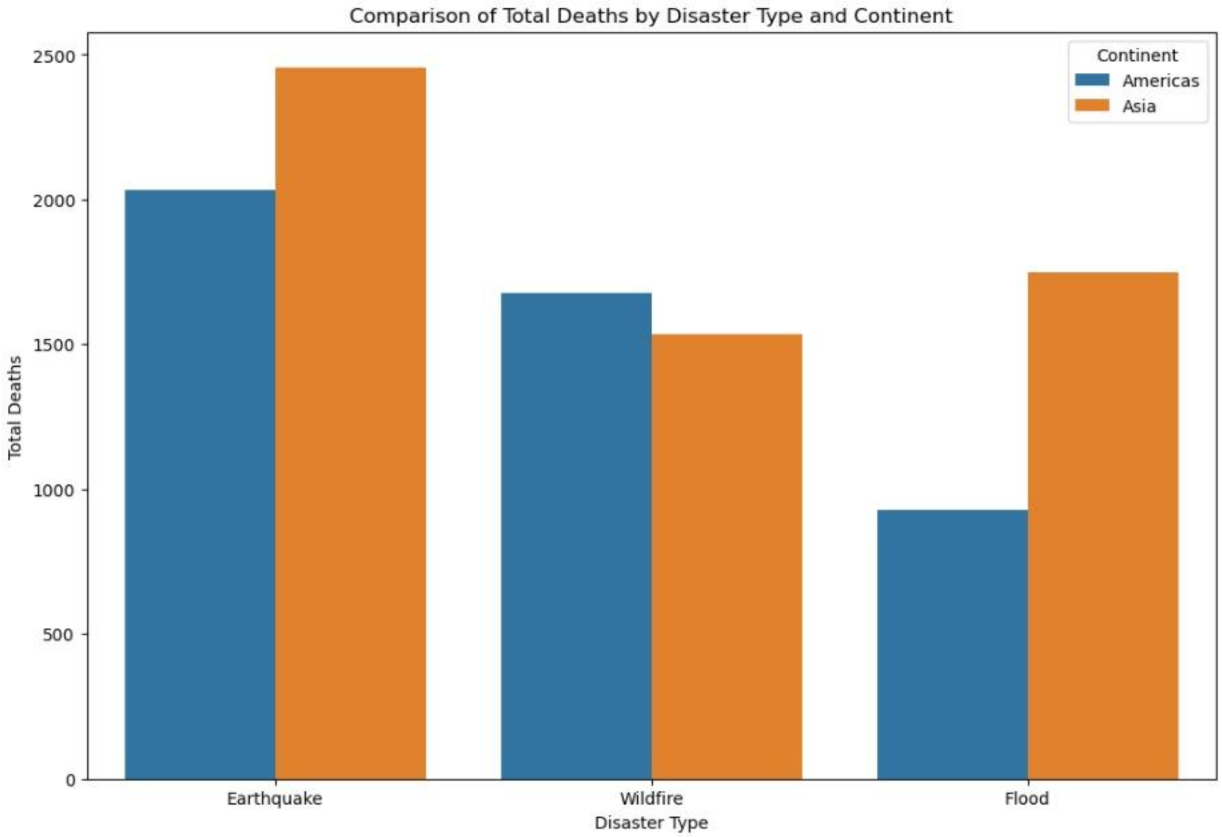


In [18]:

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import kruskal

# Assuming 'df' is your DataFrame
# Specify the encoding when reading the CSV file
df = pd.read_csv('da2.csv', encoding='ISO-8859-1')

# Assuming 'df' is your DataFrame
# Replace these column names with the actual columns in your dataset
asia_deaths_flood = df[(df['Continent'] == 'Asia') & (df['Disaster Type'] == 'Flood')]['Total Deaths'].dropna()
asia_deaths_earthquake = df[(df['Continent'] == 'Asia') & (df['Disaster Type'] == 'Earthquake')]['Total Deaths'].dropna()
asia_deaths_Wildfire = df[(df['Continent'] == 'Asia') & (df['Disaster Type'] == 'Wildfire')]['Total Deaths'].dropna()

north_america_deaths_flood = df[(df['Continent'] == 'Americas') & (df['Disaster Type'] == 'Flood')]['Total Deaths'].dropna()
north_america_deaths_earthquake = df[(df['Continent'] == 'Americas') & (df['Disaster Type'] == 'Earthquake')]['Total Deaths'].dro
north_america_deaths_Wildfire = df[(df['Continent'] == 'Americas') & (df['Disaster Type'] == 'Wildfire')]['Total Deaths'].dropna(

# Kruskal-Wallis test
result = kruskal(asia_deaths_flood, asia_deaths_earthquake,asia_deaths_Wildfire, north_america_deaths_flood, north_america_deaths

# Output the results
print(f"Kruskal-Wallis Test Statistic: {result.statistic}")
print(f"P-value: {result.pvalue}")

# Interpret the results
alpha = 0.05
if result.pvalue < alpha:
    print("Reject the null hypothesis. There is a significant difference in the number of total deaths between disaster types in
else:
    print("Fail to reject the null hypothesis. There is no significant difference in the number of total deaths between disaster

# Visualization - Bar graph for Total Deaths by Disaster Type and Continent
plt.figure(figsize=(12, 8))
sns.barplot(x='Disaster Type', y='Total Deaths', hue='Continent', data=df[(df['Continent'].isin(['Asia', 'Americas'])) & (df['Dis
plt.title('Comparison of Total Deaths by Disaster Type and Continent')
plt.xlabel('Disaster Type')
plt.ylabel('Total Deaths')
plt.show()
```

```
Kruskal-Wallis Test Statistic: 57.276978829415775
P-value: 4.4339762098680887e-11
Reject the null hypothesis. There is a significant difference in the number of total deaths between disaster types in Asia and
North America.
```

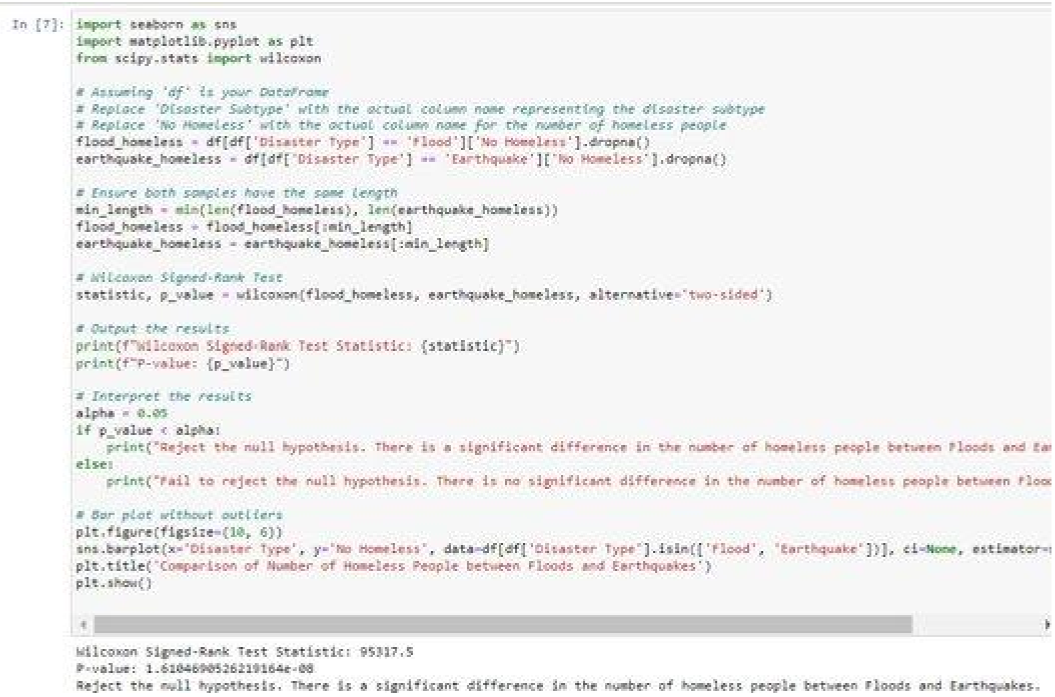
As p-value is less than 0.05 hence we reject the null hypothesis and accept the alternative hypothesis.



Comparison of Number of Homeless People between Floods and Earthquake :

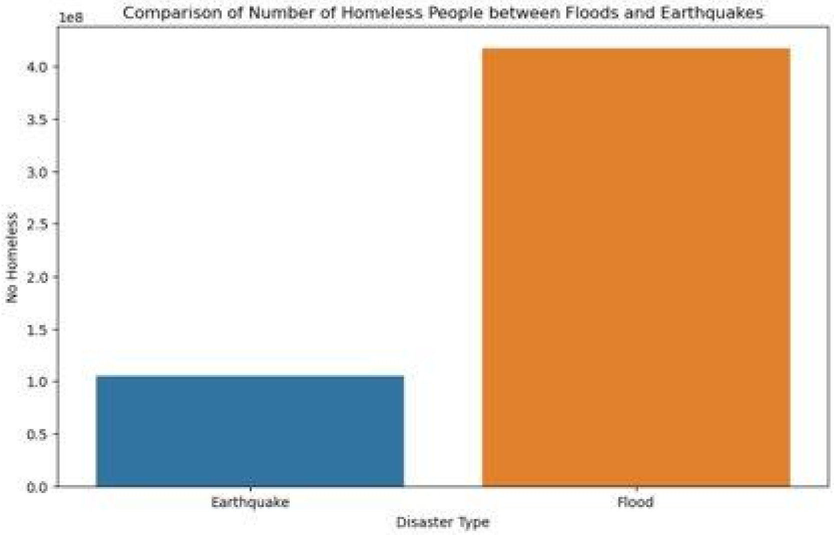
The Wilcoxon Signed-Rank Test is chosen in this context because it is a non-parametric test designed to compare two related samples.

In this case, we are comparing the number of homeless people between two types of disasters (Floods and Earthquakes), and you have pairs of related observations for the two groups.

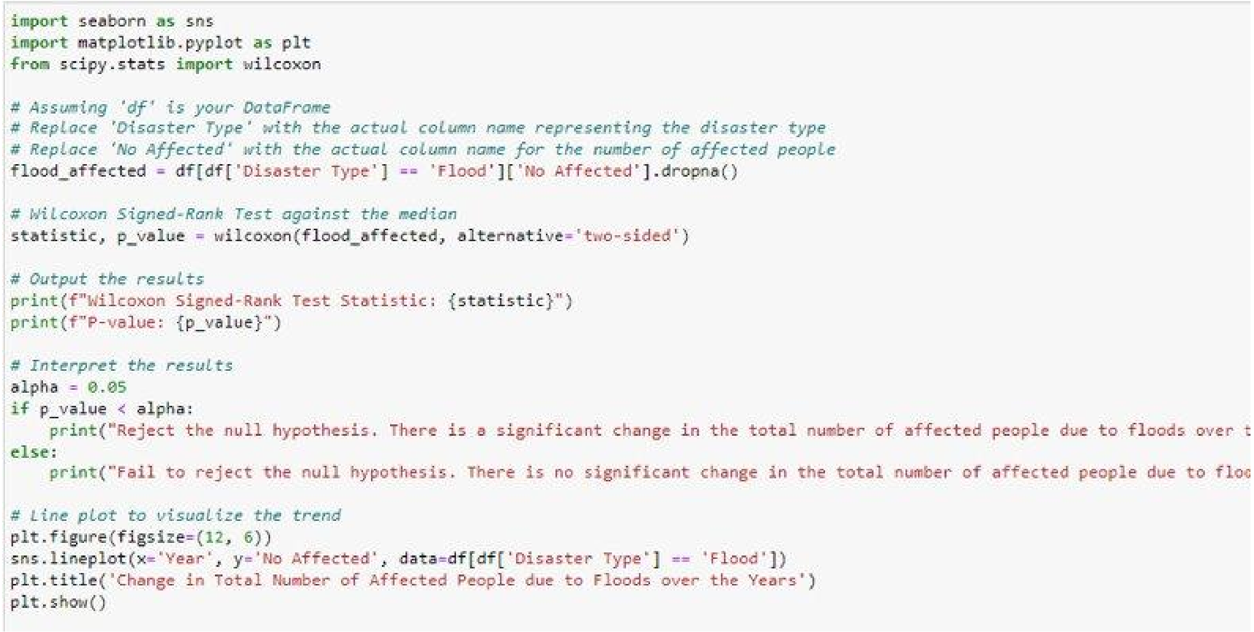


```
In [7]: import seaborn as sns
        import matplotlib.pyplot as plt
        from scipy.stats import wilcoxon

        # Assuming 'df' is your DataFrame
        # Replace 'Disaster Subtype' with the actual column name representing the disaster subtype
        # Replace 'No Homeless' with the actual column name for the number of homeless people
        flood_homeless = df[df['Disaster Type'] == 'Flood']['No Homeless'].dropna()
        earthquake_homeless = df[df['Disaster Type'] == 'Earthquake']['No Homeless'].dropna()

        # Ensure both samples have the same length
        min_length = min(len(flood_homeless), len(earthquake_homeless))
        flood_homeless = flood_homeless[:min_length]
        earthquake_homeless = earthquake_homeless[:min_length]

        # Wilcoxon Signed-Rank Test
        statistic, p_value = wilcoxon(flood_homeless, earthquake_homeless, alternative='two-sided')

        # Output the results
        print(f"Wilcoxon Signed-Rank Test Statistic: {statistic}")
        print(f"P-value: {p_value}")

        # Interpret the results
        alpha = 0.05
        if p_value < alpha:
            print("Reject the null hypothesis. There is a significant difference in the number of homeless people between Floods and Ear
        else:
            print("Fail to reject the null hypothesis. There is no significant difference in the number of homeless people between Floo

        # Bar plot without outliers
        plt.figure(figsize=(10, 6))
        sns.barplot(x='Disaster Type', y='No Homeless', data=df[df['Disaster Type'].isin(['Flood', 'Earthquake'])], ci=None, estimator=
        plt.title('Comparison of Number of Homeless People between Floods and Earthquakes')
        plt.show()
```

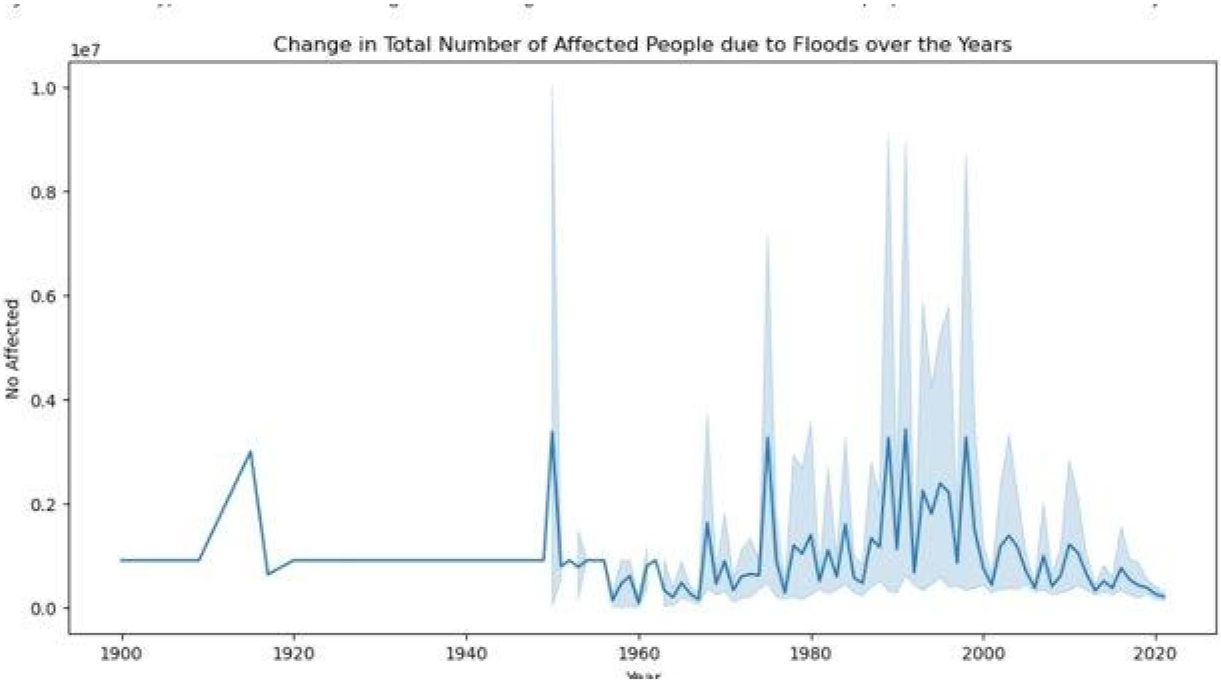
```
Wilcoxon Signed-Rank Test Statistic: 95317.5
P-value: 1.6104690526219164e-08
Reject the null hypothesis. There is a significant difference in the number of homeless people between Floods and Earthquakes.
```



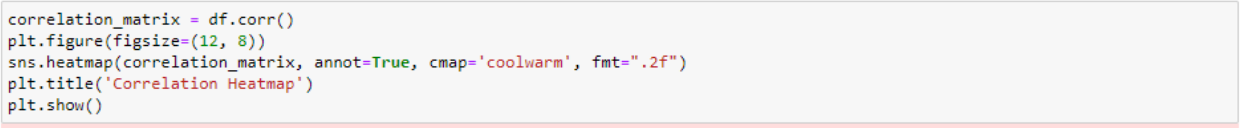
Change in Total no.of Affected People due to Floods over the years :



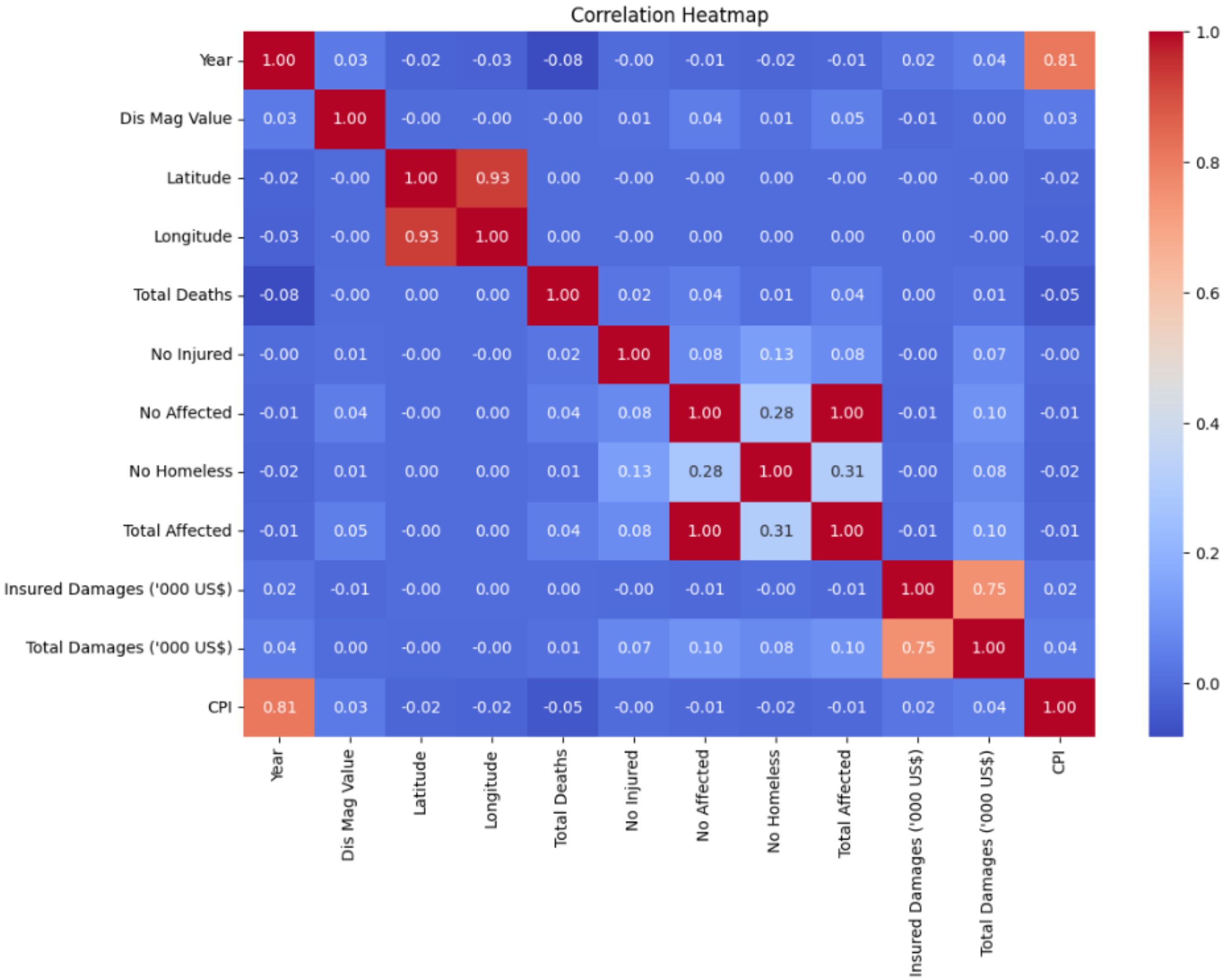
```
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import wilcoxon

# Assuming 'df' is your DataFrame
# Replace 'Disaster Type' with the actual column name representing the disaster type
# Replace 'No Affected' with the actual column name for the number of affected people
flood_affected = df[df['Disaster Type'] == 'Flood']['No Affected'].dropna()

# Wilcoxon Signed-Rank Test against the median
statistic, p_value = wilcoxon(flood_affected, alternative='two-sided')

# Output the results
print(f"Wilcoxon Signed-Rank Test Statistic: {statistic}")
print(f"P-value: {p_value}")

# Interpret the results
alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis. There is a significant change in the total number of affected people due to floods over t
else:
    print("Fail to reject the null hypothesis. There is no significant change in the total number of affected people due to floo

# Line plot to visualize the trend
plt.figure(figsize=(12, 6))
sns.lineplot(x='Year', y='No Affected', data=df[df['Disaster Type'] == 'Flood'])
plt.title('Change in Total Number of Affected People due to Floods over the Years')
plt.show()
```



Heatmap of Correlations :



```
correlation_matrix = df.corr()
plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Heatmap')
plt.show()
```



	Year	Dis Mag Value	Latitude	Longitude	Total Deaths	No Injured	No Affected	No Homeless	Total Affected	Insured Damages ('000 US$)	Total Damages ('000 US$)	CPI
Year	1.00	0.03	-0.02	-0.03	-0.08	-0.00	-0.01	-0.02	-0.01	0.02	0.04	0.81
Dis Mag Value	0.03	1.00	-0.00	-0.00	-0.00	0.01	0.04	0.01	0.05	-0.01	0.00	0.03
Latitude	-0.02	-0.00	1.00	0.93	0.00	-0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.02
Longitude	-0.03	-0.00	0.93	1.00	0.00	-0.00	0.00	0.00	0.00	0.00	-0.00	-0.02
Total Deaths	-0.08	-0.00	0.00	0.00	1.00	0.02	0.04	0.01	0.04	0.00	0.01	-0.05
No Injured	-0.00	0.01	-0.00	-0.00	0.02	1.00	0.08	0.13	0.08	-0.00	0.07	-0.00
No Affected	-0.01	0.04	-0.00	0.00	0.04	0.08	1.00	0.28	1.00	-0.01	0.10	-0.01
No Homeless	-0.02	0.01	0.00	0.00	0.01	0.13	0.28	1.00	0.31	-0.00	0.08	-0.02
Total Affected	-0.01	0.05	-0.00	0.00	0.04	0.08	1.00	0.31	1.00	-0.01	0.10	-0.01
Insured Damages ('000 US$)	0.02	-0.01	-0.00	0.00	0.00	-0.00	-0.01	-0.00	-0.01	1.00	0.75	0.02
Total Damages ('000 US$)	0.04	0.00	-0.00	-0.00	0.01	0.07	0.10	0.08	0.10	0.75	1.00	0.04
CPI	0.81	0.03	-0.02	-0.02	-0.05	-0.00	-0.01	-0.02	-0.01	0.02	0.04	1.00

In the correlation matrix, each cell color represents the strength and direction of the correlation between two variables.

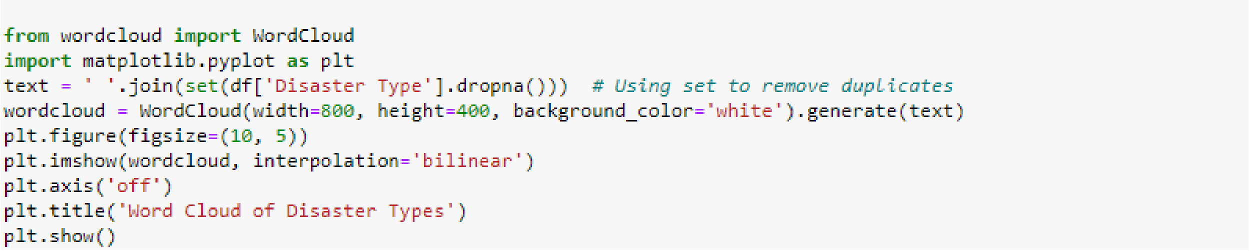
The color on the right side of the heatmap represents the strength of the correlation. Colors range from dark blue (strong negative correlation) through white(no correlation) to dark red (strong positive correlation).

The numbers inside each cell represent the correlation coefficients, which ranges from -1 to 1. Darker cells represent stronger correlations. The diagonal line consists of 1s since a variable perfectly correlates with itself.

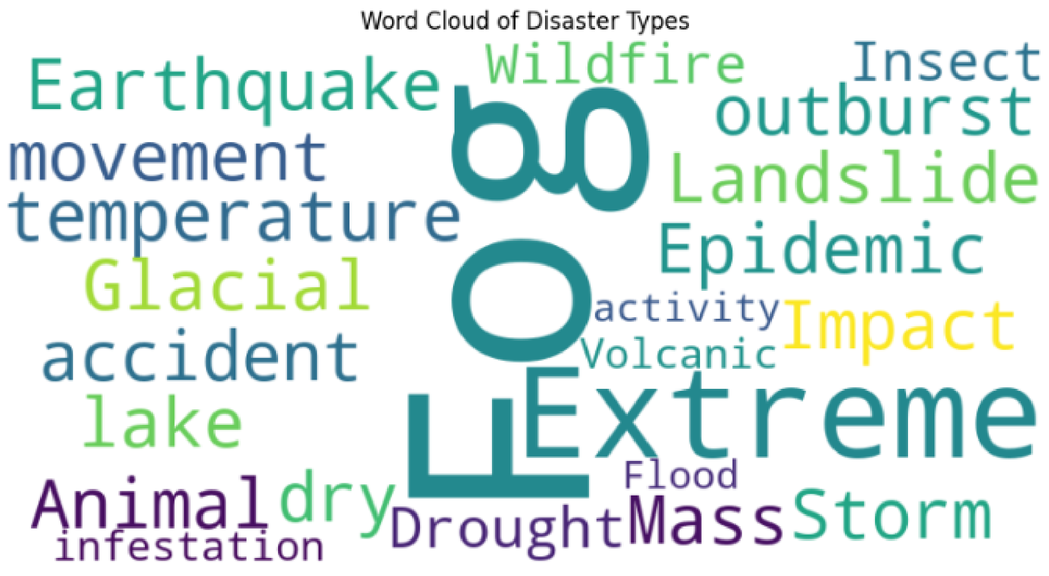
What are the most common types of Disasters?

By identifying what types of disasters occur more frequently than others. This can help in prioritizing disaster preparedness and response efforts.

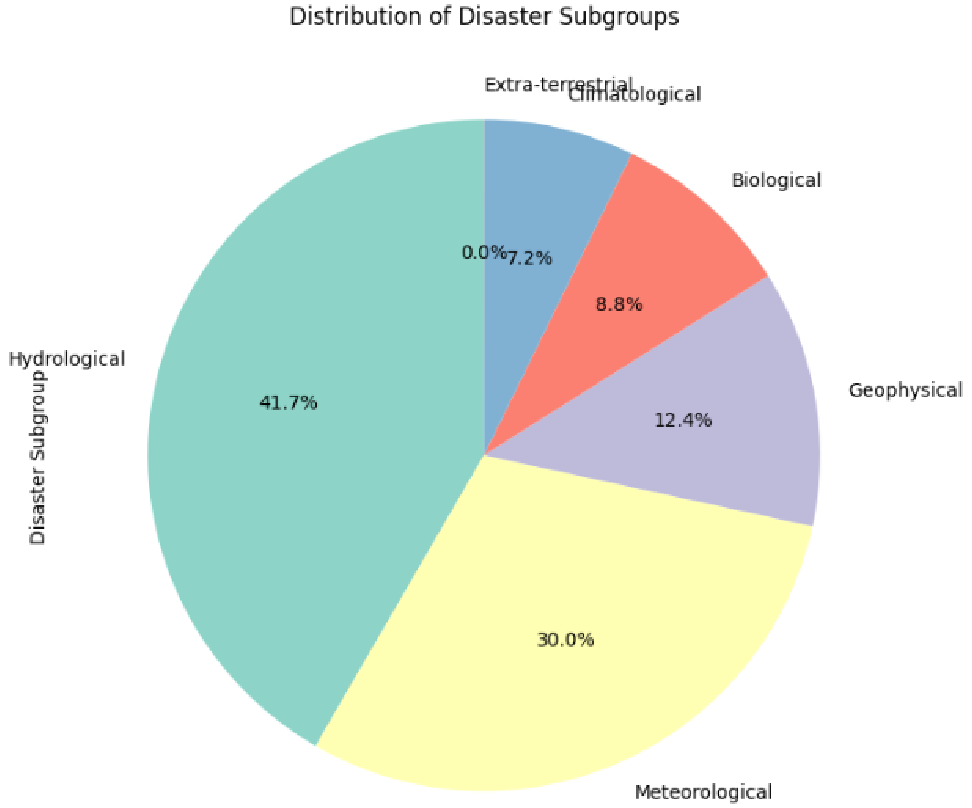
Analyze if certain types of disasters exhibit seasonal patterns. For example: Floods might be more common during the rainy season. Understanding these patterns helps in planning and resource allocation.



```python
from wordcloud import WordCloud
import matplotlib.pyplot as plt
text = ' '.join(set(df['Disaster Type'].dropna()))  # Using set to remove duplicates
wordcloud = WordCloud(width=800, height=400, background_color='white').generate(text)
plt.figure(figsize=(10, 5))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.title('Word Cloud of Disaster Types')
plt.show()
```



Distribution of Disaster Subgroups :



5. Conclusion and Future Scope

Conclusion :

Using the EM-DAT Database, we have mapped out the history of natural disasters from 1900 to 2021. Our analysis has revealed important new information on the intricate relationship between human cultures and the destructive force of nature. The histories of earthquakes, storms, and other calamities have provided a clear picture of the effects on a global scale, highlighting trends, difficulties, and the adaptability of local populations everywhere.

Future Scope :

- Predictive Modeling: Implement machine learning algorithms for predictive modeling.
- Impact of Climate Change: Take data on climate change into account when evaluating how it affects the frequency and severity of natural catastrophes.Examine any connections between climatic trends and the incidence of disasters.
- Interdisciplinary Investigations: For a comprehensive understanding, work in conjunction with social scientists, geographers, and environmental scientists. Combine information from several sources to improve the analysis.

6. References

https://www.kaggle.com/datasets/jnegrini/emdat19002021/data

https://ourworldindata.org/natural-disasters